THOUGHT LEADERSHIP

Intel's Silicon Photonics Products Could Change the World of IT

JUN 2016

Peter Christy, Research Director, Networks

The integration of optical data transmission with silicon integrated circuits has been a long-term research objective for Intel, and the first silicon photonics products are due to be announced in 2016. This report evaluates how this will impact IT as we know it today: What current problems are addressed, and what is the potential for the near future? What is still science fiction?

ABOUT 451 RESEARCH

451 Research is a preeminent information technology research and advisory company. With a core focus on technology innovation and market disruption, we provide essential insight for leaders of the digital economy. More than 100 analysts and consultants deliver that insight via syndicated research, advisory services and live events to over 1,000 client organizations in North America, Europe and around the world. Founded in 2000 and headquartered in New York, 451 Research is a division of The 451 Group.

NEW YORK
20 West 37th Street
3rd Floor
New York, NY 10018
P 212-505-3030
F 212-505-2630

SAN FRANCISCO
140 Geary Street
9th Floor
San Francisco, CA 94108
P 415-989-1555
F 415-989-1558

LONDON
37-41 Gower Street
London, UK WC1E 6HH
P +44 (0)20 7299 7765
F +44 (0)20 7299 7799

BOSTON
1 Liberty Square,
5th Floor
Boston, MA 02109
P 617-261-0699
F 617-261-0688

ABOUT THE AUTHOR

PETER CHRISTY

RESEARCH DIRECTOR, NETWORKS

Peter is the Research Director of 451 Research's Networking Practice. For more than 30 years, Peter has worked with segment leaders in a spectrum of IT and networking technologies. He managed software and system technology for companies including HP, Sun, IBM, Digital Equipment Corp and Apple. Peter was founder and VP of Software at MasPar Computer, a midrange SIMD HPC provider. He was also the Founder and Principal Analyst of the Internet Research Group.

Key Findings

Building on the technology developed during a 15-year research program, Intel is poised to produce silicon photonics electro-optical integrated circuits leveraging its investment, history and leadership in silicon integrated circuits with the potential of important cost reductions and accelerated performance and cost/performance improvements in the future.

Intel's silicon photonics parts will have immediate and material impact on datacenter networking costs just as the optical transceivers are starting to dominate overall network costs.

Silicon photonics parts enable new extended-scale system architectures (ESSAs) that span racks, rows or even the entire datacenter by enabling PCIe-bus interconnects at a distance.

ESSA in turn enables accelerated and more flexible technology use, and higher-performance datacenter-scale applications and storage systems, especially as storage is increasingly semiconductor-based with greatly reduced access latency.

Bringing ESSA to second- and third-tier service providers and enterprises poses additional challenges. Hyperscale adoption of ESSA is enabled by their excellent engineering and operations teams. Bringing the same benefits to most other organizations is challenging because of staffing differences and because the role of traditional system vendors is necessarily weakened.

Executive Summary

INTRODUCTION

2016 will be a notable year for silicon photonics as Intel is expected to launch its optical silicon platform into high-volume production. In this report, we discuss why silicon photonics is a disruptive technology, how it will change IT in the short and mid-term and why the benefit is greatest for the large-scale cloud providers. We center the discussion on Intel's offering and its role in the server and system ecology.

Silicon photonics – the application of silicon wafer planar manufacturing technology to the volume manufacture of electro-optical transceivers – enables a significant cost reduction compared to current manufacturing techniques and promises to accelerate the rate of performance and price/performance improvements as well. The technology will also have a clear impact on large datacenter cost and cost evolution and in enabling new 'rack-scale' system and application architectures.

Commercial use of silicon photonics technology couldn't come at a better time as datacenter networks upgrade capacity to try to keep up with rapidly escalating bandwidth requirements of modern cloud datacenters. In large datacenter networks, the optical transceiver cost already represents the majority of the networking spend (i.e., more than the switches).

Silicon photonics could also revolutionize datacenter system and application design by providing an optical connectivity platform or optical 'backplane' that enables new rack- and datacenter-scale architectures. Extended-scale system architecture (ESSA) is already in use by the hyperscale service operators and contributes to the value of these large services by enabling more flexible technology choice and evolution than is possible with conventional server architectures. Bringing comparable value to enterprises and second- and third-tier service providers is less certain because of the disruption in the traditional IT supply chain it requires.

METHODOLOGY

This report builds on our long-term interest in Intel's silicon photonics efforts, including discussions and ongoing briefings from the Intel rack-scale architecture (RSA) team. Research also included interviews with selected other optical technology vendors (e.g., Infinera, Fiber Mountain, Calient); discussions with hyperscale system operators including Google (Amin Vahdat) and Amazon Web Services (James Hamilton); discussions with Arista Networks (an aggressive driver of high-speed datacenter networking and optical standardization); as well as discussions with Cisco, Dell, Juniper, HPE and other datacenter network equipment vendors.

Reports such as this one represent a holistic perspective on key emerging markets in the enterprise IT space. These markets evolve quickly, though, so 451 Research offers additional services that provide critical marketplace updates. These updated reports and perspectives are presented on a daily basis via the company's core intelligence service, 451 Research Market Insight. Forward-looking M&A analysis and perspectives on strategic acquisitions and the liquidity environment for technology companies are also updated regularly via Market Insight, which is backed by the industry-leading 451 Research M&A KnowledgeBase.

Emerging technologies and markets are also covered in additional 451 Research channels, including Business Applications;

Cloud Transformation; Data Platforms and Analytics; Datacenter Technologies; Development, DevOps and IT Ops; Enterprise Mobility; European Services; Information Security; Internet of Things; Mobile Telecom; Multi-Tenant Datacenters; Networking; Service Providers; Storage; and Systems and Software Infrastructure.

Beyond that, 451 Research has a robust set of quantitative insights covered in products such as Voice of the Connected User Landscape, Voice of the Enterprise, Market Monitor, the M&A KnowledgeBase and the Datacenter KnowledgeBase.

All of these 451 Research services, which are accessible via the web, provide critical and timely analysis specifically focused on the business of enterprise IT innovation.

For more information about 451 Research, please go to: *www.451research.com*.

Table of Contents

1. Silicon Photonics – What Is It, Who Cares and Why?

Broadly, silicon photonics is the adaptation of silicon integrated circuit technology and manufacturing processes to the design and manufacture of electro-optical transceivers to be used for short-haul optical data transport (rack to 'metro' scale).

The excitement over silicon photonics is that it leverages silicon integrated circuit technology, which can be reasonably thought of as the eighth wonder of the world considering its amazing progress over the last 50 years and universal impact on our lives. Soon after transistors started to be manufactured in volume commercially in the 1950s, the idea of manufacturing many at once on a wafer of semiconductor material was conceived (the 'planar' process). Relatively soon thereafter, the idea of building entire electronic circuits at once (fabricating and interconnecting the resistors, capacitors and other circuit elements as well as the transistors) was conceived (what became known as the 'integrated circuit'). In the roughly 50 years since, integrated circuits have progressed from a few transistors in a circuit to billions. Along the way, entire computer processors were fabricated and then improved (made more powerful and more efficient) every year. The result is what we often call 'Moore's Law progress' and has led to the remarkable point where almost everything we use incorporates an integrated circuit processor, and more and more of them are interconnected over the internet.

There has never been a period of sustained innovation like that driven by silicon integrated circuits, and there may never be one again; therefore, being able to leverage what has been developed for silicon integrated circuits is very different from assuming that Moore's Law rates of improvement and investment can be replicated with other technologies. An often overlooked fact is that silicon integrated circuit progress paid its own way with each new generation expanding the market for integrated circuit products enough to more than justify the ever-increasing incremental investment required to drive Moore's Law progress forward. Over the past 50 years, hundreds of billions of dollars have been invested in the ongoing improvements in technology, manufacturing equipment and process (there was nothing easy or straightforward getting to where we are today; Moore's Law isn't a physical law, it's a shared cadence). Silicon photonics, especially as practiced by Intel (a world leader in silicon integrated circuits), is much more than the idea of doing something like building integrated electronic circuits – it's about leveraging the huge investment and remarkable progress that got us to where we are today with silicon integrated circuits.

Many silicon integrated circuits are fabricated at once on a large (300mm in diameter) wafer. The circuit is fabricated in a sequence of steps each done by a tool that acts on the whole wafer. The central process technology is photolithography – projecting patterns onto a layer of photoresist that has been deposited on the wafer – and then performing physical steps based on that pattern. With electronic integrated circuits the result is a circuit formed from with about 10 layers of wiring that interconnect the fabricated circuit elements into, for example, a modern server processor.

Moore's Law progress is marked by the ever smaller size of the transistors that can be fabricated (and the increasing number of transistors on a circuit). Photonic circuits don't depend on such small features. A small ('single mode') laser beam is about one micron across (one millionth of a meter), whereas a small transistor is about .014 micron (14nm), 70 times smaller (Moore's Law progress is usually seen in a real impact – improving the linear dimension by a factor of 70 increases the number of transistors that can be fabricated by a factor of 5,000). Intel's ability to build integrated electro-optical circuits doesn't require the small feature size but does benefit from Intel's sophistication in process technology and chip manufacturing – for example, the precise control in the verticality and roughness of etching done in silicon that is relevant to creating high-yielding and high-efficiency lasers and modulators.

THE IRONY OF USING SILICON

Silicon already plays a critical role in optical data transport because we send optical signals over fiber optic cables; those cables are made of specialty glass, and glass is largely silicon. The exact type of glass and frequencies of light used are chosen because, among other reasons, silicon does not interact with light at those frequencies (light causes the electrons of an atom to move to a higher energy state, and they emit light when they return to a lower energy state – the basics of a laser).

The irony is that silicon circuits can't be used to create or detect light at the frequencies used for transmission on optical fibers because silicon atoms cannot be used to create or detect light of that frequency for precisely the same reason. From that point of view, silicon integrated circuits are exactly the wrong beginning point despite the potential to leverage all the investment. There are a lot of ways to solve these issues but all with clear and substantial challenges and trade-offs. You can make electro-optical integrated circuits based on non-silicon materials (what Infinera has done) but then you give up a lot of the leverage in reuse of existing fabrication technology. You can bond non-silicon components onto the electro-optical circuit but then you risk losing the cost-efficiencies and yield (cost) of the integrated process. A lot of Intel's 15-year exploration and development was to find trade-off points that could leverage its experience and capability in silicon technologies while producing suitable optical parts with high yields.

WHAT HAD TO BE DEVELOPED?

An electro-optical photonic integrated circuit differs from a conventional integrated circuit because it includes electro-optical elements (means of converting electronic signals to and from optical signals) and has the means of carrying optical signals within the part ('light pipes' of some form – optical 'wires' or waveguides). With more complex optical transmission schemes, the electro-optical integrated circuit also needs the means of multiplexing/demultiplexing multiple optical wavelengths into a merged signal that is carried by a single fiber.

Figure 1: Basics of Electro-Optical Integrated Circuits

Source: 451 Research, 2016

	SILICON INTEGRATED CIRCUIT	SILICON ELECTRO-OPTICAL INTEGRATED CIRCUIT	
Base Material	Silicon wafer	(Same)	
Basic Process	Photolithography and lithographically-defined etching and deposition	(Same)	
Basic Component	Transistor		
Specialized Process/ Component		External laser stimulation light source	Silicon hybrid laser
		Embedded silicon laser structure	
		Germanium photo detector	Adaptation of existing germanium tool
		Connection to external optical fiber	
Interconnection	Deposited metal wire	Etched light pipe	
Minimum feature size	~.010 um	~1 um	

Adapting integrated circuit technologies and manufacturing processes to build electro-optical integrated circuits requires the incorporation of non-silicon elements either in the form of separate components that have to be incorporated into a finished module or within the manufacturing process. Adding new elements to an integrated circuit process is very costly and complex, and not taken lightly.

INTEL'S VERSION OF SILICON PHOTONICS AND ITS MOTIVATION

Intel's take on silicon photonics is unique because it is a world leader in silicon integrated circuit technology and manufacturing (the company has unique capabilities as a result), and because it plays a central role in both servers and networking (photonic interconnection is an Intel server strategy element; Intel is uniquely positioned to drive silicon photonics into the server ecology).

The potential value of optical data transport for integrated circuit electronics systems has been known for decades; compared to electrical 'wires,' light can send data faster and over longer distances. Intel's internal efforts started more than 15 years ago, initially to study the use of optical transport on-chip. Over the past 15 years Intel developed important technology (e.g., the wafer-scale integration of hybrid silicon lasers) and found ways to leverage its understanding of silicon process and integrated circuit manufacturing, and the objectives have evolved.

Intel's planned offerings likely include a family of electro-optical circuits that will be packaged both as conventional integrated circuit modules (suitable for integration with a printed circuit board, or PCB) and in form factors that can be used within existing photonic transceiver packages for network switches (e.g., OSPF28). Intel has disclosed more about its technology than other silicon photonics competitors. For that reason and because of its position in the computer industry, this report is Intel-centric when it comes to the technology and use discussions.

2. Optical Data Transmission – What and Why?

Signals within integrated circuits are binary, electronically encoding a 0 or a 1 as a voltage or current. Optical links take a serial electronic data stream and convert it to optical signaling (a modulated light signal) and then back.

The basics of optical data transport involve (1) the kind of laser/cable used, (2) the sophistication of the modulation of the laser light with the digital signal, and (3) whether or not more than one optical signal is carried on a fiber.

Optical transmission methods vary considerably in cost and sophistication. The most sophisticated methods are capable of sending more data over longer links. The choice of method is based on the cost of the fiber link: While an undersea cable can cost up to a billion dollars to install, the cost of the endpoints is unlikely to be significant, and any technology that reliably increases capacity is desirable. In fact, over the last decade a very significant amount of capacity has been added to installed submarine cables by improving the optical transport systems at the endpoints. The economic trade-offs in the datacenter are very different and very important.

Lasers are required for high-performance (rate, distance or both) optical data transport because they create very pure light (frequency/color). There are two categories of lasers used for optical data transport called 'single mode' and 'multi-mode.' Single-mode signals are ideal (purer frequency) but are harder to generate and have a smaller beam size that requires more precise alignment between components (e.g., more precise cable connectors, more precise alignment between the cable and the electro-optical). Single-mode and multi-mode use different cables as well. Single-mode lasers are the primary focus going forward because multi-mode links are limited in capacity and distance, even within the datacenter.

The light signal has to be modulated in some way to encode the digital signal. The simplest form calls for the laser light to be turned on for a 1 and remain off for a 0. At the most sophisticated level, the optical signal uses a complex (multi-level) modulation scheme so that more bits per second can be transmitted. (The most complex and expensive modulation is used for the longest and costliest links like undersea cables.) Datacenter networks will predictably use relatively simple forms of modulation given the trade-offs in transceiver complexity and yield.

Optical cables can carry signals encoded by multiple lasers concurrently if the lasers use different frequencies or colors (what is called wave-division multiplexing, or WDM). WDM schemes also vary in sophistication. The most sophisticated schemes – the most colors, the greatest aggregate capacity on a single fiber – are used in the most expensive long-haul cables where the value of getting more capacity out of an existing cable is highest (dense WDM, or DWDM). DWDM equipment is more expensive (e.g., more precise laser frequency control). Datacenter networking will use simpler WDM schemes (coarse WDM, or CWDM).

Electro-optical circuits that are manufactured by automated, high-yield processes (silicon photonics) will play an increasingly important role in datacenter computing going forward as data rates and link capacities are driven up due to the compounding effects of faster processors and application designs that incorporate more elements that require coordinating communication (what is called 'east-west' traffic).

As rates increase, it is desirable to use more sophisticated forms of optical data transport (more complex modulation, WDM, single-mode lasers) as long as the yield of the manufacturing process doesn't decrease dramatically with complexity, which unfortunately is the case with the legacy transceivers that are assembled from many different components. If the circuit elements used in silicon photonics integrated circuit are good enough (e.g., laser efficiency, detector sensitive) and the manufacturing process yield high enough, the hope is to attain something more like Moore's Law scalability and improvement, which will be critical as data rates continue to increase.

NON-SILICON ELECTRO-OPTICAL TRANSCEIVERS - EARLIER SOLUTIONS

As noted above, silicon circuit devices (transistors) can't be used to create or sense the optical signals sent over (silicon) fiber optic links because silicon is transparent (not interacting) at these wavelengths. There are other elements that can be used that do react atomically at the desired wavelengths. The electro-optical transceivers in the past have typically used these materials, which are sometimes referred to as '3/5' materials because of their place on the periodic table of elements.

Compared to using adapted silicon integrated circuit technologies, the creation of transceivers from composed components has advantages (the ability to use more optimized components) but clear disadvantages (the loss of the sophistication in silicon materials and processes, the complexity of assembly and the diminishing yield as complexity increases). The assembly of the composed module from disparate components is difficult because of the different characteristics of the components and the mechanical complexity of constructing the composed part. The different components have different physical properties (coefficient of expansion) and optical properties (index of refraction) requiring the use of, for example, costly precise mechanical assembly with glue bonding at the boundaries. In the case of single-mode transceivers, precise alignment of the component elements is required because of the small (~1 micron) beam size.

As the transceivers get more sophisticated (e.g., when multi light WDM transceivers are required), the modules become still more elaborate with more elements and the assembly process more complex. With many elements being (literally) glued together, as the complexity of the module goes up the yield goes down (the percentage of finished modules that function correctly) since a failure in any of the components or at any of the boundaries causes the whole module to fail. As the manufacturing yield goes down the cost of the working parts goes up. At some point the diminishing yield negates the efficiency of the increased capacity.

The most cost-competitive (to silicon photonics) module technology adapts the kind of machines developed to manufacture cell phones by the automatic assembly of electronic modules from very small components. Although using such automation increases the complexity of the modules that can be manufactured with adequate yield, industry experts estimate that Intel's initial silicon photonic parts have a roughly five times cost advantage, and if Intel process yields are good, that advantage is bound to increase with more complex transceivers.

INTEL SILICON PHOTONIC PARTS

Intel's silicon photonic parts are built on an existing Intel silicon integrated circuit production line, using the same tools and available process steps configured into a different sequence and recipe. Adapting this existing process takes advantage of the benefits of the huge investment and process sophistication of that mainstream technology. Because it is an older line (photonic circuits don't require as small feature sizes as a state-of-the-art CPUs), the process development and fab construction and fit-out costs (provisioning and commissioning of the fabrication equipment) have already been amortized by the mainstream silicon integrated circuit products. The actuarial production cost for the parts can be based on the cost of running an incremental wafer through the fab, which is considerably less than if that fab had been developed and built for this purpose.

Laser light creation and detection requires non-silicon elements. Adding new elements to an integrated circuit process is very costly because of the potential of disrupting or modifying the existing process by the effect of trace residues of the new element. The alternative of doing processing steps outside the line requires the cost and yield impact of taking the wafers in and out of the fab. Intel was able to find ways to create and detect suitable colors without introducing new elements or running processing steps outside the line.

The company is able to fabricate single-mode (purest light) lasers almost entirely within the integrated silicon part. Whereas other silicon photonics technologies use external laser sources based on 3/5 materials that are precision-attached to the integrated circuit (with the issues of differing physical and optical properties and the requirement for precise alignment discussed above), the Intel process only bonds a non-silicon gain medium (a small chip of a suitable non-silicon material that creates non-laser light) to the silicon integrated circuit and does not require precision alignment or index of refraction matching. The lasing cavity – the physical structure that generates the coherent laser beam – is etched into the silicon substrate with the frequency of the laser determined by the spacing of ridges in the cavity (analogous to an optical grating). The non-silicon light source, which incorporates specifically the 3/5 element gallium that is not part of the integrated circuit

process, is physically passivated (isolated from interacting with the rest of the circuit during fabrication) by an oxidation step (silicon oxide is a non-crystalline glass) and then the passivated light 'bar' is bonded to a matching oxide layer on the silicon wafer, after which the light that stimulates the laser action is generated by passing current through the light bar. The elimination of the use of a separate laser component eliminates cost and yield loss of precise beam alignment and the energy loss at the material boundary. Lower energy loss increases the efficiency of the laser (the percentage of electrical energy converted to laser light energy). The higher the laser efficiency, the more signaling light can be created within the overall power/heat budget of, for example, a switch transceiver module.

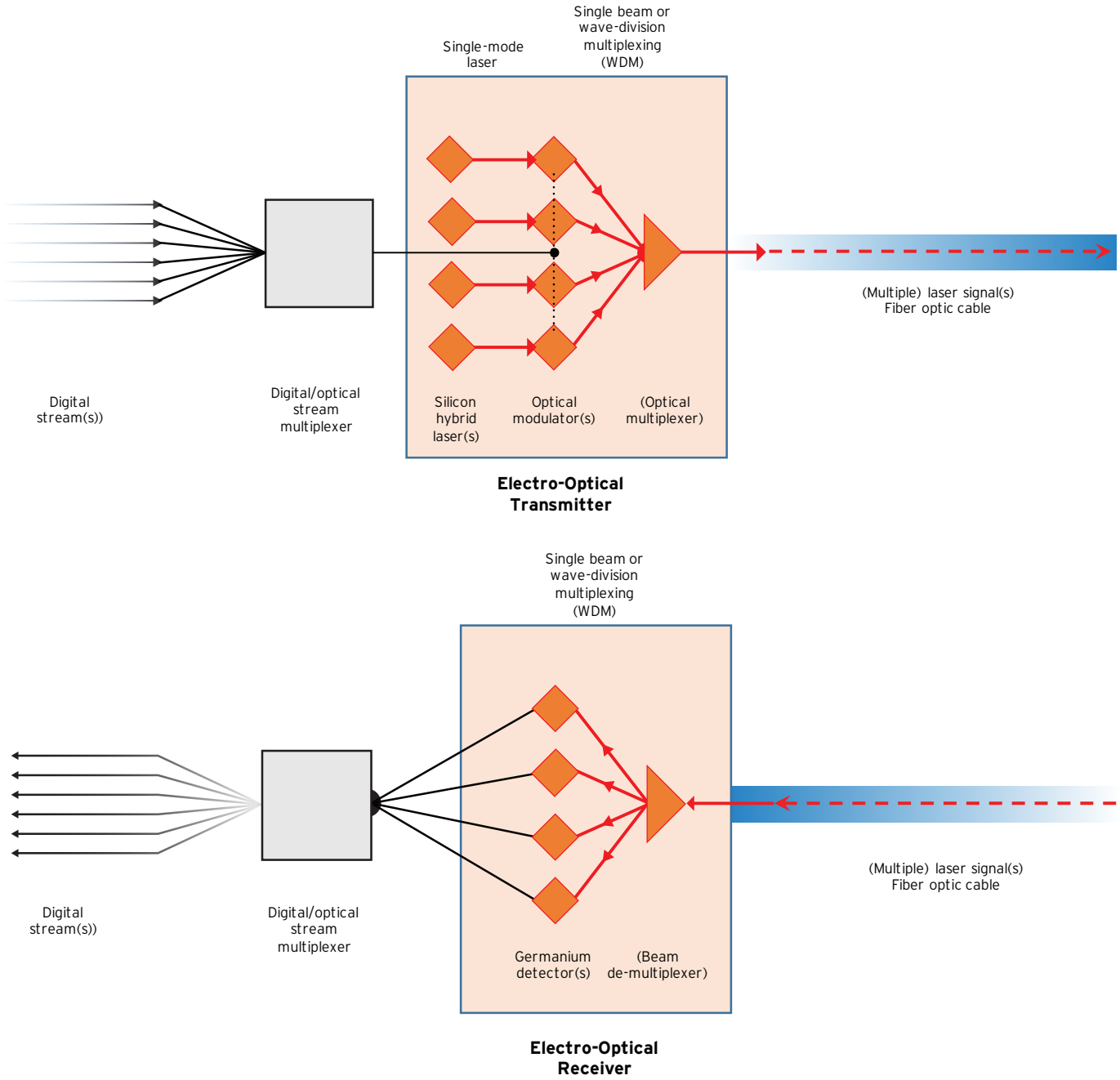
Building adequate optical detectors was simpler. It turns out that germanium, a suitable element, was already in use in the standard integrated circuit process Intel uses to fabricate silicon photonic parts. The process generation used to fabricate silicon photonic parts, and the ones that have followed, fabricates transistors on a germanium-modified silicon crystal matrix ('strained silicon') in order to make the fabricated transistors switch faster. The existing germanium processing step was adapted to create the photo-detectors used in the silicon photonics receivers. A suitable germanium processing tool was available on the line, and the process was already germanium-safe.

In the first generation of Intel silicon photonic parts, transmitter parts and receive parts were manufactured using different process recipes, and on different wafers, but both within the same production fab line.

Electro-optical transmitters or receivers (integrated electro-optical circuits) are created using a silicon planar manufacturing line to create the lasers or detectors, to etch light pipes (the equivalent of wires in a conventional integrated circuit) into the silicon, and optionally multiplex or de-multiplex signals that use different wavelengths and share a single fiber (see Figure 2). As the transmission schemes get more complex (more wavelengths), the fact that the entire transmitter or receiver can be fabricated in an integrated fashion potentially increases the cost advantage of using electro-optical integrated circuits compared to the composed component alternatives as long as the silicon photonics yield is adequate.

Figure 2: Electro-Optical Transmission and Reception

Source: 451 Research, 2016



3. Initial Intel Silicon Photonic Use Cases

There are two distinct (and quite different) important immediate uses for Intel's silicon photonics technology and components: electro-optical transceiver parts for existing datacenter networks, and parts that can be used to create the photonic connectivity for rack-scale ('disaggregated') server systems.

Datacenter switch electro-optical transceivers are packaged into small modules that plug into the switch and connect to the cable. The most popular form factor in use today is QSFP (also referred to as QSFP28 for 100G data rates) – a 8.5mm x 18mm x 72mm physical module with power consumption <3.5 watts. Creating a module with a cross-section as small as possible maximizes the number of ports on the switch front/back but also introduces power and heat limitations since the module area available for cooling is smaller. A typical QSFP module consists of multiple electronic and optical circuit elements interconnected within the module. In the silicon photonics case, there are many fewer individual components and much more of the manufacturing process is automated.

In the next few years we should see a new form of datacenter switch physical architecture emerge where the electro-optical transceivers are mounted to PCBs within the switch rather than packaged as modules at the cable interface. Arista has already developed switches of this form where a fiber optic header divides the output from a PCB-mounted transceiver to various optical ports on the switch front/back plate. The industry group Consortium for On-Board Optics (COBO) is working to establish standards. Partitioning the transceivers in this way enables optimizing port density while at the same time moving transmission rates up aggressively.

The photonic backplane parts (at least Intel's) are packaged like normal integrated circuits so they can be added to PCBs with existing manufacturing automation, with the notable difference that these parts also connect to an optical cable whereas most integrated circuits only connect to the circuit board.

OPTICAL DATA TRANSPORT - DOUBLE-CLICKING ONE LEVEL DOWN ON DETAILS

Optical data transport works by converting high-speed digital signals to modulated laser light and then sending those signals over fiber optic cables at distances ranging from tens of meters to thousands of meters, where a receiving transceiver converts them back to electric signals. Optical data transport has high value as signaling rates increase because there are hard limits on how fast and how far you can send data over copper wires (or traces on PCBs). Fiber optics has been used in networks for the last 40 years starting with parts of the telephone system. Modern optical data transport technology can send hundreds of Gbps data streams hundreds or thousands of miles over a single fiber. The internet at large is built on fiber optic links.

The demands on optical data transport technology – transceivers and cables – vary greatly depending on the use. For the undersea transoceanic that link the internet, any improvement in transceiver performance that increases the usable capacity of existing cables is affordable in the context of the $100m cost of the cable. In a modern cloud datacenter where the electro-optical transceivers already dominate the lifecycle costs of the network, the performance/cost tradeoffs are very different.

To date most of the development and investment in optical data transport has been at the scale of metro Ethernet (kilometer to tens of kilometer distances) and longer. As IT services are increasingly delivered from large-scale datacenters, the datacenter becomes a key market for optical data transport vendors and requirements deemphasize performance at a distance and focus more cost at volume and system integration. As it turns out, the performance and cost-efficiency of optical datacenters is also an important element in the continuing evolution and cost-efficacy of these cloud services as well. The needs of the datacenter drive much of the silicon photonics effort – certainly Intel's because the technical requirements are less demanding than for long-haul, and the cost-reduction more relevant at the system level.

DATACENTER NETWORKING APPLICATIONS OF SILICON PHOTONICS

The role of datacenters in IT has evolved rapidly over the last 15 years starting with the consolidation of branch office applications to datacenters after the dot-com bubble burst and enterprise IT focus returned again to cost-efficacy. Consolidation of smaller datacenters into fewer, larger datacenters to improve the economics of scale (especially operations and support) followed. In the same time frame, web-based service providers such as Google, Yahoo and Microsoft started to develop very large-scale web services.

The last 5-10 years have seen a parallel revolution in application design and use triggered by the success of the smartphone, the rising use of wireless mobile devices in enterprise IT, and the the emergence of elastic platform systems such as Amazon's AWS and Microsoft Azure that enable the server systems in back of mobile applications to be developed and deployed with minimal up-front capital investment. The mobile/cloud revolution is profoundly and rapidly changing the landscape of enterprise IT. The impact of these remarkable shifts can be seen clearly in the server CPU shipment data that Intel reports. Intel says that 2016 will be the year in which more CPUs are sold to cloud service providers – datacenter systems that serve as the basis of a service business rather than serve the needs of a particular enterprise – than directly to enterprises. Additionally, Intel says that more than 50% of the CPUs sold to cloud service providers go to the seven largest web system operators, all of which depend on the construction and operation of large-scale datacenters. The importance of optical data transport within these large-scale service datacenters, and the importance of these large-scale services to modern IT, is unarguable.

DATACENTER NETWORKING

The impact of silicon photonics for datacenter networking is relatively straightforward and directly relevant to the cost of those networks: Electrical/optical transceivers are becoming (in some cases have become) the dominating cost in the network, so the very significant cost reduction possible with high-volume silicon photonic parts is directly relevant to the cost and continuing cost/performance improvements of those networks.

Analyzing the impact of datacenter networking cost is complex and the issues change over time. The network is only a small percentage (~10%) of the overall cost of the datacenter. Operators of these datacenters work hard to maximize the useful throughput from the entire system investment so cost minimization of the datacenter network isn't by itself a goal if it compromises overall system performance (it would be 'penny wise and pound foolish'). What is changing now is the rapid growth in bandwidth requirements and the fact that optical transceivers have become the largest cost component. The fact that transceiver cost-performance progress historically has been much slower than what we think of as Moore's Law progress compounds this. Silicon photonic technology holds the potential of both reducing current transceiver costs and, more importantly, accelerating the ongoing rate of transceiver cost performance improvement.

Datacenter networks have evolved greatly both in the role networks play in IT as well as performance over the last 10 years. A decade ago, most of the datacenter network traffic was connecting the applications to the users and data sources (what is called 'north-south' traffic). More recently there has been a rapid increase in traffic within the datacenter interconnecting the various application elements and services that all reside within the datacenter (east-west traffic). At the same time, server networking capacity requirements have continued to grow with Moore's Law progress. System design heuristics such as Amdahl's Law estimate network capacity requirements grow more or less linearly with CPU power. Finally, during this period server virtualization has been broadly adopted, which adds network traffic as well. The net result in these very large datacenters has been network traffic growth rates considerably faster than Moore's Law.

Datacenter networks typically consist of one or more top-of-rack (ToR) switches (two for network redundancy) for each rack of servers, and then a hierarchy of one or more additional 'tiers' of networking that serve to interconnect the ToR switches to one another and to various external networks and the internet. Ten years ago, typical server-to-ToR connections were 1Gb Ethernet copper. Today, 10Gb is typical, with 25 and 40 starting to be used. The connection from the ToR to the next tier of switches is necessarily at a higher rate, so the switch-switch links that used to be 40Gbps are now upgrading to 100Gbps.

Ethernet cables (Cat 5/6) work well up to 5Gbps. 1G Ethernet server-to-ToR connections were just Ethernet cables. At 10Gbps, server-to-ToR switch connections today are still copper but built on a length-specific basis. Copper connections can be used pragmatically up to 25Gbps at rack-scale lengths but beyond 25Gbps or for lengths beyond a few meters optical data transport is required.

A ROADMAP FOR SILICON-PHOTONICS IN DATACENTER NETWORKS

In the first (current) phase, silicon integrated circuit technology has been adapted to enable the manufacture of electro-optical transmitters and receivers that seem likely to significantly reduce the cost and have the potential for supporting more sophisticated transmission schemes and for evolving faster with improved cost-efficiency. Today Intel leverages existing fab capabilities but builds transmitters and receivers on separate wafers (using different variations of the standard production recipe). In this first phase the impact will be largely in the form of pluggable optics (such as QSPF modules).

In the second phase the electro-optical parts will be mounted on PCBs within the switch and connected to the cables via fiber optic headers (e.g., the COBO effort described later). Moving the parts to the PCB enables higher port densities at increased transmission rates among other improvements.

In the longer term we can expect multi-chip modules that integrate electro-optical devices and switch ASIC improving the bandwidth between the parts and power efficiency.

The vision that Intel started with – integrating the electro-optical and switching logic into a single chip – still seems a long way off given the different processes used and the many engineering challenges, such as thermal interactions.

DATACENTER NETWORK ECONOMICS

Datacenter network costs have to be analyzed as depreciation (rather than purchase price) because different elements of the datacenter network are refreshed at different rates. Fiber optic cables have long life and may be amortized over up to 30 years. Switches are typically amortized over five to seven years. Electro-optical transceivers turn over most rapidly and are often amortized in as little as three years because of the need for new and faster links. When the costs are translated into yearly deprecation, the transceivers already dominate in some of the largest and highest-capacity networks.

Fiber optic cable economics are relatively static. Switches improve with Moore's Law progress. Electro-optical transceivers that are manufactured by existing methods have a hard time keeping up or, as we discussed above, may even degrade as measured in cost per bit at higher rates.

THE IMPACT OF SILICON PHOTONICS ON APPLICATION AND SYSTEM ARCHITECTURE - EXTENDED-SCALE SYSTEM ARCHITECTURE

Silicon photonics will also enable important new system and application architectures by enabling composed architectures (tightly integrated CPU/memory/storage) that span racks or an entire datacenter, not just an individual server's chassis. In traditional servers, the local storage and networking peripherals are integrated with the CPU and memory over what is called the PCIe bus, a high-bandwidth, low-latency PCB interconnection. The physical limitations of PCB signaling limit the length of a PCIe bus to just a few inches.

The core idea of a photonic PCIe backplane is to create a protocol (a specification) that defines how the PCIe bus signals are serialized and then transmitted over a fiber optic cable. Matching today's PCIe capacity requires about 60Gbps of data transport capacity, well within the 100Gbps capabilities of a current datacenter fiber optic technology. In many cases, we also want to switch the signals between different system and storage endpoints (e.g., move storage from one server to another), which means other complexities have to be implemented in the protocol because a simple PCIe bus has some amount of shared endpoint state that has to be managed.

In order to maximize the utility of a photonic backplane, we may want to be able to share the link with other protocols including Ethernet and existing serial storage protocols. A photonic backplane is a shared, switched fiber optic-based data network that replaces many of the cables used at the system or rack level today, and greatly increases the potential span (separation) of a server system.

A new form of system architecture is enabled by extending the reach of the PCIe bus – for example, what Intel calls RSA when applied to systems that fit within a single rack (where photonic integration is not required). RSA is also the structural basis of the Open Compute Project (OCP) started by Facebook. The basic idea of RSA is the disaggregation of the server into its constituent parts (CPU/memory, storage, networking) and the ability to select and change these different elements independently rather than managing them as part of an integrated system (what we think of as a server system today).

At the equipment rack, pod (multiple racks) or datacenter scale, RSA means the installable ('rackable') unit is no longer the individual, self-powered, server box but rather 'trays' of CPU/memory, storage or possibly networking, along with suitable modular power supplies.

THE MANY POTENTIAL BENEFITS OF RSA AND ESSA

The large-scale web operators have been experimenting with some form of server disaggregation for quite some time. For example, the earliest large Google datacenters were built with custom racking of powered server motherboards without any traditional server box. The economic clout and internal engineering of these 'hyperscale' services have dramatically redefined the system supply chain. The largest services operate at a scale where they increasingly buy their hardware directly from the ODM manufacturers that previously built servers for vendors such as IBM, HP and Dell (and switches for Cisco). The largest-scale service operators have adequate design competence and specific operational needs to make designing their own system components and contracting directly with the ODMs directly both possible and desirable, disintermediating the traditional server suppliers from the process. The largest of these web operators also create and maintain their own software so they can accelerate system innovation (e.g., the use of RSA or ESSA) because they are less dependent on external vendors and product ecologies to adopt the concepts (a critical issue when considering when and how the enterprise and second- and third-tier service operators will get RSA or ESSA systems).

The largest datacenters design systems and applications from the overall datacenter (layout, power, cooling) down to specifying custom variations on Intel datacenter CPUs, and everything in between. As noted above, these large-scale operators also create and own their own system and application software; as a result, they are free to move forward as they please and do it rapidly (minimal delay between the emergence and exploitation of new technology), which is unprecedented.

There are many potential benefits to RSA in terms of application and operational potential, including the ability to:

- **Optimize processor choice:** Processors can be flexibly selected or evolved based on (changing) application needs. Server CPUs range broadly in many dimensions – number of cores, speed of the core, power/efficiency of the core, and then when put into an ensemble (e.g., a rack or part of a rack), in density of compute power and the aggregate power consumption and cooling required. In an RSA, different parts of the datacenter can use different CPUs and those choices can evolve fluidly with the needs of the application system and use.
- **Scale and evolve compute power and storage independently:** A rack system can be quickly upgraded with a new generation of servers without any changes to the storage configuration, for example, while still getting the benefits of directly attached storage (rather than network-attached).
- **Independently refresh server technologies:** For example, large-scale system operators often choose the fastest processor available and aggressively move to new generations when available. Having processor trays as a separate RSA element facilitates this.
- **Share storage flexibly while optimizing performance:** Prior to RSA, storage (e.g., disk drives) could be attached to specific servers (DAS) or assembled into a storage subsystem (e.g., NAS or SAN). With RSA, these configurations can be made dynamically and changed. Increasingly, datacenter storage utilizes DAS architectures because the direct-connect performance (the rate at which the drive transfers data to the server memory) far exceeds what can be achieved over the network. With a photonic backplane we can have the best of both worlds – directly connected performance similar to DAS and network access from a distance. ESSA based on photonic integration expands the capability to share storage in this way to row and datacenter scale.
- **Create and evolve new forms of high-performance computing systems:** For example, with Hadoop big data processing, the storage is attached to individual servers and the compute is brought to the data, in contrast to earlier architectures where the data was brought to the computer. With RSA, yet another variant can be created where the data is switched between various compute clusters for different application phases (using photonic transport to move the data to the computing again), hence getting the best of both worlds.
- **Optimize flash memory storage at datacenter scale (ESSA):** Modern flash memory (e.g., Intel's forthcoming 3D XPoint memory) directly connected to a server (via PCIe) has access latency far lower than when packaged and used as an SSD. Using a switched PCIe bus in turn provides lower latency to remote flash memory than sending it over the conventional network.

The use of switched remote PCIe access makes a material difference in the ability to exploit the speed of flash in a storage system that spans the datacenter. Full support of 3D XPoint memory in Intel Xeon servers (the integrated use of DRAM as a write-back cache) is due in 2017, giving some time for Intel's RSA team to flesh out plans for how it can be supported beyond a single rack.

- **Maximize the productive work of a datacenter under peak load (ESSA):** RSA enables the greatest possible flexibility in exploiting the available compute and storage assets available at any moment in a datacenter regardless of location within the datacenter, thereby maximizing throughput and simplifying operations. The peak useful work possible from the datacenter is an important metric and contributor to the cost/performance of the service.

THE ADDITIONAL BENEFITS (AND CHALLENGES) OF AN OPEN PRODUCT ECOLOGY FOR RSA AND ESSA

For the largest service operators, the opportunity provided by RSA is multiplied by their ability to commission system elements from a diverse set of ODM and component suppliers and exploit this diverse and competitive ecology to accelerate innovation even further, assembling their RSA and ESSA web systems from many vendors.

Facebook created the OCP in part to jump-start the same ecology for smaller users of RSA (enterprises and second and third tier service operators). It established an open forum where the standards needed to build this kind of interchanging parts vendor ecology could be created.

Making an open ecology for rack-scale parts isn't simple. In the existing IT system model, for most customers the large system suppliers (e.g., IBM, Dell, HP) perform many essential functions by creating purchasable systems for which the compatibility of hardware elements and popular software systems has been designed, tested and assured by the system vendor. In the OCP model, other means must be found to assure that common software stacks work reliably with RSA systems assembled from components selected by the customer, without assuming that the customer has the engineering competence to do it themselves like the large service operators of legacy system vendors (which few do).

Until such an ecology exists, enterprises and smaller service providers will probably have to use RSA in a hybrid form, leaving a system vendor (most likely HP, IBM or Dell) in place to provide a specific set of system configurations that have been tested with specific widely used software systems (e.g., OpenStack, Windows Server, VMware vCloud). The full impact of those limitations – e.g., constrained system alternatives, limited software offerings, greater innovation delay – will only be clear when the constrained systems are brought to market which is still in the future.

INTEL'S ROLE IN RACK-SCALE ARCHITECTURE

Intel plays a unique role in the development and promulgation of RSA (and ESSA in the future). The company is a core supplier in today's server market manufacturing nearly all the server CPUs purchased, and has been at the center of the server ecology for many years as a result.

Intel has always been an architecture innovator and evangelist for the PC and server ecologies for which it provides CPUs (and other semiconductor parts). For several years it has discussed RSA, in part as response to and participation in the OCP, and in part as a result of the company's long investment in its silicon photonics technology and product development.

Intel's efforts in RSA parallel many of its earlier advanced architecture initiatives: lab research, then vision and roadmap, early prototypes and demonstrations, nourishment and sometimes investment in other parts of the ecology, joint efforts with other ecology participants, and finally customer evangelism and demand creation.

Intel's interests are direct – namely, create demand for silicon photonic parts and assure that datacenter networking limitations don't diminish the value of new and more powerful servers. They are also indirect, as evidenced by its efforts to create understanding of and demand for the advanced systems RSA enables, and the benefits they provide to the end user (e.g., rapid technology innovation, easy-to-use form of computing).

However, RSA is not something Intel can drive by itself; it requires the coordinated and, to an unprecedented degree, cooperative effort of many different potential hardware and software suppliers if its benefits are to be brought to the largest potential market and have the greatest possible benefit.

4. The Rack-Scale and Extended-Scale Conundrum – The Legacy IT Supply Chain Doesn't Help

For many, the most important issue in modern IT is the evolution of enterprise IT to the use of the public cloud driven by the importance of business agility (the ability for a business to adapt quickly to changing opportunities and conditions) and the demonstrated value of the public cloud as an agile IT platform that enables business agility rather than limiting it (as has too often been the case for legacy enterprise IT architectures).

For many reasons, most enterprises would prefer private cloud solutions to public cloud services as long as the private offering has comparable functionality and cost. Because of the cost benefits of RSA as demonstrated in large-scale cloud services – specifically rapid and flexible exploitation of technology advances – it would seem that RSA systems are needed by enterprises as well. However, unless something changes RSA will take a long time to impact enterprises.

The visible challenge is what we call the existing IT innovation supply chain that was created initially for the PC business. For the enterprise, fundamental IT innovation occurs up the supply chain, whereas with the large-scale web operators it occurs with the service itself rather than being dependent on suppliers. For example, if Intel creates a new but different CPU or system architecture, it can't be exploited by enterprises until new applications that take advantage of the innovation are available. The new applications can't be developed until the operating system and system services are available that manage and exploit the technology. The new operating system and services can't be built until new hardware platforms have been developed and made available to the system developers. With big changes, in the legacy innovation supply chain, this all gets serialized and takes a long time (years).

The existing supply chain worked perfectly for the PC industry (and later the server industry) because the most important value was software compatibility, not software and system innovation. When new PCs came out existing software continued to work. New CPUs came to market quickly because they didn't require significantly new operating systems or services.

RSA will take a long time to move through the legacy innovation supply chain, especially given that it fundamentally changes the roles of the vendors, moving system assembly and test away from the server vendor and putting it into the hands of the customer. This is the price that comes with the flexibility of choosing offerings from an ecology of compatible products from multiple vendors.

The public cloud has changed the innovation model in key ways. The largest providers have developed unprecedented internal engineering and operational competence. As a result, these largest-scale providers respond to changing market needs (adding functionality) and technological opportunity (e.g., flash memory, incorporation of GPUs for computation, photonic datacenter scale integration) with unheard of speed because the engineering and software integration are done in-house with much less dependence on other vendors in the supply ecology.

Finding a way of accelerating the innovation supply chain for products delivered to the large potential enterprise and service markets that don't and can't have the internal engineering and operational competences of the hyperscale providers is essential if these customers are going to have the full benefits of RSA. How to accomplish this isn't simple (Intel has developed reference architectures, is working to bring software and hardware vendors along, and working with diverse industry efforts including Project Scorpio, the OCP, Redfish and the Scalable Platforms Management Forum).

The *Ericsson HDS 8000* is an interesting attack on this problem of bringing full ESSA benefits to a second-tier market (mobile service operators). Ericsson has taken responsibility for the overall system design and implementation soup to nuts (or photons to applications) in a way that parallels the large-scale providers, rather than waiting for the innovation to be developed by and flow through the legacy supply and innovation chain. As a result, Ericsson is bringing these innovations to its specific customers (the mobile service operators with which Ericsson has high share), in what seems to be years in advance of when traditional enterprise customers will get them from the legacy suppliers.

The struggles of the traditional suppliers can be seen in the OCP and the related ecology. OCP imagines a new world of RSA with the system components manufactured by a diverse set of vendors (today's IT systems vendors, some of the ODMs by themselves) where the customer (an enterprise, a second- or third-tier service provider) can buy system components from this set of vendors and integrate them into a purpose-built solution optimal for their needs, believing (correctly) that unless some ecology transformation like this occurs, the smaller enterprises and service providers can't possibly keep up with the hyperscale service providers. But none of these customers have comparable internal engineering and operational resources, so the product solutions have to be 'better' (as in, less dependent on human excellence) and means must be found to replace the integration and assurance services traditionally performed by the legacy system vendors.

The open issues in OCP start with detailing what it takes to make this concept work. One interim solution would be for the traditional system vendors (e.g., Dell, HP) to provide supported rack-scale systems with specific software, all pre-tested and assured to work together. This clearly will work and solves a customer problem but in the process cleaves away much of the potential that comes with the open vendor ecology and customer freedom to use diverse software.

5. Conclusions and Recommendations

Intel's adaptation of its industry-leading silicon integrated circuit process and manufacturing technologies to the manufacturing of electro-optical transceivers has matured just in time to play an important role in datacenter networks and in enabling valuable new forms of rack-scale datacenter systems. These benefits will clearly be reaped quickly by the hyperscale web service operators. How they reach enterprises and smaller service operators is less clear and another factor impacting the rate and scale of public cloud adoption. The success of efforts like the OCP in creating a vibrant RSA commercial ecology will strongly influence the broader future success of and exploitation of photonic system integration. Silicon photonics components enable important new system and application architectures and at the same time highlight the complexity and key issues for all those that compete in one way or another with the hyperscale service providers.

6. Further Reading

Ericsson partners with AWS, Quanta to strengthen and accelerate its Hyperscale System, March 2016

Ericsson lights the way with the first commercial, rack-scale optical backbone offering, June 2015

Emerging technologies in the physical datacenter – Part one, October 2015

Infinera jumps into metro Ethernet with guns blazing, October 2015

Intel delays commercial introduction of much-anticipated silicon photonics components, February 2015

Intel Silicon Photonics a year later – soon to be juicing a datacenter near you, July 2014

Fiber Mountain shows that a glass ceiling in the datacenter isn't all bad, August 2015

CALIENT uses smoke and mirrors to keep elephants from bothering mice in the datacenter, December 2013

7. Index of Companies