Technical
University
of Munich

Making orchestration application aware: A case for augmented reality at the edge

Giovanni Bartolomeo, Nitinder Mohan, Jörg Ott
Technical University of Munich

From our work:
Bartolomeo, G., Cao, J., Su, X., & Mohan, N.
Characterizing distributed mobile augmented reality applications at the edge.
CoNEXT '23 Companion

Artifacts Available acm V1.1

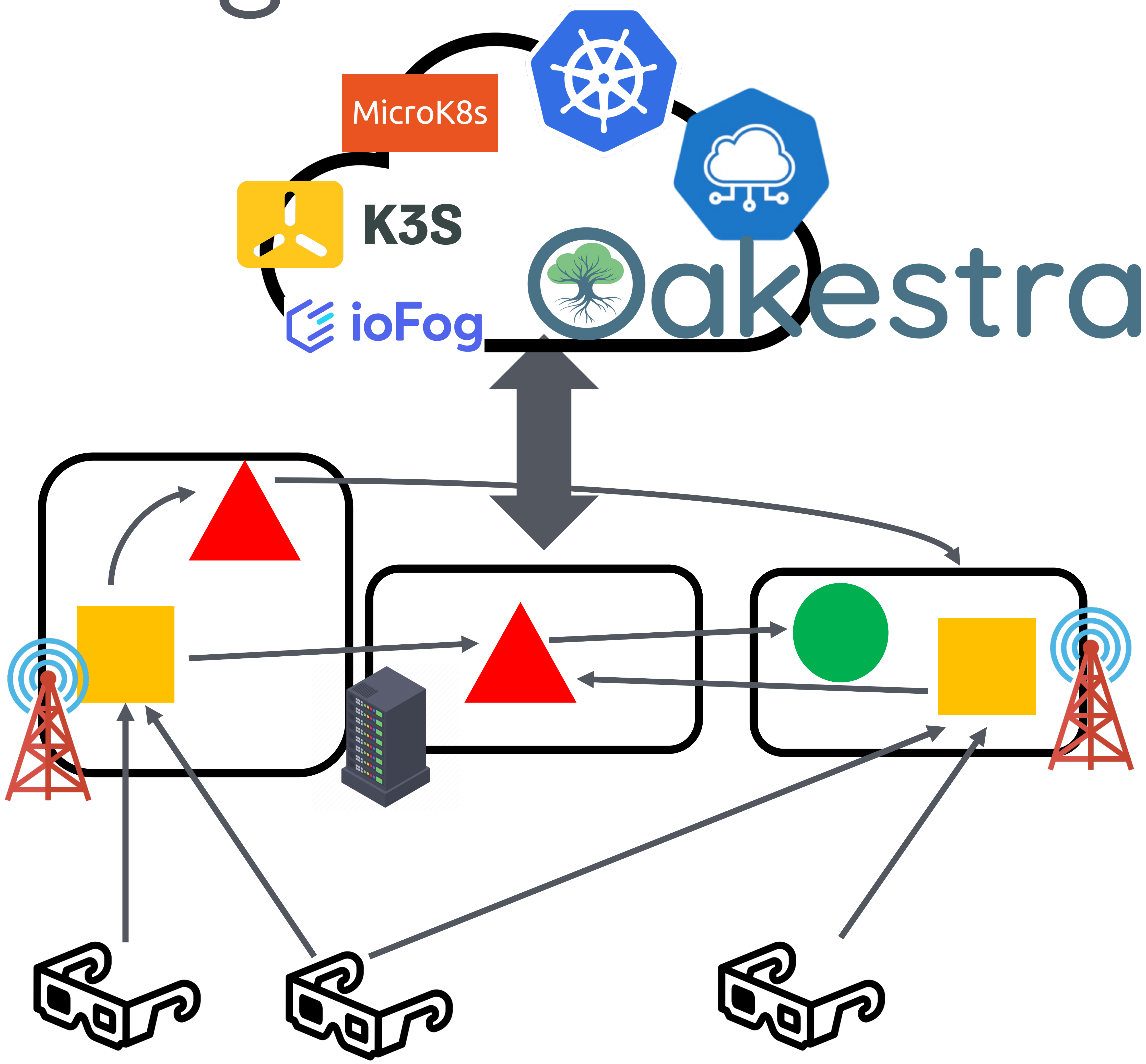
Artifacts Evaluated acm Functional V1.1

Results Reproduced acm V1.1

Distributed AR at the Edge

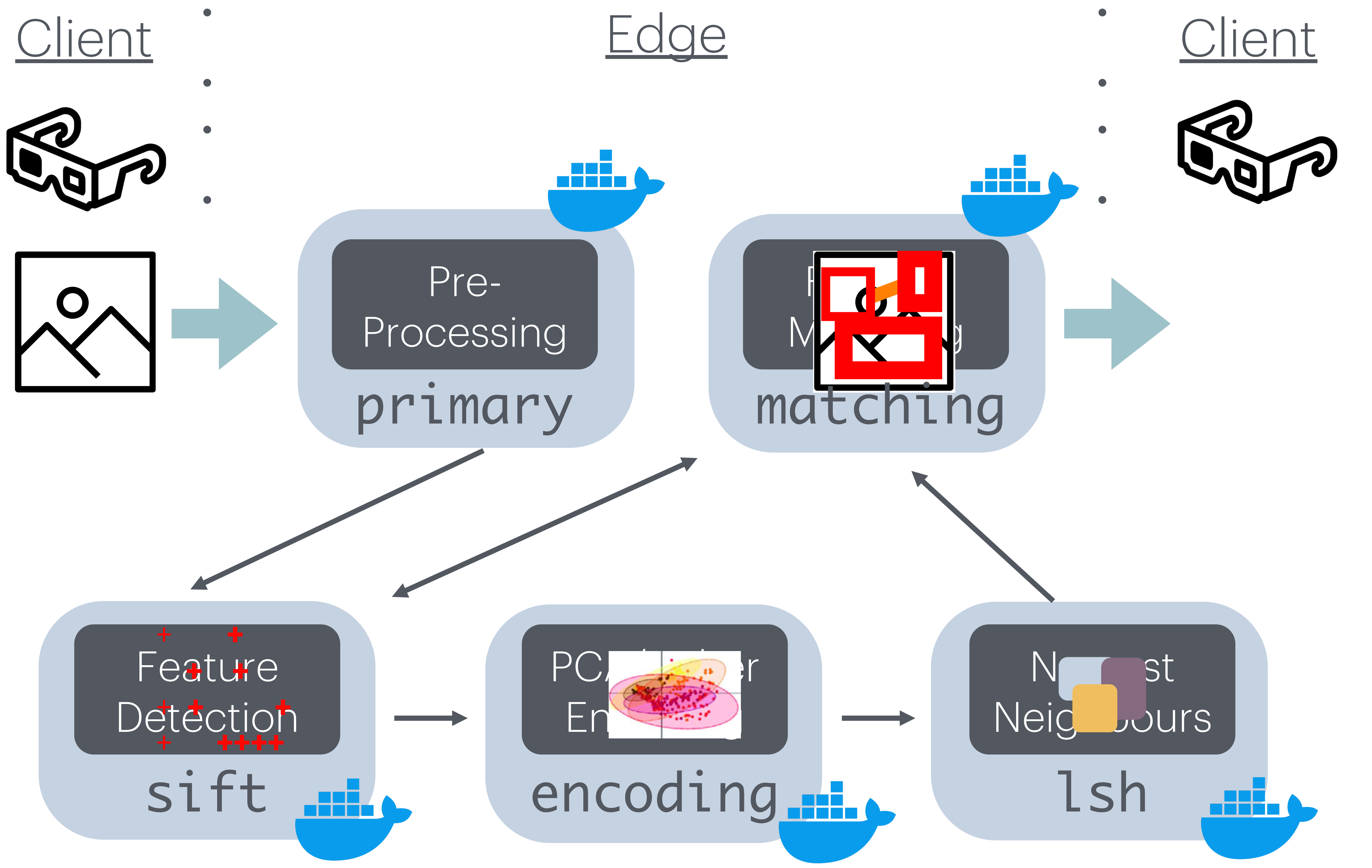
Challenges:

- Decomposition of AR functionalities
- Service placement across heterogeneous resources
  - Virtualization
  - Availability
  - CPU/GPU, Memory, Disk availability
- Collection of QoS & QoE metrics
- Server to Server network conditions



Our design approach

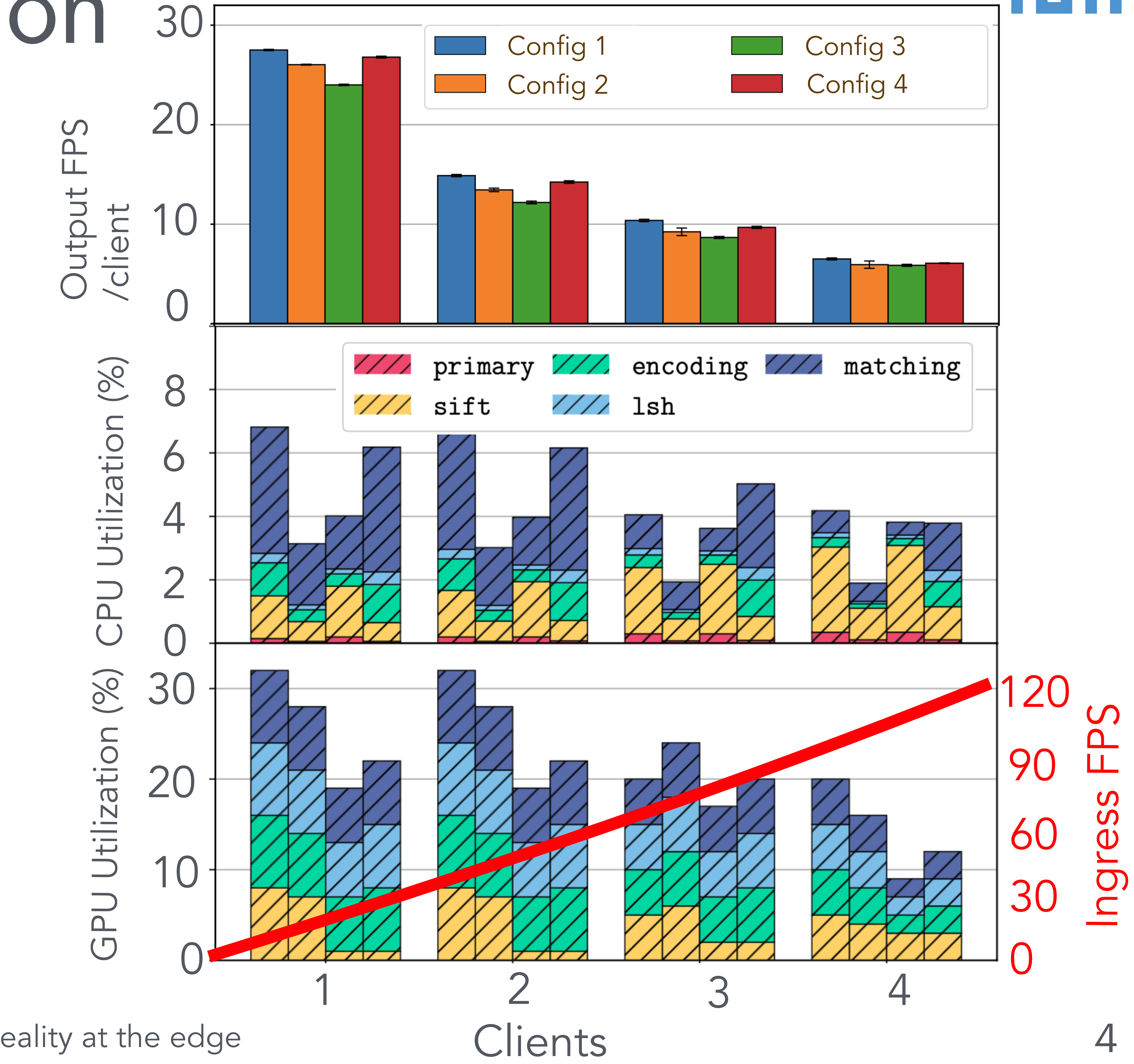
- Pipelined replicable workflow [MobiSys'18][IPDPSW][IEEECommunMag]
- Non-linear component interactions with stateless and stateful services [MMSys'23][EdgeSys'22][NSDI22]
- Full GPU offloading [MM'18]
- Multi-tenant capabilities [MobiCom'19]



Resource Consumption

The viewpoint of the orchestration system

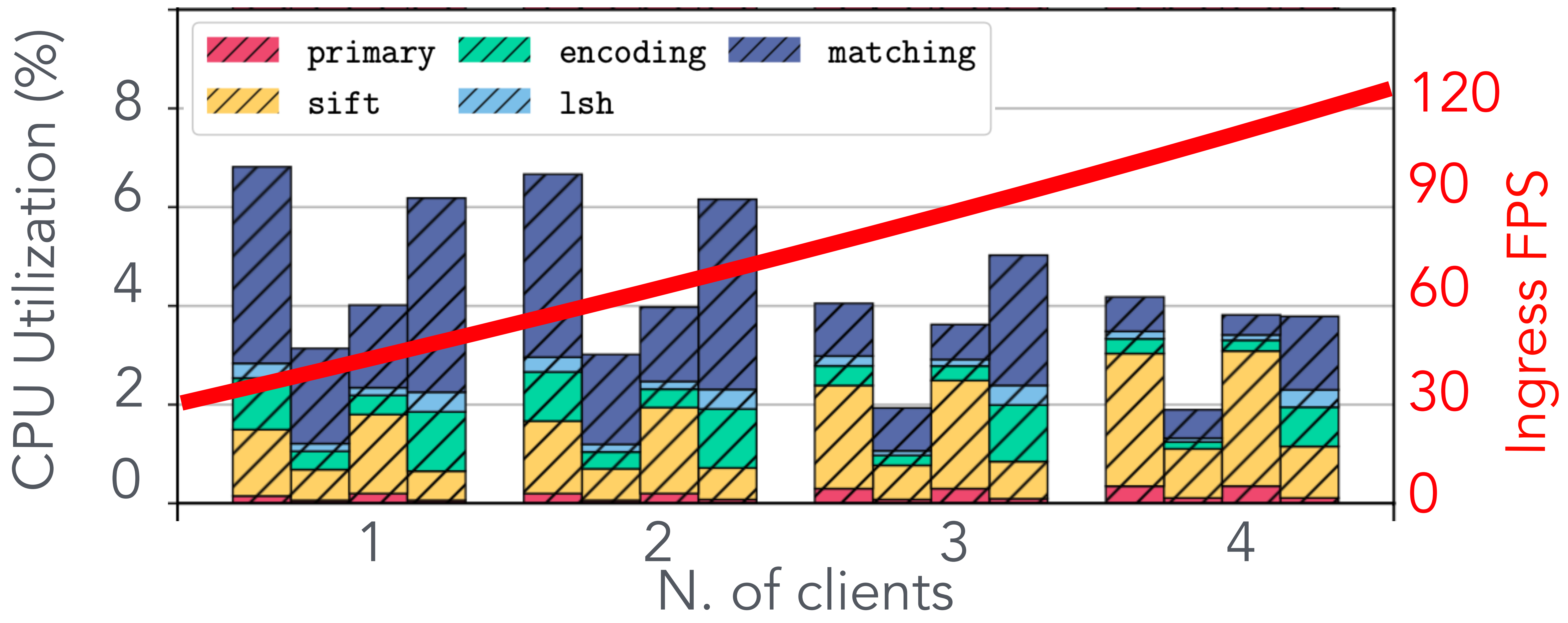
- Counterintuitive global decreasing in total CPU/GPU usage with increasing clients
- CPU% consumption only increases in sift service

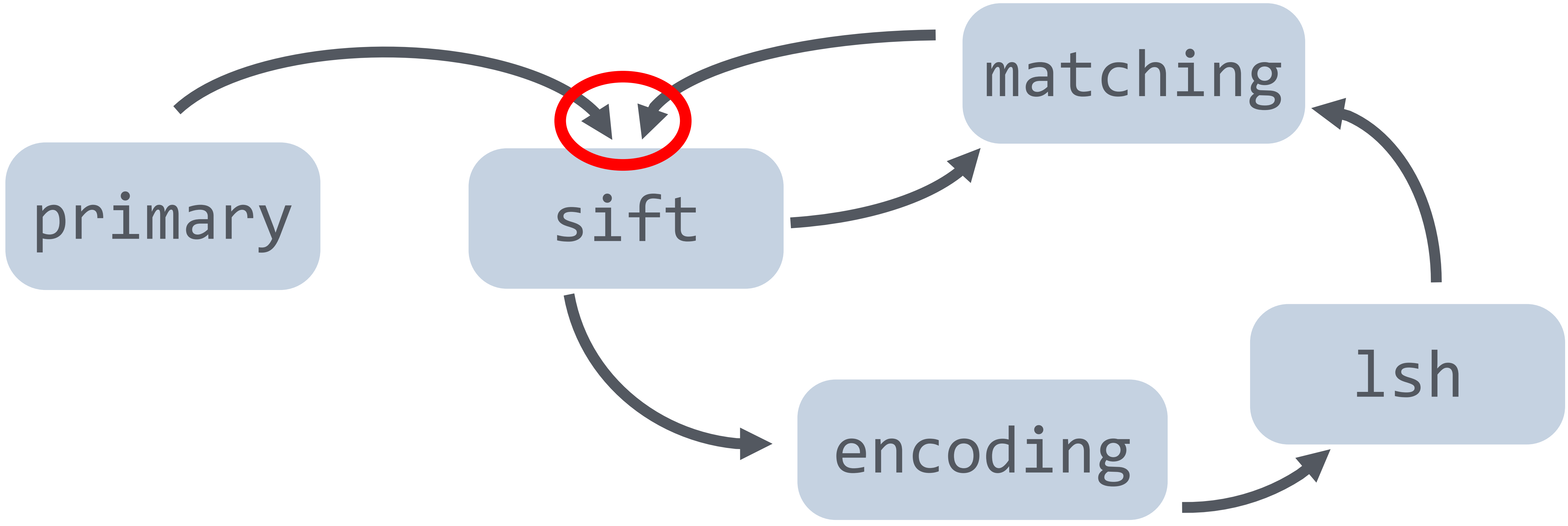


Resource Consumption

The viewpoint of the orchestration system

- Counterintuitive global decreasing in total CPU/GPU usage with increasing clients
- CPU% consumption only increases in sift service
- Careful examination reveals sift state retrieval and non-linear pipeline interactions as the main cause of congestion




matching

primary

sift

lsh

encoding

Motivation for Application Awareness

#1 Hardware utilization alone does not reflect application performance.

#2 It is extremely hard (if not impossible) to detect bottlenecks in these applications from high-level metrics

#3 Interdependence on stateful services in DSP affects the scalability.

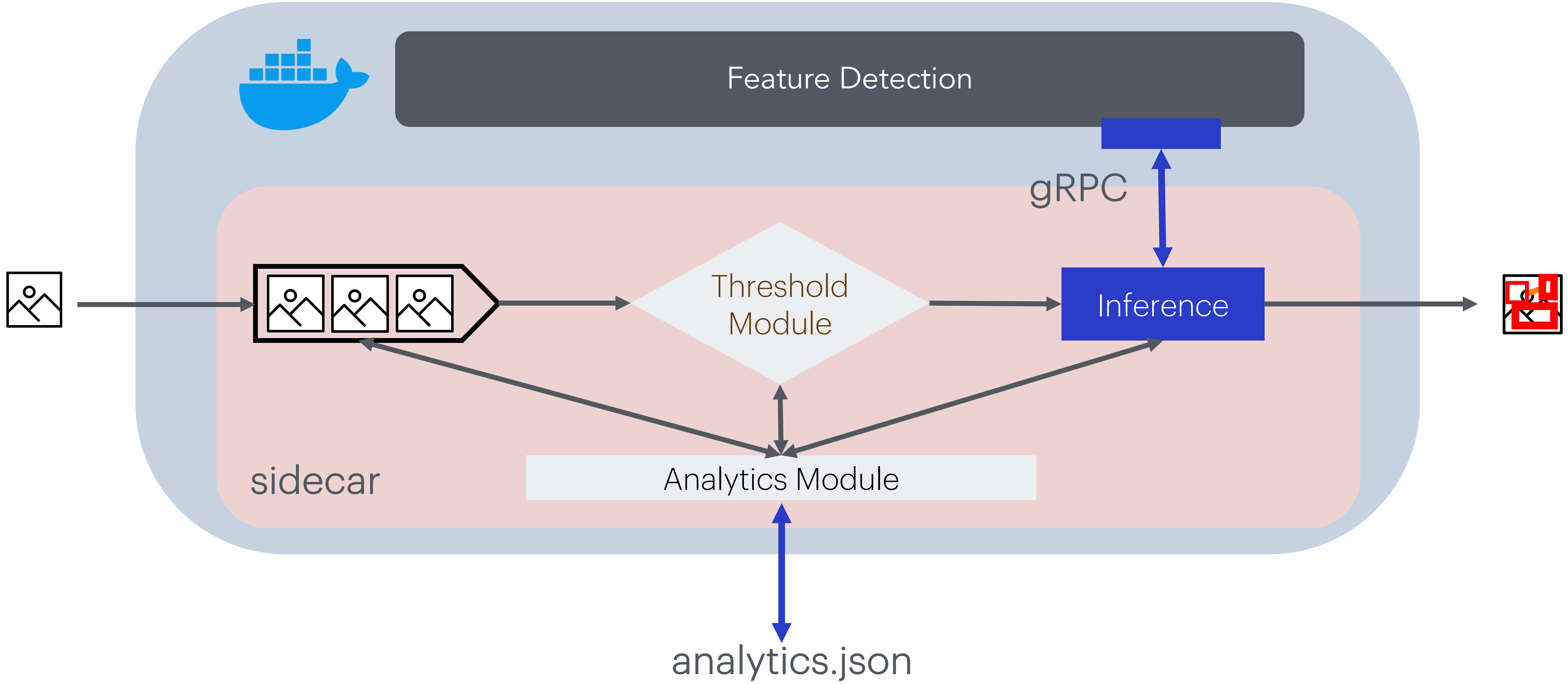
#4 There can be several other non-obvious internal application bottlenecks that orchestration remains oblivious towards

A revamped design

Pre-Processing
Feature Detection
PCA/Fisher Encoding
Nearest Neighbours
Feature Matching
primary
sift
encoding
lsh
matching

- Linearized the pipeline at the cost of increasing network data rate requirements
- Added sidecar for frames queuing and filtering
- Decoupled inference logic from communication
- Generic gRPC interface for service messages

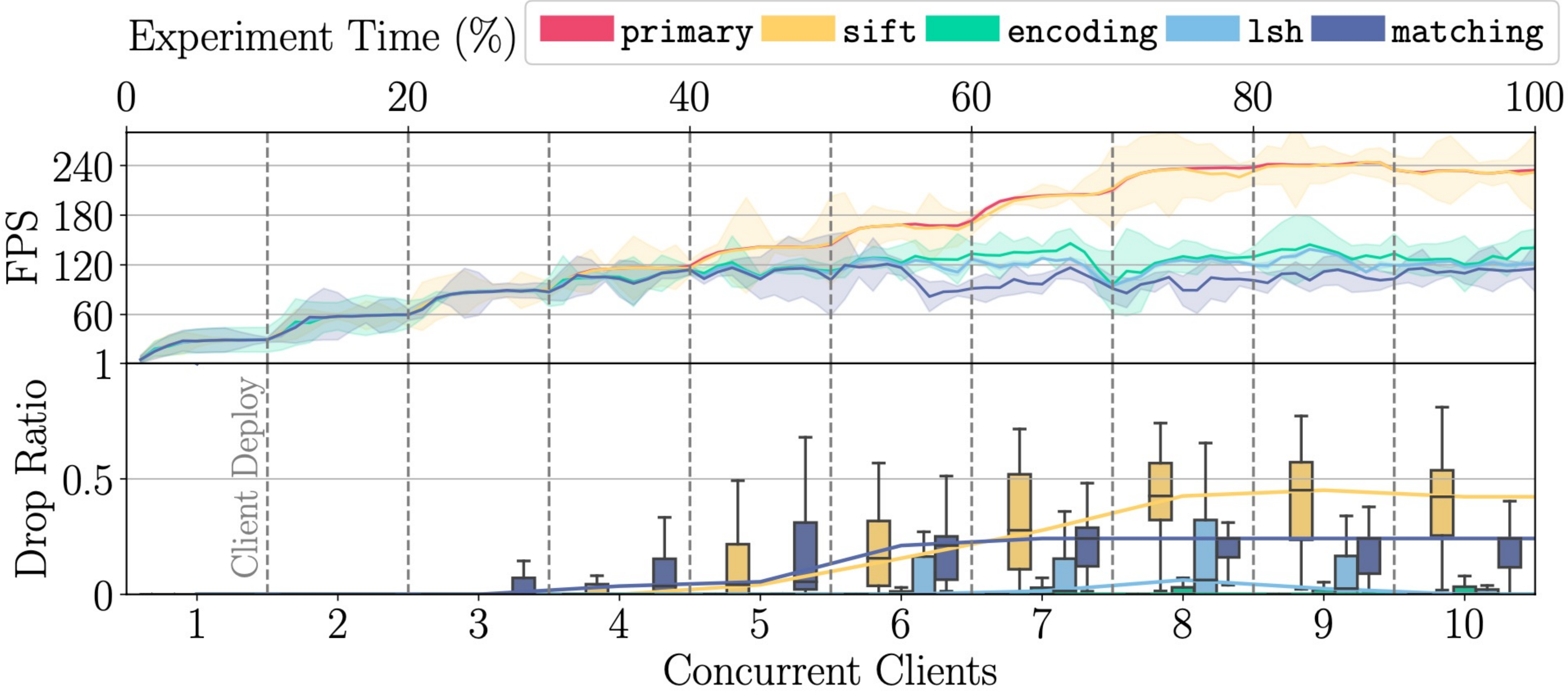
A revamped design



What do we share?

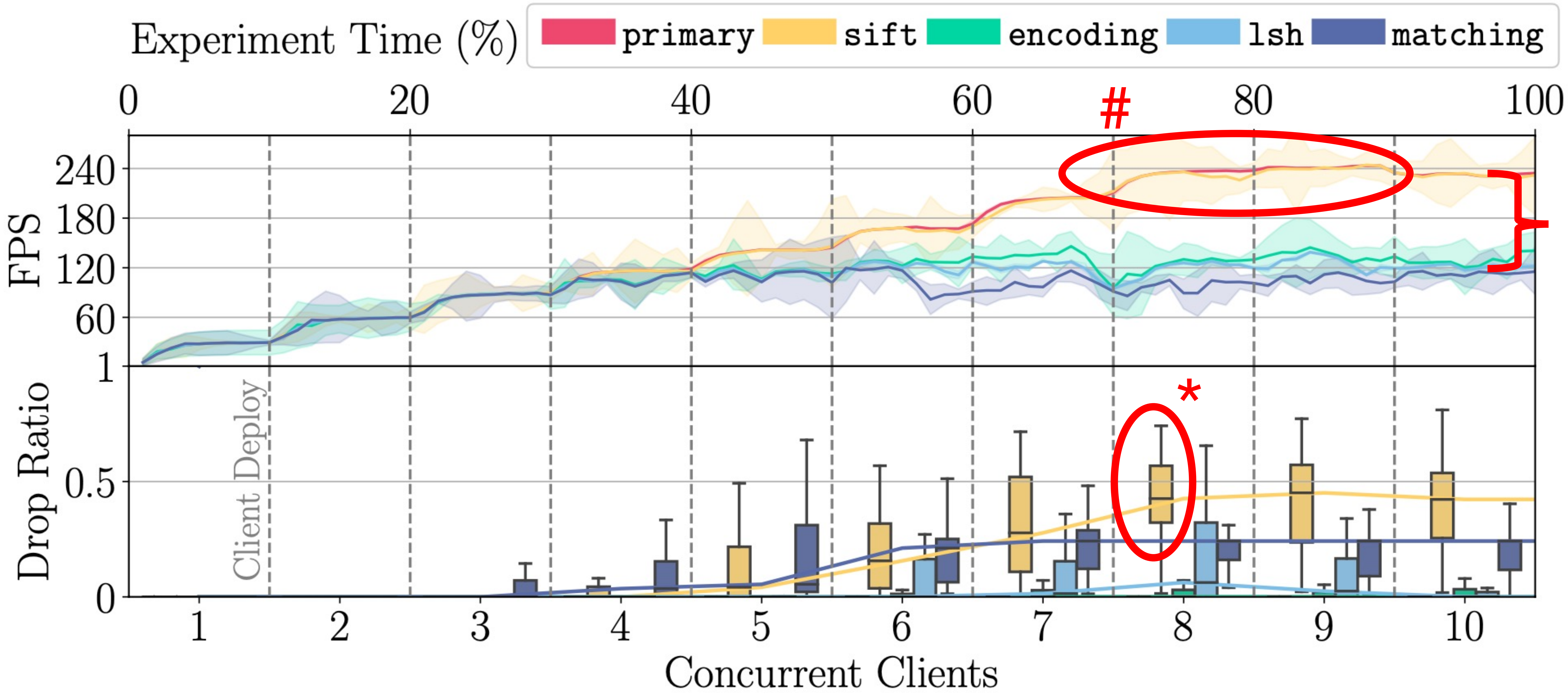
- Ingoing FPS vs Outgoing FPS
  - E.g., to detect bottlenecks in service chain
- Queuing/Dequeuing ratio ->
  - E.g., to understand build up of pressure and max throughput
- Sum of accumulated latency
  - E.g., estimate e2e latency and improvements
- Avg processing time
  - Rough estimate of per batch/per frame processing time
- Drop ratio

Sidecar Metrics



- 2.8x max frame rate improvement with 4 clients
- Max throughput from ~30 FPS to ~120 FPS
- Up to 50% frame dropped after sift with 8-10 clients due to queue threshold

What can the platform learn?



# Service throughput plateauing: Is it a network, app or resource?

} Throughput gaps: Bottlenecks? Backpressure?

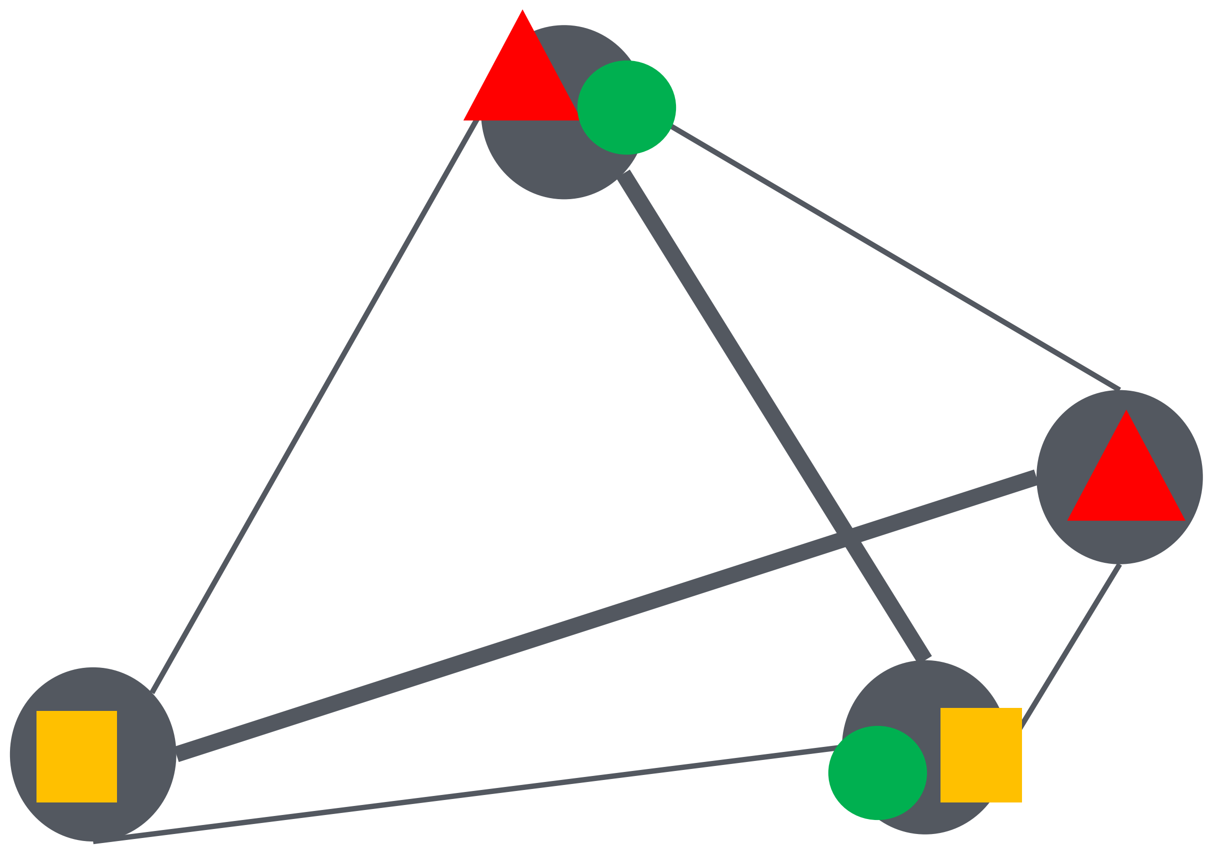
* High drop rate: Missed deadline? Accumulated latency?

Key challenges experienced

- Standardized interface to share metrics
- What to record/how to record
- Sidecar overhead

Additional Considerations

- Network aware orchestration
  - point-to-point data rate/latency
- Application aware network
  - data type, traffic priority



[CoNEXT '23] Bartolomeo, Giovanni, et al. 2023. Characterizing distributed mobile augmented reality applications at the edge.
[ATC '23] Bartolomeo Giovanni, et al. 2023. Oakestra: A Lightweight Hierarchical Orchestration Framework for Edge Computing.
[IEEE IPDPSW] Sanket Chintapalli, et al. 2016. Benchmarking streaming computation engines: Storm, flink and spark streaming.
[MobiSys '18] Luyang Liu, et al. 2018. Cutting the Cord: Designing a High-Quality Untethered VR System with Low Latency Remote Rendering.
[IEEE Commun Mag] Diego González Morín, et al. 2022. Toward the Distributed Implementation of Immersive Augmented Reality Architectures on 5G Networks.
[EdgeSys '22] Simon Bäurle, et al. 2022. ComB: A Flexible, Application-Oriented Benchmark for Edge Computing.
[MMSys '23] Jin Heo, et al. 2023. FleXR: A System Enabling Flexibly Distributed Extended Reality.
[NSDI 22] Jingao Xu, et al. 2022. {SwarmMap}: Scaling up real-time collaborative visual {SLAM} at the edge.
[MM '18] Wenxiao Zhang, Bo Han, and Pan Hui. 2018. Jaguar: Low Latency Mobile Augmented Reality with Flexible Tracking.
[MobiCom '19.] Luyang Liu, et al. 2019. Edge Assisted Real-Time Object Detection for Mobile Augmented Reality.