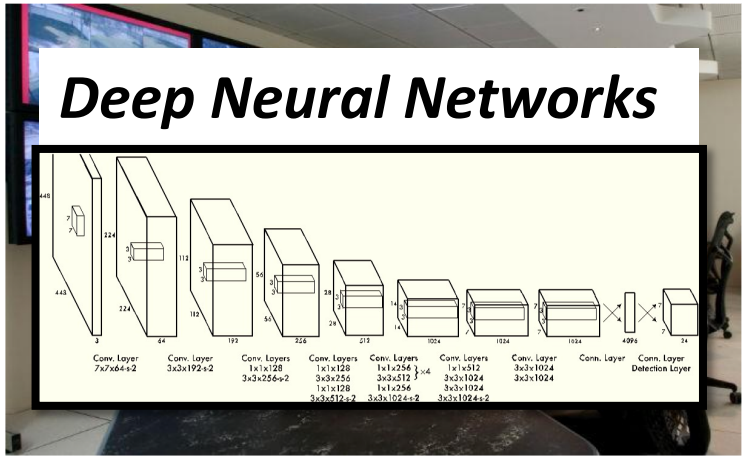
Reinventing Video Streaming for Distributed Vision Analytics

Chrisma Pakha, Aakanksha Chowdhery, Junchen Jiang

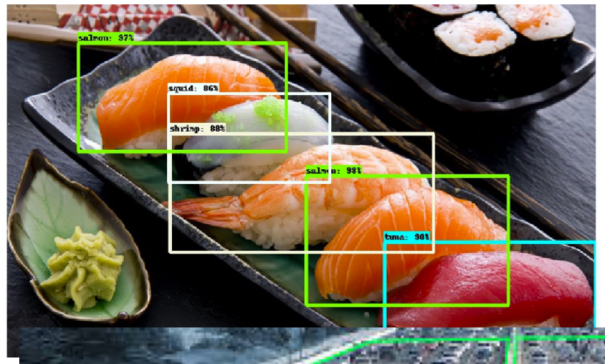


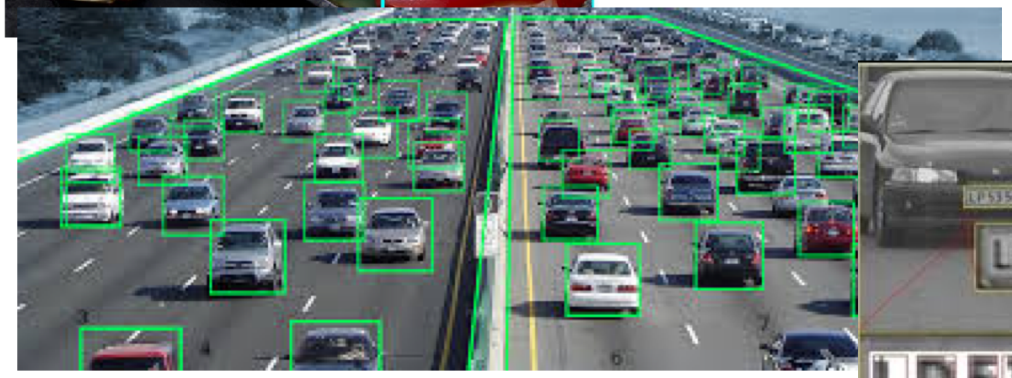


Automated video analytics everywhere

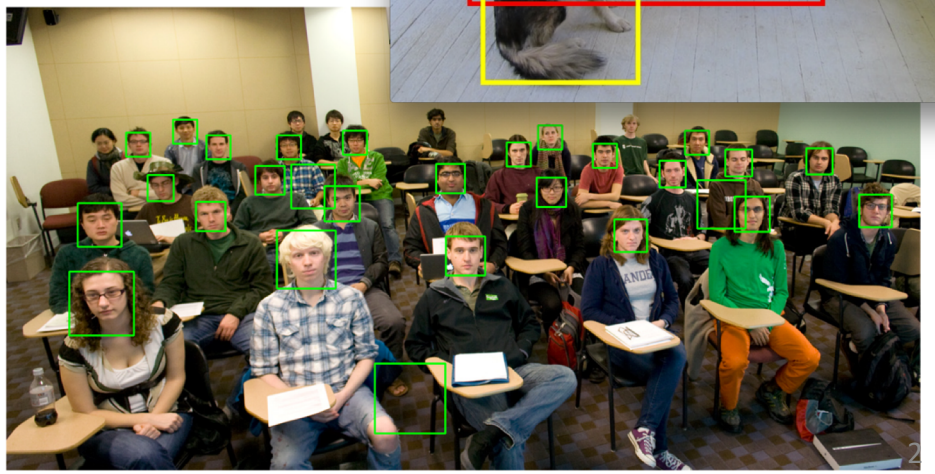


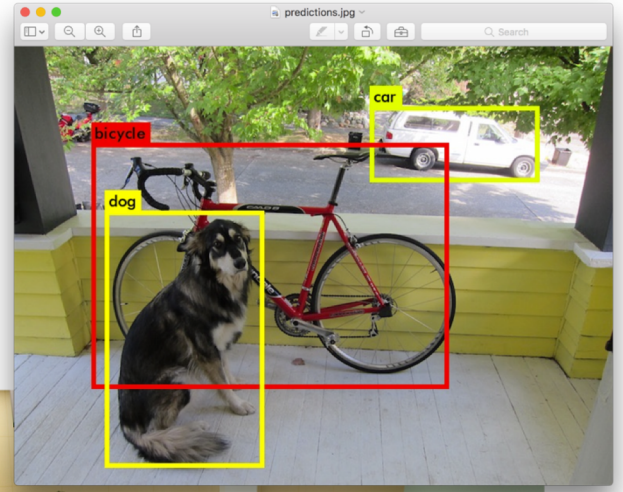
"find speeding cars on highway"
"detect an object of interest"





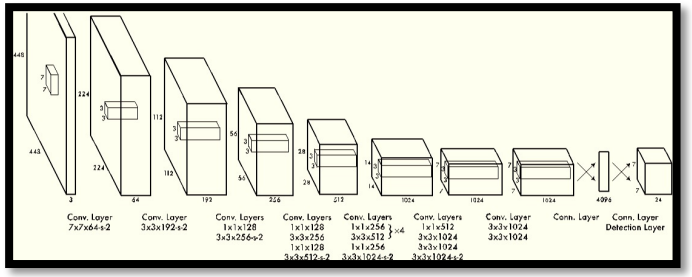




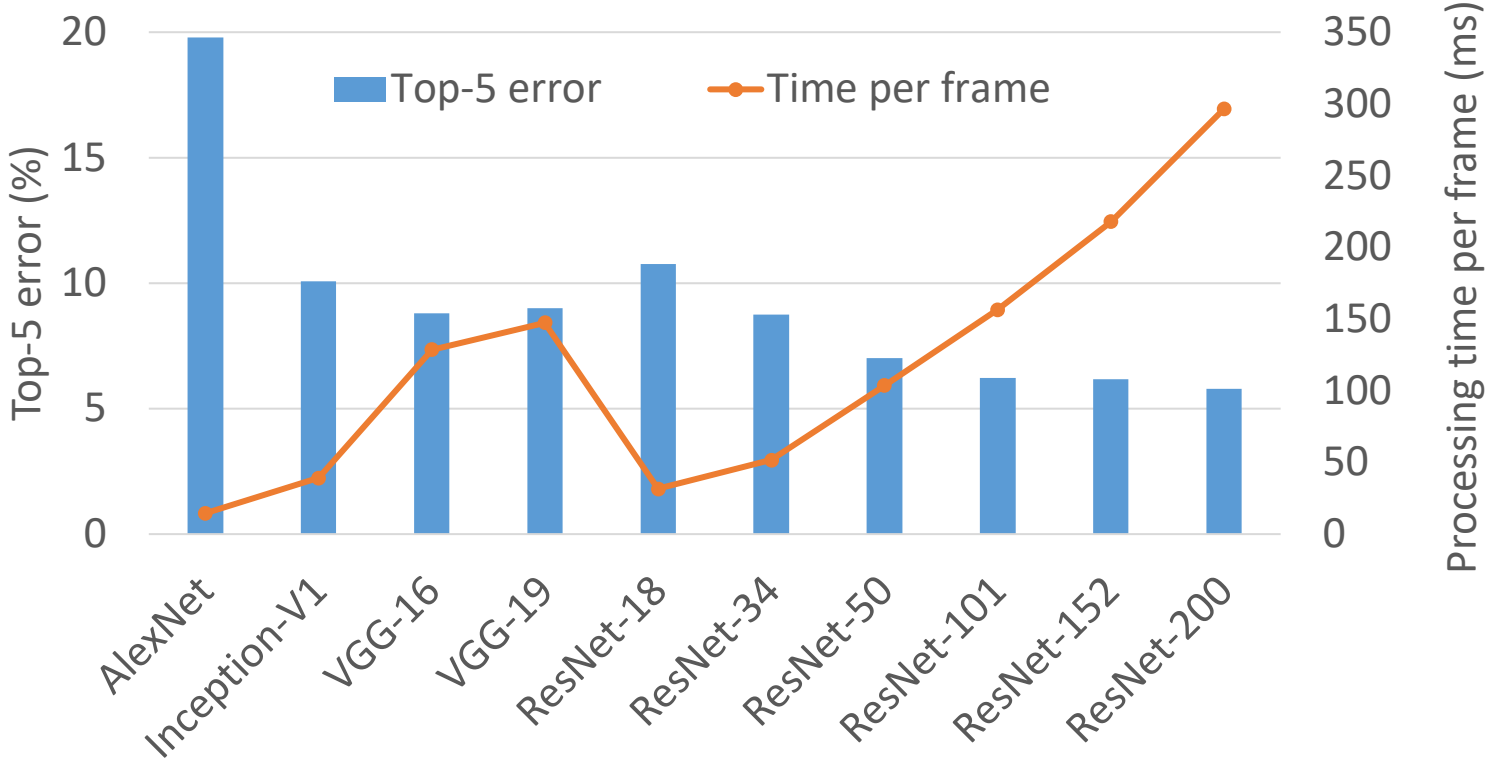


High accuracy at the expense of higher cost

Deep Neural Networks



Computationally expensive

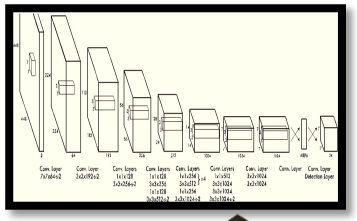


DNN + Video can be **prohibitively expensive** at scale

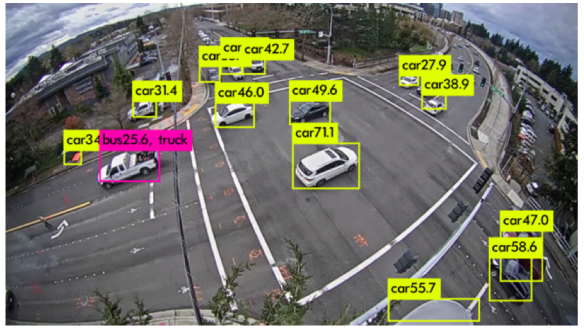
Processing one live feed can be costly











One camera of **30 fps**

Compute

(**$9,999** GPU or **$$$/mo cloud**)

Network



Storage



Distributed Vision Analytics



What if we have 1000x more cameras?

New York City has 4,176 cameras below 14th Street…



What if videos are 3840 × 2160 (4K)?

Most cameras today are only 1280 x 720

How to achieve **higher** accuracy *at* **lower** cost/bandwidth use?

Custom Video Streaming stack to explicitly balance bandwidth-accuracy tradeoffs

Video Streaming Stack: Bandwidth vs. Accuracy



Need Custom Video Streaming:
higher accuracy *at*
lower cost/bandwidth use?

MPEG for Distributed Vision Analytics



MPEG video encoding
Maximizes user-perceived QoE

- Max resolution
- Min video re-buffering/stalling
- Min frames dropped

Agnostic to DNN-perceived QoE

Client-side filtering for Distributed Vision Analytics

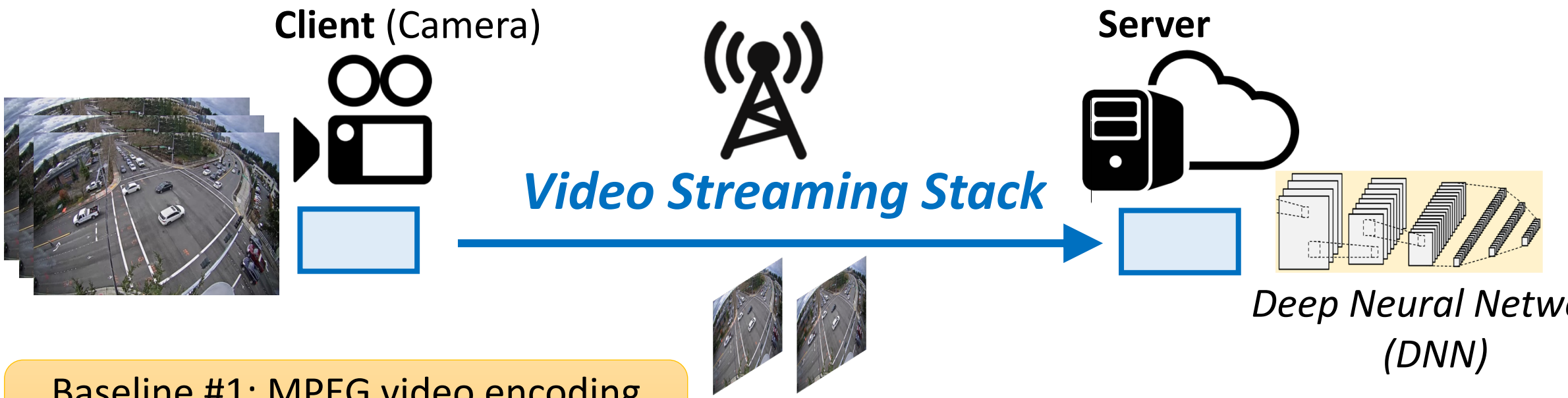


Client-side filtering
Less accurate, may miss critical details

- Heuristics: frame difference detection (Glimpse, NoScope)
- Less accurate vision models (Vigil)
- Specialized DNNs (MCDNN, NoScope)

Limited computational capability

State-of-the art approaches



Baseline #1: MPEG video encoding
Maximizes user-perceived QoE

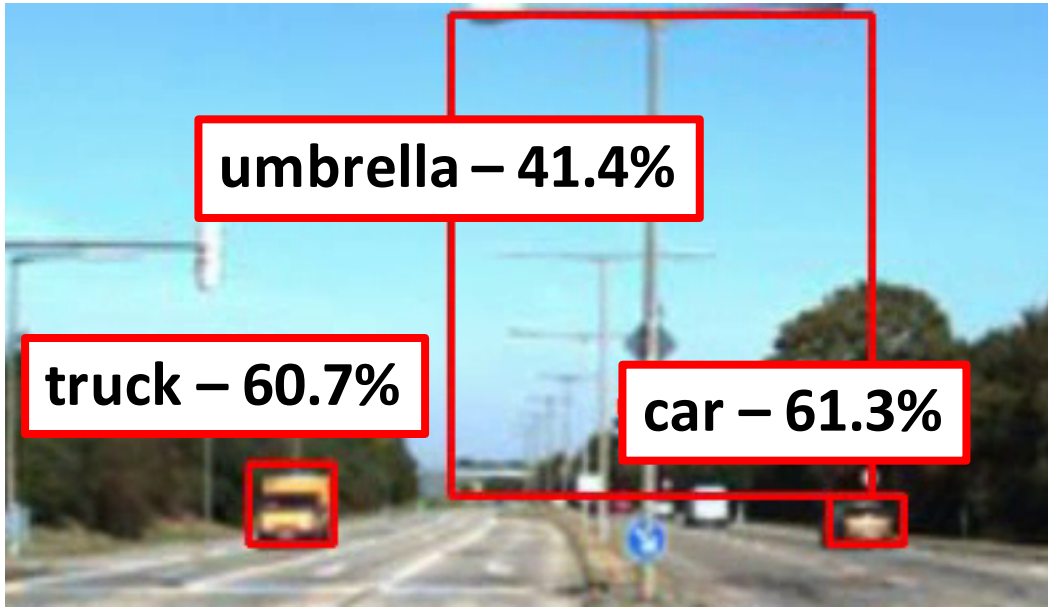
Baseline #2: Client-side filtering
Less accurate, may miss critical details

Both are agnostic to the DNN logic

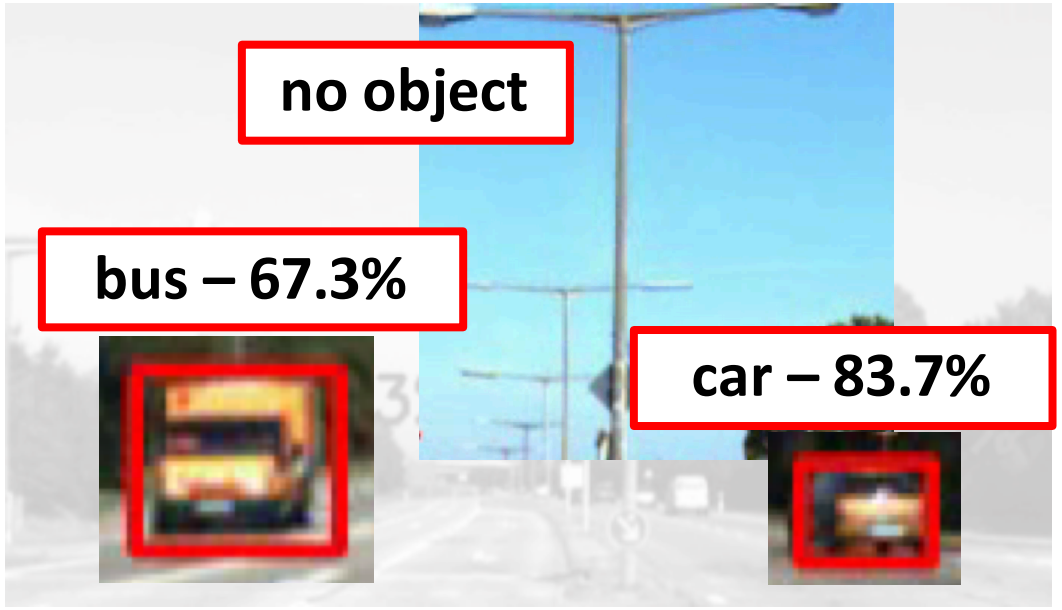
➔ **Suboptimal bandwidth-accuracy tradeoffs**

Design of **DNN-driven** Streaming

Whole image at low resolution



Cropped areas at high resolution



- Fetch video segments of interest

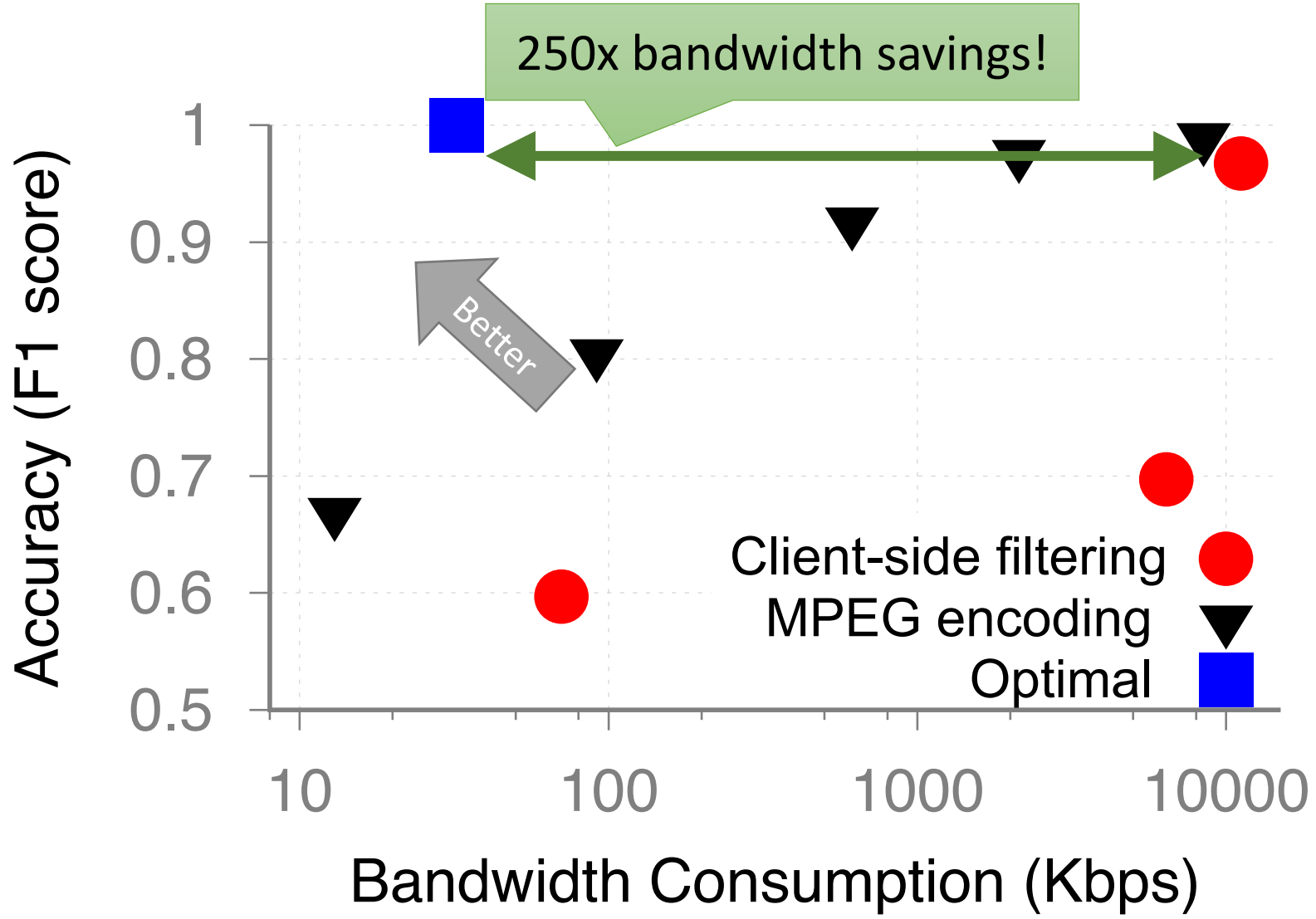
  ➔ Inference accuracy at **lower resolution**: *likely objects*

Design of **DNN-driven** Streaming

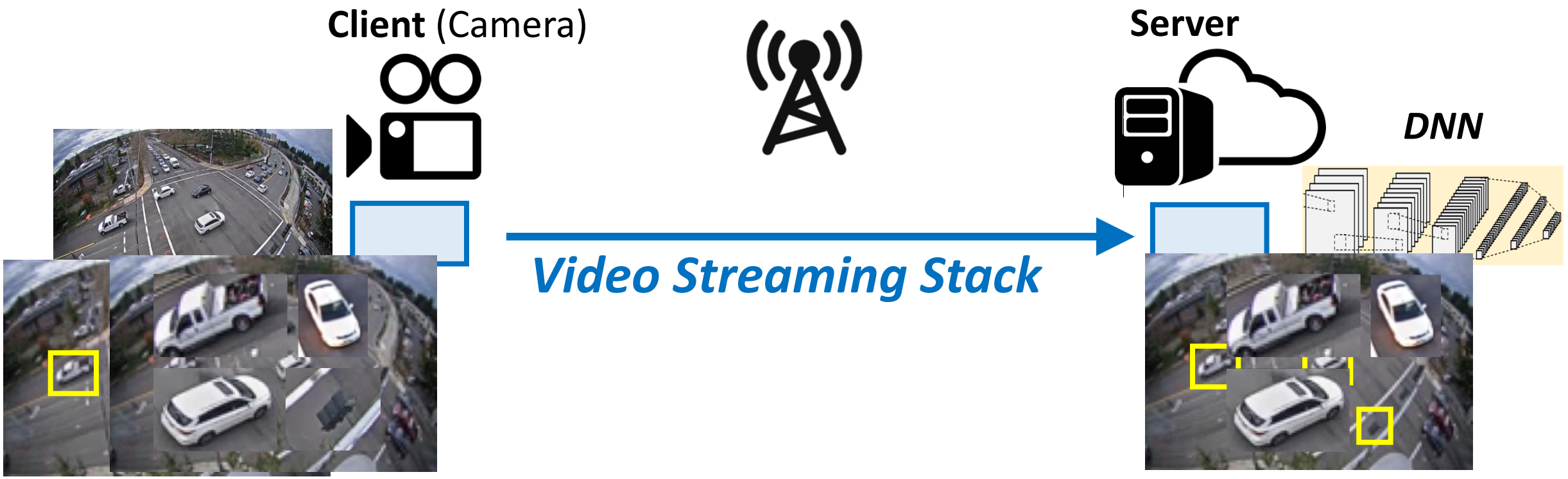


- Fetch video segments of interest
  - ➔ Inference accuracy at **lower resolution**: *likely objects*
  - ➔ Inference accuracy in **sparsely sampled frames**: *likely object locations*
  - ➔ Focus on **region-of-interest** (*Cropping*)

DNN-driven Streaming : Optimal

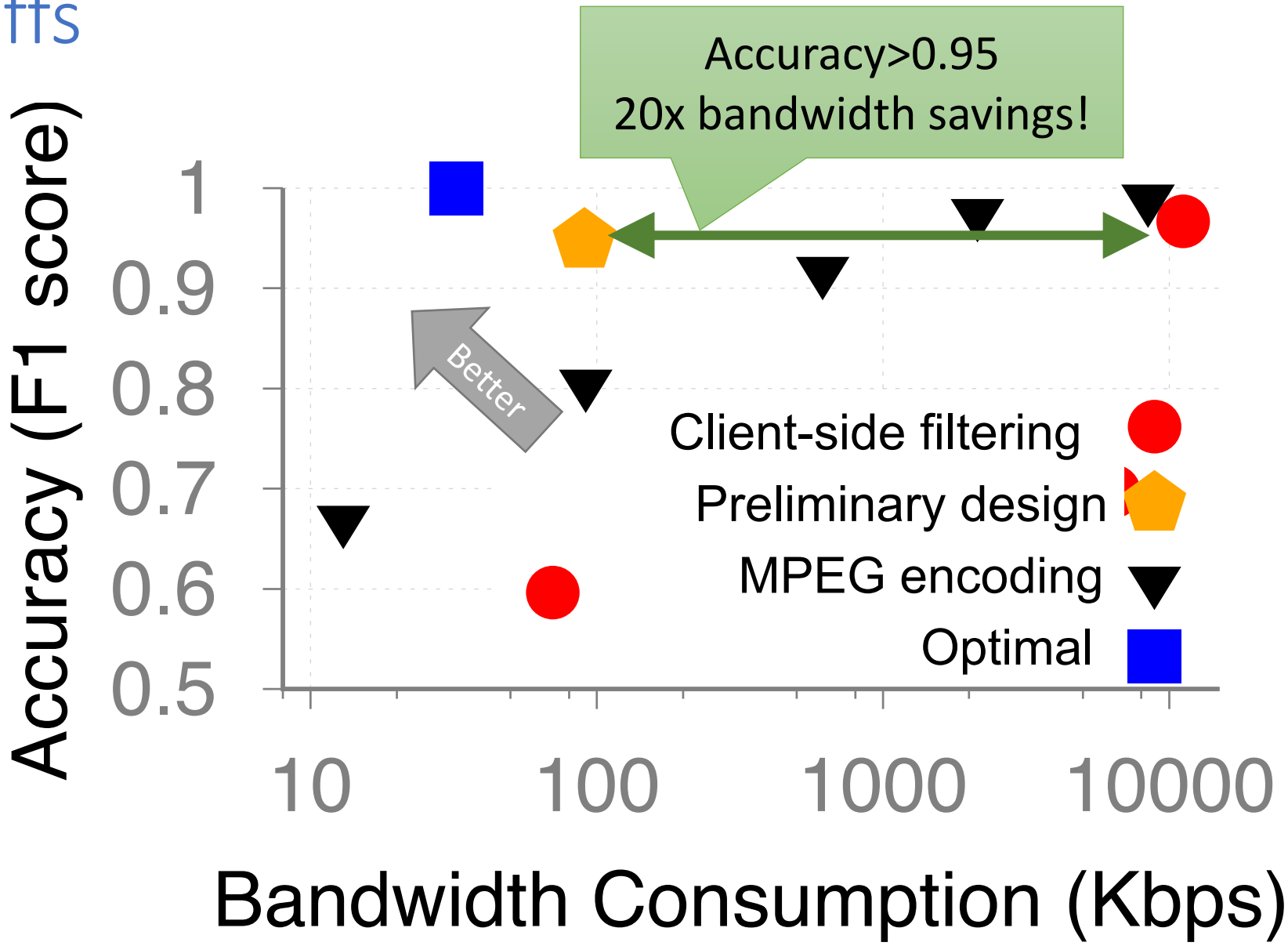


DNN-driven Streaming: Iterative Workflow

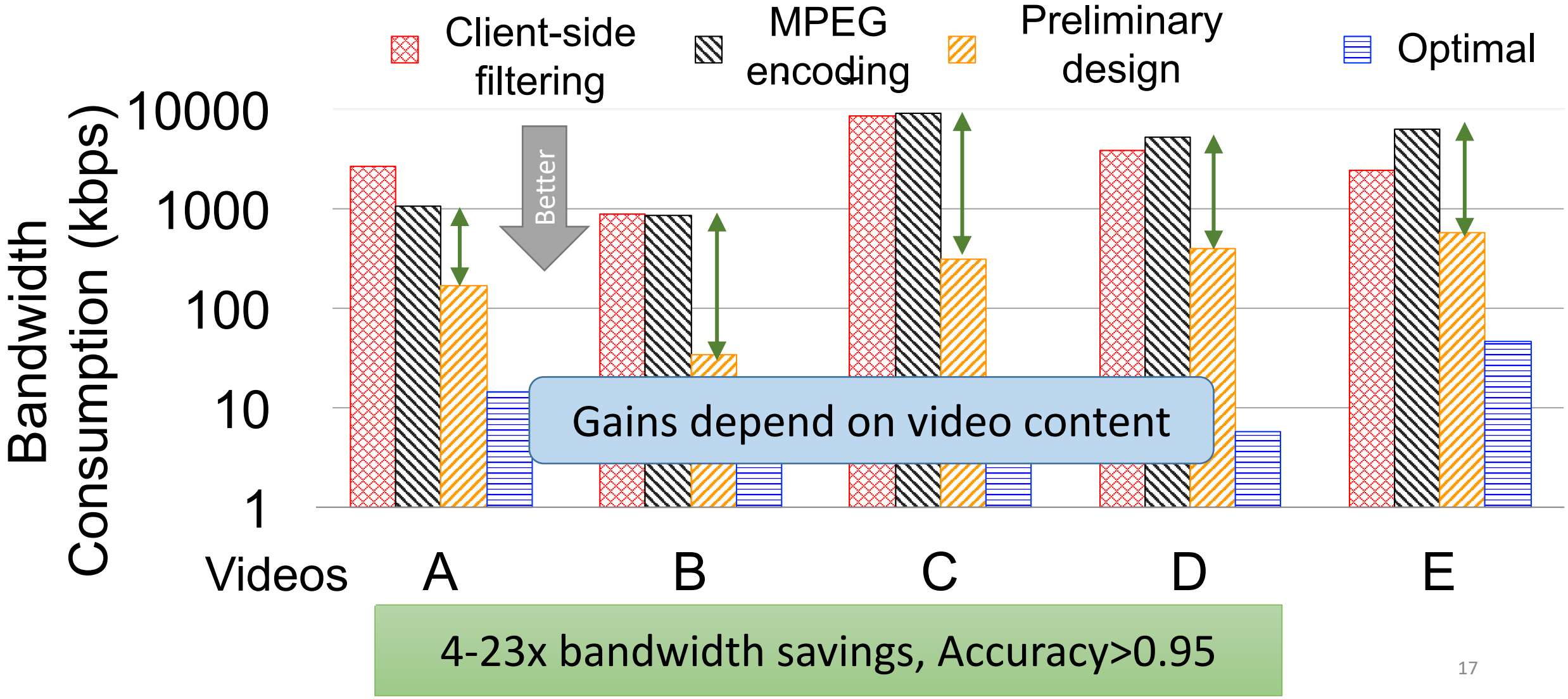


DNN-driven streaming explicitly balances between accuracy and bandwidth

Preliminary results achieve better bandwidth-accuracy tradeoffs



Preliminary results achieve better bandwidth-accuracy tradeoffs



Final remarks

- Accurate video analytics is increasingly needed!
  However, applying DNN on videos can be **prohibitively expensive** at scale

- *Better bandwidth-accuracy tradeoff* by custom video streaming stack
  Key insight: Streaming stack should be driven by the DNN logic

- Promising *order-of-magnitude* bandwidth savings!
  Several practical challenges remain…