

MoM Gen EdgeAI WG meeting on 10/23/2026

Attendees: Danilo Pau

Objective: organize the next Gen Edge AI workshop on Dec 3rd 2026

Contributions received:

Workshop 1

[Tara Khani Moorcheh, Ashutosh Kumar Intel]

Topic: From Millions of Records to Live Video: Deploying Moorcheh Highly Compressed Search Engine and Agentic Memory on Intel Core Ultra for AMRs

Abstract

Can an autonomous robot execute real-time Vision-Language-Action (VLA) policies and continuously recall past actions, visual keyframes, and part schematics—entirely disconnected from the cloud? The Intel® Distribution of OpenVINO™ Toolkit delivers high-performance acceleration for embodied policies like 0 and Action Chunking with Transformers (ACT) directly across Intel® Core™ Ultra NPUs and Arc™ iGPUs. To extend these physical AI systems to complex, long-horizon tasks, robots require deep situational history alongside fast policy inference. Conventional vector databases trigger an edge memory wall by consuming upwards of 7 GB of RAM and locking up CPU threads during graph re-indexing. Overcoming this requires Moorcheh running on the host CPU—a highly compressed search engine and agentic memory layer that scales to millions of multimodal records (video keyframes, sensor logs, images, and schematics). Moorcheh executes deterministic bit-scans directly on x86 cores in under 15 ms, maintaining massive multimodal recall without competing with the NPU or GPU for neural compute.

In this technical webinar hosted by the Edge AI Foundation, systems engineers from Intel, Robotics Company, and Moorcheh demonstrate an air-gapped, zero-cloud architecture for physical AI:

Heterogeneous Compute Split: OpenVINO accelerates continuous VLA policy inference on the Intel Core Ultra NPU and iGPU, leaving host x86 CPU cores dedicated to deterministic motion control and Moorcheh retrieval.

Millions of Multimodal Records at Sub-Gigabyte Scale: Replace bulky graph indexes with Moorcheh's Information Binarization to scale to millions of multimodal searches in under 15 ms, consuming only 737 MB of process RAM on the CPU with zero index-build delays.

Embodied Reference Platform: Review live manipulation and spatial recall running directly on Robotics Company mobile manipulation hardware, proving sub-millisecond servo determinism alongside multi-month visual memory.

References & further reading

Moorcheh

- arXiv paper 1: [*From HNSW to Information-Theoretic Binarization: Rethinking the Architecture of Scalable Vector Search* - MIB and ITS technical paper.](#)
- arXiv paper 2: [*Memanto: Stateful Agentic Memory Architecture - memory management and stateful-agent architecture*](#)
- [Open-Source Agentic Memory, Memanto, GitHub](#)
- [Moorcheh Edge Documentation](#)

Intel

- [OpenVINO toolkit: An open source AI toolkit](#)
- [New Intel Core Ultra Series 3 Processors](#)

Workshop 2

[Andrea Basso MiTo Tech] Language-Directed Semantic Navigation on a Heterogeneous MPU: Fusing Sparse Multi-Zone ToF, NPU Vision, and an On-Device Language Model

Indoor service robots increasingly must navigate around named obstacle categories on natural-language instruction ("avoid the chairs"), not geometry alone spanning mapping, recognition, and language grounding. Existing systems typically need cloud inference or GPU-class hardware, unsuited to small robots. We present a fully on-device architecture on ST's STM32MP2, pairing a Cortex-M33 real-time domain (four independent-bus VL53L9 multi-zone ToF sensors, 54×42 zones each, plus IMU) with a dual Cortex-A35 Linux domain running SLAM, NPU-accelerated YOLOv8n, and a memory-resident ~135M-parameter language model. A cross-modal tracker fuses vision, ToF distance, and intensity to compensate detector latency and recover thin-object misses. A scene-context bridge grounds the model in currently tracked objects rather than the instruction alone; a rate-limited planner escalation handles constraint infeasibility. Both are timeout-bounded with safe fallbacks, so the real-time path never depends on language-model latency. Central to the design is a confidence map, built from cross-sensor ToF disagreement and signal/ambient anomalies, that lets the language model reason about and express perceptual uncertainty rather than only object presence—a capability absent from prior language-conditioned navigation work, which grounds language only in already-clean detections. We report this as an architecture and bottleneck-analysis contribution;