Evaluating Infrastructure Options for Enterprise AI Development

WHITE PAPER

As enterprise commitment to artificial intelligence (AI) accelerates, organizations are facing a familiar and critical challenge: how and where to build and deploy AI infrastructure. There are many issues to consider, and organizations need to think beyond traditional IT architectures and plan for different types of workloads.

This paper is a blueprint to help organizations evaluate their options and make decisions that achieve the best cost efficiency, scalability, risk management and performance.

AI has rapidly made its way up the priorities list for enterprises across all industries and geographies. And it's not just multinational enterprises committing to AI: Many midsized organizations have made AI an essential part of their near- and long-term digital transformation efforts.

That has resulted in a digital gold rush of spending on AI-related solutions. IDC predicts that global expenditures on AI hardware, software and services will skyrocket from $327.5 billion in 2021 to more than $500 billion by 2024—a five-year compound annual growth rate of 17.5%.[1] IDC predicts that spending on AI-dedicated hardware will grow even faster: By 2024, AI storage will experience 31.8% growth, while AI servers will grow by 26.4% over the same period.

[1] "IDC Forecasts Improved Growth for Global AI Market in 2021," IDC, February 2021

TechTarget | Custom Media

ddn | NVIDIA

Capital expenditures of this magnitude naturally capture the attention and scrutiny of both technical and business decision-makers, all of whom want to make sure their investments are tightly aligned with overall business goals for AI and related technologies. Though many of these enterprise-class AI systems are expected to have a large footprint, organizations often demand the flexibility to start with more modest tests of concept, which can then be scaled up and out as necessary.

At first, many organizations choose to invest in cloud implementations of AI—for agility, for risk management, to avoid or defer capital expenditure, and to learn through experimentation. This can be a good way to start out, with the perception of easy access to cloud for a "start small and grow" mindset.

However, a key concept to consider is data gravity, which is the idea that data sets tend to gravitate toward each other, both increasing the size of the data sets and making them more difficult to move. As a result, there is an increasing cost to move large data sets from the point of creation to the point of use, which creates an escalating cost problem. As organizations mature their AI models, and those models grow in complexity and the data sets expand exponentially, they spend more time, money and energy to transport data from storage to where the compute systems live.

While public cloud deployment may be attractive to organizations because of the perceived cost advantages of cloud at modest scales, there is an inflection point where cloud becomes much more expensive for larger workloads. It is important for organizations to consider how they will manage the costs of data, compute and networking as their AI strategy matures, and to determine how to balance their needs among cloud, hybrid cloud and on-premises deployments at scale.

Starting the journey to enterprise-grade AI Infrastructure

Although the promise and attraction of AI is well appreciated by IT and business decision-makers alike, it's far less clear how organizations should begin their journey to enterprise-class AI.

While organizations undoubtedly have high hopes for their AI use cases, the journey from prototype to large-scale production can be daunting, and building a large-scale system is a significant undertaking. It is also essential to adopt an approach that encourages experimentation and small-scale model development and that permits scaling up as an AI initiative matures into production. It can consume a significant part of an organization's IT budget, and not many IT teams have enough in-house expertise to plan, design, deploy and manage big AI projects.

The fact is that most legacy IT architectures are not innately suited for AI from an infrastructure perspective. AI is not the kind of workload that should be undertaken with a piece-meal approach, stitching together servers from one place, storage from another, networking from yet another and the requisite software and services from elsewhere.

Even modest AI projects require large data sets to fuel AI learning and development, and over time, these projects increasingly demand more compute cycles as data sets continue to grow exponentially. That data needs to be categorized, labeled, analyzed and targeted for applications and use cases that rely on AI and machine learning algorithms. This is true whether an organization wants to deploy AI for on-premises, public cloud, private cloud, or hybrid cloud environments.

Not all organizations are ready to make the commitment to the capital expenditure associated with a dedicated AI infrastructure, and hence it is essential to carefully consider how and when to invest in AI-optimized infrastructure.

Ultimately, having a tested, proven and well-documented reference architecture—as well as a purpose-built physical solution for AI infrastructure—helps keep your options open. As much as organizations may want to go to an all-cloud infrastructure strategy for AI and other requirements, the reality is that an all-cloud approach does not suit all workloads and a more balanced approach is needed.

Key factors to consider in AI systems design

For organizations looking to make an efficient and successful journey to enterprise-class AI, starting with an AI-at-scale reference architecture can identify and prevent potential challenges during that journey, including cost, scale, integration, testing and risk mitigation.

Factors decision-makers should consider include:

- Data gravity, which drives colocation of compute, storage and data sets to support model training and learning.
- Optimized hardware and software for fast and reliable access to very large data sets.
- Tight component integration—compute, storage and networking—from the factory, with testing and validation already built in.
- Support for global deployment at any scale, from a single system to massive, interconnected systems.
- Parallel data delivery architecture for maximum redundancy, automatic failover, resiliency and data availability in the event of unexpected service interruptions.
- Networking that supports multiple protocols, including Ethernet and InfiniBand, with multiple networking interfaces on the same system to speed data transfer.
- Support for virtual and containerized applications to simplify deployment and ensure a consistent user experience.
- Multitenancy with data segregation via container-based restricted access.
- User-transparent automatic migration of files between flash storage and hard disk drive storage.
- A widely supported software stack for GPU acceleration, containers and an open source operating system, all managed from a single pane of glass.
- Validated infrastructure designs, which are purpose-built for AI and validated in similar deployments, and can be implemented more quickly than a custom design.

How NVIDIA and DDN team up to deliver integrated, optimized solutions

Organizations looking for a turnkey, pretested and preconfigured AI infrastructure platform should consider the DDN A^{3}I solution with NVIDIA DGX^{TM} A100 systems. This solution features eight NVIDIA A100 GPUs and two 2nd generation AMD EPYC^{TM} processors. The reference architecture (available at https://www.ddn.com/pod-ra) also supports the integration of NVIDIA network switches for a fully integrated infrastructure solution.

The technical collaboration between NVIDIA and DDN has resulted in a very-high-performance platform that is ideal for the scale-up and scale-out requirements of AI infrastructure for use cases such as HPC, analytics, deep learning and machine learning. For instance, the solution enables all phases of deep learning workflows, including inference, validation, curation, data ingest and simulation, to operate concurrently and nonstop.

Multiple NVIDIA DGX A100 systems access data simultaneously via an optimized, unified interface that is easily accessed from containerized applications. The solution also optimizes performance by supporting the use of a high-performance, low-latency network using HDR and EDR InfiniBand or 100 GbE.

While many initial AI use cases may start small, organizations are striving to deploy high-performance AI at scale. This means they are looking for the kind of compute performance, storage throughput and scalable capacity that are native to the DDN A^{3}I solution with DGX A100 systems. For example, as a benchmark of AI performance, independent testing indicates that integrating DDN AI200X storage appliances with the NVIDIA DGX A100 system nearly quadruples the number of images processed using PyTorch (https://pytorch.org/), the GPU-accelerated computational framework.

One of the major benefits of the NVIDIA and DDN solution is simplified and accelerated deployment, cutting down implementation and integration time from multiple months to a few weeks. This dramatically improves time to value, increases return on investment and facilitates faster learning and deeper insights, which can be applied in multiple workflows and processes throughout the organization.

The result: Organizations can quickly, efficiently and reliably build a high performance AI infrastructure in diverse computing environments, including multi cloud and hybrid cloud architectures.

Conclusion

As more organizations strive to leverage the power and utility of AI, decisions about the underlying compute and storage infrastructure become paramount. Traditional IT infrastructure options are most likely not optimized for the unique demands of AI workloads. Cloud can complement AI infrastructure, but the right approach will likely require some investment in dedicated on-prem resources placed in close proximity to the data that fuels AI development. Important issues such as performance, scalability, reliability, ease of deployment, security and integration must be part of the evaluation and selection process for the optimal AI infrastructure.

NVIDIA and DDN have collaborated to build a tightly integrated, tested and validated solution to accelerate the implementation of large-scale AI projects for a wide range of use cases, including advanced analytics and high-performance computing. The DDN A^3I solution with NVIDIA DGX A100 systems offers an end-to-end approach to increase organizational agility in a wide range of compute architectures.

For more information, please visit **www.ddn.com/a3i** and **www.nvidia.com/dgx**.

The joint reference architecture is available at **www.ddn.com/pod-ra**.