北京郵電大學
Beijing University of Posts and Telecommunications

AI Systems towards Mobile and Edge Devices

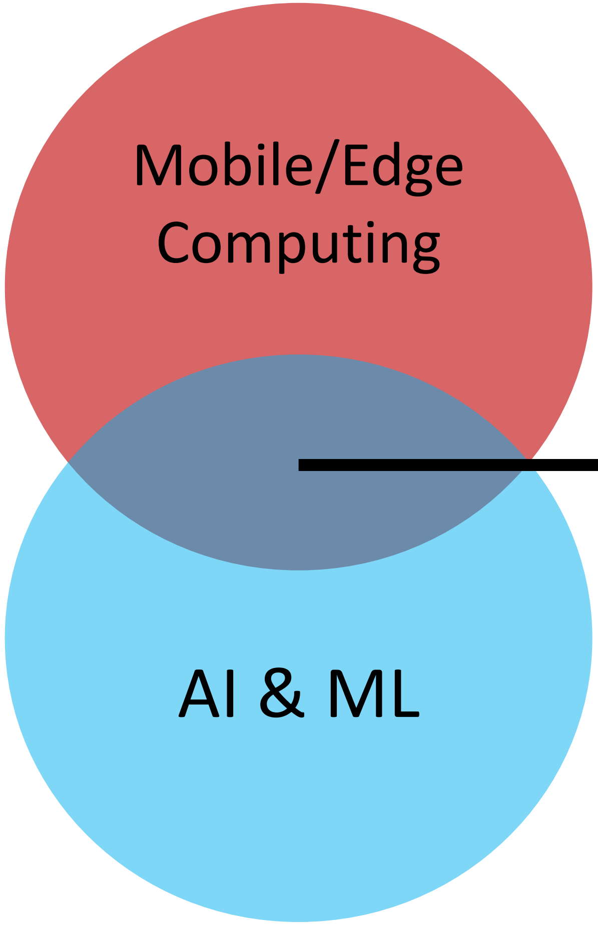
Mengwei Xu (徐梦炜)
Assistant Prof.
CS Dept @ BUPT

Outline

- Edge Intelligence: What and Why?
  - A system software perspective

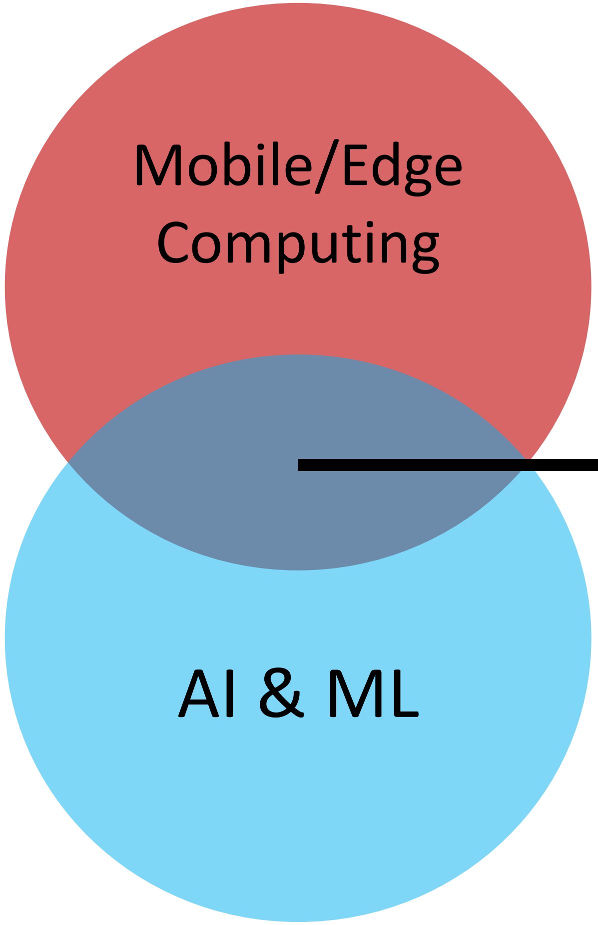
- Two pieces of my research on AIoT cameras
  - Zero-streaming Cameras (full paper under review, MobiCom'20 Demo)
  - Autonomous Cameras (MobiSys'20)

Edge + AI



- Timeliness
- Privacy
- Personality
- Scalability
- …

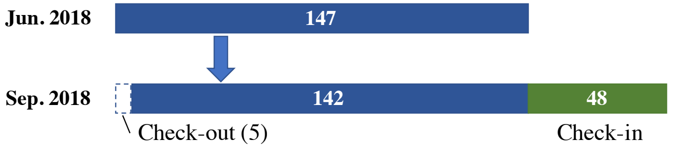
Edge + AI



- Timeliness
- Privacy
- Personality
- Scalability
- …

Some random evidences:

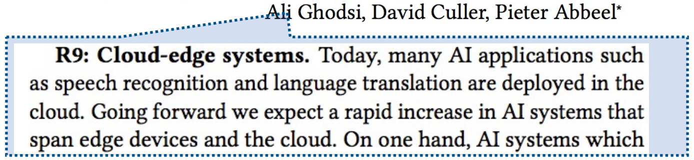
Edge AI is playing a critical role in our daily life



On Google Play, DL apps have increased by **27%** in 3rd quarter of 2018

- Statistics from our WWW'19 paper

A Berkeley View of Systems Challenges for AI

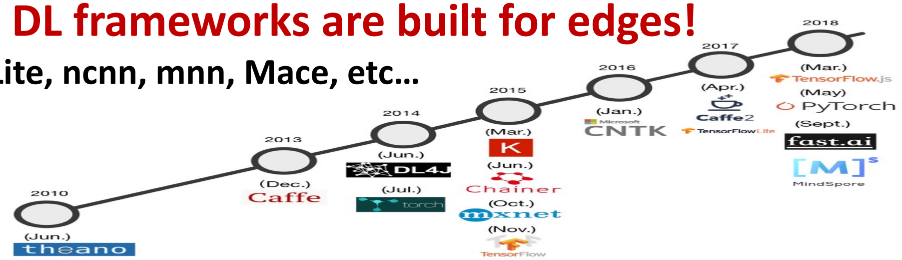


Ion Stoica, Dawn Song, Raluca Ada Popa, David Patterson, Michael W. Mahoney, Randy Katz, Anthony D. Joseph, Michael Jordan, Joseph M. Hellerstein, Joseph Gonzalez, Ken Goldberg, Ali Ghodsi, David Culler, Pieter Abbeel*

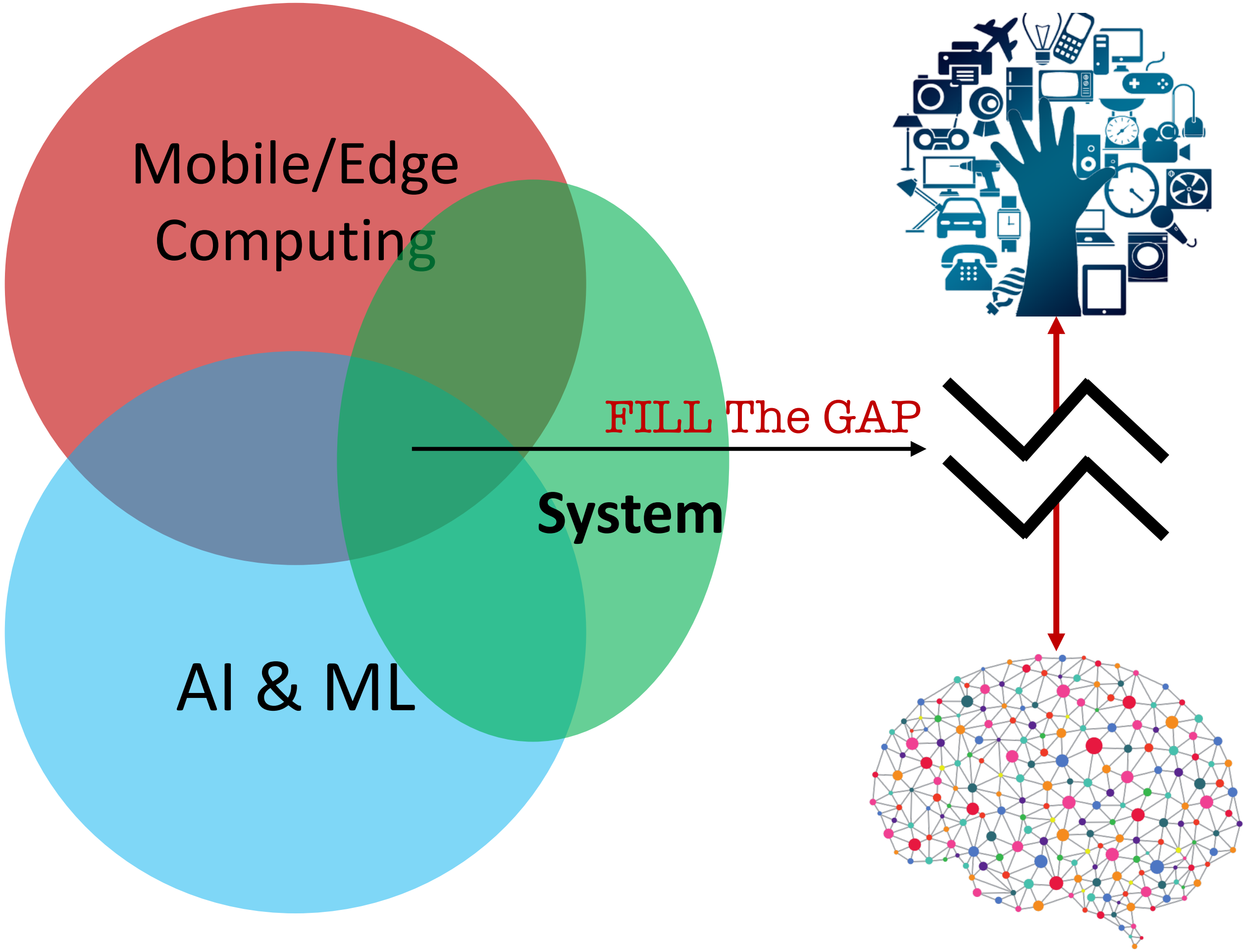
R9: Cloud-edge systems. Today, many AI applications such as speech recognition and language translation are deployed in the cloud. Going forward we expect a rapid increase in AI systems that span edge devices and the cloud. On one hand, AI systems which

More DL frameworks are built for edges!

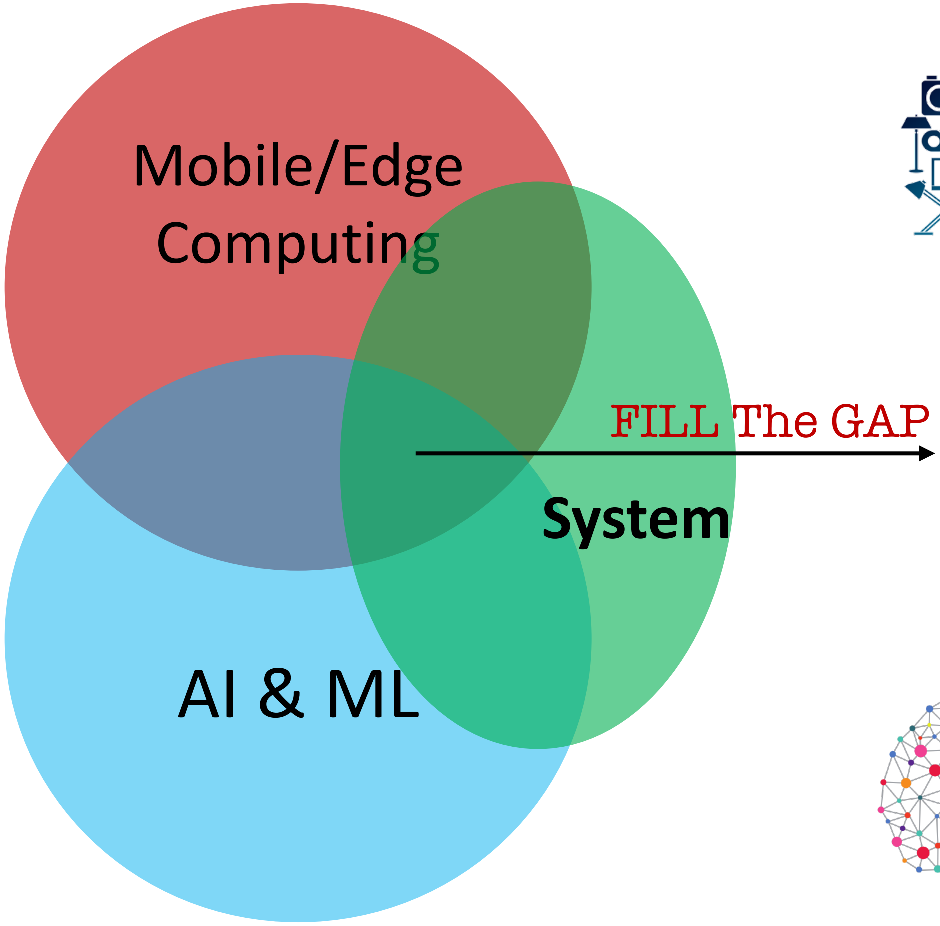
- **TFLite, ncnn, mnn, Mace, etc…**

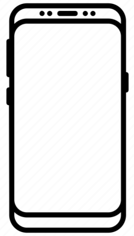


Edge + AI + System



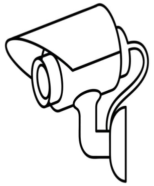
Edge + AI + System





WWW'19, MobiCom'18, UbiComp'18, TMC'17

TMC'19, EuroSys'18



MobiSys'20, arxiv' 1904.12342 (MobiCom'20 Demo)

Edge + AI + System

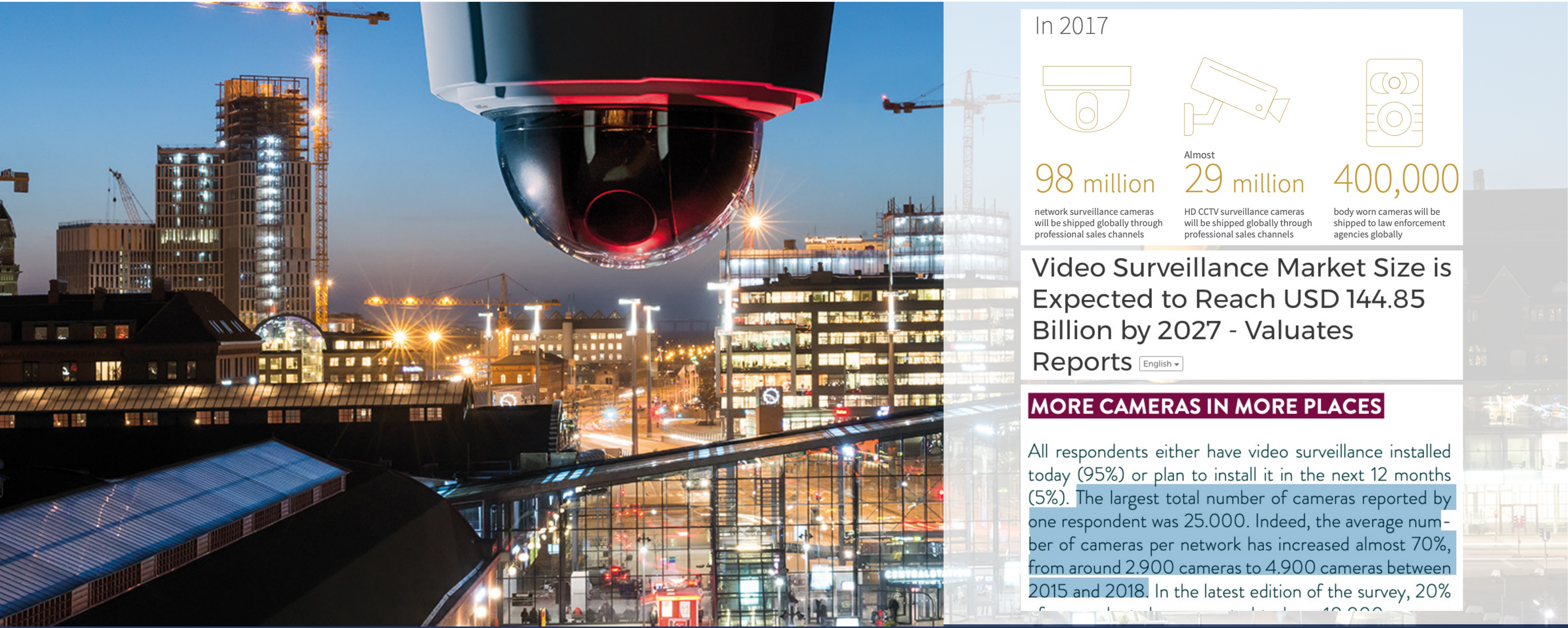
WWW’19, MobiCom’18, UbiComp’18, TMC’17

TMC’19, EuroSys’18

MobiSys’20, arxiv’ 1904.12342 (MobiCom’20 Demo)

AIoT Cameras: a key building block

AIoT Cameras: a key building block



AIoT Cameras: a key building block

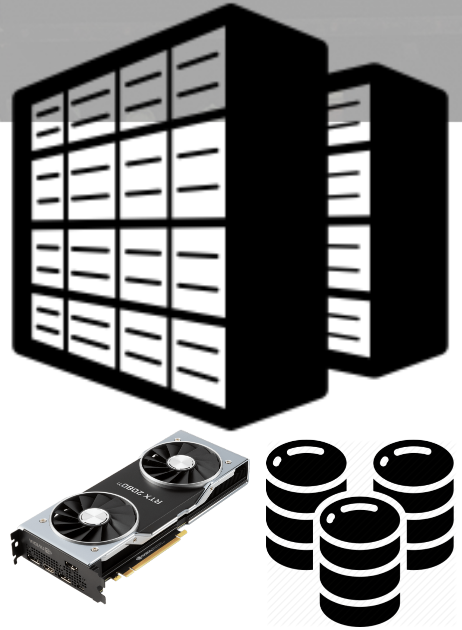
- Network
- Storage
- Compute
- Privacy

Traditional approach: cloud-centric paradigm

- [SIGCOMM’20] Reducto, [SOSP’19] Nexus, [OSDI’18] Focus, etc
- Cameras are just data sources or with dumb intelligence

AIoT Cameras: a key building block

- Network
- Storage
- Compute
- Privacy



Traditional approach: cloud-centric paradigm

- [SIGCOMM'20] Reducto, [SOSP'19] Nexus, [OSDI'18] Focus, etc
- Cameras are just data sources or with dumb intelligence

Our approach: camera-centric paradigm

Outline

- Edge Intelligence: What and Why?
  - A system software perspective

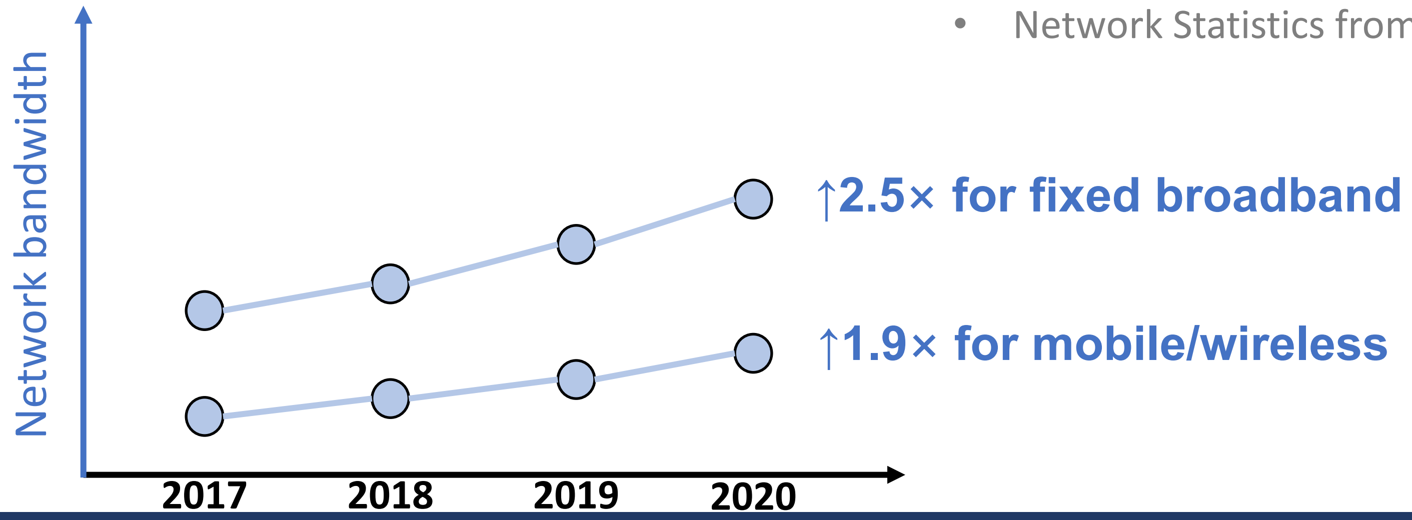
- Two pieces of my research on AIoT cameras
  - **Zero-streaming Cameras (full paper under review, MobiCom'20 Demo)**
  - Autonomous Cameras (MobiSys'20)

Zero-streaming Cameras

- Motivations
  1. Most videos are **cold**: a case study from PKU campus
     - ❑ More than 1,000 cameras deployed
     - ❑ Only <0.005% video and <2% cameras are eventually queried within 6 months

Zero-streaming Cameras

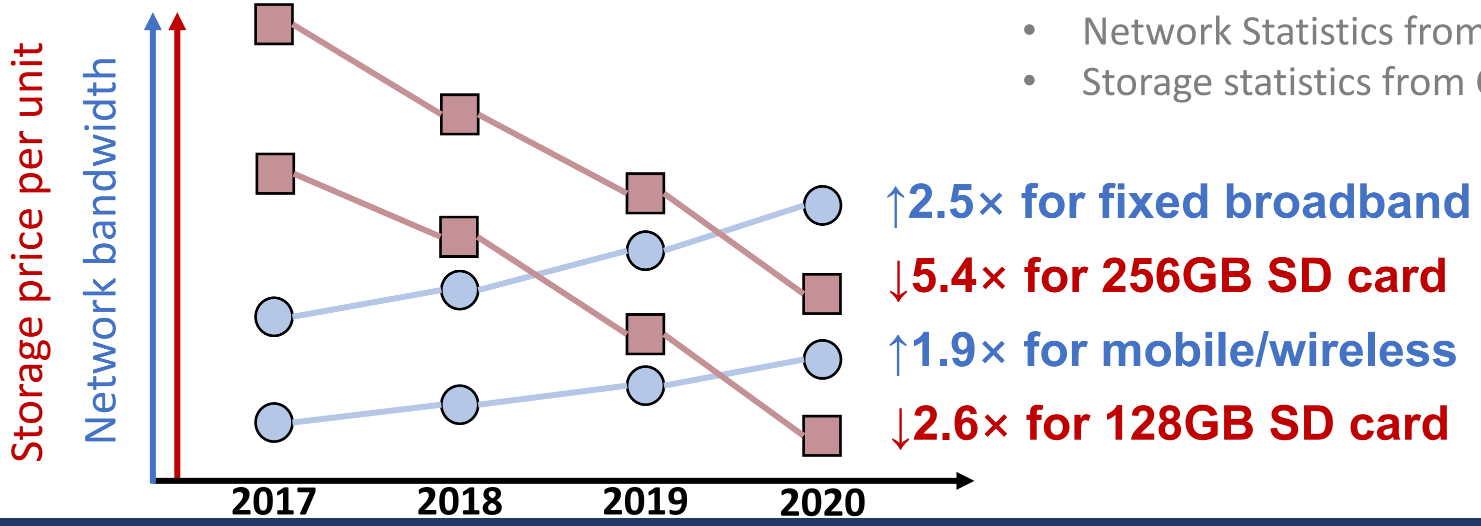
- Motivations
  1. Most videos are **cold**: a case study from PKU campus
     - More than 1,000 cameras deployed
     - Only <0.005% video and <2% cameras are eventually queried within 6 months
  2. Network bandwidth, especially wireless, remains **precious**



- Network Statistics from SpeedTest

Zero-streaming Cameras

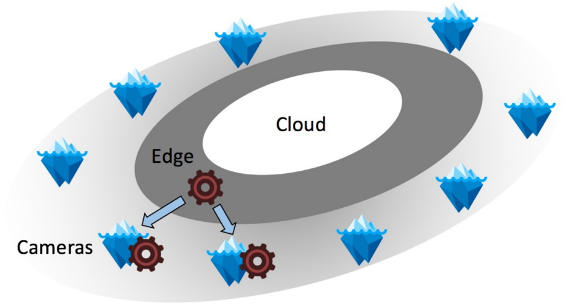
- Motivations
  1. Most videos are **cold**: a case study from PKU campus
     - ❑ More than 1,000 cameras deployed
     - ❑ Only <0.005% video and <2% cameras are eventually queried within 6 months
  2. Network bandwidth, especially wireless, remains **precious**
  3. Storage is becoming increasingly **ample**



- Network Statistics from SpeedTest
- Storage statistics from Camelcamelcamel

Zero-streaming Cameras

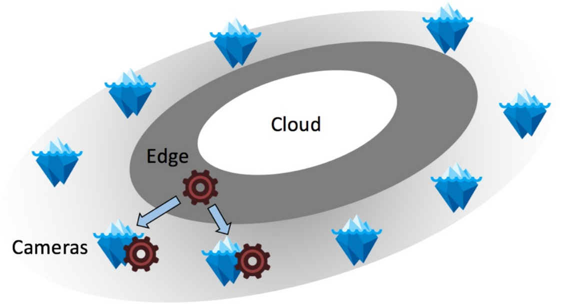
- Motivations
  1. Most videos are **cold**: a case study from PKU campus
     - ❑ More than 1,000 cameras deployed
     - ❑ Only <0.005% video and <2% cameras are eventually queried within 6 months
  2. Network bandwidth, especially wireless, remains **precious**
  3. Storage is becoming increasingly **ample**
- **Zero-streaming: shifting network constraint to camera storage**
  - Ingestion time: stored to local storage
  - Query time: camera-cloud collaboration



Cameras capture videos & keep silence
Only respond to queries

Zero-streaming Cameras

- Motivations
  1. Most videos are **cold**: a case study from PKU campus
     - ❑ More than 1,000 cameras deployed
     - ❑ Only <0.005% video and <2% cameras are eventually queried within 6 months
  2. Network bandwidth, especially wireless, remains **precious**
  3. Storage is becoming increasingly **ample**
- **Zero-streaming: shifting network constraint to camera storage**
  - Ingestion time: stored to local storage
  - Query time: camera-cloud collaboration
- **Challenge: accelerating video query**
  - Limited BW is the bottleneck: the order matters

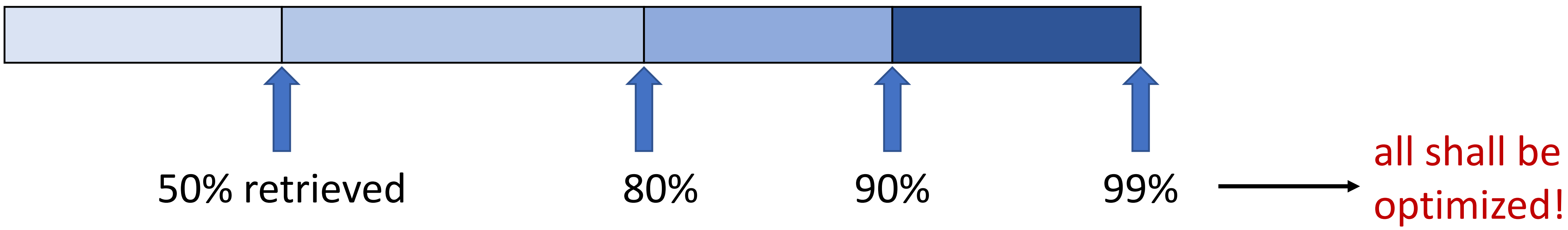


Cameras capture videos & keep silence
Only respond to queries

DIVA: a runtime for 0-streaming cameras

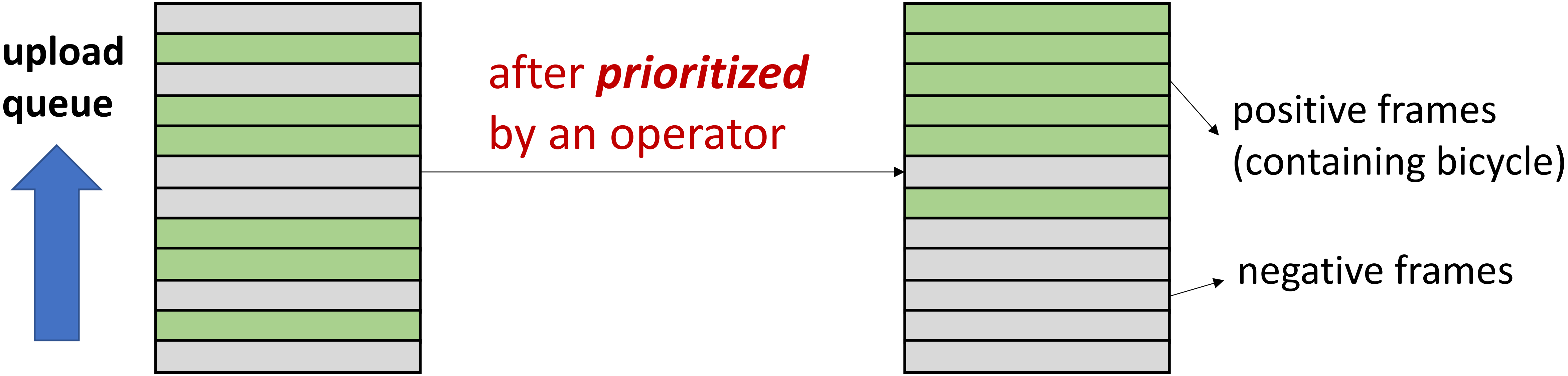
- Key idea: exploratory query with *online refinement*
  - Deliver early results to users AFAP & keep refinement

Q: retrieve all images with bicycles?



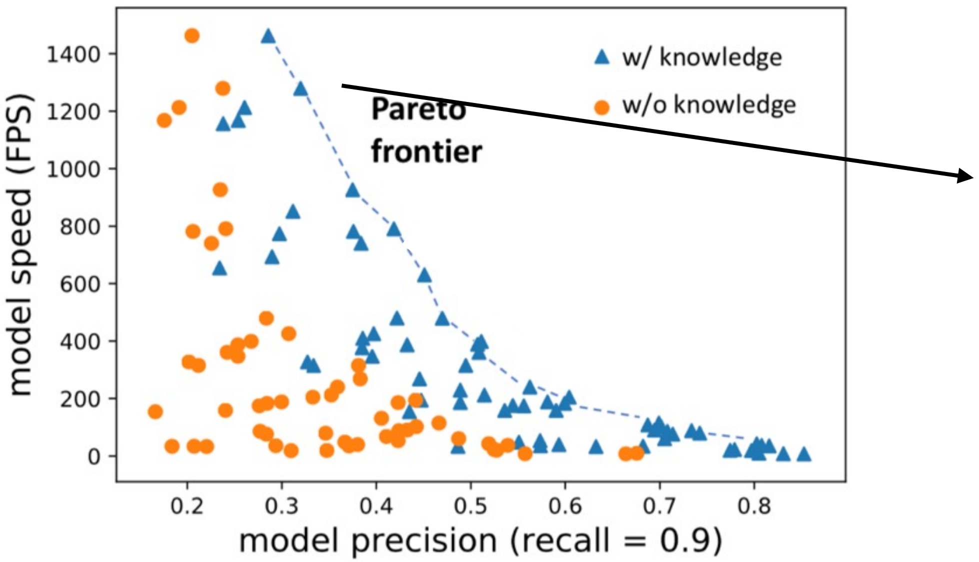
DIVA: a runtime for 0-streaming cameras

- Key idea: exploratory query with *online refinement*
  - Deliver early results to users AFAP & keep refinement
- **Key technique: multi-pass on-camera process (operator upgrade)**
  - Operator: specialized (for query) NNs, on-the-fly trained



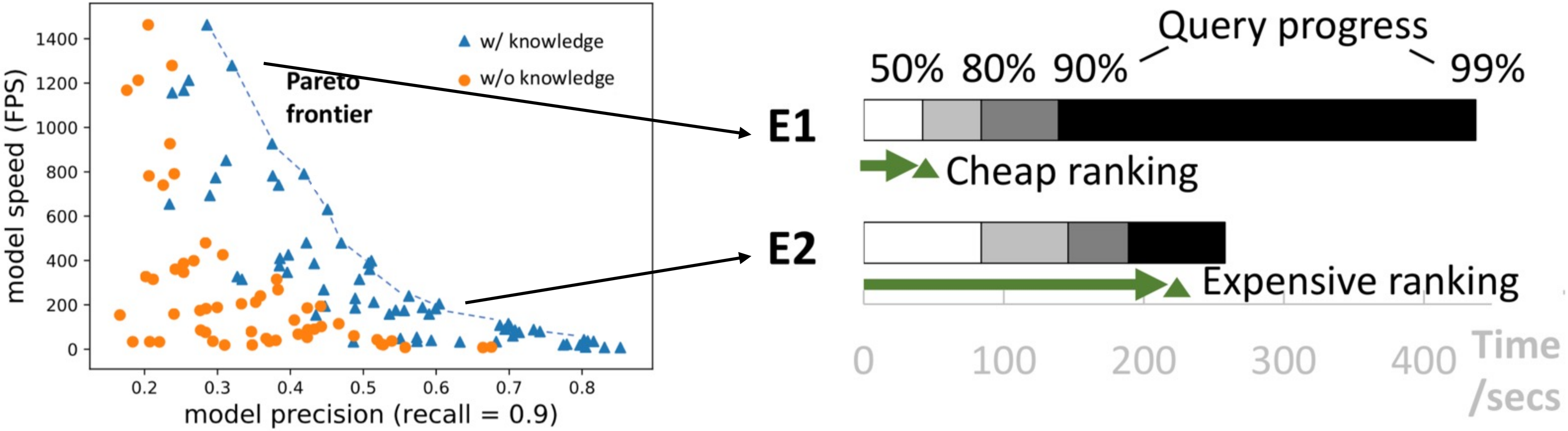
DIVA: a runtime for 0-streaming cameras

- Key idea: exploratory query with *online refinement*
  - Deliver early results to users AFAP & keep refinement
- **Key technique: multi-pass on-camera process (operator upgrade)**
  - Operator: specialized (for query) NNs, on-the-fly trained



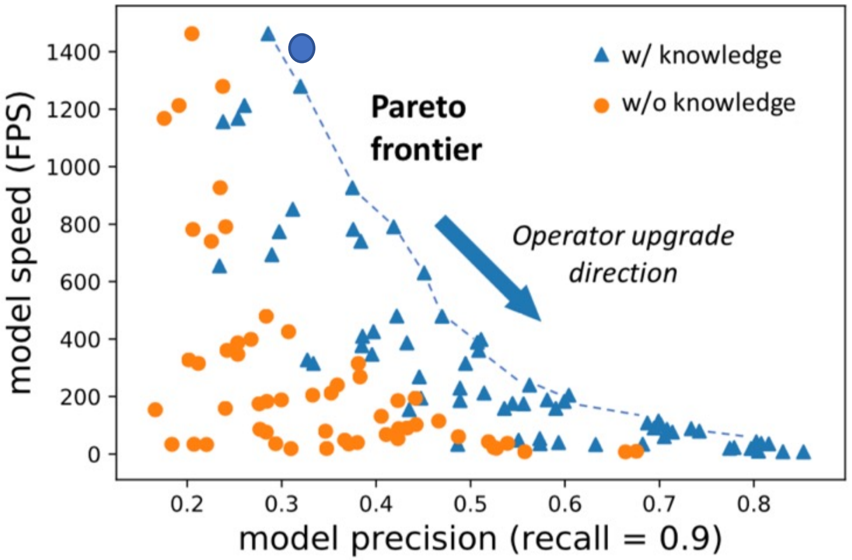
DIVA: a runtime for 0-streaming cameras

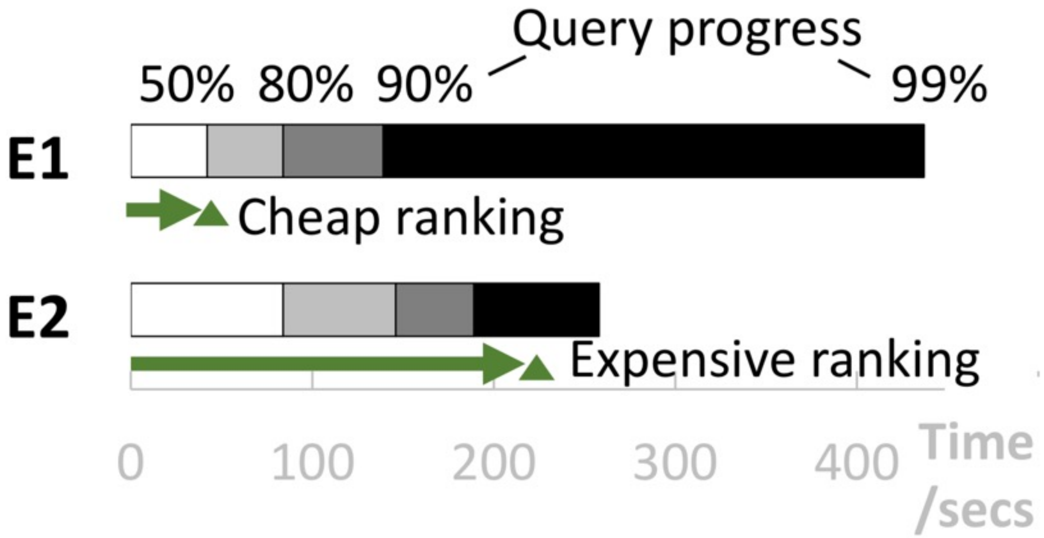
- Key idea: exploratory query with *online refinement*
  - Deliver early results to users AFAP & keep refinement
- **Key technique: multi-pass on-camera process (operator upgrade)**
  - Operator: specialized (for query) NNs, on-the-fly trained



DIVA: a runtime for 0-streaming cameras

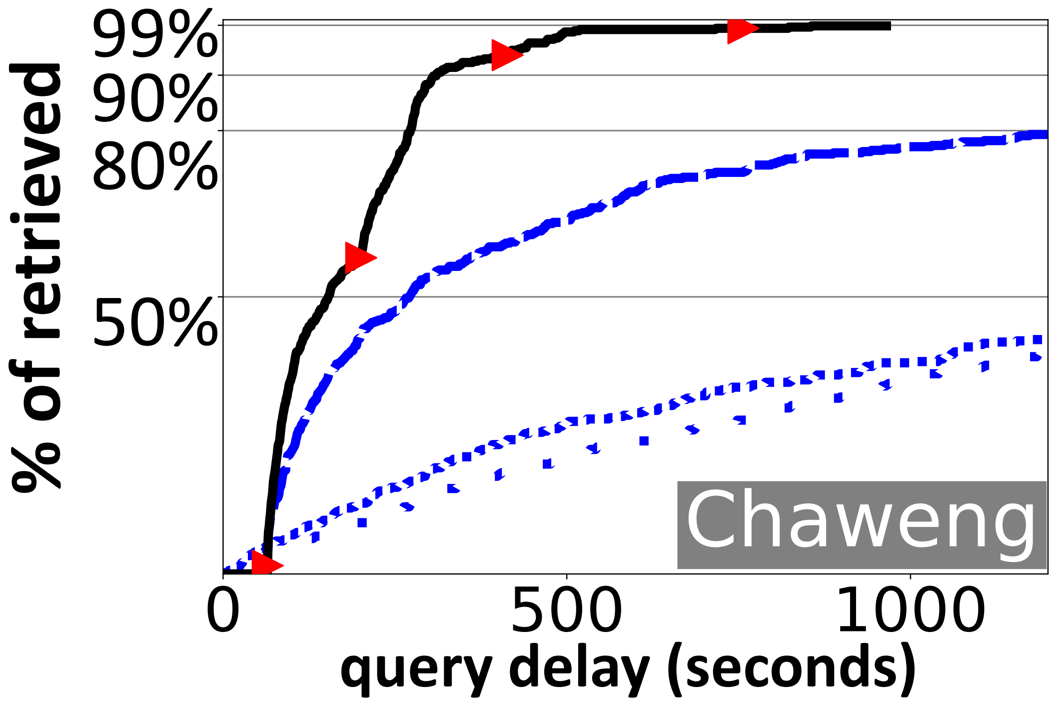
- Key idea: exploratory query with *online refinement*
  - Deliver early results to users AFAP & keep refinement
- **Key technique: multi-pass on-camera process (operator upgrade)**
  - Operator: specialized (for query) NNs, on-the-fly trained





Highlights of experiment results

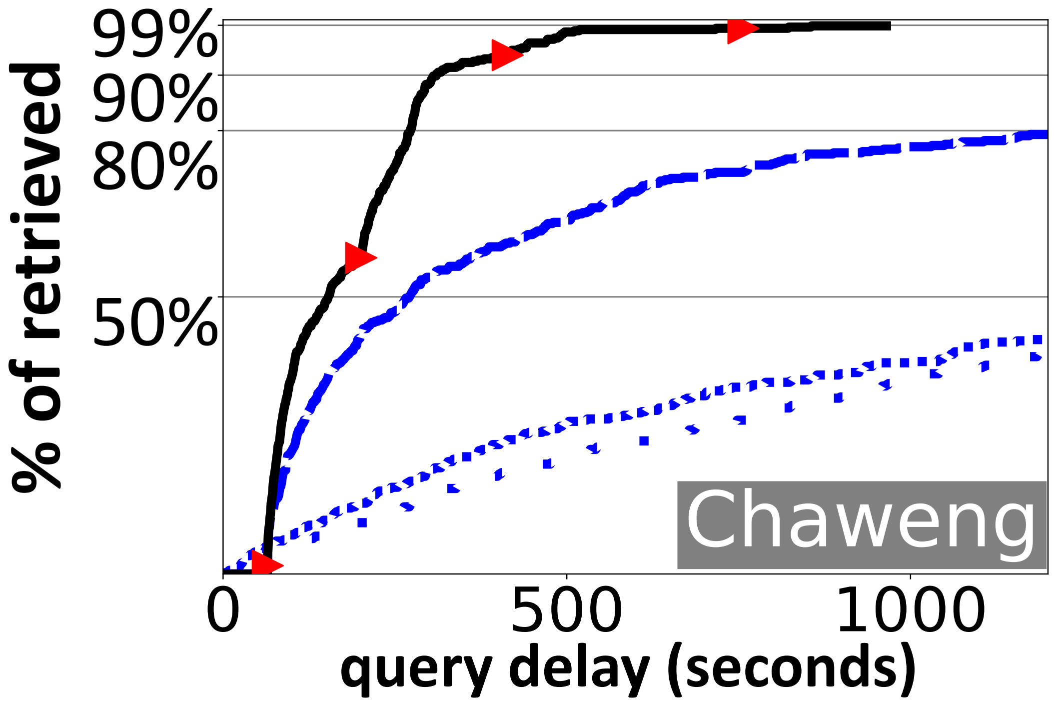
- On 15 real videos (720 hrs in total), two representative camera hardware, 3 query types
- We are 4-30X faster than competitive alternatives

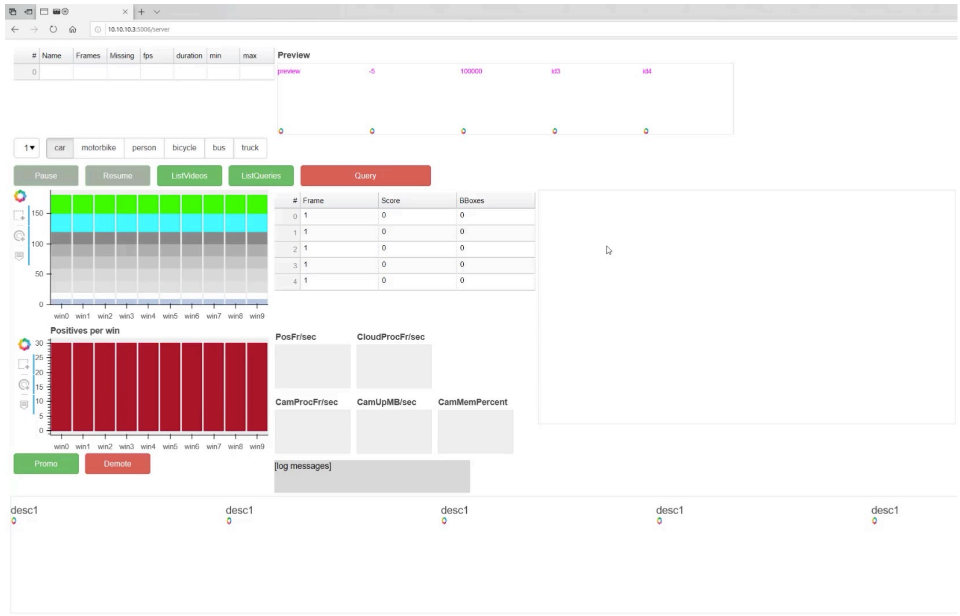


Example: How fast we can retrieve frames with bicycles to users?

Highlights of experiment results

- On 15 real videos (720 hrs in total), two representative camera hardware, 3 query types
- We are 4-30X faster than competitive alternatives





A web-based demo

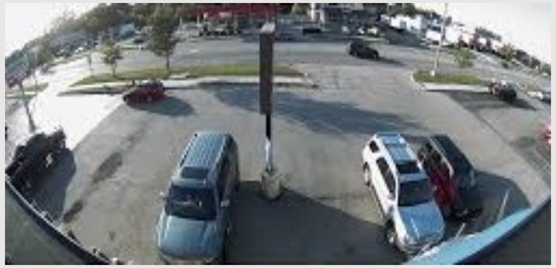
Outline

- Edge Intelligence: What and Why?
  - A system software perspective

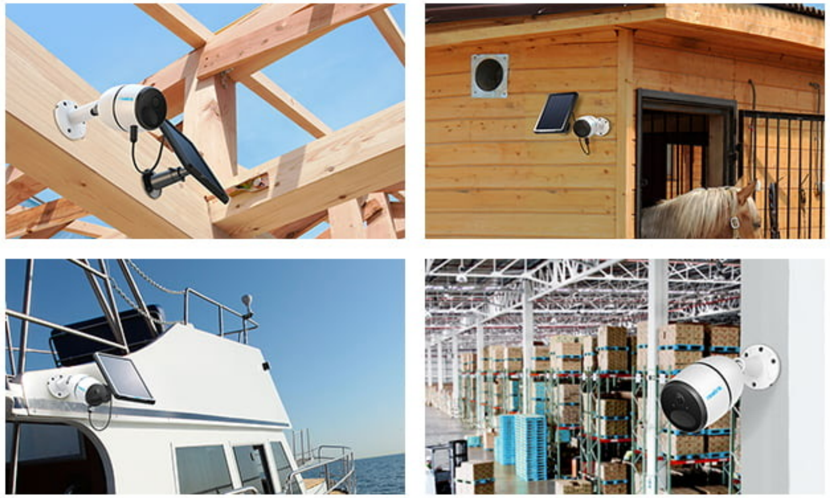
- Two pieces of my research on AIoT cameras
  - Zero-streaming Cameras (full paper under review, MobiCom’20 Demo)
  - **Autonomous Cameras (MobiSys’20)**

Autonomous Camera

- Busy cross roads
- Retailing store
- Sports stadium
- Parking lots
- …



?



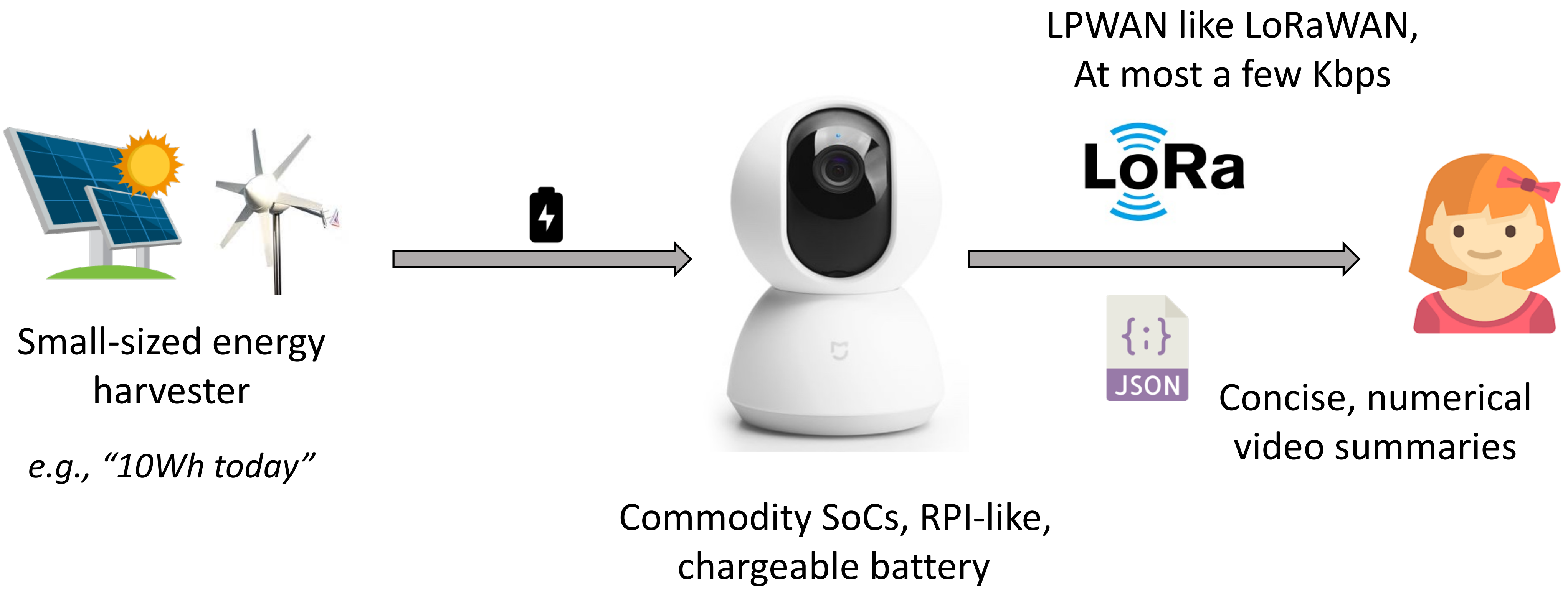
- Construction sites
- Cattle farms
- Highways
- Wildlifes
- …

Urban, residential areas

Rural, off-grid areas

Autonomous Camera

- **Energy-independent** and **Compute-independent**



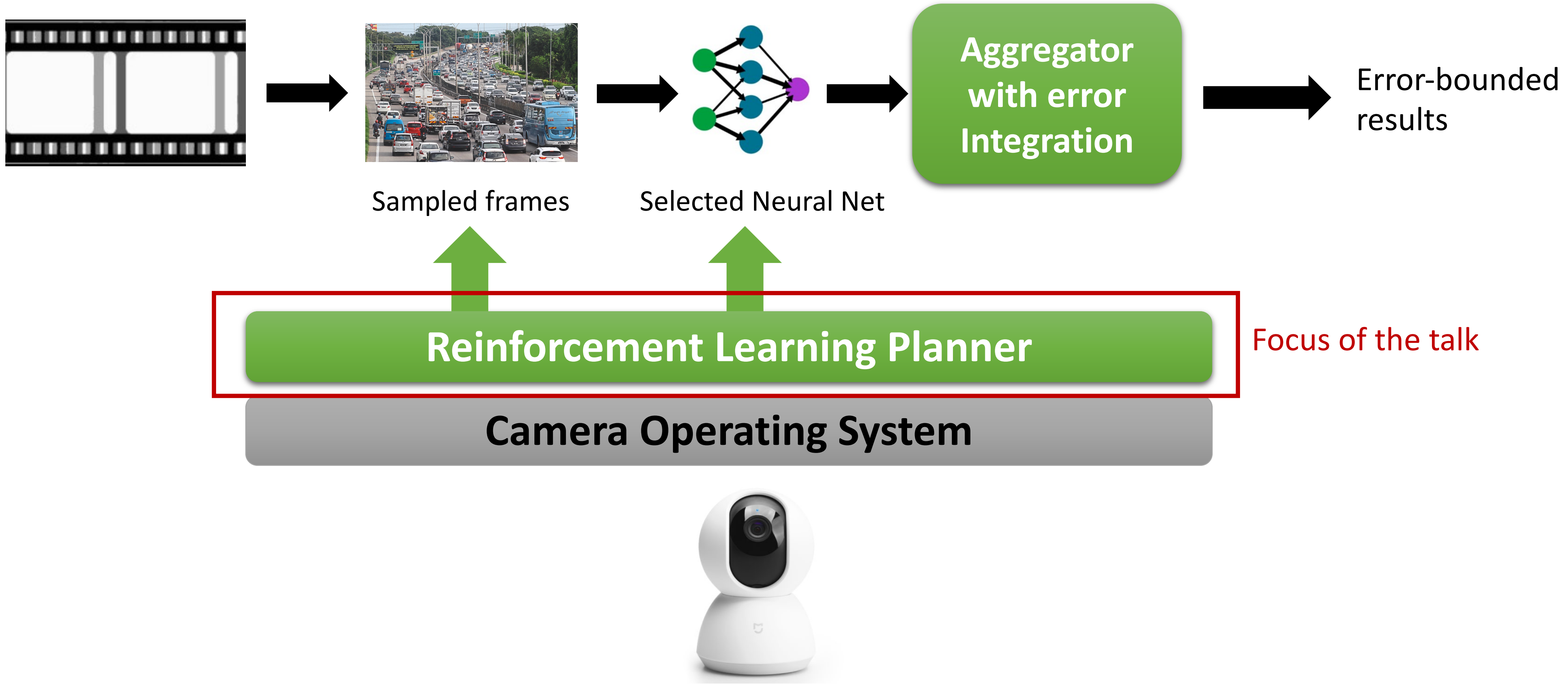
Autonomous Camera

- **Energy-independent** and **Compute-independent**

- **Target query:** summarize video based on time windows
  - With bounded error, e.g., confidence interval (CI).

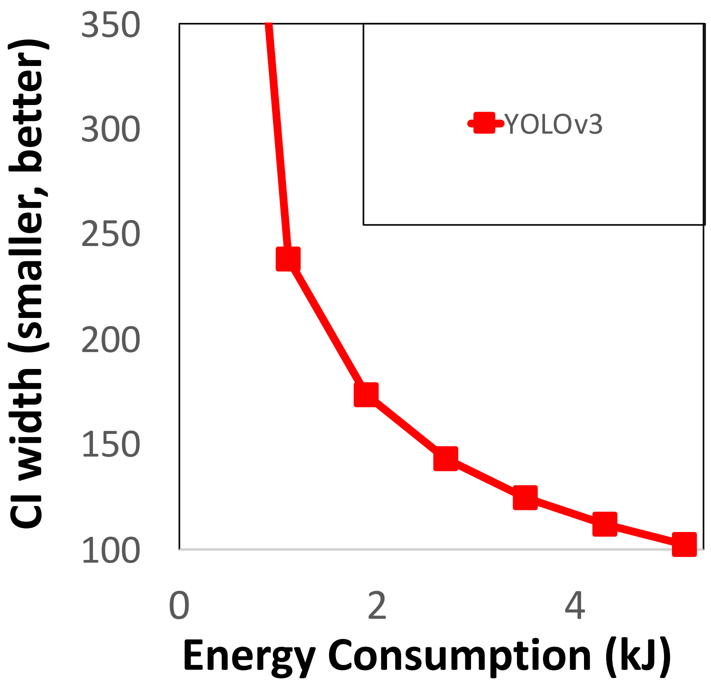
- **The central problem**: planning constrained energy (an energy budget)
  - Not enough to run the most expensive NN on every frame!
  - Key trade-offs: frame sampling and NN selection

Elf Runtime for autonomous camera



Elf tech #1: per-window planning

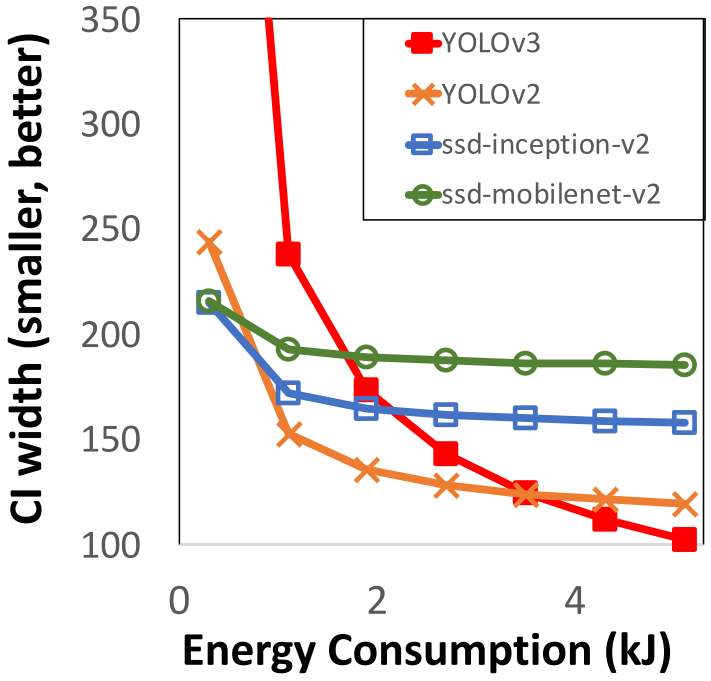
- What's the best **sampling rate** and **NN** for a window?



Energy Consumption = E(NN) * frame_num

Elf tech #1: per-window planning

- What's the best **sampling rate** and **NN** for a window? – No silver bullet

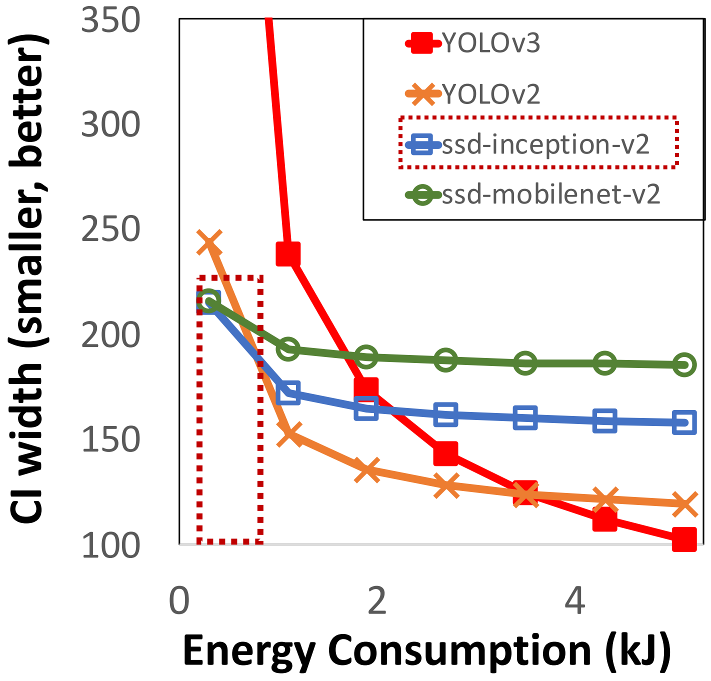


Energy Consumption = E(NN) * frame_num

NN Counters	Input	mAP	Energy
YOLOv3 (Golden, GT) [85]	608x608	33.0	1.00
YOLOv2 [84]	416x416	21.6	0.22
faster rcnn inception-v2 [86]	300x300	28.0	0.40
ssd inception-v2 [68]	300x300	24.0	0.08
ssd mobilenet-v2 [88]	300x300	22.0	0.05
ssdlite mobilenet-v2 [88]	300x300	22.0	0.04

Elf tech #1: per-window planning

- What's the best **sampling rate** and **NN** for a window? – No silver bullet

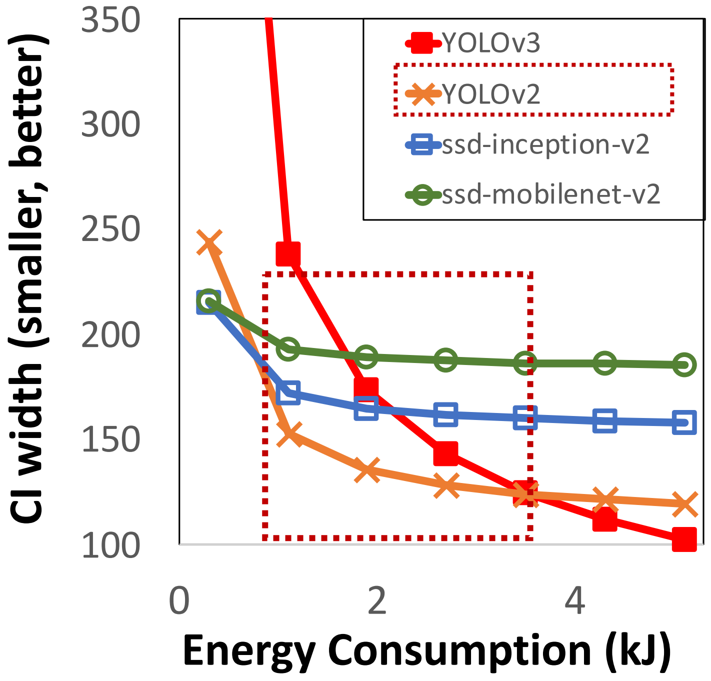


When energy is low: cheaper NNs win

- **Bottlenecked by sampling error (frame quantity)**

Elf tech #1: per-window planning

- What's the best **sampling rate** and **NN** for a window? – No silver bullet



When energy is low: cheaper NNs win

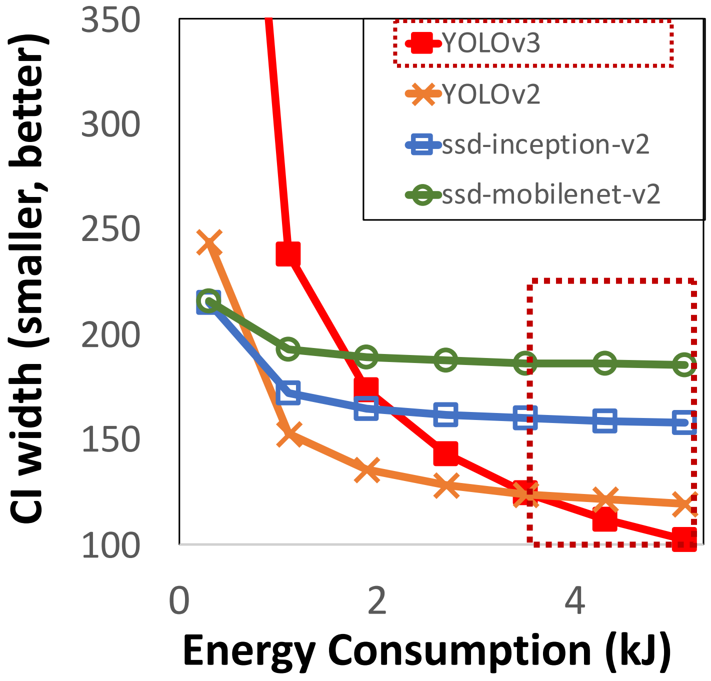
- **Bottlenecked by sampling error (frame quantity)**

When energy is high: more accurate NNs win

- **Bottlenecked by NN error (frame quality)**

Elf tech #1: per-window planning

- What's the best **sampling rate** and **NN** for a window? – No silver bullet



When energy is low: cheaper NNs win

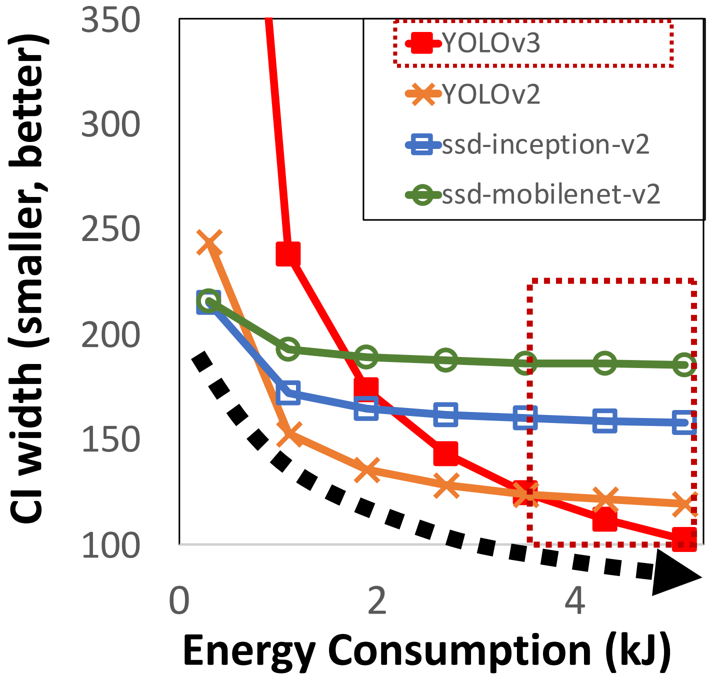
- **Bottlenecked by sampling error (frame quantity)**

When energy is high: more accurate NNs win

- **Bottlenecked by NN error (frame quality)**

Elf tech #1: per-window planning

- What's the best **sampling rate** and **NN** for a window? – No silver bullet



When energy is low: cheaper NNs win

- Bottlenecked by sampling error (frame quantity)

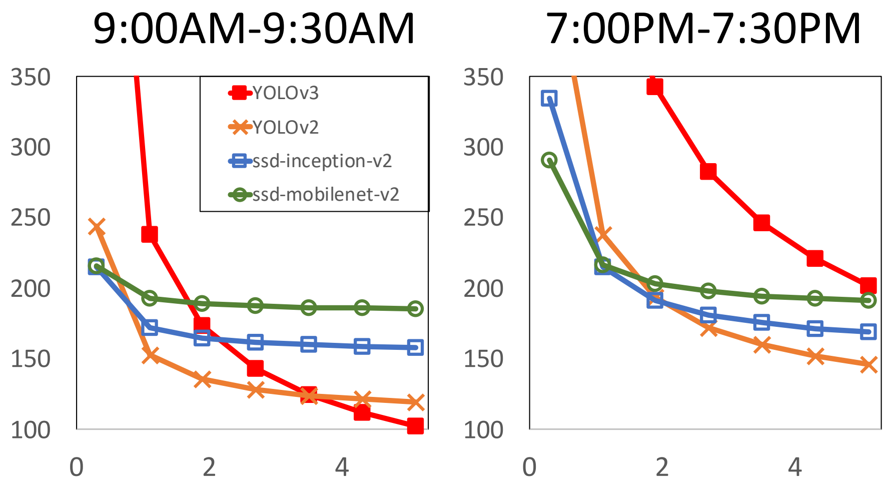
When energy is high: more accurate NNs win

- Bottlenecked by NN error (frame quality)

***Energy/CI front:** the combination of all "optimal" decisions with varied energy*

Elf tech #1: per-window planning

- What's the best **sampling rate** and **NN** for a window? – No silver bullet



Different windows have different energy/CI fronts

When energy is low: cheaper NNs win

- Bottlenecked by sampling error (frame quantity)

When energy is high: more accurate NNs win

- Bottlenecked by NN error (frame quality)

Energy/CI front: *the combination of all "optimal" decisions with varied energy*

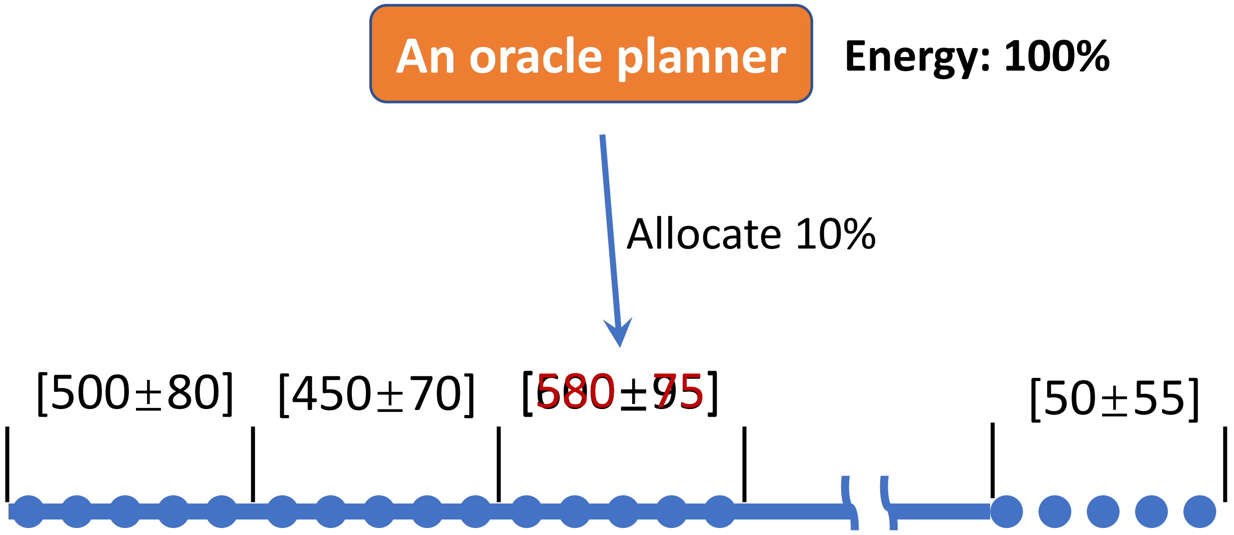
- Depends on the video characteristics

Elf tech #2: across-window planning

- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts

Elf tech #2: across-window planning

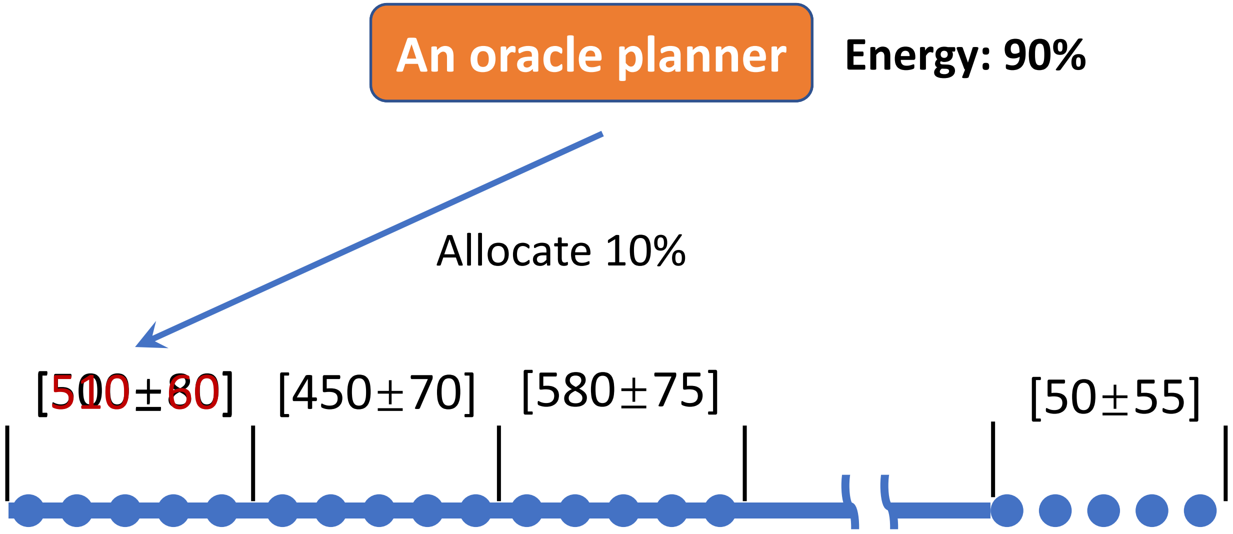
- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts



A greedy approach: giving energy to the window with the most benefit (i.e., CI width reduction).

Elf tech #2: across-window planning

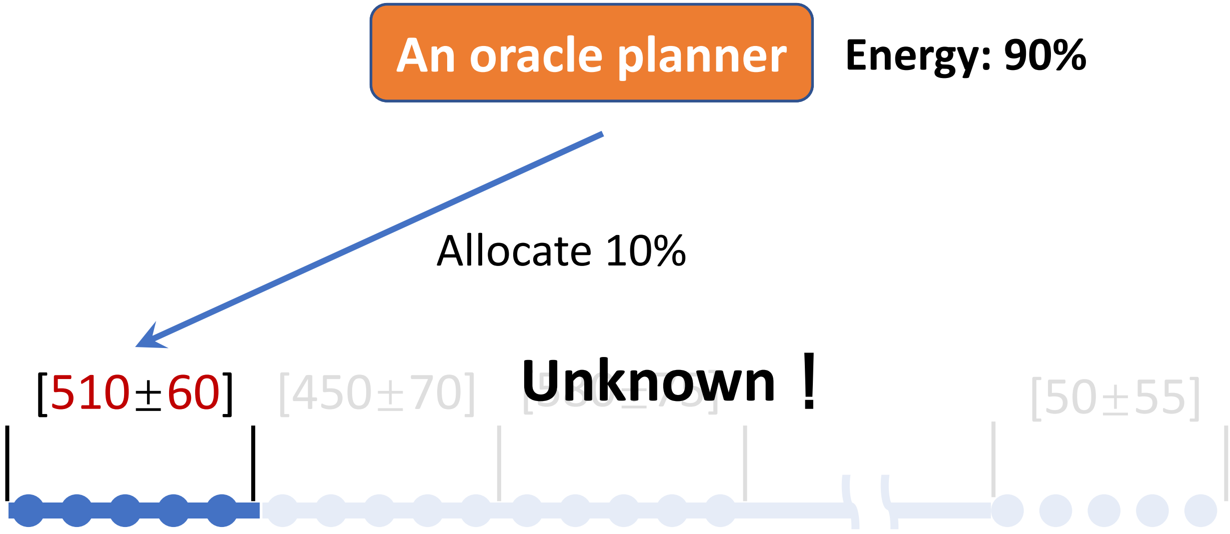
- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts



A greedy approach: giving energy to the window with the most benefit (i.e., CI width reduction).

Elf tech #2: across-window planning

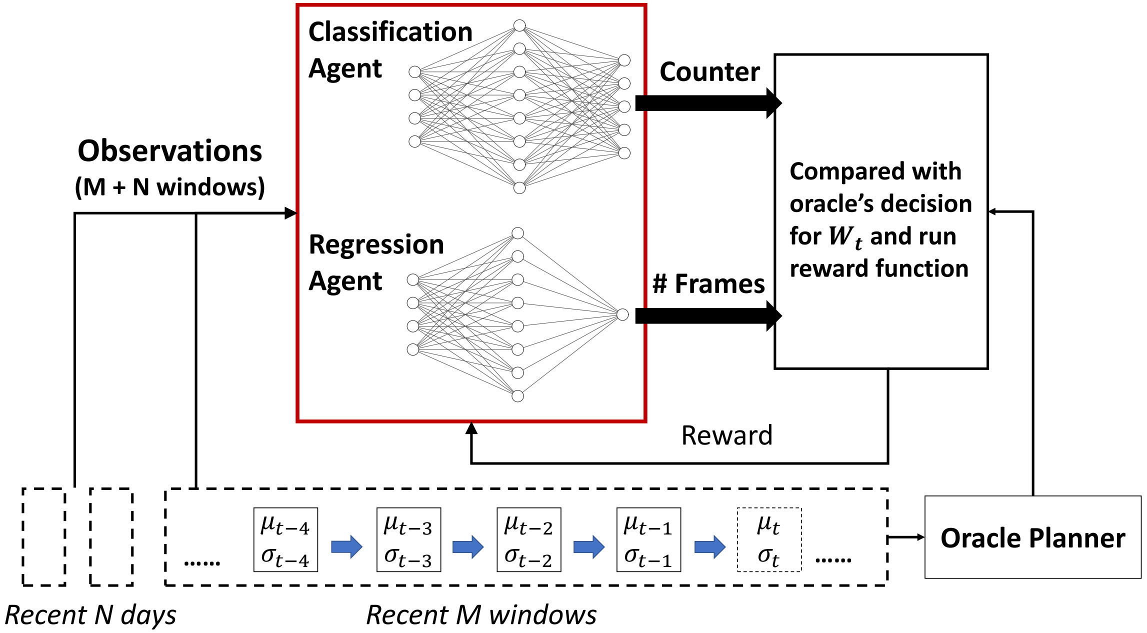
- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts



A greedy approach: giving energy to the window with the most benefit (i.e., CI width reduction).

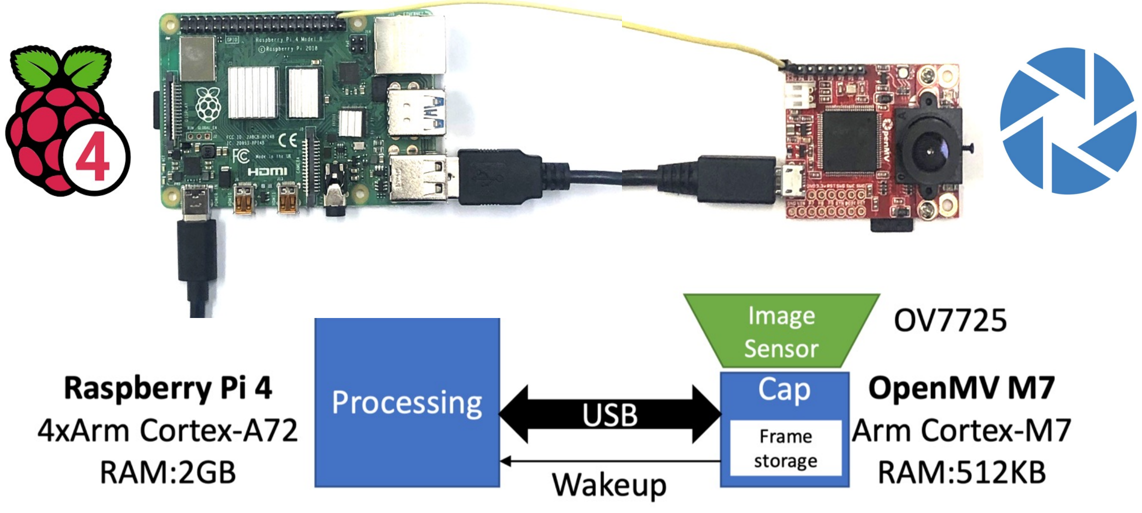
Elf tech #2: across-window planning

- An Oracle Planner: best performance but unrealistic
  - knows all energy/CI fronts
- A learning-based planner: imitating the oracle planner
  - basis: reinforcement learning
  - rationale: daily and temporal patterns
  - offline training -> online prediction
    - Two agents: NN selection and # of frames
    - Observations: knowledge of past windows
    - Penalty: deviation from oracle's decision

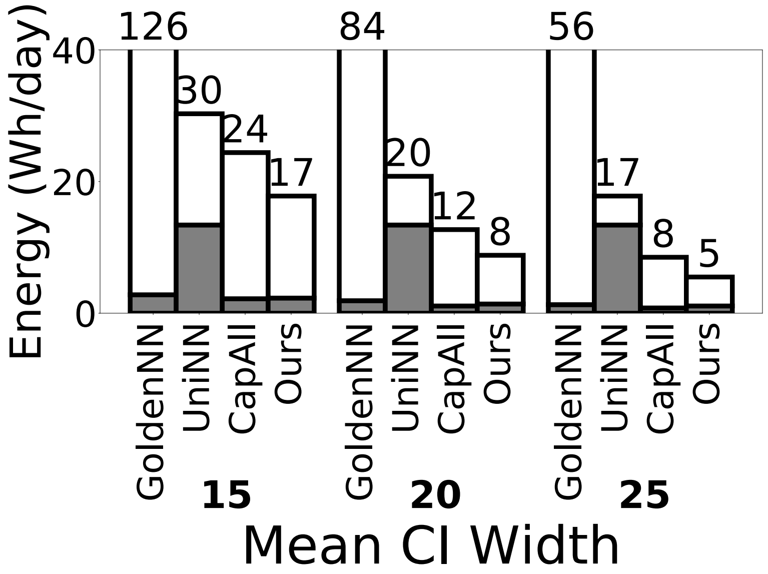


Highlights of experiment results

- **Implementation:** heterogeneous hardware



reduce energy for image capture



- **Evaluated** on over 1,000-hrs video
  - Saves up to 10X energy (to meet accuracy)

Takeaways

- Edge devices shall/will be intelligent by themselves
  - A trend of decentralization…
  - Good system support is badly needed!

- AIoT cameras are the next promising platform for edge intelligence
  - They can be zero-streaming, or even autonomous!
  - A brand new vision: *camera-as-a-service* (under major revison of IEEE Pervasive Computing)