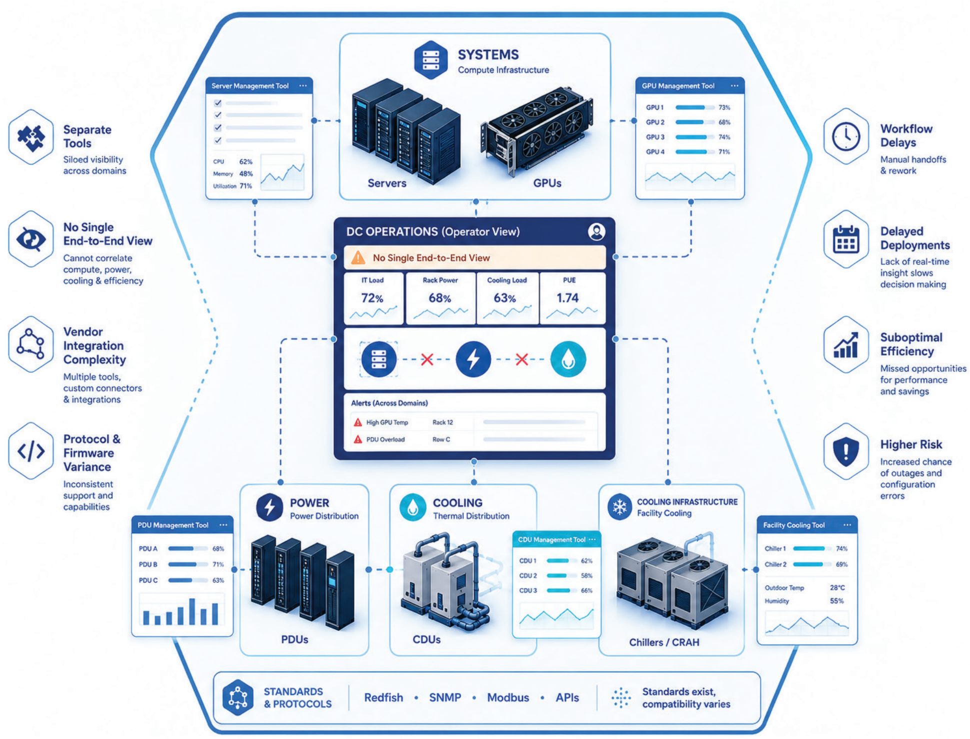
Converged Infrastructure

for Modern AI Data Centers

The Problem

Modern data centers are moving beyond traditional Data Center Integrated Management. High-density AI/HPC, and enterprise workloads now depend on intelligent power and cooling infrastructure. Conventional business compute of the past relied only on compute, networks and storage, forcing data center operators to utilize fragmented application tools to manage their operations.

Today, DC Operators often manage servers, GPUs, PDUs, CDUs, cooling infrastructure, and power sources through separate tools. This legacy paradigm often created inefficiencies in their operational model, which resulted in severely impacting their time-sensitive workflows. Data center personnel can see individual components, but they may not have a single view of how computing demand, rack power consumption, liquid cooling, and facility energy PUE affect each other.



The challenge is that most management software was not originally designed for this level of convergence. Many platforms still treat servers, power, and cooling as separate domains. At the same time, power and cooling equipment come from many different vendors, resulting in costly integration challenges and delayed data center deployments.

Even where standards such as Redfish, SNMP, Modbus, and open APIs exist, real-world compatibility still depends on vendor implementation, tested device support, variance in firmware versions, and cooperation between infrastructure OEMs; severely impacting software platform integration.

The Domains

Modern AI data center management is moving beyond a system-only view. As compute density increases, especially with GPU-accelerated platforms, the operational model now depends on a three-connected domain architecture: Systems, Power, and Cooling. Each domain has its own role, but none of them operate in isolation anymore.

SuperCloud Composer License Model

Systems: The systems domain covers the compute and network infrastructure inside the rack: servers, CPUs, GPUs, storage, switches, firmware, connectivity, topology, and hardware health. This domain is important because modern racks are no longer simple server collections; they are dense compute zones with far more components, dependencies, and interconnects rather than traditional environments a decade ago.

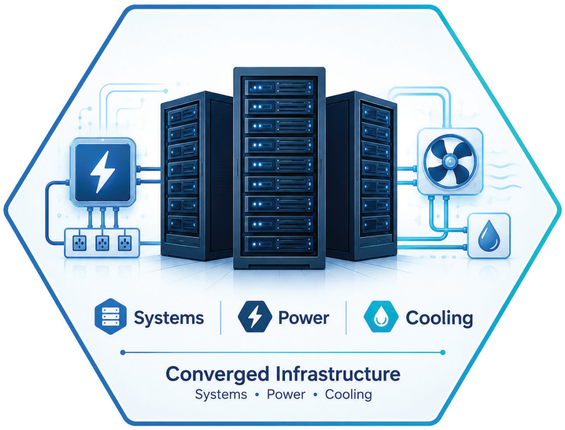
Power: The power domain covers how energy is delivered, measured, controlled, and protected across the rack and wider facility. Higher-density systems consume more power, and rack-level power awareness is becoming critical for capacity planning, availability, efficiency, and safe operation.

Cooling Domain: The cooling domain covers the thermal infrastructure needed to keep modern systems operating safely, including liquid cooling, CDUs, coolant flow, pressure, temperature, leak detection, and facility cooling. As racks become hotter and denser, cooling is no longer just a building-services concern, it is directly tied to system performance and availability.

Converged Infrastructure View

Supermicro is looking at modern data center infrastructure as a converged ecosystem, not as separate hardware silos. As compute platforms become denser and more power-intensive, the operational focus must expand beyond servers alone to include the supporting power and cooling domains that keep those systems available, safe, and performant.

As these domains become more connected, the software layer must also evolve. It is no longer enough for management platforms to monitor servers alone. Modern data center software needs the ability to discover, monitor, alert, and eventually manage additional infrastructure domains such as rack power, liquid cooling, CDUs, PDUs, and facility cooling systems. This creates the foundation for a more integrated operational model, where Systems, Power, and Cooling can be understood together rather than managed as separate environments.



Power and cooling monitoring are no longer just about uptime. They become functional needs of safety, capacity, efficiency and optimization. At rack, PDU, CDU, and cooling-system level awareness become paramount. Platforms should continuously monitor power draw, phase/load balance, outlet-level consumption, thresholds, analyze overload risk assessment where temperature, flow rate, pressure, pump status, valve status, leak detection, and cooling capacity are measured in real time.

By combining power and cooling telemetry collection into one single pane data center operatives can quickly mitigate operational outages in real time.

Converged Benefit	Operational Rationale
Sensor-Driven Intelligent Environmental Monitoring	Using data from rack sensors, PDUs, CDUs, cooling systems sensors to detect heat, airflow, liquid-cooling, or power risks to identify risks before they occur.
GPU Load Automation	Dynamically shifting, throttling, or scheduling GPU workloads based on available power and cooling, allowing GPU infrastructure to react to facility limits of constraints.
SLA Compliance	Intelligent scheduling balances performance, energy cost and infrastructure capacity while maintaining SLA commitments.
Real-Time PUE (Power Usage Effectiveness)	Use real time data to show live usage and overall efficiency.
Real time Energy Price Response via Real-Time Locational Marginal Pricing (LMP)	Help reducing operating costs and avoid stressing power capacity during peak periods.

The Solution

Supermicro has moved beyond traditional server infrastructure, investing in its wider data center ecosystem approach to support sustainable high-density compute environments. This includes not only compute systems, but also supporting cooling and power domains needed to operate AI, HPC, and enterprise workloads safely and efficiently.

SuperCloud Composer's (SCC) open distributed infrastructure management software platform brings to the industry a scalable Kubernetes control plane unifying each of the infrastructure domains into a more streamlined operational dashboard. SCC provides the foundation to discover, monitor, observe and manage systems, rack power, and cooling infrastructure through a single management experience.

SuperCloud Composer's PDU Console and LCCM can collect vital power and cooling telemetry that can be used to generate alerts, raise operational events, notify external systems, and support automated workflows. Examples include overload warnings, phase imbalance alerts, outlet-level power anomalies, cooling degradation events, leak detection alarms, pump or valve fault alerts, and capacity-risk notifications. Over time, this telemetry can also support higher-level orchestration decisions such as workload throttling, workload placement, and SLA-aware infrastructure protection.

Telemetry Trigger	Event \ Action
PDU Overload Risk	Alert operations, raise critical events, notify via email/syslog/web-hook.
Phase/Load Imbalance	Dynamically shifting, throttling, or scheduling GPU workloads based on available power and cooling, allowing GPU infrastructure to react to facility limits of constraints.
Outlet-level high power draw	Identify affected rack/server and flag abnormal consumption.
Cooling Temperature Breach	Critical thermal alert and escalation to operations.
Low Flow Rate / Pressure Issue	Cooling degradation event before thermal failure occurs.
Pump or valve fault	Maintenance alert and service workflow.
Leak detection	Critical alarm, immediate escalation, possible workload protection action.
Water\Fluid quality issue	Preventive maintenance event for liquid cooling system.
Reduced power/cooling headroom	Capacity warning and future workload placement decision.

SCC PDU Console

The SCC "PDU Console" module provides centralized visibility and management of smart PDUs by collecting key physical asset information, real-time sensor data, outlet status, and energy usage. Through an intuitive hierarchical rack view, users can quickly locate PDUs within the data center, monitor live telemetry, review historical power data through charts and tables, and perform remote outlet power on/off actions. This helps teams manage PDUs across data centers or remote edge sites, improve energy efficiency, support PUE monitoring, and make faster operational decisions for high-density HPC and AI rack environments.

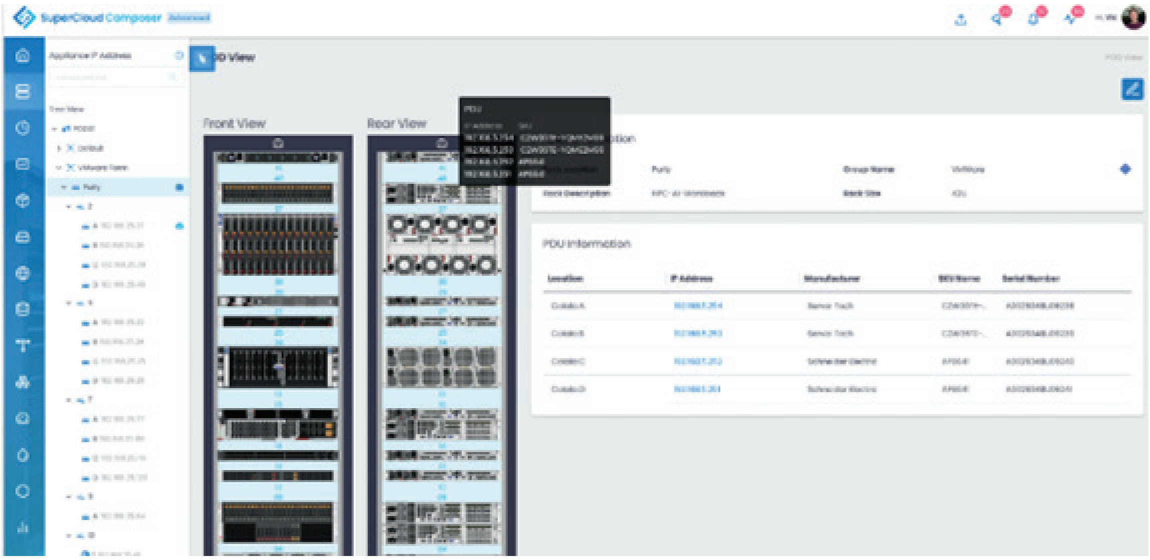


Figure 1 – SCC PDU console – displayed PDU's per rack

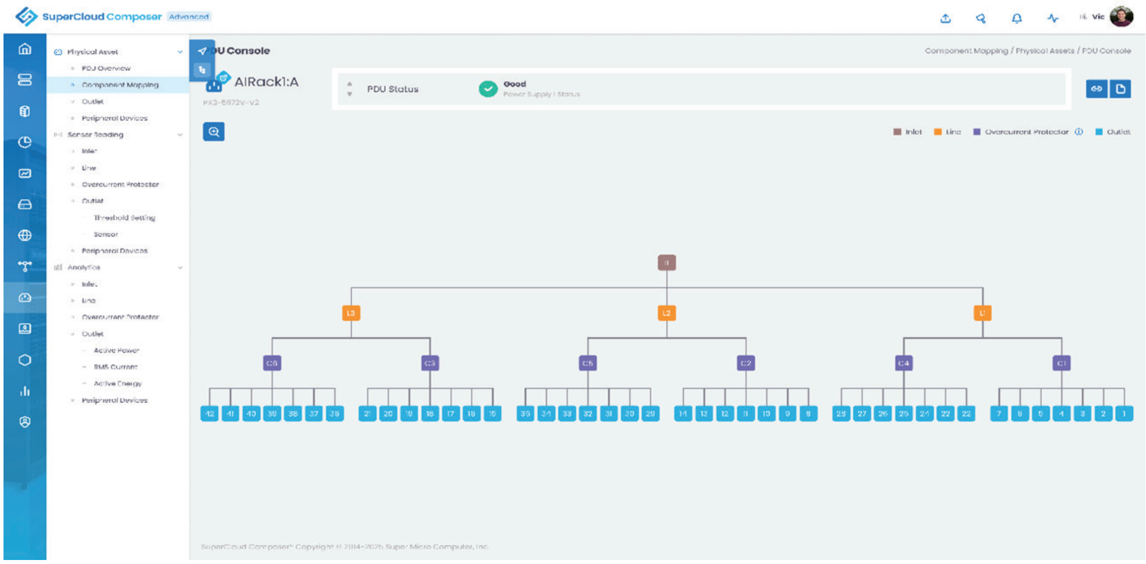


Figure 2 – SCC PDU console – Component Mapping

- Aggregates historical PDU data and presents it through charts, graphs, and tables.
- Provides granular visibility into power and energy usage over time.
- Delivers at-a-glance metrics, real-time telemetry, and historical monitoring.

SCC Cooling Infrastructure

The SCC "Cooling Infrastructure" module provides real-time monitoring across the cooling chain, from rack-level liquid cooling devices through to facility-level cooling towers.

For CDUs, SCC can monitor key operational data such as supply and return temperature, coolant flow, pressure, pump status, valve status, leak detection, and alarm conditions.

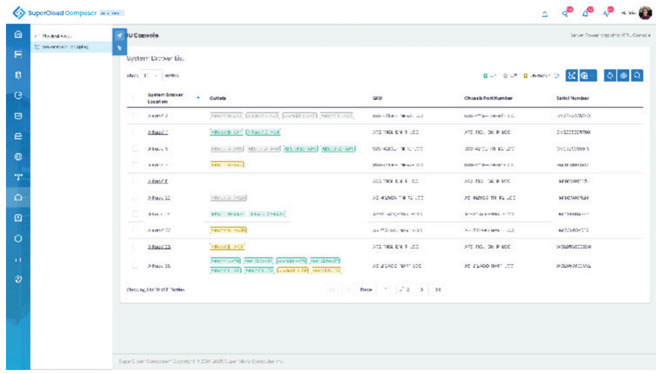


Figure 3 – SCC Cooling Infrastructure, in-rack CDU

Figure 4 – SCC Cooling Infrastructure, in-row CDU

Figure 5 – SCC Cooling Infrastructure, in-rack CDU

Figure 6 – SCC Cooling Infrastructure, in-row CDU

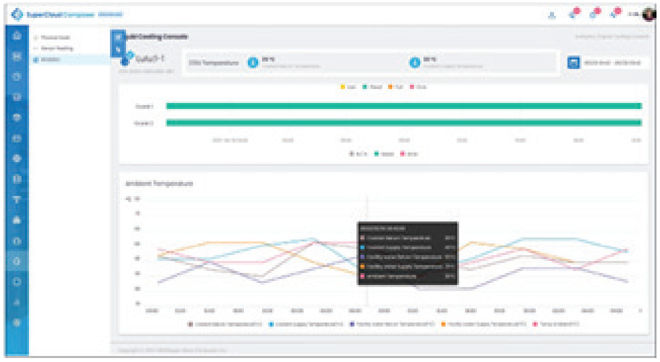


Figure 7 – SCC Cooling Infrastructure, in-rack CDU

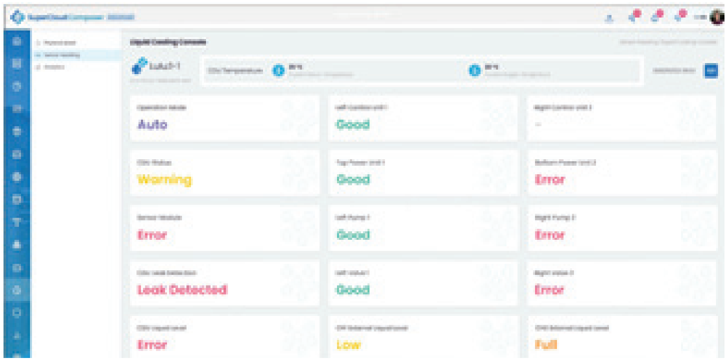


Figure 8 – SCC Cooling Infrastructure, in-row CDU

For RDHx and sidecar cooling units, SCC extends visibility into rack and row-level cooling performance, including airflow, coolant temperature and flow rate. At the facility level, cooling tower monitoring provides visibility into water temperature, pump and fan status, basin water levels, water quality, flow rate, and overall cooling capacity. This gives operators a single view of cooling health from the rack to the wider data center cooling infrastructure.

Figure 8 – SCC Cooling Tower

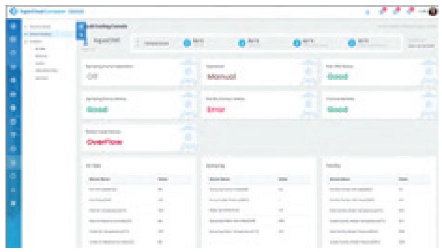


Figure 9 – SCC Cooling Tower Status

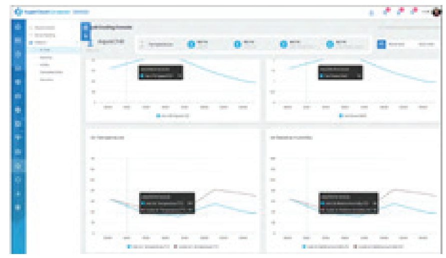


Figure 10 – SCC Cooling Telemetry

Future View

The next stage of converged data center management is orchestration, where Systems, Power, and Cooling are not only monitored together, but coordinated through policy-driven actions to manage:

Energy source orchestration

Orchestration across multiple energy sources, such as grid, BESS (battery energy storage systems), fuel cells, solar/microgrids, generators, and future power sources such as nuclear and SMRs.

Cooling-aware compute throttling

Automatically reducing CPU/GPU power draw or performance when cooling capacity, rack power, or facility conditions are degraded, helping prevent thermal failure or service disruption.

Predictive maintenance and risk prevention across power and cooling

Using telemetry from power, cooling, rack, and system infrastructure to identify degradation, capacity risk, or failure conditions before they impact workloads or data center operations.