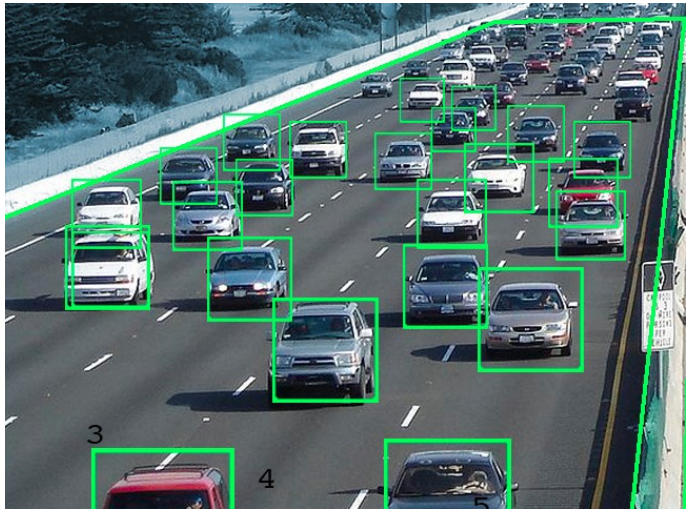
Enabling Scalable Edge Video Analytics with Computing-In-Network

Junchen Jiang

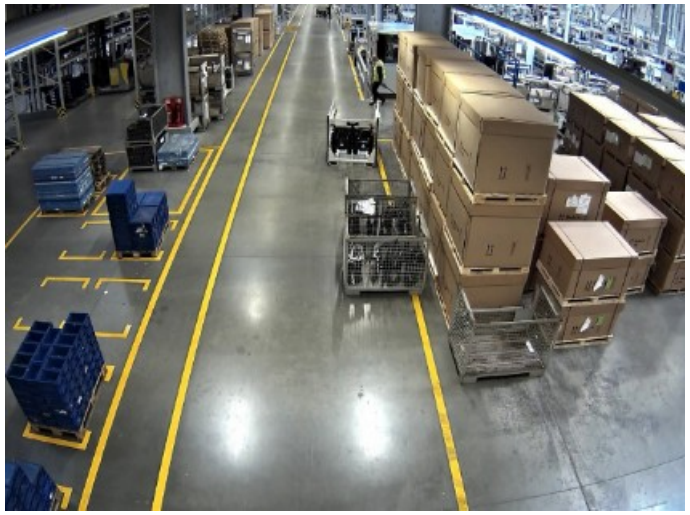
The University of Chicago

Cameras & video analysis apps are pervasive



Traffic control

Surveillance



Factory health/safety

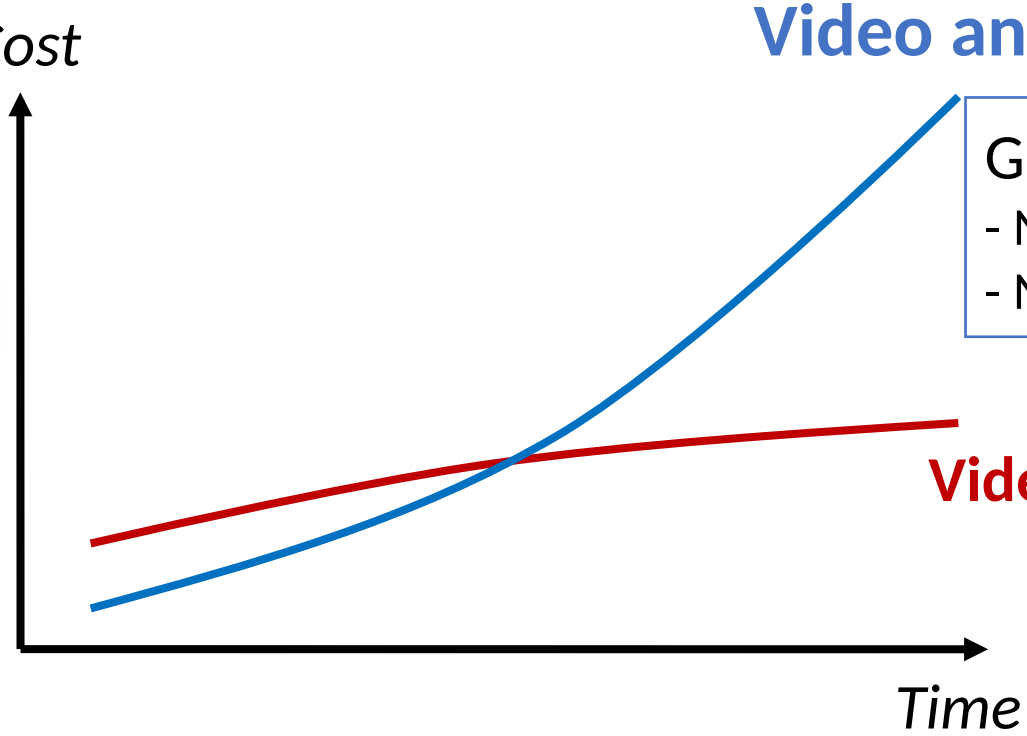
Goal: Enabling video analytics at scale

Video analytics can be **prohibitively expensive** at scale

Compute

Storage

Network



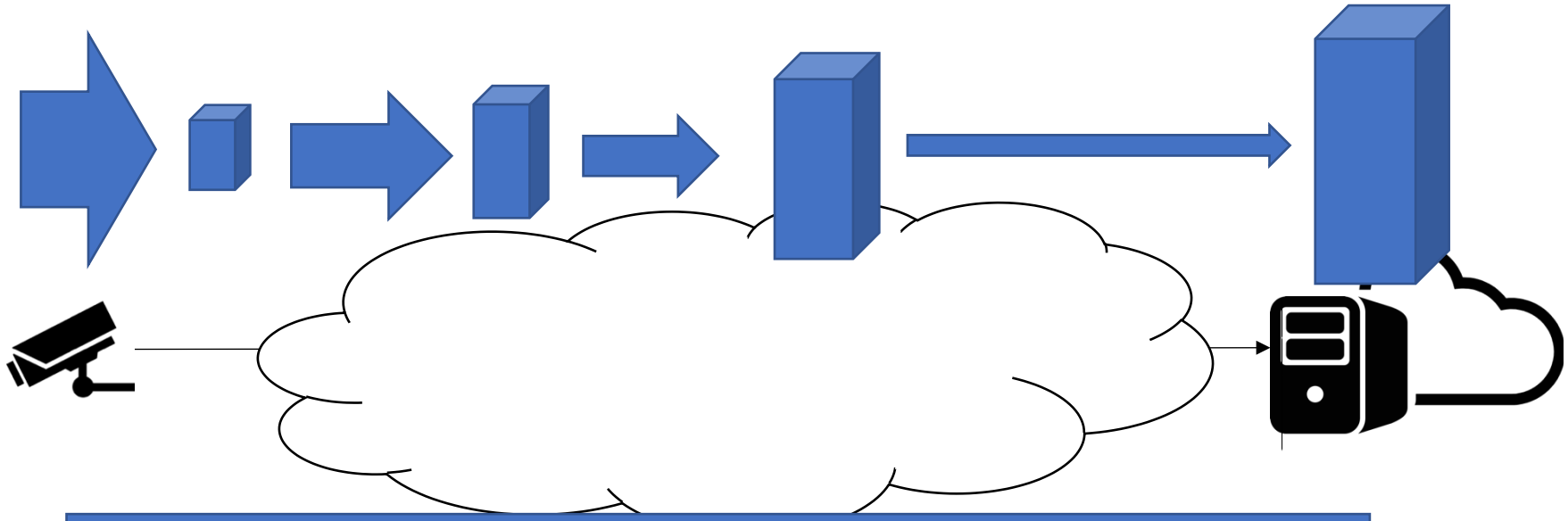
Growing faster than internet videos
- Need more resource than video streaming
- Not bounded by # of users

YouTube

NETFLIX

Today's Internet systems are built for traditional apps like video streaming, but unlikely to meet the need of video analytics.

The “Cloud-to-Edge Continuum” for Video Analytics

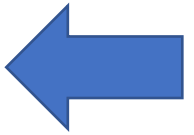


State-of-the-art: Cascading Pipelines

NoScope (VLDB’17), Glimpse (SenSys’15), FastCascading (CVPR’18), Chameleon (SIGCOMM’18), VideoStorm (NSDI’17), …

Two unique properties of video analytics

Video pipelines must be adaptive to real-time video content

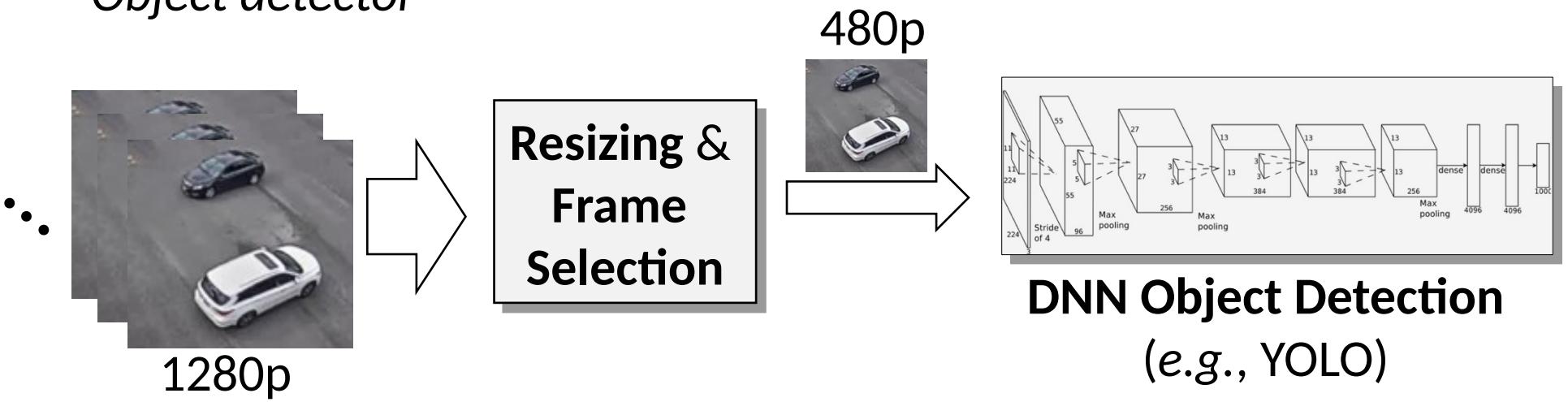


Leveraging real-time feedback from the consumer (DNN)

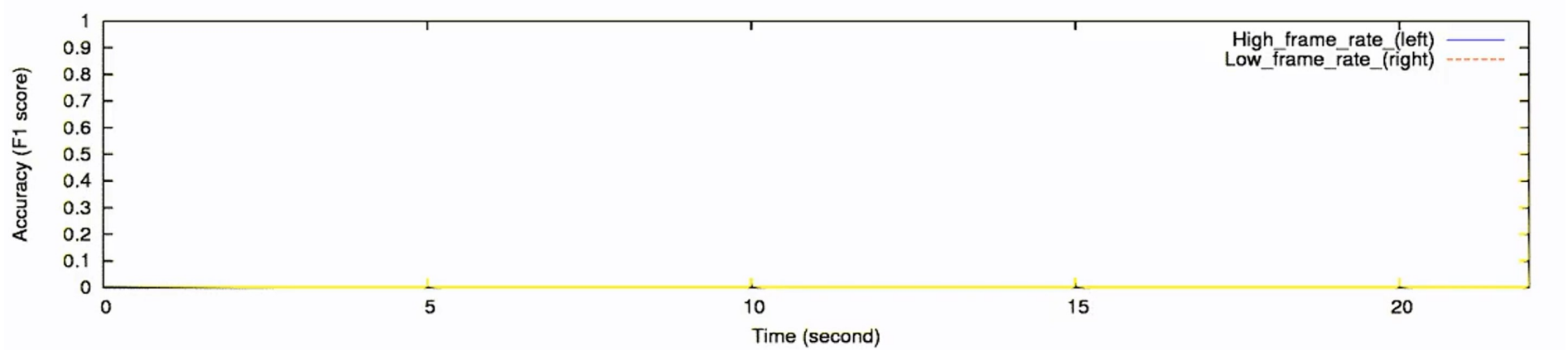
Prior work: Customize the video pipeline to the video content

Configurations:

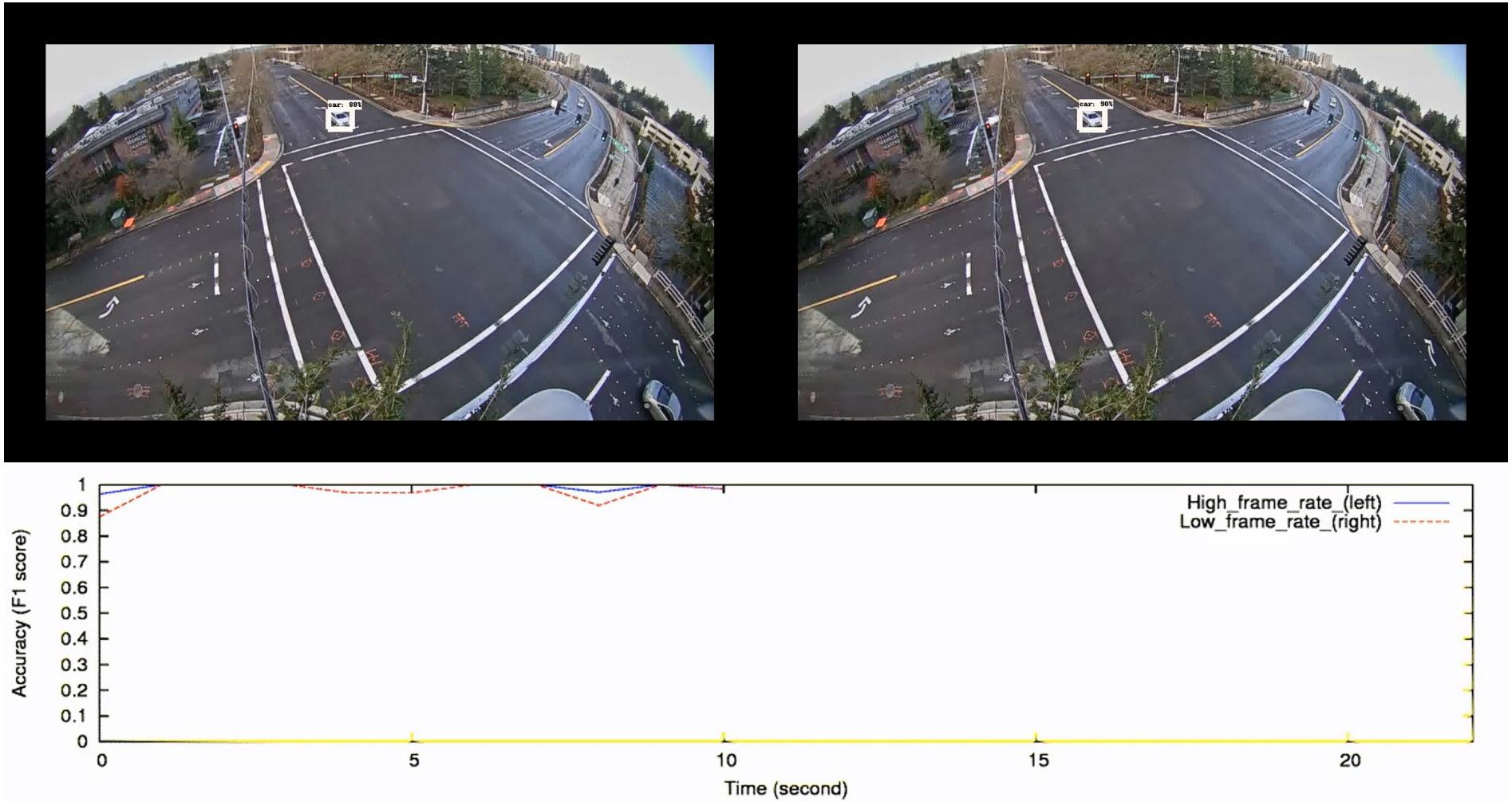
- *Resolution*
- *Frames rate*
- *Object detector*



Example: Lower frame rate



Problem: Best frame rate depends on content!



Key observation

Video content varies over time
$\not=$ best configuration varies over time

- Holds for other configuration knobs (resolution, NN classifier, etc.)
- Prior work does one-time profiling at beginning

Our approach: periodic reprofiling

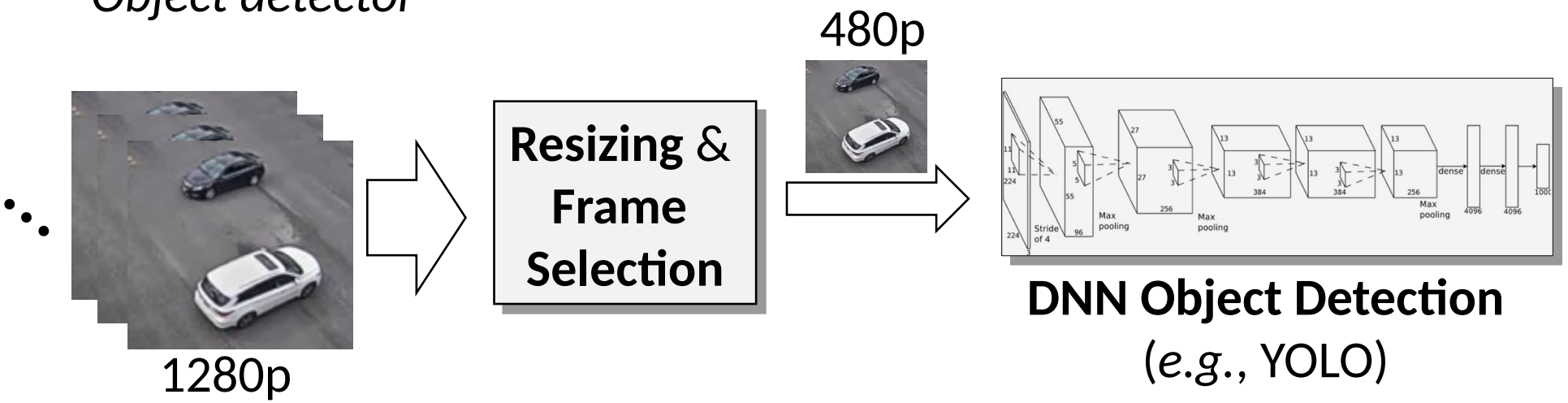
Adapt ~~Customize~~ the video pipeline to the dynamic video content

“Chameleon: Scalable Adaptation of Video Analytics via Temporal and Cross-camera Correlations” SIGCOMM 2018

Our approach: periodic reprofiling

Configurations:

- *Resolution*
- *Frames rate*
- *Object detector*

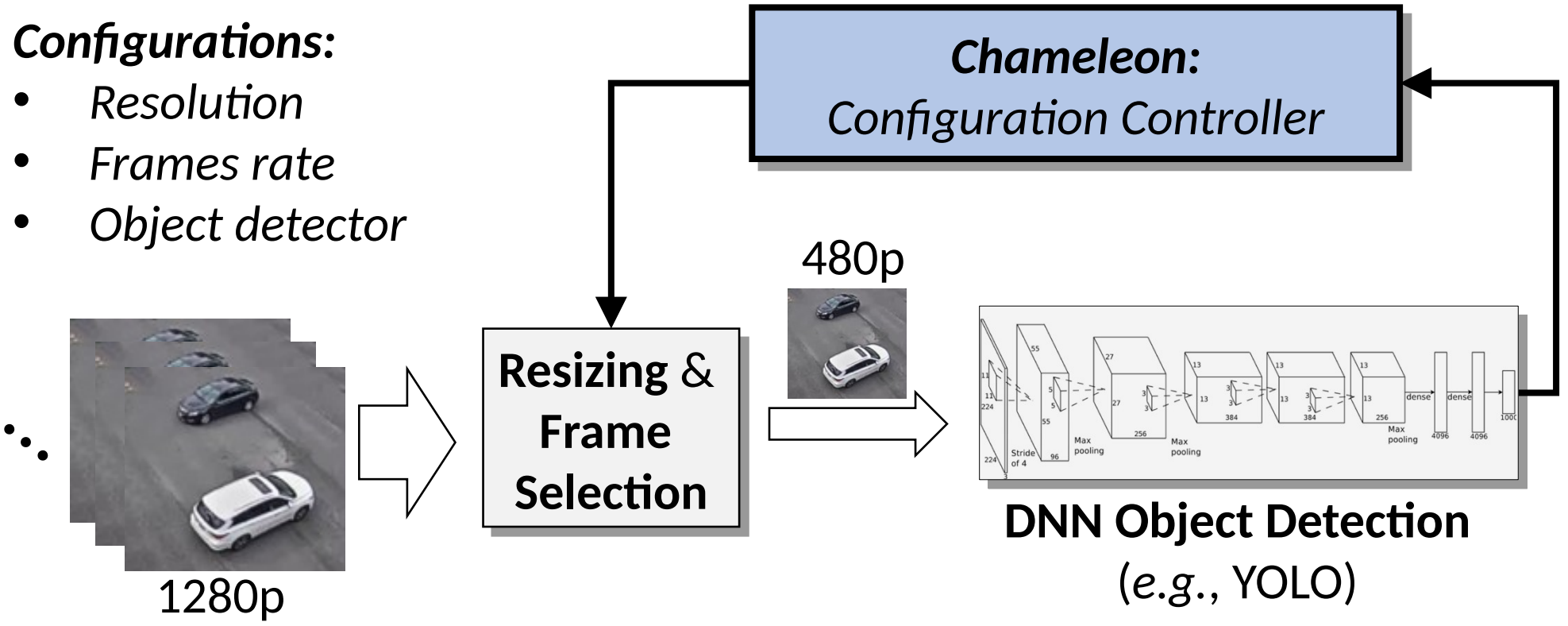


"Chameleon: Scalable Adaptation of Video Analytics via Temporal and Cross-camera Correlations" SIGCOMM 2018

Our approach: periodic reprofiling

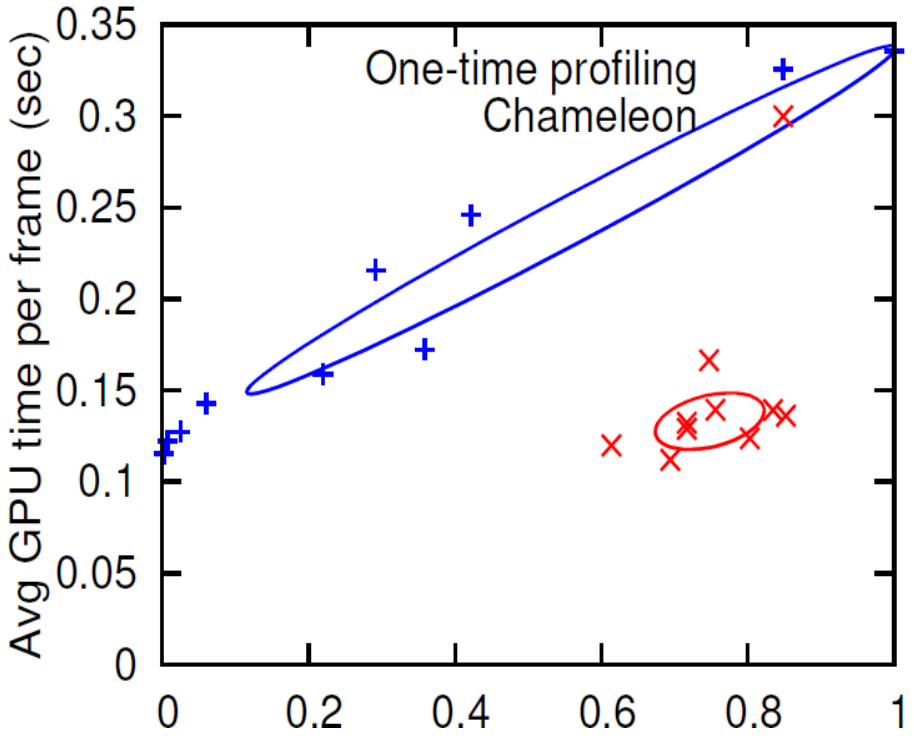
Configurations:

- *Resolution*
- *Frames rate*
- *Object detector*



"Chameleon: Scalable Adaptation of Video Analytics via Temporal and Cross-camera Correlations" SIGCOMM 2018

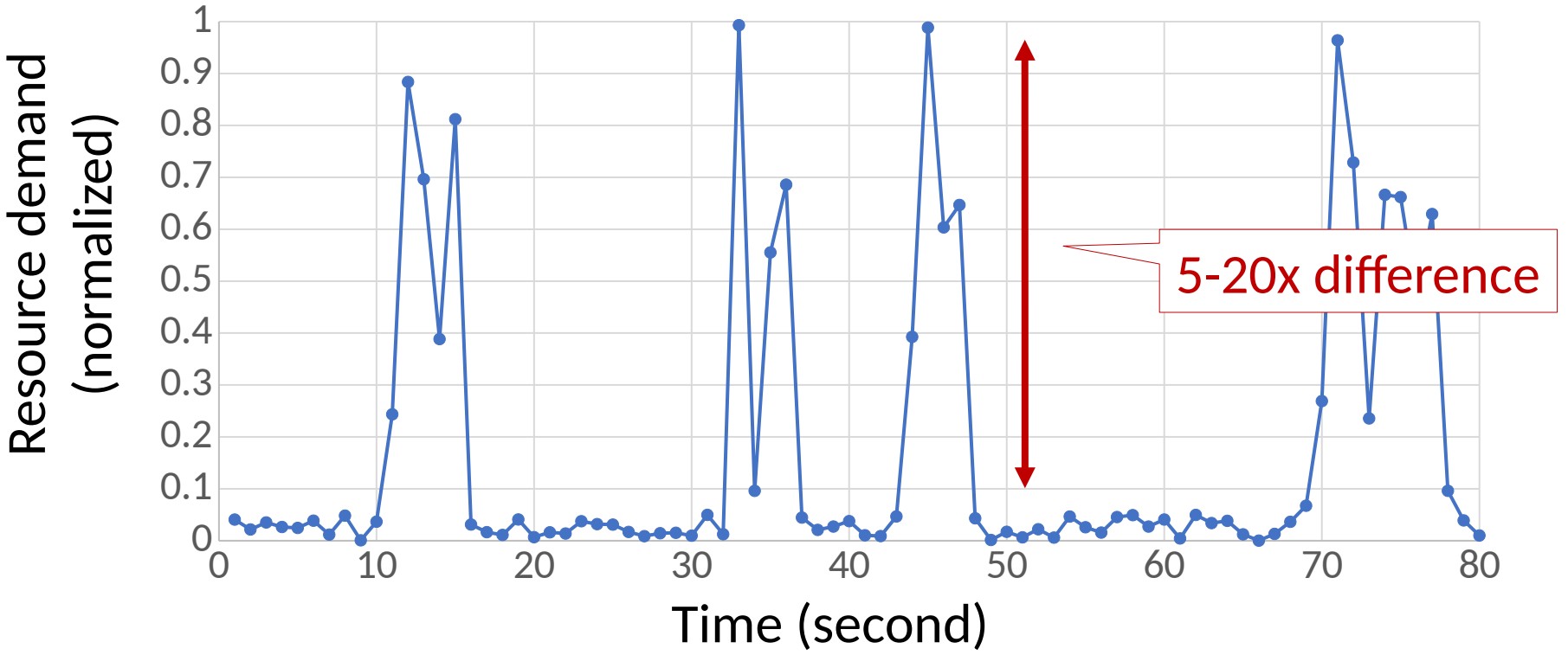
Evaluation: Chameleon improves accuracy + cost (traffic)



20-50% higher accuracy at same cost, or same accuracy at 30-50% of the cost (2-3× speedup)

"Chameleon: Scalable Adaptation of Video Analytics via Temporal and Cross-camera Correlations" SIGCOMM 2018

Implication: Spiky resource demand

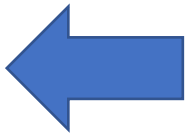


Takeaway:
In-network resource allocation must cope with spiky workload

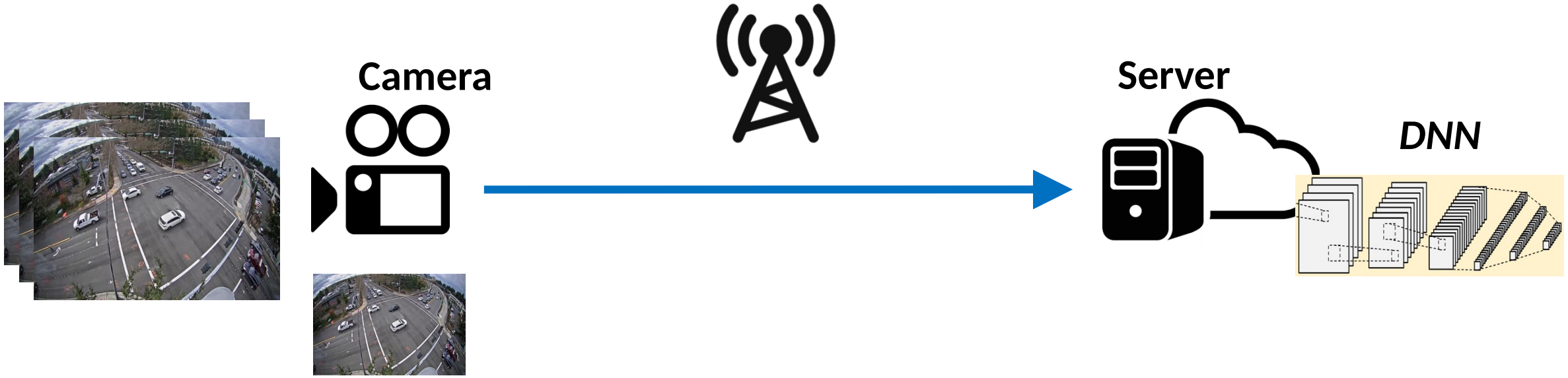
Two unique properties of video analytics

Pipelines must be adaptive to real-time video content

Leveraging real-time feedback from the analytics logic



Prior solution: Videos are sent by traditional video stack

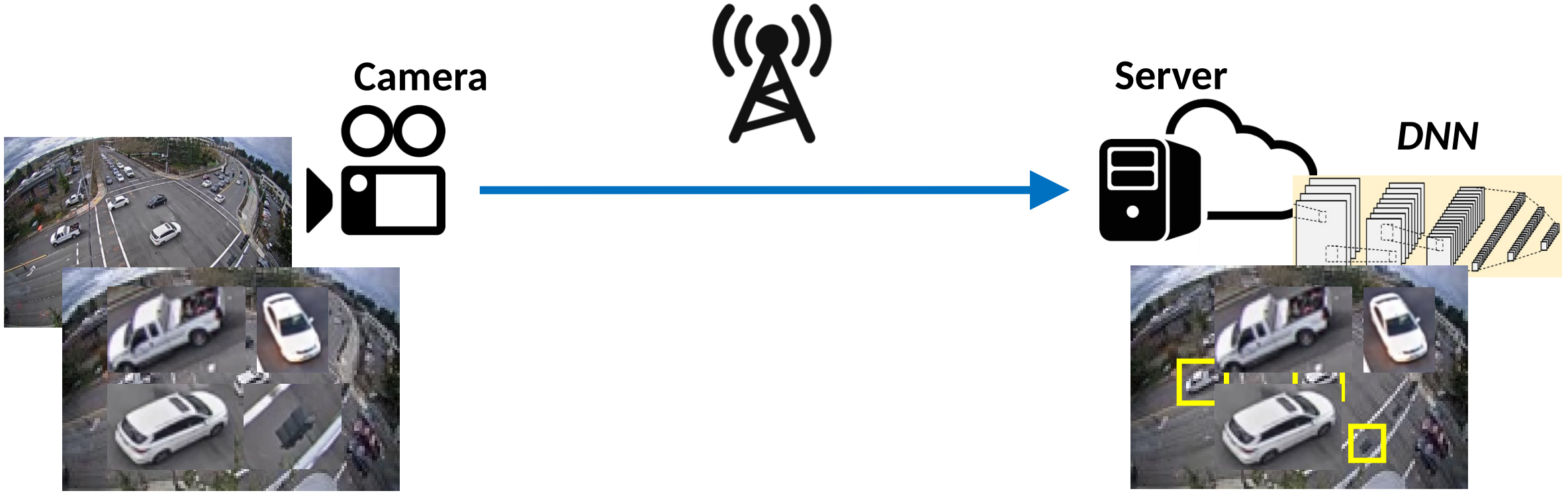


Suboptimal bandwidth-accuracy tradeoffs

Low quality ↓ Low accuracy

High quality ↓ Insufficient bandwidth

Our approach: Drive video streaming by server-side logic



This approach can save 2-5x bandwidth compared to client-side compression

"Reinventing Video Streaming for Distributed Vision Analytics" HotCloud 2018

Takeaways

Pipelines must be adaptive to real-time video content

→ Computing-In-Network should cope with spiky workloads

Leveraging real-time feedback from the analytics logic

→ Many opportunities by bringing analytics goals to the control loop