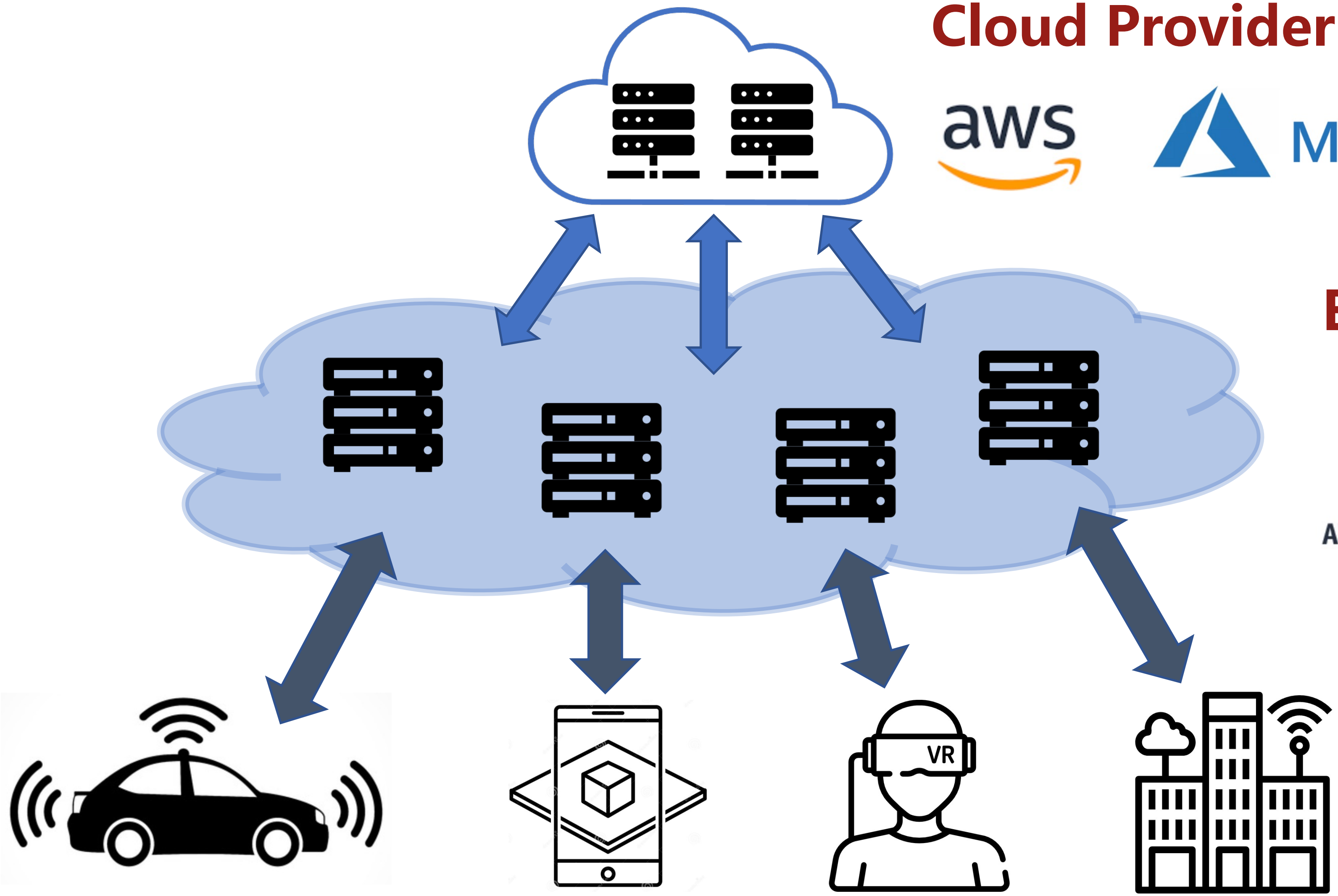
From Cloud to Edge: A First Look at Public Edge Platforms

Mengwei Xu, Zhe Fu, Xiao Ma, Li Zhang, Yanan Li, Feng Qian,
Shangguang Wang, Ke Li, Jingyu Yang, Xuanzhe Liu

UNIVERSITY OF MINNESOTA

Current Edge Platforms



Cloud Providers

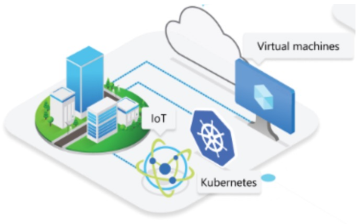
aws

Microsoft Azure

Google Cloud

Edge Platforms

AWS Local Zones



Azure Edge Zone

Edge Applications

- Autonomous driving
- AR/VR
- Smart city

Overview of NEP

- What is NEP?
  - Alibaba ENS[1]
  - A leading edge platform in China, providing IaaS, PaaS, FaaS, etc.
- What consists of NEP physical servers?
  - Built atop Alibaba CDN.
  - Cooperative third-party IDCs and network operators.
  - Mainly based on micro datacenters rather than sinking into cellular core networks.

[1] Extending the Boundaries of the Cloud with Edge Computing
https://www.alibabacloud.com/blog/extending-the-boundaries-of-the-cloud-with-edge-computing_594214

Overview of NEP

Platform	Regions / Coverage		Density ($10^6 mi^2$)	Platform	Regions / Coverage		Density ($10^6 mi^2$)
AWS EC2	24	Global	0.13	MS Azure	33	Global	0.17
	6	U.S.	1.58		8	U.S.	2.11
Google Cloud	24	Global	0.13	Alibaba Cloud	23	Global	0.12
	8	U.S.	2.10		12	China	3.23
Azure Edge Zones	5	U.S.	1.32	Huawei Cloud	5	China	1.35
AWS Wavelength + Local Zones	14	U.S.	3.70	NEP (our study)	>500	China	>135

- The deployment scale of NEP is larger than current edge/cloud platforms!

Benefits of NEP-like Edge Platforms

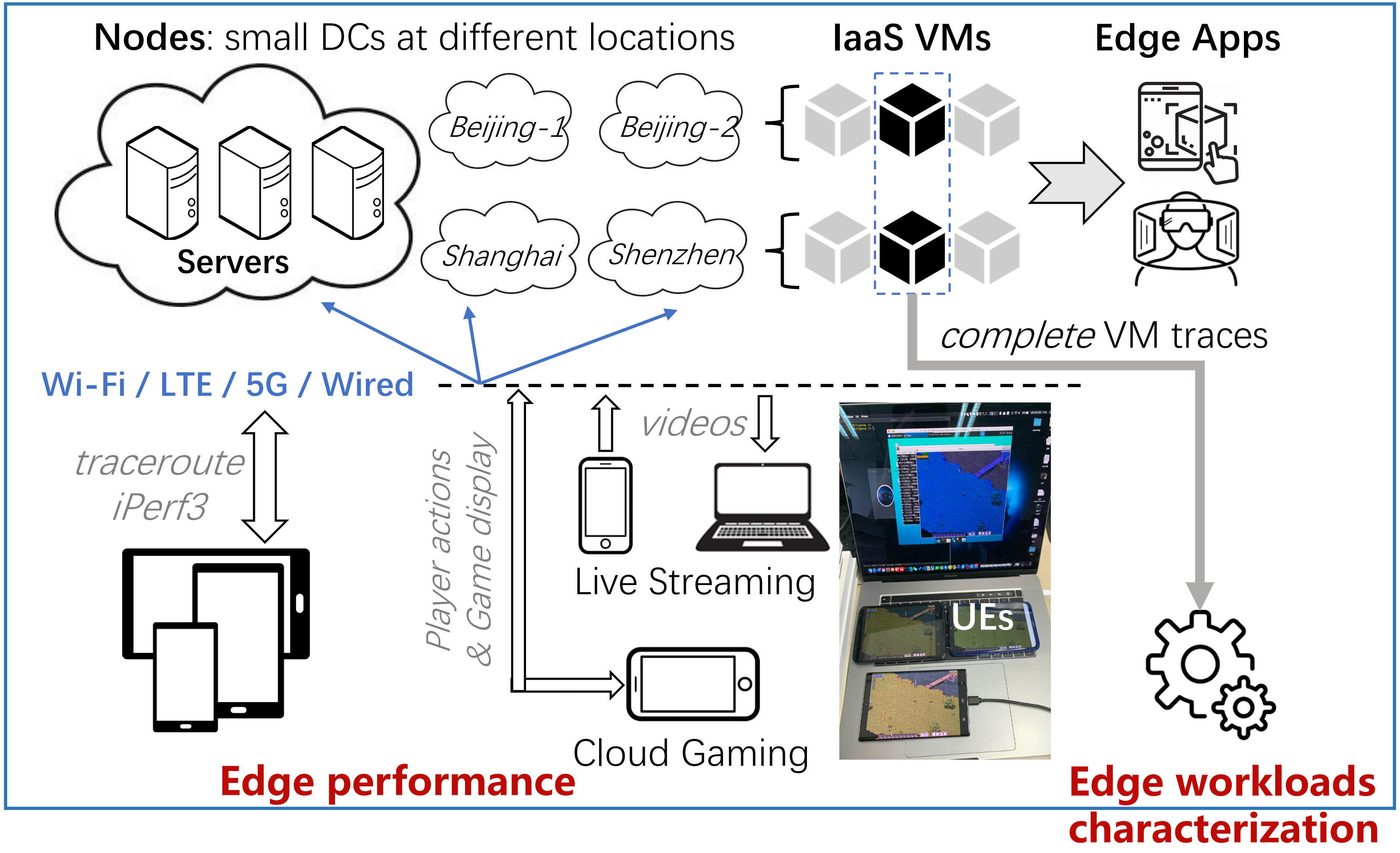
- Latency reduction, application performance improvement.
- Resource saving (e.g., bandwidth).

But all parties are still wondering:

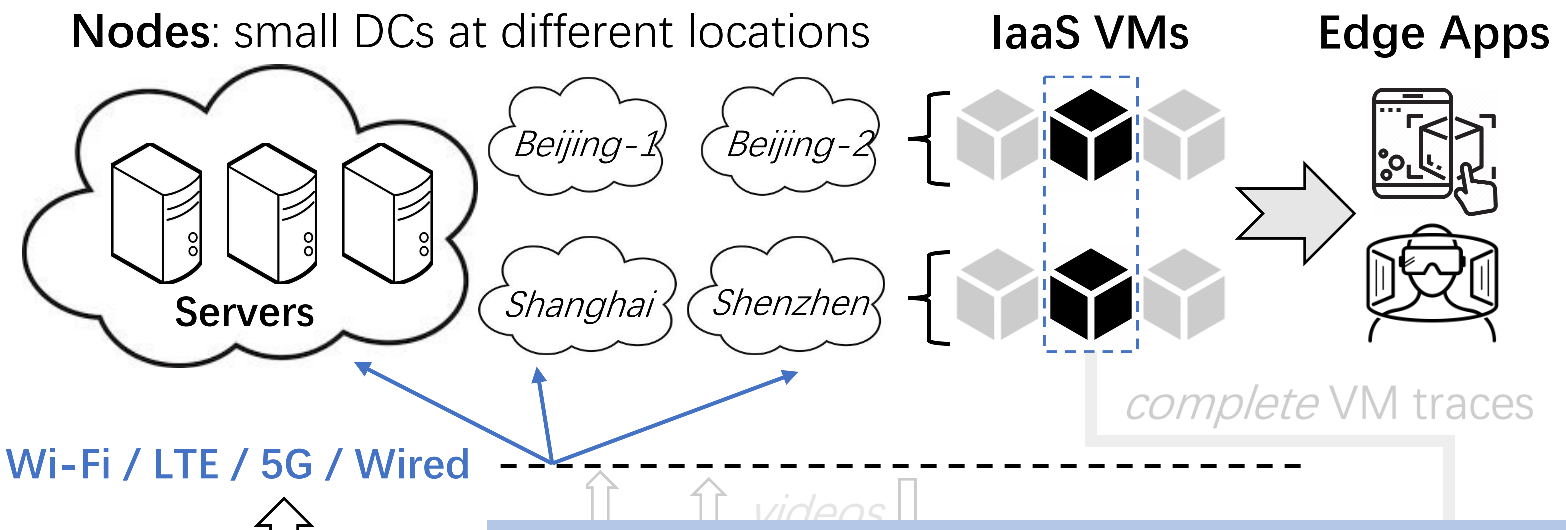
- End users: How much latency and QoE are improved?
- Cloud providers: How much bandwidth and other resources are saved?

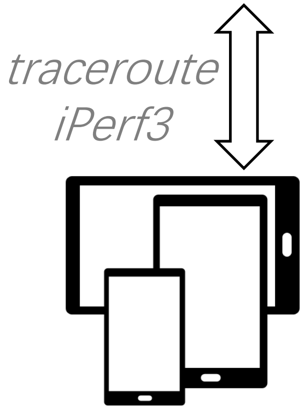
We need quantitative characteristics of NEP-like edge platforms.

Measurements over NEP



Measurements over NEP

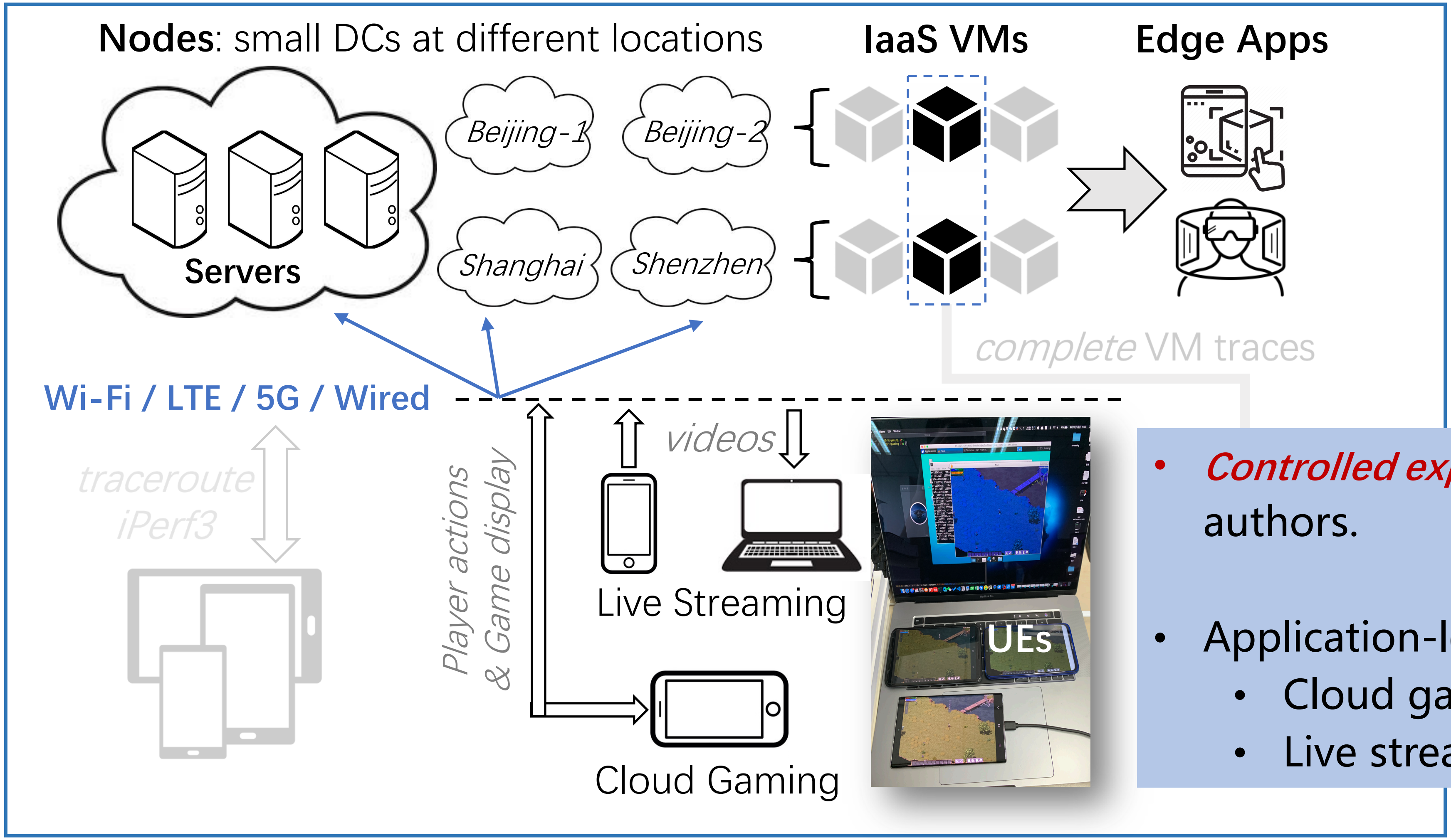




- *Crowd-sourcing experiments* from *158* participants in *41* cities across China.

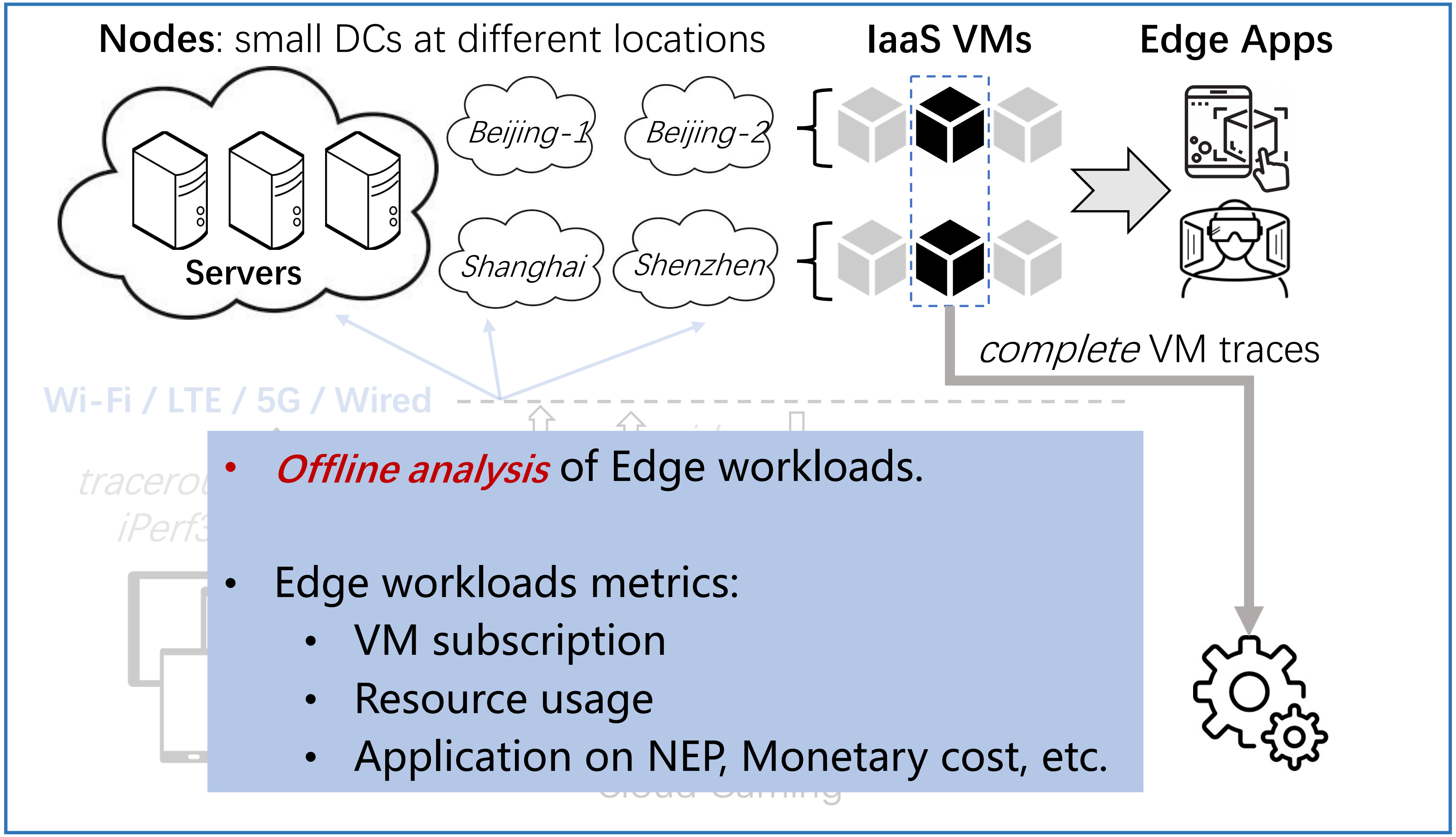
- Edge performance metrics:
  - Network latency
  - Network throughput

Measurements over NEP



- *Controlled experiments* by authors.
- Application-level QoE:
  - Cloud gaming
  - Live streaming

Measurements over NEP



- *Offline analysis* of Edge workloads.

- Edge workloads metrics:
  - VM subscription
  - Resource usage
  - Application on NEP, Monetary cost, etc.

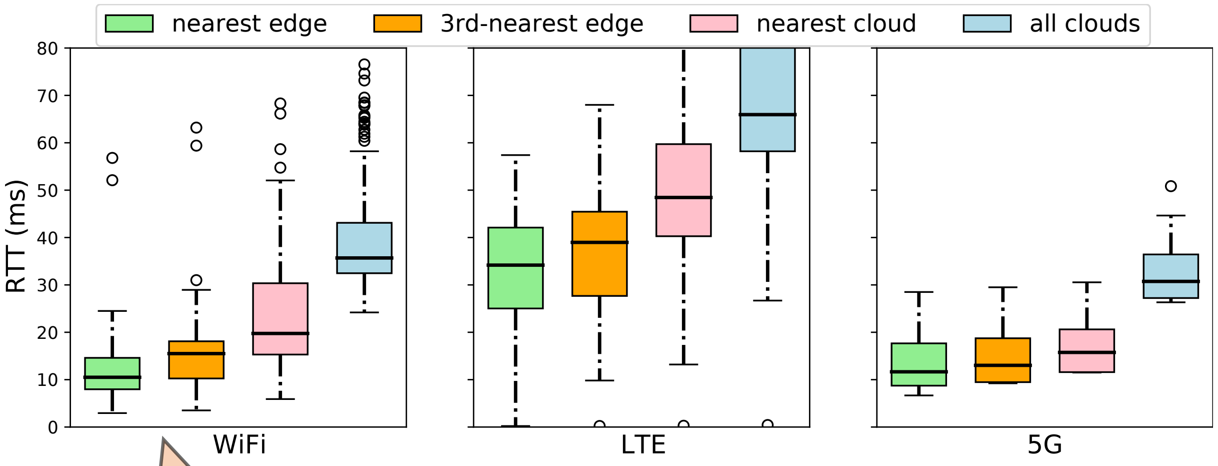
Experiment Settings

- Edge and cloud servers
  - Network latency: one VM on each edge site[1] of NEP and each cloud region of AliCloud.
  - Network throughput: 20 NEP VMs at different sites.
- Crowd-sourcing participants
  - Network latency: 158 users from 20 provinces, 41 cities in China, 59%/34%/7% of the results under Wi-Fi/LTE/5G.
  - Network throughput: 25 volunteers which is a subset from above 158 users.
- Software
  - Use traceroute (ICMP) and iPerf3 (TCP) to obtain the network latency and throughput.

[1] A site refers to a datacenter at some location, which consists of many servers, and each server hosts many VMs.

Edge Performance: Latency

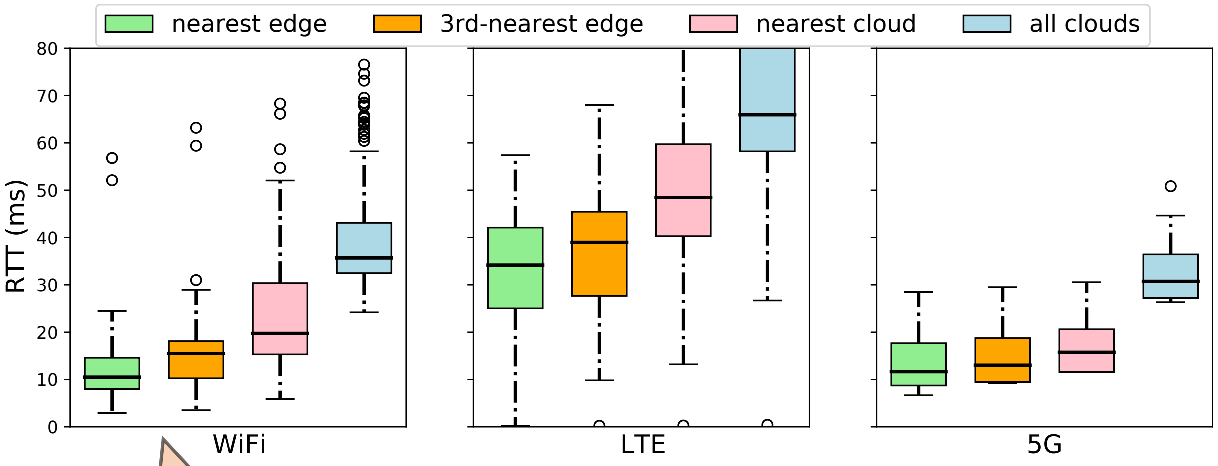
Median RTT across users



1.9X/3.4X faster than the nearest cloud/all clouds!

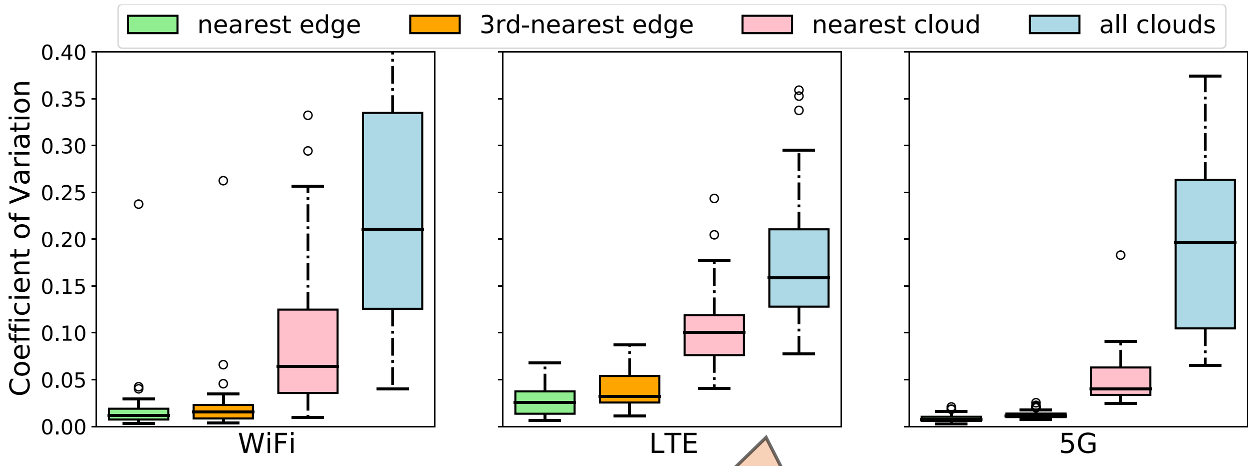
Edge Performance: Latency

Median RTT across users



1.9X/3.4X faster than the nearest cloud/all clouds!

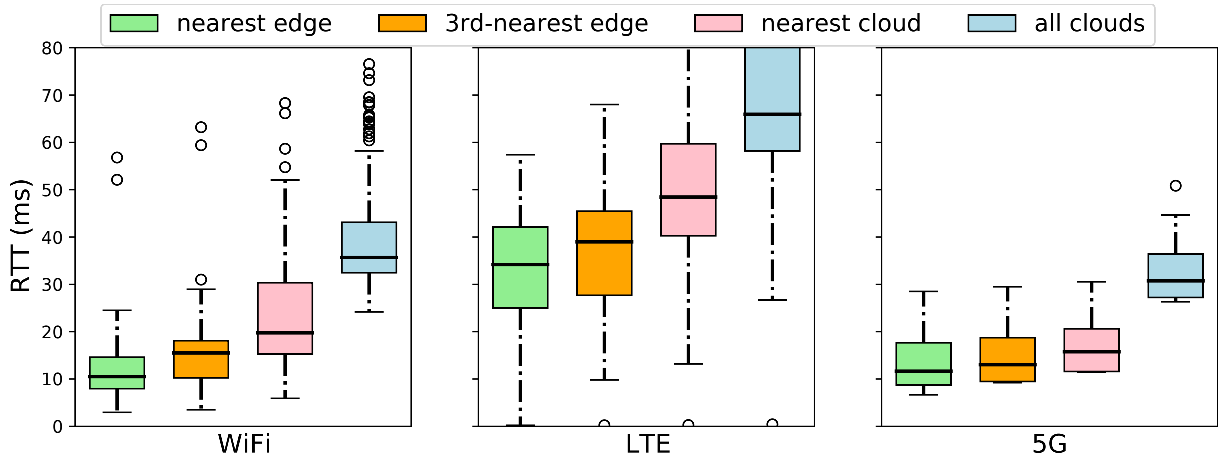
RTT coefficient of variation (CV) across users



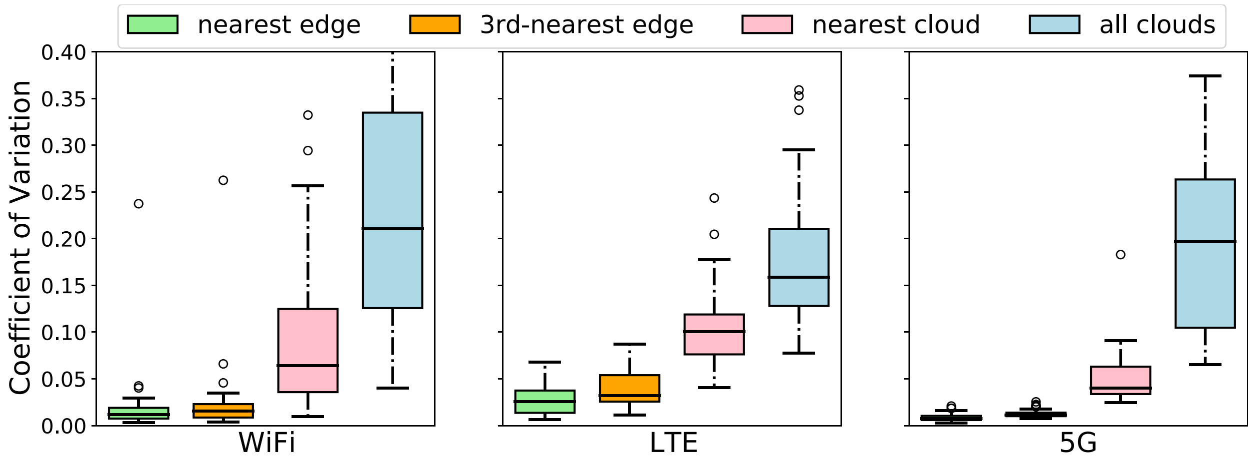
The nearest cloud has 5.8X/3.9X/5.7X higher median RTT CV under Wi-Fi/LTE/5G!

Edge Performance: Latency

Median RTT across users

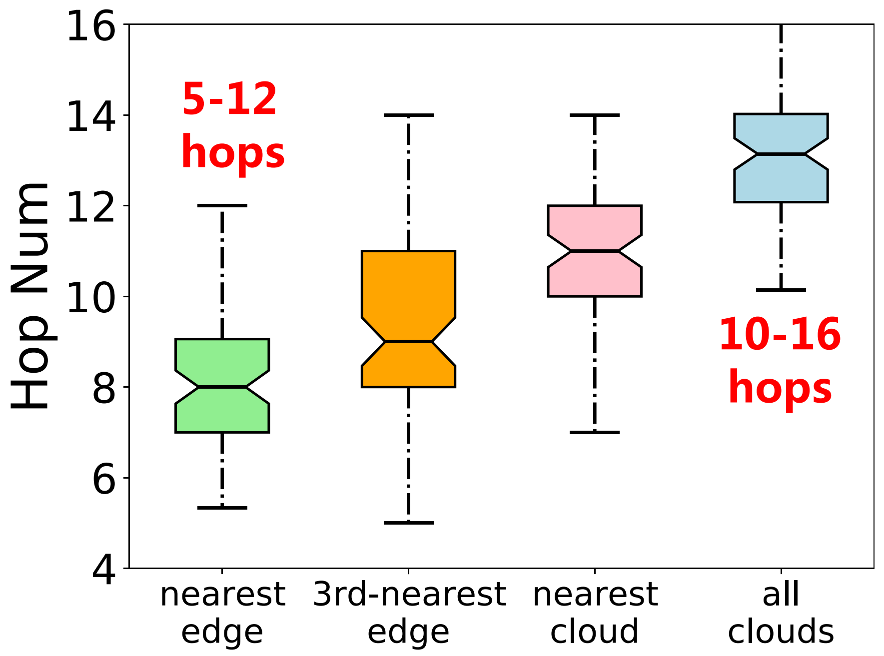


RTT coefficient of variation (CV) across users

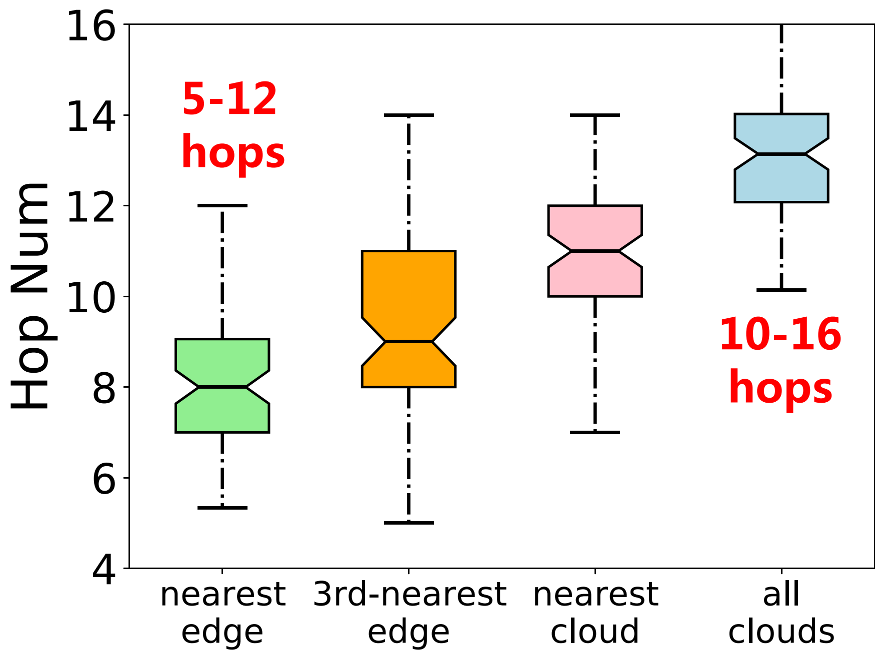


- **Edge VMs deliver lower and more stable network latency compared to Cloud VMs.**
- **The best RTT remains 10.4ms to the nearest edge VM under 5G.**

Edge Performance: Hop number



Edge Performance: Hop number

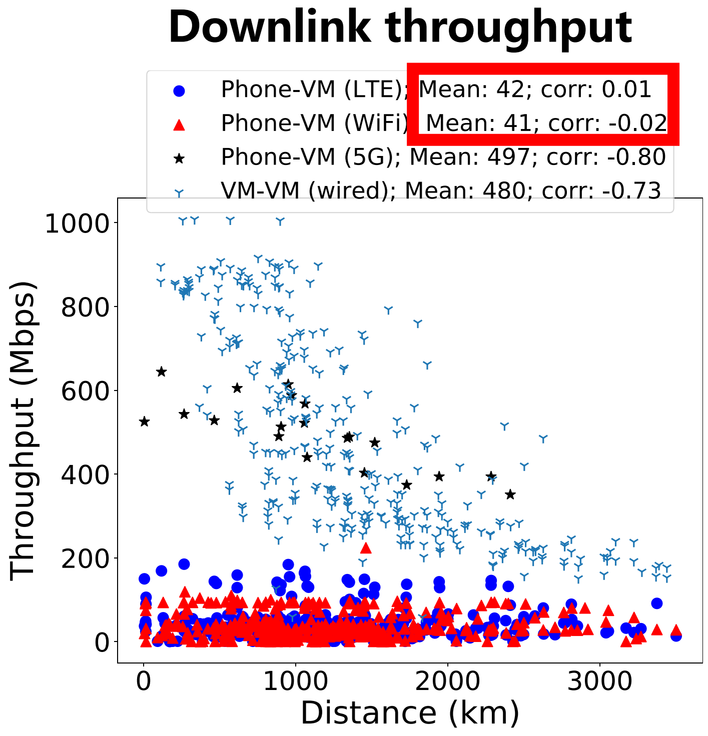


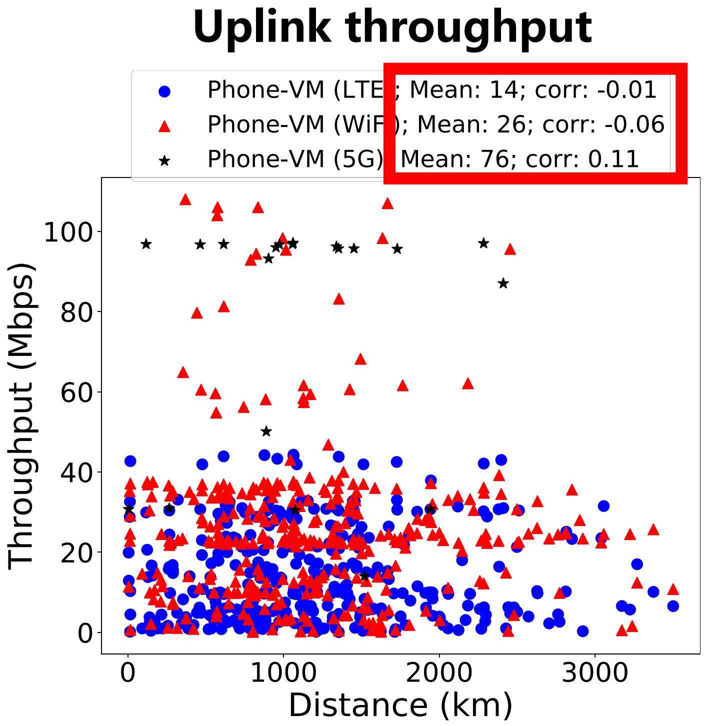
	Nearest edge site		Nearest cloud site	
	1st-2nd-3rd hop	Rest	1st-2nd-3rd hop	Rest
WiFi	44.2%-10.3%-15.1%	30.2%	30.1%-5.0%-11.5%	52.5%
LTE	10.2%-70.1%-9.4%	10.3%	10.1%-51.6%-13.1%	25.2%
5G	97.9% in total	2.1%	82.2% in total	17.8%

- The reduced hop number leads to lower network latency and jitter.
- The first 3 hops dominate more than 70 percent of the total latency for the nearest edge site.

Implication: To reached the envisioned prospects of edge computing, NEP needs to deploy denser sites and collaborate with operators to sink the edge resources into ISP 's core networks or even cellular base stations.

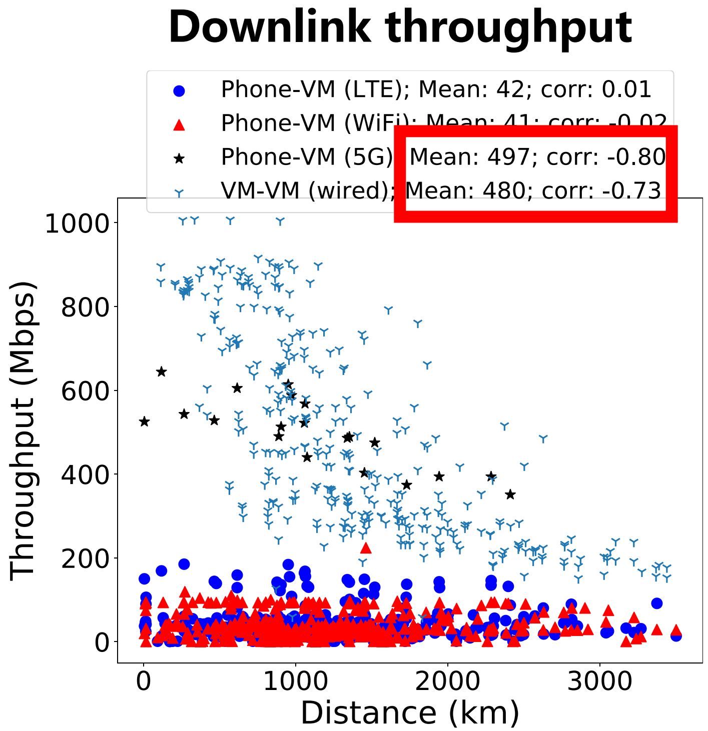
Edge Performance: Throughput

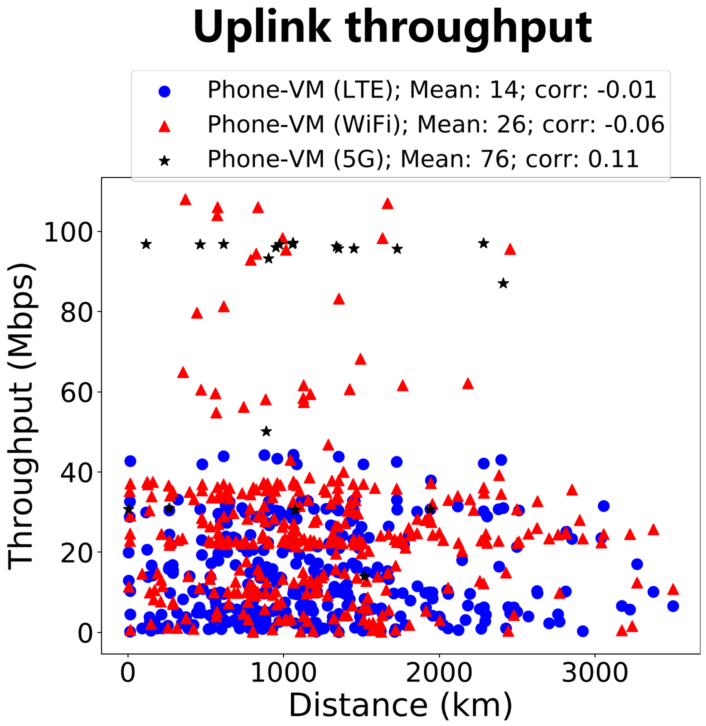




- TCP-based throughput testing.
- Running iPerf3 for 15 seconds each test.
- Corr is calculated by *Pearson correlation coefficient.*

Edge Performance: Throughput



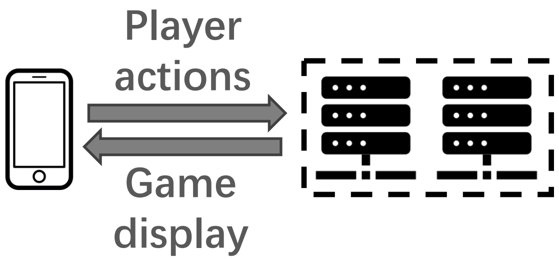


- TCP-based throughput testing.
- Running iPerf3 for 15 seconds each test.
- Corr is calculated by *Pearson correlation coefficient.*

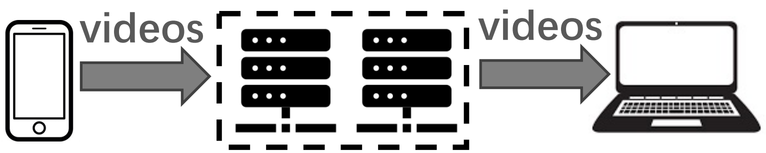
Implication: Bringing resources closer to users improves network throughput on NEP only with high bandwidth capacity at the last mile. Throughput improvement will benefit more emerging, bandwidth-hungry edge applications in the future.

Experiment Settings: Application QoE

- Deploy one nearest edge and three cloud servers (670KM/1300M/2000KM away from experiment performed!)
- Software deployment:
  - GamingAnywhere[1] for cloud gaming.
  - End-to-end streaming app using RTMP protocol.



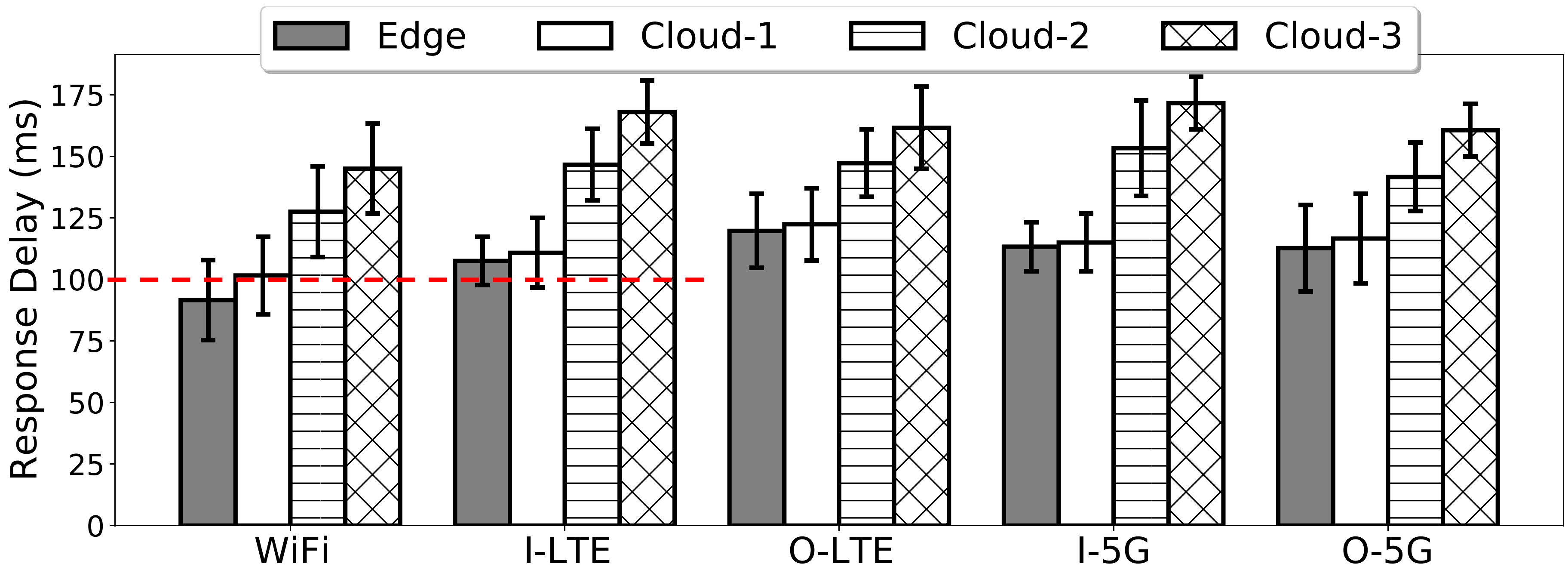
Cloud gaming



Live Streaming

[1] GamingAnywhere – An Open Source Cloud Gaming System https://gaminganywhere.org/

Edge Performance: Cloud Gaming



Implication: Placing gaming backend on nearby NEP edges help achieve less than 100ms response delay. To further enhance the experience, we need to improve the server-side gaming execution.

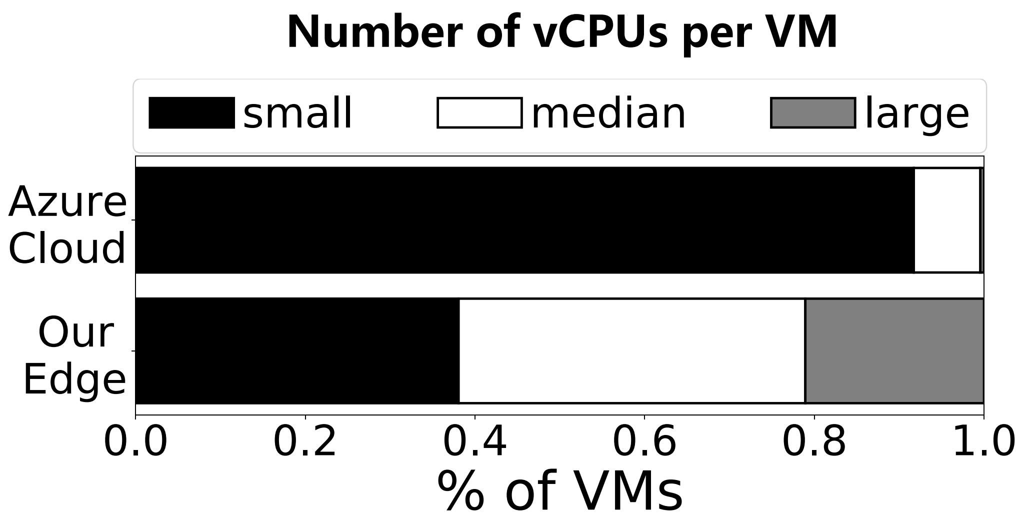
Edge Workloads: Total Analytics

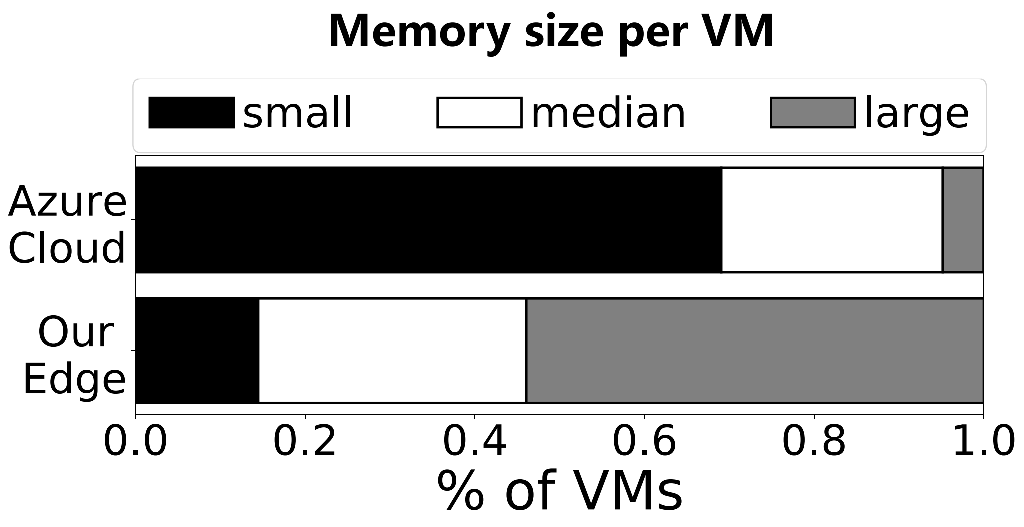
Compare with Azure cloud dataset[1]

- VM subscription
- CPU usage
- Application of NEP (e.g., Application type, VM numbers per application)
- Bandwidth usage
- Resource load balance
- Monetary of NEP

[1] Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms. (SOSP '17).

Edge Workloads: VM subscription

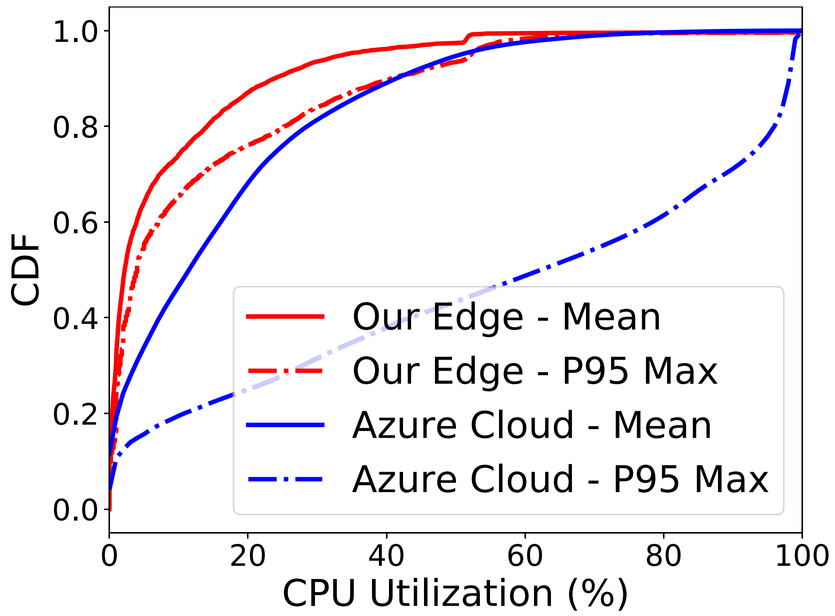


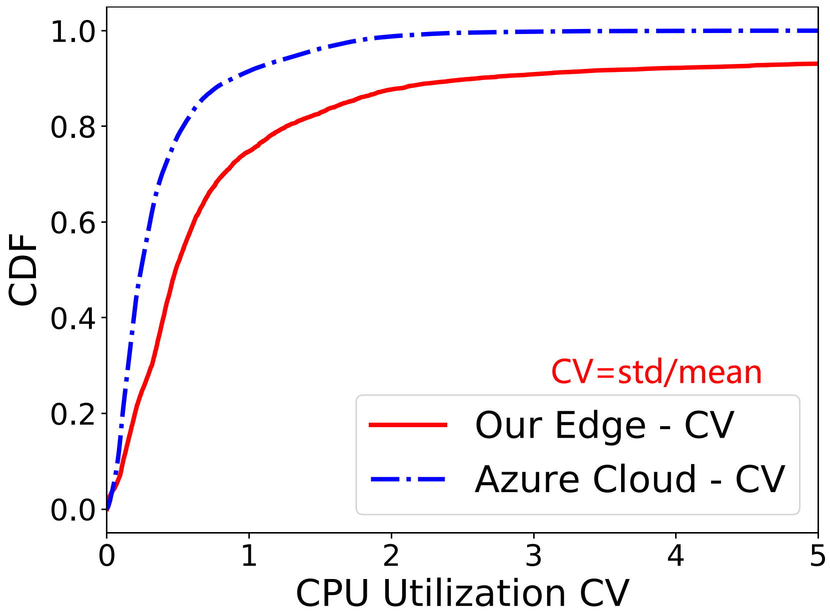


Small / median / large represents ≤4 / 5-16 / >16 CPU cores or GBs memory.

Implication: The large VM size on NEP-like edge platforms may cause severe resource fragmentation. Dynamic VM migration and resource disaggregation may help to solve this problem.

Edge Workloads: Resource usage





Implication: The relatively low but highly skewed CPU usage challenges the NEP' s VM management. To better utilize the CPU resources, NEP may need smart VM placement algorithms or employ more elastic computing forms, e.g., containers together with IaaS VMs on the same server.

Conclusions

- The first comprehensive measurement on a commercial, multi-tenant edge platform.
- Lead to insightful implications for designing future edge platforms and edge-based applications.
- Edge workloads traces are open-sourced at: https://github.com/xumengwei/EdgeWorkloadsTraces

Thanks for your listening!

Edge workloads:

https://github.com/xumengwei/EdgeWorkloadsTraces

Special thanks to Alibaba Group for their contribution to this work!