Cisco UCS C Series with High Power GPUs

Pool and Share 600W GPUs Over PCIe Fabric

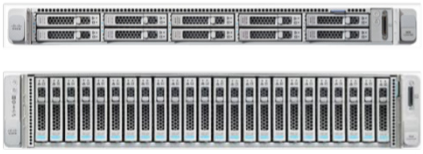
Cisco and Liqid deliver software-defined composable infrastructure that unifies GPU pooling and sharing across Cisco UCS rack servers: accelerating time to capacity, boosting efficiency, and lowering costs for AI inference, visualization, VDI, and other GPU-intensive workloads.

Beyond the core composability, the Cisco UCS C-Series further elevates its GPU capabilities through its integrated unified fabric and cloud-driven Intersight management. The unified fabric simplifies and accelerates connectivity, providing a robust network backbone for high-performance AI workloads, while Intersight offers centralized visibility, policy-driven automation, and proactive optimization for the UCS C-Series servers. This powerful combination ensures the compute foundation for GPU pooling is not only performant but also effortlessly managed and scaled, maximizing efficiency and accelerating deployment.

Cisco UCS C-Series customers rely on fabric interconnects to scale AI and GPU-centric environments that demand high capacity and peak performance. Cisco now enables enterprises to attach pools of up to 30 GPUs, including 600W-class PCIe models, to 1U and 2U C-Series servers for flexible GPU-to-CPU pairing. Powered by Liqid's high-performance PCIe fabric, these GPU pools connect to the host OS as native, bare-metal devices, enabling seamless orchestration and performance scaling on demand without thermal throttling limitations.

Enable High Power GPU Support on Traditional Servers

Cisco 1U/2U servers



Supported Cisco 1U/2U Servers include: C220 M7, C220 M8, C225 M7, C225 M8, C240 M7, C240 M8, C245 M7, C245 M8

High Power GPU Support

600W GPUs
NVLink Bridges
Full Performance

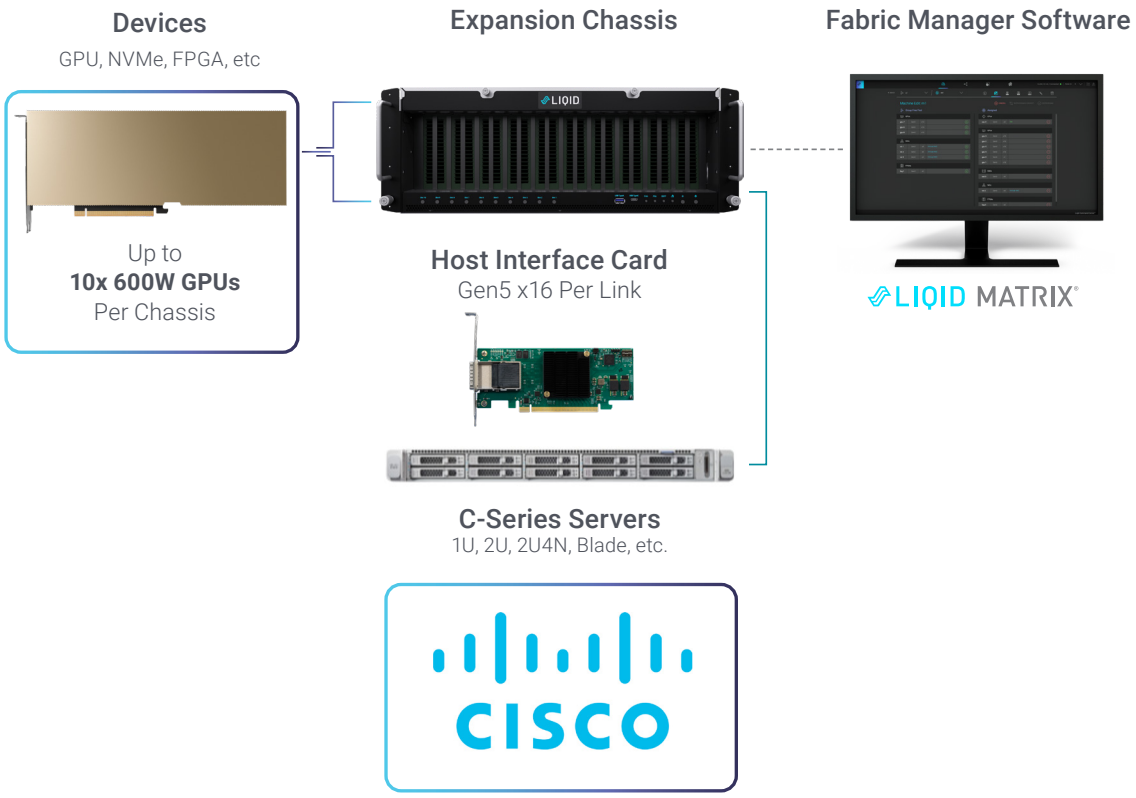
Tech Brief

Cisco Servers Supported: C220 M7, C220 M8, C225 M7, C225 M8, C240 M7, C240 M8, C245 M7, C245 M8 – AMD and Intel-based servers supported.

GPUs Supported: Nvidia (A16, A100, H100 NVL, H200 NVL, L40S, RTX PRO 6000), AMD (MI100, MI210), AMD, Intel, Groq, d-Matrix and other accelerators supported (consult the Liqid Hardware Compatibility List for details).

Note: Maximum up to 30x GPUs supported per server

UCS C-Series Servers and Liqid Software-Defined Composable Infrastructure



- » Up to 30x GPUs per UCS Server (1U/2U)
- » Supports High Power GPUs (600W)
- » Dynamic GPU Provisioning
- » Bare Metal Connectivity
- » Granular Flexible GPU Scalability
- » Simple UI, API, CLI Management
- » Slurm & K8 Orchestration Integration
- » P2P Direct GPU Performance
- » GPU Hot-Plug / Hot-Unplug
- » Multi-Vendor GPU Support

LIQID

CISCO

Use Cases

This solution is ideally suited for a wide range of enterprise and on-premises AI workloads, delivering the flexibility and performance needed for applications such as model training, inference, data analytics, visualization, and VDI. With unified GPU pooling and dynamic resource allocation, organizations can efficiently scale compute power to match evolving AI demands while maximizing infrastructure utilization and ROI.

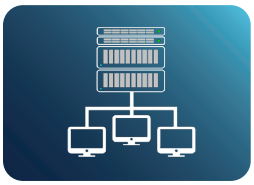
- **AI Inference & RAG Services:** Triton endpoints; low latency and higher QPS
- **VDI & Virtualization:** Fractional or full GPU allocation per VM/session
- **Rendering Farms:** High performance scale-up GPU for video and gaming
- **Transcoding & Simulations:** Assign full GPUs for windows, return to pool after
- **Data Science / Fine Tuning:** Optimize GPU jobs without requiring new servers
- **GPU-as-a-Service:** On-demand, scalable GPU power for AI intensive workloads

HPC and Research

Video, VR and Gaming

Gen AI, ML, Inference

RAG, Fine Tune, Agents



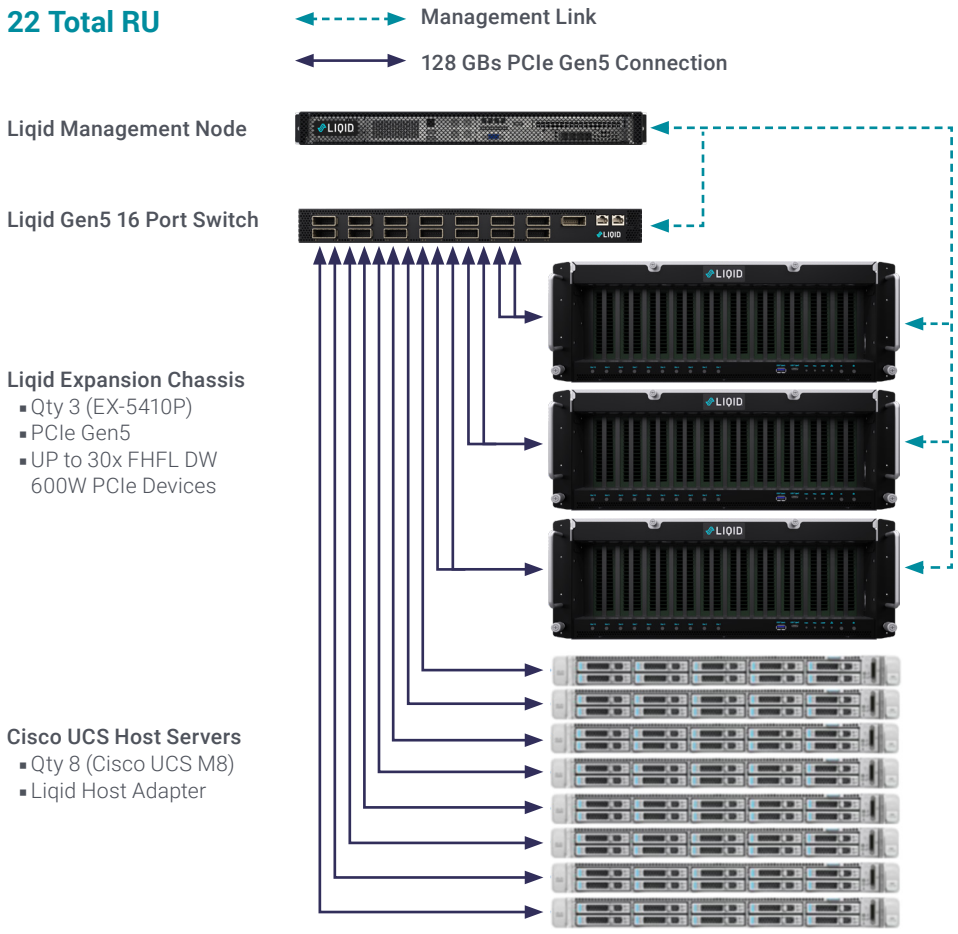
Virtualization and VDI

Analytics & Detection

GPU-on-Demand

Reference Architecture Components

- **Cisco UCS C Series (C220/C240):** Compute hosts; no app changes or drivers required
- **PCIe Gen5 Switching Fabric:** Low latency connectivity between hosts and the GPU pools
- **Liqid GPU Chassis (e.g., EX 5410P):** 10x GPUs per chassis - 600W ready power/thermal
- **Liqid Matrix Software:** Fabric managers for Orchestration, UI/API, RBAC, telemetry
- **Optical Interconnects:** Supports optional long distance low latency optical cables
- **Scalable Pod:** Scale to 30x GPUs pooled and shared with up to 8 UCS servers



Solution BOM:

- Qty 8: Cisco UCS Server M220 M8
- Qty 1: Liqid 16 Port Director
- Qty 1: Liqid 16 Port PCIe Switch
- Qty 3: Liqid EX-5410P 10 Slot Chassis
- Qty 8: Liqid Gen5 PCIe HBA + Cables
- Qty 30: GPU Power Cables
- Qty 2: Management Cable
- Qty 30: Liqid SW License, 36 months
- Qty 1: Premium Support, 36 months
- Qty 1: Professional Services

Technical Highlights and Benefits of Validated Reference Designs

Validated reference designs provide a blueprint that accelerates deployment, reduces integration risk, and ensures optimal performance. By combining proven hardware and software configurations, it enables organizations to adopt new technologies with confidence, streamline implementation, and achieve faster time to value.

- **Standard Server Compatibility:** Support dense, 600W-class GPUs in standard 1U/2U servers without thermal or power constraints.
- **Native PCIe Presentation:** GPUs appear as local Bare Metal devices — no drivers required; supports Windows, Linux, and VMware/Nutanix.
- **High Performance:** Achieve greater throughput and lower latency for AI inference and other GPU-intensive workloads.
- **Dynamic GPU Allocation:** Eliminate stranded capacity by pooling and reallocating GPUs to priority workloads on-the-fly.
- **Maximum Utilization:** Unlock up to 100% GPU efficiency through pooled and shared resource allocation.
- **Scalability and Density:** Maximize performance per rack for higher scalability and improved power efficiency.
- **Faster Time to Capacity:** Rapidly deploy inference environments and scale as workloads evolve over time.
- **Operational Simplicity:** Maintain existing OS images, tools, and workflows while enhancing performance.
- **Enterprise Integration:** Native Kubernetes/SLURM support with RBAC, audit, and API-based automation.

Resource Expansion Chassis

PCIe Fabric Switch

Host Adapter Card
Optical Cables

LIQID MATRIX®

LIQID

CISCO

Learn More

Liqid Composable GPU Solutions:
https://www.liqid.com/products/composable-gpu-solutions

Liqid Fabric and Software:
https://www.liqid.com/products/liqid-matrix-software

Liqid HCL (Hardware Compatibility List):
https://www.liqid.com/resources/library?tab=hcl-tab