Slicing in 5G Transport Networks

Paolo Monti

Electrical Engineering Department
Chalmers University of Technology
Göteborg, Sweden

Acknowledgement

Kista 5G Transport Lab

- **Goal**: To enable the 5G transport network to deliver near-ubiquitous connectivity and be a platform for service innovation
- **Funded by**: Ericsson and VINNOVA
- **Partners**: KTH, Ericsson, RISE Acreo

Optical Networks group @ Chalmers

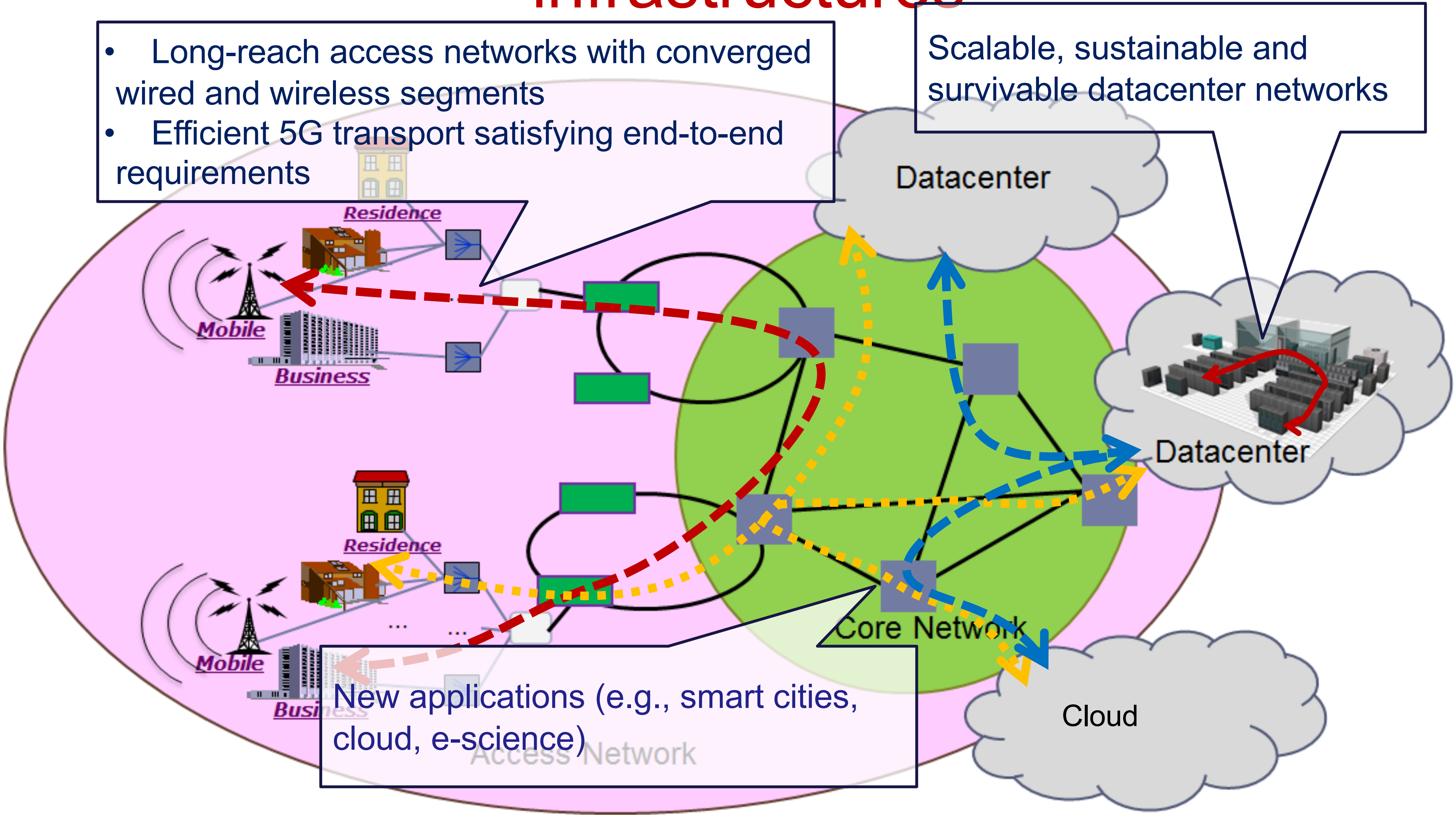
- Faculty:
  - ✓ Prof. Paolo Monti: Group Leader
  - ✓ Prof. Lena Wosinska
  - ✓ Prof. Jiajia Chen
  - ✓ Assist. Prof., Docent Marija Furdek
- 4 Postdocs
- 4 PhD students
- Visiting Ph.D. students and researchers

Optical technologies at the heart of critical infrastructures

- Long-reach access networks with converged wired and wireless segments
- Efficient 5G transport satisfying end-to-end requirements

Scalable, sustainable and survivable datacenter networks

New applications (e.g., smart cities, cloud, e-science)



ON Unit research areas

- 5G Transport Networks
- Optical Network Security
- Optical Core, Metro and Access Networks
- Optical Cloud and Datacenter Networks
- Control and Management
- Energy and Cost Efficiency
- Reliability and Disaster Recovery

Outline of the lecture

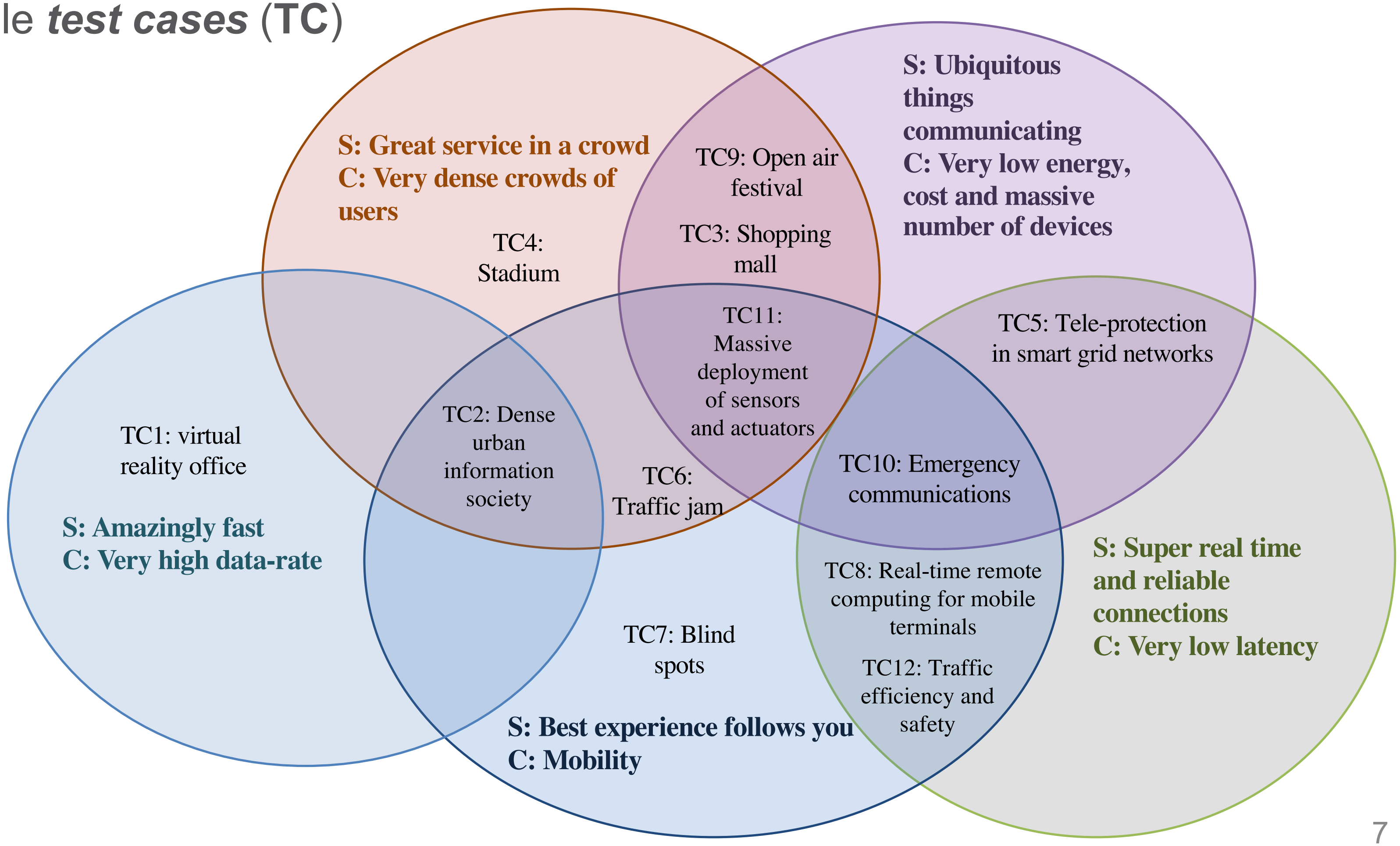
- Programmability in 5G transport infrastructures
  - SDN and NFV for E2E programmability
  - Benefits of “dynamic slicing”

- Role of machine intelligence in network orchestration
  - Traffic prediction
  - Reinforcement Learning

- A quick look at B5G and its challenges

5G challenges

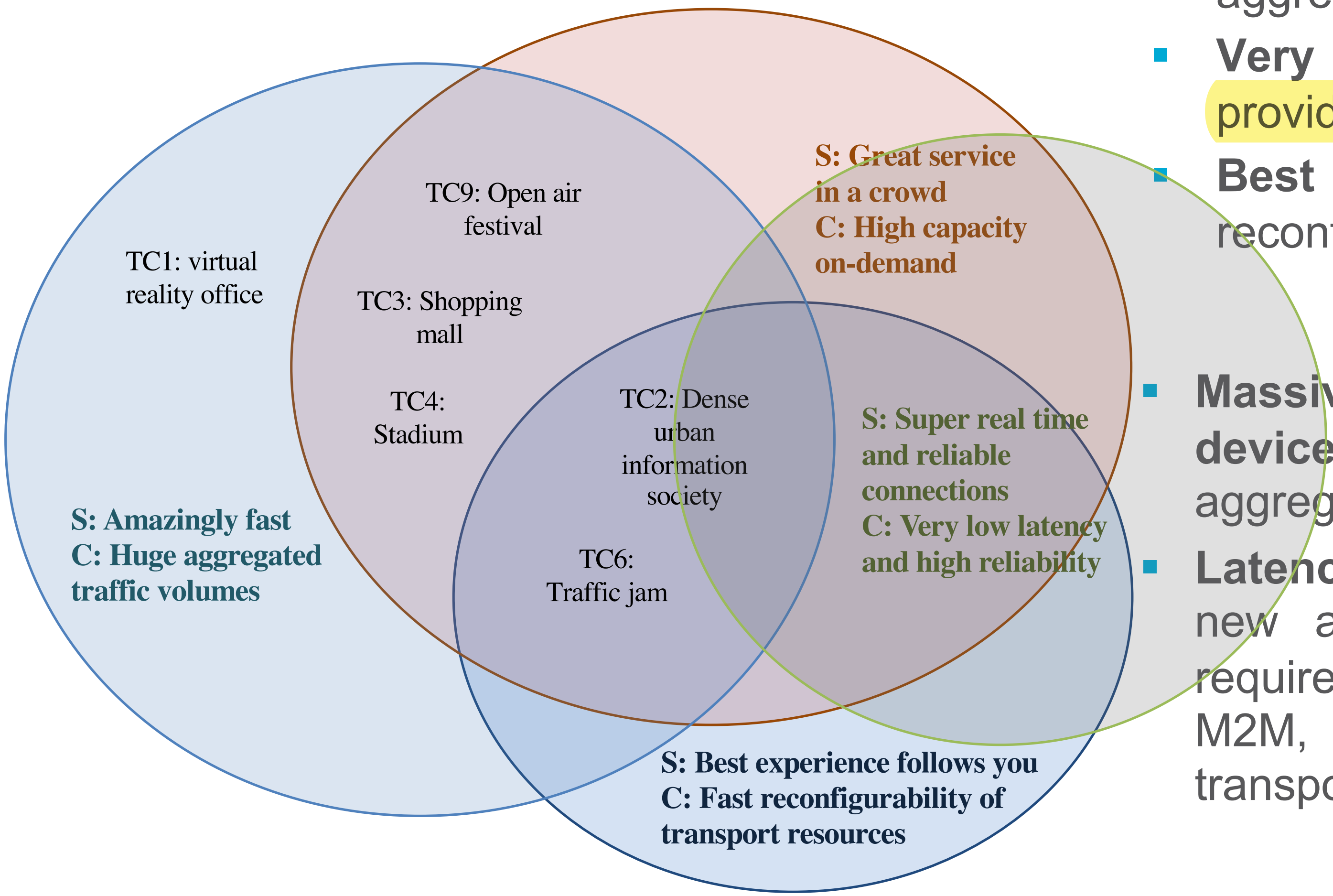
- The METIS 2020 project laid the foundation of 5G[1]
- 5G defined in terms of ***scenarios*** (**S**) supported
- Each scenario introduces a ***challenge*** (**C**)
- Each scenario multiple ***test cases*** (**TC**)



[1] METIS deliverable D1.1, "Scenarios, requirements and KPIs for 5G mobile and wireless system", April, 2013.

5G transport challenges

- The 5G challenges → transport challenges:



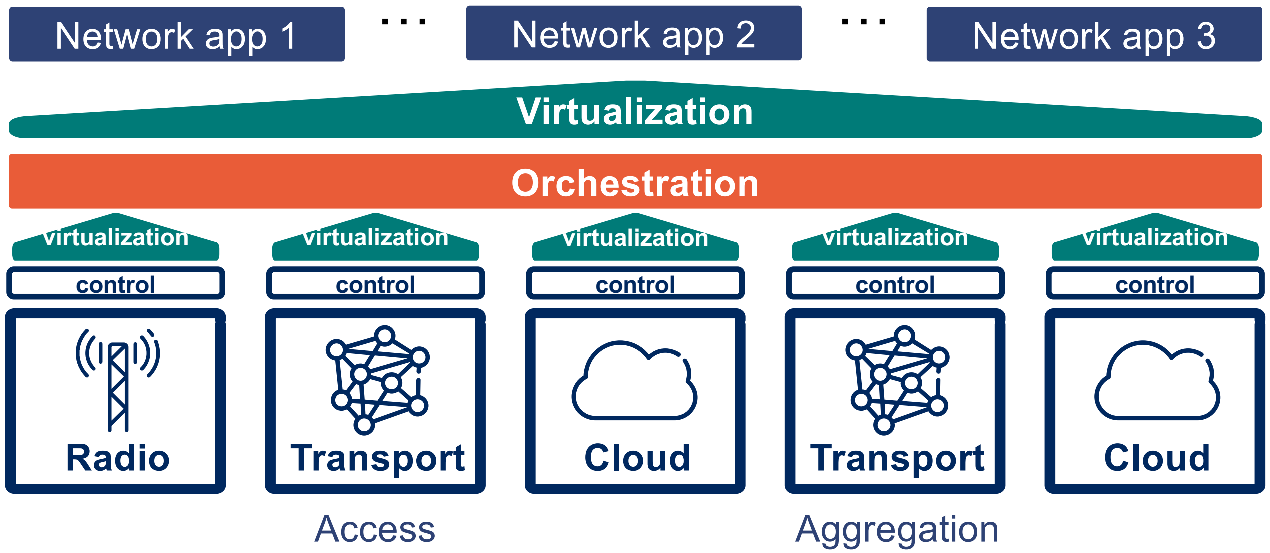
- **Very high data rate** → huge aggregated traffic volumes
- **Very dense crowds of users** → provide high capacity on-demand
- **Best experience follows you** → fast reconfigurability of transport resources

- **Massive number of connected devices** not a major issue: traffic will be aggregated
- **Latency** and **availability** are crucial: new applications with extreme delay requirements e.g., ITS, mission critical M2M, force their requirements on transport

M. Fiorani, B. Skubic, J. Mårtensson, L. Valcarenghi, P. Castoldi, L. Wosinska, P. Monti, "On the Design of 5G Transport Networks," Springer Photonic Network Communications (PNET) Journal, Vol. 30, No. 3, pp. 403-415, December 2015.

E2E programmability and flexibility

- 5G networks support a wide variety of services with a diverse set of requirements in terms of *latency*, *capacity*, and *availability*
- This requires a flexible and programmable *multi-purpose* transport network with *end-to-end orchestration* of all infrastructure resources (i.e., radio, transport, cloud)
- Software Defined Networking (SDN) and Network Function Virtualization (NFV) are promising technologies to:
  - meet the end-to-end (E2E) programmability requirements of 5G
  - allow for the sharing of infrastructure resources via *network slicing*



Slicing concept

- A key concepts in 5G networks as highlighted by the Next Generation Mobile Networks (NGMN)
- Virtualization allows for separation of software from hardware and the possibility to create *Virtual Networks (VNs)* on top of a physical infrastructure (i.e., *the slice*)
- VNs are then *allocated to applications/tenants* according to their requirements
- Slicing concept discussed by the *standardization bodies*, *vendors*, and *research community* in general

Benefits

- Scalability
- Flexibility
- Agility
- Multi-tenancy
- Resource Efficiency
- …

Challenges

- Complexity
- Upgrade traditional network management systems
- Develop open APIs
- Meet SLA of each tenant
- …

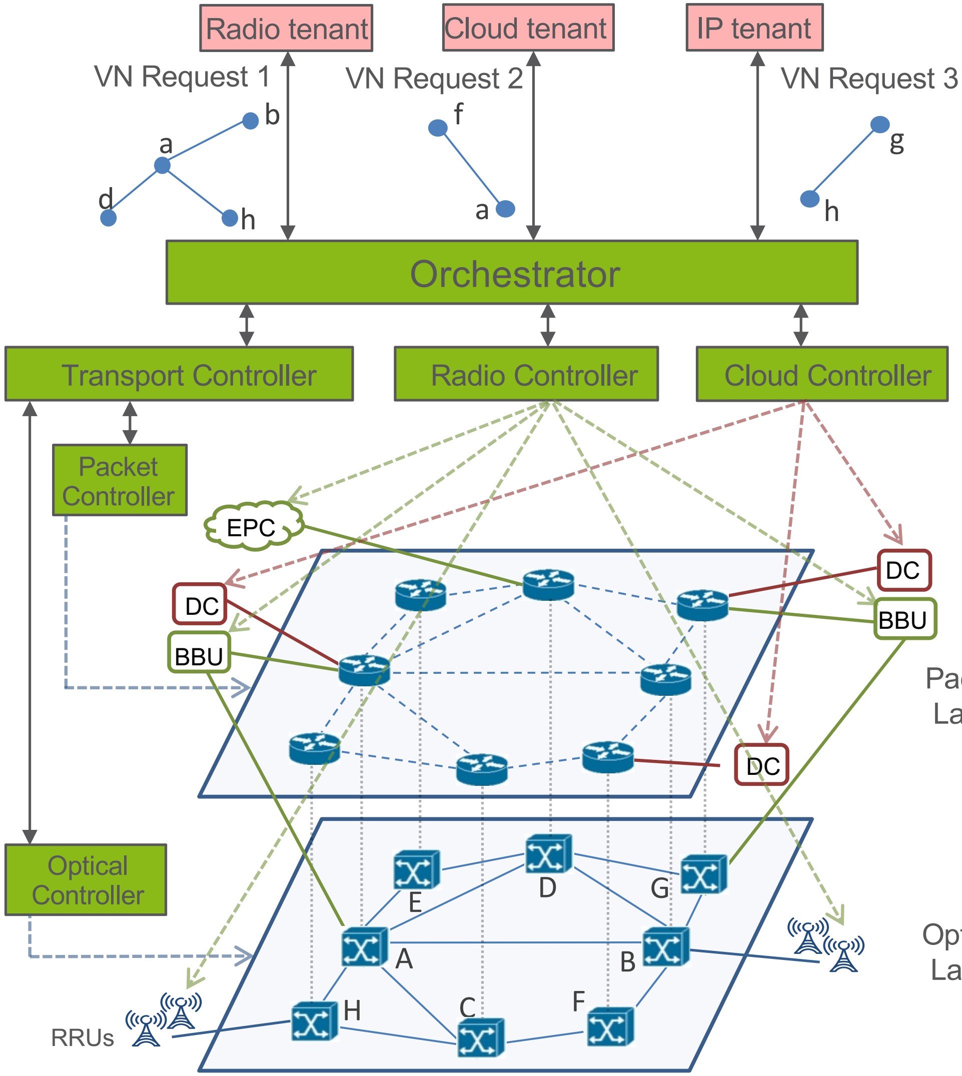
- Next Generation Mobile Networks (NGMN), "**5G White Paper**", March 2015, http://ngmn.org/5g-white-paper.html.
- Open Networking foundation (ONF) Technical Report, **"Applying SDN Architecture to 5G Slicing,"** Issue 1, April 2016.
- Nokia White Paper**, "Dynamic end-to-end network slicing for 5G,"** 2016.
- R. Nejabati , et. al., **"SDN and NFV convergence a technology enabler for abstracting and virtualising hardware and control of optical networks (invited),"** in OFC 2015.
- A. Mayoral , et. al., **"Need for a Transport API in 5G for Global Orchestration of Cloud and Networks Through a Virtualized Infrastructure Manager and Planner,"** JOCN, 2017.
- A. Autenrieth, et.al., "**Optical Network Programmability – Requirements and Applications**", in PS 2015.

Dynamic slicing

- One aspect highlighted in the literature: *network slices should be dynamic*
- Network and cloud resources should be *provisioned on the fly* and should match as close as possible the *needs* of the *tenants' services* (that in turn may *vary* in *temporal dimension*)
- This should lead to *efficient resource utilization* (when compared to a static provisioning of slices) and *significant cost savings* for the network infrastructure providers
- Will now describe strategies for provisioning and maintaining network slices considering:
  - dynamic arrival/departure of tenants' service requests
  - temporal variations of their resource requirements

Use case: multi-purpose transport

- *Orchestrator* harmonizes transport, radio, and cloud resources managed by different controllers
- *Tenants* issue *VN requests* which include connectivity information as well as specific resource constraints
- *Orchestrator* in charge of coordinating the provisioning of VNs into the physical network and notifies tenants whether services have been accepted or not



Types of tenants

- The *radio tenant* asks to connect its Remote Radio Units (RRUs) to the Evolved Packet Core (EPC) by going through a specific (or any of the available) Base Band Unit (BBU) pool

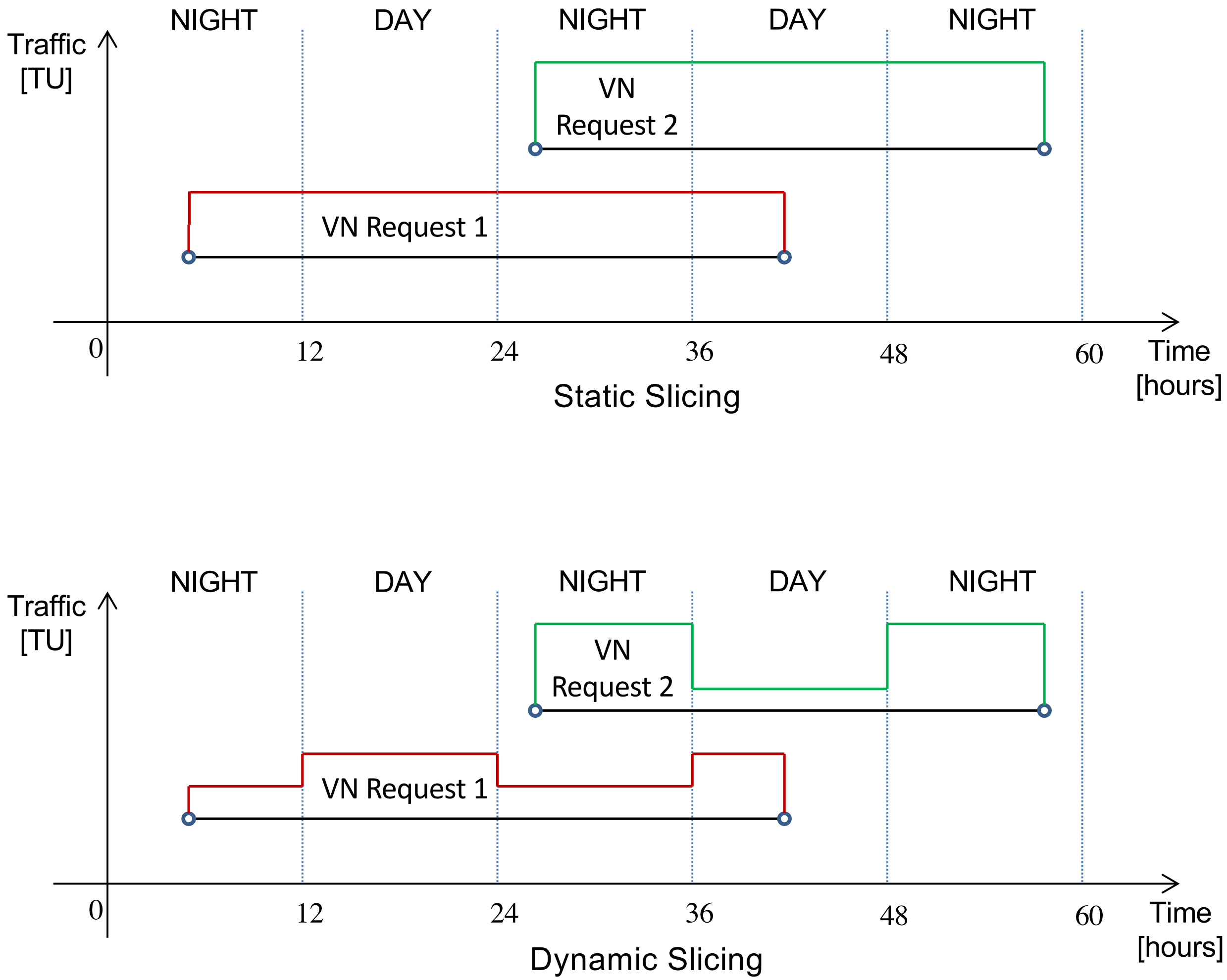
- The *cloud tenant* asks for connectivity among the Data Centers (DCs) with certain computing/storage capacity requirements (e.g., for synchronization or backup among DCs)

- The *IP tenant* asks for a certain bandwidth between two Internet Service Providers (ISPs) access points

- The bandwidth (BW) required by the virtual links of the VNs may change during time
  - Day vs. night BW requirements variations
  - Radio and IP tenants have higher BW requirement during the day
  - Cloud tenant has higher BW requirement at night, e.g., to transfer the files between the DCs

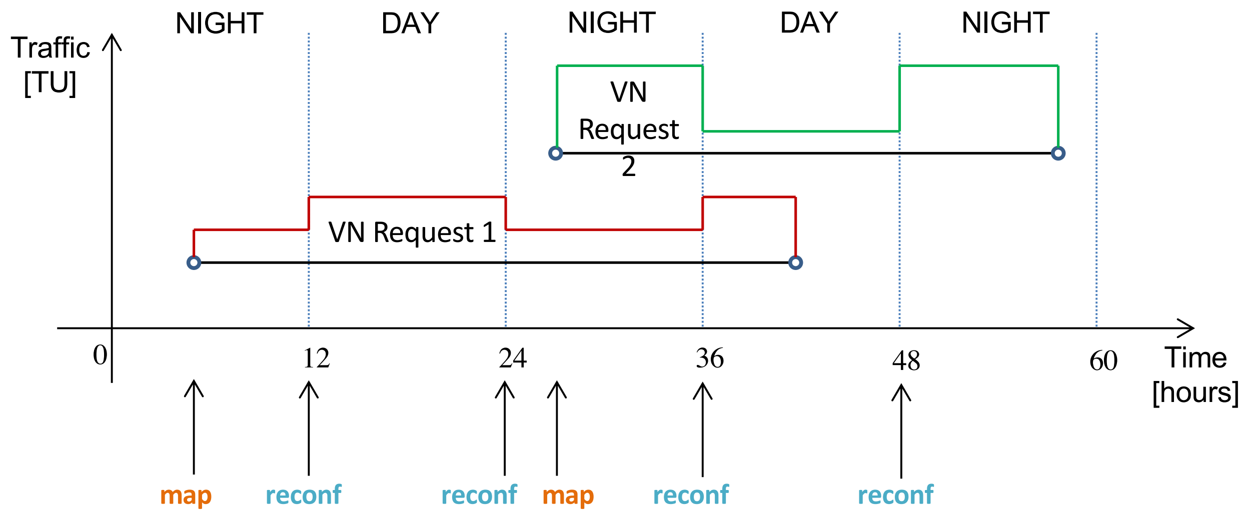
Static and dynamic slicing

- **Static Slicing:** embed the VNs into the physical network with peak BW requirements of each link
- **Dynamic Slicing:** reconfigure the VNs according to change in BW requirements during day and night



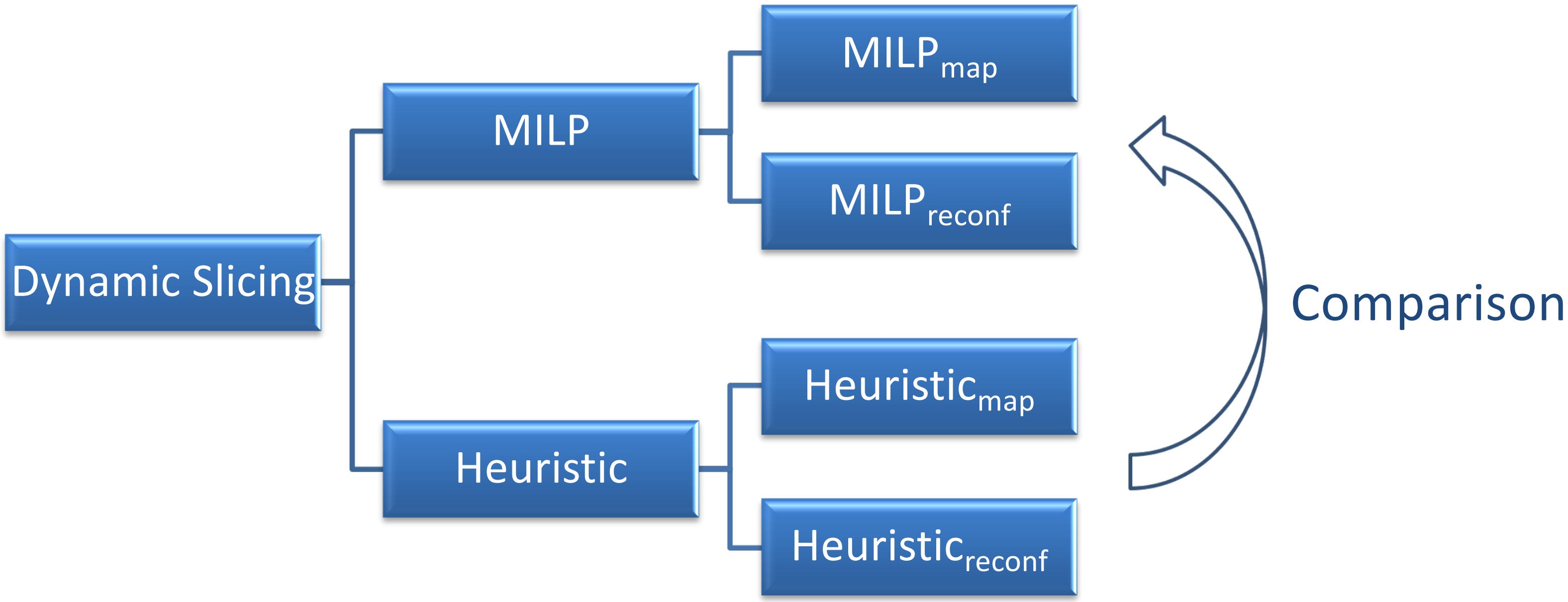
Problem description

- <u>The *dynamic slicing provisioning problem* comprises two parts:</u>
  1. **Mapping** of the VN request into the physical network considering the traffic requirements at the time of arrival (i.e., either day or night)
  2. **Reconfiguration** of all the currently mapped VNs when switching between day and night
- <u>Assumption</u>: when a VN request is rejected, it is not possible to embed it at a later time when the physical network resources become available



Proposed solutions

- *Mixed integer linear programming (MILP) formulations* for mapping and reconfiguration of the VN requests
- MILP formulations tradeoff execution time and optimality
- *Heuristic algorithm* for dynamic slicing and its performance is benchmarked with the results from MILP formulations



M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska, P. Monti, "Dynamic slicing approach for multi-tenant 5G transport networks," IEEE/OSA Journal of Optical Communication and Networking (JOCN), vol. 10, no. 1, pp. A77-A90, Jan. 2018.

MILP formulations

- The objective of **MILP$_{map}$** is to minimize the wavelength resource usage in the network

$$\min\left(\sum_{mn}\sum_{ij} z_{mn}^{ij}\right)$$

z_{mn}^{ij} = number of lightpaths from node i to node j, while passing through the link (m,n) in the physical network

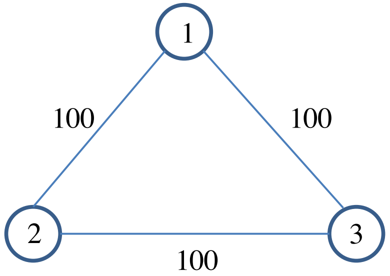
- The objective of **MILP$_{reconf}$** is to minimize the (possible) degradation of each virtual link, the number of reconfigured lightpaths, and the wavelength resource usage in the network

$$\min\left(\alpha \underbrace{\sum_{v}\sum_{be} d_{be}^{v}}_{\text{degradation of all virtual links}} + \beta \underbrace{\sum_{ij} u_{ij}}_{\text{reconfiguration of lightpaths}} + \gamma \underbrace{\sum_{mn}\sum_{ij} z_{mn}^{ij}}_{\text{wavelength usage}}\right)$$

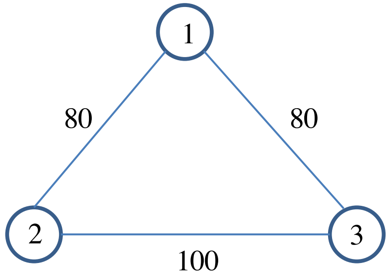
where $\alpha \gg \beta \gg \gamma$

M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska, P. Monti, "Dynamic slicing approach for multi-tenant 5G transport networks," IEEE/OSA Journal of Optical Communication and Networking (JOCN), vol. 10, no. 1, pp. A77-A90, Jan. 2018.

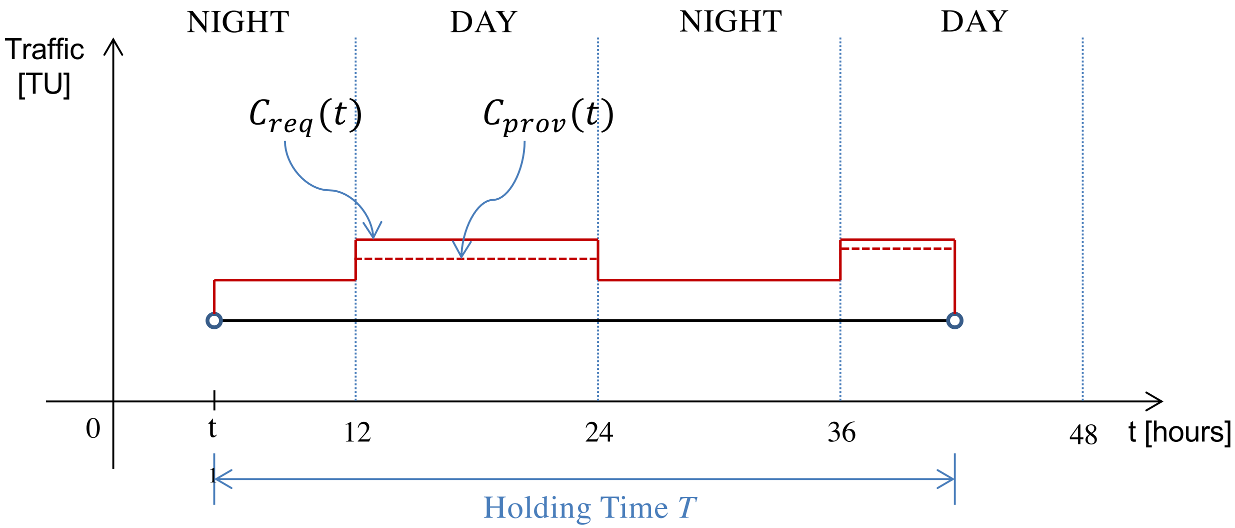
VN degradation



Total capacity required (C_{req})
at reconf = 300 Gbps



Total capacity provided (C_{prov})
after reconf= 260 Gbps

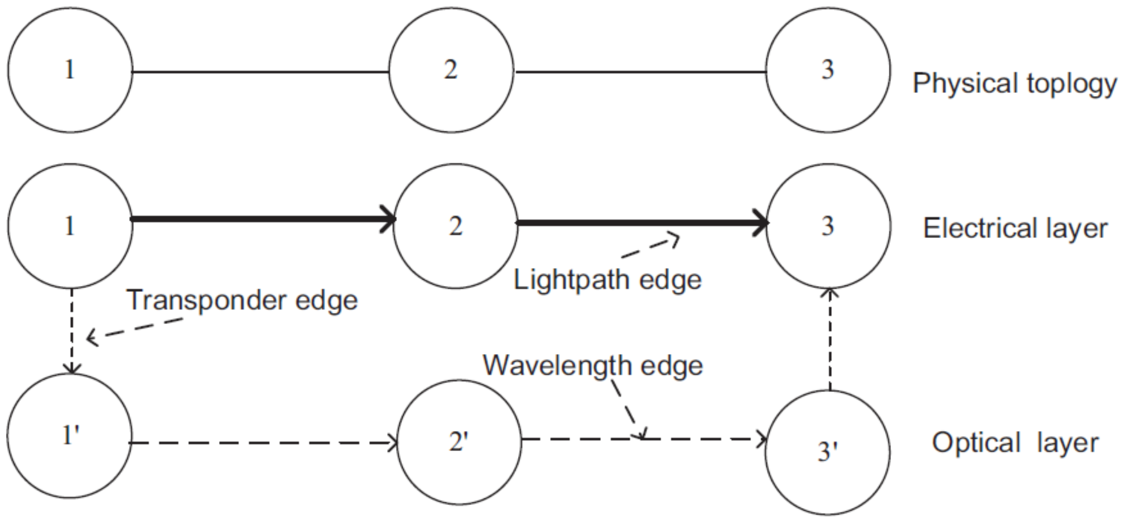


VN degradation D is computed as:

$$D = \frac{\int_{t_1}^{t_1+T} C_{req}(t) - \int_{t_1}^{t_1+T} C_{prov}(t)}{\int_{t_1}^{t_1+T} C_{req}(t)}$$

Heuristic

- Cost functions are the same as in the MILP case, but both heuristic algorithms (i.e., to map and reconfigure) are based on constructing a *grooming graph* [1] for each VN request



Grooming graph for WDM network [1]

- *Lightpath edge*: if an existing lightpath has sufficient remaining bandwidth to support the traffic demand
- *Wavelength edge*: if at least one free wavelength is available in the fiber link
- *Transponder edge*: if a transceiver is available for E/O and O/E conversion
- In order to minimize number of new lightpaths, the weights are selected as [2]:

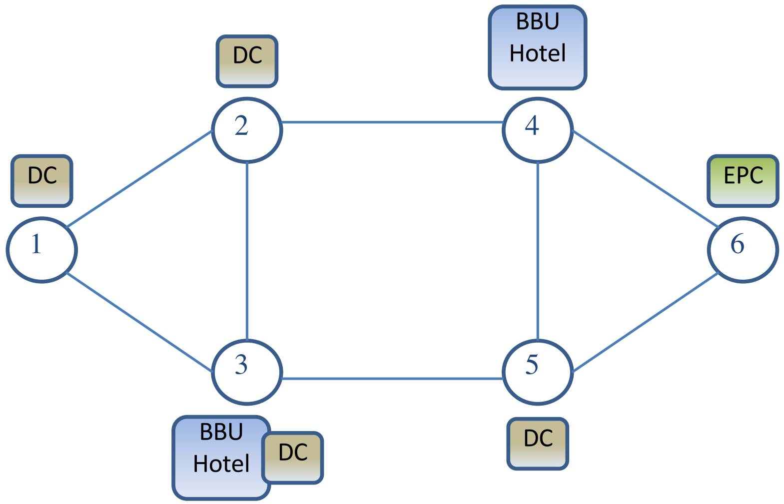
  $W_{TP} = 100$; $W_{LP} = 0.01 \times H$; $W_{WL} = 0.01$

  where H: number of hops in lightpath

[1] S. Zhang et al., **"Network virtualization over WDM and flexible-grid optical networks,"** Optical Switching and Networking, November 2013.
[2] J. Zhang et al., **"Dynamic virtual network embedding over multilayer optical networks,"** JOCN, 2015.

Performance evaluation

- 6-node network as the physical network[1]
- Each fiber link is assumed to have 80 wavelengths, with 100 Gbps capacity in each wavelength
- It is assumed that the BBU hotels and DCs have sufficient BBU ports and capacity available so that there is no VN rejection due to node mapping



[1] S. Zhang et al., "Network virtualization over WDM and flexible-grid optical networks," Optical Switching and Networking, November 2013.

VN requests

- **Radio:**

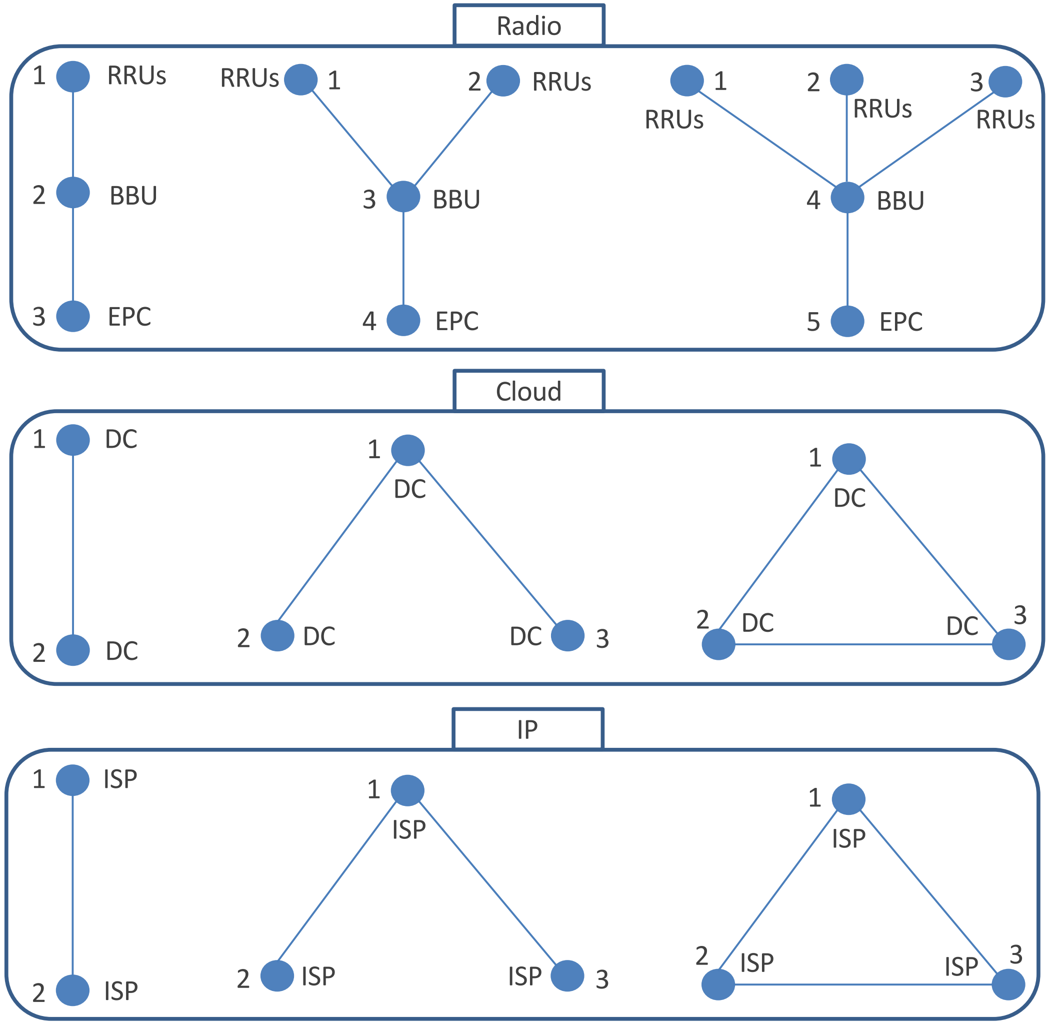
  Day/Night variation factor α = 1/8[1]
  number of RRUs per node ~ Uniform(5,15)
  link capacity per RRU (Day) = 10 Gbps
  backhaul = 10% of fronthaul

- **Cloud:**

  Day/Night variation factor β = 25[2]
  link capacity (Day) ~ Uniform(50,100) Gbps
  computational capacity per node (C) = 100 units

- **IP:**

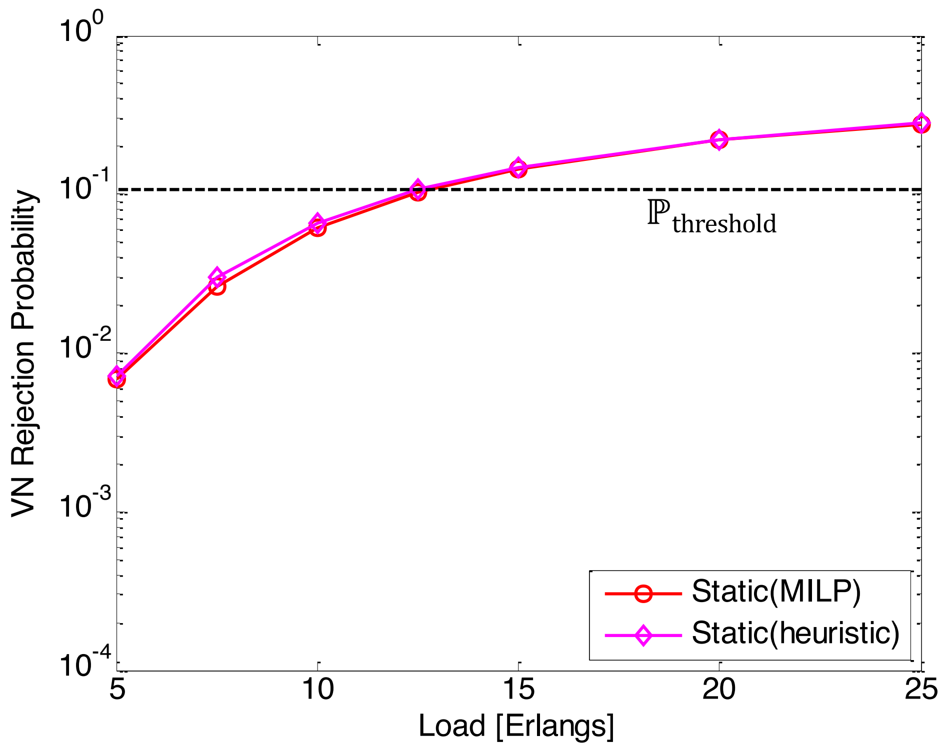
  Day/Night variation factor γ = 1/8[2]
  link capacity (Day) ~ Uniform(1200,1500) Gbps

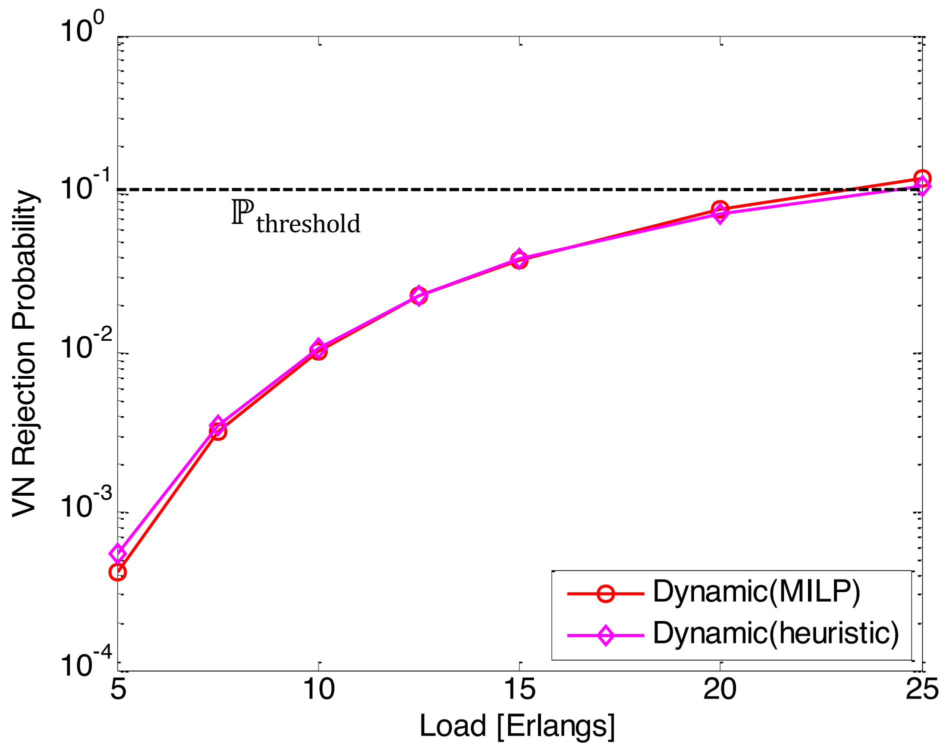


[1] EARTH Deliverable D2.3, **"Energy Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown,"** 2012
[2] F. Morales et al., **"Virtual Network Topology Reconfiguration based on Big Data Analytics for Traffic Prediction,"**, OFC 2016

VN rejection probability

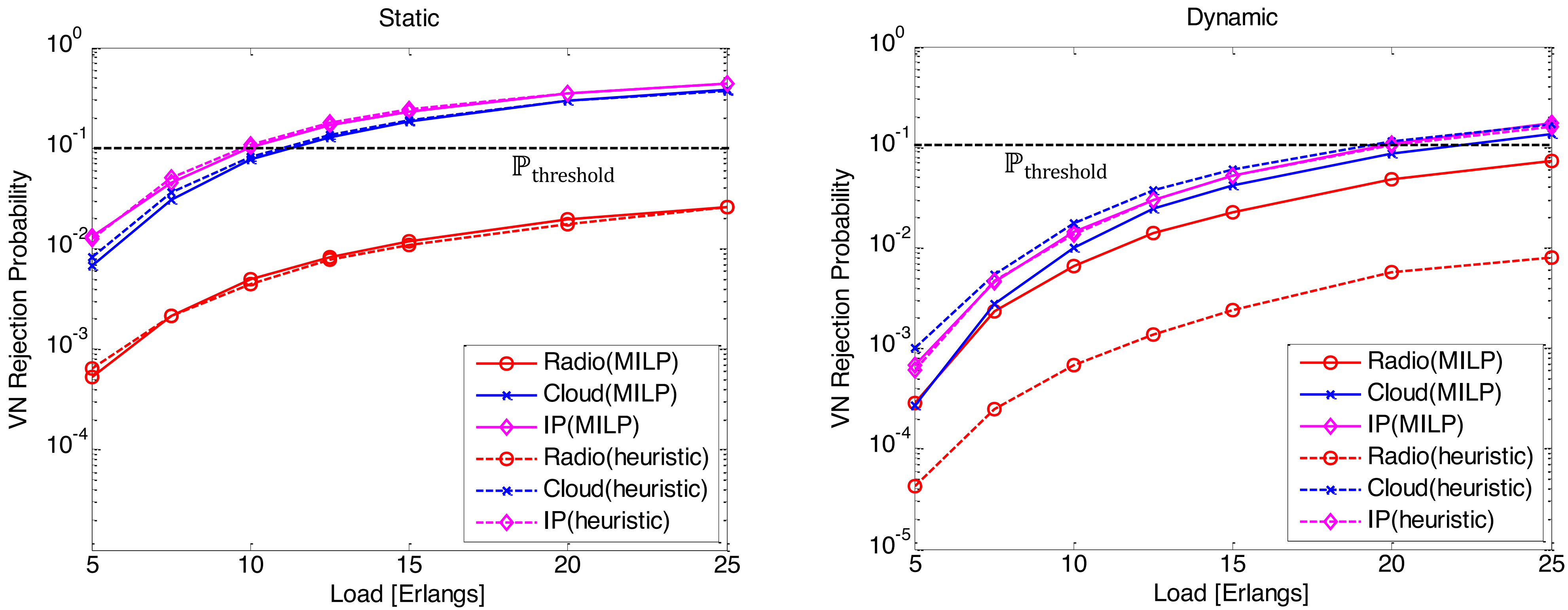
- D/N duration = 12 hours
- Mean holding time of VN requests = 50 hours (holding times and inter-arrival times exponentially distributed)
- VN requests in each experiment = 1500
- No. of experiments = 50





Per-VN type breakdown

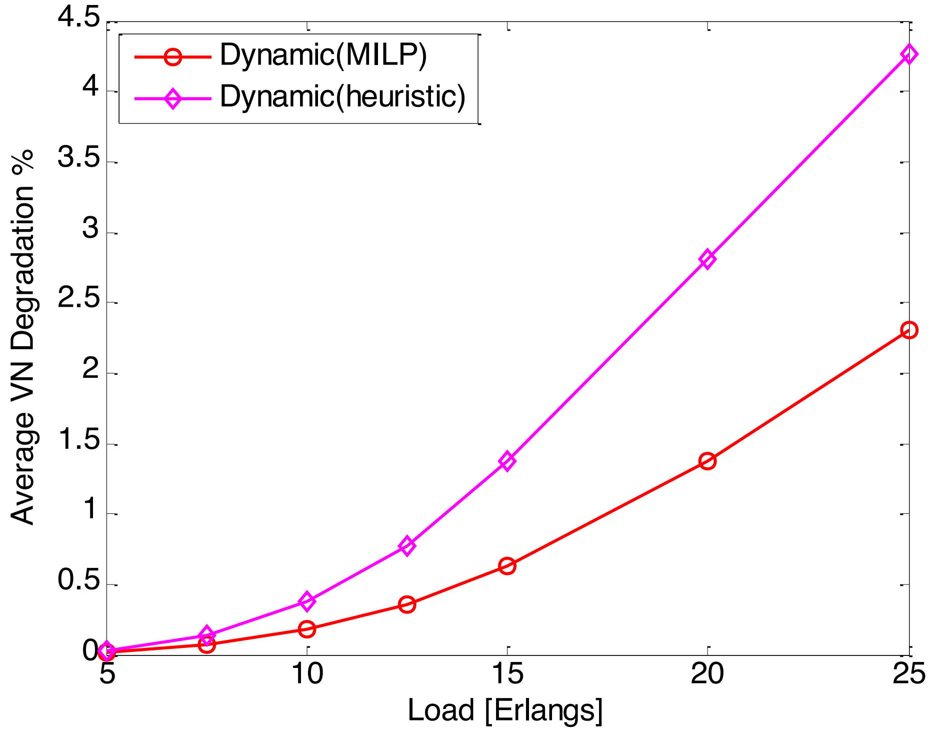
- D/N duration = 12 hours
- Mean holding time of VN requests = 50 hours (holding times and inter-arrival times are exponentially distributed)
- VN requests in each experiment = 1500
- No. of experiments = 50

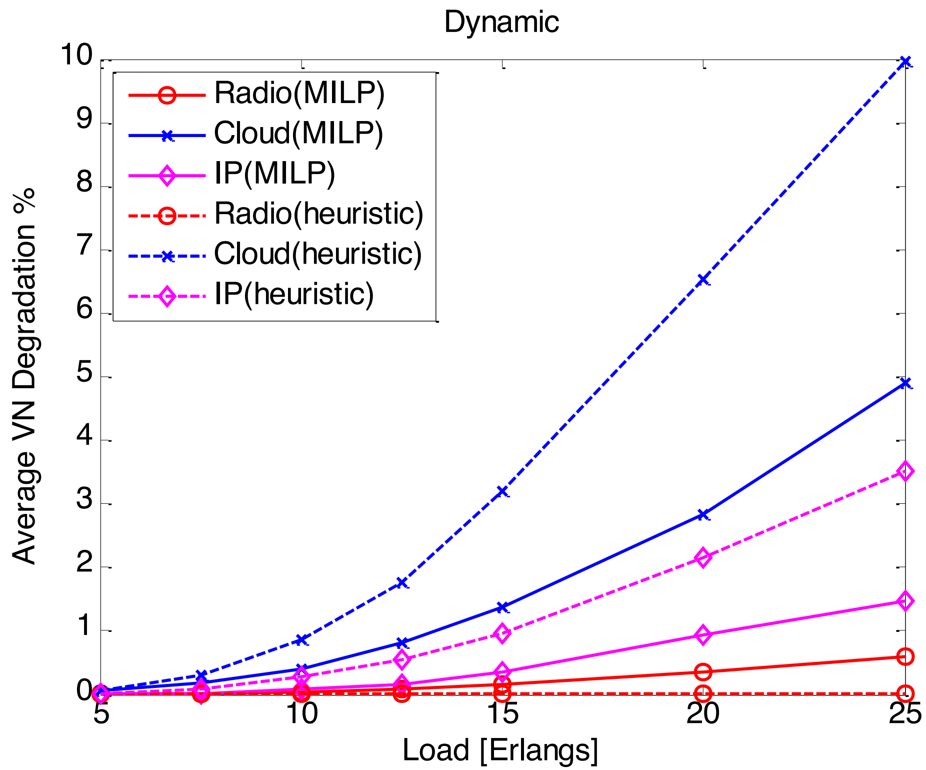


VN degradation

- D/N duration = 12 hours
- Mean holding time of VN requests = 50 hours (holding times and inter-arrival times are exponentially distributed)
- VN requests in each experiment = 1500
- No. of experiments = 50

$$D_{avg} = \frac{\sum_{i=1}^{N_{accepted}} D_i}{N_{accepted}}$$





A few thoughts about dynamic slicing

- Dynamic slicing has the potential to considerably improve the VN rejection probability
- This improvement can help infrastructure providers to increase their revenues
- However, this scheme has a trade-off in terms of degradation which is observed to be very small for acceptable levels of rejection probability

- General question: when we look at the revenue figures of an operator should all the slice requests be accepted in the first place?

Outline of the lecture

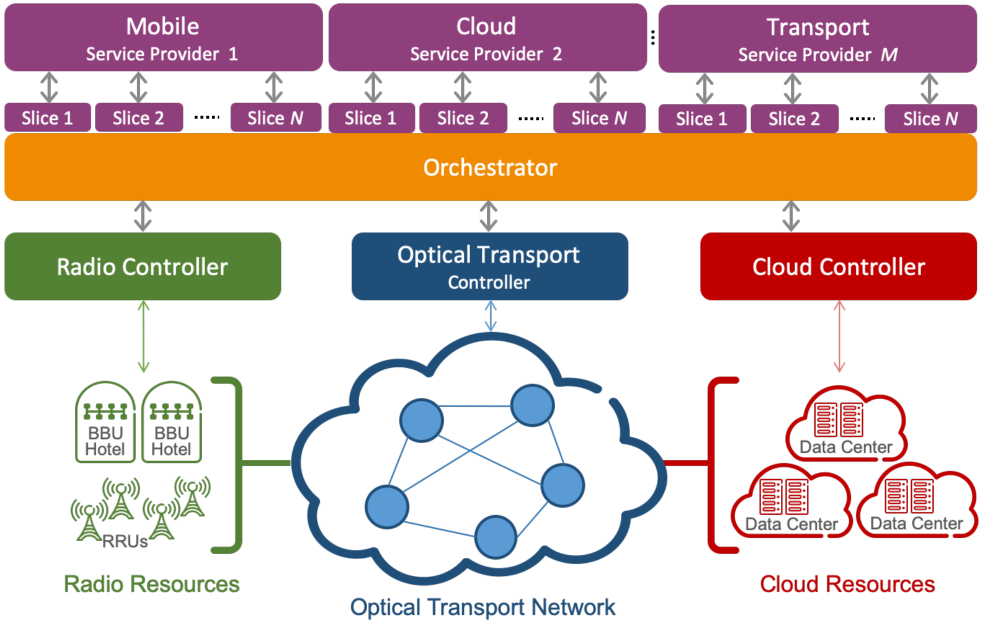
- Programmability in 5G transport infrastructures
  - SDN and NFV for E2E programmability
  - Benefits of “dynamic slicing”

- Role of machine intelligence in network orchestration
  - Traffic prediction
  - Reinforcement Learning

- A quick look at B5G and its challenges

DYNAMIC SLICING

- *Slicing*: SDN and NFV allow Infrastructure Providers (InP) to share resources among different tenants
- During *provisioning/operation* beneficial to adapt resources assigned to a slice to match time varying requirements: *dynamic slicing*
- Slice acceptance ratio can be greatly improved but with introduction of small service degradation*

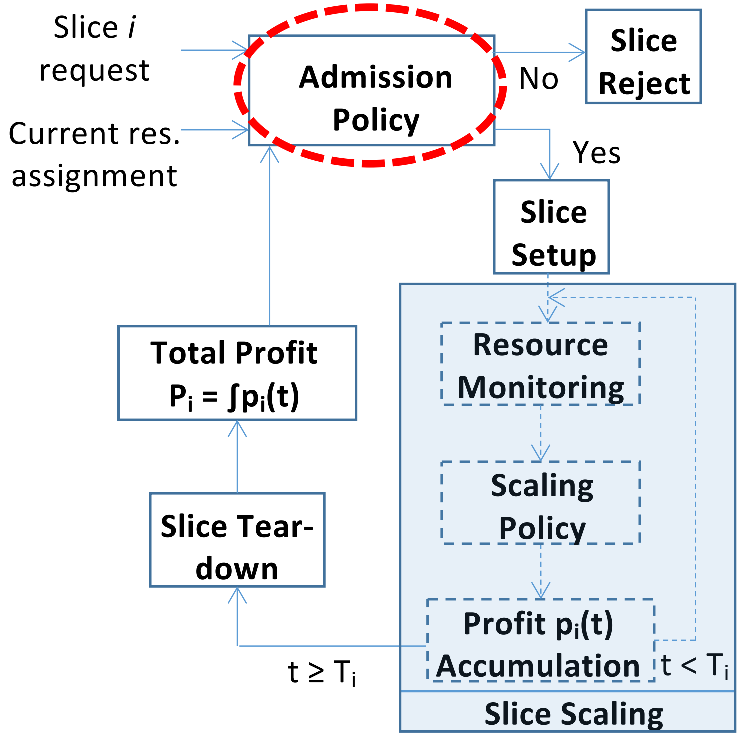


- Crucial to have intelligent policy that accepts only slices which not likely to create performance degradation
  - understand when/where resource bottlenecks might appear in the infrastructure
  - deciding which slice to accept in order to maximize the profit of an InP

* M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska and P. Monti, “Dynamic Slicing Approach for Multi-tenant 5G Transport Networks”, *in IEEE/OSA JOCN*, Vol. 10, No. 1, 2018

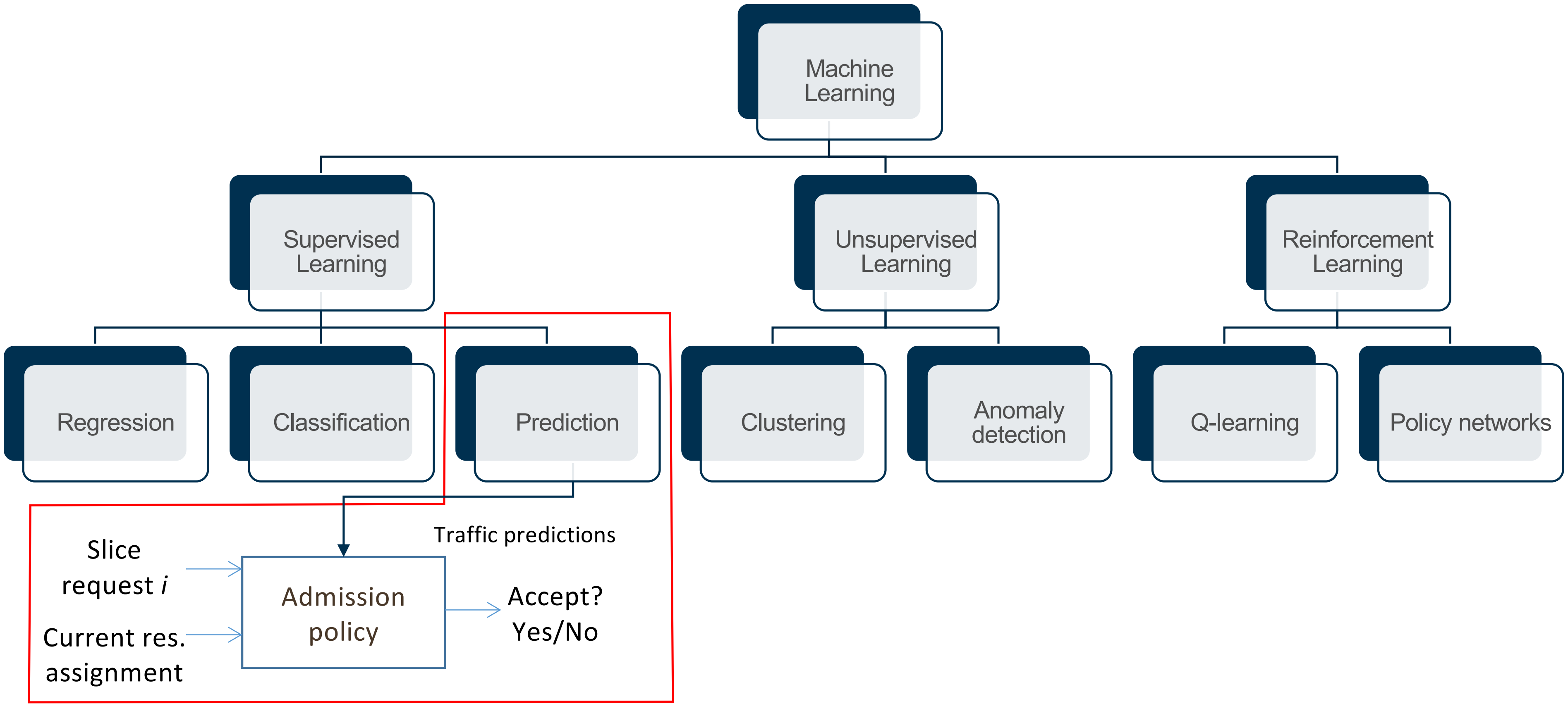
INTELLIGENT SLICE ADMISSION POLICY

- Scenario:
  - tenant(s) requests slices with different requirements and priorities
  - different priorities mean different *revenue* and *penalty* levels
- Objective:
  - *admission policy* used by InP to accepts/rejects incoming slice request with aim to maximize *Profit = revenue – penalty*
- Intuition:
  - beneficial to *proactively* reject some *low priority (low revenue)* services to make space for future *high priority (high revenue)* ones



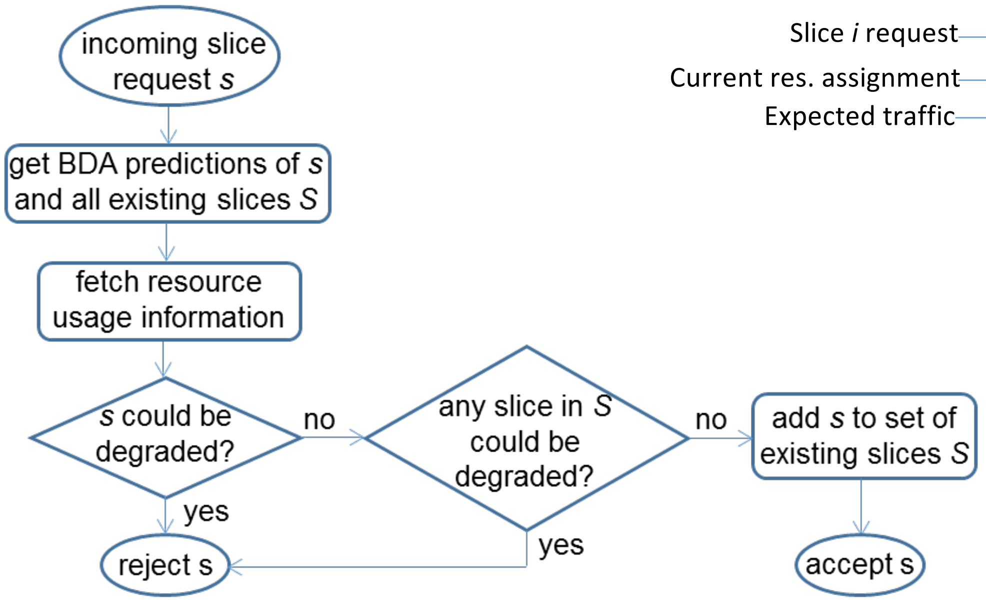
- ML can be beneficial in this process

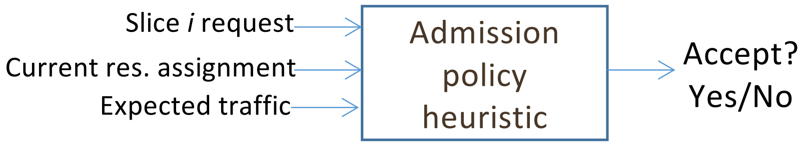
MACHINE LEARNING ALGORITHMS



SLICE ADMISSION USING TRAFFIC PREDICTION

- Use BDA for predicting temporal variation of resource requirements of each slice

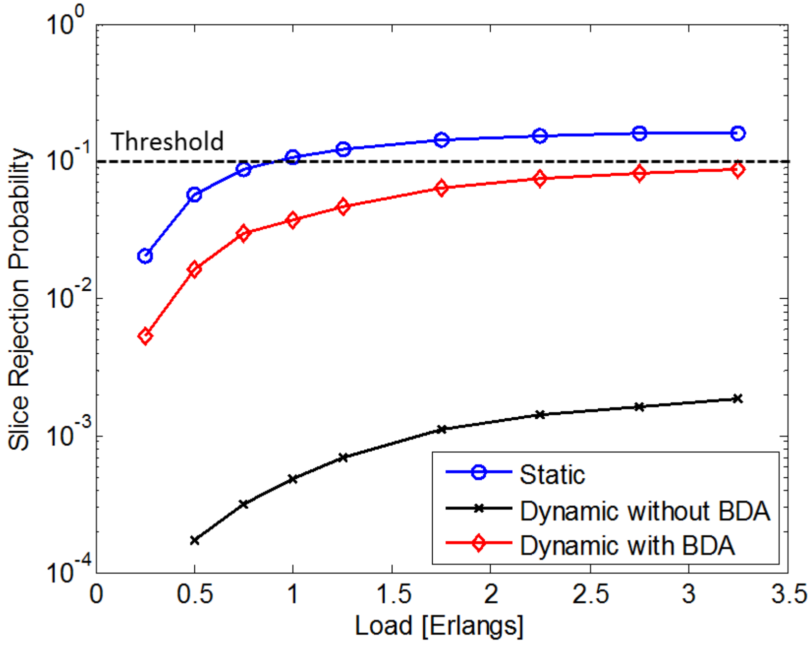


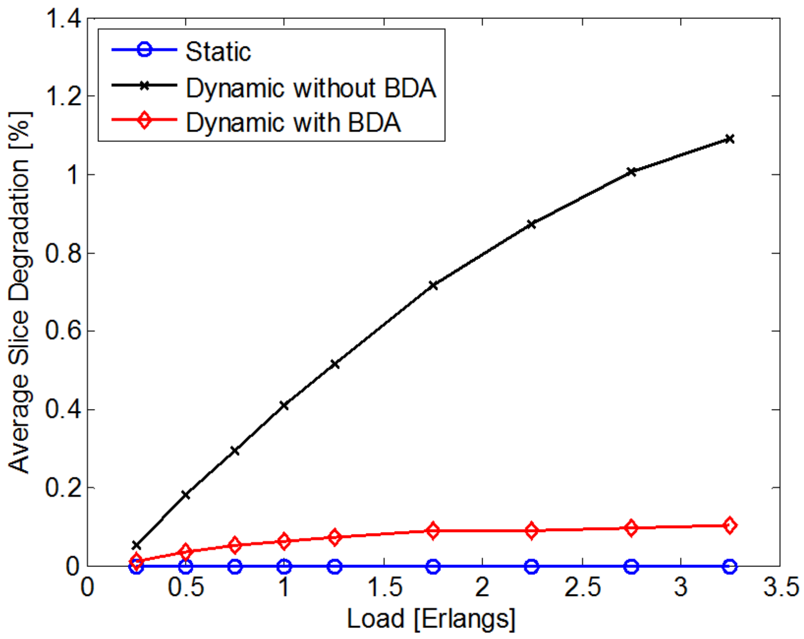


M. R. Raza, A. Rostami, L. Wosinska and P. Monti, “A Slice Admission Policy Based on Big Data Analytics for Multi-tenant 5G Networks”, in IEEE/OSA JLT, 2019.

SLICE BLOCKING RATIO VS. DEGRADATION

- Mean holding time = 2 days (holding times and inter-arrival times exponentially distributed)
- Simulation time = 365 days
- No. of experiments = 50



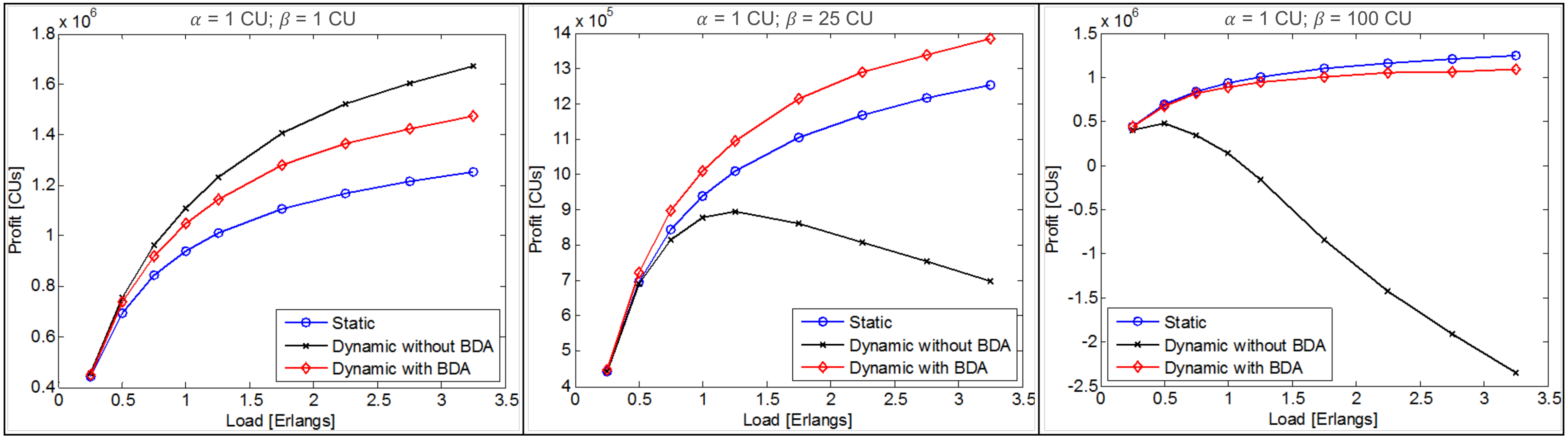


M. R. Raza, A. Rostami, L. Wosinska and P. Monti, “A Slice Admission Policy Based on Big Data Analytics for Multi-tenant 5G Networks”, in IEEE/OSA JLT, 2019.

IMPACT ON INP PROFIT

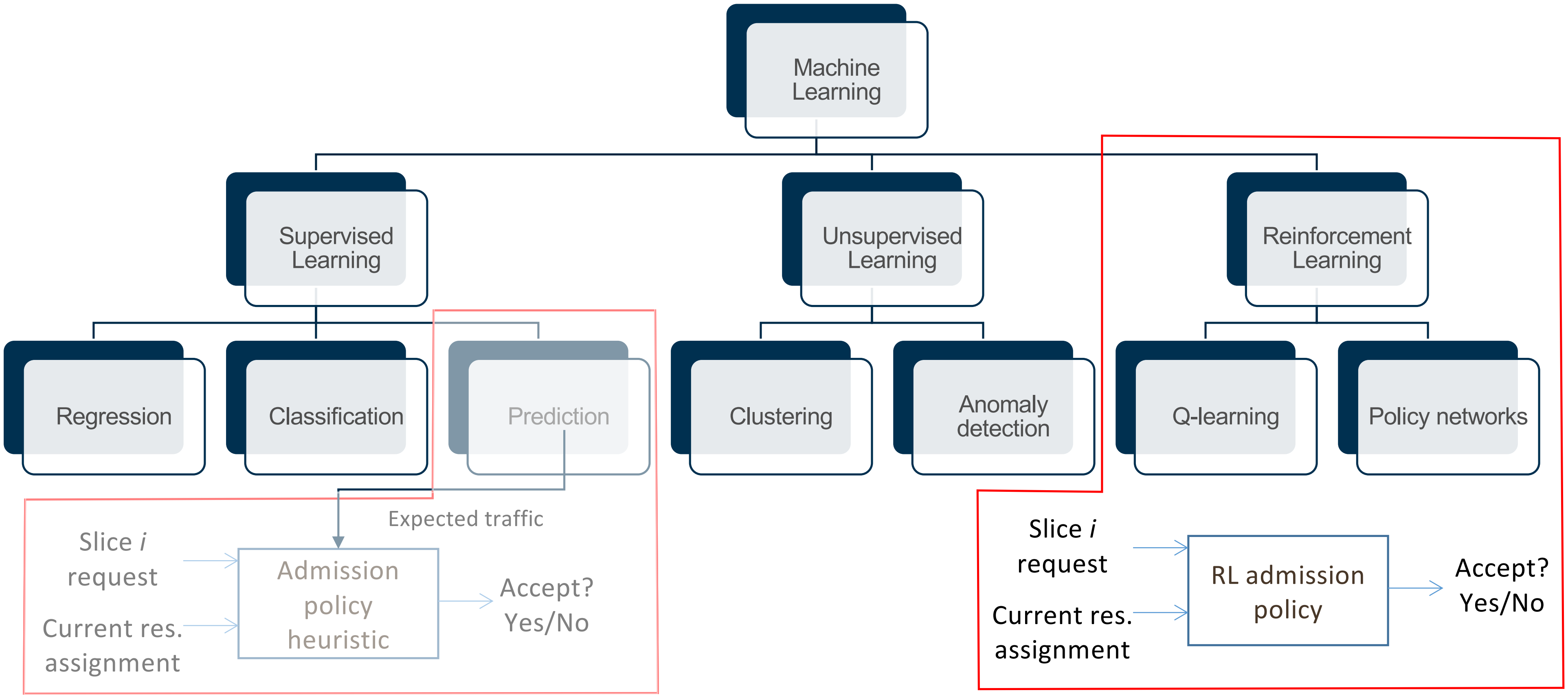
$$\underbrace{P}_{\text{profit}} = \underbrace{\left(\sum_{s \in A} \int_{t = at_s}^{at_s + ht_s} R_s(t) \right) \times \alpha}_{\text{revenue}} - \underbrace{\left(\sum_{s \in A} \int_{t = at_s}^{at_s + ht_s} D_s(t) \right) \times \beta}_{\text{penalty}},$$

- $R_s(t)$ = resources required by slice s at time t
- $D_s(t)$ = degradation of slice s at time t
- at_s = arrival time of slice s
- ht_s = holding time of slice s
- A = set of all accepted slices
- α = per-time-unit per-resource revenue factor
- β = per-time-unit per-resource penalty factor



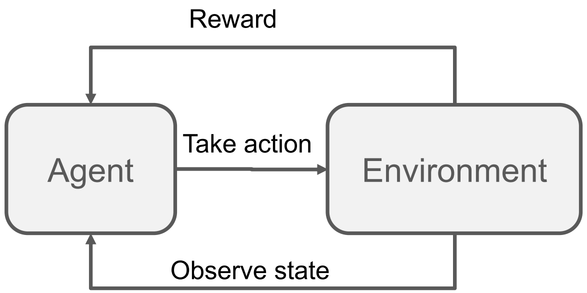
M. R. Raza, A. Rostami, L. Wosinska and P. Monti, “A Slice Admission Policy Based on Big Data Analytics for Multi-tenant 5G Networks”, in IEEE/OSA JLT, 2019.

MACHINE LEARNING ALGORITHMS

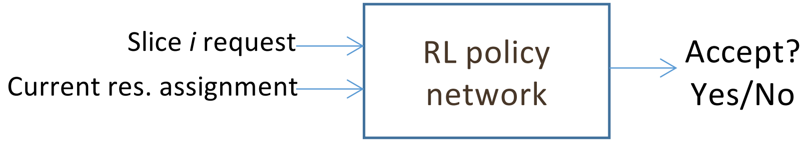


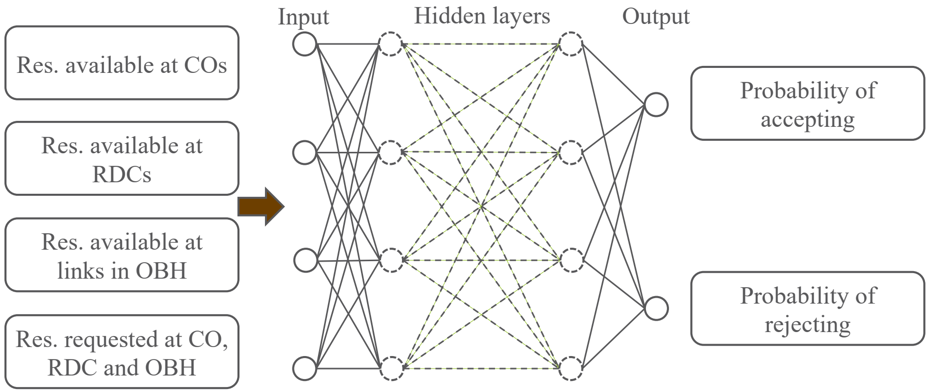
SLICE ADMISSION USING REINFORCEMENT LEARNING

- Understand how *admission decisions* impact the *reward* that is modelled as the *loss of revenue*



- Use *policy networks* to learn an admission policy that maximizes the reward over time

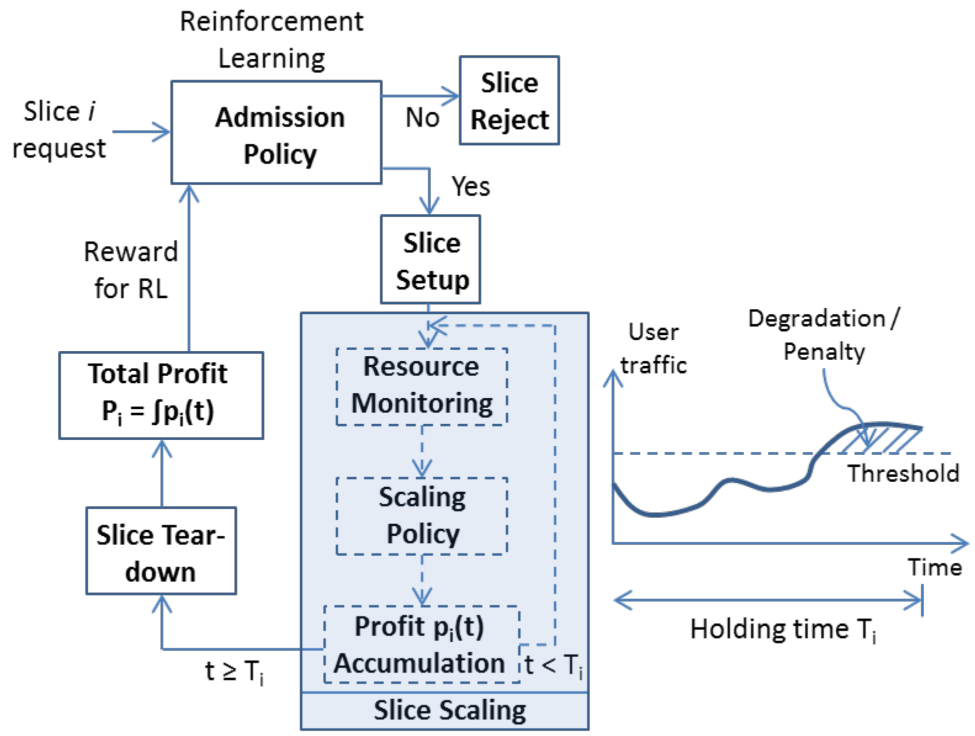




M. R. Raza, C. Natalino, P.Öhlen, L. Wosinska and P. Monti, “Reinforcement Learning for Slicing in a 5G Flexible RAN”, in IEEE/OSA JLT, 2019.

SLICE ADMISSION USING REINFORCEMENT LEARNING

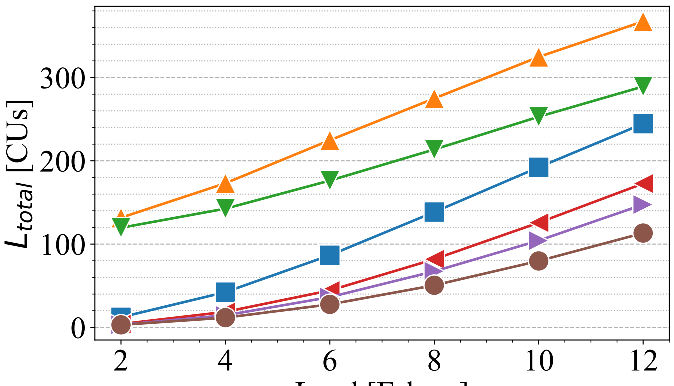
- Can the performance of dynamic slicing be improved by using reinforcement learning in the slice admission process?
- Can RL adapt to different *scaling policies* and to different *traffic composition*?



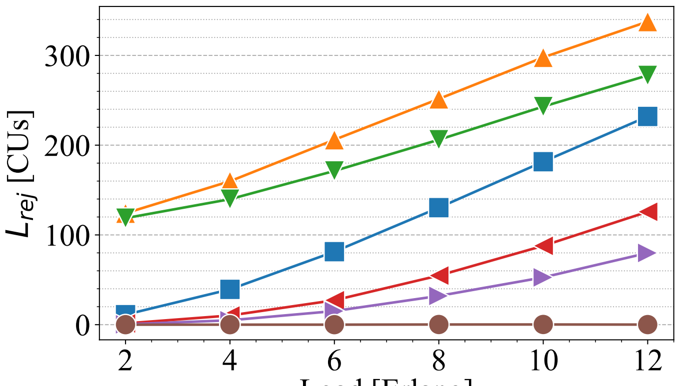
PERFORMANCE EVALUATION

- Scaling policy = high priority first (HPF), with 50% HP-50% LP services
- $L_{total} = L_{rej} + L_{deg}$ (sum of the rejection loss and degradation loss), penalty ratio = 1.5
- NN with 4 hidden layer and 40 neurons
- Test results after 2500 training iterations
- RL shows 23% improvement vs. OS-50, 60% vs. RR-50, and 53% vs. Fit

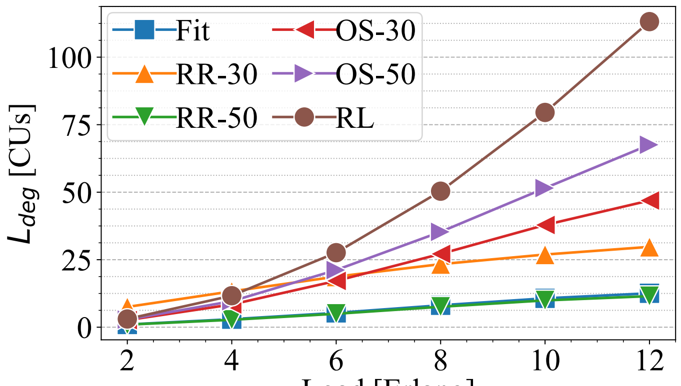
RR: Resource reservation
OS: Oversubscription



(a) Total loss L_{total}



(b) Loss from rejected services L_{rej}

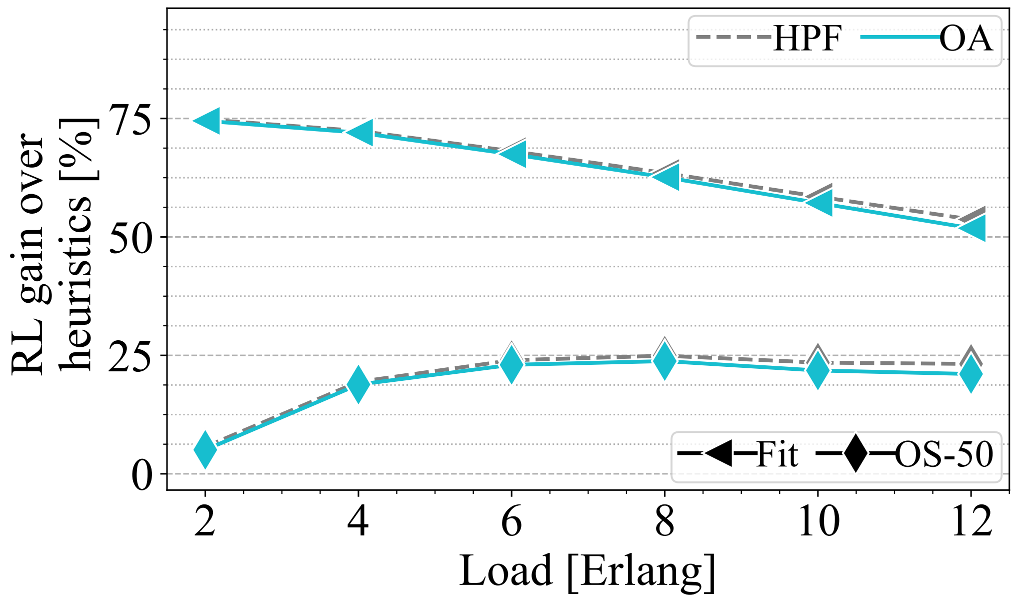


(c) Loss from degraded services L_{deg}

M. R. Raza, C. Natalino, P.Öhlen, L. Wosinska and P. Monti, “Reinforcement Learning for Slicing in a 5G Flexible RAN”, in IEEE/OSA JLT, 2019.

VARYING PENALTY RATIO AND SCALING POLICY

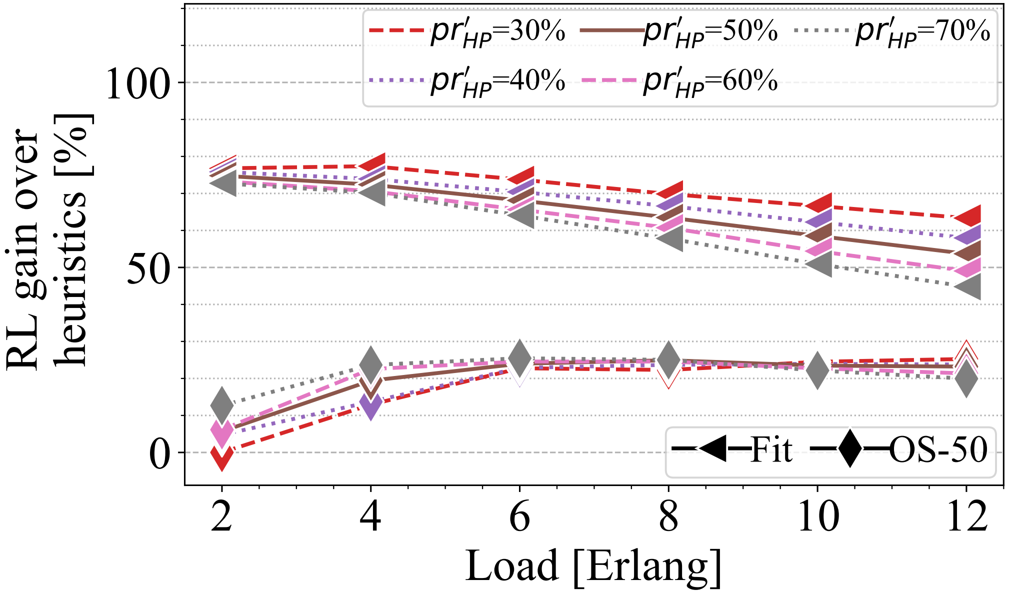
- Different scaling policies
- High priority first (HPF) and Order of Arrival
- Penalty ratio = 1.5



M. R. Raza, C. Natalino, P.Öhlen, L. Wosinska and P. Monti, “Reinforcement Learning for Slicing in a 5G Flexible RAN”, in IEEE/OSA JLT, 2019.

VARYING TRAFFIC COMPOSITION

- RL is trained with one traffic condition: 50% HP / 50% LP
- The same agent is tested with varying traffic conditions (pr'_{HP})
- When the percentage of HP services is different from the one used for the training, RL agent is able to maintain most of its gain over the heuristics



M. R. Raza, C. Natalino, P.Öhlen, L. Wosinska and P. Monti, "Reinforcement Learning for Slicing in a 5G Flexible RAN", in IEEE/OSA JLT, 2019.

A few thoughts about slice admission using ML

- ML can help to minimize the penalty experienced by infrastructure providers by taking intelligent decisions for slice admission control
- ML adapts itself according to the different scaling policies and traffic compositions

Outline of the lecture

- Programmability in 5G transport infrastructures
  - SDN and NFV for E2E programmability
  - Benefits of “dynamic slicing”

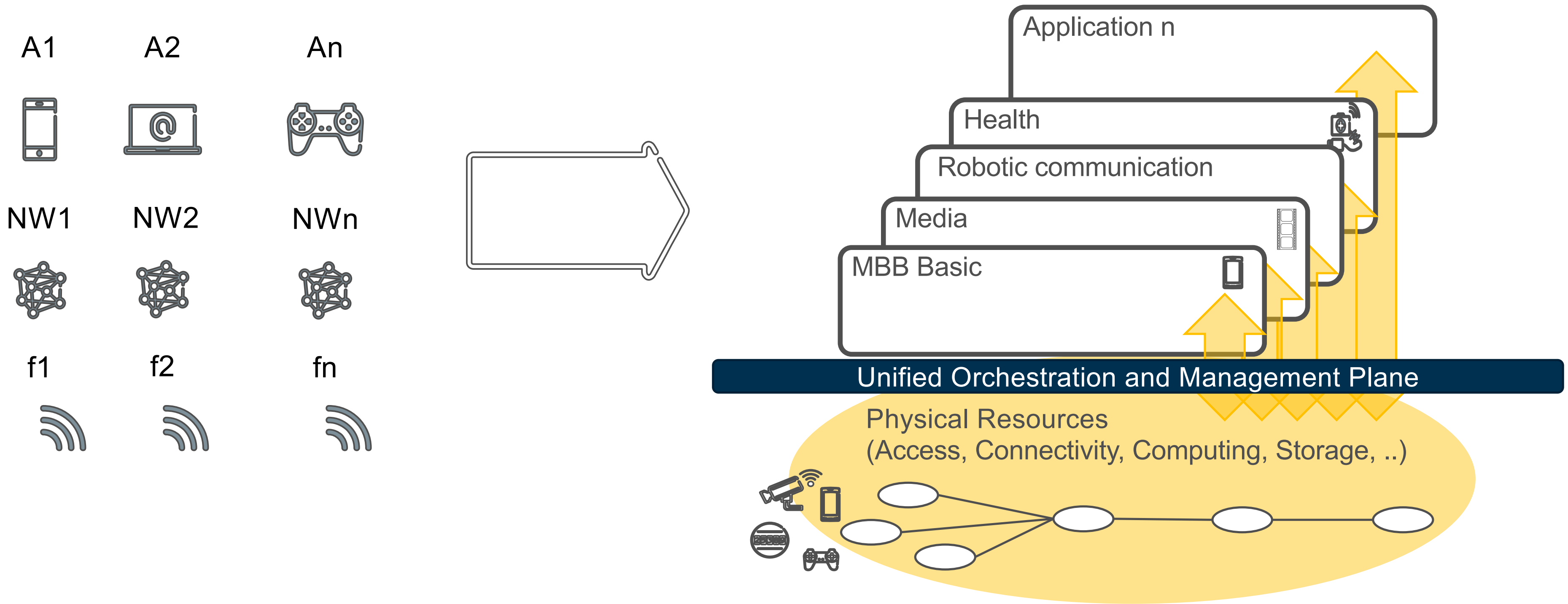
- Role of machine intelligence in network orchestration
  - Traffic prediction
  - Reinforcement Learning

- A quick look at B5G and its challenges

5G VISION: ONE NETWORK – MULTIPLE INDUSTRIES

- From dedicated physical networks with dedicated control and dedicated services/resources for different applications…

…to a "network factory" where resources and network functions are traded and provisioned: new infrastructures and services are "*manufactured by SW*"

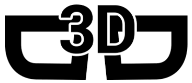


5G AND BEYOND

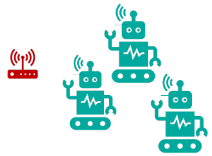
- 5G is built on a “best effort” network
    - determinism is missing
- Beyond 5G (B5G) services require much more
    - sub-ms bounded delay and delivery time-windows
    - smart and autonomous decisions
    - reliable and secure environment

EXAMPLES OF B5G SERVICES

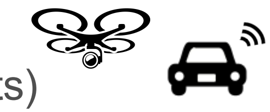
- Holographic Type Communications
  - Ultra high bandwidth
  - In-time services and with quantifiable latency



- Industrial & Robotic automation: “Industry 4.0”
  - M2M type communication
  - Require sub ms latency for closed control loops
  - Communications need to be synchronized



- Autonomic and critical infrastructures
  - Self-driving vehicles, drones, automated traffic control systems, …
  - Failsafe and autonomous infrastructure (acts and reacts rapidly to events)

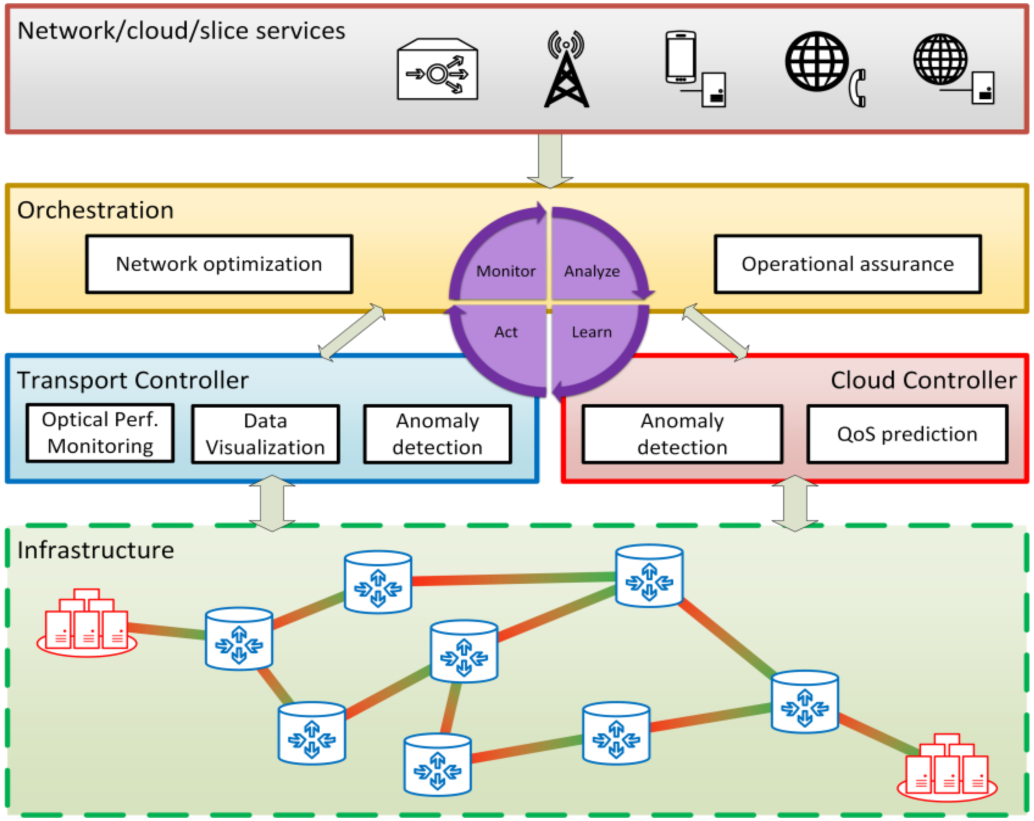


MAIN CHALLENGES FROM 5G TO B5G

- Network Service Interfaces
  - Current traffic engineering and differentiated service do not provide guarantees
  - Per-packet delivery guarantees instead of per-flow
- Automation
  - High Programmability and Agile Lifecycle (Rapid introduction of new services, adaptation to user needs)
  - SDN, NFV, slicing, programmable data plane
- Manageability
  - Higher accuracy and coverage (lower monitoring cycle time and larger number of monitoring parameters)
- Security, Resiliency, Trustworthiness
  - Reliable against failures, attacks and losses (authorization, integrity protection, encryption, backup resources, buffering)
  - Tailored for each service

NETWORK SLICE AUTOMATION

- Objective: efficient resource management (while enforcing QoS requirements), proactive reconfigurations (e.g., fault tolerance), fast setup/tear down, enabling new/better services
- Machine-learning-based policies can self-adapt, explore and learn how to operate also in the presence of previously unseen conditions
- All of these are open questions and provide a very fertile research filed for the future



SOME WORKS ON MACHINE LEARNING FOR TRANSPORT NETWORK

- L. Barletta *et al*., "QoT estimation for unestablished lighpaths using machine learning," 2017 Optical Fiber Communications Conference and Exhibition (OFC), Los Angeles, CA, 2017, pp. 1-3.
- D. Azzimonti *et al*., "Using Active Learning to Decrease Probes for QoT Estimation in Optical Networks," 2019 Optical Fiber Communications Conference and Exhibition (OFC), San Diego, CA, USA, 2019, pp. 1-3.
- F. Musumeci *et al*., "An Overview on Application of Machine Learning Techniques in Optical Networks," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383-1408, Secondquarter 2019.
- F. Musumeci *et al*., "A Tutorial on Machine Learning for Failure Management in Optical Networks," in Journal of Lightwave Technology, vol. 37, no. 16, pp. 4125-4139, 15 Aug.15, 2019.
- D. Rafique and L. Velasco, "Machine learning for network automation: overview, architecture, and applications [Invited Tutorial]," in *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 10, pp. D126-D143, Oct. 2018.
- D. M. Gutierrez-Estevez *et al*., "Artificial Intelligence for Elastic Management and Orchestration of 5G Networks," in IEEE Wireless Communications, vol. 26, no. 5, pp. 134-141, October 2019.
- …

Slicing in 5G Transport Networks

Paolo Monti

mpaolo@chalmers.se