AI Datacenters as Grid-Responsive Flexible Loads

Real-Time GPU Power Modulation

I. Problem

AI and HPC datacenters are rapidly becoming some of the most power-dense loads on electrical grids, yet they are typically operated as static, non-interruptible consumers. This creates a growing mismatch with grid requirements for fast-acting flexibility such as curtailment and demand response.

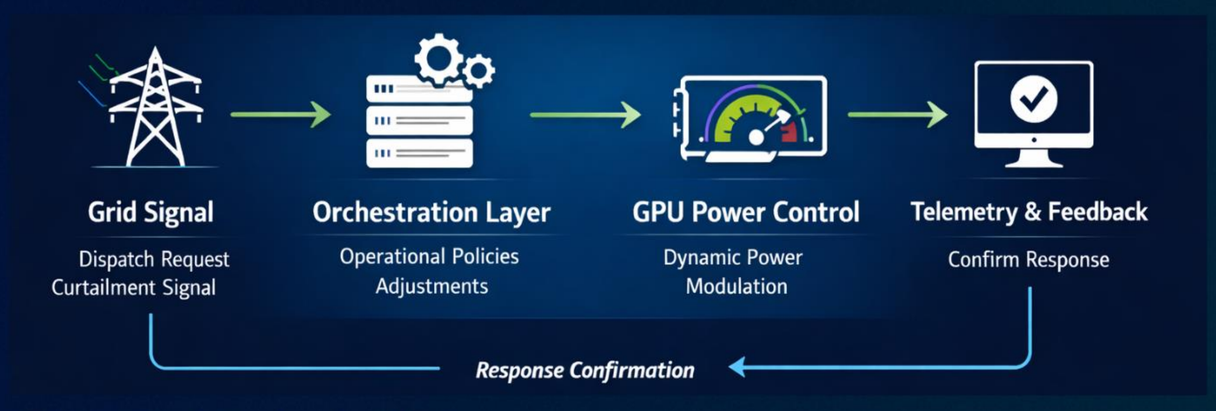
AI clusters → to GW scale
is increasing Grid volatility

II. Solution

This demonstration shows that AI datacenters can operate as grid-responsive flexible loads using existing infrastructure.

Using DCIM standards, and no hardware modifications, this software-defined orchestration enables GPU control to flex load.

III. System Architecture



Deterministic Control Loop:

1. Grid Dispatch Signal Ingestion (ISO/CSP aligned)
2. Orchestration & Constraint Adjustments
3. Real-Time GPU Power Modulation
4. Telemetry Feedback & Acceptance

The key innovation is not new hardware, but unlocking fast, predictable load flexibility through software orchestration.

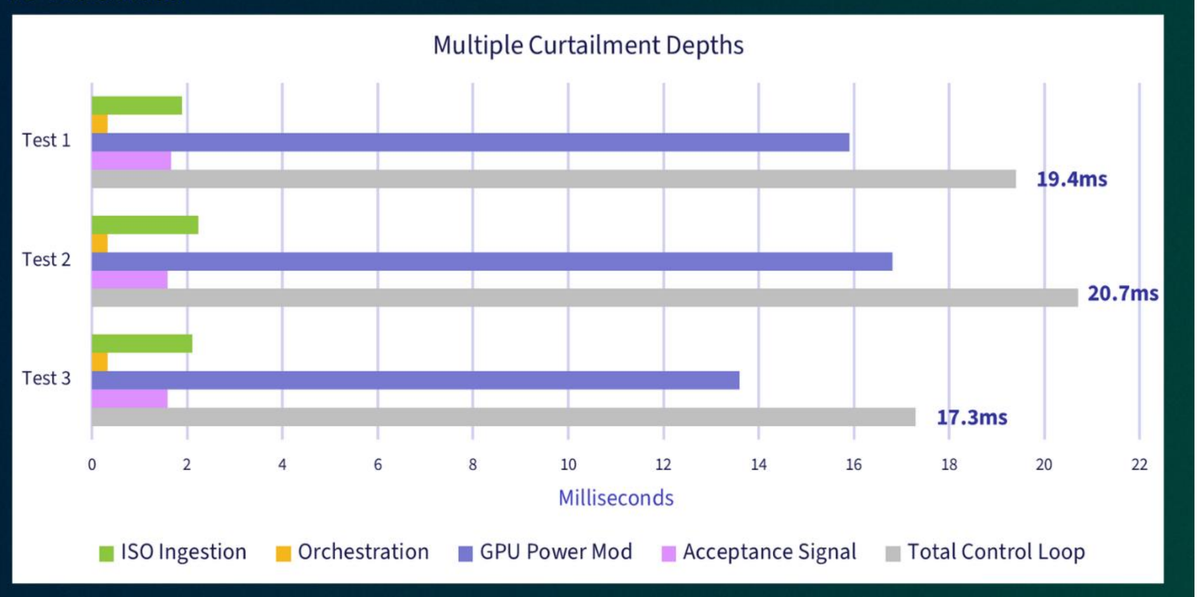
IV. Methodology

The demonstration was conducted in a controlled laboratory environment using Supermicro systems equipped with NVIDIA B200 GPUs.

The pilot focused on ultra-low latency and repeatability under grid-aligned dispatch conditions. Fast demand response signals (typically requiring sub-two-second compliance under ISO programs) were used to trigger load reduction events.

These signals were translated into real-time GPU power limit adjustments through an orchestration layer, which continuously monitored DCIM telemetry to ensure operation within vendor-supported thermal and electrical safety envelopes.

V. Results



<u>End-to-end control loop execution averaged 19 milliseconds.</u>

Operational Continuity:

- GPU power was reduced up to 75% while running workloads, without system resets or interruption, all while preserving SLA availability.

Repeatability:

- Dozens of tests confirmed bounded, predictable, and repeatable response behavior across varying curtailment signals and workload conditions.

VI. Grid & Infrastructure Impact

AI can operate as an interruptible and flexible load rather than a static demand on the grid.

- Enables demand response participation
- Faster interconnection approvals
- Supports balancing and grid stability
- Follows ISO operational requirements
- Operates within limited MW capacity
- Aligns compute growth with grid constraints

Bentaus

Robert Davidoff, Kevin Matcham, Galyn Black, Matthew Stewart, Jeremy Swyers — Bentaus
Kevin Culp, Duyen Truong — Supermicro

OCP FUTURE TECHNOLOGIES SYMPOSIUM
29-30 April, 2026
Barcelona, Spain

Leading the Future of AI.