Building Private Clouds

How a private cloud can turn your university IT department into a profit centre

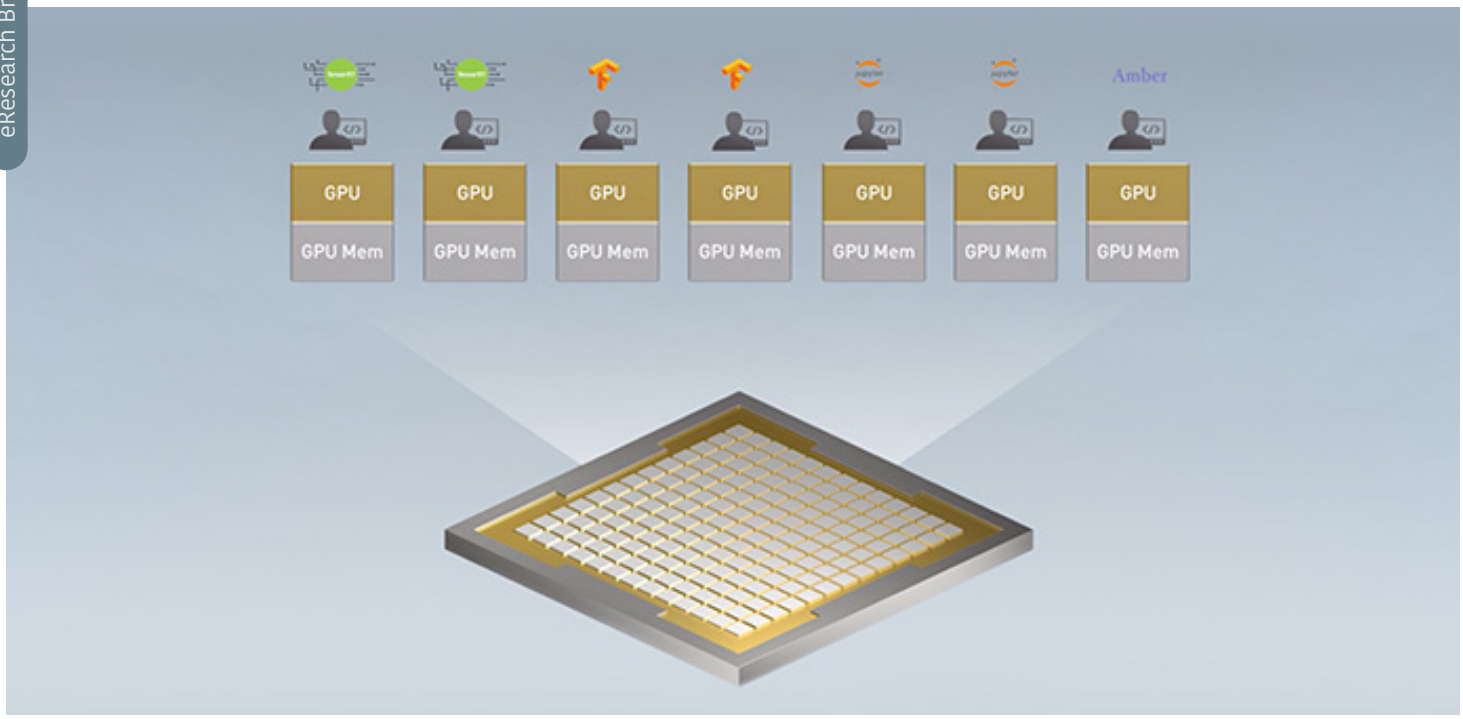
Cutting-edge IT infrastructure is not only a worthy investment, it's increasingly key to a university's ability to attract top students and postgraduates.

What if you could turn your IT cost centre into a profit centre – or at the very least structure it so that it is generating returns that go beyond its main role as the essential instrument of AI research, facilitating experimentation and discovery.

NVIDIA DGX™ systems are not only designed specifically to fast-track deep learning and data science, they can be built around a hybrid on-premises/cloud model. This makes them flexible and scalable in line with demand, and unlocks a consumption model for compute resources that can slot in with the funding models that most universities must navigate.

For example, most IT infrastructure will cycle through five-year timelines, give or take, while research grants that help to underpin resources run on much shorter cycles, usually between one and five years.

Designing your IT infrastructure so that it can be seamlessly shared between faculties, individual research projects or even external users or other institutions, enables a charge-out/user pays model for compute resources. This can be designed to fit the overall university ethos, perhaps charging only external users, or only for postgraduates on campus for short-term grant-funded research projects.



▲With MIG, jobs run simultaneously on different instances, each with dedicated resources for compute, memory, and memory bandwidth, resulting in predictable performance with quality of service and maximum GPU utilization.

DESIGNING A SHARED RESOURCE MADE EASY

While a hybrid on-prem/public cloud is one possible model, university IT departments should be wary of ballooning costs when there are enormous datasets being stored for long periods and compute loads running 24x7, as they must.

A better approach is to build an internal private cloud for your university, enabling you to leverage agile hardware and new-generation virtualisation capabilities to create the chargeback model that is right for your institution. The capabilities of NVIDIA DGX™ systems make shareability seamless, so much so that multiple institutions could collaborate to pool their budgets and create a supercomputing resource across campuses.

NVIDIA's Multi-Instance GPU (MIG) makes it simple to share compute resources across different faculties, universities or research institutes. MIG allows administrators to partition each NVIDIA A100™ Tensor Core GPU into as many as seven instances, each fully isolated with their own high-bandwidth memory, cache and compute cores. These MIG instances can be deployed independently, or joined in any combination to increase the processing power available for specific applications and workloads.

The ability to deploy right-sized GPU capacity for any AI workload, with guaranteed quality of service, ensures full optimisation of valuable resources and extends the reach of accelerated computing to more users. The availability of such precise allocation means that this valuable resource should never be sitting idle.

The MIG innovation now makes it possible to assign the GPU capacity to multiple users using the same machine at the same time.

SCALABILITY BY DESIGN

Scalability is built into the NVIDIA DGX systems, providing an affordable entry point and the potential to scale as needs grow. NVIDIA DGX™ Station A100 packs four A100 GPUs into a workstation form factor, delivering data centre performance in a box. With all the MIG features across the four A100s, the DGX Station is ideal for individuals or small teams. The next step up is the DGX A100, with eight A100 GPUs. Greater scale is available through building a cluster of DGX A100s. NVIDIA has reference architectures with tested storage and networking configurations which go from five DGX A100's in a DGX POD, to twenty DGX A100's in a DGX SuperPOD. The NVIDIA DGX SuperPOD™ can scale to 140 interconnected DGX A100's, providing a ready made architecture for exascale high performance computing. With the heavy lifting of design and cluster architecture already completed, the DGX SuperPOD allows clusters to be up and running quickly, usually in less than a month.

The ability to share resources within individual A100's, and the ability to bind multiple A100 instances together on the fly for aggregated parallel processing power delivers the flexibility required to optimise infrastructure investments.

XENON's solution specialists work in partnership with university and research institute IT teams to understand specific requirements and develop a plan to make it work, knowing that DGX systems are designed to afford scalability, modularity and shareability.

THE POWER OF THE PRIVATE CLOUD

When weighing up public versus private cloud, for an institution with on-premises IT infrastructure, a private cloud is often considered the far simpler choice logistically. For cost conscious institutions, private cloud infrastructure allows for expenses to be contained and known. Utilising resource allocation and job scheduling tools along with containerisation allows costs to be tracked to projects or faculties and linked to grants, while simultaneously delivering an agile on-demand experience for end users.

Data gravity becomes a key driver in selecting between on-premise and cloud infrastructure. AI models' accuracy and usefulness increases as the data sets they work with increase. These data sets are randomly called and need to be available in a high speed or nearline tier of storage. The costs to do this in public cloud arena quickly becomes exorbitant, especially when data sets reach 100TB+ scale. At this point, on-premise storage and AI compute is the logical, cost effective solution.

Beyond the hardware, the modern cloud experience is built on containerisation. Using containers and virtualisation, applications can be deployed and managed abstracted from the hardware layer. This brings the public cloud experience to on-premise environments, and also enables hybrid cloud solutions.

NVIDIA NGC™ catalog is an open source library of software containers and pretrained AI models optimised to run on DGX systems and NVIDIA GPUs. These NGC containers encompass a wide range of industries and applications. NGC containers can be deployed on any NVIDIA DGX system which allows for easy and predictable interoperability between form factors and even across to public cloud environments if desired.

The XENON Cluster Stack takes containerisation to a higher level, providing a modern containerised framework for high performance computing operation and management. The XENON Cluster Stack makes cluster computing easy to operate, manage and maintain across all the functions such as job scheduling, resource allocation, user access and storage management. The XENON Cluster Stack can be deployed to manage a DGX POD or SuperPOD, with the NGC containers running the AI models across the GPU instances within the XENON Cluster Stack environment. The XENON Cluster Stack delivers an agile, on-demand like cloud experience to users utilising on-premise infrastructure.

XENON AND THE WALTER AND ELIZA HALL INSTITUTE'S PRIVATE CLOUD

Australia's longest-running research organisation, the Walter & Eliza Hall Institute (WEHI), engaged XENON to help it design, install and deploy a significant overhaul of its on-premises IT infrastructure, including setting up a private cloud.

Since 1915, WEHI has been advancing medical research in cancer, diabetes, arthritis and malaria, to name just a few. The institute has more than 80 laboratories and more than 1000 staff, including 60 bioinformaticians and 30 computational biologists. WEHI runs in excess of 100 clinical trials at any given time and was facing increasing demand for HPC resources to analyse clinical data.

The demand was increasing for HPC resources, with bigger datasets and more scientists on staff. Teams sometimes had to queue up to process a 'job' on the existing cluster, and it could sometimes take up to a month to test and run. This wasn't reflective of the world-class research institute that WEHI is.

The brief was to deliver an agile, 24x7 infrastructure that would support the WEHI scientists to focus on outcomes at speed rather than cumbersome workarounds. WEHI's eResearch manager had a vision of pooling and providing resources in a virtualised environment, instead of a system of individual but shared machines. This approach of using virtual machines would ensure that all processes gain equal use of the infrastructure at any time while also increasing resource utilisation.

Working together, the WEHI – XENON team designed a private cloud solution that has allowed WEHI to transform their researchers' compute support. The solution delivered includes:

- Compute hardware, storage and networking,
- Private cloud management system and architecture,
- Batch job queuing systems,
- Self-service virtual machines.

WEHI now has an innovative shared resource model which is a commercial grade private cloud which delivers the efficient processing power the scientists require.

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.

Uitilising containerisation, the XENON Cluster Stack, and an appropriate agile private cloud architecture delivers increased productivity, resource utilisation and costs savings for universities and reseach institutes.

Explore the resources below, and get in touch with the XENON team to explore how private cloud architecture can benefit your team.

Contact XENON:

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

Talk to a Solutions Architect

MORE INFORMATION

- ***Read the XENON WEHI case study***
  https://xenon.com.au/case-studies/wehi-and-xenon-design-private-cloud-for-next-generation-cancer-disease-and-medical-research/
- ***Learn more about XENON Cluster Stack***
  https://xenon.com.au/hpc-solutions/xenon-cluster-stack-xcs/
- ***More about NVIDIA DGX systems***
  https://xenon.com.au/products-and-solutions/nvidia-dgx-systems/
- ***More about NVIDIA DGX SuperPODs***
  https://xenon.com.au/products-and-solutions/nvidia-dgx-super-pod/