Maximizing ROI on Your AI Infrastructure

For GenAI and LLMs At Scale

A Small World Big Data Solution Profile

Mike Matchett, Principal Analyst

September 2024

Small World Big Data

In partnership with:

Truth in IT

Maximizing ROI on Your AI Infrastructure

Executive Summary

Generative AI is the hottest topic in every corporate strategy meeting today - evaluating the market opportunities, outlining new business models, and looking for the right entry points and applications. With Large Language Models (LLMs) leading the way, the new AI race has started for enterprises in every market, and competition for greatest advantage is quickly ramping up. From our analyst perspective, a fundamental key to maximizing AI returns, especially at large scale, is deploying on a fully engineered AI infrastructure designed throughout for top performance, workload efficiency and operational reliability.

We recently interviewed HPC-scale AI experts from NVIDIA and DDN, collaborators in creating the world's largest and best known high-end AI systems using NVIDIA GPUs, NVIDIA BlueField® DPUs and DDN A^{3}I Storage to:

- Find out about the latest large-scale AI trends and developments
- Understand the practical impact and challenges on AI infrastructure and IT operations
- Learn how newer solutions are helping maximize AI ROI

In this brief summary, we'll also review one of the year's biggest generative AI advancements – Retrieval Augmented Generation (or RAG) along with the growing shift in AI training data sets from just textual documents to multi-modal input – and an accompanying need to develop newer models from widely distributed data while enforcing compliance requirements. These developments complicate the already demanding data storage and model processing workloads as well as raise new IT and data management concerns. The good news is that NVIDIA and DDN are helping large-scale AI adopters surmount these emerging challenges with a combination of NVIDIA BlueField DPUs and the distributed high-performance DDN Infinia Data Intelligence Platform.

"…key to maximizing AI returns…is deploying on a fully engineered AI infrastructure designed throughout for top performance, workload efficiency and operational reliability"

Overcoming AI Infrastructure Challenges

According to our analysis, the rapid adoption of large-scale AI is served well at its core with NVIDIA GPU HPC-class clusters integrated with DDN A^{3}I Storage using the world-class DDN EXAScaler parallel file system. But as the broader AI workload continues to evolve, adopters also now need additional AI infrastructure that can provide:

- Efficient access to *distributed* data sources to feed growing AI training needs
- Massive amounts of metadata to effectively and efficiently manage data locality and flows
- Increased security and data protection edge-to-edge from where the data is sourced to where it is consumed
- Larger, faster data pipelines to serve a wider and growing variety of data including images and video
- Assurance of critical AI-dependent application service levels over time

Figure 1 An NVIDIA DGX SuperPOD™ with NVIDIA DGX nodes and DDN A^{3}I storage

Today's generative AI models, already large in size and scope, are growing still larger to encompass bigger and "faster" (i.e. more real-time) training data sources. And new applications based on next generation models are consuming greater amounts of multi-modal data. For example, today we can readily find large-scale AI training on much more than text or simple image files, such as audio, multi-media/video, radar/sonar, and other more complex signals. Current trends point to overall AI data ingestion requirements quickly growing by two or more orders of magnitude to incorporate increasingly complex data.

We are also seeing a fast growing percentage of enterprise applications embedding customized AI models, multiple AI models leveraged per application, and AI processing at distributed locations in an increasingly hybrid/global IT environment. The current trend towards more complex AI data flows requires greater end-to-end data path performance and efficiency in addition to a growing core AI processing capacity.

Cost and Performance of Scaling AI Models

At scale, the cost and performance of AI implementation can quickly become a big concern. As we've seen, generative AI models are growing larger in multiple dimensions, which in turn requires more powerful accelerated computing. NVIDIA continues to innovate and deliver newer generations of GPUs and GPU architectures (e.g. NVIDIA Blackwell), but to maximize ROI, core GPU assets should be kept fully utilized. This means not only increasing data flows from wherever sources are located to feed processing and training clusters, but also optimizing end-to-end reliability over increasingly longer training timeframes. High performance storage at the core has always been critical, but now the whole larger surrounding data context needs to become highly robust and performant.

As the AI workload expands across the organization, it increases the requirements for supporting systems management. As training data is sourced from a wider distribution of data sources, security, data protection, and compliance concerns (e.g. geo-locality, IP protection) must also be addressed at the same wider scale. If poorly handled, this can greatly increase the cost and complexity of management and operations and possibly expose AI adopters to significant business risk.

Evolving AI Data Storage and Management Needs

AI data storage requirements are fast changing too. While core AI infrastructure is well serviced by integrating an HPC-class parallel file system like DDN EXAScaler, we are seeing the inclusion of new AI data sources stretching out across the wider distributed organization. Storage systems supporting high performance GPU access (e.g. leveraging NVIDIA GPUDirect®) to all training data wherever it lives will be paramount for processing tomorrow's trillion+ parameter models. In other words, the performance of the broader, distributed and often cloud-like surrounding storage architecture is becoming as critical to AI effectiveness as the core cluster storage.

As generative AI continues to grow in size and use case, AI storage concerns now also include:

- Avoiding distributed data preparation and ingestion bottlenecks for larger data sets
- Feeding "multi-epoch" training runs for larger models at high efficiency (e.g. by minimizing any latencies in the storage data path)
- Supporting faster checkpointing and checkpoint reloading during "months"-long training runs
- Efficiently managing and distributing larger models into and out of the core clusters
- Integrating with distributed cloud-like object storage (API-based versus file and file systems)
- Accelerating distributed object storage performance

Enhancing Model Accuracy with Retrieval Augmented Generation

When given an input prompt, LLMs predict the next best text results given recent context and what they've been trained on. This makes LLMs very good at generating answers that appear sensical, but LLMs aren't inherently designed to provide authoritative answers, reason or rationalize facts, or process structured database (e.g. SQL) queries. In response to demand for greater accuracy and fidelity, we see emerging solutions exploring a variety of layered, chained or nested mechanisms that can increasingly answer a wider variety of prompts with more context and relevance.

One of the new LLM methods quickly becoming popular is Retrieval Augmented Generation (RAG). In a RAG implementation, incoming user prompts are enriched with specific and timely external data before being sent on to the pre-trained and relatively static LLM. RAG implementations start by indexing quickly searchable vectors (i.e. "vector embedding") processed from any desired external data sources including domain-relevant document repositories, structured database queries, unstructured data lake searches, business applications and remote API function calls. RAG then enriches submitted queries dynamically by analyzing each prompt and mixing in contextually relevant embedded data. By incorporating detailed, factual and current input into queries, broadly trained LLMs return more useful, relevant and accurate responses.

It's important to note that RAG query augmentation occurs at AI model usage time (i.e. during operational generation, not during offline training), and can introduce overhead and latencies directly impacting the AI user experience. A critical AI application might now come to depend on (and be delayed by) RAG drawn from multiple petabyte-sized data lakes. Large-scale "on-demand" RAG will also increase the computational demands on surrounding and supporting applications and data storage.

Today's RAG is based on computing vector embedding indices over data sources similar to how internal LLM data might be processed and is still somewhat batch-oriented. Any continuous RAG indexing of updating data streams will require additional allocations of on-demand resources. We expect that overall IT demand for accelerated computing will only continue to grow as AI architectures develop more complex webs of model and enhancement hierarchy and interaction. In order to optimize complex and dynamic AI pipelines in both training and production evaluation, we predict every GPU will come to need its own individualized and fine-grained dynamic data access to almost "unbounded" amounts of stored data (i.e. data streams or sets that will be simply too large to fit in even HPC amounts of memory).

Evolving AI Infrastructure With NVIDIA and DDN Solutions

We want to mention two key solutions in this report that taken together can address new challenges with large-scale AI:

- NVIDIA BlueField DPUs
- DDN Infinia Data Intelligence Platform

NVIDIA BlueField DPU

A Data Processing Unit (DPU) is a specific optimizing network interface card that, like a GPU conceptually, offloads a great deal of work from a node's CPU and also provides specialized processing to accelerate key data-intensive tasks. While a GPU takes on a large amount of mathematical calculations over huge matrices of numbers

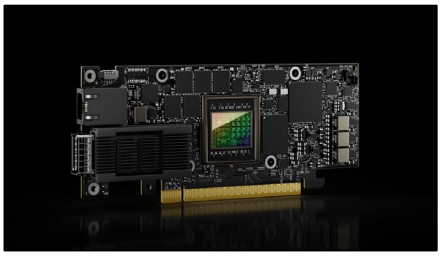

Figure 2 NVIDIA BlueField® DPU

for graphics processing or AI computing, a DPU can take on network optimization, security and data movement tasks

NVIDIA BlueField DPUs can be deployed in the nodes across an AI compute cluster to achieve a number of goals. They can:

- Provide a secure, zero trust network connectivity that is very resistant to being "hacked" even by malicious code that might be running in the host
- Offload network protocol overhead and accelerate network performance
- Internally run client-side network-facing code such as encryption, distributed security, storage and data transformation, also freeing up CPU and other system resources
- Host remote and distributed storage and data management micro-services essential to creating an accelerated intelligent data path to and from each node

There are other advantages, but this last point is perhaps the most significant for updating our AI infrastructure as it ties in to the next solution.

DDN Infinia Data Intelligence Platform

As we've reviewed, for many reasons intensive AI data and data workflows are extending from the core AI data center out to distributed locations including the cloud and remote edges. The latest advancements in AI are requiring high performance access to large distributed data sets. In order to extend high performance storage throughout an organization, it would be ideal to create it with some kind of HPC-class object-paradigm storage. And in fact, DDN has recently rolled out DDN Infinia which delivers exactly that. Infinia is a hugely scalable (e.g. high capacity) distributed object store which has been designed to work as simple cloud storage with common REST API's and autonomous operation, but is also capable of delivering *primary AI storage workload* performance.

Key to the scalability, security and performance of Infinia is the use of NVIDIA DPUs to host various Infinia storage microservices directly in storage client nodes. By running key parts of the storage system within the distributed clients, storage access is not only greatly parallelized, but also made vastly secure. Because it all runs offloaded on DPUs, each node can also accomplish more work.

The distributed bits of Infinia then help manage issues of data locality, data security, and performance. The storage system can locate, move and cache data hyper efficiently and leverage client-side RDMA acceleration. Data can be directly transferred as storage clients connect intelligently with data repositories, while security is assured through the consistent zero trust architecture. Overall DDN Infinia plus NVIDIA BlueField DPUs work together to massively reduce data path latency and ensure security over far-ranging distributed data.

While not specific to AI evolution, Infinia also presents a massively scalable metadata opportunity. Most object storage solutions offer some object metadata functionality, but Infinia raises object store metadata leverage to another level entirely. In addition to an awareness of data locality for performance, the storage is essentially physically self-aware of its data to meet resiliency goals (and more advanced security and compliance requirements).

Small World Big Data Opinion

It's clear that the pace of AI development is in high gear, and that most organizations are currently racing to figure out how to best take advantage of generative AI in one form or another. The AI workload is evolving, the data we feed into AI is growing

NVIDIA And DDN Evolving AI Infrastructure At Scale

Recently Small World Big Data's Mike Matchett got an opportunity to sit down with AI and HPC experts James Coomer from DDN and CJ Newburn from NVIDIA. You can watch our in-depth discussion about the evolution of AI workloads, the expected impacts and opportunities of new capabilities for genAI and LLMs like RAG, and how to further maximize your AI ROI.

in distribution, size and "modality", and AI infrastructure continues to develop.

The growing data scale and the time to train bigger models is a challenging aspect for competitive core AI training infrastructure. But better value extraction from source data is fast becoming important too, with methods like RAG placing significant new demands on AI operational hosting in production. And AI compliance regulations are just starting to emerge worldwide. At this point, effectively adopting AI is no longer just about building a single centralized model training facility.

We see AI adopters clamoring for more powerful IT resources like next generation GPUs, DPUs, increased memory, faster networks, and vast amounts of distributed yet performant storage. With AI, it's not just about an exponential growth in data capacity, but also in data "movement" for model training and data transformation and indexing for RAG enhancement. Both of these trends are made even more challenging by needing to distribute security (e.g. access controls) along with the data to assure data protection and compliance.

NVIDIA and DDN have perhaps the best insight and experience with the whole end-to-end AI process at the largest of scales as demonstrated by their numerous world-class AI/HPC clusters. These include some of the world's largest AI centers of excellence at national research labs, global companies, and of course NVIDIA's own Selene, Cambridge-1 and Eos (with 576 NVIDIA HGX H100™ nodes) clusters.

NVIDIA and DDN together are clearly striving to make the whole AI infrastructure optimally efficient. Beyond building large HPC core clusters, AI adopters need to make the best use of network, compute, and storage across a more distributed data estate to improve operational AI access to data as much as possible. Critical resources need to be kept highly utilized, and the biggest current AI infrastructure challenge is now one of moving massive amounts of data to available GPUs. To that end, we think the addition of NVIDIA BlueField DPUs and the DDN Infinia Data Intelligence Platform to AI architectures is not just a nice improvement, but a must-have to stay competitive and, ulitimately, recognize the maximum AI ROI.