Photonic Memory Disaggregation in Datacenters

John Shalf[1], George Michelogiannakis[1], Brian Austin[1], Taylor Groves[1], Manya Ghobadi[2], Larry Dennison[3], Tom Gray[3], Yiwen Shen[4], Min Yee Teh[4], Madeleine Glick[4], Keren Bergman[4]

[1]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California, 94720

[2]MIT CSAIL, [3]NVIDIA Corporation,4[1]Columbia University
Corresponding author: jshalf@lbl.gov

Abstract: Datacenter and HPC workloads have diverse memory capacity requirements, but typical node architectures accommodate a superset of those requirements. High bandwidth density photonic link technologies enable efficient memory disaggregation – potentially reducing wasted memory capacity by 8-10x.

1. Introduction

HPC and commercial datacenters rely on purchasing large quantities of identical nodes in order to simplify system management and gain the maximum benefit from volume purchasing. The workloads running on these nodes can have extremely diverse requirements, motivating operators to purchase nodes that can accommodate a superset of those requirements even though that can waste a lot of resources. Recent advances in high-speed, high bandwidth density photonic link technologies combined with broadband optical circuit switching enables resource disaggregation across node boundaries (rack-scale), where the connectivity between typical node-level resources such as memory, storage, CPUs, and GPUs could be configured at runtime. For example, a recent memory utilization study at the NERSC HPC center shows that 50% of the workload uses < 20% of available node memory. Rack-scale resource disaggregation can provide substantial reductions in acquisition cost and energy consumption by reducing the amount of underutilized memory capacity by 8-10x and can provide further gains if applied to right-sizing of the other resources to job requirements.

1.1. Resource Disaggregation in the Datacenter

Datacenter workloads show a large diversity in their resource demands: Training algorithms for deep machine learning stress compute and interconnect elements, in-memory databases stress integrated Non-Volatile Memory (NVM) storage bandwidth, and data-intensive analytics workloads stress memory capacity and bandwidth. For this reason, datacenter operators aspire to move towards fully "disaggregated rack" architectures able to flexibly and dynamically allocate resources such as memory, storage, and compute in response to the mixture of tasks assigned to the datacenter. Many contemporary resource disaggregation solutions are built upon Ethernet technology for the fabric. The cost, power consumption, and latency of conventional Ethernet fabrics utilized by many datacenters, however, are severe inhibitors to efficient resource sharing. In particular, disaggregating high-performance memory poses a particular challenge because of the extremely high data-rates required for the package escape bandwidth. The emergence of ultra-high-bandwidth-density photonic link technologies driven by efficient comb-laser sources open up the opportunity to break out of the package to enable more efficient rack-scale resource disaggregation.

1.2. The Challenges of Package Escape Bandwidth

To meet continued bandwidth demands, memory-intensive compute applications, such as GPUs, have largely moved to 2.5D integrated "in-package" memory technologies, which relies on higher areal density for electrical signals rather than higher signaling rates. Each signal connection runs at a comparatively low signaling rate (sub 5 GHz per IO connection), but with a much smaller pitch between the signal connections. The Heterogeneous Integration Roadmap (HIR) projects five generations of exponential improvement in bandwidth density can be realized from known packaging technology roadmaps [2] with minimal SERDES increases. This wide-and-slow approach has a natural synergy with silicon photonic ring resonator technologies where the optimal efficiency is obtained from using more channels that operate at a slower-rate per channel [3] rather than a small number of high-signal-rate SERDES rate channels. The emergence of efficient solid-state comb-laser sources [3] will take this technology to the next level, where energy efficient (1 picojoule/bit) short-reach (< 7 meters) high-bandwidth (terabyte/second) optical links can become practical and cost-effective. In particular, the bandwidth density of this emerging link technology will enable package escape bandwidths that can support these high memory bandwidths, and enable scalable disaggregation strategies for datacenters in the future.

2. PINE: Photonics for Disaggregated Datacenters

The *Photonic Integrated Networked Energy efficient datacenter (PINE)* architecture [2] takes a systems-perspective on photonics integration to build in resource disaggregation into the datacenter at a fundamental level. The PINE datacenter vision builds on three innovative pillars in Figure 1:

1. A new generation of optical links specifically optimized for energy efficiency.
2. A concept of embedded, high bandwidth density photonic interconnectivity between various types of multi-chip modules (MCMs) in a unique interposer platform.
3. A re-designed system architecture organized around a unified photonic interconnect with bandwidth steering capabilities.

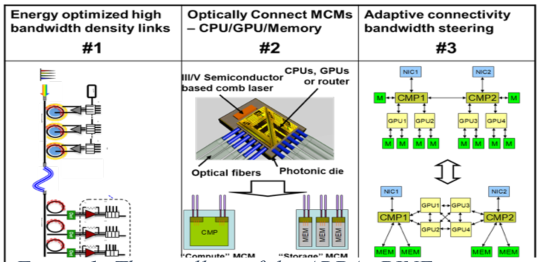

Figure 1: Three pillars of the ARPAe PINE.

The PINE architectural design disaggregates key elements of the traditional datacenter and reorganizes them around a reconfigurable network fabric. PINE disaggregation mechanisms can assign datacenter resources to workloads so that **only the required amount of computation power, memory capacity, and interconnectivity bandwidth are made available over the needed time period.** This efficient usage of resources reduces the vast amounts of wasted energy consumption of current datacenters. Within the PINE datacenter, nodes are interconnected with ultra-low power consuming photonic links uniquely co-integrated into multi-chip modules (MCMs) in our interposer platform.

3. Impact on Datacenter Operation

To illustrate the opportunities for rack-scale disaggregation, we focus on the simpler case of memory disaggregation in order to assign different quantities of memory to each node. When purchasing large-scale systems, one must select the memory capacity up-front – usually with a desire in mind to accommodate the applications with the largest memory requirements in the workload. As a result, systems are often provisioned with more memory than is required by the majority of jobs, but since the memory capacity per node cannot currently be configured at runtime, this is the best that we can do. The rule of thumb over many decades has been to purchase one byte of memory capacity per peak floating-point operation per second of compute performance (one byte-per-flop in shorthand). However, this mythical ratio has rarely been tested in practice. Indeed, Figure 2 shows data from the Cori supercomputing system at NERSC and indicates that although 15% of jobs require nearly all of the 128 Gigabytes of memory in the node, more than > 50% of CPU hours go to jobs that consume 25 Gigabytes or less.

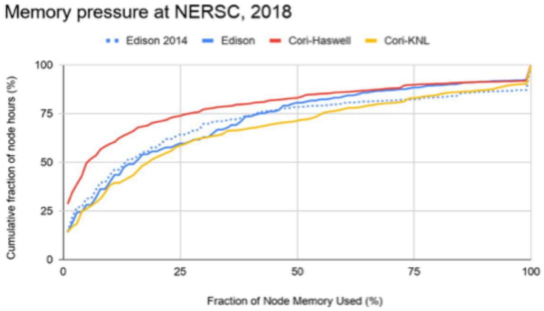

Figure 2: Memory utilization on NERSC systems.

The PINE photonic link technology offers the bandwidth densities needed to carry memory traffic at full rate, and offers the reach necessary to access as little or as much memory as is required for the job running on the node. By using broadband optical circuit switches to patch in the desired amount of memory (right-sizing) at full in-package bandwidths, memory disaggregation reduces the total amount of memory capacity purchased system-wide. Since memory consumes nearly 25-50% of total system cost and total system power consumption [4], the opportunity to reduce capital acquisition costs for systems and operating costs for power are substantial.

4. References

[1] 2019 Heterogeneous Integration Roadmap (HIR), *https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2019-edition.html*

[2] K. Bergman et. al., "PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability" in *2018 IEEE Optical Interconnects Conference (OI), June 4-6, 2018*.

[3] M. Bahadori, K. Bergman: Low-Power Optical Interconnects based on Resonant Silicon Photonic Devices: Recent Advances and Challenges. ACM Great Lakes Symposium on VLSI 2018: 305-310, 2018.

[4] S. Ghose et al. 2018. What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study. In Proc. ACM Measurement and Analysis of Computing Systems, Vol. 2, No. 3,). ACM, 2018.