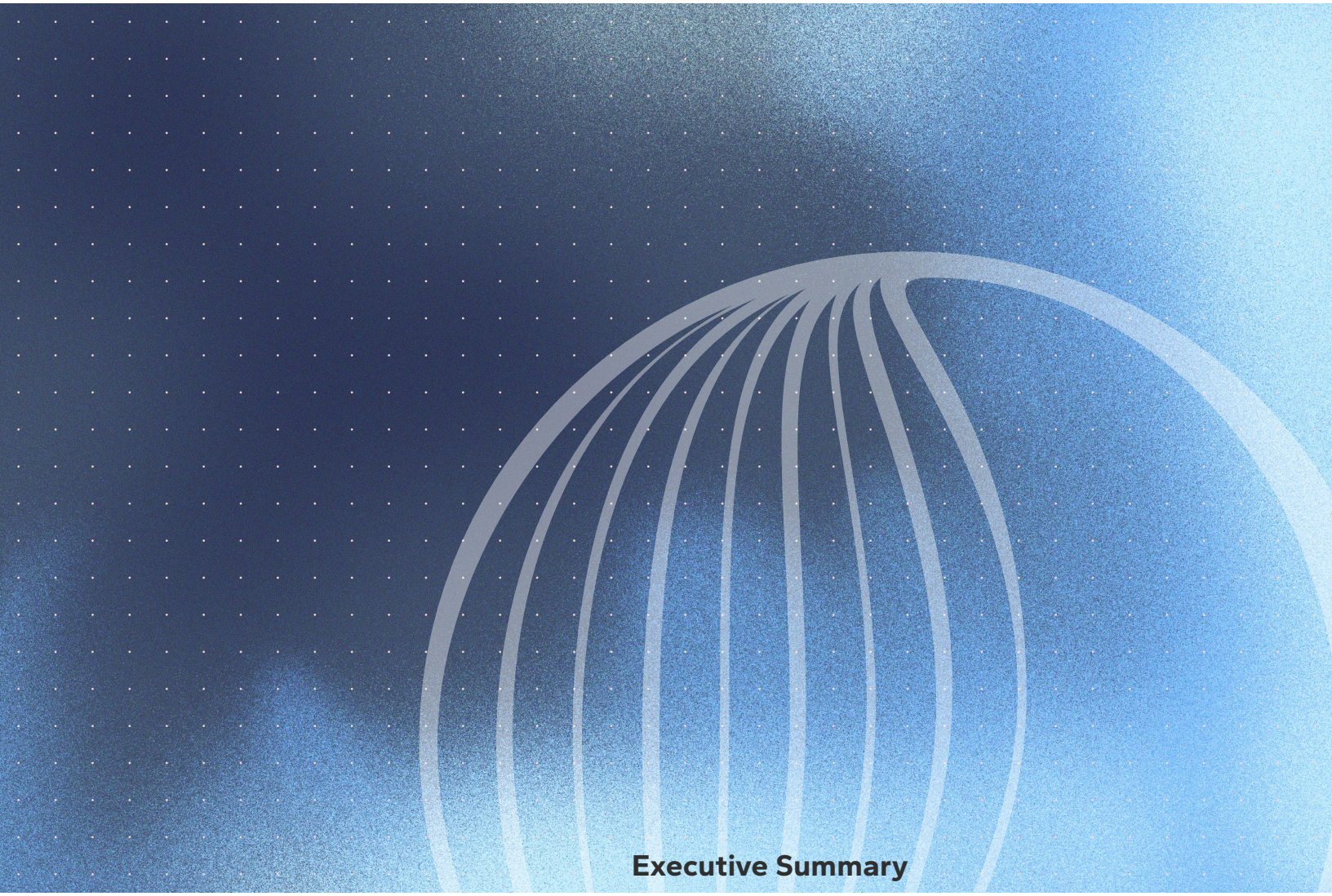
signal65



Executive Summary

Dell POC for Scalable and Heterogeneous Gen-AI Platform

AUTHOR

Russ Fellows
Head of Futurum Labs | The Futurum Group

IN PARTNERSHIP WITH

DELL Technologies

APRIL 2024

Solution Overview

As part of Dell's ongoing efforts to make industry leading AI workflows available to their clients, this paper outlines a scalable AI concept that supports heterogeneous hardware. The example created by Dell, working together with Scalers. AI and Signal65 showcases a Generative AI Large Language Model (LLM) used for inferencing, capable of operating across diverse hardware systems.

In this Proof of Concept, we investigate the ability to perform scale-out inferencing for production while utilizing a similar inferencing software stack across heterogeneous CPU and GPU systems to accommodate different production requirements. The PoC highlights the following:

- A single CPU based system can support multiple, simultaneous, real-time sessions
- GPU augmented clusters can support hundreds of simultaneous, real-time sessions
- A common AI inferencing software architecture is used across heterogenous hardware

Signal65 Comment: *The novel aspect of this proof of concept is the ability to operate across different hardware types, including Intel and AMD CPUs along with support for both Nvidia and AMD GPUs. By utilizing a common inferencing framework, organizations are able to choose the most appropriate hardware deployment for each application's requirements. This unique approach helps reduce the extensive customization required by AI practitioners, while also helping IT operations to standardize on common Dell servers, storage and networking components for their production AI deployments.*

Shown in Figure 1 below is a 3-node example that includes the hardware and general software stack

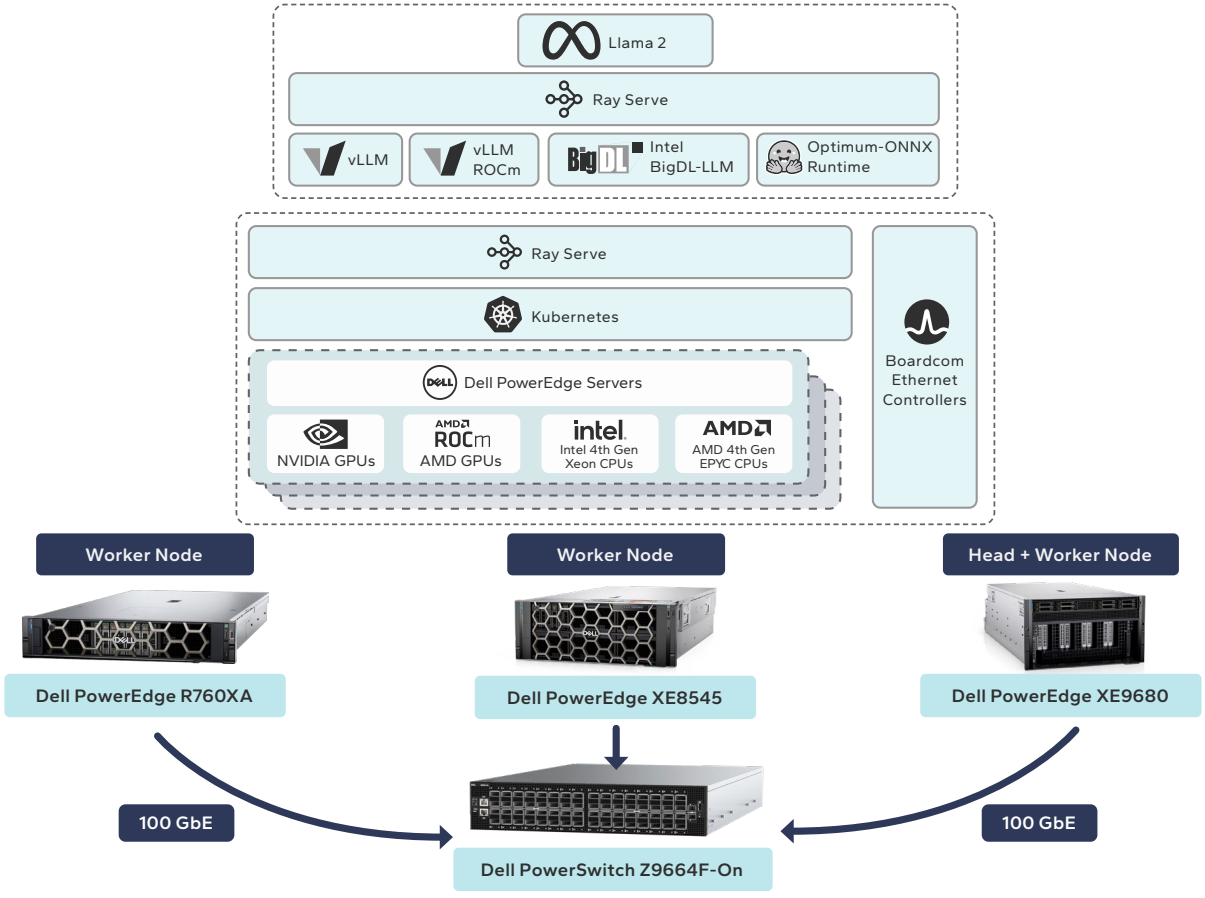


Figure 1: *General, Scale-Out AI Inferencing Stack (Source: Scalers.AI)*

There are several important aspects of the architecture that enables organizations to customize and deploy generative AI applications in their choice of colocation or on-premises data center. These 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 PowerSwitch Ethernet switches Z line support up to 400 Gb/s connectivity

Signal65 Comment: *By using a scale out inferencing solution leveraging industry standard Dell servers, networking and optional GPU accelerators provides a highly adaptable reference that can be deployed as an edge solution where few inferencing sessions are required, up to enterprise deployments supporting hundreds of simultaneous inferencing outputs.*

Interactive Performance

This PoC evaluated the effects of scaling multiple CPU and GPU configurations, along with various batch sizes on the resulting performance. For interactive performance, the rate of generative AI text generation should meet, or exceed a user's reading comprehension rate. According to our analysis, 200 words per minute can be considered a relatively fast rate for comprehending new text. Using this as a guideline results in a rate of 3.33 words per second (derived as: 200 wpm / 60 sec / minute = 3.33 words per second). A rate of 3.33 words per second (wps), is used as the desired minimum generation rate for a single interactive user or session with delay or latency calculated as 1 over this value, or 300 milliseconds.

In Figure 2 below, we show the total throughput of 3 different CPU configurations, along with the associated per word latency. As seen, a CPU only example can support over 40 words per second, significantly greater than the 3.33 word per second rate required for good interactive performance, while maintaining a latency of 152 ms., well under 300 ms.

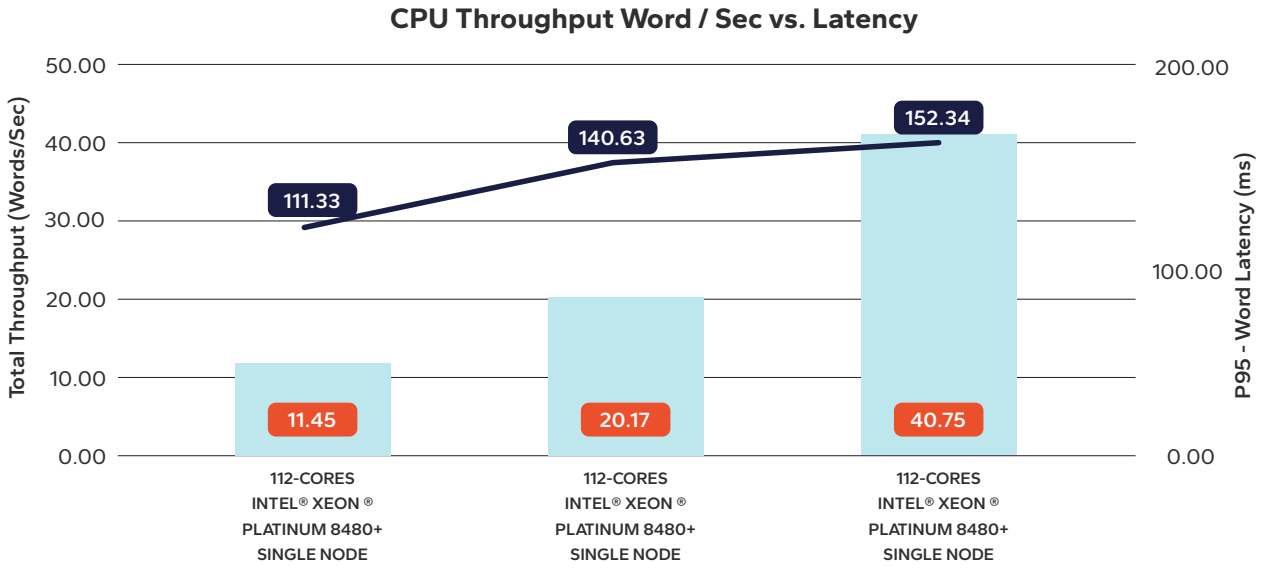


Figure 2: *Interactive Inferencing Performance for CPUs (Source: Futurum Group)*

In Figure 3 below, we show the total throughput of 4 different configurations, along with the associated per word latency. As seen, even at the rate of 1,246 words per second, latency remains at 100 ms., well below our 300 ms. threshold.

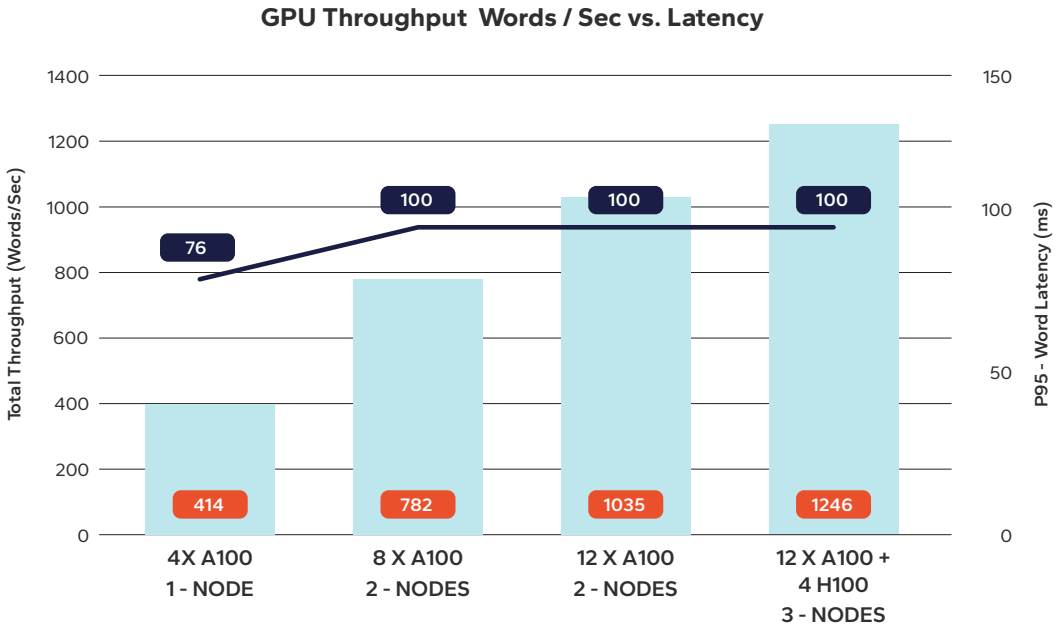


Figure 3: *Interactive Inferencing Performance for GPUs (Source: Futurum Group)*

Highlights

Perhaps the most unique aspect of this PoC is the ability to support an AI inferencing stack across CPU only and GPU enhanced hardware architectures, using an optimized stacks for each hardware type. Here, we have translated GenAI model terms such as "tokens per second" into relevant terminology that help IT operations correctly size hardware to meet expected usage. For interactive sessions, we use the rates of 3.33 wps, and a latency of 300 ms as outlined previously for baseline values for interactive sessions.

For environments requiring less than 50 simultaneous inferencing sessions:

- A current Dell PowerEdge server with a 4th generation processor can support up to 3 simultaneous interactive inferencing sessions per server.
- A 3-node CPU only cluster able to support up to 12 simultaneous sessions, at a total rate of 40 words per second across all three systems.

For use cases that require higher hundreds or thousands of simultaneous sessions:

- A single GPU based PowerEdge server with 4 GPUs can support up to 124 simultaneous interactive sessions.
- A 3-node, 16 GPU system was able to support 373 simultaneous inferencing sessions at a rate of 3.33 wps, for a total throughput of 1,245 words per second.

A key aspect of the PoC is the software stack that provides a platform for AI deployments, enabling scale-out infrastructure to significantly increase content creation rates. 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.

Summary

As artificial intelligence, and in particular Generative-AI matures, companies are seeking ways to leverage this new technology to provide advantages for their firms, helping to improve efficiency and other measures of user satisfaction. GenAI based Large Language Models are quickly showing their ability to augment some applications focused both on empowering internal users with additional knowledge and insights, and is also becoming increasingly useful for assisting clients, via chatbots or other similar interfaces. However, organizations often have concerns about becoming tied to proprietary, or cloud-based solutions, due to their privacy concerns, lack of transparency or potential vendor lock in.

As part of Dell's continuing efforts to democratize AI solutions, this proof-of-concept outlines specifically how organizations can build, deploy and operate production use of generative AI models using industry standard Dell servers. In particular, the scale-out PoC detailed in this paper showcases the ability to scale a solution efficiently from supporting a few simultaneous interactive users, up to a deployment supporting hundreds of simultaneous inferencing sessions simultaneously using as few as 3 Dell PowerEdge servers augmented by Nvidia GPUs. In an offline, or batch processing use, the same hardware example can support a throughput of nearly 3,000 words per second when processing multiple documents.

The outlined PoC utilizes the Hugging Face repository and libraries, along with hardware specific optimizations for each specific deployment of CPU or GPU type. By using a common framework, AI practitioners are better able to focus on enhancing the accuracy of the models and improved training methodologies, rather than trying to debug multiple solution stacks. Likewise, IT operations staff can utilize standard hardware, along with common IT technologies such as Kubernetes running on standard Linux distributions.

Important Information About this Report

CONTRIBUTORS

Russ Fellows
Head of Futurum Labs | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT DELL TECHNOLOGIES INC

Dell Technologies Inc engages in designing, developing, manufacturing, marketing, selling, and providing support for information technology infrastructure such as laptops, desktops, mobile devices, workstations, storage devices, software, cloud solutions, and notebooks.

ABOUT SIGNAL65

Signal65 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.

CONTACT INFORMATION
Signal65 I signal65.com