# Realizing network challenges for Network Slicing

Pablo Serrano

http://www.it.uc3m.es/pablo/

IEEE Latincom 2019, Salvador

uc3m | Universidad Carlos III de Madrid

# About me

**Dr. Pablo Serrano (IEEE SM)**

Associate Professor, Univ. Carlos III de Madrid (UC3M)

Past visiting positions at:

Univ. Massachusetts Amherst, Univ. of Edinburgh, Trinity College Dublin, Telefónica R+D Barcelona

**Current project:**

- 5G-EVE: European 5G Validation platform for Extensive trials

**Research interests:**

- 5G, Wireless Communications, Performance Analysis, Energy Efficiency, Experimental Research, Testbeds

# About UC3M

- Universidad Carlos III de Madrid (UC3M)
  - Act of the Spanish Parliament on 5 May 1989
  - First Chancellor was Professor Gregorio Peces-Barba
  - Approx. 20k students
  - Highest average grade achieved by students in Madrid
- Internationalisation
  - 20% of students at UC3M are foreign
  - Higher at both master's (30%) and doctoral (43%) levels.
  - 51% graduates have participated in international mobility programmes
- Among the top 150 best universities for employability
  - It has risen by 20 places in the QS Graduate Employability Ranking 2020
  - 92,3% found work in the first year after graduation.
- Amongst the best universities worldwide in 6 fields (incl. CompSci)

# Project involvement

- I have been involved in several H2020 projects that trailblazed a new family of concepts
- 5G-NORMA
  - Network Softwarization at all layers, including RAN (cf. xRAN and ORAN)
- 5G-MoNArch
  - Big Data Driven Networking
  - Application of AI to network management (cf. ETSI ENI)
  - Elastic resource management (cf. Rel 17 study items)
- 5G EVE, 5G VINNI
  - Large scale testbeds

# Contact

**Pablo Serrano**

Univ. Carlos III de Madrid (UC3M)

pablo@it.uc3m.es

@pablo_uc3m

http://www.it.uc3m.es/pablo/

Currently looking for motivated people

- PhD student positions available

# Outline of this talk

- Network Slicing supports the instantiation of a logical network to support a service


- Orchestration
  - What gains are expected? How to do it?
- Virtualization
  - What are the challenges? How to address them?

# WHAT IS NETWORK SLICING?

# What is 5G? (2018)

WIKIPEDIA
The Free Encyclopedia

***"5G*** is a marketing term for some new mobile technologies.[*according to whom?*] Definitions differ[*citation needed*] and confusion is common[*citation needed*]. The ITU IMT-2020 standard provides for speeds up to 20 gigabits per second and has only been demonstrated with millimeter waves of 15 gigahertz and higher frequency. The more recent 3GPP standard includes any network using the NR New Radio software." (Sept. 4th, 2018)

# What is 5G? (2019)

WIKIPEDIA
The Free Encyclopedia

"'**5G**´ is the fifth generation cellular network technology. The industry association 3GPP defines any system using "5G NR" (5G New Radio) software as, "5G", a definition that came into general use by late 2018. Others may reserve the term for systems that meet the requirements of the ITU IMT-2020." (Oct. 29th, 2019)

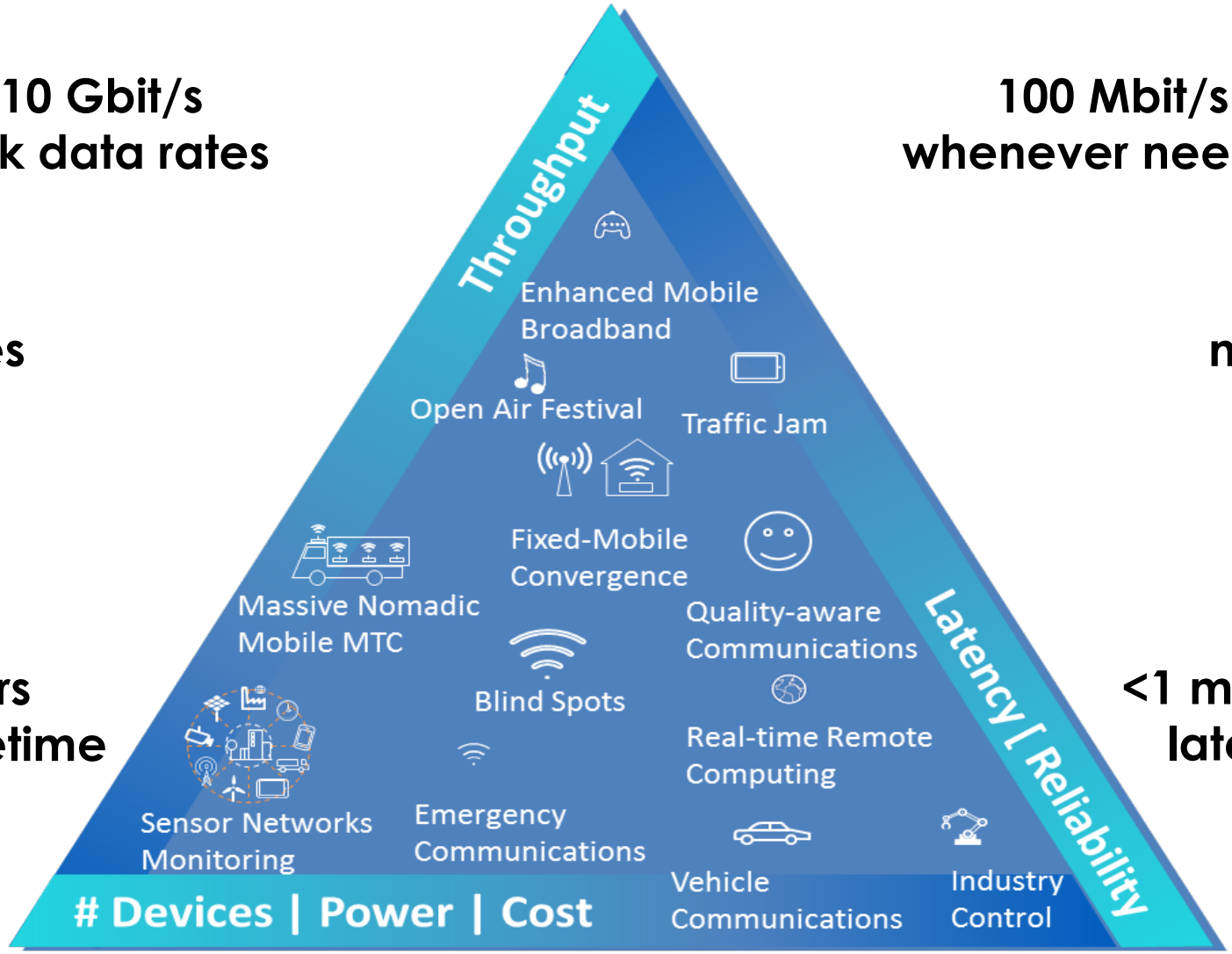
# Requirements: Heterogeneity

**>10 Gbit/s**
**peak data rates**

**10 – 100 x**
**more devices**

**M2M**
**ultra low cost**

**10 years**
**battery lifetime**




**100 Mbit/s**
**whenever needed**

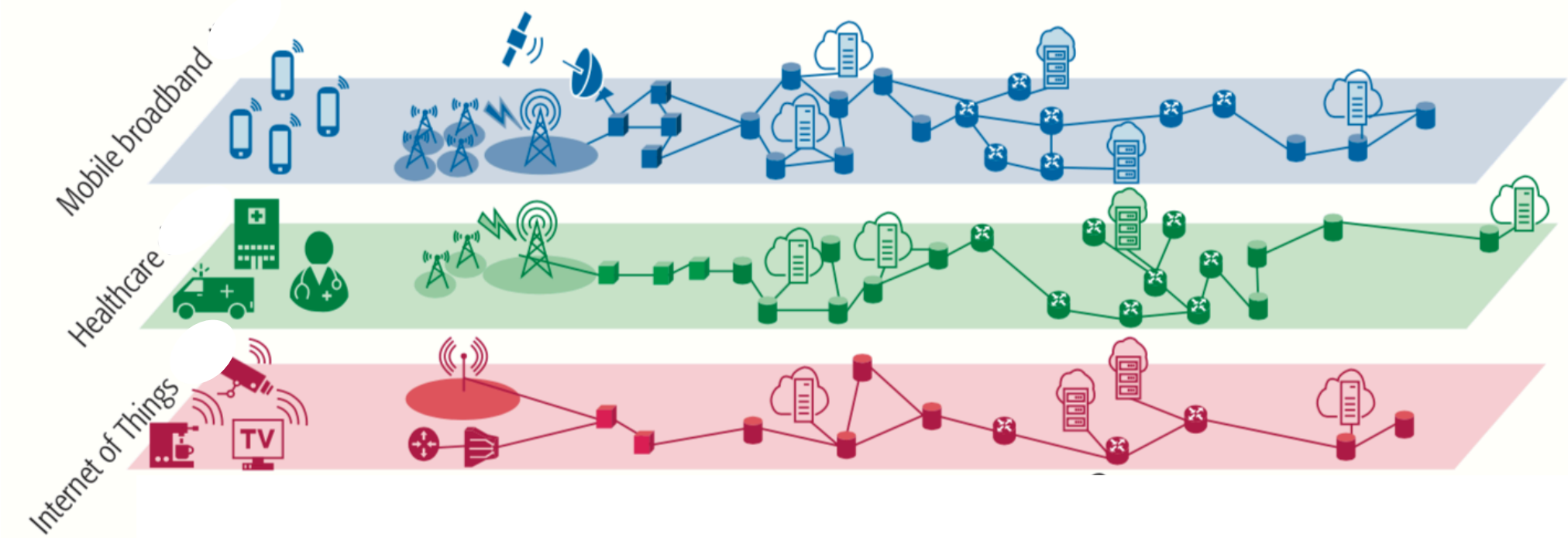
**10 000 x**
**more traffic**

**ultra**
**reliability**

**<1 ms radio**
**latency**

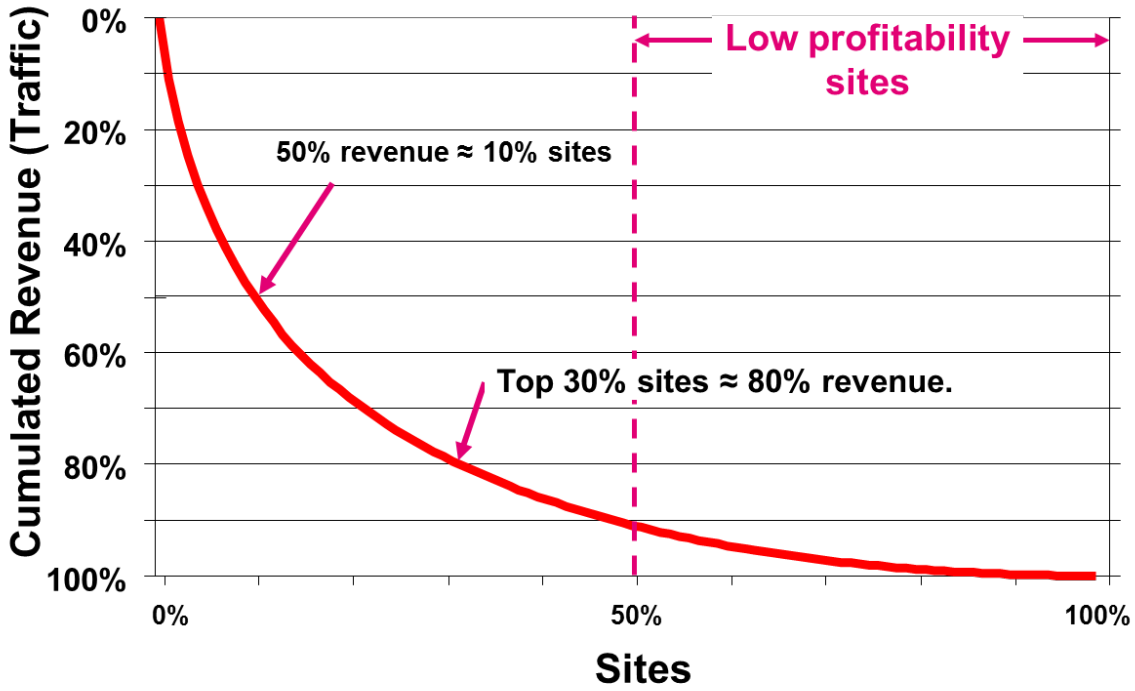
# If business as usual…

- One network per service

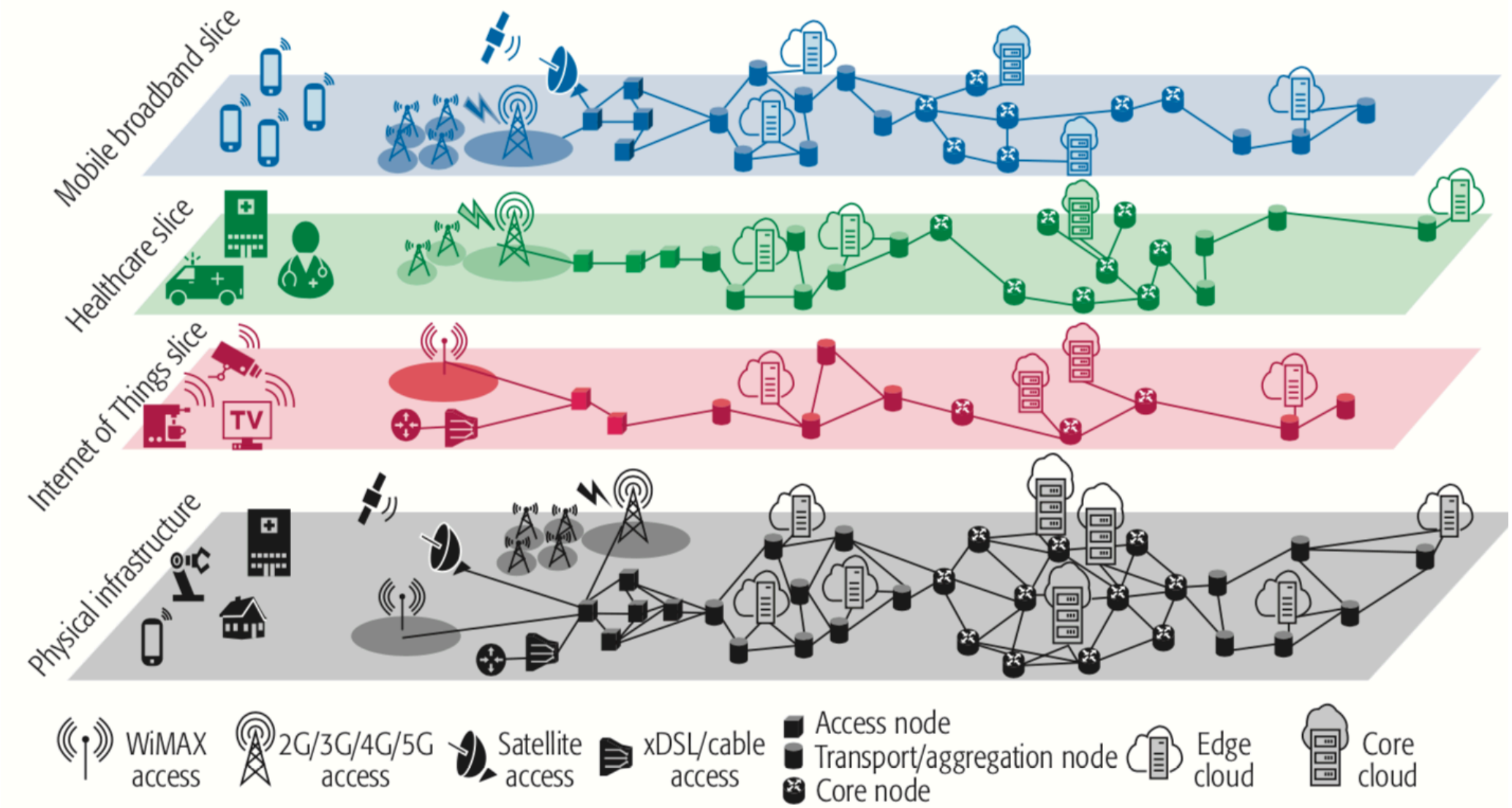


# …inefficiency

- 50% of revenue created by < 10% of sites
- Diverse traffic patterns/mobility cause **resource underutilization**

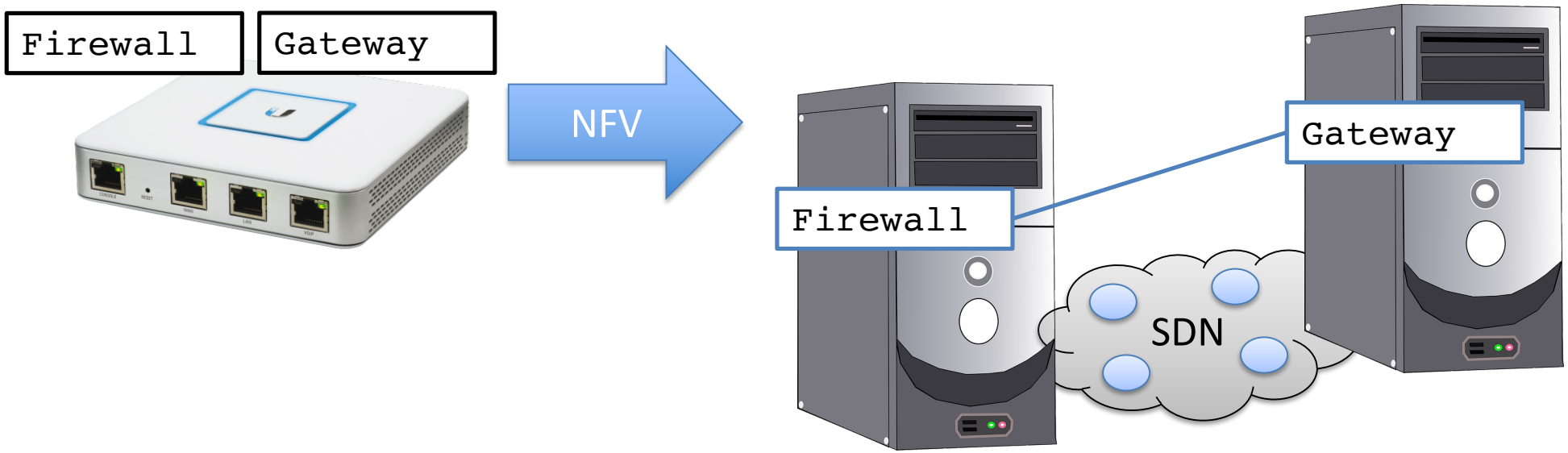


# Network Slicing: multiplexing

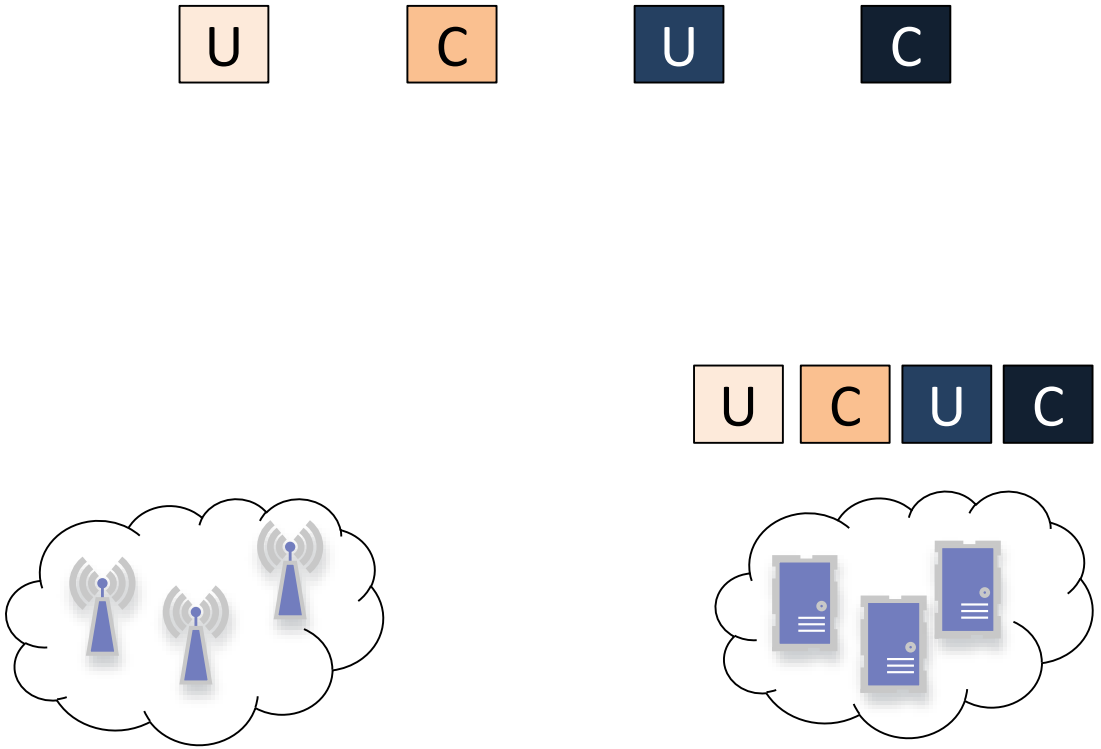



From: J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca and J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," in *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80-87, May 2017.

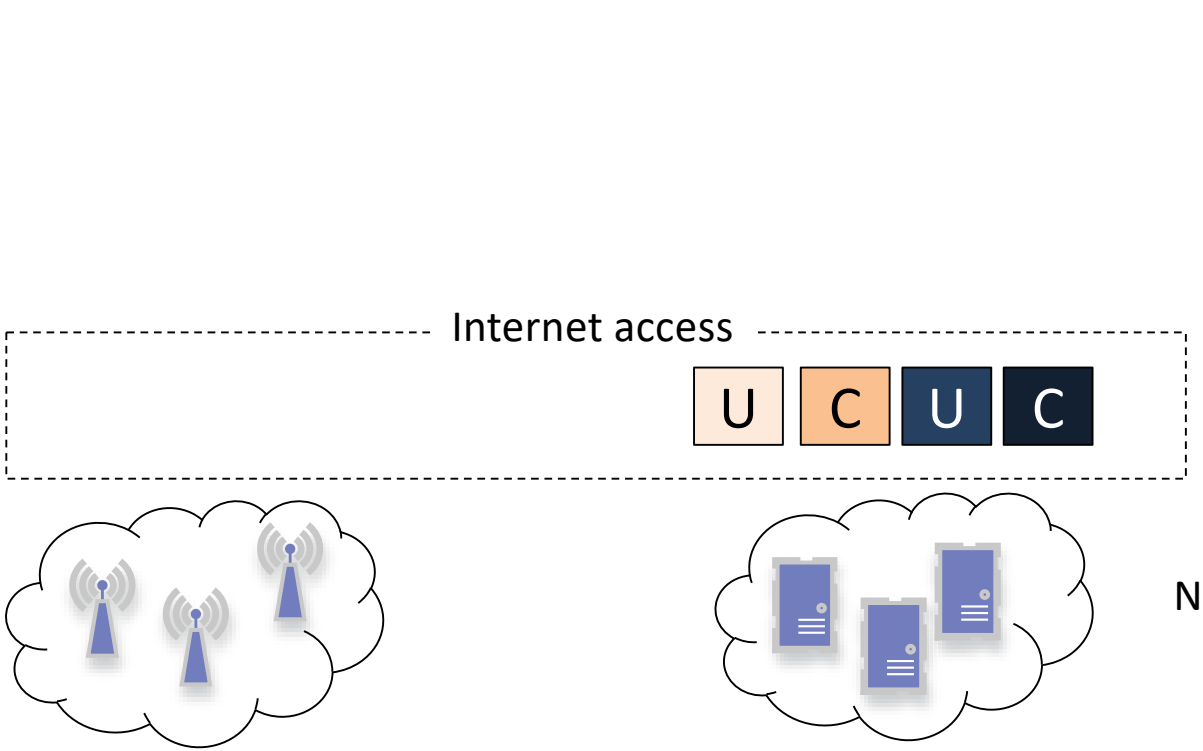
# Why now? VNF and SDN



# Orchestration of VNFs



# Orchestration of VNFs




Virtual Network Functions (VNFs)

U U-plane CN

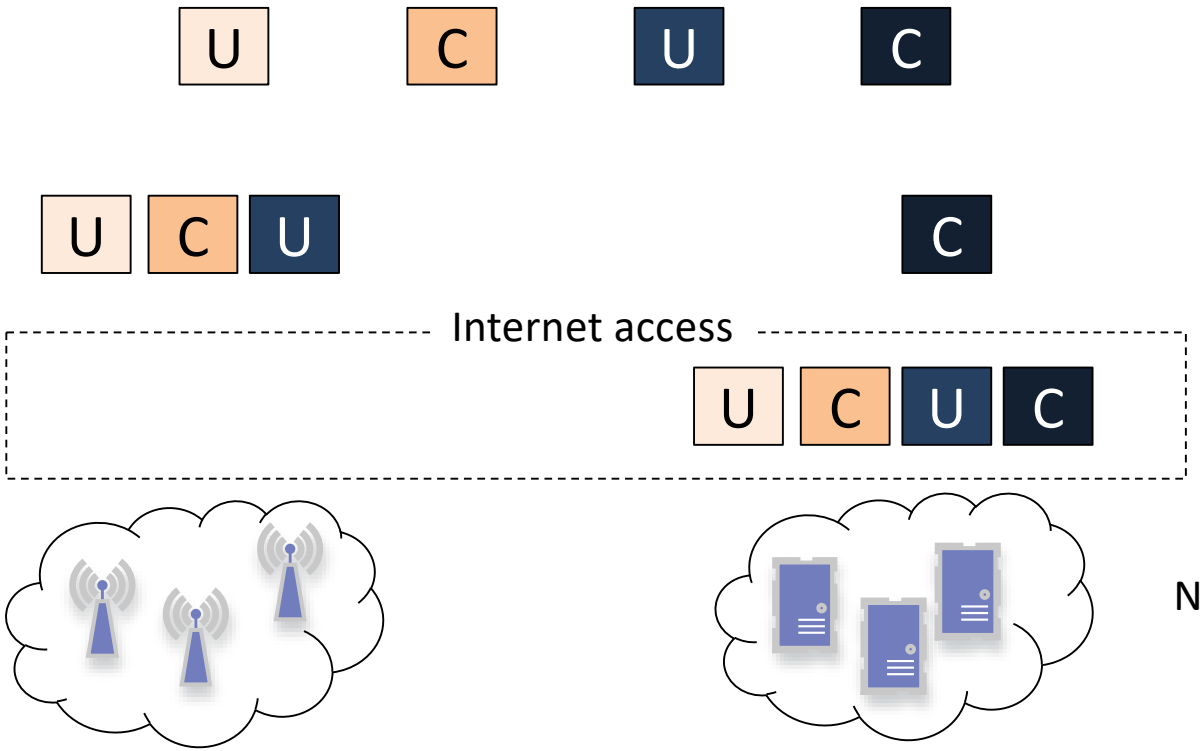
C C-plane CN

U U-plane RAN

C C-plane RAN

Nodes: general purpose processors

# Orchestration of VNFs

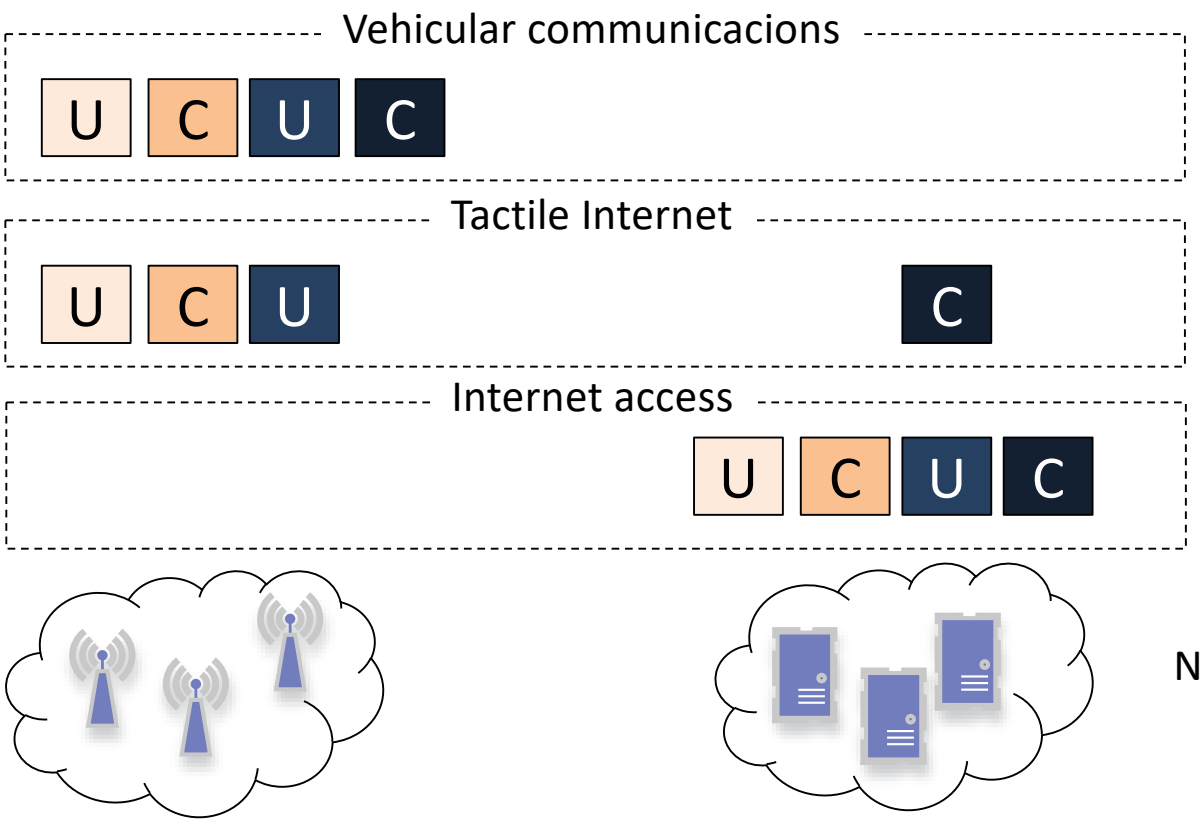



Virtual Network Functions (VNFs)

- U U-plane CN
- C C-plane CN
- U U-plane RAN
- C C-plane RAN

Nodes: general purpose processors

# Orchestration of VNFs



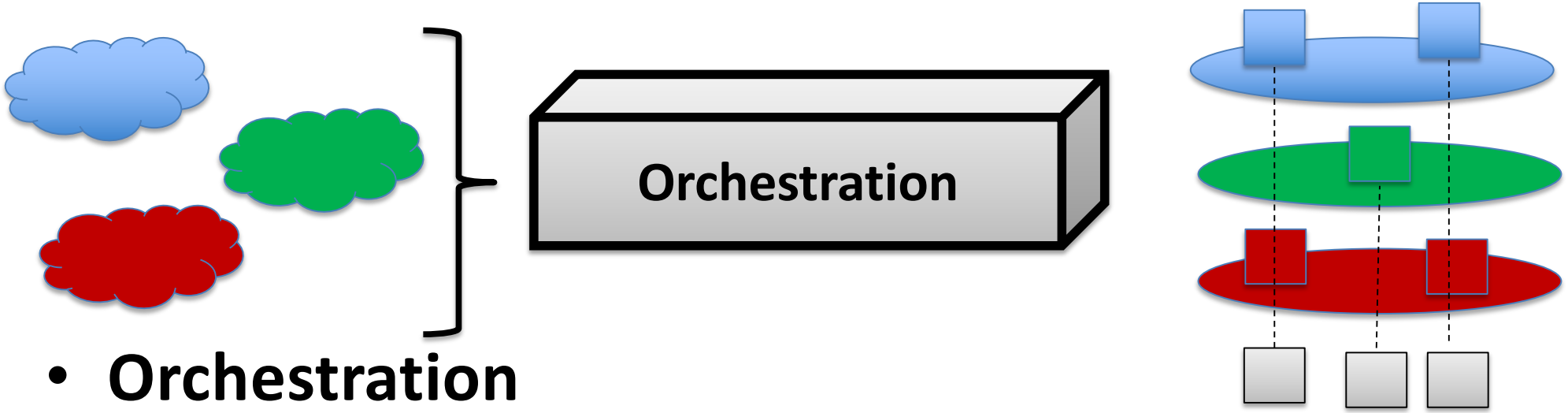

Virtual Network Functions (VNFs)

- U — U-plane CN
- C — C-plane CN
- U — U-plane RAN
- C — C-plane RAN

Nodes: general purpose processors

# Outline

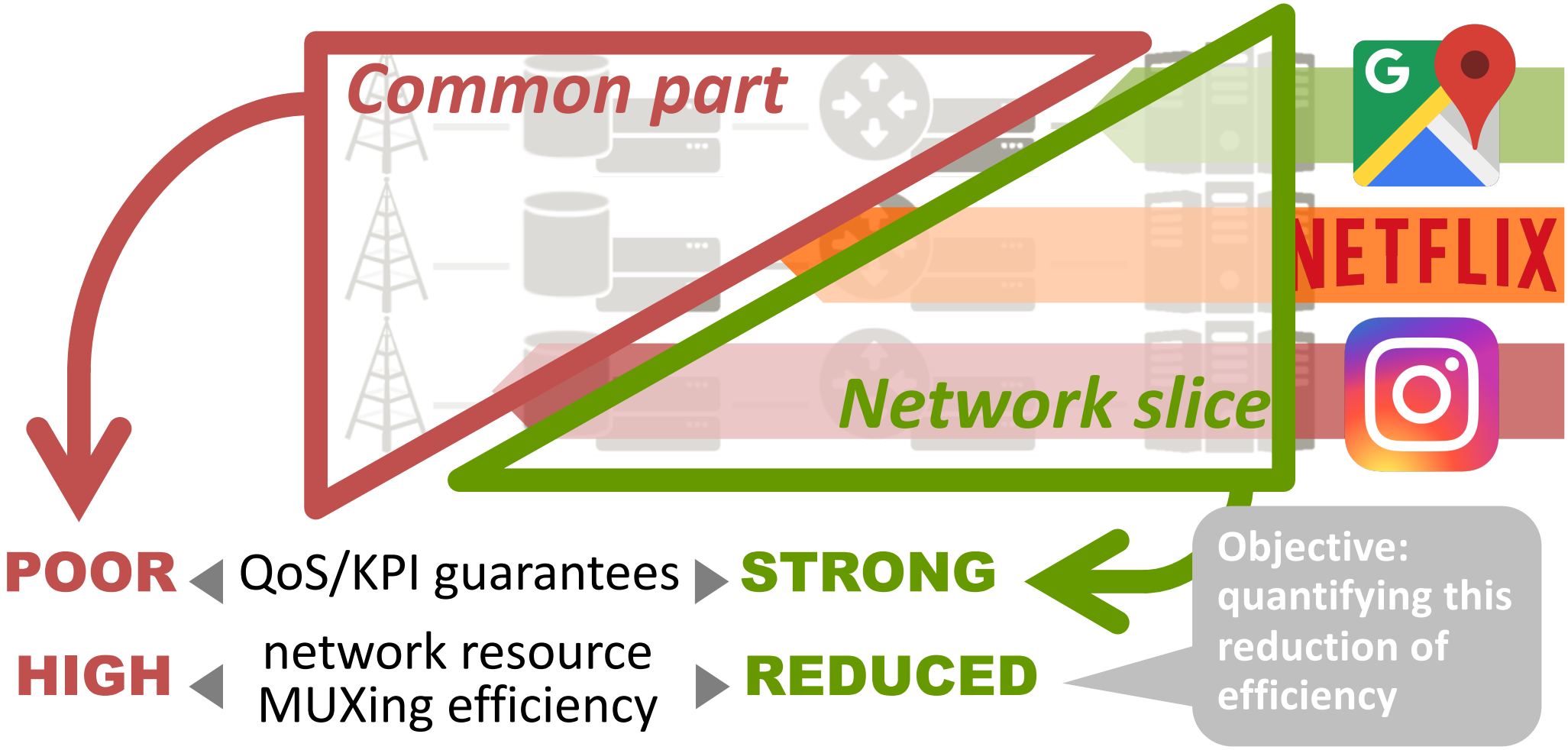
- Network Slicing supports the instantiation of a logical network to support a service




- **Orchestration**
  - **What gains are expected? How to do it?**
- Virtualization
  - What are the challenges? How to address them?

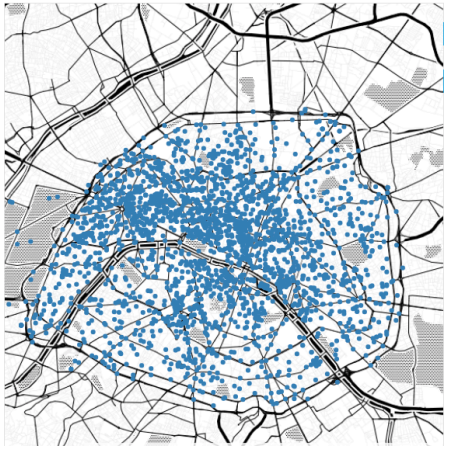
# ORCHESTRATION GAINS

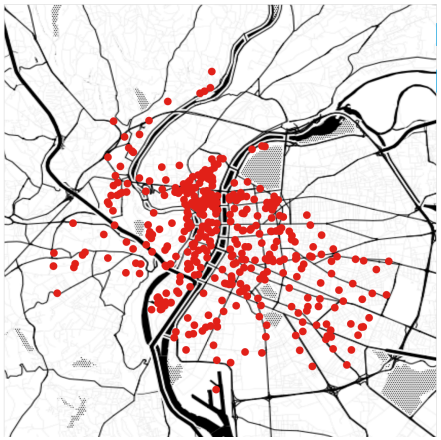
# Slicing trade-offs



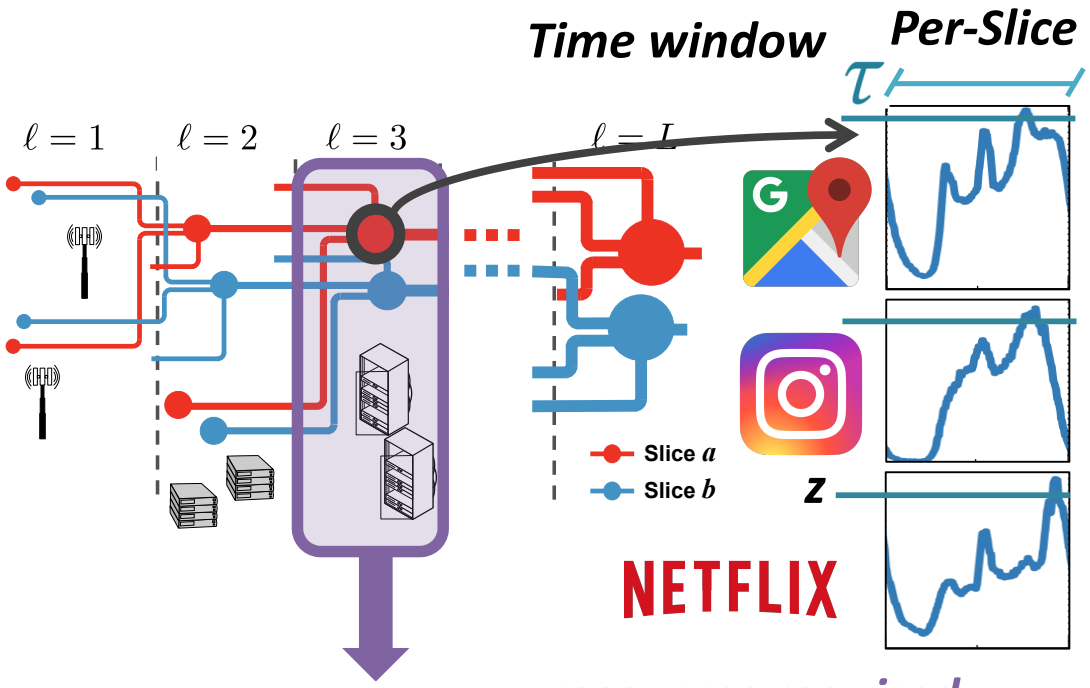
# Problem statement

- Analyse real data from two cities
- Compare the resources required to serve the traffic
- At different levels of the architecture
  - From the antenna to the core
- Using two approaches
  - With slicing
  - No slicing





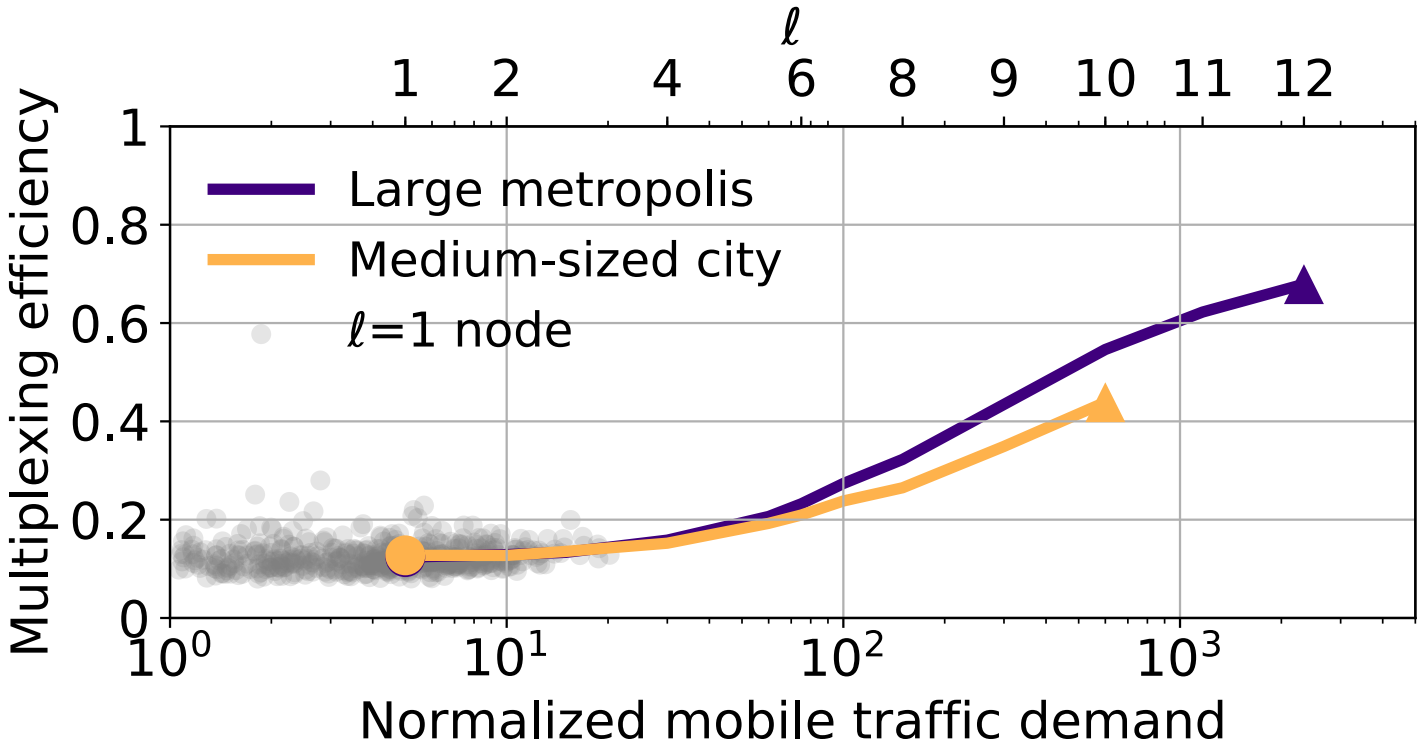
# Slicing vs. Perfect Sharing




$$\sum_{s \in \mathcal{S}} \sum_{c \in \mathcal{C}_\ell} \sum_{n \in \mathcal{T}} \tau \cdot \hat{r}^z_{c,s}(n)$$

*resources required to implement slicing at level l under z*

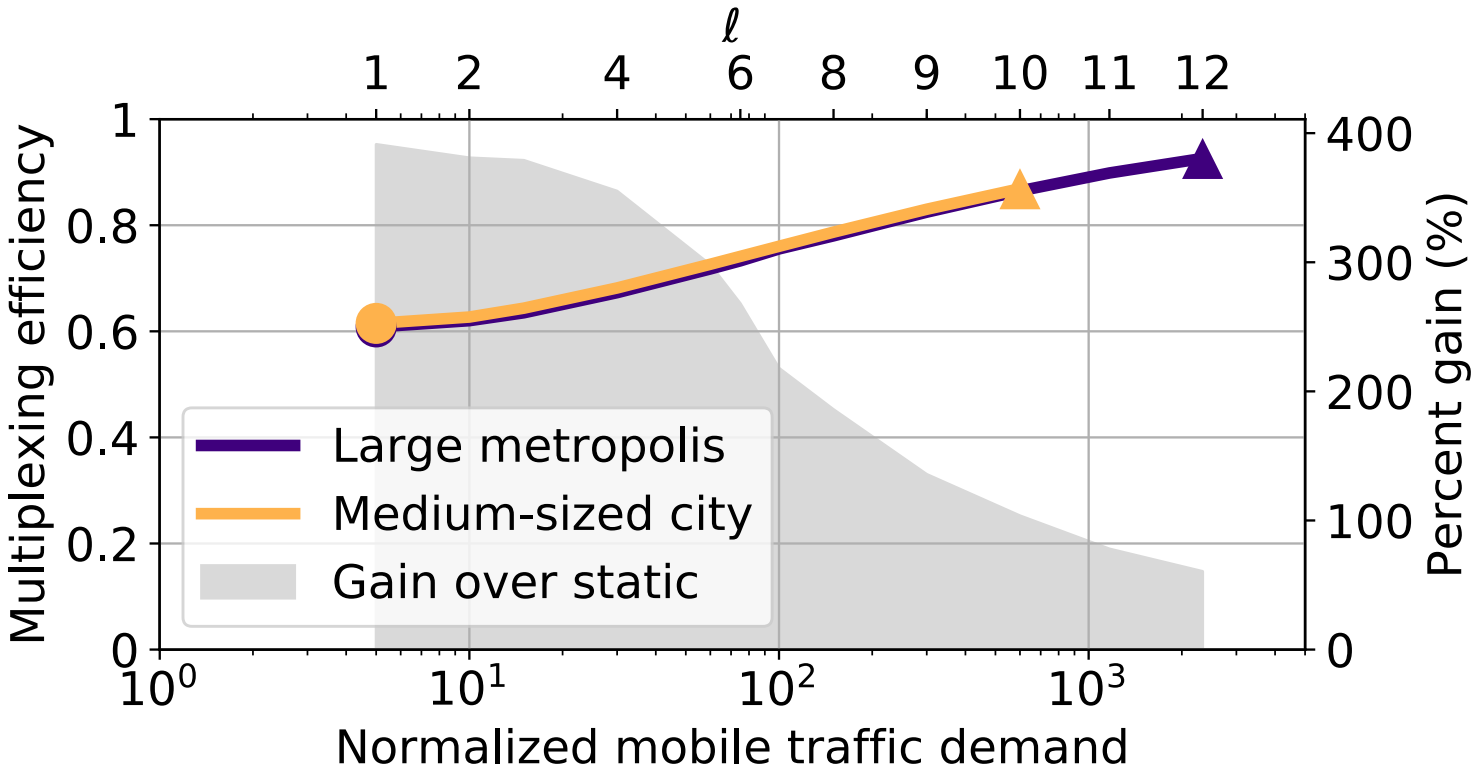
# Results: static allocation




- Resources shall be doubled even at core

# Dynamic allocation

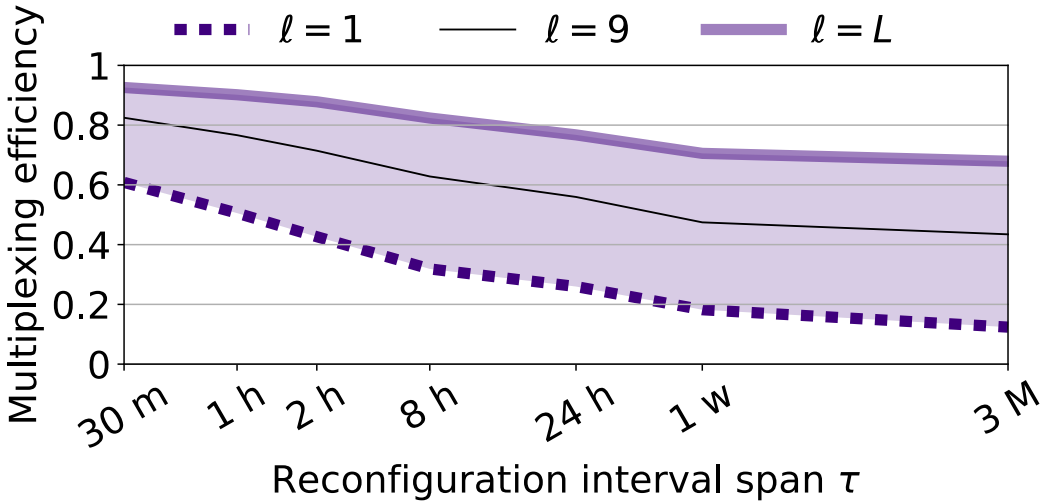
- Re-configuration every 30’




- Up to 4x improvement

# Impact of some parameters

- Timescales

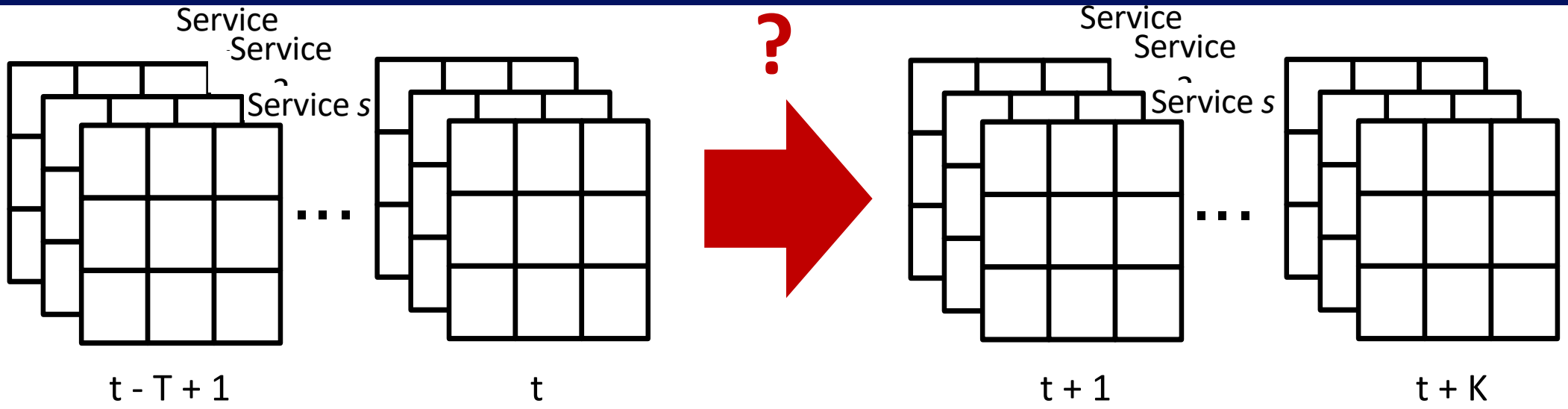


# More info

- C. Marquez et al., “How should I slice my network? A multi-service empirical evaluation of resource sharing efficiency,” ACM MobiCom 2018, New Delhi, India

# HOW TO PERFORM ORCHESTRATION

# The multi-service traffic forecasting problem

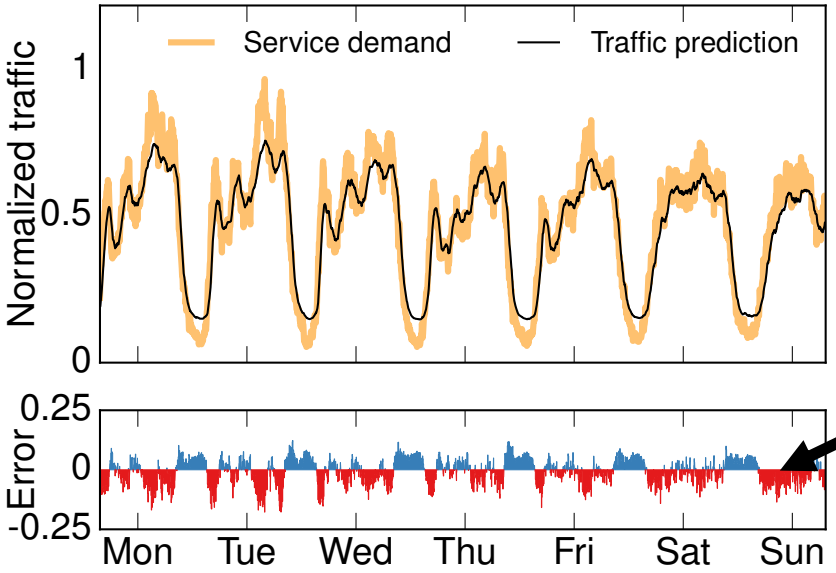



Take historical mobile traffic measurements for multiple services as input and predict future mobile traffic consumption for all services at all base stations simultaneously.

Traditional methods (ES, ARIMA) are inappropriate: Operate on individual time series, their performance degenerates considerably over time, and do not exploit correlations between different services.

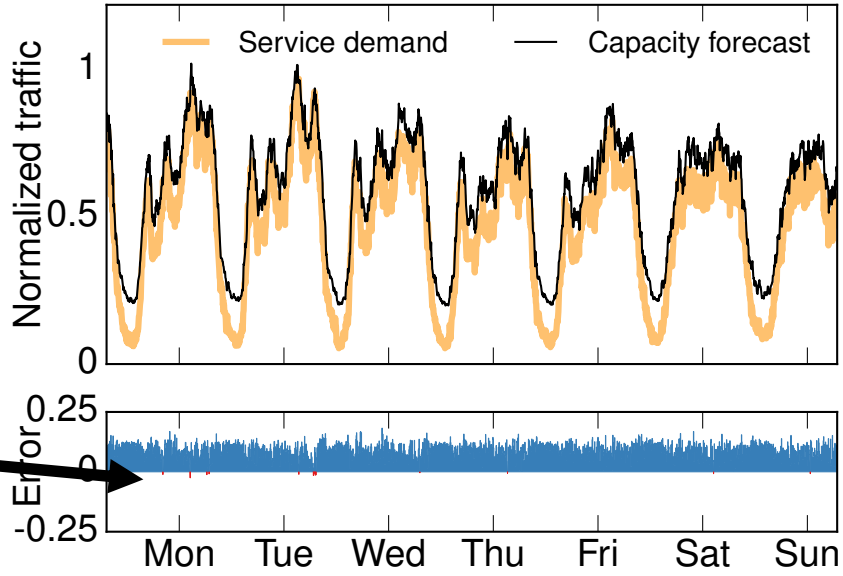
# Plus: capacity vs. traffic

- Traditional approaches deal with **demand** forecasting

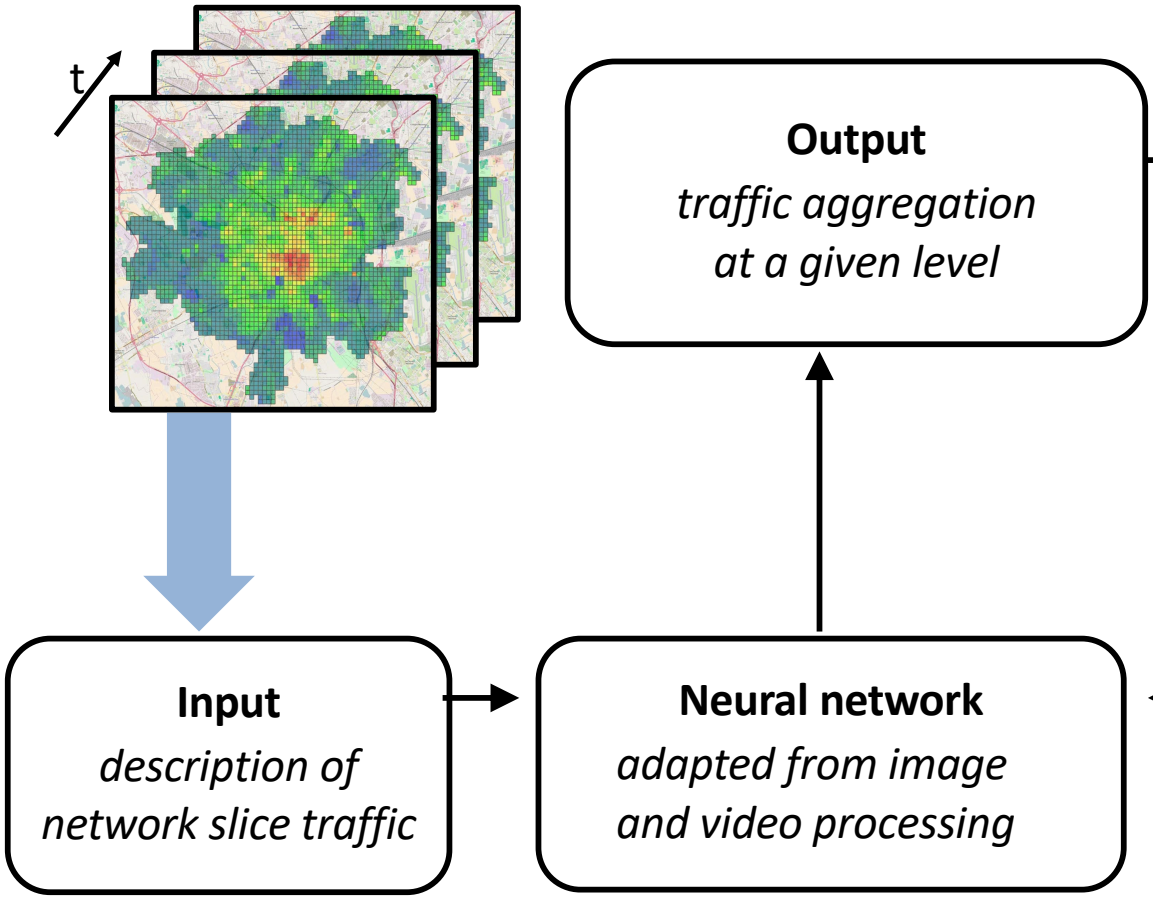



A traffic demand forecasting algorithm aims to minimize the error wrt to the original data, so underestimation is possible

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

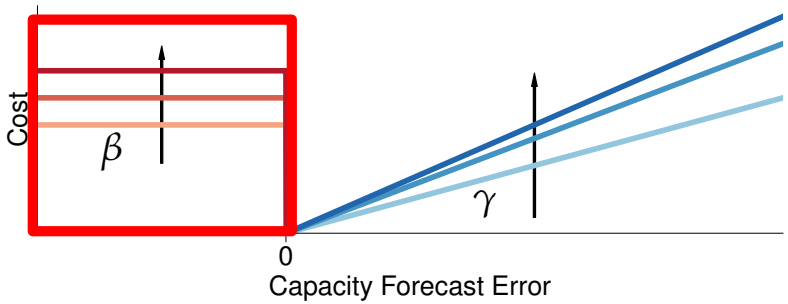


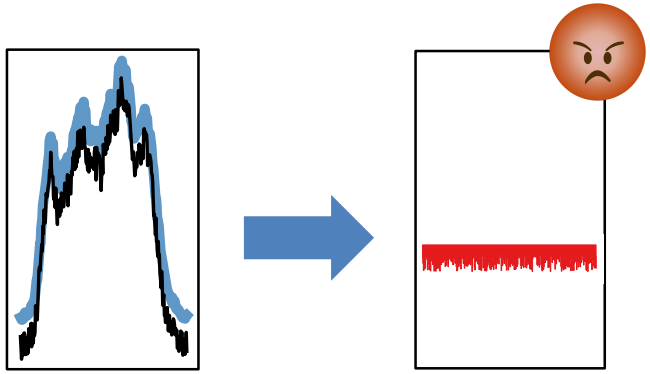
# DeepCog


t
Output
traffic aggregation
at a given level
Input
description of
network slice traffic
Neural network
adapted from image
and video processing

# Cost function

- Determines the penalty incurred when making an error
- Tailored to capacity forecast problem
- It accounts for the costs resulting from
  - SLA violations
  - Overprovisioning

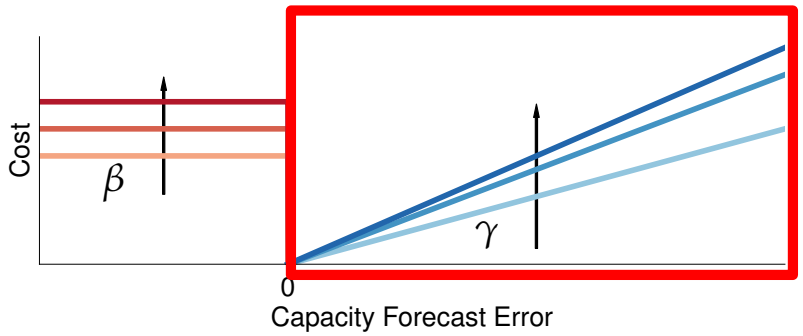


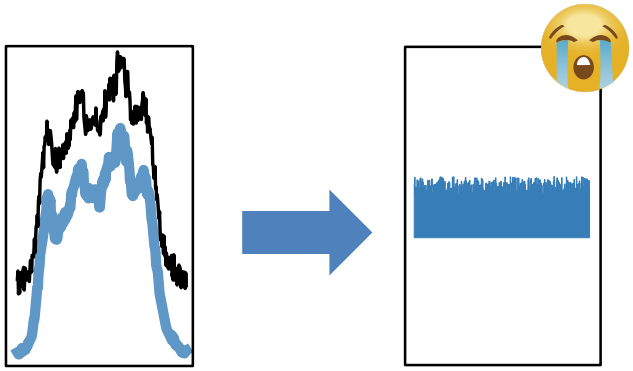



$$\ell(x) = \begin{cases} \beta & \text{if } x \leq 0 \\ \gamma \cdot x & \text{if } x > 0 \end{cases}$$

# Cost function

- Determines the penalty incurred when making an error
- Tailored to capacity forecast problem
- It accounts for the costs resulting from
  - SLA violations
  - Overprovisioning

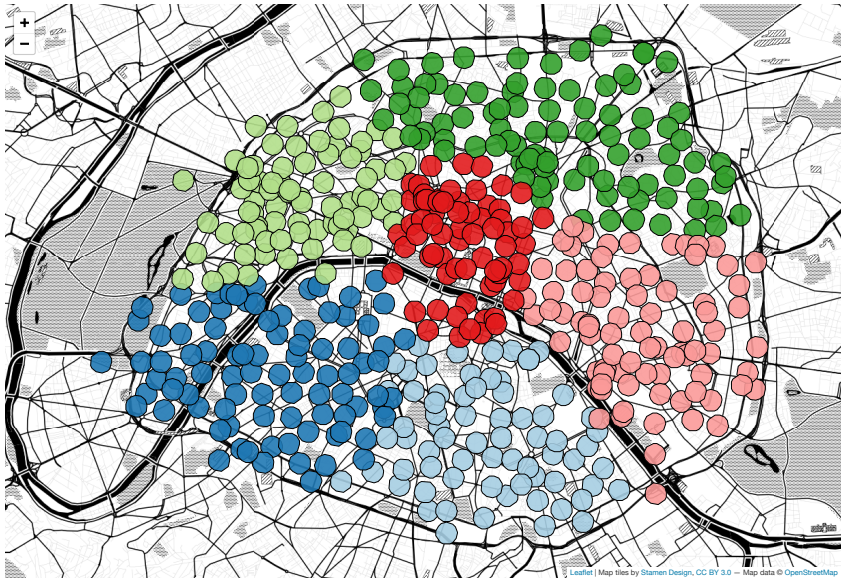


$$\ell(x) = \begin{cases} \beta & \text{if } x \leq 0 \\ \gamma \cdot x & \text{if } x > 0 \end{cases}$$

# Performance Evaluation

- Real-world demand generated by several millions of users
  - Mobile network deployed in a large city
- Different services analysed. Core, MEC, & C-RAN

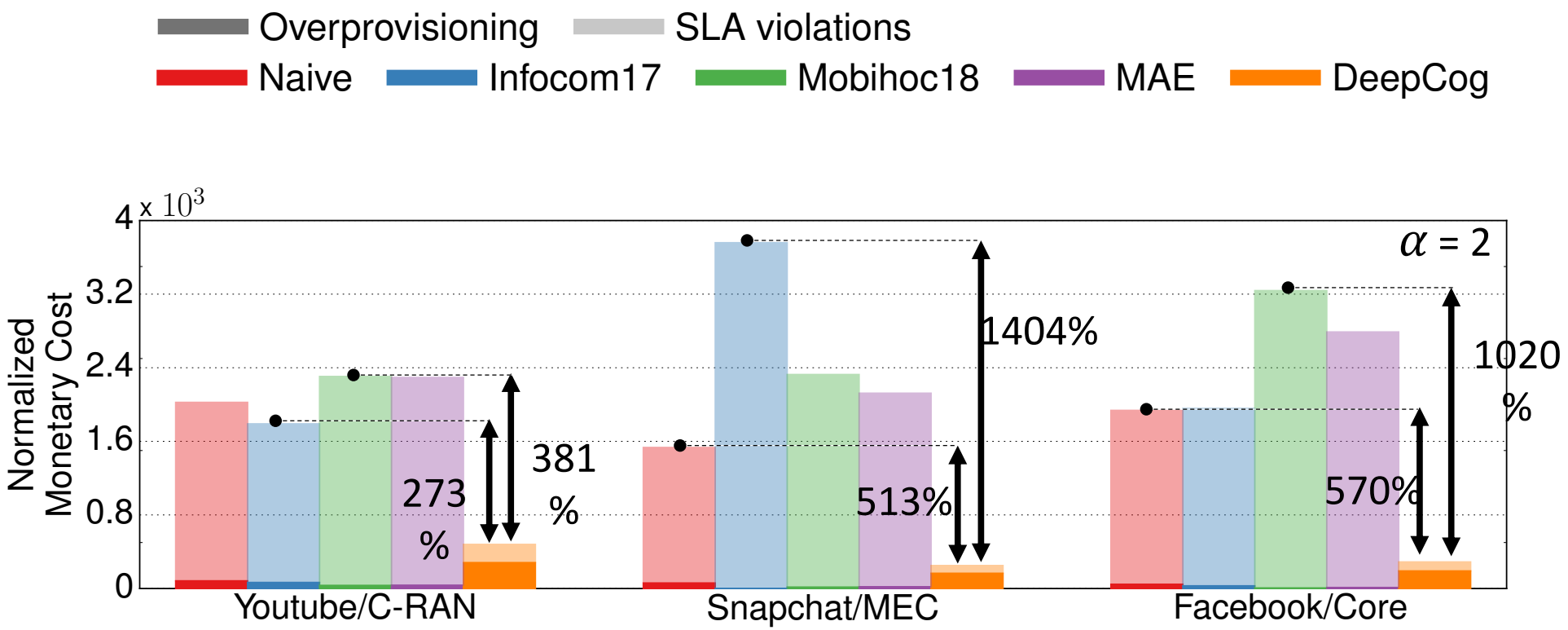


# Results - Benchmarks

- Compared against 4 benchmarks
  - Naïve – replicates the previous week demand
  - Infocom17[1] – first DL approach
  - MobiHoc18[2] – recent DL solution
  - MAE – DeepCog architecture employing Mean Absolute Error as loss function

[1] – J. Wang et al., *"Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach,"* in *Proc. of IEEE Infocom, May 2017.*
[2] – C. Zhang et al., *"Long-Term mobile traffic forecasting using Deep Spatio-Temporal Neural Networks,"* in *Proc. Of ACM MobiHoc,* Jun. 2018.

# Results


Overprovisioning
SLA violations
Naive
Infocom17
Mobihoc18
MAE
DeepCog
x 10^3
4
3.2
2.4
1.6
0.8
0
Normalized Monetary Cost
α = 2
273 %
381 %
513%
1404%
570%
1020 %
Youtube/C-RAN
Snapchat/MEC
Facebook/Core

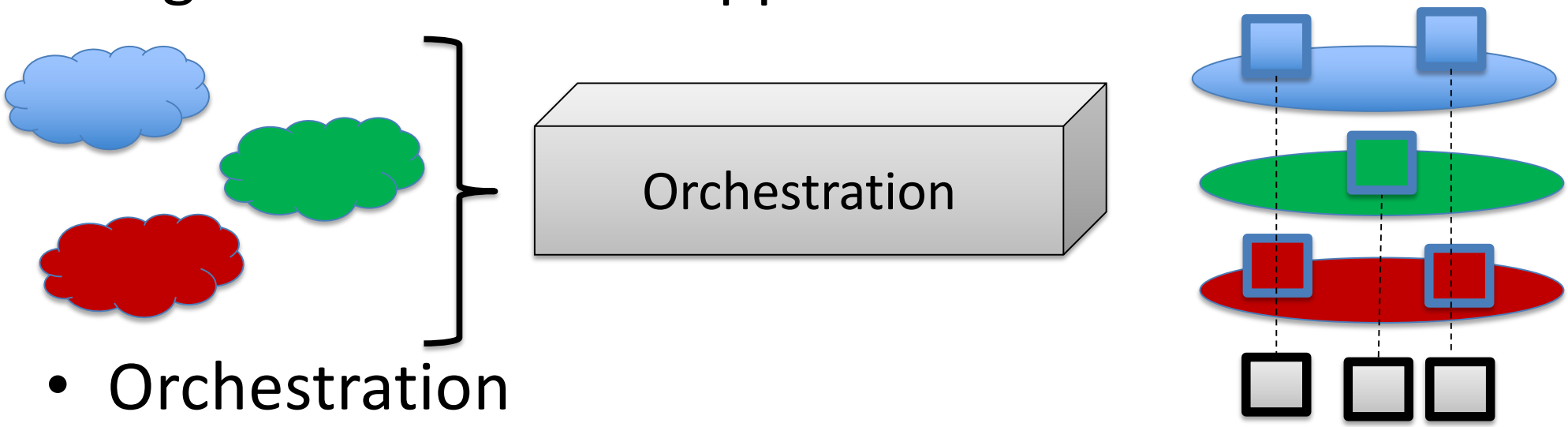
# More info & Results

- D. Bega et al., “DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning,” IEEE INFOCOM, April 2019

- C. Zhang et al. “Multi-Service Mobile Traffic Forecasting via Convolutional Long Short-Term Memories,” IEEE International Symposium on Measurements & Networking (M&N), July 2019

# Outline

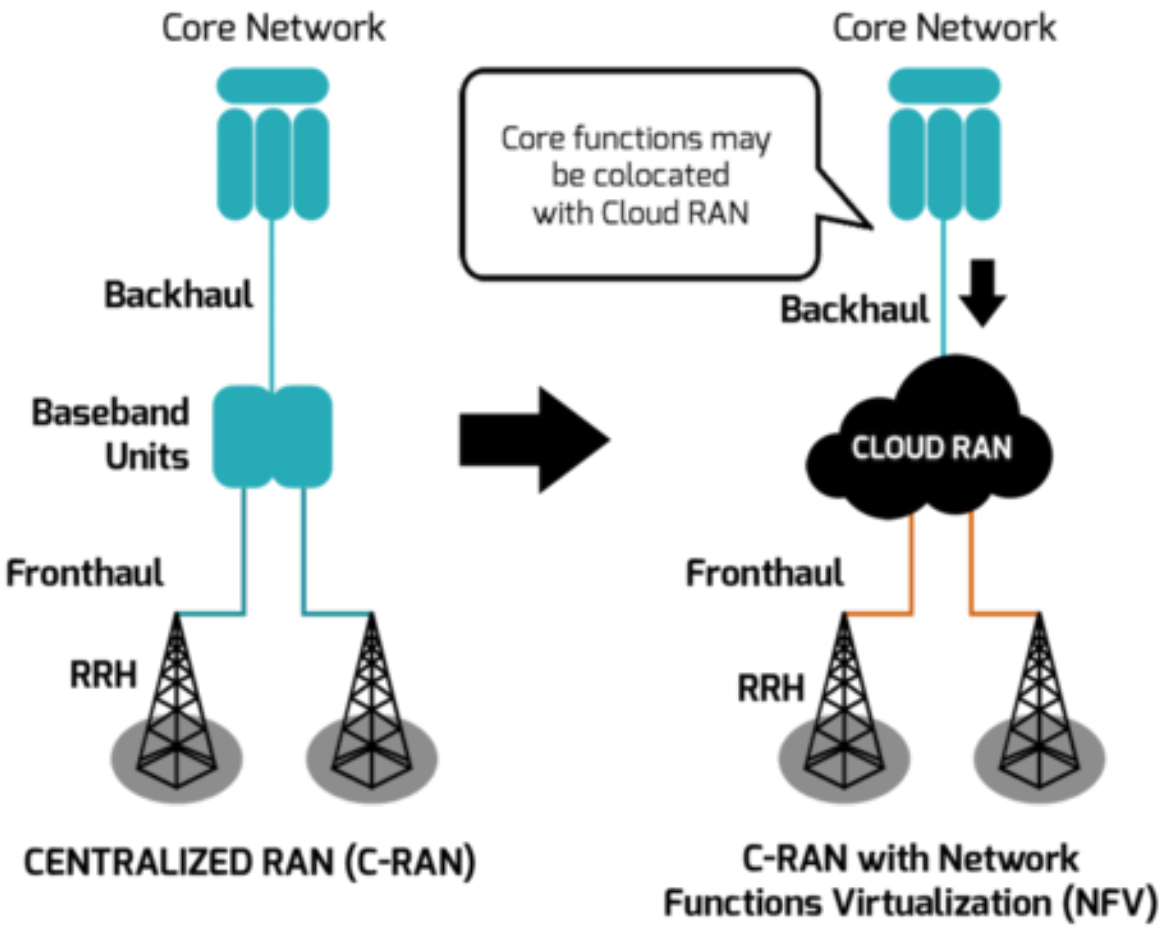
- Network Slicing supports the instantiation of a logical network to support a service




- Orchestration
  - What gains are expected? How to do it?
- **Virtualization**
  - **What are the challenges? How to address them?**

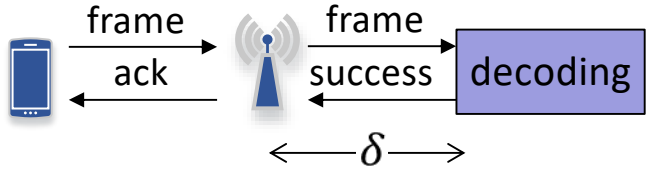
# ADDRESSING THE VIRTUALIZATION OF NETWORK FUNCTIONS

# Case Study: from C-RAN to vRAN



# Network functions in the cloud

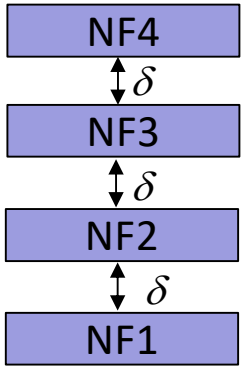
- Traditional approach: dimension for maximum capacity (i.e., all users using highest MCS)
  - Resource wastage if this rarely happens
- But some functions do not tolerate resource shortage (e.g., decoding): they are *inelastic*




- "I don't have time to decode this MCS"

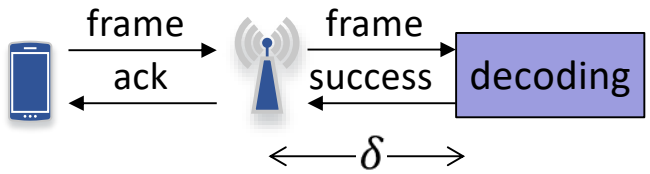
# 1) Removing tight constrains

Current stack:

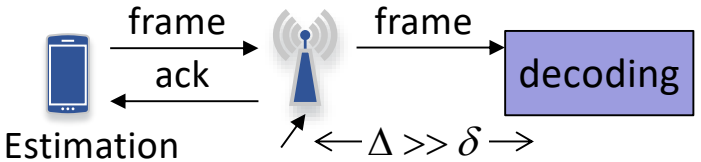



Network functions

*current HARQ*:

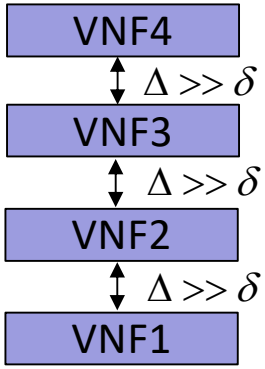



*Opportunistic HARQ*:

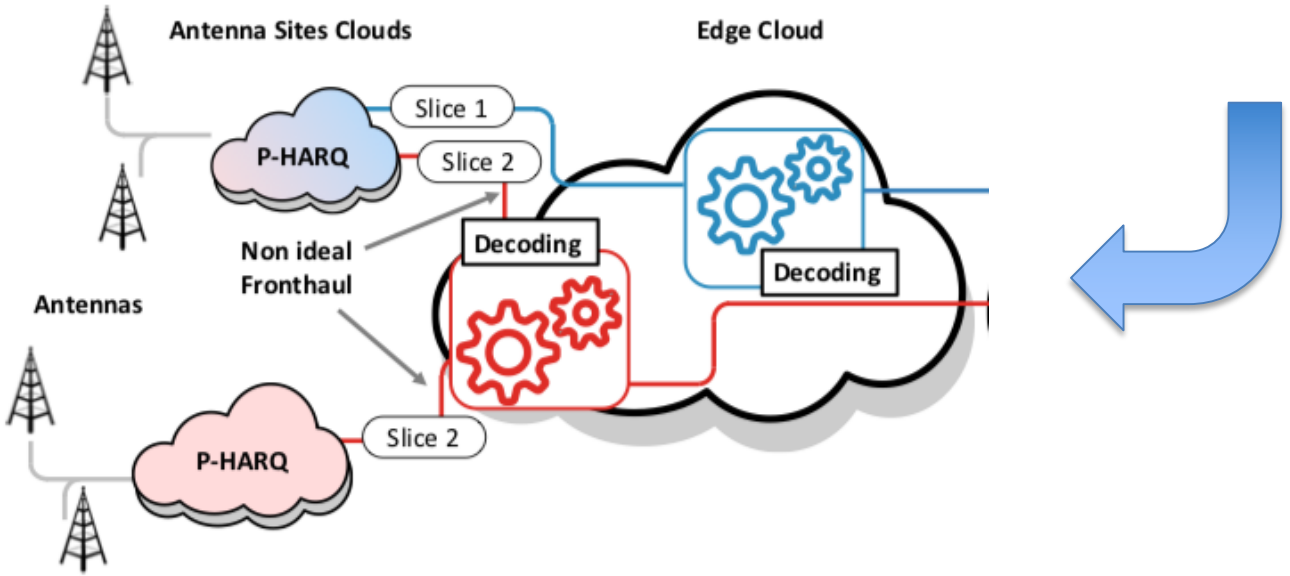



Estimation success prob.

Redesign:



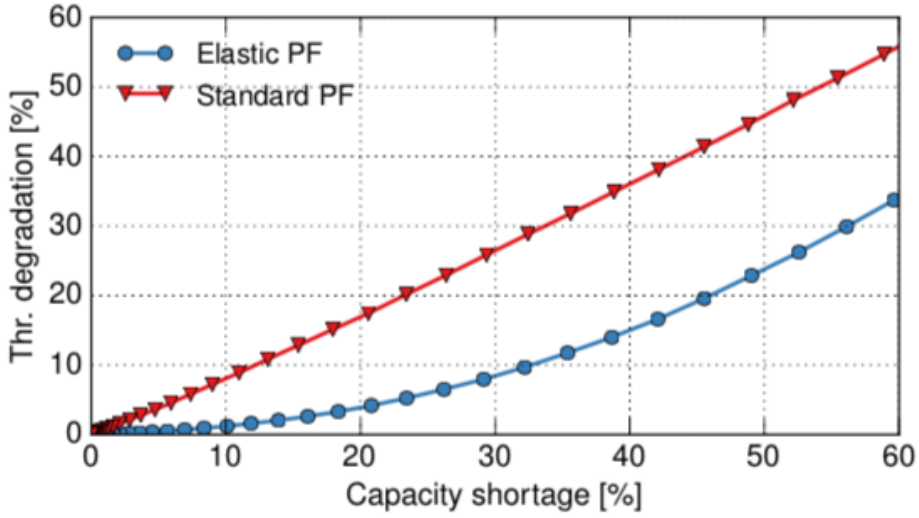

Virtual Network functions




P. Rost and A. Prasad, "Opportunistic Hybrid ARQ – enabler of centralized-ran over non-ideal backhaul," IEEE Wir. Comm. Mag

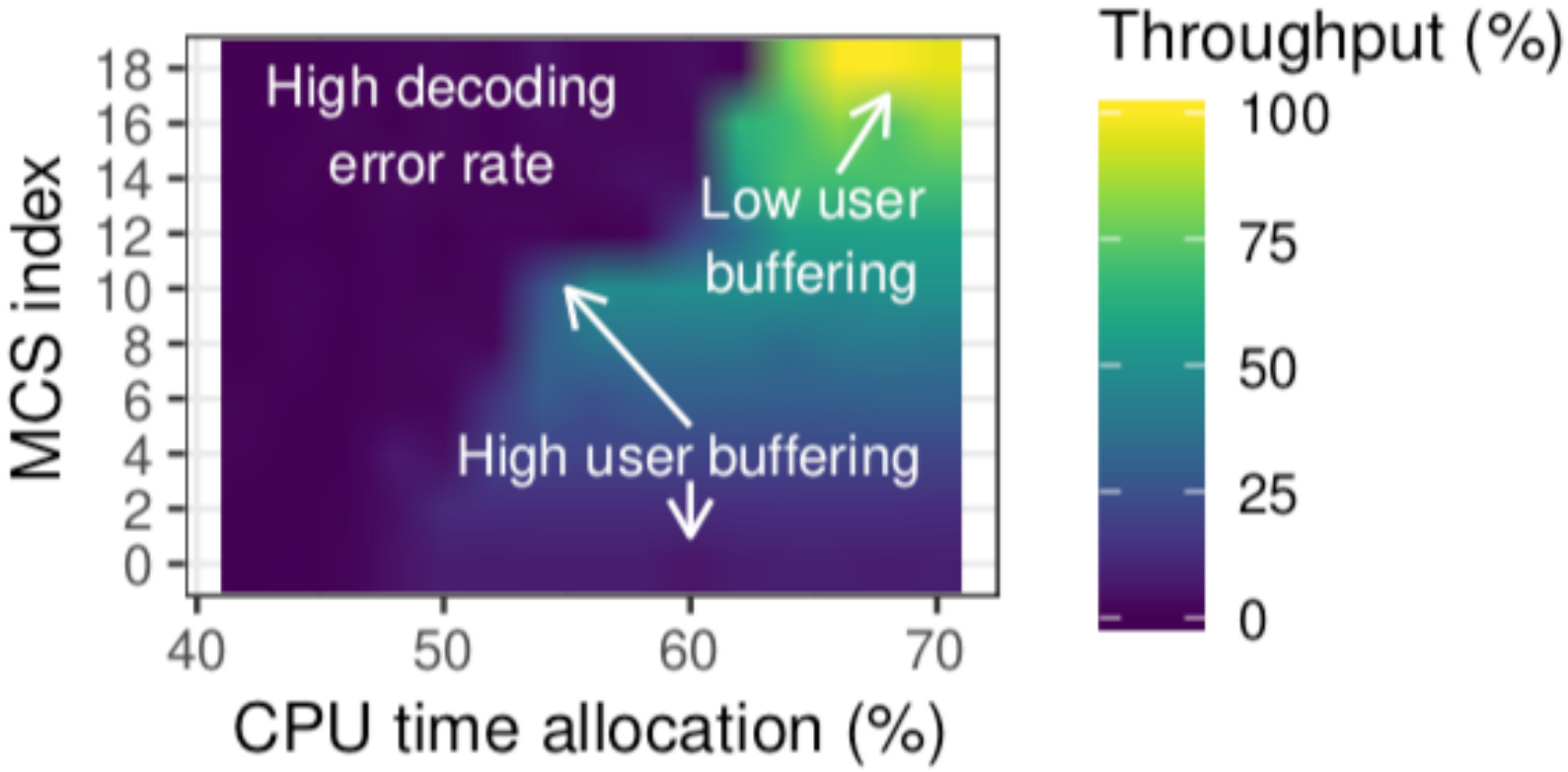
# 2) Rethinking network functions

- Different approach: assume there could be capacity shortage
- “There is no CPU for this MCS, but we could schedule this other (lower) MCS”
- Resource-elastic schedulers

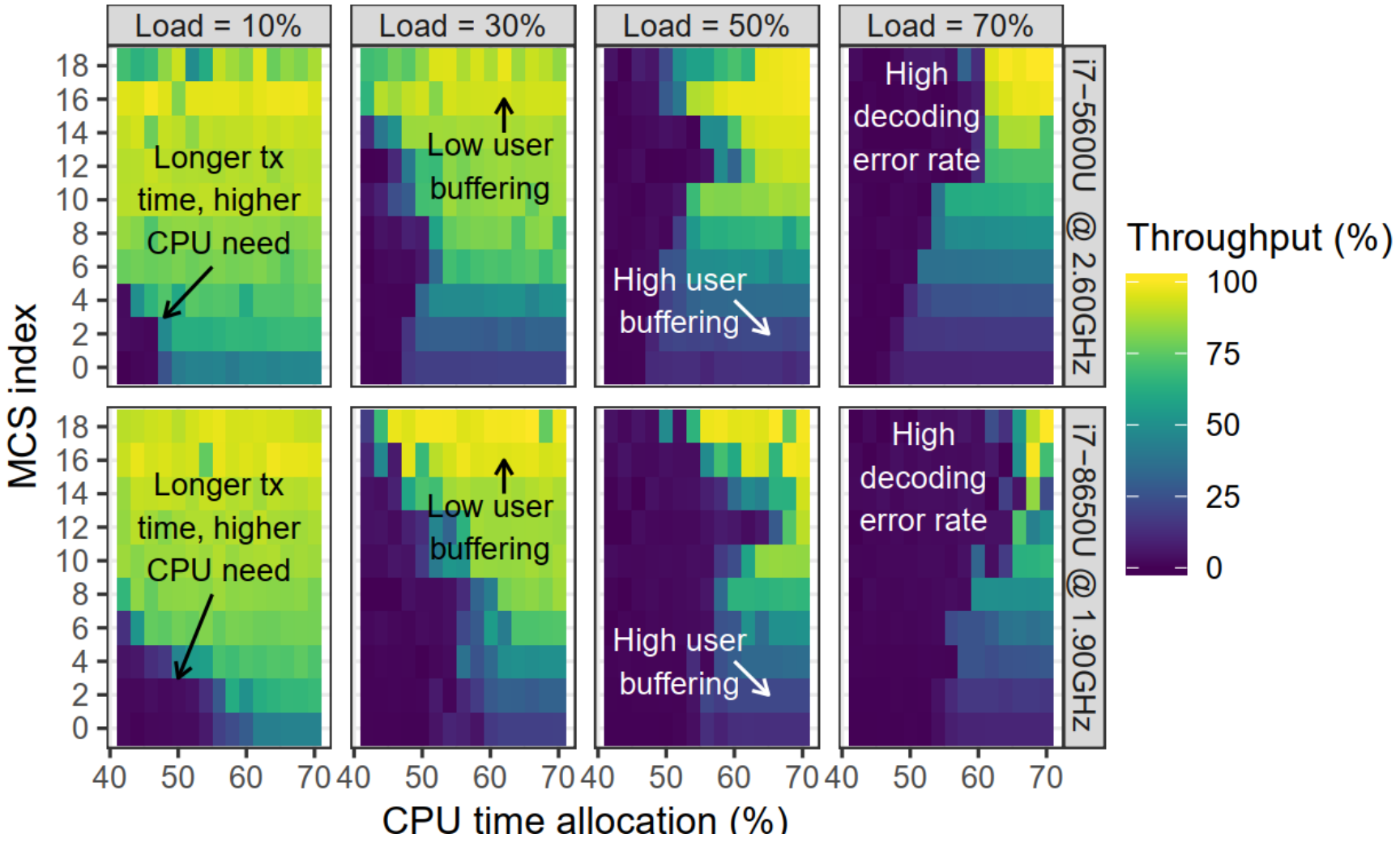


# Relation of key parameters

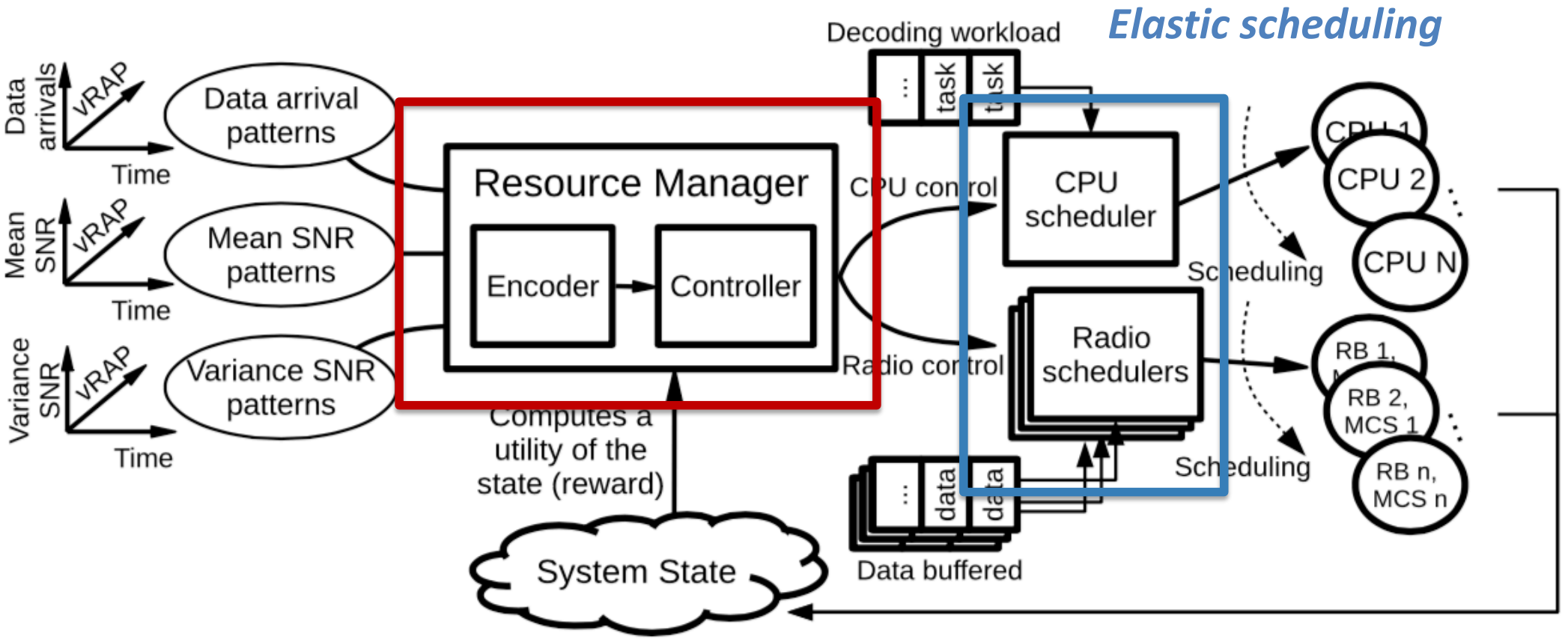
- Throughput as a function of MCS index and CPU time allocation



# Relation is complex and non-linear

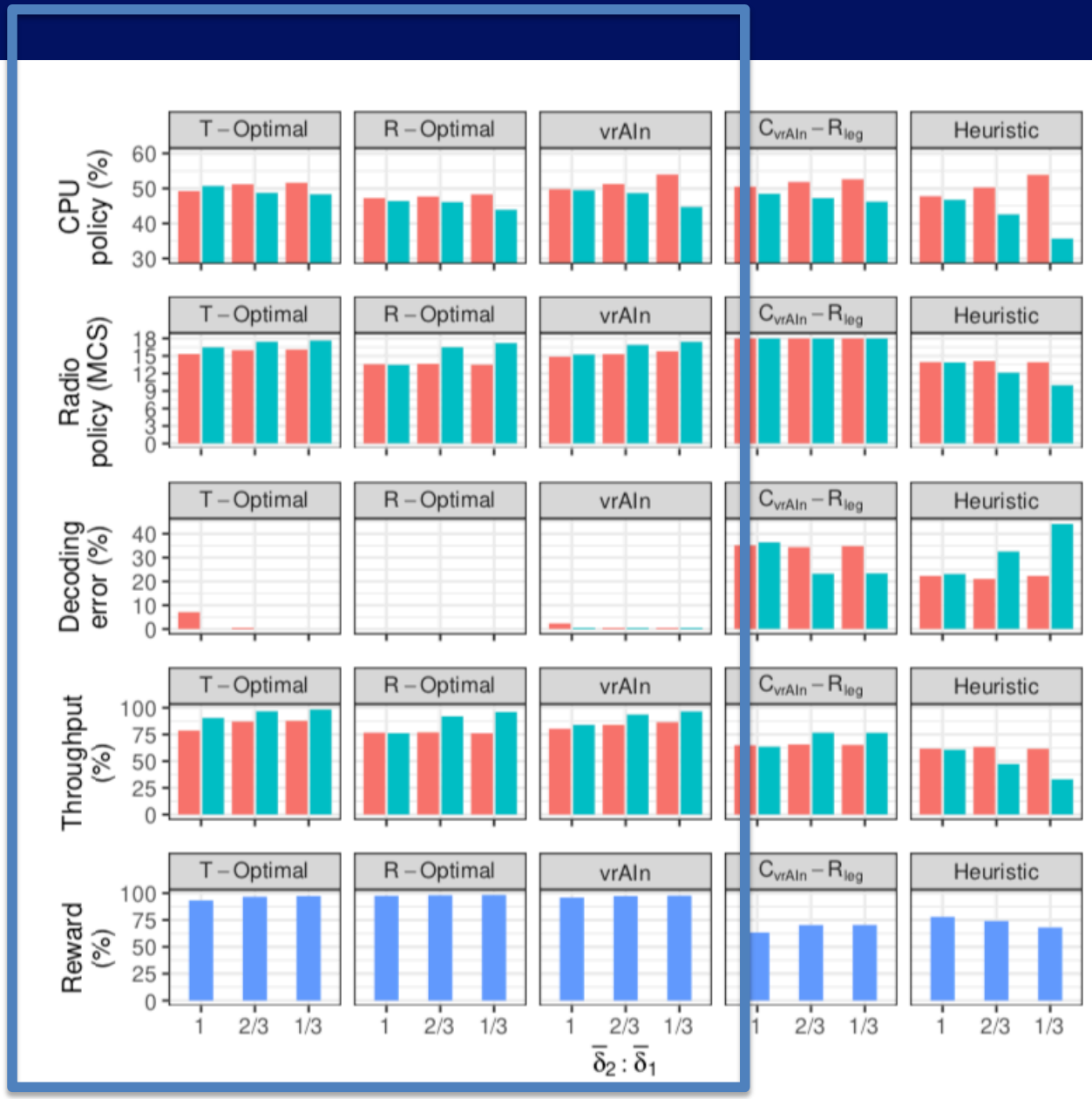


# Machine Learning



# Results

- Throughput-Optimal oracle
- Reward-optimal oracle,
- **vrAIn**
- Legacy radio
- Heuristic



# More info

- P. Serrano et al., “The path towards a cloud-aware mobile network protocol stack,” Transactions on E.T.T., May 2018
- F. Gringoli et al., “Performance Assessment of Open Software Platforms for 5G Prototyping,” IEEE Wireless Communications Magazine, Special Issue on 5G Testing and Field Trials
- J. A. Ayala-Romero et al., “vrAIn: A Deep Learning Approach Tailoring Computing and Radio Resources in Virtualized RANs,” ACM Mobicom, 2019
- M. Gramaglia et al. “The case for sustainable serverless networking,” working paper

# CONCLUSIONS

# Take away messages

- Orchestration strategies
  - Trade-off between isolation and efficiency
  - Static algorithms -> 2x resources even at the CN
  - Dynamic allocations: short timescales
  - Loose QoS requirements to boost efficiency

- Virtualization of Network Functions
  - Need to re-think their operation
  - Introduce resource awareness (PHY, CPU, load)

# The future of Network Slicing

- Several efforts aligned
  - Evolution of SDN, NFV
  - Adopted by 5G, IETF, etc.
  - Many ongoing initiatives (incl. Open source)
  - Trials (e.g., 5G-EVE, 5TONIC)
- Several challenges
  - Orchestration of these efforts
  - Security, Privacy, Resiliency
  - Use of Machine Learning / AI techniques

# Thanks! Questions?



pablo@it.uc3m.es


@pablo_uc3m

http://www.it.uc3m.es/pablo/



Many thanks to (among others): Albert Banchs, Dario Bega, Carlos J. Bernardos, Andrés García-Saavedra, Marco Gramaglia, Yan Grunenberger, Diego López, Paul Patras, Vincenzo Sciancalepore

# ADDITIONAL SLIDES

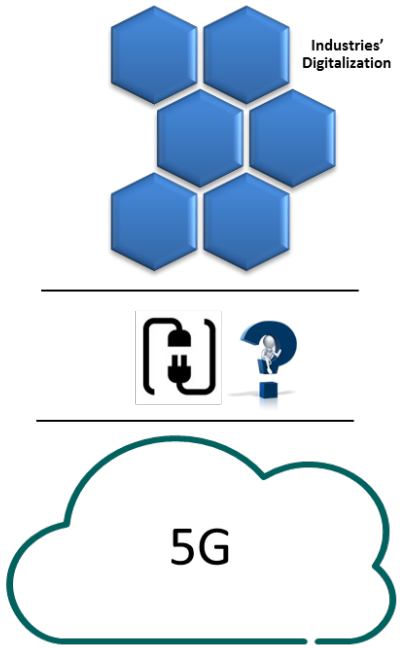
# Network Slicing: Main novelties

- **Customization**: “Underlays / overlays supporting all services equally (‘best effort’ support) are not fully representing a Network Slice”
- **Ecosystem**: Different players, domains, modules and interfaces
- **Life cycle**: [Deploy new services] ”From 90 hours down to 90 minutes”

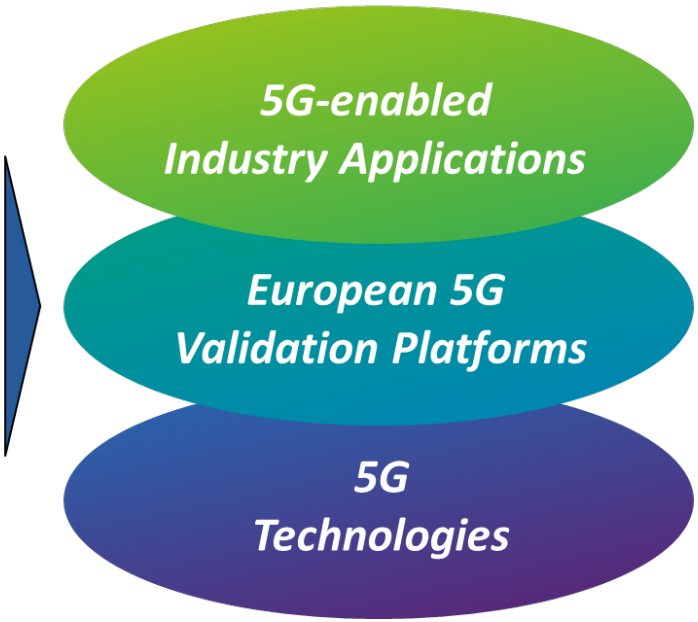
# 5G-EVE

# The vision

From: Two Worlds




To: One Innovation Ecosystem

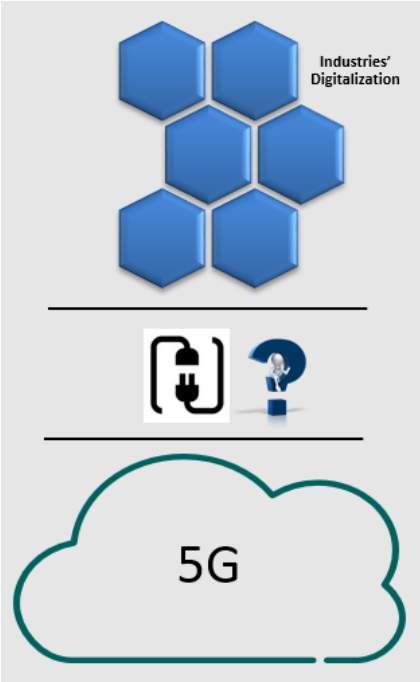


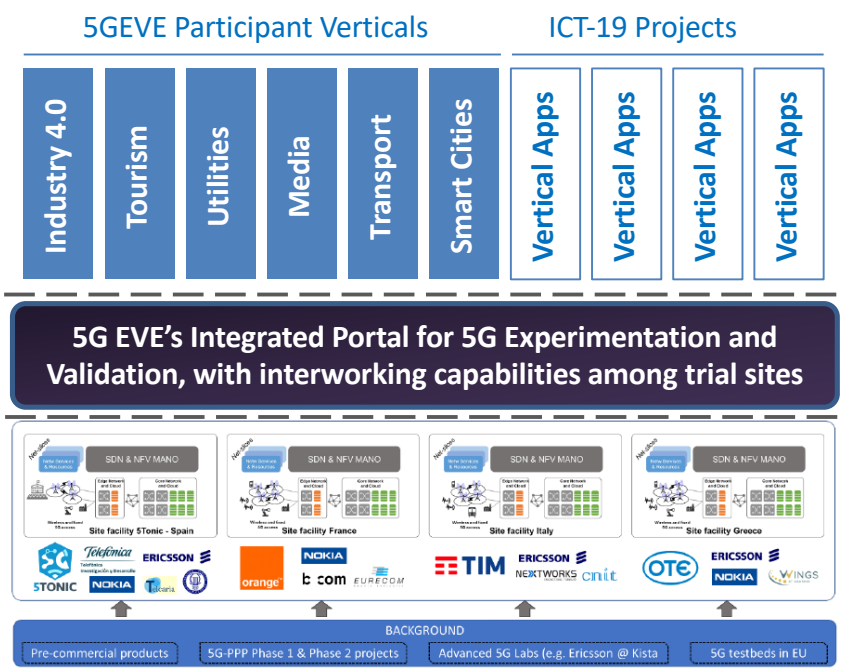

- Agile
- Diverse
- Specialized
- Transformative

- Open
- Ease-to-use
- Trustworthy
- Automated

- Performing
- Scalable
- Standard
- Secure
- Evolving

# How 5G EVE architects that vision



# 5G EVE Platform - Validation Test as a Service