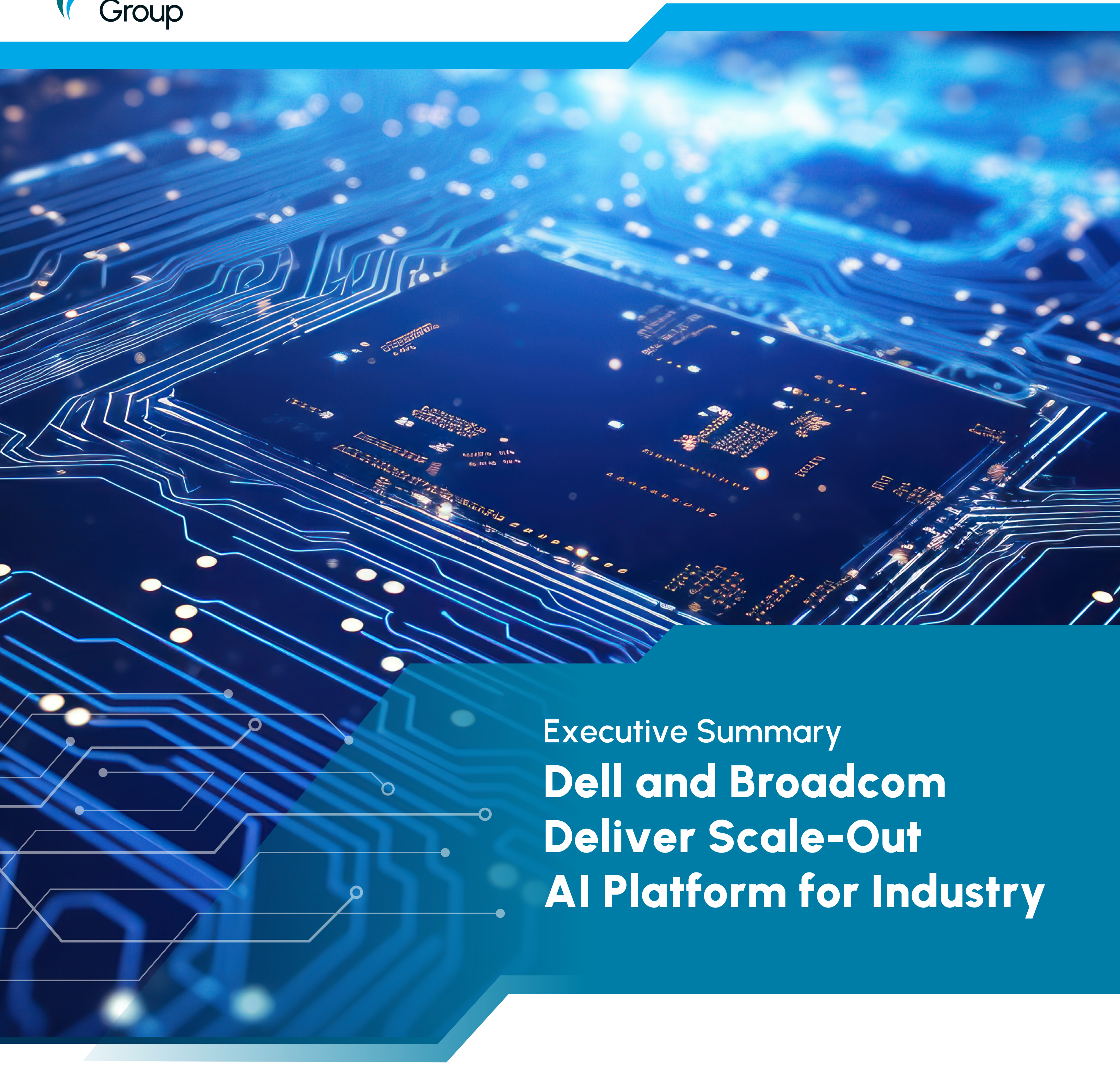

Executive Summary

Dell and Broadcom Deliver Scale-Out AI Platform for Industry

AUTHOR **Russell Fellows**
Head of Futurum Labs | The Futurum Group

JANUARY 2024

Solution Overview

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.

The solution utilizes pretrained LLM models, and then enhances or "fine-tunes" the underlying model with relevant, private data. This solves two challenges companies face, how to cost effectively train an LLM, and how to maintain private domain information within the context of a customized solution.

In this proof of concept, the distributed training cluster included three Dell PowerEdge servers with multiple GPUs, Broadcom NICs connected using a Dell Ethernet switch with training data residing on a Dell PowerScale NAS system. The key aspects of the heterogeneous Dell architecture 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.

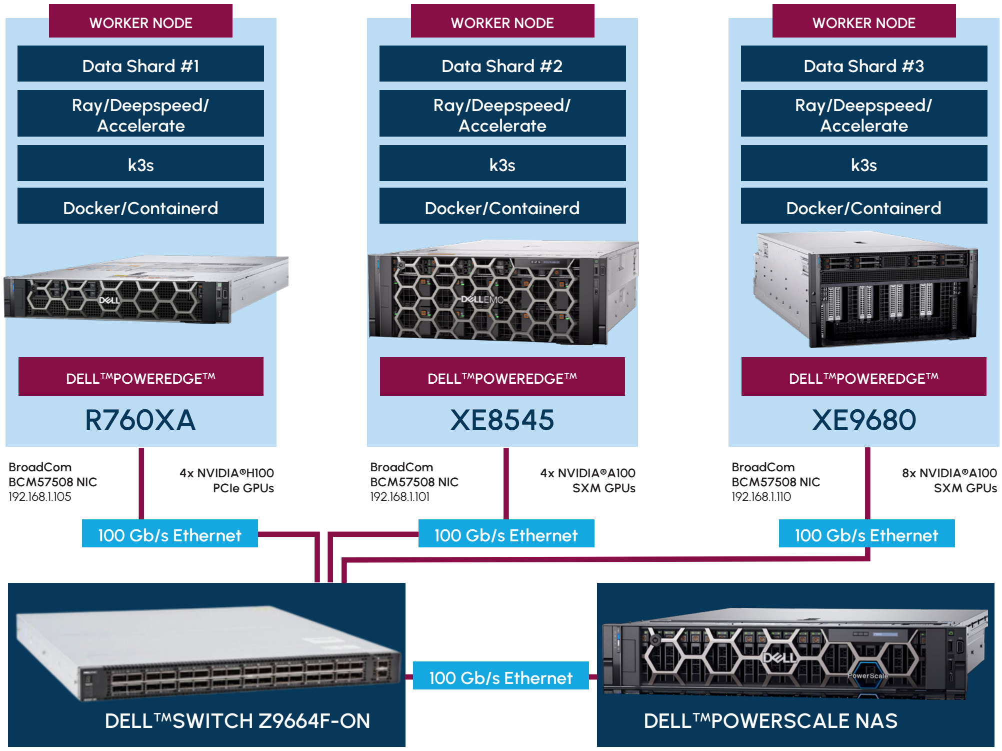

Figure 1: Distributed Scale-Out AI Hardware Platform (Source: Scalers.AI)

AI Practitioner Highlights

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, this allows firms to maintain HIPAA/HITECH compliance, and other regulations around EMR and patient records. Domain specific knowledge was provided via an open source "pubmed" data set. In a real-world deployment, it would be expected that an organization would utilize their own, proprietary and confidential medical data for fine-tuning.

The solution software stack leverages scale-out Dell / Broadcom infrastructure to reduce training time, coupled with a containerized software platform using open licensing to reduce deployment friction and cost. The Llama 2 foundational models were available from the Hugging Face repository, including three sizes: 7b, 13b and 70b. 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.

Fine-tuning occurred over multiple training epochs, using the hardware configuration outlined. For the Lama-13b model, training time on a single-node was 411 minutes, while the three-node cluster time was 78 minutes, or 5.3 times faster. The training time per epoch is shown in Figure 3, with lower (less time) indicating better results.

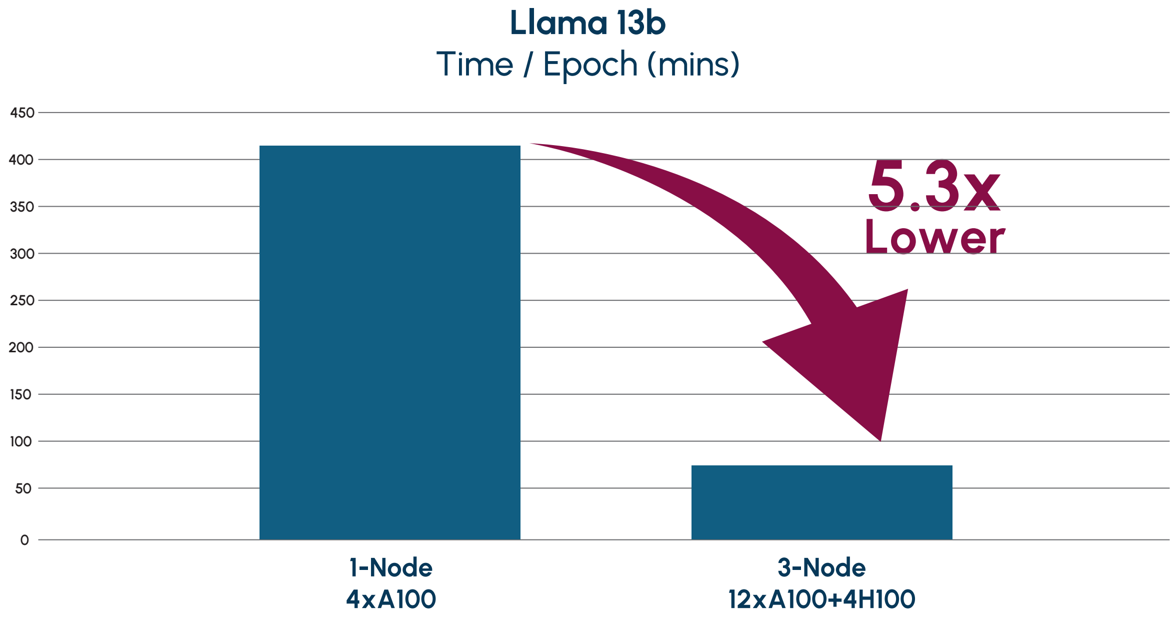

Figure 3: Scale-Out AI Software Platform (Source: The Futurum Group)

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*

This working example has been posted to Dell's GitHub Repository, available at the following URL: https://github.com/dell-examples/generative-ai

Conclusion

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 can solve several challenges that were previously available to only the very largest corporations. 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.

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.*

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.