PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability

Keren Bergman
Columbia University
bergman@ee.columbia.edu

John Shalf
Lawrence Berkeley National Lab
jshalf@lbl.gov

George Michelogiannakis
Lawrence Berkeley National Lab
mihelog@lbl.gov

Sebastien Rumley
Columbia University
sr3061@columbia.edu

Larry Dennison
NVIDIA
ldennison@nvidia.com

Monia Ghobadi
Microsoft Research
mgh@microsoft.com

***Abstract*—The cost and complexity of existing interconnects prevent designing datacenter racks tailored to emerging applications such as machine learning. We introduce the PINE interconnect (Photonically Interconnected datacenter Elements), in which compute, memory or storage modules are flexibly combined through one-model-fits-all embedded photonic connectivity and better utilize distant resources.**

I. Introduction

The recent explosive growth in data analytics that rely on machine and deep learning stress computation and communication for training, which has led to broader adoption of GPUs and accelerators. However, currently resources in datacenters are largely organized according to legacy architectures with static node configurations [1]. In applications such as machine learning, existing node architectures result in idling distant resources. Utilizing these resources efficiently calls for a disaggregated architecture that aims to provide a flexible configuration of virtual servers that extend beyond the traditional server blade (Figure 1). However, this relies on high-bandwidth, low-latency interconnects that provide equidistant communication within the elements composing the supernodes within which virtual servers are assembled. This interconnect must furthermore be energy efficient.

We aim to address these challenges and by doing so provide a 2× improvement in performance over power for future deep learning optimized datacenters. To realise this, we combine technological advances in photonics with a reconfigurable optical network that spans across nodes so as to make distant resources logically appear local by providing speed-of-light latency. Photonic Integrated Networked Energy efficient datacenter (PINE) has three key thrusts.

II. Energy-Bandwidth Optimized Optical Links

PINE builds on silicon photonics-based optical interconnect composed of ultra-low power links seamlessly interconnecting modules of diverse types, and of numerous and strategically placed low-to-medium radix optical switches. As recently demonstrated [2], lightweight, strategically disposed low optical loss switches can enable optical bandwidth steering mechanisms that dramatically improve compute performance and energy efficiency. The PINE silicon photonic physical layer is designed by our multi-component parameter optimization to deliver high-bandwidth energy-efficient links [3]. We will leverage stable wavelength comb sources based on dispersion-engineered Si3N4 microring resonator [4], develop custom, energy optimized drivers and receivers in 28nm CMOS, and fabricate optical devices within the AIM photonics manufacturing ecosystem.

This work was supported by ARPA-E ENLITENED Program (project award DE-AR00000843).

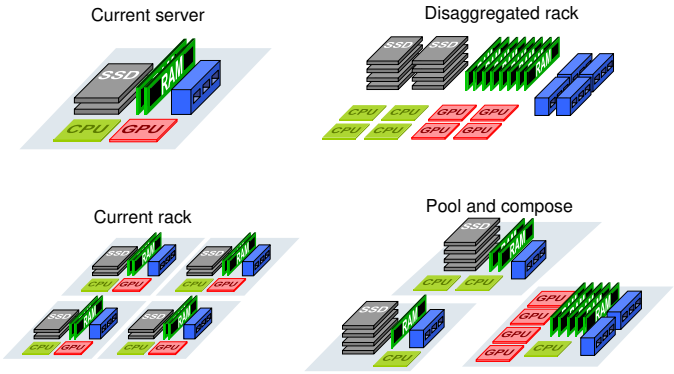

Figure 1: A disaggregated rack places resources of different types (different colors) located in different parts of the datacenter and uses high-bandwidth efficient networking to pool and compose resources together. In the bottom right figure, a logical node can be constructed from distant resources of different types shown in non-grey.

III. Embedded Silicon Photonic Connectivity into MCMs

The PINE concept fundamentally relies on optically connected optically connected MCMs (multi-chip modules),

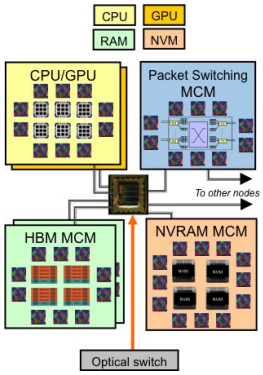

Figure 2: In an example where a node contains four different types of MCMs, PINE can assemble different nodes by configuring photonic switches inside and outside MCMs.

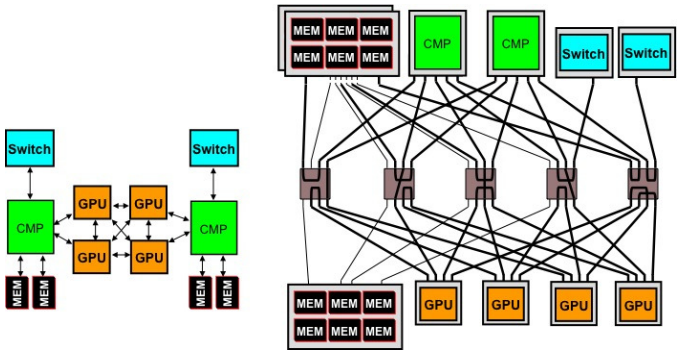

Figure 3: Optical switches inside and outside MCMs can be configured to group resources to match application demands.

which can be compute-type, memory-type, or switch-type. Building on our highly energy efficient links, we can provision abundant amounts of bandwidth around each MCM, solving most bottlenecks present in current server architectures including to memory. Speed of light latency of silicon photonics links will also allow, for instance, a CPU to directly communicate with a distant memory MCM, in case this CPU requires more memory. Optically connected MCMs open the path to totally disaggregated servers and racks, and enables datacenter tasks to straddle over multiple potentially hundreds of MCMs at almost no performance penalty compared to a single server. Thus PINE can assemble application-specific nodes out of modules of various types. Figure 2 illustrates one such example.

IV. Bandwidth Steering for Flexible Connectivity Adaptation

The PINE concept also introduces the notion of bandwidth steering through optical switching. A PINE datacenter will allow the connectivity among optically connected MCMs to be rewired by means of integrated optical switches. This reconfiguration capability could for instance be exploited to wire MCMs in a way that recreates small servers similar to the ones available in today's datacenter or turn the whole datacenter into a single server involving all MCMs. We achieve this by dynamically matching optical connections to the dominant flows of an application. We ensure full connectivity by adding optical connections as needed and routing traffic through intermediate nodes. An example is shown in Figure 3.

V. Impact

PINE offers a promising solution for handling machine learning workloads given the vast demands for interconnect latency, bandwidth, but also the dramatically different node configurations and bandwidth allocations to serve these distinct workloads or phases of the same application such as training or inference. In addition, "disaggregated rack" is a concept of catalog purchasing from a limited menu of node designs and allocating the resources dynamically from these different node types on an as-needed basis across the rack. Datacenter operators are motivated to support this kind of disaggregation because it enables more flexible sharing of hardware resources. However, a conventional Ethernet fabric is a severe inhibitor to efficient resource sharing. The ability for a task to straddle over an almost infinite number of components guarantees ultimate scalability.

Finally, now that 3D stacked memory such as HMC or HBM cost per bit for capacity is getting close to that of conventional DDR, packaging is the only barrier to a technology that can deliver the bandwidth of HBM with the capacity of DDR. The problem is that silicon carrier (the current preferred technology) has limited available surface area to provide capacity for a co-located CPU chip. As a result, we can get necessary bandwidth within the MCM, but not the desired capacity. PINE's photonic technology promises to deliver high bandwidth access to memory with little dependency on physical distance to increase capacity.

VI. Conclusion

PINE combines advances in photonics offering high-bandwidth and low-latency optical communication inside and outside of modules, with bandwidth steering to architecturally group together local and distant resources and thus exactly provision the right balance of resources for a given computation. Though these advancements PINE aims for a system-wide 2× performance over energy increase for applications such as machine learning.

References

[1] A. Singh *et al.*, "Jupiter rising: A decade of Clos topologies and centralized control in Googles datacenter network," in *Sigcomm '15*, 2015.

[2] K. Wen *et al.*, "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics," ser. SC, Nov 2016, pp. 166–177.

[3] M. Bahadori *et al.*, "Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing," ser. DATE, March 2017, pp. 326–331.

[4] J. S. Levy *et al.*, "Cmos-compatible multiple-wavelength oscillator for on-chip optical interconnects," *Nature Photonics*, vol. 4, pp. 37 EP –, Dec 2009.