uc3m | Universidad Carlos III de Madrid

NFV Orchestration in Edge and Fog Scenarios

26th October, 2021

student: J. Martín-Pérez
supervisor: C. J. Bernardos

contact: jmartinp@it.uc3m.es

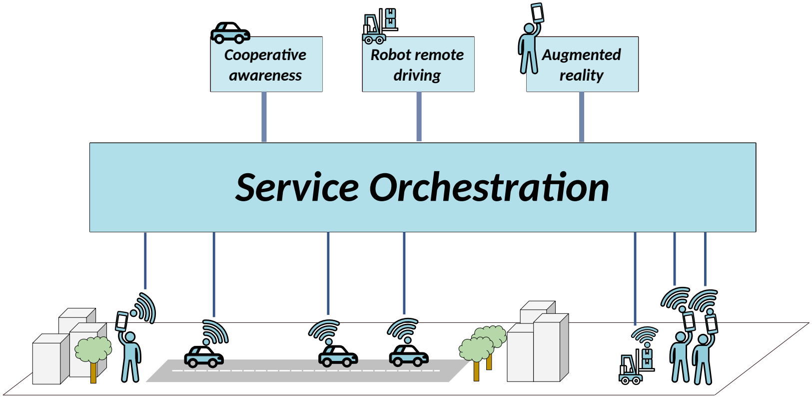


Figure 1: Orchestration of three services.

Service Orchestration:

- design network
- where services run?
- deliver to users
- monitor/scale service

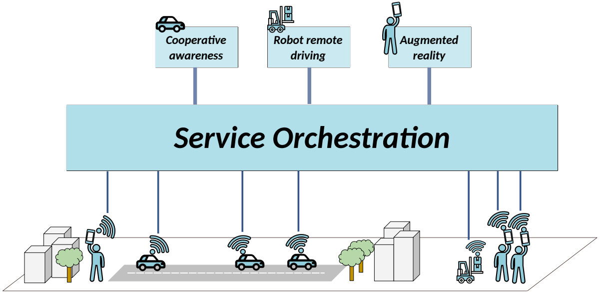


Figure 1: Orchestration of three services.

Outline

uc3m

Outline

uc3m

1. Generation of 5G infrastructure graphs
2. NFV Orchestration in federated environments
3. NFV orchestration for 5G networks: OKpi
4. Scaling of V2N services: a study case
5. Conclusions & future work

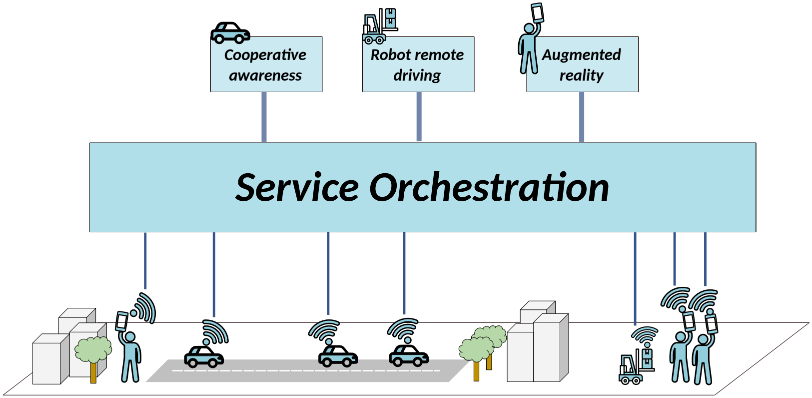


Figure 2: Orchestration of three services.

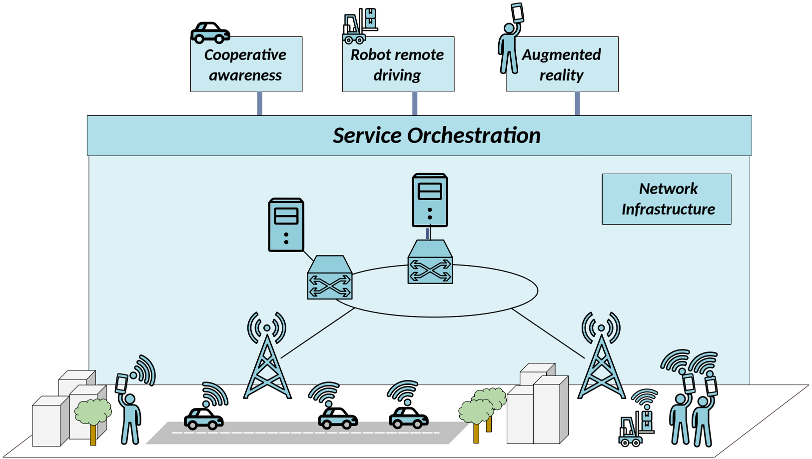


Figure 2: Network infrastructure for service providing.

uc3m

This part derives **location** of:

1. **BSs** for user coverage
2. **servers** to process traffic

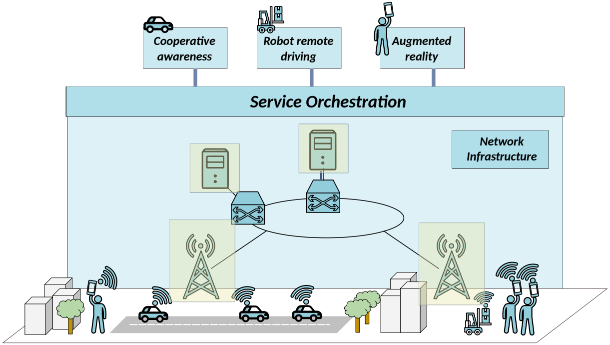


Figure 3: Network infrastructure for service providing.

This part derives **location** of:

1. **BSs** for user coverage
2. **servers** to process traffic

for **augmented reality**:

- tactile latency 1ms

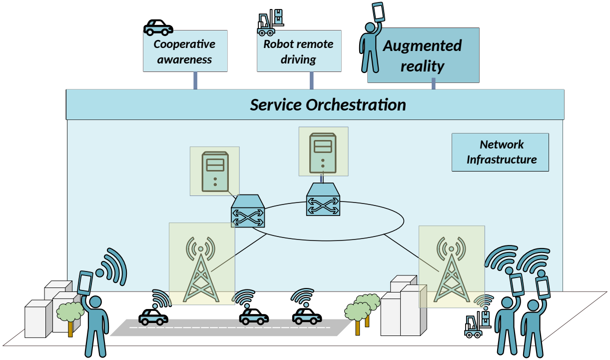


Figure 3: Network infrastructure for service providing.

New methodology in the SoA

- BS location:
  - inhomogeneous Mattérn II PPP
- Server location:
  - population census
  - access & aggregation rings
  - satisfy tactile latency 1ms

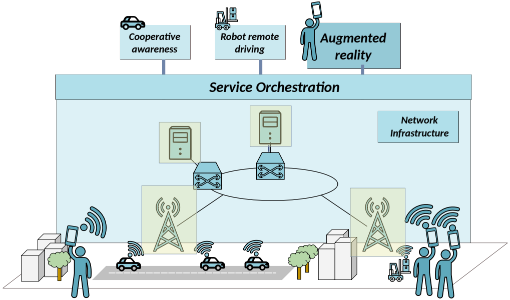


Figure 4: Network infrastructure for service providing.

Outline

uc3m

Higher gentrification $\implies$ more BSs

- R – region of interest
- C_i – area

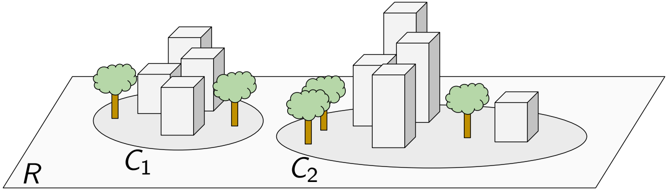


Figure 5: *Revolution functions* of a region with two building areas.

Higher gentrification $\implies$ more BSs

- R – region of interest
- C_i – area
- $f_i(x)$ – revolution func.

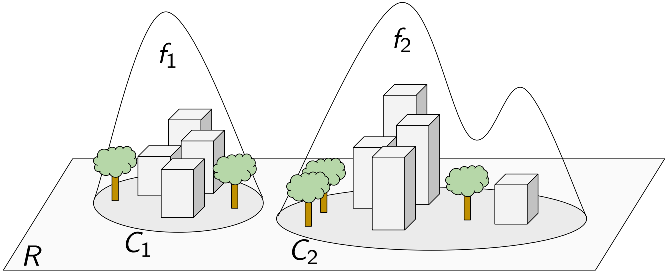


Figure 5: *Revolution functions* of a region with two building areas.

Higher gentrification $\Longrightarrow$ more BSs

- R – region of interest
- C_i – area
- $f_i(x)$ – revolution func.
- $G(x)$ – gentrification
  - $G(x) = \sum_i f_i(x)$

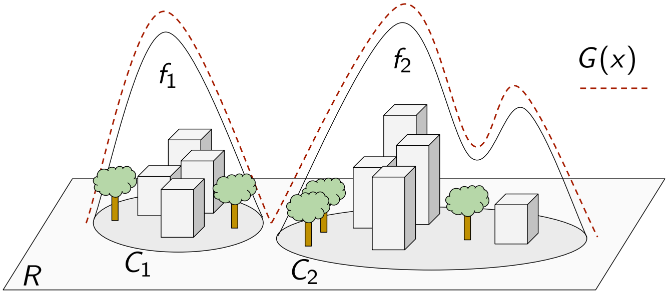


Figure 5: *Revolution functions* of a region with two building areas.

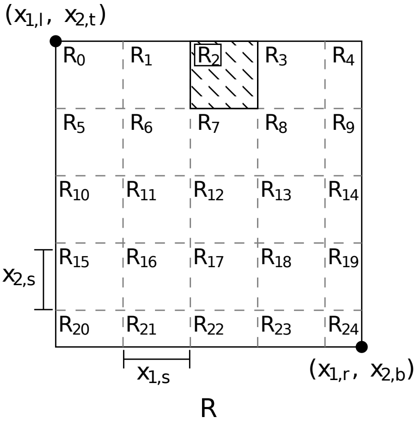


Figure 6: BS location – inhomogeneous Mattérn II process.

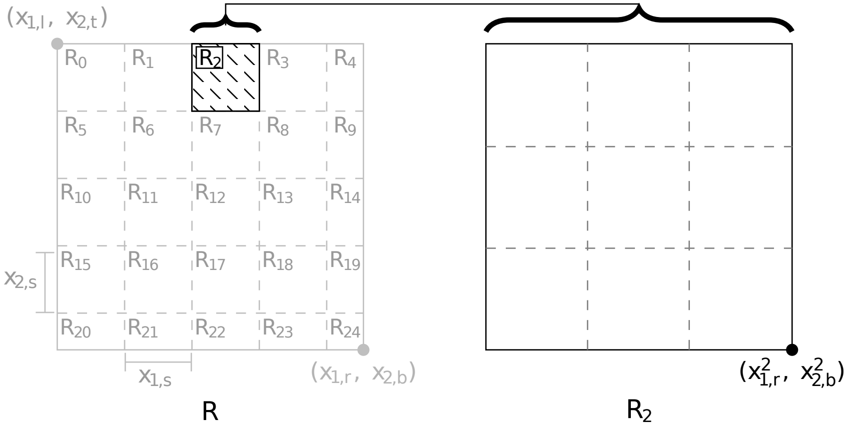


Figure 6: BS location – inhomogeneous Mattérn II process.

$\lambda(x) \sim G(x)$ probability of BS at x.

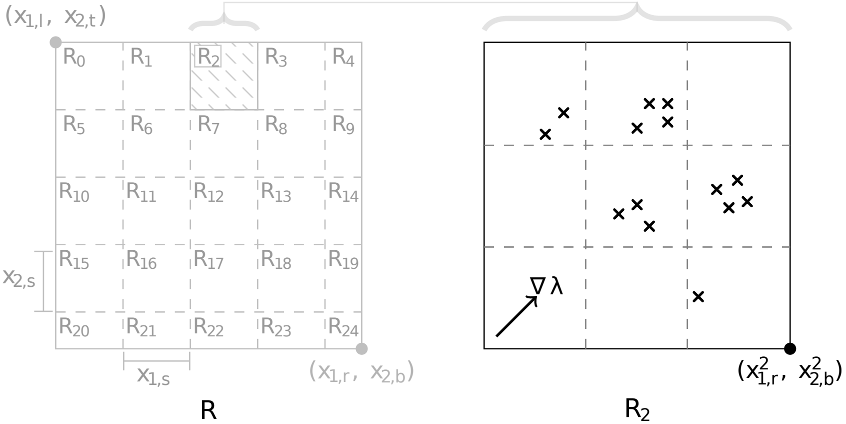


Figure 6: BS location – inhomogeneous Mattérn II process.

$\lambda(x) \sim G(x)$ probability of BS at x.

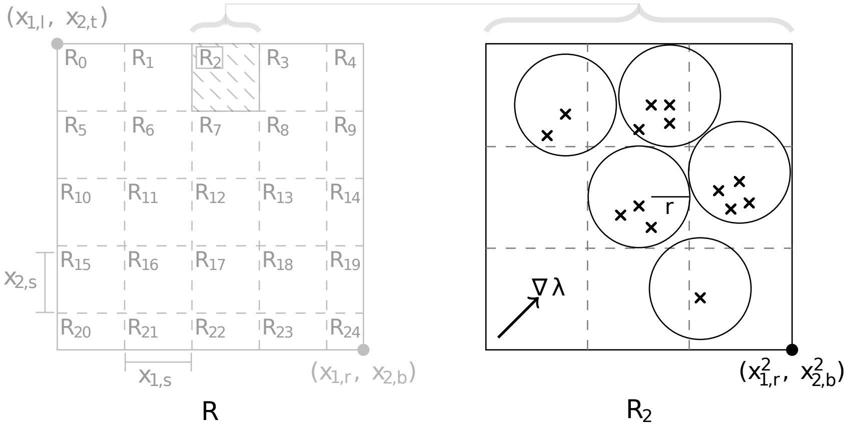


Figure 6: BS location – inhomogeneous Mattérn II process.

$\lambda(x) \sim G(x)$ probability of BS at x.

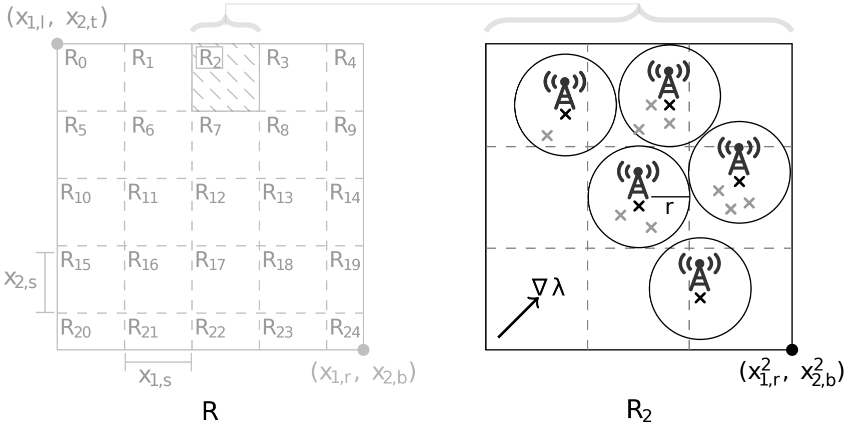


Figure 6: BS location – inhomogeneous Mattérn II process.

uc3m

Inhomogeneous Mattérn II PPs applied on:

- R: Madrid city
- $G(x)$: Madrid census
- satisfy 1ms tactile latency

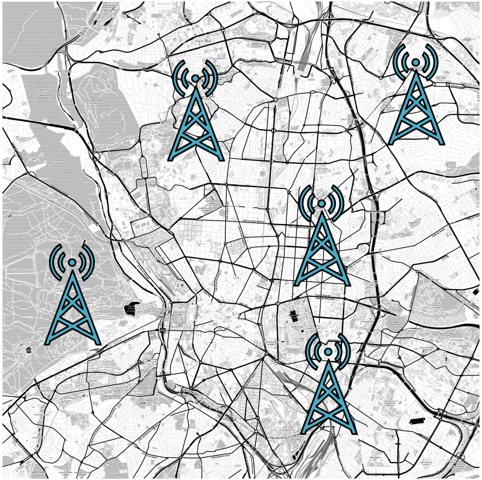


Figure 7: Location of BSs.

Inhomogeneous Mattérn II PPs applied on:

- R: Madrid city
- $G(x)$: Madrid census
- satisfy 1ms tactile latency

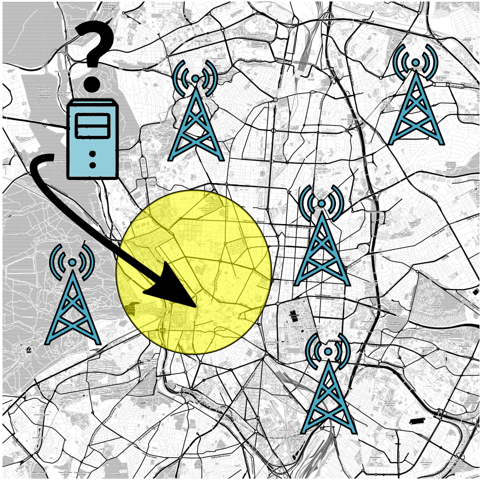


Figure 7: Location of BSs.

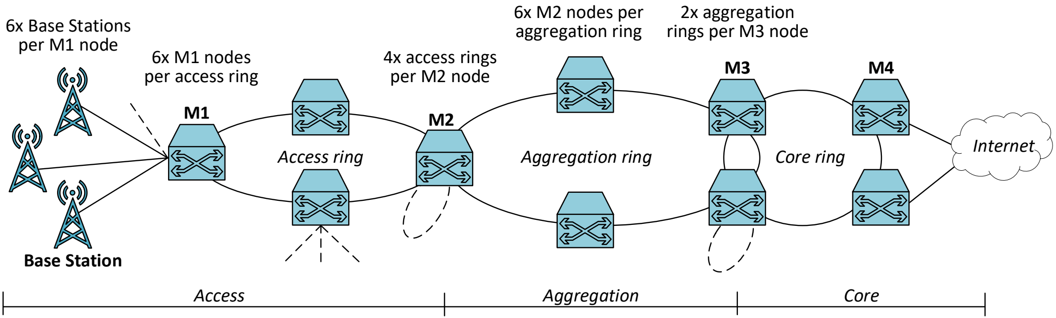


Figure 8: Reference network infrastructure as illustrated[1] in [6] and based on [10].

[1] Author: Dr. Luca Cominardi.

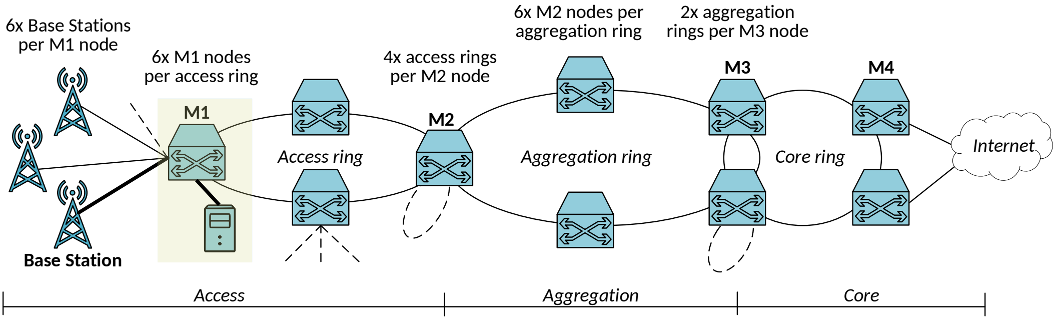


Figure 8: Reference network infrastructure as illustrated[1] in [6] and based on [10].

[1]Author: Dr. Luca Cominardi.

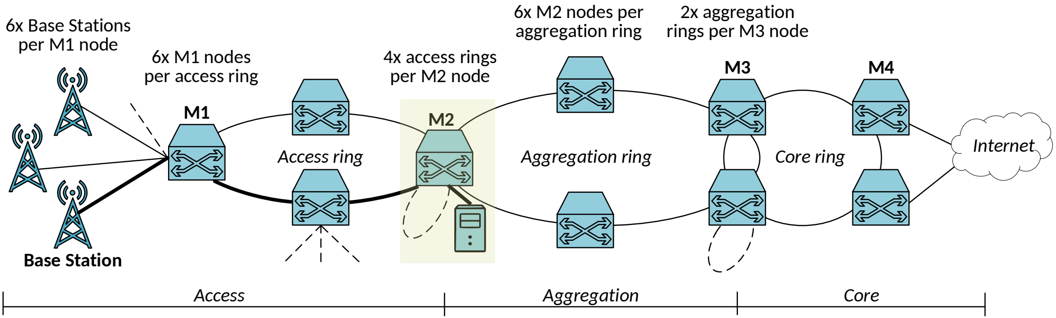


Figure 8: Reference network infrastructure as illustrated[1] in [6] and based on [10].

[1] Author: Dr. Luca Cominardi.

Derive server location s.t. $RTT \leq 1$ms (tactile latency – **augmented reality**):

$$RTT = 2d \cdot 5\frac{\mu s}{km} + 2M \cdot 50\mu s + UL + DL \tag{1}$$

- d: distance between BS and server
- M: #traversed rings (e.g., $1, 2, \ldots$)
- UL: Uplink propagation latency
- DL: Downlink propagation latency

Derive server location s.t. $RTT \leq 1\text{ms}$ (tactile latency – **augmented reality**):

$$RTT = \underbrace{2d \cdot 5\frac{\mu s}{km}}_{\text{fiber propagation}} + 2M \cdot 50\mu s + UL + DL \tag{1}$$

- d: distance between BS and server
- M: #traversed rings (e.g., $1, 2, \ldots$)
- UL: Uplink propagation latency
- DL: Downlink propagation latency

Derive server location s.t. $RTT \leq 1\text{ms}$ (tactile latency – **augmented reality**):

$$RTT = 2d \cdot 5\frac{\mu s}{km} + \underbrace{2M \cdot 50\mu s}_{\text{ring propagation}} + UL + DL \tag{1}$$

- d: distance between BS and server
- M: #traversed rings (e.g., $1, 2, \ldots$)
- UL: Uplink propagation latency
- DL: Downlink propagation latency

Derive server location s.t. $RTT \leq 1$ms (tactile latency – **augmented reality**):

$$RTT = 2d \cdot 5\frac{\mu s}{km} + 2M \cdot 50\mu s + \underbrace{UL + DL}_{\text{radio propagation}} \tag{1}$$

- d: distance between BS and server
- M: #traversed rings (e.g., $1, 2, \ldots$)
- UL: Uplink propagation latency
- DL: Downlink propagation latency

m_M: maximum distance between server at ring M and BS

Figure 9: How to select MEC PoP location.

m_2: maximum distance between server at ring 2 and BS

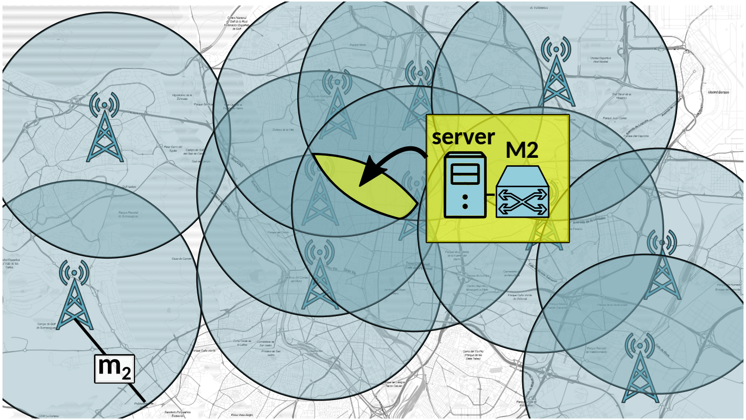
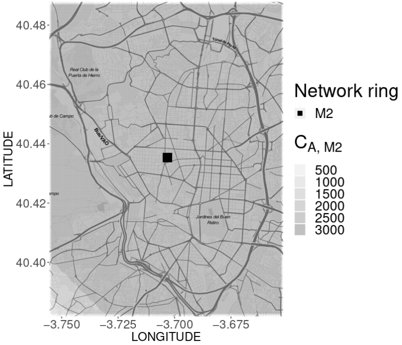


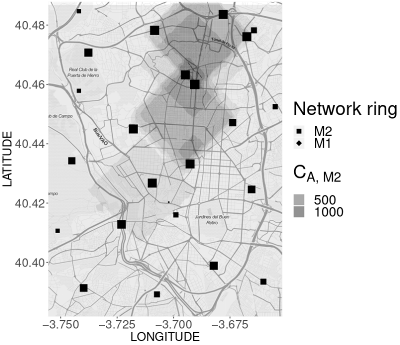
Figure 9: How to select MEC PoP location.

Experiments

- Urban, highway, industrial **scenarios**
- NR **BSs** (different UL+DL):
  - FDD 120 kHz 7s
  - TDD 120 kHz 7s
  - FDD 30 kHz 2s
- **Servers**:
  - M1 switch – access ring
  - M2 switch – aggregation ring
- **Meet**: tactile latencies 1ms



(a) FDD 120 kHz 7s



(b) TDD 120 kHz 7s

Figure 10: **Urban scenario** (Madrid city center) – $C_{A,M2}$ =covered BSs by server at M2.

uc3m

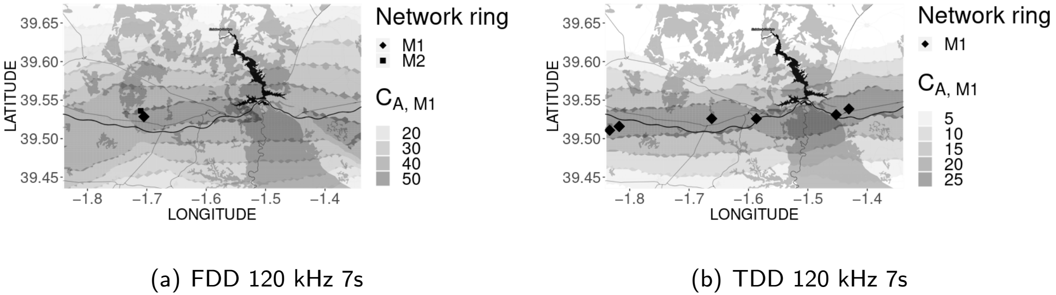


Figure 11: **Highway scenario** (Hoces del Cabriel A3) – $C_{A,M1}$ =covered BSs by server at M1.

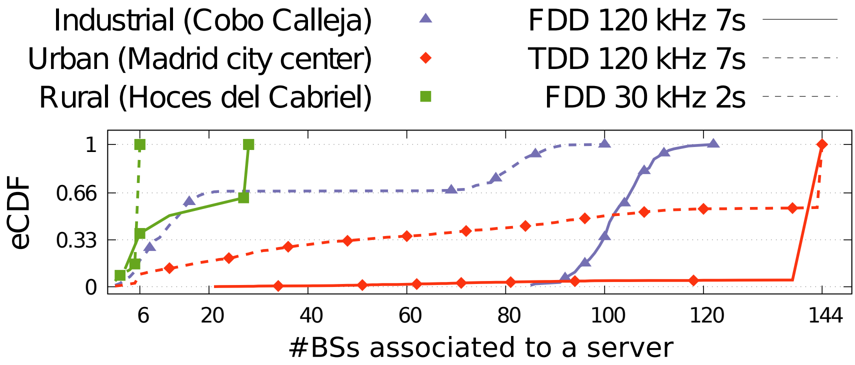


Figure 12: eCDF of the number of BSs assigned to a server in the studied scenarios.

Publications:

- **Martín-Pérez, Jorge**, L. Cominardi, C. J. Bernardos, A. de la Oliva, and A. Azcorra. "Modeling Mobile Edge Computing Deployments for Low Latency Multimedia Services". In: *IEEE Transactions on Broadcasting* 65.2 (2019), pp. 464–474. DOI: 10.1109/TBC.2019.2901406
- **Martín-Pérez, Jorge**, L. Cominardi, C. J. Bernardos, and A. Mourad. "5GEN: A tool to generate 5G infrastructure graphs". In: *2019 IEEE Conference on Standards for Communications and Networking (CSCN)*. 2019, pp. 1–4. DOI: 10.1109/CSCN.2019.8931334

Open-source:

- **BS & server generation**: `http://github.com/MartinPJorge/mec-generator/`
- **5GEN**: `https://github.com/MartinPJorge/mec-generator/tree/5g-infra-gen/`

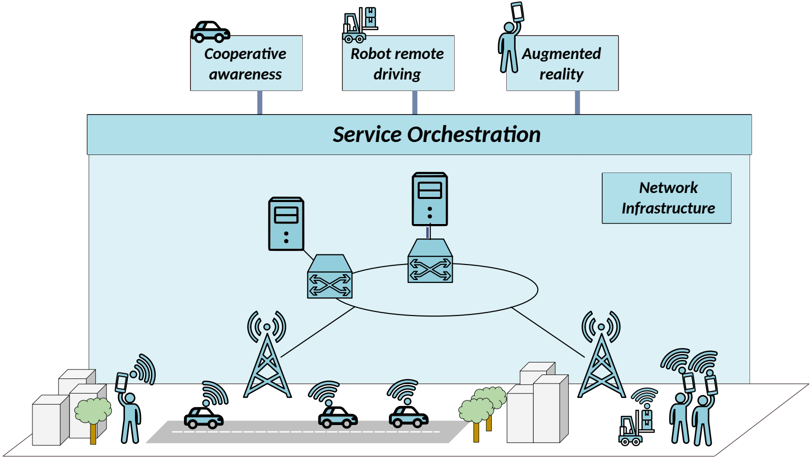


Figure 13: Infrastructure design.

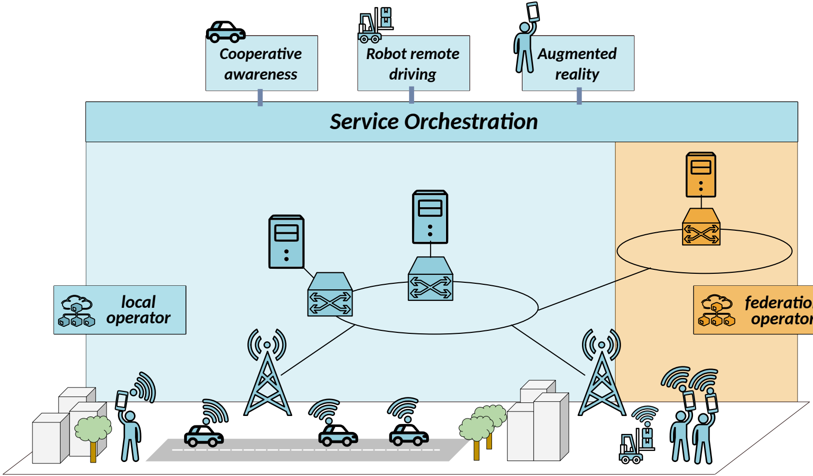


Figure 13: Local and federation operator.

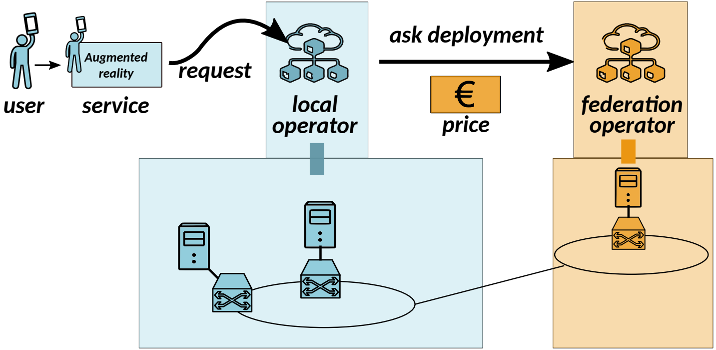


Figure 14: Service federation.

uc3m

New in SoA:

- dynamic pricing
- real-price traces AWS
- Deep Q-learning
- Telefónica scenario

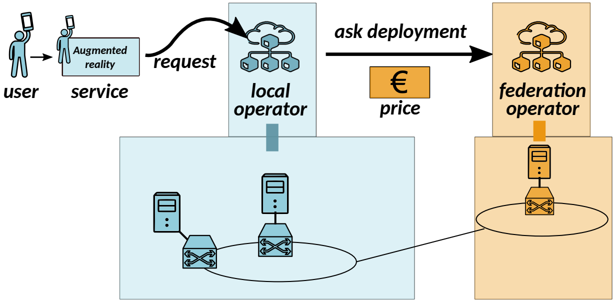


Figure 14: Service federation.

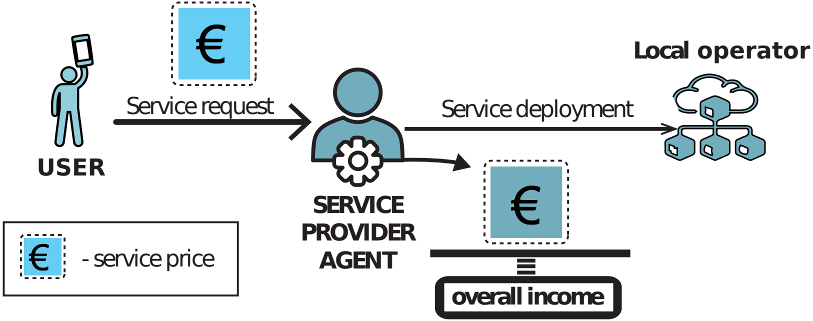


Figure 15: Business model - local deployment[2].

[2]Based on Kiril Antevski illustration

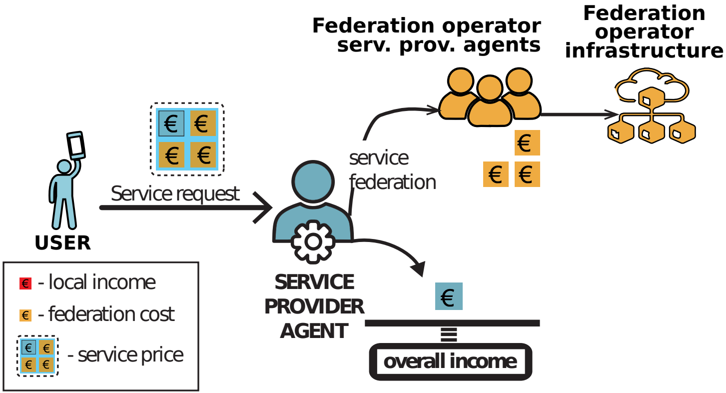


Figure 16: Business model - federate deployment[3].

[3]Based on Kiril Antevski illustration

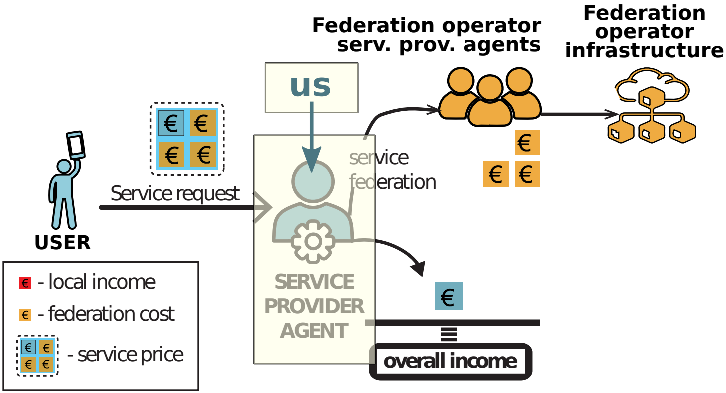


Figure 16: Business model - federate deployment[3].

[3] Based on Kiril Antevski illustration

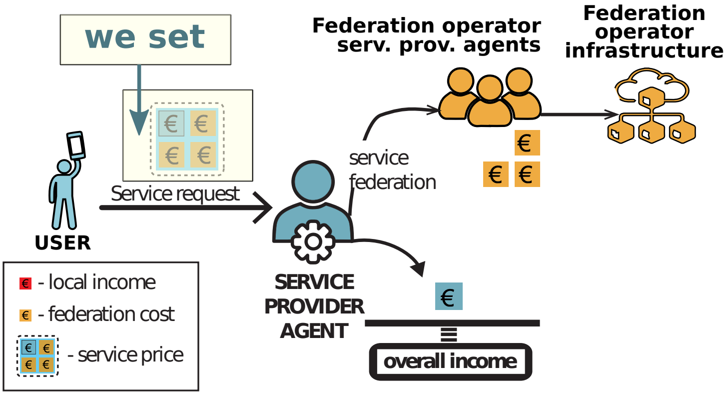


Figure 16: Business model - federate deployment[3].

[3] Based on Kiril Antevski illustration

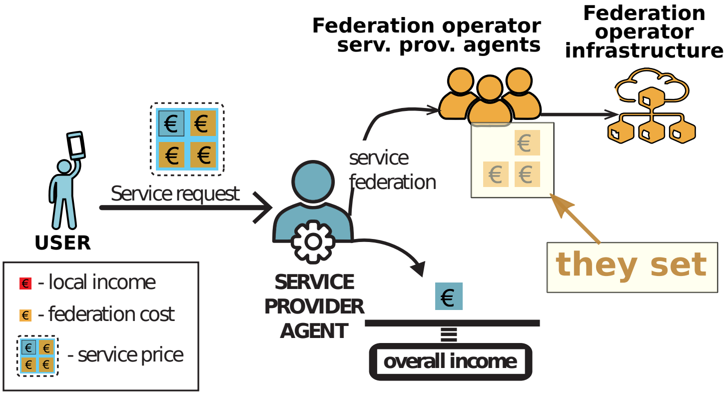


Figure 16: Business model - federate deployment[3].

[3] Based on Kiril Antevski illustration

t3a.small:

- 2 CPUs
- memory 2 GB
- storage 100 GB

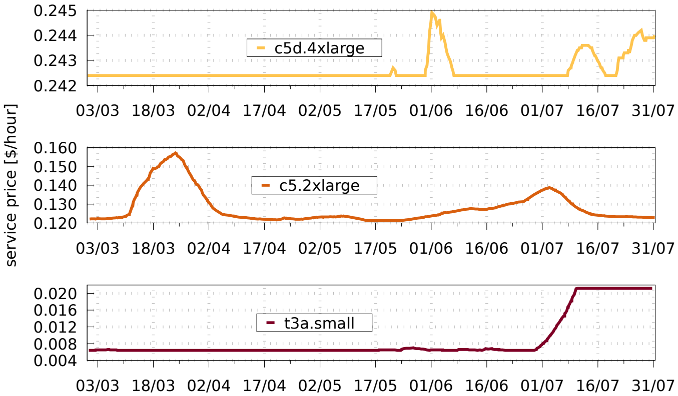


Figure 17: AWS service prices during 2020 in west Europe.

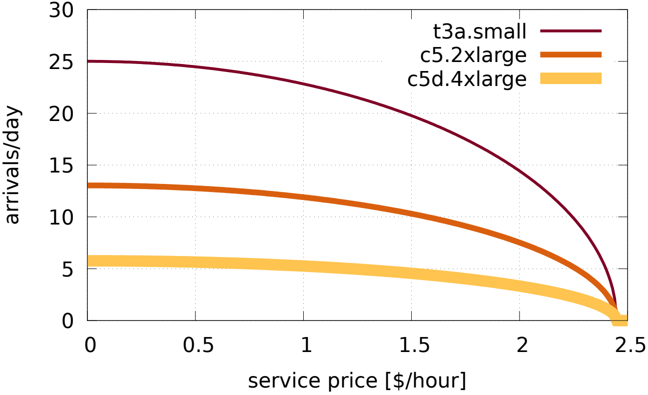


Figure 18: Impact of prices on arriving users – based on tid study [27] and [32].

Considering:

- Price changes
- Available resources (CPU, memory, disk)
- Service lifetime (e.g., 2 days)

Considering:

- Price changes
- Available resources (CPU, memory, disk)
- Service lifetime (e.g., 2 days)

For each service σ, decide / take an **action**:

- $x(\sigma) = 0$: **reject**
- $x(\sigma) = 1$: **local**
- $x(\sigma) = 2$: **federate**

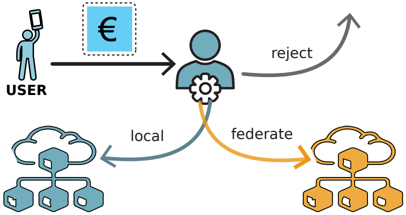


Figure 19: Possible actions.

Obtained **reward**:

$$r^{(t)}(X_t) := \sum_{\substack{\sigma:\ x(\sigma)=0\\ a(\sigma)\leq t\leq d(\sigma)}} p^{a(\sigma)}(\sigma) + \sum_{\substack{\sigma:\ x(\sigma)=1\\ a(\sigma)\leq t\leq d(\sigma)}} \left[p^{a(\sigma)}(\sigma) - f^{(t)}(\sigma)\right] \quad (2)$$

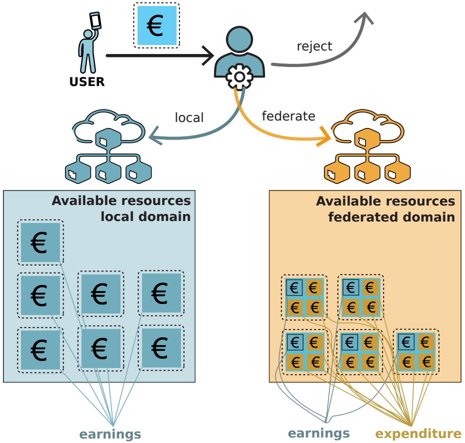


Figure 20: Environment snapshot at time t.

Obtained **reward**:

$$r^{(t)}(X_t) := \underbrace{\sum_{\substack{\sigma:\ x(\sigma)=0\\ a(\sigma)\leq t\leq d(\sigma)}} p^{a(\sigma)}(\sigma)}_{\text{local}} + \sum_{\substack{\sigma:\ x(\sigma)=1\\ a(\sigma)\leq t\leq d(\sigma)}} \left[p^{a(\sigma)}(\sigma) - f^{(t)}(\sigma)\right] \quad (2)$$

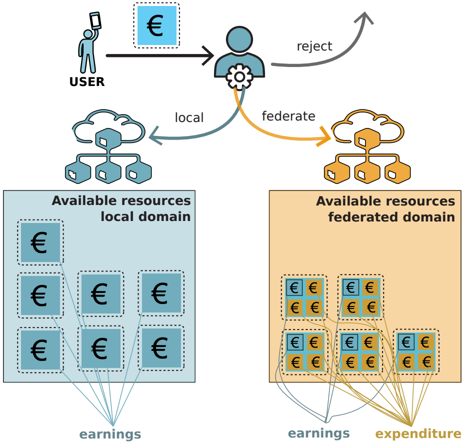


Figure 20: Environment snapshot at time t.

Obtained **reward**:

$$r^{(t)}(X_t) := \sum_{\substack{\sigma:\ x(\sigma)=0\\ a(\sigma)\leq t\leq d(\sigma)}} p^{a(\sigma)}(\sigma) + \underbrace{\sum_{\substack{\sigma:\ x(\sigma)=1\\ a(\sigma)\leq t\leq d(\sigma)}} \left[p^{a(\sigma)}(\sigma) - f^{(t)}(\sigma)\right]}_{\text{federation}} \qquad (2)$$

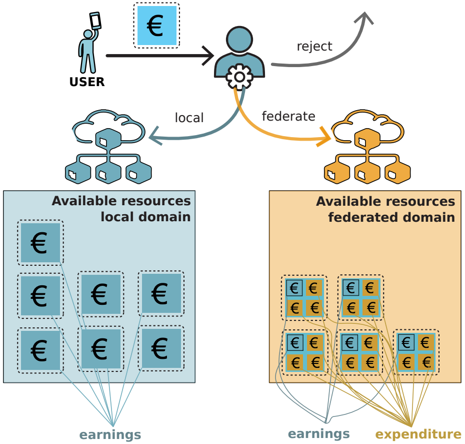


Figure 20: Environment snapshot at time t.

Online optimization problem:

- objective: $\max_{X_t} \frac{1}{T} \sum_t r^{(t)}(X_t)$
- constraints:
    - CPU
    - memory
    - disk

NP-hard: knapsack problem equivalence

Online optimization problem:

- objective: $\max_{X_t} \frac{1}{T} \sum_t r^{(t)}(X_t)$
- constraints:
  - CPU
  - memory
  - disk

NP-hard: knapsack problem equivalence

Markov Decision Problem (MDP):

- find policy π to:
  $\max_\pi \mathbb{E}_{x(\sigma)\sim\pi} \left[\sum_t \gamma^t \; r^{(t)}(\pi)\right]$
- action space $\mathcal{A} = \{0, 1, 2\}$
- state space $\mathcal{S}$:
  - available & requested resources
  - current prices
  - service lifetime
- instant reward $r^{(t)}(\pi)$

uc3m

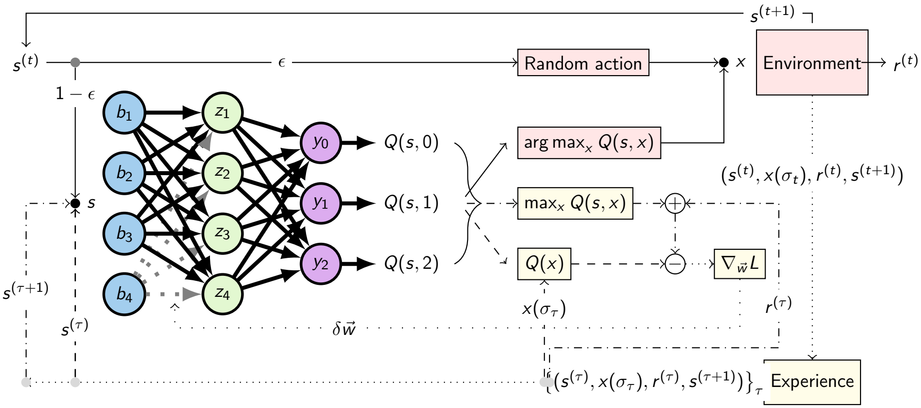


Figure 21: DQN architecture to decide rejection/local/federate.

Experimentation:

- Telefónica infrastructure & resources [27]

Experimentation:

- Telefónica infrastructure & resources [27]
- AWS prices dataset:
  - training 29/02/2020 – 02/05/2020
  - testing 03/05/2020 –31/07/2020

Experimentation:

- Telefónica infrastructure & resources [27]
- AWS prices dataset:
  - training 29/02/2020 – 02/05/2020
  - testing 03/05/2020 –31/07/2020
- Poissonian arrival of users

Comparison of:

- Optimal
- DQN
- Q-table
- Q-table explore
- greedy

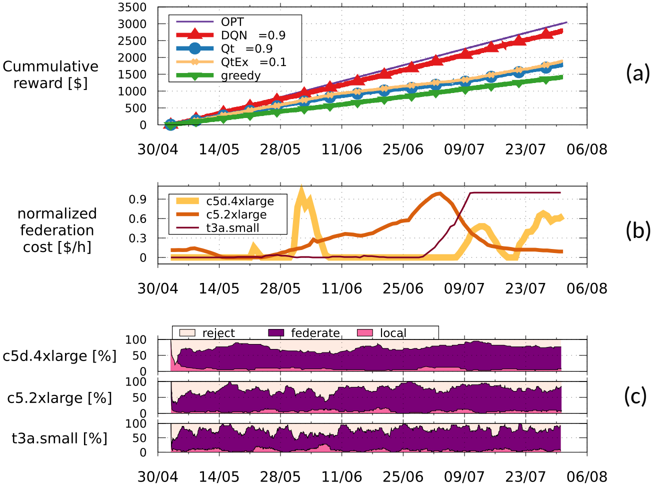


Figure 22: Federation agents' performance.

Comparison of:

- Optimal
- DQN
- Q-table
- Q-table explore
- greedy

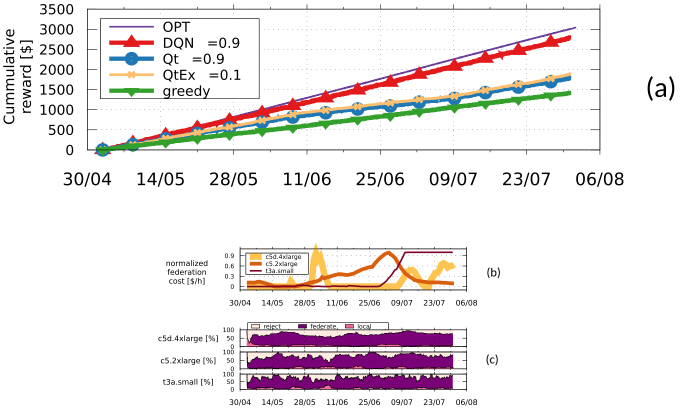


Figure 22: Federation agents' performance.

Comparison of:

- Optimal
- DQN
- Q-table
- Q-table explore
- greedy

Results:

- near-optimal

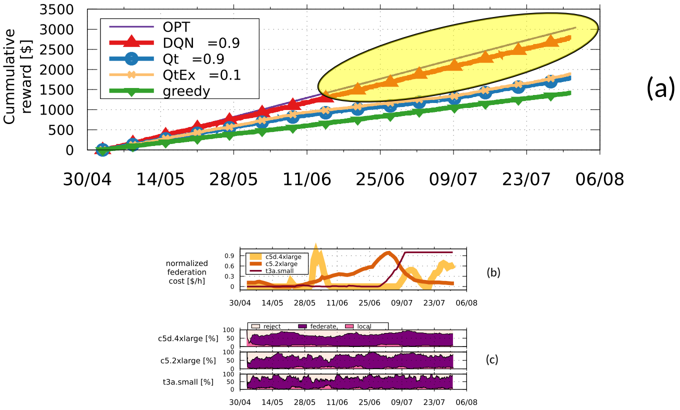


Figure 22: Federation agents' performance.

uc3m

Comparison of:

- Optimal
- DQN
- Q-table
- Q-table explore
- greedy

Results:

- near-optimal

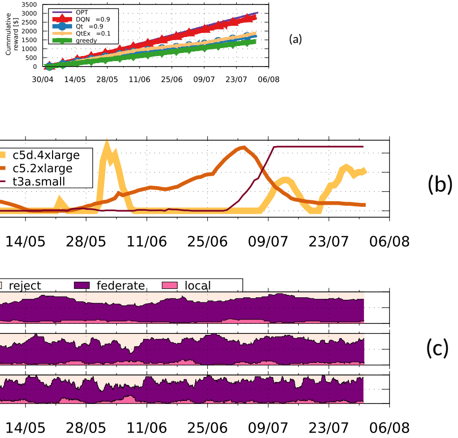


Figure 22: Federation agents' performance.

Comparison of:

- Optimal
- DQN
- Q-table
- Q-table explore
- greedy

Results:

- near-optimal
- react upon peaks

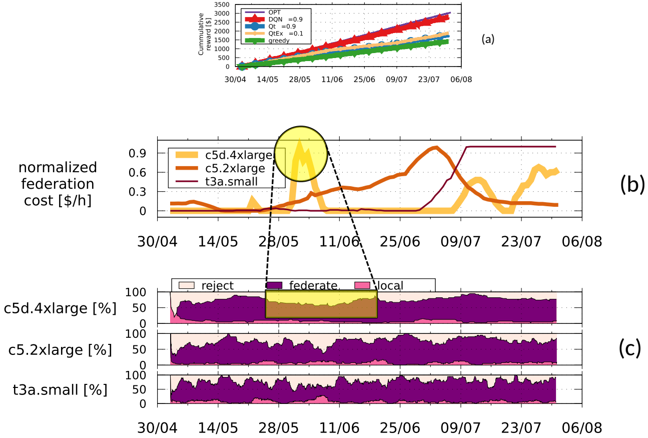


Figure 22: Federation agents' performance.

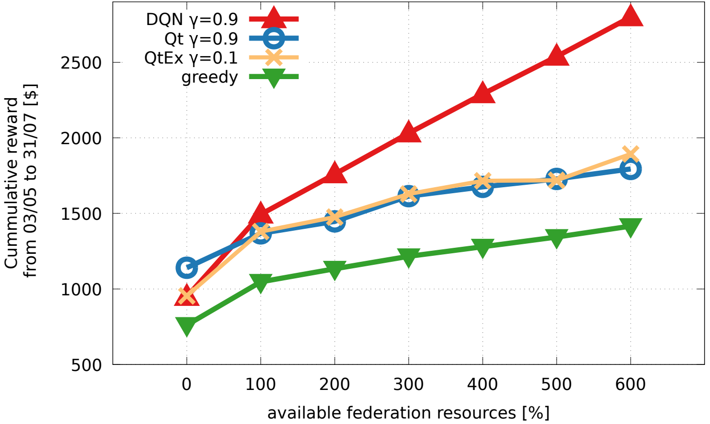


Figure 23: Cumulative reward vs. available federation resources.

Publications:

- **Martín-Pérez, Jorge** and C. J. Bernardos. "Multi-Domain VNF Mapping Algorithms". In: *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 2018, pp. 1–6. DOI: 10.1109/BMSB.2018.8436765
- K. Antevski, **J. Martín-Pérez**, A. Garcia-Saavedra, C. J. Bernardos, X. Li, J. Baranda, J. Mangues-Bafalluy, R. Martnez, and L. Vettori. "A Q-learning strategy for federation of 5G services". In: *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. 2020, pp. 1–6. DOI: 10.1109/ICC40277.2020.9149082
- **Martín-Pérez, Jorge**, K. Antevski, A. Garcia-Saavedra, X. Li, and C. J. Bernardos. "DQN Dynamic Pricing and Revenue driven Service Federation Strategy". In: *IEEE Transactions on Network and Service Management* (2021), pp. 1–1. DOI: 10.1109/TNSM.2021.3117589

Open-source:

- **DFS, BFS w/ cutoffs**: `https://github.com/MartinPJorge/placement/`
- **Q-table**: `https://github.com/MartinPJorge/5gt-federation/`
- **DQN & environment**: `https://github.com/MartinPJorge/5gt-federation/tree/extensionICC/utils/aws/`

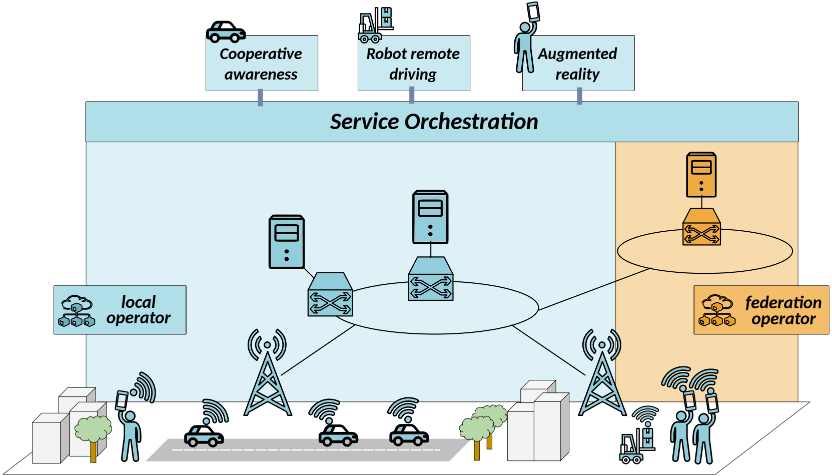


Figure 24: Local and federation operator

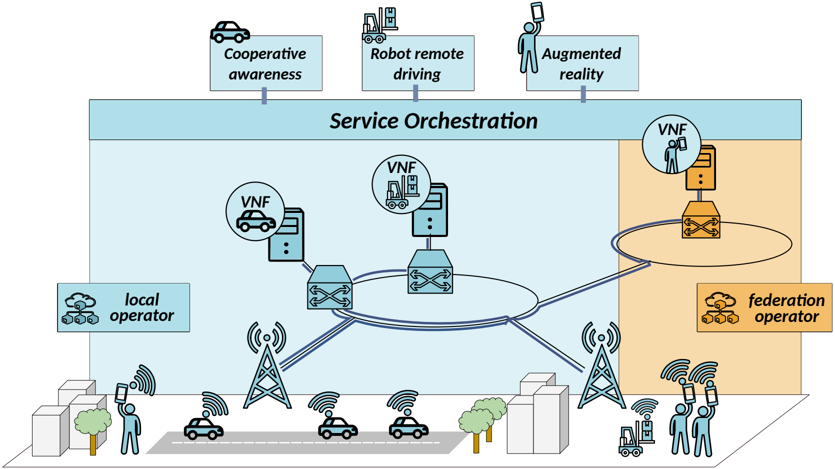


Figure 25: Services' embedding.

uc3m

OKpi **new** embedding algo. in SoA:

- latency constraints
- radio coverage
- geographical availability
- reliability

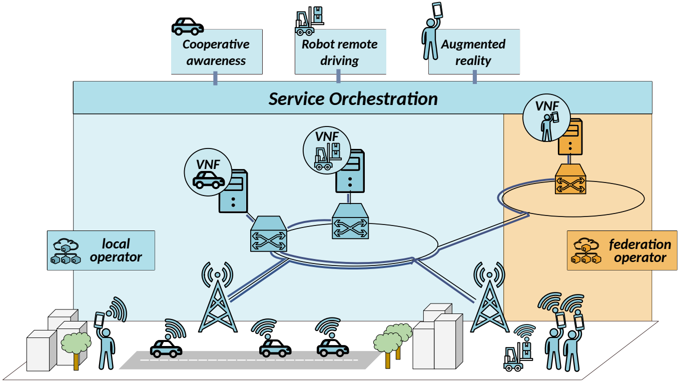


Figure 25: Services' embedding.

OKpi solves the VNE problem.

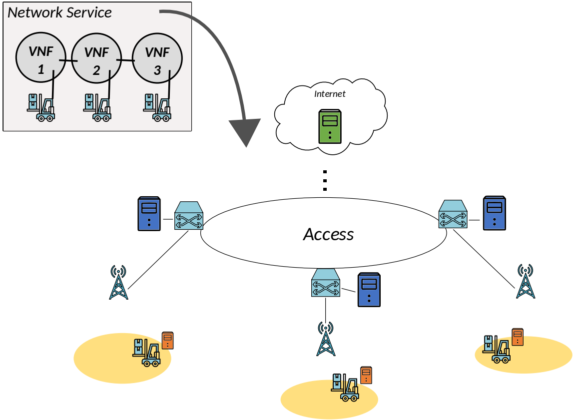


Figure 26: Virtual Network Function Embedding (VNE).

OKpi solves the VNE problem.

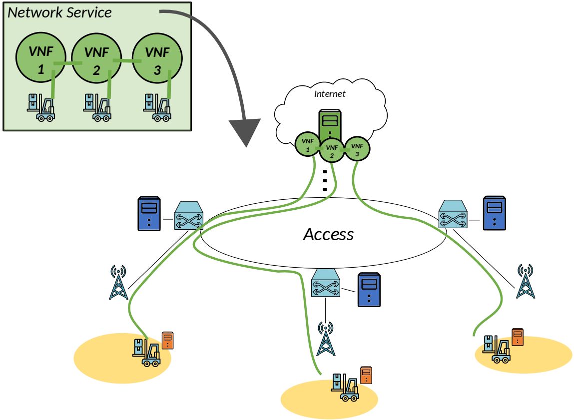


Figure 26: Virtual Network Function Embedding (VNE).

OKpi solves the VNE problem.

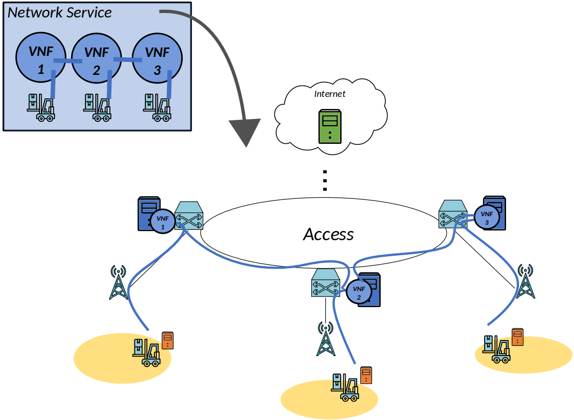


Figure 26: Virtual Network Function Embedding (VNE).

OKpi solves the VNE problem.

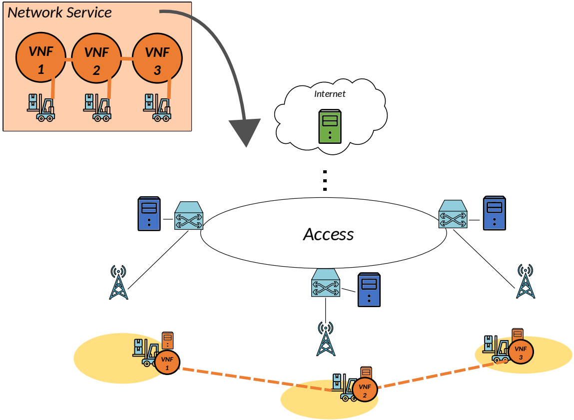


Figure 26: Virtual Network Function Embedding (VNE).

Latency constraint $D(s)$:

$$d_{\text{net}}(\psi) + d_{\text{proc}}(\psi) \leq D(s) \quad (3)$$

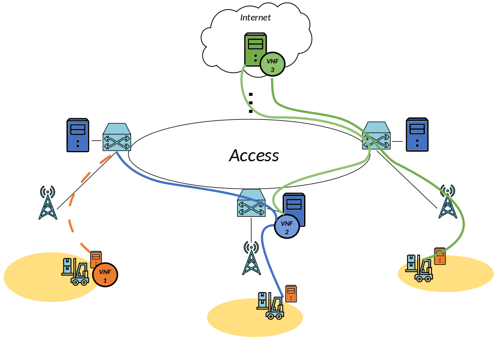


Figure 27: Service s delay.

Latency constraint $D(s)$:

$$\underbrace{d_{\text{net}}(\psi)}_{\text{propagation delay}} + d_{\text{proc}}(\psi) \leq D(s) \quad (3)$$

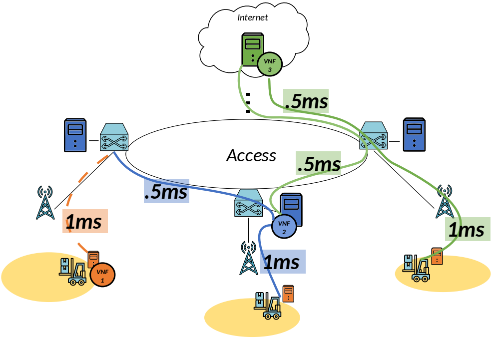


Figure 27: Service s delay.

Latency constraint $D(s)$:

$$d_{\text{net}}(\psi) + \underbrace{d_{\text{proc}}(\psi)}_{\text{processing delay}} \leq D(s) \quad (3)$$

d_{proc}: VNF as M/M/1-PS queue

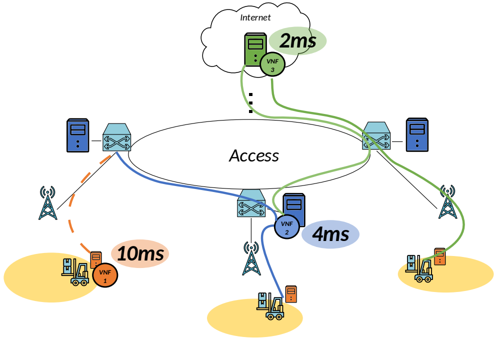


Figure 27: Service s delay.

Radio technology i constraint:

$$\rho(v,c)r_i(v) \leq R_i(c) \qquad (4)$$

- $\rho(v,c)$: VNF v is deployed in c
- $R_i(c)$: radio point of access c has radio technology i
- $r_i(v)$: VNF v needs radio technology i

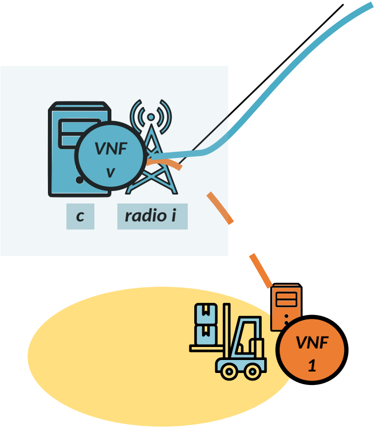


Figure 28: Radio VNF.

uc3m

Geographical availability:

$$\forall \psi = (\alpha, s), \exists c, v_1, v_2 : \tau_{\psi,c}(e, v_1, v_2) > 0 \quad (5)$$

- location α
- $\tau_{e,c}(e, v_1, v_2)$: flow (ψ, v_1, v_2) traverses link (ψ, c)

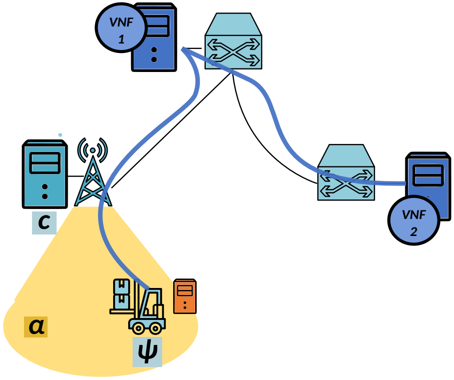


Figure 29: coverage of region α.

Service **reliability** $H(s)$:

$$\prod_{\substack{v_1, v_2 \in \mathcal{V} \\ (i,j) \in w}} \eta(j,t)\eta(i,j,t) \geq H(s) \quad (6)$$

- $\eta(j,t)$: node reliability at t
- $\eta(i,j,t)$: node reliability at t

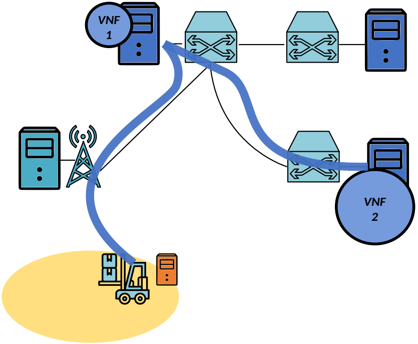


Figure 30: Traffic path.

Service **reliability** $H(s)$:

$$\prod_{v_1,v_2\in\mathcal{V}} \sum_{w\in\mathcal{W}} \underbrace{f(\psi, v_1, v_2, w)}_{\text{traffic fraction}}$$

$$\prod_{(i,j)\in w} \eta(j,t)\eta(i,j,t) \geq H(s) \quad (6)$$

- $\eta(j,t)$: node reliability at t
- $\eta(i,j,t)$: node reliability at t
- w: traffic path

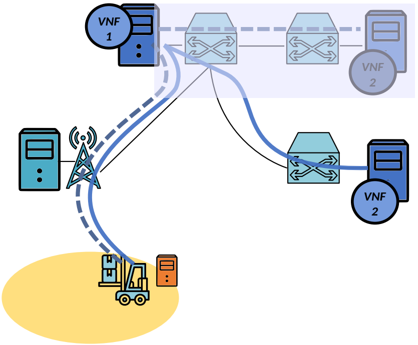


Figure 30: Fractioned traffic path.

Formulate an optimization problem:

$$\min \sum_{c} \sum_{v} \left[c_c(v) + \sum_{e} \sum_{\kappa} c_c(\kappa) a_c(\psi, v, \kappa) \right] + \sum_{(i,j)} \sum_{e} \sum_{v_1, v_2} c_{i,j} \tau_{i,j}(\psi, v_1, v_2) \tag{7}$$

$$s.t. (3) - (6) \tag{8}$$

Formulate an optimization problem:

$$\min \sum_{c} \sum_{v} \left[\underbrace{c_c(v)}_{\text{VNF cost}} + \sum_{e} \sum_{\kappa} c_c(\kappa) a_c(\psi, v, \kappa) \right] + \sum_{(i,j)} \sum_{e} \sum_{v_1, v_2} c_{i,j} \tau_{i,j}(\psi, v_1, v_2) \qquad (7)$$

$$s.t.(3) - (6) \qquad (8)$$

Formulate an optimization problem:

$$\min \sum_{c}\sum_{v}\left[c_c(v)+\sum_{e}\sum_{\kappa}\underbrace{c_c(\kappa)a_c(\psi,v,\kappa)}_{\text{assigned resource } \kappa \text{ cost}}\right]+\sum_{(i,j)}\sum_{e}\sum_{v_1,v_2} c_{i,j}\tau_{i,j}(\psi,v_1,v_2) \quad (7)$$

$$s.t.(3)-(6) \quad (8)$$

Formulate an optimization problem:

$$\min \sum_{c} \sum_{v} \left[c_c(v) + \sum_{e} \sum_{\kappa} c_c(\kappa) a_c(\psi, v, \kappa) \right] + \sum_{(i,j)} \sum_{e} \sum_{v_1, v_2} \underbrace{c_{i,j} \tau_{i,j}(\psi, v_1, v_2)}_{\text{traffic steering cost}} \tag{7}$$

$$s.t.(3) - (6) \tag{8}$$

Formulate an optimization problem:

$$\min \sum_{c} \sum_{v} \left[c_c(v) + \sum_{e} \sum_{\kappa} c_c(\kappa) a_c(\psi, v, \kappa) \right] + \sum_{(i,j)} \sum_{e} \sum_{v_1, v_2} c_{i,j} \tau_{i,j}(\psi, v_1, v_2) \tag{7}$$

$$s.t. (3) - (6) \tag{8}$$

NP-hard: bin-packing problem equivalence.

OKpi VNE idea: high $\gamma \implies$ more hops

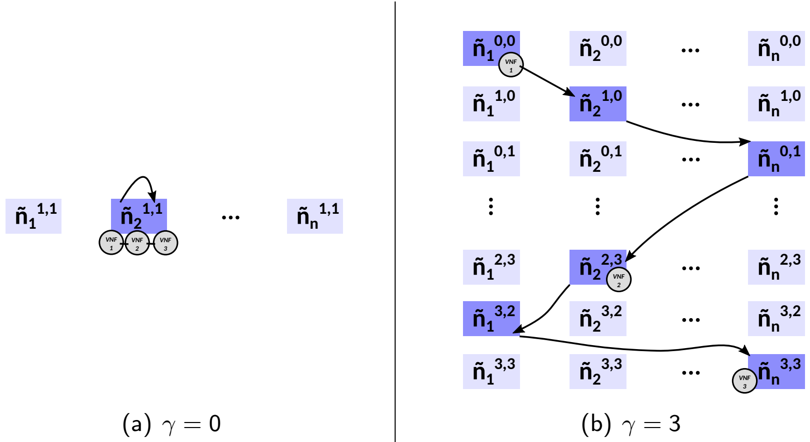


Figure 31: Impact of resolution γ on OKpi embedding.

OKpi (all KPI) solve VNE in two steps:

1. Create a decision graph
2. Create an expanded graph

Decision graph

- edge $(\tilde{n}_1, \tilde{n}_2)$ weights:

$$\left(\frac{\tilde{D}_{\tilde{n}_1,\tilde{n}_2}}{D(s)}, \frac{\log \tilde{\eta}_{\tilde{n}_1,\tilde{n}_2}}{\log H(s)} \right) \qquad (9)$$

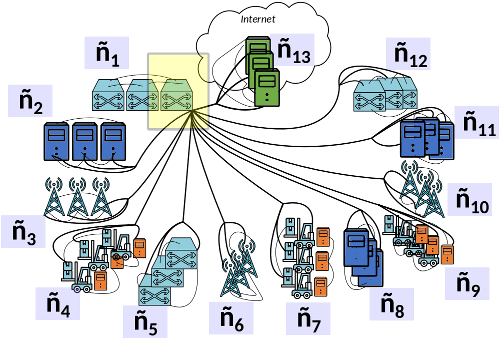


Figure 32: OKpi decision graph.

uc3m

Decision graph

- edge $(\tilde{n}_1, \tilde{n}_2)$ weights:

$$\left(\frac{\tilde{D}_{\tilde{n}_1,\tilde{n}_2}}{D(s)}, \frac{\log \tilde{\eta}_{\tilde{n}_1,\tilde{n}_2}}{\log H(s)} \right) \quad (9)$$

delay fraction

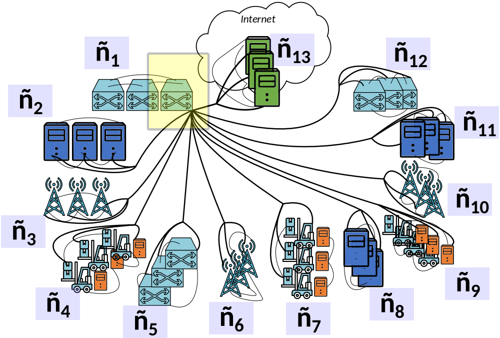


Figure 32: OKpi decision graph.

uc3m

Decision graph

- edge $(\tilde{n}_1, \tilde{n}_2)$ weights:

$$\left(\frac{\tilde{D}_{\tilde{n}_1,\tilde{n}_2}}{D(s)}, \frac{\log \tilde{\eta}_{\tilde{n}_1,\tilde{n}_2}}{\log H(s)} \right) \tag{9}$$

reliability fraction

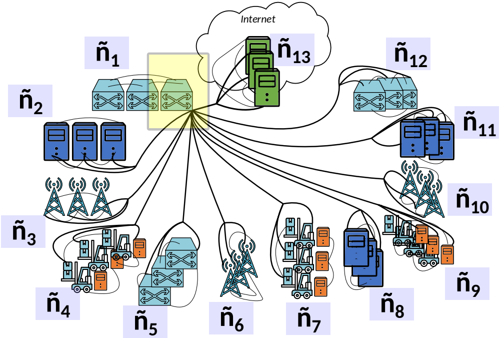


Figure 32: OKpi decision graph.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it

Figure 33: OKpi expanded graph $\gamma = 3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. do $(\gamma+1)^2$ replicas

$\tilde{n}_1^{0,0}$	$\tilde{n}_2^{0,0}$	$\cdots$	$\tilde{n}_n^{0,0}$
$\tilde{n}_1^{1,0}$	$\tilde{n}_2^{1,0}$	$\cdots$	$\tilde{n}_n^{1,0}$
$\tilde{n}_1^{0,1}$	$\tilde{n}_2^{0,1}$	$\cdots$	$\tilde{n}_n^{0,1}$
$\vdots$	$\vdots$		$\vdots$
$\tilde{n}_1^{3,3}$	$\tilde{n}_2^{3,3}$	$\cdots$	$\tilde{n}_n^{3,3}$

Figure 33: OKpi expanded graph $\gamma = 3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. do $(\gamma + 1)^2$ replicas
3. paths of up to $2 \cdot \gamma$ hops

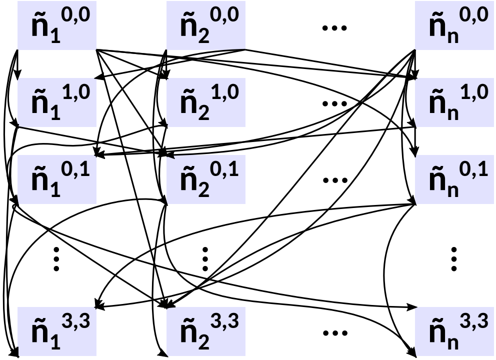


Figure 33: OKpi expanded graph $\gamma = 3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - *d*: delay to reach it
   - *r*: reliab. to reach it
2. do $(\gamma+1)^2$ replicas
3. paths of up to $2\cdot\gamma$ hops
4. embed across any path

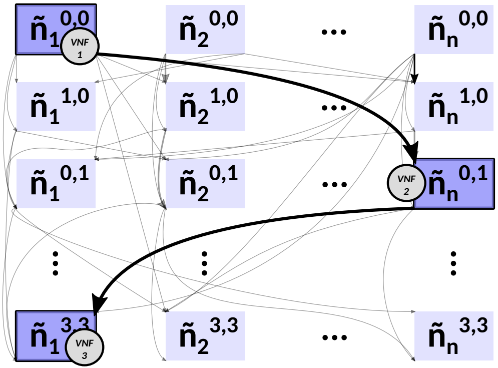


Figure 33: OKpi expanded graph $\gamma = 3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. do $(\gamma+1)^2$ replicas
3. paths of up to $2\cdot\gamma$ hops
4. embed across any path

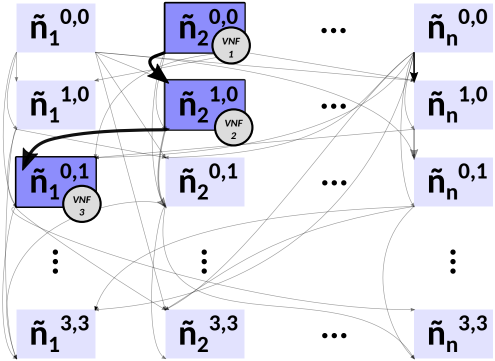
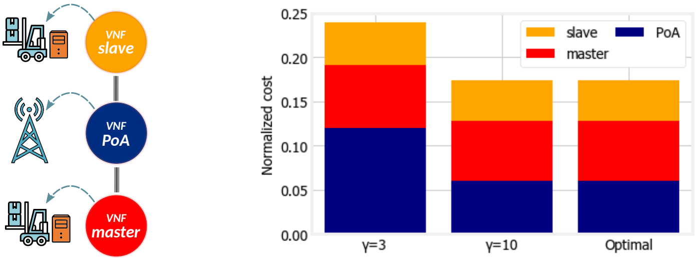


Figure 33: OKpi expanded graph $\gamma = 3$.

Simulations using realistic ITU+3GPP 5G scenarios [16]:



(a) (b)

Figure 34: (a) master-slave robotic VS illustration, and (b) optimality comparison of the VNFs' deployment costs using OKpi with $\gamma = 3, 10$.

Simulations using realistic ITU+3GPP 5G scenarios [16]:

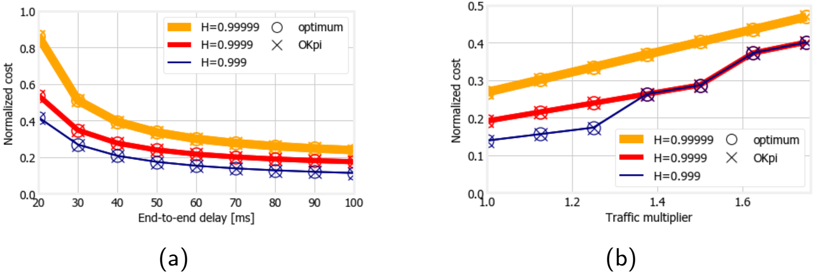


Figure 35: (a) end-to-end delay, and (b) traffic impact on deployment cost of master-slave robotic VS.

Publications:

- **Martín-Peréz, Jorge**, F. Malandrino, C. F. Chiasserini, and C. J. Bernardos. "OKpi: All-KPI Network Slicing Through Efficient Resource Allocation". In: *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 2020, pp. 804–813. DOI: `10.1109/INFOCOM41043.2020.9155263`
- J. Martín-Pérez, F. Malandrino, C. F. Chiasserini, M. Groshev, and C. J. Bernardos. "KPI Guarantees in Network Slicing". In: *IEEE/ACM Transactions on Networking* (2021), pp. 1–14. DOI: `10.1109/TNET.2021.3120318`
- B. Nemeth, N. Molner, **Martín-Pérez, J.**, C. J. Bernardos, A. de la Oliva, and B. Sonkoly. "Delay and reliability-constrained VNF placement on mobile and volatile 5G infrastructure". In: *IEEE Transactions on Mobile Computing* (2021), pp. 1–1. DOI: `10.1109/TMC.2021.3055426`

Open-source:

- **AMPLPY**: `https://github.com/ampl/amplpy/`
- **networkx**: `https://github.com/networkx/networkx/`
- **OKpi**: `https://github.com/MartinPJorge/placement/`
- **FMC**: `https://github.com/MartinPJorge/placement/`

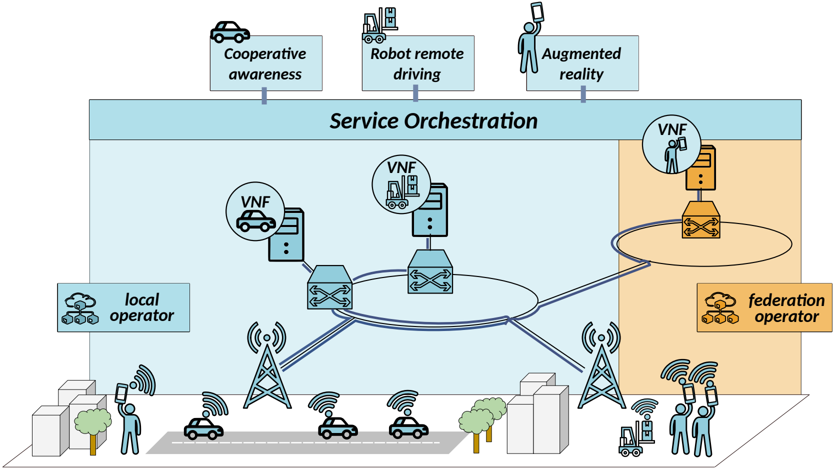


Figure 36: V2N service scaling.

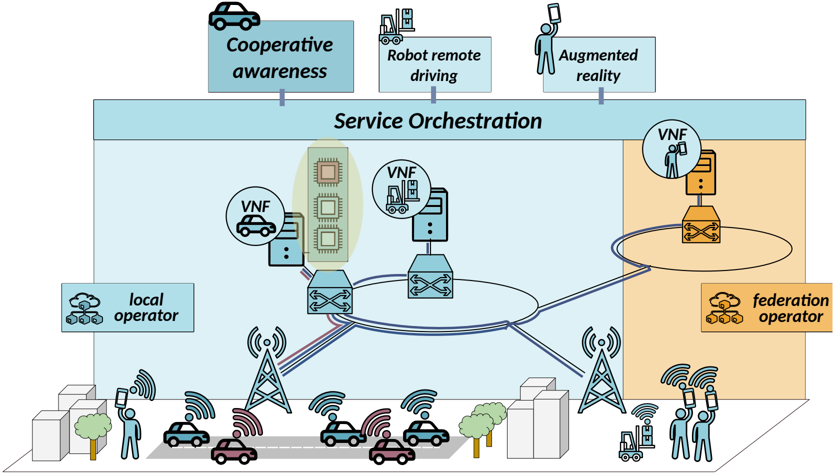


Figure 37: V2N service scaling.

uc3m

New V2N scaling using:

- time-series forecasting
- preemptive scaling based on forecast

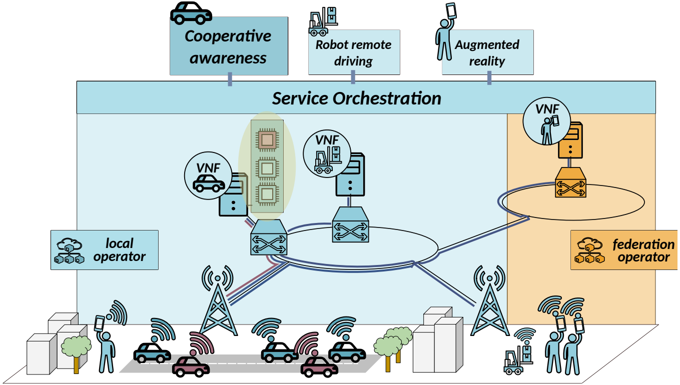


Figure 37: V2N service scaling.

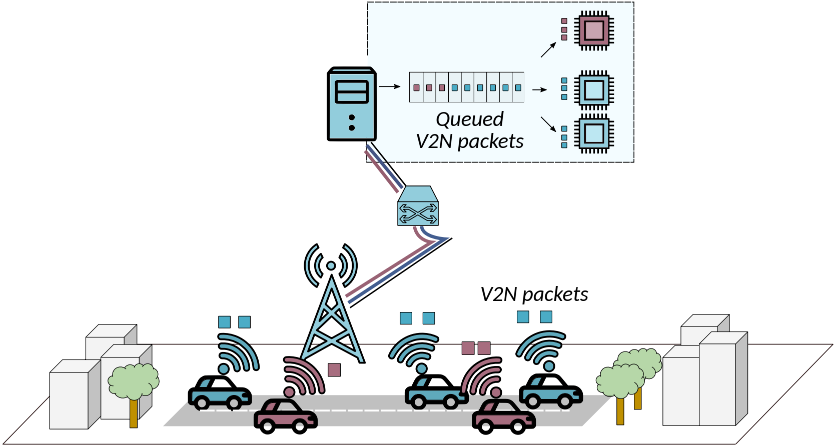


Figure 38: V2N service vertical scaling.

Server – $M/M/c$ queue:

- $\lambda(t)$: cars' arrival rate
- μ: CPU service rate
- $c(t)$: number of CPUs

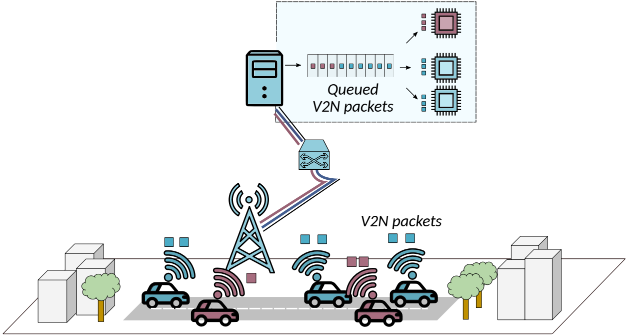


Figure 38: V2N service vertical scaling.

Dataset to derive $\lambda(t)$:

- 116 roads in Torino
- (lat,lng) of roads
- traffic [vehicles/hour] each 5 min.
- avg. speed [vehicles/hour] each 5 min.
- from 28/01/2020 – now

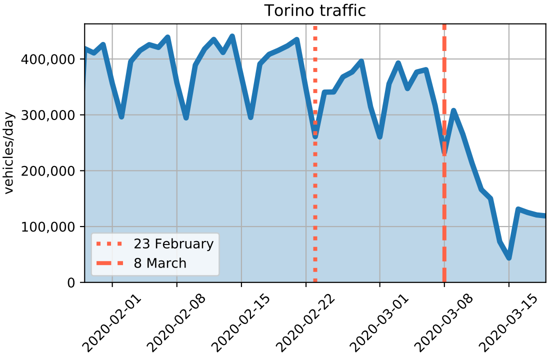


Figure 39: Traffic after COVID-19 lockdowns – 8 March.

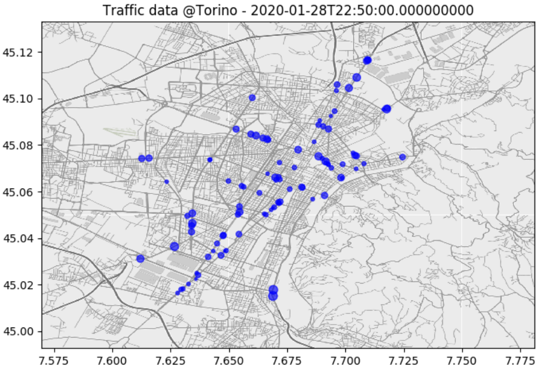


Figure 40: Vehicular traffic – wee hours @Torino.

uc3m

Predict future traffic $\lambda(t+n)$

- **time-series techniques**: DES, TES
- **proprietary**: HTM
- **neural networks**: GRU, LSTM, TCN, TCNLSTM

Patterns:

- strong **seasonality**.

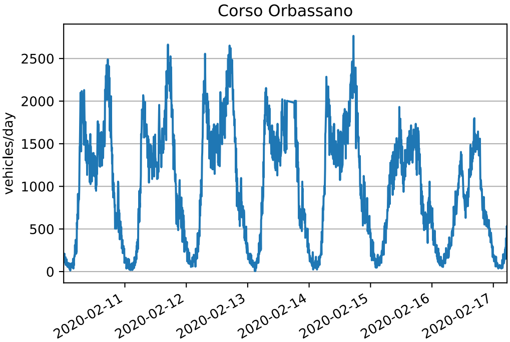


Figure 41: Weekly traffic at Corso Orbassano road.

Predict future traffic $\lambda(t+n)$

- **time-series techniques**: DES, TES
- **proprietary**: HTM
- **neural networks**: GRU, LSTM, TCN, TCNLSTM

Patterns:

- strong **seasonality**.
- week & weekend flows

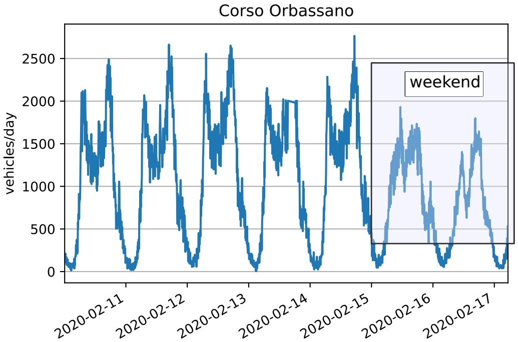


Figure 41: Weekly traffic at Corso Orbassano road.

Predict future traffic $\lambda(t+n)$

- **time-series techniques**: DES, TES
- **proprietary**: HTM
- **neural networks**: GRU, LSTM, TCN, TCNLSTM

Patterns:

- strong **seasonality**.
- week & weekend flows
- night hours,

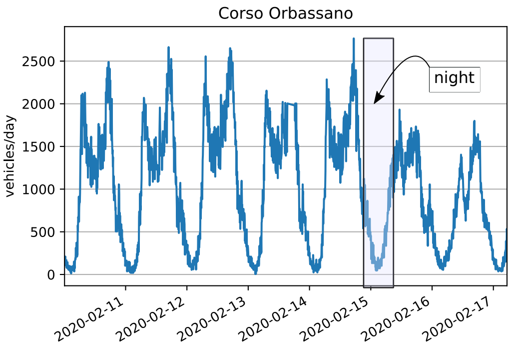


Figure 41: Weekly traffic at Corso Orbassano road.

Predict future traffic $\lambda(t+n)$

- **time-series techniques**: DES, TES
- **proprietary**: HTM
- **neural networks**: GRU, LSTM, TCN, TCNLSTM

Patterns:

- strong **seasonality**.
- week & weekend flows
- night hours, rush hours,

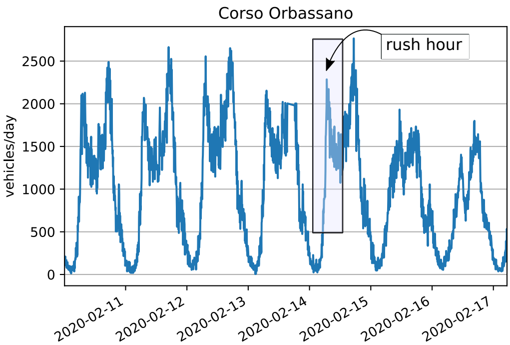


Figure 41: Weekly traffic at Corso Orbassano road.

Predict future traffic $\lambda(t+n)$

- **time-series techniques**: DES, TES
- **proprietary**: HTM
- **neural networks**: GRU, LSTM, TCN, TCNLSTM

Patterns:

- strong **seasonality**.
- week & weekend flows
- night hours, rush hours, schools' out

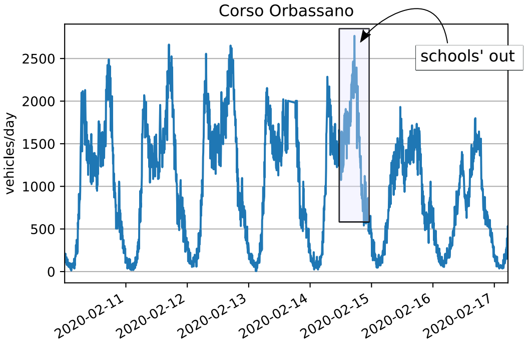


Figure 41: Weekly traffic at Corso Orbassano road.

non-COVID-19 (2020)

- training: 28thFeb - 28thMar
- testing: 29thFeb - 07thMar

COVID-19 (2020)

- training: 06thFeb - 07thMar
- testing: 08thMar - 15thMar

non-COVID-19 (2020)

- training: 28thFeb - 28thMar
- testing: 29thFeb - 07thMar

COVID-19 (2020)

- training: 06thFeb - 07thMar
- testing: 08thMar - 15thMar

Train:

- offline training
- online training

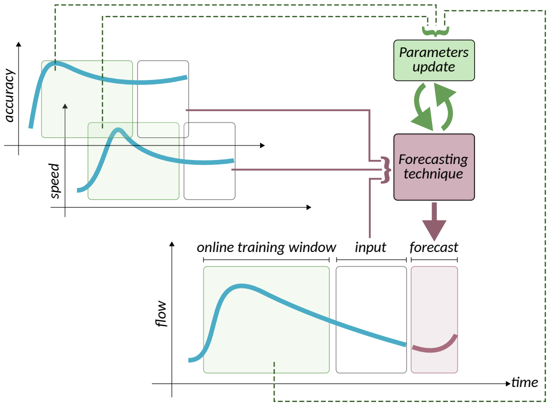


Figure 42: Online training.

Vertical scaling:

1. $\lambda(t+n)$: traffic prediction

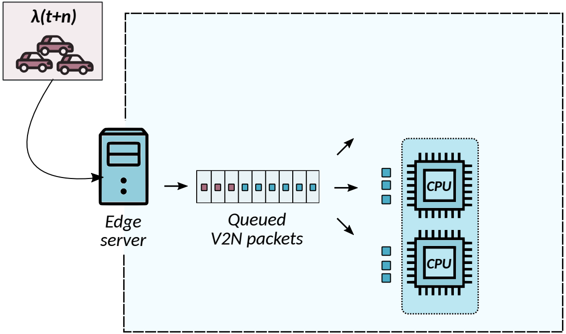


Figure 43: $M/M/c$-based scaling.

Vertical scaling:

1. $\lambda(t+n)$: traffic prediction
2. derive $c(t+n)$ s.t.:

$$\frac{1}{\mu}+\frac{P_Q}{c(t+n)\mu-\lambda(t+n)} \leq D(s) \tag{10}$$

with $D(s)$ the target delay

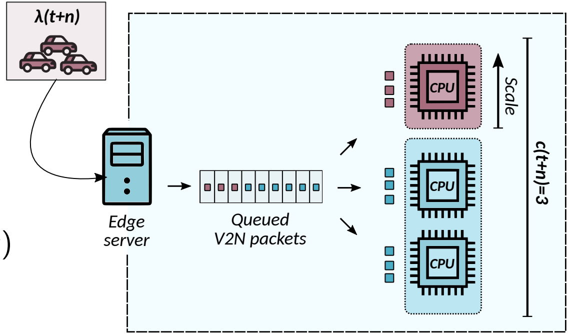
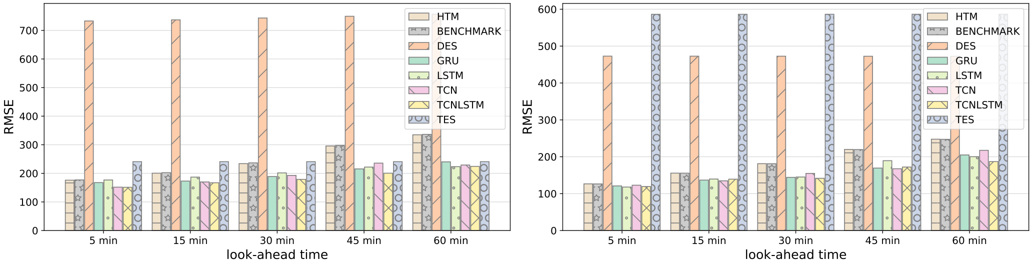


Figure 43: $M/M/c$-based scaling.



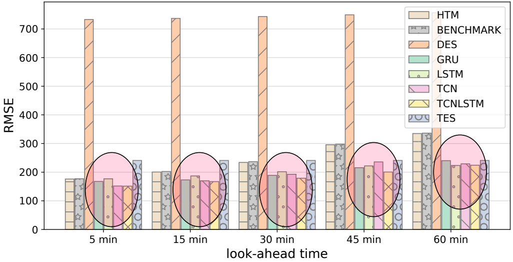
(a) non-COVID-19

(b) COVID-19

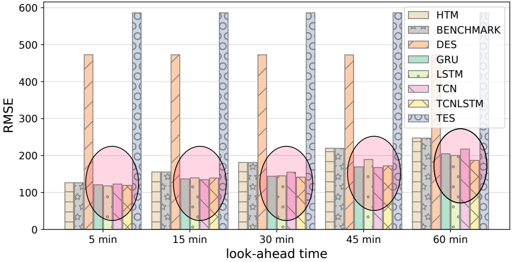
Figure 44: Prediction accuracy (offline training).

uc3m

Most accurate: **Neural Networks**



(a) non-COVID-19



(b) COVID-19

Figure 44: Prediction accuracy (offline training).

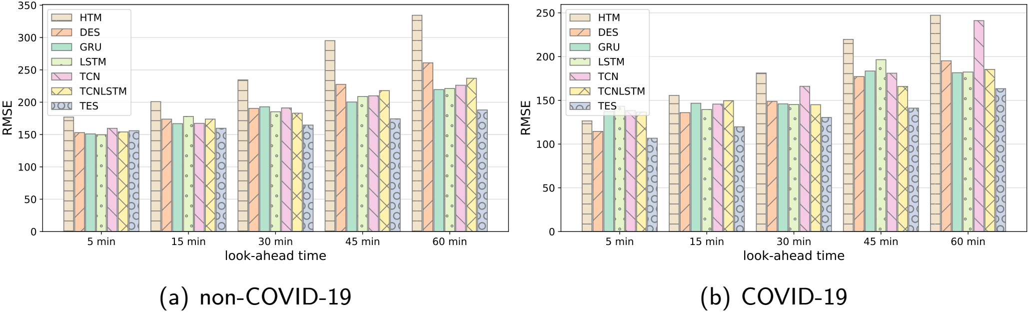
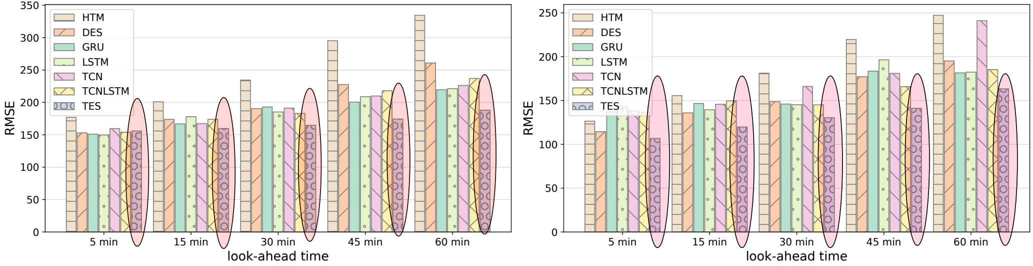


Figure 45: Prediction accuracy (online training).

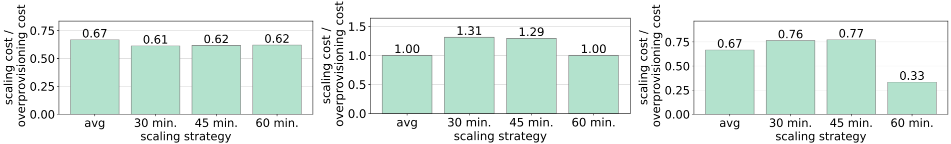
Most accurate: **TES**



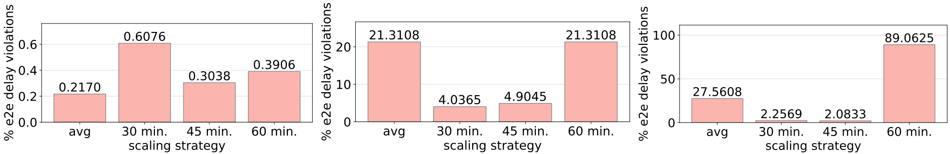
(a) non-COVID-19 (b) COVID-19

Figure 45: Prediction accuracy (online training).

uc3m



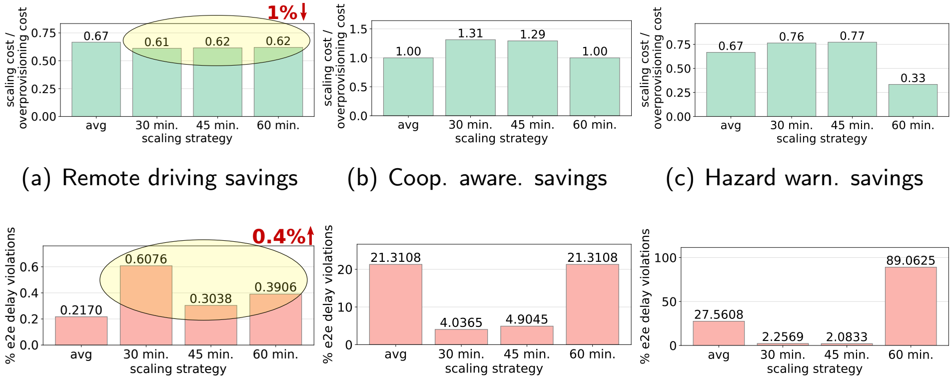
(a) Remote driving savings (b) Coop. aware. savings (c) Hazard warn. savings



(d) Remote driving delay violate (e) Coop. aware. delay violate (f) Hazard warn. violate

Figure 46: Cost savings and delay violations due to scaling – TES with online training was used.

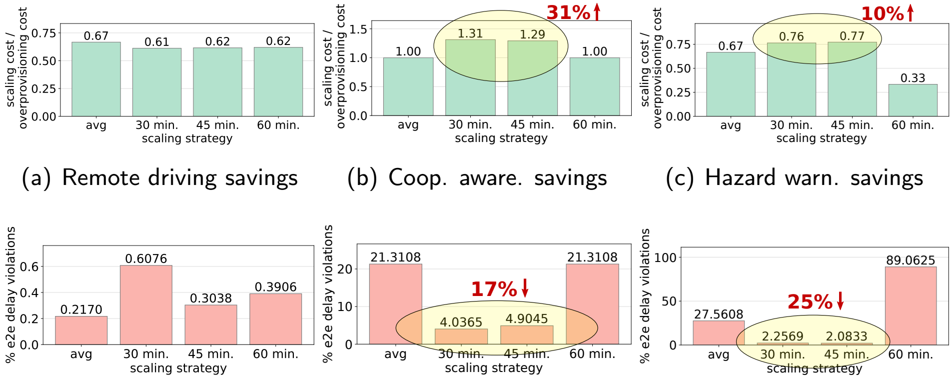
uc3m



(a) Remote driving savings (b) Coop. aware. savings (c) Hazard warn. savings

(d) Remote driving delay violate (e) Coop. aware. delay violate (f) Hazard warn. violate

Figure 46: Cost savings and delay violations due to scaling – TES with online training was used.



(a) Remote driving savings (b) Coop. aware. savings (c) Hazard warn. savings

(d) Remote driving delay violate (e) Coop. aware. delay violate (f) Hazard warn. violate

Figure 46: Cost savings and delay violations due to scaling – TES with online training was used.

uc3m

Publications:

- D. de Vleeschauwer, J. Baranda, J. Mangues-Bafalluy, C. F. Chiasserini, M. Malinverno, C. Puligheddu, L. Magoula, **Martín-Pérez, J.**, S. Barmpounakis, K. Kondepu, L. Valcarenzhi, X. Li, C. Papagianni, and A. Garcia-Saavedra. "5Growth Data-Driven AI-Based Scaling". In: *2021 EuCNC/6G Summit*. 2021, pp. 383–388. DOI: 10.1109/EuCNC/6GSummit51104.2021.9482476
- **Martín-Pérez, Jorge**, K. Kondepu, D. de Vleeschauwer, V. Reddy, C. Guimarães, A. Sgambelluri, L. Valcarenghi, C. Papagianni, and C. J. Bernardos. "Dimensioning of V2N Services in 5G Networks through Forecast-based Scaling". In: *IEEE Access* (2021). Under review

Open-source (to be released):

- `https://github.com/MartinPJorge/5growth-scaling/`
- `https://github.com/MartinPJorge/5growth-forecasting/`

uc3m

This thesis contributions:

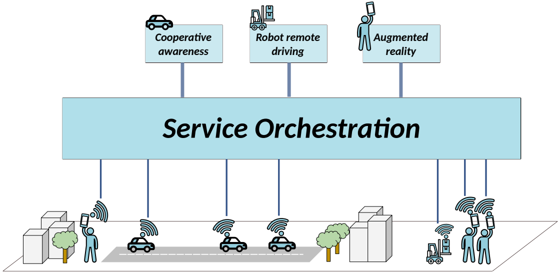


Figure 47: Service orchestration.

This thesis contributions:

1. generate **5G graphs** meeting standard requirements

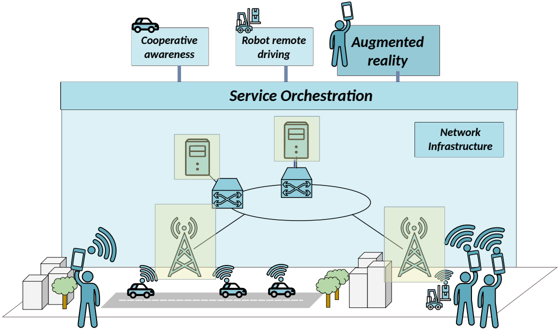


Figure 47: Service orchestration – infrastructure generation.

This thesis contributions:

1. generate **5G graphs** meeting standard requirements
2. maximize revenue under **federation** and dynamic pricing

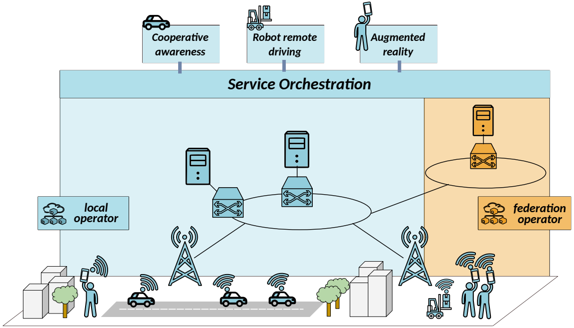


Figure 47: Service orchestration – federation.

This thesis contributions:

1. generate **5G graphs** meeting standard requirements
2. maximize revenue under **federation** and dynamic pricing
3. minimize **VNE** cost meeting: latency, reliability, and availability constraints

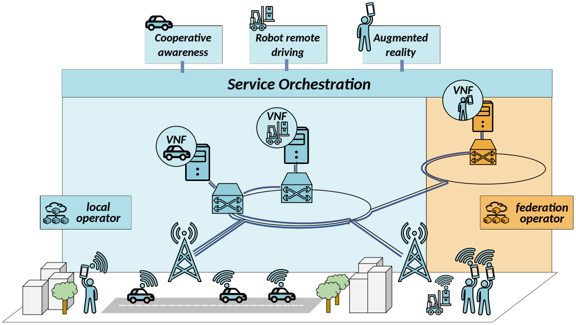


Figure 47: Service orchestration – VNE.

This thesis contributions:

1. generate **5G graphs** meeting standard requirements
2. maximize revenue under **federation** and dynamic pricing
3. minimize **VNE** cost meeting: latency, reliability, and availability constraints
4. reduce the E2E latency violations with the proposed **scaling**

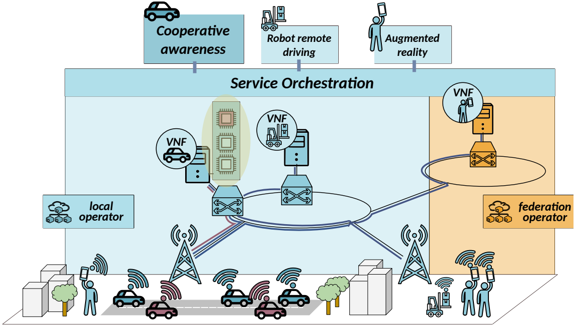


Figure 47: Service orchestration – scaling.

1. **generate federated** scenarios BSs, PoPs & datacenters, all at once

uc3m

1. **generate federated** scenarios BSs, PoPs & datacenters, all at once
2. **DQN** agent +**LSTM** layer to **predict**

1. **generate federated** scenarios BSs, PoPs & datacenters, all at once
2. **DQN** agent +**LSTM** layer to **predict**
3. **OKpi** study on γ/runtime/optimality

1. **generate federated** scenarios BSs, PoPs & datacenters, all at once
2. **DQN** agent +**LSTM** layer to **predict**
3. **OKpi** study on γ/runtime/optimality
4. **V2N scaling** to meet **99.9999%** latency quantile & use **ST-GCN**

Publications:

- Conferences:
  1. IEEE CSCN [15]
  2. IEEE BMSB [3]
  3. IEEE ICC [3]
  4. IEEE INFOCOM [18]
  5. EuCNC [18]
- Journals:
  1. IEEE TB [16]
  2. IEEE TNSM [16]
  3. IEEE TON [19]
  4. IEEE TMC [21]

Open-source (GitHub):

1. BS & server generation
2. 5GEN R package
3. DFS, BFS w/ cutoffs
4. Q-table federation
5. DQN federation + AWS env.
6. AMPLPY
7. network
8. OKpi
9. FMC

Thanks for your attention!

uc3m

[1] M. Afshang and H. S. Dhillon. "Poisson Cluster Process Based Analysis of HetNets With Correlated User and Base Station Locations". In: *IEEE Transactions on Wireless Communications* 17.4 (Apr. 2018), pp. 2417–2431. ISSN: 1536-1276. DOI: 10.1109/TWC.2018.2794983.

[2] S. Agarwal, F. Malandrino, C. F. Chiasserini, and S. De. "VNF Placement and Resource Allocation for the Support of Vertical Services in 5G Networks". In: *IEEE/ACM Trans. Netw.* 27.1 (Feb. 2019), pp. 433–446. ISSN: 1063-6692. DOI: 10.1109/TNET.2018.2890631. URL: https://doi.org/10.1109/TNET.2018.2890631.

[3] K. Antevski, **J. Martín-Pérez**, A. Garcia-Saavedra, C. J. Bernardos, X. Li, J. Baranda, J. Mangues-Bafalluy, R. Martnez, and L. Vettori. "A Q-learning strategy for federation of 5G services". In: *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. 2020, pp. 1–6. DOI: 10.1109/ICC40277.2020.9149082.

[4] A. Baddeley, C. internazionale matematico estivo, and W. Weil. *Stochastic Geometry: Lectures Given at the C.I.M.E. Summer School Held in Martina Franca, Italy, September 13-18, 2004*. Lecture Notes in Mathematics / C.I.M.E. Foundation Subseries. Springer, 2007. ISBN: 9783540381747.

[5] J. Baranda et al. "Automated deployment and scaling of automotive safety services in 5G-Transformer". In: *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*. 2019, pp. 1–2. DOI: 10.1109/NFV-SDN47374.2019.9039990.

[6] L. Cominardi, L. M. Contreras, C. J. Bernardos, and I. Berberana. "Understanding QoS Applicability in 5G Transport Networks". In: *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. June 2018, pp. 1–5. DOI: 10.1109/BMSB.2018.8436847. URL: https://e-archivo.uc3m.es/bitstream/handle/10016/27393/understanding_BMSB_2018_ps.pdf (visited on 01/10/2019).

[7] U. Fattore, M. Liebsch, B. Brik, and A. Ksentini. "AutoMEC: LSTM-Based User Mobility Prediction for Service Management in Distributed MEC Resources". In: *Proceedings of the 23rd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. MSWiM '20. Alicante, Spain: Association for Computing Machinery, 2020, pp. 155–159. ISBN: 9781450381178. DOI: 10.1145/3416010.3423246. URL: https://doi.org/10.1145/3416010.3423246.

[8] V. Frascolla et al. "5G-MiEdge: Design, standardization and deployment of 5G phase II technologies: MEC and mmWaves joint development for Tokyo 2020 Olympic games". In: *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*. Sept. 2017, pp. 54–59. DOI: 10.1109/CSCN.2017.8088598.

[9] A. M. Ibrahim, T. ElBatt, and A. El-Keyi. "Coverage probability analysis for wireless networks using repulsive point processes". In: *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. Sept. 2013, pp. 1002–1007. DOI: 10.1109/PIMRC.2013.6666284.

[10] ITU-T. *Consideration on 5G transport network reference architecture and bandwidth requirements*. Study Group 15 Contribution 0462. International Telecommunication Union - Telecommunication Standardization Sector (ITU-T), Feb. 2018.

[11] G. Li, H. Zhou, B. Feng, and G. Li. "Context-Aware Service Function Chaining and Its Cost-Effective Orchestration in Multi-Domain Networks". In: *IEEE Access* 6 (2018), pp. 34976–34991. DOI: 10.1109/ACCESS.2018.2848266.

[12] F. Malandrino and C. Chiasserini. "Getting the Most Out of Your VNFs: Flexible Assignment of Service Priorities in 5G". In: *2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*. 2019, pp. 1–9. DOI: 10.1109/WoWMoM.2019.8792983.

[13] **Martín-Pérez, Jorge**, K. Antevski, A. Garcia-Saavedra, X. Li, and C. J. Bernardos. "DQN Dynamic Pricing and Revenue driven Service Federation Strategy". In: *IEEE Transactions on Network and Service Management* (2021), pp. 1–1. DOI: 10.1109/TNSM.2021.3117589.

[14] **Martín-Pérez, Jorge** and C. J. Bernardos. "Multi-Domain VNF Mapping Algorithms". In: *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 2018, pp. 1–6. DOI: 10.1109/BMSB.2018.8436765.

[15] **Martín-Pérez, Jorge**, L. Cominardi, C. J. Bernardos, and A. Mourad. "5GEN: A tool to generate 5G infrastructure graphs". In: *2019 IEEE Conference on Standards for Communications and Networking (CSCN)*. 2019, pp. 1–4. DOI: 10.1109/CSCN.2019.8931334.

[16] **Martín-Pérez, Jorge**, L. Cominardi, C. J. Bernardos, A. de la Oliva, and A. Azcorra. "Modeling Mobile Edge Computing Deployments for Low Latency Multimedia Services". In: *IEEE Transactions on Broadcasting* 65.2 (2019), pp. 464–474. DOI: 10.1109/TBC.2019.2901406.

[17] **Martín-Pérez, Jorge**, K. Kondepu, D. de Vleeschauwer, V. Reddy, C. Guimarães, A. Sgambelluri, L. Valcarenghi, C. Papagianni, and C. J. Bernardos. "Dimensioning of V2N Services in 5G Networks through Forecast-based Scaling". In: *IEEE Access* (2021). Under review.

[18] **Martín-Peréz, Jorge**, F. Malandrino, C. F. Chiasserini, and C. J. Bernardos. "OKpi: All-KPI Network Slicing Through Efficient Resource Allocation". In: *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 2020, pp. 804–813. DOI: 10.1109/INFOCOM41043.2020.9155263.

[19] J. Martín-Pérez, F. Malandrino, C. F. Chiasserini, M. Groshev, and C. J. Bernardos. "KPI Guarantees in Network Slicing". In: *IEEE/ACM Transactions on Networking* (2021), pp. 1–14. DOI: 10.1109/TNET.2021.3120318.

uc3m

[20] B. Németh, B. Sonkoly, M. Rost, and S. Schmid. "Efficient service graph embedding: A practical approach". In: *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*. 2016, pp. 19–25. DOI: 10.1109/NFV-SDN.2016.7919470.

[21] B. Nemeth, N. Molner, **Martín-Pérez, J.**, C. J. Bernardos, A. de la Oliva, and B. Sonkoly. "Delay and reliability-constrained VNF placement on mobile and volatile 5G infrastructure". In: *IEEE Transactions on Mobile Computing* (2021), pp. 1–1. DOI: 10.1109/TMC.2021.3055426.

[22] A. Okic, L. Zanzi, V. Sciancalepore, A. Redondi, and X. Costa-Pérez. "π-ROAD: a Learn-as-You-Go Framework for On-Demand Emergency Slices in V2X Scenarios". In: *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. 2021, pp. 1–10. DOI: 10.1109/INFOCOM42981.2021.9488677.

[23] P. T. A. Quang, A. Bradai, K. D. Singh, G. Picard, and R. Riggio. "Single and Multi-Domain Adaptive Allocation Algorithms for VNF Forwarding Graph Embedding". In: *IEEE Transactions on Network and Service Management* 16.1 (2019), pp. 98–112. DOI: 10.1109/TNSM.2018.2876623.

[24] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai. "5G Radio Network Design for Ultra-Reliable Low-Latency Communication". In: *IEEE Network* 32.2 (Mar. 2018), pp. 24–31. ISSN: 0890-8044. DOI: 10.1109/MNET.2018.1700232.

[25] I. Sarrigiannis, L. M. Contreras, K. Ramantas, A. Antonopoulos, and C. Verikoukis. "Fog-Enabled Scalable C-V2X Architecture for Distributed 5G and Beyond Applications". In: *IEEE Network* 34.5 (2020), pp. 120–126. DOI: 10.1109/MNET.111.2000476.

[26] V. Sciancalepore, F. Z. Yousaf, and X. Costa-Perez. "z-TORCH: An Automated NFV Orchestration and Monitoring Solution". In: *IEEE Transactions on Network and Service Management* 15.4 (2018), pp. 1292–1306. DOI: 10.1109/TNSM.2018.2867827.

[27] A. Solano and L. M. Contreras. "Information Exchange to Support Multi-Domain Slice Service Provision for 5G/NFV". In: *2020 IFIP Networking Conference (Networking)*. 2020, pp. 773–778.

[28] V. Suryaprakash, J. Møller, and G. Fettweis. "On the Modeling and Analysis of Heterogeneous Radio Access Networks Using a Poisson Cluster Process". In: *IEEE Transactions on Wireless Communications* 14.2 (Feb. 2015), pp. 1035–1047. ISSN: 1536-1276. DOI: 10.1109/TWC.2014.2363454.

[29] V. Suryaprakash, P. Rost, and G. Fettweis. "Are Heterogeneous Cloud-Based Radio Access Networks Cost Effective?" In: *IEEE Journal on Selected Areas in Communications* 33.10 (Oct. 2015), pp. 2239–2251. ISSN: 0733-8716. DOI: 10.1109/JSAC.2015.2435275.

[30] M. Syamkumar, P. Barford, and R. Durairajan. "Deployment Characteristics of "The Edge" in Mobile Edge Computing". In: *Proceedings of the 2018 Workshop on Mobile Edge Communications*. MECOMM'18. Budapest, Hungary: ACM, 2018, pp. 43–49. ISBN: 978-1-4503-5906-1. DOI: 10.1145/3229556.3229557. URL: http://doi.acm.org/10.1145/3229556.3229557.

[31] D. de Vleeschauwer et al. "5Growth Data-Driven AI-Based Scaling". In: *2021 EuCNC/6G Summit*. 2021, pp. 383–388. DOI: 10.1109/EuCNC/6GSummit51104.2021.9482476.

[32] H. Xu and B. Li. "Dynamic cloud pricing for revenue maximization". In: *IEEE Transactions on Cloud Computing* 1.2 (2013), pp. 158–171.

[33] Q. Zhang, F. Liu, and C. Zeng. "Adaptive Interference-Aware VNF Placement for Service-Customized 5G Network Slices". In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. 2019, pp. 2449–2457. DOI: 10.1109/INFOCOM.2019.8737660.

[34] Q. Zhang, X. Wang, I. Kim, P. Palacharla, and T. Ikeuchi. "Service function chaining in multi-domain networks". In: *2016 Optical Fiber Communications Conference and Exhibition (OFC)*. 2016, pp. 1–3.

"Service orchestration is the process of designing, creating, delivering, and monitoring service offerings in an automated way." – Ericsson

Location of BSs:

- Neyman-Scott Poisson Cluster Process [28]
- Poisson Point Processes (PPPs) [4]
  - homogeneous [29, 1]
  - hard-core [9]

Location of BSs:

- Neyman-Scott Poisson Cluster Process [28]
- Poisson Point Processes (PPPs) [4]
  - homogeneous [29, 1]
  - hard-core [9]
  - **inhomogeneous & Matérn II**

Location of BSs:

- Neyman-Scott Poisson Cluster Process [28]
- Poisson Point Processes (PPPs) [4]
  - homogeneous [29, 1]
  - hard-core [9]
  - **inhomogeneous & Matérn II**

Location of MEC PoPs:

- along highways [30]
- within stadiums [8]

Location of BSs:

- Neyman-Scott Poisson Cluster Process [28]
- Poisson Point Processes (PPPs) [4]
  - homogeneous [29, 1]
  - hard-core [9]
  - **inhomogeneous & Matérn II**

Location of MEC PoPs:

- along highways [30]
- within stadiums [8]
- **population census**
- **access & aggregation rings**

Lemma

Given an inhomogeneous marked PPP X with intensity function λ, the thinning function I_2, and marks $m \sim \frac{1}{\lambda(x)}$, the resulting thinned point process, called inhomogeneous Matérn II PP, has the following average number of points at C:

$$\mathbb{E}\left[N(C)\right] := \int_C e^{-\int_{B(x,r)} \mathbb{1}(\lambda(u)>\lambda(x))\lambda(u)du} \lambda(x)\ dx \tag{11}$$

where r is the thinning radius of I_2.

with

$$I_2(x, m, X, M_X) := \begin{cases} 1 & \text{if } m = \min_{m' \in M_X} \left\{(x', m') : x' \in B(x, r)\right\} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

The RTT considered is computed as

$$RTT := 2l\left(\|x - m\|_1\right) + 2p(M) + UL + DL \tag{13}$$

We find m_M, the maximum distance from server m to the BS at position x, as:

$$\|x - m\|_1 \leq l^{-1}\left(\frac{RTT - 2p(M) - t_r}{2}\right) = m_M \tag{14}$$

with $\|\cdot\|_1$ denoting the Manhattan distance.

uc3m

Table 1: NR profiles satisfying the tactile interaction latency

Profile	**DL**	**UL**	**M1 distance**	**M2 distance**
FDD 30 kHz 2s	0.39 ms	0.39 ms	12 km	2 km
FDD 120 kHZ 7s	0.33 ms	0.33 ms	24 km	14 km
TDD 120 kHz 7s	0.39 ms	0.39 ms	12 km	2 km

- *Note*: FDD 30 kHz 2s stands for Frequency Division Duplex scheme with a subcarrier of 30 kHz and 2 symbols.
- *Note*: DL and UL values are the worst case transmission latency presented in [24].

Orchestration and **fixed pricing** in multi-domain:

- Alternating Direction Method of Multipliers (ADMM) [23]
- branching heuristic [11]
- graph-based message passing [34]
- greedy with backtracking [20]

Orchestration and **fixed pricing** in multi-domain:

- Alternating Direction Method of Multipliers (ADMM) [23]
- branching heuristic [11]
- graph-based message passing [34]
- greedy with backtracking [20]
- **cutoffs in Dijkstra, DFS and BFS [14]**
- **Q-learning federation [3]**

Orchestration and **fixed pricing** in multi-domain:

- Alternating Direction Method of Multipliers (ADMM) [23]
- branching heuristic [11]
- graph-based message passing [34]
- greedy with backtracking [20]
- **cutoffs in Dijkstra, DFS and BFS [14]**
- **Q-learning federation [3]**

Orchestration and **dynamic pricing** in multi-domain:

Orchestration and **fixed pricing** in multi-domain:

- Alternating Direction Method of Multipliers (ADMM) [23]
- branching heuristic [11]
- graph-based message passing [34]
- greedy with backtracking [20]
- **cutoffs in Dijkstra, DFS and BFS [14]**
- **Q-learning federation [3]**

Orchestration and **dynamic pricing** in multi-domain:

- **real-price traces AWS**

Orchestration and **fixed pricing** in multi-domain:

- Alternating Direction Method of Multipliers (ADMM) [23]
- branching heuristic [11]
- graph-based message passing [34]
- greedy with backtracking [20]
- **cutoffs in Dijkstra, DFS and BFS [14]**
- **Q-learning federation [3]**

Orchestration and **dynamic pricing** in multi-domain:

- **real-price traces AWS**
- **Deep Q-learning**

Orchestration and **fixed pricing** in multi-domain:

- Alternating Direction Method of Multipliers (ADMM) [23]
- branching heuristic [11]
- graph-based message passing [34]
- greedy with backtracking [20]
- **cutoffs in Dijkstra, DFS and BFS [14]**
- **Q-learning federation [3]**

Orchestration and **dynamic pricing** in multi-domain:

- **real-price traces AWS**
- **Deep Q-learning**
- **Telefónica scenario**

User pays $p^{(t)}$ for the service σ

$$p^{(t)}(\sigma) = (1 + P)l^{(t)}(\sigma) \tag{15}$$

with P the profit margin, and $l^{(t)}$ the local deployment cost (based on uncertain phenomena).

User pays $p^{(t)}$ for the service σ

$$p^{(t)}(\sigma) = (1+P)l^{(t)}(\sigma) \tag{15}$$

with P the profit margin, and $l^{(t)}$ the local deployment cost (based on uncertain phenomena).

Given the federation fee $f(\sigma)$ the **reward** is:

$$r^{(t)}(X_t) := \sum_{\substack{\sigma:\ x(\sigma)=0 \\ a(\sigma)\leq t\leq d(\sigma)}} p^{a(\sigma)}(\sigma) + \sum_{\substack{\sigma:\ x(\sigma)=1 \\ a(\sigma)\leq t\leq d(\sigma)}} \left[p^{a(\sigma)}(\sigma) - f^{(t)}(\sigma)\right] \tag{16}$$

where $X_t := \{x(\sigma)\}_{\sigma:a(\sigma)\leq t}$.

$$f(p^{(t)}(\sigma)) := \begin{cases} k\left(1-\left(\frac{p^{(t)}(\sigma)}{K\cdot M}\right)^a\right)^b, & p^{(t)}(\sigma) \leq K\cdot M \\ 0, & p^{(t)}(\sigma) > K\cdot M \end{cases} \tag{17}$$

where $M = \max_{\sigma,t}\{l^{(t)}(\sigma)\}$ is the maximum local deployment cost over time across all services σ (e.g., *t3a.small*), and K is a normalization constant to control the decay of the arrival rate.

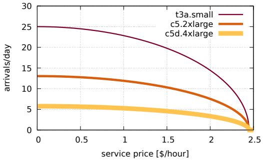


Figure 48: Impact of prices on arriving users.

Increase of P leads to:

- less user arrivals
- larger reward

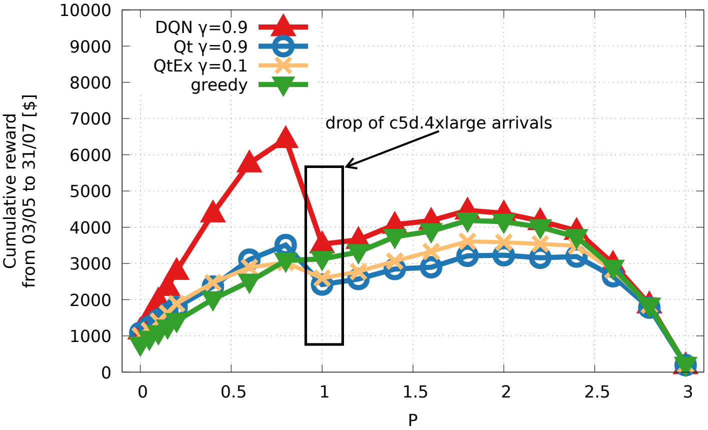
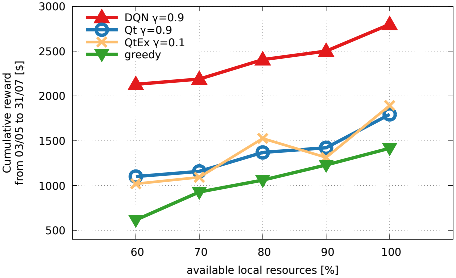
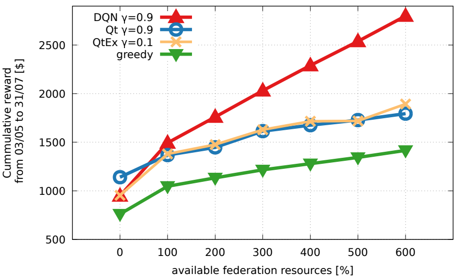


Figure 49: Impact of the marginal benefit P in the commutative reward achieved by each solution.

uc3m



(a) Increasing local resources.



(b) Increasing resources in federation.

Figure 50: Cumulative reward vs. available resources.

Existing Virtual Network Embedding (VNE) solutions:

- latency-aware, bipartite graph & Hungarian [12]
- maxZ: latency-aware, relaxed ILP [2]
- z-TORCH: KPI monitoring, k-means VNF assign [26]
- AIA: meet latency and throughput [33]

Existing Virtual Network Embedding (VNE) solutions:

- latency-aware, bipartite graph & Hungarian [12]
- maxZ: latency-aware, relaxed ILP [2]
- z-TORCH: KPI monitoring, k-means VNF assign [26]
- AIA: meet latency and throughput [33]

OKpi accounts for:

- latency constraints
- **radio coverage**
- **geographical availability**
- **reliability**

The **OKpi** (all KPI) solution:

- infrastructure as a graph
- edges with:
  - delay
  - reliability

The **OKpi** (all KPI) solution:

- infrastructure as a graph
- edges with:
  - delay
  - reliability

Solve in two steps:

1. Create a decision graph
2. Create an expanded graph

Decision graph $\tilde{G} = (\tilde{N}, \tilde{E})$.

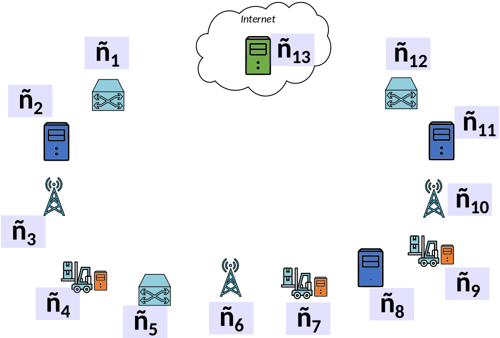


Figure 51: OKpi decision graph.

Decision graph $\tilde{G} = (\tilde{N}, \tilde{E})$.

Nodes $\tilde{N} = \{\tilde{n}_1, \tilde{n}_2, \dots\}$:

- $|\mathcal{V}| - 1$ replicas

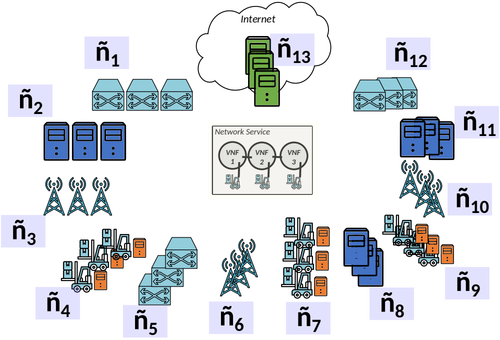


Figure 51: OKpi decision graph.

Decision graph $\tilde{G} = (\tilde{N}, \tilde{E})$.

Nodes $\tilde{N} = \{\tilde{n}_1, \tilde{n}_2, \dots\}$:

- $|\mathcal{V}| - 1$ replicas

Edges $\tilde{E} = \{(\tilde{n}_1, \tilde{n}_2), \dots\}$:

- two weights:

$$\left(\frac{\tilde{D}_{\tilde{n}_1,\tilde{n}_2}}{D(s)}, \frac{\log \tilde{\eta}_{\tilde{n}_1,\tilde{n}_2}}{\log H(s)} \right) \tag{18}$$

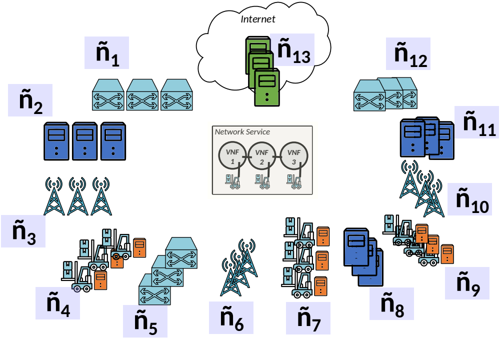


Figure 51: OKpi decision graph.

Decision graph $\tilde{G} = (\tilde{N}, \tilde{E})$.

Nodes $\tilde{N} = \{\tilde{n}_1, \tilde{n}_2, \dots\}$:

- $|\mathcal{V}| - 1$ replicas

Edges $\tilde{E} = \{(\tilde{n}_1, \tilde{n}_2), \dots\}$:

- two weights:

$$\left(\frac{\tilde{D}_{\tilde{n}_1,\tilde{n}_2}}{D(s)}, \frac{\log \tilde{\eta}_{\tilde{n}_1,\tilde{n}_2}}{\log H(s)} \right) \tag{18}$$

delay fraction

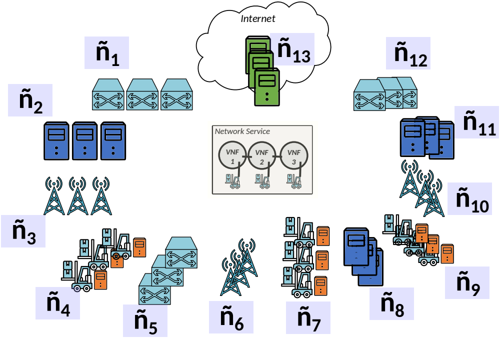


Figure 51: OKpi decision graph.

Decision graph $\tilde{G} = (\tilde{N}, \tilde{E})$.

Nodes $\tilde{N} = \{\tilde{n}_1, \tilde{n}_2, \dots\}$:

- $|\mathcal{V}| - 1$ replicas

Edges $\tilde{E} = \{(\tilde{n}_1, \tilde{n}_2), \dots\}$:

- two weights:

$$\left(\frac{\tilde{D}_{\tilde{n}_1, \tilde{n}_2}}{D(s)}, \underbrace{\frac{\log \tilde{\eta}_{\tilde{n}_1, \tilde{n}_2}}{\log H(s)}}_{\text{reliability fraction}} \right) \tag{18}$$

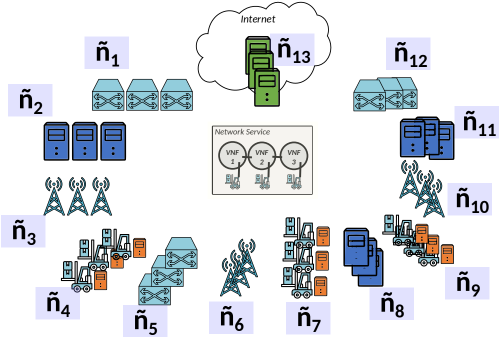


Figure 51: OKpi decision graph.

uc3m

Decision graph $\tilde{G} = (\tilde{N}, \tilde{E})$.

Nodes $\tilde{N} = \{\tilde{n}_1, \tilde{n}_2, \dots\}$:

- $|\mathcal{V}| - 1$ replicas

Edges $\tilde{E} = \{(\tilde{n}_1, \tilde{n}_2), \dots\}$:

- two weights:

$$\left(\frac{\tilde{D}_{\tilde{n}_1,\tilde{n}_2}}{D(s)}, \frac{\log \tilde{\eta}_{\tilde{n}_1,\tilde{n}_2}}{\log H(s)} \right) \quad (18)$$

- create links $(\tilde{n}_1, \tilde{n}_2)$

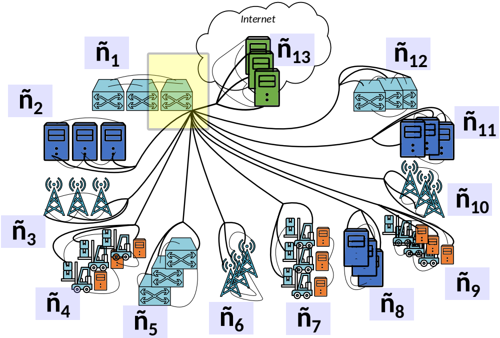


Figure 51: OKpi decision graph.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it

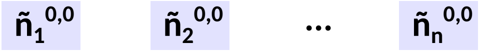


Figure 52: OKpi expanded graph $\gamma = 3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. add $(\gamma + 1)^2$ replicas

$$
\begin{array}{cccc}
\tilde{n}_1^{0,0} & \tilde{n}_2^{0,0} & \cdots & \tilde{n}_n^{0,0} \\
\tilde{n}_1^{1,0} & \tilde{n}_2^{1,0} & \cdots & \tilde{n}_n^{1,0} \\
\tilde{n}_1^{0,1} & \tilde{n}_2^{0,1} & \cdots & \tilde{n}_n^{0,1} \\
\vdots & \vdots & & \vdots \\
\tilde{n}_1^{3,3} & \tilde{n}_2^{3,3} & \cdots & \tilde{n}_n^{3,3}
\end{array}
$$

Figure 52: OKpi expanded graph $\gamma = 3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. add $(\gamma+1)^2$ replicas
3. connect $\tilde{n_1}^{d_1,r_2}$ with $\tilde{n_2}^{d_2,r_2}$

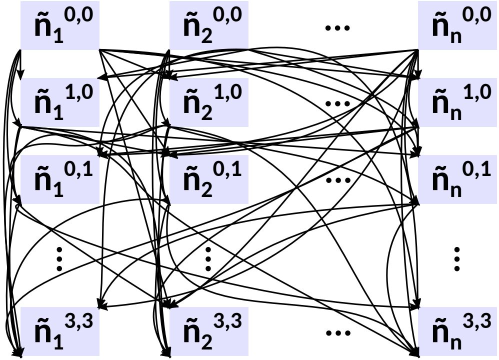


Figure 52: OKpi expanded graph $\gamma = 3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. add $(\gamma+1)^2$ replicas
3. connect $\tilde{n}_1{}^{d_1,r_2}$ with $\tilde{n}_2{}^{d_2,r_2}$
   - link $(\tilde{n}_1, \tilde{n}_2) \in \tilde{E}$
   - $d_1 + \gamma \cdot d(\tilde{n}_1, \tilde{n}_2) \leq d_2$
   - $r_1 + \gamma \cdot r(\tilde{n}_1, \tilde{n}_2) \leq r_2$

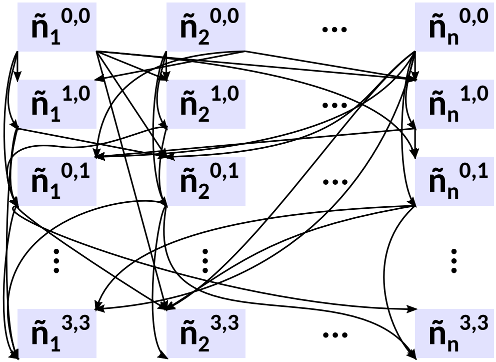


Figure 52: OKpi expanded graph $\gamma = 3$.

uc3m

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. add $(\gamma+1)^2$ replicas
3. connect $\tilde{n}_1{}^{d_1,r_2}$ with $\tilde{n}_2{}^{d_2,r_2}$
   - link $(\tilde{n}_1,\tilde{n}_2)\in\tilde{E}$
   - $d_1+\gamma\cdot d(\tilde{n}_1,\tilde{n}_2)\leq d_2$
   - $r_1+\gamma\cdot r(\tilde{n}_1,\tilde{n}_2)\leq r_2$
4. one hop per VNF

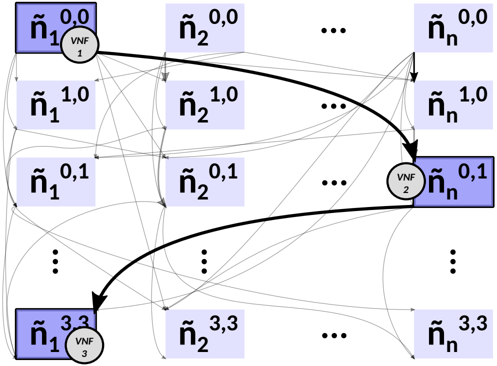


Figure 52: OKpi expanded graph $\gamma=3$.

Expanded graph:

1. add superscripts $\tilde{n}^{d,r}$
   - d: delay to reach it
   - r: reliab. to reach it
2. add $(\gamma+1)^2$ replicas
3. connect $\tilde{n}_1{}^{d_1,r_2}$ with $\tilde{n}_2{}^{d_2,r_2}$
   - link $(\tilde{n}_1,\tilde{n}_2)\in\tilde{E}$
   - $d_1+\gamma\cdot d(\tilde{n}_1,\tilde{n}_2)\leq d_2$
   - $r_1+\gamma\cdot r(\tilde{n}_1,\tilde{n}_2)\leq r_2$
4. one hop per VNF

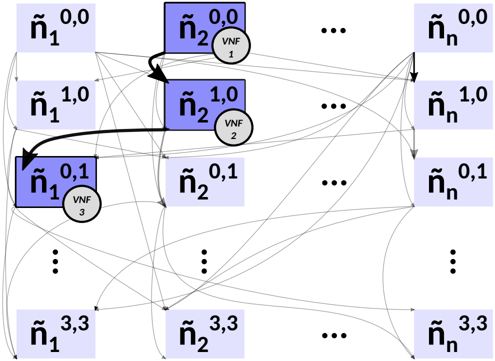


Figure 52: OKpi expanded graph $\gamma = 3$.

V2N scaling solutions:

- asign radio resource blocks [22]
- computing resources scaling:
  - threshold-based [5, 25]
  - LSTM-based [7]

V2N scaling solutions:

- asign radio resource blocks [22]
- computing resources scaling:
  - threshold-based [5, 25]
  - LSTM-based [7]
  - **compare**:
    - DES, TES
    - HTM
    - GRU
    - LSTM
    - TCN
    - TCNLSTM