Intersect360 Research White Paper: MICROSOFT AZURE DELIVERS HPC NOW INTO THE FUTURE WITH AMD

EXECUTIVE SUMMARY

While the need for scalability hasn't changed, the nature of scalability in HPC has, in applications and technologies, both on-premises and in the cloud. By combining traditional HPC and AI, researchers now strive to reach more predictive answers, more often. Data-rich applications in manufacturing, genomics, financial services, and oil exploration, among others, can be excellent fits for AI augmentation of HPC.

Cloud computing provides an entirely new dimension to scalability, beyond the confines of the data center. Cloud computing has become an increasingly popular method for reaching new levels of capability and capacity. In certain domains, organizations must be able to reproduce the same results regardless of where a given application is run.

According to Intersect360 Research, the fastest growing provider of HPC cloud solutions has been Microsoft Azure. Azure's momentum is based on completeness of services and solutions for HPC, and this is reflected in the loyalty of Azure users. Microsoft Azure has a greater proportion of its HPC users likely to increase their usage in the future than can be claimed by any other cloud provider, including Amazon Web Services (AWS). Intersect360 Research projects Azure to lead cloud providers in HPC market share gain in the coming years.

Meanwhile, in recent years, HPC has seen a new surge in compute from Advanced Micro Devices, Inc. (AMD). AMD recently launched its third-generation AMD EPYC 7003 CPU, and AMD has continued a zealous focus on high-performance applications. AMD adoption has grown substantially among HPC users.

As both AMD and Azure have surged in HPC, Microsoft is combining both with its HBv3 instances, featuring AMD EPYC 7003 processors. The HBv3 instances combine the complete HPC services of Azure with the built-for-performance aspects of EPYC. The performance-per-core characteristics can mean results are achieved with fewer processors and therefore lower licensing costs for many applications, resulting in lower total cost of ownership for HPC.

Furthermore, HBv3 Azure instances outperform existing x86 and ARM Graviton 2 instances in AWS, while supporting applications that are already optimized for x86. Replicating workhorse x86 environments on-premises and in the cloud can be the key to achieving optimal, replicable performance for a variety of HPC workloads, across multiple domains, carrying applications into the future. This is the new scalability in HPC.

White paper sponsored by Microsoft Azure and AMD. Full neutrality statement at https://www.intersect360.com/features/neutrality-statement.
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

MARKET DYNAMICS: THE NEW SCALABILITY

One persistent concept of High Performance Computing (HPC) is that it is scalable. Whatever the application, HPC is about bringing enough computing, memory, and data management capabilities together to solve the problem. HPC is a persistent need because of the very nature of scientific computing, that there are always more questions to answer, always a new insight to be gained.

While the need for scalability hasn't changed, the nature of scalability in HPC has. Problems still need to scale—that will never change—but due to advancements in computational approaches and wide scale adoption of cloud computing, there are new dimensions of scalability for HPC users to explore.

The New Scalability: Analytics and Machine Learning

Traditional scientific computation has been mathematically deterministic: input the data, do the math, get the answer. This computational approach has been applied to simulations in widespread domains, such as predicting the path of a hurricane, determining the best place to drill for oil, designing an optimal pharmaceutical, or optimizing airflow over a racecar. Accuracy improves along with refinements of the model and precision of calculation.

But brute calculation is not the only way to solve a problem. After all, a child does not learn to hit a ball by mastering physics calculations to predict the flight of the ball and the optimal swing angle, but rather through repeated experience. The same argument can be applied to almost anything a person learns, from language acquisition to driving a car. Today, this is being replicated in HPC, as machine learning approaches are revolutionizing the field of artificial intelligence (AI).

Along with the larger field of analytics, machine learning is applicable to any discipline in which patterns might be found in data. This extends to traditional HPC domains as well. We can still use calculation to compute the hurricanes path, but these models can be informed by historical data observing how past hurricanes have behaved. Neither approach is perfect. This hurricane is unique. By combining traditional HPC and AI, researchers might reach more predictive answers, more often.

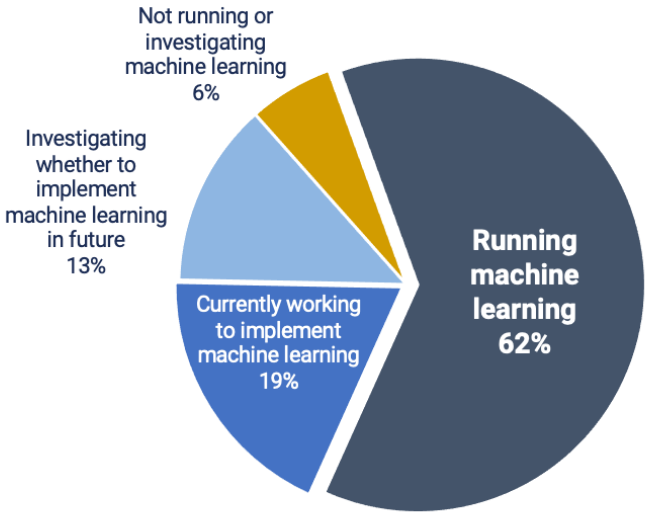
With HPC linked so inextricably to advancement, it's no surprise to find AI getting integrated into the pursuits of scientific insight and engineering breakthrough. Data-rich applications in manufacturing, genomics, financial services, and oil exploration, among others, can be excellent fits for AI augmentation of HPC.

A recent Intersect360 Research study showed that 62% of HPC users were currently running machine learning workloads as part of their HPC environments. Another 19% of respondents were working to implement machine learning, putting it in place or in planning at four out of five HPC sites. Additionally, there has been a corresponding increase in HPC budgets associated with the expansion into AI. (See charts.)

Data-rich applications in manufacturing, genomics, financial services, and oil exploration, among others, can be excellent fits for AI augmentation of HPC.

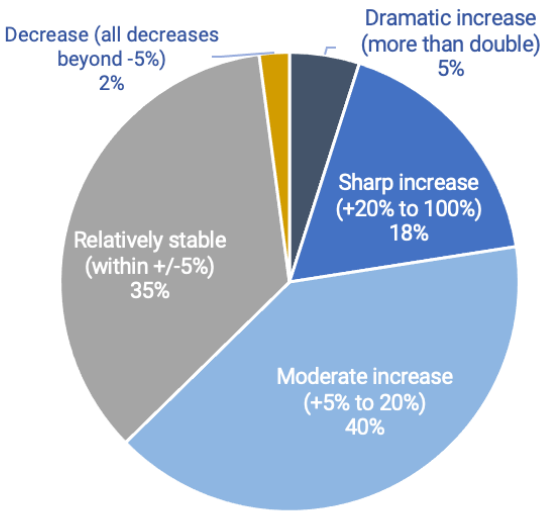
Machine Learning in HPC Environments

Intersect360 Research, 2021



Effect on High-Performance Workload Budget Related to Incorporation of Machine Learning

Intersect360 Research, 2021



The New Scalability: Cloud

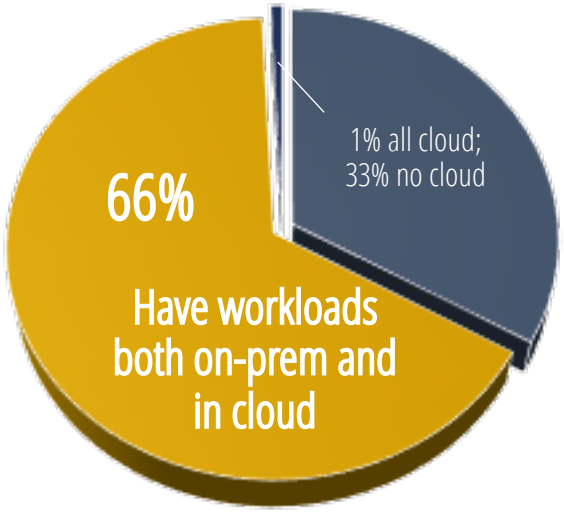
Cloud computing provides an entirely new dimension to scalability, beyond the confines of the data center. With resources that scale elastically, cloud computing has become an increasingly popular method for reaching new levels of capability and capacity. According to recent Intersect360 Research surveys, 66% of HPC users operate both on-premises and in the cloud. (See chart.) Furthermore, cloud computing is outgrowing on-premises computing, with double-digit compound annual growth. While Intersect360 Research projects that HPC

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

will still remain primarily on-premises,[1] scaling workloads from core to cloud is critical for the majority of organizations.

Proportion of Surveyed HPC Users with Workloads Both On-Premises and in Cloud

Intersect360 Research, 2021



Cloud computing presents some additional challenges along with its opportunities. Data sovereignty and stewardship have to be managed across domains. Workloads have to be balanced to optimize software licenses, costs, and efficiencies. And most importantly for certain domains, organizations must be able to reproduce the same results regardless of where a given application is run. To be sure, cloud can be a great option for specialized resources that are only needed occasionally, but to scale to a higher volume of workloads with elastic flexibility, it can be extremely important to match configurations with internal datacenters.

INTERSECT360 RESEARCH ANALYSIS

AMD Storms Back into HPC with EPYC

As HPC continues to chase new insights and discoveries, the heart of the performance race is focused on the central processing elements, with the vast majority of HPC systems being built on x86-architecture CPUs. For years this space with dominated by Intel Corporation, but recent years has seen a major surge from AMD. AMD recently launched its third-generation AMD EPYC 7003 CPU, and with each successive generation of its "Zen" x86 architecture, AMD has continued a zealous focus on high-performance applications.

To scale to a higher volume of workloads with elastic flexibility, it can be extremely important to match configurations with internal datacenters.

[1] Intersect360 Research forecast data, 2021.

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

The latest "Zen 3" architecture in the EPYC 7003 CPU delivers a ~19% improvement in operations per cycle over the previous generation, according to AMD. This performance boost comes from a combination of enhancements, such as increased bandwidth for both loads and store, and improved latency for certain calculations.

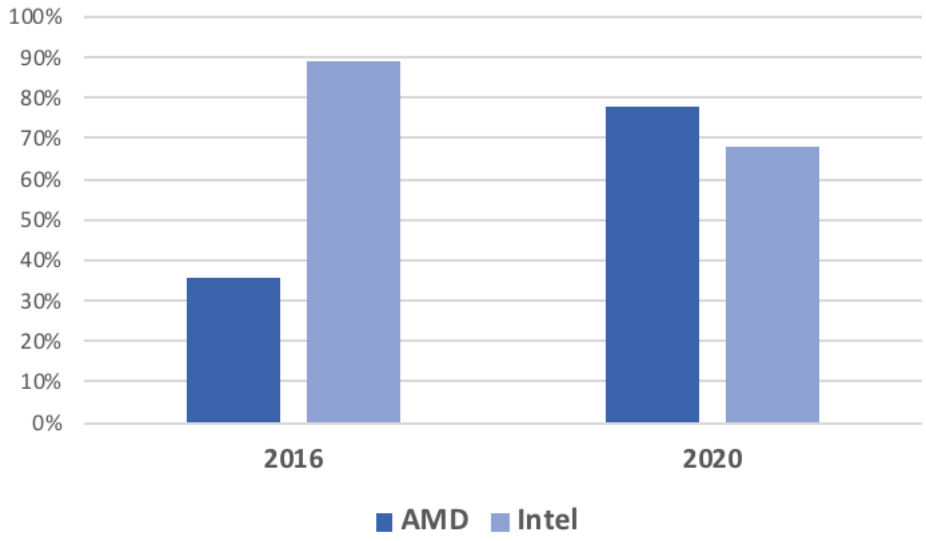
The AMD EPYC 7003 enhancements go beyond the architecture performance. The complete SOC (system on chip) has enhanced memory and cache functionality and additional security features, while maintaining socket compatibility with the previous AMD EPYC version.[2] In particular, the L3 cache is integrated into a single, large 32GB reservoir, rather than two smaller ones, allowing full cache allocation to any single core that may need it, thereby benefiting applications with databases that may fit into the single, larger cache. In one other subtle improvement, the AMD Infinity Fabric clock is now synchronous with DRAM memory. This helps reduce latency in waiting for data, resulting in an improvement for memory-sensitive applications, which are common in HPC.

To scale to a higher volume of workloads with elastic flexibility, it can be extremely important to match configurations with internal datacenters.

AMD CPUs now have a presence in 70% of HPC sites, an astonishing transition from three years ago.

Percent of HPC Users with Favorable Forward-Looking Impressions of CPUs [3]

Source: Intersect360 Research, 2021



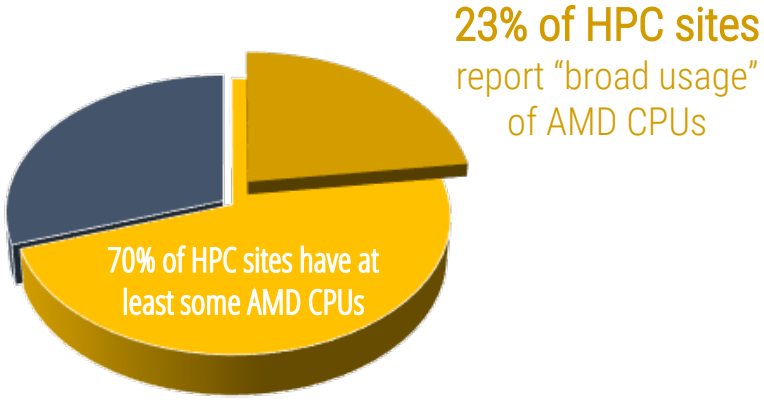
[2] AMD EPYC™ 7003 Series processors require a BIOS update from your server or motherboard manufacturer if used with a motherboard designed for the AMD EPYC™ 7002 Series processors. A motherboard designed at minimum for EPYC 7002 processors is required for EPYC 7003 Series processors.

[3] Intersect360 Research data from multiple studies. **2016:** Special study: "Processing Elements for HPC"; question, "Overall, how favorable is your forward-looking impression of each of the following, with respect to your HPC workloads? (1 = Completely unfavorable; 5 = completely favorable)"; scores are combined percentage 4 and 5 for AMD Opteron versus Intel Xeon. **2020:** "Vendor Satisfaction and Loyalty in HPC"; question, "What is your impression of each of the following vendors' future prospects for HPC?" (five-point scale); scores are combined top-two responses (Very Impressed; Impressed) for AMD EPYC CPUs versus Intel Xeon CPUs.

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

Current Penetration of AMD CPUs Among Surveyed HPC Sites [4]

Source: Intersect360 Research, 2021



With AMD now shipping its third generation of EPYC CPUs, AMD's momentum has begun to show up in real customer deployments. AMD was named as the processor vendor in only 5% of surveyed HPC systems in 2017 and 2018 combined.[5] Today, 23% of HPC users say they have AMD EPYC processors in widespread use. An additional 47% are testing or using AMD EPYC at some level, giving AMD CPUs a presence in 70% of HPC sites, an astonishing transition from three years ago. (See charts above.)

Microsoft Azure HBv3 for HPC

While AMD has been blasting forward with processing improvements for HPC, the fastest growing provider of HPC cloud solutions has been Microsoft Azure. Azure's momentum is based in large part on completeness of services and solutions for HPC, and this is reflected in the loyalty of Azure users.

Microsoft has a full range of HPC-focused instances available in Azure,[6] with a range of VM sizes,[7] and users have taken notice. Microsoft Azure has a greater proportion of its HPC users likely to increase their usage in the future than can be claimed by any other cloud provider, including Amazon Web Services (AWS). (See chart.) Based on survey responses, Intersect360 Research projects Azure to lead cloud providers in HPC market share gain in the coming years.

Microsoft Azure has a greater proportion of its HPC users likely to increase their usage in the future than can be claimed by any other cloud provider, including AWS.

[4] Intersect360 Research HPC Technology Survey, 2021.

[5] Intersect360 Research HPC User Site Census surveys across 2017 and 2018, total proportion of systems for which AMD was identified as CPU provider, including half-system credit for systems in which AMD was a shared CPU provider.

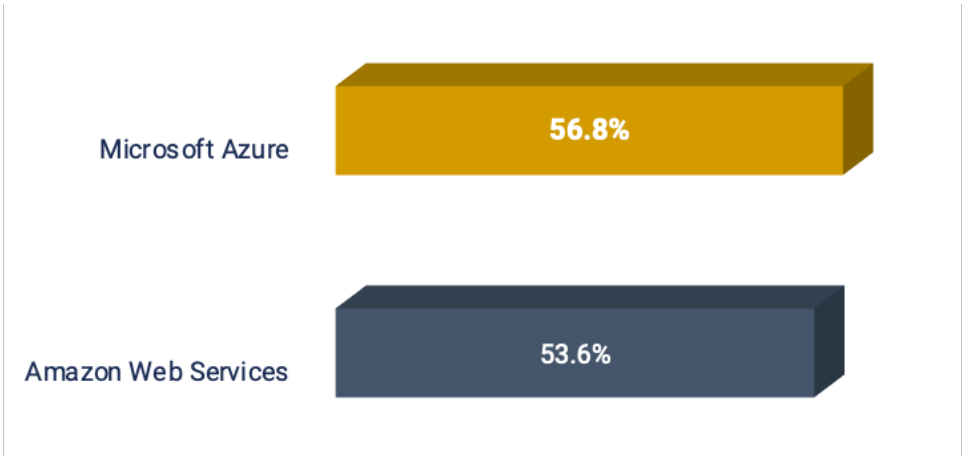
[6] https://azure.microsoft.com/en-us/solutions/high-performance-computing/

[7] https://docs.microsoft.com/en-us/azure/virtual-machines/sizes-hpc

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

Proportion of Cloud Providers' Current HPC Users Likely to Use Same Service "More" or "Much More" in Two Years

Source: Intersect360 Research, 2021



As both AMD and Azure have surged in HPC, Microsoft is combining both with its HBv3 instances, featuring AMD EPYC 7003 processors. The HBv3 instances combine the complete HPC services of Azure with the built-for-performance aspects of EPYC. HBv3 instances offer 200 gigabits per second of memory and a variety of vCPU (virtual CPU) sizes.

In a blog post online,[8] Microsoft has released benchmarking data showing HBv3 performance and scalability results, with comparisons against other HPC instance types. According to Microsoft testing, HBv3 is 2.6x faster than the Intel-based H16mr instance on a sample of small-scale HPC workloads. According to the blog, HBv3 is "capable of scaling MPI HPC workloads to nearly 300 VMs and ~33,000 CPU cores."

The advantages of AMD EPYC 7003 for HPC are already discussed above, but in the context of cloud take on yet another dimension. Its performance-per-core characteristics can mean the same results can be achieved with fewer processors and therefore lower licensing costs for many applications, resulting in lower total cost of ownership (TCO).

Microsoft Azure HB-Series vs. AWS Graviton

The diversification of processing elements and workloads make it challenging for HPC users to evaluate which choices offer the highest performance. Cloud has a role here as well, as published benchmarks of disparate instance types offer insights into HPC performance.

The most basic comparison is of CPU performance in a single VM, and here EPYC in Azure's HBv3 compares very well against a likely competitor, Amazon's ARM-architecture Graviton processor in AWS. AWS has published some results for its C6gn.16x.large instance, based on its 64-core, 2.5 GHz Graviton2 ARM processor. On single-VM floating-point performance,

[8] Jithin Jose, Jon Shelley, and Evan Burness, "HPC Performance and Scalability Results with Azure HBv3 VMs," https://techcommunity.microsoft.com/t5/azure-compute/hpc-performance-and-scalability-results-with-azure-hbv3-vms/ba-p/2206471

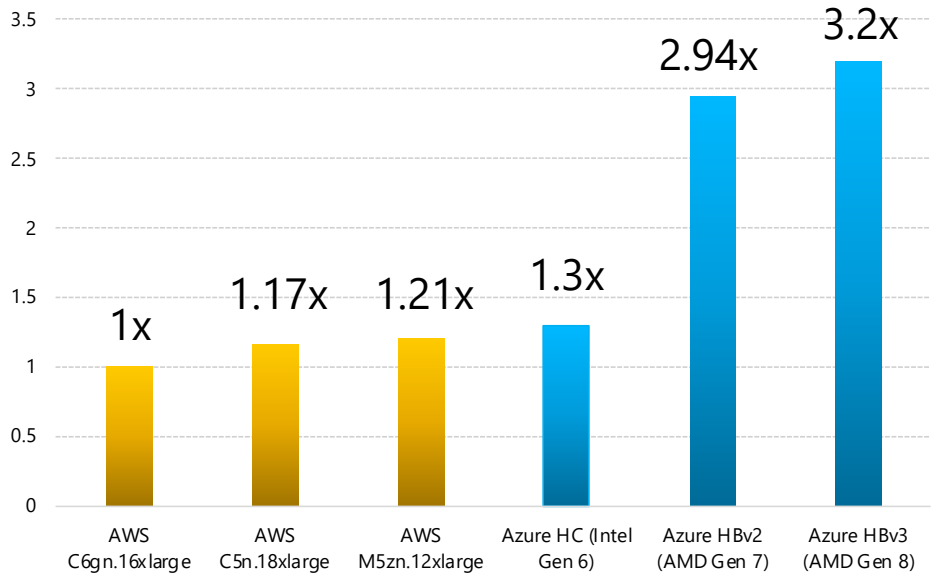
measured by SPECrate®2017_fp, Azure HBv3 offers an estimated 3.2x higher performance than AWS C6gn.16x.large. (See chart.)

Beyond the single-VM performance, there are notable difference. Azure offers InfiniBand, the most commonly used system interconnect in HPC, for its HBv3 instances, whereas InfiniBand is not an option on AWS. According to interviews conducted by Intersect360 Research, many HPC users favor Azure specifically because of the availability of InfiniBand.

Furthermore, there are potential advantages to HPC users in sticking with the x86 option of EPYC on HBv3. Existing applications that are optimized for x86 will already be compatible, and they can run the same both on-premises and in the Azure cloud. By contrast, the Graviton2 processor is available only on AWS, preventing users from matching the cloud instance to their on-premises HPC systems.

Comparison of Single-VM Floating Point Performance, AWS Instances vs. Azure Instances, Based on SPECrate®2017_fp_base (Est.) [9]

Source: Microsoft, 2021



Building into the Future

With HPC expanding in so many dimensions, there is always an emphasis on doing something new. That said, it's also important to remember where applications have been. HPC users have spent decades optimizing codes for x86. Recoding for other options takes time and effort. GPUs have been well-adopted as accelerators, and AMD has invested in open ecosystem programming and migration tools for bringing past optimizations forward.

[9] Estimated SPECrate®2017_fp_base comparison based on measured runs by Microsoft on Azure HBv3, Azure HBv2, Azure HC44rs, AWS C5n.18xlarge, and AWS M5zn.12xlarge.

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

ARM architectures are also getting increased trial and attention in HPC, and in time, they could have a significant future. For now, adoption of ARM CPUs trails far behind x86, and while experimentation is important, x86 is the industry workhorse. Replicating that environment both on-premises and in the cloud can be the key to achieving optimal, replicable performance for a variety of HPC workloads, across multiple domains, carrying applications into the future. This is the new scalability in HPC.

For more information about Microsoft Azure solutions for HPC, visit www.azure.com/hpc.

For more information about AMD solutions for HPC, visit www.AMD.com/HPC.

White paper sponsored by Microsoft Azure and AMD. Full neutrality statement at https://www.intersect360.com/features/neutrality-statement.
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com