A Research Retrospective on the AMD Exascale Computing Journey

Gabriel H. Loh, Michael J. Schulte, Mike Ignatowski, Vignesh Adhinarayanan
Shaizeen Aga, Derrick Aguren, Varun Agrawal, Ashwin M. Aji
Johnathan Alsop, Paul T. Bauman, Bradford M. Beckmann, Majed Valad Beigi
Sergey Blagodurov, Travis Boraten, Michael Boyer, William C. Brantley
Noel Chalmers, Shaoming Chen, Kevin Cheng, Michael L. Chu
David Cownie, Nicholas Curtis, Joris Del Pino, Nam Duong
Alexandru Duţu, Yasuko Eckert, Christopher Erb, Chip Freitag
Joseph L. Greathouse, Sudhanva Gurumurthi, Anthony Gutierrez, Khaled Hamidouche
Sachin Hossamani, Wei Huang, Mahzabeen Islam, Nuwan Jayasena
John Kalamatianos, Onur Kayiran, Jagadish Kotra, Alan Lee
Daniel Lowell, Niti Madan, Abhinandan Majumdar, Nicholas Malaya
Srilatha Manne, Susumu Mashimo, Damon McDougall, Elliot Mednick
Michael Mishkin, Mark Nutter, Indrani Paul, Matthew Poremba
Brandon Potter, Kishore Punniyamurthy, Sooraj Puthoor, Steven E. Raasch
Karthik Rao, Gregory Rodgers, Marko Scrbak, Mohammad Seyedzadeh
John Slice, Vilas Sridharan, Rene van Oostrum, Eric van Tassell
Abhinav Vishnu, Samuel Wasmundt, Mark Wilkening, Noah Wolfe
Mark Wyse, Adithya Yalavarti, Dmitri Yudanov

Advanced Micro Devices, Inc.

Abstract

The pace of advancement of the top-end supercomputers historically followed an exponential curve similar to (and driven in part by) Moore's Law. Shortly after hitting the petaflop mark, the community started looking ahead to the next milestone: Exascale. However, many obstacles were already looming on the horizon, such as the slowing of Moore's Law, and others like the end of Dennard Scaling had already arrived. Anticipating significant challenges for the overall high-performance computing (HPC) community to achieve the next 1000x improvement, the U.S. Department of Energy (DOE) launched the Exascale Computing Program to enable and accelerate fundamental research across the many technologies needed to achieve exascale computing.

AMD had the opportunity to contribute to the so-called "*Forward" programs from the DOE, which were a series of public-private partnerships focused on research and co-design activities covering compute architectures, interconnects, memory systems, chiplets and packaging, software stacks, applications, and more. Some of the research from these programs can now be found in the world's first exascale supercomputer, some were a little ahead of their time and may have an impact in the coming years, and others simply did not pan out. In this paper, we provide a retrospective of AMD's nearly decade-long research journey covering how we tried to predict the architecture of a supercomputer a decade into the future, what we got right, what we got wrong, and some of the insights and learnings that we discovered along the way.

CCS Concepts

• Hardware~Very large scale integration design • Computer systems organization~Architectures~Parallel architectures • Networks~Network components~Logical nodes

Keywords

Exascale, HPC, high-performance computing, supercomputing, Frontier, memory, chiplets, heterogeneous compute, accelerated processing unit, research.

1 Introduction

Supercomputing, or high-performance computing (HPC), has been at the heart of many of the world's scientific discoveries for many decades. These machines perform massive computations in support of scientific experiments that are enabling the broader research community to make new discoveries spanning many disciplines. These discoveries lead to applications that broadly impact many facets of our lives, such as designing more efficient ways to produce power; engineering safer buildings, bridges, and other infrastructure; understanding biological and chemical processes to discover new medical treatments; creating more accurate predictive weather and climate models; and probing the origins of the universe.

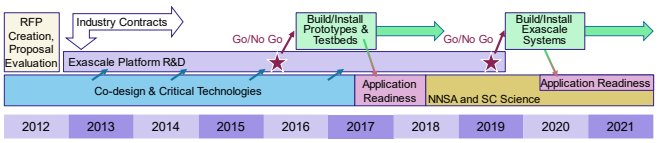
Fueled by Moore's Law and other technological innovations over the years, the growth in the computational capabilities of the world's top supercomputers historically followed an exponential trend, for example, as measured by TOP500 LINPACK performance [96]. This has taken the HPC community through multiple generations of supercomputers with key milestones demarcated by crossing the teraflop (TF, 10^{12} double-precision (DP) floating point operations per second (FLOPS)) threshold and then petaflop (PF, 10^{15} DP FLOPS) levels of performance [19]. The next target was to break the exaflop (EF, 10^{18} DP FLOPS) barrier, but by the early 2010s there were already multiple warning signs of the slowing down of technology improvements such as the end of Dennard scaling [26], challenges with maintaining Moore's Law [72], the Memory Wall [102], increasing power consumption, and more.

1.1 The View from 2012

Anticipating these growing concerns, the United States Department of Energy (DOE) launched an aggressive program to identify potential technologies that would be needed to enable the continued progress of U.S. supercomputing capabilities toward the exaflop level and beyond. Of additional concern was not just whether the necessary technologies could be developed, but also whether broader commercial priorities (i.e., outside of the HPC space) would deliver such technologies to the market on a timeline suitable for the DOE's supercomputing needs.

In 2011, the DOE issued a Request for Information (RFI) soliciting input on how a wide range of requirements for exascale technologies might possibly be met and what kinds of advanced research efforts would be required to enable it all [17]. Figure 1(a) shows the originally envisioned timeline from the 2011 RFI looking forward to enabling exascale computing in the 2019-2020 timeframe. Figure 1(b) reproduces some of the 2011 RFI's system objectives for exascale systems in the target timeframe. Not surprising is the performance target of 1000PF (an exaflop), but it is worth noting that the DOE also calls out an objective of 300PF on (to be determined) other workloads that may not be as regular or computationally intense as LINPACK [35]. This was early messaging from the DOE that a supercomputer designed to be useful to a wide and diverse set of scientific users cannot only deliver raw FLOPS, but the machine must be balanced and continue to deliver high levels of performance across a range of different workload types (e.g., memory intensive, communication heavy, irregularly structured). A machine designed only to win on a single benchmark would not be acceptable as it would have limited broader scientific utility.

Another notable constraint is the 20 megawatt (MW) power limit. The DOE recognized early on that power-performance efficiency needed to be a top priority in any of these future systems. There are multiple practical reasons for this constraint beyond environmental and sustainability concerns. Few facilities have the power infrastructure to deliver tens of megawatts of power (and note that the 20MW limit is only for the computational components and is not inclusive of storage systems or facility infrastructure such as power delivery and cooling), and the costs of upgrading facilities (if even possible, given the local capabilities of the electricity provider) for additional tens of megawatts of capacity would be substantial. In addition to the upfront facility infrastructure impact, a power consumption rate in the megawatts



(a)

Exascale System	Goal
Delivery Date	2019-2020
Performance	1000 PF LINPACK 300 PF on to-be-specified applications
Power Consumption	20 MW
Memory Capacity (incl. NVRAM)	128 PB
Node Memory Bandwidth	4 TB/s
Node Interconnect Bandwidth	400 GB/s

(b)

Figure 1. (a) Exascale timeline and (b) system objectives from the 2011 U.S. DOE exascale research and development Request for Information.

can cost the facility many millions of U.S. dollars per year for the power bill alone.

An interesting attribute of the requirements is in the system memory capacity targets. While 128PB (for the entire system) is a substantial amount of memory, it is notable that non-volatile memory (NVRAM) is explicitly called out as a possibility. Looking back at the computer architecture research environment of that era, this RFI came out when the research community was highly concerned with the (feared to be) imminent end of DRAM scaling. A significant amount of research effort was made by the community to devise ways to leverage emerging NVRAM technologies as an augmentation to, or an outright replacement of, the conventional DRAM-only memory system [62][86][103].

These RFI exascale requirements (e.g., performance, power, capacity) would likely not be achieved by hardware architects devising solutions on their own. So while not visually captured in Figure 1, the DOE was very prescient in placing a huge emphasis on the idea of "co-design" wherein technology providers needed to collaborate closely with scientists and technologists from the DOE to research and develop the new technologies to support exascale computing. The workloads and usage models of the DOE scientists are not the same as broader commercial applications, and it was a priority for the DOE to avoid funding the development of solutions that provided only marginal benefit to its HPC user base while delivering disproportionate benefits to other markets that were not directly funding or supporting the exascale mission.

1.2 The View from 2023

The 2011 RFI timeline provided a speculative, forward-looking, and perhaps aspirational roadmap for getting to exascale capabilities. Figure 2 shows a retrospective view of how events actually played out. To support the required research and development efforts to innovate and accelerate the necessary exascale technologies, the DOE funded a series of programs colloquially referred to as the "*Forward" (pronounced "star forward") programs. These programs were public-private partnerships between the DOE and various technology companies covering processors, memory, storage, networking, and software. The FastForward, FastForward 2, and PathForward programs had

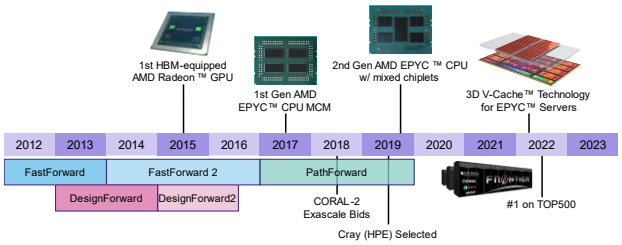


Figure 2. Timeline illustrating U.S. DOE exascale R&D programs and milestones (bottom) and key AMD technology introductions (top).

more focus on the individual components (e.g., processors, memory) [1][7][8], whereas the DesignForward and DesignForward 2 programs emphasized system-level concerns [13]. AMD was selected to participate in all five programs.

The *Forward programs also served as an effective collaboration vehicle for AMD to work more closely with system integrators. AMD designs and provides the component-level technology (CPUs, accelerators), but the system integrators are the ones responsible for building and delivering the overall completed supercomputing system. The DesignForward programs enabled AMD to gain critical insights into how our components could potentially be assembled into a larger system, and the program facilitated key interactions and collaborations with the system integrators that provided feedback for refining and changing our component architectures.

The timeline in Figure 2 shows that the DOE's original target of standing up an exascale system in the 2019-2020 timeframe was not achieved. The first official exascale score on the TOP500 came when Oak Ridge National Laboratory's (ORNL) Frontier supercomputer broke the exaflop threshold in June 2022 [75]. In the coming sections of this paper, we will share some details about technology trends and changes that affected the exascale roadmap, as well as how the continually evolving technology landscape caused AMD to repeatedly revise and refine its vision for how to enable exascale computing. Figure 2 also includes key technology introductions that we will revisit later in this paper.

1.3 Objectives of this Paper

The purpose of this paper is to provide the broader research community with an inside view into how an industrial research team defines and refines a research agenda for an aggressive long-term objective such as achieving exascale computing over a decade-long time horizon. We describe the thinking and rationale behind our initial vision for exascale computing (Section 2). We then take the reader forward in time as the broader technology landscape as well as AMD's roadmap evolved, and how those factors repeatedly impacted our approach to exascale computing (Section 3). The paper also aims to highlight specific research topics that were pursued in the course of this exascale endeavour, some of which had significant impact on AMD's exascale program and even AMD's broader product roadmap and some that are still awaiting the right technology and market conditions to arise (Section 4). While we provide a summary and overview of the final design of AMD's processors deployed in the Frontier supercomputer (Section 5), this paper is *not* a technical deep-dive or product disclosure for any of those specific components (references are provided later in the paper for those wanting more details). We hope that this paper provides the reader with meaningful, useful, and interesting insights into our research processes and thinking, highlights areas needing continued research efforts by the broader computer architecture community, and excites researchers (especially those just starting out) to continue pursuing innovative work in compute architecture. Exascale computing is a great achievement, but it is merely a milestone and not a destination. The technical challenges going forward in what is now the post-Exascale era are only getting more difficult, and we will collectively need to harness the entire community's creativity and innovation!

2 Initial Strategy for a Decade-long Program

AMD began its exascale collaboration with the DOE through the FastForward program [1][61] in 2012. The DOE had shared its system targets (Figure 1(b)), and now it was up to us to define our technical vision for a solution. How does one define research for a target that is nearly a decade away?

2.1 Known Technology

We started with what we already knew (although not necessarily public information at the time). The trend toward general-purpose GPU computing was already underway in the industry, and so the inclusion of accelerated computing as a performance engine was a natural component to consider. In 2011 in the client/consumer space, AMD had already launched its first accelerated processing unit (APU) that architecturally and physically combined the CPU and GPU into a single unified entity [1]. The APU approach provides a unified memory space for both the CPU and GPU and had the potential to greatly reduce data movement that would otherwise be incurred from copying data back and forth between the CPU host memory and the GPU video memory.

While it would not be until 2015 when AMD launched the AMD Radeon™ 300 series GPUs featuring silicon interposer technology and 3D high-bandwidth memory (HBM) [9], AMD was already well along the path of bringing these products to the market [25]. Given the node memory bandwidth target of 4TB/s, an HBM or HBM-like solution utilizing 2.5D integration was a natural starting point (2.5D refers to using 3D stacking of die on a passive silicon interposer to combine multiple chips side-by-side in a 2D arrangement) [34].

2.2 Speculating on Future Technologies

The end of Moore's Law has been forecasted many, many times over the years, and in 2012 there existed similar concerns about how silicon technology scaling would slow down by the target exascale timeframe. At the time, AMD was shipping products in a 32nm process technology [9][31]. Projecting forward, we predicted that in 2019-2020 we would be in a 10nm technology.

The physical construction of HBM uses 3D stacking of multiple DRAM dies on top of each other. While at the time there were no product plans for stacking logic or processing dies on top of each other, we speculated that by the target exascale timeframe such capabilities would be an option because the fundamental 3D stacking technology is effectively agnostic to whether the individual silicon dies are used for logic, memory, or some other purpose.

With the assumption that 3D stacking technology could be a widely available technology during the target timeframe, we also

predicted that the decades-old processing-in-memory (PIM) idea [38][43][49][57][80][93], where multiple memory dies are integrated with compute, could be considered. Minimizing data movement was seen as an important aspect of improving bandwidth while simultaneously reducing power consumption, and we believed that PIM could add significant value in this regard.

As mentioned earlier, in the 2012 timeframe, the broader computer architecture community had concerns about the industry's ability to continue technology scaling, especially for DRAM [62][86][103]. NVRAM could potentially provide a path to achieving the DOE's system-level memory capacity targets in a more scalable and/or cost-effective manner. The non-volatility of NVRAM also presented opportunities to improve the performance of checkpointing mechanisms often used in very large, long-running scientific simulations [71].

Similar to the topic of NVRAM, at the time the computer architecture community was also very active in researching technologies and applications of different types of silicon photonics [23][58][66][98]. Photonic interconnects promise to significantly increase the amount of bandwidth, at a lower effective energy-per-bit cost, than one can pull out of or feed into a single package without running against the pin limitations of a conventional electrically-connected socket.

2.3 Initial Exascale Heterogeneous Processor

Combining what we already knew about AMD's capabilities and plans together with our projections for possible technologies in the exascale timeframe, we created the initial concept for an "Exascale Heterogeneous Processor" (EHP) that would serve as the foundational computational component for our vision of a future exascale supercomputer [87].

At the heart of the EHP is a high-performance APU coupled with a 3D DRAM (e.g., HBM) memory system. Figure 3(a) shows a block diagram of the original EHP. Multiple GPU-based accelerators (labeled Vector Units) provide up to 12 DP TF of compute along with eight x86 CPU cores that execute the serial portions of applications as well as to run the operating system and other software. The CPU and GPU have their own respective L3 caches, but they are kept coherent across the system on chip (SoC) interconnect. The APU processing units share a single unified memory system consisting of 128GB of in-package DRAM delivering data at a rate of up to 4 TB/s. At the bottom left of the figure, the APU also has connections to multiple NVRAM modules outside of the package. Due to the anticipated slower access latencies for NVRAM, the APU provisions a memory-side cache as a data staging/prefetch area for each module potentially combined with other accelerators (e.g., data compression engines). Each individual NVRAM module consists of a 3D stack of NVRAM dies above a logic die that provides PIM capabilities. Finally, an integrated network interface card (NIC) with photonic interconnects provides high-speed links to the other nodes in the system.

The EHP utilizes a combination of 2.5D silicon interposer and vertical 3D stacking to assemble all of its components, as shown in Figure 3(b). The figure illustrates a CPU die stacked on top of three layers of GPU/vector units. The thinking behind this organization was that 3D stacking would be needed to maximize the compute that could fit within the limited real estate of the package, the vertical organization would help minimize the cost of data movement among the compute components, and separating the CPU and GPU

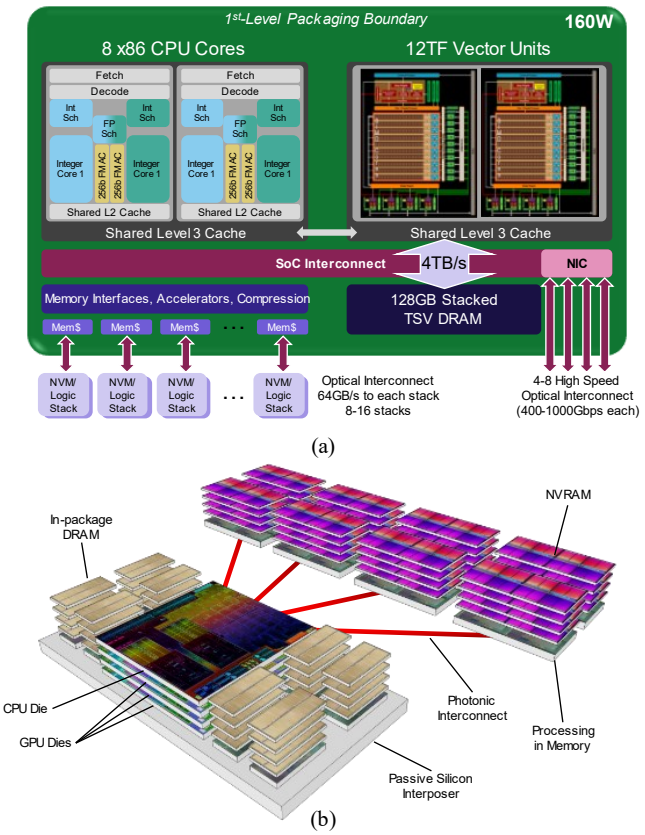


Figure 3. (a) Block diagram of the Exascale Heterogeneous Processor (EHP) concept from the original FastForward program circa 2012, (b) illustrative packaging view of the EHP.

components on to separate layers could provide opportunities to fine-tune the process technology for each layer (e.g., CPU transistors and metal layers optimized for high frequency and reduced latency). Note that in 2012, 3D meant using microbump stacking technology [30] as modern hybrid bonding techniques [95] were not yet being commercially considered. Such a 3D organization would imply significant thermal challenges, likely necessitating very aggressive liquid cooling or similar thermal solutions. The 3D-stacked APU would then be further 2.5D mounted on a passive silicon interposer along with eight stacks of DRAM with 512GB/s of bandwidth per stack. This vintage figure illustrates the DRAM stacks with four layers, but we later adjusted the concept to eight-high DRAM stacks consistent with HBM. The figure also illustrates eight external stacks of photonically-connected NVRAM.

While AMD is not a system integrator, we worked with multiple system integrator partners to explore how the EHP could be assembled into an exascale machine. At 12TF of peak compute per EHP, we would need to aggregate a minimum of 83,334 EHP nodes to reach the 1.0 exaflop target. As it would be unlikely for every node to operate at 100% efficiency, such a machine would probably have over 100,000 nodes to actually sustain an exaflop of compute (which also implies an upper bound of 200W per node given a 20MW system-level target). Providing a high-performance, scalable interconnect for 100,000 nodes would be a significant challenge,

but as that responsibility would fall to the system integrators, we do not focus on that (very interesting) problem in this paper.

3 Refining Our Exascale Node Architecture

Over the following years, we continued to conduct research into many technologies that would be needed to support our exascale node architecture (Section 4). As the various research efforts progressed, combined with updates to AMD's commercial roadmap, we revisited the EHP concept multiple times over the years to improve the design and align it with updated technology assumptions, foundry capabilities, packaging advancements, etc.

3.1 EHP Concept v2, Circa 2014

By the time we started our efforts in the FastForward 2 research program, AMD's roadmap had internally made some significant changes. As detailed in prior work [73], the projected silicon cost trend for leading-edge technology nodes was increasing at an accelerated rate, as shown in Figure 4 (a). This, among other reasons, led AMD to pioneer its chiplet technology for building our processors, with an example AMD EPYC™ server processor shown in Figure 4(b). One of the benefits of AMD's chiplet approach is the ability to reuse silicon components in multiple product configurations, and so our next iteration of the EHP was modified to utilize both the CPU chiplets (CCD) and the IO die (IOD) from our mainstream CPU server products (note that there was no concept of an IOD in the original EHP from Figure 3). Figure 5 illustrates this v2 EHP where on the left side of the package one can find the IOD with four CCDs. This decision to try to leverage the server CCD and IOD resulted in backing away from the aggressive 3D-stacked organization of the first EHP, with the consequence that the 2D-organized layout of the CCDs and IOD needed to consume a significant portion of the package real estate previously occupied by the 3D DRAM.

To offset the loss of packaging area for the 3D DRAM, we modified the GPU portion of the EHP so that the memory was directly stacked on top of the GPU. Given the trend toward breaking up the CPU into chiplets, we also posited that the GPU resources would be similarly partitioned in the future, and so the figure illustrates one DRAM stack on top of each GPU chiplet. This version of the EHP only has enough room for four DRAM stacks, and so we also doubled the height of each stack to maintain the same amount of total capacity. Normally, reducing the number of stacks would also reduce the total bandwidth, but with 3D stacking one could double the bandwidth to compensate (e.g., by doubling the number of vertical data connections). The GPU chiplets are still stacked on top of a silicon interposer to provide high bandwidth between them, but the connection to the CPU side is over the organic substrate to maintain compatibility with the electrical interface of the IOD.

In this v2 EHP, some features from the original EHP are conspicuously absent. In particular, the NVRAM was removed from the concept for a couple of reasons. Co-design activities with the DOE-application experts suggested by restructuring algorithms and codes to make use of a multi-tiered memory organization would be challenging. For example, in one study we asked DOE programmers to identify the hottest data structures in their code, which we could then pin to the in-package DRAM while other "colder" data could be left in the NVRAM. However, we discovered that in many cases this was sub-optimal because the data structures that a

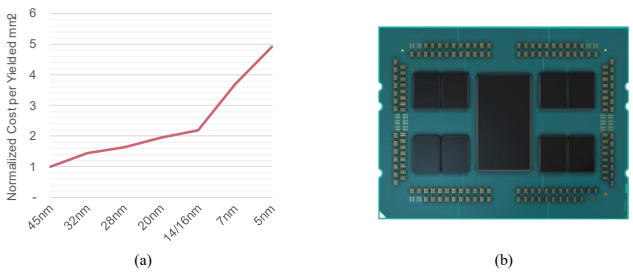


Figure 4. Silicon cost trends over time and (b) an AMD EPYC™ processor utilizing chiplets.

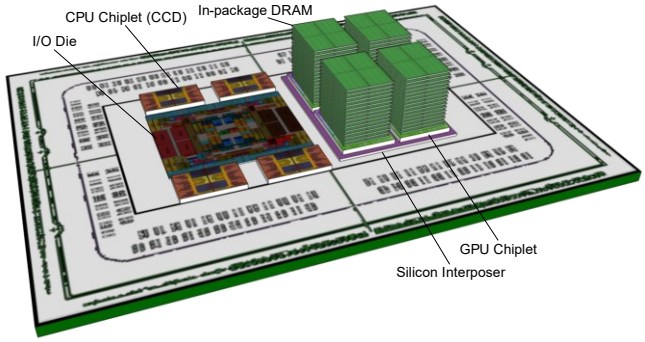


Figure 5. Refinement of the EHP (v2), circa 2014.

programmer views as most frequently accessed could, for example, have a high on-chip cache hit rate and be largely insensitive to whether the backing copy is allocated to in-package DRAM or off-package NVRAM. Around this time, we were also starting to question whether NVRAM as a DRAM replacement or augmentation would be a commercially viable option in the target timeframe. Backing off from NVRAM simplified the design by also omitting the photonic links to the NVRAM and the memory-side caches while allowing us to leverage the mainstream IOD. The combination of 3D-stacked DRAM and no NVRAM also put into question how much additional benefit PIM might provide for this specific EHP design. Note that while these features were removed from EHP v2, we continued significant research efforts on these topics.

Issues with EHP v2: This version of the EHP ended up being overly optimistic about how high the DRAM stacks would grow by the target timeframe. At the time of this writing, commercially-available HBM only implemented eight-high stacks. The GPU side is also thermally problematic, as the heat from the highly-active GPU logic would be trapped under sixteen layers of DRAM. The overall balance of CPU vs. GPU resources in this version of the EHP was also quite different from the first EHP. Given the topology of the IOD, we included four CCDs that provide a total of 32 cores (four times more than the original EHP), while the GPU silicon was reduced from three large, stacked dies to four smaller chiplets. From our other research efforts, the compute-per-mm^2 of the GPU was expected to be improved compared to our original assumptions of approximately 4TF per GPU die, but the overall balance was perhaps not ideal.

AMD packaging engineers also raised concerns about the asymmetry of the overall package (all CPU on one side, all GPU on the other). Routing I/O and other external memory would be more challenging with the IOD skewed off center. Asymmetric power

profiles between the CPU and GPU sides of the package could cause mechanical stresses on the silicon-package interface due to large temperature gradients and mismatches in the coefficient of thermal expansion (CTE) of the different die, bump, and package substrate materials potentially leading to die cracking, broken bumps, etc.

3.2 EHP Concept v3, Circa 2016

The v2 EHP just described attempted to align the CPU components to AMD's server roadmap, but this resulted in a variety of tradeoffs and concerns. In particular, it eventually became clear that 16-high memory stacks would be unlikely in the target timeframe. Figure 6 shows our third major iteration of the EHP, which AMD had previously detailed in 2017 [99].

The v3 EHP reverts to eight DRAM stacks as in the original concept, but it retains the 3D stacking of the DRAM above the GPU chiplets from the v2 EHP, which also enables a doubling of the GPU compute resources relative to the v2 EHP. To further increase compute density, the v3 EHP also proposed the usage of active interposers [54]. By moving as much non-compute resources (e.g., 2.5D die-to-die interfaces) from the GPU chiplets down to the active interposer, the GPU chiplet could be packed with more compute. Similar to how AMD's EPYC™ processors implement the IOD in an older, more cost-effective process node compared to the CCDs [73], the active interposer die (which acts like a 3D-stacked IOD) can utilize similar cost optimizations. As can be seen in the v2 EHP in Figure 5, the non-stacked IOD consumes a significant portion of the package real estate. To make room for all of the GPU resources, we also extended the active interposer organization to the CPU portion of the EHP.

Issues with EHP v3: While the height of the DRAM stacks had been brought back down to eight layers, the power density of the GPU regions still present thermal challenges as we have now replaced the passive interposer underneath the GPU chiplets with an active interposer. While technically feasible, the "triple stack" of DRAM on GPU on active interposer also significantly increases the manufacturing complexity, and in the coming years we would determine that such a capability would not be commercially available in time for the first generation of exascale platforms.

3.3 EHP Concept v4, Circa 2018

One of the main challenges with the prior iterations of the EHP was that we constrained the design to fit within the bounds of our standard server packaging (i.e., the SP3 package used by the first generations of EPYC™ servers). This created severe real estate pressure on the design that drove the aggressive use of 3D stacking (die-on-die and interposer-styled). As our research activities began in the PathForward program, we considered new packaging alternatives that could provide more room to work with. This included both internal packaging concepts as well as emerging external efforts such as what eventually became the Open Compute Project's (OCP) OCP Accelerator Module (OAM) [79].

Figure 7 shows the fourth iteration of the EHP, where we utilized the additional package real estate to back away from 3D stacking of CPU and GPU chiplets on top of active interposers. We also reconsolidated the GPU chiplets from eight smaller chiplets into two larger pieces of silicon. The reason behind this is that while there are significant cost benefits associated with smaller chiplets [73], the higher bandwidth required to support data movement and

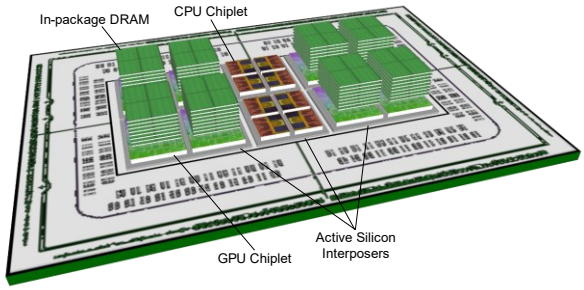


Figure 6. Refinement of the EHP (v3), circa 2016.

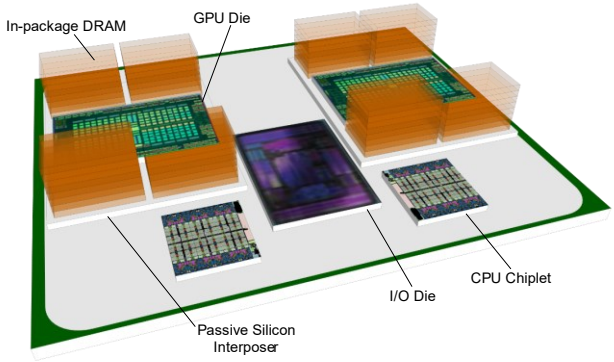


Figure 7. Refinement of the EHP (v4), circa 2018.

work distribution among the GPU compute units would be far less efficient to route among the larger number of chiplets. We still use a passive silicon interposer to connect each GPU chip with its four stacks of DRAM. At this point, we also rebalanced the CPU-to-GPU compute ratio, reducing the number of CCDs from eight in v3 down to two. This organization also allowed us to reuse the IOD and CCDs from the mainstream server parts as was done for the v2 EHP.

The two "islands" of GPU compute must contend with some physical distance between them. One option would be to route all GPU-to-GPU traffic through the IOD, utilizing the IOD's existing Infinity Fabric™ (IF) ports. However, the IODs' IF ports are provisioned for CPU levels of memory traffic (e.g., DDR) rather than what the GPU requires (i.e., HBM levels). Instead, we proposed to add multiple IF links directly across the package substrate between the GPU chips. The total distance is larger than the GPU-to-IOD spacing, but it is still manageable (comparable to the distance from the IOD to the farthest CCD in an EPYC™ CPU product).

Issues with EHP v4: This EHP represents the final version of our exascale APU concept at the end of the DOE's series of *Forward programs. The remaining concerns were less about fundamental technologies but rather more about business and other matters. One potential concern was that the EHP fixes the ratio of CPU-to-GPU compute resources for a given package, and customers with different workload requirements might prefer different ratios. Another concern is that, despite the larger package sizes assumed in this EHP concept, there still remains a substantial amount of silicon to be packed into a single module, which can make other

aspects such as power delivery and package I/O routing more difficult.

3.4 Discrete Node Architecture

From the very beginning of our exascale journey, we led with the vision of utilizing an APU architecture to provide the best of both CPUs and GPUs while making the processor easier to program by supporting a single unified memory system and cache coherence between the different processor types. However, our research also considered discrete node architectures (DNA) that could be constructed using a collection of discrete CPU and GPU/accelerator components. Figure 8 shows an example compute node with four CPUs and eight accelerators. Our proposed DNA still supports cache coherence and a flat physical address space like an APU, albeit at lower bandwidths and higher latencies given the pin and interconnect limitations between disparate packages.

Throughout the *Forward research activities, we maintained frequent discussions with multiple parties including various system integrators. The DNA was attractive to some partners because the separation of CPU and GPU components allowed them to customize platforms to provide different CPU-to-GPU ratios as well as to interoperate with other components. For example, some customers may have existing software investments in a different CPU ISA, but they are still interested in utilizing AMD's GPU accelerators.

4 Research Topics

Thus far, this paper has focused on the evolution of AMD's exascale node architecture. However, our *Forward activities conducted research in a wide range of topics in support of our broader exascale vision. This section provides a brief overview of some of this work, highlighting both successes and "deferred successes."

4.1 Compute-Optimized GPUs

A key research focus area was on optimizing and improving our GPU architectures to excel in general-purpose data-parallel compute. We asked the question: what could a "GPU" look like if it did not actually have to worry about graphics? We studied "compute-optimized GPUs" where we removed all of the specialized hardware that is only used for graphics rendering tasks (e.g., color units, spline interpolators, depth processing) and would otherwise be wasted dark silicon [39] in an HPC environment. The research activities covered other aspects of making GPUs easier to program for a wider variety of applications, such as defining scalable scoped synchronization models [52], expanding the functionality of AMD's ROCm™ open software platform, developing optimizations for DOE proxy workloads [14] on AMD GPUs, and circuit and power-optimization techniques to improve the GPU's power-performance efficiency. This compute-optimized GPU philosophy can now be seen in AMD's differentiated product lines where the RDNA™ architecture targets gaming and graphics and the CDNA™ architecture services HPC and machine learning. Researchers interested in exploring improvements to our CDNA architecture are encouraged to leverage our open-source gem5 GPU model [47].

4.2 CPU Core Microarchitecture

Although the majority of the computational horsepower for both EHP and DNA-based platforms comes from the GPU resources, in an age of widespread acceleration, Amdahl's Law cannot be forgotten. AMD's *Forward activities continued investing

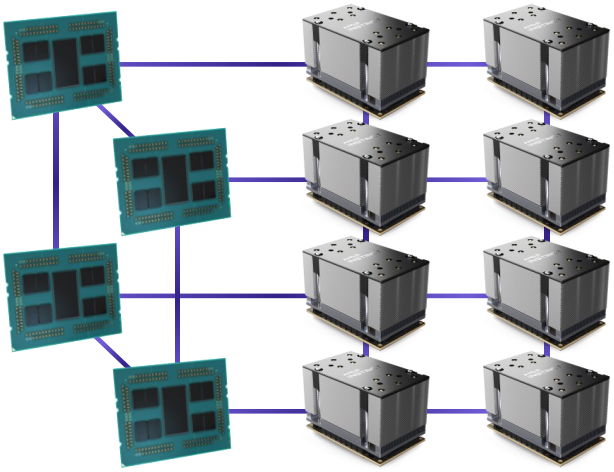


Figure 8. Discrete Node Architecture consisting of interconnected CPUs (left) and accelerators (right).

in researching techniques and enhancements for "traditional" CPU microarchitecture. Some of the work covered more general topics such as branch prediction, instruction fetch, scheduling, caching [59], and prefetching, but other activities focused on how to improve the CPU architecture specifically for the types of compute and memory patterns exhibited by the DOE's workloads, which do not always act like commercial benchmarks. A key message here is that while some in the academic research community may feel that "doing research" (i.e., publishing) in CPU microarchitecture seems harder these days, these are still topics of keen interest to industry, and we hope that the research conferences support and encourage such work.

4.3 Power-Performance Efficiency

Achieving the DOE's 1 exaflop in 20MW target (or 50 GF/W) required large improvements over the state of the practice at the time the *Forward programs began. As such, all our *Forward research plans included work on improving power-performance efficiency. This research included techniques for dynamically managing power and improving the efficiency of CPU and GPU microarchitectures [67], data movement and networks on chips [1][24], caches and memory [42][82], circuits[41] [104], and software algorithms. This research complemented the increased focus at AMD on power efficiency as a first-class design objective, which has expanded across our entire roadmap. This was broadly represented by our 25x20 initiative (which aimed to improve power-performance efficiency of AMD mobile processors by 25× from 2014 to 2020) [1], and our more recently updated 30x25 target for accelerated computing nodes [3]. We found that power-performance efficiency was a fruitful area of collaboration between research and product teams, as many mechanisms to improve power-performance efficiency in general-purpose server, consumer, and mobile designs could be directly applied to HPC designs and vice versa.

4.4 Reliability

The Reliability and Resiliency research explored three areas: understanding the nature of faults that occur in real HPC systems in the field [90][91], the development of early-stage architectural fault modeling techniques and tools for the EHP architecture and new fault modes [101], and exploring low-cost pervasive fault

detection techniques for GPUs, building on prior research on redundant execution techniques developed for CPUs [100]. All three areas influenced how we approached reliability, availability, and serviceability (RAS) design for our products and enabled several fruitful collaborations with the broader industry. The field studies were carried out in collaboration with the DOE National Labs and yielded valuable insights into CPU, GPU, and DRAM reliability. For example, the research helped AMD collaborate with the memory industry to drive improvements to the ECC architecture for HBM3 DRAM and then standardize that design at JEDEC [46]. AMD continues to study reliability field data at scale and make our findings available to the broader technical community [22]. Another key insight from our research was that targeted protection of hardware structures in the GPU, leveraging the fault modeling techniques we developed, is a cost-effective protection approach to enable their resilience at scale. More information about the overall body of research and advanced development that led to the RAS architecture of the Frontier node can be found in [21].

4.5 Programming Models and Software Optimization

Throughout the *Forward programs, AMD maintained numerous research projects in software systems. This work focused both on programming models and tools to best use platforms like the EHP, and on application- and algorithm-level research to optimize scientific codes for them. Both aspects of this software work were major drivers of co-design efforts with the DOE and were a strong bridge between AMD's research and the open-source software community.

Many of our software research projects involved co-design with the DOE to help develop programming tools for heterogeneous systems, which were in a nascent state at the beginning of the *Forward program. We developed runtimes such as ATMI [84], created models for accelerator-driven network interactions [50], ported programming languages such as Chapel [29] and APIs such as OpenMP® to AMD accelerators, and worked with the DOE to enable frameworks such as Kokkos [37] and RAJA [33] on AMD accelerators.

Large efforts went into porting DOE exascale proxy applications to these mechanisms, to provide testing of our models, feedback to proxy application creators, and training for both AMD and DOE developers about optimizing software for heterogeneous systems. Similar efforts went into porting proxy applications to languages such as OpenCL™, C++AMP, OpenACC, and HIP [32]. These porting efforts also led to research into new algorithms and optimizations to accelerate applications of interest on the envisioned exascale systems. These ranged from algorithmic primitives such as matrix computations [44][68] and graph analytics [28][32] to higher-level application work on scientific problems such as computational fluid dynamics [27][78].

4.6 Modeling and Simulation

While not a technology directly embedded in the final exascale system design, modeling, simulation, and overall projection of performance, power, silicon area, and more required significant effort in a variety of ways. The tools utilized over a decade of research spanned the gamut including spreadsheet-level models, analytical modeling, cycle-level simulation (e.g., gem5 [47][48], AMD-internal simulators), emulation (e.g., the AMD SimNow™ platform simulator), and performance and power measurements on real hardware [45]. Due to the range of analysis needed for the different studies and research topics, in many cases we had to combine results from multiple different tools, for example utilizing cycle-level simulation to characterize detailed kernel behaviors, observing scaling trends from real hardware measurements, and then synthesizing it in an analytical model to project/extrapolate to the full candidate node architecture designs. The composition of the multiple disparate tools was often imperfect, but for the research to make good progress, it is often better to have a rough answer in a short amount of time versus waiting weeks or months to perfect a much more detailed or unified tool or simulator, especially when a variety of assumptions are going to change and evolve as the research continues to progress.

The above summarizes just a few of the impactful research areas from our *Forward research. Below, we also discuss some of our research efforts that for various reasons did not make it into the first generation of exascale machines.

4.7 Multi-Level Memory

The original EHP concept (Figure 3) included a two-tier memory system consisting of in-package DRAM and external NVRAM. As discussed earlier, the NVRAM was omitted in subsequent iterations of the EHP concept. The "Holy Grail" of a multi-level memory architecture remains elusive, where one can provide the capacity (and cost) of NVRAM while delivering the bandwidth and latency of DRAM. AMD researched a wide range of hardware, software, and hybrid solutions [70][81][83][88][94], but for a range of reasons, we have not yet been entirely satisfied with any of the solutions. One key challenge is that many of the proposals work well on average, but they can still suffer from access patterns that cause performance to drop unacceptably and in a fashion that requires significant programmer effort to resolve. Finding effective multi-level memory architectures remains an important open research problem.

4.8 Processing in Memory

Data movement is a key challenge in exascale systems for both performance and power. PIM has always presented the promise of drastically reducing the cost of data movement by instead "moving your code to the data." Advances in die-stacking technology made it seem that PIM could be a possibility in the exascale timeframe. The AMD *Forward research studied many aspects of PIM including what types of compute should be co-located with memory, analyzing the types of workloads (or portions thereof) that could benefit from PIM offloading, programming .model implications, packaging and thermals, and more. The technology readiness of PIM did not align with the exascale schedules, but multiple recent industry announcements about PIM could indicate that its time may soon be coming [60][64][65].

4.9 Integrated Silicon Photonics

The desire to utilize integrated photonics in the exascale node architecture had a similar outlook to NVRAM and PIM. At the outset of our exascale research, there was already significant industrial and academic efforts prototyping and demonstrating multiple potential options for in-package optical interconnects [23][58][66][98]. Our explorations included packaging studies,

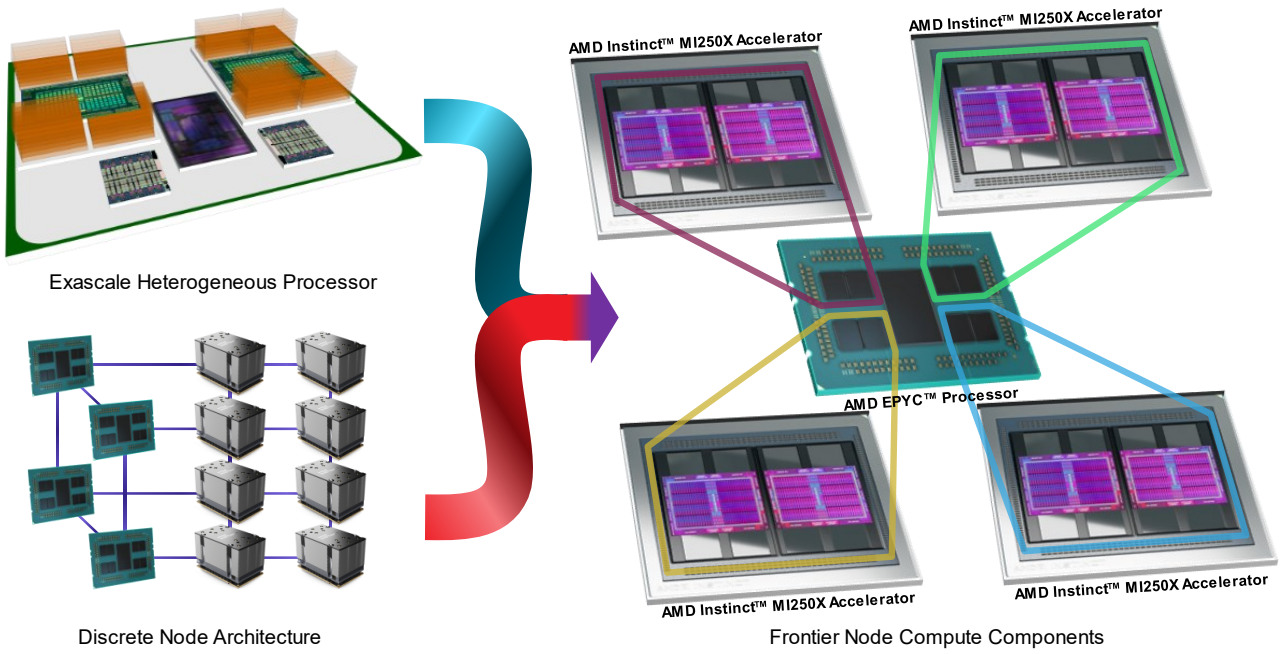


Figure 9. Synthesis of AMD exascale research concepts into the final components of the node architecture employed by the Frontier supercomputer.

architectural and memory system interfacing, usage of wave-division multiplexing, and interoperability with existing transport protocols. Similar to PIM, the amount of recent work and startups developing prototypes and testbeds brings hope that integrated photonics might not be too far off.

4.10 Asynchronous Data-dependent Tasking

While traditional bulk-synchronous execution is sufficient for many HPC applications, several emerging heterogeneous applications require extensive inter-thread communication or include tasks with a wide diversity of runtimes. Maximizing performance for these applications requires hardware and runtimes that support execution of asynchronous, multi-stream, and data-dependent tasks. Our team pursued a coordinated set of hardware [84][85] and software [15][16] enhancements to support these workloads and achieved impressive improvements for single-node HPC applications [56]. However, these enhancements have not yet impacted the official benchmark implementations used for acceptance testing because bulk synchronous implementations are better suited for current MPI primitives. Our work on eXtended Task Queueing (XTQ) [62] has the potential to simplify asynchronous task distribution across nodes, but additional hardware support for this technology is still required. As more heterogeneous accelerators may be included in future systems, developing robust support for scheduling and executing asynchronous tasks will be paramount.

5 Outcome

In May 2019, the U.S. DOE announced that it had contracted with Cray to build the Frontier supercomputer to be delivered to Oak Ridge National Laboratory [74]. The Frontier Node Architecture leverages concepts from both our EHP and DNA approaches in a way that intercepted our available technology options, desired roadmap alignment, and schedule constraints.

5.1 From EHP to the Frontier Node Architecture

The right side of Figure 9 shows the components of the Frontier compute node, which consists of an AMD EPYC™ 7A53 "Optimized 3rd Gen EPYC™" CPU cache-coherently coupled over Infinity Fabric™ links to four AMD Instinct™ MI250X accelerators. The figure visually highlights how this node design embodies a hybridization of the EHP and DNA approaches, albeit with the components distributed across different packaging boundaries than the EHP. Specifically, the primary compute components of the final v4 EHP (Figure 7) consisted of two CPU chiplets, two GPU accelerator dies, and eight stacks of DRAM, and so the Frontier node can effectively be viewed as a set of four EHP instances (highlighted by the different colored boxes on the right side of the figure) conjoined by the IOD of the EPYC™ CPU.

There were several architectural and broader technology drivers for this "disaggregated+conjoined EHP" organization. The AMD EPYC™ processor in the Frontier node extends AMD's chiplet strategy [73]; in this case we reuse the CPU chiplets from the mainstream server products, and the IOD heavily leverages existing IP with only the necessary modifications to support the shared memory and cache coherency with the GPU accelerators.

Removing the CPU and IOD components from the EHP package enabled the two GPU accelerator dies to be placed directly next to each other to maximize the efficiency and signal integrity of the high-speed die-to-die Infinity Fabric™ links (as opposed to having to cross a larger distance due to the presence of the IOD as was the case in the v4 EHP). This also greatly alleviated issues related to the limited package real estate. Thermal-mechanical concerns were also aided by maintaining symmetry in the respective CPU and GPU packages.

Placing the CPU and GPU accelerator resources in separate modules enables system integrators to mix and match components to

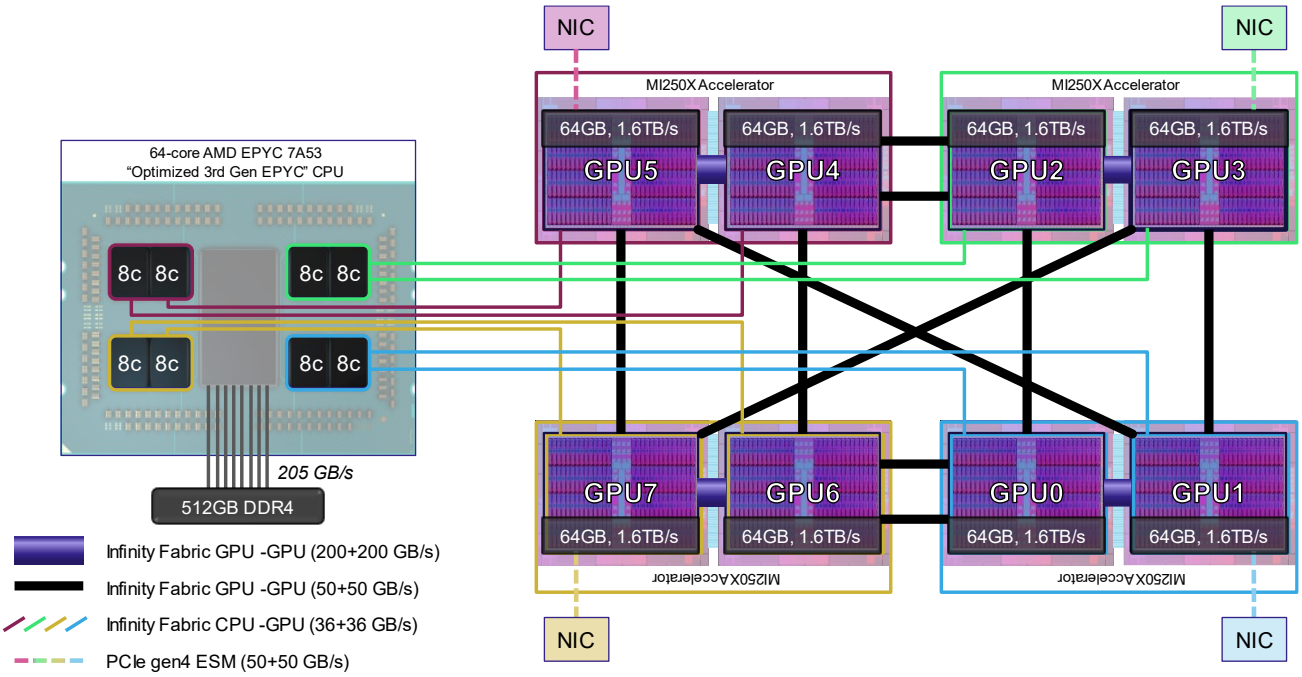


Figure 10. Block diagram of one Frontier Compute Node with peak theoretical memory and interconnect speeds. The "X+X GB/s" notation indicates X GB/s of bandwidth each for send and receive.

achieve different CPU:GPU ratios, different node sizes (e.g., 1P vs. 2P CPU configurations), and overall component interconnect topologies. While the Frontier node design utilizes a fixed CPU:GPU ratio, other systems can be designed that have other ratios to meet their respective target system requirements.

5.2 Frontier Technical Details

The overall components of the Frontier supercomputer have been detailed elsewhere [76], so here we only provide a brief summary. The overall machine provides 9,408 compute nodes housed in 74 cabinets. Each compute node has a 64-core EPYC™ 7A53 "Optimized 3rd Gen EPYC™" CPU. Each physical core provides two hardware threads (simultaneous multithreaded). The 64-core processor shares eight channels of DDR4 memory with 512GB of total capacity. Each node also has four AMD Instinct™ MI250X accelerators, each with two GPU compute dies and eight stacks of HBM2E memory supplying 128GB of capacity (64GB per GPU die). The two GPU dies implement the second-generation CDNA2 architecture [11] in a 6nm process, totaling 58 billion transistors [89]. Each AMD Instinct™ MI250X accelerator can deliver a peak vector double-precision performance of 47.9 TF and a peak of 3.2 TB/s of bandwidth from the eight DRAM stacks [10]. In addition to double-precision compute, the CDNA2 architecture also provides high-performance support for lower-precision operations common in artificial intelligence and machine learning workloads [6]. Significant advancements were also made for reliability requirements, for which a retrospective has already been published [21].

The MI250X compute dies and the HBM are co-packaged using AMD's Elevated Fan-out Bridge (EFB) technology that replaces what would otherwise be a very large passive silicon interposer (larger than reticle size) with multiple smaller silicon bridge chips built above ("elevated") the package substrate [95].

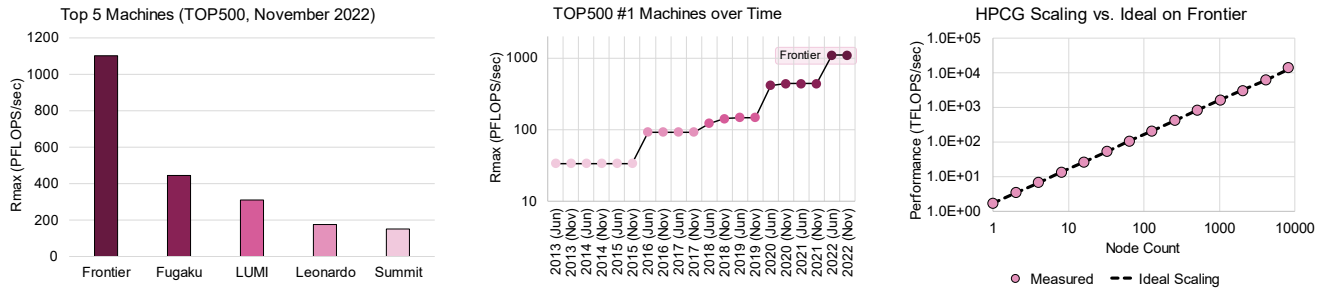
Figure 10 shows the block diagram of one Frontier node including the Infinity Fabric™ topology interconnecting the CPU and four accelerators. The eight CPU chiplets can be partitioned into four non-uniform memory access (NUMA) domains, with a pair of CCDs associated with one MI250X accelerator (effectively one logical EHP per NUMA domain). Within a NUMA domain, one CCD is associated with one of the GPU accelerator dies within the MI250X package.

The CPU and GPU accelerator components share a flat physical memory space (i.e., any processor can directly address the HBM in any of the four accelerators and all of the DDR4 connected to the CPU), and cache coherence is maintained among all processors. A consequence of extending our cache-coherent Infinity Fabric™ across the multiple packages is that it creates a single logical interconnect for both memory and I/O. Figure 10 also shows how each MI250X accelerator has a network interface (NIC) directly attached, which allows network data to be injected directly into an accelerator's local HBM without routing through the host CPU.

5.3 Software

The *Forward projects highlighted the significance of software on exascale-class systems. The systems are targeted at general scientific computation across a wide range of computational motifs and physical domains and therefore required software that could support the standard tools of HPC. These tools included network software (e.g., MPI), compilers (e.g., C++, Fortran, OpenMP®), and profiling/debugging tools. Furthermore, many applications desire *performance portability*, enabling them to compile and run an identical codebase across a range of computer architectures and systems.

AMD's answer for this is the ROCm™ open software platform [14], a full-stack open software platform encompassing



(a) (b) (c)

Figure 11. (a) LINPACK performance (RMAX) of the top five supercomputers on the TOP500 list as of November 2022, (b) the RMAX of the #1 supercomputer on the TOP500 list over the past decade, and (c) the scalability of the HPCG benchmark on Frontier (courtesy of HPE and Oak Ridge National Laboratory).

drivers/runtimes, programming models, compilers, libraries, and tools. ROCm enables programmers to quickly port existing codes to AMD hardware via its support for community standards such as OpenMP, BLAS, Tensorflow, PyTorch, etc., and porting accelerated codes from CUDA via AMD's HIP, a CUDA-like API. Another major mechanism for software portability was enabling AMD support (via HIP) for abstraction frameworks including Kokkos [37], RAJA [20], and OCCA [69]. Finally, the Frontier node architecture's Infinity Fabric™ coherent interconnect between the CPU and GPUs eases moving CPU-based codes to accelerators by reducing the code for explicit memory allocation and migrations.

5.4 Initial Results and Impacts

The operation of the Frontier supercomputer is still in its early days, and the primary focus of this paper is on the research journey and its impacts on the architecture of the machine. Nevertheless, Frontier has already delivered multiple computational accomplishments that we briefly highlight here to provide the reader with a more complete picture of what this architecture can accomplish at scale.

The LINPACK benchmark [35], often referred to as High-Performance LINPACK (HPL), is used by the TOP500 list for ranking supercomputers worldwide [96]. Figure 11(a) shows the top-five supercomputers (as of November 2022), and the results for Frontier officially exceeding the exaflop barrier as measured by HPL (RMAX). The achieved 1.1 exaflops (EF) exceeds the previous #1 supercomputer (Fugaku) by over 2×. Figure 11(b) shows RMAX (log scale) for the top supercomputers over the past decade, providing another view of the generational performance improvement that Frontier delivers.

As Frontier's main purpose is to enable scientific computations for real problems, raw computational throughput is necessary but not sufficient. While some computational problems are densely compute-bound as represented by LINPACK, many other workloads have sparser computational and data access patterns, global updates and reductions, and other behaviors that require both balanced node and system architectures that can make it challenging to scale to larger problems. The High-Performance Conjugate Gradients (HPCG) benchmark is commonly used to complement HPL to provide a more complete picture of supercomputer performance [36][51]. Figure 11(c) shows near-perfect scaling of HPCG when running across an increasing fraction of Frontier, enabled in part by having each node's NICs directly attached to the MI250X accelerators. While much of the attention on artificial intelligence and machine learning has been in the commercial sector (e.g., large-language models), there is rapidly increasing interest in utilizing AI/ML for HPC [92]. The CDNA2 architecture delivers robust performance for mixed-precision computations common in AI/ML, highlighted by Frontier placing #1 on the HPL-MXP mixed-precision benchmark rankings [53].

A key initial DOE exascale target was staying within a power budget of 20MW. At a delivered HPL performance of 1.1EF, Frontier consumes 21.1MW [96]. Normalizing this to 1.0EF, the power consumption is 19.2MW per EF. This high level of power-performance efficiency was reflected by the Frontier Test and Development (TDS) system and Frontier itself taking the #1 and #2 spots in the Green500 list when Frontier first premiered in June 2022 [97].

The Frontier supercomputer was only recently installed, and work continues on improving the overall system and software. However, new science is already being enabled on Frontier. Two different teams of scientists utilizing Frontier were finalists for the 2022 Gordon Bell Prize, with one team winning the overall prize [77]. The winning team used Frontier to perform 3D simulations of laser-matter interactions with applications in radiotherapy and high-energy physics [40]. The other finalist team used Frontier to perform natural language graph analytics on tens of millions of medical publications (going as far back as 1809) to discover previously unidentified links between different medical concepts and phenomena [55]. This medical application exceeded 1EF of sustained performance, far exceeding the original exascale RFI target of 300PF in a real workload [17].

6 Conclusions

It has been a decade-long journey from the original DOE exascale RFI to having ORNL's Frontier officially usher in the exascale era. Looking back, there are many lessons that we learned along the way. The first is that the co-design approach was very valuable. Our frequent interactions with the DOE scientists, application programmers, systems architects, and facilities managers, provided a more complete picture of the overall objectives and enabled us to better understand how to design our components to work effectively within the larger endeavor. The DOE's proxy applications also played a pivotal role in co-design, as the proxy apps provide not only code for analysis, but also served as a medium for dialogue that enabled a deeper understanding of what was really being sought in terms of computations, algorithms, and even the underlying science. Extending the co-design philosophy, we reaped huge benefits from generalizing this co-design mentality to

many of our interactions including with our system integrator partners, AMD's product and roadmap teams, university collaborators, and others.

Another lesson that we took away from our experience is the importance in having a research vision that is both bold and aggressive but can also stay flexible and adaptive given changing technology assumptions, business priorities, and other dynamic market conditions. We formulated an aggressive vision for heterogenous compute from the outset of our journey, but, as was described throughout this paper, the embodiment of that vision went through many iterations balancing the needs of pursuing aggressive targets against the practicalities of proposing something that in the end could be successfully executed by our world-class engineering teams. Related to this lesson is that we found great value in being able to pursue multiple solution paths, in particular our EHP versus DNA approaches. Key to this was avoiding a competitive scenario between the research thrusts, and instead we pursued the work in a cooperative manner that ultimately resulted in a Frontier node architecture that benefited from both efforts.

The final Frontier node architecture design was the result of a cooperative effort between our AMD Research team and multiple product teams and business units. The final design decisions were driven by various technical factors (e.g., technology availability/high-volume manufacturing readiness, schedule, risk, cost), as well as non-technical constraints (e.g., market conditions, competition, company directions, resource availability).

Our exascale adventure would not have happened without the strong public-private partnership between AMD and the DOE and other U.S. agencies. HPC is often a harbinger for technology trends that eventually impact other market segments, and so the DOE's exascale research investments not only accelerate technology advancements for HPC, but many of the benefits end up influencing other products well beyond the HPC realm.

It is incredibly exciting to have crossed the exascale threshold, but it is just as important to keep in mind that exascale is only a milestone. The world's need for ever increasing computational capabilities is not slowing down. As our collective scientific understanding improves, the problems get bigger and more difficult. The explosion of machine learning and artificial intelligence add new computational requirements to future supercomputers. This is a great time for researchers to innovate and contribute to the community's collective efforts to continue driving forward in this new post-exascale era.

Acknowledgements

This decade-long effort was enabled by many people along the way. From AMD, we thank Jane Butz, Nick George, Mark Heene, John Keaty, Andrew Kegel, Jay Owen, Rita Pillman, and Roland Schwarz for their tireless support for project/program management, budgeting, staffing, and countless other logistical matters; AMD Legal, especially Beth Apperley, Kim Vo, and Patricia Lange in their guidance in navigating contracts and other legal matters over the multiple projects; and key AMD technical leaders including Frank Helms, Nathan Kalyanasundharam, Kevin Lepak, Joe Macri, Mike Mantor, Sam Naffziger, Ben Sander, and Alan Smith. We also thank our many other AMD collaborators and contributors, interns, and postdocs, all of which are too numerous to individually name here, as well as the support from our AMD executive team. We also thank our visiting scholars Wayne Burleson, Natalie Enright Jerger, Mark Hill, Avinash Karanth Kodi, Hyesoon Kim, Mark Oskin, Apan Qasem, Matt Sinclair, Mithuna Thottethodi, Guru Venkataramani, and David Wood. We also thank our technical representatives and others in the DOE leadership including Jim Ang, Scott Atchley, Jonathan Carter, Jeanine Cook, Bronis de Supinski, Al Geist, Robin Goldstone, Si Hammond, Bill Harrod, Barb Helland, Thuc Hoang, Edgar Leon Borja, Josip Loncaric, John May, Arun Rodrigues, John Shalf, and Pavlos Vranas, as well as the many other DOE scientists who worked with us across a myriad of co-design and other interactions. We also thank our many collaborators from our system integrator partners.

We dedicate this paper to the memory of Chuck Moore, who provided the initial vision, guidance, and support in response to the DOE's initial exascale RFI and set our collective team on our first steps down this incredible journey.

References

[1] V. Adhinarayanan, I. Paul, J. L. Greathouse, W. Huang, A. Pattnaik, W. Feng, "Measuring and Modeling On-Chip Interconnect Power on Real Hardware," Proceedings of the IEEE International Symposium on Workload Characterization (IISWC), September 2016.

[2] Advanced Micro Devices, Inc., "AMD 25x20 Energy Efficiency Initiative," 2014, https://www.amd.com/en/technologies/25x20.

[3] Advanced Micro Devices, Inc., "AMD Announces Ambitious Goal to Increase Energy Efficiency of Processors Running AI Training and High Performance Computing Applications 30x by 2025," Press Release, September 2021.

[4] Advanced Micro Devices, Inc., "AMD Awarded $32 Million for 'Extreme Scale' High-Performance Computing Research Focused on I, APUs and Memory," Press Release, November 2014.

[5] Advanced Micro Devices, Inc., "AMD Fusion APU Codenamed "Llano" Demonstrated at the 6th Annual AMD Technical Forum & Exhibition 2010," Press Release, October 2010.

[6] Advanced Micro Devices, Inc., "AMD Instinct™ MI200 Adopted for Large-Scale AI Training in Microsoft Azure," Press Release, May 2022.

[7] Advanced Micro Devices, Inc., "AMD Selected by U.S. Government to Help Engineer and Shape the Future of High Performance Computing," Press Release, July 2012.

[8] Advanced Micro Devices, Inc., "AMD Selected by the U.S. Department of Energy to Help Drive Next-Generation Supercomputing Hardware Architecture," Press Release, June 2017.

[9] Advanced Micro Devices, Inc., "AMD Ushers in a New Era of PC Gaming with Radeon™ R9 and R7 300 Series Graphics Line-Up including World's First Graphics Family with Revolutionary HBM Technology," Press Release, June 2015.

[10] Advanced Micro Devices, Inc., "AMD Instinct™ MI200 Series Accelerator," 2022, https://www.amd.com/system/files/documents/amd-instinct-mi200-datasheet.pdf.

[11] Advanced Micro Devices, Inc., "Introducing AMD CDNA™ 2 Architecture," 2021, https://www.amd.com/system/files/documents/amd-cdna-whitepaper.pdf

[12] Advanced Micro Devices, Inc., "Shared Resilience: AMD Response to COVID-19," https://www.amd.com/en/corporate/amd-covid-19-response, 2020.

[13] Advanced Micro Devices, Inc., "U.S. Government Awards AMD Contract to Research Interconnect Architectures for High-Performance Computing," Press Release, November 2013.

[14] Advanced Micro Devices, Inc., "AMD ROCm Open Software Platform for GPU Compute," http://www.amd.com/ROCm, 2022.

[15] Advanced Micro Devices, Inc., "ATMI (Asynchronous Task and Memory Interface)," https://github.com/RadeonOpenCompute/atmi.

[16] Advanced Micro Devices, Inc., "DAGEE (Directed Acyclic Graph Execution Engine)," https://github.com/AMDResearch/DAGEE.

[17] Argonne National Laboratory, Request for Information (RFI) No. 1-KD73-I-31583-00, "Project: Exascale Research and Development," July 2011.

[18] M. Arora, S. Manne, Y. Eckert, I. Paul, N. Jayasena, D. Tullsen, "A Comparison of Core Power Gating Strategies Implemented in Modern Hardware," Proc. Of the ACM Int'l Conf. on Measurement and Modeling of Computer Systems (SIGMETRICS), June 2014.

[19] K. J. Barker, K. Davis, A. Hoisie, D. J. Kerbyson, M. Lang, S. Pakin, J. C. Sancho, "Entering the petaflop era: The architecture and performance of Roadrunner," Proceedings of the Supercomputer (SC) conference, 2008.

[20] D. A. Beckingsale, J. Burmark, R. Hornung, H. Jones, W. Killian, A. J. Kunen, O. Pearce, P. Robinson, B. S. Ryujin, T. R. W. Scogland, "RAJA: Portable Performance for Large-Scale Scientific Applications," International Workshop on Performance, Portability and Productivity in HPC (P3HPC), 2019.

[21] M. V. Beigi, V. Sridharan, S. Gurumurthi, "Reliability, Availability, and Serviceability Challenges for Heterogeneous System Design," IEEE International Reliability Physics Symposium (IRPS), March 2022.

[22] M. V. Beigi, Y. Cao, S. Gurumurthi, C. Recchia, A. Walton, V. Sridharan, "A Systematic Study of DDR4 DRAM Faults in the Field," IEEE International Symposium on High-Performance Computer Architecture, February 2023.

[23] A. F. Benner, M. Ignatowski, J. A. Kash, D.M. Kuchta and M. B. Ritter, "Exploitation of optical interconnects in future server architectures," IBM Journal of Research and Development, vol. 49, no. 4/5, pp. 755, July–September 2005.

[24] S. Bharadwaj, S. Das, Y. Eckert, M. Oskin, T. Krishna, "DUB: Dynamic Underclocking and Bypassing in NOCs for Heterogeneous GPU Workloads," Proceedings of the International Symposium on Networks-on-Chip (NOCS), 2021.

[25] B. Black, "Die Stacking is Happening," MICRO Keynote, IEEE/ACM International Symposium on Microarchitecture, December 2013.

[26] M. Bohr, "A 30 Year Retrospective on Dennard's MOSFET Scaling Paper," in IEEE Solid-State Circuits Society Newsletter, vol. 12, no. 1, pp. 11-13, 2007.

[27] N. Chalmers, A. Mishra, D. McDougall, T. Warburton, "HipBone: A performance-portable GPU-accelerated C++ version of the NekBone benchmark," arXiv preprint arXiv:2202.12477, 2022.

[28] S. Che, M. Orr, G. Rodgers, J. Gallmeier, "Betweenness Centrality in an HSA-enabled System," Proceedings of the ACM Workshop on High Performance Graph Processing (HPGP), 2016.

[29] M. Chu, A. Aji, D. Lowell, K. Hamidouche, "GPGPU Support in Chapel with the Radeon Open Compute Platform," at the 4th Annual Chapel Implementers and Users Workshop, 2017.

[30] K.-W. Chung, S. Shih, S.-T. Lu, T.-H. Chen, C.-T. Chen, J. Ho, J.-J. Chen, J.- P. Lin, "3D Stacking DRAM using TSV Technology and Microbump Interconnect," in the International Microsystems Packaging Assembly and Circuits Technology Conference, October 2010.

[31] P. Conway, N. Kalyanasundharam, G. Donley, K. Lepak, B. Hughes, "Cache Hierarchy and Memory Subsystem of the AMD Opteron Processor," IEEE Micro, Vol. 30, Issue 2, March-April 2010.

[32] M. Daga, M. Nutter, M. Meswani, "Efficient Breadth-First Search on a Heterogeneous Processor," Proceedings of the IEEE International Conference on Big Data, 2014.

[33] M. Daga, Z. S. Tschirhart, C. Freitag, "Exploring Parallel Programming Models on Heterogeneous Computing Systems," Proceedings of the IEEE International Symposium on Workload Characterization (IISWC), 2015.

[34] Y. Deng and W. Maly. "Interconnect Characteristics of 2.5-D System Integration Scheme," Proceedings of the International Symposium on Physical Design, pages 171–175, April 2001.

[35] J. Dongarra, "Performance of Various Computers Using Standard Linear Equations Software," Computer Science Technical Report Number CS – 89 – 85, 2022, http://www.netlib.org/benchmark/performance.ps

[36] J. Dongarra, M. A. Heroux, "Toward a New Metric for Ranking High Performance Computing Systems," Sandia National Laboratories Technical Report, SAND2013-4744, June 2013.

[37] H. Carter Edwards, Christian R. Trott, and Daniel Sunderland. "Kokkos: Enabling manycore performance portability through polymorphic memory access patterns." Journal of parallel and distributed computing 74.12 (2014): 3202-3216.

[38] D. G. Elliott, W. M. Snelgrove, and M. Stumm, "Computational RAM: A Memory-SIMD Hybrid and its Application to DSP," Proceedings of the Custom Integrated Circuits Conference, Boston, MA, May 1992.

[39] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, D. Burger, "Dark silicon and the end of multicore scaling," Proceedings of the International Symposium on Computer Architecture, June 2011.

[40] L. Fedeli, A. Huebl, F. Boillod-Cerneux, T. Clark, K. Gott, C. Hillairet, S. Jaure, A. Leblanc, R. Lehe, A. Myers, C. Piechurski, M. Sato, N. Zaïm, W. Zhang, J.-L. Vay, H. Vincenti, "Pushing the Frontier in the Design of Laser-Based Electron Accelerators with Groundbreaking Mesh-Refined Particle-In-Cell Simulations on Exascale-Class Supercomputers," Proceedings of the Supercomputing (SC) conference, November 2022.

[41] S. Ganapathy, J. Kalamatianos, K. Kasprak, S. Raasch, "On Characterizing Near-Threshold SRAM Failures in FinFET Technology," Proceedings of the Design Automation Conference (DAC), June 2017.

[42] S. Ganapathy, J. Kalamatianos, B. M. Beckmann, S. Raasch, L. G. Szafaryn, "Killi: Runtime Fault Classification to Deploy Low Voltage Caches without MBIST," Proceedings of the International Symposium on High Performance Computer Architecture (HPCA), February 2019.

[43] M. Gokhale, B. Holmes, K. Iobst, "Processing in Memory: The Terasys Massively Parallel PIM Array," IEEE Computer, 28(42):23–31, 1995.

[44] J. L. Greathouse, M. Daga, "Efficient Sparse Matrix-Vector Multiplication on GPUs using the CSR Storage Format," Proceedings of the Supercomputer (SC) Conference, 2014.

[45] J. L. Greathouse, A. Lyashevsky, M. Meswani, N. Jayasena, M. Ignatowski, "Simulation of Exascale Nodes through Runtime Hardware Monitoring," ASCR Workshop on Modeling & Simulation of Exascale Systems & Applications (ModSim), September, 2013.

[46] S. Gurumurthi, K. Lee, M. Jang, V. Sridharan, A. Nygren, Y. Ryu, K. Sohn, T. Kim, H. Chung, "HBM3 RAS: Enhancing Resilience at Scale," IEEE Computer Architecture Letters, 20(2), pp. 158-161, Dec. 2021.

[47] A. Gutierrez, S. Puthoor, B. M. Beckmann, T. Ta, "AMD gem5 APU Simulator: Modeling GPUs Using the Machine ISA," tutorial held in conjunction with the International Symposium on Computer Architecture, June 2018.

[48] A. Gutierrez, B. M. Beckmann, A. Dutu, J. Gross, M. LeBeane, J. Kalamatianos, O. Kayiran, M. Poremba, B. Potter, S. Puthoor, M. D. Sinclair, M. Wyse, J. Yin, X. Zhang, A. Jain, T. G. Rogers, "Lost in Abstraction: Pitfalls of Analyzing GPUs at the Intermediate Language Level," Proceedings of the International High Performance Computer Architecture (HPCA), February 2018.

[49] M. Hall, P. Kogge, J. Koller, P. Diniz, J. Chame, J. Draper, J. LaCoss, J. Granacki, J. Brockman, A. Srivastava, W. Athas, V. Freeh, J. Shin, J. Park, "Mapping Irregular Applications to DIVA, a PIM-based Data-Intensive Architecture," Proceedings of the Supercomputer (SC) conference, 1999.

[50] K. Hamidouche, M. LeBeane, "GPU Initiated OpenSHMEM: Correct and Efficient Intra-Kernel Networking for dGPUs," Proceedings of the Symposium on Principles and Practice of Parallel Programming (PPoPP), 2020.

[51] M. A. Heroux, J. Dongarra, P. Luszczek, "HPCG Technical Specification," Sandia National Laboratories Technical Report, SAND2013-8752, October 2013.

[52] D. R. Hower, B. A. Hechtman, B. M. Beckmann, B. R. Gaster, M. D. Hill, S. K. Reinhardt, D. A. Wood, "Heterogeneous-race-free memory models," Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems, 2014.

[53] HPL-MXP, "Mixed-Precision Benchmark," June 2022, https://hpl-mxp.org/

[54] A. Kannan, N. Enright Jerger, G. H. Loh, "Enabling Interposer-based Disintegration of Multi-core Processors," in the International Symposium on Microarchitecture (MICRO), December 2015.

[55] R. Kannan, P. Sao, H. Lu, J. Kurzak, G. Schenk, Y. Shi, S.-H. Lim, S. Israni, V. Thakkar, G. Cong, R. Patton, S. E. Baranzini, R. Vuduc, T. Potok, "Exaflops Biomedical Knowledge Graph Analytics," Proceedings of the Supercomputing (SC) conference, November 2022.

[56] A. M. Kaushik, A. M. Aji, M. A. Hassaan, N. Chalmers, N. Wolfe, S. Moe, S. Puthoor, B. M. Beckmann, "Optimizing Hyperplane Sweep Operations Using Asynchronous Multi-grain GPU Tasks," Proceedings of the IEEE International Symposium on Workload Characterization (IISWC), 2019.

[57] Y. Kim, T.-D. Han, S.-D. Kim, S.-B. Yang, "An Effective Memory Processor Integrated Architecture for Computer Vision," Proceedings of the International Conference on Parallel Processing, pages 266–269, August 1997.

[58] N. Kirman, M. Kirman, R. K. Dokania, J. F. Martinez, A. B. Apsel, M. A. Watkins, D. H. Albonesi, "Leveraging Optical Technology in Future Bus-based Chip Multiprocessors," Proceedings of the International Symposium on Microarchitecture, December 2006.

[59] A. Kokolis, N. Mantri, S. Ganapathy, J. Torrellas, J. Kalamatianos, "Cloak: Tolerating Non-Volatile Cache Read Latency," Proceedings of the International Conference on Supercomputing (ICS), June 2022.

[60] Y-C. Kwon, S. H. Lee, J. Lee, S-H. Kwon, J. M. Ryu, J-P. Son, S. O, H-S. Yu, H. Lee, S. Y. Kim, Y. Cho, J. G. Kim, J. Choi, H-S. Shin, J. Kim, B. Phuah, H. Kim, M. J. Song, A. Choi, D. Kim, S. Kim, E-B. Kim, D. Wang, S. Kang, Y. Ro, S. Seo, J. Song, J. Youn, K. Sohn, N. S. Kim, "A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications," in the International Solid-State Circuits Conference (ISSCC), February 2021.

[61] Lawrence Livermore National Laboratory, "Memory/Processor Research & Development," Request for Proposal (RFP) Number B599858, March 2012.

[62] M. LeBeane, B. Potter, A. Pan, A. Dutu, V. Agarwala, W. Lee, D. Majeti, B. Ghimire, E. van Tassell, S. Wasmundt, B. Benton, M. Breternitz, M. L. Chu, M. Thottethodi, L. K. John, S. K. Reinhardt, "Extended Task Queuing: Active Messages for Heterogeneous Systems," Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2016.

[63] B. C. Lee, E. Ipek, O. Mutlu, D. Burger, "Architecting Phase Change Memory as a Scalable DRAM Replacement," Proceedings of the International Symposium on Computer Architecture, June 2009.

[64] S. Lee, S.-H. K., J. Lee, H. Kim, E. Lee, S. Seo, H. Yoon, S. Lee, K. Lim, H. Shin, J. Kim, S. O, A. Iyer, D. Wang, K. Sohn, N. S. Kim, "Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology," Proceedings of the International Symposium on Computer Architecture, June 2021.

[65] S. Lee, K. Kim, S. Oh, J. Park, G. Hong, D. Ka, K. Hwang, J. Park, K. Kang, J. Kim, J. Jeon, N. Kim, Y. Kwon, K. Vladimir, W. Shin, J. Won, M. Lee, H. Joo, H. Choi, J. Lee, D. Ko, Y. Jun, K. Cho, I. Kim, C. Song, C. Jeong, D. Kwon, J. Jang, I. Park, J. Chun, J. Cho, "A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-in-Memory Supporting 1TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications," in the International Solid-State Circuits Conference (ISSCC), February 2022.

[66] A. Louri, A. K. Kodi, "SYMNET: An Optical Interconnection Network for Scalable High-Performance Symmetric Multiprocessors," Applied Optics, 42(17), Jun 2003.

[67] A. Majumdar, L. Piga, I. Paul, J. L. Greathouse, W. Huang, D. H. Albonesi, "Dynamic GPGPU Power Management using Adaptive Model Predictive Control," Proceedings of the International Symposium on High Performance Computer Architecture (HPCA), 2017.

[68] N. Malaya, S. Che, J. L. Greathouse, R. van Oostrum, M. J. Schulte, "Accelerating Matrix Processing with GPUs," Proceedings of the Symposium on Computer Arithmetic (ARITH), 2017.

[69] D. Medina, A. St-Cyr, T. Warburton, "OCCA: A unified approach to multi-threading languages," arXiv preprint arXiv:1403.0968, 2014.

[70] 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," Proceedings of the International Symposium on High Performance Computer Architecture (HPCA), February 2015.

[71] S. Mittal, J. S. Vetter, "A Survey of Software Techniques for Using Non-volatile Memories for Storage and Main Memory Systems," IEEE Transactions on Parallel and Distributed Systems, Vol. 27, Issue 5, pp. 1537-1550, June 2015.

[72] G. E. Moore, "Cramming More Components onto Integrated Circuits," in Electronics, Vol. 38, No. 8, April 1965.

[73] 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," in the International Symposium on Computer Architecture (ISCA), June 2021

[74] Oak Ridge National Laboratory, "No Scaling Back: DOE, Cray, AMD to Bring Exascale to ORNL," May 2019, https://www.olcf.ornl.gov/2019/05/07/no-scaling-back-doe-cray-amd-to-bring-exascale-to-ornl/

[75] Oak Ridge National Laboratory, "Frontier Supercomputer Debuts as World's Fastest, Breaking Exascale Barrier," https://www.ornl.gov/news/frontier-supercomputer-debuts-worlds-fastest-breaking-exascale-barrier

[76] Oak Ridge National Laboratory, "Frontier User Guide," 2022, https://docs.olcf.ornl.gov/systems/frontier_user_guide.html

[77] Oak Ridge National Laboratory, "Plasma Simulation Code Wins 2022 ACM Gordon Bell Prize," https://www.olcf.ornl.gov/2022/11/17/plasma-simulation-code-wins-2022-acm-gordon-bell-prize

[78] O. Obiols-Sales, A. Vishnu, N. Malaya, A. Chandramowliswharan, "CFDNet: a Deep Learning-based Accelerator for Fluid Simulations," Proceedings of the International Conference on Supercomputing (ICS), 2020.

[79] Open Compute Project, "OCP Accelerator Module (OAM) Design Specification v1.5," October 2021, http://www.opencompute.org/documents/ocp-accelerator-module-design-specification-v1p5-final-20220223-docx-1-pdf

[80] M. Oskin, F. T. Chong, T. Sherwood, "Active Pages: A Computation Model for Intelligent Memory," Proceedings of the International Symposium on Computer Architecture, pages 192–203, June 1998.

[81] M. Oskin, G. H. Loh, "A Software-managed Approach to Die-Stacked DRAM," Proceedings of the International Conference on Parallel Architectures and Compilation Techniques (PACT), October 2015.

[82] I. Paul, W. Huang, M. Arora, S. Yalamanchili, "Harmonia: Balancing Compute and Memory Power in High-Performance GPUs," Proceedings of the International Symposium on Computer Architecture (ISCA), 2015.

[83] A. Prodromou, M. Meswani, N. Jayasena, G. H. Loh, D. Tullsen, "MemPod: A Clustered Architecture for Efficient and Scalable Migration in Flat Address Space Multi-Level Memories," Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA), February 2017.

[84] S. Puthoor, A. M. Aji, S. Che, M. Daga, W. Wu, B. M. Beckmann, G. Rodgers, "Implementing Directed Acyclic Graphs with the Heterogeneous System Architecture," Proceedings of the Workshop on General Purpose Processing using Graphics Processing Unit (GPGPU), 2016.

[85] S. Puthoor, X. Tang, J. Gross, B. M. Beckmann, "Oversubscribed Command Queues in GPUs," Proceedings of the 11th Workshop on General Purpose GPUs (GPGPU), 2018.

[86] M. Qureshi, V. Srinivasan, J. A. Rivers, "Scalable High Performance Main Memory System Using Phase-Change Memory Technology," Proceedings of the International Symposium on Computer Architecture, June 2009.

[87] M. J. Schulte, M. Ignatowski, G. H. Loh, B. M. Beckmann, W. C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S. K. Reinhardt, G. Rodgers, "Achieving Exascale Capabilities through Heterogeneous Computing," IEEE Micro, July-August 2015.

[88] J. Sim, G. H. Loh, V. Sridharan, M. O'Connor, "Resilient Die-stacked DRAM Caches," Proceedings of the International Symposium on Computer Architecture (ISCA), June 2013.

[89] A. Smith, N. James, "AMD Instinct™ MI200 Series Accelerator and Node Architectures," in Hot Chips, August 2022.

[90] V. Sridharan, J. Stearley, N. DeBardeleben, S. Blanchard, S. Gurumurthi, Feng Shui of Supercomputer Memory - Positional Effects in DRAM and SRAM Faults, Proceedings of the Supercomputing (SC) conference, November 2013.

[91] V. Sridharan, N. DeBardeleben, S. Blanchard, K. Ferreira, J. Stearley, J. Shalf, S. Gurumurthi, Memory Errors in Modern Systems: The Good, The Bad, and the Ugly, Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Istanbul, Turkey, March 2015.

[92] R. Stevens, V. Taylor, J. Nichols, A. B. Maccabe, K. Yelick, D. Brown, "AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science," ANL-20/17 158802. February 2020.

[93] H. S. Stone, "A Logic-in-Memory Computer," IEEE Transactions on Computers, 19(1):73–78, January 1970.

[94] C. Sun, E. A. Leon, G. H. Loh, D. Roberts, K. Cameron, D. S. Nikolopoulos, B. R. de Supinski, "HpMC: An Energy-aware Management System of Multi-level Memory Architectures," in the International Symposium on Memory Systems (MEMSYS), October 2015.

[95] R. Swaminathan, "Case Study: AMD Products Built with 3D Packaging (Tutorial)," in Hot Chips, August 2021.

[96] TOP500, http://www.top500.org/

[97] TOP500, "Green500," June 2022, Green500, June 2022, https://www.top500.org/lists/green500/2022/06/

[98] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, J. H. Ahn, "Corona: System Implications of Emerging Nanophotonic Technology," Proceedings of the International Symposium on Computer Architecture, June 2008.

[99] T. Vijayaraghavan, Y. Eckert, G. H. Loh, M. Schulte, M. Ignatowski, I. Paul, B. Beckmann, S. K. Reinhardt, W. Brantley, J. Greathouse, O. Kayiran, M. Poremba, W. Huang, A. Karunanithi, G. Sadowski, V. Sridharan, S. Raasch, M. Meswani, "Design and Analysis of an APU for Exascale Computing," in the International Symposium on High-Performance Computer Architecture (HPCA), February 2017.

[100] J. Wadden, A. Lyashevsky, S. Gurumurthi, V. Sridharan, K. Skadron, Real-World Design and Evaluation of Compiler Managed GPU Redundant Multithreading, Proceedings of the International Symposium on Computer Architecture, Minneapolis, MN, June 2014.

[101] M. Wilkening, V. Sridharan, D. Kaeli, S. Li, F. Previlon, S. Gurumurthi, Calculating Architectural Vulnerability Factors for Spatial Multi-Bit Transient Faults. Proceedings of the International Symposium on Microarchitecture, Cambridge, UK, December 2014.

[102] W. A. Wulf, S. A. McKee, "Hitting the Memory Wall: Implications of the Obvious," ACM SIGARCH Computer Architecture News, Volume 23, Issue 1, pp. 20–24, March 1995.

[103] P. Zhou, B. Zhao, J. Yang, Y. Zhang, "A Durable and Energy Efficient Main Memory Using Phase Change Memory Technology," Proceedings of the International Symposium on Computer Architecture, June 2009.

[104] Y. Zu, W. Huang, I. Paul, V. J. Reddi, "Ti-states: Processor Power Management in the Temperature Inversion Region," Proceedings of the International Symposium on Microarchitecture (MICRO), 2016.