Federator.ai Cortex™

ProphetStor

The AI Ops Full-Stack Solution for Next-Generation GPU Data Centers

2x	**+50pp**	**90%**
GPU Efficiency	**Utilization Gain**	**Downtime Reduction**
19/19	**PUE 1.15**	**3 mo.**
NCP API Coverage	**Cooling Efficiency**	**Deployment Time**

- **Autonomous AI Operations:** 12 AI agents perform predictive remediation, causal root-cause analysis, and self-healing—reducing unplanned downtime by up to 90%. When a CDU cooling failure occurs, Cortex traces the root cause across GPU workloads, network fabric, and cooling systems in real time—something no single-layer tool can detect.
- **Every Week of Delay = $4.3M Lost Revenue:** At 10,000 GPUs running $18/hr, each month of delayed operations wastes $130M+ in potential compute revenue. Cortex compresses AI Factory deployment from 12 months to 3 months, unlocking 9 months of accelerated revenue.
- **Full NVIDIA Certification (NCP) in Weeks, Not Months:** 19/19 API categories pre-built for DGX Cloud compliance. Competitors require 12–18 months to achieve what Cortex delivers out of the box—and NCP status commands a 15–20% pricing premium.
- **Smart Liquid Cooling, AI-Driven:** Workload-aware thermal management with PID + feedforward control achieves PUE ≤ 1.15, extends GPU lifespan by 30%, and delivers 45% higher cooling throughput than manual BMS systems.
- **16 US Patents + 15 Pending:** Multi-Layer Correlation (US Patent 11,579,933), Spatial & Temporal GPU Optimization, Predictive Self-Driving Autoscaling—a deep technology moat that cannot be replicated by open-source alternatives.

The Cost of Inaction

GPU hardware failures cause 50% of all AI training interruptions (Meta LLaMA 3 study, 16,384-GPU cluster). A single CDU cooling failure costs $384K in wasted compute before anyone checks the cooling loop—because GPU monitoring and facility management live in separate worlds.

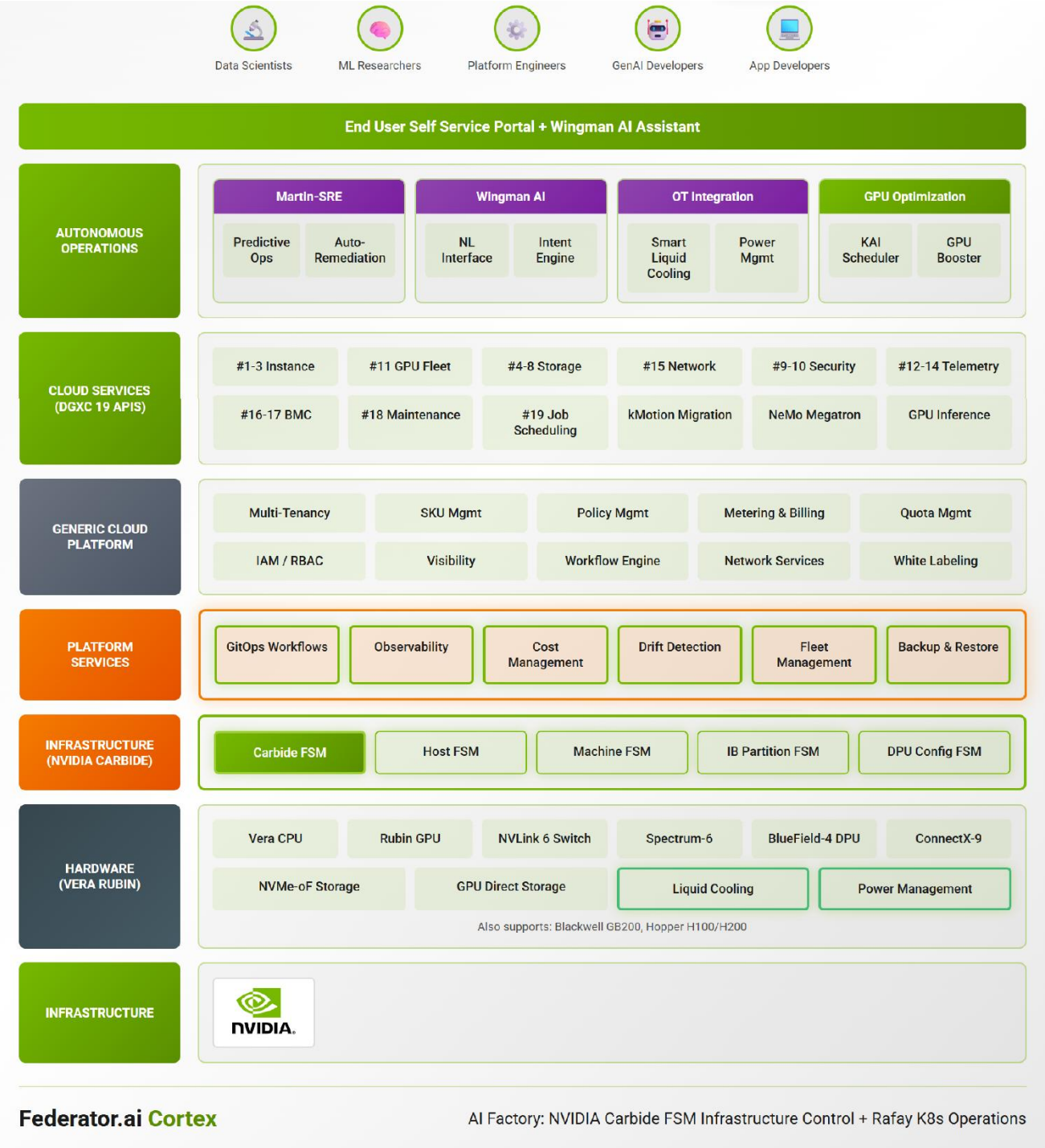
The math is unforgiving: at $25–$40K per GPU-day on H100/GB300, a 512-GPU cluster spending 6 hours chasing the wrong root cause burns $384K. Multiply across a 10,000-GPU facility, and the annual exposure exceeds $50M.

Industry GPU utilization averages below 50%. Static scheduling, thermal throttling, and fragmented tooling leave half of the most expensive compute hardware on earth sitting idle. Federator.ai Cortex exists to close this gap.

Legacy GPU Cloud vs. Federator.ai Cortex

The table below quantifies the operational transformation that Cortex delivers across every dimension of AI factory management:

Capability	Legacy GPU Cloud	Federator.ai Cortex
GPU Utilization	✗ 30–50% (industry avg)	✓ 75–95% sustained (+50pp)
GPU Scheduling	✗ Static allocation, manual	✓ Predictive 4D scheduling (patented)
Failure Detection	✗ Reactive—after incident	✓ 48-hour advance prediction, 94% accuracy
Training Failure Rate	✗ 50% caused by GPU/HBM	✓ 50% reduction via proactive remediation
Cooling Management	✗ Manual BMS, thermal guesswork	✓ AI-driven PID + feedforward, PUE 1.15
Cooling Throughput	✗ Baseline manual	✓ +45% with workload-aware optimization
Maintenance Downtime	✗ Scheduled windows, SLA impact	✓ Zero downtime (kMotion live migration)
NCP Certification	✗ 12–18 months, custom build	✓ Weeks—all 19 APIs pre-built
IT + OT Integration	✗ Siloed—separate dashboards	✓ Unified cross-layer causal analysis
Operator Interface	✗ CLI/dashboards, manual runbooks	✓ Wingman AI natural language copilot
Root Cause Analysis	✗ Single-layer, manual	✓ 12-agent Bayesian DAG, 16 failure modes
Financial Modeling	✗ Spreadsheets, external tools	✓ Built-in ROI/IRR/Monte Carlo



Federator.ai Cortex — Full-Stack AI Ops Solution: From Self-Service Portal to NVIDIA Carbide Infrastructure

Platform Architecture

Federator.ai Cortex is the only platform that delivers cross-layer causal analysis and optimization across the entire AI Factory stack. Its patented Multi-Layer Correlation engine (US Patent 11,579,933) discovers causal relationships across GPU workloads, network fabric, cooling systems, and power distribution in real time. The full-stack architecture is shown in the figure above.

DCOO — AI Factory Lifecycle Management

Cortex manages the complete AI Factory lifecycle. No other platform covers design through optimization in a single integrated system:

Phase	Key Capabilities
Design	Omniverse digital twin, CFD thermal simulation, ROI calculator, what-if scenarios, rack layout optimization
Construct	UL-certified prefab building blocks, 14-week rapid deployment, modular buildout
Operate	Autonomous operations, 12-agent Martin-SRE, Wingman AI copilot, 19/19 NCP API compliance
Optimize	Kaizen continuous improvement: Measure, Analyze, Improve, Validate, Standardize

Quantified Business Impact

- **$130M+/month Revenue Acceleration:** Every month Cortex compresses time-to-production is $130M+ in GPU compute revenue at 10,000 GPUs. Speed is not just a feature—it is the product.
- **60–80% OpEx Reduction in Engineering Headcount:** Every 10MW of AI infrastructure requires 20–40 specialized SRE engineers. Martin-SRE and Wingman AI eliminate the talent bottleneck—autonomous agents replace round-the-clock human teams.
- **15–20% Revenue Premium via NCP Certification:** NVIDIA NCP-certified AI factories command premium GPU-hour pricing. Cortex's pre-built 19 APIs deliver certification in weeks, not months—unlocking premium economics from day one.
- **$40M+/year Cooling Energy Savings:** Smart Liquid Cooling v2 reduces cooling energy by 30–40% through workload-aware flow control, eliminating overcooling during idle periods. At scale (80MW facility), savings exceed $40M annually.
- **Single Platform Replaces 8–12 Point Solutions:** One system covers monitoring, scheduling, cooling, billing, compliance, incident management, capacity planning, and financial modeling. Unified event correlation eliminates blind spots from tool fragmentation.

Deployment Specifications

Requirement	Specification
Supported GPUs	• NVIDIA H100, H200, GB200, GB300 (NVL72, NVL576, DGX SuperPOD) • Vera Rubin (roadmap) — multi-gen unified management
Orchestration	• Kubernetes v1.24+ (vanilla, OpenShift, Rancher) • Helm Charts for automated deployment, GitOps-ready
Cooling Integration	• Redfish v2.0+, IPMI v2.0, Modbus TCP/RTU • MG Cooling AC250, Supermicro SCC, generic Redfish CDUs
Networking	• NVIDIA InfiniBand 400Gb/s+, RoCE v2 • NVSwitch topology-aware scheduling, NCCL optimization
Telemetry	• Prometheus, DCGM Exporter, VictoriaMetrics, OpenTelemetry • Per-GPU: SM clock, HBM temp, NVLink, power draw
Backend	• Python 3.12+, FastAPI, SQLAlchemy 2.0 async • PostgreSQL, NATS JetStream, Redis
Frontend	• Next.js, React, TypeScript, Tailwind CSS v4 • Real-time WebSocket, glassmorphism dark-theme UI
AI/ML Runtime	• LangGraph multi-agent orchestration, Google Gemini • NVIDIA NIM, Ollama, OpenAI, vLLM (pluggable LLM provider)
Security	• JWT + API key auth, RBAC, SecretStr credential handling • Rate limiting: 120 rpm/IP, 60 rpm/tenant, TLS encryption

Industry Applications

- **AI Cloud Service Providers:** Multi-tenant GPU-as-a-Service with per-second billing, SLA enforcement, and elastic scaling.
- **Sovereign AI Programs:** National AI infrastructure with data residency, government-grade security, and national LLM training support.
- **Semiconductor & EDA:** GPU-accelerated chip design, process simulation, and yield prediction with workload-aware scheduling.
- **Healthcare & Life Sciences:** Drug discovery, genomics, clinical imaging with HIPAA-compliant multi-tenant environments.
- **Financial Services:** Low-latency risk modeling, quant trading, fraud detection with SOC2-ready compliance.
- **Research & Academia:** Foundation model pre-training, climate modeling, materials science with fair-share scheduling.