Agentic TinyML for Intent-aware Handover in 6G Wireless Networks

Alaa Saleh, *Student Member, IEEE*, Roberto Morabito, *Member, IEEE*, Sasu Tarkoma, *Senior Member, IEEE*, Anders Lindgren *Senior Member, IEEE*, Susanna Pirttikangas, *Senior Member, IEEE*, and Lauri Lovén, *Senior Member, IEEE*

***Abstract*—As sixth-generation (6G) wireless networks evolve into increasingly Artificial Intelligence (AI)-driven, user-centric ecosystems, traditional reactive handover mechanisms demonstrate limitations, especially in mobile edge computing and autonomous agent-based service scenarios. This manuscript introduces the Wireless AI Agent Network (WAAN), a cross-layer framework designed to enable intent-aware and proactive handovers in 6G networks. WAAN embeds lightweight Tiny Machine Learning (TinyML) agents as autonomous, negotiation-capable entities across heterogeneous edge nodes that contribute to intent propagation and network adaptation. To ensure continuity across mobility-induced disruptions, WAAN incorporates semi-stable Rendezvous Points (RPs) that serve as coordination anchors for context transfer and state preservation. The framework's operational capabilities are demonstrated through a multimodal environmental control case study, highlighting its effectiveness in maintaining user experience under mobility. Finally, the article discusses key challenges and future opportunities associated with the deployment and evolution of WAAN.**

I. Introduction

We are witnessing the rapid emergence of user-centric, edge-based ecosystems where sixth-generation (6G) wireless networks are expected to host an increasing number of agentic services [1]. These services, powered by autonomous Artificial Intelligence (AI) agents, introduce rising computational and communication requirements that must be met under highly dynamic and heterogeneous environments. As a result, even fundamental mechanisms such as routing, task scheduling, and offloading are evolving beyond traditional models, moving toward more adaptive, context-aware strategies capable of supporting distributed decision-making across mobile devices and edge servers [2].

This evolution is also driven by the fact that future 6G networks will be increasingly AI-driven and intent-driven [1], with decision-making and service orchestration shifting closer to the edge. As a result, traditional network functions, including the concept of handover itself, need to be revisited.

In addition to classical handovers that only transfer connectivity, emerging agentic ecosystems call for a new form of intent handover.Unlike traditional task migration, which focuses on relocating computational processes between nodes to balance resources or maintain execution, intent handover focuses on transferring the semantic and operational context of a user's intent. This process involves the transfer of intermediate computational states, learned runtime logic, and policy information required for another agent to continue execution. Consequently, it preserves semantic and computational continuity without disruption under dynamic network conditions and across heterogeneous AI agents operating over multiple layers. This enables the receiving agent to resume processing the same user intent by reusing previously generated intermediate outputs and continuing the reasoning process without recomputation.

In such ecosystems, autonomous AI agents form distributed service architectures that interact, negotiate, and perform semantic reasoning to adapt their behavior in response to user intents and environmental context. In this respect, classic reactive handovers are no longer sufficient to cope with the demands of highly dynamic environments and resource-constrained edge nodes [3]. Moreover, the combination of user mobility, traffic fluctuations, and heterogeneous device capabilities introduces additional layers of complexity [4]. This makes resilience (i.e., the ability to maintain or restore service continuity under disruption) a critical concern in these ecosystems. When users move, intermittent connectivity can easily lead to delayed or lost responses [5]. Service continuity therefore refers to the network's capability to ensure an uninterrupted user experience despite mobility, node failures, or other disruptions. These challenges highlight the need to rethink how 6G architectures and AI agents work together to deliver seamless, intent-driven services in mobile environments.

Building on this vision, in such dynamic wireless edge environments, the architectural and operational foundations of 6G systems will directly shape how AI agents operate in mobile and distributed settings. These foundations will aim to deliver user-centric services that seamlessly adapt to changes in connectivity, location, user preferences, and network conditions, ensuring responsive and personalized experiences [6]. Achieving this adaptability requires more than static mechanisms: it depends on effective inter-agent negotiation across edge servers and mobile devices [7], [8]. In addition, leveraging Quality of Experience (QoE) metrics and user feedback [9] becomes essential to refine decision-making, responsiveness, and resilience.

To illustrate these challenges, Fig.1 shows a typical workflow: when a user submits an intent, a personal agent on the user's device decomposes the request into sub-tasks and forwards them to the closest AI agent running on a nearby edge device. This AI agent processes the sub-tasks by retrieving data from relevant sources (e.g., smart building datasets con-

A. Saleh (corresponding author), S. Pirttikangas, and L. Lovén are with Center for Ubiquitous Computing, University of Oulu, Oulu 90014, Finland;

R. Morabito is with Department of Communication Systems, EURECOM, Biot 06410, France;

S. Tarkoma is with Department of Computer Science, University of Helsinki, Helsinki 00100, Finland;

A. Lindgren is with RISE Research Institutes of Sweden, Stockholm 166 40, Sweden and Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå 971 87, Sweden

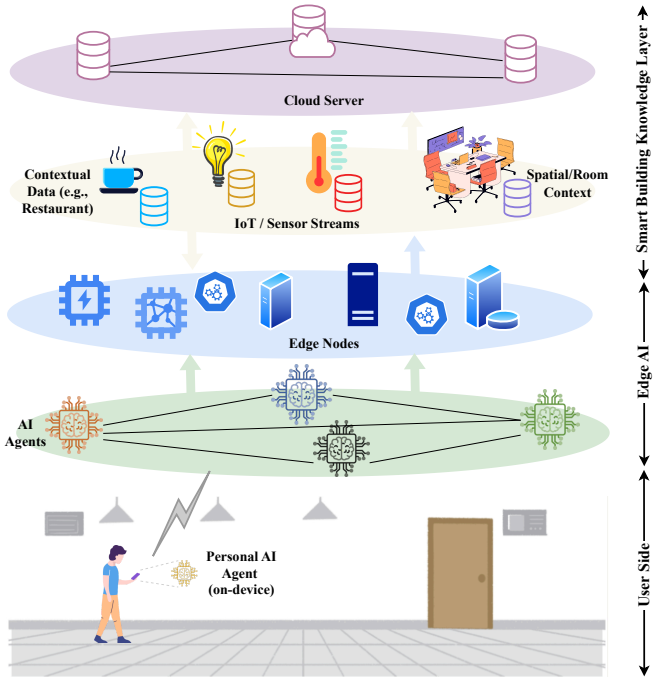

Fig. 1. Illustration of the intent-driven processing workflow in heterogeneous wireless environments.

taining information from rooms, restaurants, or local sensors) and generates a response, which is then returned to the user as a notification or action. However, user mobility can interrupt this workflow: if the user moves out of range during processing, the connection between the personal agent and the AI agent is lost. Without a mechanism for seamless continuation, the request must be resubmitted to a new AI agent, forcing redundant recomputation and causing additional latency, power consumption, and overhead.

This scenario highlights a key limitation: current systems lack mechanisms to seamlessly transfer the execution context of user intents across AI agents as users move. Overcoming this limitation requires rethinking how mobility is handled in AI-driven, intent-driven 6G networks, a challenge that we address through the design of the Wireless AI Agent Network (WAAN). The main contributions of this work are summarized as follows:

- We propose WAAN, a novel cross-layer adaptive intelligence framework that enables intent-aware and proactive handovers in 6G wireless systems.
- WAAN focuses on maintaining the semantic and computational continuity of user intents across heterogeneous and dynamic mobile edge environments.
- WAAN embeds lightweight, negotiation-capable TinyML agents across constrained edge nodes, enabling autonomous, on-device decision-making and cooperative inference to support adaptive intent propagation.
- We envision Rendezvous Points (RPs) as coordination anchors within WAAN to enhance continuity, state preservation, and resilience under mobility-induced disruptions.
- We provide a multimodal environmental control scenario, demonstrating WAAN's operational feasibility and effectiveness in maintaining user experience and service continuity under dynamic conditions.

II. Introducing the WAAN

In the WAAN, we envision an AI interconnect layer of agents operating across heterogeneous nodes to autonomously and seamlessly support the continuation of user intents across dynamic edge environments. Rather than focusing solely on connectivity, WAAN agents coordinate and negotiate task execution to maintain the semantic and computational continuity of user intents as they move. Crucially, this propagation process is not confined to the application or service layer—it operates in a cross-layer fashion, adapting to real-time wireless conditions, device capabilities, and network-level metrics [10], [11]. For instance, local AI models (e.g., Small Language Models (SLMs) or Tiny Machine Learning (TinyML) agents) can adjust task offloading, scheduling, or response generation based on variations in signal quality, congestion, or resource availability. Through this cross-layer adaptive intelligence, spanning both the device and network layers, WAAN aims to reduce redundant processing, minimize routing overhead, and support proactive intent handovers. While this represents a conceptual architecture, it addresses limitations in current systems that treat intent propagation as isolated from the underlying network dynamics.

For enabling this vision, WAAN is conceived as a loosely coupled, policy-driven architecture where agents collaborate without requiring tight synchronization, allowing the system to remain robust in dynamic, heterogeneous edge environments. A key design requirement is the deployment of adaptive intelligence on resource-constrained nodes, where lightweight AI and ML models continuously learn from local observations and neighbor interactions. This distributed intelligence allows agents to negotiate with one another, balancing their own intents and resource limitations with system-wide goals [12]. Efficient negotiation requires continuously learning agent behaviors that can autonomously manage service interruptions, reroute around failures, optimize resource usage, and dynamically adapt to evolving user intents. These capabilities are essential for maintaining a high user QoE under conditions of fluctuating connectivity, mobility, and network congestion. Central to this process is the intelligent management of how data is transmitted over the wireless channel [13]. This clearly involves the effective scheduling of wireless resources and the allocation of time–frequency radio resources to traffic flows, but also the ability to adapt to the nature of the applications generating that traffic. For example, Generative AI (GenAI)-powered agents are expected to drive a surge of uplink and downlink traffic due to interactive video assistants and immersive, multimodal applications, further straining network infrastructure. At the same time, the same GenAI models can help mitigate this load in closed-edge deployments by processing and semantically compressing data closer to the user. The possibility of extracting and transmitting only the most relevant content through these models can significantly reduce the volume of data sent over the network, alleviating congestion while maintaining responsiveness and quality [3].

Compared to existing agentic frameworks, which primarily focus on semantic reasoning and coordination among agents at the application layer, WAAN explicitly integrates cross-layer awareness and proactive intent handover. The decision-making mechanisms in agent-based orchestration frameworks are typically rule-based or reliant on Large Language Models (LLMs) reasoning. While this approach allows for flexible intent interpretation and semantic coordination, it offers limited responsiveness to real-time network dynamics. Consequently, such frameworks tend to react to service demands without considering the underlying connectivity conditions, channel variability, or device heterogeneity, and thus lack mechanisms to maintain computational continuity or preserve state during user mobility. At the same time, cross-layer handover frameworks primarily focus on reactive handovers. Although effective in sustaining connectivity and link continuity, they remain largely unaware of user intents and fail to maintain the semantic or computational context of ongoing tasks.

In WAAN, intent handover is not limited to passing control between agents of the same type: it spans different agentic domains (inter-agent systems), computing tiers (from on-device to edge to cloud), and data and knowledge sources on which these agents operate (as illustrated in Fig. 1). This enables an ongoing task to continue seamlessly even as the user moves across heterogeneous environments, devices, and datasets. The idea is to link decision-making to network state, making WAAN agents able to adapt task routing, offloading, and intent propagation strategies based on real-time wireless conditions. This combination of semantic intent handling with network-level adaptability distinguishes WAAN from current agentic systems, which largely treat networking as a passive substrate.

III. The Role of TinyML in WAAN

The realization of WAAN depends on the ability of even the most constrained devices to perform local inference and decision-making without relying on powerful cloud resources. TinyML offers lightweight, energy-efficient models that can run on a wide range of end and edge devices, from microcontrollers and low-power processors to more powerful AI-accelerated single-board computers, enabling agents to perceive context and act locally. In the WAAN framework, TinyML allows each node to autonomously evaluate conditions such as connectivity, load, and user intents, and to participate in negotiation processes without incurring prohibitive communication delays.

These models are not limited to application-layer reasoning. Instead, TinyML agents incorporate cross-layer signals, ranging from channel quality to energy availability, to inform decisions on when to offload tasks, when to handover intents, and how to manage local resources. Additionally, thanks to the possibility of relying on lightweight few-shot learning mechanisms, TinyML agents can adapt to new traffic patterns and environmental changes using minimal training data, thereby maintaining responsiveness even under previously unseen conditions [14].

A fundamental aspect of WAAN is its ability to adapt to the heterogeneous capabilities of participating devices. Extremely constrained nodes, such as microcontrollers, cannot run multiple or complex models simultaneously, which limits the type of decisions they can take locally. To cope with this limitation, WAAN introduces agent capability discovery mechanisms that allow agents to be aware of their own strengths and weaknesses and to collaborate with other agents accordingly [15]. Through this awareness and cooperation, even devices with minimal capabilities can contribute meaningfully to compound functions, relying on more capable peers for tasks beyond their local capacity. Conversely, more powerful devices, such as smartphones and edge nodes, can host multiple models and take on more complex reasoning or coordination roles, while still benefiting from distributed collaboration.

Looking at the new GenAI wave, TinyML can well complement more powerful models in WAAN by providing fast, local reactions, while complex reasoning or semantic understanding—such as intent interpretation or multimodal data summarization—can be offloaded to more capable agents. This hierarchical collaboration, deployable on emerging edge AI platforms, ensures that only the most critical or resource-intensive tasks reach the edge or cloud, while real-time decisions, such as choosing the best next hop for an intent, remain at the constrained node.

Our vision places TinyML as an active participant in a distributed agentic ecosystem. This approach is rather different than typical TinyML-based systems, which usually act in isolation for sensing or classification tasks. Through WAAN, TinyML nodes evolve from passive edge sensors to autonomous, negotiation-capable agents that play a direct role in intent propagation and network adaptation. This integration can represent an important paradigm shift for realizing intent-driven, cross-layer services in future 6G networks.

IV. Generalizable and Adaptive Intelligence: An In-Depth Look at WAAN

The WAAN forms an AI interconnect layer composed of autonomous agents distributed across heterogeneous network nodes. This intelligent layer autonomously supports user mobility and enables context-aware intent handovers through three main operational strategies: *(i)* bandwidth-aware resource allocation, *(ii)* dynamic inference offloading, and *(iii)* selective invocation of computational modules. Together, these strategies aim to optimize system performance under resource-constrained and dynamically changing conditions (see Fig. 2).

Fig. 2 provides an overview of the WAAN architecture. At the top-left, user intents are recognized and decomposed by a personal AI agent. These intents are then processed through a mesh of AI agents (center and right), where tiny agents continuously perform neighbor discovery and real-time metric collection on parameters such as (Central Processing Unit (CPU) load, memory usage, bandwidth availability, mobility patterns, and traffic type. The intent-receiver agent dynamically creates a *swarm* of tiny agents that gather these metrics from heterogeneous edge nodes to guide routing, resource allocation, and intent handovers, enabling decisions that are both context-aware, knowledge-aware, and mobility-aware.

Edge nodes, however, are resource-constrained and often exposed to new user intents and operational patterns, such

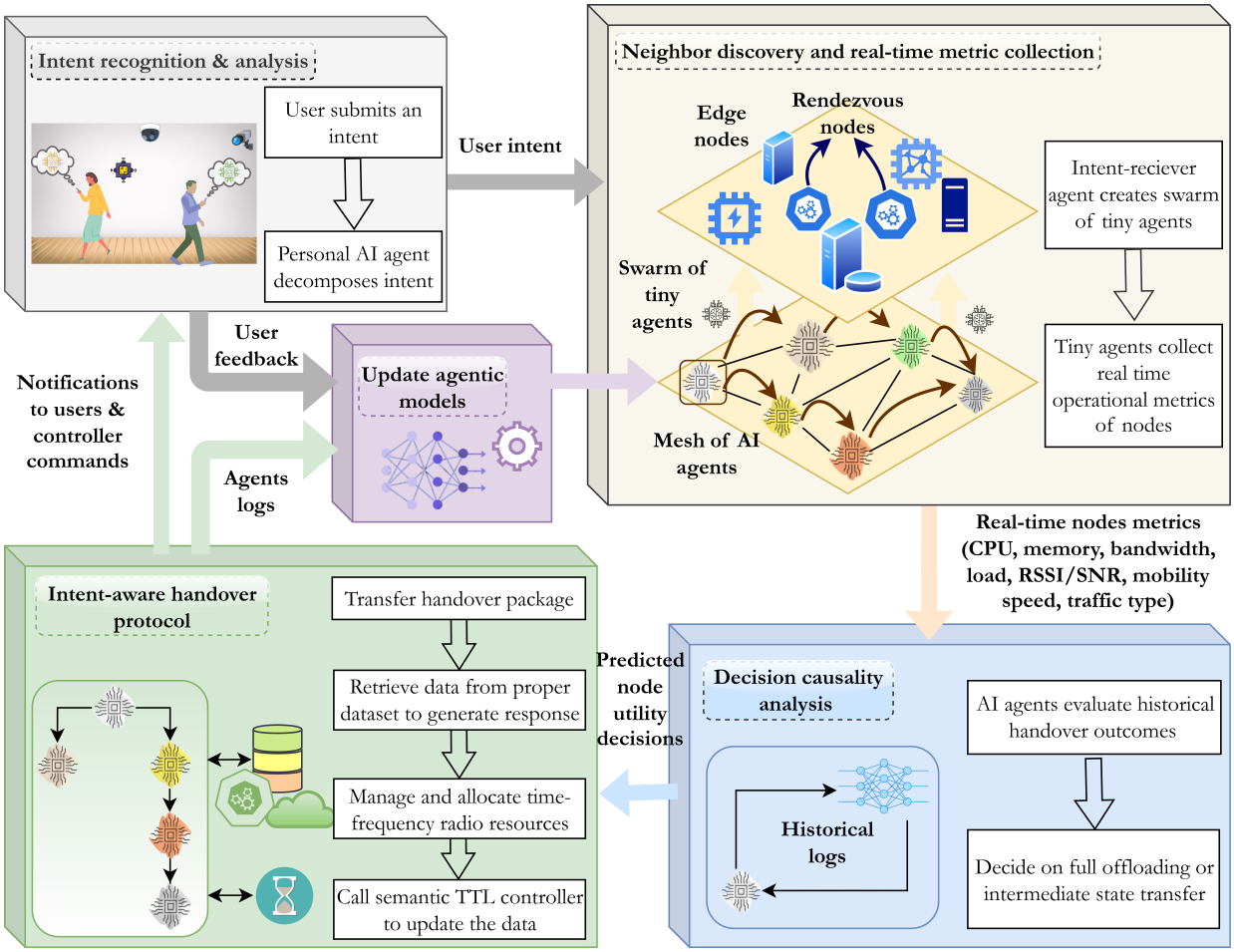

Fig. 2. Overview of the WAAN architecture for intent-aware handovers across heterogeneous edge nodes.

as sudden surges of immersive Extended Reality (XR) traffic, new types of GenAI assistants, or unexpected context changes caused by user mobility. Retraining large models on such nodes is infrequent due to overhead, which motivates the WAAN framework to rely on a self-expanding mesh of lightweight tiny-agents that collaboratively develop agentic policies guiding decision-making across the network. These agents need to track the operational requirements of each node but also predict the most suitable neighboring nodes for intent handover by analyzing the causal dependencies of past routing decisions. This capability is further enhanced by *few-shot generalization*, where minimal examples of success or failure in similar contexts allow the agents to make confident predictions for new scenarios.

The idea of embedding few-shot learning capabilities into TinyML models stems from the need to try making the intelligence capabilities of WAAN as generalizable as possible and at scale. Agents can adapt their negotiation mechanisms and improve local decision-making processes even in highly dynamic traffic conditions. Within this architecture, agents operate as autonomous control entities, responsible for scheduling tasks, managing computational offloading, and balancing edge resource loads against the predicted utility of each action to ensure resource allocation that maximizes benefit.

To support these adaptive behaviors, WAAN integrates dynamic offloading strategies, allowing tasks to be either fully offloaded or partially transferred via intermediate computational states. This enables new nodes to resume execution from prior states rather than reinitiating the task from scratch, thereby reducing latency, overhead, and energy consumption during mobility-induced disruptions. However, effective execution of such stateful handovers requires architectural elements that can maintain semantic continuity and state preservation across mobility-induced disruptions during agentic and intent handovers.

To address this, RPs in WAAN are envisioned to serve as semi-stable coordination nodes deployed at fixed edge or cloud locations. These nodes act as anchor points where user intents, data streams, and agentic policies intersect within the network. In our vision, these RPs could become the places where contextual state and intent metadata are temporarily cached or synchronized, potentially reducing the risk of interruptions and avoiding unnecessary recomputation when a handover occurs. At the same time, they could also provide an opportunity to maintain a certain level of system accountability and auditability: by concentrating part of the coordination at specific points in the network, it becomes feasible to log decisions, data exchanges, and agent behaviors, which is much

harder to achieve in a purely fully distributed setup. We see these RPs as one possible way of introducing some form of stateful coordination in what would otherwise be a completely dynamic and volatile agentic environment.

Building upon this concept, the loosely coupled, policy-driven architecture of WAAN supports horizontal scalability, allowing new RPs to be deployed dynamically as autonomous anchors without reliance on centralized control. Each RP operates as a logical coordination node rather than a bottleneck, ensuring that as the number of agents and devices increases, coordination remains tractable through hierarchical expansion.

WAAN further achieves scalability and adaptability across diverse network densities through the integration of localized, lightweight intelligence on resource-constrained devices. TinyML models embedded within WAAN nodes perform real-time inference on contextual parameters, such as bandwidth, mobility, and energy availability, enabling agents to autonomously decide when to offload tasks, reroute intents, or execute handovers. This decentralized intelligence not only minimizes reliance on centralized cloud infrastructures but also helps manage communication and computational overhead, thereby sustaining scalability as the number of connected devices increases. Moreover, selective intent propagation, guided by semantic relevance and few-shot adaptation, further reduces unnecessary transmissions and optimizes resource utilization.

In dense environments, swarms of TinyML agents balance computational loads through distributed negotiation, ensuring seamless service continuity across heterogeneous, variable-density networks and thereby realizing scalable, intent-driven intelligence. While localized TinyML intelligence ensures node-level adaptability, WAAN achieves system-level scalability through mechanisms that regulate swarm size and enable efficient agent discovery. Scaling is achieved by forming loosely coupled swarms of TinyML-powered agents that dynamically adjust their composition through adaptive policies, wherein agents autonomously expand or contract their local clusters according to network conditions and task complexity. Building upon this adaptive clustering, WAAN employs continuous, lightweight neighbor monitoring to facilitate efficient agent discovery. Through the exchange of real-time metrics, agents develop distributed awareness of neighboring capabilities. This decentralized discovery process ensures that tasks and intents are dynamically routed to the most suitable nodes.

At the same time, WAAN explicitly integrates RPs as a resilience mechanism to address failures arising from user mobility, link disruptions, or node outages. Their function in state preservation and recovery is fundamental to the framework's failure tolerance. When a mobile agent loses connectivity, a subsequent agent or node can retrieve the preserved execution context from the RP, enabling continued processing without full recomputation.

Historical logs enable the decision causality module to determine whether full offloading or intermediate state transfer is more appropriate, while experiential feedback loops continuously update the agentic models, ensuring adaptive, generalizable intelligence aligned with user QoE requirements (lower right of Fig. 2). These mechanisms allow agents to evaluate handover outcomes, recover from failed attempts, and fine-tune their internal models over time, which leads to better predictive decisions for future routing and intent placement. Additionally, fallback candidates may be pre-ranked, allowing for rapid reassignment when primary handover fails.

Finally, intent-based handovers are managed through an *intent propagation protocol* that incorporates a *semantic time-to-live (TTL)*. Unlike traditional TTLs that define the lifetime of a packet purely in temporal terms, the semantic TTL reflects contextual relevance and extends the traditional networking TTL concept by incorporating contextual relevance, intent validity, and task continuity into the lifetime of transferred state information. This mechanism ensures that the temporal and contextual relevance of information are considered when transferring intents, accounting for its semantic usefulness within ongoing interactions between agents. As a result, agents can avoid propagating stale or misaligned context that could degrade user experience. This approach ensures that handover decisions related to routing and offloading remain consistently aligned with user QoE requirements such as latency sensitivity and accuracy.

To ground the conceptual architecture of WAAN in a tangible setting, the following section details a step-by-step agentic process in operation. This illustration demonstrates how individual agents negotiate intent continuation, make context-driven decisions, and adapt across network layers to ensure uninterrupted user experience.

V. Case Study: Intent Handover with Cross-Layer Adaptation

We now illustrate how the WAAN architecture operates in practice, by considering representative case studies that highlight the role of intent handover, TinyML-driven cooperation, and cross-layer adaptivity. With these scenarios, we want to demonstrate how WAAN agents handle user mobility and heterogeneity while ensuring seamless continuation of tasks without recomputation or service degradation.

The following scenario illustrates a multimodal summarization task that is seamlessly handed over between agents as the user moves, without the need for task reinitialization. Guided by TinyML-based ranking and context transfer, this example demonstrates the complete agentic workflow within WAAN, from intent decomposition to proactive handover, encompassing the following stages: intent submission and initial processing, impending mobility detection, TinyML-based candidate ranking, intent handover, cross-layer adaptation, and completion.

Consider a user carrying a 6G-enabled smartphone who submits a complex intent: *"Provide a live multimodal summary of the room I am entering (using camera, microphone, and Internet of Things (IoT)* sensors) and adapt the environment accordingly (lighting, temperature)". The user's personal agent decomposes this intent into subtasks such as sensor fusion, multimodal summarization, and environment control, and offloads them to the closest WAAN agent based on current location. As the user moves from *Zone A* to *Zone B*, they traverse several WAAN coverage areas, each supported by edge nodes running autonomous AI agents that process these intents in real time.

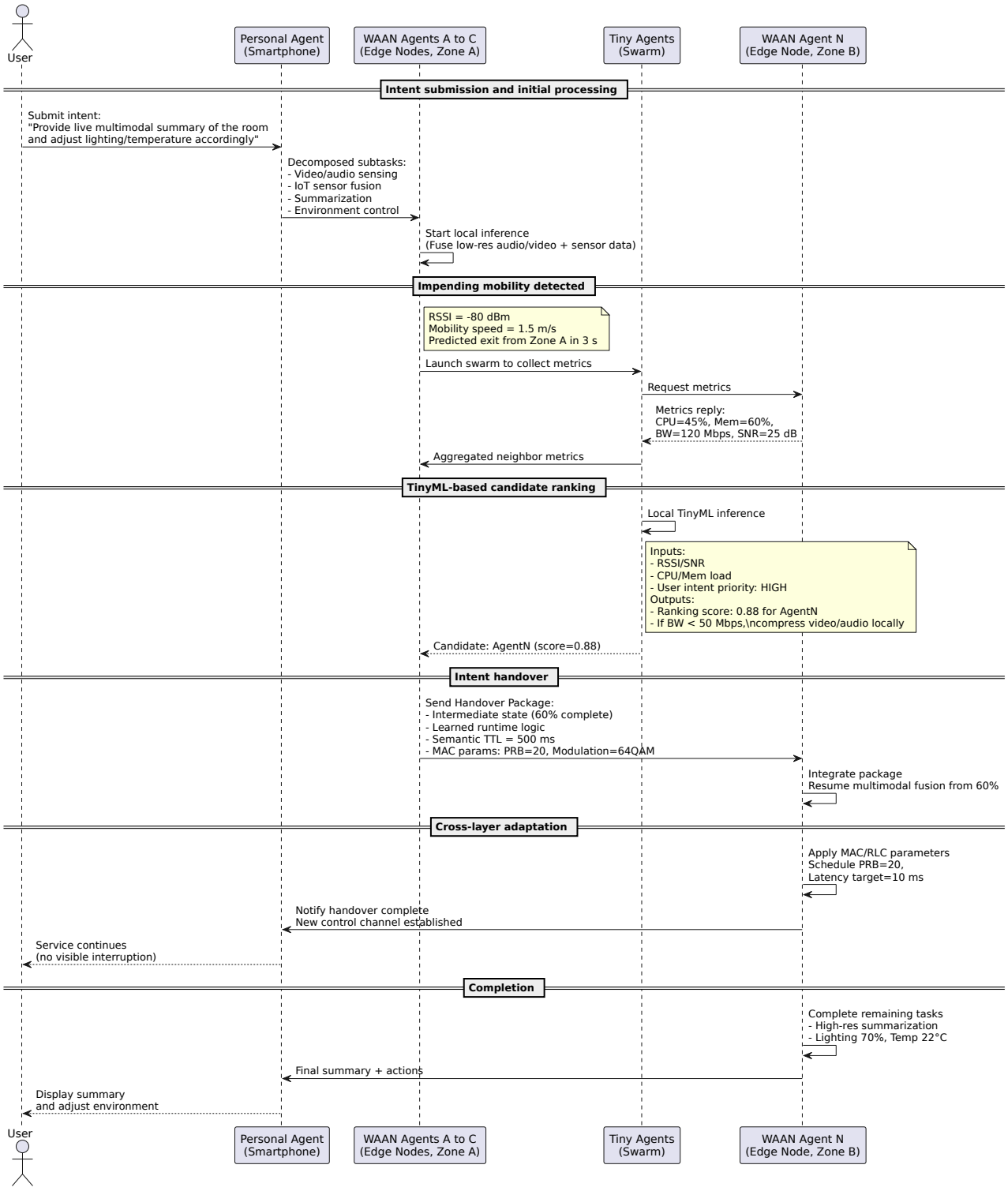

Fig. 3. Sequence diagram showing intent-aware handover with cross-layer adaptation in WAAN.

During this mobility, the initial AI agent (*Agent A*) may detect that the user is about to leave its coverage. Rather than allowing the session to drop, *Agent A* triggers an intent-aware handover protocol (illustrated in Fig. 3 and supported by the WAAN architecture in Fig. 2). This protocol uses a swarm of TinyML-powered agents to collect real-time operational metrics—such as CPU load, memory usage, available bandwidth, Received Signal Strength Indicator (RSSI)/Signal-to-Noise Ratio (SNR), mobility speed, and current traffic type—from neighboring nodes. Using these metrics, the tiny agents rank potential target agents using lightweight, local inference models. This decision incorporates both application-level requirements (e.g., low latency for multimodal processing) and network-level conditions (e.g., congestion, link

quality).

Based on this ranking, *Agent A* selects a new target (*Agent N*) and performs a knowledge-driven handover. Instead of simply transferring unfinished subtasks, *Agent A* sends a handover package that includes: *(i)* the intermediate task state (e.g., 60% of multimodal summarization already complete), *(ii)* the refined runtime logic and learned policies, *(iii)* the semantic TTL of context data, and *(iv)* relevant Medium Access Control (MAC)/ Radio Link Control (RLC) parameters. This package ensures that the receiving agent can continue execution exactly where it left off, without recomputation, while also inheriting insights from *Agent A*'s past decisions. If *Agent N* becomes unreachable during transfer, the system can quickly fall back to the next-ranked candidate, preserving progress with minimal degradation.

Upon receiving this package, *Agent N* integrates the logic into its own runtime, applies the MAC/RLC parameters for optimized scheduling of time-frequency resources, and re-establishes a control channel with the user's personal agent. This process exemplifies cross-layer adaptation: the application intent (multimodal environment awareness and control) and the lower-layer decisions (radio scheduling and resource allocation) are handled in a coordinated way, driven by TinyML at the lower tiers and larger models on edge nodes when needed.

After the handover, *Agent N* validates the relevance of the computed results using the semantic TTL, ensuring that stale data does not degrade quality. The response is then sent to the personal agent, which delivers a seamless user experience with no noticeable interruption. Through this integration of intent propagation, TinyML-driven agent cooperation, and cross-layer adaptation, WAAN enables a system that learns continuously and remains responsive under diverse mobility and environmental conditions.

This example is representative of the broader class of scenarios that WAAN targets: tasks where user mobility, heterogeneous resources, and dynamic network conditions can easily disrupt (complex) application-level experiences. It illustrates how WAAN's combination of knowledge-driven handover, TinyML-powered cooperation, and cross-layer adaptivity can preserve computational state and context while ensuring seamless service continuity. This concrete sequence shows that, instead of maintaining connectivity, WAAN preserves both semantic and computational continuity, ensuring that user intents persist. In practical terms, this enables adaptive environmental control and reduced latency for mobile AI services, demonstrating how WAAN operationalizes intent-aware intelligence in next-generation networks.

Building on this operational perspective, WAAN contributes to sustainability through energy-efficient local intelligence, adaptive resource management, and reduced computational redundancy. The deployment of lightweight TinyML models on edge and end devices minimizes reliance on power-intensive cloud infrastructures and mitigates the high energy costs associated with large-scale data transmission. Through localized decision-making and cross-layer adaptability, WAAN dynamically reallocates computational loads based on energy availability and network conditions, avoiding the energy waste of repeated recomputation. Furthermore, its intent-aware handovers and proactive context transfers enhance operational resilience, while RPs preserve computational and semantic states, allowing services to continue seamlessly and sustainably without restarting entire workflows.

Building on this representative scenario, the next sections discuss the advantages introduced by WAAN, the challenges that remain, and future research directions for intent-driven 6G edge networks.

VI. Advantages

The WAAN's explicit integration of semantic, intent-level processing with real-time wireless and network state enables agents to optimize routing, task offloading, and resource allocation. Through cross-layer adaptive propagation, WAAN reduces redundant computation, minimizes routing overhead, and supports proactive intent handovers that ensure computational continuity. During mobility events, proactive offloading and intermediate-state transfers, orchestrated through RPs as semi-stable coordination anchors, allow receiving nodes to resume execution without recomputation, thereby reducing disruption-induced delays and bandwidth overhead while maintaining user QoE under dynamic network conditions.

Furthermore, agentic behaviors, driven by continuous learning from local observations and neighbor interactions, could enhance adaptive intelligence across resource-constrained nodes. This may support dynamic intent adaptation, efficient resource management, and enhanced resilience by enabling service continuity in the presence of node failures, mobility, or other disruptions. Moreover, by incorporating cross-layer signals and relying on few-shot adaptation, TinyML may further extend autonomous control at the edge. Consequently, TinyML agents can intelligently decide when to offload, hand over intents, and allocate local resources efficiently, all while maintaining responsiveness and reducing reliance on distant compute infrastructures. Table I summarizes the key agentic operations and their corresponding effects within the WAAN framework.

VII. Challenges and Future Research

WAAN is designed to be compatible with existing 6G infrastructures and edge orchestration frameworks. Through its loosely coupled, policy-driven agentic architecture, WAAN achieves integration with heterogeneous devices and multi-tier computing environments spanning the cloud–edge continuum. Moreover, the incorporation of semi-stable rendezvous points enhances interoperability with middleware systems, enabling incremental deployment without disrupting established network operations.

Despite the advantages of WAAN in adaptability, efficiency, and autonomy, realizing this vision raises a number of fundamental challenges. These challenges arise from the heterogeneity of devices and models, the dynamics of wireless environments, and the lack of mature mechanisms for large-scale deployment and agent coordination in real-world settings.While these challenges are substantial, the integration of TinyML within WAAN offers promising avenues to mitigate

TABLE I
OPERATIONAL OVERVIEW OF TINYML AGENTS AND EFFECTS IN 6G WAAN.

TinyML Agent Operation	Operational Description	Network and System Effect
Adaptive Offloading	Fully or partially transfers intermediate computational states to neighboring or higher-tier nodes based on real-time metrics.	• Avoids device overload. • Reduces latency and energy consumption. • Maintains task continuity and responsiveness.
Intent Handover	Predictively initiates handover before link degradation, guided by signal quality, mobility, prediction, and intent priority.	• Ensures seamless connectivity. • Minimizes recomputation. • Preserves execution context across nodes.
Local Inference Scheduling	Dynamically schedules TinyML inference on constrained nodes using contextual metrics.	• Maintains service continuity. • Ensures local responsiveness.
Resource Negotiation with Peers	Coordinates with nearby agents to balance computational loads and share tasks according to capability discovery and policy rules.	• Enables load balancing. • Improves throughput. • Enhances resilience through cooperative adaptation.
Semantic TTL Adjustment	Dynamically extends or shortens the semantic TTL of intents based on contextual relevance and user QoE requirements.	• Optimizes intent propagation. • Prevents old or redundant data transfers.
Fallback Management	Activates pre-ranked backup agents when primary handover or offloading fails, utilizing cached intermediate states at rendezvous points.	• Prevents service disruption. • Increases system robustness and reliability.
Few-Shot Policy Update	Updates local decision or routing policies using few-shot learning from limited recent examples without full retraining.	• Enables adaptive learning. • Improves generalization and decision accuracy.

several of them. The distributed, energy-efficient intelligence provided by TinyML can enable real-time context awareness, adaptive coordination, and lightweight reasoning across constrained devices. However, several limitations call for new ideas and research directions in learning, orchestration, and security to fully unlock the potential of agentic intelligence in 6G, as outlined in Table II.

VIII. CONCLUSION

This article introduced WAAN as a cross-layer adaptive intelligence architecture designed to enable intent-aware and proactive handovers, representing a shift toward generalizable, intent-driven services in 6G agentic systems. By integrating lightweight TinyML agents with cross-layer decision mechanisms spanning device and network layers, WAAN achieves seamless continuity of user intents under mobility, heterogeneous resources, and fluctuating wireless conditions. The framework links decision-making to real-time network state, enabling agents to adapt task routing, offloading, and intent propagation while reducing redundant processing and routing overhead. Moreover, the integration of RPs as semi-stable coordination nodes enhances continuity in WAAN by enabling continuity and state preservation, ensuring intent-aware orchestration across dynamic 6G environments. Finally, while WAAN architecture introduces a robust basis for 6G agentic services, future research is required to address challenges related to semantic state transfer, adaptive learning under distribution shifts, and standardization of agent integration protocols.

ACKNOWLEDGMENT

This research is funded by the Research Council of Finland through the 6G Flagship (Grant Number 369116) project, and by Business Finland through the Neural Pub/Sub project (Diary No. 8754/31/2022) and the Digital Twinning of Personal Area Networks for Optimized Sensing and Communication project (Diary No. 8782/31/2022).

REFERENCES

[1] C. G. Brinton, M. Chiang, K. T. Kim, D. J. Love, M. Beesley, M. Repeta, J. Roese, P. Beming, E. Ekudden, C. Li, G. Wu, N. Batra, A. Ghosh, V. Ziegler, T. Ji, R. Prakash, and J. Smee, "Key focus areas and enabling technologies for 6g," *IEEE Communications Magazine*, vol. 63, no. 3, pp. 84–91, 2025.

[2] C. Zhou, W. Liu, T. Han, and N. Ansari, "Deploying on-device aigc inference services in 6g via optimal mec-device offloading," *IEEE Networking Letters*, vol. 6, no. 4, pp. 232–236, 2024.

[3] A. Saleh, R. Morabito, S. Dustdar, S. Tarkoma, S. Pirttikangas, and L. Lovén, "Towards Message Brokers for Generative AI: Survey, Challenges, and Opportunities," *ACM Comput. Surv.*, June 2025. Just Accepted.

[4] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6g communications," *IEEE Wireless Communications*, vol. 31, no. 6, pp. 48–55, 2024.

[5] R. Baldoni, L. Querzoni, S. Tarkoma, and A. Virgillito, *Distributed Event Routing in Publish/Subscribe Systems*, pp. 219–244. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

TABLE II
WAAN: CHALLENGES AND FUTURE RESEARCH.

Challenge Area	Challenge Description	Future Research Directions
Semantic Transfer and Intent Handover	• Coordinating the transfer of execution context in mobile environments remains a complex challenge. • This process involves the transfer of models, intermediate states, runtime logic, and learned policies across heterogeneous agents. • Emerging agent platforms include basic task handoff primitives. However, they are designed for static conditions and do not yet support intent-driven orchestration under wireless constraints.	• TinyML agents can facilitate this process by enabling cooperating with peer agents to exchange partial states or contextual knowledge. • Future work may explore the adaptation of emerging agentic standards, such as the Model Context Protocol (MCP) and Agent-to-Agent (A2A) protocols, to support TinyML-driven workloads and operate efficiently on constrained devices.
Intent Coordination and User-in-the-Loop Adaptation	• As WAANs scale, overlapping user intents will compete for shared resources and impose heterogeneous latency and compute demands.	• LLMs may act as brokers to interpret and prioritize intents. • Continuous user feedback can further refine agent behavior. Such feedback ensures that system responses align with evolving user preferences.
Cross-Layer Reasoning	• Enabling agents to perform cross-layer reasoning across application, transport, and radio layers under dynamic conditions.	• Agents must learn to reason jointly across these layers, adapting decisions to current network and resource conditions. • Designing new coordination and resource allocation strategies that balance TinyML and large-models.
Learning Under Uncertainty	• Applying WAAN in unstructured or highly dynamic environments remains challenging.	• TinyML's few-shot adaptability enables rapid local learning on constrained devices. • Due to its limited model capacity, existing few-shot learning methods must be further extended to ensure robust generalization and rapid adaptation under new mobility patterns, network behaviors, and traffic distributions.
Protocols, Security, and Compliance	• WAANs demand new standardized protocols that ensure trust, security, and regulatory compliance. • RPs and intermediate-state transfers create key security and privacy vulnerabilities, including risks of compromise, data leakage, and expanded attack surfaces.	• New protocols that integrate application-level semantics with network context. • RPs may help provide accountability and safe handling of sensitive state during intent handovers, where sensitive intent data and intermediate states are exchanged only between authenticated agents operating under secure communication protocols. • Additional mechanisms will be required to ensure compliance with privacy regulations such as the General Data Protection Regulation (GDPR) and the AI Act. • Robust cryptographic protections,end-to-end encryption , and integrity assurance for intermediate states, as well as transparent logging and explainability are needed to meet evolving regulatory standards, thereby ensuring transparent and privacy-preserving operations throughout the handover process.
Performance metrics.	• WAAN's cross-layer and agentic nature require suitable evaluation metrics.	• Intent handover efficiency should be measured via intent handover latency and intent continuity ratio. • Intent handover latency captures the time needed to transfer and resume execution of an intent between agents. • Intent continuity ratio measures the percentage of intents successfully resumed without recomputation or interruption. • TinyML inference efficiency reflects responsiveness and adaptability of lightweight agents under constrained resources. • Contextual relevance score measures the degree of semantic alignment between the transferred context and the current environmental.
Synchronization	• Loosely coupled, policy-driven architecture of WAAN enables horizontal scalable coordination, but maintaining synchronization and policy consistency across dynamically expanding RPs remains a challenge. • Continuous information exchange and need for synchronization and policy consistency can introduce significant communication and computational overhead.	• Future work should explore adaptive, event-driven synchronization strategies that minimize unnecessary coordination while preserving semantic consistency. • Compressed state sharing could reduce overhead, enabling sustainable operation on constrained devices.

[6] L. Zeng, S. Ye, X. Chen, and Y. Yang, "Implementation of big ai models for wireless networks with collaborative edge computing," *Wireless Commun.*, vol. 31, p. 50–58, June 2024.

[7] N. Xue, Y. Sun, Z. Chen, M. Tao, X. Xu, L. Qian, S. Cui, and P. Zhang, "Wdmoe: Wireless distributed large language models with mixture of experts," in *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, pp. 2707–2712, 2024.

[8] M. Xu, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, D. I. Kim, and K. B. Letaief, "When large language model agents meet 6g networks: Perception, grounding, and alignment," *IEEE Wireless Communications*, vol. 31, no. 6, pp. 63–71, 2024.

[9] H. Lee, M. Kim, S. Baek, W. Zhou, M. Debbah, and I. Lee, "Ai-driven decentralized network management: Leveraging multi-agent large language models for scalable optimization," *IEEE Communications Magazine*, vol. 63, no. 6, pp. 50–56, 2025.

[10] S. Tarkoma, *Overlay Networks: Toward Information Networking.* Auerbach Publications, 2010.

[11] A. Saleh, P. K. Donta, R. Morabito, N. H. Motlagh, S. Tarkoma, and L. Lovén, "Follow-me AI: Energy-efficient user interaction with smart environments," *IEEE Pervasive Computing*, 2025.

[12] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: a survey of

progress and challenges," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.
[13] Z. Chen, Z. Zhang, and Z. Yang, "Big ai models for 6g wireless networks: Opportunities, challenges, and research directions," *IEEE Wireless Communications*, vol. 31, no. 5, pp. 164–172, 2024.
[14] V. Rajapakse, I. Karunanayake, and N. Ahmed, "Intelligence at the extreme edge: A survey on reformable tinyml," *ACM Comput. Surv.*, vol. 55, July 2023.
[15] R. Morabito and S. Jang, "Smaller, Smarter, Closer: The Edge of Collaborative Generative AI," *IEEE Internet Computing*, pp. 1–9, 2025.

BIOGRAPHIES

Alaa Saleh (Student Member, IEEE) works as a PhD researcher at the Future Computing Group, in the Center for Ubiquitous Computing, University of Oulu, 90014, Oulu, Finland. She previously worked in the field of automated machine learning before focusing on her current research interests: publish/subscribe paradigm, edge intelligence, and Generative AI. Contact her at alaa.saleh@oulu.fi.

Roberto Morabito (Member, IEEE) is an Assistant Professor in the Communication Systems Department at EURECOM, France, and a Docent at the University of Helsinki. His research focuses on networked AI systems, with a particular emphasis on Edge AI service provisioning and lifecycle management under computing and networking constraints. He earned his PhD from Aalto University in 2019 and has previously held positions at the University of Helsinki, Princeton University, and Ericsson Research Finland. Contact him at roberto.morabito@eurecom.fr.

Sasu Tarkoma (Senior Member, IEEE) is a professor of computer science at the University of Helsinki, 00100, Helsinki, Finland, and the Dean of the Faculty of Science. He is a visiting professor with the 6G Flagship at the University of Oulu. His research interests include Internet technology, distributed systems, data analytics, and mobile and ubiquitous computing. Tarkoma received his Ph.D. degree in computer science from the University of Helsinki. He has authored four textbooks and has published over 250 scientific articles. He has 11 granted U.S. patents. He is a senior member of IEEE. Contact him at sasu.tarkoma@helsinki.fi.

Anders Lindgren is currently a senior researcher at RISE and an adjunct senior lecturer at Luleå University of Technology, Sweden. He received his Ph.D. from Luleå University of Technology in 2006, and worked at University College London and the Universtity of Cambridge from 2007-2008. Dr. Lindgren was a pioneer within DTN and ICN research and has been active in the research areas from an early stage of these research fields. He has since migrated much of this experience into distributed and edge AI research and has coordinated the EU ECSEL project DAIS with 47 partners across Europe. Anders has been involved in the work of the IRTF DTN and ICN research groups and is the co-founder of the ExtremeCom conference series. Contact him at anders.lindgren@ri.se

Susanna Pirttikangas (Senior Member, IEEE) is a Lead AI Scientist / PMTS at AMD Silo AI, D.Sc (tech), Adj. Prof. (data science). She was a deputy director for the Center for Ubiquitous Computing at the University of Oulu and the principal investigator of Interactive Edge (iEdge) research team developing adaptive, reliable, and trusted edge intelligence. Pirttikangas has extensive background in artificial intelligence related research and business. Contact her at susanna.pirttikangas@oulu.fi.

Lauri Lovén, (Senior Member, IEEE) is an Assistant Professor (tenure track), the leader of the Future Computing Group, and the vice director of the Center for Ubiquitous Computing (UBICOMP), University of Oulu, 90014, Oulu, Finland. Prof. Lovén coordinates the distributed intelligence strategic research area in the national 6G Flagship research program. He received his Title of Docent and D.Sc. degrees from the University of Oulu in 2025 and 2021, respectively, and was with the Distributed Systems Group, TU Wien, in 2022. His current research focuses on edge intelligence. Contact him at lauri.loven@oulu.fi.