Scale-Out Generative AI Solution using Dell & Broadcom Infrastructure

As part of Dell's efforts to help companies build flexible AI platforms, Dell is highlighting a scale-out architecture, built upon Dell and Broadcom equipment that can deliver the benefits of AI tools while ensuring data governance and data privacy.

- Pre-trained LLM models are fine-tuned using private data for a customized solution
- Scale-out architecture using Dell PowerEdge servers with Broadcom 100 Gb NICs
- Cost effective design utilizes heterogeneous Dell 16th Gen PowerEdge with GPU's
- Software leverages containers, Kubernetes and open-source AI libraries and orchestration

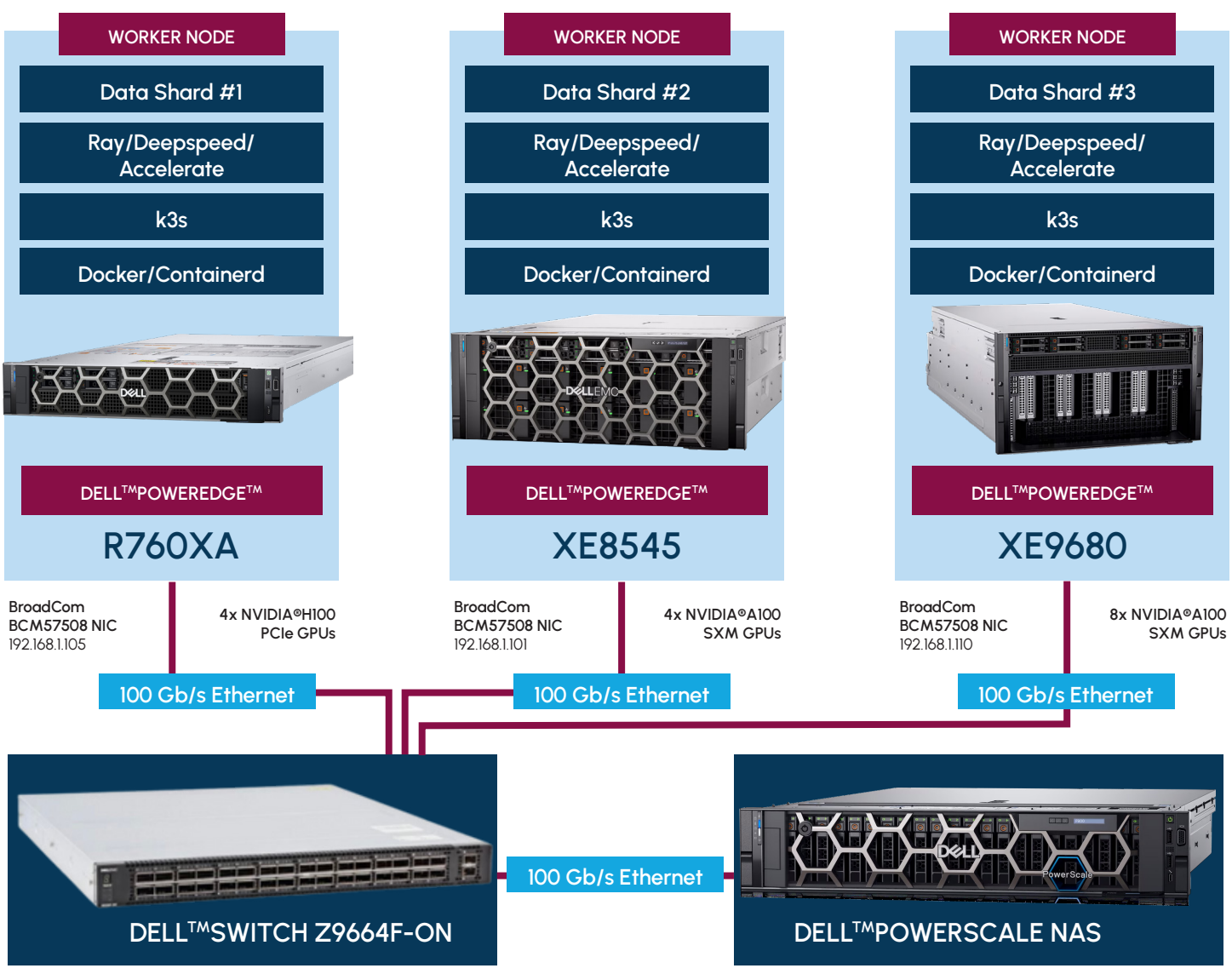
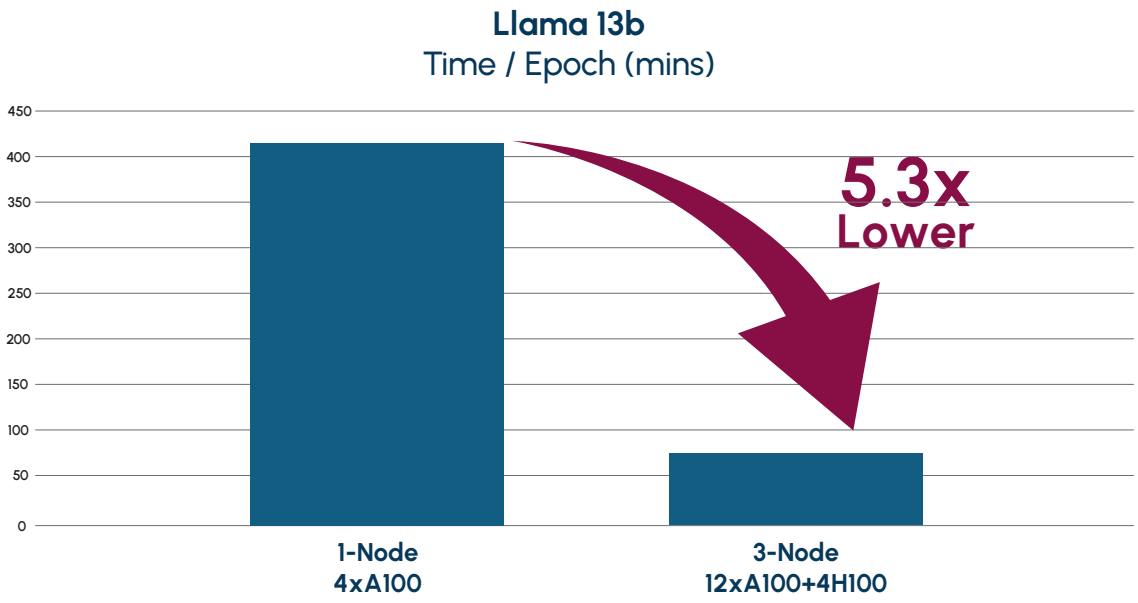


Figure 1: Distributed Scale-Out AI Hardware Platform

- The heterogeneous architecture provided nearly linear scaling on a per GPU core basis, considering the use of Nvidia A100 and H100 cards across the nodes.
- Fine tuning a single node using the Llama-7b model fell from 120 minutes to 46 minutes, or 2.6 times faster.
- Fine tuning the larger Lama-13b model, resulted in 5.3 times faster training using three nodes compared to a single node (411 minutes vs. 78 minutes)



The heterogeneous architecture provided nearly linear scaling on a per GPU core basis, considering the use of Nvidia A100 and H100 cards across the nodes.

Fine tuning a single node usinThe distributed, scale-out solution leveraged networked communications, with the training data split, or "sharded" across each node. After each step the AI model parameters are synchronized, updating model weights with other nodes using the 100 Gb/s networking connecting all systems.

- During synchronization i network bandwidth utilization spikes approached the 100 Gb/s bandwidth.
- 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.
- The AI software stack leverages open-source libraries, including:

DeepSpeed
deep-learning optimization libraries

Hugging Face
AI repository and HF-Accelerate library

PyTorch
Widely utilized AI libraries

Ray.IO
KubeRay distributed runtime management

Kubernetes
K3s container native platform

Nvidia
GPUs and CUDA drivers for fine-tuning

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Dell is a leading technology company known for its innovative computing solutions, including personal computers, servers, and storage devices. Founded in 1984 by Michael Dell, it's recognized for quality, customization, and customer service. Additionally, Dell emphasizes sustainability in its product lifecycle to minimize environmental impact.