Intersect360 Research White Paper:

Precision at Scale:

Commitment to High-Performance Computing in the Age of AI

Executive Summary

High-performance computing is in the midst of a profound transformation. The line between traditional HPC and artificial intelligence is blurring, as research institutions and enterprises alike demand infrastructure that can deliver both uncompromised numerical precision and the throughput required for modern AI. While the industry's attention has shifted toward accelerating AI, the essential role of double-precision computation in scientific discovery remains unchanged. The challenge for technology providers is clear: support both, without forcing organizations to compromise.

AMD has taken this challenge head-on. By maintaining native FP64 support across its AMD Instinct™ GPU portfolio, AMD ensures that organizations no longer have to choose between accuracy and speed. This commitment is more than a design philosophy—it's a differentiator that sets AMD apart in a market where other vendors have deprioritized high-precision computing in favor of lower-precision AI throughput.

HPC and AI Budgets Among Commercial Enterprises

Intersect360 Research HPC-AI Budget Survey, 2025

Intersect360 Research contacted over 400 companies with at least $10M in annual revenue – 92% had HPC or AI budgets.

Among commercial (non-hyperscale) use cases, AI is usually found together with HPC

Almost all HPC users are now also AI users

More often than not, HPC and AI are merged as part of the same budget; when merged, the budgets are more focused on AI than HPC

Budgets for HPC or AI	Count	Pct.
HPC only	15	3.8%
AI only	40	10.2%
Both HPC and AI	338	86.0%
Total	**393**	**100.0%**

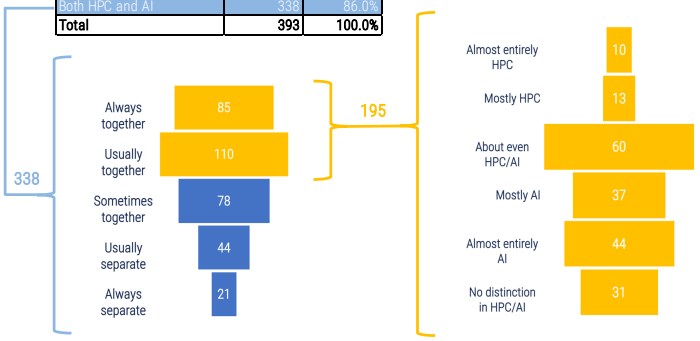

Figure 1: HPC-AI Budget Survey Results 2025 from Intersect360 Research

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

The AMD Instinct product line exemplifies this approach. The AMD Instinct MI300A APU brings together "Zen 4" CPU cores and AMD CDNA™ 3 GPU compute units with a unified pool of high-bandwidth memory, eliminating the traditional bottlenecks that have separated CPUs and GPUs. The AMD Instinct MI300X, MI325X, MI350X, and MI355X GPUs extend AMD leadership, offering more memory capacity and greater bandwidth to support the largest models and the most demanding simulations. Additionally, MI300A, MI300X, MI325X, MI350X, and MI355X GPUs all provide high-performance support for FP64.

These advances are not theoretical. AMD Instinct GPUs are at the heart of some of the world's most ambitious computing projects, including the Frontier supercomputer at Oak Ridge National Laboratory and the El Capitan supercomputer at Lawrence Livermore National Laboratory. The same technology is available to enterprise data centers, enabling organizations to scale from single-node deployments to exascale-class systems with confidence and efficiency.

All of this is supported by the open-source AMD ROCm™ software platform, which provides mature drivers, compilers, and libraries for both HPC and AI. The ROCm integration with leading frameworks such as PyTorch and TensorFlow means that developers can deploy and optimize workloads without friction, while the AMD commitment to open standards ensures that organizations are investing in a future-ready foundation.

In a landscape where precision and performance are often presented as tradeoffs, AMD delivers both. For organizations seeking to accelerate discovery, optimize infrastructure, and prepare for the next wave of scientific and AI innovation, AMD Instinct GPUs and AMD EPYC™ CPUs offer a proven, flexible, and forward-looking solution.

Market Dynamics

The modern world would not be possible without the innovations brought about by high-performance computing (HPC). Everything from a material science simulation to drug discovery relies on this technology, and many of these scientific HPC endeavors demand high numerical accuracy to ensure reliable results. Specifically, FP64 precision is required to achieve the accuracy and reliability demanded by scientific HPC workloads.

These high-precision scientific applications complement much of the artificial intelligence (AI) work currently being performed. In fact, many deep learning applications rely on formats like BF16 because it is sufficient for training and inference while speeding up the process.

While AI is certainly a valuable tool, high-precision HPC work is still absolutely essential for a variety of important tasks, and as such, vendors must consider their customers' high-precision needs alongside the growing reliance on AI.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

How HPC is Changing

The world of HPC has always been one of constant innovation, but the changes brought about by AI are starkly different from what the industry has seen in the past. HPC remains the engine behind scientific discovery, but today's HPC environments are adapting to support a new generation of AI-driven workloads alongside traditional simulation and modeling on the same physical infrastructure.

This shared, composable infrastructure is where operators take CPUs, GPUs, memory, and storage and pool their resources to dynamically allocate them to different jobs as required. Such an approach improves resource utilization and flexibility, allowing the same system to handle a mix of simulation, modeling, and AI workloads.

Such an allocation is important because of how AI workloads differ from those of traditional HPC projects. AI work often benefits from lower-precision formats like BF16 or lower. Precision is not as important for most AI workloads because deep learning models can tolerate small numerical errors without significantly affecting overall accuracy. Formats like BF16 enable faster computation and reduced memory usage, allowing for larger models and more efficient training and inference.

However, traditional HPC workloads demand a high level of accuracy and reproducibility. Climate modeling workloads requiring HPC resources are prime examples of this need for precision. These climate models simulate complex, interdependent physical processes like ocean currents or atmospheric dynamics – and these processes usually occur over vast spatial and temporal scales. Such simulations demand quintillions of calculations, and even small numerical errors can quickly accumulate and ruin the model's effectiveness. If lower precision is used – such as FP32 or BF16 – rounding errors can cascade, thereby leading to inaccurate predictions.

The Core of HPC: Native FP64 Support

With HPC workloads remaining extremely relevant, despite the advances in AI, it makes sense for organizations to continue to pursue high-precision infrastructure that can work with HPC and AI. Also, emulation does not provide the combination of accuracy and performance needed for HPC workloads. Thus, native FP64 support is non-negotiable for HPC. It is the only way to ensure the highest level of numerical fidelity needed for reproducibility and trust in scientific results, which simultaneously provide high performance.

As Scott Atchley, Chief Technology Officer at the National Center for Computational Science at Oak Ridge National Laboratory, explains: "Today, almost all our applications (i.e., > 99.9%) need floating-point numbers and almost all of them (i.e., > 99%) need more than 32-bit precision to get a correct answer."

Bronis de Supinski, Chief Technology Officer for Livermore Computing at Lawrence Livermore National Laboratory, reinforces this point

"Today, almost all our applications (i.e., > 99.9%) need floating-point numbers and almost all of them (i.e., > 99%) need more than 32-bit precision to get a correct answer."

"64-bit precision is critical to most scientific applications," de Supinski said. "They require it to get correct results. Our workload is almost exclusively such applications."

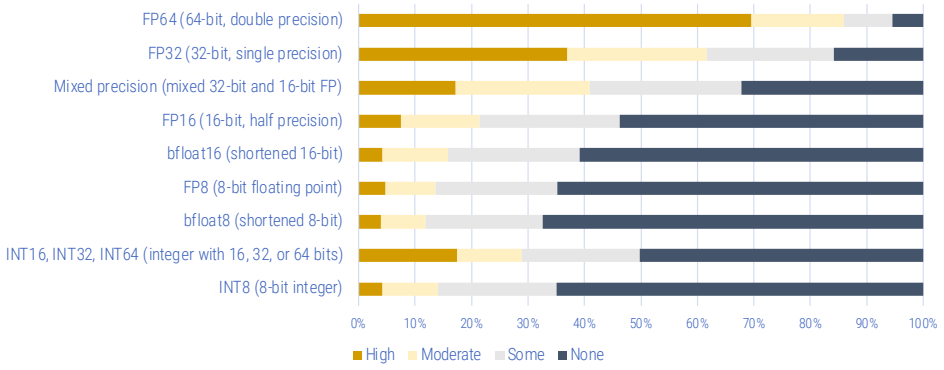

Figure 2: Future Importance of Precision Levels from Intersect360 Research

Emulation approaches such as the Ozaki method have significant limitations. Ozaki methods were developed to enable double-precision-like results on hardware where native FP64 performance is poor or unavailable. Rather than relying on FP64 or FP32, Ozaki I and II emulate FP64 GEMMs using multiple INT8 General Matrix Multiply (GEMM) operations and additional operations. This approach leverages the high throughput of INT8 GEMMs on certain GPUs to approximate FP64 results for matrix computations.

It is important to clarify that Ozaki methods are only applicable to GEMM operations and do not extend to vector operations, which make up a large portion of HPC workloads. For most scientific computing, which depends on vector math, Ozaki is not a practical solution.

Another key limitation is that Ozaki emulation does not meet IEEE floating-point arithmetic requirements—a universal standard for robust and reliable floating-point performance. This means Ozaki-based emulation may not deliver the numerical guarantees required for scientific reproducibility or regulatory compliance.

The performance of Ozaki depends on the specific hardware: if INT8 GEMM operations are much faster than native FP64 GEMM operations, Ozaki may provide a speedup for FP64 GEMM operations. However, this is only true for certain architectures and workloads and must be evaluated on a case-by-case basis.

However, a very important limitation is that the Ozaki and related methods are only applicable to General Matrix Multiply (GEMM) operations. They do not work for vector operations, which are prevalent in many HPC applications.

Emulating FP64 on lower-precision hardware in this way is certainly possible, but it generally introduces unacceptable accuracy, does not meet IEEE arithmetic requirements—a key universal standard for robust and reliable floating-point performance—and is limited in applicability since Ozaki methods are only useful for General Matrix Multiply (GEMM)

operations, whereas most HPC workloads are based on vector mathematics. Emulation using lower-precision formats can result in rounding errors that accumulate over millions of calculations, potentially leading to unstable simulations, erroneous predictions, or invalid research outcomes.

Emulation (e.g., Ozaki methods)

- Pros:
  - Enables double-precision (FP64-like) GEMM results on hardware where native FP64 is slow or unavailable, by leveraging high-throughput INT8 GEMM operations.
    Can provide performance benefits for matrix multiplication (GEMM) if INT8 GEMM is significantly faster than native FP64 GEMM on the hardware.
- Cons:
  - Only applicable to GEMM; does not work for vector operations, which are common in HPC workloads.
  - Does not meet IEEE floating-point arithmetic standards, which are critical for scientific reliability and reproducibility.
  - Limited accuracy and may introduce numerical issues calculations.
  - Difficult to implement and validate for scientific workloads.

Native FP64

- Pros:
  - High performance and accuracy for all double-precision workloads.
  - Meets IEEE floating-point arithmetic standards, ensuring robust and reproducible results.
  - Works for all HPC workloads, including matrix, vector, and scalar operations.
  - Simpler to use and validate for scientific computing.
- Cons: Can require additional silicon area depending on situation.

Emulation such as Ozaki 1 and 2 provide some value to those who lack high-performance native FP64 hardware. However, it is these emulation methods do not to match the accuracy and reliability of true double-precision support. In situations where precision and reproducibility are needed, such as in scientific computing, native FP64 remains essential.

de Supinski draws a clear distinction between the needs of scientific computing and AI workloads:

> *Scientific applications are looking to compute accurate results with known error characteristics. AI techniques look to get answers with statistical techniques that are probably right, but with little assurance that they are.*

"Scientific applications are looking to compute accurate results with known error characteristics. AI techniques look to get answers with statistical techniques that are probably right, but with little assurance that they are."

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

What's more, these emulation methods are typically limited to certain matrix operations and do not address the broader range of vector calculations that dominate HPC workloads. Thus, native FP64 remains essential for trustworthy results within HPC workloads.

Native FP64 support also enables HPC workloads to scale efficiently as research questions become more ambitious or as datasets grow larger. Atchley further highlighted the importance of precision, flexibility, and intelligent tooling.

"There was an application [on Frontier] that got a 10% speedup by avoiding promoting some data from FP32 to FP64 when it was not needed," Atchley said. "Research is needed to create tools to allow a developer to analyze their applications to determine when too much precision is being used to the detriment of performance."

In climate modeling, simulations must track the interactions of atmospheric, oceanic, and land processes over vast domains. When double precision is handled natively, researchers can run these complex, long-term simulations without sacrificing either performance or accuracy.

The ongoing commitment by AMD to native FP64 is a significant feature in the modern compute landscape. Rather than treating double precision as an afterthought or relying on weak emulation, AMD continues to engineer its hardware to deliver uncompromised FP64 performance to ensure users are not forced to compromise on performance or accuracy.

AMD Leadership in HPC Performance

This commitment to uncompromised double-precision performance is reflected in the AMD current and next-generation HPC products. Specifically, the AMD Instinct product line demonstrates the AMD commitment to advancing the field of HPC and AI.

The AMD Instinct portfolio currently features five flagship GPUs, including the latest MI350X and MI355X GPUs based on the new AMD CDNA™ 4 architecture. These products deliver impressive AI throughput and powerful FP64 performance for HPC workloads. Each GPU is designed to address the needs of both HPC and AI. The AMD Instinct MI300A APU, for instance, represents a breakthrough in converged computing by tightly integrating high-throughput AMD CDNA™ 3 GPU compute units with AMD EPYC™ "Zen 4" CPU cores and unified high-bandwidth memory. This eliminates traditional bottlenecks between CPUs and GPUs.

The AMD Instinct MI300X and MI325X GPUs, meanwhile, deliver exceptional double-precision and AI performance, with leading memory capacity and bandwidth to support the largest models and most demanding workloads. The latest AMD Instinct MI350 Series GPUs, including the AMD Instinct MI350X and MI355X GPUs, are based on the latest AMD CDNA 4 architecture and set a new standard for Generative AI and HPC. These GPUs deliver exceptional efficiency and performance for training massive AI models, high-speed inference,

and complex HPC workloads like scientific simulations, data processing, and computational modeling.

Table 1 below is a detailed dissection of each of these product offerings:

Table 1: The latest AMD Instinct™ family of GPU product offerings with featured specifications.

AMD Instinct™ Product Model	MI300A APU	MI300X GPU	MI325X GPU	MI350X GPU	MI355X GPU
Architecture	AMD CDNA 3	AMD CDNA 3	AMD CDNA 3	AMD CDNA 4	AMD CDNA 4
GPU Compute Units	228	304	304	256	256
AMD EPYC™ CPU "Zen 4" Cores	24	N/A	N/A	N/A	N/A
Matrix Cores	912	1,216	1,216	1024	1024
Memory	128 GB HBM3	192 GB HBM3	256 GB HBM3E	288 GB HBM3E	288 GB HBM3E
Memory bandwidth	5.3 TB/s	5.3 TB/s	6.0 TB/s	8. TB/s	8.0 TB/s
FP64 Vector	61.3 TFLOPs	81.7 TFLOPs	81.7 TFLOPs	72.1 TFLOPs	78.6 TFLOPs
FP64 Matrix	122.6 TFLOPs	163.4 TFLOPs	163.4 TFLOPs	72.1 TFLOPs	78.6 TFLOPs
FP32 Vector	122.6 TFLOPs	163.4 TFLOPs	163.4 TFLOPs	144.2 TFLOPs	157.34 TFLOPs
FP32 Matrix	122.6 TFLOPs	163.4 TFLOPs	163.4 TFLOPs	144.2 TFLOPs	157.3 TFLOPs
FP16/bfloat16 Matrix	980.6 TFLOPs	1.3 PFLOPs	1.3 PFLOPs	2.3 PFLOPs	2.5 PFLOPs
FP8 Matrix	1.96 PFLOPs	2.61 PFLOPs	2.61 PFLOPs	4.6 PFLOPs	4.6 PFLOPs
FP6/FP4 Matrix	NA	NA	NA	9.2 PFLOPs	10.1 PFLOPs

These AMD GPUs incorporate several advanced features that cater to the needs of the scientific community in the HPC and AI landscape. Unified memory architecture, exemplified by the AMD Instinct MI300A APU, enables both CPU and GPU cores to access a shared pool of high-bandwidth HBM3 memory. This design eliminates the need for explicit data transfers between CPU and GPU, reducing latency and simplifying programming for complex, converged workloads.

The AMD Instinct GPUs and AMD EPYC processors are at the core of some of the world's most powerful and energy-efficient supercomputers. Notably, the Frontier supercomputer at Oak Ridge National Laboratory and the El Capitan system at Lawrence Livermore National Laboratory both rely on AMD technology to achieve exascale performance. These systems not only set records for peak computational power but also maintain high precision and efficiency, enabling groundbreaking research in fields ranging from climate modeling to molecular dynamics and artificial intelligence.

Future-Ready HPC

In HPC, the only constant is change. AMD understands that today's breakthroughs are tomorrow's starting line, and the AMD Instinct MI300A APU is purpose-built for organizations that refuse to be left behind.

Based on the next-generation AMD CDNA 3 architecture, the AMD Instinct MI300A is engineered to overcome the persistent challenges that have limited traditional discrete GPU solutions. By integrating 24 "Zen 4" x86 CPU cores with 228 high-throughput GPU compute units and 128 GB of unified HBM3 memory, the AMD Instinct MI300A presents a single shared memory address space to both CPU and GPU. This coherent integration, enabled by the 4th Gen AMD Infinity architecture, is designed to eliminate performance bottlenecks from narrow CPU-GPU interfaces and to reduce the programming overhead associated with managing data movement and synchronization.

The AMD Instinct MI300 APU multi-chip architecture leverages advanced die stacking and chiplet technology to achieve dense compute and high-bandwidth memory integration. The result is a platform that delivers up to 5.3 TB/s of memory bandwidth, supporting the most demanding HPC and AI applications. Each device incorporates three "Zen 4" chiplets and six accelerated compute dies, all interconnected by 256 MB of AMD Infinity Cache™ technology. This design not only allows for power efficiency but also reduces data-movement overhead, supporting both scale-up and scale-out deployments. With up to 1 TB/s of bidirectional connectivity via eight 128 GB/s Infinity Fabric interfaces, the MI300A is built for flexible, high-performance multi-APU configurations. Virtualization support through SR-IOV enables resource sharing with up to three partitions per APU, ensuring adaptability for a range of HPC and AI workloads.

The robust, open-source ROCm software environment, combined with the architectural innovations of the AMD Instinct MI300A GPU, positions AMD to meet the evolving needs of next-generation data centers, delivering performance, efficiency, and flexibility for the future of high-performance computing.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Conclusion

The AMD commitment to HPC and AI is pushing a very clear and necessary message – organizations no longer must compromise between precision and performance. By integrating an AMD EPYC CPU and AMD Instinct GPU on one device, AMD delivers a platform that meets the rigorous demands of scientific discovery while accelerating the pace of AI innovation.

Native FP64 support, unified memory architecture, and robust AI throughput ensure that both traditional simulations and next-generation machine learning workloads are executed with uncompromised accuracy and efficiency.

FP64 support is needed for scientific computing to ensure accuracy, reproducibility, and trust HPC workloads require. And while AI continues to grow within the industry, HPC workflows will still demand high precision capability. Thus, the AMD unwavering commitment to native FP64 isn't being sacrificed to accommodate these new AI capabilities, and organizations no longer must choose between speed and precision.

Additionally, emulation methods such as Ozaki 1 and 2 are not sufficient for many important use cases. These approximate FP64 on hardware lacking high-performance, native FP64 support, but this comes with significant performance and accuracy trade-offs. Native FP64 is the only way to ensure speed, reliability, and reproducibility in scientific workloads requiring HPC.

For research institutions and enterprises seeking to optimize their HPC and AI environments and prepare for the future, AMD Instinct GPUs and EPYC CPUs offer a proven, flexible platform for HPC and AI computing.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com