POC for Scalable and Heterogeneous Gen-AI Platform

Overview

Dell, Scalers.AI, and Signal65 present a Proof of Concept for scale-out inferencing using a common software stack across heterogeneous CPU and GPU systems demonstrating:

- A single CPU based system supporting multiple, simultaneous, real-time sessions
- GPU augmented clusters supporting hundreds of simultaneous, real-time sessions
- A common software stack for inferencing across heterogeneous hardware

Distributed Inferencing Stack

We present examples that include CPU only single and multi-node systems, and GPU augmented clusters for deployments of Generative-AI LLMs on these hardware configurations.

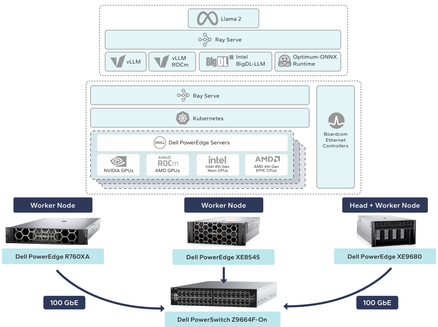

Figure 1: *General, Scale-Out AI Inferencing Stack (Source: Scalers.AI)*

Several important aspects of the architecture enable 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

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 inputs.

Performance

In Figure 2 below, we show the total throughput of 3 different CPU configurations, along with the associated per word latency.

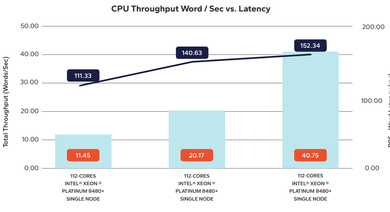

Figure 2: *Interactive Inferencing Performance for CPUs (Source: Futurum Group)*

Results showed that not all generative AI inferencing applications require a GPU for real-time deployments. With a date of 3.33 words per second and 100 ms. serving as acceptable levels for interactive performance, we achieve between 3 and 12 simultaneous streams:

- 11.45 / 3.33 = 3
- 20.17 / 3.33 = 6
- 40.75 / 3.33 = 12

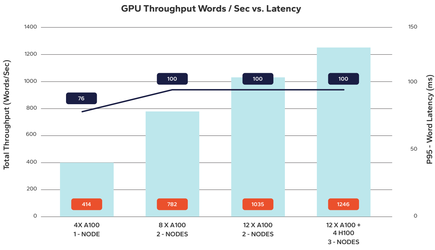

Figure 3: *Interactive Inferencing Performance for GPUs (Source: Futurum Group)*

The GPU based results significantly exceed those of the CPU based deployments. In all cases latency remains below 100ms.

- 414 / 3.33 = 124
- 782 / 3.33 = 34
- 1035 / 3.33 = 310
- 1246 / 3.33 = 374

Highlights

A key aspect of the PoC is the software stack that helps provide 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.

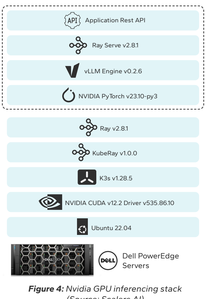

Figure 4: *Nvidia GPU inferencing stack (Source: Scalers.AI)*

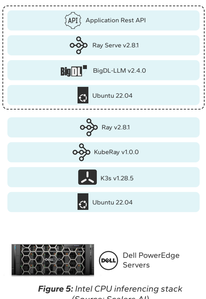

Figure 5: *Intel CPU inferencing stack (Source: Scalers.AI)*