The next frontier: Enabling Moore's Law using heterogeneous integration

By Raja Swaminathan [AMD]

The explosion of connected devices over the last 40 years in our industry has driven an explosion of semiconductor content riding the back of Moore's Law. Starting with the personal computer cycle then continuing with smartphones and Internet of Things (IoT) devices, silicon has permeated every aspect of our lives. This explosion of semiconductor content in everything from devices in our pockets to integrated into our clothing has led to the birth and growth of a new era of high-performance computing (HPC), as all the data being generated is processed into useful information to improve our lives.

From the cloud to the edge, and from artificial intelligence (AI) to 5G communications, the insatiable demand for HPC has become a driving force within the microelectronics industry and it will shape the next several generations of technology and design innovation. The demand for compute is accelerating rapidly with the doubling of HPC system performance every 1.2 years. This trend is much faster than Moore's Law, which currently has slowed to doubling of transistor density every 2-3 years; so, the compute capability is clearly driven by innovations outside of the raw silicon. The demand for HPC is not simply bragging rights of being in the top 500 supercomputers. These devices are solving problems that are pressing for humanity including drug discovery, climate models, new energy exploration and many more. Today's best compute platforms only whet our appetite for more as the possibilities for solution finding become more compelling.

Demand for computation is outpacing Moore's Law

The next bit of sobering data regards the much-discussed cracks in Moore's Law. As we know, silicon technology node introductions have been slowing down, and simultaneously delivering less benefit, while at the same time, the costs per yielded mm^2 of silicon are going up. This is particularly challenging because the semiconductor industry has thrived on delivering more performance and features in each generation by adding transistors. With these trends, the cost per transistor will stop scaling in the next few years, which creates notable economic headwinds to meeting the demand. These costs are not just a result of inflationary pressure but based on the underlying physics and complexity of these new nodes.

The next aspect of the slowdown in node introductions is that scaling factors are diverging between different intellectual property (IP) types, with static random-access memory (SRAM) and especially analog circuits lagging well behind the scale factors of logic. This leads to the chip-level view of area scaling where, with a mix of logic, SRAM and analog content, we will not be able to shrink chip designs appreciably toward the end of this decade. This illustrates that the irresistible force of compute demand is colliding with the immovable object of device physics, creating an environment where new architecture approaches and non-device innovations are critical for our ecosystem.

It is now recognized that conventional computing is approaching fundamental limits in energy efficiency. Historical trends show that general purpose CPU energy efficiency worsens with higher performance, so new approaches are required (**Figure 1**). We are also finding new approaches to reduce energy for compute. Modular design, chiplets and 3D

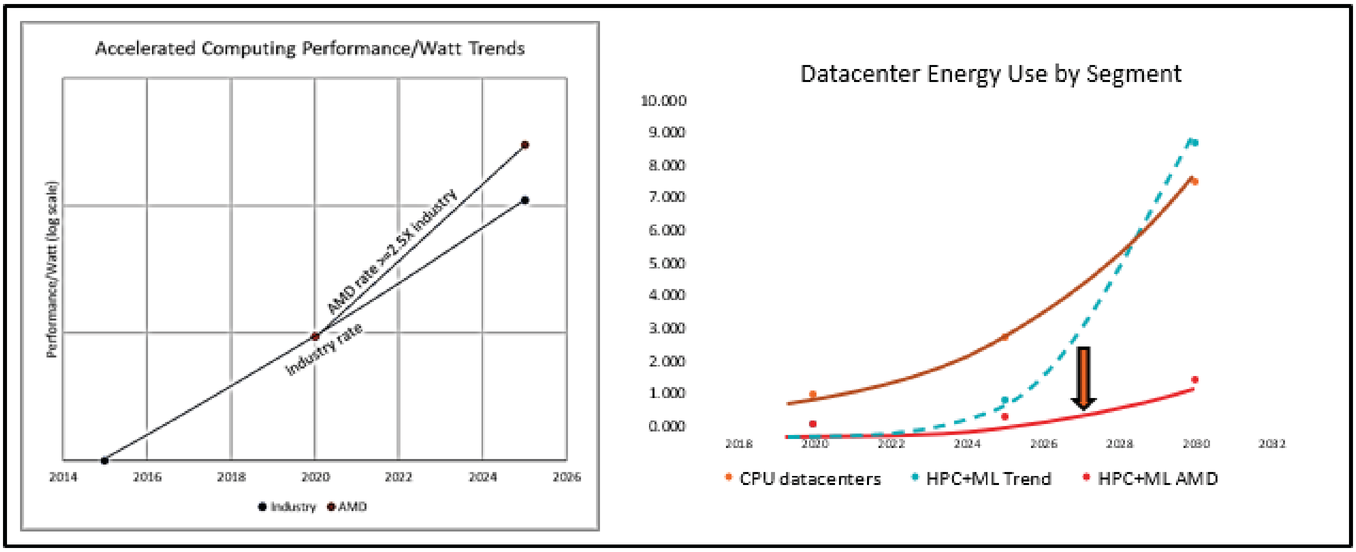


Figure 1: AMD's efficiency goal for HPC/AI applications.

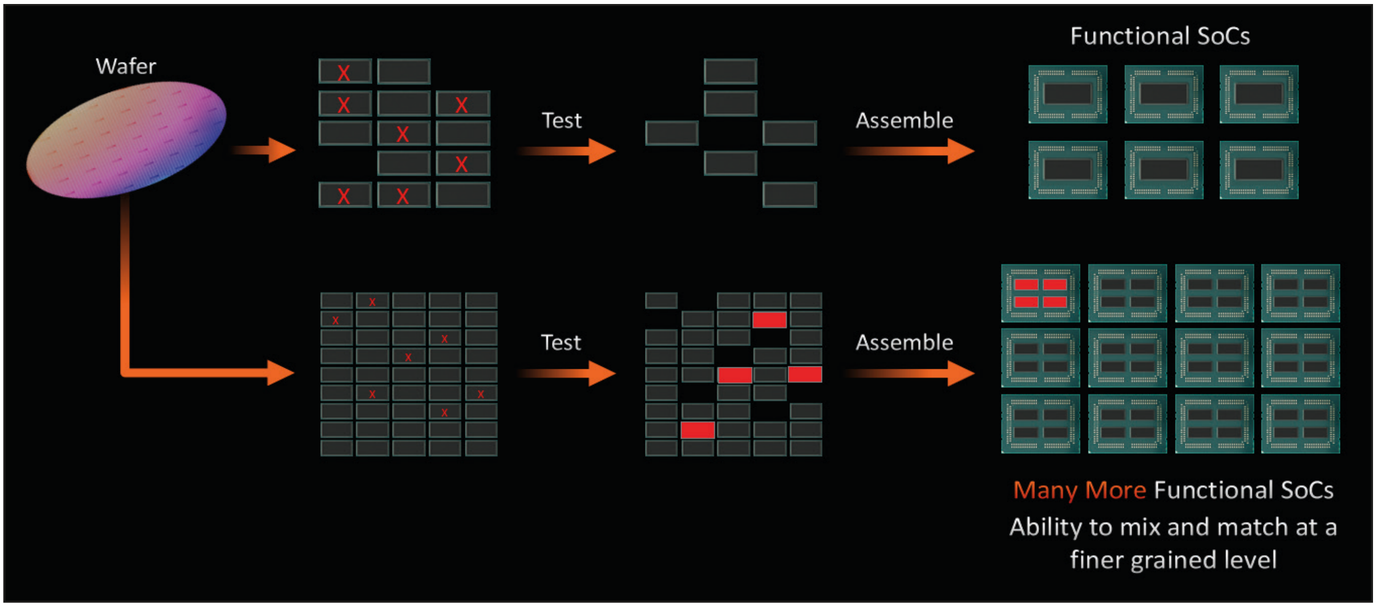

Wafer
Test
Assemble
Functional SoCs
Test
Assemble
Many More Functional SoCs
Ability to mix and match at a finer grained level

Figure 2: High-level approach to chiplets.

stacking are the next frontier for efficiency gains. Application-specific optimization provides better performance-per-Watt. The last five years show an industry efficiency improvement rate of 12X for HPC and AI nodes. The AMD goal is to dramatically accelerate this improvement rate to 30x by 2025.

So, we have the bright future of exploding compute demand and simultaneously the dark cloud of technology headwinds. The trillion-dollar question is how to architect, design, and build future systems that solve these challenges. The answer is increasingly clear that modular, multi-chip design is a fundamental enabler. Systems must be more specialized for the task they are running. General purpose is no longer generally applicable. We need efficient accelerators, and we need economically viable ways to continue to deliver this performance in the face of the formidable cost trends. Let us take a closer look at what modular design can do, and what the enabling technology requirements are.

Let us start off with the magic of chiplet-based design, which is becoming much more pervasive. AMD led the way in this approach with our heterogeneous technology server and desktop products back in 2019. An initial motivation for chiplets was economics. Back in the day, Moore's Law enabled a doubling of transistors and capability in each generation, and all was good. Lately, this has not worked out as well. With shrink factors slowing down while compute demand has not, die sizes have been growing at an unsustainable rate.

With chiplets, we can split a formerly monolithic system-on-chip (SoC) into two components to improve performance. However, this results in a non-trivial overhead associated with "chipletizing" the design. Each die needs test capability, power management, and an interface so it can talk to the other chiplets. These interfaces will not be as small, low latency, or power efficient as on-die wires; therefore, the architecture needs to accommodate new boundaries and complexity.

To illustrate benefits of the chiplet approach, let us consider the yield dynamics. With a single large die and a fixed number of defects on a wafer, we yield a small set of functional SoCs for a wafer's worth of chips (Figure 2). As soon as we split that big SoC up into, say, four chiplets, the yield dynamics start to work in our favor. The same number of defects now just take out a small chiplet, and we can use our wafer sort capabilities to select the good ones and build more functional SoCs from the same silicon. This is one factor that has helped AMD to meet market demand better when wafer supplies are so constrained.

We also gain the flexibility of building chiplet SoCs with varying numbers of chiplets to address different markets. Perhaps a less obvious benefit is that we can cherry pick faster chiplets from the wafer and assemble them into higher-performance and higher-priced SoCs for customers who want, and will pay for, the greatest performance possible.

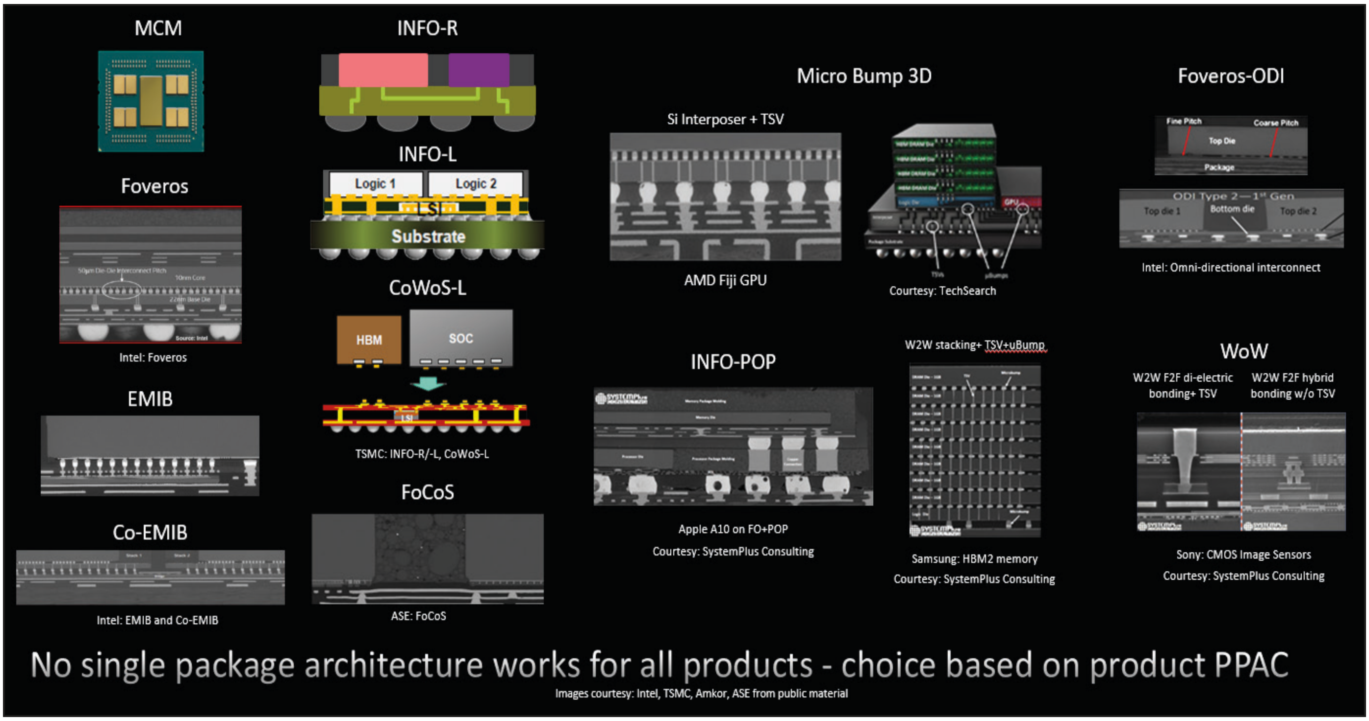
The benefits described above are substantial, though modular design is bigger than just decomposing an SoC into chiplets. We want to build tailored products for specific markets by mixing and matching chiplet types. Some chiplets can be general purpose CPUs, others can be more specialized. We can now specialize a domain-specific chiplet and include more or fewer of them for a given product. However, the success of this approach is heavily dependent on the package technologies used to assemble these dice and enable them to communicate with each other.

Package architectures

Many package architectures exist in the industry to enable die-to-die interconnections across various product segments (e.g., mobile, PC, server, and desktops) (Figure 3).

Examples include:

- Multi- chip module (MCM) architectures from AMD and other industry players.
- Other 2D architectures based on redistribution layer (RDL)-like interconnects (or 2D-organic) like integrated fan-out with redistribution layer (INFO-R), and fan-out chip-on-substrate (FoCoS).
- 2D silicon-based architectures like embedded multi-die interconnect bridge (EMIB), AMD's elevated fan-out bridge (EFB), Integrated fanout with integration of an LSI (INFO-L), and Si interposer where the die-to-die interconnect is achieved using passive Si; as well as
- 3D architectures — defined as active-on-active Si stacking, such as the AMD 3D V-Cache™, Foveros/



No single package architecture works for all products - choice based on product PPAC

Images courtesy: Intel, TSMC, Amkor, ASE from public material

Figure 3: Sample package architecture options for die-to-die chiplet interconnects.

omni-directional interconnect (ODI), wafer-on-wafer (WoW) architecture found in image sensors and the memory markets.

Chiplet package architecture choice is not a one size fits all approach, rather it is made based on specific power, performance, area, and cost (PPAC) requirements per product. A critical dimension of making this all work is driving the overhead of those interfaces down. One way to quantify this is to tabulate the linear interconnect density and the areal interconnect density of packaging approaches (**Figure 4**).

MCMs are great, low-complexity designs, but the low connection density of this technology limits its applications to specific boundaries for the chiplets. For instance, the AMD EPYC™ and Ryzen™ lines chose to put the CPU cores on one chiplet and the IO and memory interfaces on another one. This works with MCM because the CPU bandwidth requirements are relatively modest and can be supplied across highspeed SERDES routes.

To accomplish more exotic SoC chiplet configurations, higher bandwidths are required. In the middle of **Figure 4** is an example Radeon Instinct™ design, which requires high-bandwidth memory to feed the compute engines. To supply over a terabyte per second of bandwidth to memory, a higher density interconnect is required. We chose passive silicon interposers for the first instance, and most recently, the elevated fan-out bridge approach.

The holy grail of chiplet architecture is of course 3D stacking. The 3D hybrid bond approach that we have recently introduced with AMD 3D V-Cache™ provides dramatically higher bandwidth density, which has enabled us to connect a 64MB cache chiplet directly on top of the 32MB of existing cache, which required thousands of signals—so the package technology choice is very specific to the architecture. The choice can be visualized in a simplified way. The higher density package technologies are more expensive because they require more precise patterning and many more processing steps; however, with that density comes the benefits of a reduction in interface area, and of course, lower energy for data movement. Chipletizing comes with the overheads including IO area, additional design effort and complexity, additional assembly and testing steps. Getting to the right architecture requires that we must ensure that the value of our newly modular solution with its configuration flexibility and yield has benefits that more than outweigh the costs. Getting this right is a highly multi-disciplinary endeavor, requiring engineers from different domains to rapidly iterate and provide solutions in new ways.

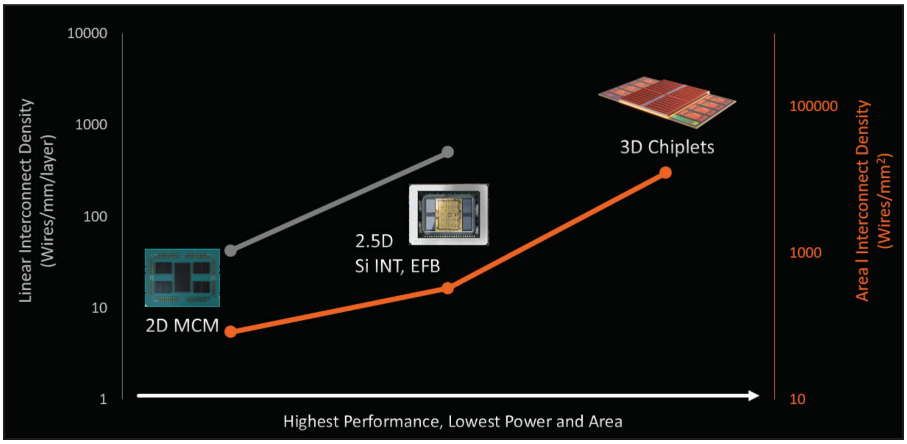


Figure 4: Improving key parameters that drive high-performance computing forward.

AMD elevated fan-out bridge architecture

To illustrate the result of one of these optimization challenges, let us focus on the implementation of a new package architecture called elevated fan-out bridge (**Figure 5**) that we recently announced for the MI200 GPU compute product. As noted earlier, these products require terabytes of memory bandwidth and therefore need denser connections than organic packages provide. One industry approach is shown on the left in **Figure 5** that embeds the silicon bridge die, containing the interconnect wires, into a cavity carved out of the organic package. This has better electrical behavior than legacy 2.5D silicon interposer approaches because it does not require through-silicon vias (TSVs) though does come with challenges associated with the substrate embedding approach.

We decided to develop a cleaner approach that elevates that silicon bridge to live in the shadow of copper pillar bumps. We can thin these silicon bridges down so that there is no significant height impact to the compute die. We now avoid having to carve out a cavity in the substrate and can also lithographically define this module as a unit without dealing with micro bumps on the substrate. Getting better placement accuracy with this method provides an example of the evolution of package technologies and the innovation going on in this space. Chiplet designs can get quite complex with eight high-bandwidth memory (HBM) stacks, two compute die chiplets, and the elevated fan-out bridges (EFBs) to connect them. By choosing a technology that is robust and manufacturable, we have been able to deploy the tens of thousands of these required for the Frontier supercomputer.

AMD 3DVCache™

As computer architects know well, large on-die L3 caches can provide instructions per clock (IPC) uplifts for CPU performance, which is especially important in today's world of ever-increasing appetites for compute and for large data sets. Not surprisingly, as we survey products across the industry over the past decades, there has been a steady increase in on-die cache sizes. So that begs the question, can this trend continue indefinitely? In fact, why is it that the on-die cache integration is starting to slow?

The answers to these questions lie in the barriers to large on-die caches. As noted earlier, Moore's Law slowdown impacts different silicon functions differently. Analog circuits have not scaled much into the advanced nodes, and SRAMs, upon which on-die caches are largely based, are also not scaling as well as logic.

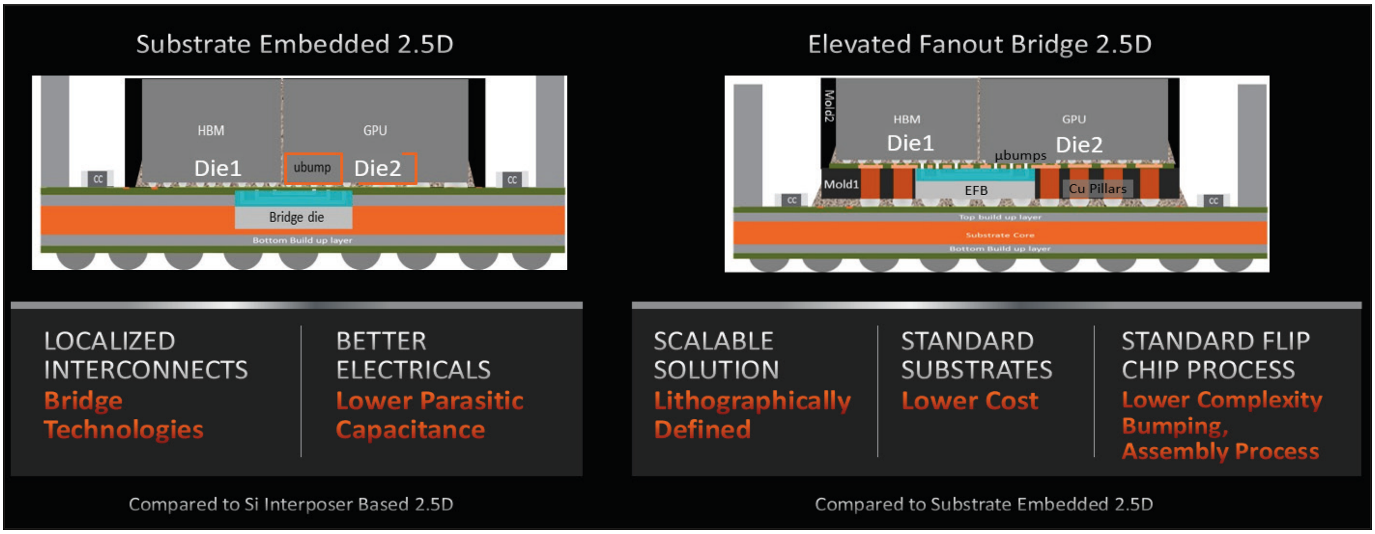
Increasing the on-die cache capacity, which also increases the die size and lowers the yield, is becoming increasingly cost prohibitive and also becomes a challenge for product flexibility. The performance afforded by large caches is important for some markets, though it can be overkill for other market segments to bear the added cost. Finally, larger area also means longer data path distances, which increases cache access latency power and can offset the performance gains.

Up to this point, chiplet integration had mostly meant 2.5D integration. For example, in a hypothetical CPU with a large cache, one can separate part of the cache into a separate die, or chiplet, and place them side by side. The smaller die sizes can improve yield, and therefore the cost, and it provides the flexibility to have the CPU die with a smaller cache as a standalone product to address different markets.

As valuable as these benefits are, extending chiplet integration to 3D can break even more barriers. By placing the dies on top of each other, you can have the added capacity without the added lateral distance, so you can keep the latency low, and the dynamic power low by freeing up valuable space inside the package. You can also fit more cores and more transistors within a given package size. All these incentives led to the creation of the AMD 3D V-Cache™—the industry's first high-performance processor product with 3D integration based on hybrid bond technology.

The AMD V-Cache™ consists of three main components. The first one is the "Zen 3" CPU core complex die (CCD). It is manufactured using TSMC 7nm FinFET technology. Each CCD contains eight cores in a core complex (CCX) and the eight cores share a 32MB L3 cache. It was able to achieve a 19% average IPC uplift over the previous "Zen 2" design, and it has a die size of 81mm^2 (**Figure 6**). What is important to point out here is that the AMD 3D V-Cache™ support, both architecturally and physically, was planned for and integrated into the CCD, from the beginning of "Zen 3" design.

The second component of the AMD 3D V-Cache™ is the extended L3 Die (L3D). Like the CCD, it was also built using TSMC 7nm FinFET technology. It has a die size of 41mm^2, which is roughly half of the CCD die size. The sizing was intentional to allow the L3D to fit over the CCD's L2 and L3 cache area. The relatively low power density of the caches allowed the thermal impact



LOCALIZED INTERCONNECTS
Bridge Technologies

BETTER ELECTRICALS
Lower Parasitic Capacitance

Compared to Si Interposer Based 2.5D

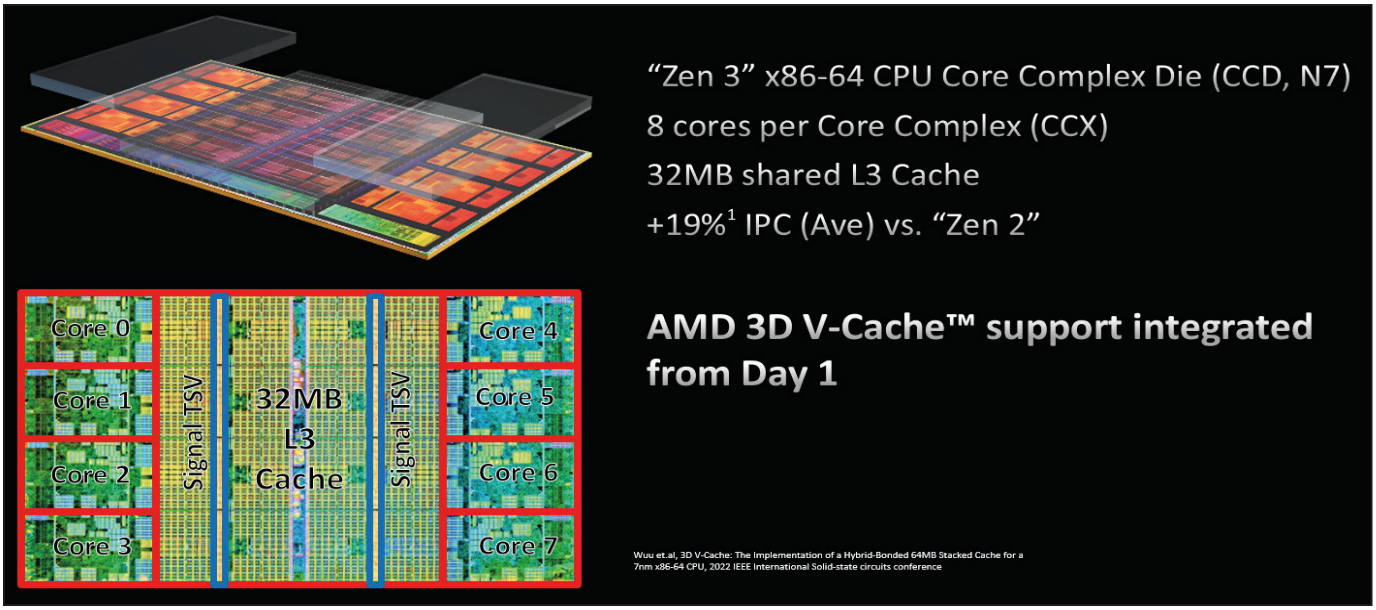
SCALABLE SOLUTION
Lithographically Defined

STANDARD SUBSTRATES
Lower Cost

STANDARD FLIP CHIP PROCESS
Lower Complexity Bumping, Assembly Process

Compared to Substrate Embedded 2.5D

Figure 5: 2.5D "bridge" architecture landscape.



"Zen 3" x86-64 CPU Core Complex Die (CCD, N7)

8 cores per Core Complex (CCX)

32MB shared L3 Cache

+19%[1] IPC (Ave) vs. "Zen 2"

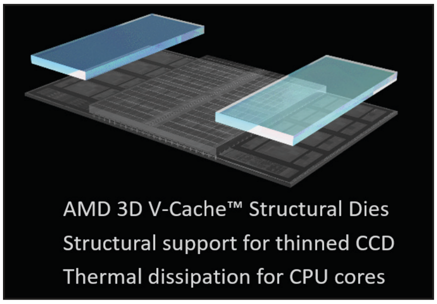
AMD 3D V-Cache™ support integrated from Day 1

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

Figure 6: AMD 3D V-Cache™ components: CCD.

because of overlapping the two dies from becoming a limiter.

The final component of the AMD 3D V-Cache™ is the structural die. Two structural dies, which are dummy silicon dies, are placed over the CCD area not covered by the L3D (Figure 7). The structural dies serve two purposes: 1) as the name implies, they provide structural support for the thinned down CCD die; and 2) because silicon is a good thermal conductor, the structural dies are also used for thermal dissipation from the high-frequency, high-power density CPU cores to the heat sinks.



AMD 3D V-Cache™ Structural Dies

Structural support for thinned CCD

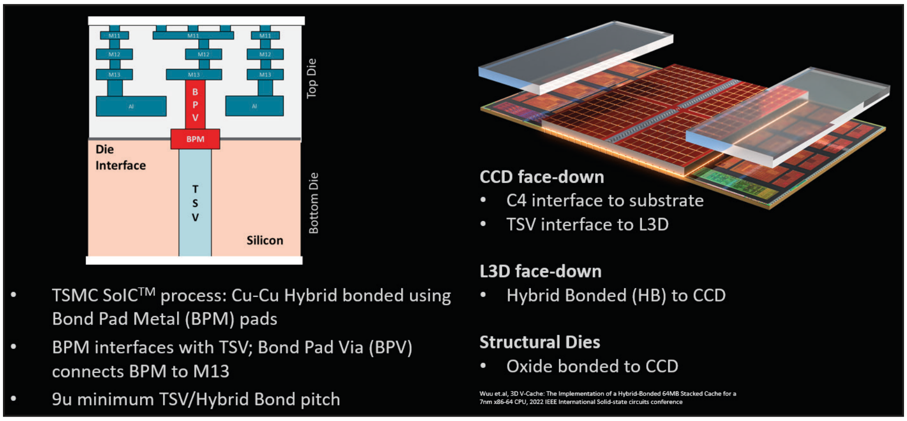
Thermal dissipation for CPU cores

Figure 7: AMD 3D V-Cache™ components: structural die.

A closer look at the AMD 3D V-Cache™ hybrid bond technology is shown in Figure 8. It uses the TSMC-SoIC™ process. The image shows the backside of the face-down bottom die, and the face-down top die hybrid-bonded onto the bottom die. The Cu interface between the dies is called bond pad metal (BPM), which connects to the TSV from the bottom die. On the other side of the BPM is the bond pad via (BPV), which is used to connect the BPM to the Cu metal 13. It is through these TSV, BPM, and BPV structures that power delivery and signals are exchanged between the top and bottom dies. The technology supports a 9µm minimum TSV pitch.

Physically, the CCD is placed face down with C4 interfaces to the substrate. The backside of the CCD is thinned down to reveal the TSVs, which serve as the interconnects to the L3D. The L3D is then also placed face-down and hybrid-bonded to the back of the CCD. Finally, the structural dies are placed on the two sides of the CCD and oxide-bonded to the CCD. Please note, this hybrid bond technology differs from the common 3D approach of connecting the dies through micro-bumps.

Now, we compare the AMD Cu-based 3D architecture versus the current best in class solder-based micro-bump 3D architecture (Figure 9). Solder-based micro-bump technology with tall TSVs is based on traditional solder-based packaging technologies and can scale from 50µm to ~36µm and is acceptable for low-bandwidth applications. AMD 3D chiplet architecture, as shown to scale relative to micro-bump technology, by contrast, uses silicon fab-like manufacturing methods with back-end design rule-based TSVs with Cu-only interconnects without the presence of solder. This is a transformational point in the industry's advanced packaging journey, where interconnect technologies are now being enabled using silicon fab-based techniques to enable extreme bandwidth architectures. As a result of the extreme scaling, we are also able to achieve >3x higher interconnect energy efficiency, >15x higher interconnect density, as well as better signal and power performance compared to micro-bump 3D architectures.



CCD face-down
- C4 interface to substrate
- TSV interface to L3D

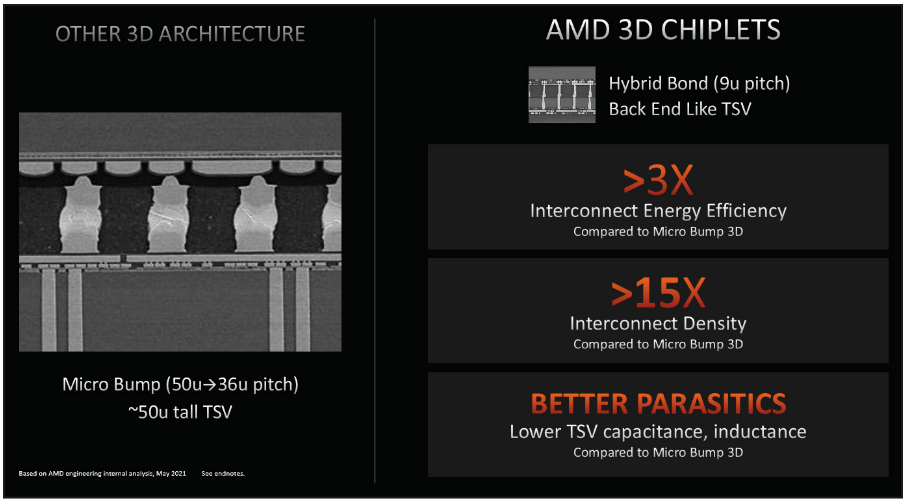
L3D face-down
- Hybrid Bonded (HB) to CCD

Structural Dies
- Oxide bonded to CCD

- TSMC SoIC^{TM} process: Cu-Cu Hybrid bonded using Bond Pad Metal (BPM) pads
- BPM interfaces with TSV; Bond Pad Via (BPV) connects BPM to M13
- 9u minimum TSV/Hybrid Bond pitch

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

Figure 8: 3D V-Cache™: bringing it together.



OTHER 3D ARCHITECTURE

Micro Bump (50u→36u pitch)
~50u tall TSV

Based on AMD engineering internal analysis, May 2021   See endnotes.

AMD 3D CHIPLETS

Hybrid Bond (9u pitch)
Back End Like TSV

>3X
Interconnect Energy Efficiency
Compared to Micro Bump 3D

>15X
Interconnect Density
Compared to Micro Bump 3D

BETTER PARASITICS
Lower TSV capacitance, inductance
Compared to Micro Bump 3D

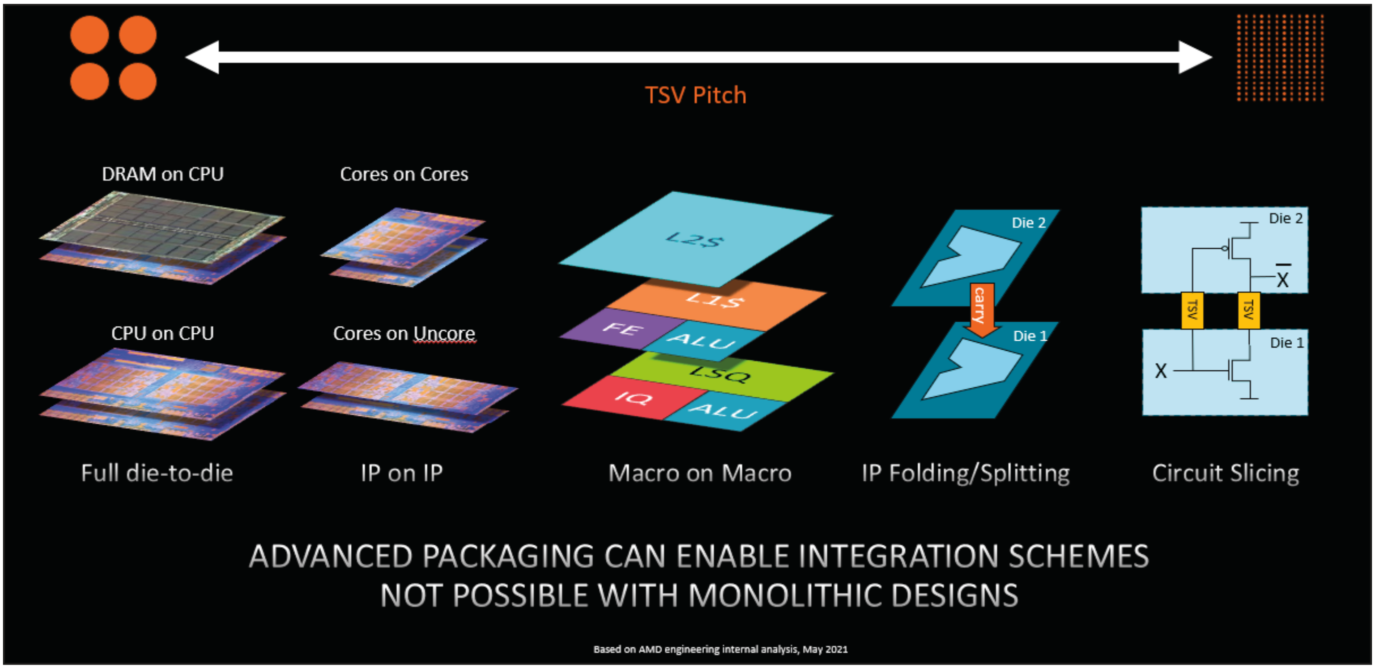
Figure 9: AMD hybrid-bonded 3D chiplet architecture comparison to solder-based 3D architectures.

Regarding "Zen 3" cache hierarchy, each core has a 32KB I-cache and a 32KB D-cache, along with a private 512KB L2 cache. There are eight cores per CCD, and all eight cores share a 32MB L3 cache. The L3 cache is 16-way set associative, with a 32B/cycle interface to each core. DECTED ECC, which can correct double bit errors and detect triple bit errors, is included for enhanced data reliability. When the L3D is bonded on top of the CCD, it expands the 32MB shared L3 cache to 96MB. The 96MB cache continues to be shared between the eight cores, and it continues to be 16-way set associative. It also maintains the L3's 32B/cycle interface to each core, which provides more than 2TB of total L3 bandwidth per second. Despite tripling the L3 size, AMD 3D V-Cache™ only adds four cycles of additional latency, which can only be achieved through 3D stacking.

Power delivery was a key architecture focus when we architected AMD V-Cache™. The CCD has three primary power supplies (**Figure 11**) – there is RVDD in orange, which is the raw, ungated supply upon which the L3 cache logic runs. Then there is VDD, which each core regulates independently from RVDD. Finally, there is VDDM, which is the supply for the L2 and L3 SRAM bit cells. Of course, there is also VSS, which is shown in grey in the diagram (**Figure 11**). When the L3D is stacked onto the CCD, both RVDD and VDDM are delivered to the L3D through power TSVs. To better convey the power delivery RDL, the construction in **Figure 11** is flipped upside down with the top L3D die on the bottom. RVDD supplies the logic portion of the L3D die, while VDDM powers the SRAM bit cells. The power TSVs are primarily placed in the channels between the SRAM macros in the CCD.

The SRAM arrays on the L3D die consist of 512 128KB data macros, and 1088 6KB tag and the (LRU) macros located near the signal TSV columns. It is a dual-rail design using VDDM for the SRAM bitcells and RVDD for the peripheral circuits. As added power can negatively impact performance in a power constrained environment, the L3D arrays are optimized not only for high density, but for low power as well. To that end, the SRAM arrays on the L3D uses extensive power reduction features.

3D interface signals are extremely simple flop-to-flop signals that can be enabled only with the use of a hybrid-bonded architecture with its low parasitics. On the transmission side, the signal after leaving the flop is buffered and sent through the TSV to the other die. On the receiving side, the signal first goes through a minimal electrostatic discharge (ESD) circuit to protect against ESD events that can occur during the 3D assembly process. The signal then goes through an isolation circuit, which properly isolates the interface signal that would be floating when the other die is not attached.



ADVANCED PACKAGING CAN ENABLE INTEGRATION SCHEMES
NOT POSSIBLE WITH MONOLITHIC DESIGNS

Based on AMD engineering internal analysis, May 2021

Figure 10: Vision for enabling new architectures with future 3D stacking innovations.

Finally, the incoming signal is captured by a flop. What is interesting here is the simplicity and the compactness of the fully-digital IO circuitry, which contributes to the power efficiency and low latency of this hybrid-bonded 3D interface. So how does this translate to performance? In a desktop gaming system, AMD 3D V-Cache™ delivered on average 15% faster gaming performance when compared with its non stacked Ryzen™ counterpart. This 15% is truly a generational leap in performance, which in the past has been enabled only by silicon node transitions.

Milan-X server implementation of the AMD V-Cache™ architecture enables three times the L3 cache compared to standard Milan processors. This additional L3 cache relieves memory bandwidth pressure and reduces latency –that in turn dramatically speeds up application performance.

3D stacking: future and challenges

3D cache stacking over CPU cores is just the beginning of the 3D journey (**Figure 10**).

The future of 3D stacking is a function of TSV pitch and can spawn many architectural innovations including IP-on-IP stacking, macro-on-macro stacking, IP folding/splitting, as well as circuit-level slicing 3D stacking technology progression. These innovations, along with other advanced packaging techniques, will enable beyond-Moore's-Law scaling this decade and enable complex heterogeneous integration schemes not possible even with monolithic designs.

There are multiple challenges to enable 3D chiplet architectures. All these chiplets need to be tested thoroughly before assembly or we throw away the entire expensive module. Stacking encounters challenges with higher power densities. This predicament comes along at the same time as Moore's Law is doing less and less for power. Managing and mitigating thermal issues is going to be an interesting and exciting area for innovation, along with power delivery solutions and high current densities across multiple dice means a 3D power grid, among other things. All our tricks of integrated regulators and power gating will need to be deployed to support the power demands of all the layers in the design. Silicon and package are merging with this architecture. Enabling the right design tools that can seamlessly move from system to package to C4 to 3D interface, to truly deliver the best-in-class DTCO, is critical.

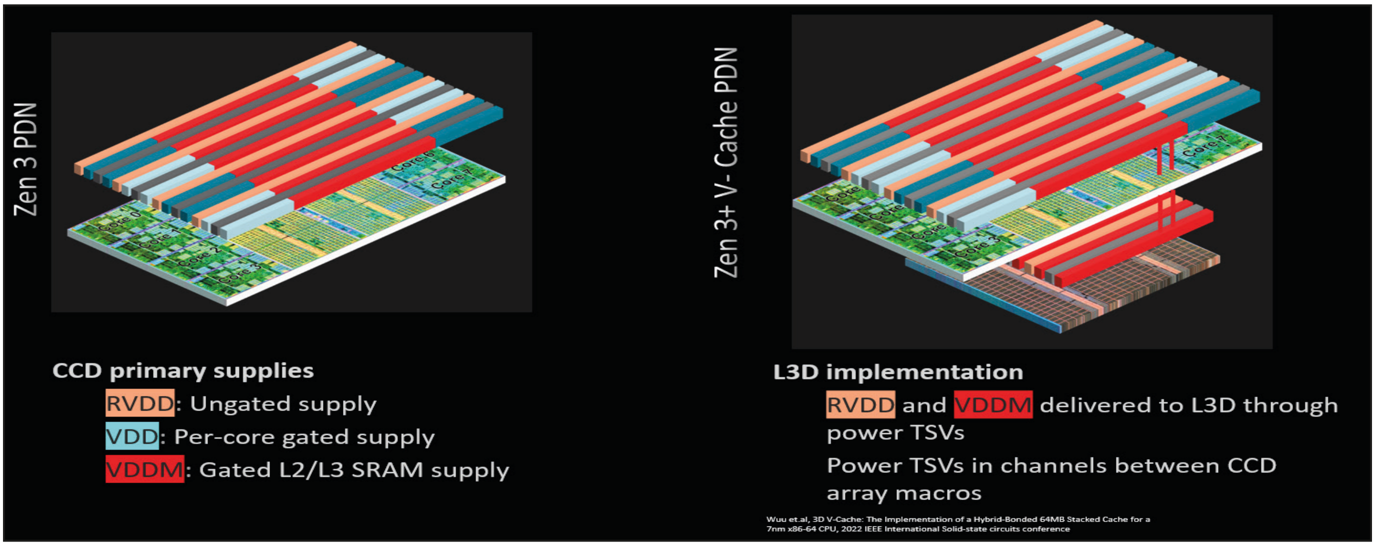
Finally, as mentioned at the outset, performance is delivered at the system level, and these heterogeneous modular SoCs will need to be connected with the right software to deliver system-level performance where an increasing amount of differentiation can be delivered.

Summary

We are truly at a new era of computing. Design and innovation must take a step up. The new paradigms will combine traditional CPU compute engines heterogeneously with accelerators, using continuously evolving and improving package technology to enable levels of integration that today are at the board level. In the future, they will be at the integrated modular silicon level. System architectures, previously only in massive supercomputers, are now coming to the masses. It will be an incredibly exciting next era of computing innovation driven by advanced packaging and I look forward to the opportunities ahead!

Endnotes

1. AMD 3D Chiplet Technology -Competition 3D architecture picture from SystemPlus. Intel Core i5-L16G7: the first utilization of Intel's Foveros Technology with Package-on-Package configuration in a consumer product.. https://www.systemplus.fr/reverse-costing-reports/intel-foveros-3d-packaging-technology/
2. MLNX-001R: EDA RTL Simulation comparison based on AMD internal testing completed on 9/20/2021 measuring the average time to complete a test case simulation. Comparing: 1x 16C 3rd Gen EPYC CPU with AMD 3D V-Cache Technology versus 1x 16C AMD EPYC™ 73F3 on the same AMD "Daytona" reference platform. Results may vary based on factors including silicon version, hardware and software configuration and driver versions.
3. MLNX-021R: AMD internal testing as of 09/27/2021 on 2x 64C 3rd Gen EPYC with AMD 3D V-Cache (Milan-X) compared to 2x 64C AMD 3rd Gen EPYC 7763 CPUs using cumulative average of each of the following benchmark's maximum test result score: ANSYS® Fluent® 2021.1, ANSYS® CFX® 2021.R2, and Altair Radioss 2021. Results may vary.
4. MLN-075A: Altair™ Radioss™ comparison based on AMD internal testing as of 09/27/2021 measuring the time to run the neon, t10m, and venbatt test case simulations using a server with 2x AMD EPYC 75F3



CCD primary supplies

RVDD: Ungated supply

VDD: Per-core gated supply

VDDM: Gated L2/L3 SRAM supply

L3D implementation

RVDD and VDDM delivered to L3D through power TSVs

Power TSVs in channels between CCD array macros

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

Figure 11: Schematic of 3D AMD V-Cache™ power delivery.

versus 2x Intel Xeon Platinum 8362. Neon crash impact is the max result test case. Results may vary.

5. MLN-080B: ANSYS® CFX® 2021.1 comparison based on AMD internal testing as of 09/27/2021 measuring the average time to run the Release 14.0 test case simulations (converted to jobs/day - higher is better) using a server with 2x AMD EPYC 75F3 utilizing 1TB (16x 64 GB DDR4-3200) versus 2x Intel Xeon Platinum 8380 utilizing 1TB (16x 64 GB DDR4-3200). Results may vary.
6. MLN-130A: ANSYS® Mechanical® 2021 R2 comparison based on AMD internal testing as of 09/27/2021 measuring the average of all Release 2019 R2 test case simulations using a server with 2x AMD EPYC 75F3 versus 2x Intel Xeon Platinum 8380. Steady state thermal analysis of a power supply module 5.3M (cg1) is max result. Results may vary.
7. MI200-01 - World's fastest data center GPU is the AMD Instinct™ MI250X. Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X (128GB HBM2e OAM module) accelerator at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), 383.0 TFLOPS peak theoretical half precision (FP16), and 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16) floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), 46.1 TFLOPS peak theoretical single precision matrix (FP32), 23.1 TFLOPS peak theoretical single precision (FP32), 184.6 TFLOPS peak theoretical half precision (FP16) floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 312 TFLOPS peak half precision (FP16 Tensor Flow), 39 TFLOPS peak Bfloat 16 (BF16), 312 TFLOPS peak Bfloat16 format precision (BF16 Tensor Flow), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1.
8. MI200-02 - Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X accelerator (128GB HBM2e OAM module) at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak double precision matrix (FP64 Matrix) theoretical, floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator resulted in 19.5 TFLOPS peak double precision (FP64 Tensor Core) theoretical, floating-point performance. Results found at: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1.
9. MI200-07 - Calculations conducted by AMD Performance Labs as of Sep 21, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e) OAM accelerators designed with AMD CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 4,096 bits times 2 die and memory data rate is 3.20 Gbps for total memory bandwidth of 3.2768 TB/s ((3.20 Gbps*(4,096 bits*2))/8). The highest published results on the NVidia Ampere A100 (80GB) SXM GPU accelerator resulted in 2.039 TB/s GPU memory bandwidth performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf
10. MI200-15A - Testing Conducted by AMD performance lab as of 10/7/2021, on a single socket Optimized AMD EPYC™ CPU server, with 4x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology, using LAMMPS ReaxFF/C, patch_2Jul2021 plus AMD optimizations to LAMMPS and Kokkos that are not yet available upstream resulted in a median score of 4x MI250X = 19,482,180.48 ATOM-Time Steps/s Vs. Dual AMD EPYC 7742@2.25GHz CPUs with 4x NVIDIA A100 SXM 80GB (400W) using LAMMPS classical molecular dynamics package ReaxFF/C, patch_10Feb2021 resulted in a published score of 8,850,000 (8.85E+06) ATOM-Time Steps/s. https://developer.nvidia.com/hpc-application-performance 19,482,180.48/8,850,000=2.20x (220%) the/1.2x (120%) faster. Container details found at: https://ngc.nvidia.com/catalog/containers/hpc:lammps Information on LAMMPS: https://www.lammps.org/index.html Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

Biography

Raja Swaminathan is a Senior Fellow & Advanced Packaging Leader at AMD, Austin, Texas, USA. Prior to AMD, he was at Apple, architecting and developing the packaging technologies for the M1x series of processors and Principal Engineer, Silicon Package Architecture at Intel. He holds 35 patents on semiconductor packaging technologies. He received his BS in metallurgy from Indian Institute of Technology, Madras, India, and a PhD in Materials Science from Carnegie Mellon U. Email: raja.swaminathan@amd.com