Real-Time Embedded AI for Deterministic, Adaptive, and Sustainable Robotics

Unifying deterministic control with adaptive intelligence for next-generation autonomous systems

SPEAKER

Rasmi Nayak

Senior Software Engineer 1

ASML

Specializing in real-time embedded systems, AI optimization, and robotics architecture for mission-critical applications. Focused on bridging deterministic control systems with adaptive machine learning capabilities in resource-constrained environments.

Conf42 DevOps 2026

CHALLENGE

The Embedded AI Paradox

Limited Compute

Constrained processing power in embedded platforms restricts complex AI operations

This often necessitates specialized hardware accelerators or highly optimized algorithms to achieve desired performance.

Tight Memory Budgets

Small memory footprints demand aggressive model optimization and efficient data structures

This also limits the size and complexity of AI models that can be deployed on device.

Millisecond Latency

Real-time requirements leave no room for inference delays or scheduling jitter

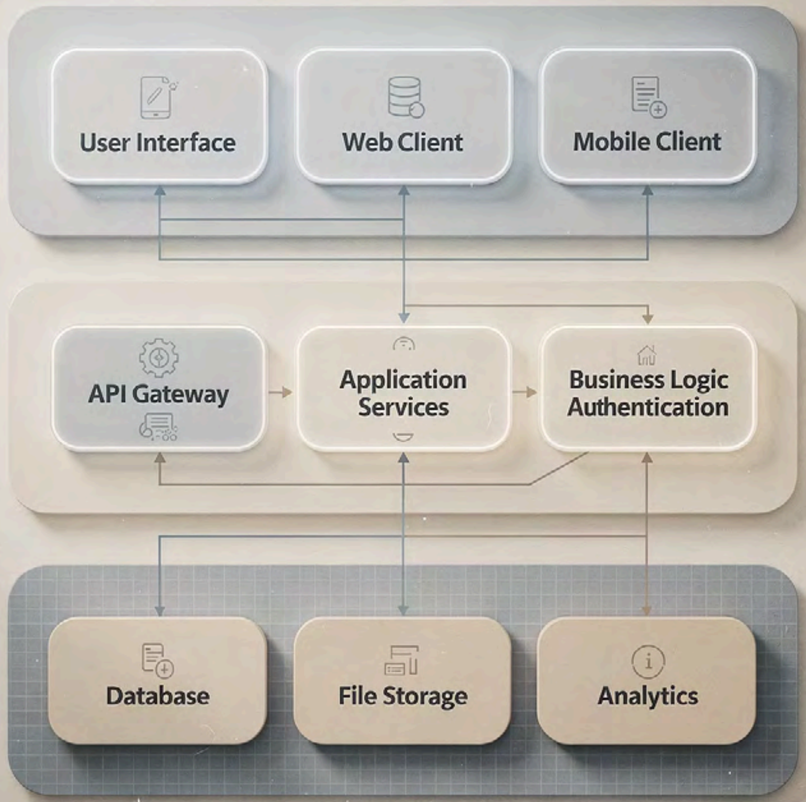
Ensuring predictable execution times is crucial for safety-critical applications and responsive user experiences.

Introducing RE-AIF

Real-Time Embedded AI Framework

A three-layer architecture unifying deterministic real-time control with adaptive AI inference for resource-constrained platforms. RE-AIF eliminates the traditional trade-off between intelligent behavior and timing guarantees.

RE-AIF delivers intelligent functionality to embedded systems without compromising reliability or responsiveness, by effectively orchestrating complex AI models within severe computational and memory constraints. It ensures real-time operations maintain critical safety and performance deadlines, preventing AI processing delays from leading to operational failures.



ARCHITECTURE

Hierarchical Design Philosophy

Perception Layer

Sensor fusion with deterministic data acquisition and preprocessing pipelines

Cognition Layer

Embedded-optimized AI inference with quantized models and hardware acceleration

Execution Layer

Jitter-free actuator control with compile-time scheduling guarantees

The Perception–Cognition–Execution design synchronizes all system components while maintaining hard real-time constraints throughout the control loop.

Hybrid Architecture Advantage

C++ Core

- Real-time control loops
- Sensor fusion pipelines
- Deterministic scheduling
- Zero-copy memory management
- Low-latency data processing
- Direct hardware interaction

Python Integration

- AI model development
- Rapid prototyping
- Data analysis tools
- Training pipelines
- Advanced analytics and visualization
- Simplified deployment of ML models

Optimized binding interfaces minimize cross-language overhead while enabling capabilities unattainable in single-language systems. The hybrid approach delivers both performance and productivity.

OPTIMIZATION

Embedded Intelligence Stack

1

Quantized CNNs

8-bit and 16-bit integer inference reduces memory footprint by 4× while maintaining accuracy within acceptable thresholds

2

Hardware Acceleration

ARM NEON SIMD instructions and GPU units leverage parallel processing for 10× inference speedup

3

Compile-Time Scheduling

Static analysis eliminates runtime overhead and guarantees deterministic execution paths

4

Lock-Free Queues

Wait-free message passing between layers removes synchronization bottlenecks and priority inversion

Energy-Aware Operation

Dynamic Voltage Scaling

Adaptive frequency adjustment based on workload reduces power consumption by up to 40% during low-demand periods

Hibernation Strategies

Intelligent subsystem shutdown during idle states extends battery life without compromising wake-up latency

Memory Pooling

Pre-allocated buffers eliminate dynamic allocation overhead and fragmentation, reducing energy per transaction

These mechanisms extend operational endurance without sacrificing timing guarantees, critical for field-deployed autonomous systems.

INDUSTRIAL

Manufacturing Applications

Precision Assembly

Vision-guided pick-and-place with sub-millimeter accuracy and 99.7% success rate in component placement. This capability minimizes manual errors and speeds up the production process significantly.

Inspection Throughput

Real-time defect detection increased throughput by 35% while reducing false positives to under 2%. This ensures higher product quality and reduces waste from flawed units.

Reliability Gains

Predictive maintenance reduced unplanned downtime by 60% through continuous system health monitoring. This proactive approach optimizes operational efficiency and extends equipment lifespan.

DEFENSE

Mission-Critical Deployment

Autonomous Navigation

GPS-denied environments present unique challenges for robotic systems. RE-AIF enables reliable localization through sensor fusion combining IMU, LIDAR, and visual odometry with millisecond update rates.

Swarm Coordination

Mission-adaptive behavior emerges from distributed decision-making across multiple agents. Each unit maintains local autonomy while coordinating through bandwidth-constrained communication channels.

Defense evaluations demonstrated robust performance in contested scenarios with dynamic obstacle avoidance and collaborative target tracking.

Implementation Deep Dive

From Theory to Practice

C++ Sensor Fusion

Deterministic loops at 1kHz with EDF scheduling

Multi-threaded data acquisition for concurrent sensor input

Adaptive Kalman filtering for state estimation

TensorFlow Lite

Quantized model deployment with 8ms inference latency

Model optimization techniques like pruning and quantization

Cross-platform compatibility for various embedded targets

Actuator Control

Jitter-free execution with <50μs variance

Robust PID control loops for precise motion

Hardware-level safety interlocks for fault tolerance

FUTURE

Next-Generation Capabilities

01

Federated Learning

Collaborative model improvement across robot fleets without centralizing sensitive data

02

Neuromorphic Hardware

Event-driven processing for ultra-low-power inference with spiking neural networks

03

Formal Verification

Mathematical proofs of safety properties for certification in regulated domains

04

Edge-Cloud Hybrid

Intelligent workload distribution between local inference and cloud-based model updates

Key Takeaways

Unified Architecture

RE-AIF eliminates the false dichotomy between deterministic control and adaptive AI, proving both can coexist in resource-constrained platforms

Embedded Optimization

Quantization, hardware acceleration, and lock-free design patterns enable real-time AI inference with millisecond latency guarantees

Sustainable Performance

Energy-aware mechanisms extend operational endurance while maintaining timing requirements critical for field deployment

Production-Ready Blueprint

Validated across industrial and defense applications, RE-AIF provides a practical framework for next-generation autonomous systems

Thank You!

Rasmi Nayak

Senior Software Engineer 1, ASML

Conf42 DevOps 2026.