



GridCORTEX

SoftServe's AI-Powered Utility Intelligence Framework
· *Powered by NVIDIA*

Solution Architecture Guide · INTERNAL ONLY

Solution Architecture

SoftServe's AI-Powered Utility Intelligence Framework | GridCORTEX: SoftServe's AI-Powered Utility Intelligence Framework, Powered by NVIDIA | Confidential

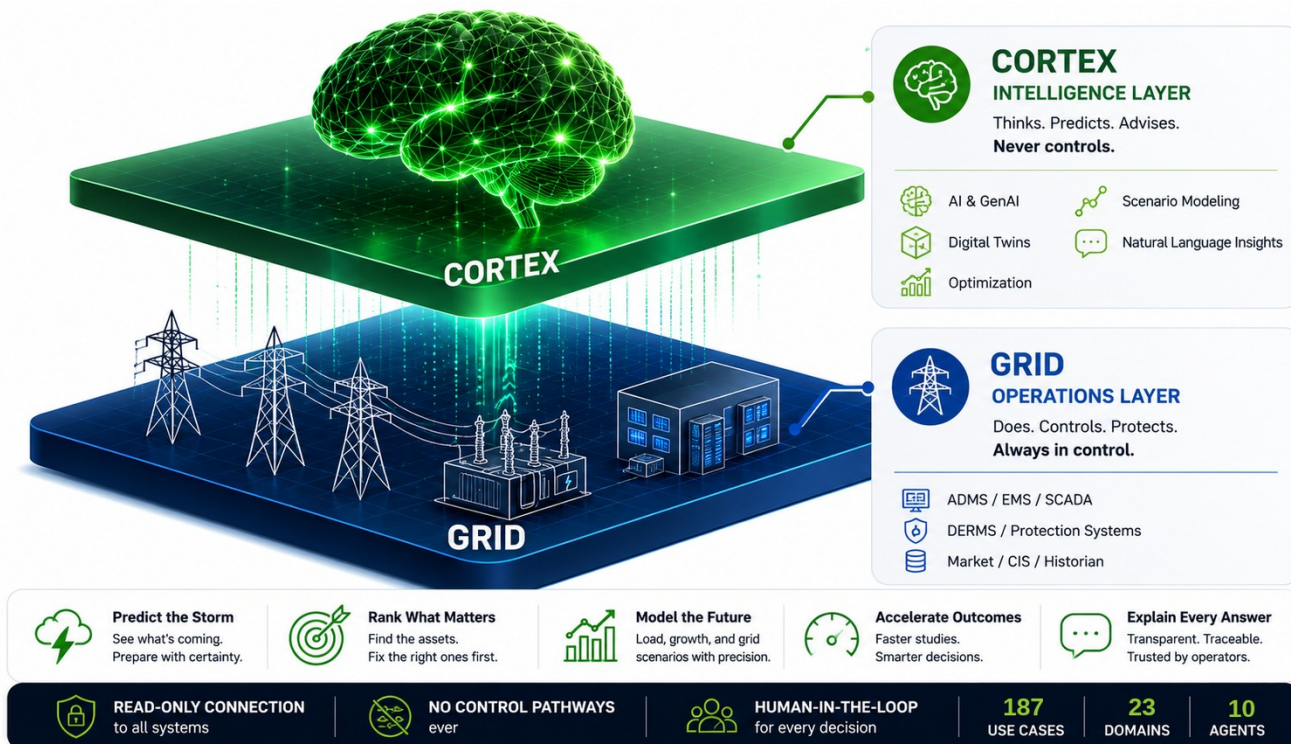
INTERNAL — SOFTSERVE + NVIDIA *For internal use and partner briefings only*

1. Executive Summary

Utilities don't buy people who know AI. They buy people who know the grid. And they don't care that your AI engineers have PhDs and worked at NASA — they want to know if you understand their pain points and can actually help solve them. GridCORTEX was built on that premise. Five forcing functions are converging simultaneously and creating the conditions for the most significant utility transformation since electrification: (1) the Data Center Load Wave — the largest load growth in a century, where utilities that answer interconnection feasibility in weeks sign the gigawatts; (2) the Aging Workforce & Infrastructure — half the experienced operators retire this decade, and aging infrastructure raises failure risk and maintenance cost at the same time; (3) Extreme Weather Mandates — storm, wildfire and flood risk compound every year, and regulators want demonstrated, auditable preparedness, not asserted plans; (4) Grid Edge & DER Transformation — electrification and DER are invalidating legacy planning, and utilities need real-time visibility and optimization at the grid edge; and (5) Regulatory Pressure & Capital Scrutiny — reliability, compliance and rate-case scrutiny are rising, and utilities must demonstrate performance and justify capital with auditable evidence. GridCORTEX is the answer to all five — simultaneously.

GridCORTEX is a ten-domain framework. GridCORTEX Robotics operates in the physical world using NVIDIA Isaac ROS 2, deploying autonomous drones, ground robots, and edge AI devices on and around live utility infrastructure. GridCORTEX Nuclear, the ninth domain, is purpose-built for nuclear operations and fleet intelligence. Using NVIDIA Isaac ROS 2, Jetson Orin, and Omniverse, Robotics deploys autonomous drones, ground robots, and edge AI devices on and around live utility infrastructure. GridCORTEX Nuclear deploys entirely in the non-safety-related domain — read-only, advisory-only, human-in-the-loop — designed to avoid NRC safety-related software qualification pathways; it should be scoped accordingly unless a client explicitly chooses a regulated safety-related use case. The library now encompasses 187 use cases across 23 operating domains.

The framework operationalizes a single thesis, proven across thirty-three years and five chairs in this industry: utilities do not have a technology problem, they have an operational adoption problem. GridCORTEX enters through the four domains where utility executives have funded budgets and urgent business cases today — Connect, Resilience, Asset, and Operations — and expands into Edge, Customer, Enterprise, and Robotics after value is demonstrated. Every architectural decision follows from that thesis: read-only OT integration, human-in-the-loop authority, explainable recommendations with citations, and operator-led design.



2. Positioning and the Gap Layer

The utility OT stack is excellent at what it was engineered for. ADMS operates the distribution grid in real time. EMS and SCADA operate transmission. OMS manages outage lifecycle. DERMS dispatches enrolled DERs. None of them was engineered for multi-day physics-informed prediction, cross-domain data fusion, population-scale optimization, natural-language reasoning over decades of operational records, or autonomous physical inspection of live infrastructure. Those capabilities live in the gap layer above OT, and that is the only place GridCORTEX lives.

Three positioning rules are inviolable and enforced in architecture, not messaging:

- Augment, never replace. GridCORTEX recommends, simulates, predicts, explains, and — through Robotics — inspects autonomously. ADMS, EMS, and SCADA execute. Control authority never moves.
- Discovery first. Every engagement begins by mapping what the utility already owns and what it does well. GridCORTEX is scoped to the gaps, never to the overlaps.
- Explainability as a feature of record. Every recommendation carries its reasoning, its data sources, and its confidence. If the people who run the grid do not trust it, nothing else matters.

3. The Ten Domain Agents



GridCORTEX Operations

The control room and field force agent. Covers control center decision support (ADMS co-pilot, alarm triage, switching validation, restoration sequencing, load shed orchestration, shift handoff), transmission operations (DLR, oscillation detection, study acceleration), and field operations (damage assessment, dispatch optimization, knowledge capture, mutual assistance, AR/XR field copilot, operator simulation trainer).

Buyer: VP Grid Operations / COO. Library domains 1, 2, 3, 18.



GridCORTEX Resilience

The extreme-event agent. Storm prediction and pre-positioning, wildfire ignition risk and PSPS scope minimization, real-time edge AI ignition detection on pole-top and tower cameras, flood and winter twins, hardening portfolio optimization, and the full vegetation management stack.

Buyer: Chief Wildfire / Resilience Officer, VP Operations. Library domains 8, 9.



GridCORTEX Asset

The fleet health and economics agent. Predictive asset twins, transformer acoustic and thermal digital twin, DGA interpretation, cable failure prediction, end-of-life economics, critical spares strategy, plus generation fleet predictive maintenance.

Buyer: VP Asset Management. Library domains 4, 10.



GridCORTEX Edge

The planning and grid-edge agent. Electrification load growth, non-wires alternatives, network model health, VVO performance audit, phase identification, DER interconnection, hosting capacity, VPP orchestration, EV siting, FERC 2222 readiness, flexible interconnection, AMI interval data intelligence, and edge AI fault anticipation at distribution automation nodes.

Buyer: VP Distribution Planning / DER. Library domains 5, 6, 17.



GridCORTEX Customer

The customer and revenue agent. Grid-aware chatbot, agent copilot, billing exception resolution, non-technical loss and revenue protection analytics, rate design simulation, vulnerable customer outreach.

Buyer: Chief Customer Officer. Library domain 7.



GridCORTEX Connect

The mission-critical large load agent and flagship wedge. Data center interconnection modeling, co-location feasibility, flexible load orchestration, tariff and contract structuring, large load resilience twins.

Buyer: SVP Strategy / Transmission Planning / Strategic Accounts. Library domain 12.



GridCORTEX Enterprise

The compliance, security, and intelligence agent. NERC CIP evidence automation, OT network monitoring, incident response, regulatory filing and DR factories, docket intelligence, markets analytics, safety analytics, supply chain intelligence, executive and board intelligence, M&A diligence.

Buyer: CISO, VP Regulatory, CFO; VP Markets / Chief Trading Officer for domain 11 (Energy Markets and Trading). Library domains 11, 13, 14, 15.



GridCORTEX Robotics

The physical AI and autonomous inspection agent. Deploys NVIDIA Isaac ROS 2-powered autonomous drones, ground robots, crawlers, and edge AI devices on and around live utility infrastructure. Use cases: transmission line drone inspection, substation autonomous inspection robot, underground cable and vault crawler, ROW

vegetation scout, robotic live-line maintenance assistant, storm damage reconnaissance drone swarm, facility security robot, solar and wind farm inspection fleet, and first-responder reconnaissance robot for downed conductors.

Packaged and sold as an add-on to GridCORTEX Asset or GridCORTEX Resilience. Standalone entry viable for Track B utilities.

Buyer: VP T&D Operations, VP Asset Management, Chief Safety Officer, Chief Wildfire Officer. Library domain 16, plus additions to domains 3 and 4.



GridCORTEX Nuclear

The nuclear fleet intelligence agent. Life extension economics optimization, RPV and primary component condition analytics, steam generator tube degradation modeling, spent fuel pool thermal management, outage critical path optimization, capacity factor and forced outage reduction, operator knowledge capture, nuclear-data center co-location and PPA structuring, INPO/WANO performance benchmarking, 10 CFR 50.59 change screening, NRC license renewal intelligence, nuclear security analytics, SMR siting feasibility, advanced reactor digital twins, and licensed operator succession planning.

Deploys entirely in the non-safety-related domain — read-only, advisory-only, human-in-the-loop. GridCORTEX Nuclear is designed for non-safety-related, advisory-only workflows and should be scoped to avoid NRC safety-related software qualification pathways unless a client explicitly chooses a regulated safety-related use case. That architectural invariant is what makes nuclear AI deployment viable without a multi-year regulatory process.

Buyer: Chief Nuclear Officer, VP Nuclear Engineering, VP Nuclear Operations, CFO. Library domain 19.



GridCORTEX ISO/RTO Markets

The system operator agent — GridCORTEX applied to the seven U.S. ISOs/RTOs as institutions. Interconnection queue study acceleration under FERC Order 2023 (intake validation, screening, restudy optimization, report drafting), market rules and tariff copilot grounded in the governing corpus, resource adequacy and extreme weather scenario twins, operator-side congestion and constraint forecasting, market surveillance and anomaly detection at population scale, Order 2222 DER aggregation operations automation, and joint blackstart restoration drill simulation for the annual multi-utility exercise.

Deploys read-only above the market stack — EMS/MMS, settlement, and study systems remain authoritative; no market decision authority. Track B entry via the tariff copilot (public documents, no OT integration); Track A expansion into study, adequacy, and surveillance workloads. Per-market positioning lives in the ISO/RTO Market Context Matrix (Doc 11).

Buyer: ISO/RTO COO, VP Transmission Planning, VP Market Operations, VP System Planning, Market Monitor. Library domain 20.

4. Reference Architecture

Layer 0: Systems of Record

ADMS, EMS, SCADA, OMS, DERMS, GIS, EAM/WMS, AMI/MDM, CIS, ERP, and market systems remain the authoritative systems. GridCORTEX consumes from them through read-only gateways and returns recommendations into their existing workflows for human review and execution. Nothing in the framework writes to a control system.

Layer 1: Data and Integration Fabric

Streaming ingestion (PMU, AMI interval data, alarms, events), batch and document ingestion (records, filings, work orders), imagery and LiDAR pipelines from both fixed sensors and autonomous robotic platforms, historian and PI reads, GIS and asset data, EAM and WMS feeds, and weather and market data — organized around a semantic grid model. Omniverse provides the common visualization environment and digital twin foundation for robotics mission planning and replay. Note: SCADA, OMS, and other OT control systems remain authoritative in Layer 0 and are connected read-only at the ingest boundary — their event and telemetry streams flow into the fabric, but the control systems themselves do not become part of it.

Layer 2: Intelligence Engines (NVIDIA Accelerated)

Earth-2 for kilometer-scale weather and climate scenarios. Isaac ROS 2 for autonomous robotics navigation and orchestration. Jetson Orin for on-device inference at edge nodes and robotic platforms. PhysicsNeMo and GPU-parallel solvers for grid physics at simulation scale. cuOpt for fleet, portfolio, and schedule optimization. RAPIDS and cuML for population-scale analytics and machine learning. Metropolis and TAO for vision AI. NeMo, NIM microservices, NeMo Retriever, and NeMo Guardrails for generative AI grounded in utility records. Morpheus for cybersecurity pipelines. TensorRT for on-device inference optimization at distribution automation nodes.

Layer 3: Domain Agents and Orchestration

Ten agents, plus the agent runtime: orchestration across agents, guardrails, role-based authority, explainability and citation enforcement, confidence scoring, and human-in-the-loop checkpoints. The Robotics agent adds a physical task coordination layer — mission planning, robot dispatch, and telemetry ingestion. Robots observe and report; humans decide and act. The Nuclear agent deploys entirely in the non-safety-related domain — advisory-only — designed to avoid NRC safety-related software qualification pathways.

Layer 4: Experience

Role-fit surfaces: control room console, storm and EOC command views, field mobile with voice and AR/XR overlay, robotics mission planning and monitoring dashboard, executive briefing, customer channels, and APIs for embedding into existing tools.

Cross-Cutting

Zero-trust security aligned to NERC CIP boundaries. MLOps and model governance with model cards and drift monitoring. Full decision audit trail including robotics mission logs and inspection records. Deployment on-premises GPU infrastructure, sovereign or commercial cloud, or hybrid. Operator-led design and adoption management as a formal workstream in every engagement.

GridCORTEX — 5-Layer Reference Architecture

Above OT • below the enterprise • read-only • advisory • human-in-the-loop



SoftServe's GridCORTEX • powered by NVIDIA • 187 use cases • 23 domains • 10 agents

5. Safety, Trust, and Governance Principles

- No control, ever. The framework holds no control authority and no control pathways exist in the architecture. Robotic platforms observe and report; all consequential actions are human-authorized. This is a design invariant, not a configuration.
- Explainability is mandatory. Every recommendation is explainable, sourced, and confidence-scored. Unexplainable outputs do not ship.
- Human authority is explicit. Named human roles hold decision authority at every consequential checkpoint, and the audit trail records who decided what on which evidence.
- Models are governed. Versioned, monitored for drift, validated against ground truth on a schedule, and retired when they degrade.
- Data sovereignty. Customer data, CEII, and CIP-scoped information are handled within the utility security boundary under utility-controlled keys.

6. Delivery and Commercial Model (Internal)

Three-step motion per domain, led by the four lead domains (Connect, Resilience, Asset, Operations): Discovery Workshop (paid, fixed-fee, \$50,000–\$75,000, two to four weeks) → Scoped Pilot (\$250,000–\$500,000, three to six months, one region, one use case, one KPI agreed in writing) → Production Deployment (\$500,000–\$1,500,000, first domain) → Managed Services (\$750,000–\$2,500,000 annually depending on utility size and scope, structured as a SoftServe managed services engagement, auto-renewing). Full commercial model and tier structure in 09_GridCORTEX_GTM_and_Managed_Services_Model. Robotics activates after lead domain trust is established, entering accounts through Asset, Resilience, and Operations rather than as a standalone initiative.

GridCORTEX Robotics pilots follow the same three-step structure. Replay pilots — running the AI model against archived helicopter or ground inspection imagery — are the fastest path to proof and require no hardware deployment to generate the first value demonstration.

The compound integration advantage: the second agent deployed costs materially less to integrate than the first because the data connections, semantic grid model, and intelligence engine infrastructure are already in place. SoftServe passes a portion of that cost reduction through to reward expansion.

Ronnie S. Mauldin | [linkedin.com/in/ronniemauldin](https://www.linkedin.com/in/ronniemauldin) | ronniemauldin.com | Internal use only — do not distribute externally

Agent Runtime & Governance — NVIDIA Agent Toolkit with NemoClaw

The ten GridCORTEX agents are delivered on the NVIDIA Agent Toolkit with NemoClaw — NVIDIA's open blueprints for building always-on, domain-specialized autonomous agents (announced June 2026). This layer sits beneath the GridCORTEX agent tier in the reference architecture and above the NVIDIA engine tier. Its strategic importance for P&U is simple: it converts GridCORTEX's operating promises — read-only OT access and human-in-the-loop approval — from architectural commitments into runtime-enforced, auditable controls.

Components and Their P&U Role

- **OpenShell Secure Runtime** — every GridCORTEX agent executes in a sandbox governed by signed policies that define what the agent may see, what it may do, and where its inference runs. Security is enforced in the environment layer, not the model or application layer. This is the direct technical answer to NERC-CIP and OT-security review: human-in-the-loop is not a promise, it is a signed policy the runtime enforces.
- **NeMo Relay Observability** — an open middleware layer that captures structured traces of every tool call, skill invocation, and model call an agent makes, in open formats (OpenTelemetry-class). For a utility this is compliance evidence produced as a side effect of operation: replayable audit trails for NERC reviews, rate-case data requests, and post-event analysis, plus 30–50%% runtime cost reduction through caching and routing optimization.
- **Nemotron 3 Open Models** — frontier-class open reasoning models (Ultra/Super/Nano) that deploy fully on-premises or in the client's own tenancy. This is the deployment answer for air-gapped nuclear environments, CIP-constrained control-room adjacency, and any client for whom cloud frontier models are a non-starter.
- **CUDA-X Skills and AI-Q** — NVIDIA libraries exposed to agents as signed, versioned skills: cuOpt for routing/scheduling/optimization, NeMo Retriever for enterprise retrieval, and the AI-Q blueprint for multi-agent deep research. AI-Q is the named blueprint behind GridCORTEX's research-heavy use cases (Regulatory Intelligence Monitor 14.3, Rate Case Data Request Factory 14.4, Market Rules & Tariff Copilot 20.2, M&A Diligence 15.6, NRC License Renewal Intelligence 19.4).
- **Skill Evolution (Hermes), P&U-safe framing** — agents may draft improvements to their own skills from operator feedback, but every skill is signed and human-approved before deployment — mapping one-to-one onto utility change management. This mechanism also powers UC 3.7: a retiring operator's judgment captured as signed, versioned agent skills.

Positioning Rule

The runtime layer is described to clients as how GridCORTEX agents are governed, never as a product the client must license or operate. GridCORTEX remains a framework delivered through engagements; NemoClaw, OpenShell, and Relay are the engineering standard underneath it. The one-sentence version for a security audience: "Human-in-the-loop isn't a promise — it's a signed policy the runtime enforces, with an audit trail."