SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Towards Cloud-Edge Collaborative Online Video Analytics with Fine-Grained Serverless Pipelines

Miao Zhang, Fangxin Wang, Yifei Zhu, Jiangchuan Liu, Zhi Wang

OUTLINE

- An Introduction to Video Analytics
- Measurement and Motivation
- CEVAS: System Design and Implementation
- Evaluation

BCKGROUND

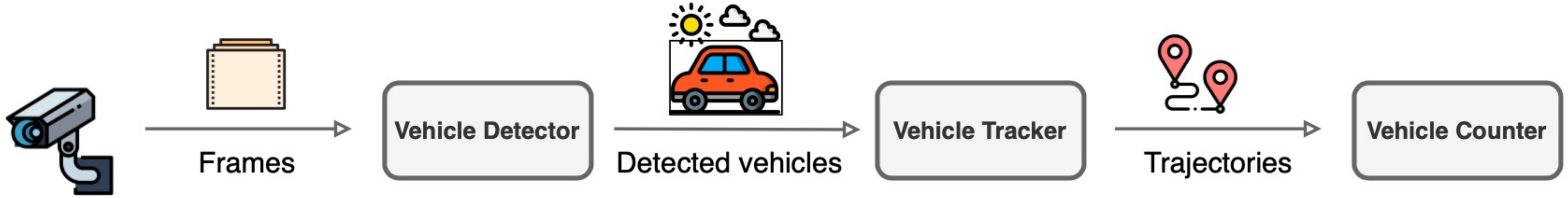
Video Analytics

Ever-growing camera deployment

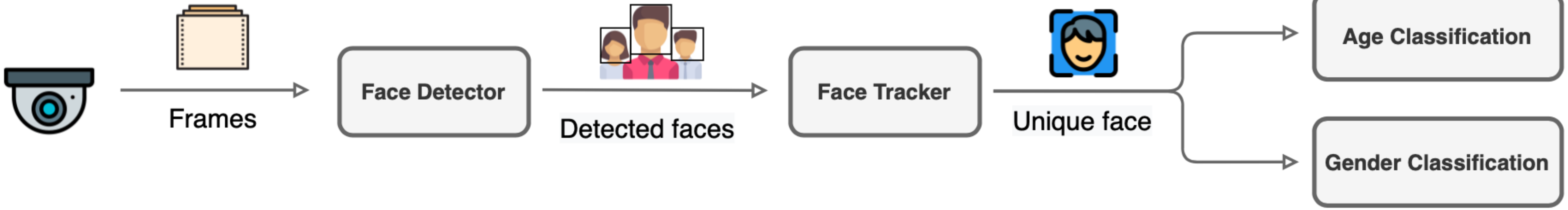
Advances in computer vision algorithms

} Video Analytics

Vehicle pipeline



Face pipeline



EXISTING EFFORTS

☐ From Retrospective to Live

Low-latency and costs queries on large datasets, e.g., Focus (OSDI’18).

Scalable real-time queries on live video streams, e.g., Chameleon (SIGCOMM’18).

☐ From Cloud to Cloud-Edge

Model compression and approximation, e.g., VideoEdge (SEC’18).

DNN model splitting, e.g., Split-brain (HotEdgeVideo’19).

Frame compression, e.g., CloudSeg (HotCloud’19).

Frame filtering, e.g., Reducto (SIGCOMM’20).

Coarse-grained and manual resource management can hardly adapt to fine-grained dynamics.

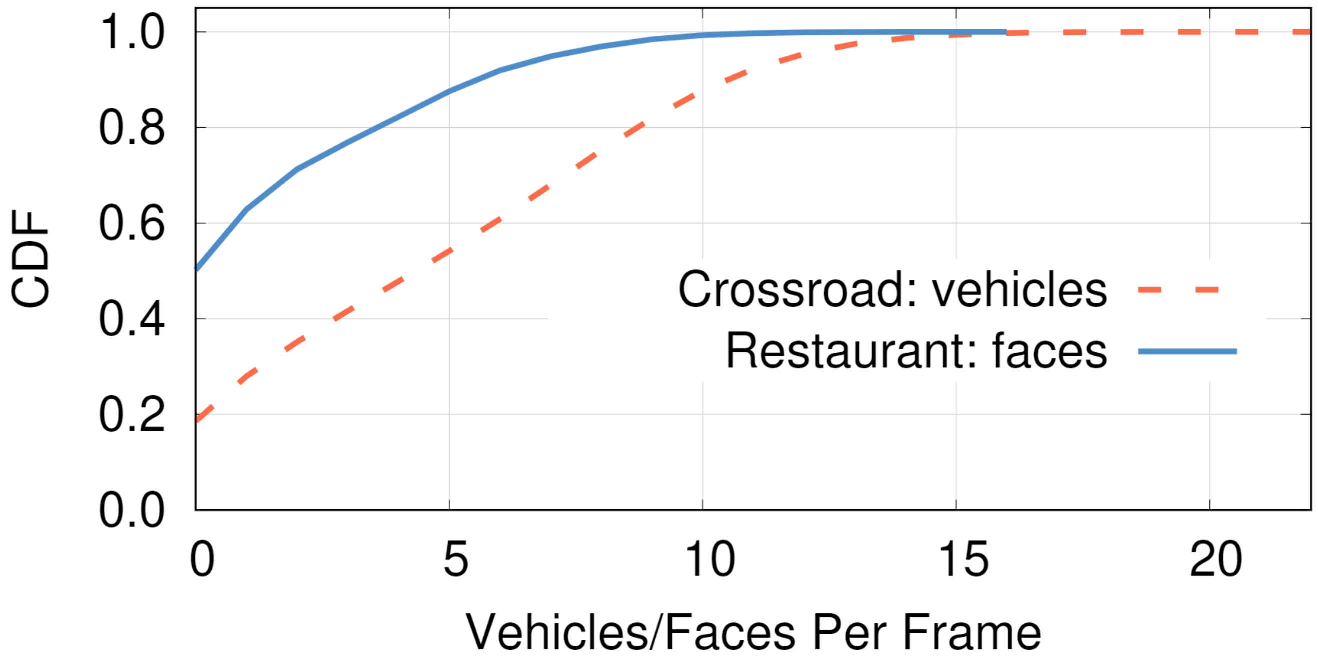
Monolithic deployment architectures hamper flexibility and scalability.

OUTLINE

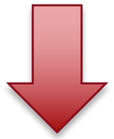
- An Introduction to Video Analytics
- Measurement and Motivation
- CEVAS: System Design and Implementation
- Evaluation

MEASUREMENT

☐ Video Analytics Statistics on Real-World Cameras



A considerable part of frames carry useless information.

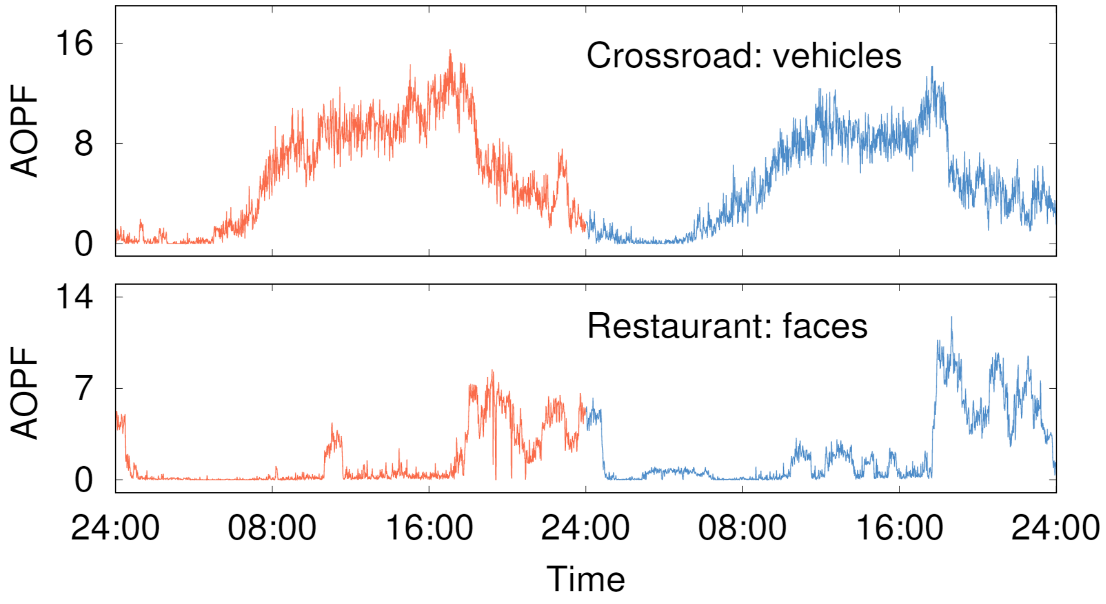


Developing content-aware resource schedulers.

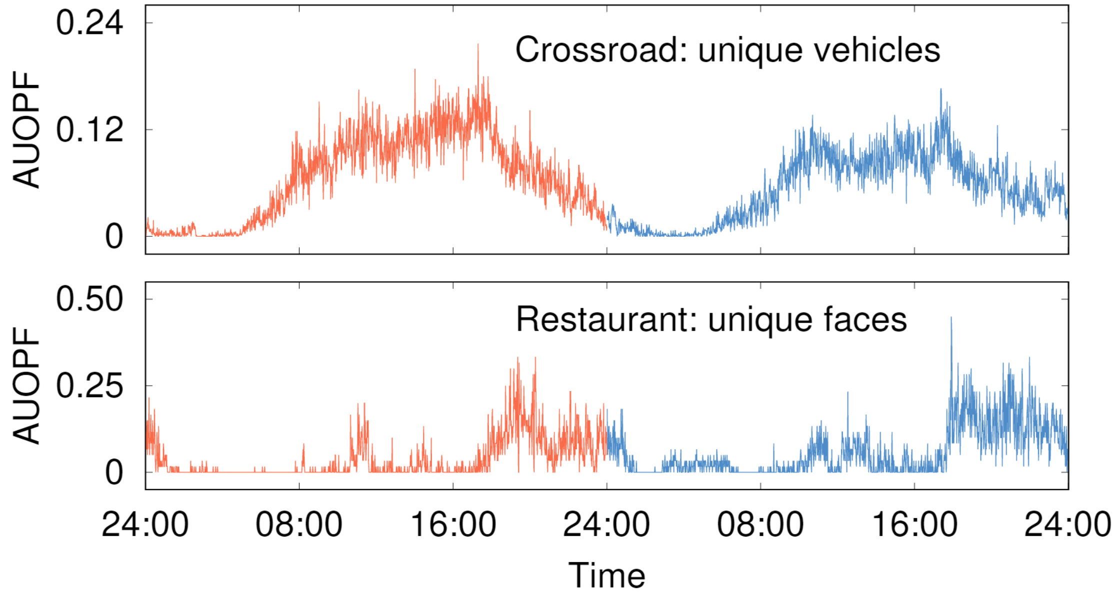
[1] Crossroad camera URL: https://www.youtube.com/watch?v=1EiC9bvVGnk
[2] Restaurant camera URL: https://www.youtube.com/watch?v=sbZNL98Z0GE

MEASUREMENT

☐ Video Analytics Statistics on Real-World Cameras



Average Objects Per Frame (AOPF)



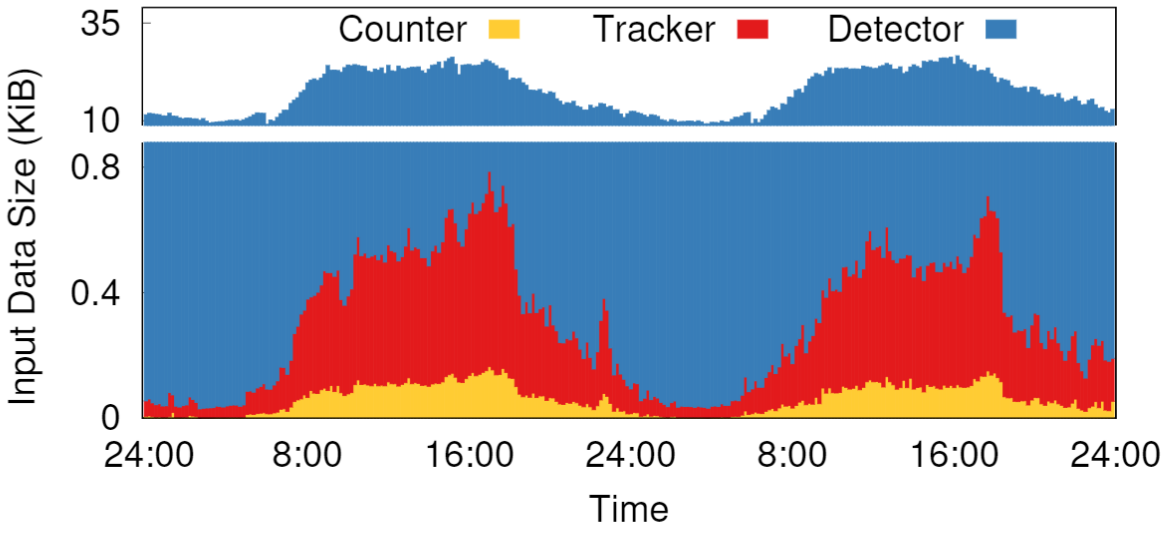
Average Unique Objects Per Frame (AUOPF)

Fine-grained dynamics can hardly be captured by one-time offline or coarse-grained online profiling.

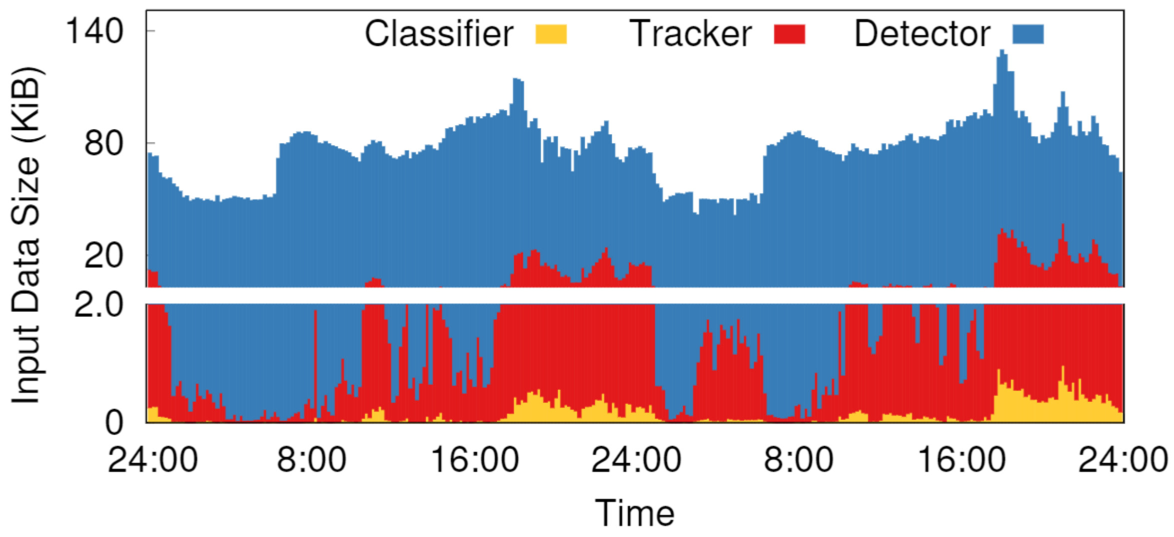
Time-series dependency information may be useful.

MEASUREMENT

☐ Video Analytics Statistics on Real-World Cameras



(a) Crossroad: vehicle pipeline



(b) Restaurant: face pipeline

Cloud-edge collaborative schemes have great potential in reducing network resource consumption.

Video content dynamics should be taken into consideration.

MOTIVATION

☐ Opportunities Brought by Serverless Computing

Function as a Service (FaaS) offerings

AWS Lambda　Google Cloud Functions　Apache OpenWhisk

Execute functions in Content Delivery Network (CDN)

Lambda@Edge

Execute functions in IoT devices

AWS IoT Greengrass Core

Fine-grained resource unit → Addressing fine-grained video content dynamics

Workload-driven resource autoscaling → Avoiding unnecessary resource provisioning

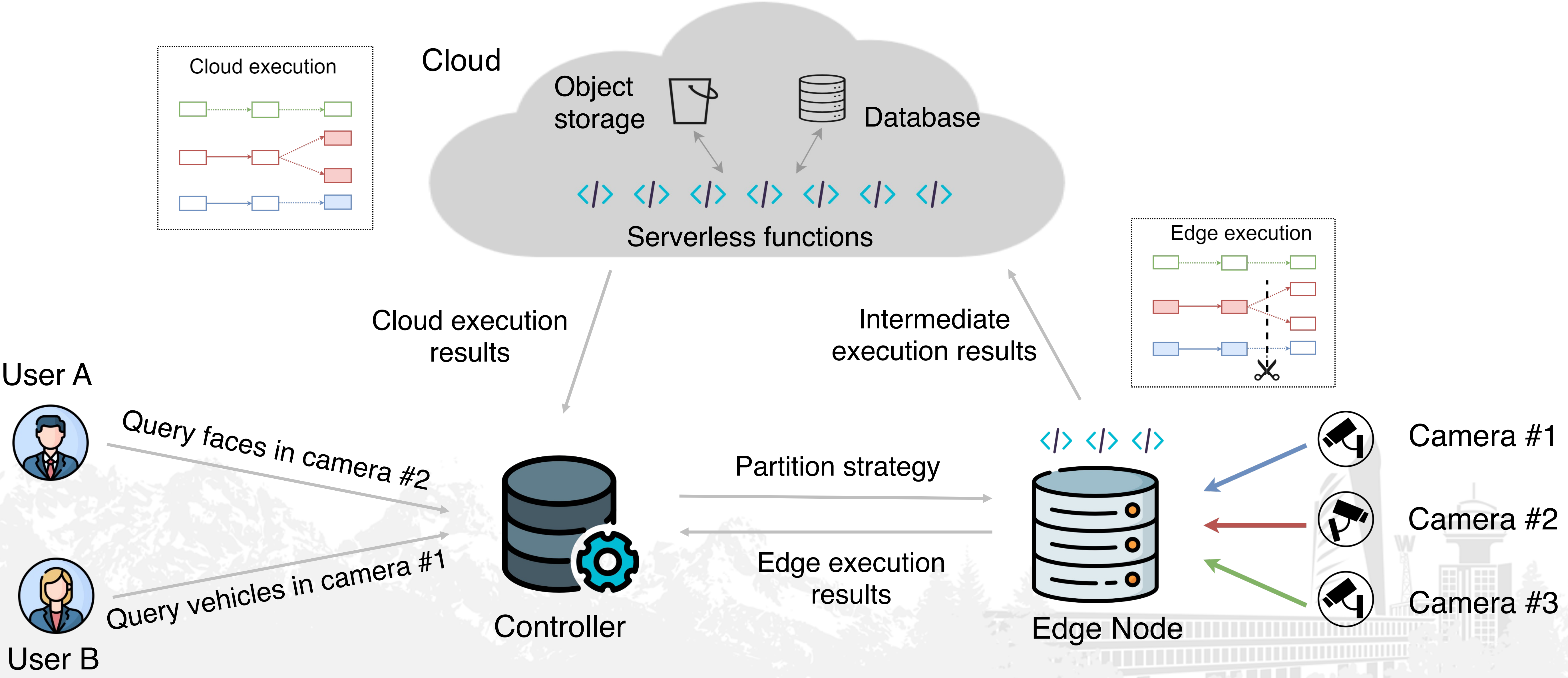
Microservice architecture → Improving flexibility and scalability

OUTLINE

- An Introduction to Video Analytics
- Measurement and Motivation
- CEVAS: System Design and Implementation
- Evaluation

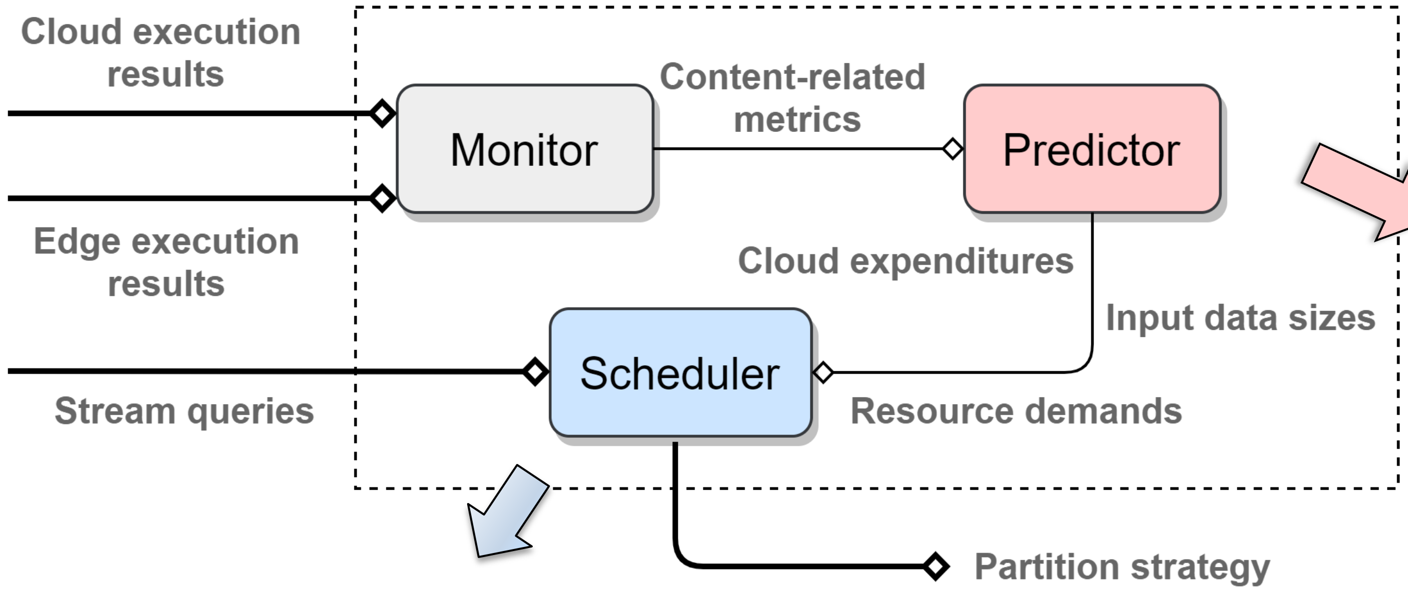
CEVAS

☐ Cloud-Edge collaborative Video Analytics with Serverless pipelines (CEVAS)



CEVAS

☐ Workload-Aware Runtime Scheduling

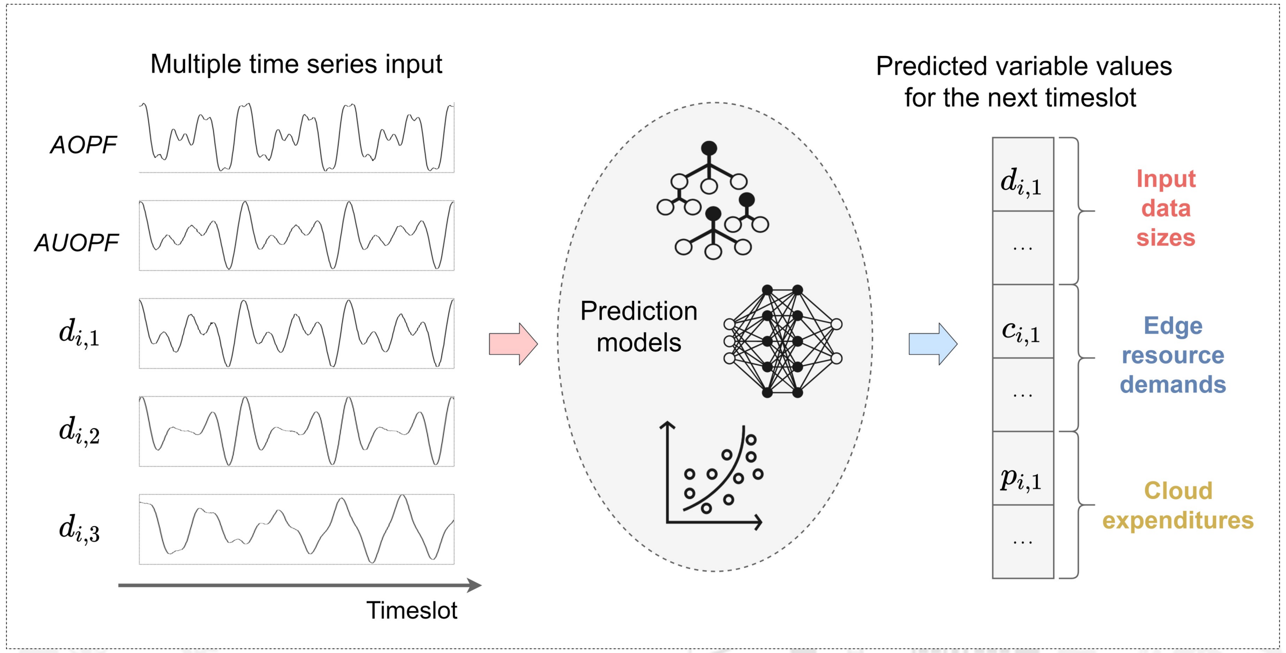


$$\min_{x_{i,j}} \sum_{i,j} (\alpha \cdot P_{i,j} \cdot x_{i,j} + \beta \cdot D_{i,j} \cdot x_{i,j})$$

$$s.t. \begin{cases} \sum_{i,j} C_{i,j} \cdot x_{i,j} \leq C \\ \sum_{i,j} G_{i,j} \cdot x_{i,j} \leq G \\ \sum_{i,j} M_{i,j} \cdot x_{i,j} \leq M \\ \sum_{i,j} M'_{i,j} \cdot x_{i,j} \leq M' \\ \sum_{j} x_{i,j} = 1, \quad x_{i,j} \in \{0,1\} \end{cases}$$

Runtime pipeline partition scheduling

Video content-aware workload, resource, and cost prediction



CEVAS

Cloud-side implementation

Video query pipelines: implemented in **Python** and deployed with **AWS Lambda**

Intermediate data: Amazon S3 for objects and Amazon DynamoDB for values

Edge Node

Hardware: NVIDIA GeForce GTX 1080 GPU, 12-core Intel Core i7-6850K CPU, 32 GB of RAM

Video query pipelines: deployed with **AWS IoT Greengrass** Core

Controller

Hardware: Off-the-shelf host

Software: AWS IoT Device SDK

OUTLINE

- An Introduction to Video Analytics
- Measurement and Motivation
- CEVAS: System Design and Implementation
- Evaluation

EVALUATION

☐ Video streams and queries

Two clips from the Crossroad camera (vehicle pipeline)

Three clips from the Restaurant camera (face pipeline)

☐ Model choice of the Predictor component

Multilayer Perceptron (MLP) models, pre-trained on corresponding camera streams.

☐ Evaluation metrics

Throughput, Cloud expenditure, Transferred data

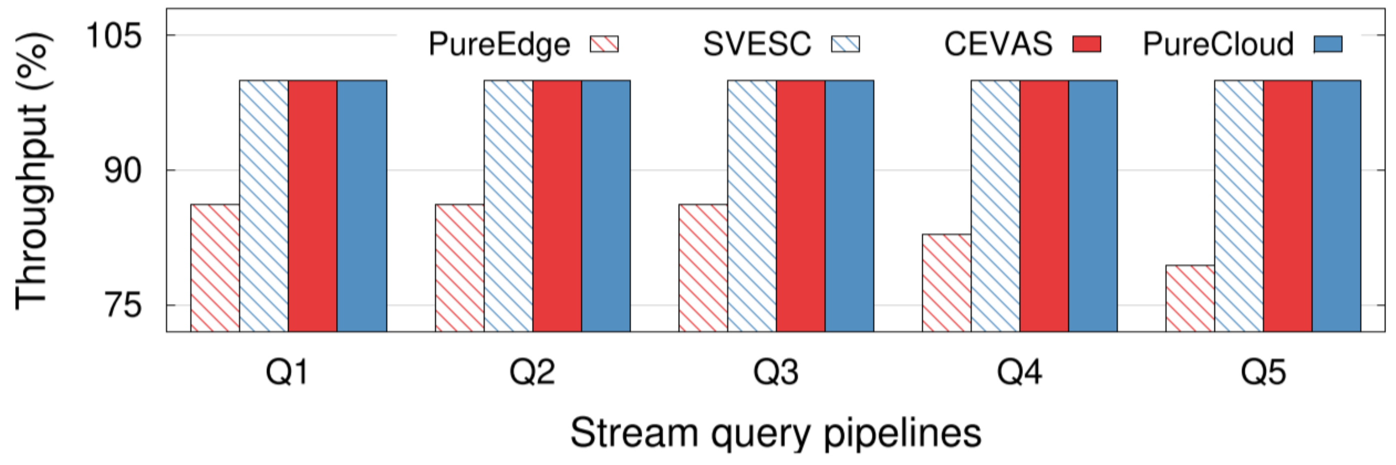
☐ Baselines

PureEdge, PureCloud

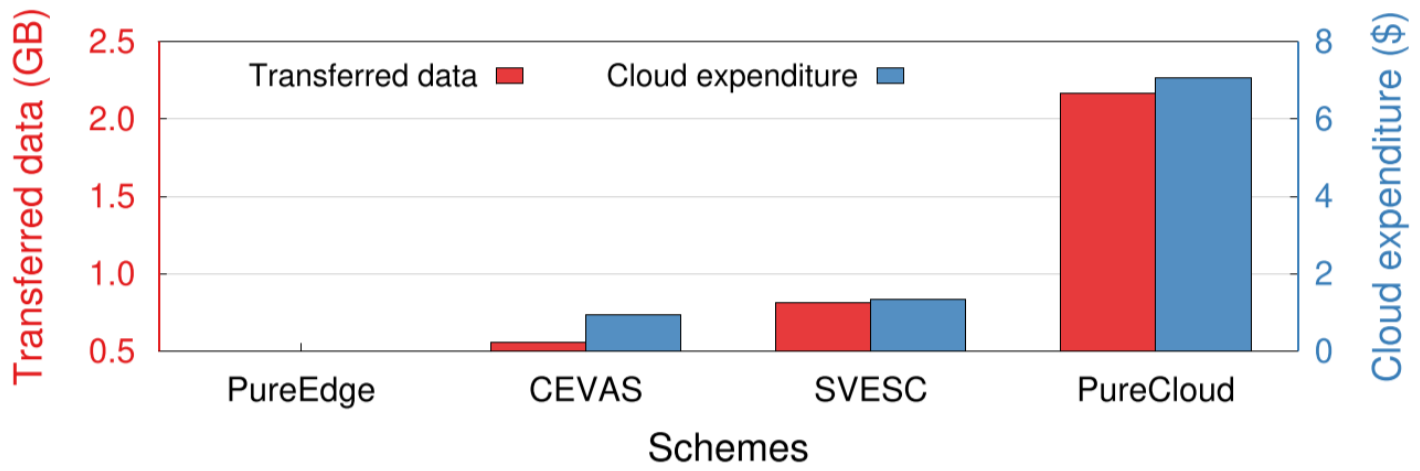
SVESC (a Slim version of VideoEdge with Serverless Computing supports)

EVALUATION

☐ Performance under persistent querying



Compared with PureEdge, CEVAS improves the throughput by up to 20.6%.

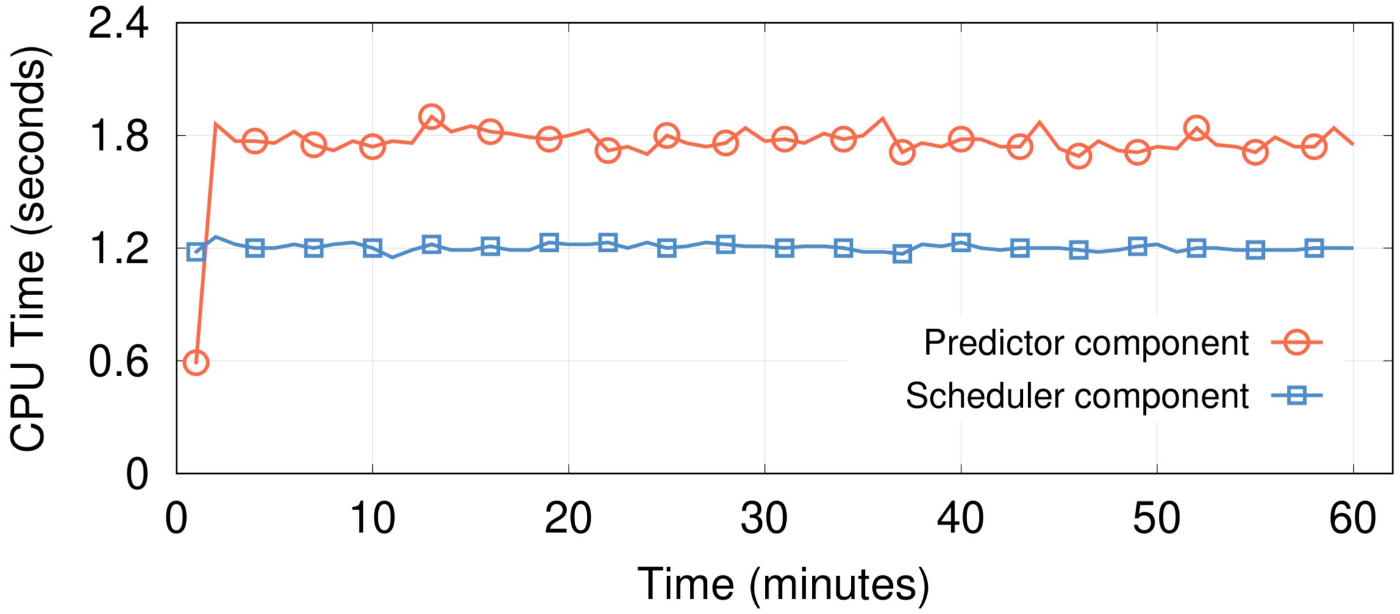


CEVAS reduces about 74.4% data transfer overhead and 86.9% cloud expenditure of PureCloud.

CEVAS reduces SVESC's data transfer overhead by 31.4% and cloud expenditure by 30.9%.

EVALUATION

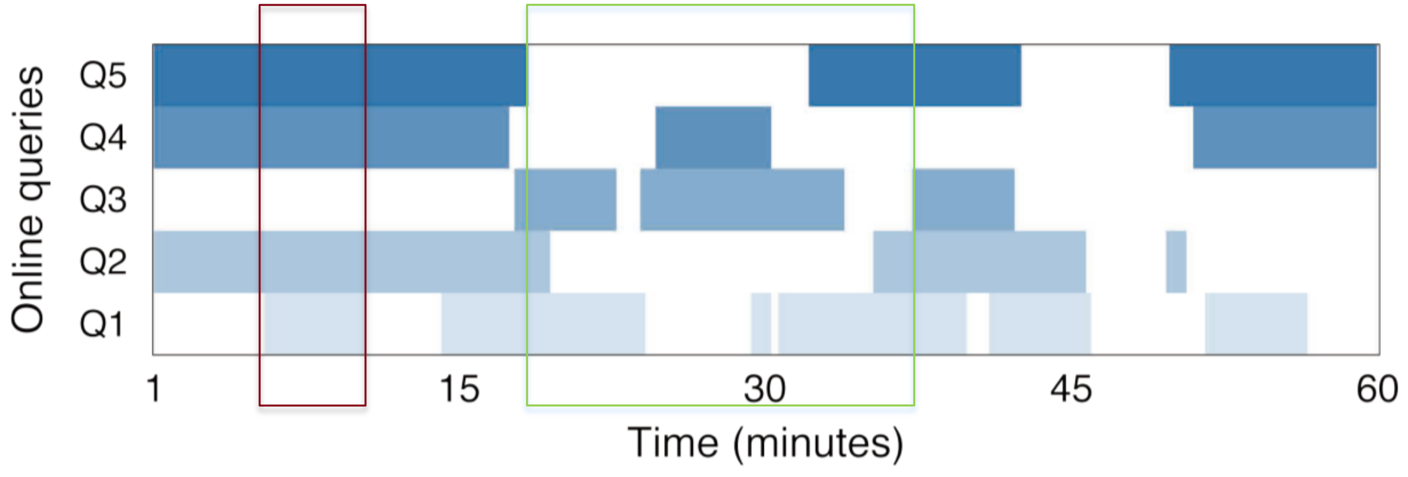
☐ System Overhead



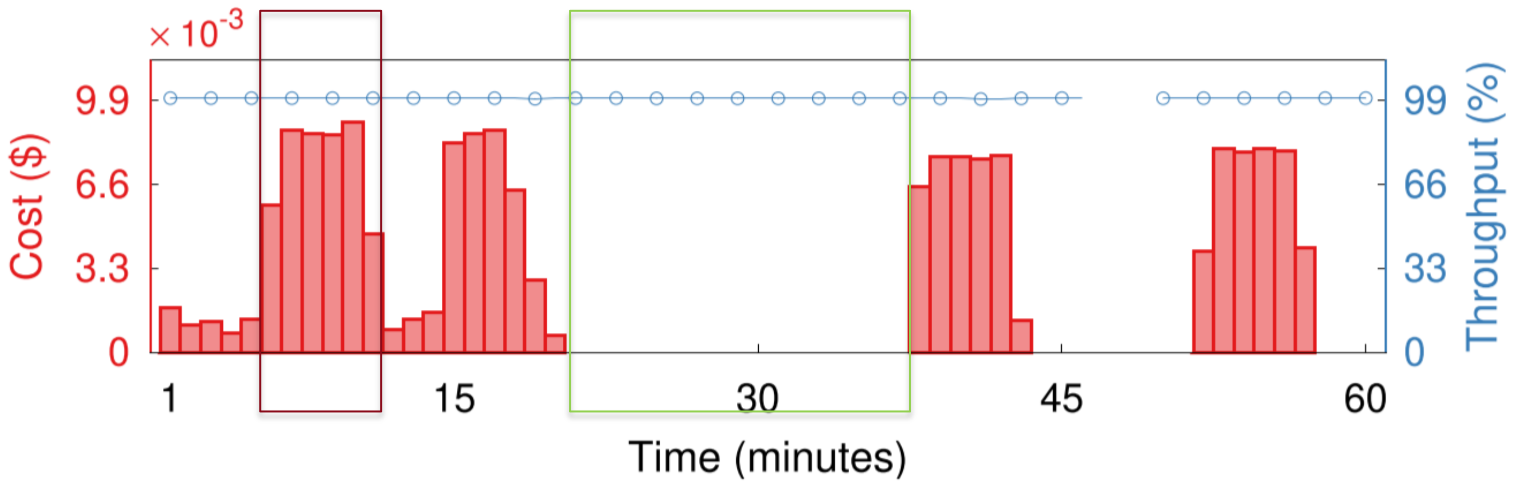
The values are accumulated CPU times for analyzing one-minute length video.

EVALUATION

☐ Performance under random querying



(a) Query arrivals and departures



(b) Performance of CEVAS

The colorized time windows in each row of (a) indicates the existence of queries on a specific video stream. (b) assumes it costs $0.1 to transfer 1GB data between the edge and cloud and sums up the money paid for data transfer and cloud expenditure to obtain the cost values.

THANK YOU

Q & A