Techno-economical Evaluation of Intra-datacenter Optical Transceiver Designs

by

Wei Yu

B.S., Materials Physics
University of Science and Technology of China, 2011

SUBMITTED TO THE DEPARTMENT OF MATERIALS SCIENCE AND ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2018

Signature of Author: ______ Signature redacted ________________
Department of Materials Science and Engineering
June 8th, 2018

Certified by: ____________________ Signature redacted ______
Lionel C. Kimerling
Thomas Lord Professor of Materials Science and Engineering
Thesis Supervisor

Accepted by: ____________________________ Signature redacted __
Donald Sadoway
John F. Elliott Professor of Materials Chemistry
Chair, Departmental Committee on Graduate Student

Techno-economical Evaluation of Intra-datacenter Optical Transceiver Designs

by

Wei Yu

Submitted to The Department of Materials Science and Engineering on May 17th, 2018 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Materials Engineering

Abstract

The evolution of data center network interconnects is based: at the component level, on cost, power, and bandwidth density; and at the system level, on cost, unit count, footprint, and implementation time. The decision for iterative vs. transformational design depends critically on the product horizon view that amortizes R&D, manufacture tooling and infrastructure preparation. An iterative platform can be implemented earlier, but it may present limited performance scalability. A transformational design can have large creation and implementation costs that are amortized over a longer time window. A framework for quantitatively capturing these variables using a new technology inventory model is developed. Scenarios for deployment of six different transceiver package platforms to meet the projected data center capacity scaling ramp are constructed. The iterative designs provide better return on investment for a 3-year time window, while the transformational designs are optimized for a 20-year window.

Thesis Supervisors:

Lionel C. Kimerling
Title: Thomas Lord Professor of Materials Science and Engineering

Acknowledgements

I want to thank my advisor Professor Lionel Kimerling, who has been unreservedly encouraging me, supporting me and believing me. I want to thank Dr. Randolph Kirchain and Professor Elsa Olivetti for guiding me and sitting in my thesis committee. I want to thank Dr. Jurgen Michel and Dr. Yan Cai from EMAT for teaching the design and processing of Germanium lasers.

I want to thank Dr. Peter O'Brien, Dr. Lee Carroll from Tyndall National Institute, Mr. Ashish Kelbar, Dr. Katharine Schmidtke, Dr. Naadar Hasani, Dr. Zhiping Yao from Facebook, Mr. Dick Otte, Dr. Madelaine Glick (now in Columbia), Dr. Terry Smith, Mr. Tom Marrapode from PSMC Roadmap TWG, Dr. Peter De Dobbelaere from Luxtera, Dr. Wei-ping Huang, Dr. Yu Zhu from Hisense, Dr. Qianggao Hu from Accelink for generously sharing their valuable knowledge and visions with me.

Lastly but not the least, I want to thank my family and friends. I feel very lucky that I can have the chance to enjoy wonderful moments with you.

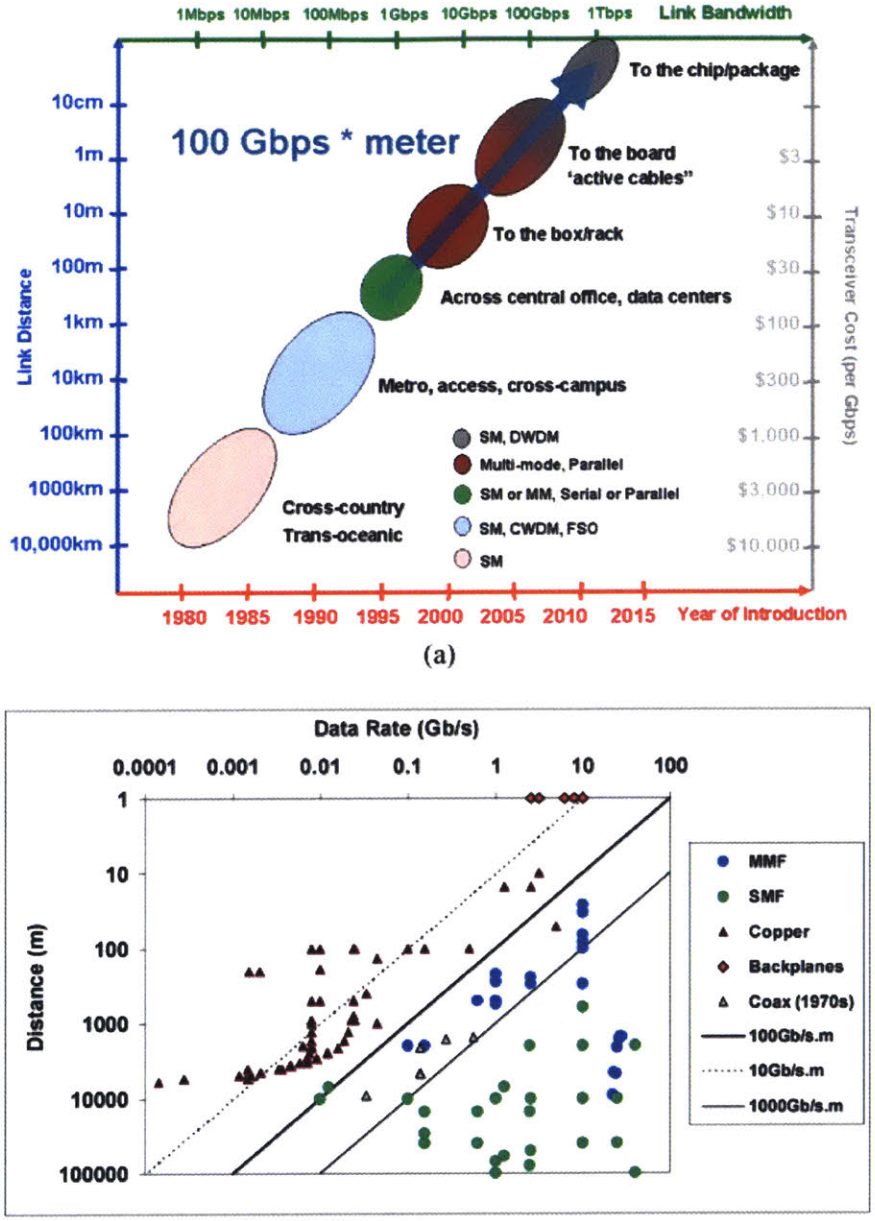
Table of Contents

Chapter 1
Introduction

Optical Transceivers

An optical transceiver is a device that sends and receives data using optical fiber rather electrical wire. The device is usually composed of two parts: a transmitter that converts electrical signals into optical signals before transmitting data and a receiver that converts optical signals back to electrical signals after receiving data. The transmitter requires a modulated light source, an electrical IC block that drives the light source or the modulator. The receiver requires a photodetector, an electrical IC block that turns current signals to voltage signals (e.g. transimpedance amplifiers) and an electrical IC block that converts analog signals into digital signals (post amplifiers). Besides, for high speed signals (e.g. 25G per channel), there are also clock and data recovery blocks to extract timing information and serializer/deserializer blocks to compensate for limited input/output. For even higher speeds (e.g. 100G per channel), digital signal processing blocks for signal modulation (e.g. PAM4) are needed. Those are the main functional blocks. To make them work together as a device, there are additional electrical interconnect components (e.g. printed circuit board and bonding wires) and optical alignment components (micro-lens, optical benches, and mechanical housings).

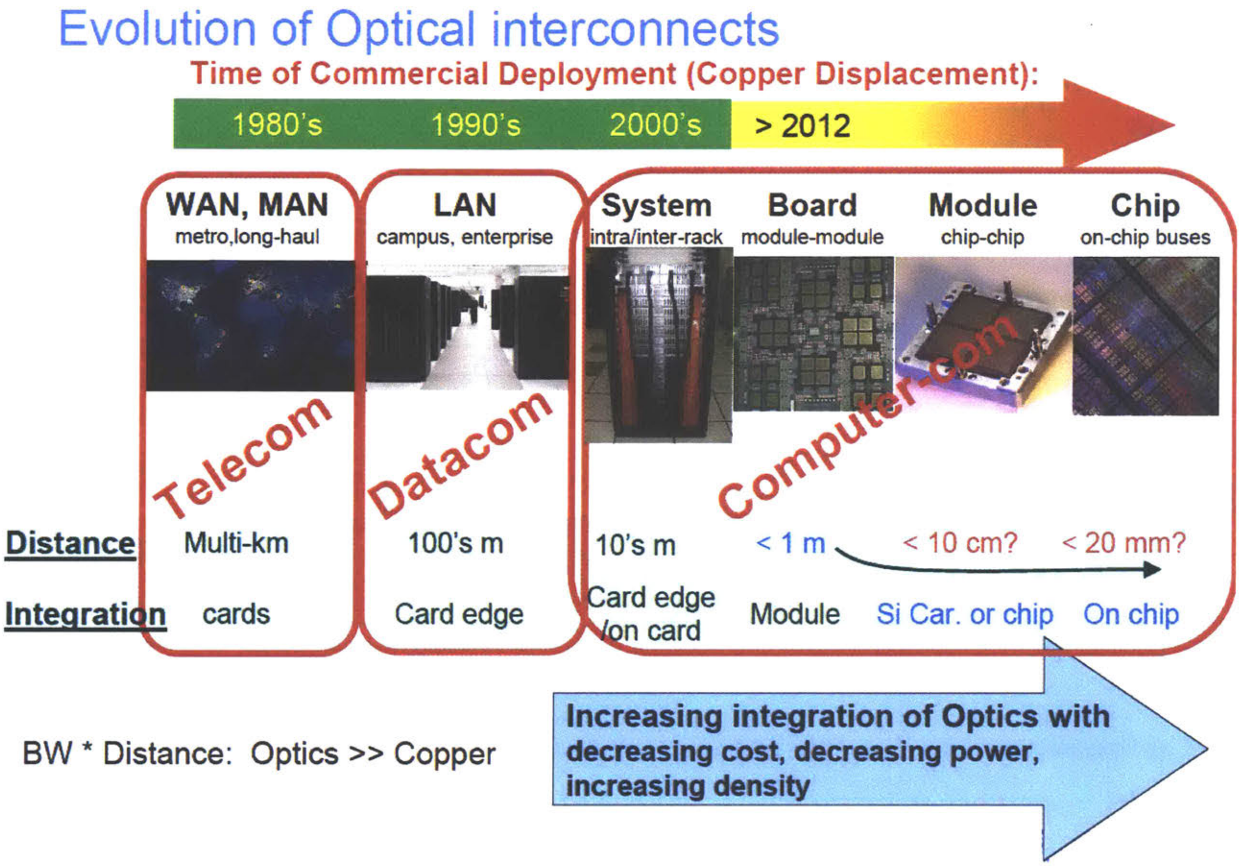
Optical transceivers arc signal converters between electrical and optical signals. Together with optical transmission media such as optical fibers and waveguides et al, they form optical interconnects between electrical circuits-based servers and network switches. By far the information computation and storage are still dominated by electrical circuits. Naturally the information transmission should be also done by electrical interconnects such as on-chip metal lines, backplane PCB, copper cables et al. However, those electrical interconnects are subject to basic physical limits such as RC delay at high bandwidth.[1] During the exponential bandwidth scaling, electrical interconnects are losing their fields to optical interconnects. [2]



(a)

(b)

Figure 1 Penetration of optical links into communications. (a) Historical roadmap for introduction of optical interconnections into digital systems. (b) Data rate versus distance of commercial electrical versus optical links. [3]



Evolution of Optical interconnects

Time of Commercial Deployment (Copper Displacement):

1980's 1990's 2000's > 2012

	WAN, MAN metro,long-haul	LAN campus, enterprise	System intra/inter-rack	Board module-module	Module chip-chip	Chip on-chip buses
	Telecom	Datacom	Computer-com			
Distance	Multi-km	100's m	10's m	< 1 m	< 10 cm?	< 20 mm?
Integration	cards	Card edge	Card edge /on card	Module	Si Car. or chip	On chip

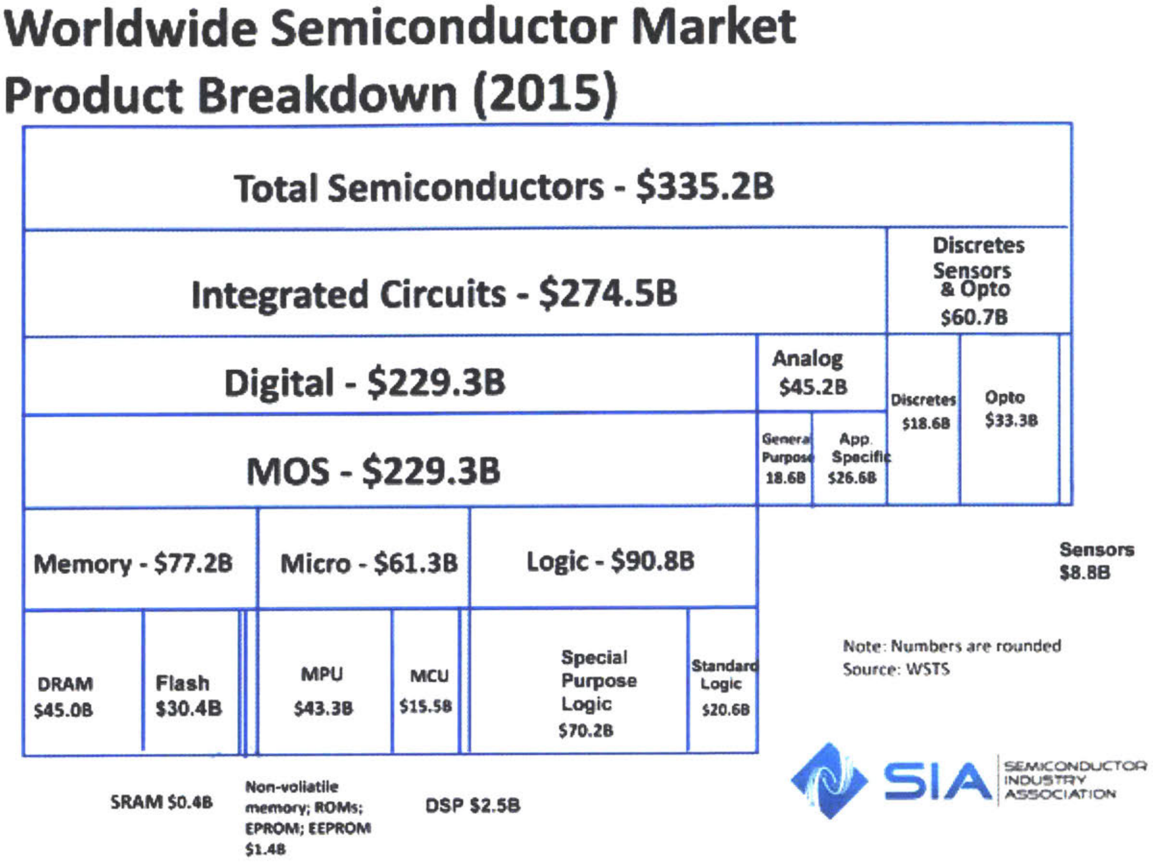
BW * Distance: Optics >> Copper

Increasing integration of Optics with decreasing cost, decreasing power, increasing density

Figure 2 Copper interconnects are being displaced by optical interconnects [4]

Barriers for entry

The forecasted growth rate for fiber-optic laser transmitters is highest among all kinds of optoelectronic product categories but the total market is still small. A market size of $1.2 billion, is far below the total semiconductor market of $335 billion [5] and the network equipment and software market of $200 billion [6]. The data communication market, the current battleground between optical and copper interconnects, is still dominated by copper [7].



Worldwide Semiconductor Market Product Breakdown (2015)

Total Semiconductors - $335.2B

Integrated Circuits - $274.5B

Discretes Sensors & Opto $60.7B

Digital - $229.3B

Analog $45.2B

Discretes $18.6B

Opto $33.3B

MOS - $229.3B

General Purpose 18.6B

App. Specific $26.6B

Memory - $77.2B

Micro - $61.3B

Logic - $90.8B

Sensors $8.8B

DRAM $45.0B

Flash $30.4B

MPU $43.3B

MCU $15.5B

Special Purpose Logic $70.2B

Standard Logic $20.6B

SRAM $0.4B

Non-voliatile memory; ROMs; EPROM; EEPROM $1.4B

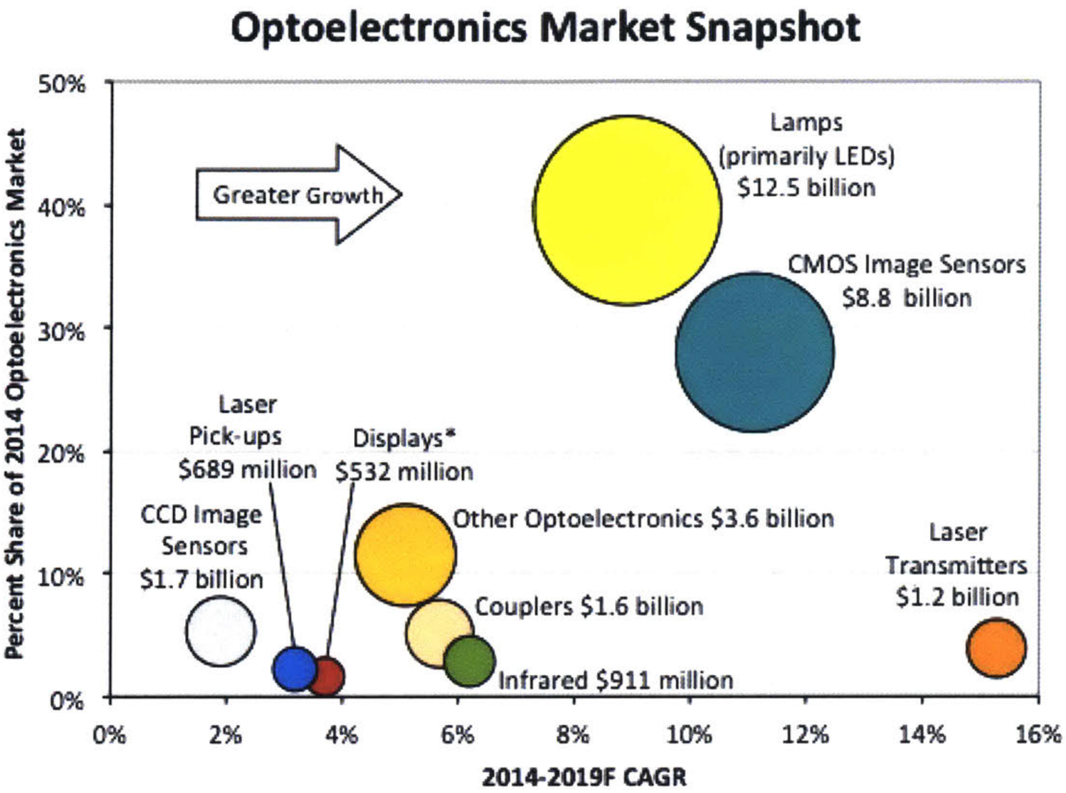
DSP $2.5B

Note: Numbers are rounded

Source: WSTS

SIA SEMICONDUCTOR INDUSTRY ASSOCIATION

Figure 3 The market share of optoelectronics is only 10% of the total semiconductors market [5]



*Includes character displays as complete assemblies but excludes LCDs and other flat-panel display technologies

Source: IC Insights

Figure 4 The growth rate for fiber-optic laser transmitters is forecasted to be as high as 15% [8]

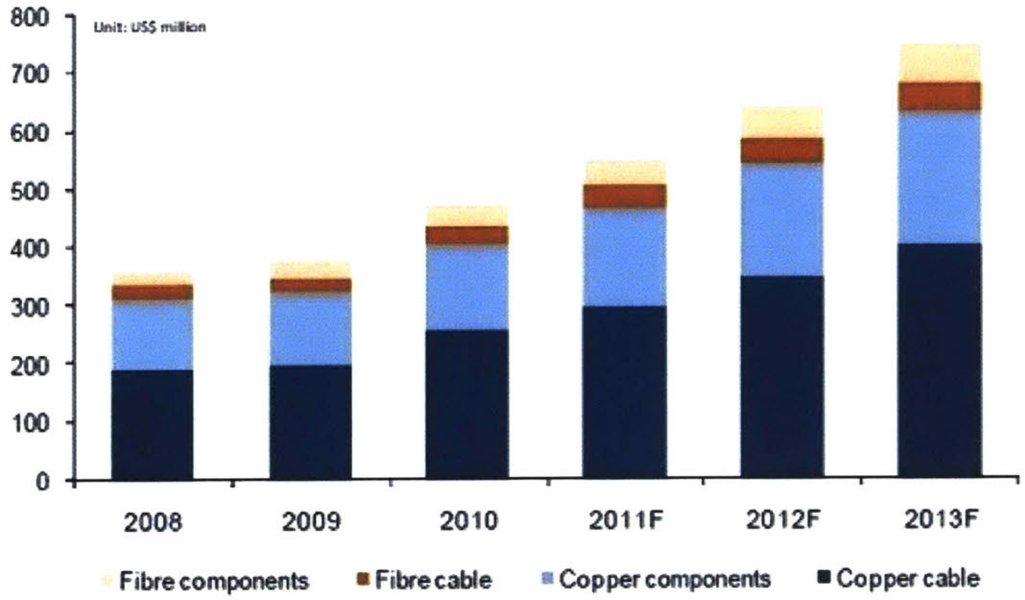


Figure 5 Copper interconnect still dominates the data communication market [7]

In 1997, Fine and Kimerling pointed out that there were three barriers for optical communication products and optoelectronic industry in general: supply chain disaggregation, insufficient government support and lack of a killer application. [9] In addition, general studies about market entry barriers can also be applied to optical interconnect products. Cost advantages of incumbents are often identified as the most crucial gating factor [10]. The communication interconnect community agrees with that point and often lists cost as the first and the important metrics for evaluation when comparing different technologies. [11] Besides, Porter's theory suggests that early market entrants have advantages in customer switching cost [12]. For example, network switch ASICs are still setting the loss margin based on copper cables. Another example is that installation and maintenance staff are still used to the flexibility that copper cables are pluggable and flexible and complain that optical fibers are fragile and sensitive to dust.

Barriers	Implications	For optical interconnects
Cost advantage of incumbents	One of the most important entry barriers, and usually results from economies of scale and learning curve effect.	Cost is the most important metrics.
Product differentiation of incumbents	Established firms have brand identification and customer loyalties due to advertising, being first in a market, customer service, or product differences.	There is not much discussion yet.
Capital requirements	Need to invest large financial resources in order to compete or enter a market constitutes	Both government and industry are making heavy investments, e.g. AIM Photonics in U.S.,

	barrier to entry, and is higher in capital-intensive industries.	Photonics21 in E.U., PETRA in Japan, A*STAR IME in Singapore, MOST plan on "Photonics and microelectronics device and integration" in China et al. .
Customer switching cost	Switching costs prevent the buyer from changing suppliers, and technological changes often raise or lower these costs.	Network switch ASICs are still setting the loss margin based on copper cables. Installation and maintenance staff are still used to the flexibility that copper cables are pluggable and flexible and complain that optical fibers are fragile and sensitive to dust.

Access to distribution channels	First or early market entrants use intensive distribution strategies to limit the access to distributors for the potential market entrants.	There is not much discussion yet.
Government policy	Government limits the number of firms in a market by requiring licenses, permits, etc.	There is not much discussion yet.
Advertising	Heavy advertising by firms already in the market increases the cost of entry for potential entrants and affects brand loyalty as well as the extent of economies of scale by causing cost per	There is not much discussion yet. Besides, copper cables firms are also investing on optical interconnects for diversification.

	dollar revenues to decline.	
Number of competitors	Market entry is expected to be more likely during periods of increasing incorporations and less likely after a lag, during periods when high numbers of business failures occur.	There is not much discussion yet.
Research and development	The barrier is usually short-lived. Incumbent firms may prevent the entry of new firms by investing effectively in R&D, which increases technological scale economies and forces the ongoing industry context to evolve in a way that	It is true that there are still investments on R&D of copper cables but the market of copper cables is keeping shrinking. (The reach of copper cables shrinks by 40% every generation.)

	would make subsequent attempts to enter even more ineffectual.	
Price	Price warfare can be a significant deterrent to entry, particularly in industries where firms are more likely to lower their prices to fill underutilized plants.	There is not much discussion yet.
Technology and technology change	Usually present in high tech industries and can actually raise or lower economies of scale, which is one of the major sources of cost advantages.	The point is incorporated in the "consumer switching cost" barrier
Market concentration	The influence and impact of concentration on entry appear to be minimal.	There is not much discussion yet.

Seller concentration	Entry is unlikely to be as easy in highly concentrated as in less concentrated markets.	There is not much discussion yet.
Divisionalization	Only expected in exceptionally profitable oligopolistic industries, Incumbent firms create new independent divisions more cheaply than potential entrants who must incur additional overhead costs for entry.	There is not much discussion yet.
Brand name or trademark	New entrants to an industry are denied the benefits of brand name created by others as a result of the exclusive rights to use given with a	There is not much discussion yet.

	trademark. Usually a weak barrier.	
Sunk costs	Contribute to entry barriers that can also give rise to monopoly profit, resource misallocation, and inefficiencies.	There is not much discussion yet.
Selling expenses	Shifts in demand functions can result from selling efforts making market entry endogenous.	There is not much discussion yet.
Incumbent's expected reaction to market entry	May deter market entry only if the incumbent firms are able to influence potential entrants' expectation about the post-entry reaction of the incumbents.	There is not much discussion yet.
Possession of strategic raw materials	Access to strategic raw materials contributes to	There is not much discussion yet. Besides,

	firms' absolute cost advantages	the key raw materials for copper interconnects and optical interconnects are different.

Table 1 List of market entry barriers from literature [10]

There have been quite a few solutions suggested to accelerate the technology diffusion and help optical interconnect products overcome those commercialization barriers. One is vertical integration of the supply chain through ownership or knowledge sharing [9], [13], [14]. However, after the dotcom bubble, the ownership approach sounds like a minefield for many U.S. companies, although Asian companies, such as Samsung, are still taking the ownership approach. As for knowledge sharing, companies are worried about intellectual properties. For example, Samsung, as a supplier of Apple's iPod and iPhone products since 2005, gained critical knowledge of technologies and market, launched its own smart phone in 2010 and has since become Apple's largest competitor. [14]

Another proposed solution is standardization [15]. However, starting from the strategic planning phase, different interest groups are arguing for different standards. The market is highly fragmented. For example, for post-100G intra-datacenter optical transceiver products, there are two speeds (200G & 400G), three form factors (QSFP-DD, OSFP and COBO) and three combinations of channels (2, 4 and 8 channels) being discussed. When there are so many standards coexisting, the benefit of standards become nominal.

New environment, new way of thinking

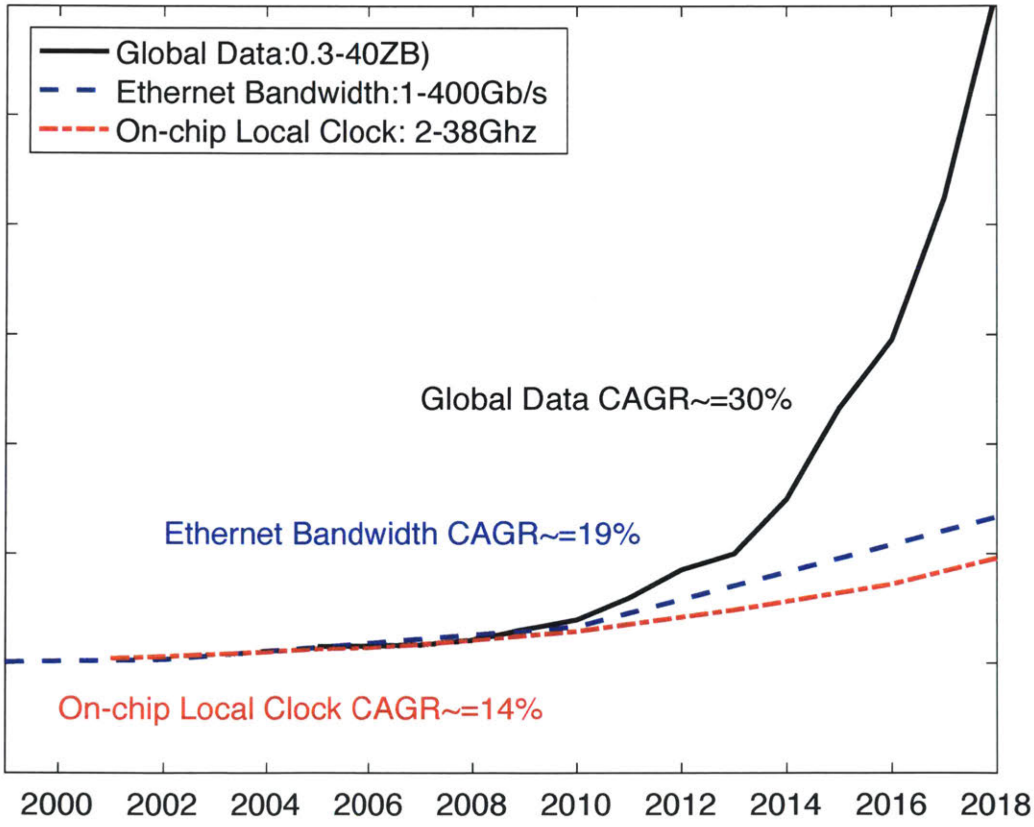


Figure 6 The increasing gap between component performances and the data volume [16]–[18]

The data communication market is new for optical transceiver vendors. The optical interconnect technology is also new for data center customers. What's more, the world of information technology is experiencing the ending of Moore's law on the side of technology supply and the booming of big data on the side of market demands. Moore's law has been giving the industry the confidence to believe that people can improve cost and power efficiency of logic and memory devices through shrinking the size of transistor. Now the technology node has been down to sub-10 nm and both the quantum effect and the fabrication challenge start to put the brake on the scaling law. [19]–[21] On the other hand, the amount of big data tasks and the size of a single task are keeping increasing [22], [23]. It requires more and more powerful logic and memory units. When the logic and memory units are failing the demanded performance scalability individually, system-level redesigns, such as GPU computing [24], [25] and distributed storage [26], [27], become necessary to realize efficient parallelism. On the algorithm level, a big task is decomposed into small sub-tasks. On the hardware level, logic, memory and storage units are re-arranged to facilitate the communication within each sub-task and between different sub-tasks.

When the environment changes, straight extrapolations of past successes may cause misunderstandings that impede future successes. Through my field studies and interviews with different stakeholders, I have discovered a few points to notice.

1. The evaluation of optical transceiver should be done at a system level.

Location	Relevant design scope
Edge of network (Internet network)	Distance, bandwidth, (latency), cost, power, size, of the link (fiber + 2 transceivers)
Edge of system (Intra-datacenter network)	Distance, bandwidth, (latency), cost, power, size of the network (link + switch)
Edge of device/chip	(Latency), cost, power, size of the system (network + CPU/GPU + memory + hard disk)

Table 2 The system scope is expanding

In a telecommunication network, the physical network topology is almost fixed and the performance and cost of optical transceivers is discussed within the scope of an interconnect link. In an intra-datacenter network, the physical network topology becomes a decision variable which is coupled with the choice of optical transceivers. The performances and costs of different optical transceivers are being examined at the network scale. For the discussion of resource-centric disaggregated datacenters, the performances and costs of network, CPU/GPU, memory and hard disks are evaluated together as a system [28], [29]. Nowadays, instead of competing with other optical transceivers only, optical transceivers are even facing the challenge of GPUs for cost-effectiveness. [30] An updated techno-economical evaluation of optical transceiver designs should take system integration into account rather than look at optical transceiver devices only.

2. The long procurement lead time of optical transceivers affects customers' adoption.

For data centers, the choice of optical transceivers, the fiber set-up and the network switches are coupled. In 2016-2017, when data centers started to install 100G optical transceivers inside a data center, the delivery of orders were delayed for six months to a year. That unexpected long procurement lead time of optical transceivers not only postponed the network production start but also put their inventories of other network components at risk. When data centers make decisions for next purchases, the length of procurement lead time is influencing their choice. For example, for next-generation network, a network customer may choose the old generation optical transceivers of 100G bandwidth over the new generation optical transceivers of 400G bandwidth to avoid supply risks.

3. The product development cost is non-trivial for intra-datacenter optical transceiver products

Due to the need of dynamically re-assignment of resources such as virtual machines and storage volumes regardless of physical distances, data centers prefer homogeneous network interconnect fabrics [31]. One hyperscale data center may install 100s of thousands of optical transceiver devices of the same design. Some people think the upfront investment on product development become trivial at such a high volume and focus on minimizing manufacturing cost only. However, the product development cost is actually non-trivial. "It takes more than a thousand wafers to qualify the process flow of silicon photonic chips in the foundry", introduced an engineering lead of a silicon photonics company which chose a two-chip design over a single-chip design to save product development cost.

An evaluation framework for integrating stakeholders

There have been several metrics proposed to compare different optical transceiver designs to provide guidance on optical transceiver research and product planning. In optical communication systems, a widely adopted metric is the bandwidth-distance product. [32]In 2005, D. Neilson et al. from Bell Labs presented a chart of data rate and data transmission distance to compare different interconnect technologies at European Conference on Optical Communication.[3] In 2011, E. Fuchs et al. from MIT estimated the manufacturing cost of different 100G optical transceiver designs.[33] In In 2011, Dr. A. Vahdat et al. from Google put forward a few requirements for data center interconnects: lower cost, larger scale, faster switching time and lower insertion loss.[34] That set of metrics introduced cost into the evaluation. In 2012, Dr. M. Taubenblatt from IBM proposed a spider plot composed of six factors to compare optical interconnects for high performance computers: 1/cost, density, 1/power, reliability, 1/BER, latency. [33] In 2013, the communications technology roadmap working group led by Professor L. Kimerling at MIT Microphotonics Center came up with three key metrics: cost, energy, bandwidth density and scalability. In 2015, D. Mahgerefteh et al. from Finisar stated that single-channel solutions would be more cost-effective than multi-channel solutions and among single-channel solutions a VCSEL-based solution would be the most cost-effective one due to its simple structure and large alignment tolerance of multi-mode light coupling.[35] In 2016, Dr. A. Chakravarty et al. mentioned that Facebook had shrunk the required temperature range and shortened the communication distance to enable lower cost points.[36] In 2017, Mr. D. Zhuo et al. from Microsoft, Columbia University and University of Washington argued that the BER of 10^{-6} was a legacy from telecom and more than needed in the data center applications. They pointed out optical links were major cost drivers of data center network and designed a system of redundant arrays of inexpensive links to stretched transceivers' reach and reduced the cost by 44%.[37] In 2017, Dr. R. Urata et al. introduced the evolution of Google datacenter network bi-section bandwidth from 1Tbps to 1000Tbps within 8 years and advocated the reduction of power and cost of optical transceivers through further integration within the device. Still, in the same year, X. Zhou et al. from Google gave out a ranked set of criterions: 1. bandwidth cost, 2. power consumption, 3. serviceability, 4. latency. [48]

With the existing literature, there are still a few questions remaining to be answered.

1) Most of the literatures agree that cost would be the most important metrics but which exact cost is the cost that matters? Is the selling price of the device? What factors other than the manufacturing cost may affect the selling price? When the adoption of different transceiver devices means setting up network topologies differently, should the system setup cost be included? To help

different stakeholders communicate efficiently and reach an agreement, a clear definition of cost is necessary.

2) How should the trade-offs among different kinds of metrics be evaluated and balanced? By far, there are many metrics that have been put forward, such as various bandwidth cost, power, size, distance, number of wavelengths, number of fibers, reliability, serviceability, scalability and latency. None of the existing solutions dominates in every aspect. To rank different solutions, people need to know an effective way to make trade-offs among different aspects.

3) Time should also be an indispensable dimension in the evaluation. In general, time-to-market is a crucial metrics for comparing different goods and studies have been done to balance the trade-off between time-to-market and product performance [38], [39]. In the optical transceiver case, time is as crucial as in general, if not more. Google's data center network bandwidth expands 1000 times within 8 years. An OFC (the Optical Networking and Communication Conference & Exhibition) report also writes that large scale Could Data Center need to upgrade networking hardware every two years [46]. To ensure a smooth large-scale network upgrade, it is critical for data centers to have the optical transceivers ready with the promised cost, performance and quantity on time.

Here I design a systematic techno-economic evaluation framework, in which there are six factors identified as the critical path elements for decision process. Those factors form a two-by-three matrix. The "two" are the two dimensions: cost and time. The "three" are the three phases: product development, mass production and system integration. Taking into account of multiple criteria in decision making, the framework integrates interests of different stakeholders. I hope such a framework can help facilitate communication among different groups and accelerate the resolution of conflicts and disagreements during strategic product planning. In the following chapters, I examine those factors and the interactions among them in detail.

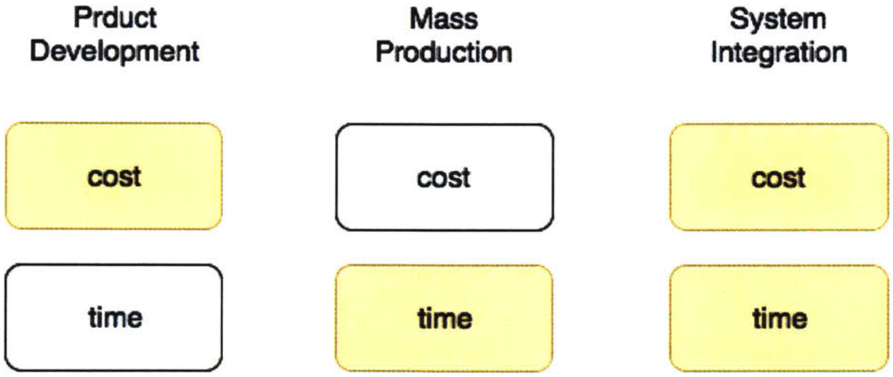


Figure 7 A six-component techno-economic evaluation framework integrating different stakeholders. The white blocks in the figure are factors that have been explored before and are the focus of previous techno-economical evaluations. The yellow blocks in the figure are factors that are often overlooked.

[1] D. A. B. Miller, “Device Requirements for Optical Interconnects to Silicon Chips,” *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, Jul. 2009.

[2] D. T. Neilson, D. Stiliadis, and P. Bernasconi, “Ultra-high capacity optical IP routers for the networks of tomorrow: IRIS Project,” in *2005 31st European Conference on Optical Communication, ECOC 2005*, 2005, vol. 5, pp. 45–48 vol.5.

[3] A. V. Krishnamoorthy *et al.*, “Progress in Low-Power Switched Optical Interconnects,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 17, no. 2, pp. 357–376, Mar. 2011.

[4] MIT Microphotonics Center, “On-Board Optical Interconnection.” Apr-2013.

[5] “The Semiconductor Market: 2015 Performance, 2016 Forecast, and the Data to Make Sense of It | SEMI.ORG.” [Online]. Available: http://www.semi.org/en/semiconductor-market-2015-performance-2016-forecast-and-data-make-sense-it. [Accessed: 14-May-2018].

[6] “Infonetics: US$1 Trillion to Be Spent on Telecom and Datacom Equipment and Software Over Next 5 Years.” [Online]. Available: https://finance.yahoo.com/news/infonetics-us-1-trillion-spent-211622815.html. [Accessed: 14-May-2018].

[7] Amclia.Liu, “How Far Will Copper Network Go,” *Fiber Optic Cabling Solutions*, 11-Jun-2013. .

[8] “Three trends driving optoelectronics market growth through 2019 | Solid State Technology.” [Online]. Available: http://electroiq.com/blog/2015/06/three-trends-driving-optoelectronics-market-growth-through-2019/. [Accessed: 14-May-2018].

[9] C. H. Fine and L. C. Kimerling, “Biography of a Killer Technology,” *Spec. Rep. Optoelectron. Ind. Dev. Assoc.*, p. 23, Jun. 1997.

[10] F. Karakaya and M. J. Stahl, “Barriers to Entry and Market Entry Decisions in Consumer and Industrial Goods Markets,” *J. Mark.*, vol. 53, no. 2, pp. 80–91, 1989.

[11] X. Zhou, H. Liu, and R. Urata, “Datacenter optics: requirements, technologies, and trends (Invited Paper),” *Chin. Opt. Lett.*, vol. 15, no. 5, p. 120008, May 2017.

[12] M. E. Porter, “Technology and competitive advantage,” *J. Bus. Strategy*, vol. 5, no. 3, pp. 60–78, 1985.

[13] C. H. Fine, *Clockspeed: Winning Industry Control in the Age of Temporary Advantage*. ReadHowYouWant.com, 2010.

[14] R. W. Seifert and O. H. Isaksson–October, “The Perks and Pitfalls of Knowledge Diffusion in the Supply Chain,” 2013.

[15] G. Tassey, "Standardization in technology-based markets," *Res. Policy*, vol. 29, no. 4–5, pp. 587–602, Apr. 2000.

[16] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper - Cisco." [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. [Accessed: 24-May-2018].

[17] "Singularity is Near -SIN Graph - Micro Processor Clock Speed." [Online]. Available: http://www.singularity.com/charts/page61.html. [Accessed: 24-May-2018].

[18] "The Ethernet community is working to introduce six new rates in the next 3 years | Network World." [Online]. Available: https://www.networkworld.com/article/3060250/lan-wan/the-ethernet-community-is-working-to-introduce-six-new-rates-in-the-next-3-years.html. [Accessed: 24-May-2018].

[19] "Moore's law," *Wikipedia*. 10-Sep-2017.

[20] T. Simonite, "Intel is putting the brakes on Moore's Law," *MIT Technology Review*. [Online]. Available: https://www.technologyreview.com/s/601102/intel-puts-the-brakes-on-moores-law/. [Accessed: 24-May-2018].

[21] J. Burt, “Memory And Logic In A Post Moore’s Law World,” 22-Mar-2017. [Online]. Available: https://www.nextplatform.com/2017/03/22/memory-logic-post-moores-law-world/. [Accessed: 24-May-2018].

[22] A. Canziani, A. Paszke, and E. Culurciello, “An Analysis of Deep Neural Network Models for Practical Applications,” *ArXiv160507678 Cs*, May 2016.

[23] S. Miao, K. Hendrickson, and Y. Liu, “Computation of three-dimensional multiphase flow dynamics by Fully-Coupled Immersed Flow (FCIF) solver,” *J. Comput. Phys.*, vol. 350, pp. 97–116, Dec. 2017.

[24] J. Nickolls and W. J. Dally, “The GPU Computing Era,” *IEEE Micro*, vol. 30, no. 2, pp. 56–69, Mar. 2010.

[25] F. Liu, N. Luehr, H. J. Kulik, and T. J. Martínez, “Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models,” *J. Chem. Theory Comput.*, vol. 11, no. 7, pp. 3131–3144, Jul. 2015.

[26] S. Ghemawat, H. Gobioff, and S.-T. Leung, “The Google File System,” in *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, New York, NY, USA, 2003, pp. 29–43.

[27] F. Chang *et al.*, “Bigtable: A Distributed Storage System for Structured Data,” *ACM Trans Comput Syst*, vol. 26, no. 2, pp. 4:1–4:26, Jun. 2008.

[28] S. Han, N. Egi, A. Panda, S. Ratnasamy, G. Shi, and S. Shenker, “Network support for resource disaggregation in next-generation datacenters,” 2013, pp. 1–7.

[29] B. Abali, R. J. Eickemeyer, H. Franke, C.-S. Li, and M. A. Taubenblatt, "Disaggregated and optically interconnected memory: when will it be cost effective?," *ArXiv150301416 Cs*, Mar. 2015.

[30] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," *ArXiv171201887 Cs Stat*, Dec. 2017.

[31] "Designing the Flat Data Center Network," *Electronic Component News*, 01-Mar-2011. [Online]. Available: https://www.ecnmag.com/article/2011/03/designing-flat-data-center-network. [Accessed: 24-May-2018].

[32] T. Li, "Optical Fiber Communication-The State of the Art," *IEEE Trans. Commun.*, vol. 26, no. 7, pp. 946–955, Jul. 1978.

[33] E. R. H. Fuchs, R. E. Kirchain, and S. Liu, "The Future of Silicon Photonics: Not So Fast? Insights From 100G Ethernet LAN Transceivers," *J. Light. Technol.*, vol. 29, no. 15, pp. 2319–2326, Aug. 2011.

[34] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, "The Emerging Optical Data Center," presented at the Optical Fiber Communication Conference, 2011, p. OTuH2.

[35] D. Mahgerefteh and C. Thompson, “Techno-economic Comparison of Silicon Photonics and Multimode VCSELs,” in *Optical Fiber Communication Conference (2015), paper M3B.2*, 2015, p. M3B.2.

[36] A. Chakravarty, K. Schmidtke, S. Giridharan, J. Huang, and V. Zeng, “100G CWDM4 SMF optical interconnects for facebook data centers,” in *2016 Conference on Lasers and Electro-Optics (CLEO)*, 2016, pp. 1–2.

[37] D. Zhuo *et al.*, “{RAIL}: A Case for Redundant Arrays of Inexpensive Links in Data Center Networks,” 2017. .

[38] R. E. Burkart, “Reducing R&D cycle time,” *Res. Technol. Manag.*, vol. 37, no. 3, p. 27, Jun. 1994.

[39] M. A. Cohen, J. Eliashberg, and T.-H. Ho, “New Product Development: The Performance and Time-to-Market Tradeoff,” *Manag. Sci.*, vol. 42, no. 2, pp. 173–186, 1996.

[40] “Cloud Data Center Evolution - From 0 to 400G | OFC.” [Online]. Available: https://www.ofcconference.org/en-us/home/about/ofc-blog/2017/november/cloud-data-center-evolution-from-0-to-400g/. [Accessed: 08-May-2018].

Chapter 2
Intra-datacenter Transceiver Designs

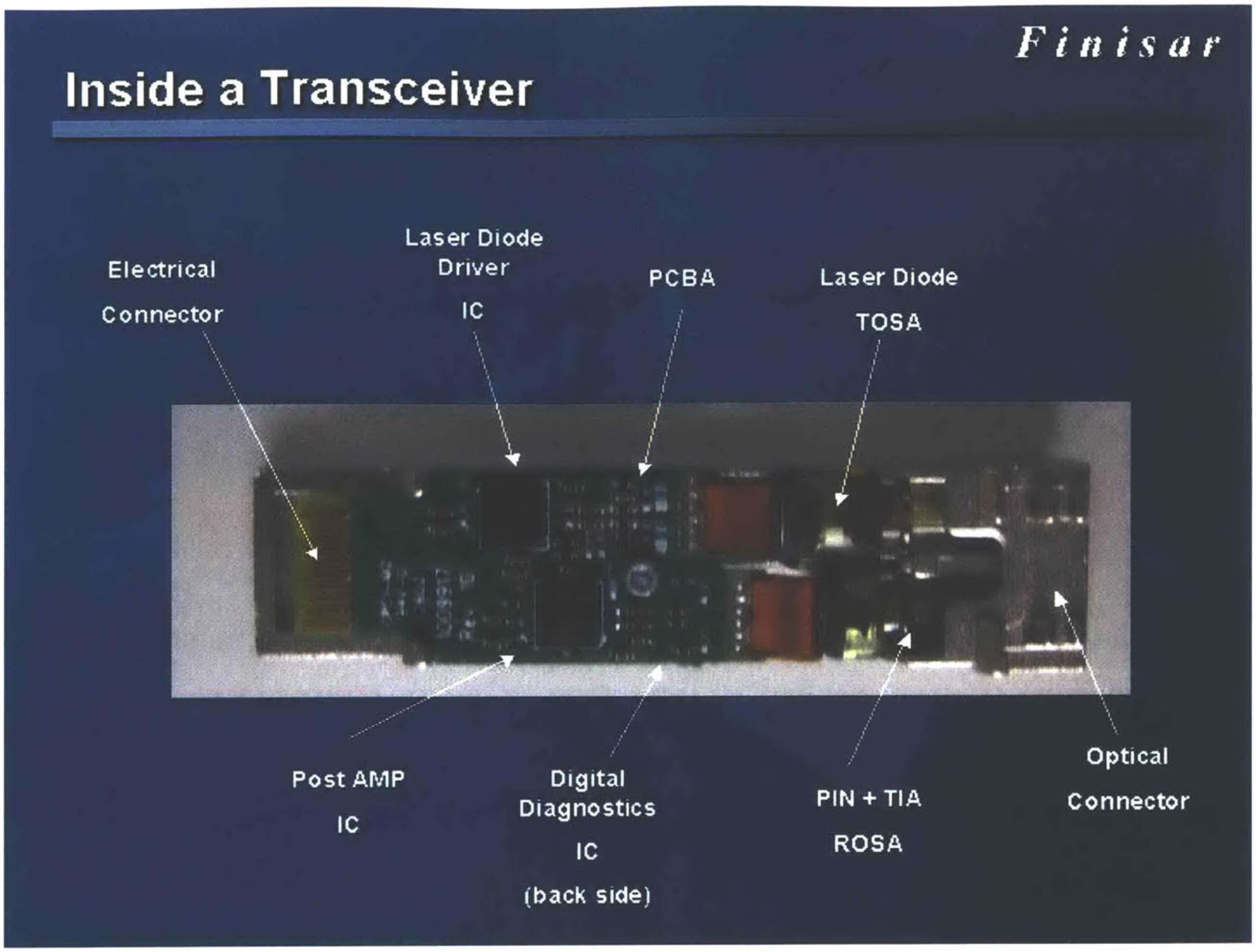
Optical Transceivers

An optical transceiver is a device that sends and receives data using optical fiber rather electrical wire. The device is usually composed of two parts: a transmitter that converts electrical signals into optical signals before transmitting data and a receiver that converts optical signals back to electrical signals after receiving data. The transmitter requires a modulated light source, an electrical IC block that drives the light source or the modulator. The receiver requires a photodetector, an electrical IC block that turns current signals to voltage signals (e.g. transimpedance amplifiers) and an electrical IC block that converts analog signals into digital signals (post amplifiers). Besides, for high speed signals (e.g. 25G per channel), there are also clock and data recovery blocks to extract timing information and serializer/deserializer blocks to compensate for limited input/output. For even higher speeds (e.g. 100G per channel), digital signal processing blocks for signal modulation (e.g. PAM4) are needed. Those are the main functional blocks. To make them work together as a device, there are additional electrical interconnect components (e.g. printed circuit board and bonding wires) and optical alignment components (micro-lens, optical benches, and mechanical housings).

Extrinsically, there are multiple ways to categorize optical transceivers, such as applications, protocols, number of fibers, data rate, distance, size etc. Intrinsically, optical transceivers come with different kinds of designs. I pick out six representative designs of intra-data center optical transceivers and describe them in the following sections.

Transceiver design 1: optical sub-assembly optical transceivers

The optical sub-assembly optical transceivers are the most traditional designs of optical transceivers. The optical and mechanical components are mainly separate parts. Those nonintegrated parts are relatively large in size but have good and reliable technical performance.



Finisar

Inside a Transceiver

Figure 8 A optical sub-assembly optical transceiver from Finisar [2]

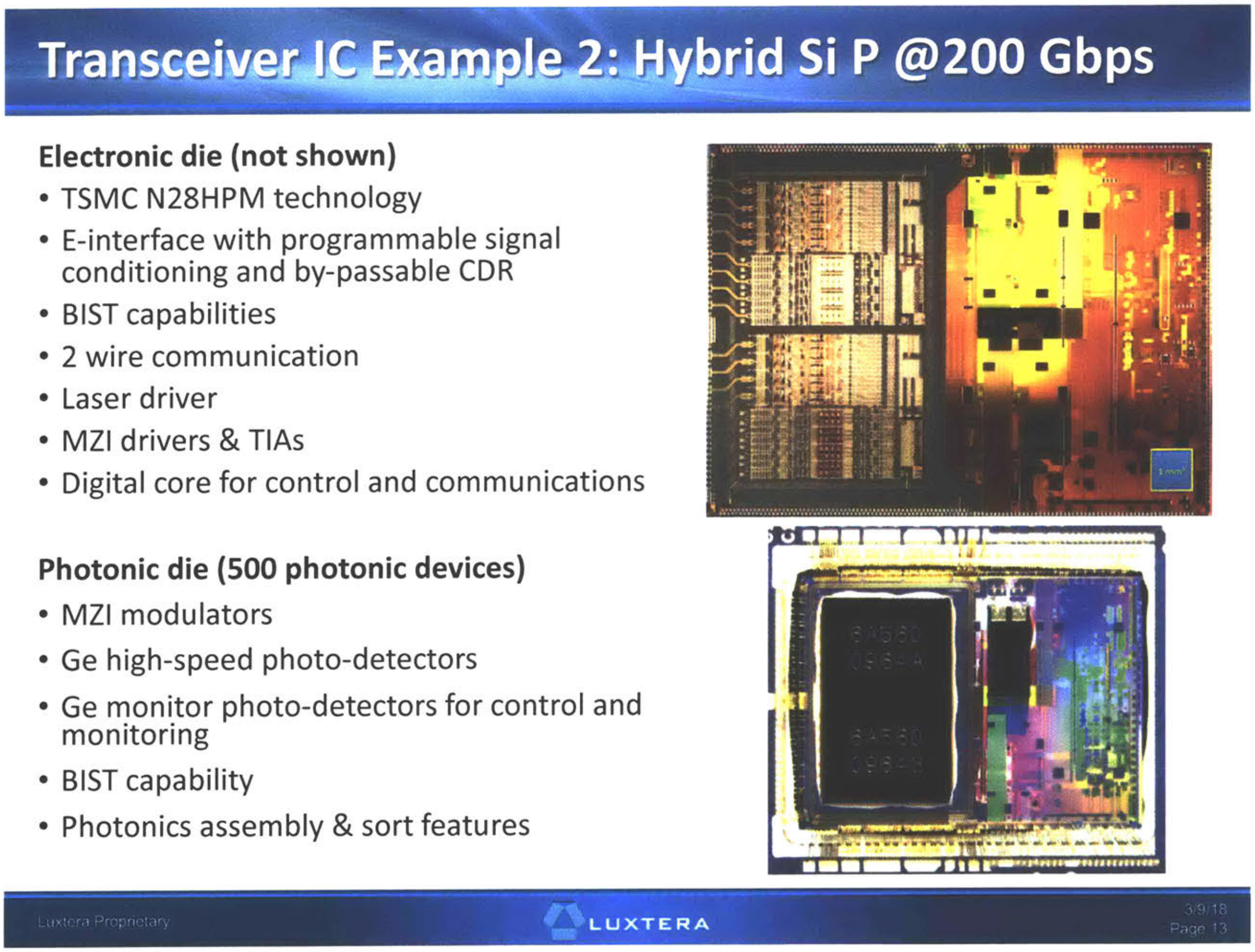
Transceiver design 2: hybrid silicon photonics transceivers

Silicon photonics is a newly emerged disruptive technology in the field of optical transceivers. There are one or several electronic dies, a photonic die and a micro-optics bench in a hybrid silicon photonics transceiver. Functionally, the electronic dies here are similar to the electronic dies in the transceiver design one. The photonic die includes modulators, photodetectors and waveguides except the optical light source. The micro-optics bench sub-assembly contains the light source, a laser chip.

A hybrid silicon photonics transceiver replaces part of the discrete optical components in optical sub-assembly optical transceivers with a silicon photonic die.

There are both upsides and downsides from that integration. The upsides include reducing the number of discrete components, shrinking the size, and shortening the interconnection distance. In the meantime, the partial on-chip integration brings difficulty for optical alignment between the chip and non-integrated optical parts because 1) the mismatch between the scale of a chip waveguide (~1 micron) and the scale of discrete parts such as micro-optics bench and optical fibers (~100 microns); 2) the planar integration on a chip breaks the axial symmetry and thus requires controlling six degrees of freedoms in the alignment. The alignment is so challenging that there is an additional ~6dB optical loss when light is coupled in and out of the photonic chip.

In addition, the integration is a partial integration because the light source is still off the chip and depends on micro-optics assembly. The size of the light source remains relatively large and thus currently only one light source can be hosted on top of the photonic chip. That gives difficultly to realize the traditional wavelength multiplexing on such a platform and thus the spatial multiplexing MSA (Multi-Source Agreement) PSM4 (Parallel Single Mode fiber 4-lane) is developed, in which there are four fiber lanes, one for each channel and all carrying the same wavelength signal. Besides, the laser output power needs to four times that the required input power for each lane. The integration upside wins over those downsides for coherent optical transceivers which involves complex optical designs but may not compensate the downsides for noncoherent single-channel optical transceivers whose optical designs are simple. [3]



Transceiver IC Example 2: Hybrid Si P @200 Gbps

Electronic die (not shown)

- TSMC N28HPM technology
- E-interface with programmable signal conditioning and by-passable CDR
- BIST capabilities
- 2 wire communication
- Laser driver
- MZI drivers & TIAs
- Digital core for control and communications

Photonic die (500 photonic devices)

- MZI modulators
- Ge high-speed photo-detectors
- Ge monitor photo-detectors for control and monitoring
- BIST capability
- Photonics assembly & sort features

Luxtera Proprietary

LUXTERA

3/9/18
Page 13

Figure 9 A hybrid silicon photonics transceiver from Luxtera

Transceiver design 3: hybrid silicon photonics transceivers with chip-bonded lasers

The design also applies a silicon photonic chip to replace part of the discrete optical parts. The difference is that the light source is a laser chip bonded on the silicon photonic chip instead of a micro-optics bench packaged laser chip. It further removes the need of optical lens and mechanical cages for aligning the laser to the silicon photonic chip and thus shrinks the size of the light source. Different from the design 2, in which only one light source can be hosted on the silicon photonic chip, design 3 is able to host multiple light sources on one silicon photonic chip. A design 3 100G transceiver can be either wavelength multiplexing or spatial multiplexing.

Transceiver design 4: single-chip silicon photonics transceivers with off-chip lasers

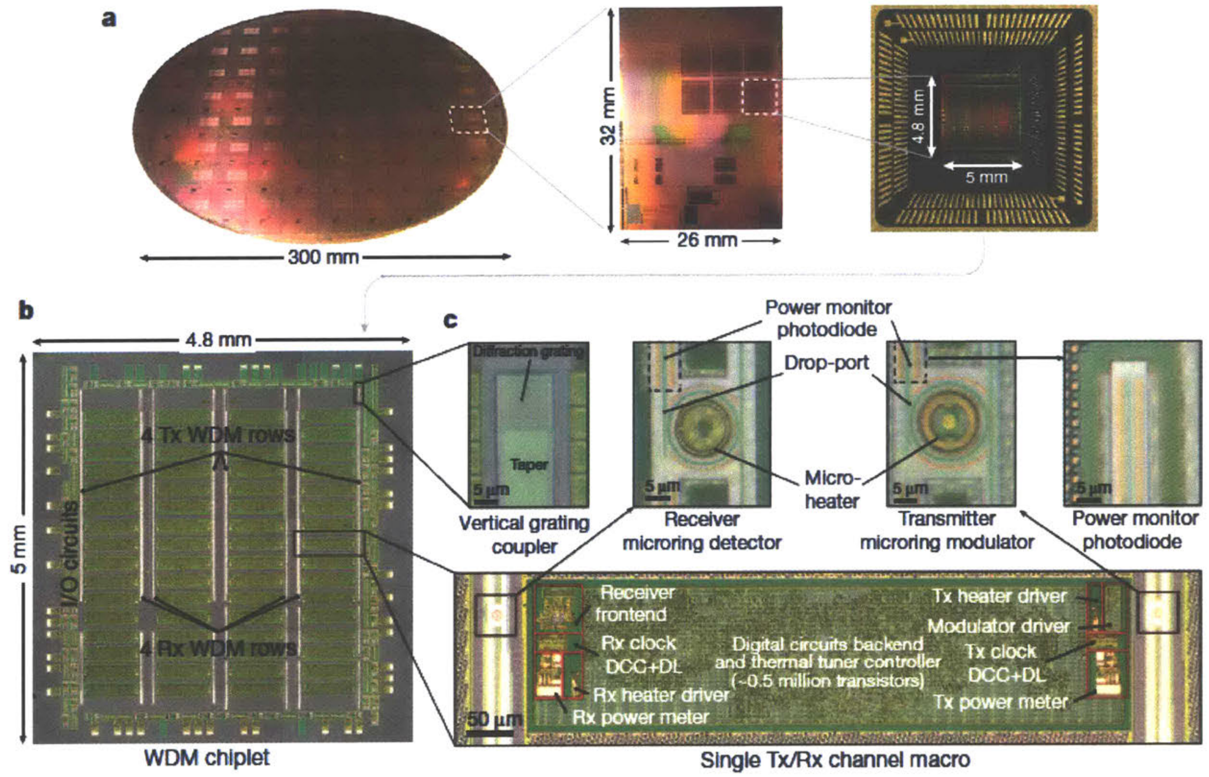


Figure 10 A single-chip silicon photonics transceiver [4], [5]

The design integrates both the photonic part and the electronic part onto the same chip except that it leaves the laser light source off the chip. On the chip, the light is traveling in the waveguides. Off the chip, the light is traveling in single-mode fibers. The light is coupled into and out of the chip through vertical grating couplers. The electronic part here includes not only the functional electronic blocks that processes the electro-optical and optical-electrical conversion but also the logical computing units. The design moves the photonic components close to the logic units and offers one tenth of the latency of those hybrid designs. The same structure has been demonstrated to realize chip-to-chip optical communications. However, currently the short latency feature is not yet highly valued by intra-data center network interconnects.

Transceiver design 5: on-board optics

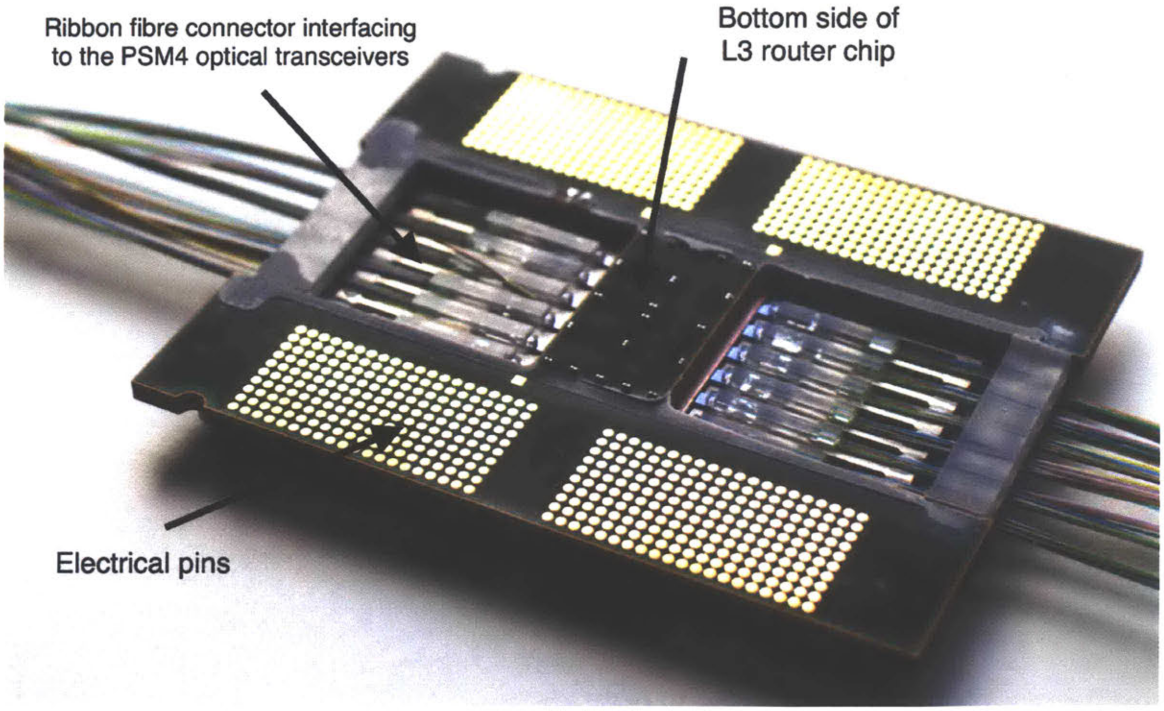

Ribbon fibre connector interfacing to the PSM4 optical transceivers
Bottom side of L3 router chip
Electrical pins

Figure 11 An in-packaged optics design from Rockley Photonics [6]

The on-board optics design targets to resolve the bandwidth density bottleneck on the network switch faceplate. Instead of innovating individual optical transceivers, the design mainly focuses on the system design that it moves optical transceivers onto the same board with the network switching chip.

Transceiver design 6: Indium-phosphide integrated optical transceivers

The design utilizes the indium-phosphide substrate instead of silicon for integrating both the light source and the other photonic functions onto the same chip. The electronic blocks are separate silicon chips. The main challenge is that the manufacturing yield on the Indium-phosphide platform is low, which limits the level of integration on a single Indium-phosphide chip.

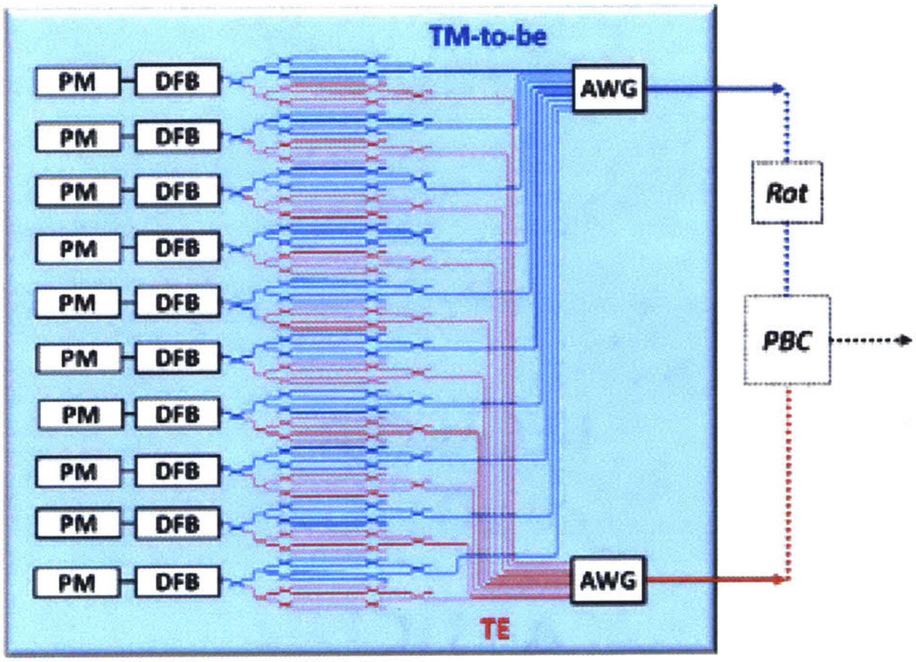


Figure 12 A 500G PIC (photonics integrated circuit chip) from Infinera [7]

[1] L. A. Barroso, J. Clidaras, and U. Hoelzle, *The Datacenter as a Computer:An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool, 2013.

[2] "defa14a." [Online]. Available: https://www.sec.gov/Archives/edgar/data/1094739/000095013404013944/f01954a1defa14a.htm. [Accessed: 08-May-2018].

[3] C. Doerr *et al.*, "Single-chip silicon photonics 100-Gb/s coherent transceiver," in *OFC 2014*, 2014, pp. 1–3.

[4] C. Sun *et al.*, "Single-chip microprocessor that communicates directly using light," *Nature*, vol. 528, no. 7583, pp. 534–538, Dec. 2015.

[5] A. H. Atabaki *et al.*, "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature*, vol. 556, no. 7701, pp. 349–354, Apr. 2018.

[6] "Gazettabyte - Home - Rockley Photonics showcases its in-packaged design at OFC." [Online]. Available: http://www.gazettabyte.com/home/2018/3/15/rockley-photonics-showcases-its-in-packaged-design-at-ofc.html. [Accessed: 08-May-2018].

[7] F. Kish, "500Gb/s and Beyond PIC-Module Transmitters and Receivers," in *Optical Fiber Communication Conference (2014), paper W3I.1*, 2014, p. W3I.1.

[8] T. Li, "Optical Fiber Communication-The State of the Art," *IEEE Trans. Commun.*, vol. 26, no. 7, pp. 946–955, Jul. 1978.

[9] D. T. Neilson, D. Stiliadis, and P. Bernasconi, "Ultra-high capacity optical IP routers for the networks of tomorrow: IRIS Project," in *2005 31st European*

Conference on Optical Communication, ECOC 2005, 2005, vol. 5, pp. 45–48 vol.5.

[10] E. R. H. Fuchs, R. E. Kirchain, and S. Liu, "The Future of Silicon Photonics: Not So Fast? Insights From 100G Ethernet LAN Transceivers," *J. Light. Technol.*, vol. 29, no. 15, pp. 2319–2326, Aug. 2011.

[11] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, "The Emerging Optical Data Center," in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2011 (2011), paper OTuH2*, 2011, p. OTuH2.

[12] MIT Microphotonics Center, "On-Board Optical Interconnection." Apr-2013.

[13] D. Mahgerefteh and C. Thompson, "Techno-economic Comparison of Silicon Photonics and Multimode VCSELs," in *Optical Fiber Communication Conference (2015), paper M3B.2*, 2015, p. M3B.2.

[14] A. Chakravarty, K. Schmidtke, S. Giridharan, J. Huang, and V. Zeng, "100G CWDM4 SMF optical interconnects for facebook data centers," in *2016 Conference on Lasers and Electro-Optics (CLEO)*, 2016, pp. 1–2.

[15] D. Zhuo *et al.*, "{RAIL}: A Case for Redundant Arrays of Inexpensive Links in Data Center Networks," 2017. .

[16] X. Zhou, H. Liu, and R. Urata, “Datacenter optics: requirements, technologies, and trends (Invited Paper),” *Chin. Opt. Lett.*, vol. 15, no. 5, p. 120008, May 2017.

[17] “Cloud Data Center Evolution - From 0 to 400G | OFC.” [Online]. Available: https://www.ofcconference.org/en-us/home/about/ofc-blog/2017/november/cloud-data-center-evolution-from-0-to-400g/. [Accessed: 08-May-2018].

Chapter 3
Cost and Time for Product Development

Product development

Product development is the set of activities beginning with the perception of a market opportunity and ending in the production, sale, and delivery of a product. The general product development process is usually composed of six phases: planning, concept development, system-level design, detail design, testing and refinement and production ramp-up. [1] For a hardware product, although the exact process may vary from company to company and from product to product, the following stream of activities is usually shared:

1) brainstorm & proof-of-concept prototype,
2) engineering validation test (EVT),
3) design validation test (DVT),
4) production validation test (PVT),
5) mass production (MP).

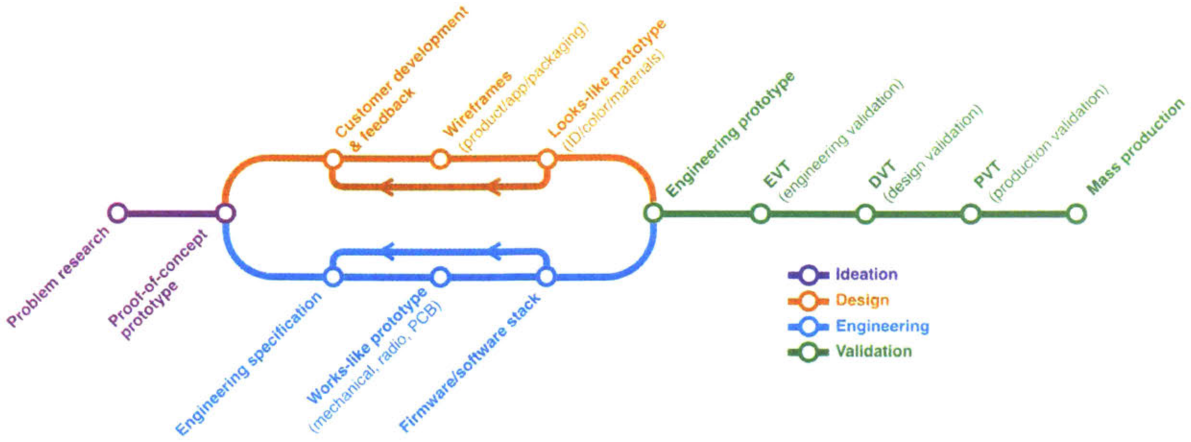

Problem research
Proof-of-concept prototype
Customer development & feedback
Wireframes (product/app/packaging)
Looks-like prototype (ID/color/materials)
Engineering specification
Works-like prototype (mechanical, radio, PCB)
Firmware/software stack
Engineering prototype
EVT (engineering validation)
DVT (design validation)
PVT (production validation)
Mass production
Ideation
Design
Engineering
Validation

Figure 13 Hardware product development process [2]

During the brainstorm and proof-of-concept prototype phase, ideas are generated and the promising ones are selected for further engineering and design. During EVT, 20~50 units with the major targeting functions, production intent materials and manufacturing processes are built and tested. The product development team can show those EVT builds to potential users and learn new requirements and feedbacks from users. For example, the transceiver samples and demos shown at OFC conference are sometimes such EVT builds. During DVT, a production environment should be in place and build 50~200 units for battery stress and regulatory testing. During PVT, 500 to thousands of units that can be sold to customers are built to finalize the tools and manufacturing processes. Right after PVT, the first full production run can be carried out to meet the minimum order quantity of most customers. During mass production, continuous efforts are made on improving the yield, qualifying new vendors and optimizing the supply chain. [3][4]

Duration and cost of product development

The duration of product development ranges from 1 year to 10 years. In terms of the production itself, the product development time depends on the product complexity, the sales lifetime and the product update cycle. In addition, it depends on the development team size and the development investment.

The cost of product development is often called as non-recurring engineering (NRE) cost. There are mainly two parts of spending. The first part is the spending on the engineering efforts, which are roughly proportional to the number of people on the project team and to the duration of the project. The second part is the investment in the tooling and equipment required for production, which is commonly taken as a fixed cost item in the production cost for in-house manufacturing. Usually, the development cost and the production investment are roughly the same.

	Ice Cream Scoop	Avalanche Probe	HP Laserjet Printer	Tesla Model S	Boeing 787
Annual production volume	10,000 units/year	1000 units/year	4 million units/year	50,000 units/year	120 units/year

Sales lifetime	10 years	3 years	2 years	5 years	40 years
Sales price	$40	$2,250	$130	$80,000	$250 million
Development time	1 year	2 years	1.5 years	3.5 years	7 years
Peak team size	6 people	18 people	175 people	2000 people	17,000 people
Development cost	$100,000	$1 million	$50 million	$500 million	$15 billion
Production investment	$20,000	$250,000	$25 million	$500 million	$15 billion

Table 3 Attributes of five products and their associated development efforts. [1], [5]

If we decompose the duration and cost of product development into different stages. The figure below presents the cumulative cash inflow and outflow over the life cycle of a typical successful product.

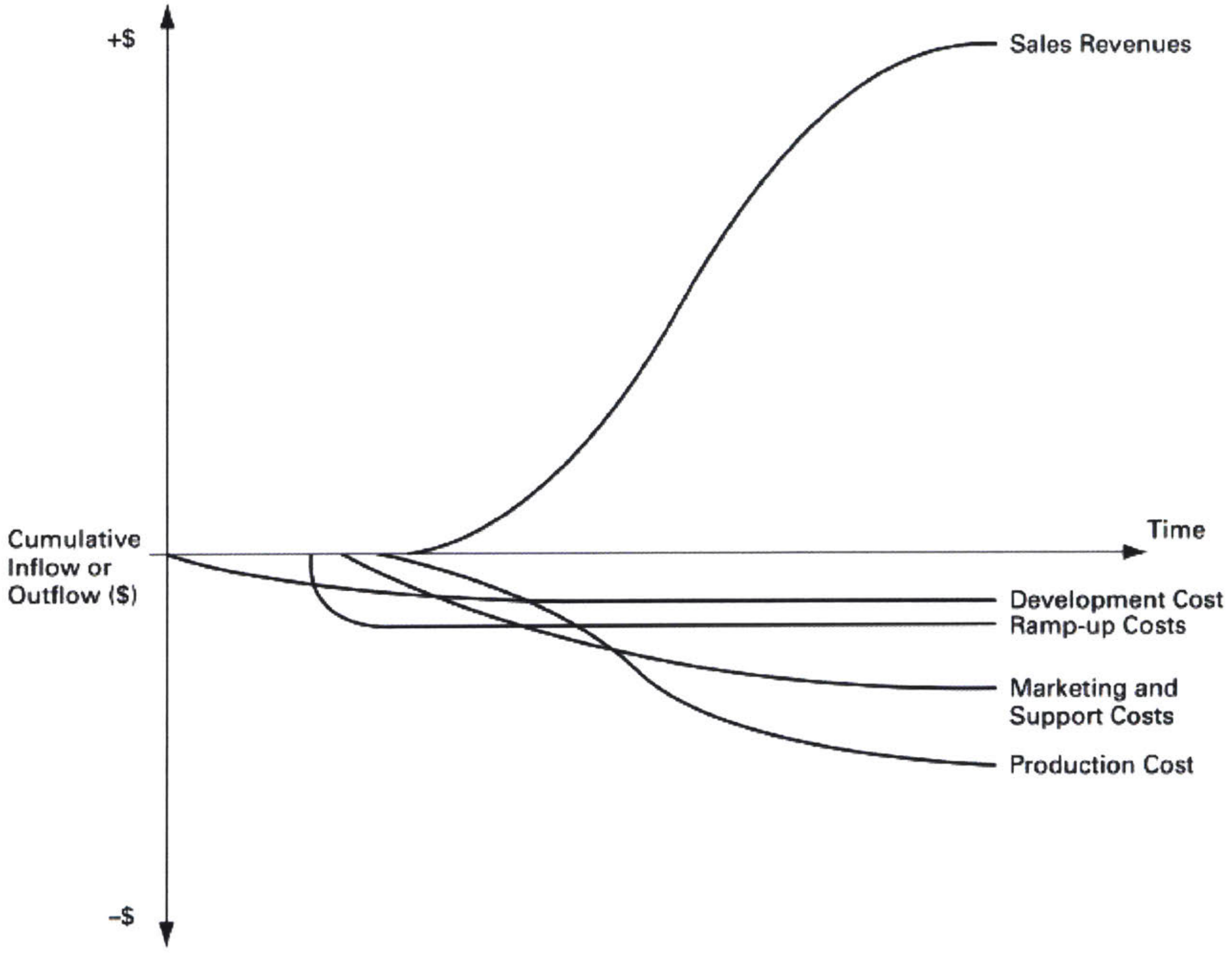


Figure 14 Typical cash flows for a new product [1]

Bottlenecks of optical transceiver product development

At the downstream of optical transceiver products, due to network system design uncertainties, it is very challenging for optical transceiver manufacturers to focus on one design in the early stage of product development. Optical transceivers are components of intra-datacenter networks. Its specification depends on the intra-datacenter network system specifications. Optical transceiver manufacturers need to wait for IEEE 802.3 standards to know the desired technical requirements for the optical transceiver from the network system perspective. However, the standard-making process involves a lot of negotiation from different parties and takes long. For example, the IEEE P802.3ba task force for 40G and 100G ethernet started in December 2007 and completed in June 2010. In addition, many important aspects including form factors are left unspecified. Several multi-source agreements (MSA), e.g. CFP, PSM4, CWDM, QSFP-DD et al., are needed to shorten the waiting of IEEE standards and fill in the gaps of technical specifications. That brings a new problem: there are many variations of standards for one application. Last but not the least, customers may alter the requirements for their own economic benefits. For example, for 100G intra-datacenter optical transceivers, Facebook has relaxed several specifications including reducing the operating case temperature to 15-55°C, reducing the distance requirements from 2km to 500m, and eased specifications on product lifetime. [6] Due to network system design uncertainties, it is very challenging for optical transceiver manufacturers to focus on one configuration in the early stage of optical transceiver product development. A common practice is that optical transceivers deliver EVT samples to customers to avoid spending on PVT and ramp-up for unsuccessful products. However, that practice makes customers to wait for PVT and production ramp-up in addition to normal order lead time.

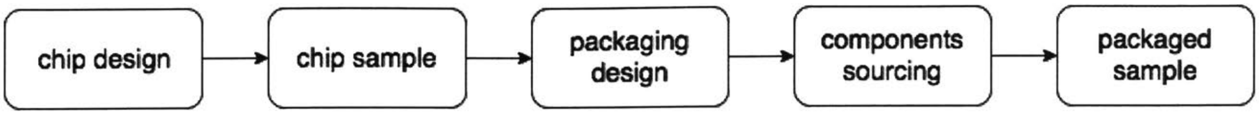
At the upstream of optical transceiver products, many optical transceiver components are made-to-order goods that cannot be purchased off-the-shelf. Those components need to be sourced from vendors or ordered from contract manufacturers. Sometimes they even require significant amount of engineering and design efforts from upstream vendors. For example, many optical transceiver companies rely on other companies' supply of lasers and those lasers are made-to-order products. That creates three challenges in the product development of optical transceivers: 1. the optical transceiver design is constrained by the availability of desired components; 2. both the product development and the mass production is commonly delayed by the supply of components.

Component	Order lead time
lasers (made-to-order)	6 ~ 12 months
silicon optical benches (made-to-order)	~ 3 months
photonic chips	6 ~ 18 months

(made-to-order)	
electrical chips (off-the-shelf)	2 ~ 3 weeks
electrical chips (made-to-order)	6 months
optical lenses (made-to-order)	~ 1 month
lensed fibers (made-to-order)	~ 1 month
printed circuit board	~1 month

Table 4 Estimated order lead time for optical transceiver components

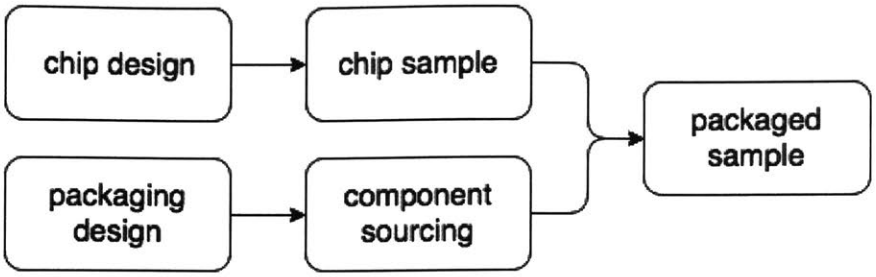
There are also bottlenecks in the ordering of the process flow. Optical transceivers are usually composed of several functional blocks and several components, which require electrical, optical and mechanical engineering and designs. For optical sub-assembly transceiver products, the product development is decomposed into parallel stages of work on different subsystems and component. For integrated design products, because the integrated chip design is so novel that the full product development won't start until the success of the chip, the product development is a sequential engineering process, which starts with chip design and test, followed by packaging design.



Because of the lack of consideration of component integration in the early stage, the sequential engineering method caused a few problems in the following packaging steps which could have been avoided with little cost in the chip design. For example,

- there was not enough space reserved on chip for optical alignment fixture;
- the optical input and output directions didn't match;
- the spacing between optical channels were too dense.

Now the situation has improved a lot. Packaging research groups are becoming involved in the chip design stage. Also, chip foundries and packaging service providers are designing packaging standards and advocating the adoption of such standards. The concurrent engineering of chip and packaging becomes more and more common.



R&D time estimation methodologies

In fast cycle industries, in which product life cycles are often three years or less, rapid product development is widely viewed as a key source of competitive edge. Datar et al. concluded that lead-time advantage affects market share positively. [1] Cohen et al. studied the trade-off between the performance and time-to-market in new product development. [2] Research efforts have also been made on developing models to estimate the product development time. Bashir et al. developed a parametric model to estimate the design hours (E) based on product complexity (PC) and severity of requirements (SR), in which

$$E = aPC^{b}SR^{c}.$$

The a, b and c are company-specific coefficients. Johnson et al. divided the product development into four stages: detailed design, formability engineering, fabrication engineering and assembly engineering. [3] For each stage, the development lead time (ST) is estimated by required design efforts in man-hours for component for a given component x (TDH_x), the total number of engineers available (TEA), the maximum development effort required for one component (MEPC) and work hours per week.

$$ST\ in\ weeks = \frac{\max\left(\frac{\Sigma_x TDH_x}{TEA}, \max(TDH_x)\ /MEPC\right)}{work\ hours\ per\ week}.$$

R&D cost estimation methodologies

Johnson et al. compared the development cost of two car component designs, a tube-based steel design and a die-cast magnesium design, by calculating the labor and IT cost of design efforts in man hours and the manufacturing cost of prototypes[4] in different stages of product development.

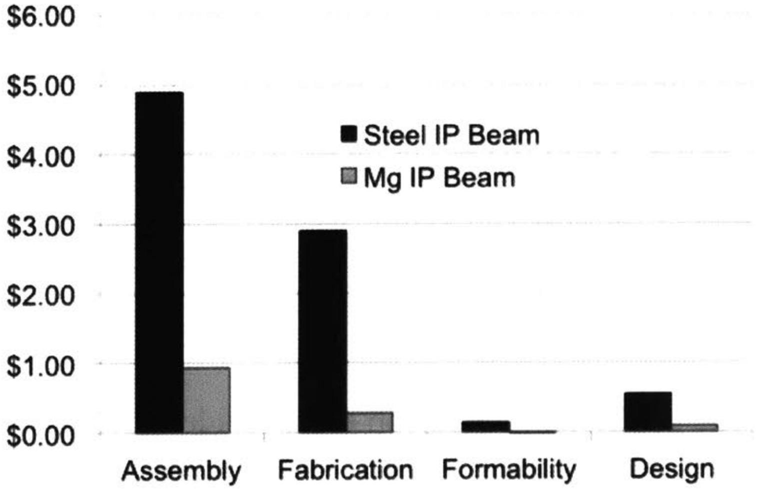


Fig. 6. Unit development costs for alternative IP beam designs at 75,000 units per year.

$$C_{R\&D} = \left(W_{engineer} + P_{software} + P_{computer}\right) * T_{Design} + C_{prototype}$$

Browning et al. estimated the new product development cost of five different product designs of uninhabited aerial vehicle through design structure matrix[5]. The study assumes different sequences of product development activities for different product architectures. The design structure matrix is used to map the correlation between different activities and estimate the probability and intensity of rework a previous step.

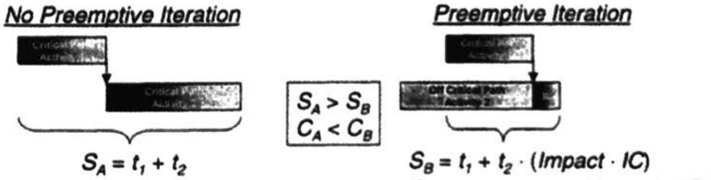


Figure 11: Effect of Preemptive Iteration on Cost and Schedule

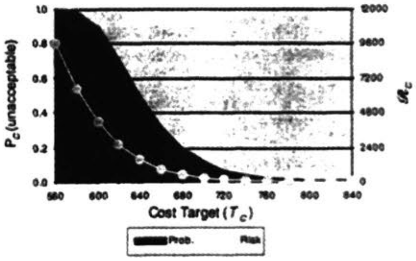
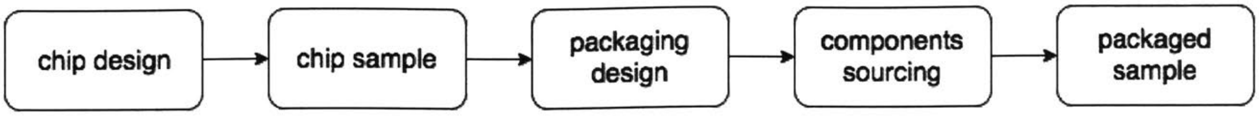


Figure 12: P_C(unacceptable) and $\mathcal{R}_C$ for Various Cost Targets (T_C)

Status of Optical Transceiver R&D

For optical transceivers, the product development process varies. In a few cases that Tyndall National Institute was facing, the process started with chip design, followed by packaging design.



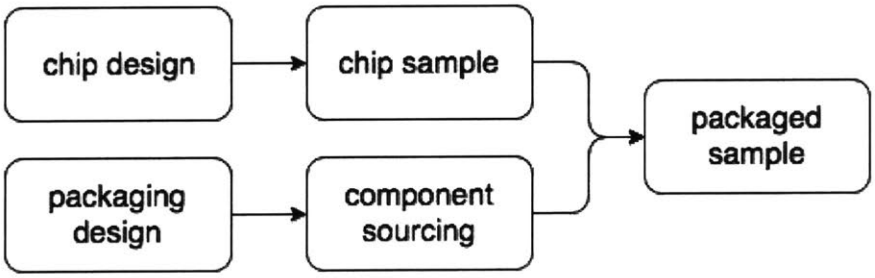
Due to the lack of consideration of integration in the early stage, the sequential engineering method caused a few problems in the following packaging steps which could have been avoided with little cost in the chip design. For example,

there was not enough space reserved on chip for optical alignment fixture;

the optical input and output directions didn't match;

the spacing between optical channels were too dense.

Now the situation has improved a lot. Packaging research groups are becoming involved in the chip design stage. Also, chip foundries and packaging service providers are designing packaging standards and advocating the adoption of such standards. The concurrent engineering of chip and packaging becomes more and more common.



Still, even in the concurrent engineering scenario, there are two known time bottlenecks:

waiting for MPW run,

waiting for customized components for packaging.

$$T_{R\&D} = \max\left(T_{chipdesign} + T_{chipmake}, T_{packagingdesign} + T_{componentsourcing}\right) + T_{packagingmake}$$

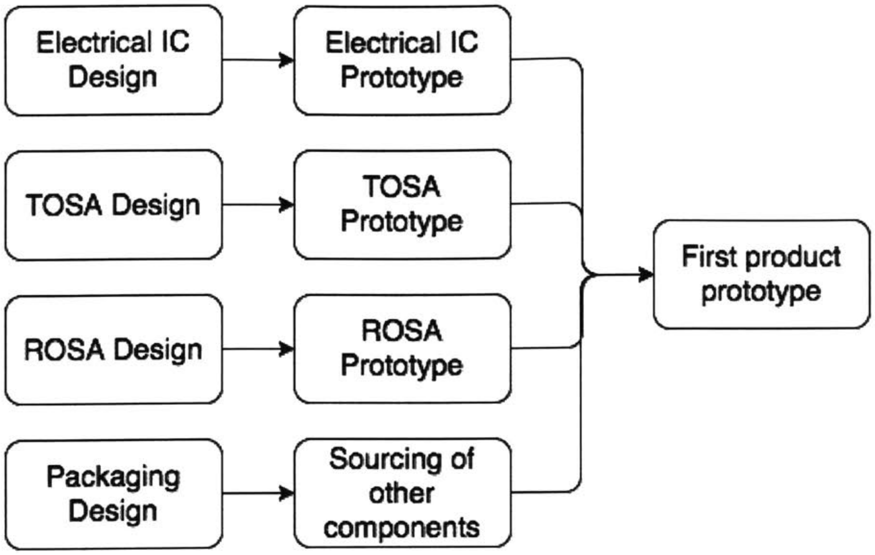
$$T_{chipmake} = T_{waitforMPWrunstart} + T_{MPWrun}$$

(Time from design to prototype product when there is no re-engineering.)

Standardization is not only an enabler of concurrent engineering of chip design and packaging design but also reduces the need of customized components and thus shorten the waiting time.

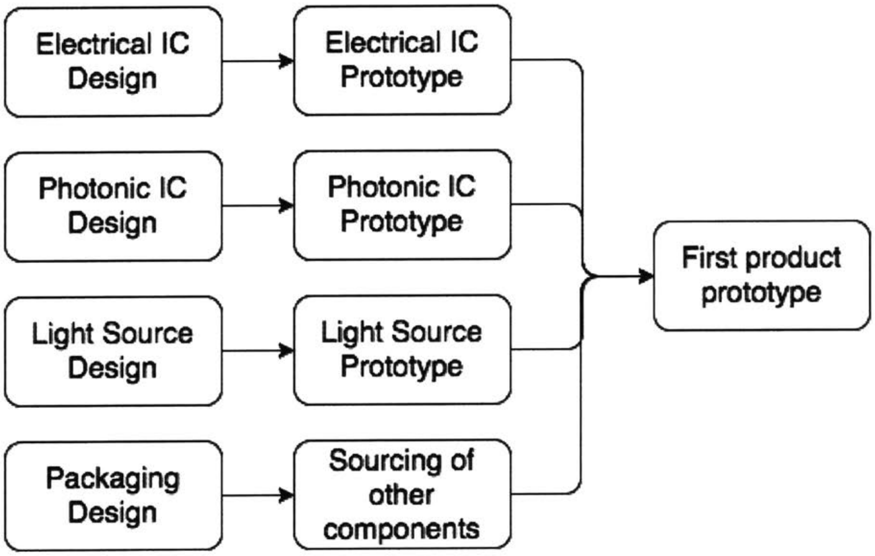
Comparison among different transceiver designs

Transceiver Design 1 - Optical sub-assembly optical transceivers



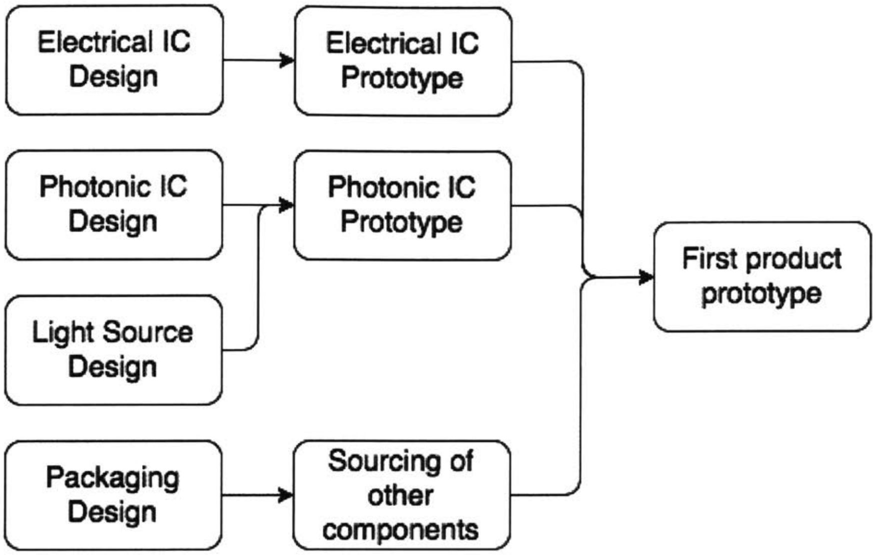
Rate limiting factor: TOSA design or Packaging design

Transceiver Design 2 - Silicon photonics transceivers with a micro-optics bench



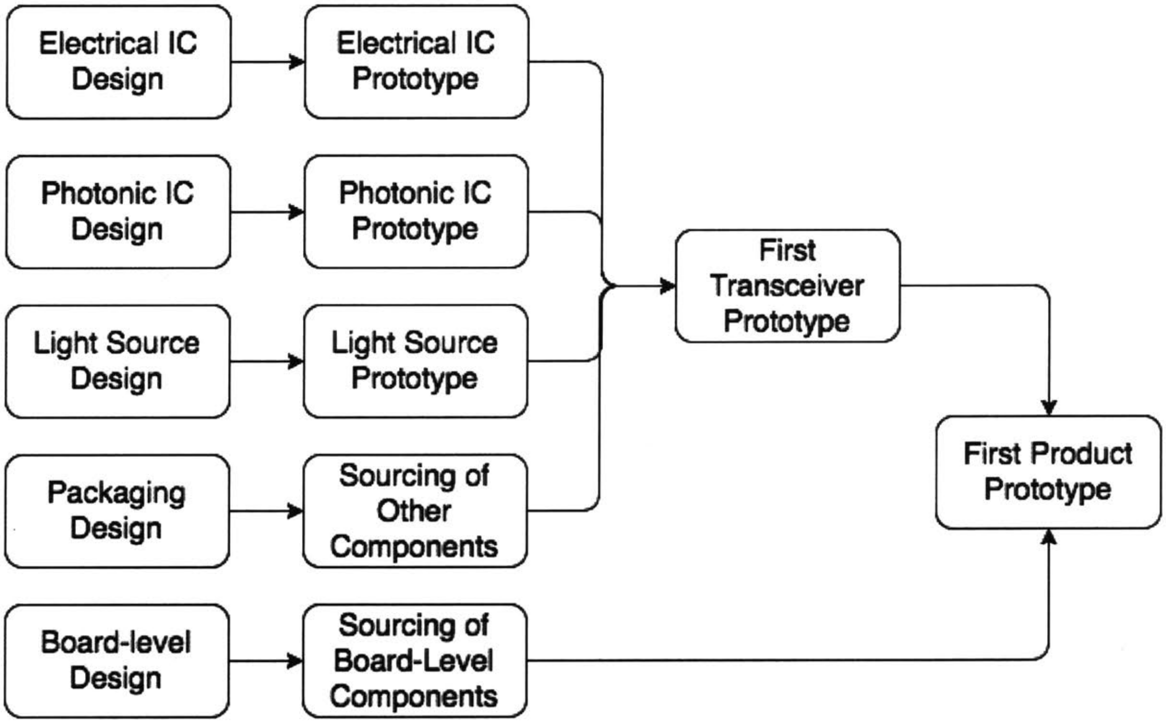
Rate limiting factor: photonic IC prototype or sourcing of other components

Transceiver Design 3 - Silicon photonics transceivers with hybrid silicon lasers

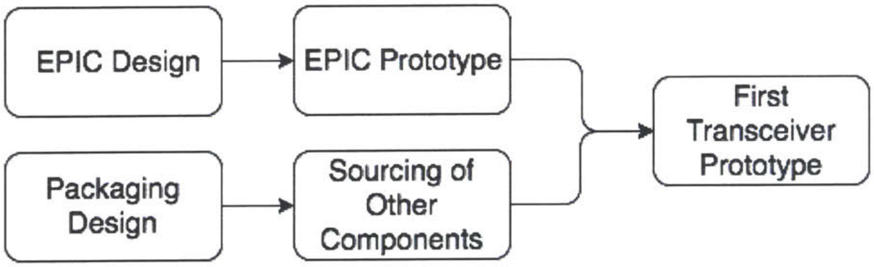


Rate limiting factor: photonic IC prototype/sourcing of other components

Transceiver Design 4 - On-board Optics

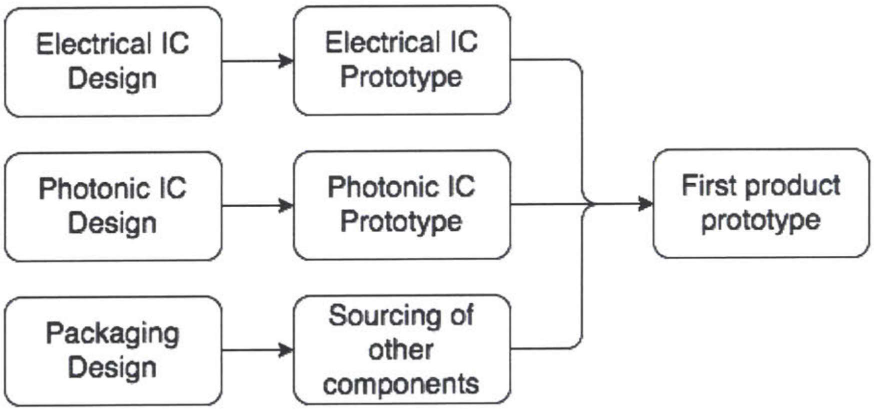


Transceiver Design 5 - Silicon photonics transceivers with off-chip lasers



Rate limiting factor: EPIC design

Transceiver Design 6 - Indium-phosphide integrated optical transceivers



Rate limiting factor: photonic IC prototype

R&D	100G	200G/400G	800G	1.6T
D1	1	1	2	2
D2	2	1	1	1
D3	2	1	1	1
D4	3	2	1	1

D5	3	2	2	1
D6	1	1	1	1

Ranking of prototype announcement speed (100G, 200G/400G, news stats. 800G, 1.6T assumption)

[1] K. T. Ulrich and S. D. Eppinger, *Product design and development*. New York, NY : McGraw-Hill Education, [2016], 2016.

[2] B. Einstein, “The Illustrated Guide to Product Development (Part 1: Ideation),” *Bolt Blog*, 20-Oct-2015. [Online]. Available: https://blog.bolt.io/the-illustrated-guide-to-product-development-part-1-ideation-ab797df1dac7. [Accessed: 09-May-2018].

[3] I. Instrumental, “Hardware engineers speak in code: EVT, DVT, PVT decoded,” *Instrumental*, 09-May-2018. [Online]. Available: https://www.instrumental.com/blog/2016/11/14/hardware-engineers-speak-in-code-evt-dvt-pvt-decoded. [Accessed: 10-May-2018].

[4] B. Einstein, “The Illustrated Guide to Product Development (Part 4: Validation),” *Bolt Blog*, 20-Oct-2015. [Online]. Available: https://blog.bolt.io/the-illustrated-guide-to-product-development-part-4-validation-1b5ab3aeaf35. [Accessed: 10-May-2018].

[5] K. T. Ulrich and S. D. Eppinger, *Product Design and Development, 5th Edition*, 5 edition. New York: McGraw-Hill Education, 2011.

[6] A. Chakravarty, K. Schmidtke, S. Giridharan, J. Huang, and V. Zeng, “100G CWDM4 SMF optical interconnects for facebook data centers,” in *2016 Conference on Lasers and Electro-Optics (CLEO)*, 2016, pp. 1–2.

[7] L. A. Barroso, J. Clidaras, and U. Hoelzle, *The Datacenter as a Computer:An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool, 2013.

[8] “defa14a.” [Online]. Available: https://www.sec.gov/Archives/edgar/data/1094739/000095013404013944/f01954a1defa14a.htm. [Accessed: 08-May-2018].

[9] C. Doerr *et al.*, “Single-chip silicon photonics 100-Gb/s coherent transceiver,” in *OFC 2014*, 2014, pp. 1–3.

[10] C. Sun *et al.*, “Single-chip microprocessor that communicates directly using light,” *Nature*, vol. 528, no. 7583, pp. 534–538, Dec. 2015.

[11] A. H. Atabaki *et al.*, “Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip,” *Nature*, vol. 556, no. 7701, pp. 349–354, Apr. 2018.

[12] “Gazettabyte - Home - Rockley Photonics showcases its in-packaged design at OFC.” [Online]. Available:

http://www.gazettabyte.com/home/2018/3/15/rockley-photonics-showcases-its-in-packaged-design-at-ofc.html. [Accessed: 08-May-2018].

[13] F. Kish, “500Gb/s and Beyond PIC-Module Transmitters and Receivers,” in *Optical Fiber Communication Conference (2014), paper W3I.1*, 2014, p. W3I.1.

[14] T. Li, “Optical Fiber Communication-The State of the Art,” *IEEE Trans. Commun.*, vol. 26, no. 7, pp. 946–955, Jul. 1978.

[15] D. T. Neilson, D. Stiliadis, and P. Bernasconi, “Ultra-high capacity optical IP routers for the networks of tomorrow: IRIS Project,” in *2005 31st European Conference on Optical Communication, ECOC 2005*, 2005, vol. 5, pp. 45–48 vol.5.

[16] E. R. H. Fuchs, R. E. Kirchain, and S. Liu, “The Future of Silicon Photonics: Not So Fast? Insights From 100G Ethernet LAN Transceivers,” *J. Light. Technol.*, vol. 29, no. 15, pp. 2319–2326, Aug. 2011.

[17] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, “The Emerging Optical Data Center,” in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2011 (2011), paper OTuH2*, 2011, p. OTuH2.

[18] MIT Microphotonics Center, “On-Board Optical Interconnection.” Apr-2013.

[19] D. Mahgerefteh and C. Thompson, “Techno-economic Comparison of Silicon Photonics and Multimode VCSELs,” in *Optical Fiber Communication Conference (2015), paper M3B.2*, 2015, p. M3B.2.

[20] D. Zhuo *et al.*, “{RAIL}: A Case for Redundant Arrays of Inexpensive Links in Data Center Networks,” 2017. .

[21] X. Zhou, H. Liu, and R. Urata, “Datacenter optics: requirements, technologies, and trends (Invited Paper),” *Chin. Opt. Lett.*, vol. 15, no. 5, p. 120008, May 2017.

[22] “Cloud Data Center Evolution - From 0 to 400G | OFC.” [Online]. Available: https://www.ofcconference.org/en-us/home/about/ofc-blog/2017/november/cloud-data-center-evolution-from-0-to-400g/. [Accessed: 08-May-2018].

Chapter 4
Cost and Time for Mass Production

Prior to the current PSMC effort, there have been relatively few published (i.e. publicly available) cost analysis studies that focused on the manufacturing of integrated photonics.

In the earliest such study, Schuelke and Pande analyzed the manufacturing cost of an optoelectronic integrated circuit chips based on a cost model developed for millimeter and microwave integrated circuits.[6] Specifically, they examined the economics of integrated four core functions (detection, preamplification, amplification & filtering, and decision making) into a single GaAs chip. Their conclusion was that integrating only two of the four functions was economically preferred because of decreases in net yield as the circuit grows in size and complexity. Marz et al. established an analytical model to estimate the relative yields and the relative costs over time of integrated optical chips to a reference chip.[7] Their detailed model would allow decision makers to estimate how costs might be expected to change with integration and therefore optimize current levels of integration. As presented this model depends on the costs of an existing reference chip. As mentioned by the authors, to be complete, the scope of the model would need to be expanded to consider packaging and assembly of the complete module. An acitivity-based cost model is applied by Stirk et al. to calculate the cost of a 2-D

VCSEL array communication module.[8] Although few details of the model are provided, this analysis makes clear that assembly yield is a critical aspect of ultimate module cost.

More recently, the research team at MIT built upon these various earlier studies. Specifically, Kirchain, Fuchs et al. from MIT Materials System Laboratory developed a process-based cost model (PBCM) with data collected from numerous firms across the optoelectronics supply chain covering front end, back end and packaging. This model allows for the user to specify the process flow, individual processing conditions, operational characteristics, and level of automation at each step.

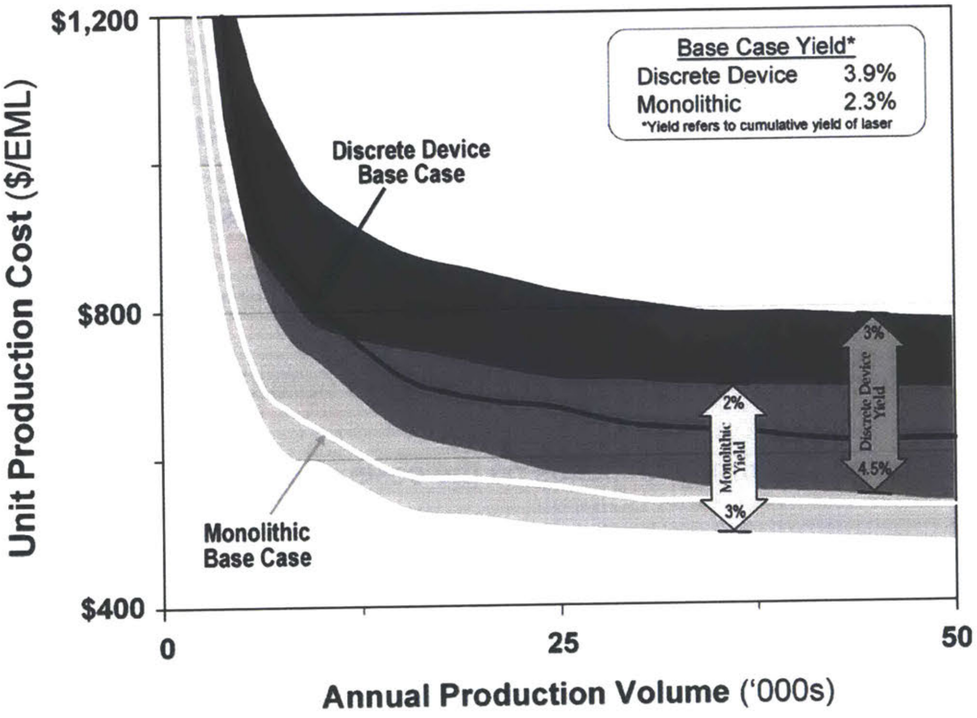


Figure 15. Cost sensitivity of production volume analysis of monolithic and discrete integration of InP-based DFB laser and EA modulator.

With the PBCM tool, the cost of integration of a 1550-nm DFB laser with an electroabsorptive (EA) modulator on an InP platform was analyzed.[9] The results suggest that a monolithically integrated design should be more cost competitive over a discrete component options regardless of production scale (see Figure 15) unless the integrated yield is particularly poor. The research also identified the dominant cost drivers as packaging, testing, and assembly – the focus of current PSMC cost modeling activities. Besides, component alignment, bonding, and metal-organic chemical vapor deposition (MOCVD) are identified as processes where technical improvements were most critical to lowering costs. It is estimated that economies of scale for manufacturing the components occurred between 30,000 and 200,000 units/year, depending on the type and complexity of the device being evaluated, which encourages photonic industry to consolidate its manufacturing sites in order to achieve economies of scale.

In a subsequent study[10], the optoelectronics PBCM model was applied to model transmitter designs with a Discrete Laser and Modulator (Prevailing Design) and an Integrated Laser and Modulator (Emerging Design) and study the impact of offshore manufacturing on optoelectronics. It was found that the economics of offshore production increases the cost advantage of the prevailing technology and therefore reduces incentives for innovation (see Figure 16).

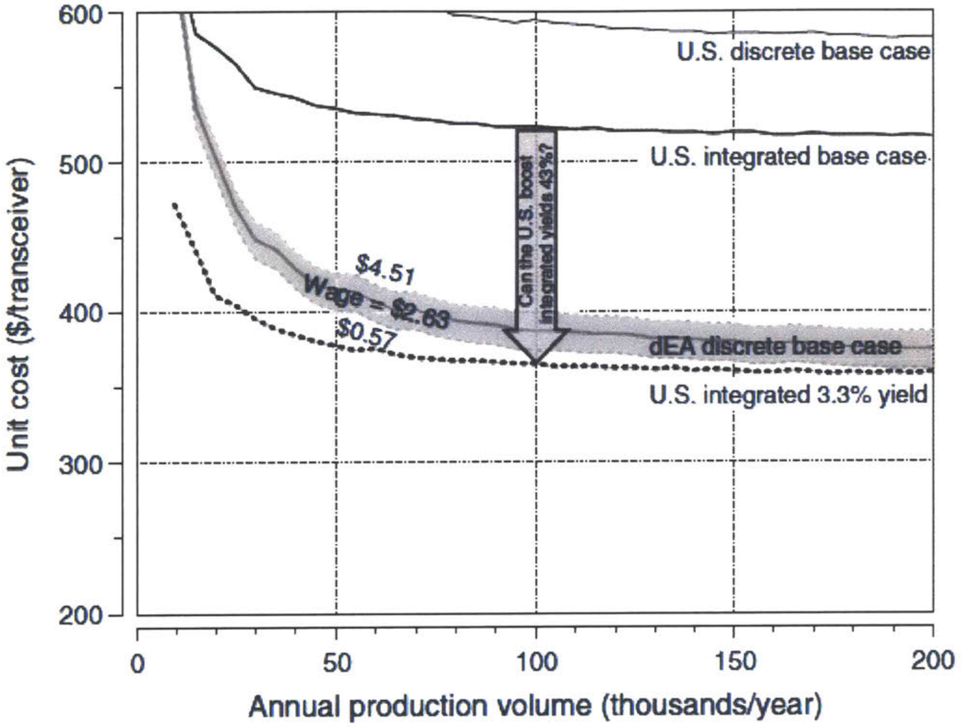


Figure 16. Cost competitiveness of U.S.-produced integrated laser modulator vs. developing east asia (dEA)-produced discrete laser and modulator design.

Finally, in a third study[11], the MIT research team explored the competitiveness of two Si photonic designs against InP-based alternatives for a 1310 nm, 100 gigabit ethernet LAN transceiver. The research suggested great promise for silicon photonics to meet cost targets in midterm server and storage area network applications and to provide even lower cost devices for future, high-volume consumer and mobile computing applications. Specifically, it was found that silicon photonics holds great potential to be cost competitive in markets with annual sales volumes above 900 000, including servers, computing, and mobile devices (see Figure 17). These results should motivate academic research into integration on the Si substrate from the perspective of cost.

These previous studies provide an important foundation on which current cost modeling efforts and tools can be built. Nonetheless, there are some important gaps which the PSMC effort is currently addressing. Previous studies primarily focus on optical transceivers for telecommunication applications. Currently, however, attention focused onto the application of optical transceivers in data centers. More pointedly, the application of transceivers for shorter distance applications inherently face stiffer cost challenges. Hence, for the current roadmap development, the model's target needs to be shifted.

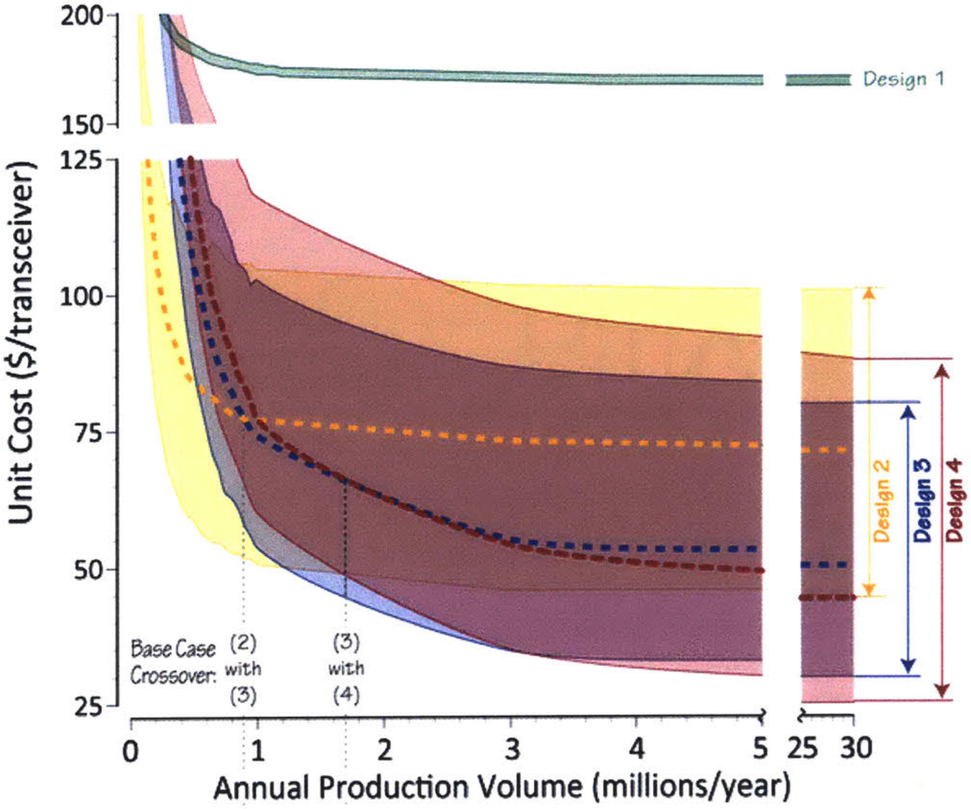


Figure 17. Total cost comparison of InP based designs (1: TOCAN and 2: DML) and Si based designs (3: Hybrid and 4: Silicon Two Chip) with yield and regional sensitivities (top and bottom bands plus base case yield bands).

This is particularly relevant because the packaging processes for data center optical transceivers differ from those used in telecommunication applications because the datacom optical transceivers are based on more compact designs. For instance, as mentioned in the Assembly & Test Chapter, for accurate assembly of small optical parts in datacenter optical transceivers (e.g. single mode optical fibers), alignment of parts to <0.1 micron of accuracy is demanded and there is a trade-off between speed and accuracy for a decision between passive alignment methods and active alignment methods. Furthermore, since these earlier studies were completed, more packaging technologies have emerged, such as high aspect ratio through silicon vias and thermocompression bonding. Last but not the least, the integrated photonics and electronics technologies are evolving quickly and thus the input data needs to be updated. Therefore, to stupport the current photonics roadmap development activity, the cost modeling team has been updating the cost model to target data center-oriented transceiver designs, add in new packaging processes, and up date the input data.

Current Model Overview

The PSMC cost modeling tool is a process-based cost model (PBCM). As such, the tool builds up cost from technical details. These details are used to estimate processing requirements (e.g., time and materials) and thereby resource requirements (e.g., equipment and personnel). From resource requirements it is possible to project future cash flows and, therefore, compute various cost-related metrics.

The model focuses on packaging processes of emerging designs for integrated optical transceivers in data centers. Because of this focus on packaging, costs of components (laser, interposer, etc.) are directly taken in as inputs instead of being calculated starting from raw material. The final cost of an integrated optical transceiver module is determined from the cost incurred for each step of the manufacturing process. The total cost is the sum of step costs and unit cost is total cost divided by production volume. For each step, the cost is projected by first mapping physical parameters of product designs to production process requirements (e.g., material and thickness requirements to their implications for cycle times, downtimes, yields). These relationships are determined using physical models or through statistical methods. The model then maps these production process requirements to the quantity of production resources (e.g., kilograms of material, person hours, and number of machines) required to meet a stated production scale target. After that, the model multiplies these resource requirements by their respective prices to determine the step costs.

During calculating step cost, the cost is further broken down into primary categories: machine cost, tool cost, building cost, material cost, labor cost and energy cost. For machine cost, the model estimates the quantity of machines required to meet production goals based on available operating time each year and the required machine time to produce the product at the targeted production capacity. The number of required machines are multiplied with prices and annualized (using a selected discount rate and conventional financial assumptions) to obtain annual machine costs. The calculation of tool cost is similar to that of machine cost except that the tool quantity is also subject to tool lifetime. In terms of building cost, there are three different types of cleanroom included in the model: Class 100, Class 1000, Class 10000 and the cost depends on the quantity of machines required and cleanroom space required by each machine. The material cost includes the cost of components (e.g. lasers, IC chips etc.), direct materials (materials that are integrated into the product e.g. metal, bonding paste etc.) and indirect or process materials (water, carrying gas etc.). To estimate labor cost, the model considers personnel of different skill levels to accommodate both highly automatized processes (that require labor with less skill) and relatively complex processes (e.g. visual inspection, active optical alignment, etc.) that require specialized skills. For energy cost, the current model considers only the energy used to power machines; energy required to power the cleanroom facility is included in the building cost as part of the building maintenance cost. In future versions, energy costs will be modeled more explicitly such that it is a function of production volume, machine power and clean room facility working hours.

As a successor of the earlier version of the MIT Photonics PBCM, the current model retains flexibility for the user to define process organization (i.e. the order of the process flow), processing conditions, operational characteristics, and level of automation at each step. Meanwhile, an improved model structure considerably reduces the work for definition of new processes and thus new processes can be conveniently added in the model and it is expedient to compare different processes.

Case Analysis

To gain insights into the production economics associated with integrated photonics, we analyze the relative economic competitiveness of three functionally equivalent datacenter optical transceivers. While the economics of the optical transceiver case will be of interest to some readers, we expect the underlying production economics to hold true beyond the specifics of the case and designs chosen.

The three transceiver designs (see Figure 18) selected to are: 1) Hybrid: laser, fiber array and IC chips mounted on active optical interposer; 2) Monolithic with hybrid layer: laser and fiber array mounted on optical and electrical integrated chip; 3) Fully monolithic: fiber array mounted on optical and electrical integrated chip with laser integrated in the chip. Across the three designs, the level of integration increases.

The cost of producing the individual chips is not modeled explicitly. Instead, we assume that chip production costs are proportional to the area of the chip and estimate that area. Also, initially we assume that yields for all chips are the same. Clearly that will not be the case. As such, we explore the sensitivity of the result to changes in yield.

CRITICAL CAVEATS

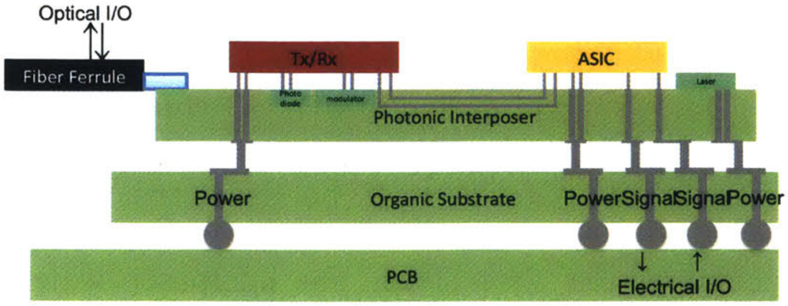
For all three designs, we limit our analysis to only the packaging of the the optical devices into a module. The costs of components are taken as inputs and are not calculating directly. Although we carry out sensitivities to understand the potential implication of various levels of component costs. The reader should view this analysis as incomplete. Although incomplete, this analysis serves to demonstrate the potential for details process-based cost analysis to critical photonics questions.

The following analysis is based on data collected from less than three firms. As such, it is not statistically significant.

Designs Analyzed

The three transceiver designs (see Figure 18) selected to are: 1) Hybrid: laser, fiber array and IC chips mounted on active optical interposer; 2) Monolithic with hybrid laser: laser and fiber array mounted on optical and electrical integrated chip; 3) Fully monolithic: fiber array mounted on optical and electrical integrated chip with laser integrated in the chip. Across the three designs, the level of integration increases. As a result, the complication level of assembly process is decreased from the design 1) to design 3). In 1) Hybrid design, Tx/Rx IC chips, ASIC (Application Specified Integrated Circuit) chips and lasers are bonded to an active photonic interposer wafer by thermocompression first. Then the interposer wafer is thinned to expose embedded TSVs (Through-silicon Via) from back surface, followed by backside metallization/UBM and wafer bumping. Next, the interposer wafer is sawed into dies. The die is mounted on an organic substrate and then the organic substrate subassembly is mounted on a PCB (Printed Circuit Board). Fiber arrays are assembled on the interposer dies at last. In comparison, in 2) Monolithic with hybrid laser design, the IC chips are integrated into the active optical interposer which becomes an optical and electrical integrated circuit (OEIC). Hence, only the laser needs to be mounted on the OEIC die before assembly with an the organic substrate. Moreover, with IC function and optical components (except light source) integrated in a single chip, high-density vertical interconnection through chip, TSV array, is no longer needed. Therefore, TSV-related fabrication steps (e.g. wafer thinning, backside metallization, etc.) doesn't exist in design 2) process flow. In 3) Fully monolithic design, the assembly process starts directly with the mounting of an OEIC with integrated laser chip on organic substrate since all the IC function and the optical components are already integrated into the single chip.

Hybrid



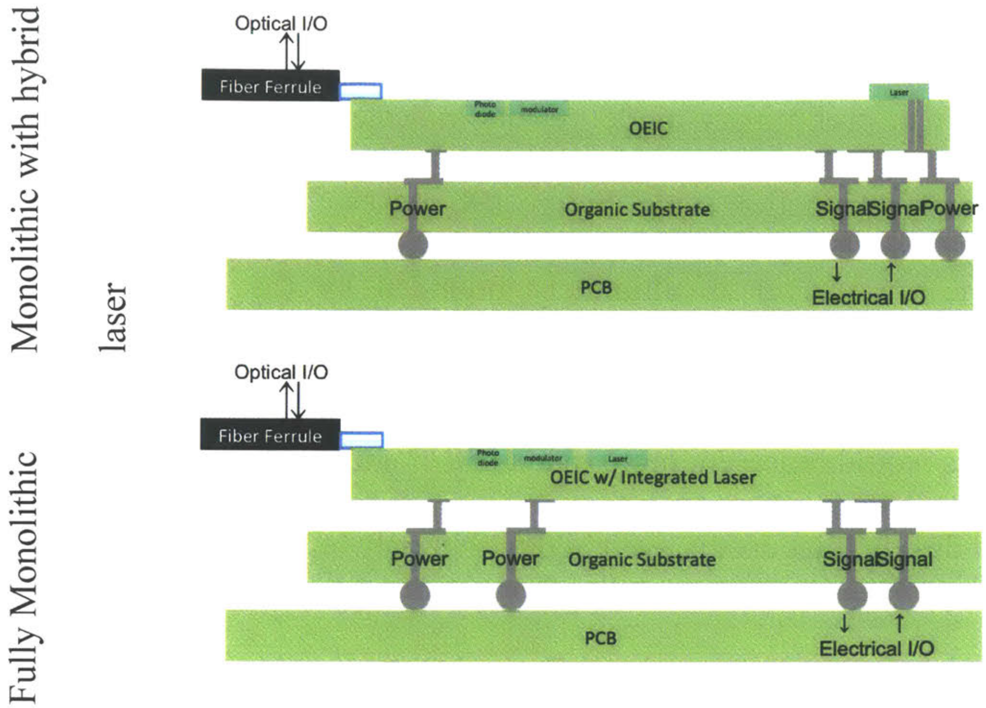


Figure 18. Cartoon sketches of the three designs described and modeled in the paper.

Baseline result

Figure 19 shows the modeled results for the baseline model conditions. These results suggest that the packaged cost of the integrated design should offer cost savings over various hybrid strategies.

A key advantage of the process-based cost modeling method is the ability to explore changes in both technological and operational conditions. One of the most cost-critical operational characteristic is production scale. Figure 19a shows that the model results predict that transceiver packaging costs are strongly a function of production volume. In fact, production volumes over 100k units per year appear to offer costs at least half of the costs at 10k units per year. Despite the fact that costs change significantly over thing range of production volume, the fundamental finding remains – these preliminary model results indicate a cost advantage for integration irrespective of production scale. Figure 19b makes clear that, at high volumes, the cost of the transceiver is ultimately bounded by the cost of the constituent components. For the fully integrated design, modeled results suggest that components would represent more than 60% of unit costs at 100k units per year. For the hybrid design, that cost is even higher, but only represents 45% of total unit cost.

Preliminary results suggest that monolithic integration of the Datacom transceiver has the potential to significantly lower packaging cost for both high and low volume production.

Although these results are preliminary, they suggest that integration offers real potential for reducing module cost. These results also raise two key questions: 1) What is driving the cost difference between the three strategies? and 2) How much do baseline conditions need to change before these results are changed?

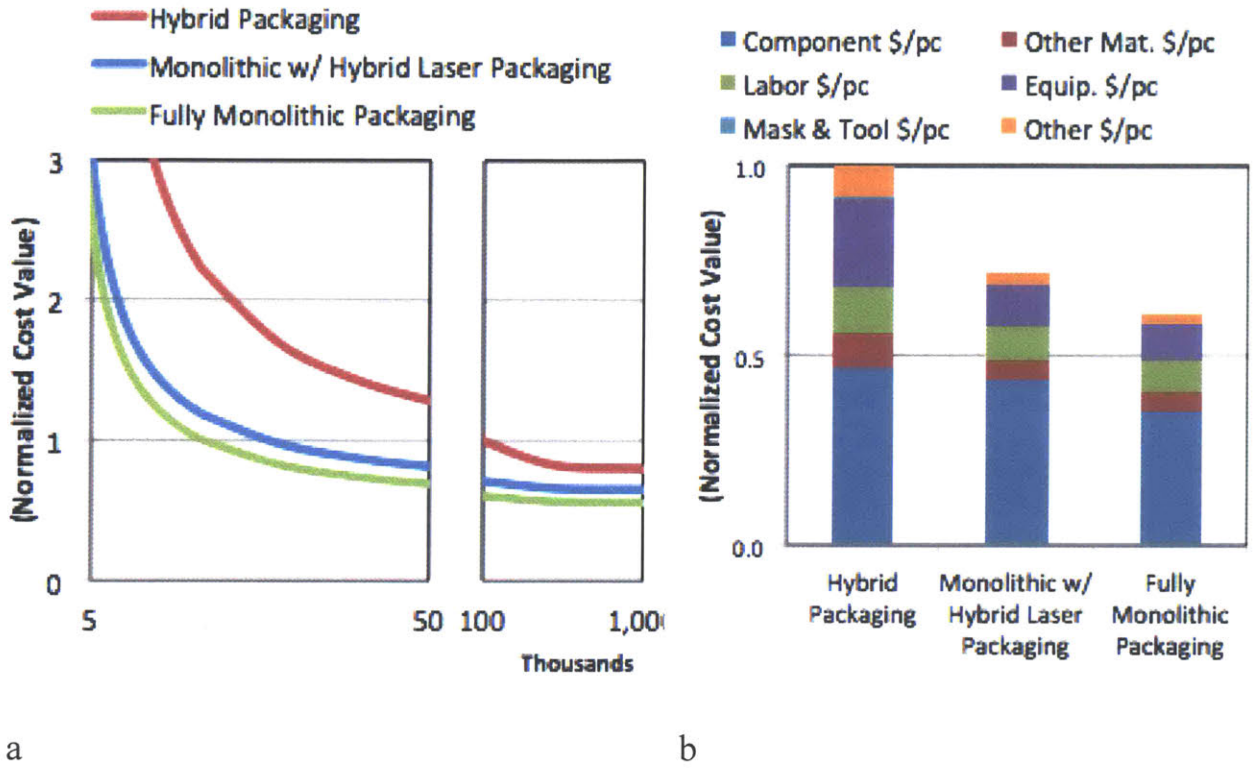


Figure 19. Baseline model result. a) Modeled unit cost versus production volume and b) Cost breakdown by cost element for the three different levels of integration

Mapping the drivers of cost difference

For roadmapping, it is important to not only quantify the expected cost difference between two alternatives, but also the drivers of those differences. As technology evolves, those drivers may be amplified or muted.

Two explore this question, we apply the cost model in two different analyses. The first is summarized in Table 5 and Figure 20. These both present different perspectives on the cost differences among the three designs.

Table 5. Breakdown of modeled cost for the three designs at 100k units per year.

		Hybrid		Monolithic w/ Hybrid Laser		Fully Monolithic	
		9 components, 17 process steps		5 components, 8 process steps		4 components, 7 process steps	
Component	PCB / Org Sub / Fiber Array	1 ea – 0.01					
	Optical Power	1 – 0.01				1 Integrated die	0.29
	Photonics	1	0.29	1 Integrated die	0.29		
	Circuits	4	0.04				
Process Step	Optical Component Assembly Steps	2 – 0.09				1	0.07

	Other Assembly Steps	12	0.33	3 – 0.07
	Test Steps	3 – 0.05		

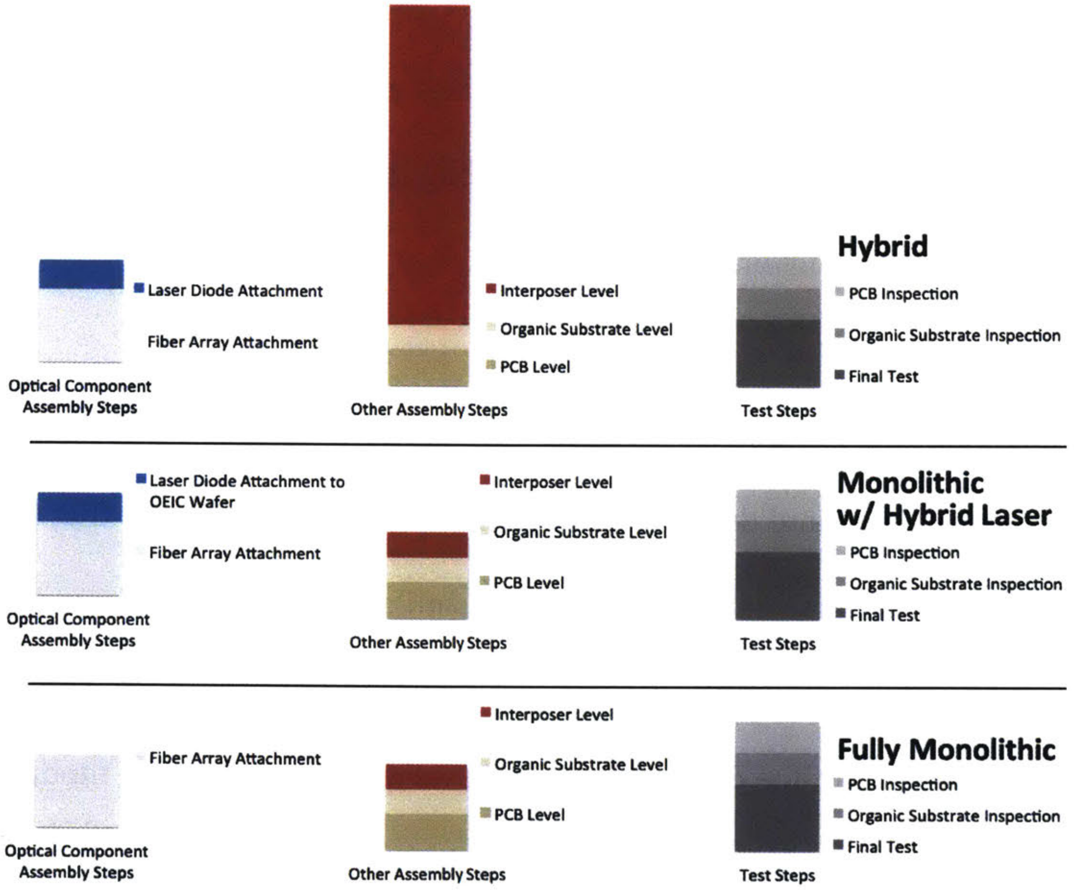


Figure 20. Breakdown of modeled costs by activity. To highlight differences, the activities are grouped by those related to i) optical component assembly, ii) other component assembly, and iii) test steps.

This analysis suggests that the largest drivers of cost difference across the three designs derives from a) the laser diode attachment (required for both hybrid designs) and b) the assembly of the interposer for the first hybrid design. As detailed in the table, the costs associated with the interposer are primarily from processing rather than components.

Preliminary model results suggest that the key cost savings opportunity for integrating in the near term derives from avoiding the expense of assembling and packaging the interposer layer.

As mentioned earlier, this current analysis very preliminary particularly regarding the cost assumptions around the optical chips. In particular, we assume that the cost of chips is constant per area and that the yield is similar for each chip. To understand the implications of these simplifications, we explored how the cost would vary depending on the yield of the chip. In particular, we quantified the total module cost at a range of chip yields for each of the three designs. Using this information we identified the point at which a lower yield for the two more integrated designs would climb to parity with the more conventional hybrid design. This analysis is plotted for a production volume of 100k units per year in Figure 21. From this, we can see that that model suggests that the cost advantage of the Monolithic with Hybrid Laser packaging would remain cost competitive even if yields fell to 60% and the Fully Monolithic packaging would remain competitive at yields well below 50%.

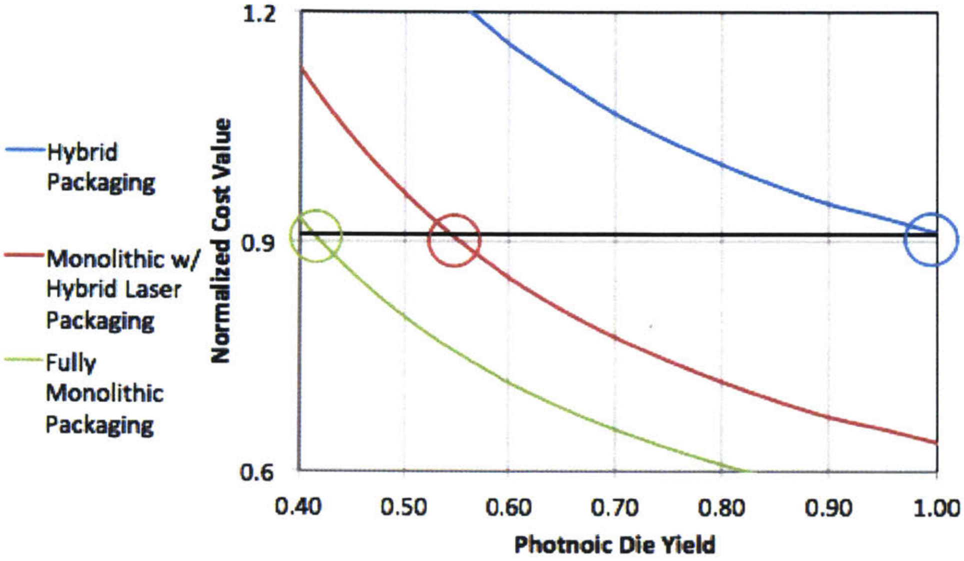


Figure 21. Yield sensitivity of three designs when annual production volume equals 100k units per year. Black line indicates packaging cost of the Hybrid design for baseline yield conditions.

Preliminary model results suggest that integration has significant cost advantages even if optical chip yields were to fall well below baseline modeled values.

Conclusions

The model results presented here are very preliminary and should not be interpreted as providing specific guidance. Nevertheless, the analysis presented in this chapter indicate the potential for such tools within the roadmapping process. In particular, these tools will allow the TWGs to isolate key points of cost leverage and to explore the cost ramification of promising technical solutions.

Manufacturing cost estimation methods

BOM

Activity-based cost modeling

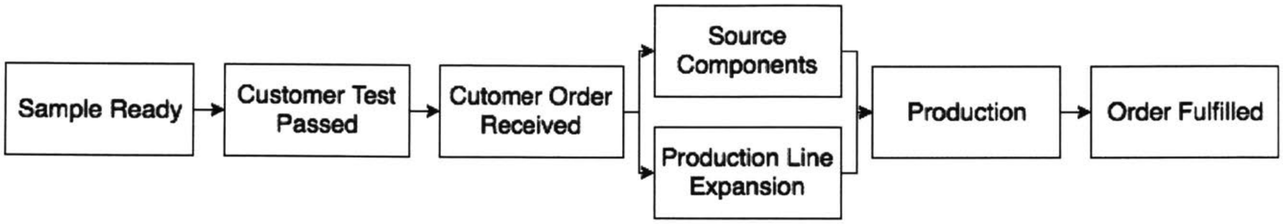
Process-based cost modeling

Process-based cost modeling

Manufacturing time (Scale-up time)

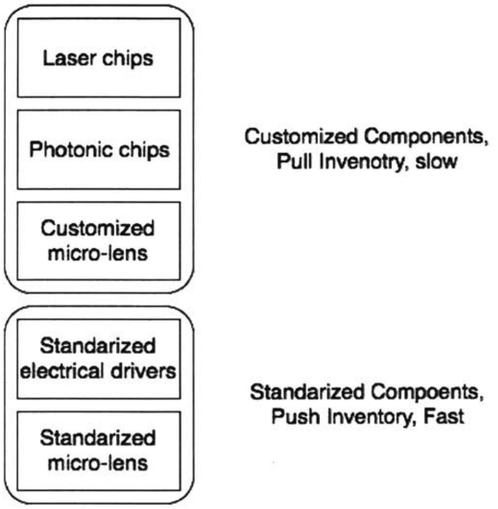
For the term time-to-market, people tend to think the main limiting factor is product and process development time. However, in the case of integrated optical transceivers, the real pain for potential buyers are the time from product prototype to volume order fulfillment, the time of scaling-up.

Due to the volatility of the market demand, many optical transceiver vendors are still taking the pull strategy for the first batch of orders. They send prototype samples to potential customers for test first. They don't source components or expand production capacity until they receive the first batch of volume orders from customers.



$$T_{scaleup} = T_{customertest} + \max(T_{component}, T_{capacityexpand}) + T_{manufacturing}$$

The step on component sourcing took the longest in the case of 100G 500m-2km reach optical transceivers. The components were micro-lens, photonic chip, laser chip and electrical driver chips. Among those, electrical driver chips were ready before the other components. They were already manufactured before volume order of transceivers were received because usually different optical transceiver vendors shared the same electrical drivers. Both photonic chips and laser chips took long. Firstly, they were customized for the optical transceivers. Their demand was so volatile that chip foundries also took the pull strategy, i.e. they didn't start planning produce those chips until they received confirmed orders. Secondly, there was a queue for manufacturing those chips in foundries. Especially the volume of those chips was relatively small for a foundry's capacity and thus foundries usually put those optical transceiver-related chips in the low priority. That explained why it took only one or two growths to produce thousands of laser chips but it took half a year for those laser chips to be delivered to optical transceiver vendors.


Laser chips
Photonic chips
Customized micro-lens
Customized Components, Pull Invenotry, slow
Standarized electrical drivers
Standarized micro-lens
Standarized Compoents, Push Inventory, Fast

Comparison among different transceiver designs

System implementation

There may be an argument why transceiver vendors should care about system implementation cost and time. A short answer is that a good salesman should know his customers' need better than themselves.

Do customers know their needs?

In the case of optical transceivers, it seems those customers (hyper-scale data centers, Google, Amazon, Microsoft, etc.) always declare what they want clearly and vendors should just listen to the customers and produce what they want. For short-term needs, the customers indeed know what they want. All the information can be exchanged through orders. However, for long-term needs, the customers are unable to announce a clear and firm demand. For example, the customers said they wanted optical transceivers coming with 100G, QSFP form factor, 2km reach and a price of $1/Gbps back in 2014[12]. Even nowadays it is still not possible to fulfill all the requirements yet. What those customers did was relaxing their requirements. They shortened the 2km reach requirement to 500m reach. They bought fewer transceiver modules than what they declared before but paid a price higher than $1/Gbps.

Customers' own long-term demand is not accurate.

There is a knowledge gap between transceiver vendors and data center customers. The knowledge gap makes it difficult for data center customers estimate what may be feasible for their vendors to deliver on time. The knowledge is not only engineering design but also the knowledge of manufacturing and vendors' suppliers.

The need of filling the knowledge gap is especially important for optical transceivers for the following two reasons. Firstly, it is a fast-cycle industry. The frequency of bandwidth expansion is currently set at every 3 years. The time for R&D and manufacturing scale-up already take longer than 1 year. To meet the industry cycle, both customers and vendors should establish a mutual understanding as early as possible. Secondly, there are too many moving pieces to define a product performance. During the group discussion in AIM Photonics for future Datacom transceivers, vendors found it very difficult for them to narrow the scope of future transceiver products. There are multiple choices for each of those following factors: number of channels, number of wavelengths, number of fibers, power density, size, bandwidth density, distance between those transceivers and ASIC, reach etc. They depend on the system requirement. System-level knowledge

Additionally, the mutual understanding can help lower the cost and speed up the technology transition. Today different data centers want different kind of optical modules.

Conclusion

Import to look things at a system level, considering all six factors.

Penalty of fragmentation

Slow, costly

If they ramp up fast, they can sell for a higher price.

Chapter 5
Cost and Time for System Integration

Introduction

Intra-data center network is the communication channel among individual servers in the same data center building. The network bandwidth limits the communication speed. To meet execution time goals for larger and larger data sets and match faster and faster servers, data centers demand larger and larger network bandwidth. Every 3~4 years hyperscale data centers expand their network bandwidth capacity.[1][2]

For each upgrade, component vendors and big data centers have been discussing about the timing and the most cost-effective technology choice should be. Different parties hold different opinions about the trade-offs between different factors. For example, Google has already decided to take the intermediate step to 200G because 400G technology won't meet Google's cost goals soon enough while other hyperscale data centers prefer to go to 400G technology directly.[3] In that example, the trade-off is made between cost and lead time. We can't help ask the following two questions:

How soon is enough for 400G technology to meet Google's cost goals soon enough? What factors driver other hyperscale data centers prefer to go to 400G technology directly? What are the critical points for decision switching?

A systematic quantitative model is necessary for us to integrate different parties' opinions and see the big picture of the market. However, by far in the world of data center network capacity planning, there are very few published quantitative economical models to answer those two questions.

Fortunately, there is a rich set of economic and operation management theories and models to quantify similar capacity planning processes in retailing and manufacturing. Here I adapt a classic inventory management model to study the intra-data center network bandwidth capacity planning problem. In the first section, I describe the classic inventory management model in capacity planning and supply chain design. In the second section, I model the data center network capacity planning process as an optimization problem based on the classic inventory management theory. In the third section, I apply the network capacity planning model to quantify the trade-off between cost and lead time for 100G, 200G and 400G technologies and explain why different hyperscale datacenters prefer different technologies.

Classic inventory management model

Inventory is a buffer stock between the input and the output of the system. For a retailer, inventory refers to those goods in the retailer's warehouse and also the goods already purchased but haven't arrived in the warehouse yet. Inventory protects against uncertainties in demand, covering the lead time of ordering, transportation and processing. However, holding inventory also causes holding cost, including storage cost, obsolescent cost and capital loss. To balance the benefits and costs of holding inventory, people need to make decisions to determine how much inventory to hold and how to replenish. The classic inventory management model is a quantitative tool to help people make such decisions.

$$Total\ Cost = \ Purchase\ Cost + Ordering\ Cost + Holding\ Cost + Shortage\ Cost \quad (2.1)$$

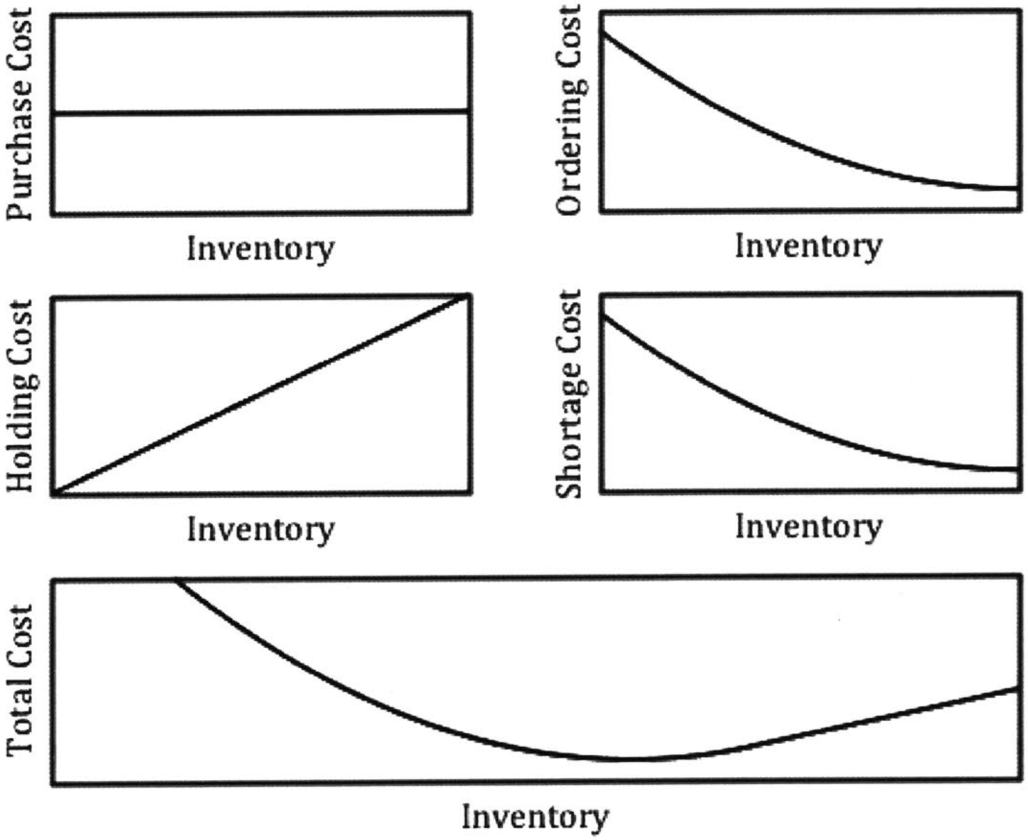


Figure 2.22 Trends of four basic cost and total cost with inventory

The objective is to minimize the total cost, which is the sum of four kinds of basic costs associated with inventories: purchase, ordering, holding, and shortage costs.[4] 1) Purchase cost is a variable cost for acquiring those sold goods. Ideally, when there is no cost in order placement, the future demand is known exactly and the lead time is zero, no inventory is needed and then purchase cost is the only expense. Determined by demand and unit cost, purchase cost doesn't change with inventory decisions. Therefore, adding purchase cost neither encourages nor discourages keeping inventory. 2) Ordering cost, also known as setup cost, is a fixed cost for placing an order. It is proportional to the times of order placement. Ordering cost keeps people from reacting to demand one by one and encourages large and few orders. Adding ordering cost favors having inventory. 3) Holding cost, also known as carrying cost, is a cost of keeping inventory. It is proportional to the value of goods in inventory. It includes costs of storage, insurance, tax, loss/shrinkage, damage, depreciation, obsolescence, and capital. Out of the four basic costs, only holding cost discourages keeping inventory. 4) Shortage cost results if an item is not available when demanded. It is also known as stock-out cost or penalty cost. When the customer is willing to wait, it is the cost of a backorder, which is the extra money spent to acquire a product. When the customer goes elsewhere for that purchase, it is the cost of lost sales, equal to the profit loss. Shortage cost also encourages holding inventory. The best inventory policy is the policy that balances the benefits and costs of holding inventory and minimizes the total Cost.

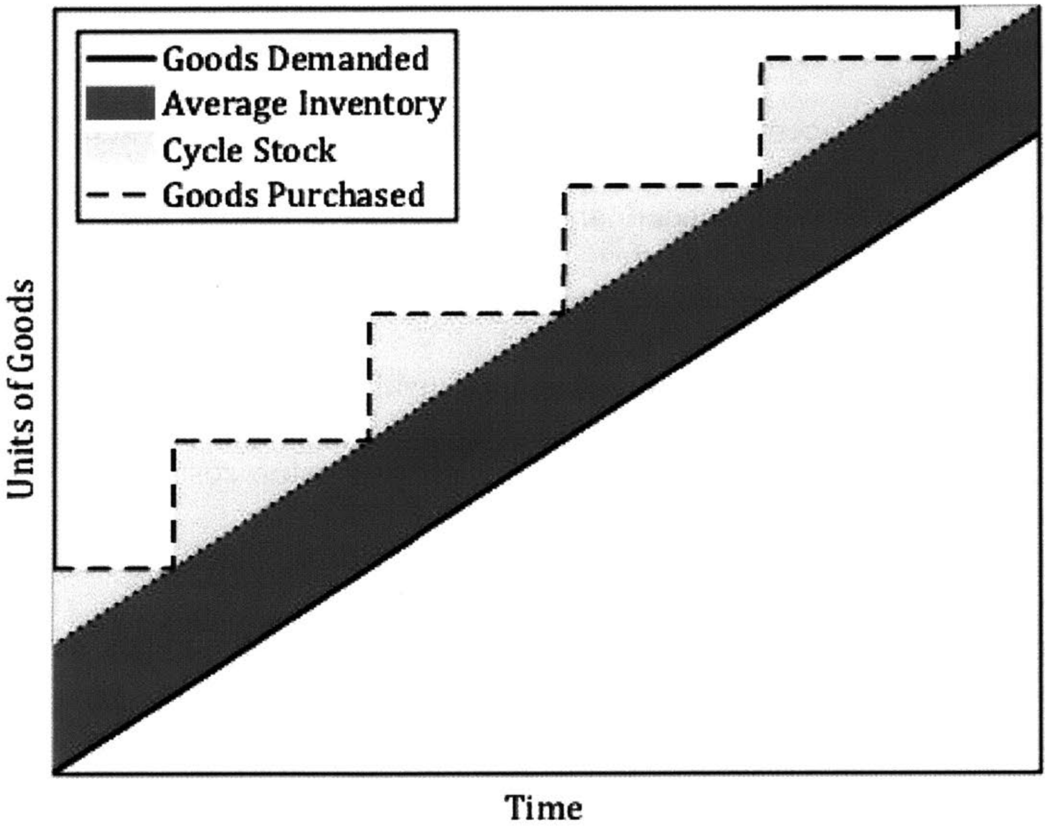


Figure 2.23 An inventory example of constant demand over time

Here is an example of a retailer's inventory. The retailer forecasts that the expected demand is constant over time. The solid line represents the quantity of goods sold until the time point. It shows the cumulative output from the retailer. The dashed line represents the quantity goods purchased until the time point. It shows the cumulative input into the retailer. The difference between the solid line and the dashed line stays in the inventory. The inventory is composed of two parts: cycle stock and average inventory. 1) The cycle stock is the periodic part. It is intended to meet the expected demand. The peak value is the quantity of goods purchased in each order, which equals to the quantity of goods demanded during each replenishment cycle. 2) The constant part is the average inventory. It includes both the pipeline inventory, which has been purchased but not arrived yet, and the safety stock, which is held to reduce the probability of stocking out due to demand uncertainty. Now that we have seen what an inventory is like, I will give out the mathematical formulation of the inventory management model. Please note that there are many variations of situations in inventory management. Different models are applied to solve the best inventory policy in different situations. Here I only describe one of the simplest models.

The monthly demand fits the normal distribution with a mean value of D units/month and a standard deviation of σ units/month. The unit cost is C $/unit. I assume the retailer orders Q units of goods in each order. For each order, the retailer pays Ct $/order upfront cost. It takes a lead time of L months from placing an order to receiving the order and having those goods ready for sale. The holding cost ratio is h $/inventory $/month, which means that each dollar value of goods in the inventory costs the retailer h $/month. Not having a unit of product on-hand to satisfy the demand costs the retailer Cs $/unit. To prevent shortage, the retailer always keeps k*σ*L units of goods as safety stock, where k is the safety factor. That is to say, the retailer places a new order of Q units of goods when the inventory on-hand is less than k*σ*L units of product. By doing so, the expected units in shortage will be σ*L *G(k)*D/Q units/month, where G(k)=φ(k)-k*(1-Φ(k)) and φ and Φ is the probability density function and cumulative distribution function of standard normal distribution.

Therefore, each month the retailer spends C*D $/month on purchasing the goods that are actually sold in the month. It needs to order D/Q times to fulfill the demand. Then the ordering cost is Ct*D/Q $/month. There are Q/2+D*L+ k*σ*L units of goods in inventory and the holding cost is h*C*(Q/2+D*L+ k*σ*L) $/month. The shortage cost is Cs*σ*G(k) $/month. We want to find the (Q, k) that minimizes the sum of purchase cost, ordering cost, holding cost and shortage cost. The mathematical formula is

$$\min_{Q,k} CD + C_t \frac{D}{Q} + Ch\left(\frac{Q}{2} + DL + k\sigma L\right) + \frac{C_s \sigma L G(k) D}{Q}. \qquad (2.2)$$

Please note that by writing the mathematical formula as above, I have two implied assumptions. The first is that the warehouse space is always large enough and thus I drop the space constraint. The second is that the transportation cost per item can be ignored or incorporated into the purchasing price and thus there is no separate transportation cost item in the objective function.

A strict solution satisfies

$$\begin{cases} \dfrac{\partial Total\ Cost}{\partial Q}|_{Q,k=Q^*,k^*} = 0 \\ \dfrac{\partial Total\ Cost}{\partial k}|_{Q,k=Q^*,k^*} = 0 \end{cases}. \tag{2.3}$$

That means

$$\begin{cases} Q^* = \sqrt{\dfrac{2D\left(C_t + C_s\sigma LG(k^*)\right)}{Ch}} \\ k^* = \Phi^{-1}\left(1 - \dfrac{Q^*Ch}{DC_s}\right) \end{cases}. \tag{2.4}$$

We can see that Q* and k* are tangled together through normal distribution probability density function and it is difficult for us to find an exact analytical expression for the final solution. To ease the decision process, usually people determine Q^* without considering shortage cost. Thus,

$$\begin{cases} Q^* = \sqrt{\dfrac{2C_tD}{Ch}} \\ k^* = \Phi^{-1}\left(1 - \dfrac{Q^*Ch}{DC_s}\right) \end{cases}. \tag{2.5}$$

The optimal order quantity Q^* is also called as the Economic Order Quantity (EOQ)[5]. Correspondingly, we can get the optimal order time interval.

$$T^* = \frac{Q^*}{D} = \sqrt{\frac{2C_t}{hCD}} \qquad (2.6)$$

Intra-datacenter network capacity planning model

In general, the intra-data center network capacity planning problem has an analogous mathematical structure as the inventory management problem. We can look at the intra-data center network bandwidth for data centers and think that it is similar to goods for retailers. A data center wants to have enough bandwidth on hand to avoid running out of bandwidth. It forecasts the demand increase first and then makes the purchase based on the forecast. The gap between the bandwidth purchased and the bandwidth actually demanded is its inventory. There are also four basic costs involved: purchase cost, ordering cost, holding cost and shortage cost. The data center needs to decide what, when and how many to buy to minimize the summed cost.

While there are still differences between network capacity planning and inventory management, we can turn the network capacity planning problem into an equivalent mathematical problem as the inventory management problem by modifying the definitions of variables. In the following, I will describe those modified definitions and then describe the results in the network case.

Demand

In the retailer case, I assume the retailer sells D units of goods each month in average with a stand standard deviation of σ units/month. From month 0 until month t, the expected number of goods sold is D*t units. Here in the network case, I assume the data center demands D Gbps additional bandwidth per month in average with a standard deviation of σ units/month. From month 0 until month t, the expected bandwidth demanded is D*t Gbps. If the replacement of end-of-life components is also taken into consideration, the assumption becomes that the sum of monthly additional demand and monthly replacement fits the normal distribution with a mean value of D Gbps/month and a standard deviation of σ Gbps/month.

Unit cost

In the network case, one unit means one set of network hardware components to support 1 Gbps non-blocking bandwidth from rack to rack. The unit cost C $/Gbps includes not only the one-time expense on purchasing and assembling the network hardware components but also the recurring expense on operating and maintaining them. All the cost items that are naturally proportional to bandwidth are included into the unit cost. In addition, unit cost is a solution-specific property in the model, which means different technology solutions have different unit cost. Last but not the least, the unit cost is normalized by performance, i.e. bandwidth in the network case.

Purchase cost

Purchase cost is the cost of acquiring the bandwidth demanded. Following the new definitions of demand and unit cost in the network case, the purchase cost in a given month is still C*D $/month.

Ordering cost

One order refers to one time of network bandwidth expansion. The ordering cost includes not only those general cost items such as the cost of placing and processing an order but also those specific cost items existing only in the network bandwidth expansion case such as the cost of designing a new set of network switch hardware and software. All the cost items that are naturally proportional to frequency of bandwidth expansion are included into the ordering cost.

Holding cost

Holding cost is the cost of carrying inventory. The traditional holding cost items include the physical cost such as warehouse cost and the capital cost of money used in buying inventory. In the fast-paced high-tech industry such as PC manufacturing in 1990s, there are also other inventory-driven cost items, such as component devaluation cost and obsolescence costs. [6] Component devaluation means that components drop in price. For example, back in 1990s, the price of a CPU might drop 40% in the first nine months after it came out to market. Similarly, in the network case, the price of an optical transceiver might drop 50% in the first year. Obsolescence means a product becomes outdated and its value needs to be written off or written down if there is recycling or scrapping benefits. For example, when a PC manufacturer decides to discontinue a particular product, the company needs to write off 100% of the value of finished goods in its inventories and the value of any components in the pipeline. In the network case, when a data center migrates from one technology to another technology, the old technology hardware, if not reused elsewhere or recycled, and the labor efforts on improving the old technology, if not transferable to new technology, will lose value. The obsolescence cost is also characterized as a risk. The risk increases as time proceeds.[7] Both devaluation cost and obsolescence cost are still monotonically increasing functions of value of goods in inventory and time. I assume both of those two costs are still linearly proportional to value of goods in inventory and time like the traditional holding cost items. Then an aggregated holding cost ratio can be applied to absorb all the holding cost items together. Currently, IT hardware usually has a product life cycle of 3 years. Thus, I assume h is equal to 0.025/S/month, which means at the end of a network component's 3-year product life cycle, its value decreases to 10% of its original value in the beginning.

Shortage cost

Since the bandwidth shortage doesn't necessarily cost money, it is difficult to quantify such an indirect cost. Every time the bandwidth is insufficient, the immediate outcome is that the customer (a web service user or a server) needs to wait longer than scheduled. The larger the bandwidth shortage is, the longer the wait is. If the wait results in failure in displaying an advertisement or loss of users, then the bandwidth shortage costs money. Here we can view the shortage cost as a kind of risk. I assume the risk is proportional to bandwidth shortage and then the shortage cost Cs $/Gbps means that every 1 Gbps bandwidth in short is expected to cost $ Cs.

Lead time

In the retailer case, lead time is simply the time interval between placing order and receiving the order. In the network case, lead time is the time interval between placing the first order of network components and getting all the components installed and functioning. It is because that after datacenter receives all the components, those components can't provide bandwidth right away until they are correctly installed and integrated into the datacenter interconnect system. Besides, same as unit cost, lead time is also a solution-specific property.

Single-period, finite horizon and infinite horizon

For inventory optimization models, there are three kinds of optimization goals in terms of time horizon: single period, finite horizon and infinite horizon. Single-period means that the goal is to optimize the cost in a predetermined time period, e.g. 1 year, 3 years, or 5 years. It ignores the cost incurred outside the given period. Finite horizon means that the goal is to optimize the summed cost in a finite number of periods and ignores the cost incurred after. The model needs to either give out the cost expression for each period, or give out the cost expression for the first period and specify the relationship from a period to its following period. One of the simplest situations is that the first period repeats itself in the following periods. Infinite horizon means that the goal is to optimize the cost in an infinite time horizon. The optimization model I have shown for the retailer case, which I will also apply to the network case, is taking an infinite horizon. It assumes a steady state condition, i.e. in the network case, it assumes current variables, including the expected demand increase speed D Gbps/month, ordering cost (set up cost) Ct \$/expansion and the unit cost of bandwidth C \$/Gbps, are constant. In practice, datacenters may look at the costs within 3~5 years and do a single-period optimization. However, the choice that one period lasts 3~5 years is actually balancing the ordering cost and the holding cost. It shows Economic Order Quantity (EOQ) thinking behind and EOQ is a result of the infinite horizon view.

With such modified definitions, the same objective function in the retailer case can also be applied to the network case.

$$\min_{Q,k} CD + C_t\frac{D}{Q} + Ch\left(\frac{Q}{2} + DL + k\sigma L\right) + \frac{C_s\sigma LG(k)D}{Q} \tag{2.7}$$

Please note that here I assume that the space is always large enough and the cooling system is always powerful to simplify the problem. In reality, it is not always the case. Instead, old datacenters usually have space and power constraints because that datacenter architects and mechanical engineers who designed datacenters ten years ago may underestimate the space and cooling power needed by network components. That is one of the reasons why datacenters limit the size and power consumption of each network component.

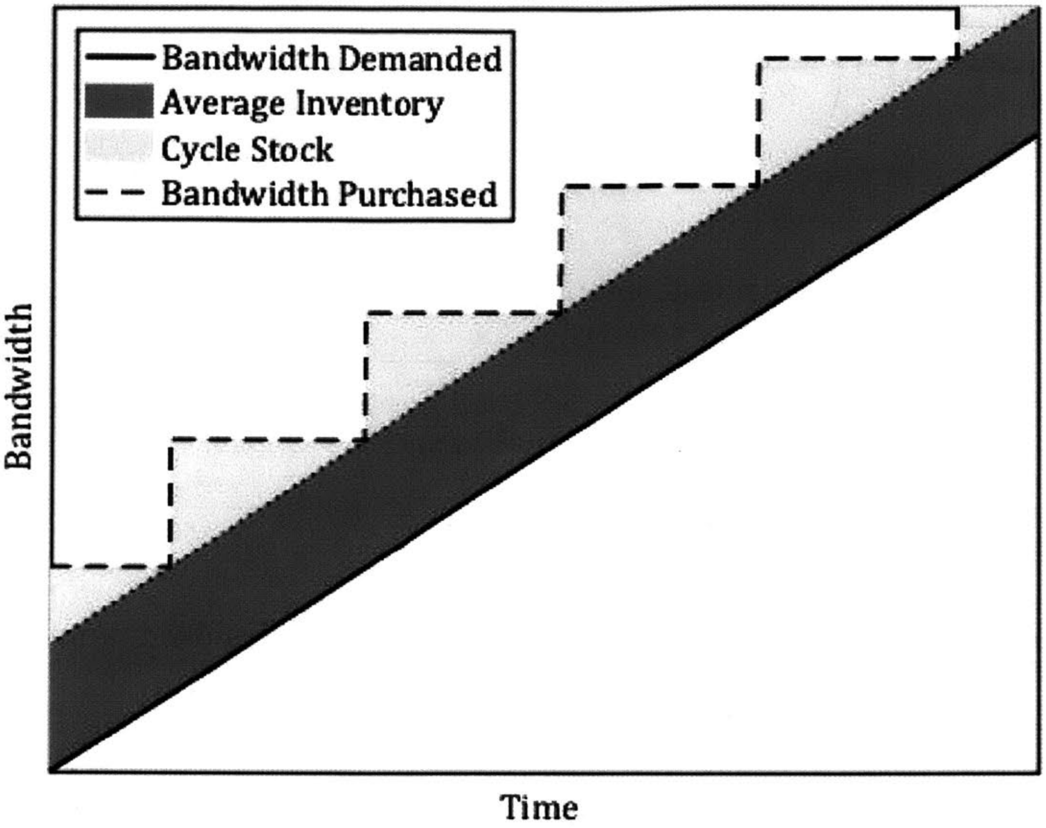


Figure 2.24 Intra-data center network bandwidth expansion

Same as the retailer case, the optimal order quantity Q^* and the safety factor k^* are

$$\begin{cases} Q^* = \sqrt{\dfrac{2C_tD}{Ch}} \\ k^* = \Phi^{-1}\left(1 - \dfrac{Q^*Ch}{DC_s}\right) \end{cases}. \tag{2.8}$$

Once $C, C_t, D, h, C_s, \sigma, L$ are given, the minimized total cost is set. And its value is

$$CD + \sqrt{\frac{CC_tDh}{2}} + Ch\left(\frac{\sqrt{\frac{2C_tD}{Ch}}}{2} + DL + \Phi^{-1}\left(1 - \frac{\sqrt{\frac{2Ch}{D}}}{C_s}\right)\sigma L\right) + \frac{C_s\sigma LG\left(\Phi^{-1}\left(1-\frac{\sqrt{\frac{2Ch}{D}}}{C_s}\right)\right)D}{\sqrt{\frac{2C_tD}{Ch}}}. \quad (2.9)$$

And still, the optimal order time interval is

$$T^* = \frac{Q^*}{D} = \sqrt{\frac{2C_t}{hCD}}. \quad (2.10)$$

Nowadays datacenters expand its bandwidth every 3~4 years. Here I take T^* as 3 years and assume that C is \$1/unit bandwidth, D is 1 unit bandwidth/month. And as mentioned in last section, the holding cost ratio h is 0.025 /\$/month. Then I back-calculate that ordering cost C_t is \$16 per expansion. The dollar and unit bandwidth here have arbitrary values.

Uncertainty evolution with time

In previous sections, the demand increase standard deviation is assumed to be constant. However, in real life, people usually have more accurate forecasts about near future demand than far future demand. That is to say that the uncertainty grows in time. For example, in the Black-Scholes option-pricing model, the instantaneous log return of future stock price is modeled as a Brownian motion (aka. random walk) and for a Brownian motion without drift, the standard deviation is proportional to the square root of time. Such uncertainty evolution model is also widely adopted by operation management researchers to describe the uncertainty evolution of demand forecast[8][9]. If we take the same approach, then the demand increase standard deviation σ is not constant any more. Instead,

$$\sigma = \sigma_u D\sqrt{L}. \tag{2.11}$$

Please note here I normalize the standard deviation by the expected value. If we assume that the forecast of demand increase may be 60% off, then $\sigma_u = \frac{0.6}{1\sqrt{36}} =$ $0.1\ /\sqrt{month}$.

After (2.11) is plugged in, the objective function becomes

$$\min_{Q,k} CD + C_t\frac{D}{Q} + Ch\left(\frac{Q}{2} + DL + k\sigma_u DL^{1.5}\right) + \frac{C_s\sigma_u L^{1.5}G(k)D^2}{Q} \tag{2.12}$$

Table 2.6 Base case: current solution

Inputs		
Name	Value	Source
Expected demand increase D	1 unit /month	Assumed.
Demand forecast volatility σ_u	$0.1\ /\sqrt{month}$	Back-calculated based on the assumption that

		forecast about 3 years later can be 60% off.
Ordering cost C_t	\$16 /expansion	Back-calculated based on the observation that datacenters expand intra-data center network bandwidth every 3 years.
Unit cost C	\$1 /unit	Assumed, a solution-specific property.
Lead time L	6 months	Assumed, a solution-specific property.
Holding cost ratio h	0.025 /\$/month	Back-calculated based on the observation that the product lifecycle of datacenter IT hardware is 3 years.
Shortage cost per unit C_s	\$4 /unit	Assumed.
Outputs		
Name	Value	Note

Optimal order quantity Q^*	36 units/expansion	Determined without considering shortage cost.
Optimal order time interval T^*	36 months (=3 years)	Determined without considering shortage cost.
Safety factor k^*	0.76	Determined after Q^*.
Purchase cost	$1 /month	
Ordering cost	$0.45 /month	
Holding cost	$0.63 /month	
Shortage cost	$0.89 /month	
Total cost	$2.97 /month	

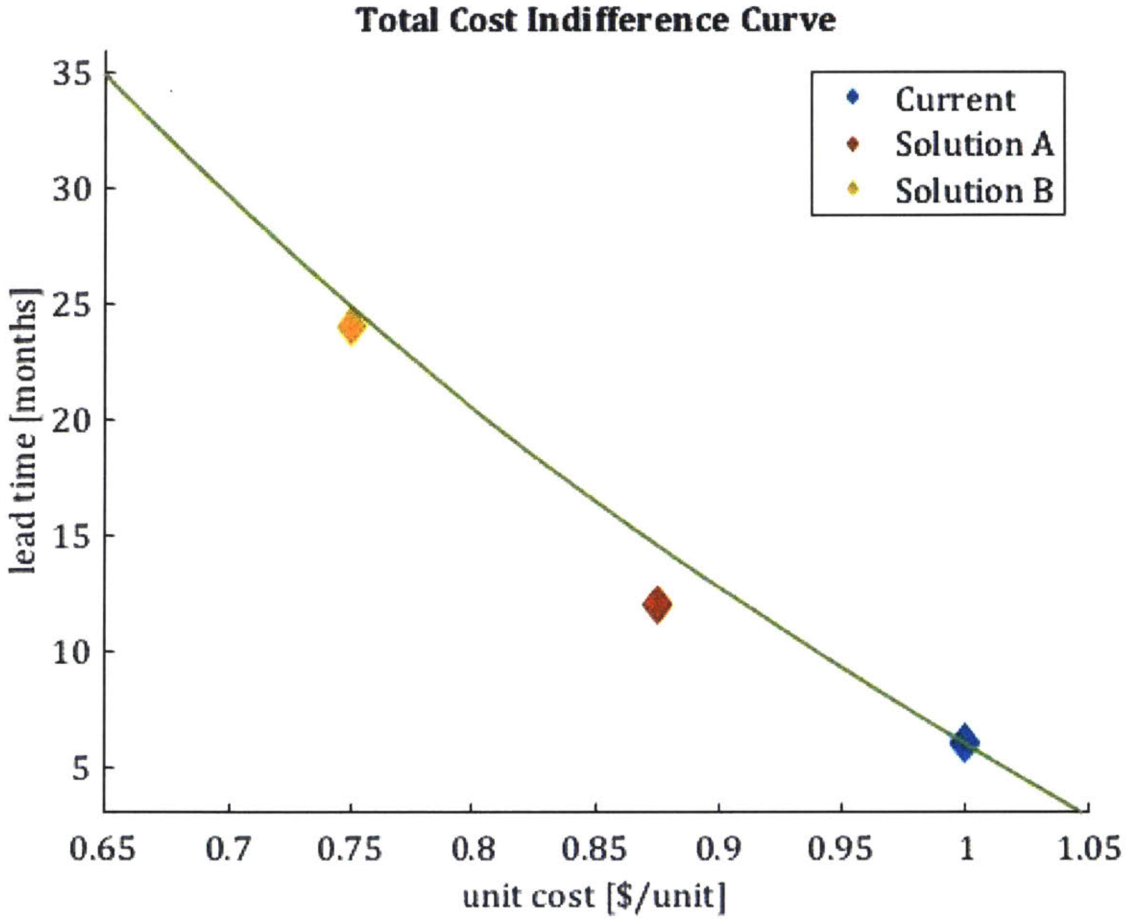


Figure 2.25 Total cost indifference curve of solutions whose total cost equals to current solution

When the minimized total cost is fixed, for each value of unit cost C, we can solve a lead time L. Then each (C, L) pair will result in the same minimized total cost. If we plot all those (C, L) pairs together, we can get a total cost indifference curve. All the technology solutions are divided into three categories by the curve: the technology solutions that cost the same as current, the technology solutions that are cheaper, the technology solutions that are more expensive. In Figure 2.25, there are three technology solutions labeled: current solution $(C, L) = (1,6)$, future solution A

$(C, L) = (0.875,12)$ and future solution B $(C, L) = (0.75,24)$. According to the total cost indifference curve, both future solutions have cheaper unit cost but longer lead time than current. From the curve, we can also see that when unit cost drops by \$0.25/unit, adding ~19 months of lead time will cancel the unit cost saving and give out the same total cost. That explains why future solution A is the cheapest in terms of total cost although solution B offers the lowest unit cost.

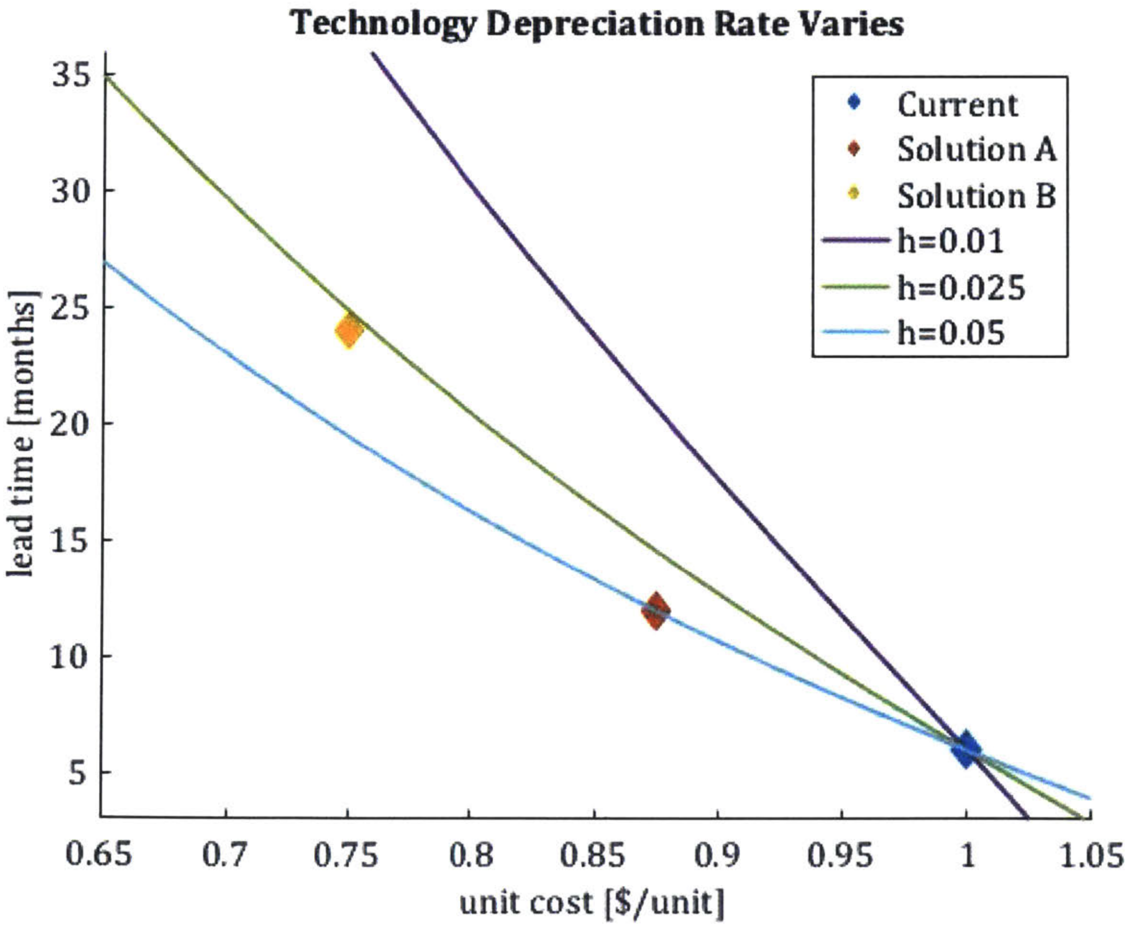


Figure 2.26 Total cost indifference curves with different technology depreciation rates

Varying technology depreciation rate

In the base case, I assume holding cost ratio h as 0.025 /$/month, which is a back-calculated result from the observation that the product life cycle of IT hardware is 3 years. What if technology depreciates faster or slower? In Figure 2.26, it shows that the faster technology depreciates, the flatter the indifference curve and the more sensitive total cost is to lead time. The result fits our common sense that a short lead time is very valuable in a world where technology changes rapidly.

Varying demand volatility

When demand is more volatile, the minimized total cost is more sensitive to lead time.

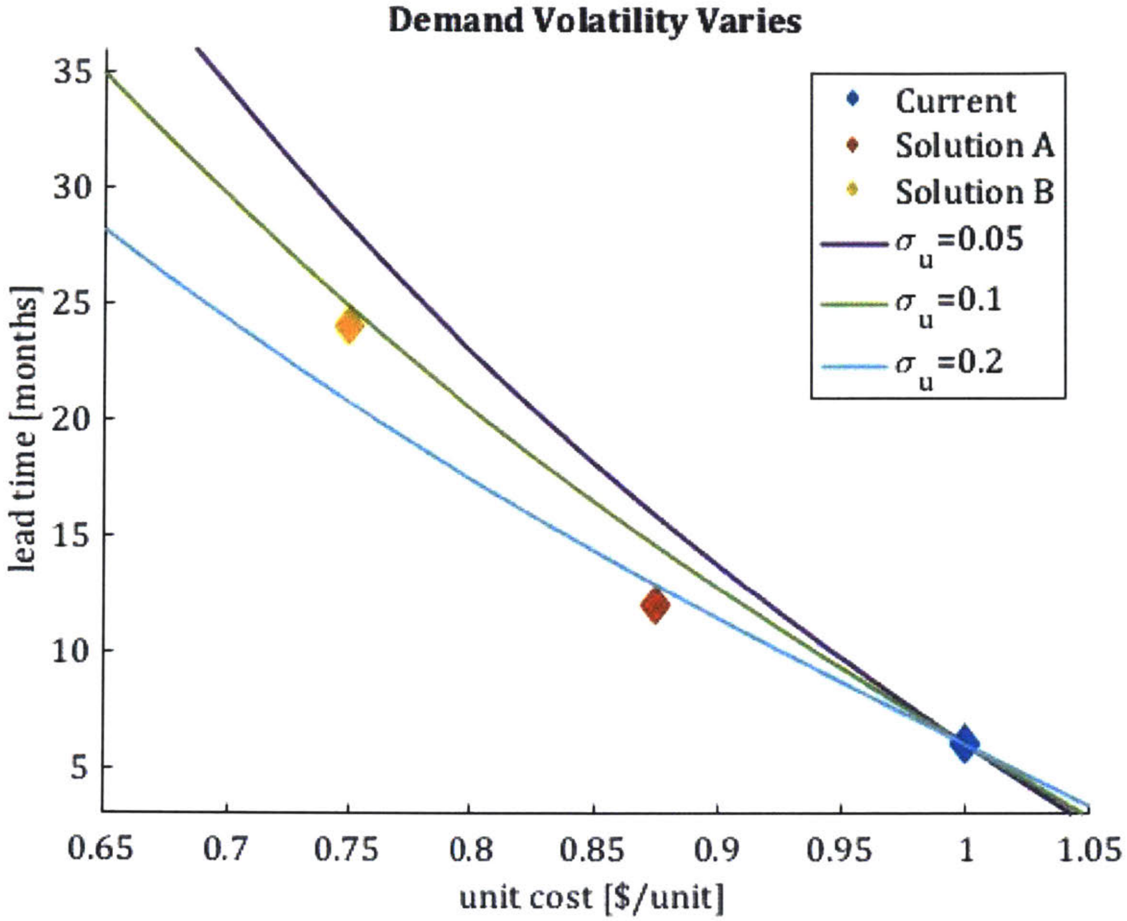


Figure 2.27 Total cost indifference curves with different demand volatility

Different ordering costs

For a large company I with 4N datacenters, ordering cost can be shared among its 4N datacenters. For a small-scale company II with N datacenters, ordering cost is burdened by only N datacenters. Then the single datacenter ordering cost of company II is four times that of company. The smaller a company is, the higher its ordering cost is. Then its optimal ordering time interval is longer. Its choice of technology solution will be less sensitive to lead time. Figure 2.28 shows that a small company prefers low unit cost solution and a large company prefers short lead time solution.

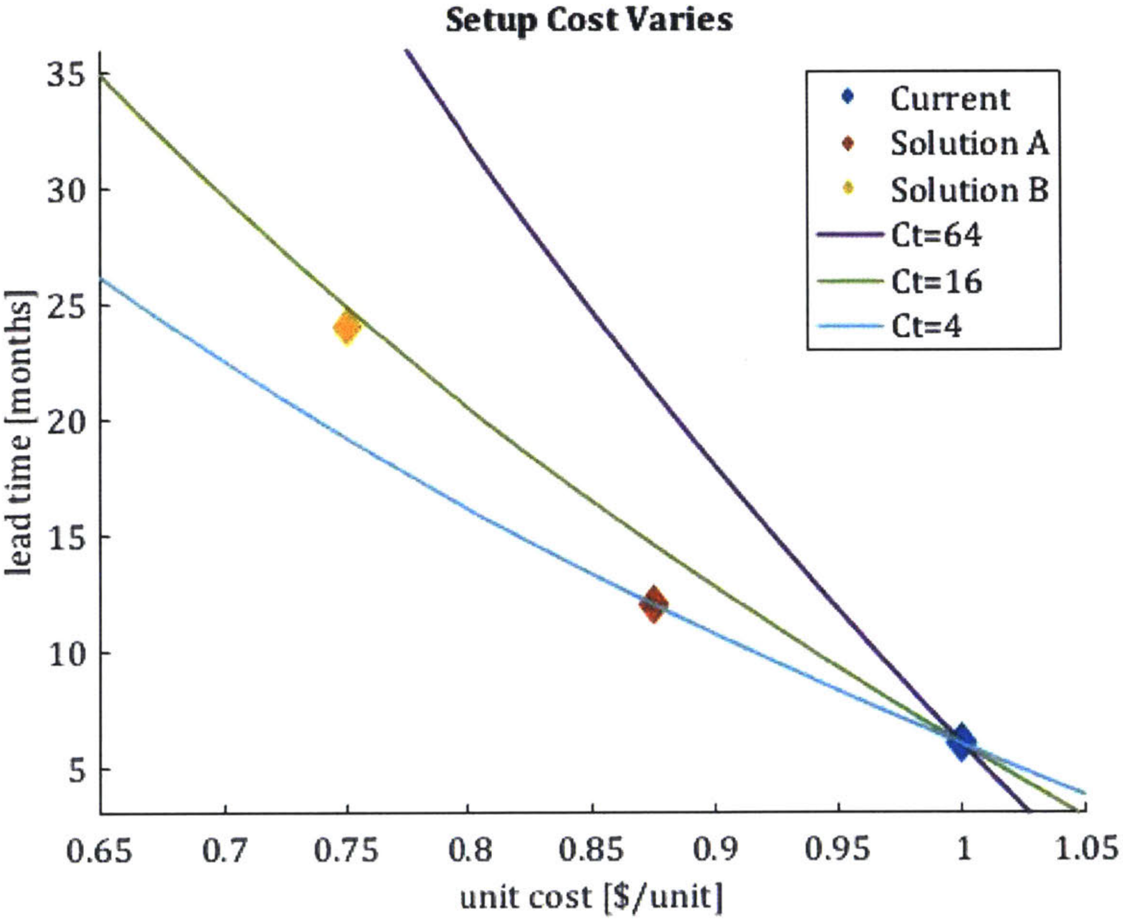


Figure 2.28 Total cost indifference curves with different ordering cost

Chapter 6
Conclusion and Future Work

In late 1980s, AT&T's Transmission Systems Business Unit (TSBU) were facing slow new product introduction although $400 million out of $3 billion revenue was spent on research and development. The key problem was because different product areas communicated infrequently and projects tended to be run as fiefdoms. To overcome the fiefdom mentality, TSBU developed the APEX (Achieving Process EXcellence) process management structure to have experts from different teams to sit down together and synthesize disjointed, and sometimes opposing views from different teams. Within three years, the new product introduction interval was shortened from 39 months to 19 months.

That example shows the power of synergy when we look at a problem from the view of system. Optical transceivers are also components of a system. People who purchase optical transceiver are not mass consumers but companies such as network system vendors, telecommunication service vendors, cloud computing platform service vendors, Internet service vendors et al. Optical transceivers are industrial goods. In front of mass consumers, the value of optical transceivers is intangible. The value of optical transceivers, melted with the contributions from network switch, CPU, memory, computer architecture design and software codes, is represented through web-page response time, file downloading speed, computation capacity et al. The improvement in optical transceivers is especially important when such improvement can improve the consumer-facing performance of the entire system. Therefore, when we evaluate those market opportunities for optical transceivers, we had better view the value chain from optical transceiver device until the data center as a system.

One future work is to apply the framework in the thesis and make projections about which will be the winner designs for next-generation (800G) intra-data center optical transceivers. Previously (<=10G), Transceiver design 1 was on the market. Currently (40G, 100G), Transceiver design 1, 2,3 are on the market. In the future (200G, 400G, 800G, 1.6T), what will happen? Will it be one platform for all the future speeds for the benefit of scalability?

Another future work direction is to quantify the impact of standardization. In terms of benefits, standardization may help reduce product development cost & time through synchronizing chip design and packaging design, reduce mass production cost through economies of scale, reduce mass production time through reducing material lead time and improving fab throughput, reduce system integration time through synchronizing component and system R&D and reduce system integration inventory risk through realizing dual sourcing. On the other side, if standards are made too hastily, the standards may become barriers because they keep creative innovations out of the candidate loop. In addition, the negotiation process of standard-making, which is sometimes political instead of being technical, may take too long. Mapping out those benefits and drawbacks quantitively would advise decision makers about how to balance the trade-off and facilitate the consolidation of disputes among stakeholders.