Case Study

VIAVI Solutions

Major Hyperscaler Builds AI-Optimized Data Center

Optimizing AI Infrastructure for Maximum Reliability and Efficiency

Customer Challenge

A major global hyperscaler was building a new AI data center designed to support the increasing computational demands of modern AI workloads, including large-scale deep learning models, generative AI applications, and real-time inference systems. With more than 20,000 GPUs and AI accelerators deployed across its data center, the customer faced a series of critical operational challenges:

- **Network Performance Bottlenecks:** Ensuring high-bandwidth, low-latency communication across a vast, GPU-dense architecture.
- **AI Workload Optimization:** Handling AI-specific workloads that demand seamless scaling and synchronization across distributed compute resources.
- **Resource Utilization Concerns:** High-cost GPU resources were being underutilized due to inefficient workload scheduling and unoptimized networking.
- **Lack of Testing and Visibility:** Prior testing relied on limited real-world benchmarking, making it difficult to detect inefficiencies and proactively optimize infrastructure.
- **High Manual Testing Costs:** The hyperscaler initially relied on actual GPU-based systems and custom software to manually create AI workloads and traffic patterns. This approach was costly to create, manage, and update, requiring extensive time and resources.

The hyperscaler needed a scalable, high-fidelity testing solution that could **emulate real AI workloads at full scale and provide actionable insights into key performance indicators** (KPIs) like latency, packet loss, tail latency, and job completion time.

Highlights

- A global hyperscaler sought to optimize its next-generation AI data center, incorporating 20,000+ GPUs and AI accelerators.
- They faced challenges in network congestion, workload optimization, and performance bottlenecks for cutting-edge AI applications.
- VIAVI AI Workload Emulation Solution provided advanced GPU emulation and AI workload traffic testing, enabling precise evaluation of network performance.
- Provided key network performance, capacity, and issue visibility via active traffic emulation tests with actionable test results, enabling users to make data-driven improvements.
- Achieved significant cost savings by improving GPU utilization efficiency, reducing unnecessary hardware expenditures.

Solution Requirements

To maximize the efficiency and cost-effectiveness of its AI data center, the hyperscaler sought a testing solution that could:

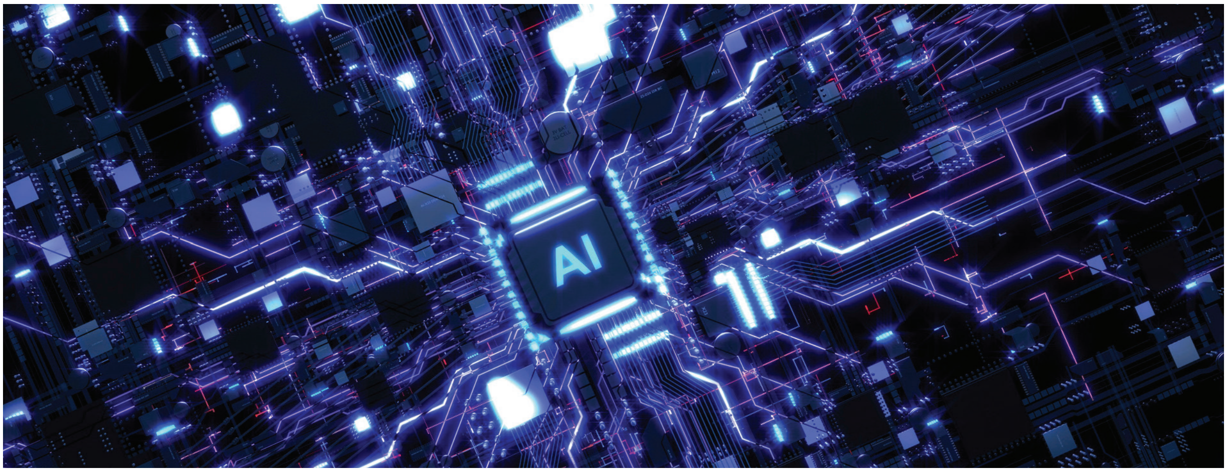
- **Emulate AI Workloads:** Accurately simulate the network impact of real AI training and inference workloads.
- **Assess Network Scalability:** Validate performance under high-throughput, distributed AI processing.
- **Analyze GPU Utilization:** Ensure that expensive GPU resources were optimally deployed and not bottlenecked by network inefficiencies.
- **Identify Performance Issues:** Provide granular visibility into latency, packet loss, and other key AI-specific network metrics.
- **Reduce Operational Costs:** Improve overall infrastructure efficiency to minimize unnecessary capital and operational expenditures.
- **Enhance AI Infrastructure Automation:** Ensure that infrastructure adjustments and optimizations could be rapidly deployed and tested.

The Solution

The hyperscaler partnered with VIAVI to deploy a comprehensive AI data center testing framework, leveraging VIAVI AI Workload Emulation Solution to ensure optimal performance at scale.

Key components of the solution included:

- **AI Workload Traffic Emulation:** Simulated realistic AI data flows to measure network responsiveness and optimize throughput.
- **xPU Utilization Testing:** Evaluated how network performance affected xPU efficiency and job completion times.
- **Latency and Tail Latency Monitoring:** Provided real-time visibility into network-induced processing delays.
- **Packet Loss and Network Congestion Analysis:** Identified potential failure points in the network to prevent performance degradation.
- **Automated KPI Benchmarking:** Continuously measured and analyzed AI-specific KPIs, providing actionable insights into data center performance.
- **Scalable Testing Capacity:** Allowed for cost-effective scaling of test environments using additional VIAVI appliances.
- **Adaptive Testing Automation:** Enabled real-time tuning of testing methodologies, enhancing responsiveness to AI workload variations.
- **Diverse AI Workload Generation:** Created a multitude of specific AI workload patterns from an expanding set of workload types, including Collective Communications Library (CCL) AI Traffic Patterns such as AlltoAll, RingAllReduce, Halving and Doubling, as well as deep learning model training, real-time inference processing, natural language processing (NLP) tasks, large-scale recommendation engine simulations, and AI-driven image recognition workloads.
- **AI Fabric Optimization:** Enabled users to characterize the performance of AI fabric and optimize the relevant network configurations, including buffer size, ECN, load-balancing algorithm, and QoS settings, resulting in significant impact and improvement of AI data center device operations and resilience.



VIAVI AI Workload Emulation Solution allows users to test networking under a variety of scenarios to understand and model performance, capacity, and other critical AI traffic KPIs. With this information, users can establish a baseline performance, make necessary adjustments, and retest to ensure continuous improvement in AI data center operations. Unlike the previous manual approach, the solution enables users to create tests in minutes rather than days and scale test capacity by cost-effectively adding additional appliances. This allows for more accurate, efficient testing with reduced capital (CapEx) and operational (OpEx) resources.

A key insight revealed through testing was that less than 1% packet loss can cause a 33% drop in xPU utilization, leading to severe performance and cost inefficiencies. Addressing such critical issues ensures AI data centers operate at maximum efficiency while minimizing operational costs.

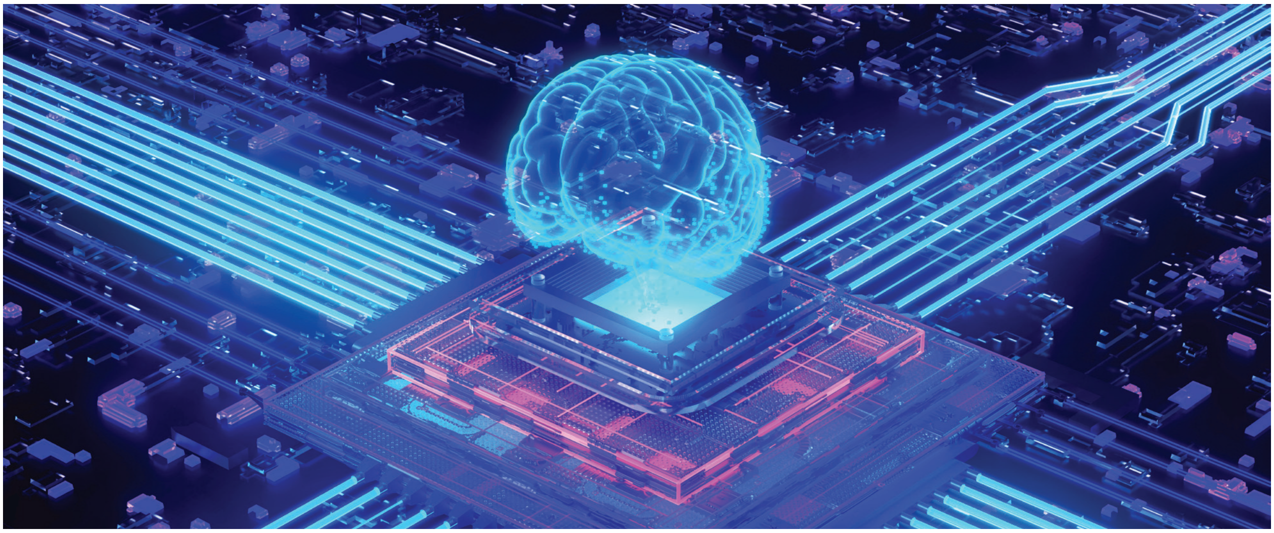
Additionally, the VIAVI tool provides not only a comprehensive solution for AI data center traffic emulation and analytics but also serves as a powerful L2-3 testing platform, significantly improving the return on investment (ROI). This dual capability ensures that users not only optimize AI workloads but also enhance overall network infrastructure testing at scale.

By integrating this solution, the hyperscaler gained a data-driven approach to optimizing network performance and xPU efficiency before full-scale deployment.

Solution Benefits

VIAVI test solution provided tangible performance improvements and significant cost savings:

- **50% Reduction in Network-Related Bottlenecks:** Optimized GPU-to-GPU communication for enhanced AI training speeds.
- **30% Improvement in GPU Utilization:** Maximized computational efficiency, reducing hardware procurement costs.
- **25% Decrease in Job Completion Time:** Enabled faster AI model training and inference execution.
- **Enhanced Network Resilience:** Minimized packet loss and congestion-related slowdowns.
- **Millions in Cost Savings:** Reduced the need for over-provisioning expensive AI accelerators and network infrastructure.
- **Rapid Deployment of AI Testing Models:** Ensured faster validation of AI workloads and infrastructure configurations.



Conclusion

As AI workloads continue to evolve, AI-centric data centers must be optimized for scale, efficiency, and cost-effectiveness. This major hyperscaler leveraged VIAVI's cutting-edge AI workload emulation and performance testing solution to ensure that its 20,000 GPU deployment delivered unprecedented efficiency and ROI.

Through real-world AI traffic emulation, in-depth KPI analysis, and optimized network tuning, the hyperscaler transformed its AI data center into a highly efficient, cost-effective, and scalable environment—ready to power the next generation of AI breakthroughs.

VIAVI

Contact Us: +1 844 GO VIAVI | (+1 844 468 4284). To reach the VIAVI office nearest you, visit viavisolutions.com/contact

viavisolutions.com