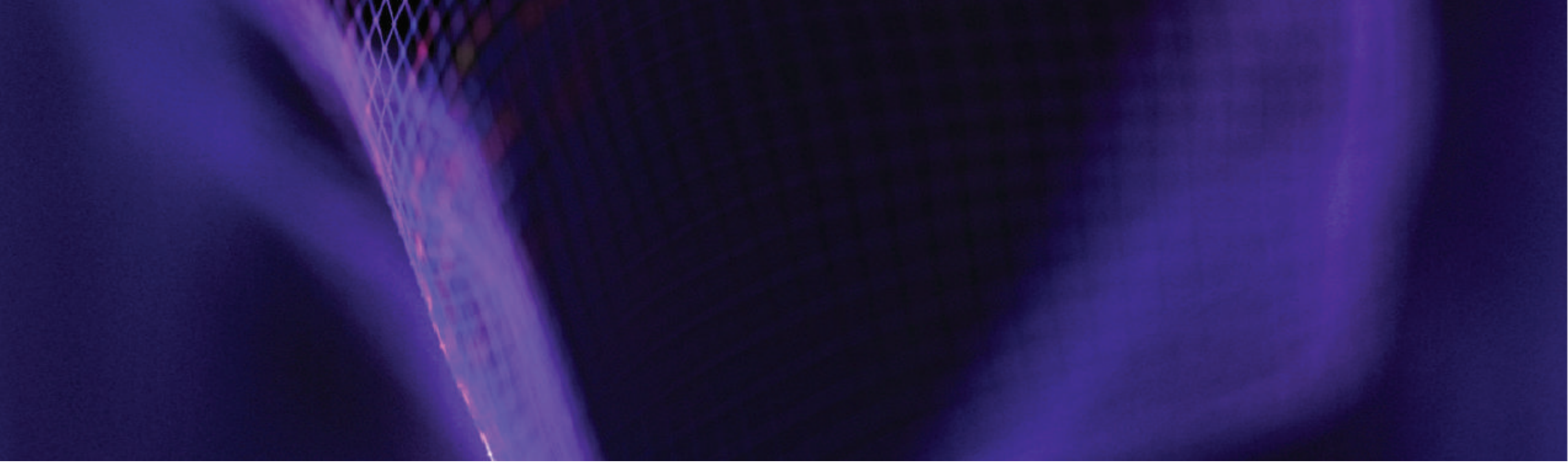
WHITE PAPER

HOW AMD IS RAISING THE BAR FOR HIGH-PERFORMANCE SOLUTIONS IN THE PUBLIC SECTOR

together we advance data center computing

Second Edition
September 2024

AMD: Harini Malik



HOW AMD IS RAISING THE BAR FOR HIGH-PERFORMANCE SOLUTIONS IN THE PUBLIC SECTOR

CONTENTS

REQUIREMENTS FOR HPC AND AI IN THE PUBLIC SECTOR

Public sector investments in exascale systems have significantly impacted technology innovation and put the United States and other leading economies in an advantageous position. Supercomputers have led to advances in energy and materials research, new manufacturing techniques, our understanding of climate, and the development of new treatments and therapeutics – all technologies critical to sustaining and advancing economic leadership. Investments in high-performance computing (HPC) are critical to maintaining a robust economy.

While HPC was once entirely the purview of large national labs (DOE, NOAA, NASA, and DOD), it has become pervasive. HPC clusters are now operated by both public and private sector organizations in diverse fields, from engineering simulation to aircraft design to life sciences to financial risk analytics. The combination of plummeting compute and storage costs, fast networks, and widespread access to cloud-based infrastructure have substantially lowered the barriers to entry.

The boom in AI, powered by modern CPUs, GPUs, and APUs, is leading to new types of systems designed for both HPC and AI. While generative AI poses public policy, safety, and ethical challenges, it also represents a tremendous opportunity. AI promises to accelerate economic development, accelerate digital transformation, and help improve the efficiency of government service delivery.

Governments and public institutions have an essential role in capitalizing on AI's promise and ensuring its use for the public good.

HPC and AI infrastructure are being deployed across public sector organizations to help realize these opportunities. To keep pace with growing demand, these organizations need to modernize their infrastructure and build a technology foundation that is scalable, flexible, sustainable, cost-efficient, and secure.

In this paper, we discuss critical requirements for high-performance systems and how foundational technologies from AMD are helping raise the bar for high-performance solutions. We also explain how public sector organizations can benefit from these investments in exascale systems, enabling them to accelerate the deployment of new applications and services and realize the transformative potential of AI.

CRITICAL REQUIREMENTS FOR HIGH-PERFORMANCE SYSTEMS

In 2008, the Exascale Study Group, supported by the Defense Advanced Research Projects Agency (DARPA), published a paper identifying the top barriers to building a supercomputer capable of executing a quintillion operations per second. Key barriers, illustrated in Figure 1, included power consumption, data movement, and achieving billion-way parallelism.[1]

Based on technologies foreseeable at the time, powering an exascale supercomputer would require an estimated **600 megawatts** of electricity, equivalent to the output of at least two small nuclear reactors.[2] The annual electricity bill for such a system would be an estimated **$600M** – well beyond the operating budget of any supercomputing facilities.

To tackle these critical barriers, the Department of Energy's (DOE) Office of Science funded leading semiconductor and systems vendors, including AMD, to develop new technologies to dramatically reduce power consumption and address other challenges impeding exascale deployments. The fact that Frontier, which as of the date of publication is the world's fastest supercomputer and first exascale system, requires just **~20 megawatts** to operate (a figure **~30x lower** than 2008 estimates) is a testament to the success of these investments and public-private partnerships.[3]

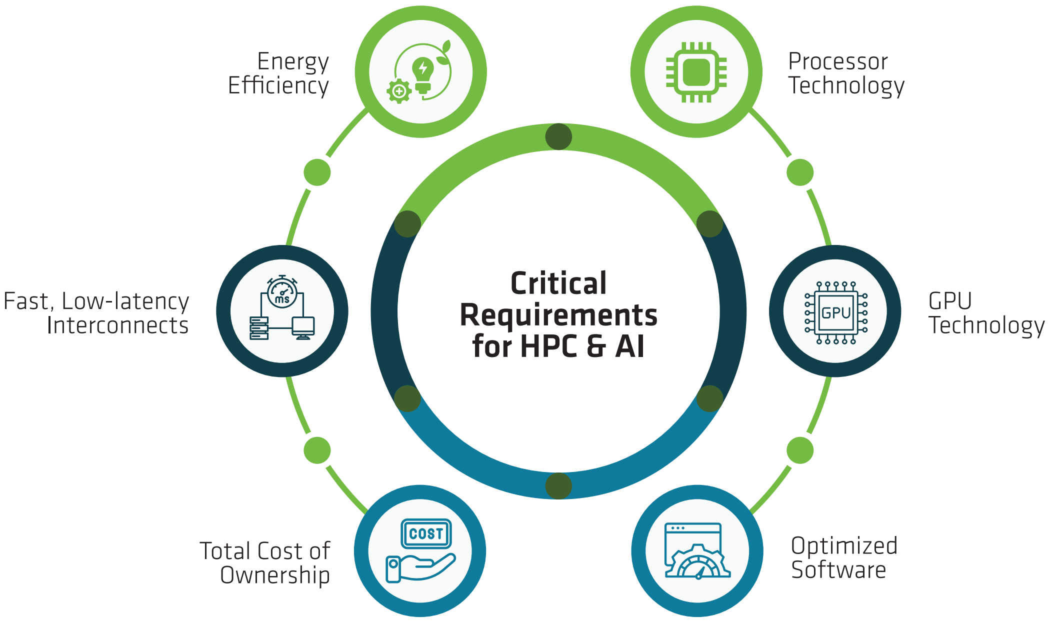


Figure 1 – Critical requirements for HPC & AI systems

With Moore's Law slowing and the end of Dennard scaling, building ever-faster supercomputers has become much more technically and economically challenging.[4] Processor technology remains critical to performance, but since the debut of the Titan supercomputer at Oak Ridge National Labs (ORNL) in 2012, performance leadership has depended on hybrid systems comprised of CPUs and GPUs.[5] Today, virtually all top supercomputers rely on hybrid CPU-GPU architectures, co-deployed using fast, low-latency interconnect fabrics to enable massive parallelism and facilitate the data ingress and egress critical to keeping GPUs fed.

Optimized software has also become critical to achieving performance goals. With the shift toward GPU accelerators, developers have needed to alter their programs and re-express algorithms to take advantage of the massive parallelism and large numbers of threads available on modern GPUs and APUs. Thanks to a decade of investment, software that can take advantage of the computational power of GPUs is now widely available.

"High-performance AI and HPC systems present an extraordinary opportunity to make a global impact. Innovation is pivotal for the long-term growth of the U.S., especially as we witness the convergence of the AI revolution with advancements in clean energy, synthetic biology, and robotics. Together, these technologies underpin a sustainable future."

Thomas Zacharia, AMD SVP Strategic Partnerships

EXASCALE LEADERSHIP DRIVES INNOVATION

Government plays an essential role in advancing research through public-private partnerships. Governments can support research in areas that are not quite ready for commercial deployment but have huge upside potential. They can absorb non-recurring engineering (NRE) costs, incenting otherwise risk-averse private companies to pursue research that would be too costly to undertake alone.

To this end, National Labs invest in energy, battery technologies, climate science, materials science, and 3D manufacturing. Many of the technical innovations we take for granted are the direct result of these public investments. As examples:

- Michael Grieves and John Vickers are credited with coining the term digital twin while working on a project at NASA.[6] The DOE also created digital twins of nuclear reactors to license them faster and improve safety and efficiency. Today, the concept of digital twins is widely used in industry.
- Horizontal drilling, which has driven oil exploration and increased extraction efficiencies globally, has its roots in simulations conducted at Sandia National Labs.[7]
- DOE programs such as FastForward, DesignForward, and PathForward, working with semiconductor and systems vendors, have delivered key advances critical to enabling the boom in generative AI and large language models (LLMs).[8]

AMD POWERS MANY OF THE LARGEST, MOST ENERGY-EFFICIENT SUPERCOMPUTERS

AMD has been a key partner working with National Labs in the research efforts described above. Today, AMD processors power 156 supercomputers on the latest Top500 list, a 29% increase year over year.[9] AMD technology-based systems include Frontier at ORNL, the first supercomputer to cross the exascale barrier. The increasing adoption of AMD technologies for the world's most demanding applications is a testament to the success of AMD investments in designing high-performance processors and accelerators.

Perhaps more significant is that **AMD CPUs power 60% of the world's top fifty most energy-efficient supercomputers**, according to the latest Green500 list.[10] Figure 2 plots energy efficiency vs. performance for the top 50 systems on the Green500 list segmented by CPU technology, illustrating that AMD technology-based systems are among the world's best-performing and most energy-efficient.[11]

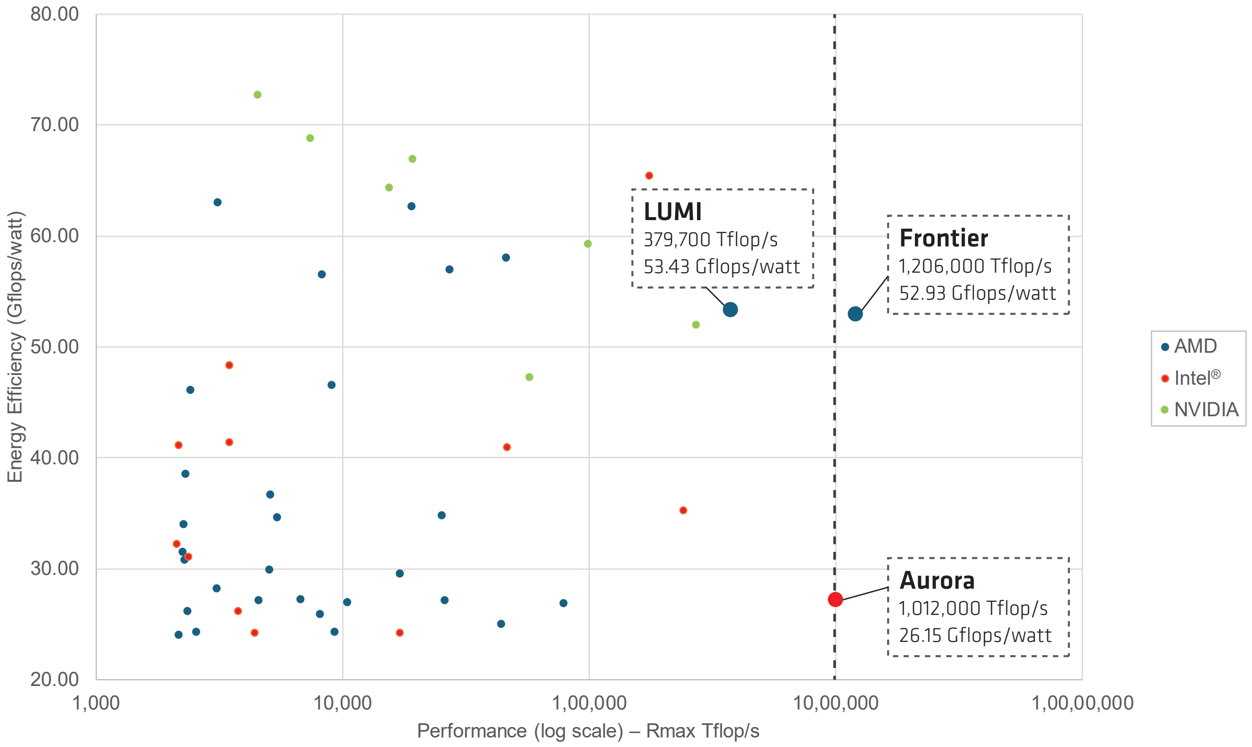


Figure 2 - AMD Powers 60% of the top 50 systems on the Green500 list

The advantage offered by AMD EPYC™ CPUs and AMD Instinct™ GPUs is illustrated by comparing the world's top two supercomputers on the June 2024 Top500 list, Frontier and Aurora, based on energy efficiency.

Frontier, powered by optimized 3rd Gen AMD EPYC processors and AMD Instinct MI250X CPUs, delivers 1,206,000 TFlop/s and consumes 22,786 kW of power for an energy efficiency of 52.93 GFlops/Watt. By comparison, the world's number 2 supercomputer, powered by competing processors and GPUs, delivers 1,012,000 TFlop/s but consumes 38,698.36 kW for an energy efficiency of only 26.15 GFlops/Watt.[12]

This comparison is shown in Figure 3. **Frontier, powered by AMD EPYC processors and AMD Instinct GPUs, delivers ~2x the energy efficiency of the next highest-performing supercomputer in the Top500 list**.[13] Public sector organizations can benefit from this leading energy-efficiency in smaller scale enterprise HPC and AI deployments to reduce costs and their carbon footprint.

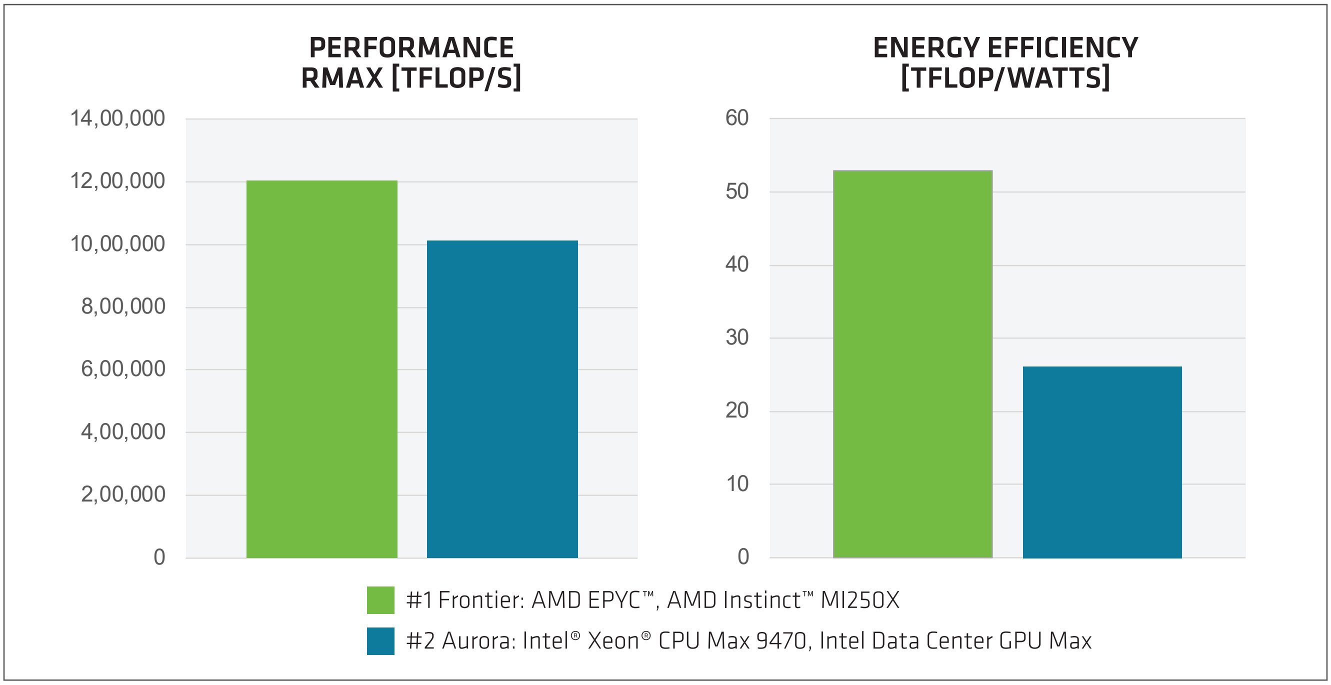


Figure 3: Comparing top exascale supercomputers based on performance and energy efficiency

EXASCALE INVESTMENTS DRIVE SUBSTANTIAL RETURNS

The impact of public investments in exascale systems is hard to overstate. The relatively modest USD 400M investments in PathForward and FastForward have arguably contributed to an industry with an estimated market capitalization of more than **USD 6 trillion**.[14]

For example, DOE's FastForward program helped fund the innovative "chiplet" architecture at the heart of modern AMD CPUs and accelerators [15], **enabling more power efficient technologies, faster interconnects and parallel processing techniques.** Exascale is a prime example of an open ecosystem turbocharging innovation both on the project itself and for broader industry.

In 2014, the Titan supercomputer delivered an energy efficiency of 2.1 Gflops/W.[16] A decade later, Frontier delivers 52.93 Gflops/W – **a 25X improvement in energy efficiency**.[17] Today, all organizations deploying server technologies benefit from these energy efficiency gains. As information technology has become critical to our daily lives, these improvements represent a giant step toward addressing the threat posed by climate change. Accelerating the transition to a sustainable low-carbon economy aims to produce benefits for economic growth, promote the health of people and our environment, and increase resilience to natural disasters.[18] Both OEMs and cloud hyper-scalers have benefited from public investments in HPC.

AMD SOLUTIONS FOR HPC AND AI

AMD EPYC™

First introduced in June 2017, AMD EPYC processors combine high core counts, large memory capacity, high bandwidth, and massive I/O to enable exceptional performance for a range of high-performance applications.

Realizing that increasing core density in monolithic designs would become more difficult with time, AMD engineers pioneered an innovative hybrid-die architecture where CPU cores and I/O functions were implemented on separate dies using different fabrication technologies.

This modular approach enables AMD to mix and match CPU and I/O dies to address a variety of application requirements.

In the latest 4th Gen EPYC processors, 'Zen 4' and 'Zen 4c' cores are produced with 5nm technology, and I/O dies are created with a 6nm process. Processor cores are combined with cache into a core complex (CCX), and these core complexes are fabricated onto a core complex die (CCD), also known as a "chiplet", as illustrated in Figure 4.[19]

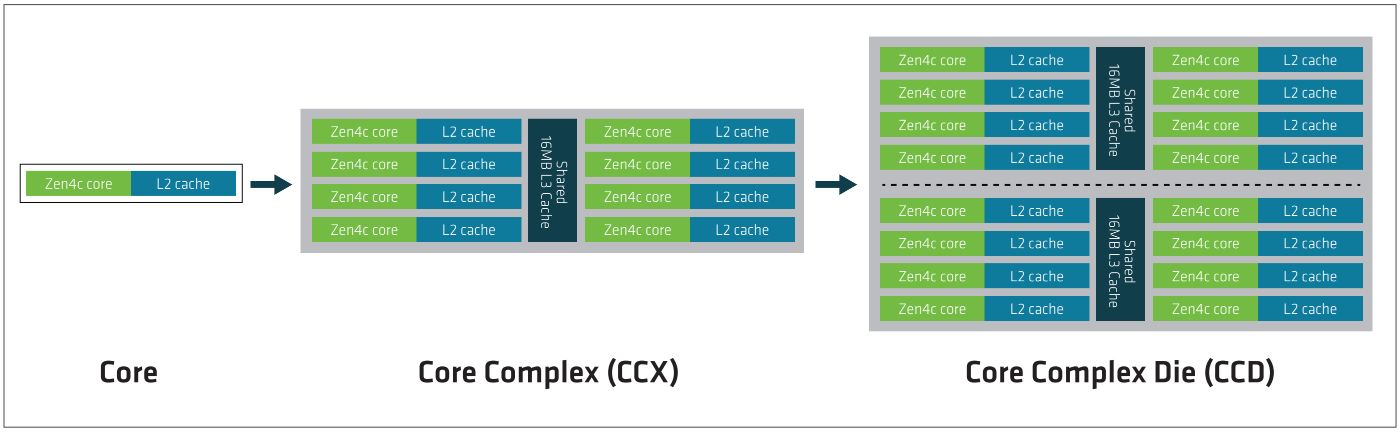


Figure 4: Relationship between core, core complex (CCD), and core complex die (CCD) or chiplet

With this modular approach, AMD can mix and match IP elements in CPUs, GPUs, and APUs to provide processors tailored to specific applications. It also helps AMD achieve economies of scale to minimize costs. Figure 5 shows the architecture of the top-off-stack AMD EPYC™ 9754 processor. In this configuration, 8 x 'Zen 4c' CPU dies, each with 16 'Zen 4c' cores and 32MB of shared L3 cache per core complex, are combined on a single AMD EPYC 9004 series processor providing 128 cores, 256 threads, and 256 MB of L3 cache for exceptional density and performance.

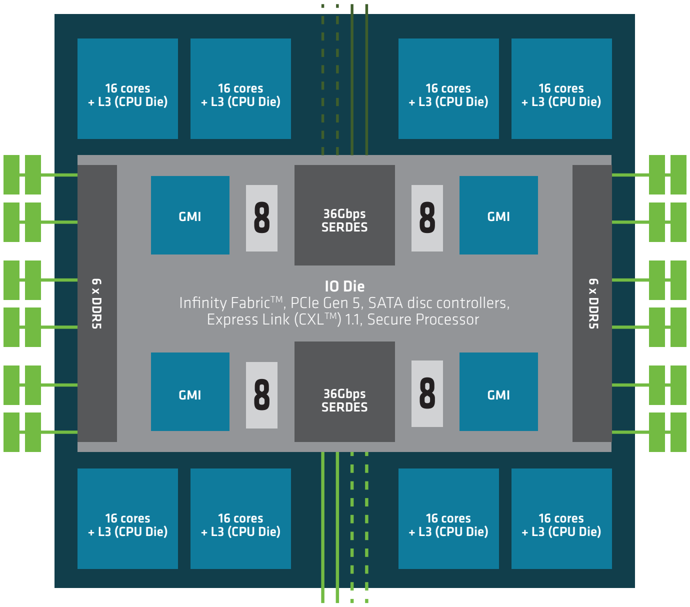

16 cores + L3 (CPU Die)
GMI
8
36Gbps SERDES
IO Die
Infinity Fabric™, PCIe Gen 5, SATA disc controllers, Express Link (CXL™) 1.1, Secure Processor
6 x DDR5

Figure 5 - AMD EPYC 9754 architecture diagram

4th Gen EPYC processors*

- Next generation 5nm technology
- Up to 128 cores/256 threads
- Up to 4.40 GHz Max boost clock speed[20]
- Up to 2 x 16MB = 32MB L3 cache per CPU die
- Up to 1152 MB L3 cache with 3D V-Cache[21]
- Infinity Fabric™ - Up to 72 Gb/s connections[22]
- 12 DDR5 memory channels per socket
- Up to 4800 MT/s DDR
- 128 PCIe® Gen 5 lanes per socket**
- CXL™ 1.1+ memory expansion
- Embedded AMD secure processor

* Specifications vary by processor SKU

** Up to 160 PCIe Gen 5 lanes in 2P configurations

For applications such as weather modeling, computational fluid dynamics (CFD), or intelligence applications that benefit from large cache, AMD offers 9004 series processors with 3D V-Cache™ technology, featuring up to 1152 MB of L3 cache per CPU.

AMD INSTINCT™ GPUS AND APUS

AMD Instinct™ MI300X accelerators are designed to deliver leadership performance for AI and HPC applications optimized to take advantage of GPUs. The AMD Instinct MI300X accelerator offers:

- 304 GPU compute units[23]
- 192 GB of HBM3 memory
- 5.3 TB/s peak theoretical memory bandwidth

Compared to competitive accelerators, this can translate into up to **2.4x the HPC performance (Peak TFLOPs)**.[24]

The AMD Instinct™ MI300A APU (accelerated processing unit) combines 24 x Zen4-based x86 CPU cores, 228 AMD CDNA™ 3 high-throughput GPU compute units, and 128 GB of unified HBM3 memory in a single package for enhanced flexibility, efficiency, and programmability.[25] The AMD Instinct MI300A is at the heart of El Capitan, which, when deployed at Lawrence Livermore National Laboratory (LLNL), is expected to be the world's fastest supercomputer.

AMD ROCM™ SOFTWARE

AMD ROCm™ is an open software stack that includes programming tools, compilers, libraries, and runtimes for software development on select AMD GPUs and APUs. The extensive set of ROCm libraries makes it easy for public sector organizations with advanced computing requirements to get to market quickly. At the heart of ROCm, is the Heterogeneous-computing Interface for Portability (HIP)—an API and runtime environment that enables developers to create applications that are portable across GPUs from multiple manufacturers.[26] HIP makes it straightforward for developers to adapt NVIDIA CUDA code to run on the AMD Instinct accelerators. Because ROCm is freely available and open source, developers have quickly embraced the platform, leading to an extensive ecosystem of AI models and HPC applications optimized for AMD hardware.[27]

A direct result of AMD leadership in Top500 systems is that many HPC and AI applications are now optimized for AMD EPYC processors and AMD Instinct accelerators. This means that public sector clients can pull freely available Docker containers from AMD Infinity Hub or ROCm Docker Hub and realize optimal performance running on-premises systems or on AMD technology-powered cloud instances.[28] AMD also publishes a catalog of hundreds of commercial and open-source applications built on ROCm across multiple disciplines to accelerate and simplify HPC and AI application deployments:[29]

- Astrophysics
- Climate & Weather
- Computational Chemistry
- Computational Fluid Dynamics
- Data Analytics
- Earth Science
- Electronic Structure
- Genomics
- Geophysics
- Machine Learning
- Material Science
- Molecular Dynamics
- Oil & Gas
- Physics
- Tools & Libraries
- Benchmarks
- ML Frameworks

THE AMD ADVANTAGE

SUPERIOR PERFORMANCE FOR HPC AND AI WORKLOADS

AMD EPYC processors and AMD Instinct accelerators excel across a wide range of workloads. AMD and its partners have published **over 300 world records** running industry-standard consortium benchmarks and ISV workloads.[30] Public sector customers can be assured that their investments in AMD processor-based systems will yield the exceptional performance and cost-efficiency for multiple applications.

Selected benchmarks, both published consortium scores and AMD test results, are highlighted below, illustrating the advantage of AMD processors over those offered by competitors for a range of general-purpose and enterprise HPC workloads.

DATA CENTER

60% faster general-purpose top-of-stack integer performance.
4th Gen AMD EPYC™ 9654 vs. 5th Gen Intel® Xeon® Platinum 8592+.[31]

1.6x VMmark® 3 general-purpose top-of-stack virtualization performance.
4th Gen AMD EPYC™ 9654 vs. 5th Gen Intel® Xeon® Platinum 8592+.[32]

2.25x better top-of-stack power efficiency with SPECpower_ssj® 2008.
4th Gen AMD EPYC™ 9754 vs. 5th Gen Intel® Xeon® Platinum 8592+.[33]

CLOUD

~70% average performance uplift and ~28% lower cloud OPEX.
Based on AMD internal testing on popular workloads – AWS M7a.4xlarge vs. AWS M7i.4xlarge instances.[34]

37% average performance uplift and 31% lower Cloud OpEx on Google Cloud.
Based on AMD Internal testing on select benchmarks – Google Cloud C3D-standard 16 vCPU vs N2-standard 16 vCPU instances.[35]

AI MODEL TRAINING & INFERENCE

~65% more AI test cases/minute on TPCx-AI SF30 based on AMD internal testing.
4th Gen AMD EPYC™ 9654 vs. 5th Gen Intel® Xeon® Platinum 8592+ 2P server comparison.[36]

~1.6x the performance on Bloom 176.
Based on AMD internal testing – 8x AMD Instinct™ MI300X vs. 8x NVIDIA® H100.[37]

DEFENSE, CLIMATE RESEARCH

Up to 47% better on SPEChpc® 2021 Tiny OMP Comparison.
2P AMD EPYC™ 9654 vs 2P Intel® Xeon® Platinum 8490H.[38]

~50% faster on Weather Research and Forecasting (WRF® 2.1).
Based on AMD internal testing – 2P AMD EPYC™ 9654 vs 2P Intel® Xeon® Platinum 8592+.[39]

19% better general-purpose top-of-stack performance on SPECrate® 2017 fp_base.
4th Gen AMD EPYC™ 9654 vs. 5th Gen Intel® Xeon® Platinum 8592+.[40]

AEROSPACE, MANUFACTURING

~50% better on ANSYS® LS-DYNA®.
Based on AMD internal testing – 2P AMD EPYC 9374F vs. 2P Intel Xeon 8562Y+.[41]

~48% better on ANSYS® CFX®.
Based on AMD internal testing – AMD EPYC 9374F vs. Intel Xeon 8562Y+.[42]

~25% better on ANSYS® FLUENT®.
Based on AMD internal testing – 2P AMD EPYC 9374F vs. 2P 32-core Intel Xeon Scalable 8562Y+.[43]

LIFE SCIENCES

~63% better general-purpose top-of-stack performance on GROMACS.
Based on AMD internal testing – 2P AMD EPYC 9654 vs. Intel Xeon Platinum 8592+.[44]

~62% better general-purpose top-of-stack performance on Quantum Chemistry with CP2K.
Based on AMD internal testing – 2P AMD EPYC 9654 vs. Intel Xeon Platinum 8592+.[45]

Figure 6 shows the relative performance of general-purpose and high-frequency 4th Gen EPYC processors versus comparable 5th Gen Intel® Xeon® Scalable Processors across common HPC workloads.[46] Comparative results are shown for Ansys® CFX®[57], Ansys® LS-DYNA®[56], GROMACS[59], CP2K quantum chemistry simulation[60], and Weather Research & Forecasting (WRF®)[54].

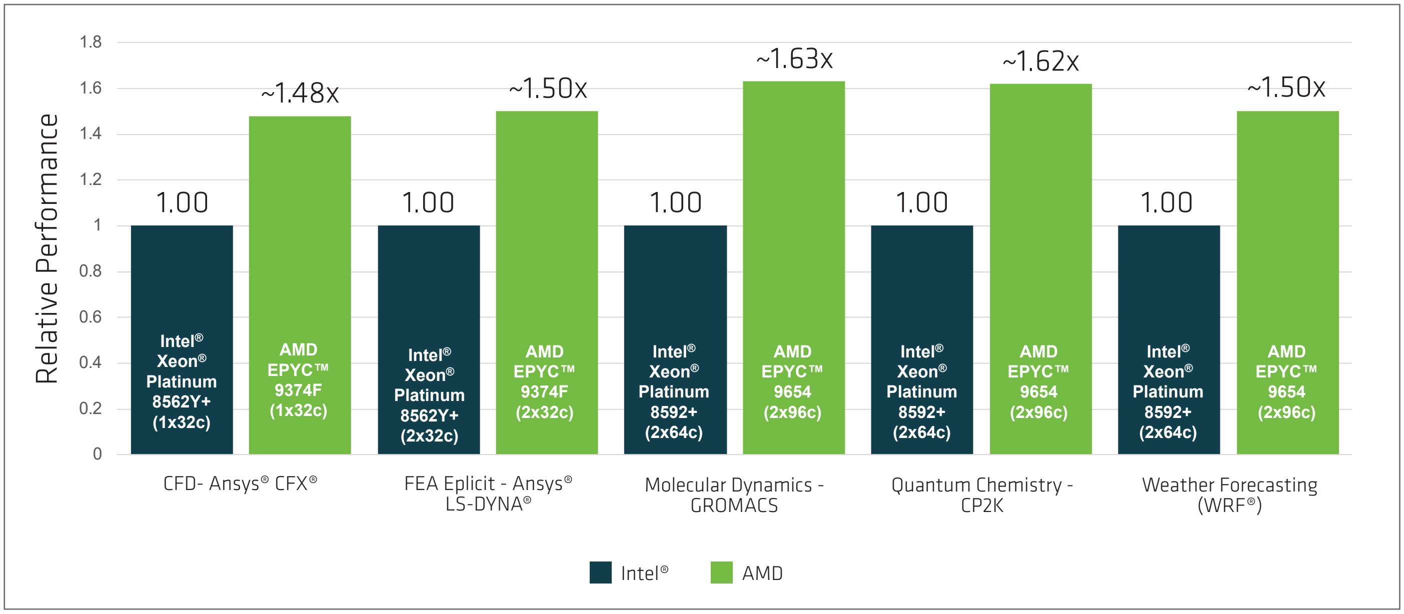


Figure 6 - Comparing 4th Gen AMD EPYC to 5th Gen Intel® Xeon® Scalable Processors across HPC workloads

ACCELERATING AI MODEL TRAINING AND INFERENCE

AMD EPYC processors can accelerate your entire AI journey, providing a host processor for GPU-accelerated machine learning and an efficient processor for AI inferencing. Servers powered by 4th Gen AMD EPYC processors and the latest AMD Instinct MI300 accelerators offer the bandwidth needed to train predictive models and arrive at the necessary parameter weights to make models function with the speed and accuracy you need. AMD Instinct accelerators provide exceptional AI performance compared to the competition. For instance, based on AMD internal testing and looking at commonly used AI precisions, the **AMD Instinct MI300X can provide up to 1.3X the performance of NVIDIA's H100 accelerators**.[47]

Once models are trained, inferences can often be made on the CPU without GPU acceleration. With a 2-socket server powered by 128-core EPYC 9754 processors running select OpenVINO™ workloads from the Phoronix Test Suite, public sector customers can obtain up to **72% higher inference FPS per watt** on average over 2P servers powered by 64-core 5th Gen Xeon 8592+ CPUs.[48]

A Zen Deep Neural Network (ZenDNN) plugin optimizes execution at the primitive level, accelerating models across diverse application types.[49] Whether you are using a computer vision model to recognize product defects, a natural language application to respond to customer prompts, or an engine to direct proactive maintenance, ZenDNN can help accelerate your inference workloads.

INDUSTRY-LEADING ENERGY-EFFICIENCY

Many federal and state governments worldwide have set greenhouse gas (GHG) reduction goals and require suppliers to disclose emissions and climate risks as part of their procurement strategies.[50] Using energy-efficient AMD EPYC processor not only helps achieve ongoing cost reductions from operations, but it can also help achieve Scope 2 GHG emission reduction targets related to the consumption of purchased energy for data center power and cooling. Similarly, for organizations operating in the cloud, using instances powered by AMD EPYC can help reduce direct cloud expenses and can reduce Scope 3 emissions resulting from the purchase of cloud services.

AMD EPYC processors power the industry's most energy-efficient x86 servers.[51] In the SPECPower® benchmark, widely used to measure energy efficiency, a dual-socket system powered by 128-core AMD EPYC 9754 processors delivered **~2.25x the energy efficiency** of a dual-socket system powered by Gen 5 Intel Xeon Platinum 8592+ processors.[48]

The innovations that make AMD processors and accelerators among the most energy-efficient platforms for supercomputing and HPC applications benefit a wide range of public sector workloads.

For organizations operating on-premises infrastructure, deploying servers based on high-performance, energy-efficient AMD EPYC CPUs can help achieve sustainability goals.

Organizations can deploy fewer servers to achieve the same application throughput, translating into less packaging and waste, lower manufacturing and shipping costs, lower power and cooling costs, savings in data center space, rack and network costs, and reduced management and administration overhead.

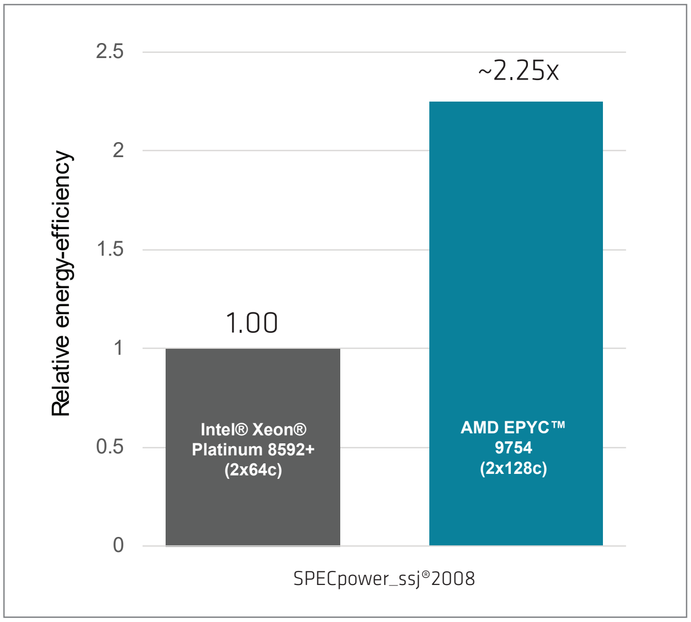


Figure 7: Relative SPECpower_ssj® power efficiency (dual-socket system)

PERFORMANCE AND ENERGY EFFICIENCY ENABLE LOWER TCO

High-performance, energy-efficient systems can help substantially reduce the total cost of ownership (TCO) for the infrastructure powering public sector HPC & AI applications.

Ansys® Fluent® is a leading computational fluid dynamics solver widely used in industry, research and academic settings. In a simulation involving the Ansys Fluent pump2 model, a 2P server populated with Intel® Xeon® Platinum 8462Y+ 32-core processors can deliver ~8,006 simulation jobs / server / day. A similar 2P server populated with AMD EPYC™ 9384X processors with 3D V-Cache™ technology delivers ~14,482 simulation jobs / server / day. **This means that 21 Intel Xeon Platinum 8462Y-based servers are required to deliver the same throughput of 165,000 jobs per day as can be achieved with just 12 AMD EPYC 9384X-based servers.**[52] These estimated savings are illustrated in Figure 8.

By substituting high-performance AMD EPYC 9384X-based servers to run this CFD workload of 165,000 jobs per day on Ansys Fluent pump2, organizations can obtain the same throughput with **~43% fewer servers, ~38% less power, and save up to ~44 U.S. tons of annual CO2e emissions annually**. AMD EPYC processors can also reduce the 3-year TCO by up to 39%.[78]

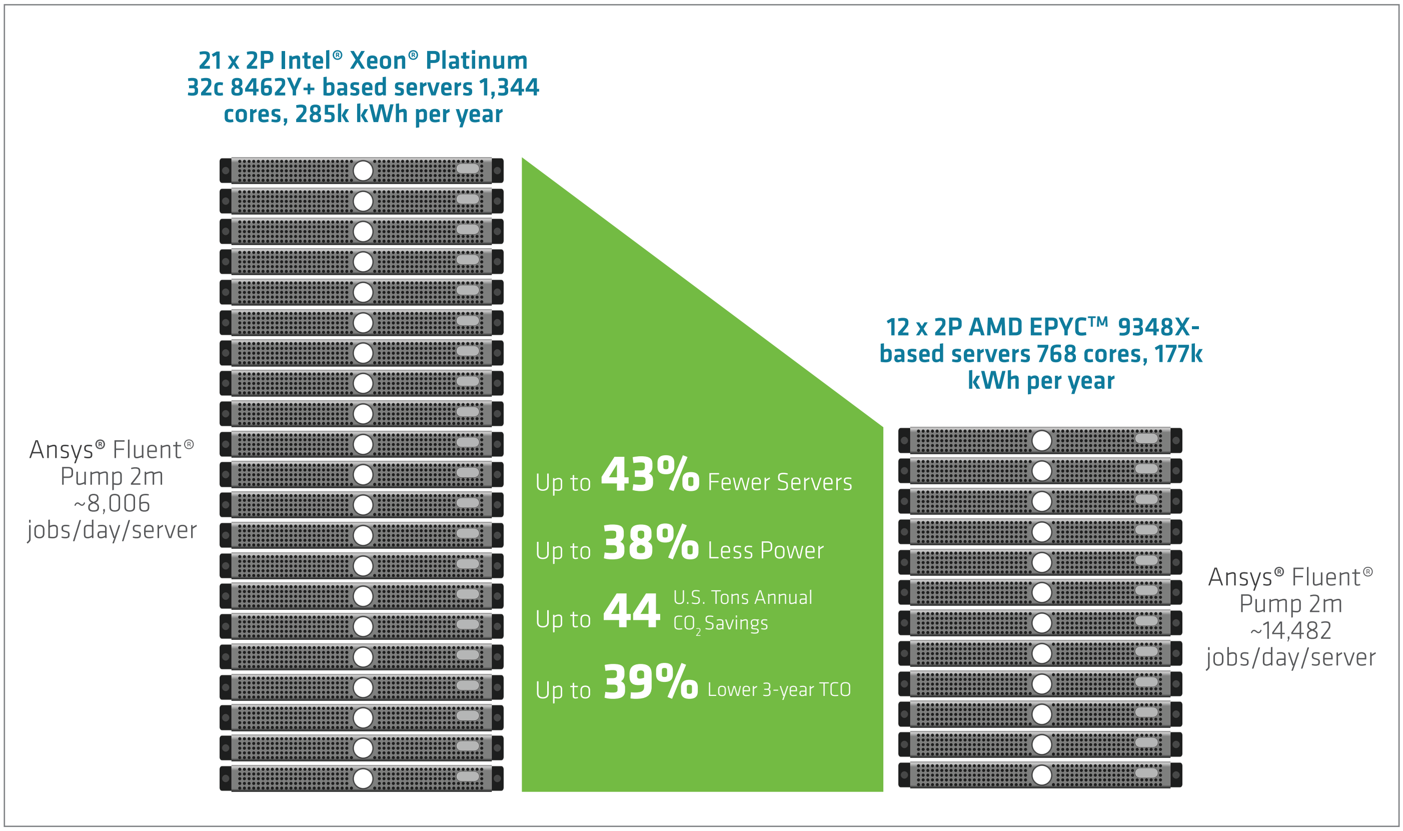


Figure 8 – Fewer servers use less power, leading to lower emissions and reduced TCO.

AN OPEN ECOSYSTEM BASED ON OPEN STANDARDS

A key design philosophy behind AMD products is a focus on open industry standards. Open standards help foster innovation and can allow for portability and interoperability. They also encourage a competitive marketplace, enabling customers to multi-source technologies based on open standards and help lower acquisition and maintenance costs. Adherence to open standards can also help improve product security by subjecting new technologies to peer review and increasing the chances that security vulnerabilities will be discovered quickly.

Open standards are not the same as open source, which refers to the availability of the source code for software. An open ecosystem does not require you to share your proprietary code, technology, or data but rather to adhere to common standards and protocols that enable seamless integration and communication among different systems. The greater the number of entities participating in an open ecosystem, the greater the interoperability and resulting benefits to innovation and security.

The recent COVID pandemic and the supply chain disruptions that ensued provide a strong warning about being dependent on a single vendor for key technologies. AMD strives to offer standards-based technologies wherever possible without compromising on performance.

OPEN INTERCONNECT TECHNOLOGIES

Today, large-scale HPC and AI clusters employ a variety of low-latency interconnects to link multiple GPU-enabled nodes and deliver performance at scale. Examples include InfiniBand™, Intel® Omni-Path, GigaIO FabreX™, Fujitsu Tofu, Cray Aries™, and cloud interconnects such as AWS Elastic Fabric Adapter (EFA).

Seven of the top ten systems on the Top500 list employ the HPE Slingshot Interconnect based on Gigabit Ethernet, providing strong evidence that standard Ethernet is scalable to the very largest supercomputers and is proven to deliver high packet rates for IP, messaging, and remote memory access.[53]

To promote a standards-based interconnect for HPC and AI at scale, **AMD is a steering member of the Ultra Ethernet Consortium (UEC), aimed at delivering an Ethernet-based open, interoperable interconnect standard** to meet the growing demands of modern workloads. UEC is backed by AI and HPC industry leaders, including Meta, Microsoft®, Intel, HPE, Dell®, Cisco®, and Broadcom®, providing a strong signal that UEC represents a great path forward.[54]

AMD is also an Ultra Accelerator Link (UALink) promoter group member, aiming to create an open, standard AI-focused interconnect based on a memory semantic fabric for GPU-to-GPU communication initially supporting up to 1,024 accelerators. Both the UEC and UALink efforts are aimed at promoting open standards and providing customers with freedom of choice for high-performance interconnects.

OPEN AND INTEROPERABLE SOFTWARE, LIBRARIES, AND TOOLS

AMD understands that customers don't want to become dependent on a particular vendor's software chain. In addition to providing open-source ROCm software for interoperability, AMD makes it easy for public sector organizations to leverage open software, libraries, and tools.

AMD provides compilers and libraries, such as AMD Optimizing C/C++ Compiler (AOCC) and AMD Optimizing CPU Libraries (AOCL), as part of AMD Zen Software Studio. AMD also feeds enhancements into gcc and other open-source compilers and libraries, recognizing that many prefer open-source tools.

Open and interoperable software is also critical for public-sector adoption of AI. Rather than requiring that customers build and train large models from scratch, AMD has invested in open models optimized for AMD EPYC processors and AMD Instinct accelerators alongside organizations such as Hugging Face, OpenXLA, Lamini, MosiacML, and others.

AMD and Hugging Face deliver high-performance transformers that work "out of the box" on AMD accelerators for training and inference models.[55] AMD is a founding member of the Google® OpenXLA project, which streamlines a developer's ability to optimize their models to target a wide variety of hardware, including AMD Instinct accelerators.[56] AMD also offers ROCm for AI, providing upstream support for leading AI frameworks, including Tensorflow and PyTorch, and dedicated libraries for machine learning, such as MIOpen and MIVisionX.[57]

These collaborations mean that public sector customers can more easily leverage the latest AI models to accelerate the development of AI systems in application areas such as data analysis, recommendation systems, knowledge management, and summarization applications.

THE X86 ARCHITECTURE

The x86-64 architecture, developed by AMD and adopted by Intel as a successor to the original x86 architecture, has enjoyed enormous market success. Customers have large sunk investments in software optimized for the x86 instruction set, particularly in HPC. Today, **x86 processors are a de-facto standard, powering over 94% of the world's top 500 supercomputers**.[58]

A key advantage of AMD EPYC is the consistency of its instruction set across generations. A consistent instruction set simplifies deployments and provides investment protection by assuring customers that software will run across the entire range of AMD EPYC processors. Customers do not need to worry about silicon-based acceleration features and instruction set extensions and whether they are implemented on particular processors.

AMD also supports open industry standards related to processors, including the Open Domain Specific Architecture (ODSA) project and the Universal Chiplet Interconnect Express (UCIe) standard.[59]

SECURITY

While HPC deployments in government labs are often "air-gapped," making security a lesser concern, security is important for all organizations. Government departments and public utilities hold sensitive information, deliver essential services, and manage critical infrastructure, placing them high on the target list of foreign adversaries and malicious actors. Public sector organizations are guarding against increasingly sophisticated threats, including malware, ransomware, phishing, and social engineering attacks.

With AMD Infinity Guard, security features are rooted in silicon, making them generally less susceptible to compromise than software-based solutions. AMD Infinity Guard serves as a technical foundation for confidential computing.[60]

At the heart of AMD EPYC processors is the AMD Secure Processor (ASP), responsible for acting as a "root of trust" and designed to maintain a secure environment. ASP functions include managing the boot process, initializing various security-related mechanisms, monitoring the system for suspicious activity or events, and implementing an appropriate response.

Figure 9 illustrates key security features of AMD EPYC processors. **Secure Encrypted Virtualization (SEV)** is designed to defend data while it is being processed. Keys are issued dynamically by the ASP for each VM, so that not even cloud providers have visibility to data—only the person or service that "owns" the data can access it. Also, because encryption occurs in hardware, it is transparent to applications, which may lessen performance impact.

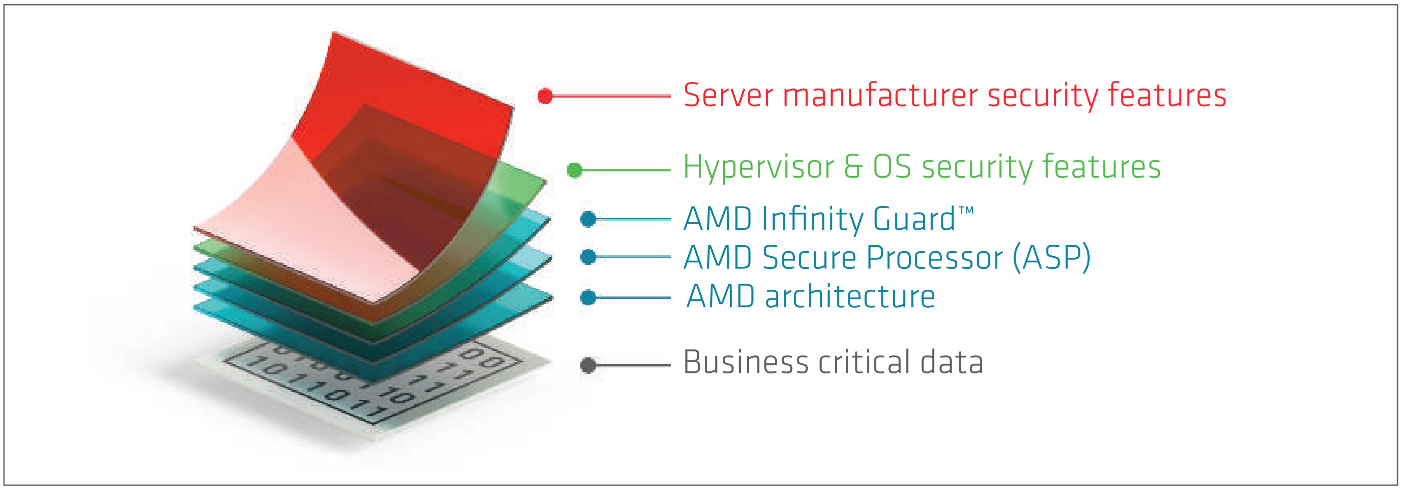

Server manufacturer security features
Hypervisor & OS security features
AMD Infinity Guard™
AMD Secure Processor (ASP)
AMD architecture
Business critical data

Figure 9 – AMD Infinity Guard is part of a layered approach to application security

Security features of AMD EPYC processors include:

- Secure Encrypted Virtualization (SEV)
- SEV Encrypted State (SEV-ES)
- SEV Secure Nest Paging (SEV-SNP)
- Transparent Secure Memory Encryption (TSME)
- AMD Secure Boot
- AMD Shadow Stack

Today, most major cloud providers, including AWS, Microsoft Azure, and Google Cloud Platform, offer confidential computing instances powered by EPYC processors and utilizing their AMD Infinity Guard features.[61] Using these confidential computing machine instances is as simple as ticking a box when deploying a cloud VM. Cloud companies and customers favor this model because it helps overcome data privacy and sovereign cloud obstacles.

Whether operating on-premises or in the cloud, AMD Infinity Guard can help public sector organizations comply with various mandates, including ITAR regulations, FedRAMP, Defense Federal Acquisition Regulation Supplement (DFARS), and FIPS requirements. European organizations have similar requirements. AMD Infinity Guard also helps simplify compliance with privacy requirements such as the EU's General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and industry-specific requirements such as the Health Insurance Portability and Accountability Act (HIPAA) by helping ensure that data is protected.

BEST PRACTICES FROM EXASCALE SYSTEMS SIMPLIFY PUBLIC SECTOR HPC DEPLOYMENTS

Public sector clients can benefit from the experience of AMD and AMD partners deploying exascale systems in several important ways for smaller scale HPC & AI deployments Clients can take advantage of:

- Optimized code for AMD CPUs and accelerators available in public Docker and Spack registries, reducing the complexity associated with deploying HPC and AI applications.[62]
- AMD offers tuning guides reflecting the company's experience deploying Top500 list supercomputing environments, which can help system administrators deploy applications more effectively.[63]
- Hardware providers offer BIOS optimizations and HPC-specific profiles in areas such as NUMA configurations, balancing devices across PCIe® busses, and power and determinism sliders, so customers needn't worry about these details.

CUSTOMER DEPLOYMENTS

High-performance systems based on AMD EPYC CPU and AMD Instinct accelerators are widely deployed in public sector organizations, including leading national and international labs. Notable examples include:

FRONTIER

Frontier is operated by Oak Ridge National Laboratory. Delivering 1.206 Exaflops, Frontier spans 74 HPE Cray Shasta cabinets and is comprised of 9,408 nodes, each powered by a single 64-core AMD optimized 3rd Gen EPYC CPU and 4 x AMD Instinct MI 250X GPUs[64]. Frontier will help researchers solve problems of critical importance, including enhancing nuclear reactor efficiency and safety, exploring the underlying genetics of disease, discovering patterns in patient data for precision medicine, and integrating artificial intelligence with data analytics, modeling, and simulation.

In addition to supporting DOE's core research, Frontier's computing facilities are made available to partners through the Accelerating Competitiveness through Computational Excellence (ACCEL) program. This program aims to help organizations enhance competitiveness by running approved simulations on Frontier. Public sector users can leverage many of the same components, interconnects, and optimized software used in Frontier in their own smaller-scale HPC & AI deployments supporting a wide range of workloads.

LUMI

LUMI (Large Unified Modern Infrastructure) is a pan-European supercomputer located in CSC's data center in Kajaani, Finland. Like Frontier, LUMI is based on 3rd Gen AMD EPYC processors (64 cores) and AMD Instinct MI 250X GPUs. Delivering over 531 peaks PF/sec, LUMI is ranked #5 globally and is one of the world's leading platforms for AI research.[65]

EL CAPITAN

When fully deployed in 2024, El Capitan is projected to be the world's most powerful supercomputer, delivering more than 2 Exaflops/second. El Capitan is a collaboration among the three National Nuclear Security Administration (NNSA) labs—Livermore, Los Alamos, and Sandia—to help ensure the US nuclear stockpile's safety, security, and reliability. El Capitan is based on AMD Instinct MI300A accelerated processing units (APUs), which combine a tightly coupled CPU and GPU in a single package.[66]

Other large deployments of AMD EPYC CPU-based systems in government and education include the National Oceanic and Atmospheric Administration (NOAA)[67], CERN's Large Hadron Collider[68], and Cornell[69] and Purdue[70] Universities.

SUMMARY

With substantial experience architecting the world's leading supercomputing environments, AMD is uniquely positioned to help public sector organizations with their own HPC and AI infrastructure deployments. With a complete line of CPUs, GPUs, APUs, FPGAs, interconnects, smart NICs, and software tools, AMD offers a complete HPC and AI infrastructure solution in collaboration with leading hardware manufacturers and cloud providers.

- AMD offers **leadership performance**, having achieved over **300 world records** along with partners running standard benchmarks for various HPC, AI, and commercial workloads.[28]
- AMD EPYC processors power the industry's **most energy-efficient x86 servers**[76], delivering up to **2.25x** the energy efficiency compared to comparable servers.[48]
- High-performance, energy-efficient processors AMD EPYC can **significantly reduce TCO** running HPC applications. In a CFD environment running Ansys® Fluent®, customers can achieve an estimated 43% reduction in server footprint, a 38% reduction in power consumption, and a 3-year TCO reduction of up to 39% compared to competing proces-sors.[78]
- AMD can help accelerate public-sector HPC and AI deployments with open-source ROCm software, an extensive catalog of optimized ISV workloads, freely available containerized applications, Spack recipes, and tuning and deployment guides.
- Finally, AMD is committed to transparency and open standards such as UEC and the emerging UALink standard, allowing for interoperability and investment protection while simplifying public sector procurement.

GETTING STARTED WITH AMD

AMD EPYC processor-based servers are available from most major computer manufacturers, including CISCO, Dell, HPE, Lenovo, and Supermicro. AMD EPYC processor-powered machine instances are also available in most clouds, including Amazon Web Services (AWS), Google Cloud Platform, Microsoft Azure, and Oracle Cloud Infrastructure (OCI).

To learn more about AMD solutions for the public sector, visit https://www.amd.com/en/solutions/public-sector.html.

To estimate how AMD EPYC processors may help reduce your TCO and total greenhouse gas (GHG) emissions, visit https://www.amd.com/en/resources/epyc-tools.html.

APPENDIX

1. Exascale computing study: technology challenges in achieving exascale systems, September 2008

2. See paragraph 6.3.4.4.3 Power Consumption Calculations in the Exascale computing study. Small modular reactors generally deliver up to 300 MW each.

3. Frontier requires 22,786 kW of power. See the June 2024 Top500 list.

4. Dennard scaling states that as transistors get smaller, their power use is proportional to the area of a semiconductor. Historically, the per-transistor power reduction afforded by Dennard scaling allowed semiconductors to raise clock frequencies as transistors became smaller without significantly increasing power consumption, but this relation broke down between 2005 and 2007 owing to challenges with current leakage.

5. Titan, a Cray XK7 system at Oak Ridge National Laboratory, debuted as #1 on the Top500 list in November of 2012. See https://top500.org/lists/top500/2012/11/highlights/

6. See Digital Twins and Living Models at NASA

7. Deployable Centralizers for Directional Drilling, Sandia National Labs Licensing and Technology Transfer.

8. See PathForward, administered by the Exascale Computing Project (ECP).

9. Press release: AMD Remains the Partner of Choice for World's Fastest and Most Efficient High Performance Computing Deployments – 05/13/2024

10. See the June 2024 Green500 list at https://top500.org/lists/green500/2024/06/.

11. Based on an analysis of the top 50 systems on the June 2024 Green500 list.

12. These figures are from the June 2024 Top500 list.

13. Aurora's energy efficiency is 52.93 GFlops/Watt. Frontier's energy efficiency is 26.15 GFlops/Watt. 52.93 / 26.15 = ~2x.

14. See Top 10 Cloud Companies in the World by Market Capitalization, March 2023.

15. See LLNL Advanced Simulation and Computing Fast Forward.

16. See Top500.org highlights – November 2014.

17. See Green500 List – June 2024. 52.93 GFlops/watt / 2.1 GFlops/watt = 25.2 GFlops/watt

18. Information Technology Industry Council (ITI), Environment & Sustainability, https://www.itic.org/policy environment-sustainability - accessed June 6, 2024.

19. See the 4th Gen AMD EPYC™ Processor Architecture Whitepaper.

20. Refers to the AMD EPYC 9174F part. Max boost speeds vary by processor. EPYC-018 - Max. boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems.

21. Applies to the AMD EPYC™ 9684X processor with 3D V-Cache technology. L3 cache varies by processor.

22. Applies only to AMD EPYC™ 9004 and 8004 series processors with four or fewer CPU dies.

23. See AMD Instinct MI300X Accelerators GPU specifications.

24. Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 163.4 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), 81.7 TFLOPs peak theoretical double precision (FP64), 163.4 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), 163.4 TFLOPs peak theoretical single precision (FP32), 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance. Published results on Nvidia H100 SXM (80GB) GPU resulted in 66.9 TFLOPs peak theoretical double precision tensor (FP64 Tensor), 33.5 TFLOPs peak theoretical double precision (FP64), 66.9 TFLOPs peak theoretical single precision (FP32), 494.7 TFLOPs peak TensorFloat-32 (TF32)*, 989.4 TFLOPS peak theoretical half precision tensor (FP16 Tensor), 133.8 TFLOPS peak theoretical half-precision (FP16), 989.4 TFLOPS peak theoretical Bfloat16 tensor format precision (BF16 Tensor), 133.8 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,978.9 TFLOPS peak theoretical 8-bit precision (FP8), 1,978.9 TOPs peak theoretical INT8 floating-point performance. Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ * Nvidia H100 GPUs don't support FP32 Tensor. MI300-18

25. See AMD Instinct™ MI300A Accelerators specifications.

26. For details on the Heterogeneous-computing Interface for Portability (HIP) see the AMD ROCm™ Software documentation

27. See https://github.com/ROCm/ROCm.

28. See the AMD Infinity Hub container collection. Also, see the AMD ROC™ Platform container collection on Docker Hub.

29. See GPU-accelerated applications with AMD Instinct™ accelerators enabled by ROCm™ software

30. See AMD EPYC™ Processor World Records

31. SPECrate® 2017_int_base results @ top-of-stack as of 08/20/2024:

- 2P EPYC 9654, Score 1810, https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240129-40896.html
- 2P Xeon 8592+, Score 1130, https://www.spec.org/cpu2017/results/res2023q4/cpu2017-20231127-40064.html

SPEC® and SPECrate® are registered trademarks of Standard Performance Evaluation Corporation. See www.spec.org for more information.

32. VMmark 3 performance @ general purpose top-of-stack as of 8/20/2024:

- 2P EPYC 9654, 40.66 @ 42 tiles, https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vmmark/2023-06-13-Lenovo-ThinkSystem-SR665V3.pdf
- 2P Xeon 8592+, https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vmmark/2024-02-20-Dell-PowerEdge-R760.pdf.

33. SP5-011F: SPECpower_ssj® 2008 comparison based on published 2P server results as of 1/12/2024. Configurations: 2P 128-core AMD EPYC 9754 (36,210 overall ssj_ops/W, 2U, https://spec.org/power_ssj2008/results/res2024q1/power_ssj2008-20231205-01347.html) is 2.25x the performance of best published 2P 64-core Intel Xeon® Platinum 8592+ (16,106 overall ssj_ops/W, 2U, https://spec.org/power_ssj2008/results/res2024q1/power_ssj2008-20231205-01349.html). SPEC®, SPECpower®, and SPECpower_ssj® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

34. SP5C-004 - AWS M7a.4xlarge max score and Cloud OpEx savings comparison to M7i.4xlarge running six common application workloads using on-demand pricing US-East (Ohio) Linux as of 10/9/2023.

- FFmpeg: ~1.9x the raw to vp9 encoding performance (52.3% of M7i runtime) saving ~40% in Cloud OpEx
- NGINX™: ~1.6x the WRK performance (61.7% of M7i runtime) saving ~29% in Cloud OpEx
- Server-side Java® multi-instance max Java OPS: ~1.4x the ops/sec performance (71.4% of M7i runtime) saving ~18% in Cloud OpEx
- MySQL™: ~1.4x the TPROC-C performance (70.4% of M7i runtime) saving ~18% in Cloud OpEx
- SQL Server®: ~1.3x the TPROC-H performance (76.0% of M7i runtime) saving ~13% in Cloud OpEx
- Redis™: ~2.2x the rps performance (44.6% of M7i runtime) saving ~49% in Cloud OpEx

Cloud performance results presented are based on the test date in the configuration. Results may vary due to changes to the underlying configuration, and other conditions such as the placement of the VM and its resources, optimizations by the cloud service provider, accessed cloud regions, co-tenants, and the types of other workloads exercised at the same time on the system.

35. SP5C-006 - MySQL™, Redis®, NGINX®, server-side Java multi-instances, and FFmpeg™ comparison of Google Cloud C3D-standard 16 vCPU to N2-standard 16 vCPU based on AMD testing on 11/02/23. OpEx savings calculated based on on-demand pricing at https://cloud.google.com/compute/vm-instance-pricing for us-central1 (Iowa) as of 11/01/2023. Configu-rations both with 64GB running Ubuntu 22.04.3 LTS. Comparisons:

- MySQL 8.0.28 HammerDB 4.2 TPROC-C (~1.2x tpm, 22% Cloud OpEx savings),
- Redis 7.2 get/set: (~1.4x rps, 32% Cloud OpEx savings),
- NGINX 1.1.9-2 WRK 4.2: (~1.2x ops/sec, 23% Cloud OpEx savings),
- server-side Java® multi instances max-OPS (~1.7x OPS, 45% Cloud OpEx savings) and
- FFmpeg 4.4.2.0 Ubuntu 22.04.1 h264-vp9, raw_h264, raw_vp9, vp9_h264 at 1080p (~1.4x frames/hr, 32% Cloud OpEx savings).

Cloud performance results presented are based on the test date in the configuration. Results may vary due to changes to the underlying configuration, and other conditions such as the placement of the VM and its resources, optimizations by the cloud service provider, accessed cloud regions, co-tenants, and the types of other workloads exercised at the same time on the system.

36. SP5-051A: TPCx-AI SF30 derivative workload comparison based on AMD internal testing running multiple VM instances as of 4/13/2024. The aggregate end-to-end AI throughput test is derived from the TPCx-AI benchmark and as such is not comparable to published TPCx-AI results, as the end-to-end AI throughput test results do not comply with the TPCx-AI Specification. AMD system configuration: Processors: 2 x AMD EPYC 9654; Frequencies: 2.4 GHz | 3.7 GHz; Cores: 96 cores per socket (1 NUMA domain per socket); L3 Cache: 384MB/socket (768MB total); Memory: 1.5TB (24) Dual-Rank DDR5-5600 64GB DIMMs, 1DPC (Platform supports up to 4800MHz); NIC: 2 x 100 GbE Mellanox CX-5 (MT28800); Storage: 3.2 TB Samsung MO003200KYDNC U.3 NVMe; BIOS: 1.56; BIOS Settings: SMT=ON, Determinism=Power, NPS=1, PPL=400W, Turbo Boost=Enabled; OS: Ubuntu® 22.04.3 LTS; Test config: 6 instances, 64 vCPUs/instance, 2663 aggregate AI use cases/min vs. Intel system configuration: Processors: 2 x Intel® Xeon® Platinum 8592+; Frequencies: 1.9 GHz | 3.9 GHz; Cores: 64 cores per socket (1 NUMA domain per socket); L3 Cache: 320MB/socket (640MB total); Memory: 1TB (16) Dual-Rank DDR5-5600 64GB DIMMs, 1DPC; NIC: 4 x 1GbE Broadcom NetXtreme BCM5719 Gigabit Ethernet PCIe; Storage: 3.84TB KIOXIA KCMYXRUG3T84 NVMe; BIOS: ESE124B-3.11; BIOS Settings: Hyperthreading=Enabled, Turbo boost=Enabled, SNC=Disabled; OS: Ubuntu® 22.04.3 LTS; Test config: 4 instances, 64 vCPUs/instance, 1607 aggregate AI use cases/min. Results may vary due to factors including system configurations, software versions and BIOS settings. TPC, TPC Benchmark and TPC-C are trademarks of the Transaction Processing Performance Council.

37. MI300-34: Token generation throughput using DeepSpeed Inference with the Bloom-176b model with an input sequence length of 1948 tokens, and output sequence length of 100 tokens, and a batch size tuned to yield the highest throughput on each system comparison based on AMD internal testing using custom docker container for each system as of 11/17/2023.
Configurations:
2P Intel Xeon Platinum 8480C CPU powered server with 8x AMD Instinct™ MI300X 192GB 750W GPUs, pre-release build of ROCm™ 6.0, Ubuntu 22.04.2. vs. An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 80GB 700W GPUs, CUDA 12.0, Ubuntu 22.04.3.
8 GPUs on each system were used in this test.
Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

38. SP5-060A: SPEChpc® 2021 Tiny OMP comparison based on published results as of 2/15/2023. Configurations: 2P AMD EPYC 9654 (192 cores, 24 base ranks, OMP parallel mode, https://spec.org/hpc2021/results/res2022q4/hpc2021-20221016-00135.html) scores 13.9 SPEChpc®2021_tny_base versus 2P Intel Xeon Platinum 8490H (120 cores, 30 base ranks, OMP parallel mode, https://spec.org/hpc2021/results/res2023q1/hpc2021-20230108-00156.html) scores 9.45 SPEChpc®2021_tny_base for 1.47x the performance. SPEChpc® is a registered trademark of Standard Performance Evaluation Corporation (SPEC). Learn more at www.spec.org.

39. SSP5-241: AMD testing as of 04/23/2024. The detailed results show the average uplift of the performance metric (Mean Time/Step) of this benchmark for the AMD EPYC™ 9654 96-Core Processor, and the INTEL® XEON® PLATINUM 8592+ running the following test on Open-Source WRF® 4.2.1: * conus2.5km: ~1.50x.
AMD System Configuration: Server: AMD Titanite; Processors: 2 x 96-Core AMD EPYC™ 9654 ; Memory: 24x 64GB DDR5-4800; Storage: SAMSUNG MZQL21T9HCJR-00A07; BIOS: RTI1009C; BIOS Settings from Default; SMT=Off, NPS=4, Determinism=Power; OS: RHEL 9.3; Kernel: Linux 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: amd_iommu=on, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance, Disable C2 States.
Intel System Configuration: Server: Lenovo Thinksystem SR650 V3; Processors: 2 x 64-Core INTEL® XEON® PLATINUM 8592+; Memory: 16x 64GB DDR5-5600; Storage: KIOXIA KCMYXRUG3T84; BIOS: ESE122V-3.10; BIOS Settings from Default: Hyperthreading=Off, Profile=Maximum Performance Profile; OS: RHEL 9.3; Kernel: 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: processor.max_cstate=1, Intel_idle.max_cstate=0, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance. Results may vary based on system configurations, software versions, and BIOS settings.

40. SPECrate® 2017_fp_base results @ top-of-stack as of 08/20/2024:

- 2P EPYC 9654, Score 1480, https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240111-40517.html
- 2P Xeon 8592+, Score 1240, https://www.spec.org/cpu2017/results/res2023q4/cpu2017-20231127-40063.html

SPEC® and SPECrate® are registered trademarks of Standard Performance Evaluation Corporation. See www.spec.org for more information.

41. SP5-246: AMD testing as of 04/23/2024. The detailed results show the average uplift of the performance metric (Elapsed Time) of this benchmark for the AMD EPYC™ 9374F 32-Core Processor, and the INTEL® XEON® PLATINUM 8562Y+ running select tests on Ansys® LS-DYNA® R13_1_1: * ls-3cars: ~1.61x, * ls-car2car: ~1.48x, * ls-neon: ~1.67x, * ls-odb10m-short: ~1.25x. AMD System Configuration: Server: AMD Titanite; Processors: 2 x 32-Core AMD EPYC™ 9374F ; Memory: 24x 64GB DDR5-4800; Storage: SAMSUNG MZQL21T9HCJR-00A07; BIOS: RTI1009C; BIOS Settings from Default; SMT=Off, NPS=4, Determinism=Power; OS: RHEL 9.3; Kernel: Linux 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: amd_iommu=on, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance, Disable C2 States. Intel System Configuration: Server: Lenovo Thinksystem SR650 V3; Processors: 2 x 32-Core INTEL® XEON® PLATINUM 8562Y+; Memory: 16x 64GB DDR5-5600; Storage: KIOXIA KCMYXRUG3T84; BIOS: ESE122V-3.10; BIOS Settings from Default: Hyperthreading=Off, Profile=Maximum Performance Profile; OS: RHEL 9.3; Kernel: 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: processor.max_cstate=1, Intel_idle.max_cstate=0, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance. Results may vary based on system configurations, software versions, and BIOS settings.

42. SP5-245: AMD testing as of 04/23/2024. The detailed results show the average uplift of the performance metric (Elapsed Time) of this benchmark for the AMD EPYC™ 9374F 32-Core Processor, and the INTEL® XEON® PLATINUM 8562Y+ running select tests on Ansys® CFX® V231. Uplifts for the performance metric normalized to the INTEL® XEON® PLATINUM 8562Y+ follow for each benchmark: * Airfoil 100: ~1.54x, * Airfoil 10: ~1.53x, * Airfoil 50: ~1.54x, * LeMans Car: ~1.40x, * Automotive Pump: ~1.37x. AMD System Configuration: Server: AMD Titanite; Processors: 2 x 32-Core AMD EPYC™ 9374F ; Memory: 24x 64GB DDR5-4800; Storage: SAMSUNG MZQL21T9HCJR-00A07; BIOS: RTI1009C; BIOS Settings from Default; SMT=Off, NPS=4, Determinism=Power; OS: RHEL 9.3; Kernel: Linux 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: amd_iommu=on, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance, Disable C2 States.

Intel System Configuration: Server: Lenovo Thinksystem SR650 V3; Processors: 2 x 32-Core INTEL® XEON® PLATINUM 8562Y+; Memory: 16x 64GB DDR5-5600; Storage: KIOXIA KCMYXRUG3T84; BIOS: ESE122V-3.10; BIOS Settings from Default: Hyperthreading=Off, Profile=Maximum Performance Profile; OS: RHEL 9.3; Kernel: 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: processor.max_cstate=1, Intel_idle.max_c-state=0, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance. Results may vary based on system configurations, software versions, and BIOS settings.

43. SP5-244: AMD testing as of 04/23/2024. The detailed results show the average uplift of the performance metric (Core Solver Rating) of this benchmark for the AMD EPYC™ 9374F 32-Core Processor, and the INTEL® XEON® PLATINUM 8562Y+ running select tests on Ansys® Fluent®: * aircraft_14m: ~1.33x, * aircraft_2m: ~1.26x, * combustor_12m: ~1.22x, * combustor_71m: ~1.27x, * exhaust_system_33m: ~1.22x, * f1_racecar-140m: ~1.29x, * fluidized_bed_2m: ~1.10x, * Fluent®-ice2: ~1.19x, * landing_gear_15m: ~1.23x, * LeMans_6000_16m: ~1.21x, * oil_rig_7m: ~1.03x, * f1_racecar_280m: ~1.24x, * pump_2m: ~1.43x, * rotor_3m: ~1.34x, * sedan_4m: ~1.39x. AMD System Configuration: Server: AMD Titanite; Processors: 2 x 32-Core AMD EPYC™ 9374F ; Memory: 24x 64GB DDR5-4800; Storage: SAMSUNG MZQL21T9HCJR-00A07; BIOS: RTI1009C; BIOS Settings from Default; SMT=Off, NPS=4, Determinism=Power; OS: RHEL 9.3; Kernel: Linux 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: amd_iommu=on, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance, Disable C2 States. Intel System Configuration: Server: Lenovo Thinksystem SR650 V3; Processors: 2 x 32-Core INTEL® XEON® PLATINUM 8562Y+; Memory: 16x 64GB DDR5-5600; Storage: KIOXIA KCMYXRUG3T84; BIOS: ESE122V-3.10; BIOS Settings from Default: Hyperthreading=Off, Profile=Maximum Performance Profile; OS: RHEL 9.3; Kernel: 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: processor.max_cstate=1, Intel_idle.max_cstate=0, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance. Results may vary based on system configurations, software versions, and BIOS settings.

44. SP5-243: AMD testing as of 04/23/2024. The detailed results show the average uplift of the performance metric (ns/day) of this benchmark for the AMD EPYC™ 9654 96-Core Processor, and the INTEL® XEON® PLATINUM 8592+ running select tests on Open-Source GROMACS. * benchPEP: ~1.70xm, * gmx_water1536K_PME: ~1.56x. AMD System Configuration: Server: AMD Titanite; Processors: 2 x 96-Core AMD EPYC™ 9654 ; Memory: 24x 64GB DDR5-4800; Storage: SAMSUNG MZQL21T9HCJR-00A07; BIOS: RTI1009C; BIOS Settings from Default; SMT=Off, NPS=4, Determinism=Power; OS: RHEL 9.3; Kernel: Linux 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: amd_iommu=on, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance, Disable C2 States. Intel System Configuration: Server: Lenovo Thinksystem SR650 V3; Processors: 2 x 64-Core INTEL® XEON® PLATINUM 8592+; Memory: 16x 64GB DDR5-5600; Storage: KIOXIA KCMYXRUG3T84; BIOS: ESE122V-3.10; BIOS Settings from Default: Hyperthreading=Off, Profile=Maximum Performance Profile; OS: RHEL 9.3; Kernel: 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: processor.max_cstate=1, Intel_idle.max_c-state=0, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance. Results may vary based on system configurations, software versions, and BIOS settings. AMD testing as of 04/23/2024.

45. SP5-239: AMD testing as of 04/23/2024. The detailed results show the average uplift of the performance metric (Elapsed Time) of this benchmark for the AMD EPYC™ 9654 96-Core Processor, and the INTEL® XEON® PLATINUM 8592+ running the following test on Open-Source cp2k. Uplifts for the performance metric normalized to the INTEL® XEON® PLATINUM 8592+ follow for each benchmark: * H2O-dft-ls: ~1.62x. AMD System Configuration: Server: AMD Titanite; Processors: 2 x 96-Core AMD EPYC™ 9654 ; Memory: 24x 64GB DDR5-4800; Storage: SAMSUNG MZQL21T9HCJR-00A07; BIOS: RTI1009C; BIOS Settings from Default; SMT=Off, NPS=4, Determinism=Power; OS: RHEL 9.3; Kernel: Linux 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: amd_iommu=on, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance, Disable C2 States. Intel System Configuration: Server: Lenovo Thinksystem SR650 V3; Processors: 2 x 64-Core INTEL® XEON® PLATINUM 8592+; Memory: 16x 64GB DDR5-5600; Storage: KIOXIA KCMYXRUG3T84; BIOS: ESE122V-3.10; BIOS Settings from Default: Hyperthreading=Off, Profile=Maximum Performance Profile; OS: RHEL 9.3; Kernel: 5.14.0-362.8.1.el9_3.x86_64; Kernel CMDLINE: processor.max_cstate=1, Intel_idle.max_cstate=0, iommu=pt, mitigations=off; Runtime Tunings: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance. Results may vary based on system configurations, software versions, and BIOS settings. AMD testing as of 04/23/2024.

46. See 4th Gen AMD EPYC™ Processors Outshine the Latest 5th Gen Intel® Xeon® Processors.

47. MI300-17: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8),

2614.9 TOPs INT8 floating-point performance. The MI300X is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 1,307.4 TFLOPS peak theoretical TensorFloat-32 (TF32), 2,614.9 TFLOPS peak theoretical half precision (FP16), 2,614.9 TFLOPS peak theoretical Bfloat16 format precision (BF16), 5,229.8 TFLOPS peak theoretical 8-bit precision (FP8), 5,229.8 TOPs INT8 floating-point performance with sparsity. Published results on Nvidia H100 SXM (80GB) 700W GPU resulted in 989.4 TFLOPs peak TensorFloat-32 (TF32) with sparsity, 1,978.9 TFLOPS peak theoretical half precision (FP16) with sparsity, 1,978.9 TFLOPS peak theoretical Bfloat16 format precision (BF16) with sparsity, 3,957.8 TFLOPS peak theoretical 8-bit precision (FP8) with sparsity, 3,957.8 TOPs peak theoretical INT8 with sparsity floating-point performance. Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core.

[48] SP5-252: Third-party testing OpenVINO 2023.2.dev FPS comparison based on Phoronix review https://www.phoronix-.com/review/intel-xeon-platinum-8592/9 as of 12/14/2023 of select OpenVINO tests: Vehicle Detection FP16, Person Detection FP16, Person Vehicle Bike Detection FP16, Road Segmentation ADAS FP16 and Face Detection Retail FP16. Testing not independently verified by AMD. Scores will vary based on system configuration and determinism mode used (Power Determinism used). OpenVINO is a trademark of Intel Corporation or its subsidiaries.

[49] See AMD Zen Deep Neural Network (ZenDNN)

[50] See Federal Supplier Climate Risks and Resilience Proposed Rule, 11/10/2022.

[51] EPYC-028D: SPECpower_ssj® 2008, SPECrate®2017_int_energy_base, and SPECrate®2017_fp_energy_base based on results published on SPEC's website as of 2/21/24. VMmark® server power-performance / server and storage power-performance (PPKW) based results published at https://www.vmware.com/products/vmmark/results3x.1.html?sort=score. The first 105 ranked SPECpower_ssj®2008 publications with the highest overall efficiency overall ssj_ops/W results were all powered by AMD EPYC processors. For SPECrate®2017 Integer (Energy Base), AMD EPYC CPUs power the first 8 top SPECrate®2017_int_energy_base performance/system W scores. For SPECrate®2017 Floating Point (Energy Base), AMD EPYC CPUs power the first 12 SPECrate®2017_fp_energy_base performance/system W scores. For VMmark® server power-performance (PPKW), have the top 5 results for 2- and 4-socket matched pair results outperforming all other socket results and for VMmark® server and storage power-performance (PPKW), have the top overall score. See https://www.amd.com/en/claims/epyc4#faq-EPYC-028D for the full list. For additional information on AMD sustainability goals see: https://www.amd.com/en/corporate/corporate-responsibility/data-center-sustainability.html. More information about SPEC® is available at http://www.spec.org. SPEC, SPECrate, and SPECpower are registered trademarks of the Standard Performance Evaluation Corporation. VMmark is a registered trademark of VMware in the US or other countries..

[52] SP5TCO-045: As of May 2023, based on AMD Internal analysis and using the AMD EPYC™ Bare Metal Server and Greenhouse Gas Emissions TCO Estimation Tool v9.33 PRO estimating the cost and quantity of 2P AMD 32 core EPYC™ 9384X powered server versus 2P Intel® Xeon® 32 core Platinum 8462Y+ based server solutions required to deliver 165,000 jobs / day with Ansys Fluent-pump2.
Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September

[53] Based on the June 2024 Top500 list. See https://top500.org/lists/top500/2024/06/.

[54] See the Ultra Ethernet Consortium website

[55] See Hugging Face and AMD partner on accelerating state-of-the-art models for CPU and GPU platforms.

[56] See OpenXLA is available now to accelerate and simplify machine learning.

[57] See AMD ROCm™ Software for AI.

[58] See the June 2024 Top 500 list. 472 of the 500 systems run AMD or Intel® x86 processors. 472 / 500 – 94.4%.

[59] See Open Domain-Specific Architecture and Universal Chiplet Interconnect Express™.

[60] AMD Infinity Guard features vary by EPYC™ Processor generations and/or series. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard.
GD-183A.

[61] Specific hardware protections vary by cloud provider and the underlying AMD EPYC processor used.

[62] Spack is a package management tool developed at Lawrence Livermore National Labs (LLNL) to deploy applications optimized for large supercomputing centers. See AMD optimized Spack recipe for HPC workloads.

[63] See AMD Documentation Hub.

[64] For details, see https://www.olcf.ornl.gov/frontier/.

[65] LUMI Supercomputer powered by AMD and HPE.

[66] El Capitan: Preparing for NNSA's first exascale machine.

[67] See NOAA completes upgrade to weather and climate supercomputer system.

[68] See AMD EPYC™ CPUs Enable Rapid Quark Detection at CERN.

[69] See Cornell Uses AMD Technology to Help Understand the Universe.

[70] See Purdue University Breaks Research Computing Barriers.