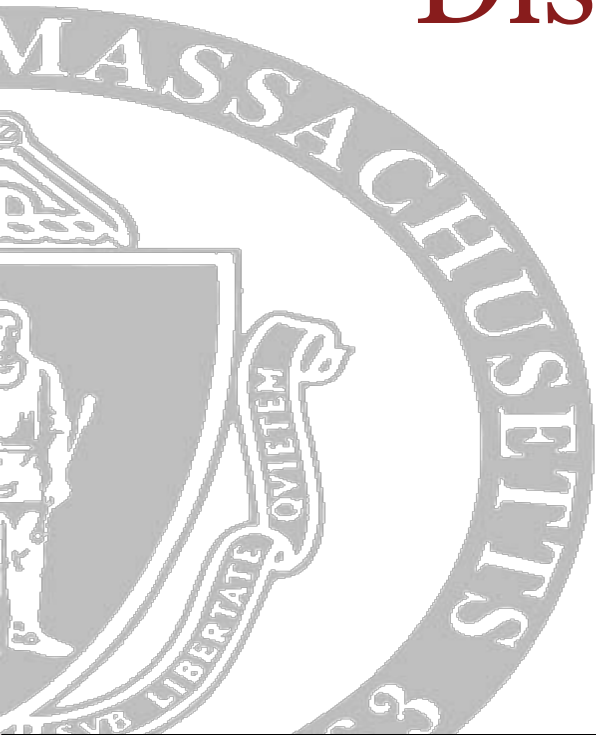


Network Assisted Content Distribution for Adaptive Bitrate Video Streaming

MMSys '17

Divya Bhat, Amr Rizk*,
Michael Zink, Ralf Steinmetz*

*Technische Universität Darmstadt



Adaptive Bitrate (ABR) Video Streaming

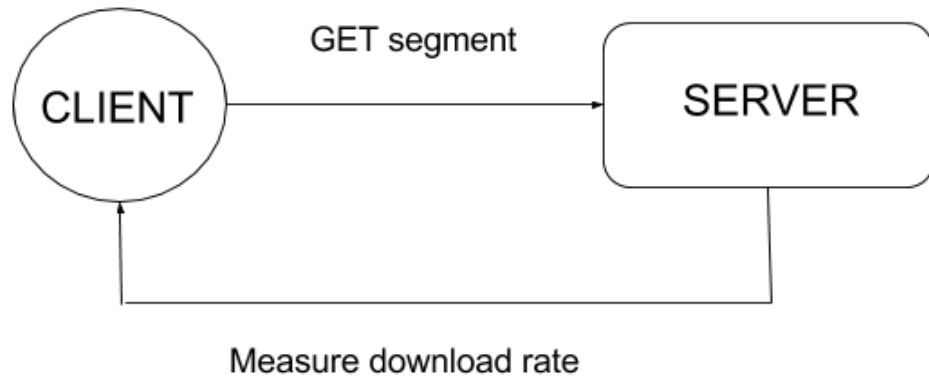


- Single Video split into multiple segments in various qualities
- **On-demand** and Live - Dynamic Adaptive Streaming over HTTP (DASH)
- Quality of Experience
 - Average Bitrate, Number of Quality Switches, Magnitude of quality switches, Rebuffering Events

[1] Image source: <https://insight.nokia.com/sites/default/files/wpuploads/2011/03/TCP-From-Data-to-Streaming-Video-Figure-1.jpg>

What are the current challenges of ABR streaming delivery?

Challenges: ABR Streaming Clients



- Media Presentation Description (MPD)
 - Qualities, segments of requested video
 - GET segment, measure download rate
- Buffer-based and Rate-based
- Key Issues:
 - Local measurement with no global knowledge of network
 - "Stale" measurements
 - Varying segment sizes (DASH)

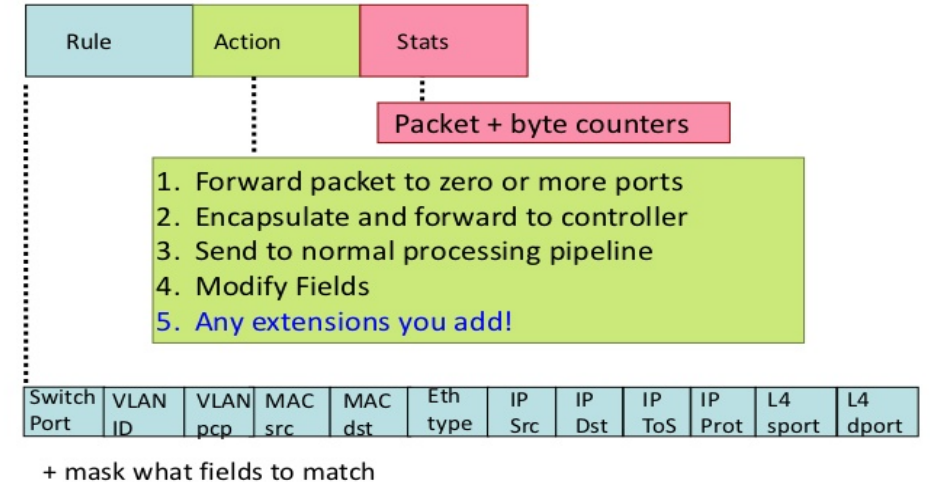
SDN and Content Distribution

- Content Distribution for ABR Delivery - Challenges
 - Multiple qualities of same video
 - Various encoding standards
 - User watching pattern
 - partially watched videos
 - ABR clients largely rely on local bandwidth estimates

- Software Defined Network (SDN)
 - Centralized vantage point
 - Traffic Engineering through dynamic routing
 - Standardized Measurement Framework APIs

OpenFlow Basics

Flow Table Entries

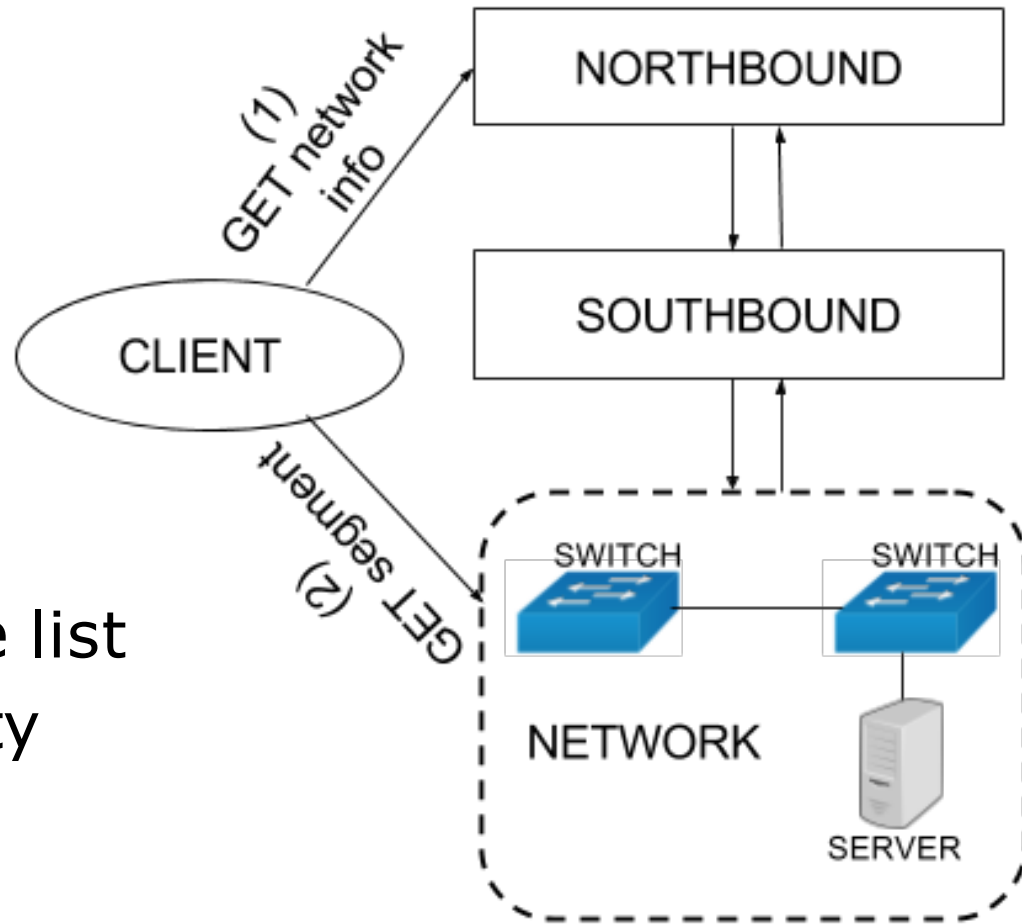


[2] Image source:
<http://image.slidesharecdn.com/openflowoverview-111116151908-phpapp01/95/openflow-overview-10-728.jpg?cb=1321460164>

How can we leverage the benefits of SDN to improve QoE for video streaming?

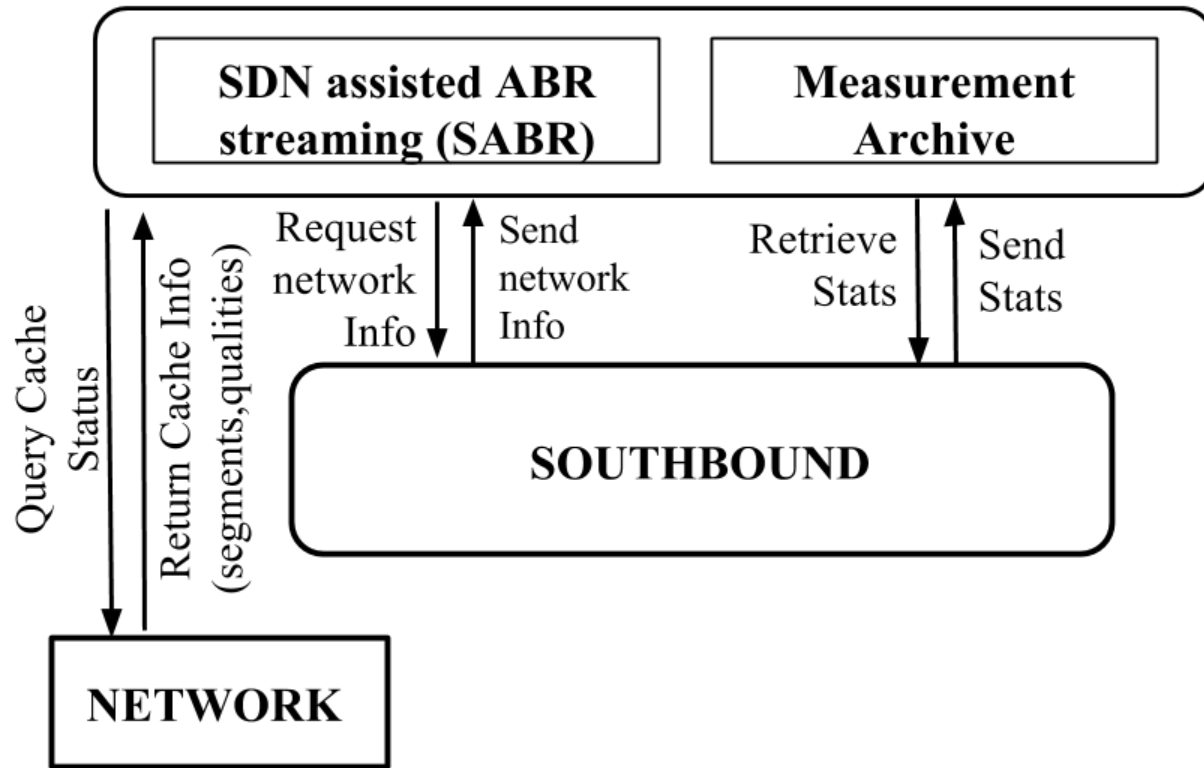
SABR - Proposed Architecture

- Client
 - Get cache list
 - Get quality



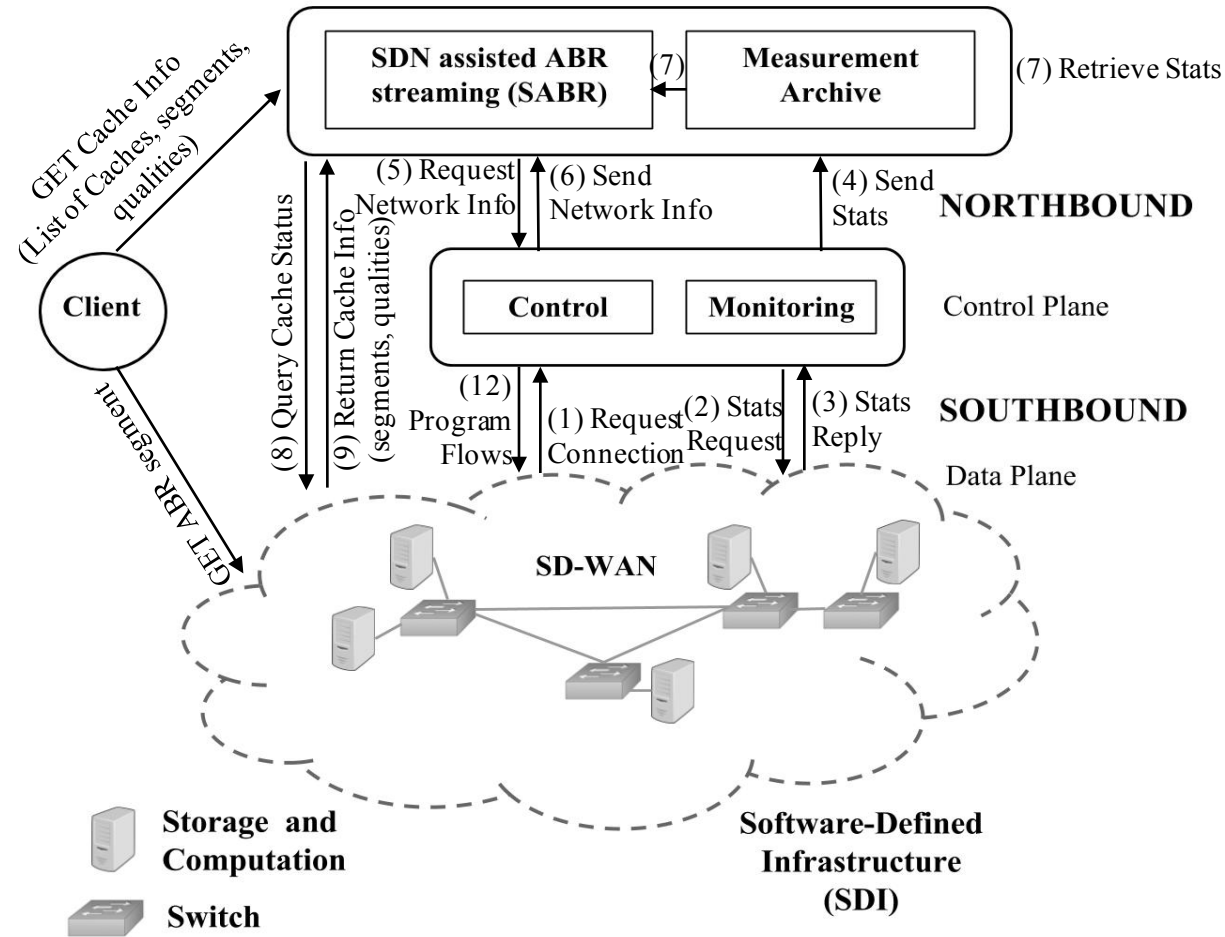
- Northbound
 - Archive stats
 - available bandwidth to each cache
- Southbound
 - Program flows
 - PoX controller
 - Measure statistics
 - ABR segment length=2s, sampling interval=1s

SABR - Northbound



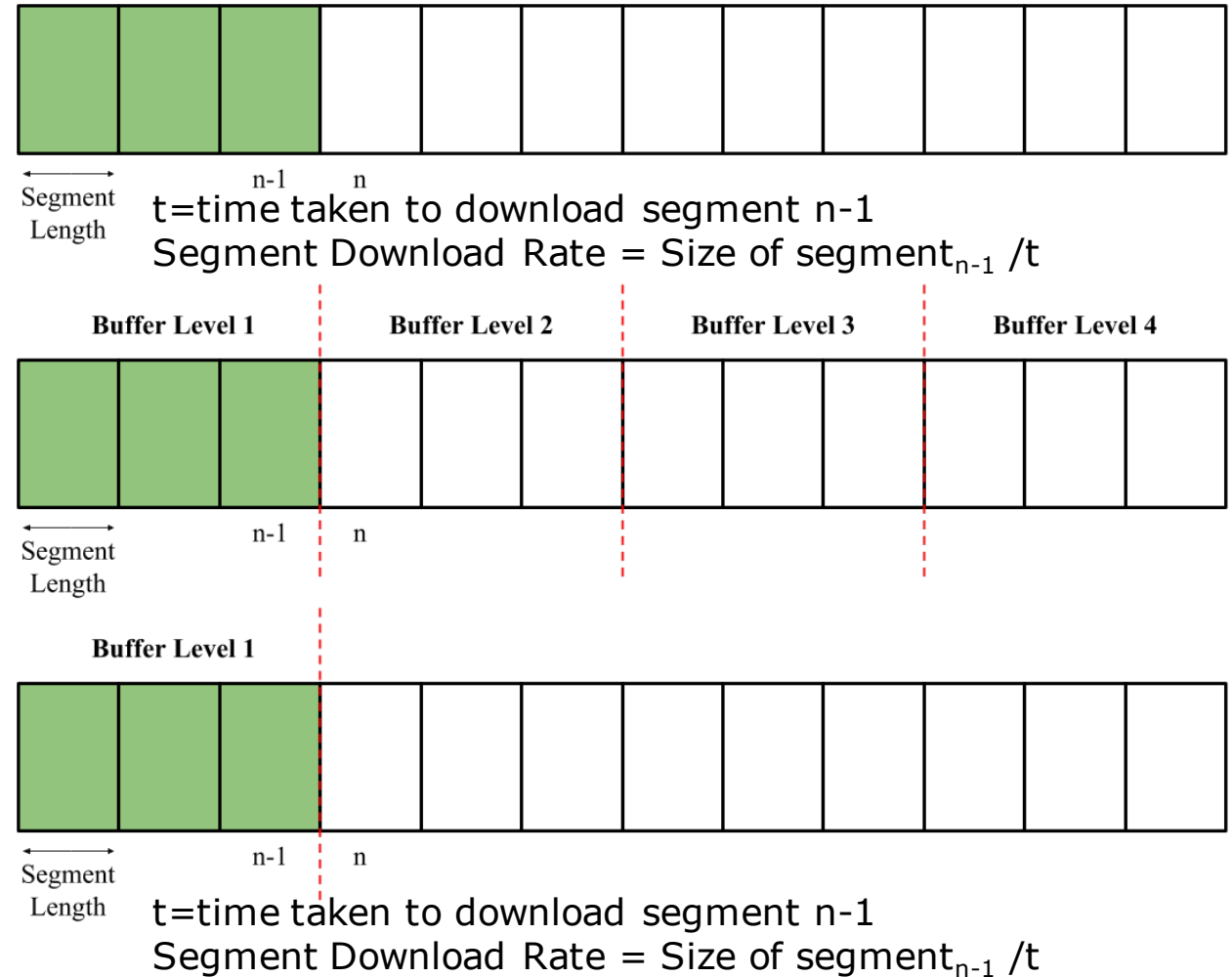
- Archive Stats
 - MongoDB
- SABR
 - Compute available bandwidth (ARIMA) forecasts
 - Get cache status
 - REST APIs

SABR – Complete Architecture



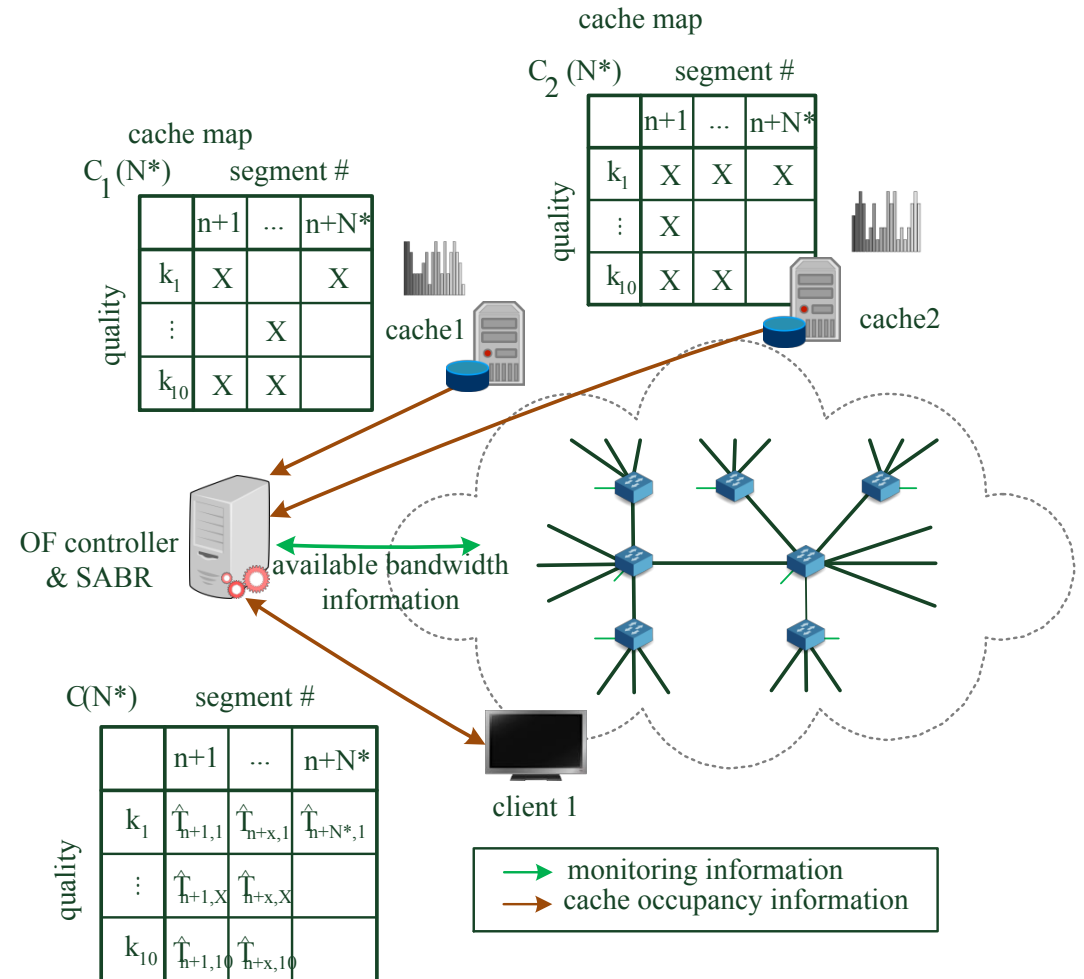
Quality Adaptation Problem

- Rate-based
- Buffer-based
- Rate and buffer-based



SABR - Client

- Rate-based
 - Quality based on available bandwidth
- Buffer-based
 - Time required $T = \text{segment size}/\text{available bandwidth}$
 - Expected buffer level = (current buffer level - T + segment length)



SABR – Caching Algorithms

- LRU Caching
 - Most common is Least Recently Used (LRU)
- Local
 - Each cache is considered independent
 - Cache content in “nearest” cache
 - Geo-location
- Global
 - Each cache is part of a global cache space
 - Full replication – miss creates n copies of the content, where n =no. of caches
 - No replication – miss creates single copy of content (“nearest” cache)
- Quality-based
 - Low quality content is cached in higher level and high quality in lower level

CloudLab Testbed setup

• Caches

- Apache2, MongoDB and scapy for WSGI emulation
- 50 videos, 5-minutes each, 5 qualities, segment length = 2s

• Client

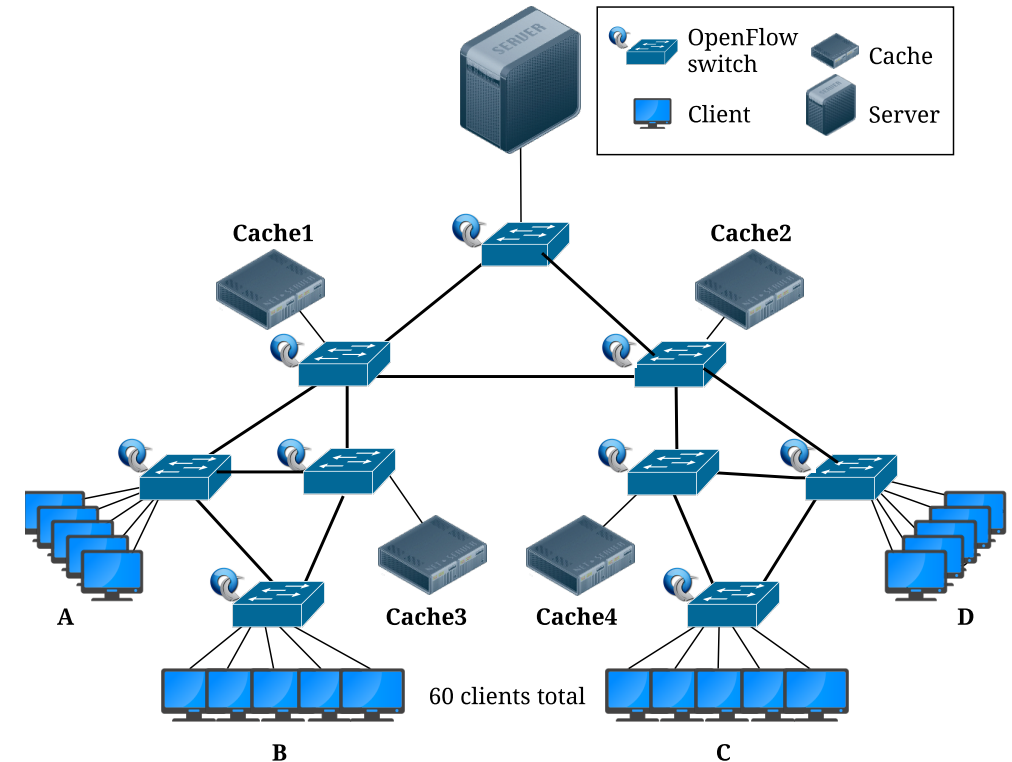
- Python DASH client [6]
- Zipf-based requests

• OpenVirtual Switches (OVS)

• OF Controller and SABR

- Control plane and Northbound application

[6] <https://github.com/pari685/AStream>



CloudLab topology

Performance Metrics

- Average Quality Bitrate
- No. of Quality Switches
- Spectrum[6]
 - Function of amplitude and no. of switches
- Rebuffering Ratio
- Cache Hit Rate
- Network Utilization
- Server Load Ratio
 - Lower indicates better hit rate

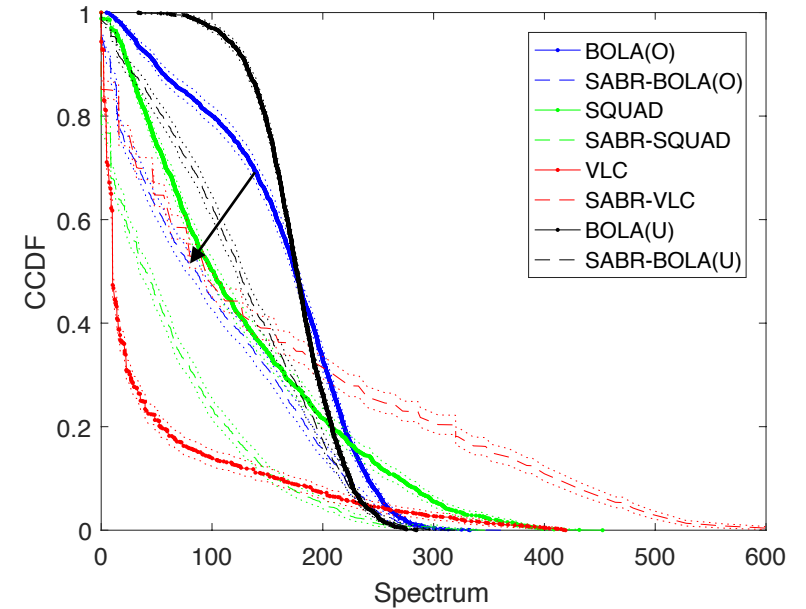
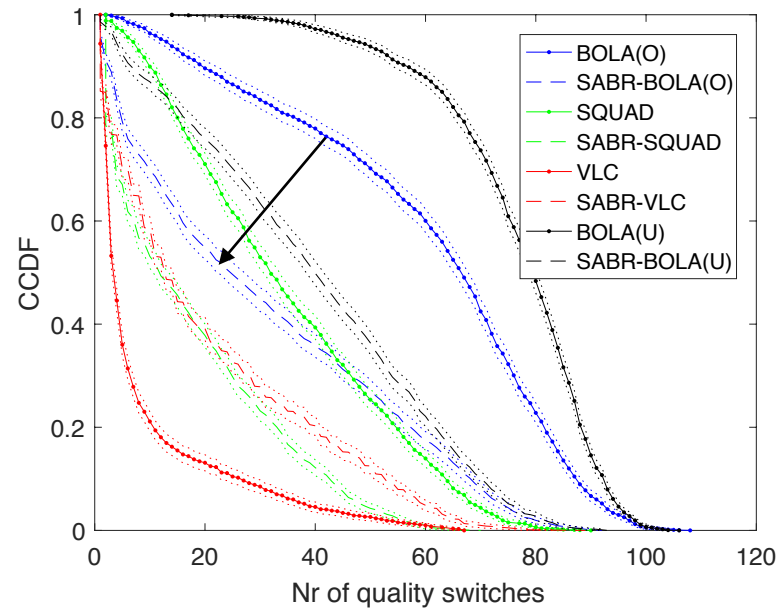
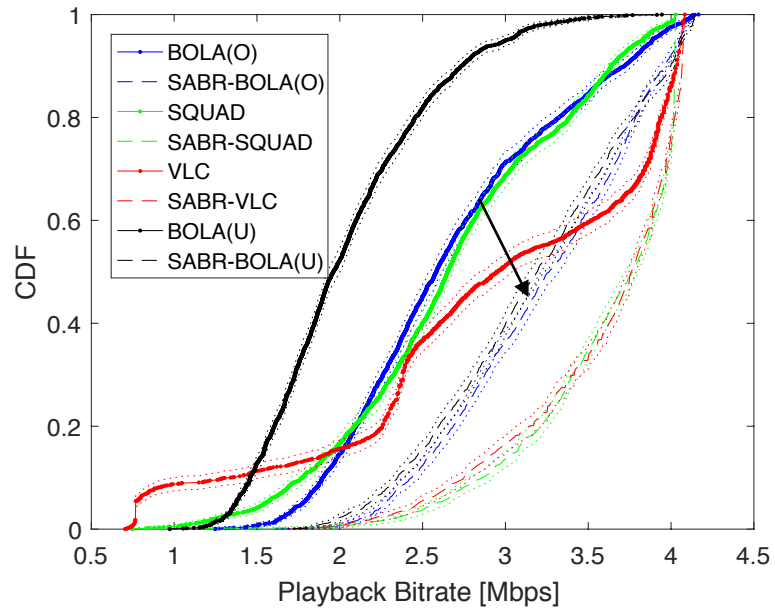


Client Metrics

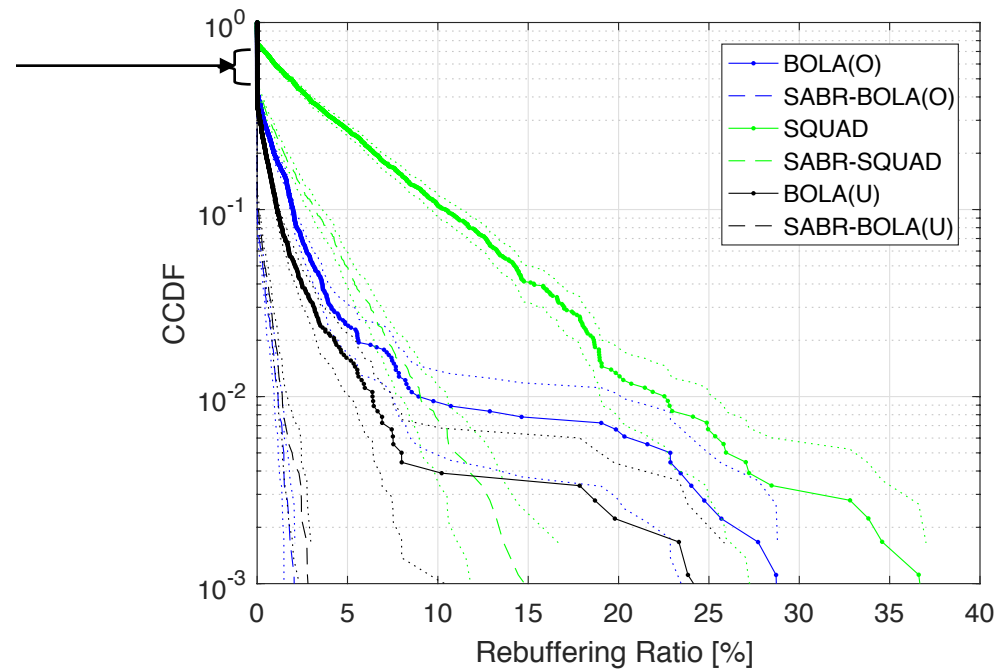
CDN Metrics

[6] Zink et al. "Layer-encoded video in scalable adaptive streaming." *IEEE Transactions on Multimedia* 7.1 (2005)

SABR: Client Performance



SABR: Client Performance



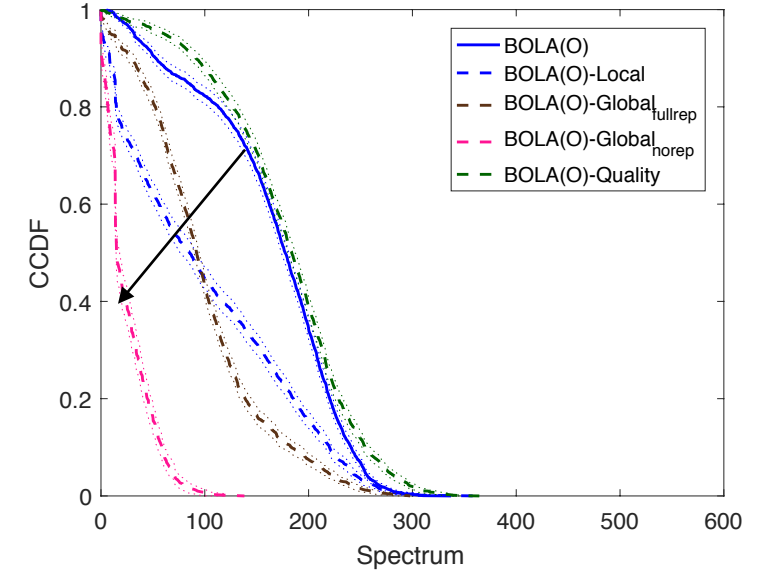
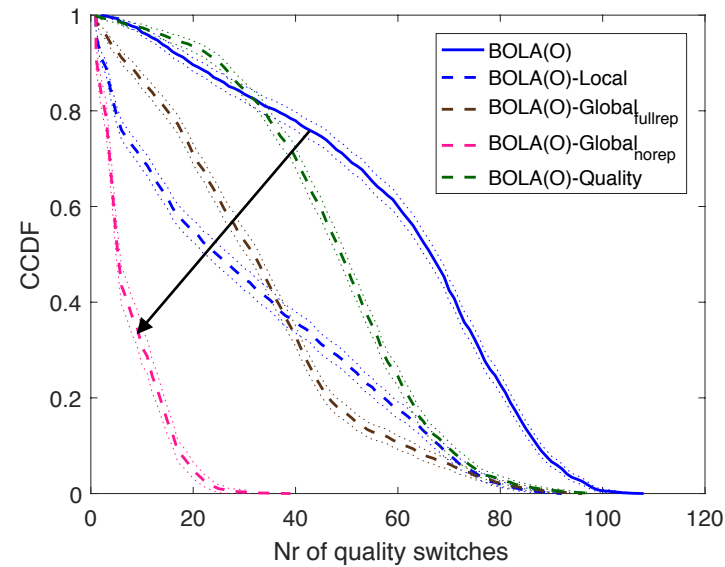
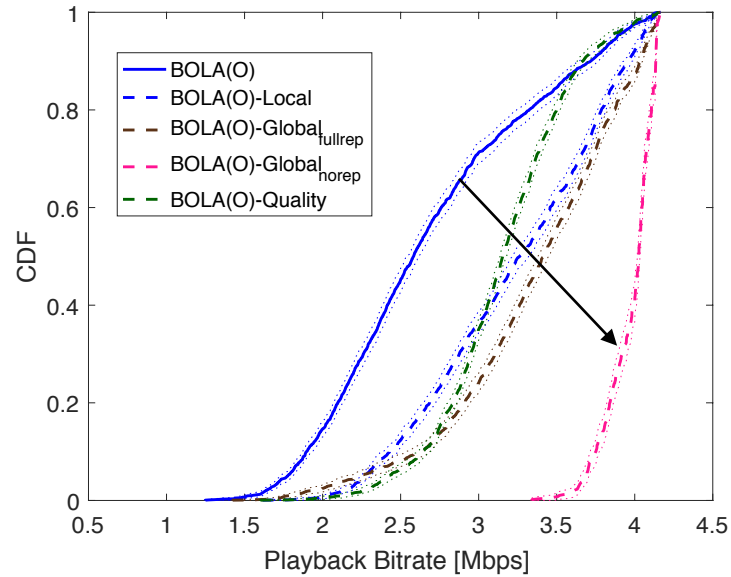
SABR: Network Performance

	VLC	SABR - VLC	SQUAD	SABR-SQUAD	BOLA(U)	SABR - BOLA(U)	BOLA(O)	SABR - BOLA(O)
Network Util. (%)	42.5	71.2	44.9	74.6	42.5	68.2	46.4	65.7
Server Load (%)	28.2	21.3	34.5	19.5	28.7	22.5	33.8	23.6

- Better network utilization – increased cache hits
- Reduced server load – lower origin offloading

How does content placement strategy impact QoE?

Caching Performance – BOLA(O)



- No replication gives best QoE
- Cache Hit rate reveals highest hit rate for Q_5

Analysis

- Quality bitrate, no. of switches and spectrum values are improved
- Cache Hit Rate, Network Utilization and Server load are improved
- Global – full replication best for SQUAD, no replication for BOLA (BOLA requests all presented qualities but SQUAD prefers some over the other)
- Quality caching means cache hit rate is improved but QoE suffers significantly

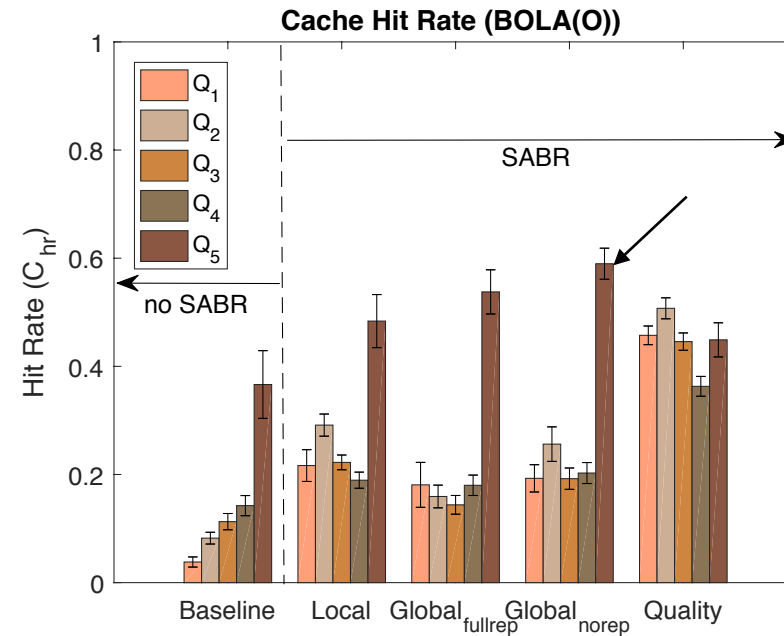
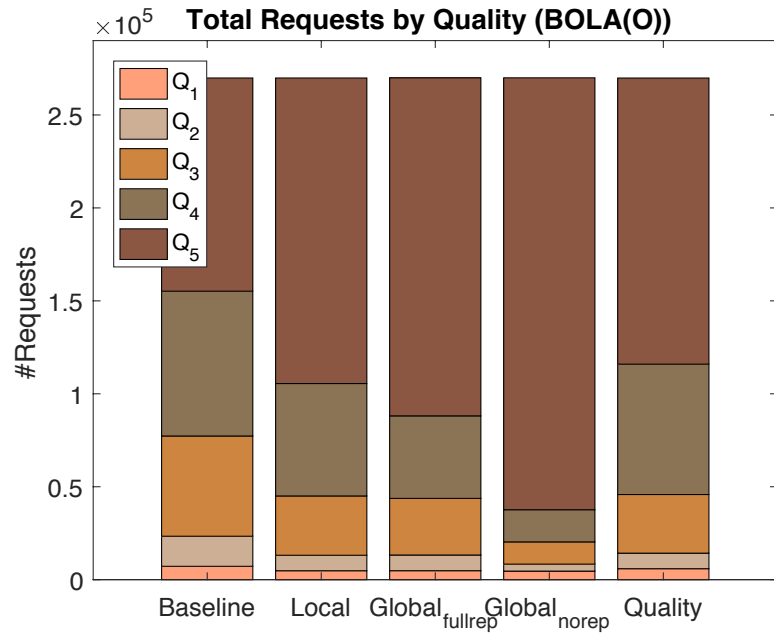
Conclusion

- Performance of ABR improves with SABR
 - Code available at: <https://github.com/dbhat/SABR.git> - Artifact Evaluated
- Benefits observed depends on:
 - Client algorithm
 - Caching strategy
- Scalability
 - Distributed frameworks to reduce load
 - Monitoring of ports that are used, sampling interval
- Other caching approaches
 - Time-to-live (TTL)
- User behavior – partially watched videos

References

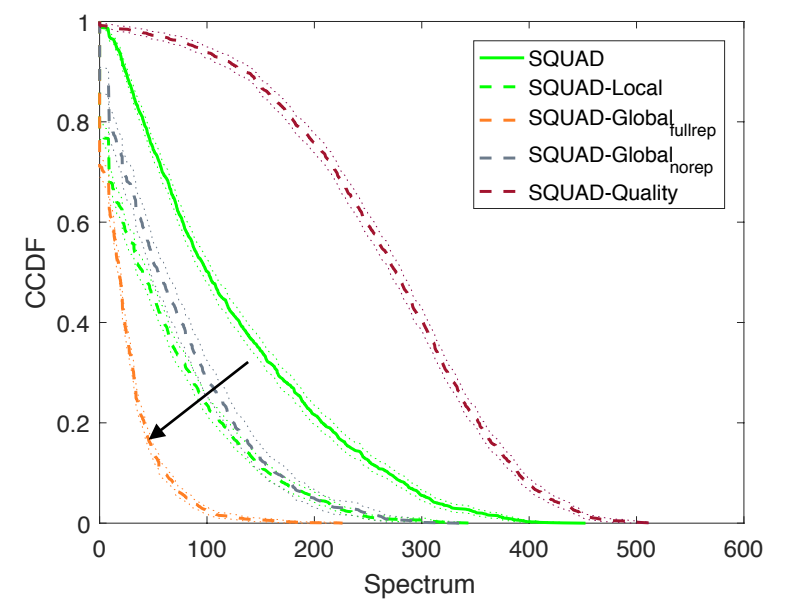
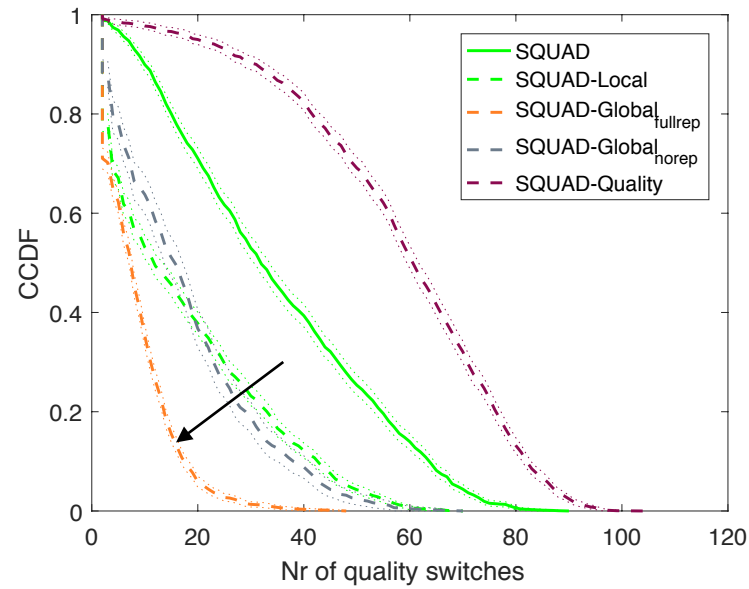
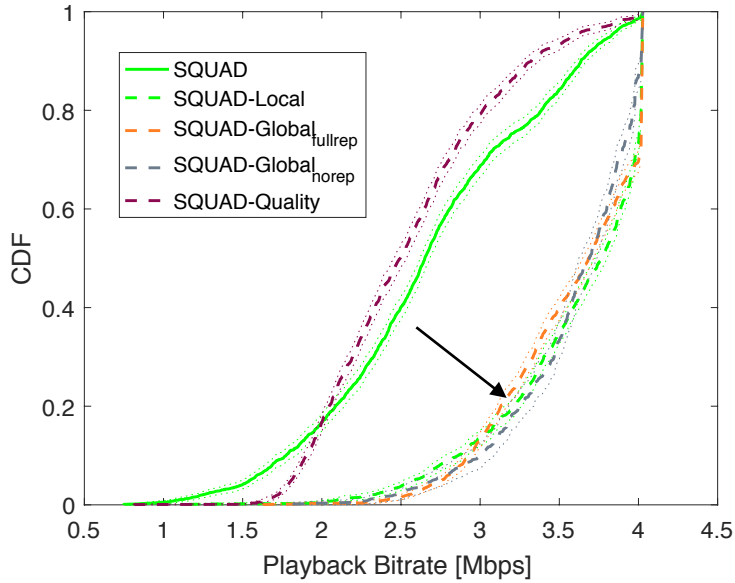
- [3] Wang, Cong, Amr Rizk, and Michael Zink. "SQUAD: a spectrum-based quality adaptation for dynamic adaptive streaming over HTTP." *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 2016.
- [4] Spiteri, Kevin, Rahul Uргаonkar, and Ramesh K. Sitaraman. "BOLA: Near-Optimal Bitrate Adaptation for Online Videos." *arXiv preprint arXiv:1601.06748* (2016).
- [5] Huang, Te-Yuan, et al. "A buffer-based approach to rate adaptation: Evidence from a large video streaming service." *ACM SIGCOMM Computer Communication Review* 44.4 (2015): 187-198.
- Benjamin Frank, Ingmar Poesе, Yin Lin, Georgios Smaragdakis, Anja Feldmann, Bruce Maggs, Jannis Rake, Steve Uhlig, and Rick Weber. 2013. Pushing CDN-ISP collaboration to the limit.

Caching Performance – BOLA(O)



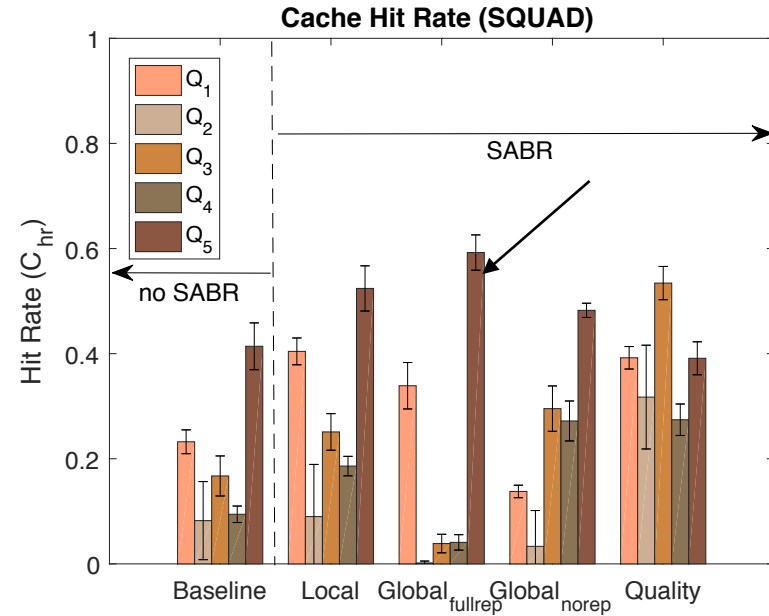
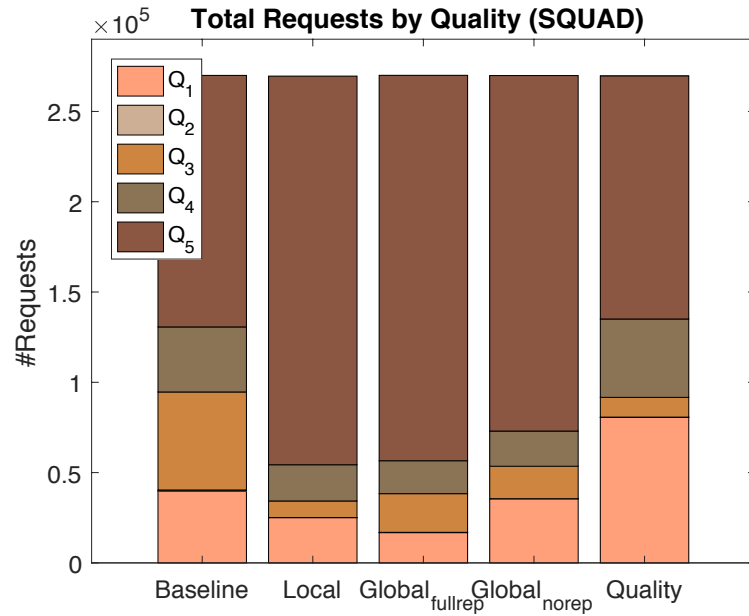
- For the no replication case, BOLA(O) sees highest hit rate for Q₅.

Caching Performance - SQUAD



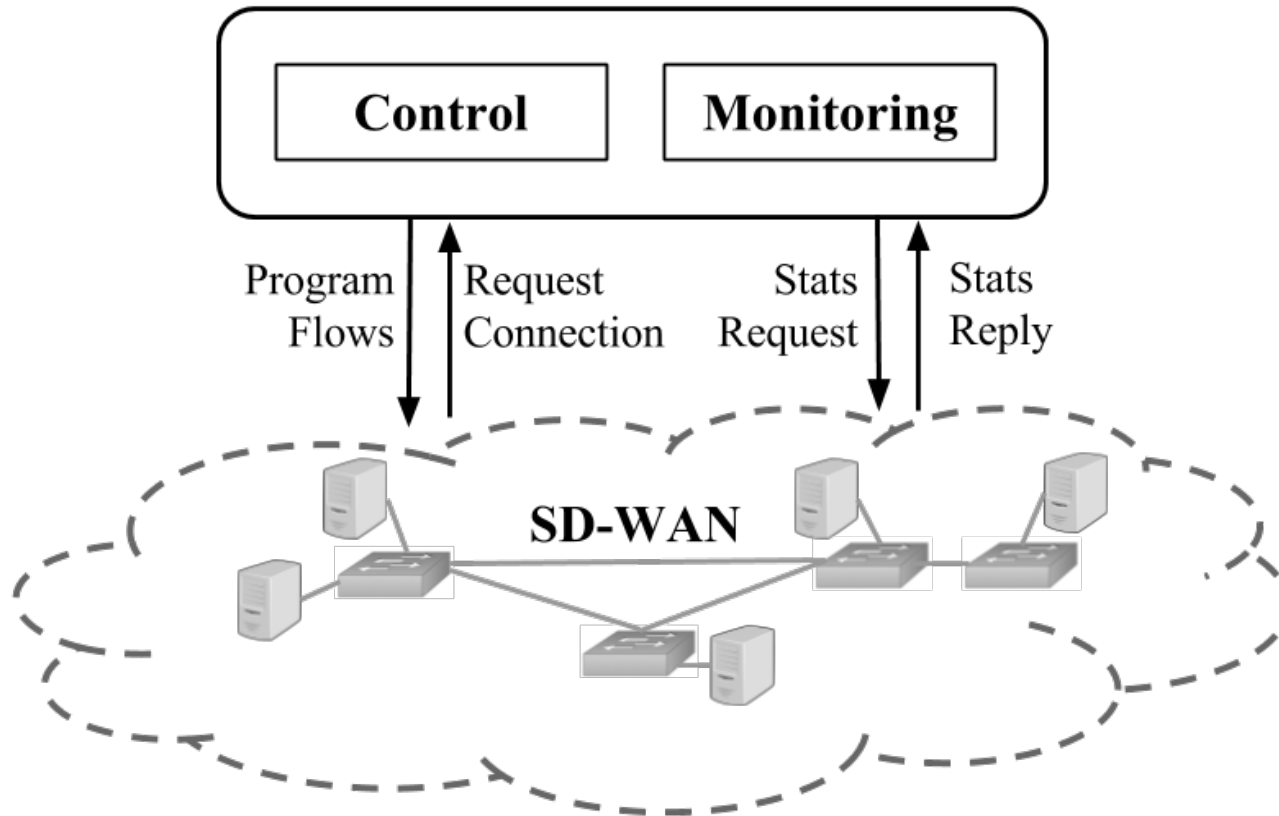
- Full replication gives best QoE

Caching Performance - SQUAD



- For the full replication case, SQUAD sees highest hit rate for Q₅.

SABR - Southbound



- Program Flows
 - OpenFlow PoX
- Measure Statistics
 - Port statistics
 - Sampling Rate limited by switch
 - Higher – more accurate
 - Lower – reduced load
 - Every second (ABR segment length=2s)

Caching – Network Performance

	BOLA (0)				SQUAD			
	Local	Global _{full} rep	Global _{no} rep	Quality	Local	Global _{full} rep	Global _{no} rep	Quality
Network Util. (%)	65.7	83.3	82.9	77.1	74.6	82.3	66.9	48.2
Server Load (%)	23.6	20.5	20.4	22.4	19.5	20.3	22.5	28.4

Spectrum Equation

$$s(v) = \sum_{t=1}^T z_t \left(h_t - \frac{1}{\sum_{i=1}^T z_i} \left(\sum_{j=1}^T z_j h_j \right) \right)^2 .$$

- h_t = no. of layers in time in slot t , $t=1, \dots, T$
- Indication of step in time slot t , $z_t \in \{0, 1\}$, $t= 1, \dots, T$
- Z_t is frequency of variations
- Higher amplitude of changes means higher spectrum