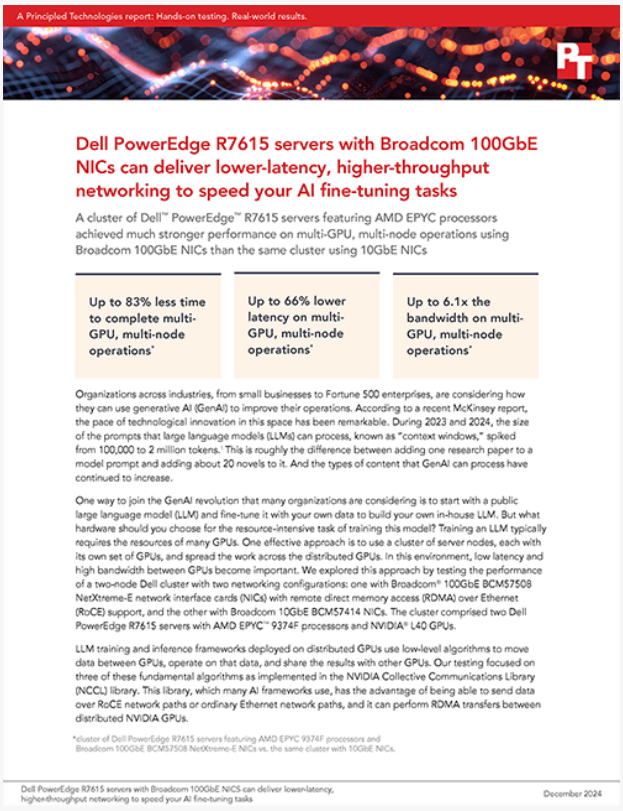
# Dell & Broadcom solution offered lower-latency, higher-throughput networking on AI fine-tuning tasks

*Dell PowerEdge R7615 servers with Broadcom 100GbE NICs using Broadcom software performed better on multi-GPU operations than the same servers with 10GbE NICs.*




Dell PowerEdge R7615 servers with Broadcom 100GbE NICS can deliver lower-latency, higher-throughput networking to speed your AI fine-tuning tasks

ROUND ROCK, TX, UNITED STATES, December 18, 2024 / EINPresswire.com/ -- As artificial intelligence (AI) continues to dominate tech news headlines, a host of organizations have already implemented AI operations or are considering doing so. One popular use case is the in-house AI chatbot, which combines a public large language model (LLM) with an organization's own data. Organizations can face a number of challenges in implementing such solutions, however. For small and medium businesses and departments within enterprises that have limited IT budgets, one challenge is determining what hardware is an appropriate choice for fine-tuning the LLM.

A recent report from third party Principled Technologies (PT) explores this question and presents a potential solution. As the test report says, "Training an LLM typically requires the resources of many GPUs. One effective approach is to use a cluster of server nodes, each with its own set of GPUs, and spread the work

across the distributed GPUs. In this environment, low latency and high bandwidth between GPUs become important."

The report goes on to explain the hardware that PT tested: "We explored this approach by testing the performance of a two-node Dell cluster with two networking configurations: one with Broadcom 100GbE BCM57508 NetXtreme-E network interface cards (NICs) with remote direct memory access (RDMA) over Ethernet (RoCE) support, and the other with Broadcom 10GbE BCM57414 NICs. The cluster comprised two Dell PowerEdge R7615 servers with AMD EPYC 9374F processors and NVIDIA L40 GPUs."

LLM training and inference frameworks deployed on distributed GPUs use low-level operations to move data between GPUs, operate on that data, and share the results with other GPUs. Testing focused on three of these operations as implemented in the NVIDIA Collective Communications Library (NCCL). Performing these operations efficiently depends on the timely transfer of data between GPUs on different servers.

PT found that the cluster with Broadcom 100GbE BCM57508 NetXtreme-E NICs performed substantially better on multi-GPU, multi-node operations, completing those operations in up to 83 percent less time than the cluster with 10GbE NICs, achieving lower latency, and supporting greater operational bandwidth. This improvement in performance could help speed AI fine-tuning tasks.

To learn more, read the test report at https://facts.pt/QAauY1Y, see the infographic at https://facts.pt/PplS5We, or review the two-page executive summary at https://facts.pt/AoOz7Np.


Infographic: Dell PowerEdge R7615 servers with Broadcom 100GbE NICS can deliver lower-latency, higher-throughput networking to speed your AI fine-tuning tasks

About Principled Technologies, Inc.
Principled Technologies, Inc. is the leading provider of technology marketing and learning & development services.

Principled Technologies, Inc. is located in Durham, North Carolina, USA. For more information, please visit www.principledtechnologies.com.

Sharon Horton
Principled Technologies, Inc.
press@principledtechnologies.com
Visit us on social media:
Facebook

X
LinkedIn
YouTube

---