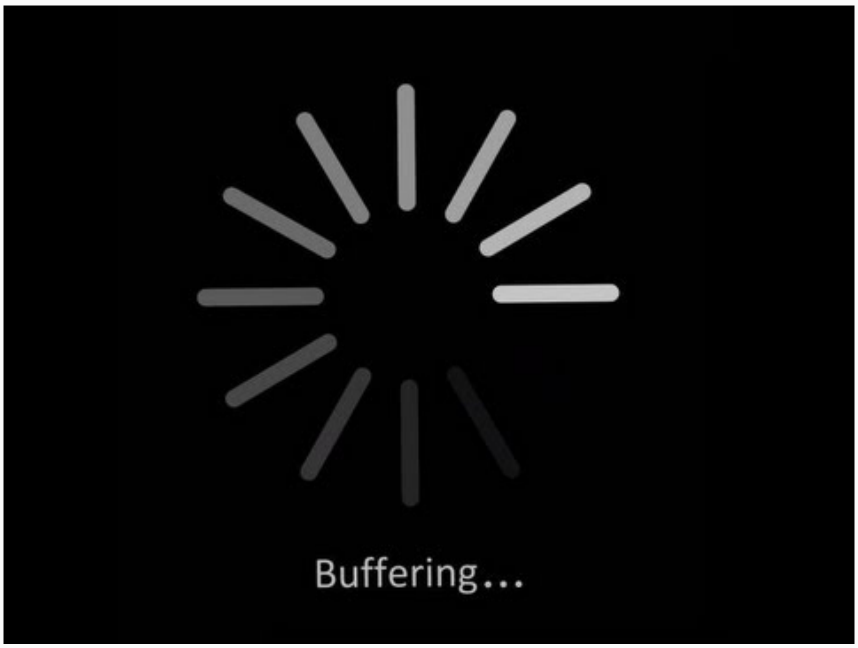
Performance Characterization of a Commercial Video Streaming Service

Mojgan Ghasemi, Akamai Technologies - Princeton University

M Ghasemi, P Kanuparthy, A Mansy, T Benson, and J Rexford
ACM IMC 2016


Buffering...

- First study to measure **e2e** video delivery
- Video makes up **70%** of the traffic!

Location	Findings
CDN	1. Asynchronous disk reads increase server-side delay.
	2. Cache misses increase CDN latency by order of magnitude.
	3. Persistent cache-miss and slow reads for unpopular videos.
	4. Higher server latency even on lightly loaded machines.
Network	1. Persistent delay due to physical distance or enterprise paths.
	2. Higher latency variation for users in enterprise networks.
	3. Packet losses early in a session have a bigger impact.
	4. Bad performance caused more by throughput than latency.
Client	1. Buffering in client download stack can cause re-buffering.
	2. First chunk of a session has higher download stack latency.
	3. Less popular browsers drop more frames while rendering.
	4. Avoiding frame drops needs min of 1.5 sec/sec download rate
	5. Videos at lower bitrates have more dropped frames

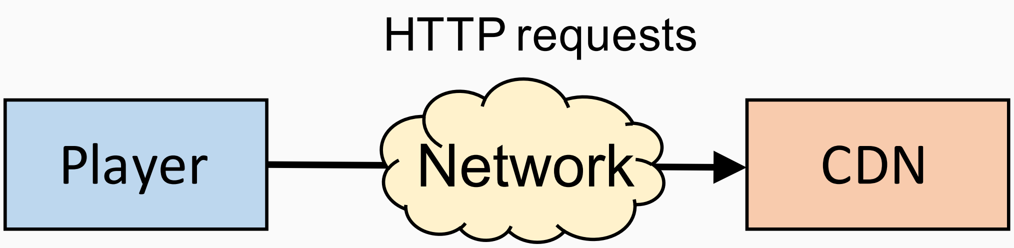
Yahoo's Video Streaming System

Yahoo's Video Streaming System

- Client receives the manifest

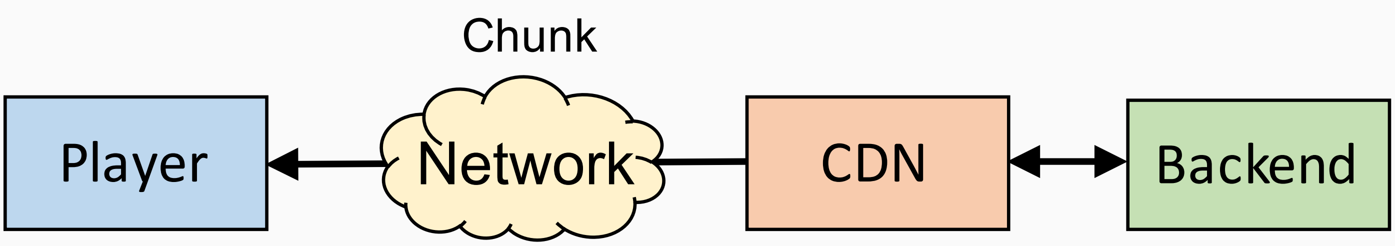
Yahoo's Video Streaming System

- HTTP requests for chunks share a TCP connection
- Each chunk is 6 seconds



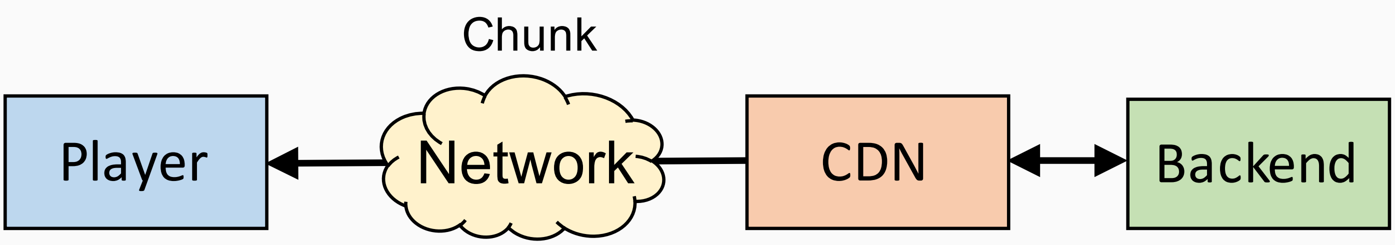
Yahoo's Video Streaming System

- CDN servers use Apache Traffic Server (ATS), LRU policy



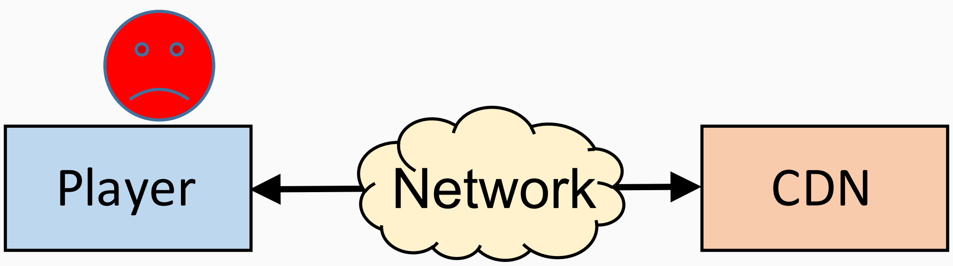
Yahoo's Video Streaming System

- Chunks pass client's "download" and "rendering" stack



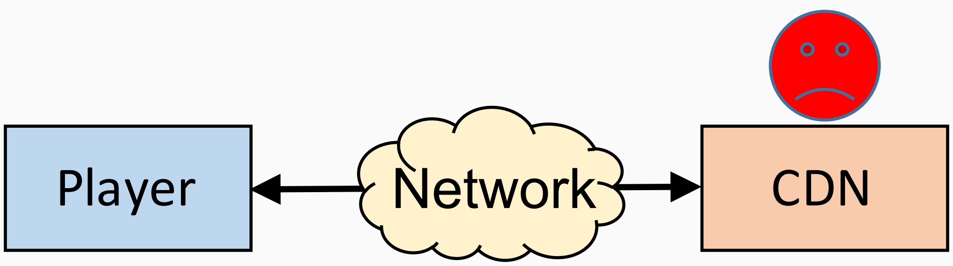
Our Goal

Identify performance problems that impact video



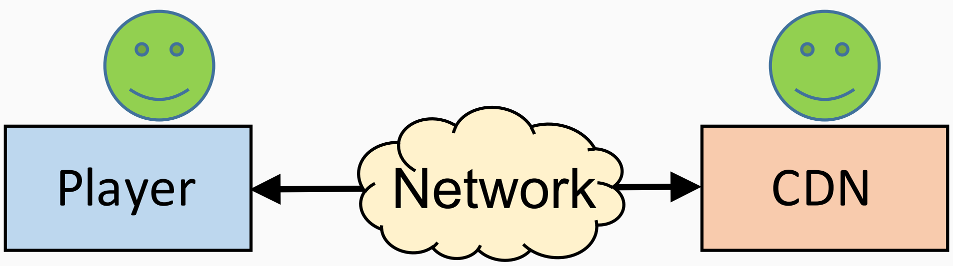
Our Goal

Identify performance problems that impact video



Our Goal

Identify performance problems that impact video



A content provider (e.g., Yahoo) controls "both sides"

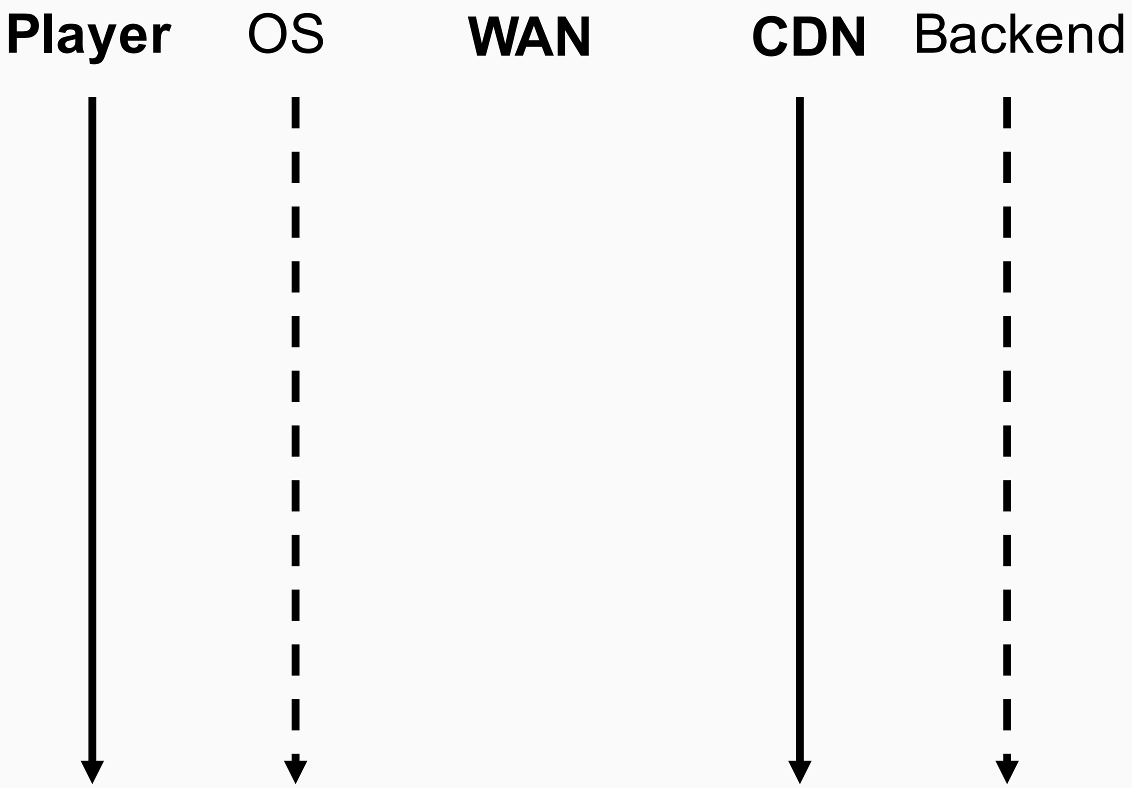
Our Approach: e2e Per-chunk Measurement

- **End-to-end**
  - Instrumenting both sides (player, CDN servers)
- **Per-chunk**
  - Unit of decision making (e.g., bitrate, cache hit/miss)
  - Sub-chunk is too expensive
- **TCP statistics**
  - Sampled from CDN host's kernel
  - Operational at scale

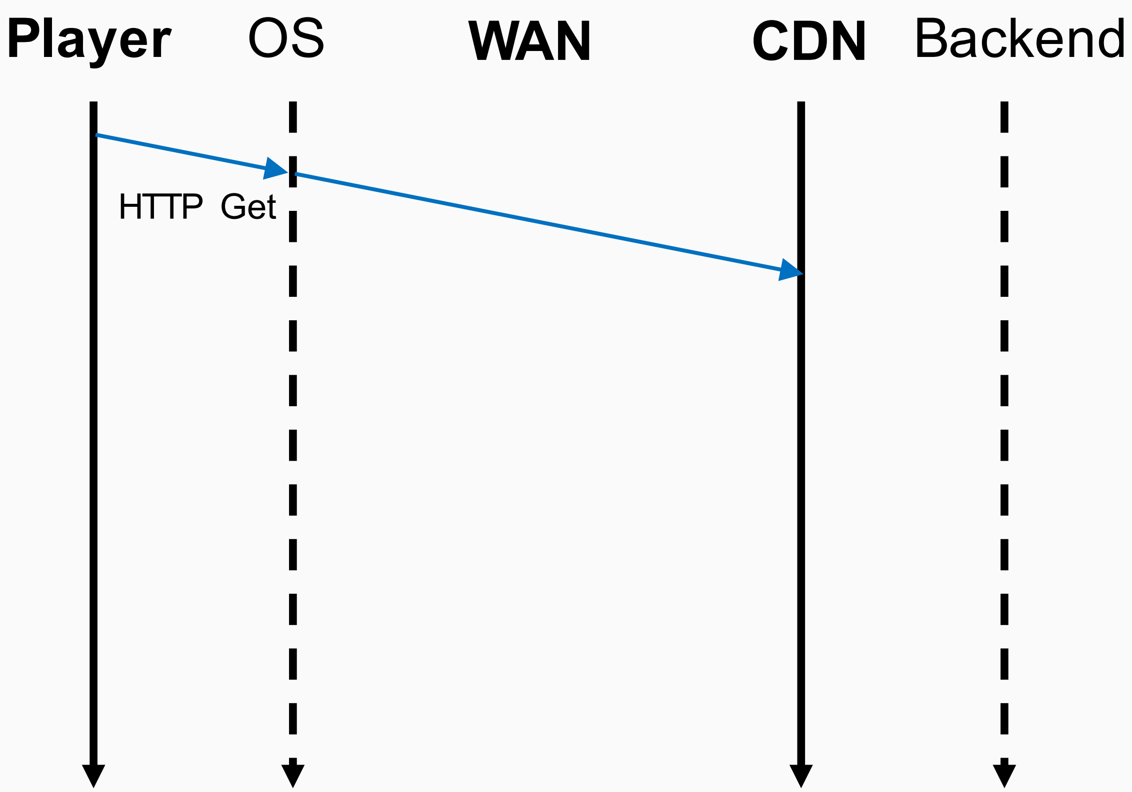
Our Approach: e2e Per-chunk Measurement

Player OS **WAN** **CDN** Backend

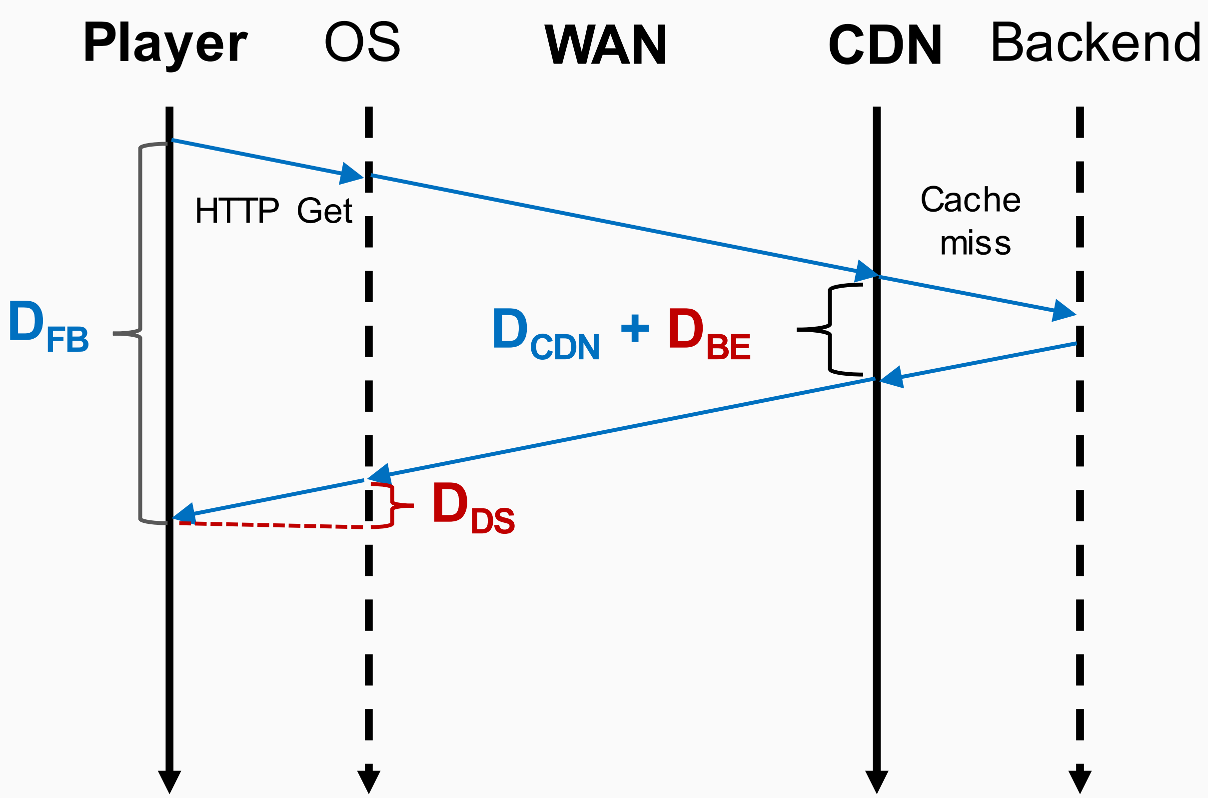
Our Approach: e2e Per-chunk Measurement



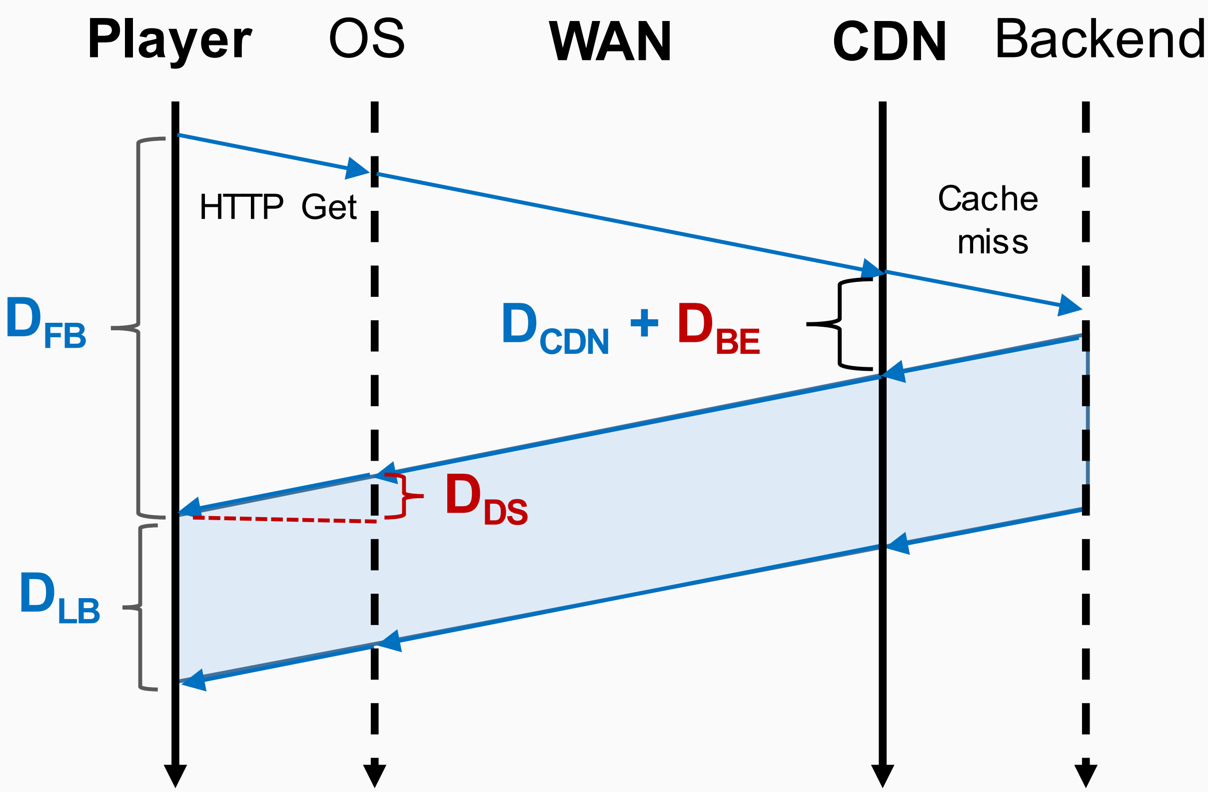
Our Approach: e2e Per-chunk Measurement



Our Approach: e2e Per-chunk Measurement



Our Approach: e2e Per-chunk Measurement



Studying QoE Factors Individually

Factors:

- Video startup time
- Rebuffering rate
- Video quality (bitrate, framerate)

We look at individual metrics, because:

- Type of content
- Length of video

Outline

- Introduction
- Measurement Dataset
- Server-side Problems
- Network Performance Problems
- Client's Performance Problems
- Take-aways and Conclusions

Our Dataset: Yahoo Videos

Yahoo Videos

YAHOO! NEWS

YAHOO! SPORTS

YAHOO! FINANCE

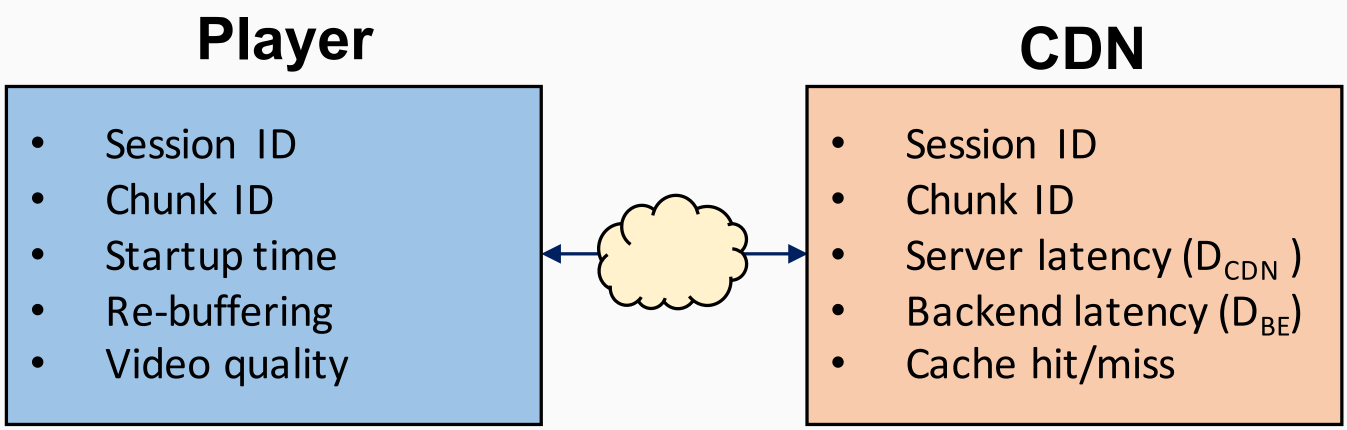
Our Dataset

- **VoD Dataset:**
  - Over 18 days, Sept 2015
  - 85 CDN servers across the US
  - 65 million VoD sessions, 523m chunks
- **Users:**
  - Non-mobile users, no proxy
  - Predominantly in North America (over 93%)
- **Video Streams:**
  - Popularity: 66% of requests for 10% of titles
  - Duration: most videos less than 100 sec

Server-side Performance Problems

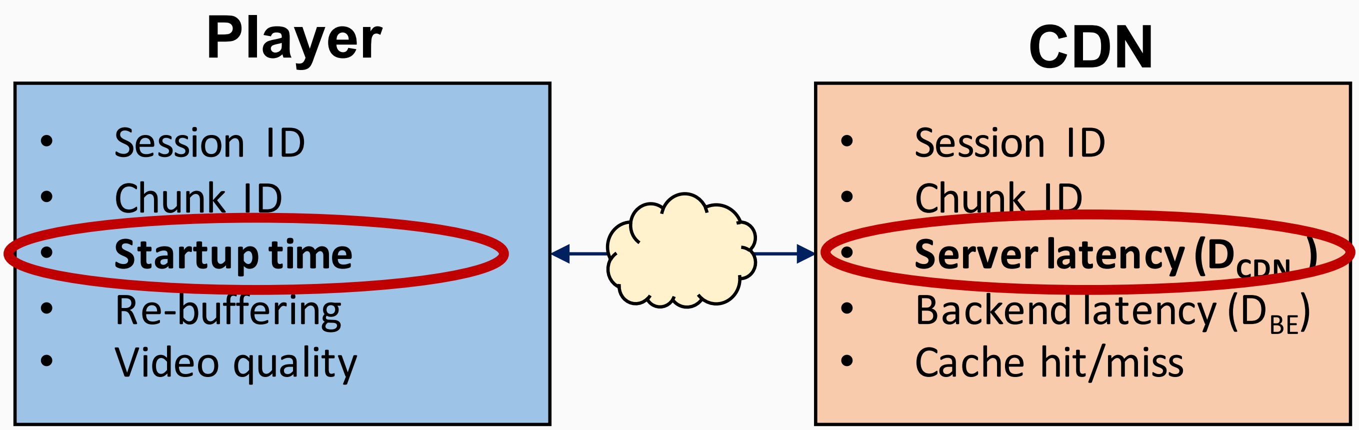
Monitoring CDN Performance

Direct measurement



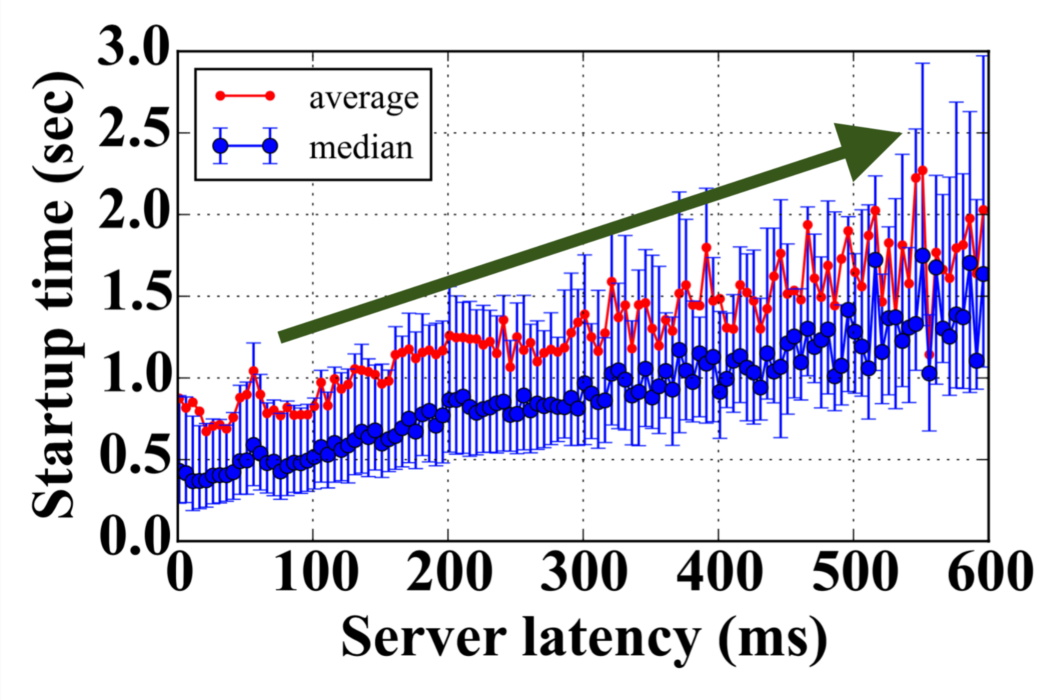
Monitoring CDN Performance

Direct measurement

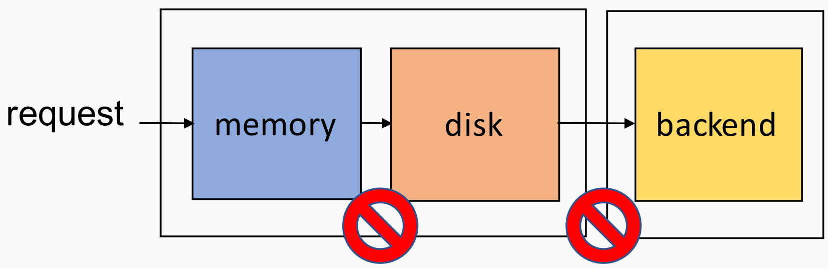


Impact of CDN on QoE

- Only possible via data from “both sides”
- Startup time vs. server latency in first chunk



1. ATS Retry Timer and Cache Misses



- ATS disk/backend retry timer
- Cache misses increase server latency
  - **40X** median, **10X** average
- Server latency can be worse than network
  - Caused by cache misses (**40%** miss rate)

2. Persistent Problems in Unpopular Videos

- Cache misses are **persistent**:
  - Average: 2%
  - After one miss: **60%**
- **Unpopular** titles have significantly higher cache misses

3. Load-Latency Paradox

Paradox: more heavily loaded servers seem to have *lower* latency

- Result of cache-focused mapping
- Less popular contents have higher latency (ATS timer, cache miss)

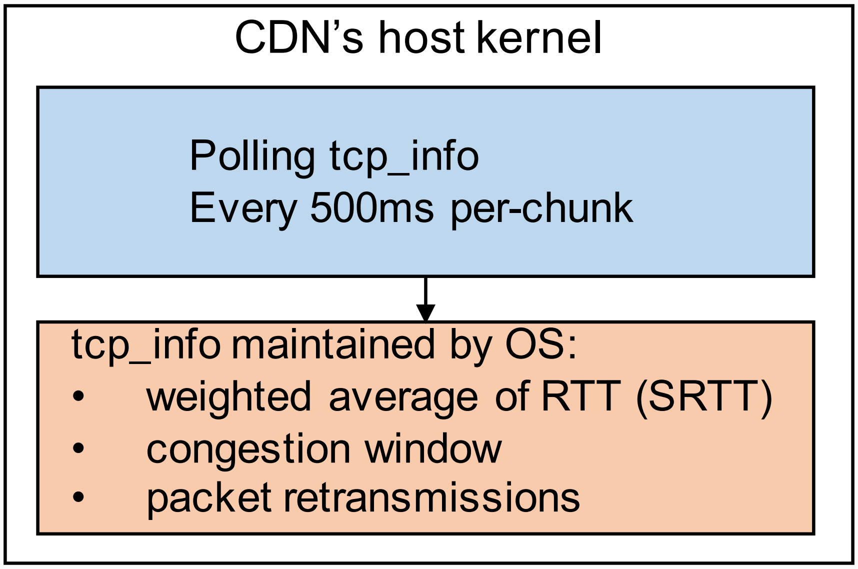
Network Performance Problems

Network Problems

- Manifestation: Packet loss, re-ordering, high latency, high variation in latency, low throughput
- Transient (e.g., spike in latency caused by congestion)
- Persistent (e.g., far away client from a server)

A good ABR may *adapt* to transient problems, but it cannot avoid bad QoE caused by persistent problems.

Network Measurement

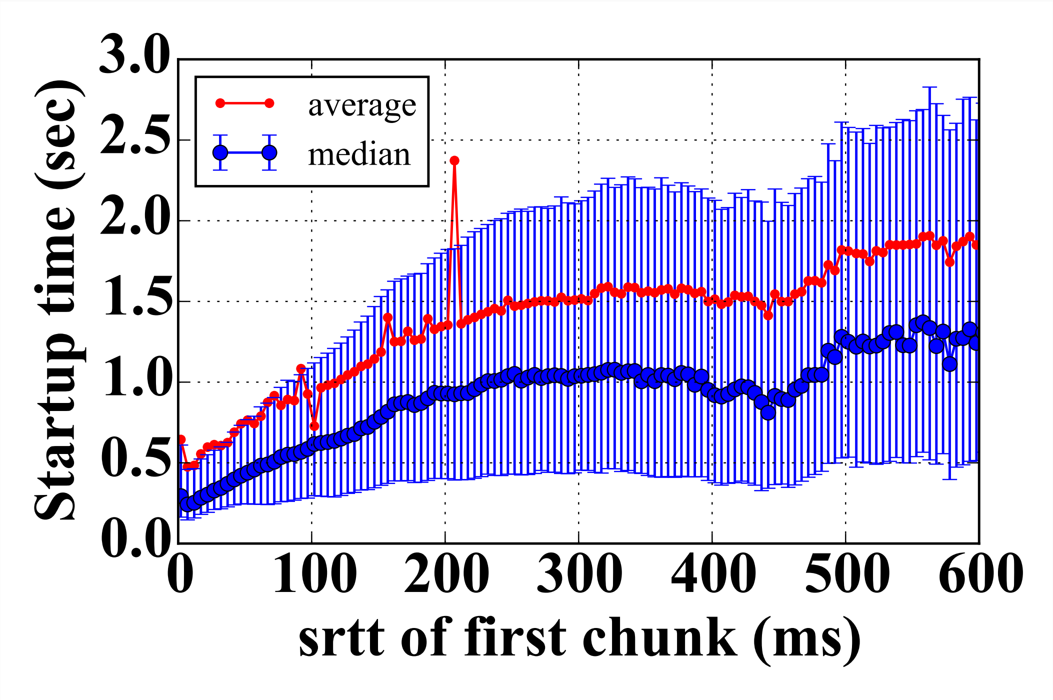


Challenges:

- Smoothed average of RTT: SRTT
- Infrequent network snapshots
- Packet traces cannot be collected

Impact of Network Latency on QoE

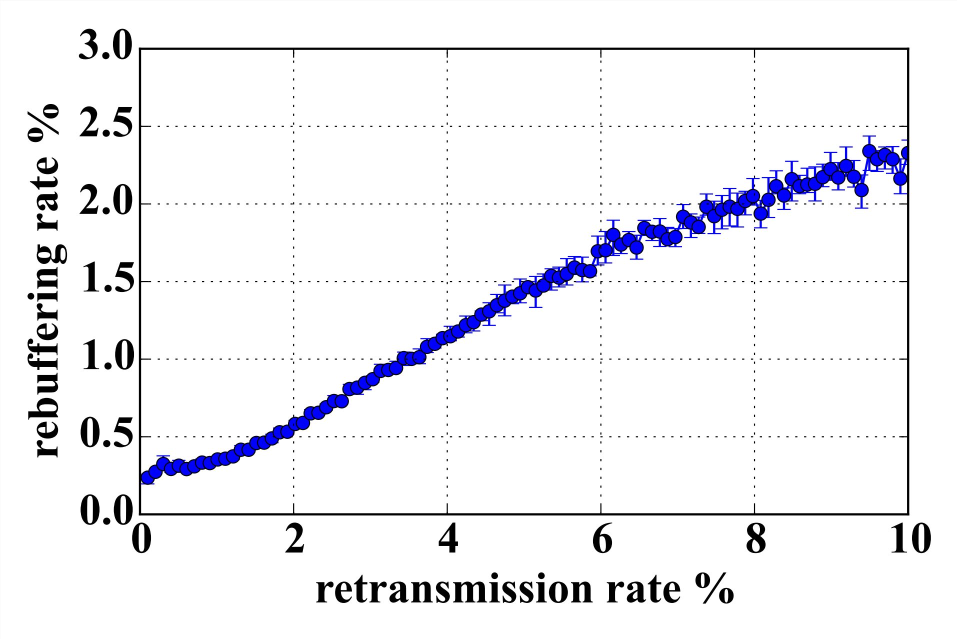
- Data from "both sides" show the impact
- Startup time vs. SRTT of first chunk
- Network latency significantly impacts video startup time



1. Network Latency Problems

- **Persistent high latency:**
  - /24 IP prefixes, recurring in 90^{th} percentile
  - **25%** of prefixes are located in the US, with the majority close to CDN nodes
- **High latency variation:**
  - Enterprise networks have higher latency variation

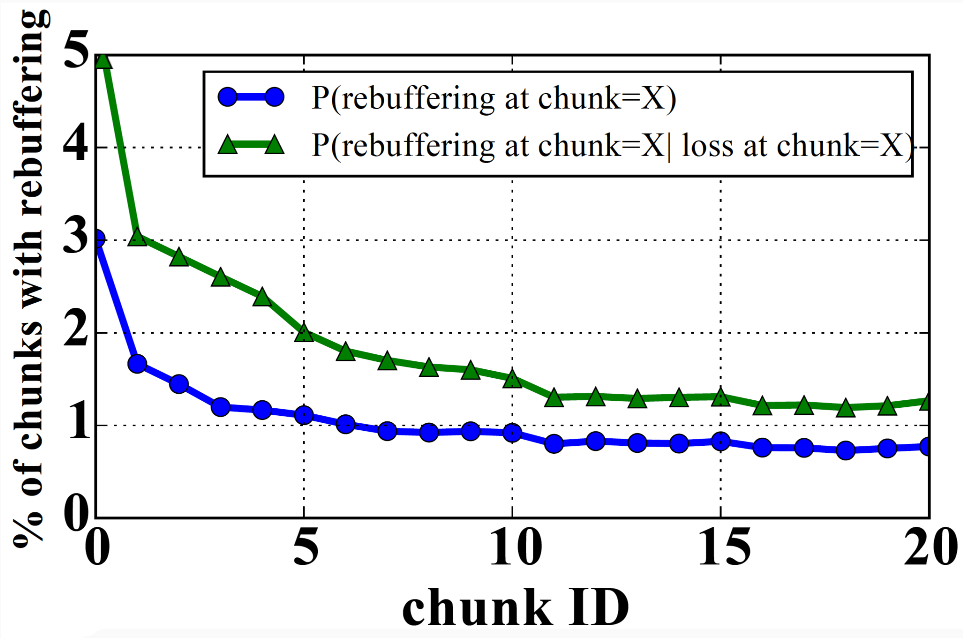
2. Impact of Packet Loss on QoE



- Higher loss rate generally indicates higher rebuffering rate.

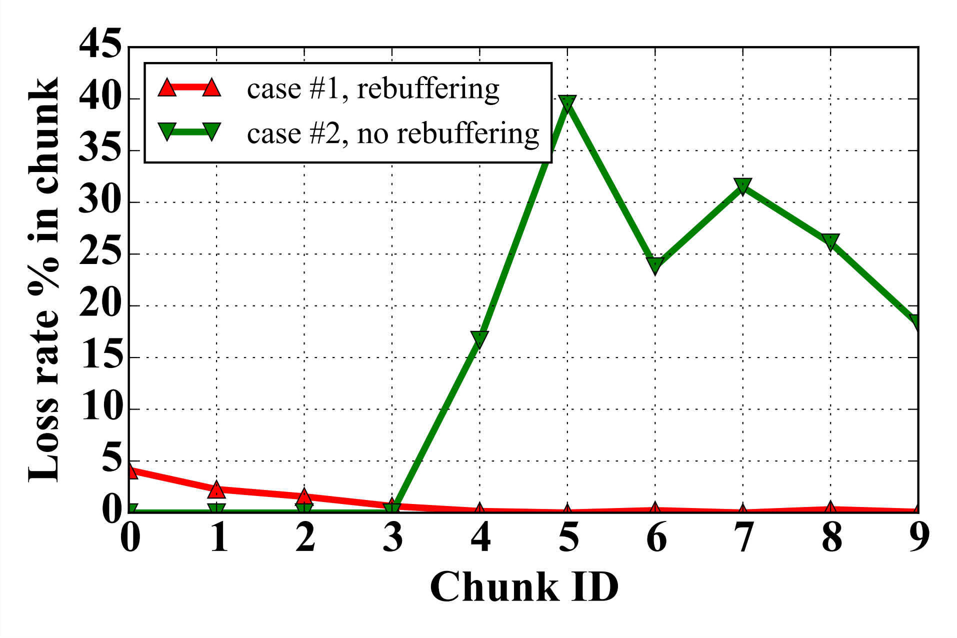
3. Earlier Packet Losses Cause More Rebuffering

- Packet loss is more common in the first chunk (4.5X)
- Packet loss in the first chunk causes more rebuffering



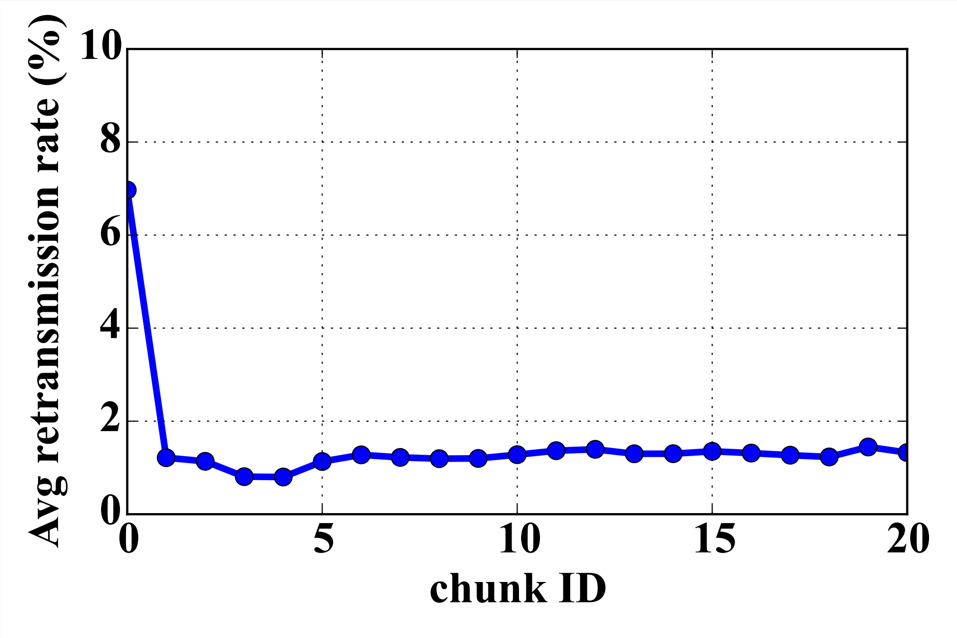
3. Earlier Packet Losses Cause More Rebuffering

Example case for Loss vs. QoE



4. Earlier Packet Losses Are More Common

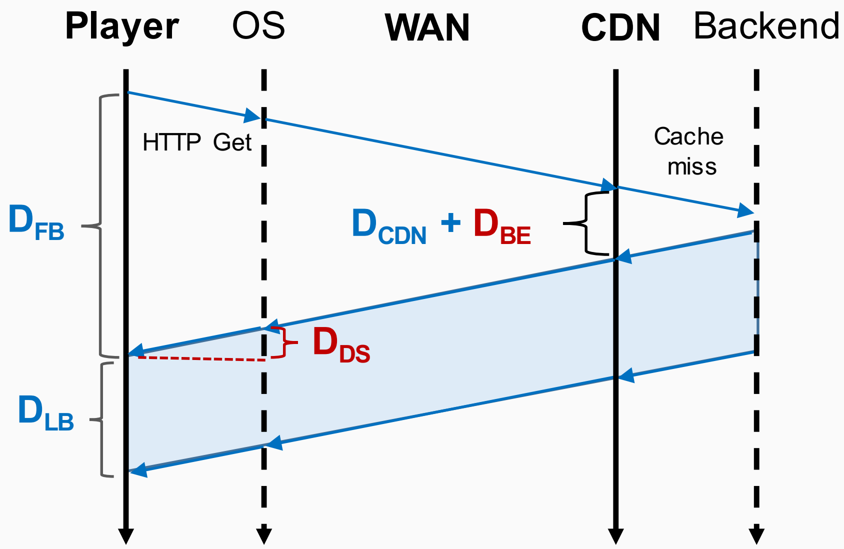
- Bursty nature of packet loss in TCP Slow Start
- Server-side pacing



5. Throughput is a Bigger Problem than Latency

$$perf_{score} = \frac{chunk\ duration}{D_{FB} + D_{LB}}$$

- D_{FB}: measure of latency, D_{LB}: measure of throughput



5. Throughput is a Bigger Problem than Latency

$$perf_{score} = \frac{chunk\ duration}{D_{FB} + D_{LB}}$$

- D_{FB}: measure of latency, D_{LB}: measure of throughput
- $perf_{score} > 1$: More than 1 sec of video delivered per sec
- $perf_{score} < 1$: Less than 1 sec of video per sec

5. Throughput is a Bigger Problem than Latency

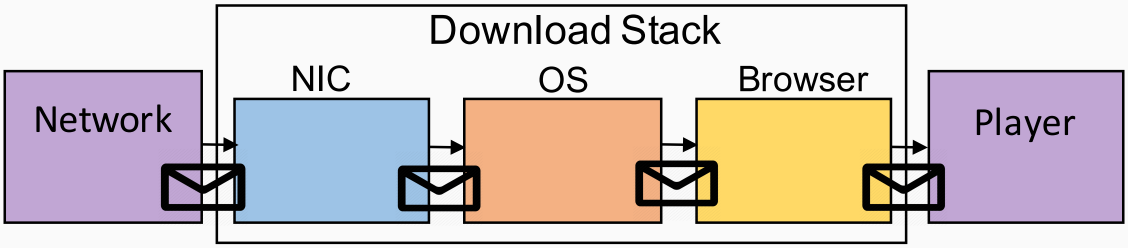
$$perf_{score} = \frac{chunk\ duration}{D_{FB} + D_{LB}}$$

- D_{FB}: measure of latency, D_{LB}: measure of throughput
- $perf_{score} > 1$: More than 1 sec of video delivered per sec
- $perf_{score} < 1$: Less than 1 sec of video per sec

D_{LB} **has a major contribution (orders of magnitude)**

Client's Download Stack Performance Problems

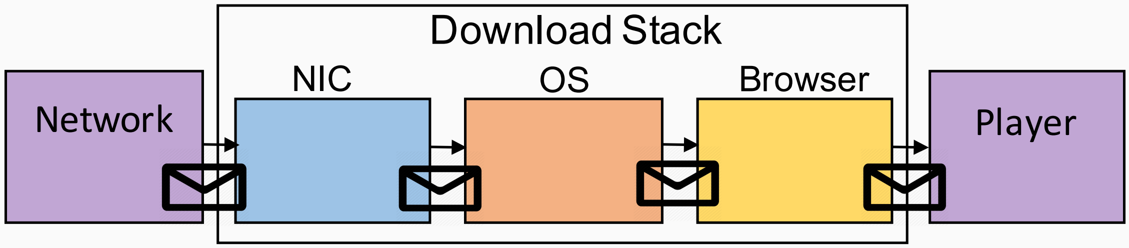
Download Stack Latency



- Cannot observe download stack latency (D_{DS}) directly
- Detecting "outliers"

$$D_{FB_i} > \mu_{D_{FB}} + 2 \cdot \sigma_{D_{FB}}$$

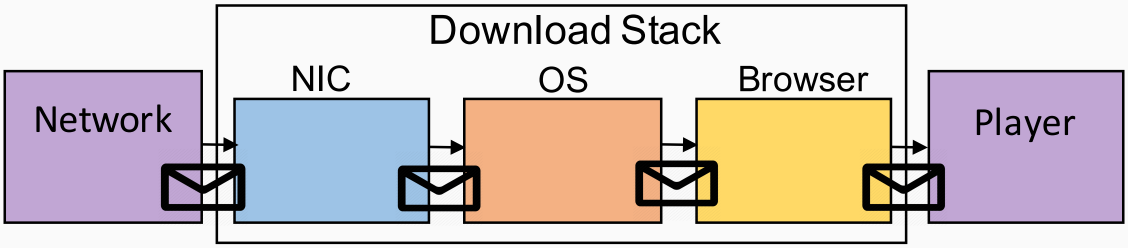
Download Stack Latency



- Cannot observe download stack latency (D_{DS}) directly
- Detecting "outliers"

$$D_{FB_i} > \mu_{D_{FB}} + 2 \cdot \sigma_{D_{FB}}$$
$$TP_{inst_i} > \mu_{TP_{inst}} + 2 \cdot \sigma_{TP_{inst}}$$

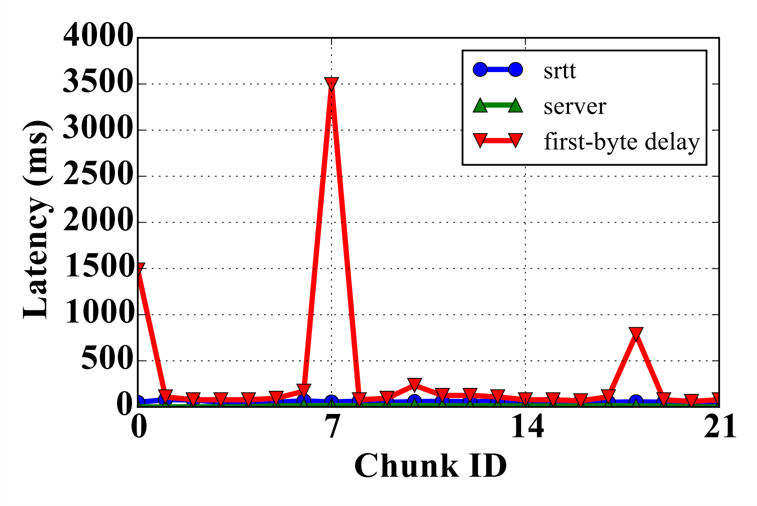
Download Stack Latency

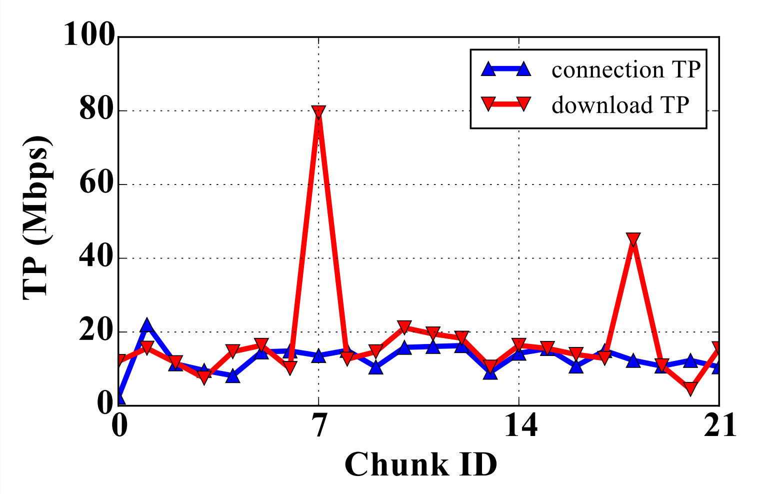


- Cannot observe download stack latency (D_{DS}) directly
- Detecting "outliers"

$$D_{FB_i} > \mu_{D_{FB}} + 2 \cdot \sigma_{D_{FB}}$$
$$TP_{inst_i} > \mu_{TP_{inst}} + 2 \cdot \sigma_{TP_{inst}}$$
$$\textit{Similar network and server performance}$$

Download Stack Latency: Case Study



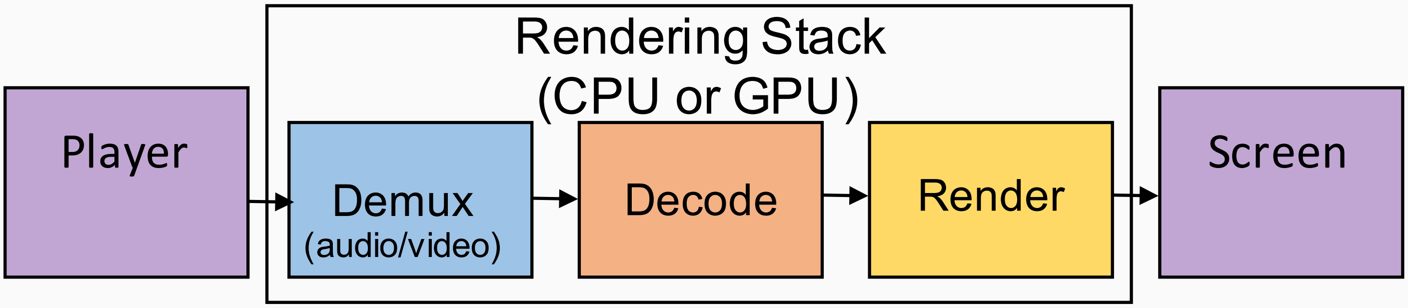


Client's Download Stack Problems

- **Transient:**
  - Outlier: 1.7M chunks (0.32%)
  - **First** chunks have higher D_{DS}
- **Persistent:**
  - In most cases, D_{DS} is higher than network and server latency

Client's Rendering Stack Performance Problems

Rendering Stack

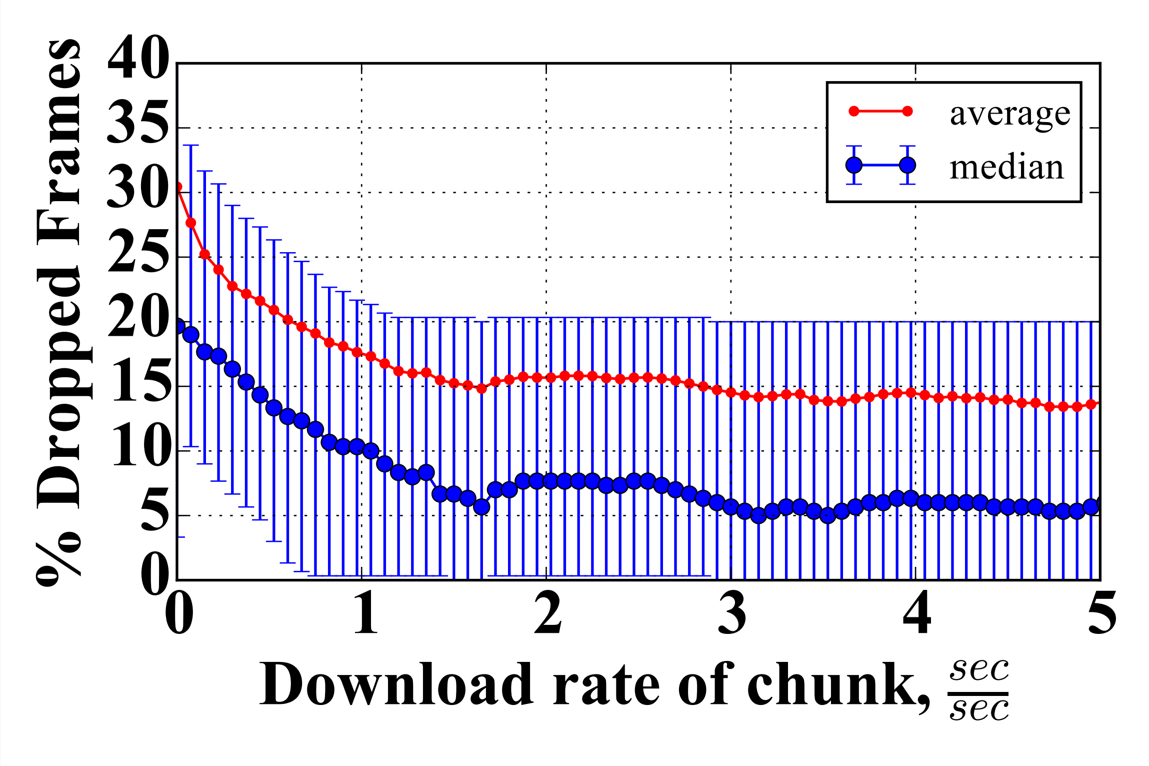


- If CPU is busy, rendering quality drops (high frame drops)
- If video tab is not visible, browser drops frames
- Per-chunk data: *vis* (is player visible?), dropped frames
- Per-session data: OS, browser

1. Good Rendering Requires $1.5\frac{sec}{sec}$ Download Rate

- De-multiplexing, decoding, and rendering takes time.

.



2. Higher Bitrates Show Better Rendering

Paradox:

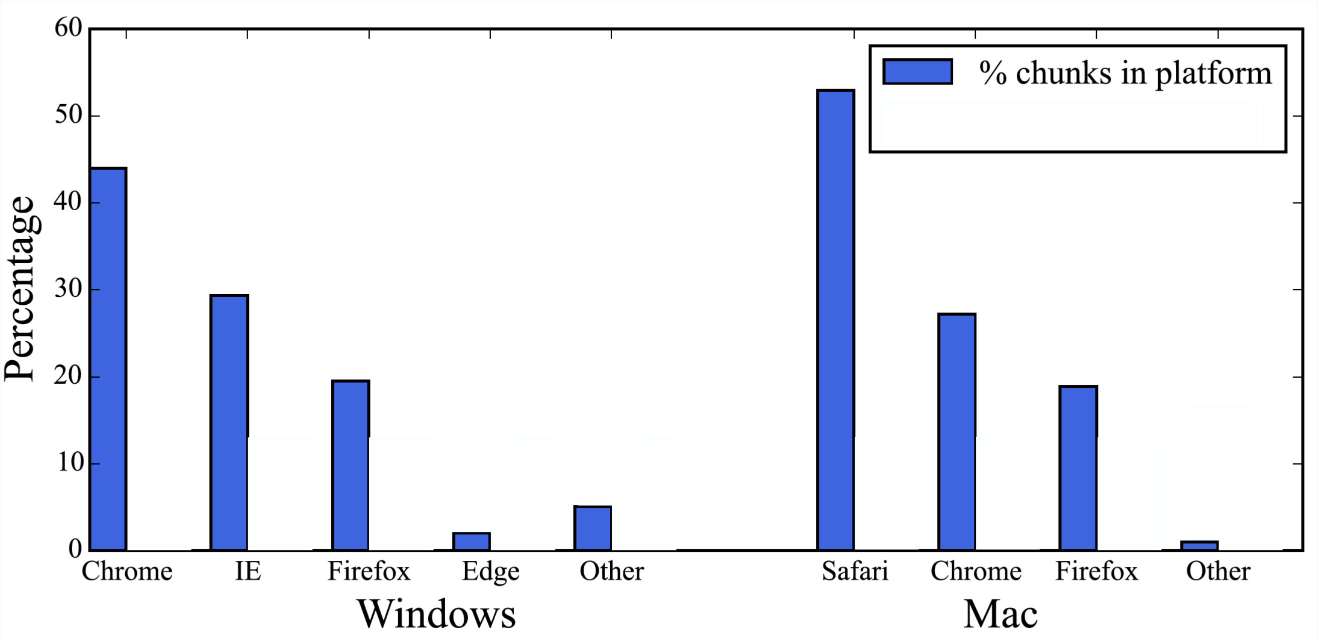
- Higher bitrates put more load on the CPU
- Showed better rendering framerate

Higher bitrates are often requested in connections:

- Lower RTT variation
- Lower retransmission rate

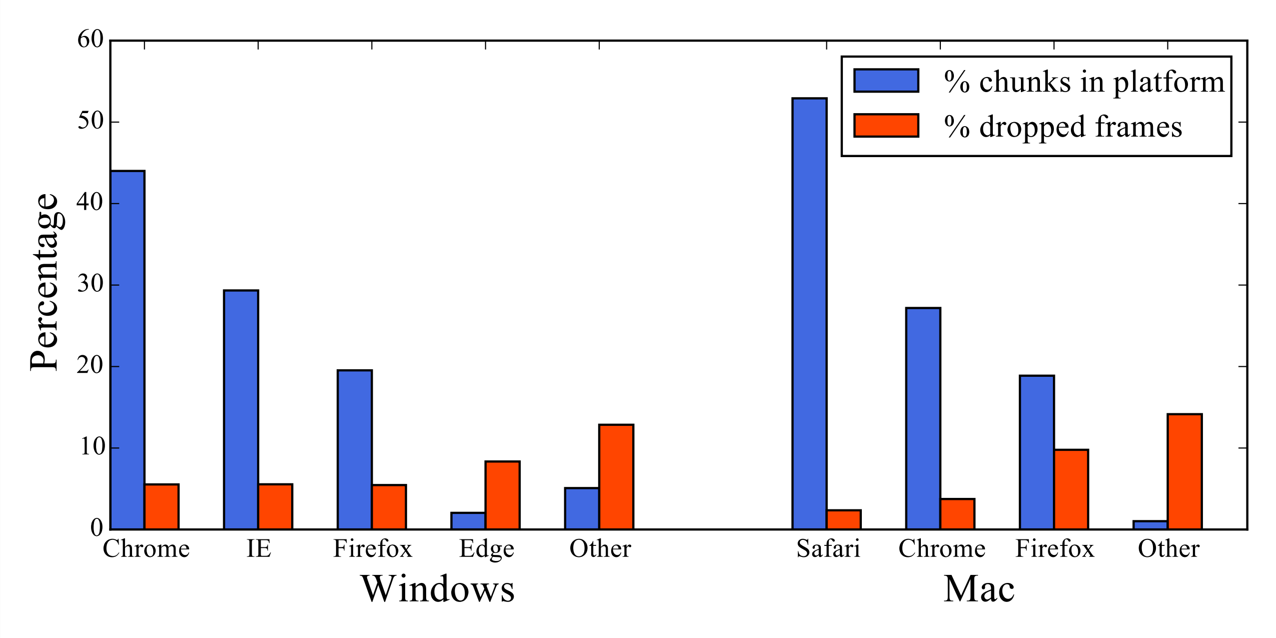
3. Unpopular Browsers Have Worse Rendering

- Chunks with good performance ($rate > 1.5\frac{sec}{sec}$)
- Player is visible (i.e., $vis = true$)



3. Unpopular Browsers Have Worse Rendering

- Chunks with good performance ($rate > 1.5\frac{sec}{sec}$)
- Player is visible (i.e., $vis = true$)



Take-aways

Take-aways: CDN

Problem	Take-away
Cache miss impact	Cache-eviction policy (e.g., GD-size, perfect LFU)
Cache miss persistence	Pre-fetch subsequent chunks
Load-Latency Paradox	Better load balancing (partition popular content)

Take-aways: Network

Problem	Take-away
Nearby clients with high latency	Avoid overprovisioning servers for nearby clients
Prefixes with persistent high latency or variation	Adjust ABR algorithm accordingly (more conservative bitrate, increase buffer size)
Earlier packet losses are more harmful (and more common!)	Use server-side pacing
Throughput is the major bottleneck	Good news for ISPs (e.g., establish better peering points)

Take-aways: Client

Problem	Take-away
Download stack latency	Can cause over-shooting or under-shooting by ABR, incorporate server-side TCP metrics
Rendering is resource-heavy	Use $1.5 \frac{sec}{sec}$ video arrival rate as rule-of-thumb
Rendering quality differs based on OS/browser	Avoid premature optimization on CDN/ISPs when the problem is client

Conclusion

- Instrumenting **both sides**
  - Uncover range of problems for the first time
- **Per-chunk** and per-session data
  - Uncover "persistent" vs. "transient" problems
- Our findings have been used to enhance performance in Yahoo

Thank You!

Questions?

gmojgan@akamai.com