Optimization of disaggregated 5G+ Networks considering Internal Traffic and QoS Requirements

M.Sc. Ameya G. Joshi, M.Sc. Tuan Khai Nguyen, Prof. Dr.-Ing. Thomas Bauschert
Technische Universität Chemnitz, Germany

Content

- Motivation
- Problem Statement
- Optimization Model
- Evaluation/Results
- Summary and next Steps

Motivation

Motivation

- Future disaggregated 5G+ networks will carry a variety of service types with different QoS requirements and traffic patterns:
  - Internet traffic (external, mainly best effort)
  - Local breakout traffic to mobile edge servers (internal, mainly with QoS requirements)
  - Machine-to-Machine (M2M) traffic (internal, mainly with QoS requirements)
- Cost optimization is particularly important for campus and industrial 5G+ networks

- Focus of our work:
  - Optimization of virtualized, disaggregated 5G+ networks with heterogeneous traffic and QoS requirements

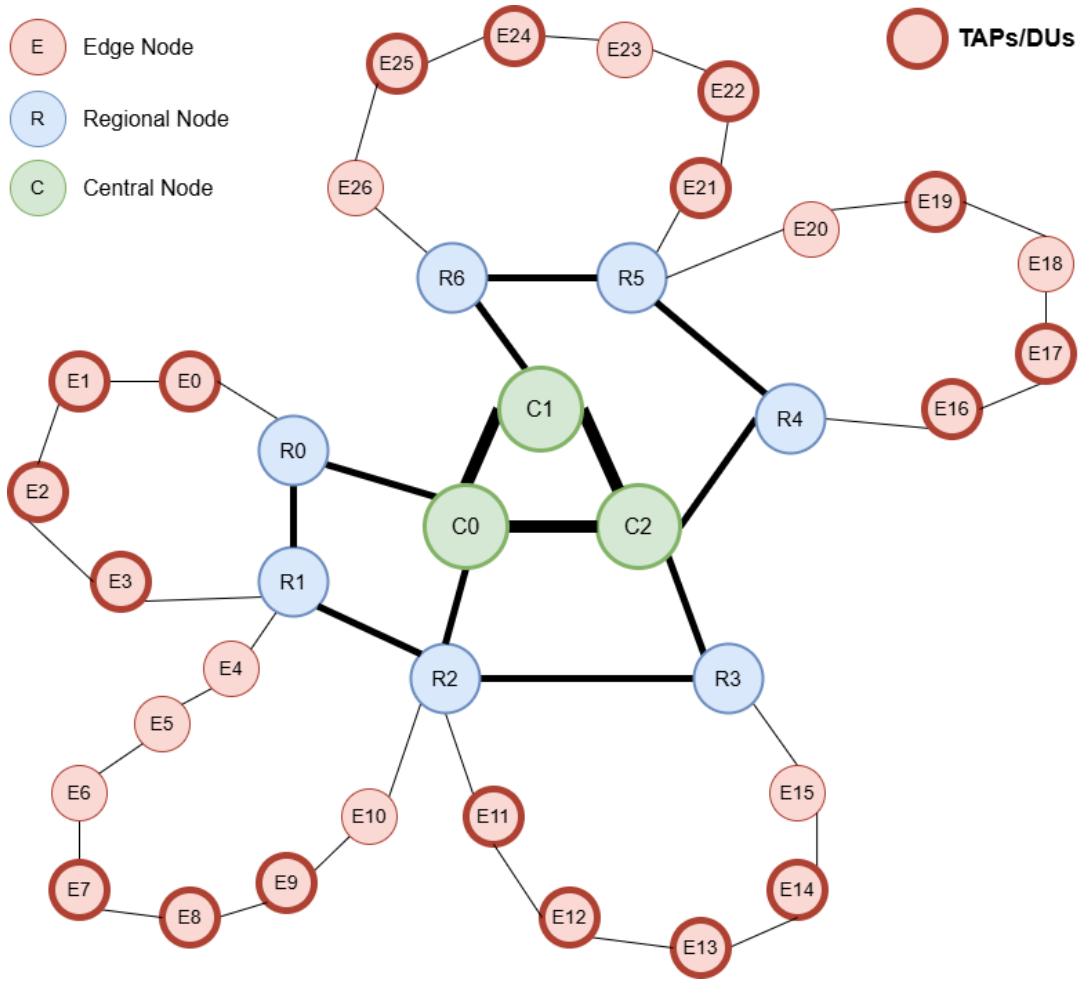
Problem Statement

Problem Statement: Optimization Problem

- Objective:

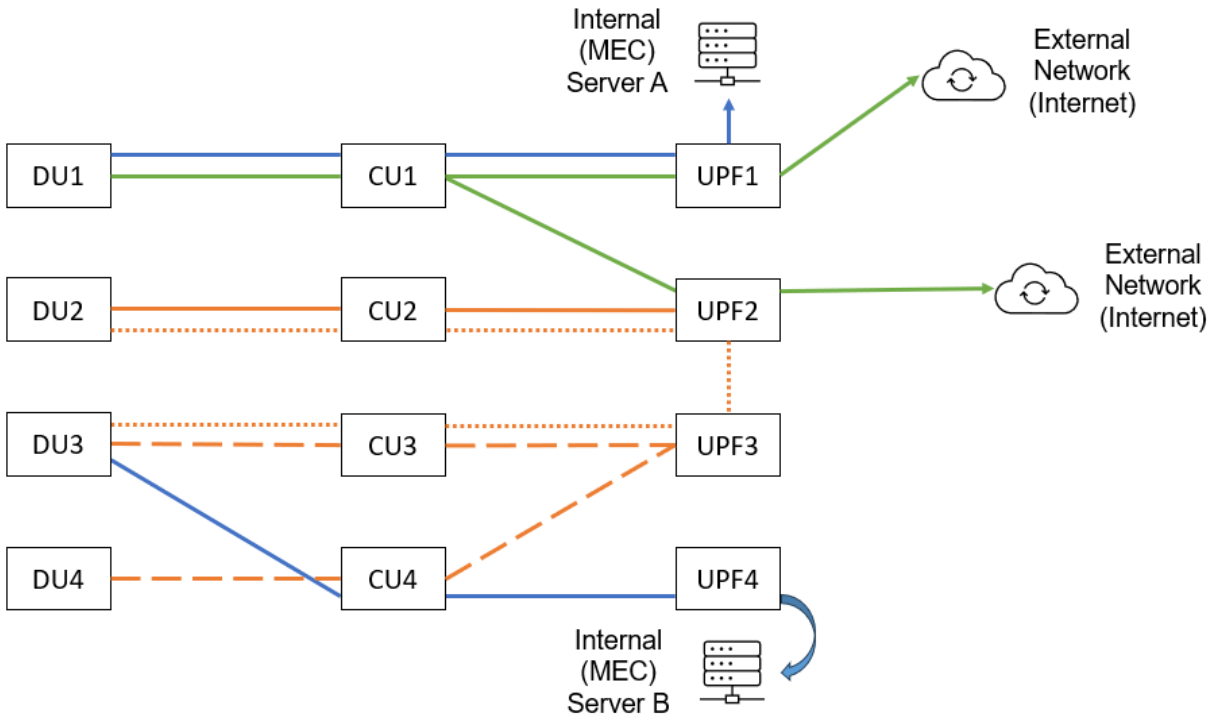
  Cost minimization of the required node (compute and switching) and link (capacity) resources of the mid- and backhaul network

- Given:
  - Mid- and backhaul network topology including maximum available node (compute and switching) and link (capacity) resources
  - Node properties: co-located mobile edge server; access to Internet
  - DU/Traffic Aggregation Point (TAP) locations
  - Traffic types with their average busy hour bitrates and QoS requirements
- Variables:
  - Virtual CU and virtual UPF location
  - Traffic routing between virtual DUs, CUs and UPFs (traffic flow splitting allowed)



Problem Statement: Traffic Model

- Traffic patterns/types*:
  - Internet traffic (external)
  - Local breakout traffic (internal)
  - M2M Traffic (internal)



- Traffic demand matrix (average bitrates in the daily busy hour):
  - Rows: Sources
  - Columns: Destinations

M2M

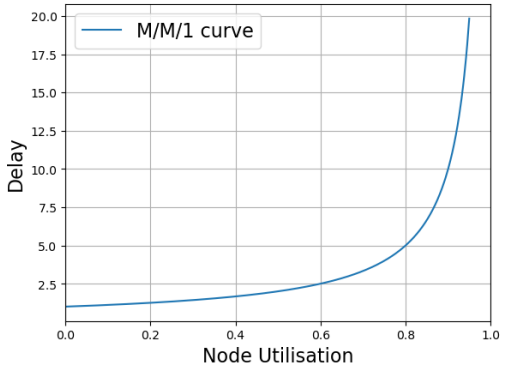
	D_1	D_2	⋯	D_n	Int	Ext
D_1	d_{11}	d_{12}	⋯	d_{1n}	d_{1i}	d_{1e}
D_2	d_{21}	d_{22}	⋯	d_{2n}	d_{2i}	d_{2e}
⋮	⋮	⋮	⋱	⋮	⋮	⋮
D_n	d_{n1}	d_{n2}	⋯	d_{nn}	d_{ni}	d_{ne}
Int	d_{i1}	d_{i2}	⋯	d_{in}	0	0
Ext	d_{e1}	d_{e2}	⋯	d_{en}	0	0

*Note: traffic flows are splittable

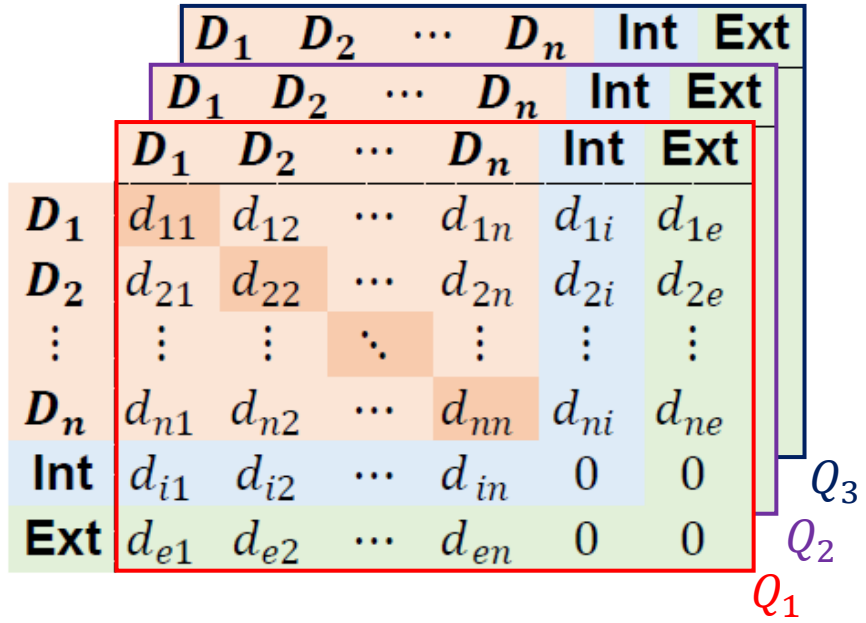
Problem Statement: QoS, Routing and Cost Model

- QoS Model: Latency

Delay Type	Model	Load dep.?
Propagation	∝ link length	No
Switching	M/M/1 Model	Yes
Processing	M/M/1 Model	Yes



Note: one traffic matrix per QoS category (i.e. required e2e latency)



- Routing Model:
  - Shortest path routing in the IP-based mid- and backhaul network
- Cost Model:
  - Cost of occupied link resources
  - Cost of node switching and processing (computation) resources
  - Upfront node costs (only if node processing resources are utilized)

Optimization Model

Optimization Model: Parameters

- $C_{q,t}$: Set of service chains for TAP/DU t and QoS category q
- g_q^{e2e}: e2e latency of category q
- N_ϖ^V : Virtual nodes of service chain $\varpi \in C_{q,t}$
- $L_\varpi^V \subset N_\varpi^V \times N_\varpi^V$: Virtual links
- e_s^v : Indicates if a particular physical node s can host VNF v

External:

$$\mathbf{DU}_{q,t}^{\mathbf{ext}} \Longleftrightarrow \mathbf{CU}_{q,t}^{\mathbf{ext}} \Longleftrightarrow \mathbf{UPF}_{q,t}^{\mathbf{ext}}$$

Internal:

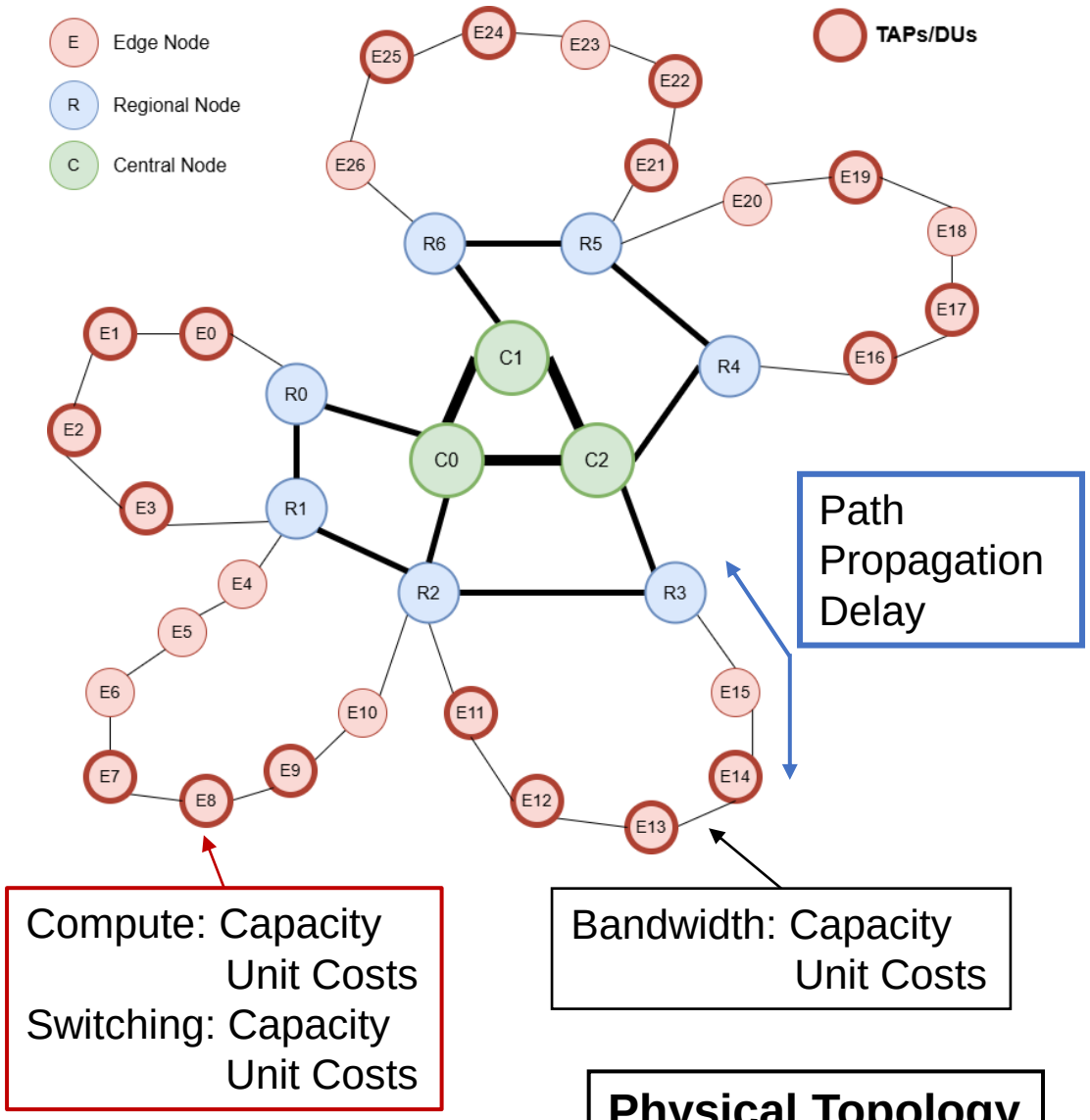
$$\mathbf{DU}_{q,t}^{\mathbf{int}} \Longleftrightarrow \mathbf{CU}_{q,t}^{\mathbf{int}} \Longleftrightarrow \mathbf{UPF}_{q,t}^{\mathbf{int}}$$

M2M:

$$\mathbf{DU}_{q,(t,t')}^{\mathbf{m2m}} \Longleftrightarrow \mathbf{CU}_{q,(t,t')}^{\mathbf{m2m}} \Longleftrightarrow \mathbf{UPF}_{q,(t,t')}^{\mathbf{m2m}}$$

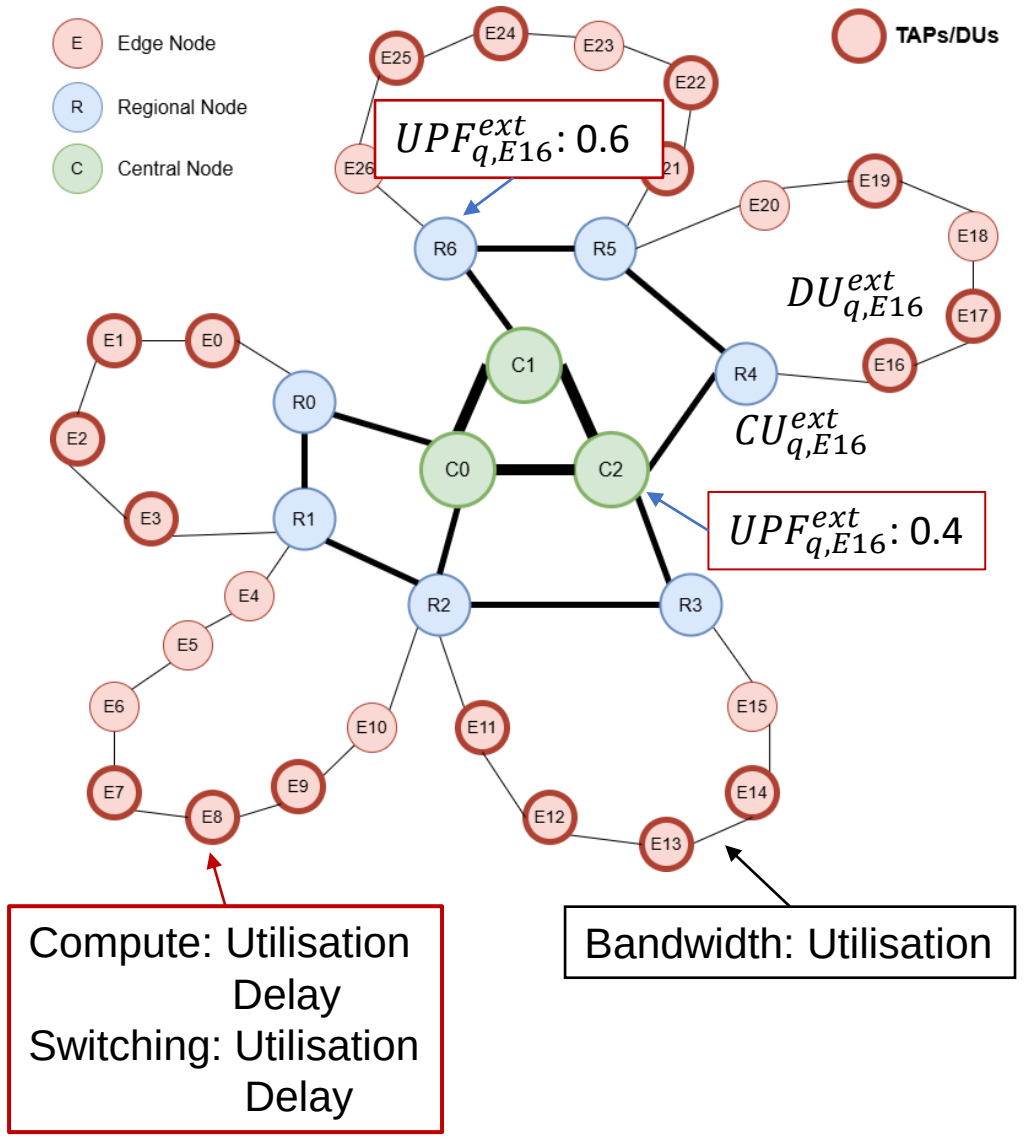
$$\Updownarrow$$

$$\mathbf{DU}_{q,(t',t)}^{\mathbf{m2m}} \Longleftrightarrow \mathbf{CU}_{q,(t',t)}^{\mathbf{m2m}} \Longleftrightarrow \mathbf{UPF}_{q,(t',t)}^{\mathbf{m2m}}$$



Optimization Model: Variables

- $\alpha_v^s \in [0,1]$: Fraction of the chain VNF $v \in N_{\varpi}^V$ hosted on physical node s
- $\theta_s \in \{0,1\}$: Indicates whether any VNF of any service chain is placed on physical node s
- $\gamma_p^{(v,v')} \in [0,1]$: Fraction of traffic between VNFs (v, v') on path p
- $\beta_p^{(v,v')} \in \{0,1\}$: Indicator variable which equals 1 if $\gamma_p^{(v,v')} > 0$, else 0
- $\tau_{(v,v')}^{pair}$: Maximum latency between a VNF pair of a given chain $\varpi \in C_{q,t}$



Optimization Model: Objective and Constraints

Objective:

$$\text{minimize} \quad a \cdot \sum_{s \in N_S} z_s^{\mathbf{act}} \cdot \theta_s + b \cdot \sum_{s \in N_S} \sum_{y \in \{\mathbf{pr},\mathbf{sw}\}} z_s^y \cdot c_s^y \cdot \mu_s^y + c \cdot \sum_{(s,s') \in L_S} z_{(s,s')}^{\mathbf{bw}} \cdot c_{(s,s')}^{\mathbf{bw}} \cdot \mu_{(s,s')}^{\mathbf{bw}}$$

Node upfront costs | Node resource costs | Link resource costs

Constraints:

- Correct placement of VNFs on suitable substrate nodes:

$$\alpha_s^v \leq \mathbb{e}_s^v, \qquad \forall s \in N_S, v \in \bigcup_{q \in Q} \bigcup_{t \in T} \bigcup_{\varpi \in C_{q,t}} N_\varpi^{\mathcal{V}}$$

$$\sum_{s \in N_S} \mathbb{e}_s^v \cdot \alpha_s^v = 1, \qquad \forall v \in \bigcup_{q \in Q} \bigcup_{t \in T} \bigcup_{\varpi \in C_{q,t}} N_\varpi^{\mathcal{V}}$$

- Node activation constraints:

$$\theta_s \geq \alpha_s^v, \qquad \forall s \in N_S, v \in \bigcup_{q \in Q} \bigcup_{t \in T} \bigcup_{\varpi \in C_{q,t}} N_\varpi^{\mathcal{V}}$$

- Flow constraints:

$$\begin{cases} \sum_{p \in P_s^{\mathbf{out}}} \gamma_p^{(v,v')} = \alpha_s^v \\ \sum_{p \in P_s^{\mathbf{in}}} \gamma_p^{(v,v')} = \alpha_s^{v'}, \end{cases} \qquad \forall s \in N_S, (v,v') \in \bigcup_{q \in Q} \bigcup_{t \in T} \bigcup_{\varpi \in C_{q,t}} L_\varpi^{\mathcal{V}}$$

Optimization Model: Objective and Constraints

- M/M/1 piecewise linearization constraints:

$$\tau_s^y \geq u_i^y + w_i^y \cdot \mu_s^y, \qquad \forall s \in N_{\mathcal{S}}, i \in \{1, \ldots, m\}, y \in \{\mathbf{pr}, \mathbf{sw}\}$$

- Latency constraints:

$$\beta_p^{(v,v')} \geq \gamma_p^{(v,v')}, \qquad \forall p \in P, (v,v') \in \bigcup_{q \in Q} \bigcup_{t \in T} \bigcup_{\varpi \in C_{q,t}} L_\varpi^{\mathcal{V}}$$

Maximum per VNF pair latency →

$$\tau_{(v,v')}^{\mathbf{pair}} \geq \tau_p^{\mathbf{prop}} + \tau_{s_p^{\mathbf{end}}}^{\mathbf{pr}} + \sum_{s \in N_{\mathcal{S}}} \mathbb{i}_p^s \cdot \tau_s^{\mathbf{sw}} + \mathcal{M} \cdot \left(\beta_p^{(v,v')} - 1\right), \qquad \forall p \in P, (v,v') \in \bigcup_{q \in Q} \bigcup_{t \in T} \bigcup_{\varpi \in C_{q,t}} L_\varpi^{\mathcal{V}}$$

Total e2e latency →

$$\sum_{(v,v') \in L_\varpi^{\mathcal{V}}} \tau_{(v,v')}^{\mathbf{pair}} \leq g_q^{\mathbf{e2e}}, \qquad \forall q \in Q, \varpi \in \bigcup_{t \in T} C_{q,t}$$

- Resource requirement constraints:

Node processing →

$$\sum_{q \in Q} \sum_{t \in T} \sum_{\varpi \in C_{q,t}} \sum_{v \in N_\varpi^{\mathcal{V}}} d_v^{\mathbf{pr}} \cdot \alpha_s^v = c_s^{\mathbf{pr}} \cdot \mu_s^{\mathbf{pr}}, \qquad \forall s \in N_{\mathcal{S}}$$

Node switching →

$$\sum_{p \in P} \mathbb{i}_p^s \sum_{q \in Q} \sum_{t \in T} \sum_{\varpi \in C_{q,t}} \sum_{(v,v') \in L_\varpi^{\mathcal{V}}} d_\varpi^{\mathbf{bw}} \cdot \gamma_p^{(v,v')} = c_s^{\mathbf{sw}} \cdot \mu_s^{\mathbf{sw}}, \qquad \forall s \in N_{\mathcal{S}}$$

Link Bandwidth →

$$\sum_{p \in P} \mathbb{j}_p^{(s,s')} \sum_{q \in Q} \sum_{t \in T} \sum_{\varpi \in C_{q,t}} \sum_{(v,v') \in L_\varpi^{\mathcal{V}}} d_\varpi^{\mathbf{bw}} \cdot \gamma_p^{(v,v')} = c_{(s,s')}^{\mathbf{bw}} \cdot \mu_{(s,s')}^{\mathbf{bw}}, \qquad \forall (s,s') \in L_{\mathcal{S}}$$

Evaluation and Results

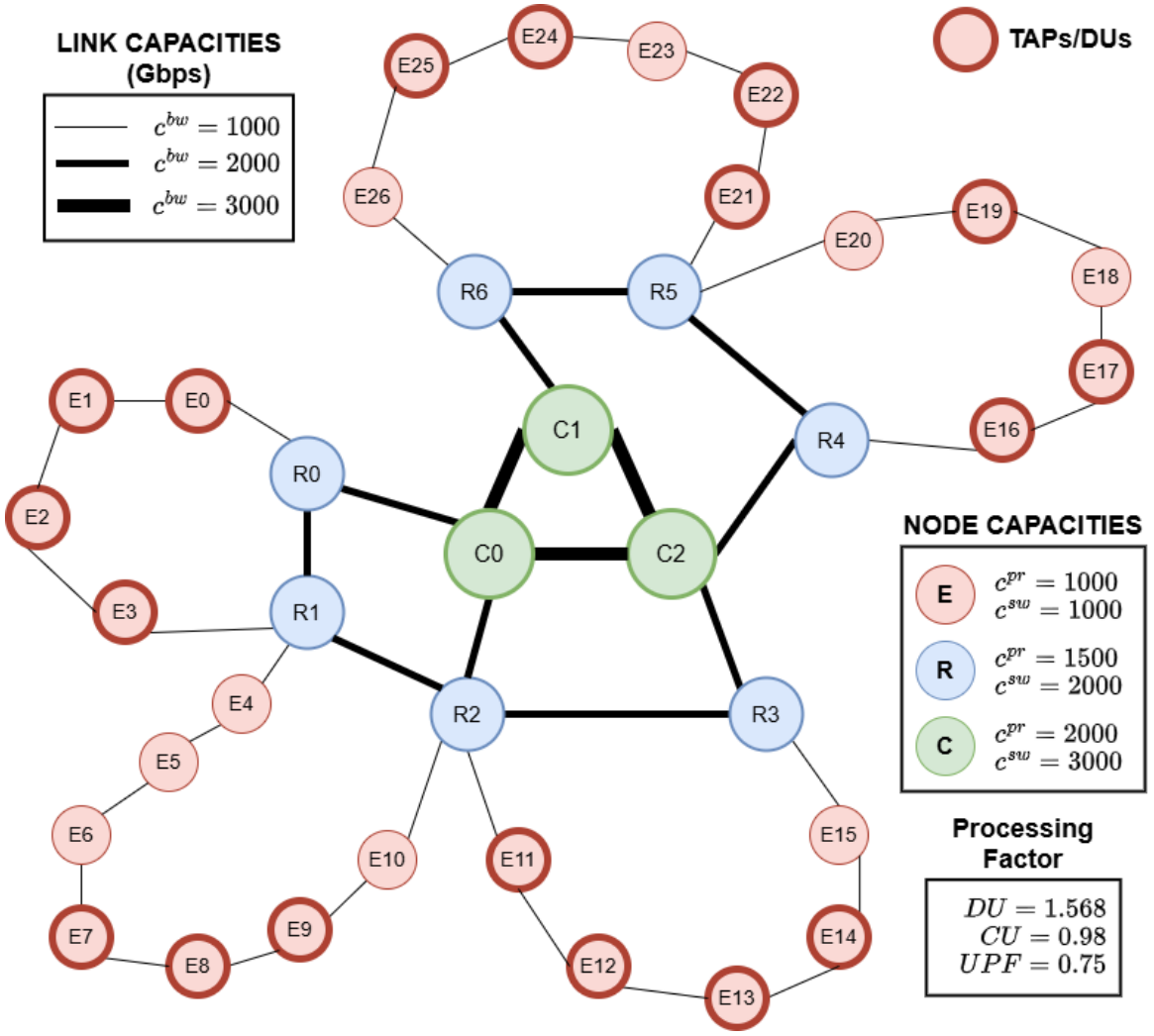
Evaluation Scenario

- Three tier network topology (see figure)
  - 27 Edge, 7 Regional and 3 Core nodes
  - 18 randomly pre-determined DUs/TAPs
- QoS categories (e2e latency in ms):

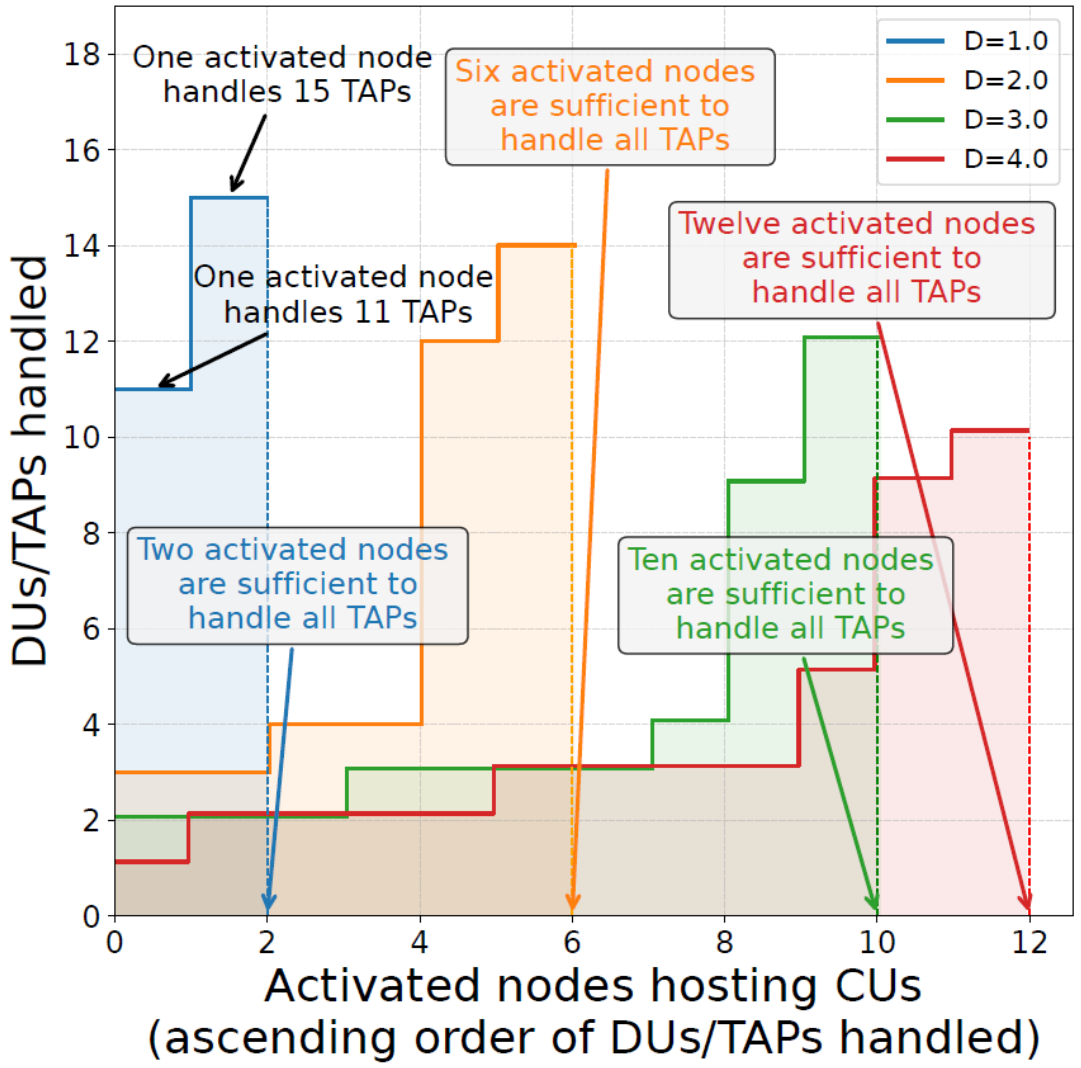
	Strict	Moderate	Relaxed
g_q^{e2e}	5	20	50

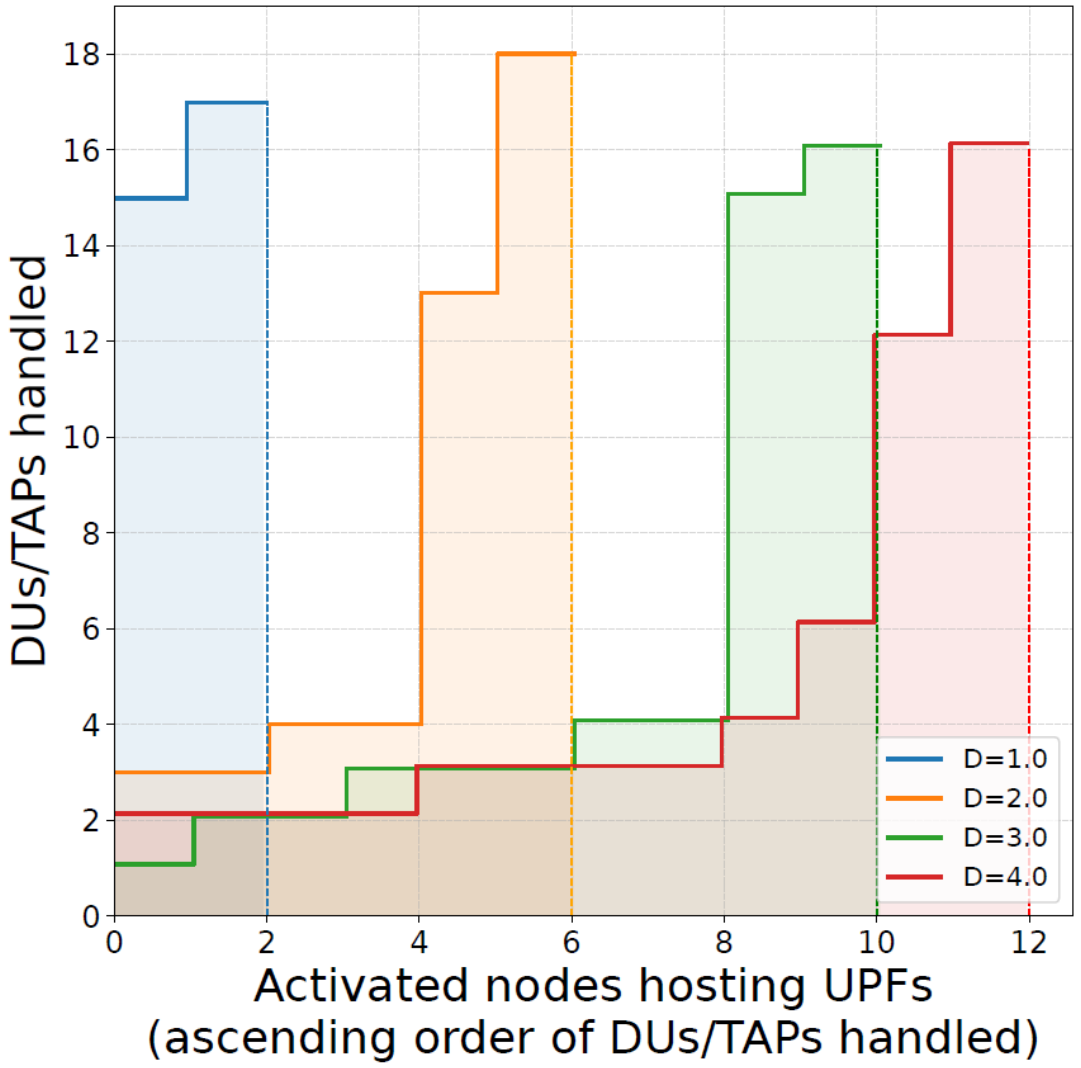
- Overall traffic demand (per QoS category):

	Strict	Moderate	Relaxed
External/Internet (per DU)	-	-	[50,60]
Internal/Local MEC (per DU)	[2,4]	[2,4]	[2,4]
Internal/M2M (per DU pair)	[1,2]	[1,2]	[1,2]

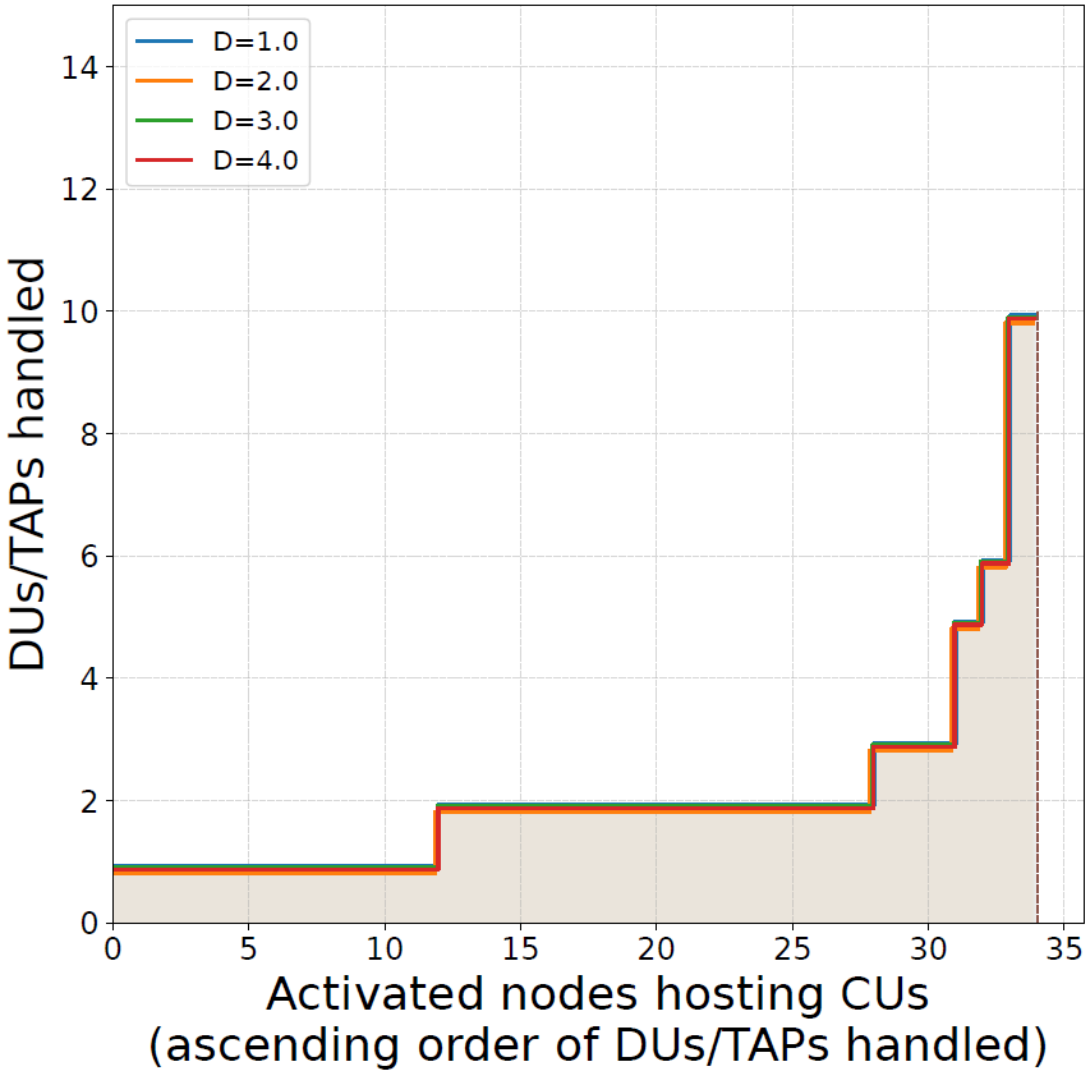


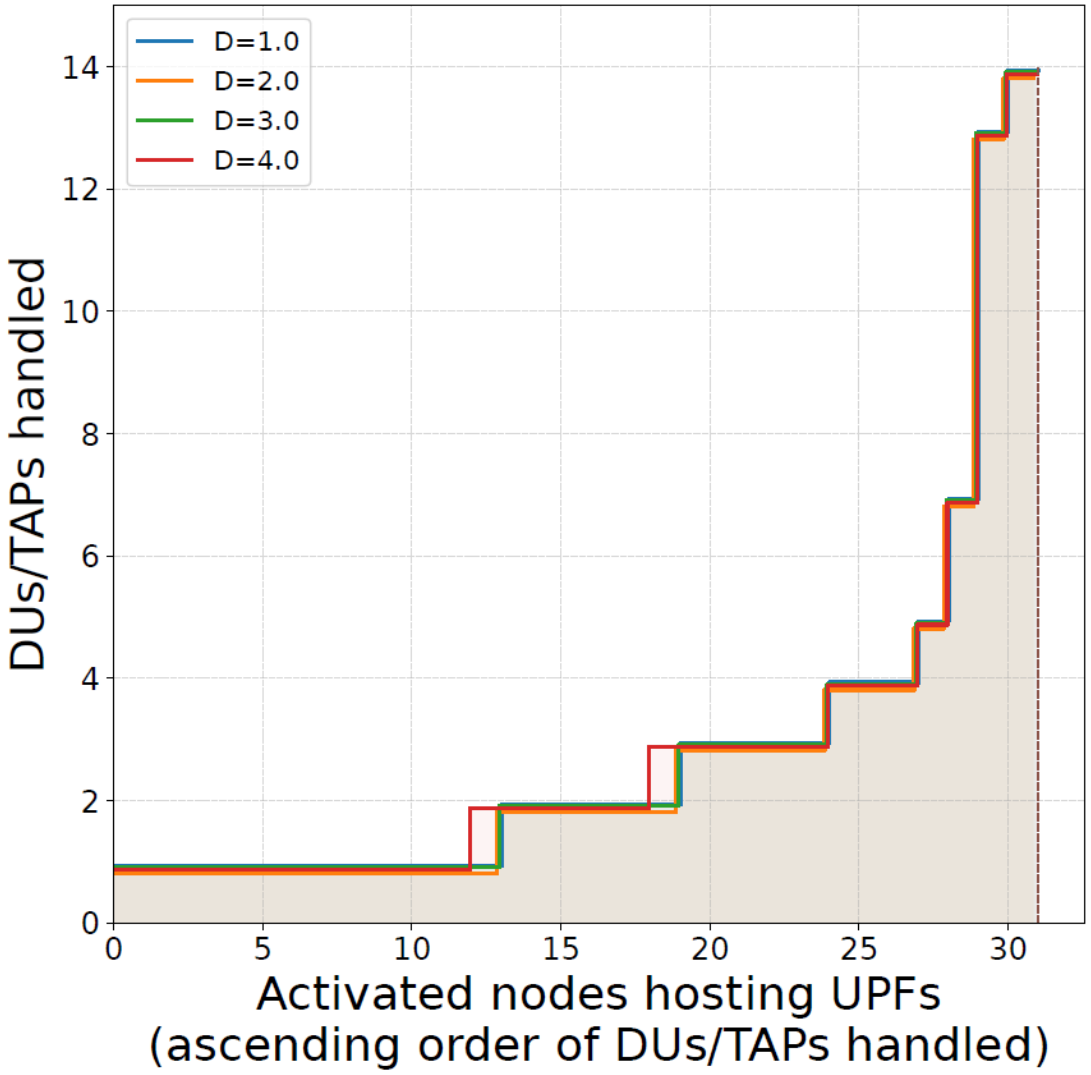
Results: DUs served per activated Node vs. Traffic Demand (w. Upfront Costs)



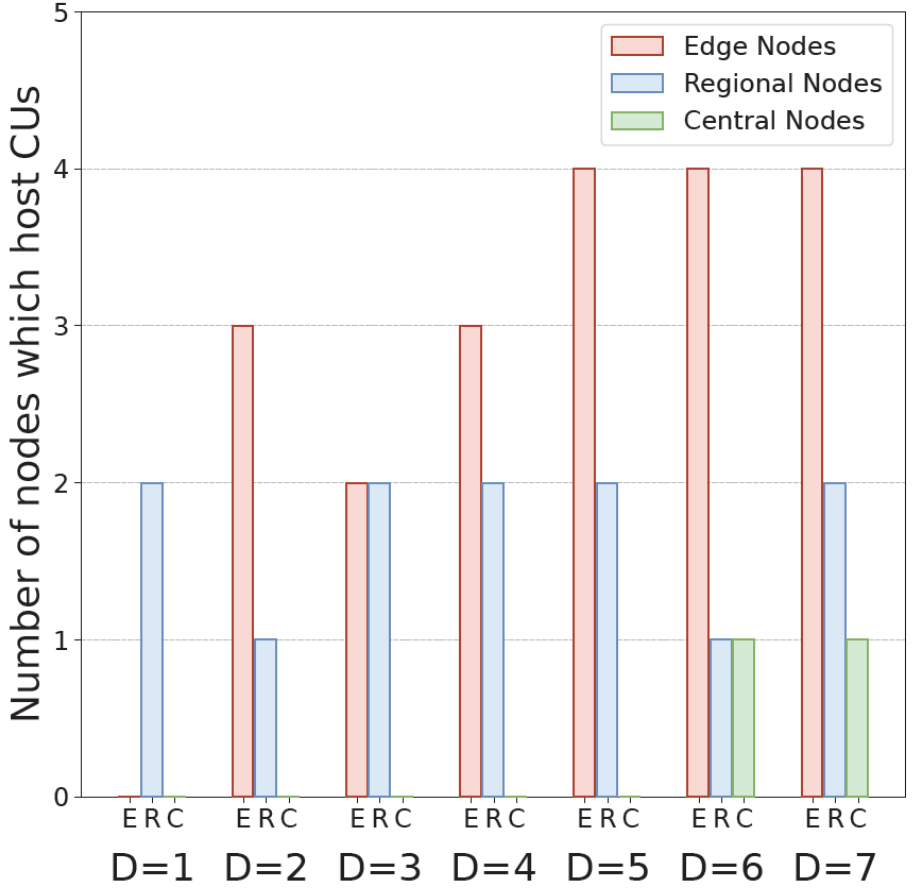


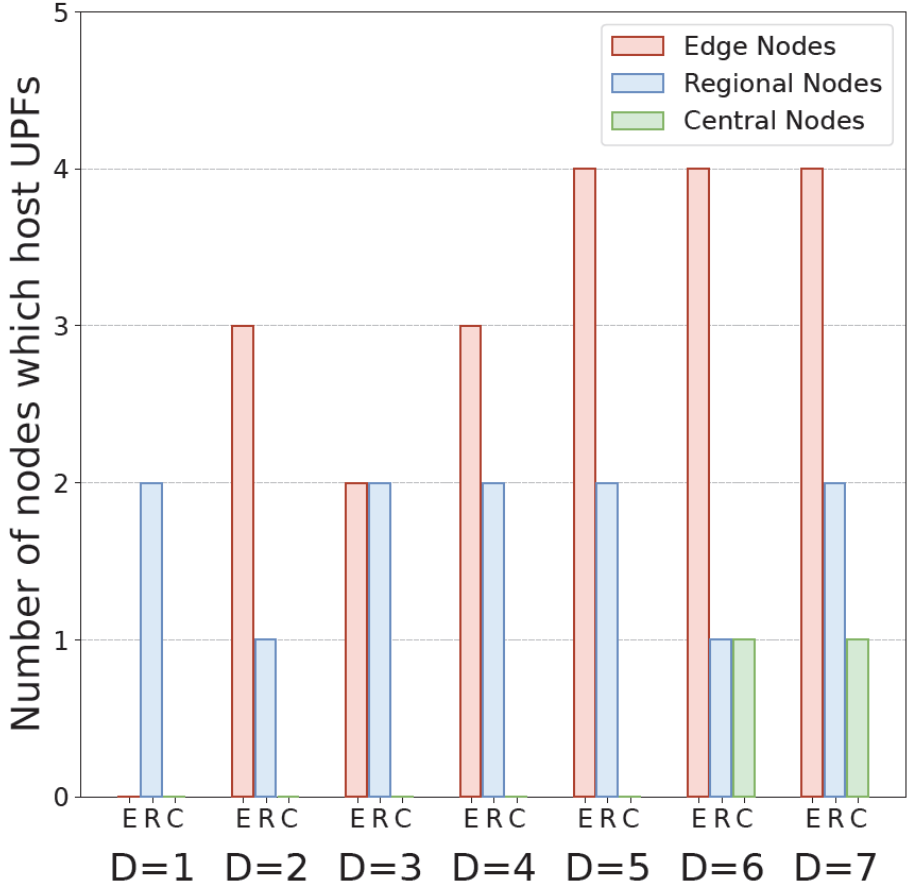
Results: DUs served per activated Node vs. Traffic Demand (w/o Upfront Costs)

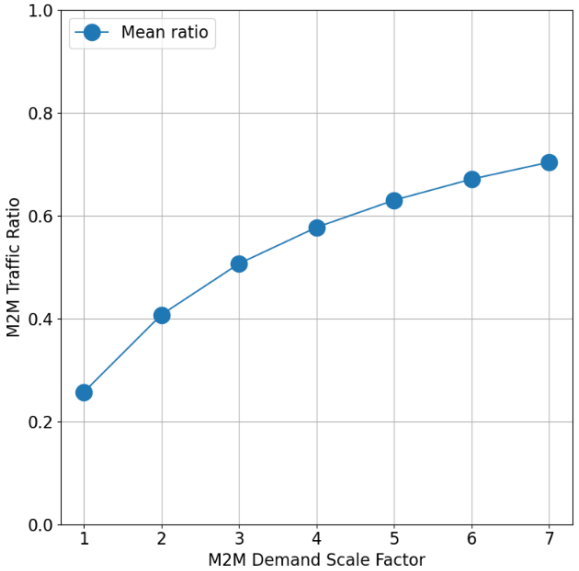




Results: Number of Nodes hosting CUs and UPFs vs. Traffic Demand







M2M traffic proportion of total traffic with increasing M2M demand

Summary and next Steps

Summary and next Steps

- Presentation of an MILP model for cost optimization of disaggregated 5G+ networks
- Consideration of different traffic types with different QoS and traffic requirements with an emphasis on internal traffic (local breakout and M2M traffic) and traffic flow splitting
- Evaluation of the impact of traffic demand scaling and upfront costs on the optimization results

- Next Steps:
  - Evaluation on further network and traffic scenarios
  - Model extension to consider robustness w.r.t. traffic demand variability and failures
  - Development of heuristic and/or DRL&GNN-based solution approaches for large scale network optimization

References

- W. Attaoui, E. Sabir, H. Elbiaze, and M. Guizani, “Vnf and cnf placement in 5g: Recent advances and future trends,” IEEE Transactions on Network and Service Management, vol. 20, no. 4, pp. 4698–4733, 2023.
- A. Baumgartner, V. S. Reddy, and T. Bauschert, “Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization,” in Proceedings of the 2015 1st IEEE conference on Network Softwarization (NetSoft). IEEE, 2015, pp. 1–9.
- ——, “Combined virtual mobile core network function placement and topology optimization with latency bounds,” in 2015 Fourth European Workshop on Software Defined Networks. IEEE, 2015, pp. 97–102.
- F. Z. Morais, G. M. F. de Almeida, L. Pinto, K. V. Cardoso, L. M. Contreras, R. da Rosa Righi, and C. B. Both, “Placeran: Optimal placement of virtualized network functions in beyond 5g radio access networks,” IEEE Transactions on Mobile Computing, vol. 22, no. 9, pp. 5434–5448, 2022.
- A. Varasteh, B. Madiwalar, A. Van Bemten, W. Kellerer, and C. Mas- Machuca, “Holu: Power-aware and delay-constrained vnf placement and chaining,” IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1524–1539, 2021.
- M. A. T. Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahmoodi, and B. H. Khalaj, “vspace: Vnf simultaneous placement, admission control and embedding,” IEEE Journal on Selected Areas in Communications, vol. 36, no. 3, pp. 542–557, 2018.

Thank You!