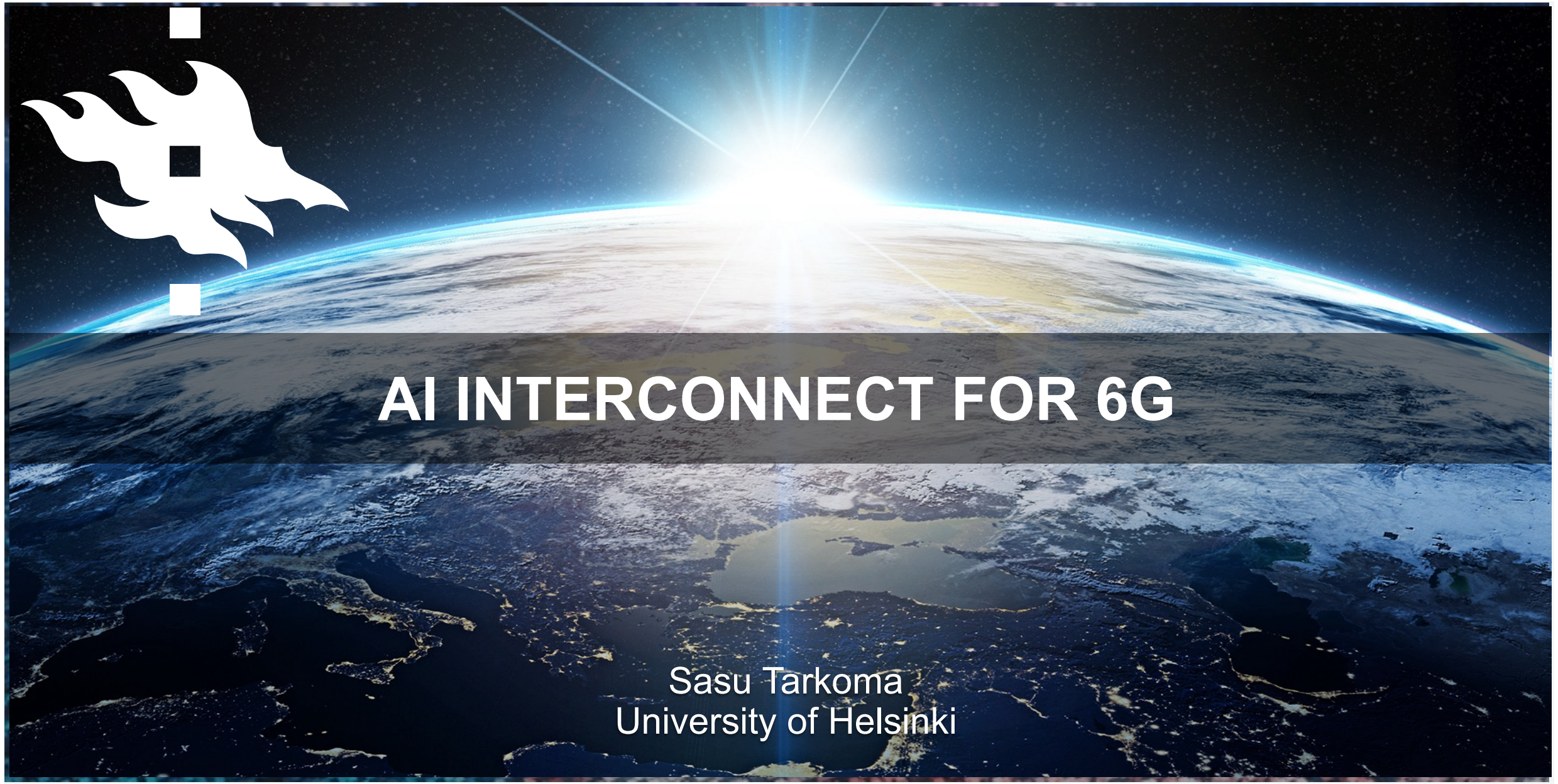


AI INTERCONNECT FOR 6G

Sasu Tarkoma
University of Helsinki

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

CONTENTS

- BACKGROUND
- TECHNOLOGY MIX: 6G AND AI
- AI INTERCONNECT AND NEURAL PUB/SUB
- EDUCATION: 6G MOOC

6G Finland

A **coalition** of world-leading Finnish companies and research institutes

Finland’s 6G network of excellence.

6GFinland.fi

NOKIA CENTER FOR ADVANCED RESEARCH (NCAR) OVERVIEW

Publications

+70 scientific papers published on high-ranking journals and conference since 2016 (https://tinyurl.com/5asazpbz)
Patents and patent applications

Events

Workshops, conferences and other dissemination events co-organized together with the FCAI and the 6G Flagship

FCAI Finnish Center for Artificial Intelligence

6G FLAGSHIP UNIVERSITY OF OULU

Externally Funded Projects and International Collaborations

Multiple national and international research projects
Key project: Academy-NSF Lean6G with professor Leandros Tassiulas at Yale University

Education

1-2 PhD degrees annually and many students working on theses.
Online education (MOOCs)

People

An international team of **+12 members** including faculty, university researchers, and postdocs currently involved in NCAR-related activities

Research

5 key research themes investigated:

- Environmental Sensing
- Edge Computing and IoT
- Low Latency Transport Protocols
- Cloud Computing and NFV
- Distributed and Cognitive Processing

https://www.helsinki.fi/en/networks/6g-research

NORDIC UNIVERSITY COOPERATION ON EDGE INTELLIGENCE (NUEI)

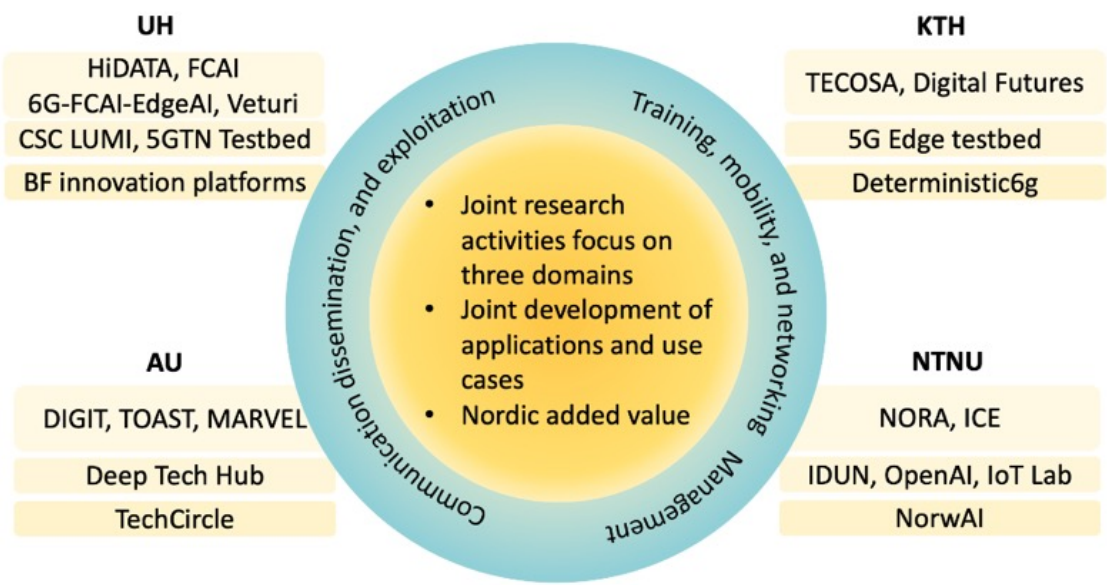
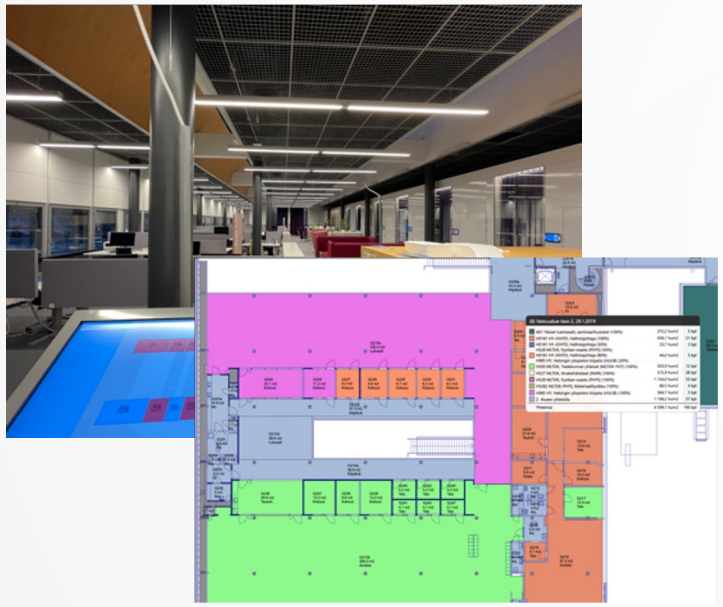


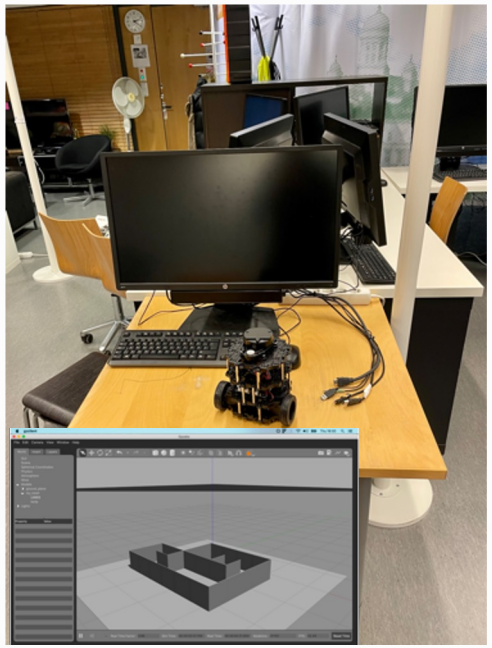
Figure 1. NUEI Cooperation mechanism

- Four partners:
- University of Helsinki (UH), lead
- KTH Royal Institute of Technology (KTH)
- Aarhus University (AU)
- Norwegian University of Science and Technology (NTNU).
- Research focus on Edge Intelligence of each partner (shown in Figure 1).
- Coherent, joint strategic approach to take a good advantage of joint cross-border collaboration.

KUMPULA CAMPUS AS ”PLAYGROUND” FOR OUR RESEARCH



UbiKampus



Software Factory

NodesLab

Heterogeneity ⟷ **Multiple Connectivity Technologies** ⟷ **Highly Dynamic**

CONTENTS

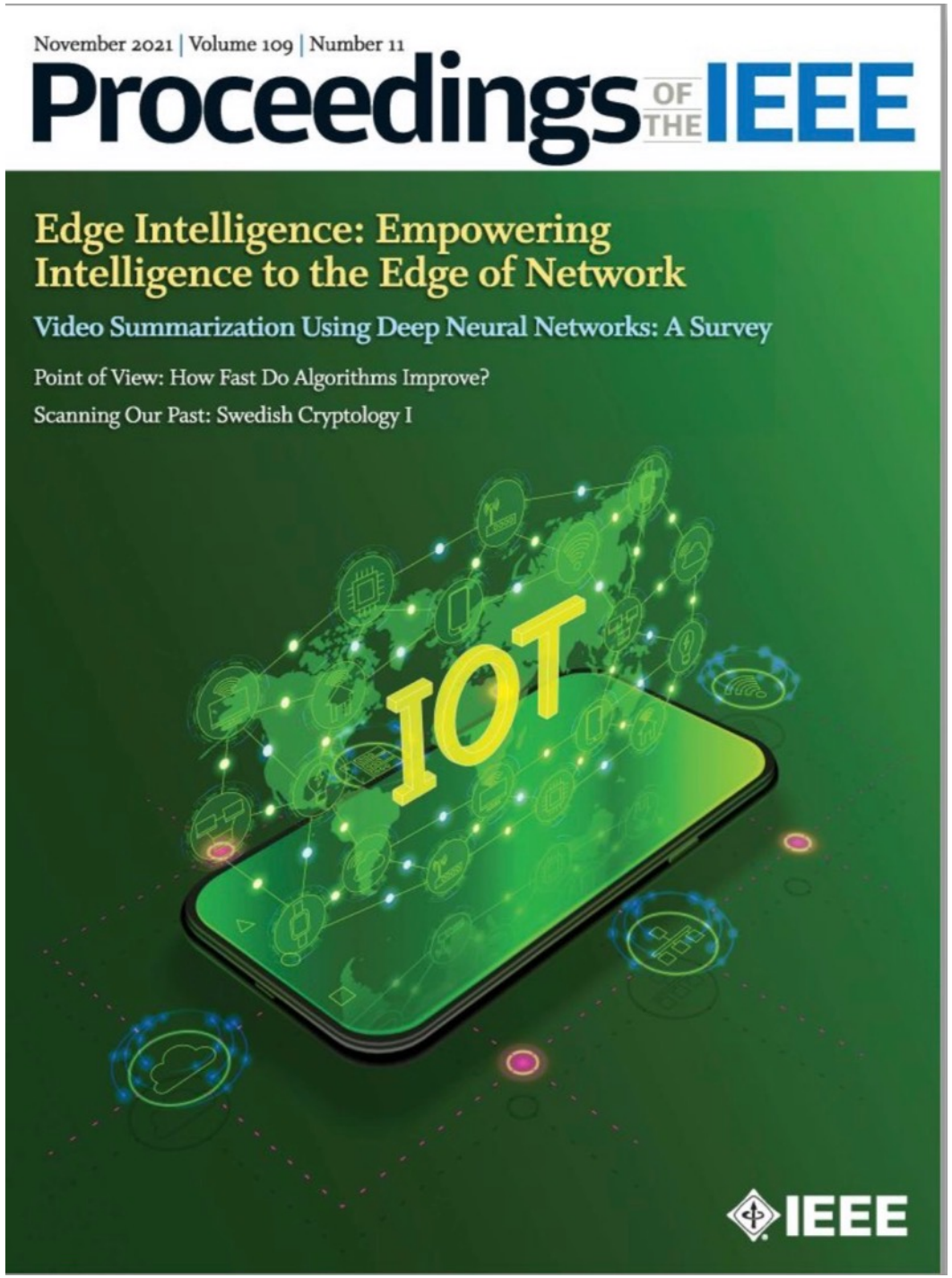
- BACKGROUND
- TECHNOLOGY MIX: 6G AND AI
- AI INTERCONNECT AND NEURAL PUB/SUB
- EDUCATION: 6G MOOC

KEY DEVELOPMENTS THAT ARE CONVERGING

Cloud computing: distributed clouds, general edge computing, DevOps

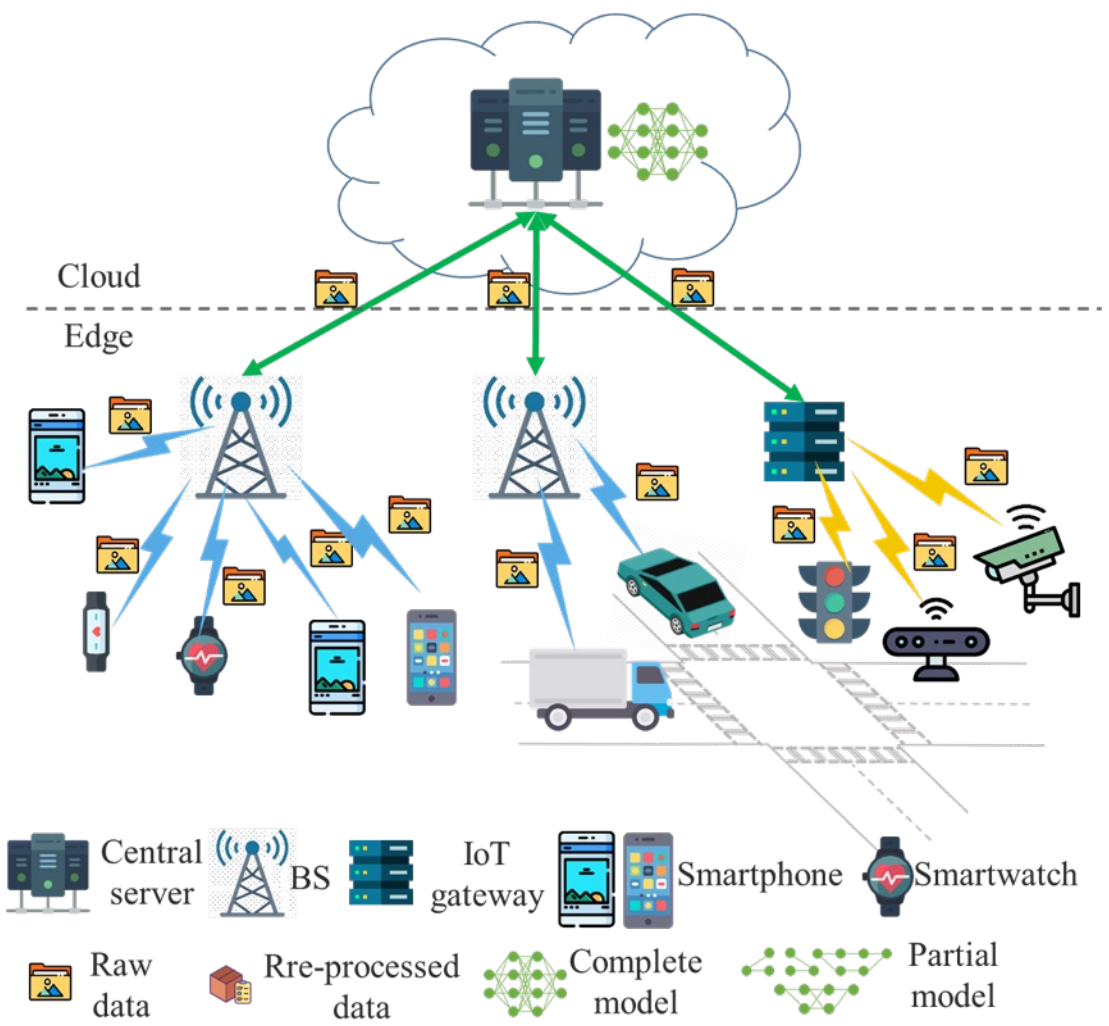
Cellular networks: cloud RAN, O-RAN, network slicing, Mobile Edge Computing (MEC), private networks

AI: distributed techniques (federated learning, split learning), transfer learning, differential privacy, LLMs, regulation

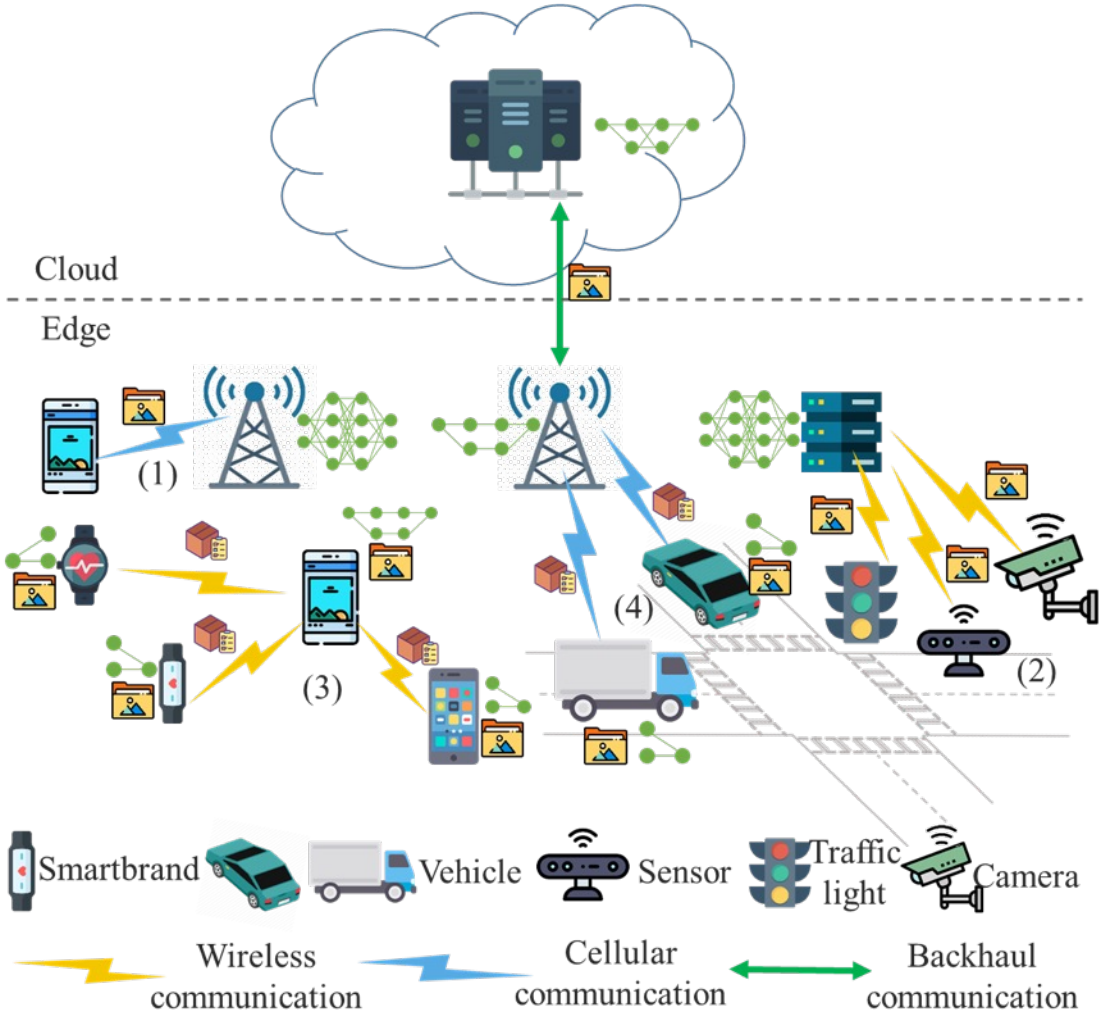


Edge Intelligence: Empowering Intelligence to the Edge of Network
D Xu, T Li, Y Li, X Su, S Tarkoma, T Jiang, J Crowcroft, P Hui
Proceedings of the IEEE 109 (11), 1778-1837.
(60 pages)

Centralized cloud for AI

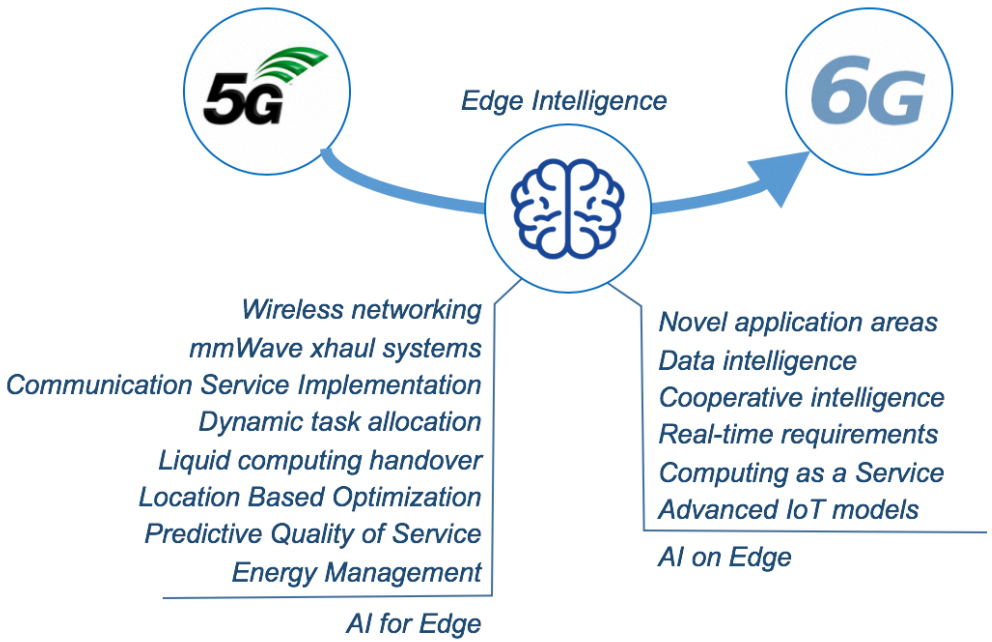


Edge Intelligence

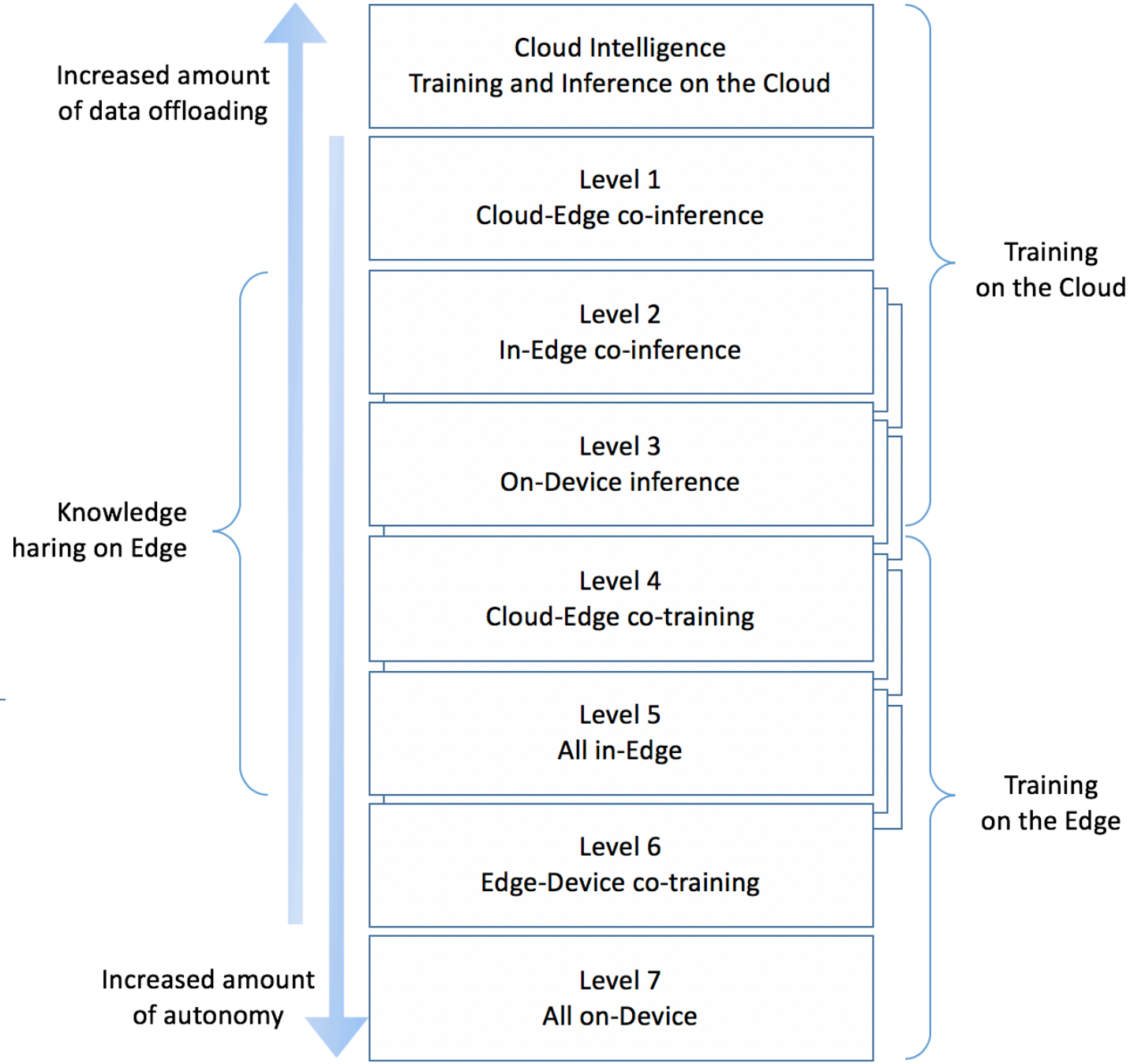


Edge Intelligence: Empowering Intelligence to the Edge of Network. D Xu, T Li, Y Li, X Su, S Tarkoma, T Jiang, J Crowcroft, P Hui Proceedings of the IEEE 109 (11), 1778-1837.

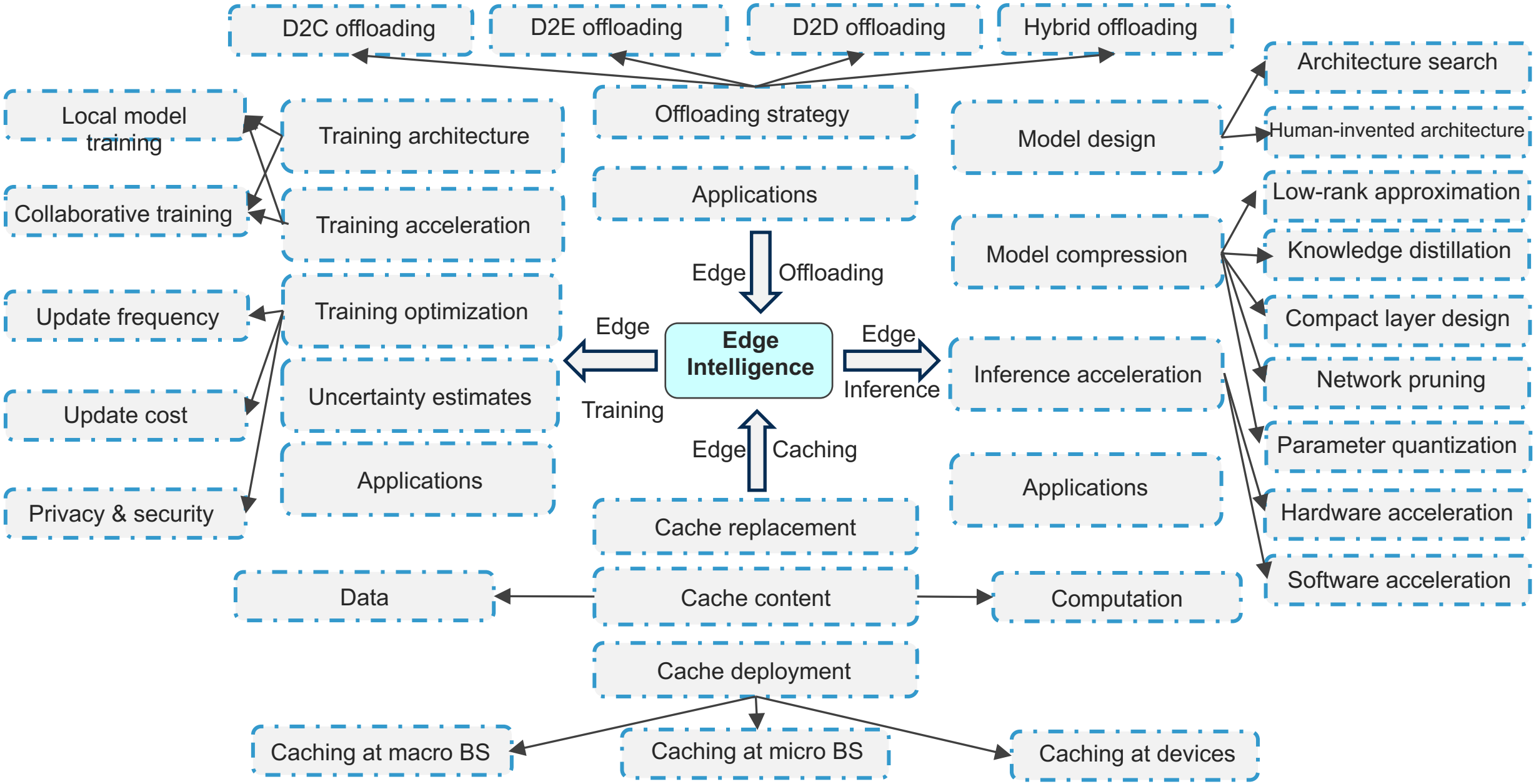
6G WHITEPAPER ON EDGE INTELLIGENCE



Whitepaper available at:
https://arxiv.org/abs/2004.14850



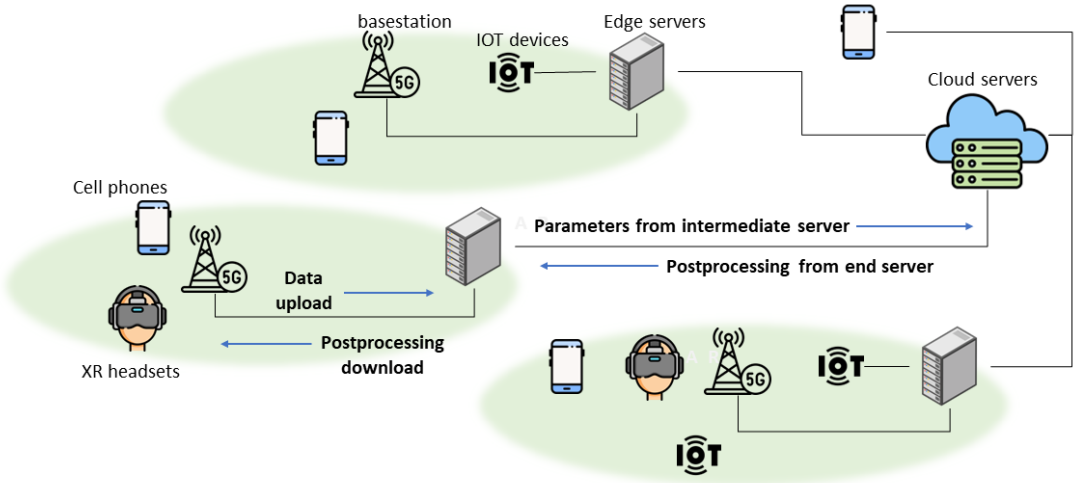
A Taxonomy


D2C offloading
D2E offloading
D2D offloading
Hybrid offloading
Offloading strategy
Local model training
Collaborative training
Training architecture
Training acceleration
Training optimization
Update frequency
Update cost
Privacy & security
Uncertainty estimates
Applications
Applications
Edge Offloading
Edge Intelligence
Edge Training
Edge Inference
Edge Caching
Model design
Architecture search
Human-invented architecture
Model compression
Low-rank approximation
Knowledge distillation
Compact layer design
Network pruning
Parameter quantization
Inference acceleration
Hardware acceleration
Software acceleration
Applications
Cache replacement
Cache content
Cache deployment
Data
Computation
Caching at macro BS
Caching at micro BS
Caching at devices

Fog Computing for Deep Learning with Pipelines

We demonstrate split learning and inference pipelines on device, cloudlet and cloud with a vision based use case (CIFAR-10).

If the fog hardware can be fully utilized, it will improve the throughput of heavy workloads.



- Fog is a hierarchical network structure.
- Resources in fog can be pooled with pipelines.
- Sparse: **S**tream **P**rocessing **A**rchitecture for **R**esource-**S**ubtle **E**nvironments[a]

[a] https://github.com/AnteronGitHub/sparse

KEY TASKS: LLMS AS CONTROLLERS

Task planning

Model selection

Task execution

Response generation

Tool dispatcher

Tool maker

Tool user

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. https://huggingface.co/papers/2303.17580
Large Language Models as Tool Makers. https://arxiv.org/pdf/2305.17126.pdf

Distributed GPTs in the Network

Drones and other autonomous sensors

Industrial systems

End-user Applications: AR and metaverse

Agents and Fine-tuned LLMs

Edge-cloud Continuum

Agents and Fine-tuned LLMs

Edge-cloud Continuum

Agents and Fine-tuned LLMs

Edge-cloud Continuum

Foundation LLM model repository and fine-tuning

Orchestration AI for agent, component, task placement

Interconnect and state management

Distributed AI based on foundation models that are incrementally fine-tuned (stacking property) for application domains

Orchestrator places and synchronizes models and state

Interconnect enables orchestration

CONTENTS

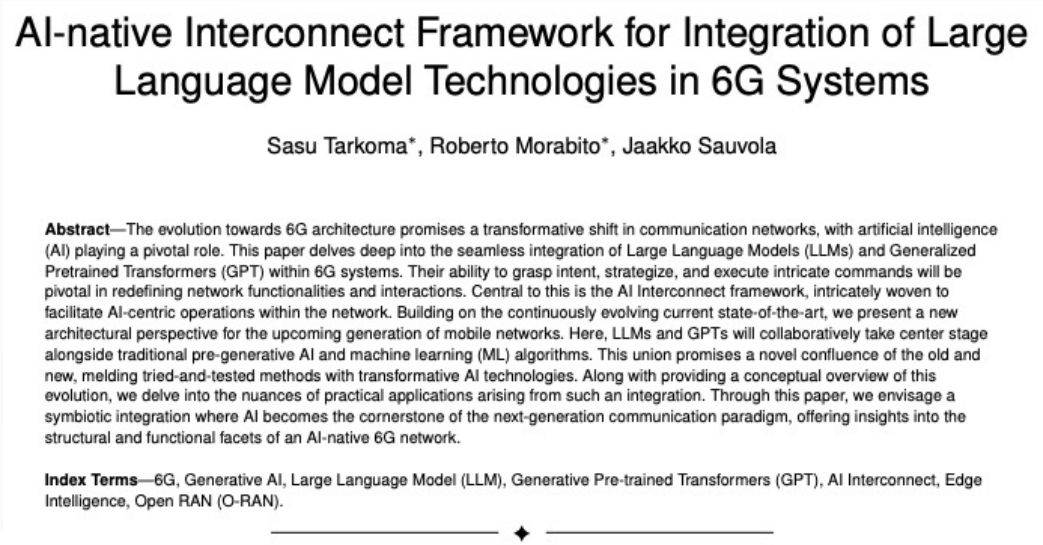
- BACKGROUND
- TECHNOLOGY MIX: 6G AND AI
- AI INTERCONNECT AND NEURAL PUB/SUB
- EDUCATION: 6G MOOC

AI INTERCONNECT – LARGE LANGUAGE MODELS FOR 6G SYSTEMS

AI-native Interconnect Framework for Integration of Large Language Model Technologies in 6G Systems

Sasu Tarkoma, Roberto Morabito, Jaakko Sauvola

https://arxiv.org/abs/2311.05842



AI-native Interconnect Framework for Integration of Large Language Model Technologies in 6G Systems

Sasu Tarkoma*, Roberto Morabito*, Jaakko Sauvola

Abstract—The evolution towards 6G architecture promises a transformative shift in communication networks, with artificial intelligence (AI) playing a pivotal role. This paper delves deep into the seamless integration of Large Language Models (LLMs) and Generalized Pretrained Transformers (GPT) within 6G systems. Their ability to grasp intent, strategize, and execute intricate commands will be pivotal in redefining network functionalities and interactions. Central to this is the AI Interconnect framework, intricately woven to facilitate AI-centric operations within the network. Building on the continuously evolving current state-of-the-art, we present a new architectural perspective for the upcoming generation of mobile networks. Here, LLMs and GPTs will collaboratively take center stage alongside traditional pre-generative AI and machine learning (ML) algorithms. This union promises a novel confluence of the old and new, melding tried-and-tested methods with transformative AI technologies. Along with providing a conceptual overview of this evolution, we delve into the nuances of practical applications arising from such an integration. Through this paper, we envisage a symbiotic integration where AI becomes the cornerstone of the next-generation communication paradigm, offering insights into the structural and functional facets of an AI-native 6G network.

Index Terms—6G, Generative AI, Large Language Model (LLM), Generative Pre-trained Transformers (GPT), AI Interconnect, Edge Intelligence, Open RAN (O-RAN).

AI INTERCONNECT – LARGE LANGUAGE MODELS FOR 6G SYSTEMS

The paper addresses **opportunities**, **challenges**, and possible **solutions** arising from the integration of Large Language Models, such as GPT, in the context of 6G systems

On-Device Cognitive Computing

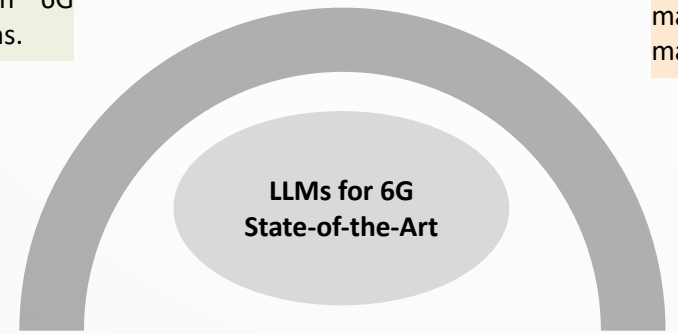
Beyond edge infrastructure: exploring the feasibility and implications of integrating LLMs directly into devices, paving the way for truly decentralized intelligence.

Edge Intelligence with LLMs

Delving into the intricacies of deploying LLMs in edge environments, addressing the challenges and innovations within 6G mobile edge computing (MEC) systems.

Demystifying Telco Through LLMs

Leveraging the natural language processing capabilities of LLMs to interpret, manage, and simplify complex telco terminologies and operations. This encapsulates the potential to make telecommunication operations and maintenance (OAM) accessible to non-experts.

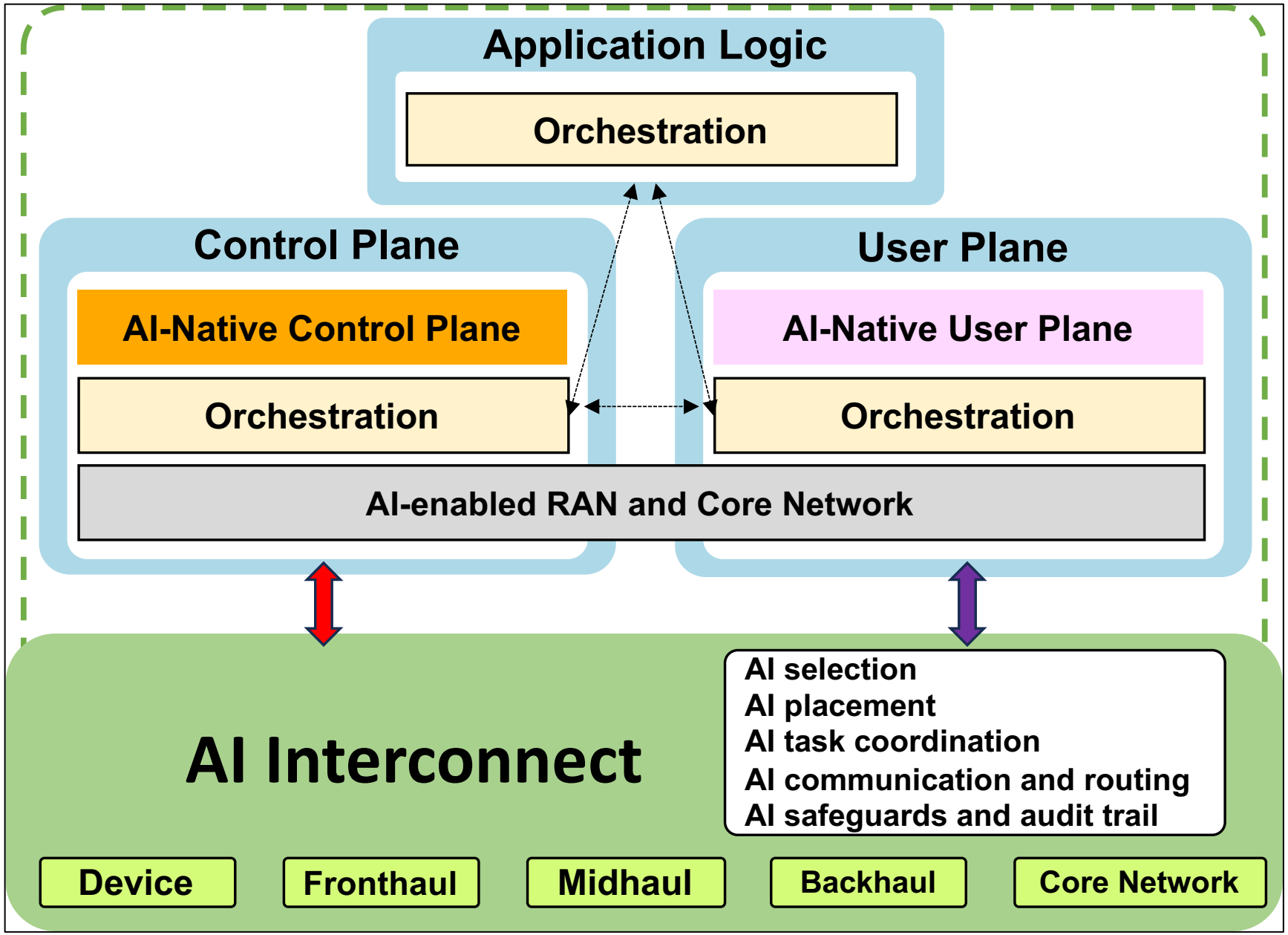


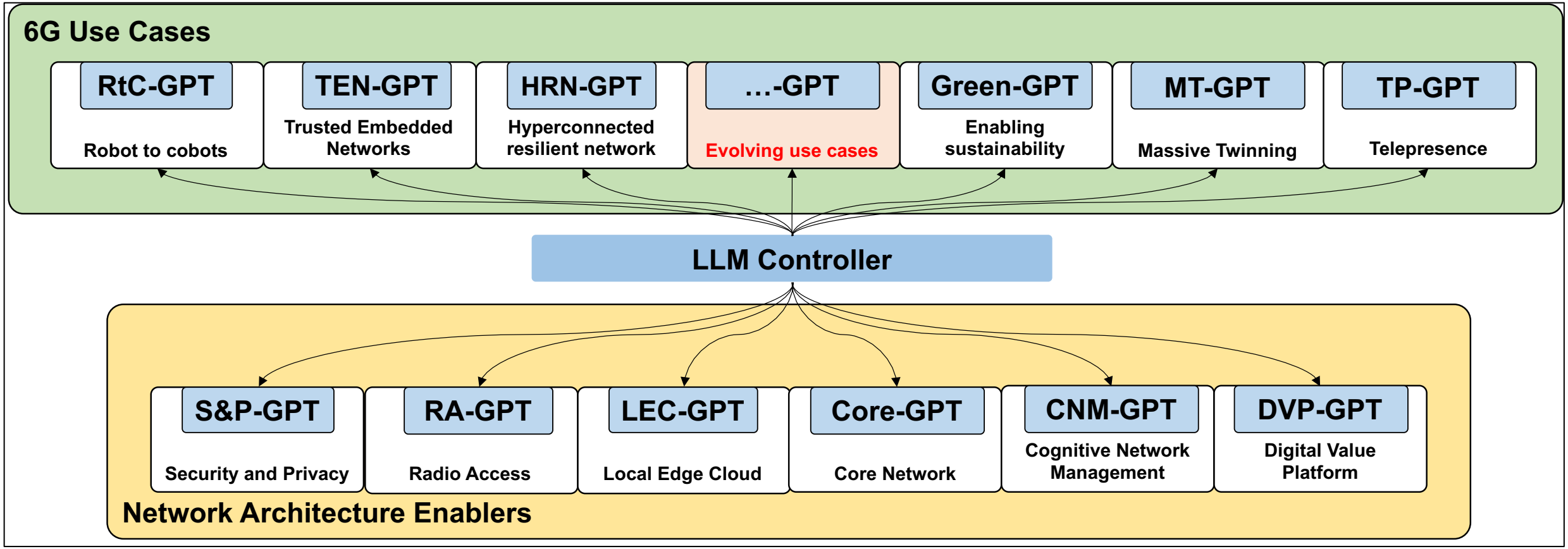
Generative AI in 6G

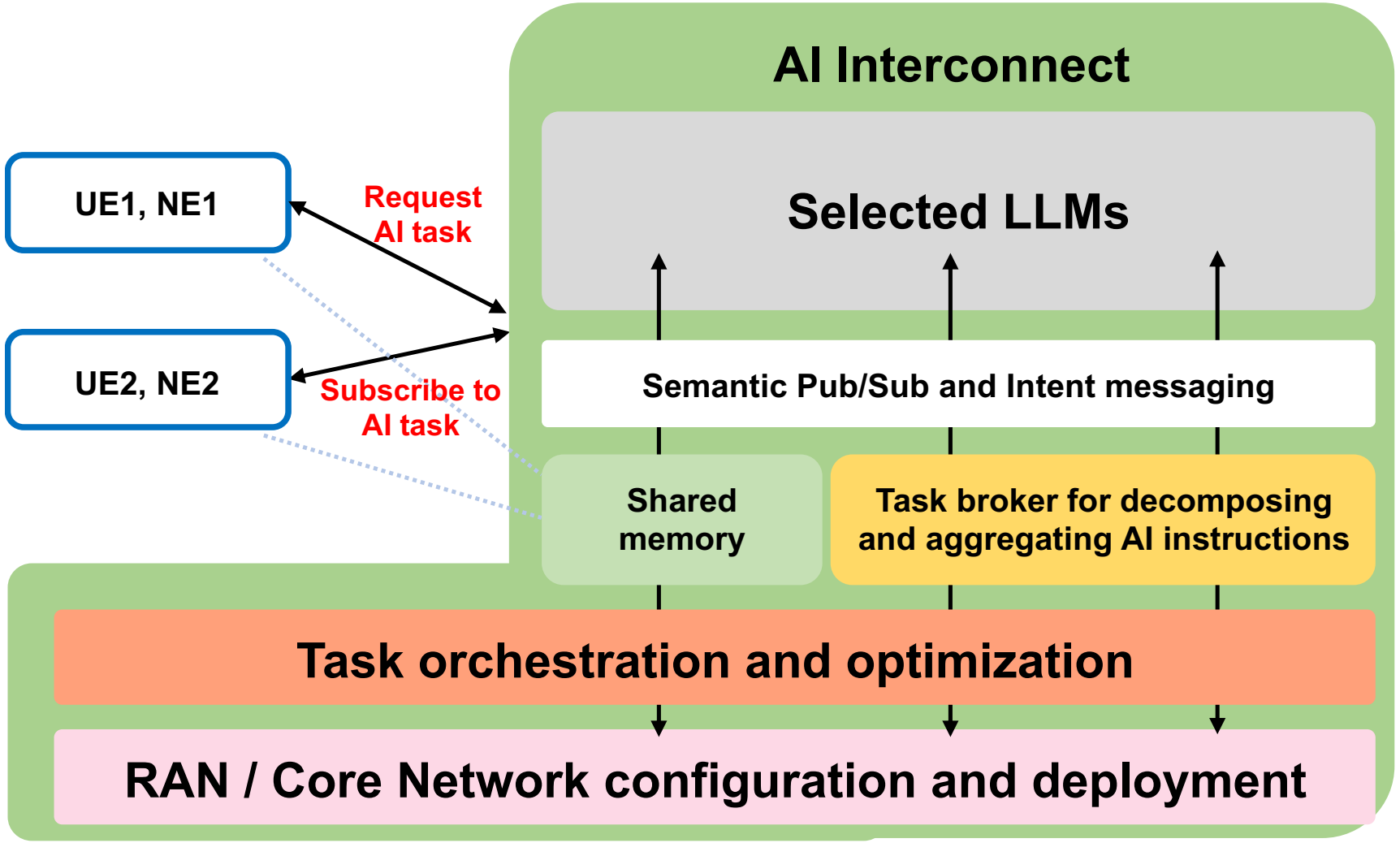
A holistic view exploring the synergy between generative large language models (LLMs) and 6G technologies, outlining the transformative potential for future wireless networks.

Diverse Applications in the 6G Ecosystem

Highlighting a spectrum of use cases from hardware acceleration, software optimization, IoT robotics, to dynamic network topology management, emphasizing the versatility of LLMs in the evolving 6G landscape.


Application Logic
Orchestration
Control Plane
AI-Native Control Plane
Orchestration
User Plane
AI-Native User Plane
Orchestration
AI-enabled RAN and Core Network
AI Interconnect
AI selection
AI placement
AI task coordination
AI communication and routing
AI safeguards and audit trail
Device
Fronthaul
Midhaul
Backhaul
Core Network


6G Use Cases
RtC-GPT
Robot to cobots
TEN-GPT
Trusted Embedded Networks
HRN-GPT
Hyperconnected resilient network
…-GPT
Evolving use cases
Green-GPT
Enabling sustainability
MT-GPT
Massive Twinning
TP-GPT
Telepresence
LLM Controller
S&P-GPT
Security and Privacy
RA-GPT
Radio Access
LEC-GPT
Local Edge Cloud
Core-GPT
Core Network
CNM-GPT
Cognitive Network Management
DVP-GPT
Digital Value Platform
Network Architecture Enablers


AI Interconnect
Selected LLMs
UE1, NE1
UE2, NE2
Request AI task
Subscribe to AI task
Semantic Pub/Sub and Intent messaging
Shared memory
Task broker for decomposing and aggregating AI instructions
Task orchestration and optimization
RAN / Core Network configuration and deployment

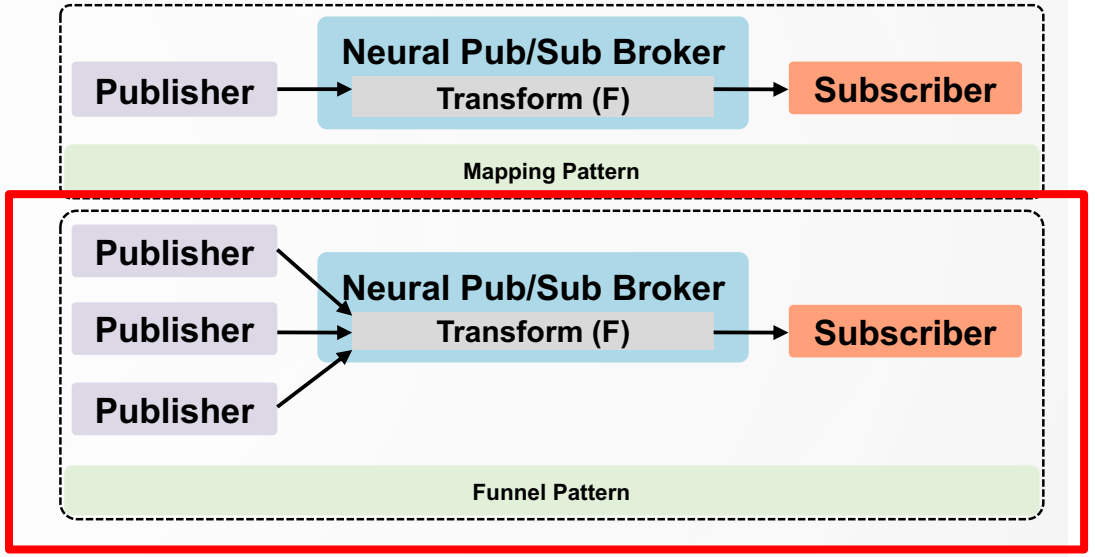
NEURAL PUB/SUB AND LARGE LANGUAGE MODELS

- Large Language Models (LLMs) like OpenAI's GPT at the forefront of 6G technological evolution.
- Enhancing network performance through intelligent management and optimization.
- **Development of a proof-of-concept using GPT for advanced network slicing management**.

Roberto Morabito et al. work in progress

NEURAL PUB/SUB FOR FUNNEL PATTERN

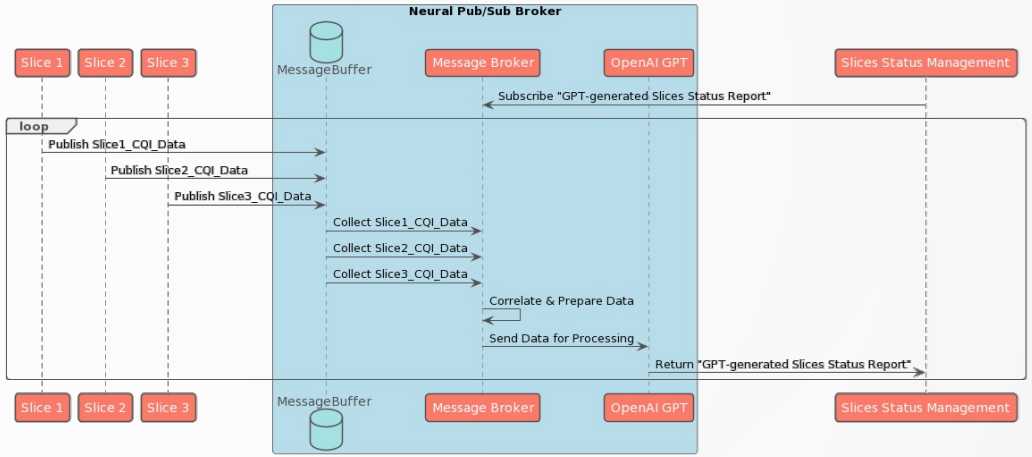
- Builds upon the mapping pattern for advanced data handling.
- Involves subscribing to one or more publications and applying a specific function (F).
- Function (F) determines the order and timing of data processing.
- Results in a single emitted publication based on the output of function F.



Funnel pattern workflow of a Neural Pub/Sub broker system designed to aggregate CQI data from three different slices.

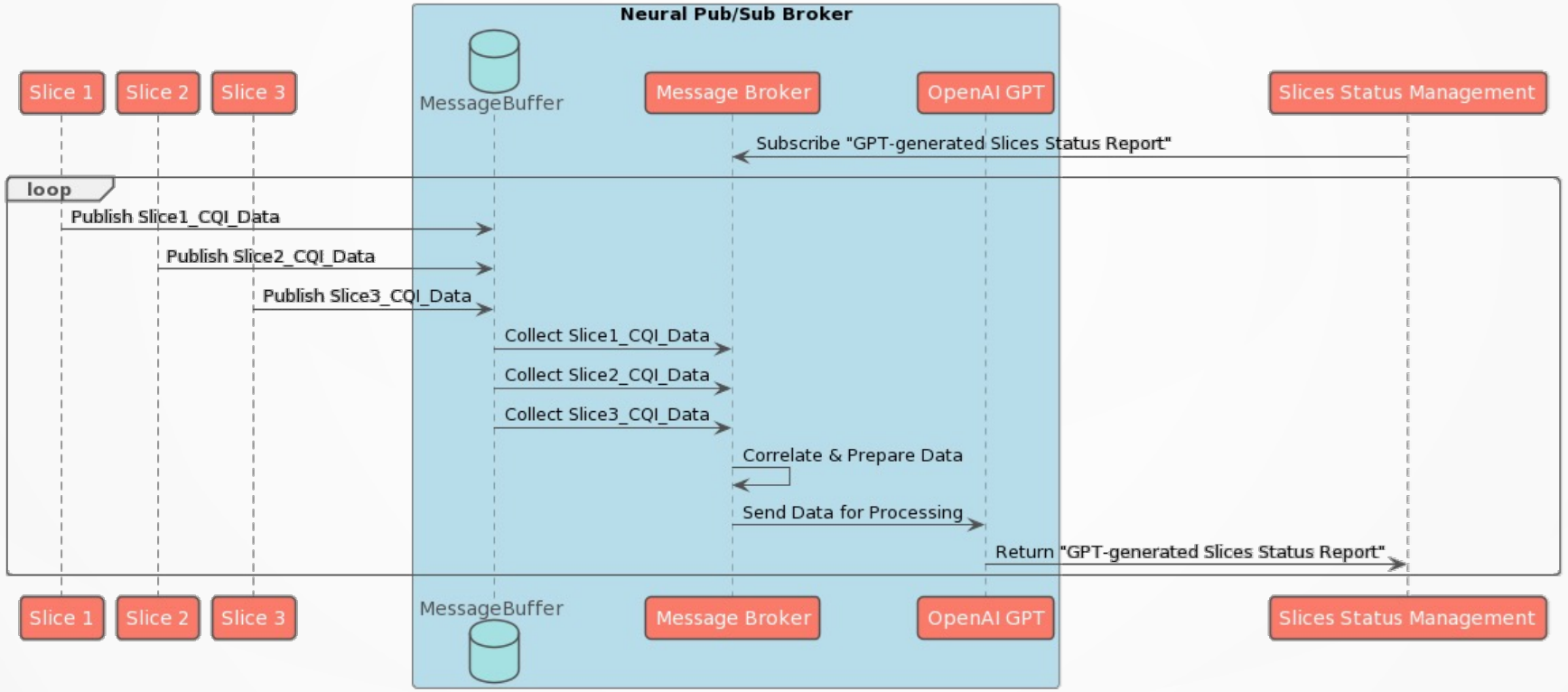
NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT

- Real-time monitoring of Channel Quality Indicator (CQI) across network slices.
- Automated generation of Slices Status Reports in natural language, offering actionable insights for network operators.
- Dynamic adjustments to network slices based on CQI data analysis to maintain optimal service quality.



Funnel pattern workflow of a Neural Pub/Sub broker system designed to aggregate CQI data from three different slices.

NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT



Funnel pattern workflow of a Neural Pub/Sub broker system designed to aggregate CQI data from three different slices.

```python
def query_openai(data):
    start_time = time.time()
    response = clientOpenAI.chat.completions.create(
        model="gpt-3.5-turbo-1106", #gpt-3.5-turbo-1106 gpt-4-32k-0314 gpt-4
        response_format={ "type": "text" },
        #stream=True,
        messages=[
            {"role": "system", "content": "The system receives CQI data for each network slice in real-time or at scheduled intervals.\
             It analyzes the CQI data to detect any values that fall below a predefined threshold which indicates suboptimal performance.\
             Upon detecting low CQI values, the system determines the best course of action, such as suggesting a reallocation of resources\
             or adjusting other slice parameters like bandwidth or priority levels. The system could also be responsible for generating alerts for network operators,\
             providing detailed reports, or executing pre-approved actions to remedy the situation."},

            {"role": "user", "content": f"We have a network slicing system where each slice is dedicated to a specific service type.\
             The system continuously monitors the Channel Quality Indicator (CQI) to maintain high service quality.\
             When a slice's CQI falls below the threshold of 10, it's an indicator that the service quality is degrading.\
             As the system, analyze the incoming CQI data and suggest actions to reallocate resources dynamically and adjust the slice parameters to improve the CQI.\
             Consider factors such as current network load, priority of the service, and historical performance data of the slice when making suggestions.\
             Generate a response that outlines potential adjustments to the network slice configuration to address the low CQI values. Consider the following {data}"}
        ]
    )

    end_time = time.time()
    processing_latency = end_time - start_time
    return response, processing_latency
```

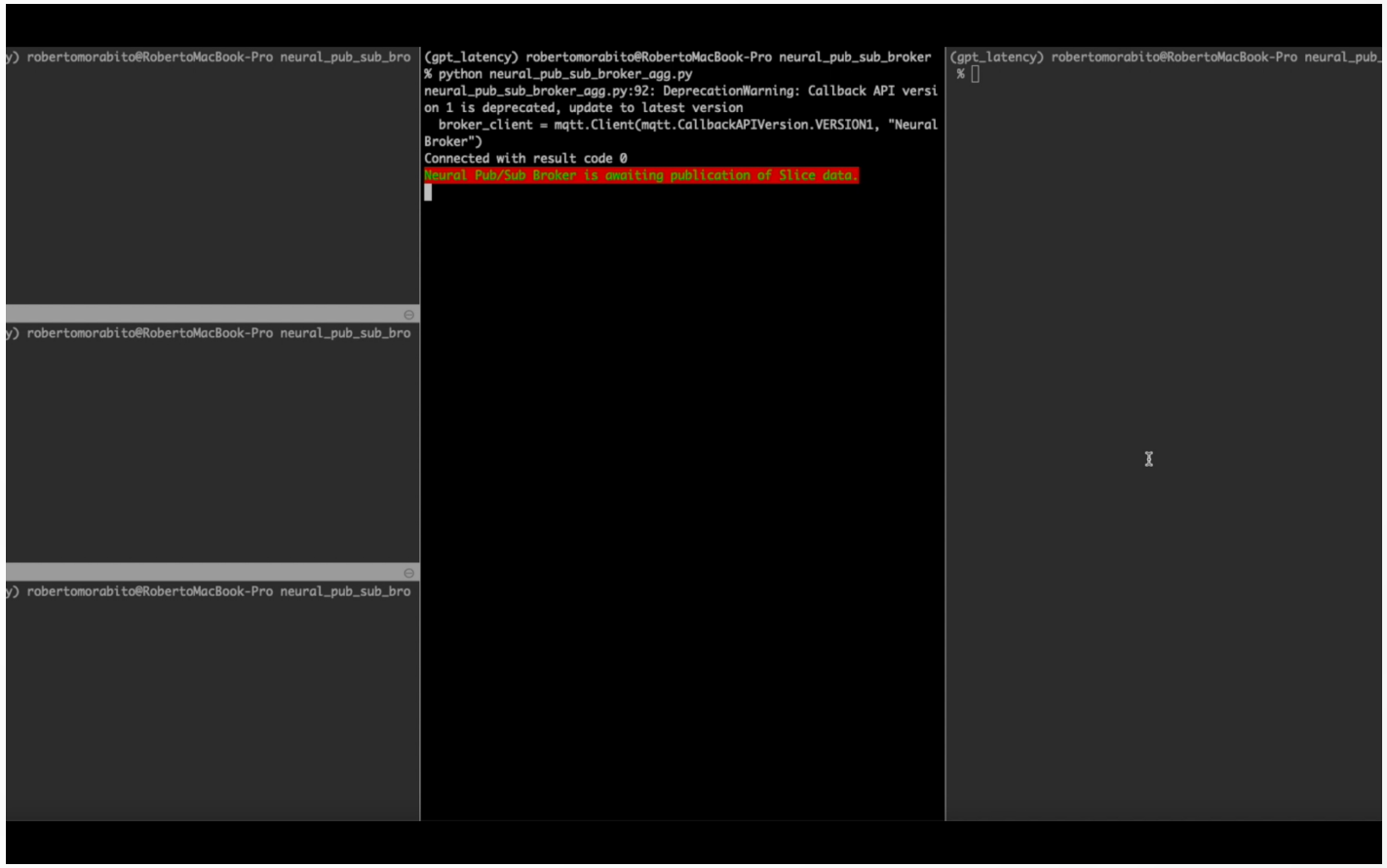
The Query Function at the Broker: The interconnect design accommodates multi-modal subscriptions and requests, complex subscriptions/tasks can be decomposed

NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT

DEMO

Phase 1:

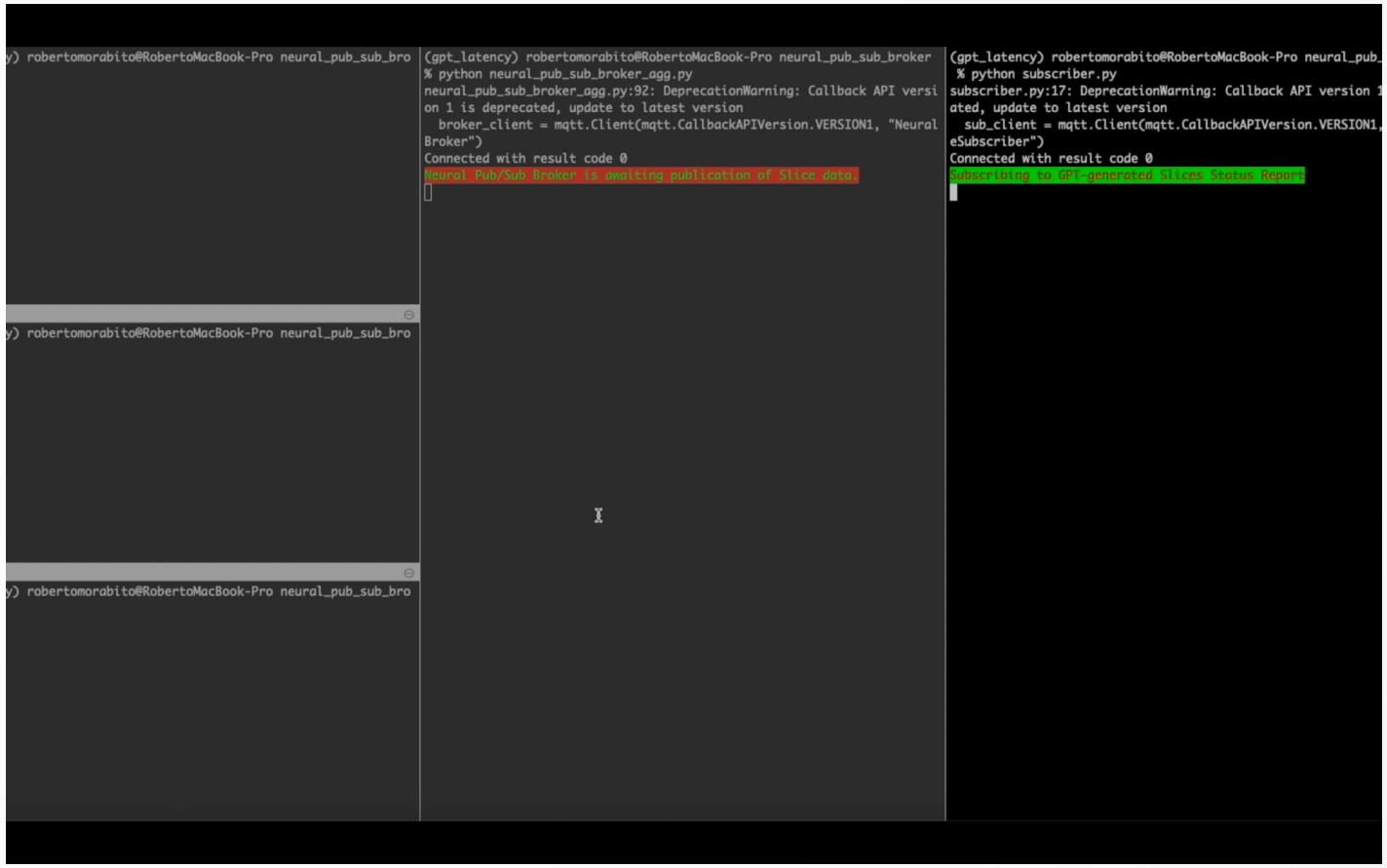
The Neural Pub/Sub Broker manages the publication of network slices data



NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT

DEMO

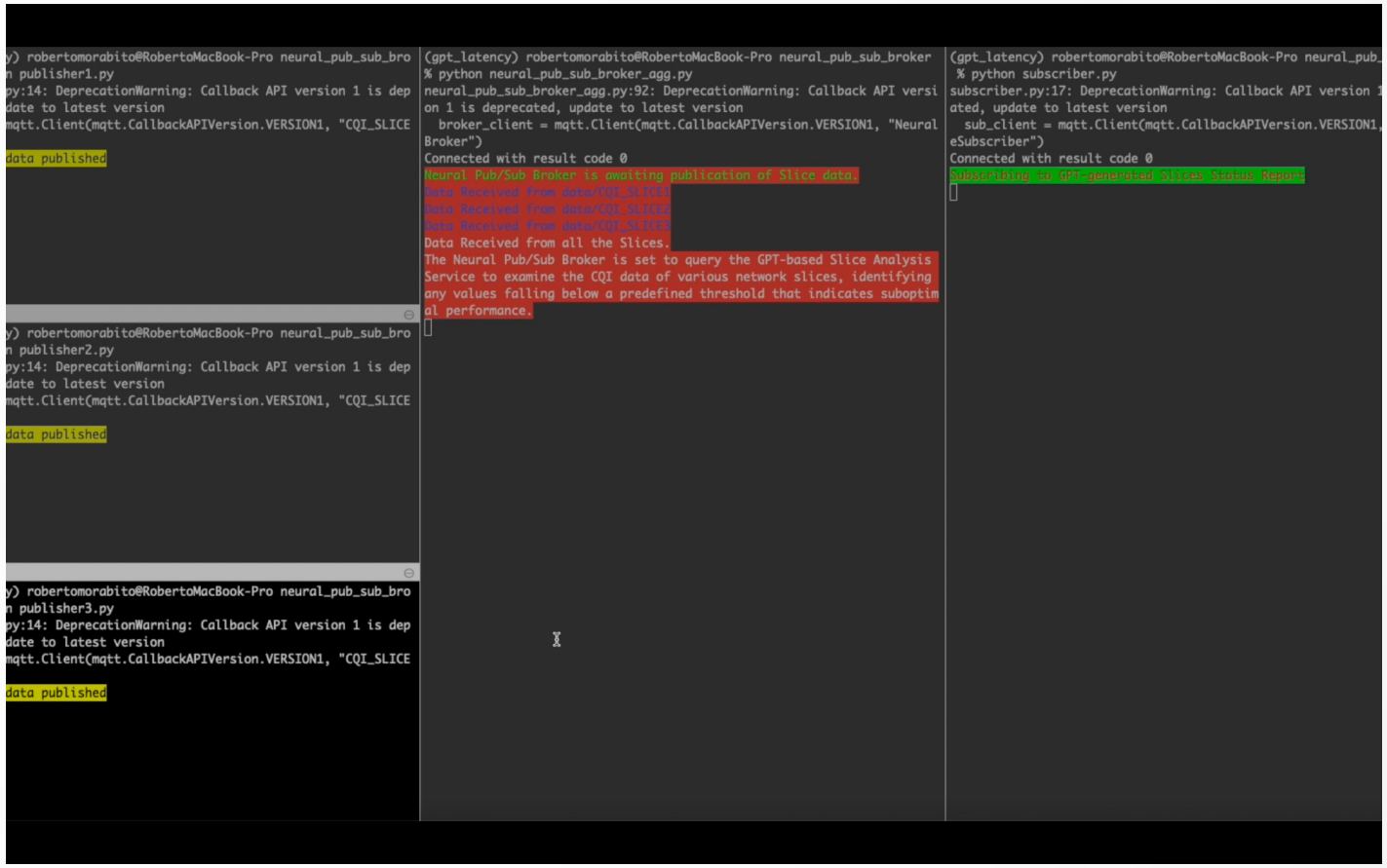
Phase 2:
The Network Operator subscribes to a GPT-generated Slices Status Report



NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT

DEMO

Phase 3:
Network Slices data is published by the three network slices

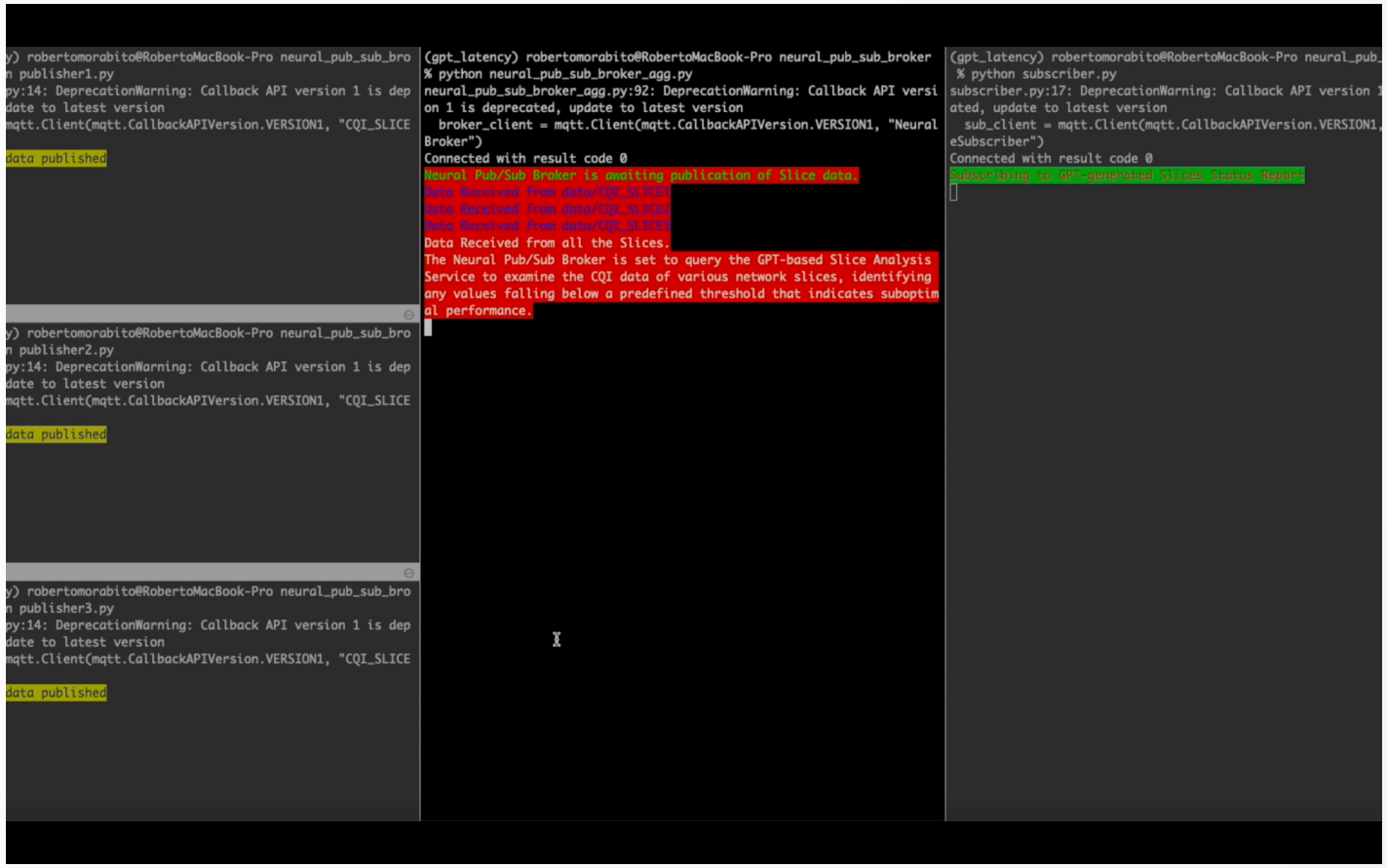


NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT

DEMO

Phase 4:

*The Neural Pub/Sub Broker query the **GPT-based Slice Analysis Service***

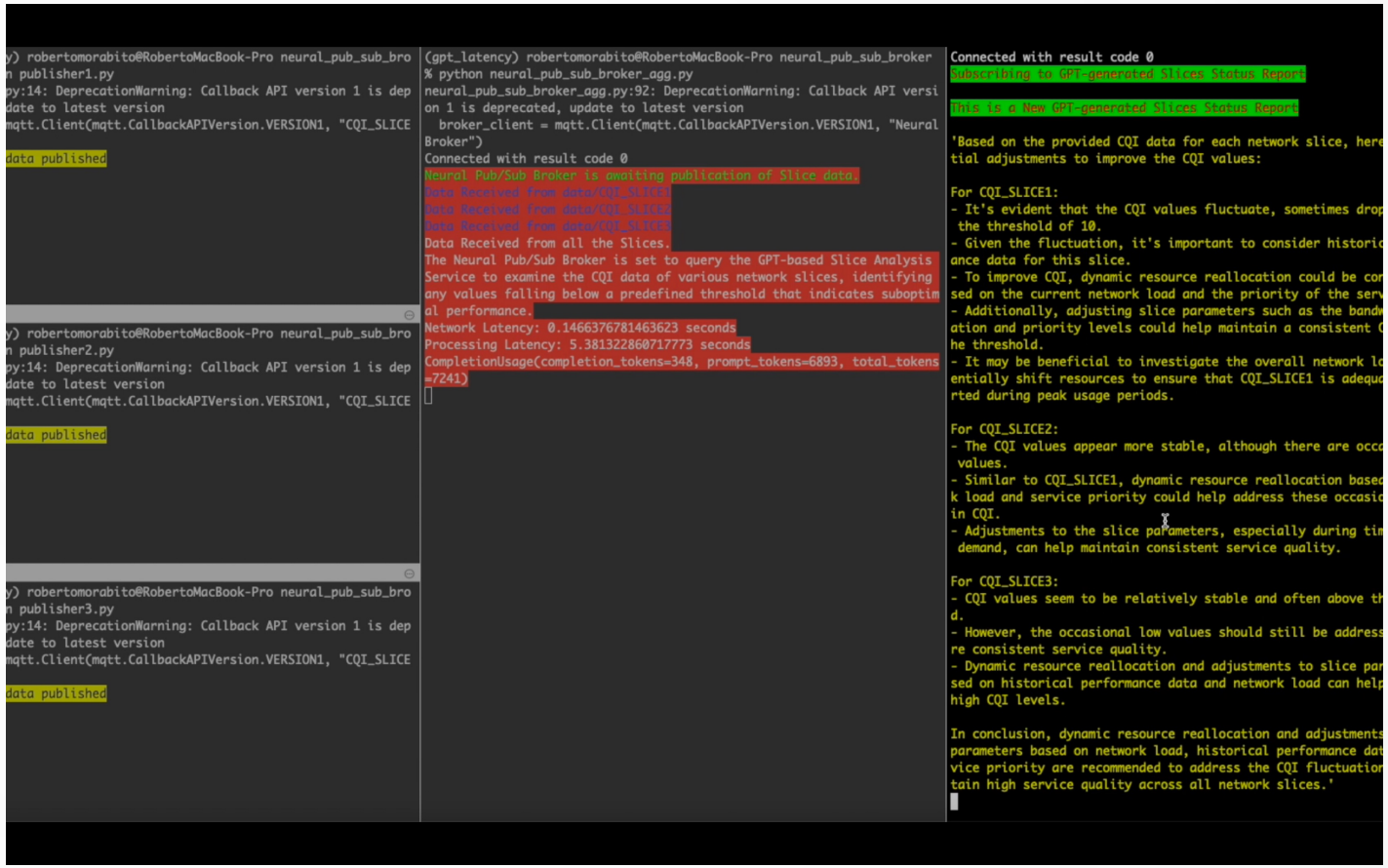


NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT

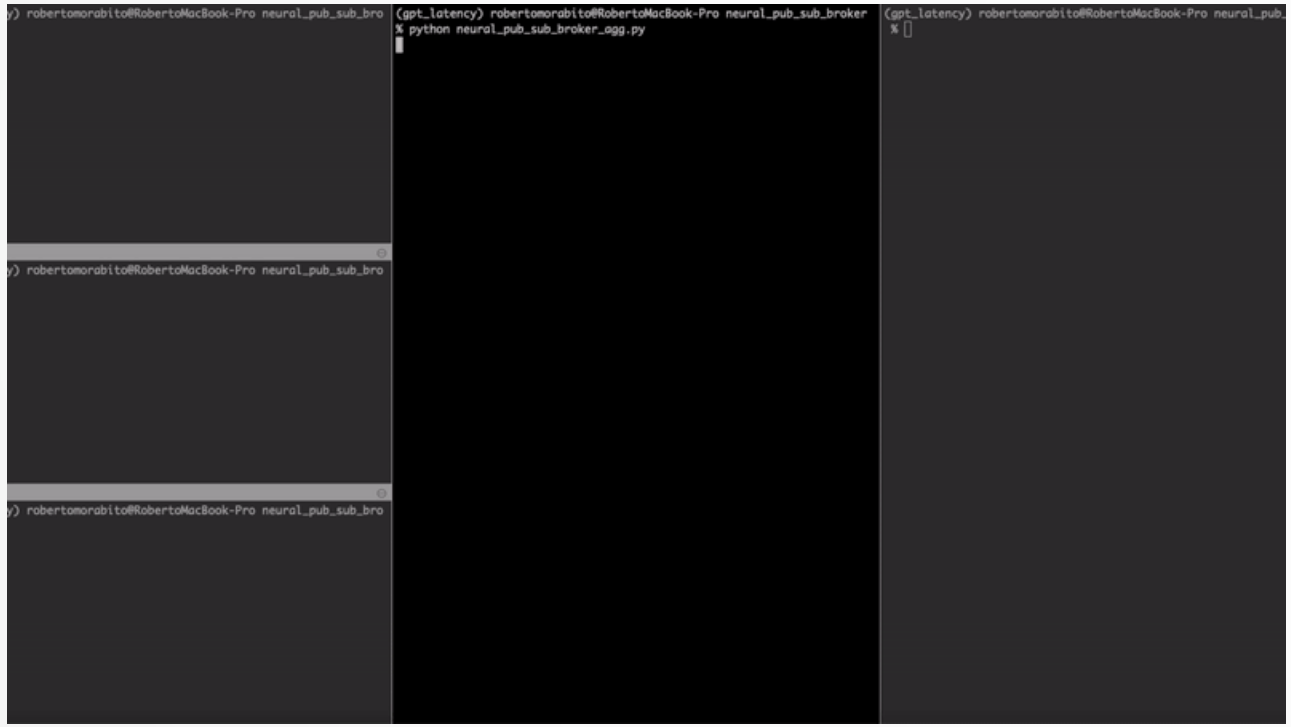
DEMO

Phase 5:

The Neural Pub/Sub Broker publishes to the Subscriber the ***GPT-generated Slices Status Report***



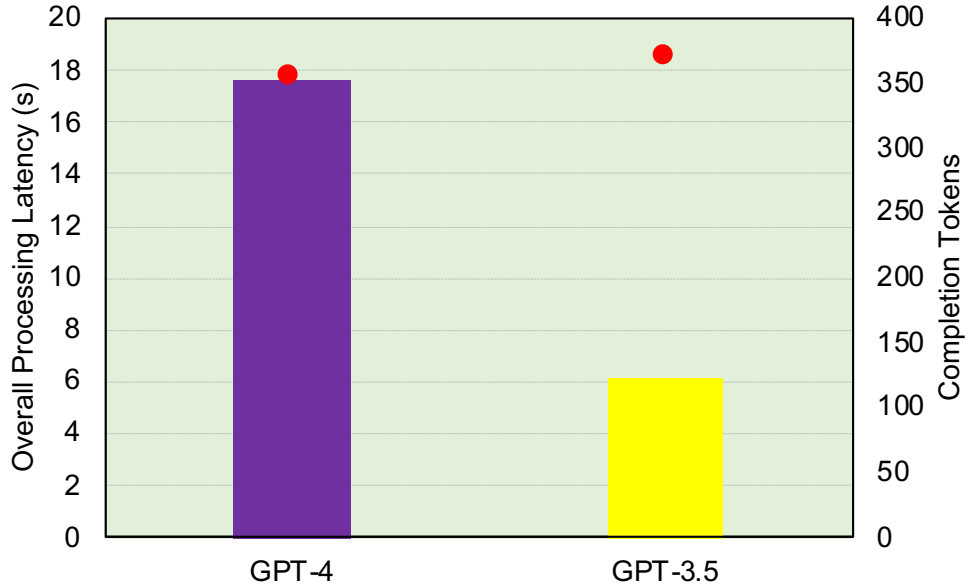
NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT



FULL DEMO

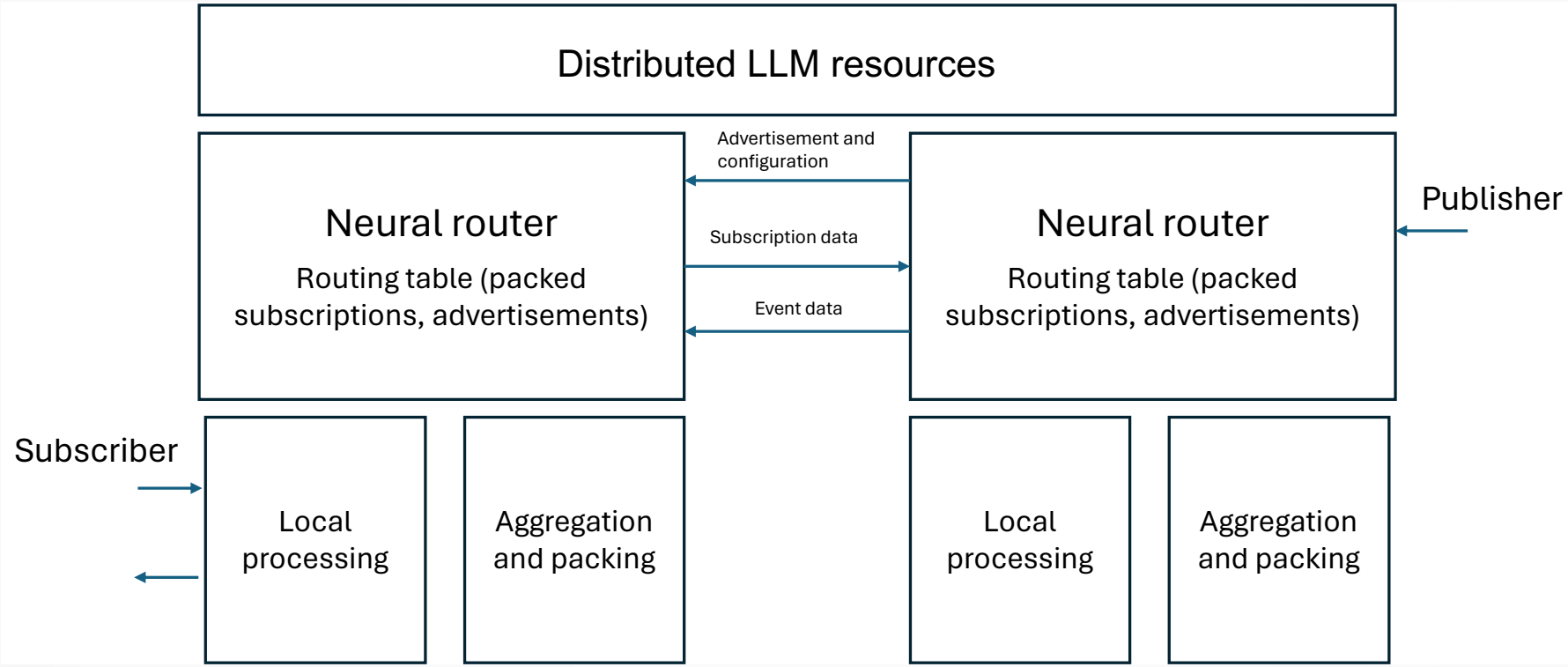
NEURAL PUB/SUB FOR NETWORK SLICES MANAGEMENT

- Demonstrated effectiveness using real data.
- Comparison between GPT-4 and GPT-3.5 models:
  - GPT-3.5 shows lower processing latency and higher verbosity, highlighting trade-offs in deployment.
- Private models (llama3, mixtral etc) are also suitable



Comparative analysis of GPT-4 and GPT-3.5, illustrating the sum of processing latency and network latency against the average completion tokens, represented by red dots for visual distinction.

NEURAL PUB/SUB ROUTER



LLM-Based Content Router

Input Subscriptions:


```
    "prompt": "Alert me whenever a firmware update fails on any IoT device, especially for our smart thermostats and lighting systems, to ensure all devices are running the
latest software for optimal performance and security."
  },
  {
    "identifier": 4,
    "prompt": "Provide real-time alerts if any industrial IoT sensors in our manufacturing plant lose connectivity for more than 5 minutes, as this could indicate critical
production issues or equipment failures."
  },
  {
    "identifier": 5,
    "prompt": "Send an alert if any environmental sensors, like those monitoring temperature, humidity, or air quality, report readings outside of our set safety thresholds in ou
storage facilities, to prevent damage to sensitive goods."
  }
]
```


Input Events:


```
[
  {
    "event": "Battery level of security camera in downtown clinic falls to 8%",
    "matching_subscription_ids": [1]
  },
  {
    "event": "Weekly data usage report shows a smart meter exceeded 600MB in a single day",
    "matching_subscription_ids": [2]
  },
  {
    "event": "Firmware update failed on smart thermostats in multiple residential buildings",
    "matching_subscription_ids": [3]
  }
```


Start the routing process.

Aggregated Subscriptions: Aggregated Subscriptions:

[{"identifier": [1, 3], "llm_output": "Notify me immediately when critical IoT devices, like healthcare monitors or security cameras, have less than 10% battery remaining or whenever a firmware update fails on any of these devices, to ensure we can proactively replace or recharge them and that all devices are running the latest software for optimal performance and security."}, {"identifier": [2, 5], "llm_output": "Send a weekly summary of data usage and real-time alerts for environmental sensors for all IoT devices, highlighting any device that exceeds 500MB of data in a day or reports readings outside of our set safety thresholds in our storage facilities, to ensure we stay within our data plan limits, identify any unusual activity, and prevent damage to sensitive goods."}, {"identifier": [4], "llm_output": "Provide real-time alerts if any industrial IoT sensors in our manufacturing plant lose connectivity for more than 5 minutes, as this could indicate critical production issues or equipment failures."}] JSON difference to original 152 and percentage 12.958226768968457Number of subs in the merged 5 Number of subs in the merged 3

Matching Events Per Subscription:

Matching: [{ "event": "Battery level of security camera in downtown clinic falls to 8%", "matching_subscription_ids": [1] }, { "event": "Weekly data usage report shows a smart meter exceeded 600MB in a single day", "matching_subscription_ids": [2] }, { "event": "Firmware update failed on smart thermostats in multiple residential buildings", "matching_subscription_ids": [3] }] Matching: [{ "event": "Battery level of security camera in downtown clinic falls to 8%", "matching_subscription_ids": [1] }, { "event": "Weekly data usage report shows a smart meter exceeded 600MB in a single day", "matching_subscription_ids": [2] }, { "event": "Firmware update failed on smart thermostats in multiple residential buildings", "matching_subscription_ids": [3] }]

Neural Pub/Sub Architecture

Publisher

Data
(text, visual, aural, telemetry, …)

Publisher Capabilities Monitoring

Node X

Neural pub/sub broker

AI pipelines scheduler
(Filters ⊕ Transformations)

AI Model Splitter and Reasoner

AI Model Discovery

Node Y

Execution Units (EU)

AI Service
(training, inference, ...)

AI Models Repository

EU Capabilities Monitoring

Node N

Node O

Subscriber

Application
(text, visual, aural, telemetry, …)

Local AI Service

AI Models Repository

Subscriber Capabilities Monitoring

Node Z

Orchestration

5G-based communication

Neural Pub/Sub testbed deployment

Hardware components

Server and Networking Equipment

Edge and End Devices

Software components

AI/ML Applications

Analytics

Monitoring

Neural Pub/Sub Broker

Orchestration

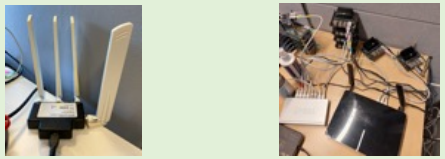
AI/ML Models

University of Helsinki site

Hardware components

Server and Networking Equipment

Edge and End Devices



Software components

AI/ML Applications

Analytics

Monitoring

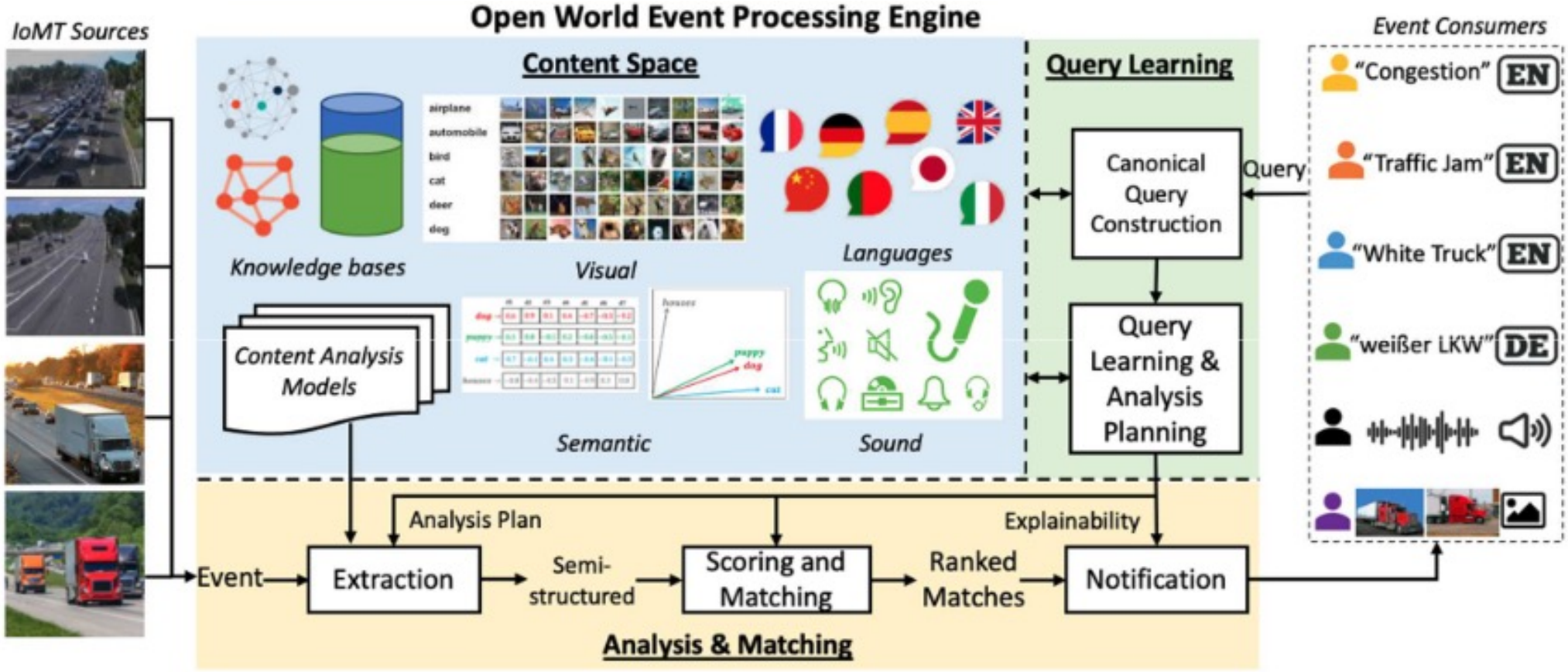
Neural Pub/Sub Broker

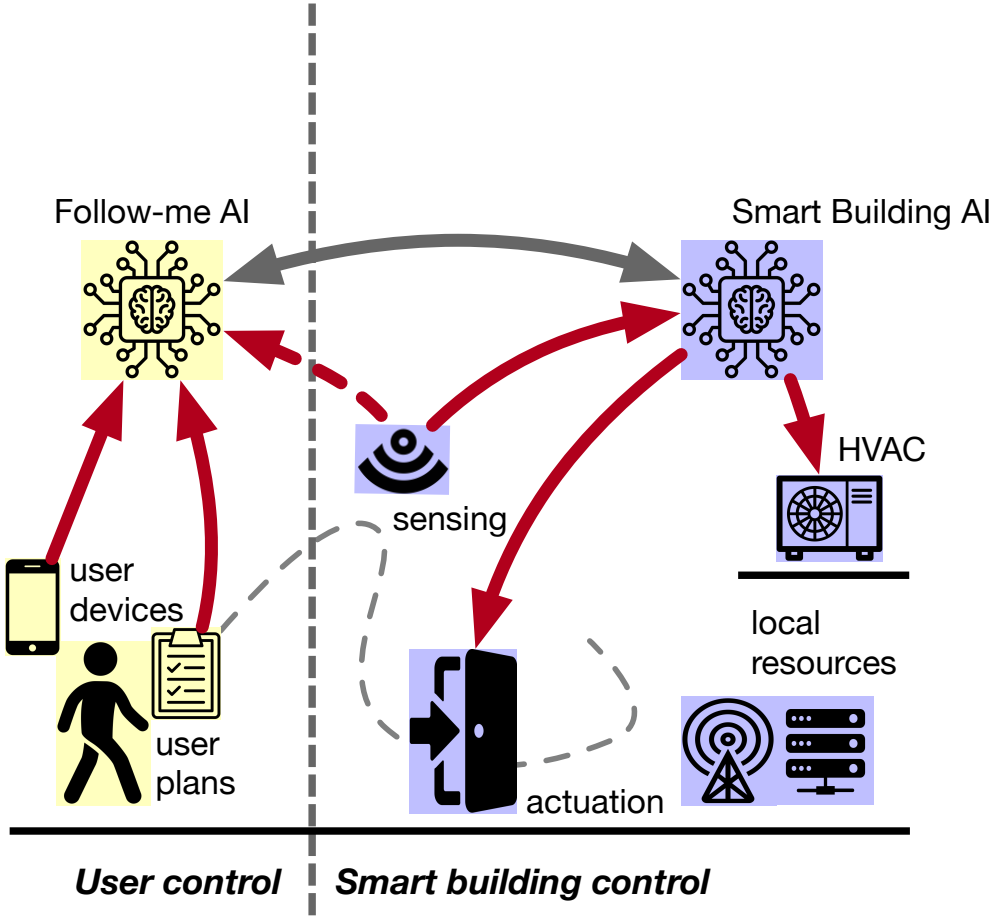
Orchestration

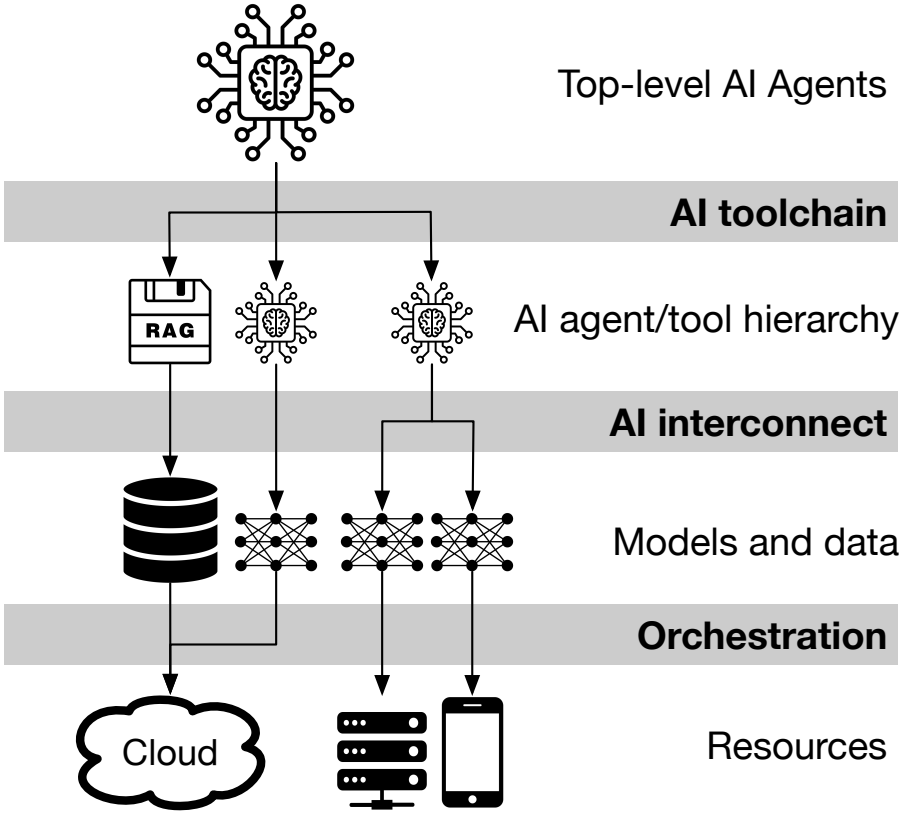
AI/ML Models

University of Oulu site

ABHISHEK KUMAR, LAURI LOVÉN, SUSANNA PIRTTIKANGAS AND SASU TARKOMA. "DATA FABRIC FOR INDUSTRIAL METAVERSE.” ACCEPTED TO 44TH IEEE INTERNATIONAL CONFERENCE ON DISTRIBUTED COMPUTING SYSTEMS (ICDCS 2024), NEW JERSEY, US.







Saleh, Alaa, Praveen Kumar Donta, Roberto Morabito, Naser Hossein Motlagh, and Lauri Lovén. "Follow-Me AI: Energy-Efficient User Interaction with Smart Environments." arXiv preprint arXiv:2404.12486 (2024).

CONTENTS

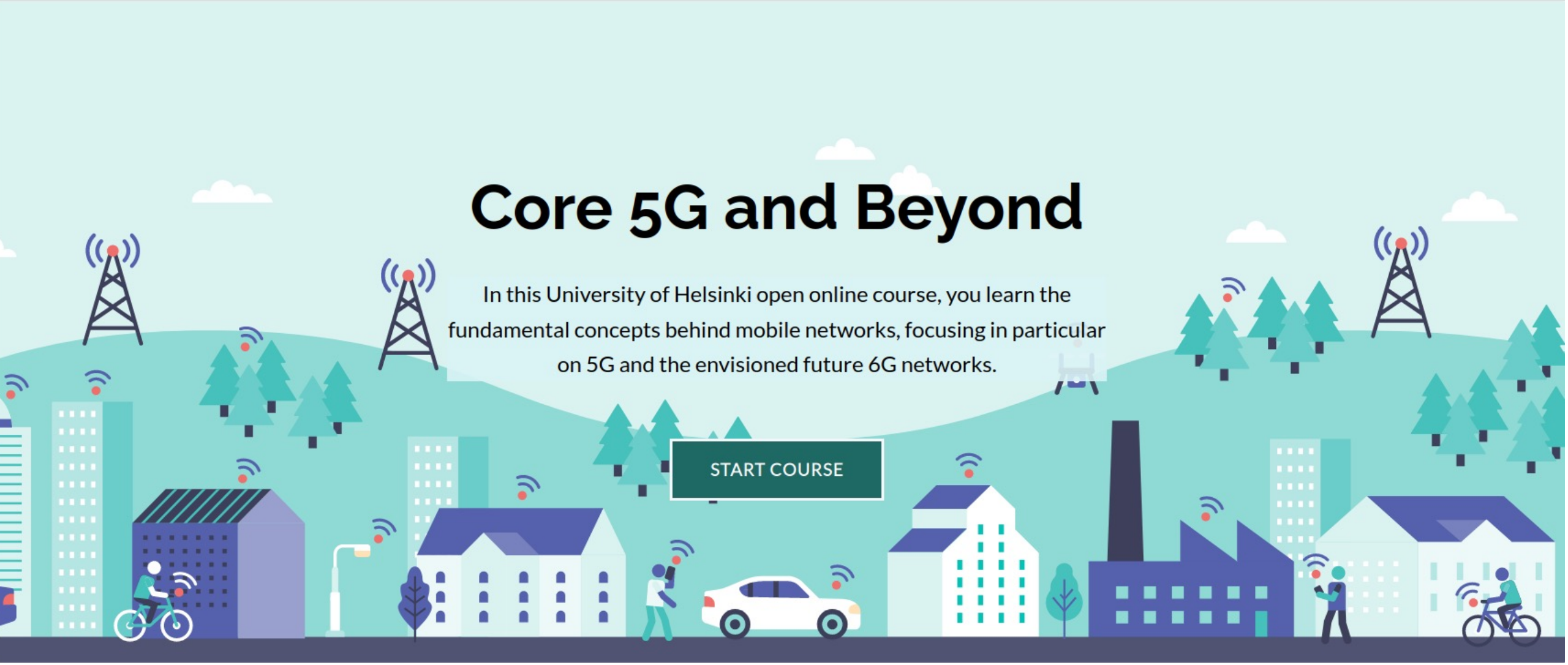
- BACKGROUND
- TECHNOLOGY MIX: 6G AND AI
- AI INTERCONNECT AND NEURAL PUB/SUB
- EDUCATION: 6G MOOC

WE NEED 6G EDUCATION AND SKILLS

Upskilling Europe and the world.
Explaining 5G and beyond.
Raising interest in technology.
For creating a safer, more sustainable and connected planet.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Core 5G and Beyond



In this University of Helsinki open online course, you learn the fundamental concepts behind mobile networks, focusing in particular on 5G and the envisioned future 6G networks.

START COURSE

UNIVERSITY OF HELSINKI

MINISTRY OF EDUCATION AND CULTURE FINLAND

FITech NETWORK UNIVERSITY

University of Helsinki together with 6G Finland invites everyone to learn about 5G and 6G.

6G.MOOC.FI

UNIVERSITY OF HELSINKI

MINISTRY OF EDUCATION AND CULTURE FINLAND

FITech NETWORK UNIVERSITY

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI