Parcus

Book a briefing

AI-NATIVE INFRASTRUCTURE

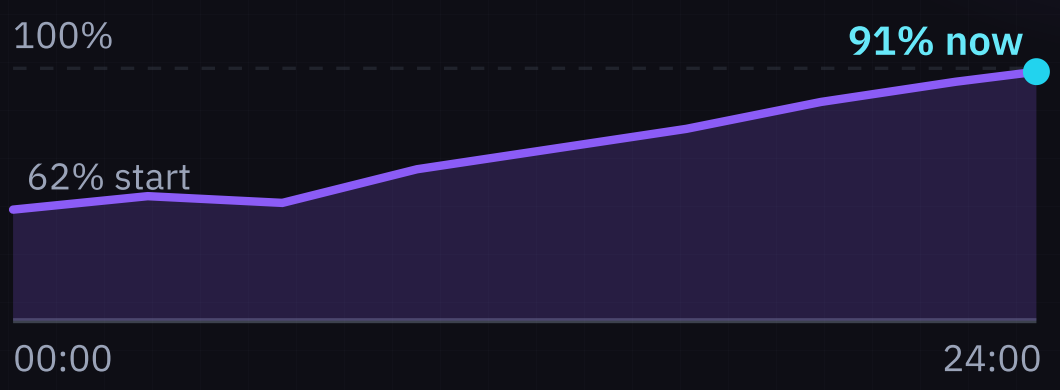
Infrastructure that thinks at the speed of your models.

Parcus designs, deploys, and operates elegant AI infrastructure — from GPU fleets and inference fabrics to the observability layer that keeps every token accountable. Premium engineering, quietly relentless operations.

Start a conversation

Explore capabilities

Live fleet telemetry
Aggregate utilization across managed clusters, last 24 hours.



99.98%	41 ms	3.2×
Fleet uptime	P50 inference	Cost efficiency

TRUSTED BY TEAMS AT HELIOS LABS Nordwind AI Cartesian Health Meridian Bank Atlas Robotics Vela Media

CAPABILITIES

Eight disciplines. One accountable partner.

Every engagement is staffed by senior engineers who own outcomes end to end — no handoff gaps, no black boxes.

GPU Cluster Engineering
Bare-metal provisioning, topology-aware scheduling, and fabric tuning for training fleets up to 4,096 accelerators.

Inference Platform Design
Low-latency serving stacks with continuous batching, speculative decoding, and autoscaling that respects your SLOs.

Model Operations & MLOps
Registry, evaluation gates, canary rollouts, and drift detection wired into a single deployment pipeline.

AI Security & Compliance
Prompt-injection defense, model access control, and audit trails mapped to SOC 2, ISO 27001, and the EU AI Act.

Observability & Cost Control
Token-level accounting, per-team chargeback, and anomaly alerts that catch runaway jobs before the invoice does.

Data Platform Engineering
Streaming and batch pipelines, feature stores, and vector infrastructure built for reproducible training data.

Hybrid & Sovereign Cloud
Workload placement across on-prem, colo, and cloud with data-residency guarantees and unified control planes.

Architecture Advisory
Independent reviews of your AI stack — capacity plans, vendor selection, and reference architectures you own.

INTELLIGENCE COMMAND CENTER

Every cluster, every model, one pane of glass.

The Parcus command center correlates utilization, latency, and spend in real time — so capacity decisions are made on evidence, not intuition.

Throughput vs. latency — managed fleet, 7-day window
Requests per second (left axis) against P95 inference latency in milliseconds (right axis).

18,400
Requests / second, peak
▲ 12% vs last week

112 ms
P95 latency, fleet-wide
▼ 9 ms after fabric retune

$0.0041
Cost per 1K tokens
▼ 18% this quarter

27
Models in production
All within SLO