e-book

# AI Beyond the Cloud: The Infrastructure of the Physical World

# AI Is Leaving The Cloud

For most of AI's commercial history, intelligence was something you reached for — a service you called, a cloud you connected to. Data traveled up, decisions came back down. That architecture made sense when the world generating the data was digital. It no longer describes the world we are building.

AI is now being embedded directly into physical systems; into vehicles, factories, hospitals, and cities. And the infrastructure required to support it looks nothing like what the cloud era produced.

Vehicles now operate as mobile AI systems, continuously interpreting sensor data to support navigation, safety, and operational intelligence. Cities deploy large-scale vision platforms that analyze video streams to improve traffic flow, detect incidents, and enhance public safety.

Industrial machines run embedded models that monitor equipment health and predict failures before they occur. Medical devices perform AI-assisted diagnostics in ambulances, clinics, and patient homes. Global supply chains rely on intelligent sensors that detect anomalies in temperature, location, and product integrity while shipments move across borders.

These environments generate massive volumes of sensor data, often in locations where latency, reliability, or governance requirements make centralized processing impractical. Decisions frequently must be made immediately and locally. A vehicle identifying a pedestrian cannot wait for a round trip to a distant cloud region. A factory detecting a safety risk must respond instantly. A medical monitoring system cannot delay clinical alerts while data traverses multiple network hops.

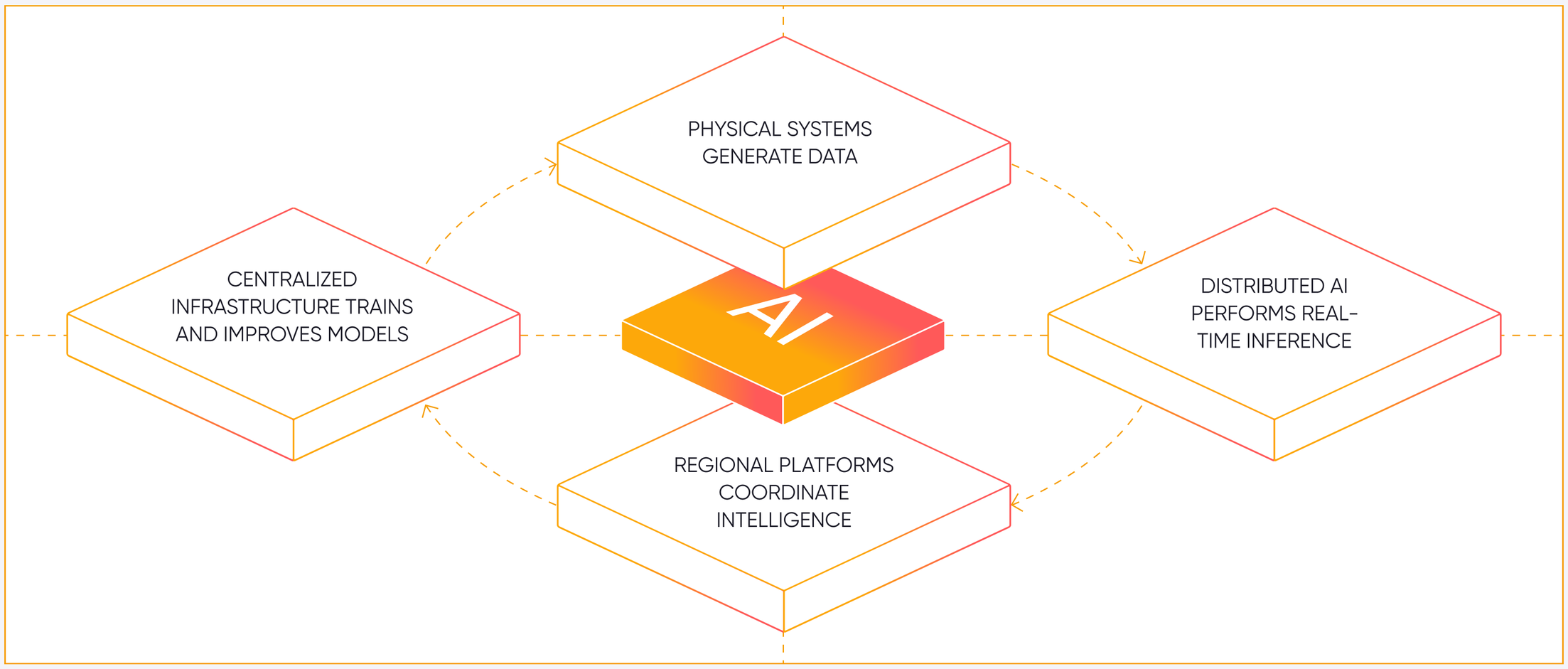
**Result**

Vehicles now operate as mobile AI systems, continuously interpreting sensor data to support navigation, safety, and operational intelligence. Cities deploy large-scale vision platforms that analyze video streams to improve traffic flow, detect incidents, and enhance public safety.

## In this model, the physical world becomes the primary source of data. In practice, this means:

- Sensors, machines, vehicles, and devices continuously observe real-world conditions, generating data volumes that make centralized collection impractical and centralized processing impossible at scale.
- Edge AI systems process that data close to where it is generated, enabling real-time inference and immediate action without reliance on a distant cloud, because a vehicle identifying a hazard, or a machine detecting a fault, cannot wait.
- Regional AI platforms aggregate operational data from thousands of distributed systems, applying larger models to detect patterns and coordinate operations while respecting the jurisdictional boundaries that govern where that data can travel.
- At the top of the stack, large-scale accelerated computing infrastructure trains and refines models that are then redistributed back across the global inference pipeline, closing the loop between real-world observation and continuous system improvement.

# Continuous Intelligence Loop




## This architectural shift to a continuous AI pipeline is already visible across many industries.

In mobility systems, vehicles perform real-time inference onboard while fleet analytics and model lifecycle management occur in regional platforms.

In smart cities, video streams are processed in metro-scale edge AI clusters to maintain local governance of sensitive data while enabling real-time situational awareness.

In healthcare environments, portable medical systems perform on-device analysis while clinical AI platforms provide specialist support and broader diagnostic insights.

Industrial environments run embedded AI within machines while site-level clusters monitor operational performance across entire facilities.

In each of these scenarios, intelligence is no longer confined to a single cloud environment. It is distributed across devices, sites, regional platforms, and large-scale training infrastructure.

This transformation is redefining what “AI infrastructure” means.

Instead of a centralized compute environment connected to devices, AI infrastructure now includes sensor systems, edge inference environments, regional AI platforms, and accelerated model training infrastructure working together as a single operational system.

The infrastructure required to support these systems must operate consistently across physical environments, regulatory boundaries, and heterogeneous compute platforms.

# AI Architecture Is Distributed By Design

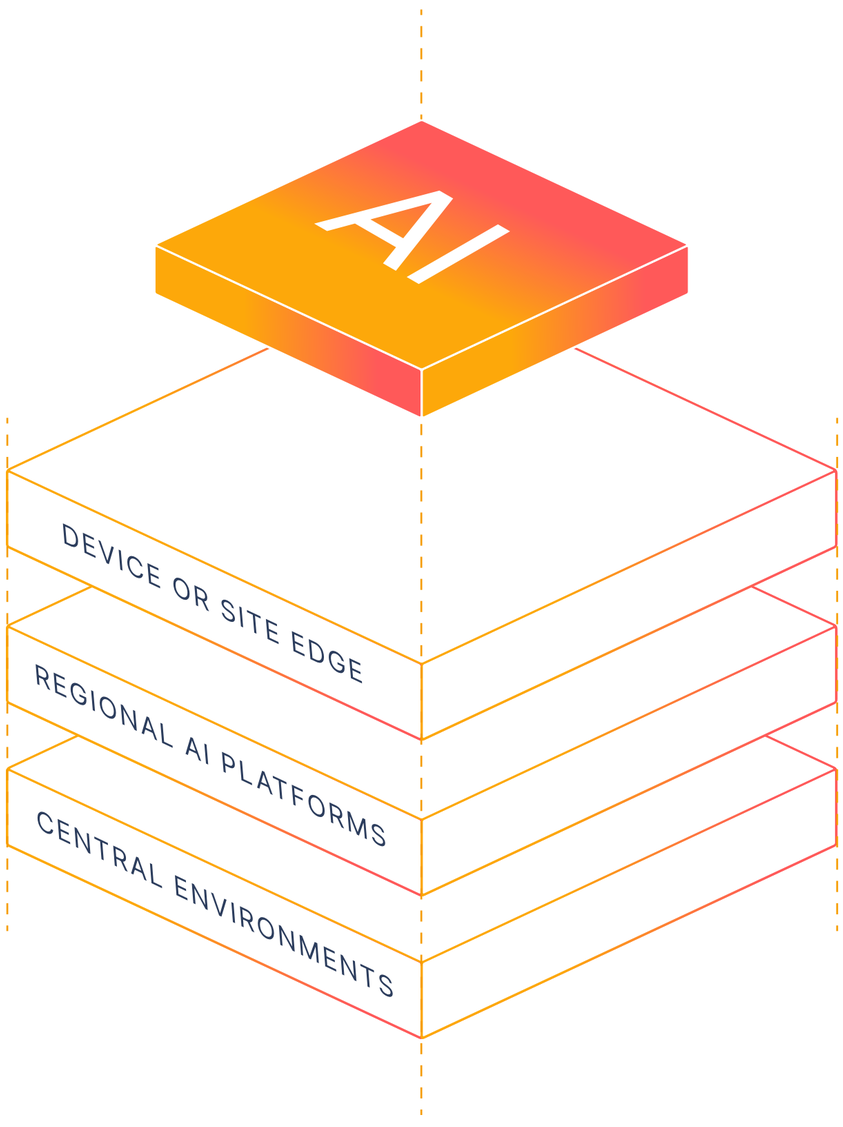
Describing AI infrastructure as "distributed" risks making it sound like an architectural preference. It is not. The distribution is a consequence of where data is generated, of the speed at which decisions must be made, and of the regulatory environments that govern what can move where.

Understanding why the stack looks the way it does requires looking at each layer not as a design choice, but as a response to a constraint.

The device edge exists because latency demands it. Regional platforms exist because data gravity and governance require it. Central cloud environments exist because model training and lifecycle management benefit from scale.

Each layer has a reason. Together, they define a new kind of infrastructure, one that is distributed by necessity, not by convenience.

### The Distributed AI Stack




## Let's look into the stack more closely:

At the foundation of this architecture is the device or site edge. Sensors, machines, vehicles, and cameras generate continuous streams of data from the physical world. Many of the most time-sensitive decisions happen directly at this layer.

Above the device layer sit near-edge or regional AI environments. These platforms aggregate data from many distributed systems and apply larger models or analytics that require greater compute resources.

Finally, at the top of the stack are central cloud environments and model lifecycle systems. These environments train large models, manage datasets, and orchestrate the continuous improvement of AI systems operating in the field. Updated models, policies, and software are then redistributed across the distributed inference pipeline.

This layered architecture is already visible across industries. Vehicles combine onboard inference with regional fleet intelligence. Smart cities process video at metro edge environments. Industrial sites run machine-level AI supported by on-site clusters. Global manufacturing networks combine plant-level intelligence with regional analytics.

AI in the physical world is therefore distributed by design, operating across multiple layers of infrastructure, rather than a single centralized cloud.

# Why Data Gravity Matters More Than Ever

Once AI systems operate across multiple layers of infrastructure, a new question becomes unavoidable: **where should intelligence actually run?**

Data gravity refers to the tendency of large datasets to remain close to where they are generated or governed. In practice, the answer is rarely determined by technology alone. It is shaped by the realities of data gravity, regulatory requirements, and operational complexity.

Data gravity refers to the tendency of large datasets to remain close to where they are generated or governed. Moving massive volumes of sensor data across networks can introduce latency, cost, and security risks. In many cases, transporting that data to a distant processing environment is simply impractical. Regulation only strengthens this gravitational pull. Many AI systems now operate in environments where data must remain within specific jurisdictions or controlled operational boundaries.

## Urban Monitoring / Smart Cities

### Example Data Type

Public video, traffic monitoring footage, facial or behavioral analytics.

### Key Governance / Compliance Frameworks

GDPR (EU), UK GDPR & Data Protection Act 2018, local municipal surveillance policies, U.S. state privacy laws (e.g., CCPA/CPRA).

### Key Governance / Compliance Frameworks

Public video must remain under municipal control. Most cities require processing to occur locally. The network cannot route this data to distant infrastructure and remain compliant.

## Healthcare Systems

### Example Data Type

Patient health records, diagnostic imaging, medical telemetry.

### Key Governance / Compliance Frameworks

HIPAA (U.S.), GDPR & UK GDPR (EU/UK), HITECH Act, regional health data residency regulations.

### Key Governance / Compliance Frameworks

Clinical AI platforms must operate inside regionally governed, compliance-certified infrastructure. The network cannot allow traffic to traverse uncontrolled jurisdictions. Data residency is a hard architectural boundary, not a preference.

## Industrial / Manufacturing Environments

### Example Data Type

Operational technology (OT) telemetry, machine performance data, safety systems.

### Key Governance / Compliance Frameworks

IEC 62443 (Industrial Cybersecurity), NIST SP 800-82, ISO 27001, sector-specific safety regulations.

### Key Governance / Compliance Frameworks

OT networks are segmented from external systems by design for safety, not convenience. AI analytics must operate within those boundaries. The network is not permitted to bridge them without strict controls.

## Pharma and Regulated Supply Chains

### Example Data Type

Temperature logs, shipment telemetry, product traceability records.

### Key Governance / Compliance Frameworks

FDA 21 CFR Part 11, EU Good Distribution Practice (GDP), DSCSA (U.S. Drug Supply Chain Security Act), GS1 traceability standards.

### Key Governance / Compliance Frameworks

Every data movement across the supply chain must produce an auditable record. That means compliant infrastructure at every handoff point the network cannot treat pharmaceutical telemetry like ordinary IoT traffic.

These constraints fundamentally shape AI architecture, and often mean AI's distributed architecture is driven not by infrastructure convenience, but by data gravity and governance requirements.

Instead of sending all data to a centralized cloud, organizations increasingly process and analyze information close to where it originates, operating within city or regional requirements, or using site-level clusters.

For organizations deploying AI across physical environments, this creates an important architectural reality. The location of data, regulatory boundaries, and operational control systems determine where inference, analytics, and model operations can run.

Infrastructure must therefore adapt to these constraints rather than attempting to override them.

# Connectivity Becomes The Control Layer

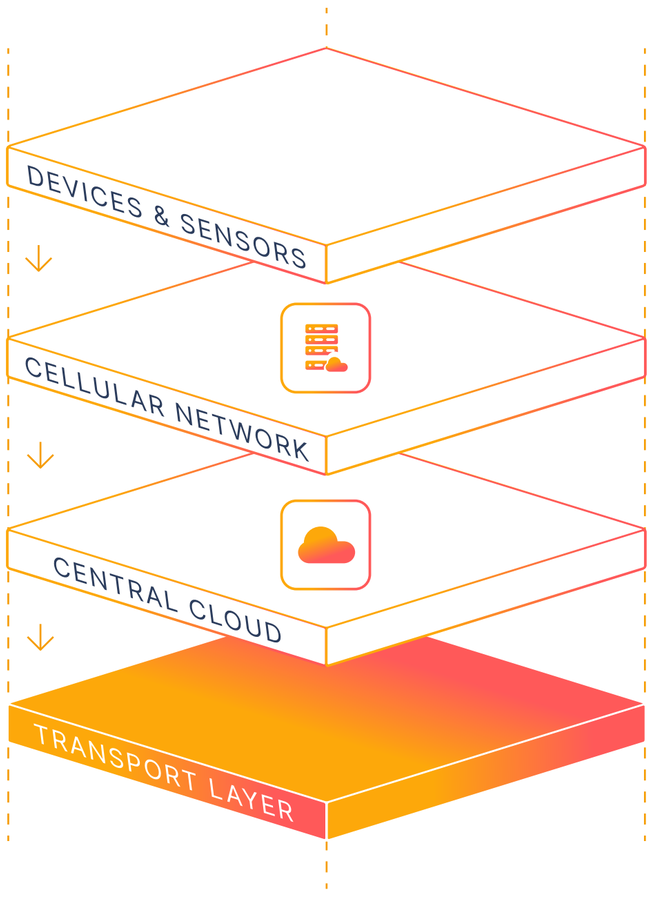
As AI systems move into the physical world, connectivity takes on a fundamentally different role.

In traditional architectures, networks primarily transported data. Devices generated information, traffic flowed through operator infrastructure, and centralized systems performed analysis.

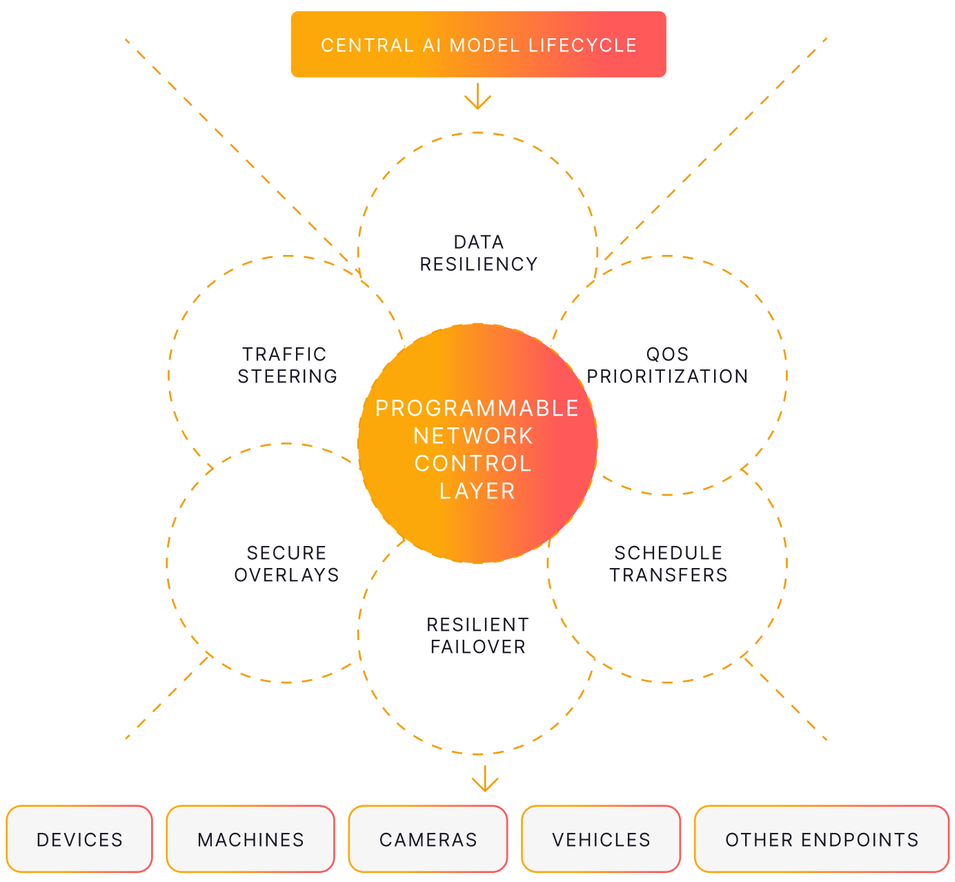
The network itself had little awareness of how applications behaved or where workloads were executed.

Distributed AI systems break this model.

## Connectivity as Transport




## Connectivity as Control Layer

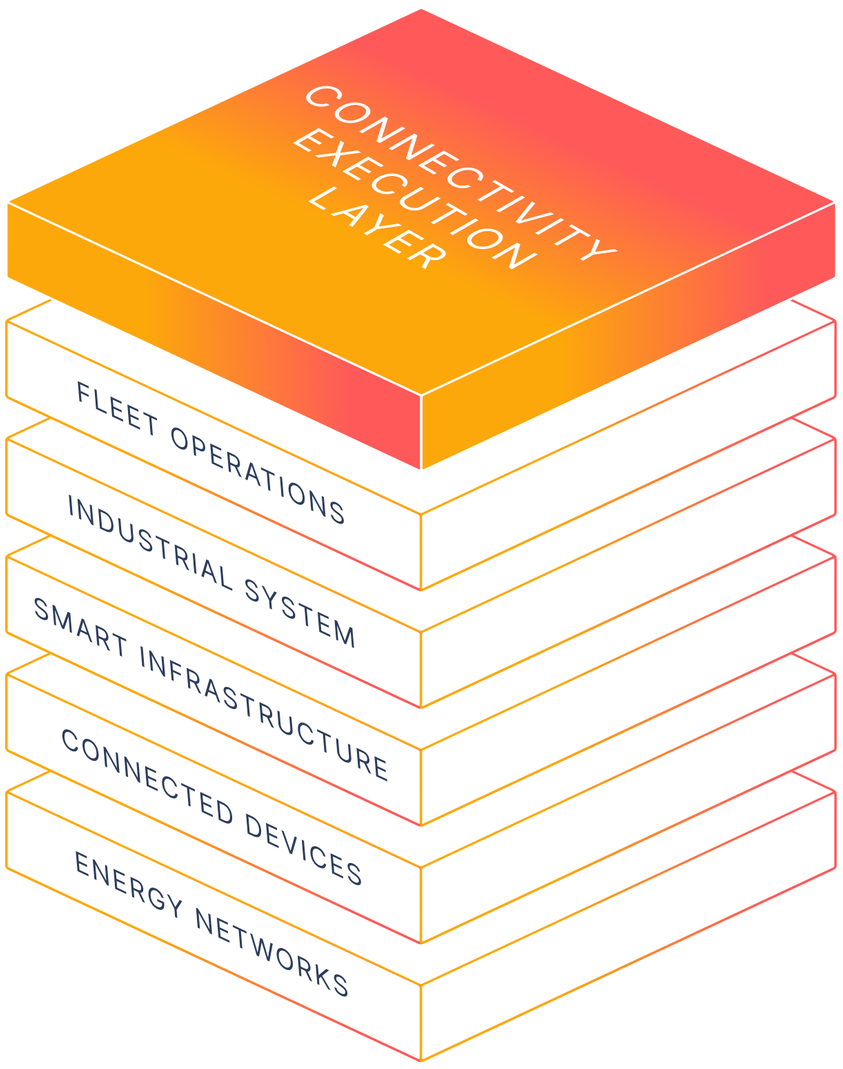



Once inference, analytics, and model management operate across multiple layers of infrastructure, the movement of data becomes a governed architectural decision, not a simple transport problem.

Data must reach specific inference environments, remain within defined jurisdictions, and interact with operational systems that cannot tolerate unpredictable latency or routing behavior.

# The Network Becomes Responsible For Enforcing The Operating Rules Of The AI System

Consider the environments where real-world AI operates. A city-scale video platform cannot simply stream raw camera feeds into distant infrastructure. Video data must often remain within municipal environments due to governance and privacy rules.




In this context, connectivity behaves less like a transport service and more like an execution layer for distributed AI systems, governing how data moves, where intelligence operates, and how real-world systems remain synchronized with the platforms that analyze and improve them.

Only alerts or derived insights may leave those boundaries. The network must enforce those constraints.

In healthcare systems, medical telemetry and diagnostic data must reach approved clinical platforms while remaining within regulated infrastructure governed by privacy frameworks such as HIPAA or GDPR. Connectivity cannot allow traffic to traverse uncontrolled jurisdictions.

Operational technology (OT) networks that control machinery are typically segmented from external networks for safety and cybersecurity reasons. AI analytics may operate across site clusters and enterprise systems, but the network must maintain strict boundaries between those environments.

When the network is responsible for maintaining the integrity of the AI architecture, that responsibility includes steering traffic toward the correct inference environments, enforcing geographic boundaries on regulated data, prioritizing time-sensitive signals generated by physical systems, and establishing secure overlays between distributed compute layers.

It may also involve scheduling large telemetry transfers, synchronizing model updates, and triggering automatic failover when connectivity conditions change.

As AI infrastructure expands beyond centralized clouds into vehicles, factories, hospitals, and cities, the network becomes something more fundamental.

It becomes the control layer that allows distributed intelligence to function coherently across the physical world.

# Five Design Constraints For Physical AI Systems

Physical AI systems do not fail because the models are wrong. They fail because the infrastructure around the models cannot support how the real world actually behaves. Connectivity drops. Data cannot cross a border. A machine must respond in 40 milliseconds, not 400. A factory floor cannot expose its operational network to external traffic.

Five design constraints appear consistently across every physical AI deployment we examine. They are not a checklist to complete, they are the structural realities that separate AI systems that work in the real world from those that work only in a controlled environment. Understanding them is the prerequisite for building infrastructure that holds.

## Placement And Latency

Many AI decisions must occur within milliseconds of a real-world event. Vehicles interpreting sensor data, industrial machines detecting anomalies, or medical devices monitoring patient telemetry cannot rely on distant infrastructure for immediate responses.

Inference therefore runs close to where data is generated, often directly on devices or within local edge environments. Regional platforms then coordinate broader analysis and system optimization.

## Data Gravity And Sovereignty

Data generated in the physical world often cannot move freely across infrastructure environments. Large datasets are expensive to transport, and many categories of information are governed by regulatory or operational controls.

Public video, healthcare records, and industrial telemetry frequently must remain within specific jurisdictions or operational environments. These constraints shape where analytics platforms and AI services can operate.

## Resilience And Continuity

Physical systems must continue operating even when connectivity conditions change. Vehicles cross network boundaries, industrial environments may experience intermittent coverage, and emergency systems cannot tolerate service interruptions.

AI infrastructure therefore requires resilient connectivity models, including multi-network access (the ability to connect across multiple operators simultaneously without reconfiguring the device), combined with automated failover paths, and the intelligence to continue local inference during temporary disconnections.

## Security And Segmentation

Many AI deployments interact directly with critical infrastructure. Industrial machinery, healthcare systems, and municipal operations require strict security boundaries between operational systems and external networks.

Connectivity must therefore support strong segmentation, encrypted communication, and controlled access to ensure that AI systems integrate safely with operational environments.

### Programmable Policy Control

Distributed AI systems operate across multiple infrastructure layers and regulatory environments. Connectivity must be able to enforce how data flows across those layers.

This includes steering traffic toward the correct AI platforms, enforcing regional data boundaries, prioritizing operational signals, and adapting network behavior as system conditions change.

Together, these constraints define the architecture of modern physical AI systems. Successful deployments do more than connect devices to the cloud.

They align infrastructure, data governance, and connectivity behavior to ensure that distributed intelligence can operate reliably across the real world.

# Deployments In The Wild: Industry Examples

The architectural patterns described in previous chapters are already visible across multiple industries.

While each deployment has unique operational requirements, the underlying infrastructure constraints remain remarkably consistent.

Across sectors, distributed AI systems combine edge inference, regional intelligence platforms, and centralized model lifecycle infrastructure. What differs is how governance, latency, and operational requirements shape the placement of those components.

The following industry clusters illustrate how these design patterns appear in practice.

## Mobility And Public Safety

Autonomous vehicle fleets and connected emergency vehicles operate as mobile AI systems, continuously generating sensor data while moving across networks, operators, and geographic regions.

Vehicles perform real-time inference directly onboard, to support navigation, safety, and system monitoring. At the same time, fleet telemetry and operational data flow to regional platforms that coordinate analytics, model updates, and system optimization.

Emergency response vehicles introduce even stricter requirements. Medical telemetry, live video, and AI-assisted diagnostics must reach hospital systems with minimal latency while remaining within regulated healthcare infrastructure.

### Infrastructure Insight

Connectivity policies must move with the asset itself, ensuring that data is routed to the correct regional platforms while maintaining continuity across networks and jurisdictions.

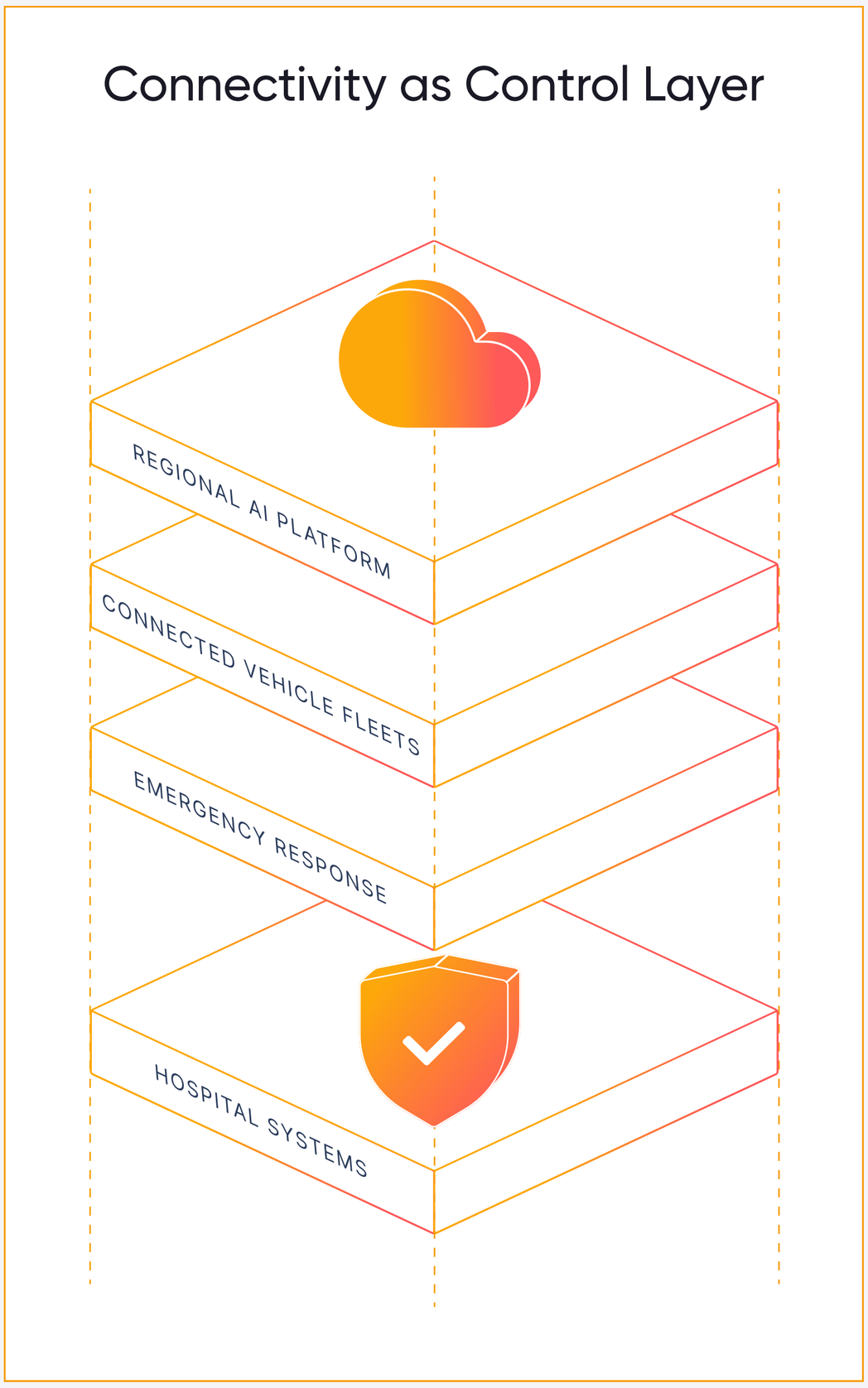
Consider a fleet of connected emergency vehicles operating across a metropolitan region. Each vehicle runs onboard AI to support navigation, patient monitoring, and real-time triage assistance.

As the vehicle moves, telemetry, including live patient vitals and AI-assisted diagnostic signals must reach the receiving hospital before arrival, routed within compliant healthcare infrastructure, regardless of which operator's network the vehicle happens to be on.

The connectivity layer is not a passive participant. It determines whether the clinical AI system functions as designed or fails at the moment it is needed most.

**Infrastructure Insight**

Policies must follow moving assets across carriers and regions.




# Vision Systems And Public Infrastructure

Large-scale vision systems deployed across cities and retail environments generate enormous volumes of visual data.

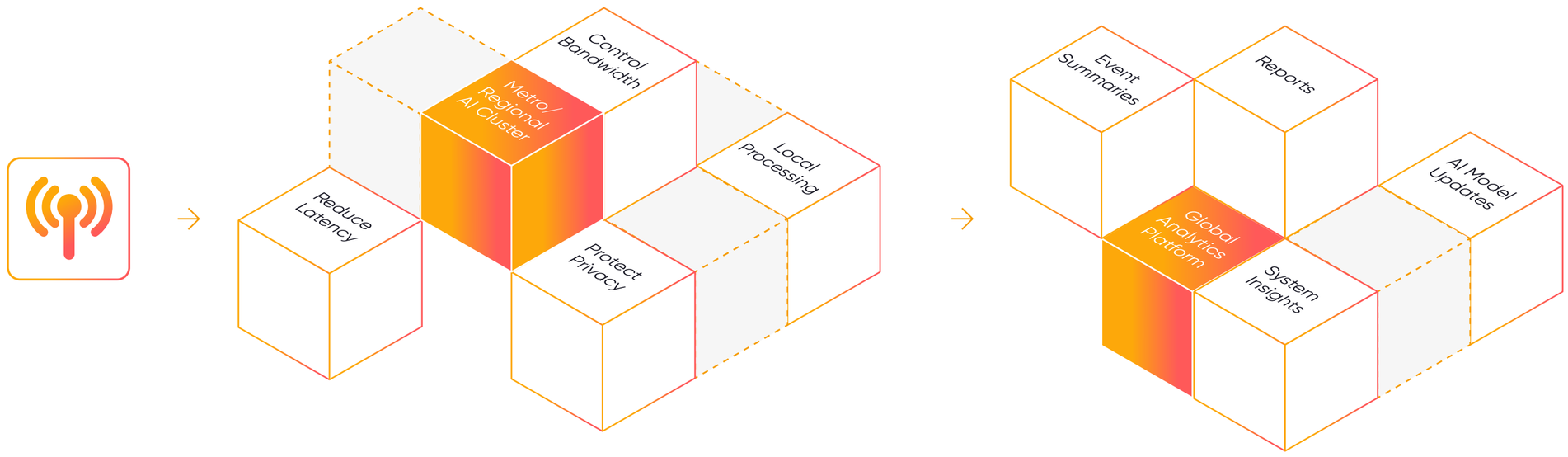
Urban camera networks support traffic monitoring, safety detection, and incident response. Retail vision platforms analyze store activity to detect theft patterns and operational inefficiencies.

In both environments, processing video close to where it is captured is essential. Raw video streams are typically analyzed in metro or regional AI clusters to reduce latency, protect privacy, and control bandwidth costs.

Only derived insights including alerts, metadata, or event summaries are transmitted to broader analytics platforms.

A city deploying 3,000 cameras across its transit network cannot afford to stream raw video to a regional data center for processing. The bandwidth cost alone makes it impractical.

The governance requirements, which may prohibit raw footage from leaving municipal infrastructure entirely, make it architecturally impossible. Instead, inference runs at metro-edge clusters distributed across the city.




## Vision System and Public Infrastructure

The network enforces those boundaries automatically, not by policy document, but by design.

Only structured outputs (vehicle counts, incident flags, dwell-time anomalies) leave those environments.

**Infrastructure Insight**

Distributed AI system often follow a simple rule: raw data stays local, intelligence travels.

# Industrial And Remote Operations

Industrial environments deploy AI across machinery, production lines, and remote operational sites.

Mining operations often run embedded AI models directly on heavy equipment while coordinating analytics through site-level compute clusters. Manufacturing plants deploy plant-edge platforms that monitor equipment performance and production quality across entire facilities.

Agricultural systems combine distributed sensors with regional AI platforms to optimize irrigation and crop management.

What connects them? Many of these environments operate in locations with intermittent connectivity and strict operational technology (OT) controls.

**Infrastructure Insight**

AI infrastructure must support resilient connectivity, local inference environments, and strict segmentation between operational systems and external networks.

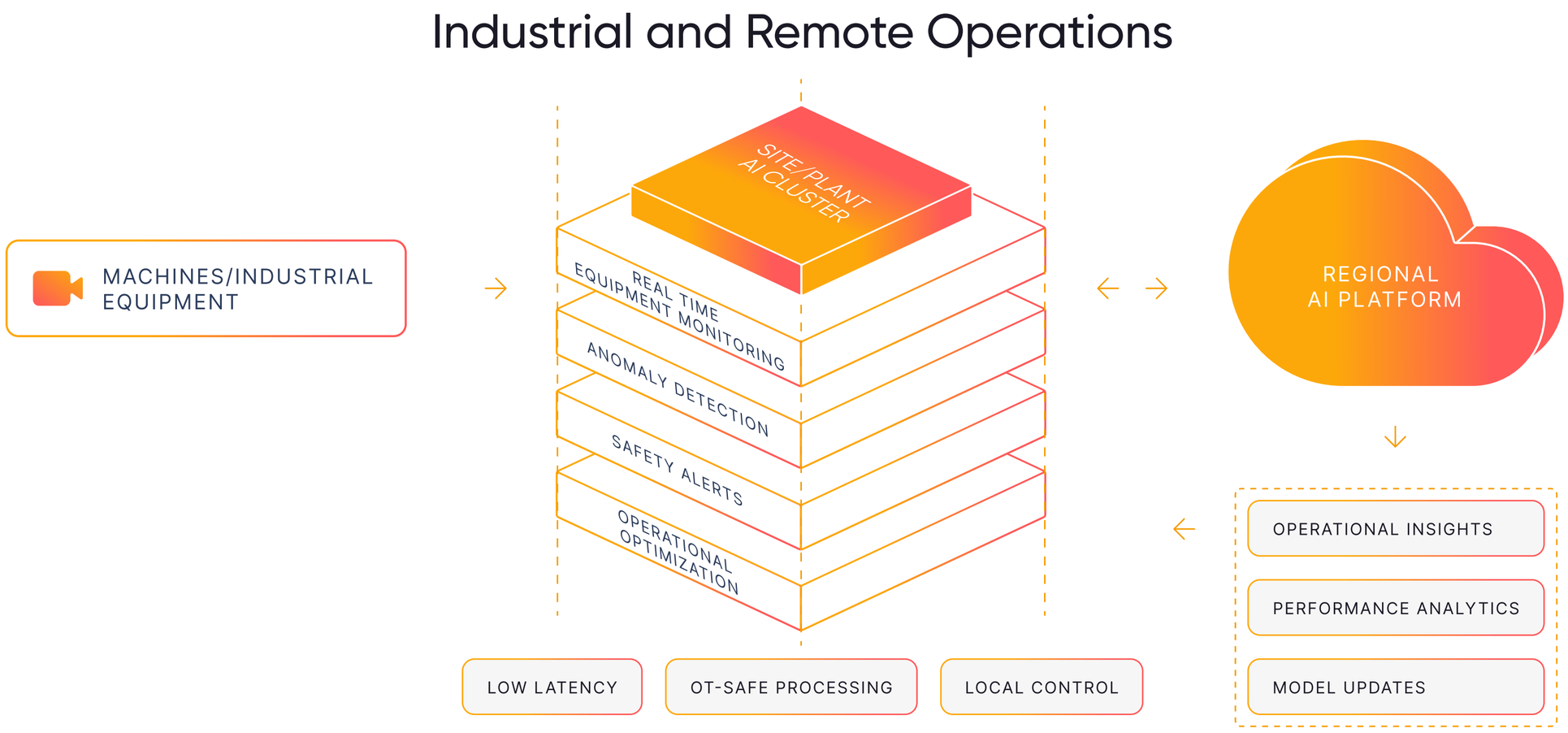
An open-pit mining operation running autonomous haulage across a remote site cannot assume continuous connectivity to a central cloud. The site may span tens of kilometers. Atmospheric conditions, terrain, and network availability are unpredictable. AI models managing route optimization and collision avoidance run directly on the equipment.

A site-level compute cluster aggregates telemetry, monitors the fleet, and coordinates operations. External connectivity is used for model updates and enterprise reporting, not for safety-critical inference.

The architecture was shaped by the environment, not by the technology preference of the team that built it.

**Infrastructure Insight**

Industrial AI runs locally to maintain control, resilience, and operational safety.




# Healthcare, Logistics, And Field Intelligence

Healthcare devices, supply chains, and field service operations all combine mobile endpoints with governance-heavy data environments. Portable diagnostic devices perform on-device inference while connecting to clinical AI platforms for deeper analysis and specialist support. Logistics systems monitor shipments across global supply chains while maintaining traceable, compliant records. Field technicians increasingly rely on AI assistants connected to enterprise knowledge systems to diagnose equipment and guide repairs.

**Infrastructure Insight**

When AI systems operate across distributed and regulated environments, connectivity must enforce governance rules while maintaining operational mobility.

A portable diagnostic device deployed in a rural clinic may perform on-device inference for common presentations, and then, when connectivity allows, transmit structured findings to a regional clinical AI platform for specialist review.

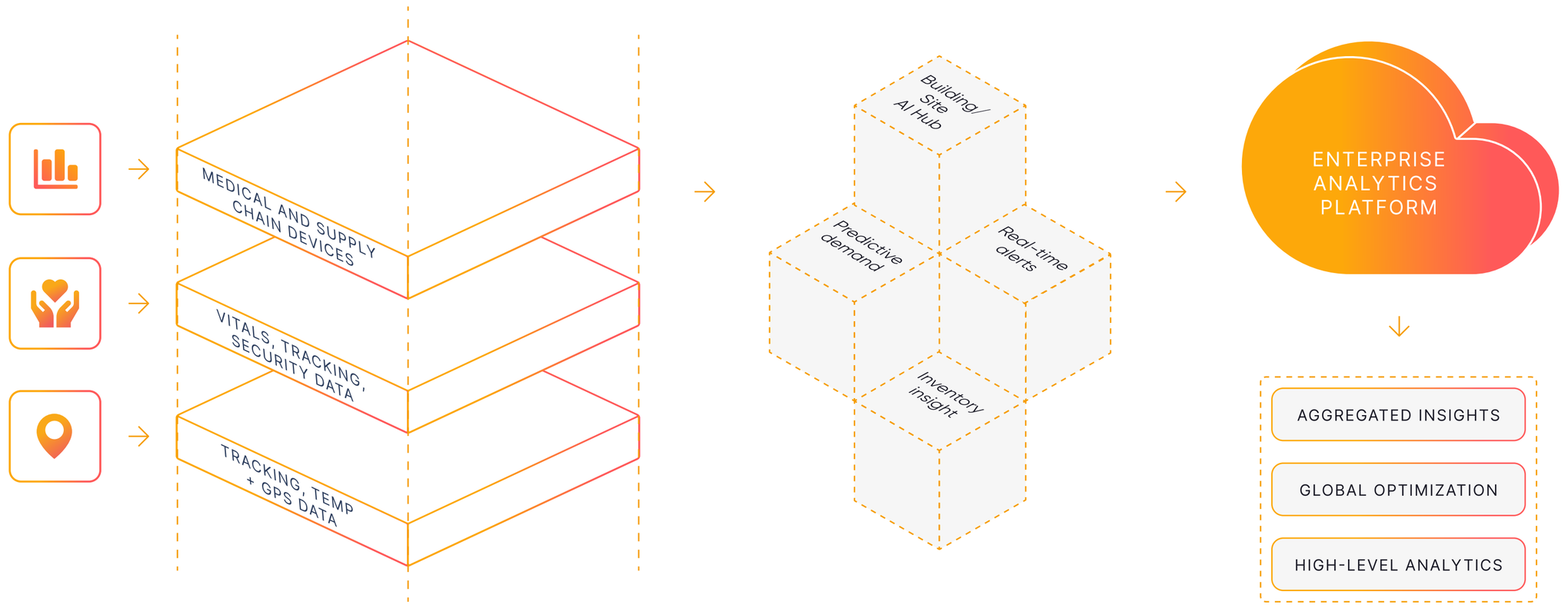
The same device, operating on a different regulatory jurisdiction the following week, may need to route that data through an entirely different compliant infrastructure path.

The connectivity layer must carry not just the data, but the governance rules that determine where the data is allowed to go.

**Infrastructure Insight**

Enforce governance while maintaining operational mobility.

## AI Empowers Healthcare and Supply Chain




# The FLOLIVE® Point Of View

The infrastructure patterns explored throughout this eBook point to one clear conclusion: AI in the physical world requires a different kind of network.

Distributed AI systems do not operate inside a single cloud or within the footprint of a single operator. They span devices, factories, vehicles, cities, and field environments.

They must comply with regional governance rules, operate across multiple networks, and support compute environments that enterprises have already chosen.

To function at scale, these systems require a connectivity layer designed not simply to connect endpoints, but to align with the architecture of distributed intelligence.

This is the perspective behind FLOLIVE®: A Network Beyond, purpose-built for the realities of global, AI-driven infrastructure.

# Keep Compute And Data Where The Customer Needs Them

Flolive is designed to support the operational realities which influence where compute and data must reside, rather than override them. Enterprises remain free to run inference at the device edge, within plant-level clusters, inside regional AI platforms, or within their preferred cloud environments.

Connectivity adapts to that architecture, allowing data to move directly to the environments where intelligence is executed. In addition to enforcing connectivity policies, the network can also provide operational telemetry that helps enterprises detect anomalies, monitor fleets, and feed AI systems with real-world infrastructure signals.

This approach ensures that AI deployments can evolve naturally as infrastructure strategies change, without forcing organizations to redesign their connectivity stack.

Flolive's Data Gravity Facilitation capability enables local breakout directly at the data lake, reducing backhaul costs for high-volume telemetry while respecting the governance constraints that determine where data can reside. The Network Telemetry Feed exposes signal health, geofencing status, and anomaly data as a structured API, giving customer AI models real-world infrastructure signals they can act on.

# Terminate Connectivity Close To The Application

Many traditional connectivity models route traffic through distant hubs or rely on roaming architectures that introduce unpredictable routing paths. For distributed AI systems, that behavior can introduce latency, compliance risks, and operational complexity.

Flolive addresses this by enabling carrier-agnostic termination close to application environments. Traffic can break out near the locations where inference, analytics, or enterprise systems actually operate.

This local behavior allows AI systems to interact directly with regional platforms and edge environments, reducing latency while preserving governance requirements. It’s global connectivity that behaves locally everywhere.

This is the function of Flolive's Near-Inference Termination capability: deploying user plane functions (UPF) wherever the customer's AI workload runs, where that’s in a cloud environment, a regional data center, or a customer-operated edge cluster, so that traffic terminates at the inference environment, not at a distant hub.

# Create One Service Plane Across Operators And Countries

Global AI systems rarely operate within a single network footprint.

Fleets move across borders, supply chains span continents, and distributed infrastructure often connects through multiple operators.

Flolive provides a single service plane across carriers and geographies, allowing enterprises to manage connectivity policies, traffic behavior, and operational controls through one consistent architecture. Differences between networks are abstracted into a unified capability layer, creating predictable behavior regardless of where devices operate.

Flolive's Model-Aware Smart Routing applies this principle at the payload level, routing training data, inference requests, and model updates via the optimal path for each payload type, across a unified service plane that spans operators and geographies. CDN Distribution extends this further, enabling model weights, firmware, and updates to be pushed from the UPF to the entire global fleet simultaneously.

# Make Connectivity Programmable Around Workload Intent

Flolive exposes programmable network controls that allow connectivity behavior to align with the intent of the workload.

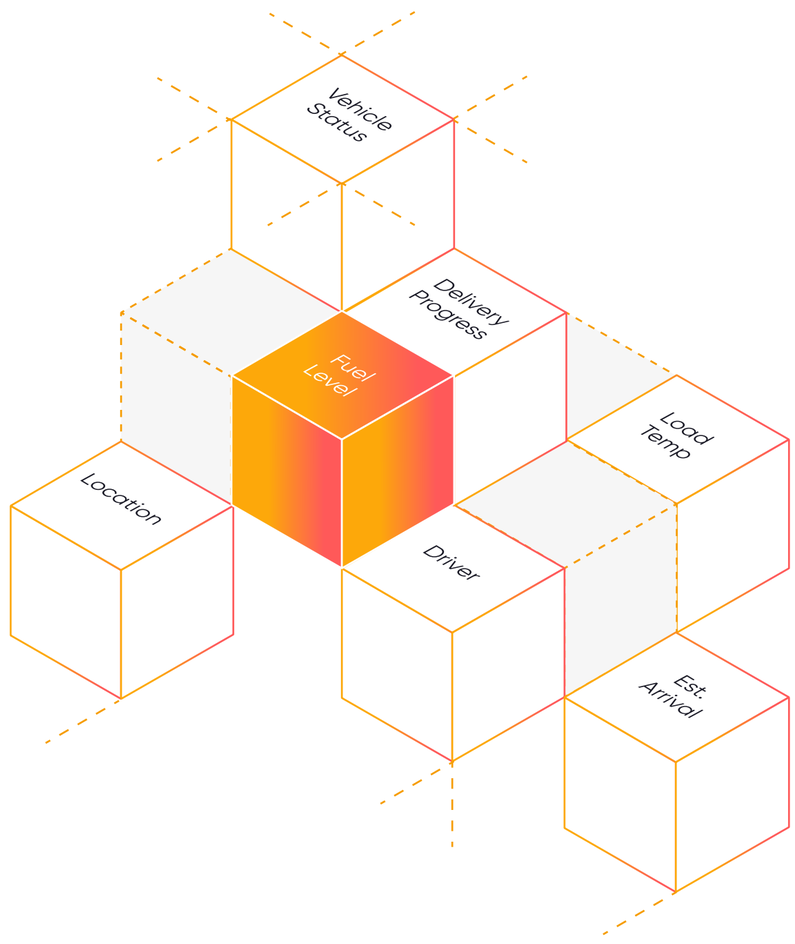
Policies can steer traffic, enforce geographic boundaries, establish secure overlays, or adjust network behavior as conditions change.

That means connectivity can respond to changing operational requirements on the fly, transforming the network from static infrastructure into an active component of the AI system itself.

Through Flolive's Programmable Network capability, AI agents can request network behaviors via API, adjusting coverage parameters, quality of service levels, or secure channel configurations in real time. The network responds to the workload, not the other way around.

Security By Design enforces zero-trust policies at the network layer, protecting models, datasets, and devices in transit without requiring application-level security to carry that burden alone.

Together, these four capabilities reflect a new model for connectivity. One that supports distributed compute, respects data gravity, and adapts to evolving infrastructure strategies across the physical world.




It is the foundation
of A Network Beyond:
A global infrastructure designed to help enterprises deploy and scale AI systems wherever intelligence is needed next.

Learn more