



Dynamic Heuristic Neuromorphic Solver for the Edge User Allocation Problem with Bayesian Confidence Propagation Neural Network

Kecheng Zhang², Anders Lansner^{1,5}, **Ahsan Javed Awan**², Naresh Balaji Ravichandran², Pawel Herman^{1,3,4}

1. KTH Royal Institute of Technology, 2. **Ericsson Research**, 3. Digital Futures, 4. Swedish e-Science Research Centre, 5. Stockholm University



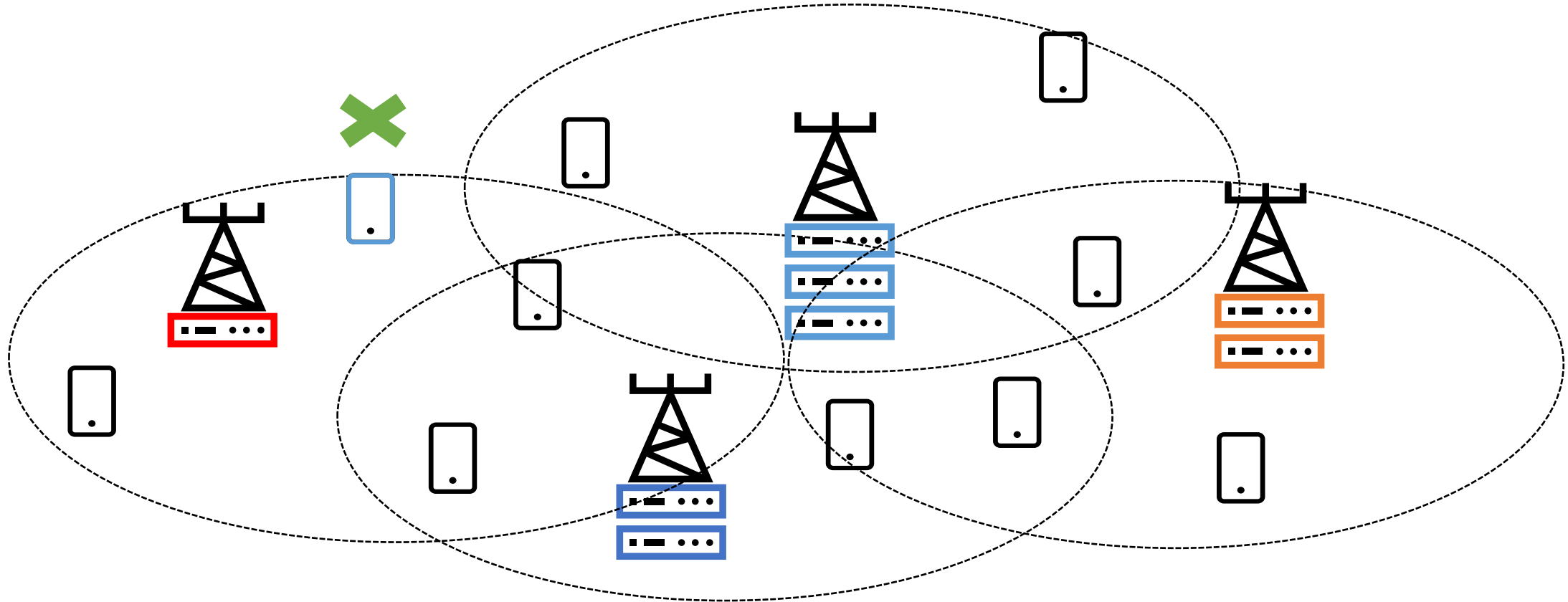
Optimal Allocation of Users to Edge Servers



Require Computational Offloading to Edge Servers



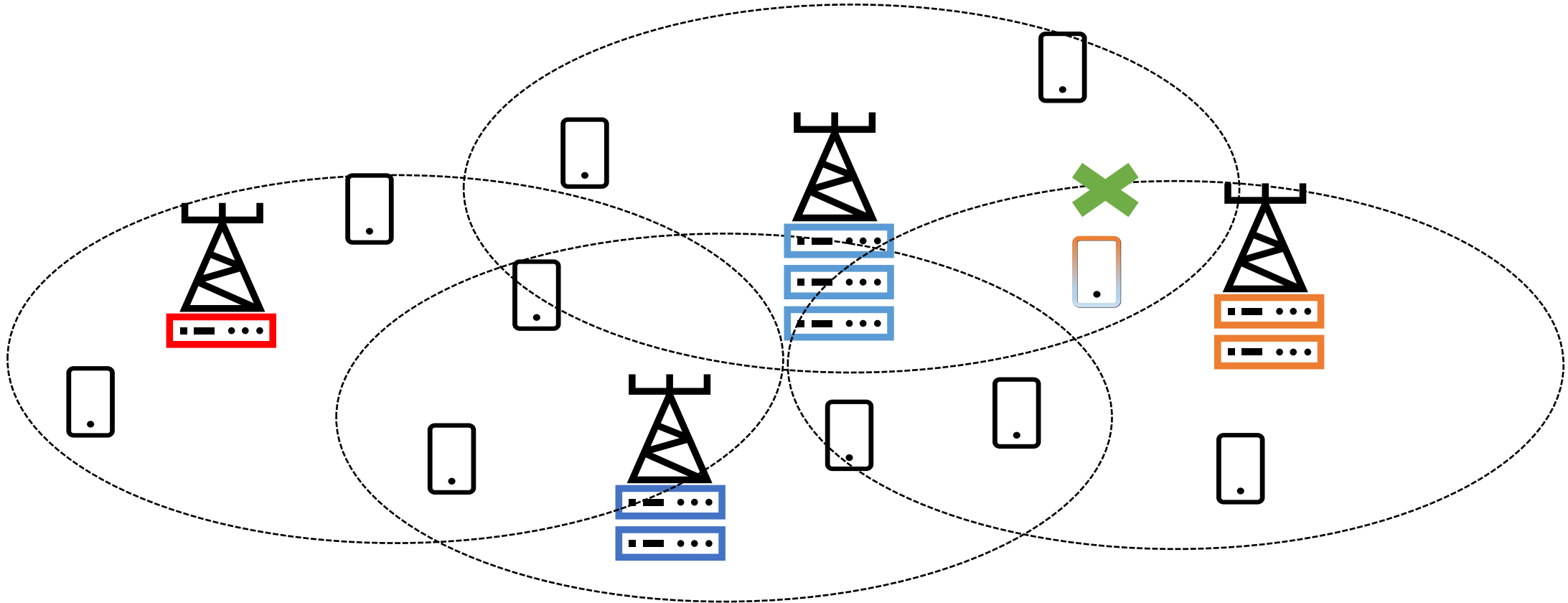
Coverage constraint



Only users located within the coverage of an edge server can be allocated to the edge server



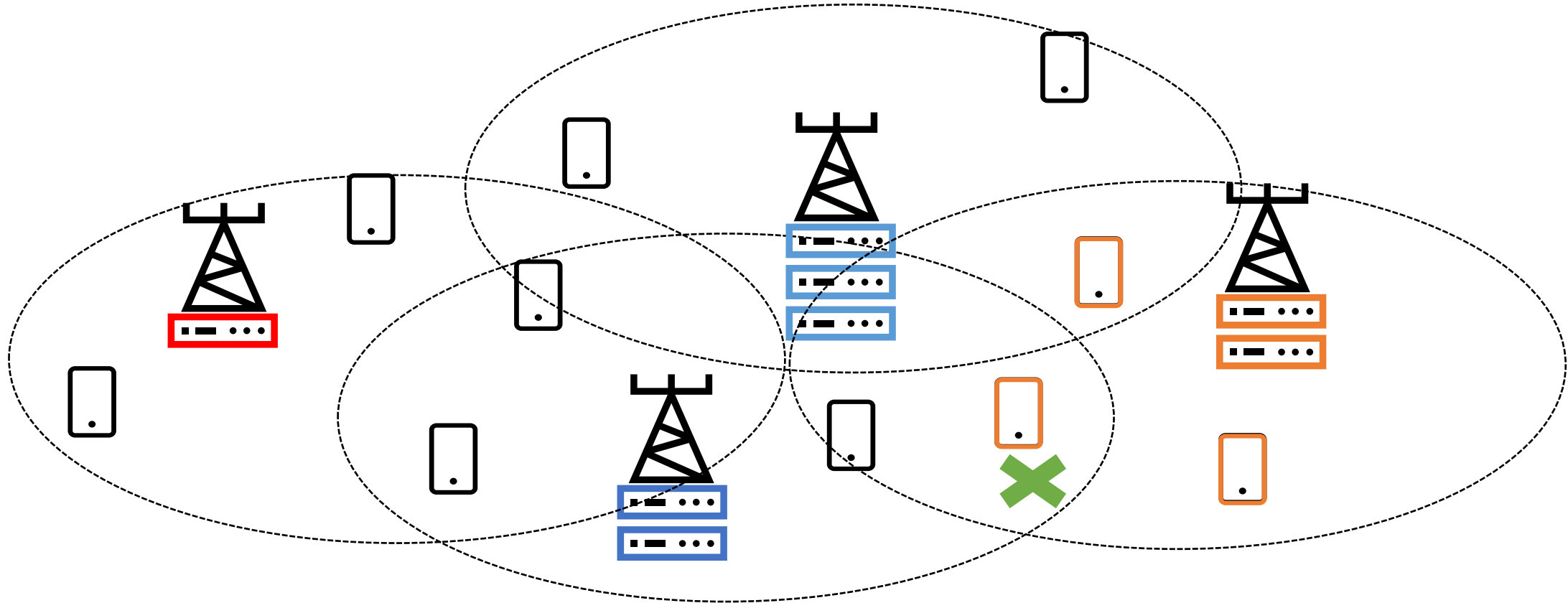
At max single association constraint



Every user is allocated to at most one edge server



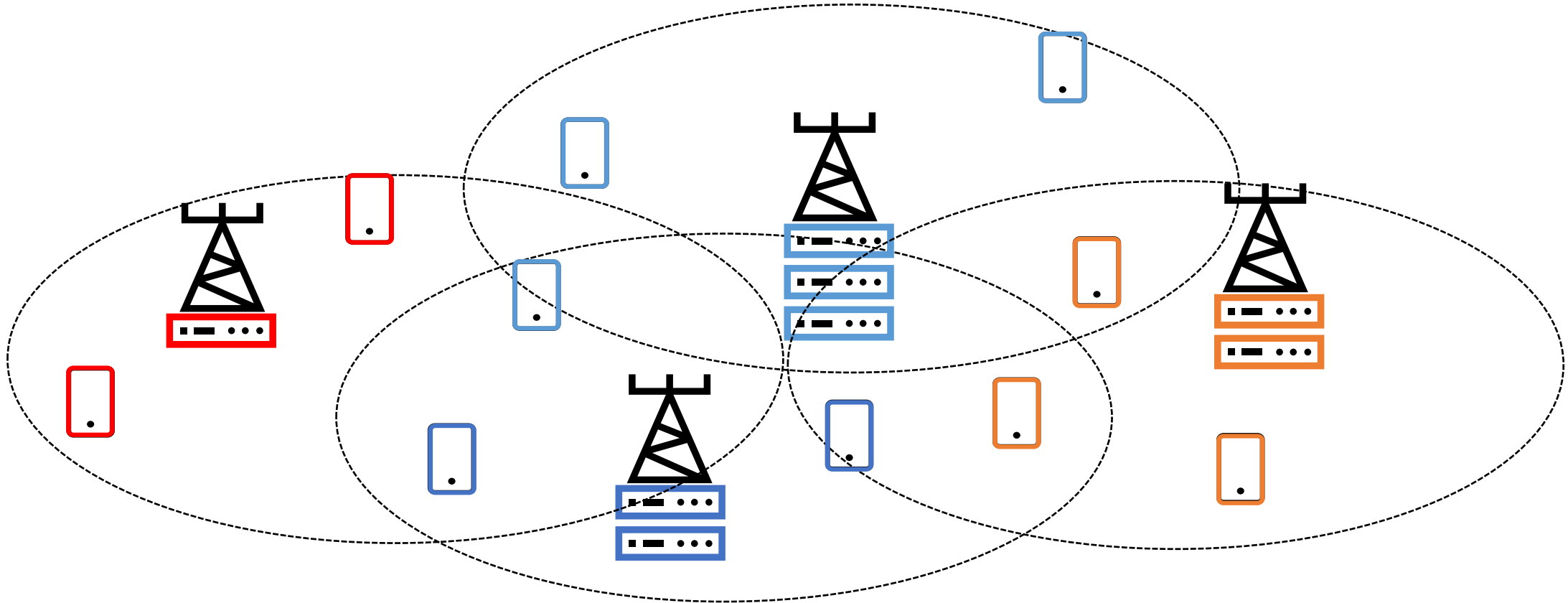
Capacity constraint



Total workload per resource must not exceed its server's remaining capacity

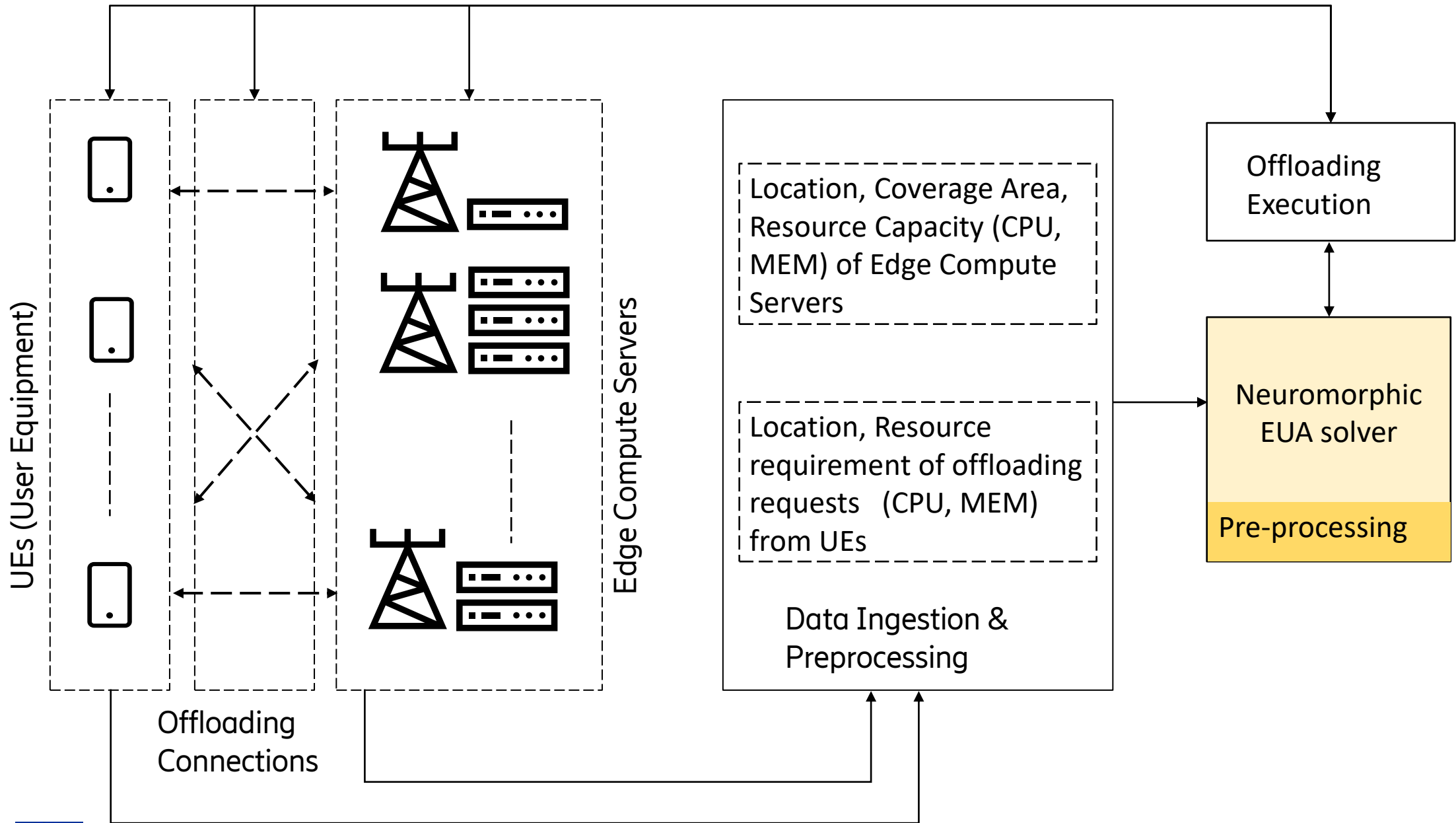


Multi objective function



Maximize the number of users served by the minimum number of edge servers





Mathematical Formulation

Integer Linear Program

$$x_{ij} = \begin{cases} 1 & \text{if user } i \text{ is assigned to server } j \\ 0 & \text{otherwise} \end{cases}$$

$$y_j = \begin{cases} 1 & \text{if server } j \text{ is used} \\ 0 & \text{otherwise} \end{cases}$$

Objective Function: Maximize the number of M users served by the minimum number of N edge servers

$$-\gamma_A \sum_{\forall i \in \text{users}} \sum_{\forall j \in \text{servers}} x_{ij} + \gamma_B \sum_{\forall j \in \text{servers}} y_j$$

Constraints:

Single Association: $\sum_{\forall j \in \text{servers}} x_{ij} \leq 1 \quad \forall i \in \text{users}$

Capacity: $\sum_{\forall i \in \text{users}} w_i^k x_{ij} \leq C_j^k y_j \quad \forall j \in \text{servers}, \forall k \in \text{resource type}$

Coverage: $d_{ij} x_{ij} \leq \text{cov}(j) \quad \forall j \in \text{servers}, \forall i \in \text{users}$

Decision Variables: (M+1)N

Quadratic Unconstrained Binary Optimization

Applying the penalty multipliers to penalize the configuration of decision variables violating the constraints.

$$E = -\gamma_A \sum_{\forall i \in \text{users}} \sum_{\forall j \in \text{servers}} x_{ij} + \gamma_B \sum_{\forall j \in \text{servers}} y_j$$

$$+ \lambda_1 \sum_{\forall i \in \text{users}} \left(\sum_{\forall j \in \text{servers}} x_{ij} + l_i - 1 \right)^2$$

$$+ \lambda_2 \sum_{\forall j \in \text{servers}} \left(\sum_{\forall i \in \text{users}} w_i^k x_{ij} + m_j - C_j^k y_j \right)^2$$

$$+ \lambda_3 \sum_{\forall i \in \text{users}} \sum_{\forall j \in \text{servers}} \left(d_{ij} x_{ij} + n_{ij} - \text{cov}(j) \right)^2$$

$$E = \mathbf{x}^T \mathbf{Q} \mathbf{x}$$

$$\mathbf{x} = \{x_{ij}, y_{ij}, q_{ij}^p, q_j^p, q_i^p : \forall i \in \text{users} \wedge \forall j \in \text{servers} \wedge \forall p \in R\}$$

QUBO for is obtained by discretizing the slack variables l_i , m_j and n_{ij} , into binary form as under, we get

$$n_{ij} = a \sum_{p=0}^{R-1} 2^{-p} q_{ij}^p - b \in [-b, 2a - b]$$

Binary Decision Variables: (M+1)N+ R(M+N+MN)

Grant no.101135809 – EXTRA-BRAIN

Problems with Prior Art

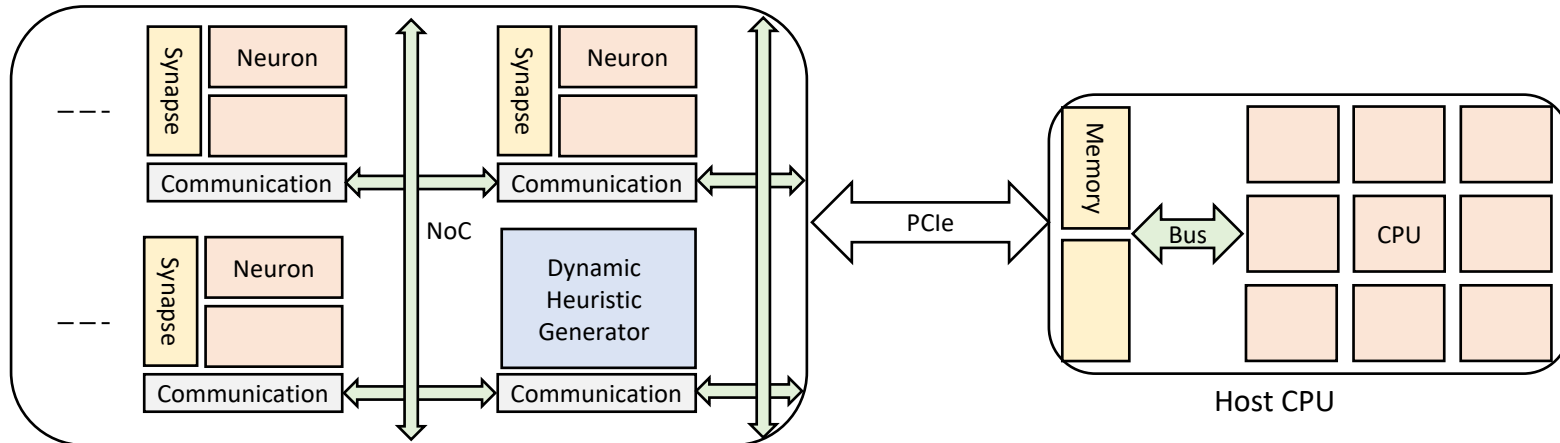
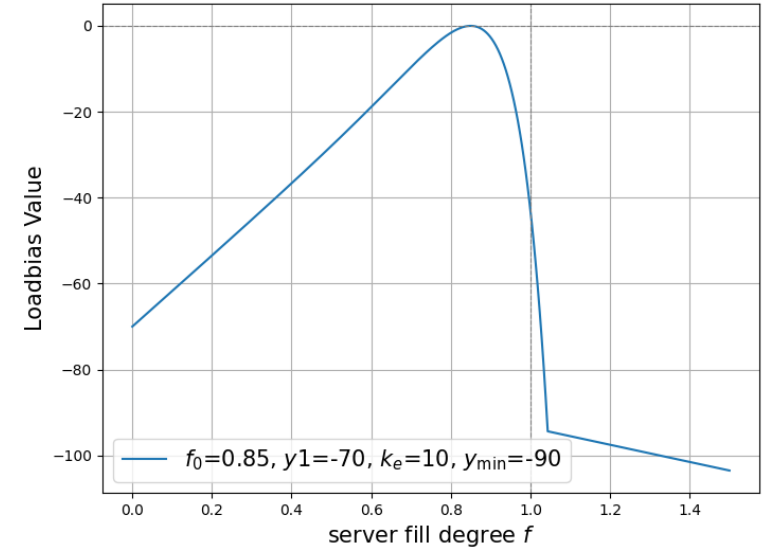
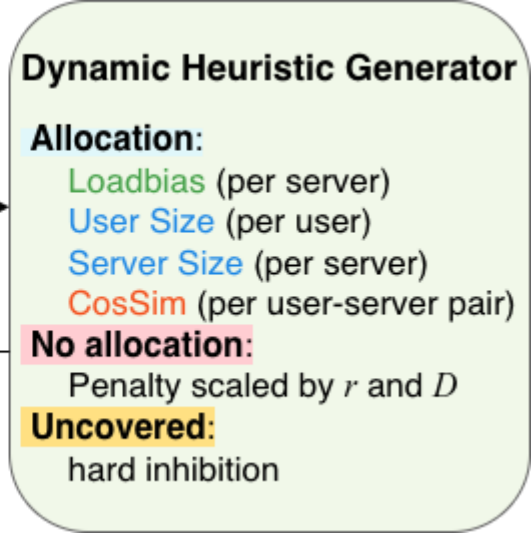
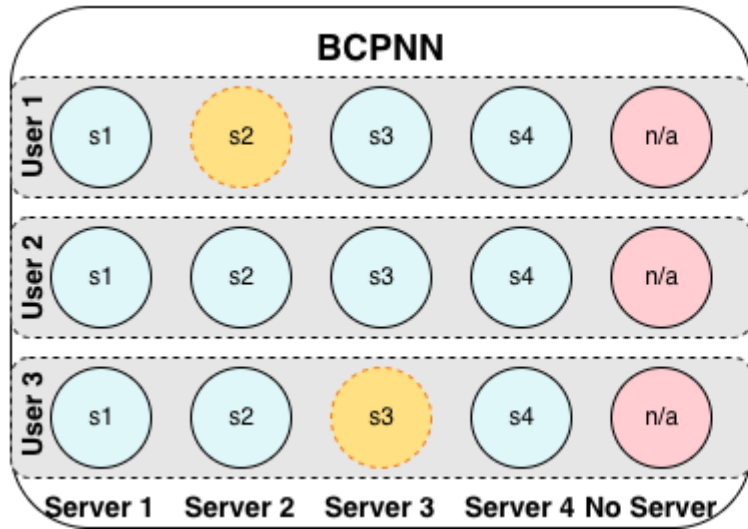
WTA circuits do not encode
"at max" constraint

Dynamic scenarios require re-
computation of the QUBO
matrix (updated penalty
coefficients)

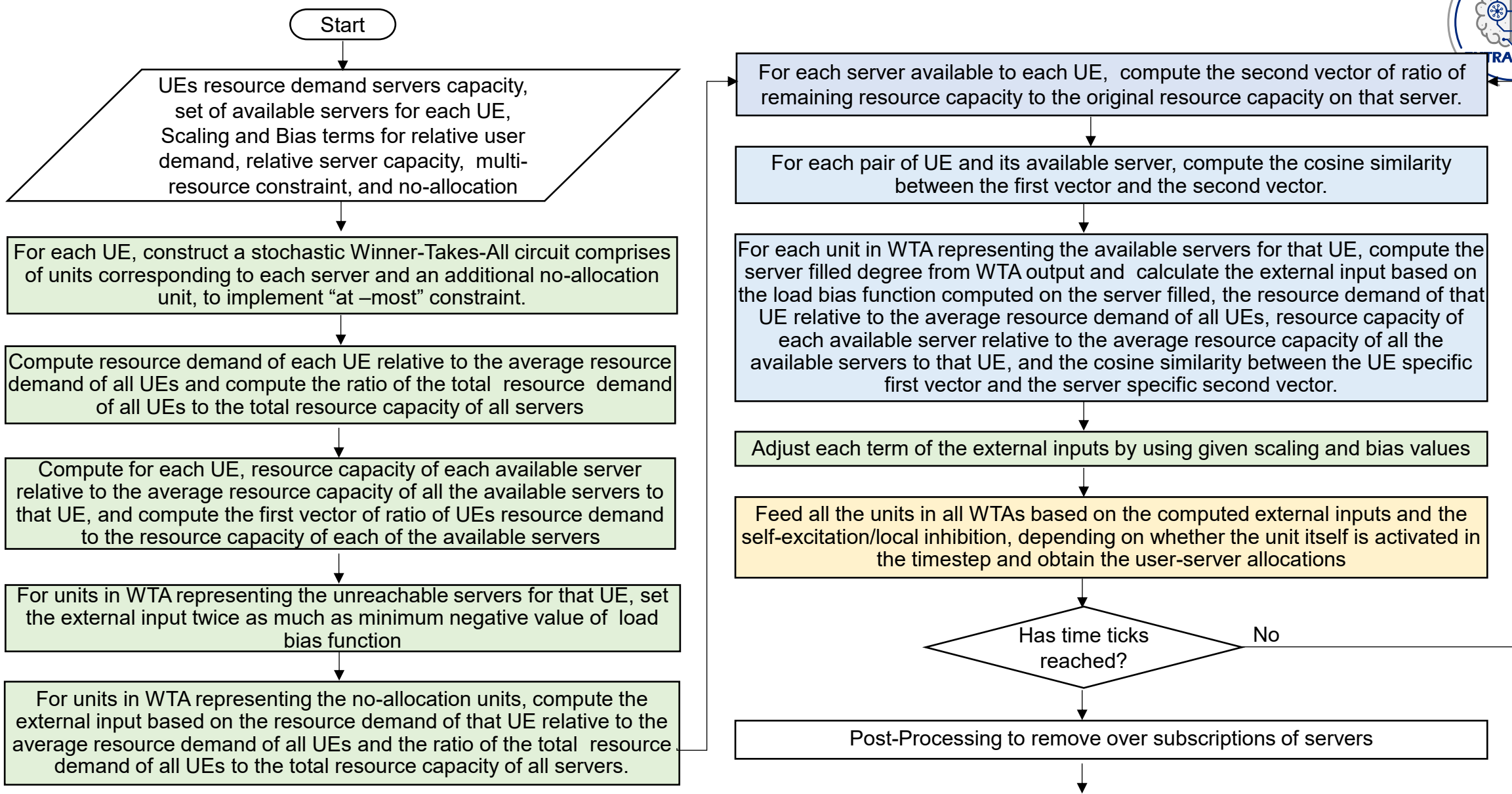
Metaheuristic parameter tuners
for stochastic SNNs are
compute and memory intensive

QUBO solvers underperform
compared to Gurobi when
solving unconstrained
reformulations of ILP problems

Our Solution



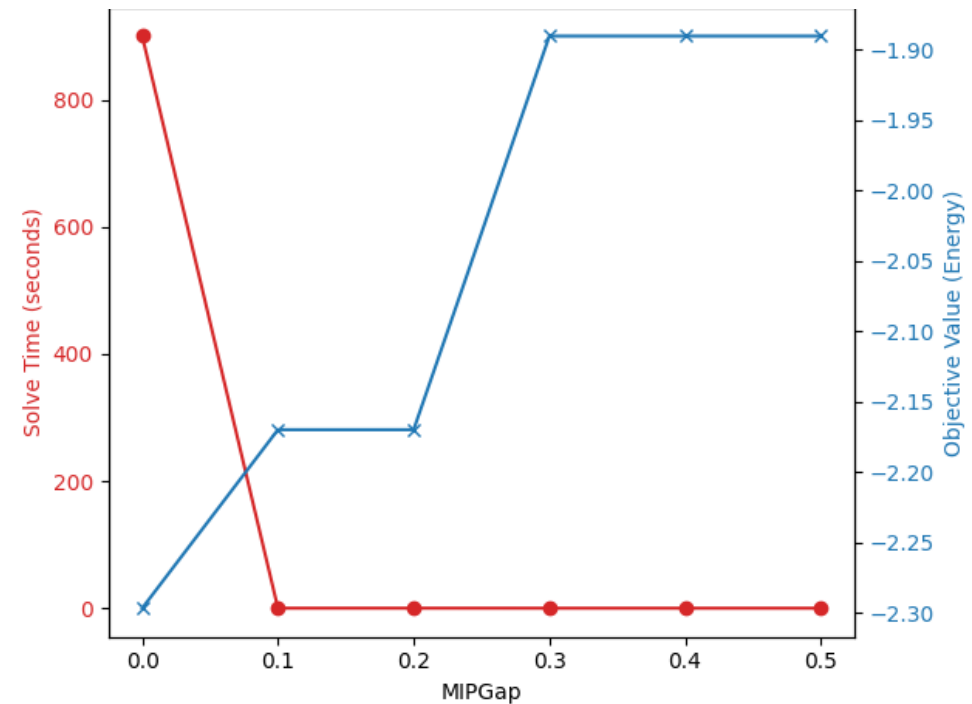
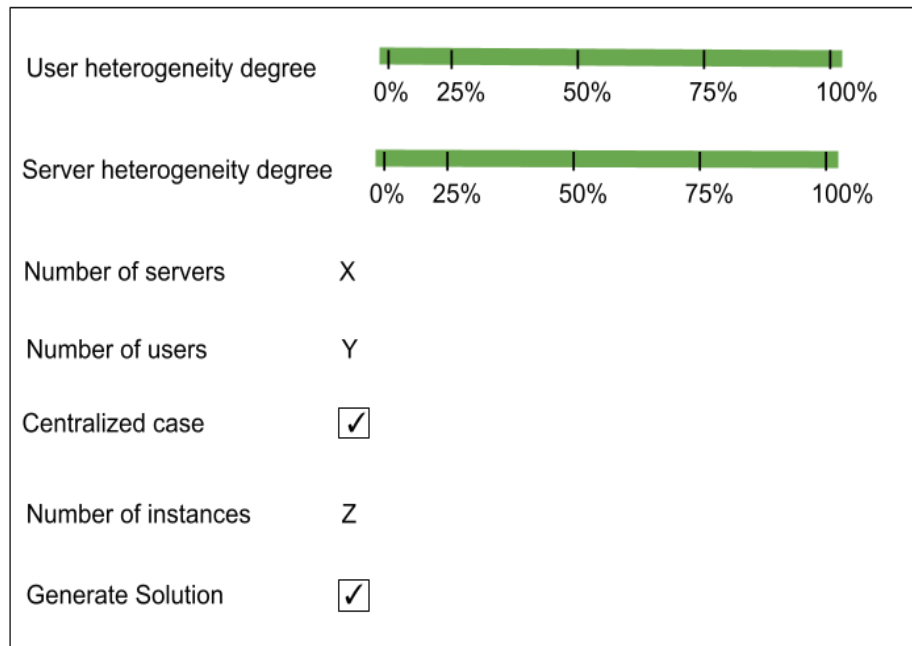
- A hybrid solver exploiting the EUA problem characteristics to compute deterministic excitation/inhibition for individual units of BCPNN inspired stochastic WTAs
- Requires HW support for Deterministic Excitation Unit



Problem Instances

- A synthetic problem-instance generator based on open-source [EUA Data Set](#) covering centralized and decentralized scenarios, each solved with Gurobi MILP solver with default parameters and 15 mins time-out limit.

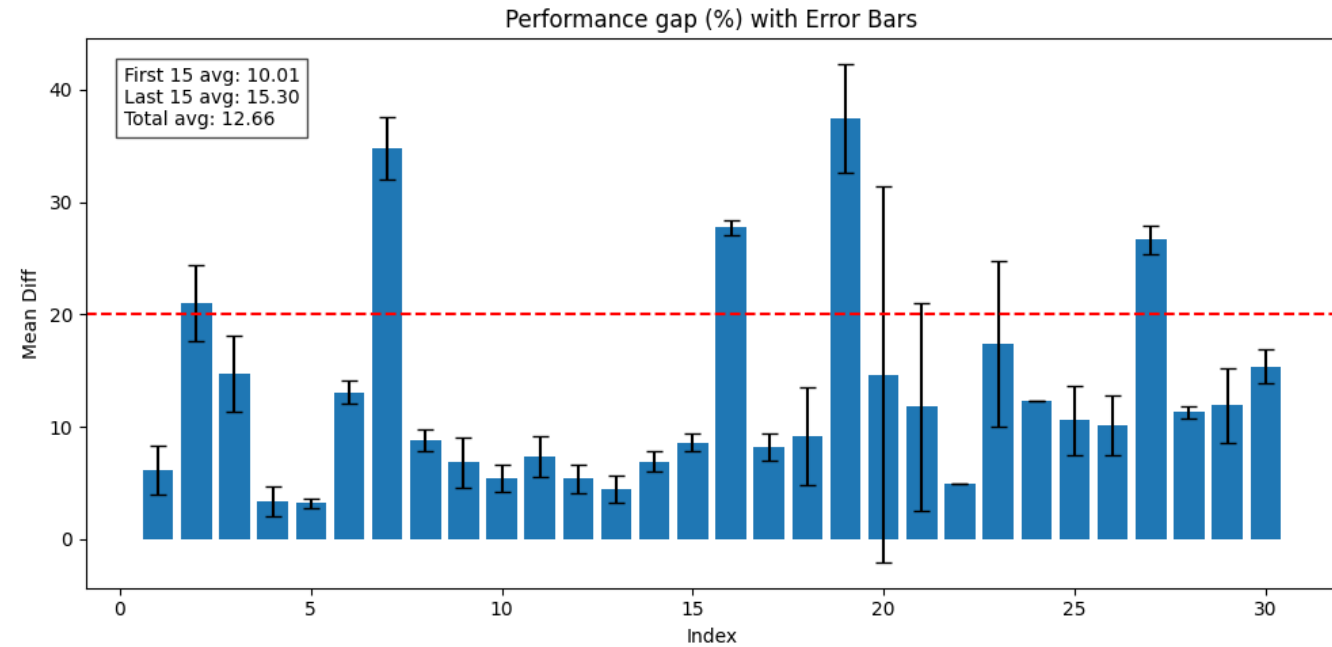
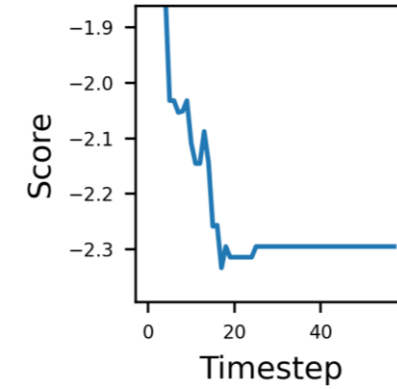
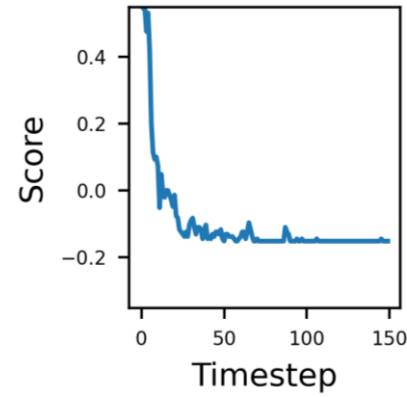
Problem Instance Generator



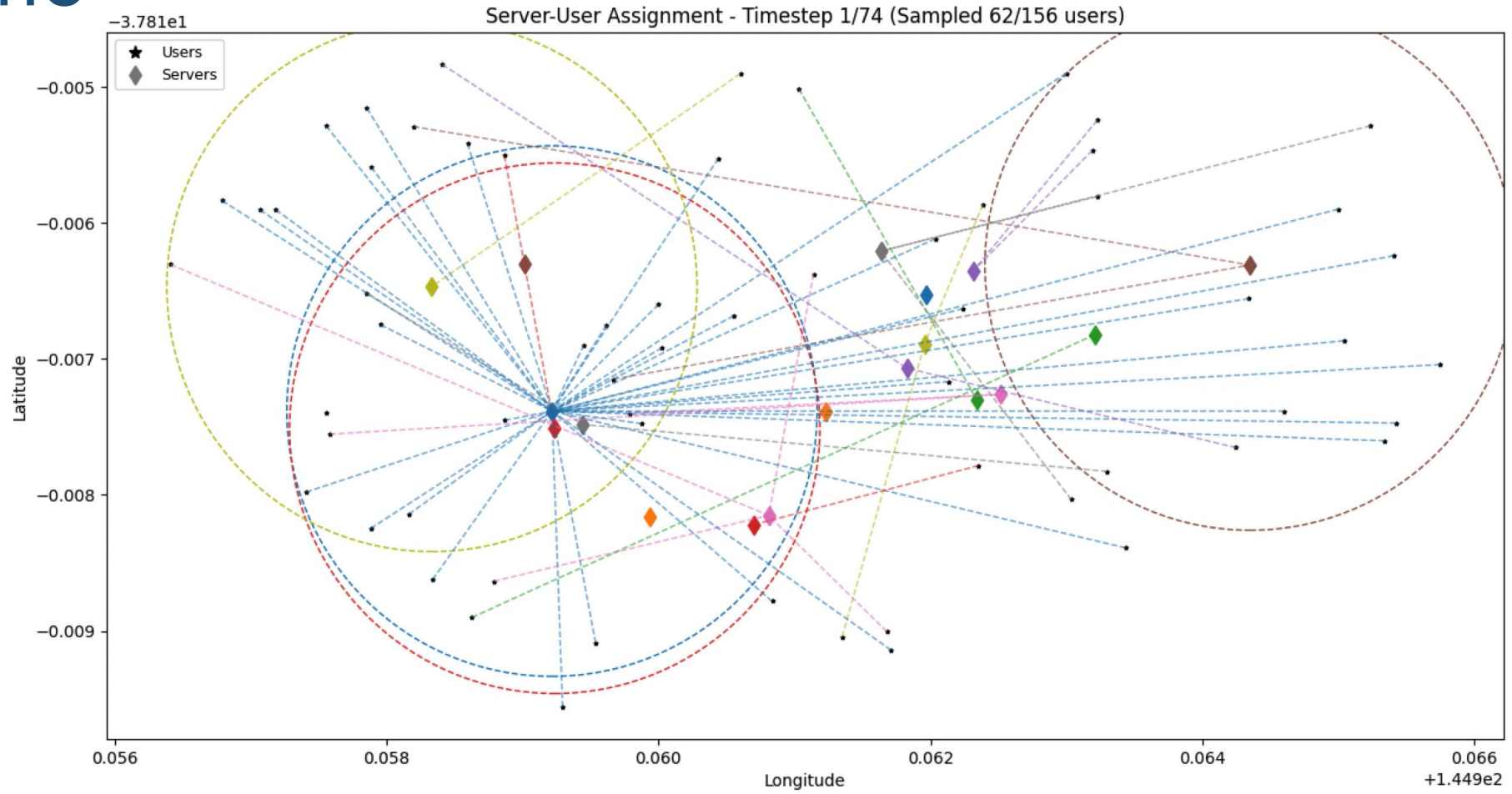
Results



Index	Users (n)	Servers (m)	User demand Heterogeneity	Server resource heterogeneity	Edge Compute Deployment
1	113	10	0.25	0.00	Distributed, City
2	111	12	0.50	0.25	
3	94	14	0.25	0.50	
4	149	16	0.5	1.00	
5	156	18	1.00	1.00	
6	145	20	0.75	1.00	
7	287	22	0.50	0.00	
8	208	24	0.75	0.75	
9	165	26	0.25	0.25	
10	173	28	0.75	0.75	
11	209	30	0.5	0.25	
12	179	32	0.25	0.00	
13	255	34	1.00	1.00	
14	239	36	0.00	1.00	
15	352	38	1.00	1.00	
16	435	10	0.75	0.50	Centralized City
17	576	12	0.75	0.00	
18	416	14	1.00	0.50	
19	256	16	0.75	0.00	
20	490	18	1.00	0.25	
21	809	20	0.00	0.75	
22	160	7	0.25	0.25	
23	224	8	0.75	1.00	
24	268	9	0.00	0.75	
25	592	11	0.75	0.50	
26	464	13	1.00	0.75	
27	624	15	0.50	0.75	
28	448	17	0.25	0.5	
29	693	19	0.5	0.75	
30	672	21	1.00	0.00	



Demo



(x, y) = (144.958938, -37.818691)

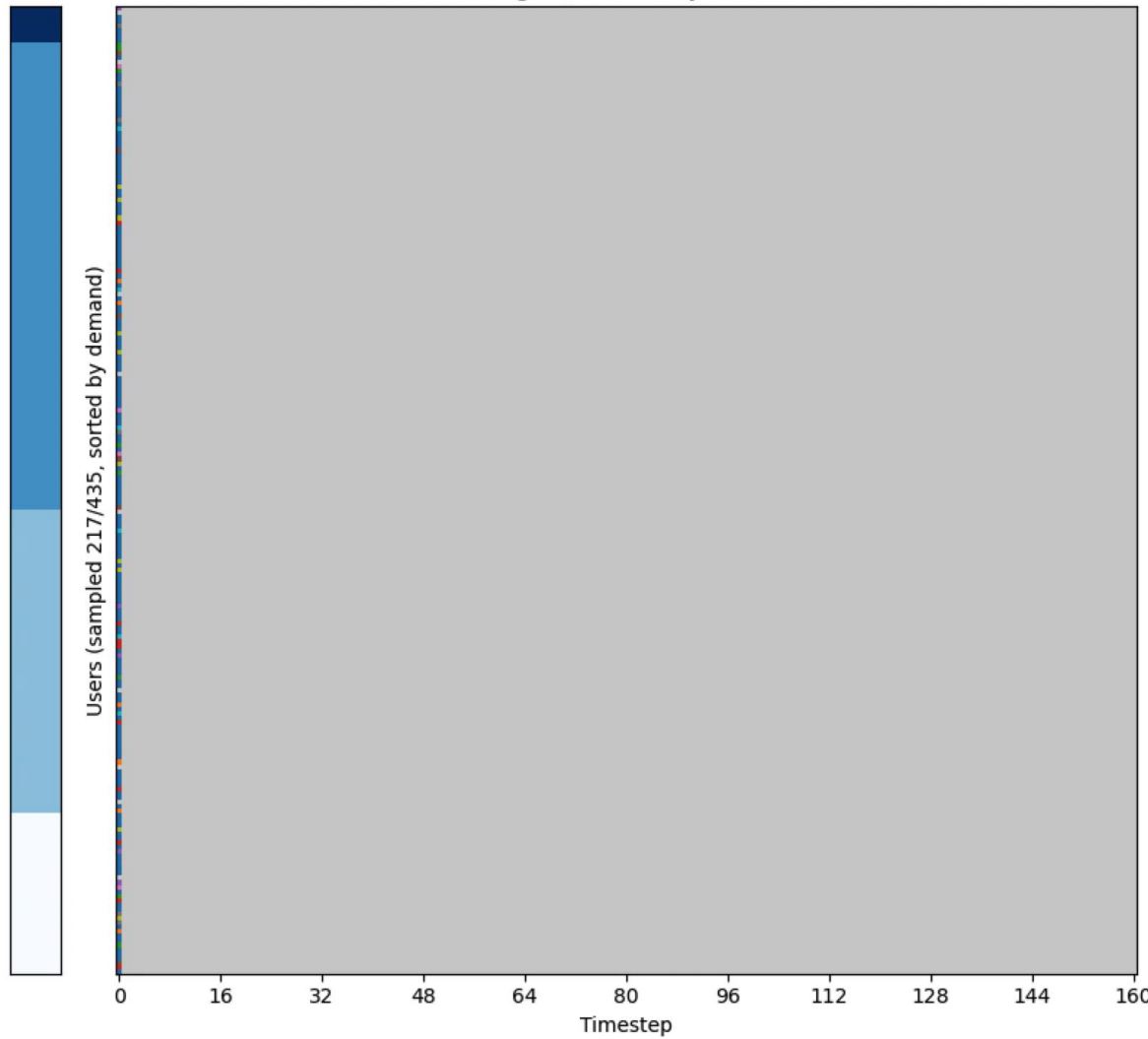


Demo



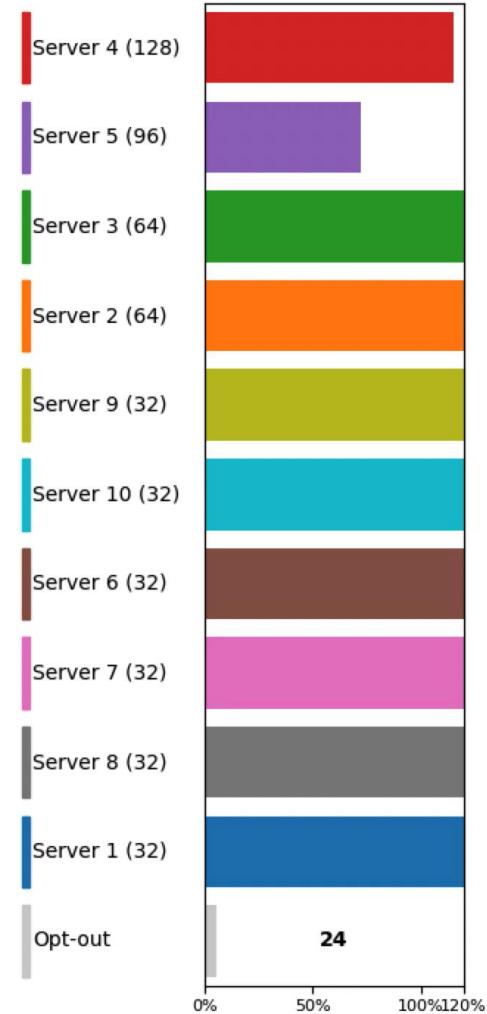
User Demand

Assignment History: 1/161



Legend

Fill Degree



(x, y) = (64.5, 52.5)
[10.0]



Demo



User 0 — Timestep 1/56 (Score: 0.503448)

