Solution overview Cisco public

Cisco – Liqid Solutions Paper

Redefining Possibilities:
Powering Tomorrow's GPU Workloads with Cisco and Liqid

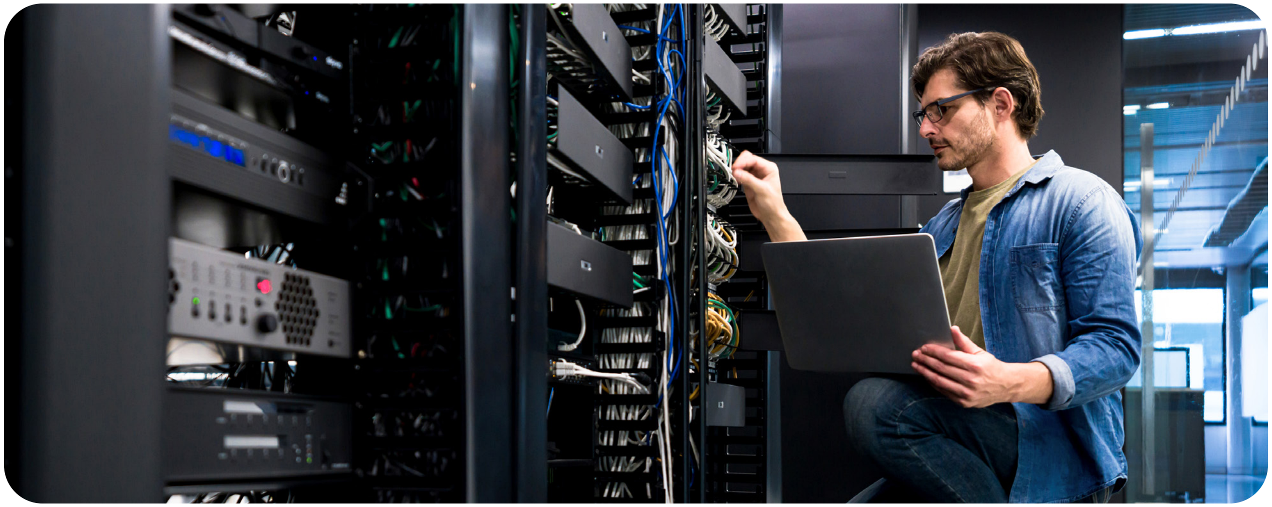
Contents

1. Introduction

Imagine an infrastructure where every watt is maximized, every GPU is fully utilized, and deploying new workloads is as effortless as clicking a button. The collaboration between Cisco UCS® servers and Liqid's PCIe composable infrastructure transforms that vision into reality. Together, we're breaking the mold of traditional data centers—delivering unprecedented GPU density and agility within a 600W power envelope, tailored for the most demanding AI, HPC, and edge inference challenges. The future of composable data centers isn't coming; it's already here.

2. The challenge: unlocking GPU potential in the AI era

Modern AI, High-Performance Computing (HPC), and data-intensive workloads demand immense GPU power. However, traditional server architectures often lead to GPU underutilization, power inefficiencies, and rigid infrastructure that struggles to adapt to rapidly changing demands. Integrating high-wattage GPUs (such as 600W models) into standard servers presents significant power, cooling, and space constraints, limiting density and driving up costs. Organizations need a flexible, scalable, and efficient solution to maximize their GPU investments and accelerate innovation.



Contents

3. Overview of Cisco UCS and Liqid PCIe solutions

Cisco UCS provides a robust and validated server platform designed to support dense GPU configurations with models such as the Cisco UCS C220 M8 and Cisco C240 M8 rack servers. These servers feature PCIe Gen5 risers, enabling integration of high-performance GPUs such as NVIDIA H100, L40S, and others. The UCS architecture unifies compute, networking, and storage to deliver a comprehensive AI solution that maximizes container density and resource efficiency, making it ideal for AI model training and inference workloads.

Complementing Cisco UCS, Liqid offers a composable infrastructure solution that utilizes PCIe Gen5 expansion chassis, switches, and director software to enable dynamic GPU pooling and sharing across multiple servers. Liqid's platform supports up to 10 PCIe devices per chassis with 64GB/s throughput and Gen5 x16 host interface cards. The Liqid Matrix software orchestrates device management and integrates seamlessly with orchestration platforms such as VMware, SLURM, and Kubernetes. This software-defined composability enhances GPU utilization and operational efficiency while optimizing power consumption.

Together, Cisco UCS and Liqid's composable PCIe fabric enable high-density GPU configurations providing 600W capacity, delivering best-in-class operations per watt and per dollar. Liqid's SmartStack technology further allows multiple servers to access composable GPU pools on demand, supporting silicon diversity and dynamic GPU assignment. This flexibility improves the return on investment by driving near 100-percent GPU utilization and extends the life of legacy infrastructure. At the edge, the composable PCIe fabric facilitates GPU-as-a-service models with dynamic allocation and scaling, optimizing critical metrics such as tokens per watt and tokens per dollar for inference workloads.

This integrated solution offers a scalable, power-efficient, and highly manageable infrastructure that meets the evolving demands of GPU-intensive applications across enterprise and edge environments.

Contents

4. Benefits of this solution from Cisco and Liqid

The Cisco UCS and Liqid PCIe solution deliver a transformative set of benefits:

- **Unmatched GPU utilization:** Achieve up to 100 percent GPU utilization by dynamically allocating resources as needed, eliminating costly GPU stranding and overprovisioning. This translates to a 2x to 4x increase in data-center resource utilization.
- **Superior power and cost efficiency:** Optimize performance per watt and per dollar, achieving up to 2x more tokens per watt and 50 percent higher tokens per dollar for AI applications. Reduce infrastructure costs by up to 50 percent by avoiding overprovisioning and extending hardware lifecycles.
- **Dynamic agility and scalability:** Instantly provision and migrate GPU resources to any server through software, without physical intervention or reboots. Scale up to 30 GPUs per host or share 30 GPUs across 20 servers, adapting to evolving workload demands in real time.
- **Accelerated time-to-value:** Rapidly deploy and reconfigure bare-metal server systems in seconds, accelerating project completion and innovation cycles.
- **Make infrastructure future-ready:** Leverage an open, standards-based architecture (PCIe Gen5) that supports multi-vendor GPUs and integrates seamlessly with existing orchestration tools like Kubernetes, Slurm, and VMware.

Overview: Cisco UCS and Liqid PCIe composable infrastructure

This solution combines the robust, high-performance compute capabilities of Cisco UCS servers with Liqid's software-defined composable infrastructure over a low-latency PCIe fabric.

- **Cisco UCS servers:** Cisco UCS rack servers provide a high-performance compute platform designed for the AI era. These servers offer independent scaling of CPUs and memory, supporting GPU-accelerated workloads and providing the foundational compute power for demanding applications.
- **Liqid PCIe composable infrastructure:** At its core, Liqid's solution disaggregates physical resources—including GPUs and NVMe storage pools that are connected through a high-speed PCIe Gen5 fabric. The Liqid Matrix software platform intelligently orchestrates these pooled resources, allowing them to be dynamically composed and reconfigured into bare-metal servers on demand.

Contents

The total solution: composable power for AI and HPC

This integration of Cisco® and Liqid solutions creates a highly flexible and efficient data-center environment:

- **Disaggregated resource pools:** Cisco UCS servers act as the intelligent compute nodes while the Liqid chassis house disaggregated pools of high-wattage GPUs and NVMe storage. These resources are connected through a low-latency PCIe fabric, allowing them to be shared across multiple servers.
- **Software-defined orchestration:** Liqid Matrix software provides a unified management interface to dynamically allocate these pooled resources to Cisco UCS servers. This means a Cisco UCS server can be instantly provisioned with the exact number and type of GPUs required for a specific workload, then have those GPUs returned to the pool when the task is complete.
- **Optimized for demanding workloads:** This architecture is ideal for AI training, inference, HPC simulations, VDI, and real-time analytics, where dynamic resource allocation and maximum GPU utilization are critical.

The advantage of 600W GPUs in a Liqid chassis

Integrating high-wattage GPUs (for example, 600W NVIDIA H200, B200, and RTX Pro 6000 and Intel® Gaudi 3) into a Liqid composable chassis offers significant advantages over traditional server-based deployments:

- **Purpose-built power and cooling:** Liqid chassis are engineered specifically to handle the substantial power and thermal requirements of multiple 600W GPUs. Unlike general-purpose servers, which have physical and thermal limitations that restrict the number of high-power GPUs they can accommodate, a dedicated Liqid chassis provides superior cooling and power delivery for high-density GPU configurations.
- **Higher density and footprint efficiency:** By centralizing GPUs in a dedicated chassis, the solution achieves far greater GPU density than is possible within individual servers. Liqid can compose up to 30 GPUs for a single host, or share them across multiple hosts, maximizing performance within a smaller physical footprint.

Contents

- **True disaggregation and independent scaling:** When 600W GPUs are in a Liqid chassis, they are truly disaggregated from the server. This means GPUs can be upgraded, replaced, or scaled independently of the server lifecycle. You can refresh your GPUs without replacing entire servers, and vice-versa, extending the lifespan of your investments.
- **Maximized utilization, minimized stranding:** A 600W GPU is an expensive asset. Placing it in a server means it is tied to that server, potentially sitting idle if the server's CPU or other resources are not fully utilized. In a Liqid chassis, these powerful GPUs can be dynamically attached to any available Cisco UCS server, ensuring they are always active and contributing to a workload, drastically reducing GPU stranding.
- **Simplified management and upgrades:** Managing and upgrading GPUs in a composable chassis is simpler. Instead of opening multiple servers, IT teams can manage a centralized pool of GPUs through software, streamlining operations, and reducing maintenance complexity.

Quantifiable Return on Investment (ROI) for the customer

The Cisco UCS and Liqid PCIe solution delivers tangible ROI by addressing both Capital Expenditure (CapEx) and Operational Expenditure (OpEx):

- **Reduced CapEx:**
  - **Eliminate overprovisioning:** Purchase only the GPU resources you need, when you need them, rather than buying GPUs for every server to meet peak demands.
  - **Higher utilization of expensive assets:** Maximize the use of high-cost GPUs, ensuring they are always working and generating value, rather than sitting idle.
  - **Extended hardware lifespan:** Disaggregation allows independent upgrades, meaning you can refresh GPUs without replacing entire server nodes, extending the useful life of your Cisco UCS investments.
- **Reduced OpEx:**
  - **Lower power and cooling costs:** Achieve up to 2x reduction in power consumption for AI workloads due to optimized utilization and efficient thermal management in Liqid chassis, leading to significant energy savings.
  - **Simplified operations:** Automated, software-defined resource allocation reduces manual intervention, freeing up IT staff and lowering operational overhead.

Contents

- **Reduced software licensing:** Fewer physical servers required to achieve the same GPU throughput can lead to lower software and licensing costs.
- **Faster project completion:** Dynamic resource provisioning accelerates the deployment of new AI and HPC projects, leading to quicker insights and faster time-to-market for new services.

By leveraging the combined power of Cisco UCS and Liqid PCIe composable infrastructure, organizations can achieve a more agile, efficient, and cost-effective data center, ready to tackle the most demanding AI and HPC workloads of today and tomorrow.

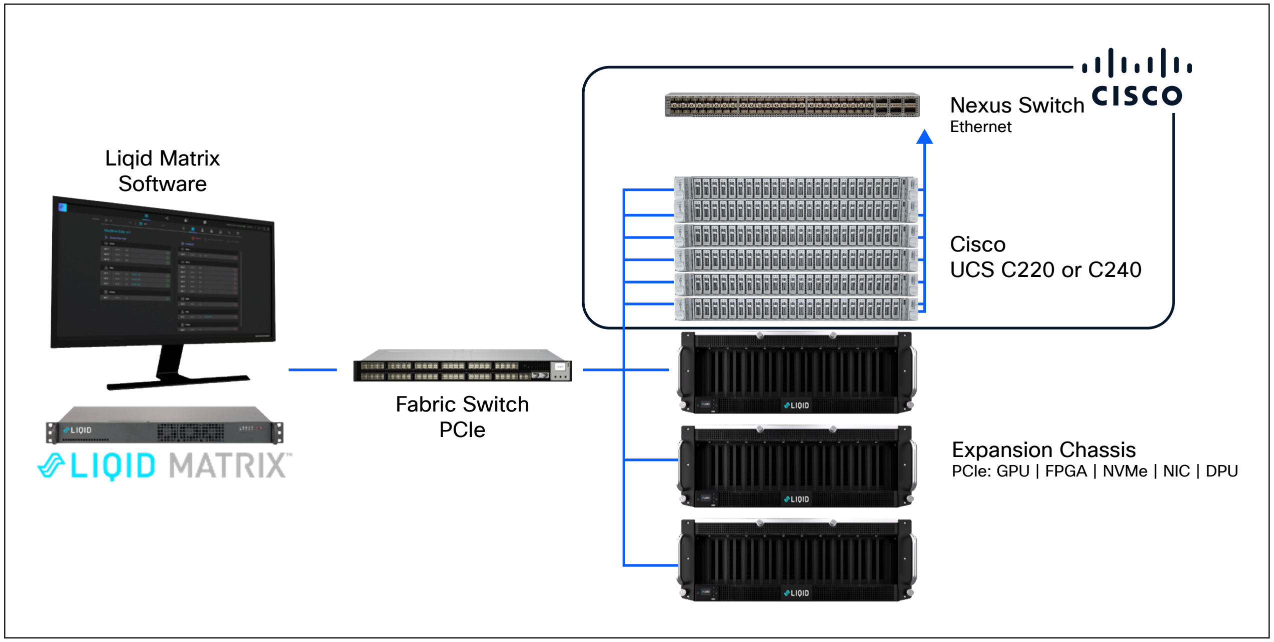


Figure 1. Cisco UCS Servers and Liqid PCIe solution.