Future Memory and Interconnect Technologies

Yuan Xie[*][†]
[*] Pennsylvania State University, USA
[‡] AMD Research, Advanced Micro Devices, Inc., USA
Email: yuanxie@cse.psu.edu

Abstract—**The improvement of the computer system performance is constrained by the well-known *memory wall* and *power wall*. It has been recognized that the memory architecture and the interconnect architecture are becoming the overwhelming bottleneck in computer performance. Disruptive technologies, such as *emerging non-volatile memory (NVM)* technologies, *3D integration*, and *optical interconnects*, are envisioned as promising future memory and interconnect technologies that can fundamentally change the landscape of the future computer architecture design with profound impact. This invited survey paper gives a brief introduction of these future memory and interconnect technologies, discusses the opportunities and challenges of these new technologies for future computer system designs.**

I. Introduction

In the past decades, the performance of computer systems has been improved by technology scaling and architecture innovations, with the dramatic improvement in CPU computing power via frequency scaling and/or multi/many-core architectures. Every computing system requires a storage and communication substrate to store and access data for its computational engines. While the computational engines (cores) have continuously benefited from the scaling offered by Moores law, it is widely recognized that future improvement of the system performance, is constrained by the well-known *memory wall* (which refers to the growing disparity of speed between cores and memories due to the limitation of communication bandwidth) and *power wall* constraints. Consequently, the memory architecture and the interconnect architecture have been a laggard/bottleneck in terms of both performance and power improvements.

Disruptive technologies are expected to help the memory hierarchy and interconnect to catch-up and tremendously scale in performance, energy-efficiency, and storage capacity to sustain the processing demands of the fast CPU cores. Three disruptive technologies are envisioned as the keys for changing the future landscape of computer system designs:

- **Emerging non-volatile memory (NVM) technologies**, such as Sing Torque Transfer RAM (STT-RAM), Phase change RAM (PCRAM), and Resistive RAM (ReRAM), have demonstrated great potentials to be the replacement of today's SRAM/DRAM memory technologies;
- **3D integration technologies** not only provide much shorter on-chip interconnects but also enable the heterogeneous integration that provides a cost-effective way to use emerging NVMs with CMOS logics;
- **Optical interconnect** provides a promising low-power and high bandwidth mechanism for both on-chip and off-chip communication.

It is important for computer architects to study such emerging memory and interconnect technologies, to understand the benefits and limitations for better utilizing them to improve the performance/power/reliability of future computing systems. This invited survey paper gives brief introduction of these technologies, reviews recent advances in such emerging memory and interconnect technologies, discusses the benefits and limitations of using these technologies at various perspective of architecture design.

II. Emerging Non-Volatile Memory Technologies

Technology scaling of SRAM and DRAM (which are the common memory technologies used in traditional memory hierarchy) are increasingly constrained by fundamental technology limits. In particular, the increasing leakage power for SRAM/DRAM and the increasing refresh dynamic power for DRAM have posed challenges for circuit/architecture designers for future memory hierarchy design.

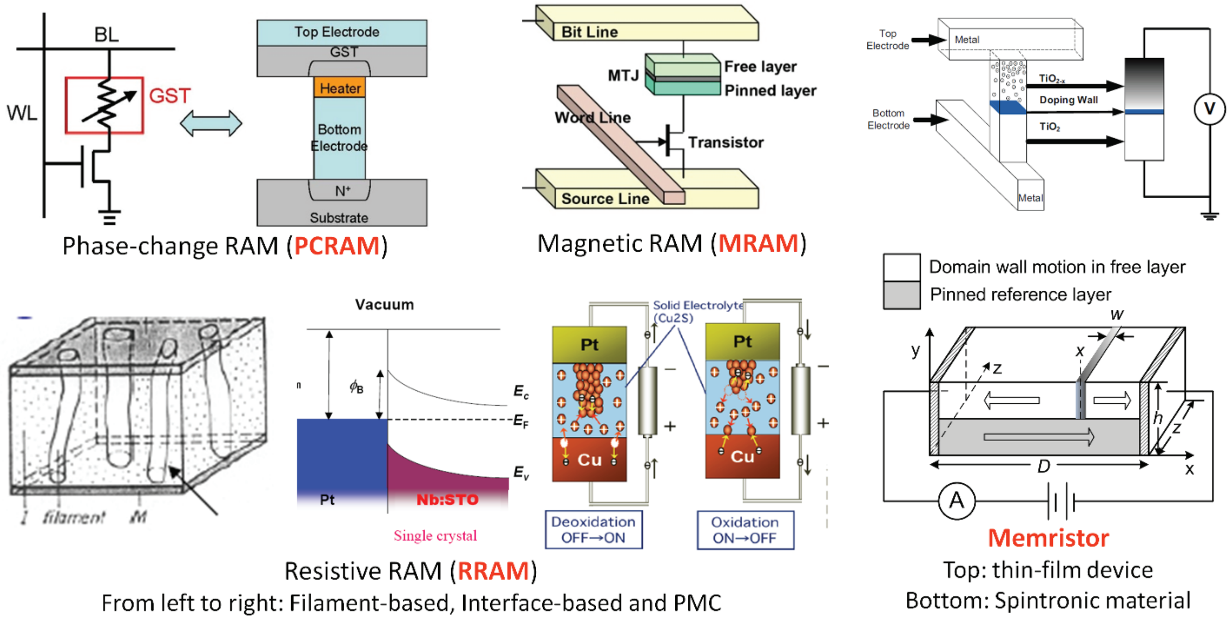
Recently, emerging memory technologies (such as STT-RAM, PCRAM, and ReRAM), are being explored as potential alternatives of existing memories in future computing systems (Figure 1). Such emerging non-volatile memory (NVM) technologies combine the speed of SRAM, the density of DRAM, and the non-volatility of Flash memory, and hence, become very attractive as the alternatives for future memory hierarchy. It is anticipated that these NVM technologies will break important ground and move closer to market very rapidly.

STT-RAM is a new type of Magnetic RAM (MRAM) [1]. The storage elements of STT-RAM is the magnetic tunneling junction (MTJ), in which a thin tunneling dielectric is sandwiched by two ferromagnetic layers, as shown in Figure 1. One ferromagnetic layer ("pinned layer") has a fixed magnetization direction, while the magnetization of the other layer ("free layer") can be flipped by a switching current. An MTJ has a low (high) resistance if the magnetizations of the free layer and the pinned layer are parallel (anti-parallel), to represent the storage of "0" or "1".

PCRAM technology is based on a chalcogenide alloy (typically, Ge_2–Sb_2–Te_5, GST) material).The data storage capability is achieved from the resistance differences between an amorphous (high-resistance) and a crystalline (low-resistance)

This work is supported in part by NSF grants 1218867, 0903432, 1017277, and SRC grants.

978-3-9815370-0-0/DATE13/© 2013 EDAA



Phase-change RAM (PCRAM)

Magnetic RAM (MRAM)

Resistive RAM (RRAM)

From left to right: Filament-based, Interface-based and PMC

Memristor

Top: thin-film device

Bottom: Spintronic material

Fig. 1: Various emerging non-volatile memory technologies.

phase of the GST node. In SET operation, the phase change material is crystallized by applying an electrical pulse that heats a significant portion of the cell above its crystallization temperature. In RESET operation, a larger electrical current is applied and then abruptly cut off in order to melt and then quench the material, leaving it in the amorphous state. PCRAM has shown to offer compatible integration with CMOS technology, fast speed, high endurance, and inherent scaling of the phase-change process at 22-nm technology node and beyond. Phase change material has a key advantage of the excellent scalability within current CMOS fabrication methodology, with continuous density improvement.

Resistive RAM (ReRAM) and Memristor. In an ReRAM cell, the data is stored as two (single-level cell, or SLC) or more resistance states (multi-level cell, or MLC) of the resistive switch device (RSD). Based on the storage mechanisms, ReRAM materials can be classified as filament-based, interface-based, programmable-metallization-cell (PMC), etc. Based on the electrical property of resistive switching, RSDs can be divided into two categories: unipolar or bipolar. Programmable-metallization-cell (PMC)is a promising bipolar switching technology. Its switching mechanism can be explained as forming or breaking the small metallic “nanowire” by moving the metal ions between two sold metal electrodes. Filament-based RRAM is a typical example of unipolar switching that has been widely investigated. The insulating material between two electrodes can be made conducting through a hopping or tunneling conduction path after the application of a sufficiently high voltage. The data storage could be achieved by breaking (RESET) or reconnecting (SET) the conducting path. Such switching mechanism can in fact be explained with the fourth circuit element, the **memristor** [2].

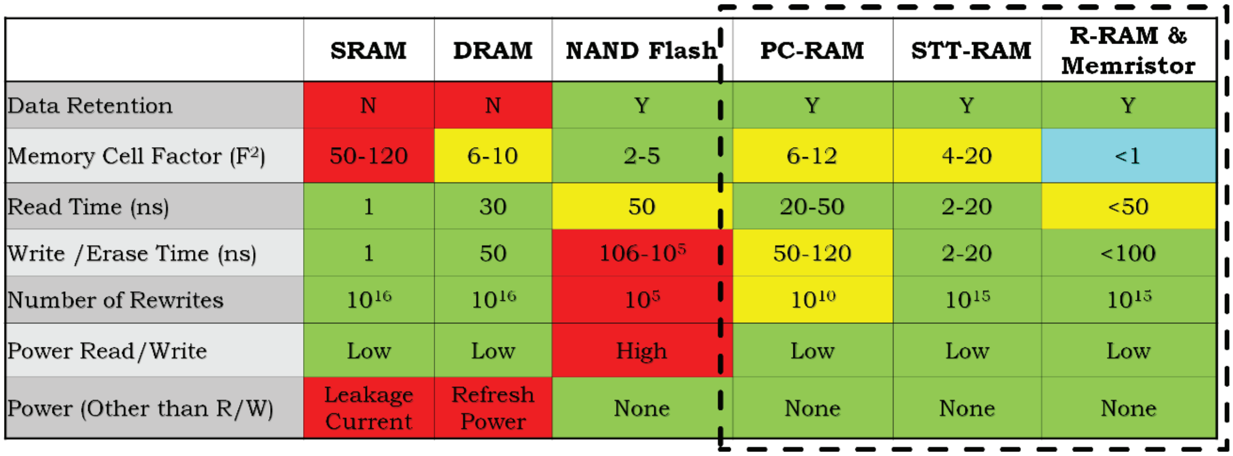
Figure 2 illustrates the comparison of emerging memory technologies against the traditional main-stream SRAM, DRAM, and NAND-based Flash memory [1].

III. Leveraging Emerging Memory Technologies in Architecture Design

As the emerging memory technologies are getting mature, integrating such memory technologies into the memory hierarchies provides new opportunities for future memory architecture designs. Specifically, there are several characteristics of STT-RAM, PCRAM, and ReRAM that make them promising as working class memories (i.e., on-chip caches and off-chip main memories), or as storage class memories: (1) Compared to SRAM/DRAM, these emerging memories usually have much higher density, with comparable fast access time; (2) Due to the non-volatility feature, they have zero standby power, and immune to radiation-induced soft errors; (3) Compared to NAND-Flash SSD, STT-RAM/PCRAM are byte-addressable. In addition, different hybrid compositions of memory hierarchy by using SRAM, DRAM, and PCRAM or MRAM can be motivated by different power and access behaviors of various memory technologies. For example, leakage power is dominant in SRAM and DRAM arrays; on the contrary, due to non-volatility, PCRAM or MRAM array consumes zero leakage power when idling but a much higher energy during write operations. Hence, the trade-off among using different memory technologies at various hierarchy levels becomes an interesting research topic.

A. NVM Architecture Modeling

For computer architects to explore new design opportunities at architecture and system levels that the emerging memory technologies can provide, architectural level cache/memory model built with NVM such as PCRAMsim [3] and NVsim [4] have been recently developed. Such architectural models provide the extraction of all important parameters, including access latency, dynamic access power, leakage power, die area, and I/O bandwidth *etc.*, to facilitate architecture-level analysis, and to bridge the gap between the abundant research activities



	SRAM	DRAM	NAND Flash	PC-RAM	STT-RAM	R-RAM & Memristor
Data Retention	N	N	Y	Y	Y	Y
Memory Cell Factor (F^2)	50-120	6-10	2-5	6-12	4-20	<1
Read Time (ns)	1	30	50	20-50	2-20	<50
Write /Erase Time (ns)	1	50	$106\text{-}10^5$	50-120	2-20	<100
Number of Rewrites	10^{16}	10^{16}	10^5	10^{10}	10^{15}	10^{15}
Power Read/Write	Low	Low	High	Low	Low	Low
Power (Other than R/W)	Leakage Current	Refresh Power	None	None	None	None

Fig. 2: The comparison of various memory technologies [1].

at process and device levels and the lack of a high-level cache and memory model for emerging NVMs.

B. Leveraging NVMs as Cache or Main Memory

Replacing SRAM-based on-chip cache or DRAM-based main memory with NVM can potentially improve performance and reduce power consumption [5] [6]. With larger on-chip cache capacity (due to its higher density), NVM-based on-chip cache can help reduce the cache miss rate, which helps improve the performance [7], [8]. Zero-standby leakage can also help reduce the power consumption in cache/memory. On the other hand, longer write-latency of such NVM-based cache may incur performance degradation and offset the benefits from the reduced cache miss rate. Although PCRAM is much denser than SRAM, the limited endurance makes it unaffordable to directly use PCRAM as on-chip caches, which have highly frequent accesses. It is therefore a more feasible approach to use PCRAM as main memory replacement.

C. Leveraging NVM to improve NAND-Flash SSD

NAND flash memory has been widely adopted by various applications such as laptops and mobile phones. In addition, because of its better performance compared to the traditional HDD, NAND flash memory has been proposed to be used as a cache in HDD, or even as the replacement of HDD in some applications. However, one well-known limitation of NAND flash memory is the “erase-before-write” requirement. It cannot update the data by directly overwriting it. Instead, a time-consuming erase operation must be performed before the overwriting.

Compared to NAND flash memory, emerging NVM has advantages of random access and direct in-place updating. Consequently, one possible solution is a hybrid storage architecture to combine the advantages of NAND flash memory and PCRAM/STT-RAM [9]. In such hybrid storage architecture, PCRAM/STT-RAM is used as the log region for NAND-flash. Such hybrid architecture has the following advantages: (1) the ability of “in-place updating” can significantly improve the usage efficiency of log region by eliminating the out-of-date log data; (2) the fine-granularity access of PCRAM can greatly reduce the read traffic from SSD to main memory; (3) the energy consumption of the storage system is reduced as the overhead of writing and reading log data is decreased with the PCRAM log region; and (4) the lifetime of the NAND flash memory in the hybrid storage could be increased because the number of erase operations is reduced.

D. Mitigation Techniques for Emerging NVM Memory

The benefits of using these emerging memory technologies in computer system design can only be achieved with mitigation techniques that can help address the inherited disadvantages that related to the write operations: (1) Because of the non-volatility feature, it usually takes much longer and more energy for write operations compared to read operations; (2) Some emerging memory technologies such as PCRAM has the wear-out problem (lifetime reliability), which is one of the major concerns of using it as working memory rather than storage class memory. Consequently, introducing these emerging memory technologies into current memory hierarchy design gives rise to new opportunities but also presents new challenges that need to be addressed.

a) Techniques to Mitigate Latency/Energy Overheads of Write Operations: In order to use the emerging NVMs as cache and memory, several design issues need to be solved. The most important one is the performance and energy overheads in write operations. The NVM has a more stable mechanism for data keeping, compared to a volatile memory such as SRAM and DRAM. Accordingly, it needs to take a longer time and consume more energy to over-write the existing data. This is the intrinsic characteristic of NVMs. PCRAM, MRAM, and ReRAM are not exceptional. If we directly replace SRAM/DRAM memory with NVM, the long latency and high energy consumption in write operations could offset the performance and power benefits, and even result in degradation when the cache/memory write intensity is high. Therefore, it is imperative to study techniques to mitigate the overheads of write operations in NVMs. Such techniques include *Hybrid Cache/Memory Architecture* [6] [10], *Read-Preemptive Novel Buffer Architecture* [6],*Redundant Write Elimination* [11], *etc.*

b) Wear Leveling Techniques to Improve Lifetime for NVMs: Write endurance is another severe challenge in PCRAM memory design. The state-of-the-art process technology has demonstrated that the write endurance for PCRAM is around 10^10 . The problem is further aggravated by the fact that writes to caches and main memory can be extremely skewed. Consequently, those cells suffering from more frequent write operations will fail much sooner than the rest. Techniques that proposed in the previous sub-section to reduce the number of write operations to NVM can definitely help the lifetime of the memory, besides reducing the write energy overhead. In addition to those techniques, wear leveling technique attempts to work around the limitations of write endurance by arranging data access so that write operations can be distributed evenly across all the storage cells [8]. Such wear leveling techniques include: Row Shifting, Word-line Remapping and Bit-line Shifting, and Segment Swapping [11].

IV. 3D Integration Technology

In a 3D integrated IC chip, multiple device layers are stacked together. The layers could be connected with wire bonding, through-silicon-vias (TSV), microbump, or even inductive/capacitive contact [12]. Figure 3 shows a conceptual 2-layer 3D integrated circuit with direct vertical interconnects called *through-silicon vias (TSV)*.

Three-dimensional integrated circuits (3D ICs) [13], [14] are attractive options for overcoming the barriers in interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology. 3D integration also provides new opportunities for future CMP design, with a new dimension of design space exploration. In particular, the heterogenous integration capability enabled by 3D integration gives designers new perspective when designing future CMPs.

Even though manufacturing/process techniques for 3D integrations are nearly mature, and 3D ICs offer tremendous benefits, there are several challenges that hinder the successful adoption of 3D architectures: (1) design space exploration at the architectural level is essential to take full advantage of 3D integration and to build a high-performance 3D multi-core systems. New opportunities brought by 3D technology can result in innovations in new architectures for future many-core chip multiprocessor (CMP) [15]. (2) there are no commercially available electronic design automation (EDA) tools and design methodologies for 3D ICs. The absence of EDA tools that can explore the design space for 3D multi-core systems is an impediment to researchers and industry practitioners in their quest for the adoption of this new technology.

V. Leveraging 3D Integration in Architecture Design

The following subsections will discuss various architecture design approaches that leverage different benefits that 3D integration technology can offer, namely, wire length reduction, high memory bandwidth, and heterogeneous integration.

A. Wire Length Reduction

Designers have resorted to technology scaling to improve microprocessor performance. Although the size and switching speed of transistors benefit as technology feature sizes continue to shrink, global interconnect wire delay does not scale accordingly with technologies. The increasing wire delays have become one major impediment for performance improvement. Compared to a traditional two dimensional chip design, one of the important benefits of a 3D chip over a traditional two-dimensional (2D) design is the reduction on global interconnects. It has been shown that three-dimensional architectures reduce wiring length by a factor of the square root of the number of layers used [16]. The reduction of wire length due to 3D integration can result in two obvious benefits: *latency improvement* and *power reduction*. For example, since interconnects dominate the delay of cache accesses which determines the critical path of a microprocessor, and the regular structure and long wires in a cache make it one of the best candidates for 3D designs, 3D cache design is one of the early design example for fine-granularity 3D partition [17], and the latency reduction can be as much as 25% for a two-layer 3D cache. 3D arithmetic-component designs also show latency benefits. For example, various designs [18] have shown that the 3D arithmetic unit design can achieve around 6%-30% delay reduction due to the wire length reduction.

Note that such fine-granularity design of 3D processor components increases the design complexity, and the latency improvement varies depending on the partitioning strategies and the underlying 3D process technologies.

B. Memory Bandwidth Improvement

Three-dimensional integration has been envisioned as a solution for future micro-architecture design (especially for multi-core and many-core architectures), to mitigate the interconnect crisis and the “memory wall” problem. It is anticipated that using 3D stacked memory with CPU in a package would be one of the early commercial uses of 3D technology for future chip-multiprocessor design [15], by providing improved memory bandwidth for such multi-core/many-core microprocessors. In addition, such approaches of memory stacking on top of core layers do not have the design complexity problem as demonstrated by the fine-granularity design approaches, which require re-designing all processor components for wire length reduction.

There are two possible approaches to leverage 3D stacked memory for future microprocessor designs. One is to stack memory on top of logic [15], [19], which requires a close collaboration between microprocessor vendors and memory vendors, and thermal management is a big challenge. The other is to place 3D stacked memory (such as Micron’s Hybrid Memory Cube) with CPUs on a silicon interposer [20] so that the CPU design and 3D stacked memory design can be decoupled, and the thermal management is much easier than the first approach.

C. Heterogenous Integration

3D integration also provides new opportunities for future architecture design, with a new dimension of design space exploration. In particular, the heterogenous integration capabil-

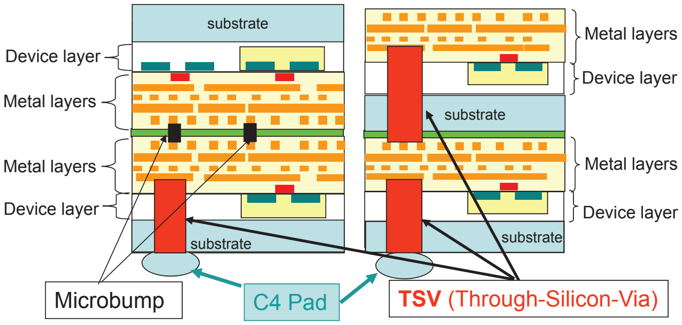


Fig. 3: Illustration of Face-to-face and Face-to-back 3D integration.

ity enabled by 3D integration gives designers new perspective when designing future CMPs.

3D integration technologies provide feasible and cost-effective approaches for integrating architectures composed of heterogeneous technologies to realize future microprocessors targeted at the ”More than Moore” technology projected by ITRS. 3D integration supports heterogeneous stacking because different types of components can be fabricated separately, and layers can be implemented with different technologies. It is also possible to stack optical device layers or non-volatile memories (such as magnetic RAM (MRAM) or phase-change memory (PCRAM)) on top of microprocessors to enable cost-effective heterogeneous integration [5] [6]. The addition of new stacking layers composed of new device technology will provide greater flexibility in meeting the often conflicting design constraints (such as performance, cost, power, and reliability), and enable innovative designs in future microprocessors.

D. Challenges for 3D Architecture Design

Even though 3D integrated circuits show great benefits, there are several challenges for the adoption of 3D technology for future architecture design: (1) *Thermal management.* The move from 2D to 3D design could accentuate the thermal concerns due to the increased power density. To mitigate the thermal impact, thermal-aware design techniques must be adopted for 3D architecture design [13]; (2) *Design Tools and methdologies.* 3D integration technology will not be commercially viable without the support of EDA tools and methodologies that allow architects and circuit designers to develop new architectures or circuits using this technology. To efficiently exploit the benefits of 3D technologies, design tools and methodologies to support 3D designs are imperative; (3) *Cost implication.* Cost analysis and cost-driven design methodologies are also needed for 3D IC technology to become main stream [21], [22]. (4) *Testing.* One of the barriers to 3D technology adoption is insufficient understanding of 3D testing issues and the lack of design-for-testability (DFT) techniques for 3D ICs [23], [24], which have remained largely unexplored in the research community.

VI. Optical Interconnect Technology

Even though 3D memory stacking can help mitigate the memory bandwidth problem, when it comes to off-chip communication, the pin limitations, the energy cost of electrical signaling, and the non-scalability of chip-length global wires are still significant bandwidth impediments. Recent developments in silicon nanophotonic technology have the potential to meet the off-chip communication bandwidth requirements at acceptable power levels. With the heterogeneous integration capability that 3D technology offers, one can integrate optical die together with CMOS processor dies. For example, HP Labs proposed the Corona architecture [25], which is a 3D many-core architecture that uses nanophotonic communication for both inter-core communication and off-stack communication to memory or I/O devices. A photonic crossbar fully interconnects its 256 low-power multithreaded cores at 20 terabyte per second bandwidth, with much lower power consumption.

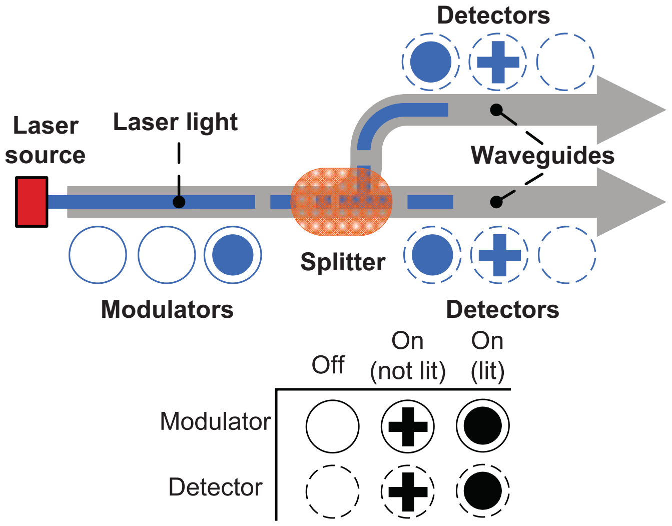


Fig. 4: A conceptual nanophotonic link, which consists of a laser source, waveguides, and micro-rings as modulators or detectors.

In addition to helping the off-chip communication, the emerging nanophotonic technology enables on-chip optical interconnects that are faster and less power-consuming than electrical wires, therefore it has been leveraged to build various on-chip interconnects.

Figure 4 illustrates the basic component of the nanophotonic interconnect, which consists of a laser source, waveguides carrying light injected by the laser source, and micro-rings to modulate and detect optical signals. With dense-wavelength-division-multiplexing (DWDM), up to 128 wavelengths can be generated and carried by the waveguides, which increases the bandwidth density to over 320Gb/s/um. Micro-rings can be electrically tuned into resonance (the on state) and remove light from waveguides; or out of resonance (the off state) and let light pass by unaffected. This mechanism is leveraged to modulate light into on-off signals. Doping Ge in a micro-ring turns it into a optical detector. When the doped micro-ring is turned on, it removes light from the waveguide and converts optical signals to electrical ones. Detecting is destructive which means if a detector is turned on then downstream detectors will not be able to detect light. A splitter is used to direct a fraction of light power to another waveguide without affecting modulated signals. It is needed to implement broadcast in nanophotonic links.

Even though optical interconnect provides many promising

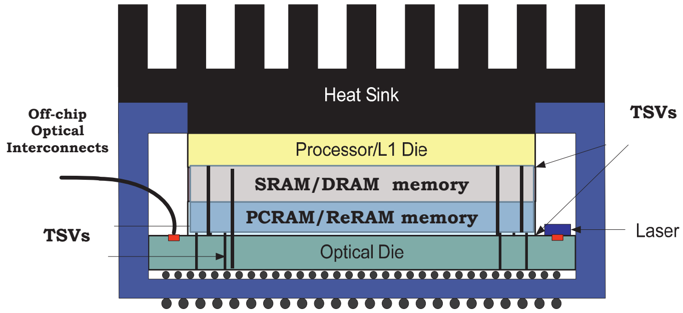


Fig. 5: The ultimate future computer system architecture with a combination of three disruptive future memory and interconnect technologies: emerging NVM, 3D integration, and on-chip/off-chip optical interconnects.

features such as low power and high bandwidth [25], [26], there are also key challenges needed to be addressed to providing robust and reliable on-chip communication. Among many challenges, the thermal sensitivity and process variations (PV) of silicon photonic devices are the key issues [27].

VII. Conclusion

While the computational engines (cores) have continuously benefited from technology scaling and multi/many-core innovations, the future improvement of the system performance is constrained by the well-known *memory wall* and *power wall*. Disruptive technologies for future memory and interconnect architectures, including emerging non-volatile memories, 3D integrations, and optical interconnects, are envisioned as possible solutions to continue the performance scaling. Figure 5 illustrate the ideal ultimate future computer system architecture with a combination of three disruptive future memory and interconnect technologies that integrates non-volatile memories and optical die together through 3D integration technology. With all the initial research efforts in recent years, we believe that the emerging of these new technologies will change the landscape of future computer system designs.

References

[1] International Technology Roadmap for Semiconductor, 2007.

[2] Dimin Niu, Yiran Chen, Cong Xu, and Yuan Xie. Impact of process variations on emerging memristor. In *Design Automation Conference (DAC 2010)*.

[3] Xiangyu Dong, Norm Jouppi, and Yuan Xie. PCRAMsim: System-level performance, energy, and area modeling for phase-change RAM. In *International Conference on Computer-Aided Design (ICCAD)*, pages 269–275, 2009.

[4] X. Dong, C. Xu, N. Jouppi, and Y. Xie. NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 31(7):994–1007, 2012.

[5] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li, and Y. Chen. Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement. In *Proceedings of Conference on Design Automation*, pages 554–559, 2008.

[6] Guangyu Sun, Xiangyu Dong, Yuan Xie, Jian Li, and Yiran Chen. A novel architecture of the 3D stacked MRAM L2 cache for CMPs. In *IEEE International Symposium on High Performance Computer Architecture, 2009.*, pages 239–249, 2009.

[7] Cong Xu, Dimin Niu, Xiaochun Zhu, Seung H. Kang, Matt Nowak, and Yuan Xie. Device-architecture co-optimization of stt-ram based memory for low power embedded system. In *International Conference on Computer-Aided Design (ICCAD)*, pages 463–470, 2011.

[8] Jue Wang, Xiangyu Dong, Norm Jouppi, and Yuan Xie. i^{2}wap: Improving non-volatile cache lifetime by reducing inter- and intra-setwrite variations. In *Intl. Symp. on High Performance Computer Architecture (HPCA)*, 2013.

[9] Guangyu Sun, Yongsoo Joo, Yibo Chen, Yuan Xie, Yiran Chen, and Helen Li. A hybrid solid-state storage architecture for performance, energy consumption and lifetime improvement. In *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2010.

[10] Moinuddin K. Qureshi, Vijayalakshmi Srinivasan, and Jude A. Rivers. Scalable high performance main memory system using phase-change memory technology. In *International symposium on Computer architecture (ISCA)*, pages 24–33, 2009.

[11] Yongsoo Joo, Dimin Niu, Xiangyu Dong, Guangyu Sun, Naehyuck Chang, and Yuan Xie. Energy- and endurance-aware design of phase change memory caches. In *Proceedings of Design Automation and Test in Europe*, 2010.

[12] W. Rhett Davis *et al.* Demystifying 3D ICs: the Pros and Cons of Going Vertical. *IEEE Design and Test of Computers*, 22(6):498– 510, 2005.

[13] Y. Xie, G Loh, B Black, and K. Bernstein. Design space exploration for 3D architectures. *ACM Journal of Emerging Technologies in Compuing Systems*, 2006.

[14] Yuan Xie, Jason Cong, Sachin Sapatneker . *Three-Dimensional Integrated Circuit Design: EDA, Design and Microarchitectures*. Springer, 2009.

[15] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, N. Vijaykrishnan, and M. Kandemir. Design and management of 3D chip multiprocessors using network-in-memory. In *International Symposium on Computer Architecture (ISCA)*, 2006.

[16] J.W. Joyner, P. Zarkesh-Ha, and J.D. Meindl. A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3D-SoC). In *Proc. 14th Annual IEEE International ASIC/SOC Conference*, September 2001.

[17] Yuh-fang Tsai, Feng Wang, Yuan Xie, N. Vijaykrishnan, and M. J. Irwin. Design space exploration for three-dimensional cache. *IEEE Transactions on Very Large Scale Integration Systems*, 2008.

[18] Jin Ouyang, Guangyu Sun, Yibo Chen, Lian Duan, Tao Zhang, Yuan Xie, and Mary Irwin. Arithmetic unit design using 180nm TSV-based 3D stacking technology. In *IEEE International 3D System Integration Conference*, 2009.

[19] G. H. Loh. 3d-stacked memory architectures for multi-core processors. In *Proc. of International Symposium on Computer Architecture,* pages 453–464, 2008.

[20] Xiangyu Dong, Yuan Xie, Norm Jouppi, and Naveen Muralimanohar. Simple but effective heterogeneous main memory with on-chip memory controller support. In *International Conference on High Performance Computing, Networking, Storage and Analysis (SC10)*.

[21] Xiangyu Dong, Jishen Zhao, and Yuan Xie. Cost analysis and cost-driven design for 3D ICs. In *IEEE Transactions on CAD (TCAD)*, pages 1959–1972, 2010.

[22] Xiangyu Dong and Yuan Xie. System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs). In *Asia and South Pacific Design Automation Conference (ASP-DAC 2009)*, pages 234–241, 2009.

[23] Yibo Chen, Dimin Niu, Yuan Xie, and Krish Chakrabarty. Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis. In *International Conference on Computer-Aided Design (ICCAD)*, pages 471–476, 2010.

[24] Jing Xie, Yu Wang, and Yuan Xie. Yield-aware time-efficient testing and self-fixing design for TSV-based 3D ICs. In *Asia and South Pacific Design Automation Conference (ASP-DAC 2012)*, 2012.

[25] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn. Corona: System implications of emerging nanophotonic technology. In *Computer Architecture, 2008. ISCA '08. 35th International Symposium on*, pages 153–164, 2008.

[26] Jin Ouyang and Yuan Xie. Enabling quality-of-service in nanophotonic network-on-chip. In *Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 351–356, 2011.

[27] Yi Xu, Jun Yang, and Rami G. Melhem. Tolerating process variations in nanophotonic on-chip networks. In *Intl. Symp. on Computer Architecture (ISCA)*, pages 142–152, 2012.