PARLON GPU FLEET SOLUTION

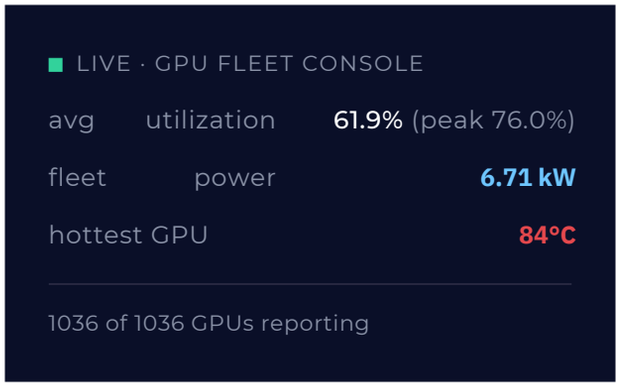
GPUs are the most expensive thing in your data center, and the hardest to see clearly.

The Parlon GPU Fleet Solution gives capacity planners and on-call engineers a live, fleet-wide view of GPU health, utilization, and waste, down to a single device.

THE PROBLEM

A GPU fleet averaging a healthy-looking 45% utilization can still be half maxed out and half sitting idle. The average hides the split, and the idle half is capacity you already paid for.

Most infrastructure monitoring wasn't built to separate a GPU that's healthy from one that's actually doing work, or to catch a device quietly throttling under heat before it costs a training job real time. Teams are left correlating utilization, temperature, and power by hand, one dashboard at a time, manageable at a handful of GPUs, unmanageable at fleet scale.



Why this matters for the AI buildout: this isn't an AI feature riding on top of infrastructure. GPUs are the compute layer AI training and inference run on, and seeing that fleet clearly, at the device level, in real time, is the infrastructure work that has to get done before an AI buildout can scale without quietly wasting the most expensive hardware in the building.

WHAT IT DOES

What the Parlon GPU Fleet Solution does

TEMP

Fleet-wide health & thermal visibility

Real-time fleet size, utilization, power, errors, and hottest GPU, with Throttle Watch correlating clock frequency against temperature to catch thermal throttling as it happens.

WASTE

Utilization & waste detection

A distribution histogram and idle GPU-hours expose stranded capacity a single average would hide, reported in percent, hours, and watts — never currency.

DRILL

Per-GPU drill-down

Click any GPU for its full identity, a plain-language health reason, and sparklines for utilization, memory, temperature, power, and clock speed.

XPRT

Expert Mode hardware deep-dive

Six collapsible subsystem views — thermal, power, compute, memory, interconnect, and errors — for GPU engineers who need to go deeper without cluttering the primary page.

47.7%

WHY IT MATTERS

of enterprises already have AI training or inference workloads deployed today. Every idle GPU behind a healthy-looking average is capacity you already paid for.

Source: EMA, 2026

DEPLOYMENT & COVERAGE

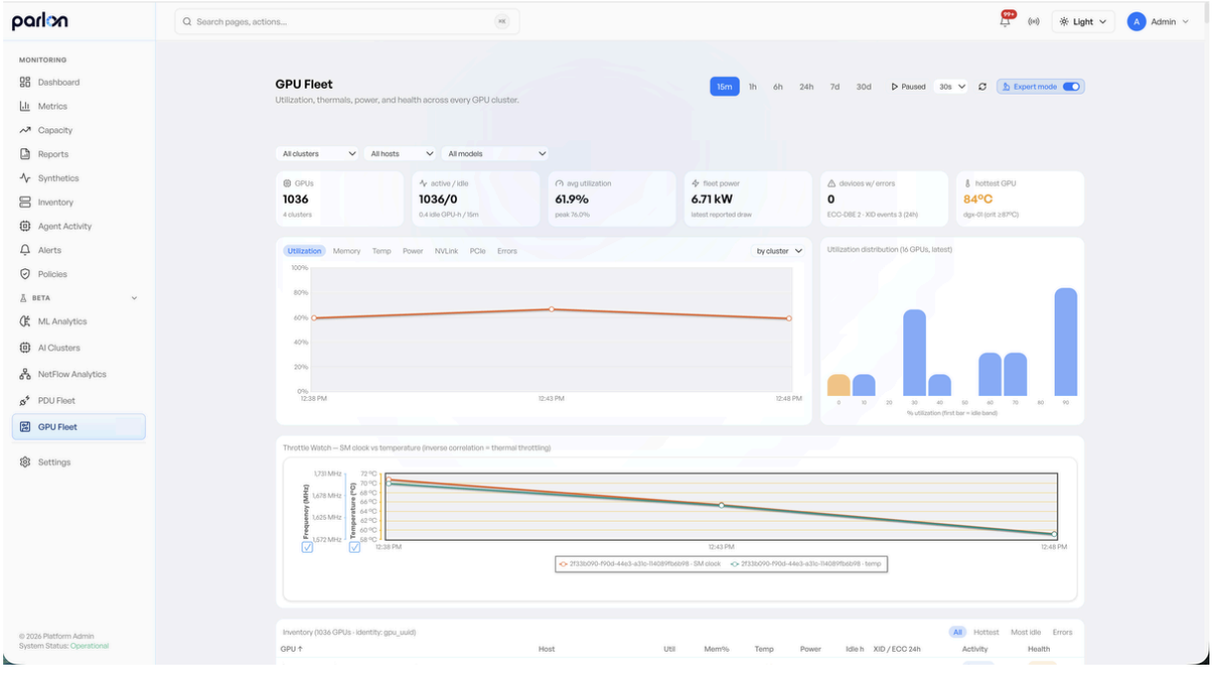
One fabric, one inventory.

GPU telemetry flows through the same Normalization Engine as every other infrastructure source. Every reporting GPU also appears in Parlon's unified asset inventory, alongside the rest of your monitored estate. Reads standard NVIDIA DCGM telemetry, so any DCGM-instrumented data-center GPU is supported out of the box.

NVIDIA DCGM

PRODUCT

The Overview tab, at a glance



The Overview tab: fleet KPI cards, utilization distribution, Throttle Watch correlation, and per-GPU inventory in one view.

LEARN MORE