A Guide to Enterprise AI Infrastructure: Accelerating Workflows with AI Data Centers

Mike Matchett | Principal Analyst

Small World Big Data

In partnership with

truthinIT

Executive Summary

We've all seen recent tech news about how emerging artificial intelligence (AI) is powering brand new end-user use cases, upsetting markets and creating exceptional new enterprise opportunities. Boosted by the popularity of applications like ChatGPT, generative AI is having a moment. We think every organization today is (or should be) seriously evaluating new AI-driven applications.

Unfortunately, the traditional infrastructure found in most enterprise data centers can't support deep AI processing at the speed and scale required to quickly take advantage. Cloud-hosting AI promises quick adoption, but for AI the usual fragmented line-of-business (LOB) cloud approach is by definition fragmented and with inherent performance limitations, which quickly becomes costly and limiting. But by consolidating AI efforts from across the enterprise into an AI center of excellence, organizations can reduce overall cost while accelerating all AI initiatives.

To meet the enterprise need for performant AI infrastructure, NVIDIA and DDN have collaborated to offer an easily consumable enterprise data center solution for competitive organizations with ambitious AI initiatives, featuring **NVIDIA DGX SuperPOD and DDN A^{3}I Storage**. The DGX SuperPOD integrates industry-leading compute, networking and AI software with a high-performance parallel file system to create a cost-effective and quick-to-deploy AI platform suitable for the modern, data-intensive, enterprise data center.

Cloud-hosting AI promises quick adoption, but for AI the usual fragmented line-of-business (LOB) cloud approach is by definition fragmented and with inherent performance limitations, which quickly becomes costly and limiting.

The Enterprise Opportunity with Artificial Intelligence

The hottest new enterprise applications, those offering notable differentiation and greatest business opportunity today, are all incorporating AI based on compute-intensive machine learning (ML) algorithms operating over significantly large data sets. The infrastructure required to support and deliver modern AI at large scale might require thousands of compute nodes and HPC-class storage to create, yet the opportunity is tremendous.

For example, everyone is talking about (and most of us have probably toyed with) recently publicly available interfaces powered by a class of neural network solutions called generative AI. This **new generation of AI-powered technology** can quickly produce professional-grade visual art from simple text prompts (e.g., Dall-E, Stable Diffusion, Crayon.ai), deeply personalize search tasks, and even author on-demand code, researched thesis or story-telling fiction (e.g., NLP-based ChatGPT, BERT). The race is on to see how and how fast enterprises can leverage these recent AI capabilities as well as develop a next generation of AI.

Of course, AI/ML applications are not new to the enterprise. Within the last 10 years, specifically designed and trained but conceptually "narrower" intelligent algorithms have been incorporated into almost every internal data analytics, IT management and business automation solution within the enterprise. But these latest larger-model AI developments have caught the broader public's attention, inspiring new use cases ripe with great business opportunity and quickly resetting expectations for what comprises a great user experience.

The successful integration and deployment of deeper AI-powered features into external- facing enterprise applications is fast becoming a critical enterprise initiative. Modern AI is enabling directly valuable, exceptionally intelligent real-time interaction with clients and customers. Large-model AI is poised to upset multiple established markets while offering amazing opportunities to AI-savvy organizations.

Despite the potential of AI to change enterprise destiny, most enterprises have struggled to support intensive AI processing requirements at larger scales and faster speeds. Legacy data center architecture simply isn't built to support large-scale AI processing that looks more like an HPC supercomputer application than a corporate database or big data warehouse.

Emerging Large-Model AI Use Cases

In recent discussions with both vendors and AI-oriented enterprises, we've noted a few interesting use cases based on recent practical applications of larger-model AI:

Medical/Healthcare Record Interpretation
Large language models (LLM) formed from Natural Language Processing (NLP) are being used to interpret, collate, translate, and codify decades of older handwritten doctors' notes to form robust electronic medical records (EMR)

Personalized Education
Generative AI in various forms are being applied to create on-demand educational materials personalized to the individual and task. Some approaches intend to provide insight through augmented reality (AR) overlays for real-time operational guidance

Personalized Entertainment
Generative AI is being applied to create uniquely personalized story-based entertainment in the form of custom generated books, podcasts and movies

Enhanced User Engagement
From smarter chatbots to dynamically morphing websites, online user engagement is being vastly enriched with large AI models applied to predictive, intelligent interaction

3D Design
Large models are able to automatically construct workable 3D digital designs from 2D input (i.e., photographs or drawings)

Internal Process Automation
While most of us are noticing external facing AI through clever new user applications, a great deal of progress is being made internally optimizing complex processes in fields like manufacturing and finance

Exploring the Best AI Infrastructure for Enterprises

We'll look a bit deeper into AI workload requirements below, but at a high level, training deep AI models and their scalable execution can require accelerated compute clusters that might range from hundreds to thousands of GPUs. Purpose-built AI processing nodes also leverage copious amounts of memory—and especially GPUs with large memory capacity—to iteratively crunch through high volumes of data. The supporting data storage and high-speed networking required to feed a large cluster of AI nodes must provide high-performance parallel access to PBs of data. This level of serious investment in AI infrastructure should optimally be sharable between multiple initiatives and projects (e.g., data science teams and application development across lines of business).

Of course, there are many cloud-hosted options for AI infrastructure as-a-service. Cloud services can work well for research projects, burst training requirements and small AI model execution. But there are critical issues with off-premises processing of any scaled-up core AI initiatives, especially those serving as the new center of enterprise differentiation. Issues can include:

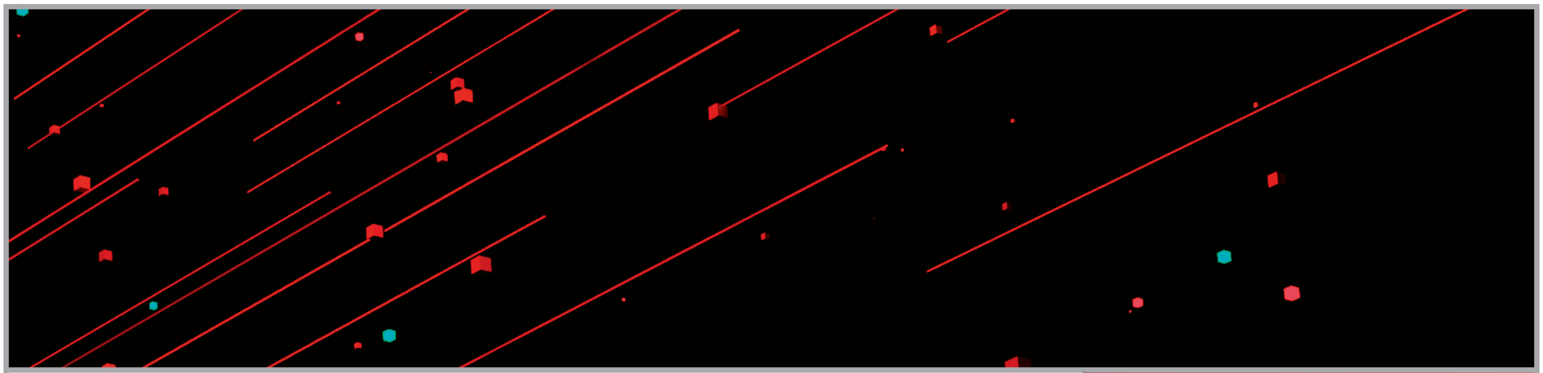
- Significant subscription costs
- Substantial data transit costs
- Sub-optimal resource sharing
- Configuration mismanagement
- Introduced latencies
- Remote data access challenges

We've also seen total cloud costs spiral in enterprises where every R&D team and LOB each instantiate their own splintered AI cloud subscription. With a fractured approach to AI, deeper cost and complexity issues quickly arise from:

- Lost, idled, and isolated resources
- Limited cloud service features
- Poor inter-organizational collaboration
- Lack of performance SLAs
- Inherent data silos, data proliferation and governance gaps
- Increased security vulnerabilities
- Proliferation of technical approaches, tooling and required IT skills

To help organizations identify compromises and cost analysis between cloud and on-premises solutions, DDN and NVIDIA have developed a **total cost of ownership calculator** to help them understand what 3-year and 5-year spending looks like for different implementation pathways.

It's clear that game-changing AI is coming (or here!), and the right organizational buy-in and AI infrastructure matters if you want to lead your competitors. But if the public cloud is not a great core AI option, can the enterprise data center step up? The many reasons why data centers were first organized and still offer value to today's critical enterprise AI applications—resource maximization, cost efficiency, centralized data and data management, best practices collaboration, enhanced security and governance, standardization, and skills leverage. In fact, today's data centers are already essentially evolving into full private clouds, offering internal shared infrastructure, utility-style computing, applications and services.



But when it comes to large-scale AI, is there an effective and efficient AI infrastructure suitable for an enterprise data center? Can enterprise IT realistically implement competitive AI infrastructure at cloud-like scale? Let's first examine evolving AI workload requirements in a little more detail.

Data Center Requirements for AI Workloads

To be clear, a lot of practical AI implementations have modest IT infrastructure requirements that can be serviced out of the more traditional enterprise data center. But at the cutting edge of AI today we need to consider AI initiatives that require high-performance parallel file storage serving expanding clusters of high-end, GPU-dense compute nodes. Generally, these kinds of resources are found in specialized HPC labs more than enterprise data centers. Something new is needed for the enterprise.

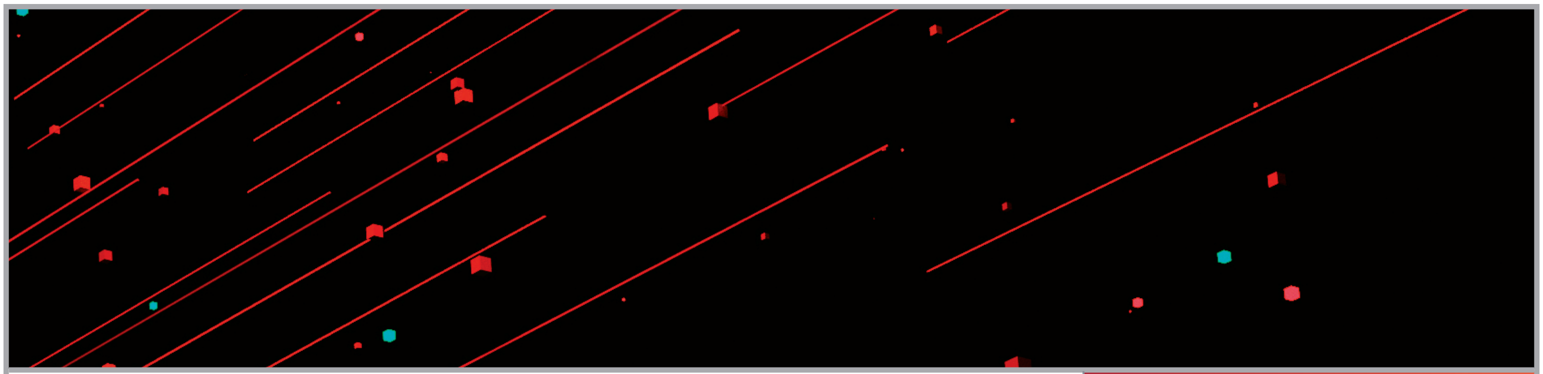
Let's start with storage. An AI data storage workload can be vastly different from an enterprise database and even big data or BI-analytical workload. A traditional enterprise database might best be served by a large SAN filled with highly structured and indexed data tables. An unstructured big data warehouse relies on partitioning data into cheap local storage across a cluster of rather plain server nodes. But AI workloads need to train deep learning models by driving PBs of data in parallel through multiple world-class GPUs in every cluster node. This data pipeline is similar to that of an HPC supercomputer in many respects, and requires similar HPC-class storage services not commonly found in existing enterprises (or natively in many clouds).

When it comes to server nodes, efficient AI computing relies heavily on GPU acceleration (i.e., advanced neural net algorithms require massively iterated floating point processing). Compute clusters for AI need to be rich in serious GPUs, and those high-end GPUs need to be optimally configured with fast, low latency network IO and IO acceleration software to be fully utilized, as collectively they represent a significant investment.

As AI clusters scale up, compute node density and overall power consumption also become important design concerns. The fast networking between compute nodes and between nodes and storage needs to accommodate growth without bottlenecking anywhere. Unfortunately, traditional server clusters (e.g., those designed for databases, virtualization or other enterprise applications) and legacy enterprise storage are simply not designed to scale significantly in these important dimensions.

As AI clusters scale up, compute node density and overall power consumption also become important design concerns. The fast networking between compute nodes and between nodes and storage needs to accommodate growth without bottlenecking anywhere.

In fact, quick scalability without requiring re-architecting is itself a key AI infrastructure design requirement. AI workloads can grow very quickly, especially when new AI-powered applications prove successful. Also keep in mind that AI data storage might need to scale significantly and independently of the compute cluster to meet sudden new demands.



Building an AI Center of Excellence

Centralizing high-performance AI infrastructure in an enterprise data center is a smart architectural approach to help meet all the AI workload challenges listed earlier. Core AI infrastructure can be a big investment, but consolidating multiple disparate AI efforts from across the organization can ultimately save big money. Anecdotally, we've heard about an enterprise that recently reduced their total AI cost outlay by 6x by repatriating siloed cloud deployments into one super-powered AI data center.

But probably the most valuable reason to bring AI infrastructure and supporting capabilities together is that it helps create an enterprise AI center of excellence. Through collaboration, centralization, sharing of resources, leveraging common tooling and management, and making data commonly accessible, an organization's AI initiatives are not only more likely to succeed and succeed faster, but also propagate deeper value back through the whole organization.

The pursuit of AI should really not be about minimizing cost, but rather maximizing business value, today and tomorrow. The realization of a game-changing AI vision can return orders of magnitude more value than an up-front AI data center investment. And next-generation AI initiatives will be built upon today's AI efforts. A center of excellence provides a strong foundation with which to launch future initiatives. Fostering a world-class AI center of excellence ensures that you are making the best AI effort possible.

Simplify AI Infrastructure with NVIDIA DGX SuperPOD and DDN A³I Storage

Building your own AI data center infrastructure from disparate vendor components can be a tough haul. Many organizations have floundered when attempting to deploy and then integrate several new-to-their-datacenter technologies like GPU-dense compute clusters, high-bandwidth networks and parallel file storage. Fortunately, the **NVIDIA and DDN collaboration** has made world-class AI infrastructure readily available to enterprise IT with a **fully-validated, ready-to-run** solution in DGX SuperPOD and DDN A^{3}I Storage. The pre-integrated design takes all the risk out of building out core AI infrastructure while supporting the largest of AI projects.

NVIDIA itself was originally faced internally with many of the same AI support challenges that their customers dealt with (see sidebar for more details). They solved the many difficult integration challenges and optimized the utilization of powerful GPUs within and between NVIDIA DGX systems.They then extended this into full AI solutions by "blueprinting" their incredibly successful configurations, storage, networking, and best practices. Now they offer these solutions as scalable **DGX BasePOD** and **DGX SuperPOD** AI infrastructure externally to enterprise data centers.

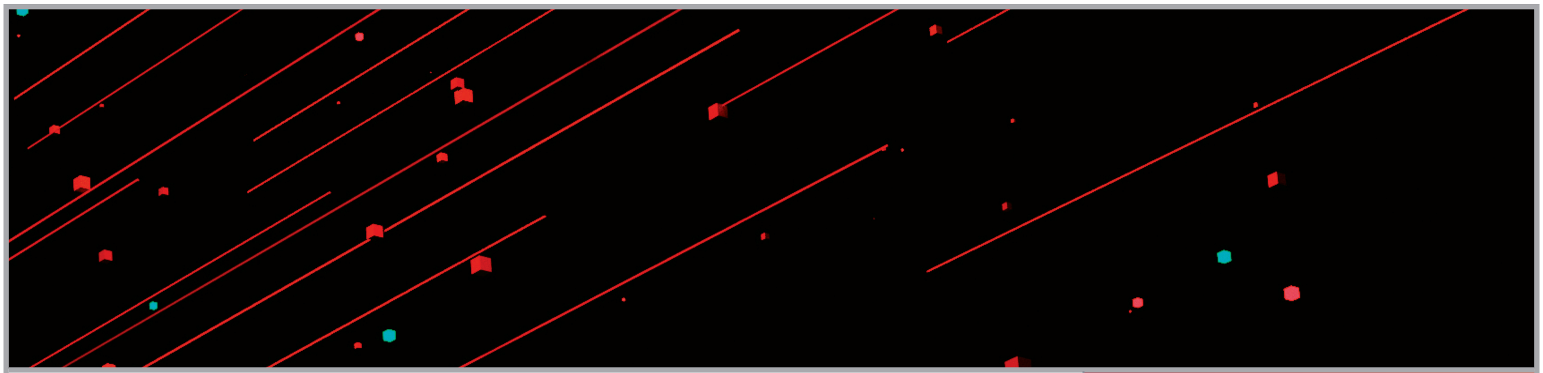
DGX SuperPOD with Two NVIDIA DGX nodes and DDN A^{3}I storage

Inside NVIDIA DGX SuperPOD with DDN A³I Storage

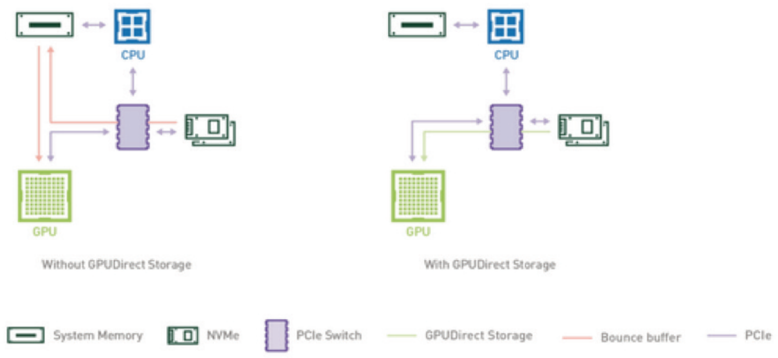
A DGX SuperPOD is essentially a full-stack data center AI platform that includes compute, storage, networking, core AI software, infrastructure management, and a white glove implementation service. The NVIDIA DGX SuperPOD with DDN A3I Storage tightly integrates and optimizes a cluster of NVIDIA GPU-powered DGX systems on NVIDIA Quantum-2 InfiniBand networking with DDN AI400X2 high-performance parallel file storage, including all the best practices and configurations needed to guarantee large AI workload SLAs out of the box.

The latest versions of DGX SuperPOD feature:

- **NVIDIA DGX systems**
  Based on either the DGX A100 or the latest DGX H100.
  Each system (node) includes for example:
  - 8 NVIDIA DGX H100 Tensor Core GPUs with 640GB total GPU memory per node
  - 10 NVIDIA ConnectX-7 400Gbps interfaces – 1 TB/s peak bi-directional network bandwidth over InfiniBand
  - Dual Intel Xeon Platinum 8480C processors
  - 2TB system memory
  - 30TB local NVMe SSD



- **DDN A^3I AI400X2 turnkey storage appliances**
  Nominally one 500TB node per every 4 DGX systems, but scalable in ratios and capacities:
  - Each appliance delivers up to 90GB/s and 3M IOPS
  - Based on DDN EXAScaler parallel shared file system,configured as all-NVMe in entry-level systems
  - Integrated with NVIDIA MagnumIO™ GPUDirect Storage™(GDS) providing direct DMA data paths between storage and GPU memory. GDS can push 57% more IO throughput to achieve GPU saturation while freeing up over 75% CPU usage for actual compute tasks



Optimization of GPU IO with NVIDIA MagnumIO GPUDirect Storage

DGX SuperPOD for Enterprise AI

Numerous organizations are already leveraging DGX SuperPOD as their in-house AI platform. Due to the reliance of advanced AI on the largest, fastest GPU accelerated systems available, it would be difficult to create a more performant AI platform than the one NVIDIA first built for itself. DGX SuperPOD has a proven rapid deployment achieved by including all the hardware, AI software, management, integrations, and implementation services necessary to implement a truly world-class AI environment. We believe that adopting DGX SuperPODs could be the absolute fastest way to create an enterprise AI Center of Excellence.

In addition, we've noted that initial deployments of DGX SuperPOD accelerate AI programs, often soon requiring (and justifying) cluster growth. The well-defined DGX SuperPOD architecture has proven to scale easily to meet the data capacity and computational growth needs of the largest AI implementations without having to rebuild or redesign at any point. Once a DGX SuperPOD is in the production data center, it also enables enterprise IT to deliver AI infrastructure internally as-a-service—another win for IT on any cloud transition/adoption initiatives.

NVIDIA is DGX SuperPOD Customer Number One

Here is a quote about the origin of DGX SuperPOD from our Small World Big Data analyst event with **Kurt Kuckein**, VP of Marketing with DDN and **Tony Paikeday**, Senior Director of AI Systems at NVIDIA:

> **Tony** . . . A lot of what goes into SuperPOD, the current generation and future iterations of it are the direct result of what we experience and learn on our internal cluster that's been growing over time, now eclipsing thousands of systems all interconnected over a high-performance network fabric. And we obviously use DDN storage. This is something day in and day out that thousands of developers around the globe are using. That's why we're very excited about the SuperPOD offer and how enterprises can use it. A lot of organizations probably don't realize that NVIDIA is actually a fairly large enterprise in terms of how we consume IT and build AI infrastructure to support our own IT and the constituents in our business units and product teams that need to use it. So we like to take all of that...and make that available in partnership with DDN. That's goodness for others who don't have to take a protracted journey to get to their center of excellence.

By developing a centralized AI service that centralizes the underlying data powering all of an enterprise's AI initiatives, we've seen enterprise IT increase their total data-value harvesting, ensure critical data integrity and enable far better data governance. Bringing together disparate AI programs scattered across LOBs with each using their own cloud subscriptions or siloed infrastructure has shown to not only save cost and accelerate AI projects by sharing world-class storage and processing resources, but also foster a collaborative organizational culture primed for ongoing AI success.

Small World Big Data Opinion

We believe that one of the first keys to success with enterprise AI adoption programs is super-powering today while future–proofing for tomorrow. Both AI model development and ongoing production "learning" can easily stretch out on insufficient or inferior infrastructure to the point where projects can stall or even flame out. If an enterprise has a strong AI vision and truly desires to dominate their market, then we see NVIDIA DGX SuperPOD with **DDN A^{3}I Storage** platform as their number one world–class option.

It's a given that advanced AI capabilities are opening up more business opportunities and are emerging faster than ever before—now in timeframes measured in weeks instead of months or years. We are noting the rapid acceleration of practical, game-changing AI applications that are well within reach of even smaller organizations, especially if they have deployed world-class AI infrastructure in their own AI center of excellence.

While not all AI/ML projects require supercomputing class infrastructure, with NVIDIA DGX SuperPOD and DDN A^{3}I Storage many more AI possibilities come within reach of historically traditional commercial enterprises. Leveraging a turn-key DGX SuperPOD, businesses can rapidly transform key aspects of their business today through the adroit application of deep learning, large language models, real-time customer engagement, dynamic content creation, and more. Here at Small World Big Data, we believe that the IT world is fast approaching an inflection point that will widely separate those that can rapidly leverage cutting-edge AI from those that will suddenly be left far behind.

To learn more, CLICK TO WATCH an on-demand webcast with an in-depth discussion on this opportunity.

About Small World Big Data

Small World Big Data creates insightful technology analysis and research for IT, Cloud and Data markets. In addition to incisive market research and opinionated analyst reports, we specialize in producing popular, consumable, expert analyst-hosted video content and podcasts. As a small agile firm with deep experience in IT and vendor marketing, our constantly refreshing research calendar provides ongoing and cost-effective opportunities for market awareness, education, demand generation, lead nurturing, and more. For schedule and information about current research and advisory services, visit SmallWorldBigData.com.

About Truth in IT

Truth in IT has published independently created informational and educational content for the IT professional since 2009. We partner with independent analysts and bloggers to create unique seminars, webinars, videos, and market research reports. We also invite sponsors to further help educate our audience about best practices and trends to help the IT professional succeed in their daily jobs and careers. Our goal is to amplify the subject matter expert's expertise, insight, and voice so our audience is able to cut through the hype to get to the Truth in IT.

Notice: The information contained herein has been obtained from multiple sources believed to be accurate and reliable, and includes personal opinions that are subject to change without notice. Small World Big Data disclaims all warranties as to the accuracy of such information and asumes no responsibility or liability for errors or for your use of, or reliance upon, such information. Company, brand, and product names referenced herein may be trademarks or their respective owners.