

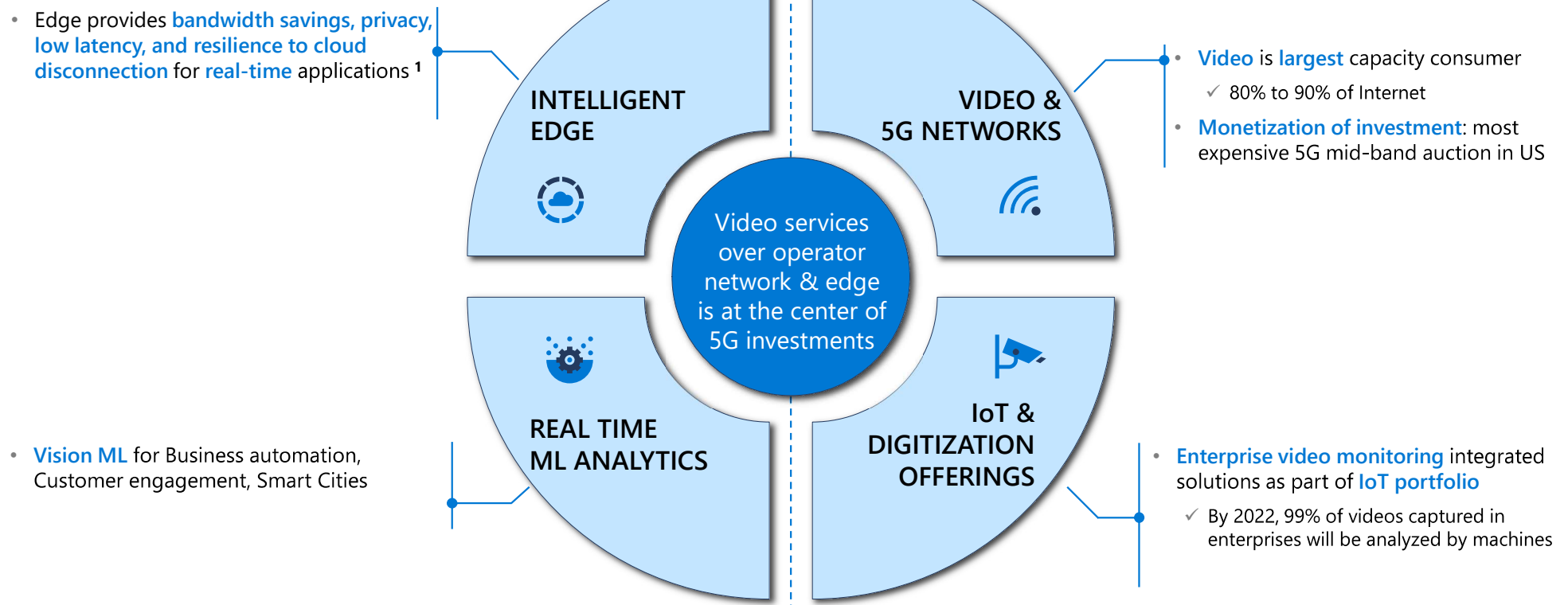


Edge Video Services on 5G Infrastructure

Ganesh Ananthanarayanan
Microsoft Azure for Operators



Real-time Video Analytics at the center of large trends



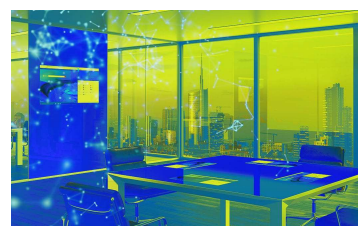
¹ The Emerging Landscape of Edge-Computing, ACM SIGMOBILE GetMobile, Mar 2020

Operators are well-positioned and keen to light up video services

KEY ENTERPRISE SCENARIOS

Advantage of Operators' position

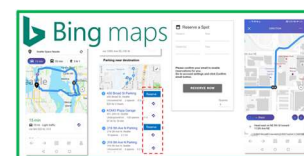
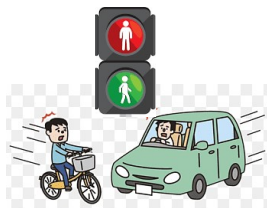
ENTERPRISE CAMERA NETWORKS



- Ability to offer packaged end-to-end security solutions with device integration

1 E2E video analytics solutions are a natural component of private 5G network enterprise solutions

SMART CITY TRAFFIC ANALYTICS



- Aligned with operators' promise in 5G networks of actions & insights based on live video

2 Wide geographic footprint to connect new cameras and upgrade existing legacy camera systems

CONNECTED FACTORIES OF THE FUTURE

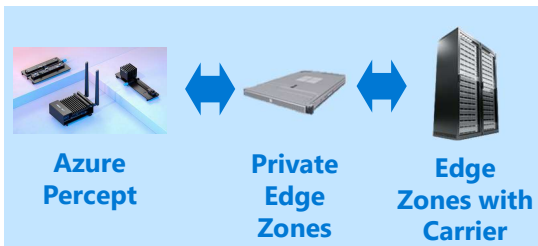


3 Existing leased lines to businesses

- Private 5G better suited than Wi-Fi for critical settings that required highly reliable connectivity
- Millimeter wave spectrum easy to deploy in private setting permitting full benefit of 5G

Designing video analytics to be operator-focused

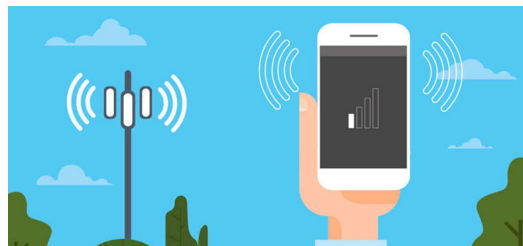
EDGE HIERARCHY WITH DIVERSE HARDWARE



Far and near edges:
On-prem, base stations, in-network

Requirement #1
Inter-edge orchestrator with
efficient video processing

DYNAMIC NETWORK CONDITIONS



Wireless & edge
connections vary with time

Requirement #2
Network monitoring & adaptation
of video processing

EXECUTE ALONGSIDE RAN WORKLOADS



RAN's demand is elastic based on
number of active users

Requirement #3
Dynamic resource allocation by the
orchestrator for video processing

[1] Smart city traffic on 5G edge hierarchy



Car/bike/pedestrian counts & near-collisions by analyzing widely-deployed traffic cameras

Dashboard & alerts



Analytics & actuation

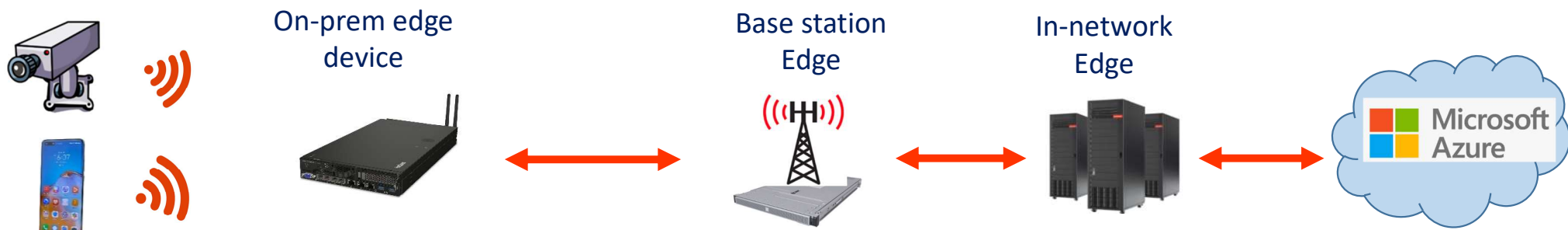


(Built up on prior work with City of Bellevue)

[1] Smart city traffic on 5G edge hierarchy



Vehicular analysis over hierarchy of edges in 5G infrastructure

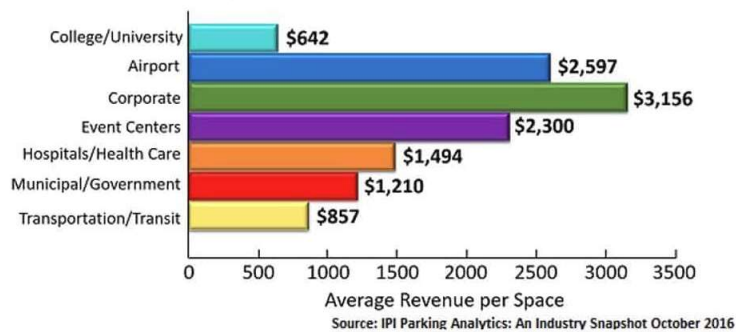


- ✓ Six-fold reduction in network traffic between the edges in the hierarchy, thus lowering the bandwidth needed to be provisioned
- ✓ Reduction in compute provisioning of edge devices via smart placement
- ✓ Vehicle counts from traffic camera videos with nearly 100% accuracy

[2] 5G Parking Services with Edge Compute

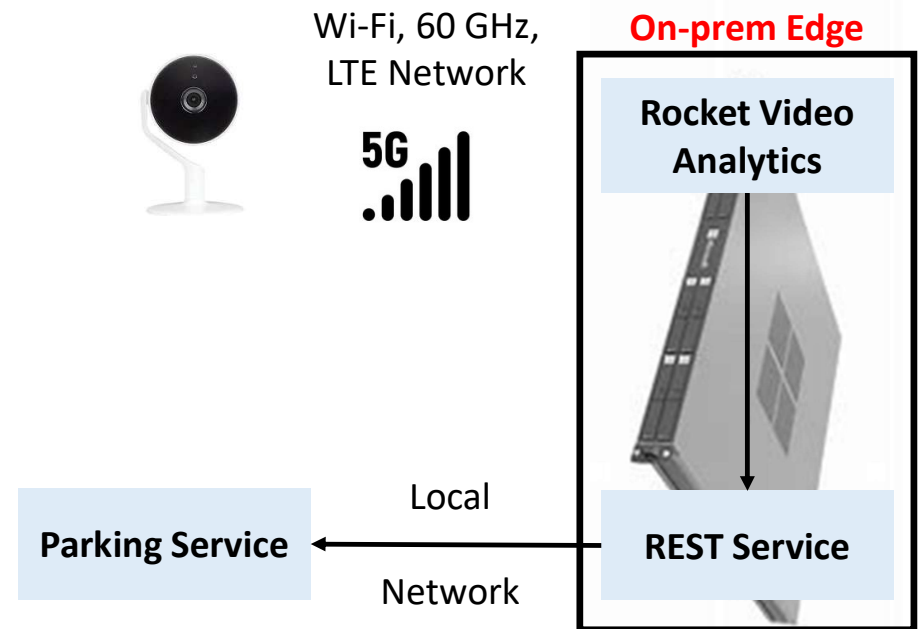
Parking Application: Finding parking can increase stress associated with traveling, CO2 emission, and traffic congestion (driving in circles)

Revenue Per Space



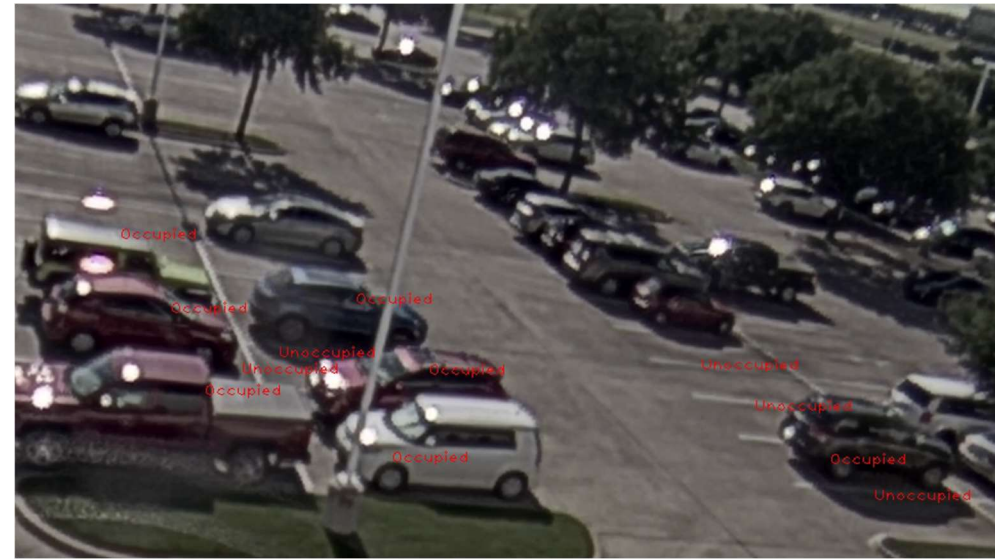
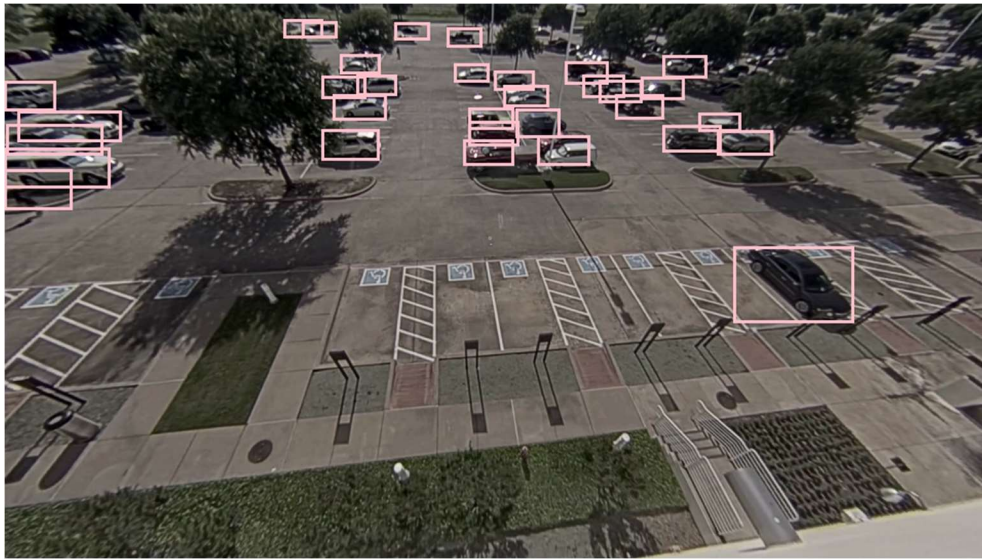
Sensors vs. Cameras

- ✓ Easily extend to other applications
- ✓ Cheap to scale up



[2] 5G Parking Services with Edge Compute

Parking Application: Finding parking can increase stress associated with traveling, CO2 emission, and traffic congestion (driving in circles)



Analyze live videos → detect vehicles → infer occupancies

[2] 5G Parking Services with Edge Compute



Parking Application: Finding parking can increase stress associated with traveling, CO2 emission, and traffic congestion (driving in circles)

tmforum




Ecosystem Catalysts

Description
Project addressing an ecosystem challenge

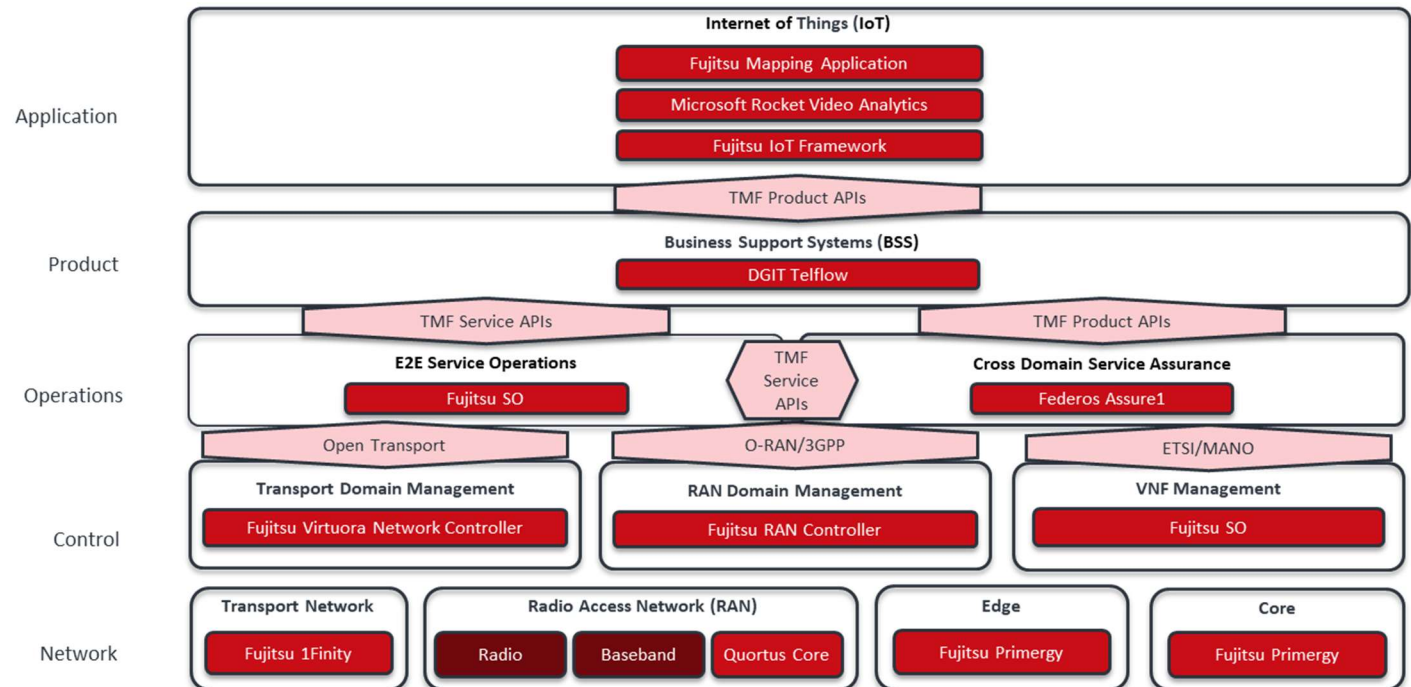
Critical Video Analytics Use Case

Catalyst Champion


City of Dublin
Catalyst Participants

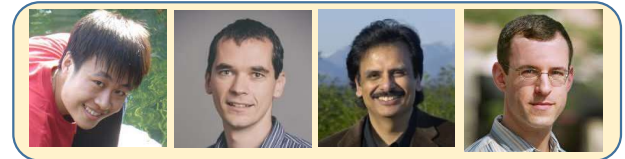
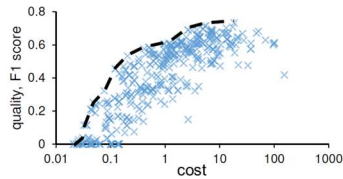
  

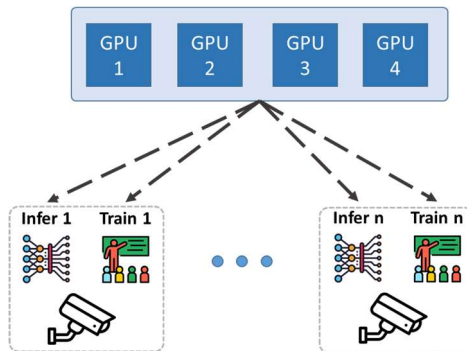


This talk will cover...

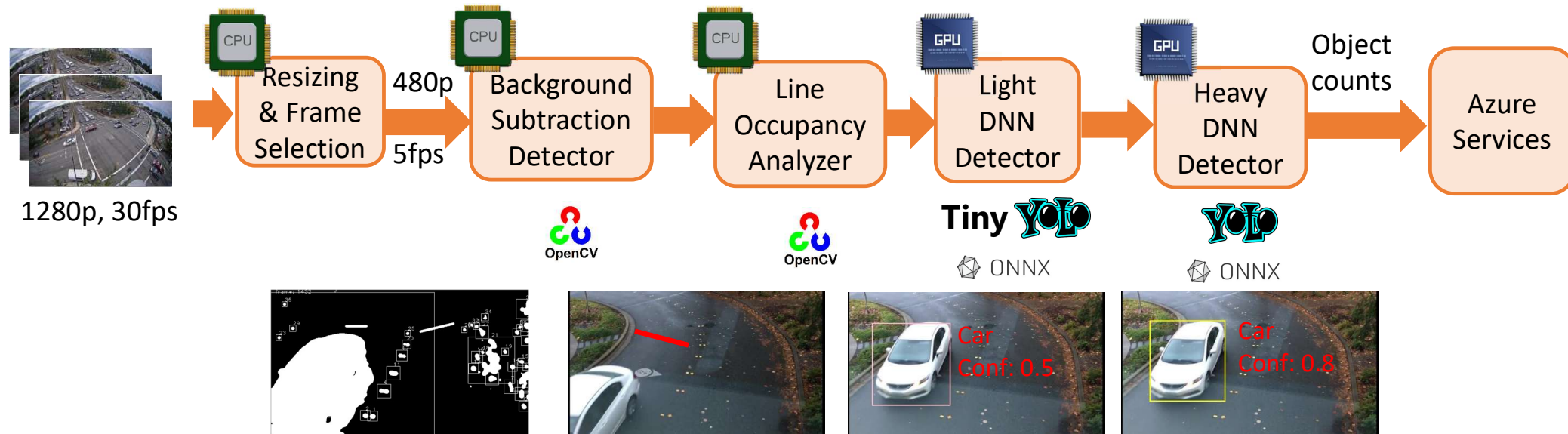
- Video analytics pipelines across edge/cloud with *approximation*



- Continuous learning of video analytics models on edge compute servers



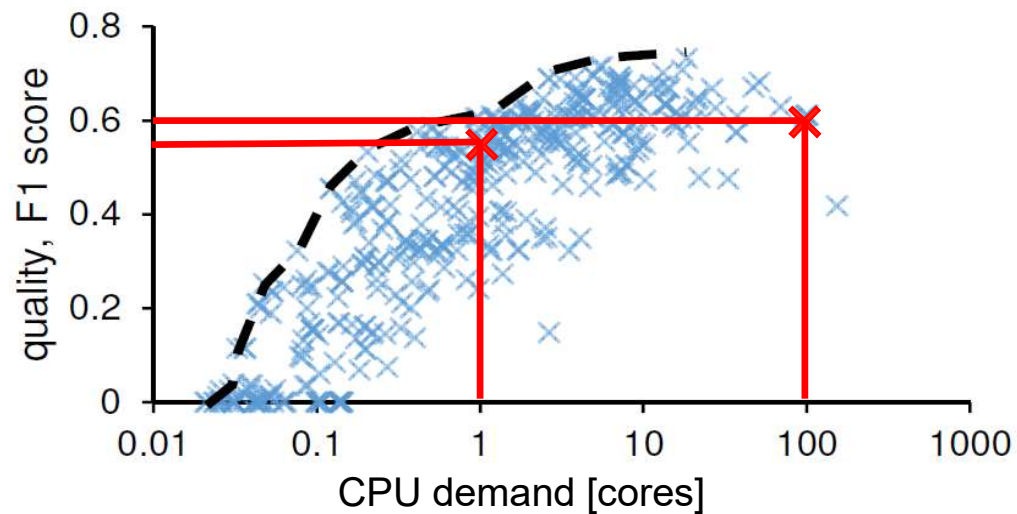
Cascaded video analytics pipeline



Configurations:

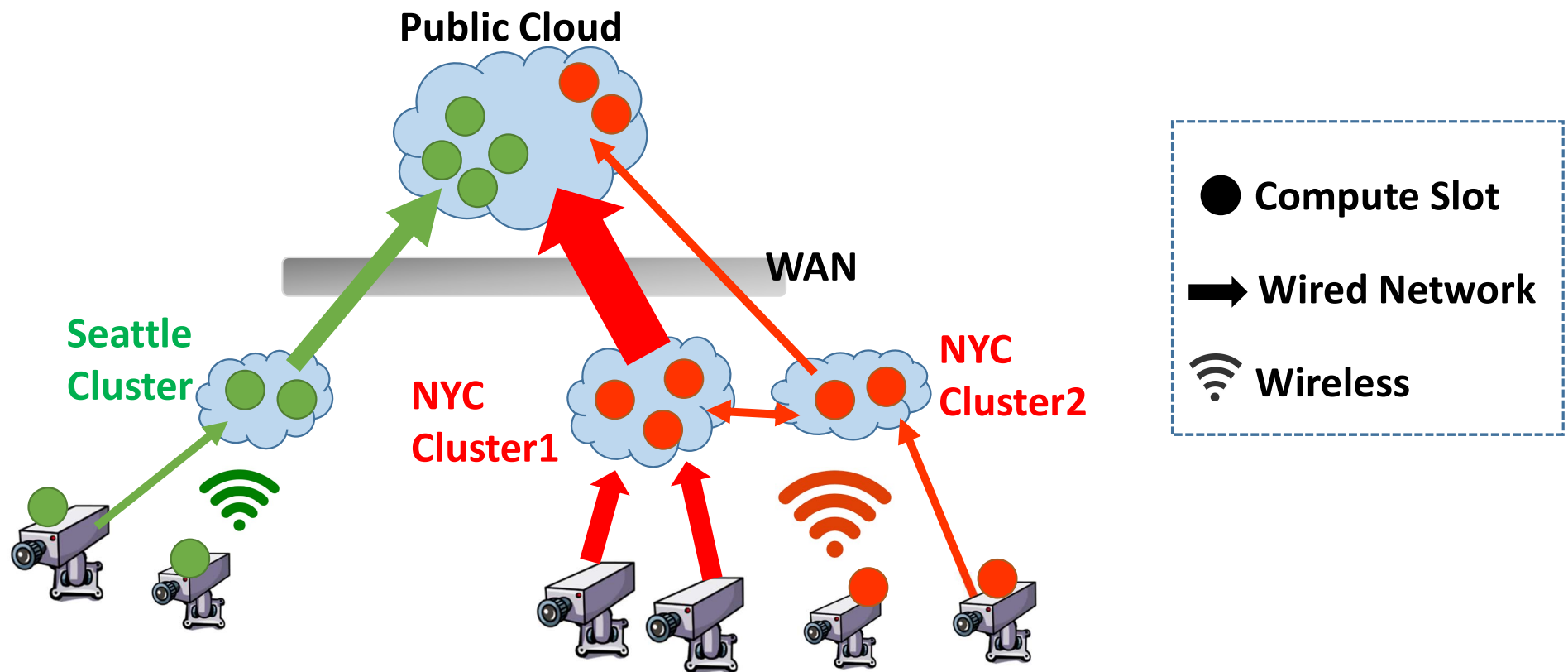
- Resolution
- Frames rate
- Object detector

How much do the *configurations* differ?

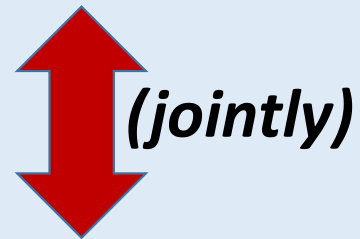


Orders of magnitude cheaper resource demand for little quality drop

Hierarchy of clusters for video analytics



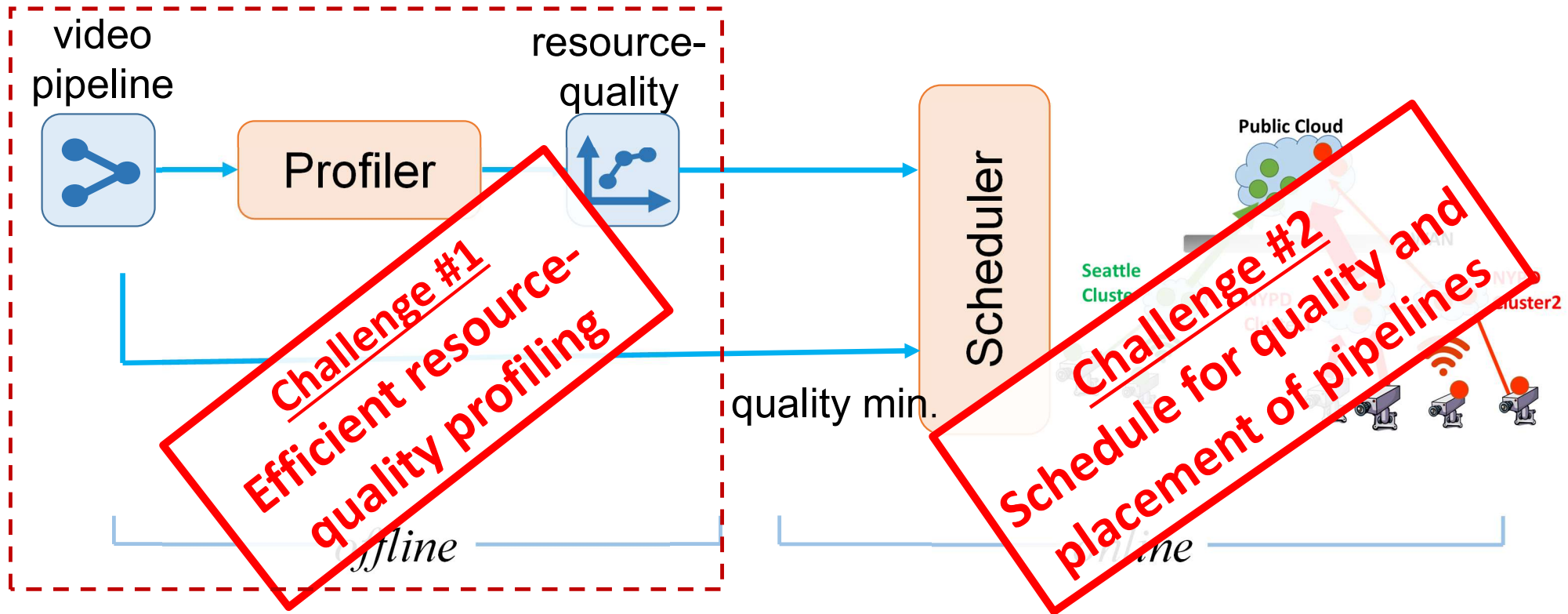
1. Pick *configurations* for video pipelines



(jointly)

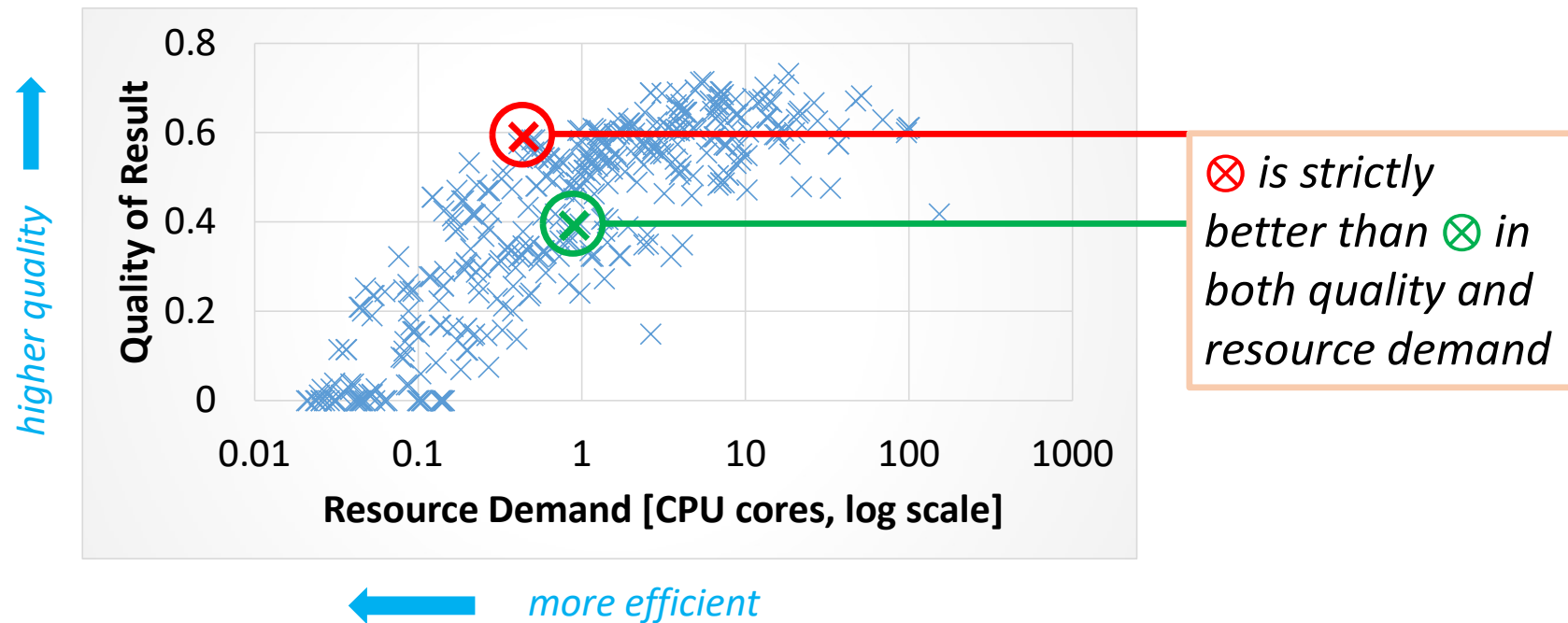
2. Place the *modules* across the hierarchy of clusters

Solution Overview



Offline: Resource-Quality Profiling

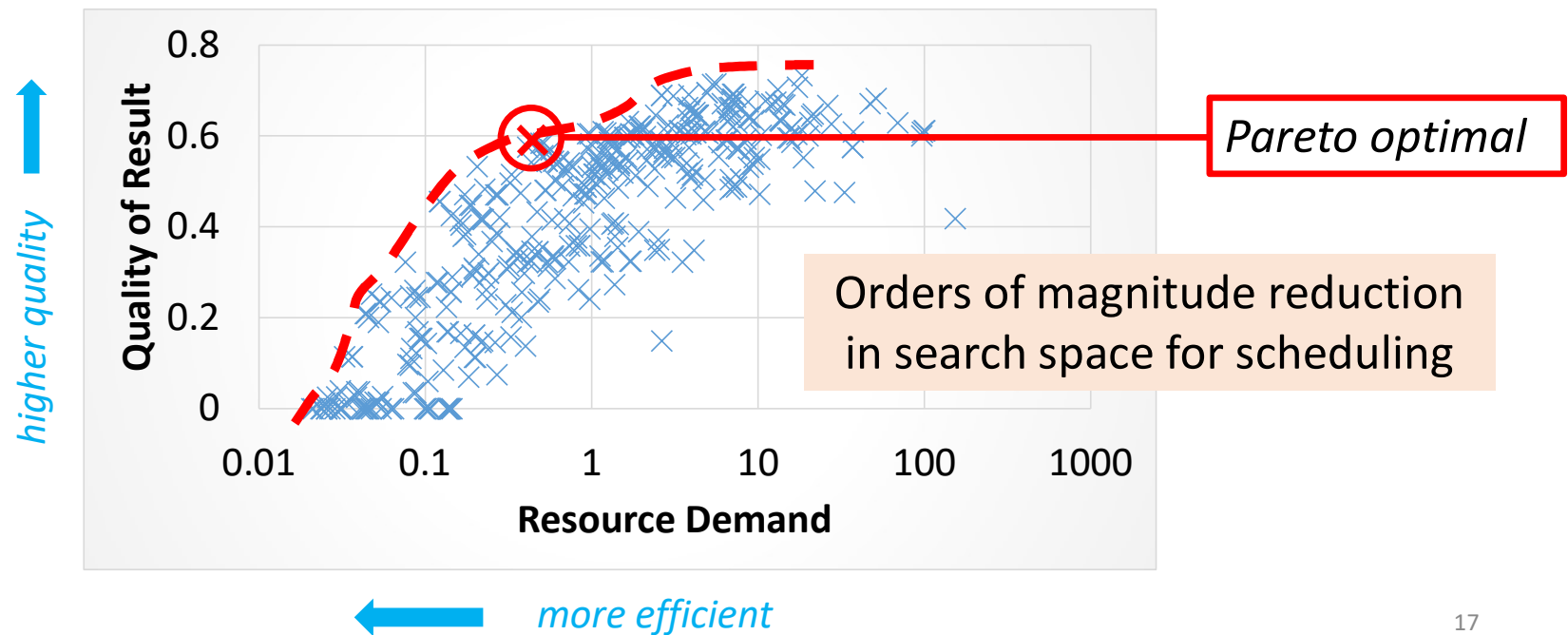
- Profile: configuration \Rightarrow {resource, quality}
 - Ground-truth: labeled dataset or results from *golden* configuration
 - Targeted search for promising configurations



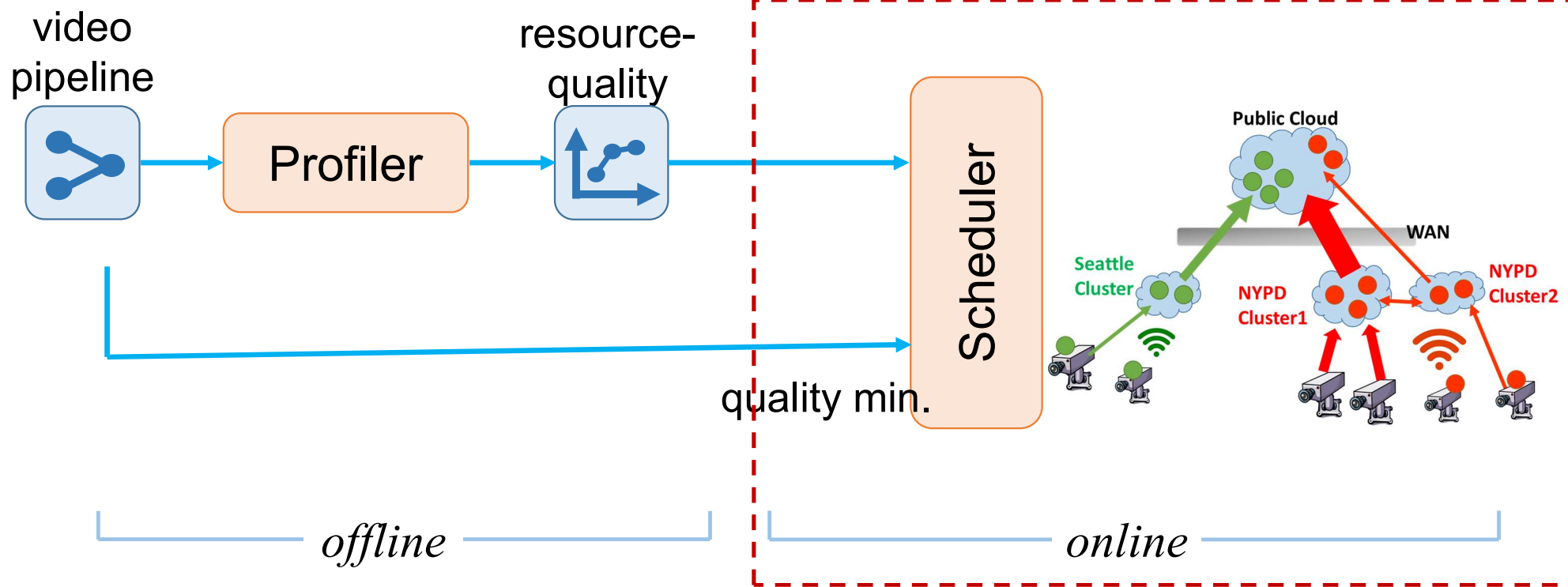
Offline: Pareto boundary

Pareto boundary: optimal configurations in resource demand and quality

- Non-Pareto plans cannot beat Pareto configs. in *both* quality & resources



Solution Overview



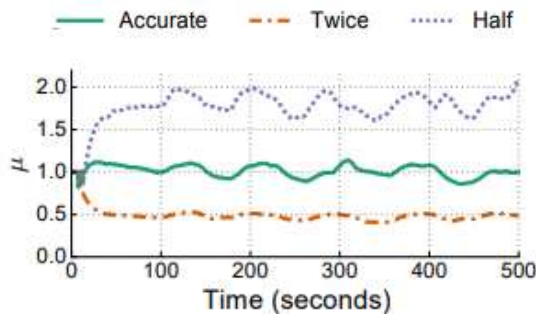
Evaluation Highlights

Workload

- Videos from traffic cameras & surveillance cameras
 - Original frame rate of 14 – 30 fps, resolution 480p – 1080p
- Workload: Object tracker, DNN classifier, Car counter, License plate reader

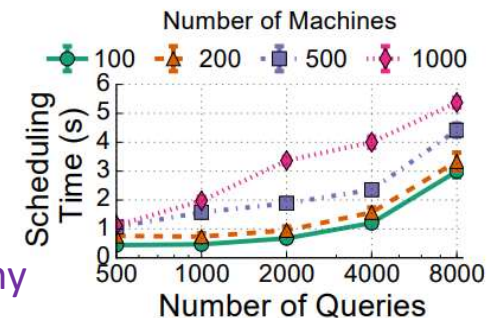
Results

- 25x better accuracy & within 6% of optimal



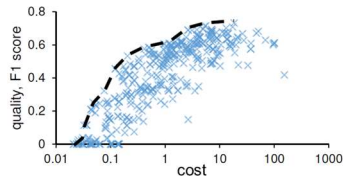
Adapts to errors
in the profile

Scales to many
1000's of queries

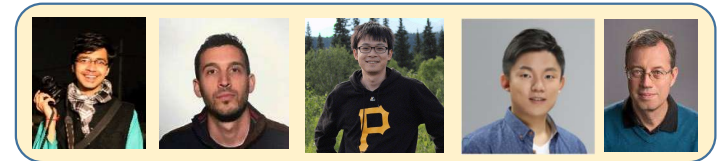
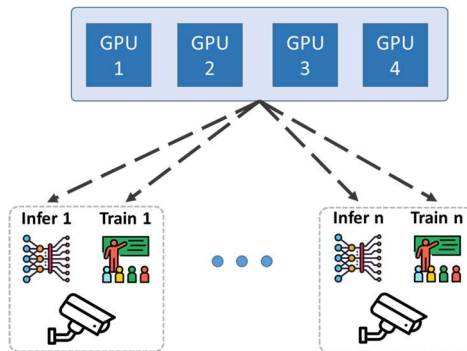


This talk will cover...

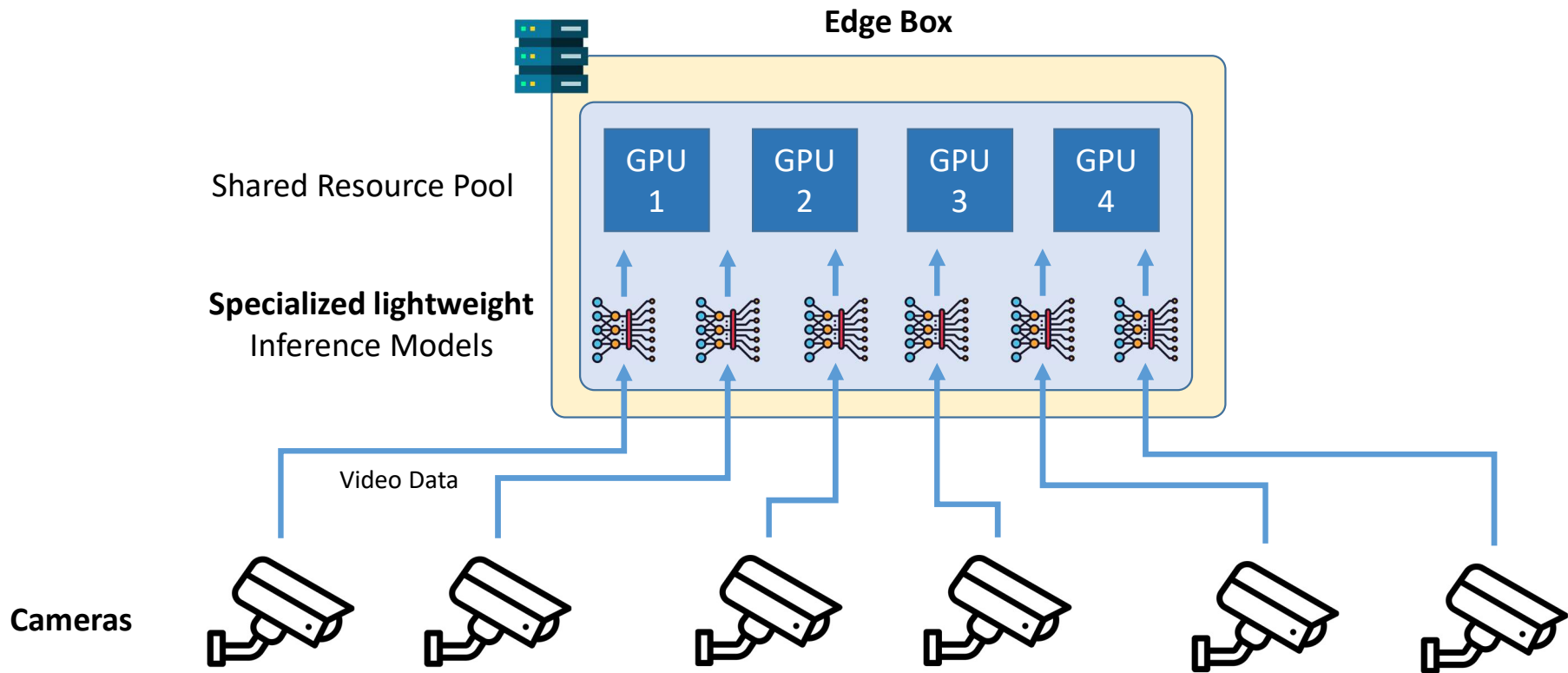
- ✓ Video analytics pipelines across edge/cloud with *approximation*



- Continuous learning of video analytics models on edge compute servers

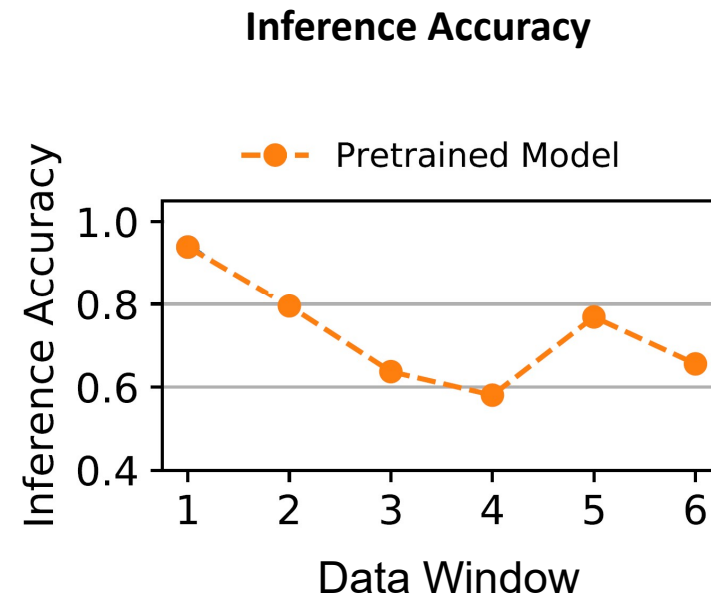
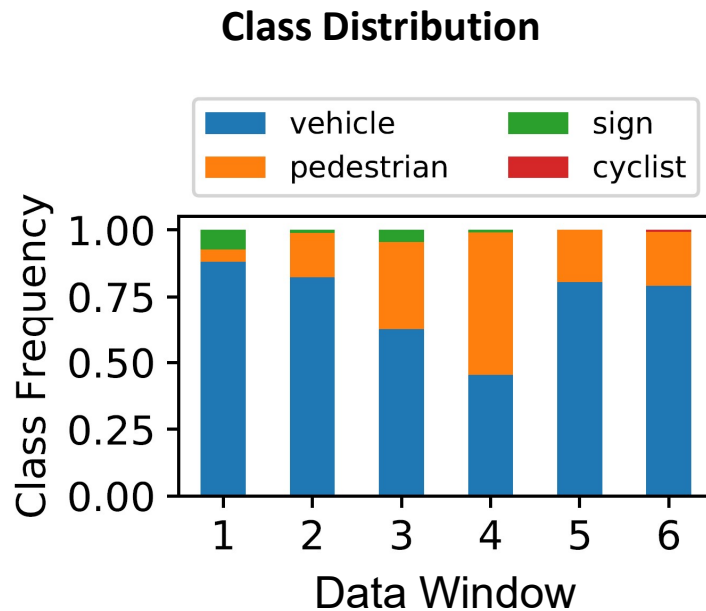


Edge Video Analytics Setup

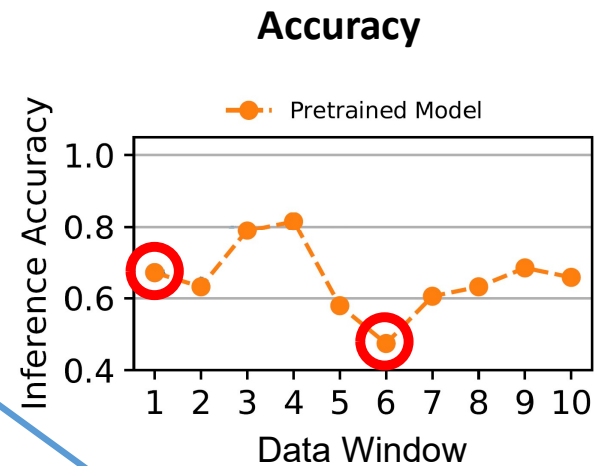
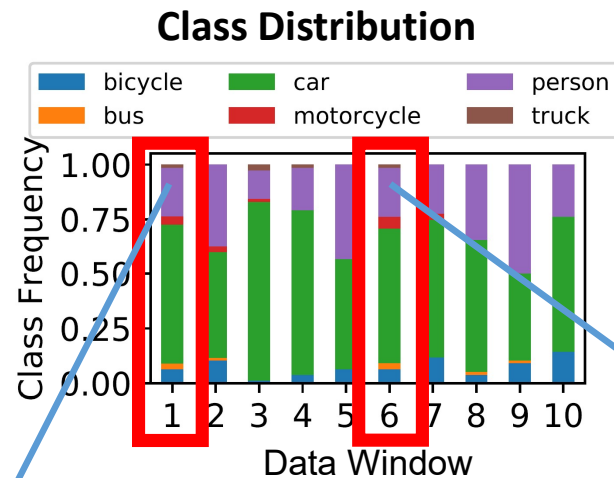


The challenge of data drift

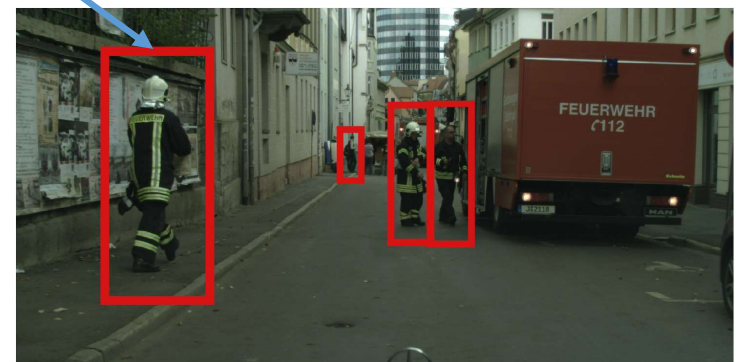
- Edge devices run lightweight models which have **limited generalizability**
- Observed data can be different than the training data, resulting in reduced accuracy
- Example – **Class Distribution Shifts**



Data Drift – Class Definition Shifts



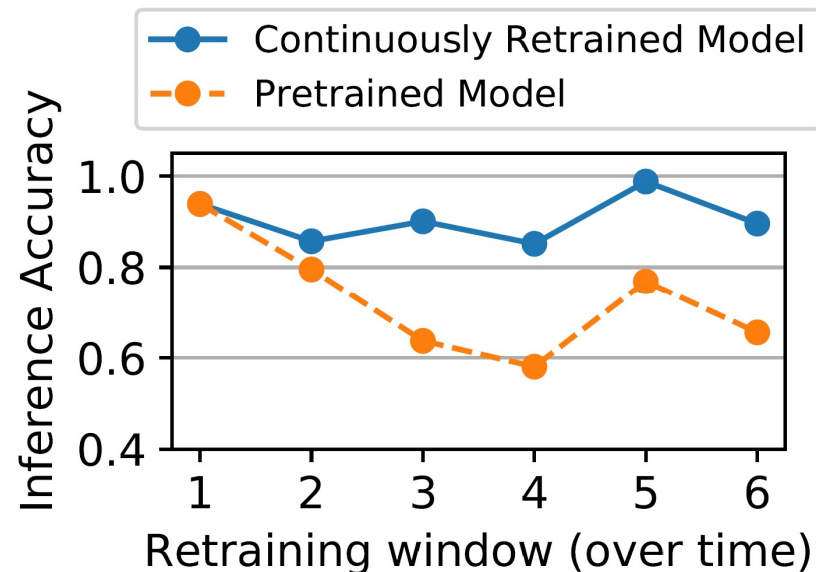
Task 1 – Person Class



Task 6 – Person Class

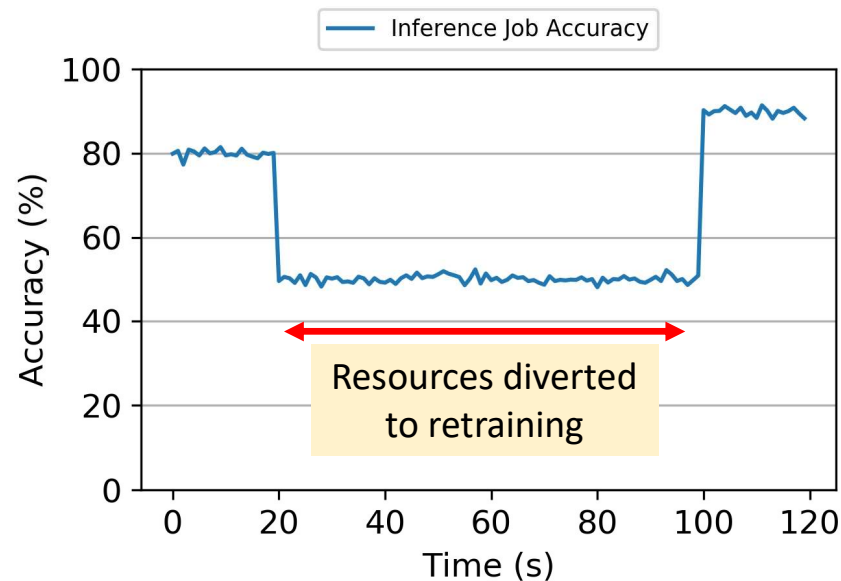
Model adaptation with continuous learning

- To counter data drift, we can adapt our models by **continuously learning** on incoming data
- Retraining is done periodically (creating “retraining windows”)



The cost of continuous learning

- Retraining models requires GPU-time, a precious quantity in resource constrained environments
- To retrain, we must **borrow resources** from inference and reallocate them to training
- Directly impacts inference accuracy

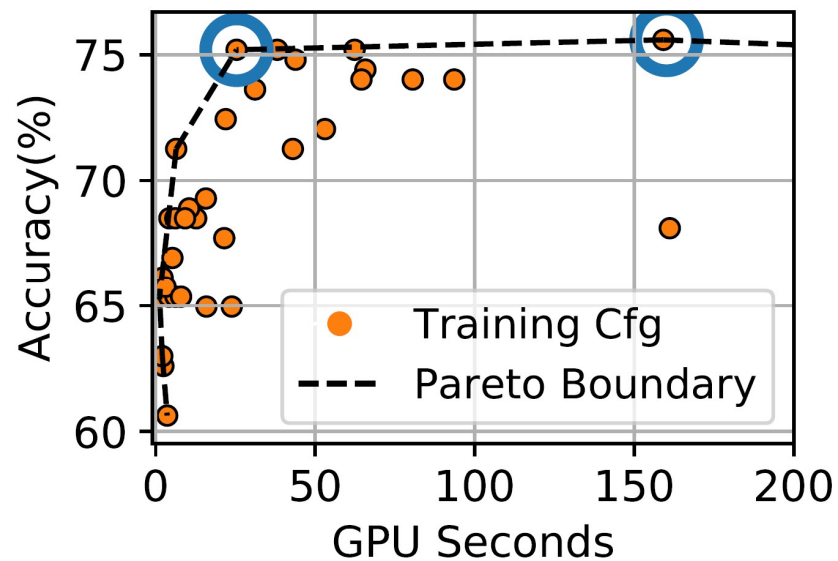


The cost of continuous learning

- The cost of retraining depends on the **configuration (hyperparameters)** chosen for retraining.

Hyperparameters:

- Layers to train
- Data sampling rate
- Learning rate
- Number of epochs
- Size of last hidden layer



Hyperparameters vs Cost

Scheduling Objective

Maximize mean inference accuracy *over the retraining window*
across all video streams

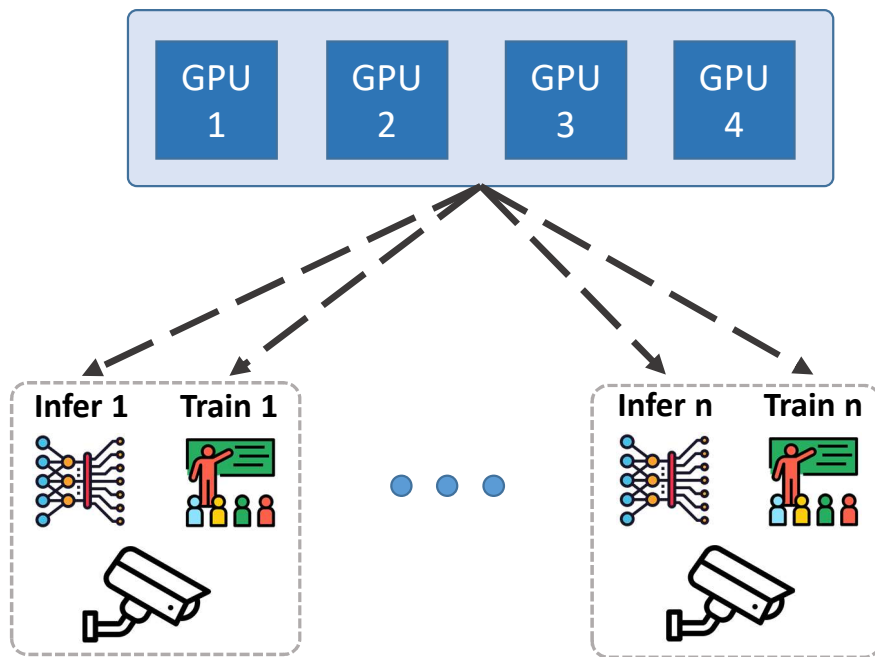
Subject to:

Resource capacity constraints
Minimum inference accuracy constraint

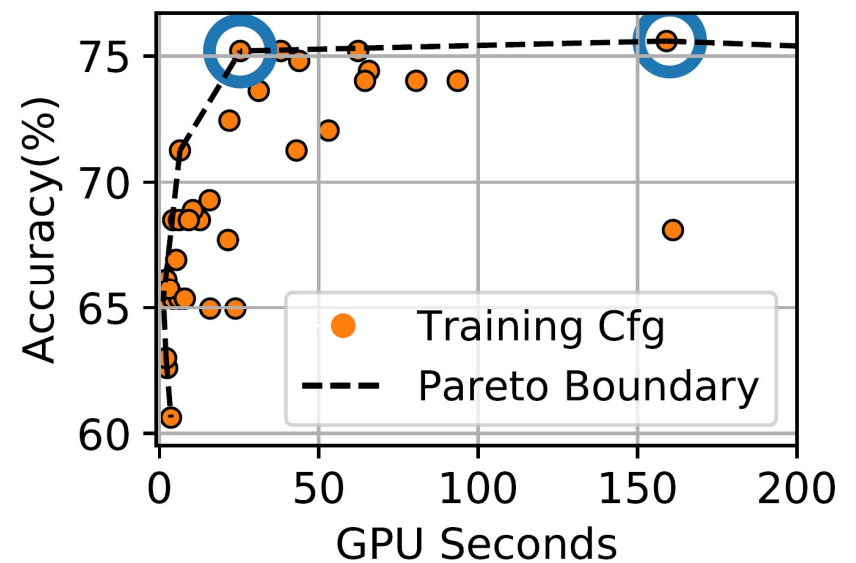
Scheduling decisions to make

Must be done jointly!

**Allocate resources between
Training and Inference
across video streams**



Pick Hyperparameters to train



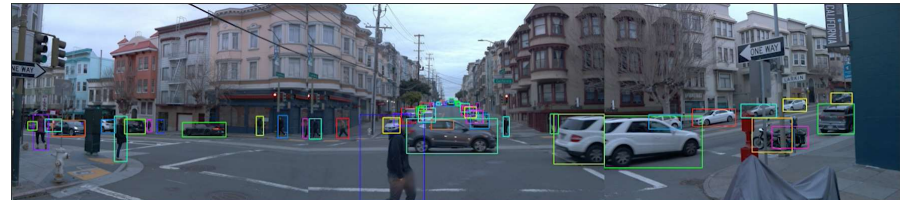
Ekya Thief Scheduler



- Start with a fair allocation to all video streams V
 - For each camera, evaluate all training configurations and pick the one which gives highest predicted accuracy
- For each thief job $j \in J$:
 - For victim job $k \in \{J - j\}$:
 - Steal a quantum of resource δ from k and allocate to j
 - Repeat configuration selection with new resource allocation
 - If expected mean inference does not improve, stop stealing from k . Else, steal again.

Evaluation

- Two datasets – Cityscapes and Waymo – both dashcam videos of driving in different cities
- Baselines – no-retraining and fair scheduler



- Ekya requires **4.3x fewer resources** to achieve the same inference accuracy as a fair scheduler

Ongoing work (*that I did not talk about*)

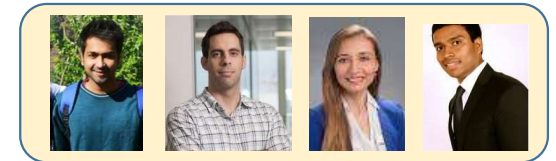
Cross-camera video analytics

- Large camera deployments in buildings, cities
- **Spatio-temporal correlations** for efficiency & accuracy



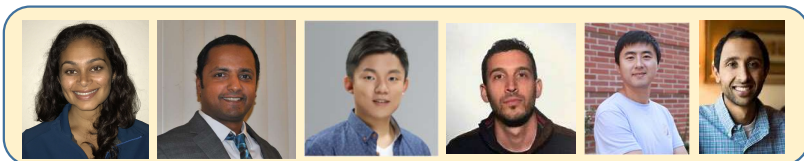
Private video analytics as a cloud service

- Side-channel attacks leak video content
- **Hybrid TEE (CPU + GPU enclaves)** design for data-obliviousness



Memory-Efficient Inference in Edge servers

- Memory is a bottleneck resource on edge servers
- **Merge layers of models** to reduce memory footprint



Multi-Hop mmWave Network of Cameras

- Mesh of cameras with HD videos for analytics
- **Data plane with mmWave networks, control plane with Wi-Fi**

