cerio

Driving the next evolution of the data center

CERIO ACCELERATED COMPUTING PLATFORM

Application acceleration is fueling everything from GenAI and machine learning to visualization, simulation and data analytics. GPUs are the currency of this computing revolution, creating the demand for a more cost-effective, efficient way to access and use these critical resources.

The Cerio Accelerated Computing Platform transforms the way systems are designed and deployed, providing full control over the utilization, choice and scale of GPU infrastructure.

Cut total cost of ownership by up to 50%

Traditional fixed systems trap GPUs and other expensive resources in servers, reducing access and driving up costs from low utilization. Even when GPUs are sitting idle, they can't be allocated until the server is finished running a workload.

By decoupling accelerators from servers and making them available in logical resource pools, Cerio delivers the flexibility to share resources and compose systems in real-time to match workload demand – all defined and implemented in software. Reductions in server complexity, power and cooling translate into significant cost savings at any scale.

Add any accelerator at any time

Specialized systems with proprietary stacks limit the choice and availability of next-generation accelerators. Multi-year upgrade cycles and depreciating value for installed systems add to the cost and complexity of managing outdated technologies.

Cerio reinvents the system model to fit the application or use case. Co-locate GPUs, mix GPUs and TPUs from different vendors, scale at any time – and build the system to fit your business.

Cerio Platform Advantages

- *Compose any GPU to any server with no BIOS changes – up to 32:1 GPU density*

- *Create shared GPU access – high mobility and over 90% utilization*

- *Use any GPU variant – mix with TPUs, DPUs, NVMe or any PCIe device*

- *Dynamically remove or add devices – with zero downtime*

- *Support industry standards and future-proofing – PCIe Gen5 to CXL 2.0*

Cerio

Experience software-defined operational agility

Running a wide variety of applications and workloads makes it very difficult and costly to schedule replacements and upgrades, especially for large-scale infrastructure. Cerio lets you share and move GPUs and other resources to meet SLAs or create tiered service levels for priority workloads. Any server can be used for any set of resources instead of waiting for the right servers to become available.

- Add or replace devices in software
- Eliminate downtime for hardware upgrades
- Deliver highly differentiated capabilities using shared infrastructure

Design and deploy your real-time system with the experts

Our dedicated customer experience team is ready to get you started with planning and system design through to testing and production deployment, providing white glove service throughout your implementation. Cerio gives you advanced platform capabilities backed by enterprise-grade support, and a team of experts in software-defined architecture.

Empower accelerated computing in any environment

As the pace of innovation accelerates, it's challenging to support the sheer diversity of applications and workloads with a single system design. Cerio provides the flexible foundation for a wide range of computing requirements, whether it's running a new inference cycle in seconds, turning static servers into smart rendering farms or reducing processing time for complex simulation workloads. GPUs can be shared, moved, dedicated or repurposed and any server can be transformed into an accelerated computing environment.

Cerio is at the center of a standards-based ecosystem helping our customers and partners envision an agile new environment designed for the next evolution in computing.
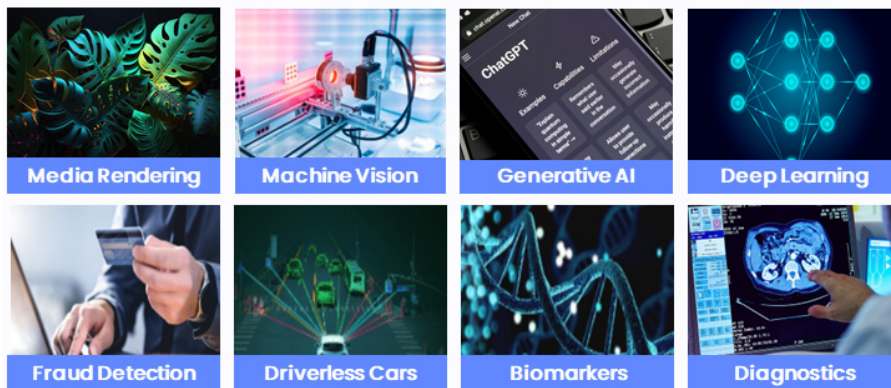
Figure 1: *Drive more value and productivity from your GPU investments for any compute-intensive application.*

How Cerio Works

SOFTWARE-DEFINED ARCHITECTURE

The Cerio Accelerated Computing Platform uses software-defined architecture to enable highly flexible systems. Three interconnected layers form the core of the Cerio model:

The **Service Management** layer provides a complete set of reusable, prebuilt services for the rapid deployment of composed systems, including:

- Dynamic Composition – creating real-time systems

- Resource Management – discovery of resources and assignment to servers

- Fault Management – identifying network issues

- Policy Management – manual and automated rules for system function

- Access Control - interface with user and authentication systems

The **Overlay** is the implementation layer for scale composability that optimizes different types of traffic and ensures full interoperability with existing system services. Overlay services include:

- PCIe compatibility – standard PCIe capabilities

- Domain translation – multiple domains of control

- Resource discovery – enabling resource management

The **Underlay** is the physical transport layer for optimized traffic flow across a high radix of paths that share no common links, ensuring resiliency, high throughput, and low latency. Underlay capabilities include:

- Distributed control plane – deadlock-free routes for efficient connectivity

- Distributed data plane – embedded switching and load balancing

- Direct interconnect – pre-wired topologies for highly scalable, entirely passive connectivity
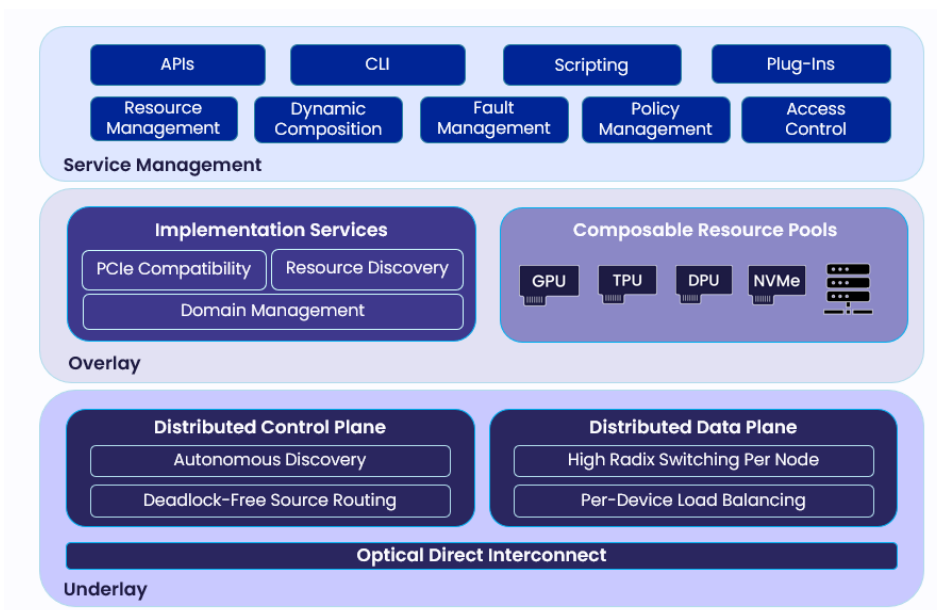
Figure 2: Create highly flexible systems using software-defined architecture.

Advanced platform technologies

The proven technologies at the core of the Cerio Accelerated Computing Platform provide an intelligent and highly reliable foundation for the design and deployment of real-time systems.
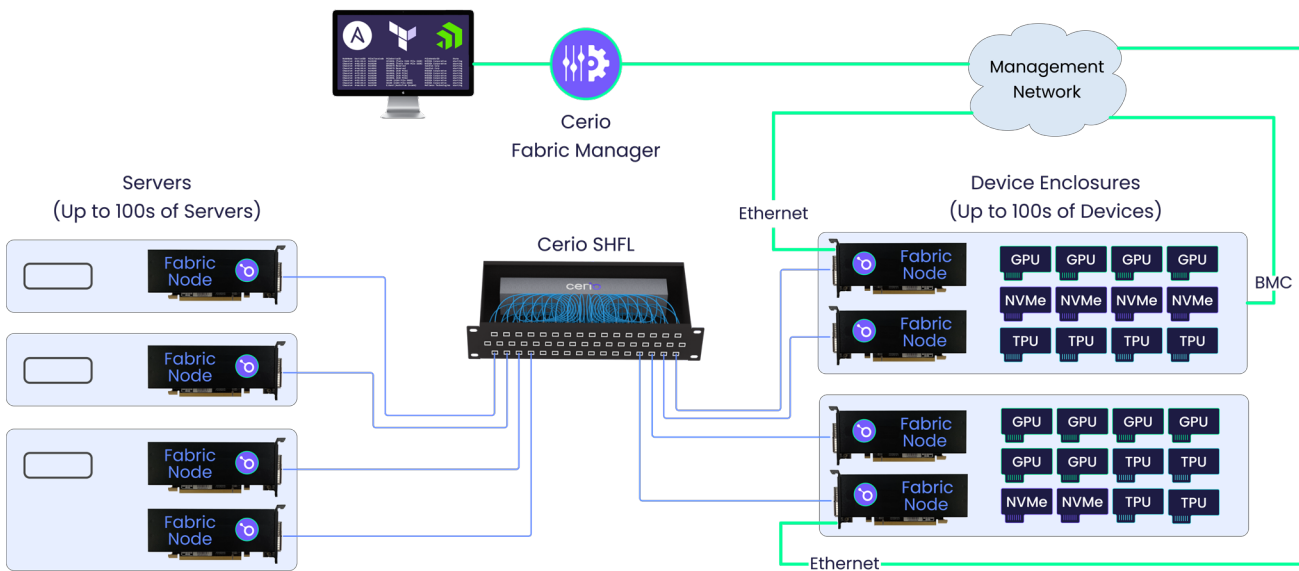
FABRIC MANAGER

The Cerio Fabric Manager is the software console that manages overlay services and device chassis using a standards-based composability data model. The Fabric Manager provides the logic and policies for system composition and resource allocation, using a rich CLI and a configurable set of Customer-Facing Services.

FABRIC NODE

The Cerio Fabric Node delivers highly optimized distributed switching and transport services for composed systems. Self-discovering, self-configuring and self-healing, the Fabric Node connects hosts and chassis inside a single high-performance domain. Modular design makes it possible to create different server attachment points for compositions, enabling high GPU utilization and agility.

FABRIC INTERCONNECT

The Cerio Fabric SHFL brings the simplicity of prewired topologies to composable systems. By decoupling the optics from the underlying switching, Cerio delivers scalable performance at a significantly lower cost and footprint using cabling that is entirely passive and requires zero power or cooling.

Figure 3: *Design the system that fits your business at any scale.*

Cerio Accelerated Computing Platform
TECHNICAL SPECIFICATIONS

SYSTEM MODEL REQUIREMENTS

Platform Component	Specification	Description
Cerio Fabric Manager Software console	Server/OS	Any Linux workstation or equivalent
Cerio Fabric Node Passively cooled, with double density Quad Small Form Factor	Form Factor	PCIe full-height, half-length
	Dimensions	111.15 mm x 167.65 mm
	Host Interface (PCIe)	PCIe Gen5 x8, Gen4 x16
	Power Consumption	40 W
Cerio SHFL Direct interconnect	Dimensions	Height: 88.4 mm Width: 482.6 mm Depth: 203.2 mm
	Weight	350 g
	Power Consumption	0 W
3rd Party Hardware Specifications		
Server Hosts	BIOS	Supports any server - no BIOS changes required
Enclosures	PCIe Chassis	Gen 5
Resources	PCI Devices	PCIe Gen 3, Gen 4 and Gen 5 Passively cooled for high-density configurations
Ethernet Interface	Ethernet Management Port	10/100/GE RJ45
Console Interface	Management Port	USB-C serial

ABOUT CERIO

Cerio empowers organizations with real-time systems to advance new discoveries and accelerate innovation at a sustainable cost and footprint. Leaders in software-defined architecture, Cerio has offices and projects spanning international markets, with Centers of Excellence in Europe and North America. Visit **cerio.io**.

Contact us:

sales@cerio.io

515 Legget Drive, Suite 500, Ottawa, Ontario K2K 3G4