Tango: Harmonious Management and Scheduling for Mixed Services Co-located among Distributed Edge-Clouds

Yicheng Feng, Shihao Shen, Mengwei Xu, Yuanming Ren,
Xiaofei Wang*, Victor C. M. Leung, Wenyu Wang

天津大学 Tianjin University
北京郵電大學 Beijing University of Posts and Telecommunications
PPIO
THE UNIVERSITY OF BRITISH COLUMBIA

ACM 52nd International Conference on Parallel Processing (ICPP), Salt Lake City, Utah, USA - August 7-10, 2023

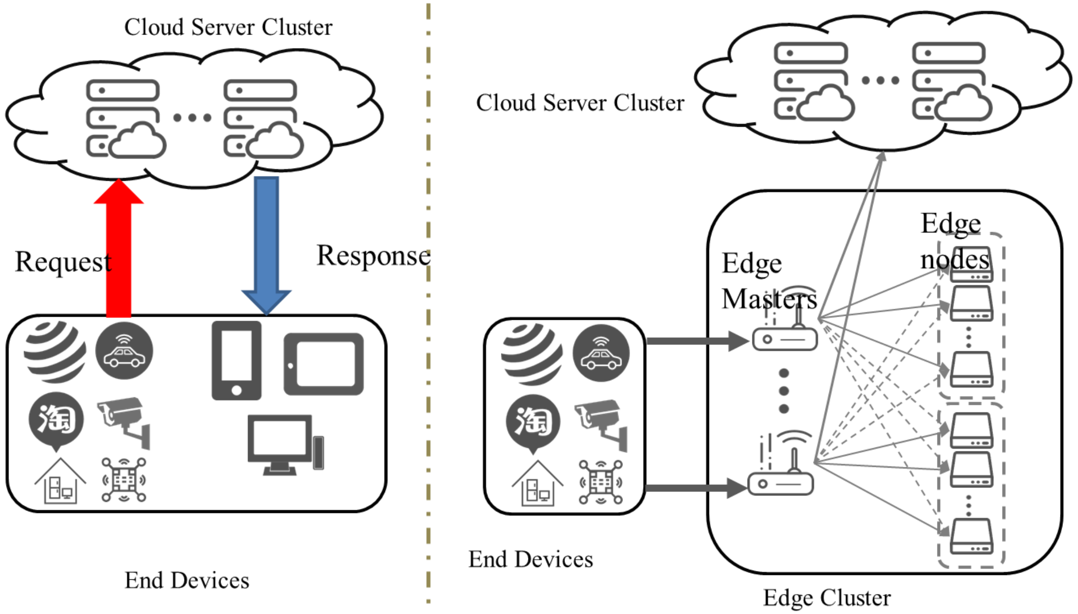
Contents

1.1 Emergence of edge computing

- ***Low latency***
  - Computing resources are deployed on edge nodes close to end devices to achieve faster response time...

- ***Bandwidth saving***
  - Data processing and analysis are performed at the edge of the network to reduce the demand for backbone network bandwidth.

- ***Data privacy***
  - Sensitive data can be processed and stored on edge devices to reduce the risk of data during transmission.



Centralized Cloud Computing

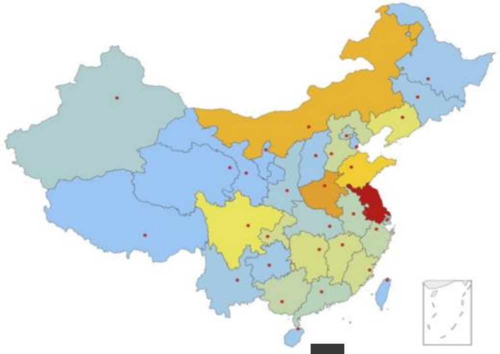
Distributed Edge Computing

1.2 Co-location of mixed services

- ***Latency-Critical (LC) service***
  - Exhibiting elevated quality of service requirements, such as low latency, and any compromise in service quality would violate SLAs, resulting in financial repercussions.

- ***Best-Effort (BE) service***
  - Exhibiting the capability to tolerate higher operational latencies and facilitate the recovery of failed services through restart mechanisms.

- ***Potential for service co-location in edges***
  - Individually hosting LC services in the edge introduces severe resource underutilization.
  - Our measurements indicate that the average utilization of edge-cloud resources is below 20%.

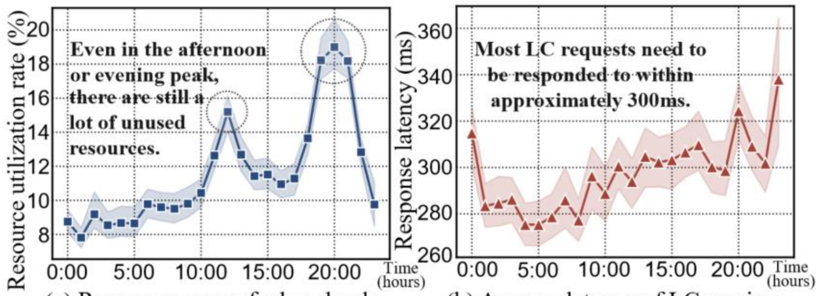


PPIO

200+ cities covered

1,000+ regions covered

10,000+ edge servers

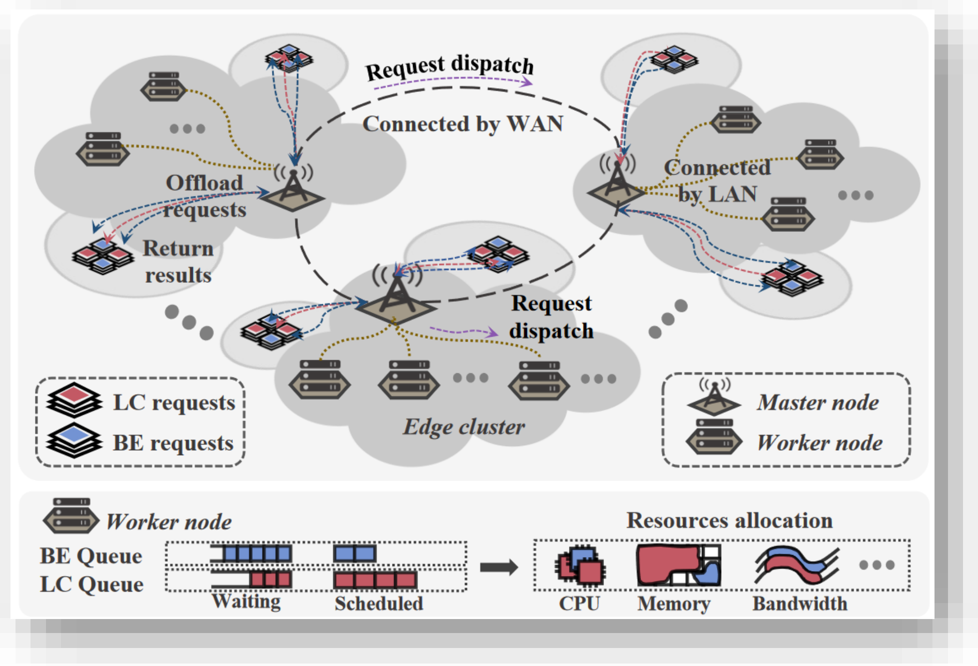


(a) Resource usage of edge clouds. (b) Average latency of LC services.

It's promising to fill the **resource usage gap** by **co-locating mixed services** on the same edge-cloud.

1.3 Unique challenges in edge-clouds

- ***Heterogeneous and resource-constrained***
  - Edge clouds are typically heterogeneous and resource-constrained, which makes competition for resources more fierce.
- ***Distributed and scattered resources***
  - Edge resources are distributed and scattered, yet loads of user requests are uneven and fluctuating across geographic locations.
- ***Divergent service quality demands***
  - How to Optimize traffic scheduling and resource allocation based on the characteristics of BE and LC services?



Co-located services in edge-clouds

1.4 Related work

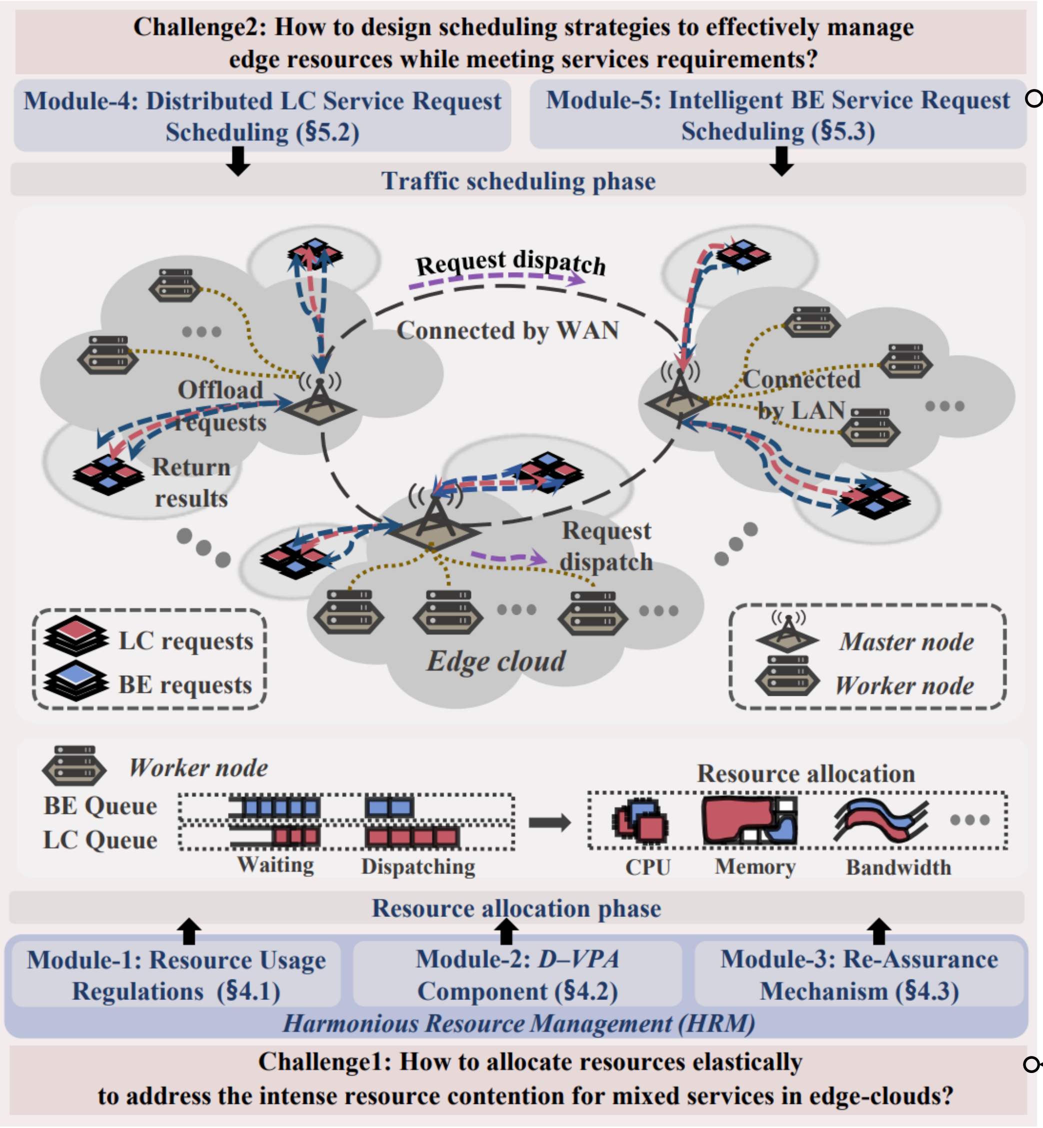
2019 S. Chen et.	2021 J.Yu et.	2021 L. Wojciechowski et.
Utilizing conventional optimization approaches, this study addresses resource isolation and dynamic allocation within cloud data centers.	By continuously monitoring the quality of mixed services, dynamic adjustments of resource allocation for both LC and BE services in cloud data centers.	Employing classical optimization methodologies, a strategy is formulated for the horizontal scaling and contraction of resources in K8s.
ACM ASPLOS	ACM ICPP	IEEE INFOCOM

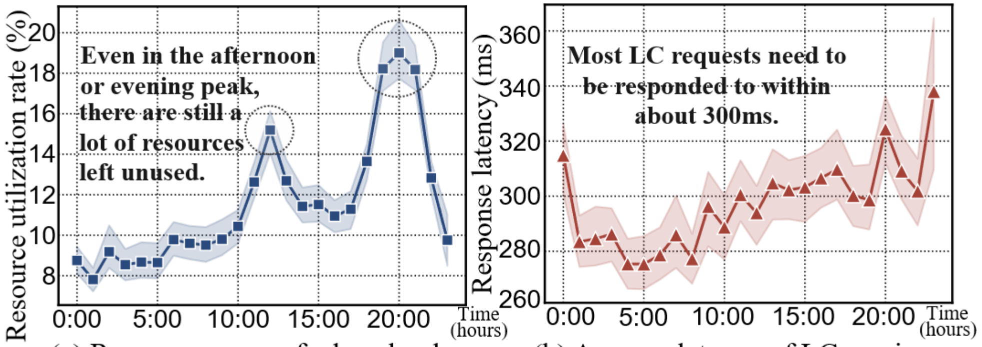
2021 P. PANG et.	2022 R.HAN et.	2022 J. Tang et.
Focused on QoS performance, a resource reconfiguration scheme is devised tailored to the co-deployment scenario within clouds.	In the context of edge cloud scenarios, a set of request scheduling strategies is devised for offline batch processing services.	Tailored for latency-sensitive services within edge cloud scenarios, a suite of request scheduling strategies is formulated.
IEEE TPDS	IEEE INFOCOM	IEEE TCC

1、S. Chen, C. Delimitrou, and J. F. Martinez, “Parties: Qos-aware resource partitioning for multiple interactive services,” in ASPLOS, 2019.
2、J. Yu, et al. "Ceres: container-based elastic resource management system for mixed workloads." Proceedings of the 50th International Conference on Parallel Processing. 2021.
3、 L. Wojciechowski, K. Opasiak, J. Latusek, M. Wereski, V. Morales, T. Kim, and M. Hong, “Netmarks: Network metrics-aware kubernetes scheduler powered by service mesh,” in IEEE INFOCOM, 2021.
4、 P. Pang, Q. Chen, D. Zeng, and M. Guo, “Adaptive preference-aware co-location for improving resource utilization of power constrained datacenters,” IEEE Trans. Parallel Distrib. Syst., vol. 32, no. 2, pp. 441456, 2021.
5、 R. Han, S. Wen, C. H. Liu, Y. Yuan, G. Wang, and L. Y. Chen, “Edgetuner: Fast scheduling algorithm tuning for dynamic edge-cloud workloads and resources,” in IEEE INFOCOM, 2022.
6、 J. Tang, M. M. Jalalzai, C. Feng, Z. Xiong, and Y. Zhang, “Latencyaware task scheduling in software-defined edge and cloud computing with erasure-coded storage systems,” IEEE Trans. on Cloud Comput. (Early Access), 2022.

Contents

2.1 First collaborative solution for edge co-location





(a) Resource usage of edge clouds. (b) Average latency of LC services.

Motivation

Aiming to improve **resource utilization**, **throughput**, and **QoS** for mixed services in edge-cloud systems.

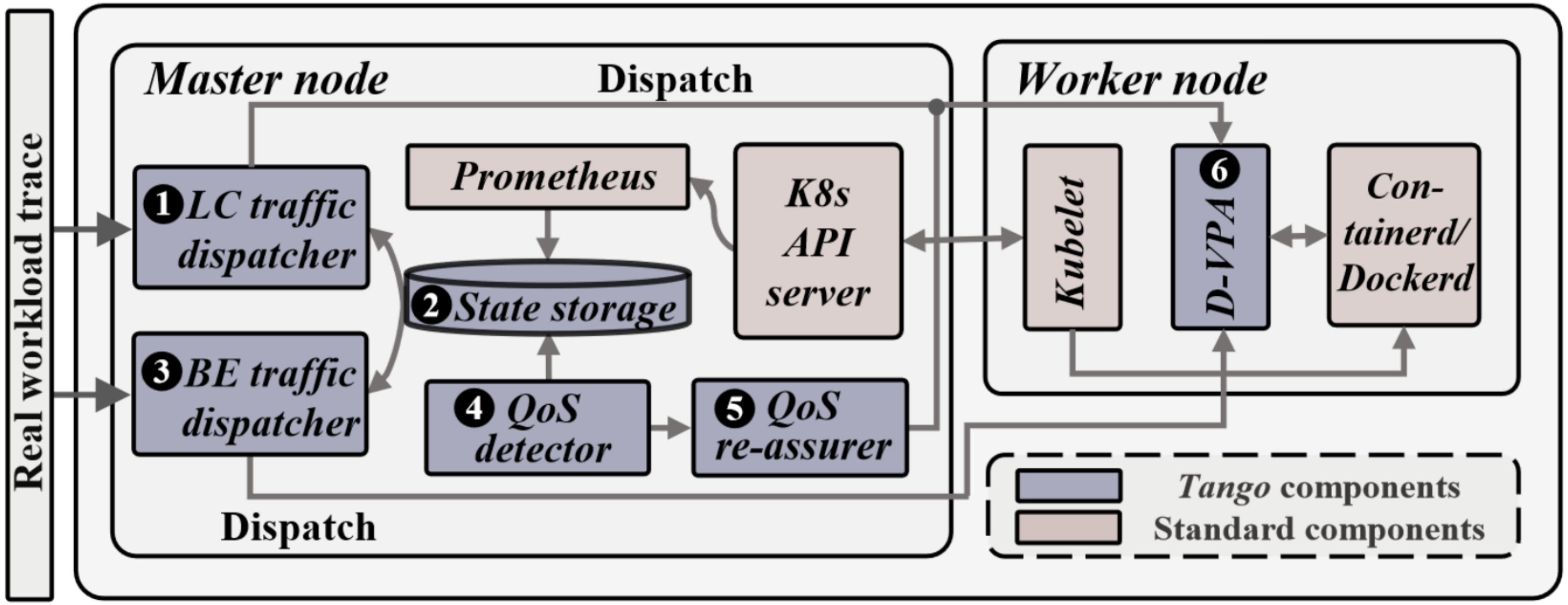
Challenge

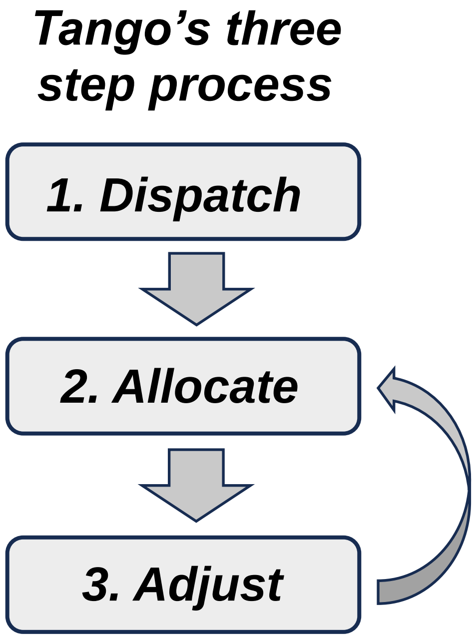
Edge resources are **distributed** and **scattered**, yet loads of user requests are **uneven** and **fluctuating** across geographic locations.

Challenge

Edge clouds are typically **heterogeneous** and **resource-constrained**, which makes competition for resources more **fierce**.

2.2 Tango design

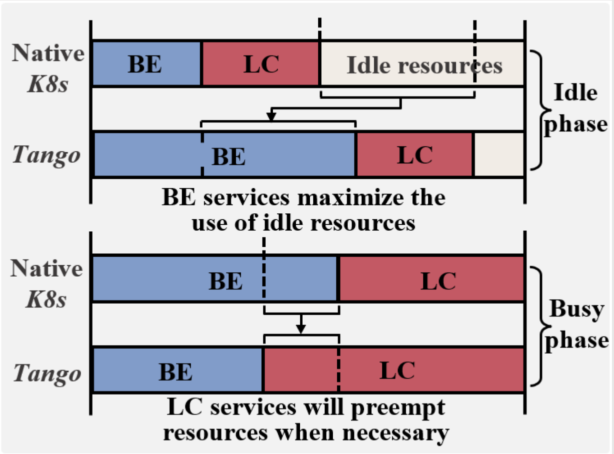




- ***Harmonious development based on K8s***
  - Dynamic vertical automatic scaling
  - Mixed traffic workload scheduling capability
- ***Harmonious design principle***
  - Fancy "dance" between BE and LC services
  - Collaborative solutions offered by Tango

2.3 Harmonious Resource Management (HRM)

①

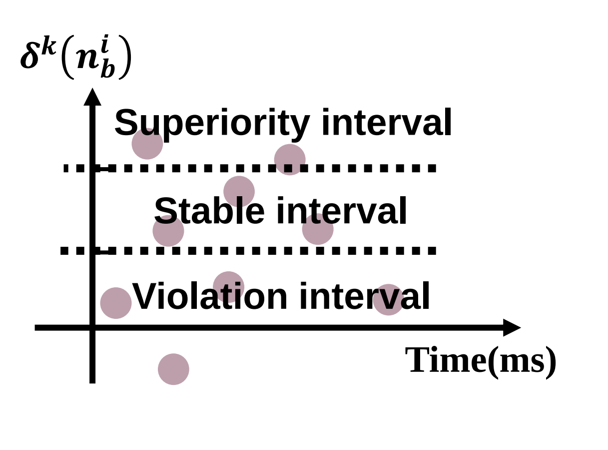


Tango Resource Balancer

K8s Qos-level Dispatcher settings

Resource usage regulations, which guides the **processing** and **offloading** by **specific priorities** to ease resource contention

②



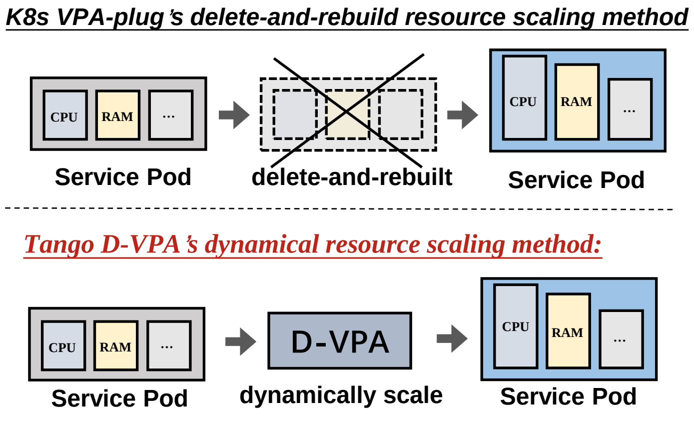
Tango Qos Detector

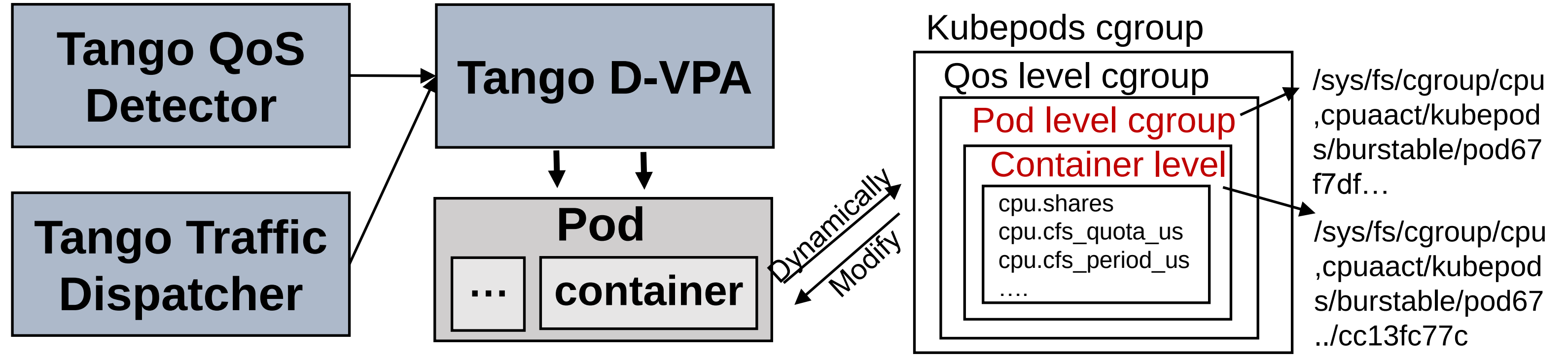
Tango Qos Re-assurer

QoS re-assurance mechanism, performing **dynamic resource tuning**

③

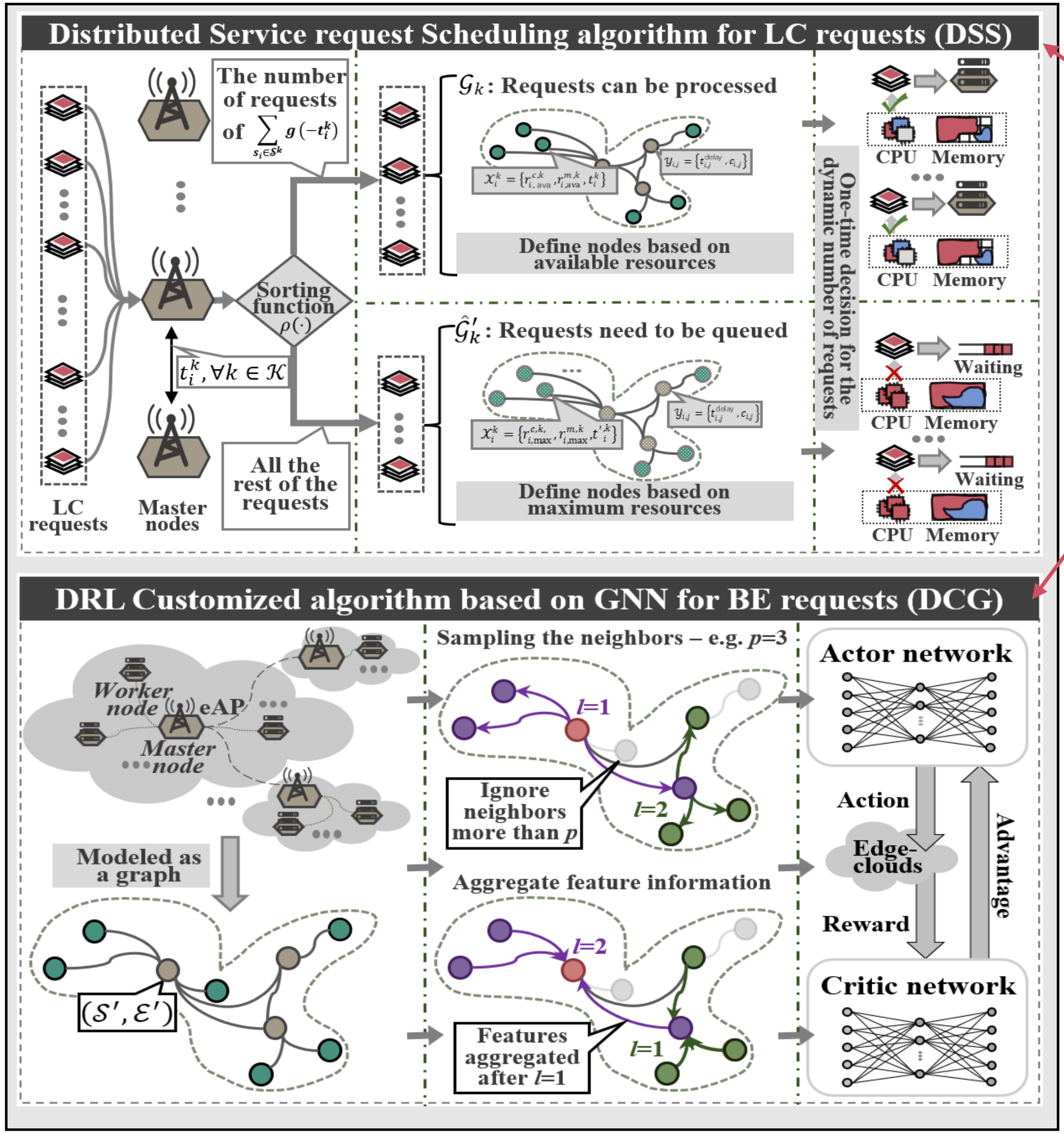
We implement the D-VPA component to solve the delete-and-rebuild resource modification problem for K8s' VPA, allowing for dynamic adaptation.

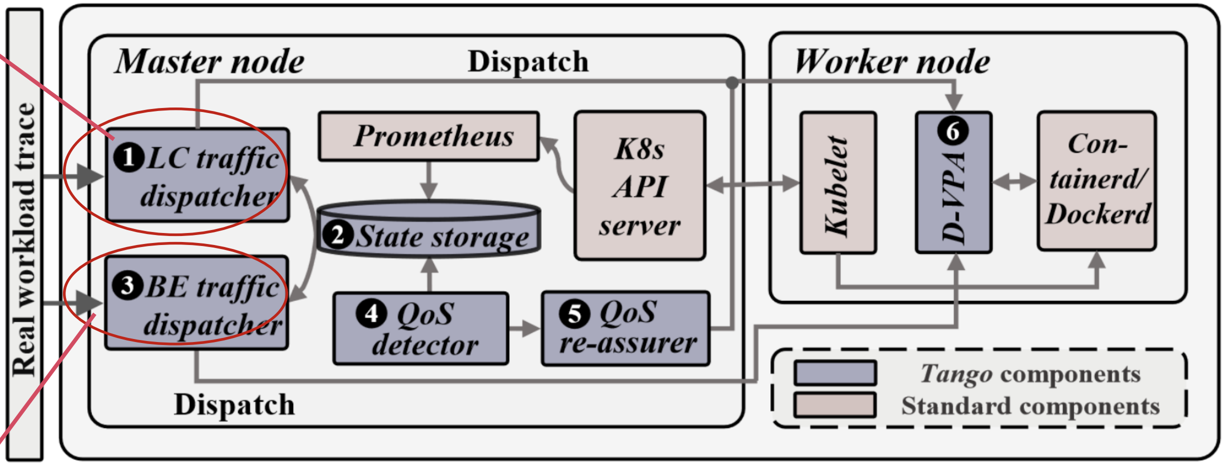




Tango's D-VPA adds a **low-level control of K8s cgroup**

2.4 Tailored traffic scheduling for service requests

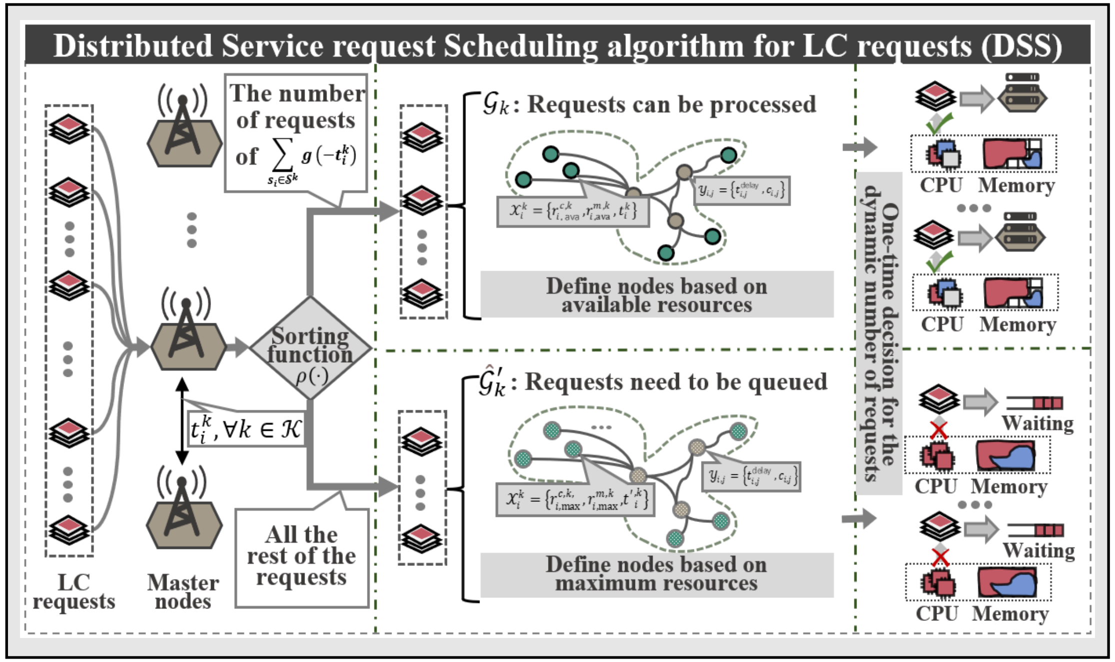




- Tango treats the <u>scheduling of LC requests</u> as a **distributed Multi-commodity Network Flow** (MCNF) problem.
- We implement a **centralized deep reinforcement learning** algorithm with adaptive adjustment capability based on **graph neural network** to make <u>BE requests scheduling.</u>

2.4.1 Tailored traffic scheduling for LC requests

- ***Distributed approach for LC requests***
  - Why not centralized scheduling? A round-trip time from the edge-cloud to the central cluster can exceed 97ms (close to 30% of the average QoS targets).
  - To make prompt scheduling decision.
  - To reduce time overhead in edge clouds.



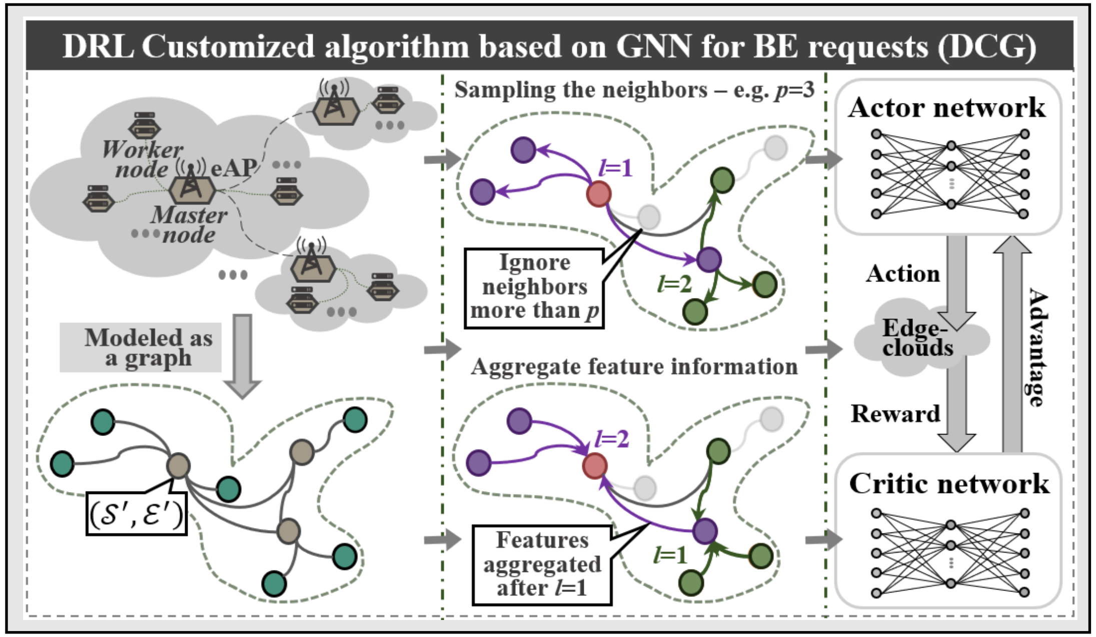
- ***Multi-Commodity network flow problem***
  - Simultaneously creates a graph G_k for each LC service request type k ∈ K.
  - DSS-LC handles requests in two case.
  - Routing paths is generated by *Ortool.*

Tango implement the <u>distributed service request scheduling algorithm</u> for LC service requests (DSS-LC).

2.4.2 Tailored traffic scheduling for BE requests

- ***Centralized approach for BE requests***
  - Allowing for an additional overhead to transfer requests to the central edge-cloud cluster.
  - Scheduling decisions based on the global state information of the edge-cloud system.
  - To reduce decision conflicts.

- ***Intelligent adaptive scheduling***
  - Graph representation learning with a Graph Neural Network (GNN) to encode the large scale network topology.
  - Deep Reinforcement Learning (DRL) to obtain the scheduling routing path.

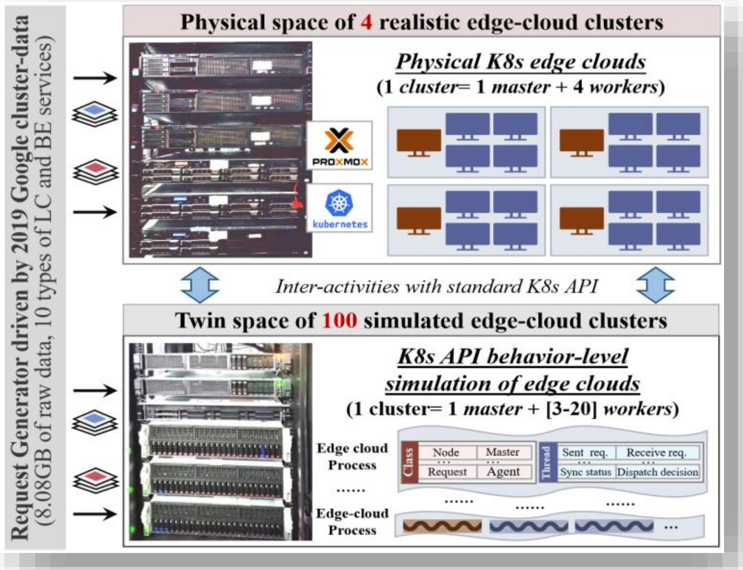


Tango propose a <u>deep reinforcement learning customized algorithm based on graph neural network for centralized</u> BE requests scheduling (DCG-BE).

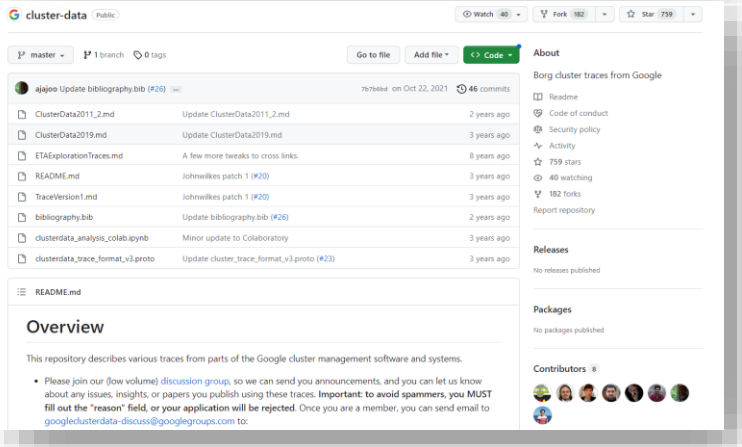
Contents

3.1 Implementation and experimental setup

- ***Dual-space edge-cloud system***
  - 4 physical edge-cloud clusters: four worker nodes (with 4 CPUs and 8GB RAM) and one master node (with 8 CPUs and 16GB RAM).
  - 100 virtual edge-cloud clusters, supported by two servers, each with 64 vCPUs and 128GB RAM.
  - Strong emphasis on asynchrony and parallelism.



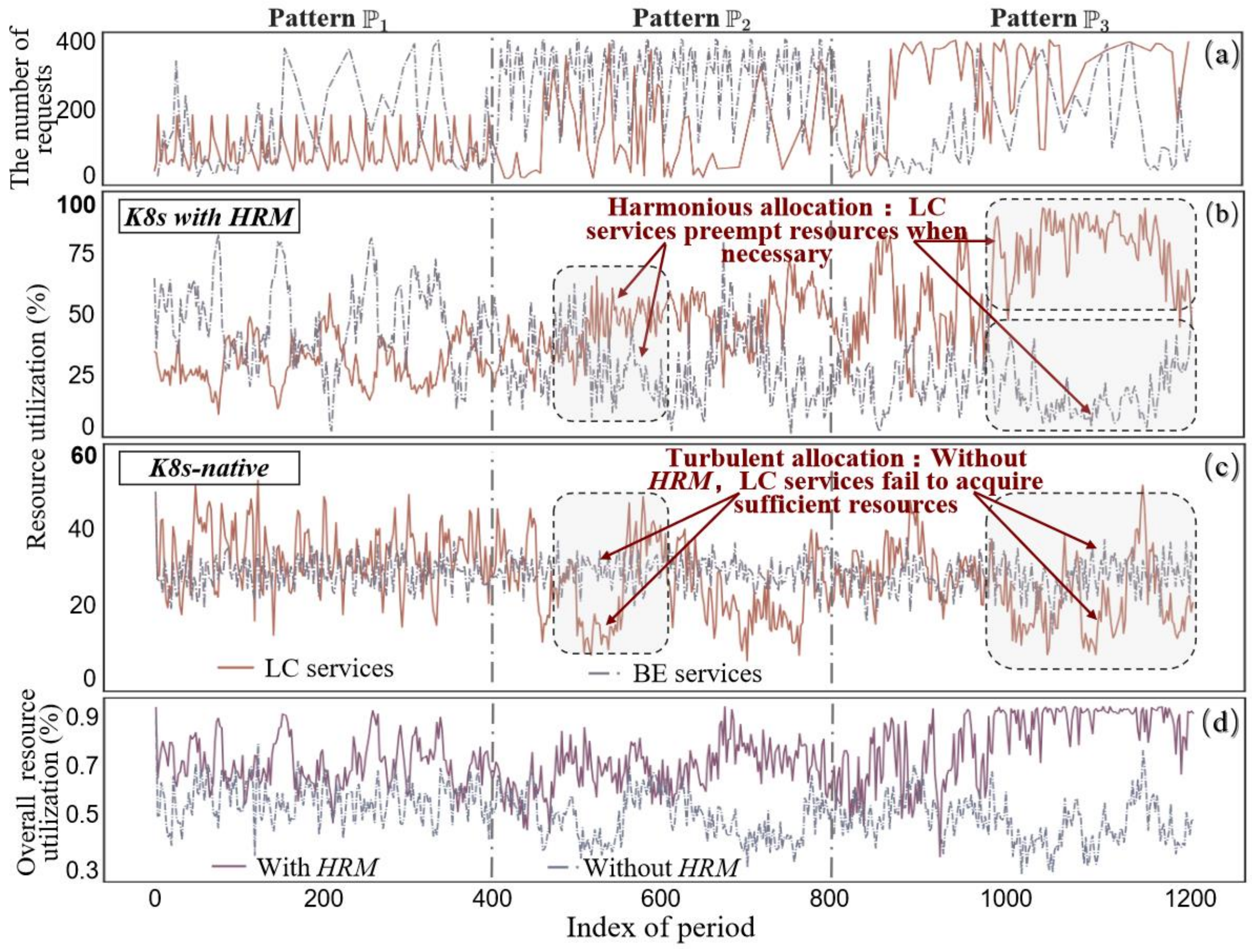
- ***Service and workload traces***
  - Workload traces of 2019 Google cluster-data.
  - One server is used as a request generator to send LC or BE service requests.
  - 10 categories of LC and BE services.

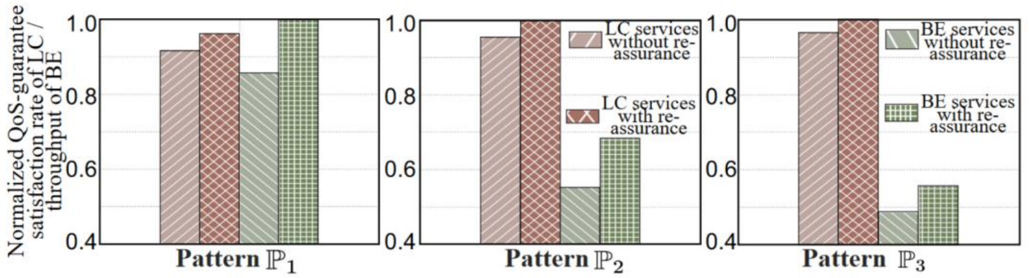


https://github.com/google/cluster-data.

https://github.com/fwyc0573/Tango

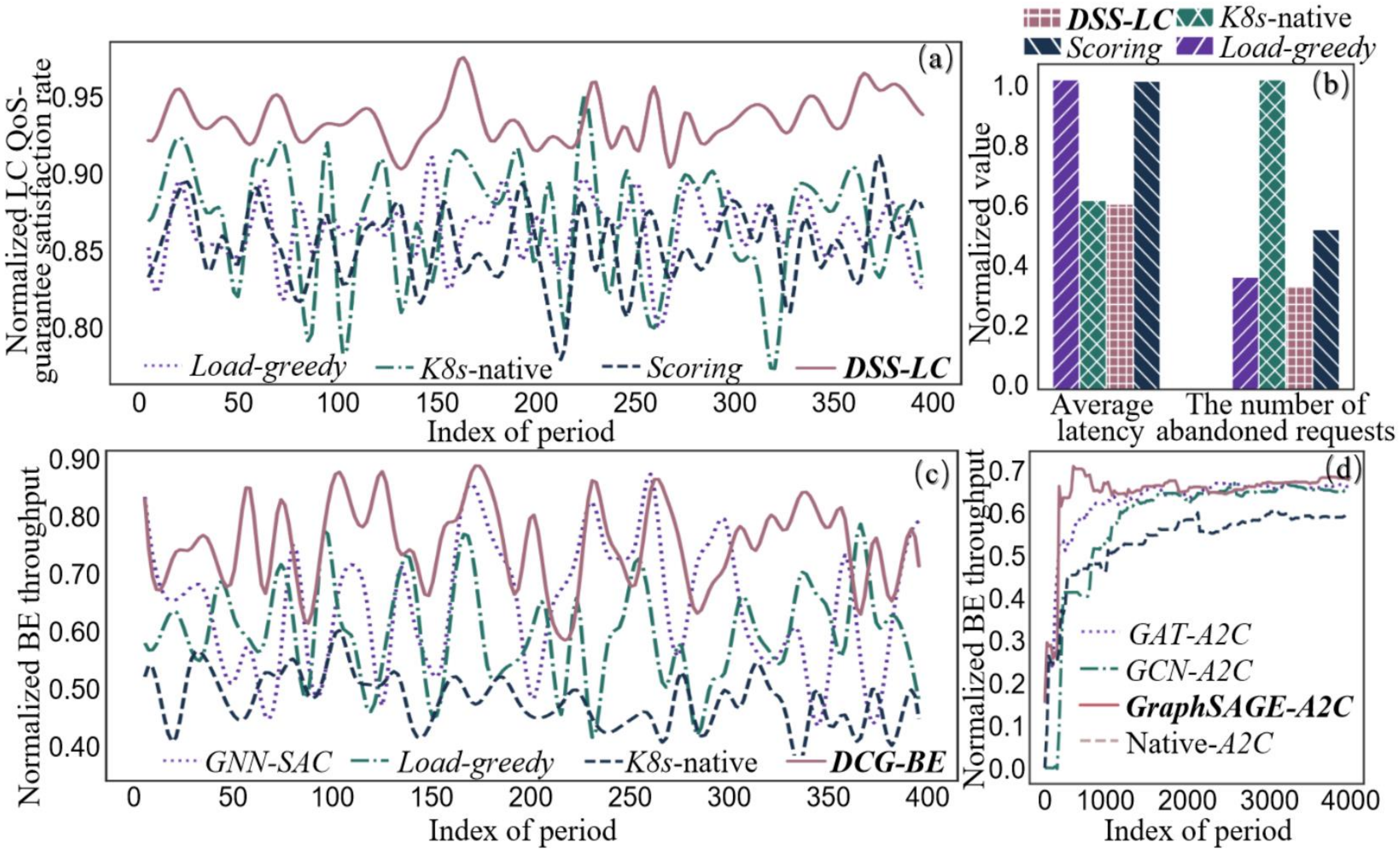
3.2 HRM Effectiveness





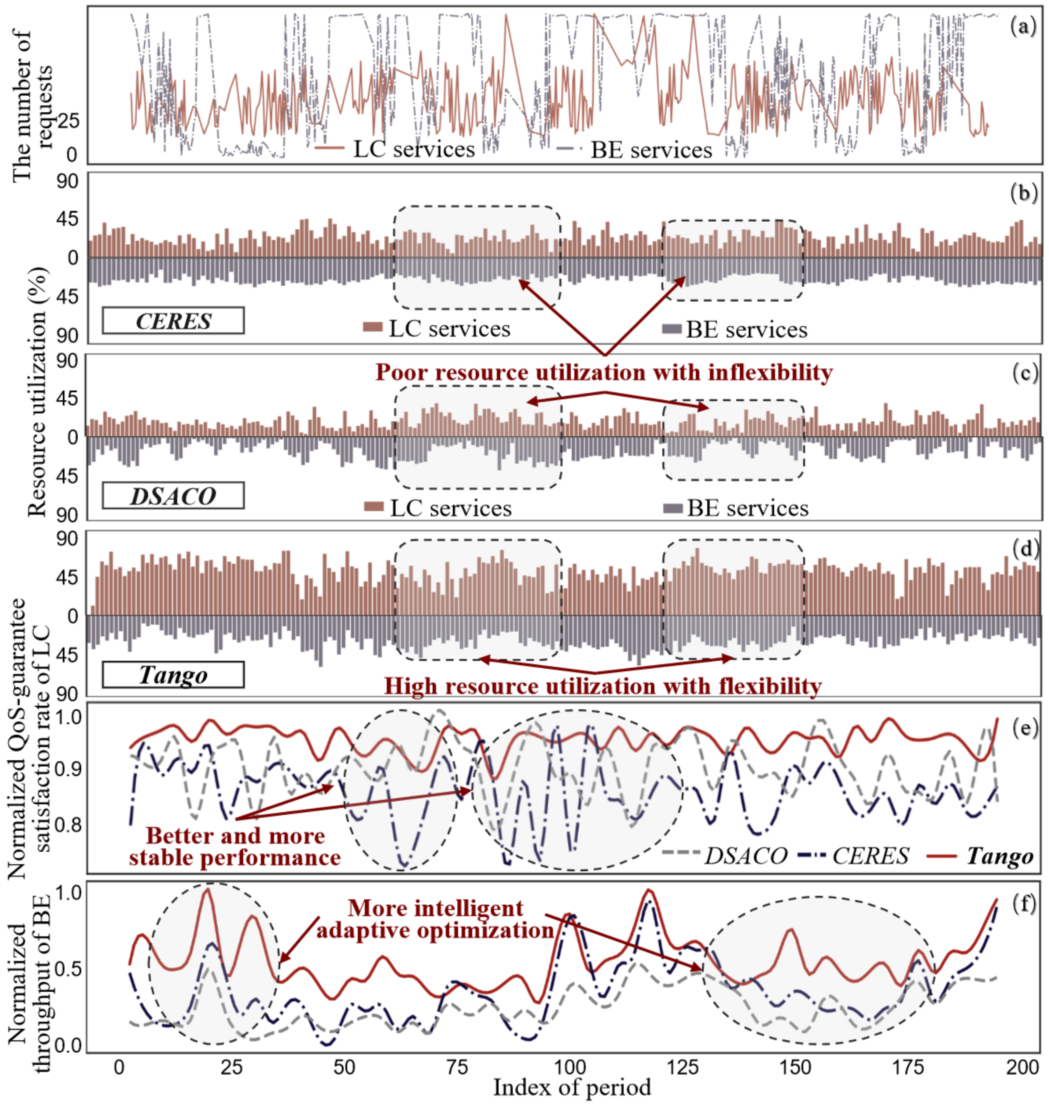
Tango's HRM mechanism effectively addresses the **resource allocation requirements** for **LC and BE services requests** across diverse workloads.

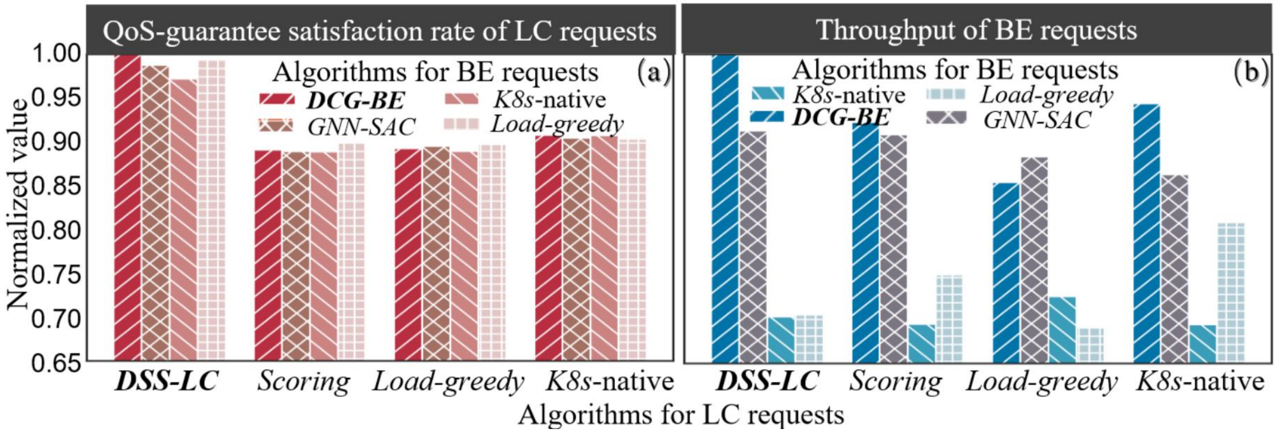
3.3 Scheduling Algorithm Performance



Compared to others, DSS-LC **significantly ensures QoS for LC service** requests across multiple metrics; DCG-BE demonstrates superior **convergence speed and throughput performance.**

3.4 Large-Scale Edge-Clouds Validation





We test different scheduling algorithm pairs and DSS-LC and DCG-BE form the **optimal algorithm combination** for Tango.

Tango improves the **system resource utilization** by 36.9%, **QoS-guarantee satisfaction rate** by 11.3%, and **throughput** by 47.6%, compared to state-of-the-art approaches.

Contents

4.1 Conclusion

We introduce Tango, a framework for harmonious management and scheduling of edge workloads. Tango is designed to improve system resource utilization and throughput while ensuring the QoS of LC services by using a series of mechanisms and components that are compatible with K8s for edge-clouds.

HRM Mechanism

Allocating edge resources elastically

DSS-LC DCG-BE

Managing edge resources effectively

Dual-Space Edge-Cloud System

Validating feasibility at scale

Thanks!

E-mail: yichengfeng@tju.edu.cn