# 5 Questions to Consider for Your AI Infrastructure

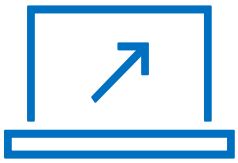
By: Andrew Glinka | April 2022

In 1884, Herman Hollerith's invention of the punched card tabulating machine was patented, leading to the 20th century being dominated by punched card data processing. I can only imagine how impossible it would seem to Hollerith that by 2020, technology would exist, such as the NVIDIA A100 GPU that would have the capability to process the equivalent of about 20 billion punch cards per second[1].

In September of 2021, the Enterprise Strategy Group (ESG) noted in their eBook, The Four Infrastructure Essentials for AI/ML Data Pipeline and Data Lake Environments, that maximizing infrastructure performance and utilization is essential for these environments. AI is complex with ever-changing requirements, Dell Technologies enables efficient sharing of resources among data scientists with VxRail, the only HCI solution for AI and MLOps certified for NVIDIA AI Enterprise[2] that is simple to provision and manage.

Organizations that address the following 5 critical AI architecture questions with a Dell Validated Design for AI or MLOps will improve their opportunities to achieve a faster time to value with better ROI vs. our public cloud competitors.

1. **Does your organization have the foresight to predict the GPU requirements for every data scientist, every trained model, and every inferencing job?** The answer is likely "no". Dell Technologies can help customers deal with their unknown AI requirements whereas public cloud providers are selling their platform services and GPU-enabled instances at a premium whether customers fully utilize the resources or not.



Dell Technologies along with NVIDIA and VMware offer an HCI platform that helps customers virtualize and maximize their GPU infrastructure[3].

With the introduction of Ampere GPU architecture, a new feature for partitioning slices of a physical GPU called Multi-Instance GPU (MIG) allows for resources of a fractional GPU. MIG enables the sharing of a GPU amongst multiple workload requirements, virtualized. This makes it easy to dedicate GPU resources where needed and on the fly. This simplifies bringing AI hardware resources to an organization's data science projects.

2. **What will be your organization's GPU utilization?** In the current business environment, there is a race to gain insights from data using AI, a global pandemic that has led to supply chain issues and the rise of crypto mining. These factors have all contributed to a scarcity of GPUs, so costs will likely

[1] Based on Dell analysis. December 2021
[2] Based on Dell analysis of systems validated as NVIDIA-Certified Systems - Data Center Servers, March 2022.
[3] New Validated Design Unlocks the Power of AI

remain high for the foreseeable future. This magnifies the need for a higher overall GPU utilization to achieve a greater overall return on investment. Many factors will come into play to determine a GPU's utilization. Will the GPU be deployed as bare metal or virtualized? Can the GPU be configured with GPUDirect to storage resources, maximizing data throughput? Multiple data science jobs will compete for the same resources with a physical deployment model. A data scientist may have their productivity in a Jupyter notebook clobbered by someone else's model training.

## In the Public Cloud, GPUDirect is limited to certain instance types for GPU-to-GPU communication, but what about GPU to storage communication for massive datasets?

In a virtualized deployment, Quality of Service can be achieved for multiple workloads sharing a GPU, and instances will have the flexibility of having their GPU resources re-allocated to other instances (though typically with the requirement of a restart). Data processing bandwidth may be throttled back because GPUDirect cannot be easily enabled -- or at all for some virtual workloads. With the Dell Validated Design featuring vSphere and NVIDIA AI Enterprise, no sacrifices are necessary to maximize GPU utilization. Instances can take advantage of a live vMotion when GPU resources need to be re-allocated, and GPUDirect can maximize data throughput for data storage.

3. **Is it too expensive for your Data Science project to succeed?** There's a good chance that piloting a Data Science project in the public cloud can be the lowest risk option. Pay as you go for GPU instances, storage, and networking helps reduce risk in trialing AI or ML initiatives. But what about when the project succeeds? A long-term commitment to a GPU-enabled instance is the most cost-effective means to run AI/ML workloads in the public cloud, but how different is that to committing to host a GPU on-premises or in a hybrid cloud? Consider that a three-year commitment in the public cloud to an A100 GPU can cost more than the procurement of an A100 GPU on-premises[4]. However, unlike in the public cloud, the ROI with on-premises improves after year three, whereas the meter keeps running in the public cloud.

4. **Can the GPU be brought to your organization's data, or do you have to bring the data to the GPU?** If the answer is that the data needs to be brought to the GPU, consider the drawbacks. Whether the requirement is copying or moving the data, all activities come with an additional element of risk. The data may be exposed to bad actors, and the cost of moving and hosting additional copies of the data needs to be considered. Even if an organization's data is already in the public cloud, there is no guarantee that the region has the necessary GPU instances available. For example, at the time of writing this blog in March of 2022, Canadian customers using the Google Cloud could not access GPU instances in the Toronto region. They could only access an NVIDIA P4 in Montreal or had to go to Iowa for a greater range of GPU options.

[4] Based on Dell Technologies analysis. December 2021

Public Cloud resources, especially GPU-enabled instances, are subject to regional availability at the discretion of the Public Cloud provider or in competition with other customers.

Customers in this situation should consider bringing the GPU to their data on-premises or via an Equinix co-location facility, for example, to avoid moving their data. Transferring data and creating copies can make the AI journey more complicated and expensive.

5. **Is the attached storage up to the task of an AI workload?** Thanks to parallelism, AI model training and inference jobs process massive amounts of data. Data needs to be fed to these jobs at scale. Single pipe throughput is not as important as file concurrency combined with throughput, which equals net output. The ideal storage platform for AI is a file storage platform that can handle file concurrency in the millions, is GPUDirect enabled, can handle NFS over RDMA while also being easy to manage. You should consider that the Dell Technologies Validated Design for AI or the Dell Technologies Validated Design for MLOps features PowerScale, a platform that is built for scale, is able to scale up to millions of open files for the data processing throughput that AI requires.

AI and machine learning have primarily been democratized with open-source tools, models, and other enabling software to help organizations achieve their data science outcomes. The decisions made about the infrastructure that will host a data science project are as important - if not more important - than the decisions about what tools are used. An infrastructure that has been pre-configured and tested specifically for AI workloads that has the simplicity of management combined with the power to scale has the advantage of faster time to value and the ability to grow with an organization's AI ambitions.

Only Dell Technologies offers an HCI solution for AI and MLOps certified for NVIDIA AI Enterprise[2] that keeps the infrastructure provisioning and management simple while also making it easy for data scientists to utilize resources with shifting processing requirements. Check out Dell Validated Designs for AI and MLOps featuring VxRail, NVIDIA AI Enterprise, and PowerScale to learn more. Much like the card readers of yesterday, modern AI environments are a resource that is in demand and require efficient deployment to be utilized to their fullest.