

Cacheless Computer Architectures: 3D Integration of Optical Interconnects and Novel Memory

Eric Cheng, S. J. Ben Yoo

As we near the end of traditional complementary metal-oxide semiconductor (CMOS) scaling, and systems are further constrained by various “walls” (e.g., power, memory, resilience), we are no longer seeing historical year-over-year exponential performance improvements. New technologies, architectures, and integration techniques are required to ensure future computing systems (from embedded to high-performance) continue to deliver the capabilities required by the scientific, business, and national security communities. Memory remains a significant limiter when it comes to the energy efficiency, scalability, performance, and productivity of computing systems. In particular, traditional memory hierarchies (i.e., the cache hierarchy) have been optimized over decades to map well to workloads with frequent, regular access and are not well-suited for the sparse, random-access workloads that are emerging and dominating key applications in the high-performance data analytics (HPDA) and graph analytics space.

Re-architecting the memory subsystem to better account for these sparse, irregular workloads and, in particular, flattening traditional memory hierarchies, can drastically reduce energy consumption, reduce memory-access latency, improve memory-access predictability, increase memory bandwidth, and enhance programmer productivity. This so-called “cacheless computer architecture” is enabled by advances in massively parallel silicon photonic wavelength-division multiplexed (WDM) interconnects, novel memory devices and architectures, and innovative electronic/photonic three-dimensional (3D) integration capabilities. This new architecture features 3D integration of optically-interconnected low-latency memory (LLM) to provide four times the memory capacity and two times the improvements in latency and energy efficiency as compared to traditional dynamic random-access memory (DRAM) while demonstrating a five times reduction in average memory access time and 60 percent reduction in access latency variability across a variety of application workloads.

Vision

Emerging applications in problem domains such as HPDA and graph analytics are placing ever-increasing demands on the computing systems available to the computing community and, in particular, stretching the capabilities of today's high-performance computing (HPC) systems. The global semiconductor industry is constantly developing new technologies to help provide the massive performance improvements and new capabilities required to meet these demands. However, effective hardware/software codesign along with intelligent integration of these available technologies into new system architectures are required to fully realize the potential for future compute systems. By leveraging advanced technologies such as massively parallel silicon photonic WDM interconnects, novel memory devices, and innovative electronic/photonic 3D integration capabilities, we can create a new "cacheless computer architecture" that is better optimized for new problem domains and provides better scalability for future HPC systems. This cacheless architecture implements a flattened memory architecture that can drastically reduce energy consumption, reduce memory access latency, improve memory access predictability, increase memory bandwidth, and enhance programmer productivity (see [figure 1](#)).

HPDA and graph analytics workloads are often dominated by memory operations (as opposed to more compute-dominated workloads) and exhibit random or irregular access to data that is sparsely distributed across a large memory footprint. With the introduction of silicon photonics (SiPh), it is possible

to begin removing levels of the cache hierarchy (i.e., flattening the hierarchy) to reduce energy consumption and access latency while increasing memory bandwidth. This new architecture also improves memory access predictability, which is key to enhancing programmer productivity, simplifying application analysis, and supporting better algorithm design.

Technologies

This cacheless computer architecture is enabled by the convergence of several technologies (incubated over the past decade) that are effectively integrated into an overall system architecture. Some of these technologies have been fully demonstrated (i.e., in the fab), while others are nearing maturity (i.e., in the lab).

In the fab

As we enter a future of heterogeneous compute [i.e., architectures with a diverse mix of compute elements ranging from central processing units (CPUs) to graphics processing units (GPUs) to fixed-function accelerators that are coupled with diverse memory and storage solutions] and resource disaggregation (i.e., logical or physical clustering/separation of distinct resources that are connected via a network), the need for highly scalable interconnection networks that provide high bandwidth, low latency, all-to-all communication (i.e., every device on the network can directly communicate with every other device simultaneously) is crucial. Massively parallel SiPh using WDM have the ability to provide this scalable and high-bandwidth

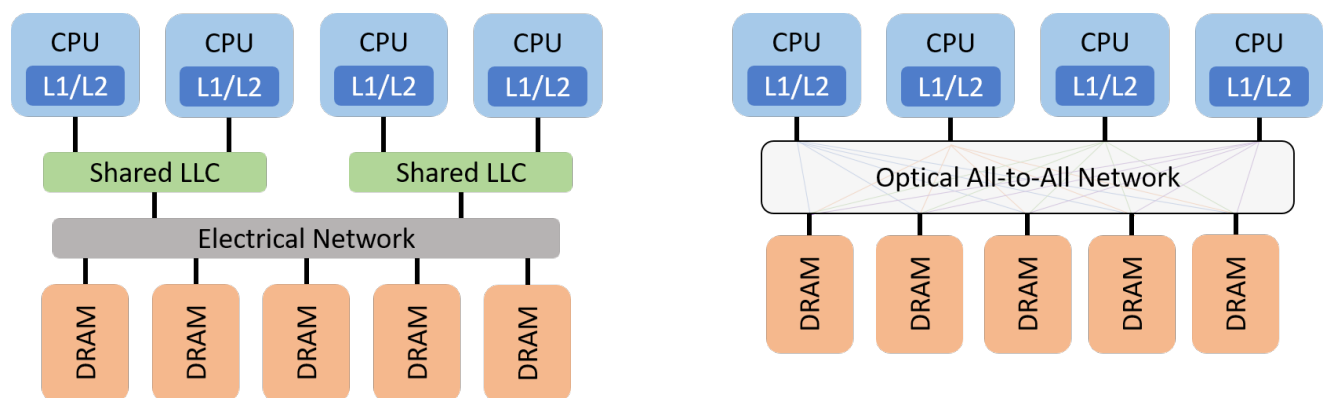


FIGURE 1. Comparison of a traditional architecture consisting of a shared last-level cache (LLC) electrically connected to DRAM (left) versus a (last-level) cacheless architecture consisting of compute elements optically connected to DRAM (right).

networking capability for computing systems. In particular, by leveraging arrayed waveguide grating routers (AWGRs) [1], we can achieve very compact interconnect fabrics that provide all-to-all connectivity between devices.

This optical interconnect fabric is created as follows: a laser source (typically an off-chip, external laser) is used to generate the various wavelengths required, which can be generated by a single frequency comb or by multiple individual lasers. The use of WDM allows for these multiple wavelengths to traverse over a single waveguide. Individual frequency-tuned modulators are used to encode data on each wavelength, and the corresponding frequency-tuned photodetector decodes the data from the corresponding wavelength. An AWGR uses this general concept to connect multiple input nodes in an all-to-all manner to multiple output nodes (i.e., every input is directly connected to every output).

Importantly, AWGRs do not just exist as a conceptual or theoretical design. Compact 8 x 8 silicon nitride (SiN) AWGRs have been fabricated and demonstrated in a compact 1 square millimeter (mm²) footprint [2] (see figure 2). Physical demonstrations of scaled networks implementing 512 x 512 SiPh AWGRs have also been fabricated [3]. Such fabricated devices demonstrate the feasibility of realizing actual systems using SiPh as well as showcase the ability to provide much better scalability, single-hop distance-independent energy-efficient communication, and higher bandwidth communication as compared to using traditional electrical-only equivalents. As the number of nodes in a system grows, the hardware cost for implementing SiPh grows linearly as opposed to quadratically for the electrical equivalents, thus providing better scalability. Furthermore, tight integration of such SiPh technologies (see later sections) hold the potential for reducing memory access energy from order 2–4 picojoules (pJ) per bit down to order 1 pJ per bit and increasing aggregate memory bandwidth from 1 gigabyte (GB) per second to 1 terabyte (TB) per second as compared to current electrical-only technologies.

In the lab

Innovation in the interconnect fabric is not the only advanced technology required to realize a cacheless architecture. While the following technologies are

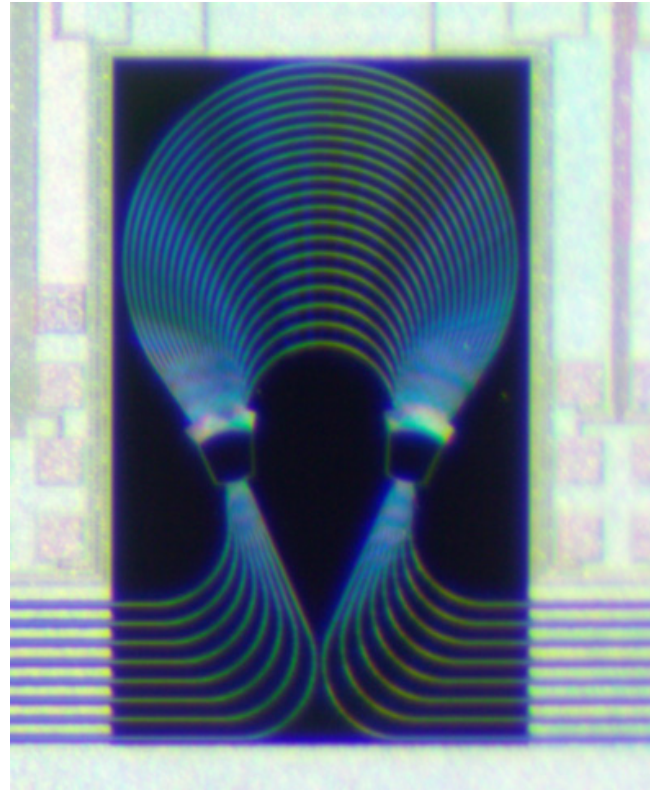


FIGURE 2. Fabricated 8 x 8 silicon nitride (SiN) arrayed waveguide grating router (AWGR).

arguably less mature, they have nonetheless demonstrated significant benefits and have been validated through detailed simulation and small-scale benchtop experiments.

Innovative electronic/photonic 3D integration techniques are necessary to provide the capabilities necessary to effectively couple SiPh technologies with conventional electronic devices. In particular, 3D stacking (i.e., stacking multiple silicon wafers/dies together to form a single integrated circuit) can greatly enhance the capabilities of a cacheless architecture by providing tight integration of compute and memory dies or by providing much greater memory capacity per packaged part. Traditional through-silicon vias (TSVs) are used to provide the connectivity across stacked dies but have bandwidth and scalability limitations due in part to TSV density constraints. An alternative is to use through-silicon optical vias (TSOVs), which are enabled by the use of ultra-compact vertical U-shaped couplers [4, 5]. TSOVs can provide additional bandwidth, reduce the via density, maintain low-energy communication across stacked

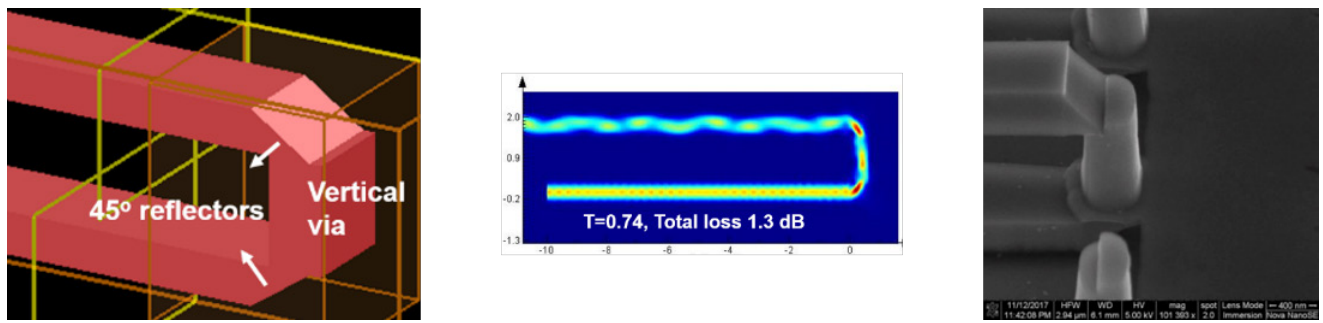


FIGURE 3. Projected view of an ultra-compact vertical U-shaped coupler (left), simulated optical propagation through the coupler (center), and scanning electron microscope (SEM) image of a fabricated coupler (right).

dies, and ensure scalability in future process nodes. Small-scale component fabrication of the U-shaped couplers, which are key to implementing TSOVs, have demonstrated the ability to achieve 1–2 micron (μm) pitches with detailed simulation demonstrating transmission losses of 1.3 decibels (dB) [4] (see [figure 3](#)). The combination of these small-scale demonstrators and detailed simulations help validate the feasibility of this electronic/photonic 3D integration technology.

Key limiters to achieving a cacheless architecture using traditional memory devices and architectures [e.g., DRAM, high-bandwidth memory (HBM), etc.] include access energy, access latency, access granularity, memory bandwidth, and overall capacity. With the assistance of optical interconnect fabrics and innovative electronic/photonic integration capabilities described earlier, novel memory devices and architectures (such as LLM) can be constructed to help overcome these limitations. Building off the conceptual fine-grained DRAM (FG-DRAM) [6] design, an extension referred to here as LLM can provide improvement across all of these parameters.

LLM leverages dedicated optical buses to memory banks to reduce contention and improve bandwidth. The use of TSOVs helps reduce access latency, increase memory capacity (through the ability to integrate a greater number of memory dies), and minimize the number of vias required to effectively connect and provide I/O (input/output) to the memory stacks (see [figure 4](#)). In the age of disaggregation, (CPU) cores would be connected directly to shared, global memory. To reduce memory access latency in this configuration, LLM leverages a dedicated memory controller at each core connected to the memory controller blocks on the memory side. The compute-side memory

controller (CMC) contains the read and write queues while the memory-side memory controller (MMC) contains the scheduling logic and command queues. The CMCs and MMCs communicate with one another via an all-to-all optical interconnect provided by the use of AWGRs. Detailed simulation has demonstrated that such an LLM design can be used to increase memory capacity by up to four times, reduce access latency and energy by two times, reduce average memory access time by five times, and reduce access latency variability by 60 percent as compared to conventional DRAM-based HBM.

Application impact

Each of the individual technology components used to realize a cacheless architecture demonstrate significant potential; however, intelligent integration and codesign is required to realize the full potential and understand the overall impact on application workloads. To more fully explore and evaluate the benefits of a cacheless computer architecture, a broad set of applications spanning traditional HPC workloads [e.g., Advanced Encryption Standard (AES), convolution, Fast Fourier Transform (FFT)] to HPDA/graph workloads [e.g., Breadth-First Search (BFS), PageRank] were evaluated.

As the motivating design point, a cacheless GPU design [i.e., a design that has removed the level 2 (L2) caches, uses conventional HBM, and leverages AWGRs] was compared to an equivalent multi-GPU system (i.e., a design with a conventional memory hierarchy, HBM, and is electrically connected) providing the same number of compute elements in both designs. Across the applications studied, a codesigned

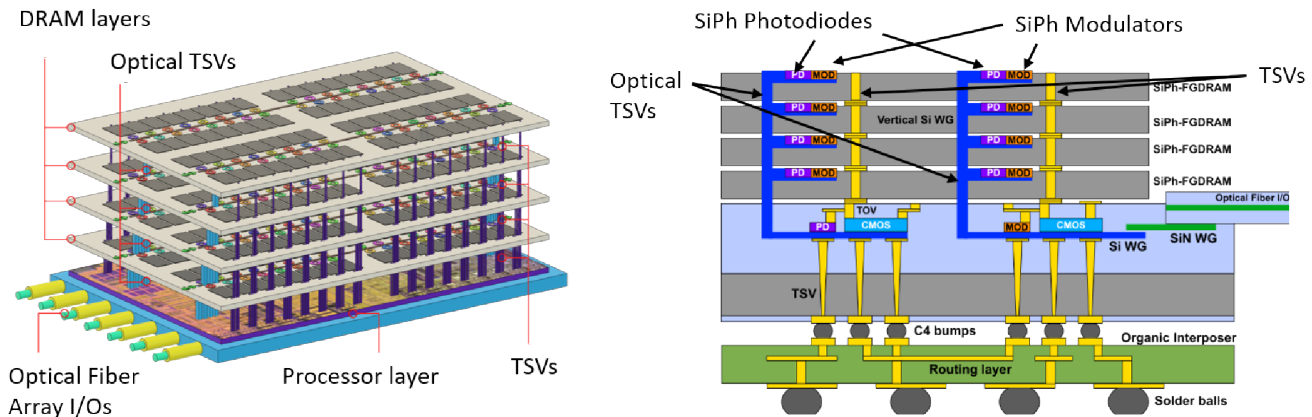


FIGURE 4. Projected view (left) and side view (right) of low-latency memory (LLM) with a SiPh base layer integrated via through-silicon optical vias (TSOVs) to the DRAM stacks.


cacheless GPU can reduce overall DRAM access latency by 10–55 percent, improve level 1 (L1) miss penalty by 2.3–5 times, and offer an overall application speedup of 1.1–1.8 times [Z].

Specifically, if we focus on two applications, FFT (a traditional HPC application) and PageRank (an important graph workload), we can concretely identify the sources of improvement. The removal of L2 caches significantly improves access latency for FFT by 31 percent and for PageRank by 10 percent. L1 miss penalty improves on FFT by 3.2 times and on PageRank by 2.8 times. Taken together, this translates to roughly a 1.4 times speedup on FFT and a roughly 1.8 times speedup on PageRank for the cacheless GPU design.

Conclusion

A cacheless computer architecture holds great potential for greatly improving the capabilities of future compute systems and, in particular, for applications in the HPDA and graph analytics domains. By leveraging innovative 3D integration techniques, optical interconnect fabrics, and LLM, future systems can be better optimized for and perform more efficiently on emerging workloads that are characterized by sparse and irregular accesses to memory. Through a combination of physically fabricated test vehicles, subcomponent demonstrations, and detailed simulations, a cacheless

computer architecture has been shown to not only be highly promising, but also well-within reach for system demonstrations at commercial foundries.

For a range of application workloads, an initial study of a cacheless computer architecture has demonstrated that it is possible to achieve a two times improvement in memory access latency and energy efficiency, a five times reduction in average memory access time, and a 60 percent reduction in access latency variability. Not only do these improvements have direct impact on the performance of applications, they also provide secondary benefits. A cacheless architecture, which provides more predictable application performance, affords an application programmer (or even an advanced compiler) a greater ability to easily reason about, analyze, and optimize applications for the cacheless architecture. This opens up the potential for additional performance gains. Finally, with the simpler and more predictable architecture, it also becomes easier to design algorithms that can better take advantage of the underlying hardware architecture (e.g., algorithms that more closely resemble underlying mathematical constructs as opposed to needing to worry about manually blocking, partitioning, and placing data to fit within the limits of conventional memory subsystems). This cacheless computer architecture opens up many possibilities for future computing systems. 

References

- [1] Grani P, Liu G, Proietti R, Yoo SJB. "Bit-parallel all-to-all and flexible AWGR-based optical interconnects." *Optical Fiber Communication Conference, OSA Technical Digest* (online). 2017. Available at: <https://doi.org/10.1364/OFC.2017.M3K.4>.
- [2] Shang K, Pathak S, Qin C, Yoo SJB. "Low-loss compact silicon nitride arrayed waveguide gratings for photonic integrated circuits." *IEEE Photonics Journal*. 2017;9(5). doi: 10.1109/JPHOT.2017.2751003.
- [3] Cheung S, Su T, Okamoto K, Yoo SJB. "Ultra-compact silicon photonic 512x512 25 GHz arrayed waveguide grating router." *IEEE Journal of Selected Topics in Quantum Electronics*. 2014;20(4). Available at: <https://doi.org/10.1109/JSTQE.2013.2295879>.
- [4] Zhang Y, Ling YC, Zhang Y, Shang K, Yoo SJB. "High-density wafer-scale 3-D silicon-photonic integrated circuits." *IEEE Journal of Selected Topics in Quantum Electronics*. 2018;24(6). Available at: <https://doi.org/10.1109/JSTQE.2018.2827784>.
- [5] Zhang Y, Samanta A, Shang K, Yoo SJB, "Scalable 3D silicon photonic electronic integrated circuits and their applications." *IEEE Journal of Selected Topics in Quantum Electronics*. 2020;26(2). Available at: <https://doi.org/10.1109/JSTQE.2020.2975656>.
- [6] O'Connor M, Chatterjee N, Lee D, Wilson J, Agrawal A, Keckler SW, Dally WJ. "Fine-grained DRAM: Energy-efficient DRAM for extreme bandwidth systems." In: *MICRO-50 '17: Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*; 2017 Oct. pp. 41–54. Available at: <https://doi.org/10.1145/3123939.3124545>.
- [7] Fotouhi P, Fariborz M, Proietti R, Lowe-Power J, Akella V, Yoo SJB. "HTA: A scalable high-throughput accelerator for irregular HPC workloads." In: *ISC High Performance 2021: High Performance Computing*. Part of the *Lecture Notes in Computer Science* book series (LNCS, volume 12728). Springer International Publishing; 2021. Pp. 176–194. Available at: https://doi.org/10.1007/978-3-030-78713-4_10.

