White Paper

Critical Infrastructure
Smart Cities

From Smart Cities to Cognitive Cities

How Agentic AI Enables Autonomous Critical Infrastructure

This paper explores how Agentic AI — deployed on Intel® Core™ Ultra processors — enables cities to evolve from reactive smart systems into self-optimizing, cognitive infrastructure ecosystems capable of autonomous decision-making at the edge.

Authors

Intel

Archit Agarwal
Edge AI Solutions Architect

Ranjan Mishra
Edge AI Solutions Architect

Ebrahim Attarwala
Edge AI Solutions Architect

Hassnaa Moustafa
Principal Engineer

Veena Mahesh
GM, Critical Infrastructure

AIREV

Muhammed Khalid
CEO & Founder

Yame Hatahet
Lead Engineer

Table of Contents

Introduction: From Smart Cities to Cognitive Cities 1
Introduction to OnDemand 2
OnDemand Capability Stack: Architecture, Intelligence, Autonomy, and Outcomes 2
Architecture Overview 3
What Makes It "Cognitive" 3
Smart City Use-Case Narrative 3
Evaluation of AIREV OnDemand on Intel® Core™ Ultra Processors 4
Key Takeaway and Conclusion 6
Appendix — Notes on Memory and Reproducibility 6

From Smart Cities to Cognitive Cities

Cities today operate some of the most intricate and interdependent environments on the planet — from energy grids and water systems to environmental monitoring networks, public-safety operations, climate-control infrastructure, and distributed field assets generating thousands of data points every second. Yet despite their importance, these mission-critical systems often remain siloed, reactive, and heavily reliant on manual intervention.

Over the past decade, the emergence of smart-city technologies introduced the first wave of automation. AI and Edge AI enabled real-time sensing, localized decision-making, and predictive capabilities across transportation, utilities, and public-safety domains. These advancements allowed cities to modernize operations, increase responsiveness, and extract insights from large volumes of data at the network's edge. But even with these innovations, most systems still functioned independently, improving efficiency within their domains without achieving true cross-system intelligence or coordinated action.

Now, the rise of Agentic AI represents a pivotal leap forward. Unlike traditional AI models that focus on analysis or prediction, Agentic AI systems can understand context, reason about complex situations, collaborate with other agents, and take goal-directed actions. For cities and critical-infrastructure operators, this means moving beyond automation toward cognitive cities — urban environments capable of self-orchestration.

In a cognitive city, interconnected AI agents continuously analyze conditions across power, water, environmental, mobility, and safety systems. They negotiate priorities, reconfigure resources dynamically, and autonomously execute operational decisions — while preserving human oversight for governance and policy control. This shift enables truly adaptive infrastructure: power grids that self-balance, water networks that detect and isolate leaks automatically, public-safety systems that anticipate risks, and environmental systems that optimize themselves based on real-time climate conditions.

As Agentic AI becomes embedded throughout critical infrastructure, cities evolve from reactive and operator-driven systems to autonomous, resilient, and self-optimizing ecosystems. This marks the next generation of urban intelligence — where cities not only sense and respond, but understand, decide, and act.

Introduction to OnDemand

As cities have adopted smart-city technologies, AI and Edge AI have delivered important gains in automation and localized intelligence — but they have not fully unified or autonomized the broader urban ecosystem. OnDemand advances this evolution by enabling cities and critical-infrastructure operators to transition from smart environments to cognitive, agent-driven systems. It provides the contextual understanding, cross-domain coordination, and autonomous decision-making required for infrastructure to continuously optimize itself, anticipate disruptions, and operate with far greater resilience.

OnDemand is an advanced Agentic AI platform designed to transform urban and critical infrastructure into cognitive systems. It enables cities to monitor infrastructure in real time, interact with assets through natural language, automate decision-making, predict failures before they happen, and execute autonomous safety protocols. In short, OnDemand transforms infrastructure from passive systems into intelligent, conversational, self-optimizing environments.

OnDemand Capability Stack

IoT Architecture

Behind OnDemand's capabilities is a robust, enterprise-grade IoT architecture built on MQTT publish–subscribe protocols. This infrastructure enables real-time telemetry from air quality, water quality, energy, temperature, and humidity sensors while supporting instant command execution to devices and actuators. With full historical data persistence and bidirectional communication, the platform forms a live digital nervous system for the city.

Analytics and Cognition

OnDemand extends beyond monitoring into cognition by continuously analyzing energy consumption, generation patterns, device runtime, and sensor trends such as CO_2, ammonia, TDS, and voltage levels. It generates insights including energy-usage analytics, sensor-trend reports, and runtime comparisons. This historical intelligence drives a shift from reactive operations to predictive and proactive intelligence — detecting anomalies, scheduling maintenance optimally, generating real-time alerts, and preventing failures before they occur.

Autonomous Agentic AI

At this stage, OnDemand becomes fully agentic. AI agents execute conditional logic autonomously — for example: shutting down a pump when water quality drops, disabling streetlights when solar generation falls below thresholds, or disconnecting power and sending alerts when hazardous acetone levels are detected. These safety and operational responses occur instantly, without human intervention.

Conversational Infrastructure Control

OnDemand introduces conversational infrastructure control. Instead of operators navigating dashboards or technical systems, they can simply speak or type natural-language commands such as "Turn on the streetlights," "Check CO_2 levels," or "Is the water pump running?" AI agents interpret intent and execute across connected systems, making municipal operations intuitive and frictionless.

Predictive Maintenance and Proactive Capabilities

OnDemand delivers proactive, intelligent, and measurable value across city infrastructure through continuous analytics that enable predictive maintenance — identifying anomalies and forecasting potential failures to reduce downtime and extend asset lifespan. Proactive capabilities include real-time alerts, optimal maintenance-window scheduling, and preventive control actions. The platform's autonomous decision-making allows AI agents to automatically execute safety protocols, such as shutting down pumps when water quality drops or disabling streetlights during low solar output, to protect both infrastructure and residents.

Operational Outcomes

OnDemand delivers measurable outcomes across efficiency, cost, and risk. Cities reduce manual oversight, accelerate response times, and streamline workflows. Autonomous maintenance and optimized energy use lower operational costs, while automated safety protocols and hazard detection mitigate risk. The system scales seamlessly — from single buildings to full metropolitan deployments — with low-code integration through MQTT, HTTP, and MCP.

Capability	Description
Conversational Interface	Natural-language operations: "Turn on the streetlights," "Check CO_2 levels."
Analytics & Cognition	Continuous analysis of energy, device runtime, and sensor trends; anomaly detection and predictive maintenance.
Autonomous Agentic AI	Goal-driven agents executing conditional logic autonomously — e.g., shut pump if water quality degrades.
IoT Telemetry & Control	Enterprise-grade MQTT publish–subscribe core; real-time telemetry across air, water, energy, temperature, and humidity.

Figure 1. OnDemand Capability Stack — from telemetry to autonomy.

Architecture Overview

OnDemand can be described through a simple four-layer narrative that maps directly to how cognitive infrastructure is built and operated: **Autonomous Agentic AI**

Layer	Description
01 — Physical Infrastructure	Houses, lighting, wind turbines, solar panels, water pumps, environmental sensors (CO_2, CO, ammonia, TDS).
02 — Real-Time Connectivity	MQTT publish–subscribe backbone with bidirectional control and persistent telemetry storage.
03 — Agentic AI Core	Natural-language processing, tool orchestration, predictive analytics, conditional automation logic, and multi-tool execution (email, SMS, reporting).
04 — Human Interaction	Conversational commands, automated reports, role-based dashboards, alerts and notifications.

Figure 2. OnDemand cognitive architecture for urban infrastructure — a four-layer model.

What Makes It "Cognitive"

A traditional smart-city system typically collects data, displays dashboards, and requires humans to make decisions. OnDemand differs by understanding intent, executing commands, learning from historical patterns, detecting anomalies, taking preventive action, and orchestrating multiple tools automatically. It moves from monitoring to analytics to prediction — and finally to autonomous action.

Smart City Use-Case Narrative

Imagine a district where energy usage is optimized daily, water quality is continuously protected, air quality is monitored in real time, and if a hazard appears the system automatically shuts down affected systems and notifies residents within seconds. That is OnDemand.

OnDemand supports:

- **Environmental monitoring** — real-time tracking of CO_2, CO, ammonia, toluene, acetone, and alcohol
- **Energy optimization** — solar-generation tracking, device-runtime analytics, and intelligent load balancing
- **Water management** — TDS-based quality monitoring with automated pump control
- **Climate control** — temperature- and humidity-driven HVAC optimization

Evaluation on Intel® Core™ Ultra Processors

This section presents the evaluation of AIREV's OnDemand Agentic AI platform on Intel® Core™ Ultra Series 3 (code-named Panther Lake, "PTL"), using real, chat-style natural-language prompts. The evaluation focuses on interacting with assets, automating decision-making, predicting failures before they occur, and executing autonomous safety protocols. Performance is assessed through interactive LLM inference across both GPU and NPU execution paths, using industry-standard KPIs including Time to First Token (TTFT), Inter-Token Latency (ITL), Tokens Per Second (TPS), and end-to-end (E2E) latency.

System Configuration

The evaluation was conducted on a single-node Dell XPS 16 (DA16260) system powered by an Intel® Core™ Ultra X7 358H processor featuring 16 cores and a 200 W TDP. The system is equipped with 32 GB of LPDDR5 memory operating at 9600 MT/s and a 1 TB KIOXIA NVMe SSD. Networking included Gigabit Ethernet and integrated Panther Lake PCH Wi-Fi.

All tests were performed on Ubuntu 24.04.4 LTS using a consistent BIOS, firmware, operating system, driver stack, and OpenVINO™ runtime configuration to ensure result reproducibility.

Methodology

Imagine a district where energy usage is optimized daily, water quality is continuously protected, air quality is monitored in real time, and if a hazard appears the system automatically shuts down affected systems and notifies residents within seconds. That is OnDemand.

The OnDemand platform was evaluated using the Qwen3 8B language model with INT4 quantization. Qwen3 8B was selected for this study because it offers a strong balance of reasoning quality and inference efficiency for the target deployment profile: it is open-weight, well-supported in the OpenVINO™ runtime, demonstrates competitive tool-calling accuracy in published benchmarks among models in its size class, and runs comfortably within the memory envelope of edge-class devices when quantized to INT4.

Two accelerator-specific model variants were deployed to optimize performance on Intel® Core™ Ultra edge hardware:

GPU-optimized model — sourced from OpenVINO/Qwen3-8B-int4-ov

NPU-optimized model — sourced from OpenVINO/Qwen3-8B-int4-cw-ov

Performance testing consisted of approximately 90 sequential, independent chat requests executed on both the GPU and NPU. Each request produced an average of ~50 output tokens, reflecting realistic interaction patterns for the target deployment scenarios.

Obtained Results

Key Performance Indicator (KPI)	GPU	NPU
Effective Throughput (tokens/sec)	20.03	16.33
Time to First Token (TTFT)	**125.26 ms**	1,213.98 ms
Decode Throughput (tokens/sec)	24.21	31.02
Reasoning Time	13.48 s	17.57 s
Fulfillment Time	1.94 s	3.06 s
End-to-End Latency (E2E)	15.43 s	20.61 s
Active Decode Time	1,821.62 ms	1,844.84 ms
Inter-Token Latency (ITL)	41.3 ms/tok	**32.2 ms/tok**
Decode Overhead Ratio	6.4%	39.7%
Time to 10 Tokens	538 ms	1,536 ms
Requests per Second (RPS)	0.065	0.049

Note 1: Results were measured and estimated based on testing performed within Intel's internal laboratory environment.

Note 2: All testing is conducted using sequential requests. Concurrent workloads or multi-accelerator configurations are outside the scope of this study.

Figure 3. GPU vs. NPU KPI comparison on Intel® Core™ Ultra Series 3.

KPI Definitions

KPI	Definition
Effective Throughput	Average token generation rate observed across the entire fulfilment period, including TTFT.
Time to First Token (TTFT)	Time from generation initiation to the first output token; captures model initialization, graph setup, memory preparation, and scheduling overhead.
Decode Throughput	Rate at which output tokens are generated during active decoding, excluding TTFT.
Reasoning Time	Time spent in pre-generation orchestration, including planning, retrieval, and tool execution.
Fulfillment Time	Duration of the generation phase, from start of decoding to the final emitted token.
End-to-End Latency	Sum of reasoning time and fulfilment time.
Active Decode Time	Portion of fulfilment time spent generating tokens after the first token is produced. (Fulfilment Time – TTFT)
Inter-Token Latency (ITL)	Average time between consecutive output tokens once decoding has started.
Decode Overhead Ratio	Fraction of fulfilment time spent before the first token is emitted. (TTFT ÷ Fulfilment Time)
Time to 10 Tokens	Time required to generate the first ten output tokens, combining TTFT and early decode speed.
Requests per Second (RPS)	Number of complete requests that can be processed per second.

Key Takeaway and Conclusion

The AIREV OnDemand platform, enabled on Intel® Core™ Ultra Series 3 processors with integrated accelerators, demonstrates performance well aligned with edge-deployment requirements, as validated through KPI-based evaluation.

For chat-style, tool-augmented workloads with short to moderate response lengths, the GPU and NPU on Core Ultra exhibit complementary performance characteristics. Together, they enable strong and consistent interactive performance under CPU orchestration.

The GPU delivers significantly lower Time to First Token (TTFT) and faster time to initial output (e.g., the first 10 tokens), resulting in markedly improved perceived responsiveness for interactive queries. In contrast, while the NPU incurs higher initialization latency, it provides highly stable and low-variance inter-token latency during the decoding phase, demonstrating strong determinism and smooth, consistent token streaming.

By combining fast startup on the GPU with stable token generation on the NPU, Core Ultra delivers balanced and reliable chat inference suitable for real-world edge AI workloads.

Appendix — Notes on Memory and Reproducibility

Higher memory capacity and memory bandwidth are associated with measurable performance benefits in the evaluated workload, highlighting the importance of memory subsystems in large-language-model execution — particularly during sustained decoding phases. The evaluation used a fixed 32 GB LPDDR5 configuration at 9600 MT/s; systems with greater DDR5 capacity or bandwidth are expected to deliver improved throughput and latency characteristics on equivalent workloads.

All measurements were taken under controlled conditions using a consistent BIOS, firmware, OS, driver stack, and OpenVINO™ runtime to ensure reproducibility. Sequential request testing isolates per-request behavior; concurrent and multi-accelerator scenarios remain outside the scope of this study and represent natural extensions for future benchmarking.

Actual results may vary based on workload, system configuration, and other factors.

intel. Airev.

[1] Performance results reflect a fixed memory configuration at 9600 MT/s; systems configured with higher DDR5 memory capacity may achieve improved performance.

[2] All testing is conducted using sequential requests. Concurrent workloads or accelerators (GPU/NPU) are outside the scope of this study.