RETROSPECTIVE: 3D-Stacked Memory Architectures for Multi-Core Processors

Gabriel H. Loh
Advanced Micro Devices, Inc.[1]

ABSTRACT

This retrospective takes a look back at the original paper fifteen years later and reflects on some of the impacts this paper has had on the computer architecture community. This short retrospective also reflects on aspects of the paper that did not quite pan out or where the community is still waiting for the technology to become commercially viable, both in terms of the 3D-stacked memory itself as well as its potential use with multi-core (CPU) processors.

1 Impact of this Work

When this paper was originally written in the spring of 2007, computer architecture conferences were already beginning to see their first crop of papers on exploring the impact of 3D die-stacking technology on CPU architecture, microarchitecture, and circuits [3][13][15][17]. However, these early papers primarily focused on logic-on-logic stacking of silicon. At this time, a few papers had also started discussing the possibilities of 3D-stacking DRAM on top of processors [5][7][9][12], but this ISCA 2008 paper seemed to capture more attention and mindshare, perhaps due to the larger potential impact of using this emerging technology to provide a step function toward dealing with the ever-challenging “Memory Wall” problems of mainstream multi-core processors [24].

In the years that followed, the computer architecture research community went from viewing die-stacking as a far-off technological curiosity to being something that a researcher could assume in their latest proposals without giving much additional thought (or justification to their paper reviewers). Die-stacking became another tool in the computer architecture researcher’s toolbox, and today we see it in various forms in a range of commercially-available products.

Independent of the specific technical proposals in the ISCA 2008 paper, this work provided a basic tutorial and introduction to die-stacking techology that helped the technological concepts reach a wider swath of the computer architecture community. This in turn accelerated the snowballing process of computer architecture researchers coming up with more ideas for how to leverage die stacking, publishing those results, and driving the virtuous cycle of innovation as their papers went on to inspire others. While I cannot speak for others for how much this ISCA 2008 paper may have inspired their subsequent research, looking back this marked an inflection point in my own research agenda for the fifteen years that followed. This included substantial efforts in pondering how exactly stacked DRAM could or should be used, especially when combined with a second tier of higher capacity memory. Key areas of follow-on research included DRAM caching [10][19], OS-visible heterogeneous/multi-level memory [14], and eventually informing and influencing our exascale research at AMD [11][21][23].

Regarding the technical proposal in the ISCA 2008 paper, a key theme in the paper was the coordinated design or adaptation of the multi-core CPU architecture to better match a high-bandwidth 3D-stacked memory system. In particular, the paper discussed how to align the address interleaving of both the last-level cache banks/slices to the address interleaving of the memory sysem to minimize unnecessary data movement. This general address-based clustering of cache and memory can be seen today in other high-bandwidth processors such as in a GPU’s cache and memory hierarchy [2].

2 Shortcomings of the Original Work

With the benefit of looking back after fifteen years of technology advancement, some of the original assumptions of the ISCA 2008 paper now seem somewhat quaint. The original paper assumed a monolithic quad-core processor with a single stack of memory on top. Today, modern server CPUs can support 96-128 cores [1] implemented across multiple chiplets [16], where chiplets were not even really a known concept (at least in academia). The original paper also only considered up to four memory channels for the entire memory stack, whereas today a single high-bandwidth memory (HBM) stack provides sixteen 64-bit (pseudo)channels [8]. Even if we had made such aggressive predictions fifteen years ago, the simulation capabilities of the day probably would not have scaled to handle so many cores and channels. However, a quick back-of-the-envelope calculation shows that today we would be able to fit somewhere around sixteen CPU cores under the area of a HBM stack. These sixteen cores with sixteen memory channels ends up matching up with the ratio one core per channel (four cores, four memory channels) assumed in the original paper.

The original paper also assumed “true” 3D memory structures where individual channels were vertically organized (similar to the concept of “vaults” introduced in Hybrid Memory Cubes [18]) and had already been proposed [22] earlier. Such true-3D memory organizations have failed to catch on thus far, where the dominant HBM organization is effectively a 3D stack of multiple memory die that are each internally 2D in naature. Maybe this will eventually change, but for now HBM remains the dominant option.

The ISCA 2008 paper also proposed a few other microarchitectural tricks including mechanisms to scalably increase the number of outstanding misses that the memory controllers could track as well as additional buffering also known as “Cached DRAM” [6] to further improve performance. The exact mechanisms have not caught on, but the underlying observation was that modifications to the multi-core architecture are needed to better extract the performance potential of a high-bandwidth memory system. In hindsight, this could be viewed as a natural consequence of Little’s Law, but this lesson of co-optimizing the end-to-end datapath from processor to memory holds to this day.

[1] This work was originally conducted while at the Georgia Institute of Technology.

3 Where's my 3D-Stacked Memory?

Multiple layers of 3D-stacked DRAM on top of a multi-core processor as originally envisioned in the ISCA 2008 paper has not yet materialized in any mainstream commercial offerings. There are several factors to this. 2.5D silicon interposer technology [4] arose as an effective means to integrate HBM in the same package as the computing logic. This enabled a large increase in available memory bandwidth while side-stepping some of the challenges of directly stacking the memory on top of the processor, such as thermals and the overheads of perforating the CPU die with enough through-silicon vias (TSVs) to deliver power, ground, and signals to the memory. Note that the passive silicon interposer is still a fundamentally 3D structure in that it includes its own TSVs (for power and IO connections to the package substrate) and the HBM and processor die are 3D-stacked on top of the interposer. It is just not the full 3D active-on-active (i.e., DRAM on processor) organization originally proposed.

Another potential challenge is that the typical CPU pipeline has primary been optimized for mainstream memory interfaces such as DDR and LPDDR, whereas HBM (letalone a true-3D DRAM architecture) can provide an order of magnitude more bandwidth. Even with modern processors supporting multiple tens of cores, there would still likely be a need for a substantial rearchitecting of the core to handle so much more bandwidth.

In the fifteen years since the paper's publication, the compute environment has become much more heterogeneous, with the rise of GPUs, FPGAs, AI/HPC accelerators, and other neural and tensor processing architectures. Compared to these, multi-core CPU memory requirements are relatively modest, and so the focus for HBM integration has targeted those platforms where the pressure for more bandwidth and energy efficiency was the greatest. Multi-core CPUs with HBM (whether 2.5D or 3D integrated) are now only starting to appear in some commercial deployments [20], and even so this is still different from the original concept in the paper of 3D stacking the memory on top of the CPU.

4 Conclusions

Despite the fact that memory-on-CPU architectures have not yet become a commercial reality, I believe that the lasting impact of this ISCA 2008 paper is that it hopefully inspired many researchers in our community to start thinking about the possibilities of 3D integration, which today really has expanded into a wider menu of advanced packaging technologies and heterogeneous integration. Even after a decade and a half, the possibilities for what we could potentially do with a combination of chiplets, 2D, 2.5D, and 3D integration are immense and seemingly continuing to expand, and it is gratifying and humbling to have had the opportunity to contribute to our research community's journey down this path.

Acknowledgments

Immense thanks go to Dr. Bryan Black who first introduced me to what were then the wild new ideas of 3D-stacked silicon, which forever changed the course of my research career. Much gratitude also goes out to the College of Computing at the Georgia Institute of Technology under which this work was done, with an additional thanks to Korea University where I actually wrote the ISCA 2008 paper while teaching abroad.

AMD, the AMD Arrow logo, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

References

[1] Advanced Micro Devices, Inc. AMD Data Center and AI Technology Premier, https://www.amd.com/en/solutions/data-center/data-center-ai-premiere.html, June 2023.
[2] Advanced Micro Devices, Inc. Introducing RDNA Architecture, https://www.amd.com/system/files/documents/rdna-whitepaper.pdf.
[3] B. Black et al. Die-Stacking (3D) Microarchitecture, MICRO 2006.
[4] Y. Deng and W. Maly, "Interconnect Characteristics of 2.5-D System Integration Scheme," in Intl. Symp. on Physical Design, April 2001.
[5] M. Ghosh and H.-H. S. Lee. Smart Refresh: An Enhanced Memory Controller Design for Reducing Energy in Conventional and 3D Die-Stacked DRAMs, MICRO 2007.
[6] H. Hidaka, Y. Matsuda, M. Asakura, and K. Fujishima. The Cache DRAM Architecture. IEEE Micro Magazine, 10(2):14–25, 1990.
[7] T. H. Kgil, S. D'Souza, A. G. Saidi, N. Binkert, R. Dreslinski, S. Reinhardt, K. Flautner, and T. Mudge. PicoServer: Using 3D Stacking Technology to Enable a Compact Energy Efficient Chip Multiprocessor, ASPLOS 2006.
[8] D. U. Lee et al. A 1.2V 8Gb 8-channel 128GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29nm process and TSV, ISSCC 2014.
[9] C. C. Liu, I. Ganusov, M. Burtscher, and S. Tiwari. Bridging the Processor-Memory Performance Gap with 3D IC Technology, IEEE Design and Test of Computers, 22(6):556–564, 2005.
[10] G. H. Loh, M. D. Hill. Efficiently Enabling Conventional Block Sizes for Very Large Die-stacked DRAM Caches, MICRO 2011.
[11] G. H. Loh et al. A Research Retrospective on AMD's Exascale Computing Journey, ISCA 2023.
[12] G. L. Loi, B. Agarwal, N. Srivastava, S.-C. Lin, and T. Sherwood. A Thermally-Aware Performance Analysis of Vertically Integrated (3D) Processor-Memory Hierarchy, DAC 2006.
[13] N. Madan and R. Balasubramonian. Leveraging 3D Technology for Improved Reliability, MICRO 2007.
[14] M. Meswani, S. Blagodurov, D. Roberts, J. Slice, M. Ignatowski, G. H. Loh. Heterogeneous Memory Architectures: A HW/SW Approach for Mixing Die-stacked and Off-package Memories, HPCA 2015.
[15] S. Mysore, B. Agarwal, S.-C. Lin, N. Srivastava, K. Banerjee, and T. Sherwood. Introspective 3D Chips, ASPLOS 2006.
[16] S. Naffziger, N. Beck, T. Burd, K. Lepak, G. H. Loh, M. Subramony, S. White. Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families, ISCA 2021.
[17] D. Nelson, C. Webb, D. McCauley, K. Raol, J. Rupley, J. DeVale, and B. Black. A 3D Interconnect Methodology Applied to iA32-class Architectures for Performance Improvements through RC Mitigation, Intl. VLSI Multilevel Interconnection Conf. 2004.
[18] J. T. Pawlowski. Hybrid Memory Cube: Breakthrough DRAM Performance with a Fundamentally Re-Architected DRAM Subsystem, Hot Chips 2011.
[19] M. Qureshi, G. H. Loh. Fundamental Latency Trade-offs in Architecting DRAM Caches, MICRO 2012.
[20] M. Sato, Y. Ishikawa, H, Tomita, Y. Komada, T. Odajima, M. Tsuji, H. Yashiro, M. Aoki, N. Shida, I. Miyoshi, K. Hirai, A. Furuya, A. Asato, K. Morita, T. Shimizu. Co-Design for A64FX Manycore Processor and "Fugaku", SuperComputing (SC) 2020.
[21] M. J. Schulte et al. Achieving Exascale Capabilities through Heterogeneous Computing, IEEE Micro Magazine, 35(4):26-36, 2015.
[22] Tezzaron Semiconductors. Leo FaStack 1Gb DDR SDRAM Datasheet, August 2002.
[23] T. Vijayaraghavan et al. Design and Analysis of an APU for Exascale Computing, HPCA 2017.
[24] W. A. Wulf and S. A. McKee. Hitting the Memory Wall: Implications of the Obvious, Computer Architecture News, 23(1):20–24, 1995.