

GPU Infrastructure Monitoring Coverage

What hardware-layer monitoring of AI clusters should cover — and why nodes vanish without it

Why AI clusters lose nodes without a clear cause

- GPUs disappear mid-training; orchestration reports a missing node, not the hardware cause
- OS and agent-based tools go silent exactly when a node overheats, loses power, or hangs
- Thermal and power faults in dense chassis are intermittent and hard to reproduce after the fact
- Multi-vendor accelerators each ship their own console — no single health view
- A single failing node can stall an entire distributed job, multiplying the cost of slow detection

The six coverage areas

GPU & accelerator health

Temperature, ECC memory errors, utilization, throttling, board status — read from the chassis, not an in-OS agent

Chassis thermals & airflow

Inlet/outlet temperature, fan speed and failure, hot-spot detection

Power & PSU integrity

Per-PSU load, redundancy state, power capping, supply degradation

Compute & memory faults

DIMM errors, CPU status, RAID/NVMe health, PCIe/NVLink link errors

Fabric & node reachability

Management-network reachability and BMC liveness — a dropped node is a hardware event, not a scheduling gap

Service & job impact

Which training jobs, inference services, or tenants depend on the affected nodes

What the out-of-band approach delivers

- Agentless, OS-independent telemetry that survives crashes and thermal events
- Multi-vendor GPU and server coverage in one health view (Redfish, IPMI, iDRAC, iLO, iBMC)
- Node faults mapped to job and service impact — respond by impact, not raw alarm