SoftServe

Edge AI Kit

- RETAIL
- AR/XR WEARABLES
- HEALTHCARE
- EMBEDDED AND INDUSTRIAL HARDWARE
- MANUFACTURING
- LOGISTICS
- CONSUMER ELECTRONICS
- SUB-300MS LATENCY
- HYBRID INFERENCE
- VALIDATED REFERENCEPIPELINES

Building a real-time AI system that runs on a device is not the same problem as building one that runs in the cloud. The hardware is constrained. The latency requirement is strict. The architecture has to decide, for every inference, whether the device handles it or the cloud does. Edge AI Kit is the reusable IP layer that makes those decisions production-ready from the start, rather than rebuilt from scratch for every new use case.

WHY REAL-TIME AI ON DEVICE IS HARDER THAN IT LOOKS

A retail team has AR glasses on the shelf floor. A shopper asks for allergen information on a product. The cloud call takes 800 milliseconds. The shopper has already moved on. The use case works in a demo and fails in a store. That is not a model problem. It is an architecture problem. The team built a cloud AI system and put it inside a wearable. The latency budget, the connectivity assumptions, and the hardware constraints of the device were never part of the design.

CLOUD AI DOES NOT FIT EDGE LATENCY BUDGETS

Real-time AI use cases require responses under 300 milliseconds. Cloud-only architectures introduce latency from network round-trips that makes interactive AI on devices unreliable, particularly in environments with inconsistent connectivity.

EACH DEPLOYMENT IS REBUILT FROM SCRATCH

There is no standardized, reusable way to design hybrid real-time AI architectures. Teams designing for AR glasses today rebuild the same pipeline decisions for handheld scanners tomorrow and industrial devices the month after. Every engagement starts from zero.

HARDWARE SELECTION IS GUESSWORK WITHOUT VALIDATED PATTERNS

Choosing between on-device inference, near-cloud compute, and full cloud processing requires validated knowledge of which models run on which chipsets at what latency. Without reference stacks, teams make architecture decisions under uncertainty and discover the limits in production.

FLEET MANAGEMENT AND ROLLOUT ARE NOT AFTERTHOUGHTS

Deploying an AI system to one device is a pilot. Deploying it to thousands of devices across multiple sites introduces fleet configuration, monitoring, model update logistics, and session continuity requirements that a single-device architecture does not address.

WHAT EDGE AI KIT IS

Edge AI Kit is a reusable IP layer that provides production-ready hybrid AI pipelines for real-time AI systems running on edge devices.

The kit is 70% pre-built. That includes pipeline structure, building blocks, and orchestration patterns. The remaining 30% is implementation work adapted to the specific use case, device, dataset, and integration requirements. The ratio means teams start from production-grade architecture, not from a blank canvas.

The kit supports three deployment modes. On-device delivers lowest latency and maximum data privacy, near-cloud via AWS Local Zones delivers consistent response times at the compute level, and hybrid orchestration routes cases where latency-critical steps run on device and heavier inference runs in near-cloud or cloud. The routing decision is built into the pipeline, not left to the development team to figure out per deployment.

BUSINESS IMPACT

<300ms

response latency for GenAI models achieved with hybrid inference

<200ms

response time demonstrated in the AR store assistant deployment on Snap Spectacles

10x

reduction in data transfer to the cloud versus cloud-only AI architectures

70%

of the kit is pre-built, reducing architecture design and delivery risk per engagement

*Latency figures are from SoftServe benchmarks and PoC deployments on Qualcomm Snapdragon AR1 Gen 1 hardware with AWS Local Zones.

HOW IT WORKS

FIVE PROPRIETARY IP LAYERS

On-device SLM orchestrator manages local context, routes requests, and coordinates on-device execution with offline and hybrid fallback logic. Model selection framework provides validated building blocks for ASR, TTS, vision, SLMs, and LLMs with performance constraints per hardware class, with reference stacks for AR/XR, handheld, and embedded deployments.

Model optimization pipeline covers fine-tuning, quantization, compression, and distillation with performance and latency optimization for constrained edge devices. Deployment strategy framework provides expert methodology for choosing edge, hybrid, or cloud topology based on latency, hardware, and connectivity requirements. Horizontal scaling blueprint covers fleet-level deployment, load balancing, session and context handling, and rollout patterns.

PRE-DEFINED MULTIMODAL PIPELINES

The kit includes a set of pre-defined multimodal pipelines connecting vision (detection, tracking, OCR, segmentation), speech (ASR, TTS), on-device SLMs, and cloud LLMs with similarity and retrieval when needed. Each pipeline implements a concrete task flow: see, detect, understand, and respond.

In the AR store assistant deployment, this means the glasses handle object recognition, promo detection, and wake-word activation on device, while the AWS Local Zone handles LLM inference for personalized recommendations, shopping list generation, and contextual data. The data flow minimizes what crosses the network. Key operations run locally. Complex reasoning runs where the compute supports it.

DEPLOYMENT PATTERNS

Edge AI Kit supports three deployment modes. The appropriate pattern is selected during the discovery and alignment phase based on latency requirements, hardware capability, and connectivity profile.

ON-DEVICE

LOWEST LATENCY. MAXIMUM DATA PRIVACY.

All inference runs directly on the device. No network dependency for AI operations. Appropriate for use cases with strict latency requirements, offline environments, or data residency constraints. Requires model optimization for constrained compute and power.

NEAR-CLOUD (AWS LOCAL ZONES)

CONSISTENT RESPONSE TIMES. FULL CLOUD COMPUTE.

Compute runs in AWS Local Zones geographically close to the deployment site. Delivers consistent sub-300ms response times for inference workloads too large for on-device execution. Appropriate for use cases requiring LLM-scale reasoning with bounded latency.

HYBRID

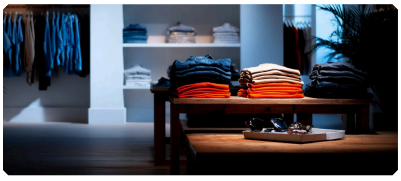
LATENCY-CRITICAL ON DEVICE. HEAVY STEPS IN CLOUD.

Latency-critical operations run on device. Heavier inference, retrieval, and content generation run in near-cloud or cloud. Routing is built into the orchestration layer. The AR store assistant deployment uses this pattern: navigation and recognition on device, personalized recommendations in AWS Local Zones.

DEPLOYMENT PATTERNS

COMPONENT	WHAT IT DOES
ON-DEVICE SLM ORCHESTRATOR	Runtime that manages local context, routes requests, and coordinates on-device execution. Handles offline and hybrid operation modes and fallback logic when cloud connectivity is unavailable or latency budgets are exceeded. Runs on Qualcomm inference-native chipsets.
MODEL SELECTION FRAMEWORK	Validated building blocks for ASR, TTS, vision, SLM, and LLM with performance constraints per hardware class. Reference stacks optimized for AR/XR, handheld, and embedded deployments. Takes model selection from guesswork to validated patterns.
DEPLOYMENT STRATEGY FRAMEWORK	Expert methodology for choosing edge, hybrid, or cloud topology based on latency requirements, hardware capability, and connectivity profile. Includes patterns for switching between modes and maintaining real-time user experience during transitions.
HORIZONTAL SCALING BLUEPRINT	Reference architecture for scaling across device fleets and user bases. Covers load balancing, session and context handling, monitoring, and rollout patterns. Addresses the architecture distance between a single-device pilot and a production fleet deployment.
MULTI-MODAL PIPELINES	Pre-defined pipelines connecting vision, speech, on-device SLMs, and cloud LLMs in concrete task flows. Pipelines are pre-built to the see-detect-understand-respond pattern and adapted to the specific use case during the 30% implementation phase.
FLEET OPERATIONS AND LIFECYCLE MANAGEMENT	Device license covers on-device inference, device configuration and rollout, fleet-level monitoring, and lifecycle operations including model updates and hardware compatibility maintenance. SLA-based incident handling with continuous performance and latency optimization.
CLOUD AND HARDWARE INTEGRATION	Integration with AWS Local Zones for near-cloud inference and Qualcomm Snapdragon chipsets for on-device inference. Cloud consumption runs in client-owned cloud infrastructure. The kit is hardware-aware and tested against specific chipset and platform configurations.

SELECTED USE CASES



In-Store AR Shopping Assistant - On AR glasses powered by Qualcomm Snapdragon AR1 Gen 1, the Store Assistant delivers voice activation, product recognition, allergen alerts, aisle navigation, and checkout guidance. On-device SLMs handle intent recognition and navigation. AWS Local Zones handle LLM inference for personalized recommendations and shopping list generation. Responses run under 200 milliseconds. Data transfer to the cloud falls by up to 10x. The use case was built for a leading global AR/VR technology company as part of an ongoing SoftServe engagement.

Industrial and Manufacturing Operator Support - Operators on production floors and field sites receive AI-assisted workflow guidance, inspection support, and compliance checks through wearables and handheld devices. On-device processing handles recognition and alert logic. Near-cloud compute handles reasoning workloads. The system operates in environments where cloud round-trips are not a viable architecture assumption.

Retail Operations and In-Store Assistance -Store associates and kiosk systems query product information, inventory status, and compliance data in real time. On-device inference handles recognition and lookup. Cloud inference handles recommendation and contextual reasoning. The architecture runs in stores where network connectivity is variable and latency requirements are strict.

Healthcare and Medical Device AI - The kit's cross-industry architecture adapts from consumer wearables to medical device deployments. On-device inference handles local data processing and visualization. Hybrid orchestration manages privacy-sensitive data flows. The model optimization pipeline produces models sized for medical-grade hardware with strict power and latency constraints.

Logistics and Smart Navigation - Handheld scanners and AR-equipped logistics workers receive real-time guidance for picking, routing, and exception handling. Edge AI Kit's hybrid orchestration routes recognition tasks on device and planning tasks to near-cloud compute. Fleet-level deployment patterns cover rollout across warehouse sites and device types.

ENGAGEMENT MODEL

DISCOVERY AND ALIGNMENT

Assess devices, backend systems, and workflows. Define the target use case and KPIs. Define the device versus cloud execution split. Select applicable Edge AI Kit pipelines. Define pilot scope and success criteria. Output: deployment blueprint based on Edge AI Kit IP.

PRODUCTION ROLLOUT

Roll out across target devices and sites. Production configuration hardening. Monitoring and observability setup. Finalize licensing and subscription scope. Knowledge transfer to client teams. Output: production-grade deployment running on Edge AI Kit.

SUPPORT AND CONTINUOUS OPTIMIZATION

SLA-based incident handling, operational monitoring, and cost control. Model and pipeline updates. Performance and latency optimization. Hardware and platform compatibility updates. Output: stable, continuously optimized hybrid AI system.

PILOT IMPLEMENTATION

Deploy runtime on selected devices. Configure hybrid routing across device and cloud. Integrate with backend systems. Run a limited pilot across a defined set of users or locations. Measure latency, stability, and cloud cost. Output: running pilot with measured KPIs and production-ready insights.

WHY SOFTSERVE

- **Cross-industry architecture.** The same kit adapts from consumer AR wearables to industrial handhelds to medical devices. The 70/30 split between pre-built IP and implementation work means the architecture is consistent across deployments while the use case remains specific to the client context.
- **Reusable IP, not a one-off build.** Edge AI Kit provides pre-built pipelines, validated hardware patterns, and production-ready orchestration. Every engagement starts from architecture that has already been tested in production, not from first principles.
- **Hardware-native expertise.** SoftServe has tested Edge AI Kit on Qualcomm Snapdragon AR1 Gen 1 and AWS Local Zones configurations used in live deployments. Hardware selection, model optimization, and latency validation are built into the methodology, not handled as ad hoc decisions.
- **End-to-end delivery.** AI, ML, on-device engineering, cloud integration, and fleet operations under one team. The engagement model covers discovery through continuous optimization in a single subscription. No separate vendors for hardware, AI, and operations.