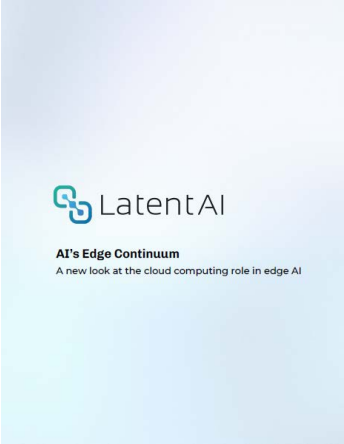
LatentAI

carahsoft.



AI's Edge Continuum:

A new look at the cloud computing role in edge AI

Thank you for downloading this Latent AI resource. Carahsoft is the distributor for Latent's AI and ML solutions available via NASA SEWP V, ITES-SW2, NCPA, and many more contract vehicles.

To learn how to take the next step toward acquiring Latent's solutions, please check out the following resources and information:

For additional resources:
carah.io/latent_resources

For upcoming events:
carah.io/latent-events

For additional solutions:
carah.io/latent-ai

For additional Artificial Intelligence solutions:
carah.io/ai-solutions

To set up a meeting:
LatentAI@Carahsoft.com

To purchase, check out the contract vehicles available for procurement:
carah.io/latent-contracts

For more information, contact Carahsoft or our reseller partners:
LatentAI@Carahsoft.com | 571-662-4235

AI’s Edge Continuum

A new look at the cloud computing role in edge AI

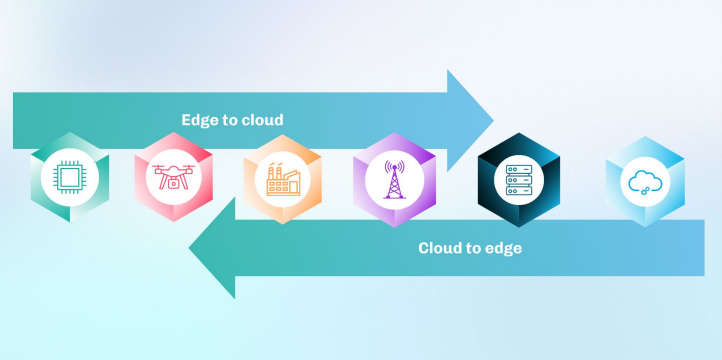
From cloud to edge: Rewiring AI's future

The unstoppable surge of data—expected to hit 175 zettabytes this year—coupled with a soaring need for instant insights, is fueling a massive overhaul in AI deployment.

The old cloud-heavy approach, once AI's cornerstone, is buckling under latency, bandwidth chokeholds, and growing privacy and security worries. The fix is a gutsy leap to the edge continuum, a distributed computing ecosystem with a full range of tools—from vast cloud computing hubs to far-edge devices like factory sensors or battlefield drones equipped with cameras. The edge continuum—a hybrid architecture—distributes AI workloads from the edge to the cloud, bringing processing closer to data sources while leveraging the cloud power for heavier tasks. Moving from "cloud to edge" means we leverage the whole stack and do not heavily rely on centralized cloud computing resources.

The unstoppable surge of data—expected to hit 175 zettabytes this year—coupled with a soaring need for instant insights, is fueling a massive overhaul in AI deployment. The old cloud-heavy approach, once AI's cornerstone, is buckling under latency, bandwidth chokeholds, and growing privacy and security worries.

The fix is a gutsy leap to the edge continuum, a distributed computing ecosystem with a full range of tools—from vast cloud computing hubs to far-edge devices like factory sensors or battlefield drones equipped with cameras.



The edge continuum—a hybrid architecture—distributes AI workloads from the edge to the cloud, bringing processing closer to data sources while leveraging the cloud power for heavier tasks. Moving from "cloud to edge" means we leverage the whole stack and do not heavily rely on centralized cloud computing resources.

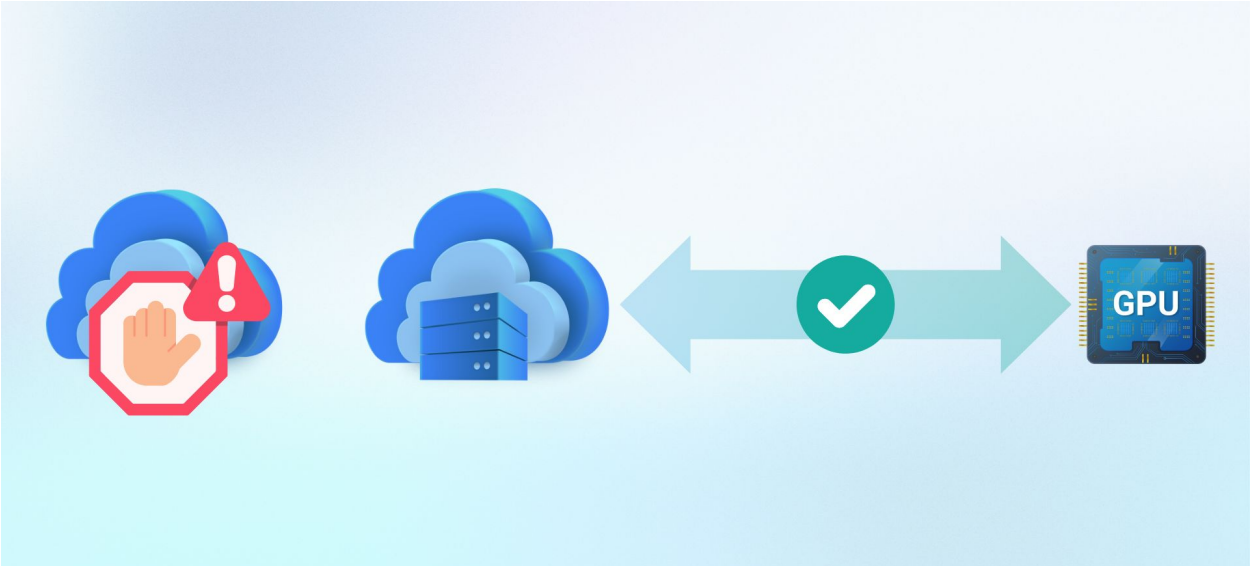
Distinction between cloud computing and edge AI

Today's cloud computing is often praised for handling intensive tasks like model training, where heavy computational resources are available. It's ideal for storing and analyzing large datasets, but it requires massive data transmission, introducing latency and bandwidth costs—not to mention the rise in security and privacy risks as sensitive data travels to and sits in the cloud.

The edge, what is often referred to as edge computing or physical AI, gets attention for managing deployment and data processing closer to data sources, cutting down on communication delays. Here, we place AI models in applications near data sources or directly embedded on local devices (like drones, cameras, smartwatches, and automobiles) that sense, monitor, and record their environment. Embedded AI at the device layer is crucial for real-time decision-making, especially in IoT applications and autonomous systems where immediate responses matter. If those same sensors have to send data back to the cloud for processing, latency occurs, which can undermine timely decisions.

Why the edge continuum is a game-changer

The edge continuum's strength lies in its distributed data processing, transforming performance, security, and costs. By placing AI near the point of data collection, organizations can process data locally—eliminating delays and enabling instant decisions. Think IoT sensors in a warehouse, drones over a disaster zone, or tools in the hands of military personnel. This speed is vital for autonomous vehicles navigating busy streets, industrial robots avoiding errors, or emergency teams in high-stakes scenarios. Plus, keeping sensitive data local enhances security by design, reducing risks of cloud transmission and meeting strict regulations like the EU's GDPR.



By distributing the workload across a flexible, multi-layered system rather than leaning on a single choke point like centralized cloud computing, this approach offers a pragmatic shift that tackles the bottlenecks of efficient scaling head-on.

Adopting an edge continuum necessitates a bold reimagining of architecture, shifting away from rigid, centralized cloud computing toward adaptable, workload-specific resource distribution. Several industries are primed to capitalize on the benefits. Today, manufacturers harness edge AI for predictive maintenance—identifying equipment degradation before breakdowns occur—while quality control mechanisms detect flaws in milliseconds, reducing waste and enhancing safety. In the energy sector, smart grids optimize supply and demand in real time, remote monitoring oversees wind turbines in extreme conditions, and predictive analytics curbs emissions—delivering gains for both efficiency and sustainability. Defense applications also seize the advantage where real-time threat detection pinpoints irregularities on the frontlines, autonomous drones adjust to evolving scenarios, and situational awareness tools refine decision-making in volatile, high-stakes settings. Each of these industries has a significant stake in benefitting from the cost and latency reduction that processing data locally, or as close to the source as possible, provides.

The hybrid approach: Edge continuum

For most, a multi-layered architecture presents a promising shift from cloud-only AI or edge-only AI driven by speed, security, and tech advancements.

The key principle of the edge continuum is to utilize distributed computing power and execute data processing as close to the source as possible while preserving the ability to pass harder problems securely and confidently up the continuum as necessary.

Real-world examples: AI in the military

Defense offers a striking example, with four edge layers: **Tactical** (frontline devices), **Operational** (field coordination), **Command** (regional oversight), and **Strategic** (high-level planning). Each layer executes the "sense, make sense, and act" cycle—sensing data, interpreting it, and responding—as close to the action as possible. A soldier's wearable might detect a threat locally, while a command center aggregates regional insights, all synced via distributed computing. This layered approach maximizes agility without sacrificing scale.

The hybrid approach: Edge continuum

Tactical Edge: The Pulse of the Frontline

The Tactical Edge layer is where the sensors turn analog into digital, working as parts of platforms like drones, underwater vehicles, or tactical kits are located—think of it as the “sensing” crew. It runs in tough, disconnected spots (DDIL mode) with small, lightweight, low-power, rugged devices that don’t need much storage or power. These edge AI devices can tweak data flow based on connection strength and quickly flag oddities to operators and the next layer. They’re built to handle efficient, accurate AI models that process heaps of data fast without hogging memory—perfect for spotting familiar signals and passing unknowns up the chain based on mission rules. Plus, they can be integrated into edge applications such as machine-to-machine and machine-to-human chats. This layer keeps the sensor-to-shooter loop tight, enabling quick autonomous or human-triggered actions.

Operational Edge: The Workflow Bridge

The Operational Edge layer sits right between the Tactical and Command Edge layers, acting as a critical link in the AI ecosystem. Its job? To take filtered data from edge devices and enhance it based on operational requirements. It can integrate additional data from other domains as needed, apply preprocessing algorithms to refine it, and then forward the processed insights to the Command layer. This layer marks the starting point for sensor fusion, where it begins combining diverse data inputs into a unified, actionable view.

The Operational Edge balances size, weight, and power (SWaP) requirements with better access to energy than the Tactical Edge, while still relying on tactical communication networks. The models here are built to handle the fast, complex data pouring in from below, delivering real-time feedback and boosting situational awareness.

The hybrid approach: Edge continuum

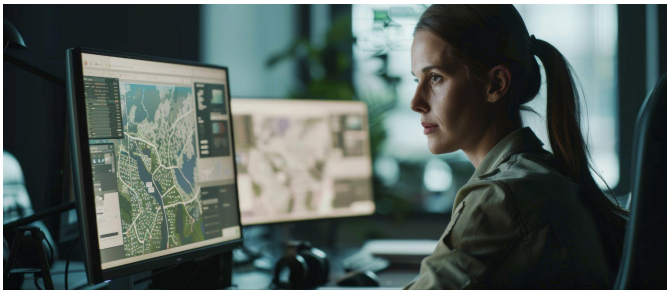
Flexibility is a big plus. The Operational Edge can either pass along its pre-filtered raw data or share key takeaways from its analysis, depending on what's needed upstream. It's also equipped for machine-to-machine and machine-to-human communication, making it a versatile player. In short, this layer doesn't just connect—it processes and refines, ensuring the right information flows efficiently to the right places.

Command Edge: Where Tactical Meets Actionable

The Command Edge sits as the final stop in the tactical communications chain. Think of it as the hub that pulls together processed data from multiple Operational Edge devices in a specific scenario—whether that's a factory floor or a field operation. Here, the models juggle a fast-moving mix of data from various sensors, all prepped and polished by downstream Operational Edge devices.

The key to the Command Edge is its flexibility—it can run stationary or go mobile, boosting its resilience depending on the situation. That adaptability means its computing power aligns with size, weight, and power (SWaP) needs, just at a different scale than the layers below. The Command Edge is about delivering clear, actionable information straight to human operators for decision-making. Beyond that, the Command Edge is used to quickly aggregate and redeploy information from all layers above it for "on the field" intelligence and decisions.

Strategic Edge: The Big-Picture Backbone



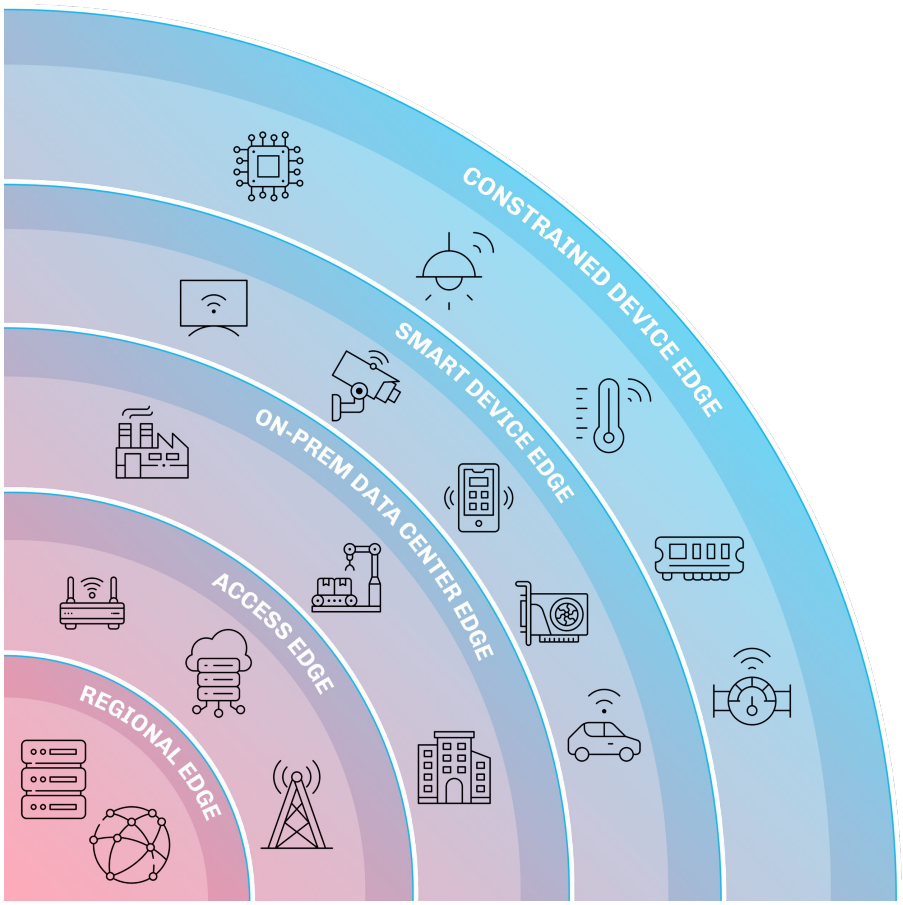
At the top tier of this architecture, this layer takes the baton from the Command Edge and collects and analyzes all that processed data. It lives within Department of Defense central data centers, national command-and-control hubs, and continuity-of-government sites, all hardwired into the broader network. It's the backbone for storing data, running deep analysis, and driving big decisions.

The hybrid approach: Edge continuum

The models at this level are powerhouse performers, built to process both real-time and historical data. That dual capability helps pinpoint strategic priorities—whether for today's operations or tomorrow's planning. This layer also doubles as the AI/ML model factory, where specialized teams focus on training, fine-tuning, securing, and rolling out those models. Across each combatant command, personnel at every layer—from Tactical to Strategic—are dedicated to keeping the models sharp, updating them with the latest operational intel and data straight from the Tactical Edge.

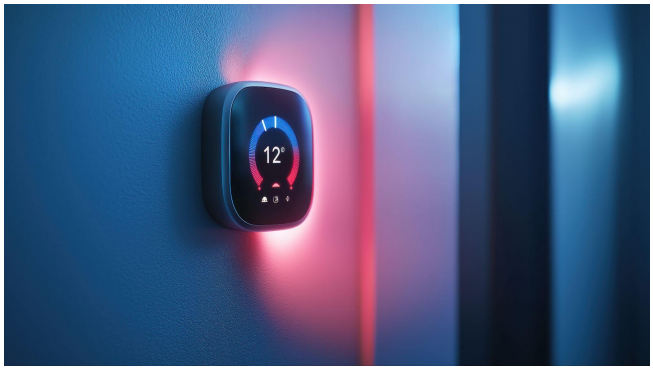
Real-world examples: AI-powered manufacturing ecosystem

Building on the military-inspired AI framework, here's how the edge continuum powers a commercial industrial setting.



The hybrid approach: Edge continuum

Constrained Device Edge: Minimal But Mighty



The Constrained Edge layer is the lean, no-frills tier of the edge continuum, home to devices like smart light bulbs, basic thermostats, or simple IoT sensors. These devices all “do one thing well”. They are tiny, ultra-low-power gadgets with minimal processing muscle and minimal programmability. Devices at this layer are hardwired for churning out small, steady streams of data. Edge AI here is stripped down, running lightweight models that handle basic pattern recognition or anomaly detection—like a light bulb dimming based on daylight or a sensor pinging an alert for odd temps. These devices lean on rigid, pre-set rules, sipping power and passing the baton to smarter layers when things get complex.

Smart Device Edge: Smarter and More Nimble

The Smart Device Edge layer hosts devices with more grit—like smartphones, wearables, or connected cameras. These devices can run flexible, mid-tier AI models, processing data on the fly with decent storage and power. They’re programmable, adaptable, and often used in low bandwidth situations so they juggle local decisions—like a smartwatch tracking vitals or a camera tagging faces—while syncing with the cloud or other layers when bandwidth allows. They shine in real-time responsiveness, balancing efficiency with smarts, and can push polished outputs (think dashboards or alerts) to humans. Compared to the tactical edge used in defense scenarios, this layer is the workhorse of the edge, turning raw data into actionable insights without breaking a sweat.

The hybrid approach: Edge continuum

On-Premises Data Center Edge: The Local Processing Hub

The On-Prem Data Center layer—think local servers humming in a nearby rack—is the edge continuum's heavy lifter. Often, it includes a layer of enhanced edge compute that sits between the nimble Smart Device edge and the heavy compute of the on-prem data center. Enhanced edge compute contains portable GPU-packed units—built for heavier AI workloads like real-time video analytics, multi-sensor fusion, or on-the-fly model inference—tasks too big for the Smart Device Edge but not yet needing a full data center's might. The Enhanced Edge Compute layer shines at preprocessing complex feeds before handing it off to the On-Prem Data Center.

The On-Prem Data Center refines raw inputs, runs heftier algorithms, and preps insights for the Regional Edge layer—or stores them for later. It's the steady backbone, balancing real-time processing with longer-term data hoarding, all while keeping latency low and security tight since it's on-site. This layer's the unsung hero for setups like factory floors, hospitals, or any spot where local control and compute power need to coexist.

The Access Edge: Linking Edge to Region

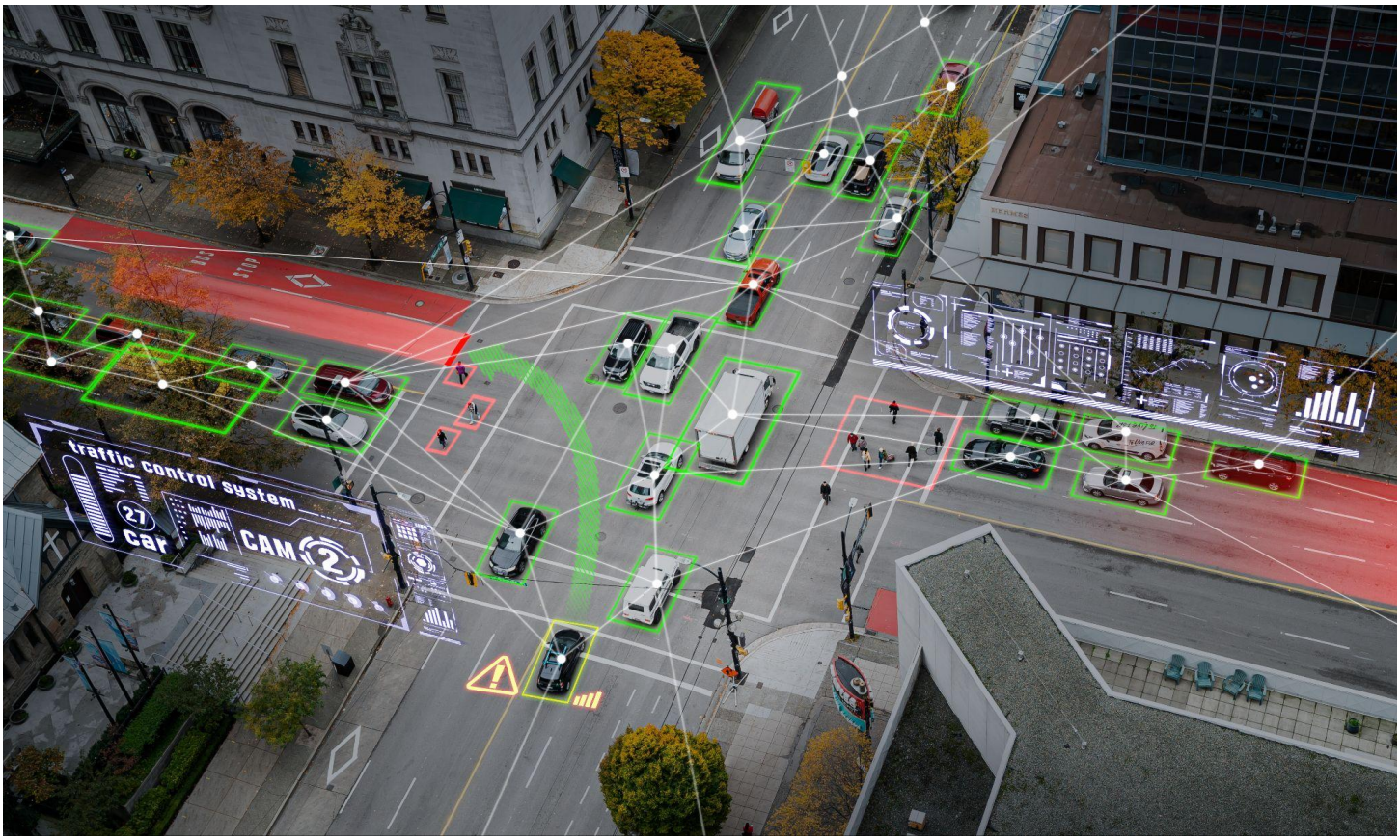
The Access Edge layer serves as the networking linchpin, channeling data from the On-Prem Data Center Edge to the broader Strategic Regional Edge. An aggregation point like switches, routers, or edge gateways stationed at the perimeter of local sites funnels processed insights from GPU-heavy data center racks to regional hubs. It's built for high-bandwidth, low-latency handoffs, smoothing the flow of AI-driven outputs—like fused sensor data or real-time analytics—over tactical or enterprise networks. Less about raw compute and more about connectivity, the Access Edge ensures seamless, secure links to the Strategic Regional Edge, where larger-scale decisions or cloud integration kick in, keeping the continuum's data pipeline tight and responsive.

The hybrid approach: Edge continuum

Regional Edge: The Corporate Brain

At the top, the Regional Edge takes the reins, collecting and analyzing data from the On-Prem Data Center across multiple plants. Housed in corporate data centers with cloud access, it's hardwired into the company's network and serves as the backbone for storage, analysis, and long-term strategy. It pulls in refined streams via the Access Edge, integrating them into a unified view for enterprise-wide decisions—like optimizing supply chains or forecasting trends. This layer leverages its hefty compute and storage to train advanced AI models, pushing insights back down the chain or archiving them for future use.

Combining strengths of edge and cloud computing



Smart cities, industrial systems, and healthcare provide examples where low latency and local processing are non-negotiable. The edge continuum is being shaped by real-world demands like privacy (keeping data on device), cost (reducing network strain), and scalability. Several large organizations are building hybrid architectures in support of AI and realizing three key benefits.

First, it spreads computational demand. The cloud handles heavy lifting—training massive models or crunching giant datasets—while edge nodes and devices take on localized inference and real-time tasks. This means you're not piling everything onto one overstretched resource. As data volume or user demand spikes—like millions of IoT sensors firing up or a city's worth of autonomous cars hitting the road—the system scales out and up. Edge layers absorb the surge locally and keep the cloud services from drowning.

Second, it's resource-efficient. Constrained edge devices, like smart cameras or wearables, can preprocess data—filtering noise or running lightweight models—before passing only what's essential upstream. This cuts bandwidth strain and lets the system handle more endpoints without needing exponentially more cloud power. Think of it as a triage: each layer does what it's best at so the whole setup can grow without choking.

Third, it's adaptable. The continuum isn't rigid—it flexes with the workload. When there is a planned or sudden AI-driven surge, up-cloud resources can easily be spun up and lightweight models can be pushed to new nodes when you need to expand to new sites. This dynamic allocation keeps costs in check and lets AI scale to new use cases without a complete overhaul.

Hidden challenges for the edge continuum

The edge continuum has its share of drawbacks. Managing multiple layers ramps up complexity, giving developers a real headache as they wrestle with diverse hardware—from low-power sensors to beefy edge servers—and unpredictable connectivity. Power consumption can become an unforeseen problem; edge devices, especially in remote or dense setups, might drain batteries or spike energy costs if not optimized tightly.

Security must be factored in at the beginning, as distributing AI across more endpoints multiplies attack surfaces. Securing every layer (from a factory sensor to a cloud hub) is more challenging than locking down a single data center.

Scaling distributed architecture demands smart orchestration tools, like Kubernetes or custom frameworks, to dynamically manage resources across layers.

How organizations can steer toward a distributed AI future

Map Your Data Terrain

Start with the source—where's your data born? Organizations should pinpoint their data origination hotspots and assess their approach. Are there gains that can be made for adjusting the architecture around latency sensitivity, bandwidth limits, or security? Match your AI workloads to the edge-to-cloud spectrum—far-edge for split-second decisions, cloud for heavy-duty analytics.

Build a Flexible Stack

Invest in a hybrid architecture that scales: lightweight AI models for edge devices (think TinyML for sensors), mid-tier processing for regional hubs, and cloud muscle for training and big-data crunching.

Implement development tools that help you scale

Select development platforms that address the current tooling gap. Look for solutions that offer:

- Model optimization capabilities
- Hardware-aware deployment tools
- Flexible customization options
- Automated quantization and pruning
- Security features, including model encryption and watermarking

How organizations can steer toward a distributed AI future

Develop Edge-Optimized Applications

Re-architect applications to leverage edge computing. Focus on lightweight, modular designs with AI/ML capabilities tailored for real-time processing and local decision-making at the edge.

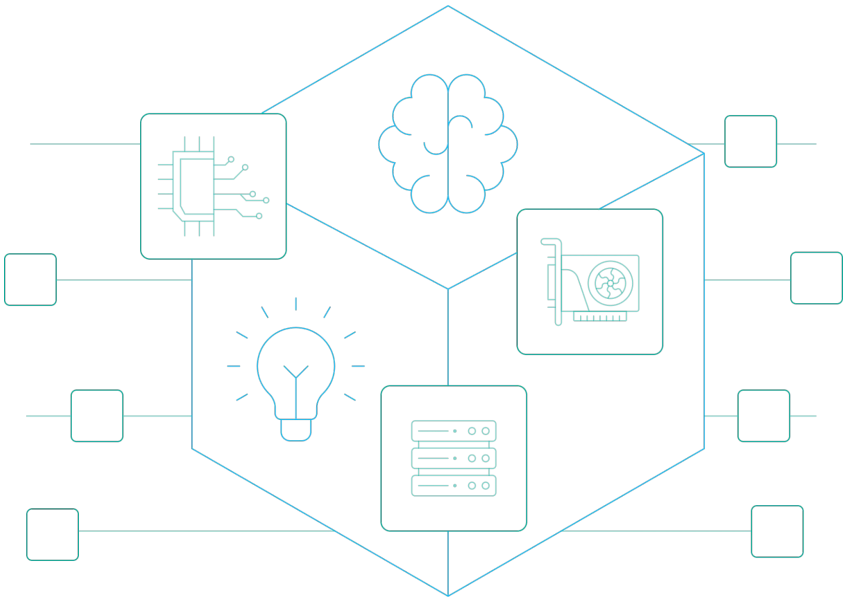
Harden the Edges

Security isn't optional when AI's running on a battlefield drone or a smart-city grid. Encrypt data at rest and in transit, lean on edge-native protocols like MQTT, and bake in zero-trust principles. The continuum's spread-out nature cuts centralized breach risks, but only if each node's locked down.

Test, Tweak, Deploy

Pilot edge AI in a high-impact, low-risk zone—say, a single warehouse or a fleet of delivery bots. Measure the wins: latency drops, cost savings, uptime boosts. Tweak the balance between edge and cloud, then scale what works. The edge continuum thrives on iteration, not perfection.

How Latent AI enables edge continuum architecture



Latent AI enables the edge continuum by delivering tools and solutions that bridge the gap between advanced artificial intelligence and the distributed, real-time demands of edge computing. As organizations increasingly rely on edge devices—from IoT sensors to autonomous systems—to process data instantly and securely, Latent AI provides a flexible, efficient framework tailored to this dynamic landscape. By optimizing AI deployment across diverse hardware and use cases, Latent AI ensures that the edge continuum thrives, enabling everything from split-second decisions in the field to robust data privacy without sacrificing performance.

- **Scalability:** Latent AI offers model- and hardware-agnostic tools, allowing developers to create appropriately sized AI models for the right hardware across the edge continuum.

- **Efficiency:** Latent AI's hardware-aware optimizations ensure efficient inference within the resource constraints at various layers of the edge continuum.

- **Interoperability:** Latent AI tools seamlessly integrate with diverse software and hardware stacks, simplifying the development and deployment of edge AI.

- **Security:** Latent AI stands out by optimizing models for both performance and security. By distributing inferences across edge devices, it keeps sensitive data local, addressing privacy concerns and bandwidth limits.

info@latentai.com

Visit latentai.com for more information.