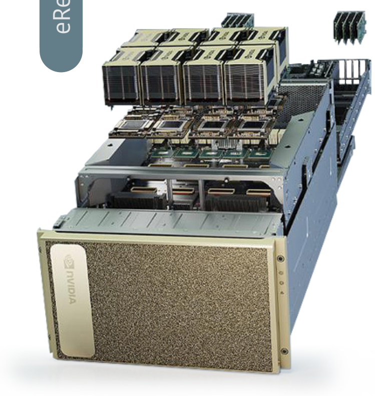
Own The Base, Rent The Spikes

Why hybrid cloud infrastructure delivers unmatched IT flexibility and scalability for AI development workflow

As data science becomes an essential tool of discovery across disciplines, universities and research institutions must have a strong vision for their IT infrastructure. Technology that will do more than just support innovation, and elevate research to a level that sets it free from the traditional constraints of compute speed and data storage. It's in the interests of broader society to ensure students and researchers have the compute power that enables them to test the limits of their own imaginations.

Artificial intelligence (AI) has long since marched out of the pages of science textbooks and into every nook and cranny of our modern world and is being fed with ever-bigger datasets. Inspired by the demands of deep learning and data analytics, NVIDIA DGX™ systems deliver groundbreaking performance and faster insights with an end-to-end portfolio of supercomputing systems optimised and purpose-built for the unique demands of AI.

These systems bring high-performance computing hardware on-premise, with a flexible, scalable suite of solutions. Beginning with NVIDIA DGX™ Station A100, data scientists and developers can get access to fast, productive experimentation with the power of a data centre delivered in a form factor that can be deployed at the desk. Built on the same software powering all NVIDIA DGX systems, work can be easily migrated from desk-side development on DGX Station A100 to a DGX A100 in the data centre, as well as the cloud. NVIDIA DGX systems are designed and built to give data scientists the most powerful tools for AI exploration, from the desk to the data centre to the cloud.



▲NVIDIA DGX A100

The NVIDIA DGX A100 can be partitioned into many GPU instances each fully isolated with their own high-bandwidth memory, cache, and compute cores, via the Multi-Instance GPU (MIG) feature. Each A100 GPU within the DGX A100 can be split into seven GPU instances using MIG. Across the eight A100 GPUs this provides up to 56 GPU instances to deploy to users or workloads. A single NVIDIA DGX A100, for example, could deliver 25 data scientists the performance equivalent of two dedicated GPUs each[1], with the processing power equivalent to a rack of enterprise CPU servers. This enables multiple workloads to run in parallel on the system, including AI training and inference workloads that traditionally use dedicated hardware.

Combining these capabilities enables the IT department to use DGX A100 as a unified building block for an AI centre of excellence, flexibly adapting to the demands of analytics, training and inference. Compared with traditional computing platforms which are often under-capacitised for AI demands, DGX A100 delivers exactly the right amount of resources for every workload. With MIG, multiple GPU instances can power AI workloads, right-sized for the needs of students and researchers or as a dual use platform for teaching and research.

AI workloads demand high processing capabilities and on-premise systems such as the NVIDIA DGX A100 deliver industry-leading accelerated computing, powered by the world's fastest GPU - the NVIDIA A100. In addition to the DGX A100 packs higher memory capacity, extreme internal and external bandwidth and an NVSwitch architecture for parallelising complex workloads across multiple GPUs.

With more and more AI workloads coming from every faculty – featuring ever bigger datasets and ever-increasing model complexity – traditional cloud and on-prem infrastructure can quickly be stretched beyond its capacity, potentially starving AI workloads of the resources they need.

Data science teams need effortless access to computing resources to power their work; they don't want to spend their time worrying about limits – they should be free to experiment and drive insight instead of worrying about infrastructure. Many organisations are seeing escalating spend in cloud in response to insufficient on-premise IT platform resources.

However, as larger and more complex workloads are executed in cloud, many organisations see an escalation in cost associated with data gravity, namely the time and cost associated with moving large datasets and complex models from where they're created, to where the compute resources lie.

For this reason, forward-leaning organisations are increasingly employing a hybrid-cloud for infrastructure, with on-premises capacity designed to address the "steady-state" demand of users and departments, combined with supplemental cloud-hosted capacity to address the "spike" in user demand based on the temporal needs that exceed on-premises capacity. This approach is embodied in the mantra - "own the base, rent the spikes" which enables organisations to enjoy the benefits of a fixed cost infrastructure that offers the lower cost per AI workload associated with on-going steady-state needs, combined with cloud capacity to ensure temporal demands are adequately serviced and users are not starved of computing capacity.

Additionally with the containerisation of the NVIDIA software stack managed and hosted in the NVIDIA NGC™ catalog – organisations can effortlessly run workloads, collaborate across platforms, and move work from cloud to on-premises. This can include seamlessly moving between systems like a DGX Station for initial prototyping to a DGX A100 for scaled AI training to even a DGX SuperPOD for the most complex jobs, all without any re-writing of code or additional effort by the user.

UNDERSTANDING THE AI ADVANTAGE OF A HYBRID ON-PREM/CLOUD SYSTEM

The difference between an on-premises system and the cloud is akin to that between buying or renting a home.

There's less capital upfront for renting. You pay as you go and if building repairs are required, it will be handled by the property owner. If you should need more – or less – space, you move on and rent somewhere more suited to your changed requirements. So it is with the cloud: It's a lower financial barrier to entry, and gives you the ability to quickly scale up or down to a different-sized computing cluster.

Buying is a one-time, fixed cost. You buy the house and stay there as long as you like (or as long as you keep paying the mortgage). And as long as everyone fits in the house, you can invite over more people to stay.

Investing in an on-premise system such as the NVIDIA DGX A100 means it can be used for as much time and as many projects as the hardware can handle, making it easier to iterate and try different methods without considering cost. Utilising the MIG feature is like to being able to the walls in the house instantly.

On-premise hardware can be used on multiple experiments, essentially without limits, and with known, stable costs and full control over configurations, security and data. The university owns the base-load compute and can run it 24x7. MIG allows the GPU resources to be easily assigned to the work that needs them at the time – for example running 7 instances within a GPU for inference work, and then overnight combining all the instances to one larger GPU for a deep learning application.

Running a sophisticated hybrid system – own the NVIDIA DGX base, leverage the cloud for the temporal spikes – unlocks the best of both worlds. It allows data scientists to run their AI workloads on a deskside or server based DGX system on-premises, and burst to the cloud for unexpected demands. There's flexibility to move from one environment to another at different points in the journey, from initial experimentation to large-scale deployment. Utilising cloud resources is ideal for initial model testing and experimentation, proof-of-concept work before the data sets grow too large. Cloud also offers the potential to spin up discrete workloads and applications on a short term basis.

When building complex AI models with huge datasets, operating costs for a long-term project can escalate, which risks causing researchers to focus on each iteration or training run they undertake, reducing their freedom to experiment. Data plays a vital role in ensuring an AI model is accurate and maintains accuracy over time as new data is added. Because of this, data gravity—an analogy of data's ability to attract additional applications and services—comes into play. As models become more complex, data sets start growing exponentially, and with more frequent model iteration, teams hit an inflection point where data gravity starts to significantly drive up costs. Organisations are starting to realise that they need to train where their data lives, using a purpose-built co-resident AI infrastructure to achieve the lowest-cost-per-training run.

Scientists and students using on-prem systems at universities or research institutes don't have to count how many hours they're racking up, or budget for how many runs they can afford over a particular timespan. An on-prem system allows unlimited iteration and testing time for a one-time, fixed cost.

If a new methodology fails at first, there's no added investment required to try a different variation of code, encouraging the 'what if' experimentation that's so essential for discovery.

The more an on-prem system is used, the greater the return on investment for the university or research institution – not that you can put a price on ground-breaking innovation.

XENON – ENABLING RESEARCHERS TO DO GREAT NEW THINGS

For over 25 years XENON has been delivering innovative solutions for universities and research teams looking to iterate faster, collaborate more, and develop solutions for tomorrow's problems today.

Balancing private and public cloud resources can assist teams to be agile, responsive and budget smart.

Explore the resources below, and get in touch with the XENON team to explore solutions for your IT challenges.

Contact XENON:

- 1300 888 030 (Australia)
- +61 3 9549 1111 (International)
- info@xenon.com.au
- www.xenon.com.au

Talk to a Solutions Architect

TECHNICAL FOOTNOTE

1. 5,000 TFLOPS per DGX A100 system / 56 MIG slices = 90 TFLOPS per MIG slice. Rounding down to account for overhead, 50 MIG slices were assumed. If you have 25 researchers, each researcher can get 2 MIG slides with 180TFLOPS, which is equivalent to having 2 NVIDIA V100 instances.

MIG enables each of the A100 GPUs in the DGX 100 to be divided into seven independent units, which can be operated individually or combined in any number. There are eight A100 GPUs in a DGX A100. The MIG feature is also available in selected PCIe form factor GPUS.

XENON®
High Performance Computing

© 2023 XENON Systems Pty Ltd. Terms of Use. XENON-Own the Base-Rent the Spikes-WP07B.pdf