




5G PhD School
The International Scientific Event Companion to 5G Italy


2 | 3 | 4
dicembre
2 0 1 9

cnit consorzio nazionale
interuniversitario
per le telecomunicazioni



5G PhD School
The International Scientific Event Companion to 5G Italy

Tutorial: Data-analytics based orchestration



Albert Banchs
Professor, Carlos III University of Madrid
Deputy Director, IMDEA Networks institute

Acknowledgement: The material of this tutorial and keynote speech has been based on the work that the presenter is doing in the H2020 5G-TOURS European project (Grant Agreement No. 856950).

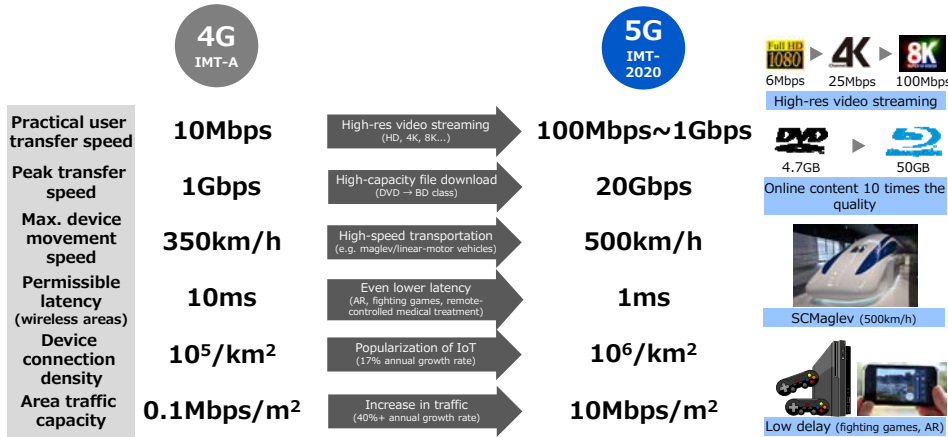
Tutorial outline

- Network slicing
- Network Softwarization
- Network orchestration
- Data analytics and AI

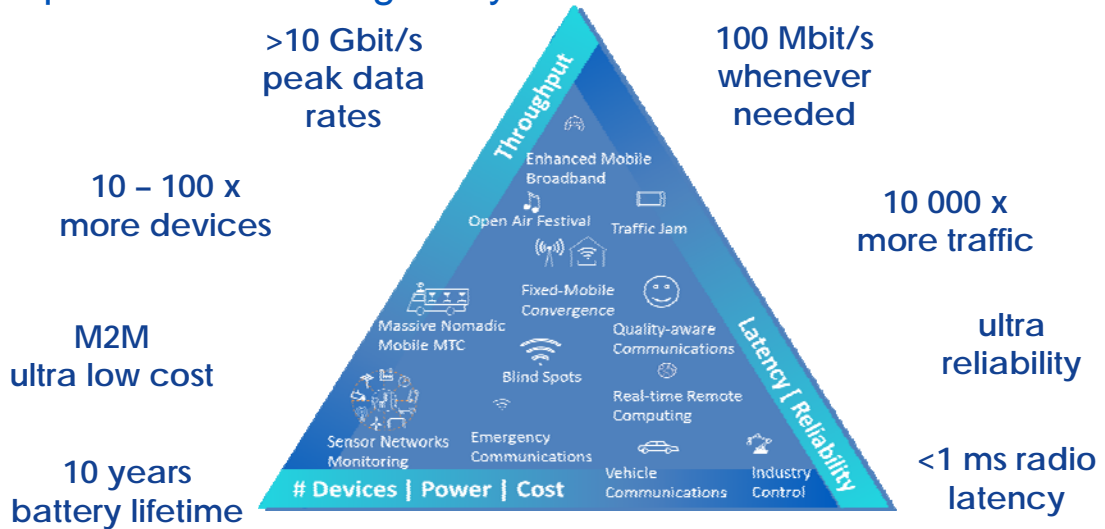
Tutorial outline

- Network slicing
- Network Softwarization
- Network orchestration
- Data analytics and AI

Requirements for 5G



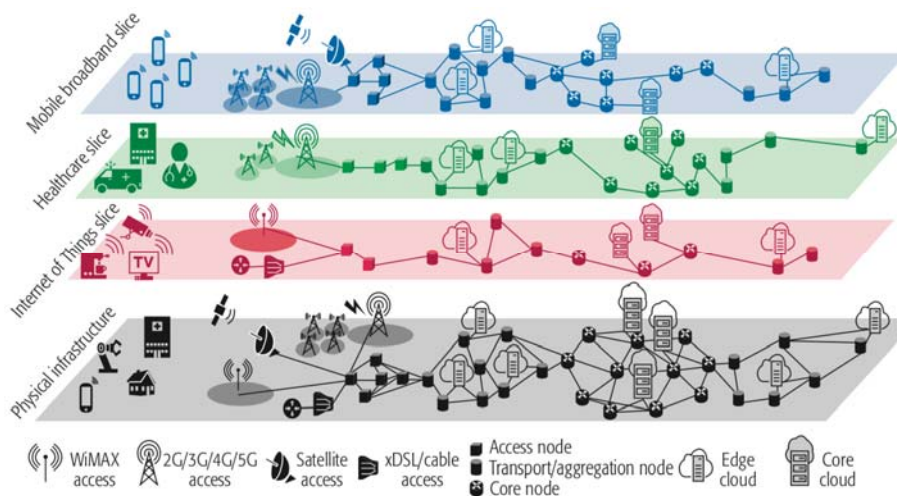
Requirements heterogeneity



What is network slicing?

- **Network Slice:** A set of infrastructure resources and service functions that has attributes specifically designed to meet the needs of an industry vertical or a service
- **Network Slicing:** A management mechanism that Network Slice Provider can use to allocate dedicated infrastructure resources and service functions to the user of the Network Slice
- **3GPP definition:**
 - “A logical network that provides specific network capabilities and network characteristics”
 - “A network created by the operator customized to provide an optimized solution for a specific market scenario which demands specific requirements with end to end scope”
 - Implemented by “slice instances”
 - Created from a “network slice template”

What is network slicing?

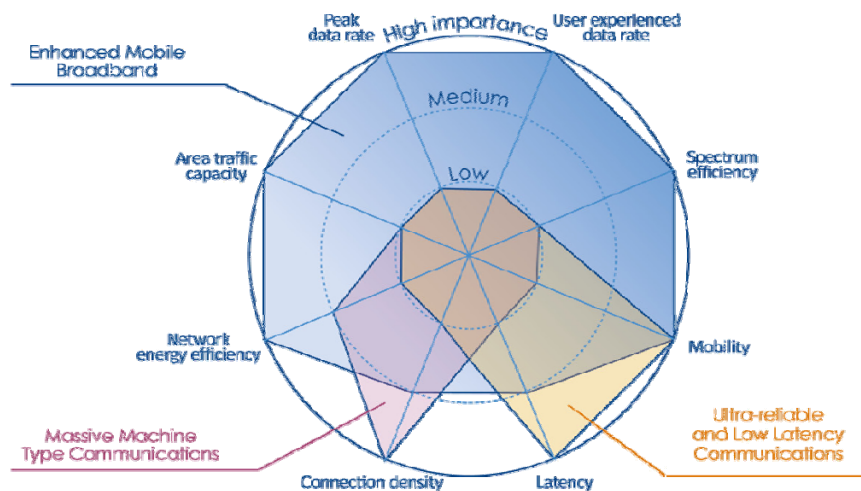


From: J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca and J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," in *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80-87, May 2017.

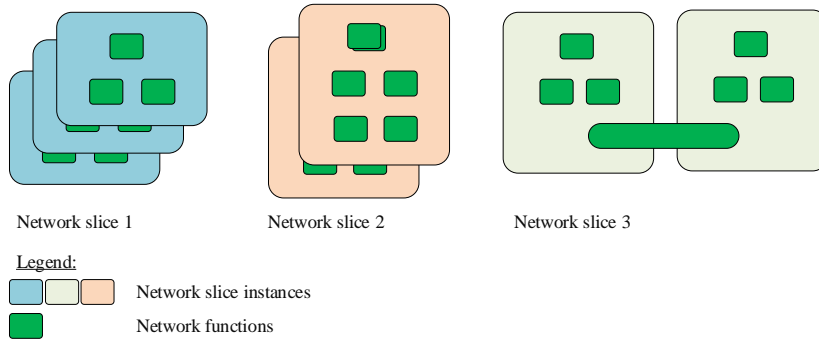
Network slice types

- **Enhanced Mobile Broadband (eMBB)** to deal with hugely increased data volumes, overall data capacity and user density
- **Massive Machine-type Communications (mMTC)** for the IoT, requiring low power consumption and low data rates for very large numbers of connected devices
- **Ultra-reliable and Low Latency Communications (URLLC)** to cater for safety-critical and mission critical applications

Slice types' requirements



Network Slices vs. Network Slice Instances



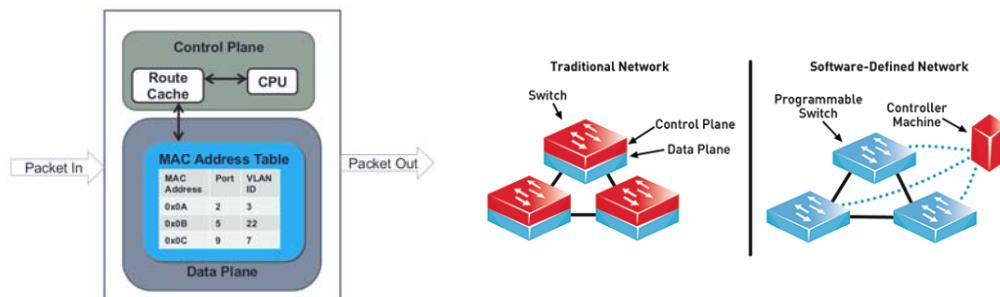
- Network Slice: A logical network that provides specific network capabilities and network characteristics.
- Network Slice instance: A set of Network Function instances and the required resources (e.g. compute, storage and networking resources) which form a deployed Network Slice.

Tutorial outline

- Network slicing
- Network Softwarization
- Network orchestration
- Data analytics and AI

Software Defined Networking (SDN)

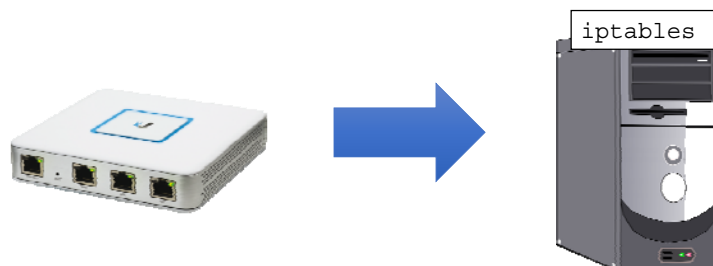
- Separate user plan from control plane, bringing controller to a centralized location
- Allow to modify the behavior of the control plane by means of “user policies”



From: McKeown et. al “OpenFlow: Enabling Innovation in Campus Networks,” ACM CCR 2008

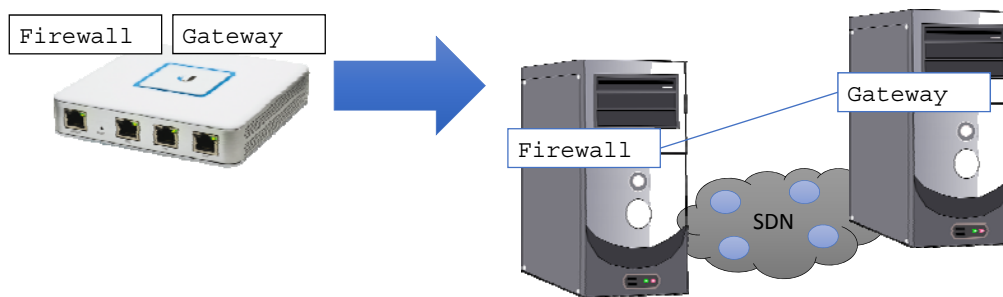
Network Function Virtualization (NFV)

- Complementary technology to SDN, which depends on SDN to deliver its benefits
- Network Function: Building block of a communication service
- E.g., gateway, load balancer, firewall



NFV definition

- Use technologies of IT virtualization to virtualize and connect network node functions



Benefits of NFV

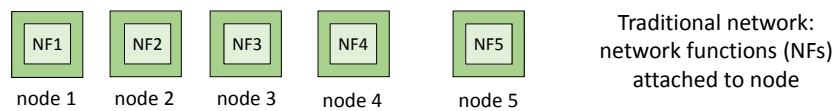
- Reduced equipment costs and reduced power consumption through consolidating equipment and exploiting the economies of scale of the IT industry
- Increased velocity of Time to Market by minimizing the typical network operator cycle of innovation
- Much more efficient test and integration
 - Production, test and reference facilities can be run on the same infrastructure
- Targeted service introduction based on geography or customer sets is possible
 - Services can be rapidly scaled up/down as required
 - Service velocity is improved by provisioning remotely

Virtualizing Network Functions



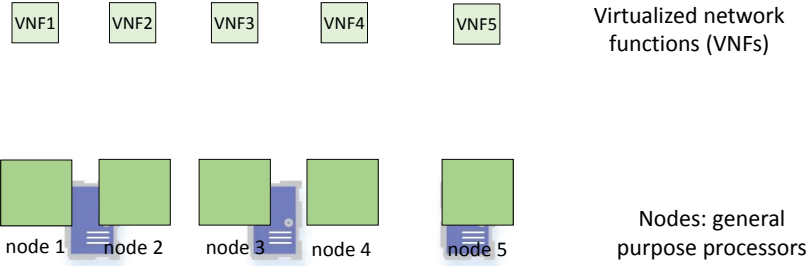
17

Virtualizing Network Functions

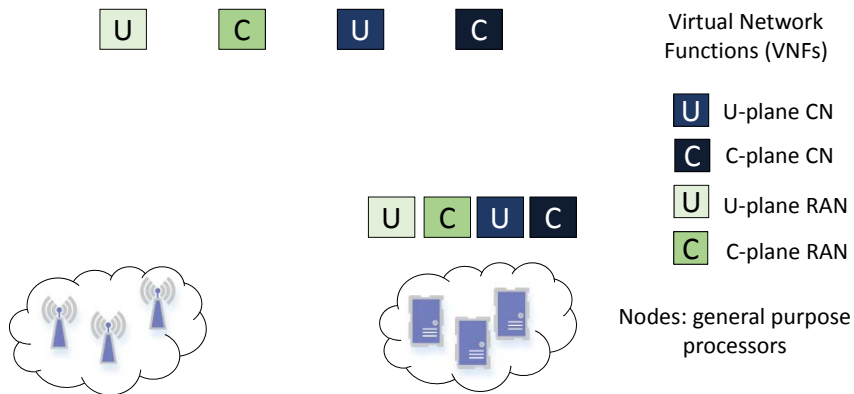


18

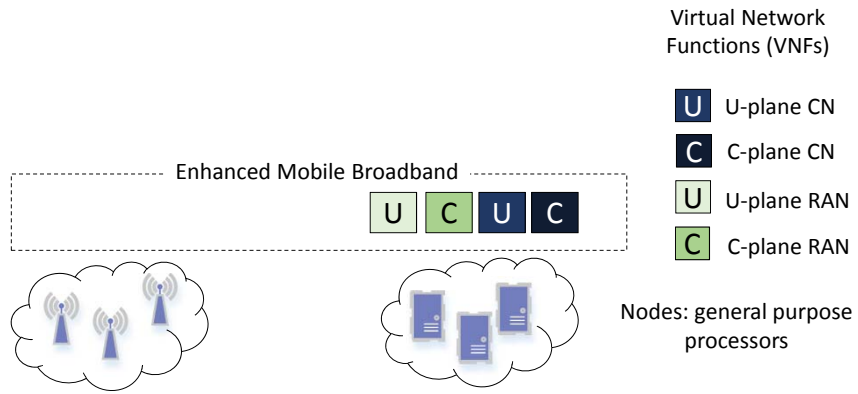
Virtualizing Network Functions



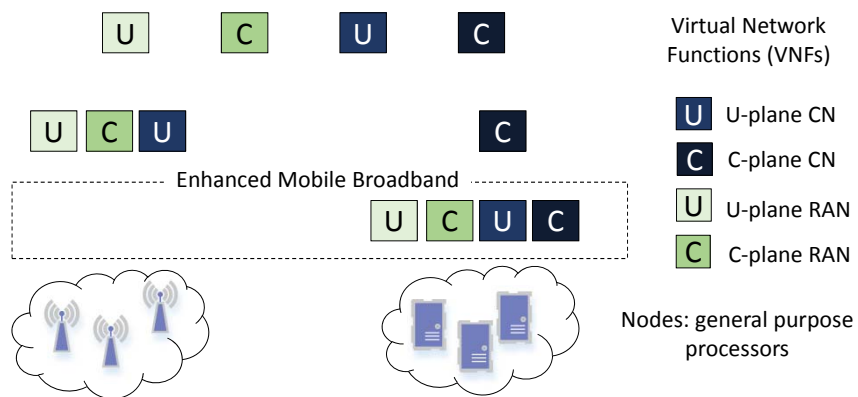
Orchestration of VNFs



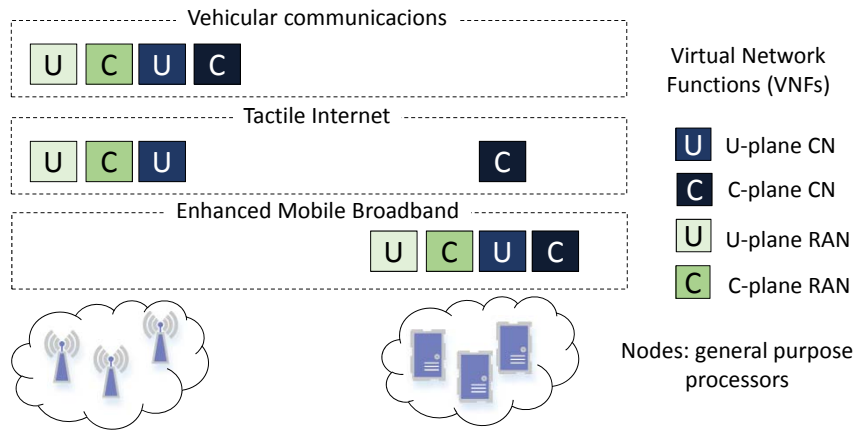
Orchestration of VNFs



Orchestration of VNFs



Orchestration of VNFs



Tutorial outline

- Network slicing
- Network Softwarization
- Network orchestration
- Data analytics and AI

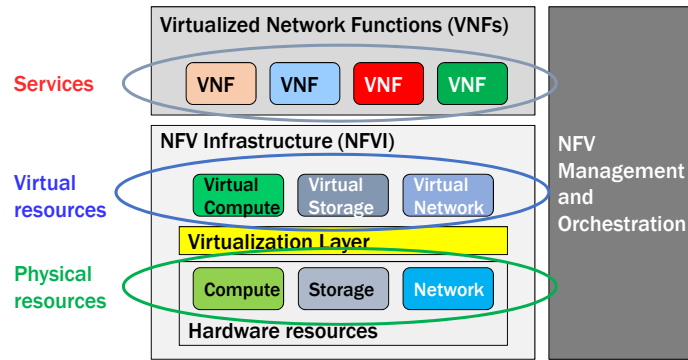
Motivation

- SDN and NFV have brought a revolutionary paradigm on network management
- This allows for enhanced network features but increases the complexity for the management of the network
- Moreover, new management functionality has to be provided
 - Network function placement
 - Resource orchestration
- Therefore, the management system in 5G needs to be heavily revisited

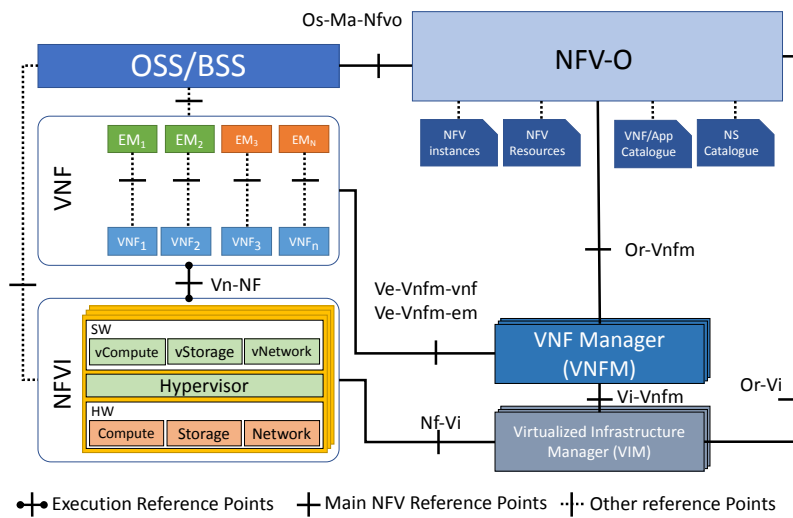
Management and Orchestration

- Also known as MANO or M&O
- Management
 - Network Function Selection
 - Network Function Configuration
 - Network Function Chaining
- Orchestration
 - Network Function Placement
 - Resource Allocation
 - Including both cloud and RAN resources

High-level NFV framework



ETSI NFV MANO Architecture

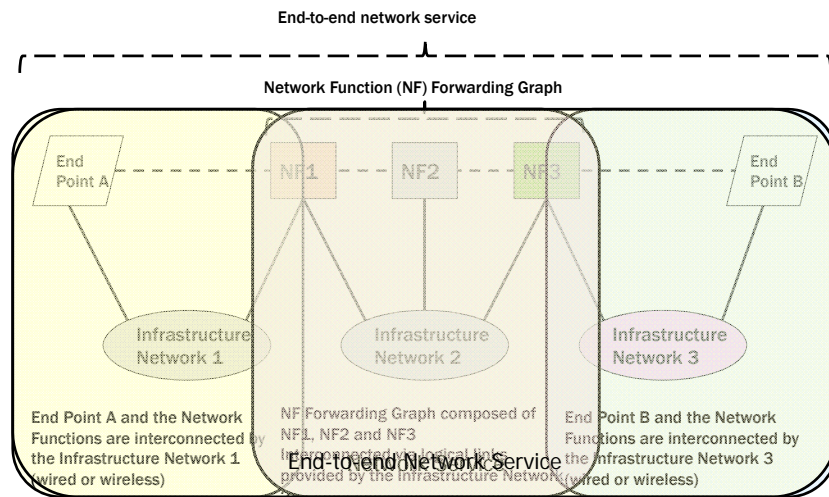


Service and Resource Orchestration

- 5G Networks relies on two different kind of orchestration
- Resource orchestration
 - Assignment to each slice of the needed resources
 - Proper configuration of the associated resources (i.e., spectrum)
 - No understanding of the “semantic” of the deployed VNFs
 - The underlying topology is also out of the scope of resource orchestration
- Service orchestration
 - Understanding the service needed by the slice and translate it into VNFs
 - Also their chaining and relation shall be provided

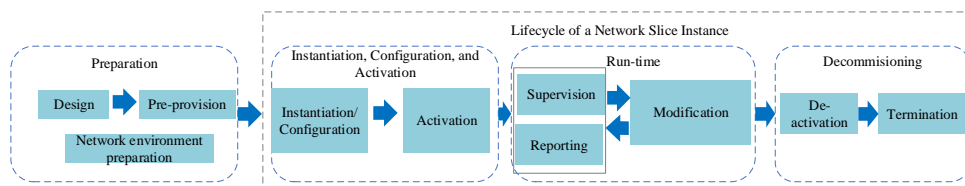
End-to-end service Orchestration

- To orchestrate an end-to-end service, the NFV Orchestrator (NFVO) instantiates the network slice as follows:
 - It issues the corresponding requests to the Software-Defined Networking (SDN) controller to instantiate connections between the different network nodes
 - It requests the Virtualized Infrastructure Manager (VIM) to reserve the virtual resources at the different network nodes
 - It commands the Virtual Network Function Manager (VNFM) to instantiate the VNFs
 - It configures the VNFs and PNFs (Physical Network Functions)



Network Slice Lifecycle management

- Preparation
 - Translation of slice requirements to network function chain
- Activation
 - Slice kick-off on the shared infrastructure
- Runtime
 - Scaling of the NF, according to the conditions
- Decommissioning
 - When the service is not available anymore



Shared and non-shared NFs

- A NSI may contain functions that are shared among network slice instances (NSIs), while other are dedicated
 - For example, a shared AMF (Access Management Function)
- When creating a new NSI, we check if there are existing shared network functions that can be used for creation of a new NSI
 - In this case they become shared NFs
 - In the case some shared network functions are available, only additional (non-shared) network functions may need to be created
 - The existing shared network functions may need to be reconfigured, and the resources supporting them may need to be added to ensure that all NSI(s) can be served
- In the case where no existing network functions are available for the new NSI, both the shared and the non shared shall be created
 - The new shared are just "shareable"
 - They only belong to the new single NSI at this point until being shared by other NSIs (where they become "shared")

Tutorial outline

- Network slicing
- Network Softwarization
- Network orchestration
- Data analytics and AI

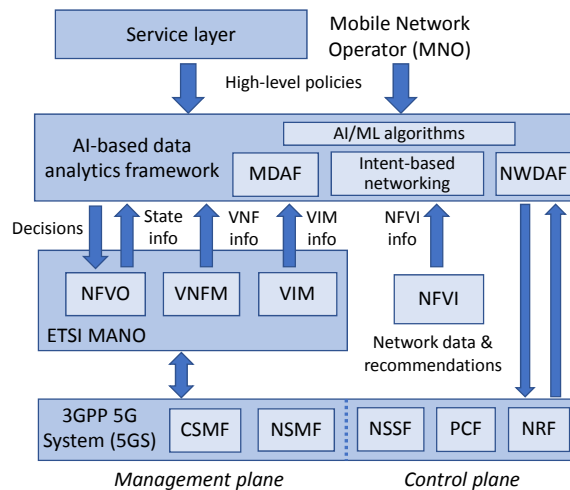
Artificial intelligence & data analytics

- AI is a computation paradigm that endows machines with intelligence
 - Aiming to teach them how to work, react, and learn like humans
 - Many techniques fall under this broad umbrella
- Machine learning enables the artificial processes to absorb knowledge from data and make decisions without being explicitly programmed
 - Data needs to be collected and made available to AI algorithms
 - Machine learning is closely related to data analytics
- Machine learning has become very popular driven by:
 - Modern challenges are “high-dimensional” in nature
 - We have rich data sources and processing power that can be use to solve problems
 - Machine learning can be integrated into working software to support products demanded by industry
- In line with the rising popularity of machine learning, this tool is being widely used for many networking problems including 5G

Data analytics and Artificial Intelligence for Orchestration

- Artificial Intelligence is a natural choice for driving orchestration decision
 - We need to make predictions, classifications and decisions based on data
- 3GPP has identified this and started efforts towards defining an AI-based Data Analytics
 - Autonomous and efficient control, management and orchestration
- Modules defined by 3GPP to this end
 - Network Data Analytics Function (NWDAF)
 - Management Data Analytics Function (MDAF)
- Standardization efforts are still ongoing
 - There is no current full-blown data analytics-assisted architecture ready

AI-based data Analytics framework



Data analytics for the control plane

- In the control plane, analytics allow NFs to optimize their behavior at run-time, typically at a much faster speed than what network management and orchestration systems allow
- NWDAF analytics can be leveraged to improve
 - Slice-level load balancing
 - Service experience and Quality of Experience (QoE)
- Examples of data analytics usage
 - NSSF: Selecting the set of Network Slice instances serving a UE
 - PCF: Unified policy framework to govern network behavior, including the QoS parameters
 - NRF: Selection of a NF instance when a certain NF type is needed

Examples of data analytics for the control plane

- NSSF: Slice selection
 - NWDAF: monitors both load status and service experience statistics and predictions
 - Slice selection and load control functionality to decide which slice optimally serves each of the new UEs arriving in the network
- PCF: QoS control
 - Informed by NWDAF analytics on UE and application service experience
 - Adapt service QoS parameters across all UEs on a slice in such a way that the slice SLA is satisfied.
- NRF: Selection of NF instance
 - Keep NF profile of all NFs belonging to a slice, including their instantaneous load
 - Pre-selection step so that not only instantaneous NF load is taken into account, but also statistics and predictions
 - Load balancing is embedded in the selection process of the new NF instance among the candidate set

Data analytics for the management plane

- Data used as input by the AI-based analytics framework
 - NFV Infrastructure (NFVI): knowledge on the computational resources' capabilities (such as the type of CPU and memory, accelerators, etc.) along with their availability (i.e., the status and utilization level)
 - MANO system: requirements of the network slices
- Decision taken
 - NFVO: NF placement and resource allocation decisions while ensuring that the resulting resource allocation satisfies the respective slice SLA
 - VNFM: Run-time up and down scaling of resources
 - CSMF (Communication Service Management Function) and NSMF (Network Slice Management Function (NSMF): Admission control of new slices

Keynote: Resource Allocation for network slicing



Albert Banchs

Professor, Carlos III University of Madrid
Deputy Director, IMDEA Networks institute

Acknowledgement: The material of this tutorial and keynote speech has been based on the work that the presenter is doing in the H2020 5G-TOURS European project (Grant Agreement No. 856950).

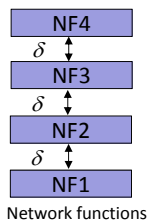
Keynote outline

- Research challenges with network slicing & orchestration
- Analysis of the benefits of dynamic orchestration
- Realizing dynamic orchestration with machine learning

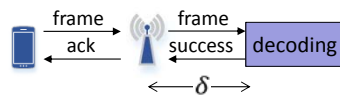
Protocol Stack redesign

- Current protocol stack
 - Designed considering that certain functions are co-located and can exchange data with no latency
 - This has introduced temporal dependencies that limit the flexibility in the placement
- Research challenge: remove tight constraints

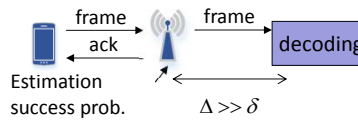
Current stack:



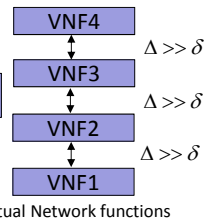
current HARQ:



Opportunistic HARQ:

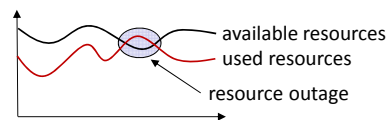
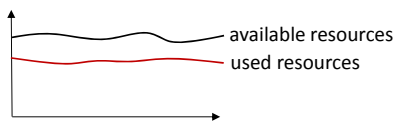


Redesign:

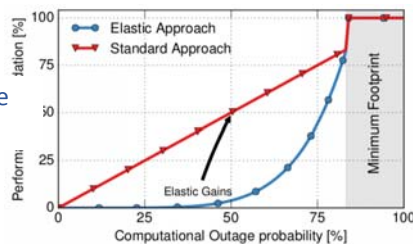


Bringing resource awareness: design of elastic VNFs

- Traditional environment for wireless functions
 - Capacity and computational load are known: Resources are always available
- Environment with virtualized wireless functions
 - Fluctuations on the capacity and load: resource outages may occur



Graceful performance degradation



Designing the algorithms

- 3GPP provides the definition of the modules and the interfaces
 - However, the algorithms run by the different modules are not specified
 - The internals of the different modules are not in the scope of the standards
 - Furthermore the standards are still at a very early stage
 - Research work is required instead to fill this gap
- These are new paradigms that require completely new algorithms
 - Algorithms to determine the required resource allocation for the different VNFs to ensure that the respective SLAs with the tenants are met
 - Algorithms to determine the best location for the different VNFs
 - We need to account both for communications and for computing resources
 - Artificial Intelligence is a natural candidate for many of these problems

Some research results

- Resource sharing for network slicing
 - "Network slicing games: Enabling customization in multi-tenant networks", [IEEE INFOCOM 2017](#)
 - "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads", [IEEE/ACM ToN 2017](#)
 - "Network Slicing Games: Enabling Customization in Multi-Tenant Mobile Networks", [IEEE/ACM ToN 2019](#)
- Resource allocation for network slicing
 - "Mobile traffic forecasting for maximizing 5G network slicing resource utilization", [IEEE INFOCOM 2017](#)
 - "RL-NSB: Reinforcement Learning-Based 5G Network Slice Broker", [IEEE/ACM ToN 2019](#)
- Admission control for network slicing
 - "Optimising 5G infrastructure markets: The business of network slicing", [IEEE INFOCOM 2017](#)
 - "Network slicing for guaranteed rate services: Admission control and resource allocation games", [IEEE TWC 2018](#)
 - "A machine learning approach to 5G infrastructure market optimization", [IEEE TMC 2019](#)
- Orchestration efficiency
 - "How should i slice my network?: A multi-service empirical evaluation of resource sharing efficiency", [ACM MOBICOM 2018](#)
 - "Resource sharing efficiency in network slicing", [IEEE TNSM 2019](#)
- Orchestration algorithms
 - "DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning," [IEEE INFOCOM 2019](#)
 - "Optimizing Resource Provisioning in Network Slicing with AI-based Capacity Forecasting", [IEEE JSAC 2019](#)
- Elastic VNFs
 - "CARES: Computation-Aware Scheduling in Virtualized Radio Access Networks" [IEEE TWC 2018](#)
 - "vrAI: A Deep Learning Approach Tailoring Computing and Radio Resources in Virtualized RANs", [ACM MOBICOM 2019](#)
- Orchestration architecture/implementation
 - "POSENS: a practical open source solution for end-to-end network slicing", [IEEE Wireless Communications 2018](#)
 - "A 5G Mobile Network Architecture to Support Vertical Industries", [IEEE Communications Magazine 2019](#)

Focus of the keynote

- Analyzing the advantages of dynamic orchestration
 - Gains resulting from dynamically adjusting the resources allocated to different slices
 - Shows the need for devising intelligent orchestration algorithms
 - Publication at ACM MOBICOM 2018
- Design of an intelligent orchestration algorithm
 - Example of how machine learning can be used to address a mobile network problem
 - Realizing the gains resulting from the above analysis
 - In line with the MDAF module considered by 3GPP
 - Publication at IEEE INFOCOM 2019

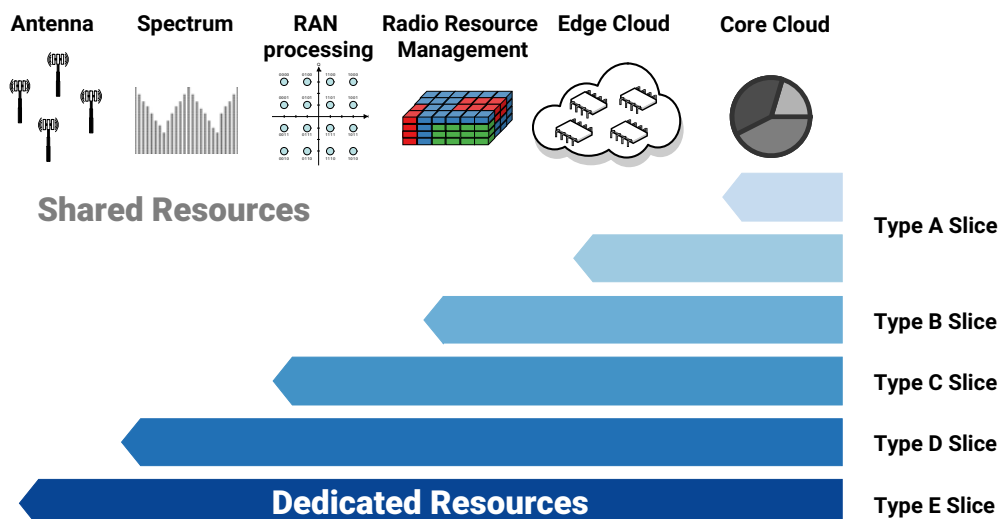
Keynote outline

- Research challenges with network slicing & orchestration
- Analysis of the benefits of dynamic orchestration
- Realizing dynamic orchestration with machine learning

Empirical evaluation of network slicing efficiency

- Following a data driven approach we want to
 - Quantify the price paid in efficiency when suitable algorithms for dynamic resource allocation are not available, and the operator has to resort to physical network duplication
 - Evaluate the impact of sharing resources at different levels of the network, including the cloudified core, the virtualized radio access, or the individual antennas
 - Outline the benefit of dynamic resource allocation at different timescales under various slice specifications
- Methodology
 - Our approach can be used for generic kinds of resource allocation
 - Still, it is not an optimization, but rather an indication of how well slices will behave

Slicing types



Network slicing model

Discrete set of levels

- Antenna $\ell = 1$
- Complete cloud $\ell = L$
- Set of network nodes C_ℓ

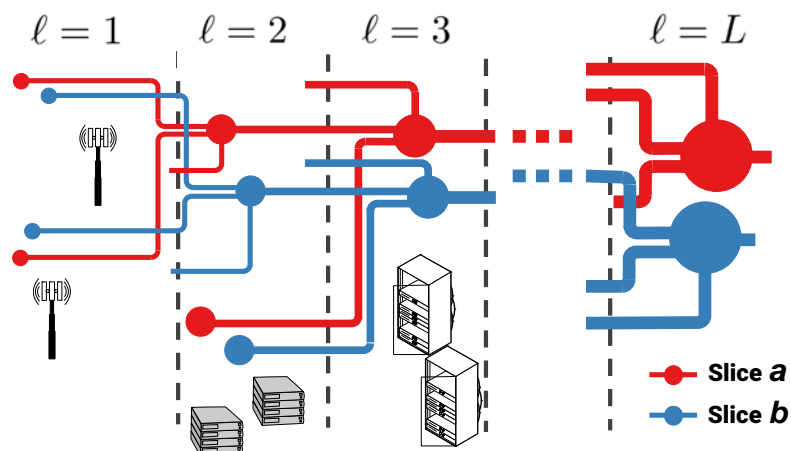
Network slice specifications \mathbb{Z} $\mathbb{Z} = (f, w)$

- Guaranteed time fraction f
 - The percentage of time the operator is engaged to fully serve a slice
- Averaging window length w
 - The above is meant over a set of discrete time intervals

Reconfiguration intervals

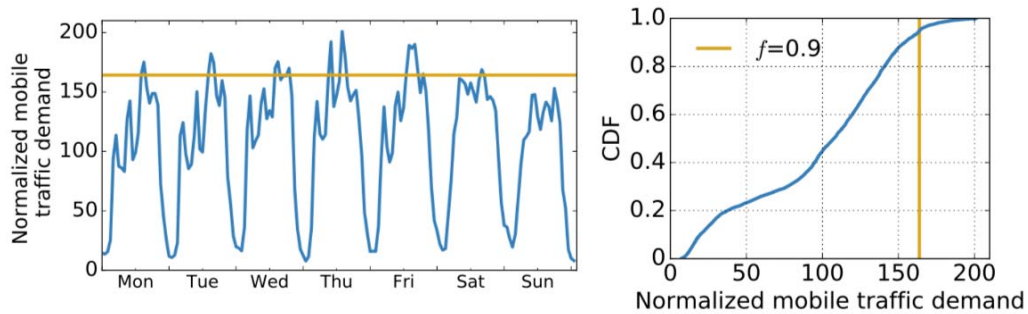
- Each slice can be re-orchestrated every interval $\tau \gg w$

Network level & Aggregation



Meeting slice requirements

f 90% w : 1 hour



Efficiency evaluation

We evaluate the efficiency of a multi-slice scenarios by comparing

- A sliced scenario in which we need to statically provision each slice with the necessary resources to meet the slice requirements

$$\mathbb{R}_{\ell, \tau}^z = \sum_{s \in \mathcal{S}} \sum_{c \in \mathcal{C}_\ell} \sum_{n \in \mathcal{T}} \tau \cdot \hat{r}_{c,s}^z(n).$$

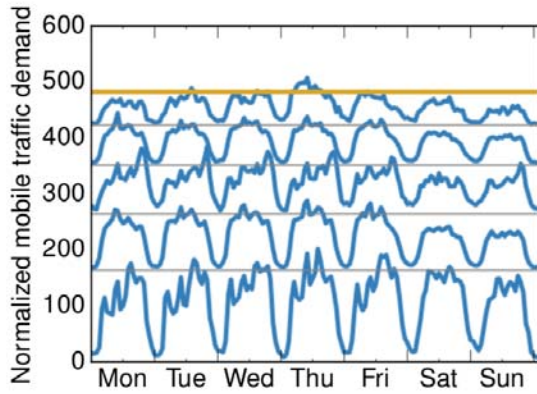
- A perfect slicing scenario, in which the exact amount of resources are shared instantaneously among all slides

$$\mathbb{P}_{\ell, \tau}^z = \sum_{c \in \mathcal{C}_\ell} \sum_{n \in \mathcal{T}} \tau \cdot \hat{r}_c^z(n),$$

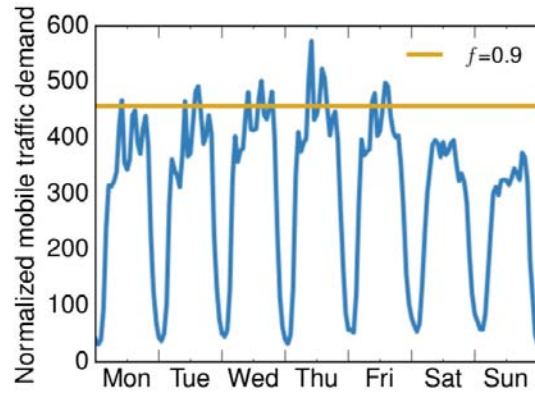
- Ideal algorithm
- Very difficult to implement in practice
- Used in this analysis as benchmark

Efficiency example

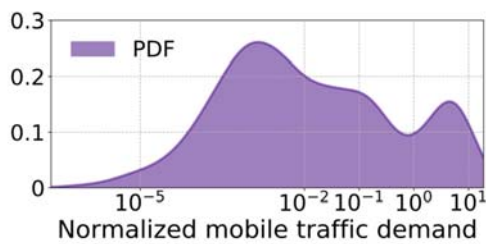
f 90%



w 1 hour

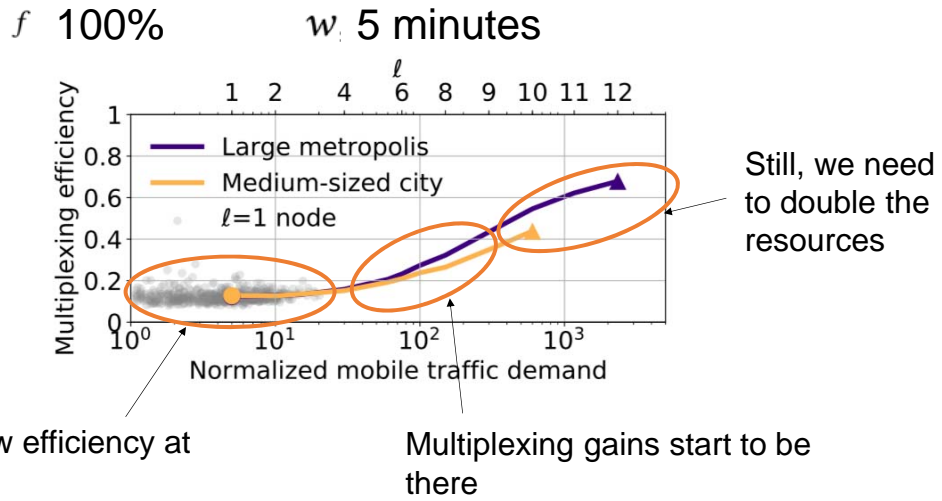


Empirical evaluation

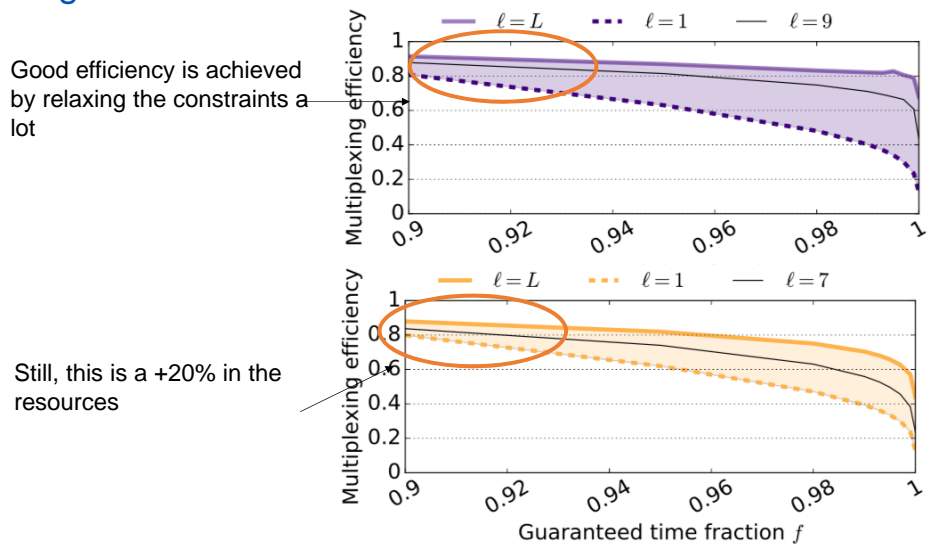


Two large cities
Three months of data
Granularity in space: sector
Granularity in time: 5 minutes
38 services in total

Global efficiency view

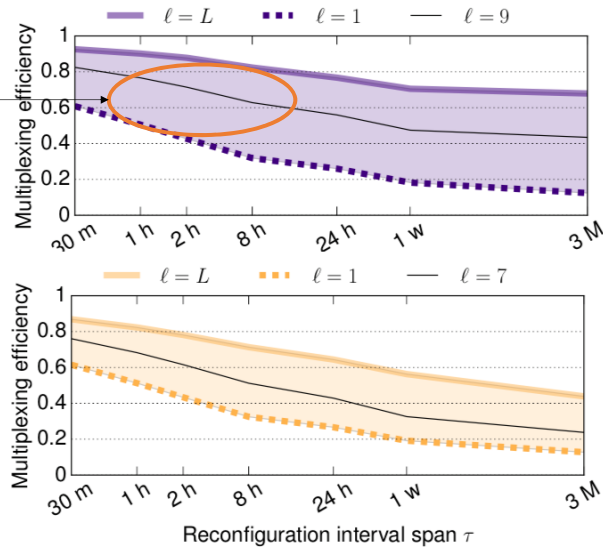


Relaxing the constraints



Reconfiguration interval impact

Reconfiguration sweetspot is here



Either we allow such timescale, otherwise we don't have much gain over static

A model for resource deployment

The previous efficiency model is good to evaluate continuous time efficiency

- OPEX scenario (i.e., maintenance, dynamic resource assignment)

Extension of the model to consider an operator point of view

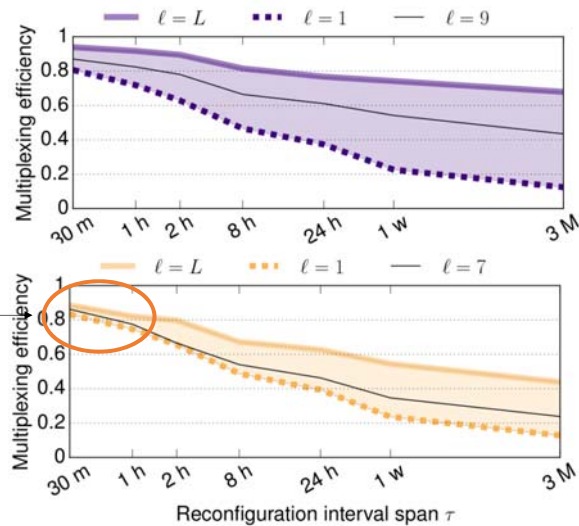
- CAPEX scenario (i.e., size of the deployed infrastructure)

$$\mathbb{R}_{\ell, \tau}^{\star z} = \sum_{s \in \mathcal{S}} \sum_{c \in \mathcal{C}_{\ell}} \max_{n \in \mathcal{T}} (\hat{r}_{c, s}^z(n))$$

$$\mathbb{P}_{\ell, \tau}^{\star z} = \sum_{c \in \mathcal{C}_{\ell}} \max_{n \in \mathcal{T}} (\hat{r}_c^z(n))$$

Deployment scenario

Similar efficiency is achieved across aggregation levels



Takeaway messages

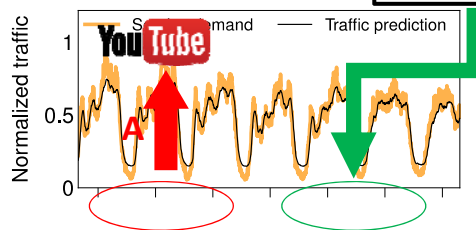
- Multi-service requires more resources
 - At least 20% more in the less challenging scenario
- Geography has limited impact
 - The two cities considered show similar trends
- Direction is a factor
 - Uplink is more challenging
- Moderating the requirements may not help
 - Good efficiency values are only achieved with non realistic service requirements
- Reconfiguration plays a key role
 - We need of orchestration algorithms that allow to dynamically re-allocate resources
 - Deployment cost may be mitigated: crucial for the 5G deployment

Keynote outline

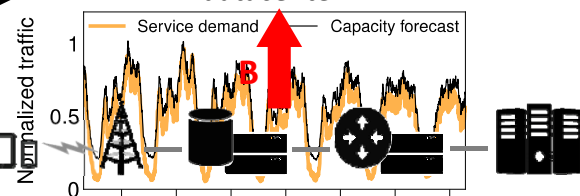
- Research challenges with network slicing & orchestration
- Analysis of the benefits of dynamic orchestration
- Realizing dynamic orchestration with machine learning

Capacity vs Demand forecasting

- Traditional approaches deal with **DeepCog** forecasting



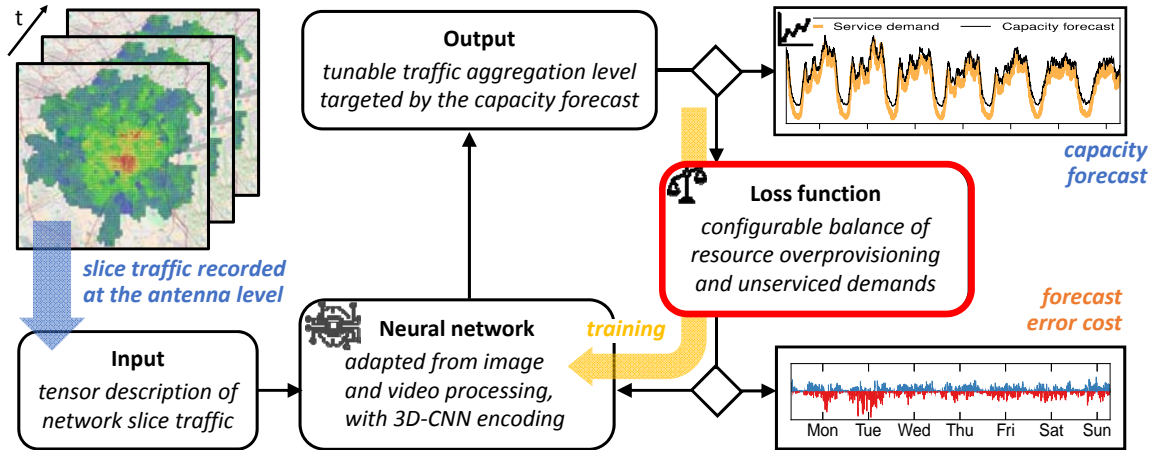
A traffic demand forecasting algorithm aims to minimize the error wrt to the original data, so **underestimation is possible**



A capacity forecasting algorithm minimizes the amount of resources needed to serve a given demand

DeepCog

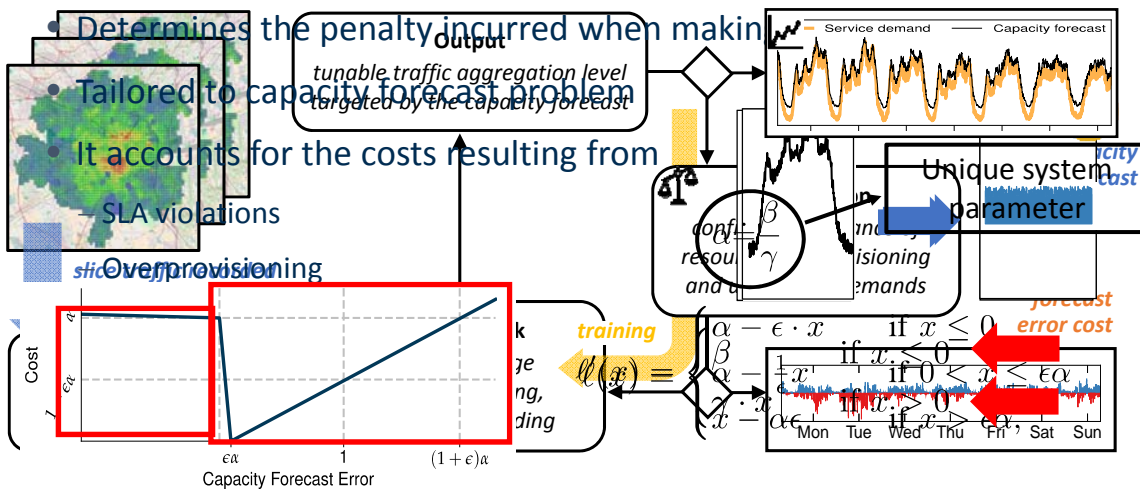
- DeepCog's design follows a deep learning approach

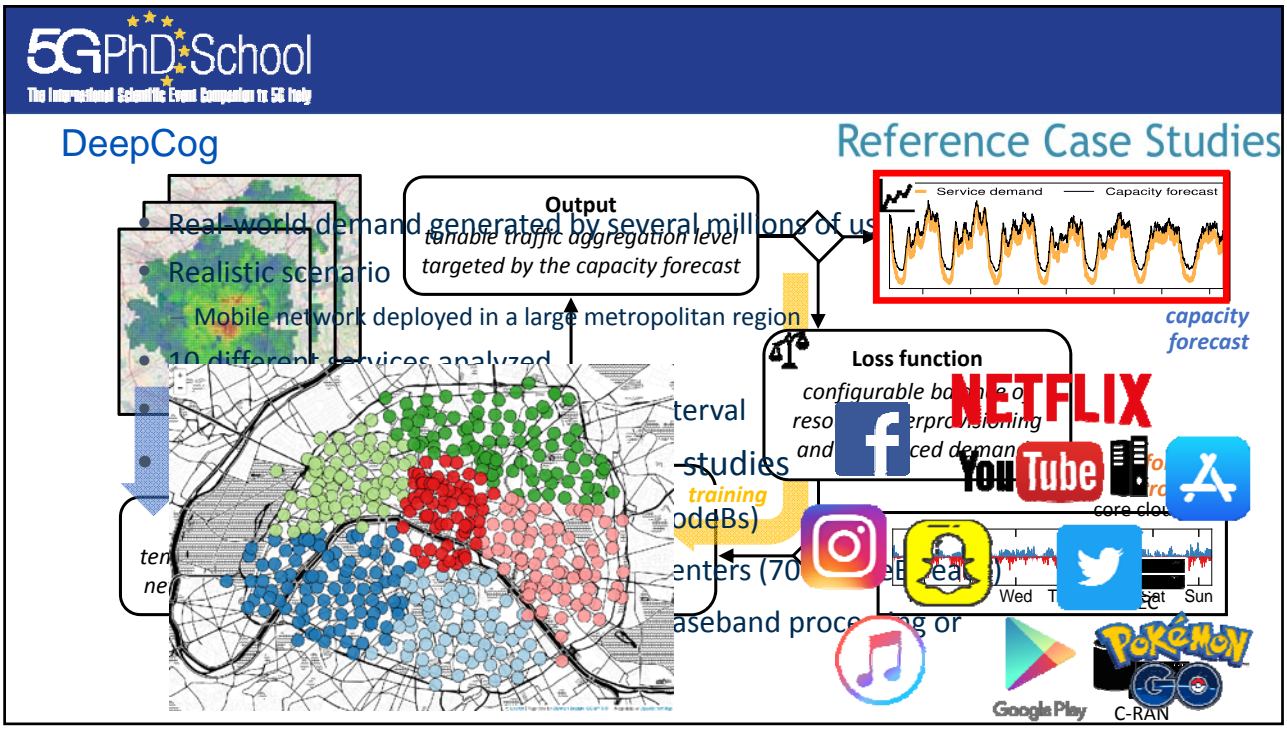
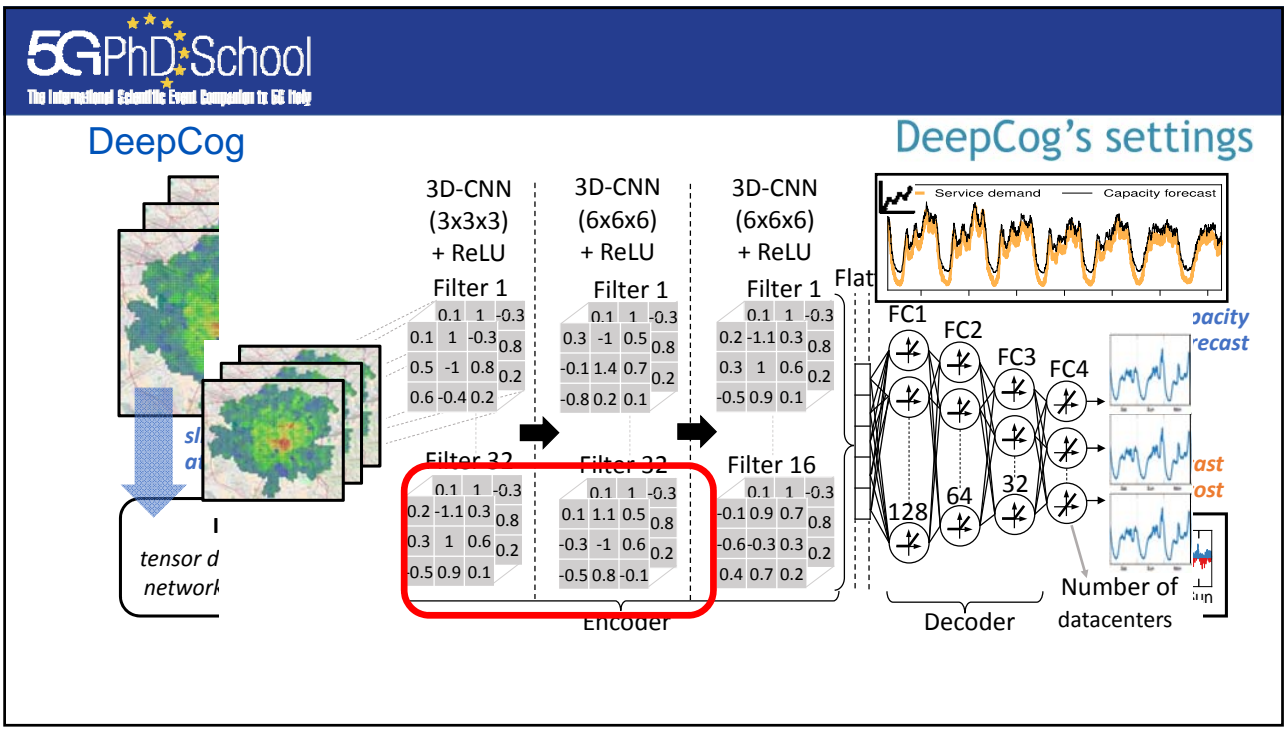


DeepCog

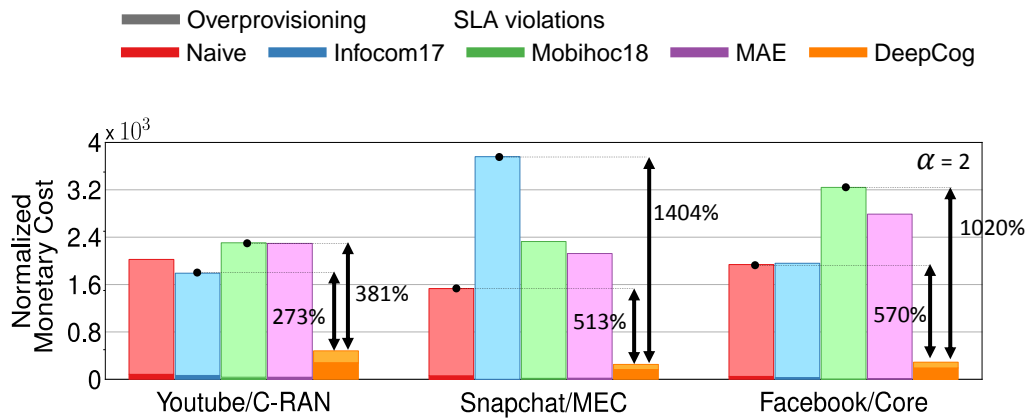
Loss function

- DeepCog's design follows a deep learning approach



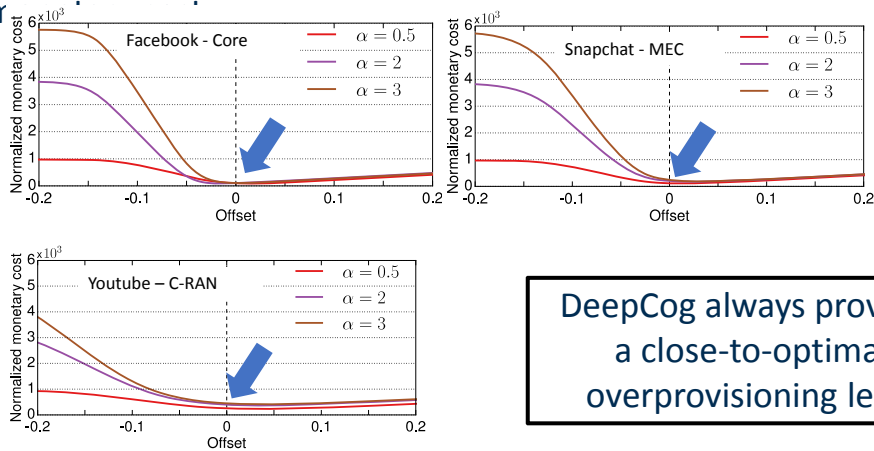


Results



Monetary cost minimization

- DeepCog provides an overprovisioning level (x-axis origin) that entails a given α



DeepCog always provides a close-to-optimal overprovisioning level

Conclusions

- DeepCog represents a novel data analytics tool for cognitive resources management in sliced 5G networks
- Leverages on Deep Neural Network structure
- Customized loss function employed aiming at capacity forecasting
- First work to date where DL architecture is explicitly tailored for mobile networks problem
- Extensive evaluations with real-world data show the substantial advantages provided by DeepCog

Questions?



Albert Banchs

Professor, Carlos III University of Madrid
Deputy Director, IMDEA Networks Institute

Thanks to Marco Gramaglia, Dario Bega, Cristina Marquez, Pablo Serrano,
Xavier Costa, Marco Fiore, Vincenzo Sciancalepore, Andres Garcia-Saavedra