ODESA: Load-Dependent Edge Server Activation for Lower Energy Footprint

Blas Gómez[1,*], Suzan Bayhan[2], Estefanía Coronado[1,3], José Villalón[1] and Antonio Garrido[1]

IEEE Wireless Communications and Networking Conference | 23 April 2024

[1]High-Performance Networks and Architectures, Universidad de Castilla-La Mancha, Albacete, Spain
[2]Faculty of EEMCS, University of Twente, Enschede, The Netherlands
[3]I2CAT Foundation, Barcelona, Spain

*blas.gomez@uclm.es

EUROPEAN UNION
European Social Fund
2019-PREDUCLM-10921

Introduction

Motivation

- **Edge Computing ➔ Proximity to the user**
  - Tens of thousands of BSs
  - Potentially thousands of edge servers
- **Need for strategies to reduce the energy footprint**
- **Edge servers are always-ready ➔ always on**
  - Capacity scaled to peak demand ➔ Idle resources in off-peak hours
  - Idle resources consume energy
- **This work only considers operational energy consumption**

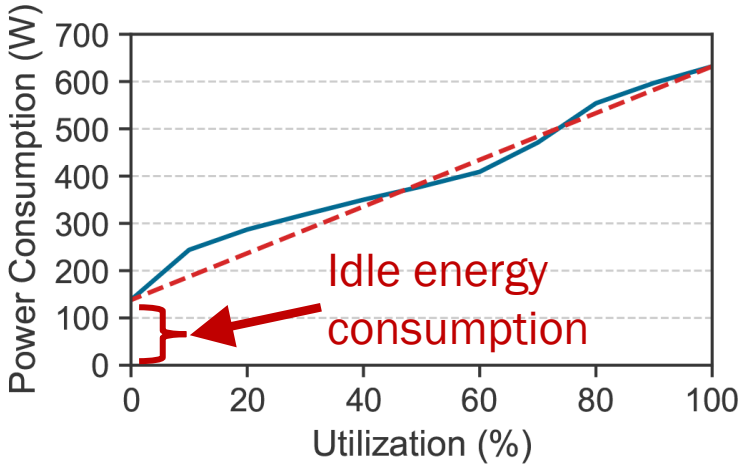
Introduction

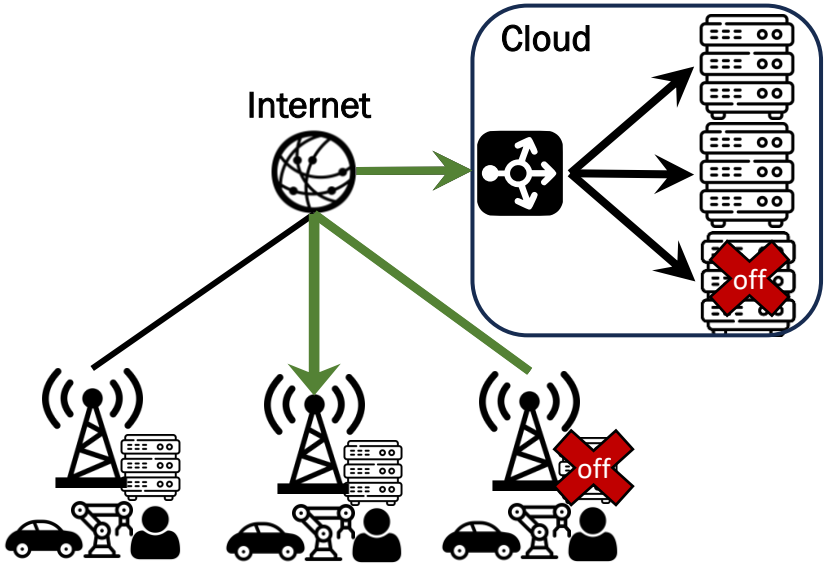
Cloud vs. Edge

- Idle resources use energy
- Cloud ➔ Resource scaling on demand
  - Servers are activated on demand
  - Logical vision of a single server
  - It does not matter which server attends a request

Can we do the same with edge servers?

- How much energy can be saved?
- What is the impact on QoS?
  - Is the advantage facilitated by the proximity to the user lost?



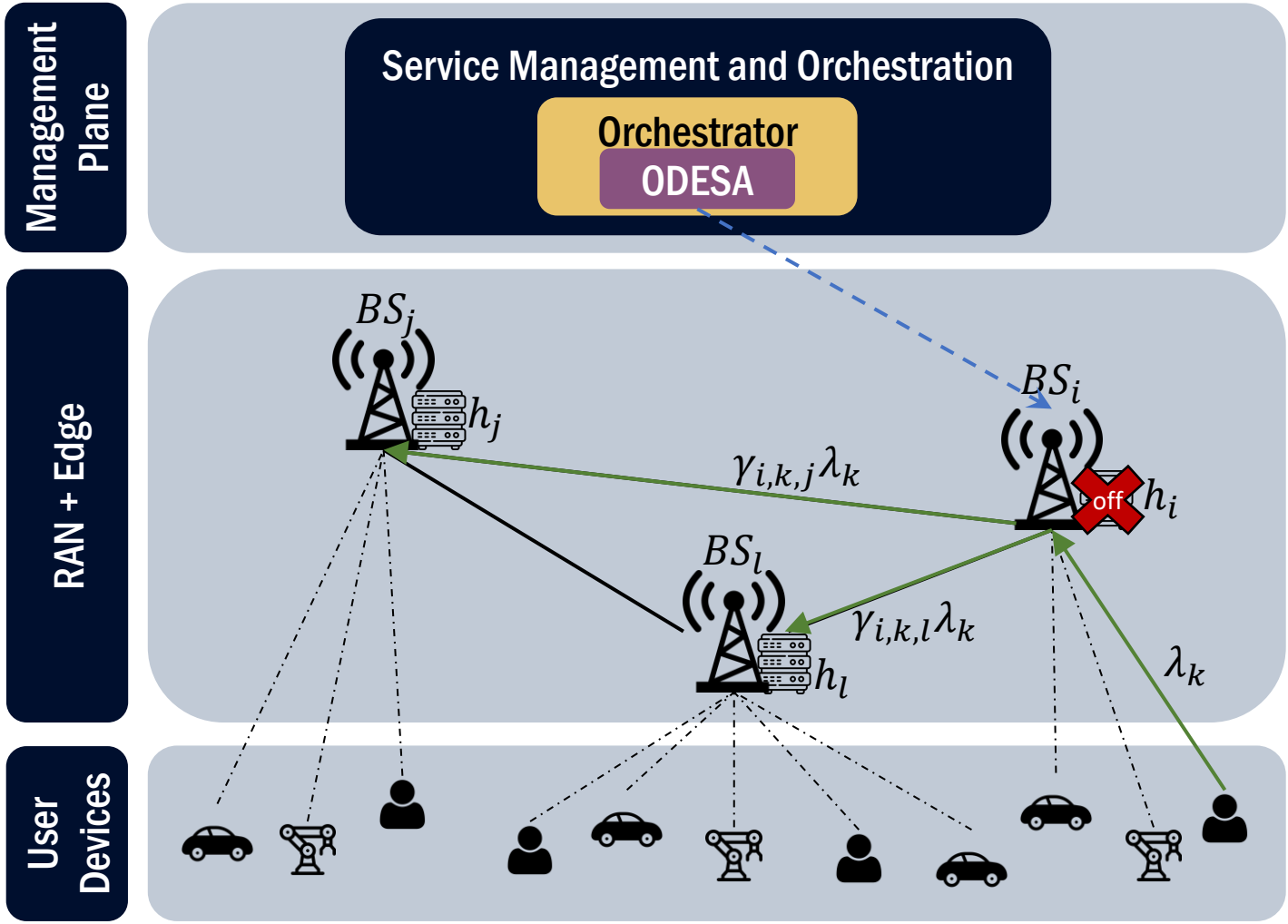


[1]Energy consumption of AMD EPYC 8534P CPU. Standard Performance Evaluation Corporation: SPECpower results.

SYSTEM MODEL

System Model

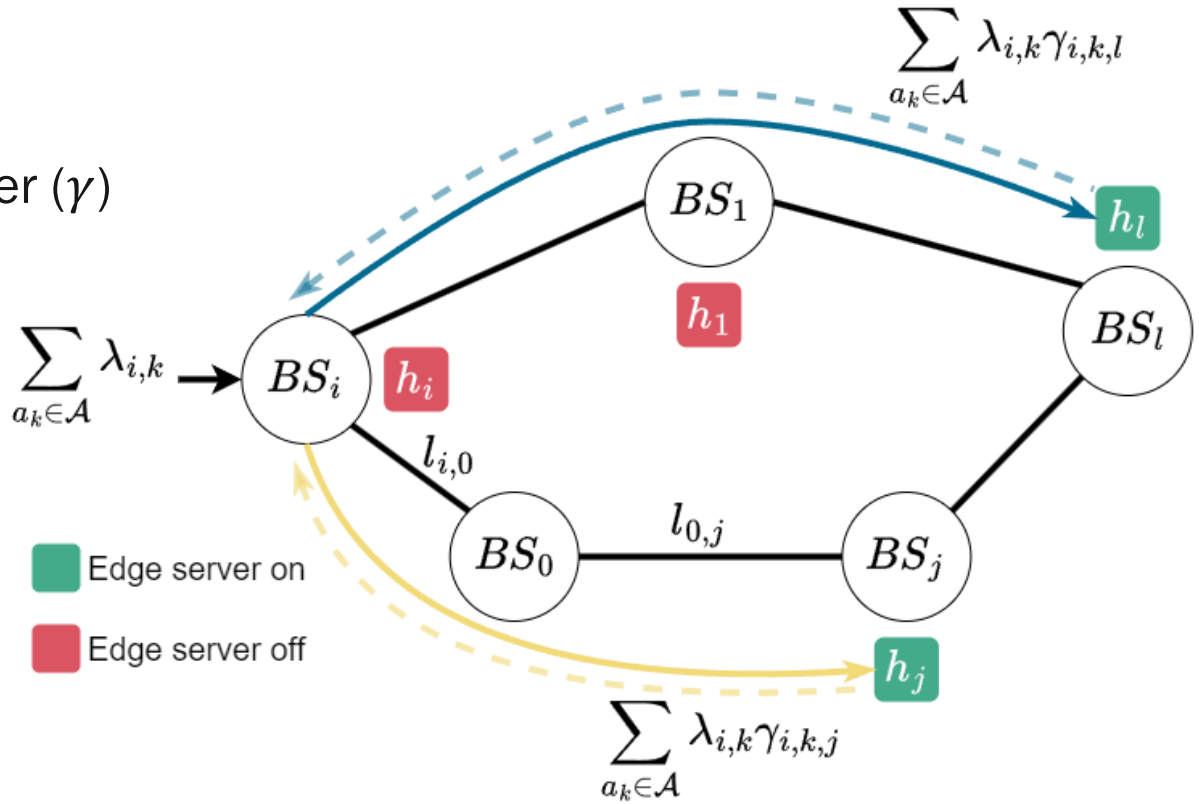
Overview



System Model

Operation

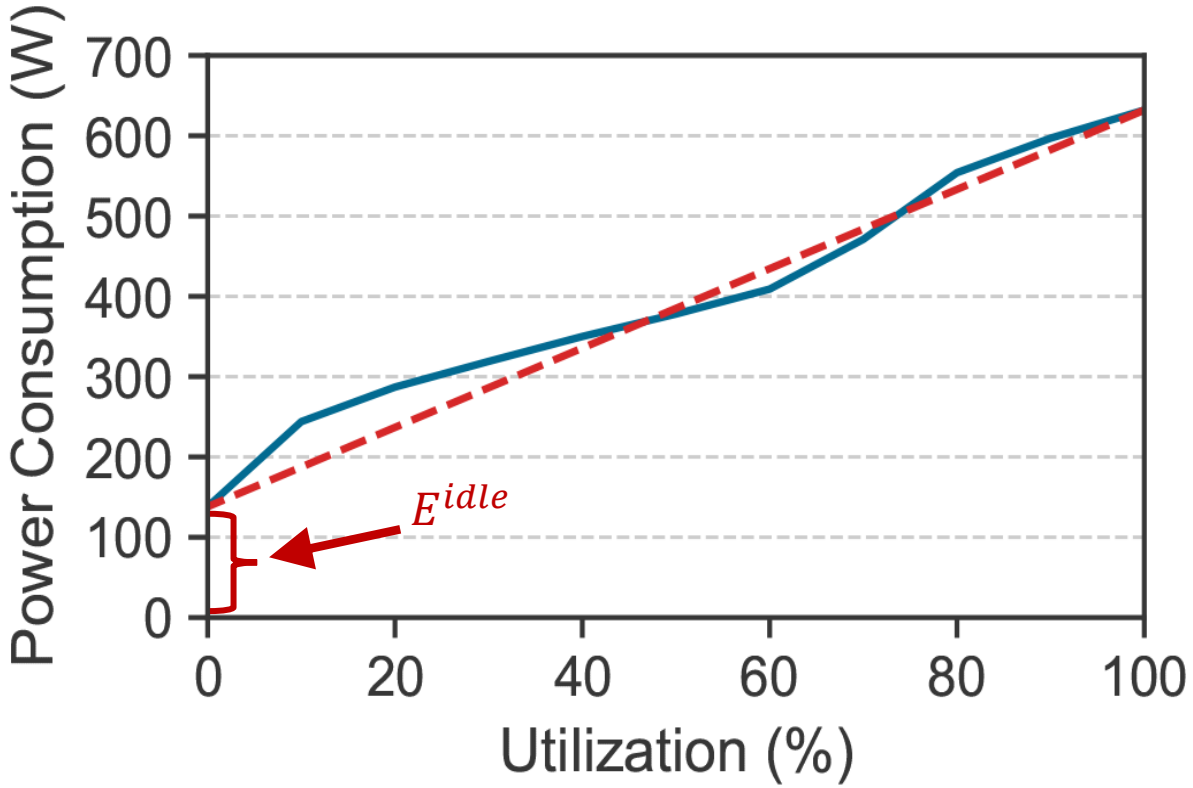
- **Computing requests**
  - Max delay: T_k^{max}
  - Arrival rate: λ_k
- **Orchestrator**
  - Select edge servers that remain active (η)
  - What requests are attended on each server (γ)
- **GOAL: Minimize energy consumption**
  - Edge servers + backhaul
- **Constraints**
  - Computing capacity at edge servers
  - Link capacity
  - Latency (Delay budget)
    - $T < T_k^{max}$



System Model

Energy Model

- **Energy Consumption**
  - Servers: $\sum_{h_m \in \mathcal{H}} \eta_m \left(E_m^{idle} + O_m E_m \right)$
    - *Idle consumption + load (ops/s) x energy per operation (J)*
  - Backhaul: $\sum_{\ell_{o,p} \in \mathcal{L}} \sigma_{o,p} V_{o,p}$
    - *Energy per bit (J) x bits transmitted*



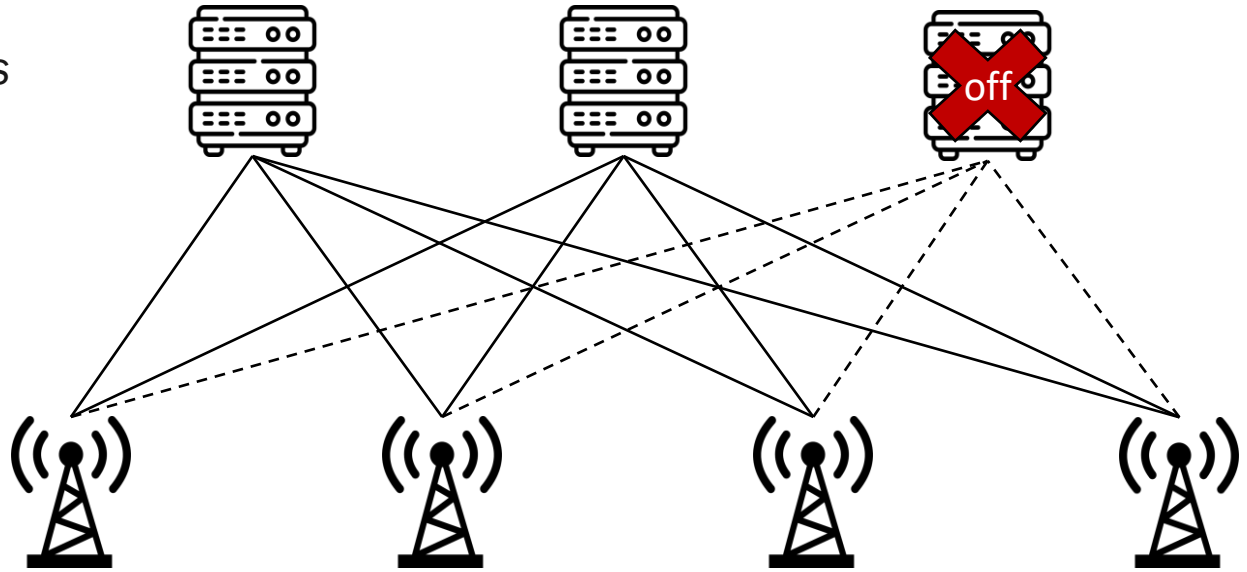
[2]P. Wiesner and L. Thamsen, 'LEAF: Simulating Large Energy-Aware Fog Computing Environments', in *Proc. of IEEE ICFEC*, Melbourne, Australia, May 2021

LOAD DEPENDENT SERVER ACTIVATION

Load Dependent Server Activation

CFLP

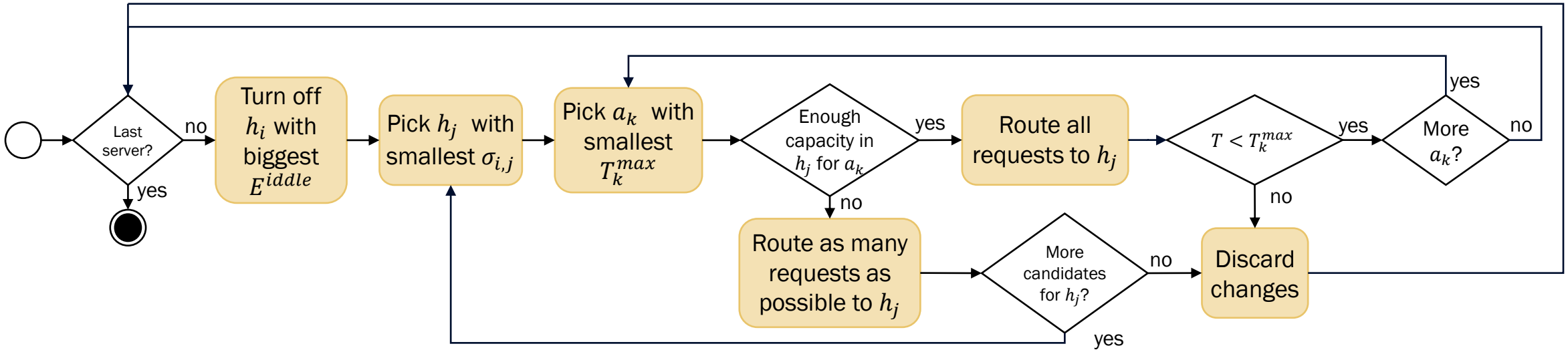
- Capacitated Facility Location Problem
  - Additional constraints:
    - *Multiple services*
    - *Limited capacity of the links*
  - Two stages:
    - *1. Decide the subset of edge servers*
    - *2. Decide the routing of requests*
  - NP-hard ➔ Heuristic solution



Load Dependent Server Activation

ODESA

- DROP-based heuristic
  - Drop servers one by one
- Re-route requests via the most efficient path
  - More efficient links are likely to be also shorter ➔ less latency
    - $\sigma_{i,j} = \sum_{l_{o,p} \in \mathcal{L}} \sigma_{o,p}\, p_{o,p}$
- Allocate services with the smallest delay budget first
- If changes are unfeasible study next server
- If the new configuration uses more energy, changes are discarded too

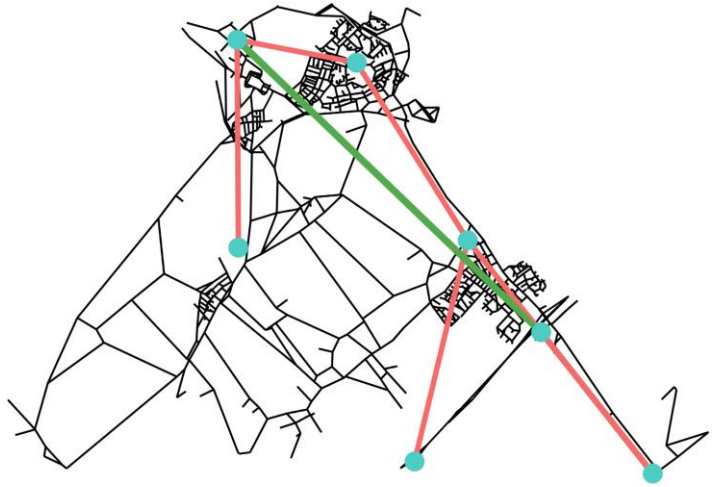


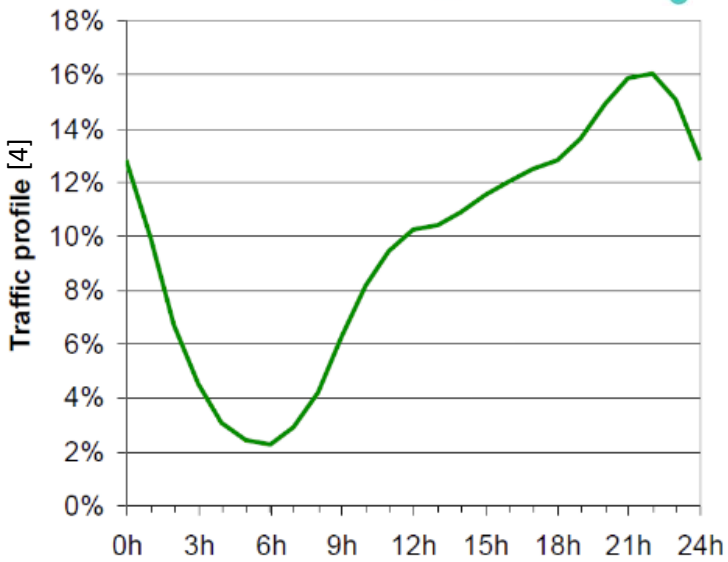
PERFORMANCE EVALUATION

Performance Evaluation

Scenario

- **Dutch MNO in Elburg, Netherlands**
  - All BSs host an edge server
  - Computing capacity scaled to meet demand in peak hours
- **Maximum 3800 users**
  - Random user mobility
  - Arrivals with Poisson distribution $\lambda_k = 15$ req/s
- **ODESA vs. Baselines:**
  - Always On
  - Sleep below 10% load threshold
- **Vehicular safety service**
  - Delay budget 5 ms



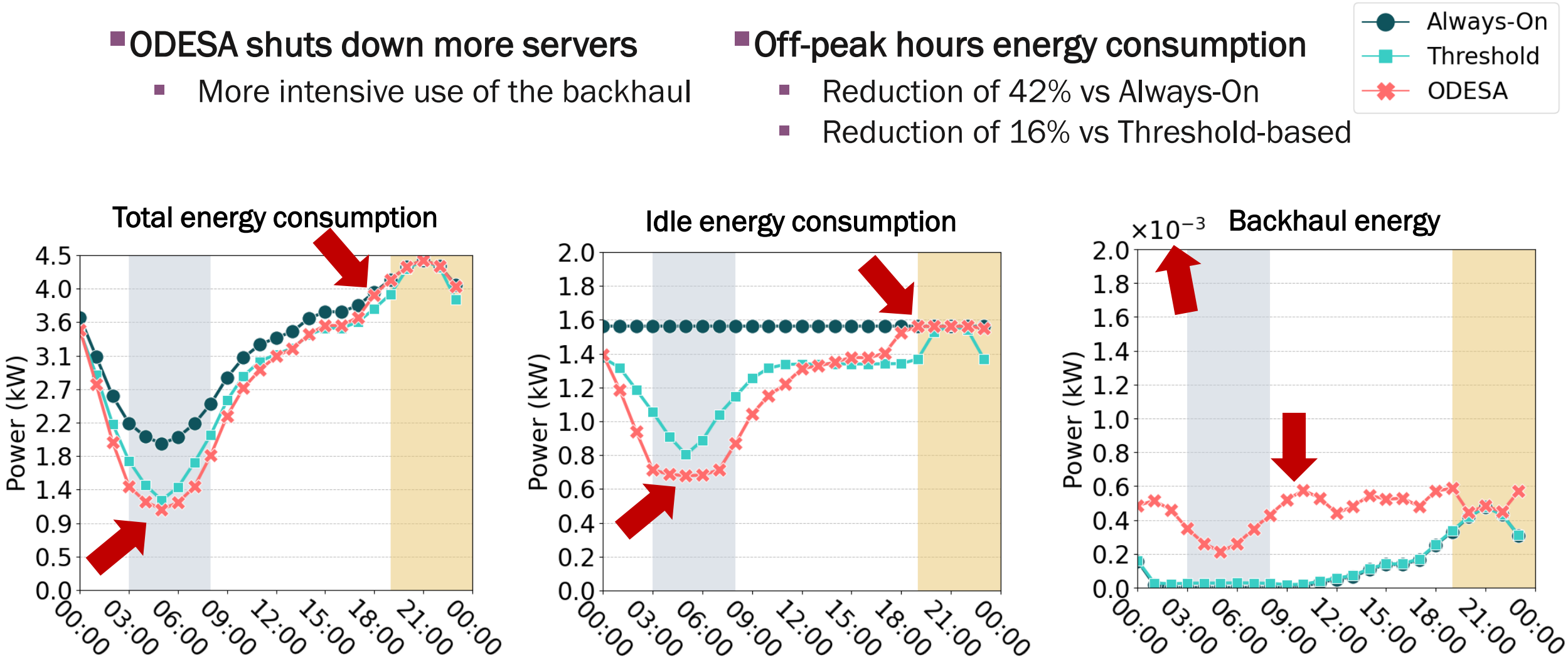


[3] METIS-II: Deliverable D2.3 Performance evaluation results

Performance Evaluation

Energy Consumption

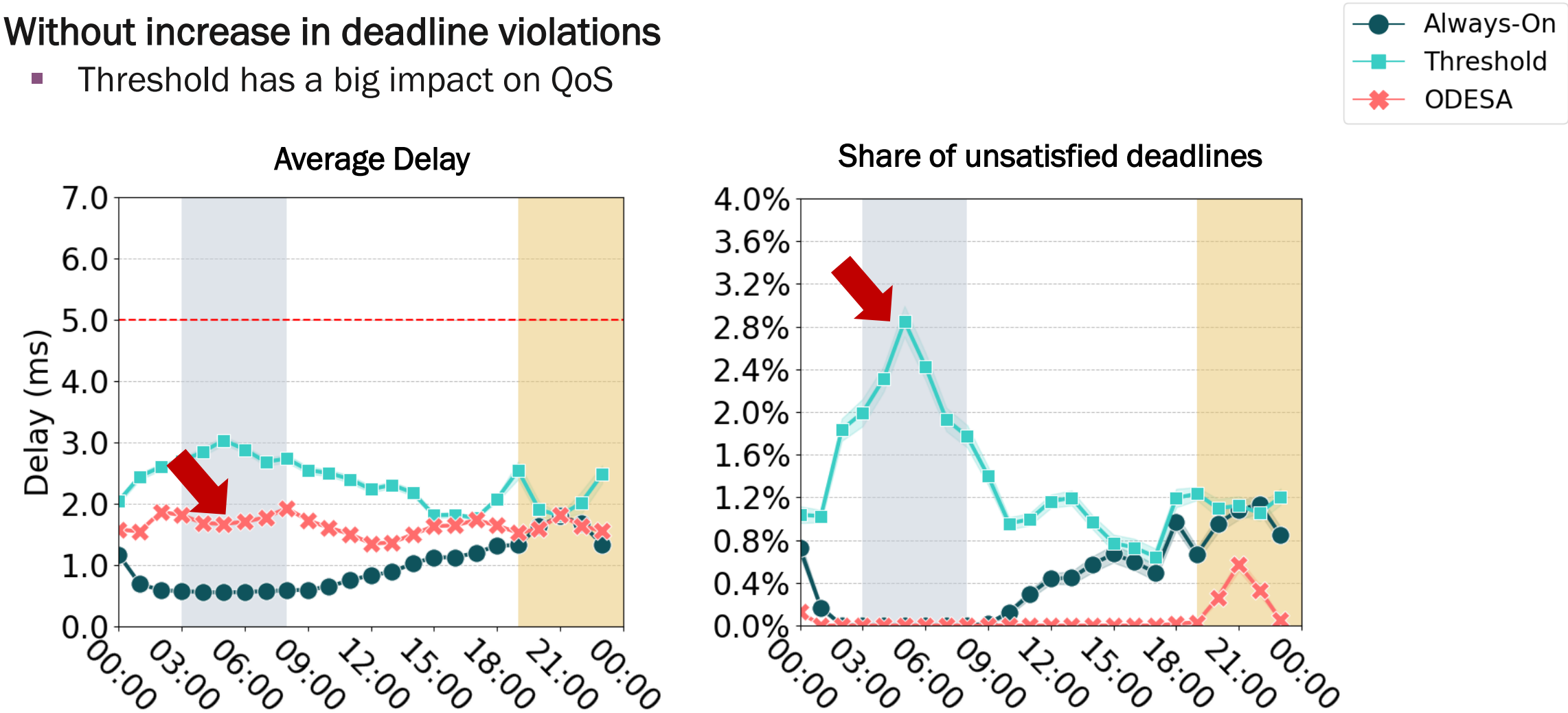
- **ODESA shuts down more servers**
  - More intensive use of the backhaul
- **Off-peak hours energy consumption**
  - Reduction of 42% vs Always-On
  - Reduction of 16% vs Threshold-based



Performance Evaluation

Energy Consumption

- Without increase in deadline violations
  - Threshold has a big impact on QoS



CONCLUSION

Conclusion

- ODESA reduces the energy consumption of edge computing
- ODESA does not cause significant QoS degradation
- Server shutdown has the potential to be a cost-effective solution to improve the energy efficiency of edge computing infrastructure
- Future work:
  - Proactive shutdowns and activations
  - Evaluation in bigger scenarios
    - *Savings in current results limited by small number of servers*
  - How do different server distributions affect the behavior of ODESA?

ODESA: Load-Dependent Edge Server Activation for Lower Energy Footprint

Blas Gómez[1,*], Suzan Bayhan[2], Estefanía Coronado[1,3], José Villalón[1] and Antonio Garrido[1]

[1]High-Performance Networks and Architectures, Universidad de Castilla-La Mancha, Albacete, Spain

[2]Faculty of EEMCS, University of Twente, Enschede, The Netherlands

[3]I2CAT Foundation, Barcelona, Spain

*blas.gomez@uclm.es

EUROPEAN UNION
European Social Fund
2019-PREDUCLM-10921