MECBENCH

A FRAMEWORK FOR BENCHMARKING MULTI-ACCESS EDGE COMPUTING PLATFORMS

Omar Naman[1], Hala Qadi[1], Martin Karsten[1], Samer Alkiswany[1,2]

UNIVERSITY OF WATERLOO

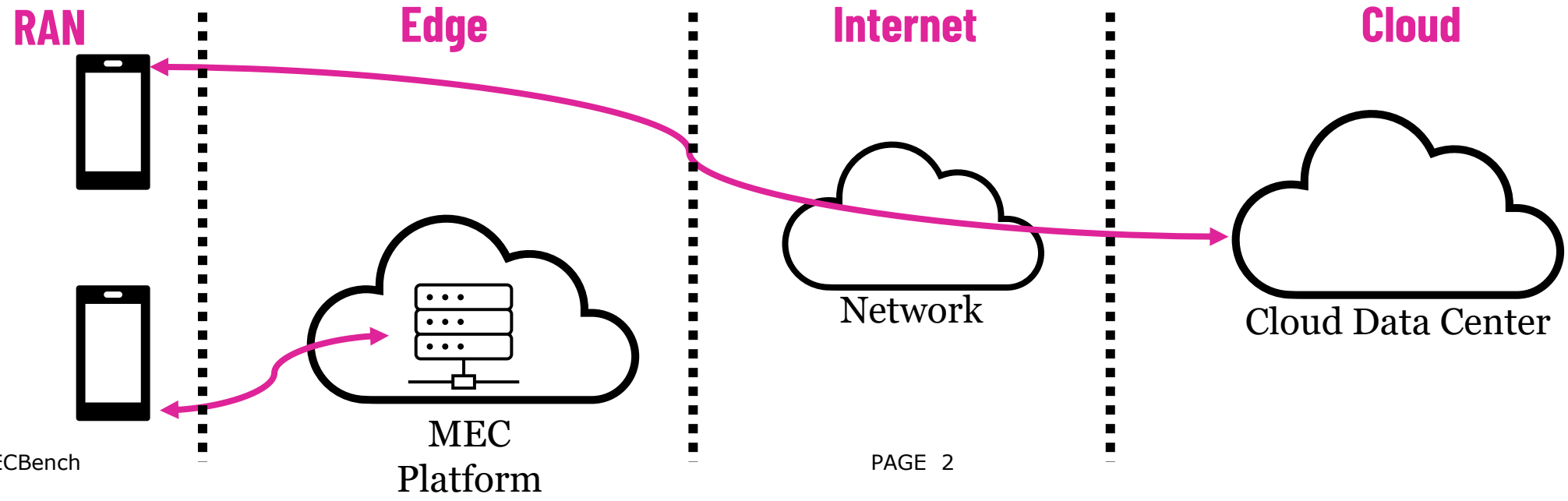
1

Acronis

2

Multi-Access Edge Computing (MEC)

- Proposed alongside fifth-generation networks.
- Masks latencies of distant data centers by existing on the edge of the network.
- Enables more granular deployment of geo-distributed applications.
- Offloads complex application logic from resource-limited devices.



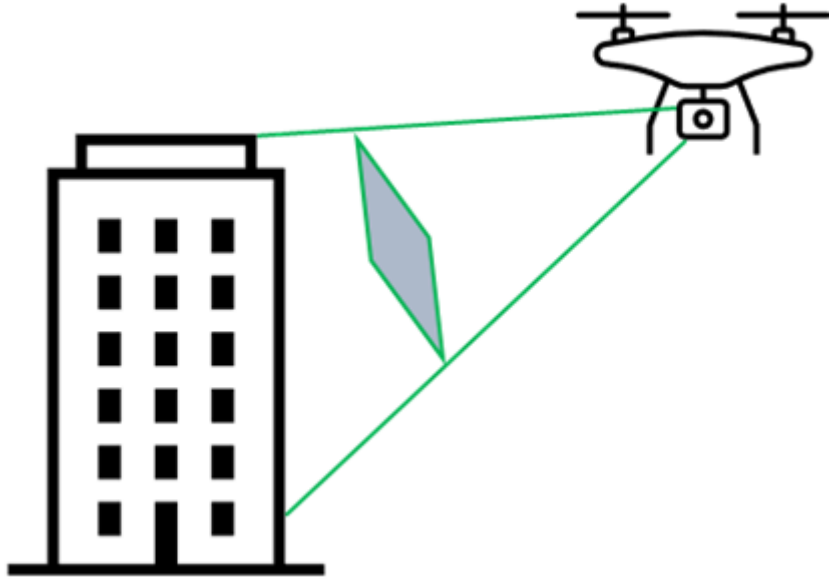
MEC Challenges

- It is ambiguous
  - No standard hardware specification
  - No standard network specification
  - No established applications

- Makes it hard for
  - Providers to design a MEC
  - Developers to design a service
  - Providers to host, bill, and configure services

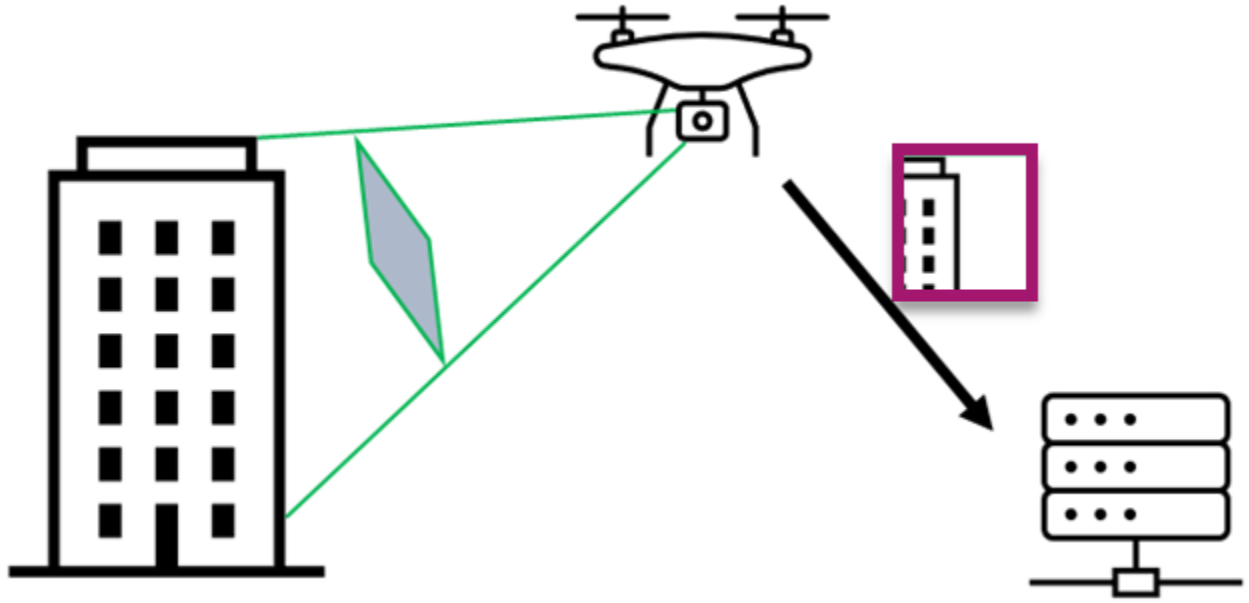
Drone Delivery Application

- Drones to deliver packages
- Use sensor data to avoid obstacles



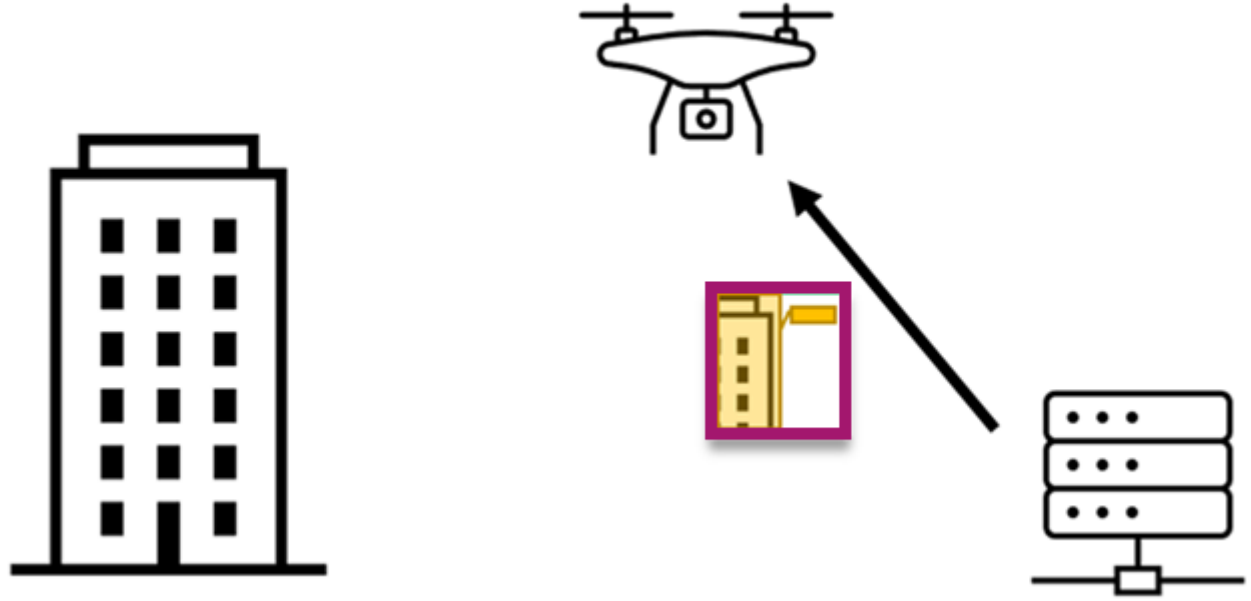
Drone Delivery Application

- Drones to deliver packages
- Use MEC to implement obstacle avoidance
  - Send data to the MEC
  - MEC uses ML-based model for object detection



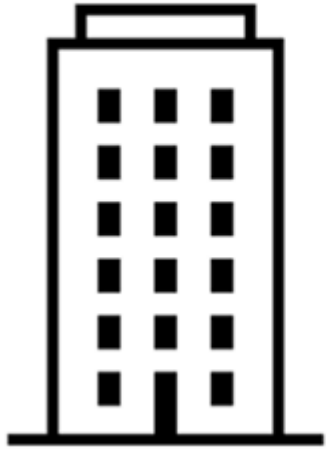
Drone Delivery Application

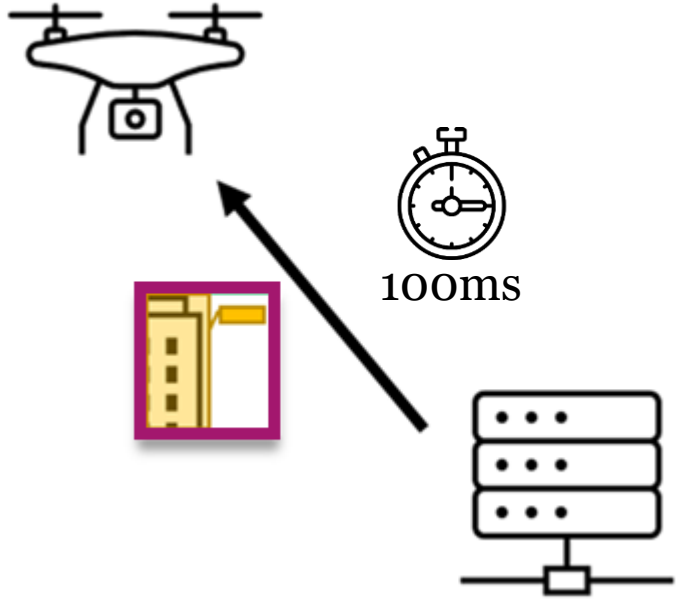
- Drones to deliver packages
- Use MEC to implement obstacle avoidance
  - Send data to the MEC
  - MEC uses ML-based model for object detection
  - MEC informs the drone of obstacles



Drone Delivery Application

Strict latency requirement: 95th percentile of response time should be under *100ms* (Service-Level Agreement)





Drone Delivery Application: Provider Questions

- Can a provider maintain the requested SLA?
- How should the deployment be designed to support the application?

- The answer depends on:
  - Hardware spec.
  - Cost of service.
  - Network spec in the target area.

Drone Delivery Application: Developer Questions

- Will data compression improve response time? (IO/Compute trade-off)
- What is the relation between drone speed and the request rate?

- How can you assess the performance trade-offs?
  - Accuracy from image resolution
  - Cost performance from increasing hardware

MECBench Introduction

- A Benchmarking framework that can help answer these questions
- MECBench assists in:
  - Deploying the application on a MEC platform
  - Mimicking network conditions
  - Generating configurable client workloads
  - If applications are not available, it can mimic application compute and I/O load

MECBench Findings

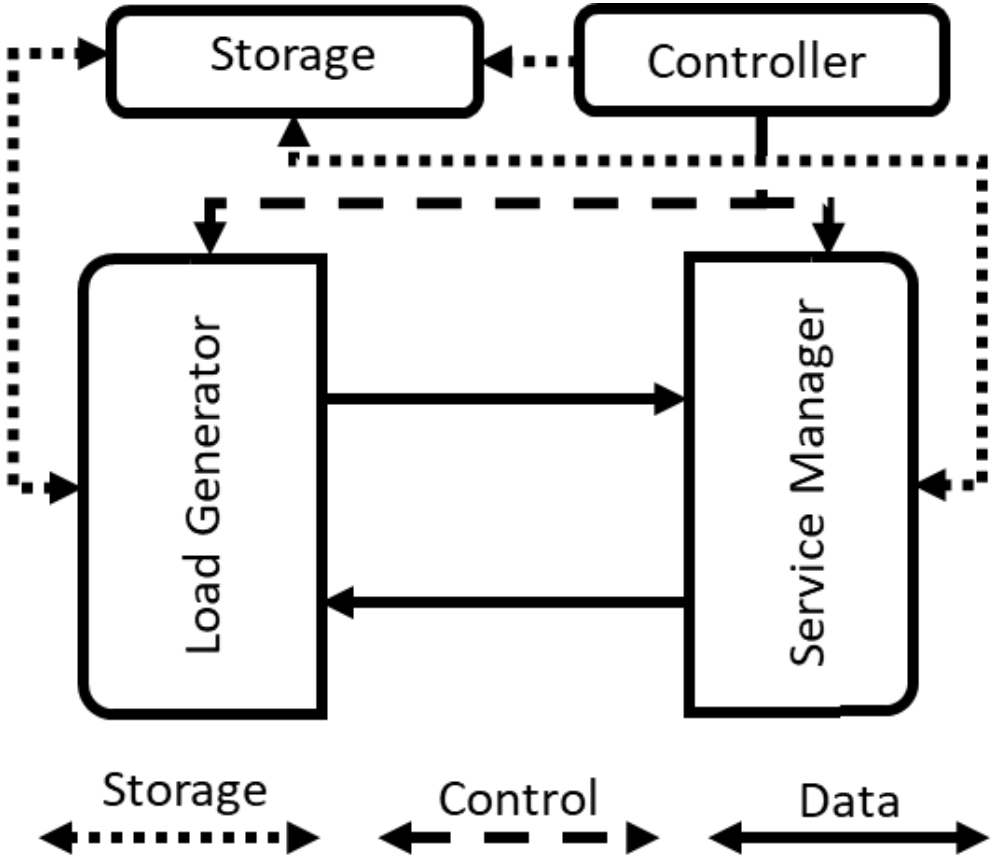
- Experimented with 2 scenarios
  - Object detection for drone navigation
  - Text processing for mobile offloading
- Evaluation on AWS
- Extract surprising findings

Outline

- Introduction
- MECBench design
  - Load generation (LoadGen)
  - Edge service and SUT
  - Control and storage
- Evaluation
- MECBench as a Service
- Conclusion

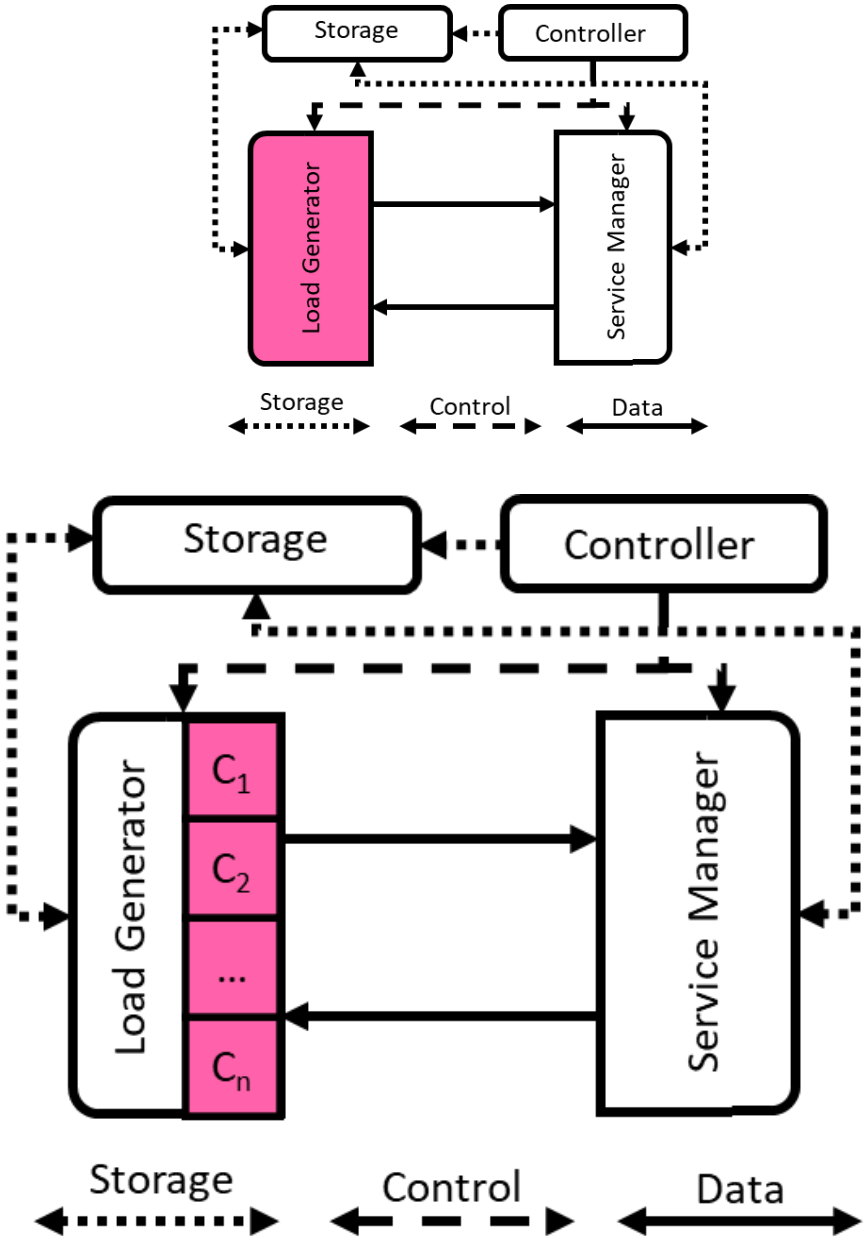
MECBench

- Load generator as client side
- Service manager as server side
- Controller and storage services



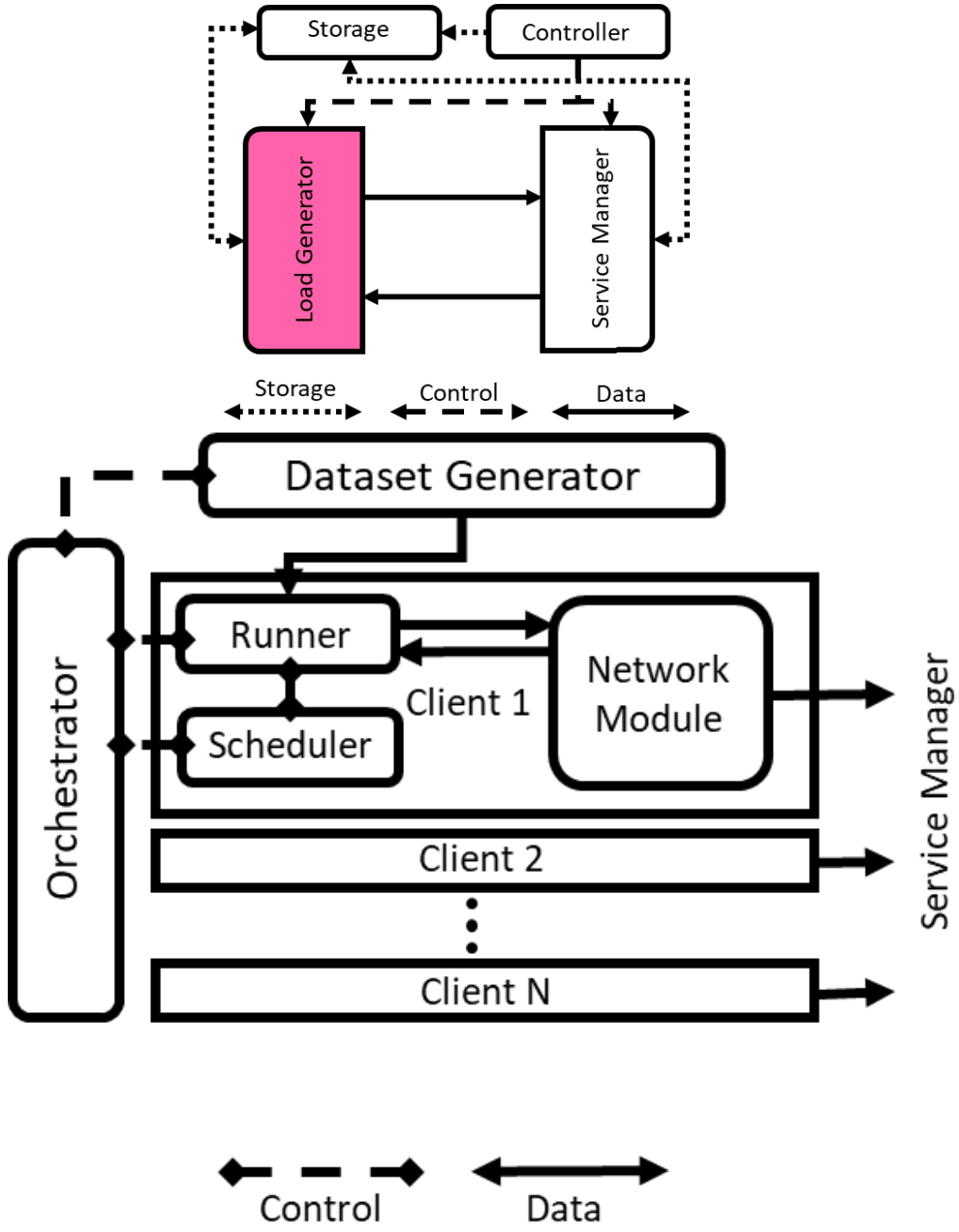
LoadGen Design

Emulates multiple clients running on the load generator



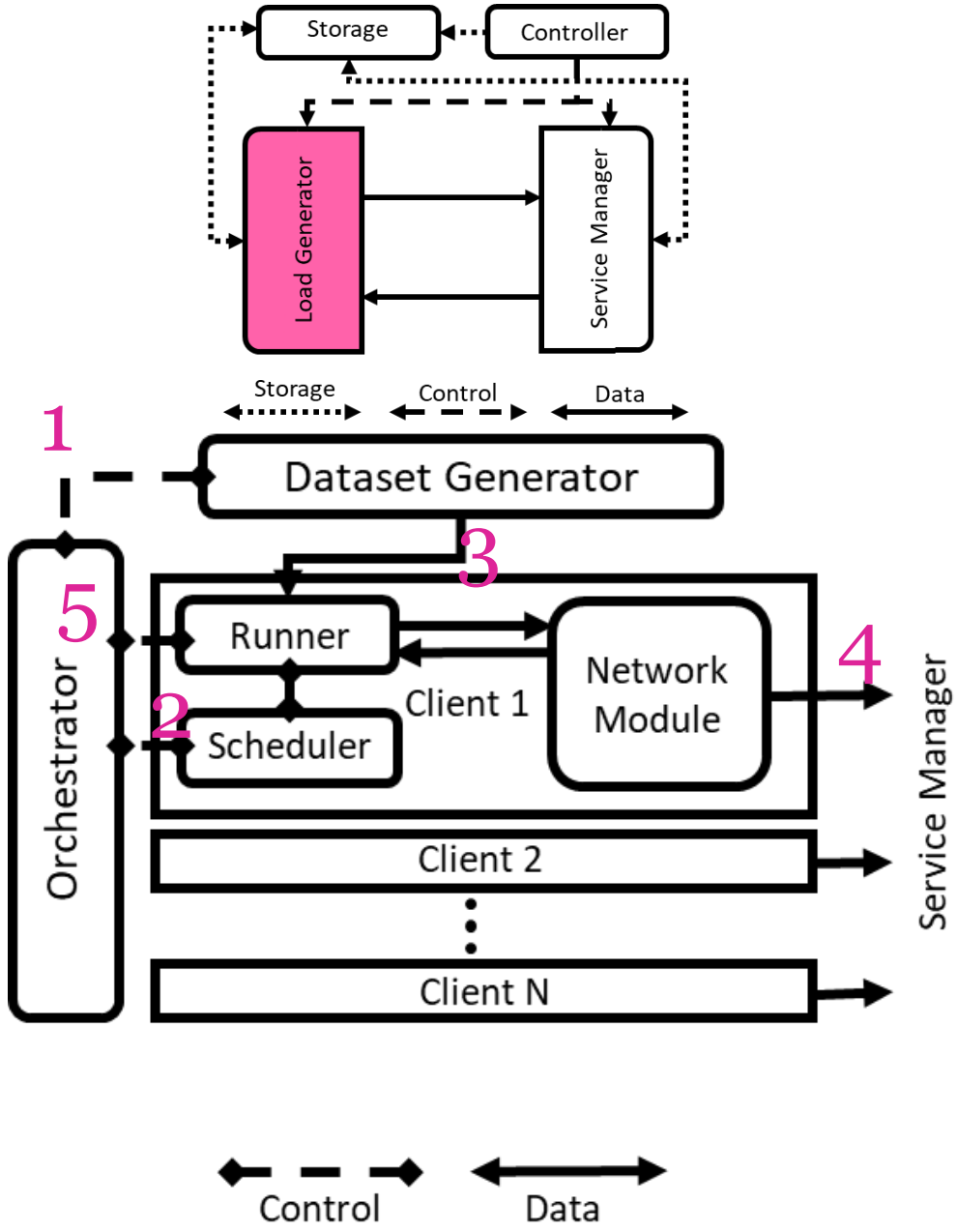
LoadGen Design

- Client load generator
  - Orchestrator
  - Dataset generator
  - Clients and Runners



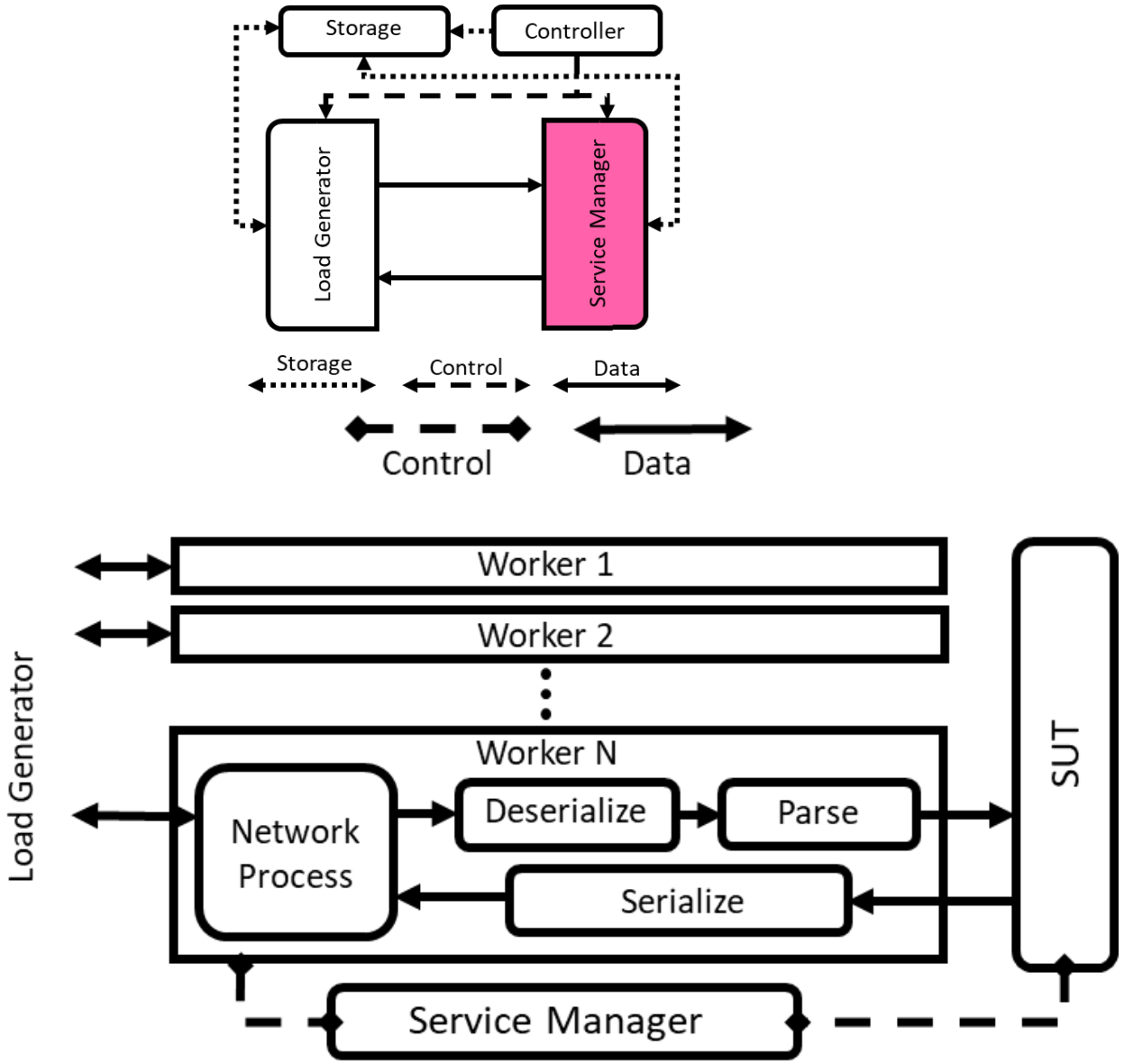
LoadGen Design

1. Load dataset metadata
2. Generate queries
3. Load dataset items
4. Send requests to the service manager and receive responses
5. Report query completion



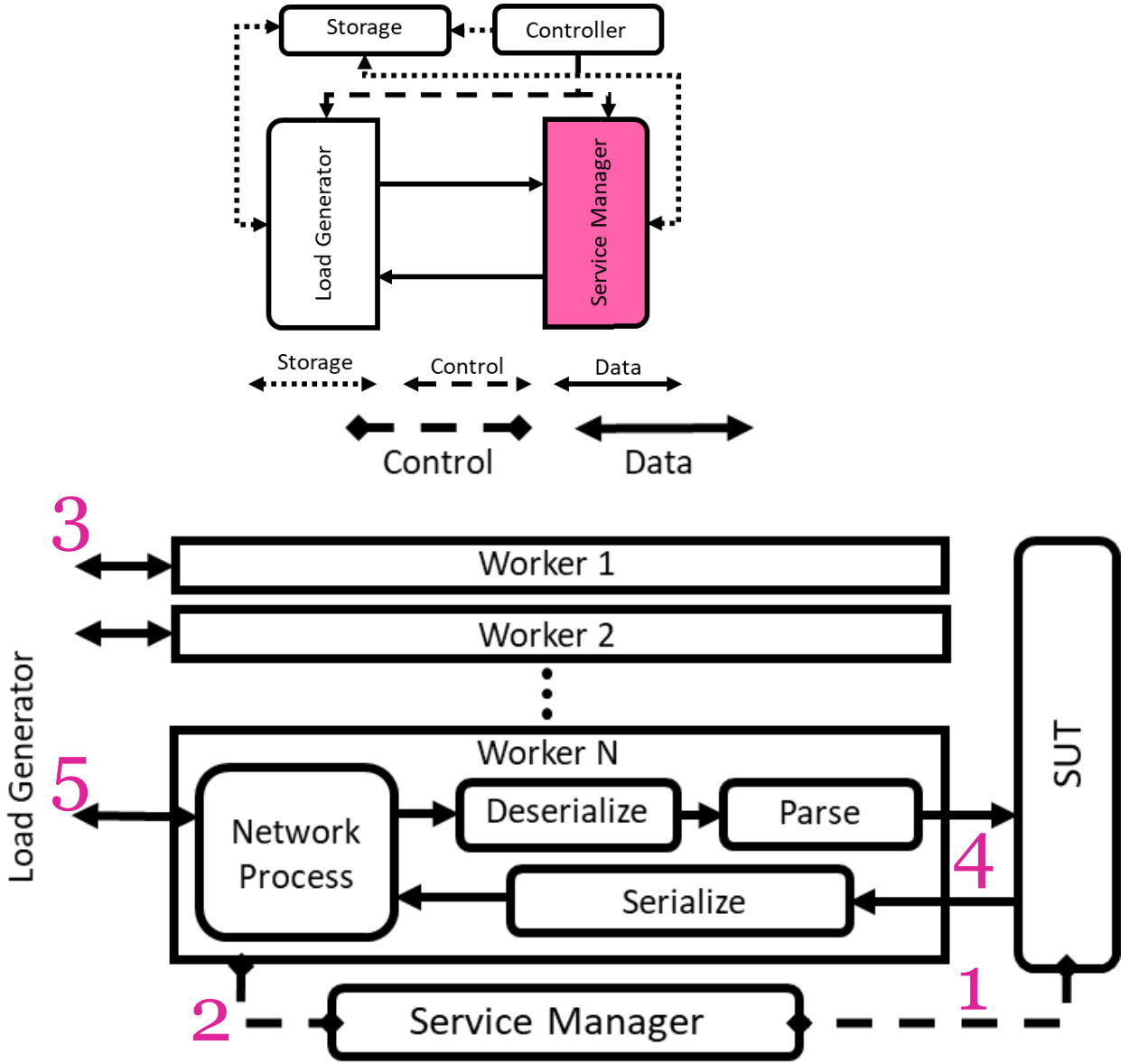
Service Manager Design

- Service manager as server side
  - Service Under Test (SUT)
  - Request parsing and serialization



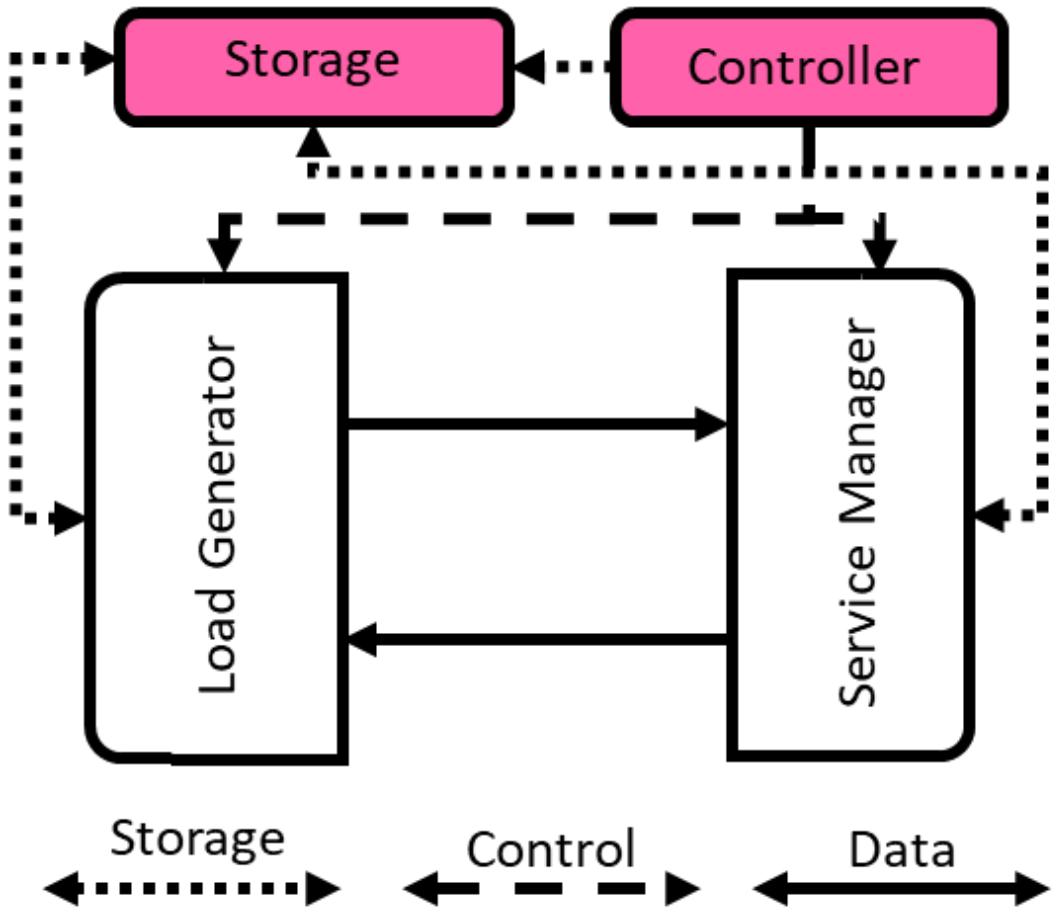
Service Manager Design

1. Load and initialize SUT
2. Spin up workers
3. Receive requests from client
4. Pass queries to SUT and receive results
5. Send results back to clients

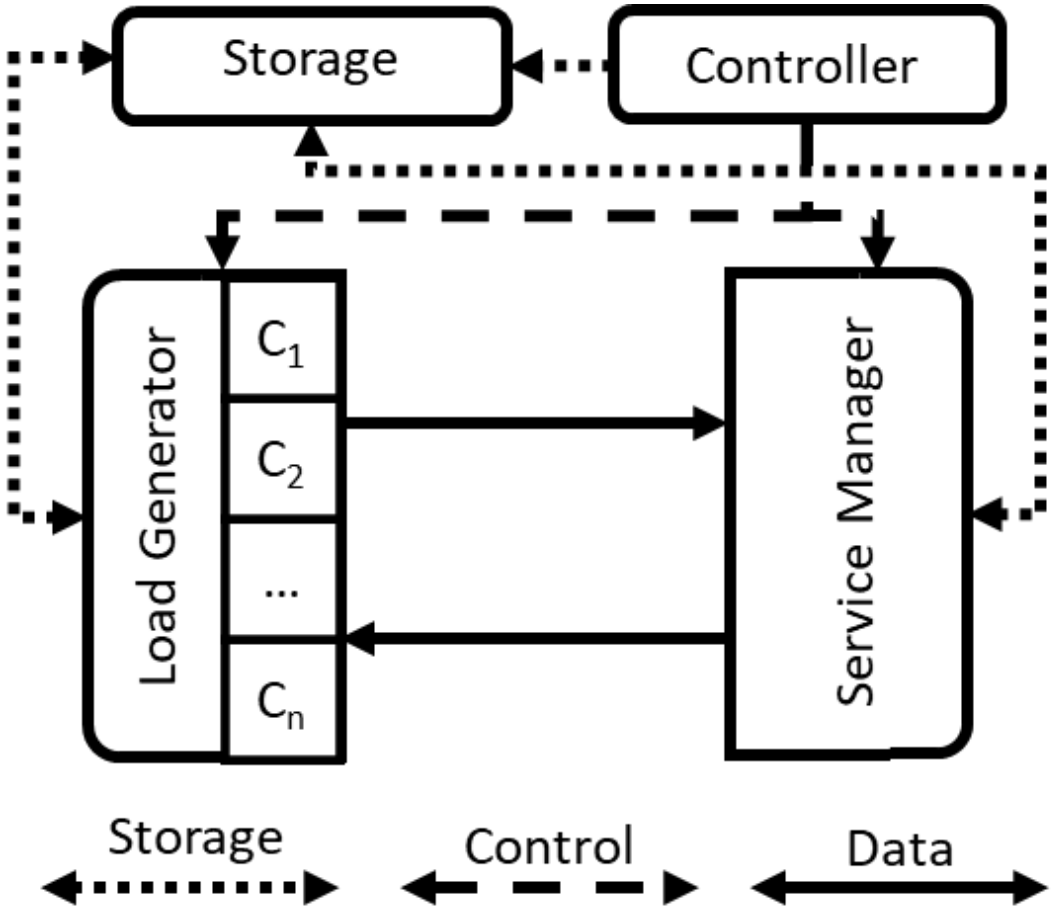


MECBench Services

- Storage service for data, results, and configurations :
  - MECBench Storage
  - Blob Storage
- Controller
  - Orchestrates MECBench's components
  - Exposes APIs for automation

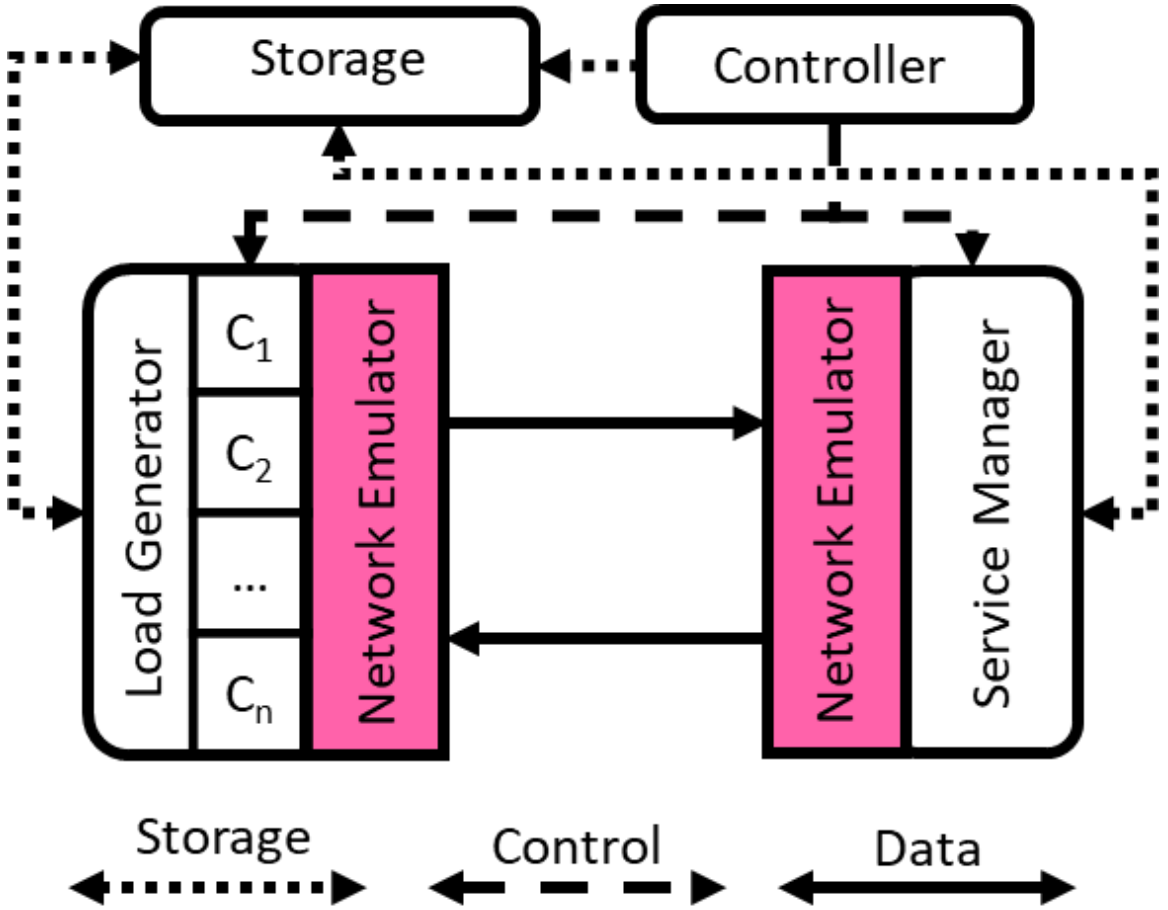


MECBench Network Communication



MECBench Network Communication

- Network emulator based on TC
- gRPC for data communication
- Stateless HTTP for control and storage communication



MECBench Configurability

- Client-side:
  - Number of clients
  - Number of queries per request
  - Workload scenarios represent different real-life use-cases
- Server-side
  - Number of threads assigned to SUT
  - Maximum concurrent workers
- Network emulation
  - Delay and jitter
  - Packet loss
  - Transfer rate
  - Packet reordering

MECBench Extensibility

- MECBench can be extended to:
  - Add new client workloads
  - Add new SUTs

LoadGen Extensibility

Dataset

- loadDataset
- loadQueryData
- getQueryData
- postprocess
- getNumberOfQueries

Runner

- runQuery
- Call
- Clone
- Init
- Constructor

Service Manager Extensibility

SUT

- Load
- parseQuery
- processQuery
- serializeResponse

Outline

- Introduction
- MECBench Design
  - Load generation (LoadGen)
  - Edge Service and SUT
  - Control and Storage
- Evaluation
- MECBench as a Service
- Conclusion

Evaluation Scenarios

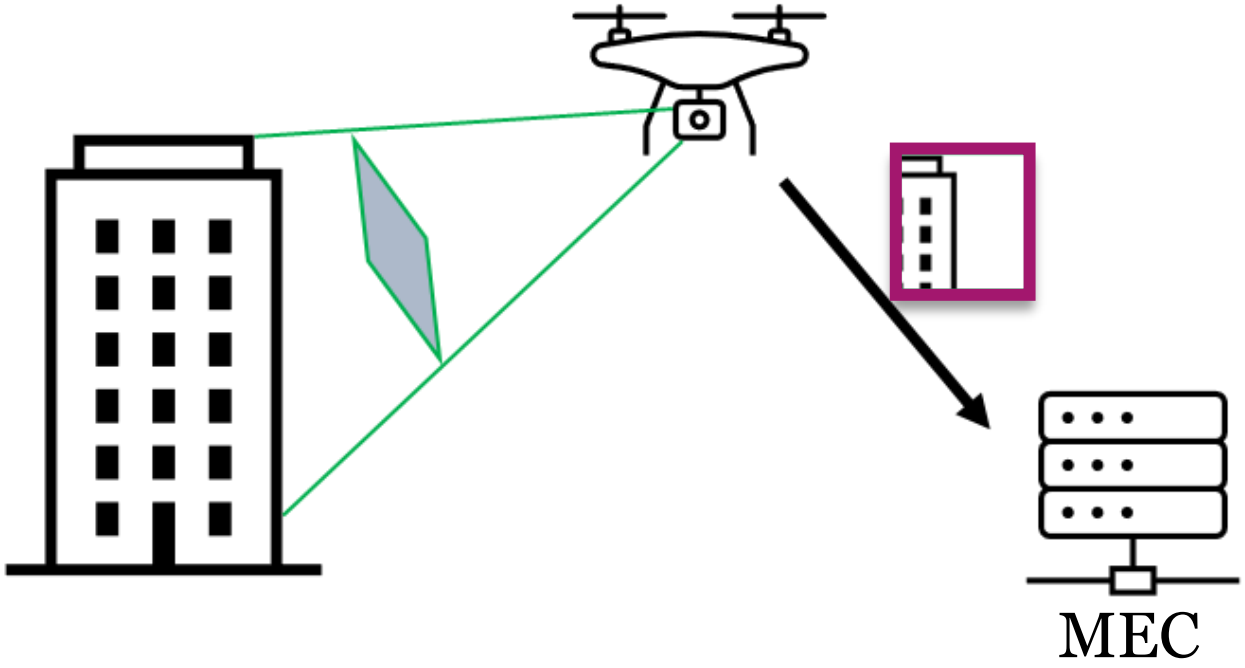
- Drone navigation
- Text-based named entity recognition

Evaluation Scenarios

- Drone navigation
- Text-based named entity recognition

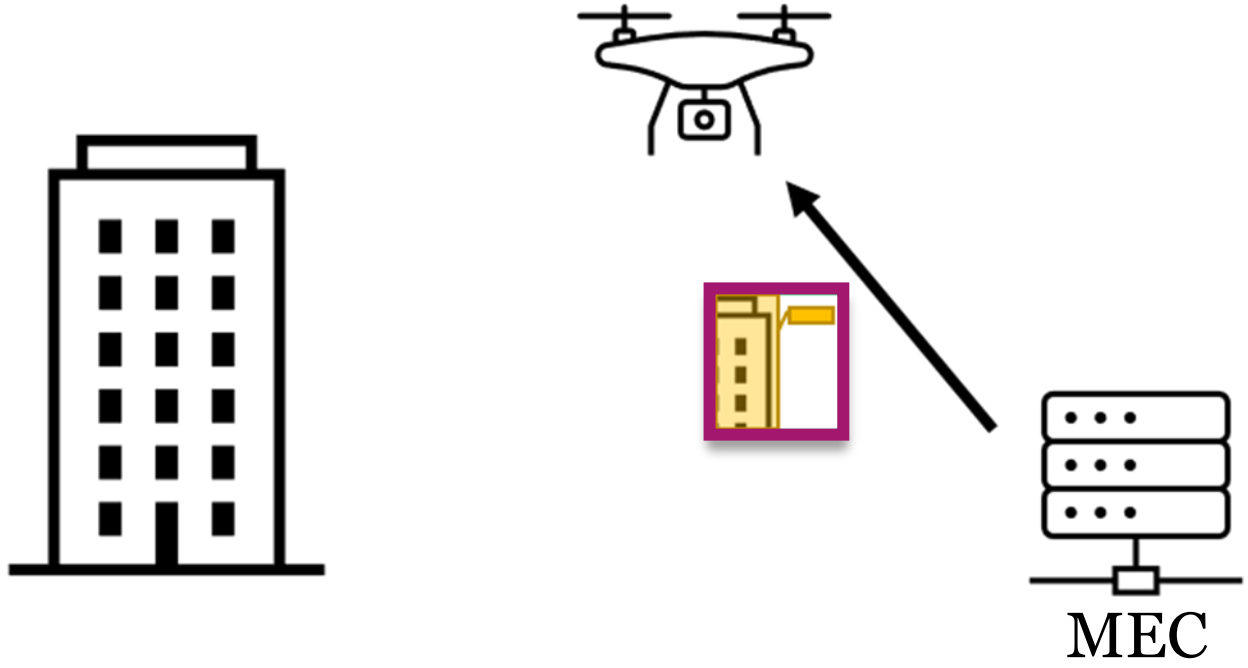
Drone Application (Refresher)

- Drone sends images to MEC for object detection



Drone Application (Refresher)

- Drone sends images to MEC for object detection
- MEC uses ML-based object detection
- MEC sends results to the drone in under 100ms



Drone Application (Questions)

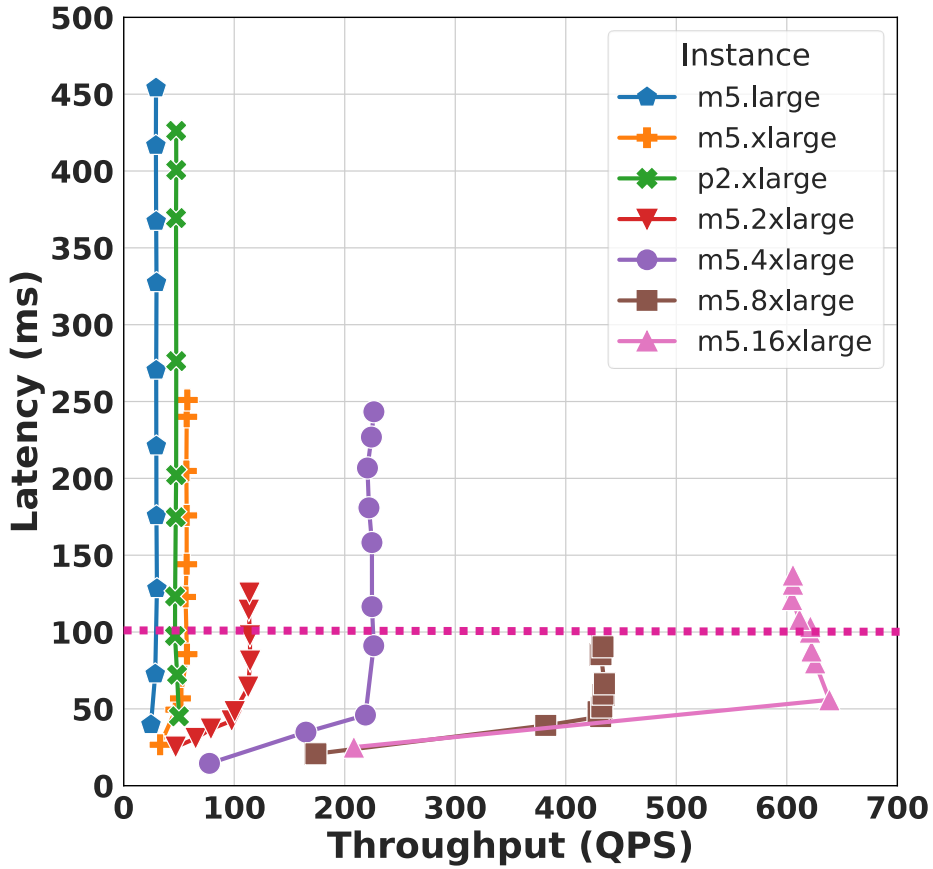
- What is the Cost/Performance trade-off of AWS instances?
- Does the application scale to use multiple cores?
- What is the impact of image resolution on accuracy and performance?
- What is the impact of the network's performance on the number of drones?
- What is the impact of data compression on application performance?
- At what speed should the drone fly under different network technologies?

Drone Application (Questions)

- Provider Questions
  - What is the Cost/Performance trade-off of AWS instances?
  - How many drones can be supported using different network technologies?
- Developer Questions:
  - What is the impact of data compression on application performance?

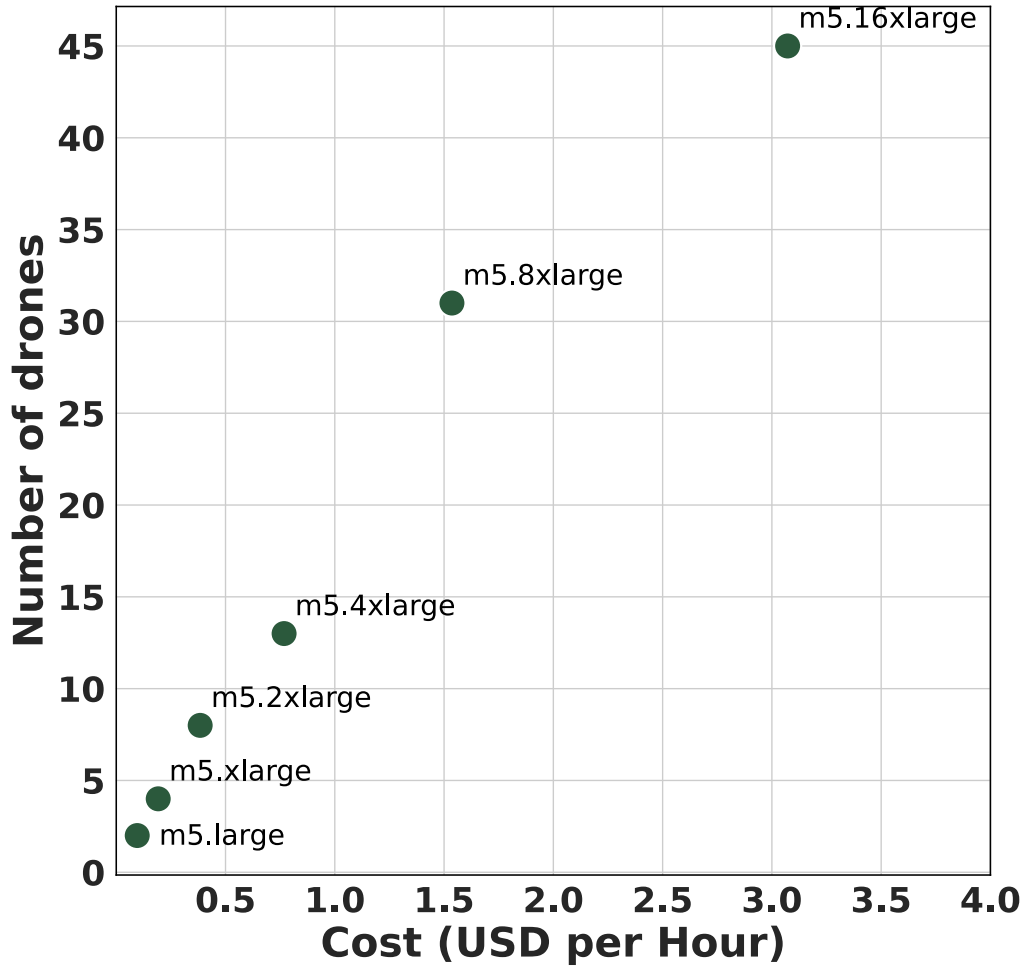
What Is the Cost/Performance Trade-off of AWS Instances?

- Closed-loop clients
- SUT running SSD-Mobilenet
- Multiple types of AWS instances
- End-to-end latency threshold: 100ms



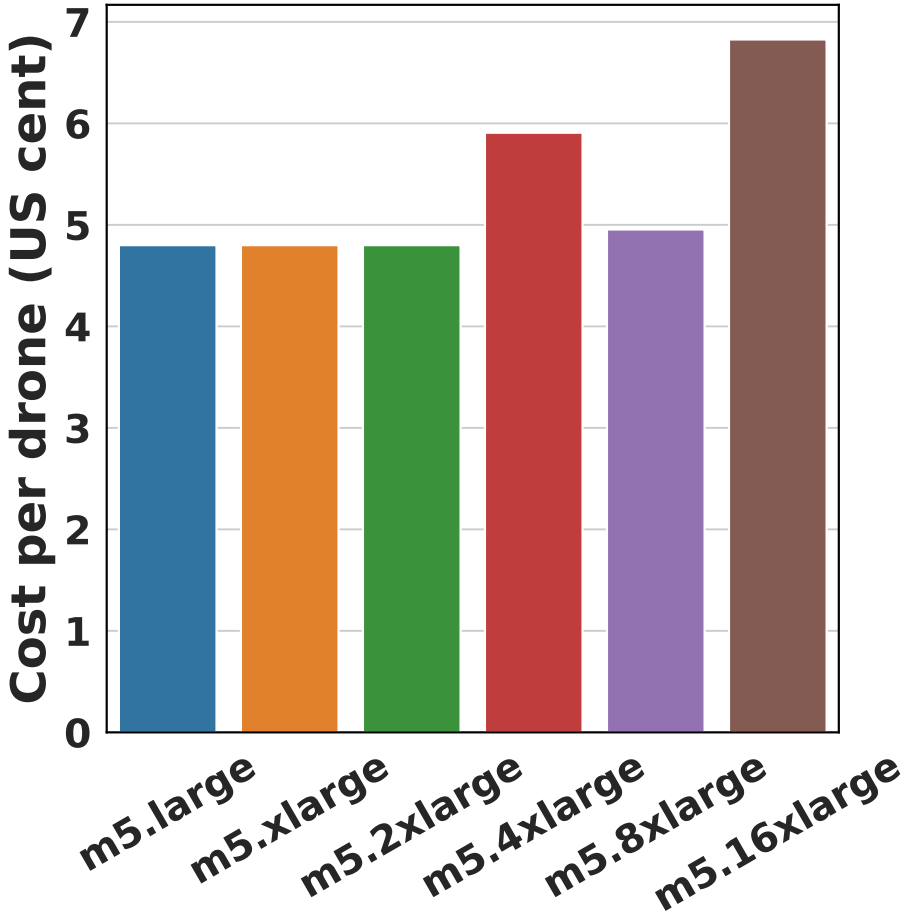
What Is the Cost/Performance Trade-off of AWS Instances?

- Closed-loop clients
- SUT running SSD-Mobilenet
- Multiple types of AWS instances
- End-to-end latency threshold: 100ms



What Is the Cost/Performance Trade-off of AWS Instances?

- Closed-loop clients
- SUT running SSD-Mobilenet
- Multiple types of AWS instances
- End-to-end latency threshold: 100ms

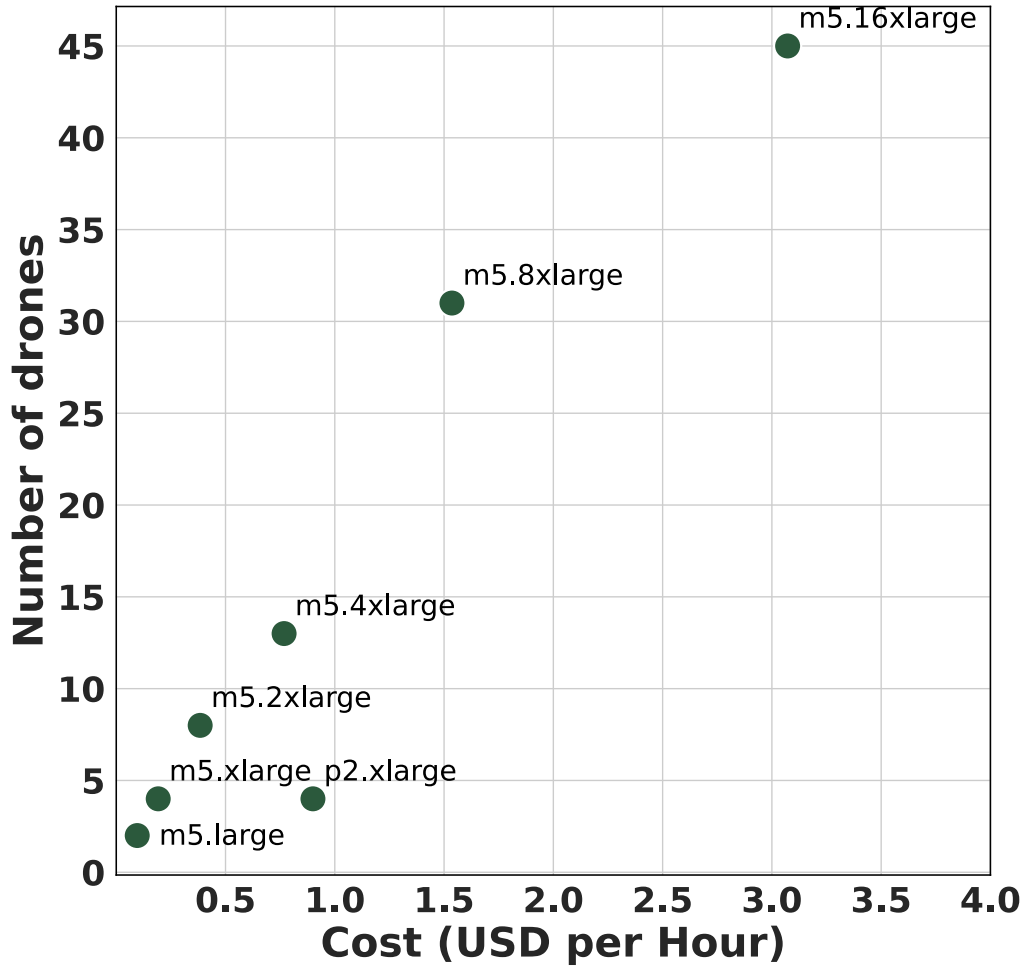


m5.4xlarge and m5.16xlarge are more expensive per drone

Providers are better off scaling other instances horizontally

What Is the Cost/Performance Trade-off of AWS Instances?

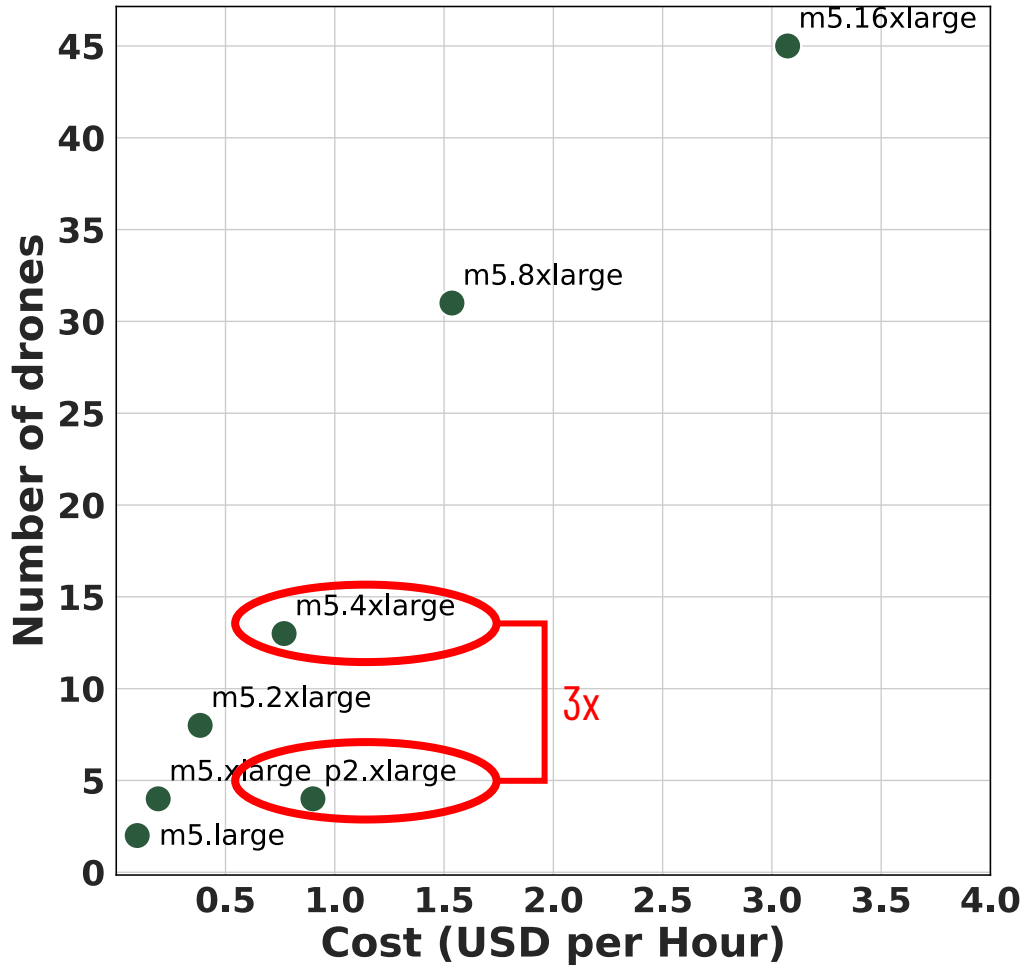
- Closed-loop clients
- SUT running SSD-Mobilenet
- Multiple types of AWS instances
- End-to-end latency threshold: 100ms
- What about GPU?



What Is the Cost/Performance Trade-off of AWS Instances?

- Closed-loop clients
- SUT running SSD-Mobilenet
- Multiple types of AWS instances
- End-to-end latency threshold: 100ms
- What about GPU?

m5.4xlarge costs the same as p2.xlarge but performs 3x better



What is the impact of the network's performance on the number of drones?

Network specifications used in our evaluation

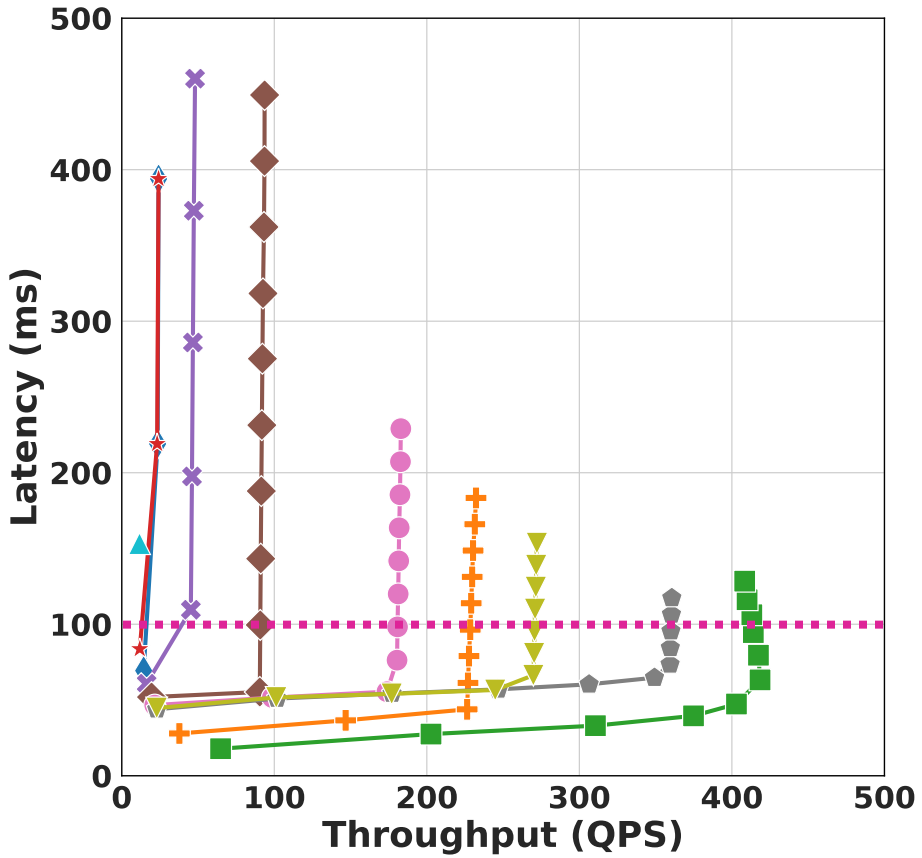
Network Condition	RTT (ms)	Download (Mbps)	Upload (Mbps)
5G	1	10,000	1,000
4G-LTE+	10	1,000	500
4G-LTE	10	100	50
WiMAX	30	128	64

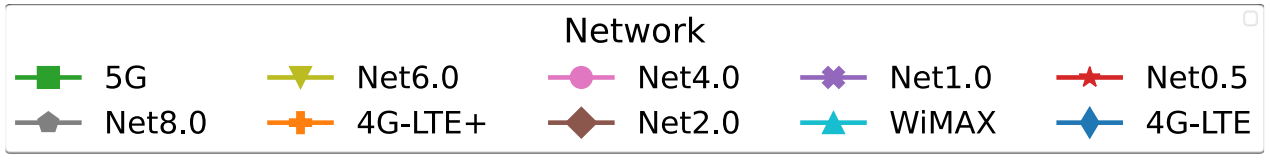
Synthetic network specifications

Network Condition	RTT (ms)	Download (Mbps)	Upload (Mbps)
Net8.0	25	8,000	800
Net6.0	25	6,000	600
Net4.0	25	4,000	400
Net2.0	25	2,000	200
Net1.0	25	1,000	100
Net0.5	25	500	50

What is the impact of the network's performance on the number of drones?

- Closed-loop clients
- m5.8xlarge instance
- Variable network specs
- RAW COCO dataset: 264KB per item
- End-to-end latency threshold: 100ms



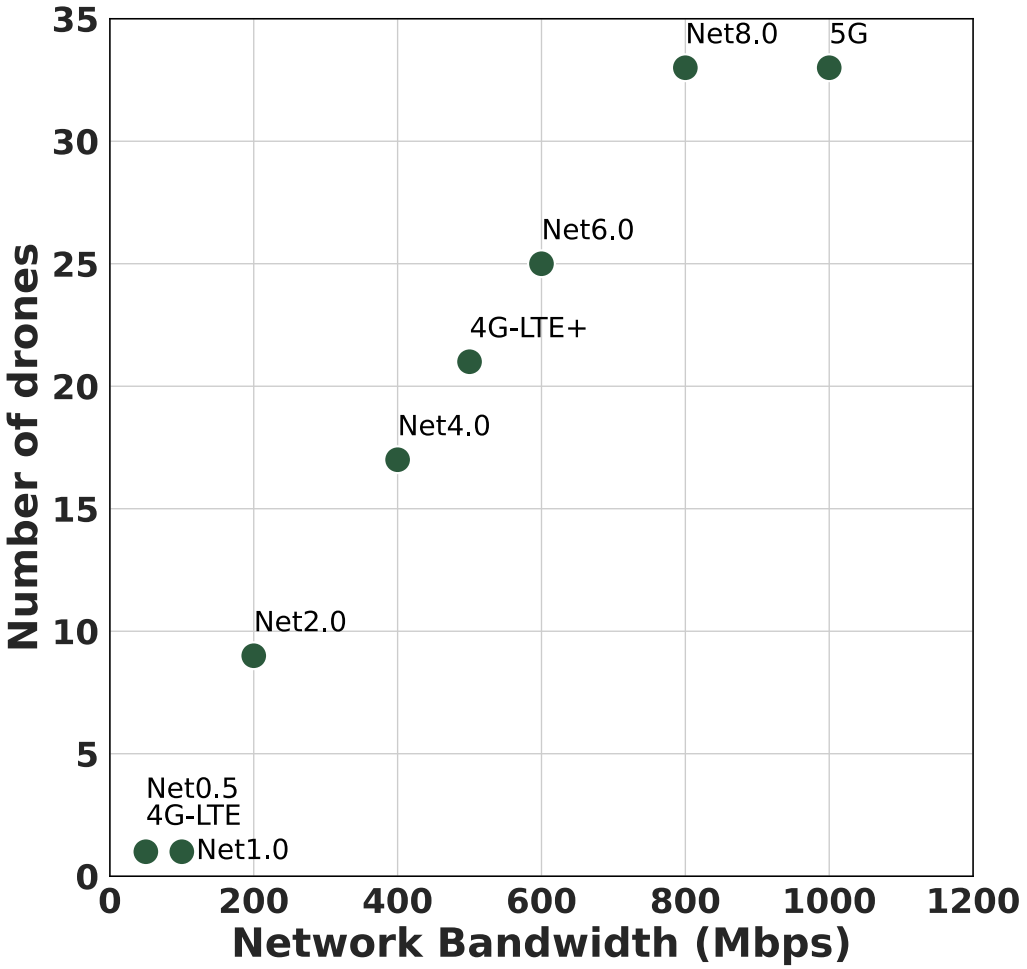


What is the impact of the network's performance on the number of drones?

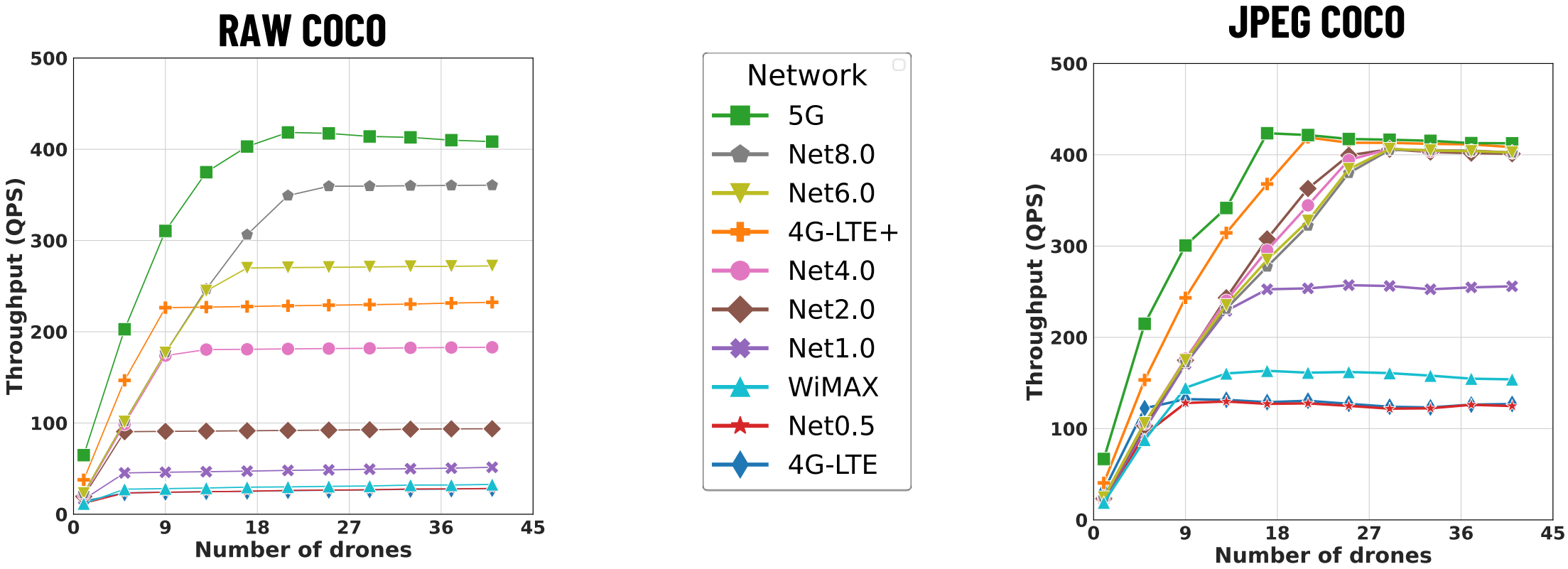
- Closed-loop clients
- m5.8xlarge instance
- Variable network specs
- RAW COCO dataset: 264KB per item
- End-to-end latency threshold: 100ms

At least 4G-LTE+ is needed to support over 5 drones

The 5G setup is limited by the SUT Server



What is the impact of data compression on application performance?



Compression helps low bandwidth networks

Compression does not make a difference in the 5G setup due to compute overhead

Outline

- Introduction
- MECBench Design
  - Load generation (LoadGen)
  - Edge Service and SUT
  - Control and Storage
- Evaluation
- MECBench as a Service
- Conclusion

MECBench as a Service

- All components are containerized and deployable through a container orchestration engine
- Provides a web interface for accessibility
- Includes 7 models and 4 datasets

Outline

- Introduction
- MECBench Design
  - Load generation (LoadGen)
  - Edge Service and SUT
  - Control and Storage
- Evaluation
- MECBench as a Service
- Conclusion

Conclusion

- MECBench
  - MEC Benchmarking framework
  - Extensible experiments and reproducible results
- Evaluated
  - Object detection service
  - Text named-entity extraction

- Can help answer many questions regarding edge applications
  - Price performance trade-offs
  - The impact of network conditions
  - Model Accuracy

Thank you! ☺