Towards Exascale HPC Systems: Exploiting Advances in High Bandwidth Memory (HBM2) Through Scalable All-to-All Optical Interconnect Architectures

Pouya Fotouhi, Roberto Proietti, Paolo Grani, Mohammad Amin Nazirzadeh, S. J. Ben Yoo
Department of Electrical and Computer Engineering, University of California
Davis, California
pfotouhi,rproietti,pgrani,manazirzadeh,sbyoo@ucdavis.edu

ABSTRACT

As we reach the limits of miniaturization in fabrication processes, the interpretation of Moore's law has changed from doubling the frequency every eighteen month to doubling the core count every three to four years (from 2 cores in 2004 to 16 cores in 2015). To reach exascale-level computation, the communication and data transfers between processors and memory is expected to increase drastically; the on-chip interconnect plays a key role in the overall system latency and energy-efficiency. Therefore, novel solutions providing one order of magnitude higher bandwidth and lower energy consumption than what possible with current electrical interconnects are needed. This paper discusses an optical interconnected compute node that makes use of embedded photonic interconnects together with emerging high bandwidth memory technologies (such as HBM and HBM2). Two different multi-processors architectures with different requirements in terms of number of lasers, high-speed SERDES and memory bandwidth per processors are discussed.

1 EXASCALE CHALLENGES

Achieving performance improvements through technology scaling appears to be impractical as we reach the limitations of miniaturization in fabrication process. Therefore, high-performance systems tend to shift toward multi to many cores. This shift turns the on-chip interconnect into the bottleneck for performance as well as energy consumption. Meanwhile, to achieve higher memory capacity and memory bandwidth on chip, 2.5D integration (connecting several dies through an interposer) and 3D stacking (vertically stacked dies) has been proposed. For instance, [7] illustrates the details of an interposer-based system with 64 cores and 4 stacks of High Bandwidth Memory (HBM). Each HBM module, as described in [5], consists of 4 vertically integrated layers of 2 Gb DRAM dies with 128 GB/s as the total I/O bandwidth required (1024 bits operating at 1 Gb/s). [6] outlines specifications of the second generation of HBM systems, indicating a total bandwidth as high as 256 GB/s.

These solutions comes with several challenges to be addresses. First of all, the bandwidth should be increased by one order of magnitude when designing interconnects for such systems. Secondly, restrictions in terms of power consumption should be respected while achieving such high bandwidth. The latter challenge turns out to be more significant considering the power budget to be shared by all the components at the chip level (*e.g.* memory, interconnection, processors), and also the thermal issues related with 3D die-stacks. Since a considerable increase in the overall power budget of the chip is hard to imagine in near future (limited by Thermal Design Power), power consumption of the interconnection network (*i.e.* energy per bit) shall be improved significantly.

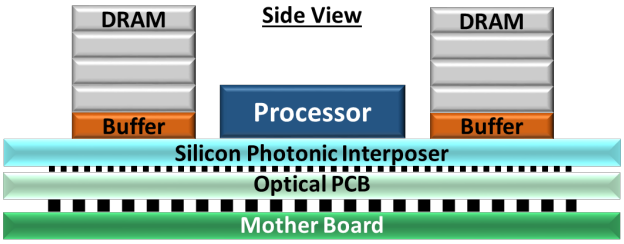

Figure 1: An architecture containing multiple silicon photonic optical interposer and electronic ICs

Electrical interconnects have poor power consumption and suffer lack of scalability due the nature of electrical signaling. Supposedly, an electrical interconnect will not be able to meet the requirements of HPC systems in future [8]. Alternatively, through use of silicon photonics, optical on-chip communication appears to be the compelling solution by providing lower and nearly distant-independent energy consumption, higher bandwidth, low crosstalk and built-in parallelism.

Figure 1 shows a system with optical silicon interposer and HBM2 as memory on chip as an example of the architecture under discussion. By using optical interposer based solution and wavelength multiplexing and routing in Arrayed Waveguide Grating Router (AWGR) [9], we propose to interconnect multiple processors and HBM modules exploiting high-bandwidth wavelength division multiplexing (WDM) optical links. Section 2 introduce the principle of AWGR routing, Section 3 discussed the details of two different compute node architectures we are currently investigating.

2 ARRAYED WAVEGUIDE GRATING ROUTER (AWGR)

The AWGR is a passive optical cross connect consisting of two star couples connected through waveguides with different lengths. Diffraction grating is obtained through linear increase in the length of the waveguides. AWGR provides a nonblocking cross connect since each output port receives the full set of wavelengths with each wavelength coming from a different input.

Figure 2 illustrates the functionality of an 5×5 AWGR, allowing simultaneously all-to-all communication among all input and output ports using different wavelengths.

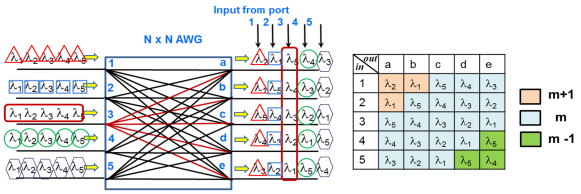

Figure 2: (TOP) Wavelength routing and wavelength assignment in a 5×5 AWGR.

3 ARCHITECTURE AND SIMULATION STUDIES

In this section we report our current studies on two multi-processors architectures with AWGR-based photonic interposer. The first one uses a 16×16 AWGRs connecting 8 processing nodes and 8 HBM modules. The second architecture takes a different approach where fewer HBM modules (HBM2) are shared among multiple processors to trade-off per-processor memory bandwidth and complexity.

3.1 Architecure 1

We assume 256 I/Os for each processing node (*i.e.* CPUs, GPUs) operating at 4 GHz resulting in a bandwidth of 1 Tb/s in each direction. Each HBM module has 1024 I/Os operating at 1 Gb/s to realize 1 Tb/s link bandwidth in each direction. We compared our performance results against recent state-of-the-art interposer based electrical NoC [4].

For the optical interconnect, two scenarios are considered. The first scenario, O-Interposer-Base, consists of 8 AWGR slices in parallel with 4 tunable lasers operating at 32 Gb/s to realize a per-processor link bandwidth of 1 Tb/s. Therefore, the network interface on processing nodes and HBMs requires 8:1 and 32:1 SERDES units respectively. The second scenario is designed to reduce the transmission clock frequency by using more parallel AWGRs. This scenario is referred as O-Interposer-Opt for the rest of this section. O-Interposer-Opt consists of 17 parallel AWGRs with 12 tunable lasers operating at 5 Gb/s instead of 32 Gb/s in O-Interposer-Base. Note that the link bandwidth is the same is previous scenario (1 Tb/s in each direction). In this case, the need for SERDES in transmitters and receivers on processing node is eliminated which can contribute significantly to power savings.

To evaluate the performance of the architecture described above, we used gem5 simulator [2] in Full-System mode running Linux 2.6.27. We modeled Chip Multi Processors with 64 core using Alpha ISA. PARSEC benchmark suit [1] were used to examine the architecture under investigation. As can be seen in Figure 3, two main scenarios are considered for simulations. One with 32 KB and the other with 16 KB as L1 cache size. In the first scenario, O-Interposer-Base outperforms E-Interposer by 25% on average. Note that O-Interposer-Opt performs slightly better compared to O-Interposer-Base but

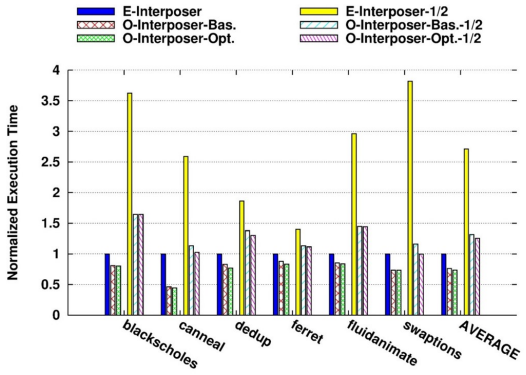

Figure 3: Execution time results

is expected to have lower power consumption. Dividing the size of L1, O-Interposer-Base outperforms E-Interposer by 50%. Also, note how shrinking the size of caches has higher impact on the electrical interposer interconnect compared to the optical solution. The performance of E-Interposer is declined by a factor of 3 (compared to full sized cache scenario) whereas the performance of O-Interposer-Base is decreased by a factor of 1.5. Also, note that the performance of O-Interposer-Base with half-size caches is only 30% lower compared to OE-Interposer with full-sized caches. Details on the experiment setup, power analysis, and further discussions on results are presented in [3].

3.2 Architecure 2

While Architecture 1 provide very high per-processor memory bandwidth (1 HBM module=1 Tb/s), the complexity in terms of number of tunable lasers, SERDES and AWGR is fairly high. As our next step, we are currently evaluating a different and simpler architecture (see Figure 4) with fewer HBM modules shared among multiple processors. 8 processing nodes are connected to two HBM dies through a 16×16 AWGR. This architecture can reduce of a factor of 4 the number of lasers and SERDES per processors at the expenses of a factor of 4 reduction in memory bandwidth. At the same time, it requires fixed-wavelength lasers (not tunable) and only one AWGR. This study aims investigating trade-offs between link bandwidth and complexity as well as power consumption when comparing with the first architecture.

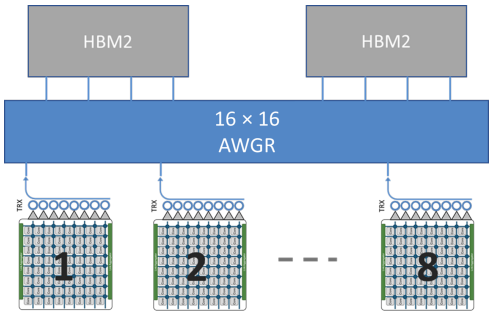

Figure 4: Architecture with HBM2 modules shared among multiple processors

4 ACKNOWLEDGEMENTS

This work was supported in part by DoD award # W911NF-13-1-0090, DoD contract # H98230-16-C-0820, and NSF grant # 1611560.

REFERENCES

[1] Christian Bienia, Sanjeev Kumar, Jaswinder Pal Singh, and Kai Li. 2008. The PARSEC Benchmark Suite: Characterization and Architectural Implications. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques (PACT '08)*. ACM, New York, NY, USA, 72–81. https://doi.org/10.1145/1454115.1454128

[2] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R. Hower, Tushar Krishna, Somayeh Sardashti, Rathijit Sen, Korey Sewell, Muhammad Shoaib, Nilay Vaish, Mark D. Hill, and David A. Wood. 2011. The Gem5 Simulator. *SIGARCH Comput. Archit. News* 39, 2 (Aug. 2011), 1–7. https://doi.org/10.1145/2024716.2024718

[3] P. Grani, R. Proietti, V. Akella, and S. J. B. Yoo. 2017. Design and Evaluation of AWGR-Based Photonic NoC Architectures for 2.5D Integrated High Performance Computing Systems. In *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 289–300. https://doi.org/10.1109/HPCA.2017.17

[4] A. Kannan, N. E. Jerger, and G. H. Loh. 2015. Enabling interposer-based disintegration of multi-core processors. In *2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 546–558. https://doi.org/10.1145/2830772.2830808

[5] D. U. Lee, K. S. Lee, Y. Lee, K. W. Kim, J. H. Kang, J. Lee, and J. H. Chun. 2015. Design considerations of HBM stacked DRAM and the memory architecture extension. In *2015 IEEE Custom Integrated Circuits Conference (CICC)*. 1–8. https://doi.org/10.1109/CICC.2015.7338357

[6] D. U. Lee, K. S. Lee, Y. Lee, K. W. Kim, J. H. Kang, J. Lee, and J. H. Chun. 2015. Design considerations of HBM stacked DRAM and the memory architecture extension. In *2015 IEEE Custom Integrated Circuits Conference (CICC)*. 1–8. https://doi.org/10.1109/CICC.2015.7338357

[7] Gabriel H. Loh, Natalie Enright Jerger, Ajaykumar Kannan, and Yasuko Eckert. 2015. Interconnect-Memory Challenges for Multi-chip, Silicon Interposer Systems. In *Proceedings of the 2015 International Symposium on Memory Systems (MEMSYS '15)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/2818950.2818951

[8] Ian O'Connor and Gabriela Nicolescu. 2012. *Integrated optical interconnect architectures for embedded systems*. Springer Science & Business Media.

[9] R. Yu, S. Cheung, Y. Li, K. Okamoto, R. Proietti, and S. J. B. Yoo. 2014. A silicon photonic chip-scale AWGR switch for high performance computing systems. In *2014 Conference on Lasers and Electro-Optics (CLEO) - Laser Science to Photonic Applications*. 1–2.