White Paper

Meeting Security & Data Sovereignty Requirements in Defense Networks

Navigating AI, Cloud, and Compliance in Mission-Critical Defense Infrastructure

Andreas Hegers
Senior Director Global Marketing
Ribbon Communications

› Your Path to Autonomous Networks

The Defense Dilemma: To Cloud or Not to Cloud

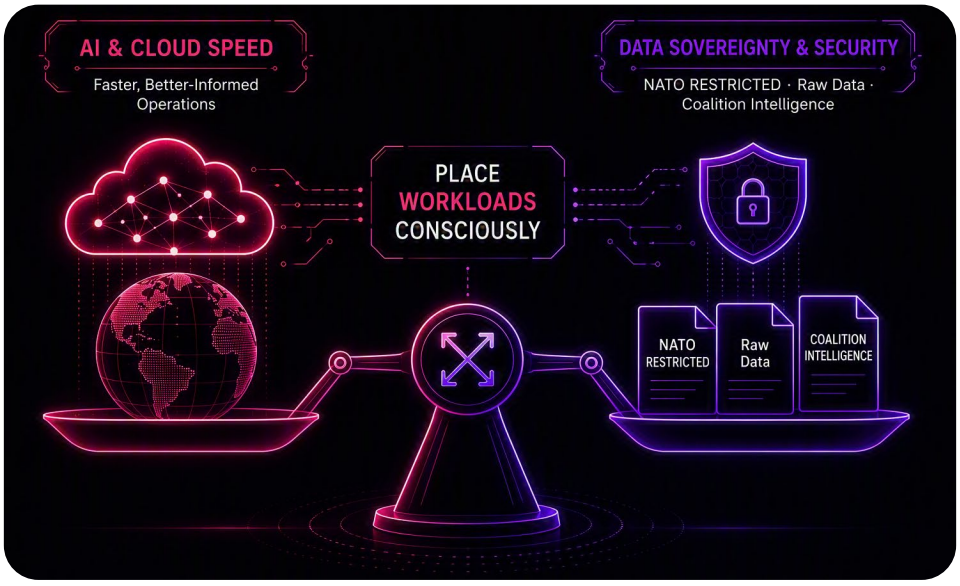
Defense ministries, armed services, and coalition partners are under constant pressure to modernize. AI-driven decision support, predictive logistics, and cloud-scale analytics promise faster, better-informed operations across the full spectrum of conflict — from garrison administration to multi-domain command and control. At the same time, security and sovereignty requirements for defense data have never been more stringent, driven by classification regimes, coalition releasability rules, and an increasingly contested cyber and electromagnetic environment.

- AI is accelerating digital transformation across defense logistics, ISR, and command networks.
- Cloud adoption continues to grow, even in government and defense communities, driven by cost and agility pressures.
- Security and sovereignty requirements — national security, operational availability, and regulatory compliance — have never been more important.

Data Sovereignty is more than location. It means knowing where data resides, who can access it, which laws and classifications apply, and maintaining unbroken control — not just where a server sits.

- National security and operational security (OPSEC)
- Mission availability, even under contested or denied conditions
- Regulatory and multinational compliance across coalition partners

The underlying paradox has to be faced. AI and cloud computing promise faster, better-informed operations and scalability but raise data residency and shared-responsibility concerns. On-premises and private cloud offer stronger governance and sovereignty, albeit at the cost of higher operational overhead.



One of the most important developments over the last years is that organizations no longer need hyperscale cloud environment access to deploy useful AI capabilities.

For defense, utilities, public safety, transportation, and other critical infrastructure operators, capable AI can now be deployed entirely within sovereign, air-gapped, off-net, or disconnected environments while maintaining full control over data, models, and operational workflows.

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

An LLM does not retrieve facts from a database; it calculates the most likely next word based on billions of learned parameters. Training data becomes internalized "knowledge," and a prompt triggers inference that yields small, decision-grade outputs — exactly the principle that makes sovereign AI practical: a 70-billion-parameter model is roughly 140 GB, small enough to move across a data diode into a classified enclave, even though its training corpus may be petabytes.

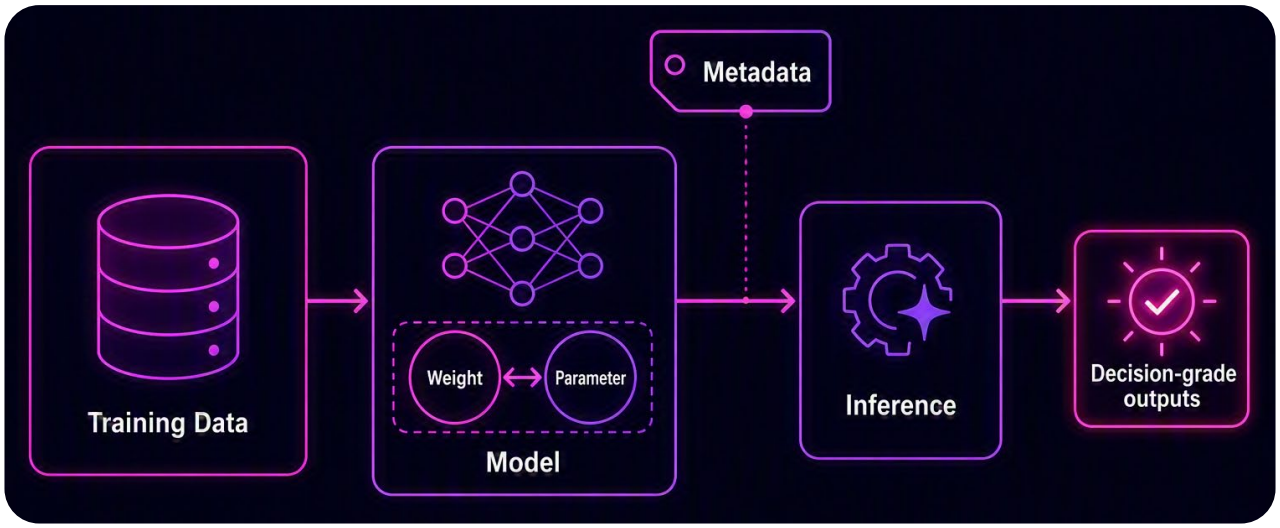


Figure 2 — LLM architecture: training data, model (weights/parameters), metadata, inference, and decision-grade outputs.

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 lives on the connections between artificial "neurons" in the network.
- **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 "The capital of Germany is ..." is usually followed by "Berlin."
- **After training:** The parameters have been adjusted so that the model is highly likely to select "Berlin."

The Physics – Policy – Trust Framework

A tailored framework helps network architects decide what may cross constrained or contested networks.

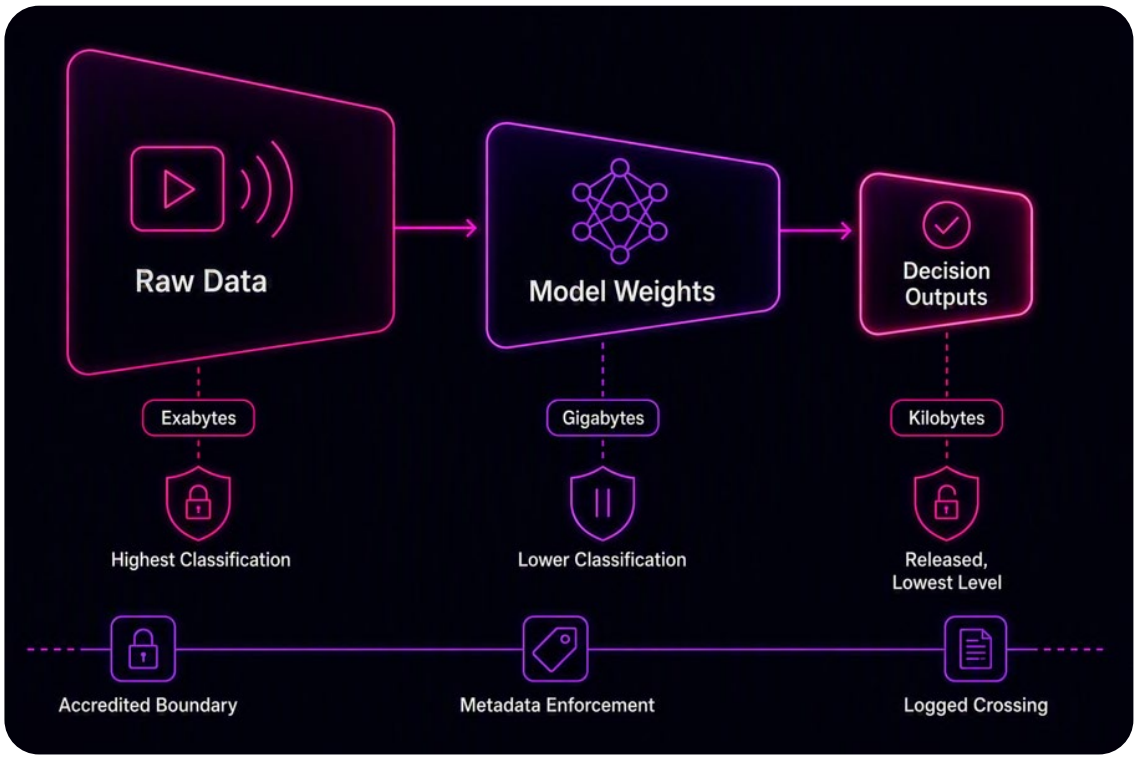


Figure 3 — Physics-Policy-Trust: raw data, model weights, and decision outputs across an accredited boundary

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 data carries the highest classification and the strictest jurisdictional constraint.
- Models trained on classified data may still be classified — but often at a different, sometimes lower, level than the source data, depending on accreditation.
- Decision outputs can be released at an even lower level after a release engine review.
- Metadata is the mechanism that makes this enforceable at machine speed — every packet, every gradient, every prompt carries its own policy label.

Trust

- Raw data never leaves an accredited boundary, so the accreditation stays valid.
- Only pre-vetted artefacts cross boundaries, so the cross-domain guard has a small, well-defined problem to solve.
- Every crossing is logged, so after-action reconstruction is possible.

AI Solutions don't have to be On-Net to deliver value

Three complementary classes of AI technology now allow defense organizations to bring meaningful intelligence to disconnected, contested, or classified environments without requiring a hyperscale cloud.

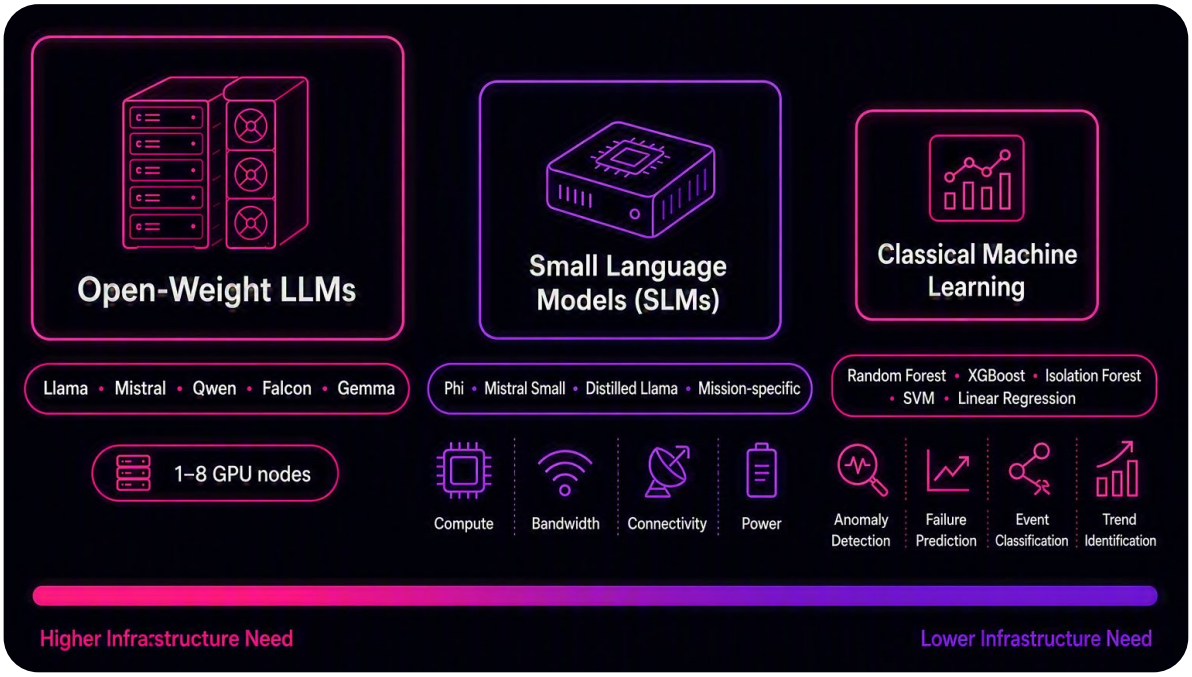


Figure 1 — Options for off-net defense environments: open-weight LLMs, Small Language Models, and classical ML.

- **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 tactical 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 Defense 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	Unclassified administrative, training, and public-facing workloads
Private Cloud	Sensitive-but-unclassified and coalition-shared enterprise applications
On-Premises	Classified, tactical, and mission command systems
Colocation	Regional garrison or coalition data centers needing resilience without full on-base 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 classified or mission-critical information. Over-rotating to fully isolated on-premises infrastructure risks higher costs, slower AI adoption, and reduced interoperability with coalition 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 defense organizations benefit from a layered approach rather than a single model — balancing AI capability, resilience, security, sovereignty, and cost, and enabling meaningful AI even in fully disconnected or mission-critical environments.

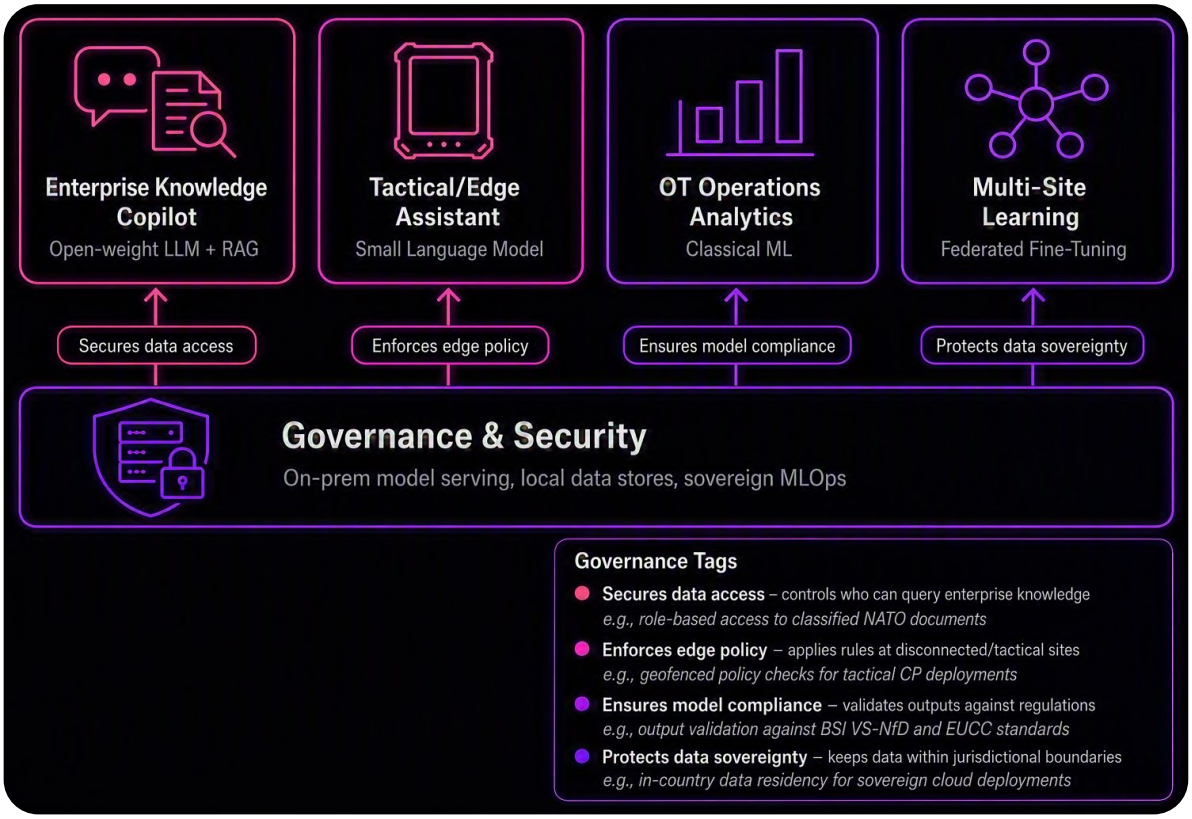


Figure 5 — Layered architecture: Enterprise Knowledge Copilot, Tactical/Edge Assistant, OT Operations Analytics, and Multi-Site Learning, governed by sovereign MLOps.

Model	Technology
Enterprise Knowledge Copilot	Open-weight LLM + Retrieval-Augmented Generation
Tactical / 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 hundreds of bases, forward sites, and coalition partners, raw data often cannot be shared. 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 infrastructure environment with hundreds of locations
- Regulations or operational policies prohibit sharing raw data

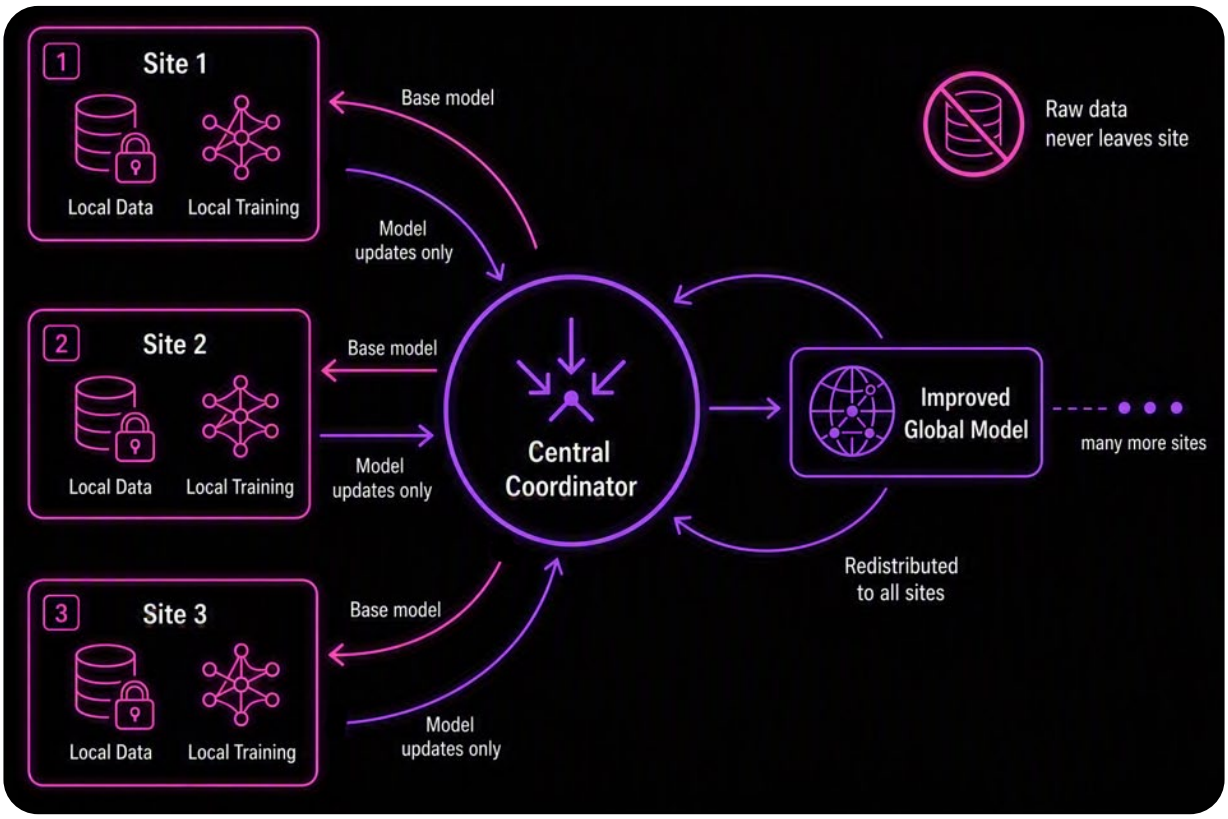


Figure 4 — Federated fine-tuning: each site keeps its local data; only model updates cross to the central coordinator.

Federated Learning to the Rescue:

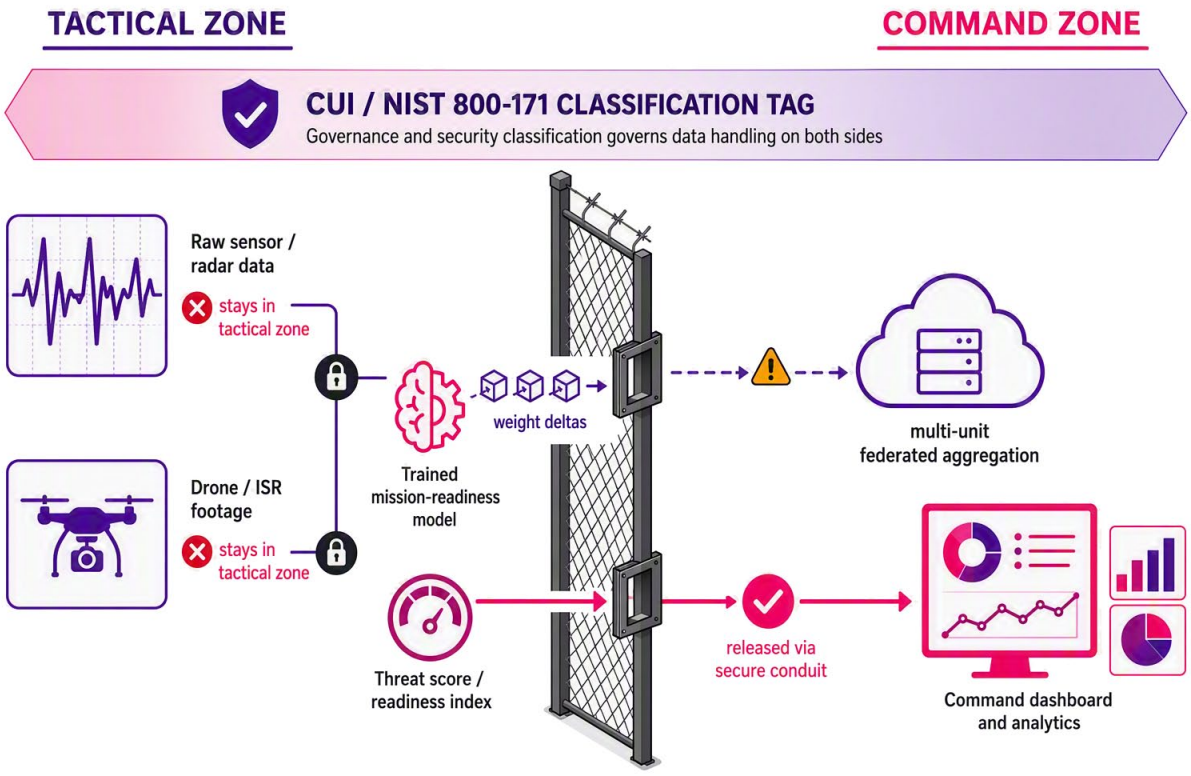
- Each site retains its local data
- A base model is distributed to all sites
- Local training occurs within the site's secure environment
- Only model updates or weight deltas are shared
- A central coordinator aggregates updates into an improved global model

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 / radar data	Forward edge node (tactical zone)	Stays in the tactical zone
Drone / ISR footage	Forward edge node	Stays in the tactical zonee
Trained mission-readiness model	Runs at the edge	Weight deltas move to multi-unit federated aggregation
Threat score / readiness index	Output of local inference	Released via secure conduit to command
Classification / zone tag (e.g., CUI, NIST 800-171)	Attached to every artifact	Governs command 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.

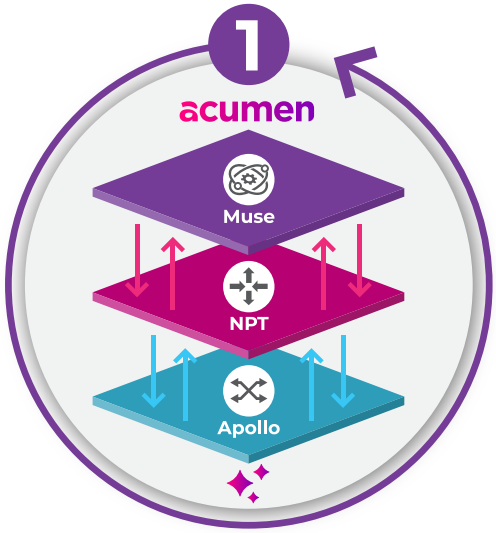


Figure 6 — Ribbon's deterministic transport fabric: Apollo optical, NPT IP/ MPLS, Muse automation, orchestrated by Acumen AIOps.

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 defense AI. Public cloud, private cloud, and on-premises deployments each have advantages, and the most resilient organizations deploy a deliberate, layered mix of all three — with data sovereignty built in from day one, not retrofitted after the fact.

With Open-Weight LLMs, SLMs at the tactical edge, classical ML for OT analytics, federated learning across distributed sites, and a deterministic transport fabric such as Ribbon's Apollo, NPT, Muse, and Acumen platforms, defense organizations can bring the full power of AI to bear — without ever compromising the sovereignty of the data that keeps warfighters and critical infrastructure safe.

- **No single deployment model wins.** Public cloud, private cloud, on-premises, and colocation each suit different workloads; most defense organizations need a deliberate mix matched to classification and mission-criticality.
- **AI no longer requires the hyperscale cloud.** Open-Weight LLMs, Small Language Models (SLMs), and classical machine learning now run entirely within sovereign, air-gapped, and disconnected environments — from garrison data centers to the tactical edge.
- **Federated learning solves the multi-site problem.** Hundreds of bases 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 accredited enclave.
- **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 Defense 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 Tactical Edge

Small Language Models — such as the Microsoft Phi family, Mistral Small, distilled Llama variants, and specialized mission-focused models — are optimized for environments where compute is constrained, bandwidth is limited, connectivity is intermittent, and power consumption matters. SLMs can run on tactical compute nodes, ruggedized servers, edge gateways, vehicle-mounted systems, portable command posts, and networks-in-a-box, enabling AI capabilities directly at the edge rather than requiring reach-back connectivity to a centralized data center.
Military and emergency response applications include:

- Mission planning assistance
- Local document search
- Intelligence report summarization
- Field maintenance guidance
- Sensor report interpretation
- Emergency response procedures

Because inference occurs locally, warfighters and first responders continue to receive AI assistance even during communications outages or contested network conditions.

Classical Machine Learning: Often the Best Choice

Despite the excitement around generative AI, many OT and critical infrastructure challenges — including the defense of tactical and garrison 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 unauthorized protocols, recognizing malicious behavior, and baselining normal operations across defense 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