

# Towards Scalable Edge-Native Applications

Junjue Wang, Ziqiang Feng, Shilpa George, Roger Iyengar (CMU), Padmanabhan Pillai (Intel Labs), Mahadev Satyanarayanan (CMU)

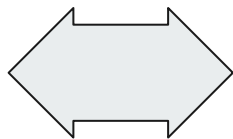
*The Fourth ACM/IEEE Symposium on Edge Computing (SEC'19)*

# Tiered Model of Computing



Tier-3

**Low-Latency  
High-BW**



**Wireless**



Luggable



Vehicular



Coffee Shop  
Mini-datacenter

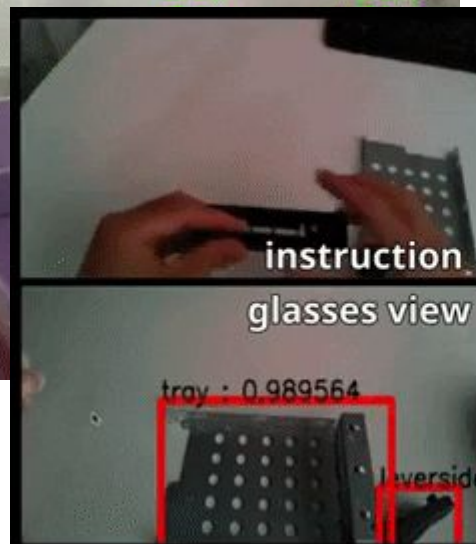
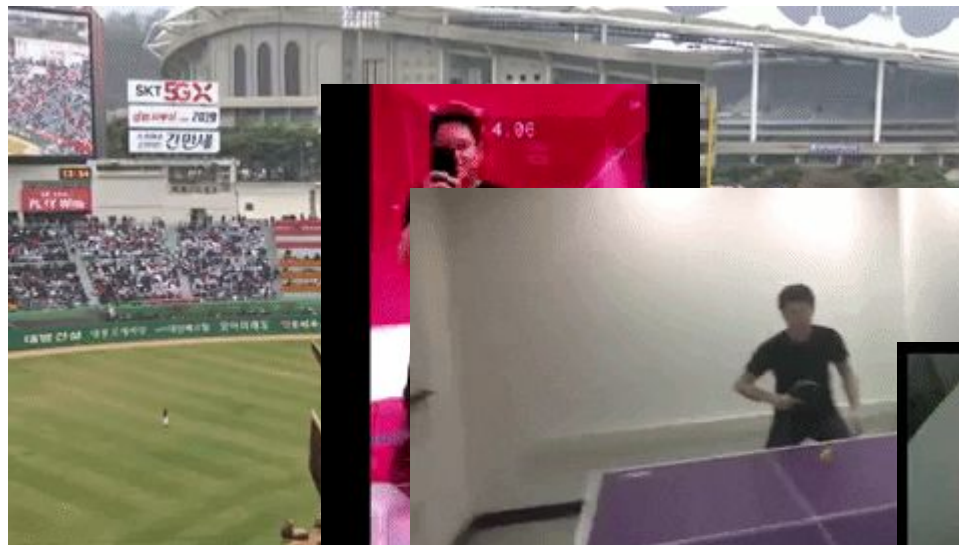
Tier-2

**Wide Area Network**



Tier-1

# Examples of Edge-Native Applications



Latency is guaranteed by over-provisioning.



---

How to build  
**multi-user**

edge computing systems that  
**preserve low latency**  
even as load increases?

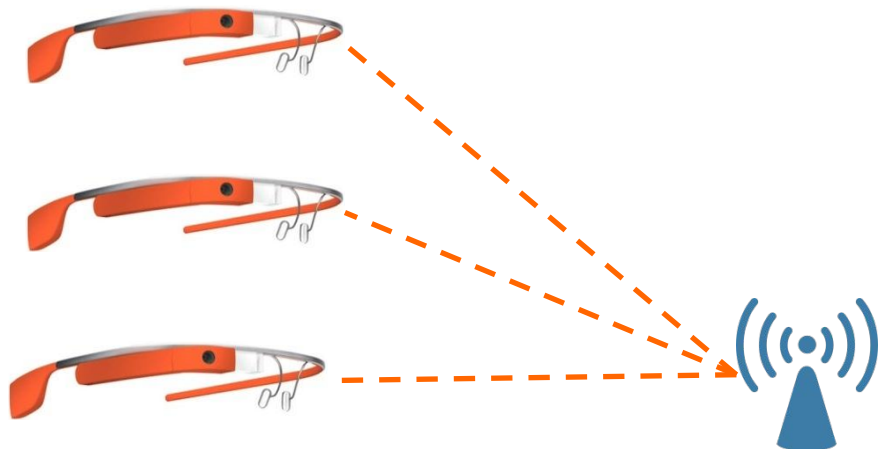


# Outline

- Motivation & Problem Statement
- Adaptation-Centric Approach and Architecture
- Workload Reduction for Wearable Cognitive Assistance
- Utility-based Cloudlet Resource Allocation
- Evaluation
- What's More in the Paper

# Adaptation Approach for Scalable Wearable Cognitive Assistance

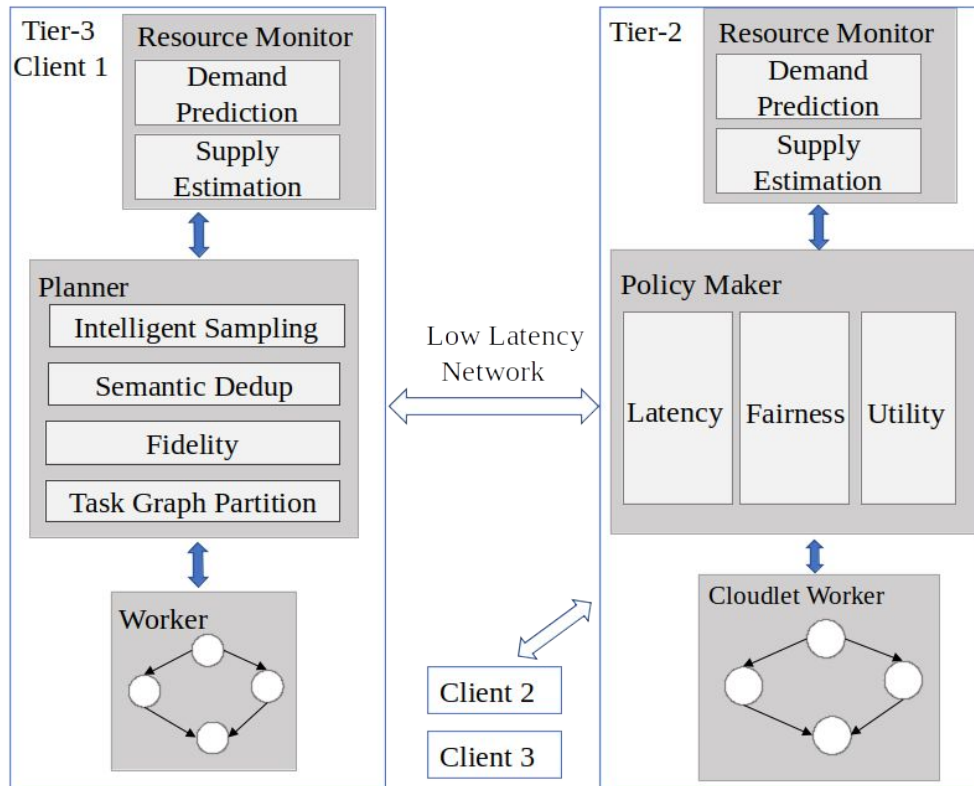
Reduce Offered Load  
(Application Assisted)



Adaptation-Centric Resource  
Management at Tier-2



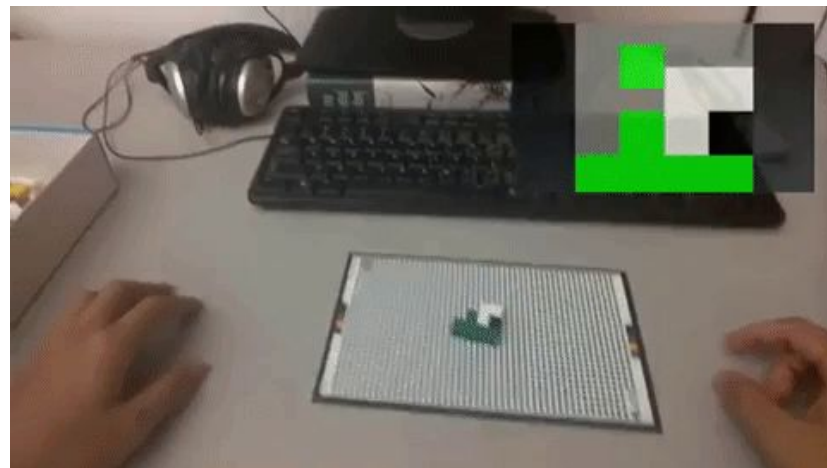
# Adaptation Architecture



## Reduce Offered Load: Adaptive Sampling

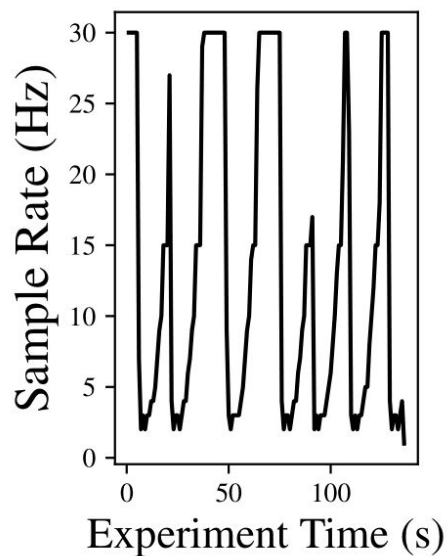


V.S.





## Reduce Offered Load: Adaptive Sampling



Trace	Sample Half Freq	Adaptive Sampling
1	50%	25%
2	50%	28%
3	50%	30%
4	50%	30%
5	50%	43%

(a) Percentage of Frames Sampled

	Guidance Delay (frames $\pm$ stddev)
Sample Half Freq	7.6 $\pm$ 6.9
Adaptive Sampling	5.9 $\pm$ 8.2

(b) Guidance Latency

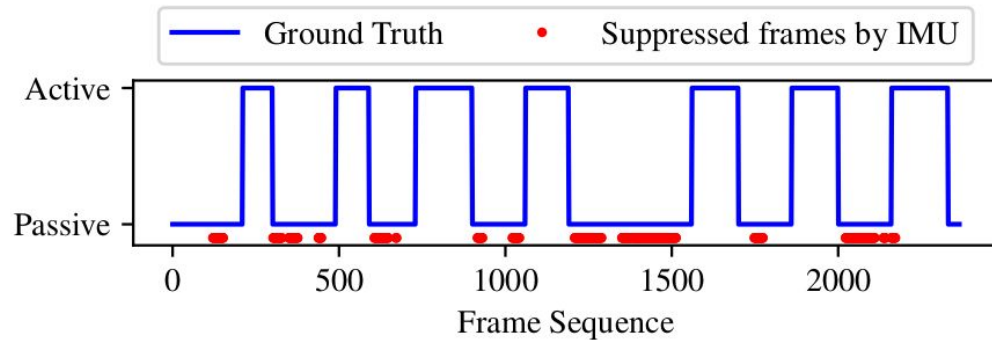
## Reduce Offered Load: IMU-based Passive Phase Suppression



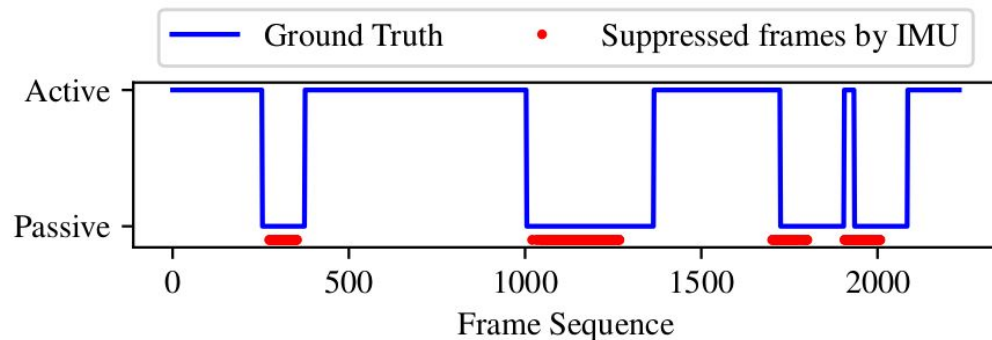
V.S.



# Reduce Offered Load: IMU-based Passive Phase Suppression



(a) LEGO



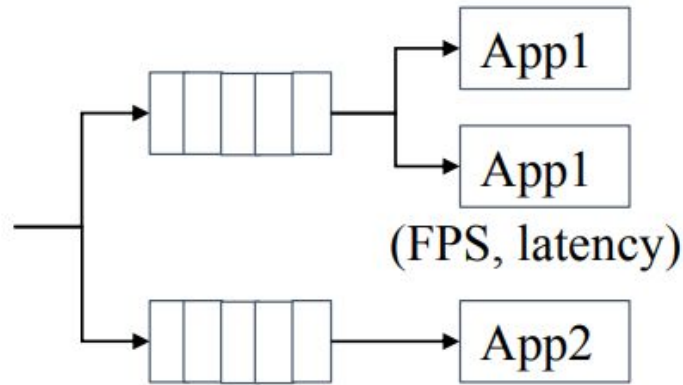
(b) PING PONG



# Developers Describe App Characteristics

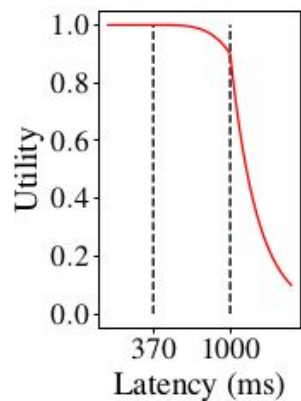
- How often are instructions given, compared to task duration? **Adaptive Sampling**
- Can IMUs be used to identify the start and end of user activities? **IMU-based Frame Suppression**
- Will a user wait for system responses before proceeding? **Key Frame Selection**
- ...

## Adaptation-Centric Resource Allocation (Tier-2)

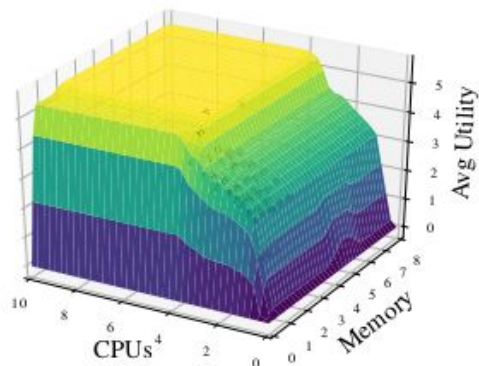


Only request flow is shown.

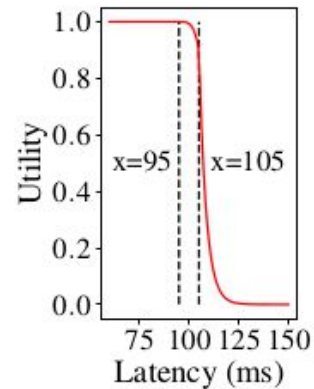
# QoS-Centric Profiling



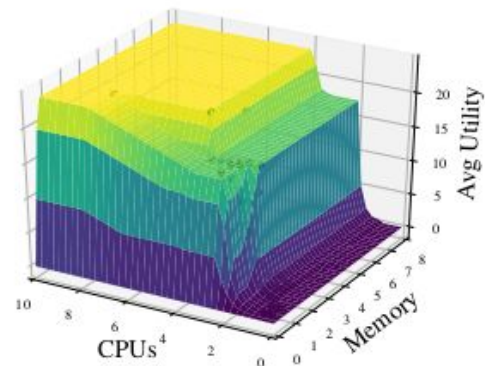
(a) Utility For FACE



(b) Profile for FACE



(a) Utility For POOL



(b) Profile for POOL



# Allocation Policy: Maximize Overall System Utility

Maximize **Total Utility of the System**

(sum of utilities of all client sessions)

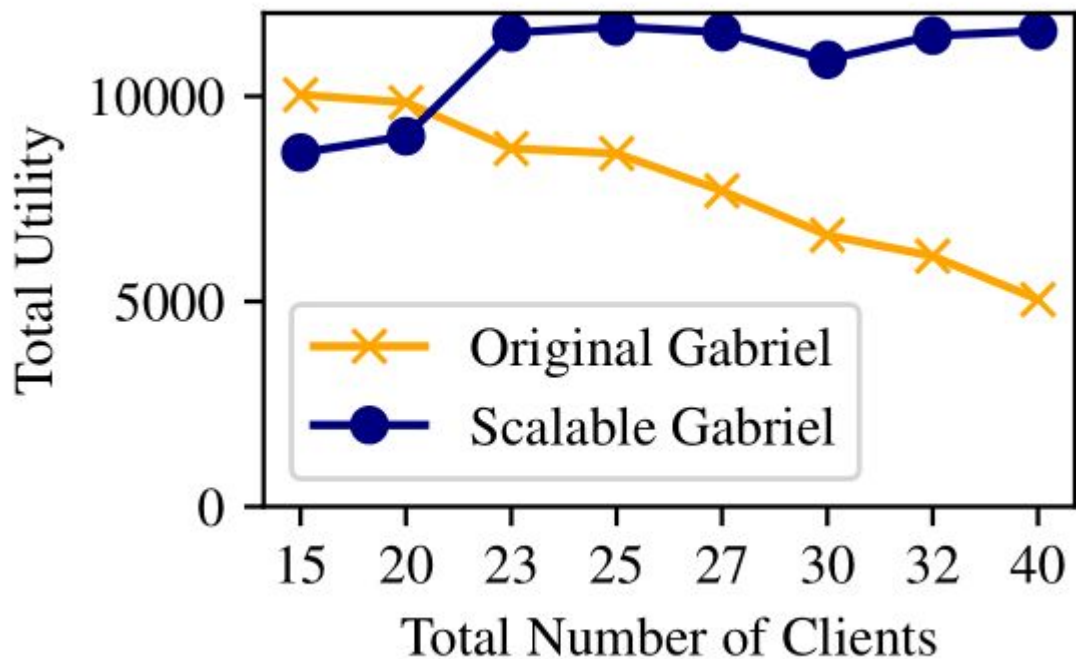
Subject to

*Total Allocated Resources  $\leq$  Total Resources*

*Allocated Resources Per client  $\geq 0$*

*Total Allocated Resources Per App  $\leq$  Upper  
Bound Proportional to Number of Client*

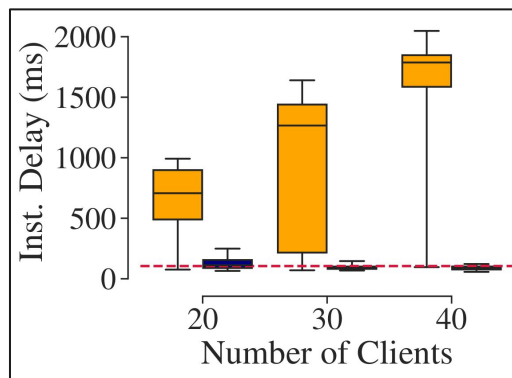
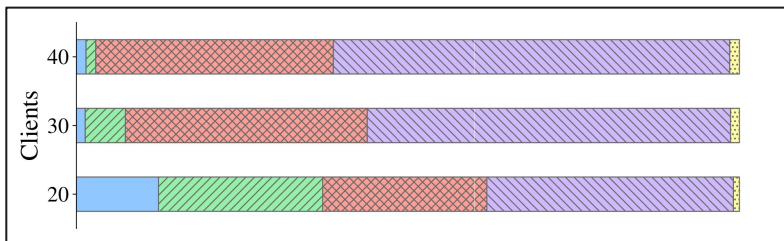
# Evaluation: Adaptation-Centric Resource Allocation



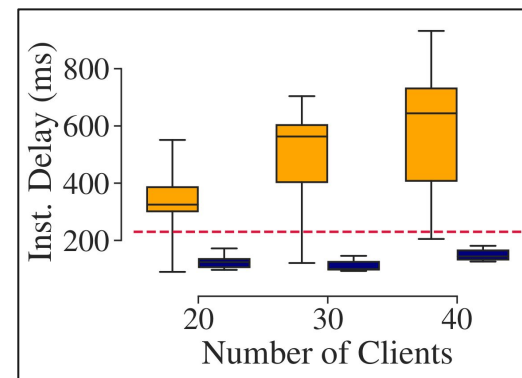


# Evaluation: User Experienced Latencies

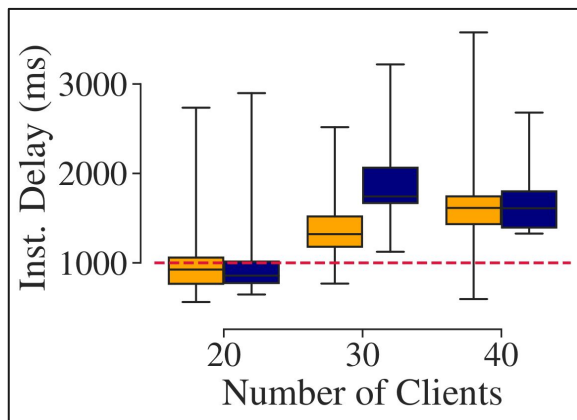
FACE LEGO PING PONG POOL IKEA



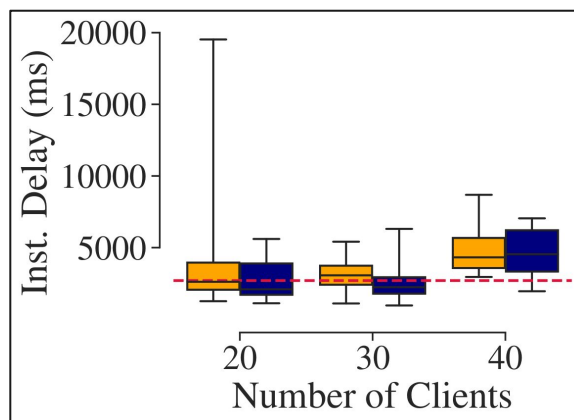
Pool



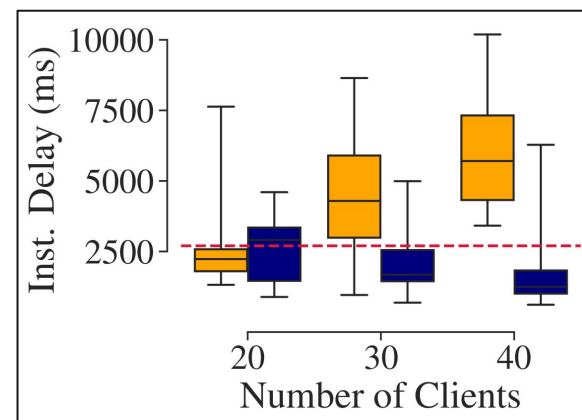
Pingpong



Face



Lego



Ikea



## More in the Paper

- Application Characteristics of Wearable Cognitive Assistance
- A Complete Taxonomy for Adaptation
- Detailed Implementation of Workload Reduction
- Detailed Evaluation of Our Cloudlet Resource Allocation Scheme
- Related and Future Work

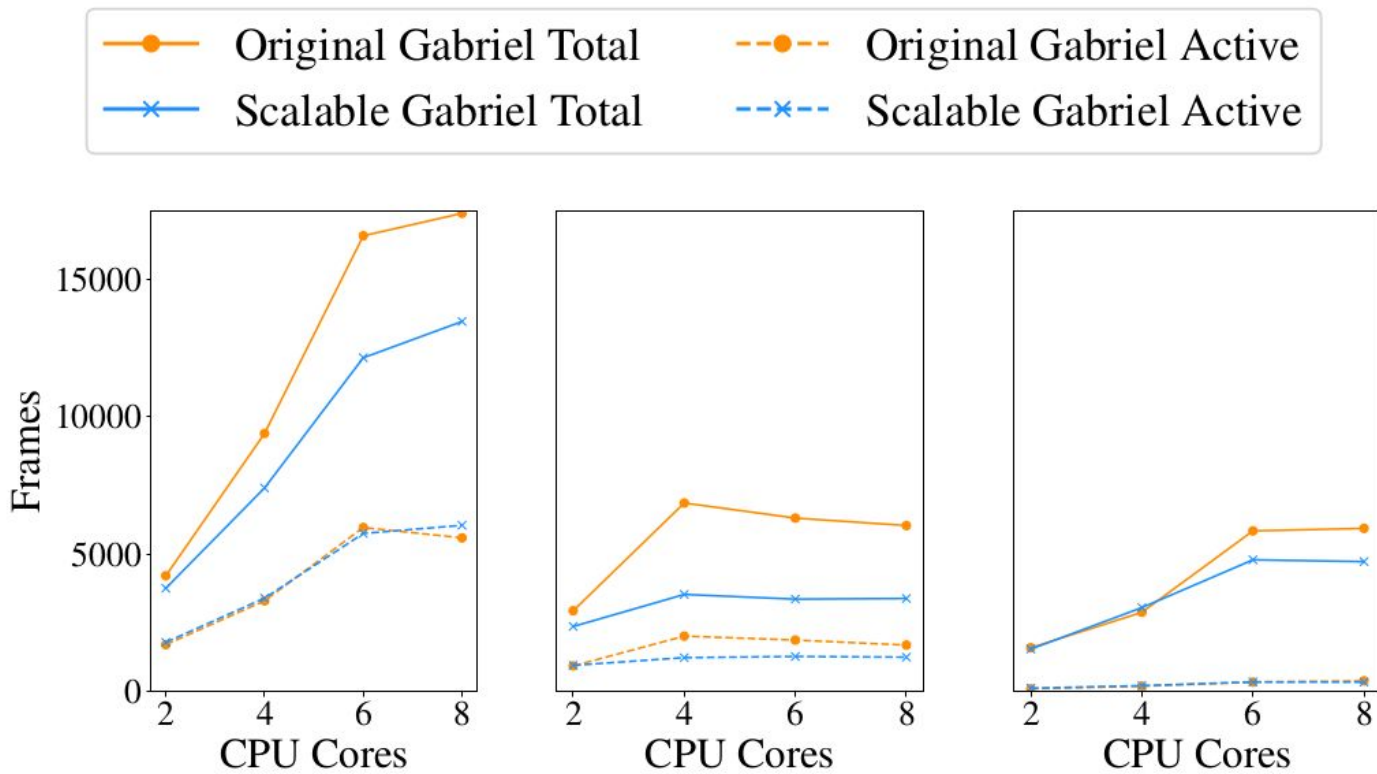


# Thank you! Questions?



# Backups and Unused Slides

# Evaluation --- Workload Reduction



# Taxonomy for Reducing Offered Load

