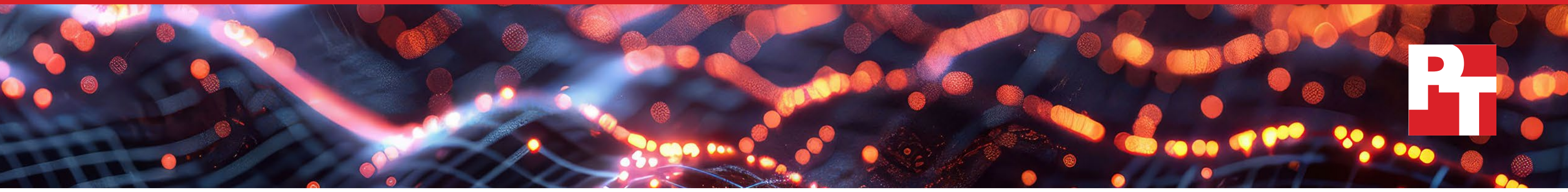

Dell PowerEdge R7615 servers with Broadcom 100GbE NICs can deliver lower-latency, higher-throughput networking to speed your AI fine-tuning tasks

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.[1] 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.

For each configuration, we studied three multi-GPU, multi-node AI computations from the NCCL test suite[2] at packet sizes ranging from 4 B to 256 MB and measured the time to complete the operation and the effective bandwidth of the network during the operation. This operational bandwidth is a combination of the very fast data transfer between GPUs on the same node, and the slower data transfer between GPUs on different nodes. Across this range of packet sizes and each of the three low-level AI operations, the cluster with 100GbE networking dramatically outperformed the cluster with 10GbE networking. Compared to the 10GbE networking configuration, the operational latency decreased by 26 percent to 67 percent, and the operational bandwidth was 3.7 to 6.1 times as high. In addition, the 100GbE cluster achieved these gains without increasing power usage.

Please note that these test do not send enough data between servers to overwhelm the networking link. Rather, these tests comprise a sequence of computational steps on each GPU, where a given step may require data from other GPUs. In such cases, a GPU can only start the next computational step once it has the data from those 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. The quality of this data transfer depends on three factors: the time to transfer small amounts of data from a GPU to the server's NIC, the time to transfer this data through the network link to the second server's NIC, and the time to transfer this data from this NIC to the second GPU.

The value of an in-house LLM for small and medium businesses

AI technologies are complex, and it would be easy to assume that only the largest organizations can utilize AI effectively and at scale. But that's not the case. In a recent survey, eight out of ten businesses with under $1M in revenue reported that they already rely on AI tools.[3] According to the Bipartisan Policy Center, which surveyed businesses on their use of digital tools, "Significant progress in connecting small business owners to AI has occurred over the last two years."[4]Just as large enterprises are building AI implementations for everything from product development to customer service, small and medium businesses (SMBs) are improving business operations using AI.

The idea of a private LLM, trained on your own organization's existing data and updated regularly as new data comes in, is particularly appealing. LLMs trained on your own data allow you to gain all the benefits of an AI chatbot while keeping your data in house, thus maintaining data privacy. SMBs could both save time and access new opportunities by building and utilizing such LLMs. Manufacturing organizations might be able to leverage their LLMs to find defects more quickly. Companies across industries could benefit from LLMs that can analyze images in ways that target specific business needs.

Building an in-house LLM requires a great deal of planning. One of the first steps in the planning process is selecting the technology solution. You'll likely want powerful computing resources and networking, and sourcing them from a manufacturer with significant AI experience could provide further benefits.

Dell: A proven partner for AI

While we highlight the performance of one specific Dell server in this paper, Dell offers a large range of AI solutions and services. In the 2024 Principled Technologies report "Meeting the challenges of AI workloads with the Dell AI portfolio," we highlight advantages Dell brings for AI. According to that report, the Dell AI portfolio offers "professional and consultative services that help customers build implementation roadmaps and prepare their data for AI models….training courses that cover machine learning (ML) concepts and other educational topics…[and] validated designs for AI to help ensure implementation success."[5]

Our approach to testing

Training LLMs with custom data typically requires many GPUs, which companies can deploy in a multi-node cluster. Modern LLM frameworks such as DeepSpeed, Megatron, and PyTorch perform fundamental arithmetic and data-transfer operations on an LLM spread across all GPUs. Low network latency and high bandwidth are necessary for performance because, e.g., the overall computation rate slows if GPUs are waiting for data.

We performed tests to determine the operational latency and throughput for three multi-node, multi-GPU tasks common to and necessary for LLM data-parallelism methods and LLM model-parallelism frameworks. We used tasks from NCCL, which uses RoCE, when present, to speed inter-node GPU communications (see the box "What are RDMA and RoCE?" to learn more). NCCL optimizes GPU communication to achieve high bandwidth and low latency over PCIe and NVLink high-speed interconnects within a node and across nodes.[6] In our tests, we used publicly available Broadcom driver modules to achieve this functionality, viz. GPUDirect, for the PCIe and RoCE interconnects.

To assess the benefits of choosing low-latency, high-speed Broadcom NICs, we tested the cluster's performance with two network configurations: one with 100GbE Broadcom BCM57508 NetXtreme-E NICs with RoCE and one with 10GbE NICs. Table 1 provides an overview of the hardware in our test configurations. For greater detail, including how we configured the network switch for RoCE, see the science behind the report.

Table 1: The two cluster configurations we tested. Source: Principled Technologies.

100GbE cluster configuration	10GbE cluster configuration
2 x Dell PowerEdge R7615 servers	
3 x NVIDIA L40 GPUs per server	
1 x AMD EPYC 9374F processor per server	
1 x Dell PowerSwitch Z9100-ON (for both 100Gbps and 10Gbps configurations)	
Broadcom BCM57508 10/25/50/100/200G NetXtreme-E NIC with RoCE	Broadcom BCM57414 10G/25G NetXtreme-E NIC
Broadcom software to extend RDMA into the NVIDIA GPUs	

Note: The Broadcom BCM57508 NetXtreme-E NIC supports the following speeds: 10GbE, 25GbE, 40GbE, 50GbE, 100GbE, and 200GbE. We used the 100GbE setting. The Broadcom BCM57414 10G/25G NetXtreme-E NIC supports the following speeds: 10GbE and 25GbE. We used the 10GbE setting.

About the Dell PowerEdge R7615

The Dell PowerEdge R7615, featuring 4th Generation AMD EPYC processors, is a 2U, single-socket server that Dell has "designed to be the best investment per dollar for your data center."[7] In another recent PT study, we found that the PowerEdge R7615 could deliver 44 percent better MySQL database performance than a legacy server, which supports consolidation and the possibility of OpEx savings.[8] Learn more at https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-r7615-spec-sheet.pdf.

Why test the impact of network speed on training?

Much of the AI activity in the news emphasizes the inference stage of the AI LLM workflow. Before inference, however, comes LLM training. Publicly available AI models are pre-trained on general sets of data. If organizations wish to use these pre-trained models, they may skip straight to using them for inference—at the cost of being unable to leverage their own in-house data while maintaining the privacy of that data. Alternatively, organizations can train the models on their own corpuses of data. This requires them to go through an additional phase of training, but at the end of that phase, the model could base its output on an organization's specific data.

Low-latency networking hardware, such as the Broadcom 100GbE BCM57508 NetXtreme-E NICs with which we tested, is especially useful in an AI training setting, as we'll explore in greater detail later in this report.

For our testing, we chose three multi-GPU, multi-node NCCL primitive operations for AI that are commonly used in GenAI frameworks performing LLM training with GPUs. Those operations are:

- **all-reduce:** Operate on the entire dataset, distribute across all GPUs in the cluster, and store the single result on each GPU
- **reduce-scatter:** Divide the data on every GPU into logical chunks, and operate on each chunk across the cluster to form partial results. Then send one partial result to each GPU and store it there
- **send-receive:** Send data from one GPU to another on the second server, and return a response

What are RDMA and RoCE?

In a multi-node cluster, where each server has its own GPU(s), the speed at which data can travel from one GPU to another—inter-GPU communications—plays a vital role in performance.

To understand the findings of our testing, it's useful to know two terms: RDMA and RoCE. Remote Direct Memory Access, or RDMA, supports moving data from application memory on one server to that on another server without any CPU involvement. Broadcom describes RoCE as follows: "RoCE (RDMA over converged Ethernet) is a complete hardware offload feature supported on Broadcom Ethernet network adapters, which allows RDMA functionality over an Ethernet network. RoCE helps to reduce CPU workload as it provides direct memory access for applications bypassing the CPU. As the packet processing and memory access are done in hardware, RoCE allows for higher throughput, lower latency, and lower CPU utilization on both the sender and the receiver side, which are critical for Machine Learning (ML/AI), Storage, and High Performance Compute (HPC) applications."[9]

The Broadcom NICs we used in our 100GbE cluster configuration support RoCE. In our testing, this direct memory transfer allowed data to travel efficiently from a GPU on one node to a GPU on another node without the CPU becoming involved, a significant factor in the performance gains we observed in this configuration. While the Broadcom BCM57414 10GbE NICs are advanced enough to support RoCE, we chose not to use then to see how non-RoCE NICs performed in an AI training environment.

About Broadcom BCM57508 10/25/50/100/200G NetXtreme-E NICs

Broadcom network adapters support RoCE, which provides direct memory access for applications, allowing them to bypass the processor and thus reduce overall CPU load. Skipping the processor can result in higher throughput, which can speed up AI training workloads. According to Broadcom, the BCM57508 10/25/50/100/200G NetXtreme-E NIC we used in our testing "builds upon the success of the widely-deployed NetXtreme E-Series architecture by combining a high-bandwidth Ethernet controller with a unique set of highly-optimized hardware acceleration engines to enhance network performance and improve server efficiency."[10]

To learn more about these NICs, visit https://www.broadcom.com/products/ethernet-connectivity/network-adapters/bcm57508-200g-ic.

What we found

Once you've decided to put your own data to use and create a tailored LLM in house, the next step is deciding which hardware you'll use to support your LLM. The servers and networking solution you choose for LLM training should be able to process data quickly to speed up the training process so you can ultimately move on to the next phase. Better performance means you can complete training operations on larger data sets faster, and get to a viable AI implementation sooner. As the test results we present in this section illustrate, the Dell solution with Broadcom BCM57508 10/25/50/100/200G NetXtreme-E can give organizations the performance they need for in-house LLM training.

Time to complete tasks

Figures 1 through 3 show our multi-GPU, multi-node performance results on the three AI fine-tuning tasks for the two networking configurations. We see the same pattern across the results: As the size of the data increased, the amount of time the configuration with the slower 10GbE networking needed increased at a much faster rate than the configurations with faster 100GbE networking. At the largest packet size we tested, the 100GbE networking configuration took approximately one-sixth the time to complete each of the tasks as the 10GbE configuration did. At this size, time to complete decreased by 82 percent on the all-reduce and reduce-scatter tasks (see Figures 1 and 2) and by 83 percent on the send-receive task (see Figure 3).

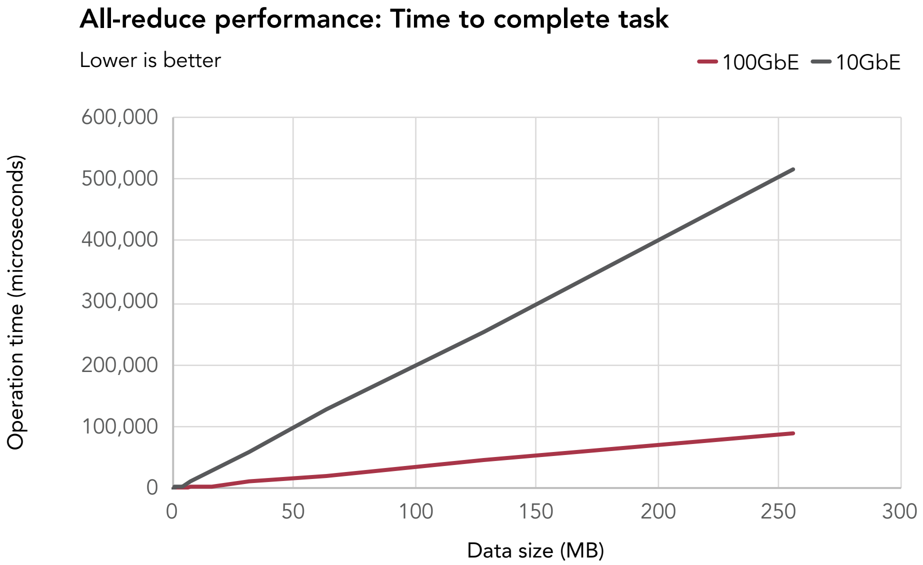

Figure 1: Performance of all-reduce multi-GPU, multi-node task in terms of time in microseconds to complete the task on datasets of multiple sizes. Lower is better. Source: Principled Technologies.

Figure 2: Performance of reduce-scatter multi-GPU, multi-node task in terms of time in microseconds to complete the task on datasets of multiple sizes. Lower is better. Source: Principled Technologies.

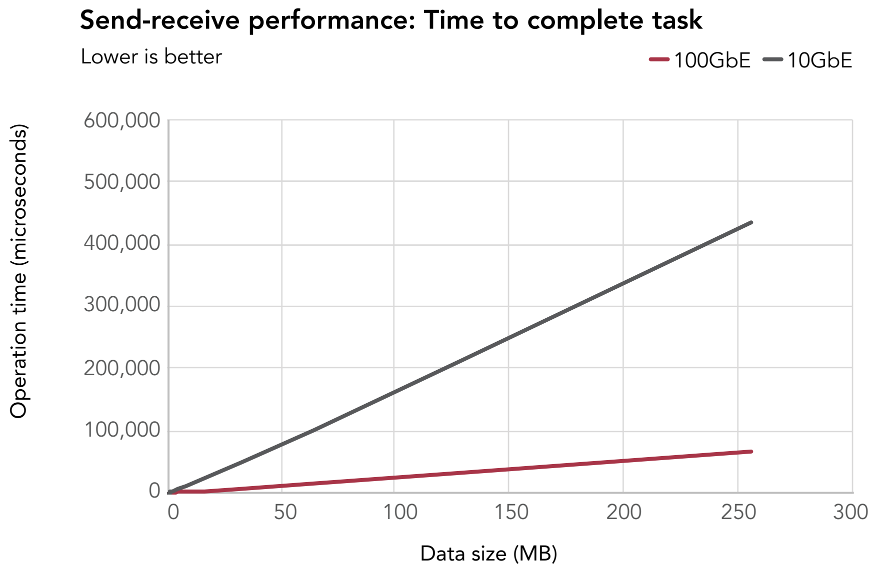

Figure 3: Performance of send-receive multi-GPU, multi-node task in terms of time in microseconds to complete the task on datasets of multiple sizes. Lower is better. Source: Principled Technologies.

About 4th Gen AMD EPYC processors

The servers we tested used AMD EPYC 9374F processors, part of the 4th Gen AMD EPYC processor family. According to AMD, this group of processors "feature the performance, scalability, compatibility, and energy efficiency to support hosting of advanced GPU AI engines."[11] EPYC processors include AMD InfinityGuard, which AMD describes as "a set of layered, cutting-edge security features that help you protect sensitive data and avoid the costly downtime cause by security breaches."[12] To learn more about 4th Gen AMD EPYC processors, visit https://www.amd.com/en/products/processors/server/epyc.html.

Latency for multi-GPU, multi-node AI tasks

We measured the latency for the distributed GPU operations by examining the completion time for small packet size (4 B for all-reduce and send-receiver and 48 B for reduce-scatter), where the on-GPU computational time was minimal and the inter-GPU communications dominated.[13] The latencies we measured are in Table 1. As it shows, using the 100GbE NIC improved latency by over 65 percent for the all-reduce and reduce-scatter tasks, and by 26.7 percent for the send-receive task.

Table 2: Latency of multi-GPU, multi-node operations. Lower latency and higher percentage improvement are better. Source: Principled Technologies.

Multi-GPU, multinode operation	**Latency (microseconds)** Lower is better		**Percentage reduction** Higher is better
	100GbE configuration	10GbE configuration	
all-reduce (packet size: 4 B)	40	123	67.4%
reduce-scatter (packet size: 4 B)	29	85	65.8%
send-receive (packet size: 48 B)	41	56	26.7%

Bandwidth for multi-GPU, multi-node AI tasks

When completing AI training workflows, the rate at which data travels across your AI solution matters. The greater the flow, or bandwidth, from a GPU on one node to a GPU on another, the better. Choosing AI solutions with greater bandwidth reduces possible performance bottlenecks and can leave you room to scale as your AI needs continue to grow.

On all three tasks, we saw dramatically greater bandwidth for multi-GPU, multi-node operations with the 100GbE network configuration. On all-reduce and reduce-scatter tasks, the bandwidth was 5 times that of the 10GbE configuration (see Figures 4 and 5). On the send-receive task, the 100GbE configuration achieved 6 times the bandwidth (see Figure 6). Note that in Figure 5, we see one data point at which bandwidth exceeds 10Gbps on the 10GbE adapter. We believe that intra-GPU traffic—data moving within a GPU—caused this.

Figure 4: Bandwidth achieved for multi-GPU, multi-node all-reduce task. Higher is better. Source: Principled Technologies.

Reduce-scatter bandwidth

Higher is better

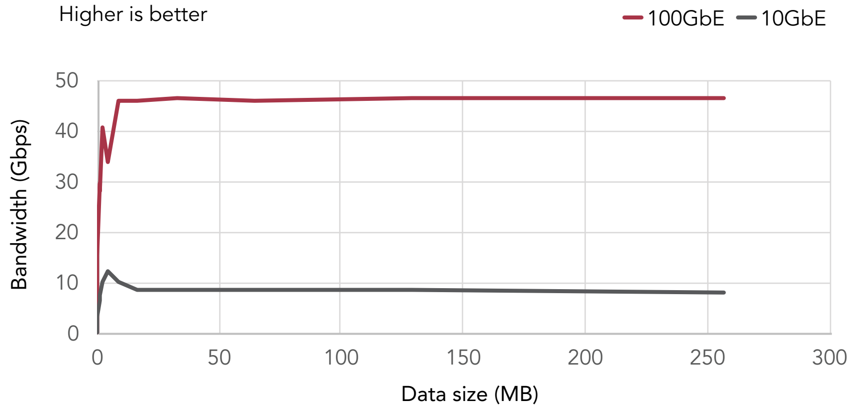

Figure 5: Bandwidth achieved for multi-GPU, multi-node reduce-scatter task. Higher is better. Note that the operational bandwidth at 4MB for the 10G network actually exceeds 10Gbps. For this packet size, the speed and amount of data transferred between GPUs on one node contributed more to the operational bandwidth than that for inter-node data transfers. Source: Principled Technologies.

Send-receive bandwidth

Higher is better

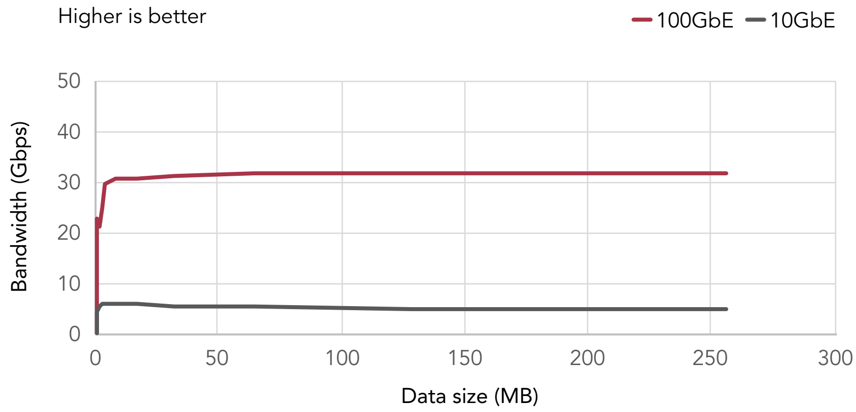

Figure 6: Bandwidth available for multi-GPU, multi-node send-receive task. Higher is better. Source: Principled Technologies.

About Dell PowerSwitch Z9100-ON Series switches

The Dell EMC Z9100-ON is a 10/25/40/50/100GbE fixed switch Dell has designed specifically to support applications in high-performance data center and computing environments. The 1RU switch offers a choice of the following: 32 ports of 100GbE (QSFP28), 64 ports of 50GbE (QSFP+), 32 ports of 40GbE (QSFP+), 128 ports of 25GbE (SFP28), or 128+2 ports of 10GbE (using breakout cable). These options let users to conserve rack space, increase footprint density, and migrate more easily to 100Gbps in the data center core.

Learn more at https://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/dell-networking-z9100-spec-sheet.pdf.

Power usage

As AI ripples through global news headlines, the world has been paying close attention to the increased power and cooling that AI workloads require. According to one *Scientific American* interview on the topic, "there's going to be a growth in AI-related electricity consumption"—although "the latest servers are more efficient than older ones."[14] Selecting servers with increased power efficiency can help you not only reduce your organization's carbon footprint but also save on operational expenditures (OpEx), lowering those hefty power and cooling bills.

We wished to see whether the higher-performing 100GbE environment required more power during our tests than the 10GbE one. It did not. As we conducted our multi-GPU, multi-node testing, we measured the power consumption of both servers. Table 3 reports the change to power usage by the two servers at three representative packet sizes, spanning four orders of magnitude. Despite the great multi-GPU, multi-node AI task performance improvements the 100GbE Broadcom card enabled, power usage did not increase significantly with its use. (Note that we did not specifically drill down into GPU power usage during testing; instead, we report the server's power usage, which includes the GPU power usage.)

Table 3: Power usage of the two network configurations we tested at three tasks and on three packet sizes: 8KB, 1MB, and 128MB. Lower is better. Source: Principled Technologies.

Power usage by the servers during each test (Watts, Lower is better)						
Packet size	All-reduce		Reduce-scatter		Send-receive	
	100GbE	10GbE	100GbE	10GbE	100GbE	10GbE
8B	1,393.2	1,396.4	1,392.3	1,410.5	1,381.2	1,380.6
1B	1,387.6	1,389.0	1,390.0	1,388.9	1,392.6	1,388.6
128B	1,405.6	1,392.8	1,405.5	1,393.0	1,416.4	1,392.3

The potential for cost savings

Every IT organization has a budget, and according to an Enterprise Technology Research report, IT budget growth is beginning to slow.[15] When you're considering purchasing a new technology solution to initiate or grow your AI implementation, you must consider its cost alongside its value. That cost is more than purchase price. Expenditures for power, cooling, and licensing are also factors in the total cost of a solution over its lifetime.

Additionally, choosing a solution that can process your in-house data quickly lets you build and fine-tune your AI model in less time and puts your data to work faster, thereby increasing your business agility and allowing you to reap the benefits of your AI implementation sooner.

Conclusion

Many companies want to do LLM training on their internal data so they can use it to solve a host of business problems. LLM training uses low-level fundamental operations over distributed GPUs. When these operations perform efficiently, your LLM training takes much less time to complete and you can have your AI implementation operational sooner. Our tests looked at the performance of fundamental operations over distributed GPUs. We found that using Broadcom 100GbE BCM57508 NICs in a cluster of two Dell PowerEdge R7615 servers with AMD EPYC processors and NVIDIA GPUs provided dramatically lower latency and greater bandwidth than using only 10GbE networking, with no increase in power usage.

1. "McKinsey Technology Trends Outlook 2024," accessed November 19, 2024, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-top-trends-in-tech#new-and-notable.
2. NVIDIA, "NCCL Tests," accessed November 16, 2024, https://github.com/NVIDIA/nccl-tests.
3. Small Business & Entrepreneurship Council, "Small Business AI Adoption Survey October 2023," accessed November 17, 2024, https://sbecouncil.org/wp-content/uploads/2023/10/SBE-Small-Business-AI-Survey-Oct-2023-FINAL.pdf.
4. Sujan Garapati, "Poll Shows Small Businesses Are Interested in and Benefit from AI," accessed November 17, 2024, https://bipartisanpolicy.org/blog/poll-shows-small-businesses-are-interested-in-and-benefit-from-ai/.
5. "Meeting the challenges of the Dell AI portfolio," accessed November 17, 2024, https://www.principled-technologies.com/Dell/AI-portfolio-vs-HPE-0124.pdf.
6. NVIDIA, "NVIDIA Collective Communications Library (NCCL)," accessed November 21, 2024, https://developer.nvidia.com/nccl.
7. "PowerEdge R7615," accessed November 17, 2024, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-r7615-spec-sheet.pdf.
8. "Improve performance and gain room to grow by easily migrating to a modern OpenShift environment on Dell PowerEdge R7615 servers with 4th Generation AMD EPYC processors and high-speed 100GbE Broadcom NICs," accessed November 17, 2024, https://www.principledtechnologies.com/clients/reports/Dell/PowerEdge-R7615-100GbE-Broadcom-NICs-MYSQL-database-0524/index.php.
9. Broadcom, "RDMA over Converged Ethernet (RoCE)," accessed November 21, 2024, https://techdocs.broadcom.com/us/en/storage-and-ethernet-connectivity/ethernet-nic-controllers/bcm957xxx/adapters/RDMA-over-Converged-Ethernet.html.
10. Broadcom, "BCM57508 - 200GbE," accessed November 21, 2024, https://www.broadcom.com/products/ethernet-connectivity/network-adapters/bcm57508-200g-ic.
11. AMD, "Nothing Stacks up to EPYC," accessed November 25, 2024, https://www.amd.com/en/products/processors/server/epyc.html.
12. AMD, "Nothing Stacks up to EPYC."
13. NVIDIA, "Performance reported by NCCL tests," accessed November 15, 2024, https://github.com/NVIDIA/nccl-tests/blob/master/doc/PERFORMANCE.md.
14. Lauren Leffer, "The AI Boom Could Use a Shocking Amount of Electricity," accessed November 17, 2024, https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity/.
15. ETR, "2024 IT Budget Growth is Slowing," accessed November 17, 2024, https://www.linkedin.com/pulse/2024-budget-growth-slowing-etr-enterprise-technology-research-t9ikf/.

Read the science behind this report at https://facts.pt/cbYq3Wb ▶

Principled Technologies®

Facts matter.®

For additional information, review the science behind this report.

This project was commissioned by Dell Technologies.