Marsal: Self-Driven orchestration and intelligent management of communication, storage and computational resources

Dr. Panagiotis Kokkinos

Institute of Communication and Computer Systems (ICCS)

National Technical University of Athens

kokkinop@mail.ntua.gr

Motivation

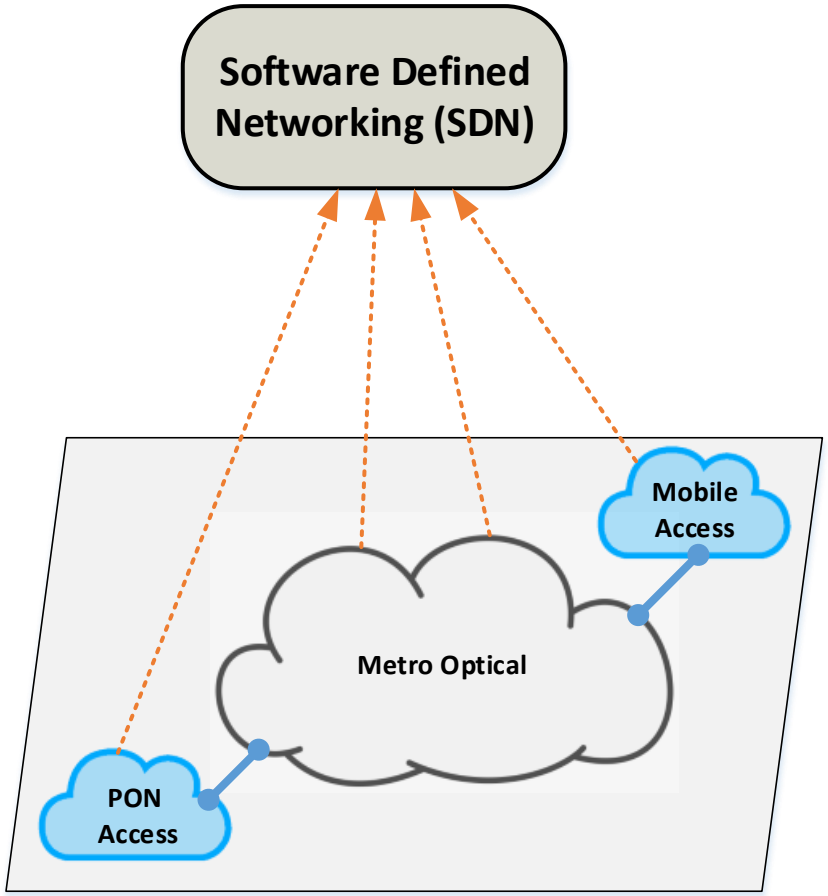
- The requirements for computing, storage and network resources is constantly increasing due to the use of popular on-line services, the Internet of Things (IoT) and smart-city applications
- Service providers tend to push more and more content, data and services as close as possible to the end-users/devices *(**to the edge**)*
- A large percentage of traffic flows, start and end within the city boundaries
- A variety of networking and computing technologies and infrastructures deployed in the cities that belong to different providers

Integrated Infrastructures: Computing and Storage

- Centralized cloud computing infrastructure services handle most processing and storage workload
- Edge computing offers computation and storage at the very edge of the network where data is produced:
  - on-device and on-premise edge
  - “far edge” (i.e., within a range of 100km from users and premises)
  - “near edge” (typically some 100s of kilometres from the users)
  - regional data centres

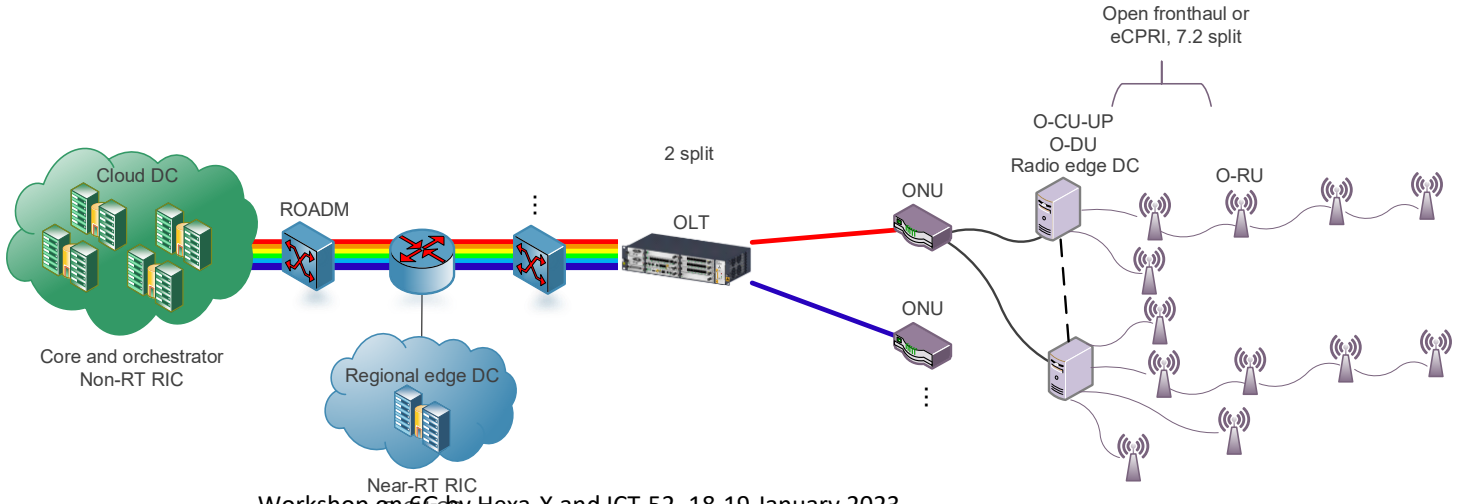
Integrated Infrastructures: Networks

- Metropolitan networks interconnect the access networks with the core communication infrastructures
- Both wired (optical) and wireless technologies can be employed in the various network segments (metro and access)
- The programmability of the infrastructures both wireless and wired is key for their efficient operation and the provision of advanced services

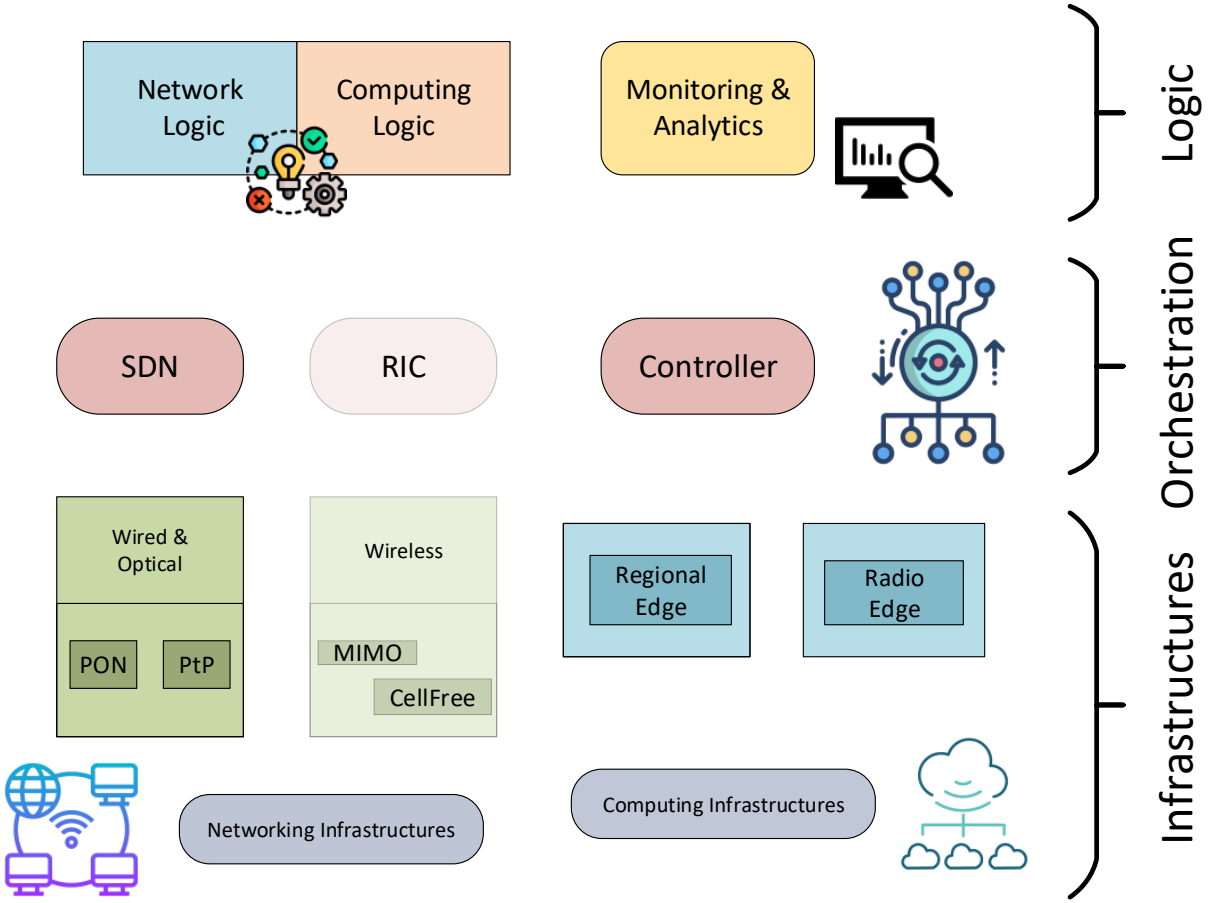


EU Project MARSAL

- **Machine Learning-Based, Networking and Computing Infrastructure Resource Management of 5G and Beyond Intelligent Networks**
  - Call: ICT-52-2020 - 5G PPP – Smart Connectivity beyond 5G
- The EU-funded MARSAL project proposes a new paradigm of elastic virtual infrastructures targeting converge computing and networking infrastructures
- In practice a virtual elastic infrastructure is a network slice composed of computing and networking resources reserved for use by an application or user
- MARSAL develops networking and computing management frameworks based on SDN and Cloud-Native Kubernetes coupled with a hierarchy of analytic and decision engines that provide close-loop operation

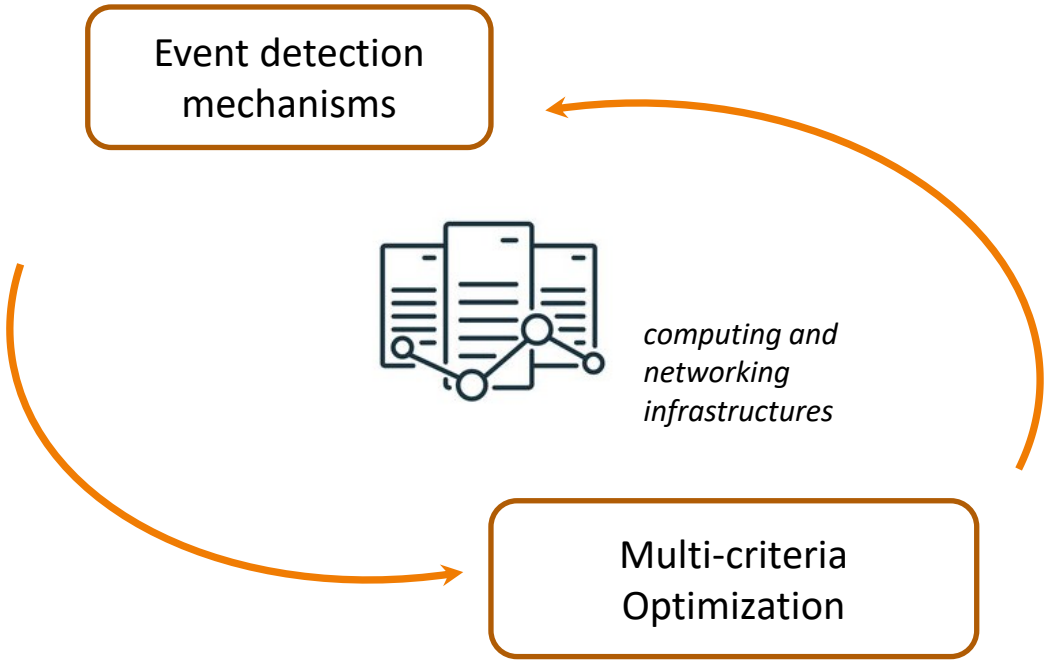


Marsal: Challenges & Innovations



Design and implement orchestration and analytics mechanisms for computing (edge & cloud) and network (Fixed Access - FAN and Radio Access Network) resources

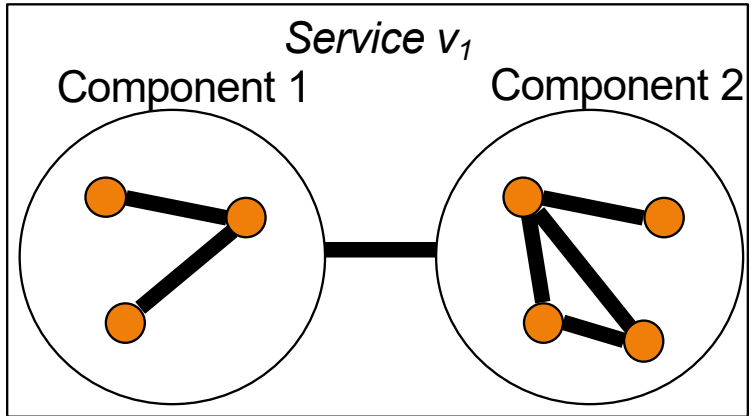
Marsal: Challenges & Innovations



- Enable a closed-loop control for computing and networking infrastructures
- Application requirements and resource characteristics change rapidly
- AI/ML analysis of the telemetry data
  - Provide load predictions
  - Detect performance degradation
  - Provide recommendation actions, etc
- Proactive and reactive re-optimization adjustments

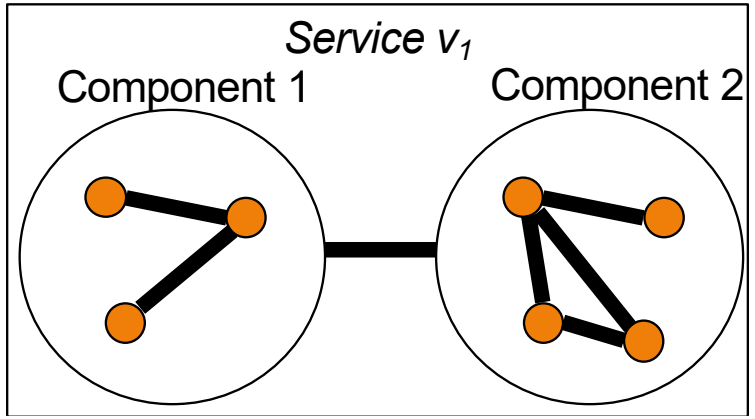
Resource allocation for cloud-native applications

- Cloud-native applications consisting of **container-based microservices** with computing requirements and networking constraints that can be represented by a graph
- Containers allow apps to be packaged and isolated with their entire runtime environment, making it easy to move them between environments while retaining full functionality
- Develop **appropriate resource allocation** mechanisms
  - Allocate microservices to the available resources
  - Consider network and cost parameters
  - Scale application's microservices vertically or horizontally based on the resource availability



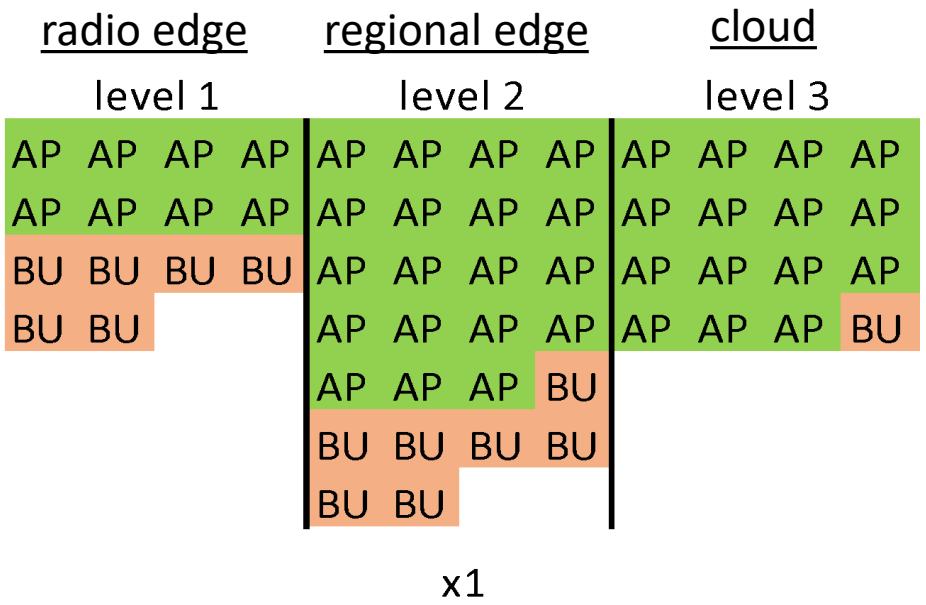
Resource allocation for cloud-native applications

- Cloud-native applications consisting of **container-based microservices** with computing requirements and networking constraints that can be represented by a graph
- Containers allow apps to be packaged and isolated with their entire runtime environment, making it easy to move them between environments while retaining full functionality
- Develop **appropriate resource allocation** mechanisms
  - Allocate microservices to the available resources
  - Consider network and cost parameters
  - Scale application's microservices vertically or horizontally based on the resource availability



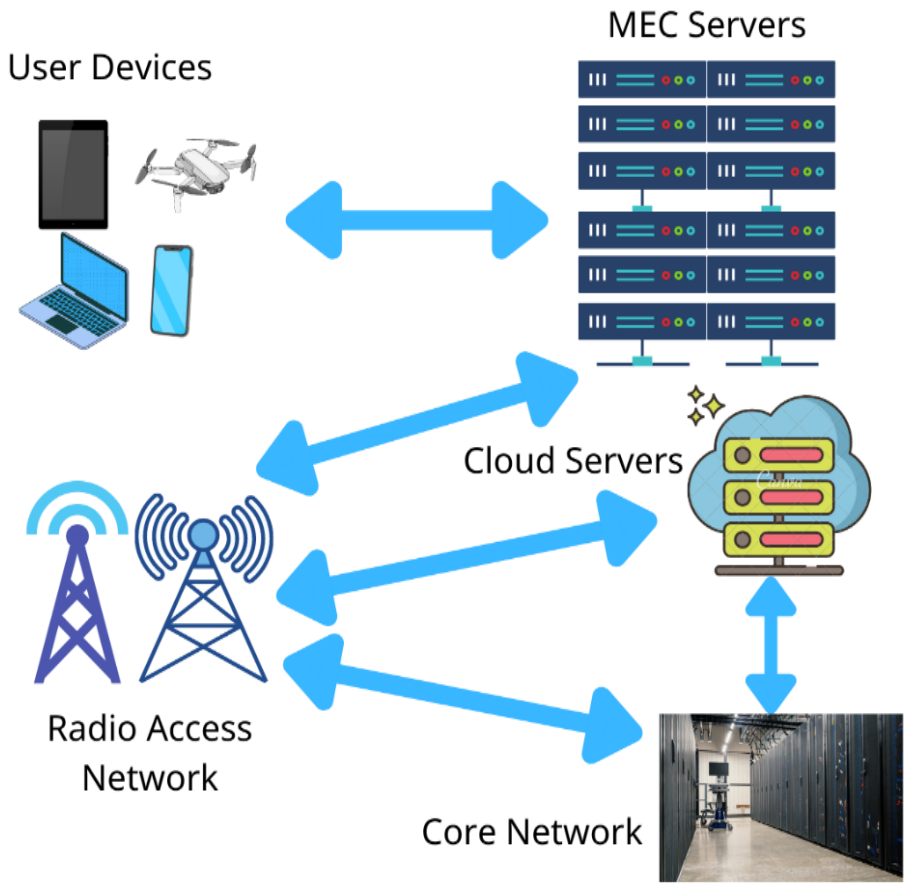
Resource allocation for cloud-native applications

- Joint allocation of baseband and generic application workload
- Baseband and application loads can be flexibly assigned to the various resources, in the same and at different levels (radio edge, regional edge, cloud)
- The baseband processing loads (BU) are divisible (CU/DU)
- The application computing loads (AP) are divisible

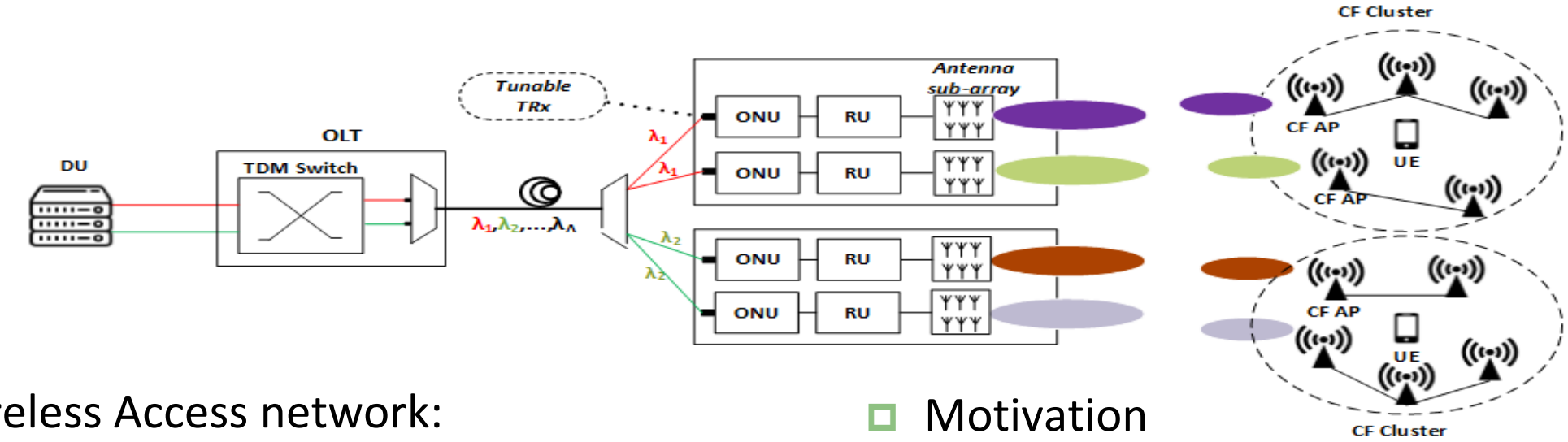


Mobile Edge Computing for B5G mobile networks

- 5G and beyond RAN utilize virtualization and edge computing
- Edge computing for mobile network can provide benefits including:
  - Lower latency, optimized network efficiency, reduced network congestion
- These benefits can assist use cases including autonomous vehicles, mobile gaming, and support for IoT application
- Cell Free networks have been proposed as a candidate solution for 6G networks.
  - Joint signal processing from many distributed access points



Resource Allocation for Joint Fiber Wireless



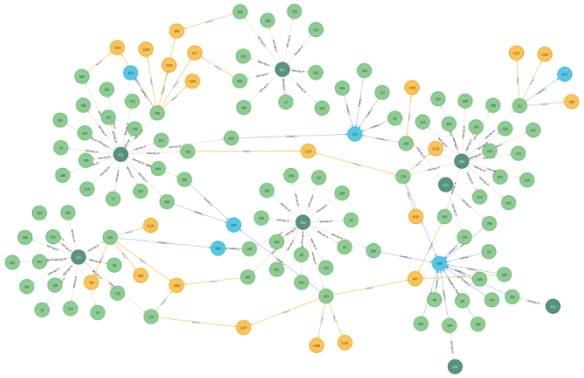
- Wireless Access network:
  - APs that operate in cell free (CF) manner
  - mMIMO BS that transmit beams towards the APs
- Wired transport network
  - TDM PON to transport the traffic to the appropriate computing nodes
- Edge computing nodes

- Motivation
  - Users are served by CF Clusters of APs
  - Clusters change dynamically in a user centric manner
  - Fronthaul traffic is served by a converged fiber wireless infrastructure
  - Develop resource allocation algorithms for the carried traffic and the required computations

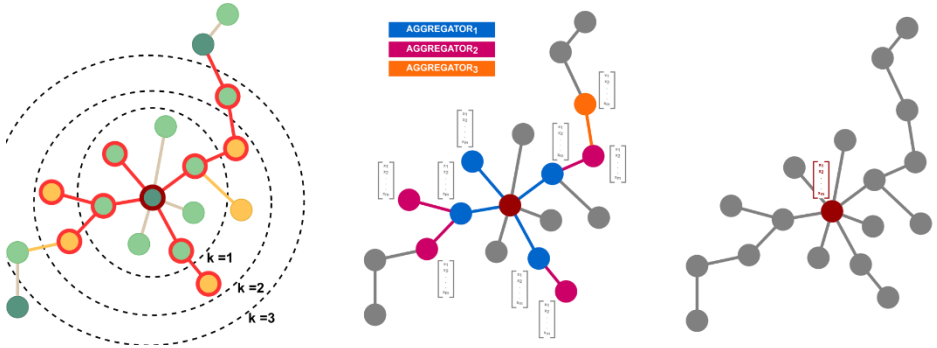
ML based event detection mechanisms

- Computing resources are subject to failures and capacity exhaustion
- Develop event detection mechanisms that can initiate re-optimization procedures
- The idea is to use knowledge graph for representing the complexities of edge-cloud continuum infrastructures
- A knowledge graph represents a network of real-world entities and illustrates the relationship between them

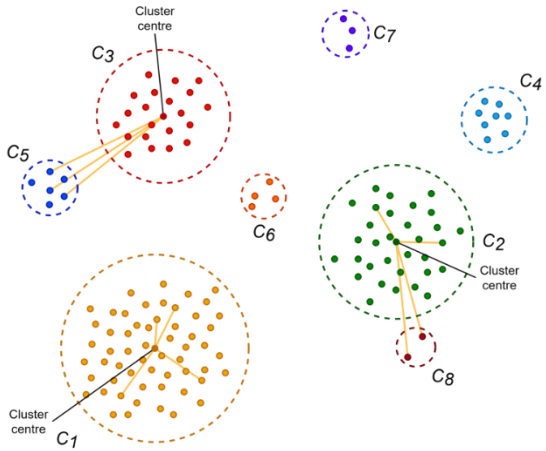
- The steps of our approach:
  - Step 1. Knowledge graph organize data of multiple sources, capture information about entities and forge connections between them
  - Step 2. Application of a graph embedding method to transform the graph entities (nodes, edges) into vectors, while preserving the graph's topology
  - Step 3. ML based Anomaly detection techniques are applied to the created embeddings
    - Density-based and distance-based algorithms



1. KG representation



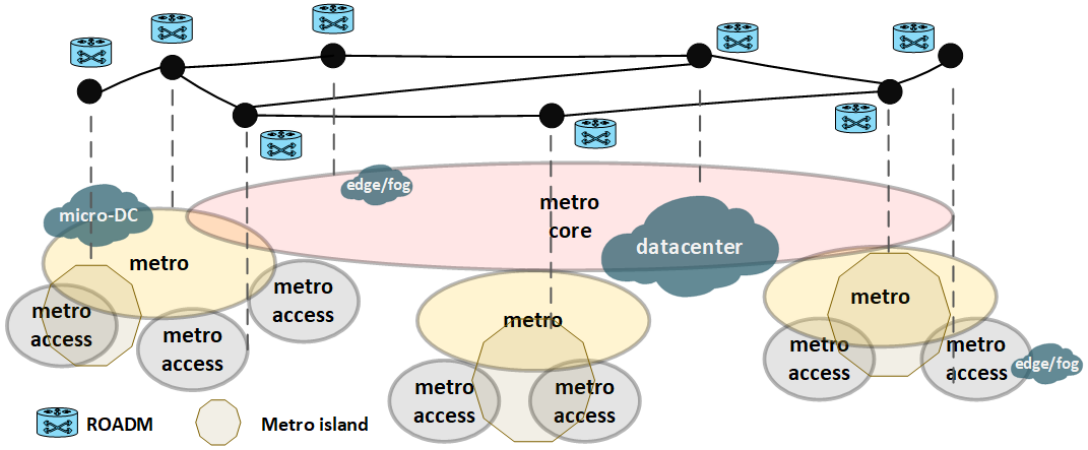
2. Graph embeddings



3. Anomaly detection

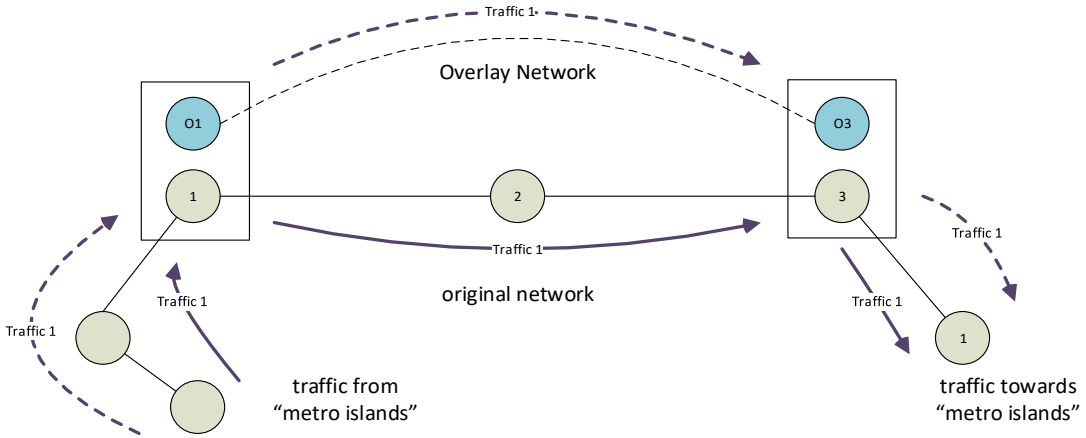
Datacenter in the City Architecture

- Address the intra-city traffic demands, interconnecting users to users and users to computing and storage resources
- The challenges are similar to the one that Datacenters face
- Provide high capacity and low latency connectivity
- The main challenges include the following:
  - Identify areas of the network with high locality – what we call metro islands
  - Create an physical or virtual overlay network for their interconnection – like highways in a city



Datacenter in the City Architecture

- Metro islands based on filterless technologies, taking advantage of the filterless broadcast and select nature
  - Wavelength waste depends on the size of the metro island
- Virtual Overlays use SDN capabilities of WSSes, ROADMS, transceivers to create the highways
  - Use wavelength and slotted (all optical) switching



Datacenter in the City Architecture

- Algorithmic Challenges
  - Selection of the nodes from the original topology where equipment of enhanced capabilities will be placed (higher capacity tranceivers, mini datacenters)
    - These nodes also aggregate the traffic from the metro island they are assigned to (sinks of the filterless network)
  - Dimensioning of metro islands capacity is important for setting up the overlay network
  - Place processing and data workloads
    - Workloads placed inside the “metro islands” are benefited by the low latency between end-devices and resources and their fast service
    - Loads that cannot be served locally and are transferred to different resources (e.g. in other “metro islands”) are benefited by the low latency and pre-established communication highways, setup through the overlay network

Questions ?

Thank You !

Machine Learning-Based, Networking and Computing Infrastructure Resource Management of 5G and Beyond Intelligent Networks

ICT-52-2020 - 5G PPP – Smart Connectivity beyond 5G

https://www.marsalproject.eu/

This Project has received funding from the European Union´s Horizon 2020 research and innovation programme under Grant Agreement No. 101017171

Institute of Communication and Computer Systems (ICCS)

High Speed Communication Networks Laboratory (HSCNL), National Technical University of Athens

http://hscnl.ece.ntua.gr/ (Director: Prof. Manos Varvarigos)

Presenter: Panagiotis Kokkinos, kokkinop@mail.ntua.gr