Distributed Satellite Communication System Design: First-Order Interactions between System and Network Architectures

by

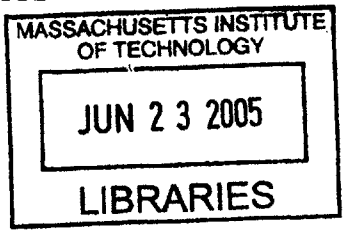
Jennifer E. Underwood

S.B. Aerospace Engineering with Information Technology
Massachusetts Institute of Technology, 2003

SUBMITTED TO THE DEPARTMENT OF AERONAUTICS AND ASTRONAUTICS IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN AERONAUTICS AND ASTRONAUTICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2005



Signature of Author _____
Department of Aeronautics and Astronautics
20 May 2005

Certified by _____
Dr. Dorothy Poppe
Charles Stark Draper Laboratory
Thesis Supervisor

Certified by _____
Professor Olivier de Weck
Department of Aeronautics and Astronautics and Engineering Systems
Thesis Advisor

Accepted by _____
Jaime Peraire
Professor of Aeronautics and Astronautics
Chair, Committee on Graduate Students

Distributed Satellite Communication System Design:
First-Order Interactions between System and Network Architectures
by

Jennifer E. Underwood

Submitted to the Department of Aeronautics and Astronautics on
May 20, 2005, in partial fulfillment of the requirements for the
Degree of Masters of Science in Aeronautics and Astronautics

Abstract

Humanity now exists in the midst of the fast-moving Information Age, a period of history characterized by fast travel and even faster information transfer. As data becomes seemingly more valuable than physical possessions, the introduction of exciting applications for communications services becomes ever more critical for the success – and in some cases, survival – of businesses and even nations.

While the majority of these innovations have occurred over cable and fiber, a number of the most socially significant have occurred due to the introduction of satellites. Terrestrial fiber and cable systems have a number of advantages, but the extent of their reach and the cost of installation – in terms of both capital and time – favor industrialized nations over more remote and underdeveloped communities.

Even as satellites offer the only real chance for ultimate communications ubiquity and true global unity, there remains a significant cost-benefit barrier. Few commercial satellite systems have succeeded economically without first falling victim to bankruptcy. The upfront capital required to implement a satellite communications system is staggering, and historically satellite companies have failed to adequately match capacity and service options to the current and actual future demand.

The design process itself is an inherent limiting factor to the achievable cost and performance of a system. Traditionally, the first step toward designing satellite communication systems – as well as terrestrial, sensor web, and ad hoc networks – has been to specify the system topology (e.g., the orbits of the satellites and the locations of the ground stations) based on the desired market and then to design the network protocols to make the most of the available resources.

Such a sequential process assumes that the design of the network architecture (e.g., protocols, packet structure, etc) does not drive the design of the system architecture (e.g., constellation topology, spacecraft design, etc). This thesis will show that in the case of Ka-band distributed satellite communication systems this fundamental assumption is not valid, and can have a significant impact on the success (cost, capacity, customer satisfaction) of the resulting satellite communication system. Furthermore, this thesis will show that how a designer values performance during the design and decision process can have a substantial impact on the quality of the design path taken through the trade space of possible joint architectures.

Technical Supervisor: Dr. Dorothy Poppe
Title: Senior Member of the Technical Staff

Thesis Advisor: Professor Olivier de Weck, Ph.D.
Title: Robert N. Noyce Assistant Professor of Aeronautics and Astronautics and Engineering Systems

ACKNOWLEDGMENT

20 May 2005

This thesis was prepared at The Charles Stark Draper Laboratory, Inc, under the Distributed Sensor Network Internal Research and Development #13152 and #12579 001.

Publication of this thesis does not constitute approval by Draper or the sponsoring agency of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

PERSONAL ACKNOWLEDGMENTS

I will forever be indebted to Dr. Dorri Poppi at Draper Laboratory and to Professor Oli de Weck at MIT for their help and guidance. They gave me the freedom to bite off way more than I could chew and yet found some way to keep my mind mostly focused on what needed to be done to finish this monstrous project on time.

Professor Murman, thank you for all of your support, academic and otherwise.

Much thanks to Joel Schindall, a major technical contributor to the Globalstar satellite system and a wonderfully kind Professor of the Practice at MIT, for putting up with all of my questions – and sleep deprivation – as I started my journey toward understanding satellite communications.

Yet more gratitude goes to Raymond Leopold, one of the founders of the Iridium satellite system, for taking time out of his busy schedule to tell me more about Iridium than I believe one can find in the literature. The groundbreaking technological achievement that is Iridium is an inspiration to those of us who design satellites and/or are geeky enough to know enough about Iridium to appreciate its beauty.

More thanks go to Phil Lin for his much-needed advice on simulating routing; and to everyone at Draper who has made my stay there an enjoyable one, especially Tom Kostas and Dorri who have made me feel like I was part of a family.

Of course, I can't forget to thank all of my friends who helped make this happen: Nick F. for his guidance and poking early on; Nirav S. for his always brilliant insights; Anna for being such a good friend and supporter throughout; Eli for insisting, even when I didn't believe it, that my research was worthwhile; and all the myriad of other friends who I may have forgotten. Last, but certainly far from least, my dear Viet, for all of your patience, editing and fact-finding assistance, and really good coffee. For all the times I had to choose thesis over spending time with you, thank you for putting up with me. You kept me going in the darkest of times and brightened my life even when I did not have one. You stubbornly made sure I would finish and for that I am eternally grateful.

I must thank my parents, John and Linda, for working so hard and sacrificing so much to enable me to make something of myself. When I was still crawling around our small trailer in middle-of-nowhere, Oklahoma, I bet you couldn't imagine how far your love and example would help me go. We faced a lot of obstacles together, from floods and earthquakes, to losing the farm, being forced to move to the strange land known as California, and the sudden loss of some of our dearest loved ones. Your strength in the face of adversity taught me so much about life and how to thrive in this rough and unpredictable world.

Mom: I can't imagine how hard it must have been juggling more than a full-time job, finishing your education, taking care of the household, and raising me and Elise...all at the same time, and managing it all with such grace and determination. You are amazing.

Dad: I couldn't have done this without you. You introduced me to science and technology, and brought it to life for me. You taught me to question everything, to be independent, and to do the right thing even if it's the unpopular thing to do. I love you.

Elise: Your turn. May your university education be fulfilling; you are an extremely capable, strong-willed, intelligent, and incredibly talented young lady. I was blessed to have you as my younger sister.

Finally, to Grandma, for believing in me and for all of the tools you gave me to succeed. I miss you so much.

Table of Contents

List of Definitions

DEFINITION 1: DISTRIBUTED SATELLITE SYSTEM:
A SYSTEM OF MANY SATELLITES DESIGNED TO OPERATE IN A COORDINATED WAY IN ORDER TO PERFORM SOME SPECIFIC FUNCTION.

DEFINITION 2: DISTRIBUTED SATELLITE COMMUNICATION SYSTEM:
A COLLECTION OF MULTIPLE SATELLITES INTERACTING IN A COORDINATED WAY TO PROVIDE COMMUNICATION SERVICES TO CUSTOMERS.

DEFINITION 3: SYSTEM:
AN INTEGRATED SET OF ELEMENTS TO ACCOMPLISHED A DEFINED OBJECTIVE. THESE INCLUDE HARDWARE, SOFTWARE, FIRMWARE, PEOPLE, INFORMATION, TECHNIQUES, FACILITIES, SERVICES, AND OTHER SUPPORT ELEMENTS.

DEFINITION 4: SYSTEMS ENGINEERING:
AN INTERDISCIPLINARY APPROACH AND MEANS TO ENABLE THE REALIZATION OF SUCCESSFUL SYSTEMS.

DEFINITION 5: NETWORK:
A SET OF THREE OR MORE INTERCONNECTED COMMUNICATING ENTITIES.

DEFINITION 6: SYSTEM KEYWORD RATIO (SKR):
THE NUMBER OF SYSTEMS KEYWORDS COUNTED IN A GIVEN PAPER DIVIDED BY THE TOTAL NUMBER OF KEYWORDS COUNTED IN THAT PAPER.

DEFINITION 7: SYSTEM PAPER RATIO (SPR):
THE NUMBER OF SYSTEMS PAPERS COUNTED CONTAINING A GIVEN KEYWORD DIVIDED BY THE TOTAL NUMBER OF PAPERS COUNTED CONTAINING THAT KEYWORD.

DEFINITION 8: PRODUCT DEVELOPMENT:
THE PROCESS OF TRANSFORMING CUSTOMER NEEDS INTO AN ECONOMICALLY VIABLE PRODUCT THAT SATISFIES THOSE NEEDS.

DEFINITION 9: COUPLED TASKS:
TASKS THAT DEPEND ON EACH OTHER FOR INPUT INFORMATION.

List of Figures

List of Tables

Acronym List

APR	Annual Percentage Rate
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
BOL	Beginning of Life
BPSK	Binary Phase Shift Keying
CDMA	Code Division Multiple Access
CPF	Cost Per Function
CRC	Cyclic Redundancy Check
CUM	Cost per User per Month
DLC	Data Link Control
DoD	Department of Defense
DSC	Distributed Satellite Communication
ELF	Extremely Low Frequency
EOL	End of Life
FCC	Federal Communications Commission
FDM	Frequency Division Multiplexing
GEO	Geostationary Earth Orbit
GTO	Geostationary Transfer Orbit
GNP	Gross National Product
ICO	Intermediate Circular Orbit
IDC	Initial Development Cost
IGRP	Interior Gateway Routing Protocol
INCOSE	International Council on Systems Engineering
ISO-OSI	International Standardization Organization/Open Systems Interconnection
LCC	Life Cycle Cost
LEO	Low Earth Orbit
LLC	Link Layer Control
MAC	Media Access Control
MEO	Medium Earth Orbit
MF-CDMA	Multiple Frequency – Code Division Multiple Access
MF-TDMA	Multiple Frequency – Time Division Multiple Access
PPP	Purchasing Power Parity
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RDT&E	Research, Development, Testing & Evaluation
RIP	Routing Information Protocol
RTD	Round Trip Delay
SKR	System Keyword Ratio
SMF	Satellite Market Fraction
SOS	Satellite over Satellite
SOTT	Small Optical Telecommunications Terminal
SPR	System Paper Ratio

SRP	Selective Repeat Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
TFU	Theoretical First Unit
TIA	Telecommunications Industry Association

Chapter 1

INTRODUCTION

1.1. Motivation

Humanity now exists in the midst of the fast-moving Information Age, a period of history characterized by fast travel and even faster information transfer. As data becomes seemingly more valuable than physical possessions, the introduction of exciting applications for communications services becomes ever more critical for the success – and in some cases, survival – of businesses and even nations.

While the majority of these innovations have occurred over cable and fiber, a number of the most socially significant have occurred due to the introduction of satellites. Terrestrial fiber and cable systems have a number of advantages, but the extent of their reach and the cost of installation – in terms of both capital and time – favor developed nations over more remote and less industrialized communities.

Satellite communication systems employ spacecraft that reflect or relay signals providing communications services to customers. An individual satellite can directly connect any two or more points visible to it, and points beyond can be connected by relaying through available terrestrial links, bouncing to a relay satellite ("bent-pipe" relaying), or transferring to an adjoining satellite via an inter-satellite link. For this reason, satellites are uniquely capable of providing communications services anywhere in the world, bridging political, physical, and even economic divides.

Even as satellites offer the only real chance for ultimate communications ubiquity and true global unity, there remains a significant cost-benefit barrier. Few commercial satellite systems have succeeded economically without first falling victim to bankruptcy. The upfront capital required to implement a satellite communications system is staggering, and historically satellite companies have failed to adequately match capacity and service options to the current and actual future demand.

Fortunately, hope for an industry revival remains. New deployment strategies are emerging to cut the necessary upfront capital and designers are becoming increasingly aware of the need for reconfigurability and real options. However, new technologies and design considerations are not the only things that drive the direction of the cost-benefit barrier.

The design process itself is an inherent limiting factor to the achievable cost and performance of a system. Traditionally, the first step toward designing satellite communication systems – as well as terrestrial, sensor web, and ad hoc networks – has been to specify the system topology (e.g., the orbits of the satellites and the locations of the ground stations) based on the desired market and then to design the network protocols to make the most of the available resources.

Such a sequential process assumes that the design of the network architecture (e.g., protocols, packet structure, etc) does not drive the design of the system architecture (e.g., constellation topology, spacecraft design, etc). This thesis will show that in the case of Ka-band distributed

satellite communication systems this fundamental assumption is not valid, and can have a significant impact on the success (cost, capacity, customer satisfaction) of the resulting satellite communication system. Furthermore, this thesis will show that how a designer values performance during the design and decision process can have a substantial impact on the quality of the design path taken through the trade space of possible joint architectures.

1.2. Why Distributed Satellite Communication Systems?

Two basic satellite system architectures can achieve a mission coverage objective. If the desired market or target zone is isolated to a single region or to a set of regions in close proximity to each other, then it is possible to use a single satellite to provide the necessary coverage and capabilities (Figure 1(a)). This type of system has universally been accomplished with a geostationary satellite; a satellite in geostationary earth orbit (GEO: 35,786 kilometer altitude) seems fixed in the sky with respect to a user on the ground. Nearly a third of the Earth's surface is visible to a single satellite at GEO.

On the other hand, if the desired market or target zone includes a set of regions that are not in close proximity to each other, or if the market is a global one, then a distributed satellite system is the way to go (see Figure 1(b)). As will be discussed shortly, a distributed system may be the best choice even for cases for which a single GEO satellite is an option.

Definition 1: Distributed Satellite System

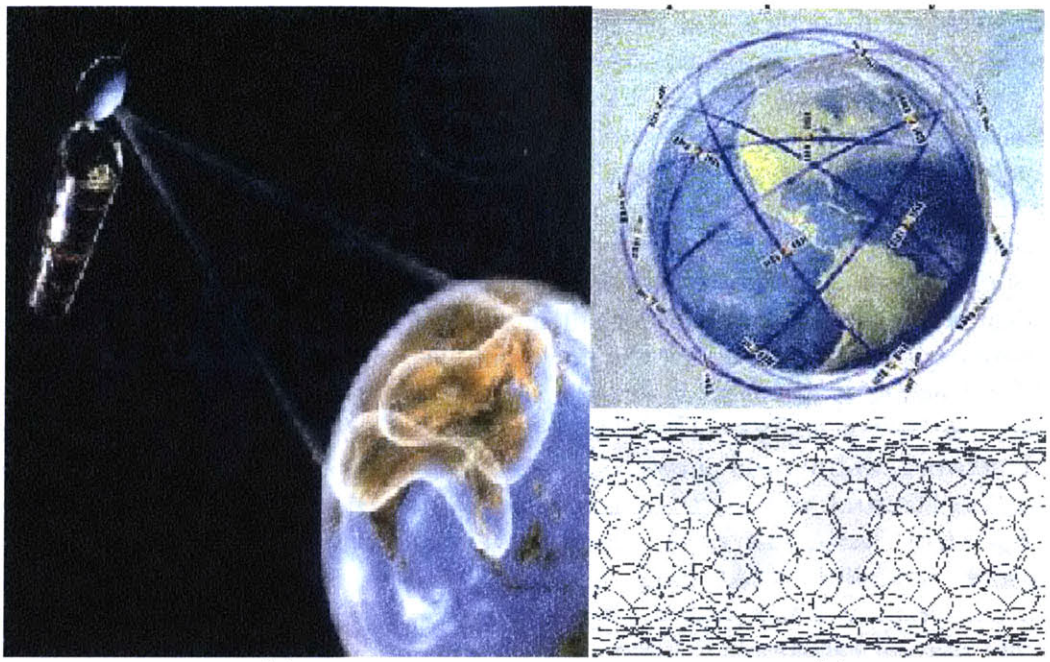
"A system of many satellites designed to operate in a coordinated way in order to perform some specific function." [Shaw99]

If the specific function of a distributed satellite system is to provide communication services, then the system is a distributed satellite communication (DSC) system.

Definition 2: Distributed Satellite Communication System

A collection of multiple satellites interacting in a coordinated way to provide communication services to customers.

This thesis focuses on DSC systems for two reasons. First, DSC systems lend themselves to far more interesting problems than single-satellite systems. Secondly, the world is entrenched in an era of fast-evolving markets requiring ever more demanding data services. This treacherous economic environment sparks a need for systems that can rapidly deploy new supporting infrastructure anywhere in the world while still being able to reconfigure and leverage off of existing assets. DSC systems have the potential to do this; single-satellite systems do not, nor do terrestrial cable and fiber optic systems.



(a) (b)

Figure 1: (a) GEO and (b) Distributed (LEO) Architecture and Footprints
If the desired market or target zone is isolated to a single region or to a set of regions in close proximity to each other, then it is possible to use a single GEO satellite to provide the necessary coverage and capabilities. On the other hand, if the desired market or target zone includes a set of regions that are not in close proximity to each other, or if the market is a global one, then a distributed satellite system is the way to go. The distributed LEO system is Iridium, which was successfully launched in under a year. The GEO and Iridium pictures taken from [Inma05] and [GEOS05].

1.2.1. DSC Advantages

DSC systems offer major advantages over terrestrial systems. DSC systems can provide global connectivity, data options for maritime and aircraft operations, natural broadcast and multiple access capabilities, isolation from local and regional disasters, improved system survivability, and improved staged deployment of network components. This section describes these advantages in detail.

Primarily, DSC systems can achieve global connectivity seemingly overnight (see Figure 1(b)). Global connectivity is the first step toward enabling complete communications ubiquity. It would be nearly impossible to realize this level of connectivity with terrestrial cable and fiber optic systems due to the cost, time, and expertise required to lay down the necessary infrastructure. After all, the current level of development only provides internet connectivity to an estimated 15% of the Earth's population [AMDn04], and in 2002 the global telephony penetration for developing countries was a paltry 28.1%, even including mobile subscribers – compare to an estimated 90% in the United States alone [WTDR03]!

Furthermore, satellites offer the only real data communications option for maritime and aircraft operations. Minimal communication is possible without satellites – the U.S. Navy uses extremely-low frequency (ELF) channels to contact their submarines [HAAR05] – but the

throughput, power and antenna size requirements of such systems limit their use to one-way transmissions of short, pre-determined codes.

DSC systems enable natural broadcast (Figure 2(a)) and multiple access (Figure 2(b)) capabilities. When a satellite beams information down to the ground, it is automatically broadcasting to some portion of the Earth's surface area, even if it is meant for a single user. Likewise, anyone can transmit to the same satellite simultaneously, as long as they are all within the area visible to the satellite (the footprint). As a result, it is possible to provide two-way service to mobile users anywhere within the satellite's footprint; the service can move with the user! This particular advantage creates a challenge to efficiently and fairly provide access to the limited resources of the satellite uplink and downlink.

Another advantage is that DSC systems are isolated from local and regional disasters. This characteristic can be especially beneficial to systems like the proposed Earth Science Enterprise sensor web network, whose main purpose is to detect, monitor, and predict weather, climate, and natural disasters [Torr02]. However, any communication system can benefit from isolation from disasters and other events that might disrupt terrestrial communication networks. A loss of communication capability could mean economic disaster for developed nations as their economies will grind to a halt within a matter of days, if not hours. While a single satellite system might prevent catastrophic communication meltdown for a locality or region, in some cases it might not provide sufficient connectivity to outside networks to bring backup networks online.

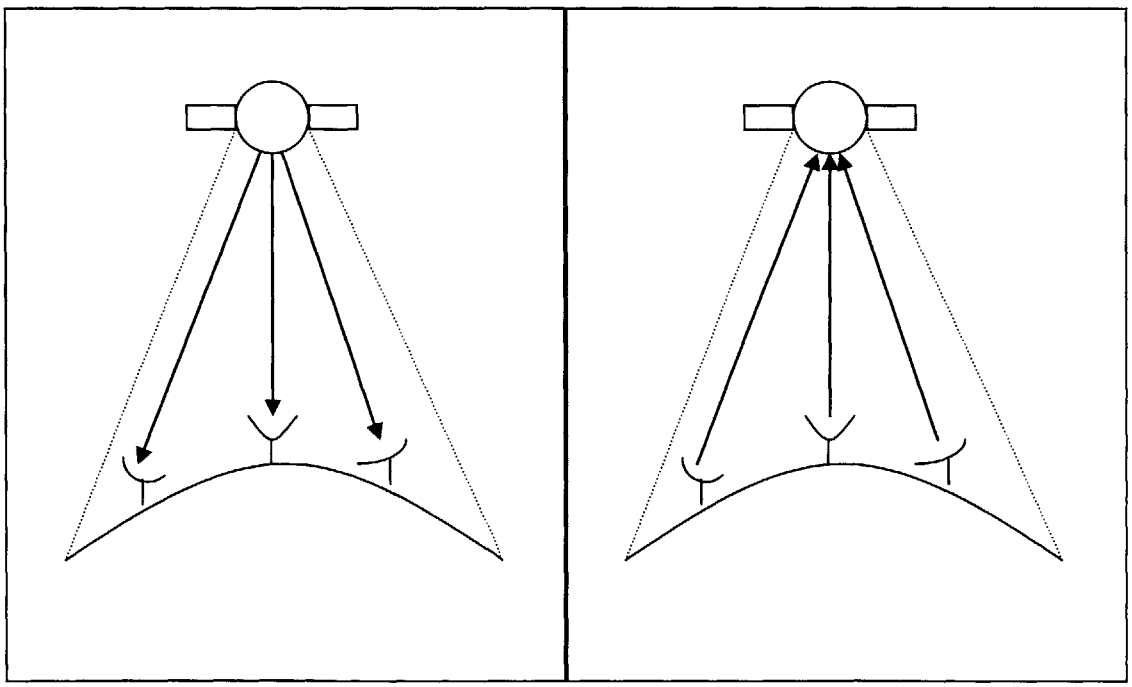
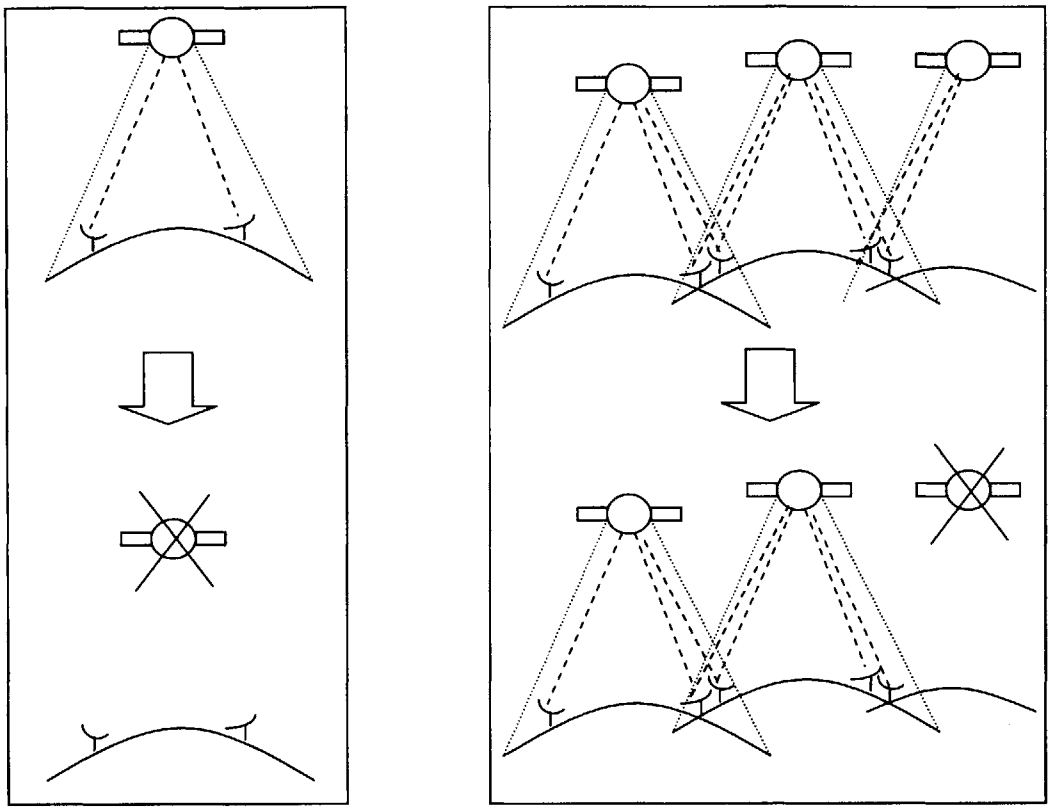


Figure 2: Satellite (a) Broadcast and (b) Multiple Access Capabilities
When a satellite beams information down to the ground, it is automatically broadcasting to some portion of the Earth's surface area, even if it is meant for a single user. Likewise, anyone can transmit to the same satellite simultaneously, as long as they are all within the area visible to the satellite (the footprint).



(a) (b)

Figure 3: Survivability for (a) GEO and (b) Distributed Communication Systems
Decentralizing resources has the effect of increasing the probability that the network and the system can still operate in the event of failures; non-distributed systems (GEO), on the other hand, fail if the centralized resource (i.e., satellite) fails.

A big plus for distributed satellite communication systems is system survivability [Shaw99]. Decentralizing resources has the effect of increasing the probability that the network and the system can still operate in the event of failures; non-distributed systems (see Figure 3(a)), on the other hand, fail if the centralized resource (i.e., satellite) fails. In a distributed system (see Figure 3(b)), if any given satellite is lost for any reason, there is a high likelihood that another pathway for communication exists through the system from a given source (say, the President of the United States) to a given destination (the Prime Minister of Great Britain, for example). The second path may not be as desirable as the first in terms of quality of service, but it still achieves availability requirements.

Another advantage inherent to satellite systems with distributed components is that the upfront investment in a system can be minimized by gradually launching satellites as they are needed. This feature is the basis for staged deployment, a real options investment strategy that will be discussed later in this chapter. As a result of staged deployment, the potential exists to better match incoming revenues with operating expenses by growing with demand (although, given the history of the LEO communication systems, this potential may not be realizable for many systems). A parallel can be drawn to terrestrial wireless mesh networks – a network in which each node in the network acts as a router, even those nodes operating as hosts for users. In

addition to the ability to grow with demand, wireless mesh networks experience increased communications reliability and network coverage as more nodes are installed in the system [JunJ03]. As we will see later on in this thesis, this holds true for distributed satellite communication systems as well.

1.2.2. DSC Disadvantages

DSC systems do have a number of disadvantages. A sufficiently large number of small satellites will cost more than a small number of large satellites; the savings implied by economy of scale may not occur since satellite systems cannot escape from the need for power and aperture on orbit. Large numbers of satellites increases the complexity of the network, increasing the complexity of the routing table and adding to the computational requirements each satellite must meet. The design of small satellites introduces its own challenges and disadvantages including increased vulnerability to failures and decreasing the achievable satellite lifetime.

In general, the more distributed the satellite constellation, the smaller and less complex the satellites [Shaw99]. This trend is liable to decrease the cost per satellite for a given system. However, just because the cost per satellite decreases doesn't mean that the overall cost of the system decreases. A distributed satellite communication system tends to require a lot of satellites, usually absorbing any reduction in cost achieved by reducing the cost per satellite.

Many applications have strict power and aperture requirements on orbit. For this reason, even having hundreds of satellites does not mean that there will be significant manufacturing savings from economy of scale [Shaw99].

As the number of satellites increases, the satellites themselves tend to become smaller and less complex; however, the same is not true of the network! The more satellites a system has, the more complex the network. As the network complexity increases, so do the protocols. The size and complexity of routing tables grow. This adds a computational expense to the satellite that isn't necessarily an issue in non-distributed GEO systems.

Furthermore, designing smaller satellites introduces its own problems. Small satellites cannot incorporate as much redundancy into the design as their larger counterparts, increasing vulnerability to single-point failures. Furthermore, small satellites have smaller payload and fuel capacity, decreasing the size and quantity of on-board instruments and transmitters, and reducing the satellite lifetime [Jill97].

This thesis will further justify the use of distributed systems by demonstrating their impact on performance in Chapter 4.

1.3. Literature Review

A significant amount of work has already been done to further our understanding of the design of distributed satellite communication systems. Figure 4 illustrates the relationships between various major components leading to the design of the overall system architecture. These components are either part of the system architecture or of the network architecture. Similarly, the distributed satellite communication system literature can be broken down into either systems or network literature.

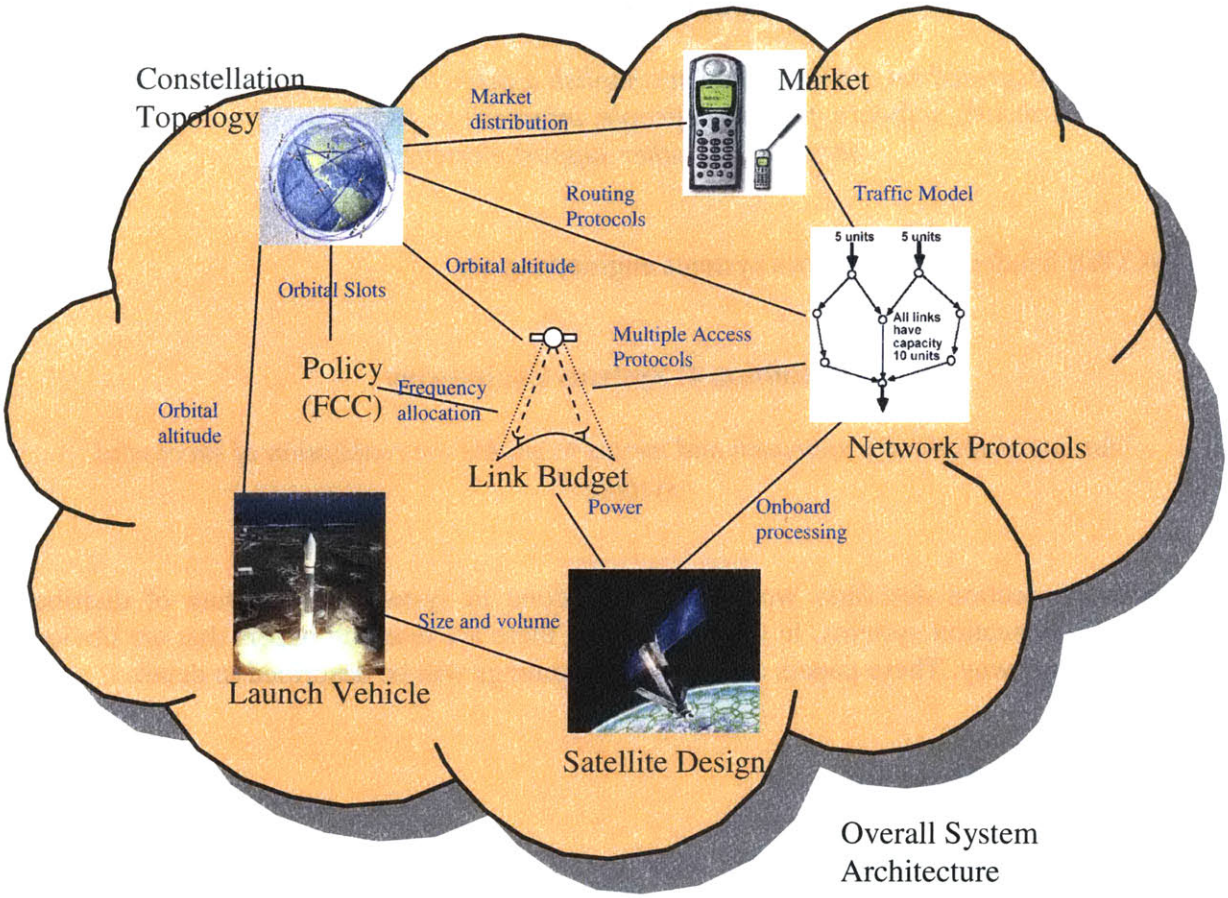


Figure 4: Overall System and Network Subsystem Architecture Relationship
This figure illustrates the relationships between various major components leading to the design of the overall system architecture. These components are either part of the system architecture or of the network architecture. The constellation topology, market, network protocol, launch vehicle and satellite design pictures taken from [GEOS05], [Modi04], [Sate05], and [Irid00].

The systems literature comprises research into satellite communications market modeling, high-level architectural trade analysis, optimal constellation design, spacecraft design, and real options.

The network literature is composed of research into satellite communication traffic modeling, traffic management, design of multiple access and routing protocols, and performance analysis.

1.3.1. Systems Literature

The International Council on Systems Engineering (INCOSE) Handbook [INCO98] defines a system as:

Definition 3: System

"An integrated set of elements to accomplish a defined objective. These include hardware, software, firmware, people, information, techniques, facilities, services, and other support elements."

The INCOSE handbook also defines systems engineering as:

Definition 4: Systems Engineering

"An interdisciplinary approach and means to enable the realization of successful systems."

The following section describes work previously done in systems engineering of distributed satellite communication systems. In some instances, there are research papers that are obviously deficient in some way. These papers and their shortcomings will be described in detail.

1.3.1.1. Market Modeling

In distributed satellite communication systems, a successful system is one that provides communication services to customers at a level that the customers find desirable. Market modeling enables system designers to understand the characteristics of the customers comprising the target market. This information is used to identify high-level functional requirements and to make initial constellation design decisions.

The market model should enable understanding of the past market trends, awareness of the current economic state of the industry, and identification of important indicators for predicting future trends. Haase, Christensen and Cate [Haas02] provides an example of this research for the global industry. Not all markets exist on a global scale. Hu and Sheriff evaluate the Satellite-Universal Mobile Telecommunications System (UMTS) terminal market in Europe [HuYF99]. Care must be taken since market predictions are notoriously difficult to get right. [Haas02] predicts strong growth in the satellite telephony services, which has yet to materialize; in fact, this market has generally failed to catch on in the midst of the cellular boom (Iridium and Globalstar, for example).

However, identification of the market and potential trends is not the whole story. The market model must be integrated into the high-level architecture trades done by the system designer. Kashitani incorporates a non-uniform market model in [Kash02]. Kashitani integrates the market model in order to explore how different distributed satellite communication system architectures succeed in matching system capacity with market demand. Kashitani's work is extended to propose hybrid architectures using elliptical orbits and multiple altitude layers to more efficiently allocated system capacity with market demand [Chan04].

1.3.1.2. Architecture Trades

It is important to understand the trade offs involved in complex systems with components that compete for resources to achieve a desired performance. Clearly, distributed satellite communication systems involve a great many trade-offs as the critical subsystems (e.g., communications, power, propulsion, etc) compete for extremely limited resources (e.g., mass, power, cost budgets, etc). Architecture trades analysis involves developing methodologies for evaluating the trade space of a complex system.

[Kash02] examines the trade-offs between the system architecture and the technical and economic performance as market demand is increased. This paper finds that LEO systems perform better in high-demand situations, while MEO systems perform better in low-demand scenarios. Evidently, there is a trade-off between the perceived potential market and the constellation architecture design; higher altitudes decrease the life cycle cost, while high demand market scenarios increase the achievable subscriber revenue for a fixed subscriber rate.

A methodology for conducting system architecture trades is presented in de Weck and Chang [deWe02]. In this particular instantiation of the methodology, the system life cycle cost of a LEO personal communication system is plotted against the system lifetime capacity, enabling analysis of the Pareto optimal front of non-dominated solutions for these two objectives.

One goal of this thesis is to understand the architectural trades between the system and network architectures in distributed satellite communication systems. Shaw [Shaw99] seems, on first glance, to analyze these relationships during the development of a generalized analysis methodology. However, the research falls significantly short of capturing the impact of the network on the overall architecture.

In [Shaw99], the satellite system is treated as an information collection and dissemination network, which is fine. However, the information network in the communication system case study is assumed to be comprised of "decoupled, parallel paths from the sources through a single satellite" to the destination; very little data is assumed to be routed through the inter-satellite links and thus they are ignored. These assumptions are extremely flawed. Even in bent-pipe satellite systems – which seem to be what is being modeled in this case – data transfers from various users are anything but decoupled. Notwithstanding the coupling inherent to the chosen routing protocol, the channel resource must be divided among all of the offered users, and this resource is finite and expensive. There is clear coupling between the data transfers of different users. Furthermore, the performance metrics used to distinguish architectures in [Shaw99]: Isolation, Information Rate, Integrity, and Availability, are all very much dependent on the network architecture. Assuming decoupled paths obscures all of the main interactions inherent in a communications network.

1.3.1.3. Constellation Design

The constellation design of a distributed satellite communication system has a huge impact on the performance and cost of the satellite system. The constellation design sets the topology of the network, which greatly impacts the design of the network protocols. It also specifies the altitude or set of altitudes, influencing the Federal Communications Commission (FCC) frequency allocations and the choice of launch vehicle. The lower the altitude, the more satellites are required to achieve full global coverage, thus increasing the system life cycle cost. Furthermore,

the constellation selection often defines the achievable system lifetime, the space environment (the van Allen radiation belts, for example), and the viewing geometry. Thus, understanding the impact of altitude and constellation design on the overall system architecture and performance is very important.

Some of the constellation design literature focuses on finding optimal constellations in terms of coverage and constellation patterns in order to minimize the number of satellites and to achieve some performance (redundancy, continuous coverage, etc). Rider [Ride85] and Adams [Adam87] find the optimal polar orbit constellation families in order to achieve redundant and continuous earth coverage above specified latitudes. On the other hand, Turner [Turner02] examines constellation designs using Walker patterns. The paper identifies "rules to assist in the rapid selection of Walker patterns." The rules are designed to minimize the number of orbital planes – to minimize the number of launches or to minimize the number of dedicated spares. The study is limited to low-altitude, circular and high-altitude elliptical constellations. Wertz and Larson [Wert99] includes a discussion on designing orbits and constellations "to meet the largest number of mission requirements at the least possible cost."

The affect of altitude on system capacity, user-to-user delay, power system design and offered communication services is examined in Gavish and Kalvenes [Gavi98]. The study examines interactions considered in this thesis, but like other similar research, it falls short on credibility and scope. While the study captures many of the physical interactions, including the Doppler effect, synchronization delay, power management considerations, and the satellite time in darkness, it fails to adequately account for the network architecture. As with [Shaw99], an implicit assumption is that data transfers are decoupled. The user-to-user delay calculations are based solely on propagation and switching delays, and it is unclear on what basis routing is conducted, other than an examination of the effect of space-based routing (user-sat-user) versus ground-based routing (user-sat-ground station-wire-user). This work seems to ignore any accounting of actual network decisions or protocols, and yet the paper examines the effect altitude has on performance metrics that are strongly related to the network architecture.

1.3.1.4. Spacecraft Design

The spacecraft design is important to the overall system architecture. Not only does the spacecraft design drive the choice of launch vehicle by specifying the volume and mass restrictions of the fairing, but it also strongly influences the link budget and network performance.

Spacecraft design is a complex process, and for the purposes of high-level spacecraft design and sizing, it is usually sufficient to make estimates on the mass and power requirements. These estimates, coupled with initial configurations, enable educated guesses about the size of the spacecraft. Wertz and Larson [Wert99] outlines a procedure for spacecraft design and sizing based on four decades of engineering design and well-defined techniques. More recently, a parametric model for communications spacecraft is developed in [Spri03] for use in trade studies performed in the early conceptual stages of design. This particular model is based on non-geostationary system data from 1990 to 1999.

1.3.1.5. Real Options

Real options characterize the decisions that a system designer can implement at various stages of the development or operation of a system. Real options give designers the ability to make changes to the system after its initial deployment, without obligating them to make a decision. Various real options strategies have been proposed, including constellation reconfigurability, staged deployment of systems, investment options, and technology portfolio management.

Constellation reconfigurability involves physically rearranging and adding and/or subtracting components to a constellation to improve some desired performance. Scialom [Scia03] proposes a method for converting an initial constellation with low capacity into a new constellation with higher capacity to take advantage of some perceived economic opportunity. As market demand increases, the capacity of a system can be increased by supplementing the existing system with additional satellites; this staged deployment Chaize [Chai03] can be done in conjunction with orbital reconfiguration. In [Chai03], paths of architectures are found in the architectural trade space rather than attempting to find optimal solutions for specific capacities.

Suzuki et. al. [Suzu03] examines the impact of new technologies on the Pareto-optimal frontier of the trade space of possible architectures. The methodology developed in [Suzu03] could be useful for analyzing exposure to the additional cost and development time required for implementation of various new technologies that provide some incremental performance and capacity improvement.

Once a system is in place, investment options become available that can leverage off of the real options discussed above. Bonds et. al. [Bond00] discusses the need to increase the amount of communications the Department of Defense (DoD) owns and leases to support warfighting. Current DoD systems will not meet the projected communications demand with dedicated military assets. The report examines the impact of buying out or leasing capacity on commercial wideband satellite systems versus buying dedicated military systems with commercial characteristics.

1.3.2. Network Literature

The network subsystem architecture is a critical piece of the overall system architecture for a distributed satellite communication system. The network must integrate with the system architecture in order to provide communication services to customers; the network provides the communication service capability. A communication network may be defined as follows:

Definition 5: Network

A set of three or more interconnected communicating entities.

Much literature has been produced concerning the design of the network architecture for distributed satellite communication systems. This section will provide an overview of the most relevant; literature that is obviously deficient in some way will be described in detail.

1.3.2.1. Traffic Modeling

Traffic models are derived from assumptions concerning the desired customer base, and are important to the design of the network architecture. The traffic models provide network engineers with some idea as to the required bandwidth, the traffic density, and the necessary end-to-end connectivity in the system.

Mohorcic et. al. ([Moho00] and [Moho03]) develop and use traffic models that consider the geographic distribution of traffic sources and destinations. The models account for traffic that is generated and received in the same geographical region (hot spot traffic) as well as for traffic that is generated in one region and transmitted to another (regional traffic). The asymmetric nature of this traffic is modeled; however, these papers do not model the time variance of the traffic.

1.3.2.2. Traffic Management

Once the projected traffic is modeled, it becomes necessary to determine how to manage the traffic loads observed by the network. Heavy load conditions can create instabilities in networks that provide alternate routing, creating the need to control the flow of information through the system. Instabilities occur when the traffic rate offered to a node or a link exceeds the rate that the node or link can process the traffic. Good traffic management can reduce the observed round-trip delay as well as the occurrence of dropped packets.

Most traffic management literature focuses on improving the performance of networks under heavy and overloaded conditions. In [Arul94], Arulambalam and Ansari enhance the performance of mesh-connected, circuit-switched satellite networks by reserving a fraction of the capacity of each link for direct-routed calls. The amount of capacity allocated is continually adapted based on the traffic load conditions. The paper finds that reserving capacity in near-overloaded conditions overcomes the instability that alternate routing scenarios cause in these systems.

1.3.2.3. Multiple Access Protocols

The communication link is a finite and expensive resource and multiple users will often request access to the same communication link at the same time. Multiple access protocols control user access to the channel; the goal of a multiple access protocol is to efficiently allocate resources among all of the users requesting access, and to do so as fairly as possible.

Time Division Multiple Access (TDMA) was one of the earliest incarnations of a multiple access scheme, standardized in 1990. Code Division Multiple Access (CDMA) was developed more recently, and was standardized by the Telecommunications Industry Association (TIA) in 1993. The standardized implementation of CDMA is detailed in Whipple [Whip93].

Although standardizations set forth by TIA describe how the various protocols should be implemented so that they are cross-compatible, Chang and de Weck [ChaD03] provides methods for calculating the basic capacity of satellite systems which use Multiple Frequency – Time Division Multiple Access (MF-TDMA) and Multiple Frequency – Code Division Multiple Access (MF-CDMA) protocols. The paper defines capacity as the number of simultaneous

duplex channels that a single satellite can support. [ChaD03] constructs models incorporating both the power and the bandwidth limits inherent to the communication channel.

1.3.2.4. Routing Protocols

In the network architecture, the routing protocols specify the metrics by which node in the network is to pass along traffic to other nodes. Some protocols determine that a particular route is good if it minimizes the number of nodes that the data must pass through before reaching its destination; others route along paths that have the least observed congestion.

There are many different methods by which to route data in a network, and a number of different ways to analyze routing performance. Some methods adapt to the changing network environment ([Moho00], [Moho02], and [Moho03]), while others route based on the traffic class (e.g., "real time" based on minimized delay, "high throughput" which maximizes data throughput, and "best effort" which has no specific requirements) [Svig02]. Often, routing studies attempt to optimize routing to achieve some overall network performance. Kukukates and Ersoy [Kucu03] endeavor to "minimize the maximum flow over a given set of shortest paths," while Werner et. al. [Wern97] routes by minimizing severe delay jitter in ATM-based systems.

In each of the previous examples, the routing strategies are optimized for a single constellation topology, which is usually based on existing or planned systems such as Celestri and Iridium. One routing paper considers adjusting the constellation topology to get benefits out of the network. The Satellite over Satellite (SOS) Network [LeeJ00] is a proposed hierarchical system architecture to solve problems involving performance issues common in traditional constellations. The results of this paper show that this novel architecture outperforms the traditional "flat" architectures in terms of quality of service metrics. In some sense, this paper is ground-breaking, providing an example of network architectural design driving the system architectural design. However, the paper misses the fundamental problem: the design process.

Wood [Wood01] examines the networking issues affecting satellite systems using complex constellation architectures. [Wood01] even analyzes the effect that the constellation topology has on end-to-end delays. This work is very important to understanding some of the interactions between the system and network architectures, but it does not go far enough. This research still fundamentally assumes that the constellation topology drives the network architecture, but does not ask what happens when the network architecture pushes back.

1.3.2.5. Performance Analysis

The term "performance analysis" refers to a methodological analysis of the quality of service performance of various network topologies, protocols, and traffic scenarios. One can think of performance analysis as being similar to trade studies undertaken by systems engineers, except applied to the network architecture.

Many papers covering performance analysis of "general" LEO satellite systems (Teledesic, Iridium, etc) focus on analyzing quality of service metrics such as the call blocking probability [Zaim02]; others focus on the mean message traversal, message delivery time, and the call loss probability [HuJi98]. Sometimes the papers focus on particular types of networks. Chotikapong et. al. [Chot00] examines the network architecture and performance of the TCP/IP and ATM

protocol suites over satellite. However, it appears that the results were abstracted from any actual satellite architectures; the results should thus be taken with a grain of salt.

Werner et. al. [Wern95] attempts to understand the basic design problems involved in LEO/ICO (Intermediate Circular Orbit) systems. Although this paper gives a good discussion of some of the design issues involved, it seems to fall short on describing the impact these design decisions have on the overall system performance. Furthermore, it does not really account for many of the complex interactions that occur. Again, it appears that much of the design is assumed to be decoupled.

1.4. Thesis Overview

Chapter 2 outlines observations made in the literature review and supports those conclusions with a lexical analysis study of the literature. Consideration of a model of the design process offers clues as to where to take these conclusions, leading to a thesis hypothesis and methodology.

Chapter 3 discusses the important design considerations that must be made from both the system and the network perspectives. This chapter also includes the development of the models and assumptions referenced in Chapters 4 and 5.

Chapter 4 begins by providing an example in which the Existence proof is satisfied. This chapter proves that the design of the system architecture drives the design of the network protocols and vice-versa. This chapter also explores the Significance proof and the insights that can be gleaned from the results. In this chapter, an argument will be made for a re-evaluation of the product development process for those satellite communication systems demonstrating strong coupling between the system and network architectures.

Chapter 5 concludes the thesis with remarks on the newfound motivation to re-evaluate the current product development process, a commentary on common assumptions found in research literature, and a brief discussion on the potential ramifications to other systems of interest.

1.5. Impact of Thesis

First, this thesis will provide an overview of the underlying first-order interactions between the system and network architectures and their importance to the overall design of distributed satellite communication systems. Chapters 2 and 3 examine many of the important aspects driving the design of these systems.

Second, this thesis identifies invalid assumptions commonly made by researchers who study distributed satellite communication systems.

Third, this thesis initiates an exploration of the important couplings impacting the design process and the design decisions on both sides of the system/network interface. For example, it is extremely important for a systems designer in charge of picking design parameter x to understand how that decision modifies or limits the choices of the other parameters. Perhaps it will be necessary to talk to the network engineer in charge of choosing design parameter z to mitigate the impact his or her decision has on parameter x. The investor paying for functional requirement y, the performance of which is determined by design parameter x and z, will want to know what these interactions mean to the bottom line.

Finally, the results of this thesis may also extend to the design of terrestrial, sensor web, and ad hoc networks since the problem and design processes parallel each other. While this thesis is not specifically focused on these types of systems, this author recognizes the similarities and suggests researchers in these areas conduct their own research to determine whether they would benefit from studies similar to this thesis.

Chapter 2

OBSERVATIONS

2.1. Introduction

This chapter expands upon the observations made in Chapter 1. Some issues were observed in the distributed satellite communication literature involving the assumed interaction between the system and network architectures in DSC systems. A common theme on the system literature side was an assumption of decoupling between user data transfers in the satellite network; an assumption that was noted to be flawed in Section 1.3.1. Furthermore, very little accounting of the network was considered, even in papers examining the relationship between system cost and performance metrics that could be considered quality of service oriented. In the network literature, it was common to see very little accounting of the system architecture, other than to assume that it had already been designed. A few papers on the network side seemed to make the realization that the design of the network architecture could potentially drive the system architecture in terms of topology, but stopped short of identifying any underlying design issues.

Chapter 2 questions these fundamental assumptions and identifies the potential underlying design issues driving the assumptions commonly found in the literature. The chapter starts by motivating why it is important to ensure that the architectural design of these systems is done correctly. Then, a lexical analysis is performed on a subset of the available literature to confirm that there are deficiencies in the research; the analysis determines whether or not the literature considers the DSC systems as an integrated whole – as it should be – or as apparently decoupled components – as the observations made in Chapter 1 seem to suggest. A model of the design process is developed in Section 2.4, further confirming the observation that DSC systems are designed and evaluated as decoupled components, and enabling scrutiny of the potential impact this assumption may have. These observations culminate in Section 2.5 in the development of a hypothesis that the design of the system architecture and that of the network architectures are coupled, in spite of the direction the research literature seems to be taking. Finally, a methodology is proposed in Section 2.6 for proving or refuting the hypothesis.

2.2. Motivation

Despite all of the benefits distributed satellite systems provide, quality of service has been a long-running complaint. Quality of service measures the performance of the network in terms of the link quality that is observed by the customer. High levels of quality of service performance indicate that the network is doing its job: providing communications services to customers. As such, it should be clear that quality of service is strongly related to the network architecture.

A 2002 study funded by Globalstar LP compared telephone quality of service between the Iridium and Globalstar satellite networks. Iridium is a LEO distributed satellite communication system implemented in the late 1990s to provide global voice, data, fax, and paging services; the Iridium architecture uses polar orbits and inter-satellite links to provide complete global coverage. Globalstar is also a LEO distributed satellite communication system launched in the late 1990s; however, the Globalstar architecture uses Walker orbits rather than polar and does not utilize inter-satellite links. Globalstar provides telephony and data services to regions within the ±70° latitude belt.

For voice calls in rural environments (United States), the Globalstar LP quality of service study found that 84% of Globalstar calls were connected on the first attempt compared to 71% of Iridium calls, and 37% of Globalstar calls were dropped at some point during the call compared to 40.7% of Iridium's. In urban environments, dropped call rates nearly doubled for both services, at 64.5% for Globalstar and 70.4% for Iridium, although first attempt connection rate (83% and 74%, respectively) did not differ greatly [Fros02]. These numbers are likely due to the improved diversity of Globalstar – on average three satellites in view at any time over the US – over Iridium at single diversity. In the Globalstar network, an obstruction between a user and one or even two satellites will not prevent a call from occurring or drop a connected call; however, in urban environments, there are often many obstructions between the user and the satellite network, leading to the increased rate of call blocking and call dropping probabilities observed above. In rural environments, the average audio quality of Globalstar – on a scale from 1 unintelligible to 5 excellent – was 3.6 compared to 2.9 for Iridium; the average audio quality (3.2 and 2.5) did not differ greatly when calls were placed in urban areas [Fros02]. A possible explanation of these results is that Globalstar incorporates a variable rate vocoder which operates at an average rate of around 2.4 kbps, but can adapt to the network environment to provide a peak rate of 9.6 kbps; Iridium's vocoder is fixed rate. Enabling variable rate vocoding can significantly improve voice legibility.

If the cellular industry is any guide, network quality of service has a huge impact on customer retention. A 2003 CNN money article references statements made by the senior director of wireless services for J.D. Power and Associates, Kirk Parsons. According to the article, customers who implied they were switching carriers within the next year also reported that on average 19% of calls suffered from static or other forms of interference, while customers who indicated they were staying with the same provider claimed only 5% of calls were problematic. The overall average rate of static and interference problems, however, was 9%, hinting that higher than average rates of static and interference are a major factor in loss of customers [Vald03]. This trend just goes to show that the user is happy when the network is transparent.

It is interesting to note that the subjective audio quality metric for the satellite systems is similar to the measurement for the cellular systems. The 2002 study noted that Iridium calls suffered from noticeable signal degradation and time lag; entire words were being dropped from conversations and voices were slurred (the so-called "Iridium drawl"). Globalstar calls, on the other hand, sometimes exhibited an echo effect; phone users had to endure listening to themselves as they spoke. Considering that the study's audio quality scale indexed cell phone voice quality as 4.5, the grossly inferior scores achieved by both satellite phone networks make it unsurprising that satellite telephony has had extremely limited market penetration.

If there is one insight to be gleaned from these examples, it is this: if the network architecture fails to perform adequately then the system fails, regardless of how well the rest of the system is

designed. If the customer does not receive the quality of service expected – generally, this means a quality of service on par with other available services – then the network architecture has failed to perform its function. For this reason, it should be clear that in a distributed satellite communication system, the communications network architecture is a highly critical piece of the infrastructure.

2.3. Lexical Analysis

The literature review in Section 1.3 brings up important deficiencies in previous work on distributed satellite communication systems. Primarily, these deficiencies involve not accounting for the relationships between the overall system architecture and the network subsystem architecture. A good body of work has been done on both architectural components; however, very little has been done in the way of considering them as a truly integrated system. To demonstrate this phenomenon, this section details a lexical analysis study on twenty-four randomly selected distributed satellite communication conference papers.

The papers were sorted into two categories: satellite systems engineering or satellite network engineering. Only papers published in the five years previous to this study were considered. For a list of papers used in the study, please refer to Appendix A.

Lexical analysis is the study of words used in a piece of writing [Poli98]. For the purposes of this analysis, the frequency of pre-determined words (keywords) was analyzed. A set of keywords characterizing the systems engineering side were chosen as well as a set of keywords specifying the network engineering side. For a list of the keywords used, please refer to Appendix B. If a keyword (or any substantial variation thereof) appeared in a given paper, a '1' was notated; otherwise, a '0' was notated. Once this was done for all of the papers, the frequencies were tabulated according to the following definitions:

Definition 6: System Keyword Ratio (SKR)

The number of systems keywords counted in a given paper divided by the total number of keywords counted in that paper.

Definition 7: System Paper Ratio (SPR)

The number of Systems papers counted containing a given keyword divided by the total number of papers counted containing that keyword.

First, the System Keyword Ratio was calculated for each paper. A histogram of these ratios is shown in Figure 5. Each SKR value was traced back to the corresponding paper; it turns out that any paper with an SKR substantially greater than 0.5 correlated with a systems paper; any paper with an SKR significantly less than 0.5 referenced back to a network paper.

The distribution is clearly bi-modal. Thus, there does not seem to be any crossover in keyword distribution among the type of paper.

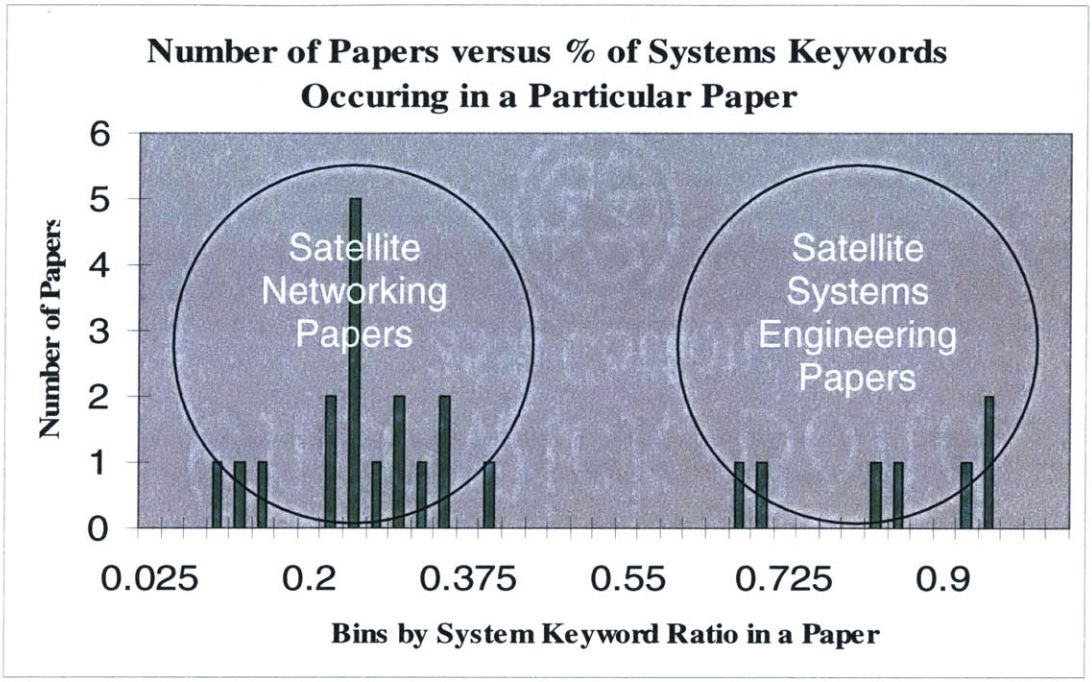


Figure 5: Histogram of the System Keyword Ratios for all Considered Papers
Each System Keyword Ratio (SKR) value was traced back to the corresponding paper; it turns out that any paper with an SKR substantially greater than 0.5 correlated with a systems paper; any paper with an SKR significantly less than 0.5 referenced back to a network paper. The distribution is clearly bi-modal. Thus, there does not seem to be any crossover in keyword distribution among the type of paper.

The second major analysis involves the System Paper Ratio. A histogram of these ratios is shown in Figure 6. Unlike with the SKR histogram, a noticeable area of crossover occurs between 25% and 75% SPR. Keywords that appear in this crossover area are:

Systems keywords:

- Architectures (54% SPR)
- Customers (50% SPR)
- Demand (33% SPR)
- Coverage (31% SPR)
- Performance (29% SPR)

Networking keywords

- Frequency reuse (50% SPR)
- Data rate (44% SPR)
- Bandwidth (35% SPR)
- Scalability (25% SPR)

The crossover terms are vague, general, and fail to convey any specific meaning.

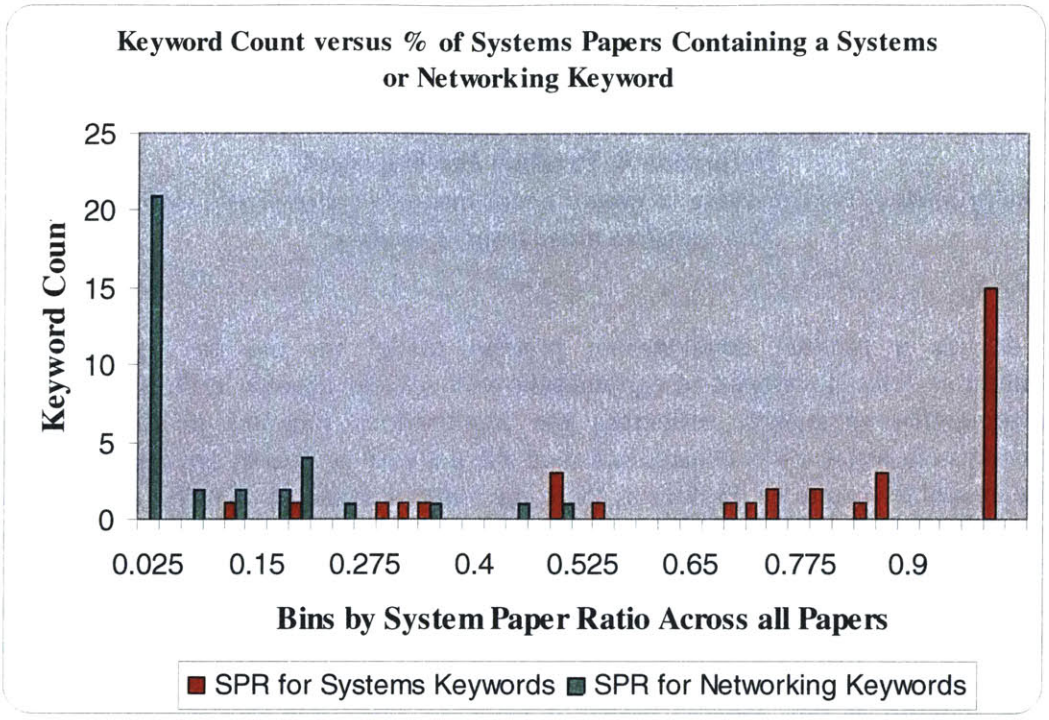


Figure 6: Histogram of the System Paper Ratios for all Considered Papers
Unlike with the SKR histogram, a noticeable area of crossover occurs between 25% and 75% System Paper Ratio (SPR). The crossover terms turn out to be vague, general, and fail to convey any specific meaning.

The lexical analysis supports the notion that there are significant deficiencies in the industry literature on these systems. This thesis will demonstrate that the two architectural components are indeed interrelated. The fact that there seems to be no substantial overlap in the literature is significant; a considerable chunk of the research underpinning the distributed satellite communication industry may contain a severe flaw.

2.4. Design Process Model

The potential importance of this apparent disconnect on the design process model is best summarized with the following quote from the International Council on Systems Engineering (INCOSE) Handbook [INCO98]:

> *"The need for a well-integrated approach to system design and development can be better appreciated when it is realized that approximately eighty to ninety percent of the development cost of a large system is predetermined by the time only five to ten percent of the development effort has been completed."*

The "development effort" mentioned above is best characterized by the product development process:

Definition 8: Product Development
"The process of transforming customer needs into an economically viable product that satisfies those needs [Jogl01]."

Figure 7 depicts a product development process model for use in distributed satellite communication systems developed in conjunction with several experts in the field. The process model is sequential in nature, reflecting the assumptions inherent in the literature; the constellation system topology is chosen and then the network protocols are created to make the most of the available resources. However, there are some indications that this design process is incomplete. For example, the footprint and link margin and power requirement inputs into the systems engineering process are potentially trades in the network design process. The system topology specified by the system engineering process is a major input into the network engineering process. And, as is clear in the literature, the system topology and the routing protocols are intimately connected since the traffic estimates and the topology drive the protocols. In looking at this process model it becomes clear that this design process model involves coupled tasks, even if this coupling is not incorporated into the process itself.

Definition 9: Coupled Tasks
Tasks that "depend on each other for input information" [Jogl01].

Concurrent Engineering is a general product development process that accounts for coupled tasks. Concurrent Engineering has two major applications: it can reduce the development time of a project without special consideration of the performance or quality of the product, or it can increase the quality and performance of the product for a given development time. There are risks associated with Concurrent Engineering; introducing coupling in the design process can result in increased communication between subsystems or an excessive number of iterations in designs between coupled subsystems. Both of these things can increase development lead time and cost. A good product development process seeks to reduce lead time, development and system cost, and improve product quality. Thus, there is an inherent trade-off between improving the overall performance and quality of the product without allowing any task to impose an unacceptable performance penalty on any other task [Jogl01].

Why should the distributed satellite communication industry consider concurrent engineering? Papers and presentations lead one to believe that systems such as Iridium and Globalstar were developed according to a process similar to the one shown in Figure 7. An enormous amount of attention was spent on the overall system topology for these systems. For example, Iridium aimed to provide complete global coverage, accomplished with inter-satellite links and polar orbits [Irid98]. As the motivation Section 2.1 shows, the network side still suffers. Teledesic seems to be a system that tried to drive the other direction. Teledesic started by specifying network characteristics that would put it on par, if not better, than competing

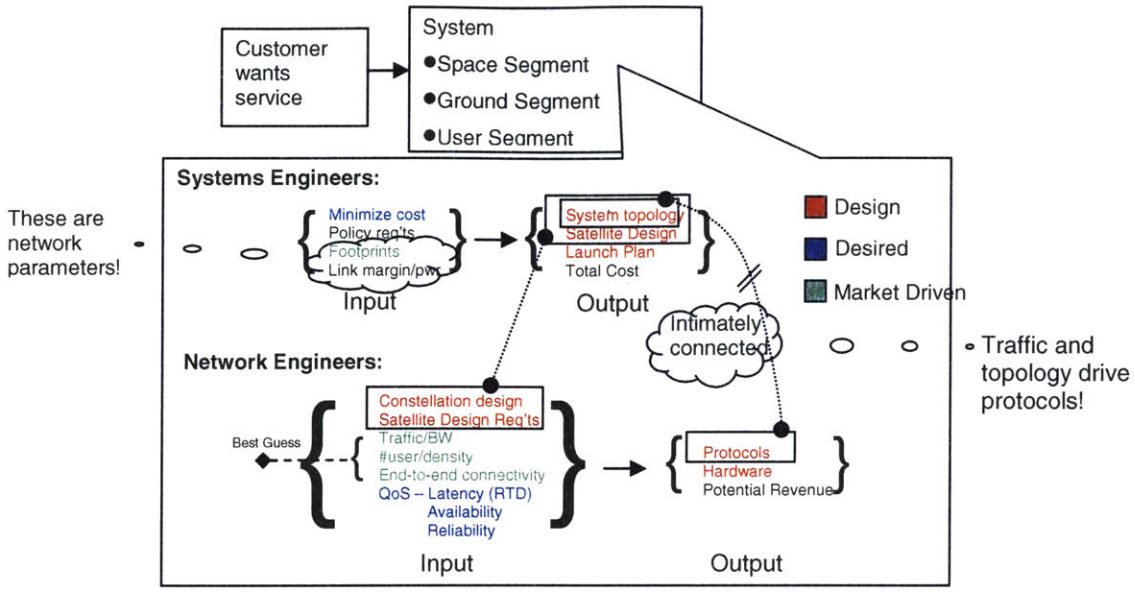


Figure 7: Example Design Process Model for DSC Systems
The process model is sequential in nature, reflecting the assumptions inherent in the literature; the constellation system topology is chosen and then the network protocols are created to make the most of the available resources. However, there are some indications that this design process is incomplete. For example, the footprint and link margin and power requirement inputs into the systems engineering process are potentially trades in the network design process. The system topology specified by the system engineering process is a major input into the network engineering process. And, as is clear in the literature, the system topology and the routing protocols are intimately connected since the traffic estimates and the topology drive the protocols. In looking at this process model it becomes clear that this design process model involves coupled tasks, even if this coupling is not incorporated into the process itself.

terrestrial systems. However, the original Teledesic design ended up with almost 900 satellites [Stur96]! Eventually, Teledesic shrank to a mere twenty some satellites and then disappeared altogether.

Given the history of satellites and of network engineering, it is unsurprising that a rift has developed between the two main architectural components of DSC systems. Although communication satellites have been around since the 1960s, network theory didn't emerge as a serious field of study until sometime in the early 1980s with the advent of the internet. The Big LEO systems like Iridium and Globalstar didn't surface until the mid-1990s, in the midst of the dotcom boom.

It is also highly desirable to avoid coupling in the design process as much as possible. Coupling introduces significant overhead to the cost of developing a system. Thus, it would have been easy to assume that the network would not introduce substantial coupling to DSC systems when previous non-distributed systems exhibited little to no coupling.

2.5. Hypothesis

Based on the observations discussed in section 1.2, the following hypothesis is made:

There exists a disconnection between the design of the system topology and the design of the network protocols in distributed satellite communication systems.

Prediction

The design of the network architecture strongly influences the design of the system topology – if this were allowed – for at least a subset of the possible designs.

2.6. Methodology

2.6.1. Distributed Satellite Communication Design Theorem

The hypothesis is based on the assertion that a disconnection exists between the design of the system topology and the design of the network protocols in distributed satellite communication systems. In order for such a division to exist, there must be some intrinsic interrelationship between the topology and the protocols. It should be apparent from the literature that the topology drives the design of the protocols; otherwise, papers developing optimal routing strategies would be considering more than one satellite constellation for their study. The fact that this situation does not appear in the literature is the reason for this thesis; hence, the hypothesis also suggests that the design of the protocols drives the design of the topology.

The Design Theorem uses this logic to translate the hypothesis into a form that can be examined in a methodical, scientific fashion.

Theorem 1: Distributed Satellite Communication Design

The design of the system architecture drives the design of the network architecture, and the design of the network architecture drives the design of the system architecture.

The following section outlines the proof of Theorem 1.

2.1.1. Proof of the Design Theorem

Proving and appreciating the implications of the Distributed Satellite Communication Design Theorem is at the heart of this thesis. To this end, the proof is broken up into two steps. The first step establishes the existence of the theorem (in other words, the design of the system and network architectures are shown to be coupled) while the second step evaluates whether or not the coupling ever becomes important to the end performance of the system.

2.1.1.1. *Proof of Existence*

One can prove an existence theorem using a constructive proof. In a constructive proof, the existence of an object with a given set of properties – usually mathematical in nature – is demonstrated by providing a method for creating said object [Cons05].

Since this Design theorem is applied rather than mathematical in nature, it is sufficient to generate – via a simulation model – a single example for which the assertion made in Theorem 1 is true. Chapter 3 enumerates the simulation model used for the proof and Chapter 4 provides the example required to prove existence.

2.1.1.2. *Proof of Significance*

Once existence is satisfied, the importance of the claim must be established. Just because coupling exists does not automatically mean that altering the design process of the network architecture will noticeably affect the design of the system architecture or the performance of the overall system. In this sense, the Significance proof only seeks to prove if the coupling is strong, weak, or depends on the design path.

If the coupling is strong, then allowing the design of the network architecture to push back on the system architecture will strongly influence the predicted performance of the entire satellite system. In this case, a product development process based on a sequential model – the topology is chosen and then the network protocols are created – is not sufficient to guarantee anything approaching an optimal design. A process model incorporating feedback will be more likely to find a design with a good trade-off between system and network requirements and performance.

If the coupling is weak, however, then allowing feedback from the network architecture will only weakly influence the predicted performance of the overall satellite system (if there is any influence at all). Thus, a process model incorporating feedback will not produce a better design than a sequential-based process.

A third possibility is that the strength of coupling depends on the design path taken. For example, strong coupling may only be apparent in systems with inter-satellite links, and not in systems using bent-pipe architectures, or in systems with the satellites in LEO versus GEO orbits. In this case, some types of distributed satellite systems should be designed incorporating feedback in the design process, and for others this will not be necessary.

The Significance proof ascertains the validity of the current communication satellite product development process as well as the various assumptions seemingly inherent to the industry's research literature. The Significance proof is constructed by exploring an extensive trade space of possible designs and analyzing the trends. This type of proof is related to computer-assisted proofs; most mathematicians consider these proofs to be valid. Chapter 3 (simulation model) and Chapter 4 (proofs) go into more detail.

2.6. Conclusions

This chapter has outlined the argument for the Distributed Satellite Communication Design Theorem. The fundamental assumptions found in the DSC literature have been questioned and the potential underlying design issues driving these assumptions have been identified. The

importance of ensuring that the architectural design of these systems is done correctly is motivated. A lexical analysis performed on a subset of the available literature has confirmed that there are deficiencies in the research; the analysis has further determined that the literature apparently considers the DSC systems as decoupled components, rather than as an integrated whole. A model of the design process further confirms the observation that DSC systems are designed and evaluated as decoupled components. A hypothesis has been made that the design of the system architecture drives the design of the network architecture and vice versa, contrary to the assumptions inherent in the research literature and the apparent design process.

Chapter 3

SIMULATION MODELS

3.1. Introduction

This thesis uses two simulation models to prove the Distributed Satellite Communication Design Theorem developed in Chapter 2: a basic model sufficient to prove existence and an advanced model designed to show significance. The basic model provides the core of the advanced one.

The Existence proof model and the corresponding performance metrics are discussed in Section 3.2. Section 3.3 provides the derivations supporting the advanced model. The metrics that will be used to show the significance of Theorem 1 are quite involved and thus are developed separately in Section 3.4.

3.2. Basic Model: Existence Proof

The Existence proof merely needs to show a single example for which the claim laid out in Chapter 2 is true. Thus, a zeroth-order model linking the satellite system constellation and the network protocols is constructed. A protocol fundamentally influenced by the system constellation (or topology) is the routing protocol: a methodology which discovers paths through the network enabling the transfer of data from a source to a destination.

Since this simulation model will also be used for the more intricate Significance proof, most of the fundamental model components will be of the appropriate complexity.

In this zeroth-order model, it is sufficient to model the satellite system constellation as a network topology, specifying the connections – communication links – between users at specific locations. The topology consists of communication nodes – satellites in orbit and users on the ground – and the communication links between these nodes. Section 3.2.1 details the constellation topologies considered for the existence proof.

Next, it is necessary to model the usage of the satellite constellation network by the customers. Section 3.2.2 develops the traffic model used in both the basic and advanced simulation models. The traffic model provides the link between the satellite system constellation and the network protocols; the routing protocols direct the transmission of user data through the topology on the basis of the traffic flow.

The routing protocols are discussed in Section 3.2.3, including the concepts of the adjacency matrix – indicating the logical structure of the network topology, and reachability – how the network determines whether any two nodes are connected, enabling the routing of data between this particular set of nodes.

Finally, Section 3.2.4 describes the performance metrics used to enable demonstration of the existence of the Distributed Satellite Communication Design Theorem.

3.2.1. Topologies

Topologies provide an understanding of the physical structure of the network that the traffic load will be transmitted through. The topology of the network specifies the connections – communication links – between users at specific locations. The topology consists of communication nodes and communication links between nodes. For the purposes of this model, only subnets – collections of users in a geographical area – rather than individual users are considered. At the most basic level, the topology guides the protocol to figure out how data packets can get from the source to the destination; one can think of the topology as a roadmap showing available routes between the city a given car is in now and the city the car desires to be in at some point in the near future.

Since the goal of the Existence proof is to prove the existence of Theorem 1, it is necessary to have a sufficiently rich network to demonstrate sensitivity to the interactions between the choice of routing protocol and the choice of constellation topology.

In order to keep the system and calculations as simple as possible, only GEO satellites are considered; thus, a first-order terrestrial network model is used to provide network richness. This terrestrial model is overlaid onto the grid of thirty latitude-longitude rectangles discussed in the Traffic Model section and is shown in Figure 8.

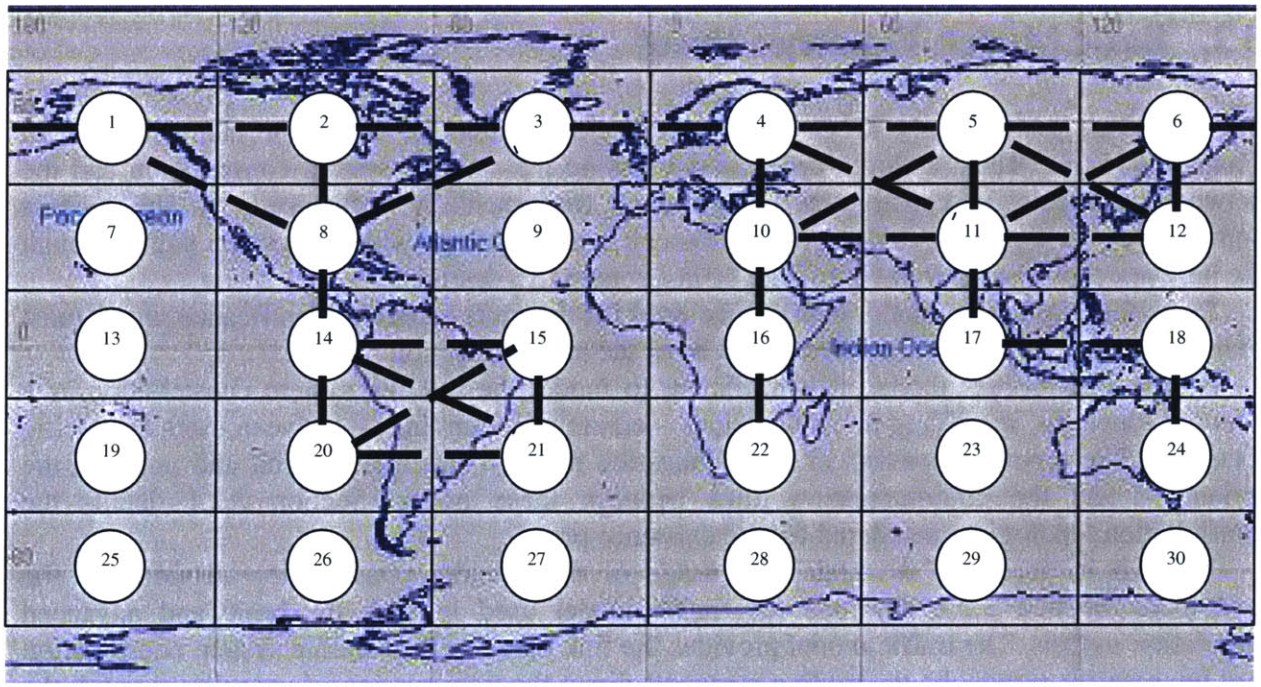


Figure 8: First-Order Terrestrial Model
In order to keep the system and calculations as simple as possible only GEO satellites are considered; thus a first-order terrestrial network model is used to provide network richness. This terrestrial model is overlaid onto the grid of thirty latitude-longitude rectangles discussed in the Traffic Model section.

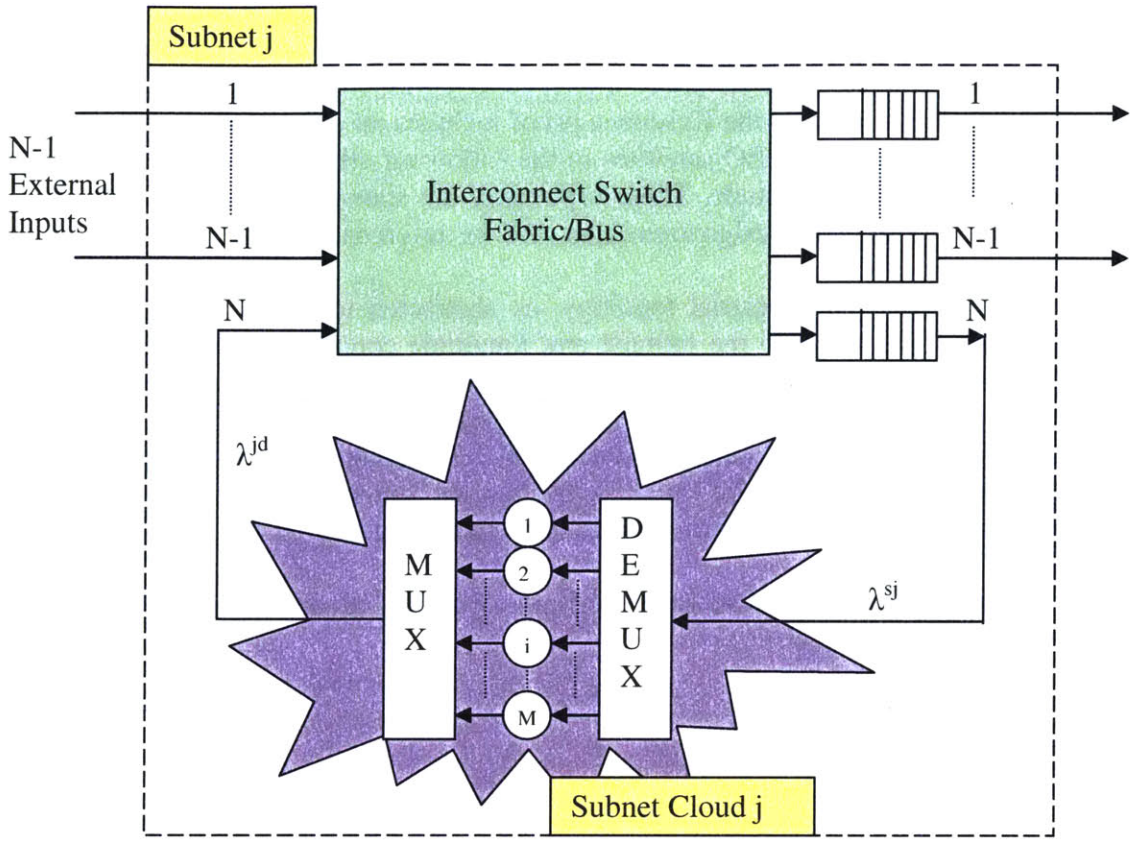


Figure 9: Model of Terrestrial Subnet

Each rectangle in the grid is assigned a terrestrial subnet, which captures all of the traffic flow occurring within that geographical area in the subnet cloud. Assuming there are M users, where M is arbitrarily large, each user will contribute some arrival rate of data to the subnet cloud. This contribution is portrayed as a multiplexing (MUX) operation. The multiplexing of arrival rates from all of the users in the subnet cloud results in the total arrival rate of packets to the subnet interconnect switch. The switch is assumed to be an output-buffered interconnect switch, which handles the routing to and from the other nodes in the network. Each subnet can process bi-directional data traffic along N independent links connected to the rest of the network. For the purposes of the Existence proof, the arrival rate λ^{jd} to each subnet from the subnet cloud is assumed to be 1. The Significance proof utilizes a model of the market demand to scale the arrival rates to each of the terrestrial subnets.

Each rectangle in the grid shown in Figure 8 is assigned to a terrestrial subnet (see Figure 9), which captures all of the traffic flow occurring within that geographical area in the subnet cloud. Assuming there are M users, where M is arbitrarily large, each user will contribute some arrival rate of data to the subnet cloud. This contribution is portrayed as a multiplexing (MUX) operation. The multiplexing of arrival rates from all of the users in the subnet cloud results in the total arrival rate of packets to the subnet interconnect switch. The switch is assumed to be an output-buffered interconnect switch, which handles the routing to and from the other nodes in the network. Each subnet can process bi-directional data traffic along N independent links connected to the rest of the network.

For the satellite-to-ground links, the simulation assumes that each such connecting link acts like a spot beam over the given geographical area and successfully multiplexes all of the traffic

between the given subnet and the visible satellite in a statistical fashion. It is assumed no packets are dropped due to congestion on the link nor are they lost due to packet errors. These assumptions will be relaxed in the advanced model covering the Significance proof.

The set of topologies used in the Existence proof includes all combinations consisting of the terrestrial network and up to 3 GEO satellites in the following orbital slots: 120° W. Longitude, 0° Longitude, and 120° E. Longitude. These topologies are shown in Figure 10 thru Figure 17 with their respective topology designations that will be referenced throughout the rest of this thesis.

Figure 10 illustrates the terrestrial topology, as abstracted to nodes and links alone. The topology is mapped according to the latitude and longitude coordinates of each subnet on the Earth; these coordinates are given in degrees. The Earth is assumed to be perfectly spherical; thus, the altitude of the terrestrial topology is zero kilometers.

The topology created by the addition of a single GEO satellite at 120 W. Longitude is shown in Figure 11. The coordinates of the satellite describe the sub-satellite location of the satellite over the Earth in degrees. The altitude is scaled such that the z-axis of the figure is 0.009 times the altitude at GEO in km.

Figure 12 through Figure 17 give the remaining topologies as the other satellites are added individually and in combination.

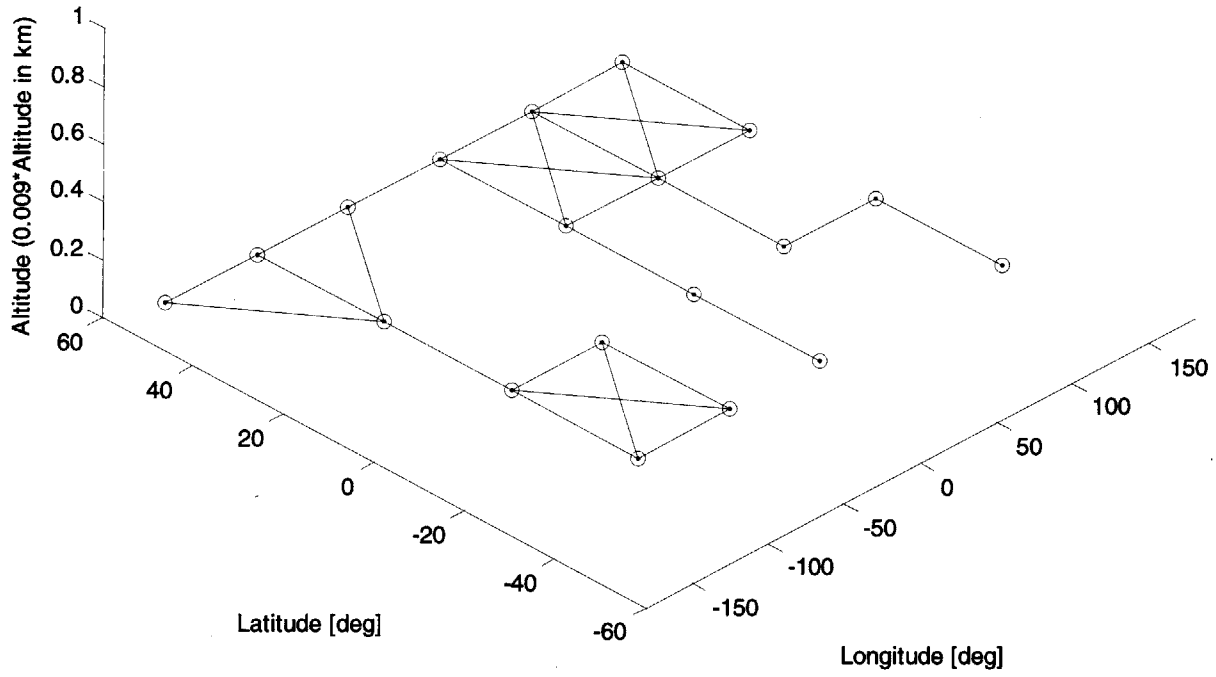


Figure 10: Topology 1: Terrestrial Topology
The terrestrial topology is shown above. The topology is mapped according to the latitude and longitude coordinates of each subnet on the Earth; these coordinates are given in degrees. The Earth is assumed to be perfectly spherical; thus, the altitude of the terrestrial topology is zero kilometers.The first-order terrestrial network model is used to provide network richness. Each terrestrial subnet (specified by a circle) captures all of the traffic flow occurring within that geographical area.

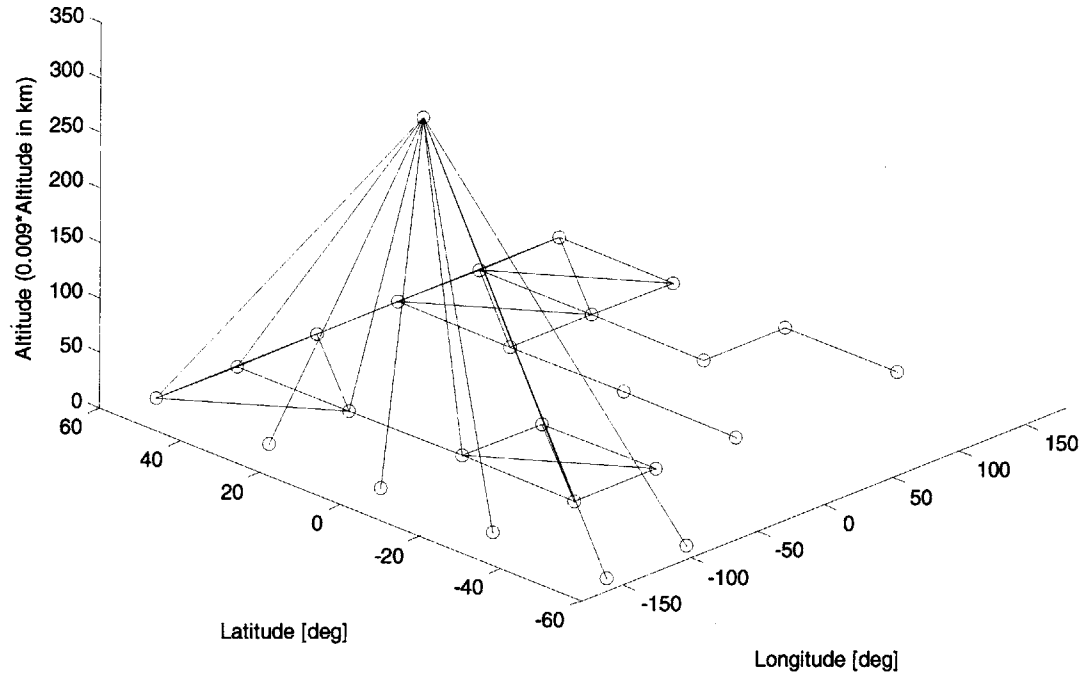


Figure 11: Topology 2: Terrestrial + 1 GEO (120° W. Long.)
In order to keep the system and calculations as simple as possible only GEO satellites are considered. For the satellite-to-ground links, the simulation assumes that each such link is like a spot beam over the given geographical area and successfully multiplexes all of the traffic between the given subnet and satellite. It is assumed no packets are dropped due to congestion on the link or lost due to packet errors.

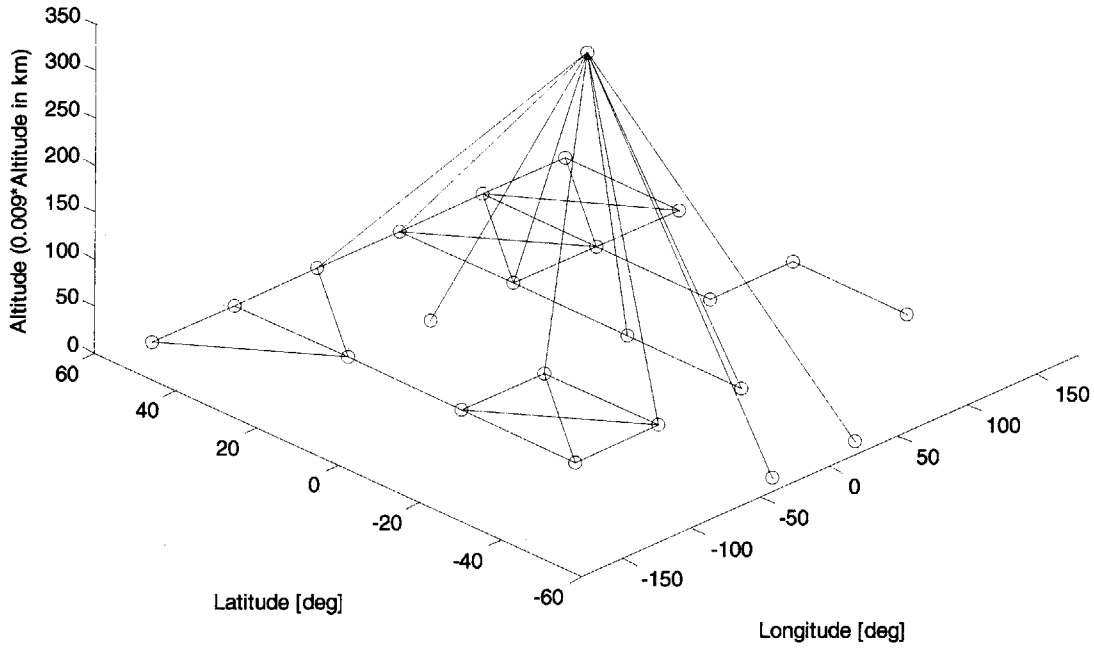


Figure 12: Topology 3: Terrestrial + 1 GEO (0° Long.)

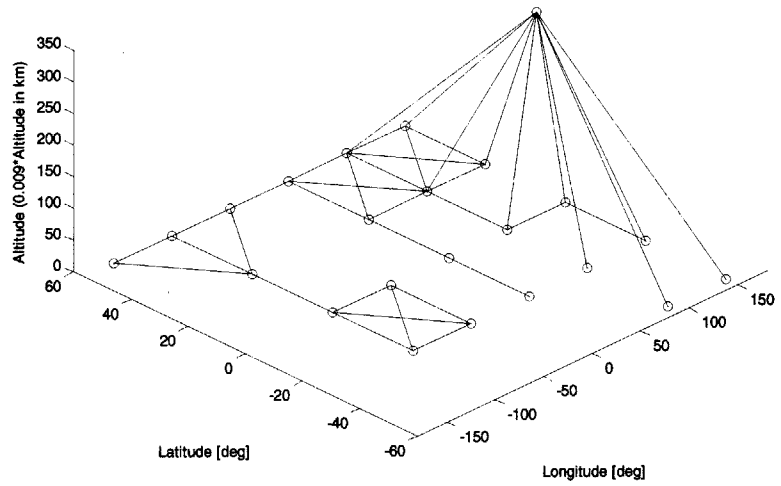


Figure 13: Topology 4: Terrestrial + 1 GEO (120° E. Long.)

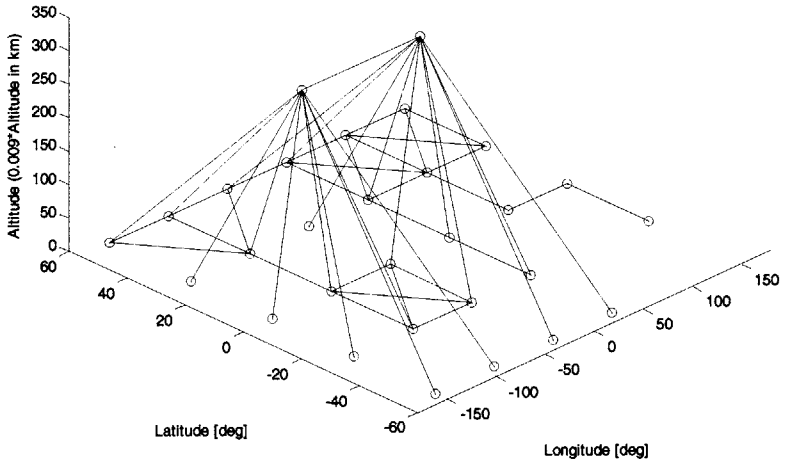


Figure 14: Topology 5: Terrestrial + 2 GEO (120° W. Long. and 0° Long.)

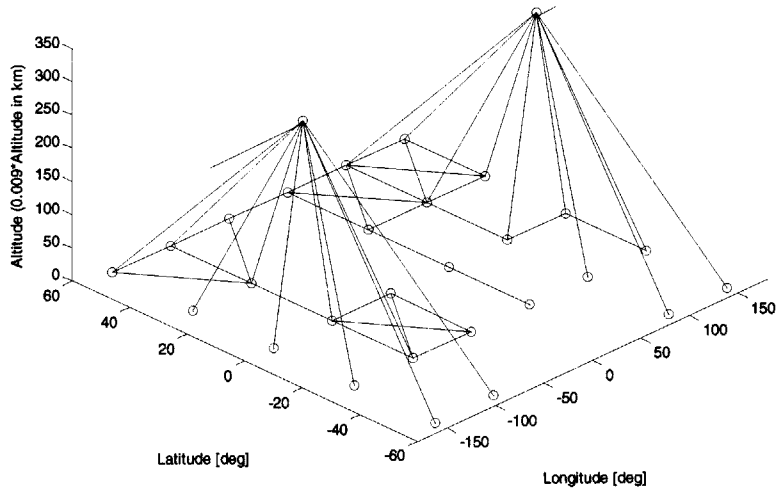


Figure 15: Topology 6: Terrestrial + 2 GEO (120° W. Long. and 120° E. Long.)

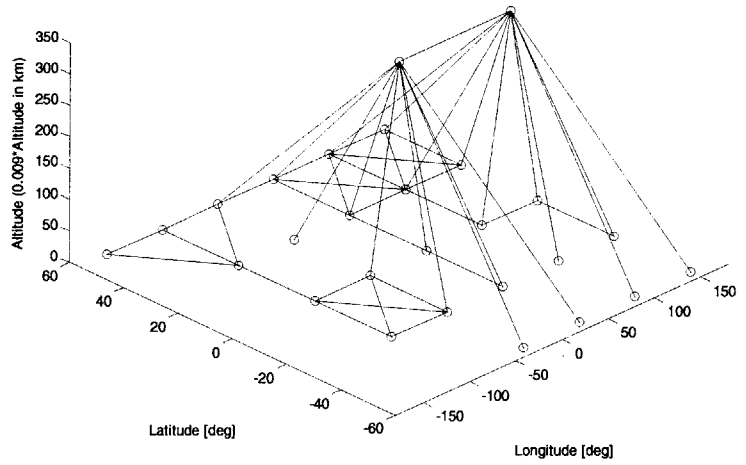


Figure 16: Topology 7: Terrestrial + 2 GEO (0 ° Long. and 120° E. Long.)

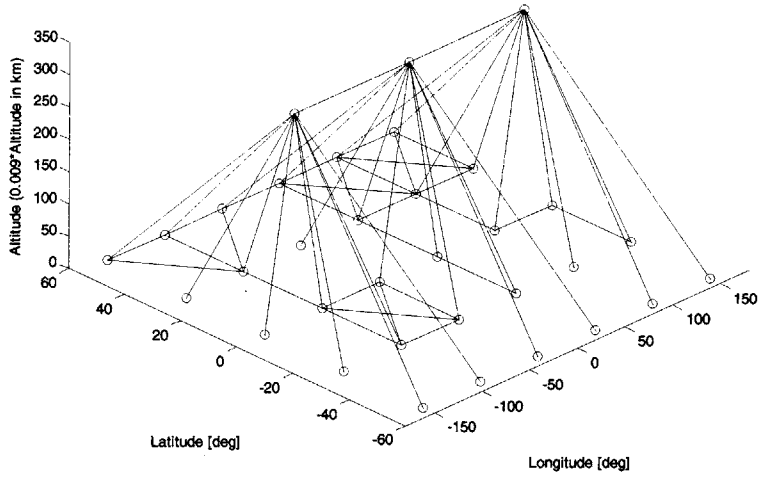


Figure 17: Topology 8: Terrestrial + 3 GEO (120° W. Long., 0° Long., and 120° E. Long.)

3.2.2. Traffic Model

To build a simulation model of this type, it is necessary to have some idea of the expected communications load in the system. This is done by building a traffic model – an estimated distribution of information transfer (such as packet data or voice circuits) in the network – the easiest of which is to assume a uniform traffic distribution; a zeroth-order model might assume a uniformly distributed global demand. As this traffic model will also be used in the Significance proof, a first-order model is required. Given that the distribution of demand across the globe – and even across a given region – is decidedly non-uniform, a reasonable first-order model would make an attempt at capturing relative concentrations and directivity of traffic. Nearly any traffic model can be justified as representing traffic allocated to a distributed satellite communication system; since traffic models are arguably arbitrary to proofs, any non-uniform market distribution should be sufficient

Fortunately, previous work has been done in the area of traffic flow dynamics. [Moho00] models the total traffic flow between source and destination regions using estimates of the

distribution of web servers (hot spot traffic) as well as estimations reflecting conditions of the telephone industry (regional traffic). Since this thesis is more interested in traffic flows on a regional basis, hot spot traffic will be ignored from now on. Table 3-1 provides the data on the percentage of total traffic flow between source and destination regions as given in this reference. The data deals in percentages, allowing the information to be used in probability form. Thus, this data provides the probability that a given data packet will travel to a given destination region from a given source region (assuming packets are generated independently).

The main issue with the data as is concerns the fact that most of the actual demand in these regions occurs on land rather than in the ocean. Figure 18 graphically depicts the regions as defined in the paper by [Moho00]; since the paper assumes the traffic to be uniform within each region, it is clear that a better construction of the regions is desired. Furthermore, this simulation model requires more geographical fine detail as the topology section demonstrates.

Fortunately, the data collected in [Moho00] can still be put to good use. Figure 19 shows the modified regional definitions, which contour themselves to follow the landmasses where the majority of the traffic is actually originating.

Table 3-1: Total Traffic Flow between Source and Destination Regions
Table given in percentages, or conditional probabilities (the probability that a given data packet will travel to a given destination region from a given source region). This traffic model attempts to capture relative concentrations and directivity of global traffic, based on the estimate reflecting conditions of the telephone industry. Table generated after [Moho00].

	Destination Region					
Source Region	N. America	S. America	Europe	Africa	Asia	Oceania
N. America	85	3	4	2	4	2
S. America	7	81	7	2	2	1
Europe	4	3	85	3	4	1
Africa	5	2	7	81	4	1
Asia	5	1	5	2	83	4
Oceania	5	1	2	1	7	84

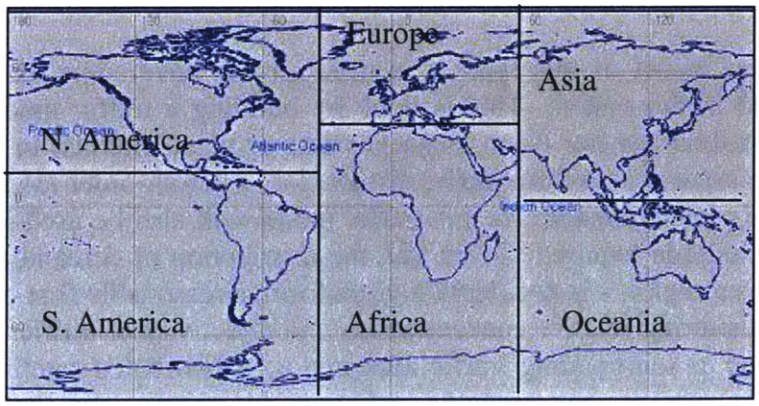


Figure 18: Definition of Geographical Regions as Given in Table 3-1
Picture created after [Moho00].

Table 3-2: Modified Total Traffic Flow between Source and Destination Regions
Table given in percentages, or conditional probabilities (the probability that a given data packet will travel to a given destination region from a given source region). To improve upon the model in Figure 18 and Table 3-1, estimates of the percentage of traffic flowing to and from the oceans are made and the remaining numbers are renormalized accordingly. Traffic is assumed to be distributed uniformly within each of these regions.

	Destination Region						
Source Region	N. America	S. America	Europe	Asia	Africa	Oceania	Ocean
N. America	83.3	2.9	3.9	3.9	2	2	2
S. America	6.8	79.4	6.8	2	2	1	2
Europe	3.9	2.9	82.5	3.9	2.9	1	2.9
Asia	4.9	1	4.9	81.3	2	3.9	2
Africa	4.9	2	6.9	4	80.2	1	1
Oceania	5	1	2	6.9	1	83.1	1
Ocean	35	5	25	30	2	2	1

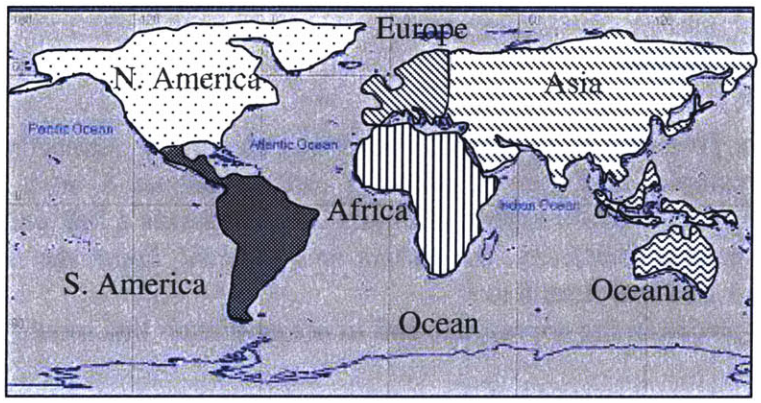


Figure 19: Definition of Modified Geographical Regions as Given in Table 3-2
The main issue with the data given in Table 3-2 and Figure 18 is the fact that most of the actual demand in these regions occurs on land rather than in the ocean. This modified traffic model differentiates the ocean traffic from that occurring in other, more populated regions.

To use the data in [Moho00], it is thus necessary to estimate the percentages of traffic flowing to and from the oceans that are internally accounted for in Table 3-1 and to renormalize accordingly. Some of the numbers for traffic flowing from the oceans appear to be quite high; it is assumed that regions supporting significant amounts of ocean and air traffic – North America, Europe, and Asia – will receive the majority of the calls from the ocean. The results of this process are documented in Table 3-2. Within each region, it is assumed that the traffic is uniformly distributed.

To reduce the computational expense of this simulation model, a grid of thirty 28° by 60° latitude-longitude rectangles were specified between ±70° latitude (the zone in which most of the world's population resides). Within each rectangle, the percentage of each geographical landmass (given as regions in Table 3-1and Table 3-2) was estimated to the nearest 1/16 of a latitude-longitude rectangle. This information was used to scale the percentages in Table 3-2 so

that the probability of destination given source in each source rectangle was properly normalized to account for the geographically-weighted probabilities.

To simply routing calculations, it was assumed the traffic could be specified as independent Poisson processes. Furthermore, the source arrival rates (λ^{jd}) to each subnet in Figure 8 were assumed to be equal to 1.

It is important to note that while traffic is time-varying, it was considered beyond the scope of this thesis to account for the impact of time of day (and thus time zones) or to be more precise in specifying usage patterns. While these do have an influence on the performance of routing protocols, they are not critical to proving the existence of Theorem 1.

3.2.3. Routing

In order to model routing successfully, it was necessary to create an adjacency matrix specifying the connectivity in the network and to employ rules for reachability – how to determine paths existing between the source and destination nodes in the network.

3.2.3.1. Adjacency Matrix

The adjacency matrix [Weis05] is a straightforward way to specify connectivity between nodes in a network. For example, in Figure 8, subnet 1 can see subnet 8, so a 1 is placed in the corresponding cell of the adjacency matrix in Figure 20; this means a link exists between subnet 1 and subnet 8. To enable directivity, the matrix is symmetric about the diagonal. This way traffic can flow from 1 to 8 and from 8 to 1.

Otherwise, a 0 is placed in the appropriate cell to notate that no link exists.

	Destination Subnet								
Source Subnet	1	2	3	4	5	6	7	8	9
1	0	1	0	0	0	1	0	1	0
2	1	0	1	0	0	0	0	1	0
3	0	1	0	1	0	0	0	1	0
4	0	0	1	0	1	0	0	0	0
5	0	0	0	1	0	1	0	0	0
6	1	0	0	0	1	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	1	1	1	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

Figure 20: Adjacency Matrix for the First 9 Subnets in the Terrestrial Model
The adjacency matrix is a straightforward way to specify connectivity between nodes in a network. For example, if subnet 1 can see subnet 8, then a 1 is placed in the corresponding cell of the adjacency matrix; this means a link exists between subnet 1 and subnet 8; otherwise, a 0 is placed in the appropriate cell to notate that no link exists.

3.2.3.2. Reachability

In order to generate a routing protocol, it is necessary to define reachability. The destination must be reachable by the source for data to travel from a source node to a destination node – a path must exist between the two nodes.

Rules:

1. If the source node and the destination node are the same, the packet stays within the subnet.
2. No node in the path may appear more than once. Otherwise, this may prevent the solution from terminating.
3. If, in a given number of hops, the source node cannot see the destination node, it is considered unreachable.

3.2.3.3. Routing Protocols

For this simulation model, the routing classification scheme used is fairly common. The routing protocols are differentiated by how the protocol defines the cost of traversing a link. The decision of which route to select is determined by which of the available paths through the network costs the least.

For simplicity, the algorithm chosen is the Floyd-Warshall [Bert92] algorithm. Floyd-Warshall is used to solve the all-pairs least-cost path routing problem. The algorithm assumes the topology is specified using a weighted, directed graph; negative weights are allowed, but not negative weight cycles (otherwise, the algorithm may not terminate). At a simplistic level, the algorithm operates by multiplying the adjacency matrix multiple times (in practice, however, this can easily violate the Reachability rules as a node in the path will appear more than once).

Floyd-Warshall acts as though the network uses perfect-knowledge distributed routing since each node, were it using this algorithm to compute its routing table, is aware of the state of every node and link in the network when it finds the least-cost path (in reality, this is a lot like centralized routing in terms of state knowledge except that in centralized routing, the routed path must find its way to the central hub – or network routing station – before it is routed out to its destination). Distributed routing algorithms usually have very limited information with which to route – as in, it knows its neighbors but not much else – but the node can send the traffic out without needing to go through a centralized hub. To first order, it is safe to assume a perfect-knowledge distributed routing scheme. More detailed studies should consider improving the fidelity of this assumption.

This simulation model examines two different routing protocols. The first (P1) uses normalized distance over a link as the link cost, while the second (P2) uses the number of hops as the link cost. The latter is known as the Routing Internet Protocol (RIP), a protocol introduced way back in networking infancy [InTH05].

Normalized Distance (P1

Equation 1 gives the link cost metric for routing protocol 1. Dividing by the maximum value of the link distance normalizes the link cost. Normalization ensures that each link has a cost on the same order of magnitude, thus guaranteeing fairness in the routing protocol decision.

$$LinkCost = \frac{Distance(i,j)}{\max(Distance(i,j))}$$

Equation 1

The distances are considered for each valid link (i,j), where i represents the source node and j the destination node.

Number of Hops (P2)

Equation 2 sets the link cost to 1 for all valid links (i,j), where i represents the source node and j the destination node. Using the adjacency matrix as the link cost guarantees that all of the links are valid.

$$LinkCost = Adjacency(i,j)$$

Equation 2

The routing cost is incremented by 1 for each link user data is transmitted over. A minimum cost path is thus one in which the number of links traversed – hops – is minimized.

3.2.4. Performance Metrics

The Existence proof considers three metrics which characterize the performance of the network:

- Minimizing the maximum number of hops
- Minimizing the congestion
- Maximizing the load balancing performance

This section discusses the reasons for, and derivations of, these performance metrics.

3.2.4.1. *Maximum Number of Hops*

A good simple performance metric is the maximum number of hops required to guarantee a path from any source to any destination, assuming such a path exists. An example of a path that does not exist would be a connection between subnet 1 and subnet 9 in Figure 8 if there is not a satellite overhead – as would be the case for Topology 1 and Topology 4. To zeroth-order,

minimizing the number of hops traversed by user data between a source and a destination will minimize the round-trip delay experienced by that data.

3.2.4.2. Congestion

One basic measure of congestion is to find the maximum offered load (λ_{max}) on any link in the network. To understand the meaning of offered load, consider the Poisson merging property illustrated in Figure 21(a). Suppose two communication links (1 and 2) feed into a subnet with one outgoing link. The arrival rates on links 1 and 2 (λ_1 and λ_2, respectively) merge together to form the offered load on the outgoing link (this load is a worst-case approximation since buffering occurs in the subnet, congestion leads to dropped packets, etc). The Poisson merging property tells us that this offered load is the sum of the arrival rates coming into it.

Likewise, since in any traffic stream user data is heading to outgoing link 1 with some probability P and to outgoing link 2 with some probability 1-P, according to the routing table and Table 3-2, we can calculate the "strength" of the arrival rate on each outgoing link by applying the Poisson splitting property (see Figure 21(b)).

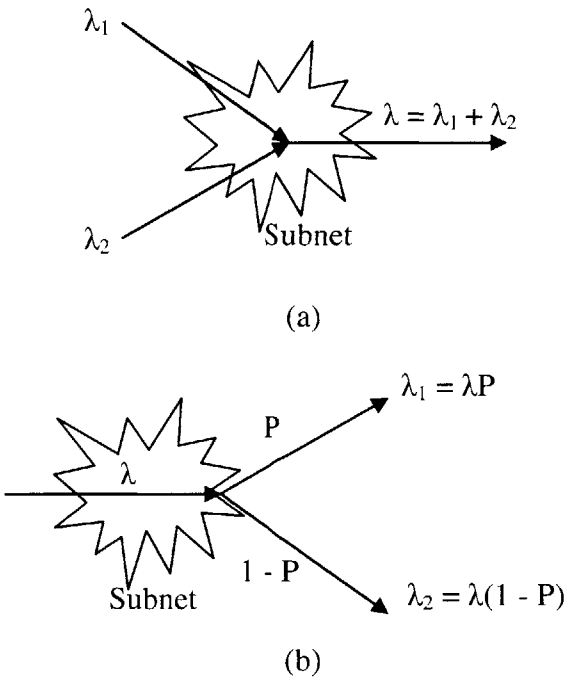


Figure 21: Example of (a) Poisson Merging and (b) Poisson Splitting Properties
(a) Poisson Merging Property: suppose two communication links (1 and 2) feed into a subnet with one outgoing link. The arrival rates on links 1 and 2 (λ_1 and λ_2, respectively) merge together to form the offered load on the outgoing link. The Poisson merging property tells us that this offered load is the sum of the arrival rates coming into it. (b) Poisson Splitting Property: some percentage of the total traffic in a traffic stream into the subnet is going to outgoing link 1 and some is going on outgoing link 2, and so on, according to the routing table and Table 3-2, we can calculate the "strength" of the arrival rate on each outgoing link according to the Poisson splitting property.

These observations tell us that our congestion calculations on all links with arrival rates λ_{ij}^{sd} from source-destination pair (s,d) on link (i,j) are subject to some constraints that enable their straightforward calculation. The total arrival rate of packets on link (i,j) is constrained by Equation 3:

$$\lambda_{ij} = \sum_{s,d} \lambda_{ij}^{sd} \text{ , for all } i,j$$

Equation 3

However, we also have to consider flow conservation at a subnet. The traffic that comes into a subnet must either go out on an outgoing link or be at its final destination within the subnet cloud (see Figure 9). Flow conservation at a subnet is defined in Equation 4:

$$\sum_j \lambda_{ij}^{sd} - \sum_j \lambda_{ji}^{sd} = \begin{cases} \lambda^{sd}, & \text{if } s = i \\ -\lambda^{sd}, & \text{if } d = i \\ 0, & \text{otherwise} \end{cases} \text{, for all } s, d, \text{ and } i,$$

Equation 4

where λ^{sd} is the arrival rate of packets at source s that are destined for destination d.

Equation 3 and Equation 4 [Rama96] can be used to calculate the offered load on every link in the network. In this case, congestion is defined as the maximum offered load on any link, such that Equation 5 can be expressed as:

$$\lambda_{ij} \leq \lambda_{\max} \text{ , for all } i,j$$

Equation 5

3.2.4.3. Load Balance

A network is load balanced if the traffic load is evenly distributed among all of the nodes in the network. In the real world, this is nearly impossible to accomplish, so it becomes necessary to specify a desired threshold value of congestion: simplistically, the greater the percentage of links ($LL_{\%}$) falling below the threshold, the better the load balancing.

A typical histogram of congestion from this simulation model is shown in Figure 22. The threshold value was arbitrarily chosen to be 0.5.

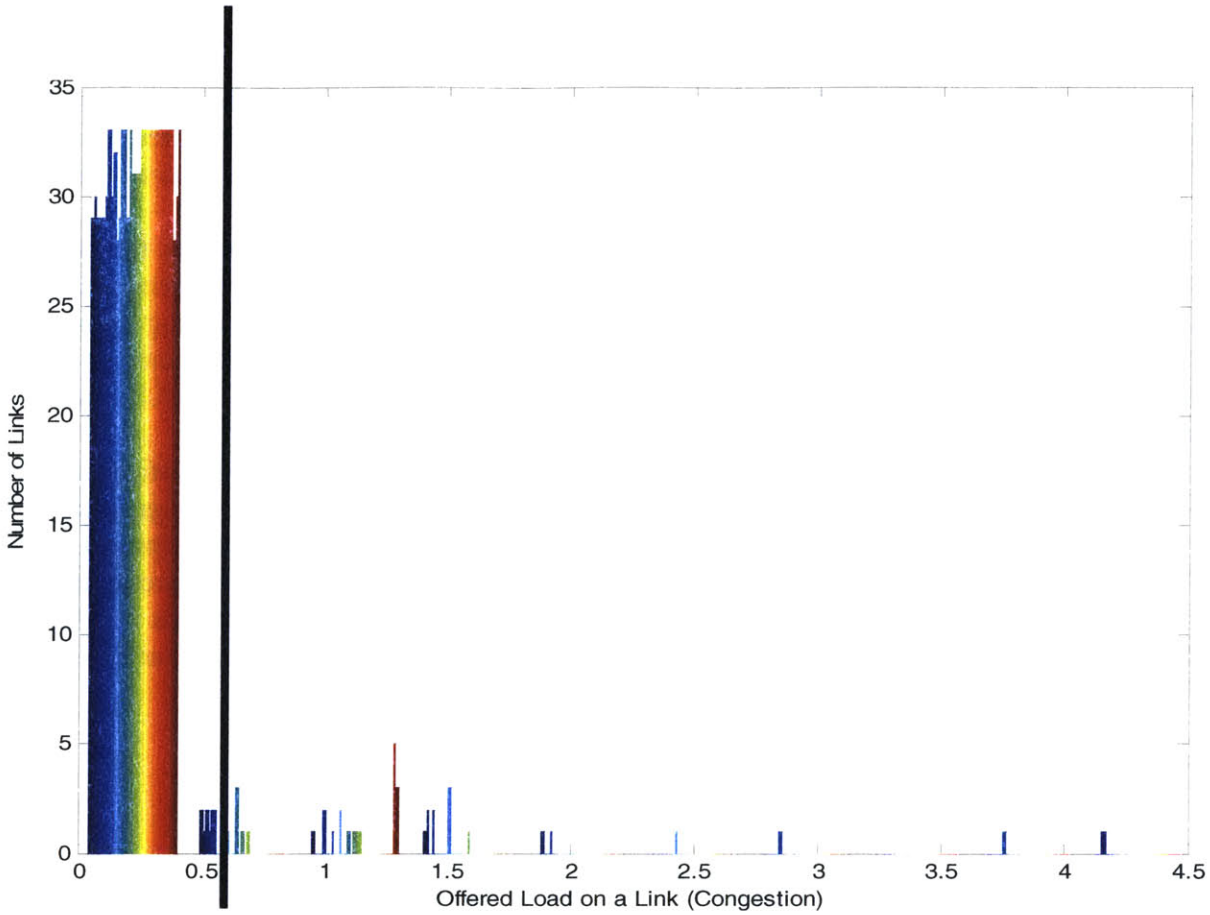


Figure 22: Typical Histogram of Congestion
A network is load balanced if the traffic load is evenly distributed among all of the nodes in the network. In the real world, this is nearly impossible to accomplish, so it becomes necessary to specify a desired threshold value of congestion: simplistically, the greater the percentage of links ($LL_{\%}$) falling below the threshold, the better the load balancing. The threshold value was chosen to be 0.5.

3.3. Advanced Model: Significance Proof

The Significance proof is necessary to understand the importance of the interactions captured in the Distributed Satellite Communication System Theorem. The Existence proof shows the truth of the fundamental claim in the Theorem, but gives no hint of whether allowing the design of the network protocols to influence the design of the system architecture is meaningful. The protocols may drive the design of the system architecture, but the degree to which it does this may be inconsequential, or an allowance in the design process for any significant feedback may incur more penalties than benefits. Therefore, the Significance proof exists to determine whether the coupling illustrated in the existence proof is weak, strong, or depends on the design path.

This section covers the design of the simulation model used for the Significance proof. Since this stage of the overall proof of the Theorem is to first-order approximation, the models described here are in general more complex and involved than those covered in Section 3.2.

This section begins with an overview of the performance metrics used in the Significance proof; the metrics described in Section 3.3.1 attempt to capture the valuation of system responses used by the system and network designers in the course of the design process. The design vector

discussion in Section 3.3.2 introduces some parts of the system that can be controlled by the designers; the values used in the design space analysis are provided. The system requirements in Section 3.3.3 describes overall simulation assumptions while the policy requirements in Section 3.3.4 captures aspects of the system that are decided by public policy.

Model development of the systems architecture begins in Section 3.3.5 with the combined Market-Traffic Model, which uses the estimated market data to scale the arrival rates of user data originating in the terrestrial subnets. A brief discussion of the cost modeling used in the simulation model is provided in Section 3.3.6; likewise, information on the spacecraft model, the launch model and the operations model can be found in Sections 3.3.7, 3.3.8, and 3.3.9, respectively.

Section 3.3.10 provides a description of the development of the constellation topology, including the overall constellation design as well as the topology and connectivity generation.

Model development of the network architecture starts with the development of the modulation scheme model in Section 3.3.11. Analytical models of the multiple access protocols are derived from [ChaD03] and [Modi04] in Section 3.3.12. Section 3.3.13 details the design of the link budget, while Sections 3.3.14 and 3.3.15 develop the routing protocol metrics and the accounting for the network overhead, respectively.

3.3.1. Simulation Objectives

Objectives are desired system responses. The following discussion outlines the reasons for choosing the objectives shown in Table 3-3 and provides general definitions for each objective as a performance metric; the actual derivations of the performance metrics will be discussed in Section 3.4.

The systems objectives include measures of performance that are important to systems designers: the monthly subscriber cost that a user can expect to pay (Cost/User/Month); the potential of the system to attract customers (Market Potential); the total cost of the system over its lifetime (Life Cycle Cost); the amount of capacity not being used by paying customers (Unused Capacity); and the number of users that the system can support simultaneously (Simultaneous Users).

Similarly, the network objectives summarize many of the performance issues that are important to network designers: the amount of billable services per unit of available spectrum (Spectral Efficiency); the amount of data lost in transit between the source and the destination nodes (Data Loss); the traffic load experienced by the network (Congestion); the ability of the network to distribute the traffic load as evenly as possible across all the nodes and links in the network (Load Balance); and the total time between when a given packet of data is queued at the source subnet and when it is received successfully at the destination subnet (Round-Trip Delay – Latency).

Table 3-3: Simulation Objectives
Objectives characterize the desired system responses. The table differentiates between the systems objectives and the network objectives.

Optimization	Systems Objective Function		Optimization	Network Objective Function
Min	Cost/User/Month		Max	Spectral Efficiency
Max	Market Potential		Max	Load Balance
Min	Life Cycle Cost		Min	Data Loss
Min	Unused Capacity		Min	Congestion
Max	Simultaneous Users		Min	Round-Trip Delay

3.3.1.1. Systems Objective Functions

Cost/User/Month

A large factor determining the success of a distributed satellite communication system is the cost to the customer. If this subscription cost is not comparable to the cost of a competing system, a given potential or current customer will likely choose the competitor. Thus, the systems designer should aim to minimize the cost per user per month. Since the metric is based on the expected total life cycle cost of the system, it represents the average cost per user per month (CUM) the customers will see. Also, this information can be used to determine the point at which the system will break even on the investment.

Market Potential

The market potential is the estimated global total subscriber base [Kash02]. This metric gives an idea of the prospective market capture in terms of the number of customers. Clearly, it is in the best interest of the company to design a system to maximize its market potential since a greater market potential increases the economic viability of the company for a fixed system cost.

Life Cycle Cost

The life cycle cost includes both the upfront capital investment for the initial deployment of the required permanent assets and the cost of maintaining the system on a day-to-day basis, which is considered recurrent cost. Minimizing this cost will reduce the amount of time and the number of customers required to break even on the investment. The life cycle cost for this thesis is calculated in thousands of 2005 dollars ($2005K).

Unused Capacity

Although maximizing the unused capacity of the system enables future expansion of the system – the ability of the system to absorb a new user increases with increased unused capacity as well as the ability to incorporate new services – the unused capacity represents the amount of system

capacity not being used by paying customers. Thus, an argument can be made that the amount of unused capacity should be minimized to improve system efficiency as well as to support the minimization of the user subscription cost; attracting uses will be very difficult if the subscription cost is prohibitively expensive.

There is a maximum arrival rate to a link that achieves link capacity (in other words, the link is fully utilized without dropping packets due to congestion). According to queuing theory, the system is stable for arrival rates less than the average service time.

The unused capacity of a link (Equation 6) is the difference between the number of packets that the link can support at capacity and the number of packets actually traversing the link. This definition includes the packets in queue at a node waiting to traverse the link.

$$UnusedCap_{Link} = SimPkts_{Link}^{Cap} - SimPkts_{Link}$$

Equation 6

The unused capacity of the system as a whole is simply the sum of the unused capacity over all of the links and nodes in the network. This relationship is captured in Equation 7.

$$UnusedCap = \sum_{i \in Links} SimPkts_i^{Cap} - SimPkts_i$$

Equation 7

Simultaneous Users

The number of simultaneous users the system can support is an important metric for system capacity in people terms. It is directly related to the total system capacity. There are two important measures of the number of simultaneous users: the number of simultaneous users that the system can support at capacity (the number of users fully utilizing all of the links in the network without dropping packets due to congestion – the ideal case), and the number of simultaneous users that are actually using the system at any given time (dependent on the number of subscribers determined by the market model), again without consideration of the packets dropped due to congestion.

For the purposes of this thesis, the number of simultaneous users will be measured at capacity. This quantity indicates the ideal user capacity of the system, while other measurements (congestion, load balance, etc) account for variations away from this ideal.

3.3.1.2. Network Objective Functions

Spectral Efficiency

Similar to the idea of cost/user/month, the spectral efficiency of the system is a measure of the amount of billable services that a wireless communication system carries per unit of available spectrum. The spectral efficiency is measured in terms of bits/second/Hertz/cell; design decisions

such as the choice of multiple access protocols, modulation methods, channel organization, resource reuse, and so forth, affect the spectral efficiency [Inte02].

Data Loss

Data Loss refers to the amount of data lost in transit between the source and the destination nodes. Clearly, it is in the best interest of the network designer to minimize the data loss as it will result in an increase in the congestion – data that is lost in the network must be retransmitted or the message is not successfully transmitted. Another way to specify the data loss is to treat it in terms of overhead efficiency, where the overhead measures the average amount of data – in addition to the user data – that must be transmitted over the system in order to transmit a single packet successfully. Thus, maximizing the overhead efficiency will minimize the data loss. It is assumed that the overhead includes data lost due to bit errors as well as due to dropped packets from congestion.

Congestion

Congestion provides a measure of the traffic load on a link, in the queue, or in the switch fabric of a node. If the congestion increases beyond the capacity of the link or node, then data packets are dropped – or lost – from the network. Hence, it is important to minimize the congestion in the network as much as possible.

For the purposes of this thesis, congestion will be measured in terms of the number of dropped packets. The more congestion a network experiences, the more packets are dropped to keep the network stable.

Load Balance

Load balancing refers to the ability of the network to distribute the traffic load as evenly as possible across all the nodes and links in the network. Thus, maximizing the load balancing of the network translates to maximizing the number of links in the network whose traffic load is below a certain threshold.

Links can be characterized as lightly loaded (less than or equal to 25% of capacity), medium loaded, or congested (packets dropped due to congestion).

This simulation model defines load balance as the percentage of links that are either lightly loaded or medium loaded in the network.

Round-Trip Delay (Latency)

The round-trip delay is the total time between when a given packet of data is queued at the source subnet and when it is received successfully at the destination subnet. Decreasing the average round-trip delay improves the quality of service seen by the customer, especially for latency-sensitive applications.

This simulation model defines the average round-trip delay to include the average extra time required for packet retransmission due to data loss.

Table 3-4: Significance Proof Simulation Design Vector
Design variables are aspects of the system that can be controlled by a designer. The table is broken down into design variables controlled by systems designers and design variables controlled by network designers.

Notation	Systems Design Variables	Units	Values
T	Orbital Period	[days]	1/9,1/5, 1
Tlife	Spacecraft Lifetime	[years]	5, 15
Emin	Minimum Elevation Angle	[deg]	15, 20
D_R	Receiver Diameter	[m]	0.05, 0.5
P_T	Transmitter Power	[kW]	4, 8
D_T	Transmitter Diameter	[m]	1.5, 3
TC	Terrestrial Capacity Real Option Flag	[-]	0 (Don't Buy), 1 (Buy)
Notation	**Network Design Variables**	**Units**	**Values**
ND	Network Routing Decision	[-]	0 (Centralized), 1 (Distributed)
RP	Routing Protocol	[-]	1 (RIP), 2 (IGRP)
MAP	Multiple Access Protocol	[-]	1 (MF-TDMA), 2 (MF-CDMA)
ARQ	ARQ Protocol	[-]	2 (Go-Back-N), 3 (SRP)
MS	Modulation Scheme	[-]	1 (BPSK), 2 (QPSK)
PS	Average Data Packet Size	[Bytes]	10, 100, 1000
Ruser	Average User Data Rate	[kbps]	10, 100, 1000

3.3.2. Simulation Design Vector

Design variables are aspects of the system that can be controlled by a designer. A rich set of design variables was selected for the significance proof; an effort was made to capture the most important variables that the system and network engineers have control over. Table 3-4 specifies the design vector used in the simulation model, including the values used in the trade space evaluation.

The following discussion defines each of the design variables and outlines the reasons for including each of them in the simulation design.

3.3.2.1. Systems Design Variables

Orbital Period

The orbital period is one way to specify the altitude of the satellite constellation. If the orbital period is specified in terms of an integer number of revolutions per integer number of days, then the satellite has a repeating ground track. One revolution per day corresponds to a GEO orbit. For k revolutions:

$$T = (1 \quad sidereal \quad day) \cdot \left(\frac{1440\,\text{min}}{day} \right) \Big/ (k \quad revolutions) \text{ [min]}$$

Equation 8

Although a sidereal day corresponds to 1,436.068167 min, this simulation assumes a value of 1440 min is sufficient. A satellite with repeating ground tracks has a subsatellite point on the Earth that repeatedly traces the same path [Wert99]. This attribute makes repeating ground tracks attractive for systems attempting to ensure precise and repeatable coverage of high-demand areas. For a simulation modeling the time-variance of a system, repeating ground tracks are especially useful since the system only need be modeled over the course of one day. This significantly reduces the computation required.

The altitude of the satellite constellation is perhaps the most significant design parameter under the control of the system designer. The altitude directly impacts the number of satellites required to meet coverage requirements specified by the desired market, as well as influencing the size and mass of the satellites and the choice of launch vehicle required to place the satellites in the specified orbit.

The choice of altitude also has a significant impact on the network performance of a system, placing strict bounds on the average round-trip-delay experienced by a user. Furthermore, an increase in the altitude reduces the number of satellites and hence the number of paths available to route packets. A reduction in the number of paths through the system introduces decreased reliability to the system as a failure in a single satellite transponder could potentially impact every single user in the system in terms of quality of service.

This simulation model will consider the effects of three altitudes corresponding to 1, 5, and 9 revolutions per day (1, 1/5, and 1/9 period [days]).

Spacecraft Lifetime

The lifetime of the spacecraft is another important design variable. The longer a spacecraft is designed to live, the longer it can support customers; this trend reduces the estimated cost per user per month by spreading the cost of system out over time. Similarly, the longer a system is in operation, the more customers it has an opportunity to attract.

On the other hand, a longer design life also has the effect of increasing the cost of the satellite. The satellite will require additional propellant to keep it on-orbit. Also, the required reliability and redundancy of the spacecraft increases since the probability of failure increases over time due to normal wear and tear. Guaranteeing a given probability of failure increases development and construction costs. Finally, longer design lives limit the adoption of new technologies; satellites systems can launch replacement satellites incorporating improved capabilities as the satellites expire.

This simulation model considers the effect of two lifetimes: 5 and 15 years.

Minimum Elevation

The elevation angle – or grazing angle as it is sometimes referred to – is important to the overall system architecture because it specifies the area observable to the satellite as the satellite travels

through its orbit. This means that the elevation angle has a large impact on the availability of the system.

A high elevation angle increases the probability of continuous coverage by reducing the impact of obstructions to the line-of-sight between the user and the satellite (trees and buildings, for example). However, the greater the elevation angle, the smaller the satellite footprint, increasing the number of satellites required to provide full global coverage, thereby increasing the overall cost of the system. Assuming full global coverage at a fixed level of diversity – the number of satellites required to be in view of a single point on the Earth at a given time – an increase in the elevation angle increases the availability of the system for a given user.

Similarly, the lower the elevation angle, the fewer satellites required to achieve full global coverage. Or for a fixed number of satellites, a decrease in the minimum elevation angle will increase the average diversity. An increase in diversity can offset the effect of line-of-sight obstructions since there are more paths available between a given user and a satellite in the network.

This thesis considers the situation in which the number of satellites required for full global coverage is calculated based on the minimum elevation angle for the case of single-satellite diversity. Thus, an increase in elevation angle can be associated with an increase in the availability but at a cost penalty to the system.

Two minimum elevation angles are considered: 15 and 20 degrees.

Receiver Diameter

The receiver diameter of the earth antenna specifies the size of the customer equipment. Normally, the receiver diameter is not considered in conjunction with the satellite transmitter diameter and transmitter power since the link budget can be sized to achieve a desired signal-to-noise ratio. The process just described would be undertaken for a system trade study aimed at finding optimal designs for a given desired market. This thesis, however, is more concerned with understanding the high-level interactions between the system and network architectures; a potentially key contributor to these interactions is the market.

The market of interest drives the size of the receiver diameter. A satellite system geared toward telephony services would do well to keep the size of the receiver diameter (here, assuming parabolic antennas) to around 0.05 meters (~2 inches), on par with cellular phones, or many customers will balk and go elsewhere for their telephony needs. Likewise, a system aimed at penetrating the broadband internet market could get away with distributing user antenna dishes on the order of 0.5 meters (~18 inches). If a broadband internet system could achieve the necessary quality of service with receiver antennas on the order of 0.05 meters, then new applications such as satellite internet on handheld devices could be enabled.

For the reasons given above, the significance model examines the impact of receiver diameters of 0.05 and 0.5 meters.

Transmitter Power

The satellite transmitter power has a significant impact on the sizing of the dry mass of the spacecraft, and contributes substantially to the signal-to-noise ratio per link achievable by the satellite system. As the transmitter power is increased, the demand on the power system of the satellite likewise grows, increasing the battery requirements as well as the requirements levied on

the solar panels. The transmitter power is important to the link budget of the system since the transmitter power is divided among all of the channels coming into and out of the satellite. If the satellite transmitter power is insufficient, then either the desired number of channels cannot be supported for a given required signal-to-noise ratio, or the quality of the transmission decreases for a given number of channels.

This thesis examines the impact of varying the satellite transmitter power by considering 4 and 8 kilowatt transmitters.

Transmitter Diameter

The sizing of the satellite transmitter diameter (here, assuming parabolic antennas) is the last key component to determining the quality of the link attainable between the satellite and the ground. The transmitter diameter is also important for sizing the spacecraft.

The satellite antenna diameter in combination with the frequency allocation provided by the FCC directly specifies the beamwidth of the satellite beam; the antenna size also limits the gain achievable at all points in the footprint. The beamwidth of the satellite beam limits the footprint, or area visible to the satellite on the ground.

This thesis considers two satellite antenna diameters: 1.5 and 3 meters.

Terrestrial Real Option

The Terrestrial real option captures the design decision of buying out capacity in the terrestrial system. A satellite system that does not buy out capacity must necessarily route all of the packets in their system via satellite. Buying out capacity may help alleviate congestion problems that could occur if the market demand is underestimated. Another advantage is that buying out capacity in the terrestrial infrastructure increases the number of paths available, which should benefit load balancing and reliability.

There are several scenarios in which buying out terrestrial capacity could be advantageous. First, it provides a staged deployment stepping stone such that market demand can be built up in certain areas as the system is being deployed and brought online. Second, buying out terrestrial capacity benefits customers in city environments where tall buildings commonly interfere with reception and transmission of information, even in systems with high diversity. If a satellite is not in view, cellular reception may still be possible and the call and/or data can be routed through the cellular network. Third, as the demand for the service grows, there exists the possibility that a satellite-only system will become overloaded if insufficient excess capacity is designed for. At this point, terrestrial capacity can be purchased to reduce the impact of overloading on the customers and/or to reduce the congestion experienced in high-demand areas.

The terrestrial real option is a design decision captured in a binary design variable. If the option to buy out terrestrial capacity is taken, then the value of the design variable is 1, otherwise it is 0.

3.3.2.2. Network Design Variables

Although the design parameters for the distributed satellite communication system infrastructure are well-defined and thoroughly studied, this is not the case for the network subsystem. The

design parameters of the network subsystem architecture are poorly-defined; designers do not yet know how to talk about the network design space. However, the network architecture does have well-defined characteristics that can be used to link the design of the system infrastructure to the design of the network subsystem. The network "design variables" chosen for this simulation model are thus characteristics of the network that can be varied, or decisions that can be made, in the course of the network architecture design.

Network Routing Decision

The network routing decision design variable captures a very key decision for the network designers: whether the routing of the network should be done in a centralized or distributed fashion. This routing architecture directly affects the ability of the network to achieve the desired congestion and load balancing metrics. This thesis deals with purely centralized or purely distributed routing only. In the real world, the routing architecture can be designed anywhere in-between.

Purely Centralized Routing

In a purely centralized routing architecture, all of the packets route through a centralized server before being passed onto their final destination. For the purposes of this thesis, the centralized server is assumed to exist at subnet 8. It is likely that picking another subnet as the location of the centralized server will greatly impact the results; examining the impact of the location of the centralized router in a satellite network would make for interesting research.

Purely Distributed Routing

In a purely distributed routing architecture, all subnets (nodes) in the network act as routers, passing the packets along the shortest path between the source and the destination. In the ideal case, every node knows the status of every other node and link in the network, thus guaranteeing that the packets are routed along the shortest available path. In the real world, it is rarely the case that a node knows the status of any other node besides its neighbors (if that).

Routing Protocol

Routing protocols determine the basis by which packets are routed from source to destination through the network. The decision of which path to route along is made based on the state of the network from the point of view of the router; the router builds routing tables based on this information. In the case of centralized routing, each node only knows the best path to the central node; if this path changes, the centralized router generally uploads the necessary changes to the entire network.

This thesis considers the effects of two routing protocols. The first is based on the Internet Protocol Routing Internet Protocol (IP RIP), a protocol that has been in use since the early days of networking. The state information that determines the path taken through the network is based on the number of hops required to get from a source to a destination: RIP seeks to minimize the number of hops. The second routing protocol under study is similar to the Interior Gateway Routing Protocol (IGRP) [InTH05]. This protocol is a general routing protocol that seeks to

minimize the "cost" of a packet traversing between a source and a destination. For IGRP, the "cost" is some combination of delay, bandwidth (capacity), reliability, and distance.

Multiple Access Protocol

Multiple access protocols specify how the available bandwidth is allocated to all of the users attempting to access the network simultaneously. There are two main categories of multiple access protocols: fixed assignment (the bandwidth is divided into channels which are then allocated – usually on-demand – to the users), and random access (users transmit whenever they wish to communicate; collisions between users are resolved with contention protocols). Clearly, the choice of multiple access protocol is a significant decision on the part of the network designer.

For simplicity, this thesis considers only fixed assignment protocols. The two protocols under consideration are: multiple-frequency time-division multiple-access (MF-TDMA) and multiple-frequency code-division multiple-access (MF-CDMA). These protocols and the development of their models will be discussed in more depth later on in this section.

Automatic Repeat Request (ARQ) Protocol

ARQ protocols specify how the network architecture corrects for corrupted packets. In the majority of cases, errors cannot be corrected internally at the destination node. However, protocols can be written directing how the destination node can acknowledge the arrival of successful packets and/or request retransmissions of packets with errors. These protocols also account for packets lost due to congestion – ones that never arrive at the destination – by incorporating a timeout process that resends packets that are not acknowledged or requested within a certain amount of time. Choosing an appropriate ARQ protocol can have a significant impact on the performance of a network.

This simulation model considers two of the basic ARQ protocols: Go-Back N and Selective Repeat Protocol (SRP). Most current ARQ protocols are based on these generalized models: the differences are mainly due to improvements on the basic protocol structure. The ARQ protocol models considered in this thesis are ideal since only packets containing errors are assumed to be retransmitted; in real systems, sometimes packets may need to be retransmitted if their window has expired. For the purposes of this thesis, packets are assumed in error if there at least a single bit error or if the packet is lost due to congestion.

Modulation Scheme

Modulation schemes determine the achievable probability of bit error on a link based on the required signal-to-noise ratio (or, vice-versa, the bit-error rate necessary to achieve a required signal-to-noise ratio). Well-designed modulation schemes can significantly impact the quality of service seen by the customers by limiting the number of packets affected by errors.

Although there are a number of different modulation schemes used in satellite communication systems, in the interest of time and ease of implementation, only two of the most basic were chosen for consideration: binary phase-shift keying (BPSK) and quadrature phase-shift keying (QPSK). Both of these modulation schemes are a good balance between achievable signal-to-noise ratio for a given BER and utilization of the available spectrum.

Average Data Packet Size

The average data packet size determines the number of user data bytes encapsulated by the various layers of the ISO-OSI network interface. Most standard protocols have a wide range of packet sizes that the protocol can support, leaving the network designer leverage on which packet size or set of packet sizes the network will utilize.

In most standard protocols, the average data packet size largely determines the amount of overhead that must be incorporated into the finished packet so that the user data can be processed correctly. Larger packet sizes tend to require an increased number of extra bytes of information; but in smaller packet sizes, the required extra bytes might be a substantial percentage of the total transmitted packet size (the overhead efficiency is low, in this case).

Also, for a given link capacity and assuming that the channel slot size corresponds to the size of the average packet, the packet size and data rate limit the number of channels available on a given link. Larger packet sizes may incur more congestion penalties than the smaller packet sizes for a fixed available bandwidth if many users attempt to access the same link simultaneous.

This simulation model considers three packet sizes: 10, 100, and 1000 bytes.

Average User Data Rate

The average user data rate characterizes the type of service the network is offering. For example, voice telephony services require only about 10 kbps uncompressed, while broadband data services require on the order of 1 Mbps. In a system with a fixed capacity, smaller average user data rates can accommodate a larger number of simultaneous users as there are more channels of a fixed size available. However, larger average user data rates increase the transparency of the network to an individual user so long as congestion factors do not become overwhelming.

This thesis considers three average user data rates: 10 kbps for telephony services, 100 kbps for radio and TV services, and 1 Mbps for broadband internet services. It is assumed that these average user data rates apply to each half-duplex channel in the network. Thus, the data transferred from the source to the destination is transmitted at this rate, and the return data from the destination to the source is transmitted at the same rate.

3.3.3. System Requirements

The simulation model requires that the average inter-satellite link data rate per channel be the same as the average user data rate. This assumption simplifies the network calculations, but does overlook the fact that inter-satellite links can support much higher data rates, which could reduce the average round-trip delay.

It is assumed that the required BER between each source and destination is 1e-2 (based on the Technical Specification of Globalstar given in [ChaD03]). For the purposes of this simulation model, the required BER is used as the probability of bit error used to find $\frac{E_b}{I_{tot}}$ in the estimation for the number of CDMA channels in Section 3.3.12.2. Thus, if the maximum BER on a link in the network is less than 1e-2, then the system architecture meets the minimum requirement; otherwise the system fails to meet this requirement and the system fails. Although the model

does not incorporate a filter to weed out designs that fail this requirement, achieving optimal – or near-optimal – network objectives should guarantee that the system meets this requirement.

The BER seen on each terrestrial link is assumed to be 1e-10, while the BER seen over each satellite uplink/downlink and ISL link is calculated based on the environmental conditions and distance.

3.3.4. Policy Requirements

There has been recent interest in satellite systems in the Ka band frequency spectrum for the large amount of untapped bandwidth. Although the Ka band suffers from greater atmospheric degradation than the Ku band – which was widely used in the Big LEO satellite systems such as Globalstar and Iridium – it does boost the effective gain of the transmitting and receiving antennas [Feng01].

This thesis assumes that the Federal Communication Commission (FCC) assigns the satellite system 5 MHz of bandwidth around a frequency of 30 GHz for the satellite uplink carrier, 5 MHz of bandwidth around a frequency of 18 GHz for the satellite downlink carrier, and 100 MHz of bandwidth around a frequency of 60 GHz for the inter-satellite link carrier.

It is further assumed that if the system designers choose to buy some of the terrestrial capacity for use by the satellite system, then the capacity per link that the satellite system purchases from the terrestrial network is 0.1% of the total capacity of each terrestrial link. Since the total terrestrial capacity per link shown in Figure 8 is assumed to be 10,000 Mbps, this translates to about 10 Mbps that the satellite system buys per link.

3.3.5. Market-Traffic Model

The market model is based on the work done by Kashitani [Kash02], while the traffic model is largely founded on the work done by [Moho00] – see Section 3.2.2. This thesis improves upon this traffic model and integrates it with the market model.

The market model attempts to capture information about the one aspect of a satellite system that cannot be controlled or predicted: the people who may, or may not, subscribe to the service the satellite system provides. For this reason, market models are difficult to produce, and even more difficult to guarantee accuracy.

Fortunately, this thesis cares about the market model only to the extent that it makes some prediction – to first-order only – of the distribution of the traffic loads entering the network. Since the traffic models are likewise difficult to predict, merely considering a combined market-traffic model accounting for the geographically non-uniform nature of demand and traffic loading is sufficient to first-order. After all, this thesis is only looking to understand the interactions between the system and network architectures, not to find an optimal system designed to meet a particular market.

The data collected and the basic market model were developed by Kashitani [Kash02]. Kashitani reasonably assumes that a customer's willingness and ability to subscribe to a telecommunications service is dependent on their economic status and exposure to technology – modeled by the world map of Gross National Product adjusted by Purchasing Power Parity (PPP) in Figure 23. Furthermore, the number of customers in an area willing to subscribe is likely also

dependent on the population density, also shown in Figure 23. Of course, there are many other factors contributing to a consumers decision to subscribe or not, but these issues seem sufficiently dominant for the purposes of this thesis.

Combining the data shown graphically in Figure 23, it is possible to construct a normalized matrix of the relative demand weightings distributed geographically. Figure 24 shows one such market demand map matrix. This information is combined with the traffic model in section 3.2 to yield the first-order joint market-traffic model used in the Significance proof.

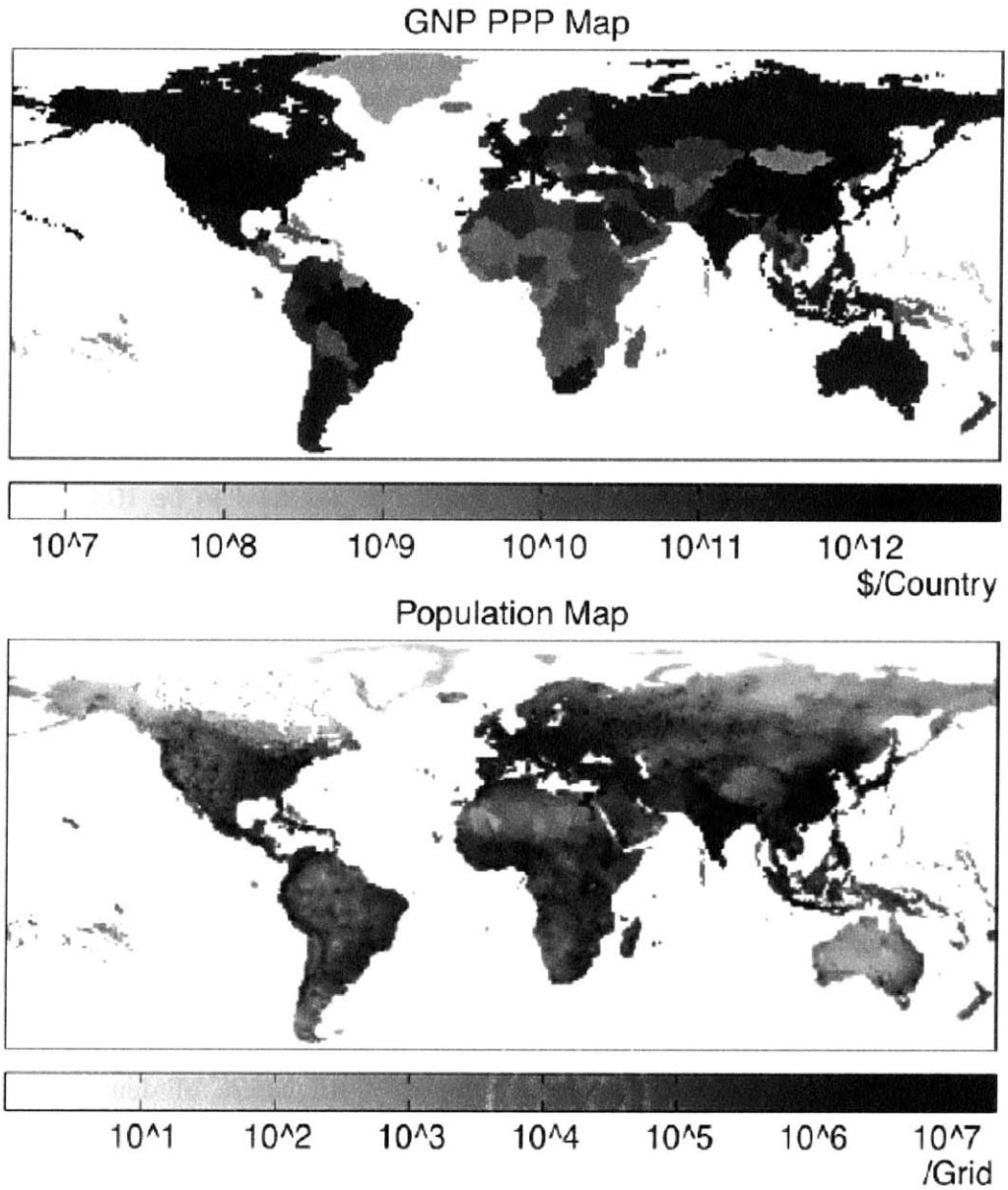


Figure 23: Global Gross National Product (GNP) and Population Distribution Maps
The market model assumes that a customer's willingness and ability to subscribe to a telecommunications service is dependent on their economic status and exposure to technology; this assumption is captured in the GNP PPP model. Furthermore, the number of customers in an area willing to subscribe is likely dependent on the population density.
Picture taken from [Kash02].

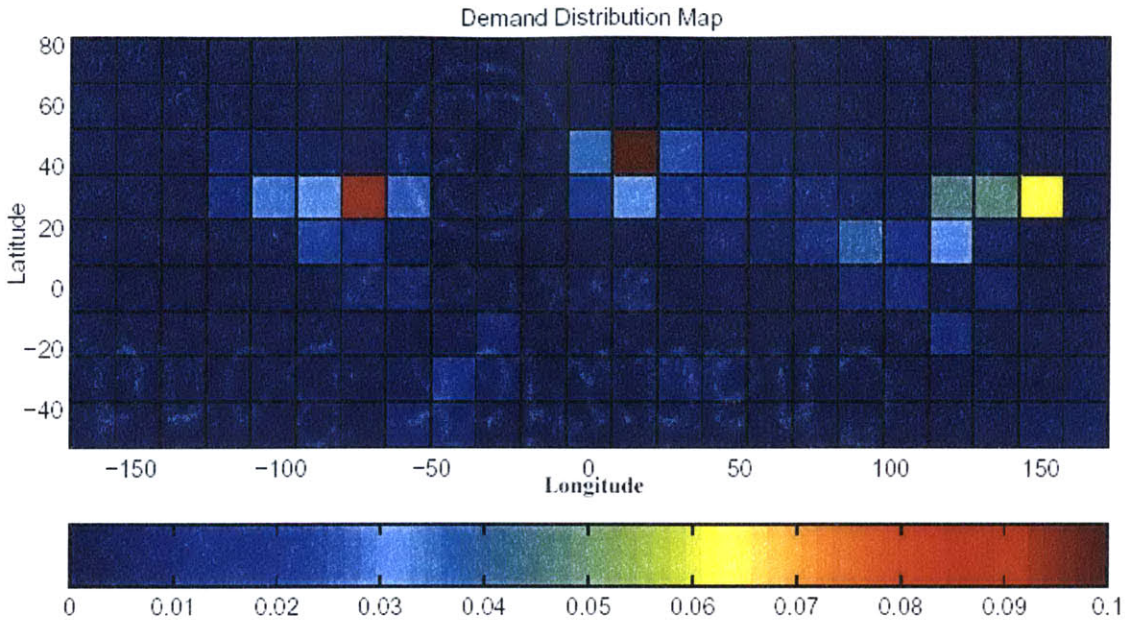


Figure 24: Market Demand Map
Combining the GNP PPP data with the population densities enables construction of a normalized matrix of the relative demand weightings distributed geographically. This information is further combined with the traffic model in Section 3.2 to yield the first-order joint market-traffic model used in the significance proof. Picture taken from [Chan04].

Given no other user characteristics, a first order approximation is made. First, it is assumed that the normalized market distribution in Figure 24 corresponds to the relative packet arrival rate weightings to each source node (λ^S) instead of simply assuming a uniform rate of 1 as the model in section 3.2 does. Combined with the probability of a packet going to a given destination D given a source S ($P(S \mid D)$) matrix given in section 3.2, we can find the normalized arrival rate weightings of packets from subnet source S destined for destination subnet D:

$$\lambda^{SD} \mid_{Normalized} = \lambda^S \cdot P(S \mid D)$$

Equation 9

Further, given the estimated market potential calculated to satisfy the market potential objective, the arrival rates (packets per second) can be scaled to model the traffic load at both the beginning and end of the satellite system lifetime. Designing the simulation model this way enables modeling of the effects of a growing subscriber base on the performance of the system – an aspect usually overlooked in the literature. To see how this system could be scaled, consider Figure 25. Figure 25(a) shows a network in terms of the total load on each link. By inspection, the total load to the system is 30 units. Figure 25(b), on the other hand, shows a network in terms of the normalized arrival rate weightings.

It is assumed that each user generates one packet per second on average. Thus, on average, the total load to the network will be the number of subscribers (in the case of Figure 25, thirty subscribers), which will change – hopefully grow – over time. Multiplying the total load to the

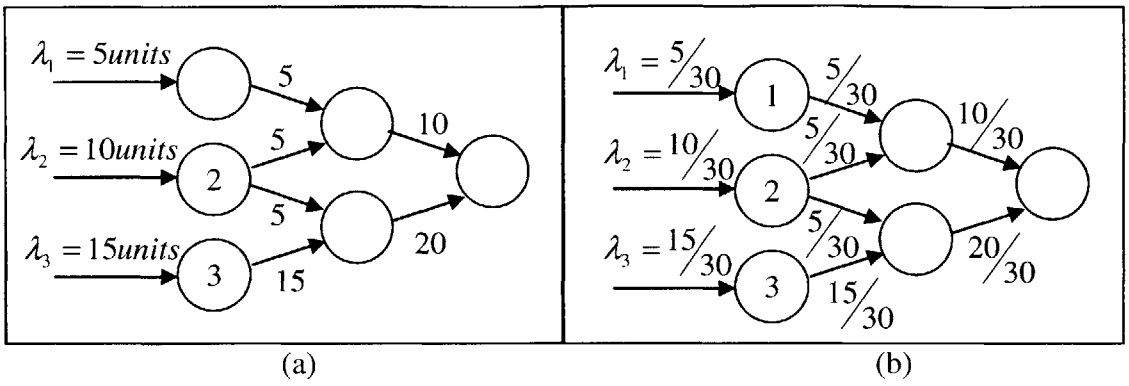


Figure 25: Conversion of (a) Actual Loading on Links to (b) Relative Loadings
Subnets 1, 2, and 3 are the source subnets to the network. The load is distributed on the links as shown in (a). By inspection, the total load on the system is 30 units. Now consider (b): multiplying the relative weightings by the total load to the system gives the total load to that particular link.

network by the normalized arrival rate weightings will generate (a) again. The two cases are identical as long as (b) is multiplied by the total load to the network. Note that this assumes that the traffic characteristics do not change over time.

Given that some users will be streaming continuous data through the network while others will hardly use it at all, the usage assumption is fair and simplifies the calculations. If anything, this assumption approximates peak loading on the network.

Within each subnet, the traffic is assumed to be uniformly distributed geographically. If a satellite can see part of the land-area covered by a subnet, then the satellite is assumed to receive the same arrival rate from that subnet as a satellite visible to the entire area. This is certainly not true: consider the case where a satellite can see ocean, but not the United States – the satellite would receive the arrival rate typical of an ocean setting but not the arrival rate typical of a populated country. However, this assumption is reasonable to first order since the aggregated arrival rates to a subnet are scaled by the geographical landmasses in that subnet (see Section 3.2 for further discussion).

All of the terrestrial subnets that can see a given satellite share the communication link between the ground and that satellite, as they would in real life. All of the terrestrial subnets and links are assumed to be identical.

3.3.6. Cost Model

The cost model used in the Significance proof is principally based on the Matlab simulation model developed by Chang [ChaD04], which roughly follows the procedure outlined in Chapter 20 of [Wert99]. This thesis incorporates consideration of the cost of buying out terrestrial channels.

The model assumes a discount rate of 15% APR, and an initial development time of 5 years. Furthermore, it is assumed that the number of spare spacecraft is equal to the number of planes in the constellation. The constellation is assumed to be developed for a commercial program. It is important to note that the life cycle cost calculation does not include user equipment cost.

Once the space mission characteristics are determined – most of which are determined by the design variables, fixed assumptions, or by calculations earlier in the simulation – the cost of various components is estimated. First, the Research, Development, Test, and Evaluation (RDT&E) hardware costs are calculated. This computation assumes the system uses a nominal new design leveraging off of heritage from existing systems. Second, the Theoretical First Unit (TFU) hardware costs are estimated, followed by the hardware costs of every unit accounting for manufacturing learning curve. Finally, the aerospace ground equipment cost for RDT&E, the total program level cost for RTD&E, TFU, and every unit, the launch operations and orbital support cost for all units, the flight software cost, the launch vehicle cost, the ground software cost, the total ground segment development cost, the initial deployment cost (IDC) assuming 60% of the expenditure occurs by the middle of the schedule, and the operation and support cost (OSC) during the life time (assuming contractor labor) are found.

The cost of buying out terrestrial channels is estimated at approximately $100 in year 2005 dollars per year per channel.

3.3.7. Spacecraft Model

The spacecraft Matlab simulation model is based largely off of the LEO constellation design work of Chang and Springmann [Spri03]. This thesis extends the work of Chang and Springmann to include non-LEO constellations.

The spacecraft model attempts to simulate the design of the communication satellites used in the constellation. Given the required satellite power and antenna size – design variables in this case – the model parametrically scales the spacecraft mass and volume.

The model assumes that all of the communication satellites in the network are identical. Each satellite is assumed to have two parabolic antennas – one for the uplink communications, the other for the downlink – and four optical inter-satellite links. Also, the model accounts for the analog/digital beam forming required if the cell duration (the amount of time a beam can see a cell) is less than thirty seconds. It is further assumed that the apogee kick motor and attitude control system is 3-axis stabilized, with a motor specific impulse of 290 seconds. The specific impulse of the stationkeeping engines, assuming liquid monopropellant fuel, is 230 seconds.

The major difference between the model used in the Significance proof and the model developed by Chang and Springmann is the extension of validity to constellations existing beyond LEO. The expansion to altitudes beyond LEO is accomplished by accounting for differences in de-orbiting protocols and calculations of required stationkeeping delta-v's.

3.3.8. Launch Model

The launch Matlab simulation model is adapted from the work done by Chang [ChaD04]. This thesis enhances the built-in model of the Atlas IIIA to include a GTO (geostationary transfer orbit) booster, enabling launches to MEO and GEO altitudes.

Given the altitude, minimum elevation angle, number of satellites, and mass and volume characteristics of the satellites, the launch model returns the best launch vehicle capable of meeting these requirements as well as estimates of the number of launches required, the total cost of launching the constellation, and data on the failure rate of the launch vehicle.

3.3.9. Operations Model

The operations model is also based on the work of Darren Chang [ChaD04].

The model attempts to account for the cost of operating a satellite communication system. Operation costs arise out of the need for systems on the ground to monitor the health of the satellite constellation (and the ground stations) and to make corrections as needed. To do this, it is assumed that there are two ground stations for each plane in the satellite constellation (communication with these ground stations is assumed to be contained within the subnet traffic models). In some satellite architectures, all calls through the satellite network must be routed through these ground stations. This simulation operates independently of whether or not this is true – the subnet architectures are transparent to the satellite network.

At each of these ground stations, personnel are required 24 hours a day for monitoring. This simulation assumes that there are three personnel per shift per ground station plus two personnel per shift per command center (it is assumed there are two command centers). Furthermore, it is assumed that a shift is 8 hours in duration.

To enable monitoring and execution of commands, each ground station and command center is equipped with specialized software. Based on GlobalStar, ARIES, ORBCOMM, and Starnet, the ground system is assumed to have 6,300 lines of code.

The dollar costs associated with these operational expenditures is accounted for in the cost model.

3.3.10. Constellation Topology

The constellation topology model incorporates a calculation of the number of satellites and planes required for full global coverage used in [ChaD04], as well as an estimation of the number of cells in a footprint given in [Lutz00]. The topology and connectivity matrix generator was developed specifically for this thesis.

The constellation topology model has several key functions: estimate the number of satellites and constellation planes required to achieve full Earth coverage; estimate the number of spot beam cells in a satellite footprint; assign satellites to initial orbital slots; and analyze the network connectivity among and between satellites and terrestrial subnets.

The simulation model makes several simplifying assumptions. First, the space network never moves relative to itself; in other words, movement at the cross-seam is ignored. Similarly, inter-satellite links are assumed to exist, except at the cross-seams. The constellation topology is assumed to be comprised of polar, circular orbits, whereby the effects seen at the poles are ignored. Furthermore, the Earth is assumed to be perfectly spherical, and circular footprints are assumed to be approximated as squares for the purposes of estimating connectivity. The constellation is assumed to have a diversity of one.

Numerically, the average Earth radius is assumed to be 6,371 kilometers, the period of the Earth (for example, the period at GEO) is 1,440 minutes, and the speed of light is 2.998e8 meters per second.

Given the orbital altitude in kilometers (h), the minimum elevation angle in radians (elev), and the radius of the earth in kilometers (R_{Earth}), the nadir angle – a measurement of the angle between the subsatellite point of the spacecraft and some target position (the edge of the footprint, in this case) – η in radians can be found using Equation 10:

$$\eta = \sin^{-1}\left(\frac{R_{Earth}}{(R_{Earth}+h)}\cos(elev)\right)$$

Equation 10

Given the nadir angle and the beamwidth in radians of a spot beam at an edge cell (θ_{edge}), the number of cells in the footprint can be estimated by Equation 11:

$$\#cells = 1.21 \cdot \frac{(1-\cos(\eta))}{(1-\cos(\theta_{edge}))}$$

Equation 11

The simulation code that assigns each satellite to its initial orbital slot automatically generates the subsatellite point coordinate array for the constellation. The first satellite is assigned a subsatellite point at coordinates (0,0), corresponding to 0° latitude, 0° longitude. The longitudinal coordinate also corresponds to the first constellation plane. Next, each plane is assigned a subsatellite longitudinal coordinate spaced equally around the globe. A single satellite is initially assigned to each of these planes at 0° latitude. Finally, the remaining satellites are placed such that their latitudinal coordinates are equally spaced around each plane. The subsatellite point latitude-longitude coordinates are now known for the initial constellation topology.

Once the subsatellite point coordinate array is found, the connectivity among and between the satellites and terrestrial subnets can be found. Finding the in-plane connectivity is easy, since each satellite is assumed to be able to see the satellite ahead and behind (the only exception is at GEO, where there are no other satellites in the plane). Since the satellite network does not move relative to itself, the intra-plane inter-satellite links never change. The initial configuration has each of the satellites in a ring aligned along the same latitudinal line, so each satellite should have connectivity to the satellite immediately to the left and the satellite immediately to the right, so long as none of these connections occur at the cross-seam. To calculate the connectivity between each terrestrial subnet and a satellite: if the satellite can see any part of the given subnet, then it is assumed that there is connectivity between the satellite and subnet.

3.3.11. Modulation Schemes

The analysis of the impact of the modulation schemes are mostly based on the work of [Orfa04] and [Proa02]. Some modifications are made in order to enable the calculation of the required signal-to-noise ratio to achieve a desired bit-error rate.

The two modulation schemes under consideration are phase-coherent Binary Phase Shift Keying (BPSK) and Quadrature Phase Shift Keying (QPSK). The structure of the simulation enables driving the modulation process both ways: given a desired signal-to-noise ratio, what is the required probability of bit error for the given modulation scheme; and given a desired probability of bit error, what is the necessary signal-to-noise ratio?

Most of the systems literature finds the required signal-to-noise ratio to achieve a desired probability of bit error on a link. However – assuming the system does not use regenerative

repeaters at each node – as a call or data packet is routed along multiple links, the signal-to-noise ratio is altered according to Equation 12 [Orfa04]:

$$\left(\frac{E_b}{N_o}\right)^{-1}_{TotalPath} = \left(\frac{E_b}{N_o}\right)^{-1}_{link1} + \left(\frac{E_b}{N_o}\right)^{-1}_{link2} + \cdots + \left(\frac{E_b}{N_o}\right)^{-1}_{linkN}$$

Equation 12

Assuming that the signal repeaters at each of the nodes in the network are non-regenerative – the signal is not rebuilt before it is amplified and sent along to the next node – is a fair assumption since the cost and the need to obtain space qualification for regenerative repeaters onboard satellites usually prevents satellite companies from employing them. The computer processing requirements for regenerative repeaters are not insubstantial.

Based on Equation 12, the actual probability of bit error experienced by the user depends on the routed path and the modulation scheme used, and merely using the probability of bit error on a link as a measure of customer quality of service is insufficient and generally leads to a significantly better expected probability of bit error than can be achieved. So, although the probability of bit error tends to be a system requirement, it is too simplistic to find the required signal-to-noise ratio on the basis of a single link.

A better method calculates the achievable signal-to-noise ratio per link and determines the probability of bit error experienced by the customer based on the total signal-to-noise ratio per routed path. This technique forms the basis for the simulation model used in the Significance proof. However, there are cases in the model for which the modulation scheme is driven the other way. For example, the terrestrial links are assumed to provide a certain bit-error rate (BER) per link; the design of these links is beyond the control of the satellite system or network designers. The expected signal-to-noise ratio for each of these links is then calculated based on the assumed BER, and is used to calculate the overall signal-to-noise ratio per routed path.

BPSK and QPSK are modulation schemes commonly found in existing satellite systems. They demonstrate good tradeoffs between BER performance – as measured by the signal-to-noise ratio achievable at a specified BER – and use of spectrum. However, both are susceptible to phase disturbances, so they are not good choices for some types of systems [Wert99]

3.3.11.1. Binary Phase Shift Keying (BPSK)

In a BPSK modulation scheme, the carrier signal phase is set to 0° to transmit a binary 0 and 180° to transmit a binary 1. The signal constellation representation of this antipodal system is shown in Figure 26 where the distance to the signals is the energy per bit [Proa02].

In general, given the distance d separating two symbols, the probability of error between them can be found by Equation 13 [Modi04]:

$$P_e = Q\left(\sqrt{\frac{d^2}{2N_o}}\right)$$

Equation 13

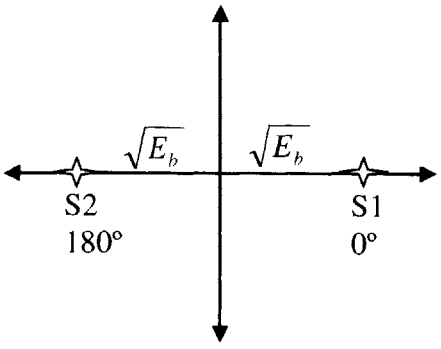


Figure 26: Signal Constellation for BPSK Modulation
Signal S1 represents the carrier phase for binary "0" and S2 represents the carrier phase for binary "1".

In the case of BPSK, d is simply $2\sqrt{E_b}$. Combining with Equation 13 gives the following probability of error:

$$P_e = Q\left(\sqrt{\frac{2E_b}{N_o}}\right)$$

Equation 14

If the system is driven the other way, then the signal-to-noise ratio per bit $\frac{E_b}{N_o}$, can be found as follows. First, start with the definition of *Q(x)*.

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt$$

Equation 15

Q(x) can also be expressed in terms of the error function *erf(x)*.

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Equation 16

The relationship between *Q(x)* and *erf(x)* can be found by change of variable. Let $u = t/\sqrt{2}$, then Equation 15 becomes:

$$Q(x) = \frac{1}{\sqrt{\pi}} \int_{\frac{x}{\sqrt{2}}}^{\infty} e^{-u^2} du$$

Equation 17

Combining with Equation 16 and rearranging gives:

$$Q(x) = \frac{1}{2}\left[1 - erf\left(\frac{x}{\sqrt{2}}\right)\right]$$

Equation 18

Thus, for BPSK systems, the probability of error is found by Equation 19.

$$P_e = Q\left(\sqrt{\frac{2E_b}{N_o}}\right) = \frac{1}{2}\left[1 - erf\left(\sqrt{\frac{E_b}{N_o}}\right)\right]$$

Equation 19

Rearranging Equation 19 enables calculation of the required signal-to-noise ratio per bit.

$$\frac{E_b}{N_o} = \left(erf^{-1}(1 - 2P_e)\right)^2$$

Equation 20

3.3.11.2. Quadrature Phase Shift Keying (QPSK)

QPSK defines four symbols corresponding to one of four carrier phases. Two bits are required per symbol as shown in Figure 27: 00 for 0° carrier phase, 01 for 90° carrier phase, 11 for 180° carrier phase, and 10 for 270° carrier phase (other implementations may differ). The reduction in symbol rate by one half of the bit rate likewise reduces the required spectrum by one half [Wert99].

Assuming a perfect estimate of the carrier phase, QPSK acts just like two orthogonal binary-phase modulation signals with a bit error probability identical to that for BPSK. However, perfect estimates are nearly impossible to attain, so a more accurate estimate of the symbol error probability must be found.

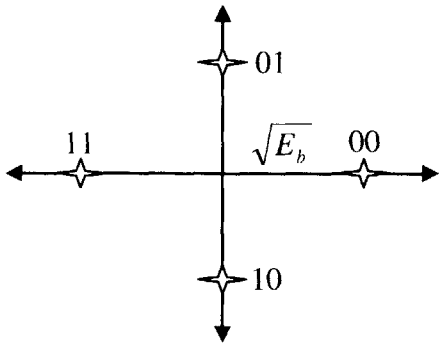


Figure 27: Signal Constellation for a QPSK Modulation Scheme
QPSK defines four symbols corresponding to one of four carrier phases. Two bits are required per symbol: 00 for 0° carrier phase, 01 for 90° carrier phase, 11 for 180° carrier phase, and 10 for 270° carrier phase (other implementations may differ). The reduction in symbol rate by one half of the bit rate likewise reduces the required spectrum by one half.

If statistical independence is assumed for the noise on the quadrature carriers, then the probability of a correct decision for the QPSK symbol is (from [Proa02]):

$$P_c^4 = \left(1 - P_e^2\right)^2 = \left[1 - Q\left(\sqrt{\frac{2E_b}{N_o}}\right)\right]^2$$

Equation 21

The probability of an incorrect decision, corresponding to the symbol error probability, is thus:

$$P_e^4 = 1 - P_c^4$$

Equation 22

Plugging in Equation 21 for P_c^4 gives:

$$P_e^4 = 2Q\left(\sqrt{\frac{2E_b}{N_o}}\right)\left[1 - \frac{1}{2}Q\left(\sqrt{\frac{2E_b}{N_o}}\right)\right]$$

Equation 23

The signal-to-noise ratio per bit $\frac{E_b}{N_o}$ can be found by rearranging Equation 23.

$$Q\left(\sqrt{\frac{2E_b}{N_o}}\right)^2 - 2Q\left(\sqrt{\frac{2E_b}{N_o}}\right) + P_e^4 = 0$$

Equation 24

Solving this quadratic equation results in Equation 25:

$$Q\left(\sqrt{\frac{2E_b}{N_o}}\right) = 2 - 2\sqrt{1 - P_e^4}$$

Equation 25

From the discussion on BPSK, it is known that:

$$Q\left(\sqrt{\frac{2E_b}{N_o}}\right) = \frac{1}{2}\left[1 - erf\left(\sqrt{\frac{E_b}{N_o}}\right)\right]$$

Equation 26

Plugging into Equation 25 and solving for $\frac{E_b}{N_o}$ results in:

$$\frac{1}{2}\left[1 - erf\left(\sqrt{\frac{E_b}{N_o}}\right)\right] = 2 - 2\sqrt{1 - P_e^4}$$

Equation 27

$$\frac{E_b}{N_o} = \left[erf^{-1}\left(1 - 4 \cdot \left(1 - \sqrt{1 - P_e^4}\right)\right)\right]^2$$

Equation 28

Both BPSK and QPSK suffer from phase distortions. To demodulate the signal, the phase of the received carrier must be measured. If the phase distortions are significant enough, then the demodulation process will not correctly retrieve the original signal. QPSK systems are more vulnerable to phase distortions than BPSK due to smaller differences between the phases.

3.3.12. Multiple Access Protocols

The multiple access protocol models developed for this thesis are based on the work done by Chang and de Weck [ChaD03]. Although the derivation process used in this thesis is nearly identical to that done in [ChaD03], the driving variables are considerably different, leading to

equations similar in form but incorporating different variables. Furthermore, the work done in Chang and de Weck is extended to include accounting of the effects of buffering and queuing in the links.

3.3.12.1. Multiple Frequency – Time Division Multiple Access (MF-TDMA)

Time Division Multiple Access (TDMA) is structured such that user access time is divided into frames – assumed to be 90 milliseconds in duration – and the time frames are further divided into time slots – assumed to be the amount of time required to send one packet worth of data; see Figure 28. A guard time is inserted between each packet.

Each packet is assumed to consist of the average user data packet Pkt_{user} plus some amount of network overhead OH associated with the required network headers.

If the duration of a packet time slot is T_{slot}, then to guarantee a required user data rate R_{user} in bps, Equation 29 must hold:

$$R_{user} = \frac{Pkt_{user}}{T_{slot}}$$

Equation 29

Similarly, if the burst data rate allowed for each TDMA carrier is R_b in bps, then the following must also be true:

$$R_b = \frac{Pkt_{user} + OH}{T_{slot}}$$

Equation 30

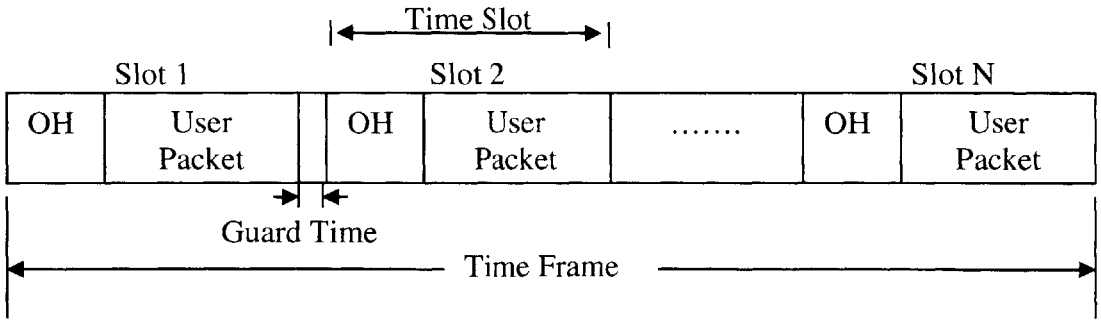


Figure 28: Structure of Time Division Multiple Access (TDMA)
Each user access time in TDMA is divided into frames – assumed to be 90 milliseconds in duration – and the time frames are further divided into time slots – assumed to be the amount of time required to send one packet worth of data. A guard time is inserted between each packet.

Combining Equation 29 and Equation 30 gives:

$$R_b = \left(Pkt_{user} + OH\right) \cdot \frac{R_{user}}{Pkt_{user}}$$

Equation 31

According to Chang and de Weck, the number of TDMA channels for a given burst data rate, assuming a frame duration T_f and guard time T_g in seconds is:

$$N_{hd} = \frac{R_b T_f}{Pkt_{user} + OH + R_b T_g}$$

Equation 32

The guard time is assumed to be 9.6 microseconds, based on the minimum inter-packet gap for a standard multiple access layer protocol. Plugging into Equation 32 for R_b :

$$N_{hd} = \frac{\left(\frac{R_{user}}{Pkt_{user}}\right) T_f}{1 + \left(\frac{R_{user}}{Pkt_{user}}\right) T_g}$$

Equation 33

Multiple Frequency-TDMA (MF-TDMA) increases the number of channels available by associating multiple TDMA carriers with different frequency channels. To avoid inter-symbol interference, the receiver is assumed to use Nyquist filtering with a filter roll-off factor $\beta = 0.35$. Thus, the channel bandwidth required for a given signal modulation level M can be found by:

$$B_{ch} = \frac{(1+\beta) \cdot R_b}{\log_2 M}$$

Equation 34

Again, plugging into Equation 34 for R_b :

$$B_T = \frac{(1+\beta) \cdot \left(Pkt_{user} + OH\right) \cdot R_{user}}{\left(\log_2 M\right) \cdot Pkt_{user}}$$

Equation 35

The total bandwidth required to support T TDMA carriers, assuming the frequency carriers occupying bandwidth B_T are separated by guard bands of size B_g as shown in Figure 29, is given by:

$$B_{TOT} = T \cdot (B_T + B_g)$$

Equation 36

The guard band is assumed to be 1.236 kHz; based on the Iridium system. Rearranging Equation 36 and plugging in for B_T:

$$T = \frac{B_{TOT}}{\left[\frac{(1+\beta)\cdot(Pkt_{user} + OH)\cdot R_{user}}{(\log_2 M)\cdot Pkt_{user}} + B_g\right]}$$

Equation 37

According to Chang and de Weck, the total number of MF-TDMA channels available given the number of TDMA channels in a time frame T_f and the number of TDMA carriers for a given bandwidth B_{TOT} is:

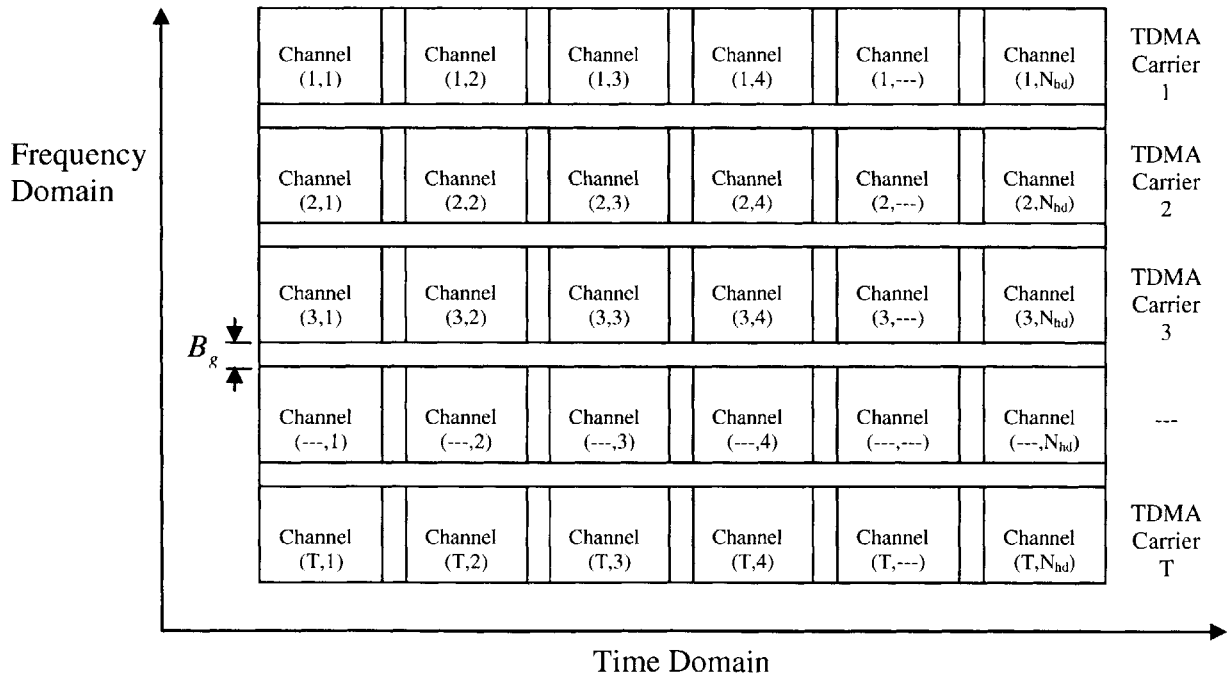


Figure 29: Structure of Multiple Frequency-TDMA (MF-TDMA)
MF-TDMA increases the number of channels available by associating multiple TDMA carriers with different frequency channels.

$$N_{ch} = T \cdot N_{hd}$$

Equation 38

Substituting in Equation 37 for T and Equation 33 for N_{hd}:

\

$$N_{ch} = \frac{B_{TOT}}{\left[\frac{(1+\beta)\cdot(Pkt_{user}+OH)\cdot R_{user}}{(\log_2 M)\cdot Pkt_{user}} + B_g\right]} \cdot \frac{\left(\frac{R_{user}}{Pkt_{user}}\right)T_f}{1+\left(\frac{R_{user}}{Pkt_{user}}\right)T_g}$$

Equation 39

The number of uplink channels is assumed to be one half of this value; the same is assumed true for the number of downlink channels. Furthermore, the system is assumed to use spot beams. Thus, the number of cells Z and the cluster size K – the number of frequency bands used in the cells – must also be accounted for:

$$N_{ch} = \left(\frac{Z}{K}\right) \cdot \frac{B_{TOT}}{\left[\frac{(1+\beta)\cdot(Pkt_{user}+OH)\cdot R_{user}}{(\log_2 M)\cdot Pkt_{user}} + B_g\right]} \cdot \frac{\left(\frac{R_{user}}{Pkt_{user}}\right)T_f}{1+\left(\frac{R_{user}}{Pkt_{user}}\right)T_g}$$

Equation 40

It is assumed that if there is a single cell in the footprint, then the MF-TDMA cluster size is one; if there is more than one cell in the footprint, then the MF-TDMA cluster size is assumed to be twelve, the same as in Iridium.

3.3.12.2. Multiple Frequency – Code Division Multiple Access (MF-CDMA)

In CDMA, unique orthogonal pseudorandom noise codes are used to differentiate users simultaneously accessing the spectrum. The transmitted signal is then spread over a larger bandwidth than is needed for transmission; the receiving station must have the correct code to retrieve the original encoded information. As the number of users grows, each individual user experiences more interference, increasing the observed data error rate. This phenomenon is captured by including the effect of the mean total noise power spectral density I_{tot}. CDMA is well-suited to satellites as it avoids the near-far problem; all transmitters are approximately the same distance from the satellite receivers. Since all of the transmitters are operating at roughly the same power levels, much of the need for strict power control is alleviated.

For C CDMA carriers, the total number of MF-CDMA channels can be found using an equation from Chang and de Weck [ChaD03]:

$$N_{ch} = C + \frac{B_{TOT} - C \cdot B_g}{R_b} \cdot \frac{1}{\alpha} \cdot \left[\left(\frac{E_b}{I_{tot}} \right)^{-1} - \left(\frac{E_b}{N_o} \right)^{-1} \right]$$

Equation 41

Given that p_b is the required probability of bit error, Equation 41 assumes the following:

$$\frac{E_b}{I_{tot}} = \left[erfc^{-1}(2 \cdot p_b) \right]^2$$

Equation 42

The number of CDMA channels, based on the number from the Globalstar system, is assumed to be 13. The expected value of the voice activity state α is 0.5 since the transmit signal is switched off to save power during pauses in speech or data transmission. Again, plugging into Equation 41 for known quantities:

$$N_{ch} = C + \frac{\left(B_{TOT} - C \cdot B_g\right) \cdot Pkt_{user}}{R_{user} \cdot \left(Pkt_{user} + OH\right)} \cdot \frac{1}{\alpha} \cdot \left[\left(\frac{E_b}{I_{tot}} \right)^{-1} - \left(\frac{E_b}{N_o} \right)^{-1} \right]$$

Equation 43

Accounting for the use of spot beams gives:

$$N_{ch} = Z \cdot \left[C + \frac{\left(B_{TOT} - C \cdot B_g\right) \cdot Pkt_{user}}{R_{user} \cdot \left(Pkt_{user} + OH\right)} \cdot \frac{1}{\alpha \cdot (1+f)} \cdot \left[\left(\frac{E_b}{I_{tot}} \right)^{-1} - \left(\frac{E_b}{N_o} \right)^{-1} \right] \right]$$

Equation 44

For MF-CDMA systems, the cluster size is one since all cells can utilize the full frequency band. Interference will increase due to users in neighboring cells; the factor $(1+f)$ accounts for this phenomenon. The value of f used in the simulation is assumed to be the lower bound for f, which is 1.36.

The structure of MF-TDMA and MF-CDMA is also important for estimating the delay seen by the average packet as well as calculating the number of packets the system can support. Queuing theory is needed to develop models for these aspects of the communication network.

In Figure 29, there are clearly $T \cdot N_{hd}$ channels in the MF-TDMA system (or $C \cdot N_{hd}$ channels in the equivalent MF-CDMA system). Each of these channels can be thought of as an M/D/1 queuing system: the arrivals are assumed to be Poisson memoryless processes (see the traffic model), the service times are deterministic, and there is one server – the channel. Thus, each of these channels can be analyzed in terms of queuing theory.

The amount of time spent in queue for a Time Division Multiplexing (TDM) system – equivalent to a slotted Frequency Division Multiplexing (FDM) system with slots of 1 time unit – is given by Equation 45:

$$W_{TDM} = W_{SFDM} = \frac{\lambda E[x^2]}{2 \cdot (1-\rho)} + \frac{E[v^2]}{2 \cdot E[v]}$$

Equation 45

where $E[x]$ is the expected value of the service time, $E[x^2]$ is the second moment of the service times, $E[v]$ and $E[v^2]$ are the corresponding values of the vacations taken when there is nothing to transmit (the channel is idle), λ is the arrival rate of packets to the queue, and ρ is the total system load ($\rho = \lambda \cdot E[x]$).

Since the service time has been strictly enforced by assuming that each time slot is the size – and hence, duration – of a single packet, then $E[x] = E[v] = D_{TP}$, the packet transmission time and $E[x^2] = E[v^2] = (D_{TP})^2$ since the value of D_{TP} is deterministic. Furthermore, it is assumed that the arrival rate to each of the channels is the total arrival rate to the queue equally shared among the channels. Thus, the amount of time spent in queue for the MF-TDMA and MF-CDMA systems can be found by Equation 46:

$$W = \frac{\left(\frac{\lambda}{N_{ch}}\right) \cdot (D_{TP})^2}{2 \cdot \left[1 - \left(\frac{\lambda}{N_{ch}}\right) \cdot D_{TP}\right]} + \frac{D_{TP}}{2}$$

Equation 46

The average amount of time a packet spends in the "system" (queuing plus service time per channel), can be found using:

$$T_{"system"} = E[x] + W$$

Equation 47

Substituting in D_{TP} for $E[x]$, and Equation 46 for W gives:

$$T_{"system"} = D_{TP} + \frac{\left(\frac{\lambda}{N_{ch}}\right) \cdot (D_{TP})^2}{2 \cdot \left[1 - \left(\frac{\lambda}{N_{ch}}\right) \cdot D_{TP}\right]} + \frac{D_{TP}}{2}$$

Equation 48

Rearranging,

$$T_{"system"} = \frac{\left(\frac{\lambda}{N_{ch}}\right)\cdot(D_{TP})^2}{2\cdot\left[1-\left(\frac{\lambda}{N_{ch}}\right)\cdot D_{TP}\right]} + \frac{3D_{TP}}{2}$$

Equation 49

The average number of packets in the "system" per channel can be found by:

$$N = \left(\frac{\lambda}{N_{ch}}\right)\cdot T_{"system"}$$

Equation 50

Plugging in Equation 49 into $T_{"system"}$,

$$N = \frac{\left(\frac{\lambda}{N_{ch}}\right)^2\cdot(D_{TP})^2}{2\cdot\left[1-\left(\frac{\lambda}{N_{ch}}\right)\cdot D_{TP}\right]} + \frac{3}{2}\cdot\left(\frac{\lambda}{N_{ch}}\right)\cdot D_{TP}$$

Equation 51

The average number of packets per MF-TDMA or MF-CDMA frame is:

$$N_{TF} = N_{ch}\cdot N$$

Equation 52

Combining Equation 52 with Equation 51 and Equation 44 gives:

$$N_{TF} = \frac{\lambda^2\cdot(D_{TP})^2}{2\cdot[N_{ch}-\lambda\cdot D_{TP}]} + \frac{3\cdot\lambda\cdot D_{TP}}{2}$$

Equation 53

3.3.13. Link Budget Design

The link budget design is based largely on the recommended downlink design procedure detailed in Wertz and Larson [Wert99], which is extended to include the uplink design as well as the link budget for the inter-satellite and terrestrial links. Additionally, some of the steps are avoided by virtue that some components are considered to be design variables for the purposes of this simulation model.

First, the carrier frequency is selected based on spectrum availability and FCC allocations. Table 3-5 shows the limitations on the allocation for frequency bands (Table 13-12 in [Wert99]).

Earlier in the discussion of the simulation model for the significance proof, it was assumed that the inter-satellite links should use a carrier frequency of 60 GHz. It is assumed that this allocation applies uniformly to all inter-satellite links in the constellation. Operating at this frequency shields the satellite-to-satellite communication from interference or jamming by signals originating on the earth. This is due to the high absorption band of oxygen at that frequency [Wert99].

Similarly, the uplink and downlink frequency allocations were assumed to be 30 GHz and 18 GHz, respectively. As the table shows, these frequencies are well within the frequency limitations of the Ka band.

Second, according to [Wert99], the satellite transmitter power P_T should be selected based on the satellite size and power limits. For the purposes of this simulation, the satellite transmitter power is assumed to be a design variable. The inter-satellite link power is assumed to be 2 kW.

Table 3-5: Frequency Allocation Limitations
The carrier frequency is selected based on spectrum availability and FCC allocations. This table gives the limitations on the allocation for frequency bands. Table modified from Table 13-12 in [Wert99].

Frequency Band	Frequency Range (GHz)		Service
	Uplink	Downlink	
UHF	0.2 - 0.45	0.2 - 0.45	Military
L	1.635 - 1.66	1.535 - 1.56	Maritime/Nav Telephone
S	2.65 – 2.69	2.5 - 2.54	Broadcast, Telephone
C	5.9 – 6.4	3.7 - 4.2	Domestic, Comsat
X	7.9 – 8.4	7.25 – 7.75	Military, Comsat
Ku	14.0 – 14.5	12.5 – 12.75	Domestic, Comsat
Ka	27.5 – 31.0	17.7 – 19.7	Domestic, Comsat
SHF/EHF	43.5 – 45.5	19.7 – 20.7	Military, Comsat
SHF/EHF	49	38	Internet Data, Telephone, Trunking
V	~60		Satellite Crosslinks

The optical link sizing is based on the Small Optical Telecommunications Terminal (SOTT), a test system implemented for GEO-to-GEO inter-satellite links (the antenna power for SOTT is 2 W, but the distance and data rate requirements are much less demanding than those for this thesis) [Opti05].

Third, radio frequency (RF) losses between the transmitter and satellite antennas should be estimated. [Wert99] notes that these values usually occur between -1 and -3 dB. Thus, it is assumed that the RF line loss L_l is approximately -1 dB. It is also assumed that a similar loss will be experienced by the optical inter-satellite links.

Normally, the required beamwidth for the satellite antenna is determined next. The beamwidth depends on the satellite orbit, satellite stabilization, and ground coverage area. However, since the satellite antenna diameter is specified as a design variable (or, in the case of the inter-satellite links, the optical antenna aperture is assumed to be 0.2 meters), it makes more sense to calculate the achievable beamwidth at the specified carrier frequency:

$$\theta_T = \frac{21}{f_{GHZ} D_T}$$

Equation 54

The maximum antenna pointing offset angle should be estimated based on the coverage angle, satellite stabilization error and desired stationkeeping accuracy. Based on Table 5-7 in [Wert99], the pointing error e is assumed to be about 0.06 degrees.

Given the achievable beamwidth for a given satellite antenna diameter, the transmit antenna gain toward the ground station can be calculated:

$$G = \left(\frac{70\pi}{\theta_T} \right)^2 \eta_T$$

Equation 55

where η_T is the transmitter illumination efficiency, assumed to be 0.55 (0.6 for the optical inter-satellite links). A pointing offset from the beam center will cause a reduction in the peak gain:

$$G_T = G - L_\theta$$

Equation 56

where L_θ can be found from the following relationship:

$$L_\theta = -12(e/\theta_T)^2$$

Equation 57

Thus, plugging in Equation 55 and Equation 57 into Equation 56, the transmitter gain in dB is determined to be:

$$G_T = \left(\frac{70\pi}{\theta_T}\right)^2 \eta_T - 12(e/\theta_T)^2$$

Equation 58

The simulation assumes a switch-feed parabolic antenna, so the uplink and downlink channels in a cell should achieve the same gain. Thus, the difference between the signal-to-noise ratio's per bit and the BER will be due to the effects of the different carrier frequencies.

The space loss in dB can be found using:

$$\begin{aligned} L_S &= 20\log(3x10^8) - 20\log(4\pi) - 20\log S - 20\log f \\ &= 147.55 - 20\log S - 20\log f \end{aligned}$$

Equation 59

The satellite orbit and ground-station locations determine the path length, S, in meters. The frequency f is in Hz.

Other losses include the propagation absorption loss due to the atmosphere; this loss can be estimated using Figure 13-10 in [Wert99] by finding the zenith attenuation. Once the attenuation factors are found, the propagation loss can be determined by dividing by the sine of the minimum elevation angle in radians as measured from the ground station to the satellite (this is assumed for simplicity to be the minimum elevation angle for the satellite regardless of ground station location). For the frequency allocations assumed in this thesis, the atmospheric propagation losses can be found in dB according to Equation 60 and Equation 61:

$$L_a^{18GHz} = \frac{-0.1}{\sin(elev)} \text{ dB}$$

Equation 60

$$L_a^{30GHz} = \frac{-0.3}{\sin(elev)} \text{ dB}$$

Equation 61

Furthermore, large ground antennas – assumed to be greater than 10 meters – will introduce a further loss of about 0.3 dB due to polarization mismatch:

$$L_P = -0.3 \text{ dB}$$

Another 1 dB loss can be expected if the system uses a radome:

$$L_R = -1 \text{ dB}$$

The ground station antenna diameter is also considered to be a design variable. It is assumed that the ground stations use autotracking, which introduces a pointing error offset from the beam center by approximately 10% of the beamwidth:

$$e = 0.1\theta_R$$
$$L_\theta = -12(0.1)^2 = -0.12 \text{ dB}$$

Equation 62

The beamwidth for the receive diameter (assumed to be the same aperture size as the transmitter for the inter-satellite links) can be found using:

$$\theta_R = \frac{21}{f_{GHZ} D_R}$$

Equation 63

The receive antenna gain toward the satellite (the simulation assumes a receiver illumination efficiency η_T of 0.55), is thus:

$$G_R = \left(\frac{70\pi}{\theta_R}\right)^2 \eta_R - 0.12 \text{ dB}$$

Equation 64

Before the signal-to-noise ratio can be calculated, the system noise temperatures in clear weather must be estimated. Using Table 13-10 in [Wert99], and assuming a linear relationship between 20 GHz and 40 GHz, the uplink system temperature is found to be:

$$T_{sys}^{30GHz} = 751.7K = 28.8 \text{ dB-K}$$

Equation 65

Similarly, for the downlink (assuming the value does not change appreciably between 18 GHz and 20 GHz):

$$T_{sys}^{18GHz} = 424K = 26.3 \text{ dB-K}$$

Equation 66

Finally, for the optical inter-satellite links:

$$T_{sys}^{60GHz} = 682K = 28.3 \text{ dB-K}$$

Equation 67

An estimate of the degradation due to rain can be incorporated into the system temperature calculations. Using Figure 13-11 in [Wert99], Table 3-6 gives the estimated losses due to rain attenuation; the inter-satellite links are not included since rain attenuation will not affect their performance.

Rain attenuation (RA) increases the antenna temperature according to the following law:

$$T_a = (1 - 10^{-RA}) T_o$$

Equation 68

where T_o, the temperature of the rain, is assumed to be 290 K, and RA is the magnitude of the rain attenuation in dB. The adjusted system temperatures can be found by adding this increase in temperature to the system noise temperatures calculated above:

$$\overline{T}_{sys} = T_{sys} + T_a$$

Equation 69

In general, the input parameters are adjusted until the link margin – the difference between the expected value of the calculated $\frac{E_b}{N_o}$ and the required $\frac{E_b}{N_o}$ – exceeds the estimated value for the rain degradation by at least 3 dB. Since the input parameters are for the most part fixed, it makes more sense to incorporate this 3 dB margin into the system temperature losses and analyze the achieved link margins later. Thus, the antenna temperature becomes:

$$T_a = (1 - 10^{-RA-3}) T_o$$

Equation 70

The signal-to-noise ratio per bit per link for the required data rate in bps can be found using [Wert99]:

$$\frac{E_b}{N_o}\Big|_{link} = 10\log_{10}(P_T) + L_l + G_T + L_p + L_R + L_S + L_a + G_R + 228.6 - 10\log_{10}(\overline{T_{sys}}) - 10\log_{10} R_{user}$$

Equation 71

Table 3-6: Estimated Loss Due to Rain Attenuation
Figure 13-11 in [Wert99] gives the estimated losses due to rain attenuation; the inter-satellite links are not included since rain attenuation will not affect their performance.

Frequency	Elevation Angle		
	10 deg	**15 deg**	**20 deg**
18 GHz	7 dB	5 dB	4 dB
30 Ghz	20 dB	15 dB	10 dB

The transmitter power is assumed to be in W and represents the total satellite transmitter power divided by the number of cells in the footprint divided by the number of channels per cell. This definition is required to accommodate the multiple access nature of the channels.

A further reduction of 1 dB in the signal-to-noise ratio per bit is allocated to account for implementation losses.

$$L_{imp} = -1 \text{ dB}$$

Given the $\frac{E_b}{N_o}$ seen over a given link and the set of routed paths, the achievable BER observed by a customer whose packet is sent along a given routed path can be calculated according to the Equation 72 in conjunction with the modulation scheme:

$$\left(\frac{E_b}{N_o}\right)^{-1}_{TotalPath} = \left(\frac{E_b}{N_o}\right)^{-1}_{link1} + \left(\frac{E_b}{N_o}\right)^{-1}_{link2} + \cdots + \left(\frac{E_b}{N_o}\right)^{-1}_{linkN}$$

Equation 72

Any link margin calculations can be used as a way of analyzing the sensitivity to the physical parameters. Comparison of the maximum BER observed by a customer in the network with the required BER of the system can be used to determine the feasibility of the system.

The link budget for the terrestrial system is designed a little bit differently. The bit energy is determined by the energy distance between two signals (see Section 3.3.11). Since this value can be varied easily in terrestrial networks, a probability of error is assumed for a single hop in the terrestrial network ($P_e = 10^{-10}$); this BER translates to an $\frac{E_b}{N_o}$ assumed to be achievable on every terrestrial link. This information is combined with the $\frac{E_b}{N_o}$ calculated per link calculated for the space segment to find the overall probability of error over a routed path.

3.3.14. Routing Protocols

The development of the routing protocols is based almost entirely on existing implementations. While the actual implementations almost certainly involve more than is discussed here, for the purposes of the simulation model, only the basic structure of the protocols is employed.

The simulation model used in the significance proof analyzes the effects of two different routing protocols. Routing protocols determine the paths packets travel over the network and guarantee the correct delivery of messages once the routes are chosen. Under high load conditions, good routing protocols increase the observed throughput – offered load minus rejected load – for a fixed value of average packet delay and decrease the average packet delay under low and moderate load conditions. Routing protocols that increase the observed throughput usually incorporate some flow control measures to control the flow of packets into

the network; due to time and complexity issues, flow control is considered beyond the scope of this thesis.

For both routing protocols, as long as there is connectivity on the link (i,j) and there is some capacity with which to carry packets on the link, then the link (i,j) is assigned a cost based on the metric specified by the routing protocol. Otherwise, the link cost is set to infinite in order to keep the link from being routed over.

3.3.14.1. Minimize the Number of Hops

Minimizing the number of hops required to transmit a packet from its source to its destination is the metric used by the IP RIP protocol. The actual protocol is limited to 16 hop networks. Thus, the model used in the simulation assumes that the number of hops only includes communications between subnet gateways.

The link cost for IP RIP is equal to 1 for every link in the path. Thus, for every communication link a packet travels over, the hop count is incremented by one. This relationship is best expressed by the adjacency matrix, which automatically accounts for connectivity on the link:

$$LinkCost = Adjacency(i, j)$$

Equation 73

3.3.14.2. *Minimize the Delay and Maximize the Capacity*

This protocol is similar to IGRP, a CISCO proprietary routing protocol with a standard equation for calculating the link cost. The standard routing metric equation considers the bandwidth BW (actually, the inverse is used in kbps multiplied by a factor of 1e7), the delay in units of 10 microseconds, the load – merely a weighting measure from 1 to 255 for which a higher number is less appealing, and the reliability in fractions of 255 for which 255 means perfectly reliable [IGRP01].

$$LinkCost = (K_1 \cdot BW) + \frac{(K_2 \cdot BW)}{(256 - Load)} + (K_3 \cdot Delay) \cdot \frac{K_5}{(\mathrm{Re}liability + K_4)}$$

Equation 74

The two measures most straightforward to calculate are the bandwidth (capacity) and the delay (in this case, it is assumed to mean channel delay). However, accounting for both metrics in the above equation requires some measure of the reliability of each link. A decent reliability metric would be a scaling factor based on the probability of bit error observed by the customer, but given the strongly coupled relationship between the routed path and the overall probability of bit error, incorporating this measure of reliability into the link cost would require an inner loop of calculations within the already computationally-expensive simulation model. For this reason, a

simplified version of the IGRP protocol was used, favoring links which maximize the capacity while minimizing the channel delay:

$$LinkCost = \frac{\frac{Capacity(i,j)}{\max(\max(Capacity))}}{\frac{ChannelDelay(i,j)}{\max(\max(ChannelDelay))}}$$

Equation 75

3.3.15. Network Overhead

The network overhead calculations contained in this section were, for the most part, conceived for the purposes of this thesis. Some of the intermediate calculations are based on examples from [Modi04], and the values used for most of the network overhead components are based on actual network layer implementations.

Network control overhead accounts for all of the encapsulation information necessary to interface between the ISO-OSI internet layers. Figure 30 shows the structure of the internet layers for a "bent-pipe" satellite system like Globalstar, while Figure 31 illustrates the structure of the internet layers for an inter-satellite based system like Iridium. The number of data bytes required to manage the network interfaces depends on the internet layer under consideration and the desired implementation.

The effect of the transport, session, presentation, and application layers will be ignored for the purposes of this thesis.

First, the network layer overhead will be estimated. The network layer governs the routing protocols and network-level flow control. Figure 32 depicts how the network overhead is implemented at each of the lowest three layers of the internet stack. The network layer takes the packet encapsulated by the higher layers and adds its own header information. Since this simulation considers only the effects of routing, only the network header $OH_{Header}^{Network}$ bytes from the routing protocols need be estimated. Fortunately, information on number of header bytes used in existing routing protocols can readily be found; the RIP routing protocol standard incorporates 24 header bytes while the IGRP standard has 26 ([InTH05], and [IGRP01], respectively).

The Link Layer Control (LLC) sublayer of the Data Link Control (DLC) internet layer increments the overhead from the network layer and above with address $OH_{Address}^{LLC}$ and control $OH_{Control}^{LLC}$ bytes. The typical value for the LLC address is 2 bytes. The number of necessary control bytes is determined by the Automatic Repeat Request (ARQ) protocol [DLLP05].

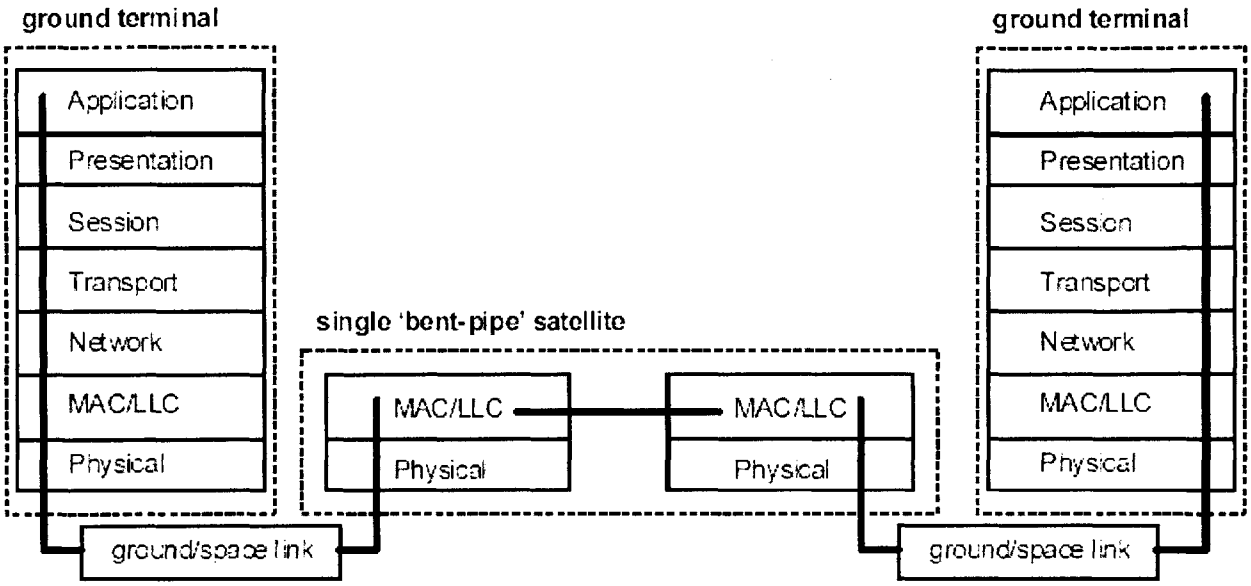


Figure 30: ISO-OSI Network Layer Model for Bent-Pipe Satellite Systems
Picture taken from [Zhan03].

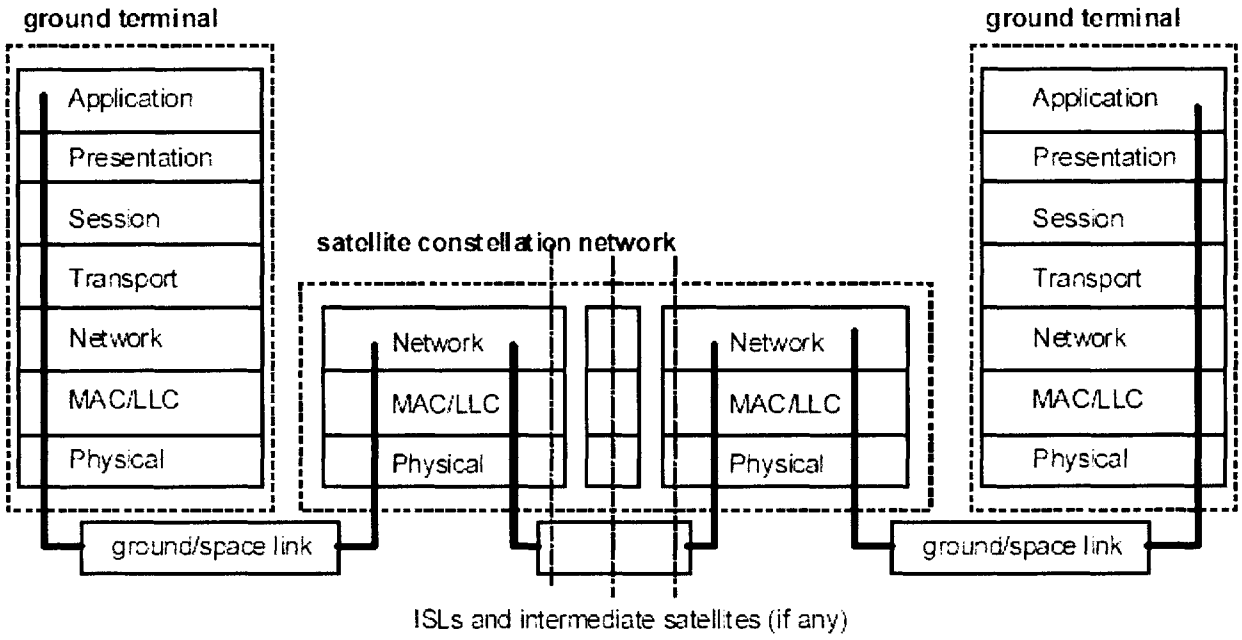


Figure 31: ISO-OSI Network Layer Model for Satellite Systems using ISLs
Picture taken from [Zhan03].

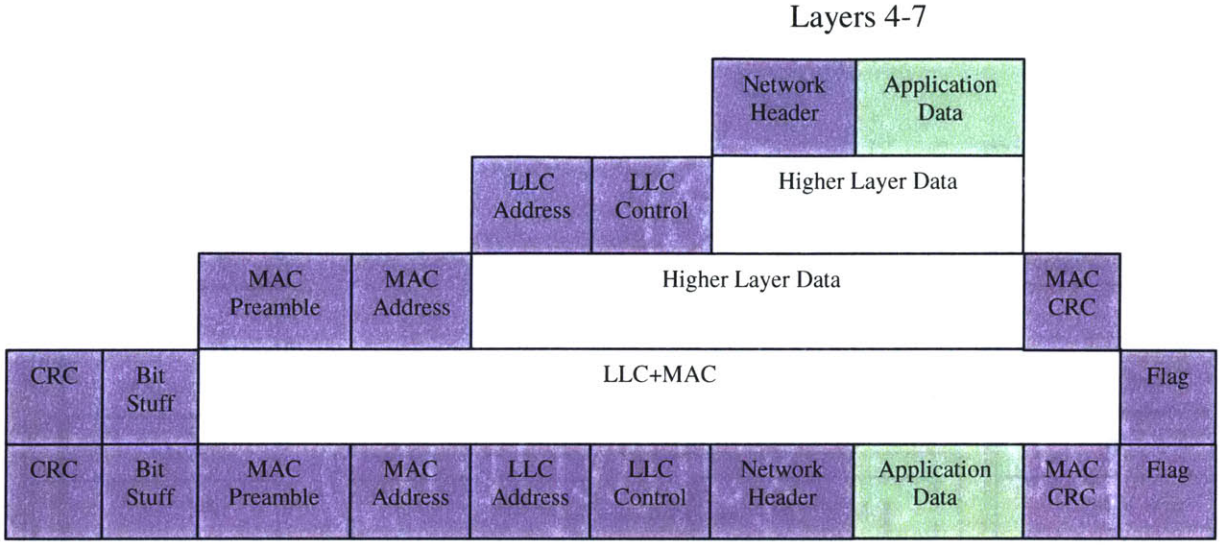


Figure 32: Representation of Overhead in the Lowest 3 ISO-OSI Layers
The network layer takes the packet encapsulated by the higher layers – application data – and adds its own header information. The Link Layer Control (LLC) sublayer of the Data Link Control (DLC) internet layer increments the overhead from the network layer and above with address and control bytes. The Media Access Control (MAC) sub-layer of the DLC internet layer further increments the overall overhead with a preamble enabling synchronization among users, an address header, and a cyclic redundancy check (CRC). Once the DLC layer has finished adding on its required overhead, then the overall packet structure must be prepared for transmission over the physical channel. This preparation is done by stuffing bits into the packet to avoid duplication of the packet stop flags as well as framing the packet with the necessary flags. The frame itself includes a CRC to protect against errors in the frame

To estimate the number of control bytes required by an ARQ protocol, it is necessary to understand a little bit about how ARQ protocols work. Depending on the implementation of the ARQ, the protocol either automatically requests a duplicate packet if a packet is corrupted or if an expected packet fails to appear, or it acknowledges correctly received packets (if the sender fails to receive an acknowledgement than an error is assumed to have occurred and the packet is re-sent). The actual implementation does not matter since they are equivalent. For the purposes of this discussion, it is assumed that acknowledgements are used. The amount of time required for a packet to be transmitted and acknowledged can be found by estimating the time the packet takes to travel from source S to destination D (see Figure 33) and the amount of time the acknowledgement or request takes to travel from the destination D to the source S.

Clearly, the amount of time to travel between the source and destination is accounted for in the packet transmission time D_{TP} and the propagation time D_P. The time the acknowledgement takes to travel between the destination and the source is similarly described by the acknowledgement transmission time D_{TA} and the propagation time D_P. Adding these times together gives the approximate amount of time, S, for a packet to be transmitted and acknowledged:

$$S = D_{TP} + 2 \cdot D_P + D_{TA}$$

Equation 76

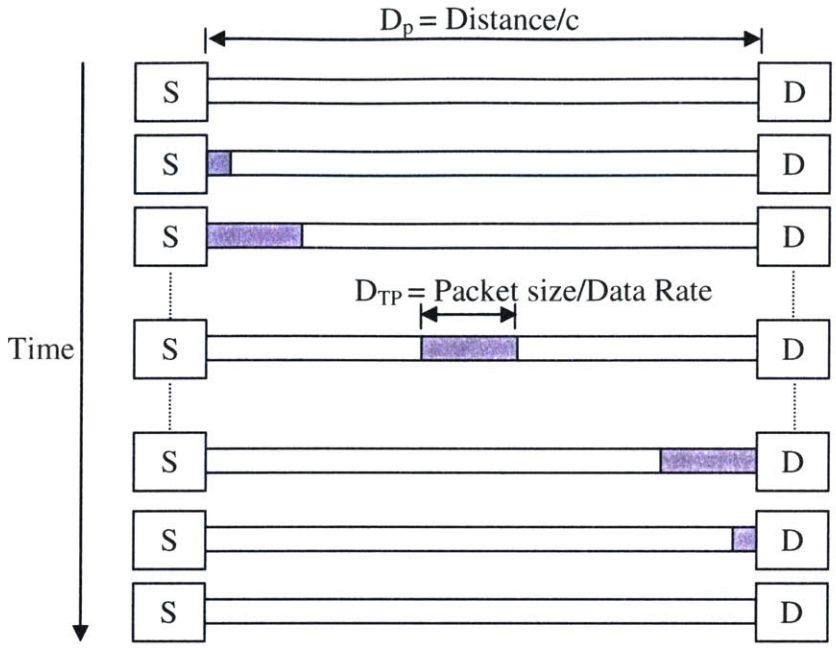


Figure 33: Diagram of Packet Traversal from Source S to Destination D
The amount of time required for a packet to be transmitted and acknowledged can be found by estimating the time the packet takes to travel from source S to destination D and the amount of time the acknowledgement or request takes to travel from the destination D to the source S.

If the acknowledgement is assumed to piggyback on top of a packet traveling in the opposite direction (a common implementation when traffic is bi-directional), Equation 76 becomes:

$$S = 2 \cdot (D_{TP} + D_P)$$

Equation 77

Most ARQ protocols incorporate sliding windows that allow new packets that fall within the window set of packets to be transmitted while waiting for earlier ones to be acknowledged; the window advances upon the arrival of acknowledgments for previous packets. The window size N should allow for continuous transmission of packets even if the first packet of the window has yet to be acknowledged. As can be seen in Figure 34, continuous transmission is enabled so long as Equation 78 holds true.

$$N > \frac{S}{D_{TP}}$$

Equation 78

This condition is satisfied if we assume:

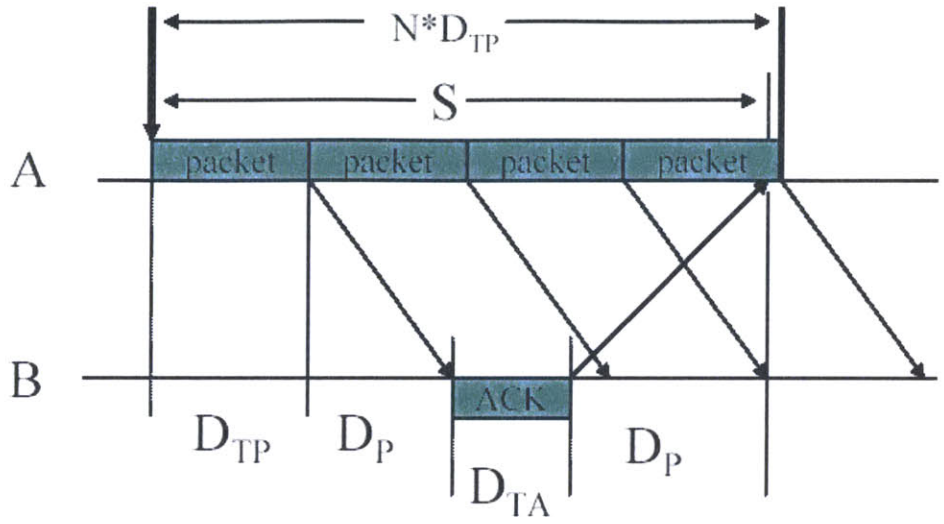


Figure 34: Illustration of the Condition on Window Size N
This condition on the window size N allows continuous transmission of packets while waiting for the first packet's acknowledgement. Picture taken from [Modi04].

$$N = \left\lfloor \left(\frac{S}{D_{TP}} \right) + 1 \right\rfloor$$

Equation 79

Assuming packets do not get out of order, this estimate of the window size guarantees correctness of the algorithm. In the event of an error, the entire window of N packets must be retransmitted; thus, making N large is advantageous only if the probability that a packet experiences an error is much less than the probability that the acknowledgement experiences an error. Since these probabilities are the same – the packets are assumed to travel the same route through the network and the acknowledgement is piggybacked on another packet – the worst-case congestion is expected.

$\frac{S}{D_{TP}}$ is found by dividing Equation 77 by D_{TP}:

$$\frac{S}{D_{TP}} = 2\left(1 + \frac{D_P}{D_{TP}}\right)$$

Equation 80

The propagation delay can be found by Equation 81.

$$D_P = \frac{Dis\tan ce_{Max}}{c}$$

Equation 81

The transmission delay is given by Equation 82:

$$D_{TP} = \frac{Pkt_{user}}{R_{user}}$$

Equation 82

.

To estimate the maximum total time for a packet to be transmitted and acknowledged, the minimum packet size, maximum distance (worst-case routing), and maximum data rate is assumed. These requirements translate to assuming the maximum distance $Dis\tan ce_{Max}$ in meters, the speed of light c in meters/sec, R_{user} in bps and Pkt_{user} in bits. Plugging these quantities into Equation 80:

$$\frac{S}{D_{TP}} = 2\left(1 + \frac{Dis\tan ce_{Max} \cdot R_{user}}{c \cdot Pkt_{user}}\right)$$

Equation 83

Next, it is necessary to derive how the control overhead relates to the choice of window size for the two ARQ protocols of interest: Go-Back-N and SRP.

In Data Networks [Bert92], it is shown that the correctness of the Go-Back-N algorithm is maintained so long as the sequence and request numbers for the original packets and the acknowledgement packets, respectively, are sent modulo m, where m is strictly greater than the window size N. As discussed above, this statement is true so long as the packets do not get out of order. Thus, in order to guarantee that the sequence and request numbers are sent modulo m, the control information in the LLC header for implementing Go-Back-N must be of the form:

$$OH_{Control}^{LLC} = \lceil \log_2 m \rceil$$

Equation 84

The constraint that m must be strictly greater than the window size N gives:

$$OH_{Control}^{LLC} = \lceil \log_2 (N+1) \rceil$$

Equation 85

For Selective Repeat (SRP), any given packet may be followed by the first packet of the window or it may be followed by the last packet in the window. Thus, the destination receiver must be able to distinguish between twice the total number of packets in a window; the modulus m must now conform to:

$$m \geq 2N$$

Equation 86

Thus, the control overhead for SRP must satisfy:

$$OH_{Control}^{LLC} = \lceil \log_2(2 \cdot N) \rceil$$

Equation 87

Plugging in for the estimations of the required window size gives:

$$OH_{Control}^{LLC} = \left\lceil \log_2\left(\left\lfloor \frac{S}{D_{TP}} \right\rfloor + 2 \right) \right\rceil$$

Equation 88: Go-Back-N

$$OH_{Control}^{LLC} = \left\lceil \log_2\left(2 \cdot \left(\left\lfloor \frac{S}{D_{TP}} \right\rfloor + 1 \right) \right) \right\rceil$$

Equation 89: SRP

The Media Access Control (MAC) sub-layer of the DLC internet layer further increments the overall overhead with a preamble $OH_{Preamble}^{MAC}$ enabling synchronization among users (typically 8 bytes), an address header $OH_{Address}^{MAC}$ (typically 14 bytes in Ethernet, but assumed to be 8 bytes for the simulation), and a cyclic redundancy check (CRC) of 4 bytes OH_{CRC}^{MAC} to enable detection of errors within the packet as encapsulated by the MAC sub-layer.

Once the DLC layer has finished adding on its required overhead, then the overall packet structure must be prepared for transmission over the physical channel. This preparation is done by stuffing bits into the packet to avoid duplication of the packet stop flags as well as framing the packet with the necessary flags. The frame itself includes a CRC $OH_{CRC}^{Framing}$ to protect against errors in the frame (assumed to be the same size of the MAC CRC, 4 bytes).

To estimate the number of framing bits, it's necessary to assume that the packet before the framing process consists of independent, identically distributed random binary variables. These binary variables are assumed to have an equal probability of being a 0 or a 1. Furthermore, the termination flag for a frame is assumed to be of the form 01^j0 for some j.

The integer value of j that minimizes the expected overhead for a given original frame length in bits ($E\{K\}$) turns out to be (from [Bert92]):

$$j = \lfloor \log_2 E\{K\} \rfloor$$

Equation 90

The difference in the value of j with $E\{K\}$ representing the packet with all of the additional overhead and $E\{K\}$representing the original user packet is very small. Thus, for a rough approximation, assume $E\{K\}$ is the average user packet size:

$$j = \lfloor \log_2 Pkt_{user} \rfloor$$

Equation 91

With this value of j, the number of framing bits required is estimated to be:

$$OH_{Flags}^{Framing} = \left| E[K] \cdot 2^{-j} + j + 1 \right|$$

Equation 92

$$OH_{Flags}^{Framing} = \left| Pkt_{user} \cdot 2^{-j} + j + 1 \right|$$

Equation 93

Once all of these overhead contributions have been estimated, the overall number of bits comprising the network overhead can be calculated as follows:

$$OH = \left| OH_{Header}^{Network} + OH_{Address}^{LLC} + OH_{Control}^{LLC} + OH_{Preamble}^{MAC} + OH_{Address}^{MAC} + OH_{CRC}^{MAC} + OH_{Flags}^{Framing} + OH_{CRC}^{Framing} \right|$$

Equation 94

3.4. Derivation of Significance Proof Performance Metrics

Due to the computational expense of this Matlab simulation model, the performance metrics are examined for a single snapshot in time for each topology; there are 3 topologies under consideration. Furthermore, to minimize the computational complexity as much as possible, it is assumed that the network performance metrics are calculated as steady-state averages.

The following discussion derives the performance metrics in terms of the models discussed in Section 3.3.

3.4.1. Cost/User/Month

The calculation of the cost per user per month metric follows the argument for a cost per function (CPF) metric used by Kashitani [Kash02]. The derivation is adapted to include accounting of some of the design variables unique to this simulation model, and the subscription fees are found per month rather than per year.

The user subscription charge for network service is calculated in terms of year 2005 constant dollars. It is assumed that this subscription charge $C_{\text{Subscription\_2005}}$ is held constant during the operational period. This subscription cost must be adjusted each year to account for the inflation rate $i_{\text{inflation}}$ (assumed to be 2.4% [Infl05]), giving the subscription cost in nominal dollars $C_{\text{Subscription\_Nominal}}$, the actual yearly cost seen by the customer. Thus, the subscription cost in nominal dollars can be found as a function of the year y in the operational period:

$$C_{Subscription\_Nominal}(y) = C_{Subscription\_2005} \cdot (1 + i_{Inflation})^{y-2005}$$

Equation 95

Discounting the nominal charge by the internal rate of return i_{IRR} every year results in the net present value subscription charge, $C_{Subscription\_NPV}$:

$$C_{Subscription\_NPV}(y) = \frac{C_{Subscription\_Nominal}(y)}{(1 + i_{IRR})^{y-2005}}$$

Equation 96

Combining Equation 95 and Equation 96 gives:

$$C_{Subscription\_NPV}(y) = C_{Subscription\_2005} \cdot \left(\frac{1 + i_{Inflation}}{1 + i_{IRR}} \right)^{y-2005}$$

Equation 97

Once the life cycle cost has been estimated, the yearly subscription charge can be adjusted to guarantee that the net present value of the system will be zero. Without any internal rate of return, this value corresponds to the system breaking even on the investment by the end of the anticipated system lifetime. With an internal rate of return of 30%, the system will actually gain 30% on the investment by achieving zero net present value by the end of life.

It is assumed that the only revenue to the system is from individual customer subscription fees; another potential source could be from advertising. Furthermore, it is assumed that the satellite system is launched in 2010 and that the system lifetime is specified by the design variable T_{life}. Thus, the net present value of the revenue R_{NPV} can be found by summing the yearly subscription revenue – found by multiplying the net present value of the subscription cost by the estimated total number of subscribers $N_{Subscribers}(y)$ for that year – over the lifetime of the system.

$$R_{NPV} = \sum_{y=2010}^{2010+Tlife-1} C_{Subscription\_NPV}(y) \cdot N_{Subscribers}(y)$$

Equation 98

Substituting in the previous expression for $C_{Subscription\_NPV}$, Equation 97:

$$R_{NPV} = C_{Subscription\_2005}\left[\sum_{y=2010}^{2010+Tlife-1}\left(\frac{1+i_{Inflation}}{1+i_{IRR}}\right)^{y-2005} \cdot N_{Subscribers}(y)\right]$$

Equation 99

To achieve zero net present value, the net present value of the system revenue must equal the total system lifecycle cost (LCC) in net present value LCC_{NPV}. The yearly subscription cost in 2005 dollars per user can be found by replacing R_{NPV} with LCC_{NPV} in the above equation and rearranging to solve for $C_{Subscription\_2005}$:

$$C_{Subscription\_2005} = \frac{LCC_{NPV}}{\left[\sum_{y=2010}^{2010+Tlife-1}\left(\frac{1+i_{Inflation}}{1+i_{IRR}}\right)^{y-2005} \cdot N_{Subscribers}(y)\right]}$$

Equation 100

This equation can be rewritten as:

$$C_{Subscription\_2005} = \frac{LCC_{NPV}}{\left[\sum_{i=1}^{Tlife}\left(\frac{1+i_{Inflation}}{1+i_{IRR}}\right)^{(2010+i-1)-2005} \cdot N_{Subscribers}(2010+i-1)\right]}$$

Equation 101

Finally, the cost per user per month can be found by dividing by the number of months in a year.

$$C_{Subscription\_Month} = \frac{LCC_{NPV}}{12 \cdot \left[\sum_{i=1}^{Tlife}\left(\frac{1+i_{Inflation}}{1+i_{IRR}}\right)^{(2010+i-1)-2005} \cdot N_{Subscribers}(2010+i-1)\right]}$$

Equation 102

3.4.2. Market Potential

The following derivations are based off of the work of Chang [ChaD04]. The major divergence is that this thesis keeps track of the market potential as a function of the year of operation, rather than considering only one value for the number of subscribers.

The simulation model assumes that the satellite system as designed satisfies the demand of one percent of the potential broadband satellite system market (also known as the satellite market

fraction, *SMF*). The simulation will not separately model the telephony or radio/TV media markets.

Either the market is modeled off of low-bandwidth systems, defined by user data rates of less than 50 kbps, or the market is modeled off of high-bandwidth systems, in which the user data rate is greater than or equal to 50 kbps. The data is based off of the Globalstar system and its experience with penetrating the telephony market. The population data is in units of 1 million.

The initial system development time *IDTime* is assumed to be 5 years.

3.4.2.1. Low-Bandwidth

In 2003, the number of potential system users is assumed to be 49.6 million users. Since the industry has a tendency to overestimate market demand, the model was shifted over to 2004 and started from the same initial number. The annual increment is 2.9677 million users.

$$N_{Subscriber}(2010) = SMF \cdot (49.60 + 2.9677 \cdot (IDTime + 1)) \cdot 1e6$$

Equation 103

Converting Equation 103 to account for the life time of the system (which is a design variable), gives Equation 104.

$$N_{Subscriber}(2010 + Tlife - 1) = SMF \cdot \sum_{i=1}^{Tlife} \left[N_{Subscribeer}(2010 + i - 1) + 2.9677 \cdot 1e6 \right]$$

Equation 104

3.4.2.2. High-Bandwidth

The number of potential users in 2003 is assumed to be 4.9158 million people, with an annual increment of 0.5159 million. Again, the model is shifted by one year to account for overestimation of the demand.

$$N_{Subscriber}(2010) = SMF \cdot (4.9158 + 0.5159 \cdot (IDTime + 1)) \cdot 1e6$$

Equation 105

$$N_{Subscriber}(2010 + Tlife - 1) = SMF \cdot \sum_{i=1}^{Tlife} \left[N_{Subscribeer}(2010 + i - 1) + 0.5159 \cdot 1e6 \right]$$

Equation 106

Fortunately, the exact numbers are not that important for this thesis. The relative growth in subscriber base and the impact of the number of subscribers on the other performance metrics is what matters.

3.4.3. Life Cycle Cost

The life cycle cost metric derivation is nearly the same as that used by Chang [ChaD04]; the differences are detailed in the cost model section.

3.4.4. Unused Capacity

The unused capacity metric was developed for this thesis, based on the concepts in network and queuing theory.

To develop an analytical model for the unused capacity, more queuing theory must be considered. For stability, the average packet arrival rate to the system λ must be less than the average service rate μ. Otherwise, the system cannot keep up with the arrivals to the system and the delay increases without bound. The stability condition is:

$$\lambda < \mu$$

Equation 107

The average service rate μ is also the rate at which a packet is transmitted, or served to use the language of queuing theory. If D_{TP} is the packet transmission time, then Equation 107 becomes:

$$\lambda < \frac{1}{D_{TP}}$$

Equation 108

However, to find the steady-state average number of simultaneous packets, it is necessary to consider the arrival rate to a single multiple-access channel on the link, not the arrival rate of packets to the entire link. If N_{Cap} is the total number of channels on a link, then:

$$\frac{\lambda}{N_{Cap}} < \frac{1}{D_{TP}}$$

Equation 109

To guarantee stability on any link (i,j), let:

$$\frac{\lambda_{Cap}^{ij}}{N_{Cap}^{ij}} = \frac{0.99}{D_{TP}}$$

Equation 110

Packets can exist in queue waiting for service along a link, in service at the transmitter of a link, or in transit on a link between nodes in the network. Accounting for all of these packets requires recalling the average number of packets per MF-TDMA or MF-CDMA time frame developed in the multiple access section:

$$N_{TF} = \frac{\lambda^2 \cdot (D_{TP})^2}{2 \cdot [N_{ch} - \lambda \cdot D_{TP}]} + \frac{3 \cdot \lambda \cdot D_{TP}}{2}$$

Equation 111

This equation, combined with the number of packets in transit along a link, and summed over all possible links in the network, can be used to find the number of simultaneous packets at capacity.

$$SimPkts^{Cap} = \sum_{i,j \in links} \left\{ \left[\frac{\lambda_{Cap}^{ij} \cdot Distance^{ij}}{c} \right] + \left[\frac{3 \cdot \lambda_{Cap}^{ij} \cdot D_{TP}}{2} + \frac{(\lambda_{Cap}^{ij} \cdot D_{TP})^2}{2 \cdot (N_{Cap}^{ij} - 0.99 \cdot N_{Cap}^{ij})} \right] \right\}$$

Equation 112

Rearranging:

$$SimPkts^{Cap} = \sum_{i,j \in links} \left\{ \left[\frac{0.99 \cdot N_{Cap}^{ij} \cdot Distance^{ij}}{c \cdot D_{TP}} \right] + \left[\frac{3 \cdot 0.99 \cdot N_{Cap}^{ij}}{2} + \frac{(0.99 \cdot N_{Cap}^{ij})^2}{0.02 \cdot N_{Cap}^{ij}} \right] \right\}$$

Equation 113

$$SimPkts^{Cap} = \sum_{i,j \in links} \left\{ \left[\frac{0.99 \cdot N_{Cap}^{ij} \cdot Distance^{ij}}{c \cdot D_{TP}} \right] + 50.49 \cdot N_{Cap}^{ij} \right\}$$

Equation 114

Similarly, assuming each subscriber transmits one packet per second:

$$SimPkts = \sum_{i,j \in links} \left\{ \left[\frac{N_{Subscriber} \cdot \lambda^{ij} \cdot Distance^{ij}}{c} \right] + \left[\frac{3 \cdot N_{Subscriber} \cdot \lambda^{ij} \cdot D_{TP}}{2} + \frac{(N_{Subscriber} \cdot \lambda^{ij} \cdot D_{TP})^2}{2 \cdot (N_{Cap}^{ij} - N_{Subscriber} \cdot \lambda^{ij} \cdot D_{TP})} \right] \right\}$$

Equation 115

The unused capacity is defined as follows:

$$UnusedCap = SimPkts^{Cap} - SimPkts$$

Equation 116

Substituting in known quantities:

$$UnusedCap = \sum_{i,j \in links} \left\{ \left[\frac{0.99 \cdot N_{Cap}^{ij} \cdot Dis\tan ce^{ij}}{c \cdot D_{TP}} \right] + 50.49 \cdot N_{Cap}^{ij} \right\}$$
$$- \left\{ \left[\frac{N_{Subscriber} \cdot \lambda^{ij} \cdot Dis\tan ce^{ij}}{c} \right] + \left[\frac{3 \cdot N_{Subscriber} \cdot \lambda^{ij} \cdot D_{TP}}{2} + \frac{(N_{Subscriber} \cdot \lambda^{ij} \cdot D_{TP})^2}{2 \cdot (N_{Cap}^{ij} - N_{Subscriber} \cdot \lambda^{ij} \cdot D_{TP})} \right] \right\}$$

Equation 117

3.4.5. Simultaneous Users

The number of simultaneous user metric was developed for this thesis based on the concepts in network and queuing theory.

The number of simultaneous users is dependent on the number of simultaneous packets (at capacity and otherwise), as well as the average round-trip delay in the system. The reason for describing these metrics in this fashion is simple: the network will obey Little's Theorem [Bert92]:

$$N = \lambda T$$

Equation 118

Where N is the steady-state average number of users (customers, packets, etc)
λ is the arrival rate of users to the system
T is the average time spent in the system.

Thus, the number of simultaneous users in the system can be found by considering the total arrival rate of packets to the system λ_{Total} and the average round-trip delay experienced by a packet in the system $\overline{RTD}$:

$$SimPkts_{Total} = \lambda_{Total} \cdot \overline{RTD}$$

Equation 119

Since each subscriber is assumed to have an average arrival rate λ_{User}, the total arrival rate of packets to the system can by found by using the Poisson merging property [Modi04]:

$$\lambda_{Total} = \sum_{i=1}^{Nuser} \lambda_{User}$$

Equation 120

Equation 120 is equivalent to Equation 121.

$$\lambda_{Total} = N_{User} \cdot \lambda_{User}$$

Equation 121

The average rate at which a user can transmit a packet is $\frac{1}{D_{TP}}$; D_{TP} is the packet transmission time. However, the user does not transmit constantly all day long. This means this value must be scaled accordingly. Assuming each user transmits, on average, one packet per second:

$$\lambda_{User} = 1$$

Plugging Equation 121 with $\lambda_{User} = 1$ into Equation 119:

$$SimPkts_{Total} = Nuser \cdot \overline{RTD}$$

Equation 122

Rearranging,

$$Nuser = \frac{SimPkts_{Total}}{\overline{RTD}}$$

Equation 123

Similarly, the number of simultaneous users at capacity can be found with Equation 124:

$$Nuser = \frac{SimPkts_{Total}^{Cap}}{\overline{RTD}}$$

Equation 124

3.4.6. Spectral Efficiency

The definition of spectral efficiency comes from [Inte02], and was extended to account for differences in the MF-TDMA and MF-CDMA systems.

In general, spectral efficiency is defined as follows [Inte02]:

$$SE = \frac{ChannelThroughput}{ChannelBandwidth}$$

Equation 125

From the previous multiple access models, the channel bandwidth is:

$$B_T = \frac{(1+\beta)\cdot(Pkt_{user} + OH)\cdot R_{user}}{(\log_2 M)\cdot Pkt_{user}}$$

Equation 126

The channel throughput corresponds to the billable (user) data throughput of each TDMA or CDMA carrier, taking into account the frequency reuse brought about by reusing frequencies in neighboring cells (cluster size). The data throughput of each carrier can be found as follows:

$$ChannelThroughput = N_{hd} \cdot R_{user} \cdot K$$

Equation 127

where N_{hd} is the number of TDMA or CDMA time slots per carrier (see Section 3.3.12), R_{user} is the user data rate in bps, and K is the cluster size. Thus, using Equation 123 gives:

$$SE = \frac{N_{hd} \cdot R_{user} \cdot K}{B_T}$$

Equation 128

Plugging in for known quantities, the spectral efficiency for the MF-TDMA system becomes:

$$SE_{MF-TDMA} = \frac{\left(\frac{R_{user}}{Pkt_{user}}\right)T_f}{1+\left(\frac{R_{user}}{Pkt_{user}}\right)T_g}\cdot(R_{user})\cdot\left(\frac{K}{B_T}\right)$$

Equation 129

The number of CDMA channels can be found by:

$$N_{hd} = 1 + \frac{B_{TOT}\cdot Pkt_{user}}{R_{user}\cdot(Pkt_{user}+OH)}\cdot\frac{1}{\alpha}\cdot\left[\left(\frac{E_b}{I_{tot}}\right)^{-1} - \left(\frac{E_b}{N_o}\right)^{-1}\right]$$

Equation 130

Plugging into the spectral efficiency equation gives the spectral efficiency for the MF-CDMA system:

$$SE = \frac{1 + \frac{B_{TOT} \cdot Pkt_{user}}{R_{user} \cdot (Pkt_{user} + OH)} \cdot \frac{1}{\alpha} \cdot \left[\left(\frac{E_b}{I_{tot}} \right)^{-1} - \left(\frac{E_b}{N_o} \right)^{-1} \right] \cdot R_{user} \cdot K}{\frac{(1+\beta) \cdot (Pkt_{user} + OH) \cdot R_{user}}{(\log_2 M) \cdot Pkt_{user}}}$$

Equation 131

3.4.7. Data Loss

The data loss metric was developed for this thesis. Derivations for the expected number of extra packets due to corrupted packets were adapted from [Modi04].

The data loss measures the amount of data lost in the system due to error. Error sources include noise that causes data corruption and network congestion resulting in dropped packets, both of which require retransmission of packets. This section outlines the development of the data loss estimation and discusses the data loss performance metric based on the idea of overhead efficiency.

The probability that a packet is dropped in the satellite network can be estimated by finding the ratio of the number of dropped packets and the total number of packets attempted or successfully pushed through the system.

$$P(DroppedPkt) = \frac{DroppedPkts}{DroppedPkts + SimPkts_{Total}}$$

Equation 132

It is assumed that the error sources – noise and congestion – occur independently, and that each bit is equally susceptible to the total bit error rate over the path BER_{TOT}. The probability of packet error over a given routed path can be estimated given the total bit error rate over the path, the total number of bits in a packet including overhead Pkt_{Total}, and the probability that a packet is dropped due to congestion. Considering just the error due to noise:

$$P(PktError) = \sum_{i=1}^{Pkt_{Total}} \binom{Pkt_{Total}}{i} BER_{TOT}{}^{i} (1 - BER_{TOT})^{Pkt_{Total} - 1}$$

Equation 133

However, finding the probability of packet error this way is very difficult. Instead, use:

$$P(NoPktError) = \binom{Pkt_{Total}}{0} BER_{TOT}{}^{0} (1 - BER_{TOT})^{Pkt_{Total}}$$

Equation 134

$$P(NoPktError) = (1 - BER_{TOT})^{Pkt_{Total}}$$

Equation 135

$$P(PktError) = 1 - P(NoPktError)$$

Equation 136

$$P(PktError) = 1 - (1 - BER_{TOT})^{Pkt_{Total}}$$

Equation 137

Accounting for errors due to congestion:

$$P(PktError) = \left\{1 - \left[(1 - BER_{TOT})^{Pkt_{Total}}\right]\right\} + P(DroppedPkt)) - \left\{\left[1 - \left[(1 - BER_{TOT})^{Pkt_{Total}}\right]\right] \cdot P(DroppedPkt)\right\}$$

Equation 138

The ARQ protocols determine how packets are retransmitted. As such, the efficiency with which data is successfully sent through the system depends greatly on the ARQ implementation. Fortunately, it is easy to calculate the number of extra packets that an ARQ protocol is expected to generate.

Let x = the number of times a packet is sent before it is successfully received. For the Go-Back-N protocol, given the probability of packet error p (based on [Modi04]):

$$x = \begin{cases} 1, & if \quad no \quad error \quad (1-p) \\ x + N, & if \quad error \quad (p) \end{cases}$$

Equation 139

The expected value of x becomes:

$$E[x] = 1 \cdot (1 - p) + (x + N) \cdot p$$

Equation 140

Rearranging:

$$E[x] = 1 + \left(\frac{N \cdot p}{1 - p}\right)$$

Equation 141

Thus, the number of extra packets required for the Go-Back-N protocol is found by:

$$ExtraPkts = \frac{N \cdot p}{1-p}$$

Equation 142

Substituting in for known values:

$$ExtraPkts = \left\lceil 2\left(1 + \frac{Distance \cdot R_{user}}{c \cdot Pkt_{Total}}\right) \cdot \left(\frac{P(PktError)}{1 - P(PktError)}\right) \right\rceil$$

Equation 143

A similar procedure exists to find the number of extra packets for the SRP protocol. Given the probability of packet error p (based on [Modi04]):

$$x = \begin{cases} 1, if \quad no \quad error \quad (1-p) \\ i, if \quad error \quad p^{i-1} \cdot (1-p) \end{cases}$$

Equation 144

The expected number of extra packets thus becomes:

$$E[x] = 1 \cdot (1-p) + 2 \cdot p \cdot (1-p) + 3 \cdot p^2 \cdot (1-p) + \cdots + i \cdot p^{i-1} \cdot (1-p) + \cdots$$

Equation 145

This expression can be simplified to:

$$E[x] = \frac{1}{1-p}$$

Equation 146

Therefore,

$$ExtraPkts = \left\lceil \frac{1}{1 - P(PktError)} - 1 \right\rceil$$

Equation 147

The data loss can be expressed in terms of the overhead efficiency; a measure of the amount of information transmitted through the network to successfully transmit user data:

$$Eff_{OH} = \frac{Pkt_{user}}{Pkt_{Total} \cdot (1 + ExtraPkts)}$$

Equation 148

3.4.8. Congestion

The following congestion metric was developed for this thesis.

The congestion metric is based on estimating the number of dropped packets in the network due to congestion. Packets are dropped if, on any link (i,j), the offered load exceeds the following condition:

$$\lambda_{Cap}^{ij} = \frac{0.99 \cdot N_{Cap}^{ij}}{D_{TP}}$$

Equation 149

where N_{Cap}^{ij} is the total number of channels on a link. The "actual" arrival rate at capacity is:

$$\lambda_{Cap}^{ij} = \frac{N_{Cap}^{ij}}{D_{TP}}$$

Equation 150

However, this definition would make the network unstable (see Section 3.4.4).

3.4.9. Load Balance

The load balance metric described below was developed for this thesis.

As noted previously, the load balance performance of a network is defined to be the percentage of links demonstrating light or medium loading. A link is considered lightly-loaded if less than or equal to 25% of the link capacity is being utilized. A link that is medium loaded has between 25% and 99% of its capacity being used. A congested link is one where packets must be dropped to guarantee stability (defined as a link in which greater than 99% of the channels are being used). Hence,

$$LoadBalance = \frac{\%LightLoad + \%MediumLoad}{Total}$$

Equation 151

3.4.10. Round-Trip Delay

The following round-trip delay metric was derived for the purposes of this simulation model.

The round-trip delay is also referred to as the latency of the network. As a general rule, satellite networks are considered high-latency systems, and since many protocols are latency-sensitive, it is difficult to design protocols that behave well over satellite networks.

The expected round-trip delay per packet includes the time associated with the switching delay at each of the nodes on the path plus the queuing, service, and propagation delays multiplied by the expected number of packets per successful transmission. It is assumed that the terrestrial and satellite switching delays D_{Switch} are each 5 milliseconds.

As determined in the data loss section, the extra packets for Go-Back-N and SRP protocols are respectively expressed as:

$$ExtraPkts = \left\lceil 2\left(1 + \frac{Distance \cdot R_{user}}{c \cdot Pkt_{Total}}\right) \cdot \left(\frac{P(PktError)}{1 - P(PktError)}\right) \right\rceil$$

Equation 152: Go-Back-N

$$ExtraPkts = \left\lceil \frac{1}{1 - P(PktError)} - 1 \right\rceil$$

Equation 153: SRP

The expression for the average round-trip delay, given the number of nodes on the path N_{Nodes} is:

$$RTD = \left(D_{Switch} \cdot N_{Nodes} + D_{Queue} + D_{Service} + D_{Propagation}\right) \cdot \left(ExtraPkts + 1\right)$$

Equation 154

From the results of the multiple access section,

$$D_{Queue} + D_{Service} = \frac{\left(\frac{\lambda}{N_{ch}}\right) \cdot (D_{TP})^2}{2 \cdot \left[1 - \left(\frac{\lambda}{N_{ch}}\right) \cdot D_{TP}\right]} + \frac{3D_{TP}}{2}$$

Equation 155

$$D_{Propagation} = \frac{Distance}{c}$$

Equation 156

3.5. Conclusion

The simulation models used in this thesis are presented in detail in this chapter. The performance metrics used to determine the "goodness" of the system and network architectures are motivated and developed. The assumptions and derivations will be used in conjunction with the results in Chapter 4 to fully understand the interactions occurring between the system and network architectures.

Chapter 4

PROOFS

4.1. Introduction

This chapter details the results of the simulation models described in Chapter 3 in the context of proving Theorem 1. Section 4.2 outlines the Existence proof which references the simulation model developed in Section 3.2, while Section 4.3 summarizes the Significance proof, making frequent use of the analytical models developed in section 3.3 to explain the resulting trends.

4.2. Existence Proof

In order for the Existence proof to be satisfied, a single example must be found demonstrating that the constellation topology drives the network protocols (in this case, routing protocols) and that the network protocols in turn drives the constellation topology. This section provides the necessary illustration to prove the existence of the Distributed Satellite Communication System Theorem.

4.2.1. Scenario

Consider the following scenario. Suppose a satellite company wishes to provide communication services to customers in North America, East Asia, and the Pacific Ocean. Such a service would have been particularly useful to the United States military during the WWII Pacific campaign, for example. Today, such a service might be geared toward US corporations engaging in large financial transactions with Japan, Taiwan, China, Korea, and India. Coverage over the Pacific could enable high-speed communication links on business jets and commercial aircraft as well as providing real-time tracking of cargo ships between the two regions.

Once the desired market is identified, the systems objectives and constraints are specified. Two high-level objectives for this system might be (in order of priority):

1. *Cost.* Minimize the cost of the system. While there are many factors contributing to the total lifetime cost of a satellite system, including the choice of launch vehicles, desired lifetime, antenna size, and the number of ground stations, the major dollar decision is the number of satellites. Thus, for the purposes of this example, minimizing cost will translate to minimizing the number of satellites.
2. *Quality of Service.* Maximize the quality of service seen by the customers. Generally, this means minimizing call blocking probability (probability that a voice or data circuit is not

available when a call or data transfer is attempted), probability of dropped calls (probability that a voice or data transfer is dropped in the middle of the communication), round-trip delay, etc.

Clearly, the major constraint should be that the system architecture meets the required coverage as specified by the desired market. What often happens in real systems is that the system architecture (including the constellation topology, the location of the ground stations, and the basic design of the satellites) is chosen based on these high-level objectives and constraints. This information is then passed onto the network engineers who must design to their own objectives within the scope of the pre-selected architecture (choice of topology and routing protocol in this example).

Now consider what would happen if the network engineers had the opportunity to select the overall system architecture. As before, the major constraint would be that the system must meet the required coverage. However, the high-level network objectives differ:

1. *Congestion*. Minimize traffic congestion. As traffic builds up on a given communication link, the expected delay as seen by a given user increases. A similar phenomenon occurs with vehicular traffic. In systems with random access protocols, an increase in congestion leads to an increase in the probability of collisions between two or more packets. In systems with fixed-access schemes, although collisions are avoided, the system can only accommodate so many simultaneous users; at some point the communication link reaches capacity and some data must be rejected. Thus, congestion leads to an increase in the average delay seen by the user as well as an increase in the probability of blocked calls and dropped calls.
2. *Maximum Number of Hops*. Minimize the maximum number of hops required. One can think of the number of hops as the number of intermediary communication links a data packet must travel through on its path from source to destination. The more hops a packet must make, the more time – in general – it takes to go from source to destination. Furthermore, increasing the number of hops may increase the probability that the packet will travel a heavily congested link; this trend is not always the case since if many packets traveling minimum hop paths try to go over the same intermediary link, the link could become congested.
3. *Load Balancing*. Maximize the load balancing of the network. As covered in Chapter 3, the degree to which a network is load balanced is the degree to which the traffic is evenly distributed among all the communication nodes. A system that is well load-balanced is a system that is stable. It is possible for a network to come to a point at which the rate at which data arrives into a link is greater than the rate at which data can leave the link. When this happens, the network deadlocks; think of what happens when a drain clogs. A system that is well load-balanced is better able to distribute the traffic among multiple links reducing the probability that the system becomes unstable (it may still become temporarily unstable if there is a sudden drastic increase in the network traffic).

Although the network objectives are quality of service measures, these are generally not the same quality of service measures used by the systems engineers to choose high-performing systems. However, one should note that traffic congestion, number of hops, and load balancing are performance issues that affect the quality of service desired by the systems engineers.

If the overall system architecture (choice of topology and routing protocol) chosen by the network engineers differs from the architecture chosen by the systems engineers, then the network protocols drive the design of the system topology and vice versa. Demonstrating this relationship will prove the Existence theorem. To see this, consider that the choice of routing protocol directly impacts the performance of the network in terms of congestion, number of hops and load balancing; each of these metrics depend on how packets are routed through the network. The constellation topology strongly influences how the packets are routed through the network.

4.2.2. Results

The architectural decisions made by the systems engineers and the network engineers are considered separately to see what design choices are made.

4.2.2.1. Systems Engineers

Based on the objectives and constraints identified above, consider the system architecture (topology and routing protocol) the systems engineers would choose for the specified scenario. For the purposes of this exercise, only the topologies discussed in section 3.2 will be considered.

The first architecture down-select would ensure the coverage constraint is met. Only two topologies out of the eight can guarantee the required coverage. Topology 6 and topology 8 are the only architectures with coverage of the Pacific, East Asia, and North America. These satellite system architectures are shown in Figure 35 for easy reference.

The second architecture down-select chooses the architecture that minimizes the cost – in this case, minimizes the number of satellites. Clearly, system topology 6 (Figure 35(a)) achieves this performance objective with two satellites compared with the three satellites required for topology 8 (Figure 35(b)).

Now that the system topology has been chosen, it is necessary to choose the routing protocol that maximizes the quality of service seen by the user. Table 4-1 and Figure 36 illustrate the quality of service performance metrics for system topology 6 in conjunction with protocol P1

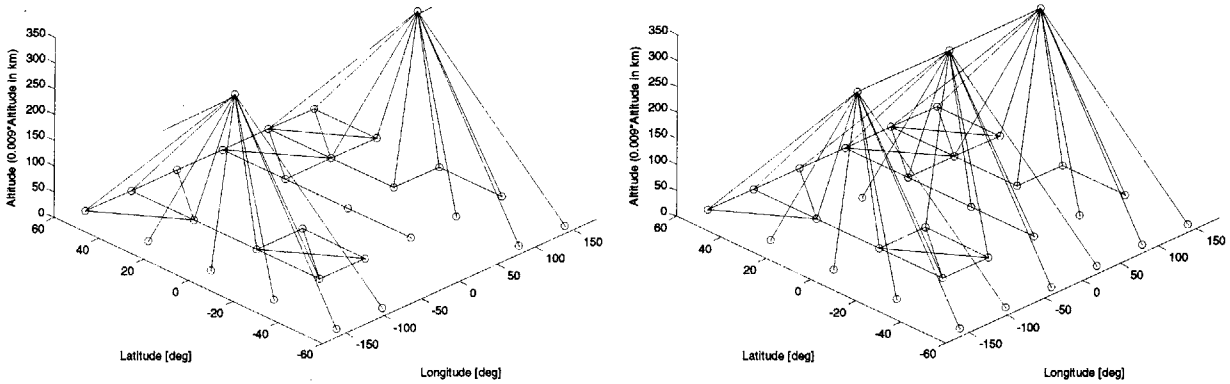


Figure 35: Architectures Meeting Required Coverage: Topology (a) 6, and (b) 8
Topologies 6 and 8 are the only two satellite system architectures that can guarantee the required coverage.

(Figure 36(a)), which routes based on the minimum distance between the desired source and destination, and with routing protocol P2 (Figure 36(b)), which routes based on the minimum number of hops between the desired source and destination.

For this illustration, maximizing the quality of service as seen by the user is the same as minimizing the maximum number of hops, minimizing the congestion, and maximizing the load balancing. As per Section 3.2, minimizing the congestion means minimizing the maximum offered load on any link in the network. Likewise, maximizing the load balancing in the network means maximizing the percentage of links in the network with the offered load falling below a threshold value of 0.5. Table 4-1 clearly shows that the optimal decision for topology 6 is to choose routing protocol 2. Figure 36 graphically depicts the difference between the two protocols by showing the relative congestion on each link.

Systems Decision: Topology 6 and Routing Protocol 2 (P2).

Table 4-1: QoS Data for Topology 6 with Routing Protocol 1 and 2
The optimal quality of service (QoS) decision for each metric is highlighted for routing protocol 1 (P1) and routing protocol 2 (P2).

Metrics	P1	P2
Max Hops	8	7
λmax	4.5373	3.1129
LL%	46.88	53.92

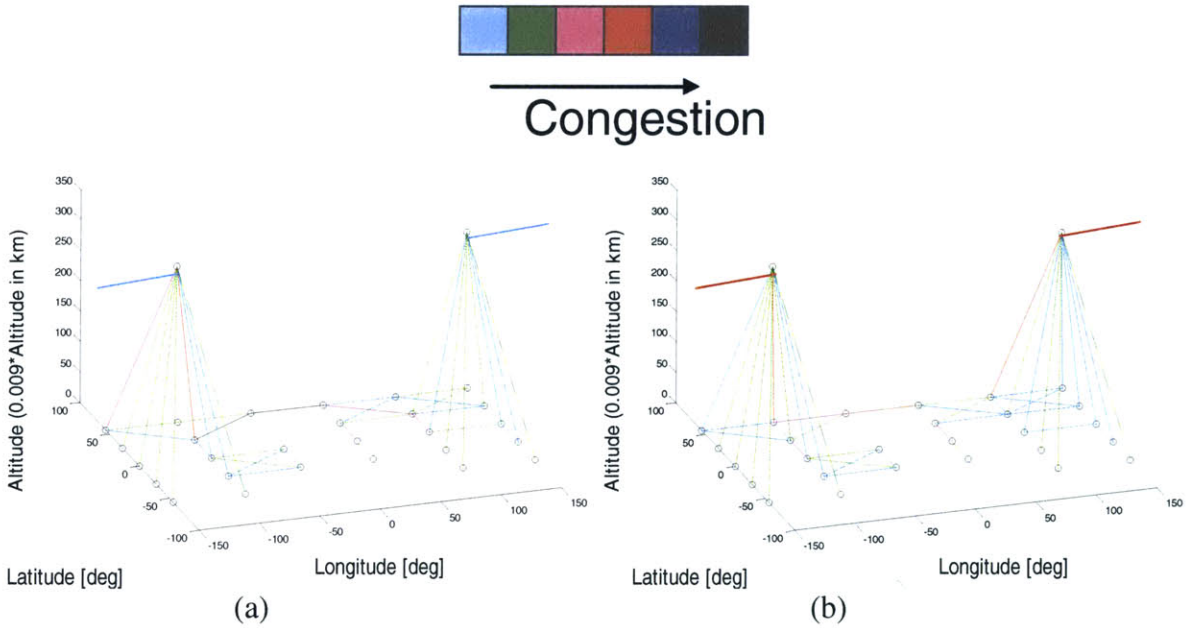


Figure 36: Basic congestion maps of Topology 6 and (a) P1, (b) P2
Routing protocol 1 (P1) minimizes the distance traveled between any source/destination pair, and routing protocol 2 (P2) minimizes the maximum number of hops required for data to traverse between any source/destination pair. The darker the color on a link, the more congestion the link sees.

*4.2.2.2. **Network Engineers***

Now consider what would happen if the network engineers chose the system architecture. As before, filtering out the architectures that fail to meet the coverage constraint leaves two possible topologies: topology 6 and topology 8.

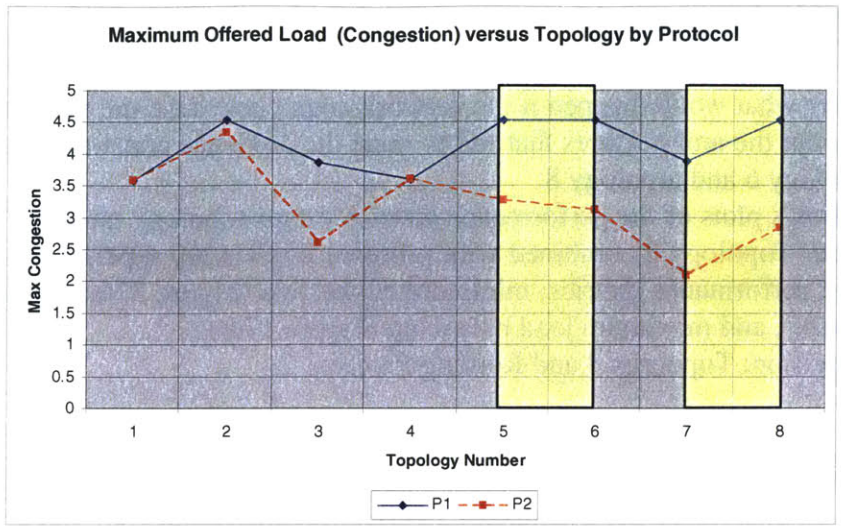
Figure 37 shows plots of the performance metrics versus topology number. Topologies 6 and 8 are highlighted. Topology 8 combined with protocol 2 is visibly superior to topology 6 for all three considered performance metrics: minimum congestion (Figure 37(a)), minimum number of hops (Figure 37(b)), and maximum load balancing (Figure 37(c)).

Network Decision: Topology 8 and Routing Protocol 2.

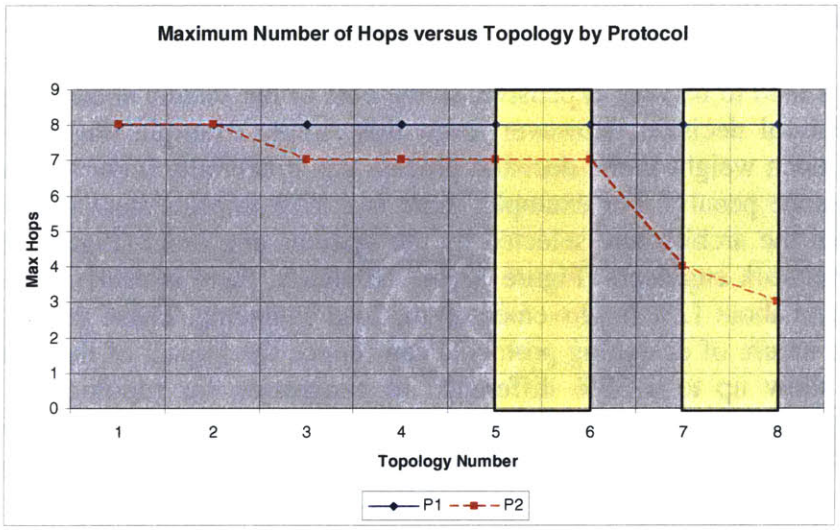
4.2.3. Conclusions

Because the architectural decisions made by the system and network engineers differ, existence for Theorem 1 has been proven; these two very different solutions are shown in Figure 38. Satellite systems tend to be very expensive, so the cost of the system should be a major factor in the end architectural decision. However, even this simple example illustrates that giving the dollar cost too much weight in the decision process can potentially incur a significant quality of service performance penalty. For example, there is a 57% improvement in the number of hops required between the architecture selected by the systems engineers (Figure 38(a)) and the one chosen by the network engineers (Figure 38(b)). Similarly, there is nearly 10% improvement in the congestion and about 12% improvement in the load balancing. These are not small numbers!

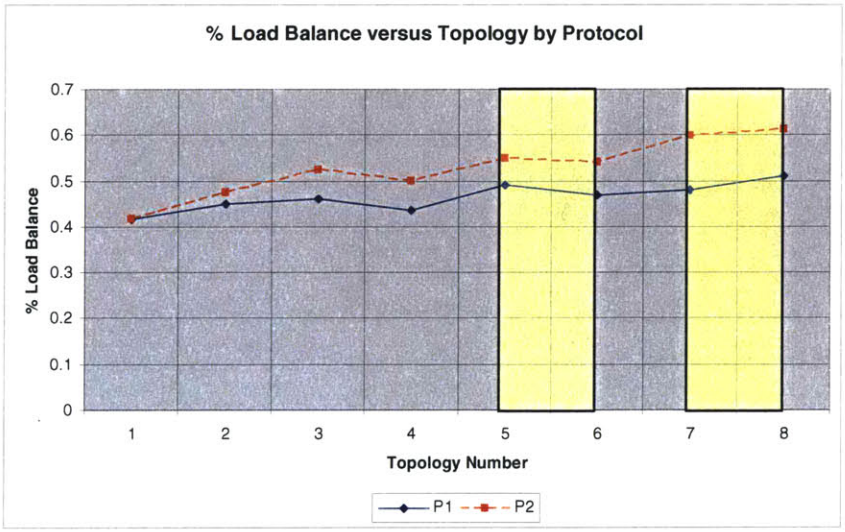
The flexible nature of designing protocols can reduce the impact of this penalty. Even these simple results show up to a 38% difference in congestion for topology 8 between routing protocols 1 and 2 (and in at least two topologies, there is no noticeable difference). However, this flexibility does not guarantee a system will be able to provide sufficient quality of service to woo customers; and without customers the system will fail.



(a)



(b)



(c)

Figure 37: (a) Congestion, (b) Max # of Hops, and (c) Load Balancing vs. Topology
These plots analyze the relationship between the three performance metrics and the topology. Topology 6 and Topology 8 are highlighted for convenience.

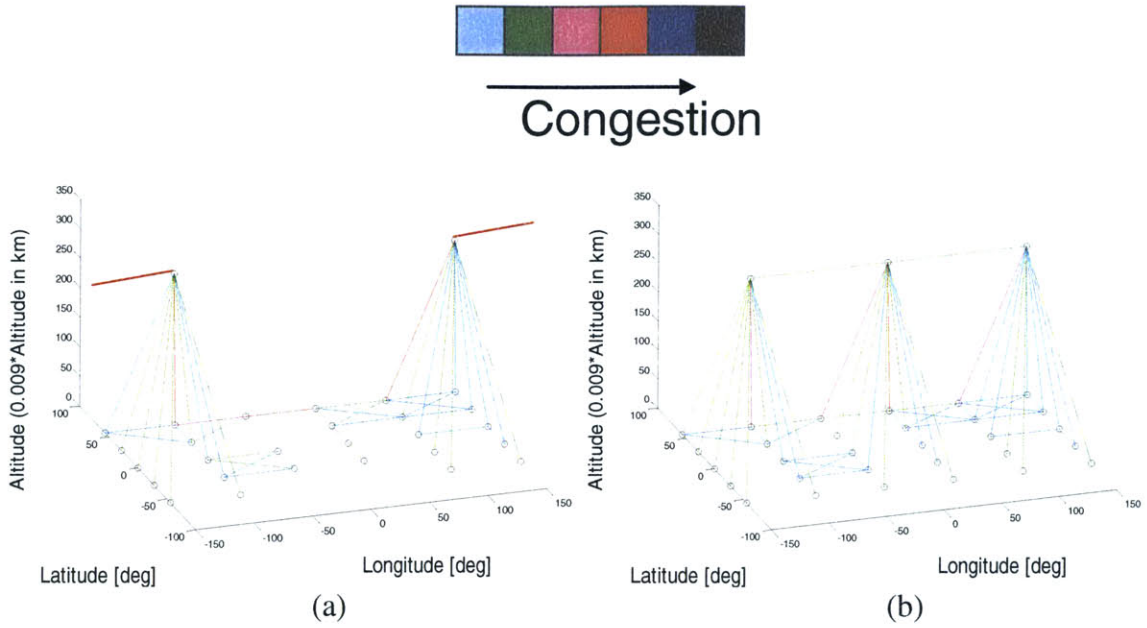


Figure 38: Architectural Solutions Chosen by (a) Systems, (b) Network Engineers
The systems engineers choose the (Topology 6 + P2) architecture while the network engineers choose the (Topology 8 + P2) architecture. The darker the color on a link, the more congestion the link sees.

4.3. Significance Proof

Now that existence has been proven for Theorem 1, it makes sense to continue with the Significance proof. It is now known that coupling indeed exists between the system and network architectures, but it is not yet known how strong this coupling can be. The Significance proof examines the cost of sequential design, a measure of the penalty incurred by assuming little to no coupling in a complex system.

If the coupling is weak, then allowing feedback from the network architecture will only weakly influence the predicted performance of the overall satellite system (if there is any influence at all). Thus, a process model incorporating feedback will not produce a better design than a sequential-based process. Weak coupling further implies that the quality of service performance penalty for designing a system sequentially is small or inconsequential.

However, if the sequential design performance penalty is substantial, then a case can be made that the system exhibits strong coupling, and alternate design methodologies should be considered. In this case, allowing the design of the network architecture to push back on the system architecture will strongly influence the predicted performance of the entire satellite system. A sequential product development process is not sufficient to guarantee anything approaching an optimal design; nor is it guaranteed to result in a system with a quality of service adequate enough to attract and retain customers. A process model incorporating feedback will be more likely to find a design with a good trade-off between system and network requirements and quality of service performance.

A third possibility is that the strength of coupling depends on the design path taken. For example, strong coupling may only be apparent in systems with inter-satellite links, and not in

systems using bent-pipe architectures, or in systems with the satellites in LEO versus GEO orbits. In this case, some types of distributed satellite systems should be designed incorporating feedback in the design process, and for others this will not be necessary.

Section 3.3 covers the development of the extensive simulation model for this proof. Section 4.3 will reference back to Section 3.3 as necessary.

4.3.1. Scenario

Consider a scenario in which the desired market is global. This market requires the satellite system to achieve full global coverage. Assume that the Federal Communications Commission has granted the system 10 MHz total in the Ka frequency band.

If the system is designed sequentially, starting with the system designers and driving toward the network engineers, the following occurs. Suppose that the system designers decide that they want to do a trade study of the possible architectures. For the reasons outlined in section 3.3, they decide on the following objectives: minimize the cost per user per month, maximize the market potential of the system, minimize the life cycle cost of the system, minimize the unused capacity, and maximize the number of simultaneous users the system can support at capacity.

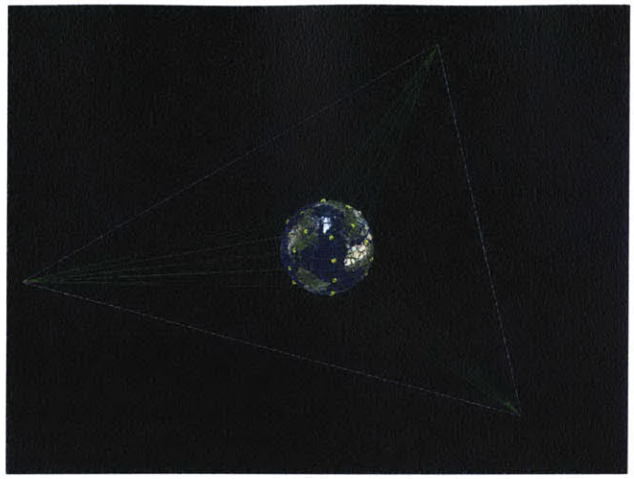
Furthermore, they decide they can adjust the following variables: altitude (specified by the orbital period, Tday), spacecraft lifetime (Tlife), minimum elevation angle (Emin), end-user receiver diameter (Dr), satellite transmitter diameter (Dt), satellite transmitter power (Dt), and a decision on whether or not to buy terrestrial capacity to support the system (TC).

Now suppose that the systems designers finish their trade study and decide on some system architecture; this architecture includes the constellation topology (the three possible topologies are shown in Figure 39) and high-level spacecraft design. They take this decision and pass it onto the network designers who must now design the network protocols.

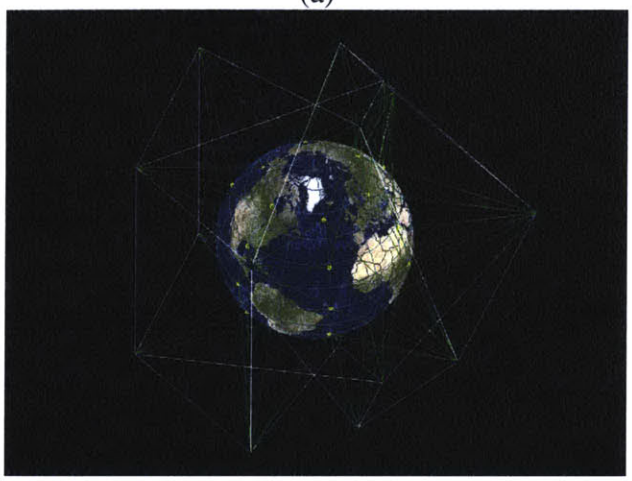
The network designers take the system topology and spacecraft design – which are now fixed – and decide to pursue their own trade study of the decisions under their control. Their design provides the actual communication service to the customers and they want to choose the best network architecture they can. For the reasons detailed in Section 3.3, they decide on the following objectives: maximize the spectral efficiency of the system, maximize the load balancing capability of the network, minimize the data loss (which translates to maximizing the network efficiency), minimize the network congestion, and minimize the round-trip delay experienced by the average user.

After some discussion, the network engineers settle on the following design variables: routing protocol (RP), multiple access protocol (MAP), error correction protocol (ARQ), modulation scheme (MS), average user packet size (PS), average user data rate (Ruser), and a decision on whether to route through a centralized server or to make each satellite a router via distributed routing (ND). The systems engineers think they want to hit a broadband market but the network designers aren't convinced the architecture handed to them can support that.

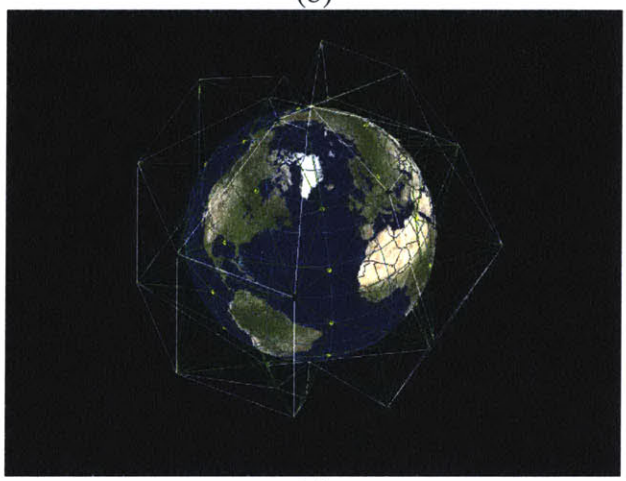
When the network designers finish, they have a network architecture that provides the best quality of service that the overall architectural design can manage given the valuations of the network objectives by the network designers.



(a)



(b)



(c)

Figure 39: Constellation Topologies with Periods: (a) 1, (b) 1/5, and (c) 1/9 Days
The set of constellation topologies considered in the Significance Proof, occurring at (a) an orbital period of 1 day (GEO) , (b) an orbital period of 0.2 days (MEO), and (c) an orbital period of 0.1111 days (low MEO). In the product development process, systems designers are responsible for choosing a system architecture which consists of the constellation topology and high-level spacecraft design.

If the system is designed sequentially, but driven from the other direction, a similar scenario occurs. The development process is nearly identical, except that it is the network designers who optimize the performance of the network protocols before passing the network architecture design – now fixed – on to the system designers who then decide on the best system topology and spacecraft design possible.

It should be noted that this particular design process cannot be realistically implemented as described. However, it is an interesting intellectual exercise to fully understand the relationship between the system and network architectures. A more practical process that emulates the behavior of this scenario is to fix very high standards for quality of service and to use these standards as constraints on the feasible system architectures. The initial designs of the Teledesic system appear to have used this process when claiming to achieve a BER on the order of 1e-10 with a constellation of nearly 900 satellites [Stur96].

The cost of sequential design compares the performance of a system designed sequentially against how the system would perform if designed concurrently under similar conditions. A concurrent design process simply enables feedback between coupled systems, facilitating a drive toward the global optimum; sequential design processes tend to find only local optimums if the processes are coupled.

The Significance proof takes this scenario and examines the cost of designing sequentially to determine the degree to which the interactions between the system and network architectures are coupled. The significance proof considers all possible designs and does not filter by desired market.

The assumptions and simplifications of the model are as defined in Section 3.3.

4.3.2. Multi-Objective Weightings

The design decisions of both the network and system engineers are modeled using multi-disciplinary design optimization. It is assumed that both the system and network designers assign some measure of importance to each of the performance metrics – objectives – under their respective control and that these weightings guide the determination of which designs are optimal. If the performance of one or more of these objectives is constrained to meet certain requirements, then the objective space of possible solutions is reduced even further. The weightings for each step in the sequential process must always sum to one; in the case of concurrent design, all of the weightings taken together must sum to one.

The Significance proof considers two examples. Example 1 models a system that values minimizing the data loss – expressed in terms of maximizing the overhead efficiency – over minimizing the average RTD; valuations that one would expect to see in a system specializing in non-real time messaging and data transfer services where data quality is very important, but the applications are not latency-sensitive. Example 2 models a system that values minimizing the average RTD over minimizing the data loss. However, the value of the data loss objective is kept at roughly the same order of magnitude as the weighting for the average RTD. These are valuations one would expect to see for a satellite telephony system, where the observed latency and data quality are very important measures of quality of service.

For both Example 1 and 2 in Table 4-2, the systems designers are assumed to place an importance of 0.75 on the overall life cycle cost, 0.225 on the cost per user per month, and 0.025 on the expected increase in market potential over the course of the satellite system lifetime.

Table 4-2 provides an overview of the assumed weightings used in both examples. The cost per user per month metric is extremely dependent on the accuracy of the market potential model, and to alleviate some of the bias introduced by the inherent inaccuracies of that model, only the difference in the market potential is considered. The weightings are assumed to be distributed in this fashion for the following reasons:

The life cycle cost of the system has traditionally been one of the major drivers in the architectural design of the system architecture. A large portion of this money must be provided up front – fixed-costs associated with start-up, research and development, and launch of the satellite constellation – and investors must be convinced that their money is being used wisely and effectively. Similarly, the life cycle cost has a huge influence on the subscription cost each user must pay each month in order for they system to break even on the investment with some internal rate of return. If the subscription costs are too high, then potential customers will likely choose a competing system. There is some trade off between the life cycle cost and the projected increase in market potential. The longer the system is in service, the more customers it is likely to attract, assuming that the cost to each customer is not prohibitively high and the quality of service is sufficient to meet the customer needs. However, the longer the system is in service, the greater the life cycle cost due to operational expenses that are incurred each year the system is flown.

A major flaw of the market potential model is that no accounting is made of the influence that quality of service and subscription costs play on market penetration and customer retention as a function of time. Thus, the major drivers involved the trade-off between life cycle cost, cost per user per month, quality of service, and the market potential are not truly captured. The results

Table 4-2: Multi-Objective Weightings for Example 1 and 2
The objective weightings for each step in the sequential process must always sum to one; in the case of concurrent design, all of the weightings taken together must sum to one.

Example #	Max RTD	System Weights			System Requirement Weights	
		LCC	CUM	Delta Market Potential	Users at Capacity	Unused Capacity
1	1	0.75	0.225	0.025	0.07125	0.07125
2	0.1	0.75	0.225	0.025	0.04875	0.04875
Example #	Max RTD	Network Weights				
		Avg RTD	Avg Overhead Efficiency	ULDL Spectral Efficiency	Load Balance	Congestion
1	1	0.05	0.38	0.1425	0.1425	0.1425
2	0.1	0.35	0.26	0.0975	0.0975	0.0975

given in this section involving cost per user per month (CUM) and the market potential should thus be viewed with a big grain of salt. Likewise, the magnitude of the CUM objective should be viewed with skepticism since there are ways to significantly reduce the cost seen by the user. For example, the internal rate of return can be reduced; other potential sources of revenue can be explored, etc.

To emphasize the influence of life cycle cost on the end design and to reduce the impact of these flaws, the relative weightings of the CUM and market potential objectives are kept small.

Two of the system objectives, maximizing the number of simultaneous users – users at capacity – and minimizing the unused capacity of the system (both taken to be at the beginning of life, BOL), are highly dependent on the choices made by the network designers. While there are ways of estimating these numbers without consideration of the network architecture, these estimations are poor at best. It is a feature of the simulation model that these objectives are calculated using information on the network architecture. For the purposes of calculating the cost of sequential design, it is assumed that these two systems objectives are system requirements levied on the network engineers; the network architecture must be designed to maximize the number of simultaneous users and to minimize the unused capacity. Since designing for the best network architecture will naturally drive these objectives in the direction desired by the system engineers, it is further assumed that the requirements levied on the network architecture will be met by assigning these objectives a small portion of the weightings allocated by the network designers.

In multi-objective design optimization, each of the objectives should be of the same order of magnitude in order for the objective weightings to maintain their relative importance. A very small weighting can effectively become a very large weighting if the objective value is orders of magnitude larger than the others.

To renormalize each of the objective values, they were divided by the maximum respective objective value occurring in the objective-space. The only exception to this procedure is the renormalizing factor used for the average RTD metric. The largest average RTD value occurring in the objective space is infinite. It was not sufficient to filter out the architectures with infinite average RTD values since there were many examples with very, very large, but not infinite RTD's. For this reason, a maximum RTD value was chosen arbitrarily small. The values of the maximum RTD were varied to examine their effect on the architectures chosen and the performance of said architectures. As the maximum RTD value is decreased, the distinction between architectures with average RTD's less than the maximum is increased; architectures with larger average RTD's than the maximum RTD factor can still be chosen but with a significant penalty on the overall objective function. As Example 1 models a system that is not latency-sensitive, the maximum RTD was set at 1 second. However, Example 2 is very latency sensitive, so the maximum RTD was set at 0.1 seconds.

4.3.3. Results

As discussed above, there are three design processes under consideration:

1. Sequential: System → Network,
2. Sequential: Network → System, and
3. Concurrent.

Table 4-3: Overall Objective Functions for Example 1 and 2
The overall objective functions are calculated for each design process.

Example #	Process Description	Overall Objective Function
1	Sequential: System->Network	0.0831
	Sequential: Network->System	-0.1075
	Concurrent	0.082
2	Sequential: System->Network	-0.5473
	Sequential: Network->System	-0.4088
	Concurrent	-0.2713

The relative performance of each of these design processes are discussed for two examples as illustration of the cost of sequential design for distributed satellite communication systems. Section 4.3.3.1 examines Example 1 while Section 4.3.3.2 studies Example 2.

The objective weightings modeling the designer decisions for both examples are shown in Table 4-2; the overall objective functions for these examples, calculated for each design process are given in Table 4-3. The resulting architectural designs – categorized by example and design process – are given in

Table 4-6; the architectural design choices and trends are analyzed in Section 4.3.3.3.

4.3.3.1. Example 1: Results and Interpretation

The system examined in Example 1 values minimizing the data loss – expressed in terms of maximizing the overhead efficiency – over minimizing the average RTD; valuations that one would expect to see in a system specializing in non-real time messaging and data transfer services where data quality is very important, but the applications are not latency-sensitive

Examining the architectural decisions selected by the system and network designers in each of the design processes shows that for a marginally small increase in cost, fairly significant increases in quality of service performance can be achieved. In Figure 40, the LCC versus RTD are plotted for the thousands of architecture permutations evaluated by the simulation model. In this plot, the life cycle cost penalty for a gain in average RTD is illustrated by comparing the 'best' architecture reached using the different design processes:

- System → Network (S→N) process: architecture decision is circled,
- Network → System (N→S) process: architecture decision is enclosed by a diamond, and
- Concurrent (C) process: architecture decision is marked by a square.

The penalties and gains are indicated on Figure 40 by tracing lines from the choice of architectural designs to their corresponding values along the objective axes.

Comparing Concurrent and System → Network Design Processes

Comparing the 'best' architecture design points for the three design processes, we see that the cost penalty for designing concurrently rather than sequentially for the System → Network case is $30.5 million. In exchange for that $30.5 million (a mere 2.29% increase in cost), the average round-trip delay decreases by 2.08 seconds (a whopping 86% improvement)! For comparison, designing concurrently rather than sequentially for the Network -> System case saves the system $108.8 million (an 8.16% improvement in cost), in exchange for a mere 0.05 second (15.25%) increase in the average round-trip delay.

Another major quality of service consideration is the average overhead efficiency (also provides measure of the average data loss experienced by the system). Figure 41 illustrates the cost of sequential design for this case. For the same cost penalty described in the discussion of the average round-trip delay, designing concurrently rather than sequentially for the System → Network case provides a 45.7% improvement in the overhead efficiency. This means that there is a 45.7% reduction in the data loss seen by the system for only a 2.29% increase in the cost! By comparison, the Network → System case provides no noticeable improvement, but does save $108.8 million.

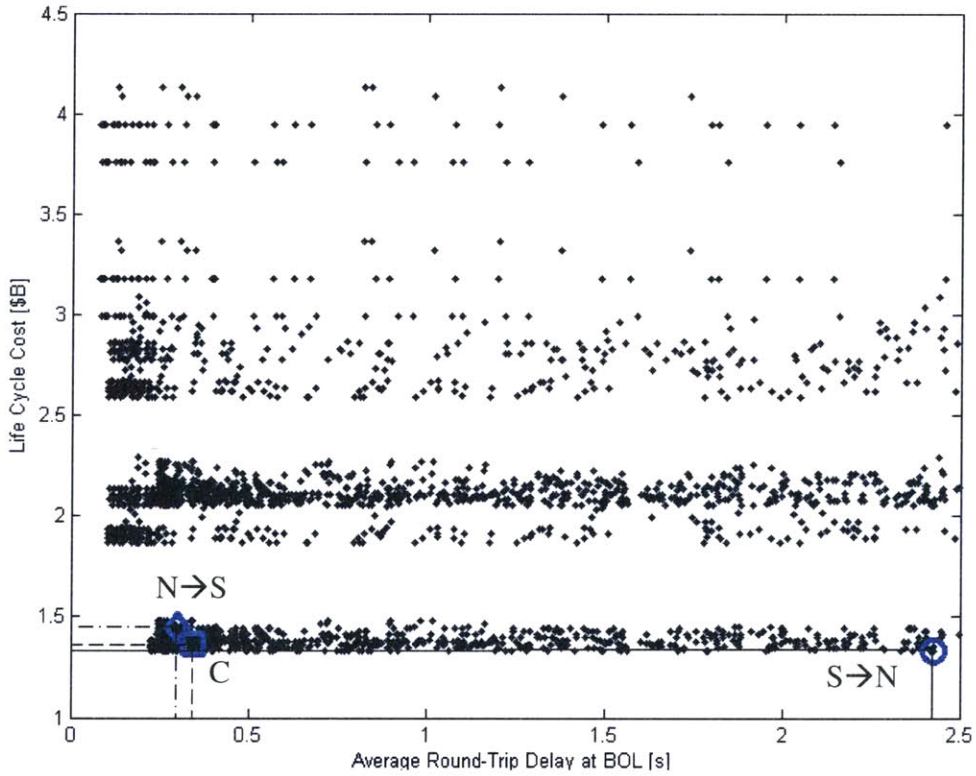


Figure 40: Ex. 1: Cost of Sequential Design for LCC vs. Avg. RTD at BOL
The figure illustrates the life cycle cost LCC [$B] penalty for a gain in average round-trip delay (RTD) in [s] at beginning of life (BOL). The dots represent the possible architectural designs considered by the simulation model. The penalties and gains are indicated by tracing lines from the choice of architectural designs to their corresponding values along the objective axes. The final architectures selected by the three design processes are highlighted: the System → Network (S→N) process architecture decision is circled, the Network → System (N→S) process architecture decision is enclosed by a diamond, and the Concurrent (C) process architecture decision is marked by a square.

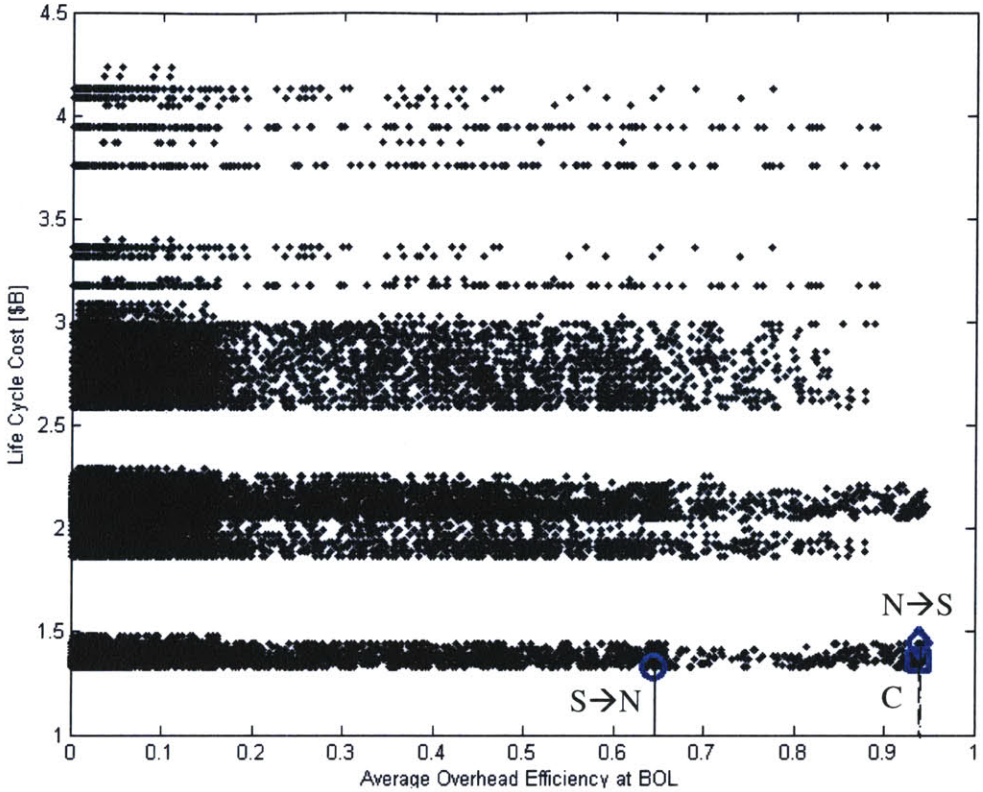


Figure 41: Ex. 1: Cost of Sequential Design for LCC vs. Overhead Efficiency at BOL
The figure illustrates the life cycle cost LCC in [$B] penalty for a gain in BOL average overhead efficiency [-].

Similarly, significant gains for the other performance metrics are shown in Figure 42 for the spectral efficiency of the uplink/downlink (65% improvement), Figure 43 for the load balancing performance (17% increase), and Figure 44 for the observed congestion (99.6% decrease).

Comparing Concurrent and Network → System Design Processes

The results are more mixed for the Network->System case. The spectral efficiency (Figure 42) of the uplink/downlink is improved by 73% for the concurrent process, but the load balancing performance and observed congestion suffer by 4% (Figure 43) and 88.4% (Figure 44) respectively.

Table 4-4 and Table 4-5 provide the performance metric data in convenient tabular form.

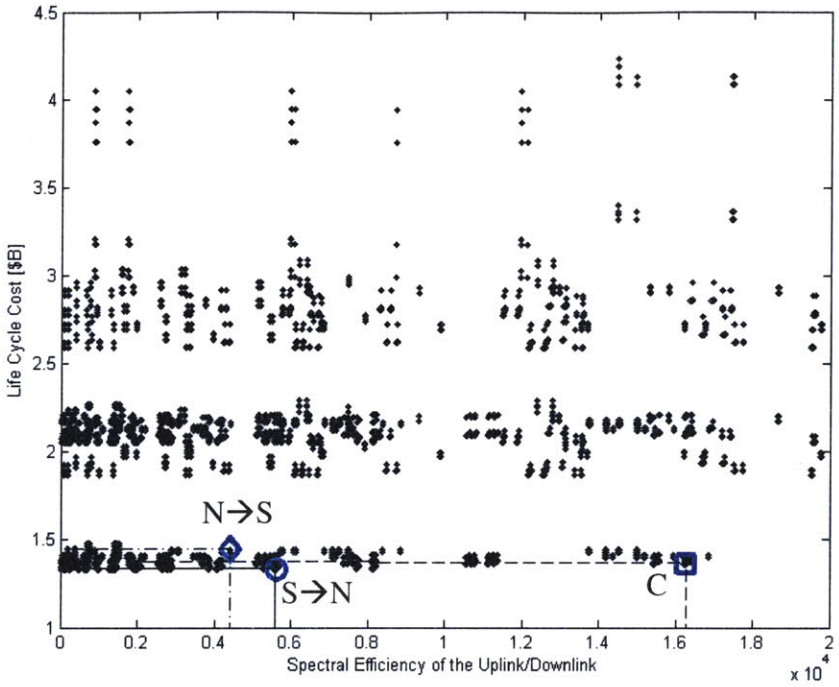


Figure 42: Ex. 1: Cost of Sequential Design for LCC vs. Spectral Efficiency
The figure illustrates the life cycle cost [LCC] in [$B] penalty for a gain in spectral efficiency of the uplink/downlink [-].

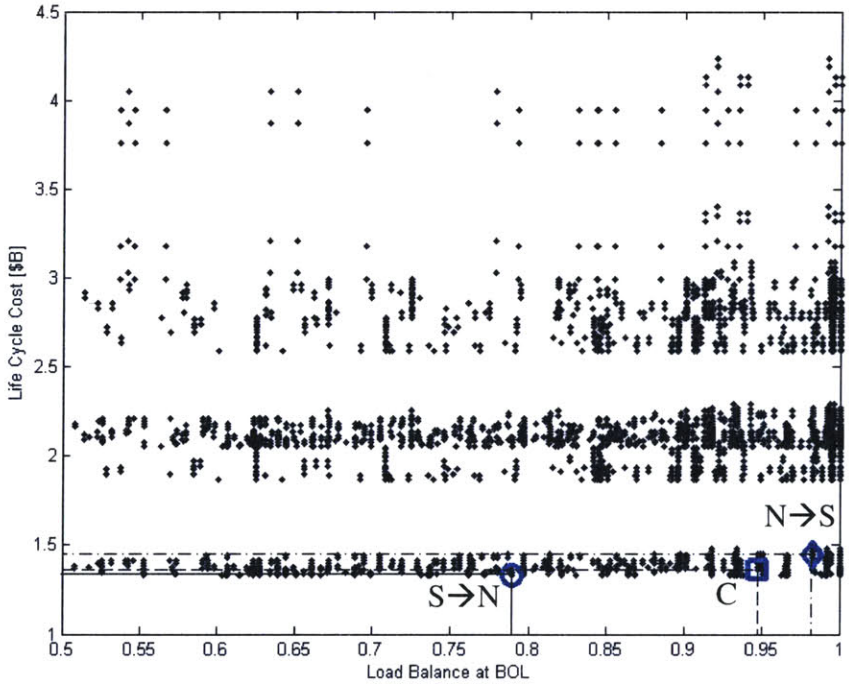


Figure 43: Ex. 1: Cost of Sequential Design for LCC vs. Load Balance at BOL
The figure illustrates the life cycle cost [LCC] in [$B] penalty for a gain in BOL load balance performance [-].

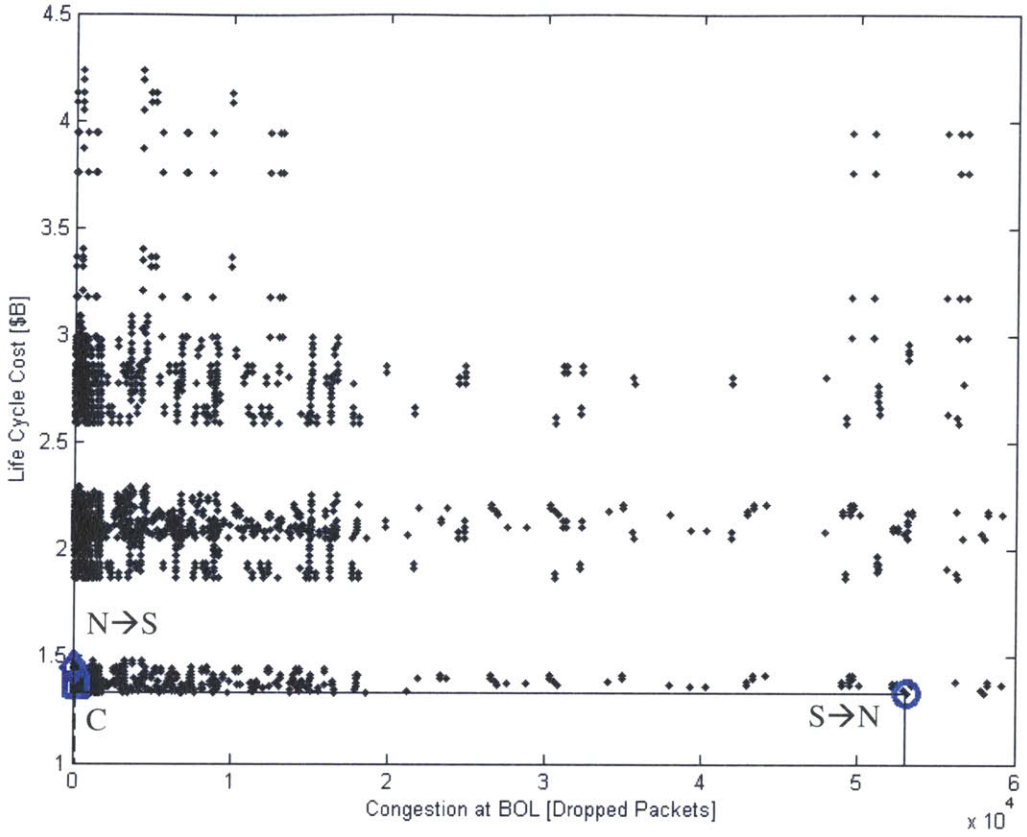


Figure 44: Ex. 1: Cost of Sequential Design for LCC vs. Observed Congestion at BOL
The figure illustrates the life cycle cost [LCC] in [$B] penalty for improved congestion avoidance in [Dropped Packets] at BOL.

Comparison Summary

This example proves the significance of Theorem 1: designing the system incurs a significant quality of service penalty if it is designed sequentially (System → Network) instead of concurrently for the design valuations given in Table 4-3. Example 1 illustrates that a sequential product development process is not sufficient to guarantee anything approaching an optimal design; nor is it guaranteed to result in a system with a quality of service adequate enough to attract and retain customers. A process model incorporating feedback is shown to be more likely to find a design with a good trade-off between system and network requirements and quality of service performance.

Example 1 further illustrates that in some cases, designing sequentially virtually guarantees that the achievable average RTD will not be sufficient to provide real-time data services (300 milliseconds is generally considered to be the cut-off for these types of services) in addition to the desired non-real time services. However, designing concurrently drives the achievable average RTD very close to the value necessary to achieve real-time data services. In other words, designing concurrently rather than sequentially could open the system up to service options that might be impossible to achieve otherwise.

4.3.3.2. Example 2: Results and Interpretation

The system examined in Example 2 values minimizing the average RTD over minimizing the data loss. However, the value of the data loss objective is kept at roughly the same order of magnitude as the weighting for the average RTD. These are valuations one would expect to see for a satellite telephony system, where the observed latency and data quality are very important measures of quality of service.

Once again, examining the architectural decisions selected by the system and network designers in each of the design processes demonstrates strong coupling between the two architectures. In this case, however, the increase in cost is substantial. On the other hand, the increase in quality of service performance is still impressive. Even though the percentage improvement is not as much as in Example 1, there are other considerations which will be discussed in more detail shortly.

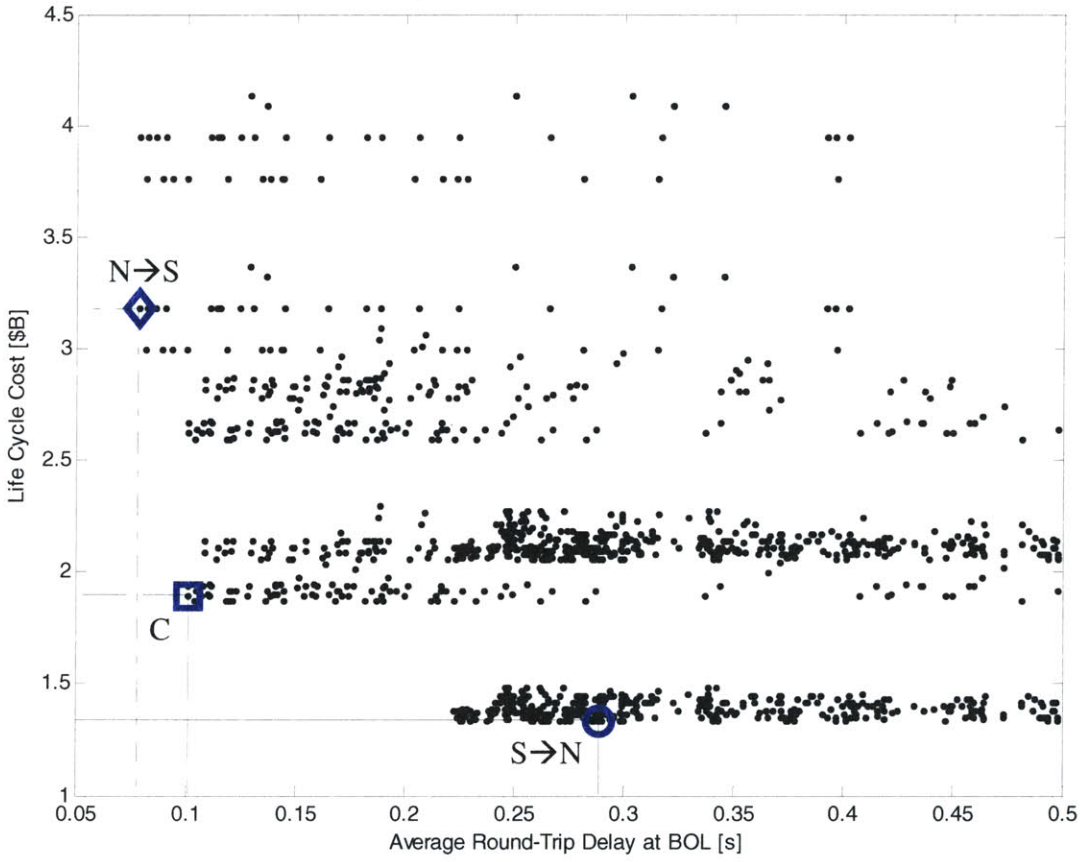


Figure 45: Ex. 2: Cost of Sequential Design for LCC vs. Avg. RTD at BOL
The figure illustrates the life cycle cost (LCC) in [$B] penalty for a gain in BOL average round-trip delay (RTD) in [s].

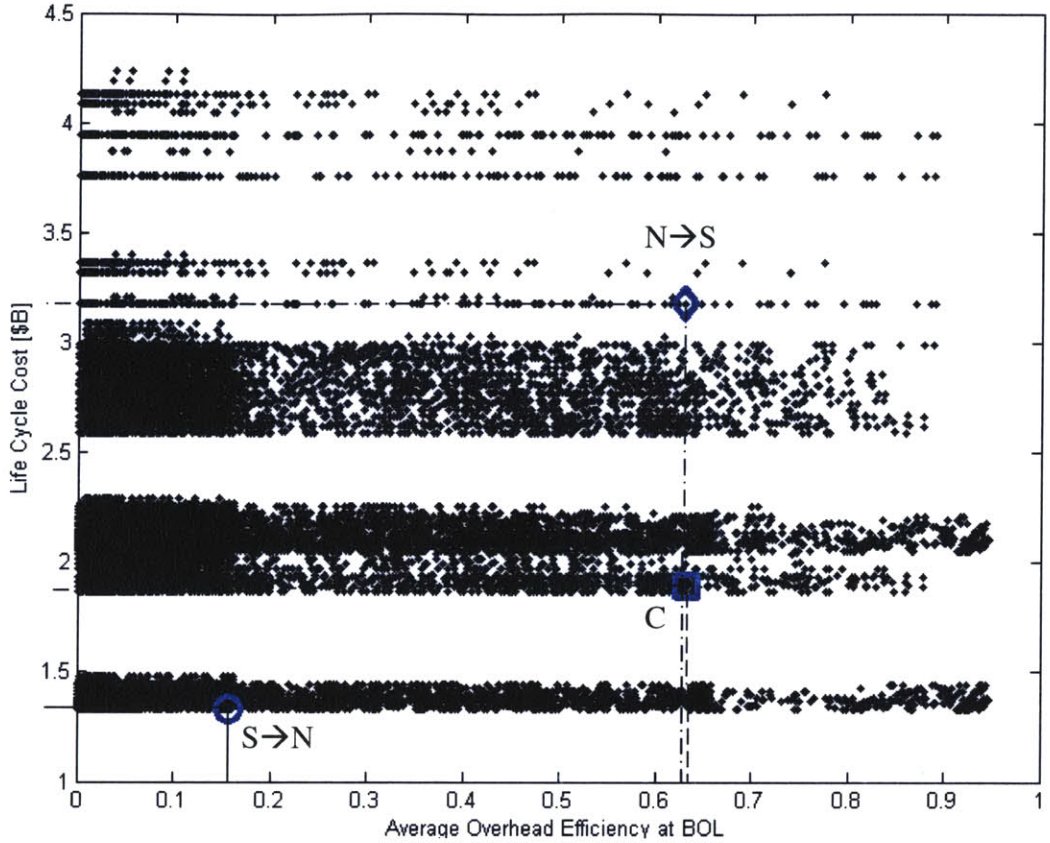


Figure 46: Ex. 2: Cost of Sequential Design for LCC vs. Overhead Efficiency at (BOL)
The figure illustrates the life cycle cost (LCC) in [$B] penalty for a gain in average Overhead Efficiency [-] at BOL.

In Figure 45, the LCC versus RTD are plotted for the thousands of architecture permutations evaluated by the simulation model. In this plot, the life cycle cost penalty for a gain in average RTD is illustrated by comparing the 'best' architecture reached using the different design processes:

- System → Network (S→N) process: architecture decision is circled,
- Network → System (N→S) process: architecture decision is enclosed by a diamond, and
- Concurrent (C) process: architecture decision is marked by a square.

The penalties and gains are indicated on Figure 45 by tracing lines from the choice of architectural designs to their corresponding values along the objective axes.

As Figure 45 through Figure 49 reveal, for an extra $559.2 million (41.95% increase in life cycle cost), the concurrent design provides a 0.1876 second decrease in average RTD from 0.2888 seconds to 0.1012 seconds (35.04% improvement), 75% improvement in the data loss, 97% improvement in spectral efficiency, with marginal tradeoffs involving 0.4% loss in load balance performance and increasing the congestion from 0 dropped packets to 4.302 dropped packets.

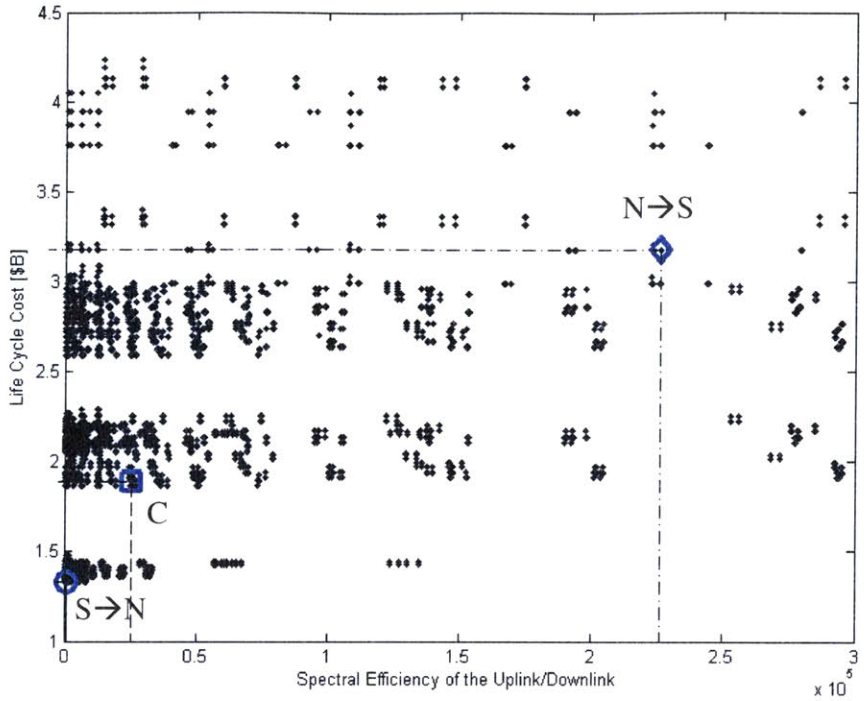


Figure 47: Ex. 2: Cost of Sequential Design for LCC vs. Spectral Efficiency
The figure illustrates the life cycle cost [LCC] in [$B] penalty for a gain in spectral efficiency of the uplink/downlink [-].

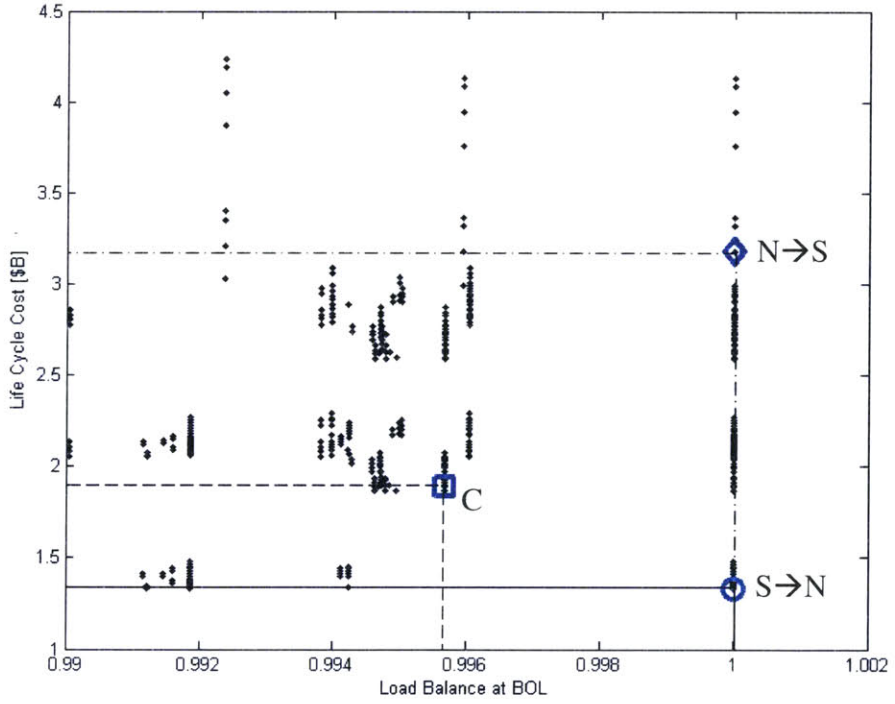


Figure 48: Ex. 2: Cost of Sequential Design for LCC vs. Load Balance at BOL
The figure illustrates the life cycle cost (LCC) in [$B] penalty for a gain in BOL load balance performance [-].

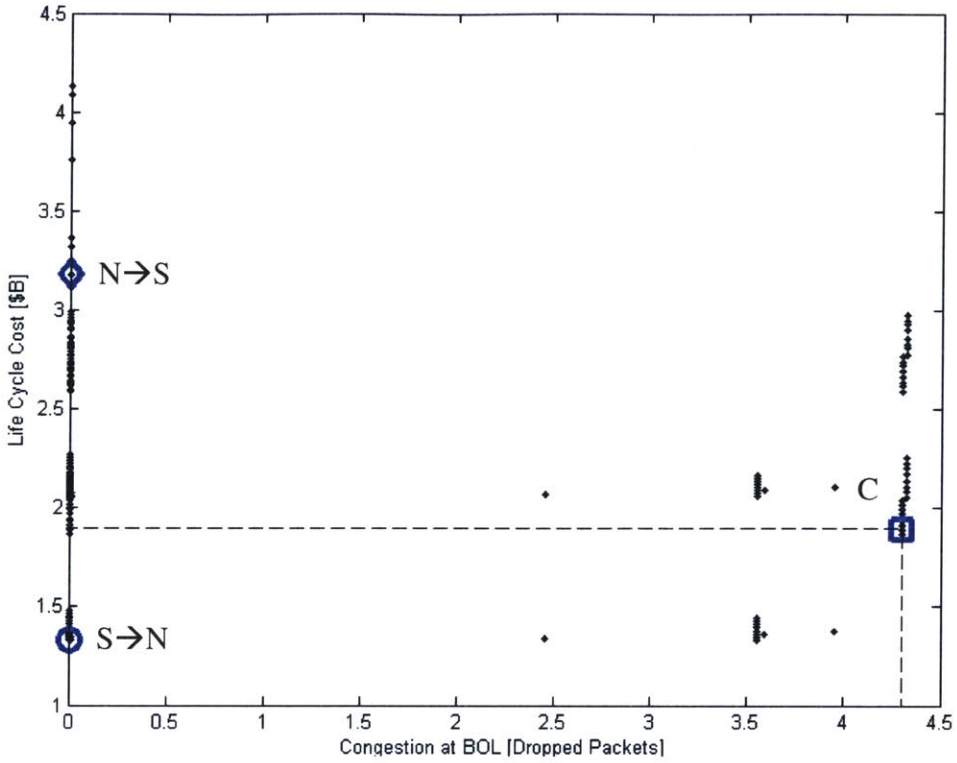


Figure 49: Ex. 2: Cost of Sequential Design for LCC vs. Observed Congestion at BOL
The figure illustrates the life cycle cost (LCC) in [$B] penalty for improved congestion avoidance in [Dropped Packets] at BOL.

Comparison Summary

This significance of Theorem 1 was proven in Section 4.3.3.1, but this does not mean more information cannot be gleaned by looking at a different situation. Example 2 models the case of a satellite telephony system, for which the observed latency and data quality (measured in terms of the overhead efficiency) are both very important measures of the quality of service.

The increased emphasis on RTD appears to push the design of the overall system architecture for the System → Network process closer to the Pareto optimal front in Figure 45 than is seen in Example 1. In this particular case, the sequential process sufficiently drives the architectural choice to a RTD that can enable real-time data transfers. Performing a concurrent analysis, however, can reduce the average RTD significantly further, enabling extremely latency-sensitive applications that the objective valuations used to make the architectural decisions in Example 1 would not have found. However, the penalty for enabling these applications is a 42% increase in the expected life cycle cost; this penalty may be greatly offset by increased market penetration.

A comparison of the results between Examples 1 and 2 suggest that how the various performance metrics, or system responses, are valued by the designer plays a large role in determining the strength of coupling between the system and network architectures. Simply increasing the importance of the RTD, without adjusting the relative percentage weightings of the other network performance metrics has demonstrated a surprisingly large affect on the strength of coupling between the architectures. More detailed analyses will need to be done to understand how to predict these interactions.

Table 4-4: Systems Performance Metrics Results[1] for Example 1 and 2
Table includes results at beginning of life (BOL) and end of life (EOL).

Example #	Process Description	LCC	Capacity BOL	Market Potential BOL	Unused Capacity BOL
		[B$]	[Users]	[Users]	[Packets]
1	Sequential: System->Network	1.33	6.94E+06	674062	7.29E+06
	Sequential: Network->System	1.44	5.64E+05	80112	8.55E+05
	Concurrent	1.36	2.04E+05	80112	3.37E+05
2	Sequential: System->Network	1.33	7.39E+06	674062	1.09E+07
	Sequential: Network->System	3.18	1.57E+07	80112	6.41E+07
	Concurrent	1.89	8.05E+05	80112	4.24E+06
Example #	Process Description	CUM	Capacity EOL	Market Potential EOL	Unused Capacity EOL
		[$]	[Users]	[Users]	[Packets]
1	Sequential: System->Network	155.13	6.50E+06	792770	7.17E+06
	Sequential: Network->System	1370	5.33E+05	100748	8.42E+05
	Concurrent	1300	1.87E+05	100748	3.26E+53
2	Sequential: System->Network	155.13	7.31E+06	792770	1.09E+07
	Sequential: Network->System	3020	1.57E+07	100748	6.41E+07
	Concurrent	1800	7.99E+05	100748	4.24E+06

[1] There is a small scaling error in the simulation code involving the number of users at capacity and the unused capacity metrics. It is unlikely that this scaling error effects the observed trends as any differences between systems should fall within the margin of error. Only the absolute values should be affected to any noticeable degree.

Table 4-5: Network Performance Metrics Results for Example 1 and 2
Table includes results at beginning of life (BOL) and end of life (EOL).

Example #	Process Description	ULDL Spectral Efficiency	Avg Overhead Efficiency BOL	Load Balance BOL	Congestion BOL	Avg RTD BOL
		[-]	[-]	[-]	[Packets]	[s]
1	Sequential: System->Network	5.64E+03	0.6445	0.7895	53050	2.4258
	Sequential: Network->System	4.43E+03	0.9392	0.9825	23.9289	0.301
	Concurrent	1.63E+04	0.939	0.9474	206.1955	0.3469
2	Sequential: System->Network	696.347	0.1575	1	0	0.2888
	Sequential: Network->System	2.26E+05	0.6297	1	0	0.0783
	Concurrent	2.56E+04	0.6336	0.9957	4.302	0.1012
Example #	Process Description		Avg Overhead Efficiency EOL	Load Balance EOL	Congestion EOL	Avg RTD EOL
			[-]	[-]	[Packets]	[s]
1	Sequential: System->Network		0.6423	0.7719	67237	2.5875
	Sequential: Network->System		0.939	0.9649	85.2371	0.3184
	Concurrent		0.9381	0.9035	360.1024	0.3792
2	Sequential: System->Network		0.1575	1	0	0.292
	Sequential: Network->System		0.6297	1	0	0.0783
	Concurrent		0.6335	0.9913	8.4218	0.1019

4.3.3.3. Architectural Design Observations

The architectural designs of Examples 1 and 2 in Table 4-6 are given in terms of the design variable values selected for each design process. Although the results from Examples 1 and 2 prove the significance for the Distributed Satellite Communication Design Theorem (the system and network architectures are strongly coupled), the design process is not guaranteed to result in distributed satellite communication systems that a company would want to fly. This result is due to the sparse nature of the design space and the assumptions built into the simulation models. Most of the design variables had only two possible values that could be chosen, and only the orbital period (Tday), packet size (PS) and user data rate (Ruser) had three. Also, it should be noted that due to time limitations, only a select portion of the design space for Tday = 0.1111 days could be evaluated. In this section we will discuss some of the reasons these results might not be real-world solutions and how the models might be improved.

As the latency sensitivity becomes more of a system driver, the altitude decreases. This result makes sense, but there is a subtle caveat. At lower altitudes, on average a given packet will require more hops to travel between a source and a destination. The number of hops has a significant impact on the quality of service. In a system without regenerative repeaters (as assumed here), the impact is even greater since the probability of packet error is greater than in the case of regenerative repeaters. Although systems with regenerative repeaters suffer less from packet errors, there is a trade-off in terms of the cost of the hardware and the extra time required to process the data.

Table 4-6: Architecture Decisions for Example 1 and 2
Table sorted by example number, choice of design process, and design variables.

Example #	Process	System Design Variables						
		Tday	Emin	Tlife	Dr	Dt	Pt	TC
		[days]	[deg]	[year]	[m]	[m]	[kW]	[-]
1	Sequential: System->Network	1	15	5	0.5	1.5	4	0
	Sequential: Network->System	1	15	5	0.5	3	8	0
	Concurrent	1	15	5	0.5	3	4	0
2	Sequential: System->Network	1	15	5	0.5	1.5	4	0
	Sequential: Network->System	0.1111	20	5	0.5	3	8	0
	Concurrent	0.2	15	5	0.5	1.5	8	1
Example #	Process	Network Design Variables						
		ND	RP	MAP	ARQ	MS	PS	Ruser
		[-]	[-]	[-]	[-]	[-]	[Bytes]	[kbps]
1	Sequential: System->Network	1	1	2	3	2	100	10
	Sequential: Network->System	1	1	1	3	1	1000	1000
	Concurrent	1	1	2	3	1	1000	1000
2	Sequential: System->Network	1	1	2	3	1	10	10
	Sequential: Network->System	1	1	2	3	1	100	1000
	Concurrent	1	1	2	3	1	100	1000

Lower elevation angles are clearly favored, but this is likely an artifact of model used, which does not account for availability requirements. Higher elevation angles – and increased diversity – increase the probability that a satellite is visible to the ground station. However, higher elevation angles and increased diversity tend to require more satellites for all other considerations held equal. Thus, to achieve greater availability, it is expected that the cost of the system will increase.

Shorter system lifetimes seem to be ideal, but as mentioned in the multi-objective weightings section this results is likely misleading. The inaccuracies of the market model fail to give a clear accounting of the interactions between quality of service performance, customer retention, market penetration, and the system lifetime.

A receiver diameter of 0.5 meters appears to be optimal for all cases, but this result is misleading. This result merely says that given the sparse design space considered, there are no scenarios with a receiver diameter of 0.05 meters providing sufficient link margin to ensure the quality of service required by the system. It is possible to improve the link margin by adjusting the gain on the receiver end by considering other types of antennas, or by boosting the gain in orbit (at higher cost).

The variation in the choices of the satellite transmitter diameter and power is best explained in terms of the average data packet size and the average user data rate and will be discussed shortly.

For the most part, the decision to buy out capacity on the terrestrial network is not favored (TC = 1). The most likely explanation for this is that it does add recurrent costs to the overall cost of the system – each year, the satellite company must pay to keep exclusive rights over the channels. There are subtle interactions with the average RTD and congestion metrics. The number of terrestrial channels that the satellite system can buy out is limited. In areas where much of the traffic is being routed to neighboring subnets, most of the traffic will attempt to travel through the terrestrial network rather than over satellite to reduce the number of hops and RTD. If the traffic load offered to these links is greater than the capacity available to the satellite system, then the links become congested, packets are dropped, and the average RTD increases. Routing protocols incorporating congestion as a metric can be used to mitigate these effects. However, the added cost of buying out the capacity may not be worth the increase in performance if the system is lightly loaded – at the beginning of life, for example. Consideration of this design choice is warranted in cases where the system loading increases later in the system lifetime or if the system has such stringent RTD requirements that the increase in congestion on some links improves the overall quality of service.

Choosing to route the traffic in a distributed fashion (ND = 1) is clearly favored. Routing traffic through a centralized subnet will increase the average RTD and the congestion. The performance penalty for centralized routing will be greater in systems requiring more hops on average, thus favoring GEO systems over those in lower altitudes. The main trade-off with the performance benefits of distributed routing is the extra cost incurred by the sizeable routing tables that would be required. Naturally, the size of the routing tables will increase dramatically with the number of satellites in the system.

As mentioned numerous times, the number of hops is a significant quality of service performance driver. This phenomenon seems to be captured in the routing protocol design choice, with a clear favoring of the routing protocol based on minimizing the number of hops required (RP = 1).

The systems considered favor using MF-CDMA (MAP = 2) as the multiple access protocol. MF-CDMA is generally considered to be more spectral efficient, thus increasing the billable information sent through the system.

The systems considered also favor using SRP (ARQ = 3) as the error-correction protocol. This result is unsurprising as SRP is the most efficient basic ARQ protocol. Although the simulation model assumes ideal SRP, more advanced ARQ protocols can achieve similar levels of performance.

The systems also seem to favor BPSK (MS = 1) systems, which is somewhat surprising. QPSK systems are far more spectral efficient, but do incur a penalty in the expected BER of the channel. The systems are assumed to be using the Ka frequency band, which tends to provide far better link margin than the Ku band, all other things equal. Still, the systems seem to suffer more from BER issues than they gain by increased spectral efficiency.

It is interesting that both of the Network → System and Concurrent design processes seem to favor average user data rates on the order of 1 Mbps in combination with packet sizes on the order of hundreds or thousands of bytes. As expected, driving up the packet size and average user data rate boosts the link margin requirements on the system end, increasing the satellite transmitter diameter and power. It is likely that the trend toward greater average user data rate is due to the emphasis on minimizing the average round-trip delay. As the system increases its latency sensitivity (moving from Example 1 to Example 2), the average data packet size decreases for a given user data rate (regardless of design process). Larger packet sizes take longer to transmit and are subject to a greater probability of packet error due to noise for a fixed user data rate.

4.3.4. Conclusions

In this section, the link between the design of the network protocols and decisions and the design of the system topology and spacecraft design was explored. It was found that the architectural design of both the network and the system are strongly coupled. However, the benefits and importance of this coupling appears to be strongly tied to the relative importance of each of the performance metrics used to down-select the architectural design for each product development process. In some sense, Example 1 exhibits far stronger coupling than Example 2; the system in Example 1 achieves huge quality of service improvements for a marginal increase in cost, while the system in Example 2 achieves moderate quality of service improvements for a much larger life cycle cost increase. On the other hand, the average round-trip delay achieved by the concurrent process in Example 2 is a 73% improvement over the average round-trip delay achieved by the concurrent process in Example 1!

These results imply that for a concurrent design process, increasing the quality of service requirements of the system incurs an increasing cost penalty, but drives the overall design closer to the achievable optimal network performance than can be found using only a sequential (System → Network) process. Also, the multi-objective design and objective spaces seem to be increasingly tricky to navigate. Fewer jointly-considered systems can meet the requirements, and it becomes less likely that a high-level trade study will locate these rare jointly-optimal systems.

4.4. Conclusion

Chapter 4 has proven the existence and significance of the Distributed Satellite Communication System Design Theorem. The results verify the existence of the Design Theorem because an example was shown in which the system designers and the network designers choose different satellite architectures (topology plus routing protocol). The significance of the Design Theorem is demonstrated by providing examples in which, for all other things held equal, a concurrent design process can greatly improve the quality of service of a distributed satellite communication system for a marginal increase in cost.

Furthermore, it was discovered that how the various performance metrics, or system responses, are valued by the designer appears to play a large role in determining the strength of coupling between the system and network architectures. The exact nature of this interaction is unknown, and more studies will be required to understand how these valuations influence the strength of coupling.

The results of the significance proof in Section 4.3 indicate that the product development process for distributed satellite communication systems should be reconsidered for some types of systems. Future systems should take care to investigate their exposure to performance penalties early on in the design process.

Although assuming a sequential design process seems to benefit the life cycle cost, the system can suffer a substantial quality of service penalty. Clearly, assuming a sequential design process is not necessarily a good assumption for distributed satellite communication systems. This result would imply that researchers studying distributed satellite communication systems from a systems perspective should make more of an effort to include the effects of the network in their high-level studies. For network engineers, rather than finding the optimal protocol for a given topology, explore how protocols can be optimized by varying the topological structure and spacecraft link margin dynamics, specifically ground and space antenna size, structure, and power.

Chapter 5

CONCLUSIONS

5.1. Overview

This thesis has proven the Distributed Satellite Communication Design Theorem, and has shown that the influence of the interactions between the system and network architectures depends on how the design decisions are made on a design process scale as well as internally to both the system and network designers in terms of how the designers value certain objectives.

The concluding chapter expands on the thesis impact discussion in Chapter 1, taking into account the results found in Chapter 4. Finally, suggestions for future research are made based on the assumptions and simplifications as well as other questions that have arisen during the course of this research.

5.2. Re-evaluation of the Product Development Process

The results of the significance proof in Section 4.3 indicate that the product development process for distributed satellite communication systems should be reconsidered for some types of systems. Future systems should take care to investigate their exposure to performance penalties early on in the design process.

While the quality of service performance improvements for concurrent over sequential design processes is staggering, there are costs to consider when deciding on which design process to follow.

First, any large-scale high-level trade study of the interactions between the system and network architectures under consideration by both the system and network designers will be massively computationally complex, and will become more so as the altitude decreases. The Matlab simulations undertaken for this thesis ran on the order of a minute each for the geostationary earth orbit (GEO) constellations, 5-8 minutes for the high medium earth orbit (MEO) constellations, and 10-20 minutes each for the low MEO systems. The computers used to run these simulations included several 2.99 GHz, 512 MB computers. Fortunately, high-end supercomputers are becoming increasingly affordable; they will be needed for the concurrent design of these systems.

Second, concurrent design processes incur people overhead. As the design process requires feedback between the coupled components (in this case, the system and network architectures), the people involved in these areas will need to communicate more. As such, it can be expected that the development time of the system will increase.

Third, it should be clear from the design architectures shown in

Table 4-6 that even small changes in the value of design variables can have a significant impact on the end performance of the overall system architecture depending on the design

process. This observation implies that the design and objective spaces are very non-linear and discrete, and that the design process chosen can have a significant impact on how well this trade space is explored.

Fourth, in the course of researching this thesis, it became apparent that there is a cultural divide in the satellite communication industry. Network designers seem to have been aware of the potential impact that their protocol designs can have on the system topology, but have been unable to communicate this to the system engineers. On the other hand, the system engineers seem to have been woefully clueless as to the impact their decisions have on the quality of service that the end network can provide, even if designed optimally for the given topology and spacecraft design. Note, these observations are unlikely to be true in all cases.

Finally, it was observed in the course of this thesis that the weightings used in the decision making process for both the system and network designers can have a significant impact not only on what design process to choose, but also on the quality of service and cost performances of the overall architecture.

Evidently, more work will need to be done on an industry scale to find a development process that reduces exposure to quality of service penalties without acquiring excessive amounts of overhead and development cost as a result.

Awareness of the issues brought up in this thesis may enable satellite communication companies to make better decisions in terms of accepting a small cost penalty in exchange for significant gains in network quality of service performance.

Truly, the communication network in distributed satellite communication systems is a critical piece of the architecture, perhaps even the most critical piece. As a result, ignoring its influence for expediency in design and the dollar bottom line may damage a satellite company's ability to generate sufficient revenue later on. In other words, the seemingly innocuous assumption of weak or unimportant interactions with the network, because of the "after all, we can design around it later" mindset inherent in the sequential design process may doom the system to economic failure later on.

5.3. Commentary on Common Assumptions in Research Literature

The results of this thesis provide ample evidence that many common assumptions found in the research literature have a significant impact on the performance of distributed satellite communication systems. A good assumption is one that simplifies the process to get to an answer without drastically altering the results. An assumption that is not valid is one that assumes away an interaction that is critical to the understanding the system. Distributed satellite communication systems are very complex, meaning even a seemingly inconsequential assumption could prove to have a large effect.

Although assuming a sequential design process seems to benefit the life cycle cost, the system can suffer a substantial quality of service penalty. Clearly, assuming a sequential design process is not necessarily a good assumption for distributed satellite communication systems. This result would imply that researchers studying distributed satellite communication systems from a systems perspective should make more of an effort to include the effects of the network in their high-level studies. For network engineers, rather than finding the optimal protocol for a given topology, explore how protocols can be optimized by varying the topological structure and

spacecraft link margin dynamics, specifically ground and space antenna size, structure, and power.

5.4. Application to Terrestrial, Sensor Web, and Ad-Hoc Systems

The architectural design of terrestrial, sensor web, and ad-hoc networks mirror that of satellite communication systems; the design process is also similar. Take a basic network found in universities and offices as an example. Figure 50 relates these design problems to satellite systems.

First, consider the relationship between terrestrial and satellite networks. The only architectural difference is the wired connection between the cable network and the wireless access points. If this wired connection could disappear, the wireless access points would behave similarly to a satellite system; all data transfer would have to flow between access points and the customers who use them. Just as in a satellite network, the access points would need to handoff communications as users move between them. Handoff is still an active area of research, one whose success directly impacts the quality of service performance of both types of systems! The main architectural difference between this scenario and satellites is that the access points don't move, while satellites can.

A sensor web network is very similar [Torr02]. Generally, sensor web networks consist of many wireless devices placed in situ into environments of interest. These wireless devices typically monitor the environment and transmit information automatically or by request to end-users by way of relay stations and pre-existing network infrastructure. It should be easy to see the relationship to terrestrial systems via Figure 50. A sensor web network is a terrestrial network that has gotten rid of the wired connection linking to the wireless access points. An important aspect of sensor web networks is they are generally stationary – the access points don't move and their location is known a priori.

Finally, one can generally think of ad-hoc systems as wireless access points whose position are not known a priori and can move randomly. Although ad hoc systems are generally located on the ground, they are more similar to satellite systems in that the access points are mobile. However, there is a major difference – even when satellites move, their position is known a priori as their future location in the orbit can be calculated very precisely. For this reason, distributed satellite communication systems can be considered a special case of ad hoc networks.

The inherent architectural relationships between these different systems means that results found for one likely apply to the others. Since distributed satellite communication systems are a special case of ad hoc, this means that ad hoc systems are more complex than even the satellite systems, and even more likely to experience strong coupling. Furthermore, it is very likely that time variability will generate greater coupling between the system and network architectures; ad hoc systems will be greatly affected by this as well!

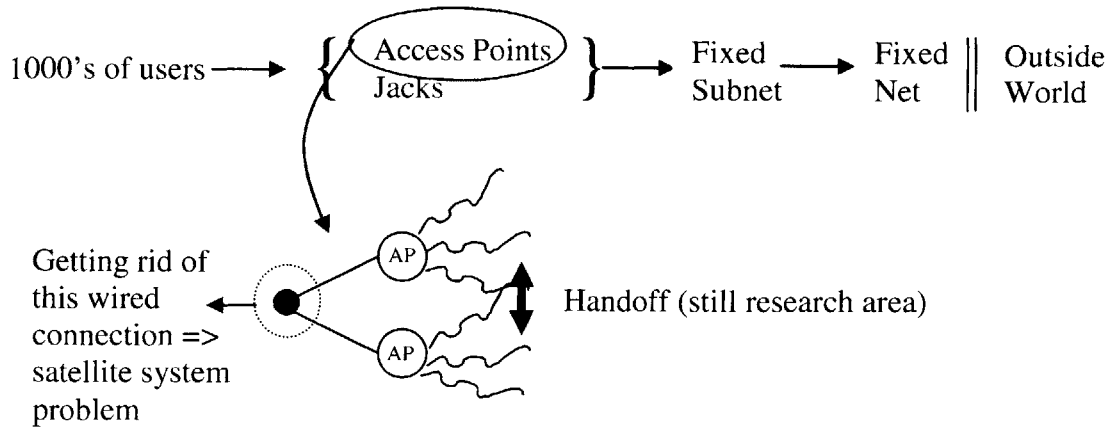


Figure 50: Relationship between DSC and Other Network Systems
This diagram illustrates the relationship between distributed satellite communication (DSC) and terrestrial, sensor web, and ad-hoc systems.

5.5. Future Work

A large motivator for this thesis is the desire to design distributed satellite communication systems that provide communication services that customers really want and at the quality they demand. This thesis has taken the first steps to understanding how to effectively integrate the system and network architectures to achieve this aspiration. However, there is much work left to be done. This section details a number of things that can be done to improve the simulation model used in this thesis, and encourages new areas of research that will be required for a comprehensive understanding of the interactions between the system and network architectures for these systems.

A major piece missing from this work is an examination of the effect of time variability on the strength of the interactions. Since the network protocols deal with phenomena that are time-varying, the performance of the network will also. Understanding how this interaction impacts the overall system design will be very important to the design of future systems.

As mentioned in the previous section, the handoff problem is still an active area of research. Efficient means of passing along customer calls and data connections as satellites or other mobile access points move out of sight (or the reverse problem of mobile users moving beyond the reach of stationary access points, such as in the cellular industry) are critical to reducing the probability of dropped connections in these systems. This thesis did not consider the impact of handoff as only steady-state averages were considered; time variability was not considered.

Similarly, there are a number of time-dependent quality of service metrics that will be important for any future work to consider. Among these is the probability of blocked calls, availability, call connectivity, probability of dropped calls, and so forth.

It will be important in future work to improve the cost models inherent in the simulation. Several deficiencies were in this thesis. For example, there are a number of costs that have not been considered due to lack of time or access to data, including the dollar cost of different multiple access schemes in terms of equipment, implementation, development, etc; and an accounting for end-user equipment cost. It may also be illuminating to consider sources of

revenue other than user subscription fees so that a more accurate estimation of the cost per user per month can be found.

On a similar note, it has become apparent during the course of this thesis that a more accurate model of potential market penetration be developed. A major flaw of the market potential model is that no accounting is made of the influence that quality of service and subscription costs play on market penetration and customer retention as a function of time. Thus, the major drivers involved the trade-off between life cycle cost, cost per user per month, quality of service, and the market potential are not truly captured. Furthermore, there are ways to significantly reduce the cost seen by the user. For example, the internal rate of return can be reduced; other potential sources of revenue can be explored, etc

Of course, there are many questions that have been raised during the course of this research. Among these are the following. How should the design process of these systems integrate the systems and network? Is it possible to find a subset of more optimal designs more quickly? Where would the industry benefit most in terms of advances in technology – antenna design, protocol design, or what? Would integrating the design of various network protocols reduce the coupling effect? How can the results of this thesis apply to terrestrial, sensor web, and ad hoc networks, and what gains can be made for these systems as a result? How does the location of the centralized router in a satellite network impact the quality of service performance?

An extremely important area of future research involves exploring how the system and network designers should value each of the performance metrics under their control in order to achieve a desired path through the design and objective spaces. Perhaps if this aspect of the overall design is understood, the computational costs can be reduced.

Finally, given the sparse nature of the design space considered, it would be prudent to expand on the design space exploration started here.

APPENDICES

Appendix A: Lexical Analysis Papers

Systems Papers	Network Papers
ChaD03	Arno97
Chai03	Arul94
deWe02	Chot00
Kash02	Feng01
Scia03	HuJi98
Shaw99	JunJ03
Spri03	LeeJ00
Suzu03	Mert98
	Moho02
	OrsT00
	SunZ00
	Svig02
	Wern95
	Wern97
	Wood01
	Zaim02

Appendix B: Lexical Analysis Keywords

Systems Keywords	Network Keywords
Economics	Routing
Market	Traffic Flow
Demand	Load Balancing
Policy	Protocols
Technology Infusion	Channels
Industry	Cells
Customers	Frequency Reuse
Business/Commercial Applications	Communication Paths/Links
Objectives	Satellite Hops
System Requirements	Queuing/Queues
System Drivers	Source/Origin
Operations	Destination/Sink
Deployment	Connectivity
Risk	Resource Management
Schedule	Scalability
Lifecycle Cost	Reachability
Conceptual Design/Stage	Overhead
Trade Studies	User Access
Trade Space	Network Interface
Trade-offs	Switching
Architectures	Mesh Topologies/Networks
System Reliability	Bottlenecks
Sensitivity	Link Capacity
Uncertainty	Network Management
Robustness	Network Control
Survivability	Link Acquisition
System Flexibility	Data Rate
Lifetime	Link Availability
Real Options	Blocking Probability
Supportability	Network Utilization
Reconfigurability	Multiplexing
Performance	Packets
System Utility	Traffic Load
Coverage	Traffic Class
Quality of Service	Network Requirements
	Bandwidth
	Mobility Management

REFERENCES

[Adam87] Adams, W.S., and Rider, L., "Circular Polar Constellations Providing Continuous Single or Multiple Coverage Above a Specified Latitude," *The Journal of the Astronautical Sciences*, Vol. 35, No. 2, pp 155-192, April-June 1987.

[AMDn04] "AMD and Cable & Wireless Enable Internet Connectivity and Computing Power Throughout the Caribbean with the Personal Internet Communicator," AMD Personal Internet Communicator News [online], December 1, 2004, URL: http://www.amdboard.com/pic_120104.html.

[Arno97] Arnon, Shlomi, and Kopeika, N.S., "Laser Satellite Communication Network – Vibration Effect and Possible Solutions," *Proceedings of the IEEE*, Vol. 85, No. 10, October 1997.

[Arul94] Arulambalam, A., and Ansari, N., "Traffic Management of a Satellite Communication Network Using Mean Field Annealing," IEEE (?), pp. 3577-3582, 1994.

[Bert92] Bertsekas, D., Gallager, R., "*Data Networks*", *2nd Ed.*, Prentice Hall, New Jersey, 1992.

[Bond00] Bonds, T., et. al., "Employing Commercial Satellite Communications: Wideband Investment Options for the Department of Defense," Project AIR FORCE, RAND, Arlington, VA, 2000.

[ChaD03] Chang, D., and de Weck, O.L., "Basic Capacity Calculation Methods and Benchmarking for MF-TDMA and MF-CMDA Communication Satellites," AIAA 21st International Communication Satellite Systems Conference, Yokohoma, Japan, April 15-19, 2003.

[ChaD04] Chang, D., "Quantifying Technology Infusion and Policy Impact on Low Earth Orbit Communication Satellite Constellations", M.S. Thesis, M.I.T., Department of Aeronautics and Astronautics, 2004.

[Chai03] Chaize, M., "Enhancing the Economics of Satellite Constellations via Staged Deployment and Orbital Reconfiguration," M.S. Thesis, M.I.T., Department of Aeronautics and Astronautics, 2003.

[Chan02] Chan, V., Lecture Notes from 16.399 Space Communications and Networks class, M.I.T., Department of Aeronautics and Astronautics, 2002.

[Chan04] Chan S., Samuels A., Shah N., Underwood J., de Weck O.L., "Optimization of Hybrid Satellite Constellations using Multiple Layers and Mixed Circular-Elliptical Orbits", AIAA-2004-3205, 22nd AIAA International Communications Satellite Systems Conference and Exhibit, Monterey, California, May 9-12, 2004.

[Chot00] Chotikapong, Y., Cruickshank, H., Sun, Z., and Evavns, B.G., "Network Architecture and Performance Evaluation of Broadband Satellite Systems, IEEE(?), 2000.

[Cons05] "Constructive Versus Existential Proofs", How to Write Proofs [online], URL: http://zimmer.csufresno.edu/~larryc/proofs/proofs.construct.html, [cited May 14, 2005].

[deWe02] de Weck, O.L., and Chang, D., "Architecture Trade Methodology for LEO Personal Communication Systems," 20th ICSSC, Montreal, Canada, May 12-15 2002.

[deWe03] de Weck, O., and Schindall, J., "Systems Architecting of an Integrated Earth Observation System for 2030," NASA Unsolicited Proposal, 2003.

[DLLP05] "LAN Data Link Layer Protocols", Protocols.Com [online resource], URL: http://www.protocols.com/pbook/lan.htm, [cited May 14, 2005].

[Feng01] Feng, Y., "Resource Allocation in Ka-band Satellite Systems," M.S. Thesis, Center for Satellite and Hybrid Communication Networks, University of Maryland, 2001.

[Fros02] "Satellite Telephone Quality of Service Comparison: Iridium vs. Globalstar," Frost & Sullivan, 2002.

[Gavi98] Gavish, B., Kalvenes, J., "The Impact of Satellite Altitude on the Performance of LEOS based Communication Systems," *Wireless Networks*, Vol. 4, pg. 199-213, 1998.

[GEOS05] "Geostationary, LEO, MEO, HEO Orbits" [online], URL: **www.geo-orbit.org/sizepgs/geodef.html**, [cited May 14, 2005].

[HAAR05] "Extremely Low Frequencies (ELF)", High Frequency Active Auroral Research Program (HAARP) Technical Details [online], URL: http://www.haarp.alaska.edu/haarp/elf.html [cited May 14, 2005].

[Haas02] Haase, E.E., Christensen, C.B., Ten Cate, H., "Global Commercial Space Industry Indicators and Trends", *Acta Astronautica*, Vol. 50, No. 12, pp. 747-757, 2002 (Published by Elsevier Science Ltd, Great Britain).

[HuJi98] Hu, J., Wu, S., Li, L., "Performance Analysis for Inter-satellite Link Networks of LEO/MEO Mobile Satellite Communication Systems," International Conference on Communication Technology, Oct 22-24, Beijing, China, 1998.

[HuYF99] Hu, Y.F., and Sheriff, R.E., "Evaluation of the European Market for Satellite-UMTS Terminals," *International Journal of Satellite Communications*, Vol. 17, Issue 5, pp 305-322, 1999.

[IGRP01] Haden, R., "IGRP", *Data Network Resource* [online resource], 2001, URL: http://www.rhyshaden.com/igrp.htm, [cited: May 14, 2005].

[INCO98] San Francisco Bay Area Chapter International Council on Systems Engineering. Systems Engineering Handbook. Technical report, INCOSE, January 1998.

[Infl05] Sahr, R., "Inflation Conversion Factors for Dollars 1665 to Estimated 2015" [online], Oregon State University, URL: http://oregonstate.edu/dept/pol_sci/fac/sahr/sahr.htm, [cited May 14, 2005].

[Inma05] "Inmarsat Service", Vietnam Telecom International website [online], URL: http://www.vti.com.vn/english/Inmarsat.html, [cited May 14, 2005].

[Inte02] "IntelliCell®: A Fully Adaptive Approach to Smart Antennas", ArrayComm, Incorporated document [online], URL: http://www.arraycomm.com/docs/IntelliCellWhitepaper.pdf, [cited May 14, 2005].

[InTH05] "Internetworking Technology Handbook", CISCO Systems [online handbook], URL: http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/, [cited May 14, 2005].

[Irid98] Nelson, R.A., "Iridium: From Concept to Reality", Reprinted [online] from *Via Satellite*, September 1998, URL: http://www.aticourses.com/iridium.htm, [cited May 16, 2005].

[Irid00] Ray, J., "Pioneering Iridium Satellite System Reaches Dead End", Spaceflight Now [online], March 18, 2000, URL: http://www.spaceflightnow.com/news/0003/18iridium/, [cited May 14, 2005].

[Jill97] Jilla, C.D., and Miller, D.W., "Satellite Design: Past, Present and Future", *International Journal of Small Satellite Engineering*, Vol. 1, Issue 1, ISSN 1360-7014, July 1997.

[Jogl01] Joglekar, N. R., Yassine, A. A., Eppinger, S. D., Whitney, D. E., "Performance of Coupled Product Development Activities with a Deadline," *Management Science*, Vol., 47, No.12, pp 1605-1620, December 2001.

[JunJ03] Jun, J., and Sichitiu, M.L., "The Nominal Capacity of Wireless Mesh Networks." *IEEE Wireless Communications*, pp 8-14, 2003.

[Kash02] Kashitani, T., "Development and Application of an Analysis Methodology for Satellite Broadband Network Architectures," 20th AIAA ISCSSC and Exhibit, Montreal, Canada, May 12-15 2002.

[Kucu03] Kukukates, R., and Ersoy, C., "High Performance Routing in a LEO Satellite Network," Computers and Communication 2003 (ISCC 2003) Proceedings, Eighth IEEE International Symposium, 2003.

[Kwok01] Kwok, Kenneth C.H., "Cost Optimization and Routing For Satellite Network Constellations," M.S. Thesis, MIT, 2001.

[Kwon98] Kwon, Y.H., Sung, D.K., "Analysis of Handover Characteristics in Shadowed LEO Satellite Communication Networks", 1998.

[Lang98] Lang, T.J., Adams, W.S., "A Comparison of Satellite Constellations for Continuous Global Coverage," J.C. Van der Ha (ed.), *Mission Design and Implementation of Satellite Constellations*, pp 51-56, 1998.

[LeeJ00] Lee, Jae-Wook, Lee, Jun-Woo, Kim, T., and Kim, D., "Satellite over Satellite (SOS) Network: A Novel Concept of Hierarchical Architecture and Routing in Satellite Network." *IEEE,* pp 392-399, 2000.

[Lutz00] Lutz, E., Werner, M., and Jahn, A., "Satellite Systems for personal and broadband Communications," Springer, 2000.

[Mert98] Mertzanis, I., Sfikas, G., Tafazolli, R., Evans, B.G., "Satellite-ATM Networking and Call Performance Evaluation for Multimedia Broadband Services," COST 252 TD(98) 38, Venice, Italy, 6 November 1998.

[Modi02] Modiano, E., Lecture Notes from 16.399 Space Communications and Networks class, M.I.T., Department of Aeronautics and Astronautics, 2002.

[Modi04] Modiano, E., Lecture Notes from 16.37 Data Networks class, M.I.T., Department of Aeronautics and Astronautics, 2004.

[Moho00] Mohorcic, M., Werner, M., Svigelj, A., Kandus, G., "Comparison of Adaptive Routing Algorithms in ISL Networks Considering Various Traffic Scenarios," 4th European Workshop on Mobile and Personal Communications, pg 72-81, London, UK, September 2000.

[Moho02] Mohorcic, M., Werner, M., Svigelj, A., Kandus, G., "Adaptive Routing for Packet-Oriented ISL Networks: Performance in Various Traffic Scenarios," *IEEE Transactions on Wireless Communications,* Vol. 1, No. 4, October 2002.

[Moho03] Mohorcic, M., Svigelj, A., Kandus, G., Hu, Y.F., Sheriff, R.E., "Demographically Weighted Traffic Flow Models for Adaptive Routing in Packet-Switched Non-Geostationary Satellite Meshed Networks," *Networks*, Volume 43, Issue 2, 7 October 2003. Pages 113-131.

[Nico03] Nicopolitidis, P., Obaidat, M.S., Papadimitriou, G.I, and Pomportsis, A.S., *"Wireless Networks,"* John Wiley & Sons, LTD., N.J., USA, 2003.

[Opti05] "Small Optical Telecommunications Satellite", *Optical Communications and Inter-satellite Links* [online article], URL: http://www.wtec.org/loyola/satcom2/03_06.htm, [cited May 14, 2005].

[Orfa04] Orfanidis, S.J., "Transmitting and Receiving Antennas", *Electromagnetic Waves & Antennas* [online], Chapter 14, pp 517, June 21, 2004, URL: www.ece.rutgers.edu/~orfanidi/ewa, [cited: May 14, 2005].

[OrsT00] Chotikapong, Y., Sun, Z., Ors, T., and Evans, B.G., "Network Architecture and Performance Evaluation of TCP/IP and ATM over Satellite," AIAA ICSSC and Exhibit, 18th, Oakland, CA, April 10-14, 2000, Collection of Technical Papers, Vol. 2.

[Poli98] Poli, C., and Carboni, G., "Lexical Analysis of Texts" [online], Fun Science Gallery, June 1998, URL: http://www.funsci.com/fun3_en/lexicon/handbook.htm, [cited May 14, 2005].

[Proa02] Proakis, J.G.,. and Salehi, M., "*Communication Systems Engineering", 2nd Ed.*, Prentice Hall, New Jersey, 2002.

[Rama96] Ramaswami, R., Sivarajan, K.N., "Design of Logical Topologies for Wavelength-Routed Optical Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 5, pg. 840-851, June 1996.

[Ride85] Rider, L., "Optimized Polar Orbit Constellations for Redundant Earth Coverage," *The Journal of the Astronautical Sciences*, Vol. 33, No. 2, April-June, 1985, pp 147-161.

[Sate05] "Iridium 9505 Portable Satellite Phone", SATWEST Phones & Pagers Product Page [online], URL: http://www.satwest.com/phone_mi9505_basic.html, [cited May 14, 2005].

[Sche96] Ronald J. Schertler, "Summary Report on key ACTS Experiments," AIAA ICSSC, 16th, Washington D.C., February 25-29, 1996, Technical Papers Pt 2 (A96-21571 04-32), Washington D.C., AIAA, 1996, pp 738-748.

[Scia03] Scialom, U., "Optimization of Satellite Constellation Reconfiguration," M.S. Thesis, M.I.T., Department of Aeronautics and Astronautics, 2003.

[Shaw99] Shaw, G. B., "The Generalized Information Network Analysis Methodology for Distributed Satellite Systems," Ph.D. Thesis, M.I.T., Department of Aeronautics and Astronautics, 1999.

[Sher01] Sheriff, R.E., and Hu, Y.F., "Mobile Satellite Communication Networks," John Wiley & Sons, LTD, 2001.

[Spri03] de Weck, O., Springmann P.N., Chang D.,"A Parametric Communications Spacecraft Model for Conceptual Design Trade Studies", Paper AIAA-2003-2310, 21st International Communications Satellite Systems Conference, Yokohama, Japan, 15-19 April, 2003.

[Stur96] Sturza, M.A., "The Teledesic Satellite System", 4th Budapest International Conference on Up-to-Date Satellite Communications, September 1996.

[SunZ00] Sun, Z., Chotikapong, Y., Chaisompong, C., "Simulation Studies of TCP/IP Performance over Satellite," AIAA, inc., AIAA-2000-1167, AIAA ICSSC and Exhibit, 18th, Oakland, CA, April 10-14, 2000, Collection of Technical Papers, Vol. 1.

[Suzu03] de Weck, O.L., Chang, D., Suzuki, R., Morikawa, E., "Quantitative Assessment of Technology Infusion in Communications Satellite Constellations, 21st ICSSC, 15-19 April 2003, Yokohama, Japan.

[Svig02] Svigelj, A., Mohorcic, M., Kandus, G., "Traffic Class Dependent Routing in Packet-Switched Non-Geostationary ISL Networks," Personal, Indoor, and Mobile Radio Communications, 2002, The 13th IEEE International Symposium on, Vol. 3, 15-18 September 2002.

[Torr02] Torres-Martinez E., Schoeberl M., Kalb M.W., "A Web of Sensors: Enabling the Earth Science Vision", IGARSS02-06-08:20, IGARSS 2002, Toronto, Canada, June.

[Turn02] Turner, A.E., "Constellation Design Using Walker Patterns," AIAA/AAS Astrodynamics Specialist Conference and Exhibit, 5-8 August 2002, Monterrey, CA.

[Vald03] Valdes-Dapena, P. "How's your cell service rate? J.D. Power study finds Verizon Wireless has best network quality while Alltel's is worst," [online], URL: http://money.cnn.com/2003/07/31/technology/cellular_survey, July 31, 2003, [cited May 13, 2005]

[Weis05] Weisstein, E.W., "Adjacency Matrix." From MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/AdjacencyMatrix.html, [cited May 14, 2005].

[Wern95] Werner, M., Jahn, A., Lutz, E., and Bottcher, A., "Analysis of System Parameters for LEO/ICO-Satellite Communication Networks. *IEEE Journal on Selected Areas in Communications,* Vol. 13, No. 2, February 1995.

[Wern97] Werner, M., Berndl, G., Edmaier, B., "Performance of Optimized Routing in LEO Inter-satellite Link Networks," *IEEE(?),* 1997.

[Wert99] Wertz, J.R., and Larson, W.J., "Space Mission Analysis and Design," 3rd Edition, Kluwer Academic Publishers, Boston, USA, 1999.

[Whip93] Whipple, D. P., "The CDMA Standard," *Applied Microwave and Wireless*, Winter 1993, pp. 27-37, 1993.

[Wood01] Wood, L., "Internetworking with Satellite Constellations," Ph.D. Thesis, University of Surrey, Centre for Communication Systems Research, 2001.

[WTDR03] "World Telecommunication Development Report 2003: Access Indicators for the Information Society", International Telecommunication Union Report [online], World Summit on the Information Society, Geneva, December 2003, URL: http://www.cnnic.net.cn/download/manual/international-report/wtdr_03.pdf, [Cited May 13, 2005].

[Zaim02] Zaim, A.H., Perros, H.G., Rouskas, G.N., "Performance Analysis of LEO Satellite Networks," E. Gregori et all (eds.), Networking 2002, LNCS 2345, pp 790-801, 2002.

[Zhan03] Zhang, Y. (ed.), Wood, L., "Satellite Constellation Networks", *Internetworking and Computing over Satellite Networks,* Chapter 2, Kluwer Academic Press, ISBN 1-4020-7424-7, pp 13-34, March 2003.