A Holistic QoS View of Crowdsourced Edge Cloud Platform

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

2023.06

Contents

1 Background

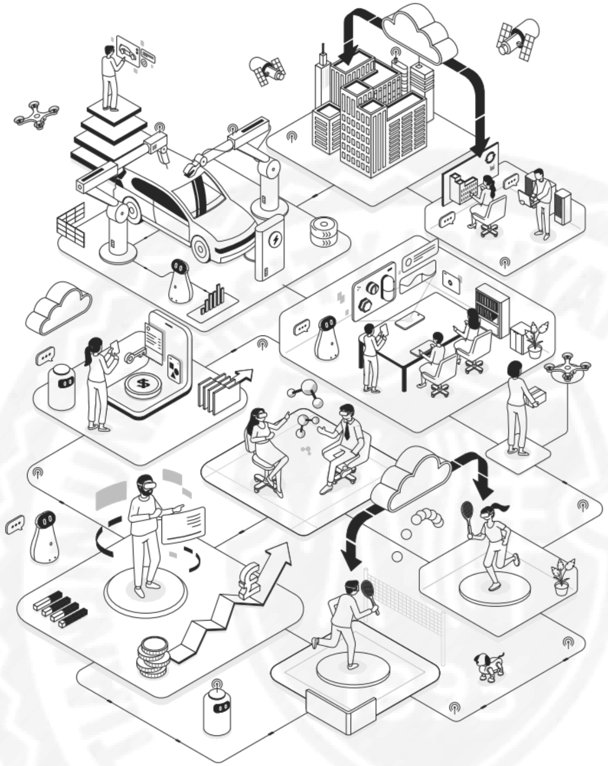
2 Overview

3 Exploring Edge Servers

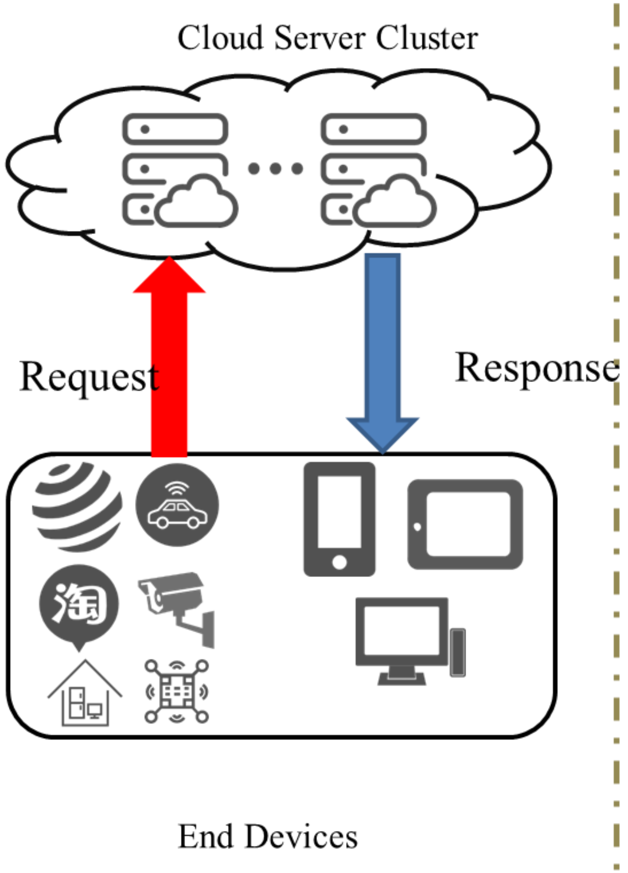
4 Exploring Containerized Services

5 Exploring User Requests

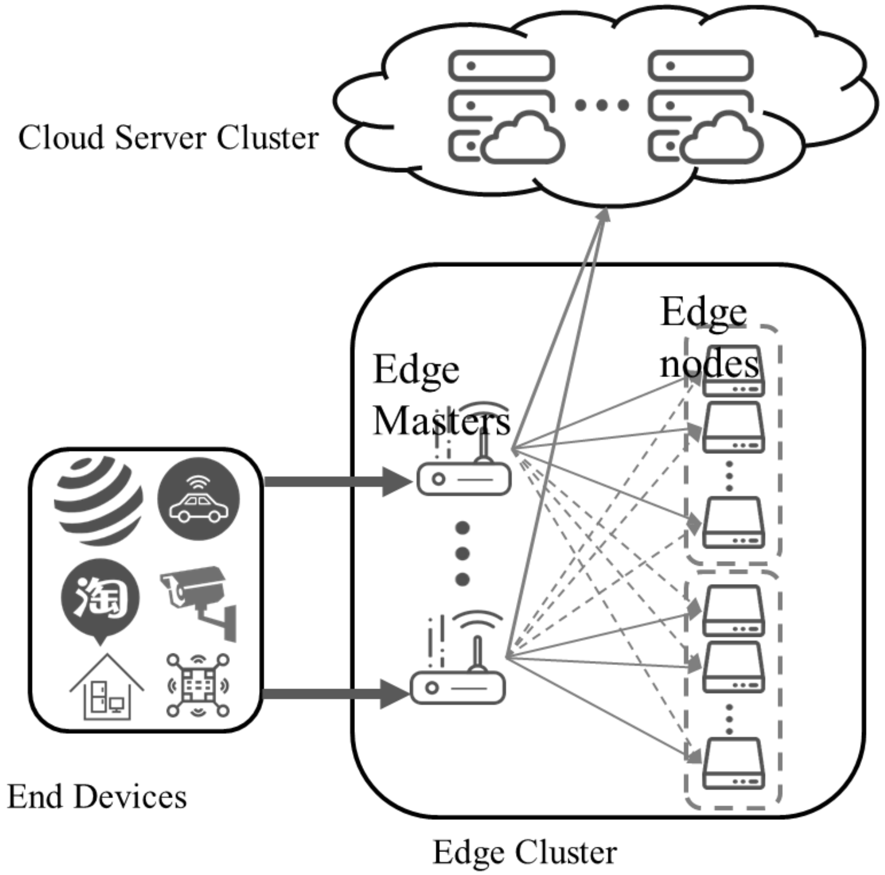
6 Conclusion



1.1 Emergence of edge computing



Centralized Cloud Computing



Distributed Edge Computing

Advantages 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.

Edge computing has unique advantages in many aspects

■ 1.2 Mainstream commercial solutions for edge computing

AWS for the Edge

https://aws.amazon.com/edge/

400+ Edge Locations and 13 Regional Edge Caches

Azure MEC

https://azure.microsoft.com/en-us/solutions/private-multi-access-edge-compute-mec/#overview

60+ Azure regions and hundreds of network POPs and data centers

Google Distributed Cloud Edge

https://cloud.google.com/distributed-cloud/edge/latest/docs/overview

187 network edge locations and 112 zones

Current solutions are still too sparse to provide proximate request responses

1.3 Crowdsourced edge cloud platform

Advantages

Decentralized-by-nature

Edge servers recruited through crowdsourcing are naturally decentralized and operating in close to users.

Cost-efficient

With crowdsourced hardware, ESP has zero expenditure in purchasing the hardware (i.e., no cold-start fee).

Carbon-friendly

By leveraging the unused hardware already manufactured, the crowdsourced edge cloud platform does not need new hardware from the manufacturer.

Challenges

Uncontrolled devices

All its infrastructure is built upon hardware out of the control of the platform.

Frequent disconnections

Servers could connect/disconnect at any time or get into failure more frequently than a datacenter-level machine.

Highly volatile resources

The hardware capacity could vary severely across time.

Crowdsourced edge cloud platform offers a new solution, but it also brings new challenges

Contents

1 Background

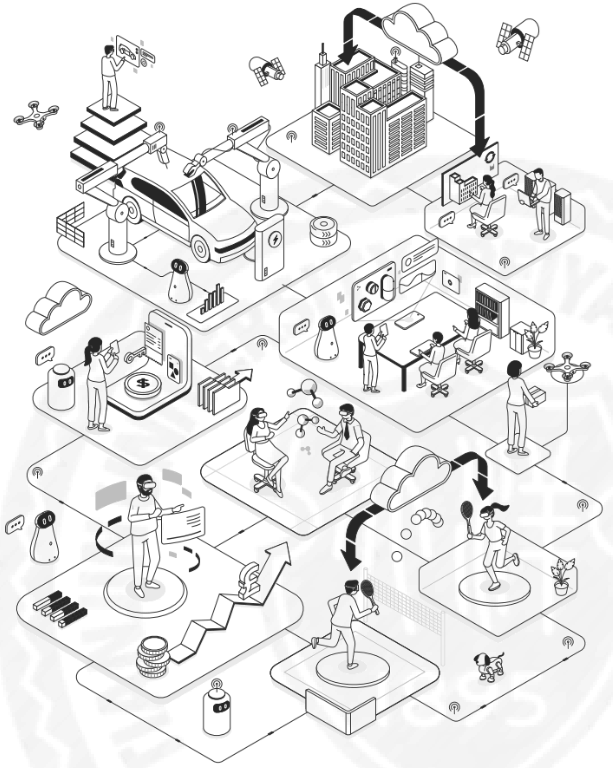
2 Overview

3 Exploring Edge Servers

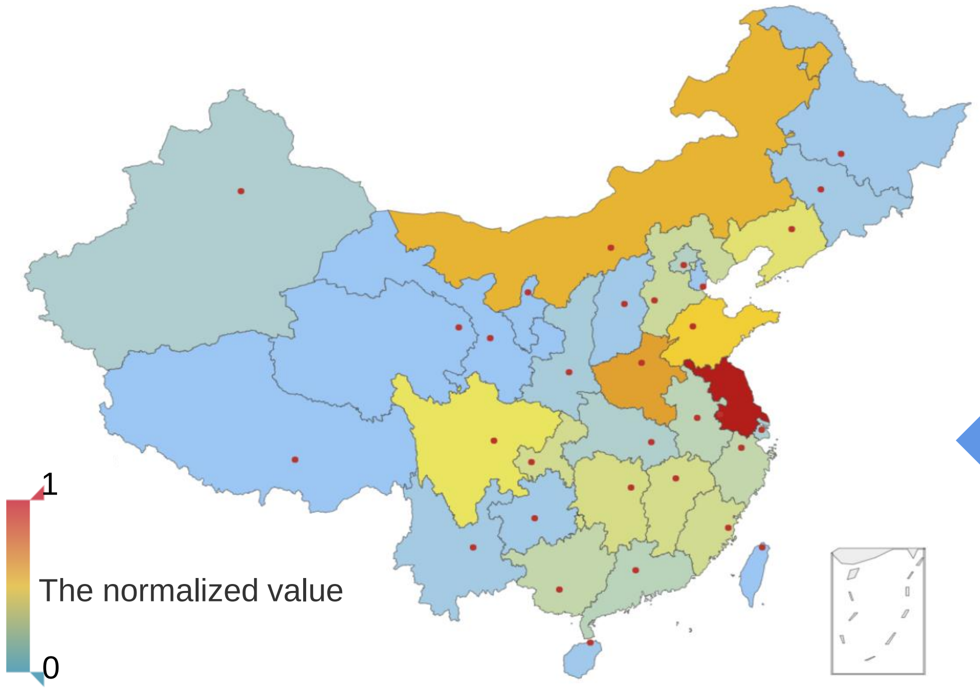
4 Exploring Containerized Services

5 Exploring User Requests

6 Conclusion



■ 2.1 Crowdsourced Edge computing Service Platform (C-ESP, i.e., PPIO Edge Cloud)



Distribution of edge server numbers

200+ cities covered

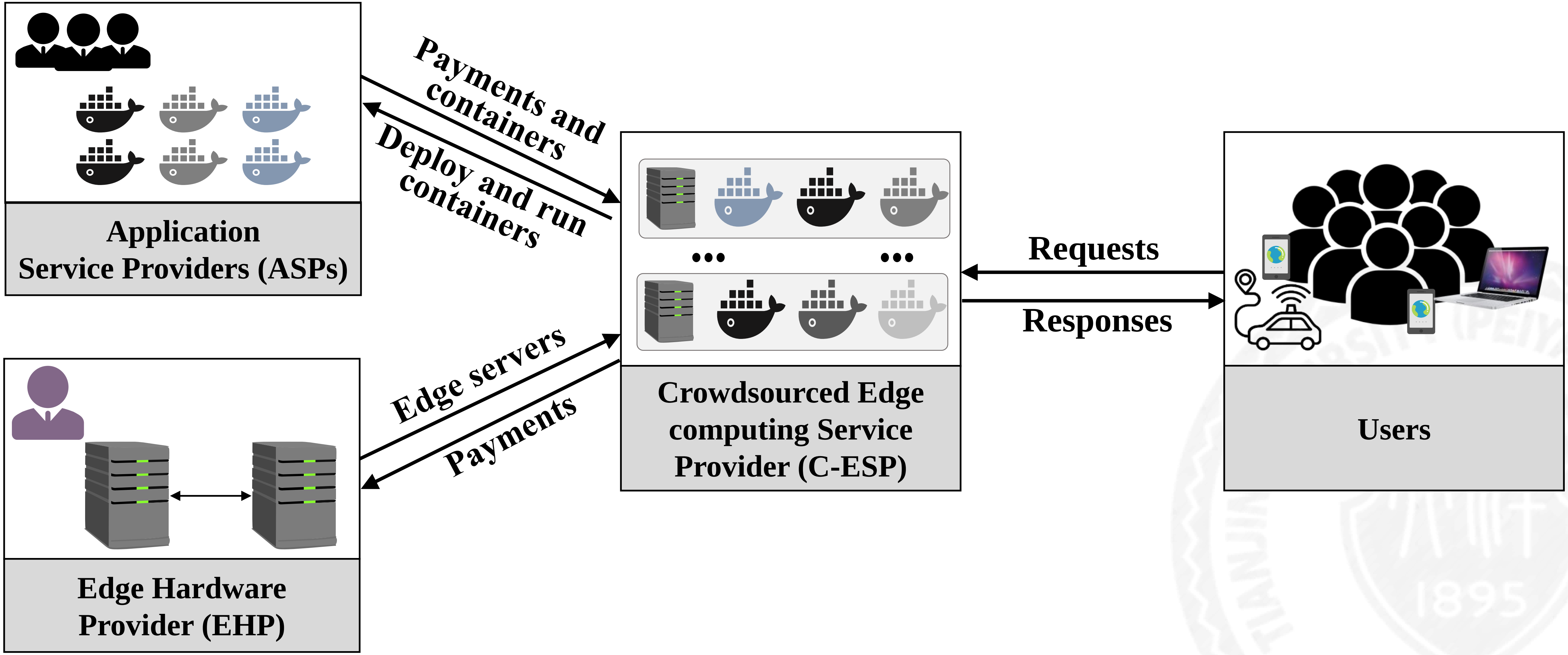
1,000+ regions covered

10,000+ edge servers

https://www.ppio.cn/

C-ESP builds an edge network that can cover almost all areas of China

2.2 Crowdsourced edge cloud platform paradigm

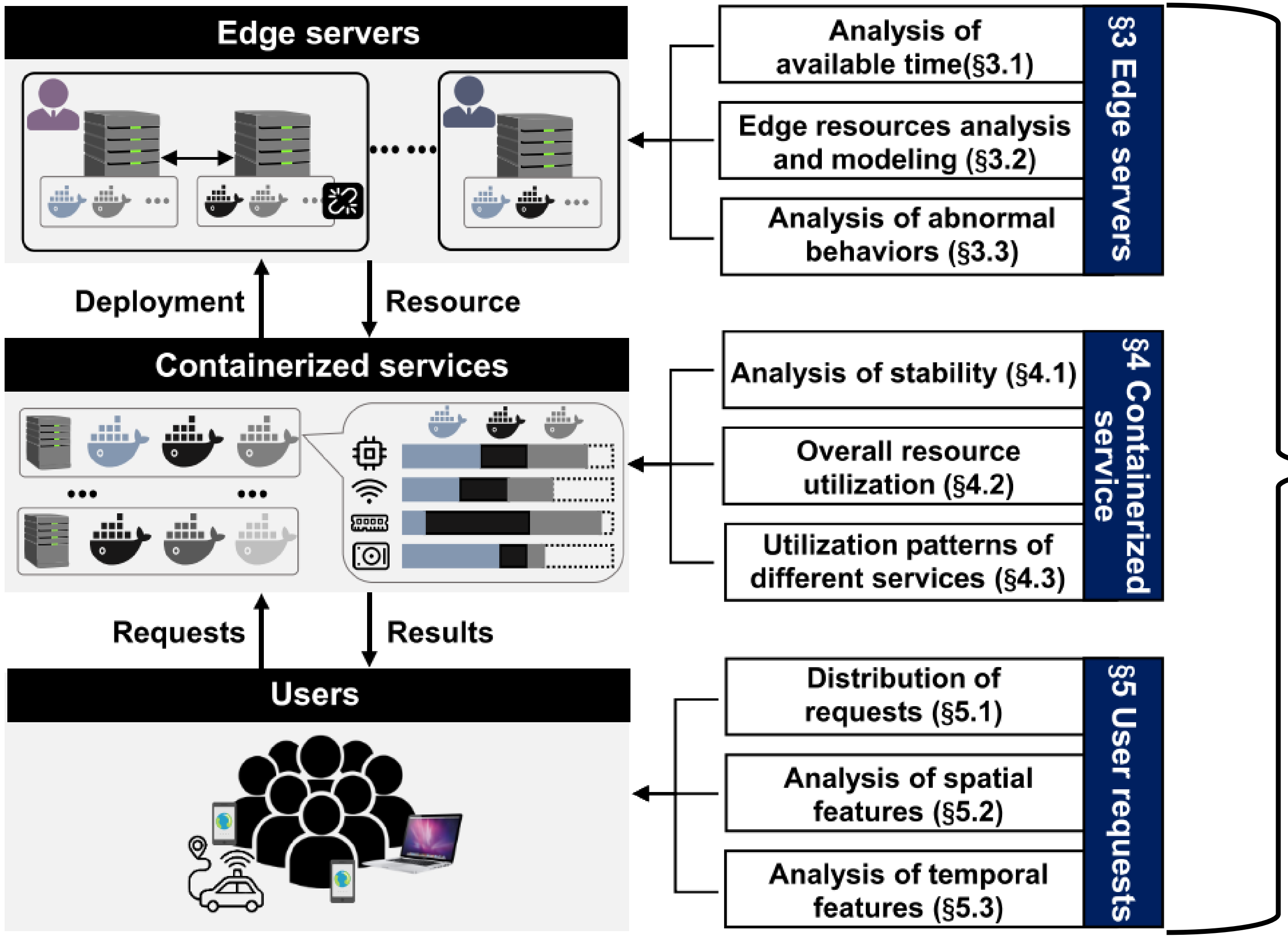


C-ESP needs to interact with three roles

2.3 Concerns of paper

Concern	Question
Concern #1	What are the quality and quantity characteristics of the hardware resources provided by C-ESP's edge servers?
Concern #2	What are the stability and utilization characteristics of the containerized services hosted on C-ESP?
Concern #3	What are the spatial and temporal distribution characteristics of user requests handled by C-ESP?

2.4 Detailed paper structure



A Holistic QoS View of Crowdsourced Edge Cloud Platform

Measurement in three dimensions: server, service and request

Contents

1 Background

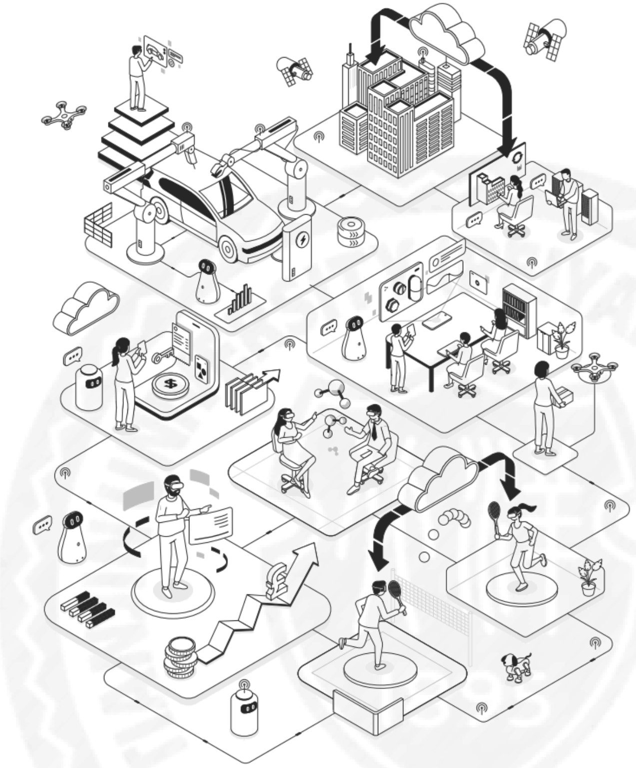
2 Overview

3 Exploring Edge Servers

4 Exploring Containerized Services

5 Exploring User Requests

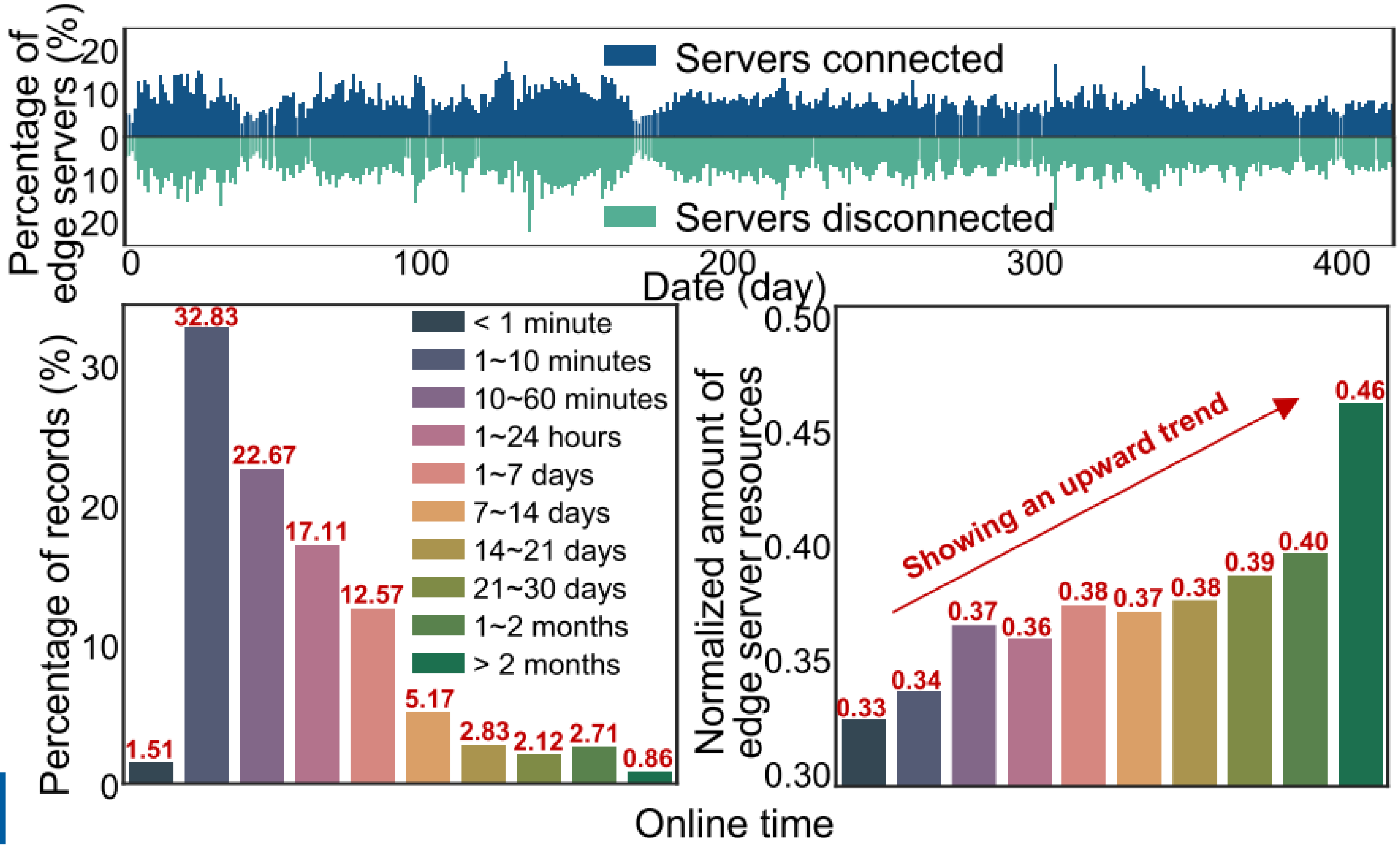
6 Conclusion



天津大学
Tianjin University

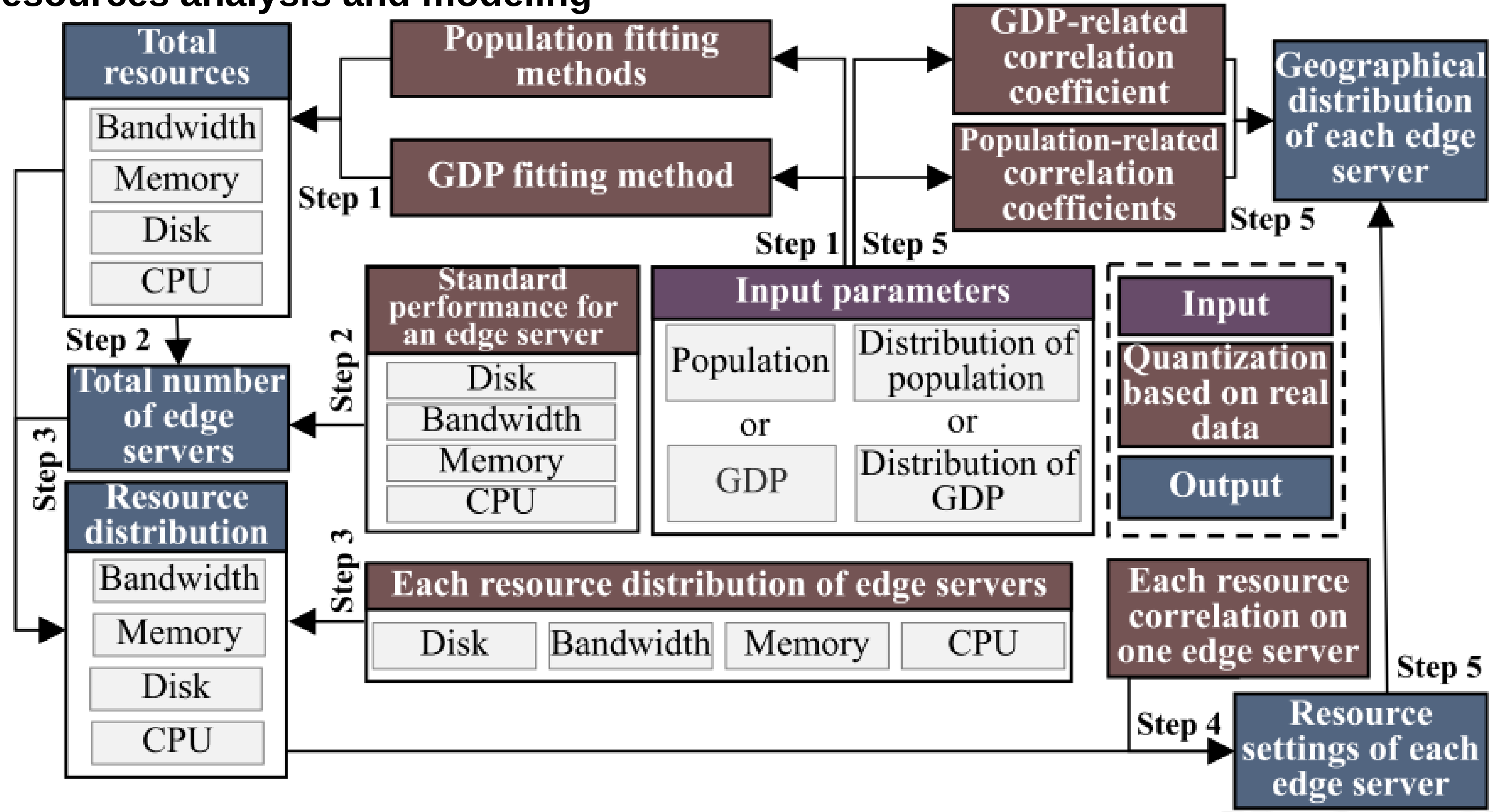
3.1 Analysis of available time

- On average, 8.53% of the total number of servers have connection records per day, while 8.30% of the total number of servers have disconnection records;
- More than half of the online time records are less than one hour, while the overall average online time is 337,782 seconds (about four days);
- Edge servers with higher resources tend to be online for longer periods.



Collected up to 418 days of server connection and disconnection records, a total of 185,698 records

3.2 Edge resources analysis and modeling



https://github.com/76481786/Flexible-Measurement-based-Modeling-Generators

Quantitative revelation of potential patterns in edge servers

3.2 Edge resources analysis and modeling

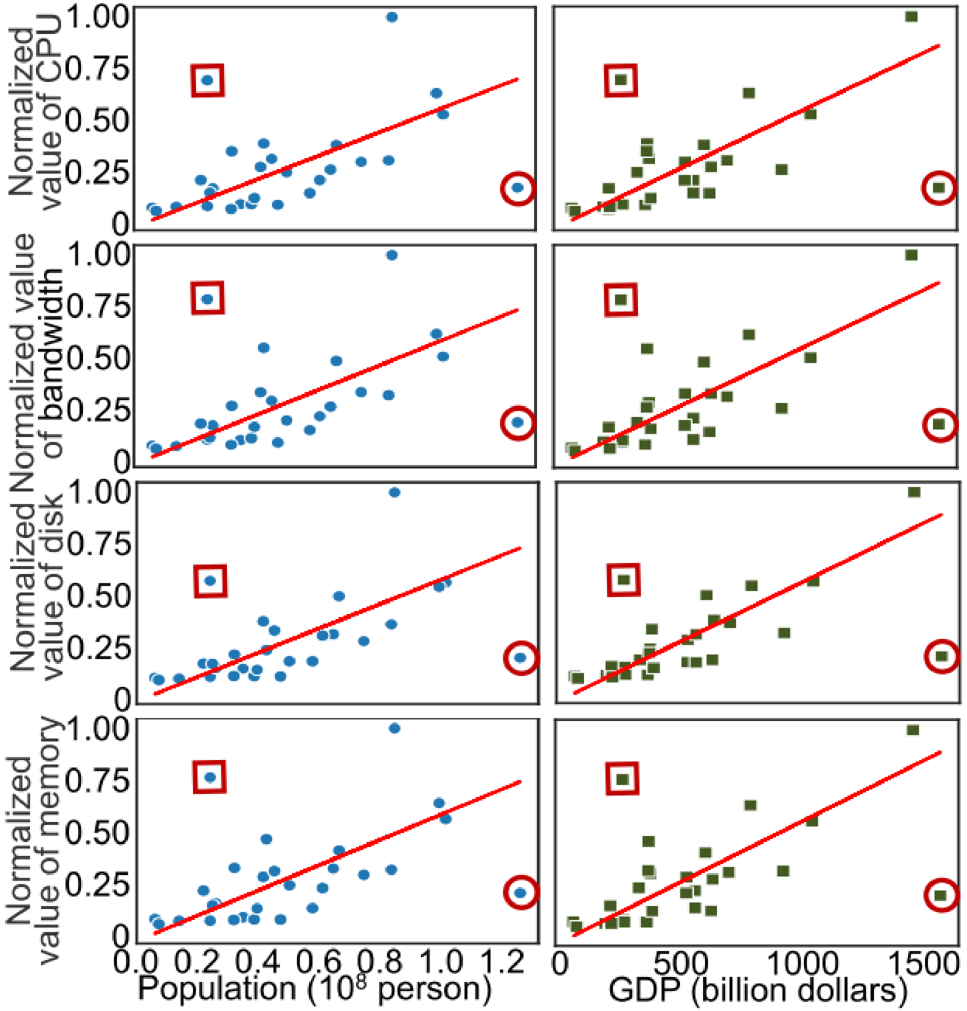


Table I
LINEAR FITTING EQUATIONS FOR POPULATION/GDP WITH EACH RESOURCE.

x	y	Linear fitting equation
Population	CPU	$y = 1.324 \times 10^{-4}x - 1.595 \times 10^{3}, r = 0.73$
	Disk	$y = 3.576 \times 10^{7}x - 6.052 \times 10^{14}, r = 0.76$
	Memory	$y = 2.104 \times 10^{5}x - 2.401 \times 10^{12}, r = 0.74$
	Bandwidth	$y = 9.145 \times 10^{3}x - 1.035 \times 10^{11}, r = 0.72$
GDP	CPU	$y = 1.206 \times 10^{1}x - 1.341 \times 10^{3}, r = 0.84$
	Disk	$y = 3.316 \times 10^{12}x - 5.657 \times 10^{14}, r = 0.89$
	Memory	$y = 1.906 \times 10^{10}x - 1.947 \times 10^{12}, r = 0.84$
	Bandwidth	$y = 8.086 \times 10^{8}x - 7.387 \times 10^{10}, r = 0.80$

- Except for marked area, the resources of other areas have a linear relationship with population/GDP;
- We derive the linear fitting equations as shown in Table I.

Population/GDP are chosen as the entry point because they are easily accessible

3.2 Edge resources analysis and modeling

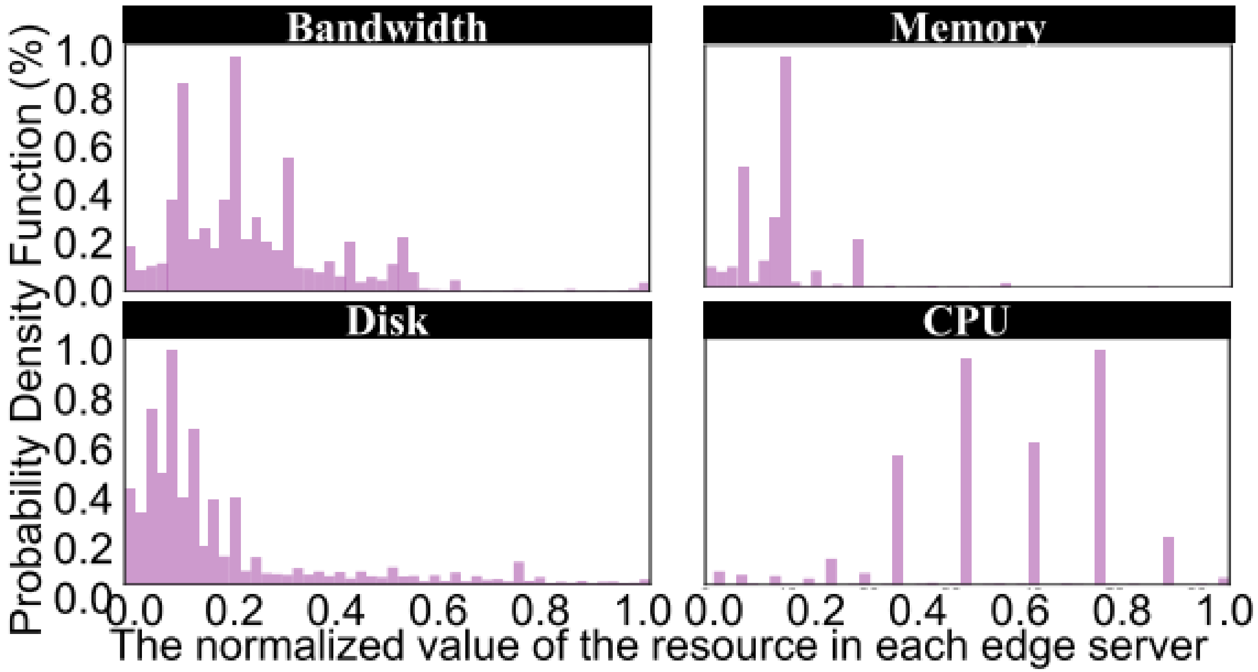


Table II
SPEARMAN CORRELATION COEFFICIENT BETWEEN DIFFERENT RESOURCES OF EDGE SERVERS.

	CPU	Disk	Memory	Bandwidth
CPU	1	0.15	0.42	0.42
Disk	0.15	1	0.29	0.36
Memory	0.42	0.29	1	0.44
Bandwidth	0.42	0.36	0.44	1

- Edge servers have different heterogeneity in CPU, disk, memory, and bandwidth;
- In addition to heterogeneity, there is also correlation between different resource.

Resource distribution is determined by quantifying the resource heterogeneity

■ 3.3 Analysis of abnormal behavior

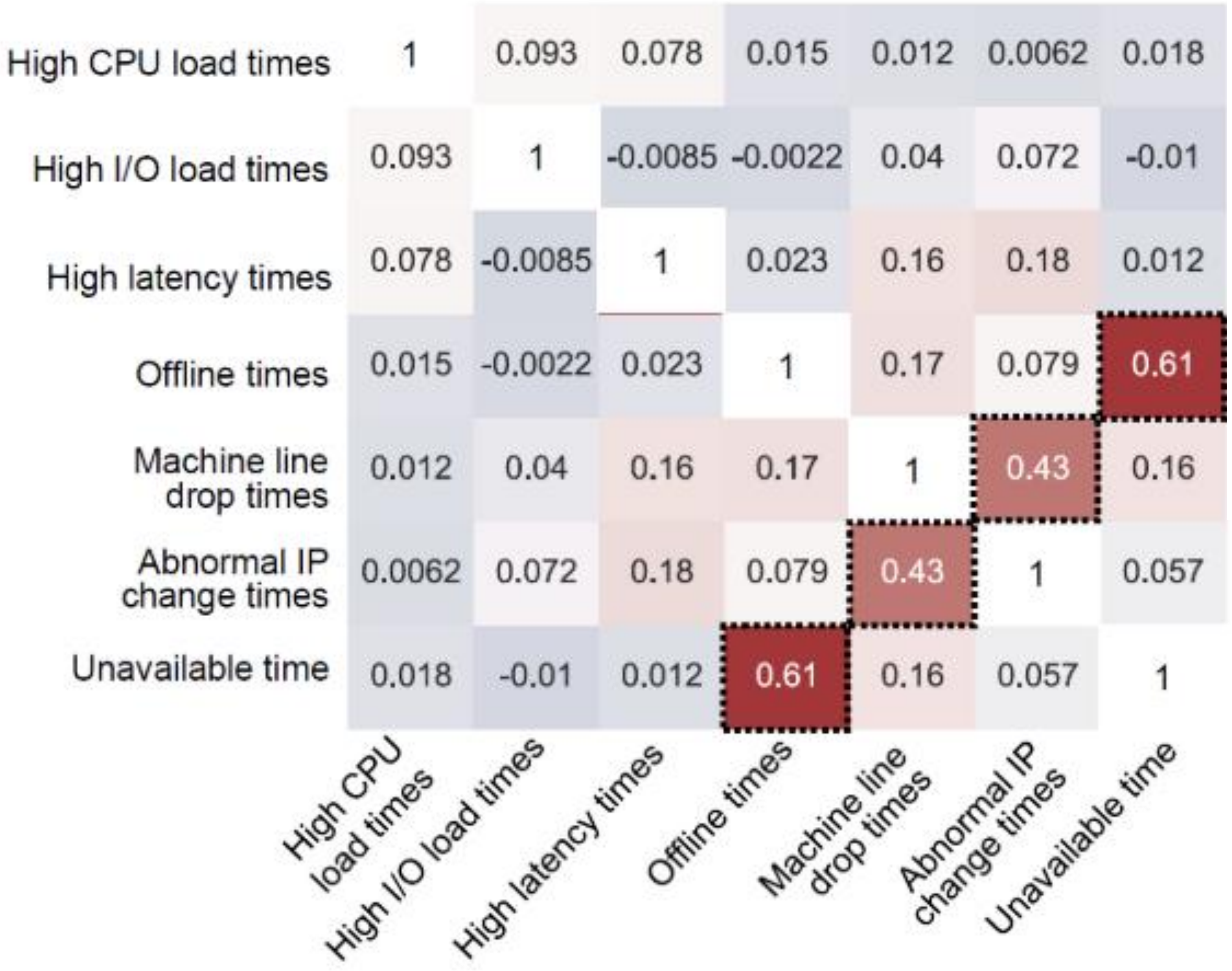


Table III
THE DISTRIBUTION OF ABNORMAL BEHAVIOR.

Abnormal behavior	Average	Percentage of zero
High CPU load times	0.532	95.57 %
High I/O load times	6.464	80.20 %
High latency times	4.084	77.18 %
Offline times	0.150	93.48 %
Machine line drop times	33.92	76.41 %
Abnormal IP change times	1.727	91.52 %
Unavailable time (seconds)	1464	93.52 %

- In addition to disk storage space, disk I/O needs more consideration when optimizing;
- It is difficult to predict the occurrence of abnormal behaviors by correlation.

we collected 428,160 operation and maintenance data for 139 days

Contents

1 Background

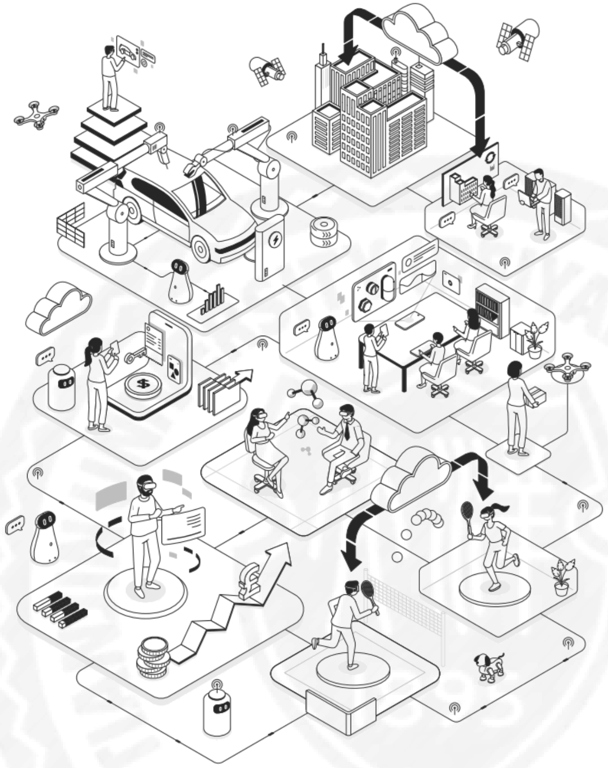
2 Overview

3 Exploring Edge Servers

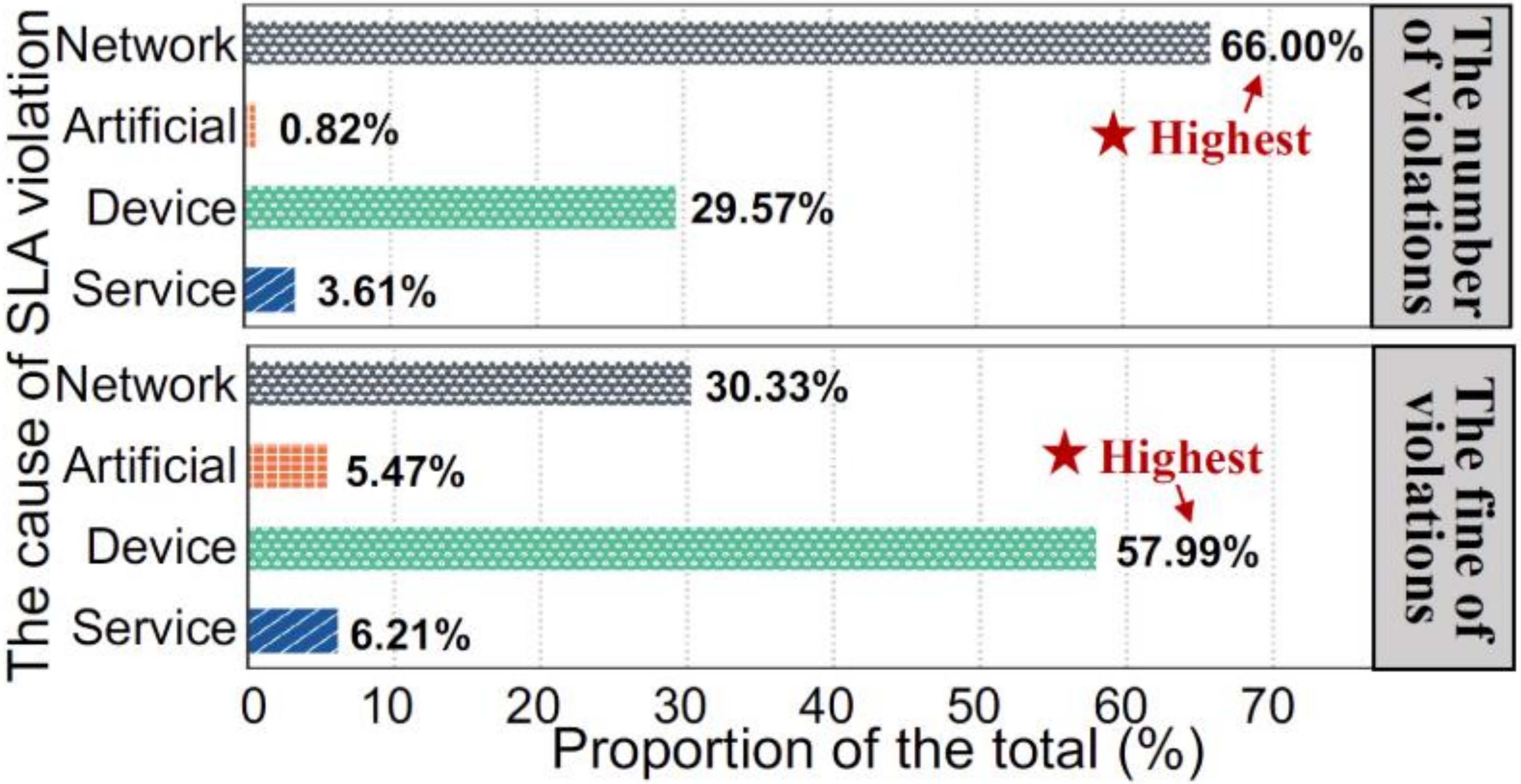
4 Exploring Containerized Services

5 Exploring User Requests

6 Conclusion



■ 4.1 Analysis of Stability

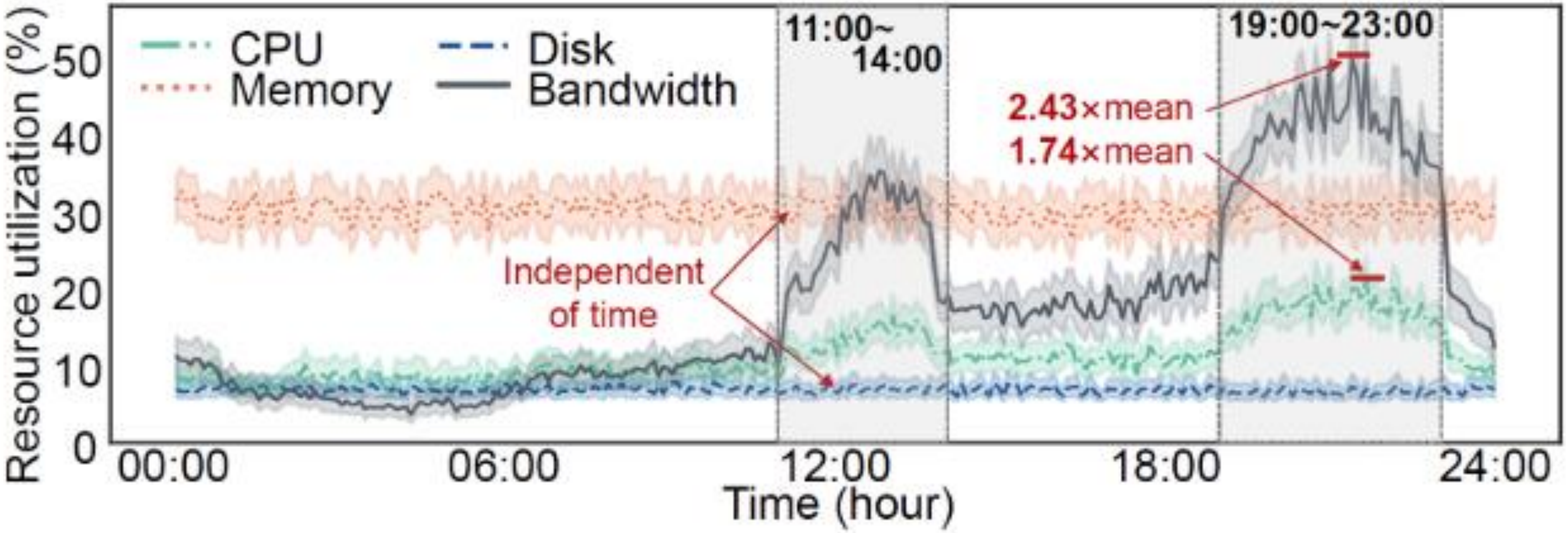


- The network is the factor that causes the most SLA violations, accounting for 66.00% of the overall;
- The device is the factor that causes the most fine of SLA violations, accounting for 57.99% of the overall.

Network and device are two of the most notable factors that lead to SLA violations

■ 4.2 Overall resource utilization

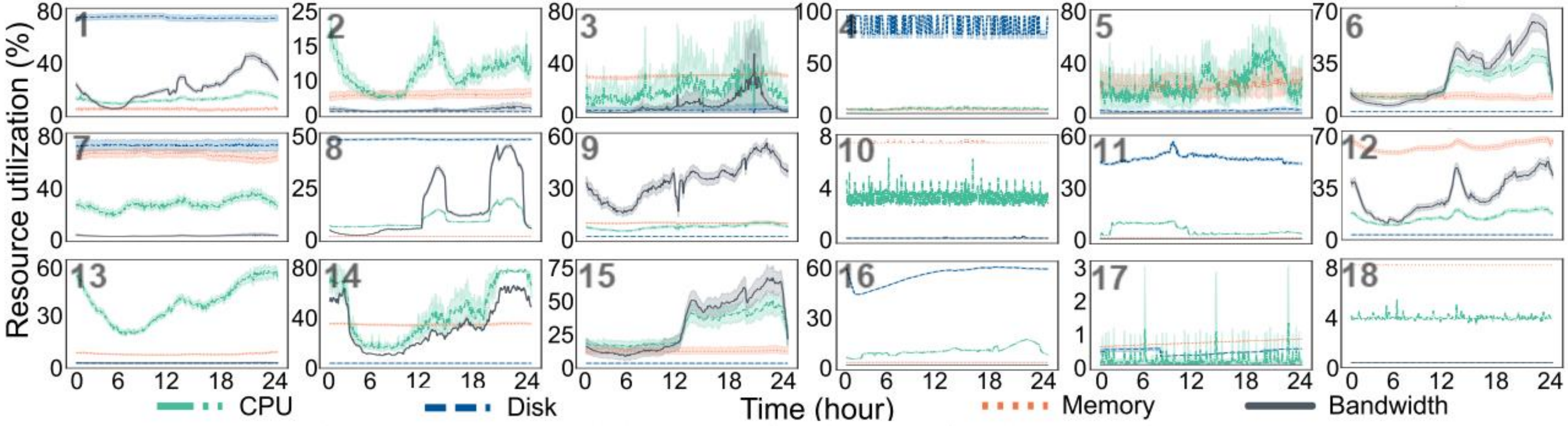
- Resource requirements for memory and disk are relatively independent of time, while resource requirements for CPU and bandwidth are time-dependent;
- There are two peaks at noon/evening, with a noon peak of 3 hours (11:00~14:00) and an evening peak of 4 hours (19:00~23:00) .
- The evening peak is higher than the afternoon peak, with its CPU and bandwidth peaks 1.74 and 2.43 times the daily average, respectively.



Data is collected every 5 minutes, i.e. 288 logs are generated in a day

天津大学 Tianjin University

4.3 Utilization patterns of different services



- Idle resources on edge servers are direct resources that can be reallocated, and idle resources in edge containers are indirect resources that cannot be reallocated;
- How to arrange containers on edge servers to make full use of resources;
- How to deploy containerized services with complementary requirements on an edge server to improve resource utilization?

The heterogeneity of different utilization patterns poses many challenges

Contents

1 Background

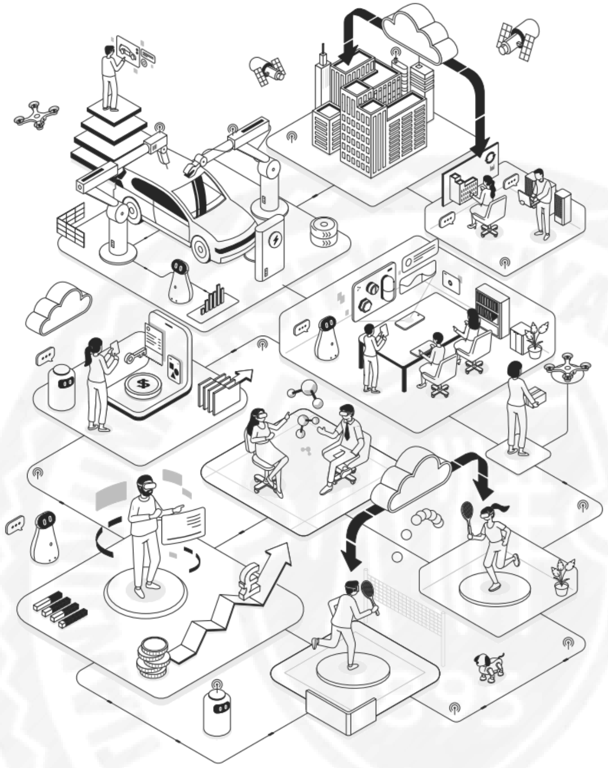
2 Overview

3 Exploring Edge Servers

4 Exploring Containerized Services

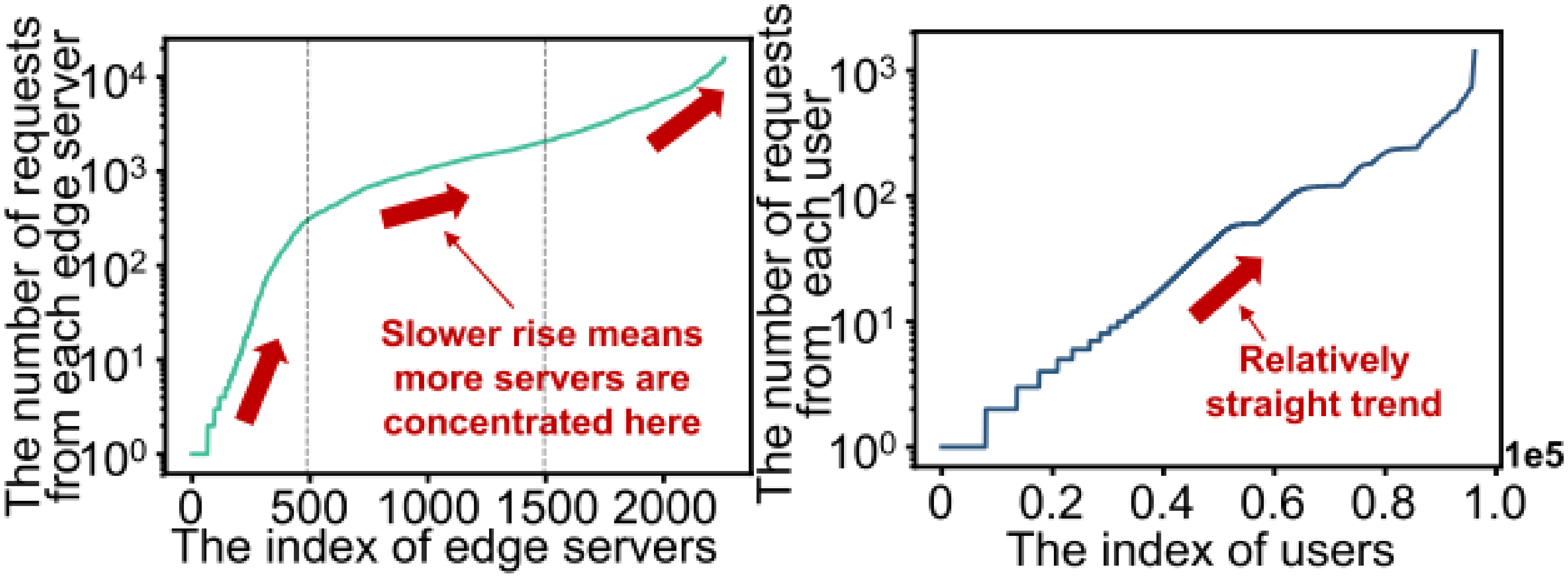
5 Exploring User Requests

6 Conclusion



■ **5.1 Distribution of requests**

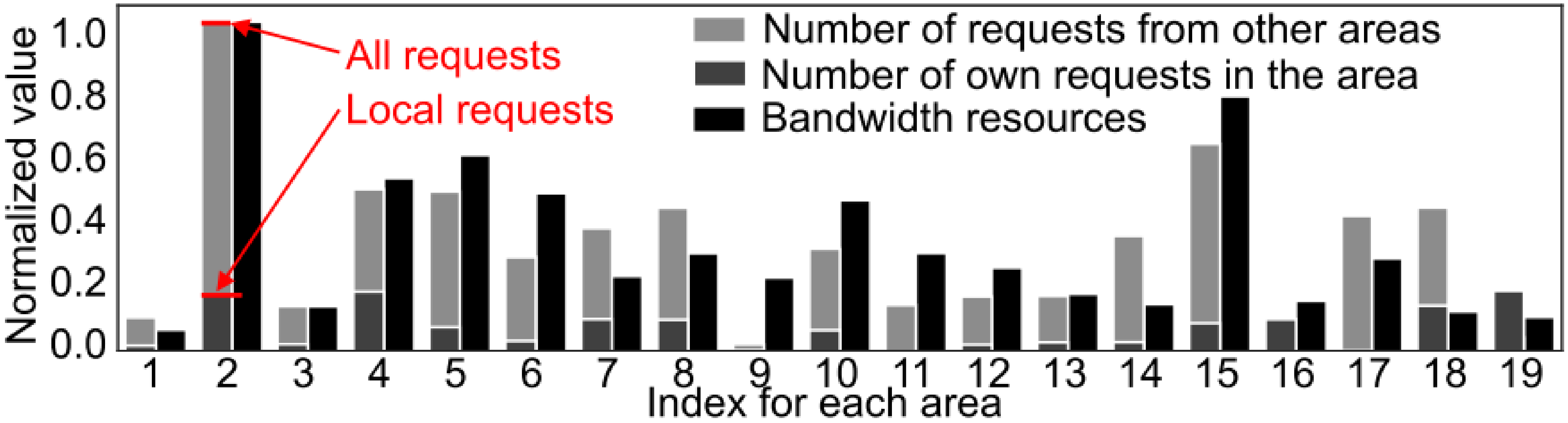
- The number of requests per user in the collected data follows the exponential distribution;
- The distribution of requests on the server is more centralized.



Differences in the distribution of requests between servers and users

5.2 Analysis of spatial feature

- The number of resources in each area is inconsistent with the number of requests generated;
- It is necessary to develop request scheduling algorithms to schedule requests from different areas.
- The resources of edge platform are distributed causing the problem of matching resources and requests.



Scheduling is necessary due to the mismatch between resources and request distribution

5.3 Analysis of temporal features

- The Poisson distribution is:

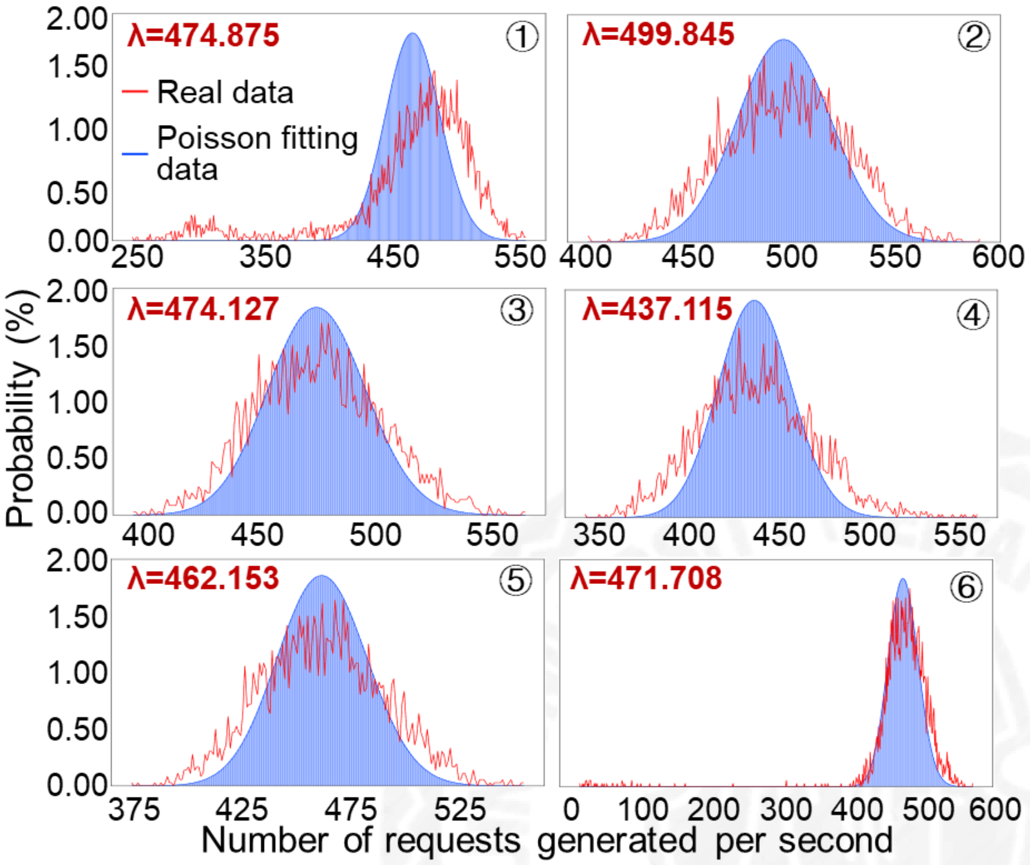
$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,\cdots$$

- λ is proportional to the number of users N:

$$\lambda = C \cdot N$$

- **C=0.0134**

Index	λ	N	C	Index	λ	N	C
1	474.875	35688	0.0133	4	437.115	33699	0.0130
2	499.845	34449	0.0145	5	462.153	34770	0.0123
3	474.127	35074	0.0135	6	471.708	34314	0.0137



Fitting analysis based on Poisson distribution

User requests are quantified based on Poisson distribution

Contents

1 Background

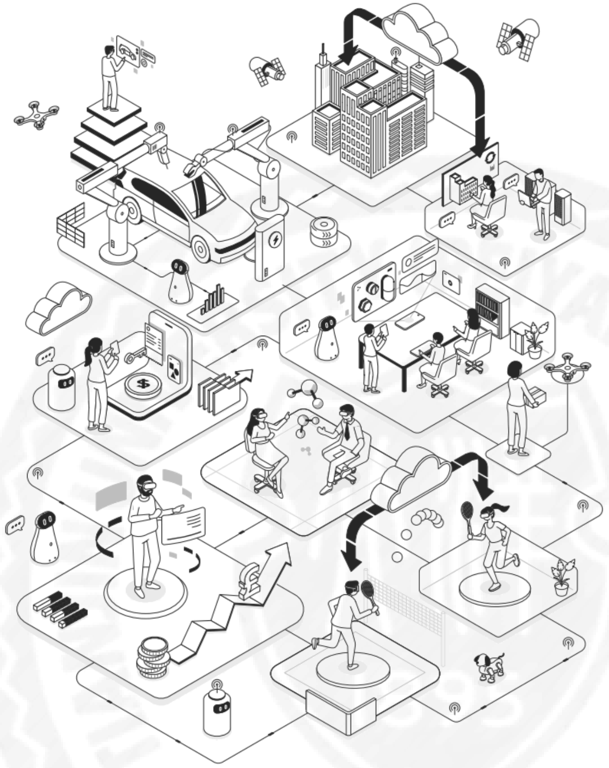
2 Overview

3 Exploring Edge Servers

4 Exploring Containerized Services

5 Exploring User Requests

6 Conclusion



We perform QoS measurements based on three dimensions: edge servers, containerized services, and user requests. Based on the above research, we aim to provide realistic experience for related research and promote solutions to the problems mentioned in this paper.

Edge Servers

Quality and Quantity

Containerized Services

Stability and Utilization

User Requests

Spatial and Temporal


Thanks Everyone!