Accelerating AI and HPC in the Cloud

Network Bandwidth in the Cloud: When InfiniBand Matters

New Scalability for AI and HPC: Supercomputing in the Cloud With InfiniBand
Many organizations have been running high-performance computing and artificial intelligence workloads for years, almost exclusively using on-premises infrastructure to meet their various needs. However, organizations that want to use a public cloud provider for their HPC and AI workloads have typically been stymied by the lack of scalability, since demands can vary widely over short periods of time.

This paper looks at how these systems are evolving, how technologies like InfiniBand are enabling new levels of cloud elasticity to meet the myriad emerging use cases, and how a unique partnership between a cloud provider and silicon vendor is stepping up to meet this long unfilled need.

The HPC Evolution

Hyperion Research forecasts the broader HPC market, including cloud services, will grow from $30 billion in 2020 to $47 billion in 2024, driven by the increasing need for efficient, scalable, high-speed data processing that can accurately process large volumes of data. HPC capability is increasingly important in a broad range of applications, including financial services, medicine, research, seismic exploration, and government and defense. HPC has become the behind-the-scenes power for everything from CT and MRI scans to household products to stock market analyses.

Many applications are simply not possible without HPC. For others, HPC offers the benefit of time-savings, enabling organizations to perform tasks in hours rather than years.
AI and HPC: Close Computing Cousins

Similarly, AI has grown into one of the most inspiring areas of tech evolution, with the global AI market projected to reach around $554 billion by 2024, according to IDC.1 AI is powering a wide array of important applications, including natural language processing, robot-powered process automation, and machine learning and deep learning. The progress has been astounding: Where just a few years ago, high-end AI and deep learning was delegated to tasks such as image classification, today’s AI solutions are pushing the envelope by augmenting human understanding, preferences, intent, and even spoken languages.

Ultimately, AI is improving our access to knowledge—providing better healthcare outcomes and making transportation more efficient, for example—and it will undoubtedly fuel transformations that are still beyond our imagination.

In the past three years, AI’s demand for compute power has grown by almost four orders of magnitude, rapidly outpacing Moore’s Law. AI companies are finding new ways to get more out of each piece of silicon using tools like mixed-precision modes, enabling them to do more with less. But this isn’t enough. The only way to satisfy the expanding ambitions and imaginations of the AI research community is through massive scale—and the only way to scale up real-world applications without slowing down performance is via fast, low-latency connections such as InfiniBand.

(Cloud) Supercomputing Challenges

Although some HPC and AI workloads differ in terms of the types of processors and communication patterns they employ, they have much in common. For example, the best real application performance for both HPC and AI workloads requires purpose-built computing platforms and could not be achieved using just generic servers. Additionally, having very fast interconnects between processing elements is key to delivering the kind of performance that today’s workloads demand. For those reasons, many organizations have made huge investments in their on-premises CPU, networking, and storage technologies to meet their existing HPC and AI needs.

1 “IDC Forecasts Improved Growth for Global AI Market in 2021,” IDC, Feb. 23, 2021
However, substantial challenges remain in enabling scalability and growth, and the casual HPC and AI use cases that may not warrant investment in on-premises infrastructure. While traditional computing workloads are easily migrated—or scaled—to the cloud, HPC and AI infrastructure has been historically lacking at public hyperscale cloud providers. What’s more, most public cloud infrastructure that is touted as HPC is often just repackaged generic servers with general-purpose networking that is capable of employing only a subset of the MPI libraries and techniques used in HPC. Jobs typically achieve a smaller scale, with higher latency and lower bandwidth between nodes—impeding scalability and often negating the cost benefits of elasticity. Worse yet, the performance at scale of these systems can fall far short of real application workload demands.

The Need for Speed

To overcome these challenges, enterprises are looking for public cloud-based HPC and AI platforms that can offer a solution highly optimized for HPC scenarios, with the highest, most consistent levels of performance, using CPUs and GPUs designed for the task. Requirements for these systems include:

- Purpose-built high-performance interconnect fabrics employing trusted technologies and library stacks that provide an experience identical to the best on-premises supercomputer-class networks.
- Cloud-native manageability.
- A right-sized, building-block approach that makes scalability—up or out—practical, easy and cost optimized for applications of varying characteristics.

Most public cloud infrastructure that is touted as HPC is often just repackaged generic servers with general-purpose networking that is capable of employing only a subset of the MPI libraries and techniques used in HPC.

- Ready access to a broad range of tools that let enterprises build solutions that meet the scope of the HPC and AI problems at hand, including both cloud-native and hybrid schedulers and job control systems.
- A full range of configurations to support workloads including:
  - Lightweight GPU platforms with a wide reach for real-time inference and geographically distributed workloads that demand a low cost of persistence per virtual machine.
  - Midrange GPU systems for loosely coupled, cost-per-throughput-optimized business scenarios.
  - Flagship, massively parallel configurations of tightly coupled GPUs, starting at eight and scaling to thousands, for unprecedented time to solution with the most demanding AI and HPC tasks.

These solutions must offer locality to enterprise data and a simple way to move data between the cloud platform and enterprise data repository.
The Cloud Built for HPC and More

Fortunately, these capabilities are available today from Microsoft Azure, which, in partnership with NVIDIA, delivers purpose-built HPC and AI hardware in the cloud to meet even the most demanding real-world application workloads at scale while meeting price-performance and time-to-solution requirements.

This product portfolio provides the best of on-premises infrastructure, previously available only to organizations at supercomputer centers, bringing an end-to-end HPC platform with all the amenities of the public cloud, including pay-as-you-go and on-demand infrastructure.

Enterprises also benefit from Azure security, including the broadest range of industry security standards, backed by Microsoft’s team of more than 3,500 cybersecurity experts who work 24/7 year-round to ensure the safety and security of enterprise data.

Organizations that need to augment existing on-premises HPC or AI infrastructure can take advantage of the ability to extend current workloads from on-premises to Azure cloud-based infrastructure. Further, every enterprise can leverage NVIDIA technologies in Azure to take advantage of HPC and AI in the cloud and gain a computational microscope to do what traditional instruments cannot in the fields of:

- Medical research, to accelerate research in diagnostics and treatments for diseases like COVID-19 and beyond
- Stellar exploration, including discovery of new exoplanets
- Industrial design, to locate hidden stress points in vehicles and other machinery, prevent wear and tear, and enable proactive, predictive maintenance
Why Azure and NVIDIA

These solutions are made possible by the power and efficiency of low-latency, high-bandwidth InfiniBand, with in-network computing offload engines and adaptive routing, enabling new levels of cloud elasticity without the application-specific network topology and placement considerations common for on-premises deployments. Unlike other clouds that offer lower performance, generic interconnects, the NVIDIA and Azure partnership offers performance levels for real-world applications that reduce time to solution and optimize overall resource consumption, resulting in a better total cost to solution.

With InfiniBand and applications that achieve near-linear scaling, enterprises are free to scale out without slowing down in Azure.

Designed to provide the best performance for tightly coupled, high-throughput workloads, these solutions are built for the needs of HPC and AI apps rather than one-size-fits-all computing use cases. Azure’s HPC VM products are highly optimized for per-server, per-cluster, per-workload performance. As a result, the Azure and NVIDIA partnership delivers SPEC benchmark test results two to three times faster than the competition.

The partnership between Microsoft Azure and NVIDIA takes the best of proven on-premises technology into cloud and is the only solution that can deliver:

- Purpose-built HPC and AI hardware for the highest performance
- InfiniBand networking and switches that deliver high data throughput, extremely low latency, and smart in-network computing acceleration engines
- The trust and security realized by working with proven industry leaders
- World-class, round-the-cloud enterprise support
- Real-world application performance at the largest scales

Next Steps

Click here for more information.