CHIST-ERA LeadingEdge Project

LeadingEdge

Holistic and foundational resource allocation framework for optimized and impactful edge computing services

Topic: SDCDN

Iordanis Koutsopoulos (Coordinator)

Athens University of Economics and Business – Research Center

(jordan@aueb.gr)

CHIST-ERA projects seminar: March 28-30, 2022

Overview of presentation

- LeadingEdge facts
- Challenges and objectives
- Partners and added value
- Main project achievements
- Dissemination actions
- Exploitation actions and systainability

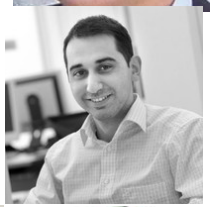
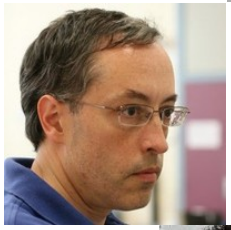
LeadingEdge facts

- **Topic**: Smart Distribution of Computing in Dynamic Networks (SDCDN)
- **Start date**: April 1, 2020
- **Duration**: 3 years
- **Funding**: 1,01MEuros
- **Coordinator**: Iordanis Koutsopoulos (AUEB-RC)
- **Webpage**: https://mm.aueb.gr/leadingedge/

- **Partners:**
  - Athens U of Economics and Business – Research Center (GR) – Coordinator
  - StreamOwl (GR)
  - University Polytechnic Catalunia (ES)
  - University of Oulu (FI)
  - EURECOM (FR)
  - Huawei Lab Paris (FR)

Partner Presentation

- **AUEB-RC** (Prof. Iordanis Koutsopoulos):
  Dynamic resource allocation, Federated Learning, online learning
- **StreamOwl** (Dr. Savvas Argyropoulos)
  Video quality assessment
- **University Polytechnic Catalunya** (Profs. Leandro Navarro, Felix Freitag):
  Distributed systems, Federated Learning, IoT, experimentation with *guifi.net* community network
- **U of Oulu** (Prof. Mehdi Bennis):
- Distributed Learning, edge AI, low-latency & reliability, drone and vehicular networks
- **EURECOM** (Profs. Adlen Ksentini, Navid Nikaien):
  AI-driven service orchestration, 4G/5G experimentation computation offloading
- **Huawei France** (Dr. Apostolis Destounis):
  Online Convex Optimization, Learning,m on-device AI

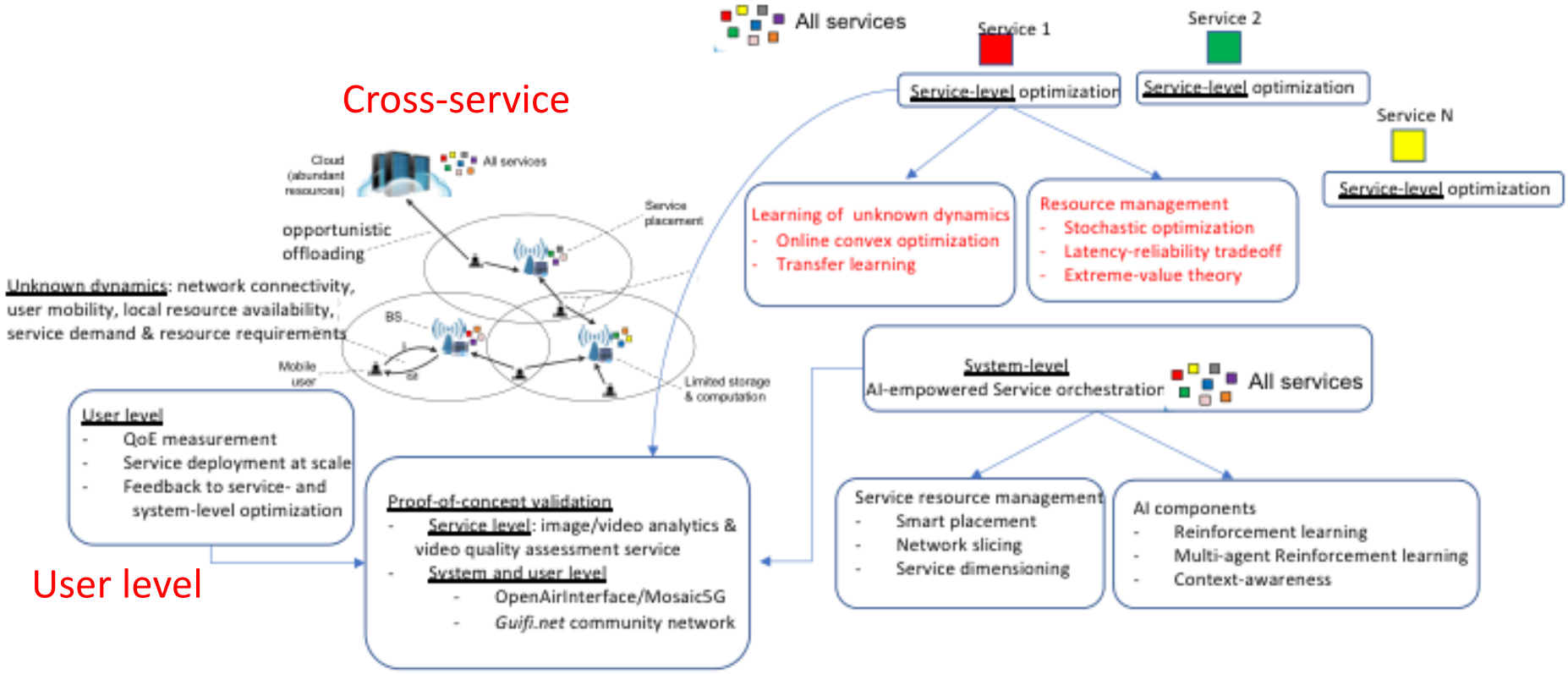







LeadingEdge in one slide

- Focus on **Edge Computing** and **Edge Analytics**
- Aspired Impact: ***Leave a firm mark on how services are managed in 5G and Beyond networks***
  - Resource-demanding services using resources at the network edge (e.g. for data analytics)
- **Pillar 1**: Intra-service resource allocation optimization at the edge
  - Edge resource consumption and management (cache capacity, compute, bandwidth, energy) within a service
  - Coping with unknown dynamics at edge (demand, mobility, topology, etc)
- **Pillar 2**: Cross-service orchestration at system level
  - AI driven service orchestration
- **Pillar 3**: User-level QoE
  - Characterization of QoE at different settings

LeadingEdge Overview



Challenges and Scientific Objectives (1)

- Optimally use edge/cloud resources at service level, through intelligent task offloading to cloud, and dynamic allocation of computation among edge nodes
  - Stochastic optimization
  - novel usage of caching resources
- Derive fundamental performance limits and devise dynamic algorithms for maximization of computation rate of data analytics, with low latency and high reliability
  - extreme value theory
- Overcome unpredictability of service demand, mobility, topology
  - online convex optimization (OCO) for learning dynamics affecting intra-service resource allocation decisions
- AI-based *system-level service orchestrator*
  - dynamic service placement,
  - network slicing, dynamic resource provisioning across services,
  - context-awareness

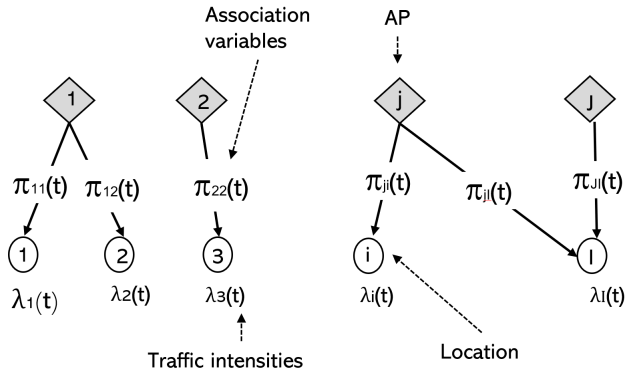
Project Objectives (2)

- Service-level proof-of-concept (PoC) validation:
  - Real-time image recognition that optimally uses resources
  - Video quality assessment solutions, with different complexity/accuracy and configuration of edge/cloud and cloud/backend resources
- System-level validation:
  - **OpenAirInterface.org** (OAI), **Mosaic-5g.io** software platforms: Real-time experimentation environment with full 4G/5G functionalities for service orchestration
  - the ***guifi.net*** community-network infrastructure to deploy services at scale and measure perceived user-level QoE

LeadingEdge WP Structure

- **WP1**: Effective learning of unknown dynamics in edge computing: HUA, AUEB-RC, UO
- **WP2**: Low-latency and reliable edge computing: UO, AUEB-RC, HUA
- **WP3**: Optimized intra-service resource allocation within services: AUEB-RC, UO, EURECOM, SOWL
- **WP4**: Service orchestration for optimized inter-service resource allocation and network slicing: EURECOM, UPC
- **WP5**: Experimentation and service deployment over community networks: UPC, EURECOM
- **WP6**: Dissemination and Exploitation: SOWL, ALL

Online Convex Optimization for user association



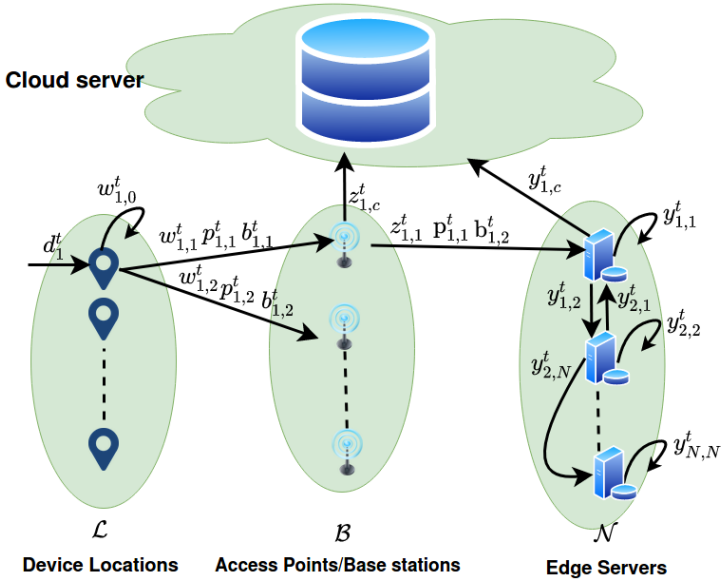
- Blind optimal user association in small-cell networks
- Learn what is the best service policy π of a set of locations through a set of Access Points
  - Demand at each location unpredictable, non-stationary, known only after association decision is made
- Devised zero-regret algorithm
- Defined new benchmark that is in between static and dynamic regret

L. E. Chatzieleftheriou, A. Destounis, G. Paschos, I. Koutsopoulos. *"Blind Optimal User Association in Small-Cell Networks"*, in IEEE INFOCOM 2021.

OCO for Edge Computing: system model

- Resource allocation decisions:
  - Fraction of the request to process locally and transmit to each AP.
  - Transmission power and bandwidth to be used. — Wireless access
  - Bandwidth for routing to the cloud and each edge server. — APs to edge
  - Allocation of computing resources at each edge server.
  - Bandwidth for rerouting to other edge servers and the cloud server. — Edge
- Incoming Requests: Request rate of computation tasks - adversary.
- Cost:
  - Delays based on the load ρ resulting from routing and allocation of resources.
  - Total cost:

$$f_t(x_t) := \beta_1 \sum_{l\in\mathcal{L}} \Big(\overbrace{c^t_{l,0}(w^t_{l,0})}^{\text{Local processing delay}} + \overbrace{\sum_{b\in\mathcal{B}} c^t_{l,b}(w^t_{l,b}, p^t_{l,b})}^{\text{Delay over the wireless access}} \Big) + \beta_1 \sum_{b\in\mathcal{B}} \Big(\overbrace{c^t_{b,c}(z^t_{b,c})}^{\text{Latency to the cloud}} + \overbrace{\sum_{n\in\mathcal{N}} c^t_{b,n}(z^t_{b,n}, \mathrm{p}^t_{b,n})}^{\text{Latency to the edge server}} \Big)$$
$$+ \beta_1 \sum_{n\in\mathcal{N}} \Big(\underbrace{c^t_{n,n}(y^t_{n,n})}_{\text{Edge processing delay}} + \underbrace{c^t_{n,c}(y^t_{n,c})}_{\text{Latency to the cloud}} \Big) + \beta_2 \Big(\underbrace{\sum_{l\in\mathcal{L},b\in\mathcal{B}} (p^t_{l,b})^2 + \sum_{b\in\mathcal{B},n\in\mathcal{N}} (\mathrm{p}^t_{b,n})^2}_{\text{Power expenditure}} \Big)$$



Delay at node i $\approx \frac{\rho_i}{1-\rho_i}$

- Introduced new concepts suitable for edge computing, as constraints in online optimization problem, e.g. total number of nodes with service outages
- Proposed OCO no-regret algorithms for objective and constraints
- Tackled case of full and limited (bandit) feedback

A. Chouayakh and A. Destounis, "Towards no regret with no service outages in online resource allocation for edge computing," in IEEE ICC, 2022

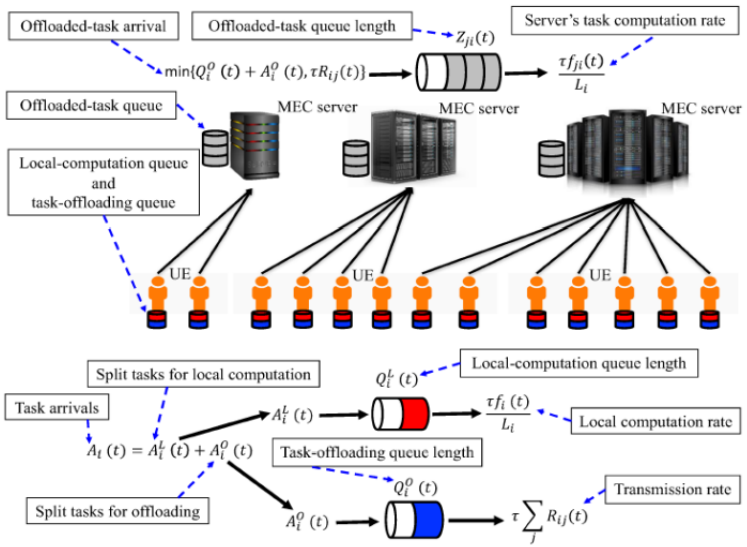
Low-latency 5G Communication

- Goal. Dynamic Task Offloading and Resource Allocation for Ultra-Reliable Low-Latency Edge Computing
- Challenges:

1. Much tighter URLLC requirements.
2. Information mismatch/freshness in factory manufacturing

Objectives:

1. Reliability-latency-energy-power tradeoff.
2. Information freshness investigation.



(a) System architecture.

1) Local vs. Remote computing
2) User-BS association under latency constraints

Cast as a Lyapunov optimization based user scheduling and RB allocation

Tail distribution characterized via Extreme Value Theory

Age-optimal IoT communication

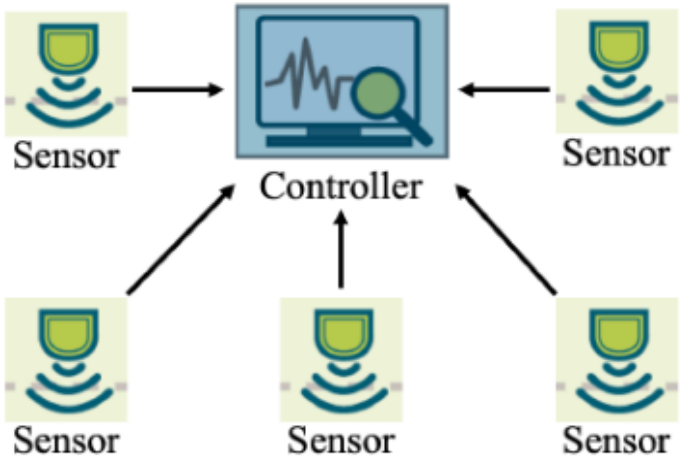
Goal: Reliable connectivity for smart factory and Industry 4.0.

Challenges:

1. Much tighter URLLC requirements.
2. Information mismatch/freshness in factory manufacturing

Objectives:

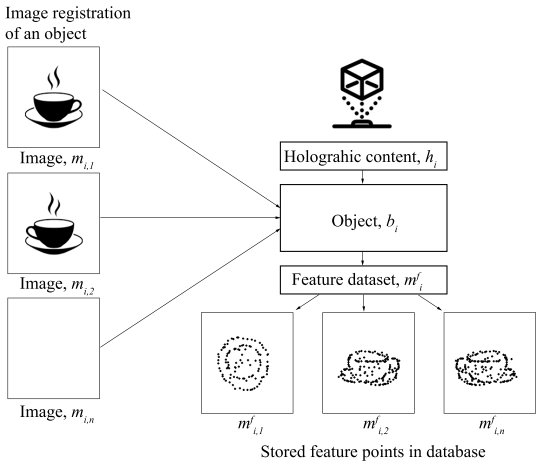
1. Reliability-latency-energy-power tradeoff.
2. Information freshness investigation.

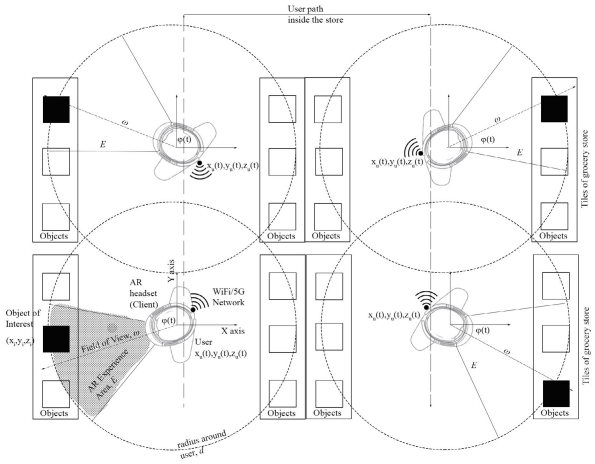


- The sensor's transmit power affects the **queuing delay**, **transmission delay**, and **peak AoI**.
- We focus on the **power minimization** problem

C.-F. Liu and M. Bennis, ''Federated learning with correlated data: Taming the tail for age-optimal industrial IoT,'' in *Proc. 19th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)* 2021.

Novel use of caches for Augmented Reality



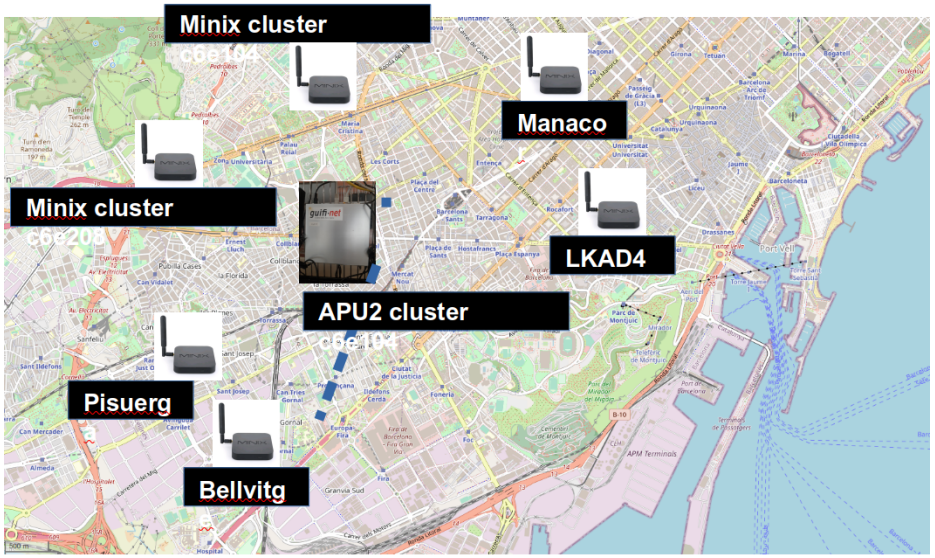


- Object image caching to reduce computation and communication costs involved in object recognition in Augmented Reality (AR)
- Caches images using a combination of metrics:
  - object popularity index (favours objects most probable to be requested for recognition)
  - percentage of times when the object label has been encountered in the past,
  - probability that an image is similar enough with already encountered images with same label
- Outcome: drastic reduction of database searches
- significant reduction of computational resources (by 90%) and induced network traffic, thus improving user experience

G. Koutitas, S. Vyas, C. Vyas, S. Jadon, I. Koutsopoulos "Practical Methods for Efficient Resource Utilization in Augmented Reality Services", in IEEE Access, Dec 2020.

Community network based services

- Community network testbed: Wireless mesh network with mini-pc small servers
- Locations: UPC campus and several in Barcelona
- eReuse testbed:
- Distributed Ledger Technology (DLT) (permissioned blockchain*) for verifiability (traceability, accountability) of sustainable (circular) edge/end-user devices (use, reuse, recycling).*
- Implemented with geth & Besu daemons, 3-5 replica nodes, transactions and Ethereum virtual machine with Solidity smart contracts
  - Two locations: UPC campus, CATNIX Internet eXchange



Future:

→ Explore user experience of eReuse: digital sustainable **product passport**, at global scale → **spin-off company**, in preparation, to provide software and services on digitalised management of edge devices in computing infrastructures

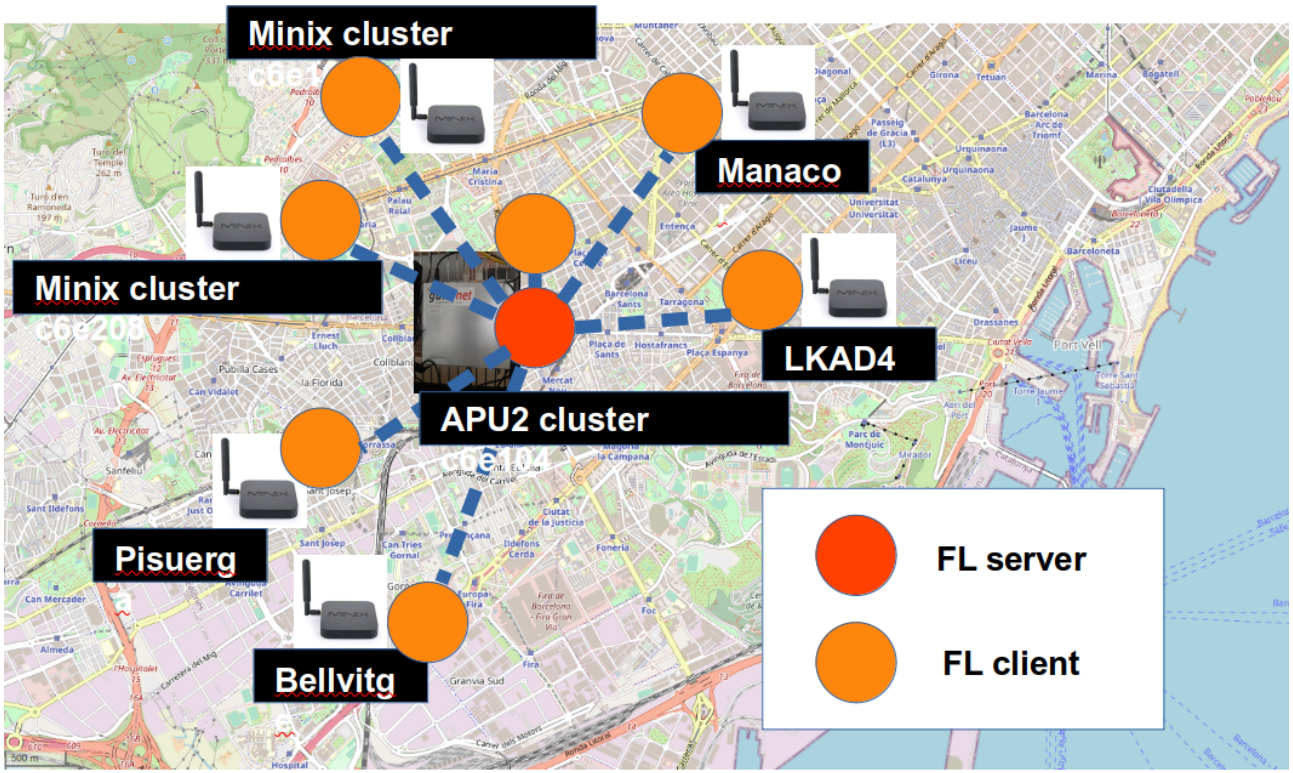
→ Contribution to **international standards**: UN ITU-T, on decentralized digital infrastructures for sustainable digital transformation of cities and communities.

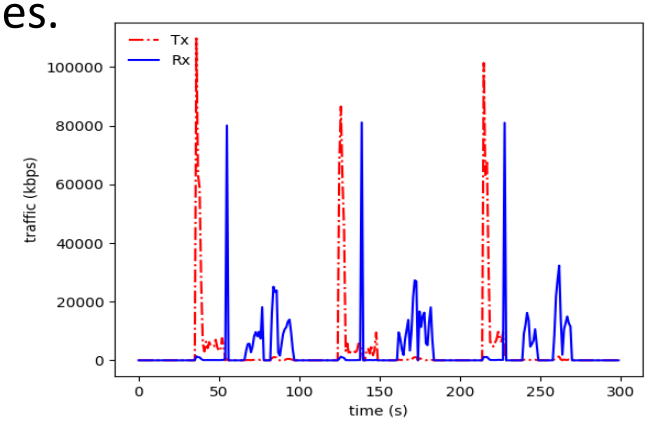
L Navarro, M Roura, D Franquesa, *eReuse poster: the circular economy of digital devices*, COMPASS '21: ACM SIGCAS Conf. on Computing and Sustainable Societies, June 2021

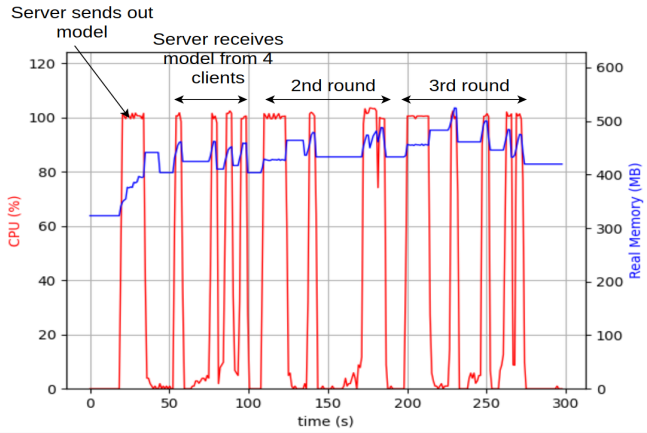
Federated Learning experimentation

Federated Learning deployment and evaluation in *Guifi.net* testbed

Develop, evaluate experimentally and analyse performance of federated learning over wireless mesh networks and low computing capacity devices.



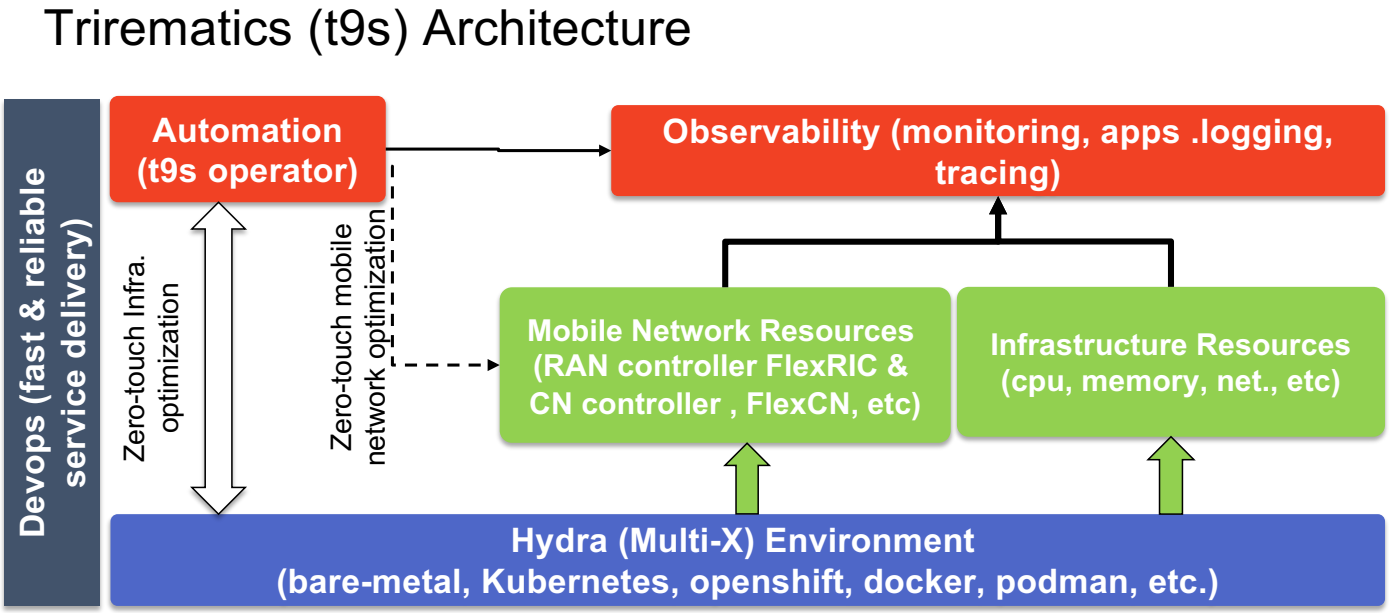




F. Freitag, P. Vilchez, L. Wei, C.H Liu, M. Selimi, I. Koutsopoulos. *Demo: An Experimental Environment Based On Mini-PCs For Federated Learning Research*. 2022 IEEE Consumer Communications & Networking Conference (CCNC), January 2022

F. Freitag, P. Vilchez, Ch. Liu, L. Wei, M. Selimi. *Performance Evaluation of Federated Learning over Wireless Mesh Networks with Low-Capacity Devices*. International Conference on Information Technology & Systems (ICITS), February 2022

Intelligent 4/5G Network Service orchestration



Cloud-Native agile 4G/5G service Operator in cloud native environment (Kubernetes)

Developed RAN network functions from multiple vendors as cloud native functions to be automated in Kubernetes via Trirematics-Operator.

Developed Trirematics-Operator (management and orchestration tool to support different deployment scenarios (e.g., monolithic and disaggregation modes for the RAN and core network)

Automated deployment and (re)configuration.

Of high interest for OpenAirInterface (OAI) and Mosaic5G communities for service provision in cloud native.

O. Arouk, A. Mohammadi, N. Nikaein. "Trirematics: An Intelligent RAN service operator in cloud native environment" OAI Workshop, December 2021; https://youtu.be/duFrpN3zQBg?list=PLvcdcD3akbdBFPZs44h2hpQDmEomruEUP

Dissemination actions

- 14 papers in international conferences
- 9 papers in international journals
- 1 book chapter, 1 demo paper
- Involvement of consortium members in organization of 2 international workshops and conferences: WiOpt 2021, Mosaic5G workshop
- 1 Tutorial and 3 Keynote presentation in international conferences
  - @IEEE ICC 2020, IEEE GLOBECOM 2020, IEEE CCNC 2021
- Collaboration with research groups out of project
- Dissemination to 3 European projects
  - 6GENESIS project https://www.oulu.fi/6gflagship/
  - 6G-Brains, 5G-VICTORI (https://www.5g-victori-project.eu/
- Participation in Video Quality Experts Group

Sustainability and exploitation (1)

- **3 Pillars of exploitation**
  - **guifi net:** service deployment and orchestration of interest to guifi.net community network
  - **Trirematics** (successor of Kube5G) developed with the possibility to orchestrate 4G/5G RAN in cloud native environment (Kubernetes), to be released as an opensource software.
    - of high interest for OpenAirInterface (OAI) and Mosaic5G communities for service provision.
  - **StreamOwl**: Edge-based versions of video quality monitoring probes based on international video quality standards.
- **New services**
  - eReuse initiative has released several software pieces for the inventory and management of digital devices in a community according to principles of circular economy**.**
  - LoRaMoto: a communications system aimed at helping civilians exchange information in the aftermath of a natural disaster

Sustainability and exploitation (2)

- **Software and prototypes**
  - **Federated Learning**: publication of software in public repositories
  - **PiGeon:** a lightweight platform for deploying QoS-sensitive services in edge clouds built of single-board devices.
  - **REDEMON**: a resilient decentralized automated **monitoring** system of the ***guifi.net*** Community Network to replace legacy monitoring system
- **Planned exploitation actions**
  - Creation of a spin-off software and services company **around the topic of the circular economy of digital devices in local communities**.
  - European patents of inventions from project work
  - Launch follow-up internal project in Huawei to leverage results and examine applicability for 6G AI services
  - Horizon Europe followup projects, discussions underway

Thank you!

LeadingEdge site:
https://mm.aueb.gr/leadingedge/

Contact: jordan@aueb.gr