Multi-Chip Technologies to Unleash Computing Performance Gains over the Next Decade

Lisa T. Su, Samuel Naffziger, and Mark Papermaster
Advanced Micro Devices (AMD) Inc., Sunnyvale, CA USA

Abstract — **Datacenter and high-performance computing capabilities have continued their exponential improvements in performance over the prior decade, driven by the proliferation of devices and data through the internet of things (IoT), and new applications in the enterprise and cloud. This trend will continue over the next decade as the demand for compute performance continues to grow with exabytes of data being created daily and new use models incorporating machine learning and artificial intelligence become more prevalent. As Moore's Law has slowed in recent years, numerous techniques including system, architectural and software innovation have been used to extend the high-performance processor performance improvements. We examine these techniques and demonstrate that although some of these will continue, new innovations are needed especially at the system level to continue the performance trend over the next decade. We believe that multi-chip technologies and system level innovations are key to unlocking the performance gains in computing over the next decade.**

I. Introduction

From smart phones to smart homes, autonomous vehicles and intelligent personal devices, billions of new devices are being deployed every year with new services and features that are driving greater demand for high-performance silicon. These devices are creating new capabilities and an explosion of data that is driving the need for a broad expansion in computing capabilities at both the device and enterprise level. With the rapid increase of computing horsepower and new artificial intelligence (AI) and machine learning capabilities, devices and servers are increasingly able to sense, act upon, and predict our needs transforming the way we live, work, communicate and play.

These trends indicate that the demand for high-performance computing will only accelerate going forward. This paper will examine the trends that have driven high-performance computing over the last 10 years and focus on key areas of innovation for continuing to drive advances in computing for the next 10 years.

II. CPU and GPU Performance Trends

High-performance computing requires both general-purpose processing which is usually accomplished by CPUs, and special-purpose processing often accomplished by GPUs, FPGAs, or application-specific accelerators. For the purposes of this paper, we will focus on the trends in CPUs and GPUs as they cover a significant portion of the high-performance computing space.

In the personal computing era, computing capability doubled every 1.5 years and similarly computing power efficiency doubled every 1.5 years [1]. As this trend has slowed in recent years due to a slowing of transistor scaling [2], numerous techniques including system, architectural, and software innovation have been used to extend the performance improvements [3].

A common metric that tracks general purpose compute capability is represented by dual-socket (2P) mainstream x86 server CPU performance measured by the SPECint_rate2006 benchmark. Fig. 1 shows that for top-of-stack servers at introduction, the performance doubles every 2.4 years (data from SPEC.org). Similarly, if you look at power efficiency of these systems at peak system output efficiency in Fig. 2, we also see the same rate of doubling every 2.4 years.

In recent years, the application of GPU technology, with highly-optimized and parallel vector architectures, has become increasingly more prevalent for compute applications [4]. The FLOPS rate (or floating point operations per second) for GPUs is typically an order of magnitude (or more) higher than a general purpose CPU, and is a good proxy for the GPU performance capability. Fig. 3 shows top-of-stack GPU FLOPS rates at time of introduction over time following a very similar, but somewhat faster trend, of doubling every 2.1 years, based on AMD-internal analysis of top-of-stack industry GPUs.

III. Technology and Design Trends

To understand the significant innovation that has occurred in high-performance computing architectures over the last 10 years, we examined the contributions of technology scaling, power scaling, die size increases, power management advances, microarchitectural improvements, and software.

It has been well discussed that Moore's Law driven gains in performance and energy efficiency are slowing [5]. Looking at the delivered gains garnered from the metrics of transistor speed, density and switching capacitance of high-volume technologies over the same time frame, we see an improvement rate in density and in energy efficiency that doubles every 3.6 years on both metrics based on AMD analysis of leading-edge industry process technologies (Fig. 4). This 3.6 years is a bit longer (slower improvement) than noted by the ITRS roadmap [6] of ~3.0 years due to our inclusion of product-level implementation realities around wire scaling, cell utilization, variation effects and other manufacturing issues that dilute the raw device improvements. This means that in the 2.1 to 2.4 year time span when the performance increased 100%, 35% (GPU) to 40% (CPU) of the improvement was due strictly to process technology.

Power management has been another significant contributor to processor performance improvements over the last decade. As high-performance processors have essentially become power constrained and manufacturing variation has become a larger factor in determining overall power consumption, a significant focus has been placed on new circuit and adaptive approaches to control on-chip power. These algorithms have become very sophisticated, including boost and sensor-based methods which are able to adapt to workloads and enable more performance in each generation [7, 8]. Essentially, power management technology for peak output is working to eliminate wasted energy during computation. Voltage variations are compensated out, temperature margins are adapted out, and manufacturing variations are tracked and adapted across chip and chip to chip [9]. Boost technologies track workload behavior and apply power when performance is required. Substantial innovations here have yielded generational performance gains in the 15% range [10]. These approaches are now table stakes in modern processor design (Fig. 5) accounting for almost a 50% reduction in net power in a complex system-on-chip (SoC) as compared to a similar design without this capability from a decade ago.

In addition to power management, we have also allowed the absolute power of server CPUs and GPUs to increase at a rate of about 7% per year over time as shown in Fig. 6. This power budget is additive to the gains from power management and has been used to deal with lagging memory performance (caches), higher levels of system connectivity (I/O), as well as supporting more computation. It's important to note that for typical servers, the fraction of power devoted to computation is about 1/3rd of the total with most of the rest consumed with data movement and caching.

Technology density improvements have enabled larger die sizes and increased parallelism through additional core count for both CPUs and GPUs. Adding more CPU cores (or GPU vector engines) does deliver more throughput performance but due to memory, I/O, and other overheads, there is about a 70% efficiency in translating more cores into performance for workloads like SPECint_rate2006. Fig. 7 shows the die size progression for both CPU and GPUs. This extra die area is being used for both additional cores as well as additional area for memory and I/O improvements to fully utilize the compute capability. Approximately half of that die size increase is used to compensate for reduced memory bandwidth per operation (Fig. 8). When coupled with the aforementioned 70% efficiency from core counts, we estimate the benefit of the die size increase to be between 9% and 15% of the total improvement in historical performance.

Finally, we have also examined the rate of improvement of microarchitectural and software improvements. When we look at SPECint_rate2006, there is roughly a 17% increase in instructions per cycle removing compiler improvements as shown in Fig. 9 over the 2.4 year period. Finally, compiler software improvements contribute approximately 2-5% of the annual performance gains for CPUs.

To summarize these trends, significant innovation has occurred in every aspect of processor design to continue the high-performance processor improvements over the last 10 years. Fig. 10 summarizes these components including technology scaling, power management, increased thermal budget, die sizes, and microarchitectural and software improvements.

IV. System Level Integration

We next explore the key trends for future processor design. A number of the techniques we used to enhance performance over the last 10 years are reaching their limits and require additional innovation at the system level. Transistor scaling continues to be very important; however, increases in die size, power management improvements, and increased thermal budgets have limited headroom for continued improvement.

We note previously that increased die size is very beneficial to improve performance through increased cores and vector units. We also need the additional area to mitigate memory bandwidth and latency, increase instructions per cycle, and other I/O. However, the cost per mm^2 in advanced technology nodes has continued to increase with each node as the technology complexity has increased. In addition, yields are poor for very large die sizes as we reach the limits of the reticle field.

A. "Chiplet" Approach Using Multi-Chip Modules

To address this problem at the system level, we have recently introduced an innovative multi-die "chiplet" architecture on our AMD EPYC server product [12].

In this case, we build a 32-core server die using 4 discrete "chiplets" combined in a multi-chip module with a combined area of 852 mm^2 which exceeds the maximum printable die area. In addition to providing more functionality than a monolithic die, splitting a large die into multiple smaller components has the added benefit of improving overall yield. Since the smaller die yields significantly better than a single large die, we estimate that manufacturing costs are reduced by approximately 40% when using mature defect densities even considering the 10% overhead that is required for the interconnect between the die (Fig. 11). Although multi-chip modules are not new, enabling the scalability required by high-performance processors with chiplets necessitates innovations such as those in EPYC, including the high-speed Infinity fabric, per-core voltage and distributed power management.

We believe the use of multi-chip modules in high-performance computing will continue to be a significant performance lever going forward. There are many creative ways to partition a large die that exploit different technology nodes for different elements of the monolithic design such as using prior-node process technology for the analog I/O components and specialized nodes for memory components.

B. 2.5D Packaging with High-Bandwidth Memory

Another example of system level innovation, is the use of high-bandwidth memory (HBM) integration in high-performance GPUs. This technique was first introduced by AMD in

2015 [17] and has continued in subsequent products including our recently introduced Vega products. Memory bandwidth is a critical factor in improving performance at the system level which has been lagging the performance gains (Fig. 8). By shortening the distance between memory and GPU from 40mm+ to within 1 mm (Fig. 12), we greatly reduce the energy per bit accessed from memory. In addition, the greater density of the micro-bump and interposer connections enable a lower frequency and simpler electrical protocol than the high-speed signaling required by traditional GDDR5 memory. These improvements contribute to a 4X reduction in energy per memory bit access (Fig. 13), providing both performance and power advantages.

C. 3D Die Stacking

As we move forward, the next level of system level innovation will likely involve 3D stacking. 2D and 2.5D approaches have many merits, but there are also some limitations going forward.

The 2D multi-chip module (MCM) approach improves total functional die area; however, there is extra power due to the high-speed I/O that is used to interconnect the chiplets. Similarly the 2.5D HBM approach delivers a significant step function improvement in memory bandwidth by bringing the memory closer to the processing unit; however, 2.5D has limitations in interface area, density and cost and there is an insatiable demand for more memory closer to the processor.

A promising future innovative approach is the use of 3D die stacking as shown in Fig. 14 to address some of these issues. Vertical stacking of processor, memory, and other components can potentially reduce the I/O power interconnecting the die and allow significantly more system level integration. Although these techniques have been demonstrated experimentally, they have yet to reach high-volume production given challenges in through-silicon via (TSV) technology, heat removal, and cost-effective high volume assembly. Further technology advances that improve thermal conductivity in 3D die stacked solutions, and potentially micro-fluidic approaches to extract heat at the micron level, are key areas for innovation, as well as holistic design-for-thermals approaches that optimize TSVs, micro-bumps and bonding methods to improve thermal conductivity [13]. We believe this is an area where significant focus will be very beneficial for future advances in the high-performance processor space.

V. Device Innovation

Last, but definitely not least, device technology that reduces the energy per operation while maintaining performance is required and needs to remain a key focus. We have relied on improved transistor performance for so many years, there is no amount of design or system-level techniques that can replace improvements in device performance. We look forward to the many promising options that are being actively worked and presented in this conference such SiGe trigate [14], and gate all around (GAA) devices [15].

While good logic devices are critical, communication off-chip consumes a large portion of the power budget as already shown in Fig. 6. This leads to the need for device innovation in the analog I/O space to both reduce power and continue the increase in signaling speeds.

Finally, lagging memory capability is an increasing source of power and performance challenge. 2.5D and 3D integration approaches mitigate this, but provide just a single-step improvement. The underlying technology needs to see improvements in density, cost per bit, and energy per bit to enable further gains. A number of non-volatile technologies such as PCM, MRAM, and ReRAM show promise in this regard [16] and need to reach maturity.

VI. Conclusions

There has been a very impressive exponential growth in compute capability in the industry, and the demand for this performance continues to accelerate. This growth has been achieved over the last decade through a number of innovations including Moore's Law driven technology improvements, improvements in power management, increases in absolute die size and power budget, and architectural and software efficiency.

As exciting as it has been, we need additional innovation over the next decade to continue this trajectory of performance improvement and satisfy the needs of high-performance computing. We see system-level integration as a very key and highly valuable field to enable enhancing performance through 2-D MCM technology, 2.5D packaging technology, and 3D stacking technology. We also believe that device innovation will remain critical to continuing the performance improvements and this includes transistor improvements in the logic, memory, and analog I/O space.

Acknowledgment

The authors gratefully acknowledge the global AMD engineering teams for contributing to the performance improvements over the past decade, and driving innovative solutions that lay the foundation for the next generation of high-performance computing.

References

[1] Jonathan Koomey et al, "Implications of Historical Trends in the Electrical Efficiency of Computing," *IEEE Annals of the History of Computing*, Vol. 33, Issue 3, pp. 46–54, 2011.

[2] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, Vol. 9, Issue 5, pp. 256–268, 1974.

[3] J. Koomey, S. Naffziger, "Moore's Law Might Be Slowing Down, But Not Energy Efficiency," *IEEE Spectrum* [Online], March 2015.

[4] John D. Owens et al, "A Survey of General-Purpose Computation on Graphics Hardware," *Computer Graphics Forum* [Online], Vol. 26, No. 1, pp 80—113, 2007.

[5] Thomas N. Theis; H. -S. Philip Wong, "The End of Moore's Law: A New Beginning for Information Technology," *Computing in Science & Engineering*, Vol. 19, No. 2, pp 41-50, 2017.

[6] ITRS 2013 data as tabulated in https://purl.stanford.edu/gc095kp2609

[7] A. Grenat et al, "Increasing the performance of a 28nm x86-64 microprocessor through system power management," IEEE International Solid-State Circuits Conference (ISSCC), pp 74 – 75, 2016.

[8] Y. Zu et al, "Adaptive guardband scheduling to improve system-level efficiency of the POWER7+" 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp 308–332, 2015.
[9] S. Sundaram et al, "Adaptive Voltage Frequency Scaling Using Critical Path Accumulator Implemented in 28nm CPU," 2016 29th International Conference on VLSI Design and 2016 15th International Conference on Embedded Systems (VLSID), pp. 565–566.
[10] S. Sundaram et al, "Bristol Ridge: A 28-nm times 86 Performance-Enhanced Microprocessor Through System Power Management," *IEEE Journal of Solid-State Circuits*, Vol. 52, No. 1, pp 89 – 97, 2017.
[11] D. Kanter, "Adjusting SPEC CPU2006 Scores, Modifications Keep Benchmark Relevant for Servers," *Microprocessor Report*, October 2016.
[12] K. Lepak et al, "The Next Generation AMD Enterprise Server Product Architecture," IEEE Hot Chips 29 Symposium, 2017.
[13] H. Oprins et al, "Experimental Characterization of the Vertical and Lateral Heat Transfer in 3D-SIC Packages," ASME 2015 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems collocated with the ASME 2015 13th International Conference on Nanochannels, Microchannels, and Minichannels. Vol. 1: Thermal Management San Francisco, California, USA, July 6–9, 2015.
[14] R. Xie et al, "A 7nm FinFET technology featuring EUV patterning and dual strained high mobility channels," 2016 IEEE International Electron Devices Meeting (IEDM), pp. 2.7.1 - 2.7.4.
[15] M. Guillorn et al, "Density scaling beyond the FinFET: Architecture considerations for gate-all-around CMOS," 2016 74th Annual Device Research Conference (DRC), pp. 1.
[16] S. Kosuke, S. Swanson, "A Survey of Trends in Non-Volatile Memory Technologies: 2000-2014," IEEE International Memory Workshop (IMW), pp. 1-4, 2015.
[17] J. Macri, "AMD's next generation GPU and high bandwidth memory architecture: FURY," IEEE Hot Chips 27 Symposium (HCS), 2015.

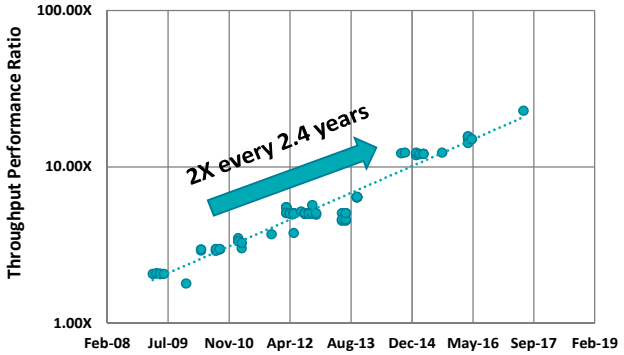

Fig 1. SPECInt_rate2006 2P server performance trend over time

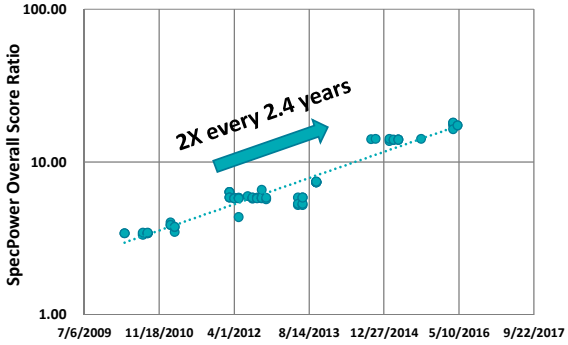

Fig 2. Performance-per-watt trend over time in server CPUs

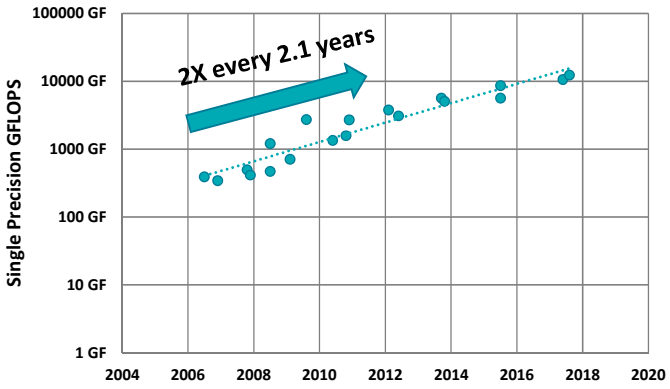

Fig 3. GPU single precision floating point operations per second trend

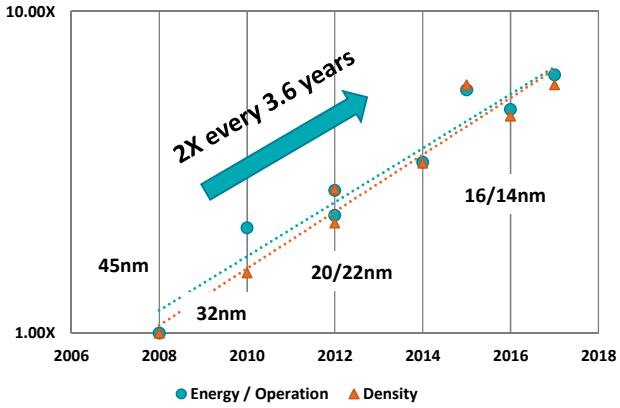

Fig 4. Technology energy efficiency and density across process nodes

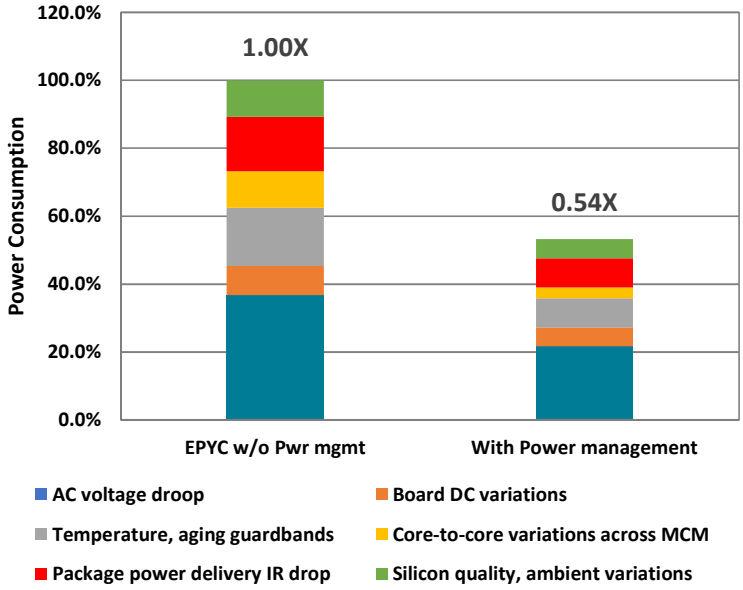

Fig 5. Power management gains for AMD EPYC Server Processor

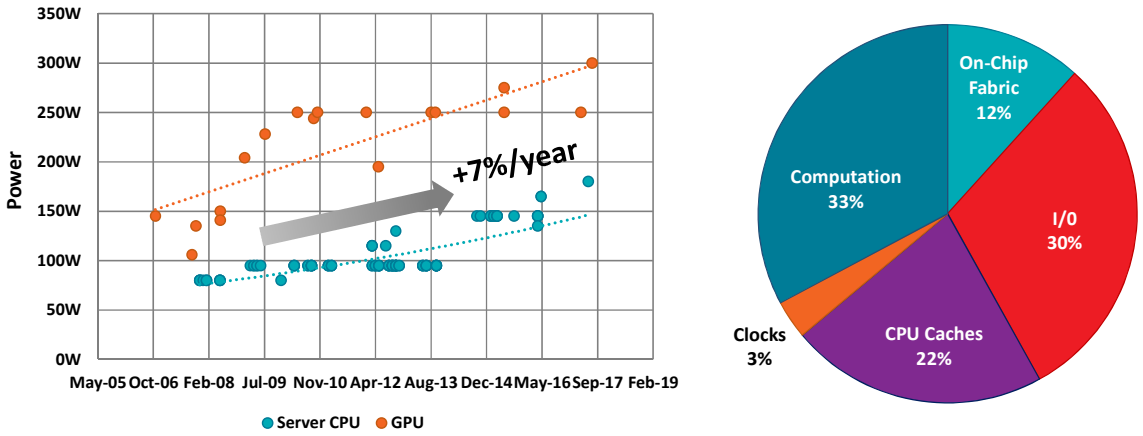

Fig 6. A) Thermal design power over time in server CPU and GPUs. B) Typical server power breakdown 2017

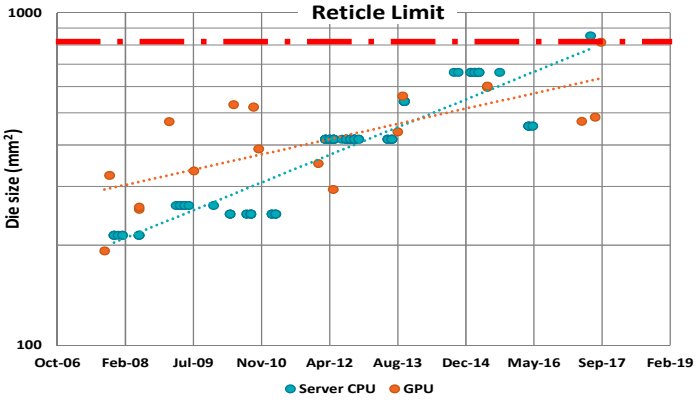

Fig 7. Die size increases over time in server CPU and GPUs

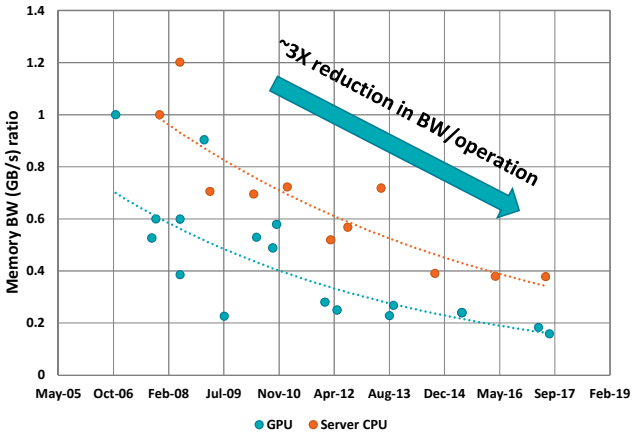

Fig 8. Memory bandwidth per operation in server CPU and GPUs

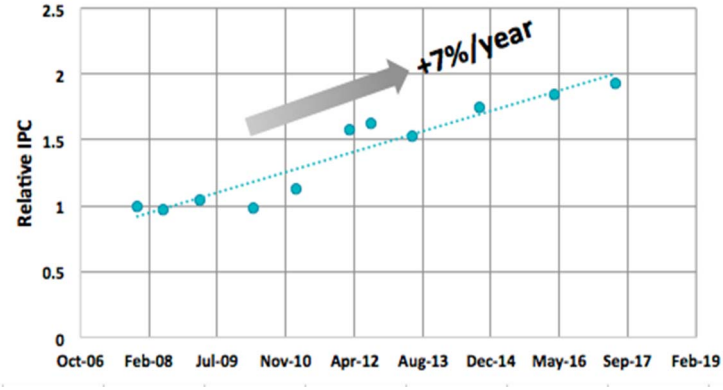

Fig 9. Server Performance / Core / MHz

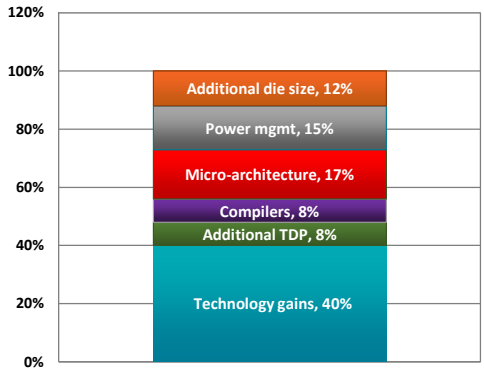

Fig 10. Elements of 2X in 2.4yr performance gain

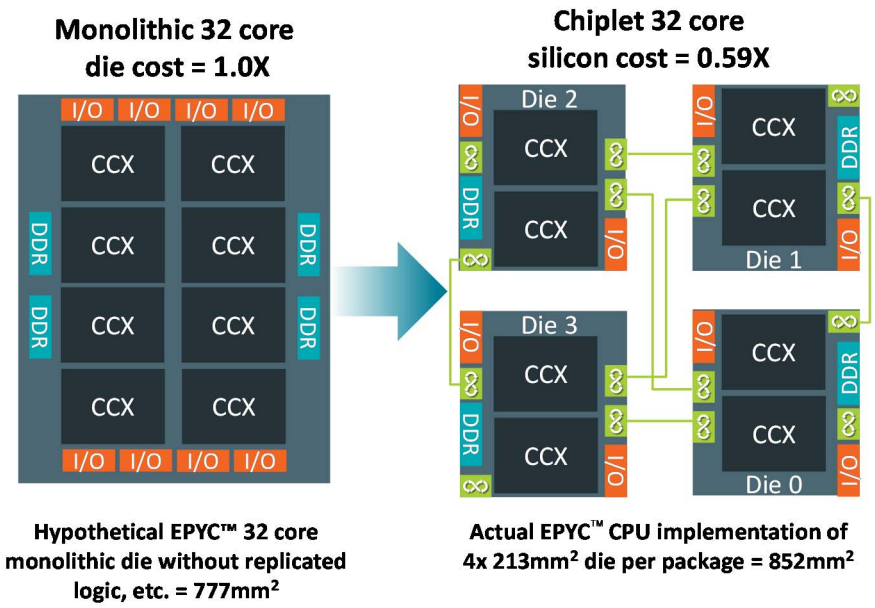

Fig 11. Benefits of using a 4-die chiplet approach versus a monolithic die approach

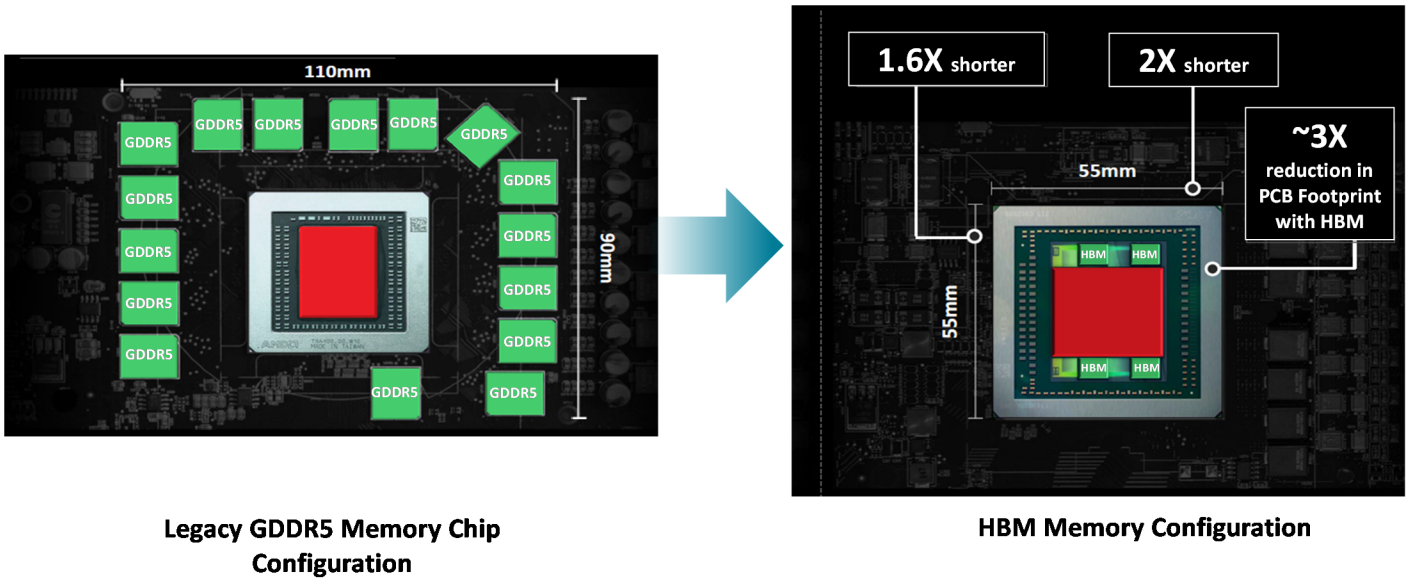

Fig 12. HBM integration brings memory closer to the processing unit and reduces overall footprint.

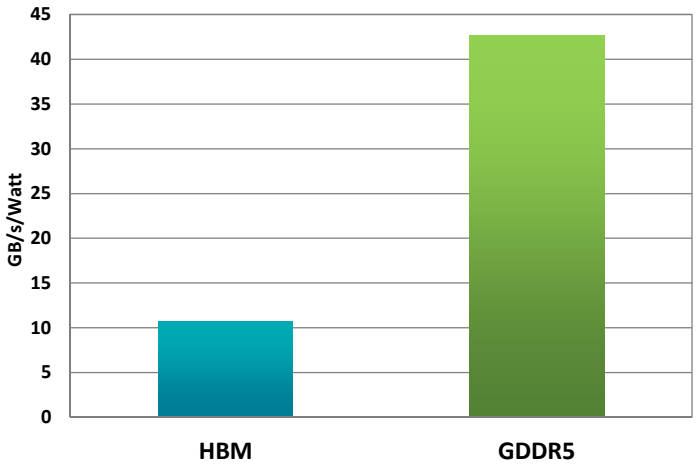

Fig 13. HBM 2.5D provides both higher bandwidth and lower power.

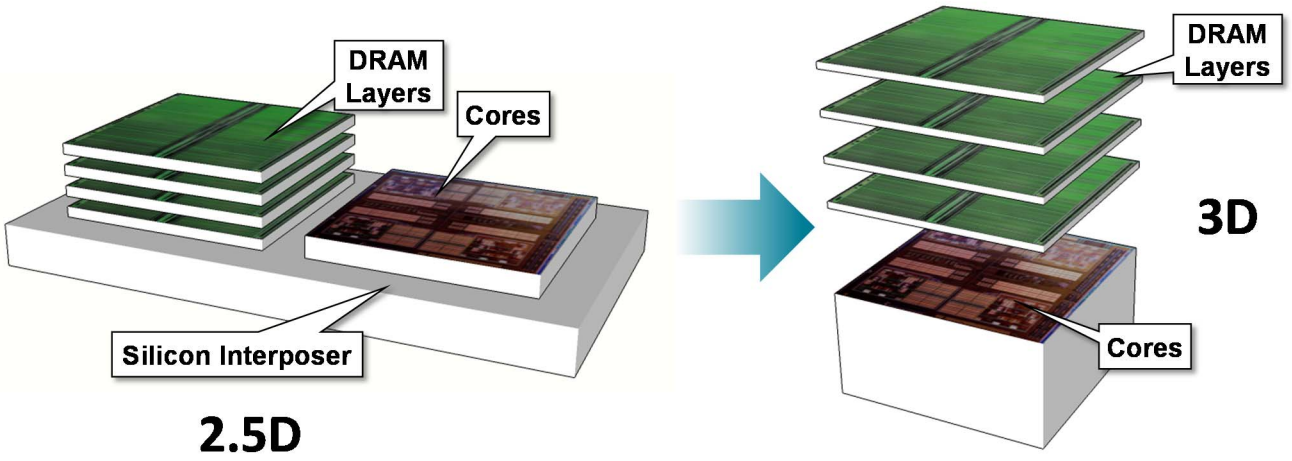

Fig 14. 3D memory stacking enables additional power savings and density gains.