# Understanding Reliability on Shared Edge

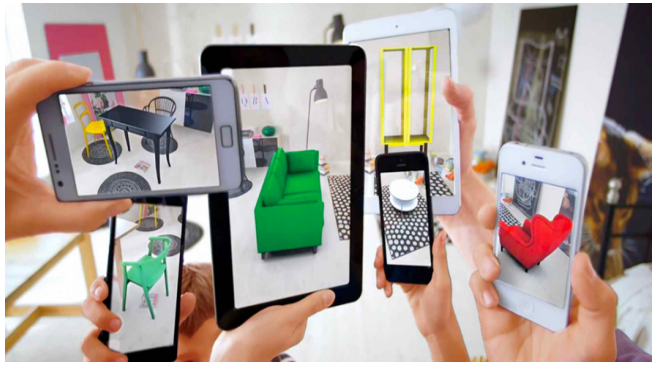
Faculty: Haryadi Gunawi, Junchen Jiang

CHI@Edge Community Workshop

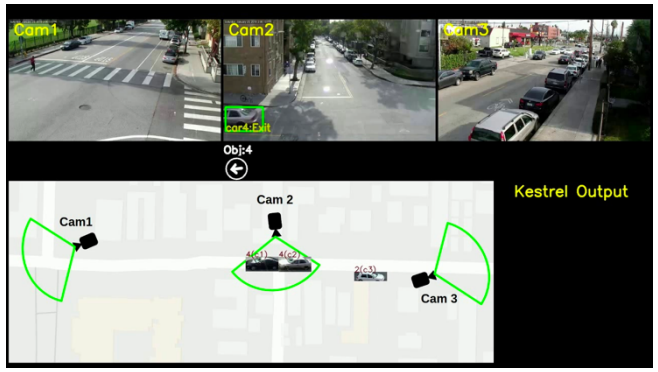
9/9/2021

THE UNIVERSITY OF CHICAGO

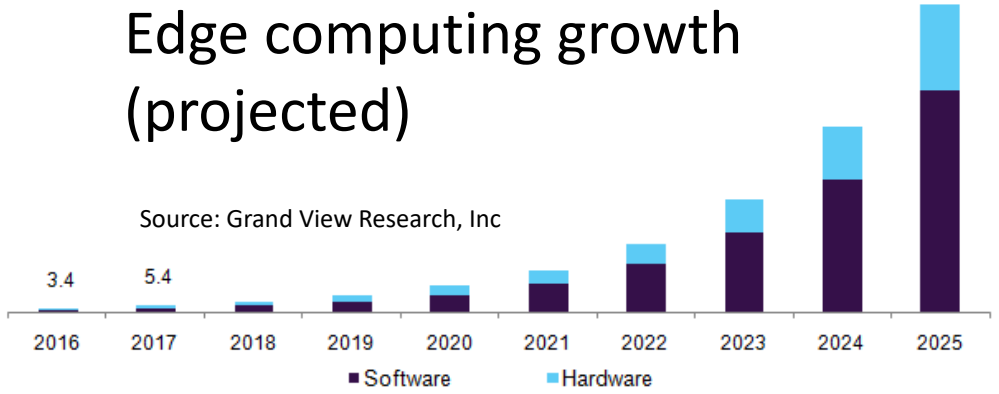
# AI-powered edge applications are growing fast!



Collaborative Augmented Reality



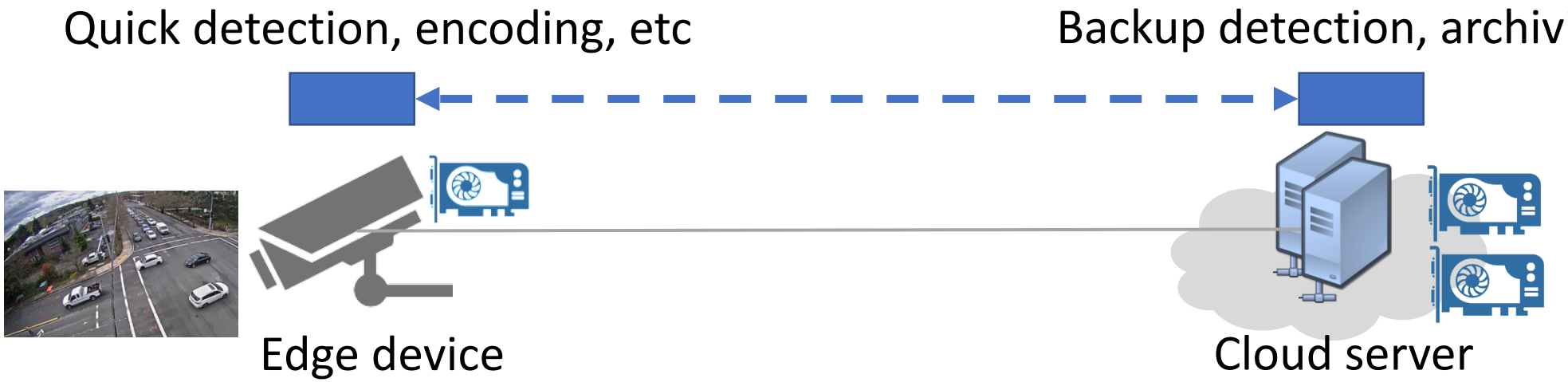

Camera-network Video Analytics



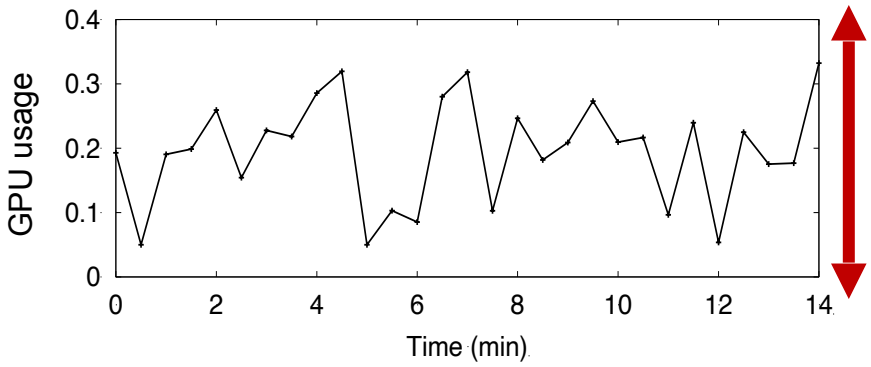

More software running on limited hardware

Goal: Addressing key systems challenges in a **shared edge** that enables real-time accuracy-driven applications

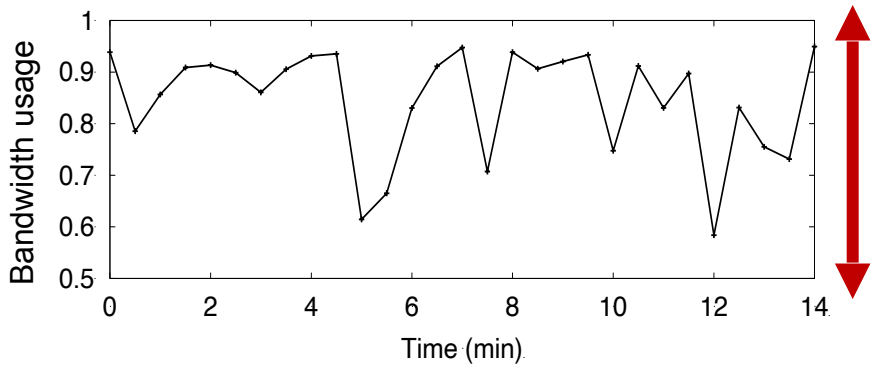
# Resource demands of video analytics pipelines




High accuracy with low **average** resource usage, but create more **bursty** resource demands




6x change in compute

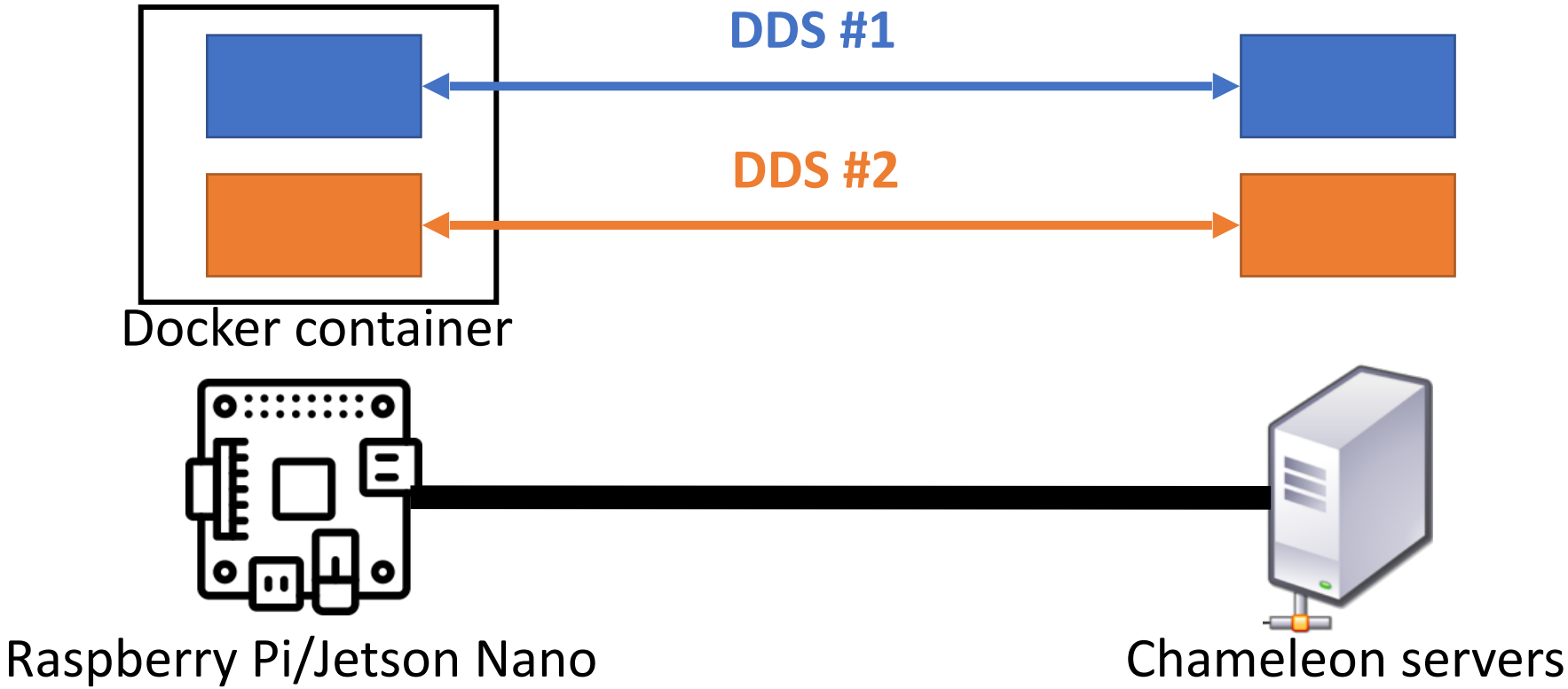



40% diff in bandwidth



"Towards Performance Clarity of Edge Video Analytics" Symposium on Edge Computing, 2021

# CHI@Edge offers a testbed for resource allocation



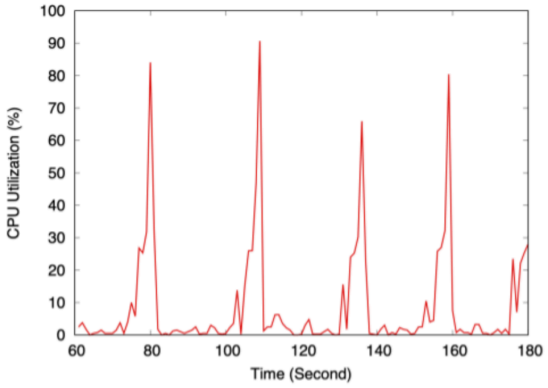

DDS [SIGCOMM'20] uses server-side feedback to iteratively encode videos at low quality but still allowing accurate DNN inference

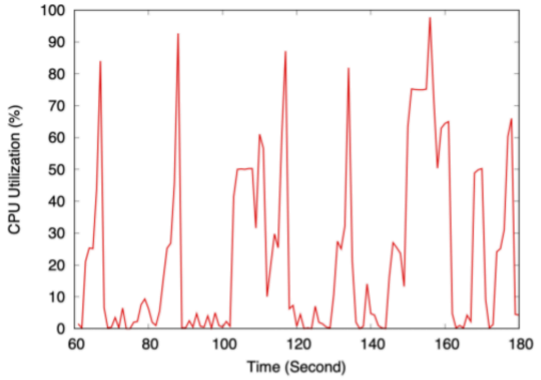
# Early example results of resource demands

Per pipeline

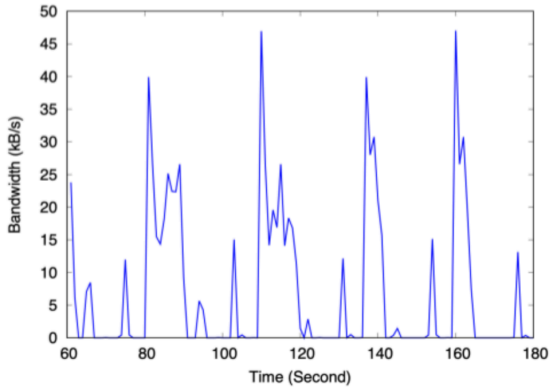
Total of two pipeline (without isolation)

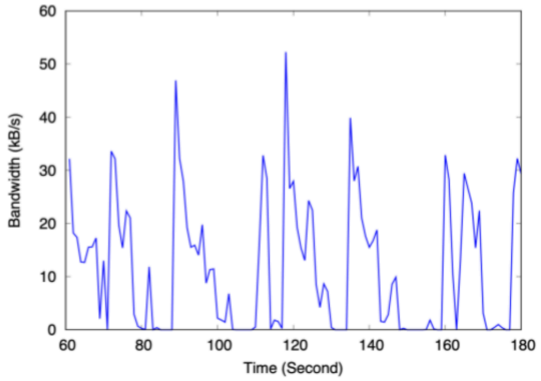
Edge CPU demands



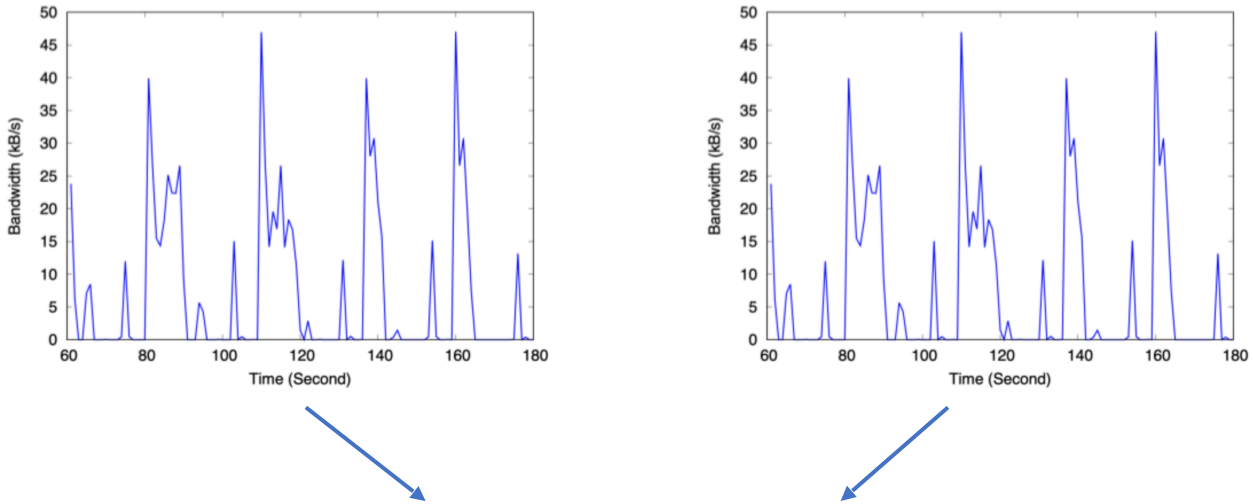




Bandwidth demands



# Challenges of today's solutions




Resource-sharing framework

Resource sharing strawmen:

- No isolation: High tail latency
- Fair allocation: resource wastage

Opportunity: Video analytics pipelines have great flexibility to trade accuracy for latency

We seek to find ways to maximize min accuracy and minimize max delay

# Technical plan

- **Continuously profiling impact of bandwidth/latency on analytical quality**
  - Analytical quality = analytics accuracy + freshness of inference results
  - Our recent work has shown this is doable by leveraging output of the analytics DNN engine (Chameleon [SIGCOMM'18], DDS[SIGCOMM'20]).

- **Two potential sharing mechanisms:**
  - Dynamically traffic shaping to ensure enough bandwidth for each pipeline's need
  - Cross-layer solution where resource-sharing logic and video analytics pipeline negotiate to minimize impact on analytics accuracy

# Experience with CHI@Edge so far

- **Early stage use cases**
  - Raspberry Pi: typical low-end device, good fit for simple vision tasks
  - Jetson Nano: beefy GPU (~50fps VGG-16) but low CPU RAM, good fit for specific vision tasks, but resource overall is limited

- **Many issues already addressed** (*big shout out to the team!!*)
  - Creation and loading of Docker images
  - TensorFlow version issues
  - …

- **More issues emerging**
  - Real-time access to (& control over) compute/network resource (e.g., linux tc)
  - …