LIQID

UltraStack Performance Whitepaper

Redefining AI Infrastructure with Pioneering Innovation

The Evolving Landscape of AI Computing

Artificial Intelligence (AI) continues to spearhead a new IT infrastructure revolution, spurring the need for increasingly advanced computing solutions to match its rapid expansion and complexity. These complex applications require extreme parallel processing capabilities, which are optimally provided by Graphics Processing Units (GPUs). As the computational intensity of tasks such as machine learning, image processing, and natural language processing expands, the quest for increased GPU density within servers becomes paramount.

NVIDIA GPUs have consistently been at the forefront of AI acceleration. Yet, the challenge of incorporating sufficient GPU resources into a single server—achieving the desired GPU density—remains a challenge. Traditional servers, often limited to eight or fewer GPUs, cannot efficiently meet the growing intensity of AI workloads.

Addressing this imperative, Liqid's partnership with Dell Technologies has given rise to UltraStack, an innovative AI server solution that rises above conventional GPU density limitations. With servers containing up to 20 NVIDIA L40S GPUs, and based on trusted Dell PowerEdge R760 and R7625 servers, UltraStack delivers an efficient, flexible, and cost-effective approach for meeting and exceeding the demands of contemporary AI initiatives.

This whitepaper delves into the strategic advantages of the UltraStack solution, examining its performance through MLPerf benchmarking and NCCL tests, and discussing its potential to revolutionize AI infrastructure with enhanced GPU integration for accelerated performance, energy and cost efficiency, and rapid market deployment.

Key Customer Challenges: Performance, Cost, and Availability

As organizations strive to implement AI, they currently face three pivotal challenges that could impact their operational effectiveness and efficiency.

Performance

Limited GPU density in most servers leads to unmet computational needs for AI.

Cost

Relying on multiple low-density servers increases purchase, power, and maintenance costs.

Availability

Current GPU shortages can delay AI project timelines, impacting time to results.

Performance

Many AI deployment plans are hindered by inadequate GPU density, typically ranging from 4 to 8 per server. The dispersion of GPUs across multiple servers can result in reduced computational efficiency and performance, key concerns for intensive AI tasks such as deep learning and real-time analytics. Addressing this issue is vital to harness the full potential of AI technologies.

Cost and Complexity

Distributing GPUs across several servers not only inflates expenditures on power, cooling, and space but also adds complexity to system management, increasing the total cost of ownership. For instance, a company expanding its AI capabilities might find the operational costs spiraling unexpectedly due to the need for additional servers and their accompanying maintenance resources.

Time

The AI industry currently faces significant wait times for high-performance GPUs. Demand often exceeds supply, leading to delays in deployment. This shortage is a major concern for organizations seeking timely and efficient AI infrastructure implementation, impacting their ability to stay competitive and innovative.

Addressing these challenges is crucial for organizations, underscoring the need for innovative solutions like UltraStack that cater to performance, cost-efficiency, and availability. The subsequent sections will explore how UltraStack addresses these pressing issues, offering a groundbreaking approach to AI infrastructure.

Introducing Liqid UltraStack: The High-Density GPU Servers

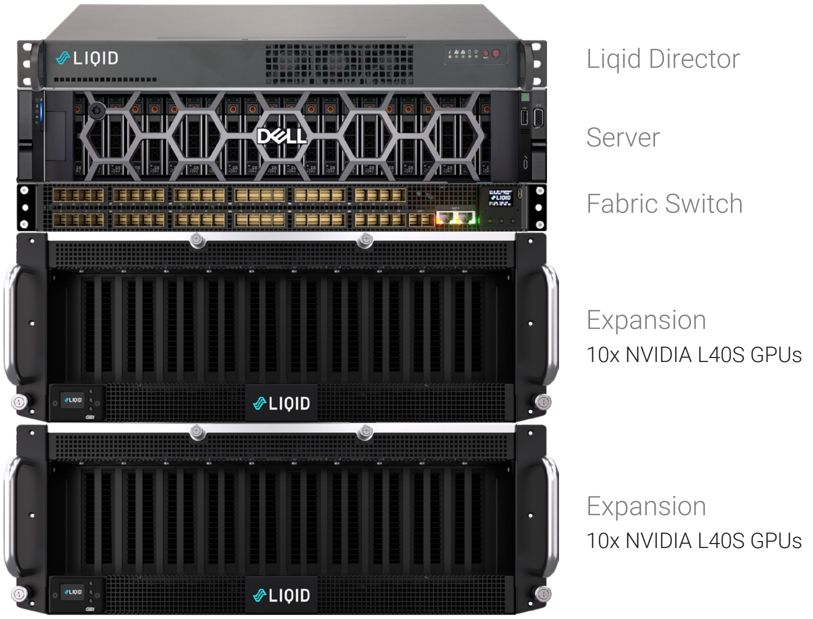
Liqid UltraStack is a groundbreaking solution to the key challenges of performance, cost, and availability in AI infrastructure deployment, focusing on high-density GPU integrated servers. The core advantage of UltraStack lies in its ability to seamlessly and transparently connect large quantities of GPUs to a standard server. This approach not only significantly enhances processing power but also provides unprecedented flexibility and scalability, enabling customers to achieve optimal GPU performance for demanding applications such as AI training and inference workloads.

UltraStack is available as stand-alone and cluster-ready servers options. The standalone servers contain up to 20 L40S GPUs. The cluster-ready servers include up to 16 L40S GPUs, in addition to Liqid IO Accelerator NVMe SSDs, NVIDIA ConnectX-7

NICs and NVIDIA BlueField-3 DPUs, for internode connectivity and efficiency in complex AI environments.

With these advanced features, Liqid UltraStack positions itself as a premier solution in the market, offering unparalleled GPU density and performance efficiency for AI-driven environments.

UltraStack UX-2020: 20-Way GPU Server



Why Liqid UltraStack for AI Workloads?

Liqid UltraStack is uniquely designed to meet the rigorous demands of AI workloads, offering unparalleled benefits in terms of performance, cost-efficiency, and reliability.

High-Density GPU Solutions for Optimized Performance

Exceptional GPU Density: UltraStack is engineered to maximize computational efficiency by incorporating up to 20x NVIDIA L40S GPUs into a Dell R760 or R7625 server. This design significantly enhances processing speeds for computationally intensive AI tasks.

Enhanced Processing Efficiency: Enhance processing efficiency with more GPUs and fewer servers. With up to 20x GPUs in a single server, GPUs and CPUs share locality, eliminating potential network bottlenecks and ensuring optimal performance.

Cost-Efficiency and Ease-of-Use

Reduced Operational Costs: Achieving target GPU density should not require purchasing additional CPUs/Servers. UltraStack's innovative design, which packs more GPUs into a single server, reduces sever footprint, for considerable savings in power consumption, cooling requirements, and space utilization. This efficient design results in a marked reduction in operational costs and total cost of ownership.

Ease of Management: The high-density configuration of Liqid UltraStack simplifies AI infrastructure by minimizing the number of servers required. This streamlined architecture makes deployment and scaling more manageable and cost-effective.

Capable and Available GPU Resources

Unparalleled AI Performance: The UltraStack utilizes the NVIDIA L40S GPU for powerful AI compute, and best in class graphics and media acceleration. With broad availability, utilizing the L40S ensures timely availability for deployments, and accelerating AI infrastructure development.

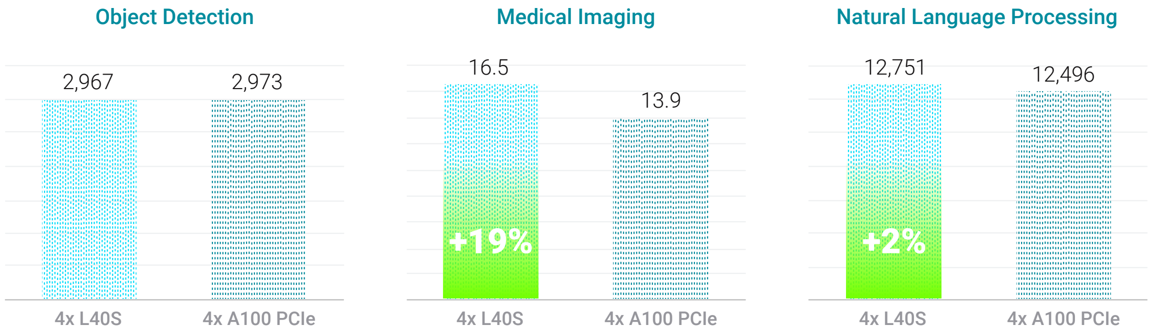
In summary, Liqid UltraStack stands as a comprehensive solution, addressing the core needs of modern AI workloads with its superior performance, cost-efficiency, and reliable GPU availability.

MLPerf Benchmark Performance: A Comparative Analysis of NVIDIA GPUs

MLPerf 3.1 Benchmark Analysis: Demonstrating the UltraStack Advantage

In the competitive landscape of AI computing, performance metrics are a pivotal factor for organizations when selecting their infrastructure. The MLPerf 3.1 benchmarks offer an industry-standard comparison of GPU performance in AI-related tasks. This section provides a comparative analysis between Liqid UltraStack integrated with NVIDIA L40S GPUs vs alternative solutions integrated with NVIDIA A100 PCIe GPUs, H100 PCIe GPUs, and H100 SXM GPUs for inference.

4x NVIDIA L40S PCIe GPUs Compared to 4x A100 PCIe GPUs: Inference



Task: Object / Data: OpenImages (800x800) / Model: Retinanet

Task: Medical Imaging / Data: KiTS19 / Model: 3D-UNet

Task: Language / Data: SQuAD v1.1 / Model: BERT

NVIDIA GPU	Object Detection Task: Object / Data: OpenImages (800x800) / Model: Retinanet	Medical Imaging Task: Medical Imaging / Data: KiTS19 / Model: 3D-UNet	Natural Language Processing Task: Language / Data: SQuAD v1.1 / Model: BERT
4x L40S PCIe GPUs	2,967	16.5 (19% Faster)	12,751 (2% Faster)
4x A100 PCIe GPUs	2,973	13.9	12,496

Table/Chart 1: MLPerf 3.1 Inference - Queries/second. Comparing 1x server with 4x L40S composed to 1x server with Server to 4x A100 PCIe GPUs composed.

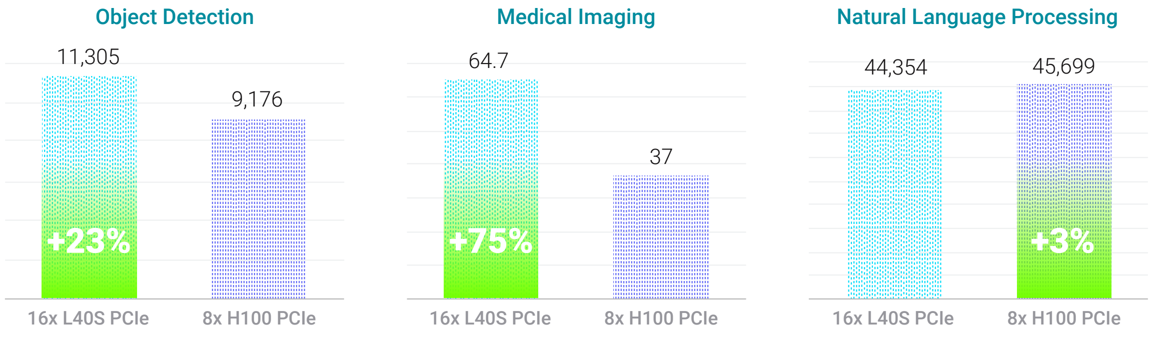
The purpose of these tests was to prove the NVIDIA L40S could at outperform the well-regarded A100 PCIe GPU in certain inference workloads. They were conducted on two servers with GPUs composed-in by Liqid Matrix. One server contained 4x L40S PCIe GPUs and the other with 4x A100 PCIe GPUs. Results indicate that the L40S GPUs not only match but in certain aspects surpass the performance of the notable NVIDIA A100 PCIe GPUs:

Object Detection: The L40S GPUs demonstrate virtually identical performance, highlighting their capability for complex visual tasks crucial in AI applications.

Medical Imaging: With a score of 16.5 over the A100's 13.9, the L40S GPUs excel in medical imaging, a field where accuracy and speed are imperative.

Natural Language Processing (NLP): The L40S GPUs outperform the A100 PCIe, evidencing superior processing of the intricate language models essential for modern AI services.

16x NVIDIA L40S PCIe GPUs Compared to 8x H100 PCIe GPUs: Inference



Task: Object / Data: OpenImages (800x800) / Model: Retinanet

Task: Medical Imaging / Data: KiTS19 / Model: 3D-UNet

Task: Language / Data: SQuAD v1.1 / Model: BERT

NVIDIA GPU	Object Detection Task: Object / Data: OpenImages (800x800) / Model: Retinanet	Medical Imaging Task: Medical Imaging / Data: KiTS19 / Model: 3D-UNet	Natural Language Processing Task: Language / Data: SQuAD v1.1 / Model: BERT
16x L40S PCIe GPUs	11,305 (23% Faster)	64.7 (75% Faster)	44,354
8x H100 PCIe GPUs	9,176	37	45,699 (3% Faster)

Table/Chart 2: MLPerf 3.1 Inference - Queries/second. Comparing 1x UltraStack w/ 16x L40S to 1x 8x H100 PCIe GPU Server.

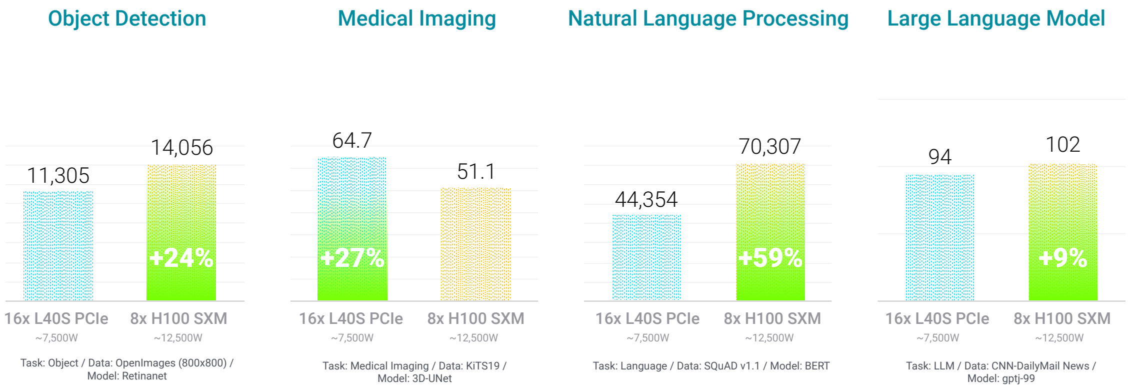
When scaling AI operations, the number of GPUs in play becomes a critical factor. In this this test, an UltraStack with 16x L40S GPUs was compared to a server with 8x H100 PCIe GPUs. In this example, the UltraStack with L40S GPUs showcases its scalability and efficiency:

Object Detection: With 16x L40S GPUs, the UltraStack delivers higher throughput in object detection tasks than the 8x H100 PCIe GPUs, a testament to its efficient scaling capabilities.

Medical Imaging: The UltraStack stands out with a significant performance advantage, suggesting an exceptional capability for the precise demands of medical AI workloads.

NLP: Despite a lower score compared to the H100 PCIe, the UltraStack's robust performance, given the higher GPU count, positions it as a viable contender for large-scale NLP implementations.

16x L40S PCIe GPUs Compared to 8x H100 SXM GPUs: Inference



NVIDIA GPU	Object Detection Task: Object / Data: OpenImages (800x800) / Model: Retinanet	Medical Imaging Task: Medical Imaging / Data: KiTS19 / Model: 3D-UNet	Natural Language Processing Task: Language / Data: SQuAD v1.1 / Model: BERT	Large Language Model Task: LLM / Data: CNN-DailyMail News / Model: gptj-99
16x L40S PCIe GPUs	11,305	64.7 (27% More)	44,354	94
8x H100 SXM GPUs	14,056 (24% More)	51.1	70,307 (59% More)	102 (9% More)

Table/Chart 3: MLPerf 3.1 Inference - Queries/second. Comparing 1x UltraStack w/ 16x L40S to 1x 8x H100 SXM GPU server.

The NVIDIA H100 SXM GPU has set the accelerator standard for AI. However, the UltraStack solution, powered by 16 NVIDIA L40S PCIe GPUs, at half of the power of the H100 SXM GPUs (~7,500W vs ~12,500W total), offers a compelling performance and efficiency profile across a range of AI tasks, that's available today. Here's an insight into the benefits it brings to different AI domains.

Object Detection: With 16x L40S GPUs, UltraStack showcases robust capabilities in object detection. Although the absolute performance is enhanced with H100 GPUs, the L40S-equipped UltraStack provides a solid foundation for detailed image analysis and recognition tasks, suitable for a variety of AI-driven applications.

Medical Imaging: The UltraStack shines in medical imaging, where it demonstrates a significant performance edge. This is particularly beneficial in healthcare AI, where the ability to accurately process and analyze medical data can directly impact diagnostic and treatment outcomes. The 22% higher score indicates the system's proficiency in managing high-resolution medical datasets.

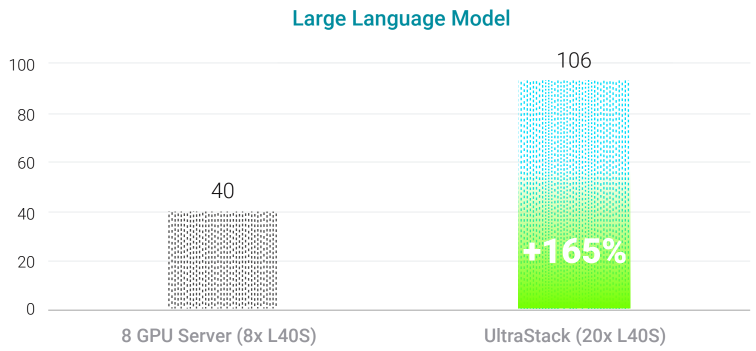
Natural Language Processing: In NLP, UltraStack equipped with L40S GPUs presents substantial performance capacity. While the H100 GPUs show higher scores, the L40S configuration offers a very competitive and capable setup for language understanding and processing tasks. This makes it suitable for businesses and research institutions looking to leverage AI for natural language applications.

Large Language Model (LLM): In the MLPerf benchmark for LLM tasks, UltraStack score is closely matched to the H100 SXM GPUs, with only a 9% difference. This highlights UltraStack's adeptness at handling the next frontier of AI applications with impressive efficiency.

Conclusion

With L40S GPUs, The UltraStack is designed to efficiently meet and exceed the demands of intensive AI workloads, offering versatility and performance. Whether it's for real-time AI analytics, advancing medical research with AI, or developing sophisticated NLP models, the UltraStack provides a powerful and reliable infrastructure, enabling organizations to push the boundaries of innovation and discovery in AI.

20x L40S PCIe GPUs Compared to 8x L40S PCIe GPUs: Inference



Task: LLM / Data: CNN-DailyMail News / Model: gptj-99

	8-GPU Server	20-GPU UltraStack
GPUs Per Server	8x L40S	20x L40S
NICs Per Server	2x	10x
GPU DRAM	384GB	960GB
CUDA Cores	145,408	363,520
ML Perf - LLM	42	106 (165% More)
SW Cost	100%	~60%
Infrastructure Power	100%	~50%

Table/Chart 4: MLPerf 3.1 Large Language Model Inference - Queries/second. Comparing 1x UltraStack w/ 20x L40S to 1x 8x L40s PCIe GPU server (MLCommons ID: 3.1-0141)

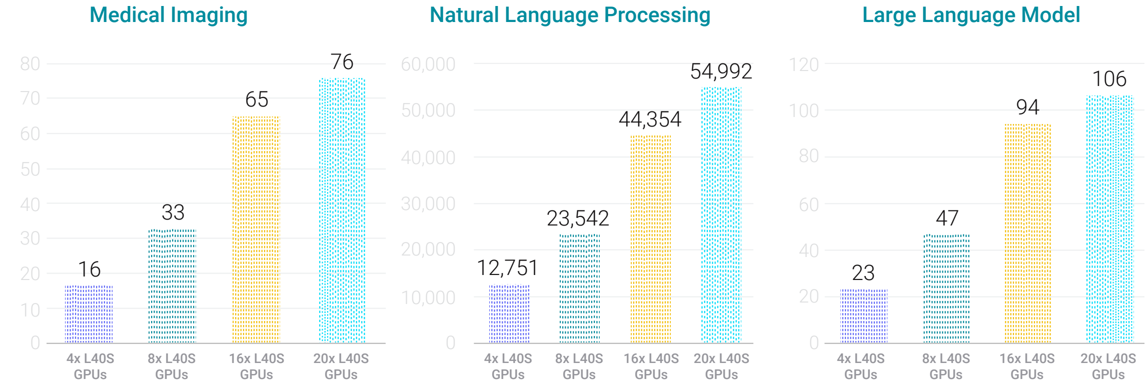
The data at hand not only illustrates a substantial leap in performance but also underscores the efficiency gains when scaling up with UltraStack's 20x L40S PCIe GPUs compared to a standard 8-GPU server option.

Performance Leap: With 20x L40S GPUs, the number of CUDA cores more than doubles, offering a massive parallel processing capability that translates into a 165% increase in ML Perf LLM inference scores. This jump in performance metrics signifies a notable enhancement in the server's ability to handle inference workloads, providing faster, more accurate AI insights.

Efficiency Advantage: The efficiency of UltraStack is evident in its software and infrastructure power costs, both of which are approximately 50-60% less than the 8-GPU server configuration. This reduction is a testament to the architectural optimizations and the inherent design efficiencies of UltraStack, which ensure that the increased computational power does not come at the cost of proportionally increased expenses.

By integrating more GPUs within a single server unit, UltraStack achieves greater computational density without a linear increase in cost or power consumption, paving the way for smarter, more sustainable AI infrastructure development.

UltraStack Delivers Near-Linear Performance Scaling


Medical Imaging

Natural Language Processing

Large Language Model

UltraStack Servers	Medical Imaging Task: Medical Imaging / Data: KiTS19 / Model: 3D-UNet	Natural Language Processing Task: Language / Data: SQuAD v1.1 / Model: BERT	Large Language Model Task: LLM / Data: CNN-DailyMail News / Model: gptj-99
4x L40S GPUs	16	12,751	23
8x L40S GPUs	33	23,542	47
16x L40S GPUs	65	44,354	94
20x L40S GPUs	76	54,992	106

A key factor in high-performance computing, particularly within AI-driven applications, is the ability to scale resources efficiently. The UltraStack platform showcases near-linear scalability, as evidenced by MLPerf benchmark inferencing results across 4-way, 8-way, 16-way and 20-way configurations for tasks in Medical Imaging, Natural Language Processing (NLP), and Large Language Model (LLM) workloads.

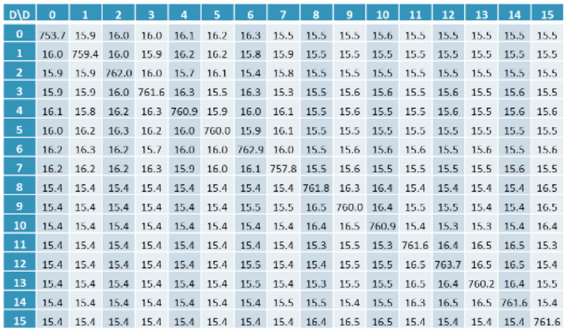
From a 4-way to a 20-way GPU configuration, the UltraStack consistently demonstrates a proportional increase in performance, with MLPerf results underscoring the system's ability to expand its processing capabilities almost directly in line with GPU quantity increases.

Peer-to-Peer (P2P) Performance: Revolutionizing Data Transfer with RDMA

In the realm of high-performance computing and AI, the efficiency of data transfer between components is as critical as the processing power itself. Liqid UltraStack addresses this with its RDMA Peer-to-Peer (P2P), for GPU-to-GPU communication, as well as between storage and GPUs, all over native PCIe. The following bandwidth and latency tests were run on a single Liqid UltraStack with 16x NVIDIA L40S GPUs, with P2P enabled. P2P increased bandwidth (GB/s) by 222% and reduces latency by 91%, for inter-GPU communication, compared when P2P is disabled (traverses the CPU).

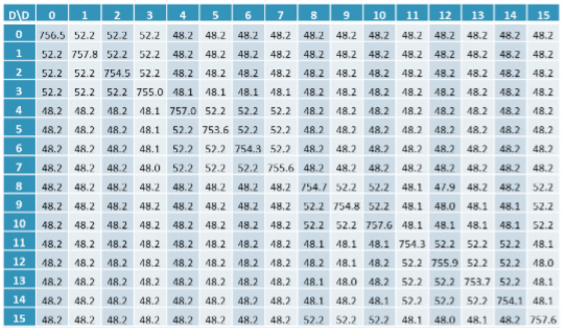
Increased bandwidth by 222%

Table 1: Bidirectional P2P=Disabled Bandwidth Matrix (GB/s), Average = 15.2GB/s



D\D	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	753.7	15.9	16.0	16.0	16.1	16.2	16.3	15.5	15.5	15.5	15.6	15.5	15.5	15.5	15.5	15.5
1	16.0	759.4	16.0	15.9	16.2	16.2	15.8	15.9	15.5	15.5	15.5	15.5	15.5	15.5	15.5	15.5
2	15.9	15.9	762.0	16.0	15.7	16.1	15.4	15.8	15.5	15.5	15.5	15.5	15.5	15.5	15.5	15.5
3	15.9	15.9	16.0	761.6	16.3	15.5	16.3	15.3	15.5	15.6	15.6	15.5	15.6	15.5	15.6	15.5
4	16.1	15.8	16.2	16.3	760.9	15.9	16.0	16.1	15.5	15.6	15.5	15.5	15.6	15.5	15.6	15.6
5	16.0	16.2	16.3	16.2	16.0	760.0	15.9	16.1	15.5	15.5	15.5	15.5	15.5	15.5	15.5	15.5
6	16.2	16.3	16.2	15.7	16.0	16.0	762.9	16.0	15.5	15.6	15.6	15.6	15.5	15.6	15.5	15.6
7	16.2	16.2	16.2	16.3	15.9	16.0	16.1	757.8	15.5	15.6	15.5	15.5	15.5	15.5	15.6	15.5
8	15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.4	761.8	16.3	16.4	15.4	15.4	15.4	15.4	16.5
9	15.4	15.4	15.4	15.4	15.4	15.4	15.5	15.5	16.5	760.0	16.4	15.5	15.5	15.4	15.4	16.5
10	15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.4	16.4	16.5	760.9	15.4	15.3	15.3	15.4	16.4
11	15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.3	15.5	15.3	761.6	16.4	16.5	16.5	15.3
12	15.4	15.4	15.4	15.4	15.4	15.4	15.5	15.4	15.4	15.5	15.5	16.5	763.7	16.5	16.5	15.4
13	15.4	15.4	15.4	15.4	15.4	15.4	15.5	15.4	15.3	15.5	15.5	16.5	16.4	760.2	16.4	15.5
14	15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.5	15.5	15.4	15.5	16.3	16.5	16.5	761.6	15.4
15	15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.4	16.4	16.5	16.5	15.4	15.4	15.4	15.4	761.6

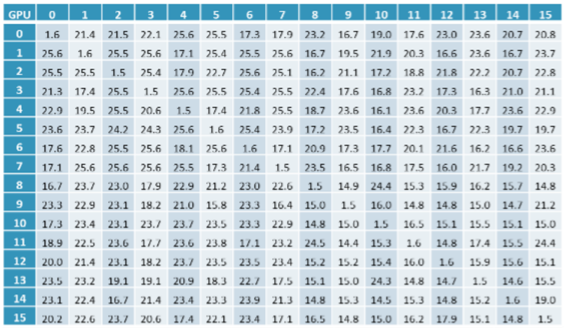
Table 2: Bidirectional P2P=Enabled Bandwidth Matrix (GB/s), Average = 49GB/s



D\D	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	756.5	52.2	52.2	52.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
1	52.2	757.8	52.2	52.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
2	52.2	52.2	754.5	52.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
3	52.2	52.2	52.2	755.0	48.1	48.1	48.1	48.1	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
4	48.2	48.2	48.2	48.1	757.0	52.2	52.2	52.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
5	48.2	48.2	48.2	48.1	52.2	753.6	52.2	52.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
6	48.2	48.2	48.2	48.1	52.2	52.2	754.3	52.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
7	48.2	48.2	48.2	48.0	52.2	52.2	52.2	755.6	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2
8	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	754.7	52.2	52.2	48.1	47.9	48.2	48.2	52.2
9	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	52.2	754.8	52.2	48.1	48.0	48.1	48.1	52.2
10	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	52.2	52.2	757.6	48.1	48.1	48.1	48.1	52.2
11	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.1	48.1	48.1	754.3	52.2	52.2	52.2	48.1
12	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.1	48.2	52.2	755.9	52.2	52.2	48.0
13	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.1	48.0	48.2	52.2	52.2	753.7	52.2	48.1
14	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.1	48.2	48.1	52.2	52.2	52.2	754.1	48.1
15	48.2	48.2	48.2	48.2	48.2	48.2	48.2	48.2	52.2	52.2	52.2	48.1	48.0	48.1	48.2	757.6

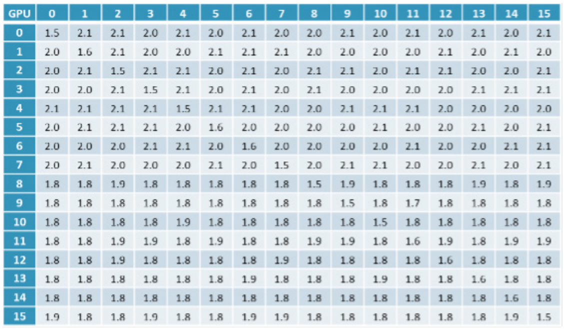
Reduced Latency by 91%

Table 3: P2P=Disabled Latency Matrix (us), Average = 20us



GPU	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1.6	21.4	21.5	22.1	25.6	25.5	17.3	17.9	23.2	16.7	19.0	17.6	23.0	23.6	20.7	20.8
1	25.6	1.6	25.5	25.6	17.1	25.4	25.5	25.6	16.7	19.5	21.9	20.3	16.6	23.6	16.7	23.7
2	25.5	25.5	1.5	25.4	17.9	22.7	25.6	25.1	16.2	21.1	17.2	18.8	21.8	22.2	20.7	22.8
3	21.3	17.4	25.5	1.5	25.6	25.5	25.4	25.5	22.4	17.6	16.8	23.2	17.3	16.3	21.0	21.1
4	22.9	19.5	25.5	20.6	1.5	17.4	21.8	25.5	18.7	23.6	16.1	23.6	20.3	17.7	23.6	22.9
5	23.6	23.7	24.2	24.3	25.6	1.6	25.4	23.9	17.2	23.5	16.4	22.3	16.7	22.3	19.7	19.7
6	17.6	22.8	25.5	25.6	18.1	25.6	1.6	17.1	20.9	17.3	17.7	20.1	21.6	16.2	16.6	23.6
7	17.1	25.6	25.6	25.6	25.5	17.3	21.4	1.5	23.5	16.5	16.8	17.5	16.0	21.7	19.2	20.3
8	16.7	23.7	23.0	17.9	22.9	21.2	23.0	22.6	1.5	14.9	24.4	15.3	15.9	16.2	15.7	14.8
9	23.3	22.9	23.1	18.2	21.0	15.8	23.3	16.4	15.0	1.5	16.0	14.8	14.8	15.0	14.7	21.2
10	17.3	23.4	23.1	23.7	23.7	23.5	23.3	22.9	14.8	15.0	1.5	16.5	15.1	15.5	15.1	15.0
11	18.9	22.5	23.6	17.7	23.6	23.8	17.1	23.2	24.5	14.4	15.3	1.6	14.8	17.4	15.5	24.4
12	20.0	21.4	23.1	18.2	23.7	23.5	23.5	23.4	15.2	15.2	15.4	16.0	1.6	15.9	15.6	15.1
13	23.5	23.2	19.1	19.1	20.9	18.3	22.7	17.5	15.1	15.0	24.3	14.8	14.7	1.5	14.6	15.5
14	23.1	22.4	16.7	21.4	23.4	23.3	23.9	21.3	14.8	15.3	14.5	15.3	14.8	15.2	1.6	19.0
15	20.2	22.6	23.7	20.6	17.4	22.1	23.4	17.1	16.5	14.8	15.0	16.2	17.9	15.1	14.8	1.5

Table 4: P2P=Enabled Latency (P2P Writes) Matrix (us), Average = 1.9us



GPU	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1.5	2.1	2.1	2.0	2.1	2.0	2.1	2.0	2.0	2.1	2.0	2.1	2.0	2.1	2.0	2.1
1	2.0	1.6	2.1	2.0	2.0	2.1	2.1	2.1	2.0	2.0	2.0	2.0	2.1	2.0	2.1	2.0
2	2.0	2.1	1.5	2.1	2.1	2.0	2.1	2.0	2.1	2.1	2.0	2.1	2.1	2.0	2.0	2.1
3	2.0	2.0	2.1	1.5	2.1	2.0	2.1	2.0	2.1	2.0	2.0	2.0	2.0	2.1	2.1	2.1
4	2.1	2.1	2.1	2.1	1.5	2.1	2.1	2.0	2.0	2.1	2.1	2.1	2.0	2.0	2.0	2.0
5	2.0	2.1	2.1	2.1	2.0	1.6	2.0	2.0	2.0	2.0	2.1	2.0	2.0	2.1	2.0	2.1
6	2.0	2.0	2.0	2.1	2.1	2.0	1.6	2.0	2.0	2.0	2.0	2.1	2.0	2.0	2.1	2.1
7	2.0	2.1	2.0	2.0	2.0	2.1	2.0	1.5	2.0	2.1	2.1	2.0	2.0	2.1	2.0	2.1
8	1.8	1.8	1.9	1.8	1.8	1.8	1.8	1.8	1.5	1.9	1.8	1.8	1.8	1.9	1.8	1.9
9	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.5	1.8	1.7	1.8	1.8	1.8	1.8
10	1.8	1.8	1.8	1.8	1.9	1.8	1.8	1.8	1.8	1.8	1.5	1.8	1.8	1.8	1.8	1.8
11	1.8	1.8	1.9	1.9	1.8	1.9	1.8	1.8	1.9	1.9	1.8	1.6	1.9	1.8	1.9	1.9
12	1.8	1.8	1.9	1.8	1.8	1.8	1.8	1.9	1.8	1.8	1.8	1.8	1.6	1.8	1.8	1.8
13	1.8	1.8	1.8	1.8	1.8	1.8	1.9	1.8	1.8	1.8	1.9	1.8	1.8	1.6	1.8	1.8
14	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.6	1.8
15	1.9	1.8	1.8	1.9	1.8	1.8	1.9	1.9	1.8	1.8	1.8	1.8	1.8	1.8	1.9	1.5

The results from our P2P bandwidth and latency tests demonstrate substantial benefits for environments where intensive data processing and high throughput are crucial. This technology not only enhances the overall performance of UltraStack but also opens new horizons for handling the most demanding computational tasks with unprecedented efficiency and speed.

Harnessing the Power of NCCL in UltraStack with NVIDIA L40S GPUs

UltraStack's integration with NVIDIA L40S GPUs and its utilization of NVIDIA Collective Communications Library (NCCL) represents a significant advancement in GPU intercommunication capabilities. Our recent NCCL test results offer compelling insights into the efficiency and speed of data processing achievable in UltraStack.

NCCL Test Overview and Results

The NCCL (NVIDIA Collective Communications Library) test serves as a rigorous benchmark for assessing the communication performance of GPU clusters in collective operations, which are vital in parallel processing and especially crucial in deep learning environments. The "all_reduce_perf" test, conducted using an UltraStack with 16x NVIDIA L40S GPUs, measures collective communication over a range of substantial data sizes, from 1GB to 8GB. Such operations are at the heart of synchronizing data among multiple GPUs, a common requirement in training large and complex neural networks.

Results

```
CUDA_VISABLE_GPUS=0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
./build/all_reduce_perf -b 1G -e 8G -f 2 -g 16
# nThread 1 nGpus 16 minBytes 1073741824 maxBytes 8589934592 step: 2(factor) warmup
iters: 5 iters: 20 agg iters: 1 validation: 1 graph: 0
#
# Using devices
#  Rank  0 Group  0 Pid  39608 on r7625-04-48 device  0 [0x27] NVIDIA L40S
#  Rank  1 Group  0 Pid  39608 on r7625-04-48 device  1 [0x28] NVIDIA L40S
#  Rank  2 Group  0 Pid  39608 on r7625-04-48 device  2 [0x29] NVIDIA L40S
#  Rank  3 Group  0 Pid  39608 on r7625-04-48 device  3 [0x2a] NVIDIA L40S
#  Rank  4 Group  0 Pid  39608 on r7625-04-48 device  4 [0x2b] NVIDIA L40S
#  Rank  5 Group  0 Pid  39608 on r7625-04-48 device  5 [0x2c] NVIDIA L40S
#  Rank  6 Group  0 Pid  39608 on r7625-04-48 device  6 [0x2d] NVIDIA L40S
#  Rank  7 Group  0 Pid  39608 on r7625-04-48 device  7 [0x2e] NVIDIA L40S
#  Rank  8 Group  0 Pid  39608 on r7625-04-48 device  8 [0xef] NVIDIA L40S
#  Rank  9 Group  0 Pid  39608 on r7625-04-48 device  9 [0xf0] NVIDIA L40S
#  Rank 10 Group  0 Pid  39608 on r7625-04-48 device 10 [0xf1] NVIDIA L40S
#  Rank 11 Group  0 Pid  39608 on r7625-04-48 device 11 [0xf2] NVIDIA L40S
#  Rank 12 Group  0 Pid  39608 on r7625-04-48 device 12 [0xf3] NVIDIA L40S
#  Rank 13 Group  0 Pid  39608 on r7625-04-48 device 13 [0xf4] NVIDIA L40S
#  Rank 14 Group  0 Pid  39608 on r7625-04-48 device 14 [0xf5] NVIDIA L40S
#  Rank 15 Group  0 Pid  39608 on r7625-04-48 device 15 [0xf6] NVIDIA L40S
#
#                                                              out-of-place
```

```
in-place
#       size         count      type   redop    root     time   algbw   busbw #wrong
time   algbw   busbw #wrong
#        (B)    (elements)                               (us)  (GB/s)  (GB/s)
(us)  (GB/s)  (GB/s)
 1073741824     268435456     float     sum      -1    95293   11.27   21.13      0
95273   11.27   21.13      0
 2147483648     536870912     float     sum      -1   190495   11.27   21.14      0
190444   11.28   21.14      0
 4294967296    1073741824     float     sum      -1   381066   11.27   21.13      0
380945   11.27   21.14      0
 8589934592    2147483648     float     sum      -1   762294   11.27   21.13      0
762002   11.27   21.14      0
# Out of bounds values : 0 OK
# Avg bus bandwidth    : 21.1346
```

Key Findings

Stable Bandwidth Across Various Data Sizes

Across the range of tested data sizes (1GB to 8GB), the bandwidth remained consistently high. The algorithmic bandwidth (algbw) was approximately 11.27 GB/s, and the bus bandwidth (busbw) was around 21.13 to 21.14 GB/s. This consistency is indicative of UltraStack's capability to handle large and complex data sets efficiently.

Efficient Data Aggregation

The test used a 'float' data type and 'sum' reduction operation, typical in many high-performance computing tasks. The results show that UltraStack can aggregate data across multiple GPUs efficiently, a critical factor in reducing training times for machine learning models or accelerating computation in data-intensive applications.

Scalability

The consistent performance across different data sizes also highlights the scalability of UltraStack when coupled with NCCL. This scalability ensures that performance gains are maintained even as the data size or complexity increases, making it suitable for a wide range of high-demand applications.

Implications for UltraStack in High-Demand Environments

In the emerging AI world where Job Completion Times are crucially important, these NCCL test results are not just numbers; they translate into real-world performance gains. For instance, in deep learning, these results mean faster model training times, leading to more rapid iterations and development. In scientific computing, it translates to quicker simulations and data processing, enabling researchers to achieve more in less time.

Conclusion: UltraStack with L40S GPUs – The Optimal AI Infrastructure Choice

UltraStack, featuring NVIDIA L40S GPUs, represents a balanced solution adept at meeting the intricate demands of AI infrastructure. It offers a blend of high performance, energy efficiency, and market readiness, positioning itself as an attractive option for organizations navigating the expanding realm of AI.

Performance Tailored for AI: Liqid UltraStack with L40S GPUs have demonstrated through MLPerf benchmarks their capability to efficiently manage a spectrum of AI tasks, from object detection to natural language processing, ensuring that computational power is never a bottleneck.

Energy and Cost Efficiency: With an energy-efficient design, the L40S GPUs enable UltraStack to deliver top-tier performance without the high energy costs, aligning with the financial and environmental considerations of businesses today.

Availability for Timely Deployment: Addressing GPU availability challenges, UltraStack with L40S GPUs is readily available, facilitating swift deployment and scalability for organizations to maintain competitive momentum.

The UltraStack Edge: More than just raw power, UltraStack is a solution built for the realities of enterprise demands, combining performance where it's needed most, cost-effectiveness where it counts, and GPU availability that accelerates time to value. For AI initiatives that require superior performance, higher efficiency, and rapid deployment, UltraStack with L40S GPUs is the strategic choice.

As AI reshapes industry landscapes, UltraStack with L40S GPUs is poised to accelerate businesses, enabling them to harness innovation, maintain sustainability, and lead the market with agility. Please visit Liqid.com to learn more.

HEADQUARTERS

11400 Westmoor Circle
Ste 225
Westminster, CO 80021

ONLINE

www.liqid.com
info@liqid.com

TELEPHONE

+ 1 303.500.1551