



# Latency-Aware Media Delivery through Software-Defined Networks

**Thorsten Herfet**

Saarland Informatics Campus  
66123 Saarbrücken, Germany



UNIVERSITÄT  
DES  
SAARLANDES

**José Manuel Menéndez**

Universidad Politécnica de Madrid  
28040 Madrid, Spain



POLITÉCNICA



- ▶ Low Latency Video Streaming
- ▶ Software-Defined Networking
  - 5G Broadcast through SDNs
  - Predictably-reliable Real-Time Transport<sup>1</sup>
  - Transparent Transmission Segmentation<sup>2</sup>
- ▶ Buffer Dynamics Stabilizer<sup>3</sup>

1. Gorius, M.: "[Adaptive Delay-constrained Internet Media Transport](#)", Dissertation, UdS, December 2012
2. Schmidt, Andreas; Herfet, Thorsten: "[Approaches for Resilience- and Latency-Aware Networking](#)", International Symposium on Networked Cyber-Physical Systems (NetCPS, Poster Session), Munich, September 2016
3. Shuai, Yongtao; Herfet, Thorsten: "Improving User Experience in Low-Latency Adaptive Streaming by Stabilizing Buffer Dynamics", IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, January 2016



- ▶ All domains are going IP
  - TelCos, BCAST con- & distribution (DOCSIS), production (Industrial Ethernet, TSN), even the holy grail automotive (OPEN)
  - Antipodes are synchronous, prioritized streaming (TSN) and dynamic adaptive streaming (DASH)
- ▶ Significant progress in theory and domains
  - [Coding with Finite Block Length](#) (Polyanskiy et.al. 2010)
    - Channel capacity with limited block length (no given time)
  - [ITU-T Watch Report](#) (Tactile Internet, 2014)
    - Extremely low latencies of 1 ms (no given reliability)



▶ Media Transmission requires (ITU-T Y.1541)

– Predictable Delay

- Application dependent

– Predictable Reliability

- Application dependent

▶ Coding delay essential

- 64 kbit block length means:

- 0.5 – n sec. for audio
- 0.005 – 0.2 sec. video

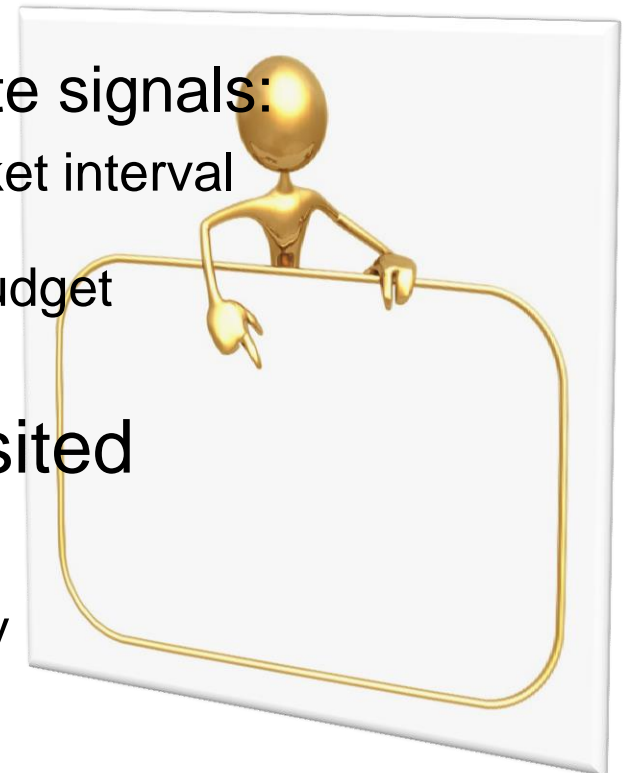
▶ And additionally...

Class	IPTD	IPDV	IPLR	IPER	IPRR	Applications (examples)
0	100 ms	50 ms	$1 \times 10^{-3}$	$1 \times 10^{-4}$	–	Real-time, jitter sensitive, low delay, highly interactive
1	400 ms	50 ms	$1 \times 10^{-3}$	$1 \times 10^{-4}$	–	Real-time, jitter sensitive, medium delay, interactive
2	100 ms	U	$1 \times 10^{-3}$	$1 \times 10^{-4}$	–	Transaction data, low delay, highly interactive
3	400 ms	U	$1 \times 10^{-3}$	$1 \times 10^{-4}$	–	Transaction data, medium delay, interactive
4	1 s	U	$1 \times 10^{-3}$	$1 \times 10^{-4}$	–	Low loss
5	U	U	U	U	–	Best effort
6	100 ms	50 ms	$1 \times 10^{-5}$	$1 \times 10^{-6}$	$1 \times 10^{-6}$	High bit rate, strictly low loss, low delay, highly interactive
7	400 ms	50 ms	$1 \times 10^{-5}$	$1 \times 10^{-6}$	$1 \times 10^{-6}$	High bit rate, strictly low loss, medium delay, interactive

Notes – U: undefined  
IPTD: IP Packet Transfer Delay  
IPDV: IP Packet Delay Variation  
IPLR: IP Packet Loss Rate  
IPER: IP Packet Error Ratio  
IPRR: IP Packet Reordering Ratio



- ▶ IP makes things worse:
  - Not „noisy“ (AWGN) but „lossy“ ( $E[\text{rasure}]$ ) channel
  - Complete IP packets get lost due to noise, contention and/or queuing
  - IP packet rate low even for high rate signals:
    - 4 Mbps SD video has ~2.5 ms IP packet interval  
(assumed 7 MPEG-2 TS packet per IP packet)
    - Small blocks already consume time budget  
(40 packets / 100 ms)
- ▶ Channel capacity has to be revisited
  - Capacity under delay constraints
    - Time dependent minimum redundancy
  - Residual error rate tolerable





- ▶ Low Latency Video Streaming
- ▶ **Software-Defined Networking**
  - 5G Broadcast through SDNs
  - Predictably-reliable Real-Time Transport<sup>1</sup>
  - Transparent Transmission Segmentation<sup>2</sup>
- ▶ Buffer Dynamics Stabilizer<sup>3</sup>

1. Gorius, M.: "[Adaptive Delay-constrained Internet Media Transport](#)", Dissertation, UdS, December 2012
2. Schmidt, Andreas; Herfet, Thorsten: "[Approaches for Resilience- and Latency-Aware Networking](#)", International Symposium on Networked Cyber-Physical Systems (NetCPS, Poster Session), Munich, September 2016
3. Shuai, Yongtao; Herfet, Thorsten: "Improving User Experience in Low-Latency Adaptive Streaming by Stabilizing Buffer Dynamics", IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, January 2016



- ▶ TV broadcasting and mobile broadband are undoubtedly essential parts of today's society. Both of them are now facing tremendous challenges to cope with the future demands.
- ▶ Video is expected to contribute ~70% of all the mobile traffic by 2018.
- ▶ Content distribution is expected to be the dominant contributor to the mobile data traffic demand, therefore content media distribution is being more and more present in everyday life communications.







- ▶ The consumption is moving towards on-demand services so that anyone can access contents anytime, anywhere and regardless of the device type.
- ▶ And quality is going further very quickly, in most European countries people have become accustomed to HDTV and are now expecting even UHD..

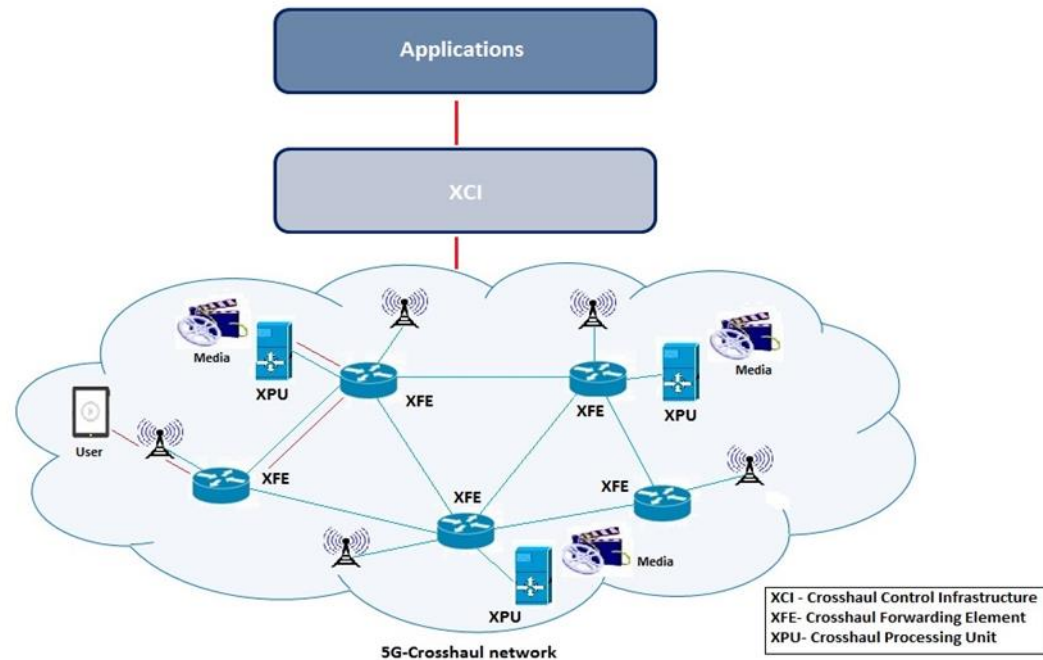
- The trends of on-demand, mobile, and Ultra HD quality impose formidable challenges for TV and the delivery network of the future to be coped for 5G capabilities.







- ▶ Media content distribution takes advantage of an adaptive and cost-efficient solution for the 5G transport network, integrating both fronthaul and backhaul segments.
- ▶ The envisioned solution requires a fully integrated and unified management of fronthaul/backhaul resources in a sharable, scalable and flexible way.



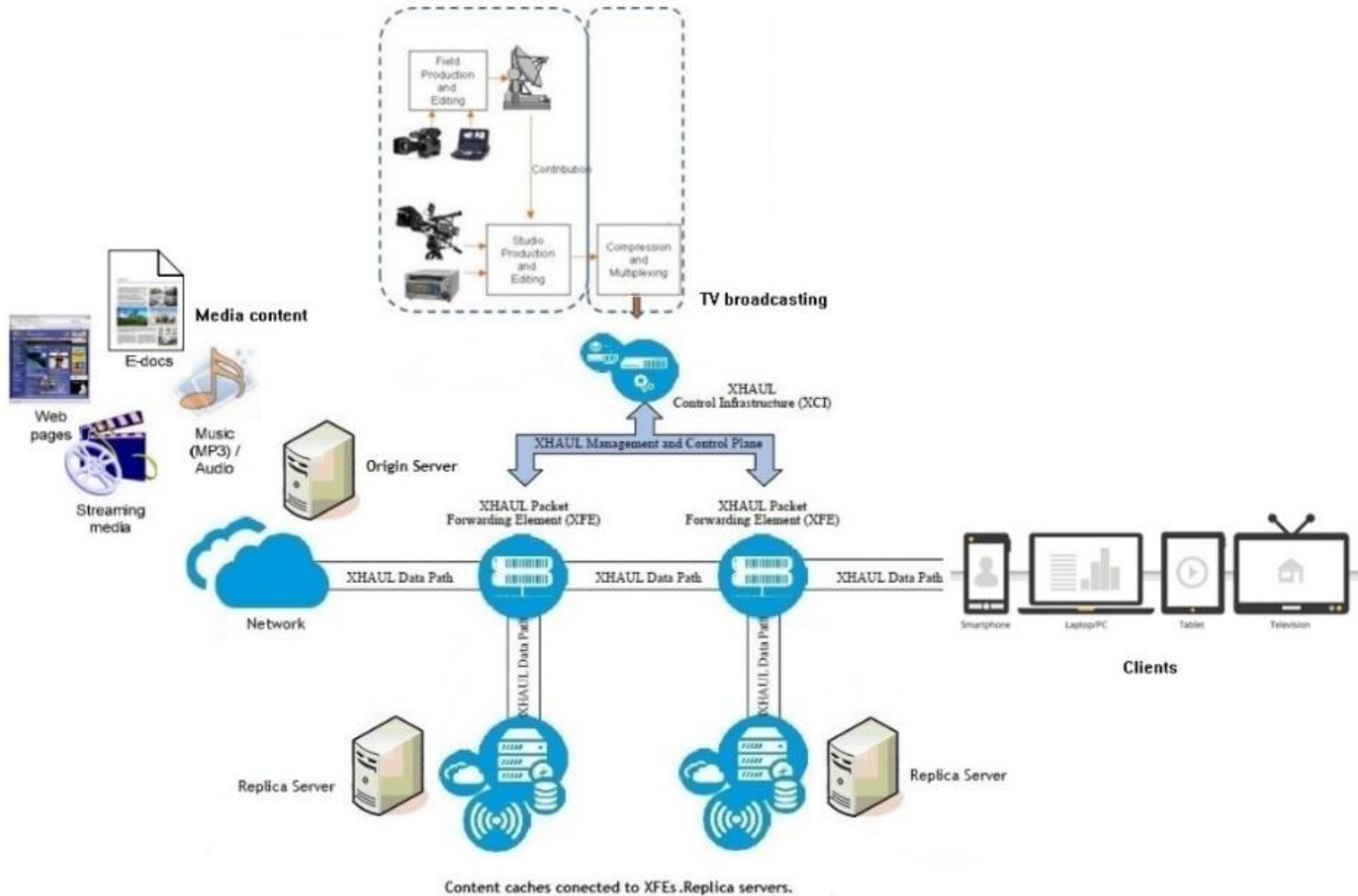


- ▶ The control and management of such an integrated transport network, will be based on the **SDN principles** and architecture defined by the Open Networking Foundation (ONF) and will adopt **Network Function Virtualisation (NFV)** concepts and mechanisms as well as aligned with the ETSI **Management and Orchestration (MANO) architecture** as a specific means to offer a subset of services of network, cloud and storage.



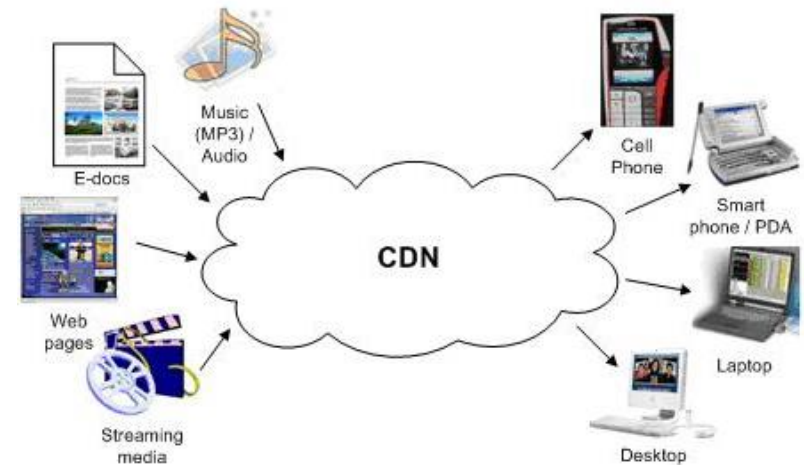


# Media content distribution scenario

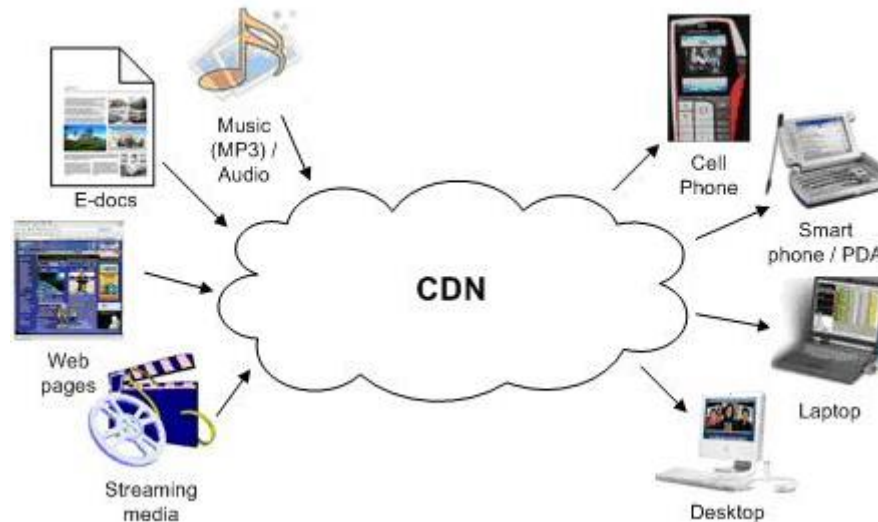




- A CDN is a combination of a content-delivery infrastructure, a request-routing infrastructure and a distribution infrastructure.



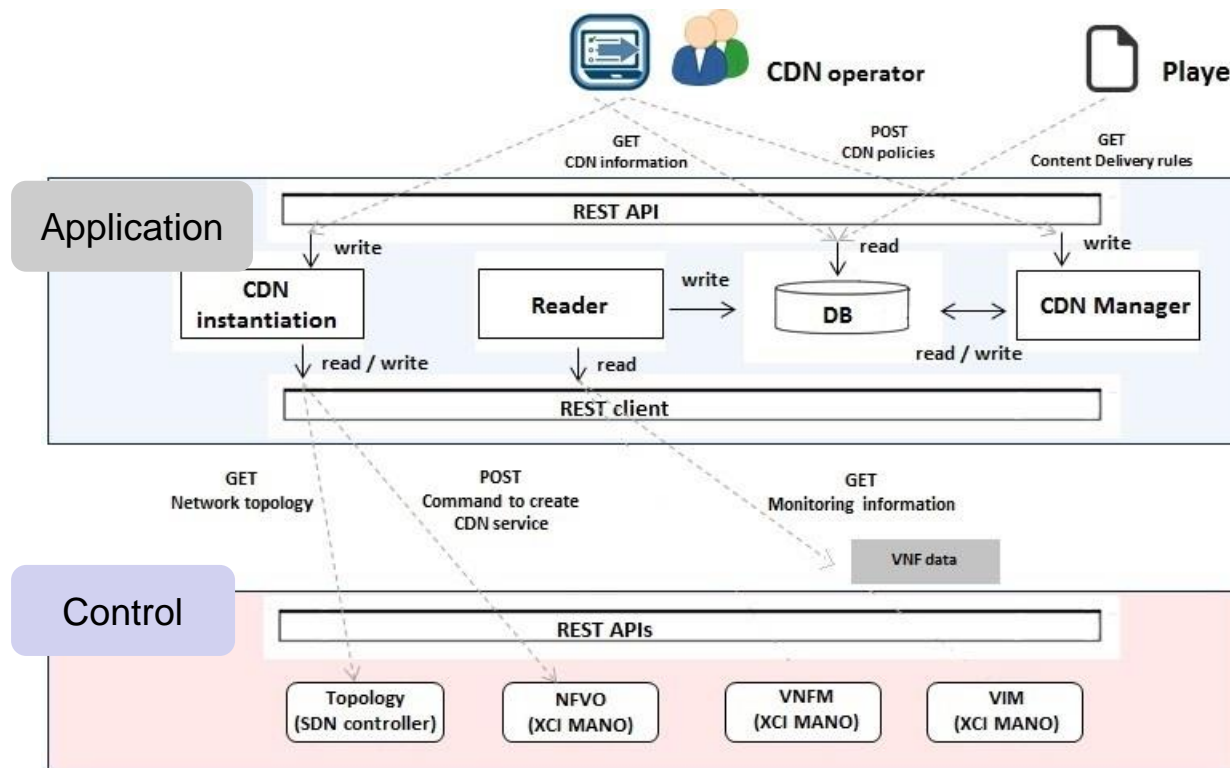
- ▶ Virtual CDNs will control the load balancing over several replica servers strategically placed at various locations in order to deal with massive content requests.
- ▶ This model will improve the content delivery maximizing bandwidth and improving accessibility through the CDN infrastructure according to the user demands.



- ▶ The managing application running on the control plane will manage the entire CDN infrastructure and it will evaluate the system performance.
- ▶ It will receive monitoring information from the infrastructure (location, latency, packet-loss, bandwidth, server load) and information from the user about his location.



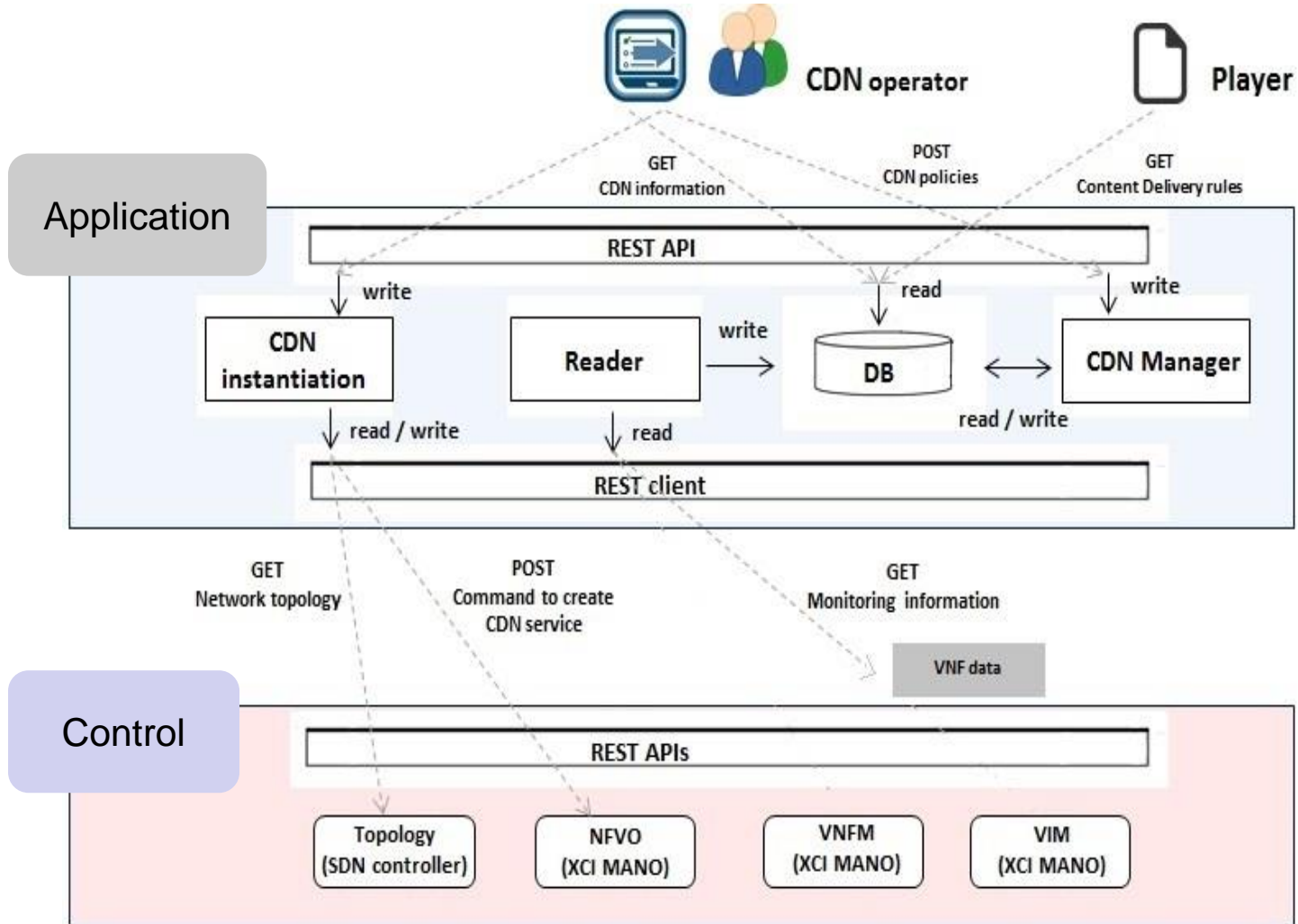
- ▶ Taking this information as input and applying optimization logic defined by the CDN operator based on the CDN and network metrics, the application strategically deploys and optimizes the replica servers in order to optimize resource utilization and deliver the content to the end users/content consumers with maximum possible QoE.







# Content Delivery Networks







- ▶ The content delivery infrastructure will be implemented via networks of content caches -replica servers- which are deployed close to the FEs across the network topology.
- ▶ The request-routing and distribution infrastructure functions will be enforced through optimal content routing and delivery based on information from 5G network.



# Content Delivery Networks

Application Plane

CDN  
Instantiation

CDN  
Manager

Control Plane

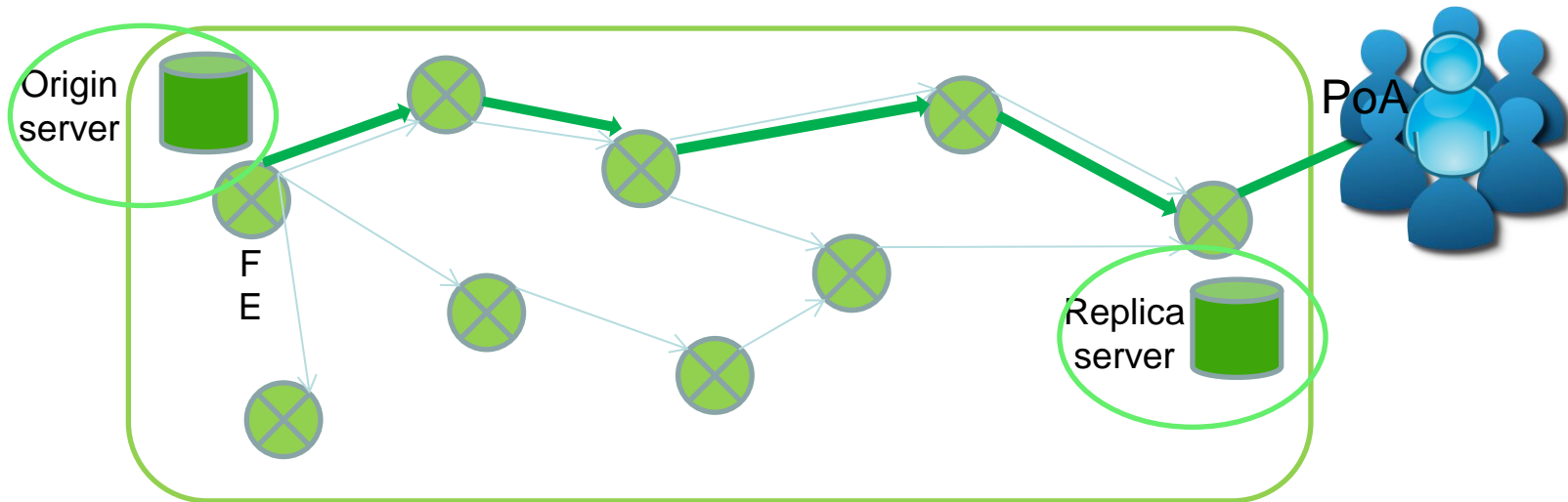
Analytics for  
Monitoring  
SDN Controller

Network re-  
configuration  
SDN Controller

Path provisioning  
SDN Controller

VIM & NFVO (MANO)

Data Plane





- ▶ Media/TV broadcasting & multicasting services utilizing the 5G using the same network with a controlled quality and offered as a Broadcast-as-a-service
- ▶ The focus is on minimizing both the cost and the spectrum consumption. Increase cost-effectiveness of transport technologies for media distribution
- ▶ Useful to demonstrate the feasibility of 5G objectives such as front and backhaul-integrated (control/planning) applications
- ▶ QoS based decision on the service provision (routing and quality adaptation) enables real-time decision



Application Plane

Video Service

Network Topology Monitoring

QoS

REST API

Control Plane

SDN Controller

NFV MANO

Path provisioning - Broadcast tree

Network reconfiguration

Data Plane

Forwarding Elements

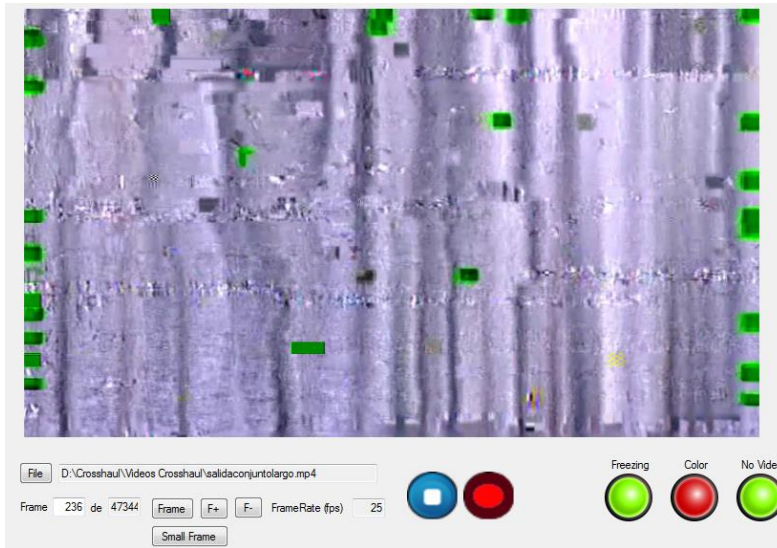
Quality Probe



- ▶ **Monitoring routing** is important to understand reachability and network paths that can affect network performance, leading to high packet loss or latency. Loss – Latency – Jitter
- ▶ **Network performance** is typically measured by the success of IP forwarding, the time and variation for packets to make it to the destination, and the theoretical capacity and actual bandwidth available.
  - Packet loss
  - Latency and jitter
  - Bandwidth
  - Undersized Path MTU and oversized TCP MSS



- ▶ **Video quality assessment** is used for detecting transmission problems on the image. Frozen frames, black frames and NR metrics are used to feed back the service





- ▶ Low Latency Video Streaming
- ▶ **Software-Defined Networking**
  - 5G Broadcast through SDNs
  - Predictably-reliable Real-Time Transport<sup>1</sup>
  - Transparent Transmission Segmentation<sup>2</sup>
- ▶ Buffer Dynamics Stabilizer<sup>3</sup>

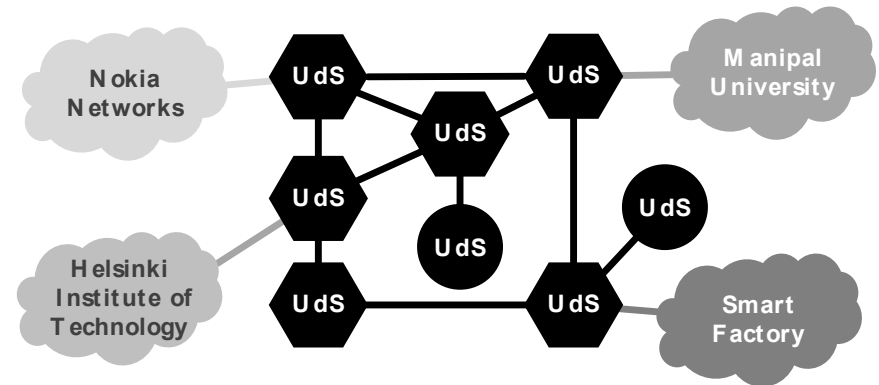
1. Gorius, M.: "[Adaptive Delay-constrained Internet Media Transport](#)", Dissertation, UdS, December 2012
2. Schmidt, Andreas; Herfet, Thorsten: "[Approaches for Resilience- and Latency-Aware Networking](#)", International Symposium on Networked Cyber-Physical Systems (NetCPS, Poster Session), Munich, September 2016
3. Shuai, Yongtao; Herfet, Thorsten: "Improving User Experience in Low-Latency Adaptive Streaming by Stabilizing Buffer Dynamics", IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, January 2016





## Devices

- ▶ Gateway (Orchestration Master, Firewall, VPN Endpoint).
- ▶ Nodes (Switching Units).
- ▶ Devices (End-Hosts).
- ▶ Relays (Tx Optimization).



## Partners

- ▶ SmartFactory (Kaiserslautern, DE)
- ▶ Nokia Networks (Munich, DE)
- ▶ HIIT (Helsinki, FI)
- ▶ Manipal University (IN)
  
- ▶ We operate [ON@UoS](https://on@uds.de)

Entities:

Node

Host

Remote Network

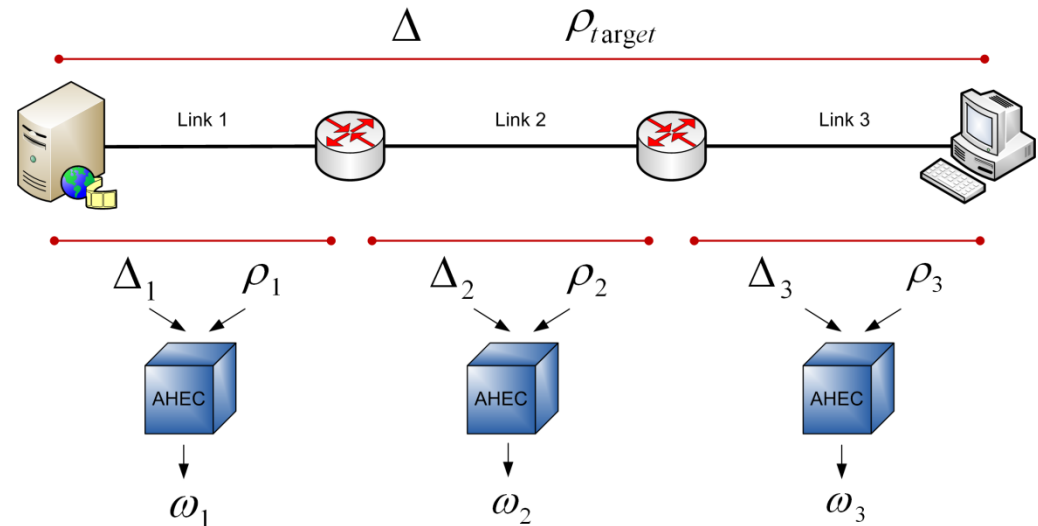
Locations:

<b>Saarland University (UoS)</b> (Saarbrücken, DE)	<b>Nokia Networks (NN)</b> (Munich, DE)	<b>Helsinki Institute of Tech. (HIIT)</b> (Helsinki, FI)
<b>Smart Factory (SF)</b> (Kaiserslautern, DE)	<b>Manipal University (MU)</b> (Jaipur, IN)	



## ▶ Multi-Link / Multi-Hop

- uses individual segment properties
- intermediate nodes act as error correction relay
- apply AHEC as atomic unit
- Expect significant coding gain



**Error Correction in „Overlay Mode“**



## Future Media Internet



Pure RTP Transmission



M-TCP Transmission



UNIVERSITÄT  
DES  
SAARLANDES



Packet Loss Rate  
Adjustment



Packet Loss Rate  
(actual)



target actual  
Coding Delay



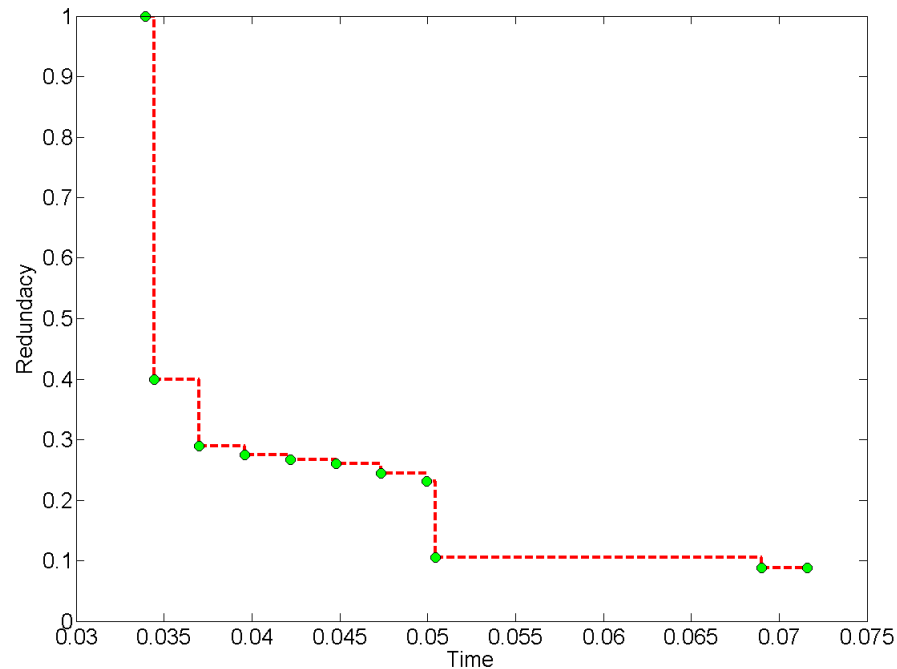
optimal actual  
Redundancy  
Information



target actual  
Packet Loss Rate  
(resulting)



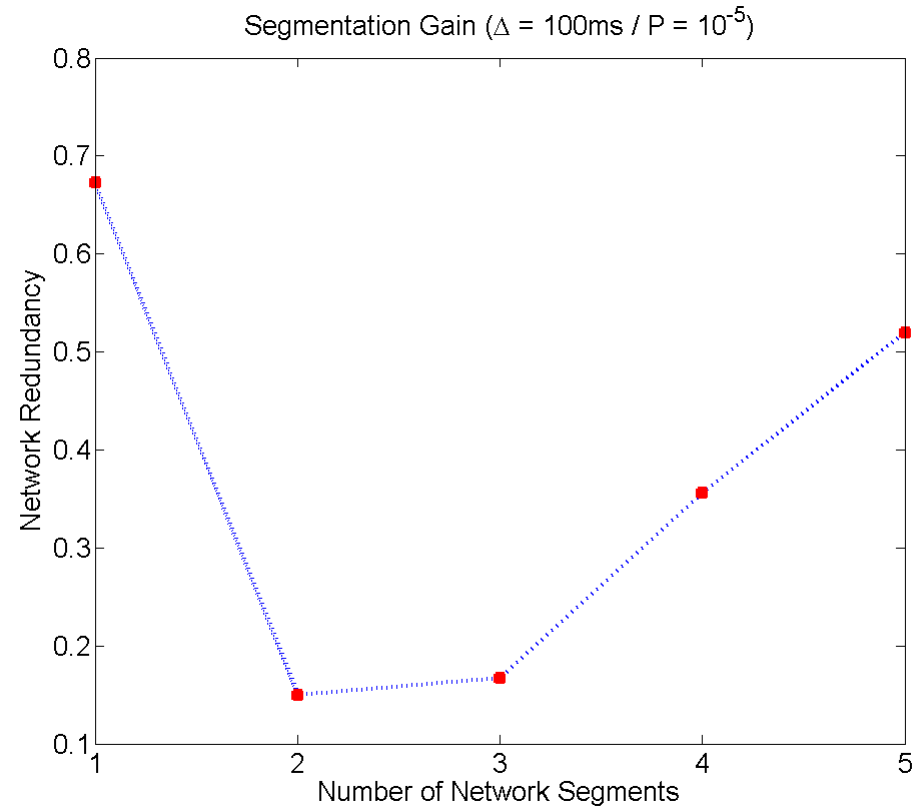
- ▶ Typical redundancy/coding-time behavior of a link:



- More time leads to a higher coding efficiency (Shannon Theorem)
- Monotonously decreasing shape
- Discrete values (time/redundancy)

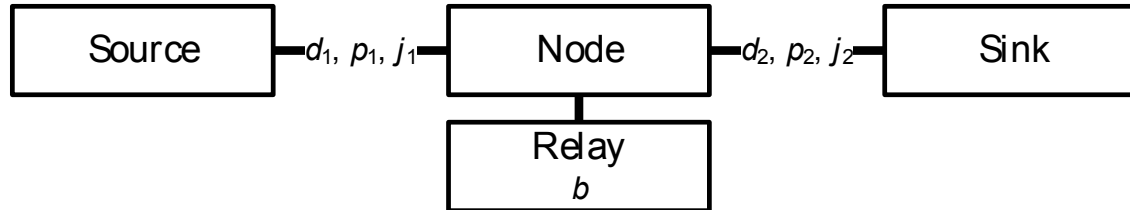


- ▶ Bathtub shape of overall coding gain due to:
  - Available time per segment decreases linearly ...
- ▶ ... whereas ...
  - Redundancy per link increases super-linearly





## Setup



## Results

- ▶ *Metric*: Transmission time [ms] for a TCP data stream.

## Parameters

- ▶ Delays:  $d_1=50\text{ ms}$ ,  $d_2=10\text{ ms}$
- ▶ Jitter:  $j_1=5\text{ ms}$ ,  $j_2=1\text{ ms}$
- ▶ Loss:  $p_1=10^{-8}$ ,  $p_2=3\%$

## Consequences

- ▶ Mean and StdDev decrease (latencies decrease and jitter reduces).
- ▶ Median and minimum increase as there is more overhead.

Case	Mean	StdDev	Median	Min	Max
E2E	146.686	121.456	122.319	110.234	4753.730
TTS	140.427	104.601	123.371	110.660	5236.300
	+4.457%	+16.114%	-0.852%	-0.384%	-9.216%



Case	$d_1$ (ms)	$j_1$ (ms)	$p_1$ (%)	$d_2$ (ms)	$j_2$ (ms)	$p_2$ (%)	b [Byte]	$A_{12}$	p
Buffer (Very Small)	50	5	1e-06	10	1	1e-06	16	0.065	<1e-04
Buffer (Small)	50	5	1e-06	10	1	1e-06	64	0.729	<1e-04
Buffer (Big)	50	5	1e-06	10	1	1e-06	1024	0.892	<1e-04
High Error Rate	50	5	1e-06	10	1	3	1024	0.917	<1e-04
No Errors	50	5	0	10	1	0	1024	0.901	<1e-04
High Jitter	50	50	1e-06	50	50	1e-06	1024	0.688	<1e-04
Natural	50	5	1e-04	5	1	1e-04	1024	0.841	<1e-04

- ▶ Scenario: Two Links, one split using 1 relay
- ▶ Parameters: Delay, Jitter, PLR, Relay Buffer Size
- ▶ Measured Metric: Transmission Time for all Data of a TCP Stream
- ▶ Performance Metric:  $A_{12}$  score (stochastic superiority<sup>1</sup>)  
( $A_{12} < 0.5$ : E2E better,  $A_{12} > 0.5$ : TTS better)

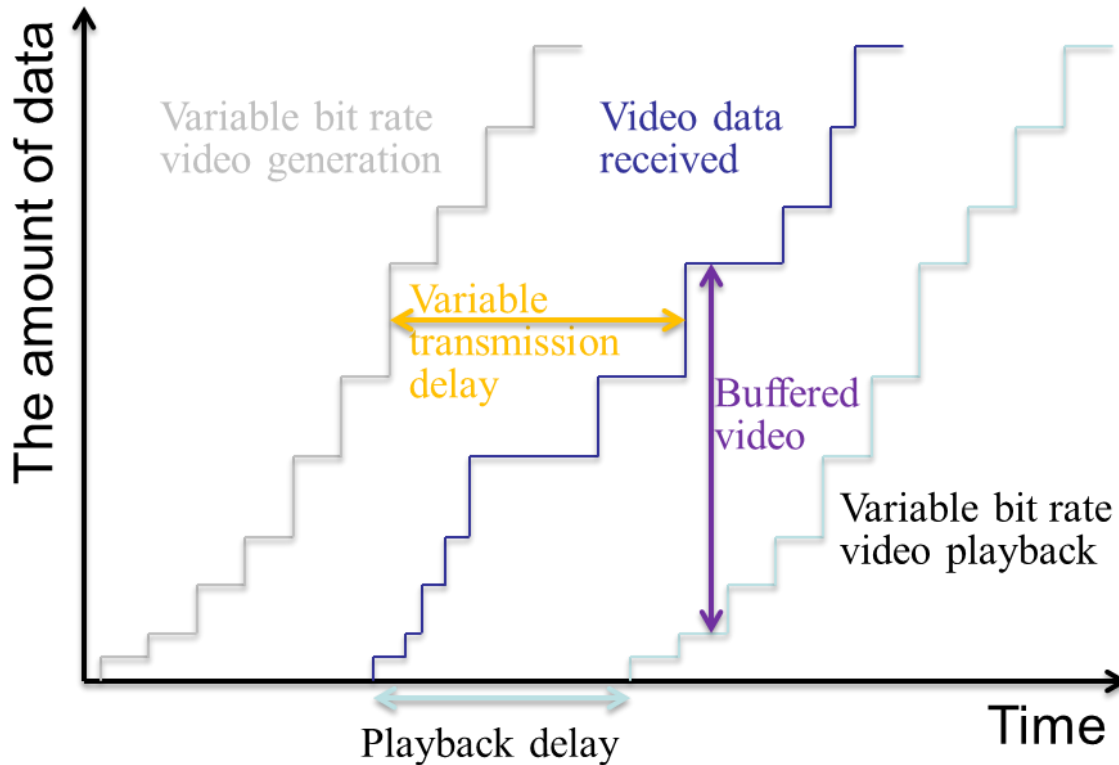
<sup>1</sup> A. Vargha and H. D. Delaney, "A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong", Journal of Educational and Behavioral Statistics, vol. 25, no. 2, pp. 101–132, 2000.



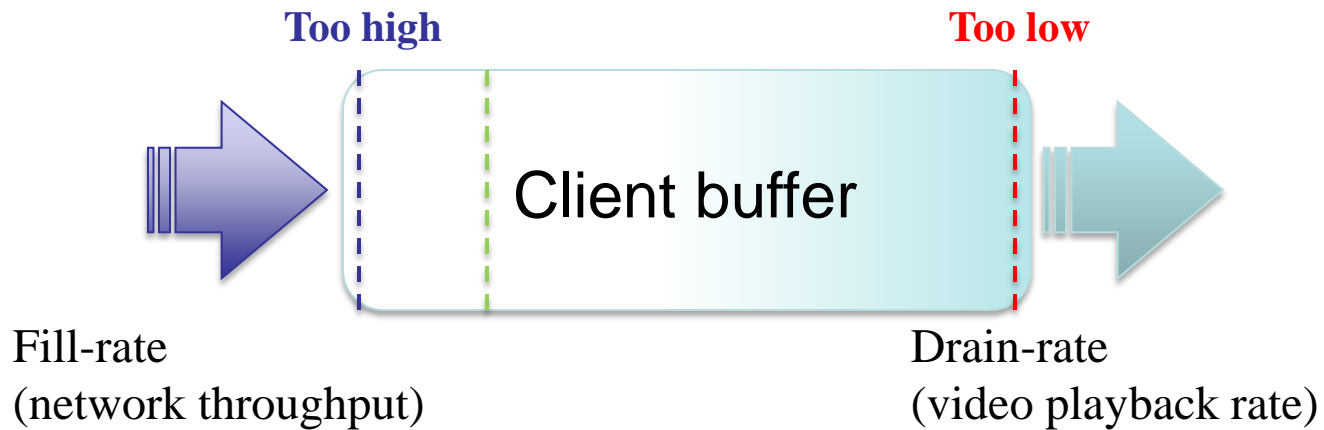


- ▶ Low Latency Video Streaming
- ▶ Software-Defined Networking
  - 5G Broadcast through SDNs
  - Predictably-reliable Real-Time Transport<sup>1</sup>
  - Transparent Transmission Segmentation<sup>2</sup>
- ▶ **Buffer Dynamics Stabilizer<sup>3</sup>**

1. Gorius, M.: "[Adaptive Delay-constrained Internet Media Transport](#)", Dissertation, UdS, December 2012
2. Schmidt, Andreas; Herfet, Thorsten: "[Approaches for Resilience- and Latency-Aware Networking](#)", International Symposium on Networked Cyber-Physical Systems (NetCPS, Poster Session), Munich, September 2016
3. Shuai, Yongtao; Herfet, Thorsten: "Improving User Experience in Low-Latency Adaptive Streaming by Stabilizing Buffer Dynamics", IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, January 2016

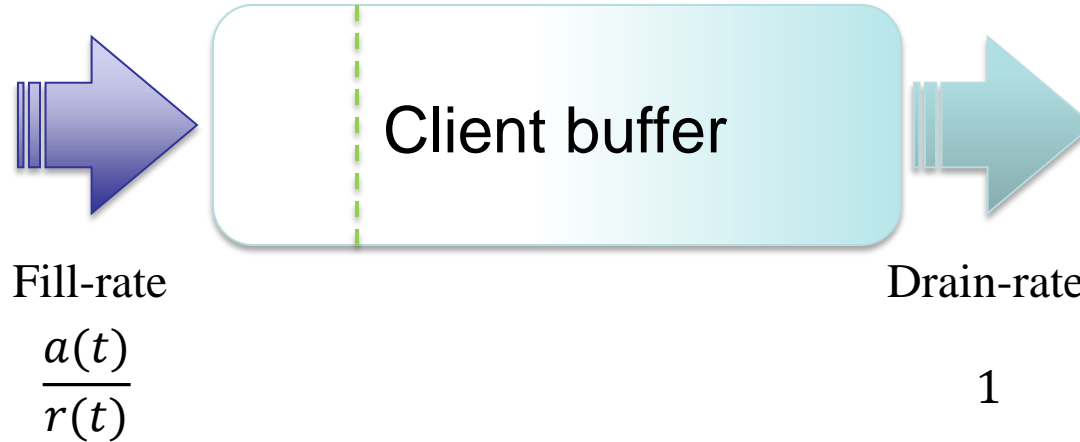


- ▶ Buffering Delay is buffered video in seconds.
- ▶ We achieve low-latency dynamic video streaming with buffering delays as low as the chunk-duration.



## Quality Selection

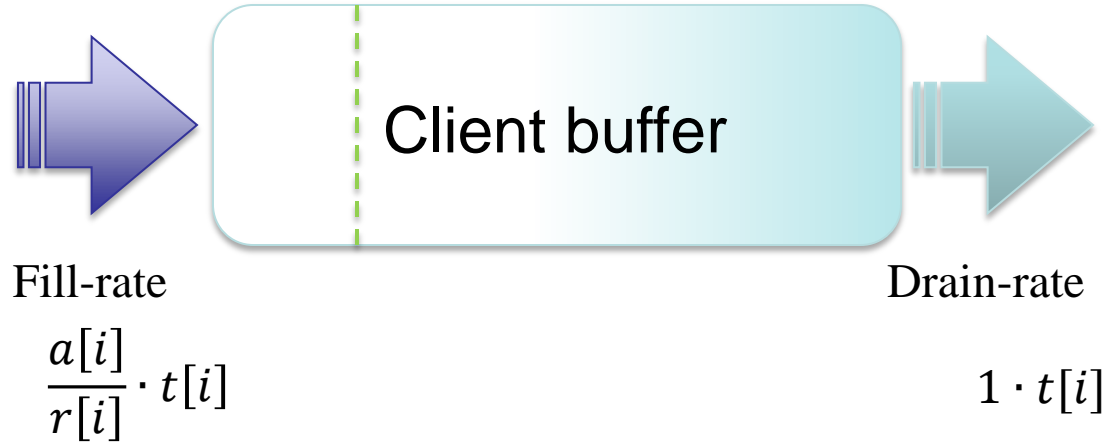
Stabilizing the buffer to the desired level by regulating the drain-rate, i.e. by selecting a video bit rate for the chunk.



Express the buffer level in **seconds of video**.

$a(t)$  : the throughput rate achieved at the time  $t$

$r(t)$  : the selected video bit rate at the time  $t$

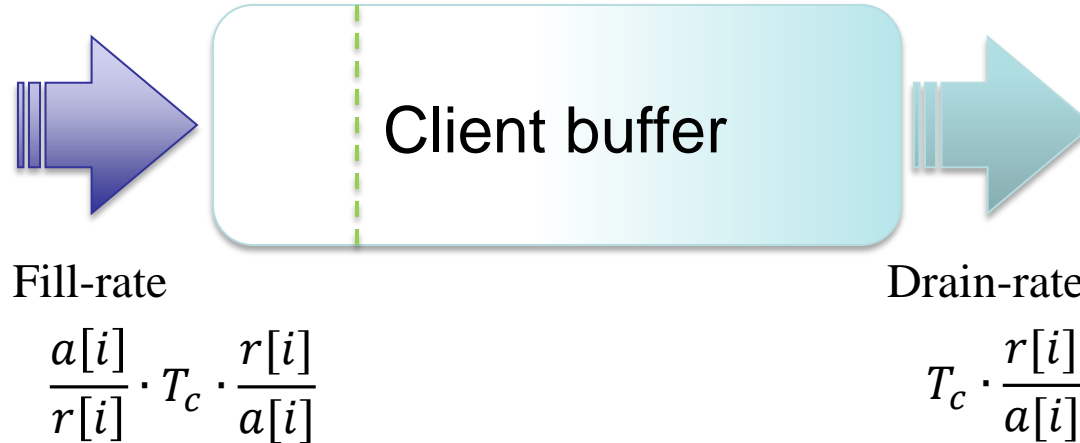


Compute the buffer level **every discrete chunk**.

$a[i]$  : the throughput rate achieved during the reception of chunk  $i$

$r[i]$  : the selected video bit rate of chunk  $i$

$t[i]$  : the reception duration of chunk  $i$

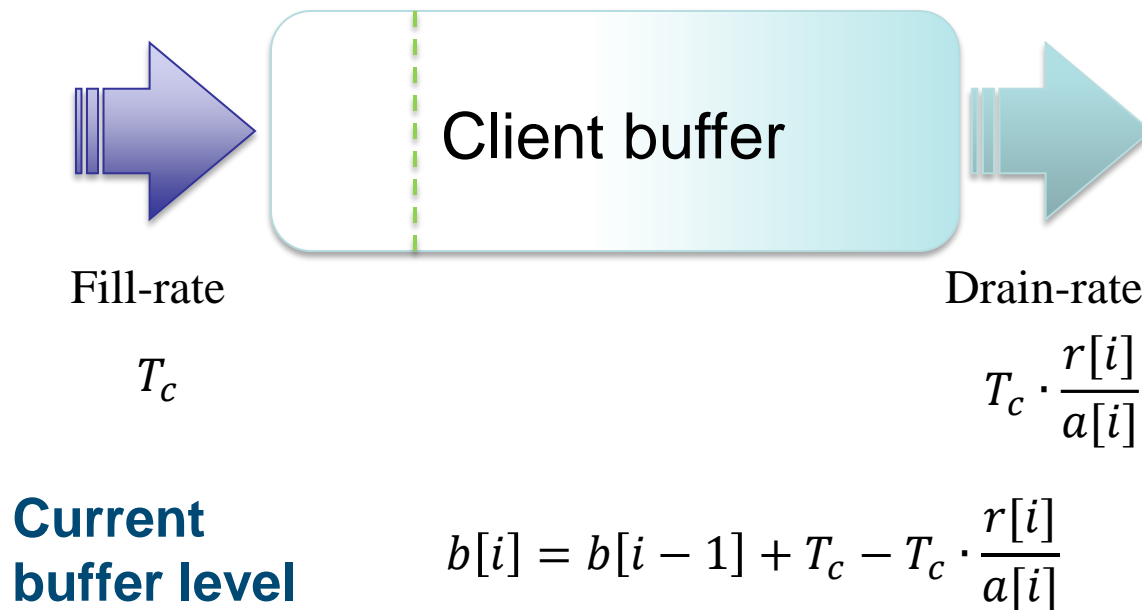


Compute the buffer level **every discrete chunk**.

$a[i]$  : the throughput rate achieved during the reception of chunk  $i$

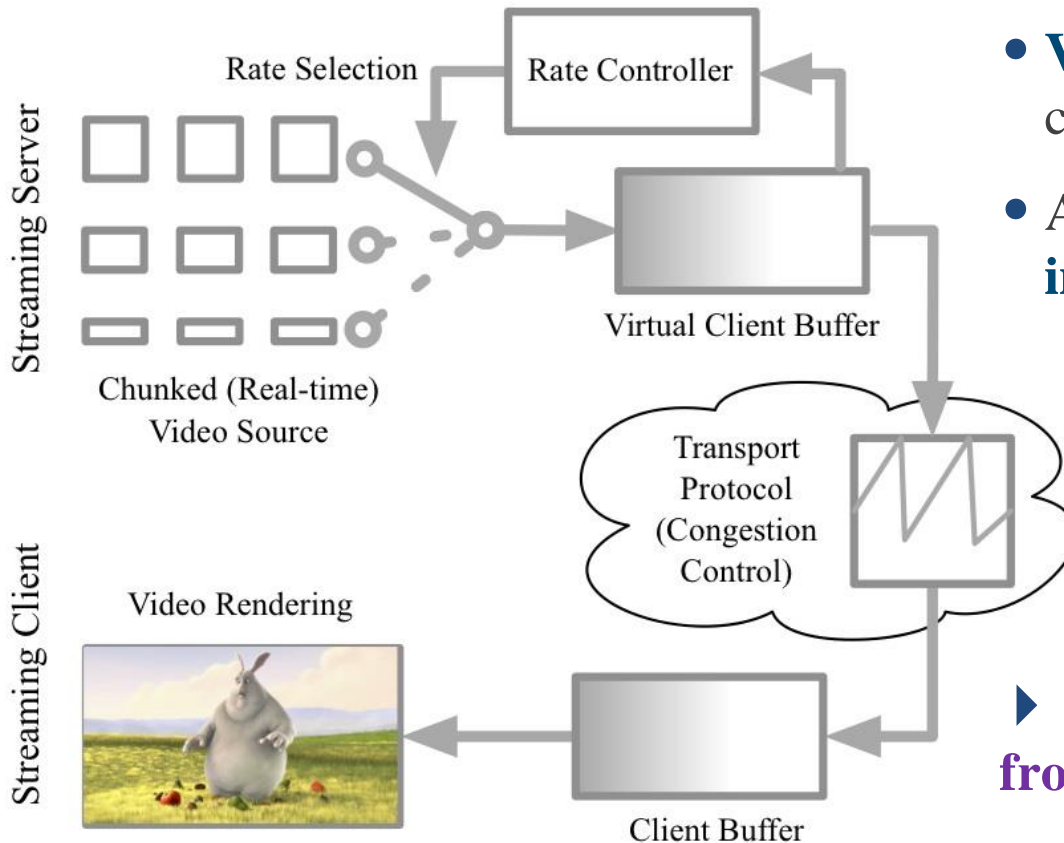
$r[i]$  : the selected video bit rate of chunk  $i$

$T_c$  : the chunk duration



$b[i]$  : the buffer level (in seconds) when the client finishes the reception of chunk  $i$





- **Virtual client buffer** simulates client buffer on the server.
- A rate control on the server offers **immediate feedback** from clients.

► Our buffer stabilization **benefits from OLAC**:

- **user-perceived quality** improves by up to 68%.



## Demo

A **demo example** of experiments is shown **with a speed of 2.5x**.

DASH

DAST

QAC

DASP

<https://youtu.be/G5R1Uj9fBLo>

<https://youtu.be/HQhIZWGPEYU>



- ▶ Video is going IP (in all domains)
  - Low Latency is extremely important
  - Latency and reliability need to be balanced
- ▶ We introduced:
  - 5G Media Broadcast through SDNs
  - Predictably Reliable Real-Time Transport
  - Transparent Transmission Segmentation
    - Optimized link-segmentation
  - Buffer Dynamics Stabilizer
    - Buffer sizes as small as a single chunk size