Lab Insight

Dell and Broadcom Deliver Scale-Out AI Platform for Industry

Authors:

Randy Kerns

Russ Fellows

November - 2023

Executive Summary

As part of Dell's ongoing efforts to help make industry leading AI workflows available to their clients, this paper outlines a solution example that leverages scale-out hardware and software technologies to deliver a generative AI application.

Over the past decade, the practical applications of artificial intelligence (AI) have increased dramatically. The use of machine learning, AI-ML has become widespread, and more recently the use of AI tools capable of comprehending and generating natural language have grown significantly. Within the context of generative AI, Large Language Models (LLMs) have become increasingly practical due to multiple advances in hardware, software and available tools. This provides companies across a range of industries the ability to deploy customized applications which can help provide significant competitive advantages.

However, there have been issues limiting the broad adoption of LLM's until recently. One of the biggest challenges was the massive investment in time, cost and hardware required to fully train an LLM. Another on-going concern is how firms can protect their sensitive, private-data to ensure information is not leaked via access in public clouds.

As part of Dell's efforts to help firms build flexible AI platforms, Dell together with Broadcom are highlighting a scale-out architecture, built upon Dell and Broadcom equipment that can deliver the benefits of AI tools while still ensuring data governance and privacy for regulatory, legal or competitive reasons.

By starting with pretrained LLM models, and then enhancing or "fine-tuning" the underlying model with additional data, it is possible to customize a solution for a particular use case. This advancement has helped solve two challenges companies previously faced, how to cost effectively train an LLM, and secondly how to utilize private domain information to deliver a relevant solution.

With fine-tuning, GPUs are utilized to produce high quality results within reasonable timeframes. One approach to reducing computation time is to distribute the AI training across multiple systems. While distributed computing has been utilized for decades, often multiple tools are required, along with customization, requiring significant developer expertise.

In this demonstration, Dell and Broadcom worked with Scalers.AI to create a solution that leverages heterogeneous Dell PowerEdge Servers, coupled with Broadcom Ethernet NICs to provide the high-speed inter-node communications required with distributed computing. Each PowerEdge system also contained hardware accelerators, specifically Nvidia GPUs to accelerate LLM training.

Highlights for IT Decision Makers

The distributed training cluster included three Dell PowerEdge servers, using multi-ported Broadcom NICs and multiple GPUs per system. The cluster was connected using a Dell Ethernet switch which enabled access to the training data, residing on a Dell PowerScale NAS system. There are several important aspects of the heterogeneous Dell architecture utilized, which provides an AI platform for fine-tuning and deploying generative AI applications. The key aspects include:

- Dell PowerEdge Sixteenth Gen Servers, with 4th generation CPUs and PCIe Gen 5 connectivity
- Broadcom NetXtreme BCM57508 NICs with up to 200 Gb/s per ethernet port
- Dell PowerScale NAS systems deliver high-speed data to distributed AI workloads
- Dell PowerSwitch Ethernet switches Z line support up to 400 Gb/s connectivity

This solution uses heterogenous PowerEdge servers spanning multiple generations combined with heterogeneous Nvidia GPUs using different form factors. Dell PowerEdge Servers included a Dell XE8545 with four NVIDIA A100 GPU accelerators, a Dell XE9680 with eight Nvidia A100 accelerators and a Dell R760XA with four NVIDIA H100 accelerators. The PE XE9680 acted as the both a Kubernetes head-node and worker-node. Each Dell PowerEdge system also included a Broadcom network interface (NIC) for all internode communications and storage access to the Dell PowerScale NAS system.

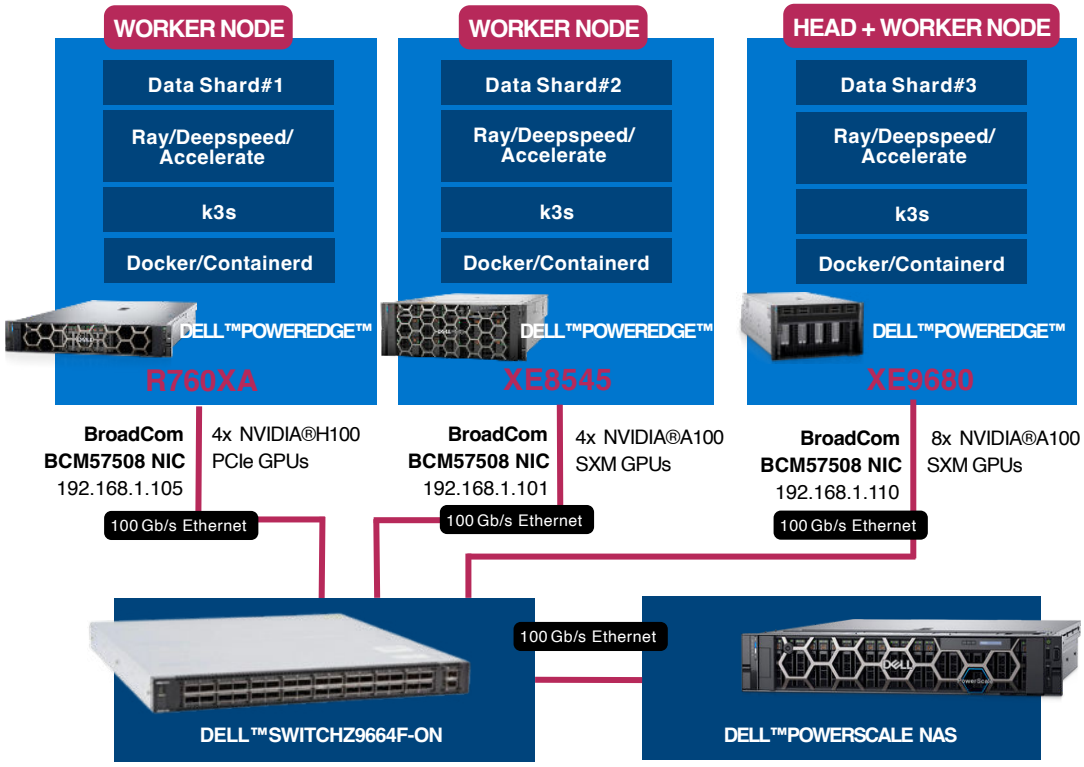


Figure 1: Distributed Scale-Out AI Hardware Platform (Source: Scalers.AI)

***Futurum Group Comment**: The hardware architecture utilized showcases the flexibility of using dissimilar, heterogeneous systems to create a scale-out cluster, connected using cost effective Ethernet, rather than proprietary alternatives. Together, Dell and Broadcom along with AI hardware accelerators provide the foundation for successful AI deployments.*

Broadcom BCM-57508 Ethernet cards are an important aspect of the solution, solving a common bottleneck with distributed systems, the inter-node communications, with both bandwidth and latency as key factors. Broadcom's Peer Direct and GPUDirect RDMA technologies enables data to bypass host CPU and memory, for direct transfer from the network into GPUs and other hardware accelerators. Without these technologies, data is driven by the CPU into local memory and then copied into the accelerator's memory – adding to latency. Broadcom's 57508 NICs allows data to be loaded directly into accelerators from storage and peers, without incurring extra CPU or memory overhead.

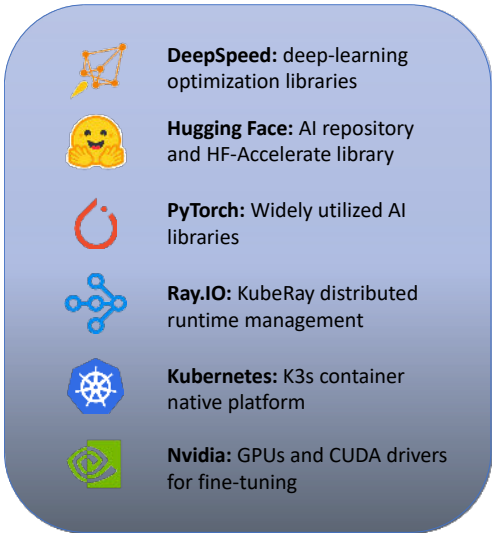
Dell PowerScale NAS storage for unstructured data, used all-flash and RDMA optimized data access to power the low-latency and high-bandwidth demands of AI workflows. PowerScale supports SMB3, NFSv3/v4 along

with S3 object access for the scale-out storage that can meet the needs of AI projects, while maintaining data privacy, and corporate control over your critical data.

Dell PowerSwitch Z-Series core switch line provide connectivity up to 400 Gb/s, with breakout options to support 100 GbE and lower as required. The Z series provides high-density data-center Ethernet switching with a choice of network operating systems for fabric orchestration and management.

Highlights for AI Practitioners

A key aspect of the solution is the software stack that helps provide a platform for AI deployments, enabling scale-out infrastructure to significantly reduce training time. Importantly, this AI Platform as a Service architecture was built using Dell and Broadcom hardware components, coupled with cloud native components to enable containerized software platform with open licensing to reduce deployment friction and reduce cost.



Futurum Group Comment: *The software stack utilized is important for several reasons. First, the support for containerized workloads on Kubernetes is a common industry best practice, along with support for PyTorch, TensorFlow and CUDA, which are widely utilized AI libraries. Finally, the use of the deep learning accelerators and libraries help automate distributed scale-out fine-tuning. Together this AI Platform plays a critical role in the overall solution's success.*

Figure 2: Scale-Out AI Software Platform (Source: The Futurum Group)

The AI platform is based upon K3s Kubernetes, Ray.IO KubeRay, Hugging Face Accelerate, Microsoft DeepSpeed, along with other libraries and drivers including NVIDIA CUDA, PyTorch along with CNCF tools such as Prometheus and Grafana for data collection and visualization. Another key aspect was the use of the Hugging Face repository, which provided the various Llama 2 models that were trained, including the 7b, 13b and 70b models containing 7, 13 and 70 billion parameters respectively.

Additionally, the solution example is being made available through Dell partners on a GitHub repository, which contains the documentation and software tools utilized for this solution. The example provided helps companies quickly deploy a working example from which to begin building their own, customized generative AI solutions.

The distributed AI training setup utilizes the Dell and Broadcom hardware platform outlined previously and is shown in the subsequent steps:

Distributed AI Training Process Overview:

1. Data curation and preparation, including pre-processing as required
2. Load data onto shared NAS storage, ensuring access to each node
3. Deploy the KubeRay framework, leveraging the K3s Flannel virtual network overlay
    a. Note: Larger clusters may utilize partitioned networks with multiple NICs to create sub-nets to reduce inter-node traffic and potential congestion
4. Install and configure the Hugging Face Accelerate distributed graining framework, along with DeepSpeed and other required Python libraries

Generative AI Training Observations

As described previously, the distributed AI solution was developed utilizing a trained, Llama 2 base model. The solution authors, Scalers.AI performed fine-tuning using each of the three base models from the Hugging Face repository, specifically, 7b, 13b and 70b to evaluate the fine-tuning time required.

The training time per epoch is shown in Figure 3, with lower (less time) indicating better results.

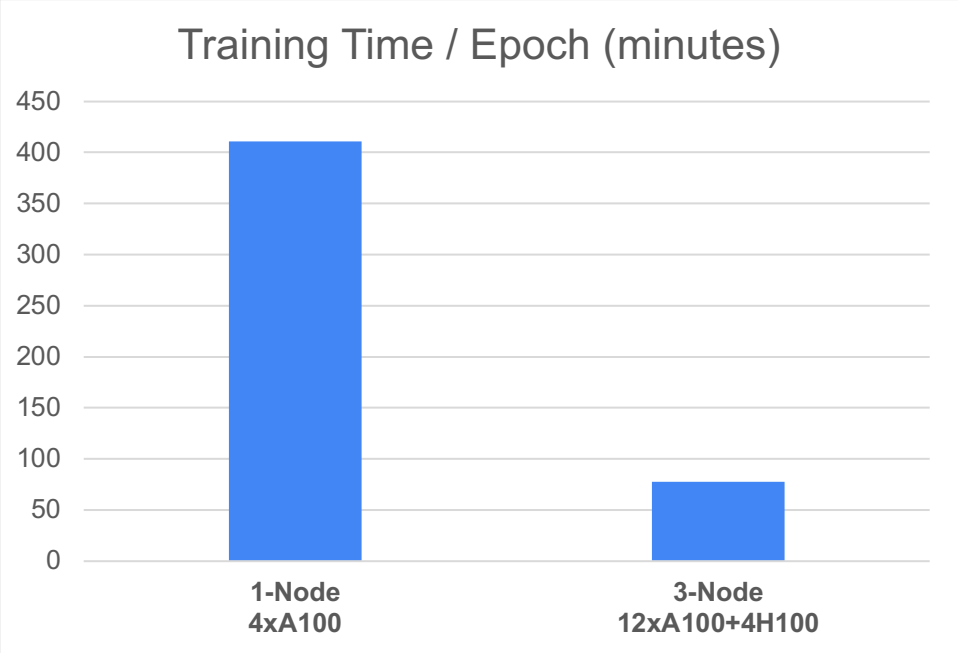


Figure 3: Scale-Out AI Software Platform (Source: The Futurum Group)

Futurum Group Comment: *These results demonstrate the significant improvement benefits of the Dell – Broadcom scale-out cluster. However, specific training times per epoch and total training times are model and data dependent. The performance benefits stated here are shown as examples for the specific hardware, model size and fine-tuning data used*

Fine-tuning occurred over 5 training epochs, using two different hardware configurations. The first utilized a single node and the second configuration used the three-node, scale-out architecture depicted. The training time for the Llama-7b model fell from 120 minutes, to just over 46 minutes, which was 2.6 times faster. For the larger Lama-13b model, training time on a single-node was 411 minutes, while the three-node cluster time was 148 minutes, or 2.7 times faster.

An overview of the scale-out architecture is shown in Figure 4.

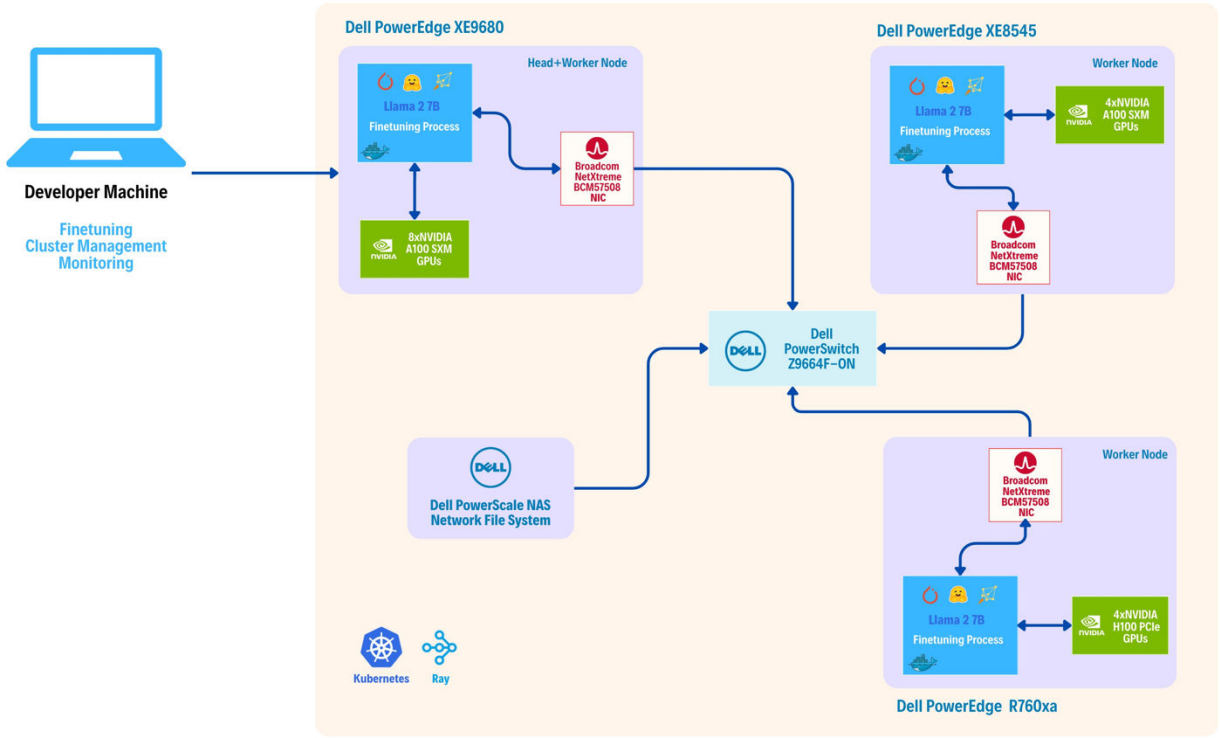


Figure 4: Scale-Out AI platform using Dell & Broadcom (Source: Scalers.AI)

A critical aspect of distributed training is that data is split, or "sharded" with each node processing a subset. After each step the AI model parameters are synchronized, updating model weights with other nodes. This synchronization is when the most significant network bandwidth utilization occurred, with spikes that approached 100 Gb/s. Distributed training, like many HPC workloads is highly dependent upon high-bandwidth and low-latency for synchronization and communication between systems. Additionally, networking is utilized for accessing the shared NFS training data, which enables easily scaling the solution across multiple nodes without moving or copying data.

In order to add domain specific knowledge, an open source "pubmed" data set was used to provide relevant medical understanding and content generation capabilities. This was used to enhance accuracy of medical questions, understanding medical literature, clinician notes and other related medical use cases. In a real-world deployment, it would be expected that an organization would utilize their own, proprietary and confidential medical data for fine-tuning.

Another important aspect of the solution, the ability to utilize private data, is a critical part of why companies are choosing to build and manage their own generative AI workflows using systems and data that they manage and control. Specifically for companies operating in healthcare, they can maintain HIPAA/HITECH compliance, and other regulations around EMR and patient records.

Final Thoughts

Recently, the ability to deploy generative AI based applications has been made possible through the rapid advancement of AI research, hardware capabilities combined with open licensing of critical software components. By combining a pre-trained model with proprietary datasets, organizations are able to solve several challenges that were previously available to only the very largest corporations. Leveraging base models from an open repository removes the significant burden of training large parameter models and the billions in dollars of resources required.

***Futurum Group Comment**: The solution example demonstrated by Dell, Broadcom and Scalers.AI highlights the possibility to create a customized, generative AI toolset that can enhance businesses operations cost effectively and economically. By leveraging heterogenous Dell servers, storage and switching together with readily available GPU's and Broadcom high-speed ethernet NICs provides a flexible hardware foundation for a scale-out AI platform.*

Additionally, the ability to build and manage both the hardware and software infrastructure helps companies compete effectively, while balancing their corporate security concerns and ensuring their data is not compromised or released externally.

The demonstrated AI model leverages key Dell and Broadcom hardware elements along with available GPUs as the foundation for a scalable AI platform. Additionally, the use of key software elements helps enable distributed training optimizations which leverage the underlying hardware to provide an extensible, self-managed AI platform that meets business objectives, regardless of industry.

The solution that was demonstrated highlights the ability to distribute AI training across multiple, heterogenous systems in order to reduce training time. This example leverages the value and flexibility of Dell and Broadcom infrastructure as an AI infrastructure platform, combined with open licensed tools to provide a foundation for practical AI development, while safe-guarding private data.

About The Futurum Group

The Futurum Group is dedicated to helping **IT professionals** and vendors create and implement strategies that make the most value of their storage and digital information. The Futurum Group services deliver **in-depth, unbiased analysis** on storage architectures, infrastructures, and management for IT professionals. Since 1997 The Futurum Group has provided services for thousands of end-users and vendor professionals through product and market evaluations, competitive analysis, and **education**.

This document was developed with funding from Dell Inc. and Broadcom. Although the document may utilize publicly available material from various vendors, including Dell, Broadcom and others, it does not necessarily reflect such vendors' positions on the issues addressed in this document.