AI at the point of ACTION

Architecting the Sovereign Edge: AI at the Point of Action

March 2026

DELL Technologies

The Federal Edge AI Imperative

Introduction: AI at the Point of Action

Federal missions increasingly depend on machine learning inference, computer vision, and autonomous decision-making at the point of action. Missile intercept calculations, border surveillance alerts, disaster response coordination. These workloads cannot wait for sensor data to travel to distant data centers, get processed, and return. The architecture must match the mission, and for federal agencies operating in contested or connectivity-limited environments, that means processing data where it is generated.

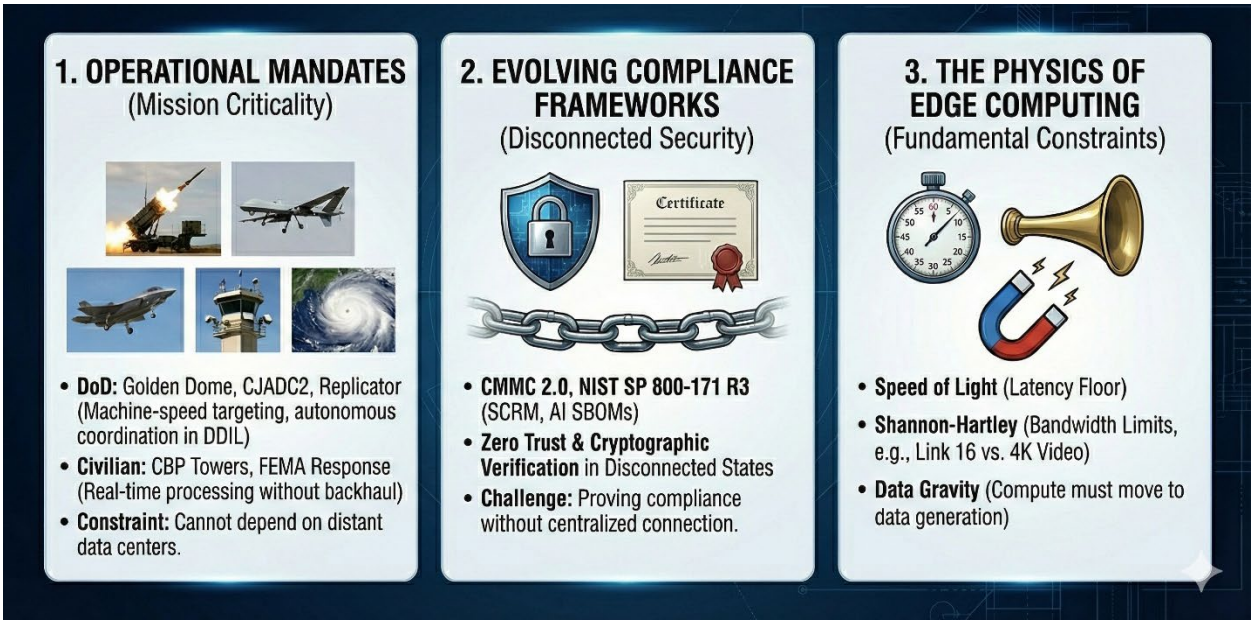
The operational constraints are unforgiving. A hypersonic threat traveling at Mach 10 covers 680 meters during a 200-millisecond round-trip to a remote data center, potentially closing the intercept window entirely. An autonomous surveillance tower monitoring remote borderlands cannot backhaul 150 Mbps of multi-camera video over a single-digit megabit satellite link shared across an entire sector. A disaster response team coordinating dozens of drone platforms after a hurricane cannot depend on centralized services when cell towers are rubble and satellite bandwidth is saturated.

These scenarios are not edge cases but the baseline operational environment for federal agencies executing time-sensitive missions in Denied, Disrupted, Intermittent, or Limited (DDIL) communications environments.

The Convergence of Strategic Imperatives

Understanding the operational environment is one thing. What makes this moment different is the convergence of active programs, evolving compliance requirements, and physics that cannot be negotiated. These represent some of the primary drivers shaping federal edge AI today, as shown in **Figure 1** below and discussed in the following sections.

Figure 1. Primary Drivers Shaping Federal Edge AI Today



Operational Mandates

Several Department of Defense programs are driving urgent requirements for edge AI infrastructure. Golden Dome missile defense[8] must process orbital sensor data and generate intercept solutions in seconds, requiring ground fusion centers with localized AI inference. Combined Joint All-Domain Command and Control (CJADC2)[2] must enable machine-speed targeting across domains where round-trip delays to distant data centers close the decision window. The Replicator Initiative[3] must coordinate

thousands of autonomous platforms via distributed mesh networking when centralized connectivity is jammed or denied. Each program faces the same architectural constraint: mission-critical AI workloads that cannot depend on connectivity to distant data centers.

Federal civilian agencies face identical technical constraints. U.S. Customs and Border Protection operates hundreds of Autonomous Surveillance Towers[9] processing multi-modal sensor data in locations where commercial cellular networks don't exist. FEMA disaster response[10] coordinates aerial reconnaissance and damage assessment when terrestrial communications infrastructure is destroyed. The technical constraints (bandwidth scarcity, latency budgets, DDIL connectivity) cut across defense and civilian missions.

Evolving Compliance Frameworks

Finalized in late 2024, the CMMC 2.0 framework establishes certification requirements for defense contractors developing AI systems that process Controlled Unclassified Information (CUI). NIST SP 800-171 Revision 3 introduced dedicated Supply Chain Risk Management controls ending the era of unvetted AI model consumption. Contractors must now produce AI Software Bills of Materials documenting training data provenance, model weight hashes, and software dependency chains. The "black box" downloading of pre-trained models from public repositories without rigorous vetting is explicitly non-compliant. The Department of Justice Civil Cyber-Fraud Initiative underscores the seriousness of these compliance obligations.

For edge AI systems, these requirements create a unique challenge. Traditional compliance frameworks assume continuous connectivity to centralized security services, audit infrastructure, and identity providers. DDIL operational environments violate these assumptions. Agencies need infrastructure that can demonstrate compliance even when disconnected, with cryptographic supply chain verification baked into hardware and Zero Trust architectures validated to operate without reaching back to centralized policy engines.

The Physics of Edge Computing

Federal edge architectures must operate within fundamental physical laws that no amount of engineering can circumvent. The speed of light establishes an absolute floor on latency to distant data centers. The Shannon-Hartley theorem defines the maximum information rate for any communications channel, which transforms bandwidth limitations from network engineering problems into fundamental physical constraints. Data gravity pulls computation to where data is generated when moving that data is expensive or impossible.

Link 16, the primary tactical data link for joint military operations, provides approximately 115 kilobits per second of effective throughput[12]. A single 4K camera stream requires 25 megabits per second, over **200 times** Link 16's capacity. A six-camera ISR pod would demand bandwidth equivalent to **over 1,300 Link 16 channels simultaneously**. Alternative communications options do not solve this mismatch. Commercial LEO satellite constellations delivering 50-200 Mbps under ideal conditions still face a three orders of magnitude deficit against autonomous vehicle sensor suites generating 50+ gigabits per second from LiDAR, cameras, and radar arrays.

These physical constraints dictate where AI processing must occur, making edge-local computation a requirement rather than an optimization.

Hybrid Cloud-Edge Architectures and Federal Mission Constraints

Technology vendors often propose "cloud-edge hybrid" solutions where lightweight edge preprocessing feeds cloud-based AI services. This architecture serves commercial IoT applications with relaxed latency requirements and reliable connectivity. Federal mission AI operates under much different constraints.

Hybrid architectures typically assume the cloud is reachable with consistent latency. Federal tactical operations often experience DDIL connectivity where these assumptions don't hold. Hybrid models typically assume continuous synchronization between edge and cloud. Electromagnetic jamming, satellite handoff gaps, and expeditionary deployments can disrupt this synchronization. Hybrid approaches may

assume sensor data can transit outside government-controlled infrastructure. Data sovereignty requirements for classified intelligence, law enforcement sensitive information, and disaster survivor PII (Personally Identifiable Information) often prohibit this data movement.

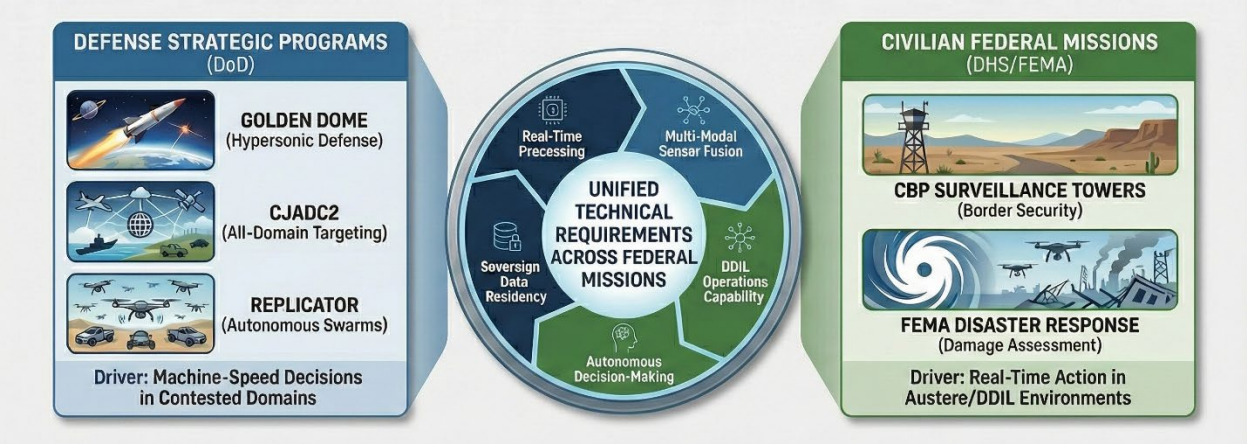
A true federal edge architecture inverts the cloud-edge relationship, implementing **edge-sovereign AI with selective cloud augmentation**. Critical inference happens at the edge using locally deployed models. Cloud resources provide strategic analytics on aggregated mission data, model retraining on classified datasets within air-gapped facilities, and centralized coordination. But this can only happen when connectivity permits and security policy allows.

The edge is not an extension of the cloud. The edge is the primary compute infrastructure for mission autonomy.

The Federal Strategic Imperative: Machine-Speed Autonomy at Scale

Active federal programs are driving urgent requirements for edge AI infrastructure, some examples are shown in **Figure 2** below. Department of Defense initiatives all demand AI processing at forward positions where centralized infrastructure cannot satisfy latency and bandwidth constraints. Federal civilian agencies face parallel requirements across border security and disaster response.

Figure 2. Unified Requirements Across Federal Missions



Taking a deeper look into these programs can reveal the technical patterns shaping federal edge AI requirements today.

Golden Dome: The $175 Billion Edge AI Forcing Function

On January 27, 2025, Executive Order 14186 formalized the United States' commitment to space-based missile defense, allocating $175 billion over three years for what the Department of Defense calls "Golden Dome"[15]. The mission is to detect and intercept hypersonic threats traveling at Mach 5+ speeds, faster than existing ground-based radar can track. Orbital sensors identify threats, ground fusion centers synthesize multi-sensor data, and interceptor batteries receive targeting solutions within seconds. There is no time to backhaul terabytes of sensor data to a strategic data center for AI processing. The compute must be where the sensors are, at the orbital edge, at ground stations, at tactical firing units.

Golden Dome surfaces the architectural challenges federal edge AI must address: distributed processing across multiple layers, coordination between those layers under bandwidth constraints, continued operation when connectivity is degraded or denied, and data sovereignty across classified environments.

Defense Strategic Programs: CJADC2 and Replicator

This architectural pattern extends across DoD as the department transforms toward machine-speed decision-making in all domains. Two additional programs validate the scale of federal edge AI requirements:

Combined Joint All-Domain Command and Control (CJADC2)

The DoD's CJADC2 strategy integrates sensors, command centers, and weapon systems across air, land, sea, space, and cyber domains. A Government Accountability Office oversight report in April 2025 (GAO-25-106454) identified distributed edge processing as essential for achieving "machine-speed targeting"[16]. Tactical units cannot depend on intermittent satellite backhaul to centralized clouds when prosecuting time-sensitive targets. Local AI inference for object detection, threat classification, and targeting must execute at forward operating bases, aboard maritime platforms, and within tactical vehicles.

The technical demands mirror Golden Dome. Massive multi-modal sensor data, millisecond decision timelines, and operations in DDIL communications environments.

Replicator Initiative: Thousands of Autonomous Systems

The Replicator Initiative aims to field thousands of autonomous systems coordinating across contested domains[17]. Each autonomous platform acts as an edge compute node, running consensus algorithms for swarm coordination and computer vision for navigation. When thousands of platforms operate in contested electromagnetic spectrum, centralized command and control cannot scale. The platforms must coordinate among themselves, making local decisions while maintaining collective mission coherence. Attritable by design, these platforms assume losses. Mission intelligence must be distributed across the swarm so the collective continues operating as individual nodes drop out. Each platform carries the onboard processing to act independently when coordination links fail.

Civilian Federal Missions: The Parallel Imperative

Federal civilian agencies face the same edge AI constraints. Border security, law enforcement, and disaster response all demand real-time processing in DDIL environments.

Customs and Border Protection: 200+ Autonomous Surveillance Towers

U.S. Customs and Border Protection (CBP) has deployed over 200 Autonomous Surveillance Towers across the southwest border, operating in remote borderlands where commercial cellular networks do not exist[18]. These systems process multi-modal sensor feeds including video, thermal, and radar in real time to detect and track threats. Satellite backhaul capacity is measured in megabits per second. Local AI processing is not an optimization. It is the only viable architecture.

CBP's $2.7 billion investment in AI-enabled border security validates edge AI as mission-critical for civilian agencies operating in austere environments.

Federal Emergency Management Agency: Disaster Response in DDIL Environments

FEMA's use of UAS for disaster damage assessment represents edge AI requirements in their most extreme form. Destroyed communications infrastructure, no electrical grid, and mobile command centers as the only computational resource. When Hurricane Helene devastated western North Carolina in September 2024, federal responders deployed UAS equipped with onboard AI to map structural damage, locate survivors, and prioritize rescue operations, all without connectivity to external cloud resources.

DDIL is not only a DoD-specific challenge. It is the baseline assumption for federal disaster response.

Unified Technical Requirements Across Federal Missions

Despite diverse mission profiles spanning hypersonic defense, cross-domain targeting, border surveillance, law enforcement, and disaster response, five technical requirements emerge consistently.

1. **Real-Time AI Processing.** Decision timelines measured in milliseconds (Golden Dome intercept) to minutes (disaster damage assessment), not hours or days.
2. **Multi-Modal Sensor Fusion.** Integration of video, thermal, radar, acoustic, and RF sensor feeds for autonomous perception and threat classification.
3. **DDIL Operations Capability.** Systems must function with Denied, Disrupted, Intermittent, or Limited communications. Zero connectivity cannot be a failure mode.
4. **Autonomous Decision-Making.** Human-supervised autonomy for perception, planning, and execution, augmenting human operators with machine-speed analysis.
5. **Sovereign Data Residency.** Classified defense sensor data, law enforcement sensitive information, and disaster survivor PII cannot leave government-controlled infrastructure.

These requirements are physically incompatible with traditional cloud-dependent AI architectures. The next section demonstrates why.

Data Gravity: Why Distant Infrastructure Cannot Serve Tactical AI

Physics and information theory establish hard limits on where AI processing can execute for time-sensitive federal missions. Whether the distant infrastructure is commercial cloud or government core data centers, the constraints are identical: the speed of light determines minimum latency, channel capacity limits determine maximum data throughput, and contested environments deny connectivity entirely. These are not engineering problems awaiting better solutions. They are fundamental constraints that dictate architectural decisions. Understanding these limits clarifies why federal tactical AI must process data where it is generated.

The Physics of Latency and Bandwidth

Enterprise AI architectures range from fully cloud-hosted to entirely on-premises, with organizations selecting based on latency tolerance, data sovereignty, and operational requirements. Federal mission AI operates at the extreme end of this spectrum, where physical constraints make edge-local processing mandatory. Three constraints define this boundary: the speed of light, channel capacity limits, and the operational reality of contested communications.

The Speed of Light Ceiling

The first constraint is time. Physics sets a hard floor on round-trip latency to any distant data center. A ground fusion center communicating with a distant data center 2,000 kilometers away faces theoretical minimum latency of 13 milliseconds, assuming perfect conditions with no network congestion or processing delays.

Reality is harsher. Typical cloud API round-trips for forward-deployed units using satellite backhaul range from 200-400 milliseconds. A hypersonic threat traveling at Mach 10 covers 3.4 kilometers per second. During a 200-millisecond round-trip, the threat moves 680 meters, potentially closing the intercept window entirely.

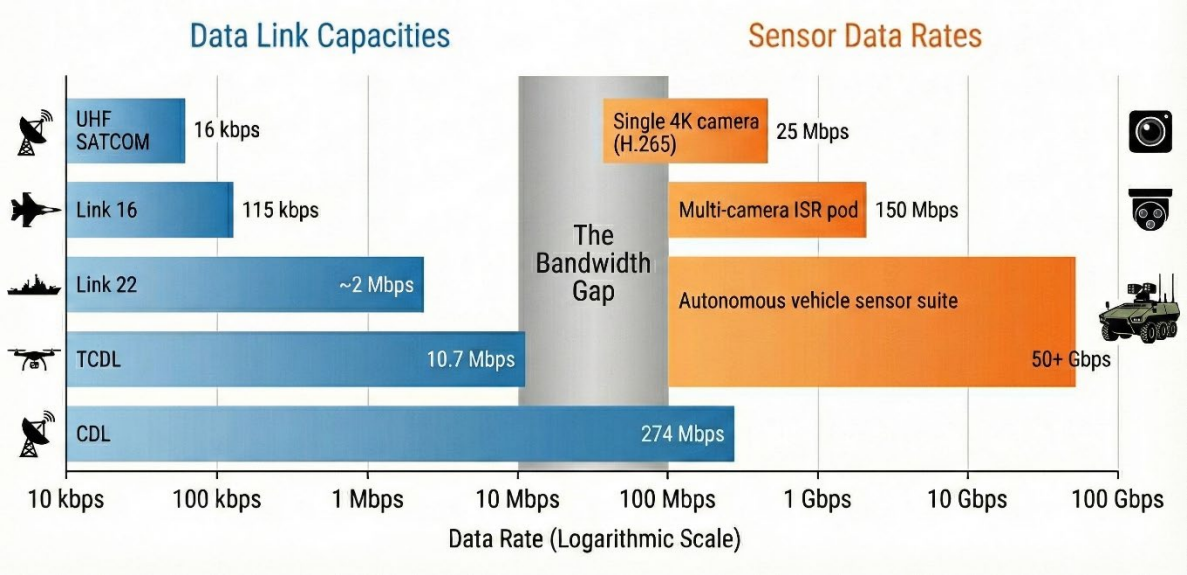
Machine-speed decision-making cannot tolerate these delays. CJADC2 targeting loops require perception-to-action cycles measured in seconds[2]. Autonomous swarm coordination demands sub-100-millisecond consensus updates. UAS collision avoidance needs 10-30 millisecond reaction times. Centralized infrastructure, whether cloud or core data center, is too far away.

The Bandwidth Wall

The second constraint is capacity. Even if latency were acceptable, data volume creates an independent limit. The Shannon-Hartley theorem establishes the maximum data rate any communications channel can carry based on its bandwidth and noise characteristics. No equipment upgrade can change this ceiling. Link 16 illustrates what this means for tactical operations[12]. As the primary data link for joint military

operations, it provides approximately 115 kilobits per second of effective throughput. A single 4K camera stream requires 25 Mbps, over **200 times** Link 16's entire capacity. A six-camera ISR pod would demand **over 1,300 Link 16 channels simultaneously**. The below **Figure 3** visualizes the gap that needs to be addressed between the bandwidth of data links and the amount of information that current sensors can generate.

Figure 3. Tactical Bandwidth vs Sensor Output



These constraints fundamentally shape architectural decisions. When sensor output exceeds backhaul capacity by orders of magnitude, the only viable response is processing data at its source.

Data Gravity in Action

These first two constraints combine to create what researchers call "data gravity," the principle that large datasets pull computation toward them. When moving data is expensive or impossible, computation must move to where data is generated.

Federal edge missions exemplify extreme data gravity. A ground fusion center processing multi-sensor inputs from orbital, terrestrial, and maritime sources cannot backhaul 500+ Gbps of aggregate sensor data over shared fiber links. The solution is local GPU clusters for sensor fusion, with distant data centers used only for strategic analytics of post-engagement data.

As a real world example of this principle, the U.S. CBP's Autonomous Surveillance Tower network generates gigabits of total sensor output across hundreds of towers[9]. Available satellite backhaul provides megabits shared across entire sectors. The solution is edge AI at each tower for threat detection, transmitting only alerts and track data rather than raw video streams.

Local AI inference transforms what travels over constrained links. Rather than transmitting raw sensor data, edge systems transmit inference results such as detected objects, threat classifications, and track files. A multi-gigabit thermal camera stream becomes a kilobit metadata feed of object bounding boxes, confidence scores, and track identifiers. This is not data compression. It is mission-relevant information extraction, executed at the edge.

DDIL: The Forcing Function

The third constraint is availability. DDIL communications represent the baseline assumption for federal tactical operations, not an edge case[13,14]. DoD planning assumes peer adversaries will target communications infrastructure. Disaster response depends on operations where infrastructure could be inoperable or destroyed. Border security consistantly patrol areas where commercial cellular coverage is limited or does not exist.

AI architectures dependent on distant data centers fail catastrophically in DDIL environments. A dropped satellite link halts computer vision when processing depends on distant infrastructure. RF jamming blinds threat detection when models reside in centralized data centers rather than near the sensor. The mission fails not because the AI failed, but because the architecture assumed connectivity that contested environments cannot guarantee.

Edge-sovereign AI architectures degrade gracefully. When connectivity is available, edge systems synchronize data, update models, and coordinate with command centers. When connectivity fails, edge systems continue autonomous operation using locally cached models. Mission capability persists.

This is not a technical preference. It is a strategic requirement formalized in DoD policy and increasingly mandated in federal civilian operations.

Edge AI Architecture for Federal Missions

A Mission-Driven Edge Framework

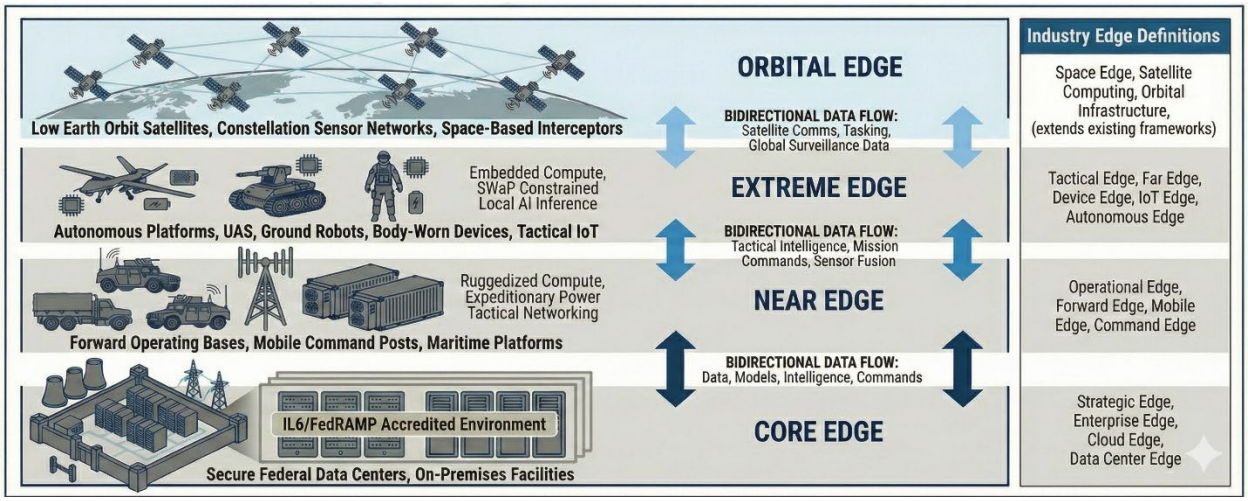
Federal edge deployments span radically different environments: climate-controlled data centers, expeditionary forward operating bases, autonomous vehicles, and orbital spacecraft. What unifies them is not physical proximity to sensors but the operational requirement for local AI processing. A four-layer architecture captures this reality, defining each layer by its mission autonomy requirements rather than geographic distance. This section presents that framework and the infrastructure requirements that span all layers.

Beyond "Near" and "Far": A Mission-Driven Framework

The Atlantic Council has articulated edge continuum architectures for defense applications, defining Tactical, Operational, Command, and Strategic layers based on the sense-make sense-act decision cycle[1]. A ground fusion center for Golden Dome is 200 kilometers from orbital sensors but must process terabytes of tracking data in real time. A CBP tactical coordination center is 50 meters from an Autonomous Surveillance Tower but cannot tolerate round-trip latency to distant infrastructure. A FEMA disaster response center operates in destroyed infrastructure where satellite connectivity measures 2 Mbps and drops every 15 minutes. Distance from the sensor is not the defining characteristic. Mission autonomy requirements are.

Industry frameworks use varied terminology for edge tiers—tactical edge, far edge, operational edge, enterprise edge, and others. This paper consolidates these into four operational layers: **Core Edge**, **Near Edge**, **Extreme Edge**, and **Orbital Edge**. The addition of Orbital Edge reflects the growing role of space-based processing in federal missions. **Figure 4** below maps these layers to common industry definitions.

Figure 4. Multi-Layer Edge Computing Architecture



Core Edge: The Sovereign Foundation

Physical Deployment: Federal data centers, on-premises facilities, IL6 air-gapped environments

Operational Characteristics: Reliable power and cooling infrastructure; enterprise networking with 10+ Gbps throughput; physical security and access control; and continuous staffing and maintenance

Mission Role: The Core Edge serves as the sovereign compute foundation for AI model training, data fusion across multiple missions, and strategic analytics. This is where federal agencies train computer

vision models on classified sensor data, where intelligence fusion centers synthesize multi-INT analysis, and where compliance-critical workloads require FedRAMP High or DoD IL6 accreditation.

Golden Dome's ground fusion centers represent Core Edge infrastructure. These facilities process multi-sensor tracking data from orbital, terrestrial, and maritime sources, running AI inference at scale while maintaining complete data sovereignty. No sensor data leaves government-controlled infrastructure. The CMMC 2.0 framework establishes requirements for defense contractors handling controlled unclassified information that often necessitate this architectural approach[19].

The Core Edge provides sovereign compute capabilities addressing federal data residency and security requirements.

Near Edge: Forward Deployed Infrastructure

Physical Deployment: Forward operating bases, tactical operations centers, maritime platforms, mobile command posts

Operational Characteristics: Intermittent or limited connectivity (satellite, tactical radio, LTE when available); expeditionary power (generators, tactical power systems); ruggedized environmental protection (MIL-STD-810H, IP65+); and reduced cooling capacity compared to Core Edge

Mission Role: The Near Edge brings AI inference and tactical decision-making to forward-deployed units operating in contested or austere environments. CJADC2 requires machine-speed targeting at forward operating bases where backhaul to strategic data centers is measured in seconds—too slow for time-sensitive targets. Border Patrol tactical coordination centers in remote sectors need real-time threat detection from 54+ Autonomous Surveillance Towers despite satellite backhaul measured in megabits per second, not gigabits[18].

The Near Edge maintains mission capability during DDIL communications. When connectivity degrades or fails, local AI processing continues. Autonomous systems synchronize when links are available, operate independently when links are denied.

This layer demands ruggedized infrastructure that survives expeditionary conditions while delivering enterprise-class compute performance. Dell's XR8000 and XR7620 systems provide modular GPU acceleration in compact, environmentally hardened form factors designed specifically for this operational context.

Extreme Edge: Autonomous Platforms and Sensor Networks

Physical Deployment: Unmanned aircraft systems, autonomous ground vehicles, body-worn sensors, tactical IoT networks

Operational Characteristics: Size, Weight, and Power (SWaP) constraints (watts, not kilowatts); battery or vehicle power, zero external cooling; fully autonomous operation in GPS-denied, communications-denied environments; and tactical mobility and rapid deployment

Mission Role: The Extreme Edge represents compute at the platform itself. UAS run real-time computer vision for autonomous navigation, tactical robots execute swarm coordination algorithms, and body-worn sensors perform threat classification at the point of detection. These platforms cannot depend on backhaul to the Near Edge or Core Edge. The perception-decision-action loop must close locally, in milliseconds.

Replicator autonomous systems executing swarm tactics cannot pause for cloud API calls. Each platform is an independent edge compute node collaborating through distributed consensus when communications allow, operating autonomously when communications fail.

SWaP constraints dominate this layer. AI inference must run on embedded GPUs consuming 10-50 watts, not data center accelerators consuming 700 watts. Model optimization, quantization, and purpose-built inference hardware become mission-critical enablers. Dell's XR5610 compact edge server delivers GPU acceleration in a 1U short-depth form factor optimized for vehicle integration and mobile deployment.

Orbital Edge: Space-Based Sensing and Processing

Physical Deployment: Low Earth Orbit satellites, constellation-based sensor networks, space-based interceptors

Operational Characteristics: Extreme SWaP constraints (every kilogram costs $10,000+ to launch); solar power with intermittent availability; radiation-hardened or radiation-tolerant compute; and no physical access for maintenance or repair

Mission Role: Golden Dome's orbital sensors represent the newest and most extreme edge: AI processing in space. Hypersonic threats traveling at Mach 5+ require detection, tracking, and targeting solutions generated in seconds. Orbital sensors identify threats, onboard AI performs initial classification and trajectory prediction, and refined targeting data transmits to ground fusion centers for final intercept decisions.

Latency to ground stations is measured in milliseconds when satellites have line-of-sight, but orbital mechanics create gaps in coverage. Space-based processing enables continuous tracking through handoffs between satellites in a constellation. The compute must survive radiation exposure, operate on solar power with battery backup during eclipse, and function autonomously for years without physical maintenance.

This layer is emerging rapidly. Commercial remote sensing constellations already perform onboard AI for ship detection, wildfire monitoring, and change detection. The DoD is extending this architecture to missile defense, space domain awareness, and multi-INT fusion.

Unified Infrastructure Requirements Across Layers

Despite radically different operational environments—climate-controlled data centers, expeditionary FOBs, autonomous vehicles, orbital spacecraft—five technical requirements span all four edge layers:

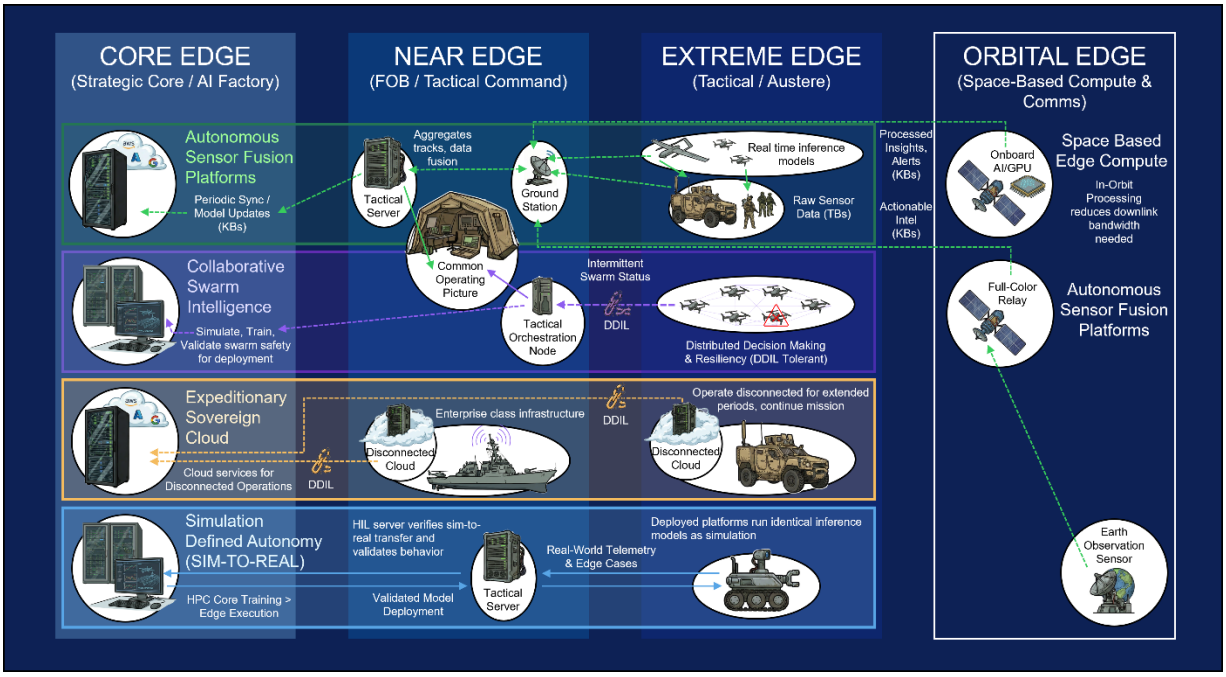
1. **Sovereign Data Residency:** Classified defense data, law enforcement sensitive information, and disaster survivor PII cannot leave government-controlled infrastructure. Every layer must support air-gapped or FedRAMP-accredited operation.
2. **DDIL Operations Capability:** Communications are never guaranteed. Each layer must function with Denied, Disrupted, Intermittent, or Limited connectivity. Zero connectivity cannot be a failure mode.
3. **Real-Time AI Inference:** Decision timelines range from milliseconds (Extreme Edge threat detection) to minutes (Core Edge strategic fusion). All layers require GPU acceleration and optimized inference software.
4. **Multi-Modal Sensor Fusion:** Integration of video, thermal, radar, RF, and acoustic sensors for autonomous perception. Edge infrastructure must ingest, synchronize, and fuse heterogeneous sensor streams in real time.
5. **Cybersecurity and Supply Chain Assurance:** CMMC 2.0 compliance, hardware root of trust, firmware validation, and supply chain verification apply across all layers. Compromised edge infrastructure creates mission failure[20].

Infrastructure spanning all four layers requires purpose-built hardware, orchestration software for distributed deployment, and end-to-end security validation including supply chain assurance. No single layer operates in isolation. The continuum works as a system, with each layer performing the AI processing it can execute locally, synchronizing with adjacent layers when connectivity allows, operating autonomously when connectivity fails.

Capability Archetypes: Mission Patterns at the Tactical Edge

The four-layer edge continuum provides the architectural foundation. What remains is understanding how federal missions actually use that architecture. Federal edge AI deployments share common architectural patterns despite diverse mission contexts. A maritime autonomous vessel surveilling contested waters, a CBP surveillance tower monitoring remote borderlands, and a tactical UAS supporting law enforcement operations implement variations of the same fundamental pattern: multi-modal sensors generating data faster than available bandwidth, onboard AI extracting mission-relevant information, and selective transmission of inference results to tactical command centers. Understanding these repeatable patterns (capability archetypes) enables agencies to design procurement strategies, systems integrators to accelerate deployment timelines, and technology vendors to align product portfolios with validated mission requirements.

Figure 5. Edge AI Capability Architypes



Four capability archetypes capture the operational patterns federal agencies deploy across DoD and civilian missions today. **Figure 5** above illustrates these archetypes and where they fit within the four operational layers. The figure also includes an emerging Orbital Edge example—space-based AI processing that reduces downlink bandwidth by transmitting insights rather than raw imagery. While orbital edge deployments are nascent, the architectural pattern mirrors terrestrial sensor fusion: compute at the source, transmit intelligence. The sections below go into detail on the four established archetypes:

Autonomous Sensor Fusion Platforms

Architectural Pattern

Mobile platform integrates multiple sensor modalities (electro-optical, infrared, radar, acoustic, RF), onboard GPU executes real-time AI inference for object detection and threat classification, and constrained satellite or radio uplink transmits inference results (not raw sensor data) to tactical operations centers.

Operational Context

Maritime autonomous vessels patrolling exclusive economic zones detect and track surface vessels, submarines, and aircraft using sensor fusion across radar, electro-optical cameras, and acoustic arrays. Continuous sensor output generates terabytes daily. Satellite uplink capacity measures 10-50 Mbps. Onboard AI reduces terabytes of raw sensor data to kilobytes of track files: object classifications, positions, velocities, confidence scores. Human operators at shore-based command centers receive actionable intelligence, not unprocessed video.

CBP's Autonomous Surveillance Towers implement this same pattern terrestrially, with hundreds deployed across the southwest border. Each tower integrates day/night cameras, thermal imagers, and ground-scanning radar. Remote borderland locations lack commercial cellular coverage. Satellite backhaul shared across multiple towers provides single-digit Mbps per site during peak utilization. Local AI processing detects human activity, vehicle movement, and threat patterns. Only alerts and track data transmit to tactical coordination centers[18].

Multi-Layer Integration

- **Extreme Edge:** Platform-mounted GPU (NVIDIA Jetson AGX, IGX) executes real-time inference models for sensor fusion and object detection
- **Near Edge:** Tactical operations center (Dell XR7620) aggregates tracks from multiple platforms, performs multi-platform fusion, generates common operating picture
- **Core Edge:** AI Factory (Dell PowerEdge XE9680) trains computer vision models on aggregated mission data, validates new models in simulation, distributes updated models to deployed platforms

Technical Requirements

- **Size, Weight, Power (SWaP) Constraints:** Platform-mounted compute limited to 10-50 watts thermal budget
- **Real-Time Processing:** 10-30 FPS inference for computer vision, sub-100ms latency for autonomous navigation
- **Model Optimization:** Quantization (FP16, INT8) and pruning to fit models in embedded GPU memory
- **DDIL Operations:** Platforms operate autonomously during communications loss, synchronize when connectivity restores

This archetype scales from individual platforms to distributed sensor networks. The fundamental pattern remains constant: compute where the sensors are, transmit intelligence not data.

Collaborative Swarm Intelligence

Architectural Pattern

Multiple autonomous platforms coordinate via distributed mesh networking, execute consensus algorithms for formation control and task allocation, share perception and intent through low-bandwidth inter-platform links, and maintain mission capability despite individual platform loss or communications disruption.

Operational Context

The DoD Replicator Initiative envisions thousands of autonomous systems coordinating across contested domains[3]: aerial swarms for ISR and jamming, surface swarms for sea lane denial, ground swarms for urban reconnaissance. Traditional command-and-control architectures (centralized mission planning with platforms as remote sensors) cannot scale to thousands of platforms operating in electromagnetically contested environments. Swarm intelligence distributes decision-making across the collective.

Each platform in an autonomous swarm acts as both sensor and decision-maker. Aerial reconnaissance swarms partition search areas dynamically based on observed threat patterns. If one platform detects a target, the swarm reallocates assets to surround and track it without human intervention. If jamming severs communications with command centers, the swarm continues mission execution using locally cached objectives and distributed consensus for coordination.

Civilian applications mirror these DoD patterns. FEMA disaster response deploying dozens of UAS for damage assessment across hurricane-impacted regions requires autonomous coordination when communications infrastructure is destroyed[11]. Individual platforms assess structural damage using onboard AI, share findings via mesh networking, and collaboratively construct comprehensive damage maps without continuous connectivity to command posts.

Multi-Layer Integration

- **Extreme Edge:** Each platform runs consensus algorithms (Raft, Byzantine Fault Tolerant) for swarm coordination, executes local mission planning
- **Near Edge:** Tactical orchestration node (Dell XR8000) monitors swarm health, updates mission parameters when connectivity permits, aggregates swarm intelligence for human operators
- **Core Edge:** AI Factory simulates swarm behaviors in digital environments, trains reinforcement learning policies for emergent coordination, validates swarm safety before field deployment

Technical Requirements

- **Distributed Consensus:** Platforms achieve agreement on formation, task allocation, and threat assessment without centralized coordination
- **Low-Bandwidth Mesh Networking:** Inter-platform communication operates on tactical radio (Link 16[12]) or line-of-sight RF mesh
- **Fault Tolerance:** Swarm maintains capability despite 30-50% platform loss
- **Emergent Behavior Validation:** Simulation-based testing ensures swarm behaviors align with mission objectives and safety constraints

Swarm intelligence represents the most challenging edge AI pattern: autonomous decision-making must be provably safe, verifiably aligned with mission intent, and resilient to adversarial disruption. The Core Edge simulation environment becomes critical, and swarm behaviors must be exhaustively tested in digital twins before field deployment.

Expeditionary Sovereign Cloud

Architectural Pattern

Deploy self-contained compute infrastructure to forward operating bases, mobile command posts, or maritime platforms, providing tactical units with sovereign AI services (model inference, data fusion, tactical intelligence) despite intermittent or denied backhaul to strategic data centers.

Operational Context

Forward operating bases in denied or austere environments require locally hosted AI services that would normally reside in strategic cloud data centers. Tactical intelligence analysts need AI-assisted intelligence fusion. Mission planners require autonomous route planning and threat prediction. Communications planners need spectrum analysis and jamming avoidance. These AI workloads cannot tolerate 200-400 millisecond satellite round-trip delays or assume continuous connectivity.

Dell pioneered this architectural pattern with the Dell Tactical Microsoft Azure Stack, developed in partnership with Tracewell Systems to bring Azure-consistent cloud capabilities to operating environments where network connectivity is unavailable or unreliable[26]. The ruggedized, field-deployable platform enables fully disconnected operation without requiring connectivity to public cloud infrastructure. Designed for remote and rugged circumstances where mobility and high portability are required, the solution addresses scenarios from forward operating bases to maritime platforms where mission success depends on local compute autonomy.

The U.S. Navy's Flank Speed program demonstrates this pattern at scale. In February 2025, Navy PEO Digital reported successful deployment of Flank Speed to the tactical edge using hyper-converged infrastructure through Azure hubs as a "deployable package" both ashore and afloat[27]. The infrastructure provides local caching and resiliency for high-intensity operations, paired with Flank Speed Wireless for sailor connectivity at the edge. Flank Speed achieved 151 of 152 Zero Trust requirements, validating that expeditionary infrastructure can meet enterprise security standards. PEO Digital is now extending this

architecture into IL6 classified environments, demonstrating a pathway from unclassified through controlled unclassified information (CUI) to classified workloads on common infrastructure.

FEMA mobile command centers coordinating disaster response implement identical patterns. Hurricane relief operations require AI-driven damage assessment, logistics optimization, and resource allocation. When cellular towers are destroyed and satellite bandwidth saturated by commercial traffic, FEMA cannot depend on cloud services. The mobile command center becomes a self-contained AI platform running inference workloads locally.

Maritime afloat edge represents the most extreme case: aircraft carriers, amphibious assault ships, and surface combatants operate as floating data centers thousands of kilometers from terrestrial infrastructure. Carrier air wings generate terabytes of mission data daily from dozens of aircraft. Shipboard tactical operations centers require AI-assisted threat fusion, mission planning, and logistics optimization. Satellite bandwidth cannot support architectures dependent on distant infrastructure. The ship itself must provide enterprise-class AI infrastructure.

Multi-Layer Integration

- **Near Edge:** Expeditionary data center (Dell XR8000 cluster) provides local AI inference, tactical data storage, multi-tenant orchestration
- **Core Edge:** During connectivity windows, expeditionary infrastructure synchronizes with strategic AI Factory for model updates, aggregates mission data for strategic analysis, and receives intelligence updates
- **Disconnected Operations:** Expeditionary cloud operates autonomously for days to weeks using cached models, continues mission support despite zero backhaul

Technical Requirements

- **Ruggedized Infrastructure:** Equipment survives expeditionary transport, operates in austere environmental conditions (heat, dust, shock, vibration)
- **Zero-Touch Deployment:** Infrastructure provisions without specialized IT personnel. Tactical units deploy, power on, and mission workloads execute automatically
- **Multi-Tenant Security:** Brigade-level infrastructure supports multiple units with zero-trust isolation satisfying NIST SP 800-171 access control requirements[21]
- **Store-and-Forward Synchronization:** AI models, mission data, and system updates queue during disconnection, synchronize automatically when connectivity restores

The expeditionary sovereign cloud inverts traditional cloud economics. Rather than paying per-use for remote services, agencies invest capital in deployable infrastructure providing unlimited local compute. For sustained operations in DDIL environments, capital investment outperforms operational expenditure.

Simulation-Defined Autonomy (Sim-to-Real)

Architectural Pattern

Train AI models and validate autonomous behaviors in high-fidelity simulation environments at Core Edge, transfer validated models to edge platforms through controlled deployment pipeline, maintain bit-for-bit compatibility between simulated and physical systems enabling continuous validation.

Operational Context

Autonomous systems operating in safety-critical missions (targeting, navigation, swarm coordination) cannot be validated through trial-and-error field testing. An autonomous targeting system that misidentifies a school bus as a military vehicle during testing creates catastrophic outcomes. Traditional software testing (unit tests, integration tests) cannot verify emergent AI behaviors or adversarial robustness. Simulation provides the validation environment.

The DoD employs simulation-based validation for all autonomous weapons systems. Computer vision models for target recognition train on millions of synthetic images generated in simulation, covering conditions impossible to replicate in physical testing: extreme weather, adversarial camouflage, novel threat vehicles. Autonomous navigation algorithms for GPS-denied environments validate in simulated

urban terrain with bit-for-bit identical sensor models. Only after exhaustive simulation testing do models deploy to physical platforms.

Commercial providers offer automotive-grade simulation for autonomous vehicle validation, the same methodology federal agencies require for tactical autonomy. Train perception models on billions of simulated miles. Test edge cases: pedestrians emerging from occlusion, sensor failures during critical maneuvers, adversarial attacks on camera inputs. Achieve validation coverage impossible through physical road testing.

Multi-Layer Integration

- **Core Edge:** AI Factory runs HPC simulation clusters (Dell PowerEdge, NVIDIA Omniverse), generates synthetic training data at petabyte scale, validates autonomous behaviors across millions of scenarios
- **Near Edge:** Tactical validation testbeds (hardware-in-the-loop rigs with production sensors and compute) verify sim-to-real transfer, ensure simulated behaviors match physical performance
- **Extreme Edge:** Deployed platforms run identical inference models as simulation, telemetry feeds back to Core Edge for continuous validation and model improvement

Technical Requirements

- **High-Fidelity Physics:** Simulation replicates sensor physics (camera optics, radar propagation, LiDAR returns) with sufficient accuracy to ensure sim-to-real transfer
- **Scalable Synthetic Data Generation:** Render billions of training images covering operational diversity (weather, terrain, threat vehicles)
- **Hardware-in-the-Loop Validation:** Production sensors and edge compute integrated into simulation loop verify bit-for-bit compatibility
- **Continuous Integration Pipeline:** Model updates flow from Core simulation → Near validation → Edge deployment with automated testing gates

Simulation-defined autonomy represents the only scalable path to autonomous system validation. Physical testing cannot achieve the coverage that simulation can. The Core Edge becomes the digital proving ground where autonomous systems earn trust before field deployment.

Dell Technologies Edge AI Solutions

Dell's Edge AI Architecture: Sovereign Infrastructure from Core to Extreme Edge

The mission patterns above share a common requirement: infrastructure that operates across the full edge continuum while maintaining data sovereignty, cybersecurity, and supply chain assurance. Building this infrastructure demands enterprise-grade compute that moves with your data, from sovereign training facilities to forward operating bases to autonomous platforms at the tactical edge. Dell Technologies delivers this capability through a unified architecture spanning core AI Factory to extreme edge integration, validated through decades of federal partnerships.

Figure 6. Dell's AI Offerings for the Federal Edge

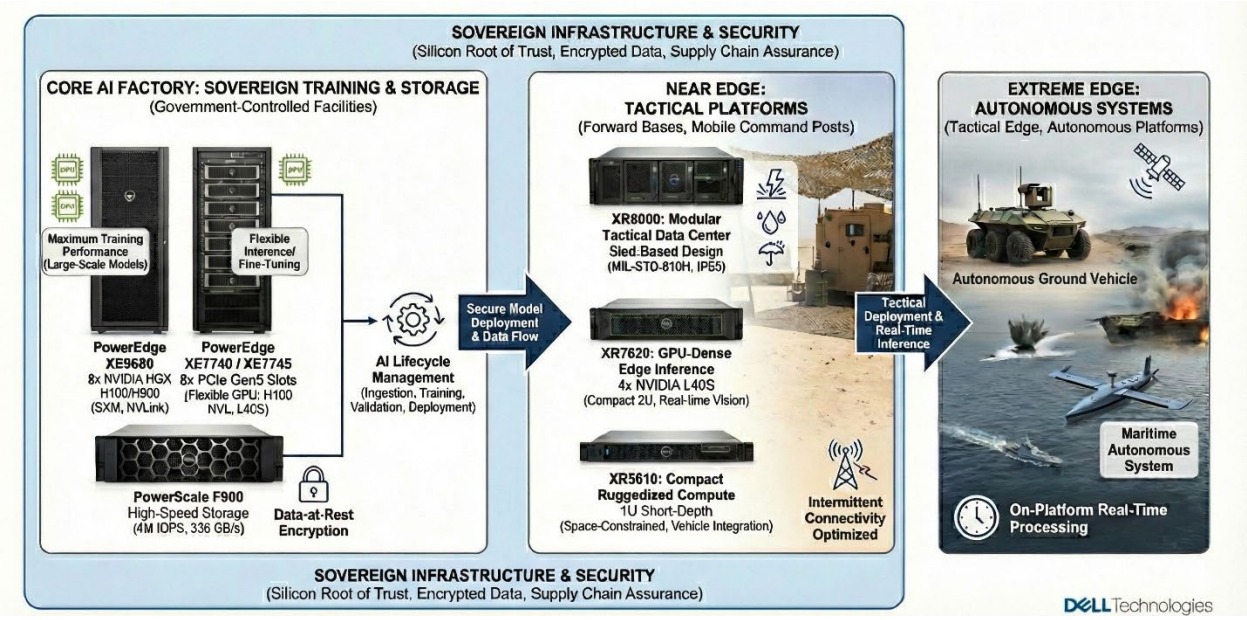


Figure 6 above gives an overview of some Dell's infrastructure offerings for the Federal edge, the sections below will go into more detail around their specific capabilities.

Core AI Factory: Sovereign Training and Storage

The Core Edge establishes the foundation for federal AI sovereignty: infrastructure where classified sensor data, law enforcement sensitive information, and disaster response records never leave government-controlled facilities Dell's AI Factory architecture provides a secure, government-ready path for the complete AI lifecycle—encompassing data ingestion, model training, validation, and deployment—within government-controlled facilities[23].

The PowerEdge XE series provides GPU-dense compute optimized for federal AI workloads[25]. Two architectural approaches address different mission requirements:

PowerEdge XE9680 / XE9780: Maximum Training Performance

The PowerEdge XE9680 is Dell's flagship 8-GPU Hopper-generation training node, integrating eight NVIDIA HGX H100 or H200 SXM GPUs (as well as other high-end accelerators such as AMD Instinct MI300X and Intel Gaudi 3) in a dense 6U chassis fully interconnected with NVLink/NVSwitch. This SXM/OAM architecture creates a single high-bandwidth GPU domain optimized for large-scale

data-parallel and model-parallel training, where fast, in-fabric collective operations matter more than raw PCIe bandwidth.

The PowerEdge XE9780 extends this architecture into the Blackwell generation, using NVIDIA HGX B300 (Blackwell Ultra) GPUs and offering both air-cooled (XE9780) and direct liquid-cooled (XE9780L) variants for rack-scale deployments. In Dell's AI Factory with NVIDIA, XE9780-class systems are designed to deliver significantly higher LLM training throughput than prior XE9680 configurations, while retaining the same 8-GPU, NVLink-connected building-block model for scaling out clusters.

For Golden Dome's multi-sensor fusion models—integrating orbital infrared tracking, terrestrial radar, and maritime observations—this interconnect density is critical, allowing terabyte-scale training datasets and intermediate feature maps to move across the GPU fabric without bottlenecks.

PowerEdge XE7740/XE7745: Flexible PCIe Architecture

For workloads that prioritize GPU flexibility and broad accelerator choice over maximum NVLink bandwidth, the PowerEdge XE7740 and PowerEdge XE7745 provide a PCIe-based alternative with up to eight GPUs in a 4U chassis. These systems support NVIDIA H100 NVL, H200 NVL, L40S, L4, and NVIDIA RTX 6000 Ada / RTX 6000 PRO GPUs, giving agencies a wide range of options for inference, fine-tuning, and mixed training-and-inference deployments.

The PCIe architecture enables straightforward GPU swaps and upgrades as new accelerators become available, helping protect infrastructure investments across multiple GPU generations.

Common Security Foundation

Both architectures integrate **Silicon Root of Trust**, **Secured Component Verification**, and **data-at-rest encryption**, aligning with federal security requirements and helping agencies maintain a hardened infrastructure posture.

PowerScale Enterprise Storage: Capacity and Throughput for AI training

A single PowerScale F900 node delivers 4M IOPS and 336 GB/s bandwidth, sufficient to feed multi-node GPU training pipelines without storage bottlenecks. Backed by the OneFS scale-out architecture, F900 clusters can linearly scale performance and capacity to support petabyte-scale AI datasets and high-concurrency access patterns. PowerScale integrates data-at-rest encryption (FIPS 140-2 validated), access control, and audit logging, satisfying NIST SP 800-171 requirements for CUI protection.[21].

Near Edge: Tactical Platforms for Forward Operations

The Near Edge brings AI inference to forward operating bases, ground fusion centers, and mobile command posts where connectivity is intermittent and environmental conditions are harsh. Dell's XR series portfolio delivers enterprise compute in ruggedized form factors.

XR8000: Modular Tactical Data Center

The PowerEdge XR8000 uses a sled-based modular design providing superior thermal headroom for GPU acceleration. A single chassis integrates up to 16 compute sleds, each supporting dual Intel Xeon processors or NVIDIA GPUs[5]. Environmental certifications validate tactical readiness: MIL-STD-810H for shock/vibration/temperature extremes, IP65 ingress protection against dust and moisture. Individual sleds deploy, upgrade, or replace without chassis-level maintenance, critical when the nearest depot is days away.

XR7620: GPU-Dense Edge Inference

The PowerEdge XR7620 integrates up to four NVIDIA L40S GPUs in a compact 2U chassis. Each L40S delivers 362 TFLOPS of AI inference throughput, sufficient for real-time computer vision and multi-modal sensor fusion. Tactical coordination centers processing distributed surveillance feeds require this GPU density to run threat detection models locally.

XR5610: Compact Ruggedized Compute

For vehicle integration and space-constrained installations, the PowerEdge XR5610 provides GPU acceleration in a 1U short-depth chassis[6]. Autonomous ground vehicles and maritime autonomous systems integrate the XR5610 when full XR7620 capacity exceeds size, weight, and power budgets.

Extreme Edge: Integration with Autonomous Platforms

The Extreme Edge represents compute at the platform itself: autonomous vehicles, UAS, and tactical robots where size, weight, and power constraints dominate. Dell does not manufacture embedded GPUs or vehicle-integrated compute modules. Instead, Dell's Extreme Edge strategy focuses on ecosystem integration: enabling data pipelines from NVIDIA Jetson/IGX embedded platforms to Dell Near Edge and Core Edge infrastructure.

Autonomous platforms running embedded GPUs for onboard perception generate mission data requiring tactical fusion at Near Edge servers and strategic analysis at Core AI Factory infrastructure. Dell's value at the Extreme Edge is not hardware; it is integration across the continuum.

Security and Operations for DDIL Environments

Deploying AI infrastructure to federal tactical edge environments introduces security challenges absent from traditional data center IT. Edge nodes operate in contested electromagnetic spectrum, platforms experience DDIL connectivity for hours to weeks, yet federal compliance requirements apply uniformly regardless of where infrastructure resides. This section addresses the key areas federal programs must consider: regulatory frameworks governing workload placement, Zero Trust architectures adapted for disconnected operations, AI-specific threat mitigations, supply chain assurance, and unified orchestration for dispersed deployments.

Regulatory Frameworks

Classification and regulatory requirements shape where federal AI workloads execute and how infrastructure must be architected:

- **FedRAMP High SA-9(5):** External Information System Services control requires all data, processing, and AI services for high-impact federal systems reside on U.S. soil[22]
- **DoD Impact Level 6 (IL6):** SECRET classification workloads demand air-gapped infrastructure with no internet connectivity
- **ITAR (22 CFR 120-130):** International Traffic in Arms Regulations prohibits foreign national access to defense-related AI models and training data

These frameworks apply whether infrastructure resides in government data centers, accredited commercial clouds, or tactical edge deployments.

Tactical Zero Trust

The DoD Zero Trust Strategy mandates "never trust, always verify" across all domains, but traditional Zero Trust assumes always-on connectivity to centralized identity providers[4]. When satellite links drop or jamming denies RF spectrum, cloud-based Zero Trust fails. Dell's DoD-validated Zero Trust solution implements Tactical Identity, Credential, Access Management (T-ICAM):

- **Pre-Deployment:** Units synchronize credentials, access policies, and certificate revocation lists before deploying to DDIL zones
- **Disconnected Operations:** Each tactical enclave operates its own policy decision point, making local access control decisions
- **Reconnection:** Edge platforms automatically synchronize audit logs and receive policy updates when connectivity restores

AI-Specific Security

The MITRE ATLAS framework catalogs adversarial AI threats requiring attention[24]. Model poisoning (adversaries injecting malicious training data) requires training data provenance verification and

cryptographic model signing. Model extraction (reverse-engineering weights through repeated queries) requires authentication of inference requests and rate limiting. Dell AI Factory integrates these mitigations: PowerScale maintains immutable snapshots of training datasets, model pipelines generate signed artifacts, and NativeEdge verifies signatures before deploying models to edge nodes.

Supply Chain Assurance

NIST SP 800-171 Revision 3 introduced Supply Chain Risk Management controls requiring verification of AI component provenance[21]. Dell's Secured Component Verification implements cryptographic verification at every supply chain stage: silicon-level hardware root of trust, firmware cryptographic signing, and component-level attestation.

Unified Orchestration

Dell NativeEdge provides the management plane for dispersed tactical operations[7]: zero-touch provisioning for rapid deployment, remote lifecycle management for firmware and software updates, and security posture monitoring with automated remediation. When satellite links to forward operating bases operate 4-6 hours daily, NativeEdge queues updates for execution during connectivity windows.

Dell's Federal Edge AI Capabilities

Federal agencies evaluating edge AI solutions have multiple options. Dell brings specific capabilities informed by federal mission requirements:

End-to-End Integration

Dell provides integrated capabilities across multiple edge layers: Core AI Factory training infrastructure, Near Edge tactical servers, orchestration for dispersed deployments, and integration pathways to Extreme Edge platforms.

Federal Heritage

Decades of DoD, Intelligence Community, and federal civilian partnerships provide institutional knowledge of compliance requirements and mission criticality. FedRAMP authorizations, DISA validations, and DoD program integrations are contract vehicles and active deployments.

Open Ecosystem Architecture

Dell's AI infrastructure supports accelerators from NVIDIA, AMD, and Intel, allowing agencies to select GPUs based on mission requirements, existing software investments, and procurement considerations rather than vendor lock-in. This open approach extends to validated configurations with enterprise software partners, enabling agencies to deploy AI workloads on infrastructure aligned with their operational and compliance needs.

Operational Support

Dell ProSupport Enterprise delivers 24/7/365 support with federal-cleared personnel, on-site spares for critical deployments, and logistics networks reaching forward operating bases globally.