Optics for Volume Servers

Terry Morris
Hewlett-Packard Company, 3000 Waterview Parkway, Richardson, TX 75080
terrel.morris@hp.com

Abstract: Photonic interconnects provide a solution to the problems of power consumption and diminishing communication radius in volume servers, but cost considerations dictate a very different photonic infrastructure than is found in other applications.

OCIS codes: (200.4650); (200.2605)

1. Introduction

Volume servers utilize industry-standard components to provide the bulk of compute solutions for the planet. The nomenclature is validated by numbers of shipments, greater than 1.6 million units per quarter [1], even during a recession. The majority of volume servers are utilized in data centers, housed within racks or blade enclosures. Within each server are processors and associated memory elements, as well as connections to local and remote storage elements and networking.

The primary motivations for considering optics in volume server applications include the mitigation of power consumption associated with large numbers of inefficient electrical interconnections, reversal of shrinking communication radius, and development of server architectures that are not constrained by the current electrical packaging infrastructure.

2. The (Literal) Power of the Mouse-click

Consider the typical internet search. A few keywords are entered into a search engine, and at the click of a mouse, kilobytes of data traverse the internet to arrive in a host data center. Keywords are routed to a succession of servers, where they are compared to index tables. The resultant low-level communication between processor cores, caches, memory, and I/O on multiple servers causes an explosion of data transfer activity on inefficient electrical interfaces. Board-to-board, chip-to-chip, and on-chip communications are involved in the search. Final results are concatenated and transferred across the internet to the requestor. The significant power consumption for the search occurs in the many GB/TB of low-level communications rather than the few KB of long-distance communications

The relationship between the lengths of specific interconnections and the number of these interconnections occurring in a data center is important. Bautista [2] points out that as interconnections grow progressively shorter, the numbers of interconnections grows progressively larger. Thus there are many more interconnections between racks of systems than between data centers, many more between boards than between racks of systems, many more on a board than between boards, and many more on a chip than between chips.

Looking now at the nature of these interconnections, it can be found that they operate inefficiently when considered as a system. Even though the state of the technology has continually advanced, and power consumption in terms of mW/Gb/S has been reduced from previous generations of technology, it remains that when considered as a system, high-speed electrical interconnections generally operate at less than 2% efficiency.

$$\eta = \frac{P(tf)}{P(t)}$$

Where:

- Eta is the efficiency of the interconnect

- P(tf) is the actual power required to change the state of the final receiving transistor
- P(t) is the total power into the interconnect

It should be noted that energy consumed by any form of signal termination, signal conditioning, ESD structure, and channel loss is counted as waste, as the useful work done by the interconnect system is the successful transmission and reception of data. Considering now a two-socket server, up to 40% of the total energy is consumed in the low-level shuffling of data between compute elements [3]. Considering the projected growth of digital information in the coming decade and the impact of global server power consumption [4], there is sufficient motivation to pursue interconnect power reduction as a fundamental tenet of server design.

3. Mitigating Decreasing Communication Radius

Effective computer system topologies must comprehend the physical realities of system and subsystem design. The combination of compute elements comprising a solution must occupy finite space, and as data rate is increased, the system must revert to a hierarchy of connectivity, thus ensuring that certain elements will be placed further apart than other elements. At the same time, successive generations of electrical interconnections are spanning progressively shorter distances and are less tolerant of connector boundaries.

The net result of these effects is a set of increasingly strict constraints affecting which system elements can be connected at a given distance, limiting the size and performance of the system solution. Furthermore, the variables available for optimization of electrical channels, such as improved driver and receiver signal conditioning, better channel materials, and unique channel constructions have an impact on total system design, cost, and power.

4. Constraints Imposed by Electrically-Optimized Infrastructures

The current system construction of chip packages, DIMMs, boards, and backplanes is the result of four decades of successive electrical optimization. While boards or blades have six possible surfaces for system interconnection, only one or at most two of these surfaces are typically used. New axes of connectivity are possible.

As SerDes electrical channels have become pervasive, system designs have morphed into hierarchies of packet switches, with switch chips forwarding data from one inefficient electrical link to the next until the final destination is reached. Within these complexes, the same data packet is repowered numerous times as it crosses the boundaries associated with communication radius limitations. Additionally, the switches are themselves neither able to compute nor store data, but consume power in order to provide a communication network.

5. Desired Capabilities of Photonic Interconnections

Photonic interconnections have consistently provided communications at longer distances than electronics for equivalent power and cost. Over the last two decades, the tradeoff distance, namely that distance for which electrical and photonic solutions have equivalent cost (combined operating and acquisition cost) and capability at a given data rate, has diminished from kilometers to tens of meters. This capability of extending communication radius within a given bit rate, power, and cost envelope is becoming attractive at distances well under 1M. Additionally, more efficient interconnect hierarchies, such as passive optical busses and coupled matrices of waveguides [5], are promising alternatives to the present packet-switched electrical implementations.

From a power perspective, a point-to-point network has a 1:1 transmitter-to-receiver ratio, thus each connection needs to be independently powered. Alternatively, an optical bus or matrix structure can amortize the power associated with a single transmitter across N receivers, where N is a function of system design, power splitting, and receiver sensitivity.

The elimination of requirements for DC balanced data streams is strongly desired as means to effectively interface photonic interconnects with existing electrically-based compute elements, and free-space interconnections between compute elements allows all sides of a board or blade to be a potential connection point.

6. Cost Considerations

In volume server interconnections, cost is relative to the value provided. If the solution merely replaces an existing or proposed copper link, then it can cost no more than the link it replaces. Alternatively, if the solution provides additional capability, or obviates expensive electronic components, it can withstand costs in line with the value it provides. A good example of this, but over a longer distance, is Infiniband active optical cables, which today obviate multiple "hops" through switch equipment for longer paths. This solution provides superior system-level cost and power metrics when compared to the electrical alternative.

At an industry conference in 2004 [6], a competitive price target of $1/Gb/S was discussed for photonic interconnections within computer systems. It is important to note that the cost of bandwidth for SerDes electronic interconnections in computer applications declines at a rate of approximately 20-25% per year as a function of successive performance gains combined with modest channel cost increases. Carrying the same target forward from 2004, one can see that the new target for and end-to-end connected link is approximately $0.13/Gb/S for 2013.In order to approach this level of cost reduction, the photonic sources, packaging, connectors, and fibers will need to undergo drastic changes relative to their data communication predecessors.

It is interesting to note that photonic and electronic connector costs have remained relatively flat during this transition, riding the gains provided by active components. Unfortunately, this trend cannot continue in the volume server space, because as the projected costs of photonic engines have come down over time, the cost of connectors now exceeds the cost of the photonic solutions they are connecting. The industry will find this situation unsustainable, and new and more cost-effective solutions will be demanded.

More cost-effective photonic sources are also required in this space. The present disparity between the cost of mouse VCSELs and those made for data communications is not supported by a similar disparity in their capability when applied to sub-10M applications.

7. Conclusion

Volume servers represent an outstanding opportunity for low-cost, high-performance photonic interconnections. Demand for lower power consumption and extended communication radius provide one path to market, while demand for new and more capable system topologies provides an alternate route. Solutions that address both needs will be particularly appealing. Standard data communications components are unlikely to meet the cost targets and desired functionality for this demanding application. Market-optimized packaging and connector solutions will be required in order to meet cost requirements, as total solution cost will be compared with the value of the solution at the system level.

[1] Gartner Press Release, Sept 2 2009.

[2] Bautista, J., presentation to OIDA Interconnects Forum, 2004

[3] Astfalk, G. "Why Optical Data Communication and Why Now" Appl Phys A (2009) 95: 933–940, DOI 10.1007/s00339-009-5115-4

[4] J.G. Koomey, "Estimating total power consumption by servers in the US and the world" (Stanford University, 2007)

[5] Beausoleil, R. et al, "A nanophotonic interconnect for high-performance many-core computation." IEEE LEOS Newslett. 22(3), 15–22 (2008)

[6] Kuo, H.,"Free-space optical links for board-to-board interconnects Appl Phys A (2009) 95: 955–965 DOI 10.1007/s00339-009-5144-z

[7] OIDA Conference "Thinking Inside the Box" Forum Report October, 2004