Joint Configuration Adaptation and Bandwidth Allocation for Edge-based Real-time Video Analytics

Can Wang, Sheng Zhang, Yu Chen, Zhuzhong Qian,
Jie Wu, and Mingjun Xiao

Outline

- Background
- Related work and motivation
- Key idea and algorithms
- Theoretical analysis
- Some evaluation
- Summary

Background

- Massive video recordings are happening everywhere

traffic control

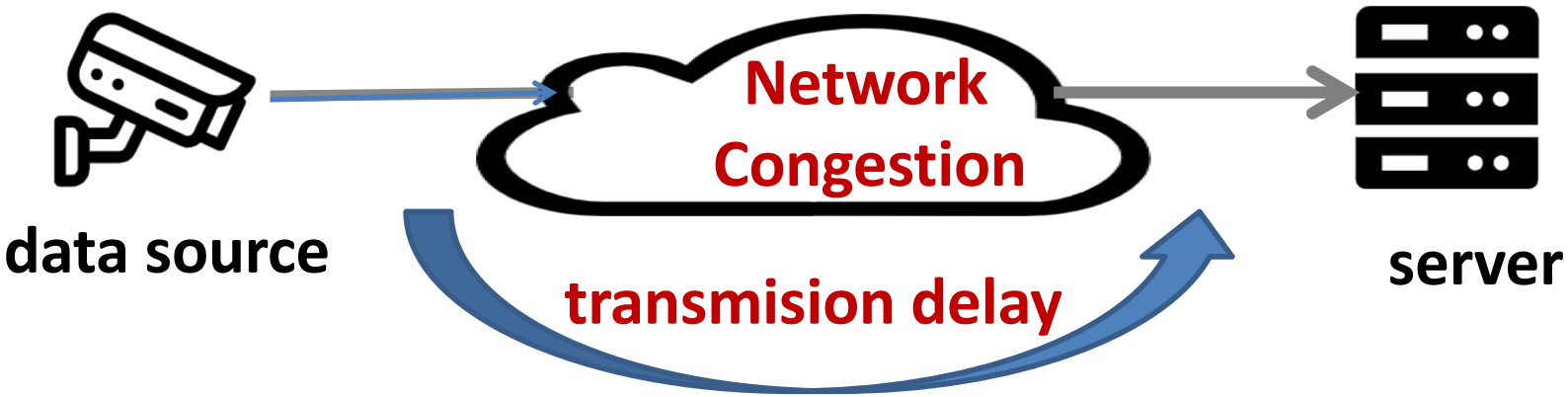
crime prevention

business intelligence

AR/VR

Background

- Real-time video analytics are expensive in resource usages
  - Best car tracker — 1 fps on an 8-core CPU
  - DNN for object classification — 30GFplops
- Cloud based solution incurs long delay



Related work and motivation

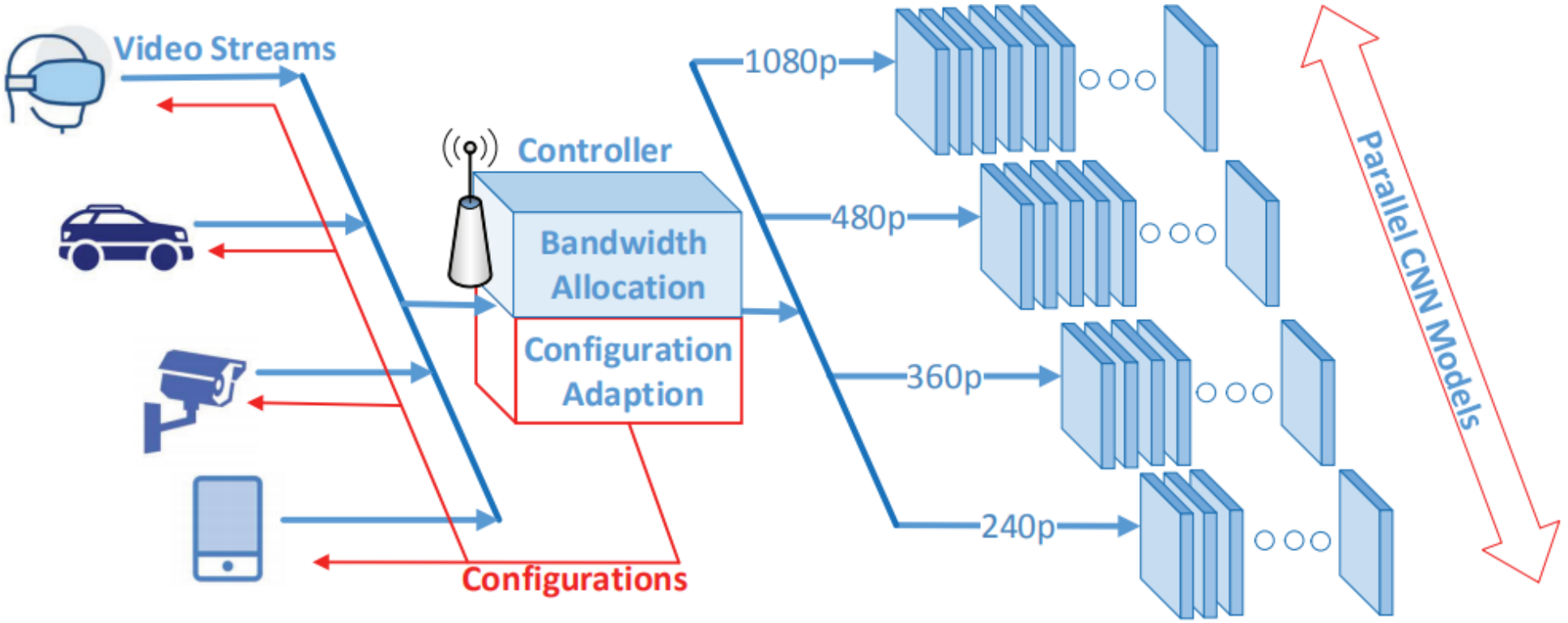
optimizing service delay:
video crowdprocessing [INFOCOM 18]

balancing between delay and acuracy:
edge network orchestrator[INFOCOM 18] Deepdecision [INFOCOM 18]

choosing the best configuration:
Chameleon[SIGCOMM 18] AWSream [SIGCOMM 18]

goal:
determine the optimal offloading strategy for video analytics tasks

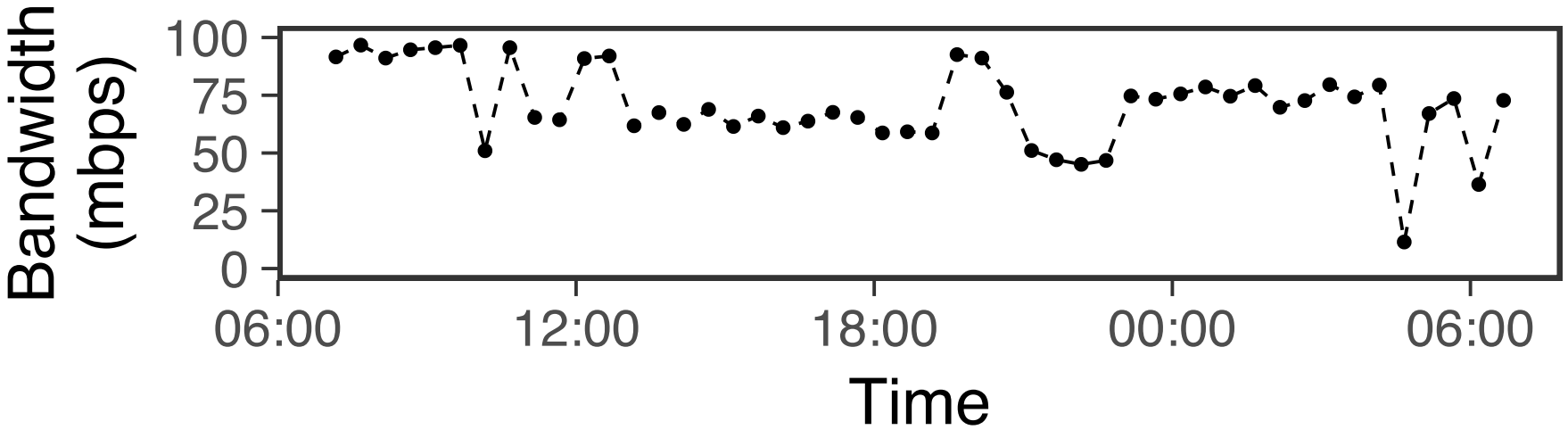
Edge-assisted Video Analytics System



Task Scheduling:
Joint Configuration Adaptation and Bandwidth Allocation

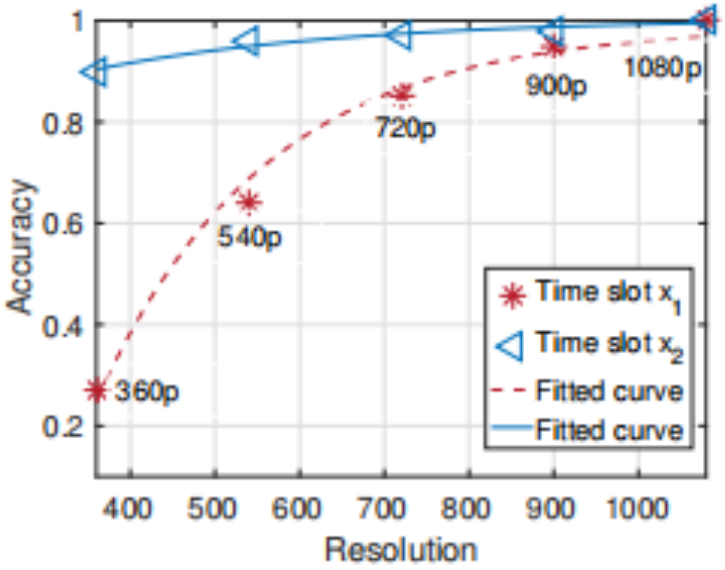
Challenges

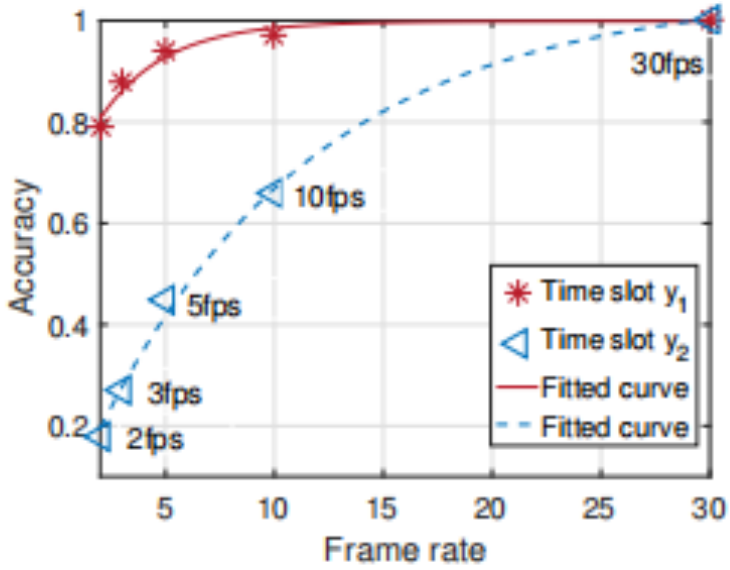
- The best offloading configuration varies over time.
  - optimize the trade-off between accuracy and energy consumption
- Network bandwidth is often unpredictable.

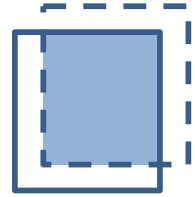


Key idea

- How to model analytics accuracy?
  - the relationship between resolution/framerate and accuracy can be formulated as concave functions
  - frame resolution and frame sampling rate independently impact accuracy







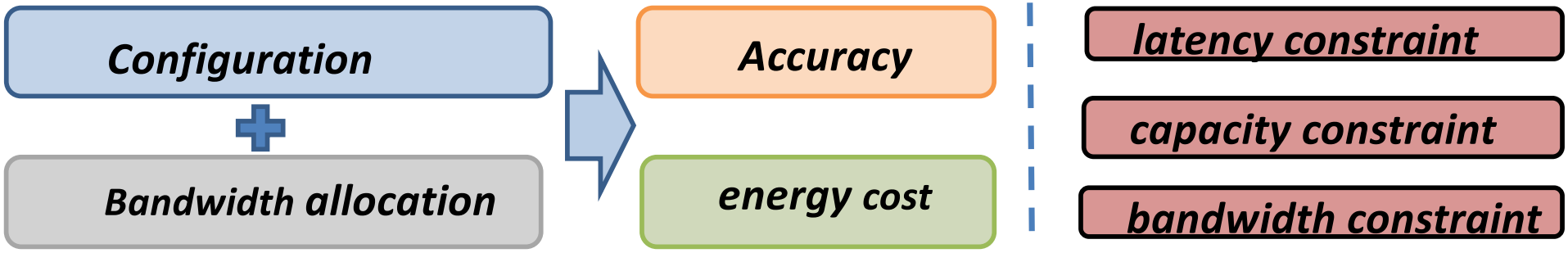
IOU=0.7

positive if ***intersection over union (IOU)*** larger than 0.7

the accuracy of the configuration can be formulated as the prduct of these two concave functions

Key idea

- long term optimization problem
  - achieving desirable analytics accuracy under the ***long-term latency constraint***
    - data transmision latency + data processing latency
  - Keeping energy cost as low as possible
    - data transmission & local CNN processing



Key idea

- Problem transformation using ***Lyapunov Optimization***
  - introduce a ***virtual queue*** as a historical measurement of the exceeded latency
  - it is crucial to keep the latency queue stable
    - we attempt to minimize the supremum bound for the ***drift-plus-penalty*** function
- One slot optimization problem
  - Only rely on the **current system information**
  - The new problem is the **weighted sum of latency, accuracy and energy cost**, which is NP-Hard in general.

Algorithms

- The latency queue guides us to follow the long-term latency constraint thereby enabling ***online decision making***.

The JCAB Algorithm

for t = 0 **to** T

Profile accuracy function of resolutions

Profile accuracy function of frame rates

Selecting the best ***model selection policy***, ***bandwidth allocation scheme***, and ***frame rates*** by solving the one slot optimazation problem.

update the virtual queue

Solving online optimization problem

- Once model selction variables are fixed, ***two sub-problems*** left to be sloved:
  - optimizing bandwidth allocation to reduce latency.
  - adapting frame rates to maximize configuration utility.

  (***optimal*** bandwidth allcation and frame rates can be derived using ***convex optimization*** techniques)

- How to find the best model selection policy?
  - ***Markov optimization*** based method.

Solving online optimization problem

partially depending on the objective value difference of the old and the new solution

One Slot Optimization for JCAB

Repeat

Randomly pick a user k and change its model

obtain optimal bandwidth allocation and frame rates

With ***probability η***, user k accepts the new model

With probability (1 - η), user k keeps model unchanged

until no significant improvment can be achieved

13

Theoretical analysis

Theorem 1: JCAB achieves the following performance bounds on the time-averaged utility and queue backlog:

average objective value

optimal value of the original problme

control parameter

$$\lim_{T\to+\infty}\frac{1}{T}\sum_{t=0}^{T}E[a_t-\omega e_t]>=\nu_{opt}-B/V,$$

$$\lim_{T\to+\infty}\frac{1}{T}\sum_{t=0}^{T}E[l_t]\leq\frac{B}{\varepsilon}+\frac{V}{\varepsilon}(\nu_{opt}-\nu_{min})+L_{max}$$

average latency

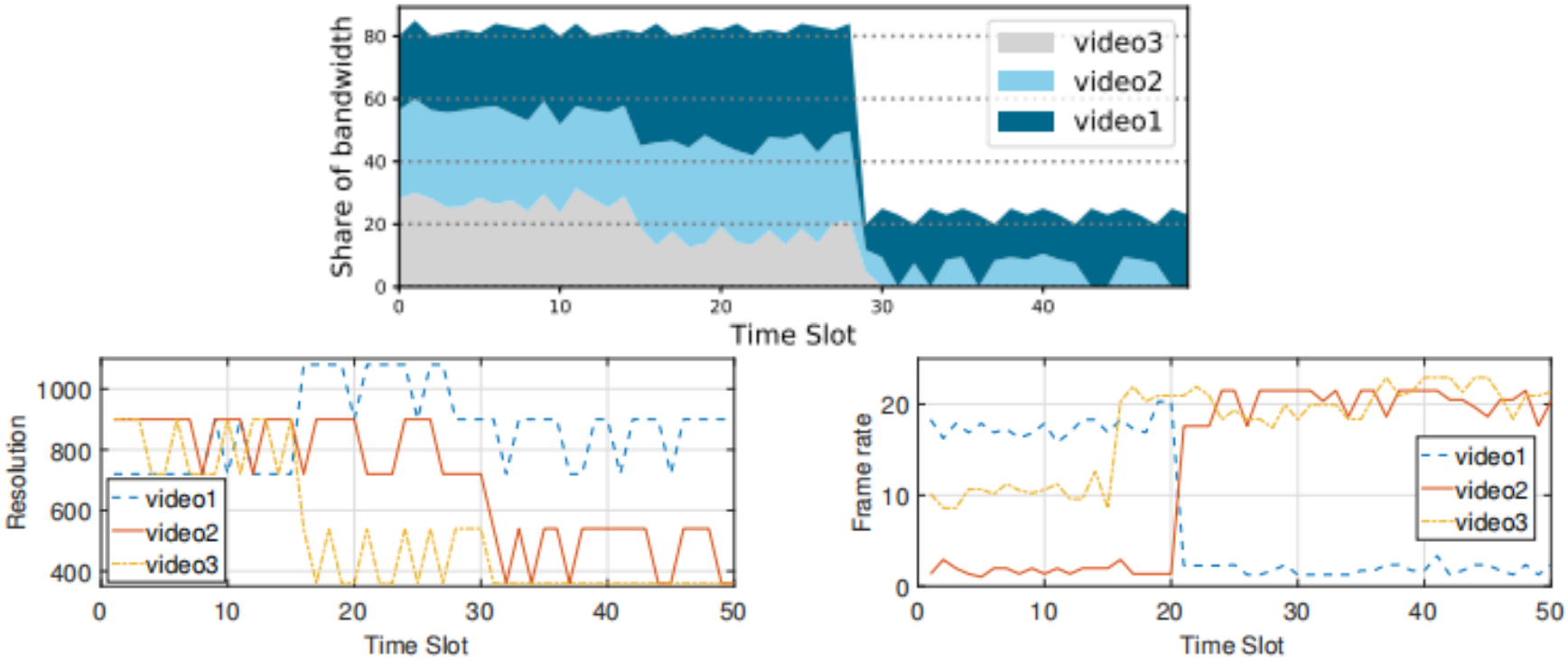
worst solution of the original problem

latency constraint

- utility delay tradeoff is characterized within ***[O(1/V),O(V)]***

Some evaluation

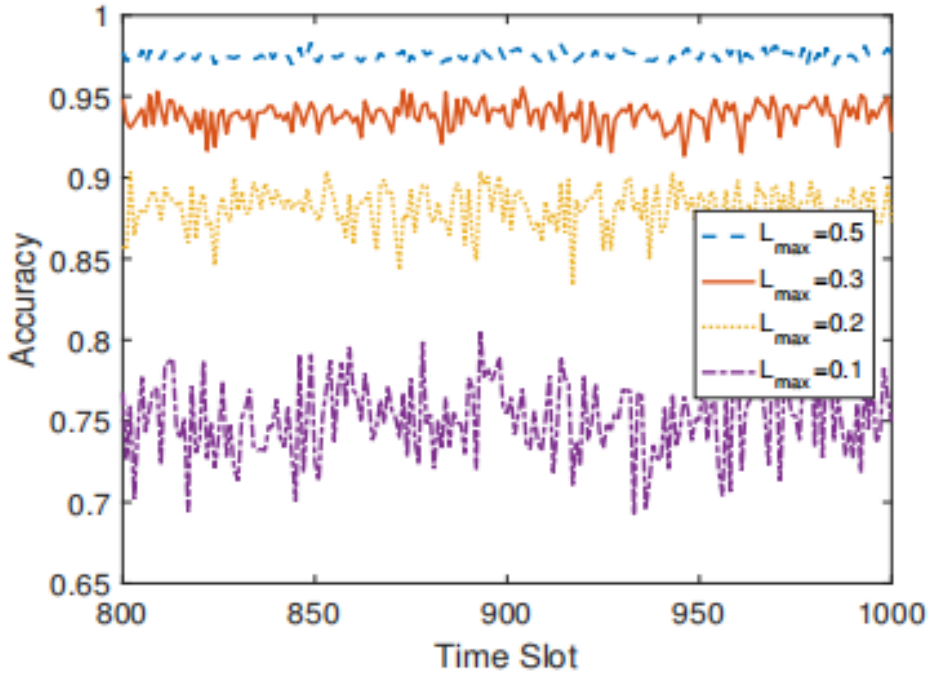
- Setting:
  - CNN models: 360p, 540p, 720p, 900p and 1080p

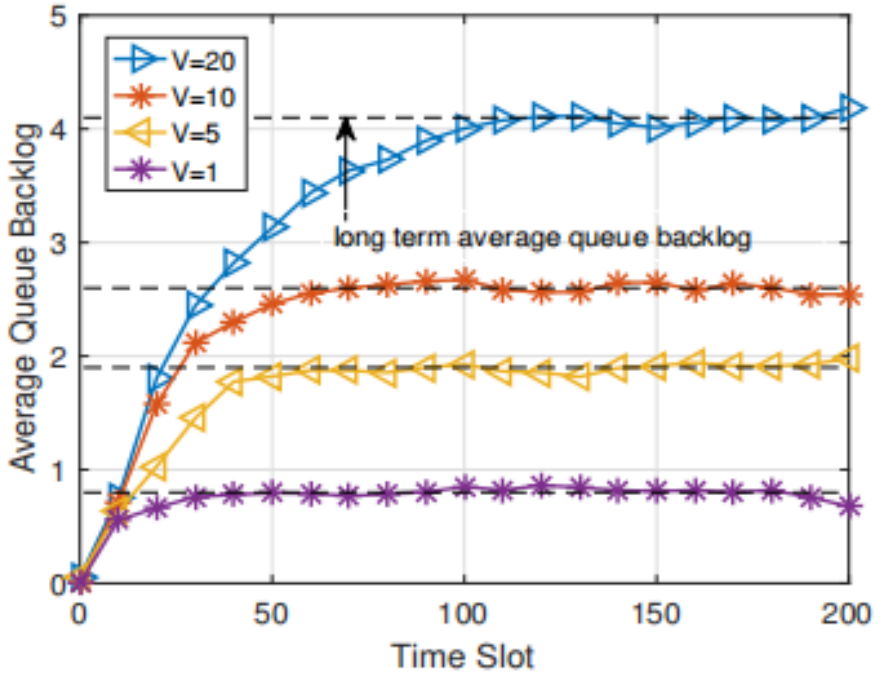


when available bandwidth decreases dramatically, and all video streams subsequently lower the resolution to reduce the bandwidth requirement.

Some evaluation

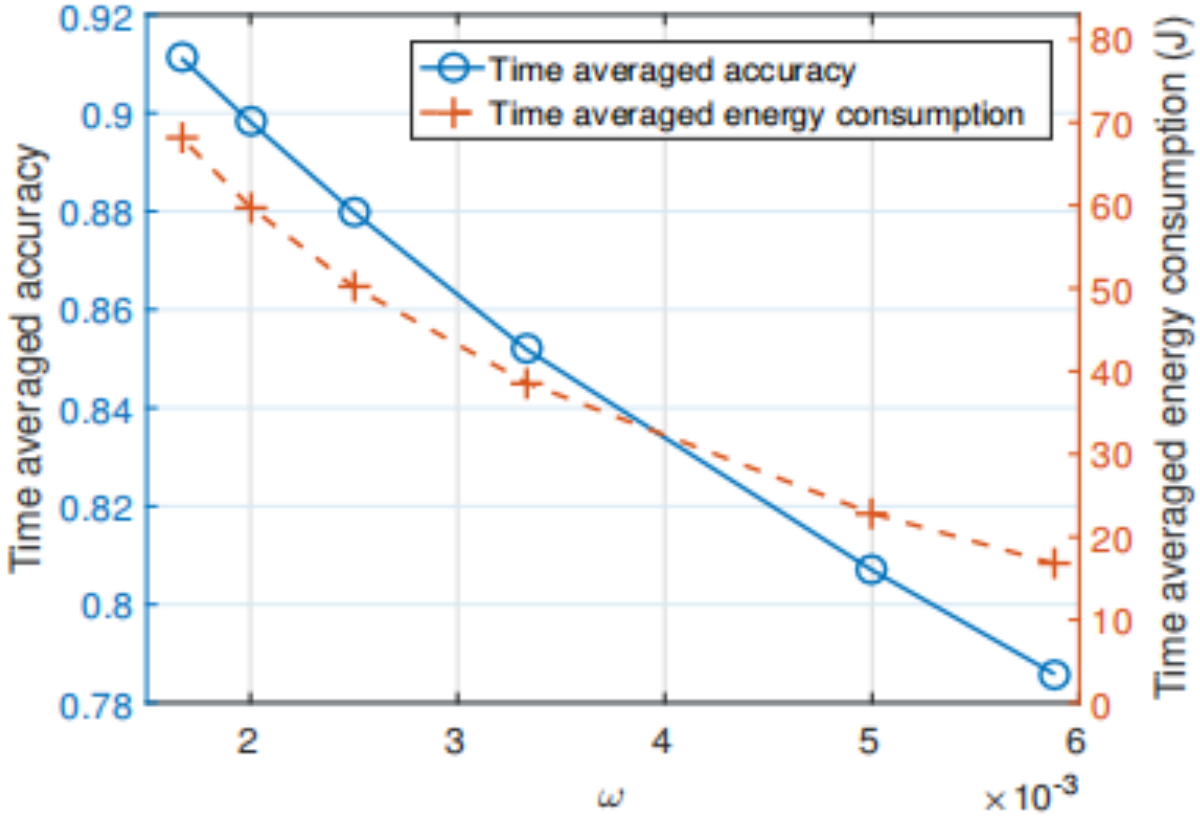
- L_{max} and V control the ***Latency-accuracy tradeoff***:





Some evaluation

- Accuracy-energy tradeoff:

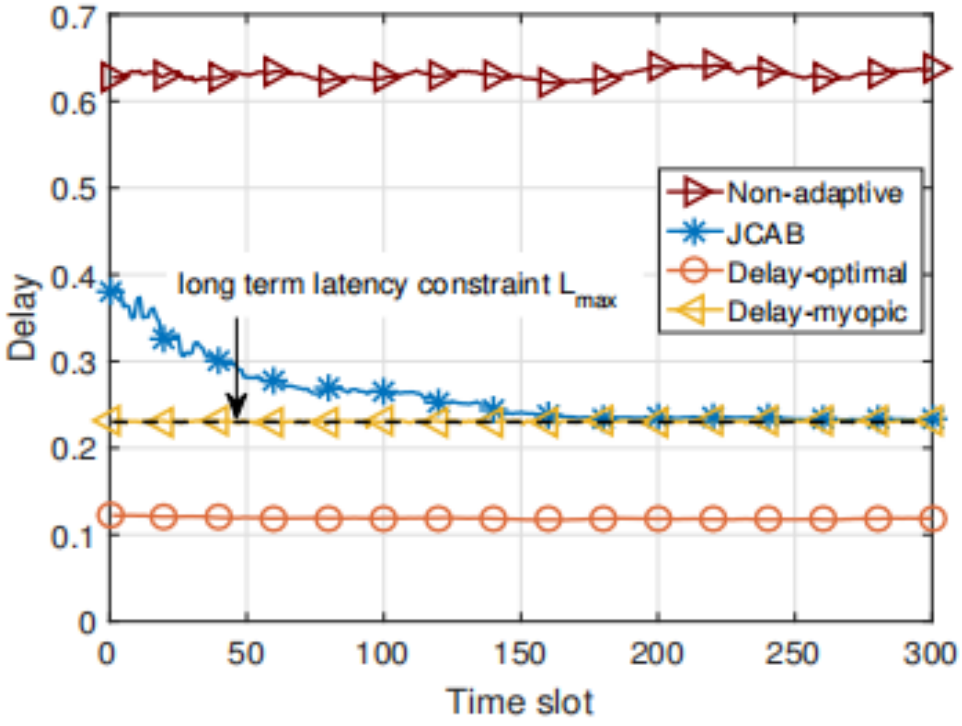


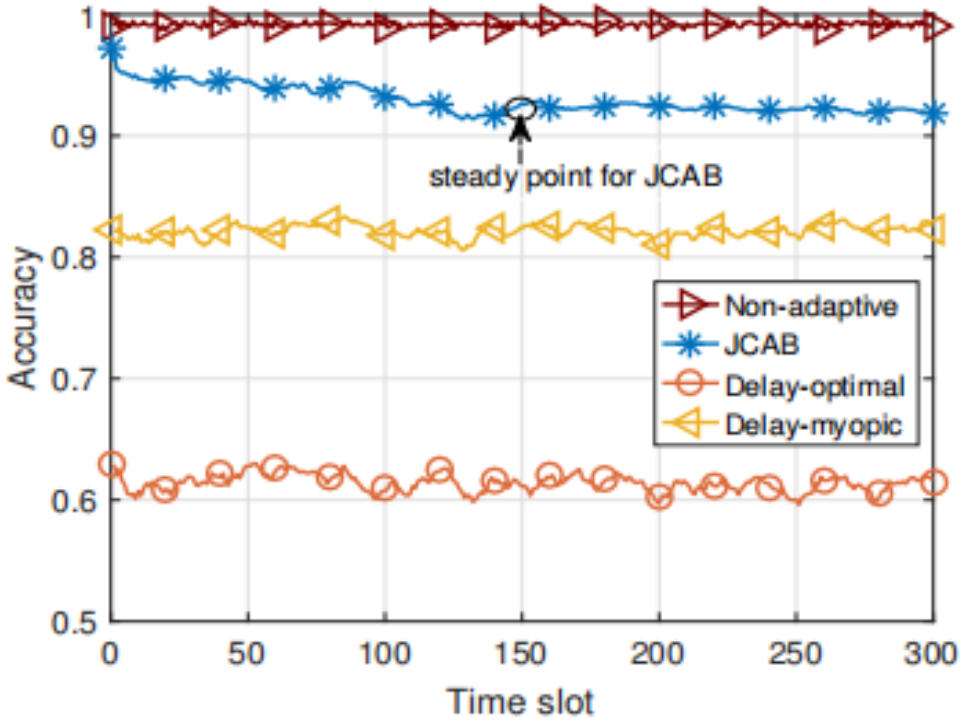
JCAB efficiently save energy consumption while maintaining a desirable accuracy

when increasing ω from 0.001 to 0.003, the algorithm gains up to 44% energy consumption reduction with only 4% loss of the analytics accuracy

Some evaluation

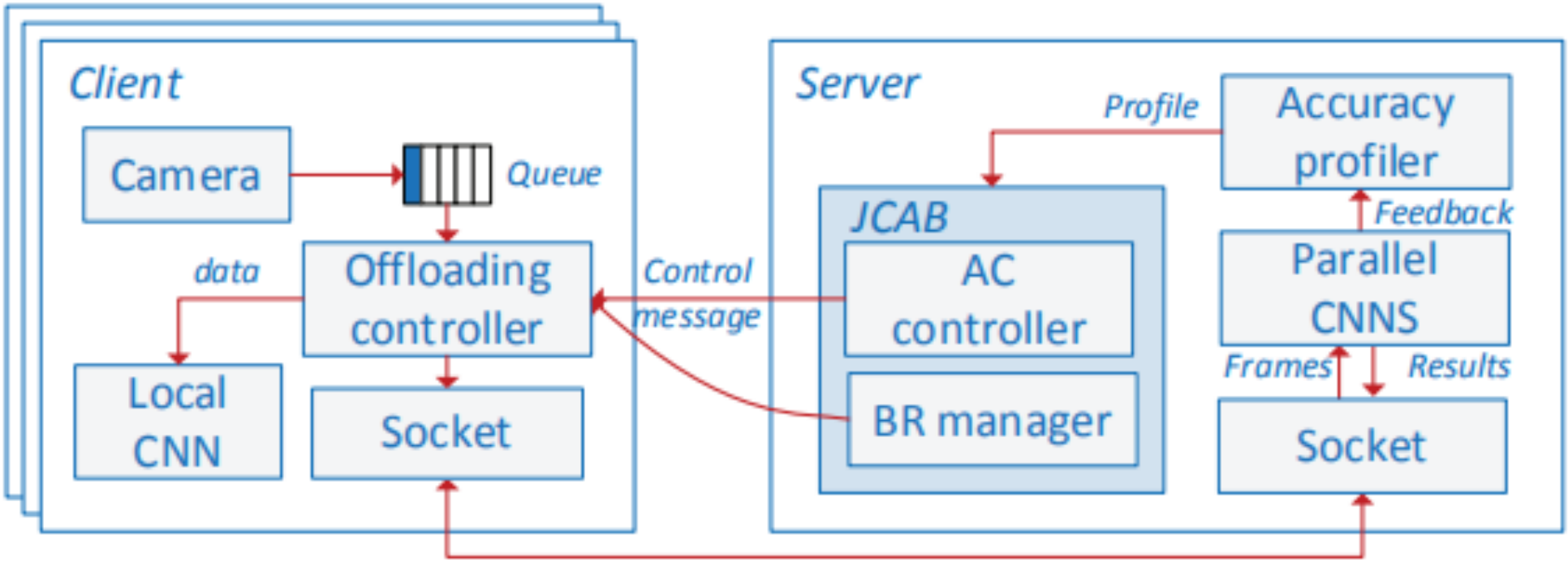
- Algorithm Comparison:





JCAB has a ***convergence process***, during which the algorithm gradually finds the optimal trade-off between latency and accuracy. Generally, ***JCAB achieves desirable average accuracy while closely following the long-term energy constraint***.

Solution Overview



- We present JCAB
  - focuses on ***configuration adaption*** and ***bandwidth allocation*** for multiple video streams
  - takes ***energy consumption***, ***system latency***, ***analytics accuracy*** into consideration.
  - ***works online*** without requiring future information
  - achieves a ***provable performance bound***

Thank you!
Q&A