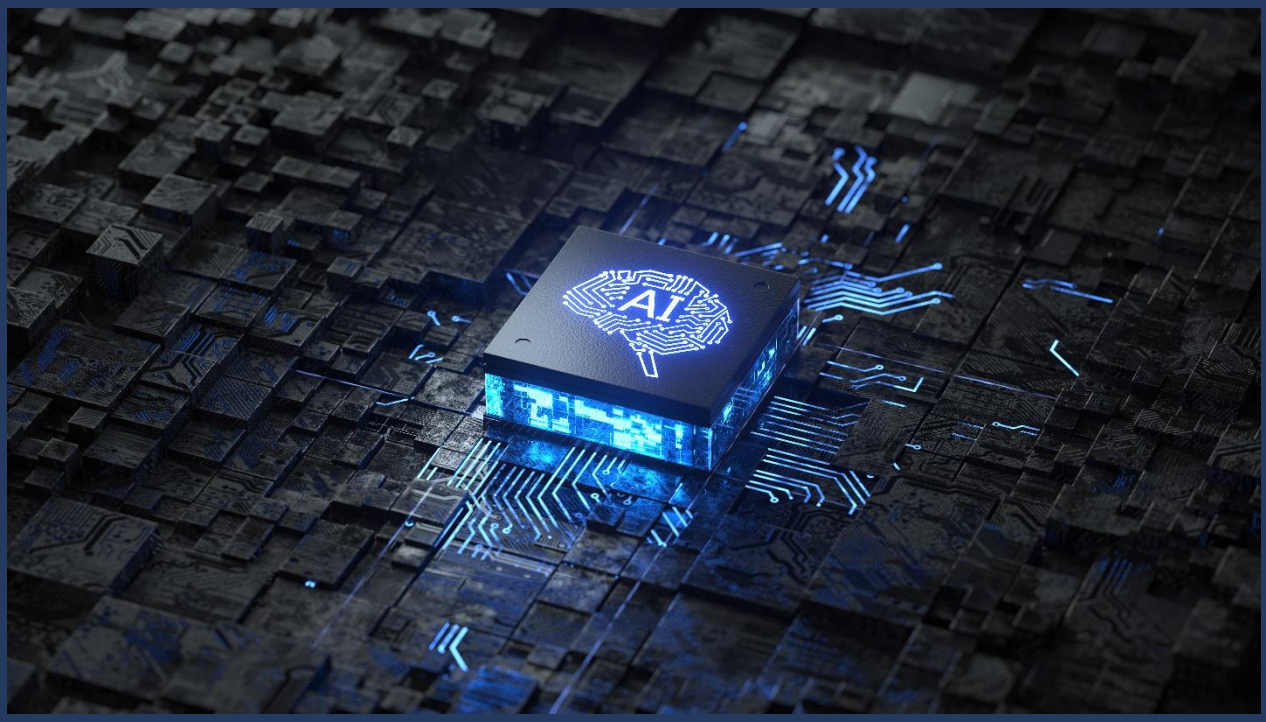


Supercomputing performance in the cloud

Microsoft Azure and NVIDIA Bring Top 30[1] Supercomputer Performance to Public Cloud Services

Microsoft | NVIDIA

[1] *Based on June 2021 TOP500*

AI is changing the face of HPC, accelerating it in ways never considered possible.

The use of high-performance computing (HPC) platforms to power use cases such as AI, machine learning and deep learning continues to push the bounds of available infrastructure performance. As data scientists and technologists develop increasingly complex and demanding models, access to supercomputer-class infrastructure determines how quickly these applications can deliver insights and business value.

However, building infrastructure that can deliver TOP500 (a biannual survey of the highest performant supercomputers) level performance is costly and takes time to test and deploy. Utilizing the Azure cloud, customers are able to access supercomputer type systems without having to physically build out the system and infrastructure. This also allows them access to newer technologies when they become available without being locked to a physical cluster for the life of that cluster (3-5+ years typically).

As a result, there is an increasing desire to make use of public cloud services that can provide infrastructure for demanding HPC use cases as it is needed. The ability to "pay by use" to leverage existing software and containers with the additional advantage of a broad and comprehensive stack makes the cloud a competitive option. Until recently, HPC cloud instances didn't quite stack up to on-premises HPC platforms. An effective platform for demanding HPC workloads must deliver a highly performant cloud service that is on equal footing with the most capable on-premises options.

According to a recent study by Hyperion Research study[2], 20% of HPC workloads are currently running in the public cloud. However, much of this was cloud augmentation of on-premises infrastructure to provide additional resources. The real growth in public cloud usage will be driven by supercomputer-class cloud HPC instances or virtual machines, and by access to the most up-to-date libraries and software that support AI, deep learning, and machine learning workloads. And this demand will come from companies of all sizes. A recent Forrester study found that 39% of small and medium-sized firms plan on running HPC and AI workloads on public cloud services.

[2] Hyperion Research Study April 2020

Microsoft's Azure and NVIDIA bring true Supercomputer performance to the cloud

Microsoft and NVIDIA have long worked together to deliver technology solutions that push the boundaries on the high end of the performance spectrum. They are now collaborating to deliver the most capable cloud AI service in the market today. Consider just some of the accomplishments of this partnership to deliver HPC solutions:

- ✓ ***Delivering six generations of NVIDIA silicon to customer through Azure***
- ✓ ***Azure provides customers the choice of 12 VM sizes with NVIDIA technologies***
- ✓ ***Working together to provide advanced visualization support for customers***

This performance-centric approach has resulted in impressive metrics. Today, the joint Azure/NVIDIA service provides the most performant A100 offering in the cloud. Simply put, Azure is designing and shipping purpose-built cloud services designed to let our most demanding customers scale up and scale out in the most performant and cost-effective way.

The addition of the latest Azure VM size, NDv4, delivers an entirely new level of HPC performance for cloud services. The performance data below can be viewed in more detail on a video produced by Microsoft and NVIDIA. It is the first top 30 class cloud service, and as of this writing is the fastest supercomputer in the public cloud based on June 2021 Top500 results.

High-Performance Linpack (HPL)

This benchmark looks at uniformly dense linear equations and is a key component of determining the TOP500 list. The Microsoft and NVIDIA engineers collaborated to optimize the hardware and software, along with HPL, to generate the results provided here. Once the optimization efforts were completed, the cloud service was able to deliver linear performance at scale, topping out at 16.59 PF, using 164 nodes (linear scaling shown in Figure 1). This score would place it 26th on the June 2021 TOP500. The cluster also achieved 148 PF on the HPL-AI.

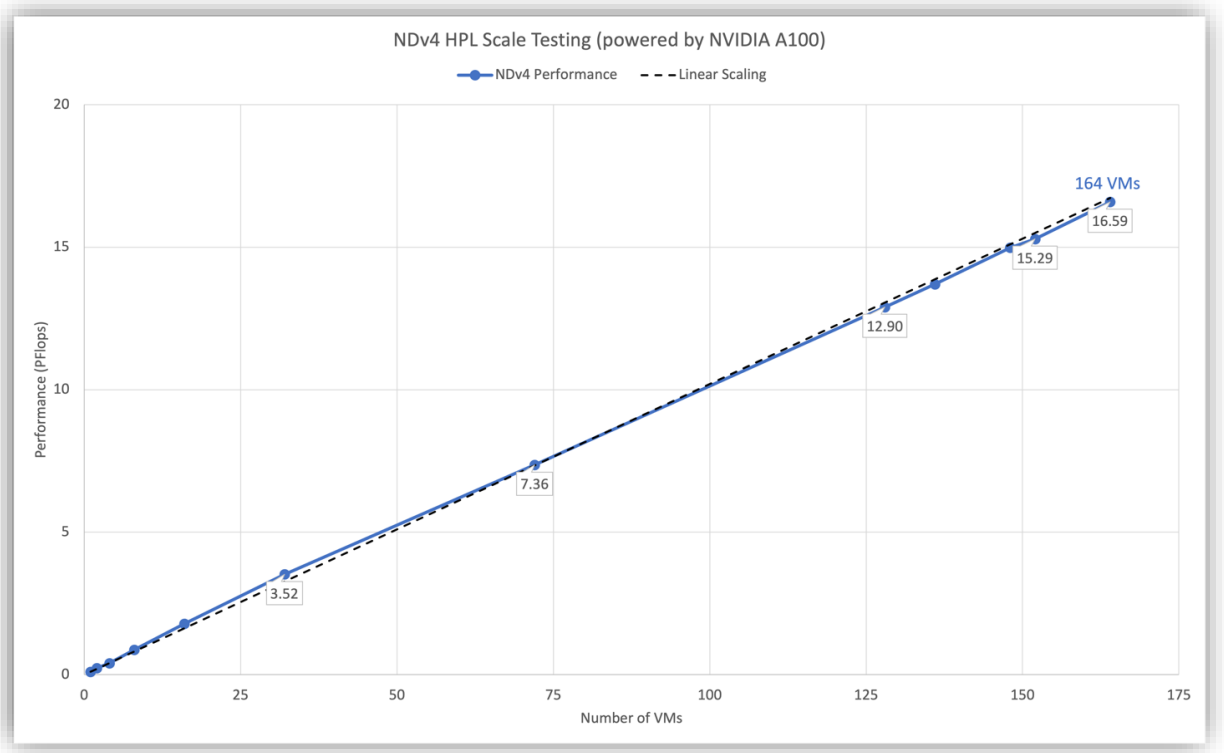


Figure 1:

NDv4 HPL Scaling Performance 1BERT-Large Natural Language Processing (NLP) Deep Learning Training

This deep learning model was developed by Google's AI team. Deep learning models use huge data sets and highly complex models and are the foundation of the benchmark. BERT-Large training is used as a standard performance test for large cluster deployments to measure how well different systems perform training to convergence. Some BERT-Large benchmark results show that training to convergence can take many hours. The new Azure VMsize, NDv4, did far better. Using the optimized MLPerf BERT Training recipe, with 32 nodes (256 A100 GPUs), training to convergence took less than four minutes. Using 128 nodes (1,024 GPUs) brought the time to train down under 90 seconds. Figure 2 shows the scaling performance of MLPerf BERT-Large Training on up to 128 nodes. Again, all processing was done in the cloud as was the case for the HPL benchmark.

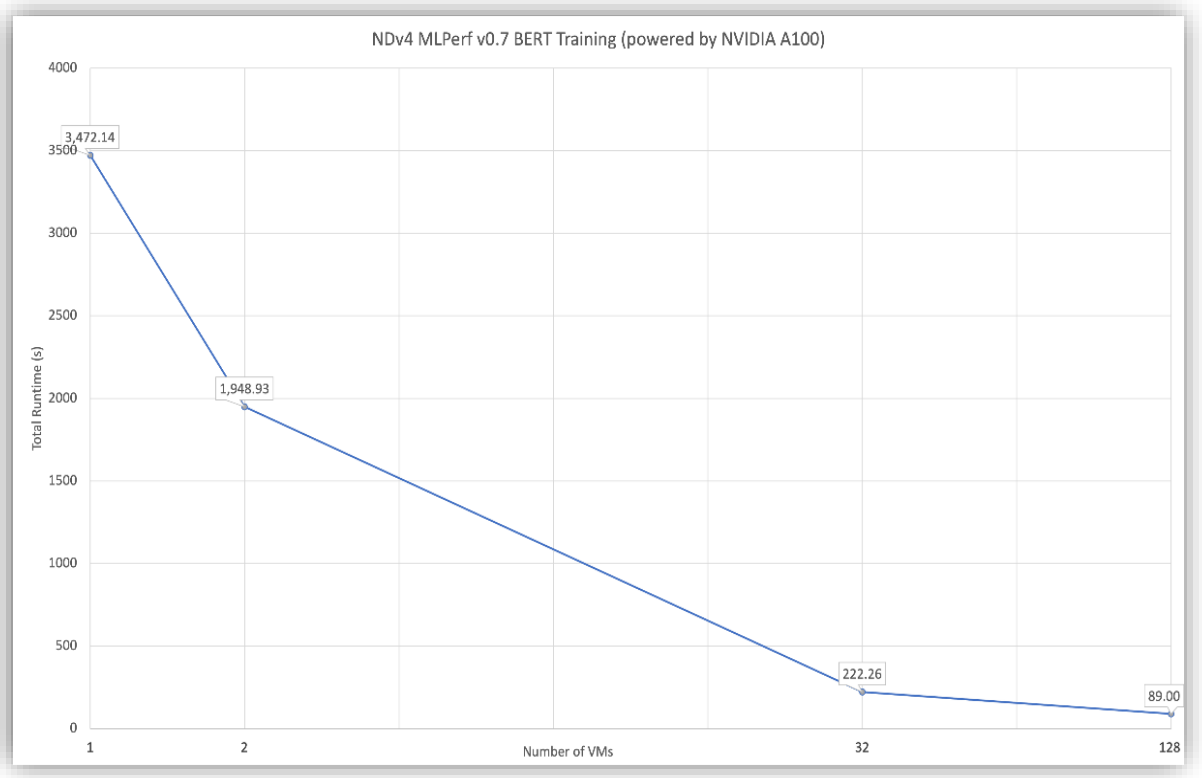


Figure 2:

NDv4 MLPerf BERT Training at Scale

The technology behind the Azure supercomputer-in-the-cloud service

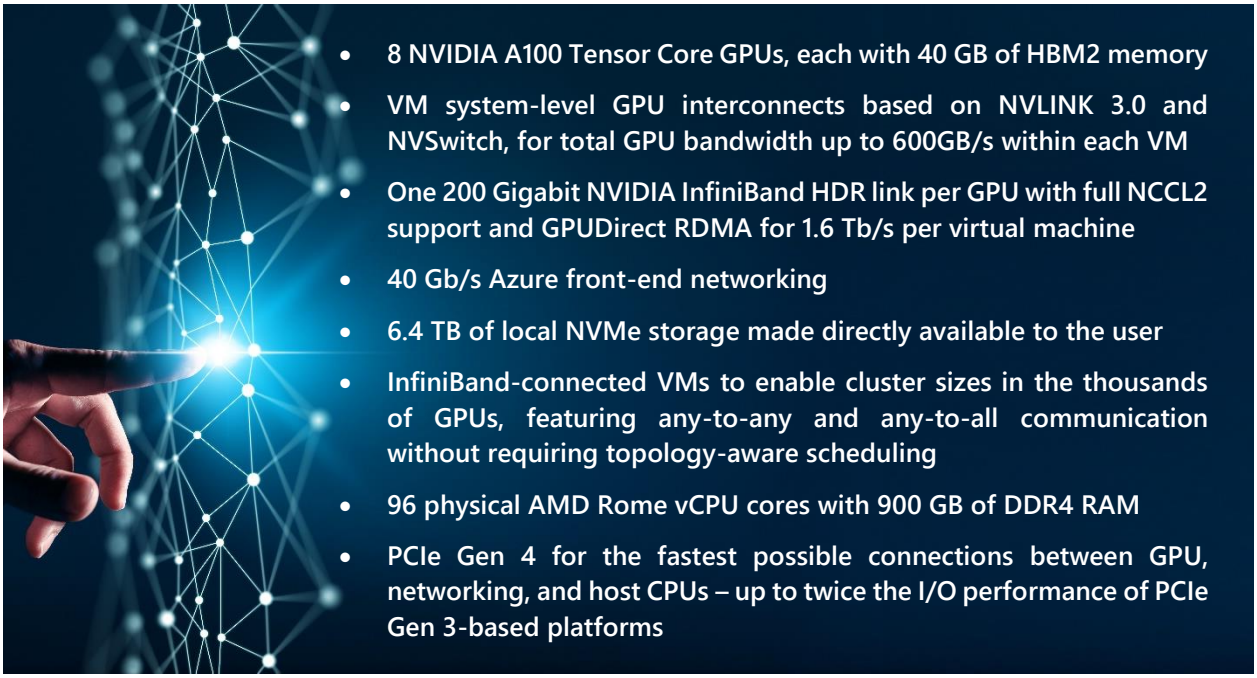
Delivering this kind of game-changing HPC performance in a cloud service required an uncompromised approach to creating a purpose-built solution that would redefine the possibilities of HPC performance in a cloud service. NDv4's comprehensive design includes a complete set of balanced infrastructure components, each of which delivers the highest levels of performance currently available. Key design points of this new VM size include:

- ✓ *A full range of GPU and CPU capabilities that can scale applications to tens of thousands of cores and thousands of GPUs.*
- ✓ *The fastest InfiniBand interconnects, including edge to cloud connectivity. Use of the latest Mellanox network connectivity solutions delivered the performance necessary to make a cloud service competitive.*
- ✓ *Local high-performance storage.*

This infrastructure leverages known and familiar software tools and processes to support endto-end workflows. NDv4 supports scaling up and scaling down the use of the cloud resources to optimize efficiency. Intelligent services, such as Azure ML, are provided to support AI, machine learning and deep learning at scale.

Azure NDv4 VM instance details

State-of-the-art technology components and building blocks come together to power the Azure NDv4 VM instances. This VM leverages advanced best-in-class components and technologies to deliver the performance shown in the benchmarks detailed above. The specifications include:



This type of computing capability would form the foundation of a very capable on-premises HPC deployment. However, it is now available as a cloud service, with all the benefits, agility, and flexibility that entails.

Users of this public cloud service can leverage the NVIDIA NGC software catalog to get workloads running more quickly on the NDv4 VM size. This catalog brings together key NVIDIA technologies to lower the barrier to building AI models. This hub of GPU-optimized AI, HPC and data analytics software simplifies and accelerates creating end-to-end workflows. The catalog provides enterprise-grade containers, pre-trained AI models and industry-specific SDKs that can readily be used on the NDv4 instances. These software resources enable organizations to achieve faster time to market, and time to value with their HPC and AI applications.

Key takeaways

The current and future generations of HPC workloads are more demanding than those that were common just 12-24 months ago. These demanding applications make many of the existing HPC-centric public cloud services a poor match for these new workloads. Together, Microsoft and NVIDIA have achieved an industry first with NDv4: a public cloud instance that delivers top 30 performance within the TOP500. With GPU VMs, customers can optimize their spending, have access to the latest GPUs and supporting technology, and optimizing utilization levels.

The NDv4 has brought together a set of leading-edge technologies in a well-balanced design that can scale applications to thousands of GPUs. NDv4's unprecedented levels of HPC and AI advancing a wide range of scientific simulations and AI-driven applications,

For more information on the Azure NDv4 VM, please go to https://bit.ly/3hsELpu.