SAILING THROUGH THE DATA DELUGE WITH PERVERSIVE OPTICAL CONNECTIVITY
Abstract
Keeping up with growth in global bandwidth demand requires coping with the stress this growth exerts along the dimensions of cost, density (footprint) and power. The total amount of dollars, cubic meters, and Watts available for a technology upgrade cycle of a given IT facility remains fairly constant, whereas the amount of compute, storage, and networking capacity that must fit within those envelopes increases exponentially. Technology improvements must continue to make up for this discrepancy.

Processors, memories, and storage devices are still predominantly electronic, whereas communication has become predominantly optical. We can picture a perimeter around the electronic domain demarcating the border between electrical and optical communication. Inside the perimeter, electrical interconnects are the most cost-efficient solution, whereas outside optics rule. This perimeter is progressively shrinking because electrical signaling rates are climbing to meet bandwidth demands. Simultaneously, bandwidth density requirements are increasing because of increases in the speed and number of high-speed serial links per unit of area or volume. The specific challenge addressed in this paper is that optical communication needs to close these gaps, but today’s modus operandi of manufacturing and deploying optics is ill suited to the task.

The ever-tighter planar integration of photonics devices into photonic integrated circuits and the integration of electronic and photonic circuits inside a single package are the emerging technologies to meet this challenge. These technologies enable hybrid opto-electronic devices we refer to as optoASICs. They will enable pervasive optical connectivity across a wide range of applications, increasing bandwidth density, lowering energy consumption, and driving down cost with respect to conventional optical links. Moreover, they will unlock the potential for broad system-level innovation by tearing down the bandwidth-distance walls that limit architectural choices.

Voracious Global Appetite for Bandwidth
Demand for communication bandwidth keeps growing at an exponential pace across the globe, fueled by surging numbers of connected devices, increasingly bandwidth-intensive consumer applications (HD video, gaming), and a proliferation of “machine-to-machine” traffic owing to the Internet of Things and vehicular communications [1][2]. On the enterprise front, continuing migration to “serverless” cloud-hosted IT is adding further demand.

Inside data centers, incoming traffic volume gets multiplied, as each incoming request can trigger a flurry of internal server-server traffic [3]. Each user-accessible front-end service is typically composed of many separate back-end applications. These applications may each be highly distributed and require a highly connected infrastructure. An example of such a communication-intensive workload is machine learning [4], which has become so important so quickly that an arms race has erupted to build domain-specific hardware [5].

These trends in applications are boosting demand in global connectivity (WAN, MAN), in data center connectivity (LAN, SAN), and in intra-server connectivity (system buses including I/O, memory, multi-processor, accelerator, etc.). To meet this demand, communication data rates need to increase accordingly. We can increase the number of physical connections (“channels”), increase the data rate of each connection, or do both.

Electrical Bandwidth Scaling Out of Steam
We have now gotten to a point where further bandwidth scaling in the electrical domain is becoming increasingly difficult. Scaling along the dimension of parallelism means increasing the number of channels, which in turn means that more pins on a chip’s package need to be allocated to high-speed electrical I/O. Such increase is constrained by practical limits on chip package size, mostly because of package warpage when soldering it down onto a printed circuit board. Power is another important constraint: High-speed electrical I/O consumes a substantial amount of power. As the fraction of total power consumed by the I/O increases, less remains for the chip’s other functions. Higher total power also means that more chip pins need to be allocated to power and ground and, more critically, that the thermal system budget is much harder to meet.

Figure 1: Reach of twin-ax copper cables versus data rate per channel, as specified by IEEE 802.3 Ethernet standards.
Scaling along the data rate dimension is also running into limitations because of the frequency-dependent signal losses in an electrical channel. These consist of conductor losses caused by the skin effect of rough conductor surfaces and dielectric losses. Improved dielectric materials and smoother conductor traces help minimize the overall loss, but a system can tolerate only a certain amount of loss to achieve a target bit error ratio (BER). As data rates keep doubling, these losses severely limit the distance that a signal can travel. This trend is reflected, for instance, by decreasing maximum reach for copper twin-axial Ethernet cabling, which has shrunk from 7m for 10Gbps signaling down to 5m for 25Gbps signaling and just 3m for 50Gbps (Figure 1).

One technique to mitigate the signal losses is to apply sophisticated, but power-hungry, equalization techniques to compensate for signal distortion. A complementary approach is to encode more than a single bit in each symbol; this way, more information can be transferred without increasing the symbol rate. For example, PAM-4 signaling encodes two bits per symbol. The trade-off is that the receiver needs to distinguish between four different symbols. This entails an effective decrease in signal-to-noise ratio, because technological limits prevent the signal amplitudes from being increased accordingly, causing a corresponding increase in BER.

An additional technique that has become widespread is Forward Error Correction (FEC), which enables a substantial decrease in BER by encoding redundant information in the transmitted bit stream to enable not only detection but also—to a limited extent—correction of bits received in error. Unfortunately, FEC incurs substantial overhead in terms of chip area (cost, power) and especially latency.

The limited reach of electrical links is starting to constrain system designs. In the context of datacenter networks, one example is the number of servers that can be connected to a Top-of-Rack switch, which is now limited by copper reach rather than switch capacity. A single high-end Ethernet switch ASIC with 256 channels at 50Gbps per channel could connect several racks worth of servers, but because copper reach is limited to 3m, this is no longer feasible. The consequence is a network architecture that is sub-optimal in terms of cost as well as performance.

Optics Closing In
The trends described above imply that there exists a crossover point in terms of bandwidth-distance product beyond which electrical links get replaced by optical links. The latter inherently—owing to the physics of photons versus electrons—support a much larger bandwidth-distance product [6]. Historically, this crossover point has hovered around 100Gbps*m [7], suggesting that as link speeds move to 100Gbps, their reach will not be more than 1m.

Optical communication solves the issue of limited reach. It also offers much higher bandwidth density, because fibers are much thinner (and lighter) than copper cables, and because optics can exploit the wavelength dimension by using multiple carriers at different frequencies to simultaneously transport many channels across a single fiber pair.

Optics have been moving from the outside in and must continue to do so as electrical reach continues to shrink. This trend can’t progress without cost-effective, high-density, low-power optical connectivity. But right now, optical link technology does not yet tick all these boxes. What are the hurdles and how do we overcome them?

Pros and Cons of Modular Optics
The present model for deploying optical links is largely based on modular faceplate-mounted transceivers. The photonic devices (lasers, modulators, detectors, filters, etc.) and associated electronic circuitry are assembled and packaged in pluggable modules providing a limited number of channels. These modules conform to certain standards or multi-source agreements (MSAs) in terms of their physical form factor (e.g. QSFP-DD, OSFP), their high-speed electrical (e.g., IEEE 802.3 400GAEUI-B) and optical interfaces (e.g., IEEE 802.3 400GBASE-DR4), and their management interface (e.g. SFF-8636). The following factors have contributed to the success of pluggable optical modules:

- **Pay as you grow**: Added capacity can be deployed when needed, instead of having to deploy and pay for all optical ports up front.
- **Flexibility**: Network operators can choose the best-fit per-port option in terms of requirements versus cost based on their deployment scenario.
- **Serviceability**: When a pluggable module transceiver fails or an upgrade is desired, it can be replaced in the field.
- **Multi-vendor**: Standardized form factors and interfaces created a healthy ecosystem with a multitude of vendors for each type of module, avoiding “single-source” dependencies.
This deployment model also has its share of disadvantages.

- **Density:** Bandwidth density is low, as each module supports from one to a maximum of eight channels. This is primarily determined by the maximum number of electrical channels that can be grouped together into a given form factor.

- **Cost:** The cost of optical modules (per capacity) is much higher than that of electrical links. These modules are complicated micro-optical systems, assembled from many discrete components. Assembly and packaging account for a large chunk of the total costs. Low integration density (Figure 2) implies that these costs are not amortized over a sufficient number of optical channels. Also, single-mode optics are considerably more expensive than multi-mode optics, which are only suitable for ranges up to about 100m.

- **Power:** Modular deployment adds two chip-to-module (C2M) electrical links to the system: one from the host ASIC to the module on the transmit side and one from the module to the host ASIC on the receive side (Figure 3). These links cause much more signal degradation than the optical link by itself. As data rates increase, these channels need sophisticated transmit and receive equalization. If the distance between host ASIC and transceiver becomes too large, additional discrete retimer devices are required to meet the link budget. These add significant cost and power overhead.

- **Thermal management:** Transceiver modules are mounted such that they cover almost the entire faceplate, e.g. in a 1RU rack-mounted Ethernet switch. The modules dissipate a substantial amount of heat while obstructing the airflow, necessitating the use of powerful fans. This setup drastically increases the air temperature for the host ASIC, posing a major thermal challenge for systems with ≥25Tbps of capacity.

To highlight the penalty associated with the chip-to-module channels, note that the end-to-end link budget is a very lopsided affair: At 100Gbps per channel, the optical part of the link, spanning from 500m to 2km, has a typical loss budget of 3-4dB. The electrical channels to cover a distance of less than 10” on either side between host ASIC and optical module incur a loss anywhere from 12 to 16dB (IEEE 802.3ck).

The cost issue provides a substantial economic incentive to minimize long-reach capacity, which manifests itself in two ways. The first approach is to deploy less long-reach capacity altogether. Oversubscription is common practice in data center networks, where capacity at subsequent network tiers is reduced to compensate for the difference in cost between downlinks and uplinks. The alternative is to shrink the physical distances such that electrical reach still suffices for a large share of the connections by packaging nodes much closer together. The former approach entails performance compromises, especially for highly-connected data-intensive workloads such as machine learning. The latter approach entails customized high-density system packaging and thermal management.

**On the Marriage of Electronics and Photonics**

All our data collection, processing, and storage nodes are surrounded by a perimeter within which electrical links are the most cost-efficient solution. We need to deal with the shrinking of this perimeter. Optical communication is the most viable alternative to fill
this void, moving ever closer to the electronic circuits. Aiming to minimize the number of intermediate components between the electrical and optical domains ultimately leads to direct integration of optical and electronic circuits in the same package.

This realization has spawned considerable efforts towards achieving the marriage of electronic and optical circuits. Already three decades ago [8] researchers demonstrated that the basic materials system used to implement highly-integrated electronic circuits, i.e., silicon-based semiconductors (e.g. CMOS), is also suitable for photonic integrated circuits (PICs). This gave rise to silicon photonics [9]. Almost all devices that are necessary for optical connectivity, including waveguides, modulators, photodetectors, wavelength multiplexers and demultiplexers, and fiber couplings, can be manufactured in a silicon photonics platform. As lasers require a different material, e.g. indium phosphide (InP), these need to be manufactured separately and integrated at a later stage, either by directly bonding the laser dies onto the silicon photonic circuit, or by coupling the light in from an external laser via additional “optical power supply” waveguides or fibers.

Silicon photonics has not yet been able to replicate the decades of exponential improvements that large-scale integration has brought to the micro-electronics industry. Optical communication has for a long time been used for long-haul links in the tele-communications industry. However, as channel rates continued to increase, optical links became prevalent in the burgeoning market for inter- and intra-datacenter networks to the extent where now all network links except those to the servers are optical.

Silicon photonics enables the creation of electronic and photonic integrated circuits in the same technology platform. On the surface, this is an alluring proposition. Continuous miniaturization may be expected to bring the same benefits (in terms of density, performance, power, cost) to photonics as it has brought to electronics. Why, then, has silicon photonics so far has failed to deliver on its promise?

**The Importance of Platform Selection**

The choice of silicon photonics platform plays a critical role here, especially with respect to the waveguide dimensions. A dichotomy exists between sub-micron and multi-micron platforms. In an effort to reuse existing capital-intensive CMOS foundry infrastructure, the main established foundries have converged on ~220nm waveguides [10]. Another important reason is that sub-micron waveguides exhibit “pure” single-mode behavior, meaning that the light travels through the waveguide in only one electromagnetic mode. Single-mode optics offer greater range and the ability to perform efficient wavelength division multiplexing (WDM).

However, shrinking the waveguide dimensions incurs trade-offs [11]. Compared to multi-micron waveguides, they exhibit higher propagation loss, have lower power handling capability, and have higher device sensitivity to process variations, for instance in terms of phase error or wavelength registration. Larger waveguides are more amenable to high-throughput integration of other materials systems (e.g., III-V lasers or modulators) due to much lower coupling losses. Another major advantage of large waveguides is that they can form intersections on the die with negligible loss, ensuring very dense optical layouts. Perhaps most importantly, larger waveguides enable easier and more efficient fiber coupling, as the waveguide dimensions are within the same order of magnitude as single-mode fiber’s core diameter. Furthermore, it turns out that multi-micron waveguides can be engineered to exhibit single-mode behavior [12], overcoming the main obstacle.

Some silicon photonics developers have pursued monolithic integration, which entails implementing photonics and electronic circuits in one and the same process. Although this arguably constitutes a “holy grail” objective, there are several important problems. Cutting-edge bulk CMOS processes are not suitable for silicon photonics for a variety of reasons, including the need for silicon-on-insulator (SOI), the etching depth required for large waveguides, incompatible process steps, v-groove etching, etc. Even if they were suitable, it would be costly and challenging to port all of the photonics-specific aspects of the process to each new CMOS node. Moreover, it is questionable whether sacrificing extremely valuable CMOS die area to implement comparatively large photonic devices is an economically sound decision. As photonic devices are not scalable the same way electronics are, they can’t fully benefit from advanced CMOS technology.

Conversely, monolithic integration could be performed on the basis of older technology nodes. This is fine from the perspective of the photonics but will negatively affect the electronic circuits with respect to performance and power consumption. Compound yield is perhaps the biggest challenge, because the tried-and-true ways of compensating...
process variations for electronic circuits do not readily apply to photonic circuits. As yield is a critical factor in overall cost per good die, adding process-variation-sensitive photonics circuits to high-value CMOS dies is likely to drive overall cost up, which runs counter to one of the main objectives.

In-Package Optics: A Complete Optical Connectivity Solution

Taking into account the learnings from over three decades of silicon photonics, Rockley Photonics has developed a complete solution for ubiquitous optical connectivity. This solution is a technology platform that enables co-packaging of third-party digital IP cores with silicon-photonics-based optical I/O. Our In-Package Optics solutions are based on the following guiding principles.

Use the most suitable process technology for each component based on system requirements

A dedicated silicon photonics process in a technology node optimized for photonics, not electronics avoids compromising photonic or electronic performance. Using advanced BiCMOS for analog/mixed-signal drive and receive circuitry optimizes performance and reduces power. Supply optical power either via hybrid integration of III-V materials or through external lasers (as opposed to monolithic integration) to enable using known good, burnt-in lasers, facilitate better thermal laser management, and enable field serviceability.

Adopt a multi-micron silicon photonics platform

Platform functionality and performance: Our platform supports all requisite photonic devices, enables high-density integration of a large number of components, provides high-speed, compact, low-power modulator and detector technologies, and tight waveguide bends. It also provides single-mode waveguides, offers low optical loss, high power handling, low sensitivity to process variations, and a broad (>300nm) operating wavelength range. These waveguides enable low-loss integration of electro-absorption modulators, which have 40% lower power and are much smaller than monolithically-integrated Mach-Zehnder Interferometers.

Cost reduction: The fabrication of passives and actives in silicon lowers cost because the PIC integrates expensive BOM components, e.g. PLC demultiplexers, lenses, etc. High-yield wafer processing with low sensitivity to process variations, efficient integration of other materials, low-loss (≤1.5dB) fiber coupling, and high-yield high-throughput fiber assembly reduces manufacturing and assembly cost.

Disaggregate optical transceiver modules

First separate a transceiver module into its constituent parts, i.e., optical engine (optical and high-speed electrical I/O), optical and electrical power supply, and control. Then,

a) consolidate the optical power sources in a shared laser module to reduce cost (simpler transmit assembly, laser sharing), increase laser reliability and serviceability,

b) consolidate the individual small engines into one high-channel-count optical engine to amortize the cost of assembly and packaging over many more channels than in a conventional transceiver, and

c) consolidate the management interfaces for a laser module and its associated optical engines in one control IC to reduce the number of individual control circuits.

Deploy integrated optics as an “optical engine” sub-assembly

The optical engine is a modular unit of deployable bandwidth to support from less than 2 Tb/s to tens of Tb/s in a single optoASIC package (Figure 4). The engine has a channel count that optimizes the trade-off in per-channel cost between yield and assembly/packaging cost. The engine’s electrical interface is compatible with existing host ASIC I/O. This eliminates the need for specialized I/O and enables wrapping any third-party IP core with optical I/O. If tighter integration is desired, the electrical I/O can be optimized to further reduce power.

The engine is attached to a common organic substrate through a micro-electrical connector, which avoids the need for reflow soldering, allows partial reworking of the assembly and simplifies system testing. It has been designed to operate reliably in the challenging thermal environment of hot-running host ASICs.
Pursue inter-operable migration from modular to integrated optics
The initial transition to co-packaged optics must adhere to established standards to prove feasibility and viability. Unproven proprietary approaches are unlikely to meet with acceptance, being perceived as too risky, despite potentially improving on certain metrics. Once initial adoption has taken place, new standards tailored to integrated optics need to be established, for instance to optimize the electrical channels between host ASIC and optical engines or to define new single-mode optical interfaces for server connectivity. Proprietary deployments, e.g., in mega-scale data centers or high-performance computing systems, may benefit from optimizations departing from the standards.

On-board optics (OBO) overcome the mechanical limitations of the faceplate, but do not significantly improve airflow and heat distribution with respect to pluggable optics. OBO also still suffers from a high-loss electrical channel from the ASIC to the module. Unlike In-Package Optics, OBO does not enable a fundamental improvement in system architecture and functional reallocation.

In-Package Optics technology fundamentally changes the economics of reach. The number of discontinuities in the relationship between cost and distance is going to shrink accordingly. This will drastically flatten and simplify the traditional interconnect hierarchy. It is important to realize that In-Package Optics is more than a mere substitute for modular optics: The simultaneous “liberation” from the constraints of electrical reach on one hand and the cost and density constraints of modular optics on the other opens up ample opportunities for rethinking system and network architectures [13][14].

OptoASIC Use Cases
The most obvious application of In-Package Optics is in chips with high bandwidth requirements: merchant silicon Ethernet switch ASICs are the canonical example. The world’s first practical realization of an optoASIC with integrated single-mode optical I/O is Rockley Photonics’ Topanga™ 100GigE switch [14], see cover image. This technology demonstrator confirmed the feasibility of the technology, enabling a 1RU Ethernet switch box (Figure 5) with 12 optical ports (4 channels of 25Gb/s each on parallel single-mode fiber) without any modular optics: the transceivers for all ports were integrated in a single package with the CMOS switch die.

Based on the same silicon photonics platform, we have architected a solution that can scale to support switch ASICs with hundreds of high-speed (50Gbps, 100Gbps) optical I/O channels to match the merchant switch silicon roadmap. As the optical engine’s architecture is modular and complies with existing electrical and optical interface standards, it can be used with switch ASICs of varying capacity across different vendors. In addition, it offers the flexibility to only partially populate a switch’s I/O with optics or combine different optical interface standards tailored to the deployment scenario (for instance, to account for different rate or reach requirements on a switch’s downlinks versus its uplinks). Figure 6 shows an application of our In-Package Optics platform to simplify conventional transceivers.

Figure 5: 1RU Ethernet switch box with 12x100GE ports. The Topanga device implements a Layer-3 switch ASIC along with parallel single-mode fiber optics for all ports. External laser modules supply the optical power. This switch consumes less than 3W per port, which is lower than the typical per-module power of a 100GE optical module alone, but includes all switching and routing functionality.

Figure 6: Rockley’s transmit-receive optical sub-assembly (TROSA) for data communication AOCs and transceivers.
OptoASIC technology will enable high-density (1RU), power-optimized fabric switches (leaf and spine) with all optical interfaces based on the coming generations of 25.6 and 51.2Tbps switch ASICs.

The recent resurgence in artificial intelligence and machine learning (ML) has spawned a new generation of domain-specific ASICs [5] in addition to the more conventional GPUs commonly used to accelerate ML workloads. As ML is a communication-intensive workload requiring high bandwidth in terms of network and memory, these chips have substantial I/O requirements—several of the aforementioned ones even have integrated networking capabilities. Applying In-Package Optics would bring the benefits of integrated optical I/O to ML architectures.

The domain of scientific and technical High-Performance Computing (HPC) has traditionally had much more demanding networking requirements than commercial data center workloads. These needs have been met by specialized network architectures, which were often highly customized and proprietary. Nowadays, this space is dominated by InfiniBand at the high end—with some bastions of custom interconnects persisting—and Ethernet for systems with more relaxed networking requirements. The highest-end systems often resort to ultra-dense packaging to ensure that lots of bandwidth between adjacent nodes can be delivered in a cost-effective manner. The advent of integrated optical I/O opens up alternative paths, delivering high bandwidth at low power and cost, while relaxing packaging constraints.

In a next step, we see intra-rack copper cable connectivity being replaced by integrated optics. Direct Attach Copper cable is commonly used to connect servers to the Top-of-Rack switch located in the same rack. As explained before, Ethernet copper reach continues to shrink, limiting the physical radius into which to pack a growing number of servers. Replacing copper by fiber solves this problem, giving rise to a fiber to the server architecture (Figure 7). From an economic standpoint this is a big challenge as the integrated optics need get close to the per-capacity cost of copper cables. However, this migration has deeper consequences that shift the cost intersection of copper vs fiber to the server when considering total network cost. The overall network architecture can become more efficient, with fewer network layers, fewer network boxes, lower latency, and higher throughput.

The above architectural shift also suggests opportunities for In-Package Optics in end-point connectivity, such as Network Interface Cards (NICs). These typically have a much lower channel count, so the benefits of integration are not as apparent. Still, integrated optics could be applied to NIC ASICs, or lower the cost of individual transceiver modules to render fiber to the server viable.

![Figure 7: Cost-effective optical links not only reduce the cost of communication itself, but also enable architectural changes at the network level. The transition from copper (top) to fiber to the server (bottom) illustrates this point: Connecting more servers per edge switch can eliminate a network tier.](image)

Further down the road, we may see conventional computer system buses (which haven’t been buses in the traditional sense for quite a while)—such as PCIe, QuickPath™, HyperTransport™, NVLink™, as well as emerging contenders such as GenZ, OpenCAPI, and CCIX, adopting integrated optical I/O. Clearly, replacing intra-rack or intra-box electrical wires is a much taller order than replacing modular optics. However, the advent of In-Package Optics will undoubtedly lead to a thorough reevaluation of the system-level cost-performance trade-offs.

**Conclusion**

In-Package Optics technology is here to address the demand for low-cost, high-density optical connectivity, filling the gap created by shrinking electrical reach on one hand and the limitations of conventional modular optics on the other. Based on our silicon photonics platform, optimized for optical performance and high-yield high-throughput manufacturing, we have overcome the technological challenges to kick off the optoASIC revolution.
REFERENCES
[2] IEEE 802.3 New Ethernet Applications Ad Hoc, Ethernet Bandwidth Assessment, Part II.

The Integrated Optics Company

Rockley Photonics is a fabless supplier of silicon photonics chipsets, IP and custom designs for high-volume optics applications.

Learn more at www.rockleyphotonics.com