

Revenue Management and Resource Allocation for Communication Satellite Operators

by

Markus Guerster

M.Sc., Technical University of Munich (2017)

B.Sc., Technical University of Munich (2015)

Submitted to the Department of Aeronautics and Astronautics
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

© 2020 Massachusetts Institute of Technology. All rights reserved.

Author
Department of Aeronautics and Astronautics
June 30, 2020

Certified by
Prof. Edward F. Crawley
Professor, Aeronautics and Astronautics
Thesis Supervisor

Certified by
Dr. Peter Belobaba
Principal Research Scientist, Aeronautics and Astronautics
Member, Thesis Committee

Certified by
Henry Weil
Senior Lecturer, Sloan School of Management
Member, Thesis Committee

Certified by
Dr. Joël Grotz
Senior Manager, Adaptive Resource Control System Lead, SES Engineering
Member, Thesis Committee

Certified by
Prof. Nicholas Roy
Professor, Aeronautics and Astronautics
Member, Thesis Committee

Accepted by
Zoltan Spakovszky
Professor, Aeronautics and Astronautics
Chair, Graduate Program Committee

[This page is intentionally left blank]

Revenue Management and Resource Allocation for Communication Satellite Operators

by

Markus Guerster

Submitted to the Department of Aeronautics and Astronautics
on June 30, 2020, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Aeronautics and Astronautics

Abstract

This dissertation presents the first academic study on Revenue Management (RM) for broadband communication satellite (satcom) operators. It proposes a satcom RM framework to structure, automate, and optimize operators' demand and capacity management.

New entrants, increasing demand for data, digital payloads, and new phased array technologies are likely to remake the current satcom landscape. One of the challenges, old and new, operators face is how to manage demand and capacity. This work finds that airlines' tiered pricing and seat inventory control (known as RM) offers insights to the satcom market.

The satcom industry shares many characteristics with the airline industry, such as inflexible capacity, low marginal sales cost, perishable inventory, heterogeneous customers, and variable and uncertain demand. Generally, those characteristics favor the implementation of an RM system.

However, four unique challenges are discovered that require the extension of existing RM frameworks by a resource management part: First, the unit of capacity (Watts) is not the unit of demand (Mbps). Second, the resource allocation is an optimization problem itself. Third, available capacity is uncertain based on resource usage. And fourth, existing Service Level Agreements (SLAs) do not fully leverage the new satellites' flexibility.

The dissertation proposes algorithmic solutions to each of these four challenges. It specifically focuses on the resource allocation process and its optimization of user terminal grouping, routing, frequency assignment, and power allocation.

Finally, the value of the proposed satcom RM framework is demonstrated by applying it to a satellite operator's data. The results show that dynamic resource allocation frees up considerable capacity (38% in the analyzed scenario), which operators can monetize into additional revenues. More sophisticated RM algorithms lift the revenues between 4-7% compared to heuristic pricing policies.

Thesis supervisor: Prof. Edward F. Crawley
Title: Professor, Aeronautics and Astronautics

Acknowledgements

This dissertation closes the final chapter of my Ph.D. journey, a 3-year stage in my life that I will be forever grateful for. I believe there is no better word than “journey” to describe what a Ph.D. is for many. It is a collection of ups and downs that I had the fortune to share with truly inspiring people. While the “ups” are what is detailed in this document, the “downs” are what genuinely refined me as a person and made this journey so precious.

Even though the dissertation bears my name, it could not have existed without the support of many. I want to express my sincere gratitude to the following people:

First, I want to thank my advisor, Edward Crawley, who took a chance with me as a visiting student, and convinced me to go on the Ph.D. journey, an opportunity I would have regretted to miss. His undoubted trust and confidence in my abilities have been a catalyst for much of the work I achieved. I am especially grateful for Ed’s wisdom beyond research that he synthesized down to precise principles. They gave me guidance, and I am sure they will support me in many future decisions. Thank you, Ed, and I trust you will pass on your wisdom to many others!

Furthermore, I want to recognize the central role that Bruce Cameron played. He has been my go-to place for pragmatic and structured advice when my thoughts were unclear and unfocused, especially in the earlier stages of my research. I also want to thank Bruce for his service as a thesis reader, whose treasured comments improved the document significantly. Thank you, Bruce, for all the support that you gave me, and I am sure your advice will substantially shape the research of many more!

Moreover, I want to show my appreciation for my committee members, Peter Belobaba, Henry Weil, Joël Grotz, and Nicholas Roy:

Peter, whose dissertation was the first on the topic of Revenue Management for the airline industry, taught me much about Revenue Management. It has been my fortune to have him on my committee. His expertise and input to my dissertation are invaluable. Thank you, Peter, and I hope you are as excited as I am to observe how the future of Revenue Management might unfold in a new industry!

Henry opened up my horizon with his acquaintance of economic and business principles in the airline and telecommunication industry from which I could draw many analogies. His patience and willingness to discuss with me the dynamics between technology and markets helped to shape the intersection between

engineering and economics research. Thank you, Henry, for your time commitment and the viewpoints I learned from you!

Joël brought the practical industry view into this dissertation. His ability to give balanced guidance between research and industry needs has been instrumental in my journey. He always carved out the time of his busy schedule to meet with me, generously sharing his thinking and vast satellite communication know-how. Thank you, Joël, without your involvement, this research would not have left academic contemplations!

Nick shaped my thinking of academic research. He taught me how to identify research gaps, frame research, and make appreciated contributions. Thank you, Nick, for taking the time to teach and for serving on my committee!

A special acknowledgment also goes to my second thesis reader, Vincent Chan, who shared his tremendous knowledge about telecommunication with me and encouraged my research. Thank you, Vincent!

Additionally, I want to show my sincere appreciation for SES. They did not only fund my research but also opened their doors, continuously provided feedback along the journey, and gave me access to crucial data. In particular, I want to thank Valvanera Monero for her relentless efforts to make this sponsor relationship exemplary. She introduced me and my ideas to many people within SES. Each of these conversations shaped my dissertation. Especially, I want to acknowledge the discussions with Jean-Pierre Choffray, Hira Muzammil, Javier Trujillo, Markus Gross, Stefano Andrenacci, Ruy Pinto, Marek Wojcik, Stefan Brak, Chris Pendleton, Emil Ibrahim, James Mowat, Antonio Bove, Gilles Schutz, Yuvan Bhikajee, Gianluigi Morelli, and Steven Chacko. Thank you, Valva, and thank you to everybody else at SES!

I also want to take the opportunity to thank the team at MIT with whom I had the treasure to work on this project Gemini. Special thanks go to Juanjo Garau Luis. He was there from the beginning and was instrumental in supporting the navigation of the uncertainties associated with a newly launched project and team. Thank you, Juanjo, I have no doubt you will create something unique in the future! Furthermore, I want to recognize the contributions made by Nils Pachler. He developed algorithmic solutions to central problems and helped me streamlining my ideas in numerous intellectual stimulating conversations. Thank you, Nils, and I cannot wait to see how the project moves forward with you! Additionally, I want to express my appreciation for Rubén Alinque Diane, Damon Jones, Skylar Eiskowitz, Thomas Finn, and Willian Torous who were part of the project over the years.

I would also like to thank the people from the AeroAstro department, especially from 33-409. They made the office a warm and inspirational place for sometimes 7 days a week. I want to mention particularly Íñigo del Portillo Barrios, who freely shared his knowledge about satellite communication with me. Further thanks go to George Lordos, Johannes Norheim, Suhail Alsalehi, Matthew Moraguez, and Alex Trujillo; and to the more recent members Sydney Dolan, Katie Carroll, and Yaniv Mordecai. Thanks to you all for making 33-409 a second home. To Amy Jarvis, Beth Marois, and Ping Lee, thank you for the invaluable administrative support.

I am very fortunate to be able to share this Ph.D. experience with wonderful friends. Axel Garcia Burgos is one of the most driven people I know. His entrepreneurship and ability to navigate is something I am still aspiring to. Axel, I am lucky to have you as my friend. Yaroslav Menshenin was the first person I met when I arrived at MIT, and I am thankful for his mentorship, friendship, and thoughtfulness. A big thank you also to my friends at A&M, who made Texas home when I was away from MIT. And thank you to all of my roommates at 177 over the years. Appreciation also goes to my friends back in Germany. A special mention deserves Florian Mueller. He has been a friend for many years and a consistent source of unlimited support, thoughtfulness, and joy. Flo, thank you! Thanks also to Thomas Hofmeister for his unconditional friendship throughout the years.

Finally, and very importantly, I want to thank my family: my parents, Petra and Alfred, and my brother Tobias. My mum gave me everything as a child without ever asking for something in return. My dad passed on the engineering spirit and set an example to aspire to. I have learned, and I am still learning so many things from him. Tobias, your drive to accomplish is inspiring, I cannot wait to see your future unfold! Ultimately, I want to thank my partner Samalis Santini, who brings the joy of life into my day and is my daily support. Samalis, I cannot wait to go on the next journey with you!

Reading Guide

This dissertation is a multidisciplinary work that contains contributions in different areas. While the expressed concepts of the framework are interconnected, the format of a written document requires a linear and sequential discussion. Therefore, the chosen structure might not fit every reader's objective. Depending on their background and interest, they might wish only to read parts of the document or change the order:

Satcom executives and other executive summary readers might wish to start with Section 3.2 and then transition directly to the simulation summary Sections 8.7 – 8.9. If the reader desires a deeper understanding of how the framework works, Section 3.5 provides an introduction.

Satcom capacity managers/researchers might decide to start with Section 1.5 to get an overview of the research objectives, skim or skip Chapter 2 and come back later to it, and focus in Chapter 3 on the satcom aspects and the proposed framework (i.e., skim Section 3.3). At the end of Chapter 3, the reader should recognize which of Chapters 4 to 7 is of most value. One option is to continue directly with the analysis Chapter 8 to grasp the framework's possible benefits and challenges.

HTS satcom resource allocation managers/researchers might choose to start straight with the resource allocation in Chapter 5 (especially if the reader is more technically focused). They might then refer back to Chapter 4 on the satcom simulator, which could be a useful and practical discussion of how to structure a satcom simulation environment. If the reader wants to expand the scope, Section 1.5 and the first two Sections of Chapter 3 provide an introduction to the economic and business connections to the resource allocation challenge.

Revenue Management researchers might want to start with the introduction Chapter 1 to give them relevant satcom background. Their key Chapter is likely 3, which discusses the satcom Revenue Management framework, especially Sections 3.2 – 3.5. The reader might further find the key results and conclusions from Chapter 8 noteworthy.

Market dynamics managers/researchers might wish to read the introduction Chapter 1 first and then transition to market dynamics discussion in Chapter 2. Some readers might find it interesting to continue with Chapter 7 on novel SLAs and market segmentation. An introduction of the complete picture of the framework is given by Sections 3.2 and 3.5. With that background, Chapter 8 provides some additional market-related discussion, in particular Sections 8.5.2 and 8.8.

Table of Content

| | |
|---|-----------|
| Acknowledgements | 4 |
| Reading Guide | 7 |
| Table of Content | 8 |
| 1 Introduction | 13 |
| 1.1 The satellite communication industry | 14 |
| 1.2 Technology push – phased arrays and digital processors..... | 16 |
| 1.3 Pull from demand – high growth for bidirectional and bursty demand | 18 |
| 1.4 Changing landscape of satellite operators..... | 19 |
| 1.5 General objectives and expected contributions | 23 |
| 1.6 Outline of the dissertation..... | 26 |
| 2 Satcom Market Dynamics | 29 |
| 2.1 Description of the model and assumptions..... | 32 |
| 2.2 Validation..... | 38 |
| 2.3 Sensitivity analysis | 39 |
| 2.4 Effects of changing the delays | 42 |
| 2.5 Vertically integrated market..... | 47 |
| 2.6 Conclusions and implications..... | 48 |
| 3 Satcom Revenue Management Framework | 51 |
| 3.1 Primer on Revenue Management..... | 52 |
| 3.2 Applicability of RM for satcom operators..... | 55 |
| 3.3 Review of RM systems in other industries..... | 58 |
| 3.3.1 Generic across industries..... | 58 |
| 3.3.2 Airline | 59 |
| 3.3.3 Hotels..... | 60 |
| 3.3.4 Rental cars | 61 |
| 3.3.5 Air cargo | 64 |
| 3.3.6 Internet Services..... | 67 |
| 3.3.7 Telecommunications | 68 |
| 3.4 Summary of reviewed industries and identification of the challenges | 71 |
| 3.4.1 Weatherford and Bodily’s taxonomy..... | 71 |
| 3.4.2 Taxonomy to compare RM characteristics across industries | 72 |
| 3.4.3 Four major challenges in satcom RM | 76 |
| 3.5 Proposed Satcom Revenue Management framework..... | 77 |
| 3.5.1 Demand management..... | 78 |
| 3.5.2 Resource management..... | 80 |
| 3.5.3 Interaction between demand and resource management..... | 82 |
| 3.6 Illustration of the working principles and potential gains | 83 |
| 3.6.1 Setup of the simulation and assumptions | 83 |
| 3.6.2 Resource allocation | 84 |
| 3.6.3 Satcom simulator..... | 85 |
| 3.6.4 Customer usage history | 86 |
| 3.6.5 Available capacity forecaster..... | 87 |

| | | |
|----------|---|------------|
| 3.6.6 | Customer purchase history..... | 88 |
| 3.6.7 | Parametrized SLA menu | 88 |
| 3.6.8 | Customer elasticity estimation | 89 |
| 3.6.9 | Pricing optimization..... | 90 |
| 3.6.10 | Results..... | 91 |
| 3.6.11 | Discussion | 92 |
| 3.7 | Two main applications of the RM framework | 94 |
| 3.8 | Summary and contributions | 95 |
| 4 | Satcom Simulator | 97 |
| 4.1 | Constellation, satellites, amplifiers..... | 99 |
| 4.2 | Beams, carriers | 101 |
| 4.3 | Link budgets..... | 101 |
| 4.4 | Gateways, user terminals, users, SLAs..... | 106 |
| 4.5 | Simulation, results | 107 |
| 4.6 | Validation..... | 109 |
| 4.7 | Summary..... | 111 |
| 5 | Resource Allocation..... | 113 |
| 5.1 | Literature review | 114 |
| 5.1.1 | Cognitive radios | 114 |
| 5.1.2 | Scheduling for data relay satellites..... | 116 |
| 5.1.3 | Power allocation..... | 116 |
| 5.1.4 | Frequency assignment..... | 118 |
| 5.1.5 | Joint power allocation and frequency assignment..... | 120 |
| 5.1.6 | User terminal grouping..... | 121 |
| 5.1.7 | Beam shape | 122 |
| 5.1.8 | Metrics..... | 123 |
| 5.1.9 | Optimization techniques | 125 |
| 5.2 | Summary of literature and identification of gaps..... | 128 |
| 5.3 | Specific objectives..... | 130 |
| 5.4 | Resource allocation process | 131 |
| 5.5 | Grouping of user terminals..... | 135 |
| 5.5.1 | Finding the minimum number of beams through edge clique cover | 135 |
| 5.6 | Routing..... | 138 |
| 5.6.1 | Algorithm description..... | 138 |
| 5.6.2 | Results | 143 |
| 5.7 | Frequency assignment..... | 148 |
| 5.7.1 | Frequency assignment by first-fit heuristic | 148 |
| 5.7.2 | Extension to include gateway beams | 151 |
| 5.8 | Power..... | 152 |
| 5.9 | Application of the resource allocation process | 154 |
| 5.9.1 | Simulation setup..... | 154 |
| 5.9.2 | Results and discussion | 156 |
| 5.9.3 | Conclusions and insights..... | 162 |
| 5.10 | Summary and contributions..... | 163 |
| 6 | Available Capacity Forecaster | 165 |
| 6.1 | Probabilistic regression of user traffic | 166 |

| | | |
|----------------------|--|------------|
| 6.1.1 | Literature review | 167 |
| 6.1.2 | Gaussian process regression..... | 169 |
| 6.1.3 | Stochastic process of normal random variables..... | 170 |
| 6.2 | Determine power density at satellite level..... | 172 |
| 6.3 | Summary and contributions | 174 |
| 7 | Novel SLAs..... | 175 |
| 7.1 | Review of current SLAs in satcom..... | 176 |
| 7.2 | Challenges..... | 180 |
| 7.2.1 | Operator perspective | 180 |
| 7.2.2 | Customer perspective..... | 182 |
| 7.3 | Specific objectives..... | 183 |
| 7.4 | Review of SLAs in similar industries..... | 183 |
| 7.5 | Novel SLAs | 185 |
| 7.6 | Market segmentation and mapping to novel SLAs..... | 189 |
| 7.6.1 | Identification of the bases of customer segmentation in satcom | 193 |
| 7.6.2 | Derived segmentation | 194 |
| 7.6.3 | Mapping of novel SLAs to segmentation | 199 |
| 7.7 | Summary and contributions | 202 |
| 8 | Application of the Satcom RM Framework..... | 203 |
| 8.1 | Input data and assumptions | 206 |
| 8.1.1 | User traffic data..... | 206 |
| 8.1.2 | Elasticities | 208 |
| 8.1.3 | Gateways | 212 |
| 8.1.4 | Space segment..... | 212 |
| 8.2 | Resource allocation for the baseline | 213 |
| 8.3 | General pricing optimization approach | 219 |
| 8.3.1 | Sorted list of prices and binary search | 220 |
| 8.3.2 | Algorithmic approach for computing the ordered list of prices..... | 221 |
| 8.4 | Monetizing the available capacity through existing customers..... | 226 |
| 8.4.1 | Selling more capacity through existing SLAs | 226 |
| 8.4.2 | Selling additional products | 232 |
| 8.5 | Monetizing the available capacity through new customers..... | 238 |
| 8.5.1 | Unlocking affordability elasticity | 238 |
| 8.5.2 | Increasing market share | 249 |
| 8.6 | Challenge of uncertain elasticity information..... | 256 |
| 8.7 | Summary of the four analyses | 260 |
| 8.8 | Implications on the market when operators adopt RM | 265 |
| 8.9 | Summary and conclusions | 271 |
| 9 | Conclusions, Contributions, and Future Work..... | 273 |
| 9.1 | Conclusions..... | 274 |
| 9.2 | Contributions | 278 |
| 9.3 | Future work | 279 |
| Appendix..... | | 281 |
| A. | What is system dynamics? | 281 |
| B. | Weatherford's [5] taxonomy..... | 283 |
| C. | Industry comparison tables..... | 283 |

| | |
|---|------------|
| D. Primer on price elasticity of demand | 285 |
| E. Frequency allocation for t = 3 hours | 287 |
| F. Service providers SLA menu | 289 |
| G. List of used MODCODs | 291 |
| H. Result of user grouping and frequency assignment for SpaceX's Starlink | 292 |
| List of Acronyms..... | 293 |
| List of Figures | 294 |
| List of Tables | 299 |
| List of Algorithms | 301 |
| List of Boxes | 301 |
| Bibliography..... | 303 |

1

Introduction

In 1978, after the deregulation of the airline market, the entrance of low-cost carriers resulted in severe competition for all major airlines. American Airlines responded with the introduction of the Ultimate Super Saver fares to stay competitive with the new low-cost entrants in 1985 [1].

On the one hand, selling all seats at this lowest fare would lead to maximum utilization of the aircraft. However, it does not necessarily maximize revenues, since passengers with a higher willingness-to-pay divert to lower fares [2]. On the other hand, selling seats only at the highest fair would maximize yield, but might lead to an underutilized aircraft, and thus not maximizing revenues either. The non-trivial optimum lies in-between these two extremes. Realizing this, American Airlines implemented what is now known as *Revenue/Yield Management (RM)*. RM is commonly defined as “*the process of allocating the right type of capacity to the right kind of customer at the right price so as to maximize revenues or yield*” [3].

After the airline industry, hotels and rental cars were the early adopters of RM with many others following [4]. According to Weatherford [5], Talluri [6], and Kimes [3], the *six conditions* that favor RM are:

- Capacity is inflexible
- Capacity costs are high compared to marginal sales cost
- Inventory is perishable
- Customers are heterogeneous and can be segmented
- Demand is variable and uncertain
- Organization has data and information system infrastructure

We find that these six conditions apply to broadband satellite communication (satcom) operators (rationalization in Section 3.2), therefore suggesting that RM has the potential to provide value to satcom operators. To the best knowledge of the author, this is the first academic work on RM for broadband satcom of this scope.

Despite the fit of these RM conditions to satcom, we cannot easily apply existing frameworks from other industries due to the unique characteristics of satcom. We identify *four unique challenges* that an operator needs to overcome to make an RM system work (and dedicate one Chapter of this dissertation to each):

1. unit of demand is not unit of capacity
2. resource allocation is an optimization problem itself
3. uncertain available capacity based on resource usage
4. existing SLAs do not fully leverage the new satellites' flexibility.

As we discuss in the following, several disruptive changes are underway in satcom, making it an ideal time for an integrated capacity management and pricing approach, such as Revenue Management – the central idea of this dissertation.

1.1 The satellite communication industry

We divide the satellite communication industry into two main sectors: first, the unidirectional broadcasting of mostly video content (e.g., direct-to-home television). Second, the growing sector of bidirectional broadband connectivity. The first area is often called Broadcasting Satellite Services (BSS) and we group Fixed Satellite Services (FSS) and Mobile Satellite Services (MSS) into bidirectional broadband [7]. Figure 1-1 shows a comparison of those two areas.

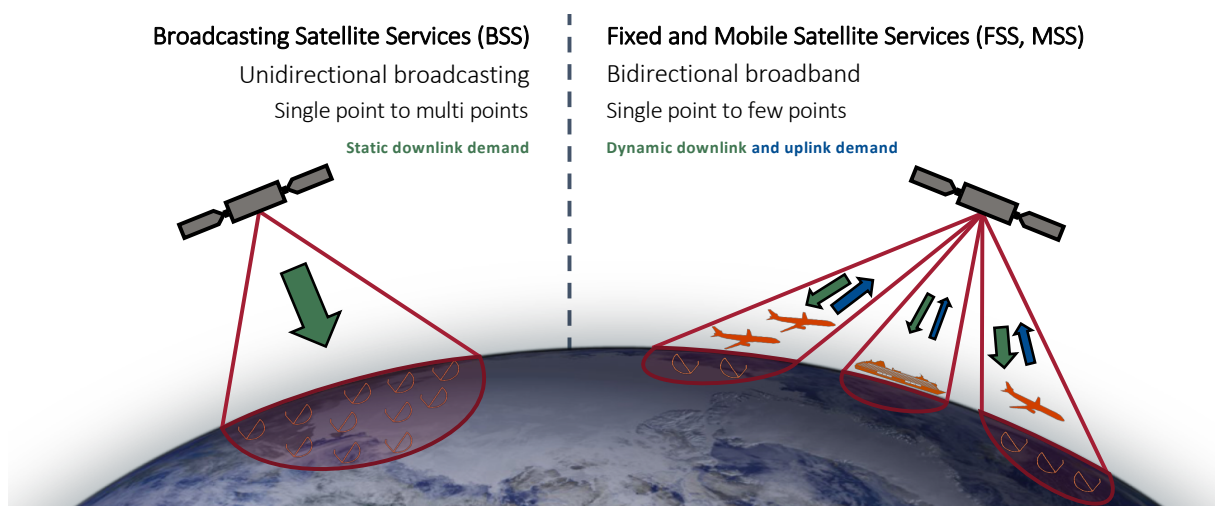


Figure 1-1: Comparison of the two major segments in the satellite communication industry

Traditionally, the main business of satcom operator is the broadcasting of television content. An additional feature that distinguishes the two sectors is the time dependency of the data rates. For broadcasting, the unidirectional downlink data rate is predictable and static (e.g., the number of television channels is known and does not change daily).

In contrast, for broadband connectivity, the more unpredictable demand of individual users results in fluctuating data rates. As shown in Section 3.2, the users' demand behavior is especially characterized by diurnal patterns. Therefore, the dynamic control of resources onboard satellites is becoming increasingly necessary to follow changing demand in the future.

Over the last decade, the consumers' consumption is shifting from broadcasted content towards on-demand streaming, resulting in a shrinking satcom broadcasting sector. Therefore, the whole industry is moving towards the growing broadband sector, i.e., providing Internet access to airplanes, cruise and cargo ships, backhauling of Wi-Fi and 4G/5G hotspots in remote areas, and internet backbone trunking. The focus of this dissertation is the growing sector of bidirectional broadband Internet access through satellites, which undergoes disruptive changes (more details in Section 1.4).

In particular, over the last few years, that transition towards broadband gained momentum with several companies entering the market. The Federal Communications Commission (FCC) received 11 applications from commercial companies for new high-throughput (HT) non-geostationary satellites [8], specifically large LEO constellations, such as from Telesat [9], OneWeb [10] and SpaceX [11]. In addition to those newly proposed mega-constellations, established players in the communication market expect to launch Medium Earth Orbit (MEO) and Geostationary Orbit (GEO) satellites to provide broadband connectivity specifically. ViaSat is planning to launch Viasat-3 to provide multiple Mbps broadband access by 2020 [12]. SES is launching a fleet of O3b mPower spacecrafts to supplement their O3b MEO satellites within the next years [13], as well as SES-17 – a high-throughput GEO satellite that will provide connectivity to America and the Atlantic Ocean [14].

All of these new constellations fall into the class of high-throughput satellites (HTS), which have “of the order of 100 Gbps” according to the International Telecommunication Union (ITU) [15, p. 3]. These new satellites will operate in various orbits and in different sizes. They encompass one critical technological development that will change how satcom operators manage their satellites: *flexible payloads*. Transparent digital processors and phased arrays enable this flexibility and allow for the dynamic

allocation of resources. Depending on the technical implementation, the degree of flexibility ranges from adjustable power only to full control over frequency assignment, beam pointing, and beam shape.

In the 1990s, companies such as Iridium, Globalstar, Orbcomm, and Teledesic launched large projects to launch a large LEO constellation [16-19]. At this time their main focus was voice services, but also rather slow (compared to speeds today) broadband Internet access. However, market uptake was poor. Several digests [16-18] identified a combination of causes: unexpected large cellular build-out shrank the target market, technological limitations and poorly identified customers (line-of-sight required between satellite and handheld device), and generally poor operational execution. However, over the past two decades, we observe a technology push (Section 1.2) that complements a pull from demand (Section 1.3). These trends indicate that both the technology and market are poised for competitive satellite broadband connectivity.

1.2 Technology push – phased arrays and digital processors

The first commercial satellite, Intelsat I, was launched in 1965 [20, 21] with a mass of 70 kg, a power of 40 W and a data throughput equivalent to one television (TV) channel (see Figure 1-2).

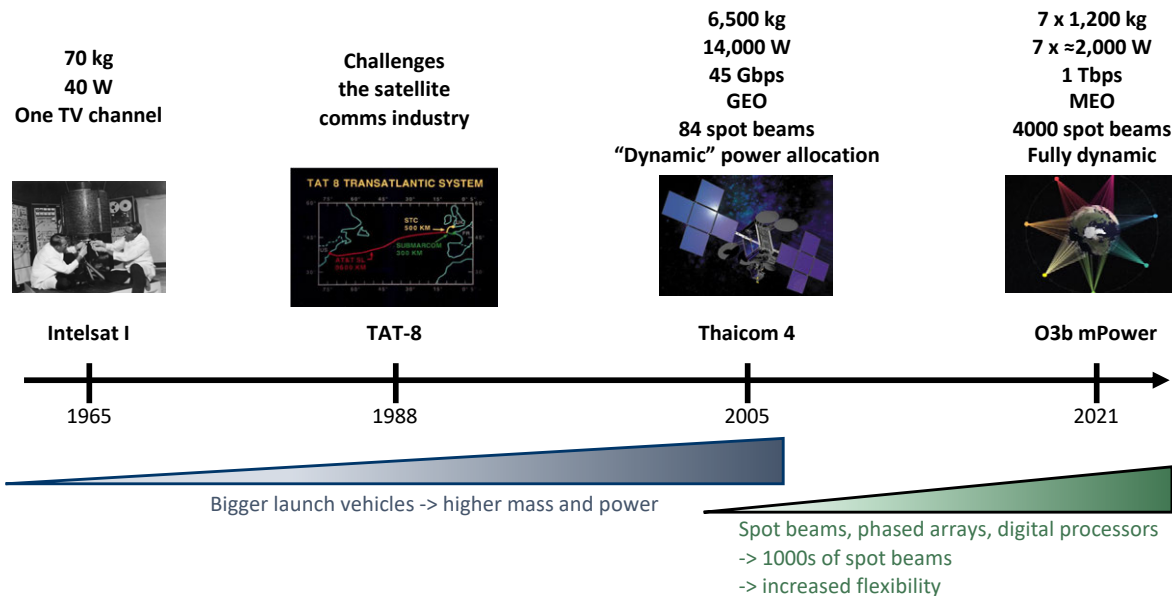


Figure 1-2: Development of satellite technology over the last five decades

Its primary usage was providing trans-Atlantic telephony services. During that time, satellites were the most cost-effective way to connect over the Atlantic. At around 1988, TAT-8 [22], a trans-Atlantic copper cable, provided a more cost-efficient way for communication services between Europe and North America. Such terrestrial solutions are the primary competition for satellite communications today. The worldwide revenues of communication services were roughly \$1,250 Billion in 2016 [23]. The entire

communication satellite industry (unidirectional broadcasting and bidirectional broadband) generated around \$125 Billion [24], so around 10% of the total market. Out of that 10% in satcom, the television broadband accounts for most of it, with broadband connectivity having a (growing) share of $\approx 16\%$ (annual global revenues of around \$20 Billion). Combining these two shares, satellite broadband connectivity accounted for only $\approx 1.6\%$ of the total generated revenues for communication services in 2016.

Forty years after Intelsat I, Thaicom 4 marked another milestone as the first high-throughput satellite [25].

There are two trends to point out:

First, compared to Intelsat I, the mass increased by almost two orders of magnitude with an even higher increase in power. The throughput increased four orders of magnitude from one TV channel (around 3.5 Mbps) to 45 Gbps. One of the key technologies that allowed for that increase is the availability of more capable launch vehicles [22]. They supported a higher mass, which allows for more power generation on-board the satellite and, therefore, higher throughput (note that there were also additional technological advancements that supplemented that trend).

Second, the number of beams increased by almost two orders of magnitudes from 1 (Intelsat I) to 84 spot beams enabling more efficient bidirectional broadband connections. In addition to this development, Thaicom 4 had the first version of dynamic power allocation: 20% of the power was available as a pool to be shared amongst beams when needed (e.g., in a rain fade event).

While the trend of increasing mass characterized the first 40 years after Intelsat I, we observe that this trend flattened out. Indeed, the industry has reversed that trend over the last fifteen years and is moving to smaller and lighter satellites to reduce costs [26]. On the other side, the second trend of multiple beams and more flexible resource allocation is continuing after the initiation by Thaicom 4 [27]. One example of the reflection of those two trends is the O3b mPower constellation [28]. It consists of 7 MEO (Middle Earth Orbit) satellites providing a combined throughput of over 1 Tbps [29] – almost two orders of magnitude more than Thaicom 4. However, the combined mass and power has not changed significantly. Most of the increased throughput is due to two orders of magnitude increase in the number of beams. The digital communication payload and multi-beam phased array allow the operator to allocate resources dynamically [30].

These developments lead to performance increases (in terms of data throughput) while at the same time reducing the per-bit cost of broadband communication [31]. This cost reduction could increase the competitiveness of satellite broadband access relative to terrestrial broadband solutions.

1.3 Pull from demand – high growth for bidirectional and bursty demand

In 2017, Morgan Stanley released an investment report predicting the future of the space economy until 2040 [32]. They expect the global demand for data to grow at an exponential rate due to an increase in the global population, autonomous cars, Internet of things, artificial intelligence, virtual reality, and video applications. Several other sources [33-37] expect similarly high rates of growth with an often-cited doubling of data volume every two years. Morgan Stanley [32] and a 2019 Northern Sky Research (NSR) report [38] believe that the satellite broadband market can leverage that growth. Morgan Stanley forecasts an increase in broadband revenue opportunity of two orders of magnitude by 2040. The NSR sees a tripling of the number of satcom consumer broadband subscribers until 2027 for a low growth scenario and a tenfold increase for the high growth scenario. An SES report [39] cites a particularly high growth for the mobility sector, in particular for airplanes and maritime, which aligns with the new market opportunities considered by Morgan Stanley: airplanes, maritime, trains, and trucks, and automobiles. Kota [40] and Farserotu [41] specifically discuss the mobility sector and the integration of satellite networks. As an additional future market, Hosseini [42] reviewed how satellites can be integrated into a 5G landscape to play an essential role in providing control and data links to Unmanned Aerial Vehicles (UAVs).

A critical characterization of the growing demand is its ratio of received (downloaded) to sent (uploaded) data. If the ratio is close to 1, i.e., both data rates are close, we can talk about *bidirectional*. Whereas if the ratio is far off from 1, we call the connection *unidirectional*, in which either the received or uploaded data is dominant.

If we go back to before the creation of the Internet, content creation and distribution was an expensive and challenging process [43]. Therefore, only larger institutions had the resources to create content and distribute widely. The data flow was typically unidirectional from institutions to individuals. The users were mostly *consumers*. Take, for example, television (TV), large corporate- and government-controlled media generate content. Capital-intensive communication satellites then broadcast the content (see Figure 1-1 unidirectional broadcasting). Users can choose between different TV channels, but they cannot directly influence the streamed content or create their own content.

With the arrival of the Internet, content distribution was made much easier, faster, and cheaper [43]. Thus, allowing users to be not only consumers but also *creators* [44]. The connection becomes increasingly bidirectional. An example here is the interaction of a passenger on a flight with a social media platform. The passenger might create content by sharing photos while consuming content by scrolling through the

newsfeed. Hence, communication satellites must be able to handle dynamic bidirectional connections (see Figure 1-1). Another example is YouTube, where users create and consume at the same time. Bloggers (creators) are upload-heavier and consumers download-heavier.

Video streaming services, such as Netflix, are mostly unidirectional and download heavy. There is a small upload part where the user requests the streaming of a specific movie. Compared to traditional TV, consumers can influence the streamed content.

To summarize, with the expected high growth of demand for data, the need for satellite communication is expected to grow as well [32, 38, 39]. That is especially true for the mobility sector, where terrestrial alternatives perform more poorly [32, 39, 42]. Since the creation of the Internet and the drop in cost for creating and sharing content, users are no longer only consumers but also become creators. Thus, most of the demand becomes bidirectional and more variable. These demand trends are narrowly interweaved with the technological advancement describe in Section 1.2, resulting in favorable conditions for a growing role of communication satellites for bidirectional broadband connectivity. However, this opportunity did not pass by unnoticed. New players plan to enter the market and are likely to disrupt the current landscape.

1.4 Changing landscape of satellite operators

As a response to the disruptive changes underway, new entrants enter the market, and current market leaders react. The current landscape is likely to change over the next years until the impact of LEO constellations become clear. For example, since the initiation of this dissertation, one major player entered (Amazon), while two others filed for bankruptcy (LeoSat and OneWeb).























Overview of the key players and new market entrants

Traditionally, the main business of satellite operators was the unidirectional broadcasting of television content. Especially in earlier days with government sponsored media, the government often launched their satellites to distribute content. Therefore, the market for satellite communication was split up by region. During the last decades, satellite operators were privatized and merged, and hence coverages became more global. Many of those companies already started to add broadband services (FSS and MSS, see Figure 1-1) to their portfolios.

We combined data from several sources [32, 45, 46] with companies' homepage data to constructs an overview of the current market landscape, with focus on broadband satellite operators with continuous coverage (that excludes systems like ORBCOMM). In particular, in Table 1-1, we categorize the main

players by two Quality of Service (QoS) metrics: their round-trip latency (direct related to orbit) and their throughput. The logos with a green-dashed rectangle are companies that applied to the FCC but are currently not operational (we call them new entrants). LeoSat, with a red-dashed rectangle, went out of business in 2019, and OneWeb filed for bankruptcy in March 2020 [47]. We base the categorization on the company’s current operational satellite or constellation with the highest throughput and lowest latency.

Table 1-1: Overview of the key players in the broadband communication market (FSS and MSS). A green-dashed rectangle indicates companies who filed an application with the FCC by April 2019 but have no operational assets. We base the categorization on the companies’ current operational satellite with the highest throughput.

| | Round-trip latency | | |
|-------------------|---|--|---|
| Throughput | ≈50 ms (LEO) | ≈150 ms (MEO) | ≈1000 ms (GEO) |
| Mbps/Kbps |   |  | |
| 10-100 Gbps |   |      |     |
| Multiple 100 Gbps | |   |  |
| Tbps |    |   | |

Looking at the landscape of current operators (excluding the new entrants), we can identify two main groups. First, there are low-throughput and low-latency LEO constellations from Globalstar and Iridium. Second, we note medium-throughput and high-latency GEO satellites from Inmarsat, Hispasat, SES, Embratel, Telesat, Thaicom, Intelsat, JSAT, Eutelsat, Echostar, and ViaSat (ordered by increasing throughput). We found that particular the GEO satellites in the 10-100 Gbps range are hybrids between broadcasting and broadband. Jupiter-2 from Echostar and ViaSat-2 are more focused on broadband

services, resulting in multiple 100 Gbps of throughput. O3b's constellation is a unique system outside of these two main groups with providing solely broadband connectivity from an MEO (SES acquired O3b in 2016).

The market for the Globalstar and Iridium is mainly voice or low data-rate services for mobile users in the form of satellite phones. Demand felt far behind expectations, and hence those companies had many setbacks. Nevertheless, they are now able to operate successfully by addressing niches: regions without cellular coverage or for special applications, e.g., connections with stringent security requirements.

All of the major entrants propose LEO constellations, with some of them not being traditional satellite companies (such as Amazon or SpaceX). According to their filing, all of them aim to enter the broadband market, but their proposed constellations vary widely in size and throughput. On one end is SpaceNorway with two satellites and throughput in the Mbps area, and on the other side of the spectrum is SpaceX with 4,425 and > 20 Tbps of throughput. The majority of the proposals plan for Tbps of throughput combined with fiber-like latency resulting in a better Quality of Service (QoS). We expect severe competition for the existing Gbps broadband operators with the risk that new entrants take away their market share for broadband connectivity – leaving them with a shrinking broadcasting business. However, existing GEO satellite operators are unlikely to stand by and watch. We discuss an overview of their response in the following (note that is an overview; a detailed analysis of each competitor's strategy is outside the scope of this dissertation).

The response from existing satellite operators

From our research, we did not find any indication that either Globalstar or Iridium will consider significant changes to their strategy (all arguments based on publicly available information as of February 2020). Their low-throughput voice service might be less susceptible than the GEO satellites' medium-throughput services. In general, we found three types of responses:

No strategic change. Similar to Globalstar and Iridium, we did not find any substantial change in the strategy of Embratel and Thaicom. They seem to expand their broadband capacities slowly but keep their primary focus on broadcasting.

Merging and partnering. The relationship between SES and O3b is a prime example. SES supported O3b financially and technically over the years until SES acquired O3b entirely in 2016. Similarly, the existing operators Hispasat and JSAT invested in LeoSat [48, 49]. In 2018, Intelsat's shareholders [50] did not support a planned merger between Intelsat and OneWeb (but Intelsat remains as an investor). In the same

year, Inmarsat [51] rejected a merger with EchoStar. EchoStar's Hughes division was partnering with OneWeb to build its ground network [52]. In 2018, Eutelsat decided to forego jointly funding a satellite with ViaSat, leaving them as competitors for the broadband market in Europe [53]. All of these activities in a single year show that the industry undergoes disruptive changes. Existing operators try to either partner with new entrants or increase their size and competitiveness by merging with another existing operator.

Taking a risk with new technology. Over the last decades, new technology was adapted rather slowly by satellite operators. To keep up with the new entrants, we observe that operators are willing to take more risks with new technology. For example, Telesat is heavily investing in its own LEO constellation of 117 satellites. SES is launching a new generation of 7 MEO satellites called O3b mPower, with phased arrays and digital processors. Eutelsat is launching a highly flexible satellite Quantum with customizable beams that can dynamically move capacity [54]. Moreover, Eutelsat plans to launch a new high-throughput satellite (HTS) Konnect as well as an upgraded version Konnect very-HTS with up to 500 Gbps. Besides a much higher capacity, a common theme of these new satellites is a significant increase in the flexibility of resource allocation.

To conclude, we expect that the landscape of satellite operators will undergo significant changes over the next years (Chapter 2 provides an analysis of total market demand and capacity, and their dynamic interaction). The market entrants generate new dynamics between the existing operators themselves and the new entrants. We saw that new partnerships were built, and companies merged. Compared to previous years, existing satellite operators adopt new technologies faster and take more risks. In particular, flexible payloads have the opportunity of better utilization of the satellite under variable demand (this is true not only for existing operators, but we have also seen indications that new entrants are likely to adopt this technology). Proper management of the flexibility is likely to become an important differentiator in the marketplace.

1.5 General objectives and expected contributions

While new flexible payloads offer unique opportunities to gain a competitive edge, they also come with challenges, in particular operational challenges. Controlling the many degrees of freedom is no longer manually feasible and requires automation that spans multiple parts of an operator’s organization and its value chain. That can go as far as changing the way operators are making business (especially for organizations that transition from traditional satcom to flexible HTS at an accelerated pace). We illustrate the challenging environment an operator is facing in Figure 1-3 with flexible HTS¹.

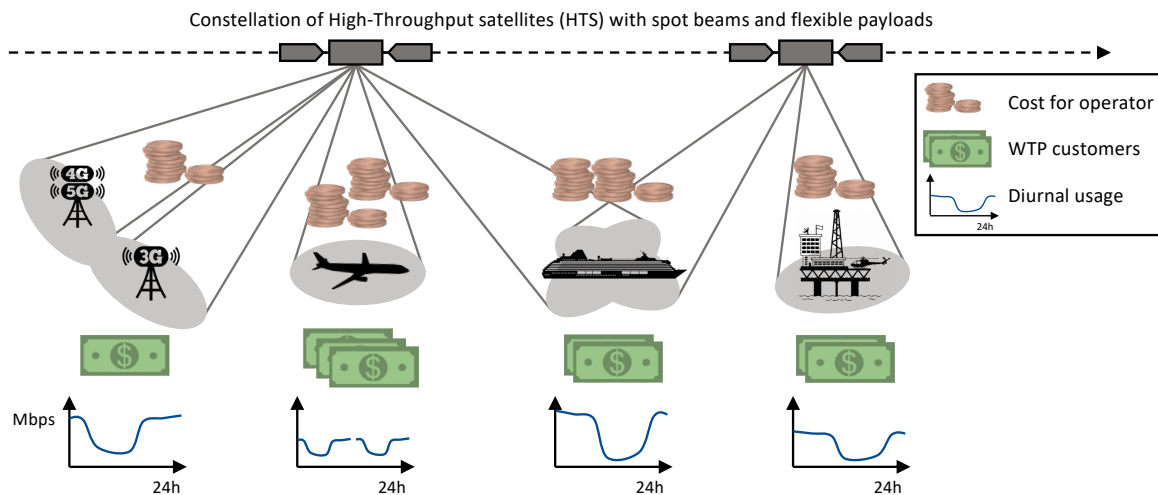


Figure 1-3: Conceptual drawing of the challenging environment in which flexible HTS are operating. The four different segments are backhauling, aviation, maritime, and energy. The brown coins represent the cost for the operator, the green bills the willingness-to-pay (WTP), and the blue line illustrates the diurnal usage of data rate.

In the example, the satellite operator has a constellation of satellites either in GEO or Non-geostationary orbit (NGSO). For this argument, we say that there are four segments: backhauling, aviation, maritime, and energy. Customers in each segment have a different willingness-to-pay (WTP) on one side. On the other side, operators have a different cost of serving each segment (and even each customer). Backhauling user terminals might have a 2.4m dish with a G/T of 27dB while an aviation terminal has a 0.6m phased array with only 15dB G/T. Hence, to achieve the same data rate, the operator has to compensate for the 12dB difference by increasing power (or bandwidth). Moreover, each user has different diurnal behavior. It might have a more considerable variation between day and night, more burstiness, and is more or less predictable. To further complicate, links are not independent of each other, and therefore, increasing power and bandwidth for one user constraints the ability to do the same for the other users.

¹ While this drawing focuses on new flexible HTS, some of the ideas apply to traditional satcom as well. However, since the objective of this dissertation is flexible HTS, we will not detail this further.

The challenge becomes for the operator to navigate this environment and find the right, coupled trade-offs between different costs of serving, WTP of customers, and their different usage behavior. Since the new digital payloads and phased arrays come with a cost premium, it is critical for satcom operator to understand how they can leverage this technology economically, which is a combination of the inherent advantages of the technology and its smart management. The research gap is around the question:

How can satcom operators monetize the flexibility of digital payloads and phased arrays?

Addressing this research question is the general objective of this dissertation. At first sight, the question might seem straightforward. However, when diving deeper into the details, we find that it opens up a series of other questions:

1. How do the dynamics, the uncertainties, and the trends in the satcom market affect the strategy of satcom operators to monetize freed up capacity? (Chapter 2)
2. Given the tight coupling between economics and technical aspects of the questions, can we leverage Revenue Management techniques from other industries? (Chapter 3)
3. What challenges need to be overcome when adopting a framework from another industry to satcom? (Chapter 4 – 7)
4. How, and how much value can satcom operator extract by implementing such a framework? (Chapter 8)

The last question ties back to the main research question and provides an answer on how operators can leverage and navigate the environment outlined in Figure 1-3. To arrive at this answer, we must address the other three questions first. We have to understand how the market might evolve in the coming years, given the uncertain success of the new entrants and if greater flexibility can influence the dynamics of the market. The outcome will frame the remainder of the dissertation and motivate the competitive advantages of a more automated demand and resource management system. We hypothesize that the benefits from the implementation of *Revenue Management* in other industries are transferable to the satcom industry. While building these analogies, we discovered four unique satcom challenges to overcome. We propose solutions to these challenges and implemented them to test the proposed satcom Revenue Management framework with real data from a satcom operator.

We dedicate a separate Chapter to the first two and the last question, and one Chapter for each of the four discovered challenges from the third question. When applicable, we review the literature, define the specific objectives, and summarize the contributions within the Chapters. An overview of the state-of-the-

art, the gap, and the contributions of each Chapter is provided by Table 1-2. Note that Chapter 4 detailing the satcom simulator is not listed since it does not contain a substantial scientific novelty (even though it is crucial for the overall framework).

Table 1-2: Summary of the state-of-the-arts, the gaps, and the expected contributions made in this dissertation. Chapter 4 (the satcom simulator) is not listed in the table as the Chapter does not contain a strong scientific novelty.

| Chapter | State-of-the-art | Gap | Contribution |
|--|---|--|--|
| 2. Satcom Market Dynamics | <ul style="list-style-type: none"> No academic market studies for satcom exists. | <ul style="list-style-type: none"> No public study and market dynamics model to understand the dynamics of the satcom market. | <ul style="list-style-type: none"> Developed a satcom market dynamic model on which future research can be built. Analyzed how the flow of capacity, and vertical integration affects the bottom lines in the future given the significant uncertainties in the market. |
| 3. Satcom Revenue Management Framework | <ul style="list-style-type: none"> Some literature on generic frameworks Specific literature for various industries, e.g. airlines, hotels, car rentals, air cargo, Internet services, telecom, ... | <ul style="list-style-type: none"> No work on Revenue Management for satcom | <ul style="list-style-type: none"> Identified Revenue Management as highly applicable to satcom. Contrasted six industries with satcom using our own taxonomy. Identified that resource management is a key dimension for satcom RM not considered by current RM research. Proposed a satcom RM framework that captures the complexity of satcom. Identified four challenges of satcom RM: unit of demand not unit of capacity, resource allocation, available capacity forecaster, and novel SLAs. |
| 5. Resource Allocation | <ul style="list-style-type: none"> Increasing amount of literature on individual aspects of resource allocation | <ul style="list-style-type: none"> No work done on formalizing and decomposing the resource allocation process | <ul style="list-style-type: none"> Formalized and decomposed the resource allocation process into four sub-problems. Developed a closest-first and balanced gateway allocation algorithm for the routing sub-problem. |
| 6. Available Capacity Forecaster | <ul style="list-style-type: none"> General time-series forecasting State-of-practice is deterministic | <ul style="list-style-type: none"> No probabilistic forecasting that captures the uncertainty in demand usage. | <ul style="list-style-type: none"> Developed an approach of how to probabilistic forecast the available capacity based on historical time-series user traffic. |
| 7. Novel SLA | <ul style="list-style-type: none"> Static, long-term satcom SLAs | <ul style="list-style-type: none"> No study of novel SLAs and how they can benefit operators and customers. | <ul style="list-style-type: none"> Reviewed current satcom SLAs and built analogies to other industries. Proposed a novel set of SLAs that are beneficial to operators and customers. |
| 8. Application of the Satcom RM Framework | | | <ul style="list-style-type: none"> Proved the proposed RM framework's value by implementing Chapters 4 – 7 and testing it with real data from a satellite operator. Showed revenue gains of 0-39% are possible and that more sophisticated RM outperforms pricing heuristic consistently by 4-7%. |

1.6 Outline of the dissertation

To guide the reader through the main body of this dissertation, we use the proposed Revenue management framework developed and justified in Section 3.5 (see Figure 1-4). Chapters 3 - 7 each cover one component of the framework. We discuss in Chapter 3 in detail why Revenue Management (RM) is a promising framework and why we decompose the satcom framework the way we did. Nevertheless, we briefly describe the working principles here beforehand to provide the reader with a crude sense of how the framework is functioning, how it supports answering the four research questions, and how it structures the document.

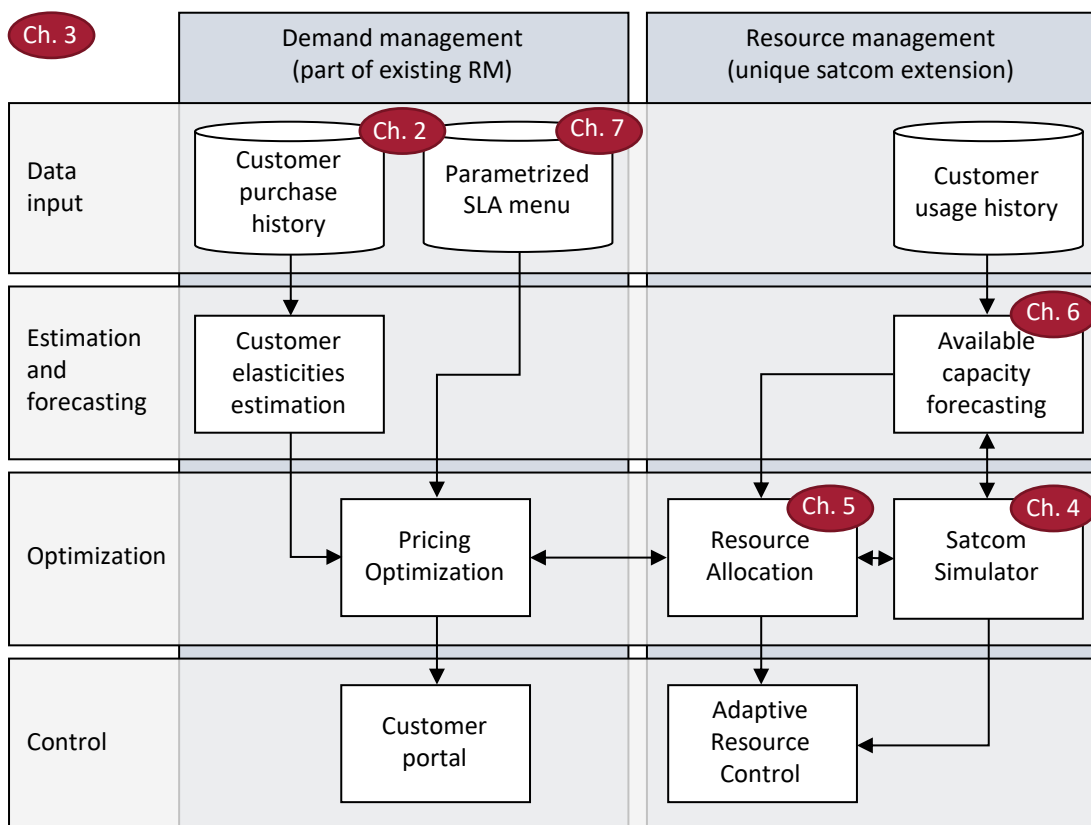


Figure 1-4: Preview of the proposed framework from Section 3.5 as a guide through the remainder of this dissertation

The framework consists of four layers: data input, estimation and forecasting, optimization, and control. Both the demand and resource management parts of the framework represent all four layers.

In the demand management part, the customer purchase history allows for an estimation of elasticities (conceptually the green bills in Figure 1-3). The pricing optimization uses these elasticities together with

the SLA menu (Chapter 2 and 7) to find the set of prices for each customer that maximizes the operator's revenues. Then, the control layer communicates these prices through a portal to the customers.

On the resource management part, the data input consists of a database storing the historical usage of customers (conceptually the diurnal usage in Figure 1-3). An available capacity forecaster uses that data to estimate the capacity used and still available (Chapter 6). The forecaster interacts with the resource allocation (Chapter 5) and the satcom simulator (Chapter 4). The objective of these two components is the minimization of resource usage for a given data rate usage of the user (resource being the cost in Figure 1-3). It turns out that this is a particular challenge, and hence Chapter 5 is one of the critical parts of our work. In a final step, the found solution updates the constraints of the adaptive resource control, which reacts to short-term variations.

The interface between the pricing optimization and resource allocation tightly couples the demand and resource management. The pricing optimization computes new prices and, therefore, quantities of data rate in each iteration. The resource allocation then coordinates with the satcom simulator and the available capacity forecasting to find the minimum amount of resources required to support the new set of data rates. This feedback flows back to pricing optimization.

The linkage between the four research questions, the framework, and the Chapters of this document is as follows:

- Chapter 2 *provides an input* to the framework. We primarily answer the broader, first research question: what are the dynamics, uncertainties, and trends in the satcom market. Furthermore, the Chapter sets the market context for the customer purchase history and the SLA menu.
- Chapter 3 *develops the overall framework*. We examine the applicability of RM to satcom, review RM in six industries, and develop the framework, as shown in Figure 1-4. Furthermore, in this Chapter, we identify the four challenges of satcom RM. The Chapter answers the second research question.
- Chapters 4 – 7 *present solutions to the four challenges* of adopting RM from other industries to satcom. The solutions are represented by separate components in the framework and answer the third research question. We dedicate separate Chapters to the four challenges:
 - Unit of capacity is not unit of demand requires a satcom simulator deliberated in Chapter 4

- Resource allocation is an optimization problem itself commands algorithmic solutions presented in Chapter 5
- Uncertain available capacity based on resource usage is addressed by the available capacity forecaster covered by Chapter 6
- Existing SLAs might not fully leverage the new satellite flexibility require novel and parametrized SLAs introduced in Chapter 7
- Chapter 8 *uses the framework*. We implement each component of the framework and integrate them into an executable RM simulator. Using real data from a satellite operator, we analyze several use cases and evaluate the value of the proposed framework.

We start with the following Chapter 2, in which we build a market dynamic model to analyze questions that have important implications for the remainder of the dissertation.

2

Satcom Market Dynamics

We divide the current satcom landscape into three layers: operators, service providers/value-added resellers, and customer/end-users. Figure 2-1 shows an overview of the key players. We did not find any public information about SpaceX, OneWeb, and Amazon strategies and whether they are partnering up with existing service providers or going directly to the customer (note that OneWeb filed for Chapter 11 bankruptcy in March 2020 [47]). The black arrows indicate business relationships between companies. The red arrows are relationships between operators and customers without an intermediary service provider. Most of the data is obtained by researching the companies' webpages and their annual 10-K reports. We base some links on news articles from SpaceNews and Via-Satellite, and we found other relationships through interviews with various people within the industry. The markets (verticals) on the customer level are from left to right: voice and low data for remote locations, land mobility, maritime, cruise ships, oil & energy, aviation, backhauling, and residential broadband (more on customer segments in Section 7.6).

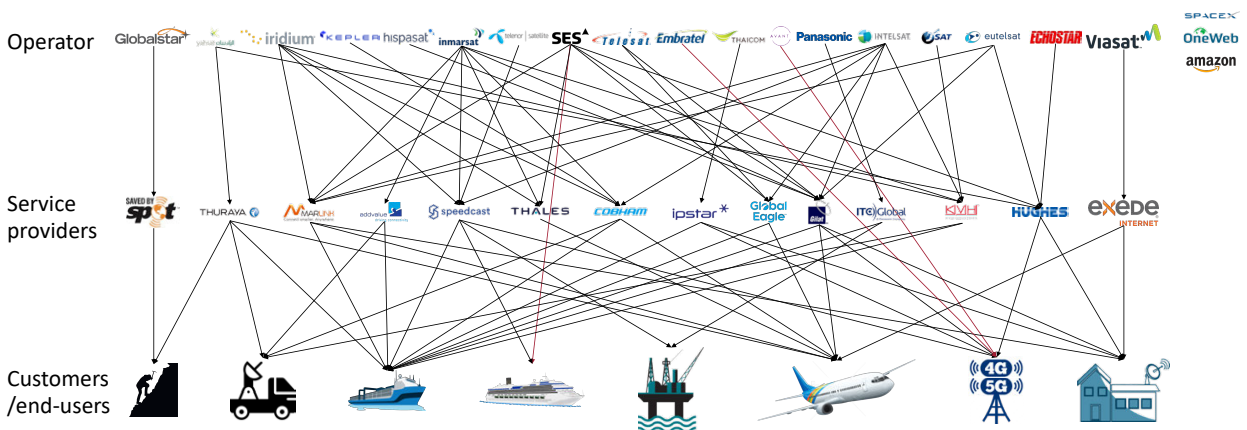


Figure 2-1: Overview of the different players in the market, divided into operators, service providers, and customer/end-users. The connections illustrate publicly available information about business relationships.

Interviews with several professionals in the industry revealed the following capacity dynamics: companies in each horizontal layer buy capacity in bulk with different durations, as shown in Table 2-1. The operator launches new satellites or constellations with a lifetime of 10-15 years (for GEO), whereas for MEO and LEO, this cadence is more in the 3-10 years range. Service providers then buy from this bulk capacity with 1-3 years long SLAs. They sell this bulk capacity to customers with a contract duration of usually years. In individual cases, operators sell *occasional use* capacity with duration as short as multiple hours.

Table 2-1: Overview of the current purchasing behavior

| | Capacity purchase behavior | Capacity contract durations | Dynamics over time |
|--------------------|--|------------------------------------|--------------------|
| Satellite operator | Bulk through launch of satellites | 5-15 years | |
| Service provider | Bulk from operator | 1-3 years | |
| Customer | Classical SLAs (see Section 7.1 for definition) from service providers | Usually years, can go down to days | |

This bulk purchase behavior results in a mismatch between demand and capacity for each level. Since the purchases have different time constants, the dynamics differ (not the whole horizontal as discussed later in the market dynamic model). The right column of Table 2-1 shows that schematically:

- Satellite operators launch their satellites with the highest available capacity in the beginning. Due to solar array degradation and therefore reduced power availability, the capacity declines over time until operators decommission satellites and replace them with a new satellite. During an initial ramp-up phase of 2-3 years, the available capacity is sold.
- Service providers buy in bulk, expecting lower utilization initially, as the demand has not materialized yet. Over time, demand is growing until the service providers need to rebuy more capacity, which results in a drop in the utilization.

- Customer dynamics are different. They are mostly due to diurnal patterns. Customers do not make use of all their purchased capacity (here with a Committed Information Rate (CIR)) with particular low usage during night hours. We will discuss this usage behavior in greater detail in a later chapter of this dissertation and support it with real data.

Building on these observations, we develop a parametric market dynamic model in the remainder of this Section to understand in greater depth the dynamics in the satcom market.

There are some commercial market models that, for example, NSR and Euroconsult, have developed to forecast data in their reports. However, these are not openly accessible. Hence, we build the model that we describe in full detail to be reusable and extendable by the community. Our purpose in building the market dynamic model is to address the following questions, amongst other aspects systematically:

- How do new entrants change the market?
- How do the uncertainties, e.g., the success of the new entrants, affect the market?
- How does the market react if contract durations are shortened?
- How does the market react if vertical integration continues?

Our approach to addressing these questions is to use the proposed market dynamic framework from Weil [55, 56] and adapted it to the global broadband satcom market. We use the System Dynamic language that lets us represent and solve coupled, nonlinear, integral equations (see a short introduction to System Dynamic in Appendix A). In that, we leverage the modeling capabilities of Vensim and the flexibility of Python, resulting in the workflow depicted in Figure 2-2. We use the PySD package to translate the model.

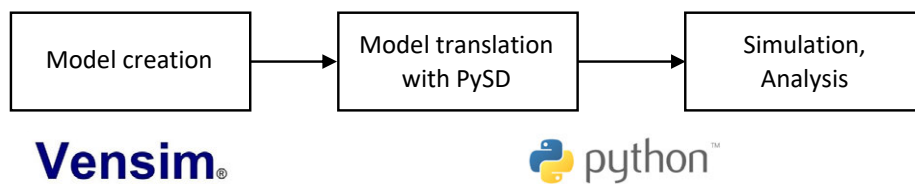


Figure 2-2: Workflow of market dynamic model creation, translation, and analysis

In the following Section 2.1, we describe the model, the assumptions, and the input data. Section 2.2 describes the validation of the model. The global sensitivity analysis of the model is described in Section 2.3 and the effects of changing contract durations in Section 2.4. Section 2.5 addresses the question of vertical integrations. We state the conclusions and the implications for the remainder of the dissertation in the final Section 2.6.

2.1 Description of the model and assumptions

Figure 2-3 depicts an overview system dynamic diagram. Per our previous description of the satcom market, we decompose the satcom market into three horizontals: operators, service providers, and customers and make them explicit in our market dynamic model.

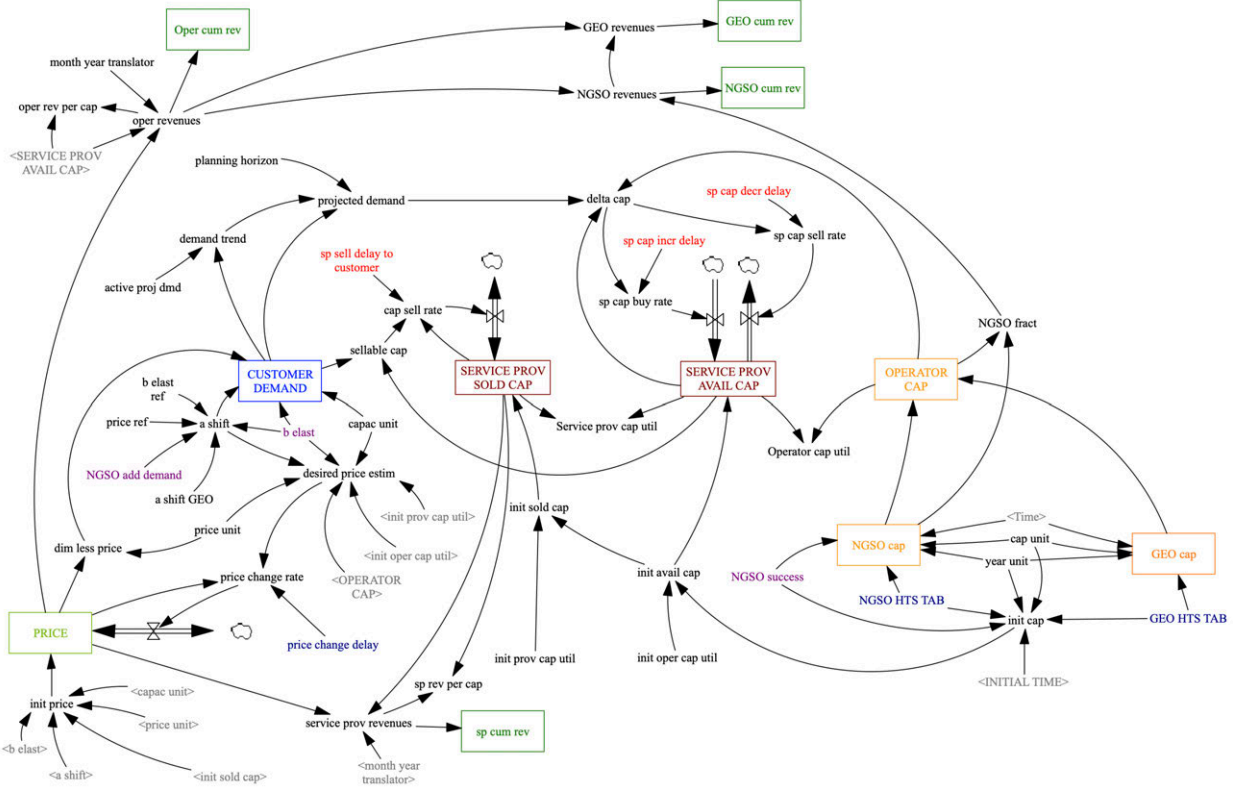


Figure 2-3: system dynamic diagram of the market model

The main feedback loop in the system is spanning from the customer demand level to the available capacity of the service provider. Service providers project demand based on historical trends and calculate the desired capacity based on targeted utilization. Finally, they obtain the delta capacity that it buys from the operator to increase the capacity (or sell back to the operator to decrease overcapacity). Since the buying- has a different dynamic than the selling-process, we model both separately and hence can influence them independently. Both effect the available capacity of the service provider. Separate from that is the sold capacity of the service provider, which is changed by the capacity sell rate. The sellable capacity is the customer demand and capped by the available capacity. The ratio of sold over available capacity defines the service providers' capacity utilization.

The price drives customer demand through a demand elasticity function. The price itself is changed by the price change rate that is informed by the desired price estimation. We base the estimation on a target utilization and the available capacity of the operator.

The assumption about capacity development in the coming years drives the operator capacity. We split up the capacity between NGSO and GEO HTS capacity, where NGSO is representative for LEO constellations despite that some capacity is also from legacy operators, such as SES’s O3b MEO fleet. For the model, we assume that GEO and NGSO capacities are equal. We hence sum them up to one operator capacity that is available to be bought by the service providers. Furthermore, the control variable V multiplies the NGSO capacity. By varying it from 1, we simulate situations where the large LEO constellations fall behind (lower than 1) or are more successful than expected (larger than 1).

As data input, we refer to source [57], which summarized the supply, demand, and price trend based on data points from Euroconsult and NSR in 2017 (see Figure 2-4). Over the last years, there has been a sharp decline in prices, which lead so far to a modest increase in demand due to relatively inelastic demand. However, analysts from Euroconsult and NSR [58] expect that prices below \$200/Mbps/month trigger sensitive points in the demand elasticities and therefore increase demand considerably.

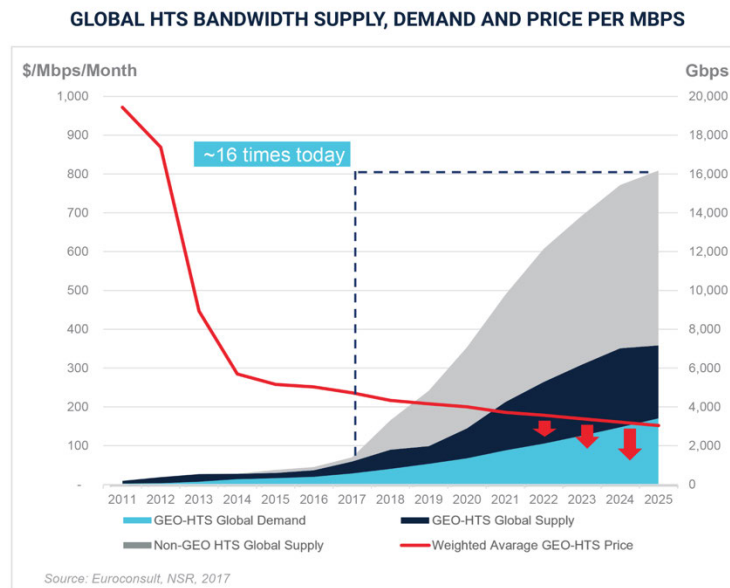


Figure 2-4: Supply, demand, and price trend analysis obtained from [57], which is based Euroconsult and NSR.

The capacity development of NGSO and GEO HTS is extracted by the digitalization of Figure 2-4 so that we get Table 2-2, where C_{oper} is the total global satcom HTS capacity. As the data indicates, the NGSO

capacity outgrew the GEO capacity in the year 2019. After that, the data forecasts an increase for GEO capacity is with an average rate of around 0.75 Tbps/year and NGSO with 1 Tbps/year.

Table 2-2: HTS capacity, supply, and price forecast for years 2017-2025 extracted from [57], which consolidated 2017 data from NSR and Euroconsult. Capacity and demand are in Gbps, price in \$/Mbps/Month.

| | $C_{GEO}(t)$ [Gbps] | $C_{NGSO}(t)$ [Gbps] | $C_{oper}(t)$ [Gbps] | $D_{GEO}(t)$ [Gbps] | $p_{GEO}(t)$ [\$/Mbps/Month] |
|-------------------|------------------------|-------------------------|-------------------------|------------------------|---------------------------------|
| 2011 ⁺ | 240 | - | 240 | 69 | 969 |
| 2012 ⁺ | 430 | - | 430 | 136 | 870 |
| 2013 ⁺ | 580 | - | 580 | 202 | 451 |
| 2014 | 580 | - | 580 | 321 | 285 |
| 2015 | 580 | 60 | 640 | 362 | 257 |
| 2016 | 749 | 190 | 939 | 403 | 252 |
| 2017* | 1,240 | 198 | 1,438 | 625 | 240 |
| 2018 | 1,799 | 1,574 | 3,374 | 848 | 219 |
| 2019 | 1,970 | 2,872 | 4,842 | 1,105 | 210 |
| 2020 | 2,867 | 4,308 | 7,175 | 1,362 | 203 |
| 2021 | 4,249 | 5,502 | 9,750 | 1,700 | 189 |
| 2022 | 5,388 | 6,730 | 12,118 | 2,000 | 180 |
| 2023 | 6,000 | 7,600 | 13,600 | 2,342 | 170 |
| 2024 | 6,975 | 8,408 | 15,383 | 2,892 | 160 |
| 2025 | 7,145 | 8,962 | 16,107 | 3,443 | 152 |

* year of data; ⁺ not considered for the demand elasticity regression

The following provides a summary of the input and output parameters into the model. After that, we describe the mathematics of the model in more detail.

Input

We divide the input parameters of the model into three categories: control parameters with which we can change the dynamic of the system, input parameters with significant uncertainty, and other input parameters.

Control parameters (red font in Figure 2-3)

- Service provider sell delay to customer $\Delta t_{SP,sell}$
- Service provider capacity increase delay $\Delta t_{SP,incr}$
- Service provider capacity decrease delay $\Delta t_{SP,decr}$

Parameters with significant uncertainty (purple font in Figure 2-3)

- Elasticity of customer demand to price $D(t) = f(p(t))$
- Success of the NGSO constellations s_{NGSO}

Miscellaneous parameters (blue font in Figure 2-3)

- NGSO capacity development throughout the simulation time $C_{NGSO}(t)$
- GEO capacity development throughout the simulation time $C_{GEO}(t)$
- Price change delay Δt_{price}

Output

The model produces various outputs. We are particularly interested in the following main outputs:

- Revenues over time and cumulative revenues for service provider and NGSO and GEO operator $\Pi_{SP}(t), \Pi_{NGSO}(t), \Pi_{GEO}(t), \Pi_{SP,cum}, \Pi_{NGSO,cum}, \Pi_{GEO,cum}$
- Capacity utilization over time for service providers and combined operators $\eta_{SP}(t), \eta_{oper}(t)$
- Price and customer demand development over time $p(t), D(t)$

Mathematical model description

The total operator capacity is obtained by Eq. (2-1) using the success of NGSO constellations s_{NGSO} , which draws data from the lookup Table 2-2. Eq. (2-2) gives the fraction of NGSO capacity $x_{NGSO}(t)$.

$$C_{oper}(t) = C_{GEO}(t) + s_{NGSO} \cdot C_{NGSO}(t) \quad (2-1)$$

$$x_{NGSO}(t) = \frac{s_{NGSO} \cdot C_{NGSO}(t)}{C_{oper}(t)} \quad (2-2)$$

We model the demand elasticity of the customers regarding price by fitting a function to the data points obtained from Figure 2-4 (and printed in Table 2-2). We consider the prices from 2014 onwards as the steep drop in the years before misaligns the regression for the lower price ranges, which are more relevant in the later years of the simulation. The underlying assumption is that the current demand elasticity b for GEO HTS is representative of total demand (GEO + NGSO HTS). Furthermore, we assume that the NGSO constellations create additional demand modeled by $m_{NGSO,demand}$ that equals 1 for the baseline (same demand generated by NGSO than there is for GEO). We find a log-linear model to achieve the best results ($R^2 = 0.98$) with the parameters $a_{GEO} = 3.33 \cdot 10^{13}$ and $b = -4.11^2$. The resulting two equations are:

$$D(t) = (1 + m_{NGSO,demand}) \cdot a_{GEO} \cdot p(t)^b \quad (2-3)$$

² This elasticity value is considerably more elastic than the upper range of -2 we find for the telecommunication industry in Section 8.1.2. As we discuss in the later Section, the reason is likely the influence of backhauling on the total market.

$$p(t) = \left(\frac{D(t)}{(1 + m_{NGSO,demand}) \cdot a_{GEO}} \right)^{\frac{1}{b}} \quad (2-4)$$

with for the baseline $a_{GEO} = 3.33 \cdot 10^{13}$

$$b = -4.11$$

$$m_{NGSO,demand} = 1$$

Note that for simulating the uncertainty in the elasticity b , we adjust a such that the price and demand are the same for the point of $p = \$200$. Mathematically, we write $\tilde{a} = a \cdot p^{b-\tilde{b}}$ where \tilde{a} is the new intersection for the sampled elasticity \tilde{b} . Using Eqs. (2-3) - (2-4) the desired price for the customers is estimated. The assumption is that the operator capacity C_{oper} , and the initial utilization of service provider $\eta_{SP,init}$ and operator $\eta_{oper,init}$, drive the price. The rationale for that is that the unsold capacity of operators affects the pricing. The resulting equation reads:

$$\hat{p}(t) = \left(\frac{C_{oper}(t) \cdot \eta_{oper,init} \cdot \eta_{SP,init}}{(1 + m_{NGSO,demand}) \cdot a_{GEO}} \right)^{\frac{1}{b}} \quad (2-5)$$

The actual price $p(t)$ is then changed by integrating the price change rate. The initial price p_{init} is set based on the initial capacity of simulation start inserted into Eq. (2-5).

$$p(t) = p_{init} + \int \frac{\hat{p}(t) - p(t)}{\Delta t_{price}} dt \quad (2-6)$$

In turn, we compute the customer demand by Eq. (2-3). The service provider projects the demand $\hat{D}(t)$ into the future to make capacity management decisions (modeled as aggregation). That is modeled by calculating the average trend over the past Δt_{avg} average years and using this to project $\Delta t_{planning}$ into the future. The official documentation of Vensim [59] offers further details on the *trend* function implemented. The formal equation is:

$$\hat{D}(t) = D(t) \cdot \left(1 + \frac{\Delta t_{planning}}{trend(D(t - \Delta t_{avg}; t))} \right) \quad (2-7)$$

We use $\Delta t_{avg} = 2$ and $\Delta t_{planning} = 3$ throughout all simulations. The desired capacity then depends on the projected demand $\hat{D}(t)$ and the already available capacity $C_{SP,avail}(t)$. Therefore, we get for the delta capacity $\Delta \hat{C}(t)$ with the purchasable capacity limited by $C_{oper}(t)$:

$$\Delta\hat{C}(t) = \min\left(\hat{D}(t), C_{oper}(t)\right) - C_{SP,avail}(t) \quad (2-8)$$

If $\Delta\hat{C}(t) > 0$, then the service provider aims to purchase capacity from the operator, if $\Delta\hat{C}(t) < 0$ a capacity decrease is desired. Both cases have different delays denoted with $\Delta t_{SP,incr}$ and $\Delta t_{SP,decr}$. Hence, we write

$$C_{SP,avail}(t) = C_{SP,avail,init} + \int \max\left(\frac{\Delta\hat{C}(t)}{\Delta t_{SP,incr}}, 0\right) + \min\left(\frac{\Delta\hat{C}(t)}{\Delta t_{SP,decr}}, 0\right) dt \quad (2-9)$$

and the operator utilization then becomes:

$$\eta_{oper}(t) = \frac{C_{SP,avail}(t)}{C_{oper}} \quad (2-10)$$

There is one dynamic loop left: the selling process of capacity from service provider to customer. The sellable capacity is the customer demand $D(t)$ capped by the available capacity $C_{SP,avail}(t)$. The sell rate is influence by the delay $\Delta t_{SP,sell}$. Finally, the sold capacity $C_{SP,sold}$ and the utilization $\eta_{SP}(t)$ reads:

$$C_{SP,sold}(t) = C_{SP,sold,init} + \int \frac{\min\left(D(t), C_{SP,avail}(t)\right)}{\Delta t_{SP,sell}} dt \quad (2-11)$$

$$\eta_{SP}(t) = \frac{C_{SP,sold}(t)}{C_{SP,avail}(t)} \quad (2-12)$$

Eqs. (2-1) - (2-12) are the full set required to describe the dynamics of the systems as implemented in Vensim. The monetary outputs are described by the following equations.

$$\Pi_{NGSO}(t) = x_{NGSO}(t) \cdot C_{SP,avail}(t) \cdot p(t) \quad (2-13)$$

$$\Pi_{GEO}(t) = \left((1 - x_{NGSO}(t))\right) \cdot C_{SP,avail}(t) \cdot p(t) \quad (2-14)$$

$$\Pi_{SP}(t) = C_{SP,sold}(t) \cdot p(t) \quad (2-15)$$

$$\Pi_{NGSO,cum}(t) = \int \Pi_{NGSO}(t) dt, \quad \Pi_{GEO,cum}(t) = \int \Pi_{GEO}(t) dt, \quad \Pi_{SP,cum}(t) = \int \Pi_{SP}(t) dt \quad (2-16)$$

Summary of assumptions

During the description of the model and its dynamics, we stated several assumptions. These, together with the additional assumptions that we made are:

- The market dynamics model is on global broadband HTS satcom capacity, demand, and price
- The demand elasticity of GEO HTS is representative of the global demand and NGSO generate additional demand that has the same volume as GEO HTS

- The price of capacity for GEO and NGSO, as well as customer, service provider, and operator are the same (hence assuming there are no margins along the value chain, which does not represent reality but still allows us to capture the dynamics of the market).
- The forecasts of capacity, demand, and price from [57] and Table 2-2 are representative
- The price is driven by available operator capacity
- Capacity can be bought incrementally
- No insolvency considered
- There is negligible capacity sold directly from operator to customers, and everything goes through the service provider (note, that we will run a scenario in Section 2.5 of the dynamics of a completely vertically integrated market)
- The capacity expansion of GEO and NGSO operator and price do not affect each other until 2025 (rationale is that this capacity is already ordered, delivery time is five years, and cancellation is not an economical option)
- The simulation duration is from 2014 to 2025, where we use the interval from 2014-2019 as validation of the model (see next Section 2.2) and 2020-2025 for predicting.

2.2 Validation

We validate the model by comparing the results obtain from 2014 to 2019 with the historical data from Table 2-2, a data point from Prasad, NSR [60], and a price trend regression of SES data. As metrics for validation, we use the GEO HTS price $p_{GEO}(t)$ and demand $D_{GEO}(t)$ since no demand-data is available for NGSO. We achieve this by running the model with $s_{NGSO} = 0$ and $m_{NGSO,demand} = 0$. Figure 2-5 displays the resulting plots for the demand and price.

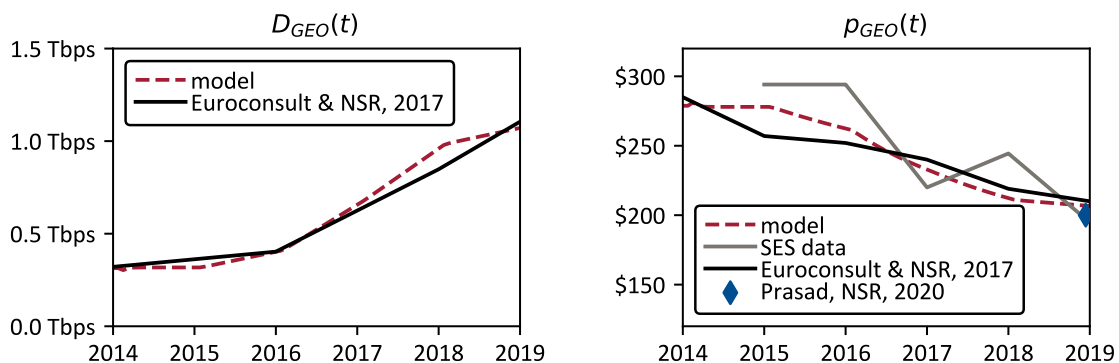


Figure 2-5: Validation of the market dynamic model against the data from Table 2-2, Prasad, NSR [60], and a price trend regression of SES data. Price $p(t)$ is in \$/Mbps/Month.

The black line is the validation data from Table 2-2, and the red line the output of the market dynamics model. The prices are in \$/Mbps/Month. We tune the initial utilizations of the operators $\eta_{oper,init}$ and service providers $\eta_{SP,init}$ to match the validation data sets.

For both metrics, the model and the validation data align well. The price trend obtained from SES data confirms the price decrease. Prasad [60] mentions in an online article for NSR in January 2020 that the price reached \$200/Mbps/Month, which what the 2017 forecast suggested, the model predicted, and the SES data confirms. For the Euroconsult & NSR, 2017 data, it is not surprising that a good fit for the demand also results in a good fit for the price since we obtained the demand elasticity of the model by regression of the validation data set. However, the defined operator capacity increase yields to the shown dynamics in demand is a validation of the model. The model represents the central dynamic: the almost linear downward trend in prices and the resulting over-proportional increase in demand. The underlying buying and selling delay dynamics cause the small differences in both lines.

2.3 Sensitivity analysis

The purpose of this global sensitivity analysis is to understand what impact certain parameters have on the metrics of interests. In particular, we consider the following parameters to be particularly uncertain: the success of NGSO constellations s_{NGSO} (multiplier for the NGSO capacity see Eq. (2-1)), the demand elasticity b , and the added demand through NGSO constellations $m_{NGSO,demand}$. Table 2-3 outlines the baseline value as well as the ranges in which we let the parameters vary. Worst and best case refer here to the achieved total revenues in the market. For example, no additional demand ($m_{NGSO,demand} = 0$) is the worst case, and twice the GEO demand is the best case.

Table 2-3: Overview of the considered uncertainty ranges for the success of NGSO constellations s_{NGSO} , the demand elasticity b , and the added demand through NGSO constellations $m_{NGSO,demand}$.

| Parameter | Baseline value | Worst case | Best case |
|--|----------------|------------|-----------|
| Success NGSO s_{NGSO} | 1 | 0 | 2 |
| Demand elasticity b | -4.11 | -2 | -5 |
| Additional demand NGSO $m_{NGSO,demand}$ | 1 | 0 | 2 |

We conduct two analyses in this Section: (1) quantification of the relative impact of the uncertainty in each of these parameters on the results, and (2) the dispersion in the results that these uncertainties cause.

Relative impact

Since our goal is to get a sense of the relative global sensitivities, we chose a variance-based Sobol sensitivity analysis [61]. For that, we use a Saltelli sampler [62] to generate a distribution of possible values for the three parameters s_{NGSO} , b , and $m_{NGSO,demand}$. The resulting matrix of this sampler contains $N \cdot (2 \cdot D + 2)$ rows, where N is the desired number of samples and D the number of parameters. In our case, we set $N = 100$ and therefore generate a matrix with 800 rows. We rerun the market dynamics model for each of these cases and the conduct one Sobol analysis for each metric on the distribution of values in the final time step t_{final} . We record the total-order indices for the price and the cumulative revenues for both operators and the service providers (see Figure 2-6).

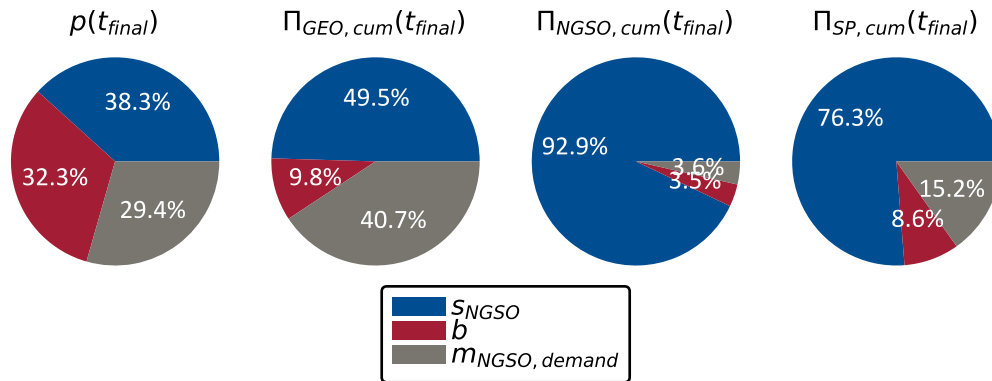


Figure 2-6: Contributions of the three uncertain parameters success of NGSO constellations s_{NGSO} , demand elasticity b , and additional NGSO demand $m_{NGSO,demand}$ on the price $p(t_{final})$, cumulative GEO revenues $\Pi_{GEO,cum}(t_{final})$, cumulative NGSO revenues $\Pi_{NGSO,cum}(t_{final})$, and cumulative service providers revenues $\Pi_{SP,cum}(t_{final})$. We use the total-order indices of the Sobol analysis.

The percentages in the pie plots quantify the relative contribution of each parameter to the total uncertainty of the parameter. Both, for the price $p(t)$ and the cumulative revenues of the GEO operator $\Pi_{GEO,cum}(t_{final})$, the uncertainty of all three parameters have a similar impact. The success of NGSO has the highest contribution with around 30-40%, while the demand elasticity and the additional demand generated by NGSO constellations are 20-30%.

The results draw a different picture of the cumulative revenues of NGSO operators and service providers. NGSO operators are almost insensitive to the demand elasticity and the additional demand they generate. The success of their capacity expansion drives them. That makes sense since, without capacity, NGSO operators cannot capture any demand to generate revenues. The actual shape of the demand (b and $m_{NGSO,demand}$) is then a second-order influence as the results confirm.

Service providers show a similar dependency surprisingly on the success of NGSO constellations (assuming no vertical integration). The uncertainties about the demand shape have only an approximately 9-15% impact. These results suggest that the uncertainty of the growth on the supply side has a more substantial influence than the uncertainties on the demand side. That illustrates a non-linearity of the dynamics in the market. The log-linear demand elasticity causes un-proportional demand growth for lower prices. When the total capacity is not larger enough, the high growth demand area is not hit and therefore resulting in smaller revenues for the service provider. However, if NGSO constellations are successful, the demand grows exponentially with success. In contrast, the parameters $m_{NGSO,demand}$ and b affect the demand growth only in a linear way.

Dispersion of the results

We use the same sampled data set and simulate the model for each of the 800 parameter combinations. We statistically analyze the recorded distributions of the simulation results with their mean, 68%, 80%, and 95% confidence intervals. Figure 2-7 shows plots over time of the price $p(t)$, the demand $D(t)$ and the revenues for GEO operators, NGSO operators, and service providers.

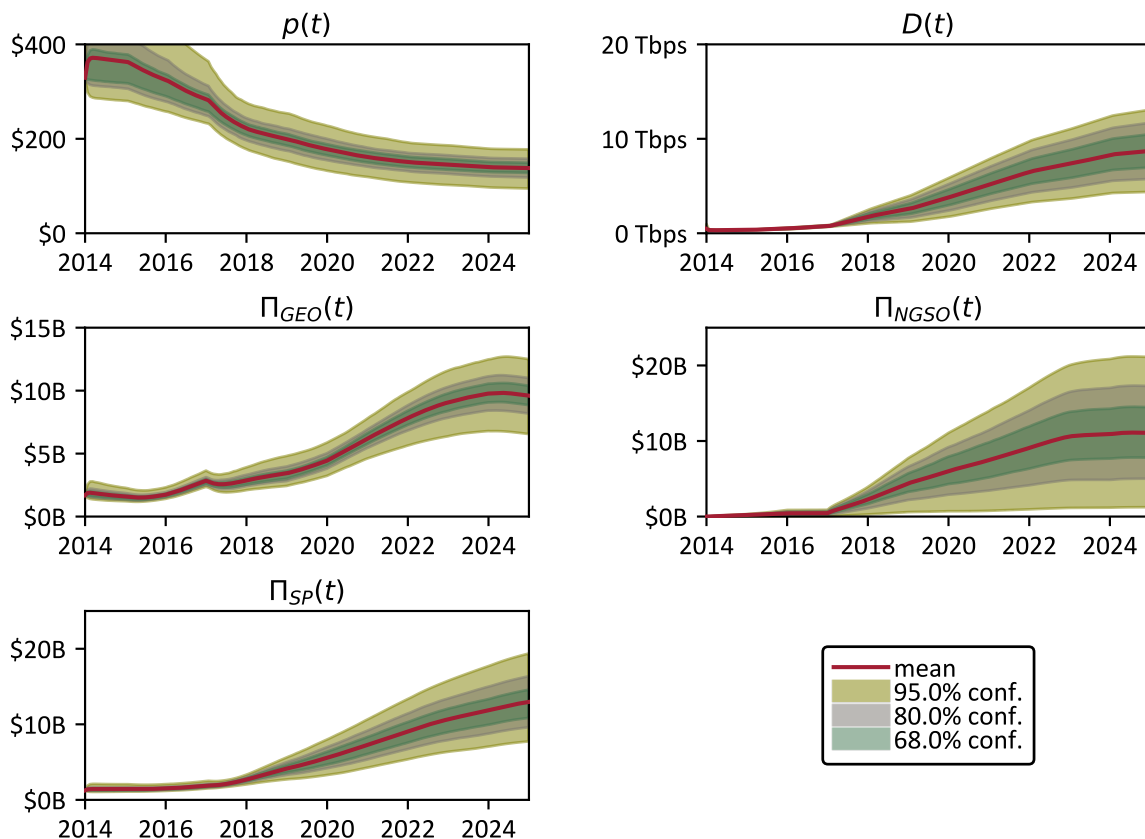


Figure 2-7: Dispersion of $p(t)$, $D(t)$, $\Pi_{GEO}(t)$, $\Pi_{NGSO}(t)$, $\Pi_{SP}(t)$ given the uncertainties in s_{NGSO} , b , and $m_{NGSO,demand}$

First, the results show that the trend of declining prices continues well below the \$200/Mbps/Month mark in the years 2020-2025. The model suggests a fairly narrow band of uncertainty around that trend. Indeed, it gets smaller, the lower prices become. In the first years until 2017, NGSO constellations have almost no capacity, and the GEO capacity drove the prices. Therefore, during that time, the uncertainty is mainly caused by the demand elasticity b^3 . That causes negligible dispersion in the other metrics $D(t)$, $\Pi_{GEO}(t)$, $\Pi_{NGSO}(t)$, and $\Pi_{SP}(t)$. The revenues for GEO operators even slightly decline between 2014 and 2015 as prices drop and the caused demand increase cannot compensate that.

In 2017, NGSO constellations start to significantly expand their capacity triggering a dip in prices and GEO operator revenues. The uncertainty in the parameters $m_{NGSO,demand}$ and s_{NGSO} become a significant cause of the dispersions (in particular for the NGSO and service provider revenues as concluded from Figure 2-6). The increase in capacity leads to a further decrease in prices. Through the non-linear relationship, the demand grows exponentially, i.e., smaller decreases in price yield to larger increases in demand. The uncertainties have the most significant impact on the revenues of the NGSO operators. As the Sobol analysis revealed, the success of the capacity expansion s_{NGSO} mainly drives that uncertainty. The revenues of the service providers are strongly correlated with this success.

Despite the further expansion of capacity after 2023, the revenues for both GEO and NGSO operators reach a plateau and even start to trend downwards for GEO. Again, the non-linearity in the demand elasticity is the origin of this behavior. In contrast, the revenue for service providers continues to grow as they increase the capacity they are selling. The delay in this capacity increase process causes the service providers to “run behind”. For example, the NGSO capacity increase of 2017 takes until around 2018 to result in a revenue boost.

After analyzing the behavior of the market given the uncertainties in s_{NGSO} , b , and $m_{NGSO,demand}$, we study in the following Section, how a change in the delays affects the dynamics and the bottom line of operators and service providers.

2.4 Effects of changing the delays

The delays $\Delta t_{SP,sell}$, $\Delta t_{SP,incr}$, and $\Delta t_{SP,decr}$ are the control parameters in the market that we consider have the potential to be actively influenced by the operators and the service providers. Our approach is

³ if we would repeat the Sobol analysis for one of these time steps, s_{NGSO} and $m_{NGSO,demand}$ would have very small percentages

to change one parameter at a time, observe the dynamics of the utilization of operator and service provider, and report the final cumulative revenues. Table 2-4 summarizes an overview of the parameters, their baseline values, and first and second reduction values. The baseline values are 1 year for the service provider sell-delay to customers and the capacity increase (i.e., the re-buying from the operator). Service providers are assumed to have a delay of 2 years to decrease their capacity. Within the market dynamics model, the term delay is equivalent to the contract duration. For the first reduction, we say that the selling and capacity increase can be speeded up to every month and eventually every week. The first step of improvement of capacity decrease is down to 1 year, and then 1 month.

Table 2-4: Overview of the baseline values of the parameters and the values for a first and second reduction.

| Parameter | Baseline value | 1 st reduction | 2 nd reduction |
|---|----------------|---------------------------|---------------------------|
| Service provider sell delay to customer $\Delta t_{SP,sell}$ | 1 year | 1 month | 1 week |
| Service provider capacity increase delay $\Delta t_{SP,incr}$ | 1 year | 1 month | 1 week |
| Service provider capacity decrease delay $\Delta t_{SP,decr}$ | 2 years | 1 year | 1 month |

We run the model for the 6 cases separately. For each case, the Saltelli sampler generates the 800-row matrix representing the uncertainties. We present two sets of results. First, we compare the utilization of operators and service providers over time in Figure 2-8 to understand the impact on the dynamics, and second, we record the mean value and standard deviation of the distribution of the cumulative revenues for the final time step in tabular form (Table 2-5 and Table 2-6).

Impact on dynamics

Figure 2-8 shows six plots where each horizontal set of two plots corresponds to one parameter. The left plots show the utilization of the service provider over time, and the right ones the utilization of the operator. We distinguish the different cases by the color of the lines: the baseline in black, the first reduction in blue, and the second reduction in red. The lines represent the deterministic simulation cases without uncertainty.

We start with the sell-delay to customers $\Delta t_{SP,sell}$. It affects the sold capacity of the service providers and, therefore, directly their utilization (see Figure 2-3 and Eqs. (2-11) - (2-12)). The upper left plot of Figure 2-8 also illustrates that. The shorter the delay, the higher the utilization and the more dynamic the service provider can respond to changes in prices and demand: the 2017 capacity expansion and the price drop is reflected in the utilization for a delay of one month and one week while it is smoothed in the 1-year baseline. In contrast, a more dynamic selling process of the service provider to the customer has no impact on the utilization of the operator since the model decouples these two feedback loops.

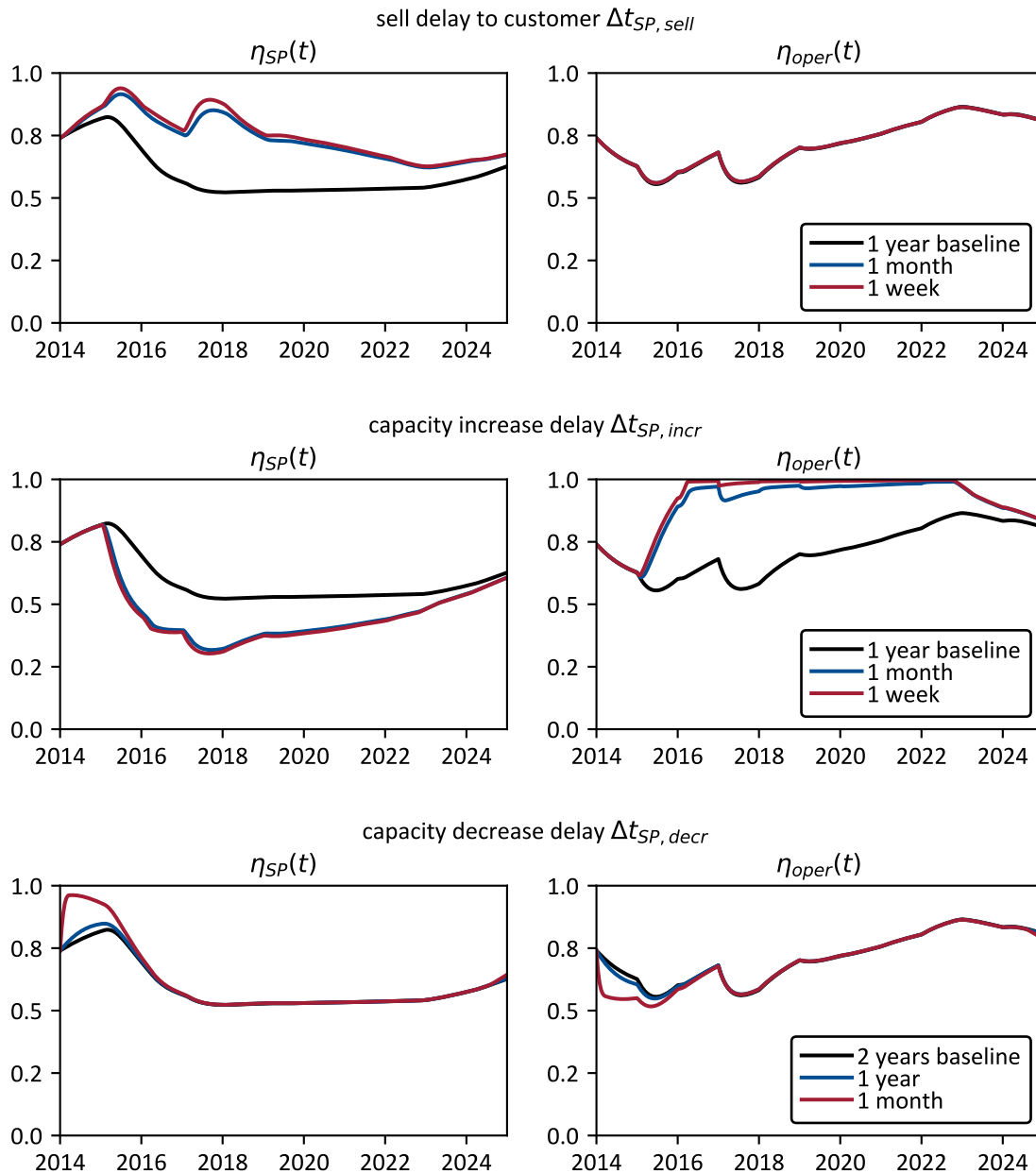


Figure 2-8: Impact of delay reduction on the utilization of service provider and operator.

The capacity increase delay $\Delta t_{SP,incr}$ shows a different picture: it impacts both utilizations. The delay for the service provider to buy capacity from the operator is equivalent to the delay for the operator to sell capacity. On one side, as seen by $\Delta t_{SP,sell}$, a more dynamic selling results in higher utilization. On the other side, the service provider's utilization decreases since $\Delta t_{SP,sell} > \Delta t_{SP,incr}$, and therefore the service providers buy capacity in faster intervals than it can sell to the customers. When comparing the difference in baseline to the first and second reduction, we observe a diminishing return – reducing from

1 year to 1 month yields most of the benefit. In contrast, a further reduction to 1 week only shows minor changes. Furthermore, we observe that until 2015, $\Delta t_{SP,incr}$ does not influence the utilization, meaning that the service provider did not increase their capacity but decreased it as is confirmed by the final set of plots for $\Delta t_{SP,decr}$.

Reducing the delay for the capacity decrease has the opposite effect of the increased delay on the utilization. A shorter delay increases the utilization of the service provider and reduces it for the operator. $\Delta t_{SP,decr}$ impacts the market dynamics only if service providers sell back capacity to the operator (in our model between 2014 and 2015). Since the baseline is 2 years and the first and second reduction is down to 1 year and 1 month, the impact is largest between 1 year and 1 month.

In the next subsection, we will discuss the impact of reducing the delay for $\Delta t_{SP,sell}$, $\Delta t_{SP,incr}$, and $\Delta t_{SP,decr}$ on the bottom line, including the distribution caused by the uncertainties.

Impact on bottom lines

In this part of the analysis, we summarize the distribution of possible results from the 6 cases. We compare the cumulative revenues of the final time step for GEO and NGSO operators as well as the service provider. We use two statistical measures: the mean value and the standard deviation, displayed in Table 2-5 and Table 2-6, respectively. We compare each case with the baseline and calculate the relative difference in percentage points. We highlight the numbers in the tables that show a significant change.

Table 2-5: Mean value of the 6 cases using 800 samples of the three uncertainties.

| | | $\Pi_{GEO,cum}(t_{final})$ | | $\Pi_{NGSO,cum}(t_{final})$ | | $\Pi_{SP,cum}(t_{final})$ | |
|---|-----------------|----------------------------|------------------|-----------------------------|------------------|---------------------------|------------------|
| | | [billion \$] | rel. to baseline | [billion \$] | rel. to baseline | [billion \$] | rel. to baseline |
| | baseline | 55.3 | | 57.4 | | 63.1 | |
| Service provider sell delay to customer $\Delta t_{SP,sell}$ | 1 month | 55.3 | 0.0% | 57.4 | 0.0% | 77.2 | 22.3% |
| | 1 week | 55.3 | 0.0% | 57.4 | 0.0% | 78.3 | 24.1% |
| Service provider capacity increase delay $\Delta t_{SP,incr}$ | 1 month | 67.7 | 22.3% | 69.6 | 21.3% | 63.1 | 0.0% |
| | 1 week | 68.8 | 24.5% | 70.5 | 22.9% | 63.1 | 0.0% |
| Service provider capacity decrease delay $\Delta t_{SP,decr}$ | 1 year | 55.2 | -0.2% | 57.3 | 0.0% | 63.1 | 0.0% |
| | 1 month | 54.7 | -1.1% | 57.3 | -0.2% | 63.0 | -0.2% |

Similar to the observations from the previous subsection, the sell-delay of service providers to customers has only an impact on the service provider as this feedback loop is decoupled from the operator. By

lowering the delay from 1 year to 1 month, 22% more cumulative revenues can be generated, and an additional 2% more when reducing $\Delta t_{SP,sell}$ to 1 week.

A startling behavior is the impact of $\Delta t_{SP,incr}$. It increased the utilization of the operators while reducing it for the service providers. The bottom lines of the operators reflect this increase by an additional 22/21% to 24/23% of cumulative revenues. However, the service providers’ revenues are unaffected by the lower utilization since their sold capacity is not affected. It is the available capacity that increases faster and therefore causes lower utilization. Since we use the sold capacity for revenue calculation, the cumulative revenues are insensitive to $\Delta t_{SP,incr}$. Nevertheless, the available capacity is tight to cost, which is not captured by the market dynamics model. Hence, costs would increase $\Delta t_{SP,incr}$ while revenues remain constant, leading to reduced profits.

For the capacity decrease delay $\Delta t_{SP,decr}$, the impact is slightly negative for the GEO operator. As we saw in the previous subsection, this delay is only relevant for the first year, and it decreases the utilization of the operator, resulting in the -1% cumulative revenue drop. The impact for the NGSO operators and the service providers is negligible as capacity is growing in the market, and the capacity increase process dominates. If this trend would revert, we expect greater importance of the capacity decrease delay $\Delta t_{SP,decr}$.

The final question that we address is if a reduction in the delays has an impact on the robustness against the uncertainties. We quantify the robustness with the standard deviations printed in Table 2-6.

Table 2-6: Standard deviation of the 6 cases using 800 samples of the three uncertainties.

| | | $\Pi_{GEO,cum}(t_{final})$ | | $\Pi_{NGSO,cum}(t_{final})$ | | $\Pi_{SP,cum}(t_{final})$ | |
|---|-----------------|----------------------------|------------------|-----------------------------|------------------|---------------------------|------------------|
| | | [billion \$] | rel. to baseline | [billion \$] | rel. to baseline | [billion \$] | rel. to baseline |
| | baseline | 8.2 | | 31.0 | | 15.1 | |
| Service provider sell delay to customer | 1 month | 8.2 | 0.0% | 31.0 | 0.0% | 19.0 | 25.4% |
| $\Delta t_{SP,sell}$ | 1 week | 8.2 | 0.0% | 31.0 | 0.0% | 19.3 | 27.5% |
| Service provider capacity increase delay | 1 month | 9.6 | 17.0% | 37.8 | 21.7% | 15.1 | 0.0% |
| $\Delta t_{SP,incr}$ | 1 week | 9.8 | 18.3% | 38.3 | 23.4% | 15.1 | 0.0% |
| Service provider capacity decrease delay | 1 year | 8.2 | -0.5% | 31.0 | 0.0% | 15.1 | 0.0% |
| $\Delta t_{SP,decr}$ | 1 month | 8.1 | -1.4% | 31.0 | -0.1% | 15.2 | 0.2% |

We expect the standard deviation increases as the absolute value of the mean increases as well. An increase in robustness would display itself by a difference between the relative differences in the standard

deviation and the mean. We do not notice any significant difference by comparing these numbers between Table 2-5 and Table 2-6, implying that changing the delays does not impact the robustness.

In this Section, we analyzed how a reduction in the delays for selling, increasing, and decreasing capacity of the service provider impacts the dynamics of the utilizations and the bottom lines. We saw that a reduction from 1 year down to 1 month of the selling delay increases the utilization of the service provider and increase the cumulative revenues by 22%. The same shortening of the capacity increase delay increases the utilization of the operators and increases their cumulative revenues by 21%. The capacity decrease delay has a minor impact on the utilizations and bottom lines. In the next Section, we rerun the model for a vertically integrated market.

2.5 Vertically integrated market

As discussed at the beginning of this Chapter, the market trend is greater vertical integration and stronger partnerships between operators, service providers, and customers. We use our market dynamics model to investigate what impact such vertical integration could have. We model the dynamic effects of integration by reducing the time delays $\Delta t_{SP, sell}$, $\Delta t_{SP, incr}$, and $\Delta t_{SP, decr}$ to 1 day (the benefits of reducing the margins along the value chain are not modeled but it would contribute to the value of vertical integration). Similar to the previous Section, we record the cumulative revenues of the final time step (see Table 2-7).

Table 2-7: Comparison table of baseline versus a more vertically integrated market. The integration is modelled by reducing the delays $\Delta t_{SP, sell}$, $\Delta t_{SP, incr}$, and $\Delta t_{SP, decr}$ to 1 day.

| | $\Pi_{GEO, cum}(t_{final})$ | | $\Pi_{NGSO, cum}(t_{final})$ | | $\Pi_{SP, cum}(t_{final})$ | |
|-----------------------------|-----------------------------|------------------|------------------------------|------------------|----------------------------|------------------|
| | [billion \$] | rel. to baseline | [billion \$] | rel. to baseline | [billion \$] | rel. to baseline |
| Baseline | 56.4 | | 64.4 | | 67.5 | |
| Vertical integration | 69.4 | 23% | 78.5 | 22% | 83.9 | 24% |

The vertical integration yields an increase of 22% - 24% for both operators and the service providers. These numbers are naturally similar to the ones obtained in the previous Section. The changes to the delays have a mostly independent influence on the cumulative revenues. Therefore, Table 2-7 shows very similar results to a superposition of the individual contributions. This Section presented the final analysis done with the market dynamics model. In the following, we summarize the main conclusions and discuss the implications for the remainder of this dissertation.

2.6 Conclusions and implications

After describing the market dynamic model, we conducted a Sobol sensitivity analysis to understand the relative impacts of the three uncertainties: s_{NGSO} , b , and $m_{NGSO,demand}$. Furthermore, we discussed the dispersion in the results caused by the uncertainties. The next analysis investigated how a reduction in the delays $\Delta t_{SP,sell}$, $\Delta t_{SP,incr}$, $\Delta t_{SP,decr}$ impact the utilization of service provider and operators and their bottom lines. Finally, we studied what effect a higher vertical integration has on the market's revenues. We draw these main conclusions:

- The three uncertainties s_{NGSO} , b , and $m_{NGSO,demand}$ have a similar impact on the price and the cumulative revenues of the GEO operator; the uncertainty in the cumulative revenues of the NGSO operator and service provider is mainly driven by the supply side through s_{NGSO} .
- Prices decline further as NGSO constellations build out their capacity; NGSO operators have the most uncertain revenue development; the operators' revenues reach a plateau in 2024 and even slightly decline.
- A decrease in the selling delay $\Delta t_{SP,sell}$ bumps up the utilization of the service provider but does not affect the operator. Shortening the capacity increase delay $\Delta t_{SP,incr}$ reduces the service provider utilization but increases the operator utilization. Reducing the capacity decrease delay $\Delta t_{SP,decr}$ only affects the market in the first year, while capacity expansion is still slow.
- Reducing $\Delta t_{SP,sell}$ from 1 year to 1 month adds 22% additional revenues to the service providers. Shortening $\Delta t_{SP,incr}$ from 1 year to 1 month, adds 21% additional revenues to the operators. Changing $\Delta t_{SP,decr}$ does not meaningfully affect the bottom line.
- Vertical integration of the market ($\Delta t_{SP,sell}$, $\Delta t_{SP,incr}$, $\Delta t_{SP,decr}$ set to 1 day) boost the revenues by 22% - 24% for operators and service providers.

With these observations, we can answer the questions raised at the beginning of this Chapter. First, the new entrants will shake up the market by introducing considerable uncertainty, not only for their revenues but also for the price development and corresponding demand response. Reducing delays cannot mitigate this uncertainty. However, shortening the contract durations benefits all horizontals in the market. Vertical integration achieves similar gains.

The insights from the market dynamics model have central implications for the remainder of this dissertation. Since we are particularly interested in the operators' level, a reduction in $\Delta t_{SP,incr}$ will be most beneficial to them. New satellites with flexible payloads will have the necessary technological

capabilities to support more dynamic allocation of capacity. A reduction in the contract durations can also be the selling of novel SLA with more short-time nature as elaborated in Chapter 7 in detail. The demand and resource management decisions have to be automated to realize shorter-term SLA. The next Chapter 3 portrays that Revenue Management contains principles that support the structuring and automation of these decisions.

3

Satcom Revenue Management Framework

We introduced the broadband satcom landscape in the first Chapter and identified the market dynamics in Chapter 2. The outcome of this Chapter is the overall framework for Revenue Management in satcom, which builds the foundation for the remaining of the dissertation (see Figure 3-1 for the mapping between the framework's components and Chapters). The subsequent Chapters 4 - 7 suggest solutions to the four main challenges we identify. Chapter 8 describes the implementation and usage of the framework with actual data when possible.

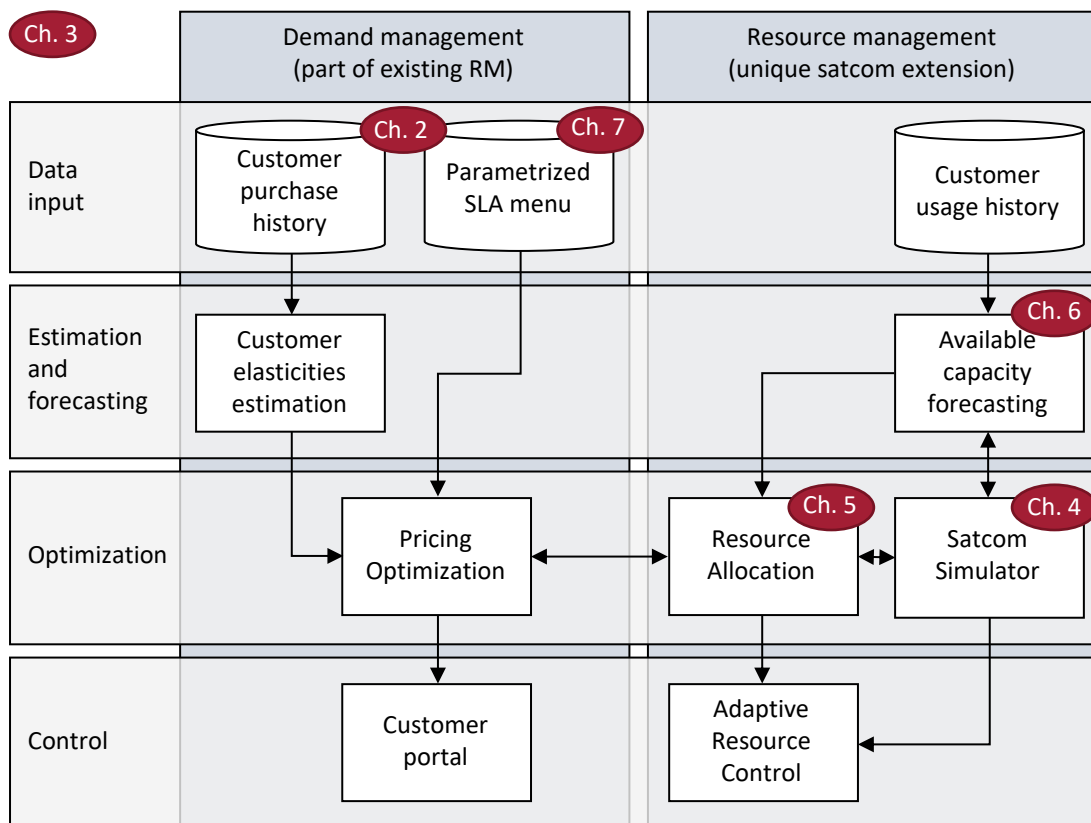


Figure 3-1: Preview of the proposed framework and document guidance, copied from Figure 1-4

In developing the satcom framework, we first shortly introduce the main concepts of Revenue Management in Section 3.1, check the applicability of the concepts to satcom in Section 3.2. Since to the best of our knowledge no work exist on HTS satcom RM, Section 3.3 reviews the characteristics of six other industries with successful Revenue Management systems. The review is summarized in Section 3.4 and compared to the satcom industry using our taxonomy. Finally, this yields to our proposed framework, which we present in Section 3.5. We illustrate its working principles in Section 3.6, and conclude the Chapter with a discussion of two application of the framework in Section 3.7 and a summarizes the main contributions in Section 3.8.

3.1 Primer on Revenue Management

Note that this Section is adapted from the author's original publication [63].

Revenue Management (RM) or also Yield Management⁴ is *“the process of allocating the right type of capacity to the right kind of customer at the right price so as to maximize revenues or yield”* [3]. More formally, RM is a conceptual framework for making sophisticated demand-management decisions⁵ in three categories: structural decisions about the selling format, pricing decisions, and quantity decisions [6].

The airline industry pioneered the development of many RM concepts. American Airlines claimed in 1992 that RM increased its revenues by \$900M annually (in 2019 dollars) [1, 2]. The underlying intuition is that empty seats are lost revenues and seats are priced different depending on the booking class. At one extreme, selling all seats at just above marginal cost would result in a high proportion of seats filled (i.e., load factor), but not in maximization of revenues. At the other extreme, selling only a few seats for very high prices is also not optimal, as many seats would remain empty [64]. Therefore, to maximize revenues, there is a trade-off to be made between the load factor and the average price (yield).

The stochastic nature of the problem adds further complexity. There are no-shows and late cancellations of bookings, which introduce uncertainty as to the actual number of empty seats remaining for sale. If customer segmentation is imperfect, lowering prices intending to attract customers from a segment with lower willingness-to-pay can lead to “buy-down” by customers with a higher willingness-to-pay. An RM

⁴ we are using throughout this dissertation Revenue Management; when authors refer to Yield Management, we rename it to Revenue Management to remain consistent.

⁵ this is the common understanding, but as we will discuss, we find a satcom RM includes resource management decision in addition to the demand management aspects

system manages these decisions through optimization and forecasting. It uses historical data when available and learns continuously to make better decisions for future flights.

At the highest level, RM models are either quantity- or price-based [6]. The difference is in the output of the optimization (the customer sees in both cases the price): a maximum quantity for a product with a given price, or the price for a product given an estimated quantity. Historically, distributing changing prices was impractical, and therefore airlines implemented quantity-based RM systems. However, with modern information technology, this is no longer a limitation, and notably some new airlines have moved directly to price-based RM. In the following, we illustrate with a basic example of the basic working principle of deterministic *price-based* RM (inspired by Talluri [6]).

Let us consider two separate customer segments with the two demand elasticity functions $d_1 = -p_1 + 230$ and $d_2 = -0.1 \cdot p_2 + 25$ with p_1 and p_2 being the prices. We further denote marginal revenues as

$$J(d) = \frac{\partial}{\partial d}(p \cdot d). \tag{3-1}$$

Therefore, we get $J_1(d_1) = -2 \cdot d_1 + 230$ and $J_2(d_2) = -20 \cdot d_2 + 250$. We indicate the total revenues using Π and assume that capacity is limited to 100, i.e., $d_2 = 100 - d_1$. The objective is to find the optimal prices p_1^* and p_2^* so that Π is maximized. That is achieved when the marginal revenues J_1 and J_2 are equal (any increase in quantity for one customer for an additional increment of revenues would require a decrease in the quantity of the other customer, resulting in a larger revenue increment lost). This problem can be solved by sweeping through all combinations of d_1 and d_2 as shown in Table 3-1.

Table 3-1: Example for price-based RM

| d_1 | d_2 | p_1 | p_2 | $J_1(d_1)$ | $J_2(d_2)$ | Π |
|-----------|-----------|---------------|---------------|------------|------------|---------------|
| 100 | 0 | 130.00 | 250.00 | 30 | 250 | 13,000 |
| 95 | 5 | 135.00 | 200.00 | 40 | 150 | 13,825 |
| 90 | 10 | 140.00 | 150.00 | 50 | 50 | 14,100 |
| 85 | 15 | 145.00 | 100.00 | 60 | -50 | 13,825 |
| 80 | 20 | 150.00 | 50.00 | 70 | -150 | 13,000 |
| 75 | 25 | 155.00 | 0.00 | 80 | -250 | 11,625 |
| 70 | 30 | 160.00 | -50.00 | 90 | -350 | 9,700 |

We find that the revenue maximizing prices are $p_1^* = 140$ and $p_2^* = 150$ (at which point the marginal revenues are equal). The RM will then communicate these prices to the two customer segments (assuming this is possible). Given the price-elasticities, we expect a demand of $d_1 = 90$ and $d_2 = 10$, clearing the capacity of 100. This example illustrates two segmentation scenarios:

One is where the selling takes place at a single time. In this scenario, the customers' characteristics define the segmentation bases, e.g., the geographical location. With the example above: customer segment 1 might be from a more price-sensitive country than customer segment 2.

In the second scenario, the customer segments arrive during two separate periods. For example, leisure air travel passengers tend to book earlier, and business travelers later. This division maps to the example's customer segments 1 and 2, respectively. The airline RM system will post a price that results in a demand of 60 for period 1 from leisure travelers with the expectation that it can charge more in period 2 for business travelers to clear the remaining 40. In a stochastic RM, these numbers have an underlying probabilistic distribution [65].

In reality, RM often makes use of both segmentation strategies. Airlines used a Saturday-night-stay requirement to prevent business travelers from buying-down to cheaper fares even if they book early. Moreover, they have an advance purchase requirement for the cheaper fares, effectively raising prices for bookings closer to departure day. Hotels and rental car companies exercise similar pricing schemes (see discussion in Section 3.3).

After the airline industry, hotels and rental cars were early adopters of RM with many others following [4]. But what are the industry characteristics that favor the use of RM? And what does that mean for satcom operators? Various authors (e.g., Weatherford [5], Talluri [6], and Kimes [3]) derived sets of conditions that favor RM. While the sets are slightly different, the common six conditions in favor are:

- Capacity is inflexible
- Capacity costs are high compared to marginal sales cost
- Inventory is perishable
- Customers are heterogeneous and can be segmented
- Demand is variable and uncertain
- Organization has data and information system infrastructure

The goal of the following Section 3.2 is to test the suitability of RM for satcom operators against these conditions.

3.2 Applicability of RM for satcom operators

Note that this Section is adapted from the author's original publication [63].

We use each of the six conditions that we introduced earlier to make as a qualitative test of the potential applicability of RM for satcom operators. We describe these conditions one by one below and summarize our findings at the end of the Section.

RM Condition 1: Capacity is inflexible

The total capacity of communication satellites is defined during the design process and is fixed for the lifetime of the satellite (ranging between a few years and 15 years). In bent-pipe designs, the total capacity is fixed, as well as is the capacity distribution over the covered regions. With new flexible satellites, there is the option to reallocate capacity from one region to another (the total capacity stays the same). Constellations with many smaller satellites have some flexibility to expand their capacity by launching new satellites or withdraw capacity by stopping to replace decommissioned satellites.

Nevertheless, such changes are likely to take multiple years to result in significant capacity changes. As a result, we consider the capacity to be inflexible, which makes it virtually impossible for a satcom operator to match the capacity to unpredictable short- and medium-term changes in demand. RM can reduce this mismatch by managing the demand side of the equation.

RM Condition 2: Capacity costs are high compared to marginal sales cost

The capacity costs (both in terms of time and money) of launching a large communication satellite or a constellation of smaller satellites are substantial. They range between many hundreds of million USD and several billion USD. To assess the marginal sales costs, we consider two changes to demand. First, demand changes for existing customers. Given the flexible nature of new satellites and automated resource management, the marginal sales cost of this change is negligible. Second, adding a new customer: the marginal sale cost is mainly driven by the customer terminal, which ranges between multiple 100 USD for small GEO Very Small Aperture Terminal (VSAT) terminal and several hundred thousand USD for high-end 3-antenna systems for MEO [66]. Emerging flat panel antennas have the potential to provide lower-cost alternatives [67]. Low marginal sale costs allow RM to change demand more frequently and therefore be more reactive. However, as we illustrated, terminal costs are not negligible and hence limit the frequency and extent of changes. This observation leads us to the separation into more frequent changes to existing customers and the less frequent adding of new customers. As we will see in Section 3.3, this characteristic has a strong analogy with air cargo.

RM Condition 3: Inventory is perishable

The inventory of satcom operators is the power and frequency spectrum available on their satellites. Perishability means that when a resource is unused at the moment, it cannot be stored and used at a later time. The resource is spoiled and has an opportunity cost. The perishability property is right for spectrum, but not necessarily for power: a satellite has onboard batteries. The storage capabilities are limited and depend on design, orbit, and demand patterns. The partial imperishability of power introduces a time-dependency into the RM formulation on the capacity level. While this is not a limitation per se, it adds complexity. A reasonable first approximation is to assume a constant average power limit per orbit (to account for the smoothing capabilities of onboard batteries) and consider the power to be perishable.

RM Condition 4: Customers are heterogeneous and can be segmented

Satcom operators have a variety of customers. An example segmentation is along the following three dimensions. One is the type and size of the customer: a single terminal end-customer, a multi-location end-customer, or a wholesaler. Another can be based on geographical attributes, such as the country or the location of the terminals: land, sea, or air. Third, the gateways can be owned and operated by the satcom operator or by the customers themselves. All of these three dimensions are directly observable by the satcom operator and hence allow for the potential of effective segmentation. A successful segmentation approach is identifiable, stable over time, and has a substantial, homogenous customer base within each segment [68]. The more different the segments are between themselves, the better the RM can differentiate between them. Section 7.6 discusses segmentation in more details.

RM Condition 5: Demand has variability and uncertainty

Variability is changing over time. It can occur on different timescales (hours, days, months) and can be periodic, repeating, or single events. Depending on the nature of the demand, it is more or less precisely predictable. Additionally, there is uncertainty about future demand, such as how the market evolves, how competitors behave, and how technology develops. To understand the variability and uncertainty of demand for satcom, we observed and analyzed data from SES's customers [69]. We detected that depending on the customer, the variability and uncertainty is considerably different. The relative uncertainty becomes smaller if more demand is aggregated. We also note, that the most distinct periodic pattern is diurnal with differences in day and night usage exceeding 50% in many cases, see Figure 3-2 for an example (normalized and shifted in time). We fitted a Gaussian Process to one month worth of actual measurements. We randomly sampled one thousand points and used a kernel with a sum of a radial basis

function and white noise. The drop in night usage closely correlates with the local time-zone and, therefore, longitude. From the view of the satellite, this results in an additional geographical variation along the longitude. RM can exploit these variations and uncertainties by multiplexing to utilize the satellite better.

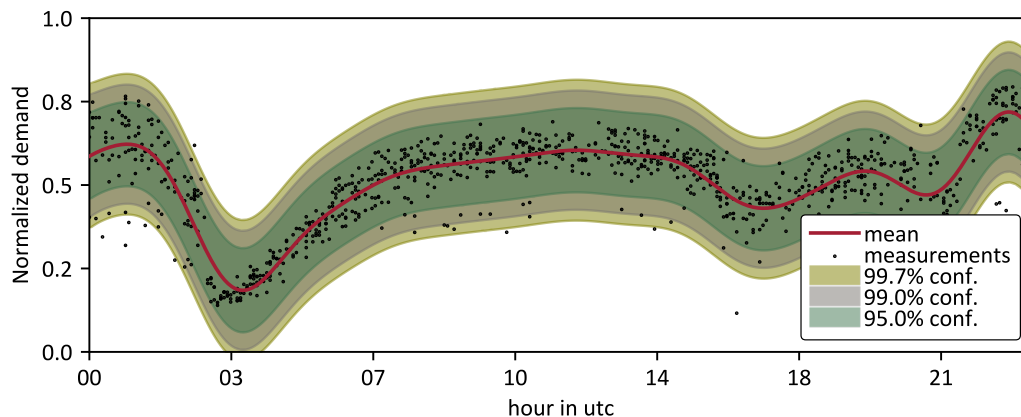


Figure 3-2: Normalized 24h data rate variation; one-month of sampled demand data fitted by a Gaussian process using a radial basis and white noise kernel.

RM Condition 6: Organization has data and information system infrastructure

While the first five conditions are at a market level, the last condition is organizational. We cannot make a general assessment for all satcom operators, but we use the example of SES S.A. The company plans to launch the highly flexible MEO constellation O3b mPower in 2021. To operate and control the satellites' many degrees of freedom, dynamic resource management (also referred to as Adaptive Resource Control (ARC)) is under development. A significant part of this development effort is to integrate data streams from ground segments and satellites into a commonly accessible place. The optimization engine of this tool then uses this data to generate a new configuration automatically. The tool commands this configuration then to the satellites. The deployed data and information system infrastructure are the foundation for the development of RM for satcom. The synergies between RM and automation tools like ARC help to shorten development time and reduces cost and risk (see Figure 1-4 for the overview and interaction between RM and ARC).

Summary of findings

We examined the satellite industry against the six conditions that favor the implementation of RM. We conclude that satcom meets all conditions, which supports our proposal of RM for satcom operators. In the following Section 3.3, we review the implementations of RM systems in six other industries and compare it to satcom in Section 3.4.

3.3 Review of RM systems in other industries

In this Section, we review RM frameworks from six industries to understand which components compose a RM system. Based on the categorization by Chiang [4], we divide the review into generic work in Section 3.3.1, work for the traditional RM industries: airline, hotels, and rental cars (Sections 3.3.2 - 3.3.4), and the non-traditional industries: air cargo, Internet services, and telecommunications (Sections 3.3.5 - 3.3.7).

3.3.1 Generic across industries

While the majority of work is done on RM for airlines, Talluri’s book on “The Theory and Practice of Revenue Management” [6] aims to provide a general view of RM. Talluri distinguishes between quantity and price-based RM approach and states that all RM decisions are of a demand management nature: “RM addresses the structural, price, timing, and quantity decisions a firm makes in trying to exploit the potential of this multidimensional demand landscape” [6 p.12]. Furthermore, in Talluri’s view, RM generally consists of the four functional steps: (1) data collection, (2) estimation and forecasting, (3) optimization, and (4) control. Figure 3-3 depicts the detailed process flow through these four steps.

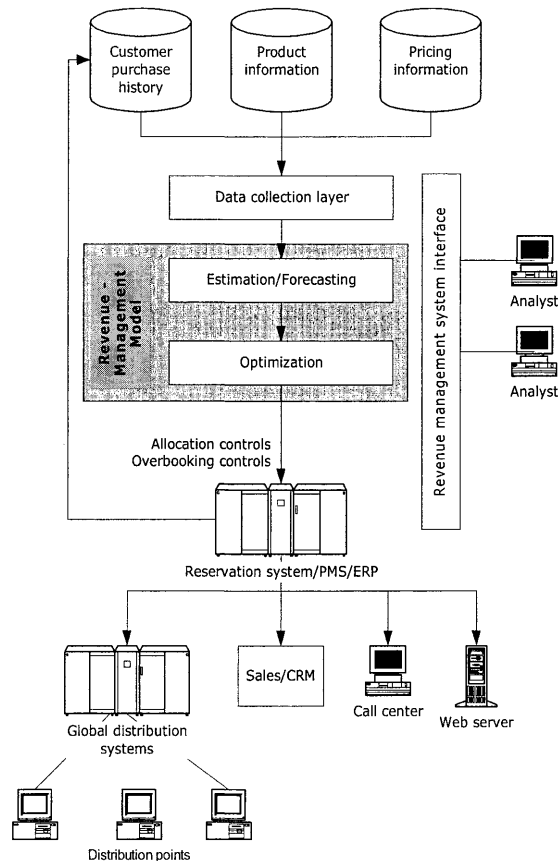


Figure 3-3: Talluri’s [6] representation of the RM process flow

The data collection layer gathers data from three databases: customer purchase history, product information, and pricing information. The RM model consists of estimation/forecasting and optimization and produces an allocation (quantity-based) or prices (price-based), which the reservation system stores. It is further distributed over several channels and fed back to the customer purchase history. Analysts can access the RM model to review suggestions made by the RM system. We will group the industry-specific RM system into quantity- and price-based wherever applicable and discuss their functionality in terms of (1) data collection, (2) estimation and forecasting, (3) optimization, and (4) control.

3.3.2 Airline

The research community has done exhaustive research for RM in the airline industry. Indeed, RM systems are used by airlines for over 30 years in their daily operation. Around half of the airlines develop their RM systems, while the other half purchases from external RM vendors (PROS, Sabre, and Amadeus are the largest) [70]. Due to the relatively long history, the airline industry has the most sophisticated RM systems, their high-level architecture has converged, and the challenges are well understood. Much of the research (in particular, the research with Operations Research (OR) emphasis), focuses on developing more intelligent and accurate segmentation, forecasting, and optimization algorithms to solve the known challenges.

Since our primary objective is to compare the high-level conceptual frameworks of different industries, we limit our review in this Section to a single author representative for the airline industry. Belobaba’s 1987 doctoral dissertation [71] is widely recognized as the first Ph.D. dissertation published on the topic of airline yield management. In his 2015 book on the global airline industry [64], he refers to three different generations of RM systems: the first developed in the early 1980s, the second in the mid-1980s, and the third generation in the late 1980s (see Figure 3-4).

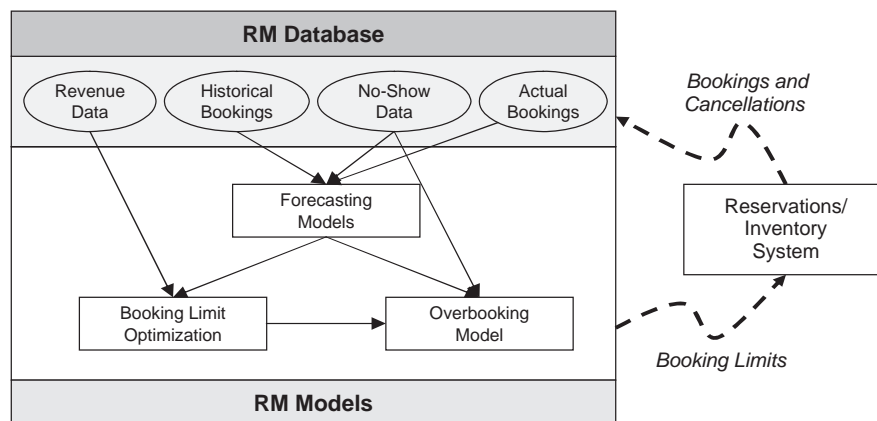


Figure 3-4: components of typical third-generation airline RM system obtained from Belobaba [64]

The framework combines historical and actual bookings with no-show data into the forecasting model. The booking limit optimization then uses this forecast together with revenue data. The forecasts flow into the overbooking model that uses the no-show data to determine an optimal overbooking. Belobaba lists four main capabilities of the third-generation airline RM system:

- “Collects and maintains historical booking data by flight and booking class.
- Forecasts future demand by flight departure date and booking class.
- Makes use of mathematical models to optimize total expected flight revenues, by determining both the optimal overbooking levels by aircraft compartment and the optimal booking class limits by booking class within each compartment (e.g. first, business, and economy classes).
- Provides interactive decision support for RM analysts, allowing them to review, accept, or reject the overbooking and booking limit recommendations.” [64, p.101]

Talluri’s general decomposition adopted many components of Belobaba’s. The data input layer consists of the RM database, followed by forecasting, and then RM models perform the optimization. The reservation/inventory system is the control layer. As we will see in the following Section, many of these components are found in other itinerary based industries as well.

3.3.3 Hotels

Hotels were the second industry where RM showed substantial gains. Kimes [3, 72] has done fundamental research to map the analogies. She generally decomposes RM into four main problems, which we will find in most of the reviewed RM systems:

- **Demand patterns for various rates/fares.** This problem is the best match for the historical booking component from Belobaba and Talluri’s customer purchase history. Kimes also includes the forecasting functionality, which is a separate component for Belobaba and Talluri.
- **Overbooking policy.** For Talluri, this is part of the optimization block, while for Belobaba, it is a separate block.
- **Demand elasticities.** The reaction of demand on price changes is captured in the Pricing Information database (Talluri) and the revenue data (Belobaba).
- **Information system.** The RM related functionality on the information system is reflected in the Reservation/Inventory System block from Belobaba and the control layer from Talluri.

Another relevant paper was authored by Vinod [73] from Sabre in 2004. It provides a functional decomposition of an implemented RM system for the hotel application (see Figure 3-5).

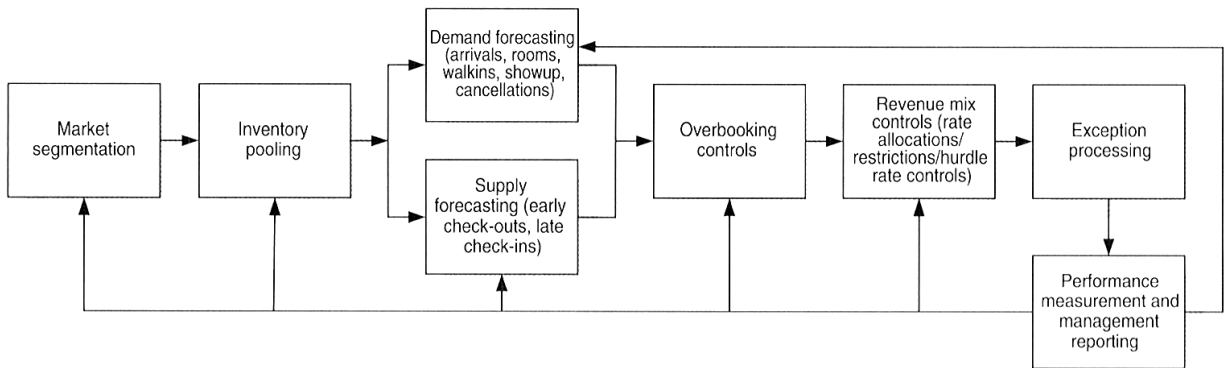


Figure 3-5: Vinod's [73] functional decomposition of the hotel RM system

Most of the functions and components can be easily mapped and found with slightly different wording in the research discussed before. However, there are three functionalities made explicit by Vinod [73]: market segmentation, inventory pooling, and supply forecasting. Airline RM research commonly segments demand into leisure and business travelers. Vinod proposes the two classes of qualified and non-qualified customers for Hotels. Qualified are customers who have corporate pre-negotiated rates. In Vinod's approach, corporate customers override RM decisions, reduce revenues during peak demand but increase utilization and average daily rate during off-peak days. The inventory pooling then groups the customers into 6-10 classes. Each class has a defined price (indicating a quantity-based RM) and its forecast. However, as additional functionality, the supply is also forecasted. Early check-outs and late check-ins cause the nonfixed capacity. The remaining steps are similar to the functionalities examined before.

3.3.4 Rental cars

Following airlines and hotels, rental cars was the third traditional industry of RM. In the early 1990s, automobile manufacturers flooded the market with excess capacity, which reduced prices. Car rental was a secondary business for manufacturers: it was more important to accommodate excess car production than making the car rental business profitable [74]. In this hypercompetitive environment, pure rental car companies had to reinvent themselves to survive. The implementation of an RM system helped to achieve the required competitive edge by improved inventory management and pricing [74, 75]. We review two successful RM systems, one by Carroll and Grimes [75] for Hertz, and the other by Geraghty and Johnson [74] for National Car Rental.

Note the analogy for satcom: LEO constellations cause a similar threat to legacy satcom players by flooding the market with capacity. If successful and demand is unable to catch up, they will add considerable excess capacity to the market. In that case, prices will continue to drop, placing considerable pressure on

incumbents and entrants. As the rental car case suggests, the implementation of an RM system could provide a way out.

Carroll and Grimes [75] describe the RM system developed and implemented by Hertz (1995). The system decides the availability of Hertz’s product combinations over time. For the particular case of Hertz, the authors list four major linked decisions:

- How should Hertz purchase and dispose cars over time for its fleet-level planning?
- How should Hertz deploy their cars?
- What mix of product availability maximizes Hertz’s net revenues?
- What mix of services should Hertz offer?

Hertz had a variety of other decision support systems for fleet planning, Daily Planning and Distribution Aid (DPDA), rates for their product offering, and cost allocation. The RM system (called Yield Management System YMS) builds on top of these systems (see Figure 3-6). The YMS has a graphical user interface and a control layer with reservation functionality.

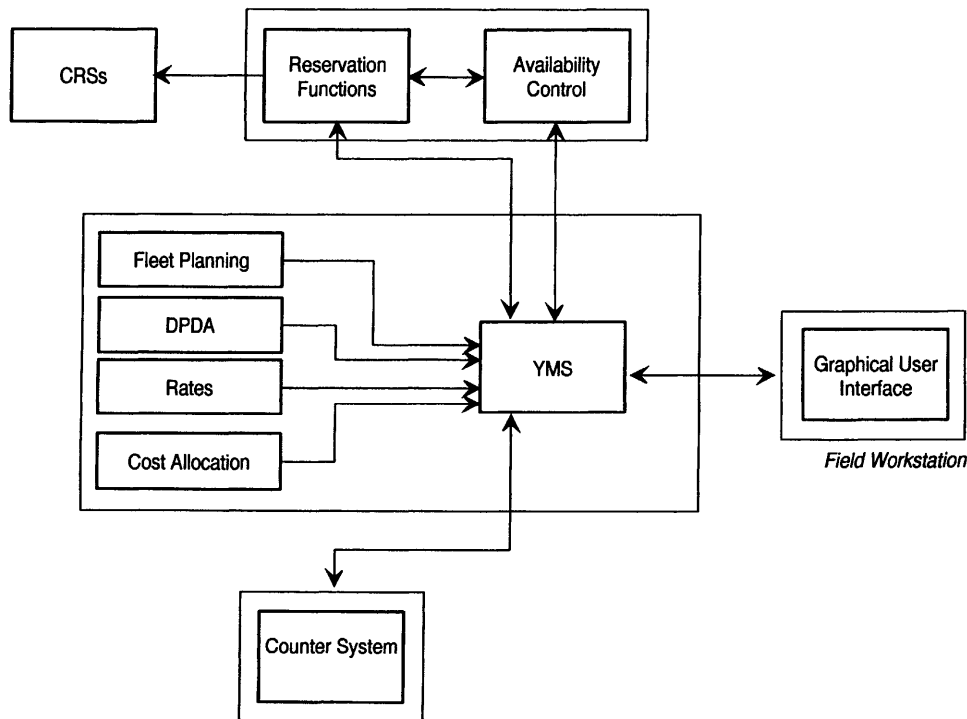


Figure 3-6: Carroll and Grimes [75] implemented car-rental RM system for Hertz; CRS=Computerized reservation system; DPDA= Daily Planning and Distribution Aid; YMS=Hertz’s Yield Management System.

Even though Carroll and Grimes [75] use different terminology, the functionality of their RM system represents all of the common components. The data input layer consists mainly of the four previously

existing decision-support tools, the YMS itself has forecasting and optimization capabilities, and availability control, reservation functions, and Computerized Reservation System (CRS) handle the control layer. One difference that we have not observed yet is the fleet planning component. The capacity of a rental-car company is fixed in the short-term. However, the mid-term (weeks/months) regional capacity is variable by moving inventory between locations; and selling and buying cars adjust the total long-term capacity. While the change of total capacity is more constrained for satcom operator, the change of regional capacity has the analogy to reallocation of resources from one spot beam to another.

Geraghty and Johnson described in their paper [74] how the implementation of an RM system saved National Car Rental and returned it to profitability in 1994. The implemented RM system supported three main functionalities: capacity management (addressing a high fleet utilization), pricing (leveraging customers’ different price sensitivities), and reservations control (accept and reject bookings). Figure 3-7 depicts an overview of the processes that National Car Rental implemented.

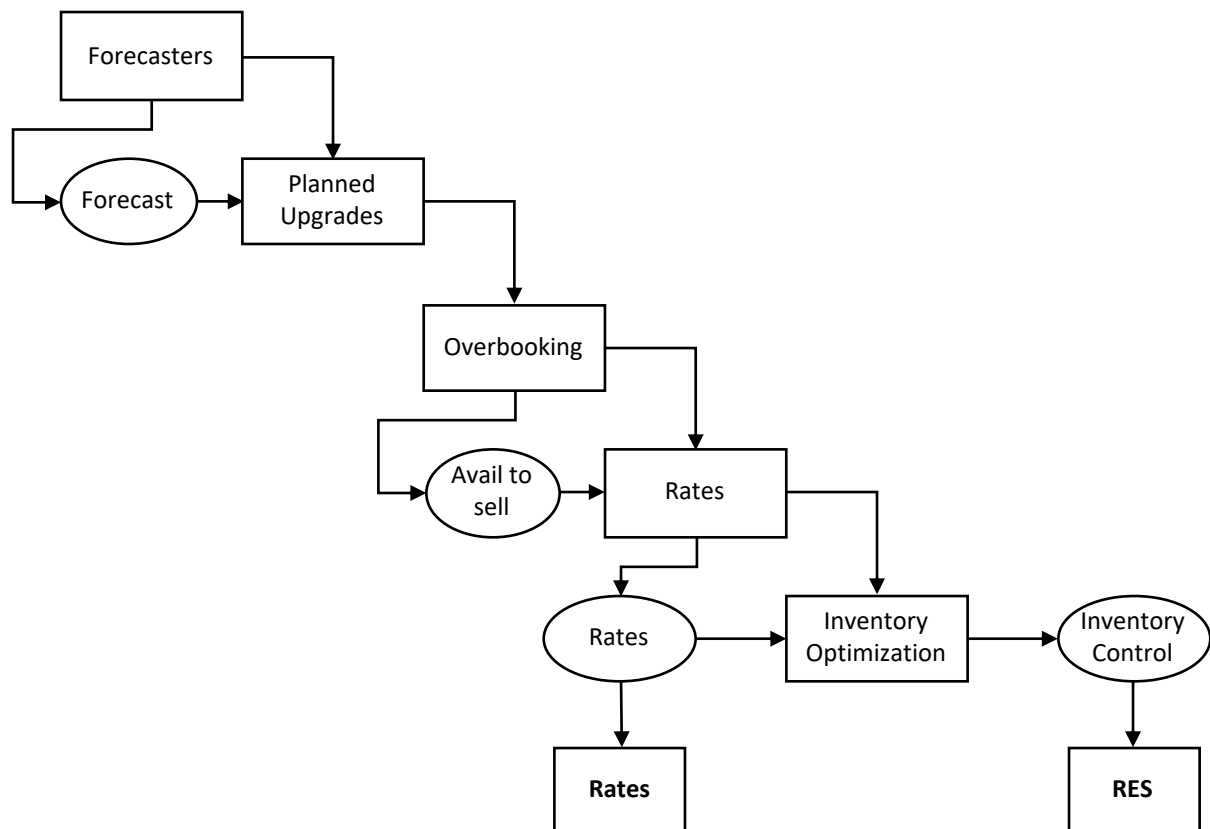


Figure 3-7: The processes and components of the RM system implemented by National Car Rental; obtained from Geraghty and Johnson [74]; RES=Reservation System.

The process represents a waterfall approach where each step depends on the outputs of the previous steps and is triggered whenever the output of the previous step changes. There are no iterative loops, but

an analyst can influence the RM system at the point denoted by the ovals and trigger the downstream computations. The first step is the forecasting of several variables as we have seen before. A uniqueness of Geraghty and Johnson [74] is that the planned upgrade step, which ensures that no customer receives a downgrade. As in airline RM, the overbooking accounts for no-shows and late cancellations.

In contrast to Carroll and Grimes [75], the rates are not predefined but are an integral part of the RM system. This pricing step tightly interacts with the inventory optimization, making it a joint price- and inventory-based RM system. In the final step, the reservation system (RES) posts the rates and controls the acceptance or rejection of bookings.

3.3.5 Air cargo

Air cargo is the first non-traditional RM industry that we review. We start with Kasilingam [76, 77] who highlighted the differences between passenger RM and cargo RM: first, the uncertain and a three-dimensional capacity (weight, volume, position). Second, cargo RM is service versus itinerary based, and third, cargo RM has reserved allotments for crucial customers (business-to-business (b2b) vs. mainly business-to-customer (b2c) for airlines, hotels, and rental cars). These differences are also found by Billings [78], and Slager and Kapteijns [79]. The differences require additional components for a cargo RM. Kasilingam's [76] proposed model consists of the seven components shown in Figure 3-8.

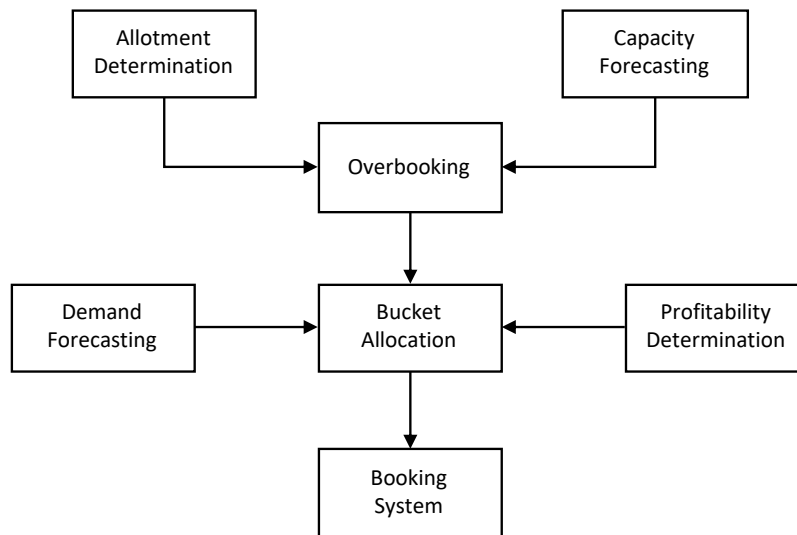


Figure 3-8: Kasilingam's [76] cargo RM model (redrawn for better readability)

The author divides the cargo RM process into four steps (not directly reflected in the figure) with the last two steps being similar to airline RM: (1) forecast the available cargo capacity, (2) allocate space for allotments, (3) overbook the remaining capacity, and (4) maximize revenue by allocating capacity to

different markets and products (quantity-based). While the essential functional components of data input, forecasting, optimization, and control are represented in cargo RM as well, the differences result in additional components that Kasilingam highlights. These are relevant to satcom. The total capacity of a satellite is known, but the regional capacity is uncertain. The capacity has two dimensions: power and bandwidth (however, we propose an approach to reduce them to one dimension). Satcom products are service-based through SLAs and have durations of 1-3 years. Few major customers reserve a significant portion of the capacity, and most external relationships are b2b.

Slager and Kapteijns describe in their practice paper [79] how KLM successfully implemented an RM system in 2002. The authors divide the selling of cargo capacity into two categories: contracted with a capacity guarantee (similar to Kasilingam’s [76] allotments) and free-sale booking without a guarantee (see Figure 3-9).

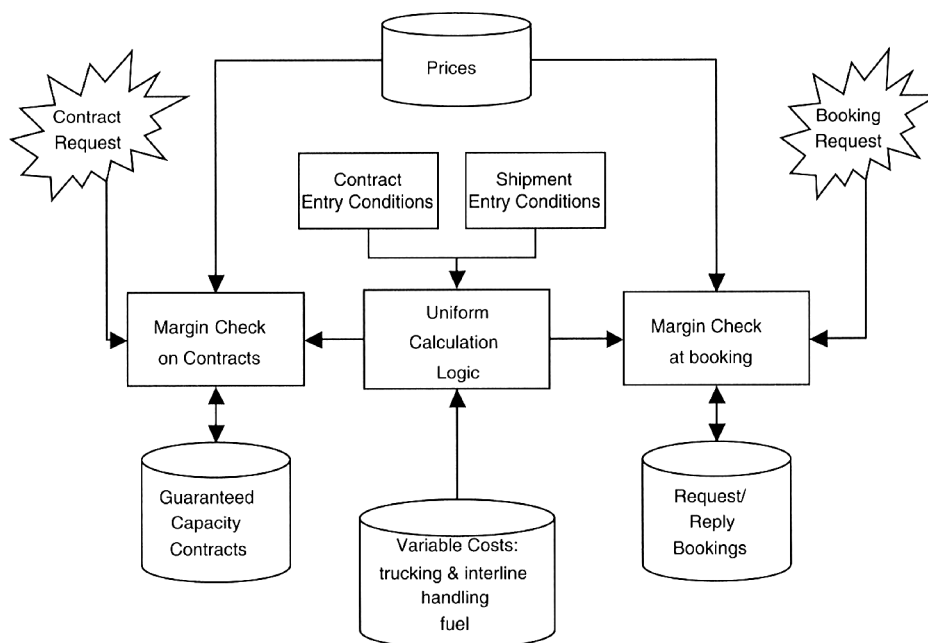


Figure 3-9: KLM’s cargo RM system (obtained from Slager and Kapteijns [79])

In cargo, the contracted capacity is renegotiated twice per year for the summer and winter season. During these periods, the contract and shipment conditions are entered into the RM system. The uniform calculation logic determines the optimal mix between contracted and free-sale bookings. The contracts are ordered based on their margin and checked against operational considerations. The contracts with the best margins and operational fit are signed before the season begins. If customers request a contract during the season, a decision is made based on a comparison against a minimum required financial

margin. A similar process is followed for booking requests regarding the free-sale capacity. They are checked against margin and then accepted or rejected.

In contrast to Kasilingam, Slager and Kapteijns optimize the mix between allotments and free-sale capacity. For both, accept/reject decisions are made based on the margin of incoming contracts or bookings. Both RM systems are quantity-based.

Becker and Dill [80] specifically focus on the complexity drivers for cargo RM. In their words, the goal of an RM is “to optimize the revenue and the profit of a flight event, a round trip and the whole network” [80 p.176]. Figure 3-10 illustrates the authors’ decomposition of the tasks that are necessary to achieve that goal.

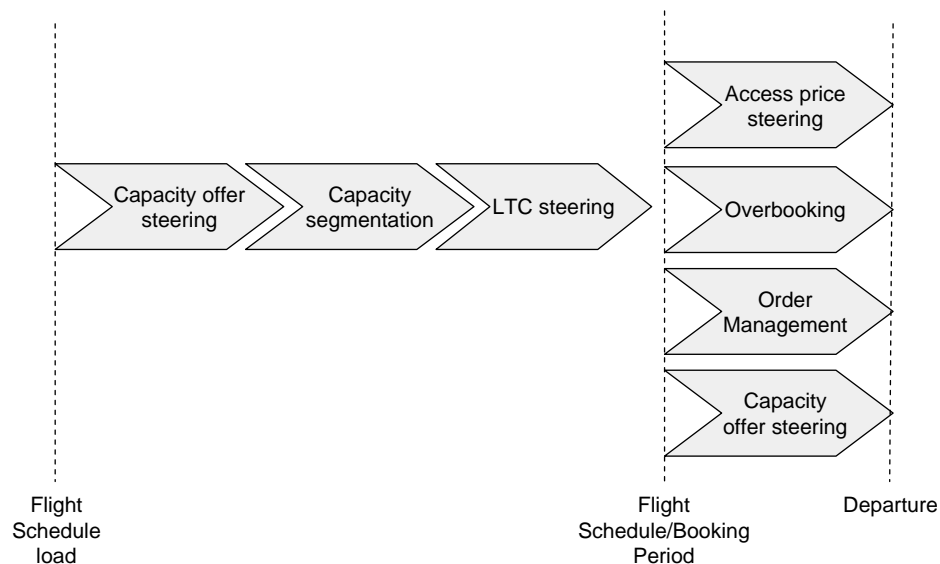


Figure 3-10: Outline of the RM process with its tasks: LTC=Long-Term Contract; obtained from [80].

Becker and Dill order the tasks based on their time of execution. Two timeframes are defined: between flight schedule load and booking period, and then booking period until departure. The first task is the capacity offer steering, which is the preliminary planning of the total available capacity (assuming cargo operators have this flexibility). Then, like Slager and Kapteijns [79], the capacity is segmented into long-term allotments and non-allocated capacity (sold in the last 30 days before departure). The Long-Term Capacity (LTC) steering trades-off the security of having a longer-term base utilization for a lower price versus more uncertain short-term bookings for a higher price. The remaining four steps are in parallel during the last 30 days before departure. The access price steering determines the optimal pricing for each segment by interacting with the capacity offer steering (making it the only price-based cargo RM reviewed in this Section). The overbooking accounts for no-shows and late cancellations. Based on the

prices, most bookings can be accepted or rejected automatically, and the order management supports those who require manual intervention.

3.3.6 Internet Services

Nair and Bapna [81] studied strategies of how RM principles can help to utilize better the network capacity of Internet Service Providers (ISPs, they provide the interface between customer and the Internet). Nair and Bapna define two levels, where RM systems make decisions. *Tactical level* (aggregate capacity, market segmentation policy, price-setting) and *operational level*, which are day-to-day decisions such as accepting and rejecting requests (most similar to the control layer). The authors point out three significant differences between ISPs and airlines and hotels:

1. “The ISP problem is inherently continuous both in state and time. True capacity is “modem-hours” rather than number of modems, and customers draw upon this capacity in a continuous fashion. There is no natural cutoff time which could be used as a time horizon to solve the problem. Airlines use flight takeoff time and hotels use 6 p.m. For ISPs, customers log-on and -off all day long.
2. Service is determined by the time it takes to get on the network. Thus, the request and the service happen simultaneously. This is not the case in airlines and hotels where the request is made at one time (making the reservation) and the capacity is used up at another (the flight taking off or the hotel room gets occupied).
3. For the above reason, overbooking is not an issue in YM for ISPs.” [81 p.352]

The first point certainly applies to satcom: satellite capacity and time is continuous. A cutoff time is defined for airlines, hotels, rental cars, and cargo. However, for satcom, the satellite capacity is contracted continuously over time with customers coming and going continuously (except for rather rare discrete events such as a satellite launch or decommissioning). However, the second point does not apply to satcom; the request and service happen at different times. A contract is signed, and then the service begins for a duration of time afterward. With that, the argument for the third point vanishes. As we will motivate later in more detail, overbooking is indeed a core functionality of a satcom RM (even though not based on no-shows or late cancellations).

Zhu et al. [82] and Byde et al. [83] worked on market-based resource allocation in data centers. Dube et al. [84, 85] researched in a similar direction for using RM on computing resources. We reference here Dube et al. work as their 2014 patent “Methods and Apparatus for Managing Computing Resources based on Yield Management Framework” [85] comprises, amongst other things, a flow diagram of their RM

methodology (see Figure 3-11). The methodology encompasses many functional components that we have already seen. It first builds predictions based on historical demand data. The following step 204 predicts the available capacity. The prices for given service levels are set, the total revenues computed, and then in the final step, the optimal prices, service-levels, and quantities are computed. The authors' focus on different prices and resource fits of products (or "price-service-level" combination) is a unique aspect of their methodology. Since satcom is service-based, this becomes relevant for our proposed RM system and a price-service-level" combination is close to an SLA.

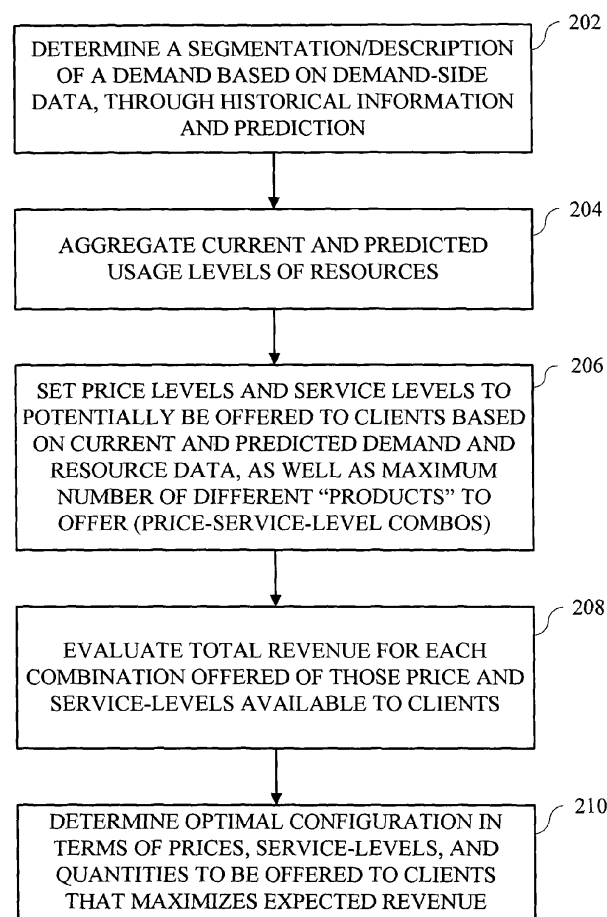


Figure 3-11: Flow diagram of the RM methodology for computing resources protected by patent US 8,788,310 B2 [85]

3.3.7 Telecommunications

Humair claims that his 2001 dissertation [86] was the first comprehensive study of how to make RM practical for telecommunications. He found the following four problems to be most prevailing on airline RM: forecasting, over-booking, seat-inventory control, and pricing. Humair proposes to optimize usage of

the spare capacity only while considering the base load as given. In analogy to cargo, the author’s definition of spare capacity would correspond to free-sale bookings. Since the current services are unsuited for the usage of spare capacity, the author proposes six services that make use of the spare capacity by creating flexible demand. As we will discuss in Chapter 7, current satcom SLAs are also not optimal. Therefore, we develop a novel set of satcom services inspired by Humair’s work.

The work from Goodman and Mandayam [87], and Saraydar et al. [88] focus specifically on resource management (see Chapter 5), they found that pricing could be an effective way to control power and guide users to a more efficient operating point of the system. Saraydar et al. use of pricing goes beyond an admission control mechanism and is setup as a noncooperative power control game. However, their approach is not directly applicable due to the different nature of satcom. Future work could be to adapt this work for the Adaptive Resource Control component of the framework (see Figure 3-13).

Jallat and Ancarani [89] link RM with dynamic pricing and Customer Relationship Management (CRM). The authors distinguish between RM and dynamic pricing, but dynamic pricing can also be interpreted as a part of price-based RM. Jallat and Ancarani link the strategies and tools to managerial issues, as illustrated in Figure 3-12.

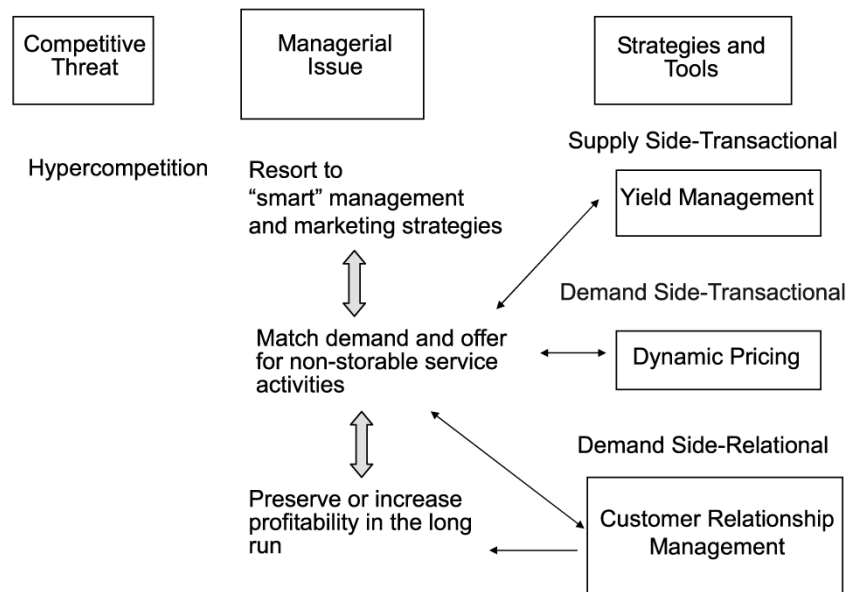


Figure 3-12: Integration of RM, dynamic pricing, CRM, and profitability as outlined by Jallat and Ancarani [89]

The center managerial issue outlined is to match demand and offer (capacity) for non-storable (perishable) service activities. As we saw throughout the previous Sections, RM and dynamic pricing are tools that can help to solve that issue. The new component is CRM, which impacts the shorter-term

mismatch between demand and capacity, but also the long-run profitability. CRM plays a crucial role, especially for b2b markets, with only a few customers. Since we observed the tendency of greater vertical integration in the satcom industry (such as SES's cruise ship segment), the nature of an operator's CRM could change dramatically. Even though it might remain mostly b2b, the number of customers and the variety of products might increase while contracts are updated more frequently.

3.4 Summary of reviewed industries and identification of the challenges

We reviewed six industries and discussed their characteristics in terms of RM systems. Our goal for this Section is to summarize and categorize the differences. By including the satcom industry in the comparison, we identify four major challenges. We start by using the taxonomy developed by Weatherford and Bodily (Section 3.4.1), use our own taxonomy for the comparison in Section 3.4.2, and summarize the challenges in Section 3.4.3.

3.4.1 Weatherford and Bodily’s taxonomy

Weatherford and Bodily [5] developed a 14-element taxonomy for perishable-asset RM problems. See Appendix B for the overview table, including proposed descriptors. We try to use their work to organize the RM systems that we reviewed. We find that the last ten elements describe the implemented RM system more than the specifics of the industry (which is our focus). Therefore, we concentrate on the first four elements (see Table 3-2).

Table 3-2: Summary of the RM reviewed using Weatherford’s and Bodily’s [5] taxonomy and their descriptors.

| | Airline | Hotel | Rental cars | Air cargo | Internet services | Telecom |
|-----------------------------|------------------|---|--|--|--|---|
| | Belobaba [64] | Vinod [73] | Carroll and Grimes [75], Geraghty and Johnson [74] | Kasilingam [76], Slager and Kapteijns [79], Becker and Dill [80] | Nair and Bapna [81], Dube et al. [85] | Humair [86], Jallat and Ancarani [89] |
| A Resource | Discrete (seats) | Discrete (rooms) | Discrete (cars) | Continuous (weight, volume), Discrete (position) | Continuous (computing resources, time) | Continuous (data rate, time) |
| B Capacity | Fixed | Fixed (but uncertain) | Total fixed, regional nonfixed | Nonfixed (and uncertain) | Fixed | Fixed |
| C Prices | Predetermined | Predetermined | Predetermined, Set jointly | Predetermined, Set jointly | Predetermined, Set jointly | Set jointly |
| D Willingness to Pay | Buildup | Buildup (and by product) | Buildup (and by product) | Buildup (and by contract duration) | N/A (by contract duration) | N/A (by product) |
| <i>Notes</i> | | <ul style="list-style-type: none"> Supply forecasting Market segmentation | <ul style="list-style-type: none"> Fixed short-term capacity Nonfixed medium-term by regional reallocating cars Nonfixed long-term by buying and selling cars Upgrade but no downgrade | <ul style="list-style-type: none"> Multi-dimensional capacity Optimize mix of allotment & free sale Service based contracts with b2b focus Access price steering | <ul style="list-style-type: none"> No cutoff time defined Different service-level combinations | <ul style="list-style-type: none"> Novel SLA products CRM integration |

We discover that the taxonomy does not capture the whole complexity and differences in the industry's relevant RM characteristics. Therefore, while using the suggested descriptors from Weatherford and Bodily, we further provide details in parentheses. The notes row contains a summary of the additional components as introduced by the corresponding authors.

Various other authors provided similar industry comparison tables: Carroll and Grimes [75] between airline, hotel, and rental car; Billings et al. [78] between airline and cargo; and Nair and Bapna [81] between airline, hotel, and ISPs (see Appendix C for their comparison tables). While Nair's and Bapna's base their table on the taxonomy from Weatherford and Bodily [5], they added the element cutoff time, overbooking, and split reservation demand into arrival and departure pattern. Carroll and Grimes, and Billings et al. proposed their own elements. We hypothesize that these other authors found it similarly challenging to fit the characteristics of the industry into the taxonomy (established initially to compare RM problems, not industries). Therefore, we develop our elements that capture the differences between the six reviewed industries more explicitly, as described in the following.

3.4.2 Taxonomy to compare RM characteristics across industries

Our goal by selecting the elements is to highlight the differences that are relevant for RM. While we believe that the taxonomy below is more suited to compare the six industries and highlight the challenges of satcom, we do not claim that it generalizes to every other industry for which RM is applicable. Furthermore, for a comparison of the specific RM implementation in traditional industries, Weatherford and Bodily's taxonomy is likely to highlight more details.

Our taxonomy has eight elements without dedicated descriptors. They are numbered with Latin numerals to avoid confusion:

- I. **Capacity unit.** This element describes the dimensions of measurement for capacity. We further specify if this dimension is continuous or discrete in parentheses to capture element A from Weatherford and Bodily.
- II. **Demand unit.** Similar to (I), this element illustrates how demand is measured. We specifically separated the measurement of capacity and demand to highlight that these two units can be different (such as it is the case for satcom while they are the same in the other six industries).
- III. **Resource allocation.** This element describes how capacity maps to demand. For airlines, if one unit of demand is sold (one seat), one unit of capacity is reserved (one seat). That is significantly

different for satcom: the reservation of one unit of capacity (e.g., power) for a given unit of demand (e.g., data rate) is not trivial.

- IV. **Contracted use.** This element describes how long a customer uses a resource. This element can contain a list of possible lengths for different products. It might also be referred to as the cycling-time of the service/product, i.e., for how long a customer occupies the capacity.
- V. **Type of business.** We consider two descriptors here: b2c and b2b businesses. There are some characteristics implicit: b2b have a smaller number of more significant customers, and CRM is often more personal, manual, and less automated.
- VI. **Capacity flexibility.** With this point, we aim to describe how flexible the capacity is during a given *booking period* within a *relevant scope*, i.e., a plane for a given route, a single hotel, a single car rental center, a regional telecom service, a geographical area in the field of view of the satellite/constellation. It is equivalent to Weatherford's and Bodily's element B.
- VII. **Cause of uncertain available capacity.** This element lists the uncertainties that the RM system leverages for *overbooking*. We chose not to name it overbooking because we want to highlight that both the total as well as the used capacity can be uncertain. The available capacity is the difference between these two and therefore captures uncertainties originated from both.
- VIII. **Base for differential pricing.** Since RM leverages the heterogeneity of customers, this element lists dimensions along which different prices can be charged. These can be different products, different booking times before service starts, different contract volumes, and different base willingness-to-pay (WTP) of verticals and geographical regions.

We use these eight elements to compare the six industries and satcom, as depicted in Table 3-3. Our focus in describing the table is to discuss the characteristics of the satcom industry and its differences.

Starting with (I), the capacity is continuous in power and bandwidth (bandwidth might be discrete depending on the technical specification of the payload). Similar to cargo, satcom has a multi-dimensional capacity with power and bandwidth (however, we will discuss later how we can make the capacity one dimensional in power for flexible payloads). Generally, we divide satcom into two generations of satellites: traditional satcom with almost no payload flexibility and new satcom with fully flexible payloads.

For traditional satcom, the main allocated capacity is bandwidth, and design fixes power. That is also why we included bandwidth as a demand unit in (II). Without flexibility in power, operators split the bandwidth between users within one wide beam. These satcom systems are often *bandwidth-constrained*. With

flexible payloads and spot beams, most of the services are managed services with demand being in data rate or volume. These systems are mostly *power-constraint*.

Table 3-3: Comparison of the industries’ characteristics along our eight elements.

| | Airline | Hotel | Rental cars | Air cargo | Internet services | Telecom | Satcom |
|--|----------------------------------|---|--|--|---|--|---|
| | Belobaba [64] | Vinod [73] | Carroll and Grimes [75], Geraghty and Johnson [74] | Kasilingam [76, 77], Slager and Kapteijns [79], Becker and Dill [80] | Nair and Bapna [81], Dube et al. [85] | Humair [86], Jallat and Ancarani [89] | |
| I Capacity unit | Seat (discrete) | Room (discrete) | Car (discrete) | Weight, volume (continuous), position (discrete) | Computing resources (continuous) | Data rate (continuous) | Power, bandwidth (continuous) |
| II Demand unit | Seat | Room | Car | Weight, volume | Computing resources | Data rate/volume | Bandwidth, data rate/volume* |
| III Resource allocation | Trivial ⁺ | Trivial ⁺ | Trivial ⁺ | Routing | Trivial | Routing | Resource optimization |
| IV Contracted use | One itinerary | One or multiple nights | One or multiple days | Allotment for half a year or one itinerary | Allotment for multiple years, or on demand use [§] | Allotment for multiple years | Allotment for multiple years or occasional use [#] |
| V Type of business | b2c | b2c | b2c | b2b | b2c | b2c | b2b |
| VI Capacity flexibility | Limited (Demand-driven-dispatch) | Fixed | Limited (moving cars between locations) | Fixed | Fixed | Fixed | Fixed |
| VII Cause of uncertain available capacity | no-show, late cancellation | no-show, late cancellation, early/late check-outs | no-show, late cancellation, early/late return | no-show, late cancellation, variable tendering | usage levels of resources | usage levels of resources | usage levels of resources |
| VIII Base for differential pricing | WTP buildup, compartment type | WTP buildup, room type | WTP buildup, car type | Capacity guarantee, contract duration, weight, volume | Service guarantee, contract duration, computing resources | Service guarantee, contract duration, data rate/volume | Service guarantee, contract duration, bandwidth or data rate/volume, vertical, region |

* bandwidth mostly for traditional satcom, all are continuous; ⁺ upgrade possible; [§] based on Amazon’s EC2 products [90]
[#] current practice but does not leverage new satellites’ optimally

The fact that demand has a different unit than capacity is a unique characteristic of satcom not found in any of the other reviewed industries. Its challenge becomes clear with (III), where the resource allocation of capacity to demand is an optimization problem itself. It is not trivial to find the optimal required power for the demanded data rate. As defined by Talluri [6], all RM decisions are of the demand management

nature. However, as we see here, the resource allocation decisions for satcom are of *resource management* nature, adding a new dimension to RM. The elements (IV) and (V) describe the business characteristics of the industry. Satcom is mostly b2b (with some b2c for residential broadband, e.g., through Exede by ViaSat), and the duration of use is multiple years for allotments and some short-term, occasional use. As we will detail in Chapter 7 the long-term contracts were well suited for traditional allocation of bandwidth, but they do not leverage the full potential of flexible payloads.

The next element (VI) captures how flexible the capacity is during the booking period in the proper scope. For satcom, the total capacity is fixed since the satellites are in orbit. For new satellites, the reallocation of total capacity to regional capacity is flexible, however.

The leading cause of uncertain available capacity (VII) in satcom is the usage of resources by customers. The analogy to cargo is variable tendering. Internet services and telecommunication experience similar uncertainties. Due to the different units of capacity and demand, available capacity does not directly translate to sellable demand. Since this relationship is non-linear, the uncertainties in demand also reflect differently on the available capacity, making the forecasting of the available capacity particularly challenging in the satcom context.

The last element (VIII) lists dimensions for differential pricing. Similar to more service-based industries (cargo, Internet services, telecommunication), service guarantee is a crucial differentiator for satcom. Longer contract durations and higher volume usually result in price discounts. In satcom, managed services are priced higher than pure bandwidth services. The vertical often defines the user terminal size and usage behavior. For example, aviation has a smaller user terminal with more variable demand making it more expensive (in terms of capacity) to serve than e.g., trunking. Hence the cost depends on the vertical, and different prices should be charged for each vertical to achieve a similar revenue per capacity unit. Another unique feature of satcom is its global access to demand with the same capacity. Therefore, the RM can leverage the differences in the WTP across regions, which is mainly driven by the Purchasing power parity (PPP) and available terrestrial alternatives.

3.4.3 Four major challenges in satcom RM

In the previous Section, we discussed the differences. Before we propose our satcom RM framework, we summarize the four unique challenges that are not found in another industry (see Table 3-4). Based on our distinction between traditional and new satcom, we assess in the table if the challenge is relevant.

Table 3-4: Summary of the four challenges in satcom and mapping them to traditional and new satcom. A checkmark ✓ denotes that the challenge is relevant.

| Challenge | Traditional satcom (Inflexible payloads, capacity: bandwidth, demand: bandwidth) | New satcom (Flexible payloads, capacity: power/bandwidth, demand: data rate/volume) |
|--|--|---|
| 1. Unit of capacity is not unit of demand | | ✓ |
| 2. Resource allocation is an optimization problem itself | | ✓ |
| 3. Uncertain available capacity based on resource usage | ✓ | ✓ |
| 4. Existing SLAs do not fully leverage the new satellites' flexibility | | ✓ |

In particular, the combination of challenge 1 and 2 add an essential dimension to the common understanding of RM systems. They contain decisions about the management of resources, not only demand, which will drive the proposed framework in the following Section. Dedicated Chapters address each of the four challenges.

3.5 Proposed Satcom Revenue Management framework

As described, Talluri [6] decomposes RM into four layers: data input, estimation and forecasting, optimization, and control. Data is collected from different locations, conditioned if needed, and stored in one place. From this, an estimator is built that also can be used to forecast. Once optimization is triggered, with these estimates and forecasts, the system finds a set of optimal policies. The final control step makes decisions based on the optimal policies and manages the day-to-day booking transactions.

The conducted literature review suggests that these layers are indeed a good description of what functions the RM systems are performing in various industries. We identified four challenges that are unique to satcom. Each is addressed by a separate component, which we include in our satcom RM framework. The objective of this Section is to compose a framework including all building blocks required for a satcom RM and group them into the four layers. Figure 3-13 provides an overview of the building blocks and the red numbers map to the four challenges from Table 3-4.

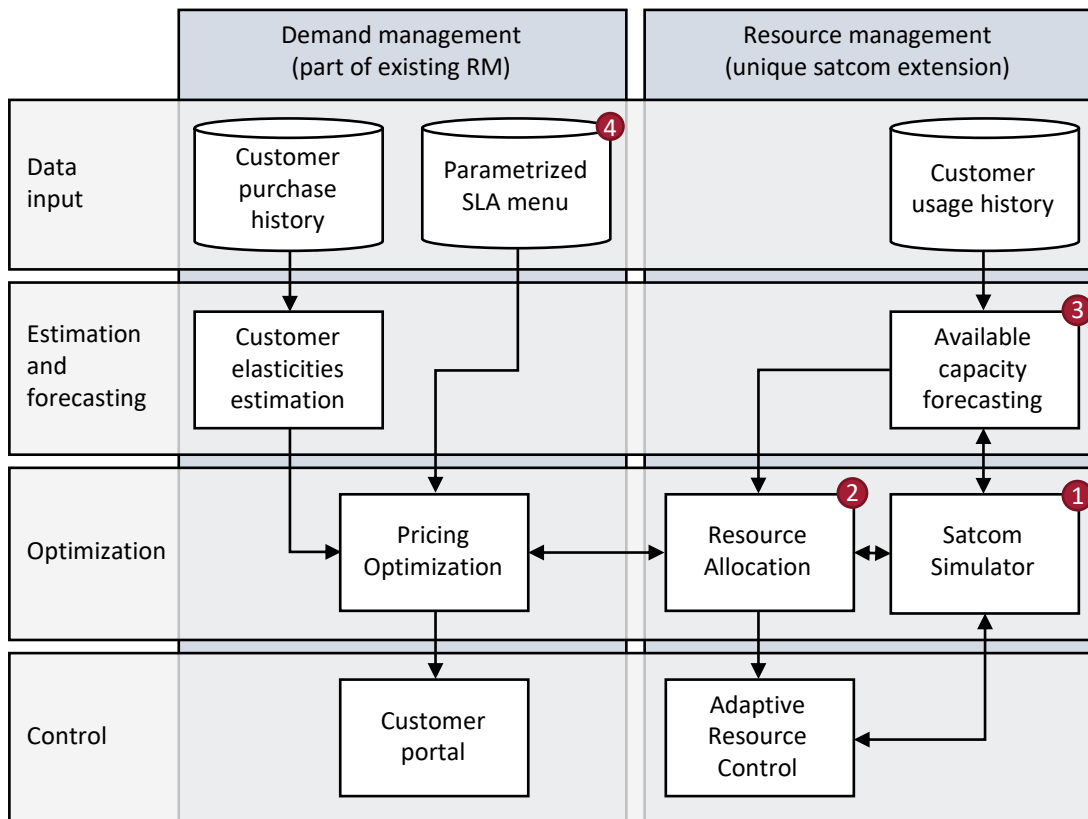


Figure 3-13: Overview of our proposed satcom RM framework mapped to the four challenges from Table 3-4: (1) unit of capacity is not unit of demand, (2) resource allocation is an optimization problem itself, (3) uncertain available capacity based on resource usage, and (4) existing SLAs do not fully leverage the new satellites’ flexibility.

We describe each block in detail below and, if applicable, we outline from which author and RM system we borrowed the block. The following Sections are structured first by demand and resource management and then by the four layers.

3.5.1 Demand management

The demand management aspect of the satcom RM is similar to already existing systems. It consists of a customer purchase history that informs the elasticity estimation and a pricing optimization taking the elasticities and the parametrized SLA menu to determine optimal prices, which are then made available to the customer, e.g., through a customer portal.

3.5.1.1 Customer purchase history and parametrized SLA menu

While for a real implementation, it is a challenge itself to collect, store, and maintain the data, it is of little academic value to detail the various interfaces the databases have inside and outside of the company. Therefore, we assume that all the data is already in a clean and consistent form available and stored. We describe the two databases that we consider in the demand management part below.

Customer purchase history

This term is borrowed from Talluri [6] and contains historical contractual information. The analogy to airlines is Belobaba's Historical Bookings component. The database includes but is not limited to price, product, country, start time of service, contract duration, and customer names. The objective is to keep a collection of data for the following elasticity estimations. In the beginning, this database will mostly contain data on longer-term contracts and some for occasional use. It grows over time as the operator adds products.

Parametrized SLA menu

The products that satcom operators currently sell are not necessarily parameterized but often written documents and individualized [66]. For RM to work, the products need to be parametrized, through which a high degree of individualization is achievable. Since this is one of the four challenges introduced in Section 3.4.3, we devote Chapter 7 to filling this database. This database is similar to Talluri's product information database and Dube's et al. price-service-level combos (step 206 in Figure 3-11).

In comparison to the customer purchase history, this database is small in volume, containing only templates for the parametrization of various SLAs. Note that in contrast to most of the reviewed frameworks, we optimize prices and do not predetermine them, making our framework a price-based RM.

That has the advantage to explicitly exploit the trade-offs between price, service-level, cost, and elasticity explicitly.

3.5.1.2 Customer elasticities estimation

Generally, estimation is descriptive: it finds the parameters of a model to fit observed data best. Forecasting is predictive: it uses the model to calculate future unobserved values [6]. The input is conditioned data, and the output is a probability distribution or a point estimate.

In quantity-based RM, the elasticity of demand regarding price is implicit in the distribution of the arrival process. For example, customers accepting a full-fare price arrive in at a lower rate than those only accepting discounted fares. In the satcom case, prices are set depending on service level, cost of service, and estimated elasticities. Since an understanding of how customers will respond to pricing in different segments is key to every RM system, we made the elasticity estimation component explicit in our framework. Fed by the customer purchase history, this component's objective is to estimate various elasticities.

3.5.1.3 Pricing optimization

The type (or author) of optimization techniques for quantity-based RM vary: Booking Limit Optimization (Belobaba), Inventory Optimization (Geraghty and Johnson), or Bucket Allocation (Kasilingam) to name a few examples. They assume predetermined pricing. For service-based RM, Dube et al. is an example that determines the optimal pricing, service-levels, and quantities. We propose something similar for our price-based RM.

A relevant work for pricing optimization in satcom context is the work from Thraves [91]. He studied a two-part tariff pricing problem and contrasted a heuristic that makes use of a mixed-integer program with a dynamic programming approach. Both methods outperform the current pricing strategy – even for misspecifications in the assumptions. However, the work has critical limitations: Thraves does not consider different capacity costs for the operator and does not consider flexible HTS, where the reallocation of resources is not trivial. His work focuses only on the demand management of our satcom RM framework and does not address the resource management part.

The pricing optimization component takes customer elasticities and the product information into account and coordinates with the resource allocation component to understand the capacity costs associated with every unit of demand allocated. We discuss this particular interface in Section 3.5.3 in more detail.

3.5.1.4 Customer portal

Once the pricing optimization found an optimal set of prices, these move onto the control layer. On the demand management side, there is a customer portal that communicates the pricing information to the customer and can accept or reject bookings. Since satcom is a b2b business, this process is likely to be heavily augmented by manual intervention, in particular, during the first time. When a new customer comes on board, the customer purchase and the usage history databases are updated.

3.5.2 Resource management

The second aspect of the satcom RM is resource management, which is a unique extension to other RM systems. In this part, the customer usage history informs the available capacity forecaster, which interacts with the satellite network simulator. The forecasts are then passed to the resource allocation optimization that uses the satellite network simulator together with the pricing optimization to find an allocation that satisfies all constraints and optimizes multiple objectives. The adaptive resource control takes this solution.

3.5.2.1 Customer usage history

This database is directly fed by the modems that measure the data rate at the terminal of long-term allotment customers. The characteristics of the data streams vary by modem vendor, so all data needs to be brought into a standard form before storing. Adding information about special events or anomalies helps to build a more accurate estimator. Given the amount of information, this database is the largest of the three in the input layer.

3.5.2.2 Available capacity forecasting

The customer usage history database directly feeds this forecaster. The available capacity is the difference between the fixed capacity and the usage of the sold capacity. It is the primary source of uncertainty in satcom and an indispensable lever for overbooking. The goal of the forecaster is to provide the optimization algorithm with statistics of available capacity for any given time. A similar component is the supply forecasting from Vinod, the capacity forecasting from Kasilingam, and step 204 by Dube et al.

This challenge is the only one that is relevant for traditional satcom as well. However, the nature of the forecasting is considerably different, given a wide beam and no dynamic allocation of resources. The statistical multiplexing of traffic becomes a significant driver, whereas, for new satellites, the majority of multiplexing is achieved through power reallocation. We dive into the details of this component in Chapter 6.

3.5.2.3 Resource allocation and satellite network simulator

The optimization layer of the resource management part decomposes to the resource allocation and the satellite network simulator components.

Resource allocation

The resource allocation is a tremendous challenge of the satcom RM framework. It encompasses several optimization steps that have to be separated to make the problem tractable. The steps include the grouping of user terminals, the routing between user terminals and gateways, the frequency assignment, and the required power computation. Each of these steps has further sub-steps. Some are computationally expensive due to the dimensionality and non-linearity of the underlying physics. Chapter 5 depicts this component in detail together with proposed algorithms.

The resource allocation coordinates with the pricing optimization to determine the capacity cost of additional served demand (further discussed in Section 3.5.3). The available capacity forecaster informs this component and uses the satellite network simulator to model the response of the real satellite system given different resource allocations.

Satcom simulator

As a unique quality of the satcom RM, capacity and demand have different units. If either the satellite or the terminal is moving, the relationship is non-linear and a function of time. A sophisticated simulator is built to model the behavior of the real-world system. The simulator is called by the available capacity forecaster, the resource allocation, and the adaptive resource control.

3.5.2.4 Adaptive resource control

After the resource allocation finds an optimal solution, the adaptive resource control changes the configuration of the satellites in real-time to adapt to a changing environment. The solution of the resource allocation is robust enough so that if these changes are within the predicted ranges, the satcom system can be controlled to ensure contracted service-levels. Note that we referred in earlier work to this component as a real-time engine (RTE) [92], and Garau et al. compared several artificial intelligence optimization algorithms addressing this challenge [93, 94]. While this is an exciting and vital field of research, it is not the emphasis of this dissertation (apart from the short Section 5.8). We refer the interested reader to the references [92-94].

3.5.3 Interaction between demand and resource management

Since resource management is a unique component, the interaction with the traditional RM part needs attention. The information flows in both directions between the pricing optimization and the resource allocation component. The pricing optimization varies the amount of new demand considered for a specific price. The optimization needs to know how much capacity a unit of demand requires. Since this relationship is not trivial, it triggers the resource allocation with a set of new potential demand. The resource allocation then takes this additional demand into account, runs the optimization algorithm, and returns the amount of capacity the new potential demand would require. With that feedback, the pricing optimization can determine the resulting revenues. We repeat this cycle iteratively until the pricing optimization converges, and we find the revenue-maximizing set of prices (which directly translates to a set of capacities).

While this cycle is logically straightforward, there are computational challenges that come with it: the resource allocation optimization becomes the evaluation function of the pricing optimizer. As we will discuss in Chapter 5, the resource allocation optimization can take up to several hours to run, making just a few evaluations computationally challenging to track. Generally, there are three possible solution directions: (1) make the resource allocation optimization faster, (2) reduce the number of function evaluations required by the pricing optimization, and (3) approximate the resource allocation evaluation with a surrogate that is fast to compute. We decided not to dedicate a separate Chapter but cover these aspects in the relevant Sections.

In the following Section, we aim to illustrate the working principles of the proposed framework. We set up an example that has all of the components but with reduced complexity to highlight the crucial aspects of our framework.

3.6 Illustration of the working principles and potential gains

The focus of this Section is on the process of our satcom RM framework. We illustrate the procedure by using an example that captures all of the components. Furthermore, the results of the example will give us the first sense of potential revenue gains. Chapter 8 outlines four more exhaustive analyses with varying assumptions and uncertainty consideration. The following example represent the analysis of selling capacity through new SLAs (see Section 8.4.2). We start by introducing the setup of the example and the assumptions in Section 3.6.1, describing the resource allocation and the satellite network simulator in Sections 3.6.2 and 3.6.3. We outline the assumptions for the customer usage history database in Section 3.6.4 and detail the implemented available capacity forecaster in Section 3.6.5. For the demand management part, the customer purchase history, parametrized SLA menu, and customer elasticity estimation are covered in Sections 3.6.6 - 3.6.8. Section 3.6.9 describes the used pricing optimization algorithm. The following two Sections, 3.6.10 and 3.6.11, present the results and a discussion. Note that we adapt the text and figures from the author’s original publication [63].

3.6.1 Setup of the simulation and assumptions

We consider a GEO satellite located at 0° latitude and 0° longitude with an altitude of 35,629 km (see Figure 3-14). We focus on the forward downlink from satellite to customer. The satellite serves one customer from each of the five segments: aviation, backhauling, maritime, trunking, and vsat with the longitude and latitude coordinates listed in Table 3-5.

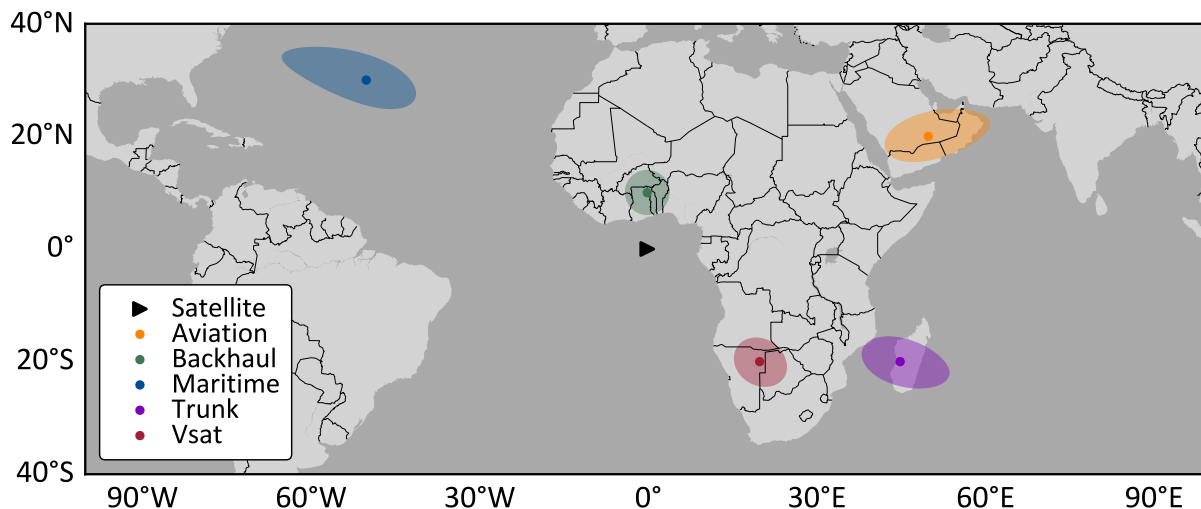


Figure 3-14: Overview map of the location of the five customers and the GEO satellite.

Each customer segment has a different characteristic which we describe in the following (see Section 7.6 for more details on market segmentation):

Aviation services provide broadband connectivity to commercial or private aircrafts. The end-customers are the passengers on board. All passengers are multiplexed and perceived as a single customer by the satellite.

Backhauling services connect a subnetwork with the fiber backbone of the Internet. Often the subnetwork is a cellphone tower or Wi-Fi hotspot in regions with no or low-speed terrestrial connection. Backhauling is often on larger scales and is mainly considered a viable solution for developing countries.

Maritime services are similar to aviation. The slower movements add fewer dynamics to the communication links. For large cruise ships, thousands of end-customers are multiplexed. The cruise ship cooperation often acts as an intermediary between satellite operators and end-customers.

Trunking services are similar to Backhauling with the focus on using satellites for excess demand in terrestrial networks or contingency scenarios. The traffic volume is often considerable, and less variable and uncertain since the link multiplexes many end-customers.

Vsat services address private end-customer and business solutions with light and variable traffic demands. The satellite operator often directly sells their services to the end-customer and provides connectivity to the Internet backbone.

Table 3-5: Assumptions about the customers input parameters.

| | Technical parameters | | | | | Demand parameters | | | | SLA parameters | | | |
|--------------------|----------------------|--------------|------------|----------------|-----------------|-------------------|--------------|-------------------|-------------------|----------------|------------------------------|-------------------------|------------|
| | Lon [°] | Lat [°] | D [m] | G/T [dBK] | Γ [-] | μ_b [-] | ν [-] | Δt [h] | σ_b [-] | CIR [Mbps] | p_{CIR} [\$/month/Mbps] | ϵ_{CIR} [-] | A [-] |
| Aviation | 50 | 20 | 0.6 | 15 | 1 | 0.4 | 0.30 | 7.33 | 0.3 | 70 | 300 | -0.5 | 0.99 |
| Backhauling | 0 | 10 | 1.2 | 21 | 2 | 0.5 | 0.30 | 4.00 | 0.2 | 150 | 100 | -0.9 | 0.99 |
| Maritime | -50 | 30 | 1.2 | 21 | 2 | 0.4 | 0.20 | 0.67 | 0.4 | 100 | 200 | -0.7 | 0.99 |
| Trunking | 45 | -20 | 2.4 | 27 | 3 | 0.7 | 0.05 | 7.00 | 0.1 | 250 | 100 | -0.8 | 0.99 |
| Vsat | 20 | -20 | 1.2 | 21 | 2 | 0.3 | 0.20 | 5.33 | 0.5 | 50 | 200 | -0.7 | 0.99 |

3.6.2 Resource allocation

We set up the example in such a way that the resource allocation is straightforward. The user terminals are separated far enough that interference is not a concern, and the satellite has enough available beams to give each customer their beam with a half-cone angle of 0.7° . Furthermore, we assume that there are no reuse constraints on the satellite payload, i.e., the center frequency is the same for all beams (20 GHz). The amount of bandwidth is then chosen based on a desired spectral efficiency Γ , as defined in Table 3-5)

for the mean usage. To not exceed the highest modulation and coding scheme (MODCOD), we let the spectral efficiency not exceed a certain threshold Γ_{max} , here chosen to be 5 (equals 64APSK5/6, see Appendix G). Formally, the amount of bandwidth is then obtained by:

$$B_i = \max\left(\frac{\mu_b \cdot CIR}{\Gamma}, \frac{CIR}{\Gamma_{max}}\right) \tag{3-2}$$

We assume that B_i does not vary over time and, therefore, does not depend on the demand management part of the RM system. Nevertheless, the required power does depend on demand. We formulate an analytical approximation in the following Section.

3.6.3 Satcom simulator

The relationship between the demanded data rate R and the power capacity C of the satellite is defined by the link budget [95]. Since we consider a GEO satellite, the geometry is not time-dependent and simplifies the model considerably. For analytical simplicity, we use the Shannon limit $R = B \cdot \log_2(1 + C/N)$ so we can write for customer i in linear form

$$R_i(t) = B_i \cdot \log_2\left(1 + \frac{C_i(t) \cdot G_{Tx} \cdot G/T_i}{L \cdot k_B \cdot B_i}\right) \tag{3-3}$$

with L being the free space loss that we assume here to be constant with 209.5 dBi. The gain of the satellite is G_{Tx} and assumed to be 51 dBi. The variable k_B is the Boltzmann constant with -228.6 dBK. Solving for $C_i(t)$ gives us Eq. (3-4) in which we define Z_i to be the first term.

$$C_i(t) = \underbrace{\frac{L \cdot k_B \cdot B_i}{G_{Tx} \cdot G/T_i}}_{Z_i} \cdot \left(2^{\frac{R_i(t)}{B_i}} - 1\right) \tag{3-4}$$

The exponent in the parentheses introduces a non-linearity that together with the customer specific properties G/T_i and B_i lead to significant differences in the capacity cost per data rate $C(t)/R(t)$. Table 3-6 illustrates that with using the $R(t) = CIR$.

Table 3-6: different capacity costs of the five customers

| | Specific capacity cost [mW/Mbps] | Cost increase relative to trunking |
|--------------------|-------------------------------------|---------------------------------------|
| Aviation | 5.7 | 682% |
| Backhauling | 2.9 | 348% |
| Maritime | 4.8 | 572% |
| Trunking | 0.8 | 100% |
| Vsat | 4.8 | 572% |

The first column shows the specific capacity cost in $mW/Mbps$, and the second column shows the cost increase relative to trunking, which is the “cheapest”. Given our assumptions, backhauling is 3x, maritime and Vsat are almost 6x, and aviation is almost 7x times more expensive to serve than trunking. The most significant driver for this is the difference in their G/T . The optimization will trade-off the higher cost with the higher willingness to pay of the segments.

3.6.4 Customer usage history

We model the customer usage history by parametrizing a cosine function for the five representative customers, as shown in Table 3-5 and illustrated in Figure 3-15. We choose the cosine function as it represents well the diurnal usage behavior (see Figure 3-2). The mean data rate $\mu(t)$ of the 24h demand pattern is modeled by a cosine with three parameters and is scaled by the CIR. Eq. (3-5) describes the relationship

$$\mu(t) = CIR \cdot \left(\mu_b + v \cdot \cos\left(\frac{t - \Delta t}{12} - \pi\right) \right) \quad (3-5)$$

with t being the hour of the day (in Coordinated Universal Time UTC), μ_b the normalized base mean, v the variation of the cosine around μ_b , and Δt the location of the minimum usage, i.e., the hour of the night drop with respect to UTC. The uncertainty around the mean is assumed to be proportional to the mean with a base sigma σ_b :

$$\sigma(t) = \sigma_b \cdot R(t) \quad (3-6)$$

Furthermore, we assume the demand has a normal distribution and therefore we define a stochastic process as a collection of independent random normals:

$$\{X_t\}_{t \in T} \quad \text{with} \quad X_t \sim N(\mu(t), \sigma(t)^2) \quad (3-7)$$

with X_t being a normal random variable and T being the 24 hours of the day. For the computations, we discretize T into 15 minutes. The demanded data rate $R(t)$ is then a sample from X_t .

Applying the parameters listed in Table 3-5 to Eqs. (3-5) - (3-7), we get Figure 3-15, which displays the demand pattern of the five customers introduced. All customers are modeled with a drop in night usage over a seven-hour interval. Trunking is the largest customer with a CIR of 250 Mbps, with a relatively small variation v and uncertainty σ . In contrast, Vsat has the lowest CIR with 50 Mbps and a higher uncertainty around the mean. Aviation and backhauling are in-between with aviation being more uncertain but uses, on average 40%, of the CIR while backhauling uses 50% ($\mu_b = 0.5$). Maritime is modeled with a higher uncertainty due to the more bursty characteristic of the customer segment.

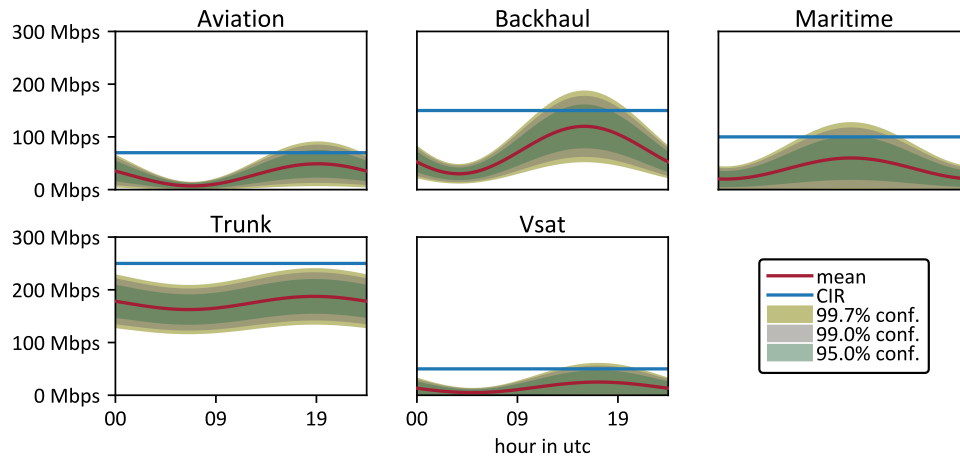


Figure 3-15: Demand over a 24 h period; confidence intervals are function of the mean (see Eq. (3-6)).

3.6.5 Available capacity forecaster

The available capacity forecaster determines how much capacity is available. It sums up the required capacity from Eq. (3-4) for each customer i and subtracts it from the maximum capacity C_{max} . In the deterministic case, we formulate

$$C_{avail}(t) = C_{max} - \sum_{i=1}^n C_i(t) \tag{3-8}$$

with n being 5 in this example. We set C_{max} to the power required if capacity is allocated based on CIR, i.e., $C_{max} = 1.77 W$. As we work with stochastic processes, we sample 50,000 times $R_i(t)$ from the demand $X_{i,t}$ and get a distribution of possible $C_{avail}(t)$ (assuming demand is not correlated). Since the distribution is no longer a normal random variable, we work with an empirical distribution function for each t . We denote the resulting random variable as Y_t . Figure 3-16 shows the resulting process $\{Y_t\}_{t \in T}$ including its confidence intervals.

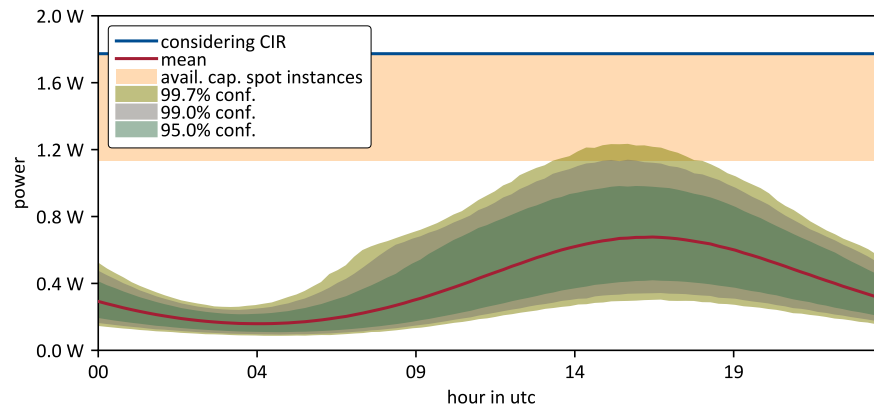


Figure 3-16: Plot of the used capacity modelled by the stochastic process $\{Y_t\}_{t \in T}$ after sampling $X_{i,t}$ 50,000 times for each i and t . The blue line illustrated the maximum capacity which is equal to an allocation based on CIRs.

The blue line is the capacity that would have to be reserved if allocating for CIR, i.e., without having flexibility in the payloads to follow the demand pattern and reallocated capacity. Due to the non-linearities, the uncertainty is no longer symmetric around the mean, particularly for higher data rates at Hour 15. The daily pattern follows that of the input demand data with a high-drop around Hour 5. While the 99 % upper confidence interval of the demand was touching the CIR line for the high usage period (see Figure 3-15), we can see from Figure 3-16 the gains of multiplexing customers. There is capacity available that can be used, even in the high usage period.

To calculate the available capacity in the stochastic case, we can rewrite Eq. (3-8) with the promised availability A from Table 3-5:

$$\mathbb{P}(\mathbf{Y}_t < C_{max} - C_{avail}(t)) \geq A. \quad (3-9)$$

We solve this equation for $C_{avail}(t)$ by inverting the empirical cumulative distribution function of \mathbf{Y}_t . This gives us how much of the capacity is available for reallocation at each time t while ensuring the promised availability A of the allotment SLA is met.

3.6.6 Customer purchase history

We assume the customer purchase history that is given by Table 4-1 through CIR and p_{CIR} . Furthermore, we assume that the elasticity in these points are known which gives us the elasticity estimation in Section 3.6.8.

3.6.7 Parametrized SLA menu

For this example, we define one product that makes use of the available capacity. We call it *spot instance* (in analogy to the alike Amazon Web Service product [90]). It is an add-on to the existing SLAs and is sold to the *same* customers. It works with the customers' price elasticity to generate additional demand. The price is below the long-term allotments, but the operator can quit the service at any time. That achieves a win-win situation for both sides. The customer receives additional data rate for a lower price, and the operator can safely overbook the satellite: if the usage of the long-term allotment changes, the operator can adjust to these changes by quitting spot instances. In contrast, if the operator overbooks with long-term allotments and usage behavior changes, adjustments can only be made after the long-term SLAs expire or are renegotiated. We provide more details regarding SLAs in Chapter 7.

The parameters of the *spot instance* are similar to the long-term allotment SLA. The product has a data rate and an associated price. We assume perfect customer segmentation is possible, and therefore, the parameters are different for each customer i : $R_{spot,i}$ and $p_{spot,i}$.

For further simplification, we make the conservative assumption that all customers use their full $R_{spot,i}$ throughout the 24 h window. With that, the most congested point is during peak hour of the existing SLA allotments. We get the available capacity by finding the minimum of all $C_{avail}(t)$. Formally we have:

$$C_{avail} = \min(C_{avail}(t) \forall t \in T) \tag{3-10}$$

which results in $C_{avail} = 0.63 W$ (see orange area in Figure 3-16) in our example. Note, that there is still the white area remaining between the 99% confidence interval line and the orange area. Hence, we only consider approximately half of the theoretical free capacity as available – the rest is the undulating white space shown below the orange area and above the green used capacity.

3.6.8 Customer elasticity estimation

For the customer demand elasticity, we assume an exponential behavior that goes through the price point defined by the long-term allotment (see Appendix D for a relevant background on demand elasticity). With $R_{spot,i}$ being the additional data rate contracted, we get the price $p_{spot,i}$ by

$$p_{spot,i} = \frac{1}{b_i} \cdot \ln\left(\frac{a_i}{CIR_i + R_{spot,i}}\right). \tag{3-11}$$

The elasticity ϵ for the exponential demand at the CIR price point is defined by Eq. (3-12) [6].

$$\epsilon_{CIR,i} = -p_{CIR,i} \cdot b_i \cdot e^{-1} \tag{3-12}$$

We calculate the parameter b_i by using the elasticity defined for $p_{CIR,i}$ from Table 3-5. With the values for $p_{CIR,i}$ and $R_{spot,i} = 0$ we obtain the parameter a_i . The resulting relationships are plotted in Figure 3-17. Aviation is modelled as the most price-insensitive customer segment, following by maritime and vsat, trunking, and backhauling. The upwards pointing triangle is the assumed price point of the long-term SLAs.

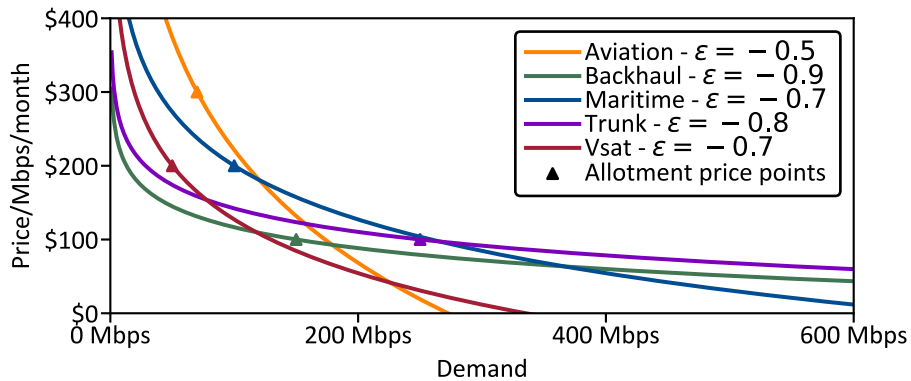


Figure 3-17: Lin-log demand functions for the five customers. The triangle is the price point of the allotments for which the elasticity ϵ is defined.

Furthermore, with the known parameters a_i and b_i , the additional revenue $\Pi_{spot,i}$ is obtained by multiplying Eq. (3-11) by $R_{spot,i}$:

$$\Pi_{spot,i} = \frac{R_{spot,i}}{b_i} \cdot \ln\left(\frac{a_i}{CIR_i + R_{spot,i}}\right) \quad (3-13)$$

The total revenues Π_{tot} are then defined as

$$\Pi_{tot} = \Pi_{CIR} + \Pi_{spot} \quad (3-14)$$

with $\Pi_{CIR} = \sum_{i=1}^n p_{i,CIR} \cdot CIR_i$ and $\Pi_{spot} = \sum_{i=1}^n \Pi_{spot,i}$ for customer i .

3.6.9 Pricing optimization

For revenue maximization we use a gradient-based greedy algorithm. We analytically derive the gradient $\partial\Pi_{spot,i}/\partial C_{spot,i}$ by inserting Eq. (3-3) into Eq. (3-13) (thus assuming the *spot instance* is provided through a new beam to the same terminal). For the purpose of clearer notation, we omit the customer index $spot, i$. With the defined variable $Z = \frac{L \cdot k_B \cdot B}{G_{Tx} \cdot G/T}$ from Eq. (3-4) we get:

$$\frac{\partial\Pi}{\partial C} = \frac{\partial}{\partial C} \left(\frac{B \log_2\left(\frac{C}{Z} + 1\right)}{b} \ln\left(\frac{a}{CIR + B \log_2\left(\frac{C}{Z} + 1\right)}\right) \right) \quad (3-15)$$

With the substitutions

$$D_1 = \frac{C}{Z} + 1 \quad \text{and} \quad D_2 = B \cdot \frac{\ln(D_1)}{\ln(2)} + CIR \quad (3-16)$$

the final gradient reads

$$\frac{\partial\Pi}{\partial C} = \frac{B \cdot \ln\left(\frac{a}{D_2}\right)}{b \cdot Z \cdot \ln(2) \cdot D_1} - \frac{B^2 \cdot \ln(D_1)}{b \cdot Z \cdot \ln^2(2) \cdot D_1 \cdot D_2}. \quad (3-17)$$

We calculate these gradients for every customer i and then follow the procedure outlined in Algorithm 3-1. The gradients are initialized with a capacity $C_i = 0$. Then, iteratively the capacity is increased by ΔC for the customer i with the steepest gradient, i.e., the highest marginal revenue. For the step size we choose $\Delta C = 0.001 W$. The algorithm terminates if either all gradients are negative (selling more capacity would reduce revenues) or the capacity limit C_{avail} is reached (capacity cleared). The result is the revenue maximizing allocation $C_{spot,i}^*$ for each customer i . With Eq. (3-3), we transform this capacity back into $R_{spot,i}^*$ and with the demand function in Eq. (3-11) we get the optimal prices $p_{spot,i}^*$. Since the marginal

revenue function $\partial \Pi_{spot,i} / \partial C_{spot,i}$ is monotonically decreasing, the found solution is ensured to be the global optimum [96].

Algorithm 3-1: incremental gradient-based optimization adapted from [96].

```

Input:  $J_i(C_i) = \partial \Pi_i / \partial C_i$ 
Input: step size  $\Delta C$ 
Input: available capacity  $C_{avail}$ 
Output: optimal allocation  $C_i^*$ 
1: Initialize  $J_i(C_i)$  with  $C_i = 0$ 
2: while  $\sum_{i=1}^n C_i < C_{avail}$  do           // repeat until capacity limit reached
3:   if any  $J_i(C_i) > 0$  then           // check if any gradient positive
4:      $C_{argmax_i(J_i)} += \Delta C$        // incr. capacity for steepest grad.
5:   else
6:     stop                               // stop when all gradients negative
7:   end if
8: end while

```

3.6.10 Results

The final results of the simulation are summarized in Figure 3-18 and Table 3-7. The algorithm terminated after allocating the complete available capacity. The upper left part of Figure 3-18 shows the spot revenues per month $\Pi_{spot,i}(C_i)$ in thousands of USD (\$k) as a function of the capacity, whereas the upper right part shows the spot revenues as a function of the data rate $\Pi_{spot,i}(R_{spot,i})$. The lower left subplot is the demand elasticity from Figure 3-17. The location of the circles in all three subplots denote the optimal solutions obtained for $C_{spot,i}^*$, $R_{spot,i}^*$ and $p_{spot,i}^*$.

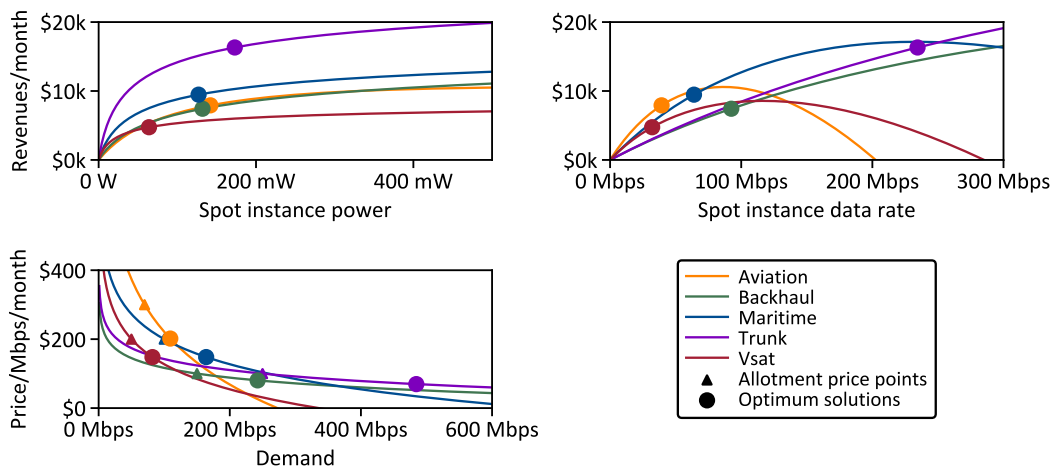


Figure 3-18: On the upper left, the additional revenues $\Pi_{spot,i}$ per month are plotted as a function of the spot instance power $C_{spot,i}$ and data rate $R_{spot,i}$ on the upper right. The lower left shows the same demand elasticity as Figure 3-17. The circles indicate the optimized solutions for $C_{spot,i}^*$, $R_{spot,i}^*$ and $p_{spot,i}^*$.

For the initialization of $C_{spot,i} = 0$ the circles in the lower left plot match the allotment price points. We can see in the upper left plot that the gradients $\partial\Pi_i/\partial C_i$ are all equal at the optimum solution points. Due to the strong non-linearity of the link budget (see Eq. (3-3)), the functions and the gradients with respect to R are of a very different form. For example, trunking has a steep gradient in the beginning for power, whereas it is almost linear with the data rate. The linearity can be explained by the flattening out of the demand elasticity for a cumulative demand above 300 Mbps. According to the Shannon limit, this linearity becomes the non-linear behavior seen in the upper left plot. An optimization on the gradients of $\partial\Pi/\partial R$ would lead to a different non-optimal solution. This observation underscores the importance of the translation between the data rate demand unit and power capacity unit.

Table 3-7: Results after optimization

| | $C_{spot,i}^*$ [W] | $R_{spot,i}^*$ [Mbps] | $P_{spot,i}^*$ [\$] | $\frac{p_{CIR,i} - P_{spot,i}^*}{p_{CIR,i}}$ | $\Pi_{CIR,i}$ [\$k] | $\Pi_{spot,i}$ [\$k] | $\frac{\Pi_{spot,i}}{\Pi_{spot}}$ |
|-------------|-----------------------|--------------------------|------------------------|--|------------------------|-------------------------|-----------------------------------|
| Aviation | 0.14 | 39 | 202 | 33% | 21 | 7.92 | 17% |
| Backhauling | 0.13 | 92 | 80 | 20% | 15 | 7.43 | 16% |
| Maritime | 0.13 | 64 | 148 | 26% | 20 | 9.47 | 21% |
| Trunking | 0.17 | 234 | 70 | 30% | 25 | 16.31 | 36% |
| Vsat | 0.06 | 32 | 148 | 26% | 10 | 4.75 | 10% |
| | | | | \sum_i | 91 | 46 | |
| | | | | Π_{tot} | 137 | +51% | |

Overall, under the assumption that all customers buy the *spot instances* for the posted price and quantity, they add \$46k in revenues per month to the baseline of \$91k (see Table 3-7). This equals an increase of 51%. The greatest revenues (around one third) come from trunking, with vsat contributing the least at 10% (see column $\Pi_{spot,i}/\Pi_{spot}$). The capacity is split almost equally between aviation, backhauling, maritime, and trunking (Vsat receives around half of that). The prices $p_{spot,i}^*$ are between 20% and 33% lower than $p_{CIR,i}$. The additional data rate is the highest for trunking with 234 Mbps.

After optimization, these determined prices would be passed on to the booking system (see Figure 3-13). The price-based RM system would communicate the prices for the *spot instance* product to the customers.

3.6.11 Discussion

Our objective of this Section was to illustrate the working principles of a possible realization of an end-to-end satcom RM system given the framework developed in the previous Sections. We described a statistical capacity estimator, which makes use of customer multiplexing, i.e., overbooking. The model is based on a Shannon link budget approximation, and we illustrate its non-linear behavior and the implications. For

the novel SLA, we discussed one possible product: spot instances that the operators sell to the same customer for a discounted price, but the operator can quit service anytime.

The most critical assumptions we made were around the novel SLA. The specification of the product has implications on the demand elasticities and the available capacity considered. While we assumed here that the elasticities are anchored around the price point of the long-term allotments, this is not necessarily the case or even known. An analogy is new flights for airlines. The demand curves are initially unknown, and the RM system learns them over time.

Moreover, we considered the conservative scenario in which customers make full use of the purchased data rate of the spot instances. It is likely that they show similar behavior to the long-term allotments and leave a considerable amount unused during the night. One approach can be to design a product that specifically makes use of this night drop (e.g., a discounted over-night backup plan for non-time critical data). Depending on the satellite's technical capabilities, power can be stored during off-times and used during peak hours.

The purpose of this Section was to give the *intuition* about the working principles of a satcom RM system. For a more complete picture of the analysis and the RM benefit, the reader is referred to Chapter 8.

3.7 Two main applications of the RM framework

We illustrated in the previous Section 3.6, the working principles of the framework. Indeed, we set up the example to highlight one of the two applications of the framework, the *optimal booking for a single time instance* (see Table 3-8 for an overview). The assumptions are that no new customers arrive, and operators sell the available capacity to existing customers through new products. The gains are achieved by leveraging the mismatch between actual and contracted usage and the customer segments' heterogeneity. Proper statistical analysis ensures that the overbooking level is compliant with the SLAs.

The second application of the RM framework is the *optimal filling of the satcom system over time*. Throughout time, new customers arrive, and existing customers depart. The RM system supports the decision of accepting and rejecting arriving customers, e.g., based on the expected value of the capacity that operators would need to allocate. Due to the long timescales in satcom (10-15 years of satellite lifetime, 1-3 years of contract duration), any forecast is likely to carry significant uncertainty. The gains are primarily based on the segments' heterogeneous arrival and WTP.

Table 3-8: the two application of our satcom RM framework

| | Optimal booking for a single time instance | Optimal filling of the satcom system over time |
|----------------------------|---|---|
| Customer arrive and depart | No | Yes |
| Considered timescale | multiple weeks/month depending on contract durations | many years |
| RM determines | available capacity and corresponding short-term prices | long-term prices to accept/reject arriving customers |
| RM leverages mainly | mismatch between actual and contracted usage | segments' heterogeneous arrival and WTP |

Even though we distinguished between both applications, they are complementary, and their gains are additive. The optimal overbooking is the foundation for the optimal filling. Vice versa, knowledge about the optimal filling informs the overbooking about how much of the available capacity to sell (assuming that some products cannot be as easily cancelled). The framework, its components, and the challenges are valid for both applications.

The remainder of this dissertation focuses on the *optimal overbooking for a single time instance* for two reasons. First, as suggested by the example, gains of over 50% are possible, which is not achievable by the *optimal filling*. Second, *optimal overbooking* is the basis for the *optimal filling*, and therefore solving the first enables future research of the latter.

3.8 Summary and contributions

In this Chapter, we made a case for Revenue Management as an applicable framework for satcom. Specifically, we argue that RM concepts leverage the flexibility of new digital payloads. We reviewed six industries and compared them amongst each other using our taxonomy. We contrasted the characteristics of these six industries with satcom and identified four key challenges that we need to overcome to make RM work for satcom. These challenges are capture in a satcom RM framework for which an example setup showed revenues gains of over 50%.

We make the following contributions:

- Identified Revenue Management as highly applicable to satcom.
- Contrasted six industries with satcom using a developed taxonomy.
- Identified that resource management is a key dimension for satcom RM not considered by current RM research.
- Proposed a satcom RM framework that captures the complexity of satcom.
- Identified four challenges of satcom RM: the unit of demand is not the unit of capacity, resource allocation, available capacity forecaster, and novel SLA menu.

4

Satcom Simulator

The satcom simulator component addresses the first challenge: the unit of demand is not the unit of capacity. Therefore, the simulator's task is to translate between Mbps and Watts. The simulator is part of the framework's resource management optimization layer (see Figure 4-1). The component has interfaces with the resource allocation, the available capacity forecaster, and the adaptive resource control. The first two are addressed by the following Chapters 5 and 6.

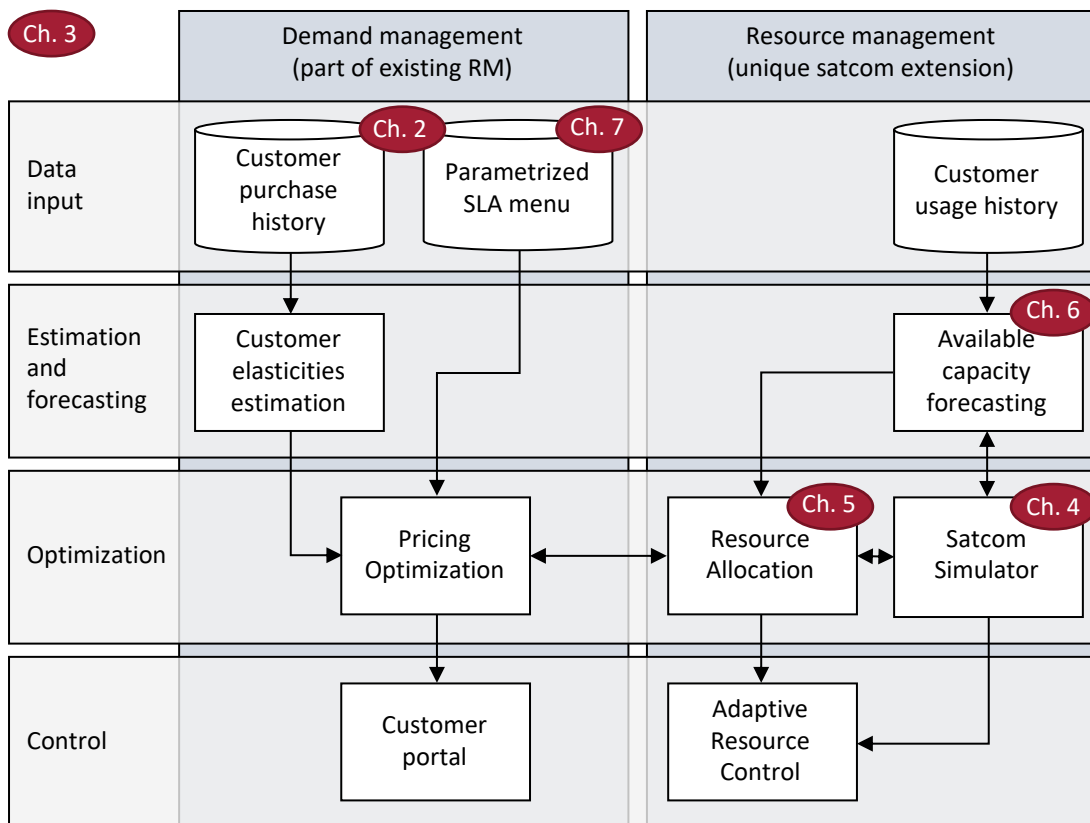


Figure 4-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4

Some parameters of the satcom simulator are set by the resource allocation component; others are constant. The model is implemented in Python and consists of over 10,000 lines of code with unit-tests of the critical modules. Around half of the model is based on work done by Portillo [97]. The author developed the other half. For completeness of this thesis, we describe the complete model and identify the parts being mainly the work of Portillo.

Figure 4-2 exhibits an abstracted overview of the simulator. It shows 12 major building blocks represented as classes/objects in Python and contains a hierarchy with cardinal descriptors. The red connections are time dependent. The blocks marked with an asterisk are mainly the work of Portillo. The bottom two objects span across the hierarchy and contain simulation relevant information such as start, end, current time and discretization. The result class keeps track of the relevant data during the simulation and has logical aggregation functionality.

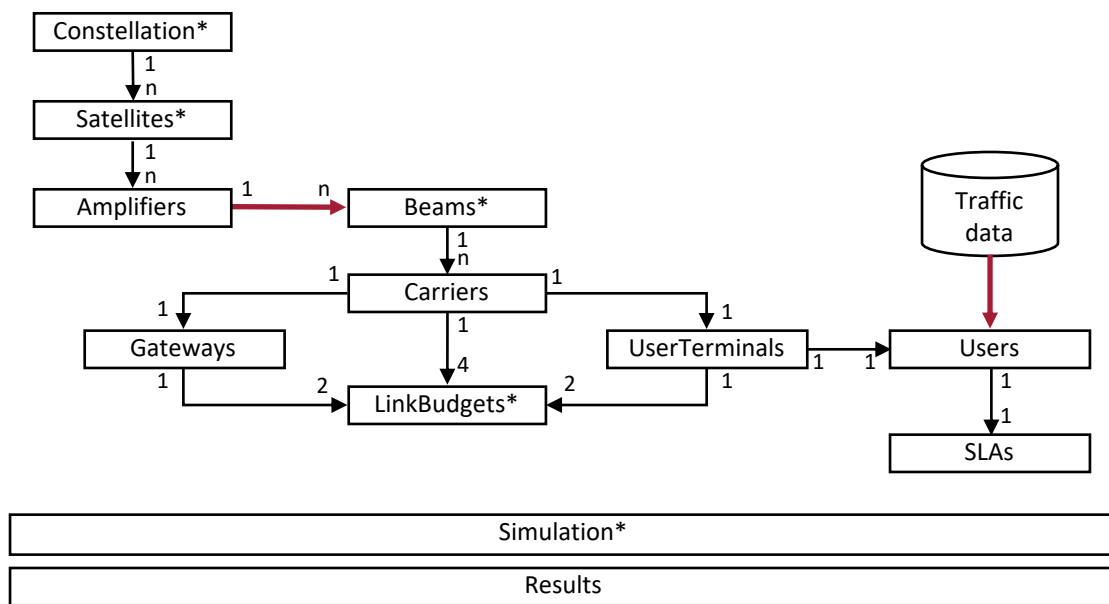


Figure 4-2: Overview of the satcom simulator. The numbers indicate the cardinality of the hierarchy, the red arrows are connections that are a function of time, and the objects with an asterisk * are majority work of Portillo [97].

The structure of the model is that a constellation consists of several satellites which themselves have amplifiers. In the case of digital payloads, multiple beams connect to an amplifier (for phased arrays, the number of beams is a decision variable as well). Beams are a container of carriers. A carrier connects a user terminal with a gateway. It has four link budgets: forward uplink from the gateway to satellite, forward downlink from satellite to user terminals, return uplink from user terminal to satellite, and return downlink from satellite to gateway. For multiplexing outside of the frequency domain, we aggregate the traffic into one user terminals object. Each user terminal has one user object that has an SLA. Throughout

the simulation, the user queries traffic data for the current simulated timestep. The simulation works by discretizing time between a start and an end time. All relevant quantities are iteratively computed for the time instance.

The central assumptions and scoping decisions we made for the simulator are:

- A user terminal does not connect to multiple carriers (except during handovers).
- A user terminal is connected to not more than one gateway
- No atmospheric losses
- No interference if beams are separated by two times the full-cone angle
- User terminal system temperature defined through G/T
- MODCOD as a combination between DVB-S2 and DVB-S2X, see Appendix G
- Altitude of all terminals is at sea level
- Link margin is 0.5 dB
- Roll-off is constant with 0.1
- Minimum elevation angle is 10 deg

It is not the focus of this dissertation to describe the simulator and its functionality in detail. Instead, we spotlight the main building blocks and functions that are most relevant for the understanding of the remainder of this dissertation. We divide this Chapter based on the 12 classes depicted in Figure 4-2 and group them as follows:

- Section 4.1: Constellation, satellite, amplifier
- Section 4.2: Beam, carrier
- Section 4.3: Link budget
- Section 4.4: Gateway, user terminal, user, SLA
- Section 4.5: Simulation, result

A validation follows the description of these blocks in Section 4.6, and Section 4.7 summarizes the Chapter.

4.1 Constellation, satellites, amplifiers

A constellation is a container object that stores information from several downstream objects. It stores the complete set of beams and the frequency allocation. For each timestep, the beams with their frequency allocation are mapped to the satellites according to the routing.

Besides several upstream functionalities from the satellites, the primary function of the constellation is to prepare the complete simulation for a specific time step, as defined by the simulation. This function triggers several routines of other objects. The major ones are:

- Propagation of the satellites' orbit to the desired epoch
- Mapping of beams to satellites either by closest distance or by the gateway allocation plan (more details in Section 5.6). The mapping triggers an update of the beam projection and gain profile given the new satellite position, which updates the slant range and the gain seen by each user terminals.
- Creation of a vectorized single link budget for each satellite based on the mapped beams (more details in the description of the satellite object)
- Mapping of the frequency allocation to the satellites' beams and the generated single link budgets, in particular, center frequency and bandwidth.

Portillo encodes the first routine while the other three are the work of the author. Since the constellation acts as the parent object, it also contains several plotting functions as used for the figures throughout the thesis. One of the vital child objects is the satellite. It describes the orbital position, contains information about amplifiers, and has a list of the relevant beams for the current time step. Power constraints and gains are additional attributes. Several vectors and rotational matrixes allow for geometrical computations. In the initialization of the simulation, beams can either be added through the outcome of the grouping of the user terminal algorithm (see Section 5.5) or by adding a user terminal with its corresponding beam separately (as used in the example of Section 3.6).

The main functionality of the satellite object is to gather all relevant data from other objects (e.g., demand from users, Tx/Rx information from user terminals, and the satellite itself, slant ranges and position). Based on that data, we create a single vectorized single link budget used for several computations. The simulator recreates these single link budgets for every time step as the number of beams connected to each satellite varies over time. However, this approach reduces the required computational resources almost proportional to the dimension of the link budget, i.e., a 500-fold reduction for a satellite with 500 beams. The reason is that without vectorization, the time scales linearly, with vectorization, it remains almost constant. The center frequencies and available bandwidth is set based on the frequency allocation passes on from the constellation.

Depending on the technology of the satellite, an amplifier might connect to more than one beam. This introduces several constraints, most notably a maximum power level per amplifier. Therefore, the power setting for a beam depends on the other beams. In that case, an additional constraint is necessary between beams and amplifiers. For these satellites, the design commonly fixes this mapping and can be imported from external files. The amplifier object internally stores that mapping and has the functionality to compute its utilization and flag if attributes exceed limits.

4.2 Beams, carriers

Since a beam can have multiple carriers, we keep this parent/child relationship also in the simulator. The four types of beams are (distinguished through attributes, not considering inter-satellite links):

- forward uplink from a gateway to satellite
- forward downlink from satellite to user terminal
- return uplink from the user terminal to the satellite
- return downlink from satellite to gateway

Each beam has a symbolic link to the constellation, the satellite, and a potential gateway. The geometrical attributes are the longitude and latitude of the beam center and the half cone angle based on a 3dB gain loss at the cone's perimeter. A functional relationship defines the gain for user terminals off-centered. Paris [98] first implemented the projection of the beam onto the earth based on a set of formulas described by Siocos [99] for equatorial satellite orbits. Portillo [97] generalized it to inclined orbits. The author added an adjustment of the footprint depending on the minimum elevation angle of the user terminals. The remaining functions and properties are for data accessing and more efficient geometrical computations

The carrier object establishes the hierarchical connection to the user terminals. Multiple user terminals can share one carrier, and multiple carriers can be within a beam. We assume that gateways do not share carriers. The link budget computes on the carrier level. We follow the approach that each carrier has a link budget with the relevant parameters, which the vectorized link budget extracts as described in Section 4.1.

4.3 Link budgets

The link budget is the crucial computation element of the simulator. For a given timestep, it takes in information from the other blocks and the resource allocation component and translates Mbps into W or

the other way around. The link budget object is optimized to reduce computational resource consumption by (1) allowing vector and matrix computation of all parameters (as used by the single link budget) and (2) re-calculating only the downstream equations that are affected by a parameter change. The following outlines the main equations of the link budget and elucidates some of the computational considerations.

Radiofrequency chain

The link budget generally considers the complete radio frequency (RF) chain from the transmitter (Tx) to the receiver (Rx). Depending on how data is available, and the desired level of detail, we choose different demarcation points in this chain. In its purest form, the carrier-to-noise ratio (C/N) is computed by the Equivalent Isotropically Radiated Power ($EIRP$), the antenna gain-to-noise temperature (G/T), all losses L between Tx and Rx, the Boltzmann constant (-228.6 dBK), and the bandwidth B . Eq. (4-1) summarizes the relationship with all parameters being in dB.

$$\frac{C}{N} = EIRP + \frac{G}{T} - L - k_B - B \quad (4-1)$$

The Tx's Effective Isotropically Radiated Power ($EIRP$), the Rx's G/T and the losses L can be further decomposed as described in the following three subsections.

Tx EIRP

The EIRP is a measure of the combination of the transmitted power and antenna gain. It is the theoretical power that would have to be emitted by an isotropic antenna. The EIRP is computed in dB by

$$EIRP = P_{Tx} + G_{Tx} - L_{Tx} \quad (4-2)$$

with L_{Tx} being the loss in the transmitter part of the RF chain, which includes the power-amplifier efficiency and its output back-off (OBO). The OBO depends on the characteristics of the amplifier, the roll-off, and the modulation and coding scheme (MODCOD). The MODCOD itself depends on the power level and the bandwidth of the link. The most common flexibility in new digital payloads is the setting of individual power levels P_{Tx} for each carrier over time.

The antenna gain G_{Tx} is particularly complicated for new phased arrays antenna. Since these are flat panels, the gain is a function of the area. Hence, the gain depends on where the beam is pointing as viewed from the satellites. We call this the scanning angle α . The more it points away from the nadir vector, the smaller is the effective area. These losses become more substantial the smaller the semi-major axis of the satellites. We approximate the phased area with a parabolic antenna with the diameter D . For

nadir pointing, we have a maximum diameter of D_{max} . The diameter as a function of the scanning angle α then becomes:

$$D(\alpha) = D_{max} \cdot \cos(\alpha) \quad (4-3)$$

We obtain the maximum gain for a parabolic antenna with the diameter D with the efficiency η , the frequency f , and the speed of light c :

$$G_{Tx,max} = 10 \cdot \log_{10} \left(\eta \left(\frac{\pi \cdot f \cdot D}{c} \right)^2 \right) \quad (4-4)$$

An additional factor that comes into play is the pointing loss if a user terminal is not in the center of the beam. The gain pattern of the beam defines this loss. It mainly depends on the distance between the user terminal and the center of the beam (measured by the angle Θ). We approximate the gain distribution across the beam's projection by the following parabolic expression [95]:

$$G_{Tx}(\Theta) = G_{Tx,max} - 12 \left(\frac{\Theta}{\Theta_{3dB}} \right)^2 \quad (4-5)$$

However, the angle for the 3dB line Θ_{3dB} is a function of the diameter of the parabolic dish with the following empirical approximation by Gérard [95]:

$$\Theta_{3dB} = 70 \cdot \frac{c}{f \cdot D} \quad (4-6)$$

Combing these equations, we get the desired gain as a function of the scanning angle α and the off-center angle Θ :

$$G_{Tx}(\alpha, \Theta) = 10 \cdot \log_{10} \left(\eta \left(\frac{\pi \cdot f \cdot D(\alpha)}{c} \right)^2 \right) - 12 \left(\frac{\Theta \cdot f \cdot D(\alpha)}{70 \cdot c} \right)^2 \quad (4-7)$$

The reference diameter D_{max} is computed by inverting Eq. (4-4) with a known $G_{Tx,max}$ for $\alpha = 0$.

Combining the above, we get Eq. (4-17) that highlights the main dependencies and shows that it is implicit in $P_{Tx}(t)$.

$$EIRP(t, \alpha, \Theta, B) = P_{Tx}(t) + G_{Tx}(\alpha, \Theta) - L_{Tx}(MODCOD(P_{Tx}(t), B)) \quad (4-8)$$

In reality, there are further dependencies, such as degradation, active elements, temperature, and influence from other beams generated.

Rx G/T

The antenna gain-to-noise temperature G/T is a quality measure of user terminals. The higher the gain and the lower the system temperature, the less power the user terminals needs. The mathematical equation is (with G/T denoting a parameter name and not a division):

$$\frac{G}{T} = G_{Rx} - T_{sys} \quad (4-9)$$

If the user terminal has a flat panel phased array, there will be an additional time-dependent gain G_{Rx} depending on the angle that the beam is pointing towards the panel. The relationship has similar characteristics to Eq. (4-7). The system temperature accounts here for the losses on the receiver part of the RF chain. We compute it by the Friis transmission equation with T_{ant} , T_{atm} , T_w being the antenna, atmospheric, and waveguide temperature. L_{RF} denotes further losses down the chain and A_t accounts for the atmospheric losses.

$$T_{sys} = T_{ant} \cdot 10^{-\frac{L_{RF}}{10}} + T_{atm} \cdot 10^{-\frac{A_t + L_{RF}}{10}} + T_w \cdot \left(1 - 10^{-\frac{L_{RF}}{10}}\right) \quad (4-10)$$

Losses L

The losses L account for the reduction in EIRP by several factors until it reaches the user terminal dish. Most notably these are the free-space-loss L_{FSPL} and the atmospheric losses L_{atm} so that we get

$$L(t) = L_{FSPL}(t) + L_{atm}(t) \quad (4-11)$$

We included dependency in time, as both losses are time-dependent in particular for NGSO constellations. The following equation [95] computes the $L_{FSPL}(t)$:

$$L_{FSPL}(t) = \left(\frac{4 \cdot \pi \cdot d(t)}{\lambda}\right)^2 \quad (4-12)$$

with $d(t)$ being the slant range between satellite and user terminal or gateway and λ the wavelength of the carrier. We use a spherical coordinate system with lat_{sat} , lon_{sat} , a_{sat} being the intersection of the nadir vector of the satellite with the earth's sphere and its semi-major axis, and lat_{ut} , lon_{ut} , a_{ut} the position of the user terminal and its distance from the center of the earth. We then get for the slant range

$$d(t) = \sqrt{a_{sat}^2 + a_{ut}^2 - 2 \cdot a_{sat} \cdot a_{ut} \cdot (\cos(lat_{sat}) \cdot \cos(lat_{ut}) \cdot \cos(lon_{sat} - lon_{ut}) + \sin(lat_{sat}) \cdot \sin(lat_{ut}))} \quad (4-13)$$

All of these components can be a function of time; however, some specific special cases simplify the equation. We want to highlight here one, as it is relevant for the simulation in this dissertation: stationary user terminal and equatorial satellite with no eccentricity. In that case $lat_{sat} = 0$ and $a_{sat}, a_{ut}, lat_{ut}, lon_{ut}$ are not a function of the time, but only lon_{sat} . All objects can update their time-dependent parameters, hence generalizing the computation to inclined satellites and mobility customers.

Due to the satellite movement and changing slant range, the atmospheric losses are affected as well. Most commonly, these are (1) rain, (2) cloud, (3) scintillation, and (4) gaseous. However, for simplicity, we assume all atmospheric losses are zero (reference [97] describes further details and an ITU compliant Python package).

Before we establish the relationships between the C/N from Eq. (4-1) and the achieved data rate R of the link, we include noise from potential interference products.

Interference

The $C/(N + I)$ accounting for interference is computed by

$$\frac{C}{N + I} = \left(\frac{1}{C/N} + \sum_i \frac{1}{C/I_i} \right)^{-1} \quad (4-14)$$

with C/I_i denoting an interference contribution, such as adjacent channel interference (CACI), intermodulation products (C3IM), adjacent satellite interference (CASI), and cross-polarization interference (CXPI). In the case of a downlink (dl), an end-to-end (ete) $C/(N + I)$ might also get interference contribution from the uplink (ul), so that we get:

$$\frac{C}{N + I} \Big|_{ete} = \left(\frac{C}{N + I} \Big|_{dl} + \frac{C}{N + I} \Big|_{ul} \right)^{-1} \quad (4-15)$$

Since the interference depends on the power level, the bandwidth, and the MODCOD, it introduces a dependency between each adjacent beam, therefore linking the individual link budgets on carrier level together. Because the interference calculation is computationally expensive, our practical approach is to make constant conservative assumptions (which also decouples the link budgets from each other).

Data rate

The Shannon limit is an optimistic approximation for the channel capacity in Mbps. It has an advantage for analytical derivations, as conducted in Section 3.6. Commonly the Shannon limit is used with the signal-

to-noise ratio that does not consider interference; however, this is not a necessary condition. We denote here the more general case with $C/(N + I)$:

$$R = B \cdot \log_2 \left(1 + \frac{C}{N + I} \right) \quad (4-16)$$

For the calculations that follow in the remainder of this dissertation, we use the more accurate approach of selecting the right MODCOD to maximize the provided data rate (known as adaptive coding and modulation (ACM)). Section 5.8 introduces a direct strategy for MODCOD selection. We rewrite Eq. (4-16) to

$$R = \frac{B}{1 + \alpha_r} \cdot \Gamma \left(\frac{E_s}{N} \right) \quad (4-17)$$

with α_r being the roll-off (assumed to be 0.1) and Γ the spectral efficiency of the MODCOD, which is a function of the ideal signal to noise ratio E_s/N . This relationship is tabulated in Appendix G for selected DVB-S2 [100] and DVB-S2X MODCODs [101]. The relationship with $C/(N + I)$ is

$$\frac{E_s}{N_0 + I_0} = \frac{C}{N + I} \cdot \frac{B}{R} \cdot \log_2 M \quad (4-18)$$

$$= \frac{E_b}{N_0 + I_0}$$

with M being the number of alternative modulation symbols. The selection of MODCOD with the maximum spectral efficiency Γ is then constraint by the following inequality.

$$\frac{E_s}{N} \Big|_{MODCOD} + \gamma \leq \frac{E_s}{N_0 + I_0} \quad (4-19)$$

with γ being the desired margin of the link (assumed to be 0.5 dB).

The non-linearity of the relationship between demand in Mbps R and power P_{Tx} in W becomes clear. The ACM even introduces an internal optimization loop and the, in P_{Tx} , implicit Eq. (4-8) makes solving the equations not trivial. Complex time dependencies for NGSO make analytical derivation impractical. Hence, many optimization techniques have been tried, as reviewed in Section 5.1. These techniques are trying to set the power P_{Tx} directly, compute the link budget in the order presented, and provide feedback to the algorithm through R . We propose a method that works the other way around, starting from a requested R , selecting the right MODCOD, and then computing P_{Tx} directly (further described in Section 5.8).

4.4 Gateways, user terminals, users, SLAs

The gateways and user terminals are both ground stations with geographical (latitude, longitude, and altitude) and antenna attributes. Furthermore, they have an associated country which is either provided

by the input dataset or computed based on the longitude and latitude. For both types, possible constraints on the traffic are considered, for example, in- and out-of-country. This constraint applies for countries in which the regulatory body enforces that the traffic has to “land” in the same country. We assume that the constraints apply to 100% of the traffic, and therefore, we can store the constraints at the gateway and user terminal level. Both have symbolic links to beams and carriers.

Since one carrier can have multiple user terminals, we implement a function that aggregates these user terminals into a “super user terminal” by aggregating the traffic within a carrier (amongst other things). Several functions allow filtering a subset of the complete imported user terminals ordered by identifications or evenly distributed around the longitude.

The user terminal has a further symbolic link to the user. The primary attributes of the users are related to demand and traffic. We distinguish between forward (fwd) and return (rtn) traffic and split these two types further into real-time and not-real-time (as necessary for the Two Classes of Service SLA in Chapter 7). A function-call can shift the not-real-time part in time to model the delay of the non-real time traffic. As we focus our simulation on a 24 hours window, we capture the stochastic nature of the demand by allowing a time-series of random variables for the different traffic types. Each user object has the functionality to return the demand for the current time of the simulation.

Furthermore, the user terminals class allows us to input a cosine demand function with the mean, variation, sigma, and time offset, as used in Section 3.6. The user has a function to generate an actual realization of the demand by sampling the random variables for the duration of the simulation. Import and sampling can make use of computers with multiple cores through parallelization. The simulator pickles and caches several imports to reduce the time needed for the creation of the simulator (a few seconds for 18,000 user terminals).

We create a separate class for the SLA, which links directly to the user. It captures all of the different types of SLA, their pricing functions, and their commitments. Chapter 7 provides further details. The object furthermore keeps track, e.g., of customer credits, accumulated revenues, used data, and usage.

4.5 Simulation, results

In comparison to the other classes, the simulation and result classes are more general and orchestrate the other objects and extract relevant data, respectively. The core tasks of the simulation object are to define the simulation start, end, and time step, and to keep track of the current simulation time. Furthermore, it coordinates the export into a Cesium [102] compatible czml file, creates the result folder

structure, and supports transformations between coordinate systems for objects with geometrical attributes. Because we set up the simulation at the beginning of scripts, it also supports flags like the treatment of warnings and the suppressing of showing figures.

The result-block consists of two main classes, the result object itself and the result-wrappers. Since the simulator supports many different analyses, a versatile storing of results is critical. The result object inherits the hierarchical structure of the simulator as depicted in Figure 4-2 (constellation, satellites, amplifier, beam, carrier, user terminal). A multilevel dataframe is constructed based on that hierarchy, and a storing function fills this dataframe. The function call triggers all result attributes, as defined in the result-wrappers. The result attributes can access every attribute of every class of the simulator. To enable hierarchical and temporal aggregation, we allow the definition of these functions on the result attribute level. The most common functions are summation, averaging, and integration. As an example, if the average power of a satellite over the first 6 hours of the simulation is of interest, the result class hierarchically sums the power (in a linear form, not dB) from user terminals up to the satellite of interested. Then the temporal aggregation of averaging is applied to the first 6 hours, and we return the value. Finally, the result class can plot multiple variations of figures, export data to csv, and save/load its dataframe. Despite that the dataframe can reach substantial size (> 1 GB, depending on the number of user terminals, the constellation, the duration, and discretization of the simulation time), we did not notice any performance falloff.

4.6 Validation

For validation of the simulator, we use reference data provided by SES. Since the link budget is the key computational element responsible for translating power into data rate and vice-versa, we focus on this part here. We randomly choose ten links with different equatorial satellites and user terminals' position to validate the orbit propagator and the geometrical computations (see Table 4-1).

Table 4-1: Overview of the 10 inputs for validation of the link budget relationships

| <i>sat lon</i> [deg] | <i>user lat</i> [deg] | <i>user lon</i> [deg] | <i>B</i> [MHz] | <i>f</i> [MHz] | <i>EIRP</i> [dBW] | <i>G/T</i> [dB/K] | <i>G</i> [dB] | <i>CACI</i> [dB] | <i>C3IM</i> [dB] | <i>CXPI</i> [dB] |
|-------------------------|--------------------------|--------------------------|-------------------|-------------------|----------------------|----------------------|------------------|---------------------|---------------------|---------------------|
| 0.0 | 28.9 | 24.3 | 14 | 17711.4 | 25.9 | 21 | 44.7 | 16 | 40 | 30 |
| 0.0 | 28.9 | 24.3 | 56 | 17733.8 | 31.9 | 21 | 44.7 | 16 | 40 | 30 |
| 0.0 | 19.1 | 342.8 | 25 | 17718.8 | 30.7 | 15 | 38.7 | 16 | 40 | 30 |
| 102.9 | 24.5 | 77.3 | 40 | 17726.3 | 30.4 | 21 | 44.7 | 16 | 40 | 30 |
| 257.1 | -3.8 | 280.7 | 20 | 17711.4 | 28.5 | 15 | 38.7 | 16 | 40 | 30 |
| 257.1 | 42.0 | 239.5 | 40 | 17726.3 | 31.5 | 21 | 44.7 | 16 | 40 | 30 |
| 308.6 | 37.8 | 288.9 | 200 | 17823.4 | 38.5 | 21 | 44.8 | 16 | 40 | 30 |
| 308.6 | -15.4 | 307.7 | 20 | 17711.4 | 30.3 | 15 | 38.7 | 16 | 40 | 30 |
| 308.6 | -9.0 | 315.6 | 20 | 17711.4 | 29.7 | 15 | 38.7 | 16 | 40 | 30 |
| 308.6 | 3.7 | 287.8 | 20 | 17711.4 | 28.5 | 15 | 38.7 | 16 | 40 | 30 |

Note that this approach also validates the logical and hierarchical correct linking of the simulator. Other input parameters are the bandwidth B and the frequency f , which are actively computed by the frequency assignment step described in Section 5.7 (but are there provided by SES). The power is expressed with an $EIRP$, meaning that we test the power to data rate conversion of the link budget here. G/T and gain G gives the user terminal characteristic. We consider the adjacent channel interference (CACI), intermodulation products (C3IM), and cross-polarization interference (CXPI) with the numbers provided in the table (higher numbers indicate lower interference levels).

Based on these inputs we compute several results with our simulator and compare them to SES's data (summarized in Table 4-2). We report the numbers for the resulting data rate R , the spectral efficiency Γ , the $C/(N + I)$ measured for the end-to-end link and the downlink, the C/N without interference, the losses L (here solely free space L_{FSPL}), and the slant range d . The percentage value is the relative difference between our value and the reference data point.

Table 4-2: comparison table between the results of our simulator and the reference provided by SES. The percentages are the relative difference which are averaged over the 10 links in the last row.

| | R [Mbps] | | Γ [-] | | $\frac{C}{N+I} _{ete}$ [dB] | | $\frac{C}{N+I} _{at}$ [dB] | | $\frac{C}{N}$ [dB] | | L [dB] | d [km] | | |
|----------------|---------------|-------------|-----------------|-------------|--------------------------------|--------------|-------------------------------|--------------|-----------------------|--------------|-------------|-------------|-------|-------------|
| Reference | 21.0 | | 1.5 | | 5.8 | | 6.0 | | 6.5 | | 197.5 | 10114 | | |
| Our simulator | 21.8 | 3.9% | 1.6 | 3.9% | 5.8 | -0.1% | 6.0 | -0.1% | 6.5 | 0.0% | 197.5 | 0.0% | 10114 | 0.0% |
| Reference | 84.0 | | 1.5 | | 5.8 | | 6.0 | | 6.5 | | 197.5 | 10114 | | |
| Our simulator | 87.2 | 3.9% | 1.6 | 3.9% | 5.8 | -0.1% | 6.0 | -0.1% | 6.5 | 0.0% | 197.5 | 0.0% | 10114 | 0.0% |
| Reference | 25.0 | | 1.0 | | 3.3 | | 3.5 | | 3.7 | | 196.6 | 9111 | | |
| Our simulator | 27.0 | 8.0% | 1.1 | 8.0% | 3.3 | -0.1% | 3.4 | -0.1% | 3.7 | -0.1% | 196.6 | 0.0% | 9111 | 0.0% |
| Reference | 64.0 | | 1.6 | | 6.0 | | 6.2 | | 6.7 | | 197.3 | 9899 | | |
| Our simulator | 62.3 | -2.6% | 1.6 | -2.6% | 6.0 | -0.1% | 6.2 | -0.1% | 6.7 | 0.0% | 197.3 | 0.0% | 9899 | 0.0% |
| Reference | 20.0 | | 1.0 | | 2.3 | | 2.4 | | 2.6 | | 196.5 | 8989 | | |
| Our simulator | 19.8 | -1.0% | 1.0 | -1.0% | 2.3 | -0.1% | 2.4 | -0.1% | 2.6 | -0.1% | 196.5 | 0.0% | 8989 | 0.0% |
| Reference | 64.0 | | 1.6 | | 6.2 | | 6.4 | | 6.9 | | 198.2 | 10896 | | |
| Our simulator | 64.7 | 1.1% | 1.6 | 1.1% | 6.1 | -0.1% | 6.4 | -0.1% | 6.9 | 0.0% | 198.2 | 0.0% | 10896 | 0.0% |
| Reference | 320.0 | | 1.6 | | 6.3 | | 6.6 | | 7.1 | | 198.0 | 10592 | | |
| Our simulator | 323.6 | 1.1% | 1.6 | 1.1% | 6.3 | -0.1% | 6.6 | -0.1% | 7.1 | 0.0% | 198.0 | 0.0% | 10592 | 0.0% |
| Reference | 26.6 | | 1.3 | | 4.5 | | 4.6 | | 5.0 | | 196.0 | 8471 | | |
| Our simulator | 24.0 | -9.6% | 1.2 | -9.6% | 4.5 | -0.1% | 4.6 | -0.1% | 5.0 | -0.1% | 196.0 | 0.0% | 8471 | 0.0% |
| Reference | 24.0 | | 1.2 | | 4.1 | | 4.2 | | 4.5 | | 195.8 | 8293 | | |
| Our simulator | 24.0 | 0.2% | 1.2 | 0.2% | 4.1 | -0.1% | 4.2 | -0.1% | 4.5 | -0.1% | 195.8 | 0.0% | 8293 | 0.0% |
| Reference | 20.0 | | 1.0 | | 2.5 | | 2.6 | | 2.8 | | 196.3 | 8797 | | |
| Our simulator | 19.8 | -1.0% | 1.0 | -1.0% | 2.5 | -0.1% | 2.6 | -0.1% | 2.8 | -0.1% | 196.3 | 0.0% | 8797 | 0.0% |
| Average | | 0.4% | | 0.4% | | -0.1% | | -0.1% | | -0.1% | | 0.0% | | 0.0% |

We see that the geometrical calculations for d and L have no noticeable relative error. There are minor differences for the signal-to-noise computations in the range of plus/minus 0.1%. A larger error is introduced by the ACM MODCOD selection process (3rd and 8th link). If the E_s/N is on the edge between two MODCODs, then minor errors in the C/N calculations lead to a different selection and hence the difference in the spectral efficiency Γ and data rate R . However, the important point is here the average of the relative differences printed in the last row: it averages out over the ten links.

In sum, the average relative error is below 0.4% (in absolute terms). Given the uncertainties in the other parameters and datasets (e.g. demand, elasticities, ...), this difference deemed sufficient to use the simulator for this dissertation. We also want to point out that the reference data is computed by a model as well and it is undetermined which result-set represent reality more accurately.

4.7 Summary

In this Chapter, we described the high-level functionality and implementation of the satcom simulator that translates between data rate and power. Figure 4-2 showed an overview of the 12 building blocks labeling what the work of Portillo [97] was and work from the author. The subsequent Sections describe these blocks and some computational considerations. Validation against the dataset provided by SES shows that our simulator has an average relative error below 0.4%. We aimed to provide a descriptive summary of the simulator that helps to understand the remainder of this dissertation, in particular the following Chapter 5.

There are many practical contributions regarding the implementation, such as the presented hierarchical decomposition, the time-dependent mapping between satellite and beams, the vectorized single link budget, the handling of demand data, and the storing and aggregation of result data from the simulator. While they can be generalized and reused by future research, these contributions are mainly simulator specific and should be seen as necessary to produce the results throughout the dissertation.

5

Resource Allocation

The centerpiece of the resource management aspects of the RM framework is the resource allocation component. It closely interacts with the previously described satcom simulator (see Figure 5-1). The component is the central optimization engine of the resource management part of the framework. Information is exchanged with the available capacity forecaster (introduced in the following Chapter 6). The resource allocation passes on its solution to the adaptive resource control, which controls the communication system reactively to shorter term changes.

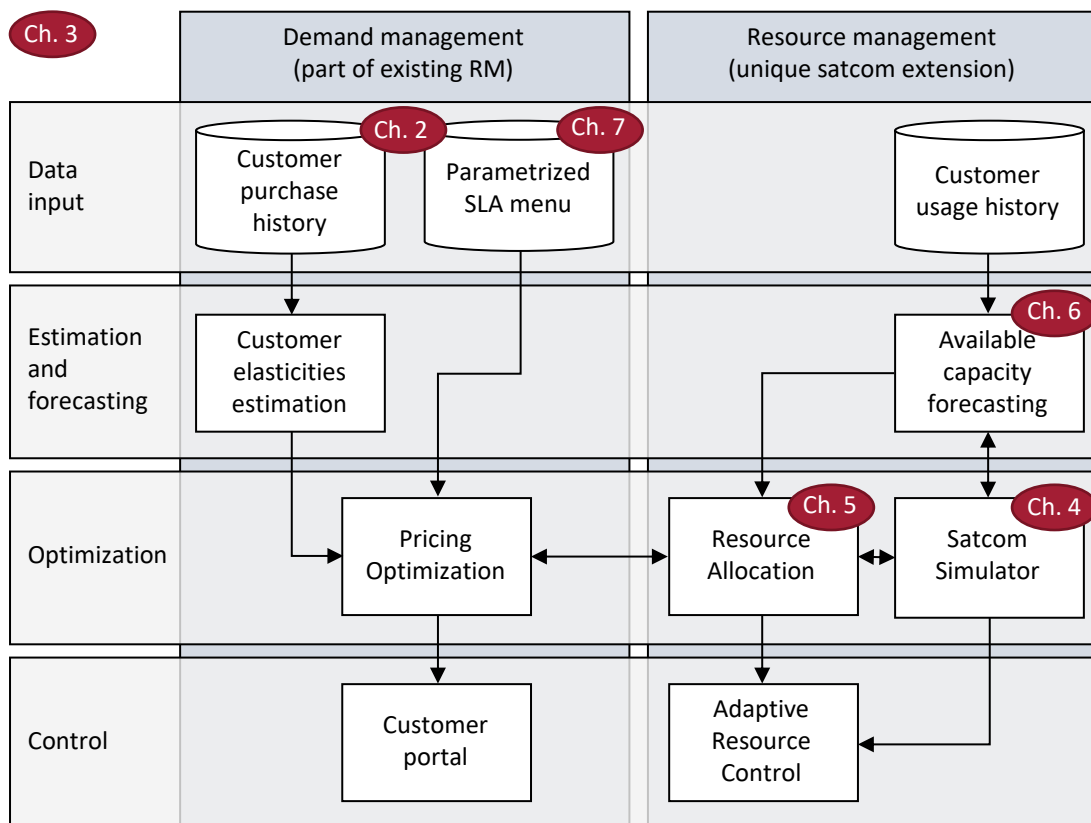


Figure 5-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4

The purpose of the resource allocation is to find the optimal resource allocation that satisfies the demanded rate. The resources are a pool of power, a pool of bandwidth, and the number of beams. There are several constraints that couple the allocation of resources for user i with that of user j . The resource allocation process consists of several steps that we define as the grouping of user terminals, routing, frequency allocation, and power computation (formalized and justified in Section 5.4).

In this Chapter, we first review the relevant literature regarding each aspect of resource allocation, the used optimization techniques, and the considered metrics (Section 5.1). Section 5.2 summarizes this review. We identify the gaps and detail the specific objectives in Section 5.3. The following Section 5.4 formalizes the resource allocation process, and the subsequent four Sections 5.5 - 5.8 illustrate our proposed algorithmic solutions to each of the four steps. We describe in Section 5.9 the results of the application to a MEO constellation. We end with summarizing the contributions of this dissertation in the area of resource allocation in Section 5.10.

5.1 Literature review

We divide the literature review on resource allocation into four Sections:

- Cognitive radios (Section 5.1.1)
- Scheduling for data relay satellites (Section 5.1.2)
- Power allocation, frequency assignment, user terminal grouping, and beam shape (Sections 5.1.3 - 5.1.7)
- Metrics (Section 5.1.8)

We organize them first to give the higher-level context of cognitive radios, for which resource allocation is a subproblem. Then, we introduce the well-studied problem of scheduling for data relay satellites and discussed its applicability. In the Sections 5.1.3 - 5.1.7, we discuss the work done in the four areas of power allocation, frequency assignment, user terminal grouping, and beam shape. In Section 5.1.8, we group the sources from Sections 5.1.1 - 5.1.7 based on their considered metrics.

5.1.1 Cognitive radios

Cognitive radios (CR) are software-defined radios that have additional intelligence [103] (mostly terrestrial). CRs emerged over the last decades as a response to higher data transmission needs. The goal is to improve the efficiency and utilization of the available frequency spectrum by providing more dynamic spectrum access. According to Abbas et al. [104], the tasks of CRs can be divided into spectrum

environment sensing, analyzing, and decision-making aspects. The latter includes decisions about the allocation of resources such as power and spectrum.

Hossain et al. [103] give an extensive survey of CR research activities. They discuss challenges, various approaches, and simulation tools. Of particular interest is the Section on resource allocation, which points to Zhang et al. [105], which provide a survey on dynamic resource allocation in CR networks. The focus lies on interference-power/interference-temperature constraint approaches. Zhang et al. conclude that convex optimization is an efficient way of solving this kind of resource allocation problem.

Researchers have done significant work to use Artificial Intelligence (AI) and learning techniques in CR networks. Abbas et al. [104] present a survey on the application of AI to CR networks. In particular, they provide an overview including strengths, limitations, and challenges of fuzzy logic, genetic algorithms, neural networks, game theory, reinforcement learning, support vector machines, case-based reasoning, decision trees, Bayesian approaches, Markov models, multi-agent systems, and artificial bee colonies. They conclude that all techniques can be applied to the decision-making aspects of CR, while only some of them apply to the spectrum sensing aspect.

Sharma et al. [106] discussed the specifics of CR for satellite communication systems. In particular spectrum sensing, interference modeling, beamforming, and beam-hopping techniques. They argue that the resource management problem for satellites differs from a terrestrial isolated wireless system. In their 2015 paper [107], Sharma et al. investigate joint carrier allocation and beamforming for a combination of broadcasting and broadband satellite services. They conclude that throughput can be increased, and link availability improved.

Ferreira et al. [108-110] have developed a series of reinforcement learning techniques with deep neural networks for cognitive satellite communications. In their latest paper [110], they consider a single beam satellite with the adaptable parameters being coding and modulation scheme, bandwidth, symbol rate, and energy to noise ratio. Using Ferreira's approach, Hackett et al. [111] describe the results of testing this cognitive engine on-board the International Space Station.

Kandeepan et al. [112], Lagunas et al. [113], and Vassaki et al. [114] study an integrated cognitive satellite and terrestrial networks. All three show that there are benefits in considering integrated resource allocation. Lagunas et al. [113] present an approach to increase the throughput of satellite systems. Vassaki et al. [114] focus on the development of a power allocation algorithm for the terrestrial network while considering QoS aspects of the satellite network.

5.1.2 Scheduling for data relay satellites

Work on *resource scheduling* traditionally focused on data relay satellites with little to no flexibility. Very few beams characterize these satellites (for example, the U.S. Tracking and Data Relay Satellite (TDRS) has only two high data rate beams [115]). Hence, careful scheduling of users is of significant importance as multiple users share one beam. For the next generation of satellites considered in this dissertation, resource scheduling remains as a sub-problem for beams with a large number of users.

There has been work done in the scheduling of resources for the TDRS [116]. The scheduling focuses on allocating a timeslot for each user, in which the user has access to the satellite resources. The objective is to maximize satellite utilization by meeting all users' requirements. The outcome is a resource distribution diagram used for mission scheduling. However, given the inflexible TDRS technology, resources remain unused and reallocation is challenging.

A two-phase process is proposed by Deng et al. [117] to schedule satcom resources. In the first phase, a Genetic Algorithm (GA) finds an initial schedule. In the second phase, a preemptive dynamic scheduling algorithm (see also [118]) tries to improve the scheduling under dynamic disturbance factors. The paper shows significant performance improvements.

Wang et al. [119, 120] integrate the mission scheduling problem with the users' behavior for more general data relay satellites. They assume users are selfish and submit more requests than needed (to receive more attention and resources). To solve this, Wang et al. construct a repeated game to maximize users' payoff and reduce resource conflicts. Furthermore, a mechanism is proposed that incentivizes users to submit their actual resource requirements (in pricing theory, this is called incentive-compatible [121]).

5.1.3 Power allocation

Garau et al. [93] conducted a comparison between a broad range of AI algorithms for power allocation: GA, SA, PSO, DRL, and hybrid approaches. The authors use different demand scenarios to quantify the robustness of each approach during changing traffic. The paper concludes that the strength of the DRL is the computational speed where a solution can be obtained in seconds. However, the DRL uses between 5-100% more power with more unmet demand. The hybrid PSO-GA achieves the lowest unmet demand performance. The GA excels with its robustness in scenarios when demand is changing often.

Wang et al. [122] used convex optimization to find the trade-off between total system capacity and fairness of power allocation amongst users. The authors formulate the problem based on the link budget and prove that this formulation can be optimally solved with an iterative algorithm based on duality

theory. Wang et al. show that his approach improves the fairness of power allocation amongst users (compared to a uniform or proportional power allocation). They claim that the proposed algorithm runs in linear time with respect to the number of beams and users.

Hong et al. [123] used a Lagrange multipliers to find the optimal power allocation for ten spot beams while meeting all SLAs. The method uses a binary search to find the optimal Lagrange multiplier. Hong et al. model the SLAs by defining a minimum traffic demand. Moreover, the authors use a total power constraint and formulate the link budget equation by using the Shannon bound.

Neely et al. [124] formulated a power-control problem and developed throughput maximizing power allocation and routing algorithms. The authors consider a system with a defined number of beams and queues (e.g., data transfer requests). The coding and modulation schemes are modeled by piecewise linear interpolation to generate a continuous convex curve (in contrast to using the convex Shannon bound directly). Therefore, the power allocation problem becomes convex and can be solved by a bisection-type algorithm very efficiently. The defined power allocation policy stabilizes the queues whenever the arrival rate vector lies within the capacity region. The algorithm allows for independent power and routing decisions by each user based on local channel and queue information.

Aravanis et al. published several papers [125-127] on power allocation using a hybrid technique combining the Genetic Algorithm (GA) and Simulated Annealing (SA). In [127], Aravanis et al. used the unmet system capacity together with the total power consumption as multiple optimization objectives. They propose a two-stage optimization, with the first stage consisting of the hybrid GA-SA and the second stage being a GA only. He shows that by using multiple objectives, the on-board power consumption can be traded-off with the system capacity. Furthermore, Aravanis showed that the power allocation problem is NP-hard [127].

He et al. [128] propose a new traffic-aware dynamic power resource allocation algorithm, which is mainly based on a GA variant known as the non-dominated sorting genetic algorithm II (NSGA-II). They use two conflicting metrics: the call completion ratio (a proxy for the QoS) and the total system capacity. The authors consider two types of traffic with different priorities: voice with higher priority on call completion ratio and video with higher priority for throughput. Furthermore, He et al. prove the NP-hardness of his problem description. The authors conclude that their proposed NSGA-II derivation reduces the computational complexity by 95% compared to the pure NSGA-II.

Durand et al. [129] studied power allocation based on Particle Swarm Optimization (PSO). They specifically showed that power reallocation to beams with rainfade improves the overall energy efficiency (compared with uniform power allocation). Durand et al. conclude that PSO has lower computational complexity than exhaustive search, direct matrix inversion, and GA. Their implementation's runtime scales with the squared of the number of beams: $\mathcal{O}(n_{beams}^2)$.

Furthermore, Destounis et al. [130] and Srivastava et al. [131] studied the impact of rain attenuation on power allocation. Destounis et al. focus on dynamic reconfiguration during rainfade events. They claim that their algorithm is simple and can be implemented on an on-board microprocessor. In a later paper, Srivastava et al. show that by grouping users with similar power requirements, the system can serve more users. Furthermore, the authors use a stochastic model for rain attenuation prediction.

5.1.4 Frequency assignment

In general, the spectrum can be reused in the frequency domain (*beam coloring*) or in the time domain (*beam hopping*). We define the *frequency assignment* as consisting of the center frequency assignment, allocated frequency spectrum, and polarization for each beam (and in the case of beam hopping a switch matrix, see Section 5.4 for details). If the satcom system uses frequency division multiple access (FDMA) for multiple users within a beam, carriers divide the frequency spectrum further (the carrier is a decomposition of a beam as defined in Chapter 4). Some literature refers to this dividing process as *carrier allocation* or *channel allocation*. That applies to both the beam coloring and beam-hopping

Reuse in the frequency domain - beam coloring

Park et al. [132] studied the spectrum allocation for a 20-beam satellite system with a shared pool of 1 GHz with a binary search heuristic. They compare their solution with the *water-filling* method in which the algorithm allocates the spectrum uniformly. The water-filling method results in a maximum use of system capacity [133, 134]. However, unmet demand is not minimal for a non-uniform demand distribution. Park's proposed scheme reduces the unmet demand while sacrificing total system capacity, resulting in better proportional fairness for users. In a similar paper, Park et al. [135] propose an active beam selection algorithm that selects a subset K out of N beams (and therefore drops the users which are not in K). They show that this can result in a total system capacity above the water-filling approach. However, the approach sacrifices the fairness for the inactive beams and leaves demand unmet.

While Park et al. use a binary search heuristic, Wang et al. [136] formulate the same problem as convex optimization and shows that their solution reaches the lowest unmet demand compared to the water-filling method and proportional spectrum allocation.

Hu and Liu [137, 138] developed a Deep Reinforcement Learning (DRL) algorithm to allocate channels for a 37 beam satellite system to maximize the number of customers served while meeting their QoS requirements. The *state* is reformulated as an image tensor to represent both the spatial and temporal features. Hu and Liu compare their DRL result with an interference measure-based allocation [139] and a proposed GA-SA hybrid [140]. They conclude their DRL approach can achieve similar performance, but the computational complexity can be reduced compared to the GA-SA hybrid as the DRL only needs to evaluate the neural network.

Solving the NP-complete subproblem of center *frequency assignment* with a gradual neural network (NN) was studied by Funabiki et al. [141]. Salcedo-Sanz et al. proposed a hybrid combination of a NN with SA [142] or GA [143]. The results show that this combination achieves better results, especially for larger-scale problems. The drawback is an increase in computational costs.

In a recent 2019 paper, Li et al. [144, 145] address spectrum management for satellite communication systems with a hunger marketing approach. The paper focuses on uplink resource management. He proposes a two-stage Stackelberg game model that achieves an optimal balance between spectrum utilization and inter-beam interference. The approach finds the profit-maximizing interference pricing for a 16-beam satellite system. Furthermore, the authors discuss the potential risk of spectrum shortage due to the hunger marketing approach. In another paper [146], Li et al. use a market-driven pricing mechanism to support more efficient spectrum trading between a satellite operator and multiple terrestrial network operators.

Sun et al. [147, 148] describe another pricing-based approach. They consider a spectrum allocation based on users' demand and shows that for a concave, strictly increasing, and continuously differentiable user utility function, an equilibrium price can be defined. For this point, the users' throughput is Pareto optimal, and the total throughput of the system can be improved.

Reuse in the time domain - beam hopping

Anzalchi et al. [149] investigated the advantages of beam hopping compared to conventional systems. The paper described a system optimization loop, including power allocation. They conclude the approach

can reduce the unmet demand and power consumption by around 50% at the same time. Kyrgiazos et al. [150] made a similar comparison concluded that beam hopping increases the overall throughput.

Alberti et al. [151] compare a conventional system with two flexible systems, namely one that has power and spectrum allocation flexibility, and one that has beam hopping capabilities. An iterative algorithm is deployed that ensures convergence for any valid input. The results indicate that both flexible systems outperform conventional systems, but the authors did not find any significant difference between the two flexible approaches. The authors state that the final selection should depend on cost and complexity, which is beyond the scope of their paper. Similar findings are published by Lei et al. [152, 153], concluding that beam hopping has a slightly better performance for non-real time services.

Shi et al. [154] proposes a joint optimization of power allocation and beam hopping. The considered system includes a smart gateway with Q/V feeder links. The results show that combined optimization yields better results than single variable optimization.

Angeletti et al. [155] uses a GA to optimize the time plan of a 100-beam system with a four-color reuse scheme. Thirty-five beams illuminate at the same time. The power allocation is uniform and static. The paper shows that this scenario can achieve a capacity lift of 30%.

5.1.5 Joint power allocation and frequency assignment

Lei et al. [156] propose an iterative two-step algorithm to jointly optimize power and carrier allocation, including minimization of co-channel interference. The first step is to allocate an optimal carrier within each beam by using a Rayleigh quotient problem. In the second step, power is directly allocated. The method repeats these steps until the algorithm converges. The results show a convergence after a few tenth iterations. The authors conclude that the dynamic allocation of resources improves the total power, and the traffic matching ratio significantly compared to uniform allocation.

Wang et al. [157] investigated the power and spectrum allocation for smart transponders, which demodulate, decode, re-encode, and modulate the signal. The paper specifically focuses on the trade-off between allocating power and spectrum for up- and downlink. Wang et al. conclude that the combined dynamic allocation results in performance gains compared to linear proportional allocation, dedicated spectrum, or power allocation schemes.

Cocco et al. [158, 159] proposed an algorithm based on SA that simultaneously allocates power and spectrum. Moreover, the paper defines a new metric, the Satisfaction-Gap Measure, that we further discuss in Section 5.1.8. The authors compare three different dynamic systems: total power and spectrum

allocation flexibility, only spectrum allocation, and only power allocation. The results illustrate that full flexibility can follow the dynamic traffic pattern the best, followed by spectrum only. The power allocation only slightly improves the results compared to fixed allocation. We believe that this is mostly due to Cocco's system model, where multiple beams connect to an amplifier, and the power setting is on the amplifier level. As we described in more detail in Chapter 4, power settings in the new generation of satellites are usually on carrier level.

Pachler et al. [160] studied a PSO and hybrid PSO-GA implementation for the joint the joint power and spectrum allocation. The multi-objective formulation trades-off the service rate with the power consumption. The results indicate a fast convergence of the PSO-only but gets stuck in local optima. The hybrid PSO-GA dominates the PSO-only in terms of power consumption and service rate. Furthermore, the PSO-GA outperform the GA-only for short run-times with up 85% power reduction and 5% more service rate.

París et al. [98, 161] used a GA to allocate power and spectrum jointly. The model takes atmospheric attenuation and interference into account. Two results are presented: for a 37-beam setting with four colors, and Viasat-1. The authors conclude that allocating the spectrum in addition to power results in up to 40% lower unmet demand. Furthermore, for demand scenarios with high variability, the additional flexibility of spectrum allocation is particularly valuable.

Ji et al. [162] described an algorithm that allocates power and channels for heterogeneous services, a real-time and non-real-time one. The real-time services consider delay constraints. The authors based the algorithm on dual decomposition. First, the power and channel allocation are derived given relaxed delay constraints. Based on this initial solution, an optimal allocation is found that satisfies all delay constraints. The result is a trade-off between throughput and delay performance.

5.1.6 User terminal grouping

The problem of *user terminal grouping* describes the mapping between gateways, satellites, and users. Specifically, the mapping between satellites and users is defined by the *beam placement* or *beam pattern*. In this Section, we consider beams of a fixed shape and discuss the specifics of dynamically varying the *beam shape* in the following Section.

Choi et al. [134, 163-165] made several contributions. In their 2002 paper [163] and 2005 article [134], the authors argued that optimizing the design of agile antenna multibeam patterns could improve the efficiency of transmission and power management. They show that increasing the number of beams can

accomplish significant gains in spectral efficiency. Choi et al. conclude that the algorithm should not only allocate resources to maximize system capacity but based on traffic demand and fairness. In 2006 [164], they specifically worked on scheduling for phased arrays. The real-time algorithm performs user selection, antenna gain pattern, power allocation, and admission control. It furthermore suppresses interference by choosing between space-division multiplexing and time-division multiplexing. The cost function depends on the throughput, delay, and channel condition. The results show that the algorithm can achieve up to 94% of the analytical result for random traffic. In a follow-up article in 2009 [165], the authors compare phased arrays with multiple beam antennas. They conclude that the more flexible power allocation for phased arrays provides better performance, especially in high-density areas.

Jahn [166] proposes graph theory algorithms to solve the resource allocation problem of frequency planning and beam placement. He compared several different degrees of flexibility in resource allocation and showed that graph theory could solve them. The particular problem of assigning a minimum number of fixed size beams to a defined number of users is also known as *clique cover*. That is a very well-studied problem for which probably optimal algorithms exist [167-172].

Camino et al. [173, 174] published work on the optimal integrated design of satellite payloads using a mixed-integer linear program [173] and a greedy approach [174]. The problem under consideration is to find a set of different sized beams to cover an area of non-uniform traffic for optimal load balancing (4 coloring scheme). While Camino et al. phrased this as a design problem for fixed beam satellites, the problem is transferable to phased-array, where beam placement is a more dynamic parameter. The authors conclude that the mixed-integer linear program outperforms the greedy for the smaller sized problem (few hundred users), while the greedy approach dominates for larger sized problems due to the computational complexity of the mixed-integer linear program.

5.1.7 Beam shape

Qian et al. [175] investigate dynamic beam coverage adjustment to balance the load between beams. The considered scenario is a heavy-loaded center beam with six surrounding beams and a three-colored reuse scheme. The paper considers two sets of beam coverages, a wider and a narrower center beam. The number of beams is different for each set, requiring a reallocation of spectrum to deconflict beams (due to the limited three colored reuse scheme). The results show that system throughput can be increased by dynamically switching between those two sets. Qian et al. mention that a more flexible change of beam size would be desirable (not just switching between two sizes).

Wenqian et al. [176] investigate the effect of dynamically adjusting the beams' shape in addition to power and spectrum allocation. The baseline scenario is a 7-beam system with a heavy loaded center beam (similar to Qian et al. [175]). Narrowing the center beam and widening the other beams could balance the load more equally. Moreover, due to the higher gain in the center of beams, this change increases the total system capacity (for the same demand scenario).

Schubert and Holger [177, 178] derive an analytical framework for the joint power and beamforming optimization for general downlink communication links. They divide the problem into the maximization of the jointly achievable signal-to-interference-plus-noise ratio (SINR), and the minimization of the total power while satisfying users' constraints. The iterative algorithm is proven to reach the global optimum.

Kyrgiazos et al. [179] proposed a joint optimization of spectrum allocation and adjusting the beam size. In particular, they use the time domain to distribute spectrum (beam hopping). The 200-beam transparent system has two different beam shapes. The results show that by using irregular beams, the total throughput can be improved by 11% compared to equal beam sizes.

Sharma et al. [107] studied the spectral coexistence of Fixed Satellite Service downlink and Broadcasting Satellite Services *feeder links* in Ka-band. He combines two allocations: carrier allocation and beamforming, with a focus on minimizing the output energy. The results show that a combined allocation improves the system throughput compared to carrier or beamforming only allocations.

5.1.8 Metrics

We classify the metrics used by the cited sources into the following categories: total system capacity, total power consumption, unmet demand, number of users, amount of interference, and fairness. We notice that none of the sources consider economic measures such as revenues.

Total system capacity

We define total system capacity as the sum of the provided data rates. If demand variation over time is considered, it becomes the integral of the provided data rate (and a data volume). In general, we desire a maximization of the total system capacity. If all provided data rates have equal value, then it links directly to revenues. The following authors use it as a metric: Wang et al. [122], Neely et al. [124], He et al. [128], Sun et al. [147, 148], Kyrgiazos et al. [150], Shi et al. [154], Angeletti et al. [155], Wang et al. [157], Choi et al. [134, 163, 164], Jahn [166], Camino et al. [173, 174], Qian et al. [175], Wenqian et al. [176], Kyrgiazos et al. [179], and Sharma et al. [107].

Total power consumption

We define the total power consumption as the sum of the required power. Similar to total system capacity, if variables are time-dependent, the metric becomes the integral: total energy. For all practical applications, we desire a reduction in overall power consumption. The following authors aim to reduce total power consumption: Durand et al. [129], Destounis et al. [130], Srivastava et al. [131], Anzalchi et al. [149], Shi et al. [154], Lei et al. [156], París et al. [98, 161], Schubert and Holger [177, 178].

Unmet demand

We define unmet demand as the differences between the requested demand and the provided data rate. Some authors also refer to it as *unmet system capacity* or *traffic matching*. Wang et al. [122], and Hong et al. [123] use the square of the difference. Others use the sum of the positive differences: Aravanis et al. [127], Wang et al. [136], Anzalchi et al. [149], Alberti et al. [151], París et al. [98, 161], and Lei et al. [156].

Number of users

The number of served users can be a proxy for revenues (under the assumption that all users have equal value). Hence a maximization is a looked-for direction. This metric does not capture the effect of different values of the users. Destounis et al. [130], Srivastava et al. [131], and Hu and Liu [137, 138] use it.

Amount of interference

The amount of interference is commonly used for the decisions about frequency plan and link configuration. This metric assesses the goodness of the solution, whereas a smaller amount of interference is generally better. It is used by: Funabiki et al. [141], Salcedo-Sanz et al. [142, 143], Li et al. [144], Lei et al. [156], and Choi et al. [134, 163, 164].

Fairness

Chitre et al. [180] provided an overview of standardization efforts in the area of Asynchronous Transfer Mode (ATM) networks in 1999. The goal is to design for highly bursty Internet and multimedia traffic uniquely. He also proposed a resource management functionality that dynamically allocates bandwidth based on demand, available capacity, QoS, and fairness. In particular, Prakash considers three types of fairness: outgoing, incoming, and system fairness, but does not specify them further. He et al. [128] (call completion ratio) and Ji et al. [162] (delay) use additional QoS metrics.

Cocco et al. [158, 181] introduce an example of a quantitative measure of fairness. The authors define a metric called satisfaction-gap measure (SGM) that measures the mismatch between provided and

requested demand. Underserving and overserving might be weighted differently in the SGM. Through a transformation into a complex plane, a 2D plot can display all properties of the SGM.

Park et al. [132] investigates a trade-off when allocating resources, i.e., between the maximum total capacity and fairness amongst users. He uses proportional fairness without further specification.

Courcoubetis et al. [121] describe a model for proportional fairness, including a primal and dual algorithm to solve it. The authors describe two additional metrics for fairness. First, weighted proportional fairness, which is a weighted sum of the proportional rate changes. Second, the max-min fairness where it is not possible to increase some flow without simultaneously decreasing another flow that is already smaller – and therefore, give absolute priority to smaller flows.

5.1.9 Optimization techniques

So far, we have mainly organized the literature review based on domains. If applicable, we described the optimization technique for each cited source. In this Section, we aim to re-organize the literature based on optimization techniques. At highest level, we divide the literature into Classical optimization, Artificial Intelligence (AI), and miscellaneous. We furthermore split AI into Metaheuristics, Machine Learning (ML), Game theory, and Graph theory. The following describes briefly how each technique is defined, and how it maps to the discussed literature.

Classical optimization – mathematical programming

In general, mathematical programming tries to find a solution to problems by finding the extrema of functions under constraints [182]. Operations research often uses it. Depending on the form of the functions, it can be a linear, quadratic, convex, discrete/integer, or stochastic program.

Kimes [72] discusses (probabilistic) linear and dynamic programming, and networks for Yield Management. She notes that it is computationally too expensive to solve them (referring to 1989, so it is more tractable today). Wang et al. [122, 136, 157], Neely et al. [124], Keon et al. [183] use convex optimization. For equality constraints, Hong et al. [123] and Choi et al. [134, 163, 164] use Lagrange multipliers to solve a constraint optimization problem by explicit parameterization. Ji et al. [162] propose a two-step process where the first step solves a problem with relaxed constraints. Camino et al. [173] use mixed-integer programming, and Lei et al. [152, 153, 156] propose an iterative algorithm.

Artificial Intelligence

Metaheuristic

Metaheuristics are used when an exhaustive sampling of the solution space is infeasible. They use procedures to guide the algorithm from a few sample points to the optimum. We found three categories: Genetic Algorithm (GA), Simulated Annealing (SA), and Particle Swarm Optimization (PSO). He et al. [128], Angeletti et al. [155], and París et al. [98, 161] use a GA; Cocco et al. [158, 159] use a SA derivative; and Durand et al. [129] propose a PSO. Aravanis et al. [125-127] develop a hybrid combination of GA and SA. Salcedo-Sanz et al. combine a Neural Network (NN) with SA [142] and a GA [143].

Machine learning

Machine learning is a technique that relies on pattern recognition and inference. It represents a family of learning algorithms that improve their ability to recognize patterns and infer by means of a data-driven training process. We found two used techniques: first, neural networks such as by Funabiki et al. [141], second Deep Reinforcement Learning (DRL) by Hu and Liu [137, 138], and Ferreira et al. [108-110]. Salcedo-Sanz et al. [142, 143] combine a neural network with SA or GA.

Game theory

The area of game theory is concerned with the interaction between different players (often decision-makers). There are numerous different types of games. Li et al. [144] use a Stackelberg game where the players move sequentially. Vieira et al. [184] use a price-based approach, and Fiaschetti et al. [185] a static cooperative game.

Graph theory

Graphs study the pairwise relationship between objects. Each graph consists of vertices and edges. Jahn [166] and Kyrgiazos et al. [179] build upon this theory to solve the frequency plan problem by using *clique cover* algorithms. A rich literature on probably optimal algorithms for the clique cover problem can be found here [167-172].

Miscellaneous

We group techniques into miscellaneous that do not fit into any of the above categories. Park et al. [132, 135], Sun et al. [147, 148], Schubert and Holger [177, 178], Destounis et al. [130], Srivastava et al. [131], and Corbett et al. [186] derived direct analytical solutions. Anzalchi et al. [149] and Kyrgiazos et al. [150] use iterative algorithms. Similarly, Alberti et al. [151] and Sharma et al. [107] developed a system

optimization loop approach. Using search to find solutions is done by Shi et al. [154]. Camino et al. [174] and Qian et al. [175] developed greedy approaches.

5.2 Summary of literature and identification of gaps

We summarize the reviewed literature in Table 5-1. We group authors with similar work and assess the sources' content in two dimensions: (1) which dynamic parameters the authors *consider*, and (2) which optimization technique they *use*. We use a three-stage scale: red-filled means the parameters are not considered to be dynamic, or the authors do not use the algorithm. Yellow-filled with a minus indicates that the authors partially consider/use it, and green-filled with a x that they consider/use it.

With the table, we identify four main literature gaps for resource management:

- 1. Gap in formalizing the resource allocation process and its steps.** No work exists that describes the complete resource allocation process formally and breaks it down into steps that can be solved independently. Every author defines the problem differently, making it challenging for the community to build upon previous research knowledge, provide consistent comparisons, and assemble the four steps into a coherent process.
- 2. Gap in an algorithm for the routing step.** We found no work tailored to the end-to-end routing between user terminal, satellite, and gateway for NGSO satellites.
- 3. Gap in a consistent set of solutions to the resource allocation process.** Most of the reviewed literature addresses one dynamic parameter, some two at a time. No work provides consistent algorithms for all four parameters.
- 4. Gap in a comparison between algorithms.** We found no clear correlation between the problem (= dynamic parameters considered) and the used optimization technique. For example, the power allocation problem is solved by mathematical programming (Wang et al. [122], Hong et al. [123], Neely et al. [124]), but also with metaheuristics (Aravanis et al. [125-127], He et al. [128], Durand et al. [129]). Little work exists in comparing the performance of algorithms for each dynamic parameter.

Table 5-1: Summary of the conducted literature review based on dynamic parameters considered and optimization techniques used.

| | | Dynamic parameters considered | | | | | Optimization technique used | | | | | |
|--|--|-------------------------------|-----------------|------------|------------|--------------|------------------------------------|-----------------|--|-------------|---------------------|------|
| | | Power | Frequ. plan | Group- ing | Beam shape | Eco- nomical | Classical Mathematical Programming | Meta- heuristic | Artificial Intelligence Machine learning | Game theory | Graph theory | Misc |
| Power alloc. | Garau [93, 187] | x | | | | | | x | x | | | |
| | Wang [122], Hong [123], Neely [124] | x | | | | | x | | | | | |
| | Aravanis [125-127], He [128], Durand [129] | x | | | | | | x | | | | |
| | Destounis [130], Srivastava [131] | x | | | | | | | | | | x |
| Frequency assignment | Park [132, 135] | | x | | | | | | | | | x |
| | Wang [136] | | x | | | | x | | | | | |
| | Hu, Liu [137, 138], Funabiki [141], Ferreira [108-110] | | x | | | | | | x | | | |
| | Salcedo-Sanz [142, 143] | | x | | | | | x | x | | | |
| | Li [144] | | x | | | - | | | | x | | |
| | Sun [147, 148] | | - | - | | x | | | | | | x |
| Joint power and bandwidth | Anzalchi [149], Kyrgiazos [150], Alberti [151] | x | x | | | | | | | | | x |
| | Lei [152, 153], Lei [156], Wang [157] | x | x | | | | x | | | | | |
| | Shi [154] | x | x | | | | | | | | | x |
| | Angeletii [155] | | x | | | | | x | | | | |
| | Cocco [158, 159], Paris [98, 161], Pachler [160] | x | x | | | | | x | | | | |
| | Ji [162] | x | x | | | | - | | | | | |
| Group- ing | Choi [134, 163, 164] | x | | x | | | x | | | | | |
| | Jahn [166] | | x | x | | | | | | x | | |
| | Camino [173, 174] | | | x | - | | x | | | | | x |
| Beam shape | Qian [175] | | | - | - | | | | | | | x |
| | Wenqian [176] | x | x | | - | | | | | | | |
| | Schubert, Holger [177, 178] | x | | | x | | | | | | | x |
| | Kyrgiazos [179] | | x | | - | | | | | x | | |
| | Sharma [107] | | - | | - | | | | | | | x |
| Pricing | Keon [183] | | | | | x | x | | | | | |
| | Courcoubetis [121], Corbett [186] | | | | | x | | | | | | x |
| | Vieira [184] | -§ | | | | x | | | x | | | |
| | Fiaschetti [185] | | | | | x | | | x | | | |
| § considers resource allocation more generally | | | | | | | | | | | | |
| Legend: | | x | Considered/used | | | - | Partially considered/used | | | | Not considered/used | |

5.3 Specific objectives

This Section defines the specific objectives based on the reviewed literature and the identified four gaps. With this dissertation, we contribute to the first three gaps. At the same time, work is has been done by Garau et al. [93, 187] on the fourth gap in parallel to this dissertation: the comparison of Artificial Intelligence algorithms for the power allocation dynamic parameter using the model described in Chapter 4.

The specific objectives of this Chapter become:

1. Formulate a generic mathematical description of the resource allocation process and decompose it into steps that can be solved independently.
2. Develop an algorithm to solve the routing step.
3. Describe a consistent set of solutions to the resource allocation process using the notation defined by (1) and the routing algorithm developed by (2).

Note that the first objective (making the steps independent) comes with suboptimalities, since the metrics for each step are only proxies of the optimal solution input for the next steps. However, given the complexity of the problem, we believe that separating the steps is the only feasible approach to make the resource allocation process tractable. The literature review also confirms this approach: the steps are solved separately, and only a few authors worked on the joint power and amount of spectrum allocation (which is a subproblem of the frequency assignment).

The remainder of this Chapter has the following organization: Section 5.4 formalizes the resource allocation process and decouples the problem into four steps. The following four Sections, 5.5 - 5.8, describe the solutions to each step. We summarize work done by Pachler [188] under the supervision of the author for the grouping of user terminals and frequency assignment in Section 5.5 and 5.7 and extend it to include gateways. Section 5.6 describes our algorithm for the routing (ties back to second specific objective), and we outline a direct way of the power computation in Section 5.8. Section 5.9 discusses the results generated by the described resource allocation process for a constellation similar to SES's O3b mPower. In the final Section 5.10, we summarize the Chapter and our contributions.

5.4 Resource allocation process

The decisions in the resource allocation problem are in four areas for a given time instance t . Figure 5-2 illustrates the first three steps as we define them. The initial representation has two satellites 1 and 2 in the same orbital plane, contains four user terminals A, B, C, and D, as well as two gateways I and II. In a first step, the users are grouped into beams, i.e., the beams placed, and their shape defined. Step 2 is to determine the routing between users, satellites, and gateways. In Step 3, the center frequency, amount of spectrum, reuse group, and polarization are assigned.

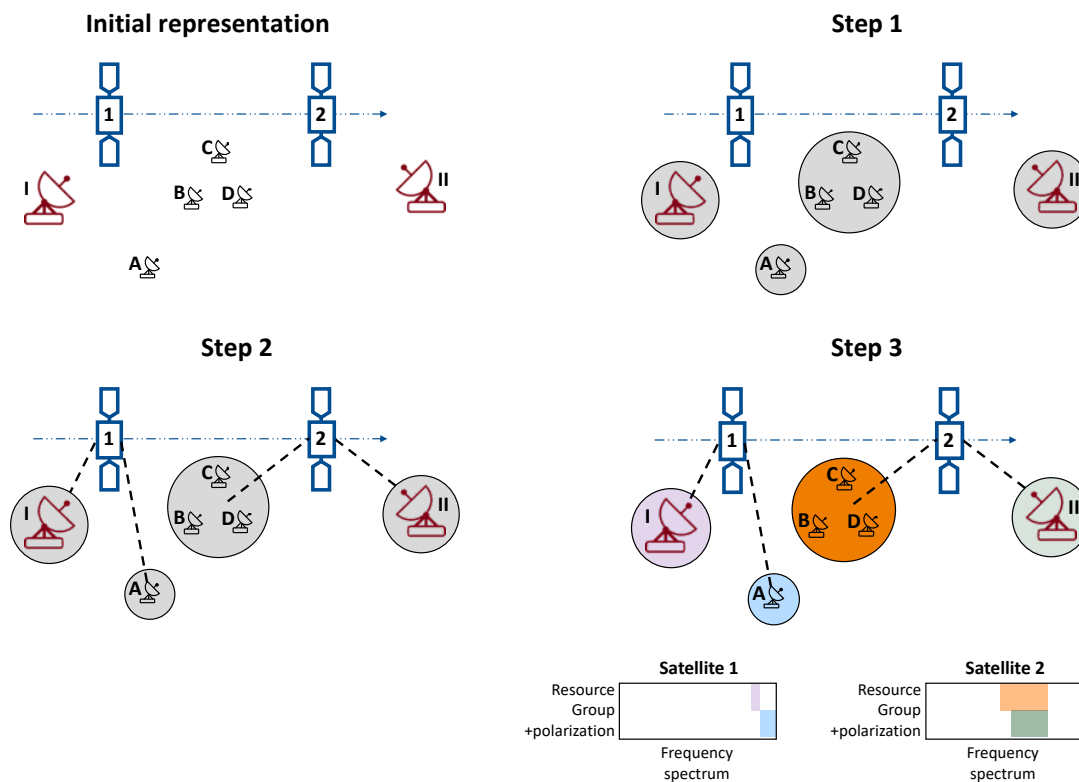


Figure 5-2: The resource allocation process showing two satellite 1 and 2 in the same plan, four user terminals A, B, C, and D, as well as two gateways I and II. Step 1: grouping user terminals; Step 2: routing; Step 3: frequency assignment.

In the remainder of this Section, we first introduce the terminology and notation of the resource allocation process and then describe the solution approaches in Sections 5.5 - 5.8. We focus on the downlink beams only.

Step 1. Grouping user terminals

The first decisions are about how many beams to place, where to place them, and with which shape. That directly implies the grouping of users within a beam. We define two different types of beams: user beams

and gateway beams. We assume that gateway beams do not point to multiple gateways at the same time, and the beam's position centers over the gateway. User beams can contain multiple user terminals, and their pointing longitude and latitude do not necessarily coincide with the ones for the user terminals that it is covering. Each user has a forward demand from the gateway to the user and a return demand, in the opposite direction. In general, demand is a function of the time. Formally, the input is a set of users, as defined in Eq. (5-1) below, with latitude and longitude attributes and demand in both directions.

$$\text{Users } \mathcal{U} = \{u_i = (\phi_{u_i}, \psi_{u_i}, D_{FWD,u_i}, D_{RTN,u_i})\} \forall i \in \{1, \dots, N_u\} \quad (5-1)$$

with $N_u = \text{number of users}$

$\phi_{u_i}, \psi_{u_i} = \text{latitude, longitude of user } u_i$

$D_{FWD,u_i}(t) = \text{demand forward of } u_i \text{ at time } t$

$D_{RTN,u_i}(t) = \text{demand return of } u_i \text{ at time } t$

For the beam set (see Eq. (5-4)), the decision variables are the latitude ϕ_{b_i} and longitude ψ_{b_i} and the number of beams N_b and their shape S_{u_i} . We can also interpret the problem as a partitioning problem of dividing users into beams. We formally define the set \mathcal{U}_{b_i} that is a subset of the users \mathcal{U} , which are into beam b_i . Eqs. (5-2) and (5-3) express the disjoint union property. In other words, a user can only belong to a unique beam.

$$\bigcup_{i=1}^{N_b} \mathcal{U}_{b_i} = \mathcal{U} \quad (5-2)$$

$$\mathcal{U}_{b_i} \cap \mathcal{U}_{b_j} = \emptyset \text{ for } i \neq j \quad (5-3)$$

The outcome of the beam grouping problem is the set of user beams \mathcal{B}_u as defined in Eq. (5-4).

$$\text{User beams } \mathcal{B}_u = \{b_{u_i} = (\mathcal{U}_{b_i}, \phi_{b_i}, \psi_{b_i}, S_{u_i}, D_{FWD,b_i})\} \forall i \in \{1 \dots N_b\} \quad (5-4)$$

with $N_b = \text{number of beams}$

$\mathcal{U}_{b_i} = \text{set of users grouped to beam } i$

$S_{u_i} = \text{shape of beam of user beam } i$

$\phi_{b_i}, \psi_{b_i} = \text{latitude, longitude of beam } i$

$D_{FWD,b_i}(t) = \text{demand forward of beam } i \text{ at time } t$

$D_{RTN,b_i}(t) = \text{demand return of beam } i \text{ at time } t$

The demand in the beam becomes a sum of the users' demand for both the forward and return direction (for ease of notation we do not include the time dependency t):

$$D_{FWD,b_i} = \sum_{u_i \in \mathcal{U}_{b_i}} D_{FWD,u_i}, \quad D_{RTN,b_i} = \sum_{u_i \in \mathcal{U}_{b_i}} D_{RTN,u_i} \quad (5-5)$$

Furthermore, we define a set of downlink gateway beams \mathcal{B}_g that have a shape S_{g_i} and the return demand of the users $D_{RTN,b_i}(t)$. Since each user beam connects to a gateway, the cardinality equals the one for the user beams N_b . In the routing step, we will map these gateway beams to gateways, which defines their position.

$$\text{Gateway beams } \mathcal{B}_g = \{b_{g,i} = (S_{g_i}, D_{RTN,b_i}(t))\} \quad \forall i \in \{1 \dots N_b\} \quad (5-6)$$

Step 2. Routing

The routing step consists of two main decision: (1) the mapping (or assignment) of user beam (and therefore set of users) to gateways, and (2) the decision of which satellites establishes this connection as a function of the time (especially in case of NGSOs). For that we define a set of satellites and gateways:

$$\text{Satellites } \mathcal{S} = \{s_i = (o_i)\} \quad \forall i \in \{1 \dots N_s\} \quad (5-7)$$

$$\text{Gateways } \mathcal{G} = \{g_i = (\phi_{g_i}, \psi_{g_i})\} \quad \forall i \in \{1 \dots N_g\} \quad (5-8)$$

with $N_s = \text{number of satellites}$

$o_i = \text{orbit of satellite } i$

$N_g = \text{number of gateways}$

$\phi_{g_i}, \psi_{g_i} = \text{latitude, longitude of gateway } i$

The decision variable is the mapping between a gateway beam and a gateway written as $b_g \sim g$, so the output is a set of mappings:

$$\text{Mappings } \mathcal{M} = \{m_i = (b_g, g) \mid \forall b_g \in \mathcal{B}_g, \forall g \in \mathcal{G}, b_g \sim g\} \quad \forall i \in \{1 \dots N_b\} \quad (5-9)$$

with $b_g \sim g \leftrightarrow \text{beam mapped to gateway}$

And then, the routing is defined as the representation of the mapping of set \mathcal{M} to the set of satellites \mathcal{S} .

$$\text{Routing } R: \mathcal{M} \rightarrow \mathcal{S} \quad (5-10)$$

Step 3. Frequency assignment

From step 2, the mapping defines the target latitude and longitude of the gateway beams (the geographical position), and the routing defines the origin point (the satellite). Hence, the geometry of both beam sets \mathcal{B}_u and \mathcal{B}_g is fully described. That information is necessary to compute interference constraints between beams to inform the frequency assignment step 3. We first combine all beams into one set \mathcal{B} that now has the cardinality $2 \cdot N_b$.

$$\mathcal{B} = \mathcal{B}_u \cup \mathcal{B}_g = \{b_i\} \quad \forall i \in \{1 \dots 2 \cdot N_b\} \quad (5-11)$$

The output of this step is a frequency assignment that sets the center frequency and bandwidth for each beam. Additionally, if the satellite allows for frequency reuse, the reuse group and polarization are assigned in this step as well. Furthermore, for frequency reuse in the time domain (beam hopping), a flag $\zeta_i(t)$ defines if the beam is active or not for the considered time instance. Eq. (5-12) formalizes the definition of the frequency assignment set \mathcal{A} .

$$\text{Frequency assignment } \mathcal{A} = \{a_i = (b_i, f_i, B_i, rg_i, pol_i, \zeta_i)\} \quad \forall i \in \{1 \dots 2 \cdot N_b\} \quad (5-12)$$

with $f_i =$ center frequency of assignment a_i

$B_i =$ bandwidth of assignment a_i

$rg_i =$ resource group ($\in \{1, 2 \dots N_{rg}\}$)

$pol_i =$ polarization ($\in \{"L", "R"\}$)

$\zeta_i =$ active or not active ($\in \{True, False\}$)

Step 4. RF power allocation

The final step in the resource allocation process is the computation of the required RF power on the satellite for all downlinks, i.e., forward link to user terminals and return link to the gateway. This computation requires the outputs from the previous steps \mathcal{S} , \mathcal{B} , \mathcal{A} , and R . Moreover, further parameters of the environment need to be defined as detailed further in Chapter 4. The output is then a set of powers for each beam $b_i \in \mathcal{B}$:

$$\mathcal{P} = \{p_i\} \quad \forall i \in \{1 \dots 2 \cdot N_b\} \quad (5-13)$$

While there is some work done on these four steps, the existing literature is falling short of addressing all aspects of the resource allocation process (as we reviewed exhaustively at the beginning of this Chapter and summarized in Table 5-1). We discuss our solutions in the following four Sections in more detail.

5.5 Grouping of user terminals

The objective of this work is to develop an algorithm to cover the set of users \mathcal{U} with a set of beams \mathcal{B}_u in a best manner, whereas “best” is not trivial to define here. Since the goal is to decouple the resource allocation process and tackle each subproblem individually, we need a definition that does not require the execution of the steps downstream. For example, minimizing the power consumption would require finding the routing, the frequency assignment, and the RF power allocation.

As shown in the literature review from Section 5.1, little work has been done in the area of user terminal grouping. The most recent work is done by Pachler et al. [188], under the author’s supervision and guidance. A pending patent by SES protects parts of that work. The optimality criterium of his work is to minimize the number of beams N_b for a given set of users \mathcal{U} . For completeness of this dissertation, we first summarize his work and then discuss the limitations of the chosen criterium. In doing so, we adapt the text and figures from his original publication [188] and align the notation with the one used in this dissertation.

5.5.1 Finding the minimum number of beams through edge clique cover

As a first step, Pachler et al. [188] compute the angular separation angles α_{ij} between users u_i and u_j as viewed from the satellite. For NGSO satellites, this is done for discrete time-instances t until the ground track repeats, so we get α_{ij_t} . In that case, we are interested in the maximum angle of all time-instances $T: \alpha_{ij} = \max(\alpha_{ij_t} \forall t \in T)$ where T is the set of all time instances. This angle is then compared with the aperture/cone angle δ , and a Boolean adjacency matrix is constructed based on the condition $\alpha_{ij} \leq \delta$, i.e., the user terminals u_i and u_j are close enough together to fit into a beam with the full-cone angle δ . This method is limited to circular beams.

In the next step, this Boolean adjacency matrix is transformed into an undirected graph, where an edge represents that two users can be in the same beam, i.e., $\alpha_{ij} \leq \delta$ holds. The objective of finding the minimum number of beams becomes the search of the shortest list of maximum cliques \mathcal{C}_{max} , also known as *edge clique cover* problem. The optimal solution to this problem is NP-hard [168]. Therefore, Pachler developed a heuristic approximation that scales with $\mathcal{O}(N_{c,max} \cdot N_u)$ where $N_{c,max}$ is the size of the largest clique and N_u the number of users. Algorithm 5-1 outlines the method. First, the list of maximum cliques \mathcal{C} in the graph is sorted in descending order by size (and randomized for the same size). An empty set initialized the solution set \mathcal{C}_{max} , and the allowed collision order $o_{collision} = 0$. A collision is defined when a user is inside two cliques. Then, in line 5, the algorithm iterates through the cliques and checks if

the collision order is lower than the allowed $o_{collision}$. If that is true, the algorithm adds it to the solution \mathcal{C}_{max} . After iterating through \mathcal{C} , the collision order is incremented by 1. If the solution covers all terminals ($|\mathcal{C}_{max}| < N_u$), the while loop terminates, and the algorithm returns the solution set \mathcal{C}_{max} .

Algorithm 5-1: heuristic approximation of beam placement, adapted from Pachler [188], protected under pending patent by SES.

| | |
|--|--|
| <p>Input: cliques \mathcal{C}</p> <p>Input: N_u</p> <p>Output: \mathcal{C}_{max}</p> <p>1: $Sort(\mathcal{C})$</p> <p>2: $\mathcal{C}_{max} = \{\}$</p> <p>3: $o_{collision} = 0$</p> <p>4: while $\mathcal{C}_{max} < N_u$ do</p> <p>5: for c in \mathcal{C} do</p> <p>6: if $c \cup (U\mathcal{C}_{max}) > U\mathcal{C}_{max}$ and $c \cap (U\mathcal{C}_{max}) \leq o_{collision}$ then</p> <p>7: $\mathcal{C}_{max} = \mathcal{C}_{max} \cup (c - (U\mathcal{C}_{max}))$</p> <p>8: end if</p> <p>9: end for</p> <p>10: $o_{collision} = o_{collision} + 1$</p> <p>11: end while</p> | <p>// set of maximal cliques of the graph</p> <p>// number of users</p> <p>// set of maximal cliques that form the solution</p> <p>// sort cliques by size in descending order</p> <p>// Initialize empty solution</p> <p>// Initialize collision order</p> <p>// check solution completeness</p> <p>// If clique is not in solution and the collision order is lower than // the threshold</p> <p>// Add maximum non-colliding clique to the solution</p> <p>// Increase collision order</p> |
|--|--|

Pachler [188] compares his algorithm against two other approaches: brute-force full enumeration that ensures optimality, and a grid approach that lays out a grid of beams and removes these with no users in it. The test case was SpaceX’s Starlink constellation with a traffic model consisting of 18,712 user terminals. Figure 5-3 shows the results with a sweeping subsets of user terminals on the horizontal axis. The vertical axis represents the execution time on the left plot and the number of beams on the right plot. NR stands for the number of runs of the heuristic algorithm. That is necessary since the sorting in line 2 of Algorithm 5-1 is non-deterministic.

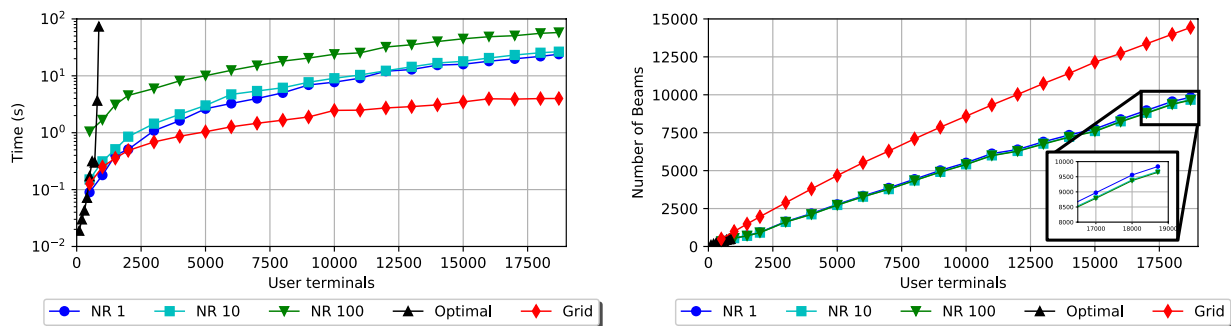


Figure 5-3: benchmarking results of the Pachler’s heuristic approximation against a brute-force and grid approach; NR = number of runs of the algorithm. Obtained from [188].

The results show that the optimal brute-force enumeration is intractable for more than 850 terminals. The grid approach is around one order of magnitude faster but results in around 40% more beams. Re-running the Pachler’s algorithm multiple times increases the quality of the solution, but with diminishing return. Overall, 9,633 beams cover 18,712 user terminals. Note that this is different from the results in Section 5.9 because Starlink’s beams have a smaller footprint.

Limitations

We see the main limitation of Pachler’s approach in the objective of minimizing the number of beams N_b . The advantages of this metric are that it reduces the consumption of the resource “beam” and makes the frequency assignment easier as there are fewer beams to deconflict. However, the power consumption increases, as user terminals are off-centered from the peak gain. In the worst case, two users are at the 3-dB border of the beam resulting in a doubling of the power consumption (see the red beam in the top right of Figure 5-4 for an example).

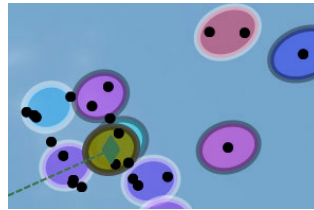


Figure 5-4: Example of a suboptimal beam placement. Black dots represent user terminals; the filled ellipses are the projection of circular beams onto the Earth’s surface; the edge and fill color represent the assigned center frequency and polarization.

On the other hand, if we aim to maximize the gain the users receive, the optimal solution would be to provide every user with their beam and center them for peak gain. However, particularly in high-density areas, there will be too many overlapping beams that cannot be deconflicted in frequency anymore. Hence, user grouping becomes a multi-dimensional optimization problem that, on the one hand, aims to minimize the number of beams and, on the other hand, maximize the received gain for the users.

As a follow-on work of Pachler, Alinque [189] elaborates on the trade-off between the number of beams and the gain. Furthermore, he proposes a *joint* gradient-based optimization of beam placement *and* shape. Alinque compares the results for multiple realistic scenarios and concludes that especially dense regions benefit from the joint optimization. However, he also states that further investigations are necessary to quantify the dependency on the density of user terminals in more detail.

5.6 Routing

We define routing as the mapping of user terminals over satellites to gateways (see Eq. (5-10)). For NGSO, this mapping is time-dependent, but it can also be a function of the time for GEO with dynamic gateway balancing strategies. The algorithms in this Section apply to all constellations with repeating ground tracks, i.e., the mapping is a stationary, repeatable plan. One example of such a constellation is SES’s O3b mPower, for which each satellite revisits the same location on Earth every six hours and, therefore, four times per day. If the constellation has a non-repeating ground track such as specific versions of inclined SES O3b mPower satellites, or SpaceX’s Starlink, then the algorithms below need to run in real-time and are strongly coupled with other constraints.

We desire a mapping that consumes the least amount of resources, is robust, and operationally implementable. The general concept of the algorithms in this Section has four stages:

1. Calculating access windows of all users and gateways to all satellites and identify all possible combinations, i.e., their overlap is sufficient for continuous coverage
2. Applying constraints to the set of possible combinations, e.g., in and out of country traffic requirements
3. Selecting the best routing
4. Selecting the start and end time of the beam task with a specific satellite to connect through

While we develop one approach for stages 1, 2, and 4, for stage 3, we develop and compare two different heuristics: *closest-first* and *balancing allocation*. In the following Section 5.6.1, we describe all four stages and the two heuristics in more detail. Section 5.6.2 compares both heuristics and discusses the results.

5.6.1 Algorithm description

Stage 1. Computing access windows and all possible combinations

In this calculation, we use the combined set of user and gateway beams \mathcal{B} and the constellation consisting of the set of satellites \mathcal{S} . Geometrical line-of-sight calculations return the access windows (start and end time for a clear view from user to satellite). The simulator propagates the orbit and stores for each discrete timestep the elevation angle, as seen from the users. We set the duration of the simulation to the ground repetition of the constellation. With a given elevation angle threshold (here 10 deg), we compute the access window, which is the time during which the elevation angle is above the threshold. We define an access window $\mathcal{W}_{u/g}$ as a set of start time t_0 , end time t_1 , beam, and satellite. The sets of access windows for users \mathcal{W}_u and gateways \mathcal{W}_g are with the cardinalities $N_{w,u}$ and $N_{w,g}$.

$$\mathcal{W}_u = \{w_{u,i} = (t_{u,0,i}, t_{u,1,i}, b_u, s)\} \forall i \in \{1 \dots N_{w,u}\} \quad (5-14)$$

$$\mathcal{W}_g = \{w_{g,i} = (t_{g,0,i}, t_{g,1,i}, g, s)\} \forall i \in \{1 \dots N_{w,g}\} \quad (5-15)$$

Given these access windows, we iterate through all combinatorial combinations between \mathcal{W}_u and \mathcal{W}_g (i.e., all routings between users and gateways) and save all feasible combinations that have a sufficient long overlap to allow for continuous coverage. In the case of the seven O3b mPower satellites with a 6-hour ground track, the overlap needs to be more than $6 \text{ hours}/7 = 51.4 \text{ minutes}$. We consider an additional margin $\Delta t_{handover}$ to allow for handover so that we formally get:

$$\Delta t_{min} = \frac{\Delta t_{ground-track}}{N_s} + \Delta t_{handover} \quad (5-16)$$

The overlap time between \mathcal{W}_u and \mathcal{W}_g can be calculated by subtracting the latest start time from the earliest end time:

$$\Delta t_{u,g} = \min(t_{g,1}, t_{u,1}) - \max(t_{g,0}, t_{u,0}) \quad (5-17)$$

With that, we can enumerate all possible combinations and check if they fulfill $\Delta t_{u,g} \geq \Delta t_{min}$. We formally now get a set of all possible combinations $\mathcal{M}_{possible}$ with cardinality $N_{m,possible}$ to map a user to a gateway. Additionally, we store the access window duration $\Delta t_{u,g}$.

$$\begin{aligned} \mathcal{M}_{possible} = \{m_i = (b_g, g, \Delta t_{u,g}) \mid \forall b_g \in \mathcal{B}_g, \forall g \in \mathcal{G}, b_g \sim g\} \forall i \in \{1 \dots N_{m,possible}\} \\ b_g \sim g \leftrightarrow \Delta t_{u,g} \geq \Delta t_{min} \end{aligned} \quad (5-18)$$

Stage 2. Applying constraints

Depending on the specific requirements for the resource allocation process, there might be constraints regarding which user connects to which gateway. For example, some countries might only allow traffic to be routed back into their own country, i.e., the user and gateway have to be in the same country. We differentiate this case by calling these cases in-country gateways versus out-of-country gateways where the traffic can land in any gateway. We apply these constraints to the list of possible mappings $\mathcal{M}_{possible}$ to obtain a new reduced set of allowed mappings $\mathcal{M}_{allowed}$:

$$\mathcal{M}_{allowed} \subseteq \mathcal{M}_{possible} \text{ such that } b_g \sim g \text{ satisfies all constraints} \quad (5-19)$$

The in- and out-of-country traffic is only one example of a constraint. There can be many more constraints, such as:

- some customer might want to exclude routing to a specific subset of countries due to security

- if customers own their gateways, they might prefer routing through them

All hard constraints, such as the in- and out-of-country traffic requirement, can be applied straight to Eq. (5-19) and reduce the set of allowable mappings $\mathcal{M}_{allowed}$.

Stage 3. Selecting the best mapping

In this stage, the algorithm selects the best mapping \mathcal{M} from $\mathcal{M}_{allowed}$. It is a multi-objective optimization problem with the following desired objectives:

1. maximize the number of connected users
2. robustness against uncertain demand and unexpected events (balanced allocation margins)
3. minimize the distance between user and gateway to reduce the path loss and, therefore, power consumption; this also increases the overlapping time of the access windows and therefore results in higher operational margins.

The approach we take in developing the algorithm is greedy heuristic. The highest priority of the algorithm is the first objective. Given that the maximum number of users can be connected, the heuristic tries to optimize the second and third objective. We design and implement two heuristics that we describe in the following and compare using the three objectives.

Closest-first

This algorithm maps each user to the closest gateway without considering the gateways' fill rates. This approach reduces the Euclidian distance between user and gateway, and hence, reduces the power consumption. Algorithm 5-2 shows the pseudo-code. It takes the allowed mapping $\mathcal{M}_{allowed}$ and the gateways \mathcal{G} as inputs and returns the desired mapping \mathcal{M} . The procedure then iterates through all user beams \mathcal{B}_u and finds the closest gateway to that user beam $b_{u,i}$ that $\mathcal{M}_{allowed}$ allows. The corresponding gateway beam $b_{g,i}$ for the same user i is then added together with the closest gateway $g_{closest}$ to the set of mappings \mathcal{M} .

Algorithm 5-2: closest-first routing algorithm

Input: $\mathcal{M}_{allowed}, \mathcal{G}, \mathcal{B}_g, \mathcal{B}_u$
Output: mapping \mathcal{M}

- 1: **Init** $\mathcal{M} = \{\}$
- 2: **for** $b_{u,i}$ in \mathcal{B}_u **do**
- 3: $g_{closest} = \text{closestGateway}(b_{u,i}, \mathcal{M}_{allowed}, \mathcal{G})$ // finds the closest gateway for a user beam given $\mathcal{M}_{allowed}$
- 4: $\mathcal{M} = \mathcal{M} \cup (b_{g,i}, g_{closest})$ // adds the mapping of gateway beam to gateway
- 5: **end for**

We will see in the results at the end of this Section that the allocation is unbalanced – some gateways are beyond their capacity limits, and some of them are almost empty. Motivated by this, we developed a second algorithm described below that balances the load between different gateways while still trying to map to the closest gateway.

Balancing allocation

The idea behind this second algorithm is to fill up gateways evenly by iteratively raising a threshold, after which a gateway is considered full. We use the generic unit of the average spectral efficiency Γ as a measure of the “fullness” of the gateway. Γ is a normalized measure making comparisons possible across different gateway terminal sizes, polarization, and amount of processable spectrum. Also, Γ is directly proportional to the needed RF power.

The actual spectral efficiency is a complex function of the dynamics of the constellation and the actual usage of the users. However, we can estimate an upper, worst-case bound Γ_{wc} by assuming CIR consumption denoted as D_{FWD,CIR,b_i} for beam b_i . We, therefore, get with the available bandwidth B_{g_i} the upper bound for the gateway i with

$$\Gamma_{wc,i} = \frac{\sum_{all\ b_g\ mapped\ to\ g_i} D_{FWD,CIR,b_i}}{B_{g_i}} \quad (5-20)$$

The equation sums across all gateway beams b_g mapped to gateway g_i . The algorithm obtains that through the mapping set \mathcal{M} . By fixing the bandwidth and spectral efficiency, we can directly compute the allowable maximum data rate in the beam using Eq. (5-20). We use $\Gamma_{wc,i}$ to determine if we consider a gateway to be full or not. For that, we introduce the variable Γ_{allow} that sets this limit. We write the following conditions:

$$\begin{aligned} g_i\ not\ full &\leftrightarrow \Gamma_{wc,i} \leq \Gamma_{allow} \\ g_i\ full &\leftrightarrow \Gamma_{wc,i} > \Gamma_{allow} \end{aligned} \quad (5-21)$$

With this notation, the pseudo-code in Algorithm 5-3 describes the resulting method. First, we iterate through all user beams and allocate these users that can only be connected to one gateway (lines 2-7). These allocated beams are then removed from the set of remaining beams $\mathcal{B}_{remaining}$. Next, we sort the remaining beams based on their CIR traffic in descending order starting with the beam having the highest demand. This heuristic ensures that the higher-traffic beams are allocated first and the lower-traffic beams can be filled into the remaining gaps. In addition, this sorting favors smaller distances between

user and gateway for higher-traffic beam as they are allocated first and it is more likely that a closer gateway is still available.

Algorithm 5-3: Balancing routing algorithm

```

Input:  $\mathcal{M}_{allowed}, \mathcal{G}, \mathcal{B}_g, \mathcal{B}_u$ 
Input:  $\Gamma_{max}, \Gamma_{allow}, \Delta\Gamma$ 
Output: mapping  $\mathcal{M}$ 
1: Init  $\mathcal{M}_{single} = \{\}, \mathcal{B}_{remaining} = \mathcal{B}_u$ 
2: for  $b_{u,i}$  in  $\mathcal{B}_u$  do
3:   if  $numberOfGateways(b_{u,i}, \mathcal{M}_{allowed}) == 1$  then           // only one option for the mapping
4:      $\mathcal{M}_{single} = \mathcal{M}_{single} \cup (b_{g,i}, g)$                        // add mapping
5:      $\mathcal{B}_{remaining} = \mathcal{B}_{remaining} \setminus b_{u,i}$                  // remove allocated user beam from set
6:   end if
7: end for
8:  $Sort(\mathcal{B}_{remaining})$                                            // sort based on  $D_{FWD,CIR,b_i}$  in descending order
9: while  $\Gamma_{allow} < \Gamma_{max}$  do
10:   $\mathcal{M} = \mathcal{M}_{single}$ 
11:  for  $b_{u,i}$  in  $\mathcal{B}_{remaining}$  do
12:     $g_{closest} = closestNotFull(b_{u,i}, \mathcal{M}_{allowed}, \mathcal{G}, \Gamma_{allow})$  // get the closest gateway that is not full
13:    if  $g_{closest}$  exists do
14:       $\mathcal{M} = \mathcal{M} \cup (b_{g,i}, g_{closest})$ 
15:    end if
16:  end for
17:  if  $length(\mathcal{M}) == N_b$  then                                   // check if all beams are assigned
18:    break
19:  else
20:     $\Gamma_{allow} = \Gamma_{allow} + \Delta\Gamma$                        // increase the allowed spectral efficiency
21:  end if
22: end while

```

In line 9, the while loop iteratively finds the lowest spectral efficiency Γ_{allow} for which the algorithm can allocate all gateway beams. We give an initial is given as input (e.g., 0.5) and the increment $\Delta\Gamma$ (e.g., 0.1). The while loop ends if Γ_{allow} exceeds a maximum Γ_{max} , which 5.9 for the highest MODCOD 256APSK3/4 considered (see Appendix G). Note that this limit is likely to be lower in reality as most communication satellites are power constrained, and therefore achieving such a MODCOD cannot be sustained for an extended period. In that case, the traffic volume on the user side is too large for the number of gateways. In each iteration of the while-loop, the algorithm iterates through $\mathcal{B}_{remaining}$ and tries to find a gateway that is allowed by $\mathcal{M}_{mapping}$ and not full. If there is more than one gateway, the closest one is selected.

If a gateway can be found, the algorithm adds $(b_{g,i}, g_{closest})$ to the mapping \mathcal{M} . The while loop terminates with the minimum Γ_{allow} if we map all of the gateway beams ($length(\mathcal{M}) == N_b$).

Stage 4. Selecting start and end time of beam task

The final stage is to select a start and end time of the beam task, more formally the routing $R: \mathcal{M} \rightarrow \mathcal{S}$, which defines how the mapping \mathcal{M} corresponds to the satellites \mathcal{S} over time. For many LEO constellations, there are multiple satellites in the Field of View (FOV) resulting in an additional decision: to which satellite to connect. Pachler et al. [188] discusses one possible approach. They divide the constellation into satellite groups. Each of these groups provides global coverage, and it reduces the dimensionality of the decision problem greatly. The algorithm has to decide which satellite group to connect to. Then, the satellite in this group is selected on a closest-first basis. We refer for more details to reference [188] and focus our formulation on the case where there is only one possible satellite in the FOV for the whole duration, and the ground track is repeating.

Due to the repeating ground pattern of 6 hours for O3b mPower and its seven satellites, the duration of the beam task is 51.4 minutes without a handover margin $\Delta t_{handover}$, as derived by Eq. (5-16). Therefore, we cut down the access windows to this duration and define the beam start $t_{b,0}$ and end time $t_{b,1}$. To allow for the most robust operation, we aim to center the beam task windows over the overlapping access windows of the user terminal and gateway. Formally, we write by using Eq. (5-16) and Eq. (5-17).

$$\begin{aligned}
 t_{b,0} &= \max(t_{g,0}, t_{u,0}) + \frac{\Delta t_{u,g} - \Delta t_{min}}{2} \\
 t_{b,1} &= \max(t_{g,0}, t_{u,0}) + \frac{\Delta t_{u,g} + \Delta t_{min}}{2} = t_{b,0} + \Delta t_{min}
 \end{aligned}
 \tag{5-22}$$

With the access windows \mathcal{W}_u and \mathcal{W}_g , we can identify the corresponding satellite s_i and combine it with the mapping \mathcal{M} to the routing projection that now has the information about the start and end time of the beam task, and the user terminal and gateway targets. This information is required for the frequency assignment computation.

5.6.2 Results

We use the traffic model of 18,712 users described in Section 5.9 and the solution obtained by implementing the user grouping algorithm from Section 5.5 for a constellation similar to O3b mPower as a basis for the routing algorithm (further details in Section 5.9). We compare the two heuristics closest-

first and balanced allocation concerning the desired objectives: maximize the number of mapped users, robustness against uncertainty, and minimize the distance between user and gateway.

The user grouping algorithm results in 2014 beams covering the 18,712 user terminals. Out of these 2014 beams, both heuristics can connect 1932 with 181 gateways. For the remaining 82, either no in-country gateway can be found, or the overlapping time between gateway and user is not sufficient, i.e., $\Delta t_{u,g} < \Delta t_{min}$. We observe this especially for high latitudes users to which equatorial satellites have a significantly reduced access window.

We report the results as one bar plot per heuristic (see Figure 5-5). We color the 181 gateways based on the attribute of in- or out-of-country. The height of the bars is the aggregated return CIR demand for all beams that connect to the gateway g_i : $\sum_{all\ b_g\ mapped\ to\ g_i} D_{FWD,CIR,b_i}$.

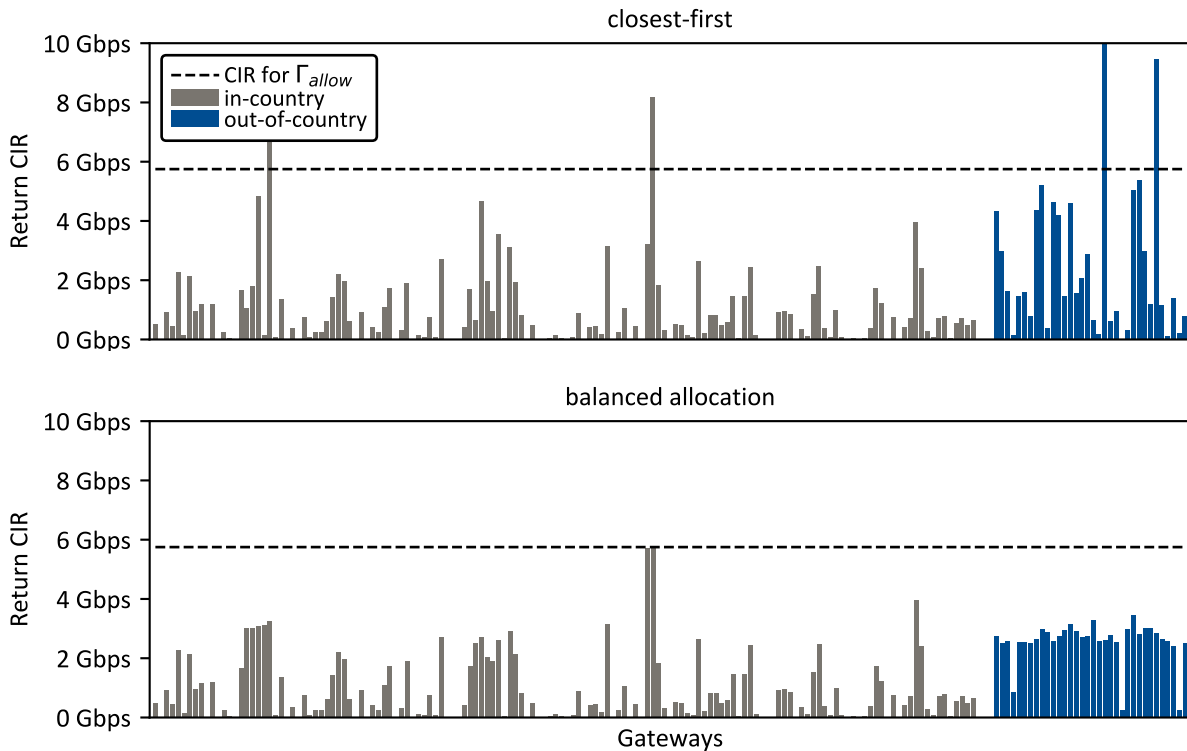


Figure 5-5: comparison of the resulting allocation for the closest-first and balanced allocation heuristic. The horizontal shows the 181 gateways colored by in-country and out-of-country. The black dashed line represents the CIR for Γ_{allow} from the balanced allocation algorithm.

The average return CIR per gateway is 1.18 Gbps for both heuristics as the number of mapped user beams are the same. However, the standard deviation is higher for the closest-first heuristic (1.78 Gbps vs. 1.23 Gbps). The balanced heuristic finds a solution with a maximum return CIR of 5.7 Gbps (results in a

spectral efficiency of 2.28 for a bandwidth of 2.5 GHz) while the maximum for the closest-first is 12.5 Gbps. Assuming a 2.5 GHz bandwidth, the maximum of 12.5 Gbps for the closest-first would result in a high required spectral efficiency of 5 needing considerable power and leaving little margin for contingencies. Figure 5-5 highlights that the balanced allocation reduces the peak and distributes them to other gateways when possible. One example is in the center of the plot, where the closest-first heuristic connects all user terminals to the closest gateway. The balancing algorithm distributes the load between the two in-country gateways (in this case, Mexico). Another example is the out-of-country gateways, where generally, the re-allocation is more flexible. Except for three low utilized gateways in France, New Zealand, and Argentina, the load is evenly balanced.

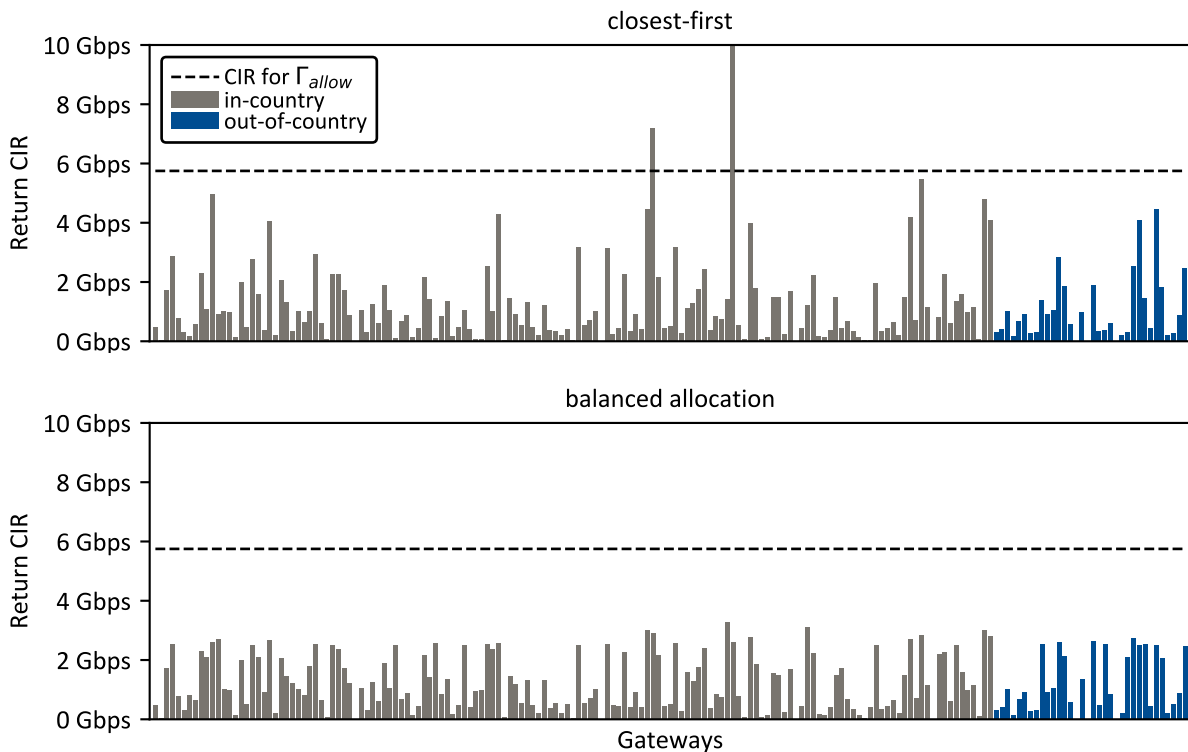


Figure 5-6: Running the routing algorithm without applying the in- and out-of-country constraints resulting in a better load balancing. The black dashed line is for comparison from Figure 5-5.

We further run both heuristics without applying the in- and out-of-country constraints with the plots shown in Figure 5-6. As expected, the additional flexibility allows the algorithm to balance the load better. The load for the two gateways in Mexico with the highest utilization is cut in half, reducing the maximum spectral efficiency from 2.28 to 1.10.

To quantify the third objective of minimizing the distance between user terminals and gateways, we use the slant range and the resulting free space loss. Both are a function of the time, so we simulate one complete orbit for the constellation of 6 hours with a 5-minute discretization. We record for the 72 timesteps the metrics for all 1932 beams and repeat the simulation for both heuristics. The mean and the standard deviation in Table 5-2 summarize the results.

The minimum slant range is when the satellite is nadir pointing (in the simulation 8075 km). The resulting mean for the closest first is 9279 km and around 0.5% more for the balanced allocation (with a 2.4% increase in the standard deviation). The mean free space loss shows little difference in the mean, but the standard deviation increases by 1.7% (note that we compute here the relative difference in dB, which requires a careful interpretation). The required RF power is log-linear with free space loss. Therefore, the balanced allocation only needs minimal more RF power than the closest-first heuristic.

Table 5-2: Comparison of the mean and standard deviation of the slant range and the free space loss for both routing heuristics. Data is generated by a 6-hour simulation with 5-minutes discretization. The percentages are the relative difference between closest-first and balanced.

| | slant range [km] (lower is better) | | | | free space loss [dB] (lower is better) | | | |
|----------------------|------------------------------------|------|-----------|------|--|------|-----------|------|
| | mean | | std. dev. | | mean | | std. dev. | |
| closest-first | 9279 | | 874 | | 197.391 | | 0.839 | |
| balanced | 9326 | 0.5% | 896 | 2.4% | 197.433 | 0.0% | 0.854 | 1.7% |

We compute the solutions of the routing algorithm within a few seconds on a single core of a standard laptop. Stage 1 is the most expensive part as the simulator propagates the whole constellation and updates the geometry of all user terminals and gateways to all satellite (scales linear with the number of discrete steps, the number of beams, and the number of satellites). The closest-first algorithm scales linear with the number of beams, and the balanced heuristic is linear with the number of beams and the size of the increment $\Delta\Gamma$.

In summary, both heuristics equally maximize the number of connected users, given the constraints. The balanced allocation strategy reduces peak loads effectively and distributes them evenly to other gateways. Even though the results are likely to be close to optimum, the heuristic does not ensure optimal balancing. It sorts the list of beams to allocate based on high demand to low. An extension could be to randomize further this list, similar to the sorting of cliques with equal length in Section 5.5. Finally, the balanced allocation solution has less than a 1% longer distance between user terminals and gateways,

resulting in negligible power differences. We conclude that, for most cases, the balanced allocation heuristic is the preferred implementation.

5.7 Frequency assignment

This step assigns the center frequency f_i , bandwidth B_i , reuse group rg_i , polarization pol_i , and the active flag ζ_i for beam b_i . An optimal solution uses all of the available bandwidth without causing interference between users. Since for a given data rate, the power is inversely proportional to the bandwidth, using the maximum bandwidth equals the use of the least power. Note here that due to the non-linear relationship between power and data rate, the split of the bandwidth pool affects the power consumption. Indeed, the system consumes the minimum power if all bandwidth is used, and the spectral efficiency Γ is equal across all beams⁶.

Pachler [188] has done work on developing a first-fit heuristic that aims to maximize the used bandwidth. Similar to Section 5.5 on the grouping of the user terminals, we summarize his work here for completeness and adjust the notation to match the one introduced in this dissertation (see Section 5.7.1). Pachler developed the approach for user beams \mathcal{B}_u only, but we will extend his algorithm to include gateway beams \mathcal{B}_g in Section 5.7.2. Therefore, we generally use the notation of \mathcal{B} for beams.

5.7.1 Frequency assignment by first-fit heuristic

Pachler's approach has three main steps: it takes a non-assigned beam, finds an assignment that respects constraints, and then assigns it to the beam. Three implemented algorithms direct the search for the next assignment:

1. **First-fit.** Assigns the first channel available and tries to reuse frequency if possible.
2. **Recoloring.** All beams are assigned the same channel in the beginning, and the first-fit tries to reassign the channels of colliding beams.
3. **Random assignment.** Tries to assign a random channel and reuse. If there is a collision, it tries again up to 100 times.

There are two constraints: the interference and the frequency reuse constraint. Pachler bases the interference constraint calculation on the angular separation angles β_{ij} between two beams i and j . Since the minimum distance between these two angles drives the interference constraint, the minimum angle over the satellite pass is taken: $\beta_{ij} = \min(\beta_{ijt} \forall t \in T)$. A constraint is when $\beta_{ij} \leq \epsilon$, where ϵ represents a minimum distance threshold, e.g., two times the cone angle $\epsilon = 2 \cdot \delta$. The reuse constraint ensures that

⁶ This is true since the marginal spectral efficiency with respect to power is monotonically decreasing.

no other beam uses the same frequency, resource group, and polarization on the same satellite at the same time.

Algorithm 5-4 provides an overview of the procedure for all of these three algorithms. The input is the union set of user and gateway beam \mathcal{B} . An allocation factor ρ scales the allocated bandwidth “chunks” based on an initialization (see Pachler et al. [188] for more details). If the algorithms cannot find a valid assignment, it lowers this parameter to find a solution. The output is a valid assignment set \mathcal{A} and not assigned beams \mathcal{B}_{NA} .

Algorithm 5-4: Pseudo code for the recoloring, first-fit heuristic, and random frequency assignment algorithm, adapted from Pachler [188]

| | |
|---|---|
| Input: beams \mathcal{B} , allocation factor ρ | // beams ordered descending by number of constraints |
| Output: assignment \mathcal{A} , not assigned beams \mathcal{B}_{NA} | |
| 1: $\mathcal{B}_{NA} = \{\}$ | // Init not assigned beams |
| 2: if recoloring then | |
| 3: for b_i in \mathcal{B} do | |
| 4: $\mathcal{A} = \mathcal{A} \cup \text{AssignFirst}(b_i)$ | // Assign the first free channel and satellite-group |
| 6: end for | |
| 7: end if | |
| 8: for b_i in \mathcal{B} do | |
| 9: $\mathcal{C}_b = \{\}$ | // Initialize empty constraint set |
| 10: for c_b in b_i do | |
| 11: $\mathcal{C}_b = \mathcal{C}_b \cup c_b$ | // Add constraint to set |
| 12: end for | |
| 13: $a_i = \text{AssignChannel}(b_i, \mathcal{C}_b, B_i \cdot \rho)$ | // Get assignment (first-fit, random) with \mathcal{C}_b & bandwidth $B_i \cdot \rho$ |
| 14: if a_i is empty then | // No assignment can be found |
| 15: $\mathcal{B}_{NA} = \mathcal{B}_{NA} \cup b_i$ | // Add b to non-assigned beams \mathcal{B}_{NA} |
| 16: else | |
| 17: $\mathcal{A} = \mathcal{A} \cup a_i$ | // Add assignment to set |
| 18: end if | |
| 19: end for | |

In the case of recoloring, the procedure assigns all beams \mathcal{B} the first available channel and satellite group. In all other cases, the assignment set \mathcal{A} remains empty. Then, in line 8, the algorithm iterates through all beams and first extract all relevant constraints for this particular beam and adds them to a temporary constraint set \mathcal{C}_b . Then, using this constraint set, a channel is assigned for beam b_i with the bandwidth $B_i \cdot \rho$. Pachler estimates the desired bandwidth B_i such that the satellites use all of the bandwidth. To achieve a fair allocation, i.e., similar spectral efficiency for all beams, B_i is chosen to be proportional to the forward demand in the beam D_{FWD, b_i} (see Pachler [188] for the details).

$$B_i \sim D_{FWD, b_i} \tag{5-23}$$

Since there might be constraints that prohibit the allocation of this desired bandwidth, the allocation factor ρ reduces this bandwidth until a valid assignment is found (for discrete bandwidth channels, ρ is discrete and has a lower bound). If the assignment is not successful and empty, the beam is added to the not assigned beam set \mathcal{B}_{NA} . Otherwise, the procedure adds valid assignments to the set \mathcal{A} . The implemented search procedure finds the highest allocation factor ρ for which all beams are assignable, i.e., $\mathcal{B}_{NA} = \emptyset$. The resulting solution satisfies all constraints and uses as much bandwidth as possible, given the implemented procedure.

Pachler applied the three algorithms to the solution he obtained with the algorithm described in Section 5.5 for the 18,712 user terminals and SpaceX’s Starlink (Appendix H contains the resulting plot). Figure 5-7 depicts the resulting computation time and unassigned beams. All three heuristics scale with the number of beams $\mathcal{O}(N_b)$, and the first-fit approach achieves better results (lower number of unassigned beams) in a shorter amount of time.

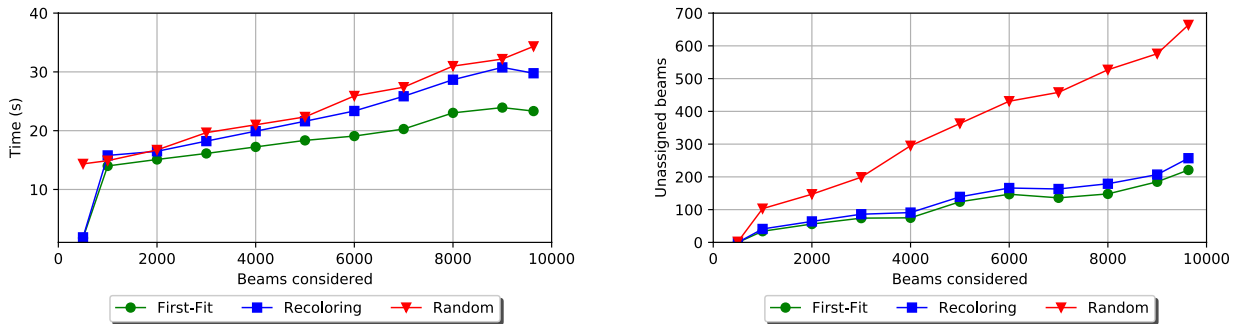


Figure 5-7: Solution comparing the three implemented heuristics for the frequency assignment by Pachler. Obtained from [188].

Limitations

We aim for maximum use of the bandwidth and equal distribution of spectral efficiency across all beams. While these principles guide the heuristic, it does not ensure optimality. Furthermore, the allocation factor affects globally all beams, which results in further suboptimalities. The result of the algorithm is a stationary plan, which means that the assignment is the same for all time steps. We see further research directions in two areas: (1) exploring other optimization techniques such as mathematical programming and ML-based techniques, and (2) considering a dynamic frequency assignment with changes during satellite-to-satellite handover or the same satellite mid-pass handovers.

5.7.2 Extension to include gateway beams

This extension is work done by the author and is necessary for all satcom systems where the frequency spectrum for gateway and user links is overlapping. In that case, the gateway beams \mathcal{B}_g cause interference with the user beams \mathcal{B}_u if they are not deconflicted and are within the minimum distance ϵ . Therefore, we apply Algorithm 5-4 on the whole set \mathcal{B} . The difference is in the estimation of the desired bandwidth B_i for the gateway beams. The relevant demand is D_{RTN,b_i} , as this is the return traffic transmitted through the downlink to the gateway and Eq. (5-23) becomes $B_i \sim D_{RTN,b_i}$. With that representation, we deconflict all downlinks from satellite to gateways and users. The following Section discusses considerations for the uplink.

Uplinks considerations

One uplink corresponds to each downlink. One is a forward uplink from the gateway to the satellite, denoted as $\mathcal{B}_{g,up}$. The other is a return uplink from the user terminal to the satellite $\mathcal{B}_{u,up}$. If both uplinks are close to each other, they can cause interference, and the resource allocation needs to deconflict them. Assuming that downlinks and uplinks have similar interference behavior, the solution of Algorithm 5-4 is valid for the user uplinks $\mathcal{B}_{u,up}$ if $D_{FWD,b_i} \geq D_{RTN,b_i} \forall b_i \in \mathcal{B}_u$. That implies that the bandwidth needed for the return link is always smaller than the one required for the forward downlink. Hence a deconflicted forward downlink entails a possible deconflicted solution for the return uplinks $\mathcal{B}_{u,up}$. For the gateway uplinks $\mathcal{B}_{g,up}$, the opposite must hold: $D_{FWD,b_i} \leq D_{RTN,b_i} \forall b_i \in \mathcal{B}_u$ with the same justification. Eqs. (3-3) and (3-4) summarize this.

$$\mathcal{B}_{u,up} \text{ valid} \leftrightarrow D_{FWD,b_i} \geq D_{RTN,b_i} \forall b_i \in \mathcal{B}_u \quad (5-24)$$

$$\mathcal{B}_{g,up} \text{ valid} \leftrightarrow D_{FWD,b_i} \leq D_{RTN,b_i} \forall b_i \in \mathcal{B}_u \quad (5-25)$$

We can see that both uplink beam sets $\mathcal{B}_{u,up}$ and $\mathcal{B}_{g,up}$ become valid if the equality $D_{FWD,b_i} = D_{RTN,b_i}$ holds. However, usually, the forward link carries more traffic than the return link (often a factor of 2 up to 5 in some cases). Therefore, the user uplinks are valid with the solution obtained by Algorithm 5-4, but not necessary for the gateway uplinks $\mathcal{B}_{g,up}$. A conservative assumption is to take the maximum traffic of the uplink/downlink combination and use this for the desired bandwidth B_i estimation, i.e., Eq. (5-23) becomes $B_i \sim \max(D_{FWD,b_i}, D_{RTN,b_i})$. While this ensures validity, it might be overly conservative and results in a solution that leaves valuable resources unused. Quantifying that suboptimality and updating Algorithm 5-4 to consider uplinks is an area of future research.

5.8 Power

The computation of the required power is the final step in the resource allocation process. The results from the previous steps set the parameters of the link budget class of the simulator (see Section 4.3). Since Eq. (4-8) is implicit in the antenna power P_{Tx} , researchers explored several metaheuristics, machine learning, and mathematical optimization techniques. These works have in common to set the power directly, compute the resulting data rate, compare it to the demand, and provide feedback to the algorithm. In much of the research, the resulting Pareto front is a trade-off space between power and unmet demand. We propose in this Section an approach that works differently. It starts with the desired data rate (which, e.g., minimizes unmet demand) and computes the required power.

First, we describe the implemented MODCOD selection method when the algorithm sets the power directly. We use a *guess-and-search* strategy that estimates the MODCOD by an approximation and then uses that as a starting point for a search (see Algorithm 5-5).

Algorithm 5-5: power guess and search procedure for computing the data rate based on power

| | |
|--|--|
| Input: $\mathcal{P}, \mathcal{B}, R, \mathcal{A}, \text{MODCODs}$ | // non-dominated MODCODs ordered |
| Output: \mathcal{R} | // set of data rates |
| 1: for b_i in \mathcal{B} do | // iterate through all beams (or vectorized) |
| 2: $\text{MODCOD} = \text{MODCODwithoutOBO}(p_i, b_i, r_i, a_i)$ | // guess the MODCOD |
| 3: $\gamma = \text{RunLink}(p_i, \text{MODCOD}, b_i, r_i, a_i)$ | // compute the margin with the guess MODCOD |
| 4: while γ is not <i>sufficient</i> do | // check of margin is sufficient |
| 5: $\text{MODCOD} = \text{NextLower}(\text{MODCOD}, \text{MODCODs})$ | // pick the next lower MODCOD from the list |
| 6: $\gamma = \text{RunLink}(p_i, \text{MODCOD}, b_i, r_i, a_i)$ | |
| 7: end while | |
| 8: $r_i = \text{ComputeDataRate}(p_i, \text{MODCOD}, b_i, r_i, a_i)$ | // compute the data rate with the right MODCOD |
| 9: end for | |

Going back to Section 4.3 and Eq. (4-8), the EIRP computation depends on the MODCOD through the OBO. Our approach for the estimation is to set the loss caused by OBO to zero. Since the loss is always negative, this guess is *optimistic*. We select the matching MODCOD for this point, run the link, including the OBO, and check if the link reaches the desired margin. If so, the algorithm terminates, if not, the search procedure starts. We start by selecting a *lower* MODCOD from the ordered, not-dominated list of MODCODs. We repeat that procedure until the link closes with a sufficient margin. On average, the search terminates after around one iteration, i.e., the MODCOD below the guessed, optimistic one. The average complexity, therefore, is $\sim \mathcal{O}(2)$, i.e., running the link twice. In contrast, a binary search without guessing has an average complexity of $\mathcal{O}(\log n_{\text{MODCODs}})$, with n_{MODCODs} being the number of MODCODs. While

the pseudo-code describes the iterative way with computing the link for each beam separately, it is easily vectorizable.

The second, direct approach, turns around the link budget and starts with a given data rate R . With a bandwidth B obtained by the frequency assignment step, and Eq. (4-17), the necessary spectral efficiency Γ is computed. With that, the algorithm selects the MODCOD and computes the needed power. This requires to run the link only once and hence has a complexity of $\mathcal{O}(1)$. Note that we assume here that the power for each link can be compute independently. If that is not the case, e.g., when beams share amplifiers or interference between beams is computed pair-wise, then there are additional convergence loops within the power computation.

It depends on the assumptions about the coupling of beams and the problem setup if the power optimization approach, or the direct power computation based on the data rate, is more desirable. For the remainder of this dissertation, we use the latter approach since we are interested in the power necessary for the requested data rate and desire zero unmet demand under nominal operation.

5.9 Application of the resource allocation process

The goal of this Section is to illustrate the application of the approach and algorithms described in the previous Sections 5.4 - 5.8. That includes all four steps of the resource allocation process: grouping of user terminals, routing, frequency assignment, and power allocation. We model a constellation of MEO satellites similar to SES's O3b mPower serving 5,000 worldwide distributed user terminals with 30 gateways (all out-of-country). First, we describe the simulation setup in the subsequent Section 5.9.1, the results follow in Section 5.9.2, and we summarize the conclusions in Section 5.9.3.

5.9.1 Simulation setup

The space segment of the simulation is a constellation of seven MEO satellites in an equatorial orbit with no eccentricity. Table 5-3 summarizes the assumptions. Each satellite visits the same location on earth every 6 hours, which equals an orbit with a semi-major axis of 14,446 km or an altitude of 8075 km for a 6371 km radius of the earth. We assume each satellite possesses a phased array and a digital payload, and power and bandwidth are available as a pool. The satellite generates a beam with a $\Theta_{3dB} = 0.7 \text{ deg}$ half cone angle and a peak gain of $G_{Tx,max} = 35 \text{ dB}$. The digital payload has 2 GHz of bandwidth discretized into 200 beamchannels, which is reused 20 times with two polarizations. Hence, the satellites are capable of generating up to 4000 beams with a bandwidth of 10 MHz.

Table 5-3: Overview of the constellation's parameters

| | Value |
|--|--|
| Number of satellites | 7 |
| Orbit | Equatorial without eccentricity at 8075 km |
| Repeating ground track | every 6 hours |
| Half-cone angle Θ_{3dB} | 0.7 deg |
| Peak antenna gain $G_{Tx,max}$ | 35 dB |
| Total available bandwidth | 2 GHz |
| Number of beamchannels | 200 with 10 MHz |
| Number of reuses | 20 including L/R polarization |
| Maximum number of beams | 4,000 |

For the traffic model, we use the one provided by SES. It has 18,712 user terminals and we used the traffic model in some of the earlier analysis. We select 5,000 distributed user terminals with the locations seen in Step 0 of Figure 5-10. The hot spots are in Central America, Brazil, Central Africa, Madagascar, India, and Indonesia. All user terminals combined have a forward CIR of 198 Gbps with a mean per terminal of 40 Mbps. The smallest terminal has a CIR of 5 Mbps and the largest 300 Mbps. The return CIR is 62 Gbps, so a ratio of 3.2:1. Figure 5-8 shows a histogram of the forward CIRs. Most of the users have a CIR under

100 Mbps with the highest concentration of around 20-30 Mbps. The model has seven segments: Aviation, Backhaul, Energy, Enterprise, Government, Maritime, and Trunk, with the relative fraction based on CIRs depicted in the right of Figure 5-8. Enterprise makes up almost 50% of the traffic with backhauling coming second. The other segments split the remaining 35% fairly evenly (except Energy, which is with under 3% the smallest segment).

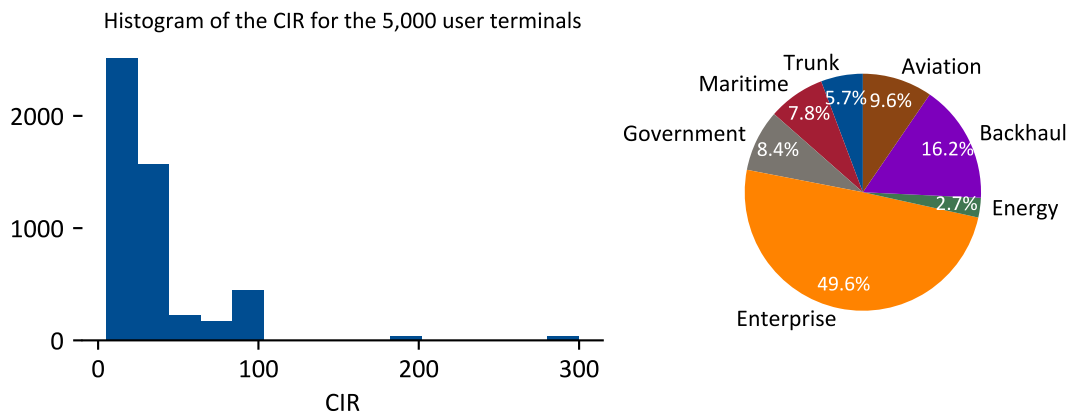


Figure 5-8: Statistics of the 5,000 user terminals. The left plot shows a histogram of the CIR and the right a distribution of the CIR amongst the seven segments.

Besides the CIR, the model has 24-hour traffic for each user terminal. Figure 5-9 depicts the traffic for one user from each segment. The diurnal usage pattern is clearly visible. Note that since we pick these randomly, no conclusions about the general behavior of the segments should be drawn.

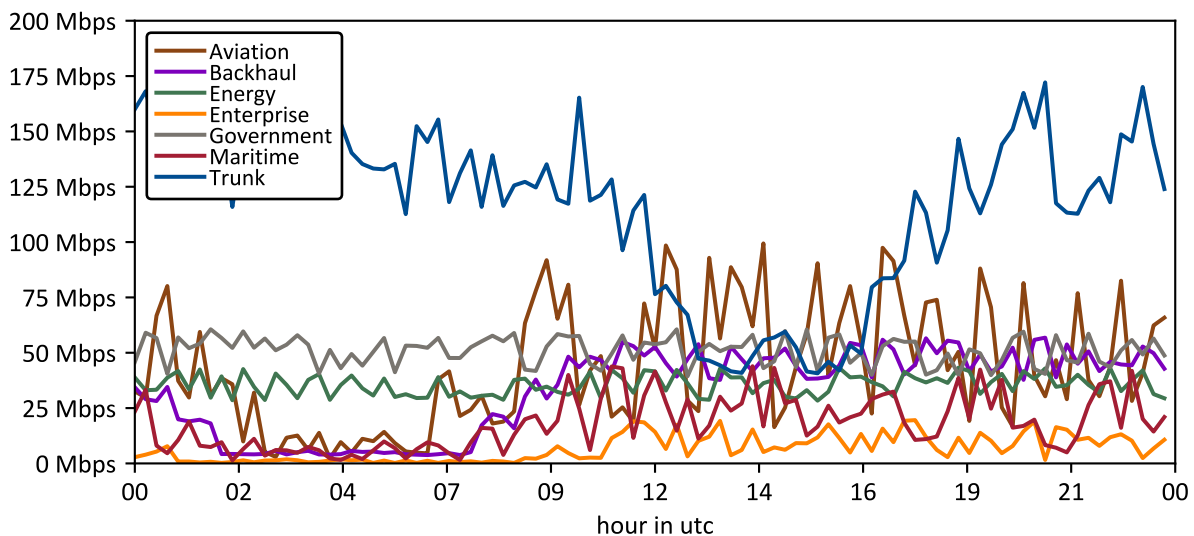


Figure 5-9: Sample of the diurnal demand pattern for the seven segments. Discretized into 15 minutes intervals.

Finally, we consider 30 worldwide gateways (see Figure 5-10, more evident in step 2 with the olive gateway beams). We assume that each gateway supports the whole range of the satellites' available

spectrum. The following Section discusses the results by applying the resource allocation process to this model.

5.9.2 Results and discussion

The results are the outcome of integrating the resource allocation algorithms with the satcom simulator from Chapter 4. Since the four steps in the resource allocation process are independent, we execute them sequentially. Figure 5-10 displays intermediary solutions after each step. The initial representation consists of seven MEO satellites, 30 out-of-country gateways, and 5,000 user terminals.

First, for the grouping of user terminals, we run the edge clique cover algorithm to find the minimum number of beams that cover the 5,000 user terminals. The procedure finds 1378 user beams. Step 1 shows the projection of these based on a mapping to the closest satellite. The most user terminals per beam are in the high-density areas of Central Africa and Indonesia.

Second, we apply the balancing routing algorithm that maps gateways to user terminals. The plot in Step 2 of Figure 5-10 shows the user beams in grey and the 1378 gateway beams in olive. The olive, dashed line between beam and satellite indicate that traffic flows through that gateway over the satellite. The darker the color of the gateway beam, the more beams point to that gateway. Seven terminals cannot be allocated to any gateways since the overlapping access windows are not sufficient. These are, as expected, in the higher latitude areas: Falkland Islands, United Kingdom, Ireland, and near the Bering Sea (see Step 1 of Figure 5-10). There are three potential options to cover these terminals nevertheless: relax the constraint on the minimum elevation angle of the terminals (currently at 10 deg), build gateways closer in the longitude to these locations, or move the traffic to another satellite or constellation. The balancing algorithm achieves to allocate all user terminals with an average return CIR traffic of 2.1 Gbps, a minimum and maximum of 0.4 Gbps, and 2.8 Gbps, respectively. On average, 45 return beams land on a gateway (minimum is 7, and the and maximum is 86).

Third, we compute the frequency assignment. We use the first-fit heuristic and consider an interference threshold of two times the full-cone angle, i.e., we assume they do not interfere if one beam fits between two beams. Step 3 in Figure 5-10 shows the resulting coloring of the beams, and Figure 5-11 plots the current assignment to the satellites. The edge color of the beam indicates the polarization (white for left and black for right polarization). The fill color is the frequency spectrum based on an HSV colormap. The gateway beams in Figure 5-11 are olive independent of their polarization or frequency. The printed solution is for the time instance of $t = 0$.

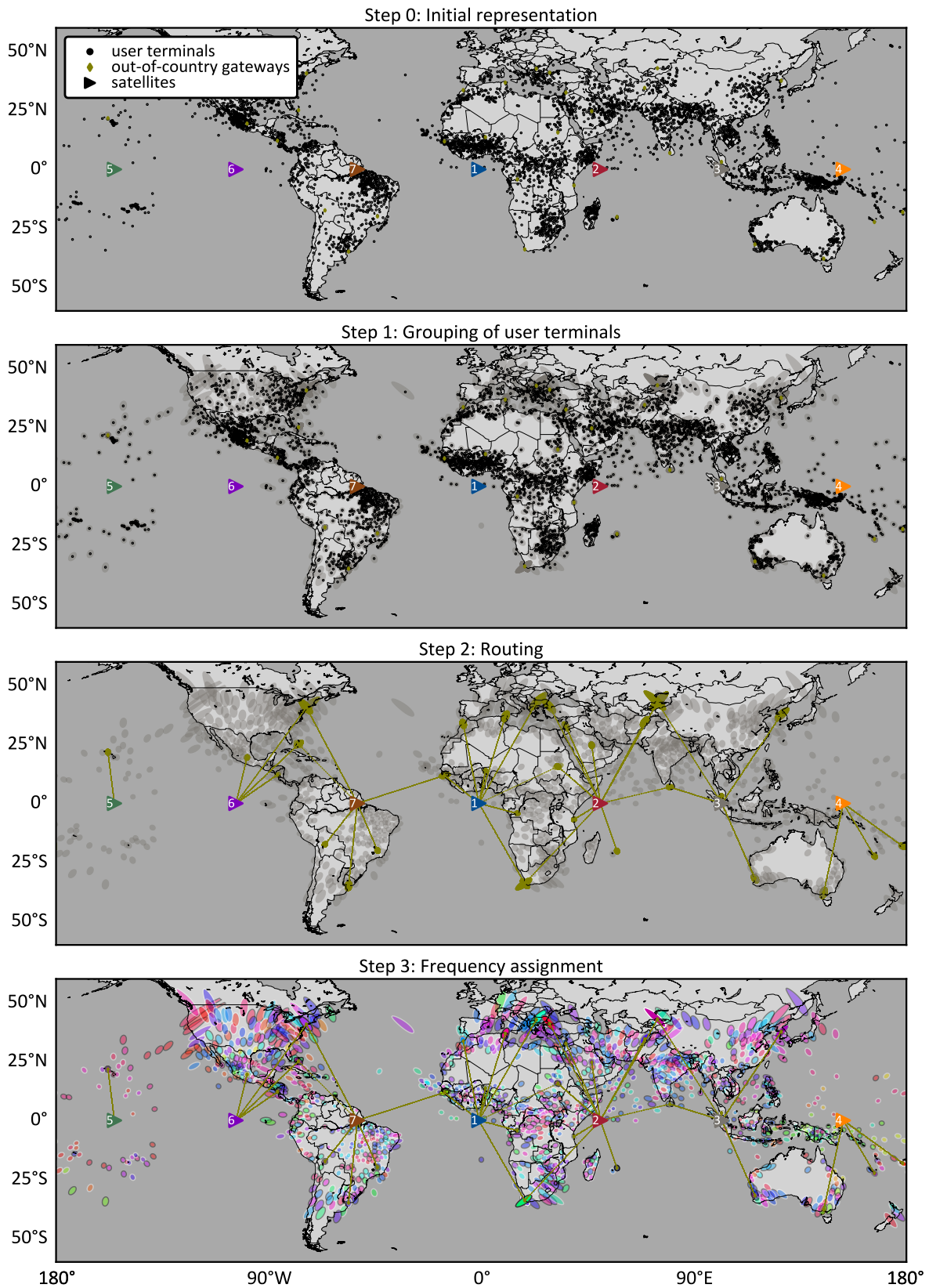


Figure 5-10: Final result plot of steps 1-3 for a constellation similar to O3b mPower with 5,000 user terminals at t_0

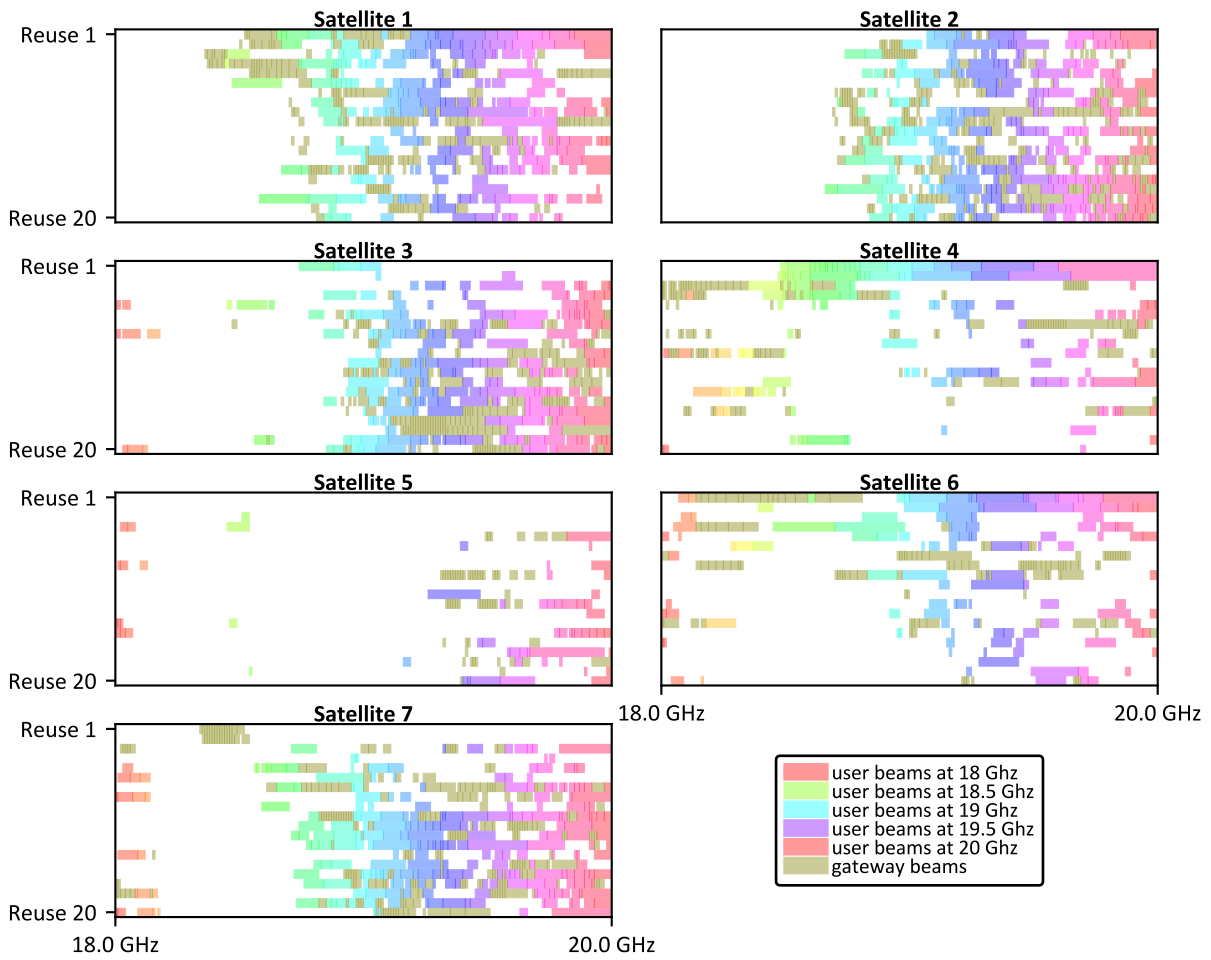


Figure 5-11: Frequency allocation for the first time step. Each block represents one beam (gateways are brown).

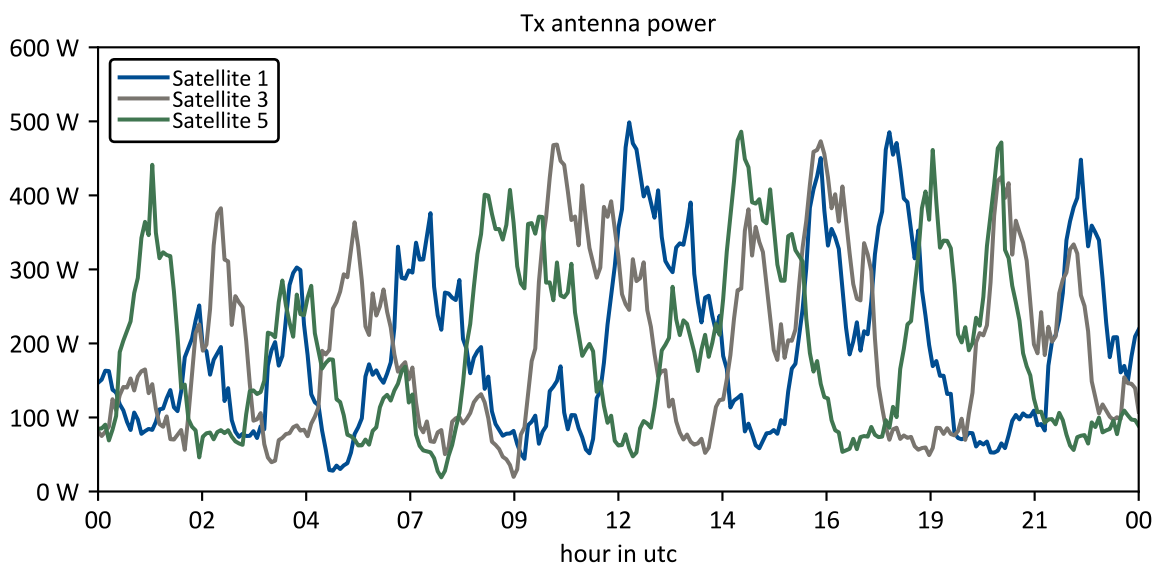


Figure 5-12: Aggregated power on satellite level over the 24h simulation period for the user beams.

Since the used frequency allocation algorithm produces a stationary solution, the frequency and bandwidth of the user terminals do not change over time. What changes is the mapping to the satellite, i.e., the chunks of Figure 5-11 are moving throughout time. The satellites are propagating into the right direction of Figure 5-10, and hence the beams are handed over from right to left satellites. To ensure continuous connectivity, the satellite hands over user and gateway beams at the same time (see Appendix E for the same plots at $t = 3$ hours).

We earlier discussed that the optimal frequency assignment is the one that uses all of the available bandwidth and the spectral efficiency is homogenous across all beams. The first objective translates to seven filled satellites in Figure 5-11 (if there are no constraints). Indeed, the current solution only utilizes, on average, 22% of the available bandwidth (8.8 GHz out of 40 GHz). That is in part because of the low traffic over the Pacific (see satellite 5). However, even for the higher demand region of the current location of satellite 1, the utilization only reaches 29%. The other part is due the suboptimality of the heuristic and the hundreds of thousands of constraints between beams, such as interference and reuse-constraints (for further details on the constraints refer to Pachler [188]). These constraints limit the choices the algorithm has for placing the chunks. It then attempts to reduce the size of the chunks with the expectation that they fit more easily. When the algorithm needs to reduce the size of the chunks many times, the utilization of the bandwidth becomes lower, as seen by this example. Given the complexity and dimensionality of the problem, the best solution is remarkably challenging to find. Therefore, it is unclear how far away from the optimum the presented solution is.

In this example, the user and gateway downlinks use the same spectrum, and therefore, highly utilized gateways constrain the local area. In the extreme case, where all the gateways use the complete spectrum, the constellation cannot serve any user terminal that is closer than the interference threshold. This observation couples the routing and the frequency assignment steps. The link is in contrast to our specific objective to separate the resource allocations steps and desires further attention in future research. We defined the objectives of the routing first to balance the load between gateways and second to minimize the distance between user terminals and gateways. That approach does not consider that the frequency assignment benefits from low utilized gateways in dense areas (if user and gateway downlinks share the same spectrum). That highlights one of the many trade-offs in the resource allocation process. The solutions are particularly problem-specific.

Finally, as the fourth step, we compute the RF antenna power of the user beams using the direct approach from Section 5.8. We simulate 24 hours with 5-minute discretization and report the aggregated power in

Figure 5-12 (we only show satellites 1, 3, and 5 to avoid cluttering). Each line represents one satellite, and the power is in Watts. We observe a diurnal variation even though the satellite is serving global demand. The reason is that the world has more demand between -90 deg and +90 deg longitude than on the other half of the globe (mostly the Pacific). Hence, the peak hours are in the afternoon and evening UTC, as Figure 5-12 confirms. The average power consumption per satellite is 194 W (see Table 5-4). The power is fluctuating between 19 W, and over 500 W. The low is during the satellite pass over the Pacific during the night, and the high is during the day over Europe/Africa. The fluctuations are a combination of three effects at different timescales (see also Figure 5-13 for the fluctuations in demand and provisioned data rate). First, the slant range varies throughout a satellite pass (51.4 minutes plus handover), causing different free space losses for which the satellite has to compensate with power. Second, the satellite is connected to different sets of user terminals along its 6-hour orbit until it revisits the same earth location. Third, the users' demand-usage behavior has a distinct 24-hour pattern. All three effects repeat after 24 hours, and therefore, the power consumption of Figure 5-12 repeats until a new service starts or one terminates (assuming deterministic demand). The implications are that each satellite has a slightly different average power consumption throughout its life (see Table 5-4).

Table 5-4: Average power consumption of the seven satellite. The percentage numbers are the relative deviation from the mean over all satellites.

| | Average power [W] | Minimum power [W] | Maximum power [W] |
|--------------------|------------------------------|------------------------------|------------------------------|
| Satellite 1 | 193.96 (-0.1%) | 28.08 | 498.64 |
| Satellite 2 | 193.10 (-0.5%) | 33.20 | 500.32 |
| Satellite 3 | 196.70 (1.3%) | 19.75 | 473.12 |
| Satellite 4 | 195.28 (0.6%) | 21.39 | 499.06 |
| Satellite 5 | 193.20 (-0.5%) | 19.09 | 486.07 |
| Satellite 6 | 192.53 (-0.8%) | 18.12 | 486.08 |
| Satellite 7 | 193.91 (-0.1%) | 22.50 | 503.94 |
| mean | 194.10 | | |

While the differences for this example are below 1% for most satellites, the operator might want to strategically position the satellites based on the results from the acceptance tests. Satellite 3 needs to provide the highest average power, while it has the lowest maximum power. Around 5% higher peak is required from satellite 7 with 504 W.

An additional result that we want to discuss is the quality of the power allocation. Figure 5-13 shows how the provided data rate follows the requested demand. The provided data rate is always above the requested demand, resulting in zero unmet demand. The amount of overprovisioned data rate is

unavoidable due to the discretization of the MODCODs. With the approach from Section 5.8, we ensure that the algorithm always meets the demand except for the cases where the spectral efficiency of the highest available MODCOD is insufficient. The algorithm minimizes power consumption by choosing the lowest MODCOD that has enough spectral efficiency to achieve the data rate with the available bandwidth.

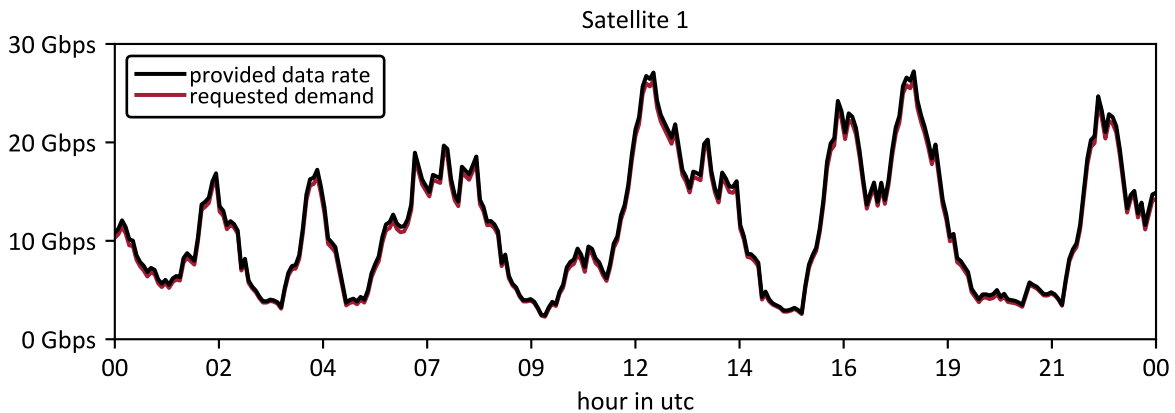


Figure 5-13: Aggregated data rate and requested demand of satellite 1 over the 24-hours simulation period. The satellite meets the demand at all times and only overprovides a negligible data rate.

We illustrate the dynamics of the MODCOD selection in Figure 5-14. From the 5,000 user terminals, we pick a backhauling customer on the Canary Islands with a CIR of 100. The vertical axis is ordered based on spectral efficiency from low to high, and each jump in the plot is the re-selection of a new MODCOD. Again, the night drop is apparent where the usage drops down to 7 Mbps, and a BPSK4/15 provides with 0.24 enough spectral efficiency. During the day, the customer peaks up to 86 Mbps for which the algorithm chooses a 16APSK4/5 (spectral efficiency of 2.9). The fastest changes are every 5 minutes (the discretization of the simulation).

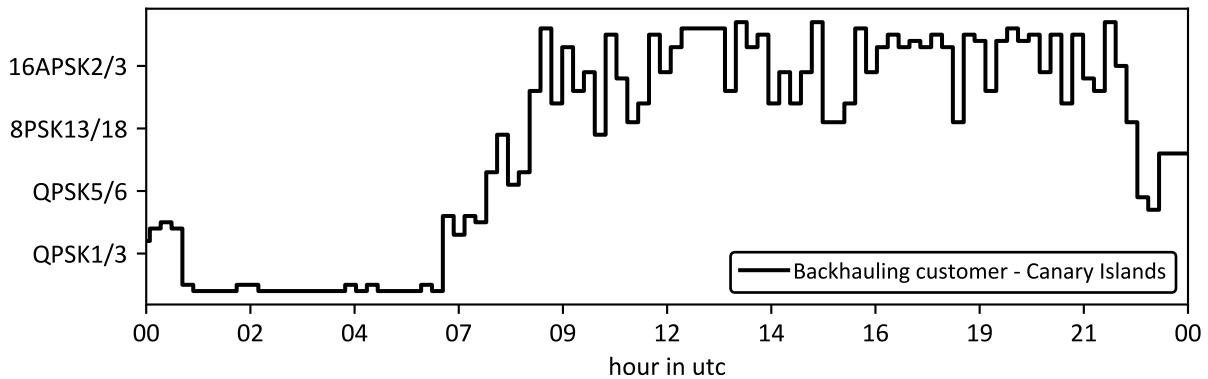


Figure 5-14: Selected user downlink MODCODs for a customer on the Canary Islands at around 28 deg latitude and -16 deg longitude. The customer has a 1.2m dish with a G/T of 21 dB and a 100 Mbps CIR SLA.

5.9.3 Conclusions and insights

In summary, we applied the resource allocation process to an MEO constellation with seven satellites serving 5,000 worldwide user terminals through 30 gateways. The algorithm covered these terminals with 1,378 user beams resulting in 2,756 beams in total. The frequency assignment found a plan using, on average, 22% of the available bandwidth. The power allocation resulted in average consumption of 194 W per satellite. We discovered the following insights:

- When a satellite is sharing the spectrum for user and gateway downlinks, there is an additional trade-off between the balancing of gateways and the blocking of frequency for gateways in high-density regions. The frequency assignment favors for the routing to shift the traffic from gateways in dense regions to isolated gateways.
- In the described scenario, the satellites do not use the majority of the bandwidth resources. It is part of future research to investigate if this low usage is improvable by a more optimal frequency assignment algorithm or if it is the nature of the problem. If the latter, constraint relaxation strategies could be explored.
- The power allocation algorithm from Section 5.8 minimizes power consumption and overprovisioning while ensuring to meet the demand at all times.
- The ground repeating pattern with four visits per day results in a slightly different average and peak power for each satellite. The operator might wish to position the satellites accordingly based on the results of the acceptance tests.

5.10 Summary and contributions

The purpose of this Chapter was to elaborate on the resource allocation component as the most vital piece of the satcom RM framework. We reviewed relevant literature regarding resource allocation in satcom and noticed that existing work only addresses subproblems. Therefore, we formalized the complete resource allocation process and described a coherent solution process consisting of four steps. While applying the solution process, we examined the impact of the decisions made in each subproblem in an end-to-end fashion. For two of steps of the resource allocation process, we draw from work done by Pachler [188] under the supervision of the author, and for the other two steps, we develop algorithms. We then apply the solution process to a constellation similar to SES's O3b mPower.

Our contributions are:

- Formalized and decomposed the resource allocation process into four sub-problems.
- Extend the frequency plan algorithms from Pachler et al. [188] to include gateways.
- Developed a closest-first and balanced gateway allocation algorithm for the routing sub-problem.
- Developed a direct RF power allocation strategy.
- Demonstrated the functioning of the resource allocation process with an application to a seven satellite MEO constellation and 5,000 user terminals.

6

Available Capacity Forecaster

We dedicate this Chapter to the challenge that available capacity is uncertain based on resource usage. The component *available capacity forecaster* encapsulates the solution. It is a statistical forecaster that considers customer usage history for predicting power usage at the satellite level (see Figure 6-1). In contrast to deterministic forecasting, we consider uncertainties, leading to a more accurate picture of the available capacity. The available capacity is proportional to the additionally sellable Mbps and hence revenues. Underestimation leads to lost revenue opportunity, and overestimation results in contention.

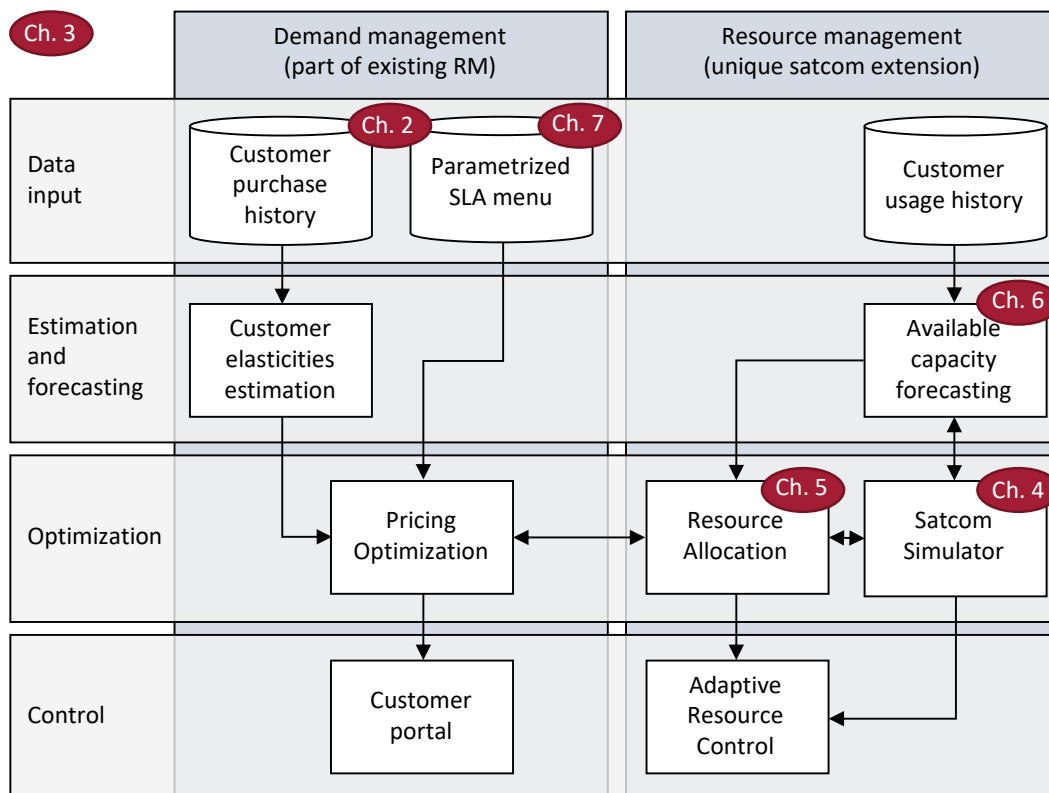


Figure 6-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4

Furthermore, the available capacity forecaster extrapolates the usage history of existing customers to new customers. That enables simulating various combinations of customers to fill the capacity. The pricing optimization triggers that simulation capability. Since the forecaster translates the Mbps on customer level to power at the satellite level, it queries the satcom simulator and, if necessary, the resource allocation.

The available capacity forecaster follows a two-step approach. First, it builds a stochastic regression for each customer's usage history. Second, the forecaster samples from the regression and computes the power consumption on the satellites for each combination of samples. The following two Sections, 6.1 and 6.2, cover the two steps. Section 6.3 concludes the Chapter.

6.1 Probabilistic regression of user traffic

The database stores the user traffic as a time-series of forward and return Mbps. The granularity varies depending on the technical capabilities of the platforms, and possibly, practical data volume limitations of the database. The recorded data is on a per-user-terminal basis and requires preprocessing, e.g., cleaning invalid measurements and adjusting for seasonality. Jones [190] offers a more detailed discussion on preprocessing and short-term traffic estimation for satcom user terminals. His focus is the forecasting of a single point per user terminal with a defined lead time. Jones uses 15-minutes head time as a reference and compares Long short-term memory (LSTM) and XG-boosted trees with a baseline that uses the last known value as the forecast.

While Jones' work on traffic forecasting is fundamental for the adaptive resource control component (it allows for active instead of reactive control), the objectives of the available capacity forecaster are different. It requires knowledge about the user's *typical day* traffic profile, *including probabilistic density*, i.e. an uncertainty distribution around each point forecast over time.

We choose a day as the desired timeframe, as it is the smallest common denominator of all dynamics' frequencies (in fact for MEO, such as O3b and mPower, the dynamics repeat every 6 hours, while the demand is mainly changing in a 24-hour cycle). Since the simulation is triggered by the pricing optimization multiple times, keeping the computational execution time of the simulation as short as possible makes the problem computationally tractable. Therefore, we desire a probabilistic forecast of a typical day for each user that condenses the historical traffic data.

A *typical day* is a forecast 1-day ahead that reflects the uncertainties of the past usage. For example, consider a probabilistic density forecast for 4 pm for the typical day. The mean is the average of all

historical measurements for 4 pm and assumes that the future behavior is similar to the past. The uncertainty around that mean then expresses the variation around that mean of all historical data points. If we now take the 99% quantile, 99% of historical measurements at 4 pm were within that value. We assume that future demand behaves similarly.

We choose to represent this by a stochastic process, i.e., a collection of the random variables \mathbf{X}_t (see (6-1)).

$$\{\mathbf{X}_t\}_{t \in \{1 \dots T\}} \tag{6-1}$$

The objective is to estimate the stochastic process given the observations, which we denote by the matrix M (which is nothing more than a stacked time series, where each row is a new day). It has the dimension $N \times T$ where N is the number of observations, i.e., the number of days, and T is the 24h time window with the desired granularity.

$$M = \begin{bmatrix} r_{11} & \dots & r_{1t} \\ \vdots & \ddots & \vdots \\ r_{n1} & \dots & r_{nt} \end{bmatrix} \quad \forall n \in \{1 \dots N\}, \forall t \in \{1 \dots T\} \tag{6-2}$$

Before we discuss the algorithms, we review the relevant literature in the following Section.

6.1.1 Literature review

The general literature on time-series forecasting is vast. Most work focuses on point forecasting with little literature existing on density methods. A detailed review of it is outside the scope of this dissertation. Instead, we focus on density methods for the electricity industry as justified below. Three starting points on general forecasting for the interested readers are: (1) a 2006 review article of the last 25 years on time-series forecasting from De Gooijer et al. [191], (2) a comprehensive review book about forecasting methods by Chatfield [192], and (3) the International Journal of Forecasting.

To the best knowledge of the author, there is no industry-specific literature on traffic estimation in satcom on a per users basis besides Jones’s work [190]. Hence, we review literature in other industries that show similar characteristics and draw analogies back to satcom. Many industries use time-series forecasting broadly, e.g., finance, stock markets, hydrology, and electrical consumption [193]. In particular, the latter industry has the following similarities with satcom:

Electrical consumption users are private households as well as small and larger cooperation. All users exhibit a varying amplitude of diurnal variations with uncertainties. The energy infrastructure is built for many years with a relatively fixed total capacity. On one hand, short-term forecasting is vital for robust

control of the energy grid. On the other hand, long-term forecast informs the capacity expansions strategy to minimize unmet demand and overcapacity.

Motivated by these similarities, we choose to review the electricity industry in more detail in the following subsection.

Electricity load forecasting

Accurate load forecasting is essential for the planning and operation of the electric power industry. Research and industry practice has evolved for over more than 100 years [194]. The main focus was and still is point forecasting but has expanded to density forecasting in the last years. Hong et al. [194] provide a tutorial review of the load forecasting research. They and Weron et al. [195] note that probabilistic forecasting is a small research field compared to point forecasting, but argue that the supply planning can benefit from a density-based approach given increased competition (note the analogy to satcom).

Arora et al. [196] leverage the data generated by smart meters per individual site (similar to traffic per user terminal in the satcom context). They aim to generate a density estimate for each smart meter and use this for dynamic pricing. For the forecast, Arora et al. use kernel density and conditional kernel density estimations with a decay parameter. Half-hourly data is available for which forecasts with lead times from a half-hour to a week are generated. Figure 6-2 shows a density forecast for a residential consumer of a kernel density with two intraday cycles and exponential smoothing (we adopt the figure from Arora et al. [196], Fig. 6 on P.18). The Mean Absolute Error (MEA) of the point forecast ranges between 0.02 and 0.08. While the diurnal pattern is identifiable, this forecasting method is not necessarily a regression of a typical day.

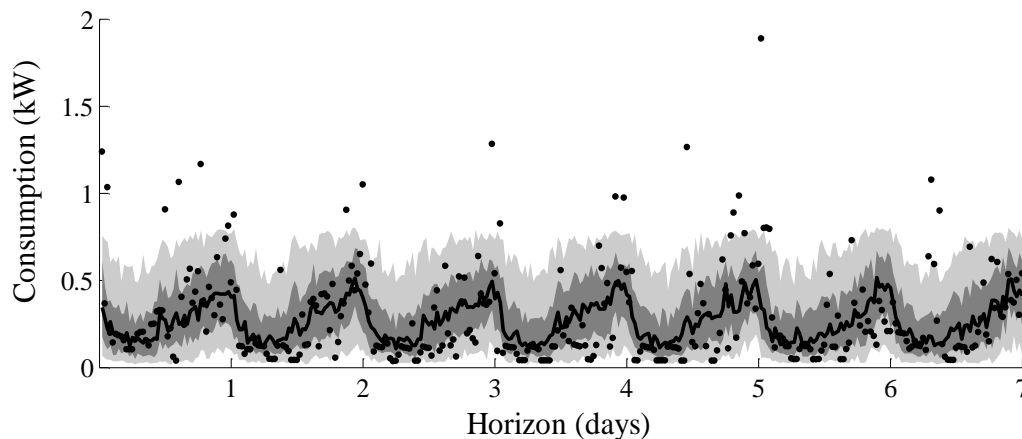


Figure 6-2: Density forecasts for a residential consumer. Adopted from [196], P.18. Black dots are observations, black line is forecast, shading is uncertainty distribution in forecast.

Another widely explored area is Gaussian process regression. Sheng et al. [197] propose a method that deals with the effect of outliers on the fitting (e.g., the series of black points well above the fit in Figure 6-2). In contrast to other studies, they integrate the outlier detection with the regression by weighting outliers lower. The authors refer to their method as a weighted Gaussian Process Regression.

Bonilla et al. [198] first introduce multi-task learning for Gaussian process prediction. Zhang et al. [199] show that leveraging the correlation between nearby cities, they can increase prediction accuracy. The authors propose a new algorithm that makes the multi-task Gaussian process computationally more tractable compared to the proposed algorithm from Bonilla et al. [198]. While Zhan et al. [199] focused on the short-term power forecast for a small number of cities, Fiot et al. [200] investigate mid-term predictions at the smart meter level.

Yang et al. [201] note that many sources of uncertainties exist in power load predictions, and their objective is to quantify that uncertainty in the forecasting. Therefore, the authors propose probabilistic forecasting using a Gaussian process quantile regression. Yang et al. claim that the advantage of their approach is that it is not parametric compared to kernel methods and, therefore, can model “arbitrarily complex systems given enough data” [201, p. 500].

The discussed works on Gaussian processes all predict multiple time horizons instead of focusing on a typical day estimation. Gaussian process regression is only a sub-area of stochastic density forecasting, and multiple other algorithms exist. While we believe some can be transformed into a “typical day” regression, the detailed study is outside the scope of this dissertation. Since the quality of the forecast directly influences the available capacity and hence revenues, we encourage future research in this area. Nevertheless, we discuss two approaches that we implement and use for the analyses in Section 3.6 and Chapter 8. We start with a Gaussian process regression in Section 6.1.2 and then introduce a collection of normal random variables in Section 6.1.3.

6.1.2 Gaussian process regression

We have shown in Figure 3-2 from Section 3.2 a Gaussian process regression of real data to illustrate the diurnal variation of user traffic (Figure 6-3 shows the same plot for convenience). The Gaussian process is a stochastic process where every finite collection of this process has a multivariate normal distribution [202]. The prediction is not only the mean but also the density with the shape of a normal distribution. A regression analysis fits a kernel to the observed data point. We then use the kernel for predicting.

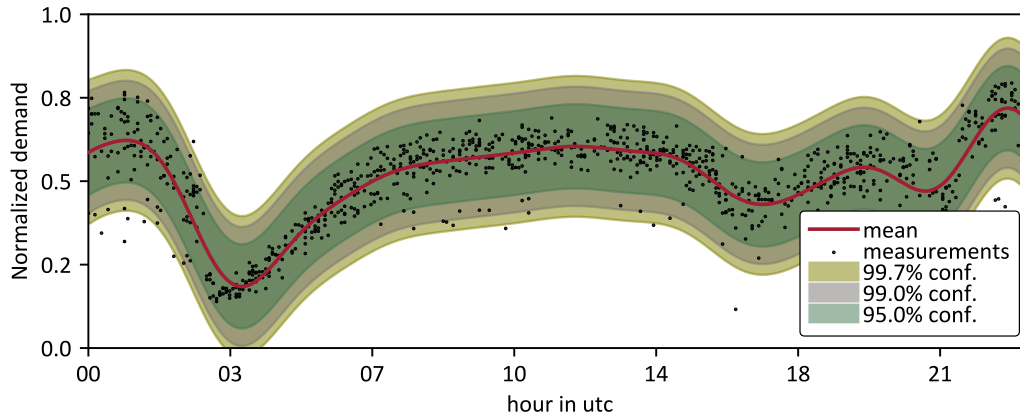


Figure 6-3: Gaussian process example of 1,000 samples of one-month traffic data fitted by a Gaussian process using a radial basis and white noise kernel. Copied from Figure 3-2.

For the example in Figure 6-3, we took 1,000 samples of one month of traffic and used a kernel composed of the sum of a constant, and the product of an Radial basis function (RBF) and a white noise kernel. We use the sklearn toolkit for fitting, which uses Cholesky factorization from Algorithm 2.1, Page 19 of Rasmussen and Williams [202].

While the result is a smooth fit, it is computationally expensive, especially for multiple thousands of samples. Additionally, the density is constant throughout time, and therefore the confidence intervals are more a global measure than a local. For example, in Figure 6-3, the uncertainty at Hour 3-4 is lower than at Hour 1-2. The Gaussian process regression does not differentiate and tries to find a global best fit for the uncertainty, i.e., the hyperparameters of the white noise kernel.

However, for our purpose, we are interested in the uncertainty at every time instance. Therefore, we propose a different approach described in the following Section.

6.1.3 Stochastic process of normal random variables

As the second approach, we use a stochastic process with a collection of independent normal random variables (see Eq. (6-3)). Figure 3-15 and Figure 8-4 illustrate examples.

$$\{\mathbf{X}_t\}_{t \in T} \quad \text{with} \quad \mathbf{X}_t \sim N(\mu(t), \sigma(t)^2) \quad (6-3)$$

This approach has a few practical advantages:

- The independent discretization allows us to simulate multiple realizations during each time step of the simulation. For our satcom simulator, this is computationally orders of magnitude more efficient compared to simulating all time-instances for one realization at a time.

- The input data for the analysis in Chapter 8 is available as discretized means and standard deviations. From that format, we can create the collection of independent normals directly.
- The regression is robust, and we can compute essential parameters, such as means and confidence intervals explicitly.

Because of these factors, we use this approach throughout the thesis for user traffic regression. In the next Section, we present the approach of using the regressions at the user level in Mbps and translating them to the power level at the satellite.

6.2 Determine power density at satellite level

The goal of this Section is to determine the power density for each satellite throughout a typical day based on the traffic density estimation from the previous Section. Given the complexity and non-linearity of the simulator, deriving the solution is not feasible analytically. Therefore, we fall back to a sample-based approach.

In that, we sample the stochastic process $N_{samples}$ times. The satcom simulator translates from Mbps into W for each sampling point and user. In all simulations this dissertation presents, changing the power is sufficient to accommodate the diurnal variation of the user, i.e., the grouping of users and frequency assignment is stationary. When the flexibility of the latter two steps is required, the resource allocation process is called in addition to the satcom simulator to translate the samples from Mbps to W.

We record the samples in the power dimension and fit them with a stochastic process consisting of an independent collection of *empirical* random variables (denoted with $\{Y_t\}_{t \in T}$). We are interested in computing the used capacity $C_{used}(t)$ at any time instance, such that the probability $Y_t < C_{used}(t)$ is larger than the contracted availability A . The following equation formalizes this:

$$\mathbb{P}(Y_t < C_{used}(t)) \geq A. \quad (6-4)$$

We solve this equation for $C_{used}(t)$ by inverting the empirical cumulative distribution function of Y_t . With that, we compute the whole-day available capacity (further defined in 7.2.1) with the maximum capacity C_{max} :

$$C_{whole-day} = C_{max} - \max(C_{used}(t)) \quad (6-5)$$

We illustrate the sensitivity of the used capacity and the whole-day available capacity on the contracted availability A in Figure 6-4. The two capacity levels $\max(C_{used})$ and $C_{whole-day}$ are recorded for A varying between 80% and 99.9%. As expected, the available capacity declines with increasing A . It becomes highly non-linear above 98%. Going from 98% to 99%, $C_{whole-day}$ drops from 39.4% to 35.7%. And further, from 99% to 99.9% available capacity decreases from 35.7% to 28.5%.

The non-linearity has significant implications for the SLA contracting: providing a service with promised availability of 99.9% allows the operator to reallocate 28.5% capacity, whereas, for $A = 98\%$ ⁷, they can

⁷ These availabilities do not include other probabilities of failure (gateway failure, satellite outage, ...). They must be considered when using this approach and stating availabilities to the customers.

reallocate 39.4%, i.e., 11% more. We use for the remainder of this dissertation $A = 99\%$ but acknowledge the sensitivity of the available capacity to this number.

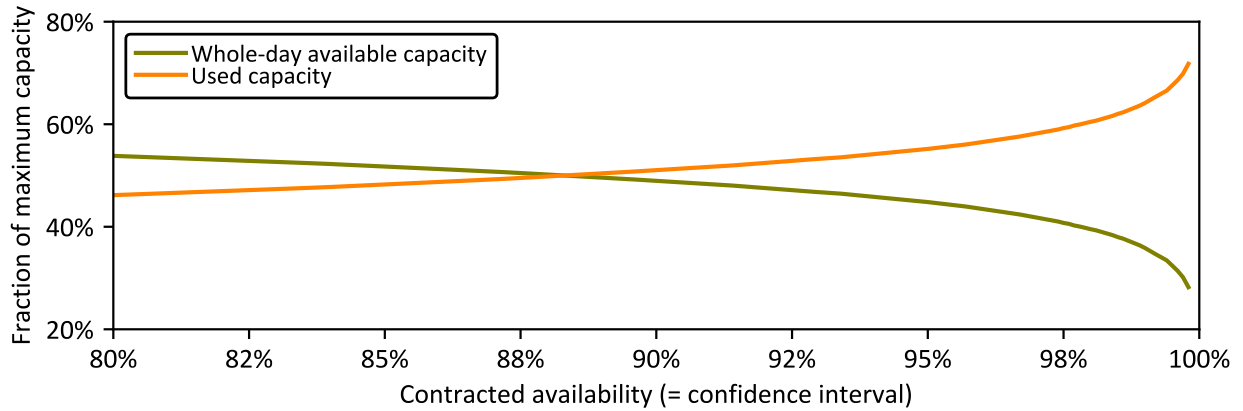


Figure 6-4: Example of the impact of the contracted availability A on the used and the available capacity as a fraction of the maximum capacity.

Convergence

We use the example from Section 3.6 to develop an understanding of the convergence concerning $N_{samples}$. Figure 6-5 shows the absolute error relative to the results for 3,000 samples. For sample sizes below 250, the error is around 10% and from then on decreasing. After 1,000 samples, the results are converging with errors below 1-2%, and we use this number for all computations following.

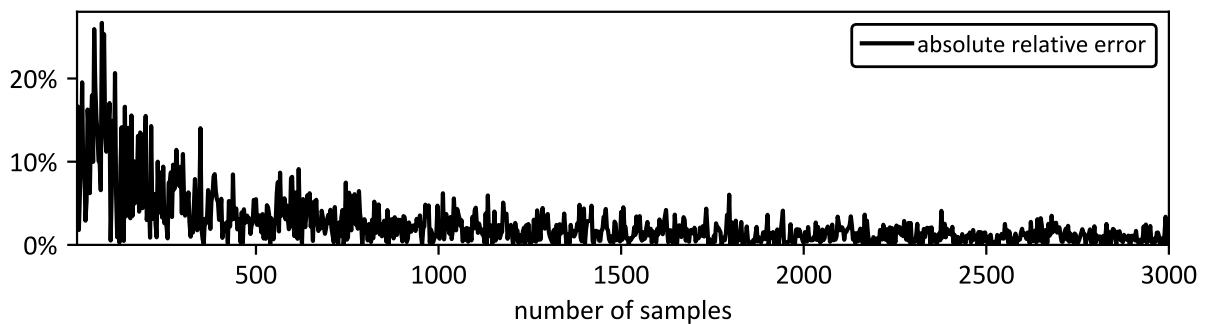


Figure 6-5: Convergence analysis of the absolute relative error as a function of the number of samples $N_{samples}$

Note that for some figures that show the density of power on the satellite level, we use 50,000 samples to have the confidence intervals visually sharp with less noise (e.g., Figure 3-16). Due to the vectorization of the link budget as described in Section 4.3, the computational expenses (in the order of a second) only increase slightly with the number of samples.

6.3 Summary and contributions

In this Chapter, we discussed our approach to the available capacity forecaster that considers the uncertainty introduced by users' traffic patterns. Our proposed method consists of two steps:

First, a probabilistic regression provides a forecast of the density of a typical day on the user level. In that, we reviewed selected algorithms used in electricity load forecasting. Based on that, we studied two approaches: Gaussian process regression and collection of independent normal random variables. The latter has practical advantages that make it our preferred approach. As we just touched the surface of this area, we see various future research opportunities.

The second step is to determine the power density at the satellite level by sampling the regression at the user level. We showed that after 1,000 samples, the results show convergence. Furthermore, we investigated the non-linear impact of the contracted availability on the available capacity. In the example, a 98% availability results in 38% more available capacity than 99.9% availability.

We made the following contributions:

- Identified the need for a *typical day* forecasting for satcom RM
- Drew the analogy between electricity load and satcom user traffic forecasting
- Proposed a computationally inexpensive method: a collection of independent normal random variables with several practical advantages
- Quantified the sensitivity of the available capacity to the SLAs' contracted availability

7

Novel SLAs

The objective of this Chapter is to review existing SLAs, identified their limitations, and explore new SLAs that leverage the new satellites' flexibilities and create a win-win situation for customers and operators. The result of this chapter is a set of parametrized SLAs, which provide the input to the framework's pricing optimization (see Figure 7-1).

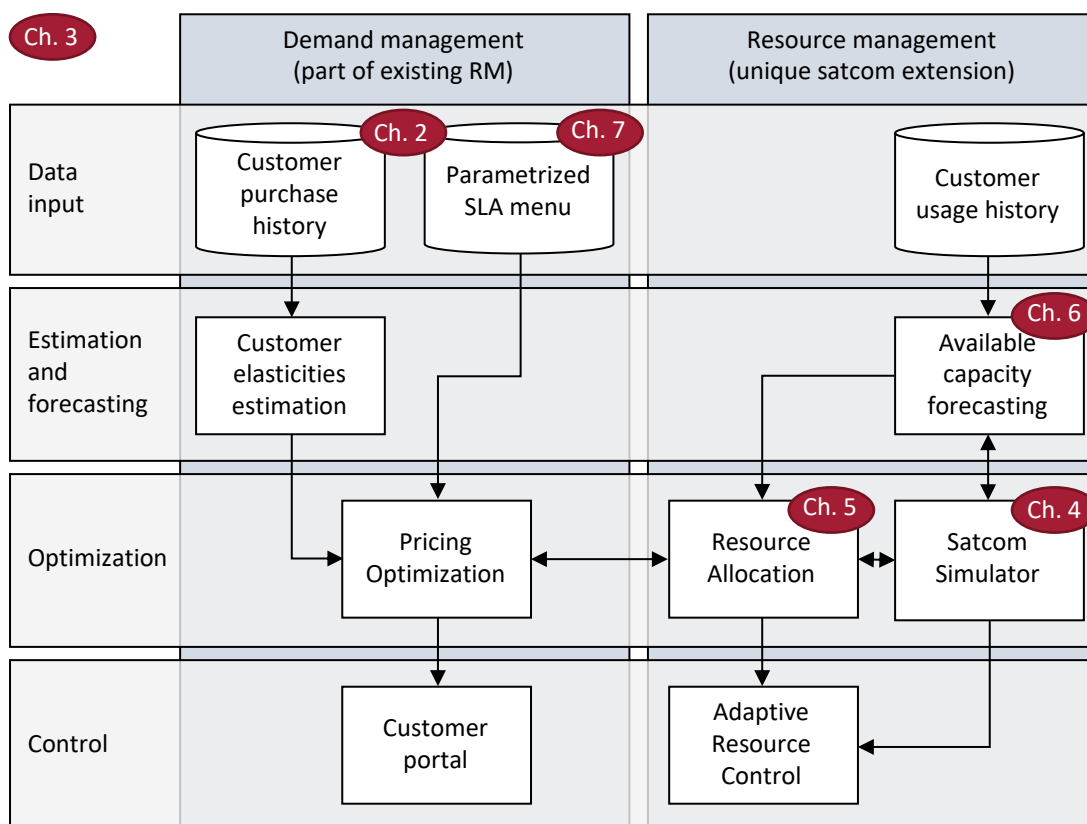


Figure 7-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4

We use a methodology based on Crawley's et al. [203] translation of needs into goals. The authors decompose needs into served and latent needs (a further discussion of latent needs can be found here

[204-206]). In addition to this differentiation, we define the third category: *observed, but unserved needs* (see Figure 7-2). The rationale behind this is to capture the needs that are observable but not yet served (in comparison to latent needs, which are unknown). Once a latent need is identified, it transitions to the observed but unserved category. Existing products (we call them here classical SLAs) fulfill the served needs of the two considered stakeholder groups: customers and operators. Our goal is to come up with novel SLAs that target the observed but unserved-needs category. Flexibility in the products (enabled through RM and flexible payloads) allows adjusting to latent needs when they surface.

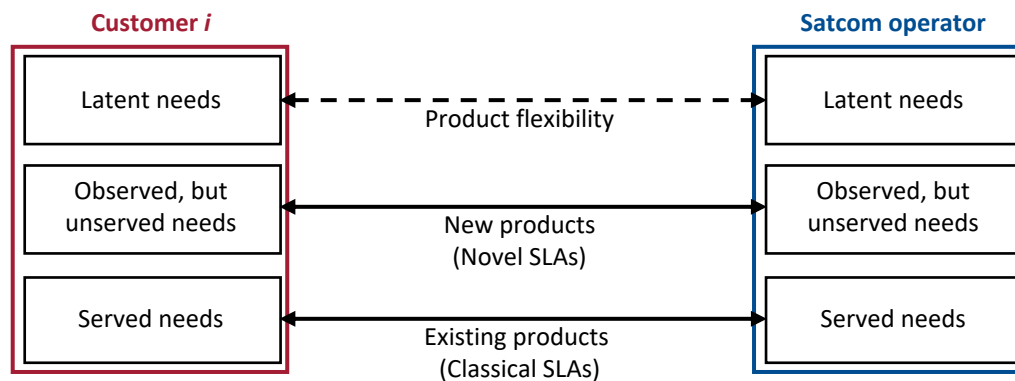


Figure 7-2: Overview of the decomposing of needs into served, observed but unserved, and latent needs.

The source of information for this Chapter are company homepages, satellite magazines, and several interviews with professionals from four companies who wish to remain anonymous. While an exhaustive customer need analysis is out of the scope of this dissertation, we focus our investigation on SLAs that fit well into the context of RM.

The research in this Chapter is mainly qualitative. A quantitative analysis of the value of novel SLAs is partly done in Section 3.6 and mainly in Chapter 8. Section 7.1 reviews the state-of-practice of SLAs in satcom. The challenges of these SLAs are outlined in Section 7.2 and the resulting specific objectives are presented in Section 7.3. The subsequent Section 7.4 reviews SLAs in other industries from which we draw analogies to inform the design of novel SLAs in Section 7.5. Since not all SLAs are equally suited for each segment, we discuss customer segmentation in the following Section 7.6 and map it to the proposed SLAs. We finish the Chapter by summarizing the contributions in Section 7.7.

7.1 Review of current SLAs in satcom

As discussed, we divide the broadband satcom industry into traditional (selling MHz) and new (selling managed Mbps) satcom. In the first case, the demarcation point for operators is at the satellite, and it is up to the customer to transform the MHz into Mbps. For managed Mbps, the operator's demarcation

point extends to the user terminal – this includes a promised uptime for the Mbps and not just the satellite’s MHz. Traditionally the MHz are commonly sold to service providers, who then sell managed Mbps and value-added services to the customers or end-users. Some operators, e.g., SES with the legacy O3b constellation, skip the service provider layer and go directly to the customer offering Mbps services. We observe that the general SLA structure is the same for sales between operator and service provider, service provider and customers, and operator and customers. Therefore, this review applies to all three levels.

The most common SLA is what we refer to with *Classical SLA*, see Box 7-1. The main parameter of this product is either the CIR or a bandwidth B accompanied by a promised availability (mainly used for managed Mbps services)⁸. The monthly recurring revenues are in \$/Month. This price is generally higher for a higher promised availability. If the operator violates the SLA, the customer receives a credit. Under this contract, provisioning a data rate above CIR or provisioning more bandwidth B is not tied to additional revenues.

Box 7-1: Classical SLA, the most commonly used product for both MHz and Mbps satcom

Classical SLA

- Traffic parameters
 - CIR/bandwidth with promised availability
- Pricing
 - Monthly price for the CIR: p_{CIR} , or bandwidth B : p_{MHz}
- Monthly revenues: $\Pi = p_{CIR} \cdot CIR$ or $\Pi = p_{MHz} \cdot B$

Additionally, we find that for some Mbps based CIR contracts, there are three additional parameters:

- Peak/Maximum Information Rate (PIR/MIR): usually, the maximum the link can deliver under “best” conditions, but not tied to any monetary compensation for the Classical SLA as defined above.
- Data volume per month for a given maximum speed
- The monthly price for the data volume (fixed and not pay-as-you-go)

We purposely do not include the PIR/MIR in the Classical SLA, as this parameter does not impact revenues directly. The pricing depends on the CIR, and additional PIR/MIR is courtesy of the operator (however, this is an integral part of negotiations, as well as customer acquisition and retention). With the flexibility of

⁸ Note that we primarily focus on the forward link as it usually dominates, but all the parameters can also be defined for the return link or FWD/RTN ratios may be used.

reallocating regional capacity dynamically, it becomes challenging to define the PIR/MIR accordingly. However, it is a meaningful limit for the data volume SLA. In this case, the volume is directly tied to pricing, so that we separate this aspect into a *Data volume SLA* summarized in Box 7-2. The parameters are the monthly volume and the PIR. The pricing is mainly driven by the volume allowance with a secondary effect from the peak rate PIR.

Box 7-2: Data volume SLA with revenues as a function of the monthly volume and corresponding price

Data volume SLA

- Traffic parameters
 - Monthly GB allowance: V_{month}
 - Maximum speed: PIR
- Pricing
 - Monthly price per month for fixed volume V_{month} : p_{month}
- Monthly revenues: $\Pi = p_{month} \cdot V_{month}$

While Classical and Data volume SLAs are bought separately, they can also be combined into what we call a Dual SLA. Box 7-3 summarizes the key characteristics. Mainly, service providers offer this Dual SLA product. For example, KVH Industries offers Dual-channel plans of an unlimited low-speed and a limited volume high-speed part [207], Speedcast has Ku-band VSAT plans [208], and Marlink advertises VSAT Sealink [209] (further details in Appendix F).

Box 7-3: Dual SLA combining the Classical and the Data volume SLA

Dual SLA

- Traffic parameters
 - CIR with promised availability
 - Monthly GB allowance: V_{month}
 - Maximum speed: PIR
- Pricing
 - Monthly price for the CIR p_{CIR} , plus
 - Monthly price per month for the volume V_{month} : p_{volume}
- Monthly revenues: $\Pi = p_{CIR} \cdot CIR + p_{month} \cdot V_{month}$

Table 7-1 shows a further illustration of three discussed SLA types (Mbps based). We assume for the calculations a monthly price of \$100/Mbps/Month and \$5/GB for a PIR of 200 Mbps [66]. Note that a constant use of 1 Mbps equals 324 GB/month. We set the parameters for all three types so that the monthly recurring revenues are \$3000/month. It follows a CIR of 30 Mbps for the Classical SLA, an allowance of 600 GB for the Data volume SLA, and for example, 20 Mbps and 200 GB for the Dual SLA.

Table 7-1: Example contracts of the three existing SLA types

| | CIR (Mbps) | V_{month} (GB) | PIR (Mbps) | p_{CIR} (\$/Mbps/month) | p_{volume} (\$/GB/month) | Π_{total} (\$/month) |
|------------------------|-----------------|---------------------|-----------------|------------------------------|-------------------------------|-----------------------------|
| Classical SLA | 30 | | | \$100 | | \$3,000 |
| Data volume SLA | | 600 | 200 | | \$5 | \$3,000 |
| Dual SLA | 20 | 200 | 200 | \$100 | \$5 | \$3,000 |

Based on this example, the Classical SLA is best suited for customers who use their CIR most of the time. In the extreme case where they use their CIR all the time, their monthly recurring cost would be $30 \cdot 324 \text{ GB/month} \cdot \$5/\text{GB} = \$48,600/\text{month}$ when using a Data volume SLA. The data volume SLA is best suited for customers who have unpredictable, variable traffic and require a high-speed connection (though the speed is contractually not ensured). Ensuring this high-speed with a Classical SLA would cost them $200 \text{ Mbps} \cdot \$100/\text{Mbps}/\text{Month} = \$20,000/\text{month}$. The Dual SLA is for customers with requirements in-between.

The parameters of these three classic SLAs are commonly defined and fixed for the timespan of the contract, typically 1-3 years⁹. These contracts were developed during a time where the resource allocation of satellites was fixed by design. As of the writing of this dissertation, almost all current satellites fall into this category. Operators contract their capacity through one of these three SLAs¹⁰.

However, with modern flexible payloads and increased automation, resources can be reallocated more dynamically. Therefore, the question arises if classical SLAs are still adequate in this new environment or if operators and customers miss opportunities. The following Section discusses the challenges before we translate them into the specific objectives of this Chapter.

⁹ The contract might define time periods for re-negotiation of the prices, which is desired by the customers due to declining market pricing as illustrated in Chapter 2.

¹⁰ Due to the b2b nature of satcom with a few large customers, the actual SLA is often further customized around the basic SLA.

7.2 Challenges

Since we scope the RM system for satcom operators, we specifically discuss here the contracting of capacity between operators and their customers (which can be service providers or end-users). We explore the operator and customer perspective separately in Sections 7.2.1 - 7.2.2 and then derive the specific objectives in Section 7.3.

7.2.1 Operator perspective

We highlighted some of the limitations of classical SLAs in Section 3.6 when illustrating the working principles of the RM. We saw that customers' usage behavior leaves significant capacity unused. Figure 7-3 shows a conceptual plot based on the data from the example's Figure 3-16.

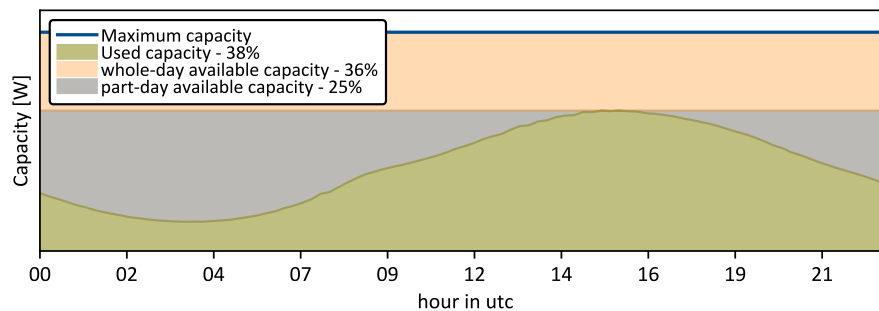


Figure 7-3: Conceptual view of the used, available whole-day and part-day capacity. Numbers are based on Figure 3-16 from the example in Section 3.6.

We separate three areas: the used capacity in olive, an available whole-day capacity in orange, and available part-day capacity in grey. The blue line gives the maximum capacity. The whole-day capacity can be sold throughout the entire day, while the available part-day capacity is a function of the time of day. For the example from Section 3.6 (which is representative for a GEO satellite), over the 24h period, the used capacity is 38%, with 36% being available as whole-day, and 25% as part-day. Note that for a NGSO constellation (as simulated in Chapter 8), the part-day available capacity has a lower percentage fraction of somewhere in the 10% area.

We assume here that all of the capacity is contracted through Classical SLAs with CIRs. Hence, considering CIRs, the operator's satellite would be 100% utilized (this is how we define the maximum capacity line). However, the actual usage behavior and power multiplexing yield to a considerably lower effective utilization (38% in this example). From an operator perspective, they would like to increase that effective utilization.

One possibility is to sell the available whole-day capacity under Classical SLA terms. That would result in an *overbooking* when considering CIRs, which is generally not an issue as long as the customers' daily

usage pattern follows the operator's forecast, and the diurnal pattern is constant during the contract duration. However, when the behavior changes, the operator risks that the used capacity outgrows the maximum capacity, i.e., the olive area goes above the blue line in Figure 7-3. This situation is not desirable as SLAs are violated and cannot be resolved until the multiple years SLAs terminate. While some operators might make a conscious decision to take some risk and overbook their system with Classical SLAs, we refer to this practice as *unsafe overbooking*.

While overbooking reduces the available whole-day capacity, the 26% part-day capacity remains unutilized (assuming the added customers show similar daily behavior). Its percentage fraction will increase with decreased whole-day capacity. At the extreme of no whole-day capacity, the olive and the grey area will go all the way up to the blue capacity limit (see Figure 7-3), and the part-day capacity fraction becomes $26/(36 + 26) = 42\%$. The specific shape and extent of the night-drop depend mainly on two factors: (1) the longitudinal distance between terminals connected to the satellite, and (2) the hour of the night-drop of the connected customers.

The orbit of the satellite gives the first factor. GEO satellites can reach up to 163° of the earth's longitudes (assuming 0° elevation angle). With equally distributed time zones, this would be a maximum time difference of 11 hours between the outermost user terminals. With an 8000 km MEO orbit, the covered longitudes decrease to 127° and 8 hours, and for a 1000 km LEO orbit, there is an upper limit of 60° and 4 hours. If all customers have the same night-drop, optimal cancellation of peak and off-peak demand occurs for a shift of 12 hours. In particular, NGSO satellites see smaller regional areas with smaller differences in the time zone.

The second factor is the hour of the night-drop of the connected customers. Like in many industries, usage is higher during the day than during the night-time. Depending on the segment, the extent of the night-drop varies, and the peak shifts. For example, business customers have peak hours between 9 am - 5 pm, whereas residential customers peak during the evenings. However, most of the traffic still follows the general day-night pattern. The combination of this and the first factor leads to an overall *ill-timed daily pattern*, as seen by the olive area in Figure 7-3.

To summarize, operators wish to sell the whole-day available capacity through *safe overbooking* and reduce the part-time available capacity by smoothening the aggregated daily pattern. In the following, we discuss the perspective of the operators' counterparts, the customers.

7.2.2 Customer perspective

For the analysis of the customer perspective, we rely on transcripts from over 60 interviews of worldwide customers across various segments made available by SES. We consolidate this information and identify the following five needs ordered by priority:

1. **Lower prices.** To be competitive with terrestrial alternatives.
2. **Better availability of more flexible capacity.** Being able to offer end-users more service types.
3. **Lower latency.** Desired, but does not limit growth as end-users developed techniques to deal with higher latency.
4. **Lower barrier to entry.** Mainly driven by cost for terminals.
5. **Better customer support.**

While all of these needs are key, we focus on the first two most critical points.

Satellite communication is often a last resort solution as the prices are (still) not competitive with terrestrial alternatives. As we saw in the market dynamics Chapter 2, prices have dropped over the last few years and are likely to decline further. This decay is welcomed by the customers as it reduces their cost for communication. Higher utilization of the satellite reduces the per Mbps cost for the operator. Assuming that the operator passes on the cost reduction to the customer by lowering prices, customers have an indirect incentive to support the operator in achieving a high utilization.

The second point is the better availability of more flexible capacity. Customers wish to provide their customers/end-users with more creative offerings to differentiate themselves, which require more flexible capacity and SLAs. This need substantiates flexible payloads and shorter contract duration, as also identified as beneficial by the market dynamics Chapter 2. Other interviews confirm that inflexible capacity is challenging for customers. It is difficult (or even impossible) for customers to predict accurately 1-3 years ahead. If they overestimate demand, they have unused capacity. If they underestimate demand, they missed revenue opportunities. This is particularly important for customers, which are not end-users but service providers.

Overall, customers would like to have low prices and more flexible capacity. We combine these observations with the ones from the operator perspective and define the specific objectives in the following Section.

7.3 Specific objectives

Recognizing the challenges discussed, our specific objective for this Chapter is to come up with a set of novel SLAs that have the following main characteristics:

- Allow the operator to overbook their capacity safely
- Smoothen out the aggregated diurnal variation
- Lower prices, i.e., help the operator to increase their utilization
- Allow flexibility in the capacity

Our approach for architecting suitable SLAs is to review similar industries from which we draw analogies. Therefore, in the following Section 7.4, we review the cloud service industry and the telecommunication industry. Using these insights, we develop a set of novel SLAs in the Section after.

7.4 Review of SLAs in similar industries

We find that the cloud service and telecommunication industries offer the most insights into novel SLA types. They are both service-based, have inflexible capacities, and a large variety of customers ranging from single end-users to large resellers.

Cloud services

Amazon Web Services (AWS) published their menu for their four Amazon's Elastic Compute Cloud (EC2) products [90] consisting of four product categories:

1. **On-Demand.** Customers pay per hour, there are no commitments, and the prices are the highest. As an example, a1-medium for \$0.0255/h (taken June 2019 for East Coast US).
2. **Spot instance.** Spot prices posted by AWS based on demand vs. capacity. AWS can quit the instances anytime. Lowest price, e.g. a1-medium: \$0.0049/h (19% of on-demand price -> 81% discount)
3. **Reserved instances.** Reserve the whole capacity, not shared. Standard contract for 1yr: 40% discount, for 3yr: 60% discount of from on-demand price.
4. **Dedicated hosts.** Dedicated physical infrastructure with pricing not directly available.

If we match this product menu to the classical SLAs that we described in the previous Section 7.1, we discover that the *Reserved instance* and *Classical SLA* are most similar. The dedicated host is more similar to buying a complete transponder on the satellite, which we interpret as a MHz-based Classical SLA for as long as the duration of the satellite's lifetime. We could not identify an AWS product that is similar to the

Data volume-based, i.e., a computational volume. On the flip side, in satcom, there is no *on-demand* and *spot-instance* product, making them exciting candidates for novel SLAs.

Telecommunications

We refer to Courcoubetis [121] as he provides a summary of possible service contracts (= SLAs) for telecommunications. The two general categories are:

- **Guaranteed Service.** Service for which there is a contract between provider and customer. The provider agrees to provide a service as long as the customer's traffic satisfies certain constraints. They come with a QoS guarantee, which implies that some resources must be reserved.
- **Best-effort Service.** The network tries to provide the best quality to each of its customers without guarantees.

Notice that the *Guaranteed Service* comes closest to the *Classical SLA* that we described above. The *best-effort service* maps to the PIR/MIR aspect as these are not guaranteed, but customers can burst up to them if capacity is available. Courcoubetis [121] further provides three examples of how these two types can be combined:

- **Time-of-day pricing.** A regulated (e.g., through fair use policies) best-effort service with no strict guarantee. Network sets the prices to utilize the available capacity fully.
- **Combined Guaranteed and Best-effort Service.** Priority traffic receives a guarantee, and if traffic exceeds the guaranteed traffic, it becomes best-effort.
- **Minimum Guarantees and Uncertainty.** Customer from a defined pool share capacity in a defined manner. Each customer gets a minimum guarantee, and the additionally available throughput for each customer depends on what the other customers use. That creates a negative externality effect. In satcom, the minimum guarantee would be the CIR, and the additionally available throughput is the PIR/MIR with the same negative externality effect.

The time-of-day pricing tries to shape the demand by communicating to the customer through prices that the opportunity cost is high of providing service during certain times. We expect from price-sensitive customers to adapt their usage behavior and shift their demand to less expensive hours (e.g., a private person might schedule to back up their laptop during the night instead of the evening). The result is less variation in the aggregated, used capacity curve. The takeaway from the last two points is that not all traffic is the same. The time-critical voice might have the highest priority, whereas a delayed cloud synchronization might not deteriorate the quality of experience of the end-user.

We will combine the insights from the review with the desired characteristics defined in the specific objective into a set of novel SLAs described in the following Section.

7.5 Novel SLAs

We propose three novel SLAs: *spot instance*, *time-of-day pricing*, and *Two Classes of Service*. The first SLA is a shorter-term contract for customers who already have a user terminal. The other two are longer-term contracts specifically focused on smoothening out the diurnal variation. Table 7-2 maps the three novel SLAs to the four desired characteristics. The checkmark in parentheses between the two-classes of service and the overbooking indicates a fit that depends on the details.

Table 7-2: Overview of the novel SLAs and their mapping to the desired characteristics

| | | Shorter-term | Longer-term | |
|-------------------------|--|---------------|---------------------|------------------------|
| | | Spot instance | Time-of-day pricing | Two Classes of Service |
| Desired characteristics | Allow the operator to safely overbook | ✓ | | ✓ |
| | Smoothen out the diurnal variation | | ✓ | ✓ |
| | Lower prices | ✓ | ✓ | ✓ |
| | Allow for flexibility in capacity | ✓ | | |
| Technical requirements | Enabled by dynamic resource management | ✓ | | |
| | Requires deep-package inspection | | | ✓ |

Furthermore, we map the technical requirements to the novel SLAs. Due to the more frequent contracting through spot instances, the capacity allocation should be automated through a dynamic resource management system.

Implementing a time-of-day pricing SLA does not require dynamic satellite technology per se. Given pricing that reassembles the revenues of the Classical SLA, its main benefit to the operator is the smoothening of the diurnal variation and therefore, “translation” of the part-day into whole-day capacity (see more details in the following). How much this is a concern for the operator depends on geographical user distribution and usage behavior, satellite technology, and resource management capabilities. For example, an MEO constellation with large batteries can smoothen out the variation on the satellite side. In contrast, a GEO satellite with limited flexibility and a smaller FOV benefits more from that SLA (comparing Figure 3-16 with Figure 8-9).

For the two classes of service, all points for the time-of-day pricing apply with the addition that the modems must be able to perform deep package inspections.

In the following three Sections, we discuss each of the newly introduced SLAs in more detail.

Spot instance

The spot instance SLA is for customers who already have a user terminal. That might be existing customers who want additional capacity, occasional use customers, or customers served by another operator but need a shorter-term capacity increase. Either side can terminate the spot instance at any time. This achieves a win-win situation for both sides. The customers have increased flexibility, and operators can safely overbook the satellite: if the usage behavior of the Classical SLA customers changes, the operator can adjust to these changes by quitting spot instances. The parameters of the *spot instance* are similar to the Classical SLA see Box 7-4.

Box 7-4: Spot instance SLA without a minimum or maximum contract duration

Spot instance SLA

- Traffic parameters
 - CIR with promised availability
- Pricing
 - Monthly price for the CIR p_{spot}
- Monthly revenues: $\Pi = p_{spot} \cdot CIR$

Time-of-day pricing

We design the spot instance to allow a safe overbooking for the operator and flexible access to capacity for the customers. The time-of-day pricing SLA aims to affect the demand pattern of customers such that on an aggregated satellite level, the part-time available capacity is reduced and transformed into available whole-day capacity. The main idea is to use pricing as a way to communicate resource scarcity to customers. In general, prices are higher if less capacity is available (peak hours) and prices are lower when capacity is in excess (night-drop). There can be a multitude of prices per day, e.g., day/night, hourly (see Δt_{chunk} in Box 7-5).

Box 7-5: Time-of-day pricing SLA

Time-of-day pricing SLA

- Traffic parameters
 - CIR with promised availability for each discretization of the day t
- Pricing
 - Monthly price for the CIR at time t with the price p_t
- Monthly revenues: $\Pi = \sum_t p_t \cdot CIR_t$

The exact price for each chunk is an optimization problem that considers price sensitivities, capacity costs, and customer relationships. There are two main effects: (1) price sensitivity customers might try to shift traffic to cheaper hours of the day, and (2) price-insensitive customers (or for the traffic that cannot be

shifted) might accept the higher price resulting in a revenue increase (or they switch to another operator that offers cheaper prices). Potential mitigation of the negative aspect of the second effect can be to only consider price discounts.

Two Classes of Service

In the Two Classes-of-Service SLA, modems separate the traffic on the user side into different categories by deep package inspection [210]. One example that we propose here is splitting into *real-time* and *not real-time*. The top plot of Figure 7-4 schematically introduces that. It is a cosine traffic pattern for 48 hours, and we assume the traffic is 50% real-time. The blue line illustrates the CIR for a Classical SLA if the two traffic types are not split. Since we assume the traffic follows a cosine behavior and it is split equally, a delay of the traffic by 12 hours results in a perfect cancelation of the daily pattern (see the bottom plot of Figure 7-4). If that is possible, the operator can transform the available part-time capacity into a capacity that is available for the whole day.

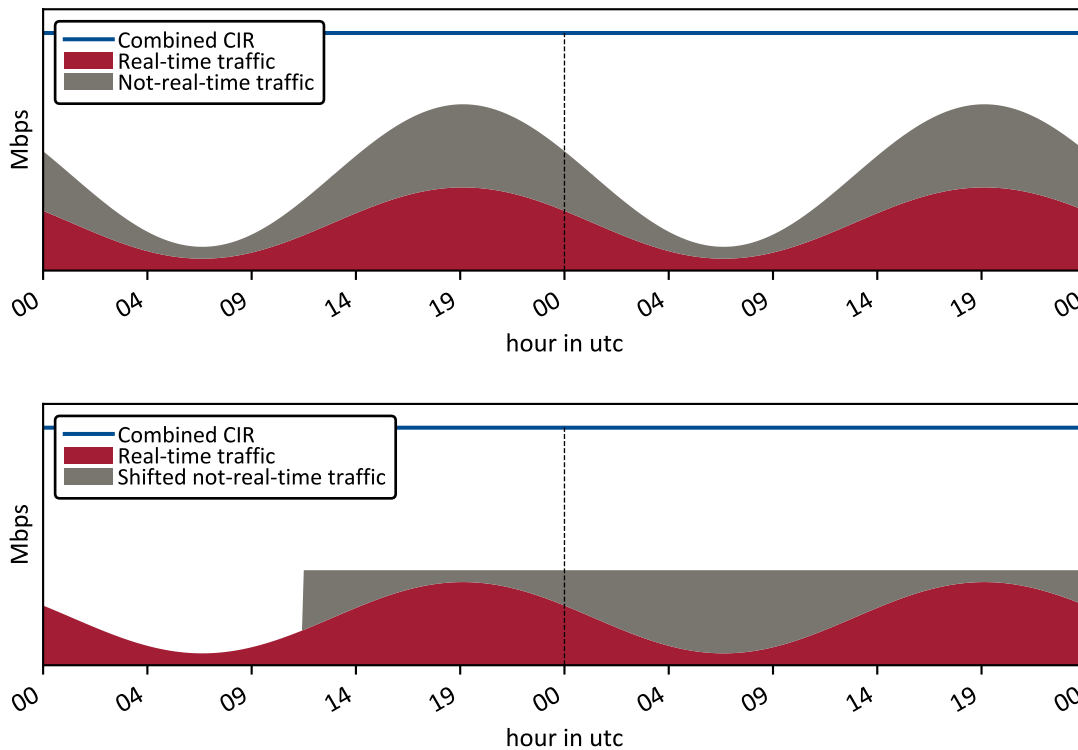


Figure 7-4: Top plot is equally split cosine traffic but not shifted; bottom plot shows the same traffic, but the not-real-time traffic is shifted by 12 hours resulting in perfect cancelation.

The keys to this SLA are that a) the traffic can be separated by the operator, and b) parts of the traffic can be delayed for some hours. The applicability of these criteria depends strongly on the customer segment,

the daily pattern, and even the individual customer. We summarize this SLA in the following Box 7-6. The parameters for the real-time part are similar to the Classical SLA defined in Box 7-1. A percentage defines the split, and the maximum delay is specified. Both classes of service have a different monthly price so that the revenues are the sum of the two pieces.

Box 7-6: Two Classes of Service SLA

Two Classes of Service SLA

- Traffic parameters
 - CIR with promised availability for the real-time traffic
 - Average percentage (or total amount) of not real-time traffic
 - Maximum allowable delay $\Delta t_{not-real-time}$
- Pricing
 - Monthly price for the CIR p_{CIR}
 - Monthly price for the not real time traffic $p_{not-real-time}$ delayed by $\Delta t_{not-real-time}$ and capped by a volume V_{month}
- Monthly revenues: $\Pi = p_{CIR} \cdot CIR + p_{not-real-time} \cdot V_{month}$

7.6 Market segmentation and mapping to novel SLAs

Market segmentation is commonly an activity of the Marketing department. Segmentation divides a market into “distinct groups of customers, with different needs, characteristics or behavior, who might require separate products or who may respond differently to various combinations of marketing efforts” [211, 212, p. 10]. Successful segmentation enables companies to position their products and service to groups that are most valuable to serve. Cooil et al. [212] divide methods for market segmentation into three categories:

- **A-priori segmentation:** segments are identified based on customer and product characteristics before customer data is collected.
- **Post-Hoc segmentation:** segments are classified based on a customer data analysis.
- **Hybrid segmentation:** these methods combine a-priori and post-hoc approaches.

The goal is to have segments with preferably homogenous groups. Kotler and Armstrong [211] define five criteria for a successful segment:

- **“Measurable:** size, purchasing power, and profiles can be measured
- **Accessible:** can be effectively reached and served
- **Substantial:** must be large or profitable enough
- **Differentiable:** two segments that react the same way are not actually separate segments
- **Actionable:** effective programs can be designed for the segment, i.e. matches company capabilities” [211, p. 37]

The market segmentation literature refers to the dimensions for classification as *bases*. One or multiple bases segment the market. Some commonly used bases for the consumer market (b2c) are geographic, demographic, psychographic, and behavioral [211]. In the business market (b2b), customer operating characteristics, purchasing approaches, situational and personal factors are additional bases. The relationship between supplier and customer can be an additional base [213], such as the criticality of the service and relationship for the customer, the fraction of the customer’s expense, and competition amongst suppliers. Vertical segmentation is alongside different industries (or countries), and horizontal segmentation across industries (or countries).

The majority of the literature on segmentation focuses on domestic markets with frequently purchased goods [214]. In contrast, satcom is an international service and mostly b2b. There is some work focusing

on switching behavior of international service customers (Ganesh et al. [215], Heide et al. [216], Keaveney et al. [217]).

The same market can be segmented in various ways depending on the dimension homogeneity is measured (for example, for marketing purposes and product targeting, or differential pricing). But also, segmentation performed for pricing (as essential for differential pricing of the RM system). We focus on the latter first and then discuss the segmentation for product targeting in the following Section 7.6.1.

As argued by Steenkamp et al. [218], price-elasticity is often a central driver for segmentation. Bolton et al. [214] studied both horizontal and vertical segmentation for international b2b based on price-elasticity. The authors [214] showed that the main driver for the difference in price elasticity across segments are: the customization of service quality (responsiveness, reliability, assurance, and empathy) and the type of customer's organization (how services are purchased). Bolton et al. [214] pose and validate six hypotheses on the elasticity for services:

1. More responsive service leads to less price sensitivity.
2. More reliable service leads to less price sensitivity.
3. More assurance and empathy lead to less price sensitivity.
4. Customers who purchase fewer services are less price sensitive.
5. If the service is highly critical to the customer, the customer is less price sensitive.
6. National and regional variables moderate service quality and organizational characteristics

The first five lead to horizontal and the sixth to vertical segmentation. The paper combines both segmentation and identifies the different behavior of countries to service quality dimensions. As an example, customers in Korea became less price sensitivity with a lower level of assurance and empathy. At the same time, in Asia Pacific, Europe, and North Korea, the company needs to offer a higher level of assurance and empathy to achieve the same insensitivity.

While we did not find specific academic literature on market segmentation for the satcom broadband market, we discover some research in the telecommunication (telco) industry: In 2010, Bayer [219] published a practitioner article on the change of segmentation in the telco industry. She notes that "telco operators have gone way beyond traditional segmentation based only on standard market criteria, such as prepaid versus postpaid, and consumer versus business. Advanced use of segmentation allows each customer to be part of a micro-segment." [219, p. 256]. The increasing granularity in the segmentation

allows businesses to customize products more precisely to customers, resulting in a win-win situation for businesses and customers.

Xevelonakis and Som [220] conducted a study with two unusual bases. They argue that getting an overview of the variety of offers is time-consuming when making a purchasing decision. Hence, customers tend to go back to ask friends for recommendations. As a consequence, Xevelonakis and Som propose a segmentation by the social network size and level of connectivity of each customer. They show that the churn rate¹¹ decreased, and spending increased by specifically addressing connectors and networkers with target offers and loyalty programs.

Băcilă et al. [221] used K-mean clustering for a post-hoc segmentation of 10,000 prepaid subscribers into 9 clusters. Their bases are the customers' spending on credits for calls, text messages, and data. While they describe the characteristics of each cluster, the authors did not discuss if targeted marketing is adequate for these clusters and how they compare to other segmentations.

While the discussed three papers are only a brief screening of the literature on segmentation in telecommunication, we observe that the transfer of segmentations to satcom is not trivial (however, some of the methods are). The segments are specific to the industry and vary significantly amongst authors, even in the same industry. As a result, we review two more satcom specific sources: (1) the International Telecommunication Union (ITU) as an agency of the United Nations responsible for information and communication technologies, and (2) Northern Sky Research (NSR) and Euroconsult as global leaders in satellite market research.

The ITU publishes radio regulations [222], which cover definitions of radio services. From there, we extract the following relevant satcom services:

- Fixed-satellite service (FSS): “A *radiocommunication service* between *earth stations* at given positions, when one or more *satellites* are used; the given position may be a specified fixed point or any fixed point within specified areas.” [222, p. 9]
- Mobile-satellite service (MSS): “A *radiocommunication service*: between *mobile earth stations* and one or more *space stations*, or between *space stations* used by this service; or between *mobile earth stations* by means of one or more *space stations*.” [222, p. 9]

¹¹ Churn rate is the ratio of customers who leave a company during a given time-period.

- Land mobile-satellite service: “A *mobile-satellite service* in which *mobile earth stations* are located on land.” [222, p. 9]
- Maritime mobile-satellite service: “A *mobile-satellite service* in which *mobile earth stations* are located on board ships.” [222, p. 9]
- Aeronautical mobile-satellite service: “A *mobile-satellite service* in which *mobile earth stations* are located on board aircraft.” [222, p. 10]
- Broadcasting-satellite service (BSS): “A *radiocommunication service* in which signals transmitted or retransmitted by *space stations* are intended for direct reception by the public.” [222, p. 11]

The last point is a unidirectional broadcasting service. The other two main points are bidirectional broadband connectivity (the scope of this dissertation). Regarding the bases, the ITU uses the following two: first, the movement of the earth station, i.e., the customer. If the station is stationary, the service is an FSS, and when it is moving, it is an MSS. For the moving stations, the ITU uses the geographical attribute as a further base: land, maritime, and aeronautical. Excluding the broadcasting service, these leads to four segments. Note, that the ITU does not specify if the satellite is moving relative to the Earth, (GEO or NGSO constellation).

In contrast to the ITU, NSR and Euroconsult use different segmentation depending on the report they publish. We focus on their pricing bottom-lines (e.g., the NSR 2020 edition [223] or Euroconsult [224]). Their most commonly used bases for broadband are:

- *Type of customer*: consumer, enterprise, mobility, government, land, maritime, aeronautical, energy, residential, trunking/backhauling [224-226]
- *Frequency band*: HTS, Ku, C [224, 225]
- *Region*: for example NAM (US and Canada), LAM (Latin America), WEU (Western European Union), CEEU (Central and Eastern Europe), MEA (the Middle East and Africa), EA (Eastern Africa), SA/SEA (South Asia, South East Asia), POR, AOR, IOR¹² [223]

To the best knowledge of the author, the ITU, NSR, and Euroconsult have not published an overview of their segmentation bases. We aim to fill this gap in the following Section 7.6.1 and discuss resulting segmentations in Section 7.6.2.

¹² NSR uses not standardized country groupings, and hence we were unable to identify what POR, AOR, and IOR stand for.

7.6.1 Identification of the bases of customer segmentation in satcom

By reviewing the NSR and Euroconsult reports, we already identified three bases: type of customer, frequency band, and region. In the following, we concentrate specifically on HTS broadband satcom. According to Frank et al. [227] and Hague and Harrison [228], bases are commonly divided into directly and not-directly observable groups. We follow this approach and further divide the directly observable group into firmographics and *behavior* (see Table 7-3). A morphological matrix [203] lists the bases together with possible categories of discretization.

Table 7-3: Segmentation bases for HTS broadband satcom

| Segment basis | Categories | | | | |
|----------------------------------|-------------------------|--------------------------|--------------------------|------------------|------------|
| | 1 | 2 | 3 | 4 | |
| Firmographics | Type of business | Single terminal consumer | Multi terminals business | Service provider | Government |
| | Location type | Air | Sea | Land | |
| | Region | Region 1 | Region 2 | ... | |
| | Terminals existing? | No | Yes | | |
| | Terminal size | 0.6m | 1.2m | ... | |
| Directly observable behavior | Usage per terminal | < 10 Mbps | 10 – 100 Mbps | > 100 Mbps | |
| | Daily pattern variation | Low | Medium | High | |
| Not-directly observable behavior | Price elasticity | Inelastic | Unit elastic | Elastic | |

In total, we find eight bases: five firmographics, two directly observable, and one not-directly observable behavior. The subsequent paragraphs describe and rationalize them in detail.

Type of business makes explicit what business is behind the “type of customer” as used by NSR and Euroconsult [224-226]. For example, their type “consumer” implies that it is mostly one single end-user. In contrast, “maritime” can be a single terminal consumer (private boat), multiple terminals business (cruise ship cooperation), a service provider, or a government customer. With our basis, this separation becomes explicit.

Location type aligns with the ITU’s second basis for segmentation and the “type of customer” basis of NSR and Euroconsult. The customer can be in the air, in the sea, or on land.

Region is the geographical basis and taken from the NSR and Euroconsult review. The granularity can vary from as small as counties or cities to groups of countries or oceans. The basis correlates with the location type: the location type limits the possible regions the customer can be (and vice versa). For example, a land customer cannot be on the ocean; however, an aeronautical customer can be both over a country and over the ocean.

Terminal existing? segments customers with an existing terminal from these without. If a compatible terminal already exists, this indicates a lower switching cost for the customer. Service offerings and marketing strategies might be different.

Terminal size has central implications for the capacity cost. A 0.6m dish has a 12 dB lower G/T than a 2.4m antenna (see Table 8-1). To achieve the same data rate, the operator needs to provide over 15 times the power.

Usage per terminal defines the average throughput of the customer. This basis is no longer a firmographic, but an observable behavior. It is on a user level, and we use here the average as an aggregation to the customer level. The three categories refer to the CIR, which correlates with the terminal size. Higher usage customers usually have larger terminals.

Daily pattern variation quantifies the difference between day and night usage. More considerable variations allow the operator to reallocate more capacity. The combination of all user's daily pattern shape defines the required capacity over time. Since the operator measures the usage, this basis is directly observable.

Price elasticity is the not-directly observable. The operator has to rely on historical contractual data and insights into the customer. Due to the small number of bookings and the b2b nature of satcom, price elasticity estimation is a challenge (more on that in Sections 8.1.2 and 8.6). Nevertheless, and especially for pricing decisions, it is crucial. According to Steenkamp et al. [218], elasticity is a central driver for segmentation. While we use here the three broad categories: inelastic (larger than -1), unit elastic (-1), and elastic (smaller than -1), smaller granularity is often desired (see Appendix D for further background information on demand elasticity).

7.6.2 Derived segmentation

In this Section, we discuss the derived segmentation based on the presented eight bases. We start by representing commonly used segments in the satcom industry with our eight bases. Then, we derive another more detailed segmentation.

We discussed in the previous Section that NSR and Euroconsult [224-226] use various segments, and we reviewed in Chapter 2 companies’ webpages and for publicly traded companies their annual 10-K reports. By combining this research, we identify seven central segments: aeronautical, backhauling, maritime, trunking, energy, residential, business, and government. For each segment, Table 7-4 maps the corresponding characteristic of the eight bases (when characteristic if defined). Note that the characteristics are only common areas, and some customers in the segment might have different characteristics. Furthermore, as we showed in Figure 2-1 in Chapter 2, service providers act currently as an intermediary in most segments with a trend of operators to move directly to the end-user. The eight segments do not specify the *type of business* basis (so is the region). As we did not find publicly available research on the price elasticity, we leave it unspecified. However, we provide a further discussion in Sections 8.1.2 and 8.6.

Table 7-4: Characteristics of each commonly used segment’s bases from the operator perspective

| Segment | Firmographics | | | | Directly observable behavior | | Not-directly observable | |
|---------------------|------------------|---------------|--------|--------------------|------------------------------|--------------------|-------------------------|------------------|
| | Type of business | Location type | Region | Terminal existing? | Terminal size | Usage per terminal | Daily pattern variation | Price elasticity |
| Aeronautical | - | Air | - | Yes | 0.6 – 1.2m | 10 – 100 Mbps | High | - |
| Backhauling | - | Land | - | - | 1.2 – 4.5m | > 10 Mbps | Medium | - |
| Maritime | - | Sea | - | - | 1.2 – 2.4m | > 10 Mbps | High | - |
| Trunking | - | Land | - | - | 2.4 – 4.5m | > 100 Mbps | Low | - |
| Energy | - | Sea | - | - | 1.2 – 2.4m | 10 – 100 Mbps | Medium | - |
| Residential | Single | Land | - | - | 0.6 – 1.2m | < 100 Mbps | High | - |
| Business | Multiple | Land | - | - | 0.6 – 1.2m | < 100 Mbps | Medium | - |
| Government | Govt. | - | - | - | 0.6 – 2.4m | < 100 Mbps | High | - |

Aviation services provide broadband connectivity to commercial or private aircraft. The end-customers are the passengers on board. All passengers are multiplexed and perceived as a single customer by the satellite.

Backhauling services connect a subnetwork with the fiber backbone of the Internet. Often the subnetwork is a cellphone tower or Wi-Fi hotspot in regions with no or low-speed terrestrial connection. Backhauling is often on larger scales and is especially considered a viable solution for developing countries.

Maritime services are similar to aviation. Due to slower movements, the communication link has fewer dynamics. For large cruise ships, the link contains thousands of multiplexed end-customers. The cruise ship company often acts as an intermediary between the satellite operator and the end-customer.

Trunking services are similar to Backhauling with the focus on using satellites for excess demand in terrestrial networks or contingency scenarios. The traffic volume is often considerable, and less variable and uncertain as many end-customers are multiplexed.

Energy services target oil and gas offshore customers. When their location is far from the shore, satcom is the only solution to provide connectivity. Depending on the size of the platform, traffic volume varies.

Residential broadband addresses a single terminal consumer to connect their house with the internet. This service is particularly attractive in remote with slow or no internet.

Business broadband is residential broadband but addressed to small or medium businesses. While there are some businesses with only a single terminal, the majority has multiple terminals.

Government is the broadest segment since the terminals can range from small sizes on UAVs to larger terminals on a remote base. All terminals would fit into the above segments; however, the government is usually considered a separate segment since the contracting process and security requirements are different from b2b or b2c.

We show another possible classification of these eight segments in Figure 7-5 (inspired by and adjusted from [229, 230]).

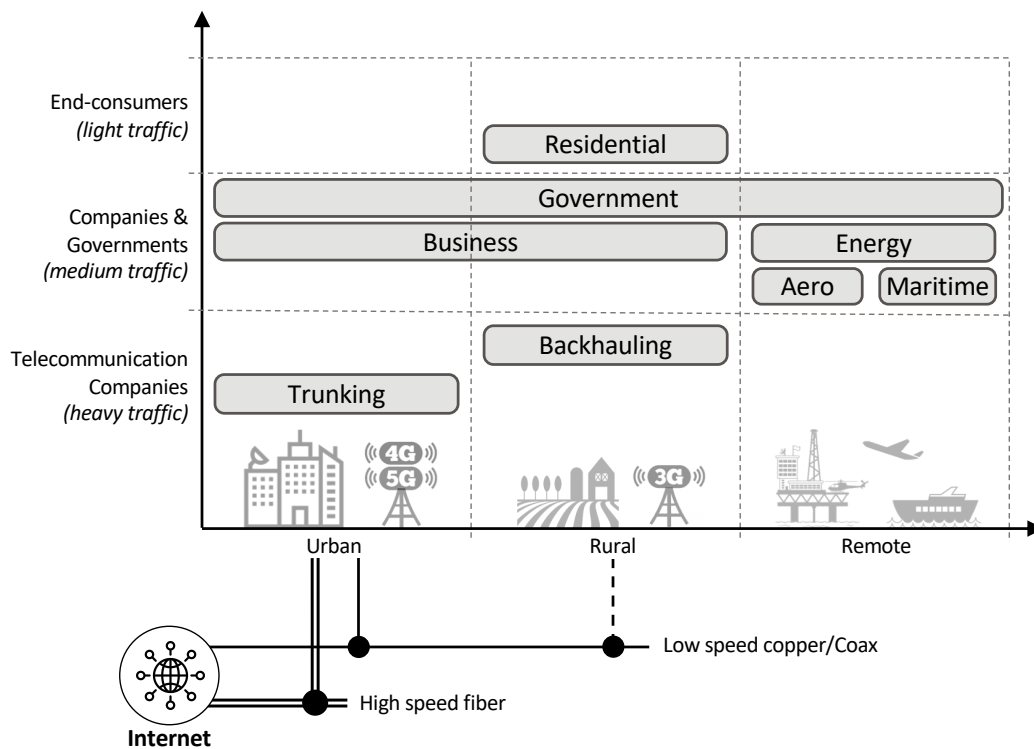


Figure 7-5: Classification of the seven segments based on location and traffic volume, inspired by [229, 230]

The horizontal axis is the location of the terminals ranging from urban with high-speed fiber to remote areas with no terrestrial alternative. The vertical axis separates the traffic volume of the terminals into heavy, medium, and light, which roughly corresponds to telecommunication companies, companies and government, and end-consumers, respectively. Despite available connectivity in urban areas, businesses or the government might choose satcom for its reliability and security; hence the segments range multiple locations. The importance of the latency service dimension depends on the traffic mix of each customer. Since Aviation sees less real-time traffic like calls and teleconferencing a greater latency might be acceptable compared to backhauling with a larger real-time traffic fraction.

While the industry uses the discussed segmentation, the difficulty in assigning specific characteristics to each segment (see Table 7-4) is an indication that the customers in each segment are *heterogeneous*. That is not surprising given the diversity of satcom customers (worldwide, different industries, businesses and consumers, various security and traffic requirements). Nevertheless, homogenous customers characterize a “good” segmentation [211]. Figure 7-6 is an attempt to use the bases *type of business*, *location type*, and *usage per terminal* to build segments systematically. The blue font identifies the final segment, where bold are segments that Table 7-4 does not list.

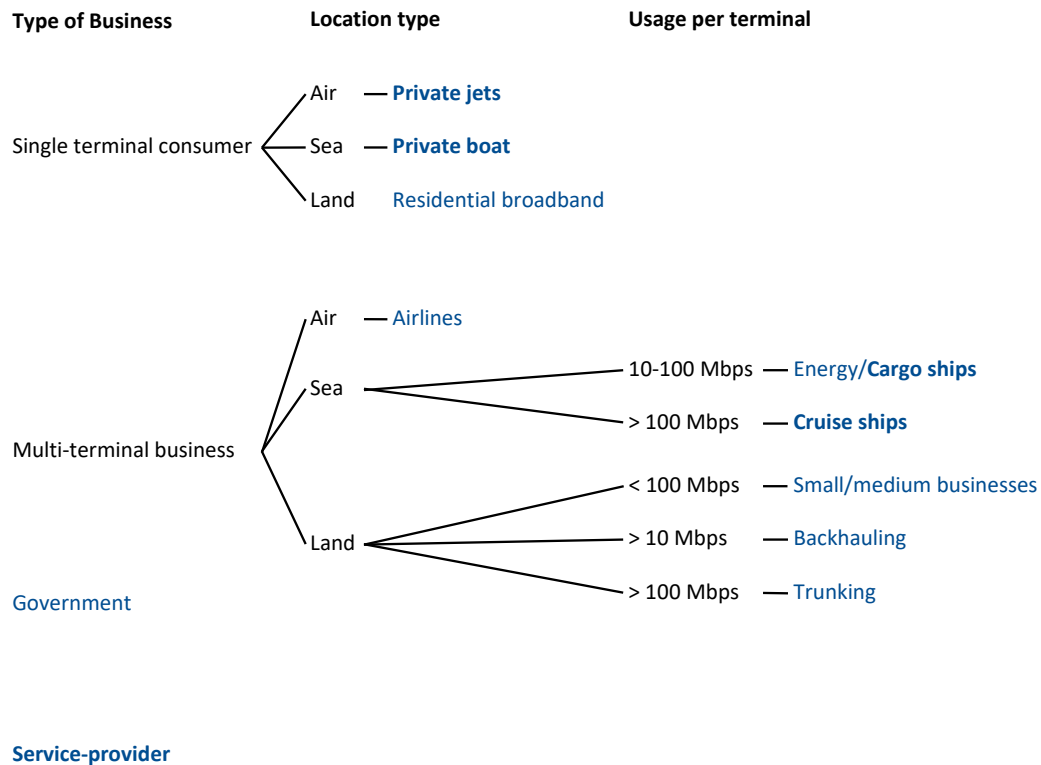


Figure 7-6: Segmentation of the market by using the three bases: type of business, location type, usage per terminal. Bold segments are these not identified in Table 7-4.

The first level is the type of business: single terminal consumer, multi-terminal business, government, or a service provider. We finish the decomposition of the latter two as both often serve a heterogeneous user base. Location type further segments the single terminal consumer and multi-terminal business into air, sea, and land. We then further decompose the sea and land multi-terminal businesses depending on their usage (note that some options are not exclusive). This process leads to the characterization of private jets and boats, cargo and cruise ships, and service providers.

While we use for the segmentation only three bases, the other five bases can be used as follows:

- *Region* applies to all the blue segments. It captures culture, behavioral, and willingness-to-pay gradients between countries.
- *Terminal existing?* applies to all blue segments, but customers across segments might have a different sensitivity to the Capital expenditures (CAPEX) for a user terminal, e.g., for residential consumers, the user terminal cost plays a much more significant role than for airlines.
- *Terminal size* tightly connects to the CAPEX of user terminals. Larger terminals are more expensive and sometimes are difficult to install due to volume limitations. Still, the operator can offer cheaper service in return, resulting in a CAPEX vs. Operating expenditures (OPEX) trade-off for the customer. This trade-off is different in every segment (and also depends on the expected traffic).
- *Daily pattern variation* is, in fact, a result of the segmentation, e.g., private jets with occasional use has a larger variation (that might not even be diurnal), whereas higher throughput terminals show less variation.
- *Price elasticity* is commonly a central segmentation basis [218]. However, due to the lack of data, we are unable to test to which segments a price elasticity-based segmentation leads.

As our result has more segments, it should be no surprise that customers are expected to be more homogenous within a segment (to the extreme where each customer has its segment). However, there is a trade-off with accessibility, substantial enough, and differentiable (depending on the use case of the segmentation). In our example, the segment of private jets might not be substantial enough to warrant a separate treatment.

Overall, the results suggest that the commonly used segments are indeed a logical first level of segmentation. Yet, there is future work that can build on the presented discussion on customer segmentation in satcom – particularly in the context of RM.

7.6.3 Mapping of novel SLAs to segmentation

In this Section, we map the current and novel SLAs to the identified market segments. We use the scale of -, ·, + with the meanings not optimal, ok, and a good fit, respectively. Table 7-5 summarizes the results of our assessment. We discuss and rational our thought-process in the following structured by segment.

Table 7-5: Our assessment of the fit of the six SLAs in the ten segments (bb = broadband).

| | Private jets/boats | Residential bb | Airlines | Energy/Cargo ships | Cruise ships | Small/medium bus. | Back-hauling | Trunk-ing | Government | Service provider |
|------------------------|--------------------|----------------|----------|--------------------|--------------|-------------------|--------------|-----------|------------|------------------|
| Classic SLA | - | · | · | · | · | · | + | + | + | · |
| Data volume SLA | · | + | + | · | · | · | · | - | - | · |
| Dual SLA | + | + | + | + | + | + | + | · | · | + |
| Spot instance | + | · | · | · | + | · | · | - | - | + |
| Time-of-day pricing | · | + | · | + | · | · | + | · | · | + |
| Two Classes of Service | + | + | · | + | + | + | + | + | - | + |

Private jets and boats have not only high diurnal variations but also seasonality. Boats might be used during vacation time and jets used on a short-term and varying flight schedule. Because of that, we suppose that the classical SLA is not an optimal fit for this segment since the purchased CIR is likely unused for long periods. The data volume SLA can be a better fit since monthly volume smooths the variations in usage. However, volume-based SLAs do not offer any speed guarantee. That might be desired, especially from rather price-insensitive business jets owners. On one side, occasionally used jets and boats can benefit from the short-term flexibility of spot instances, which can be bought just before the trip. On the other side, for more frequently used jets and boats, the time-of-day pricing can be an attractive option. And last, Two Classes of Service can support boat owners to keep communication expenses low during longer trips.

Residential broadband is, in contrast to private jets and boats, always connected. It yields a more predictable daily and weekly pattern (usage over the weekend is generally higher). However, residential consumers are generally more price-sensitive than private jet and boat owners. Therefore, time-of-day pricing can be an attractive offer for residential broadband customers to keep prices low. Due to the repetitive pattern, spot instances are likely less attractive, while the classical SLA can be a satisfying offer. Since minimum speed requirements might be lower, we consider the data volume SLA to be a good fit.

Airlines are the business counterpart to private jets. They have multiple airplanes with defined flight schedules, commonly repeating daily. Due to this repeating pattern, spot instances are likely less preferred. End-users connect for a comparable short amount of time, therefore separating the classes of service by a time of hours is not an optimal fit. Nevertheless, shorter delaying of traffic can be beneficial to smoothen out bursty spikes, which leads us to an “ok” assessment. The rationale behind the first three current SLA is similar to the residential broadband discussion.

Energy and Cargo ships are the business type of residential broadband. Therefore, the matching is very similar except for the data volume SLA, which can be a less good fit for business since a no minimum speed is promised.

Cruise ships show an apparent seasonal behavior over the year (depending on their place of operation). Due to a larger number of end-users on board, the traffic is considerable but not always well predictable. For example, special events or driving by a scenic island spikes up traffic as end-users desire to share the experience over the Internet. Spot instances can be an option for cruise ships cooperation to support the higher traffic without the need to purchase a Classic SLA with a higher CIR.

Small and medium businesses are the land version of energy and cargo ships. The difference that we make is time-of-day pricing. Since on energy platforms and cruise ships, the employees live where they work, their communication is mostly business-related during the day and private during the evenings. This situation usually is not the case for small and medium businesses; the time-of-day pricing is less good of a fit.

Backhauling is a unique segment. Similar to cruise ships, many end-users are connected, but traffic has less seasonal variation. Special events are also rarer (e.g., some significant sport events). With that, the Classic SLA is a good fit. End-users are residential consumers or small businesses and, therefore, more price sensitive. Hence time-of-day pricing and Two Classes of Service might be attractive options.

Trunking is the segment with the highest and smoothest traffic. A common decision argument for customers to choose trunking over terrestrial alternatives is the security and reliability of satcom. There are also some use cases where satcom trunking is considered as a backup if there is a failure in the terrestrial network; in that case, the arguments here do not hold. The classic SLA with ensured CIR is a favorable choice, and so can be a two-classes of service that separate real-time from not real-time with a more attractive price in return.

Government has similar requirements to trunking. The communication link is required to be secure and reliable. Government customers tend to be less price-sensitive. One difference to trunking is the two-classes of service, which requires deep package inspection. Due to frequently stringent security requirements, this might often not be possible to do, and hence the classes cannot be separated from the operator.

Service provider is the last segment that we discuss. It is challenging to make an assessment here, since depending on the segment the service providers serve, the requirements vary widely. The service providers' capacity purchasing decisions are constraint by the products that the operator sells. A wider variety of SLAs and better flexibility in them allows the service provides to match their capacity closer to uncertain and varying demand. The operators support their customers (service providers) to serve the needs of their customers.

To summarize, across segments, the following rules seem to apply for the mapping:

- Classic SLA excel where reliability and speed are essential, traffic volume is high, and variation is low.
- When consumers are price-sensitive, data volume SLAs, time-of-day pricing, and Two Classes of Service are affordable options.
- Dual SLAs are a good compromise across all segments.
- When the traffic patterns are not repeating on a daily or monthly basis, spot instances offer the desired flexibility.

7.7 Summary and contributions

This Chapter first reviewed commonly used SLAs in satcom and identified their challenges from the operator as well as the customer perspective. We studied SLAs in the cloud service and telecom industries and drew analogies to satcom. Combining these insights with the challenges, we proposed three new SLAs: *spot instance*, *time-of-day pricing*, and *Two Classes of Service*. Finally, we assessed the classical and the novel SLAs regarding their benefits to the different customer segments.

The contributions we make in this Chapter lie within:

- Identified and described currently used SLAs in satcom.
- Characterized limitations of existing SLAs in the context of new flexible payloads and RM.
- Built analogies to the cloud service and telecom industry.
- Proposed a novel set of SLAs that are beneficial to operators and customers
- Discussed market segmentation bases and mapped the resulting ten segments to the three classical and three novel SLAs

8

Application of the Satcom RM Framework

In the previous Chapter 4 – 7, we described our solution to the four main challenges of satcom RM. The objective of this Chapter is to bring the pieces together and illustrate the benefits of implementing and using the satcom RM framework based on several simulated scenarios. We fill the databases with real data from SES where possible and discuss the remaining components of the framework: customer elasticities estimation and pricing optimization (see Figure 8-1).

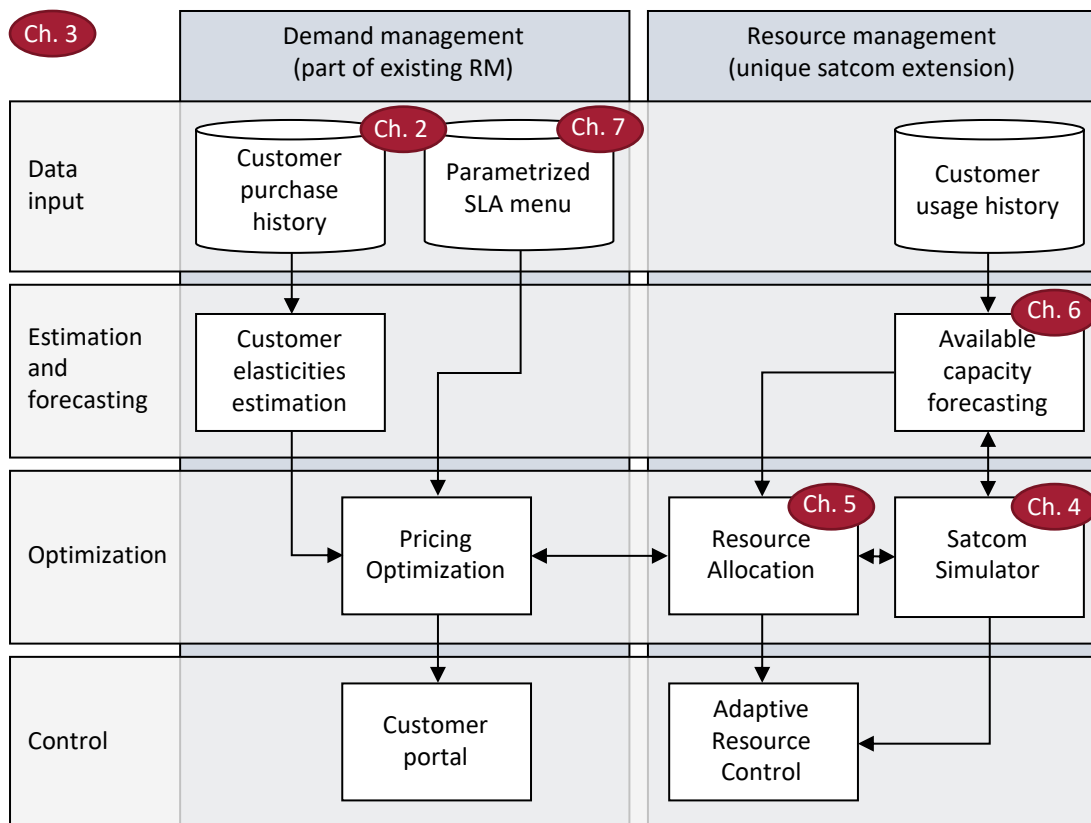


Figure 8-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4

The simulation and analysis in this Chapter are significantly extending Section 3.6. We consider a higher number of user terminals, an O3b mPower like NGSO constellation, and four different ways to monetize the available capacity. Furthermore, we analyze the effects in a competitive environment of both dynamic resource allocation and setting prices through the RM framework. Throughout the analysis, we separate the lift of these two effects when possible.

In the following Section 8.1, we describe the input data that is the same to all simulations of this Chapter. That includes details on user terminals, gateways, constellation, traffic usage, and elasticities. Figure 8-2 logically structures the four subsequent Sections 8.2 - 8.5.

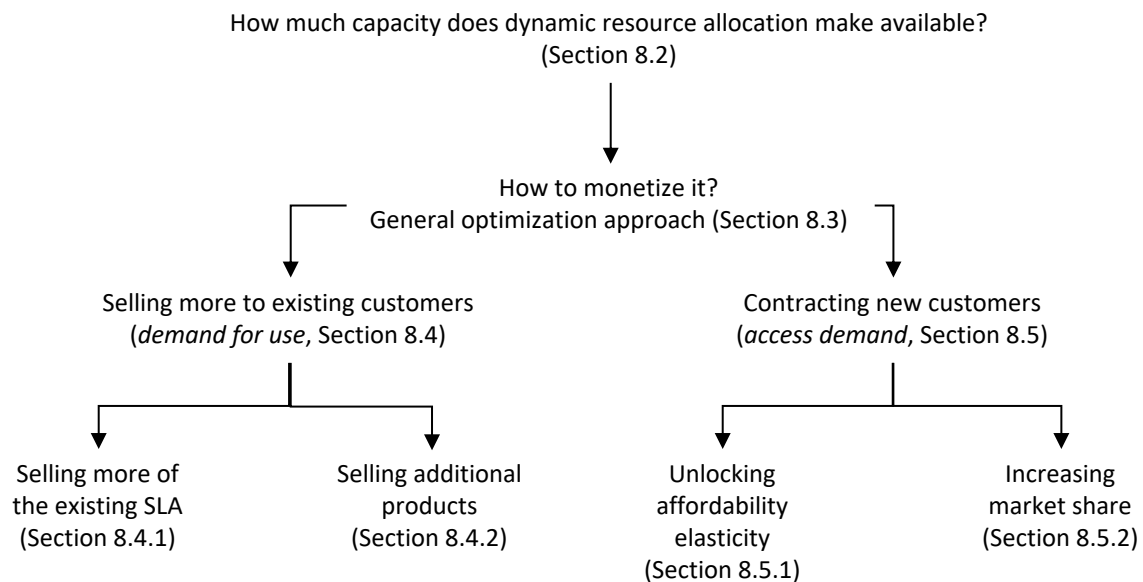


Figure 8-2: Overview of the structure of the analysis for Sections 8.2 - 8.5

We start by analyzing how much dynamic allocation reduces resource consumption, i.e., how much capacity becomes available (Section 8.2). With this metric, we measure the improvement from dynamic resource allocation in terms of power. The next step is to address the question of how this lift can be monetized, or put differently, the “translation” of available power into additional revenues. Section 8.3 describes our general optimization approach to that.

One option that we consider is the selling of more capacity to *existing* customers (Section 8.4). That can be achieved by selling more volume of the existing SLA (Section 8.4.1, referred to as *demand for use* in terrestrial telecommunication [231]) or provide customers with additional products (Section 8.4.2).

The other option is contracting *new* customers (Section 8.5). We further separate this option into two possibilities. First, price adjustments unlock the affordability elasticity of the segments (referred to as

access demand in terrestrial telecommunication [231], Section 8.5.1). And second, customers switch from competitors, which results in an increasing market share (Section 8.5.2).

We see that results are sensitive to the elasticity, the estimation of which can be challenging. Hence, we dedicate a separate Section, 8.6, to discuss approaches and learning strategies. Section 8.7 summarizes the results from the four analyses and draws conclusions. We discuss the market implication of adopting RM in Section 8.8. The final Section 8.9 summarizes and concludes this Chapter.

8.1 Input data and assumptions

8.1.1 User traffic data

We start this Section with the description of the available data from selected SES's current users (note that we use in this Chapter user and customer interchangeably). For each customer, we have the following features:

- Segment (encoded)
- CIR
- Monthly recurring revenues from which we compute the price in \$/Mbps/Month
- SLA status: active, ended, closed but not yet active
- SLA service start and end
- Location information: approximate latitude and longitude, country name, and iso-a2 code
- Terminal size from which we derive the G/T (see Table 8-1)
- Preprocessed usage data: mean usage over 24h in 1-minute granularity and uncertainty expressed in a sigma value for the average of the means (we scale the sigma for each time step according to Eq. (3-6))

Table 8-1: Terminal EIRP, G/T, and G assumptions based on the terminal diameter

| terminal size [m] | EIRP [dB] | G/T [dBK] | G [db] |
|----------------------|--------------|--------------|-----------|
| 0.3 | 46 | 9 | 32.7 |
| 0.6 | 52 | 15 | 38.7 |
| 1.2 | 61 | 21 | 44.7 |
| 2.4 | 70 | 27 | 50.7 |
| 4.5 | 80 | 35 | 58.7 |
| 7.3 | 85 | 40 | 63.7 |

We sanitize the data to comply with confidentiality agreements. In particular, we adjust all prices by multiplying with a constant number and provide approximative information of the latitude and longitude. We use this dataset to build our two databases, *customer usage history*, and *customer purchase history* (see Figure 8-1). We describe the usage history here below.

We consider 80 *active* worldwide distributed user terminals with an accumulated CIR of 22.7 Gbps. The mean CIR per user terminals is 296 Mbps, with the smallest having 6 Mbps and the largest 1500 Mbps. These user terminals belong to four different *customer groups*, encoded as A, B, C, and D. Figure 8-3 depicts a histogram of the CIR and the size of each segment based on their CIR.

We generated the same histogram for the traffic model in Chapter 5 (see Figure 5-8). Figure 5-8 had most of the terminals below 100 Mbps with the highest concentration around 20-30Mbps. The current active customers of SES have higher CIRs with considerable density for user terminals above 700 Mbps (see the left plot of Figure 8-3).

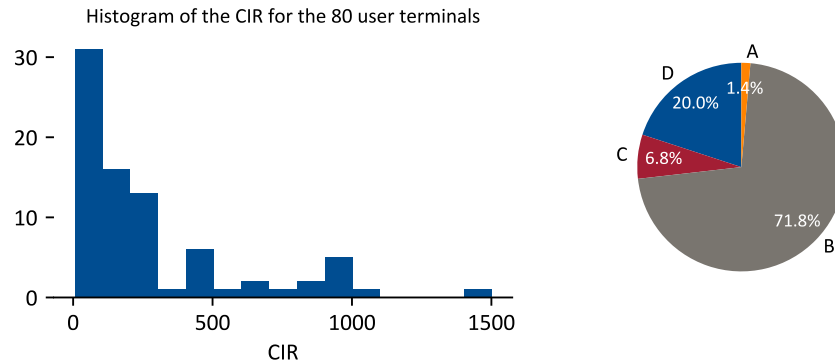


Figure 8-3: Statistical description of the 80 users

The right plot of Figure 8-3 shows the relative size of the four customer groups measured by the CIR. Group A is the smallest, with only accounting for 1.4% of the total CIR. Group B is the largest, with 72%, followed by group D and C with 20% and 7%, respectively.

Figure 8-4 depicts four usage profile estimations for an exemplary user terminal for each segment. The user terminals are from various longitude locations, so the dip in usage occurs at different UTC hours. The available base sigma σ_b is multiplied by the mean for each time step to obtain an uncertainty that considers the absolute usage (e.g., lower uncertainty around Hour 9 for group B than later in the day). The plots show all data normalized to the CIR of each user terminal.

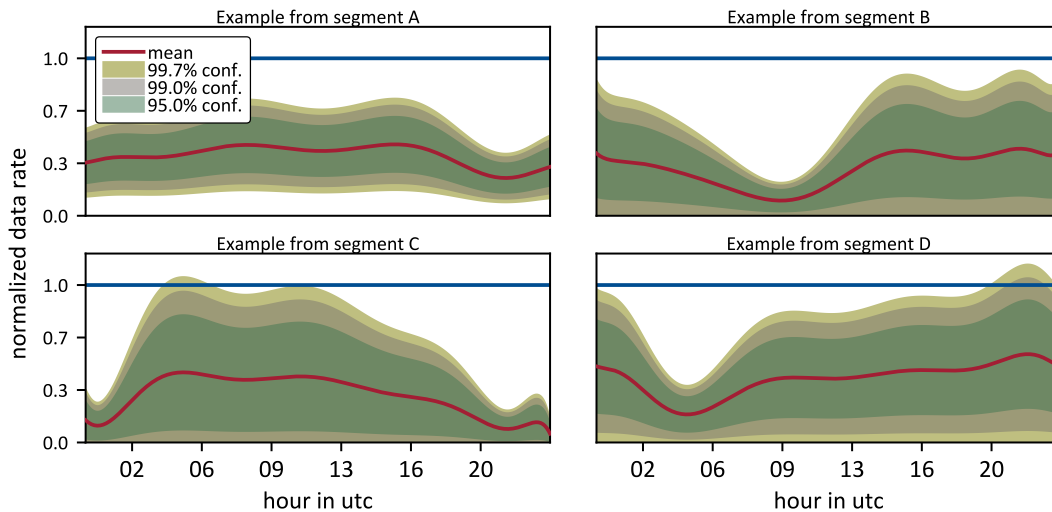


Figure 8-4: Traffic estimation for one example user from each segment

8.1.2 Elasticities

We show in the various simulations that the results are sensitive to the elasticity. Therefore, understanding the elasticities in each segment is crucial (see Appendix D for a primer on demand elasticity). We start this Section by estimating them based on past purchases. However, we deem the results from the regression as inaccurate, and we, therefore, do not use them for our analyses. As a second approach, we review the academic literature on elasticities in the telecommunication industry with the belief that the results are transferable to satcom (as there is no academic literature on broadband satcom elasticities). We combine the outcomes from both approaches to find a *range* of elasticities for sensitivity simulations in this Chapter. Section 8.6 discusses further approaches regarding the learning of elasticities.

Elasticity estimation based on customers' purchase history

For the first approach, we use 800 datapoints that are made available to us as a *customer purchase history* database. The data contains contracts between 2015 and 2023. As we showed in Chapter 2 (and also confirmed by this dataset), prices are declining in the last years. This trend would bias the price-elasticity estimations. Therefore, we correct for the price decline by adjusting all data points to a 2019 reference year. In that, we first build a linear regression model of the original data points (see Figure 8-5). Then we adjust all data points by the difference between the function evaluation of the regression for the reference year 2019 and the year of the data point. The results are the adjusted blue data points, which we use from here onwards. We notice the considerable error of the linear regression, giving the first indication that this approach could be challenging.

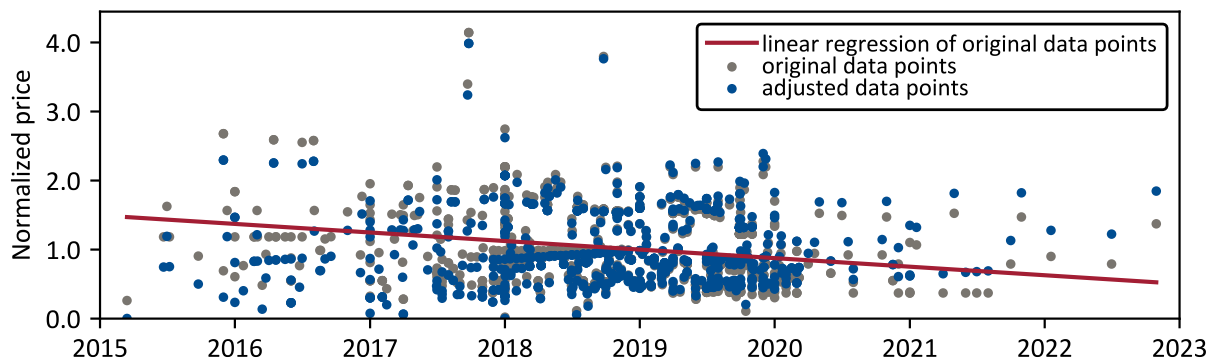


Figure 8-5: Linear regression of price decline and adjustment of the data points for this trend. The prices are normalized to the function value of the linear regression for the reference year 2019

The 800 data points are from four different customer groups, encoded with A, B, C, and D. As this grouping is the most obvious, we use it for the purpose of this Section. We build separate demand elasticity regressions for each of these four groups. Our prior is a power/log-linear curve with constant elasticity:

$$R = a \cdot p^b \quad \text{or} \quad p = \left(\frac{R}{a}\right)^{\frac{1}{b}} \tag{8-1}$$

where a and b are fitting parameters with b having the specific meaning of the *elasticity*. When applying the logarithmic function to the equation, we get the linear Eq. (8-2) for which we can compute the coefficients by minimizing the residual sum of squares with linear approximation.

$$\log(R) = \log(a) + b \cdot \log(p) \tag{8-2}$$

Figure 8-6 shows the resulting regression in blue for each of the four segments. Group A has the most elastic demand with -1.42, followed by group C and B with around -0.7, and then group C with -0.31. We report the summary of that together with the fitting parameter a and the R^2 in Table 8-2. While the elasticity is independent of the scale of the price axis, the values for the parameter a are for the normalized price. The values for the quality of the fit measurement (R^2 coefficient) indicate a far from optimal fit (optimal fit has value 1). Given the distribution of the data points, we expect other priors yield R^2 values in a similar range.

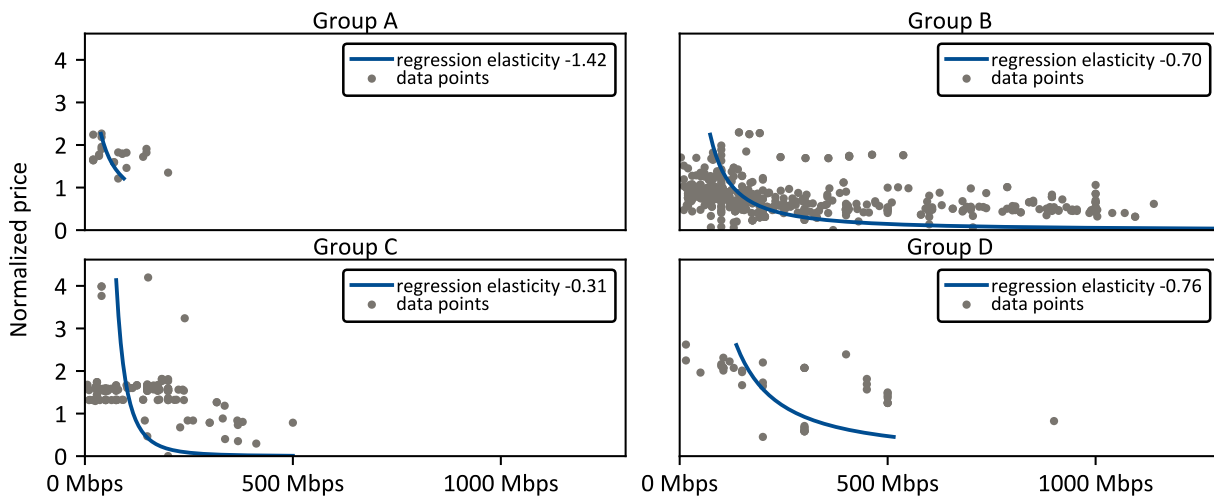


Figure 8-6: Power regression of the price elasticity for each group. Normalized prices are based on Figure 8-5

Table 8-2: regression parameters for $R = a \cdot p^b$

| Group | a | b | R^2 |
|-------|----------|-------|-------|
| A | 1,709.55 | -1.42 | 0.08 |
| B | 19.72 | -0.70 | 0.13 |
| C | 1.53 | -0.31 | 0.05 |
| D | 62.28 | -0.76 | 0.20 |

Given both the low R^2 measure and the visual impression from Figure 8-6, we can say that regression works “poorly”. At best, we can identify the range from -1.42 to -0.31 with the tendency that A is more

elastic, B and D are similar, and C is a more inelastic group. As a consequence, we do not use this regression data for the following analyses and the conclusions of this Chapter are independent. Future research can be to elaborate on the elasticity regression to achieve better results that can be used in the simulations. However, for this dissertation, we follow a different approach as outlined in the next subsection.

Elasticity estimation based on analogy to telecommunication literature

Econometrics is the discipline that studies, amongst other things, the estimation of economic parameters such as the elasticity [231-234]. From a customer perspective, the terrestrial telecommunication industry is the closest analogy to satcom. Hence, we review here selected econometric papers on the price elasticity estimation of customers in the telecommunication industry. Our goal is to find a typical range.

Aldebert et al. [231] use a translogarithmic indirect utility function to analyze residential demand for different traffic destinations (2004). Besides the price elasticity, they found that there is a considerable income elasticity. In analogy to satcom, this suggests that geographical segmentation is critical. Aldebert et al. consider the five traffic destinations: local, national, international, towards-mobile, and other traffic. They describe that the general understanding of telecommunication is that long-distance elasticity is higher than for local traffic.

Comparing Aldebert et al. [231] results with other works, Aldebert et al. found some differences. For local traffic, the authors found an elasticity of -1.44 compared to -0.09 by Park et al. [235] and -0.88 by Wolak [236]. For long-distance traffic, Gatto et al. [237] compute -0.72, while Wolak [236] has the highest elasticity with -2.07. These numbers compare to -1.33 for national traffic from Aldebert et al. [231]. They attribute the discrepancies to considering short- (under one year) vs. long-run (over one year) elasticities and different modeling approaches. For international traffic Aldebert et al. estimate a fairly inelastic demand with -0.11, while Garin-Muñoz and Perez-Amaral [238] compute values between -1.31 and -0.32 varying between countries.

Ouwensloot and Rietveld [233] focus their review on the distance dependency of the price elasticity in telecommunication. They found that a doubling in the distance leads to a 0.07 increase in the elasticity. They categorized their review into local, national, and international calls with ranges from -0.05 to -0.75, -0.24 to -2.57, and -0.03 to -2.19, respectively. These scopes are similar to the ones discussed by Aldebert et al. [231] and further illustrate the challenge associated with computing elasticities.

Hackl et al. [232] concentrate on the international telecommunication between Sweden and Germany, the United Kingdom, and the USA. They conclude that short and long-term price elasticities are in a similar

range to other studies. The short-term elasticity ranges from -0.09 to -0.98, and the long-term from -0.19 to -1.61. The authors furthermore analyze how the elasticity changed over time. They discuss various challenges associated with this task but compute a likely behavior (see Figure 8-7). The elasticity transitions from being elastic to inelastic as telecommunication becomes more widely available, and the service develops into a commodity. Satcom is currently expected to be on that path as well (given declining prices and increasing capacity).

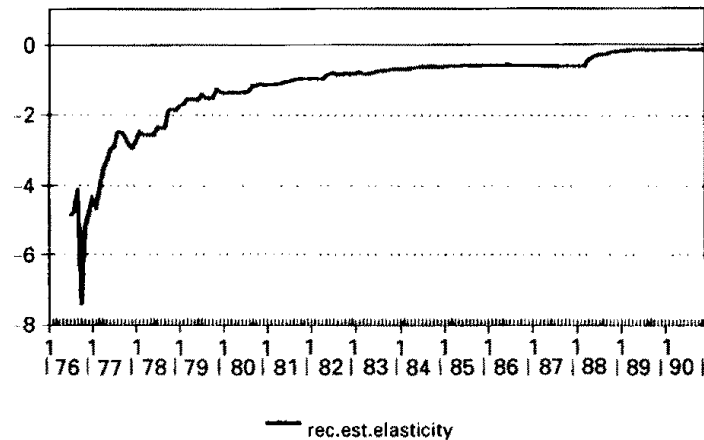


Figure 8-7: recursively estimated price elasticity for telecommunication obtained from [232, p. 255]

Palerm wrote an article for NSR [58] that contradicts the trend from elastic to inelastic over time for satcom. He notes that at the current price point, satcom is inelastic but argues that a further price decline can push the satcom industry into an elastic area of demand. The leading cause of that is cellular backhaul with an apparent demand elasticity effect. At current prices, Mobile Network Operators (MNOs) fall back to backhauling to comply with regulatory requirements (for which demand is inelastic). However, when prices fall, MNOs are commercially driven to expand on backhauling for which demand is estimated to be elastic.

In summary, we reviewed multiple econometric papers on the price elasticity of demand in telecommunication and found that the difference in the reported numbers is considerable. With our regression on satcom elasticity, we found similar challenges. These observations motivate us to treat the elasticity as an uncertain parameter on which we conduct several sensitivity analyses. By combining the values from both approaches, we pick a range from -2 to -0.5 that covers the range of elastic, unit elastic, and inelastic demand.

8.1.3 Gateways

We assume eight gateways distributed around the world, as shown in Table 8-3. They are in Hawaii, Texas, Azores, Sau Paulo, Cape Town, Colombo, and Auckland. All gateways serve out of country traffic and have a terminal size of 4.5m, giving them a G/T of 35 dB (see Table 8-1). Figure 8-8 depicts the location visually.

Table 8-3: Overview of the locations and sizes of the seven gateways considered

| Unique ID | Lat [deg] | Lon [deg] | country code [iso-a2] | country name [-] | terminal size [m] | in/out country [-] |
|-------------|-----------|-----------|-----------------------|--------------------------|-------------------|--------------------|
| GW_Hawaii | 19.9 | -155.6 | US | United States of America | 4.5 | out |
| GW_Texas | 30.6 | -96.3 | US | United States of America | 4.5 | out |
| GW_Azores | 37.7 | -25.7 | PT | Portugal | 4.5 | out |
| GW_SauPaulo | -23.6 | -46.6 | BR | Brazil | 4.5 | out |
| GW_CapeTown | -34.4 | 18.3 | ZA | South Africa | 4.5 | out |
| GW_Colombo | 6.9 | 79.8 | LK | Sri Lanka | 4.5 | out |
| GW_Auckland | -36.5 | 174.6 | CK | New Zealand | 4.5 | out |

8.1.4 Space segment

Since SES's O3b legacy MEO constellation serves these customers, we consider for this analysis a space segment is similar to the new O3b mPower MEO constellation. Table 8-4 lists the detailed parameters.

Table 8-4: Parameters of the MEO constellation, which is O3b mPower like. Compared to Table 5-3, the reuses are reduced from 20 to 4 to account for the lower traffic (198 Gbps vs. 22.7 Gbps forward CIR)

| | Value |
|--------------------------------|--|
| Number of satellites | 7 for MEO |
| Orbit | Equatorial without eccentricity at 8075 km (MEO) |
| Repeating ground track | every 6 hours |
| Half-cone angle Θ_{3dB} | 0.7 deg |
| Peak antenna gain $G_{Tx,max}$ | 35 dB |
| Total available bandwidth | 2 GHz |
| Number of beamchannels | 200 with 10 MHz |
| Number of reuses | 4 including L/R polarization |
| Maximum number of beams | 800 |

Compared to the analysis in Section 5.9, the traffic volume of the customers in this Chapter is lower (22.7 compared to 198 Gbps). Hence, we also reduce the number of frequency reuses from 20 to 4 to have a similar utilization of the satellites. The other parameters are equal to the ones from the previous Chapter. Due to the reduction in reuses, the maximum number of beams is 800, with the 10 MHz beamchannels in contrast to 4,000 in Section 5.9.

8.2 Resource allocation for the baseline

The goal of this Section is to establish a baseline. We use the preprocessed input data about the user usage, gateways, and the MEO constellation from Section 8.1.

We follow the resource allocation process developed in Chapter 5 with the following adjustments:

- We provide each user with its beam making the first step of grouping the user terminals obsolete
- The routing is based on the balanced allocation algorithm
- The frequency assignment includes the return downlinks to the gateways assuming that all downlinks are in the same frequency band
- We compare three different power allocation scenarios: (1) fixed by design, i.e., constant per beam, (2) stationary for CIR, (3) dynamic for actual usage

We depict the resulting resource allocation solution in Figure 8-8. The upper part of the figures shows the users and gateways distributed around the world. The olive lines are active beams between satellites and gateways. The bottom plot shows the frequency assignment for the seven MEO satellites. As we have already seen in Section 5.9, satellites over the Pacific see less demand. This factor, together with the suboptimality of the algorithm results in only a 25% utilization of the frequency spectrum overall.

Using the solution of the first three steps of the resource allocation, we compute the power for three different scenarios:

1. **Fixed by design.** In this scenario, the satellite payload does not have any flexibility to adjust its power setting over time (traditional satcom). The design process determines the amount of transmitted power. We (optimistically) assume that the power is set to a constant level per beam meeting CIR at the worst case. The worst case is when the free space losses are the highest, which is when the user terminal and the satellite is separated the furthest. We assume in this scenario that beams can be activated and deactivated if not connect to a user.
2. **Stationary for CIR.** The second scenario considers a flexible payload where power is available as a pool. The satcom operator is conservative and provides the power levels that provided each user with their contracted CIR.
3. **Dynamic for actual usage.** In the last scenario, the satcom operator allocates dynamically the power that is needed by the user for any given time.

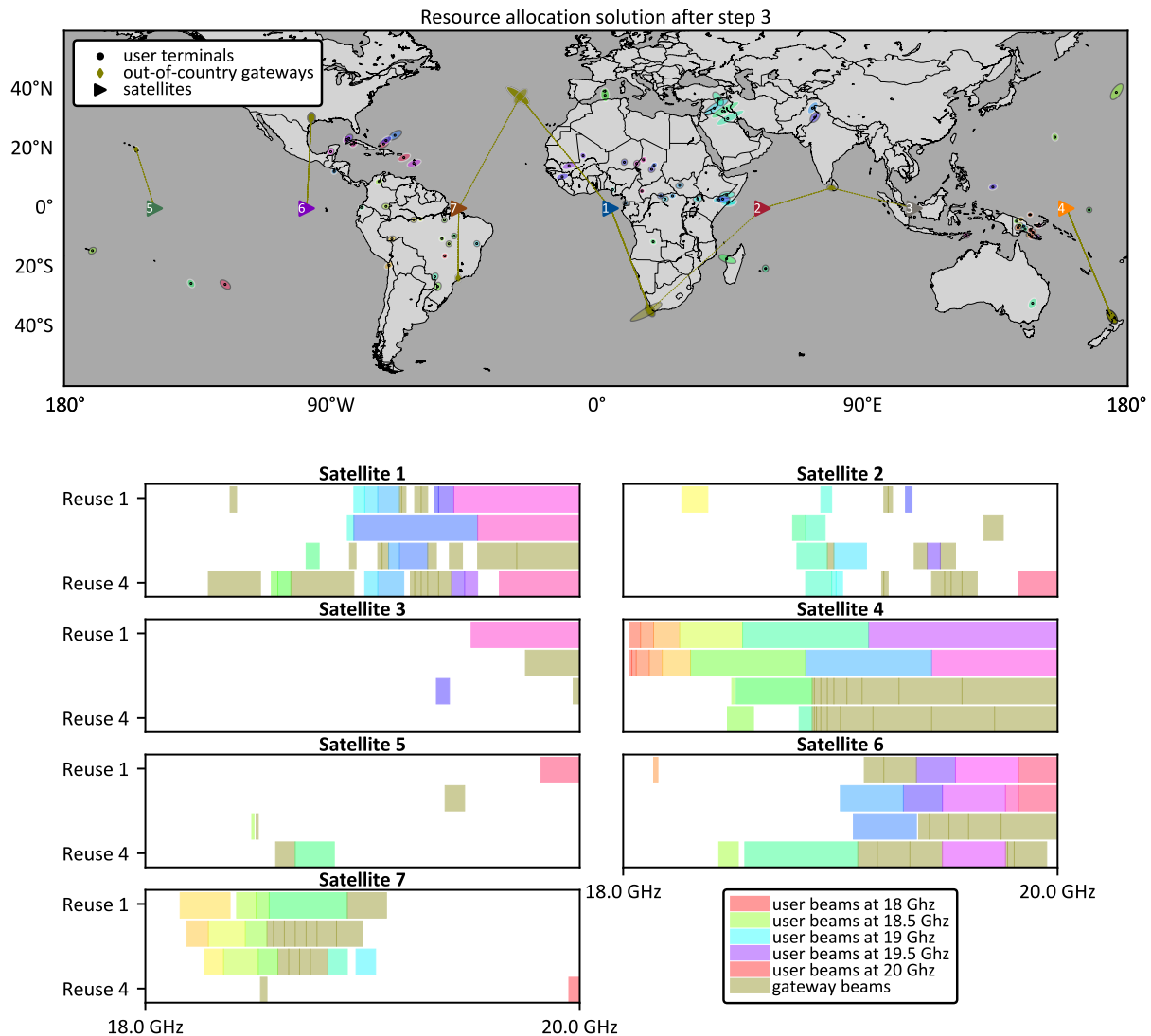


Figure 8-8: Solution after the first three steps of the resource allocation process for the MEO O3b mPower like constellation. The usage of the frequency spectrum is on average 25%.

We simulate the power levels for all three scenarios with a 1-minute granularity over a 24-hour window and record the results. Figure 8-9 summarizes the aggregation of the data for satellite #3 (fixed by design in green, stationary for CIR in blue). The results for the other satellites are the same but shifted in time based on their initial longitude. Each of the black vertical lines separates a 6-hour orbit.

We start with the first scenario: *fixed-by-design power allocation*. Since the allocation uses the CIR, the pattern repeats every orbit. The variation in the total power consumption is due to activating and deactivating beams. There are three main hot spots of high demand per orbit. The first one is at around Hour 00:30 over Papua New Guinea, the second one over Central America at 03:00, and the third one over

Europe at 04:00. We also compute the mean of the allocation over the orbit. The assumption behind is that the satellite’s batteries can compensate for fluctuations throughout the orbit, and therefore the mean is a meaningful measure to compare the different scenarios. Since the allocation for fixed by design does not change between orbits, the average per orbit stays the same as well. The average consumed power is 8.06 W across all 80 users over one orbit (see Table 8-5). We report in the table also the relative percentage of the numbers concerning this fixed by design allocation.

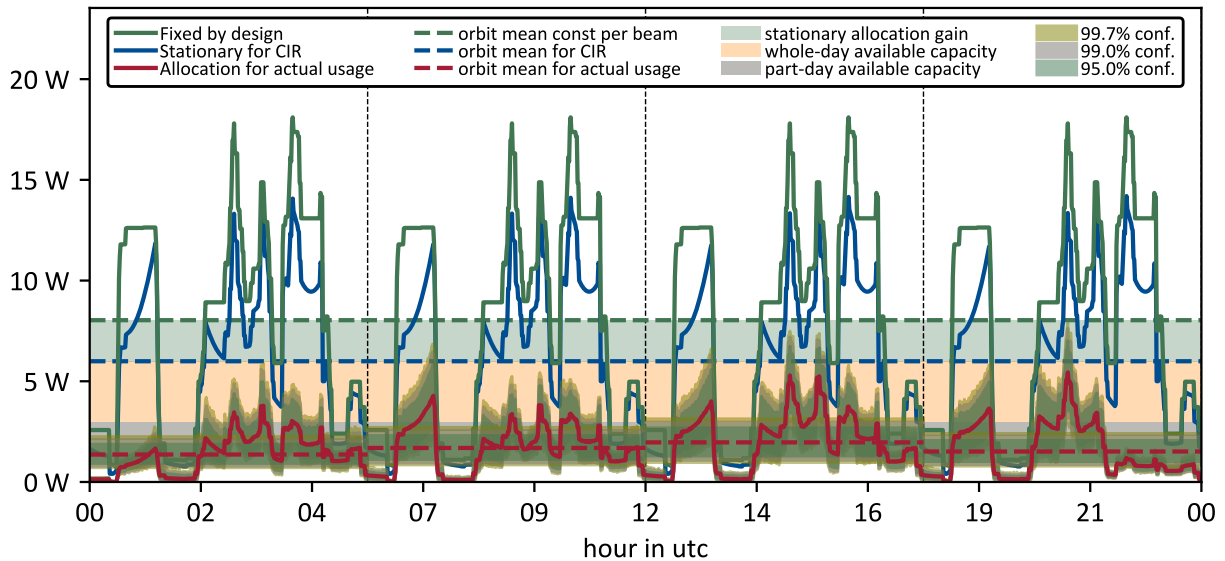


Figure 8-9: Power allocation for the three different scenarios. 1-minute granularity over 24 hours separated into four 6-hour orbits. The plot shows the power allocation for satellite #3.

Table 8-5: Summary of the mean orbit power consumption of the three different power allocation scenarios and delta capacities. The percentages are relative to the fixed by design allocation.

| | | Orbit 1 | | Orbit 2 | | Orbit 3 | | Orbit 4 | |
|---------------------|--|------------------|-----|-------------------|-----|--------------------|-----|--------------------|-----|
| | | $t \in [0h, 6h]$ | | $t \in [6h, 12h]$ | | $t \in [12h, 18h]$ | | $t \in [18h, 24h]$ | |
| The three scenarios | Fixed by design (capacity maximum) [W] | 8.06 | | 8.06 | | 8.06 | | 8.06 | |
| | Stationary for CIR [W] | 6.01 | 75% | 6.01 | 75% | 6.01 | 75% | 6.01 | 75% |
| | Dynamic for actual usage [W] | 2.10 | 26% | 2.55 | 32% | 2.97 | 37% | 2.28 | 28% |
| Available capacity | Stationary allocation lift from stationary for CIR to fixed by design [ΔW] | 2.05 | 25% | 2.05 | 25% | 2.05 | 25% | 2.05 | 25% |
| | Whole-day available capacity [ΔW] | 3.02 | 38% | 3.02 | 38% | 3.02 | 38% | 3.02 | 38% |
| | Part-day available capacity [ΔW] | 0.87 | 11% | 0.42 | 5% | 0.00 | 0% | 0.69 | 9% |

For the second scenario, the *stationary allocation for the CIR*, we see that it follows the same shape as the green line but is always below. Since the CIR does not change over time, the pattern repeats every orbit. The mean is 6.01 W, which is 75% of the maximum capacity, i.e., reducing the power usage by 25%. This is what we refer to as pure *stationary allocation lift*. The users receive their CIR independent of usage, but the flexibility in the power allocation adjusts for the geometrical varying free space loss. Figure 8-10 illustrates that further. It shows the allocated power for a single user.

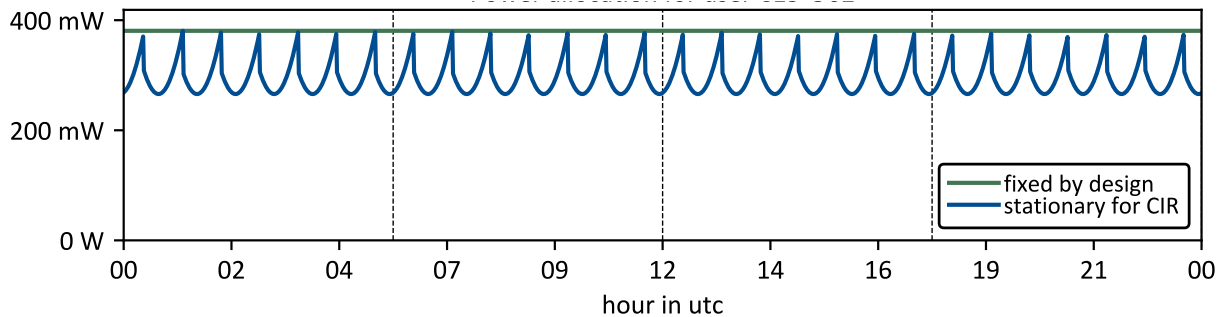


Figure 8-10: Comparison between the transmitted power on user level between the fixed by design scenario and stationary allocation for the CIR.

The fixed by design allocation cannot change the power of the beams over time (green line). Each U-shape in the blue line represents one satellite pass. There are seven passes per orbit (one with each satellite). The edges of the U are the positions where the beam gets handed over between two satellites and, therefore, has the most extended slant range. As a consequence, the free-space loss is high, resulting in a higher required power to close the link. At this point both, power allocation scenarios intersect. When the satellite propagates, the slant range becomes shorter, and therefore the link needs less transmitted power to close. The aggregation of this effect across all users results in the 26% stationary allocation improvement. The exact number depends on the geometrical properties of the communication system, such as the number of satellites, orbit altitude, and inclination. This lift is zero for GEO constellation since the geometry does not change over time (assuming stationary users).

The third scenario is the *power allocation for actual usage*. Since the traffic for each user is uncertain around a diurnal mean (see Figure 8-4), the aggregation on the satellite is uncertain as well. Therefore, we sample each timestep 10,000 times and create an empirical random variable to record the results (see Section 3.6.5 for more details). To keep the simulation computational tractable, we vectorize the second dimension of the link budget with the samples, i.e., for each satellite, the link budget is of the form $N_{samples} \times N_{users}$. Parallelization of the time dimension reduced computational costs further: 20 threads build the random variables within a couple of minutes.

In Figure 8-9, the red line is the mean, and the variation of the green colors are the confidence intervals. The general trend of the power allocation follows fixed by the design and stationary allocation for CIR scenarios. Nevertheless, the absolute numbers are smaller. Since the traffic has a diurnal variation, the orbital means vary throughout the day. For the means, we also compute the uncertainty around it. The most significant traffic is during the third orbit between 12:00 and 18:00 UTC with using 2.97 W or 37% of the maximum capacity (see Table 8-5). The lightest traffic is during the first orbit, using 26% of the maximum capacity averaged over the orbit¹³.

The lower power usage of the allocation for the actual usage scenario comes from the diurnal traffic pattern of the users. They do not use the CIR most of the time. Figure 8-11 further illustrates that by showing the CIR and actual usage of for an example user and the corresponding allocation scenarios. For the CIR allocation, the provided data rate remains constant (the required power to support that data rate is changing, as shown in Figure 8-10). It is slightly above the CIR due to the MODCOD discretization and the link budget margin (these errors are per link and aggregate on satellite level without canceling each other out). In the allocation for the actual usage scenario the provided data rate follows the requested traffic closely. Compared to the dynamic CIR allocation, the reduction results in part-day and whole-day available capacity, as discussed in the paragraph below.

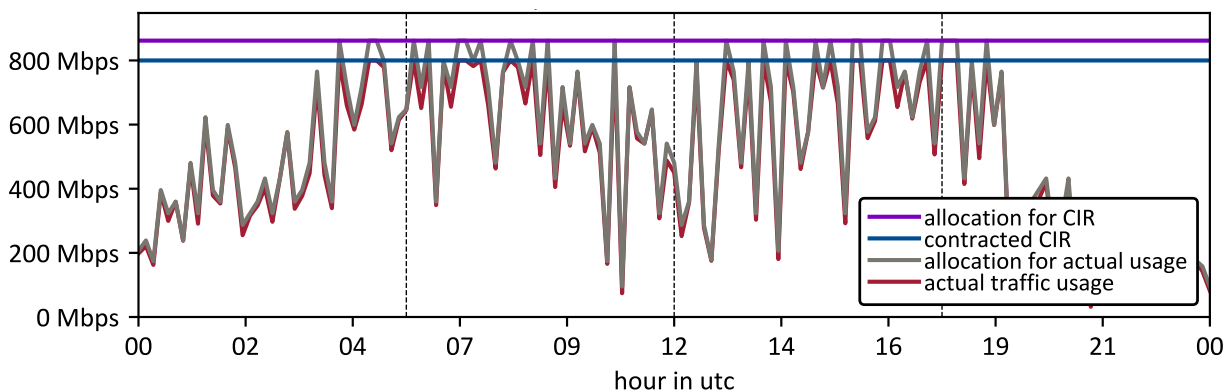


Figure 8-11: Comparison of the data rate allocated and requested between the stationary allocation for CIR and for actual usage. The curves for actual usage are plotted every 10 minutes to avoid cluttering.

In contrast to Section 3.6, we compute the available whole-day and part-day capacity based on a per orbit basis (to account for the smoothening effect of on-board batteries). We compute the whole-day capacity based on the mean orbital distribution for which the actual usage is the highest. That is in the example

¹³ The UTC Hour of the maximum and minimum traffic is obviously different for each satellite.

orbit 3, resulting in 3.02 W available capacity (or 38% of the maximum capacity). This number is close to the 36% obtained from our toy problem in Section 3.6. During the orbits for which the traffic is lower part-time available capacity becomes available (see Table 8-5). The maximum is during the first orbit with 0.87 W or 11%. Compared to the toy problem and the conceptual plot in Figure 7-3, the part-time available capacity is lower. The reason for that is that the constellation in this Section is NGSO, and we compute the part-time capacity on a per orbit basis. Both factors have a smoothing effect on power consumption.

In sum, the results indicate that stationary allocation for CIR reduces the capacity usage by 25% (100% - 75%) compared to the fixed by design allocation, while dynamic allocation for actual traffic reduces it by 63% (100% - 37%). The following Section discusses our optimization approach for the translation of the 38% available whole-day capacity and up to 11% part-day capacity into additional revenues.

8.3 General pricing optimization approach

This Section describes the elements of the RM optimization that are common across all five analyses of this Chapter. In contrast, the specifics of each optimization are described in the analyses Sections 8.4 - 8.5.

The demand and resource management part of the satcom RM framework needs to exchange information (see Figure 8-1). The pricing optimization computes new prices that updates the expected traffic. The resource management allocates resources for this new traffic and returns the used and available capacity to the demand management part. Hence, the pricing algorithm calls the resource allocation simulation every iteration.

Figure 8-12 depicts the process that is common across all analyses. We initialize the optimization with the baseline from Section 8.2. The iteration starts with an estimation of the customer and market price elasticities. For the market share analysis in Section 8.5.2, we evaluate the competitive environment at this stage.

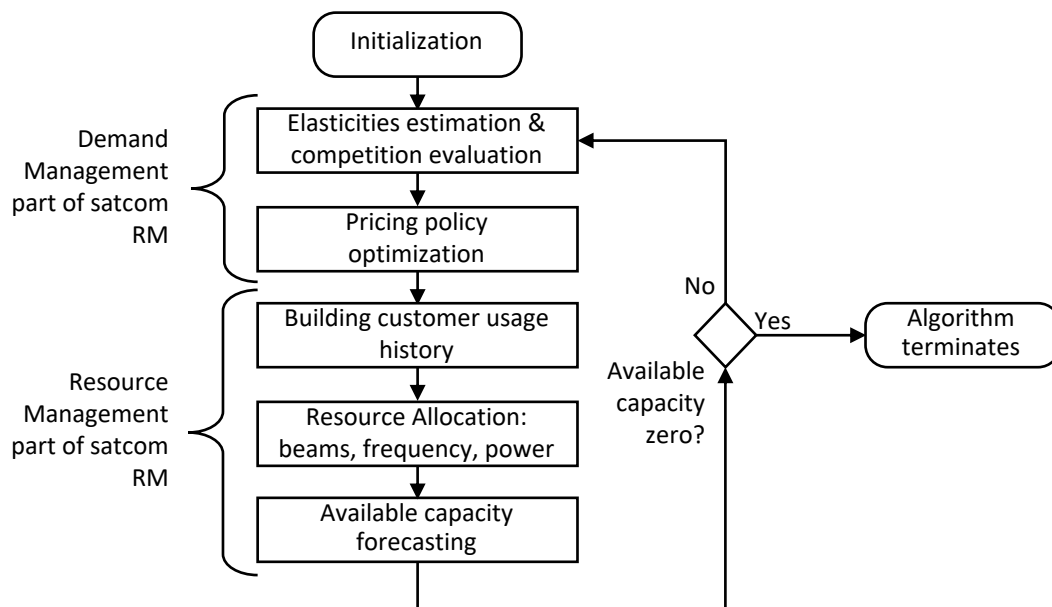


Figure 8-12: General approach for integrated optimization of pricing and resource allocation.

The results inform the pricing policy optimization, which computes its next iteration of prices. Based on these, the optimization updates the expected traffic volume and builds customer usage history. Depending on the significance of the change in the traffic, the resource allocation adjusts the power levels or additionally revises the frequency plan. Based on the results, the available capacity forecaster determines the used capacity and subtracts it from the total capacity. If the available capacity is zero (or

as close as possible given the discretization), the algorithm terminates. Otherwise, the next iteration begins with the elasticities' estimation and evaluation of the competition.

8.3.1 Sorted list of prices and binary search

Due to the monotonic behavior of the price elasticities and the link budget, lowering prices does not decrease demand, and increasing prices does not increase demand. With this characteristic, we can build a *sorted list*, which is, when sorted by prices, also inversely sorted by demand, and therefore by used capacity (lower price yields more demand, more used capacity, less available capacity). Table 8-6 illustrates a schematic example for a satellite with single user i . The maximum capacity is 6 Watts, bandwidth 50 MHz, and C_{used} is computed proportionally with the Shannon limit according to $C_{used} = 2^{(R_i/50MHz)}$. The bold line indicates the price for which the available capacity is minimized (and revenues maximized if demand is elastic).

Table 8-6: schematic example of the ordered list of price p_i and the resulting demand R_i for a single user i (elasticity is -2 at $p_i = 200$ and $R_i = 100$). Used capacity computed through the Shannon limit and available capacity with a maximum capacity C_{max} of 6 Watts.

| p_i [\$/Month/Mbps] | R_i [Mbps] | C_{used} [W] | C_{avail} [W] |
|--------------------------|-----------------|-------------------|--------------------|
| 283 | 50 | 2.00 | 4.00 |
| 258 | 60 | 2.30 | 3.70 |
| 239 | 70 | 2.64 | 3.36 |
| 224 | 80 | 3.03 | 2.97 |
| 211 | 90 | 3.48 | 2.52 |
| 200 | 100 | 4.00 | 2.00 |
| 191 | 110 | 4.59 | 1.41 |
| 183 | 120 | 5.28 | 0.72 |
| 175 | 130 | 6.06 | -0.06 |
| 169 | 140 | 6.96 | -0.96 |
| 163 | 150 | 8.00 | -2.00 |

Building the list of declining prices is the outcome of the pricing policy optimization and computationally orders of magnitude cheaper than the mapping between prices and used capacity. Since the list is sorted, we use a *binary search* to find the element where the available capacity is zero, i.e., the bold line in Table 8-6. An element comprises a set of prices for each customer. The logarithmic complexity with the length of the list N_l is $\mathcal{O}(\log N_l)$. Hence the algorithm only makes a few evaluations of the available capacity for a longer array of prices (e.g., three comparisons for $n = 1,000$, and four for $n = 10,000$).

Formally, we define the ordered list \mathcal{L} with the length N_l where p_j is the price for user j .

$$\mathcal{L} = \{l_i = (p_j \forall j \in 1 \dots N_u)\} \forall i \in \{1 \dots N_l\} \quad (8-3)$$

With that definition, we outline the pseudo-code for a recursive binary search algorithm [239] in Algorithm 8-1. Line 4 is computationally the most expensive part of the algorithm since it involves a rerun of the resource allocation process.

Algorithm 8-1: Pseudo code for the recursive binary search algorithm for pricing optimization

Input: \mathcal{L} with declining prices, i.e., increasing capacity
Output: l

```

1: def binarySearch( $\mathcal{L}, lb, ub$ ):           // function definition with lower and upper bound (lb, ub)
2:    $mid = 1 + (ub - lb) // 2$                 // floor division to obtain the middle index
3:   if  $ub \geq lb$  do
4:      $C_{avail} = availableCapacity(\mathcal{L}[mid])$  // compute available capacity for set of prices
5:     if  $C_{avail} == 0$  do                  // if available capacity zero, return current prices
6:       return  $\mathcal{L}[mid]$ 
7:     else if  $C_{avail} > 0$  do             // if available capacity larger zero, search to the right
8:       return binarySearch( $\mathcal{L}, mid, ub$ )
9:     else do                               // if available capacity smaller zero, search to the left
10:      return binarySearch( $\mathcal{L}, lb, mid$ )
11:    end if
12:  else
13:    return  $\mathcal{L}[mid]$ 
14:  end if

```

8.3.2 Algorithmic approach for computing the ordered list of prices

For the computation of the ordered list of prices, we compare mainly two algorithms (for some analyses, we consider additional approaches). The first algorithm is an equally decreasing heuristic, which might be considered as closest to the current manual approach. When capacity is available, the operator discounts the CIR prices for all customers by the same percentage points Δp . The second algorithm is a gradient optimizer based on marginal revenues. It considers the complete chain from price elasticities to link budgets.

8.3.2.1 Equally decreasing heuristic

This heuristic is an algorithm that does not consider any demand or resource aspects. Algorithm 8-2 outlines its pseudo code. The hyperparameter inputs are lower and upper bounds on the price $p_{j,min}$ and $p_{j,max}$ and the increment of the price discount Δp . Additionally, we provide the number of users N_u . The outcome of the procedures is a sorted list of sets of prices. Line 1 initializes the list with the maximum

prices. Then the prices are reduced for the users who have not reached the minimum. The algorithm terminates if all users are at their minimum.

Algorithm 8-2: Pseudo code for the equally decreasing heuristic

Input: $p_{j,min}, p_{j,max}, \Delta p, N_u$
Output: \mathcal{L}

- 1: **Init** $\mathcal{L} = \{l_0 = (p_{j,max} \forall j \in 1 \dots N_u)\}$ // Init with highest price
- 2: **while** $any(p_j > p_{j,min} \forall j \in 1 \dots N_u)$ **do** // terminates when all prices are below the minimum bound
- 3: $\mathcal{N}_{update} = \{p_j > p_{j,min} \forall j \in 1 \dots N_u\}$ // get indexes for which prices are still to be reduced
- 4: $\mathcal{L} = \mathcal{L} \cup \{(p_j \cdot (1 - \Delta p) \forall j \in \mathcal{N}_{update})\}$ // append the new price update to the ordered list
- 5: **end while**

8.3.2.2 Gradient optimizer based on marginal revenues

The idea behind this algorithm is to build on the gradient-based approach we introduced in Section 3.6.9. We know that for monotonically decreasing functions, this algorithm is (incremental) optimal [96]. Our objective is to leverage this property and transform the model accordingly. In contrast to the toy problem from Section 3.6, for the baseline in this Chapter, we can no longer find an analytical expression for the revenue gradient concerning the power (see Eq. (3-17)). Hence, we follow a three-step method. First, we extract samples from the precomputed cache to compute revenues as a function of the capacity $\Pi(C)$. The second step is to fit an analytical function through $\Pi(C)$. Finally, we analytically compute the gradient $\partial\Pi/\partial C$ for each user and prove that the resulting marginal revenues are indeed monotonically decreasing. The following details each step before we describe the adapted optimizer.

Samples for revenues as a function of capacity $\Pi(C)$

We precompute a cache that has the mapping of power to the price for each user, i.e., $C(p)_i$. We inverse this mapping to obtain $p(C)_i$. The price elasticity provides a functional relationship between price and data rate $p(R)$ (as well as the inverse $R(p)$). With that inverse, we compute the revenues as a function of the rate by $\Pi(R(p(C)) \cdot p(C))$. We use this equation to transform the $C(p)_i$ samples into samples of $\Pi(C)_i$.

Regression model of revenues as a function of capacity $\Pi(C)$

Since we want to have an analytical, monotonically decreasing expression for $\partial\Pi/\partial C$, we perform a regression on the samples. We find that a second-order power function of the following form yields the best results for elastic demands. For unit elasticity a linear regression has better fitting quality.

$$\Pi(C) = a \cdot C^b + c \quad (8-4)$$

We leverage a non-linear least-square optimization to compute the parameters that best fit this function to the samples for each level (depending on the analysis, can be user or segment). Figure 8-11 shows an example of $\Pi(C)$ with the original sample points and the fitting functions with an average R^2 of 0.997, a min of 0.979, and a maximum of 1.00.

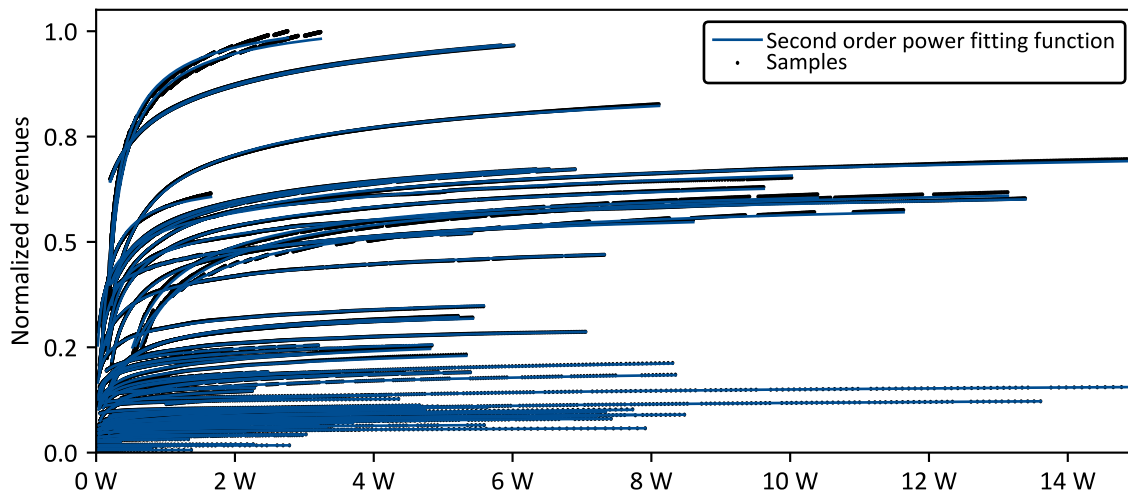


Figure 8-13: Resulting fit function in blue of the sample points in black for each user

As expected, the function $\Pi(C)$ has a diminishing return due to two factors. First, the link budget is not linear (marginally smaller data rate for each additional unit of power). Second, the elasticity of demand (here -1.5), which reduces the additional revenues generated by lowering prices for additional data rates. Depending on the combination, the curves have different shapes. The higher the non-linearity of both factors, the sharper the bend in the curve. In this example, each curve starts at the power level defined by the baseline and ends when the link reaches its theoretical limit with the given frequency allocation (or in other words, goes from higher to lower price, or smaller demand to higher demand). The majority of the fitted functions are visually not distinguishable from the samples (lower left area). The most challenging users are the ones with higher throughput shown in the upper left area of the figure. Towards higher powers, the fit underpredicts the samples slightly. Nevertheless, we consider the regression of appropriate quality as its purpose is to guide the algorithm, and we use the simulator to compute the exact numbers for the final solution.

Monotonically decreasing property of the marginal revenues $\partial\Pi/\partial C$

The marginal revenue is the analytical derivation of Eq. (8-4) with the parameters a , b and c known from the regression:

$$\frac{\partial \Pi(C)}{\partial C} = a \cdot b \cdot C^{(b-1)} \quad (8-5)$$

To ensure that the gradient optimizer finds the global optimum, we prove that the function is monotonically decreasing within the proper bounds. As the function is one dimensional, we compute the second derivation and show that it is negative for every point between the bounds:

$$\frac{\partial^2 \Pi(C)}{\partial C^2} = a \cdot (b - 1) \cdot C^{(b-2)} < 0 \quad (8-6)$$

We numerically prove this equation by sampling sufficiently and take the maximum of the recorded results. For the example, all marginal revenues are monotonically decreasing.

Optimization algorithm

The optimization algorithm is an incremental gradient approach. Algorithm 8-3 exhibits the pseudo-code of the implementation. In addition to the inputs of the equally increasing heuristic from Algorithm 8-2 (the hyperparameters of lower and upper bounds on the price $p_{j,min}$ and $p_{j,max}$ and the increment of the price discount Δp), we pass the marginal revenues for, depending on the analysis, each user or segment $J_i(C_i)$. Line 2 and 7 initialize the power as a function of the new prices (fetched from the precomputed cache) and evaluates the marginal revenues at these points. The optimization while loop has the same termination criteria as the heuristic. However, in addition, if all marginal revenues are no longer positive, the algorithm stops as well. Line 5 identifies the user i with the steepest gradient and adjust the price p_j by the increment Δp .

Algorithm 8-3: Pseudo-code for the gradient optimizer based on marginal revenues. Prices are assigned to user or segment depending on analysis type.

```

Input:  $p_{j,min}, p_{j,max}, \Delta p, N_u, J_i(C_i) = \partial \Pi_i / \partial C_i$ 
Output:  $\mathcal{L}$ 
1: Init  $\mathcal{L} = \{l_0 = (p_{j,max} \forall j \in 1 \dots N_u)\}$  // Init with highest price
2: Init  $J_j(c_j) \forall c_j \in \mathcal{C}(l_0)$  // compute the marginal revenues
3: while any( $p_j > p_{j,min} \forall j \in 1 \dots N_u$ ) do // terminates when all prices are below the minimum bound
4:   if any  $J_i(c_i) > 0$  then // check if any gradient positive
5:      $i = \text{argmax}(J_j(c_j))$  // get user  $i$  with the steepest gradient
6:      $p_j = p_j \cdot (1 - \Delta p)$  // Adjust price  $j$  by  $\Delta p$ 
7:      $l = (p_j \forall j \in N_u)$ 
8:      $J_j(c_j) \forall c_j \in \mathcal{C}(l)$ 
9:      $\mathcal{L} = \mathcal{L} \cup l$  // append the new price update to the ordered list
10:  else
11:    stop // stop when all gradients non positive
12: end while

```

We discussed our approach to pricing optimization in this Section. Since the resource allocation is computationally expensive, an ordered list of prices is computed by either a decreasing heuristic or gradient based. A binary search on this list results in a tractable number of evaluations even for larger lists > 1,000. The next two Sections 8.4 and 8.5 use this approach to compute results for selling the additional capacity to existing and new customers, respectively.

8.4 Monetizing the available capacity through existing customers

By filling the capacity with existing customers, their demand needs to increase. For that, we consider two options: selling more quantity through the existing SLA and selling additional products. The following two Sections discuss them, for which we make the following assumptions:

- Demand at the same price remains constant throughout the simulation, with no growth nor shrinking.
- Price does not affect the entrance of new users (no affordability elasticity, considered in Section 8.5.1)
- Customers do not switch operator when prices are changed (no competition, considered in Section 8.5.2)
- Elasticity is known per customer (discussed in Section 8.6)
- Price is the only lever to control demand.
- Separate prices can be charged per customer.
- Recomputing the frequency plan does not improve the result and power adjustments are sufficient for the changes in demand (required for Sections 8.5.1 and 8.5.2)

Each of the two following Sections first lists the specific approach and the assumptions, then discusses the results, and finishes with conclusions and limitations.

8.4.1 Selling more capacity through existing SLAs

Specific approach and assumptions

As discussed in Section 7.1, the vast majority of current customers are contracted by the Classical SLA with a price for the CIR. Hence, we assume that all customers have this contract type, with the baseline being their current price points for the CIR. Selling more quantity of the Classical SLA translates into contracting a higher CIR. We make the assumptions here that the usage behavior of users changes slower than users' SLA are re-negotiated to allow for safe oversubscriptions. If an operator has 80 users with an average SLA duration of 2 years, then on average, a little over one SLA expires per week. If the total traffic of the users does not increase more than the SLA's CIR per week, then the operator can safely oversubscribe.

We model the dependency between the CIR and the price with a price elasticity for the demand-for-use. Since most terminals have their CIRs, we compute the elasticity curve on a user terminal level. We use the same log-linear relationship from Eq. (8-1), where the current price point p_{CIR} for the CIR determines the parameter a through the following equation:

$$a = CIR/p_{CIR}^b \tag{8-7}$$

Figure 8-14 illustrates three demand elasticity (-2, -1, -0.5) for an example customer with a CIR of 100 Mbps for a price of \$200/month/Mbps. For the unit elasticity -1, reducing prices to \$100 increases the demanded data rate to 200 Mbps (and therefore no change in revenues). If the demand is more elastic with -2, the same price leads to twice the demanded data rate, 400 Mbps. The opposite holds for an inelastic demand of -0.5; the CIR goes down to 141.5 Mbps.

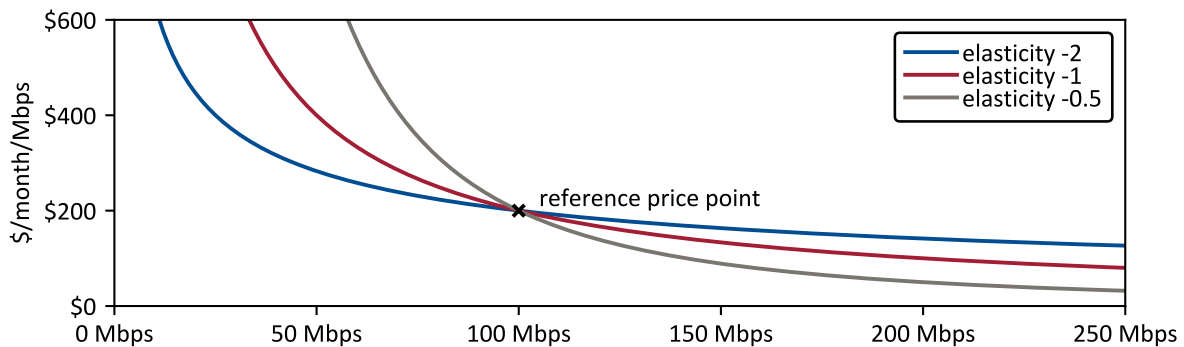


Figure 8-14: Example of a -2, -1, and -0.5 price elasticity for a customer with a 100 Mbps and \$200/month/Mbps

For the analysis, we compare three different elasticity scenarios. For two of them, all customers have the same elasticity that is either elastic with -2, or unit elastic with -1. In the third case, the elasticity is different for each customer varying between -2 and -1. The assignment of elasticity to a customer is random from this interval.

Discussion of the results

We start the discussion of the results on the operator level with Figure 8-15 and then go the customer level to understand the behavior of the algorithms and the results in more detail. Figure 8-15 shows the monthly revenues (normalized to the baseline), the average price (normalized to the average 2019 price, see Figure 8-5), and the total contracted CIR capacity. For each of these metrics, we compare the two algorithms with the baseline across the three elasticity cases. The percentage above the red bar is the relative improvement of the gradient approach compared to the heuristic.

For the unit elastic case, the monthly revenues do not change between the different algorithms. That is ordinary and validation of the simulation, since we expect that behavior from a unit elastic demand. Compared to the baseline, the gradient optimization does not make changes to the pricing. At the same time, the equal decreasing heuristic decreases average prices across all customers by 32%, resulting in 35% more sold capacity.

For the elastic cases, -2, and random between -2 and -1, the gradient optimization results in higher revenues gains. The 38% available capacity is translatable into 27% more revenues compared to the baseline for -2. This number is 18% for the random elasticity between -2 and -1 (which is on average around -1.5). The gradient approach can contract 13-14% more total capacity than the heuristic. The improvement from a gradient RM optimization compared to the heuristic pricing approach is between 5-7% depending on the elasticity.

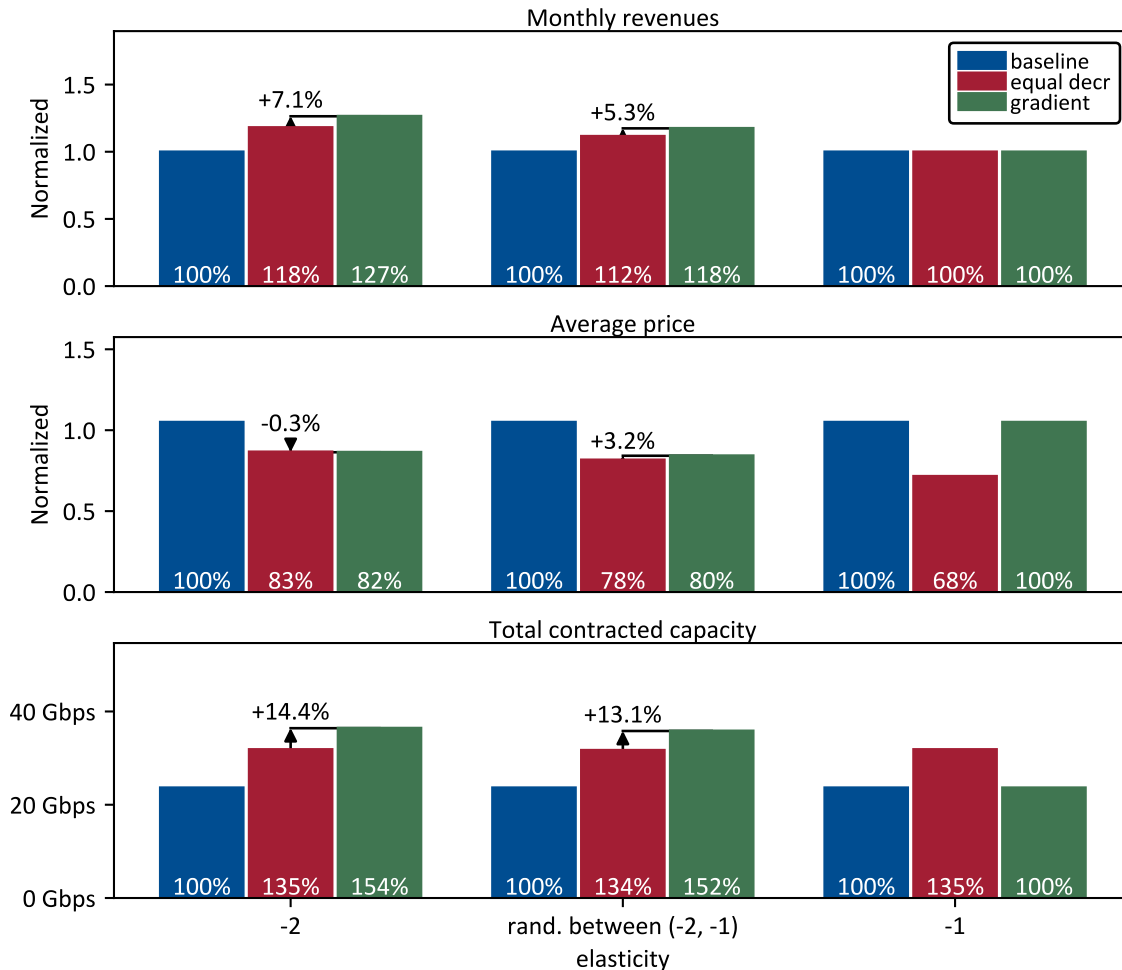


Figure 8-15: Monthly revenues, average price, total contracted CIR capacity for the three elasticity cases compared between the baseline and the two pricing algorithms.

In the remainder of the result discussion, we aim to understand why the gradient optimization outperforms the heuristic. In that, we start with Figure 8-16, which shows the normalized prices for the elastic case. The blue bars are the baseline; the red and green ones are the heuristic and gradient approach. Compared to the baseline, the heuristic lowers equally the prices as expected. Note that there are exceptions, e.g., users 16-24 with the same initial price but slightly different prices after the equal

decreasing heuristic. The reason for this that the link reaches its limit, and therefore no higher data rate is supported.

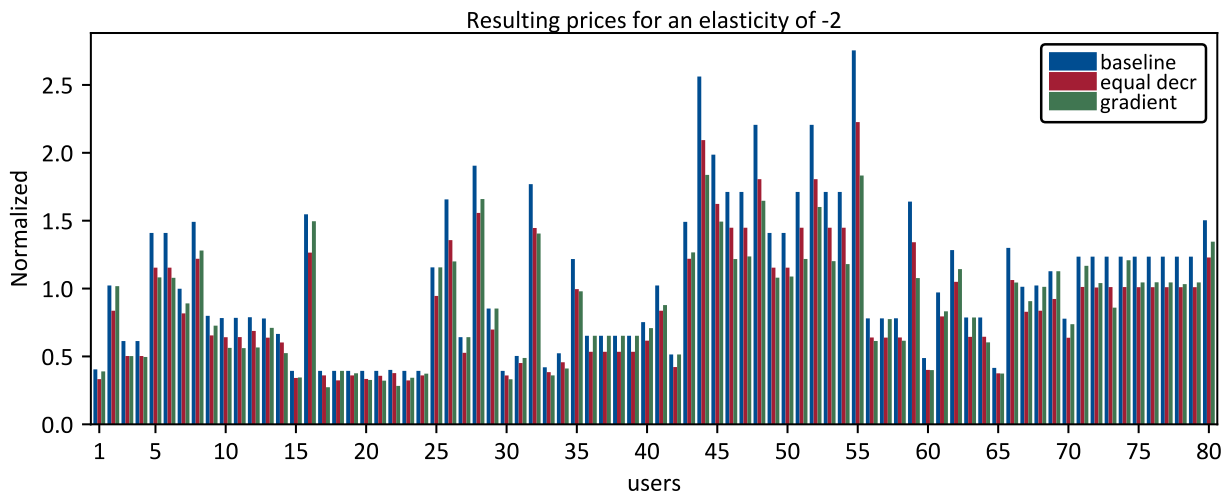


Figure 8-16: Comparison of the prices between the two algorithms and the baseline for the elastic case with -2.

For the gradient approach, we cannot identify a pattern for the price reduction based on the figure. Some users are treated equally, such as 75-79, while 71-74 show different prices. In one case of 60 and 65, the gradient solution matches the heuristic. And in another case, the gradient algorithm keeps prices at the baseline level (36-39).

To further understand why the algorithm makes these pricing decisions, we build a pair-wise correlation between seven selected attributes. These are relative price reduction in percentage points, terminal size, baseline price, customer group (A, B, C, D encoded as 0, 1, 2, 3), longitude, CIR, latitude, and traffic uncertainty. Table 8-7 shows the resulting table for the elastic case, comparing the gradient solution with the baseline. Positive correlations are formatted red and negative correlations blue. We order the attributes descending according to their absolute correlation with the relative price reduction.

The segment type moderately influences the relative price reduction and shows similar correlation with the terminal size and the baseline price. The longitude correlation is unexpected and seems to be mainly driven by the segment type and is more an outcome than a cause of the pricing. Surprisingly, the latitude shows little correlation with the relative price reduction while it interacts with the baseline price. Given the higher cost of serving customers on higher latitudes, our initial expectation was a stronger correlation. However, the correlation with the segment type indicates that latitude is similar to the longitude and outcome, not a cause. As the last attribute, the traffic uncertainty does not correlate with the pricing and only a moderate negative one with the terminal size and the CIR as rationalized before.

Table 8-7: Pair-wise correlation between selected attributes. Red colored fields are positive correlation, blue ones are negative correlation. Values are computed for the elastic -2 case comparing the gradient with the baseline.

| | relative price reduction | terminal size | baseline price | customer group (A, B, C, D) | longitude | CIR | latitude | traffic uncertainty |
|-----------------------------|--------------------------|---------------|----------------|-----------------------------|-----------|------|----------|---------------------|
| relative price reduction | 1.0 | 0.5 | 0.5 | 0.4 | -0.2 | 0.2 | 0.2 | 0.0 |
| terminal size | 0.5 | 1.0 | -0.1 | 0.4 | 0.1 | 0.8 | 0.1 | -0.2 |
| baseline price | 0.5 | -0.1 | 1.0 | 0.4 | -0.3 | -0.4 | 0.3 | -0.0 |
| customer group (A, B, C, D) | 0.4 | 0.4 | 0.4 | 1.0 | -0.3 | -0.0 | 0.4 | 0.0 |
| longitude | -0.2 | 0.1 | -0.3 | -0.3 | 1.0 | 0.2 | -0.2 | -0.0 |
| CIR | 0.2 | 0.8 | -0.4 | -0.0 | 0.2 | 1.0 | -0.2 | -0.3 |
| latitude | 0.2 | 0.1 | 0.3 | 0.4 | -0.2 | -0.2 | 1.0 | -0.1 |
| traffic uncertainty | 0.0 | -0.2 | -0.0 | 0.0 | -0.0 | -0.3 | -0.1 | 1.0 |

Conclusions and limitations

In this Section, we studied how the operator can sell the additional capacity to the same customer through the same SLA. We compared two algorithms, a heuristic, and a gradient optimization across three elasticity cases: elastic (-2), elastic range (from -2 to -1), and inelastic. From the simulation, we conclude the following main points:

- The 38% available capacity is translatable into 0-27% additional revenues depending on the elasticity. More elastic demand results in higher improvements.
- The gradient optimization outperforms the heuristic approach by 5-7%.
- The price reduction is greater for users with a larger terminal and a higher baseline price because of more beneficial capacity costs. Latitude shows no correlation.

Besides the assumptions outlined at the beginning of Section 8.4, the analysis revealed some further limitations of the simulation:

- The gradient optimization is only stable for elastic demand; for inelastic demand, the optimum is to sell the smallest increment of capacity. The log-linear price elasticity function that we use has the property that the elasticity is constant at all points on the curve. Other fitting functions with

varying elasticity as a function of the price might be used for better representations of the extremes.

- The assumption for the simulation that all power adjustments are sufficient limited the demand increase for a few users since more demand would require additional frequency spectrum (e.g, user 19 in Figure 8-16). Including a re-computation of the frequency plan in the simulation could potentially improve the optimized revenues.
- We use the heuristic approach as an approximation of how an operator might adjust their prices without using RM. However, this assumption is undoubtedly an oversimplification as an operator would base their decision on multiple factors and does not discount all users equally. Nevertheless, similar to the airline industry before the implementation of RM, these decisions are based on heuristic and expert judgment. As we saw, understanding why the gradient algorithm sets the prices it does is challenging to grasp and hence tricky to translate into heuristics.

8.4.2 Selling additional products

Specific approach and assumptions

In contrast to the previous Section 8.4.2, this Section discusses the selling of the available capacity through additional products. These products can be novel SLAs, as discussed in Chapter 7. One example here is the spot instance that an operator might want to use an *add-on*. It gives the customer additional flexibility to purchase capacity, which is especially valuable for customers who cannot predict their traffic well in advance. We leave the existing SLAs untouched in this Section, and therefore, assume that the products are priced and differentiated accordingly such that no internal cannibalization occurs.

One central question is how operators can price these new products and what the demand elasticity is. Since almost no information is available, we model the price point and elasticity explicitly with three parameters (two for the price point and one for the elasticity) and conduct a broad sensitivity analysis.

The first parameter that we introduce is the fraction of revenues $k_{rev,fract}$ that the additional product delivers compared to the existing Classical SLA. The second parameter is the adjustment parameter k_{adj} that shifts the reference point through which the elasticity curve is defined (see Figure 8-17). The following three Eqs. (8-8) - (8-10) describe the mathematical relationships between these two parameters, and the existing SLA attributes p_{CIR} and CIR. The result is the reference point defined by $p_{ref,add}$ and $R_{ref,add}$. We take the route of $k_{rev,fract}$ in the first two equations, so that $k_{rev,fract}$ affects the additional revenues in Eq. (8-10) linearly.

$$p_{ref,add} = p_{CIR} \cdot k_{adj} \cdot \sqrt{k_{rev,fract}} \quad (8-8)$$

$$R_{ref,add} = CIR \cdot \frac{1}{k_{adj}} \cdot \sqrt{k_{rev,fract}} \quad (8-9)$$

$$\Pi_{add} = p_{ref,add} \cdot R_{ref,add} = k_{rev,fract} \cdot p_{CIR} \cdot CIR \quad (8-10)$$

To capture the range of possible prices for the new products, we set up nine combinations of the two parameters, $k_{rev,fract}$ and k_{adj} (see Table 8-8).

Table 8-8: Test plan for testing of the sensitivity to the reference price point of the additional product. Each combination is simulated for four elasticity cases: -2, random between -2 and -1, -1, and -0.5.

| | Comb. 1 | Comb. 2 | Comb. 3 | Comb. 4 | Comb. 5 | Comb. 6 | Comb. 7 | Comb. 8 | Comb. 9 |
|-----------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $k_{rev,fract}$ | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 | 0.5 | 0.5 | 0.5 |
| k_{adj} | 0.2 | 0.5 | 1 | 0.2 | 0.5 | 1 | 0.2 | 0.5 | 1 |

The fractions of revenues ranges in three steps: 10%, 20%, and 50%. We take the median step, 20%, from analogy to Amazon’s pricing of their AWS spot instance (the price was 20% of the longer-term commitment for the sample we took (see Section 7.4)). The second parameter k_{adj} has the steps: 0.2, 0.5, and 1.

Figure 8-17 provides a visual understanding of the impact of different values, including the elasticities. The black dashed line is the unit elasticity from the previous Section’s example in Figure 8-14 with its reference price point of \$200/Mbps/month for 100 Mbps. We set $k_{rev,fract}$ to 20% and vary the elasticity from -0.5 to -2.

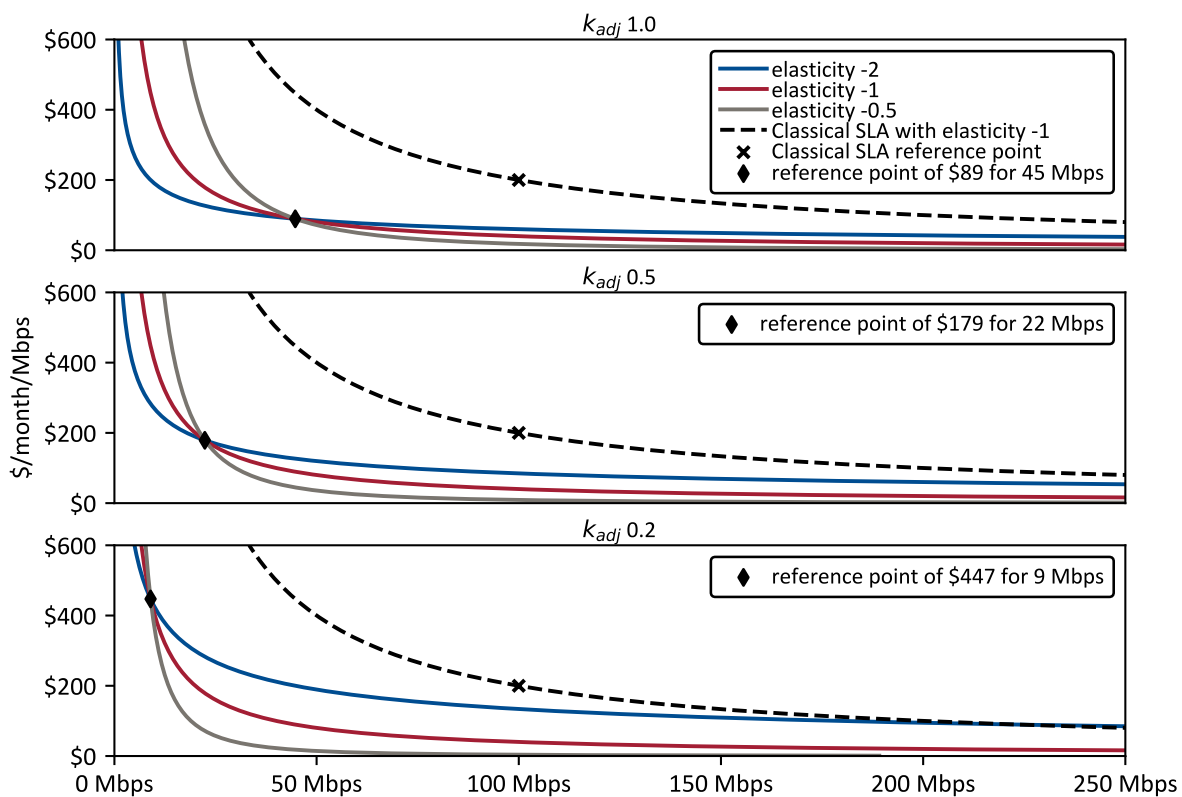


Figure 8-17: Behavior of the price elasticity function as a function of the reference point and elasticity parameter. The plots are shown for a price 20% that of the Classical SLA for the unit elasticity.

By definition, the red unit elasticity lines are *independent* of the parameter k_{adj} . That can be seen through Eq. (8-7). For $b = -1$, the equation for a becomes $a = CIR \cdot p_{CIR}$, and since k_{adj} is in the numerator for p_{add} and in the denominator for R_{add} (see Eqs. (8-8) . (8-9)), k_{adj} cancels out. However, elasticities different from the unit elasticity are *dependent* on k_{adj} . A higher value implies that the inelastic effect occurs at higher Mbps, whereas the elastic effect appears at a lower price (can be seen at the behavior of

the blue line for large Mbps numbers). This behavior might be a good approximation for a product that a larger quantity is sold for a lower price. In contrast, for a lower value of k_{adj} , the inelastic case constrains the Mbps strongly while in the elastic case, the operator can sell larger quantities for a higher price. An example product has an initial high price for a smaller quantity.

Discussion of the results

We simulate each of the nine combinations from Table 8-8 with four different elasticity cases: elastic with -2, random between -2 and -1, unit elastic at -1, and inelastic with -0.5. The results are reported for the baseline as well as for the decreasing heuristic algorithm and gradient optimization.

Figure 8-18 displays the results in the same format as Figure 8-15 for one of the nine combinations with $k_{rev,fract} = 0.2$ and $k_{adj} = 0.5$.

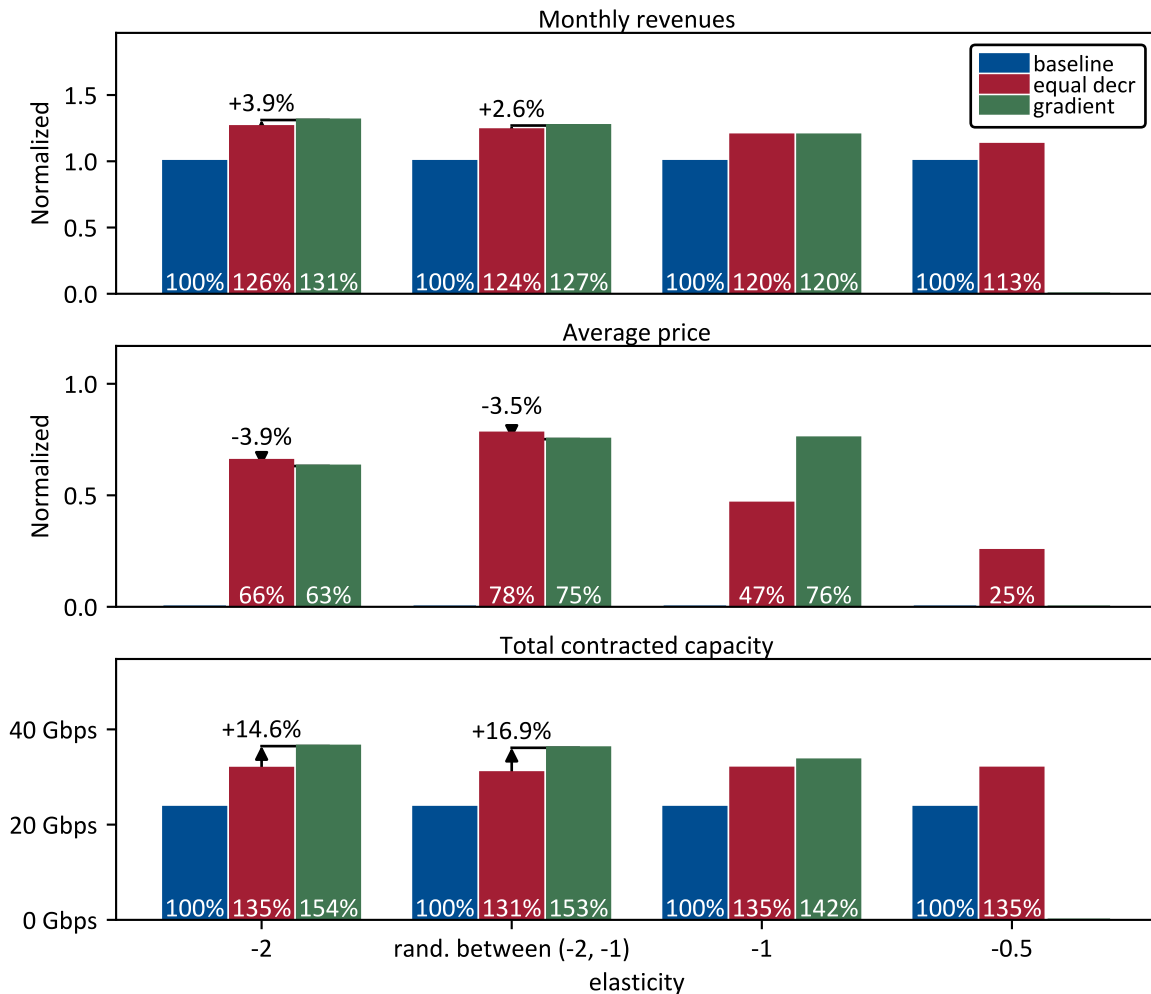


Figure 8-18: Results for the combination $k_{rev,fract} = 0.2$ and $k_{adj} = 0.5$.

The bar plot compares the two algorithms with the baseline, and the percentage points above the red bar point out the lift from the gradient optimization. The baseline has no average prices since the products are not part of the baseline. Therefore, we report the percentage points of the two algorithms concerning the normalized scale. Additionally, the gradient optimization does not produce consistent results for the inelastic demand for the reasons discussed in the limitation part of the previous Section.

The revenue lift of unit elasticity case is 20% above the baseline, which validates our modeling of the parameter $k_{rev,fract}$ (0.2 for the results in the figure). For the elastic case, the revenue lift is 24-31%, while it is 13% for the inelastic -0.5 case with the heuristic. The gradient optimization achieves a 3-4% increase. Throughout the cases, the algorithms increase the total contracted capacity by 31-54%.

While Figure 8-18 only presents the results for one of the nine combinations, Table 8-9 compares the revenue lifts across the nine combinations. The columns are first ordered by $k_{rev,fract}$ and then by k_{adj} . The first level of sorting for the rows is by the elasticity and then by the algorithm. The percentage increases in the parentheses for the gradient rows are the raises compared to the equally decreasing heuristic (same as the numbers above the red bars in Figure 8-18).

Table 8-9: Revenue lifts across the nine combinations, four elasticities, and two algorithms.

| | | $k_{rev,fract}$ 0.1 | | | $k_{rev,fract}$ 0.2 | | | $k_{rev,fract}$ 0.5 | | |
|------------------------|-------------------|---------------------|-----------------|-----------------|---------------------|-----------------|-----------------|---------------------|-----------------|-----------------|
| | | k_{adj} 0.2 | k_{adj} 0.5 | k_{adj} 1 | k_{adj} 0.2 | k_{adj} 0.5 | k_{adj} 1 | k_{adj} 0.2 | k_{adj} 0.5 | k_{adj} 1 |
| elasticity ↓ | baseline | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| | -0.5 equal. decr. | 102% | 105% | 109% | 105% | 113% | 126% | 120% | 151% | 202% |
| -1 | equal. decr. | 110% | 110% | 110% | 120% | 120% | 120% | 150% | 150% | 150% |
| | gradient | 110% (+0.0%) | 110% (+0.0%) | 110% (+0.0%) | 120% (+0.0%) | 120% (+0.0%) | 120% (+0.0%) | 150% (+0.0%) | 150% (+0.0%) | 150% (+0.0%) |
| rand. between (-2, -1) | equal. decr. | 118% | 113% | 111% | 133% | 124% | 119% | 170% | 151% | 141% |
| | gradient | 121% (+2.5%) | 115% (+1.6%) | 112% (+1.2%) | 138% (+3.7%) | 127% (+2.6%) | 121% (+2.0%) | 180% (+5.9%) | 158% (+4.3%) | 146% (+3.4%) |
| -2 | equal. decr. | 125% | 116% | 111% | 142% | 126% | 119% | 183% | 152% | 137% |
| | gradient | 129% (+3.7%) | 119% (+2.5%) | 113% (+1.9%) | 149% (+5.5%) | 131% (+3.9%) | 122% (+3.0%) | 198% (+8.5%) | 162% (+6.5%) | 144% (+5.1%) |

The case of $k_{rev,fract} = 0.1$ and $k_{adj} = 1$, with the equally increasing heuristic, yields the smallest additional revenues with 11% (just one percentage point above the baseline). The highest lift of 98%

occurs for the -2 elastic case and $k_{rev,fract} = 0.5$ and $k_{adj} = 0.2$. The lift from the gradient optimization ranges between 1.2% and 8.5%.

On the one hand, for all elastic cases, smaller k_{adj} combinations increase the revenues across algorithms. Referring back to Figure 8-17, that behavior makes sense since more quantities can be sold at a higher price. On the other hand, larger values for k_{adj} result in more additional revenues in the inelastic case.

Conclusions and limitations

In this Section, we analyzed what revenues gains are achievable by selling the available capacity through new products, such as a spot instances add-on. Given the considerable uncertainty about the pricing of these options, we explored a wide range parameterized by $k_{rev,fract}$ and k_{adj} . Based on the discussion of the results, we conclude the following points:

- The 38% available capacity is translatable into 2-102% additional revenues, which is more than the range of 0-27% we computed for the selling of the capacity through the existing Classical SLAs from the previous Section.
- The improvement from the gradient optimization is with 1-9% in a similar range than before with 5-7%.
- Products with a higher initial price for a smaller value of Mbps perform greater if the demand is elastic. The vice versa is true for products with lower reference prices for a larger value of Mbps and inelastic demand.

We identify these limitations for the analysis presented in this Section:

- We base the definition of the products on the two parameters: price for a corresponding data rate. That applies to the spot instance, but not a complete definition of the other two novel SLAs discussed in Chapter 7 (time-of-day pricing and Two Classes of Service). The inclusion of these two products requires an adjustment of the simulation and optimization. Furthermore, since more parameters define these SLAs, more assumptions have to be made.
- The algorithms vary prices without considering the impact on internal cannibalization. It is not an issue to include internal cannibalization in the optimization formulation. However, the critical point is to have data available on the amount of internal cannibalization as a function of price points.

- The up to 11% available part-day capacity remains unallocated for after the optimization, leaving room for additional revenues. The time-of-day pricing and Two Classes of Service are novel SLAs that can smoothen out the diurnal variations and therefore reduce part-day capacity.

8.5 Monetizing the available capacity through new customers

In contrast to the previous Section 8.4, we analyze in this Section, filling the capacity by contracting new customers. In our simulation, this is achieved by two options: unlocking affordability elasticity and increasing market share in a competitive environment, described in the following two Sections 8.5.1 and 8.5.2, respectively. For both analyses, we assume:

- Demand at the same price remains constant throughout the simulation, with no growth nor shrinking.
- Price changes do not affect the quantity users demand (no demand for use elasticity, considered in Section 8.4.1)
- Price is the only lever to control demand.
- Customers are contracted through Classical SLAs without add-on products (discussed in Section 8.4.2).
- Separate prices can be charged per segment.

Like in Section 8.4, we start each of the Sections by describing the specific approach and assumptions, followed by presenting the results, and lastly, summarizing the central conclusions and limitations.

8.5.1 Unlocking affordability elasticity

Specific approach and assumptions

In the article that Palerm wrote for the NSR [58], he argues that backhauling becomes elastic when prices further fall since MNOs can deploy backhauling more economically (we discussed the article in more detail in Section 8.1.2). The elasticity that Palerm discusses is for the access demand, i.e., more customers entering the market when prices fall. That is in contrast to the demand for use, where customers consume more when prices fall. The previous Section 8.4.1 discussed the latter, and this Section focuses on the former, which we also refer to as the affordability elasticity. The difference between the two approaches has three important implications for the setup of the simulations:

1. We need to define segments in which we can describe the affordability elasticity. We assume that the elasticity is known per segment (further discussed in Section 8.6).
2. With (1) the definition of the elasticity, it changes to be on segment instead of customer level.
3. The number of user terminals is changing during the pricing optimization. Therefore, the frequency allocation needs to be re-computed for every iteration. This change especially impacts the gradient approach.

We discuss the three implications in the following. Furthermore, similar to the previous two analyses, we assume that customers do not switch operator when prices change (no competition, considered in Section 8.5.2).

Segmentation

The segmentation is commonly driven by the marketing department and from the customers’ point of view (see Section 7.6). However, the selection of the bases for the segments in this Section is *driven by the simulation*. We leverage our work in the market segmentation Section 7.6, and pick the following four bases:

- The customer groups are A, B, C, D as available through the data set (they encode the type of business and location type bases from Table 7-3).
- The geographical region with the granularity of countries (32 different countries)
- The terminal size with the categories: 1.2m, 2.4m, and 4.5m
- The CIR with 47 categories

These bases group the 80 customers into 64 segments (for example, “B, Papua New Guinea, 1.2m, 100 Mbps”). From a simulation perspective, this segmentation provides all information to add user terminal to the segment: the customer groups (A, B, C, and D) and the CIR define the usage behavior. The country gives the latitude and longitude (we randomize the location within the country), and the terminal size provides the G/T.

Affordability elasticity

Given the segments, we can now define the affordability elasticity (not considering the demand for use elasticity). We use the same log-linear prior and fit it through a reference point. Figure 8-19 displays an example for a segment with two user terminals at a current price of \$200/month/Mbps.

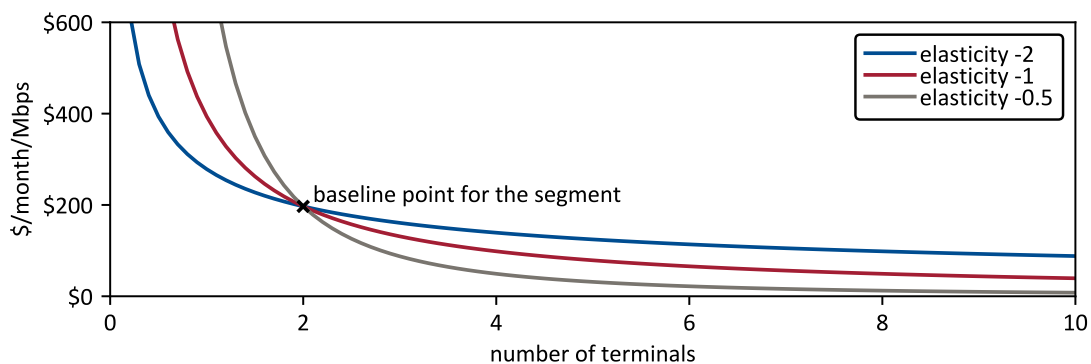


Figure 8-19: Example of the affordability elasticity for a segment with two user terminals at \$200/month/Mbps

If the price now drops to \$100/month/Mbps, the model predicts that two additional user terminals appear for the unit elastic case. Hence, the segment would end up with four terminals at \$100/month/Mbps each. While we draw the elasticity lines continues, the number of terminals is discrete, and therefore, for a price of \$101, the segment only has demand for three terminals.

Implications of changing the number of terminals

When the number of user terminals changes, the resource allocation process needs to be restarted. The decision about the user grouping is straightforward here (one beam per user terminal). Nevertheless, the computation of the frequency assignment and power is computationally more expensive. Our objective is to find the optimum combination of prices that fill the available capacity. The affordability elasticity translates the prices into the number of user terminals, for which we can compute the available capacity through the resource allocation process and the available capacity forecaster.

The general algorithmic approach outlined in Section 8.3 remains unchanged. The different algorithms compute an ordered list of prices, and a binary search algorithm finds the element where the available capacity is smallest. Note that line 4 in the binary search Algorithm 8-1 now includes the frequency assignment step in addition to power computation. The most significant change is in $\Pi(C)$ as part of the gradient optimizer. With the change from customer to segment level in the elasticity, the regression has to switch onto the segment level as well. $\Pi(C)$ is no longer the revenue per customer as a function of the power provided to the customer but is now revenue *per segment* as a function of the power provided to the segment.

When the demand changed per user terminal in Section 8.4, we swept through the possible maximum data rates and recorded revenues Π and power C to build the samples for the $\Pi(C)$ regression. In this Section, the approach is to have the number of terminals as a sweeping parameter. We can straightforwardly calculate the revenues as a function of the number of terminals per segment. However, the computation of the power as a function of user terminals is not trivial. The power depends on the assigned frequencies to the user terminal in the segment. Moreover, the frequency assignment links the frequencies for each segment. Therefore, the power as a function of the number of user terminals in one segment becomes also a function of the number of user terminals in all other segments. This combinatorial coupling results in an intractable number of simulations. As a result, we approximate the coupling by two heuristics: the *density penalty* and *satellite utilization penalty*. With that, the gradient optimization becomes a *heuristic gradient* approach.

We compute a base power $C_{base,i}$ for a single user terminal of segment i . Then for any additional terminal in the segment, we apply the two heuristic penalty factors $k_{dens,pen}$ and $k_{satutil,pen}$. Formally, the power $C_i(N_{u,i})$ for the segment as a function of the number of user terminals in the segment $N_{u,i}$ reads:

$$C_i(N_{u,i}) = C_{base,i} \cdot N_{u,i} \cdot k_{dens,pen}(N_{u,i}) \cdot k_{satutil,pen}(N_{u,i}) \quad (8-11)$$

The intuition behind the *density penalty* $k_{dens,pen}$ factor is that a concentration of many user terminals in a small area makes the frequency assignment more challenging. It is more likely that the allocated frequency spectrum is smaller, and therefore the power is higher. We model this relationship with a linear function between the penalty factor and the number of terminals. After a threshold value $N_{u,i,thresh}$ the linear function increases with the slope $s_{dens,pen}$ (which is a tuning parameter):

$$k_{dens,pen}(N_{u,i}) = \begin{cases} 1 & \text{if } N_{u,i} \leq N_{u,i,thresh} \\ a_{dens} + s_{dens,pen} \cdot \frac{N_{u,i}}{N_{u,i,thresh}} & \text{otherwise} \end{cases} \quad (8-12)$$

with $a_{dens} = 1 - s_{dens,pen}$. We compute the threshold number of user terminals for segment i as a function of the area of the country's segment $A_{seg,i}$, the number of segments in the country $N_{seg,i}$, and a tuning parameter A_{thresh} . Furthermore, we take the maximum, so the threshold number is always larger than 1, which means there is at least one terminal per segment that is not penalized.

$$N_{u,i,thresh} = \max\left(\frac{A_{seg,i}}{N_{seg,i} \cdot A_{thresh}}, 1\right) \quad (8-13)$$

Figure 8-20 shows the result of Eqs. (8-11) - (8-13) for Germany and Mali in Africa. The two tuning parameters are set to $A_{thres} = 300,000 \text{ km}^2$ and $s_{dens,pen} = 1$. That results in a threshold $N_{u,Ger,thresh} = 3$ for Germany and $N_{u,Mali,thresh} = 12$ for Mali. For 20 terminals in Germany, the penalty factor is $k_{dens,pen} = 5.9$. In contrast, the penalty is only 1.7 for Mali (and 1 for larger countries such as the US or Brazil).

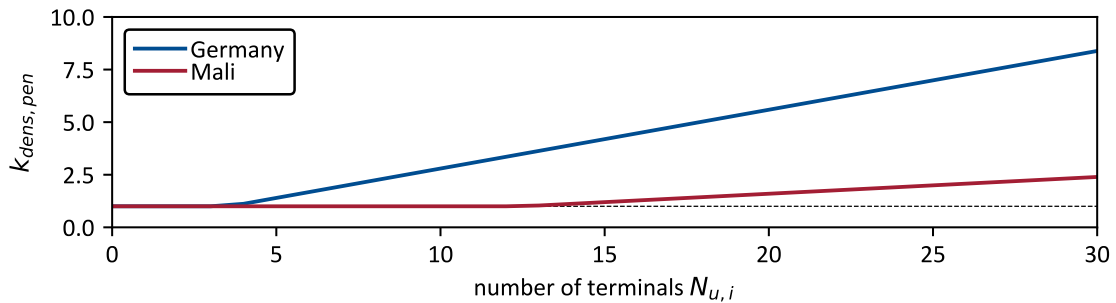


Figure 8-20: density penalty $k_{dens,pen}$ as a function of the number of terminals for Germany and Mali in Africa. The threshold is 3 for Germany and 12 for Mali with $A_{thres} = 300,000 \text{ km}^2$ and a slope of $s_{dens,pen} = 1$.

The second parameter $k_{satutil,pen}$ penalizes segments more that connect to higher utilized satellites. The rationale is that similar to the density penalty: computing the frequency assignment is more challenging when the satellite provides service to more user terminals. We use a linear relationship with the tuning parameter being the maximum penalty $k_{satutil,pen,max}$ for a relative utilization $\eta_{rel,util} = 1$ (see Figure 8-21). The utilization is relative to the highest utilized satellites in terms of CIRs.

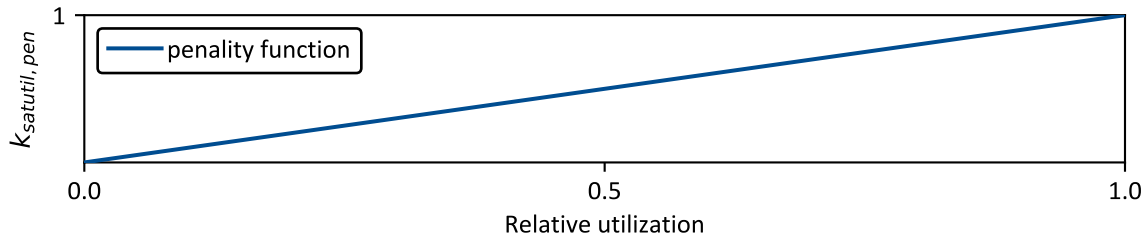


Figure 8-21: linear relationship between the relative utilization and the satellite utilization penalty

With the two heuristics, we introduce three hyperparameters, the threshold area A_{thresh} , the slope $s_{dens,pen}$, and the maximum satellite utilization penalty $k_{satutil,pen,max}$. Since the performance of the heuristic gradient optimization depends on these parameters, we tune them for each elastic case.

Heuristic gradient hyperparameter tuning

Our approach for tuning the three hyperparameters A_{thresh} , $s_{dens,pen}$, and $k_{satutil,pen,max}$ is screening through reasonable combinations. We choose six values for the maximum satellite utilization penalty (0, 0.5, 1, 1.5, 2, 3), nine values for the slope of the density penalty (0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4), and nine values for the area threshold (500, 1,000, 3,000, 5,000, 10,000, 50,000, 100,000, 1,000,000, 10,000,000). The first two parameters are unitless, and the unit of the latter is km^2 . We report the results relative to the baseline revenue. Figure 8-22 shows it for the -2 elasticity and Figure 8-23 for the elasticity range between -2 and -1. The optimum values for a given $k_{satutil,pen,max}$ is printed bold, and the global optimum has a red font. For combinations where the cells are blank, the simulation did not finish.

First, we focus on the -2 elasticity case. The most notable revenue lift is for $k_{satutil,pen,max} = 2$, $A_{thresh} = 5,000 km^2$, and $s_{dens,pen} = 1.5$. We use this setting for the remaining results in this Section. We note that the local optimums are less than 2% smaller than the global maximum. The best values for A_{thresh} vary between 5,000 and 50,000 km^2 across slopes and utilization penalty.

Another evident trend is that the influence of the density penalty parameters $s_{dens,pen}$ and A_{thresh} decreases with increasing $k_{satutil,pen,max}$ (the color gradients are less significant for larger $k_{satutil,pen,max}$). This observation implies that both penalties are closely coupled. A higher maximum

utilization penalty makes it more difficult for the density penalty to improve the results, as for example, high density areas are correlated with the utilization of the satellite. The incentivization of contracting users for low utilized satellites has a comparable effect than incentivizing lower density countries. However, the two heuristics are different if the density gradient is along the latitude. The utilization does not differentiate, but the density penalty does.

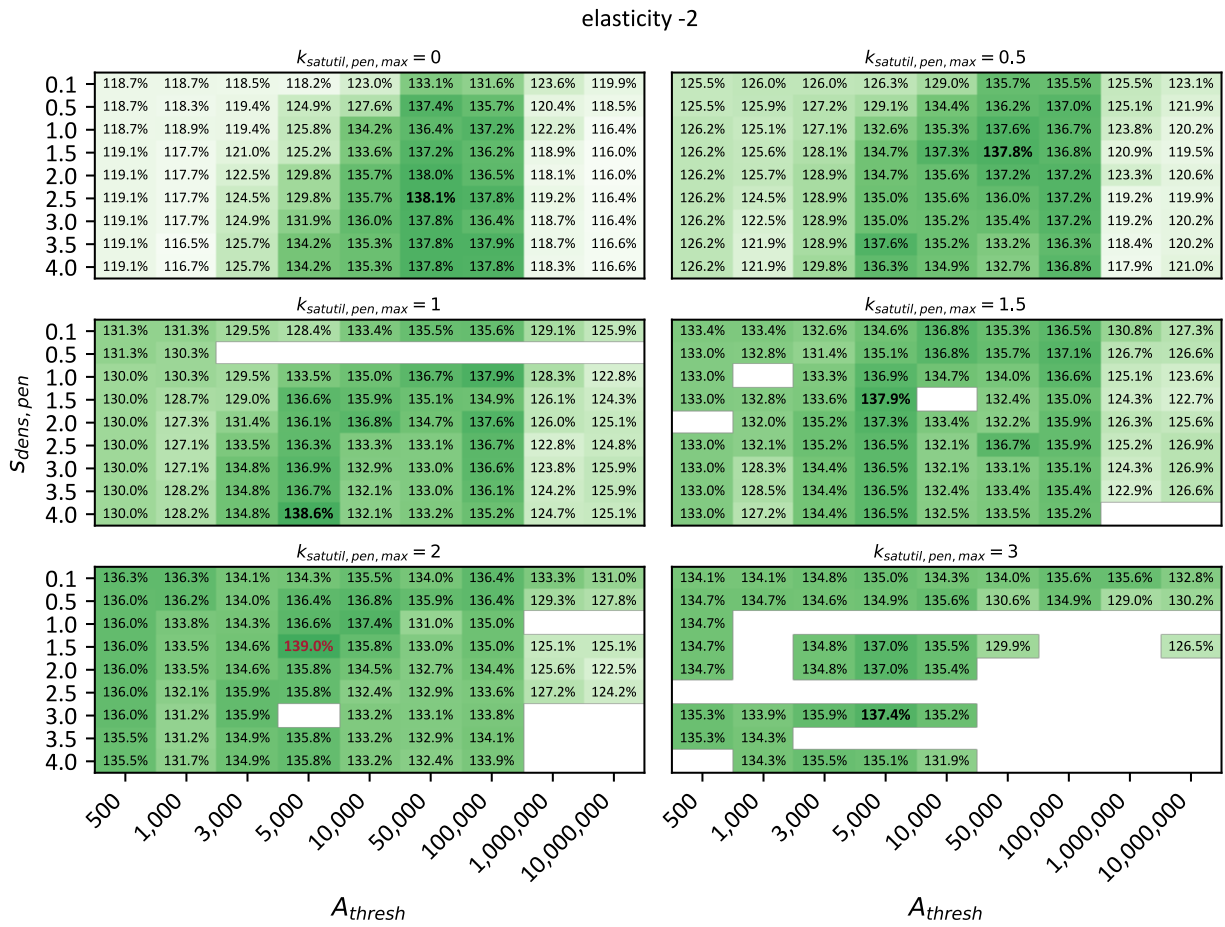


Figure 8-22: hyperparameter tuning of the heuristic gradient algorithm for the elasticity case -2. Numbers are relative to the baseline revenues. Bold values are the maximum in each $k_{satutil,pen,max}$ case and the red font indicates the global maximum.

Turning to the case with elasticity varying between -2 and -1, we observe similar tendencies. Nevertheless, the optimum result is for a different hyperparameter setting: $k_{satutil,pen,max} = 0.5$, $A_{thresh} = 10,000 \text{ km}^2$, and $s_{dens,pen} = 1.5$. In both cases, the hyperparameter tuning achieves over a 15% improvement compared to the worst-performing combination of parameters. Note that our screening approach does not guarantee that we found the optimal combination of parameters.

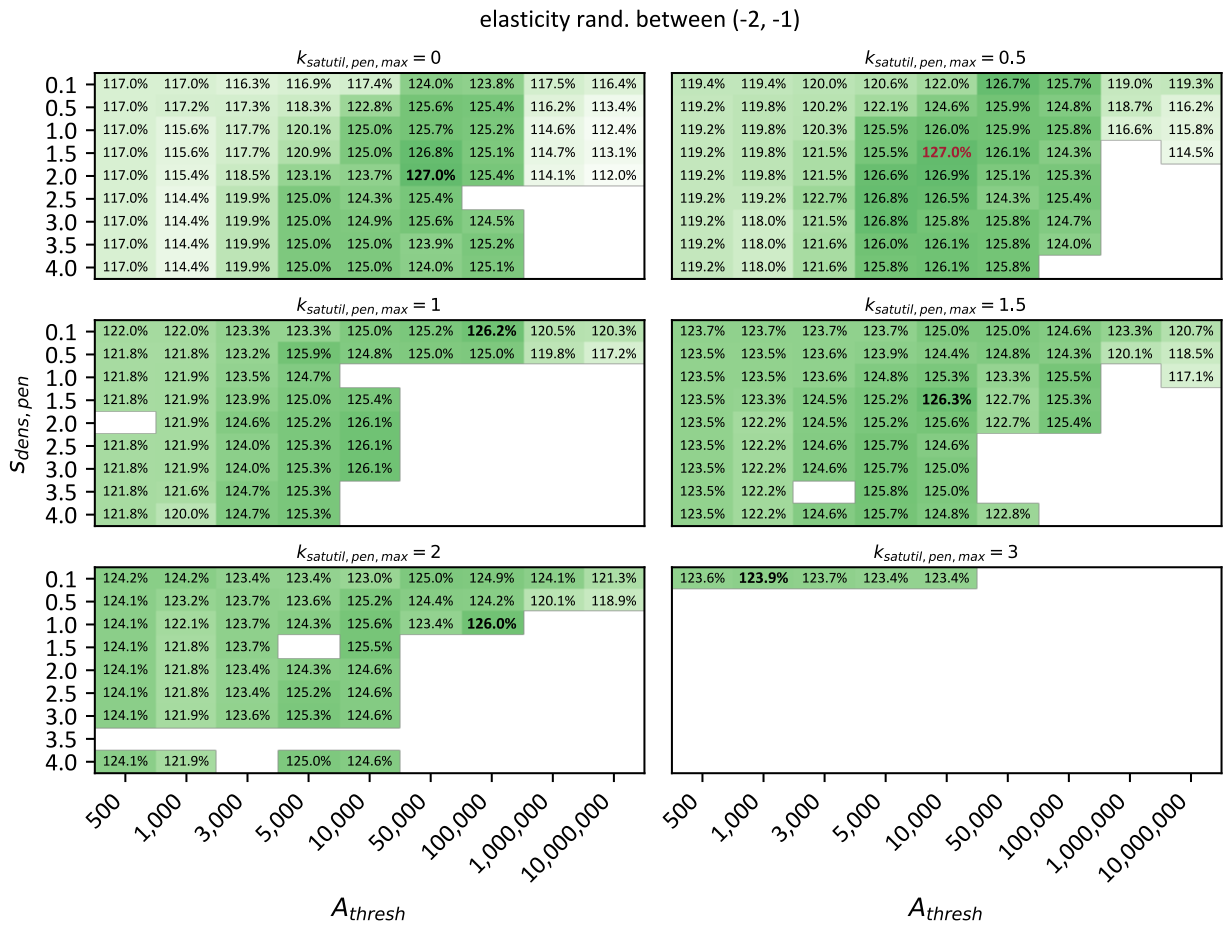


Figure 8-23: hyperparameter tuning of the heuristic gradient algorithm for the case with elasticities varying between -2 and -1. Numbers are relative to the baseline revenues. Bold values are the maximum in each $k_{satutil,pen,max}$ case and the red font indicates the global maximum.

Lastly, before we discuss the results, we introduce a new heuristic that makes pricing decisions based on the *highest yield*, where we define yield as \$/month/Mbps (which is price). Therefore, the prices are adjusted for the segment with the highest yield first. Since most prices decline in this simulation, this approach is stable in the sense that reducing prices for the highest yield first eventually allow for this segment to be at a lower yield than others and therefore leveraging out the price decline. Like the equally decreasing heuristic, a list of prices \mathcal{L} is produced, which is optimized to use up all of the available capacity by a binary search. The heuristic gradient approach differs in how the list of prices \mathcal{L} is generated. The algorithm computes the marginal revenues and adjust prices for every increment such that the list of prices is optimally sorted (see Section 8.3.2.2). While the highest yield only considers the price, the heuristic gradient trades-off the capacity cost, the elasticities, and the price.

Discussion of the results

We present the results in the same format as in the two previous Sections, 8.4.1 and 8.4.2 but include the highest yield algorithm (see Figure 8-24). The baseline is the same as throughout the previous two analysis in Section 8.4 and represents existing customers. Any lift compared to the baseline is attributed to new customers. The percentage points above the gray bar are the relative change concerning the heuristic gradient approach.

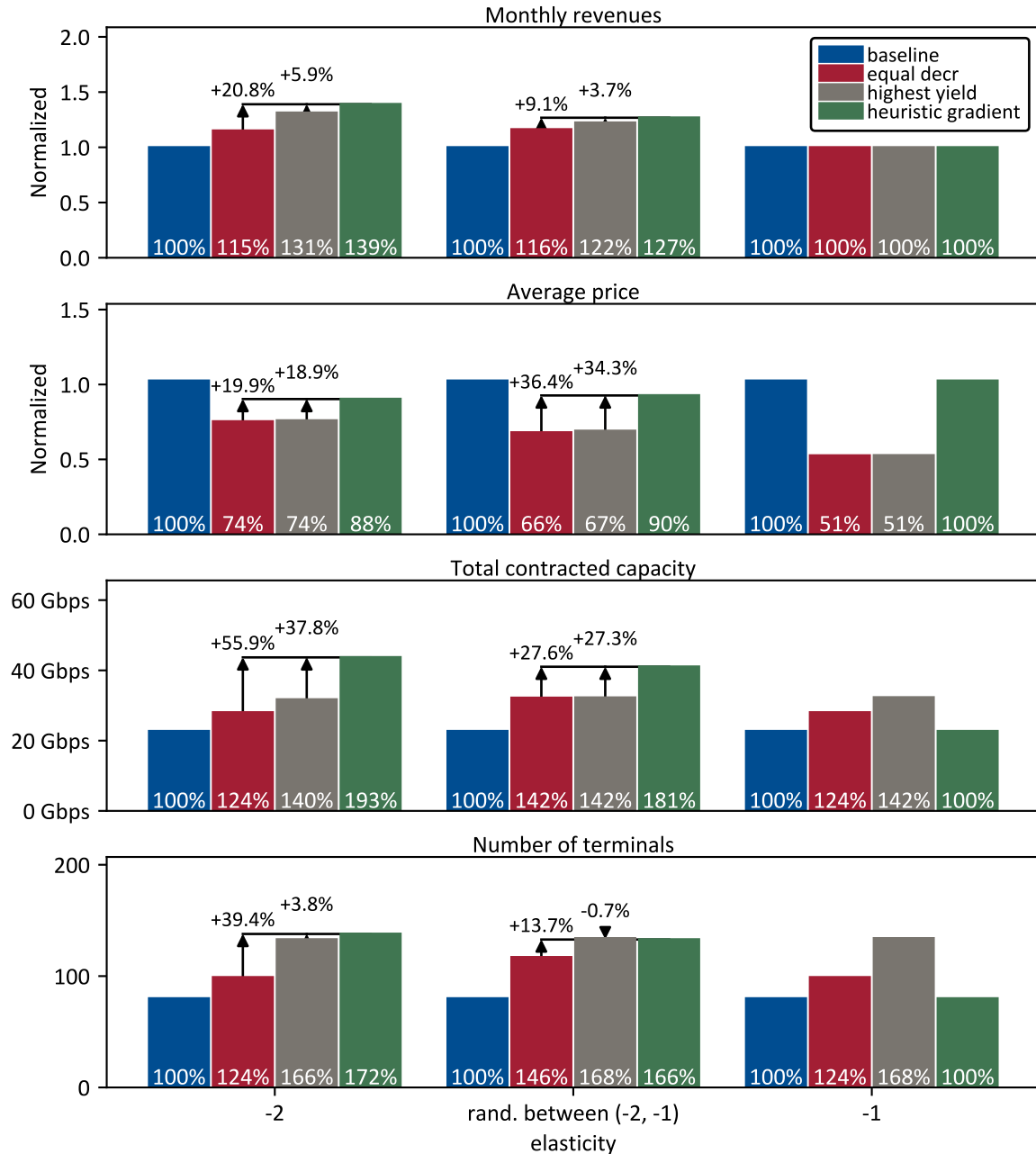


Figure 8-24: results for the monthly revenue increase, average price, total contracted capacity and number of terminals for the three algorithms for three elasticity cases.

Again, we validate the algorithms with the unit elastic case; the revenues remain unchanged even when the equal decreasing and highest yield heuristics reduce prices (the gradient approach does not reduce price since it does not result into any revenue lift, i.e., the marginal revenues are non-positive, see line 4 in Algorithm 8-3). As expected, the revenues increase across algorithms when the affordability demand is more elastic. The heuristic gradient outperforms the other approaches and achieves a 39% lift for the -2 elastic case and 27% for the case where elasticity is between -2 and -1. Furthermore, the average prices remain more stable, the total contracted capacity increases, and the number of user terminals is similar to the highest yield heuristic. The highest yield consistently performs better than the equally decreasing approach and comes close to the heuristic gradient algorithm with only 4-6% behind.

To further understand what decisions the heuristic gradient algorithm makes that let it outperform the highest yield, we compute a pair-wise correlation shown in Table 8-10. We used a similar table in the previous Section 8.4.1 to understand customer pricing decisions compared to the baseline. This time we compare the revenues per segment achieved by the heuristic gradient with the highest yield heuristic.

Table 8-10: Pair-wise correlation between selected attributes. Red colored fields are positive, blue ones are negative correlation. Values are computed for the elastic -2 case comparing the heuristic gradient with highest yield.

| | rev. incr. of gradient over highest yield | segment baseline price | terminal size | typical mean to CIR fraction | CIR | customer groups (A, B, C, D) | latitude | longitude | typical traffic uncertainty |
|---|---|------------------------|---------------|------------------------------|------|------------------------------|----------|-----------|-----------------------------|
| rev. incr. of gradient over highest yield | 1.0 | -0.4 | 0.4 | -0.3 | 0.2 | 0.2 | 0.1 | 0.1 | 0.0 |
| segment baseline price | -0.4 | 1.0 | -0.3 | 0.1 | -0.5 | 0.3 | 0.3 | -0.2 | 0.0 |
| terminal size | 0.4 | -0.3 | 1.0 | 0.1 | 0.8 | 0.2 | 0.1 | 0.2 | -0.3 |
| typical mean to CIR fraction | -0.3 | 0.1 | 0.1 | 1.0 | 0.0 | 0.2 | 0.2 | -0.1 | -0.5 |
| CIR | 0.2 | -0.5 | 0.8 | 0.0 | 1.0 | -0.1 | -0.2 | 0.2 | -0.3 |
| customer groups (A, B, C, D) | 0.2 | 0.3 | 0.2 | 0.2 | -0.1 | 1.0 | 0.3 | -0.1 | 0.0 |
| latitude | 0.1 | 0.3 | 0.1 | 0.2 | -0.2 | 0.3 | 1.0 | -0.1 | -0.1 |
| longitude | 0.1 | -0.2 | 0.2 | -0.1 | 0.2 | -0.1 | -0.1 | 1.0 | -0.0 |
| typical traffic uncertainty | 0.0 | 0.0 | -0.3 | -0.5 | -0.3 | 0.0 | -0.1 | -0.0 | 1.0 |

We correlate nine attributes from the segment with the revenue increase: the segment baseline price, the terminal size, the typical mean usage to CIR fraction, the CIR, the segment type, latitude, longitude, and a typical traffic uncertainty. Some of these correlations are similar to the ones we discussed earlier. Hence, we focus only on the ones that are particularly relevant here.

The segment baseline price has a strong negative correlation with the revenue increase of the heuristic gradient. That means that the gradient heuristic extracts more value from segments with a *lower* baseline price compared to the highest yield. We explain this by the heuristic behind the highest yield algorithm. It gives priority to the customers with the highest yield but does not consider the implications on capacity usage.

The terminal size positively correlates with a value of 0.4 (compared to 0.6 in the previous correlation analysis). Since the gradient approach considers the capacity cost in terms of power consumption, it extracts more revenues from segments that have larger terminals even if their yield is lower. The algorithm finds the optimal trade-off.

The third attribute, the typical mean data rate to CIR fraction, shows a considerable negative correlation. That means that the gradient approach prefers segments with a lower fraction. The reason is that a lower fraction means a lower resource usage (on average), and therefore, more resources can be reallocated.

The remaining attributes show little to no correlation, which is a similar result to the discussion in Section 8.4.1. Likewise, the correlations between the attributes themselves show no new trends. We summarize the conclusions and limitations in the following.

Conclusions and limitations

The goal of this Section was to analyze the monetization of the available capacity by unlocking the affordability elasticity. We defined the segments and their elasticity. For the gradient approach, the change from providing more to existing users to providing service to more users posed challenges. We developed two heuristics, the density, and satellite utilization penalty. In total, they have three hyperparameters, which we tune by screening for each elastic case. Finally, we discussed the results and provided insight with correlation analysis. The key conclusions are:

- The 38% whole-day available capacity is translatable into 0-39% additional revenues depending on algorithm and elasticity. This range compares to 2-102% from additional products and 0-27% for selling the capacity through existing Classical SLAs.

- The heuristic gradient optimization achieves 9-21% higher revenues than the equally decreasing heuristic, and 4-6% more compared to the highest yield algorithm with more stable average prices. The lift in the other Sections was in the range of 1-9% compared to the equally decreasing.
- Compared to the highest yield heuristic, the gradient-based approach extracts additional revenues from segments with a lower baseline price, larger terminals, and lower typical mean to CIR fraction.

In addition to the assumption listed at the beginning of this Section, the analysis has the following limitations (some can produce interesting future research):

- In the inelastic case, the operator would want to sell the minimum number of terminals per segment. Growing total demand over time can be another way to monetize the available capacity. This analysis can utilize the approach presented in this Section.
- The simulation drives the segmentation based on practical considerations. That might not be ideal from a customer or marketing perspective.
- For the heuristic gradient approach, we can no longer prove that the algorithm is optimal. While the results confirm that the chosen two heuristics work well, we expect that a more profound study could yield at least a few more additional percentage points in revenues.

8.5.2 Increasing market share

Specific approach and assumptions

In this fourth analysis, we *include* competition. The approach is to monetize the additional capacity by lowering prices to increase the market share, while the total addressable market remains the same size. That statement implies that prices do not impact either the demand for use (more to existing customer case, Section 8.4.1) nor the access demand through unlocking the affordability (see Section 8.5.1). That means that the elasticity of the segment plays no role. Besides these points, all other assumptions outlined at the beginning of Section 8.5 hold.

We set up two competitors, A and B, with the same technology and initial customers, as outlined in Section 8.1. They differentiate themselves by the resource allocation and RM principles they use. Competitor A is the technological leader in the market, and B follows.

At the start of each simulation, the markets are in equilibrium, current prices clear the capacities, and each competitor has a market share of 50%. We consider that competition occurs on the segment level and that the service from both competitors is indistinguishable. We analyze and elaborate on different types of imperfect competition assumptions, market share stickiness, and customer loyalty.

For this analysis, prices do not impact demand. Nevertheless, there is still a relationship between prices and the number of user terminals on segment level due to the competitive element. In that, we introduce a discount threshold $k_{discount}$ that defines how much lower the price needs to be that customers switch operators. This factor is influenced by market share stickiness (=customer loyalty) and terminal cost. It is in our simulation a parameter on which we perform a sensitivity analysis. Figure 8-25 illustrates that: when the price is below the threshold price p_{thres} , customers are doubled within a segment. The price p_{seg} is reduced by $k_{discount}$ through $p_{thres} = p_{seg} \cdot (1 - k_{discount})$. Since we have two competitors with the same initial user terminals $N_{u,A} = N_{u,B}$, the number doubles and all users from a segment switch.

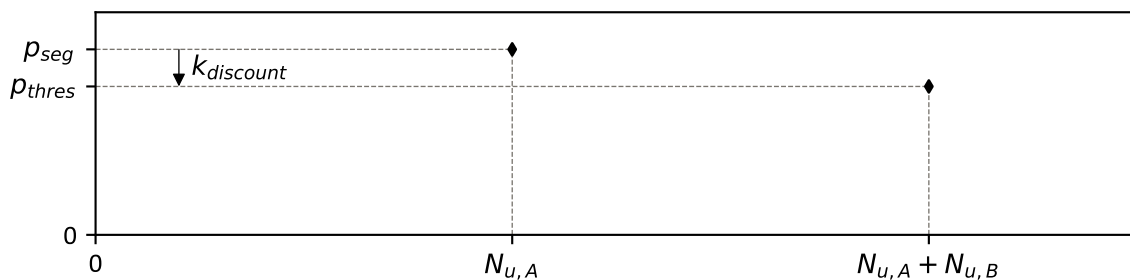


Figure 8-25: schematic drawing of the modeling of the discount factor $k_{discount}$ on segment level (here 20%)

The pricing decisions that the algorithms need to make is equivalent to a binary decision per segment: the prices can be kept at the initial level, which means the same set of user terminals or the prices can be reduced to the threshold level p_{thres} and then the users double (assumption is that competitor knows the threshold price and does not have to do price discovery). Using this representation, we can leverage the algorithms that we developed, except for the equally decreasing heuristic. The reason we cannot include the equally decreasing heuristic is that this heuristic would reduce prices for *all* segments by $k_{discount}$, which would mean a doubling from 80 to 160 users, exceeding the available capacity. Hence, the following results only show the highest yield and heuristic gradient approaches. For the hyperparameters of the heuristic gradient optimization, we use the ones from the elastic -2 case of Section 8.5.1 ($k_{satutil,pen,max} = 2$, $A_{tresh} = 5,000 \text{ km}^2$, and $s_{dens,pen} = 1.5$).

Discussion of the results

We report the final numbers in Figure 8-26. Instead of the elasticity cases as in the analyses before, the horizontal axis shows four different cases of the discount factor $k_{discount}$ (5%, 10%, 20%, and 50%). The other metrics and labels are the same.

The validation case is for $k_{discount} = 50\%$. With the doubling of the user terminals, this is precisely the definition of the unit elasticity: a 50% lower price increases demand by a factor of two. The revenues of the two algorithms match the baseline, despite that the average price, total contracted capacity, and the number of terminals differs. 50% also defines an upper boundary. If customers in a segment require more than a 50% discount to switch, this is not beneficial for the operator (from a pure revenue perspective).

For lower price discount values, the number of terminals and total contracted capacity remains unchanged. The reason is that the discount is homogenous across all segments, and therefore there is no differentiation that the algorithms can leverage. The prices need to be reduced to p_{thres} if the market share in the segment wants to be increased. The heuristic gradient can achieve 59% more total capacity compared to the highest yield, and an almost similar increase in the number of terminals.

Since the set of terminals does not change for different discount cases, the price closely correlates with the monthly revenues. Lower discount values keep the prices more stable and increase the lift in revenues from 52% for a 20% discount, to 78% for a 5% discount. The average price does not reflect the total discount as the algorithms only discount the price for a subset of segments (the available capacity is not sufficient to serve all the demand). However, it comes closer to the gradient approach as the number of

terminals increases by 76%. Since the highest yield has a lower number of additional terminals for which segments need to be discounted, the average price stays closer to the baseline.

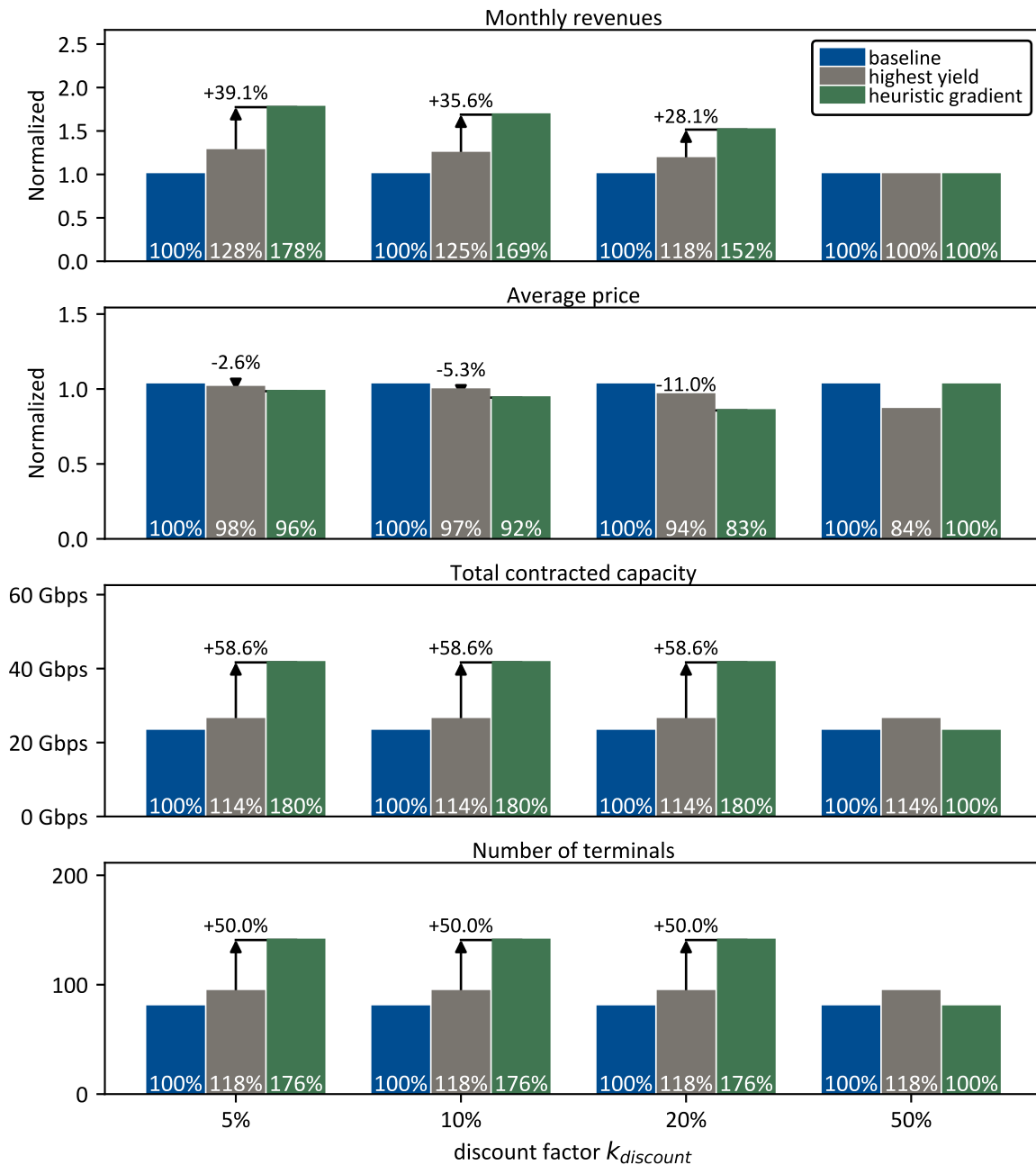


Figure 8-26: results for highest yield and the heuristic gradient optimization for different discount threshold $k_{discount}$ cases

The heuristic gradient optimization outperforms the highest yield by 28-39%, depending on the discount factor. This lift is larger than the numbers we obtained from the other analyses. We explain this by the behavior of the highest yield. It locks itself in by going after the segment with the highest price, which

happens to be the ones with the most substantial numbers of user terminals already. Taking all customers from this segment fills up the capacity quickly. Hence, not considering users from other segments with a lower yield but a peak demand that is during a time when more capacity is available. Figure 8-27 provides support. It shows the number of terminals per segment for customer A comparing the baseline and the two algorithms.

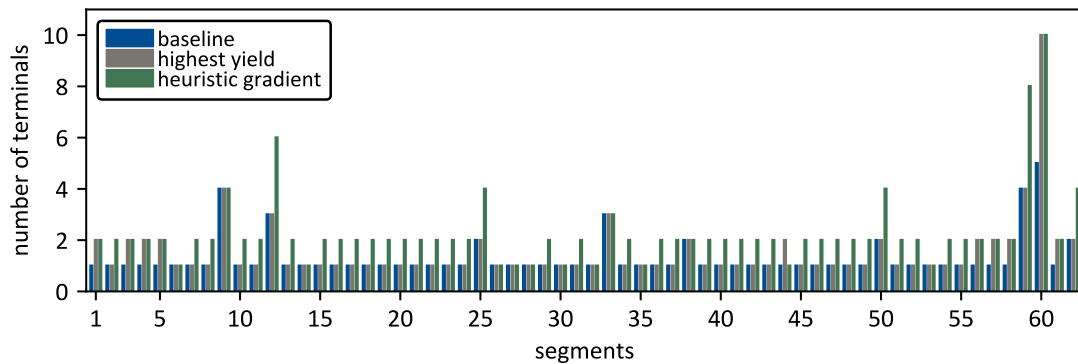


Figure 8-27: number of terminals for each segment for the baseline, the highest yield solution, and the heuristic gradient optimization. If number of terminals is different from the baseline, it is always a doubling.

The highest yield takes the customers from segment 44 and 60, while the heuristic gradient does not reduce the price there. It does this more evenly across the segments and also for smaller segments with lower yields such as 15-24.

Time dependency

So far, we showed results that are for a specific instance in time without considering the time-dependent effects. This subsection addresses these. First, the customers of competitor B have current contracts, for which the *switching costs are high* (not necessarily financial, e.g., the effort of setting up a new terminal; similar to mobile phone contracts and telecommunication services [231]). Therefore, we can expect that customers do not switch before the contract expires with B. Figure 8-28 adds this time dimension to the results presented in Figure 8-26.

Since the contract dates are actual values, we normalize the time axis to the average SLA duration and compute the delta to the year 2020. The total contract durations vary, in normalized terms, from 0.2 to 2.8 average durations. The figure shows the monthly revenues and the number of terminals versus time for both algorithms. When a contract expires, A reduces the price in the segments it targets and provides service to the new user terminals. That incrementally builds up to the revenue lifts Figure 8-26 displays. Since the gradient approach can support more terminals, it is more likely that terminals have contract

expiration dates closer to the simulation initialization time. While the plot confirms that the behavior is ambiguous, it depends on the specifics of the customers and the reference date.

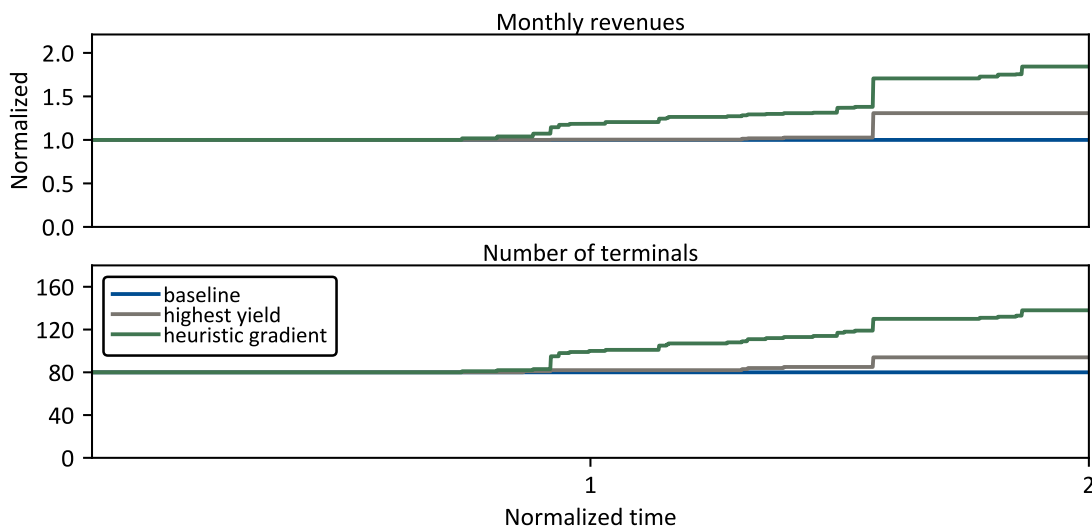


Figure 8-28: Time dependency of the monthly revenues and number of terminals for the two algorithms. Time is normalized to the average contract duration.

Nevertheless, the main objective of this figure is to illustrate that there is a definite time dependency with the current, year-long contracts (see discussion in Chapter 2 on shorter contracts and the benefits to the market). The heuristic gradient outperforms the highest-yield heuristic. However, that is not generalizable, and a possibly fruitful future research to explore algorithms that additionally consider the time dimension. That direction ties back to the second application of RM in satcom that we discussed in Section 3.7, the *optimal filling of capacity with customers over time*.

Market share stickiness and loyalty

The final discussion of this result part is on the market share stickiness and loyalty. So far, we assumed that competitor B does not react to pricing decisions made by A. Certainly, this assumption only applies, if at all, to the short-term when competitor A can sustain an information asymmetry [240].

If B decides to match prices, then in a perfect competition model, the market share is split equally, and the market is at its initial condition but with lower prices. If the demand (access demand and demand for use) is inelastic, everyone is worse off than before. However, if demand is elastic, the market grows, and all competitors increase their revenues equally. Given the uncertainty in the elasticity, this race to the bottom is risky for the market as a whole. Competitors try to find a way out of this price-based competition by differentiating their offerings, and by moving into niches [241].

Furthermore, operators (and also down the value chain, service-providers) try to increase their *stickiness*, i.e., make it harder for customers to switch by locking them in (such as frequent flyer programs in the airline industry [242]) or by convincing them that there are intangible qualities associated with their service. Various factors can increase stickiness, such as price, technical and functional service quality, switching cost, and loyalty programs [243]. For example, longer contract duration increases the switching costs because they have a cost of exiting.

In the context of this analysis, our results can be interpreted as simulating stickiness. Our approach is to vary the stickiness to understand its impact on the results. We argue that there exists a price threshold, after which customers switch despite their loyalty to the original competitor. The parameter $k_{discount}$ models that relationships. Hence, the required discount is directly proportional to the stickiness of the customers in a segment.

A possible strategy for competitor A might be to be the first to lower prices and unlock the affordability elasticity. Hence, the overall market grows, and the competitor grows its revenue while the revenues of competitor B remains unchanged. Then, A works to make the new customers sticky to avoid switching. As shown in Figure 8-24 of Section 8.5.1, sophisticated RM algorithms achieve higher revenues and more stable average prices out of the same amount of available capacity. Or put differently, sophisticated RM makes more efficient use of capacity and reduces per-unit cost.

If competitor B is to follow the pricing of A, its resource allocation must be equally sophisticated. If that is not the case, B can support fewer user terminals than A, and therefore the unit costs are higher, reducing the return on investment of the satellite.

Conclusions and limitations

In this Section, we analyzed monetizing the available capacity by increasing the market share in a competitive environment. The simulation has two competitors with the same technology and initial customer set. A discount factor $k_{discount}$ accounts for the delta in price that makes loyal customer switch operators. We computed results for a variety of different values and elaborated on the time dependency of the results. Furthermore, we deliberate on the market share stickiness and its implications. We reach these conclusions:

- The 38% available capacity is translatable into 0-78% additional revenues depending on the stickiness. This range compares to 0-39% from unlocking the affordability elasticity, 2-102% from additional products, and 0-27% for selling the capacity through existing Classical SLAs.

- The heuristic gradient approach outperforms the highest yield algorithm by 28-39% in terms of additional revenues.
- A discount factor of 50% is an upper bound in a two-player market under our assumptions. Higher values of discount factor decrease the revenues for the operator.
- More sophisticated RM uses the available capacity more efficiently, and hence reduces the per-unit cost. This benefit can increase the operator's competitive advantage compared to operators who use less advanced techniques.

When it comes to limitations of the analysis, these are the central ones:

- The multiple year-long contracts add a critical time dependency on the analysis that is an exciting area for future research.
- Given the timeframes of the contracts, competitors are likely to respond to the new prices. One approach can be to model this dynamic by game theory approaches in combinations with system dynamics as used for the market dynamics investigation in Chapter 2.

8.6 Challenge of uncertain elasticity information

The results of the previous two Sections, 8.4 and 8.5, illustrated the impact of different elasticities on the achievable revenue improvement. The objective of this Section is to discuss the challenge of customer elasticity estimation (see Appendix D for a primer on demand elasticity). The component is part of the satcom RM demand management's estimation and forecasting layer (see Figure 8-1).

In the data input Section 8.1.2, we outlined the challenge associated with estimating the elasticity based on the customers' purchasing history. Furthermore, we saw significant ranges in the more data-rich telecommunication industry, suggesting that even with more sophisticated methods, the uncertainty in the estimation remains considerable. Hence, we conclude that the elasticity is an *impactful and uncertain* input parameter to the pricing optimization that merits further attention.

In that, this Section consists of two parts. First, we analyze the impact of the information mismatch between the believed and the actual elasticity. Second, we review customer elasticity learning methods that are potential candidates to reduce the uncertainty in the estimations over time.

Impact of mismatch between believed and actual elasticity

The analyses presented in the previous Sections assume that the algorithms know the elasticity and that this estimation is accurate. In contrast, the analysis in this Section examines the impact on the final results when the elasticity assumed by the algorithms differs from the actual elasticity, i.e., the error of the elasticity estimation. Our approach is to use a Monte-Carlo simulation in which we vary the actual elasticity up to a given percentage of the believed. For example, a believed -1 elasticity has, with a mismatch of up to 50%, an actual elasticity between -1.5 and -0.5. We simulate 1,000 samples and report the statistical results with boxplot in Figure 8-29.

We compare four different values for the mismatch: 10%, 25%, 50%, and 100% across the two elastic cases -2, and random between -2 and -1. The scenario is to sell more to existing SLAs from Section 8.4.1. We parallel the gradient optimization with the equally decreasing heuristic and the baseline. The improvement of the gradient approach is in relative percentage points concerning the medians. We normalize the revenues to the baseline value.

The results for a 10% mismatch and -2 elasticity are close to those obtained in Section 8.4.1 with no uncertainty, i.e., a 0% mismatch. With increasing mismatch, the results become more uncertain to the point where some simulations achieve revenues below the baseline (see lower right plot). For a 100% mismatch, the median shrinks from 127% to 113% for the gradient algorithm and -2 elasticity. Moreover,

the uncertainty increases substantially to over 40% of the baseline revenues. Furthermore, the lift of the gradient approach remains consistent between 4-8% – almost independent of the mismatch.

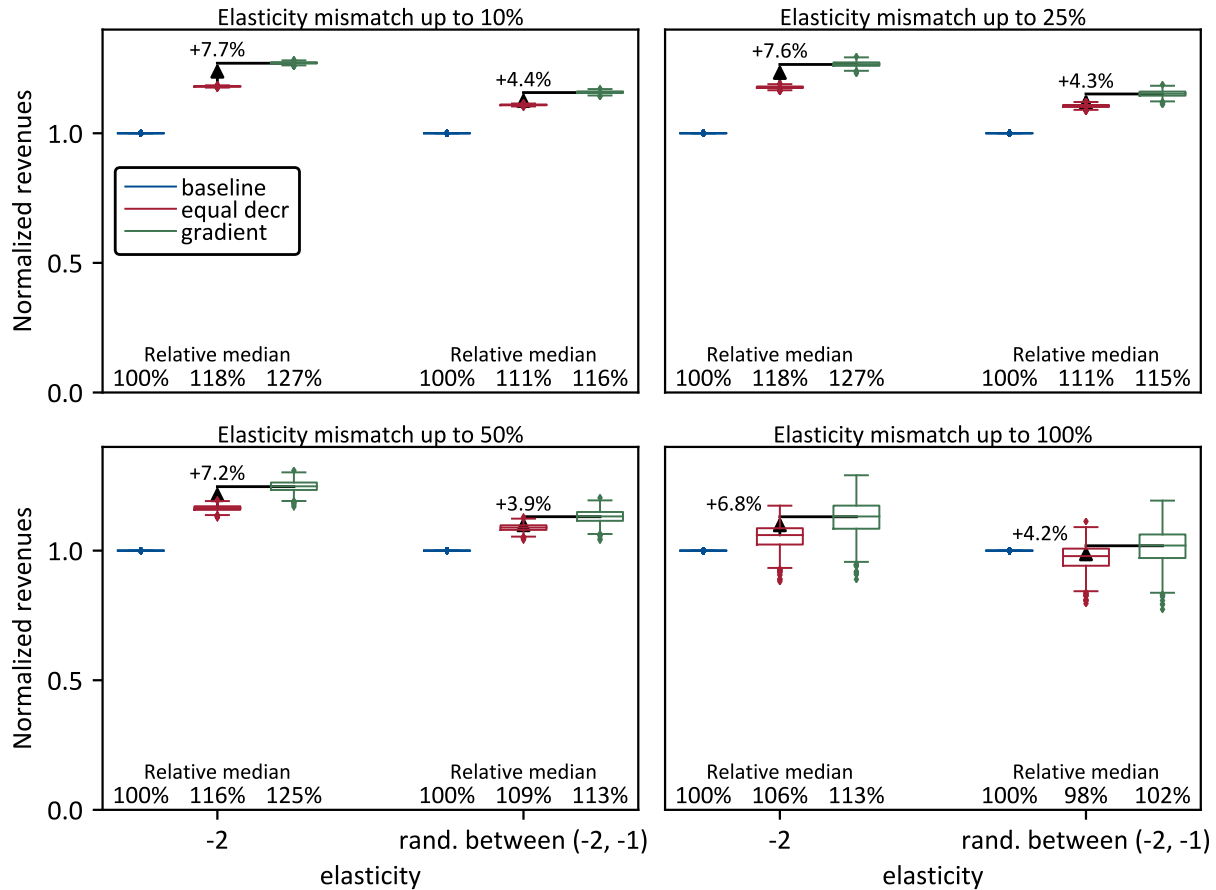


Figure 8-29: Results for a mismatch between the believed and the actual elasticity for two elastic cases and the two algorithms. The scenario is selling more through existing SLAs from Section 8.4.1.

The results reveal the impact of the estimation error on the uncertainty in the achievable revenues, but they also illustrate the difference in the expected medians. The reduction of the elasticity estimation error from 100% to 50% yields 12% more revenues for the gradient approach and -2 elasticity (from 113% to 125%). A further decrease from 50% to 25% results in another 2% lift. The behavior follows a diminishing return function. The pricing algorithms seem to be robust against an elasticity estimation error under 25%.

In sum, the analysis of the elasticity estimation error supports three conclusions:

- Both pricing algorithms appear to be robust for errors below 25%
- Reducing the error follows a diminishing return function; therefore, early efforts in reducing the elasticity estimation error are especially beneficial.

- The gradient approach provides a consistent 4-8% lift, which emphasizes the dominance of more sophisticated RM over more basic pricing approaches.

Our elasticity regression attempt in Section 8.1.2, based on actual historical purchasing data, showed limited success with R^2 numbers between 0.05 and 0.2. Hence, it is conceivable that a 100% mismatch between believed (=computed) and actual elasticity exists. The following subsection reviews literature, which addresses learning methods for elasticity estimation.

Review of elasticity learning method

The Operations Research community contains extensive work on the joint learning-and-pricing problem, i.e., the trade-off between exploration (=learning) and exploitation (=revenue generation). The two survey papers from Aviv and Vulcano [244] and Boer [245] give a broad overview of this research field [246].

When substantial uncertainty exists in the demand elasticities, *price experimentation/discovery* is an effective method to generate additional data points, learn from them, and reduce the uncertainty. Some central papers on this topic are Besbes and Zeevi [247, 248], Boyacı and Özer [249], and Wang et al. [250].

However, as Cheung et al. [246] point out, these authors do *not* consider practical business constraints in price experimentation, e.g., in particular, the number of price changes. In contrast, Cheung et al. propose a method that takes the maximum number of price changes m during T periods. The authors' objective is the minimization of the *regret*, i.e., the difference between the revenues with full information versus the actual achievable numbers with uncertain information given m price changes. They demonstrate a pricing policy where the regret scales with m iterations of the logarithm $\mathcal{O}(\log^m T)$. Figure 8-30 illustrates the phasing of the pricing policy with $m - 1$ learning phases (exploration) and one final earning phase (exploitation).

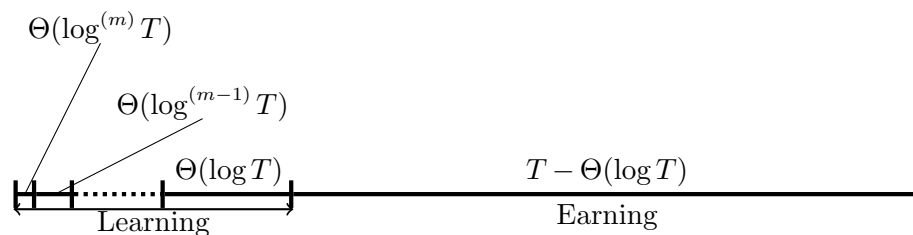


Figure 8-30: schematic of the pricing policy proposed by Cheung et al. Figure duplicated from the authors original publication [246, p.18].

Cheung et al. [246] collaborated with Groupon to test their algorithms. When Groupon launches new deals, they have a high demand uncertainty since no previous sales data is available. The analogy to

satcom is the introduction of new service offerings, such as novel SLAs. Before the work from Cheung et al., Groupon set a fixed price on each deal for the complete duration of the offer (usually between a few weeks and many months). Groupon desired that the price in only changed once for each deal due to fearing a negative customer response or customer confusion (minimizing price changes becomes even more relevant in b2b like satcom with longer contract durations). The authors construct a finite, linear demand function set for the initialization. When the algorithm suggests a price change, the new price is always *lower* (aligning well with the direction of prices in satcom). Over 1,000 deals show that, on average, the number of bookings increases by 116% relative to the first price set, with a revenue increase of 22%.

One critical insight is the diminishing return behavior (through the logarithm) between the number of price experiments and regret. As the field experiment with Groupon shows, a *single price change* ($m = 1$) *already achieves a significant revenue lift*. This result is promising for satcom since price experimentation opportunities are limited in the current b2b satcom market with rather rigid long-term SLAs.

Nevertheless, future research is vital to understand the details and limitations of applying such approaches to satcom. For example, Cheung et al. [246] assume T periods in the order of weeks and months, and repeated purchases in each period. In satcom, T is on longer timescales of multiple years, and customers make purchases less frequently. Hence, external conditions between purchases might change significantly, requiring additional steps to make purchases comparable.

8.7 Summary of the four analyses

The previous two Sections, 8.4 and 8.5, presented four different approaches for monetization of the available capacity. We compared three algorithms: equally decreasing heuristic, highest yield approach, and gradient optimization to mimic different sophistication of the RM system. The reported results contain two central metrics: first, the additional revenue improvement extracted from the 38% available capacity (made available by dynamic resource allocation). Second, the lift of the more sophisticated gradient optimization RM compared to the second-best heuristic. Table 8-11 and Figure 8-31 summarizes the data.

Table 8-11: Comparison of the additional revenues potential and the revenue lift from sophisticated RM across the four analyses.

| | Selling more to existing customers | | Contracting new customers | |
|---|--|--|---|--|
| | Selling more capacity through existing SLAs | Selling additional products | Unlocking affordability elasticity | Increasing market share |
| 38% available capacity is translatable into how much additional revenues | 0-27% | [2-102%] # | 0-39% | [0-78%] # |
| Critical assumption | | No internal cannibalization | | Competitors do not respond |
| Range depends on | <ul style="list-style-type: none"> • Demand for use elasticity • Sophistication of pricing algorithm | <ul style="list-style-type: none"> • Demand elasticity • price point for add. product • Sophistication of pricing algorithm | <ul style="list-style-type: none"> • Affordability elasticity • Sophistication of pricing algorithm | <ul style="list-style-type: none"> • Stickiness of customers with competitor • Sophistication of pricing algorithm |
| Separate price for | Customer | Customer | Segment | Segment |
| Price affects | <ul style="list-style-type: none"> • Demand for use | <ul style="list-style-type: none"> • Demand for add. product | <ul style="list-style-type: none"> • Access demand | <ul style="list-style-type: none"> • Market share |
| Price does not affect | <ul style="list-style-type: none"> • Access demand • Market share | | <ul style="list-style-type: none"> • Demand for use • Market share | <ul style="list-style-type: none"> • Demand for use • Access demand |
| Lift in revenues through more sophisticated RM over heuristic pricing (gradient optimization over second best heuristic) | 5-7%* | [1-9%] ** | 4-6%* | [28-39%] ** |
| # ranges in brackets should be taken with care due to the listed critical assumption | | | | |
| * ranges exclude the unit elastic case, for which the algorithm cannot influence the revenues | | | | |

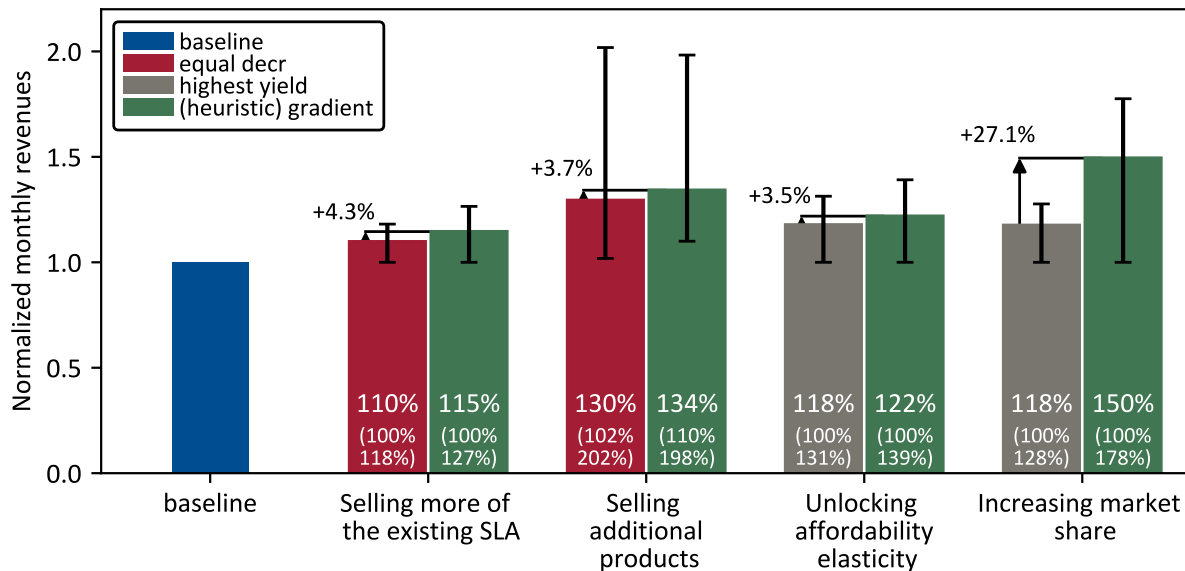


Figure 8-31: Comparison of the results across the four analyses. Besides the (heuristic) gradient approach, the second-best heuristic is plotted. The percentage numbers in parentheses indicate the error bar numbers, which are the minimum and maximum numbers. The bars show the means.

The wide ranges in the numbers are due to considerable variation in our assumptions about elasticities and stickiness. The boundary case is a unit elastic demand, in which the additional revenues are 0% (except for the additional product as discussed in the corresponding Section).

The selling of additional products has the most extensive upside range. However, this number should be taken with care for two reasons: first, our assumptions about the uncertainty vary the most, and hence we expect a broad range of possible revenues. Second, we did not account for the effect of internal cannibalization, which is likely to affect the revenue generated by the base Classical SLAs. If there is a considerable internal cannibalization effect, the value of additional products is smaller. We expect that cannibalization varies between products and segments and might be best estimated through sales expert knowledge. A realistic assessment of internal cannibalization would allow for tightening the range of possible revenue gains and allowing for more nuanced conclusions regarding additional products. We believe it is a valuable future research direction, as new products could provide an effective way to prevent (or at least slow down) a potential race-to-the-bottom. It allows operators to differentiate themselves and make service offerings stickier to customers, giving the operator a first-mover advantage with novel SLAs.

The most directly comparable numbers are between the selling-more-through-existing-SLAs analysis and the unlocking of the affordability. The reason is that we make the same assumptions about the elasticities

(ranging from -2 to -1). The difference is that, on one side, the elasticity is regarding the demand for use; and on the other side, the analysis leverages the affordability elasticity (access demand). We see that for the upper value of the ranges, unlocking the affordability elasticity is more valuable for the same elasticity assumptions. The cause of that behavior is the non-linearity between the data rate and power through the link budget. For every increment more data rate, the link requires an unproportionally larger increment of power. Another way to phrase that relationship is that the data rate has a diminishing return concerning power. In the analysis where the operator increases the data rate for existing customers, the starting point on this curve is already at a higher power and data rate level. For new customers, this starting point is at the minimum MODCOD, and therefore close to zero power and data rate. For an equal increase in the data rate, more power is required in the existing customer case than for the new customer. The aggregation of these effects makes the unlocking of affordability elasticity more valuable.

The increasing-of-the-market-share analysis has the second-highest range. That is because of the considered discount factors: the price reduction is smaller than for unlocking the affordability elasticity analysis. It can be seen as a cheaper way to get new customers. However, we assume that the competitors do not respond, which is an idealistic assumption and does not represent the long-term. Additionally, the revenue increase is associated with time dimensions as customers are expected to switch not before their contracts expire. Because of these two factors, the number should be taken with care and considered more an upper bound. Future research can be to extend the market dynamics model from Chapter 2 to a single operator level and link it to the market-share analysis. With this setup, multiple scenarios of competitive behavior could be analyzed, and the range of potential revenue gains further narrowed. If the results indicate that all competitors in the market are likely to respond timely with short time-delays in the dynamics, a potential race-to-the-bottom would be accelerated. We discuss the first mover advantage of RM with corresponding satellite technology in the competitive environment further in Section 8.8.

In the following, we switch our focus to the discussion of the second metric: the lift through more sophisticated RM techniques. The increasing market share approach is here an exception with a range between 28% and 39%. We examined the reasons in the corresponding Section, and we believe that better heuristics can reduce this gap considerably. More commonly, we observe a lift between 1% and 9%. Note that these ranges do *not* include the unit elastic case (or the 50% discount case in the market share analysis) since the algorithms do not influence the revenues. However, the average lift numbers shown in Figure 8-31 include this case. The interpretation of these numbers should be how *efficient* the pricing principles are with the available capacity. Our intention with the equally decreasing and highest

yield heuristic is to approximate the current pricing policies operators might follow. Comparing to this baseline, a more sophisticated RM system can then use the capacity more efficiently, and hence, increase the revenues by 1-9%.

So far, we looked at the four analyses separately. Our assumptions separated the effects of price adjustment to a subset: either the increase of demand for use or the access demand. We discuss interaction effects in the subsequent.

Interactions between the analyses

Generally, selling additional products is more decoupled from the other three analyses as the prices are set for a different product, in contrast to setting prices for the Classical SLA¹⁴. Price changes have the following two central effects: the demand for use and the access demand vary along with their elasticities (there are many more, but these are the two that we modeled). Selling more of the existing SLA covers the demand for use. In contrast, we further separate the access demand into unlocking the affordability elasticity and affecting the switching behavior of the competitors' customers. As a result, these three analyses are undoubtedly linked through the price. We believe that developing an integrated algorithm that trades off all these factors is a prized and challenging future research direction. We hypothesize that the integrated optimization has the potential to increase the lift from the gradient approach further.

Despite the coupling between the analyses, the operator can make decisions about the additional customer demand it wants to contract. For example, the operator reduces prices to unlock the affordability elasticity, and the price reduction also increases the demand for use. The operator can decide only to contract the new customers and leave the existing SLAs at their current CIR (since new customers are less expensive to serve than to increase throughput to existing). In this scenario, the demand is greater than the capacity, and the markets are no longer in equilibrium. The four analyses aim to provide insight into some of these trade-offs.

In sum, we draw the following conclusions:

- The monetization of the 38% available capacity depends significantly on the demand-for-use and affordability elasticities, as well as customer stickiness and varies between 0-39%.

¹⁴ We assume here that the unlocking and increasing market share occurs through the Classical SLA. If additional products play a role there as well, then the coupling with the selling of additional products is greater.

- The unlocking affordability approach has a more beneficial benefit-to-cost relationship than the selling of more demand to existing customers.
- Selling additional products has the highest potential but also the greatest uncertainties.
- Reducing prices to increase market share is an option that might be considered
- More sophisticated RM techniques, such as the developed gradient optimization, lift the revenues between 4-7% over the best heuristic pricing approach.

8.8 Implications on the market when operators adopt RM

The purpose of this Section is to discuss market implications when operators decide to implement an RM system. Is there a first-mover advantage with implementing an RM system, and if so, is it sustainable over time? What happens if everyone is becoming more sophisticated? Can a sophisticated RM system be a long-term differentiator?

We first address the first-mover advantage of implementing an RM system and then transition to a scenario in which all operators adopt more sophisticated RM. The central components that define the sophistication of our satcom RM system are *resource allocation*, *available capacity forecasting*, *customer elasticities estimation*, and *pricing optimization* (see Figure 8-1). Improvements in any of these components produce a better, more sophisticated RM system that extracts greater revenues from a given capacity and demand:

- Better resource allocation algorithms for user terminal grouping, routing, frequency plan, and power allocation reduces the used capacity of the baseline scenario (see Chapter 5 and Section 8.2).
- More accurate forecasting of the available capacity reduces the uncertainty of the used capacity for the baseline and therefore increases the available capacity (see Chapter 6 and Section 8.2).
- More accurate customer elasticity estimation reduces uncertainty and increases the median revenues (see Section 8.6).
- Better pricing optimization algorithms increase the revenues extracted (see Section 8.3 and Table 8-11 in Section 8.7: 4-7% lift of gradient approach over best pricing heuristic).

We can summarize the first two points into *more efficient use of capacity* (resource management), and the last two points into a *more efficient monetization* of the available capacity (demand management). Independent of the market conditions, and the competitors' response, more efficient use of capacity is never a disadvantage. The edge case of no advantage is if global and local excess capacity exists, and less efficient use of capacity would provide the same QoS to customers. However, more efficient use of capacity does not impact the bottom line as long as the operator does not sell the available capacity to existing or new customers.

The picture of more efficient monetization is more convoluted. Increased efficiency is associated with excess capacity, and therefore, a price decline (measured on customer/segment level). Depending on elasticity and competitive environment, there might be a first-mover advantage, which could be

sustainable in the long-term. To discuss these cases, we consider two competitors A and B with the same hardware, i.e., satellite communication system but different management of demand and capacity (similar to Section 8.5.2). Furthermore, we study two different market conditions: imperfect competition, where operators have pricing power and price information is not necessarily available to competitors, and perfect competition with no pricing power and complete price transparency. The former condition represents the majority of the satcom market segments today (few operators and limited price transparency due to individually negotiated and confidential SLAs [66]), while the latter condition is close to the airline industry and one potential future for satcom.

Table 8-12 summarizes the implications for competitor A and B. We consider two different phases. First, competitor A more efficiently monetizes the available capacity compared to competitor B. Then, competitor B catches up and adapts the same algorithms as competitor A. We discuss each of these phases under the two market conditions: imperfect and perfect competition.

Table 8-12: competitive advantage of more sophisticated pricing algorithms in different scenarios

| | Imperfect competition (pricing power and no clear price transparency) | Perfect competition (no pricing power and complete price transparency, A is price leader) |
|---|--|--|
| Competitor A more efficiently monetizes the available capacity than competitor B | <ul style="list-style-type: none"> • A has first-mover advantage and can sustain it • B has potentially lower revenues than before • If total demand is inelastic: A has lower revenues in the short-term but might benefit in the long-term • If total demand is elastic: A has higher revenues, and B might be able to sustain its revenues. | <ul style="list-style-type: none"> • A has first mover-advantage but cannot sustain it • B matches prices of A and both end up in same situation (A might be worse off given the expense of the RM system) • If satellite system different: A optimizes for its cost structure and when B matches, prices are less optimal for B. A is likely better off and can sustain advantage. • If total demand is inelastic: A and B are worse off than before • If total demand is elastic: A and B are both better off than before |
| Competitor A and B have the same efficient monetization algorithms | <ul style="list-style-type: none"> • A and B end up in same situation but there is a risk of a “race to the bottom” • If total demand is inelastic: A and B are both worse off than before • If total demand is elastic: A and B are both better off than before | |

When competitor A decides to implement a more sophisticated RM than B, A might have a sustainable first-mover advantage in an *imperfect environment*. The pricing from A is not transparent to B, and therefore B cannot easily match it. Over time, B might discover the actual prices, but A potentially keeps optimizing and changing their prices, making it challenging for B to catch up. When the total demand is

inelastic, A has lower revenues in the short-term due to price reductions. However, customers from B might want to switch to A due to better pricing, leaving B with a decision to make: lowering prices or lose market share.

Decreasing prices could have a delayed effect as some customers might have already switched to A and signed a 1-3 years long SLA. Additionally, B's required price reduction to keep its customers might not be precisely known to B as price transparency is limited, and SLAs are individually negotiated and confidential. When B chooses prices that are too high, B could continue to lose market share, and if they are too low, B missed revenue opportunities. Given the delays, a race-to-the-bottom in an imperfect environment might take place over a more extended period.

If B does not match prices, A could increase its revenues in the long-term due to the increased customer base – despite potentially inelastic demand. If demand is elastic, A potentially increase its revenues in the short- *and* long-term, and B might be able to sustain its revenues since the whole market is growing.

We suspect that the reaction of B is somewhere between these two options. Since not all segments might be equally important to A and B, B could decide to decrease prices and retain its customers in some segments while accept losing market share in others.

If the *competitive environment is perfect*, A still has the first-mover advantage since it reaches the customers first with the new prices. However, it is not sustainable since B can directly observe and match the prices of A. Both end up in the same situation, while A might be financially slightly worse off due to the RM system's expenses. A different picture shows if the satellite system itself is different. A optimizes and reduces the prices for its cost structure (which depends on the satellite system), and if B decides to match these prices, they are likely less optimal for B (since their cost structure is different from A). Since both competitors are matching prices, they end up with lower revenues if total demand is inelastic and more revenues if demand is elastic. B might choose to not match prices in all segments and create their own pricing based on its less sophisticated RM. In that case, the relative advantage (or potentially, but less likely, disadvantage) of A over B depends on the specific relationships between the cost structures and prices across the segment.

Different products, such as discussed in Chapter 7 and partly analyzed in Section 8.4.2, add another dimension to the competition between A and B. Product features allow for price differentiation within a segment.

If B now decides to become equally sophisticated in their RM than A, the competitive environment plays a smaller role when the satellite systems are the same (because A and B optimize their prices based on a similar cost structure). They both end up in a similar market share situation but with lower prices than before. A and B have lower revenues when total demand is inelastic and higher revenues for elastic total demand (since the entire market is growing in terms of revenues). The situation becomes more differentiated when we consider that A and B have different satellite systems. In an environment with limited price transparency, both might specifically focus on optimizing the prices to fit their cost structure. When the prices are more transparent, the market might drive pricing to a larger extent. A or B win customers in the segments where their cost structure is more beneficial. Hence, the cost structure, i.e., resource management, becomes a crucial competitive advantage.

With these insights, we can answer the question we raised at the beginning of this Section:

Is there a first-mover advantage, and if so, is it sustainable over time? Yes, there is a first-mover advantage in a perfect and imperfect competitive environment. It is sustainable over time if the market has limited price transparency. This is the case in most satcom segments today since SLAs are individually negotiated and confidential. An exception might be residential broadband where prices are more publicly available.

What happens if everyone is becoming more sophisticated? All competitors end up in the same situation when they have the same satellite system and are at risk of entering a downwards price spiral (potentially a “race to the bottom”). All increase their revenues if demand is elastic and reduce their revenues if demand is inelastic. The efficiency of using the capacity (resource allocation) becomes a crucial competitive differentiator.

Can a sophisticated RM system be a long-term differentiator? Yes, RM’s demand management aspects provide benefits in markets with limited price transparency and between operators with different satellite systems. If the market drives prices, the RM’s resource management aspect becomes the differentiator.

In general, we expect that the market moves from the “upper left” to the “lower right” of Table 8-12. We believe it is more likely that the competitors match the RM sophistication before the market transitions into more price transparency (so, first “down” and then to the “right”). Since the airline industry is in the latter, lower right condition, we want to loop back and draw some final comparisons between satcom and the airline industry, which was our initial motivation for RM.

Comparison of satcom RM and airline RM

Market implications

Airlines deployed RM “to ensure that low-fare leisure passengers do not consume all of the seats on high-demand flights ... [RM systems] forecast demand and calculate the number of seats to be made available to each fare type, with the goal of maximizing total flight revenues” [251, p.3]. The optimum point is a trade-off between the load factor and yield. The first studies looked at one airline without considering the competitive or market context. They commonly quoted a lift of 2-5% [1, 2, 5] (comparable to the 4-7% we found for satcom). In 1997, 10 years after the first academic work on RM [2], Belobaba and Wilson [251] studied the impact of RM in a competitive environment. They addressed similar questions to the ones outlined at the beginning of this Section. How does the introduction of RM affect the airline who introduces it and its competitor? Is there a first-mover advantage? And how does it affect the total market revenues when all airlines implement an RM system? Is RM a zero-sum game?

Aligned with our conclusion for satcom, Belobaba and Wilson [251] found that RM offers a first-mover advantage, and airlines without RM have lower revenues than before. They analyzed a similar two-competitor environment. The authors conclude that if both airlines adopt RM systems, their total revenues increase. Airline RM is, therefore, *not* a zero-sum game (which it would not be if only market shares shift, such as in our fourth analysis in Section 8.5.2). Our first three analyses in Sections 8.4.1 – 8.5.1 provide insights into the satcom analogy. In these, we do not change the market share and extract the additional revenues by using the capacity more efficiently through a more sophisticated pricing algorithm. Satcom, therefore, is also *not* a zero-sum game. When all operators implement an RM system, we expect each operator’s revenues to increase if total demand is elastic.

In airline RM, the revenue increase comes from filling previously excess capacity with lower fares (i.e., increasing load factor from an average 67% in 1995 to 84% in 2018 [252]), a better fare mix, and from protecting seats for high-fare passengers, which tend to book later [251]. In satcom RM, the revenue increase comes from more efficient use of the capacity. However, for satcom RM, the customer elasticity plays a key role. If the demand is unit elastic, the demand management aspects of the RM system provides no value.

Excess demand or excess capacity

As a further comparison between airline and satcom RM, we want to discuss the difference in the initial market conditions. In the airline industry during peak travel periods, *demand is greater than the capacity*

for the discounted fares. Amongst the functionalities of airline RM is to protect seats for later booking of higher paying customers.

The load factor of airlines averaged around 55% in 1965 [253], which is similar to the range we find in the initial conditions of our satcom simulations. Under this scenario, demand is not necessarily greater than capacity. Especially in the early life of a satellite/constellation, *capacity is greater than demand*. Another example is improvements in the resource allocation algorithms that free up capacity (the analogy to airline would be technological improvements that allow more seats per plane while offering the same customer experience). The question then becomes how to monetize that extra capacity.

In the final months of this dissertation, the COVID-19 crisis severely affected the airline industry, with an 80% decline of international flights in May 2020 [254]. The demand suddenly dropped and became *smaller* than capacity, creating demand-supply conditions similar to the satcom industry. On the one hand, some of the concepts discussed in this dissertation could become relevant for airline RM (as long as the demand decline prevails). On the other hand, the aftermath of how airlines reacted and how they adjusted their RM systems might provide valuable insights for future satcom RM research.

To summarize this Section, we discovered the following conclusions of RM on a competitive market level:

- Satcom RM has similar to airline RM a first-mover advantage. This advantage is sustainable if the market has limited price transparency.
- Satcom RM bears the risk of a “race to the bottom” since dynamic resource allocation creates available capacity. If demand is inelastic, the revenues of all competitors would decrease.
- Satcom RM can be a long-term differentiator. The demand management aspects generate benefits when price transparency is limited, and satellite systems are different. The resource management aspects become the central RM differentiator when operators are price takers.
- Satcom RM is *not* a zero-sum game. When all operators implement more sophisticated management of demand, every operator increases their revenues.
- Satcom RM generates its revenue lift by more efficient use of the capacity.

8.9 Summary and conclusions

In this Chapter, we brought the pieces of the satcom RM framework together and analyzed its value under different assumptions. We outlined a pricing optimization approach that works with a computationally expensive evaluation function, such as our resource management part of the framework. We set up a simulation that contains actual data when possible and used this as a baseline throughout the Chapter. From there, we first quantified the benefit of dynamic power allocation to be 38% whole-day available capacity. Second, we looked at four different scenarios of monetizing this 38%: selling more of existing SLAs, selling additional products, unlocking affordability elasticity, and increasing market share. We synthesized this information and discussed the interactions between the four analyses. Furthermore, we studied the impact of the elasticities' estimation errors and discussed the implication of an RM system on the market level.

The results let us draw the following conclusions:

- Stationary CIR allocation reduces capacity usage by 25% compared to fixed by design allocation. Dynamic allocation for actual traffic reduces it by 63% and frees up 38% whole-day available capacity compared to the stationary CIR allocation.
- The 38% available capacity is translatable into 0-39% of additional revenues, mainly depending on customers' price elasticity.
- The approach of selling capacity to new customers has a more beneficial benefit-to-cost relationship than selling capacity to existing customers
- Monetizing the available capacity through additional products has the highest potential but also the most significant uncertainties. It comes with the risk of internal cannibalization.
- Reducing price to increase market share is an option that might be considered but requires careful balancing. It bears the risk of initiating a "race to the bottom".
- More sophisticated pricing algorithm within the RM framework lift revenues between 4-7% over the best heuristic pricing approach. This benefit is sustainable if the market has limited price transparency or competitors' satellite systems are different.
- Satcom RM is *not* a zero-sum game since it improves revenues by using the capacity more efficiently.

9

Conclusions, Contributions, and Future Work

In the first Chapter of this dissertation we asked the question of how broadband satcom operators can monetize the flexibility of digital payloads and phased arrays. By aiming to answer this question, we raised four other questions: how do the dynamics, the uncertainties, and the trends in the satcom market affect the strategy of satcom operators to monetize freed up capacity? Given the tight coupling between economics and technical aspects of the questions, can we leverage Revenue Management (RM) techniques from other industries? What challenges need to be overcome when adopting a framework from another industry to satcom? How and how much value can satcom operator extract by implementing such a framework?

We answered these four questions in the seven preceding Chapters 2-8. Chapter 2 developed and discussed the market dynamics between customers, service providers, and operators. We reviewed RM frameworks extensively to build our satcom RM framework in Chapter 3. The Chapters 4-7 provide solutions to the four challenges of a satcom RM system. Besides the RM framework, the resource allocation challenge in Chapter 5 is the second centerpiece of the dissertation. We comprehensively applied and analyzed the opportunities and challenges of the RM framework with four different simulations in Chapter 8.

While all Chapters have their concluding remarks, the objective of this final Chapter is to summarize and connect the conclusions (Section 9.1), reiterate on the contributions of this work (Section 9.2), and discuss a range of opportunities for future research that this dissertation unwraps (Section 9.3).

9.1 Conclusions

We group the conclusions by the four questions and refer to the Chapter or Section that supports the statement.

How do the dynamics, the uncertainties, and the trends in the satcom market affect the strategy of satcom operators to monetize freed up capacity?

- Prices can be expected to decline as operators continue to increase supply further. (Chapter 2)
- The main three uncertainties in the satcom market are the success or failure of large NGSO constellation, the total demand elasticity, and the additional demand generated by NGSO constellations. All three impact the price and revenues of GEO operators similarly. At the same time the demand side (the success of NGSO constellation) drives the revenues of NGSO operators, which is also the most uncertain. (Chapter 2)
- Vertical integration and shortening the SLA durations between customers, service providers, and operators can boost total market revenues by 22-24% for operators and service providers. (Chapter 2)
- Satcom RM has a first-mover advantage where price-changes can affect the market share between operators. RM becomes a strategic tool for an operator to compete and to use its capacity efficiently. Since an efficient RM frees up capacity, there is a risk of a “race to the bottom” (Sections 8.5.2 and 8.8).
- Satcom RM is *not* a zero-sum game since it improves revenues by using the capacity more efficiently. (Section 8.8).

Given the tight coupling between economics and technical aspects of the questions, can we leverage Revenue Management (RM) techniques from other industries?

- Yes, Revenue Management is highly applicable to satcom based on these six conditions (Section 3.2):
 - capacity is inflexible
 - capacity costs are high compared to marginal sales cost
 - inventory is perishable
 - customers are heterogeneous and segmentable
 - demand is variable and uncertain
 - the organization has data and information system infrastructure

What challenges need to be overcome when adopting a framework from another industry to satcom?

- Four central characteristics of the satcom industry are different from other RM industries (Section 3.4):
 - the unit of demand (Mbps) is not the unit of capacity (W)
 - the resource allocation is an optimization problem itself
 - the available capacity depends on the customers' usage level
 - existing SLAs do not fully leverage the new flexibilities
- Our proposed satcom RM framework includes one solution component for each challenge (in addition to the more common RM components, Section 3.5):

1. The satcom simulator component supports the bi-directional translation between demand in Mbps and capacity in W. (Chapter 4)

- Validation showed an average relative error below 0.4%
- Vectorization of the main computational steps provides significant computational improvements

2. The resource allocation component finds the beam pointing, the frequency assignment, and the power levels for a set of user terminals and their traffic pattern. (Chapter 5)

- The presented resource allocation process and the algorithms assign the resources beams, frequency, and power to the user terminals successfully.
- The heuristic beam placement achieves a lower number of beams than any other approach found.
- The heuristic frequency assignment algorithm only uses around 25% of the spectrum – leaving significant room for improvement.
- The balanced gateway allocation reduces the maximum CIR per gateway by around 50% while having less than 1% more free space loss.
- The power allocation is computationally efficient and ensures demand is met at all times.
- When a satellite is sharing the spectrum for user and gateway downlinks, there is an additional trade-off between balancing gateways and blocking of frequency for gateways in high-density regions.
- The periodicity between the ground repeating pattern (in the MEO example, four visits per day) and the customers' traffic usage seasonality (here 24 hours) results in a slightly different average and peak power for each satellite. The operator might wish to position the satellites accordingly based on the results of the acceptance tests.

3. The available capacity forecaster probabilistically estimates the available capacity given the customers' usage history. (Chapter 6)

- Research on electricity load forecasting provides a valuable analogy for the time-series estimation part of this challenge.
- The Gaussian process regression on the data performed well but was not suited for our purposes. Therefore, we use a stochastic process of normal random variables for the estimation of a *typical day*.
- We use a stochastic process of empirical normals to estimate the used capacity on the satellite level.
- The contracted availability is an essential driver of the available capacity: a reduction from 99.9% to 98% yields an additional 38% more available capacity.

4. Novel SLAs and customer segmentation provide alternatives to the existing SLAs. (Chapter 7)

- The existing SLAs can be classified as Classical SLA, Data volume SLA, and Dual SLA.
- The shortcomings for the operator are that they do not allow for safe overbooking and do not incentivize to smoothen the daily pattern.
- The shortcomings for the customer are that they are expensive and not flexible.
- By reviewing of the telecommunication and cloud computing industries, we found the following novel SLAs through analogy: spot instance, time-of-day pricing, and Two Classes of Service.
- Segmenting the market and tying it back to the SLAs reveals these conclusions
 - Classic SLA excel where reliability and speed are essential, traffic volume is high, and variation is low.
 - When consumers are price-sensitive, data volume SLAs, time-of-day pricing, and Two Classes of Service are affordable options.
 - Dual SLAs are a good compromise across all segments.
 - When the traffic patterns are not repeating on a daily or monthly basis, spot instances offer the desired flexibility.

How and how much value can satcom operator extract by implementing such a framework?

- The dynamic resource allocation part of the framework frees up 38% whole-day available capacity (Section 8.2).
- We analyzed four different ways to monetize this 38%: selling more of existing SLAs, selling additional products, unlocking affordability elasticity, and increasing market share (Chapter 8).

- The 38% available capacity is translatable into 0-39% of additional revenues, mainly depending on customers' price elasticity (Section 8.7).
- The approach of selling capacity to new customers has a more beneficial benefit-to-cost relationship than selling capacity to existing customers (Section 8.7).
- Monetizing the available capacity through additional products has the highest potential but also the greatest uncertainties. It comes with the risk of internal cannibalization (Section 8.7).
- The effect of price reduction on market share is an option that might be considered but requires careful balancing. It bears the risk of initiating a "race to the bottom" (Section 8.7).
- More sophisticated pricing algorithm within the RM framework lift revenues between 4-7% over the best heuristic pricing approach. This benefit is sustainable if the market has limited price transparency or competitors' satellite systems are different (Section 8.7).

9.2 Contributions

We start this Section by listing our contributions in Table 9-1 grouped by Chapter. While Chapter 4 is necessary for the satcom RM framework to function, it does not contain novel scientific contributions, and hence we do not list it in the table.

Table 9-1: Summary of contributions by Chapter.

| Chapter | Contributions |
|--|---|
| 2. Satcom Market Dynamics | <ul style="list-style-type: none"> • Developed a satcom market dynamic model on which future research can be built. (t) • Analyzed how the flow of capacity and vertical integration affects the bottom lines in the future, given the significant uncertainties in the market. (p) |
| 3. Satcom Revenue Management Framework | <ul style="list-style-type: none"> • Identified Revenue Management as highly applicable to satcom. (t) • Contrasted six industries with satcom using our taxonomy. (t) * • Identified that resource management is a crucial dimension for satcom RM, not considered by current RM research. (t) * • Proposed a satcom RM framework that captures the complexity of satcom. (t) • Identified four challenges of satcom RM: the unit of demand is not the unit of capacity, resource allocation, available capacity forecaster, and novel SLAs. (t) |
| 5. Resource Allocation | <ul style="list-style-type: none"> • Formalized and decomposed the resource allocation process into four sub-problems. (t) • Extended the frequency plan algorithms from Pachler et al. [188] to include gateways. (t) • Developed a closest-first and balanced gateway allocation algorithm for the routing sub-problem. (t) • Developed a direct RF power allocation strategy. (t) • Demonstrated the functioning of the resource allocation process with an application to a seven satellite MEO constellation and 5,000 user terminals. (p) |
| 6. Available Capacity Forecaster | <ul style="list-style-type: none"> • Identified the need for a <i>typical day</i> forecasting for satcom RM. (t) • Drew the analogy between electricity load and satcom user traffic forecasting. (t) • Proposed a computationally inexpensive method: a collection of independent normal random variables with several practical advantages. (t) • Quantified the sensitivity of the available capacity to the SLAs' contracted availability. (p) |
| 7. Novel SLA | <ul style="list-style-type: none"> • Identified and described currently used SLAs in satcom. (t) • Characterized limitations of existing SLAs in the context of new flexible payloads and RM. (t) • Built analogies to the cloud service and telecom industry. (t) • Proposed a novel set of SLAs that are beneficial to operators and customers. (t) • Discussed market segmentation bases and mapped the resulting ten segments to the three classical and three novel SLAs. (t) |
| 8. Application of the Satcom RM Framework | <ul style="list-style-type: none"> • Proved the proposed RM framework's value by implementing Chapters 4 – 7 and applying it to several scenarios using real data from a satellite operator. (p) • Developed a gradient-based pricing optimization approach that works with expensive evaluation functions such as the resource management part of the framework. (t) • Showed that a 38% available capacity can be translated into 0-39% additional revenues and that the more sophisticated gradient-based approach consistently outperforms pricing heuristics by 4-7%. (p) |

* Contribution to RM research; (t) theoretical contribution; (p) practical contributions

We identify the type of contribution with (t) being a theoretical, and (p) a practical contribution. The majority of the contributions are in the satcom research field. However, two center theoretical contributions are in the field of RM research (marked with an asterisk): contrasting the six industry with a taxonomy we developed, and following from that, the identification of resource management as a key dimension of satcom RM. This discovery provides a theoretical feedback loop to the RM research community.

9.3 Future work

Future work lies mostly in the market dynamics, Chapter 2, in the components of the satcom RM framework, and the application of the framework. The following three subsections list the future work that could follow this dissertation.

Satcom market dynamics

- Analyzing the market dynamics when competitors react differently with capacity expansion and pricing strategies.
- Coupling the market dynamics with the RM framework, especially for a more detailed analysis of the RM's competitive effect. A study could include several scenarios of operators implementing RM systems with various sophistication levels.
- Analyzing the reduction of SLA durations in various competitive contexts.

Components of the satcom RM

- Resource allocation
 - Integrated optimization of the complete resource allocation process
 - Improving the filling rate of the frequency plan
 - Defining the trade-off space between gateway balancing and frequency assignment
- Available capacity forecaster
 - Exploring more sophisticated approaches to estimate a typical day based on a historical time series.
 - Considering the covariance between the random variables in the stochastic process
- Novel SLAs
 - Quantifying the value of the novel SLAs to the customer. Then, simulating the impact on the operator through RM simulations.

- Conducting a more exhaustive customer survey to identify needs and testing responses to different SLAs and pricing.

Application of the framework

- When flexible HTS satellites are launched, collecting real data on usage, power consumption, and pricing responses for further analyses.
- Exploring the applicability and the value of the framework to *traditional* satcom, where flexibility is more limited. Can RM help to extract value?
- Studying the application of the framework for filling the capacity of the satellite with customers over time.
- Deepening the analysis of how the satcom RM framework provides a competitive advantage under different competitive environments.

A. What is system dynamics?

Inspired from [255], the text enclosed in quote marks & cited below is the official description of System Dynamics from the System Dynamics Society [256] which adapted it from Richardson [257].

“Overview

System Dynamics is a computer-aided approach to policy analysis and design. It applies to dynamic problems arising in complex social, managerial, economic, or ecological systems—literally any dynamic systems characterized by interdependence, mutual interaction, information feedback, and circular causality.

The System Dynamics Approach

The approach begins with defining problems dynamically, proceeds through mapping and modeling stages, to steps for building confidence in the model and its policy implications.

Modeling and Simulation

Mathematically, the basic structure of a formal System Dynamics computer simulation model is a system of coupled, nonlinear, first-order differential (or integral) equations. Simulation of such systems is easily accomplished by partitioning simulated time into discrete intervals of length dt and stepping the system through time one dt at a time.

Feedback Thinking

Conceptually, the feedback concept is at the heart of the System Dynamics approach. Diagrams of loops of information feedback and circular causality are tools for conceptualizing the structure of a complex system and for communicating model-based insights.

Loop Dominance and Nonlinearity

The loop concept underlying feedback and circular causality by itself is not enough, however. The explanatory power and insightfulness of feedback understandings also rest on the notions of active structure and loop dominance.

The Endogenous Point of View

The concept of endogenous change is fundamental to the System Dynamics approach. It dictates aspects of model formulation: exogenous disturbances are seen at most as triggers of system behavior; the causes are contained within the structure of the system itself.

System Structure

These ideas are captured in Forrester's (1969) organizing framework for system structure:

Closed boundary

Feedback loops

Levels

Rates

Goal

Observed condition

Discrepancy

Desired action

Levels and Rates

Stocks (levels) and the flows (rates) that affect them are essential components of system structure. Stocks (accumulations, state variables) are the memory of a dynamic system and are the sources of its disequilibrium and dynamic behavior.

Behavior as a Consequence of Structure

The System Dynamics approach emphasizes a continuous view. The continuous view strives to look beyond events to see the dynamic patterns underlying them. Moreover, the continuous view focuses not on discrete decisions but on the policy structure underlying decisions. Events and decisions are seen as surface phenomena that ride on an underlying tide of system structure and behavior.”

B. Weatherford’s [5] taxonomy




Table Appendix 1: The elements and descriptors of Weatherford’s taxonomy

| Elements | Descriptors |
|-------------------------------------|--|
| A Resource | Discrete/Continuous |
| B Capacity | Fixed/Nonfixed |
| C Prices | Predetermined/Set optimally/Set jointly |
| D Willingness to Pay | Buildup/Drawdown |
| E Discount Price Classes | 1/2/3/.../l |
| F Reservation Demand | Deterministic/Mixed/Random-independent/Random-correlated |
| G Show-Up of Discount Reservation | Certain/Uncertain without cancellation/Uncertain with cancellation |
| H Show-Up of Full-Price Reservation | Certain/Uncertain without cancellation/Uncertain with cancellation |
| I Group Reservations | No/Yes |
| J Diversion | No/Yes |
| K Displacement | No/Yes |
| L Bumping Procedure | None/Full-price/Discount/FCFS/Auction |
| M Asset Control Mechanism | Distinct/Nested |
| N Decision Rule | Simple Static/Advanced static/Dynamic |

C. Industry comparison tables

Carroll and Grimes [75]

Table Appendix 2: Carroll and Grimes industry comparison table between the airline, hotel, and rental car industry

| Varying Complexities among Airlines, Hotels and Rental Cars |  |  |  |
|---|---|--|---|
| Inventory | Seat | Room | Car |
| Number of Unit Types | 1-3 | 1-10+ | 5-20+ |
| Total Units By Location | Fixed | Fixed | Variable |
| Mobility of Inventory | Small | None | Considerable |
| Rates Per Unit | Many (3-7+) | Few (2-3+) | Many (4-20+) |
| Duration of Use | Fixed | Variable | Variable |
| Corporate Discounts | No | Yes | Yes |
| Inventory Managed | Central | Central/Local | Central/Regional/Local |

Billings et al. [78]

Table Appendix 3: Comparison table of Billings et al. between cargo and passenger air

| Table 1: Attributes of capacity, demand and optimisation between cargo and passenger revenue management | | |
|--|---|--|
| | <i>Cargo</i> | <i>Passenger</i> |
| Capacity | Multi-dimensional — weight, volume, positions Variable (each flight has different capacity) | Single-dimensional — seats Cabin is either fixed or semi-fixed (ie movable-curtain cabin) |
| Demand | Irregular | Comparatively smooth based on seasons |
| Optimisation | Multi-dimensional Service-level-bound Short-term commitments — bookings Medium-term commitments — contracts, allocations, block space Multiple routings to meet service commitments Operational restrictions | Mostly one-dimensional Itinerary-bound Short-term commitments — bookings |

Nair and Bapna [81]

Table Appendix 4: Nair and Bapna’s industry comparison table between airline, hotels, and ISP

| Elements | Airlines | Hotels | ISP |
|---------------------------------|---------------|---------------|---------------------|
| Resource | Discrete | Discrete | Continuous |
| Capacity | Fixed | Fixed | Fixed |
| Cutoff time | Yes | Yes | No |
| Prices | Predetermined | Predetermined | Predetermined |
| Willingness to pay | Buildup | Buildup | Not applicable |
| Discount price classes | k | k | k |
| Arrival pattern | Stochastic | Stochastic | Stochastic |
| Departure pattern | No | No | Stochastic log-offs |
| Show-up of discount reservation | Certain | Certain | Stochastic |
| Show-up of full price | Certain | Certain | Stochastic |
| Group reservations | Yes | Yes | Not applicable |
| Overbooking | Yes | Yes | Not applicable |
| Diversion | No | No | Possible |
| Displacement | No | Downgrading | Not applicable |
| Bumping procedure | None | None | None |
| Asset control mechanism | Nested | Nested | Nested |
| Decision rule | Dynamic | Dynamic | Dynamic |

D. Primer on price elasticity of demand

The following primer on the price elasticity of demand is inspired by the lecture notes from Van Zandt [258]. The price elasticity E measures the sensitivity of demand Q to price P , i.e., if the price falls by some percentage points, how does the demand react:

$$E = \frac{\% \text{ change in } Q}{\% \text{ change in } P} \tag{D-1}$$

Since, generally, the quantity increases for declining prices, the elasticity E has a negative sign and has no units. Note that some textbooks and papers report the absolute numbers; however, the convention used in this dissertation is mathematically more precise. An increase in the elasticity (e.g., from -2 to -1.5) means that the elasticity becomes *less* elastic.

The elasticity varies along a demand curve and is also referred to as the local property of demand. That raises challenges of where to set the reference points of the initial and final values. The computation of the price elasticity is commonly done by regression. Algorithms fit parameters of a prior to empirical data points. The log-linear/power/constant elasticity demand function $Q = a \cdot P^E$ has proven to be a good approximation for many cases (see example in Figure Appendix 1 for three different elasticities). It has the advantages that the elasticity does *not* vary along the demand curve (can be easily seen by deriving the function by P). Furthermore, the function simplifies the regression into a linear form by applying the logarithm: $\log(Q) = \log(A) - E \cdot \log(P)$.

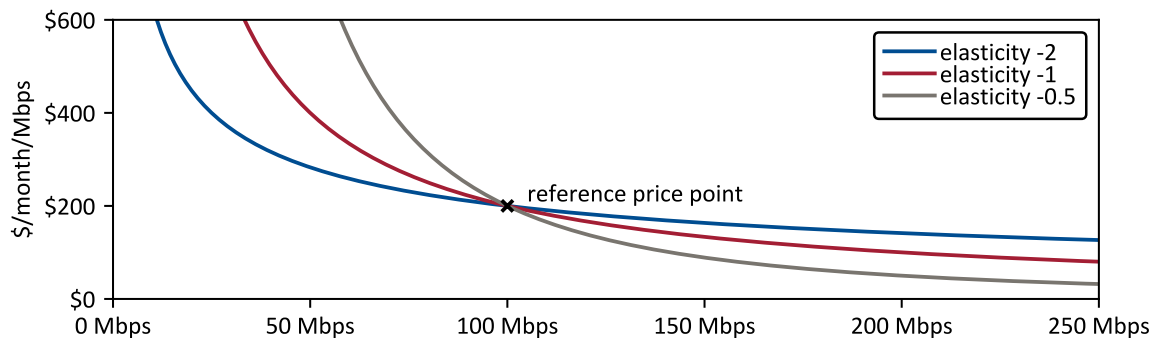


Figure Appendix 1: example of a log-linear demand elasticity

The research community and industry group commonly use the terminology outlined in Table Appendix 5. This division is central to the impact of the changing prices on the resulting revenues. Hence, we illustrate in Table Appendix 6 changes in the prices for the three different elasticity from Figure Appendix 1 and report the resulting revenues ($\Pi = Q \cdot P$). We highlight the reference point of $Q = 100$ and $P = 200$ in bold.

Table Appendix 5: Categorization of the elasticity into elastic, unit elastic, and inelastic

| | |
|----------|--------------|
| $E < -1$ | elastic |
| $E = -1$ | Unit elastic |
| $E > -1$ | inelastic |

Table Appendix 6: example of the impact of inelastic and elastic demand on the revenues

| $E = -2$ (elastic) | | | $E = -1$ (unit elastic) | | | $E = -0.5$ (inelastic) | | |
|-----------------------|------------|---------------|----------------------------|------------|---------------|---------------------------|------------|---------------|
| Q | P | Π | Q | P | Π | Q | P | Π |
| 50 | 283 | 14,142 | 50 | 400 | 20,000 | 50 | 800 | 40,000 |
| 60 | 258 | 15,492 | 60 | 333 | 20,000 | 60 | 556 | 33,333 |
| 70 | 239 | 16,733 | 70 | 286 | 20,000 | 70 | 408 | 28,571 |
| 80 | 224 | 17,889 | 80 | 250 | 20,000 | 80 | 313 | 25,000 |
| 90 | 211 | 18,974 | 90 | 222 | 20,000 | 90 | 247 | 22,222 |
| 100 | 200 | 20,000 | 100 | 200 | 20,000 | 100 | 200 | 20,000 |
| 110 | 191 | 20,976 | 110 | 182 | 20,000 | 110 | 165 | 18,182 |
| 120 | 183 | 21,909 | 120 | 167 | 20,000 | 120 | 139 | 16,667 |
| 130 | 175 | 22,804 | 130 | 154 | 20,000 | 130 | 118 | 15,385 |
| 140 | 169 | 23,664 | 140 | 143 | 20,000 | 140 | 102 | 14,286 |
| 150 | 163 | 24,495 | 150 | 133 | 20,000 | 150 | 89 | 13,333 |

The elasticity $E = -1$ is a unique value because it is the edge case where revenues remain *constant* independent of changing prices. In the elastic $E = -2$ case, the revenues increase with lower prices, while the opposite applies to inelastic demand. The implications for Revenue Management are that lower prices grow the overall market (in terms of revenues) when the demand is elastic but shrink it when demand is inelastic.

E. Frequency allocation for $t = 3$ hours

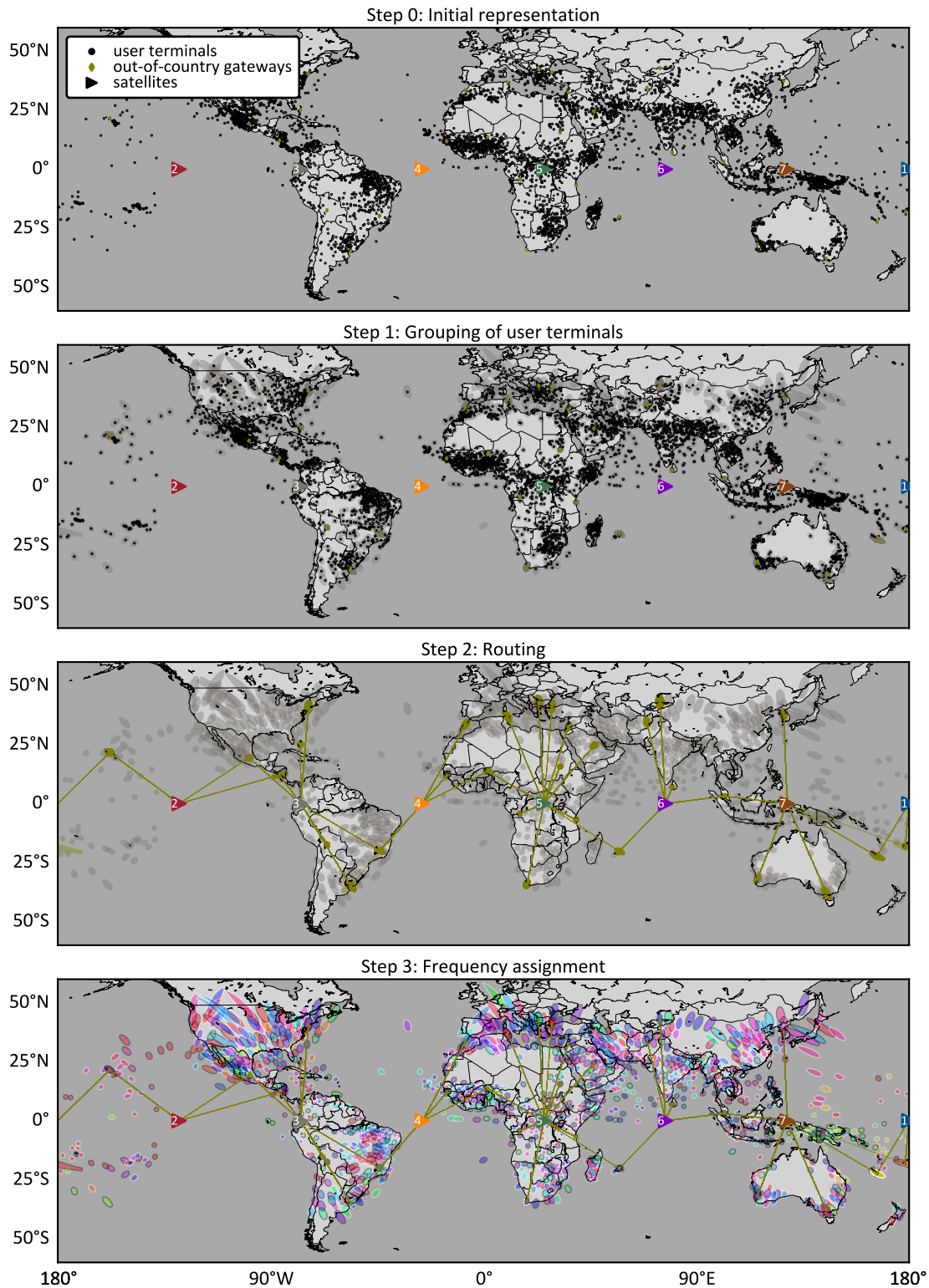


Figure Appendix 2: Resource allocation process solution for $t = 3$ hours for the constellation from Section 5.9

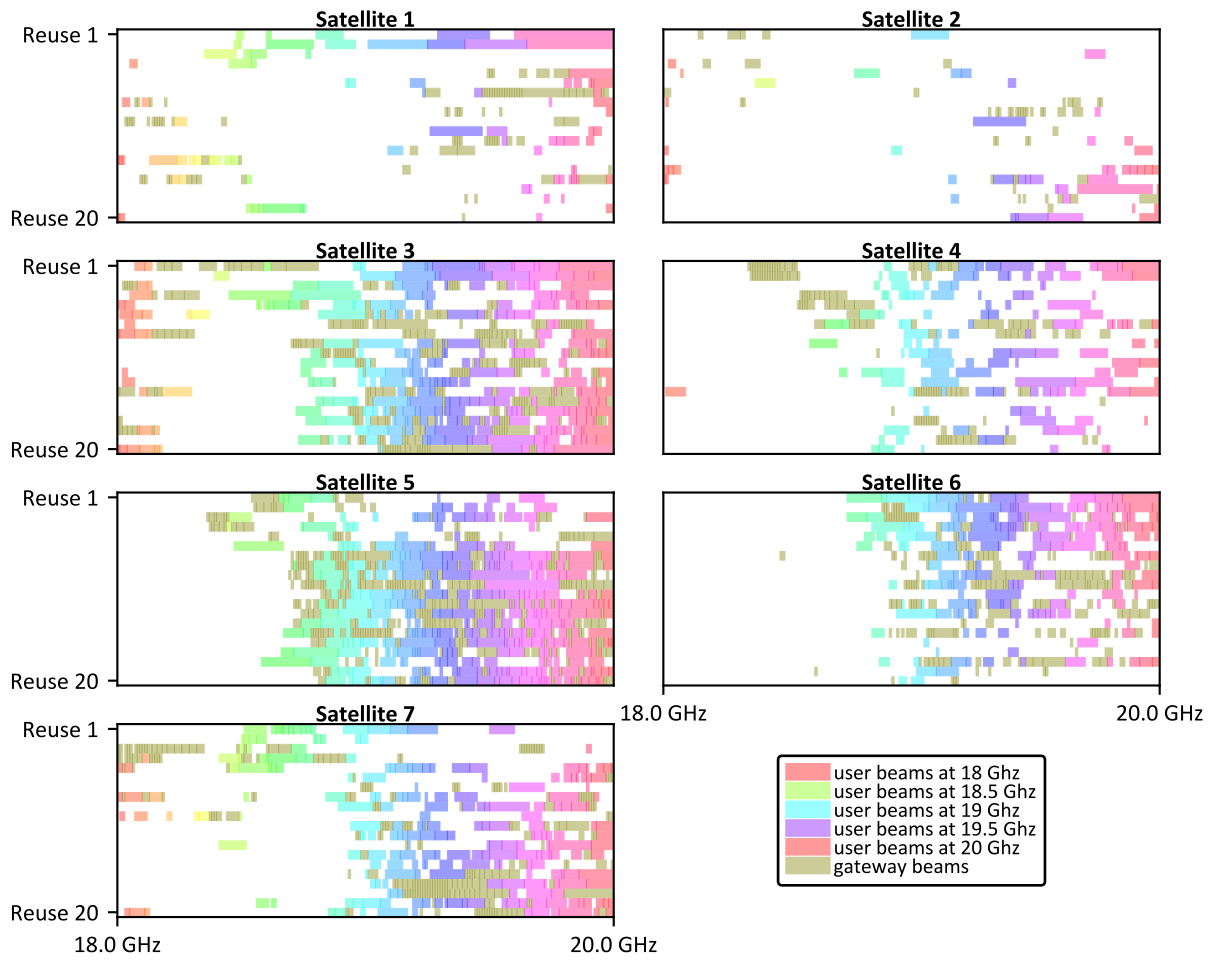




Figure Appendix 3: Frequency assignment for $t = 3$ hours solution of the constellation from Section 5.9

F. Service providers SLA menu

KVH

Ideal Applications:
 Anti-virus & software updates
 Vessel alerts & reporting
 Weather services data
 Automated document transmission
 Engine monitoring & reporting
 Web browsing
 Social media
 VPN
 Crew usage (Unallocated)
 Instant messaging (e.g., WhatsApp)
 Onboard server communications to shore


**LAN:
Unlimited Use**


**LAN:
High-speed**

Ideal Applications:
 Videoconferencing
 Web browsing
 Video chat (e.g., FaceTime, Skype)
 Dedicated media stream
 Telemedicine
 Social media
 Crew usage (Allocated)
 On-demand document transmission
 Security cameras

Select one plan for each

Step 1. Select your Unlimited Use Rate Plan:

| DATA SERVICE (Ku-band) | | | |
|------------------------|---|-------------|----------|
| Monthly Data Plans | Monthly Rates (USD) | Data Speeds | |
| PLAN | \$/MONTH | DOWNLOAD | UPLOAD |
| V7C-UL1 | Rates from \$599! Contact KVH Sales for custom pricing | 128 Kbps | 64 Kbps |
| V7C-UL2 | | 256 Kbps | 64 Kbps |
| V7C-UL3 | | 512 Kbps | 128 Kbps |
| V7C-UL4 | | 1 Mbps | 256 Kbps |
| V7C-UL5 | | 2 Mbps | 512 Kbps |
| V7C-UL6 | | 4 Mbps | 1 Mbps |
| V7C-UL7 | | 8 Mbps | 2 Mbps |

- Speed-based plans with a fixed monthly price
- Restrictions may be imposed on access to data-intensive applications and protocols

Step 2. Select your High-speed Rate Plan:

| DATA SERVICE (Ku-band) | | | | |
|------------------------|-------------------------------|---------------------|---|--------------------|
| Monthly Data Plans | Rates (USD) | Data Plan Allotment | Overage Rates (USD) | Data Speeds |
| PLAN | \$/MONTH | GB/MONTH | \$/MB | DOWNLOAD/UPLOAD |
| V7C-HS-2GB | FREE! | 2 | Rates as low as \$0.10 Contact KVH Sales for details | 10 Mbps/ 3 Mbps |
| V7C-HS-5GB | | 5 | | |
| V7C-HS-10GB | Upcharge as low as \$300! | 10 | | |
| V7C-HS-20GB | | 20 | | |
| V7C-HS-40GB | Contact KVH Sales for details | 40 | | |
| V7C-HS-80GB | | 80 | | |
| V7C-HS-150GB | | 150 | | |

- Large data allowances, highest speeds, and most affordable rates
- Unrestricted access to all applications and protocols
- Prioritization on the network with "Business Class Service"
- 100% transparency with easy-to-use tools for alerts, system configuration, data allocation, and data usage control and visibility

*Excludes BitTorrent and all such protocols

Comes complete with 2 phone lines available for Enhanced Voice Service, with calls to land or mobile for as low as \$0.05/min.

Figure Appendix 4: KVH Dual SLA menu [207]

Speedcast

Maritime and Offshore Ku-band VSAT satellite plans



Quota

Speedcast's Quota service is a metered VSAT service sold in a gigabyte (GB) packages from a half GB up to a 30 GB plan. This ku-band VSAT service is delivered with a guaranteed minimum CIR to ensure onboard communications & crew services while staying within budgetary guidelines. If the vessel exceeds the GB plan for the month, the ship can purchase additional data or suspend onboard usage for the remainder of the month.



Burst

Speedcast's Burst service provides a guaranteed minimum rate (CIR) with a higher maximum information rate (MIR). The ratio between CIR and MIR is large: CIR gives a vessel the guaranteed minimum bandwidth needed for specific business applications or data transfers, and MIR permits the service to reach a higher service level when demanded. And all the flexibility provided at a budget conscious price. Burst offers unlimited usage with no hidden fees or restricted service levels.



Guaranteed

Speedcast's Guaranteed service is a CIR only service for a minimum required bandwidth. Specific applications and data transfers require a minimum guaranteed bandwidth to ensure the application is running or the data transfer occurs in all conditions. Guaranteed provides unlimited usage with no hidden fees allowing a vessel to send and receive as much data as needed in a month with guaranteed minimum service. When customer operations onboard require a bandwidth assurance, Guaranteed will meet that need.



Professional

Speedcast's Professional service can be tailored to meet all onboard requirements. This approach to Ku-band services provides a vessel with the opportunity to choose from a wide variety of service levels all with a CIR and MIR. A vessel will have the guaranteed minimum CIR needed for data transfers or applications and the ability to reach the MIR levels when sending or receiving data. Professional is an unlimited service with no hidden fees or throttling of service levels, thus allowing the required minimums for data services and extra bandwidth when communications or crew need it.

Figure Appendix 5: Speedcast VSAT satellite plan [208]

Marlink



SEALINK ALLOWANCES



Sealink Allowances comprise a choice of competitively priced data allowance plans for 2-4 voice lines, always-on Internet, email and IP-based services ideal for business and crew communications in one package. With data allowance plans from 1 Gigabyte (GB) up to 80 GB per month, and data speeds of up to 6 Mbps, this low-cost Internet service is ideal for basic business requirements and for ship operators wishing to boost crew morale and optimise crew retention. As your business grows, our packages enable your VSAT to easily grow with you, since packages can be simply upgraded or topped up via our online portal.



SEALINK BUSINESS



Guaranteed data at a budget friendly price

Sealink Business offers 2-8 voice lines and your 13 CIR levels from 32 Kbps to 1 Mbps, in addition to a burstable Maximum Information Rate (MIR) up to 6 Mbps.

In accordance with the defined CIR, a dedicated amount of bandwidth is supplied for unlimited data usage in accordance with our Fair Access Policy, so business critical applications are always available.

Available as a regional or global service, Sealink Business enables a constant, quality service with burstable capabilities, at a budget friendly price. This service is ideal for ship operators seeking cost-effective, always-on communications for business and crew.

Figure Appendix 6: Marlink VSAT sealink [209]

G. List of used MODCODs

Table Appendix 7: List of selected MODCODs from the DVB-S2 [100] and DVB-S2X [101] standards. All MODCODs are dominant concerning the spectral efficiency and E_s/N . Further granularity can be achieved by including additional DVB-S2X MODCODs.

| Name | Spectral efficiency Γ | Ideal E_s/N |
|----------------|------------------------------|---------------|
| QPSK1/4 | 0.49 | -2.35 |
| QPSK1/3 | 0.66 | -1.24 |
| QPSK1/5 | 0.79 | -0.30 |
| QPSK1/2 | 0.99 | 1.00 |
| QPSK3/5 | 1.19 | 2.23 |
| QPSK2/3 | 1.32 | 3.10 |
| QPSK3/4 | 1.49 | 4.03 |
| QPSK4/5 | 1.59 | 4.68 |
| QPSK5/6 | 1.65 | 5.18 |
| 8PSK3/5 | 1.78 | 5.50 |
| QPSK9/10 | 1.79 | 6.42 |
| 8PSK2/3 | 1.98 | 6.62 |
| 16PSK8/15 | 2.09 | 6.93 |
| 16APSK-L8/15 | 2.10 | 6.55 |
| 8PSK3/4 | 2.23 | 7.91 |
| 16APSK26/45 | 2.28 | 7.51 |
| 16APSK3/5 | 2.37 | 7.80 |
| 16APSK28/45 | 2.46 | 8.10 |
| 16APSK23/36 | 2.52 | 8.38 |
| 16APSK2/3 | 2.64 | 8.97 |
| 16APSK25/36 | 2.75 | 9.27 |
| 16APSK13/18 | 2.86 | 9.71 |
| 16APSK3/4 | 2.97 | 10.21 |
| 16APSK7/9 | 3.08 | 10.65 |
| 16APSK4/5 | 3.17 | 11.03 |
| 16APSK5/6 | 3.30 | 11.61 |
| 16APSK77/90 | 3.39 | 11.99 |
| 32APSK32/45 | 3.51 | 11.75 |
| 32APSK11/15 | 3.62 | 12.17 |
| 32APSK3/4 | 3.70 | 12.73 |
| 32APSK7/9 | 3.84 | 13.05 |
| 32APSK4/5 | 3.95 | 13.64 |
| 64QAM | 4.50 | 14.00 |
| 64APSK7/9 | 4.60 | 15.47 |
| 64APSK4/5 | 4.74 | 15.87 |
| 64APSK5/6 | 4.93 | 16.55 |
| 256APSK-L29/45 | 5.07 | 16.98 |
| 128APSK3/4 | 5.16 | 17.73 |
| 256APSK-L31/45 | 5.42 | 18.10 |
| 256APSK32/45 | 5.59 | 18.59 |
| 256APSK3/4 | 5.90 | 19.57 |

H. Result of user grouping and frequency assignment for SpaceX's Starlink

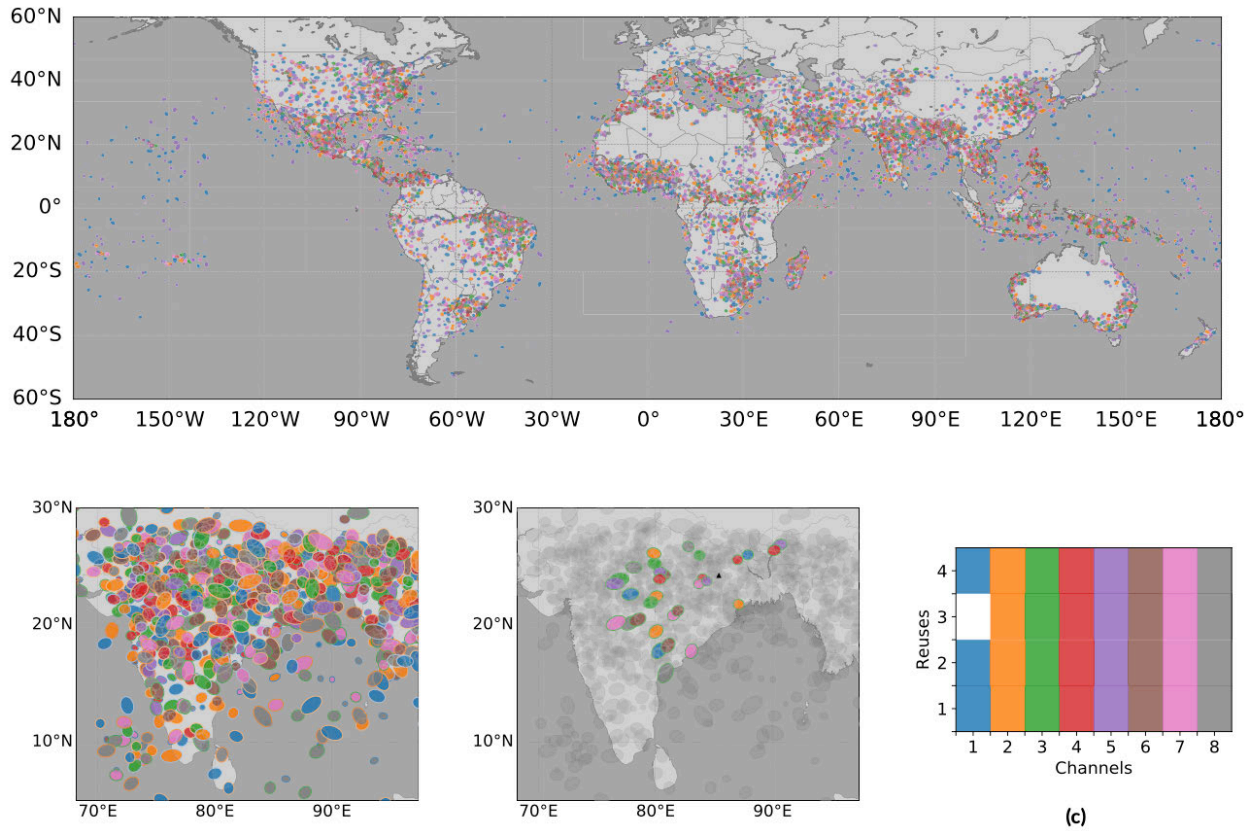


Figure Appendix 7: Result of the clique algorithm for the user grouping and the first-fit frequency assignment for SpaceX's Starlink. Obtained from Pachler [188]

List of Acronyms

| | | | |
|-------|--|--------|---|
| ACM | Adaptive Coding and Modulation | MEO | Medium Earth Orbit |
| AI | Artificial Intelligence | MIR | Maximum Information Rate |
| ARC | Adaptive Resource Control | ML | Machine Learning |
| ATM | Asynchronous Transfer Mode | MNO | Mobile Network Operator |
| AWS | Amazon Web Services | MODCOD | Modulation and Coding Scheme |
| BSS | Broadcasting Satellite Service | NAM | US and Canada |
| CACI | Adjacent Channel Interference | NGSO | Non-Geostationary Orbit |
| CAPEX | Capital expenditures | NSGA | Non-dominated Sorting Genetic Algorithm |
| CASI | Adjacent Satellite Interference | NSR | Northern Sky Research |
| C3IM | Intermodulation Products Interference | OBO | Output Back-Off |
| CEEU | Central and Eastern Europe | OPEX | Operating Expenditures |
| CIR | Committed Information Rate | PIR | Peak Information Rate |
| CRM | Customer Relationship Management | PPP | Purchasing Power Parity |
| CXPI | Cross-Polarization Interference | PSO | Particle Swarm Optimization |
| DPDA | Daily Planning and Distribution Aid | QoS | Quality of Service |
| EA | Eastern Africa | RBF | Radial Basis Function |
| EC2 | Elastic Computing Cloud | RF | Radio Frequency |
| EIRP | Effective Isotropically Radiated Power | RM | Revenue Management |
| FCC | Federal Communications Commission | RTE | Real-Time Engine |
| FDMA | Frequency Division Multiple Access | RTN | Return |
| FOV | Field of View | SA | South Asia |
| FSS | Fixed Satellite Services | SD | Systems Dynamic |
| FWD | Forward | SEA | South East Asia |
| GA | Genetic Algorithm | SES | Société Européenne des Satellites |
| GEO | Geostationary Orbit | SGM | Satisfaction-Gap Measure |
| GW | Gateway | SINR | Signal-to-Interference-plus-Noise Ratio |
| HT | High-Throughput | SLA | Service Level Agreement |
| HTS | High-Throughput Satellites | TDRS | Tracking and Data Relay Satellite |
| ISP | Internet Service Provider | UAV | Unmanned Aerial Vehicle |
| ITU | International Telecommunication Union | US | United States |
| LAM | Latin America | USA | United States of America |
| LEO | Low Earth Orbit | VSAT | Very Small Aperture Terminal |
| LSTM | Long Short-Term Memory | WEU | Western European Union |
| LTC | Long-Term Contract | WTP | Willingness-To-Pay |
| MAE | Mean Absolute Error | YM | Yield Management |
| MEA | the Middle East and Africa | YMS | Yield Management System |

List of Figures

| | |
|--|----|
| Figure 1-1: Comparison of the two major segments in the satellite communication industry..... | 14 |
| Figure 1-2: Development of satellite technology over the last five decades | 16 |
| Figure 1-3: Conceptual drawing of the challenging environment in which flexible HTS are operating. The four different segments are backhauling, aviation, maritime, and energy. The brown coins represent the cost for the operator, the green bills the willingness-to-pay (WTP), and the blue line illustrates the diurnal usage of data rate. | 23 |
| Figure 1-4: Preview of the proposed framework from Section 3.5 as a guide through the remainder of this dissertation | 26 |
| Figure 2-1: Overview of the different players in the market, divided into operators, service providers, and customer/end-users. The connections illustrate publicly available information about business relationships..... | 29 |
| Figure 2-2: Workflow of market dynamic model creation, translation, and analysis | 31 |
| Figure 2-3: system dynamic diagram of the market model | 32 |
| Figure 2-4: Supply, demand, and price trend analysis obtained from [57], which is based Euroconsult and NSR. | 33 |
| Figure 2-5: Validation of the market dynamic model against the data from Table 2-2, Prasad, NSR [60], and a price trend regression of SES data. Price $p(t)$ is in \$/Mbps/Month. | 38 |
| Figure 2-6: Contributions of the three uncertain parameters success of NGSO constellations $sNGSO$, demand elasticity b , and additional NGSO demand $mNGSO, demand$ on the price $p(tfinal)$, cumulative GEO revenues $\Pi_{GEO}, cum(tfinal)$, cumulative NGSO revenues $\Pi_{NGSO}, cum(tfinal)$, and cumulative service providers revenues $\Pi_{SP}, cum(tfinal)$. We use the total-order indices of the Sobol analysis. ... | 40 |
| Figure 2-7: Dispersion of $p(t)$, $D(t)$, $\Pi_{GEO}(t)$, $\Pi_{NGSO}(t)$, $\Pi_{SP}(t)$ given the uncertainties in $sNGSO$, b , and $mNGSO, demand$ | 41 |
| Figure 2-8: Impact of delay reduction on the utilization of service provider and operator..... | 44 |
| Figure 3-1: Preview of the proposed framework and document guidance, copied from Figure 1-4..... | 51 |
| Figure 3-2: Normalized 24h data rate variation; one-month of sampled demand data fitted by a Gaussian process using a radial basis and white noise kernel..... | 57 |
| Figure 3-3: Talluri’s [6] representation of the RM process flow | 58 |
| Figure 3-4: components of typical third-generation airline RM system obtained from Belobaba [64] | 59 |
| Figure 3-5: Vinod’s [73] functional decomposition of the hotel RM system..... | 61 |
| Figure 3-6: Carroll and Grimes [75] implemented car-rental RM system for Hertz; CRS=Computerized reservation system; DPDA= Daily Planning and Distribution Aid; YMS=Hertz’s Yield Management System. | 62 |
| Figure 3-7: The processes and components of the RM system implemented by National Car Rental; obtained from Geraghty and Johnson [74]; RES=Reservation System..... | 63 |
| Figure 3-8: Kasilingam’s [76] cargo RM model (redrawn for better readability) | 64 |
| Figure 3-9: KLM’s cargo RM system (obtained from Slager and Kapteijns [79]) | 65 |
| Figure 3-10: Outline of the RM process with its tasks: LTC=Long-Term Contract; obtained from [80]. | 66 |
| Figure 3-11: Flow diagram of the RM methodology for computing resources protected by patent US 8,788,310 B2 [85] | 68 |

Figure 3-12: Integration of RM, dynamic pricing, CRM, and profitability as outlined by Jallat and Ancarani [89] 69

Figure 3-13: Overview of our proposed satcom RM framework mapped to the four challenges from Table 3-4: (1) unit of capacity is not unit of demand, (2) resource allocation is an optimization problem itself, (3) uncertain available capacity based on resource usage, and (4) existing SLAs do not fully leverage the new satellites’ flexibility..... 77

Figure 3-14: Overview map of the location of the five customers and the GEO satellite..... 83

Figure 3-15: Demand over a 24 h period; confidence intervals are function of the mean (see Eq. (3-6)). 87

Figure 3-16: Plot of the used capacity modelled by the stochastic process $Y_{tt} \in T$ after sampling $X_{i,t}$ 50,000 times for each i and t . The blue line illustrated the maximum capacity which is equal to an allocation based on CIRs. 87

Figure 3-17: Lin-log demand functions for the five customers. The triangle is the price point of the allotments for which the elasticity ϵ is defined. 89

Figure 3-18: On the upper left, the additional revenues $\Pi_{spot,i}$ per month are plotted as a function of the spot instance power $C_{spot,i}$ and data rate $R_{spot,i}$ on the upper right. The lower left shows the same demand elasticity as Figure 3-17. The circles indicate the optimized solutions for $C_{spot,i}^*$, $R_{spot,i}^*$, and $p_{spot,i}^*$ 91

Figure 4-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4 97

Figure 4-2: Overview of the satcom simulator. The numbers indicate the cardinality of the hierarchy, the red arrows are connections that are a function of time, and the objects with an asterisk * are majority work of Portillo [97]. 98

Figure 5-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4 113

Figure 5-2: The resource allocation process showing two satellite 1 and 2 in the same plan, four user terminals A, B, C, and D, as well as two gateways I and II. Step 1: grouping user terminals; Step 2: routing; Step 3: frequency assignment..... 131

Figure 5-3: benchmarking results of the Pachler’s heuristic approximation against a brute-force and grid approach; NR = number of runs of the algorithm. Obtained from [188]. 136

Figure 5-4: Example of a suboptimal beam placement. Black dots represent user terminals; the filled ellipses are the projection of circular beams onto the Earth’s surface; the edge and fill color represent the assigned center frequency and polarization. 137

Figure 5-5: comparison of the resulting allocation for the closest-first and balanced allocation heuristic. The horizontal shows the 181 gateways colored by in-country and out-of-country. The black dashed line represents the CIR for Γ_{allow} from the balanced allocation algorithm. 144

Figure 5-6: Running the routing algorithm without applying the in- and out-of-country constrains resulting in a better load balancing. The black dashed line is for comparison from Figure 5-5..... 145

Figure 5-7: Solution comparing the three implemented heuristics for the frequency assignment by Pachler. Obtained from [188]. 150

Figure 5-8: Statistics of the 5,000 user terminals. The left plot shows a histogram of the CIR and the right a distribution of the CIR amongst the seven segments..... 155

Figure 5-9: Sample of the diurnal demand pattern for the seven segments. Discretized into 15 minutes intervals..... 155

Figure 5-10: Final result plot of steps 1-3 for a constellation similar to O3b mPower with 5,000 user terminals at t_0 157

Figure 5-11: Frequency allocation for the first time step. Each block represents one beam (gateways are brown). 158

Figure 5-12: Aggregated power on satellite level over the 24h simulation period for the user beams. .158

Figure 5-13: Aggregated data rate and requested demand of satellite 1 over the 24-hours simulation period. The satellite meets the demand at all times and only overprovides a negligible data rate. 161

Figure 5-14: Selected user downlink MODCODs for a customer on the Canary Islands at around 28 deg latitude and -16 deg longitude. The customer has a 1.2m dish with a G/T of 21 dB and a 100 Mbps CIR SLA. 162

Figure 6-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4 165

Figure 6-2: Density forecasts for a residential consumer. Adopted from [196], P.18. Black dots are observations, black line is forecast, shading is uncertainty distribution in forecast. 168

Figure 6-3: Gaussian process example of 1,000 samples of one-month traffic data fitted by a Gaussian process using a radial basis and white noise kernel. Copied from Figure 3-2. 170

Figure 6-4: Example of the impact of the contracted availability A on the used and the available capacity as a fraction of the maximum capacity. 173

Figure 6-5: Convergence analysis of the absolute relative error as a function of the number of samples $N_{samples}$ 173

Figure 7-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4 175

Figure 7-2: Overview of the decomposing of needs into served, observed but unserved, and latent needs. 176

Figure 7-3: Conceptual view of the used, available whole-day and part-day capacity. Numbers are based on Figure 3-16 from the example in Section 3.6. 180

Figure 7-4: Top plot is equally split cosine traffic but not shifted; bottom plot shows the same traffic, but the not-real-time traffic is shifted by 12 hours resulting in perfect cancellation. 187

Figure 7-5: Classification of the seven segments based on location and traffic volume, inspired by [229, 230] 196

Figure 7-6: Segmentation of the market by using the three bases: type of business, location type, usage per terminal. Bold segments are these not identified in Table 7-4. 197

Figure 8-1: Proposed satcom RM framework and document guidance, copied from Figure 1-4 203

Figure 8-2: Overview of the structure of the analysis for Sections 8.2 - 8.5 204

Figure 8-3: Statistical description of the 80 users 207

Figure 8-4: Traffic estimation for one example user from each segment 207

Figure 8-5: Linear regression of price decline and adjustment of the data points for this trend. The prices are normalized to the function value of the linear regression for the reference year 2019 208

Figure 8-6: Power regression of the price elasticity for each group. Normalized prices are based on Figure 8-5 209

Figure 8-7: recursively estimated price elasticity for telecommunication obtained from [232, p. 255].. 211

Figure 8-8: Solution after the first three steps of the resource allocation process for the MEO O3b mPower like constellation. The usage of the frequency spectrum is on average 25%. 214

Figure 8-9: Power allocation for the three different scenarios. 1-minute granularity over 24 hours separated into four 6-hour orbits. The plot shows the power allocation for satellite #3. 215

Figure 8-10: Comparison between the transmitted power on user level between the fixed by design

scenario and stationary allocation for the CIR. 216

Figure 8-11: Comparison of the data rate allocated and requested between the stationary allocation for CIR and for actual usage. The curves for actual usage are plotted every 10 minutes to avoid cluttering. 217

Figure 8-12: General approach for integrated optimization of pricing and resource allocation. 219

Figure 8-13: Resulting fit function in blue of the sample points in black for each user 223

Figure 8-14: Example of a -2, -1, and -0.5 price elasticity for a customer with a 100 Mbps and \$200/month/Mbps 227

Figure 8-15: Monthly revenues, average price, total contracted CIR capacity for the three elasticity cases compared between the baseline and the two pricing algorithms. 228

Figure 8-16: Comparison of the prices between the two algorithms and the baseline for the elastic case with -2. 229

Figure 8-17: Behavior of the price elasticity function as a function of the reference point and elasticity parameter. The plots are shown for a price 20% that of the Classical SLA for the unit elasticity..... 233

Figure 8-18: Results for the combination $krev, fract = 0.2$ and $kadj = 0.5$ 234

Figure 8-19: Example of the affordability elasticity for a segment with two user terminals at \$200/month/Mbps 239

Figure 8-20: density penalty $kdens, pen$ as a function of the number of terminals for Germany and Mali in Africa. The threshold is 3 for Germany and 12 for Mali with $Athres = 300,000 km^2$ and a slope of $sdens, pen = 1$ 241

Figure 8-21: linear relationship between the relative utilization and the satellite utilization penalty.... 242

Figure 8-22: hyperparameter tuning of the heuristic gradient algorithm for the elasticity case -2. Numbers are relative to the baseline revenues. Bold values are the maximum in each $ksatutil, pen, max$ case and the red font indicates the global maximum. 243

Figure 8-23: hyperparameter tuning of the heuristic gradient algorithm for the case with elasticities varying between -2 and -1. Numbers are relative to the baseline revenues. Bold values are the maximum in each $ksatutil, pen, max$ case and the red font indicates the global maximum. 244

Figure 8-24: results for the monthly revenue increase, average price, total contracted capacity and number of terminals for the three algorithms for three elasticity cases. 245

Figure 8-25: schematic drawing of the modeling of the discount factor $kdiscount$ on segment level (here 20%) 249

Figure 8-26: results for highest yield and the heuristic gradient optimization for different discount threshold $kdiscount$ cases..... 251

Figure 8-27: number of terminals for each segment for the baseline, the highest yield solution, and the heuristic gradient optimization. If number of terminals is different from the baseline, it is always a doubling. 252

Figure 8-28: Time dependency of the monthly revenues and number of terminals for the two algorithms. Time is normalized to the average contract duration..... 253

Figure 8-29: Results for a mismatch between the believed and the actual elasticity for two elastic cases and the two algorithms. The scenario is selling more through existing SLAs from Section 8.4.1. 257

Figure 8-30: schematic of the pricing policy proposed by Cheung et al. Figure duplicated from the authors original publication [246, p.18]. 258

Figure 8-31: Comparison of the results across the four analyses. Besides the (heuristic) gradient approach, the second-best heuristic is plotted. The percentage numbers in parentheses indicate the error bar numbers, which are the minimum and maximum numbers. The bars show the means..... 261

Figure Appendix 1: example of a log-linear demand elasticity 285

Figure Appendix 2: Resource allocation process solution for t = 3 hours for the constellation from Section 5.9 287

Figure Appendix 3: Frequency assignment for t = 3 hours solution of the constellation from Section 5.9 288

Figure Appendix 4: KVH Dual SLA menu [207] 289

Figure Appendix 5: Speedcast VSAT satellite plan [208] 289

Figure Appendix 6: Marlink VSAT sealink [209]..... 290

Figure Appendix 7: Result of the clique algorithm for the user grouping and the first-fit frequency assignment for SpaceX’s Starlink. Obtained from Pachler [188]..... 292

List of Tables

Table 1-1: Overview of the key players in the broadband communication market (FSS and MSS). A green-dashed rectangle indicates companies who filed an application with the FCC by April 2019 but have no operational assets. We base the categorization on the companies’ current operational satellite with the highest throughput. 20

Table 1-2: Summary of the state-of-the-arts, the gaps, and the expected contributions made in this dissertation. Chapter 4 (the satcom simulator) is not listed in the table as the Chapter does not contain a strong scientific novelty. 25

Table 2-1: Overview of the current purchasing behavior 30

Table 2-2: HTS capacity, supply, and price forecast for years 2017-2025 extracted from [57], which consolidated 2017 data from NSR and Euroconsult. Capacity and demand are in Gbps, price in \$/Mbps/Month. 34

Table 2-3: Overview of the considered uncertainty ranges for the success of NGSO constellations $sNGSO$, the demand elasticity b , and the added demand through NGSO constellations $mNGSO, demand$ 39

Table 2-4: Overview of the baseline values of the parameters and the values for a first and second reduction..... 43

Table 2-5: Mean value of the 6 cases using 800 samples of the three uncertainties. 45

Table 2-6: Standard deviation of the 6 cases using 800 samples of the three uncertainties..... 46

Table 2-7: Comparison table of baseline versus a more vertically integrated market. The integration is modelled by reducing the delays $\Delta tSP, sell, \Delta tSP, incr$, and $\Delta tSP, decr$ to 1 day..... 47

Table 3-1: Example for price-based RM 53

Table 3-2: Summary of the RM reviewed using Weatherford’s and Bodily’s [5] taxonomy and their descriptors. 71

Table 3-3: Comparison of the industries’ characteristics along our eight elements..... 74

Table 3-4: Summary of the four challenges in satcom and mapping them to traditional and new satcom. A checkmark ✓ denotes that the challenge is relevant. 76

Table 3-5: Assumptions about the customers input parameters..... 84

Table 3-6: different capacity costs of the five customers 85

Table 3-7: Results after optimization 92

Table 3-8: the two application of our satcom RM framework 94

Table 4-1: Overview of the 10 inputs for validation of the link budget relationships..... 109

Table 4-2: comparison table between the results of our simulator and the reference provided by SES. The percentages are the relative difference which are averaged over the 10 links in the last row. 110

Table 5-1: Summary of the conducted literature review based on dynamic parameters considered and optimization techniques used..... 129

Table 5-2: Comparison of the mean and standard deviation of the slant range and the free space loss for both routing heuristics. Data is generated by a 6-hour simulation with 5-minutes discretization. The percentages are the relative difference between closest-first and balanced..... 146

Table 5-3: Overview of the constellation’s parameters 154

Table 5-4: Average power consumption of the seven satellite. The percentage numbers are the relative deviation from the mean over all satellites. 160

Table 7-1: Example contracts of the three existing SLA types 179

Table 7-2: Overview of the novel SLAs and their mapping to the desired characteristics 185

Table 7-3: Segmentation bases for HTS broadband satcom 193

Table 7-4: Characteristics of each commonly used segment’s bases from the operator perspective..... 195

Table 7-5: Our assessment of the fit of the six SLAs in the ten segments (bb = broadband)..... 199

Table 8-1: Terminal EIRP, G/T, and G assumptions based on the terminal diameter 206

Table 8-2: regression parameters for $R = a \cdot pb$ 209

Table 8-3: Overview of the locations and sizes of the seven gateways considered..... 212

Table 8-4: Parameters of the MEO constellation, which is O3b mPower like. Compared to Table 5-3, the reuses are reduced from 20 to 4 to account for the lower traffic (198 Gbps vs. 22.7 Gbps forward CIR) 212

Table 8-5: Summary of the mean orbit power consumption of the three different power allocation scenarios and delta capacities. The percentages are relative to the fixed by design allocation..... 215

Table 8-6: schematic example of the ordered list of price p_i and the resulting demand R_i for a single user i (elasticity is -2 at $p_i = 200$ and $R_i = 100$). Used capacity computed through the Shannon limit and available capacity with a maximum capacity C_{max} of 6 Watts. 220

Table 8-7: Pair-wise correlation between selected attributes. Red colored fields are positive correlation, blue ones are negative correlation. Values are computed for the elastic -2 case comparing the gradient with the baseline. 230

Table 8-8: Test plan for testing of the sensitivity to the reference price point of the additional product. Each combination is simulated for four elasticity cases: -2, random between -2 and -1, -1, and -0.5. ... 232

Table 8-9: Revenue lifts across the nine combinations, four elasticities, and two algorithms. 235

Table 8-10: Pair-wise correlation between selected attributes. Red colored fields are positive, blue ones are negative correlation. Values are computed for the elastic -2 case comparing the heuristic gradient with highest yield. 246

Table 8-11: Comparison of the additional revenues potential and the revenue lift from sophisticated RM across the four analyses. 260

Table 8-12: competitive advantage of more sophisticated pricing algorithms in different scenarios..... 266

Table 9-1: Summary of contributions by Chapter. 278

Table Appendix 1: The elements and descriptors of Weatherford’s taxonomy 283

Table Appendix 2: Carrol and Grimes industry comparison table between the airline, hotel, and rental car industry 283

Table Appendix 3: Comparison table of Billings et al. between cargo and passenger air..... 284

Table Appendix 4: Nair and Bapna’s industry comparison table between airline, hotels, and ISP 284

Table Appendix 5: Categorization of the elasticity into elastic, unit elastic, and inelastic..... 286

Table Appendix 6: example of the impact of inelastic and elastic demand on the revenues 286

Table Appendix 7: List of selected MODCODs from the DVB-S2 [100] and DVB-S2X [101] standards. All MODCODs are dominant concerning the spectral efficiency and ES/N . Further granularity can be achieved by including additional DVB-S2X MODCODs. 291

List of Algorithms

Algorithm 3-1: incremental gradient-based optimization adapted from [96]. 91

Algorithm 5-1: heuristic approximation of beam placement, adapted from Pachler [188], protected under pending patent by SES..... 136

Algorithm 5-2: closest-first routing algorithm..... 140

Algorithm 5-3: Balancing routing algorithm..... 142

Algorithm 5-4: Pseudo code for the recoloring, first-fit heuristic, and random frequency assignment algorithm, adapted from Pachler [188]..... 149

Algorithm 5-5: power guess and search procedure for computing the data rate based on power..... 152

Algorithm 8-1: Pseudo code for the recursive binary search algorithm for pricing optimization..... 221

Algorithm 8-2: Pseudo code for the equally decreasing heuristic 222

Algorithm 8-3: Pseudo-code for the gradient optimizer based on marginal revenues. Prices are assigned to user or segment depending on analysis type. 224

List of Boxes

Box 7-1: Classical SLA, the most commonly used product for both MHz and Mbps satcom 177

Box 7-2: Data volume SLA with revenues as a function of the monthly volume and corresponding price 178

Box 7-3: Dual SLA combining the Classical and the Data volume SLA..... 178

Box 7-4: Spot instance SLA without a minimum or maximum contract duration 186

Box 7-5: Time-of-day pricing SLA 186

Box 7-6: Two Classes of Service SLA..... 188

Bibliography

- [1] B. C. Smith, J. F. Leimkuhler, and R. M. Darrow, "Yield management at American airlines," *Interfaces*, vol. 22, no. 1, pp. 8-31, 1992.
- [2] P. P. Belobaba, "Survey Paper—Airline yield management an overview of seat inventory control," *Transportation science*, vol. 21, no. 2, pp. 63-73, 1987.
- [3] S. E. Kimes, "The basics of yield management," *Cornell Hotel and Restaurant Administration Quarterly*, vol. 30, no. 3, pp. 14-19, 1989.
- [4] W.-C. Chiang, J. C. Chen, and X. Xu, "An overview of research on revenue management: current issues and future research," *International journal of revenue management*, vol. 1, no. 1, pp. 97-128, 2007.
- [5] L. R. Weatherford and S. E. Bodily, "A taxonomy and research overview of perishable-asset revenue management: Yield management, overbooking, and pricing," *Operations research*, vol. 40, no. 5, pp. 831-844, 1992.
- [6] K. T. Talluri and G. J. Van Ryzin, *The theory and practice of revenue management*. Springer Science & Business Media, 2006.
- [7] J. N. Pelton, S. Madry, and S. Camacho-Lara, *Handbook of satellite applications*. Springer, 2017.
- [8] I. d. Portillo, B. G. Cameron, and E. F. Crawley, "A Technical Comparison of Three Low Earth Orbit Satellite Constellation Systems to Provide Global Broadband," presented at the 69th International Astronautical Congress (IAC), Bremen, Germany, 1-5 October, 2018, 2018.
- [9] T. Canada, "Telesat Ka-band NGSO constellation FCC filing SAT-PDR-20161115-00108," 2018. [Online]. Available: http://licensing.fcc.gov/myibfs/forwardtopublictabaction.do?file_number=SATPDR2016111500108
- [10] W. S. Limited, "OneWeb Ka-band NGSO constellation FCC filing SAT-LOI-20160428-00041," 2018. [Online]. Available: http://licensing.fcc.gov/myibfs/forwardtopublictabaction.do?file_number=SATLOI2016042800041
- [11] S. E. Holdings, "LLC, SpaceX Ka-band NGSO constellation FCC filing SAT-LOA-20161115- 00118," 2018. [Online]. Available: http://licensing.fcc.gov/myibfs/forwardtopublictabaction.do?file_number=SATLOA2016111500118
- [12] Viasat. "Going Global - Viasat-2 and the Viasat-3 Platform Will Take Our Service Around the World." <https://www.viasat.com/news/going-global> (accessed October, 2018).
- [13] SES. "Exponentially More Opportunities With O3b mPOWER." <https://www.ses.com/networks/o3b-mpower> (accessed October, 2018).

- [14] SES. "SES Selects Arianespace for Launch of SES-17." <https://www.ses.com/press-release/ses-selects-arianespace-launch-ses-17> (accessed October, 2018).
- [15] R. Mehrotra, "Regulation of global broadband satellite communications," *Broadband Series*, ITU, 2012.
- [16] S. Finkelstein and S. H. Sanford, "Learning from corporate mistakes: The rise and fall of Iridium," *Organizational Dynamics*, vol. 29, no. 2, pp. 138-148, 2000.
- [17] J. Lim, R. Klein, and J. Thatcher, "Good technology, bad management: A case study of the satellite phone industry," *Journal of Information Technology Management*, vol. 16, no. 2, pp. 48-55, 2005.
- [18] E. W. Ashford, "Non-Geo systems—where have all the satellites gone?," *Acta Astronautica*, vol. 55, no. 3-9, pp. 649-657, 2004.
- [19] G. Comparetto and N. Hulkower, "Global mobile satellite communications-A review of three contenders," in *15th International Communications Satellite Systems Conference and Exhibit*, 1994, p. 1138.
- [20] Gunter. "Intelsat-1." https://space.skyrocket.de/doc_sdat/intelsat-1.htm (accessed March, 2019).
- [21] Intelsat. "Brief History of Satellite Communications." <https://www.telesat.com/about-us/why-satellite/brief-history> (accessed March, 2019).
- [22] D. J. Whalen. "Communications Satellites: Making the Global Village Possible." <https://history.nasa.gov/satcomhistory.html> (accessed March, 2019).
- [23] Statista. "Global Telecommunications Services Market Value from 2012 to 2019, by Region (in Billion Euros)." <https://www.statista.com/statistics/268636/telecommunications-services-revenue-since-2005-by-region/> (accessed March, 2019).
- [24] B. Space, "Technology," "2017 State of the Satellite Industry Report," *Satellite Industry Association*, 2017.
- [25] Gunter. "iPStar 1 (Thaicom 4, MEASAT 5, Synertone 1)." https://space.skyrocket.de/doc_sdat/ipstar-1.htm (accessed March, 2019).
- [26] H. Helvajian and S. Janson, *Small satellites: past, present, and future*. American Institute of Aeronautics and Astronautics, Inc., 2009.
- [27] C. Balty, J.-D. Gayraud, and P. Agnieray, "Communication satellites to enter a new age of flexibility," *Acta Astronautica*, vol. 65, no. 1-2, pp. 75-81, 2009.
- [28] SES. "O3B MPOWER." <https://www.ses.com/networks/networks-and-platforms/o3b-mpower> (accessed March, 2019).
- [29] Gunter. "O3b 21, ..., 27 (O3b mPower)." https://space.skyrocket.de/doc_sdat/o3b-21.htm (accessed March, 2019).

- [30] S. Egami, "A power-sharing multiple-beam mobile satellite in Ka band," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 2, pp. 145-152, 1999.
- [31] J. D. Rosario, "Pricing in satellite markets," NSR, 2017. [Online]. Available: <https://www.nsr.com/pricing-the-satellite-markets/>
- [32] M. Stanley, "Space: Investment Implications of the Final Frontier," 2017. [Online]. Available: https://fa.morganstanley.com/griffithwheelwrightgroup/mediahandler/media/106686/Space_%20Investment%20Implications%20of%20the%20Final%20Frontier.pdf
- [33] A. Szalay and J. Gray, "2020 Computing: Science in an exponential world," *Nature*, vol. 440, no. 7083, p. 413, 2006.
- [34] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *Journal of Parallel and Distributed Computing*, vol. 74, no. 7, pp. 2561-2573, 2014/07/01/ 2014, doi: <https://doi.org/10.1016/j.jpdc.2014.01.003>.
- [35] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, journal article vol. 19, no. 2, pp. 171-209, April 01 2014, doi: 10.1007/s11036-013-0489-0.
- [36] J. Gantz and D. Reinsel, "Extracting value from chaos," *IDC iView*, vol. 1142, no. 2011, pp. 1-12, 2011.
- [37] E. Team, "The Exponential Growth of Data." [Online]. Available: <https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/>
- [38] NSR, "VSAT and Broadband Satellite Markets," 2019, vol. 17th Edition. Accessed: November 2019. [Online]. Available: <https://www.nsr.com/research/3855-2/>
- [39] S. Networks, "Unleashing the Potential of an Empowered - World with the Launch of O3b mPOWER," 2017. [Online]. Available: https://www.ses.com/sites/default/files/2017-09/170908_SES%20Launches%20O3b%20mPOWER_FINAL.pdf
- [40] S. L. Kota, "Broadband satellite networks: trends and challenges," in *Wireless Communications and Networking Conference, 2005 IEEE*, 2005, vol. 3: IEEE, pp. 1472-1478.
- [41] J. Farserotu and R. Prasad, "A survey of future broadband multimedia satellite systems, issues and trends," *IEEE Communications Magazine*, vol. 38, no. 6, pp. 128-133, 2000.
- [42] N. Hosseini, H. Jamal, D. W. Matolak, J. Haque, and T. Magesacher, "UAV Command and Control, Navigation and Surveillance: A Review of Potential 5G and Satellite Systems," *arXiv preprint arXiv:1812.02792*, 2018.
- [43] G. Blank, "WHO CREATES CONTENT?," *Information, Communication & Society*, vol. 16, no. 4, pp. 590-612, 2013/05/01 2013, doi: 10.1080/1369118X.2013.777758.
- [44] M. Hardey, "Generation C: Content, Creation, Connections and Choice," *International Journal of Market Research*, vol. 53, no. 6, pp. 749-770, 2011, doi: 10.2501/ijmr-53-6-749-770.

- [45] S. Dubey, "Role of Satellite in 5G - The 26-28 GHz India 5G Spectrum Workshop," 2018.
- [46] F. Pinto, "High Throughput Satellites and Oil & Gas 'Big Data'," 2015.
- [47] C. Henry, "OneWeb files for Chapter 11 bankruptcy," in *SpaceNews*, ed, 2020.
- [48] D. Werner, "LeoSat and partners out satellites on a diet," in *SpaceNews*, ed, 2018.
- [49] C. Henry, "LeoSat gains Hispasat as second investor, drops demo satellite plans," in *SpaceNews*, ed, 2018.
- [50] C. Henry, "OneWeb formally ends Intelsat merger," in *SpaceNews*, ed, 2018.
- [51] C. Henry, "Inmarsat rejects second EchoStar merger proposal," in *SpaceNews*, ed, 2018.
- [52] C. Henry, "Echostar, now building OneWeb ground network, says company not a competitor," in *SpaceNews*, ed, 2018.
- [53] C. Henry, "Eutelsat pivots for competition with Viasat on European broadband," in *SpaceNews*, ed, 2018.
- [54] C. Henry, "Eutelsat weighs adding more Quantum satellites to fleet," in *SpaceNews*, ed, 2017.
- [55] H. B. Weil, "Commoditization of technology-based products and services: a generic model of market dynamics," 1996.
- [56] H. B. Weil and M. D. Stoughton, "Commoditization of technology-based products and services: the base case scenarios for three industries," 1998.
- [57] D. Oren, "Perception Versus Reality...Satellite-based cellular backhaul." [Online]. Available: <http://www.satmagazine.com/story.php?number=528901969>
- [58] L. P. f. NSR, "No pain, no gain: activating demand elasticity." [Online]. Available: <https://www.nsr.com/no-pain-no-gain-activating-demand-elasticity/>
- [59] Vensim. "TREND(input, average time,initial trend)." https://www.vensim.com/documentation/fn_trend.htm (accessed January, 2020).
- [60] V. S. Prasad, "Fixed VSAT: When will Price Elasticity be Activated?," *The Bottom Line*. [Online]. Available: <https://www.nsr.com/fixe-vs-at-when-will-price-elasticity-be-activated/>
- [61] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and computers in simulation*, vol. 55, no. 1-3, pp. 271-280, 2001.
- [62] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices," *Computer physics communications*, vol. 145, no. 2, pp. 280-297, 2002.

- [63] M. Guerster, J. Grotz, P. Belobaba, E. Crawley, and B. Cameron, "Revenue Management for Communication Satellite Operators - Opportunities and Challenges," presented at the IEEE Aerospace Conference, Big Sky, Montana, 2020.
- [64] P. Belobaba, A. Odoni, and C. Barnhart, *The global airline industry*. John Wiley & Sons, 2015.
- [65] P. P. Belobaba, "OR practice—application of a probabilistic decision model to airline seat inventory control," *Operations Research*, vol. 37, no. 2, pp. 183-197, 1989.
- [66] SES, "Personal communication," ed, 2019.
- [67] NSR, "Flat Panel Satellite Antennas," 2019, vol. 4th Edition. Accessed: November 2019. [Online]. Available: <https://www.nsr.com/research/flat-panel-satellite-antennas-4th-edition/>
- [68] M. Wedel and W. A. Kamakura, *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media, 2012.
- [69] SES. *Proprietary data*.
- [70] P. Belobaba, "Personal communication," ed, 2019.
- [71] P. Belobaba, "Air travel demand and airline seat inventory management," Massachusetts Institute of Technology, 1987.
- [72] S. E. Kimes, "Yield management: a tool for capacity-considered service firms," *Journal of operations management*, vol. 8, no. 4, pp. 348-363, 1989.
- [73] B. Vinod, "Unlocking the value of revenue management in the hotel industry," *Journal of revenue and pricing management*, vol. 3, no. 2, pp. 178-190, 2004.
- [74] M. K. Geraghty and E. Johnson, "Revenue management saves national car rental," *Interfaces*, vol. 27, no. 1, pp. 107-127, 1997.
- [75] W. J. Carroll and R. C. Grimes, "Evolutionary change in product management: Experiences in the car rental industry," *Interfaces*, vol. 25, no. 5, pp. 84-104, 1995.
- [76] R. G. Kasilingam, "Air cargo revenue management: Characteristics and complexities," *European Journal of Operational Research*, vol. 96, no. 1, pp. 36-44, 1997.
- [77] R. G. Kasilingam, "An economic model for air cargo overbooking under stochastic capacity," *Computers & Industrial Engineering*, vol. 32, no. 1, pp. 221-226, 1997/01/01/ 1997, doi: [https://doi.org/10.1016/S0360-8352\(96\)00211-2](https://doi.org/10.1016/S0360-8352(96)00211-2).
- [78] J. S. Billings, A. G. Diener, and B. B. Yuen, "Cargo revenue optimisation," *Journal of Revenue and Pricing Management*, journal article vol. 2, no. 1, pp. 69-79, April 01 2003, doi: 10.1057/palgrave.rpm.5170050.
- [79] B. Slager and L. Kapteijns, "Implementation of cargo revenue management at KLM," *Journal of Revenue and Pricing Management*, vol. 3, no. 1, pp. 80-90, 2004.

- [80] B. Becker and N. Dill, "Managing the complexity of air cargo revenue management," *Journal of Revenue and Pricing Management*, journal article vol. 6, no. 3, pp. 175-187, September 01 2007, doi: 10.1057/palgrave.rpm.5160084.
- [81] S. K. Nair and R. Bapna, "An application of yield management for internet service providers," *Naval Research Logistics (NRL)*, vol. 48, no. 5, pp. 348-362, 2001.
- [82] X. Zhu and S. Singhal, "Optimal resource assignment in internet data centers," in *MASCOTS 2001, Proceedings Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, 2001: IEEE, pp. 61-69.
- [83] A. Byde, M. Sallé, and C. Bartolini, "Market-based resource allocation for utility data centers," *HP Lab, Bristol, Technical Report HPL-2003-188*, 2003.
- [84] P. Dube, Y. Hayel, and L. Wynter, "Yield management for IT resources on demand: analysis and validation of a new paradigm for managing computing centres," *Journal of Revenue and Pricing Management*, vol. 4, no. 1, pp. 24-38, 2005.
- [85] P. Dube, Z. Liu, and L. Wynter, "Methods and apparatus for managing computing resources based on yield management framework," 2014.
- [86] S. Humair, "Yield management for telecommunication networks: defining a new landscape," Massachusetts Institute of Technology, 2001.
- [87] D. Goodman and N. Mandayam, "Network assisted power control for wireless data," *Mobile networks and applications*, vol. 6, no. 5, pp. 409-415, 2001.
- [88] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Efficient power control via pricing in wireless data networks," *IEEE transactions on Communications*, vol. 50, no. 2, pp. 291-303, 2002.
- [89] F. Jallat and F. Ancarani, "Yield management, dynamic pricing and CRM in telecommunications," *Journal of Services Marketing*, vol. 22, no. 6, pp. 465-478, 2008.
- [90] Amazon. "EC2 Spot Instances." <https://aws.amazon.com/ec2/spot/> (accessed October, 12, 2019).
- [91] C. M. Thraves Cortés-Monroy, "New applications in Revenue Management," Massachusetts Institute of Technology, 2017.
- [92] M. Guerster, J. J. G. Luis, E. Crawley, and B. Cameron, "Problem representation of dynamic resource allocation for flexible high throughput satellites," presented at the IEEE Aerospace, Big Sky, MT, 2019.
- [93] J. J. G. Luis, N. Pachler, M. Guerster, I. del Portillo, E. Crawley, and B. Cameron, "Artificial Intelligence Algorithms for Power Allocation in High Throughput Satellites: A Comparison," presented at the IEEE Aerospace, Big Sky, Montana, 2020.
- [94] J. J. G. Luis, M. Guerster, I. del Portillo, E. Crawley, and B. Cameron, "Deep Reinforcement Learning Architecture for Continuous Power Allocation in High Throughput Satellites," *arXiv preprint arXiv:1906.00571*, 2019.

- [95] G. Maral and M. Bousquet, *Satellite communications systems: systems, techniques and technology*. John Wiley & Sons, 2011.
- [96] A. Federgruen and H. Groenevelt, "The greedy procedure for resource allocation problems: Necessary and sufficient conditions for optimality," *Operations research*, vol. 34, no. 6, pp. 909-918, 1986.
- [97] I. d. Portillo, "Space and Aerial Architectures to Expand Global Connectivity," Ph.D., Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 2020.
- [98] A. París i Bordas, "Power and bandwidth allocation in multibeam satellite systems," Universitat Politècnica de Catalunya, 2018.
- [99] C. Siocos, "Broadcasting-Satellite Coverage-Geometrical Considerations," *IEEE Transactions on Broadcasting*, no. 4, pp. 84-87, 1973.
- [100] E. Etsi, "302 307:" Digital Video Broadcasting (DVB)," *Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications*, vol. 1, p. 2, 2006.
- [101] E. ETSI, "302 307-2 Digital Video Broadcasting (DVB)," *Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications*.
- [102] *A fast, simple, end-to-end platform for tiling, visualizing, and analyzing 3D geospatial data*. (2019). [Online]. Available: <https://cesium.com>
- [103] E. Hossain, D. Niyato, and D. I. Kim, "Evolution and future trends of research in cognitive radio: a contemporary survey," *Wireless Communications and Mobile Computing*, vol. 15, no. 11, pp. 1530-1564, 2015.
- [104] N. Abbas, Y. Nasser, and K. El Ahmad, "Recent advances on artificial intelligence and learning techniques in cognitive radio networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 174, 2015.
- [105] R. Zhang, Y.-C. Liang, and S. Cui, "Dynamic resource allocation in cognitive radio networks," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 102-114, 2010.
- [106] S. K. Sharma, S. Chatzinotas, and B. Ottersten, "Cognitive radio techniques for satellite communication systems," in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, 2013: IEEE, pp. 1-5.
- [107] S. K. Sharma, S. Maleki, S. Chatzinotas, J. Grotz, J. Krause, and B. Ottersten, "Joint carrier allocation and beamforming for cognitive SatComs in Ka-band (17.3–18.1 GHz)," in *2015 IEEE International Conference on Communications (ICC)*, 2015: IEEE, pp. 873-878.
- [108] P. Ferreira *et al.*, "Multi-Objective Reinforcement Learning for Cognitive Radio--Based Satellite Communications," in *34th AIAA International Communications Satellite Systems Conference*, 2016, p. 5726.

- [109] P. V. R. Ferreira *et al.*, "Multi-objective reinforcement learning-based deep neural networks for cognitive space communications," in *Cognitive Communications for Aerospace Applications Workshop (CCAA), 2017*, 2017: IEEE, pp. 1-8.
- [110] P. V. R. Ferreira *et al.*, "Multi-objective Reinforcement Learning for Cognitive Satellite Communications using Deep Neural Network Ensembles," *IEEE Journal on Selected Areas in Communications*, 2018.
- [111] T. M. Hackett, S. G. Bilén, P. V. R. Ferreira, A. M. Wyglinski, R. C. Reinhart, and D. J. Mortensen, "Implementation and On-Orbit Testing Results of a Space Communications Cognitive Engine," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 825-842, 2018.
- [112] S. Kandeepan, L. De Nardis, M.-G. Di Benedetto, A. Guidotti, and G. E. Corazza, "Cognitive satellite terrestrial radios," in *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, 2010: IEEE, pp. 1-6.
- [113] E. Lagunas, S. K. Sharma, S. Maleki, S. Chatzinotas, and B. Ottersten, "Resource allocation for cognitive satellite communications with incumbent terrestrial networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 3, pp. 305-317, 2015.
- [114] S. Vassaki, M. I. Poulakis, A. D. Panagopoulos, and P. Constantinou, "Power allocation in cognitive satellite terrestrial networks with QoS constraints," *IEEE Communications Letters*, vol. 17, no. 7, pp. 1344-1347, 2013.
- [115] NASA, "Three Newly Designed Tracking and Data Relay Satellites To Help Replenish Existing On-Orbit Fleet," 2001. [Online]. Available: https://www.nasa.gov/sites/default/files/97440main_TDRS_fs_9.18.pdf
- [116] K. Baohua, F. Ligang, M. Li, L. Yuheng, X. Xiaoshen, and C. Liyu, "TDRSS resource scheduling modes and service modeling based on mission planning," in *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 19-20 Dec. 2015 2015, pp. 806-809, doi: 10.1109/IAEAC.2015.7428668.
- [117] B. Deng, C. Jiang, L. Kuang, S. Guo, J. Lu, and S. Zhao, "Two-phase task scheduling in data relay satellite systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1782-1793, 2018.
- [118] B. Deng, C. Jiang, L. Kuang, S. Guo, N. Ge, and J. Lu, "Preemptive dynamic scheduling algorithm for data relay satellite systems," in *2017 IEEE International Conference on Communications (ICC)*, 2017: IEEE, pp. 1-6.
- [119] L. Wang, C. Jiang, L. Kuang, S. Wu, H. Huang, and Y. Qian, "High-Efficient Resource Allocation in Data Relay Satellite Systems With Users Behavior Coordination," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12072-12085, 2018.
- [120] L. Wang, C. Jiang, L. Kuang, X. Zhu, J. Yan, and L. Fei, "Repeated game based cooperation mechanism for antenna beam resource allocation in TDRSS," in *2018 IEEE International Conference on Communications (ICC)*, 2018: IEEE, pp. 1-6.

- [121] C. Courcoubetis and R. Weber, *Pricing communication networks: economics, technology and modelling*. John Wiley & Sons, 2003.
- [122] H. Wang, A. Liu, X. Pan, and J. Yang, "Optimization of power allocation for multiusers in multi-spot-beam satellite communication systems," *Mathematical Problems in Engineering*, vol. 2014, 2014.
- [123] Y. Hong, A. Srinivasan, B. Cheng, L. Hartman, and P. Andreadis, "Optimal power allocation for multiple beam satellite systems," in *Radio and Wireless Symposium, 2008 IEEE*, 2008: IEEE, pp. 823-826.
- [124] M. J. Neely, E. Modiano, and C. E. Rohrs, "Power allocation and routing in multibeam satellites with time-varying channels," *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 138-152, 2003, doi: 10.1109/TNET.2002.808401.
- [125] A. Aravanis, G. Danoy, P. Arapoglou, P. Cottis, and B. Ottersten, "Multi-objective optimization approach to power allocation in multibeam systems," in *30th AIAA International Communications Satellite System Conference (ICSSC)*, 2012, p. 15202.
- [126] A. Aravanis, B. Shankar, G. Danoy, P.-D. Arapoglou, P. Cottis, and B. Ottersten, "Power allocation in Multibeam satellites-A hybrid-Genetic Algorithm approach," in *ESA Workshop on Advanced Flexible Telecom Payloads*, 2012: European Space Agency, pp. 1-5.
- [127] A. I. Aravanis, B. S. MR, P.-D. Arapoglou, G. Danoy, P. G. Cottis, and B. Ottersten, "Power allocation in multibeam satellite systems: A two-stage multi-objective optimization," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3171-3182, 2015.
- [128] Y. He, Y. Jia, and X. Zhong, "A traffic-awareness dynamic resource allocation scheme based on multi-objective optimization in multi-beam mobile satellite communication systems," *International Journal of Distributed Sensor Networks*, vol. 13, no. 8, 2017.
- [129] F. R. Durand and T. Abrão, "Power allocation in multibeam satellites based on particle swarm optimization," *AEU-International Journal of Electronics and Communications*, vol. 78, pp. 124-133, 2017.
- [130] A. Destounis and A. D. Panagopoulos, "Dynamic Power Allocation for Broadband Multi-Beam Satellite Communication Networks," *IEEE Communications Letters*, vol. 15, no. 4, pp. 380-382, 2011, doi: 10.1109/LCOMM.2011.020111.102201.
- [131] N. K. Srivastava and A. K. Chaturvedi, "Flexible and Dynamic Power Allocation in Broadband Multi-Beam Satellites," *IEEE Communications Letters*, vol. 17, no. 9, pp. 1722-1725, 2013, doi: 10.1109/LCOMM.2013.080113.130615.
- [132] U. Park, H. W. Kim, D. S. Oh, and B. J. Ku, "A dynamic bandwidth allocation scheme for a multi-spot-beam satellite system," *Etri Journal*, vol. 34, no. 4, pp. 613-616, 2012.
- [133] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

- [134] J. P. Choi and V. W. S. Chan, "Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2983-2993, 2005, doi: 10.1109/TWC.2005.858365.
- [135] U. Park, H. W. Kim, D. S. Oh, and B. J. Ku, "Flexible Bandwidth Allocation Scheme Based on Traffic Demands and Channel Conditions for Multi-Beam Satellite Systems," in *2012 IEEE Vehicular Technology Conference (VTC Fall)*, 3-6 Sept. 2012 2012, pp. 1-5, doi: 10.1109/VTCFall.2012.6399225.
- [136] H. Wang, A. Liu, X. Pan, and L. Jia, "Optimal bandwidth allocation for multi-spot-beam satellite communication systems," in *Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, 2013: IEEE, pp. 2794-2798.
- [137] X. Hu, S. Liu, R. Chen, W. Wang, and C. Wang, "A Deep Reinforcement Learning-Based Framework for Dynamic Resource Allocation in Multibeam Satellite Systems," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1612-1615, 2018.
- [138] S. Liu, X. Hu, and W. Wang, "Deep Reinforcement Learning Based Dynamic Channel Allocation Algorithm in Multibeam Satellite Systems," *IEEE ACCESS*, vol. 6, pp. 15733-15742, 2018.
- [139] M. Umehira, S. Fujita, Z. Gao, and J. Wang, "Dynamic channel assignment based on interference measurement with threshold for multi-beam mobile satellite networks," in *2013 19th Asia-Pacific Conference on Communications (APCC)*, 2013: IEEE, pp. 688-692.
- [140] A. Tirmizi, R. S. Mishra, and A. Zadgaonkar, "An efficient channel assignment strategy in cellular mobile network using hybrid genetic algorithm," *International Journal of Electrical, Electronics and Computer Engineering*, vol. 4, no. 2, p. 127, 2015.
- [141] N. Funabiki and S. Nishikawa, "A gradual neural-network approach for frequency assignment in satellite communication systems," *IEEE transactions on neural networks*, vol. 8, no. 6, pp. 1359-1370, 1997.
- [142] S. Salcedo-Sanz, R. Santiago-Mozos, and C. Bousoño-Calzón, "A hybrid Hopfield network-simulated annealing approach for frequency assignment in satellite communications systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 2, pp. 1108-1116, 2004.
- [143] S. Salcedo-Sanz and C. Bousoño-Calzón, "A Hybrid Neural-Genetic Algorithm for the Frequency Assignment Problem in Satellite Communications," *Applied Intelligence*, journal article vol. 22, no. 3, pp. 207-217, May 01 2005, doi: 10.1007/s10791-005-6619-y.
- [144] F. Li, K. Lam, J. Hua, K. Zhao, N. Zhao, and L. Wang, "Improving Spectrum Management for Satellite Communication Systems with Hunger Marketing," *IEEE Wireless Communications Letters*, pp. 1-1, 2019, doi: 10.1109/LWC.2019.2893659.
- [145] F. Li, K.-Y. Lam, X. Liu, J. Wang, K. Zhao, and L. Wang, "Joint Pricing and Power Allocation for Multibeam Satellite Systems With Dynamic Game Model," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2398-2408, 2018.

- [146] F. Li, K. Lam, N. Zhao, X. Liu, K. Zhao, and L. Wang, "Spectrum Trading for Satellite Communication Systems With Dynamic Bargaining," *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4680-4693, 2018, doi: 10.1109/TCOMM.2018.2837909.
- [147] J. Sun and E. Modiano, "Channel allocation using pricing in satellite networks," in *2006 40th Annual Conference on Information Sciences and Systems*, 2006: IEEE, pp. 182-187.
- [148] J. Sun, "Dynamic channel allocation in satellite and wireless networks," Massachusetts Institute of Technology, 2007.
- [149] J. Anzalchi *et al.*, "Beam hopping in multi-beam broadband satellite systems: System simulation and performance comparison with non-hopped systems," in *2010 5th Advanced Satellite Multimedia Systems Conference and the 11th Signal Processing for Space Communications Workshop*, 2010: IEEE, pp. 248-255.
- [150] A. Kyrgiazos, B. Evans, and P. Thompson, "Smart gateways designs with time switched feeders and beam hopping user links," in *2016 8th Advanced Satellite Multimedia Systems Conference and the 14th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, 2016: IEEE, pp. 1-6.
- [151] X. Alberti *et al.*, "System capacity optimization in time and frequency for multibeam multi-media satellite systems," in *2010 5th Advanced Satellite Multimedia Systems Conference and the 11th Signal Processing for Space Communications Workshop*, 13-15 Sept. 2010 2010, pp. 226-233, doi: 10.1109/ASMS-SPSC.2010.5586902.
- [152] J. Lei, "Multi-beam satellite resource allocation optimization for beam hopping transmission," Universitat Autònoma de Barcelona, 2011.
- [153] J. Lei and M. Á. Vázquez-Castro, "Multibeam satellite frequency/time duality study and capacity optimization," *Journal of Communications and Networks*, vol. 13, no. 5, pp. 472-480, 2011, doi: 10.1109/JCN.2011.6112304.
- [154] S. Shi, G. Li, Z. Li, H. Zhu, and B. Gao, "Joint power and bandwidth allocation for beam-hopping user downlinks in smart gateway multibeam satellite systems," *International Journal of Distributed Sensor Networks*, vol. 13, no. 5, p. 1550147717709461, 2017.
- [155] P. Angeletti, D. Fernandez Prim, and R. Rinaldo, "Beam hopping in multi-beam broadband satellite systems: System performance and payload architecture analysis," in *24th AIAA International Communications Satellite Systems Conference*, 2006, p. 5376.
- [156] J. Lei and M. A. Vazquez-Castro, "Joint power and carrier allocation for the multibeam satellite downlink with individual SINR constraints," in *Communications (ICC), 2010 IEEE International Conference on*, 2010: IEEE, pp. 1-5.
- [157] H. Wang, Z. Liu, Z. Cheng, Y. Miao, W. Feng, and N. Ge, "Maximization of link capacity by joint power and spectrum allocation for smart satellite transponder," in *2017 23rd Asia-Pacific Conference on Communications (APCC)*, 2017: IEEE, pp. 1-6.

- [158] G. Cocco, T. De Cola, M. Angelone, and Z. Katona, "Radio resource management strategies for DVB-S2 systems operated with flexible satellite payloads," in *Advanced Satellite Multimedia Systems Conference and the 14th Signal Processing for Space Communications Workshop (ASMS/SPSC), 2016 8th*, 2016: IEEE, pp. 1-8.
- [159] G. Cocco, T. d. Cola, M. Angelone, Z. Katona, and S. Erl, "Radio Resource Management Optimization of Flexible Satellite Payloads for DVB-S2 Systems," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 266-280, 2018, doi: 10.1109/TBC.2017.2755263.
- [160] N. Pachler, J. J. G. Luis, M. Guerster, E. Crawley, and B. Cameron, "Allocating Power and Bandwidth in Multibeam Satellite Systems using Particle Swarm Optimization," presented at the 2020 IEEE Aerospace Conference, Big Sky, Montana, 2020.
- [161] A. Paris, I. d. Portillo, B. Cameron, and E. Crawley, "A Genetic Algorithm for Joint Power and Bandwidth Allocation in Multibeam Satellite Systems," presented at the IEEE Aerospace, Big Sky, Montana, 2019.
- [162] Z. Ji, Y. Wang, W. Feng, and J. Lu, "Delay-aware power and bandwidth allocation for multiuser satellite downlinks," *IEEE Communications Letters*, vol. 18, no. 11, pp. 1951-1954, 2014.
- [163] J. P. Choi and V. W. Chan, "Optimum multibeam satellite downlink power allocation based on traffic demands," in *Global Telecommunications Conference, 2002. GLOBECOM'02. IEEE, 2002*, vol. 3: IEEE, pp. 2875-2881.
- [164] J. P. Choi and V. W. S. Chan, "An Efficient Resource Scheduling Algorithm for Phased Array Antenna Satellites," in *MILCOM 2006 - 2006 IEEE Military Communications conference, 23-25 Oct. 2006 2006*, pp. 1-7, doi: 10.1109/MILCOM.2006.302416.
- [165] J. P. Choi and V. W. S. Chan, "Resource management for advanced transmission antenna satellites," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1308-1321, 2009, doi: 10.1109/TWC.2009.071131.
- [166] A. Jahn, "Resource management model and performance evaluation for satellite communications," *International journal of satellite communications*, vol. 19, no. 2, pp. 169-203, 2001.
- [167] M. Cygan, S. Kratsch, M. Pilipczuk, M. Pilipczuk, and M. Wahlström, "Clique cover and graph separation: New incompressibility results," *ACM Transactions on Computation Theory (TOCT)*, vol. 6, no. 2, p. 6, 2014.
- [168] M. Cygan, M. Pilipczuk, and M. Pilipczuk, "Known algorithms for edge clique cover are probably optimal," *SIAM Journal on Computing*, vol. 45, no. 1, pp. 67-83, 2016.
- [169] J. Gramm, J. Guo, F. Hüffner, and R. Niedermeier, "Data reduction, exact, and heuristic algorithms for clique cover," in *Proceedings of the Meeting on Algorithm Engineering & Experiments, 2006: Society for Industrial and Applied Mathematics*, pp. 86-94.
- [170] J. Gramm, J. Guo, F. Hüffner, and R. Niedermeier, "Data reduction and exact algorithms for clique cover," *Journal of Experimental Algorithmics (JEA)*, vol. 13, p. 2, 2009.

- [171] R. McIntyre and M. Soltys, "An improved upper bound and algorithm for clique covers," *Journal of Discrete Algorithms*, vol. 48, pp. 42-56, 2018.
- [172] Q. Wu and J.-K. Hao, "A review on algorithms for maximum clique problems," *European Journal of Operational Research*, vol. 242, no. 3, pp. 693-709, 2015.
- [173] J. Camino, C. Artigues, L. Houssin, and S. Mourgues, "Mixed-integer linear programming for multibeam satellite systems design: Application to the beam layout optimization," in *2016 Annual IEEE Systems Conference (SysCon)*, 18-21 April 2016 2016, pp. 1-6, doi: 10.1109/SYSCON.2016.7490613.
- [174] J.-T. Camino, S. Mourgues, C. Artigues, and L. Houssin, "A greedy approach combined with graph coloring for non-uniform beam layouts under antenna constraints in multibeam satellite systems," in *2014 7th Advanced Satellite Multimedia Systems Conference and the 13th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, 2014: IEEE, pp. 374-381.
- [175] C. Qian, S. Zhang, and W. Zhou, "Traffic-based dynamic beam coverage adjustment in satellite mobile communication," in *2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP)*, 23-25 Oct. 2014 2014, pp. 1-6, doi: 10.1109/WCSP.2014.6992014.
- [176] B. Wenqian, W. Weidong, L. Shuaijun, and C. Gaofeng, "Beam Coverage Dynamic Adjustment Scheme Based on Maximizing System Capacity for Multi-beam Satellite Communication System," Singapore, 2018: Springer Singapore, in *Space Information Networks*, pp. 288-298.
- [177] H. Boche and M. Schubert, "Solution of the SINR downlink beamforming problem," in *Conference on Information Sciences and Systems*, 2002: Princeton University.
- [178] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18-28, 2004.
- [179] A. Kyrgiazos, B. Evans, and P. Thompson, "Irregular beam sizes and non-uniform bandwidth allocation in HTS multibeam satellite systems," in *31st AIAA International Communications Satellite Systems Conference (ICSSC)*, 2013.
- [180] P. Chitre and F. Yegenoglu, "Next-generation satellite networks: architectures and implementations," *IEEE Communications Magazine*, vol. 37, no. 3, pp. 30-36, 1999.
- [181] G. Cocco, T. De Cola, M. Angelone, Z. Katona, and S. Erl, "Radio Resource Management Optimization of Flexible Satellite Payloads for DVB-S2 Systems," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 266-280, 2018.
- [182] E. o. Mathematics. "Mathematical programming." http://www.encyclopediaofmath.org/index.php?title=Mathematical_programming&oldid=38945 (accessed).

- [183] N. Keon and G. A. Anandalingam, "A new pricing model for competitive telecommunications services using congestion discounts," *INFORMS Journal on Computing*, vol. 17, no. 2, pp. 248-262, 2005.
- [184] F. Vieira and M. Á. Vázquez-Castro, "Dynamic Price-based resource allocation mechanism for ACM systems," in *25th AIAA International Communications Satellite Systems Conference (organized by APSCC)*, 2007, p. 3142.
- [185] A. Fiaschetti, M. Fiaschetti, A. Pietrabissa, and M. Petrone, "Congestion pricing for dynamic bandwidth allocation in satellite networks: A game-theoretic approach," in *Satellite Telecommunications (ESTEL), 2012 IEEE First AESS European Conference on*, 2012: IEEE, pp. 1-6.
- [186] C. J. Corbett, D. Zhou, and C. S. Tang, "Designing supply contracts: Contract type and information asymmetry," *Management Science*, vol. 50, no. 4, pp. 550-559, 2004.
- [187] J. J. G. Luis, "A Comparison of Artificial Intelligence Algorithms for Dynamic Power Allocation in Flexible High Throughput Satellites," M.Sc., Massachusetts Institute of Technology, 2020.
- [188] N. Pachler, M. Guerster, I. del Portillo, E. Crawley, and B. Cameron, "Static beam placement and frequency plan algorithms for LEO constellations," *International Journal of Satellite Communications and Networking*, 2020.
- [189] R. Alinque, "Joint optimization of Beam Placement and Shaping for Multi-Beam High Throughput Satellite systems using Gradient Descent," B.Sc., Massachusetts Institute of Technology, 2020.
- [190] D. Jones, "Short-Term Traffic Forecasting for a Smart Satellite Communications System," M.Sc., Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 2020.
- [191] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443-473, 2006, doi: 10.1016/j.ijforecast.2006.01.001.
- [192] C. Chatfield, *Time-series forecasting*. CRC press, 2000.
- [193] G. Simon, A. Lendasse, M. Cottrell, J.-C. Fort, and M. Verleysen, "Time series forecasting: Obtaining long term trends with self-organizing maps," *Pattern Recognition Letters*, vol. 26, no. 12, pp. 1795-1808, 2005.
- [194] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914-938, 2016.
- [195] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030-1081, 2014/10/01/ 2014, doi: <https://doi.org/10.1016/j.ijforecast.2014.08.008>.
- [196] S. Arora and J. W. Taylor, "Forecasting electricity smart meter data using conditional kernel density estimation," *Omega*, vol. 59, pp. 47-59, 2016.

- [197] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-term solar power forecasting based on weighted Gaussian process regression," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 1, pp. 300-308, 2017.
- [198] E. V. Bonilla, K. M. Chai, and C. Williams, "Multi-task Gaussian process prediction," in *Advances in neural information processing systems*, 2008, pp. 153-160.
- [199] Y. Zhang, G. Luo, and F. Pu, "Power load forecasting based on multi-task Gaussian process," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 3651-3656, 2014.
- [200] J.-B. Fiot and F. Dinuzzo, "Electricity demand forecasting by multi-task learning," *Ieee T Smart Grid*, vol. 9, no. 2, pp. 544-551, 2016.
- [201] Y. Yang, S. Li, W. Li, and M. Qu, "Power load probability density forecasting using Gaussian process quantile regression," *Applied Energy*, vol. 213, pp. 499-509, 2018.
- [202] C. E. Rasmussen and C. K. Williams, "Gaussian Processes for Machine Learning the MIT Press," *Cambridge, MA*, 2006.
- [203] E. Crawley, B. Cameron, and D. Selva, *System architecture : strategy and product development for complex systems*. Hoboken, NJ: Pearson Higher Education, Inc., 2016, pp. xiii, 465 pages.
- [204] T. Ahola, "How to deliver value to customers with latent needs in a business-to-business project delivery context: empirical illustration from the construction industry," in *Proceedings of the 22nd IMP-conference*, 2006.
- [205] L. Carlgren, "Identifying latent needs: towards a competence perspective on attractive quality creation," *Total Quality Management & Business Excellence*, vol. 24, no. 11-12, pp. 1347-1363, 2013.
- [206] C.-C. Yang, "An analytical methodology for identifying the latent needs of customers," *Total Quality Management & Business Excellence*, vol. 24, no. 11-12, pp. 1332-1346, 2013.
- [207] "Dual-channel Plans." KVH. <https://www.kvh.com/airtime-and-content/commercial-mini-vsat-broadband-airtime/hts-dual-channel-plans> (accessed November, 2019).
- [208] Speedcast. "VSAT Satellite." <https://www.speedcast.com/our-solution/connectivity/vsat/> (accessed February, 2020).
- [209] Marlink. "Maritime VSAT." <https://marlink.com/category-products/sealink/> (accessed February, 2020).
- [210] S. Dharmapurikar, P. Krishnamurthy, T. Sproull, and J. Lockwood, "Deep packet inspection using parallel bloom filters," in *11th Symposium on High Performance Interconnects, 2003. Proceedings.*, 2003: IEEE, pp. 44-51.
- [211] P. Kotler and G. Armstrong, *Principles of marketing*. Pearson education, 2010.

- [212] B. Cooil, L. Aksoy, and T. L. Keiningham, "Approaches to Customer Segmentation," *Journal of Relationship Marketing*, vol. 6, no. 3-4, pp. 9-39, 2008/01/14 2008, doi: 10.1300/J366v06n03_02.
- [213] C. H. Fine, "Are you modular or integral? Be sure your supply chain knows," *Strategy+ Business*, vol. 39, no. 2, pp. 1-8, 2005.
- [214] R. N. Bolton and M. B. Myers, "Price-based global market segmentation for services," *Journal of Marketing*, vol. 67, no. 3, pp. 108-128, 2003.
- [215] J. Ganesh, M. J. Arnold, and K. E. Reynolds, "Understanding the customer base of service providers: an examination of the differences between switchers and stayers," *Journal of marketing*, vol. 64, no. 3, pp. 65-87, 2000.
- [216] J. B. Heide and A. M. Weiss, "Vendor consideration and switching behavior for buyers in high-technology markets," *Journal of marketing*, vol. 59, no. 3, pp. 30-43, 1995.
- [217] S. M. Keaveney, "Customer switching behavior in service industries: An exploratory study," *Journal of marketing*, vol. 59, no. 2, pp. 71-82, 1995.
- [218] J.-B. E. Steenkamp and F. Ter Hofstede, "International market segmentation: issues and perspectives," *International journal of research in marketing*, vol. 19, no. 3, pp. 185-213, 2002.
- [219] J. Bayer, "Customer segmentation in the telecommunications industry," *Journal of Database marketing & customer strategy management*, vol. 17, no. 3-4, pp. 247-256, 2010.
- [220] E. Xevelonakis and P. Som, "The impact of social network-based segmentation on customer loyalty in the telecommunication industry," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, no. 2, pp. 98-106, 2012.
- [221] M. F. Bacila, A. Radulescu, and I. L. Marar, "Customer segmentation based on the value of consumption patterns in telecommunications," in *The Proceedings of the International Conference "Marketing-from Information to Decision"*, 2013: Babes Bolyai University, p. 36.
- [222] I. T. Union, "Radio Regulations Articles," 2016, vol. Edition of 2016. [Online]. Available: <http://search.itu.int/history/HistoryDigitalCollectionDocLibrary/1.43.48.en.101.pdf>
- [223] NSR, "Satellite Capacity Pricing Index," 2020, vol. 6th Edition. Accessed: April 2020. [Online]. Available: <https://www.nsr.com/research/satellite-capacity-pricing-index-6th-edition-2020/>
- [224] Euroconsult, "FSS Capacity Pricing Trends." [Online]. Available: <https://i2.wp.com/www.broadbandtvnews.com/wp-content/uploads/2019/12/image001.jpg?ssl=1>
- [225] G. A. f. NSR, "The satellite capacity price conundrum." [Online]. Available: <https://www.nsr.com/the-satellite-capacity-price-conundrum/>
- [226] G. A. f. NSR, "Satellite capacity pricing: is the retail era just getting started?." [Online]. Available: <https://www.nsr.com/satellite-capacity-pricing-is-the-retail-era-just-getting-started/>

- [227] R. E. Frank, W. F. Massey, and Y. Wind, *Market segmentation*. Prentice Hall, 1972.
- [228] P. Hague and M. Harrison, "Market segmentation in B2B markets," *B2B International Ltd. Path*, 2002. [Online]. Available: <https://www.b2binternational.com/publications/b2b-segmentation-research/>.
- [229] W. NetWorld's–SatCom, "The role of satellites in 5G," *white paper*, July, 2014.
- [230] O. Acker, F. Pötscher, and T. Lefot, "Why satellites matter - The relevance of commercial satellites in the 21st century – a perspective 2012-2020," Amsterdam, Brussels, Frankfurt, Vienna, 2012.
- [231] M. Aldebert, M. Ivaldi, and C. Roucolle, "Telecommunications demand and pricing structure: An econometric analysis," *Telecommunication Systems*, vol. 25, no. 1-2, pp. 89-115, 2004.
- [232] P. Hackl and A. H. Westlund, "Demand for international telecommunication time-varying price elasticity," *Journal of Econometrics*, vol. 70, no. 1, pp. 243-260, 1996.
- [233] H. Ouwersloot and P. Rietveld, "On the distance dependence of the price elasticity of telecommunications demand; review, analysis, and alternative theoretical backgrounds," *The annals of regional science*, vol. 35, no. 4, pp. 577-594, 2001.
- [234] C. Garbacz and H. G. Thompson Jr, "Demand for telecommunication services in developing countries," *Telecommunications policy*, vol. 31, no. 5, pp. 276-289, 2007.
- [235] R. E. Park, B. M. Wetzel, and B. M. Mitchell, "Price elasticities for local telephone calls," *Econometrica: Journal of the Econometric Society*, pp. 1699-1730, 1983.
- [236] F. A. Wolak, "Can universal service survive in a competitive telecommunications environment? Evidence from the United States consumer expenditure survey," *Information Economics and Policy*, vol. 8, no. 3, pp. 163-203, 1996.
- [237] J. P. Gatto, H. H. Kelejian, and S. W. Stephan, "Stochastic generalizations of demand systems with an application to telecommunications," *Information Economics and Policy*, vol. 3, no. 4, pp. 283-309, 1988.
- [238] T. Garín-Muñoz and T. Perez-Amaral, "Econometric modelling of Spanish very long distance international calling," *Information Economics and Policy*, vol. 10, no. 2, pp. 237-252, 1998.
- [239] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.
- [240] B. D. Bernheim and M. D. Whinston, *Microeconomics* (The McGraw-Hill series in economics). New York, NY : McGraw-Hill/Irwin, 2014.
- [241] G. A. f. NSR, "A (LEO) race to the bottom?." [Online]. Available: <https://www.nsr.com/a-leo-race-to-the-bottom/>
- [242] L. Jiale and D. Huiying, "Study on airline customer value evaluation based on RFM model," in *2010 International Conference On Computer Design And Applications*, 2010, vol. 4: IEEE, pp. V4-278-V4-281.

- [243] R. Lee and J. Murphy, "From loyalty to switching: exploring the determinants in the transition," *ANZMAC 2005*, 2005.
- [244] Y. Aviv and G. Vulcano, "Dynamic list pricing," in *The Oxford handbook of pricing management*, 2012.
- [245] A. V. den Boer, "Dynamic pricing and learning: historical origins, current research, and new directions," *Surveys in operations research and management science*, vol. 20, no. 1, pp. 1-18, 2015.
- [246] W. C. Cheung, D. Simchi-Levi, and H. Wang, "Dynamic pricing and demand learning with limited price experimentation," *Operations Research*, vol. 65, no. 6, pp. 1722-1731, 2017.
- [247] O. Besbes and A. Zeevi, "Dynamic Pricing Without Knowing the Demand Function: Risk Bounds and Near-Optimal Algorithms," *Operations Research*, vol. 57, no. 6, pp. 1407-1420, 2009, doi: 10.1287/opre.1080.0640.
- [248] O. Besbes and A. Zeevi, "On the (surprising) sufficiency of linear models for dynamic pricing with demand learning," *Management Science*, vol. 61, no. 4, pp. 723-739, 2015.
- [249] T. Boyacı and Ö. Özer, "Information acquisition for capacity planning via pricing and advance selling: When to stop and act?," *Operations Research*, vol. 58, no. 5, pp. 1328-1349, 2010.
- [250] Z. Wang, S. Deng, and Y. Ye, "Close the gaps: A learning-while-doing algorithm for single-product revenue management problems," *Operations Research*, vol. 62, no. 2, pp. 318-331, 2014.
- [251] P. P. Belobaba and J. L. Wilson, "Impacts of yield management in competitive airline markets," *Journal of Air Transport Management*, vol. 3, no. 1, pp. 3-9, 1997.
- [252] W. S. Swelbar and P. P. Belobaba, "Airline data project," *Massachusetts Institute of Technology*, 2010.
- [253] J.-P. Rodrigue, *The geography of transport systems*. Taylor & Francis, 2016.
- [254] D. S. f. Aislelabs, "How Airports Globally are Responding to Coronavirus (Updated Frequently)," 2020. [Online]. Available: <https://www.aislelabs.com/blog/2020/03/27/how-airports-globally-are-responding-to-coronavirus-updated-frequently/>
- [255] G. Lordos, M. Guerster, B. Cameron, O. de Weck, and J. Hoffman, "Tradespace Exploration of Space Settlement Architectures Using Long-term Cost and Benefit Metrics."
- [256] S. D. Society. "What Is SD?" <https://www.systemdynamics.org/what-is-sd> (accessed January, 2020).
- [257] G. P. Richardson, "System Dynamics," in *Encyclopedia of Operations Research and Management Science*, S. I. Gass and M. C. Fu Eds. Boston, MA: Springer US, 2013, pp. 1519-1522.
- [258] T. Van Zandt, "Chapter 8 — Elasticity of Demand," 2012. [Online]. Available: <https://faculty.insead.edu/vanzandt/pm/Session07/FPM-08.pdf>