

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

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

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



Principled Technologies®

A Principled Technologies report: Hands-on testing. Real-world results.

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

A cluster of Dell™ PowerEdge™ R7615 servers featuring AMD EPYC processors achieved much stronger performance on multi-GPU, multi-node operations using Broadcom 100GbE NICs than the same cluster using 10GbE NICs

Up to 83% less time to complete multi-GPU, multi-node operations*	Up to 66% lower latency on multi-GPU, multi-node operations*	Up to 6.1x the bandwidth on multi-GPU, multi-node operations*
---	--	---

Organizations across industries, from small businesses to Fortune 500 enterprises, are considering how they can use generative AI (GenAI) to improve their operations. According to a recent McKinsey report, the pace of technological innovation in this space has been remarkable. During 2023 and 2024, the size of the prompts that large language models (LLMs) can process, known as "context windows," spiked from 100,000 to 2 million tokens. This is roughly the difference between adding one research paper to a model prompt and adding about 20 novels to it. And the types of content that GenAI can process have continued to increase.

One way to join the GenAI revolution that many organizations are considering is to start with a public large language model (LLM) and fine-tune it with your own data to build your own in-house LLM. But what hardware should you choose for the resource-intensive task of training this model? 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. 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 algorithms to move data between GPUs, operate on that data, and share the results with other GPUs. Our testing focused on three of these fundamental algorithms as implemented in the NVIDIA Collective Communications Library (NCCL) library. This library, which many AI frameworks use, has the advantage of being able to send data over RoCE network paths or ordinary Ethernet network paths, and it can perform RDMA transfers between distributed NVIDIA GPUs.

*Cluster of Dell PowerEdge R7615 servers featuring AMD EPYC 9374F processors and Broadcom 100GbE BCM57508 NetXtreme-E NICs vs. the same cluster 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. December 2024

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

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/PpIS5We>, or review the two-page executive summary at <https://facts.pt/AoOz7Np>.

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](#)

Dell PowerEdge R7615 servers with Broadcom BCM57508 NICs can accelerate your AI fine-tuning tasks

A cluster of Dell[®] PowerEdge[™] R7615 servers featuring AMD EPYC processors achieved much stronger performance on multi-GPU, multi-node operations using Broadcom 100GbE NICs than the same cluster using 10GbE NICs.

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. Our testing focused on three of these operations as implemented in the NVIDIA Collective Communications Library (NCCL): memory, allreduce, reduce-scatter, and broadcast. This testing, which may vary by framework, can send data over NICs network paths or utilize Ethernet network paths and remote direct memory access (RDMA) between NVIDIA GPUs.

We tested a two-node cluster of Dell PowerEdge R7615 servers with AMD EPYC[™] 9374F processors and NVIDIA[®] L40 GPUs 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
- one with 10GbE NICs

For each configuration, we studied three multi-GPU, multi-node AI computations from the NCCL test suite at different packet sizes and measured the time to complete the task, latency, and the effective bandwidth of the network during the operation. The cluster with 100GbE networking dramatically outperformed the cluster with 10GbE networking across all packet sizes and tasks without increasing power usage.

Power usage that does not rise enough to offset data between servers is considered for networking task. Below these tests complete a portion of computational steps on each GPU, when a given step may require data from other GPUs. In each test, a GPU sends only the next computational step once it has the data from the other GPUs, even if that data is as small as a single byte. The operational bandwidth depends on the timely transfer of data between GPUs on different servers.

Up to 83% less time to complete multi-GPU, multi-node operations

Send-receive performance: Time to complete task

Packet Size (B)	100GbE Configuration	10GbE Configuration	Percentage reduction in time
1024	40	123	67.4%
10240	24	88	72.9%
102400	41	54	24.1%

Up to 40% lower latency on multi-GPU, multi-node operations

Packet Size (B)	100GbE Configuration	10GbE Configuration	Percentage reduction in latency
1024	40	123	67.4%
10240	24	88	72.9%
102400	41	54	24.1%

Up to 4.1x the bandwidth on multi-GPU, multi-node operations

Send-receive bandwidth

These three multi-GPU, multi-node NCCL primitive operations for AI we used for testing are:

- allreduce: Operates on the entire dataset, distributed across all GPUs in the cluster and sends the single result to each GPU.
- reduce-scatter: Splits the data on every GPU into equal chunks, and distributes each chunk across the cluster to form partial results. These send one partial result to each GPU and receive it from each GPU.
- broadcast: Sends data from one GPU to another on the second server, and receive a response for all sending chunks and results, read out for usage.

Learn more at <https://facts.pt/QAauY1Y>

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

X

LinkedIn

YouTube

This press release can be viewed online at: <https://www.einpresswire.com/article/769796766>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2024 Newsmatics Inc. All Right Reserved.