Optimizing Internet video through support from the network edge

Sanjay Rao

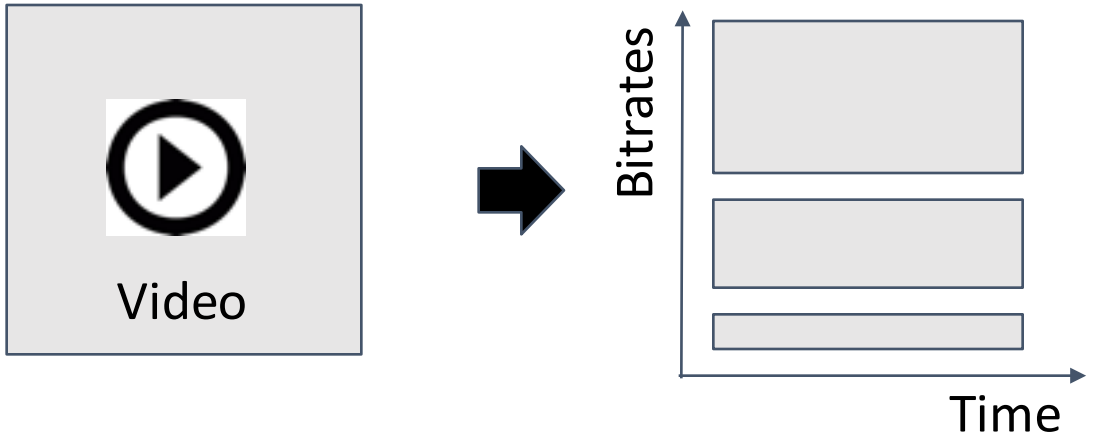
School of Electrical and Computer Engineering

Purdue University

Project goals and motivation

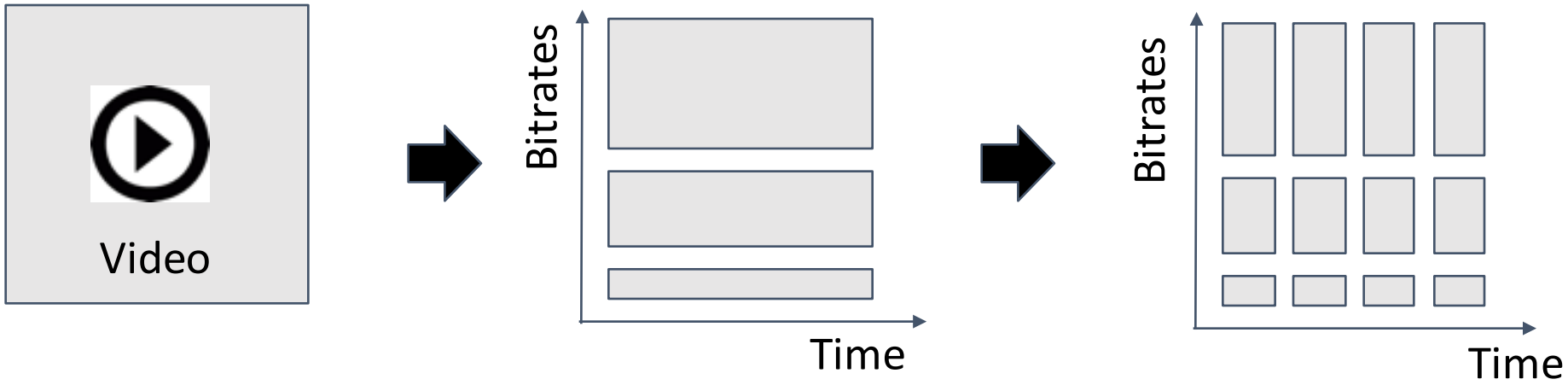
- Video streaming: dominant form of network traffic
  - Expected to account for 82% of all Internet traffic by 2022 [Cisco forecast]
  - Performance critical for user engagement.
  - 1% increase in rebuffering may lead to 3 minutes reduction in user view time [Sigcomm11]
- Video performance still a challenge
  - Higher bit rate video (4K video => 50Mbps)
  - Wide disparity in broadband quality across users (e.g., US FCC Report)
  - Growth in live video
- Project overview: Novel approaches to video streaming enabled by value added services at the network edge

Background: Adaptive Bitrate Streaming



A video clip is encoded
with multiple qualities (bitrates)

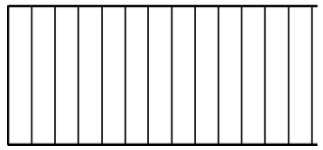
Background: Adaptive Bitrate Streaming

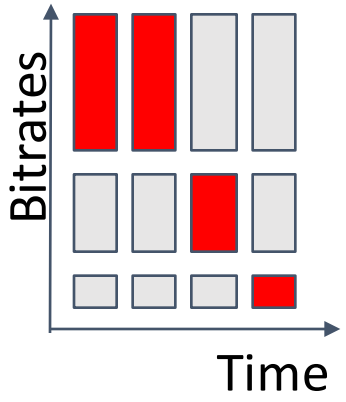


Each bitrate is split into chunks

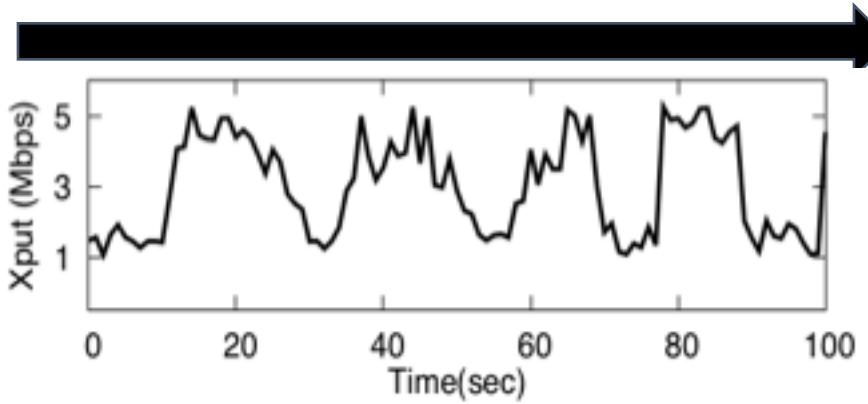
Background: Adaptive Bitrate Streaming

Video Client





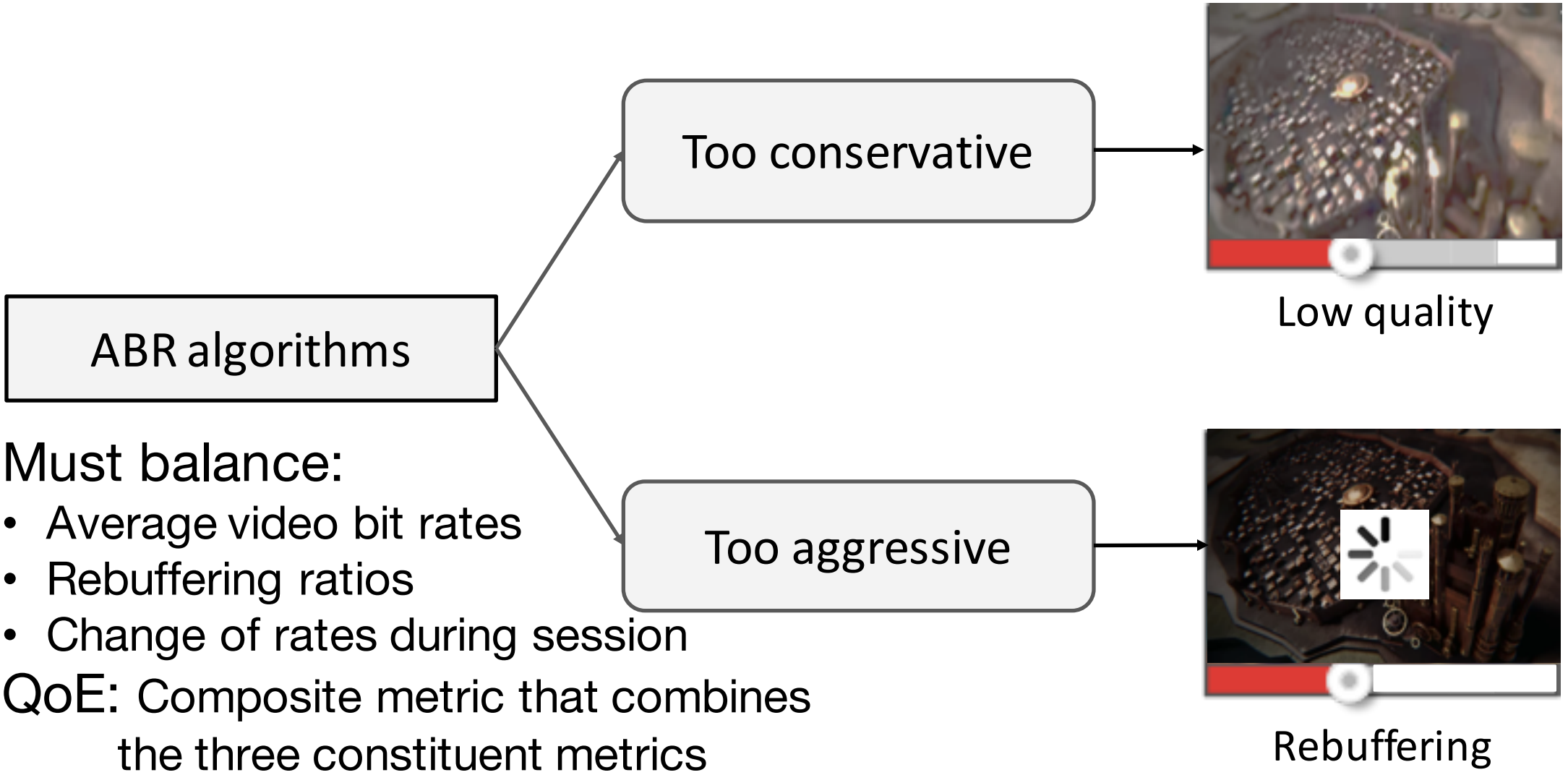
Request n*th* chunk at bitrate *r*



Video Server

Adaptive Bitrate Algorithms(ABR)

Background: Adaptive Bitrate Streaming



Must balance:

- Average video bit rates
- Rebuffering ratios
- Change of rates during session

QoE: Composite metric that combines the three constituent metrics

Challenges and our research

- Key limitation of ABR algorithms today:
  - Rely on throughput prediction based on local (end-host based) inference of network state – often erroneous
  - Interactions across multiple adaptive streaming players
- Our ongoing research:
  - Can awareness of hierarchical CDN architecture help ABR algorithm design?
  - Sharing information across other video flows going through the edge.
    - Collaboration with Prof. Marco Mellia, Politecnico di Torino
  - Developing richer throughput prediction frameworks

Research #1. ABR and hierarchical CDNs

- CDNs => hierarchical cache structure.
- Objects may be served from the CDN edge, or a higher level CDN cache, or may miss altogether in the CDN.
- Today's ABR algorithms are agnostic of where in the CDN objects are served from
- Our work:
  - Measurement study to understand how chunks within the same video session are served in hierarchical CDN settings
  - Explore how video player perceived throughput varies based on where chunks are served from
  - Evaluate the implications of CDN hierarchies for ABR algorithms

Measurement Methodology

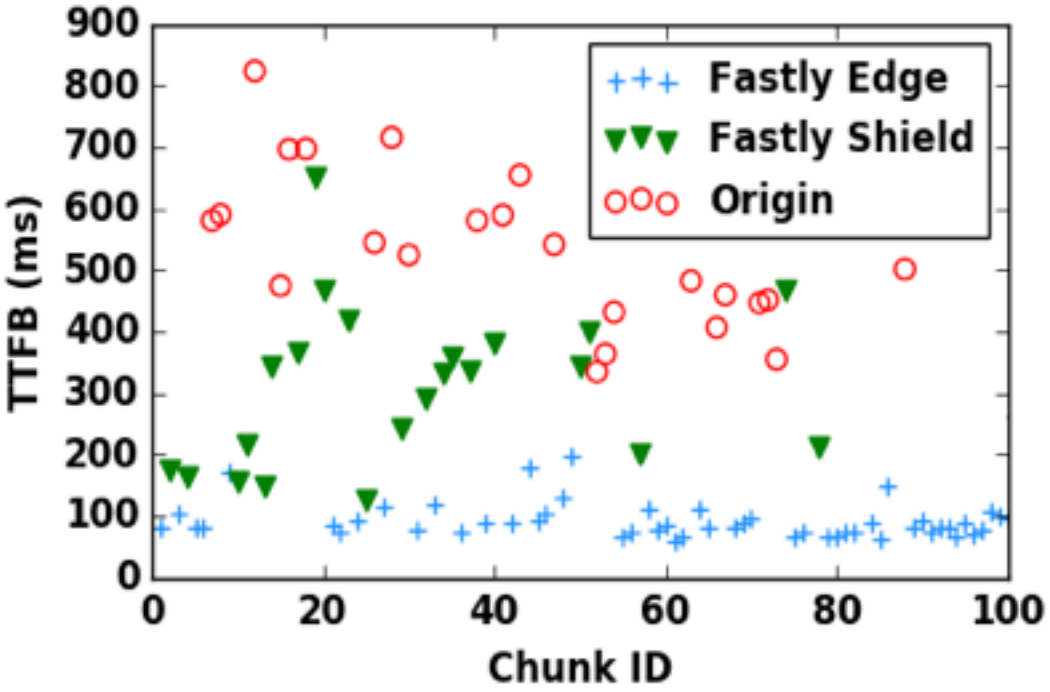
- How to know in an end-to-end fashion where in the CDN hierarchy a video chunk is served from?
- Our methodology: Leverage CDN pragmas

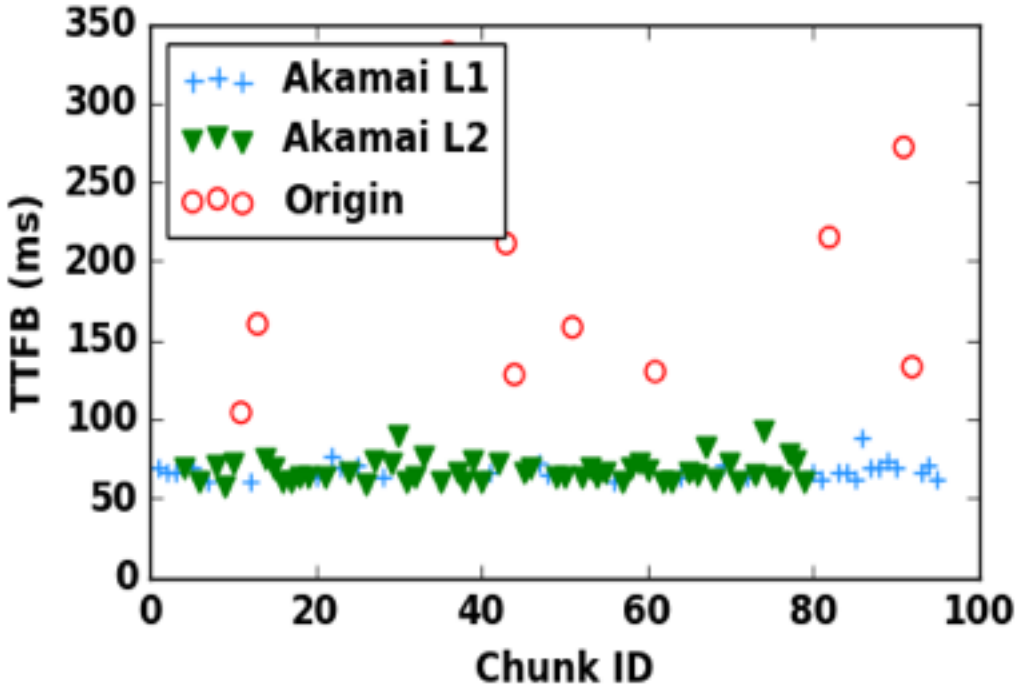
CDN	Required header
Akamai	pragma:akamai-x-cache-on pragma:akamai-x-cache-remote-on
Fastly	Fastly-Debug:1
CloudFront	No header required

CDN	HTTP response headers	Value	Server
Akamai	X-Cache X-Cache-Remote	TCP_HIT or TCP_MEM_HIT (Any response)	Akamai L1 Akamai L2
		TCP_MISS (All responses)	Origin
Fastly	X-Cache	HIT,HIT	Edge
		MISS,HIT	
		HIT	Shield
		HIT,MISS	
		MISS	Origin
		MISS,MISS	
CloudFront	X-Cache	Hit from cloudfront	Cloudfront
		Miss from cloudfront	Origin

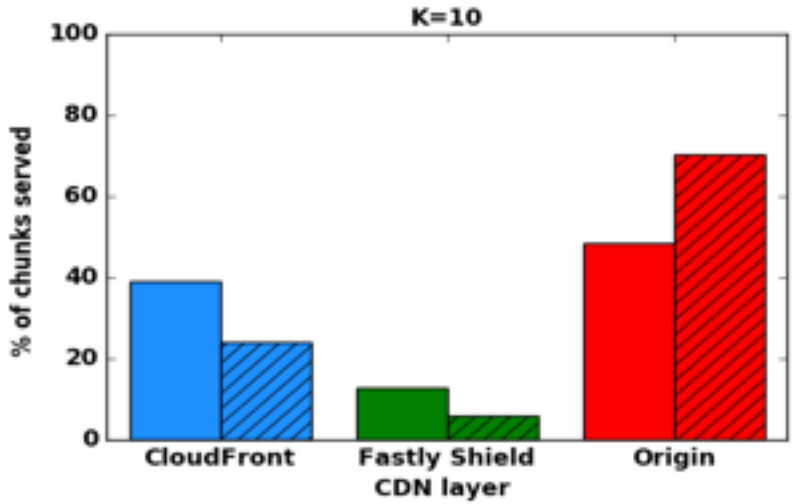
Finding: Chunks in the same video session are served by different points in the CDN hierarchy for multiple popular video publishers and multiple CDNs

Results with Vimeo (popular US video publisher and different CDNs

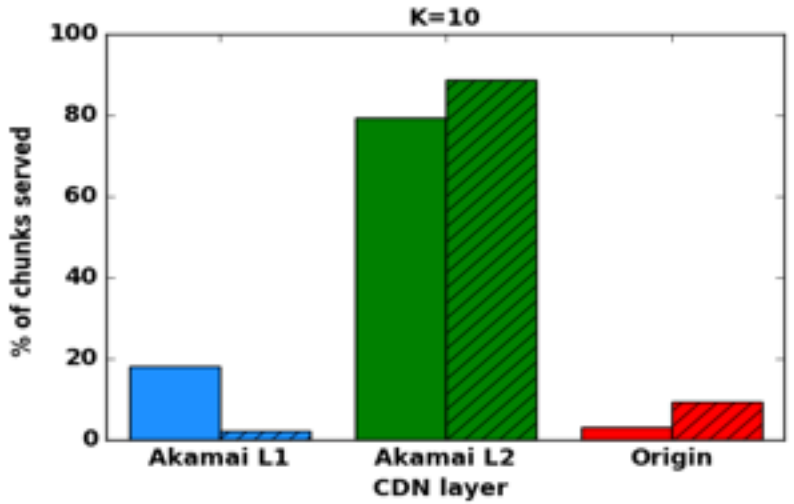




Summary of more detailed analysis



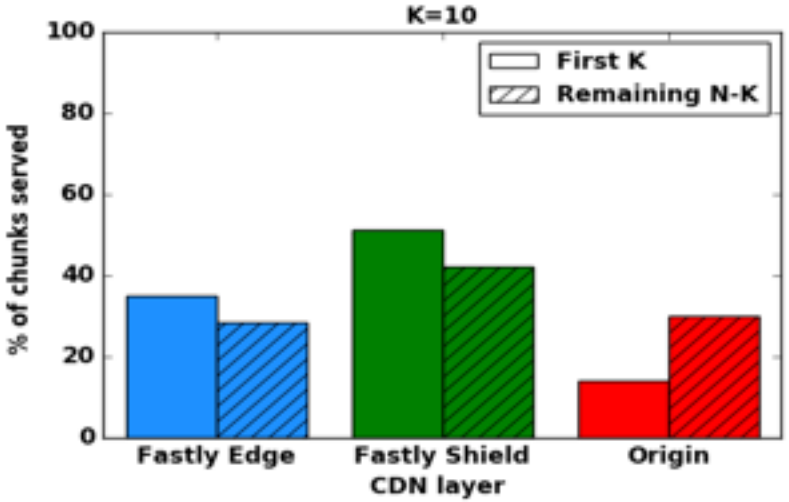
Twitch: CloudFront/Fastly.



Vimeo: Fastly.

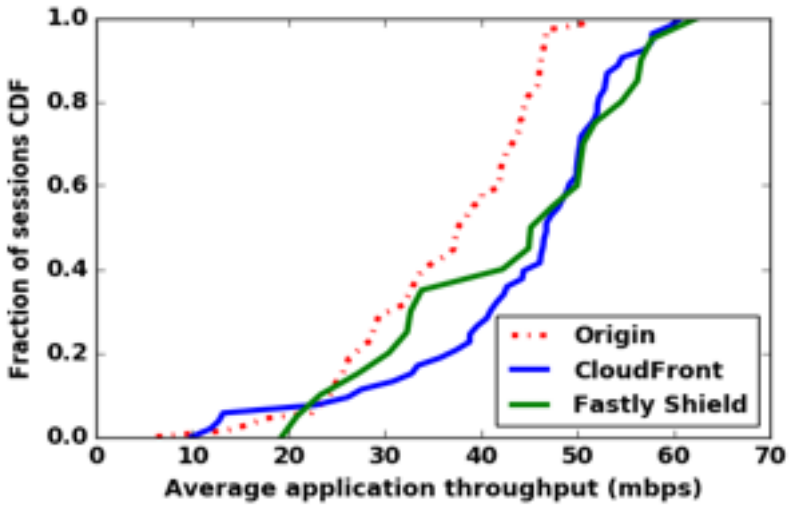
- Objects at start and middle of session both show diversity in terms of where they are served from

- Significant switching in serving location could occur from chunk to chunk

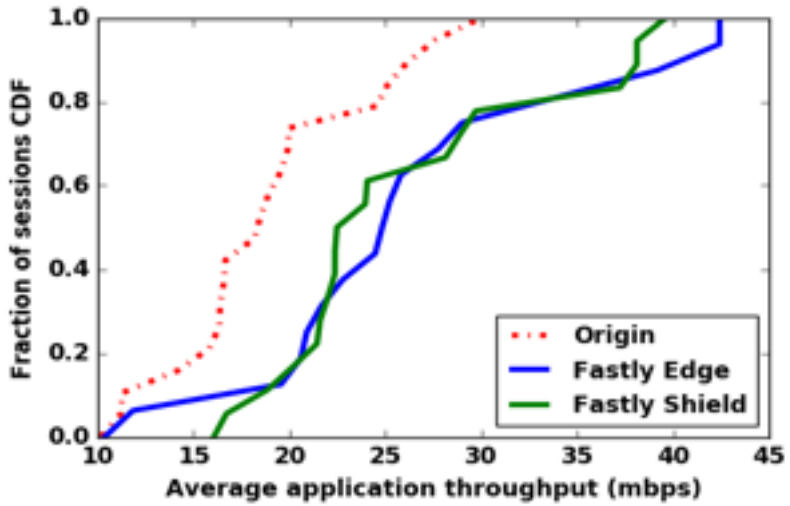


Vimeo: Akamai.

Serving location can have a significant impact on application perceived throughput



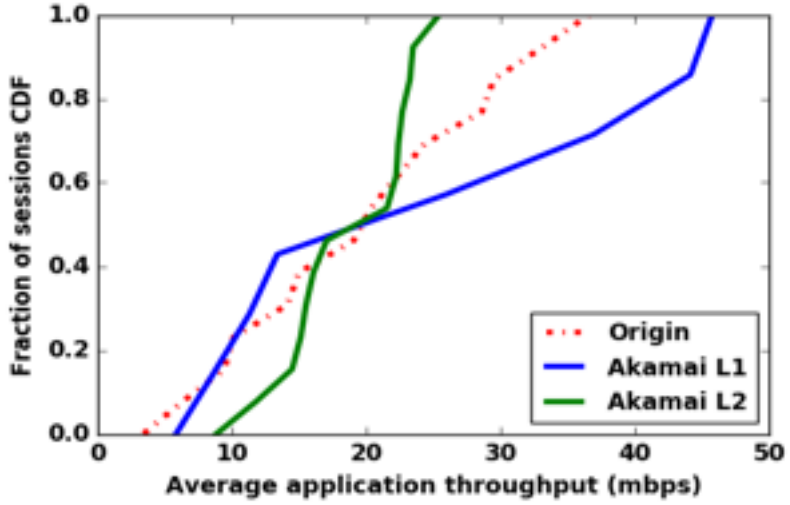
Twitch: CloudFront/Fastly.



Vimeo: Fastly.

Caveat:

Precise trends are sensitive to client location, publisher, CDN, whether certain CDN optimizations are turned on

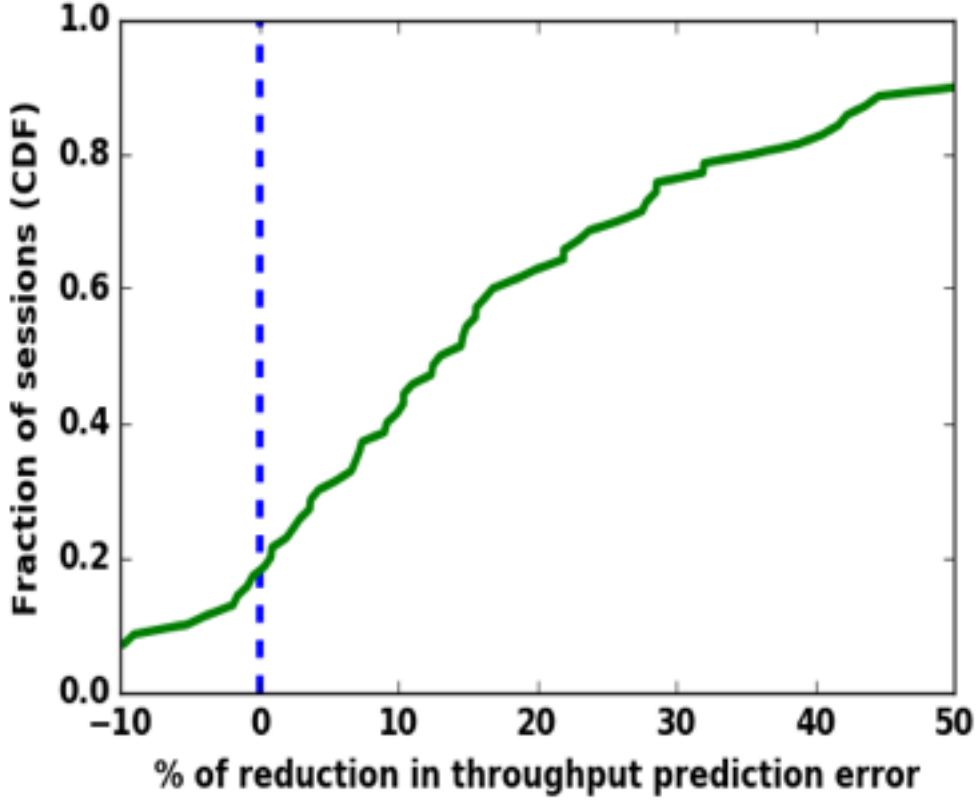


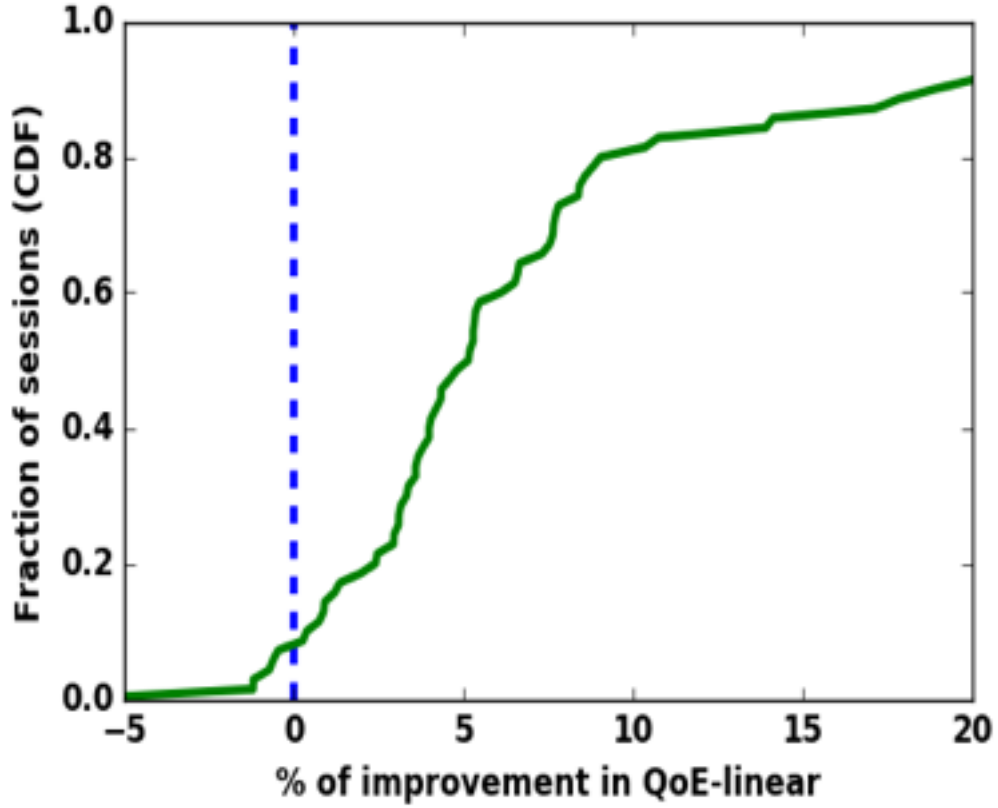
Vimeo: Akamai.

Implications for ABR algorithms

- Today's ABR algorithms are agnostic of CDN hierarchy
  - Chunks can be served from different locations
  - Chunk throughput can vary significantly based on serving location
  - Agnostic => source of errors
  - Can explicit awareness of CDN hierarchy help to optimize ABR performance?
- Study in the context of Carnegie Mellon's ABR algorithm (Sigcomm 2015) that uses a Model Predictive Controller (MPC)
  - Evaluate benefits of explicit hints that indicate where next chunk is served from

Emulation results: does making ABR algorithms CDN-aware help?





Over 80% of video sessions see reduction in throughput prediction error

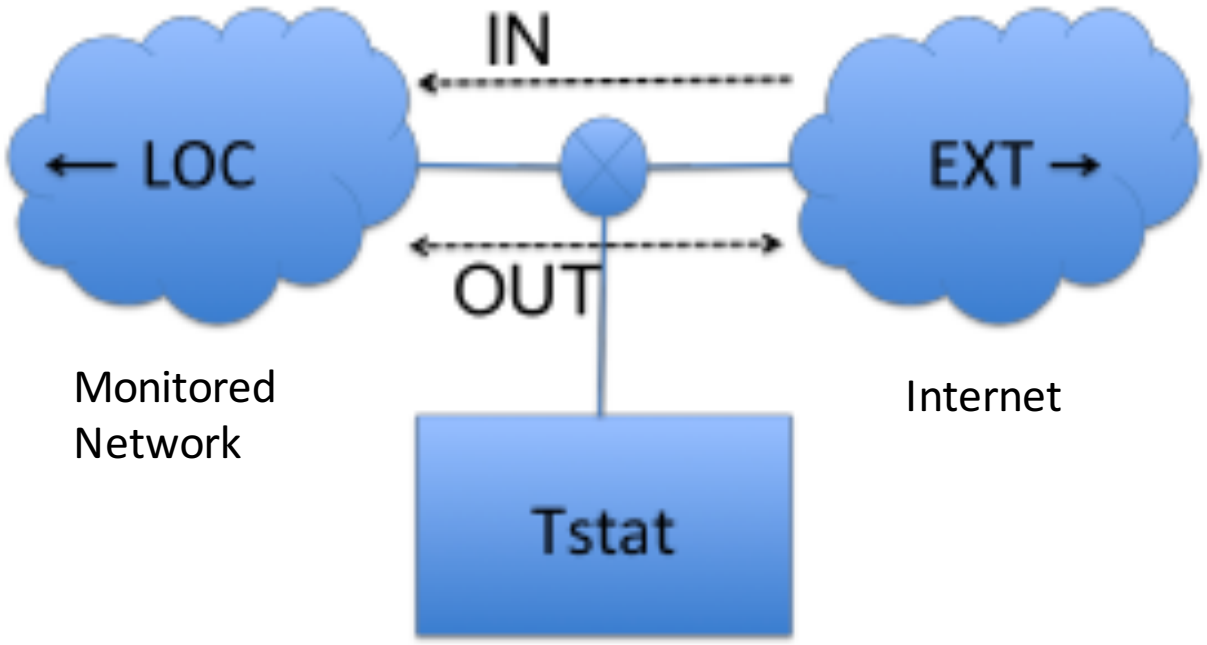
Nearly 90% see an improvement in the QoE metric (which combines bit rate, rebuffering and bit rate change)

Research #2. Sharing information across video flows going through the CDN edge

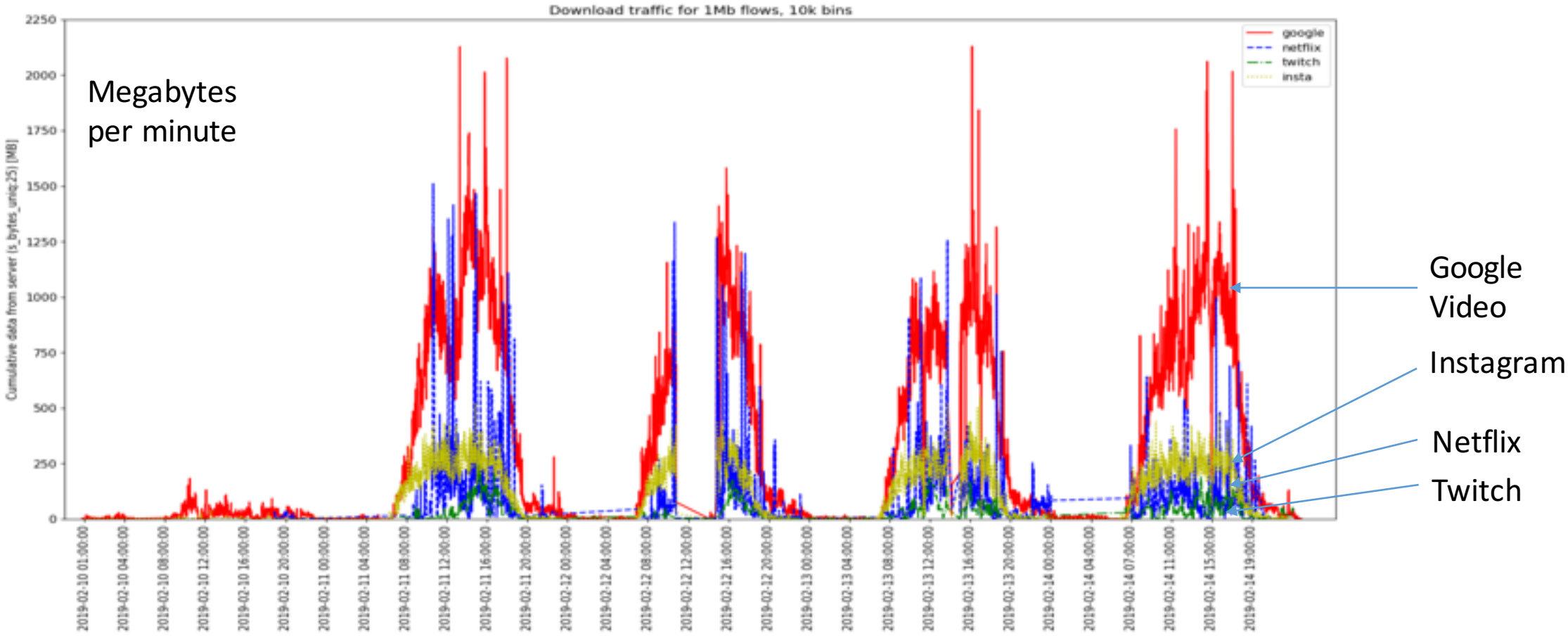
- Observe traffic across multiple flows, not just given flow
- Predict throughput taking into account such visibility
  - Does the global view help? How much? What are the challenges?
- Conducting such research difficult: Need access to data
  - Collaboration with Professor Marco Mellia, and his group (Politecnico Di Torino)
  - Polito expertise: Traffic data collection (tstat), traffic characterization
  - Data being collected on going basis in the Polito campus network

Data collection (led by Polito team)

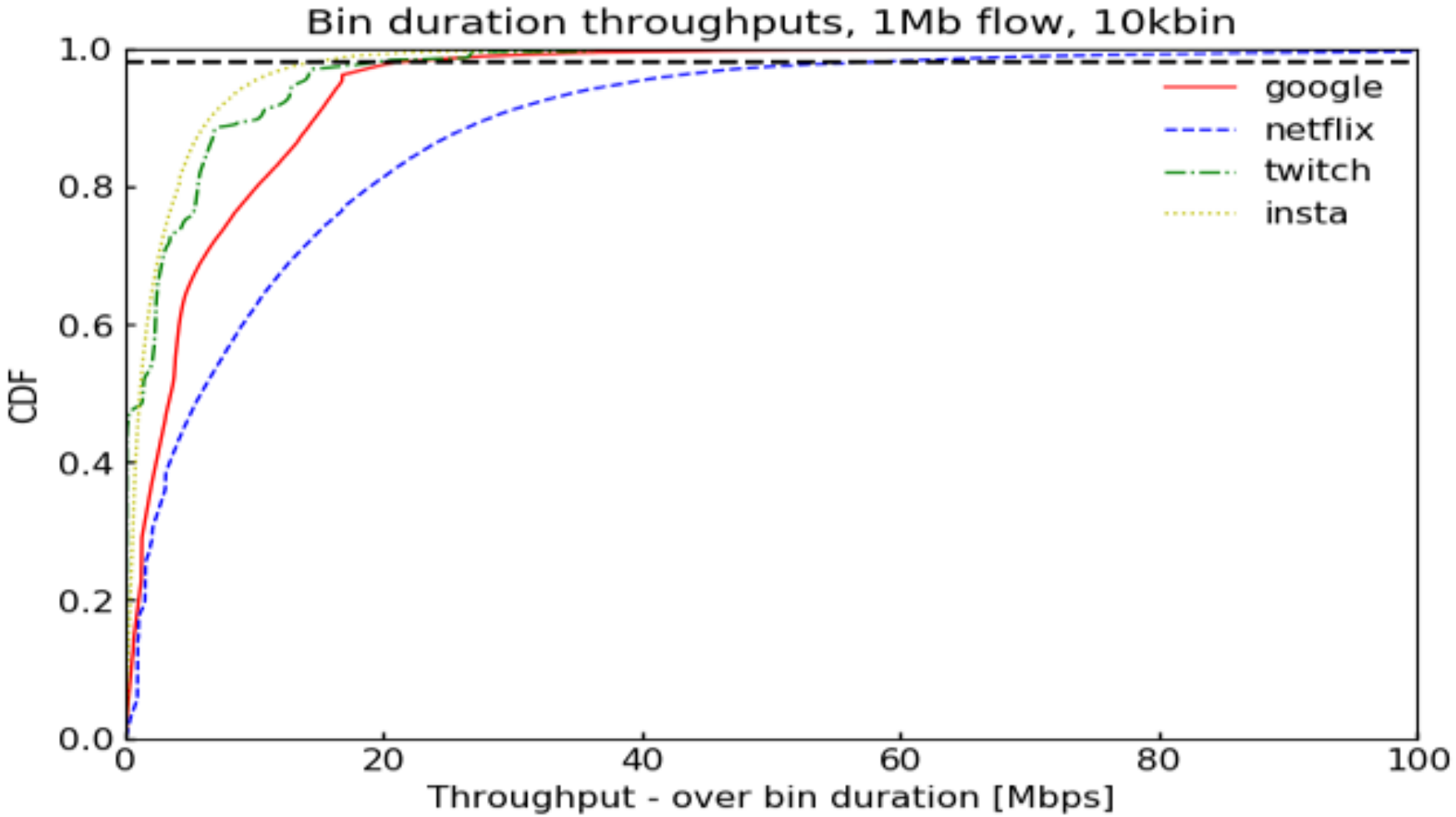
- Passive traffic monitoring tool
- Collects statistics for each flow
- Separately collect client to server, server to client
- Statistics include RTT, loss rate, throughput
- Not just average, but also at finer granularities



Data volume over 5 day period in 2019



Throughput of video flows over 1-sec bins

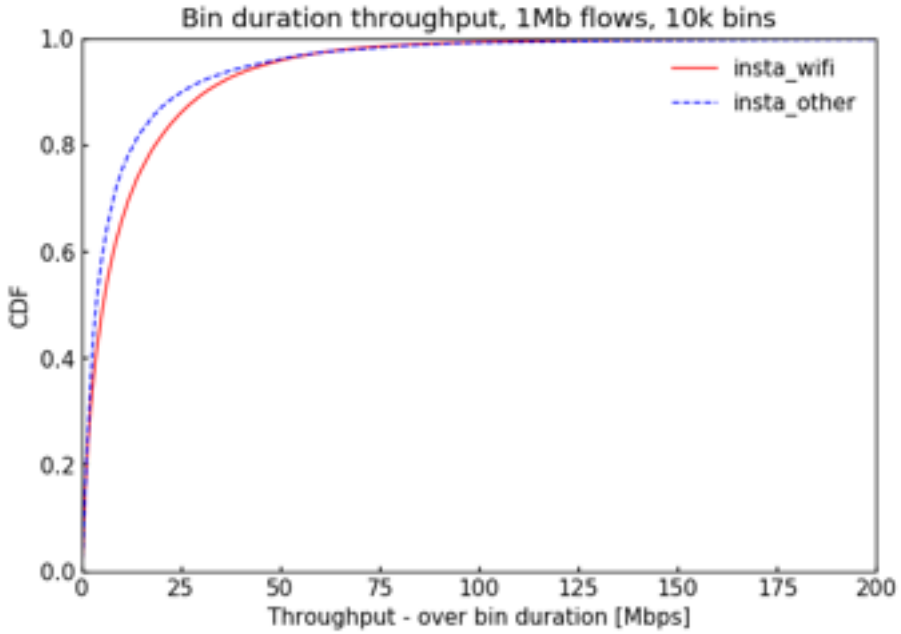


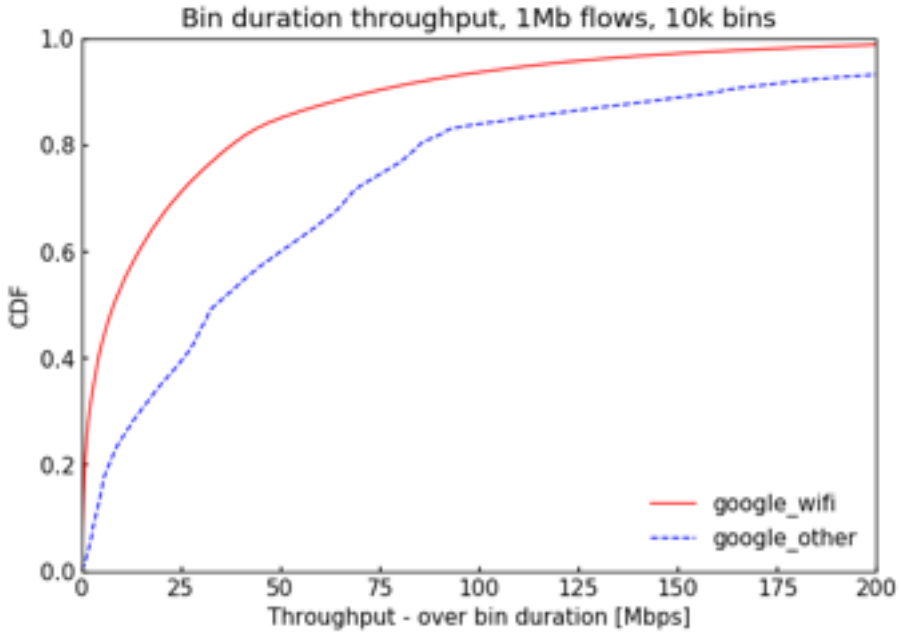
Throughput varies significantly across publishers.

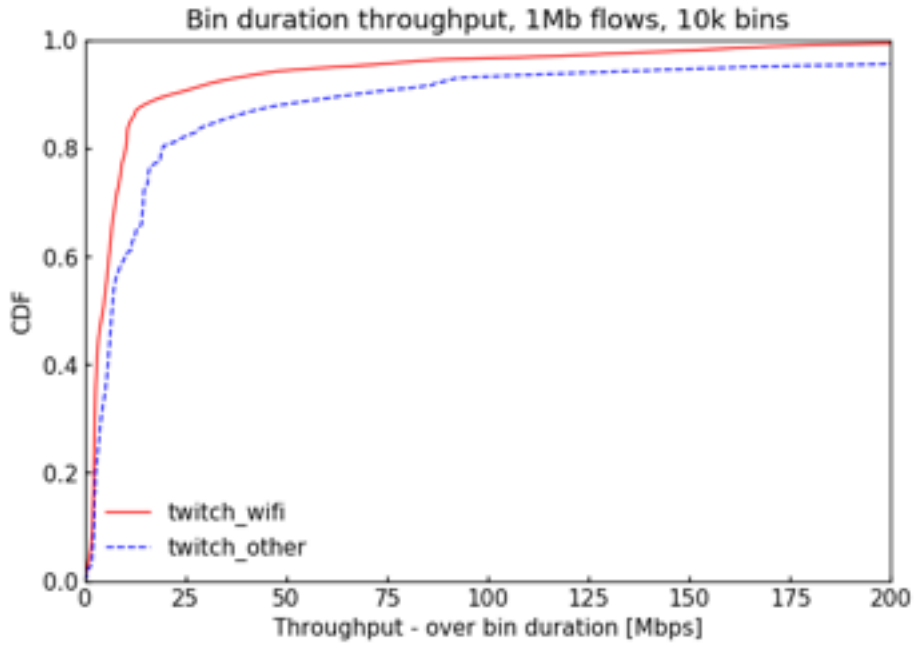
Impacted by server location, Live vs. VoD, Video bitrates, On/Off Patterns

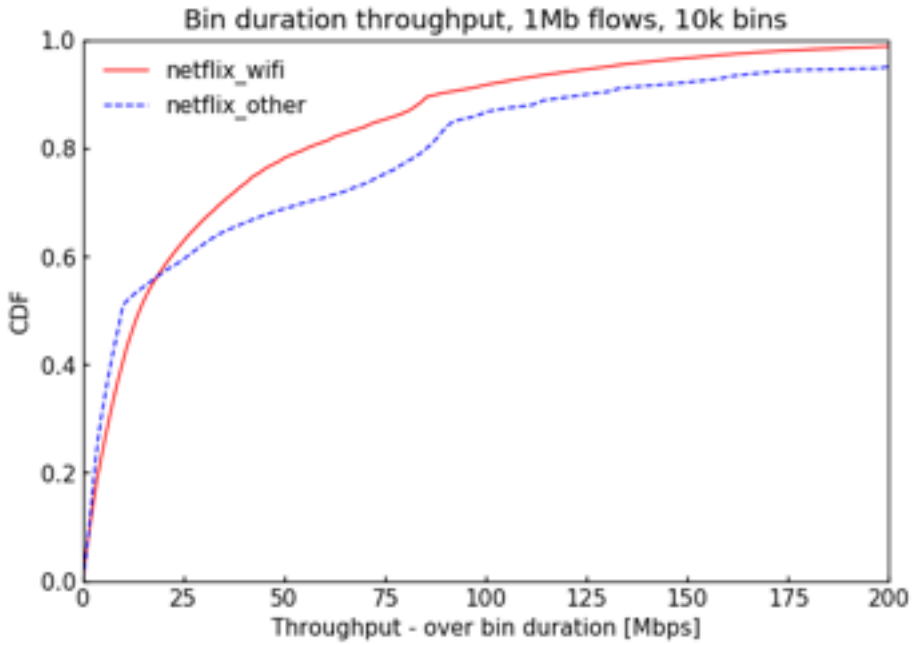
Impact of access technology (WiFI vs. Wired)

WiFi clearly achieves less throughput but differences sensitive to publisher



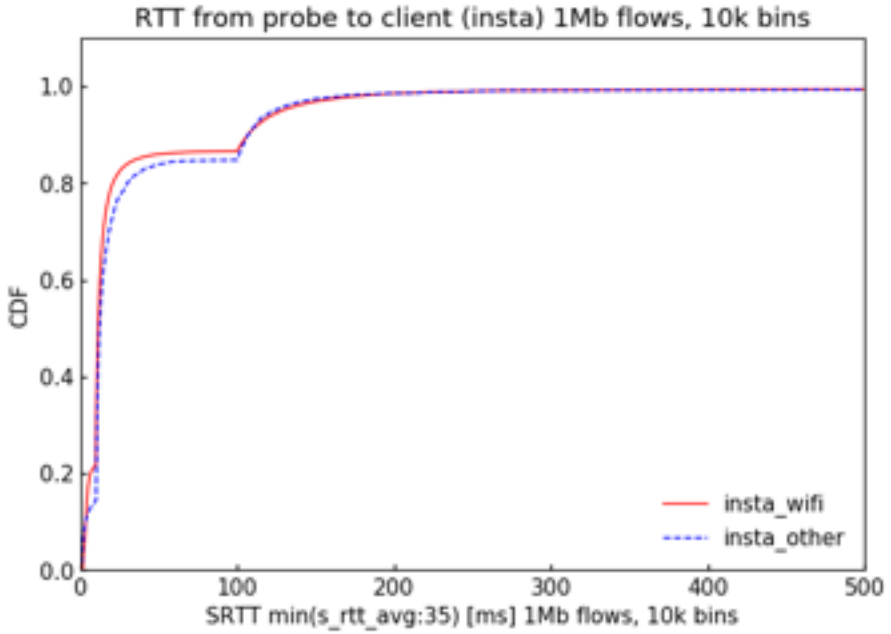


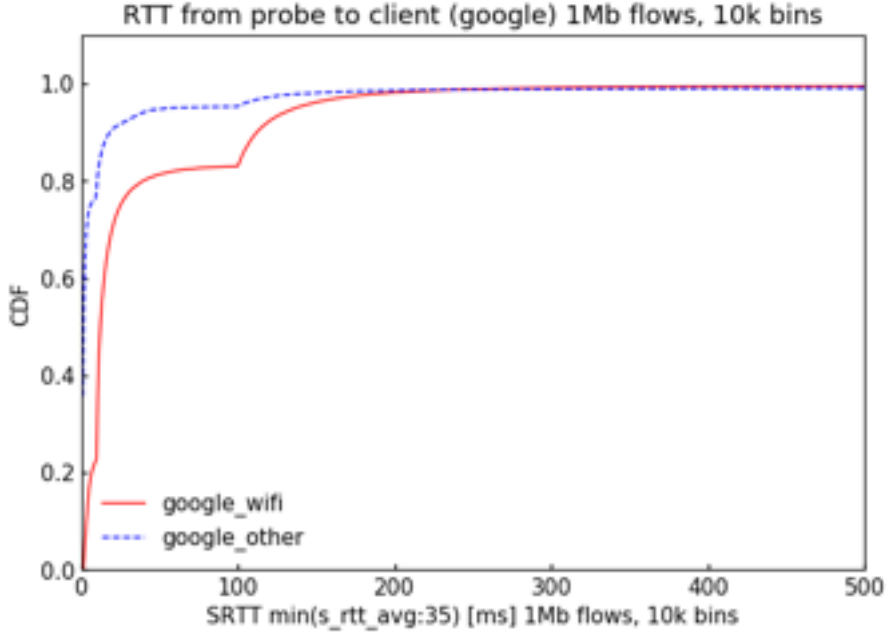


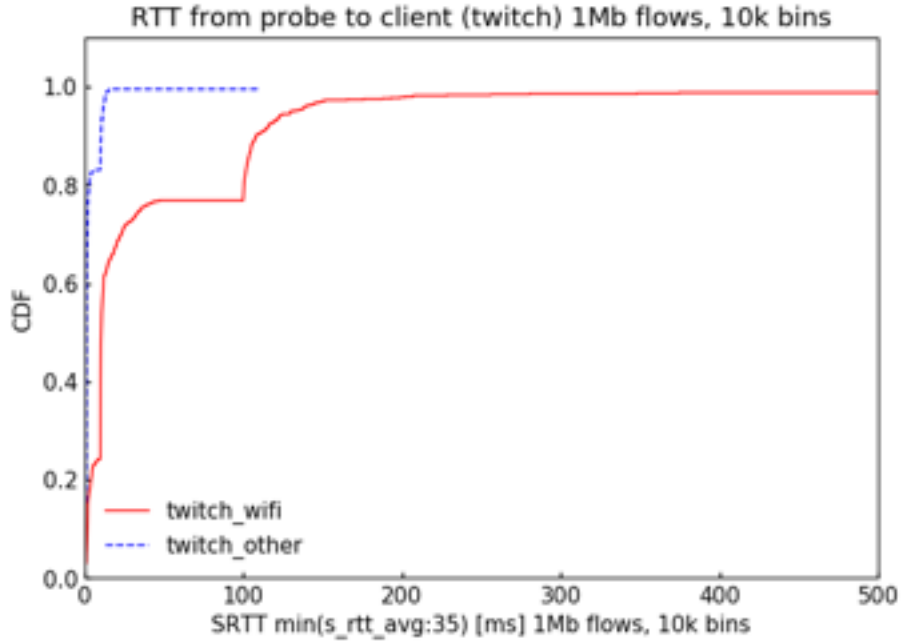


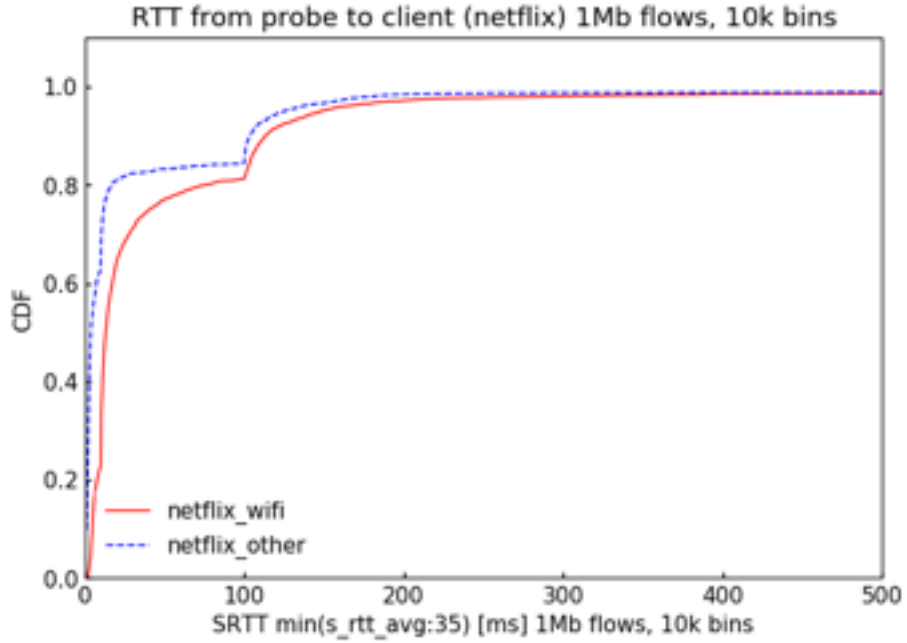
RTT from client to probe:

Wifi clients have significantly higher RTT

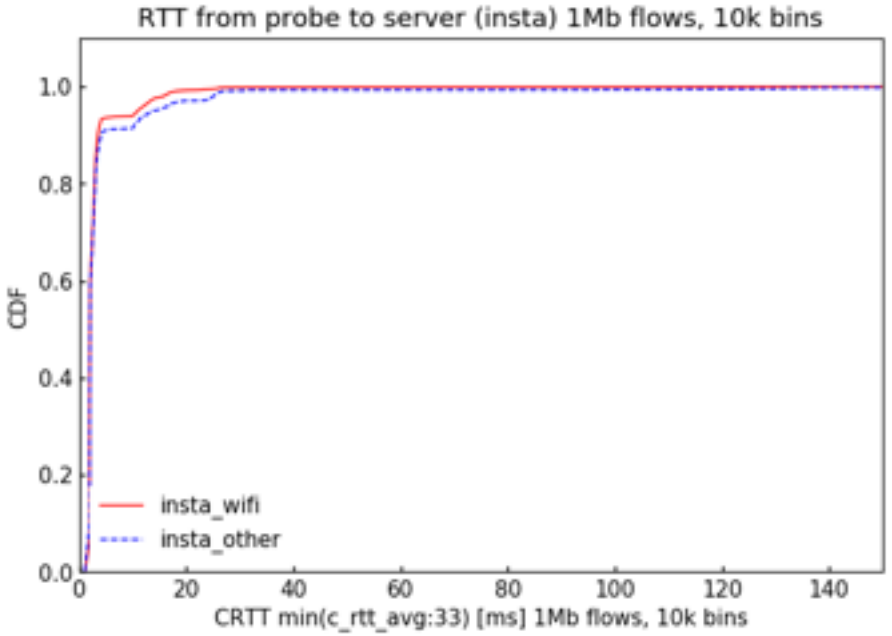


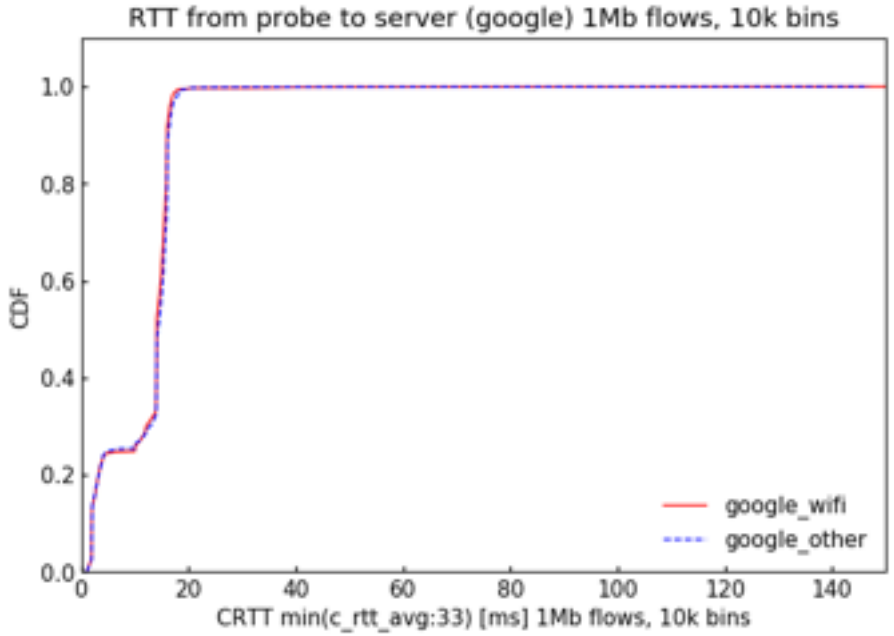


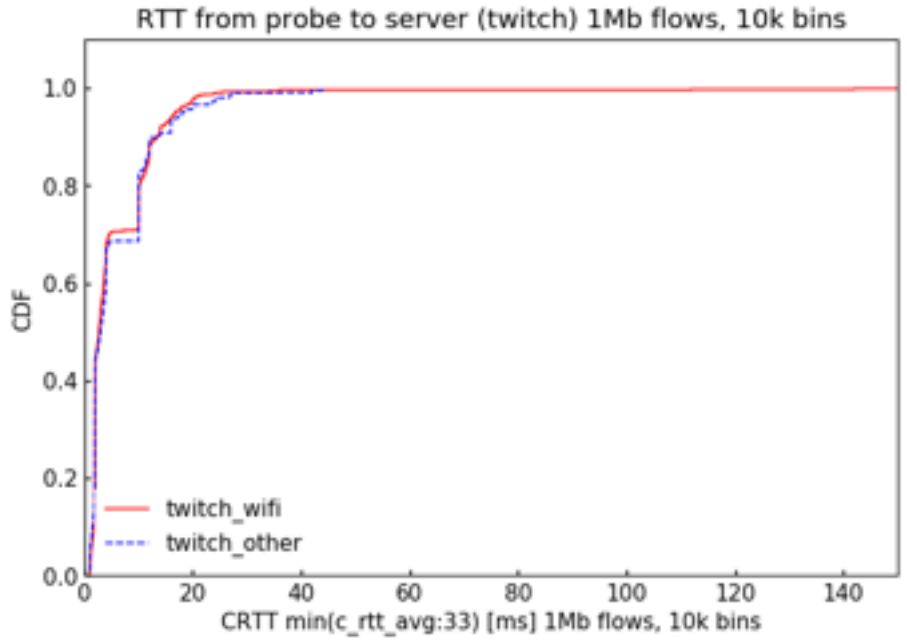


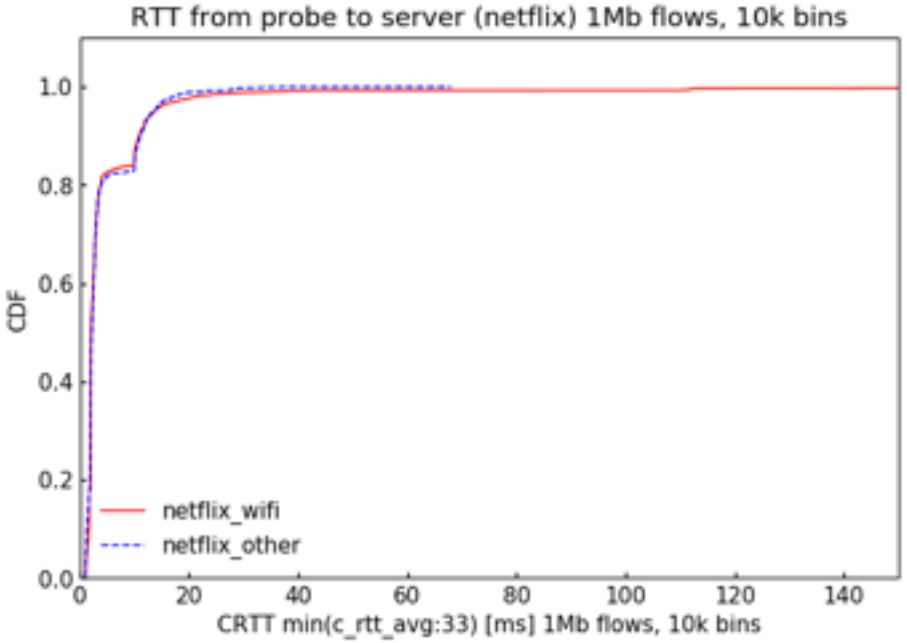


RTT from probe to server. Kinks in the curve correspond to different CDN server placements









Next steps

- Understand the extent to which using information across flows can help prediction
- See what features are needed (e.g., publisher, access technology)
- What are the likely patterns if move to 4K video?

Other Research: Richer throughput prediction frameworks

- State-of-the-art in throughput prediction for video
  - CS2P: Hidden Markov Model based scheme
  - Clusters data based on ISP,CDN etc; then builds a HMM per cluster
  - Does not consider impact of TTFB, video chunk size.
- Our research:
  - Neural network model based on LSTMs
  - Novel architecture to combine static features (e.g., ISP, CDN) with time-varying features (TTFB, throughput)
- Results:
  - Throughput prediction accuracy improved over 23.8% relative to CS2P

---- Backup----