Edge AI and Beyond

Designing Robust AI Architectures for the 6G Era

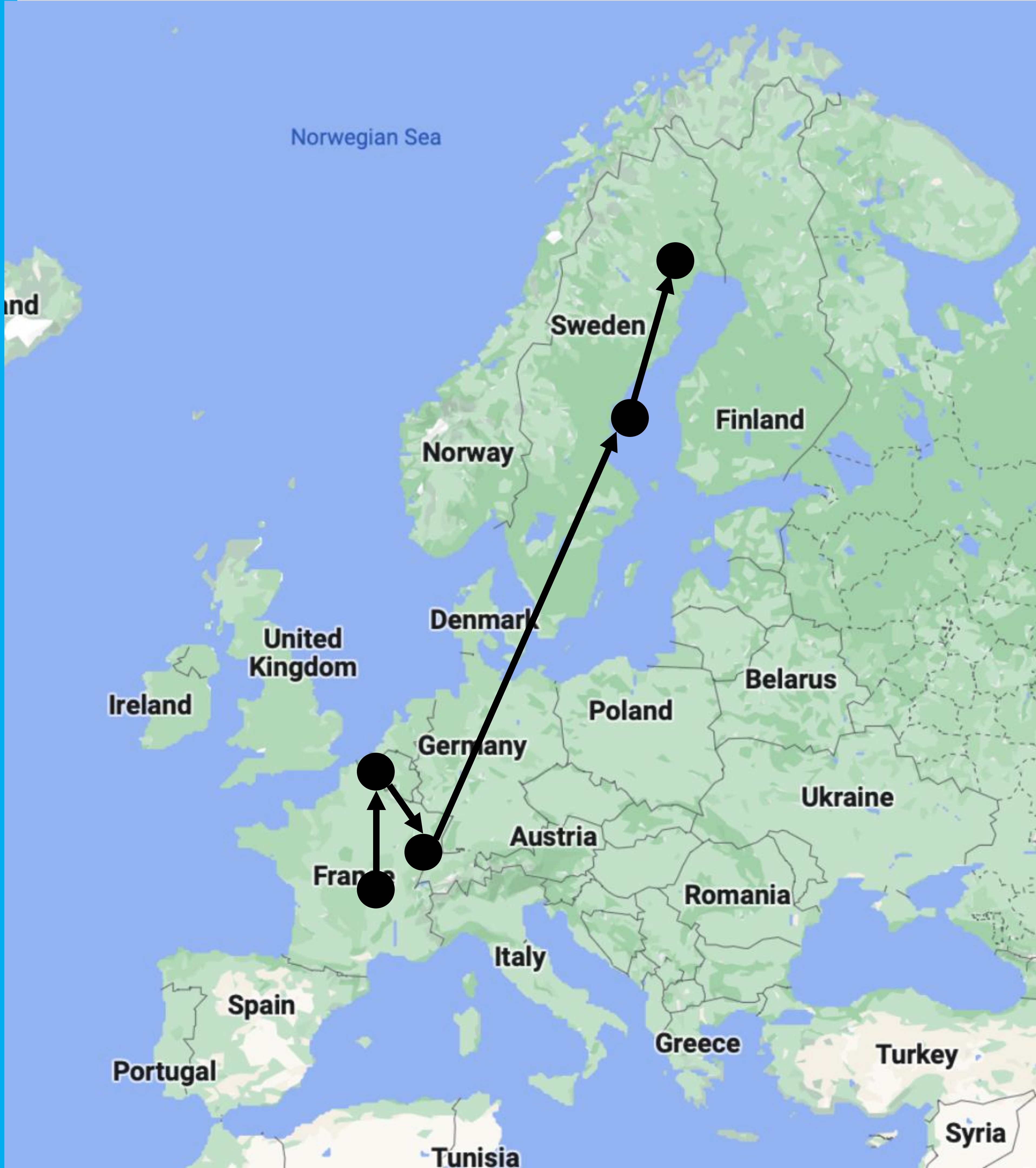
Prof. Davide Taibi

University of Oulu

About me

Full Professor @ University of Oulu

Head of the M3S Cloud Research Group



The M3S Cloud Research Group

- 3 Post-doctoral
- 6 Ph.D. students
- 3 Lecturers

M3S Cloud Research Group

Legacy System Refactoring

- Monolithic to Microservice
- Decision Framework for Migration
- Slicing / Decomposition
- Business Process Optimization

Cloud Architectural Quality

- Architectural Degradation
- Metrics for MS Coupling and Cohesion
- Service Architectural assessment
  - Reconstruction Tools
  - Organizational/architectural analysis

M3S Cloud

Software Quality

- Open Source Evaluation
  - Sustainability
  - Life Expectance
  - License compliance
- Static Analysis tools Configuration

Cloud to Edge – EdgeAI

- EdgeAI Orchestration/Offloading
- Green Edge
- Architectural Patterns
- Visualizations

Current Projects

Legacy System Refactoring

Cloud Architectural Quality

M3S Cloud

Software Quality

Cloud to Edge – 6G Soft

2023-2026. 1M euro
MuFaNo. Anomaly Detection for Microservices

Oulu Foundation 2022-2027. 300K euro
Architectural Degradation for Cloud-Native Systems

ABB **2022-2026. 320K euro**
Architectural Reconstruction and Degradation

BUSINESS FINLAND **2024-2026. 200K euro**
Edge Continuum Architectures

BUSINESS FINLAND **2024-2026. 1.5M euro**
6GSoft. *Software Eng. for 6G highly distributed Edge-AI systems*

Funded By

My Research Roadmap on Edge-AI

Edge-AI Architectures	Edge AI Orchestration
• Best/Bad Practices	• Load balancing
• Patterns / Anti-Patterns	• Scheduling
• Tools pipelines	• Resource planning
• Extremely distributed AI	• Optimization
• Quantum as a Service	• Quantum as a Service

Distributed intelligent computing

Hyperlocality, orchestration & trust

The 6GSoft Project (1.5M)

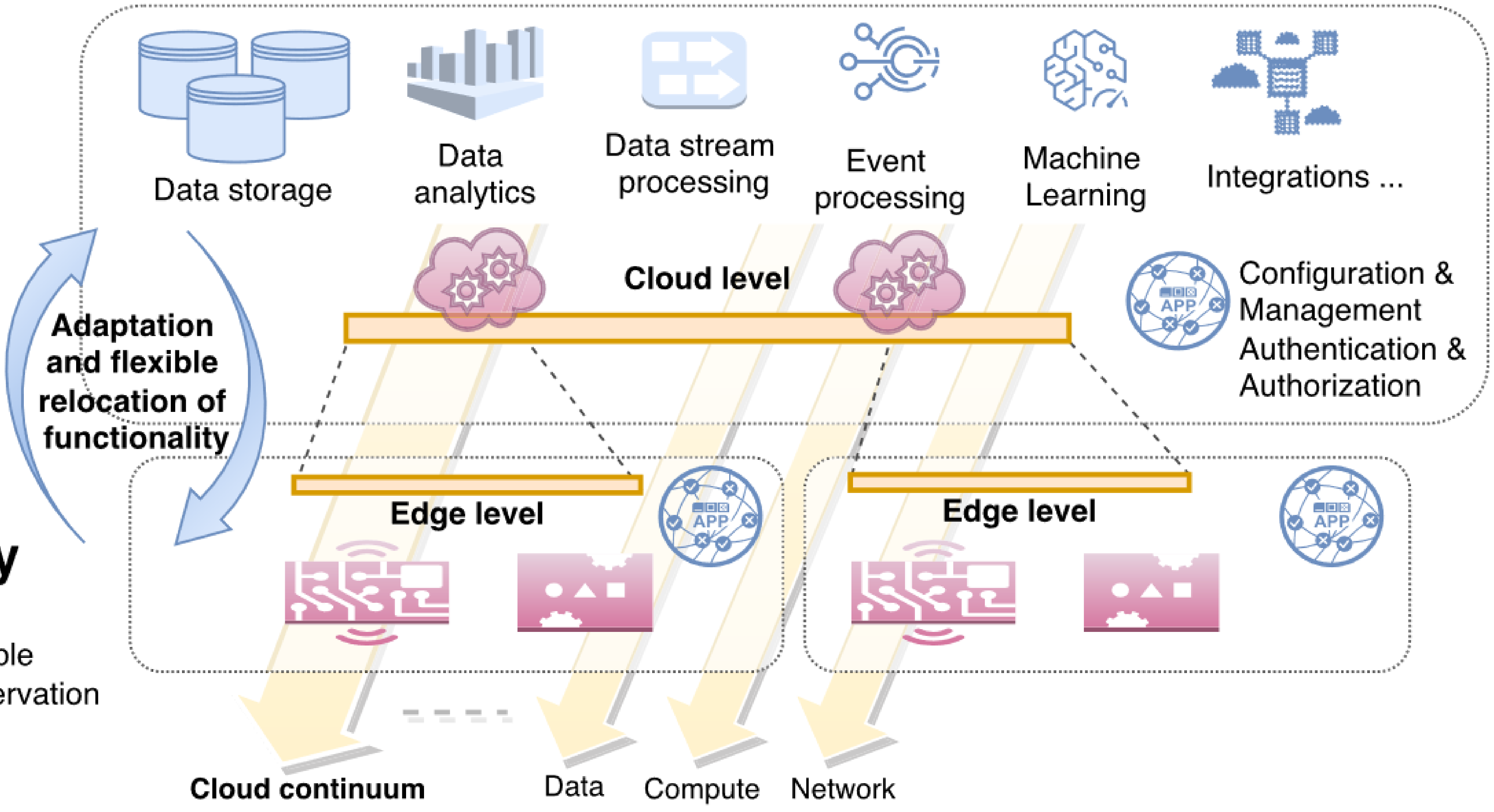
Cloud scalability

Infinitely scalable resources, wider integration possibilities

Fog flexibility

Edge dependability

Low-latency and deterministic communication capabilities, reliable communications, local data preservation



Avision of the distant future

Everything…

- Sensing
- Connected
- Autonomous
- Multimodal
- Personalized

- **Smart connected devices** are omnipresent
- Assisted automatic common tasks
- Highly personalized devices
- Centralized and coordinated behaviour

- **Research opportunities** in interactions, biometrics, analyzing behaviour, predicting intentions

INTELLIGENT, AI BASED CONNECTED DEVICES

The trend: ***Ultra-densification of wireless systems***

” (…communications had) approximately millionfold capacity increase since 1957.

Breaking down these gains shows

- a 25× improvement from wider spectrum,
- a 5× improvement by dividing the spectrum into smaller slices,
- a 5× improvement by designing better modulation schemes, and
- a whopping **1600×** gain through reduced cell sizes and transmit distance

The enormous gains reaped from smaller cell sizes arise from efficient spatial reuse of spectrum or, alternatively, higher area spectral efficiency”.

We should expect ultradense networks !!

Near future application 1: *Road safety, overtaking cars*

Seeing through buildings at city street corners, other vehicles and curves of the road

Ultra-dense wireless communications networks needed.
But how to make a scalable application?

Near future application 2: *Augmented environmental information*

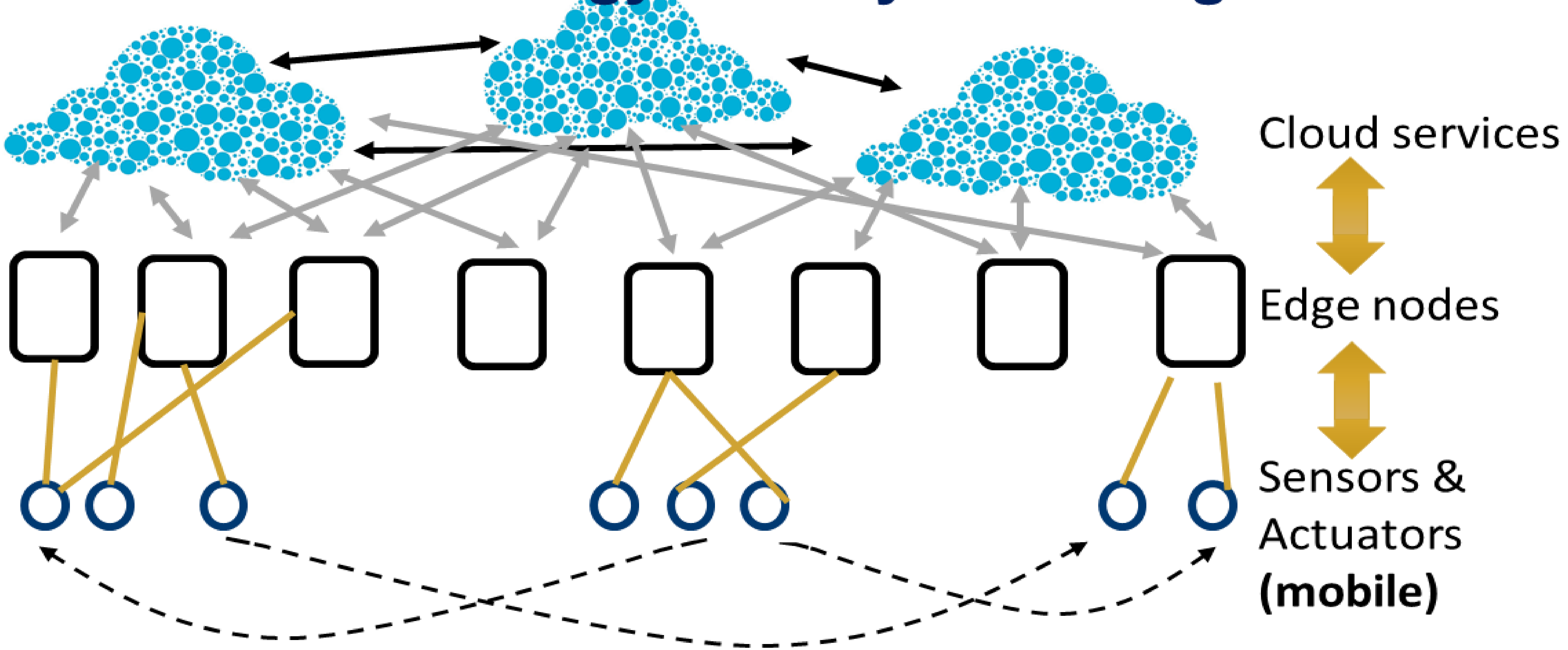
Super-situational awareness for an individual (and about the individual)

A secure digital twin of an individual follows the person, and connects with relevant sensors in the proximity

- not only just environment information, but also ”eyes in the back”
- tracking of health related information
- continuous aggregation and personalization of models

Ultra-dense communications and sensor networks that supports the mobility of individuals. **But how?**

Technology is really on its edge!



Implementing apparently straightforward application scenarios involves **big** **logistical challenges** concerning

- distribution of machine learning,
- seamless application mobility,
- information security and privacy.

Distributed Intelligence Overview

Computing on the edge-to-cloud continuum

- Evolution of edge-cloud continuum
- Serverless edge computing
- Trustworthy distributed service management
- Containers, life-cycle and autonomy
- On-device processing
- Edge computing extensions to 5GTN

Distributed AI

- Weak coupling and decentralized decision-making
- Federated learning
- Transfer learning
- Data provenance
- Split learning
- Single and multi-agent reinforcement learning
- Emergent communication protocols
- Learning via invariant causal mechanisms
- Semantic communication

Multimodal Sensing & modelling

- View-to-communicate Communicate-to-view
- THz imaging
- Backscattering
- Radar+Communication sensing using RL
- Multimodal 3D modeling
- Non-line-of-sigth sensing
- Radio propagation simulation & modeling

App 1 — Distributed vision without line of sight *(seeing through walls)*

App 2 — Remote biosignal acquisition *(rPPG & rBCG)*

App 3 — Radio & Vision Cooperative systems *(Multimodal cameras + radios)*

. . .

App N — Drones, driving, energy, …

ULTRADENSIFICATION

Distributed Intelligence Overview

App 1

App 2

App 3

. . .

App N

Computing on the edge-to-cloud continuum

- Evolution of edge-cloud continuum
- Serverless edge computing
- Trustworthy distributed service management
- Containers, life-cycle and autonomy
- On-device processing
- Edge computing extensions to 5GTN

Distributed AI

- Weak coupling and decentralized decision-making
- Federated learning
- Transfer learning
- Data provenance
- Split learning
- Single and multi-agent reinforcement learning
- Emergent communication protocols
- Learning via invariant causal mechanisms
- Semantic communication

Multimodal Sensing & modelling

- View-to-communicate Communicate-to-view
- THz imaging
- Backscattering
- Radar+Communication sensing using RL
- Multimodal 3D modeling
- Non-line-of-sigth sensing
- Radio propagation simulation & modeling

Distributed vision without line of sight *(seeing through walls)*

Remote biosignal acquisition *(rPPG & rBCG)*

Radio & Vision Cooperative systems *(Multimodal cameras + radios)*

. . .

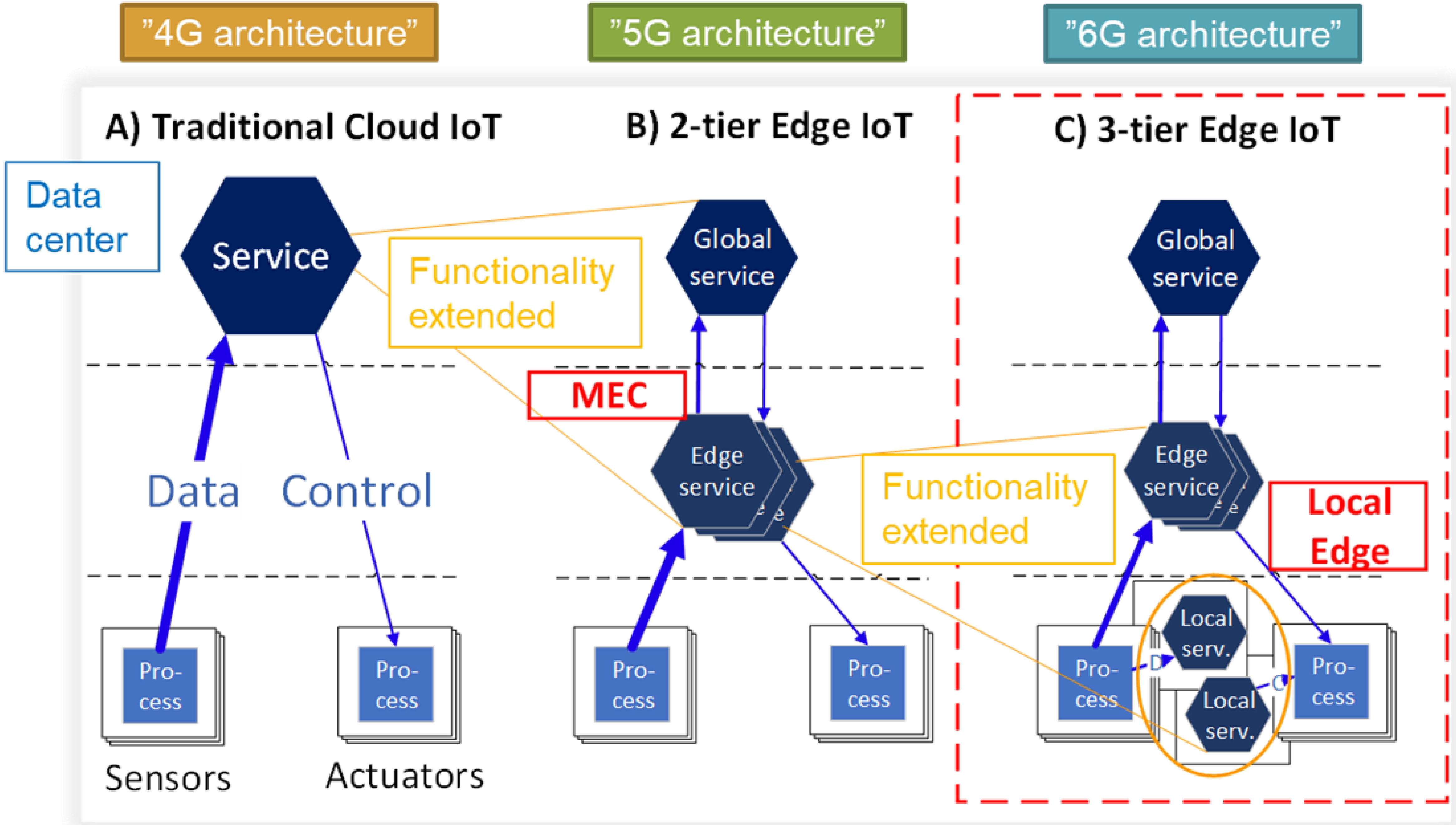
Drones, driving, energy, …

ULTRADENSIFICATION

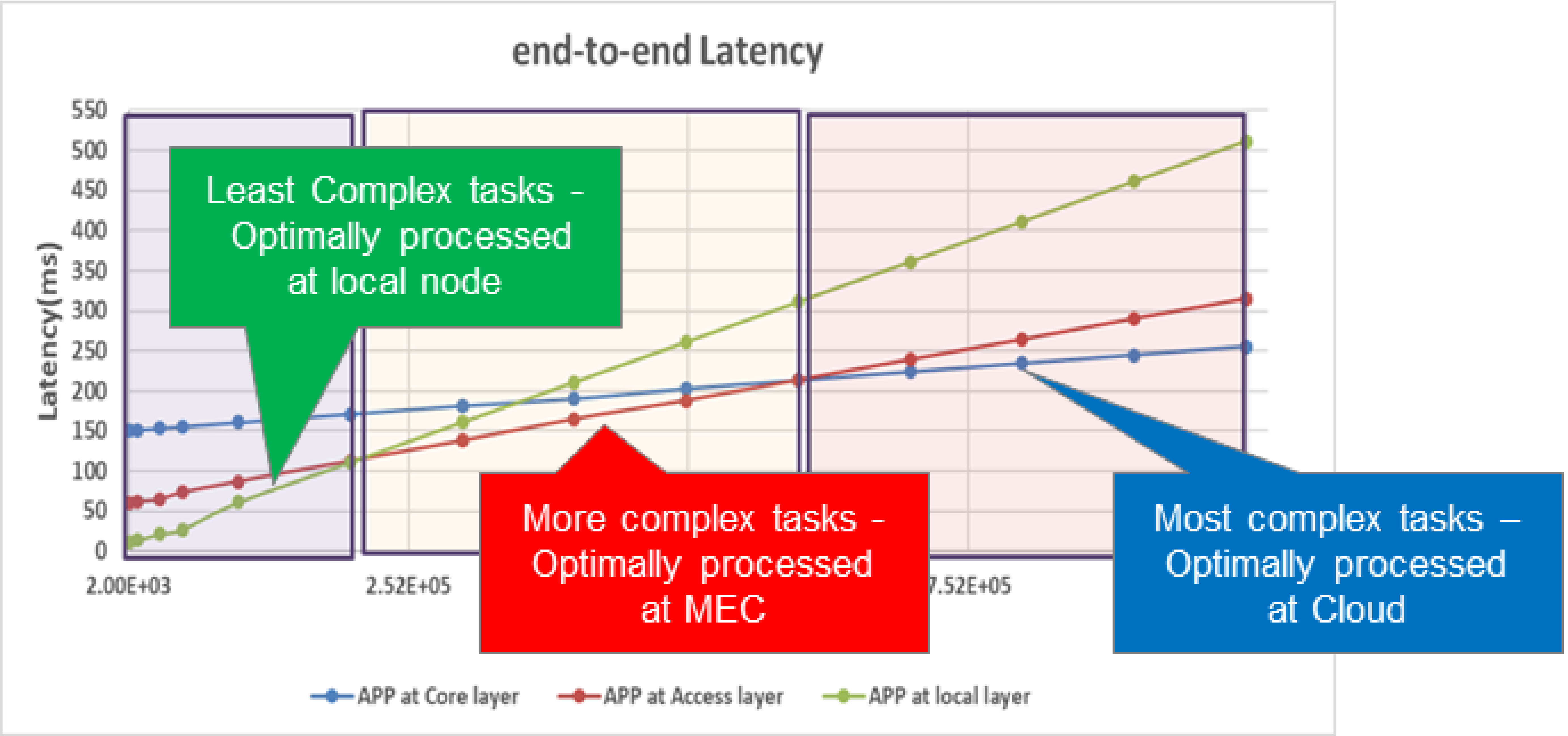
Operating Cloud Native AI based Systems

Computing on the edge-to-cloud continuum

- The three-tier edge architecture helps optimizing latency, network usage and power consumption by allowing capacity-aware placement of computational tasks on different tiers.
- Time varying model splitting and compression, accounts for uncertainties in processing & communication, lowers energy consumption and CO2 emission
- Resource consumption and service deployment & initiation times on a feasible level.



Computing on the edge-to-cloud continuum



Computing on the edge-to-cloud continuum

Orchestration:

Local orchestrator deploys lightweight granular microservices in containers

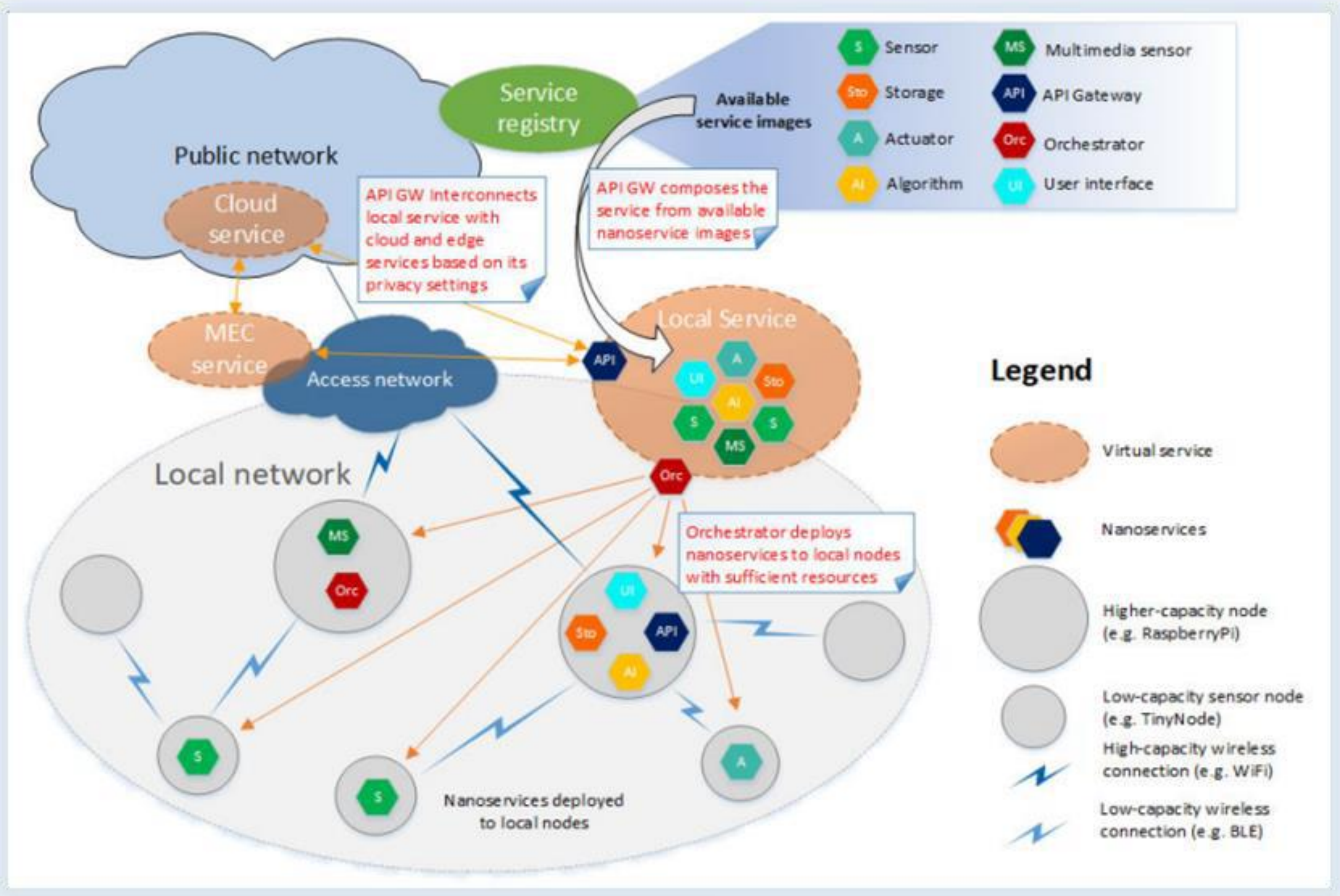
Local API gateway can compose local services of on small virtual microservices (nanoservices).

Orchestrator takes care of deployment, redeployment and undeployment

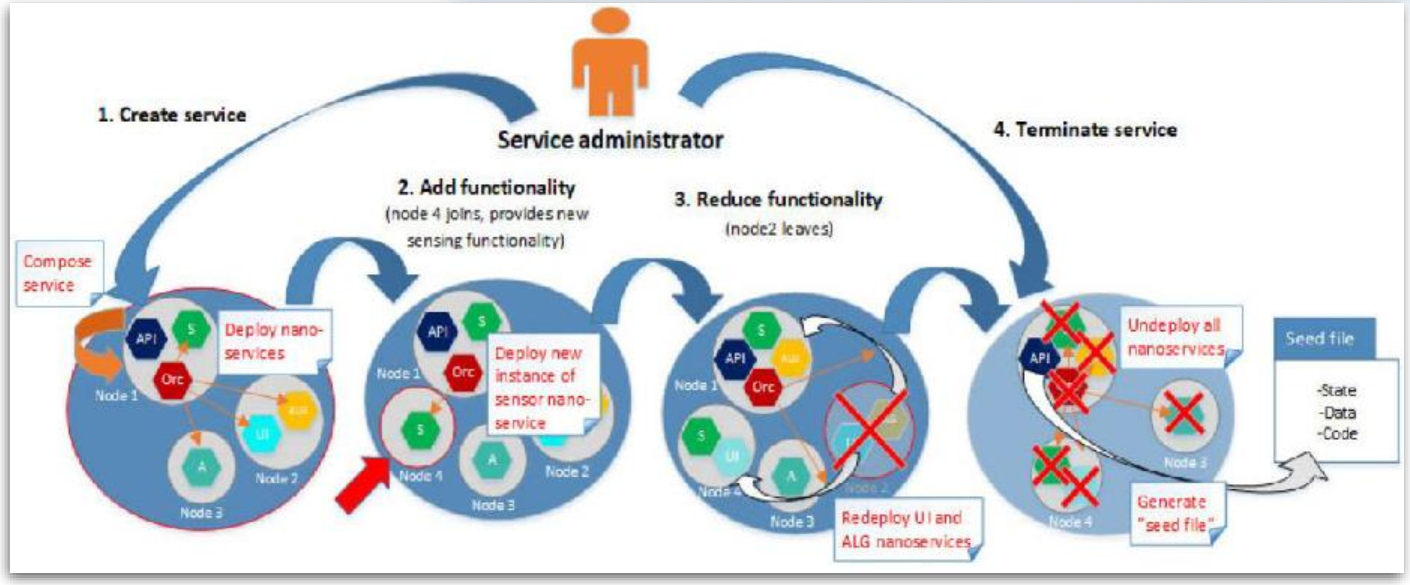
- New devices with new capacity can join the system and existing nodes may leave the system while service is running.

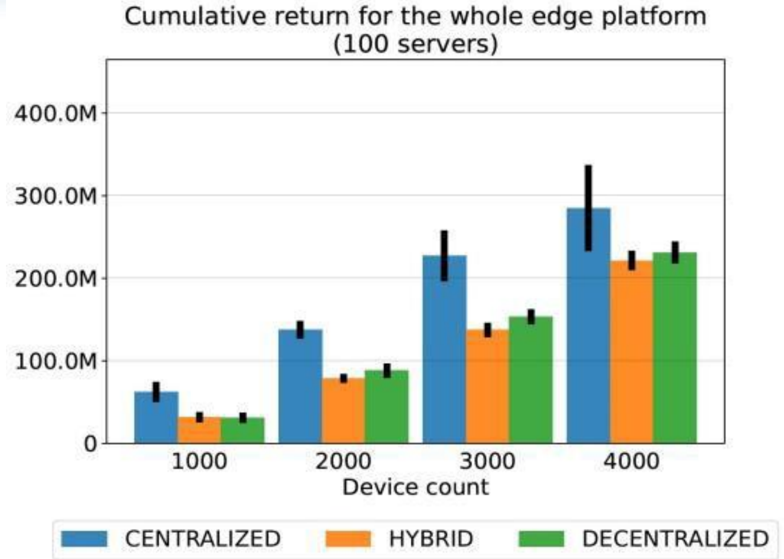
Orchestrator can be simulated !!

Serverless edge computing

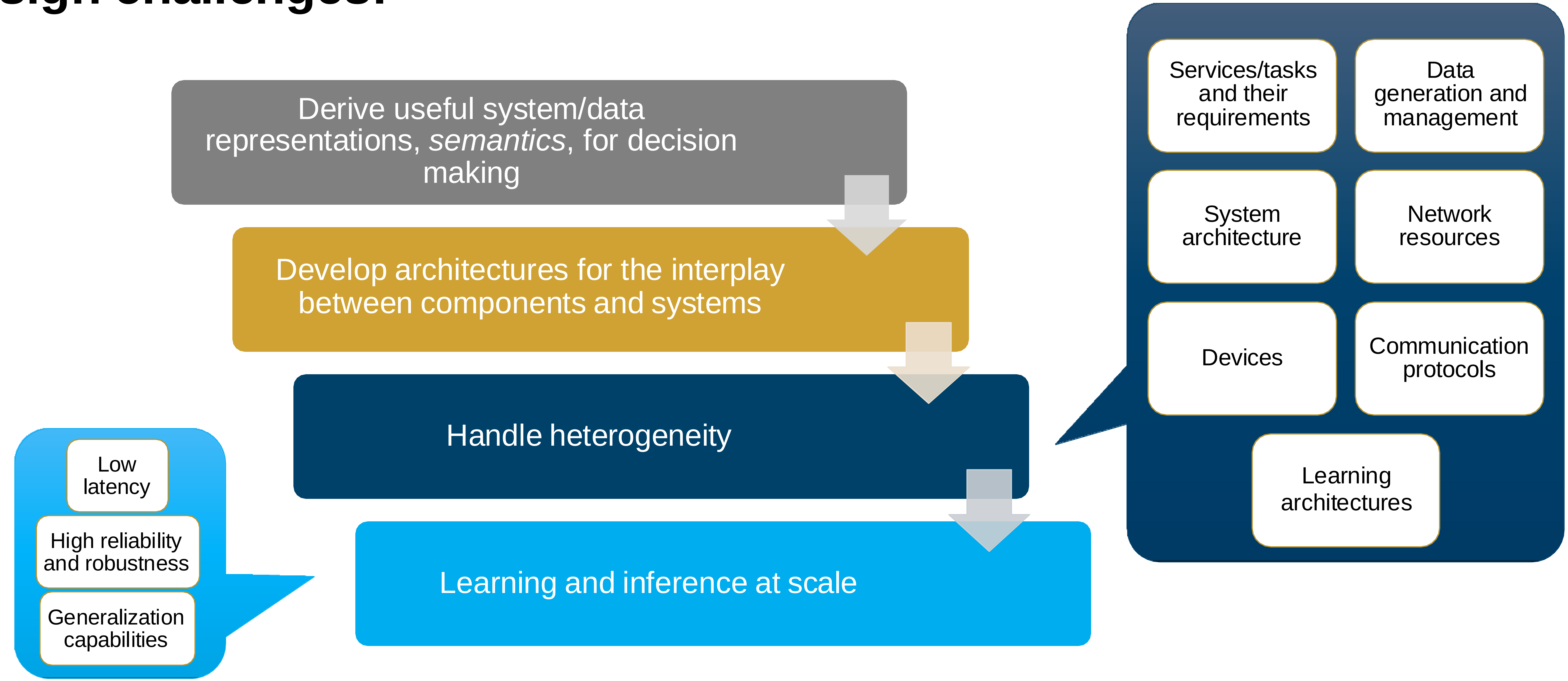


- S. Moreschini, F. Pecorelli, X. Li, S. Naz, D. Hästbacka and D. Taibi, "Cloud Continuum: The Definition," in IEEE Access, vol. 10, pp. 131876-131886, 2022
- A. Droob, D.Morratz, F.Langkilde Jakobsen, J. Carstensen, M.Mathiesen, R.Bohnstedt, M.Albano, S.Moreschini, D.Taibi. Fault Tolerant Horizontal Computation Offloading," 2023 IEEE International Conference on Edge Computing and Communications. 2023
- Cognitive Cloud: The Definition. 19th International Conference on Distributed Computing and Artificial Intelligence, 2023





Design challenges:



6G Software for Extremely Distributed and Heterogeneous Massive Networks of Connected Devices

Our Research on Edge Computing

”Investigate sustainable software solutions that are robust, scalable, and energy-efficient.”

1. Implement energy-aware (EA) orchestration and scalability models
2. Software architecture for energy-aware extremely distributed systems

May 2023 - April 2026

BUSINESS FINLAND

Energy-Aware orchestration and scalability models

- Investigated auction-based orchestration methods
  - Isomorphic implementations
- Investigated multi-layered cloud-native systems architecture reconstruction

Energy-Aware software architecture

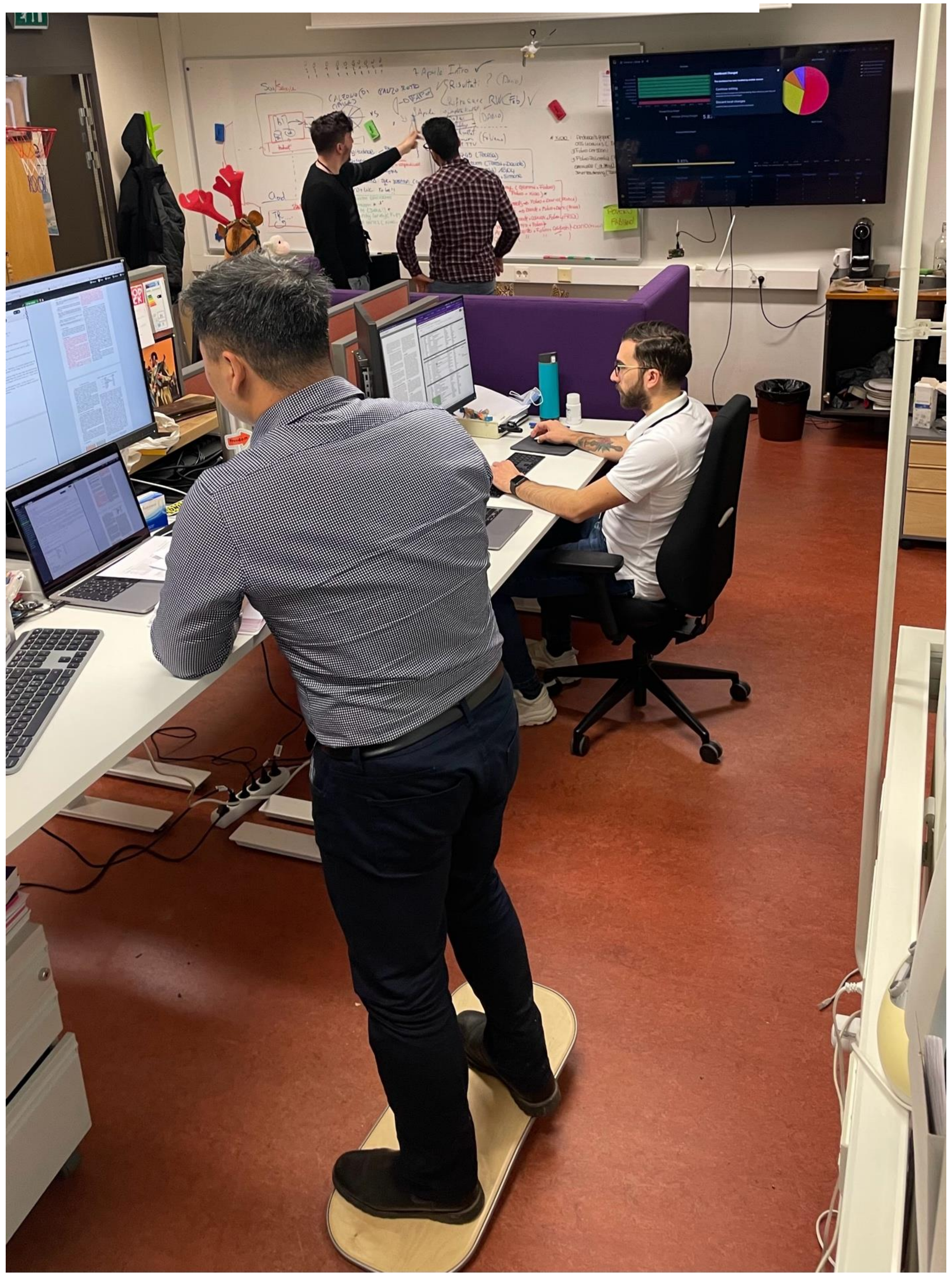
- Prototypes for next generation software applications
  - Decentralised deployment in the continuum
  - SW Architecture for AI-Based models

Demo space: Multimodal Sensing Lab

Demo Space: Office Wellbeing Use Case

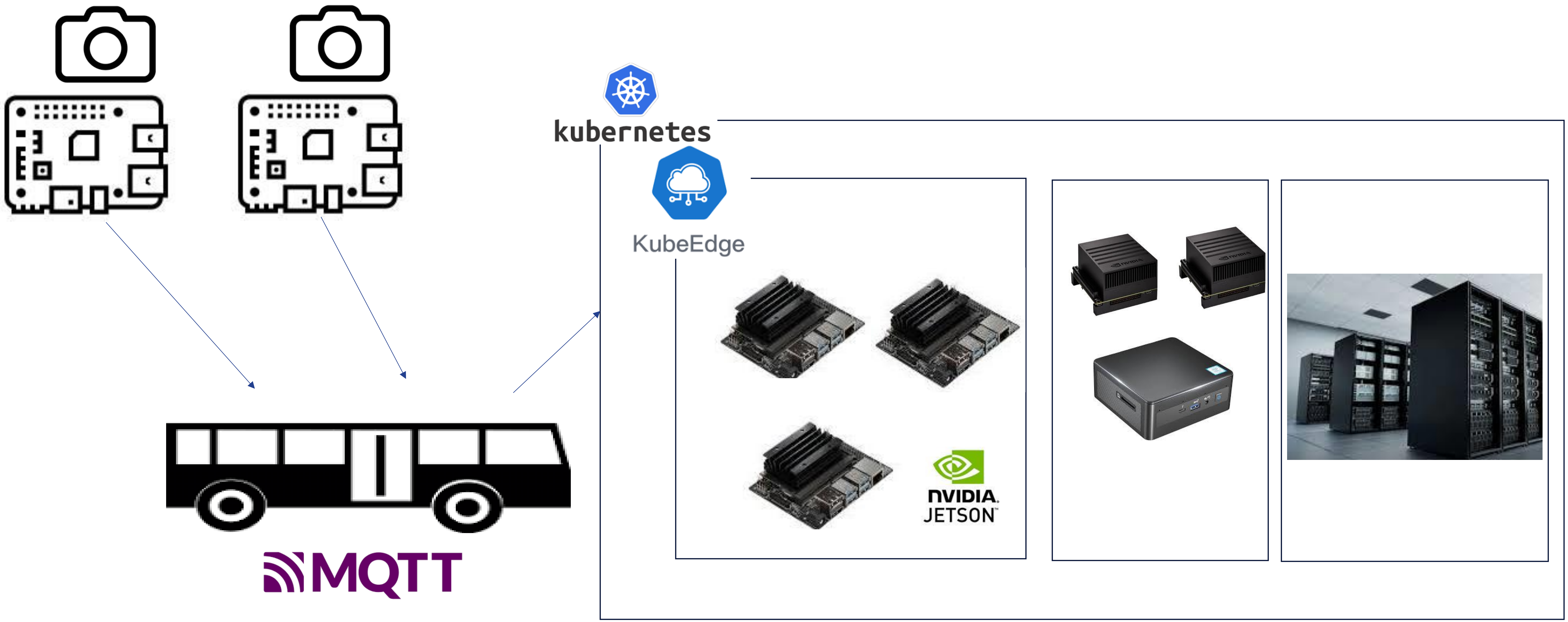
Use case of Edge-AI on office wellbeing

- Posture monitoring (cameras)
- Torso mobility (Balance Boards)
- Time Standing/seating/balance board
- Heart rate (stress)

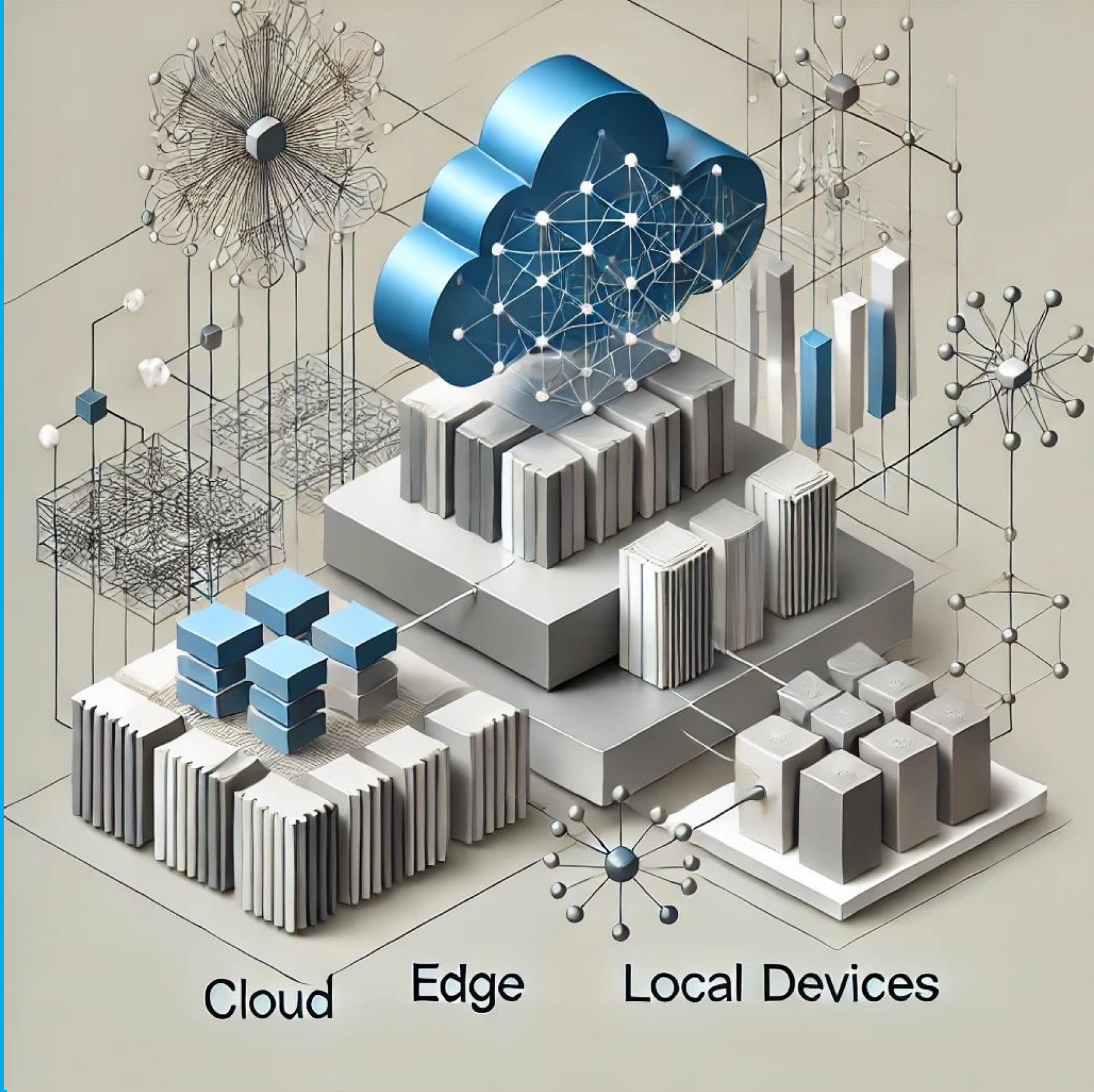


BUSINESS FINLAND

Our Multimodal Lab Architecture



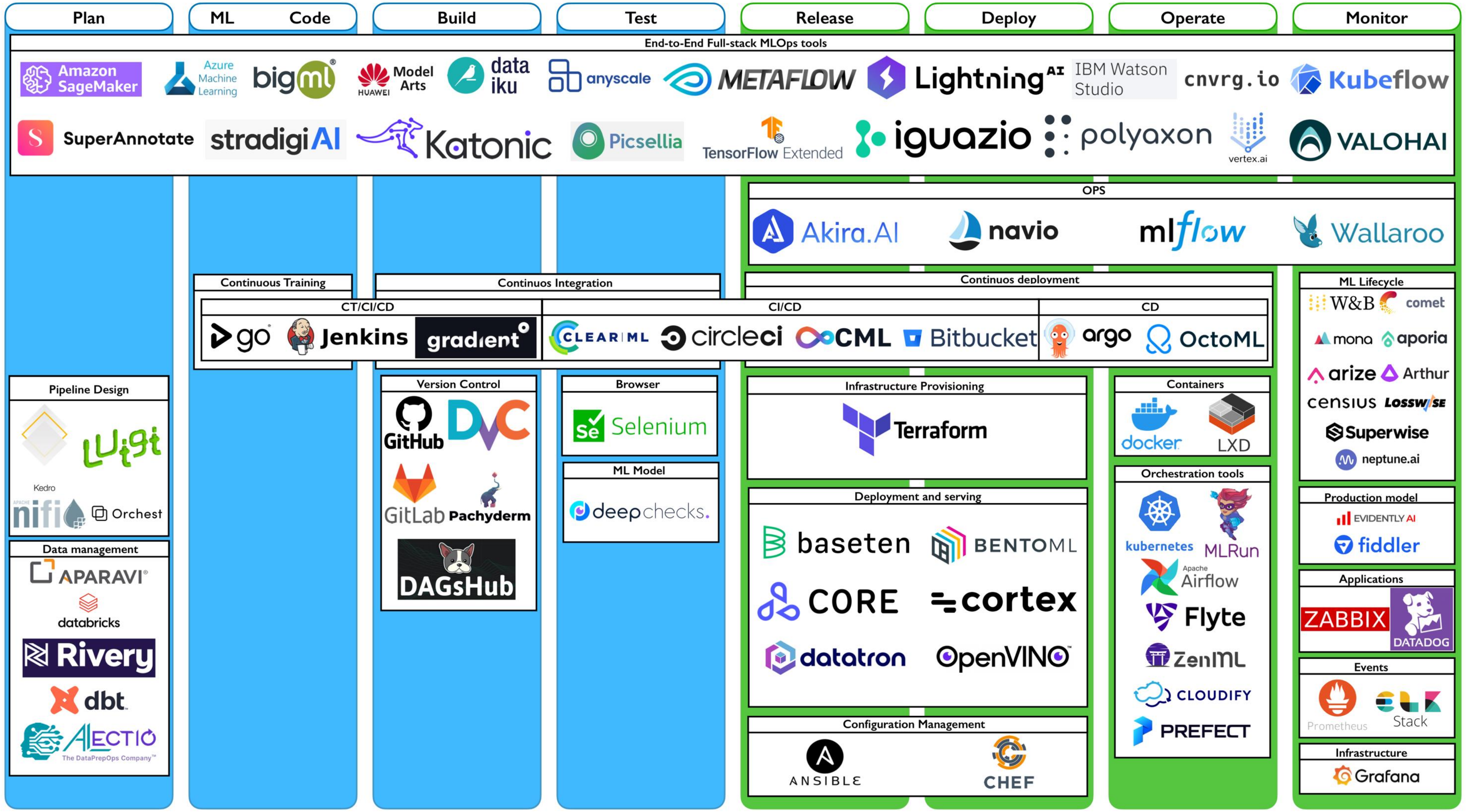
Architecting Edge-AI Systems



Challenges

- No standardized AI-based system reference architecture
- No standard tools pipeline
- Ad-hoc research studies

Tools Pipelines: AIOps tool Map



Sergio Moreschini, Elham Younesian, David Hästbacka, Michele Albano, Jiří Hošek, Davide Taibi,Edge to cloud tools: A Multivocal Literature Review, **Journal of Systems and Software**. Volume 210, 2024,

Best Practices and Bad Practices

Microservices


Processes, Motivations, and Issues for Migrating to Microservices Architectures: An Empirical Investigation

Davide Taibi, Valentina Lenarduzzi, and Claus Pahl

Serverless


Serverless Edge Computing: Vision and Challenges

Serverless: What it Is, What to Do and What Not to Do

Jussi Nupponen, Davide Taibi

Serverless Computing—Where Are We Now, and Where Are We Heading?

Davide Taibi, Josef Spillner, Konrad Wawruch

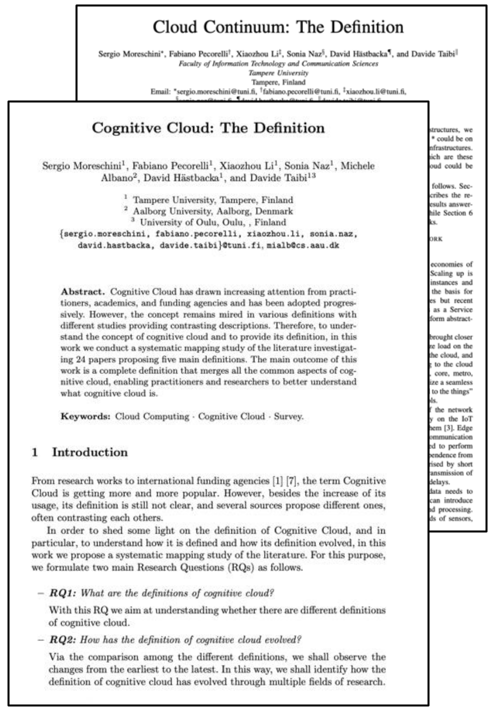
Micro-Frontends


Information and Software Technology

Motivations, benefits, and issues for adopting Micro-Frontends: A Multivocal Literature Review

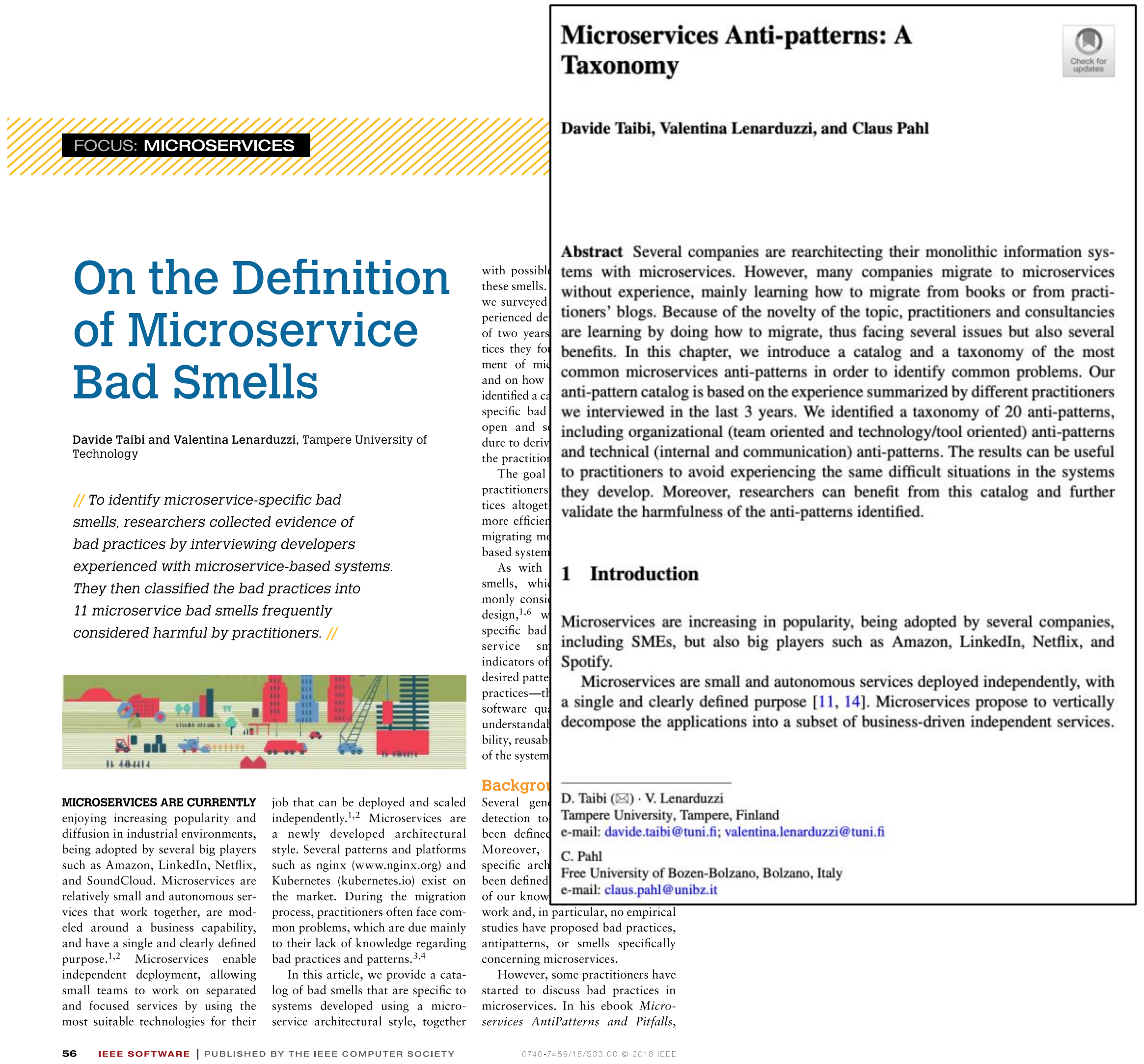
Severi Peltonen, Luca Mezzalira, Davide Taibi

Cognitive and Continuum Cloud


Cloud Continuum: The Definition

Cognitive Cloud: The Definition

Patterns and Anti-Patterns



FOCUS: MICROSERVICES

On the Definition of Microservice Bad Smells

Davide Taibi and Valentina Lenarduzzi, Tampere University of Technology

// To identify microservice-specific bad smells, researchers collected evidence of bad practices by interviewing developers experienced with microservice-based systems. They then classified the bad practices into 11 microservice bad smells frequently considered harmful by practitioners. //

Microservices Anti-patterns: A Taxonomy

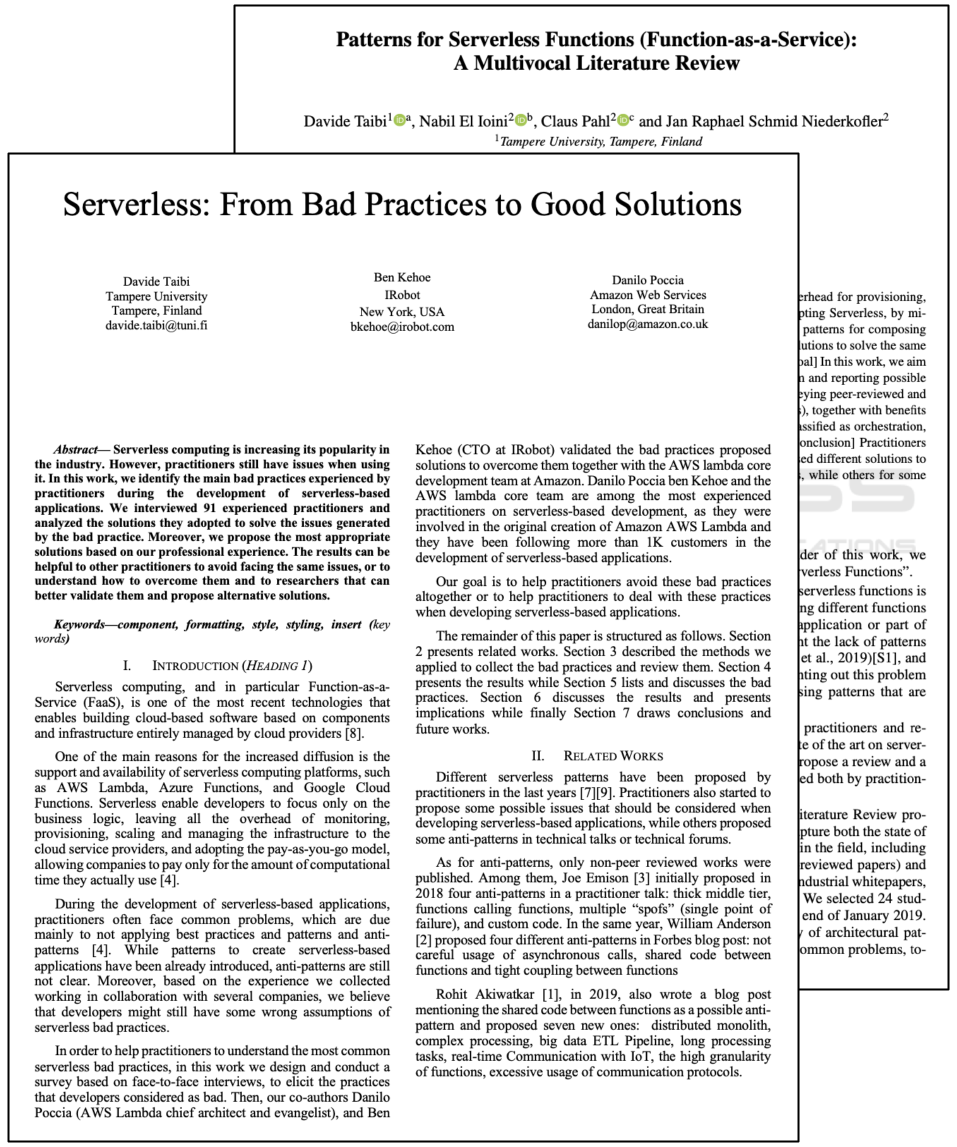
Davide Taibi, Valentina Lenarduzzi, and Claus Pahl

Abstract Several companies are rearchitecting their monolithic information systems with microservices. However, many companies migrate to microservices without experience, mainly learning how to migrate from books or from practitioners' blogs. Because of the novelty of the topic, practitioners and consultancies are learning by doing how to migrate, thus facing several issues but also several benefits. In this chapter, we introduce a catalog and a taxonomy of the most common microservices anti-patterns in order to identify common problems. Our anti-pattern catalog is based on the experience summarized by different practitioners we interviewed in the last 3 years. We identified a taxonomy of 20 anti-patterns, including organizational (team oriented and technology/tool oriented) anti-patterns and technical (internal and communication) anti-patterns. The results can be useful to practitioners to avoid experiencing the same difficult situations in the systems they develop. Moreover, researchers can benefit from this catalog and further validate the harmfulness of the anti-patterns identified.

1 Introduction

Microservices are increasing in popularity, being adopted by several companies, including SMEs, but also big players such as Amazon, LinkedIn, Netflix, and Spotify.

Microservices are small and autonomous services deployed independently, with a single and clearly defined purpose [11, 14]. Microservices propose to vertically decompose the applications into a subset of business-driven independent services.



Patterns for Serverless Functions (Function-as-a-Service): A Multivocal Literature Review

Davide Taibi, Nabil El Ioini, Claus Pahl and Jan Raphael Schmid Niederkofler

Serverless: From Bad Practices to Good Solutions

Davide Taibi, Tampere University, Tampere, Finland
Ben Kehoe, IRobot, New York, USA
Danilo Poccia, Amazon Web Services, London, Great Britain

Technology Assessment Framework

Information and Software Technology 137 (2021) 106600

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

From monolithic systems to Microservices: An assessment framework

Florian Auer[a,*], Valentina Lenarduzzi[b], Michael Felderer[a,c], Davide Taibi[d]

[a] University of Innsbruck, Austria
[b] LUT University, Finland
[c] Blekinge Institute of Technology, Sweden
[d] Tampere University, Finland

A R T I C L E I N F O

Keywords:
Microservices
Cloud migration
Software measurement

A B S T R A C T

Context: Re-architecting monolithic systems with Microservices-based architecture is a common trend. Various companies are migrating to Microservices for different reasons. However, making such an important decision like re-architecting an entire system must be based on real facts and not only on gut feelings.
Objective: The goal of this work is to propose an evidence-based decision support framework for companies that need to migrate to Microservices, based on the analysis of a set of characteristics and metrics they should collect before re-architecting their monolithic system.
Method: We conducted a survey done in the form of interviews with professionals to derive the assessment framework based on Grounded Theory.
Results: We identified a set consisting of information and metrics that companies can use to decide whether to migrate to Microservices or not. The proposed assessment framework, based on the aforementioned metrics, could be useful for companies if they need to migrate to Microservices and do not want to run the risk of failing to consider some important information.

1. Introduction

Microservices are becoming more and more popular. Big players such as Amazon,[1] Netflix,[2] Spotify,[3] as well as small and medium-sized enterprises are developing Microservices-based systems [1].

Microservices are autonomous services deployed independently, with a single and clearly defined purpose [2]. Microservices propose vertically decomposing applications into a subset of business-driven independent services. Each service can be developed, deployed, and tested independently by different development teams and using different technology stacks. Microservices have a variety of different advantages. They can be developed in different programming languages, can scale independently from other services, and can be deployed on the hardware that best suits their needs. Moreover, because of their size, they are easier to maintain and more fault-tolerant since the failure of one service will not disrupt the whole system, which could happen in a monolithic system. However, the migration to Microservices is not an easy task [1,3]. Companies commonly start the migration without any experience with Microservices, only rarely hiring a consultant to support them during the migration [1,3].

Various companies are adopting Microservices since they believe that it will facilitate their software maintenance. In addition, companies hope to improve the delegation of responsibilities among teams. Furthermore, there are still some companies that refactor their applications with a Microservices-based architecture just to follow the current trend [1,3].

The economic impact of such a change is not negligible, and taking such an important decision to re-architect an existing system should always be based on solid information, so as to ensure that the migration will allow achieving the expected benefits.

In this work, we propose an evidence-based decision support framework to allow companies, and especially software architects, to make their decision on migrating monolithic systems to Microservices based on the evaluation of a set of objective measures regarding their systems. The framework supports companies in discussing and analyzing potential benefits and drawbacks of the migration and re-architecting process.

* Corresponding author.
E-mail addresses: florian.auer@uibk.ac.at (F. Auer), valentina.lenarduzzi@lut.fi (V. Lenarduzzi), michael.felderer@uibk.ac.at (M. Felderer), davide.taibi@tuni.fi (D. Taibi).

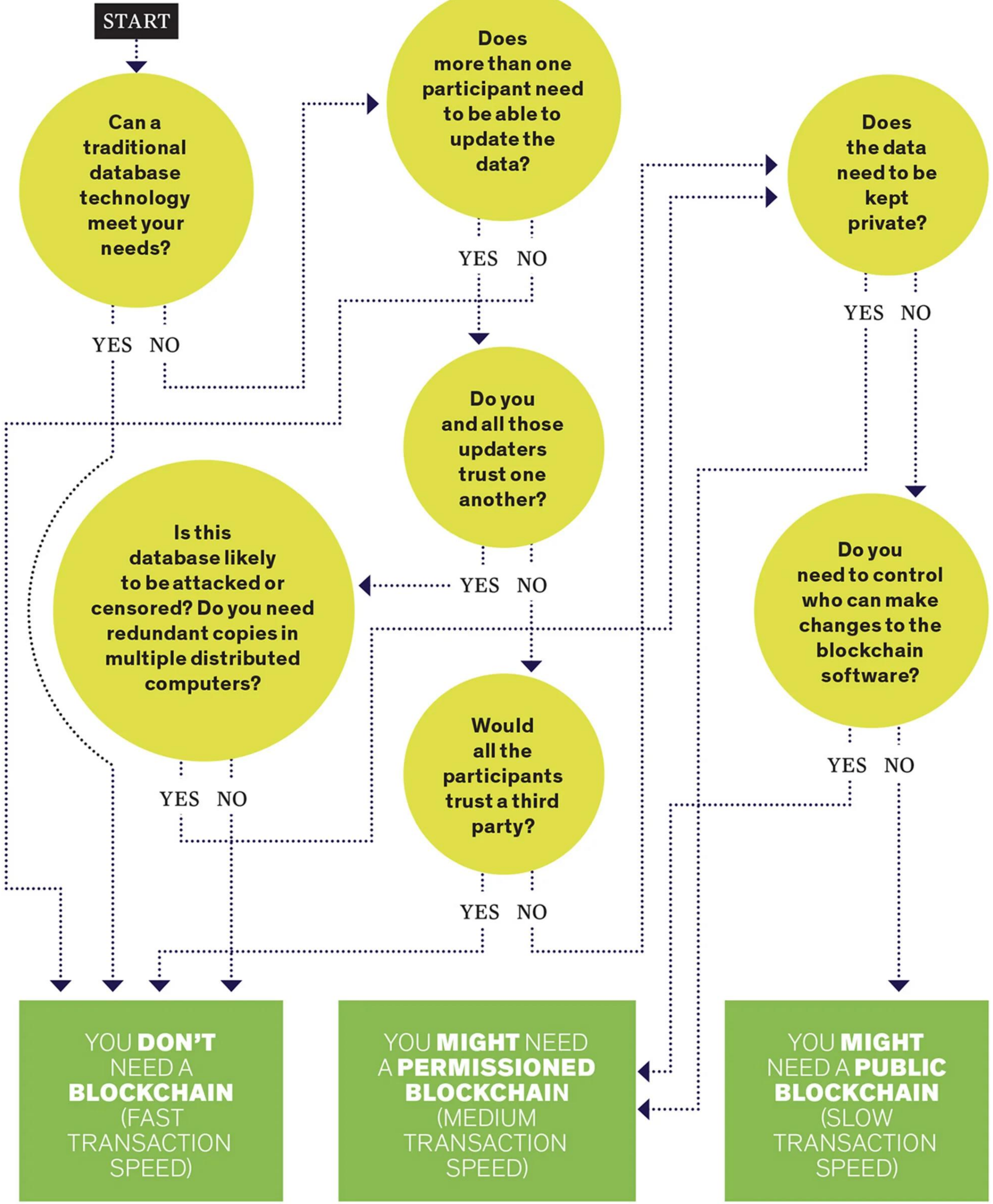
[1] https://gigaom.com/2011/10/12/419-the-biggest-thing-amazon-got-right-the-platform/
[2] http://nginx.com/blog/Microservices-at-netflix-architectural-best-practices/
[3] www.infoq.com/presentations/linkedin-Microservices-urn

https://doi.org/10.1016/j.infsof.2021.106600
Received 31 January 2020; Received in revised form 1 March 2021; Accepted 14 April 2021
Available online 30 April 2021

Technology Assessment

My favorite example (Blockchain)

M.E. Pech DO YOU NEED A BLOCKCHAIN?
IEEE Spectrum. September 2017.
https://spectrum.ieee.org/do-you-need-a-blockchain

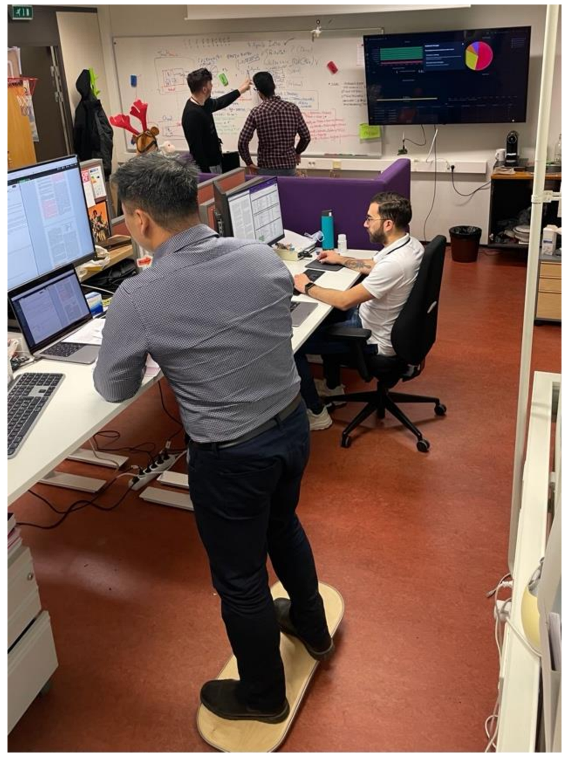


Future Goals

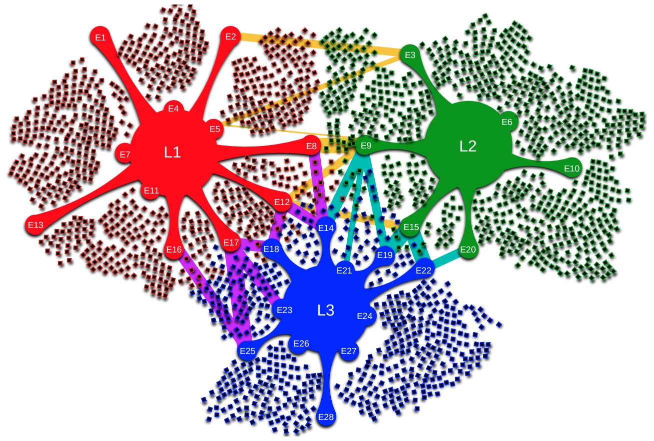
- Anomaly detection
- Energy analysis

- Orchestration optimization
  - Energy
  - Performance
  - QoS
  - …
- Quantum computing into the cloud continuum

- Visualizations



BUSINESS FINLAND



Edge Nodes Upgrade Optimization

NOKIA

Conclusions

Morale of the story!

- AI-Based Systems are not only AI

- Ai Devs and Ops need to be synchronized

- Orchestration is fundamental

- AI Deployment is expensive
  - AI-Ops is not enough!

- AI Operational costs might be unexpectedly high

Thanks!

Are you interested in our research?
Do you wish to work with us?

Get in touch!

davide.taibi@oulu.fi
https://m3s-cloud.github.io