White Paper

Meeting Security & Data Sovereignty Requirements in Mission-Critical Utility Networks

Navigating AI, Cloud, OT Security, and Compliance in Mission Critical Utility Infrastructure

David Stokes
Head of IP Portfolio and Solutions
Ribbon Communications

› Your Path to Autonomous Networks

The Mission-Critical Utility Dilemma: To Cloud or Not to Cloud

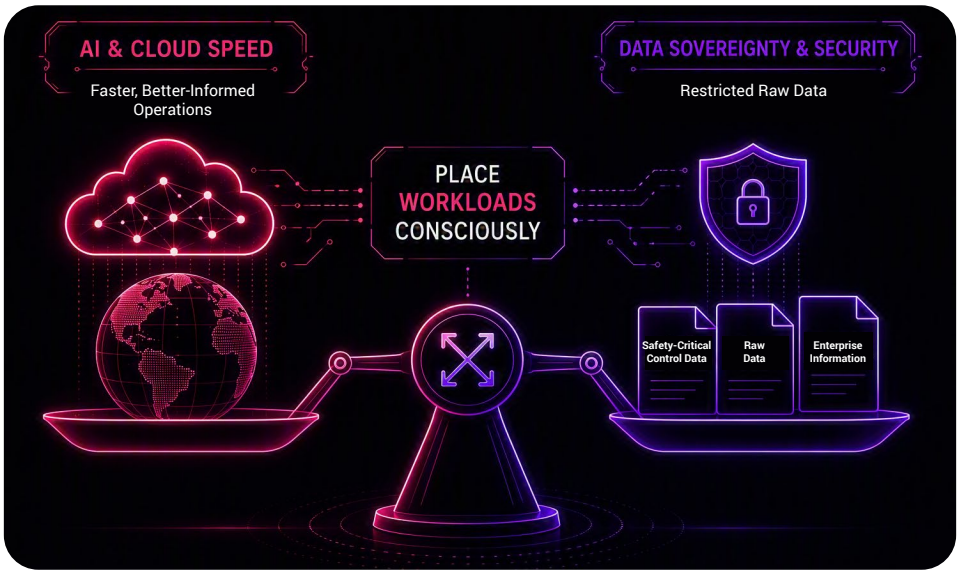
Electric, water, gas, transport and other mission-critical utility operators are under pressure to modernise. AI-driven decision support, predictive maintenance, automation and cloud-scale analytics promise faster, better-informed operations across control centres, field operations, substations, plants, pipelines and distributed assets. At the same time, security, sovereignty, resilience and regulatory requirements have never been more important for utility data and operational technology (OT) environments.

- AI is accelerating digital transformation across utility operations, asset management, field service, grid control, water networks and other critical infrastructure domains.
- Cloud adoption continues to grow, but mission-critical utility workloads must be balanced against OT security, operational availability, data residency and regulatory obligations.
- Security and sovereignty requirements — public safety, service continuity, regulatory compliance and operational control — must be designed into the architecture from the start.

Data sovereignty is more than location. For utilities, it means knowing where operational data resides, who can access it, which laws and industry obligations apply, and how control is maintained across IT, OT and cloud environments.

- Operational technology security and cyber resilience
- Service continuity, even during outages, cyber incidents or constrained connectivity
- Regulatory, national infrastructure and sector-specific compliance

The underlying paradox has to be faced. AI and cloud computing promise faster operational insight and better automation, but they can raise data-residency, control-plane and shared-responsibility concerns. On-premises, private cloud and edge deployments offer stronger governance and resilience, albeit with more operational ownership.



One of the most important developments in recent years is that utilities no longer need hyperscale cloud access to deploy useful AI capabilities.

For mission-critical utilities, capable AI can be deployed entirely within sovereign, air-gapped, off-net or disconnected environments while maintaining control over data, models, operational workflows and service continuity.

Under the Hood: Models, Metadata, and Trust – LLMs Architecture

A LLM does not retrieve facts from a database; it calculates the most likely next word based on learned parameters. Training data becomes internalised knowledge, and a prompt triggers inference that yields small, decision-grade outputs. This principle makes sovereign practical for utilities: a model can be moved into a controlled environment, while sensitive operational data remains where it is governed and protected. Where outputs, model updates or metadata must cross from an OT or utility zone to enterprise systems, a data diode can enforce one-way transfer, ensuring information can flow out through a controlled boundary without creating a return path into the operational environment.

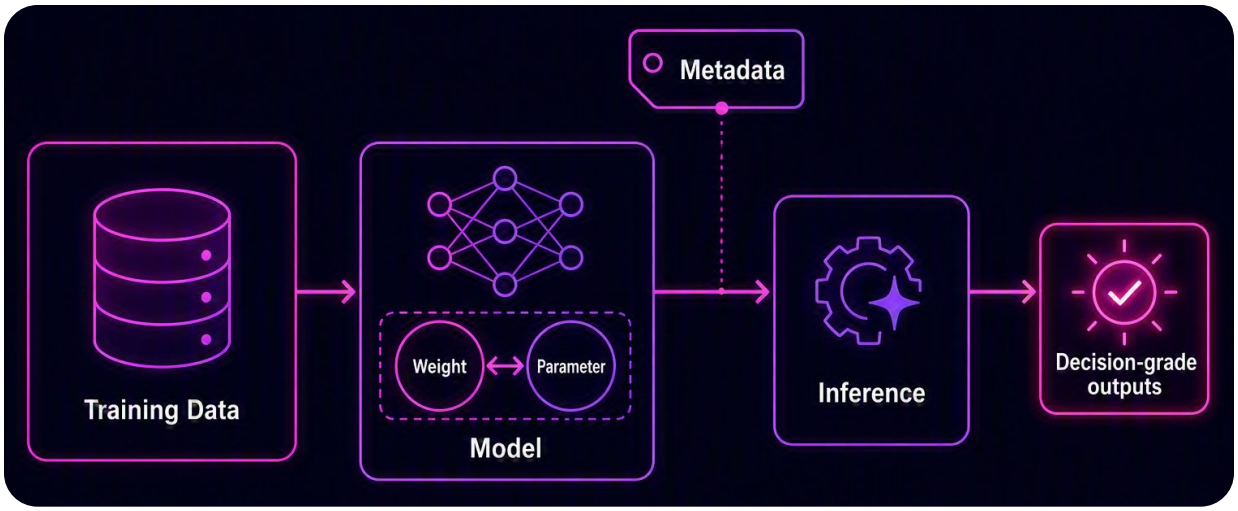


Figure 2 — LLM architecture: training data, model (weights/parameters), metadata, inference, and decision-grade outputs.

Core Building Blocks

- **Training Data** — the large corpus of text, telemetry, and documents the model learns from.
- **Model** — the trained neural network itself: a large file containing everything the AI has learned. It doesn't store facts or text directly; it stores a structure of interconnected values that together let it predict language, patterns, and relationships.
- **Parameter** — internal numerical values of an AI model, learned during training, representing "knowledge" and extracted patterns.
- **Weight** — a specific type of parameter that sits on the connections between artificial "neurons" in the network. A weight controls how strongly one part of the model influences another; higher or lower values make certain words, patterns, or relationships more or less likely to be selected during prediction.
- **Metadata** — the classification, provenance, and policy wrapper that travels alongside the model or data, governing what can move where.
- **Inference** — the model applies its learned parameters against a prompt or input.
- **Decision-grade outputs** — the final, lightweight, actionable result — embeddings, alerts, or recommendations — negligible compared to the raw data that started the flow.

Imagine a language model as a massive network of billions of connections. Each connection has a numerical value, often called a "weight." During training, these values are constantly adjusted so that the model can make better predictions:

- **Before training:** The model does not know that "A transformer is used to …" is usually followed by "change voltage levels in an electrical grid."
- **After training:** The parameters have been adjusted so the model is highly likely to associate transformers with voltage conversion and grid operation.

The Physics – Policy – Trust Framework

A tailored framework helps utility network architects decide what may cross constrained or regulated environments, and what must stay within the operational boundary.

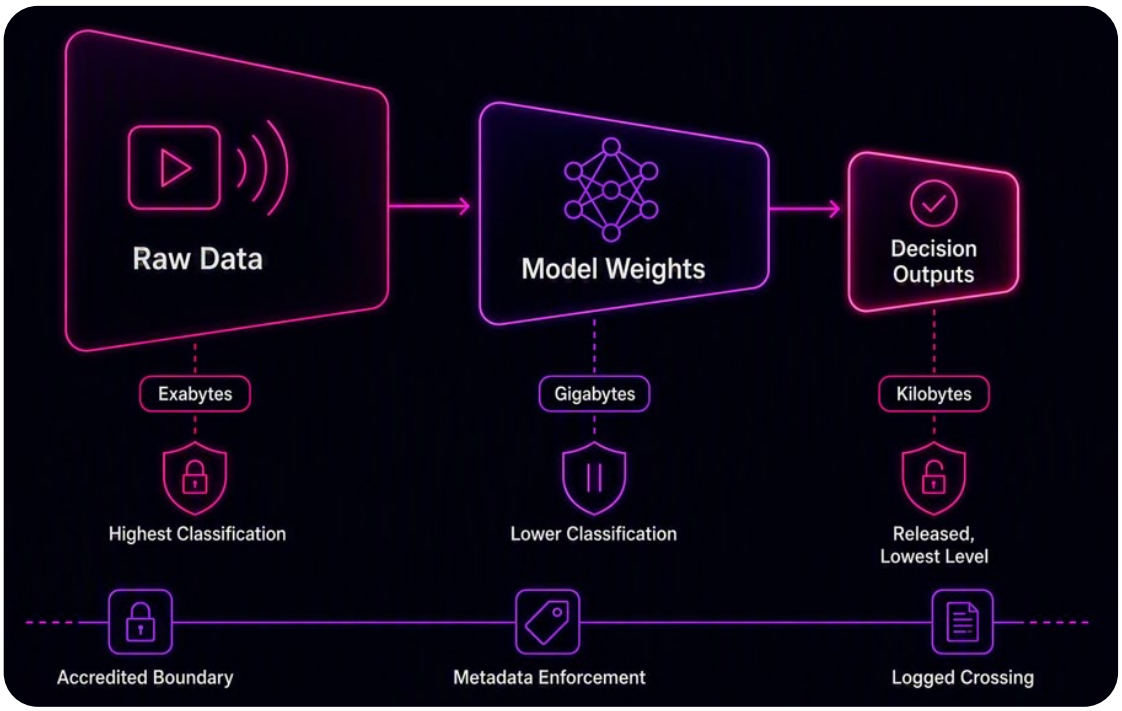


Figure 3 — Physics-Policy-Trust: raw data, model weights, and decision outputs across an accredited boundary

Physics

- Raw data is exabyte-scale and stays put
- Model weights are gigabyte-scale and can move under policy
- Decision outputs are kilobyte-scale and release at the lowest appropriate classification.

Policy

- Raw operational and telemetry data can be high-volume and should usually stay put
- Model weights are smaller and can move under policy
- Decision outputs are lightweight and can be released to the right operational systems under governance.
- Metadata is the mechanism that makes this enforceable at machine speed — every packet, every gradient, every prompt carries its own policy label.

Trust

- Raw operational data carries strict security, privacy, safety and regulatory constraints.
- Models trained on sensitive utility data may require controls based on their training source, operational use and risk profile.
- Decision outputs can be released more broadly when policy and trust controls allow it.
- Metadata is the mechanism that makes this enforceable at machine speed — every flow, model update, prompt or result carries its own policy label.
- Raw data remains inside the operational boundary, so governance and operational control stay intact.
- Only pre-vetted artefacts cross boundaries, so integration and security controls have a smaller, well-defined problem to solve.
- Every crossing is logged, so audit, compliance and incident review are supported.

AI Solutions Do Not Have to Be Fully On-Net to Deliver Value

Three complementary classes of AI technology now allow utilities to bring meaningful intelligence to disconnected, constrained or highly regulated environments without depending entirely on hyperscale cloud.

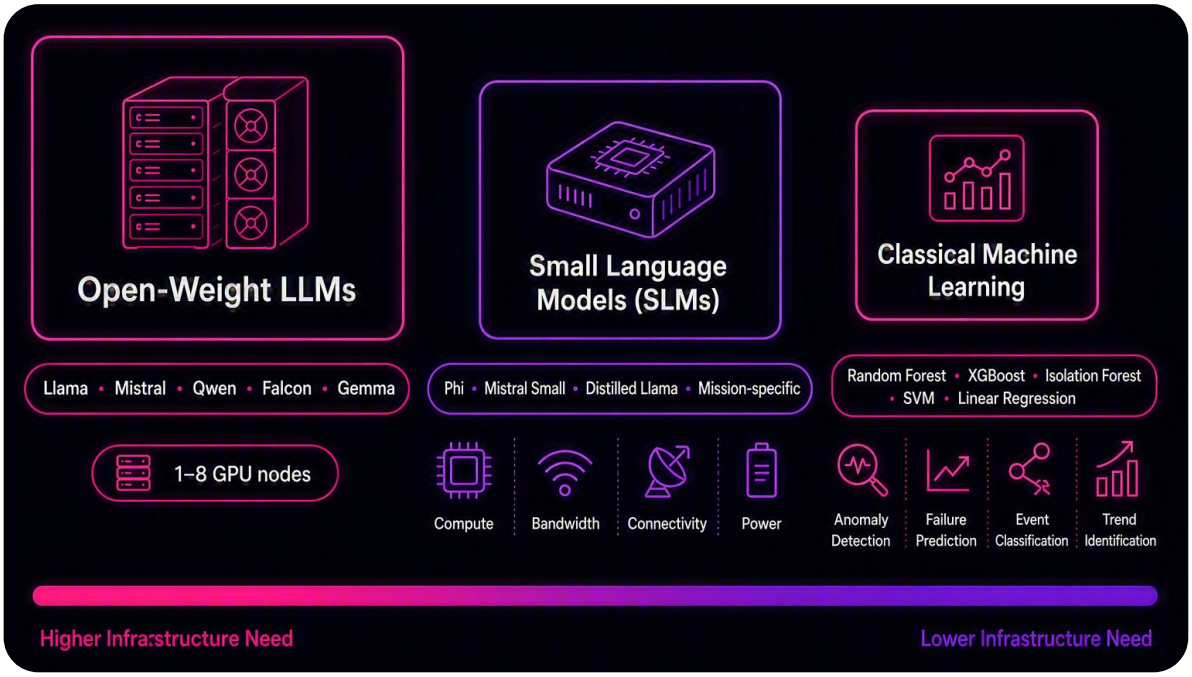


Figure 1 — Options for off-net defense environments: open-weight LLMs, Small Language Models, and classical ML.

- **Open-Weight LLMs** (Llama, Mistral, Qwen, Falcon, Gemma) run on 1–8 GPU nodes — ideal for knowledge assistants, security analytics, and OT/SCADA decision support.
- **Small Language Models (SLMs)** run on field compute nodes, ruggedized servers, and vehicle-mounted systems — enabling mission planning, report summarization, and field guidance even during contested connectivity.
- **Classical Machine Learning** remains the best choice for anomaly detection, predictive maintenance, and traffic classification — lightweight, explainable, and deployable almost anywhere.

Check out Annex A for more Details - Options for Off-Net Utility Environments Explained

AI can be deployed entirely within sovereign, air-gapped, off-net environments — maintaining full control over data, models and operational workflows.

Deployment Choices

There is no single right answer for where critical AI workloads should run. Most organizations use a deliberate mix of all four models, matched to classification and mission-criticality.

Model	Best Suited For
Public Cloud	Unsensitive administrative, training, customer-facing and non-operational workloads
Private Cloud	Sensitive enterprise applications and shared operational analytics requiring stronger governance
On-Premises	Mission-critical OT, control-room, grid, plant, pipeline and safety-related systems
Colocation	Regional utility data centres or resilient edge locations needing sovereignty and availability without full site build-out

Getting this balance wrong carries real cost: over-rotating to public cloud risks data residency violations, regulatory exposure, and loss of operational control over sensitive or mission-critical information. Over-rotating to fully isolated on-premises infrastructure risks higher costs, slower AI adoption, and reduced interoperability with interconnected operators and ecosystem partners. A layered, sovereignty-first architecture avoids both extremes.

There is no one-size-fits-all model. Data sovereignty must be part of every AI and cloud strategy from day one — not retrofitted after the fact. Security, sovereignty, and AI can coexist.

The Layered AI Architecture

Most utilities benefit from a layered approach rather than a single model — balancing AI capability, resilience, security, sovereignty and cost, and enabling meaningful AI even in disconnected or mission-critical utility environments.

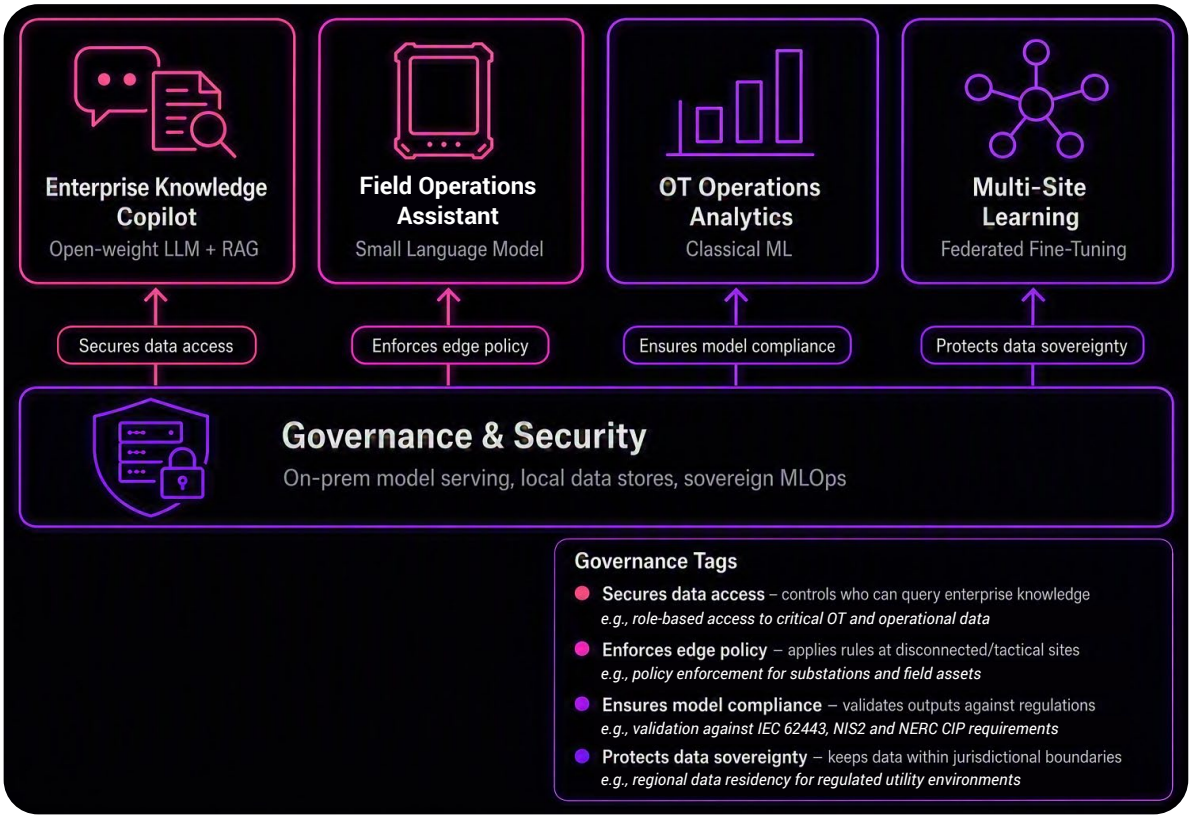


Figure 5 — Layered architecture: Enterprise Knowledge Copilot, Tactical/Edge Assistant, OT Operations Analytics, and Multi-Site Learning, governed by sovereign MLOps.

Model	Technology
Enterprise Knowledge Copilot	Open-weight LLM + Retrieval-Augmented Generation
Field / Edge Assistant	Small Language Model (SLM)
OT Operations Analytics	Classical Machine Learning
Multi-Site Learning	Federated Fine-Tuning
Governance & Security	On-premises model serving, local data stores, sovereign MLOps

MLOps — Machine Learning Operations — is the set of practices, tools, and processes that combine machine learning, DevOps, and data engineering to reliably and efficiently deploy, monitor, and maintain ML models in production. It is what DevOps is to software engineering, but applied to the ML model lifecycle: training pipelines, versioning of data, code, and models, automated testing and validation, deployment, drift monitoring, and retraining. In defense environments, sovereign MLOps ensures this entire lifecycle can run inside an accredited boundary.

Solving the Multi-Site Problem – Multi-Site Learning

Across substations, plants, control centres, pumping stations, renewable sites and field locations, raw operational data often cannot be shared freely. Federated learning solves this: each site trains locally, only model updates or weight deltas are shared, and a central coordinator aggregates them into an improved global model — raw operational data never leaves the site.

The Data Sovereignty Challenge:

- Critical utility environments with hundreds or thousands of distributed assets
- Regulations, safety obligations or operational policies restrict sharing raw data

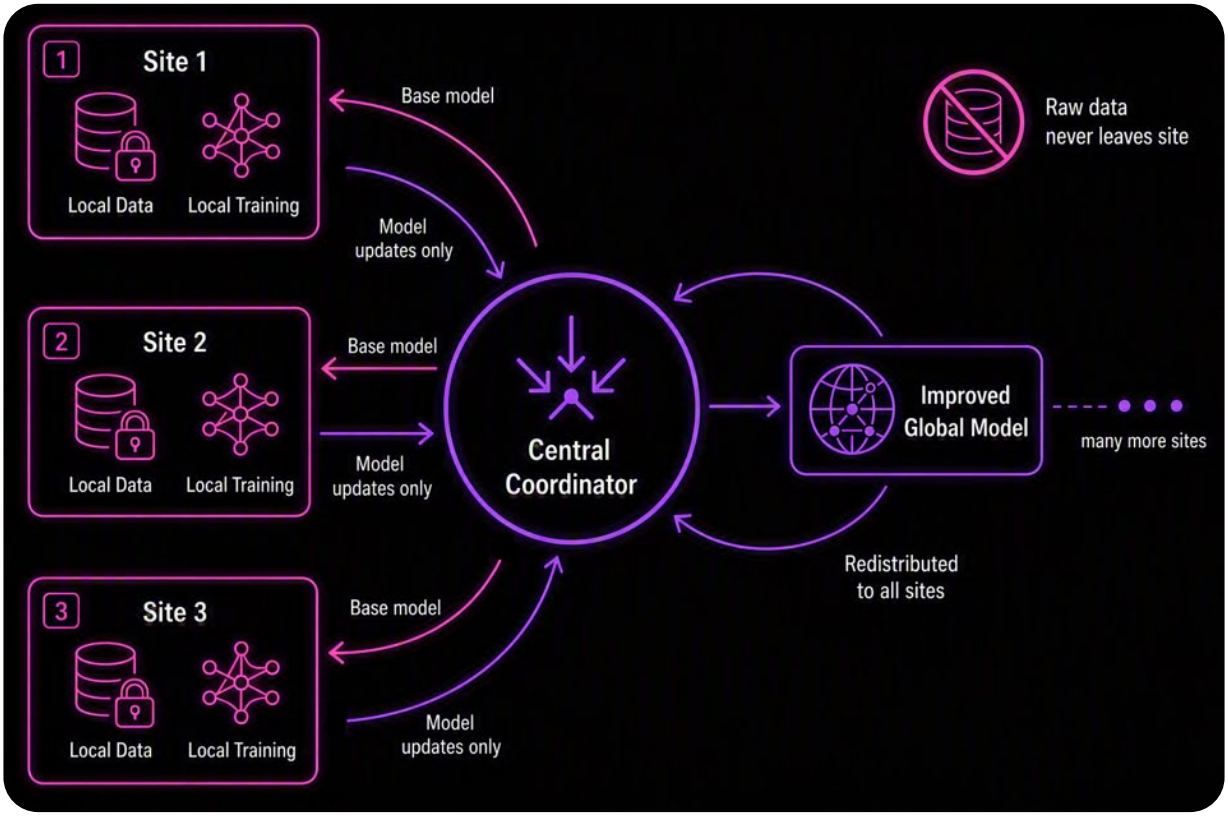


Figure 4 — Federated fine-tuning: each site keeps its local data; only model updates cross to the central coordinator.

Federated Learning to the Rescue:

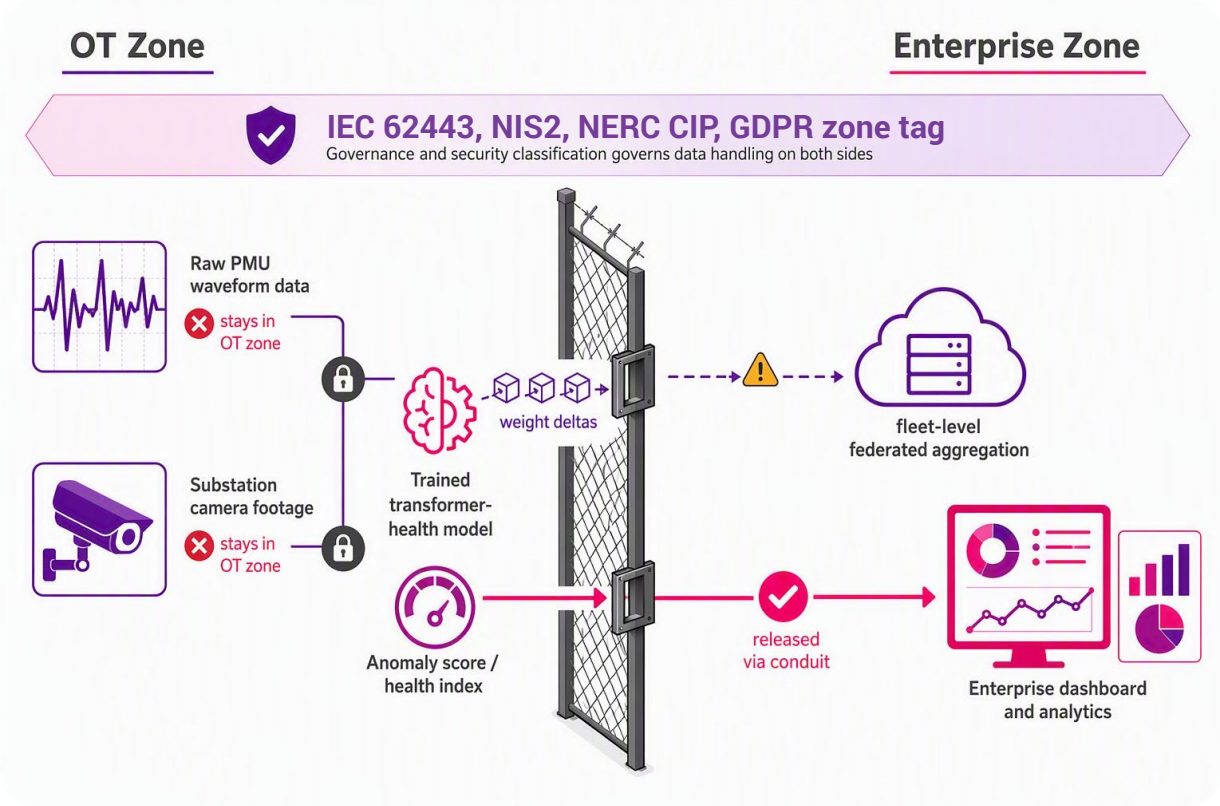
- Each site retains its local data
- A base model is distributed to all sites
- Local training occurs within the site's secure environment
- Only model updates or weight deltas are shared
- A central coordinator aggregates updates into an improved global model

Data stays; Models and Metadata move

Eventually, it comes down to the fundamental difference between commercial and sovereign models:

- **The commercial cloud model:** data moves to the compute. Raw data — documents, telemetry, images, logs — is uploaded to a cloud region where the provider's GPU clusters and pre-trained models process it, and results come back.
- **The sovereign AI model:** compute moves to the data. Only models and metadata cross boundaries. The raw data — sensitive, classified, or regulated — never leaves the enclave where it was created and where it is legally allowed to reside.

The obvious conclusion can only be that sensitive or classified information should not be handled in the public cloud, as neither the on-ramp network nor the cloud itself can be controlled.



Element	Where It Lives	Does It Move?
Raw sensor / PMU waveform data	Forward edge node (OT / field zone)	Stays in the OT / field zone
Sensor / camera footage	Forward edge node	Stays in the OT / field zone
Trained health-monitoring model	Runs at the edge	Weight deltas move to fleet-level federated aggregation
Anomaly score / health index	Output of local inference	Released via conduit to the enterprise
Classification / zone tag (e.g. IEC 62443, NIS2, NERC CIP)	Attached to every artifact	Governs enterprise handling

Mind the Transport Fabric

A sovereign AI architecture is only as strong as the network carrying it. Model distribution, federated gradient sync, and decision-output release depend on a transport fabric that is deterministic, secure, and automated — so model-sync traffic never disturbs SCADA, protection, or command channels.

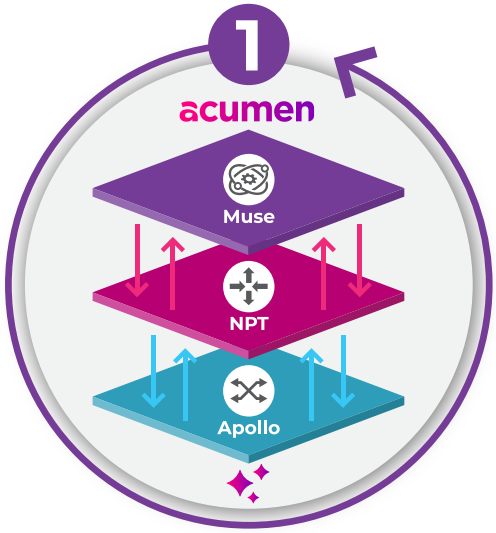


Figure 6 — Ribbon's deterministic transport fabric: Apollo optical, NPT IP/ MPLS, Muse automation, orchestrated by Acumen AIOps.

Muse™ Multilayer Automation Platform – Seamless Operation and Automation across all layers
- Intent-Based Provisioning and Rapid service activation
- Network Slicing for Mission-specific Isolation and QoS across all layers
- Advanced Analytics
- Cybersecurity & Compliance
- Cloud-Native & On-Premises
- Multi-Vendor Integration

NPT Multiservice IP/Packet platform – Resilient Multi-Protocol Support
- Redundancy and High Availability Architecture
- Deterministic Transport
- Legacy Migration Support
- MACsec, IPsec, and segment routing across latest and legacy protocols

Apollo Optical platform - Scalable, Adaptive, Secure and Programmable
- High-Capacity Future-Proof
- Mission-Critical Security
- Resilience & Survivability
- Multi-Service Support

Acumen AIOps - Real-time analysis converting data into insights and automated action
- Operates across multiple network layers and suppliers
- Analyzes real-time data from multiple sources for end-to-end observability
- Converts data into insights then actions, using automated AIOps workflows, at scale
- Leverages autonomous AI agents and models to detect, decide, and act

Conclusion

There is no one-size-fits-all model for utility AI. Public cloud, private cloud and on-premises deployments each have advantages, and the most resilient organisations deploy a deliberate, layered mix of the three - with data sovereignty built in from day one.

With open-weight LLMs, SLMs at the edge, classical ML for OT analytics, federated learning across distributed assets, and a deterministic transport fabric such as Ribbon's Apollo, NPT, Muse and Acumen platforms, mission-critical utilities can bring the power of AI to bear - without compromising the sovereignty, safety or resilience of the operational data that keeps essential services running.

- **No single deployment model wins.** Public cloud, private cloud, on-premises, and colocation each suit different workloads; most utilities need a deliberate mix, matched to operational criticality and regulatory risk.
- **AI no longer requires the hyperscale cloud.** Open-weight LLMs, Small Language Models (SLMs) and classical machine learning can run within sovereign, air-gapped, and disconnected environments — from utility data centres to substations, plants and field sites.
- **Federated learning solves the multi-site problem.** Distributed assets can jointly improve a shared model by exchanging only model updates — raw operational data never leaves the site.
- **"Data stays; models and metadata move."** Only vetted model weights, deltas, and small decision-grade outputs cross classification boundaries — raw data remains in its governed environment.
- **The network is an active enforcer, not a passive carrier.** Ribbon's Apollo (optical), NPT (IP/MPLS), Muse (automation), and Acumen (AIOps) platforms deliver the deterministic, secure, automated transport fabric that makes sovereign AI operationally viable

The real question is not "to cloud or not to cloud," but which workloads belong where, and how do we maintain control while enabling innovation.

Contact Us Contact us to learn more about Ribbon solutions.

Annex A
Options for Off-Net Utility Environments Explained

Open-Weight LLMs on Sovereign Infrastructure

Modern open-weight large language models such as Llama, Mistral, Qwen, Falcon, and Gemma have dramatically reduced the infrastructure barrier for enterprise AI deployments. Rather than requiring thousands of GPUs, many practical deployments can operate on a single GPU server or a small cluster of 2–8 GPU nodes. These models are well suited to:

- **Knowledge assistants** (manuals, operational playbooks) — the model never needs internet access because all source information resides within the organization's own environment.
- **Security analytics** — correlating security events, explaining alerts, summarizing incident reports, investigating attack chains, and generating remediation recommendations, significantly reducing analyst workload in SOCs and NOCs.
- **OT and SCADA operations** — explaining alarms, summarizing network health, interpreting telemetry, and supporting operator training, with the LLM serving as a decision-support tool rather than controlling operational systems directly.

Small Language Models (SLMs) at the Field Edge

Small Language Models — such as Microsoft Phi, Mistral Small, distilled Llama variants and specialised operational models — are optimised for environments where compute is constrained, bandwidth is limited, connectivity is intermittent and power consumption matters. SLMs can run on edge gateways, ruggedised servers, field devices, mobile service platforms and networks-in-a-box, enabling AI capabilities directly at the operational edge rather than requiring reach-back connectivity to a central data centre. Utility applications include:

- Outage and restoration planning assistance
- Local document and procedure search
- Incident report summarisation
- Field maintenance guidance
- Sensor and telemetry interpretation
- Emergency response procedures

Because inference occurs locally, operators and field teams continue to receive AI assistance even during communications outages or constrained network conditions.

Classical Machine Learning: Often the Best Choice

Despite the excitement around generative AI, many OT and critical infrastructure challenges — including electricity, water, gas, transport and other utility networks — are still best solved using traditional machine learning techniques. Detecting anomalies, predicting failures, classifying events and identifying trends frequently do not require an LLM at all.

- **Anomaly detection** — abnormal power flow patterns, equipment degradation, unusual SCADA activity, communication failures, and cyber intrusion indicators, using lightweight, explainable statistical and machine learning models.
- **Predictive maintenance** — forecasting transformer failures, router degradation, optical transport issues, mechanical wear, and battery replacement cycles, moving operators from reactive to proactive maintenance.
- **Traffic classification** — identifying traffic patterns, detecting unauthorised protocols, recognising malicious behaviour and baselining normal operations across utility networks, often with only modest compute resources.

About Ribbon

Ribbon Communications (Nasdaq: RBBN) is a global provider of voice communications software, IP routing, and optical networking to mobile and wireline service providers, enterprises, critical infrastructure and defense sectors. We support our customers' Path to Autonomous Networks by leveraging the latest AIOps automation platforms and Agentic AI technologies, helping them deliver better customer experiences, reduce operational costs, and achieve sustainable growth. To learn more about Ribbon visit rbbn.com.

Copyright © 2026, Ribbon Communications Operating Company, Inc. ("Ribbon"). All Rights Reserved. v0926