Evaluating Reliable Streaming in Mobile Networks

Dissertation Presentation

Florian Metzger

University of Vienna // University of Duisburg-Essen

2015/04/28

Section 1

Introduction

Introduction

Impact of (Reliable) Video Streaming Traffic

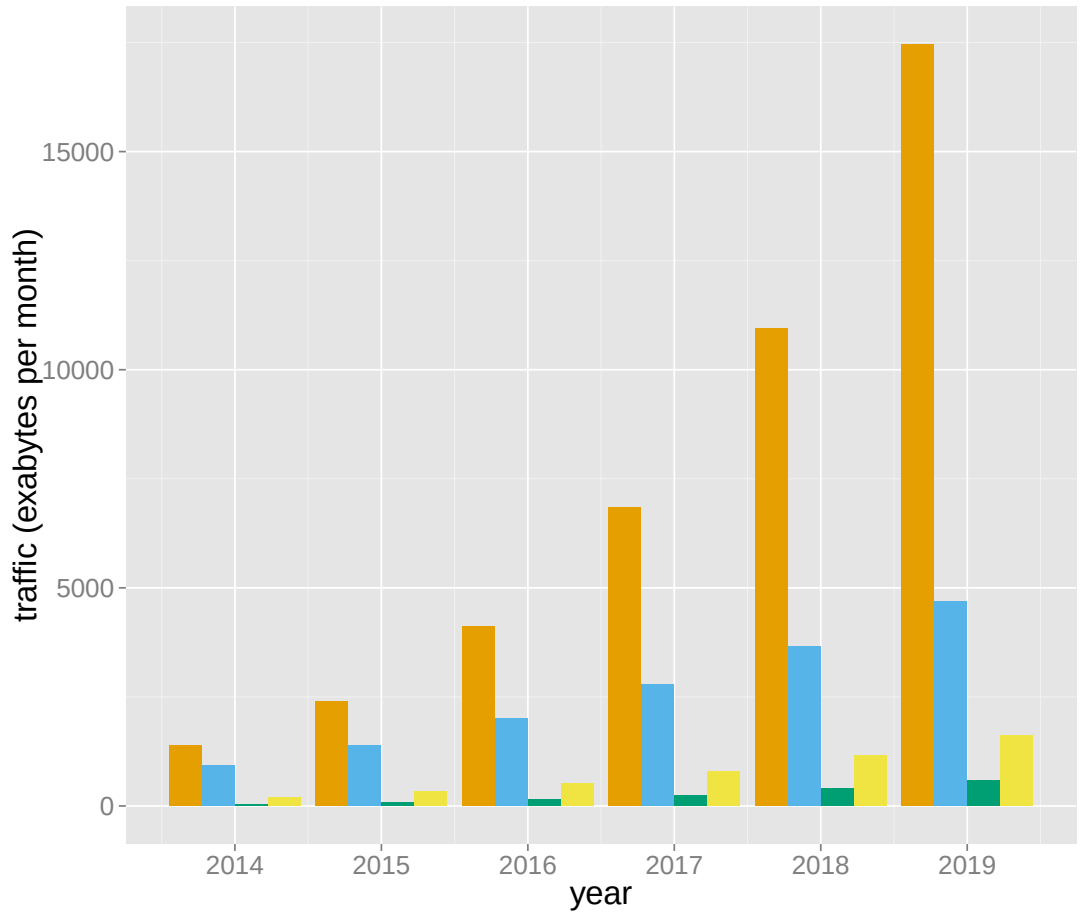
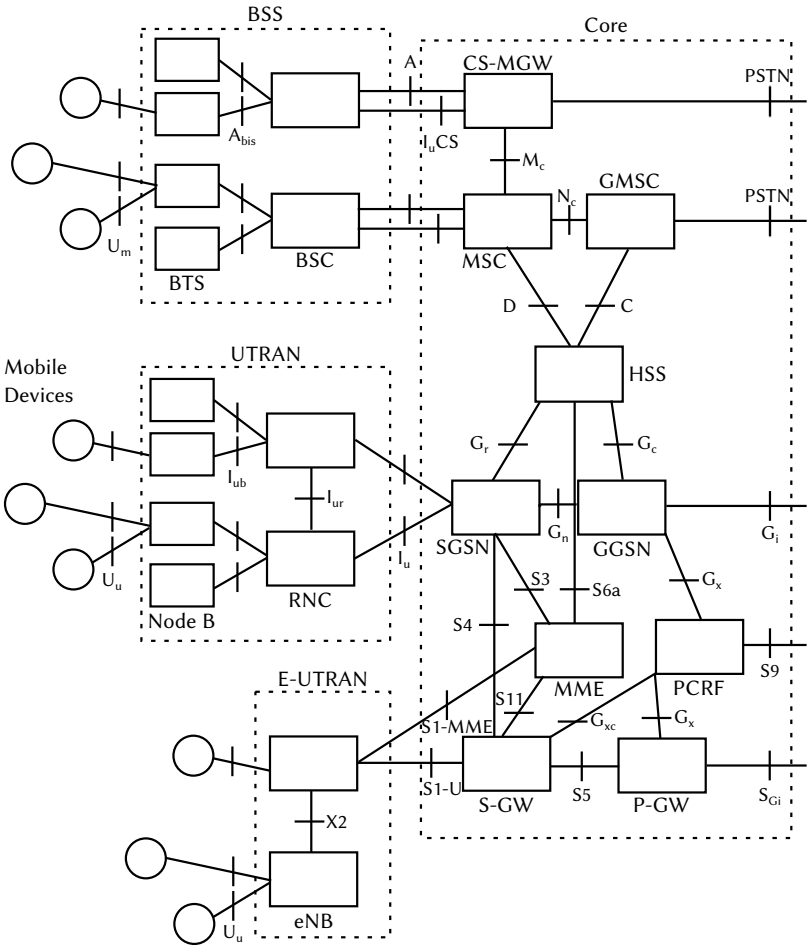


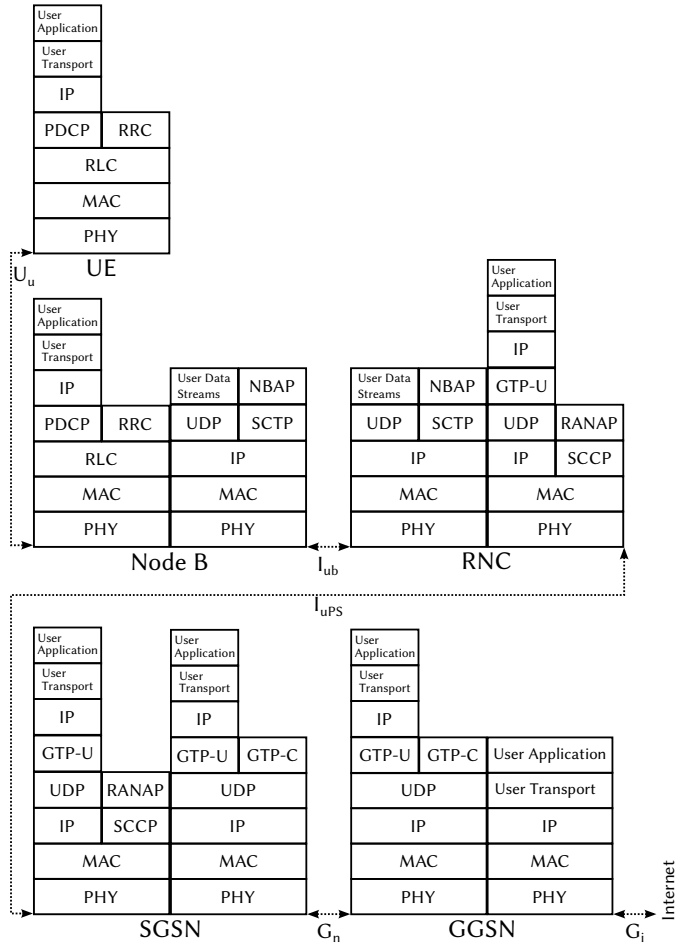
Figure: Cisco's global mobile data traffic forecast (data source: [Cis15]).

Introduction

And Reliable Streaming in Mobile Networks?

Complexity of current mobile network architectures

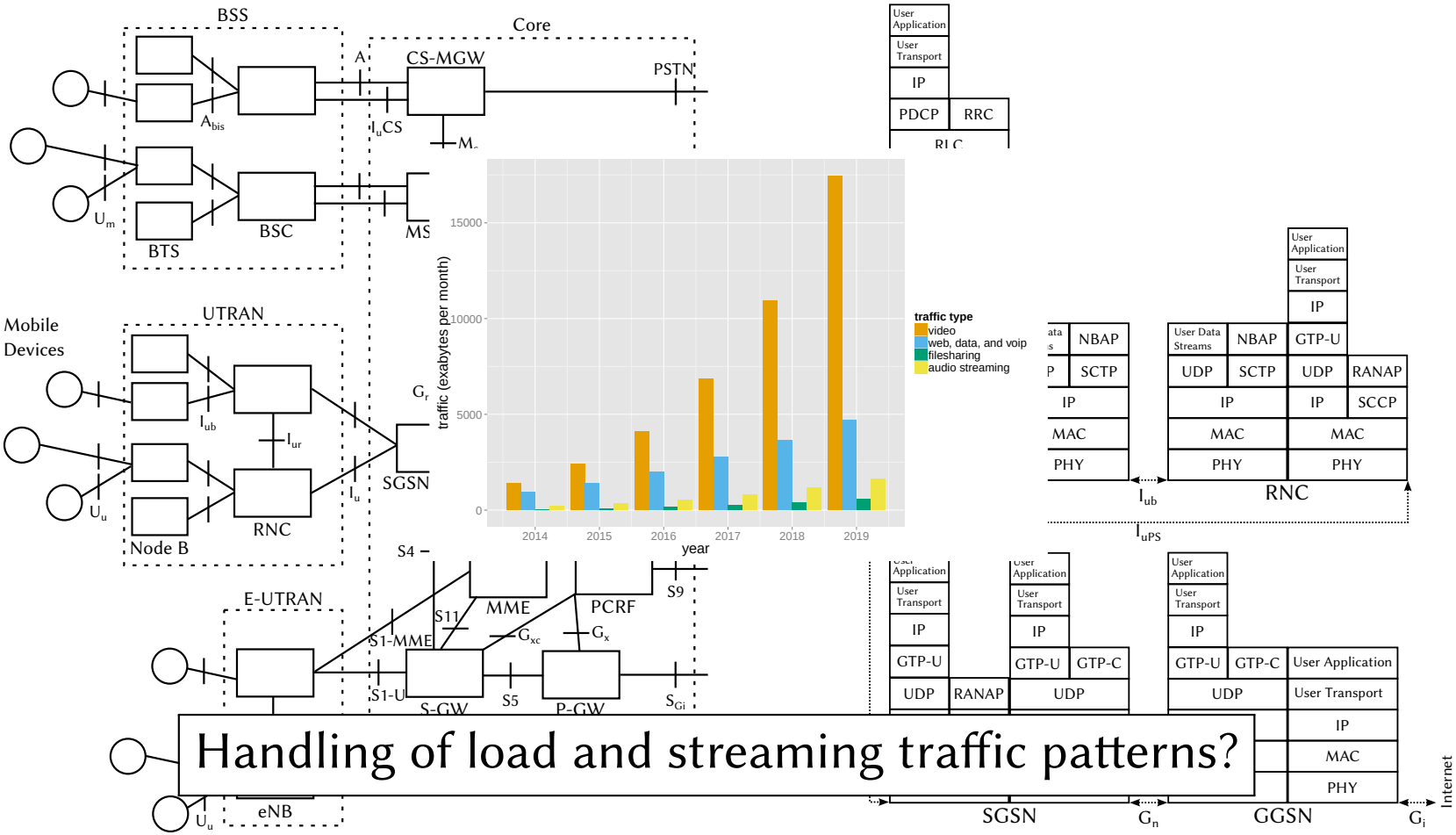




Introduction

And Reliable Streaming in Mobile Networks?

Complexity of current mobile network architectures



Handling of load and streaming traffic patterns?

Introduction

Current State of Affairs

- Both mobile network and streaming traffic shares increasing rapidly
- Large differences between RTP/UDP and reliable streaming
  - Traffic patterns
  - Control and quality adaptivity
- Impact of mobile core network and control plane not well understood
- Interactions between reliable streaming and mobile networks?

Introduction

Current State of Affairs

- Both mobile network and streaming traffic shares increasing rapidly
- Large differences between RTP/UDP and reliable streaming
  - Traffic patterns
  - Control and quality adaptivity
- Impact of mobile core network and control plane not well understood
- Interactions between reliable streaming and mobile networks?

Investigation approach

- Model for reliable streaming measurements and categorization
- Discussion of influence factors on reliable streaming
- CN control plane load model derived from passive measurements
- Creation of easy mobile streaming measurement approaches

Introduction

Thesis Structure

Thesis consists of two intersecting angles

1. Evaluation of a 3G core network
   - Investigation and evaluation of the control plane
   - Modeling and simulating load

Introduction

Thesis Structure

Thesis consists of two intersecting angles

1. Evaluation of a 3G core network
   - Investigation and evaluation of the control plane
   - Modeling and simulating load
2. Investigation of TCP-based video streaming techniques
   - Protocol survey and classification
   - Deriving a model
   - Measurements with the model

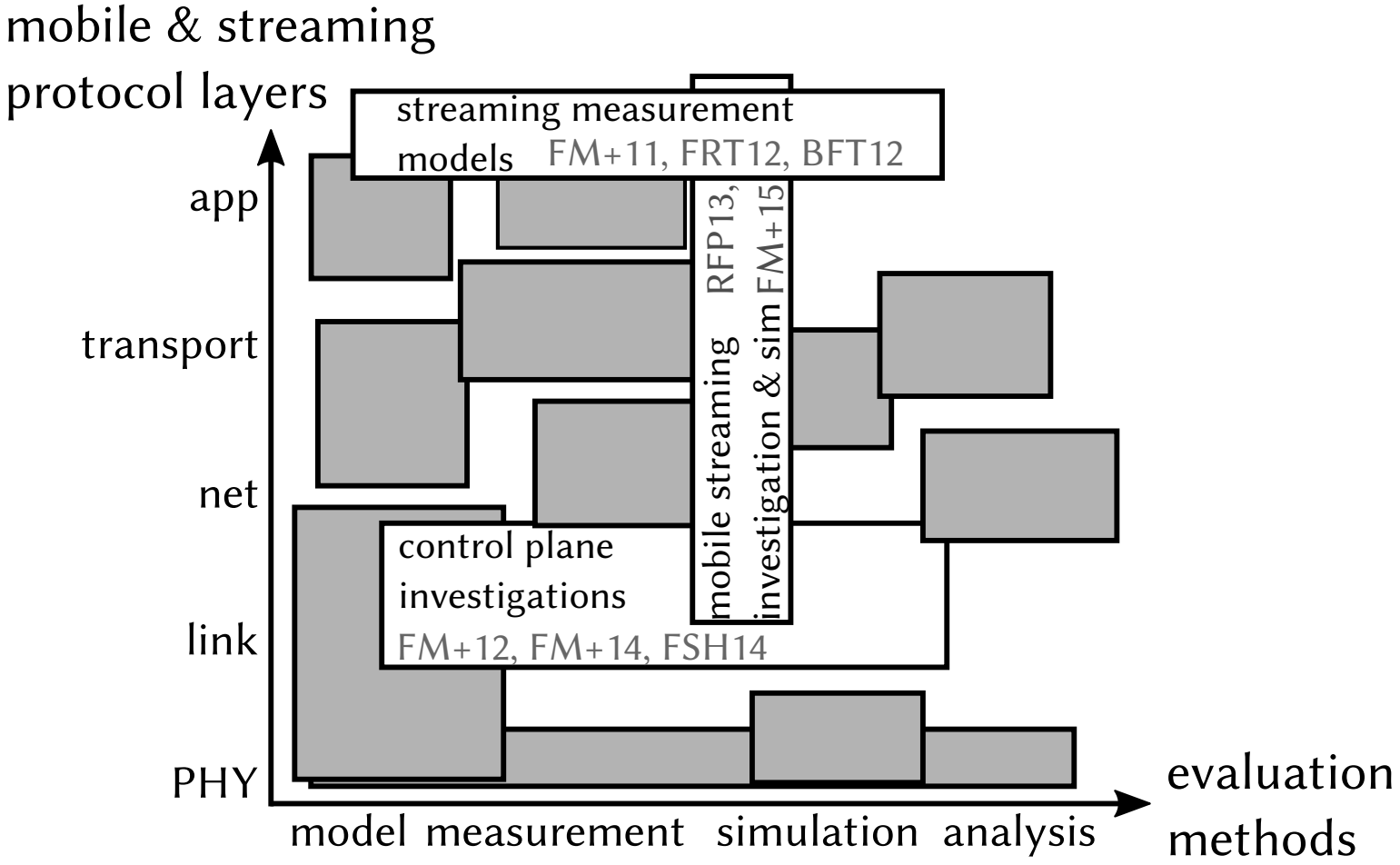
Introduction

Thesis Structure

Thesis consists of two intersecting angles

1. Evaluation of a 3G core network
   - Investigation and evaluation of the control plane
   - Modeling and simulating load
2. Investigation of TCP-based video streaming techniques
   - Protocol survey and classification
   - Deriving a model
   - Measurements with the model
3. Measuring video streaming in a 3G network
   - Investigation of influence factors, potential issues, and solutions
   - Comparison of evaluation approaches (active measurements and simulation)
   - Mobile streaming simulation framework

Placing the Thesis

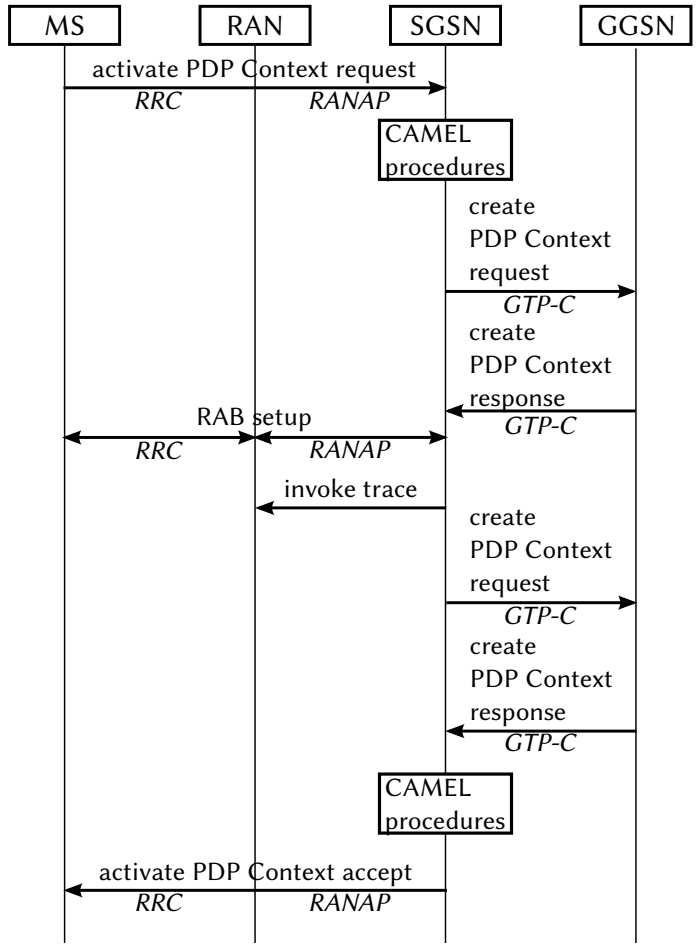


Section 2

Mobile Core Network Architecture

3G Core Network and Control Plane Overview

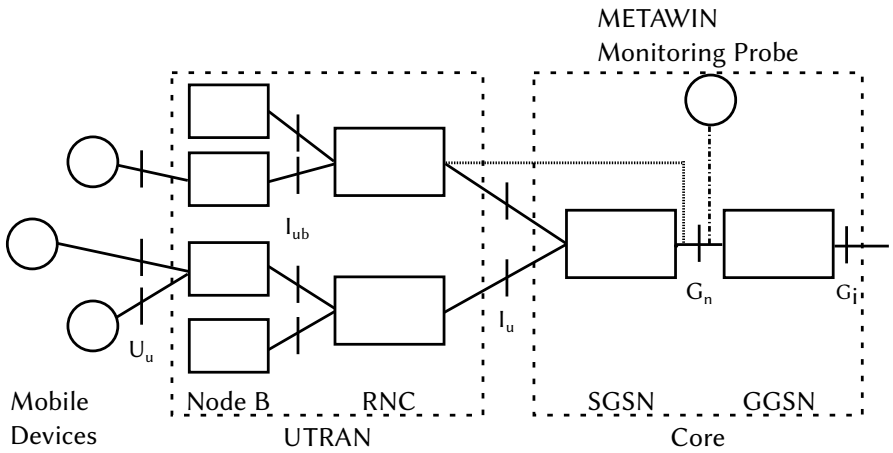
- Everything is connected (cross-node, cross-layer)
  - Action at mobile device will trigger procedures deep inside the network
- Typical action: checking for new mail
  - Activate radio only very briefly to save energy
  - Very little actual traffic
  - Triggers at least PDP context activation and deactivation proc's
  - "signaling storm" phenomenon

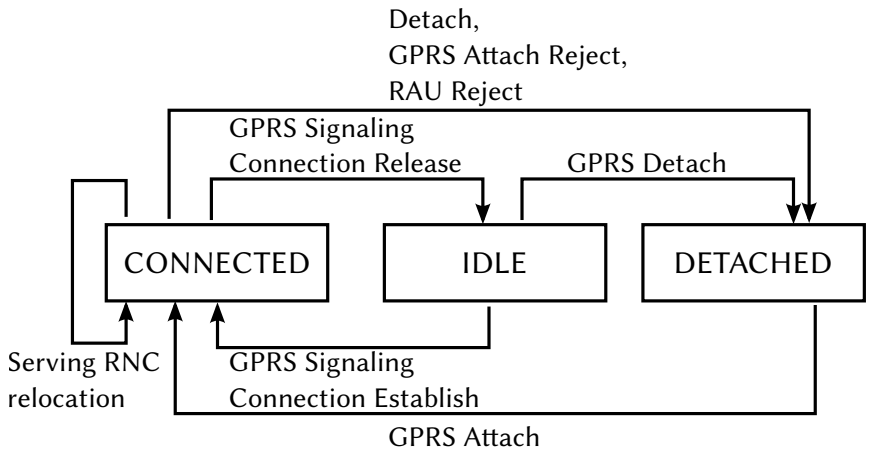


PDP Context activation procedure

GTP Tunneling Concept

- 3G user traffic is encapsulated into tunnels
  - Radio bearer to the device
  - GPRS Tunneling Protocol (GTP) at L4 SGSN ↔ GGSN
- State held at and signaled between core nodes
- Additional overhead and load
  - Network: signaling messages, headers
  - Memory: state keeping
  - CPU: state changes





Mobile Network Load

Definitions

- Traditionally user traffic volume
  - Control plane signaling often not factored in
- Factor in control plane and core network
  - Control plane tasks at each node contribute to total load
  - One factor (CPU, mem, net constraints) at one node will be the bottleneck
  - GGSN as a chief candidate

Mobile Network Load

Definitions

- Traditionally user traffic volume
  - Control plane signaling often not factored in
- Factor in control plane and core network
  - Control plane tasks at each node contribute to total load
  - One factor (CPU, mem, net constraints) at one node will be the bottleneck
  - GGSN as a chief candidate

Reasoning for GGSN as load bottleneck

- Single instance responsible for whole (or complete region) network
- Handles PDP Context signaling and state for each device/tunnel
- Involved in almost all network-wide signaling procedures
- GTP Mobility Management signaling from SGSN

Load Influencing Factors

Many sources of influence, both internal as well as external
Network-centric factors

- Specific interaction of state machines and protocols
- Inactivity timers (e.g. moving device from connected to idle/detached)

User-centric factors

- Type and OS of device, access technology
- Installed and active applications
- Specific usage, time-of-day, and traffic patterns

Load Influencing Factors

Many sources of influence, both internal as well as external
Network-centric factors

- Specific interaction of state machines and protocols
- Inactivity timers (e.g. moving device from connected to idle/detached)

User-centric factors

- Type and OS of device, access technology
- Installed and active applications
- Specific usage, time-of-day, and traffic patterns

Load evaluation approach

Investigate load at **GGSN** on the basis of **GTP tunnel characteristics**

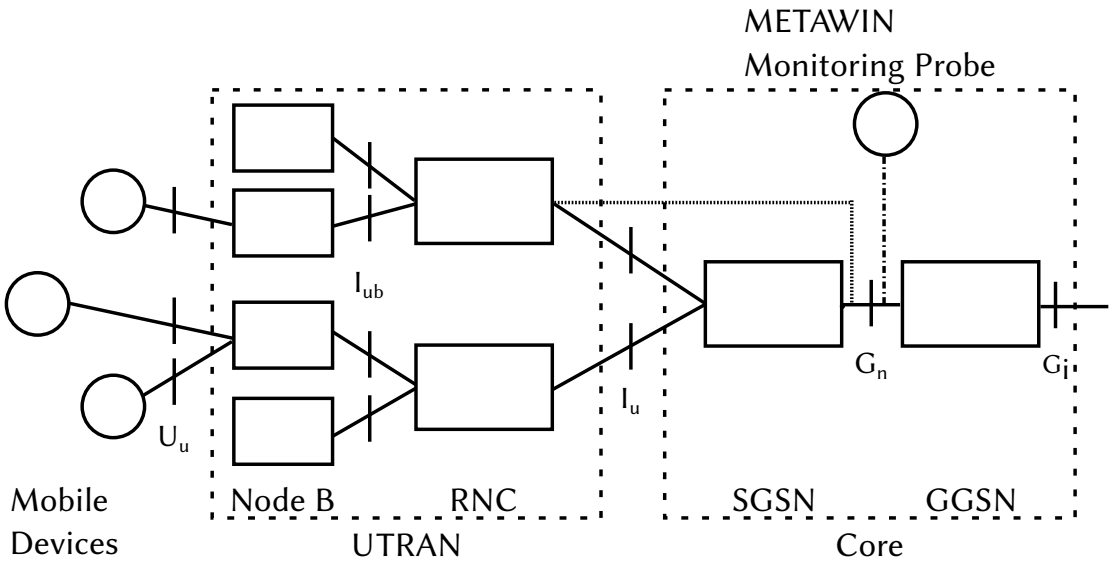
Section 3

Evaluating Mobile Signaling Traffic and Load

Trace Evaluation

Recorded dataset

- One week long passive measurements in an operator's core network (METAWIN, April 2011)
- 2.2Bn user traffic records, 410M GTP tunnel management messages
- Trace data anonymized, reduced to select fields per signaling interactions
- Device and OS identification through TAC



Trace Evaluation

Initial Findings

- Link load of control traffic insignificant
  - GTP messages contribute max. 0.7 % to total traffic
- 18 % of devices cause signaling without transmitting any user traffic
- High user traffic volume ≠ high amount of signaling messages
  - E.g. 3G dongles responsible for 75 % of traffic, but only 10 % of GTP signaling messages

Trace Evaluation

Initial Findings

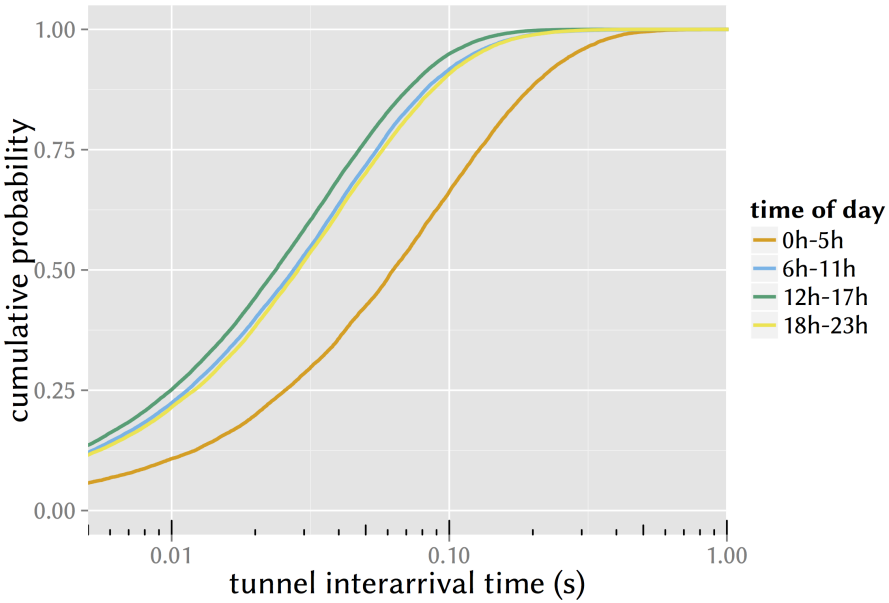
- Link load of control traffic insignificant
  - GTP messages contribute max. 0.7 % to total traffic
- 18 % of devices cause signaling without transmitting any user traffic
- High user traffic volume ≠ high amount of signaling messages
  - E.g. 3G dongles responsible for 75 % of traffic, but only 10 % of GTP signaling messages

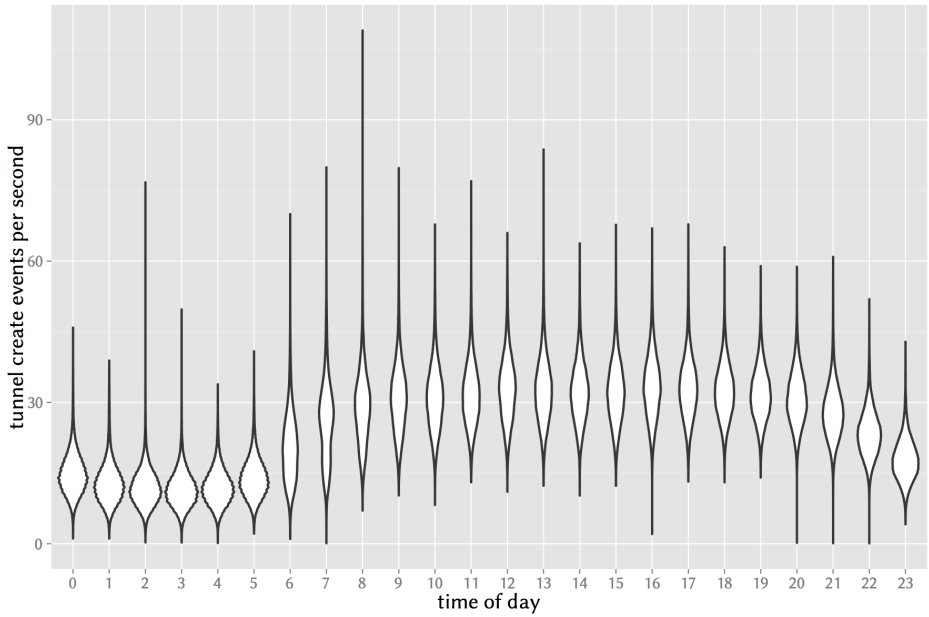
Further investigations for load characterization

- Build a queuing theory system model for GTP tunnels at the GGSN
- Closely investigate tunnel arrival and service process
- Derive distributions for these processes

Tunnel Arrivals

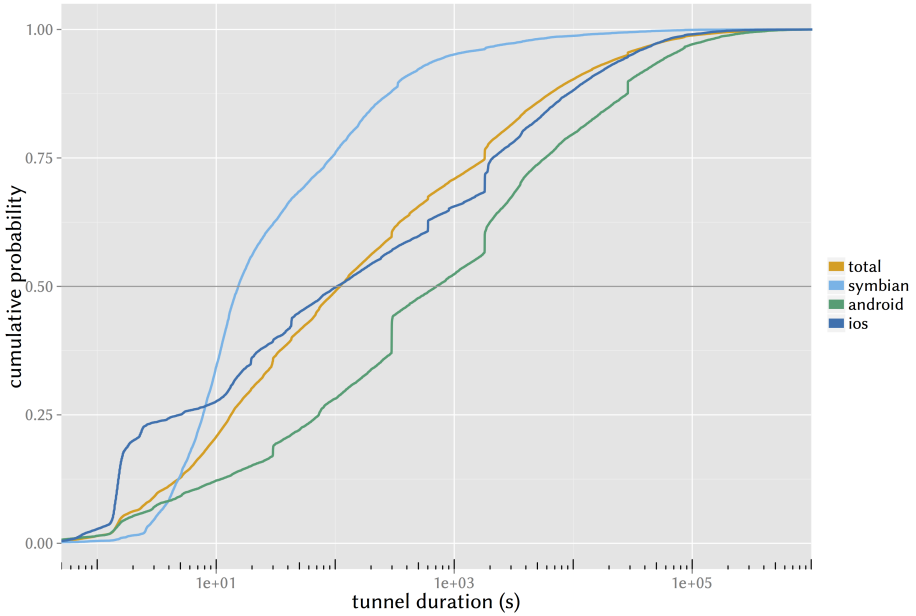
By Time of Day

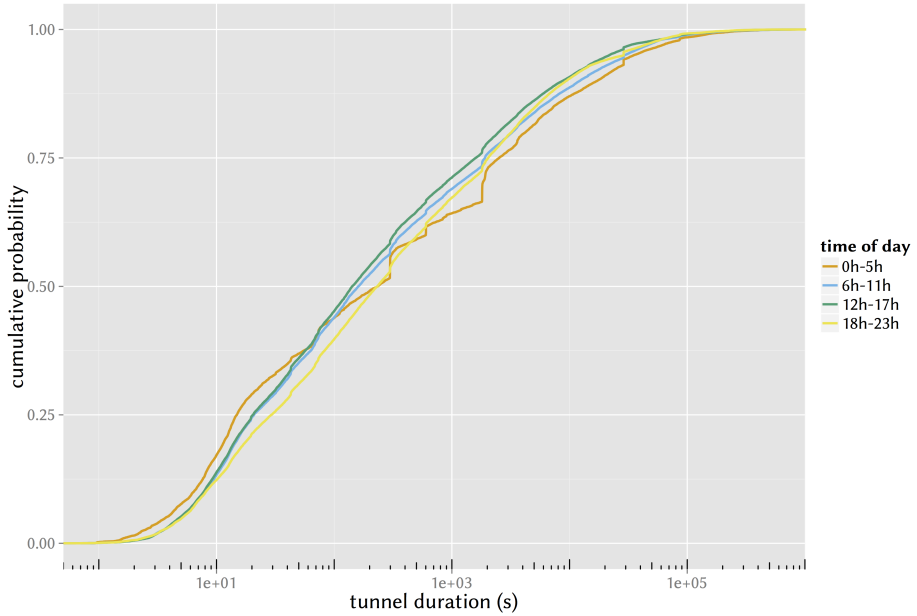




- Strong time of day dependence with busy hour in the early afternoon
- Bimodal character of arrival rate over the total time of day
- Poisson tunnel arrival process with rate $\lambda(t)$ fitted with trace data, $R \geq 0.98$

Tunnel Durations

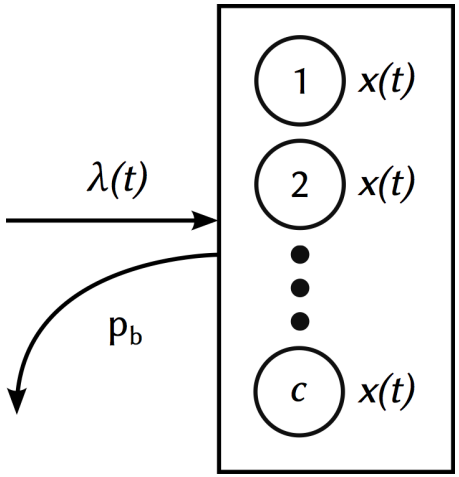




- Strong influence of user device type and OS on duration
- "Signalling storm" visible in CN: many short tunnels for specific devices
- Also dependent on time of day
- Tunnel duration $x(t)$ fitted with rational function, $R \geq 0.99$

Deriving a Model

Monolithic GGSN Queuing Model

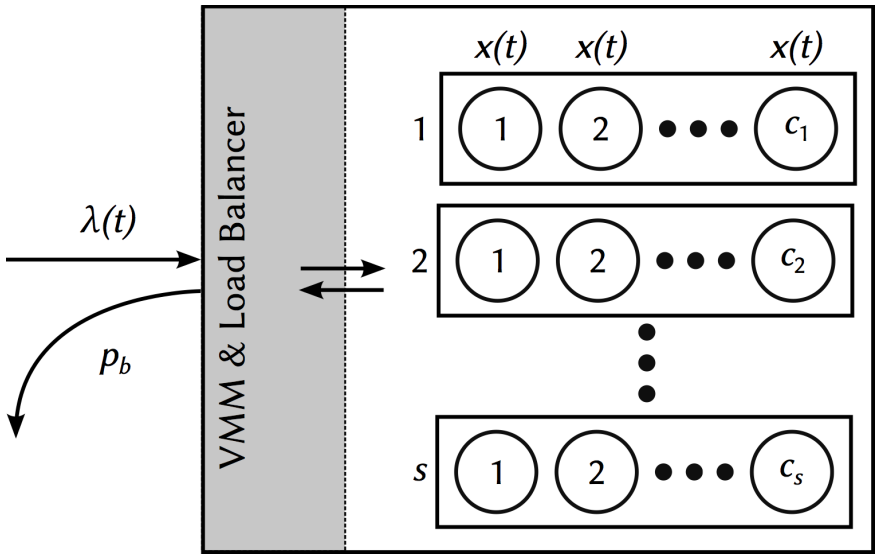


- Model for GGSN load investigation based on tunnel arrivals
- Factors in time of day, but not device type
- GGSN can serve c tunnels in parallel, limited by network/processing load and signaling/state overhead
- If GGSN is full, reject new tunnels, leads to blocking probability p_b

→ Non-stationary Erlang loss model $M(t)/G(t)/c/0$

Expanding the Model

Virtualized GGSN Queuing Model



- Suggestion for a virtual GGSN architecture for 3G networks
- Same arrival and serving time process, no queue
- Load balancing strategies for VM management and tunnel placement
- Could incur additional blocking (virtualization overhead)
- Additional dimension for scaling compared to monolithic model

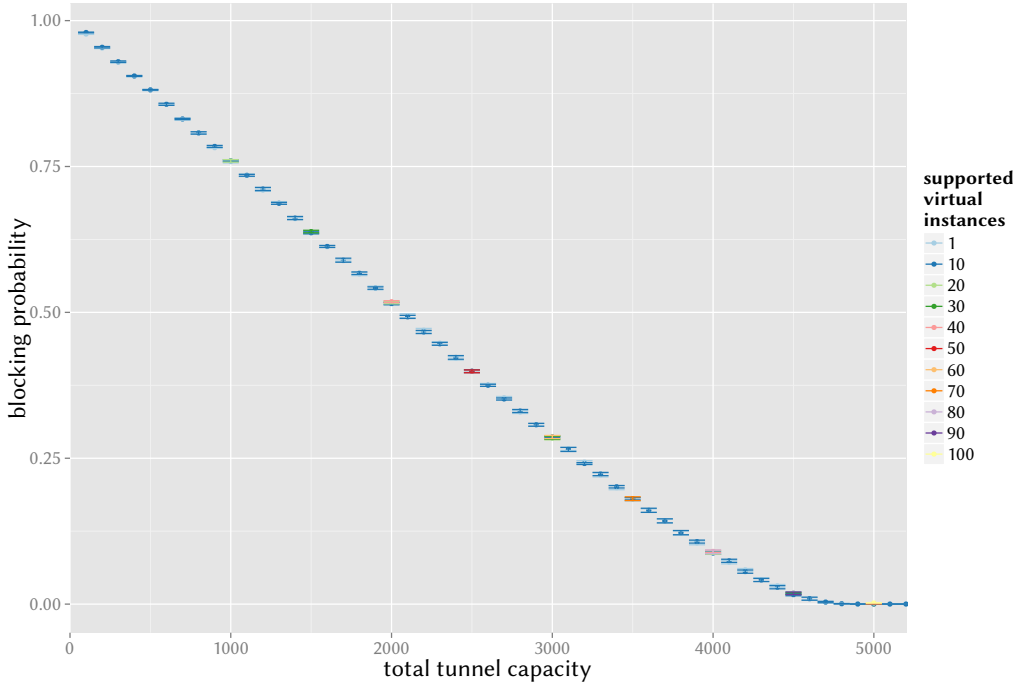
→ $M(t)/G(t)/\|\vec{c}\|_1/0$ ($|\vec{c}|_1$: total tunnel capacity of all s VMs)

Model Evaluation/Simulation

- No tractable analytic solution for $M(t)/G(t)/c/0$ models
- Use queuing simulation with the fitted distributions
- Evaluate blocking probability and VM usage based on tunnel capacity

Model Evaluation/Simulation

- No tractable analytic solution for $M(t)/G(t)/c/0$ models
- Use queuing simulation with the fitted distributions
- Evaluate blocking probability and VM usage based on tunnel capacity



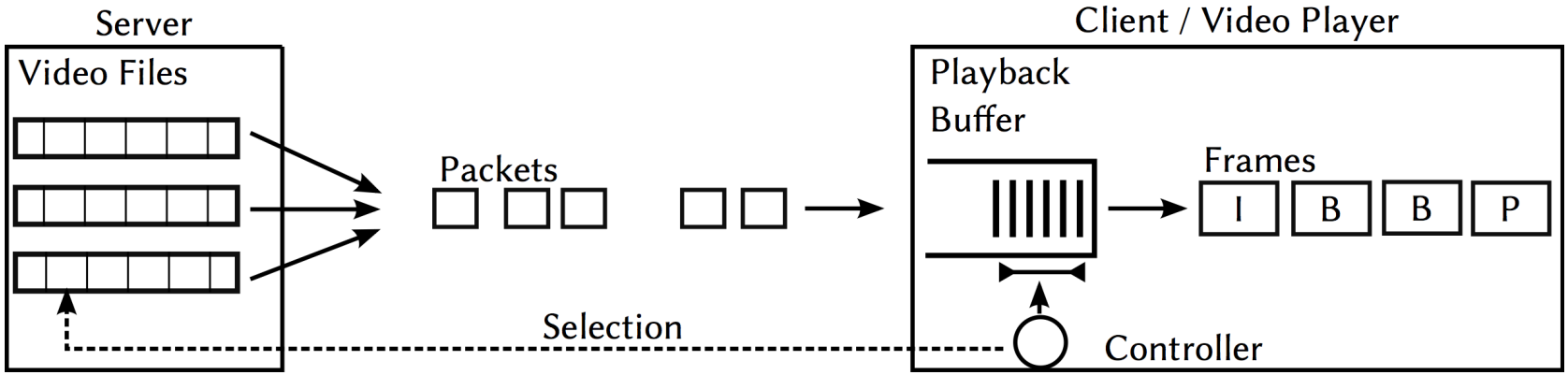
- Monolithic and virtualized GGSN scale equally with supported tunnels
- Negligible virtualization overhead, but potential energy efficiency gains
- Usage as load prediction tool for operators

Section 4

Modeling Reliable Streaming

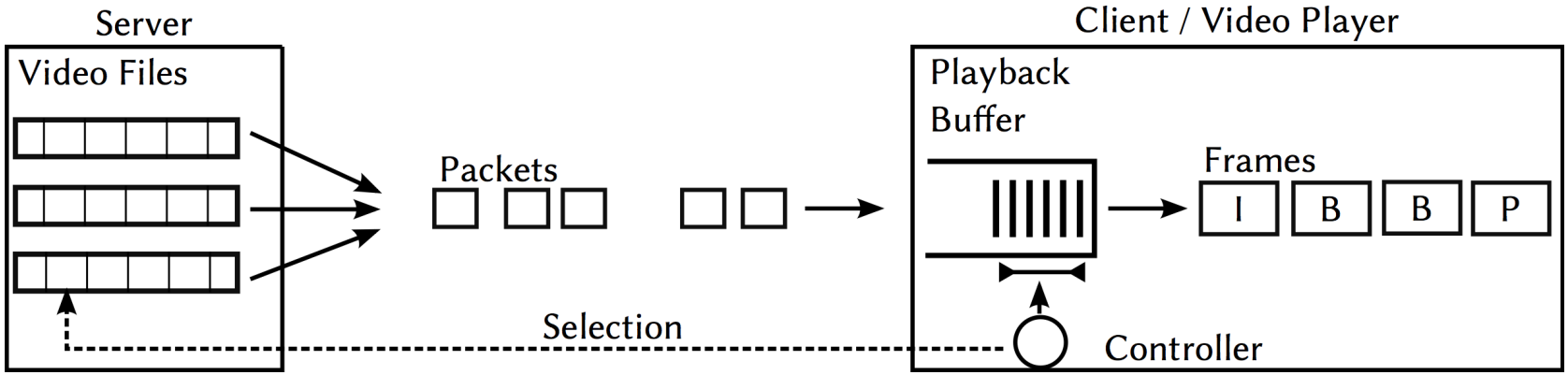
“Reliable” Video Streaming?

- Uses TCP+HTTP instead of “classic” UDP+RTP
- Offers different properties to RTP
  - Reliability, pull-based control, explicit quality adaptation
- No standardized approach, difficult to compare



“Reliable” Video Streaming?

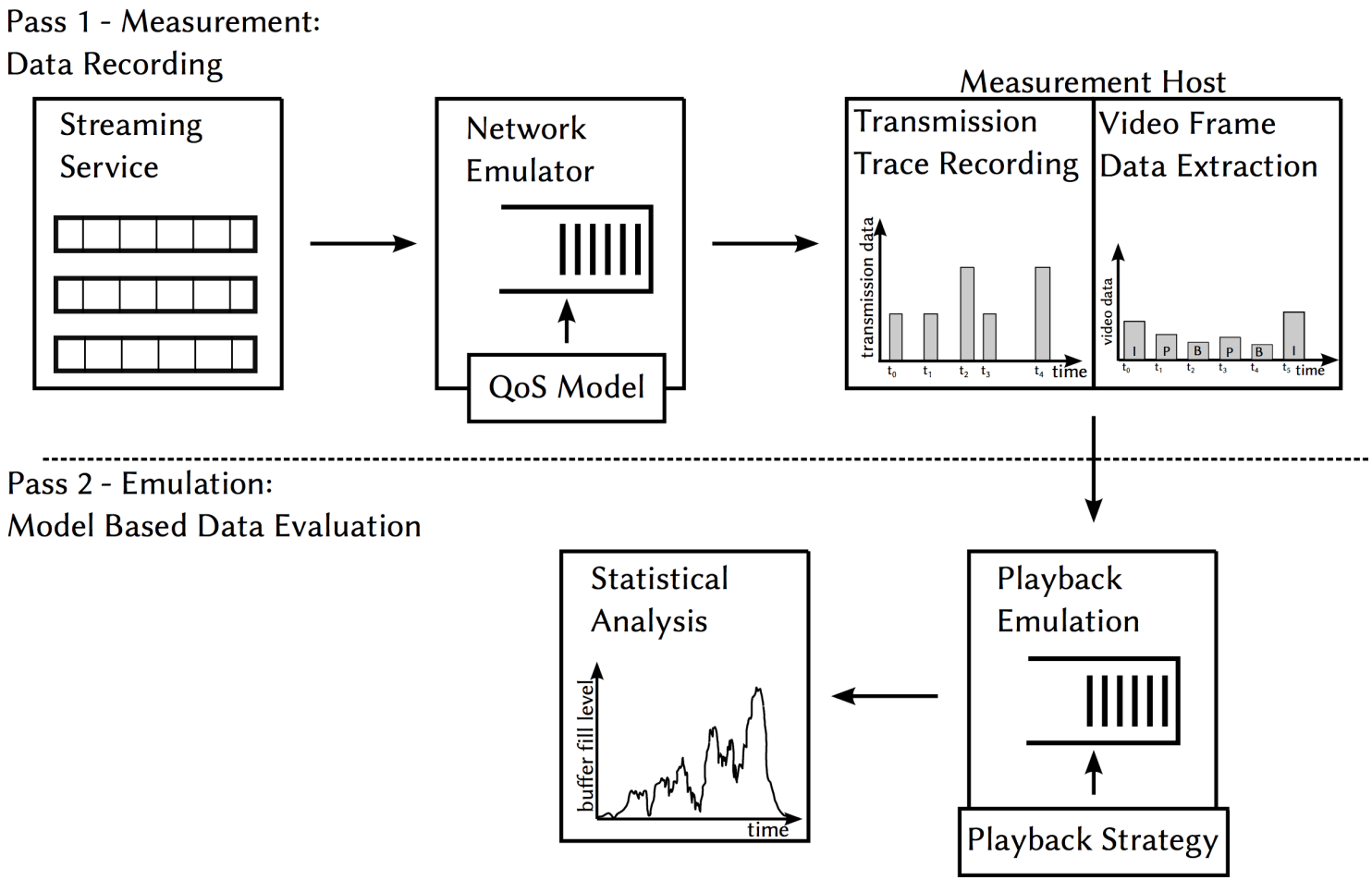
- Uses TCP+HTTP instead of “classic” UDP+RTP
- Offers different properties to RTP
  - Reliability, pull-based control, explicit quality adaptation
- No standardized approach, difficult to compare



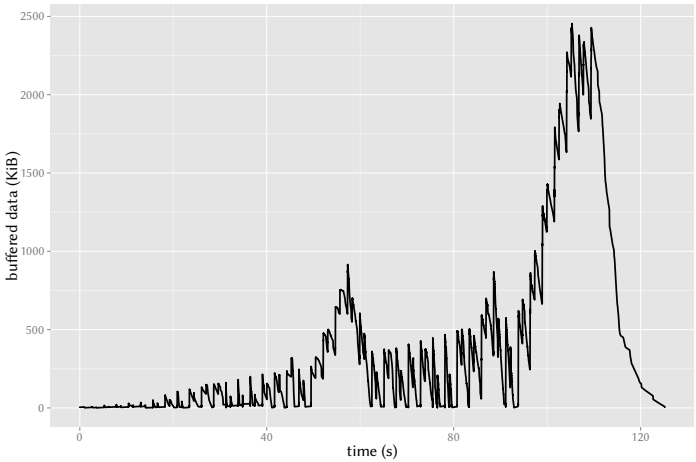
Measurement Approach

- Transmission process has no influence on the image quality (reliability)
- Evaluate streaming solely on the basis of buffering behavior
- Number and duration of stalls as metric for non-adaptive streaming

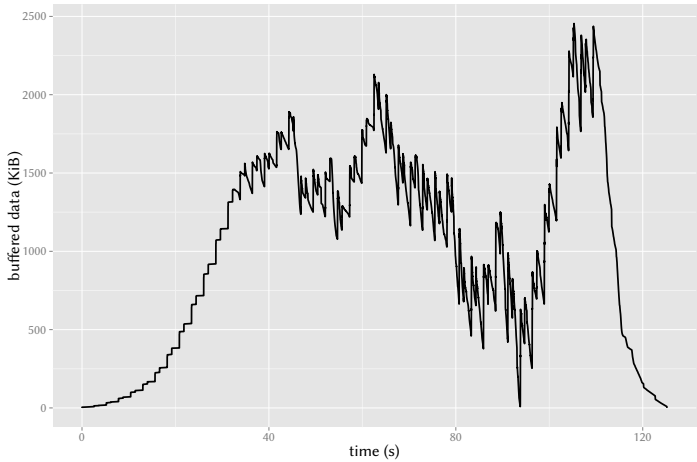
Progressive Streaming Measurement Model



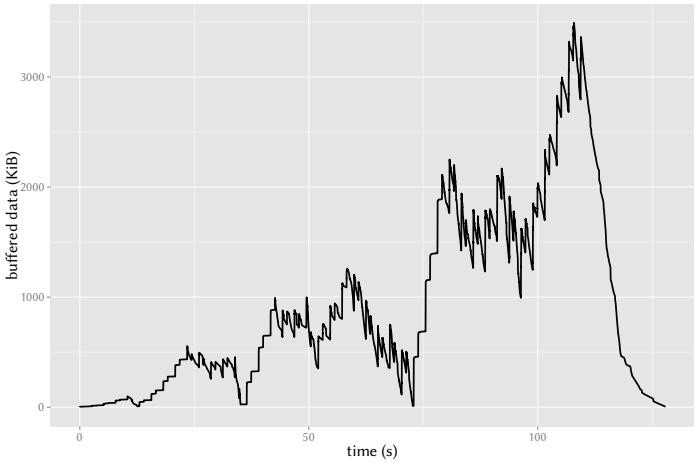
Non-adaptive Playback Strategies



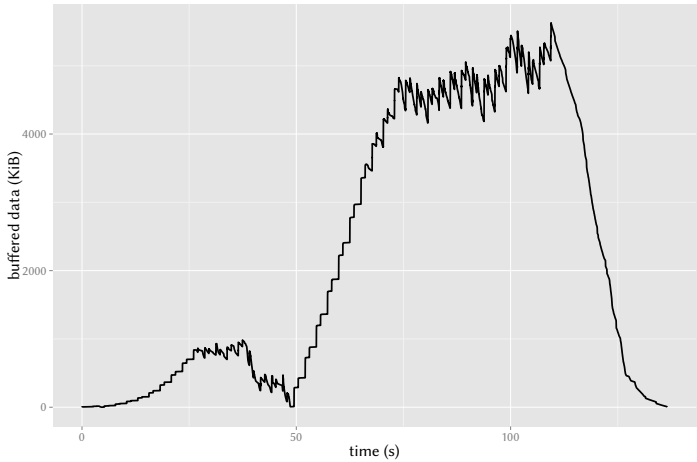
Null Strategy



Delayed Playback



YouTube Flash

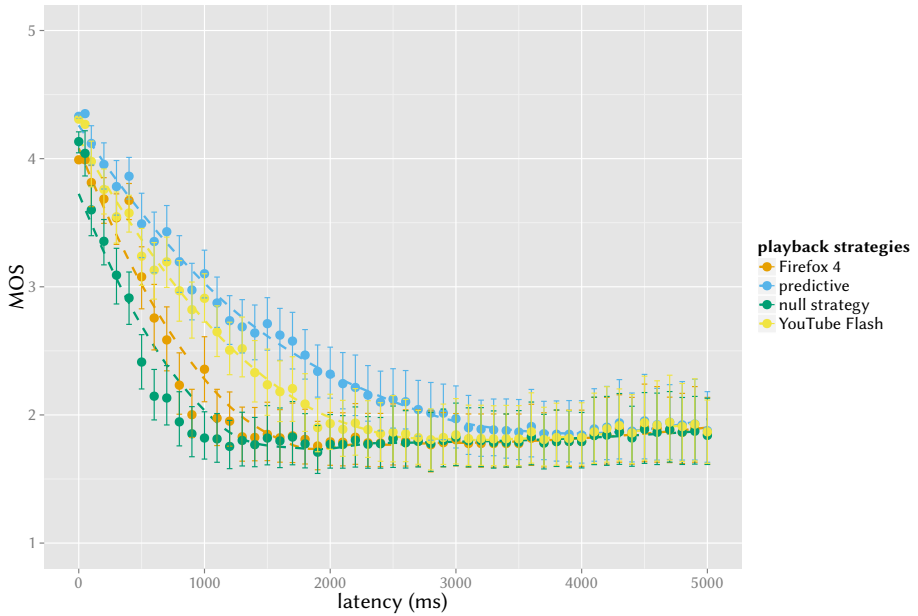


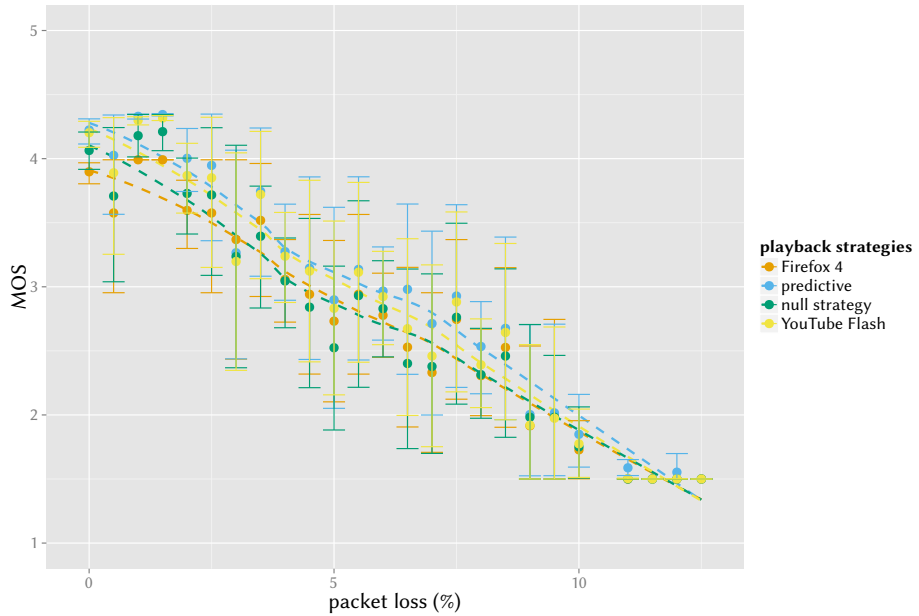
Firefox 4 HTML5

Testing the Model

YouTube Measurement Campaign

- Subjecting the playback strategies to YouTube transmission traces under different network QoS (latency, loss)
- Calculate MOS according to [Hoß+13]





- If network QoS is insufficient, no strategy can ensure QoE
- TCP interacts unpredictably with loss, strategies have little influence
- QoE optimization potential for increased latency

Section 5

Reliable Streaming in Mobile Networks

Reliable Streaming in Mobile Networks

Interactions and Influences

- Reliable streaming mostly designed for wired conditions
- Complex mobile signaling interacts with higher protocol layers
- Influences of the network stack on reliable streaming in particular?
- Possible sources on every layer, need to be understood before measuring
  - But exploitable through cross-layer information exchange

Reliable Streaming in Mobile Networks

Interactions and Influences

- Reliable streaming mostly designed for wired conditions
- Complex mobile signaling interacts with higher protocol layers
- Influences of the network stack on reliable streaming in particular?
- Possible sources on every layer, need to be understood before measuring
  - But exploitable through cross-layer information exchange
- Feedback loops of many layers operate on the same time scale as (adaptive) streaming decision processes
  - Interact with and influence streaming process
  - E.g. GTP tunnel duration and signaling
- Layer/protocols constantly changing and evolving
  - E.g. TCP: complex interactions and state machines, different on every OS and version

Evaluating Reliable Streaming in Mobile Networks

- Intent to understand these influences and interactions
- Passive core network traces are a rare opportunity
  - Numerous privacy, business interest, NDA issues
  - Focus limited to network-wide effects, not specific users

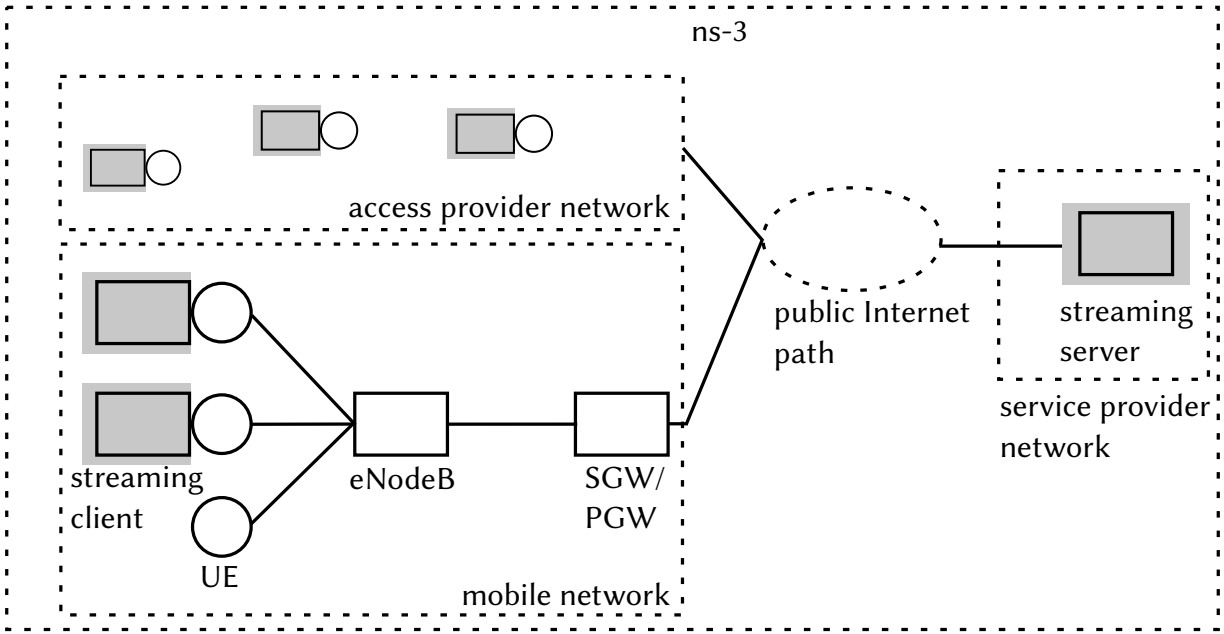
Evaluating Reliable Streaming in Mobile Networks

- Intent to understand these influences and interactions
- Passive core network traces are a rare opportunity
  - Numerous privacy, business interest, NDA issues
  - Focus limited to network-wide effects, not specific users
- Provide two more angles to conduct measurements
  - Enhancing active measurements with mobile-specific metadata
    - Privacy-aware sensor collection app for large scale mobile measurement campaigns (Sensorium, not shown here)
  - Simulating streaming in mobile networks using the introduced streaming measurement model
    - Selected the ns-3 LTE model as a basis after a survey of various mobile network simulators
    - Limitations due to the absence and accuracy of most control plane procedures

Mobile Streaming Simulation

with ns-3

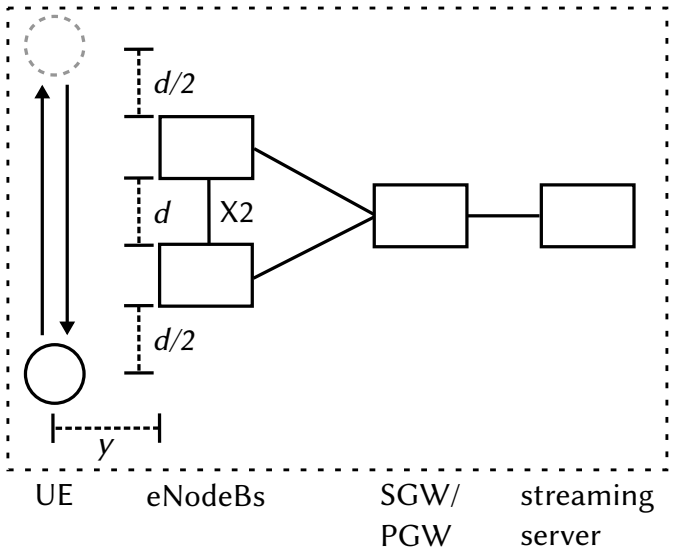
- Port the reliable streaming model to ns-3
- ns-3 does not provide HTTP
  - Directly use TCP and segmented transmissions to stream
- Implement example playback strategies
- Create a reliable streaming in an LTE network scenario
- Use it as a blueprint for future setups



Simulation Evaluation

Mobility Scenario

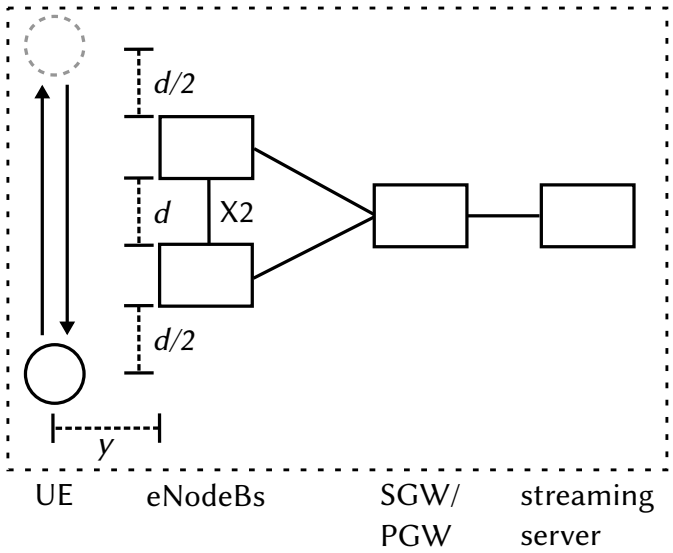
- Simple dual eNodeB scenario
- Moving back and forth between them



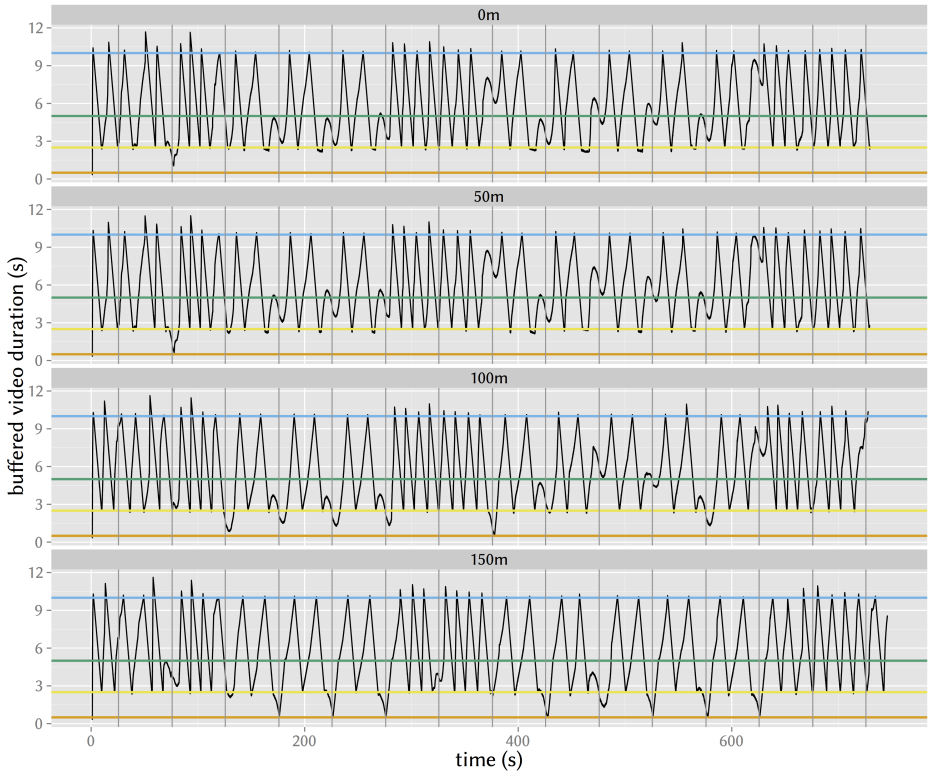
Simulation Evaluation

Mobility Scenario

- Simple dual eNodeB scenario
- Moving back and forth between them



- Stall events at coverage edges and due to handovers



Section 6

Conclusions

Summary and Contributions

- Investigate interworkings and influences of reliable streaming and mobile networks
- Provide angles to wholly evaluate streaming and in particular mobile reliable streaming
  - Two phase streaming measurement model
  - Mobile active measurement and simulation approaches
  - Impact of mobility on streaming
- Streaming design guidelines for service providers can be derived

- Better understanding of mobile control plane for future architectures
  - Operation and impact of GTP tunnels and signaling
  - Control plane load depends on traffic patterns not volume!
- Tools for better network dimensioning, planning, and traffic management for operators
  - Load Model, Queuing Simulation, and Virtual GGSN
- Models published in several papers, open sourced tools and frameworks

Future Work

- Mobile Control Plane
  - Compare to load in LTE/EPC networks
  - Look at other potential load bottlenecks, e.g. MME/HLR or PCEF/PCRF
  - Investigate specific load influence sources, esp. more user traffic aspects
  - Improve future 3GPP specs for reduced control plane complexity and better scaling
- Reliable Streaming
  - Investigate influences of different transport and application layer protocols, e.g. QUIC, TLS, HTTP/2, TCP variants
  - QoE optimization of playback strategies for current context factors
  - Streaming centric mobile measurements and simulations with focus on the interaction of adaptivity with mobility and other context factors

Thanks!

Questions!

Publications I

A. Biernacki, **F. Metzger**, and K. Tutschku. “On the Influence of Network Impairments on YouTube Video Streaming”.
In: *Journal of Telecommunications and Information Technology (JTIT)* 2012.03 (2012).

F. Metzger, A. Rafetseder, D. Stezenbach, and K. Tutschku. “Analysis of web-based video delivery”.
In: *FITCE Congress (FITCE), 2011 50th*. 2011, pp. 1–6.

F. Metzger, A. Rafetseder, P. Romirer, S. Gebert, K. Salzlechner, and K. Tutschku. “Research Report On Signaling Load and Tunnel Management in a 3G Core Network”.
In: *University of Würzburg Institute of Computer Science Research Report Series* 484 (2012).

Publications II

F. Metzger, A. Rafetseder, P. Romirer-Maierhofer, and K. Tutschku. “Exploratory Analysis of a GGSN’s PDP Context Signaling Load”. In: *Journal of Computer Networks and Communications* (Feb. 2014).

F. Metzger, A. Rafetseder, and K. Tutschku. “A performance evaluation framework for video streaming”. In: *Packet Video Workshop (PV), 2012 19th International*. 2012, pp. 19–24.

F. Metzger, C. Schwartz, and T. Hoßfeld. “GTP-based Load Model and Virtualization Gain for a Mobile Network’s GGSN”. In: *Communications and Electronics (ICCE), 2014 IEEE Fifth International Conference on*. July 2014, pp. 206–211.

Publications III

F. Metzger, C. Steindl, and T. Hoßfeld.
"A Simulation Framework for Evaluating the QoS and QoE of TCP-based Streaming in an LTE Network". In: *27th International Teletraffic Congress (ITC 27)* (2015). submitted, under review.

A. Rafetseder, **F. Metzger**, D. Stezenbach, and K. Tutschku.
"Exploring YouTube's Content Distribution Network Through Distributed Application-layer Measurements: A First View".
In: *Cnet 2011 : International Workshop on MODELING, ANALYSIS, AND CONTROL OF COMPLEX NETWORKS*. Cnet '11. San Francisco, California: International Teletraffic Congress, 2011, pp. 31–36. ISBN: 978-0-9836283-1-6.

Publications IV

A. Rafetseder, **F. Metzger**, D. Stezenbach, and K. Tutschku. “Network Federation as a Provider Concept: From Today’s Measurement to Tomorrow’s Architecture”.
In: *12th Würzburg Workshop on IP: ITG Workshop ”Visions of Future Generation Networks” (EuroView2012)*. July 2012.

A. Rafetseder, **F. Metzger**, and L. Pühringer. “Sensorium – A Generic Sensor Framework”.
In: *PIK - Praxis der Informationsverarbeitung und Kommunikation* 36.1 (Feb. 2013), p. 46.

Source Code

Thesis, this presentation, and associated scripts

https://github.com/fmetzger/thesis

Video Streaming Emulation & Evaluation

https://github.com/fmetzger/videostreaming-bufferemulation

GGSN Queuing Simulation

https://github.com/fmetzger/ggsn-simulation

Sensorium

https://github.com/fmetzger/android-sensorium

Mobile Streaming Simulation

https://github.com/Steindi01/lteNS3

References I

A complete list of references can be found in the written thesis.

3GPP. *Radio Resource Control (RRC); Protocol specification.* TS 25.331. 3rd Generation Partnership Project (3GPP), Sept. 2012.

3GPP. *General Packet Radio Service (GPRS); Service description; Stage 2.* TS 23.060. 3rd Generation Partnership Project (3GPP), Sept. 2013.

Y. Cheng, J. Chu, S. Radhakrishnan, and A. Jain. *TCP Fast Open.* Internet-Draft. Internet Engineering Task Force, Mar. 2014.

Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019.* Feb. 2015.

References II

T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau. "Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience". English.
In: *Data Traffic Monitoring and Analysis*.
Ed. by E. Biersack, C. Callegari, and M. Matijasevic. Vol. 7754.
Lecture Notes in Computer Science.
Springer Berlin Heidelberg, 2013, pp. 264–301.
ISBN: 978-3-642-36783-0.

S. Ha and I. Rhee. "Taming the elephants: New TCP slow start".
In: *Computer Networks* 55.9 (2011), pp. 2092–2110.
ISSN: 1389-1286.

References III

S. Ha, I. Rhee, and L. Xu.
"CUBIC: A New TCP-friendly High-speed TCP Variant".
In: *SIGOPS Oper. Syst. Rev.* 42.5 (July 2008), pp. 64–74.
ISSN: 0163-5980.

K. Ma, R. Bartos, S. Bhatia, and R. Nair.
"Mobile video delivery with HTTP".
In: *Communications Magazine, IEEE* 49.4 (2011), pp. 166–175.
ISSN: 0163-6804.

K. Nichols and V. Jacobson.
Controlled Delay Active Queue Management. Internet-Draft.
Internet Engineering Task Force, Mar. 2014.

L. Eggert and F. Gont. *TCP User Timeout Option*.
RFC 5482 (Proposed Standard).
Internet Engineering Task Force, Mar. 2009.

References IV

M. Allman, V. Paxson, and E. Blanton. *TCP Congestion Control.* RFC 5681 (Draft Standard).
Internet Engineering Task Force, Sept. 2009.

P. Sarolahti, M. Kojo, K. Yamamoto, and M. Hata. *Forward RTO-Recovery (F-RTO): An Algorithm for Detecting Spurious Retransmission Timeouts with TCP.* RFC 5682 (Proposed Standard).
Internet Engineering Task Force, Sept. 2009.

M. Allman, K. Avrachenkov, U. Ayesta, J. Blanton, and P. Hurtig. *Early Retransmit for TCP and Stream Control Transmission Protocol (SCTP).* RFC 5827 (Experimental).
Internet Engineering Task Force, May 2010.

References V

V. Paxson, M. Allman, J. Chu, and M. Sargent.
Computing TCP's Retransmission Timer.
RFC 6298 (Proposed Standard).
Internet Engineering Task Force, June 2011.

A. Ford, C. Raiciu, M. Handley, and O. Bonaventure.
TCP Extensions for Multipath Operation with Multiple Addresses.
RFC 6824 (Experimental).
Internet Engineering Task Force, Jan. 2013.

D. Farinacci, V. Fuller, D. Meyer, and D. Lewis.
The Locator/ID Separation Protocol (LISP).
RFC 6830 (Experimental).
Internet Engineering Task Force, Jan. 2013.

References VI

M. Mathis, N. Dukkipati, and Y. Cheng.
Proportional Rate Reduction for TCP. RFC 6937 (Experimental).
Internet Engineering Task Force, May 2013.

J. Chu, N. Dukkipati, Y. Cheng, and M. Mathis.
Increasing TCP's Initial Window. RFC 6928 (Experimental).
Internet Engineering Task Force, Apr. 2013.

Sandvine. *Sandvine Global Internet Phenomena Reports*.
2011–2014. URL:
`https://www.sandvine.com/trends/global-internet-phenomena/`.

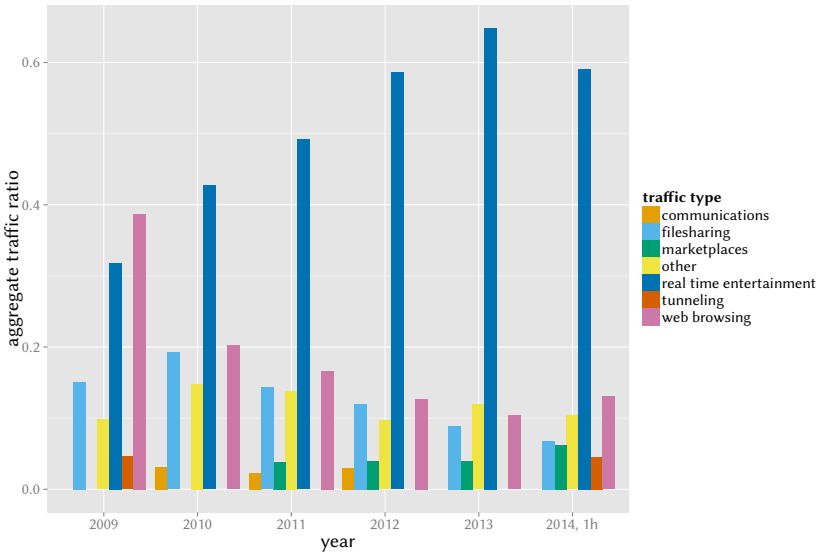


Figure: Traffic composition of North American peak fixed access aggregate traffic (data source: [San14]).

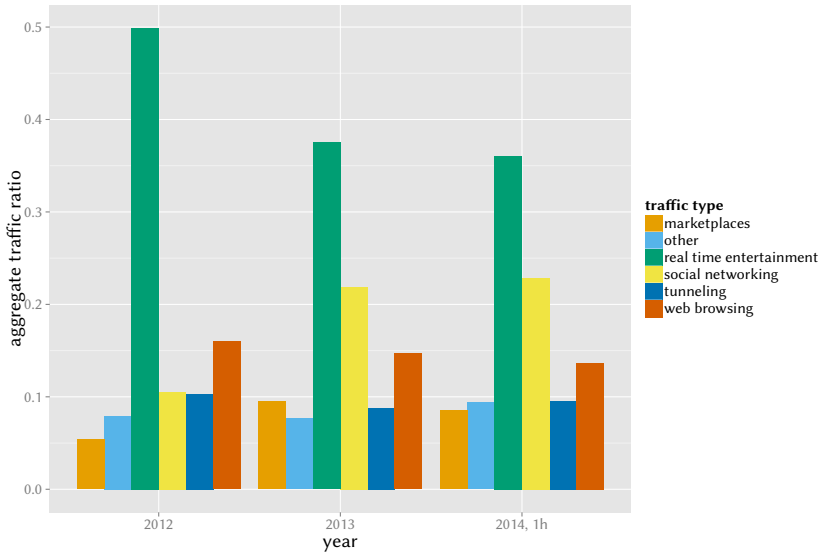


Figure: Traffic composition of North American peak mobile access aggregate traffic (data source: [San14]).

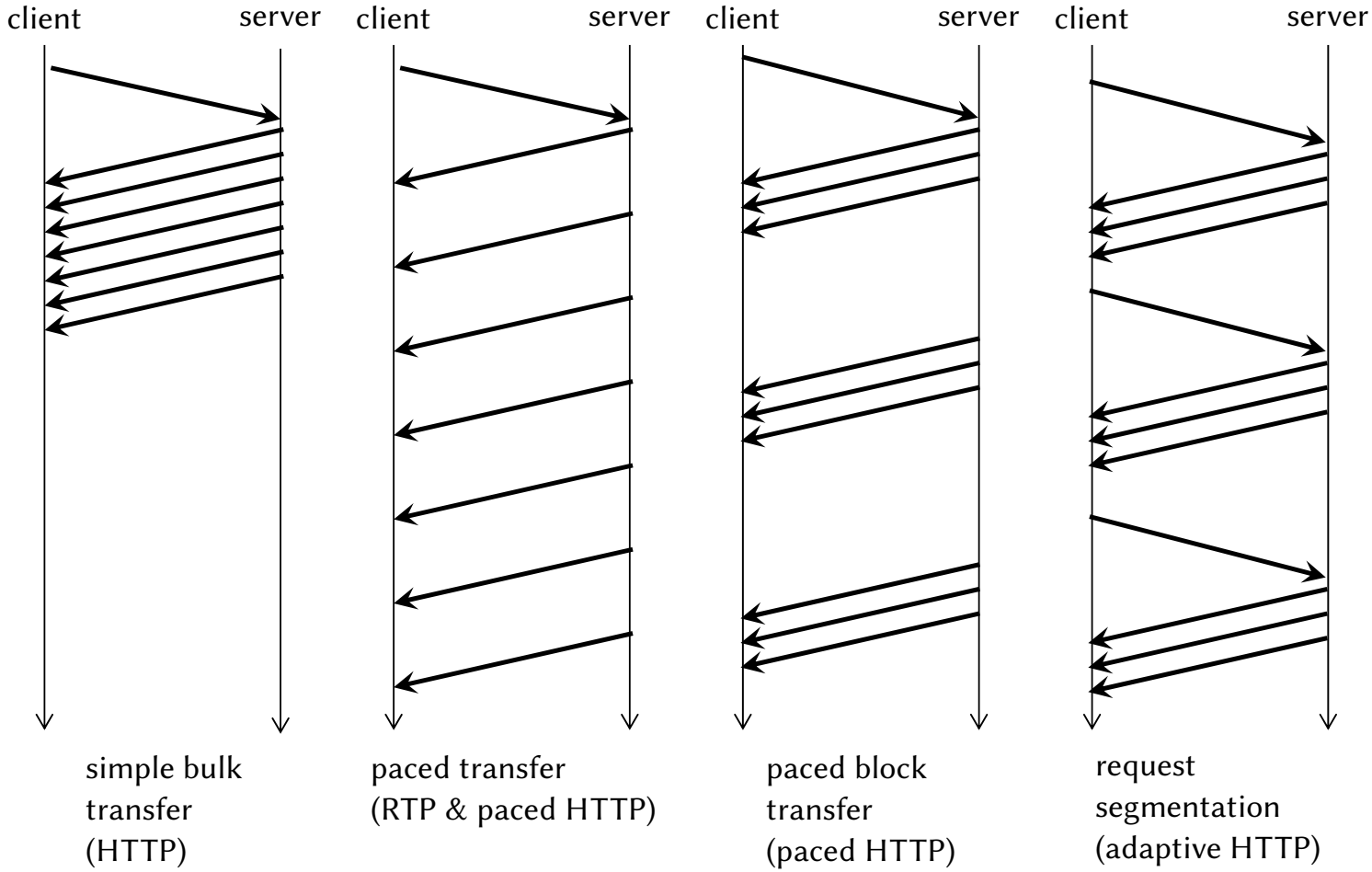


Figure: Comparison of several possible streaming transmission modes depicting the timing of the sent packets (source: [Ma+11]).

Protocol Classification Matrix

Protocol	**Vertical Location of Control**	**Horizontal Location of Control**	**Reliable Transport**	**Video Type**	**Adaptivity**	**Multicast**
RTP	out-of-band, application layer protocol	server-side and limited intermediary (translators and mixers)	unreliable (UDP)	low delay live streaming	server-side adaptation (transcoding)	using IGMP
simple HTTP	in-band, streaming application	client-side	reliable (TCP)	stored, not live	none	emulated through CDN
adaptive HTTP (e.g. DASH)	in-band, streaming application	client-side	reliable (TCP)	stored and near-live	client-side with file segmentation	emulated through CDN

Further Playback Strategies

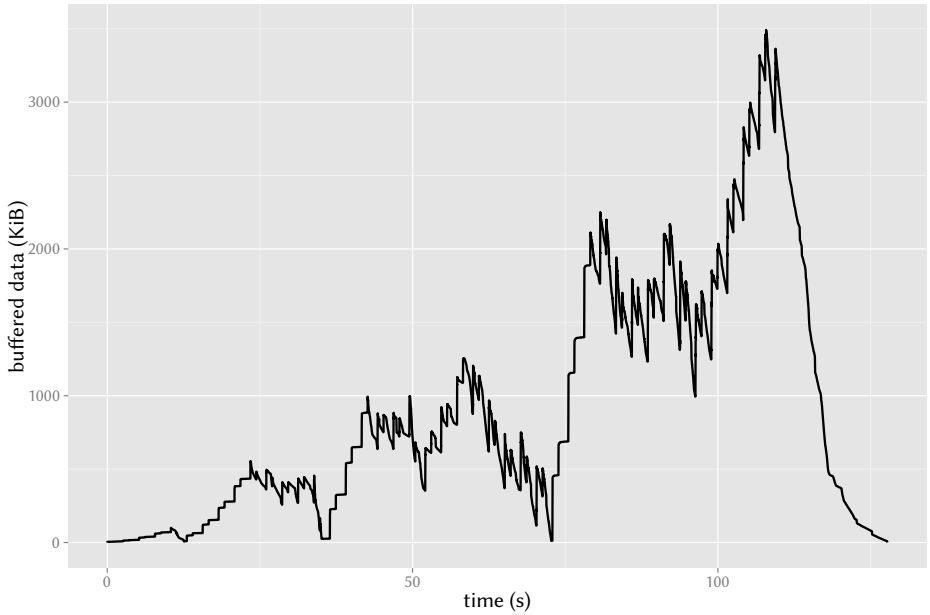


Figure: Sample buffer fill level for a 5 s buffered video duration threshold strategy with an additional 2 s initial threshold; 34 s total stalling.

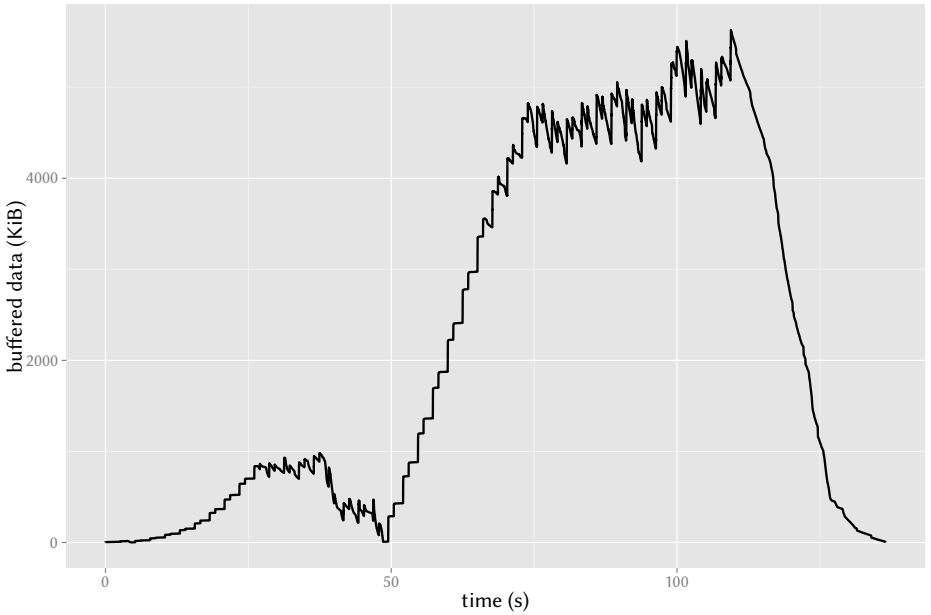


Figure: Sample buffer fill level for the Firefox 4 strategy, 44 s total stalling.

Firefox 4 Playback Strategy

if $s_{MA} > v_{MA}$ **then**
  $c \leftarrow (b_b = 20s \vee b_T = 20s)$
else
  $c \leftarrow (b_b = 30s \vee b_T = 30s)$
end if

To estimate the current and future rates, the moving average of the transmission rate s_{MA} and the video bitrate v_{MA} are calculated. The condition c Firefox uses to start and resume the playback process is given in the algorithm, with the buffered video duration b_b, and the duration spent buffering b_T.

Latency Measurement Series

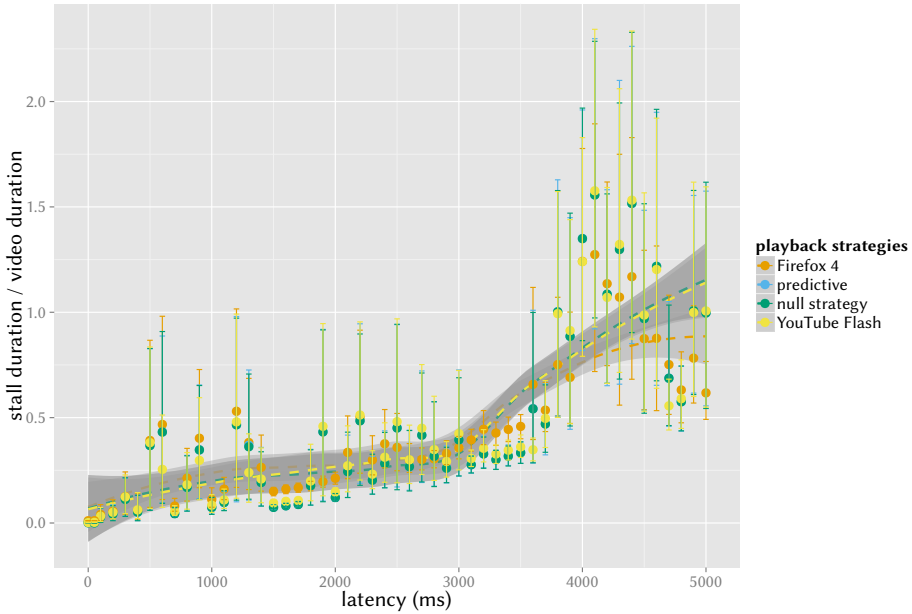


Figure: Stalling duration in relation to transmission latency with a local polynomial least-squares fit.

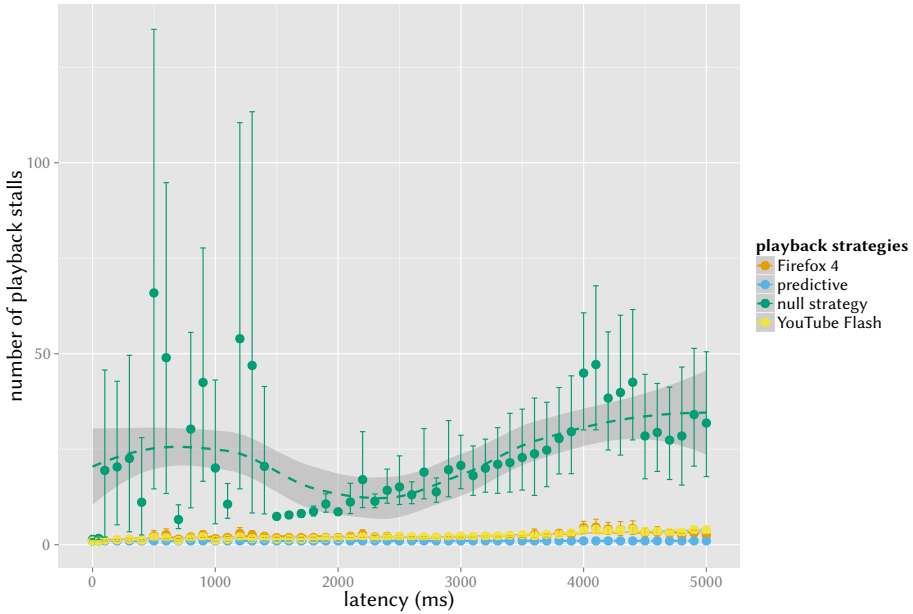


Figure: Number of stalls in relation to transmission latency with a local polynomial least-squares fit.

Loss Measurement Series

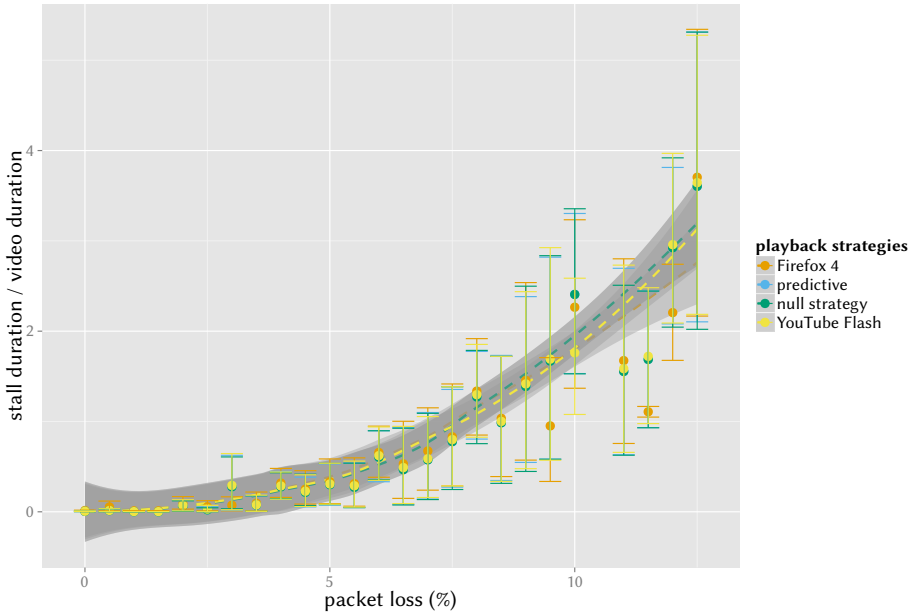


Figure: Stalling duration in relation to the packet loss with a local polynomial least-squares fit.

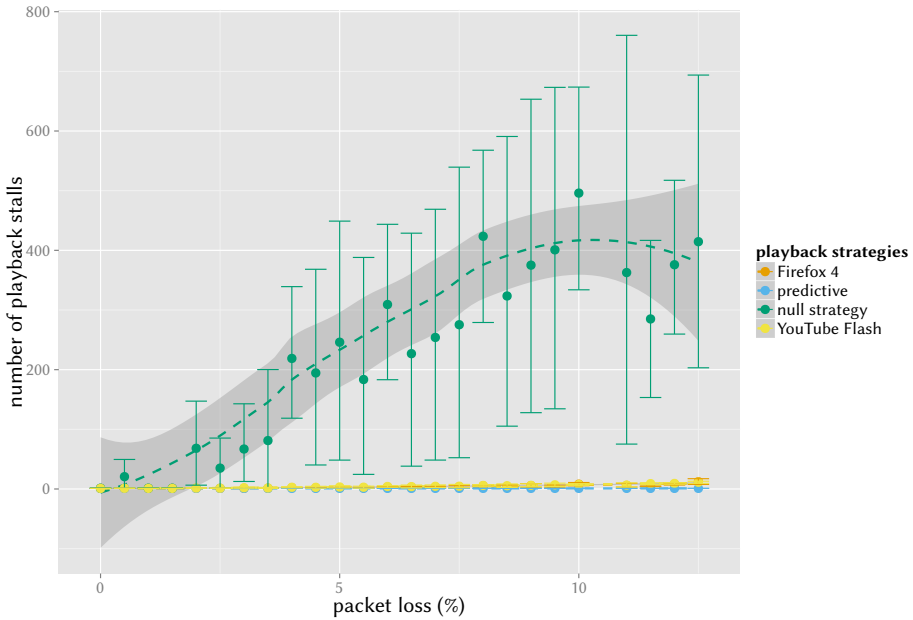


Figure: Number of playback stalls in relation to packet loss with local polynomial least-squares fit.

Loss Measurement Series QoE

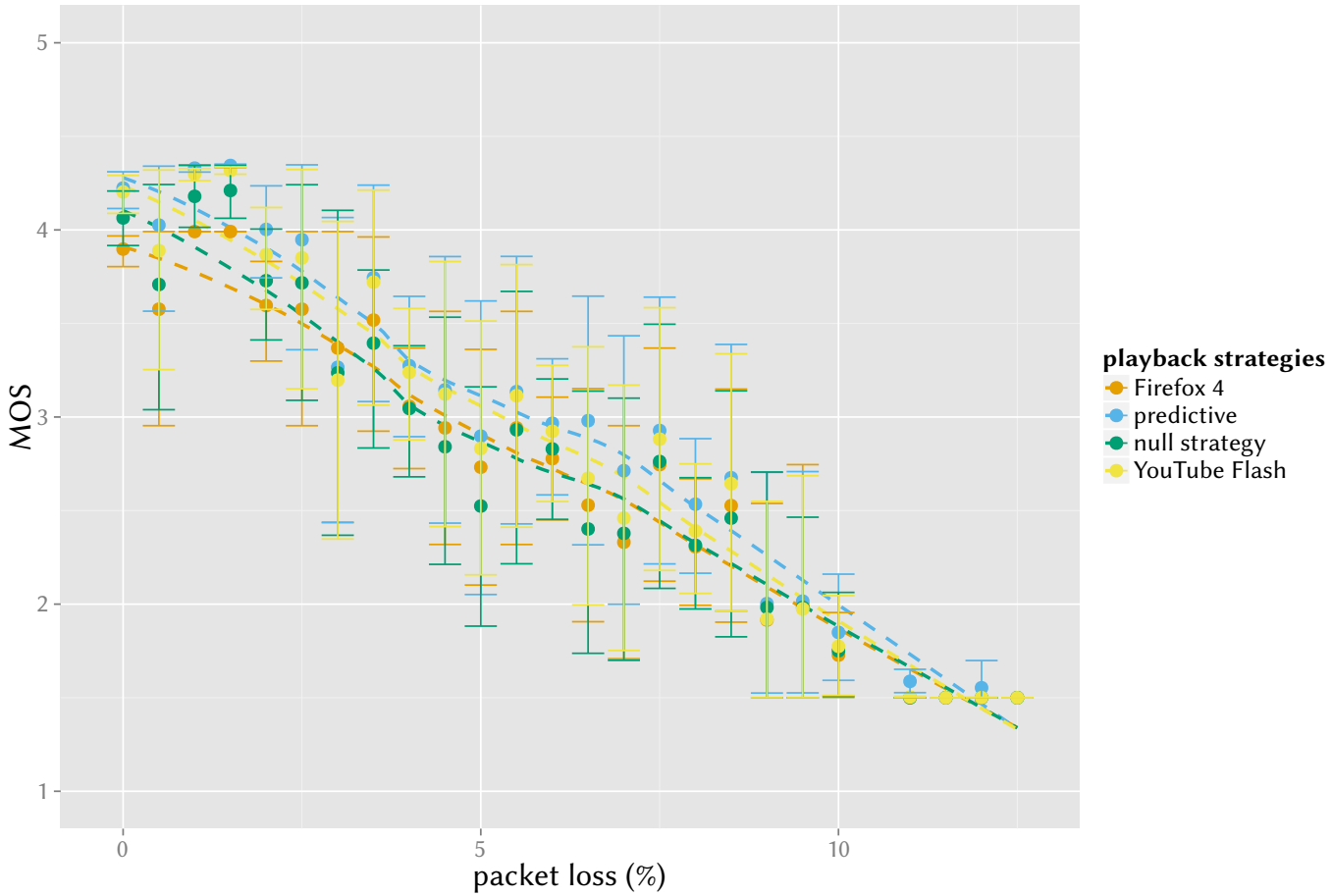


Figure: Calculated MOS for the loss measurement series.

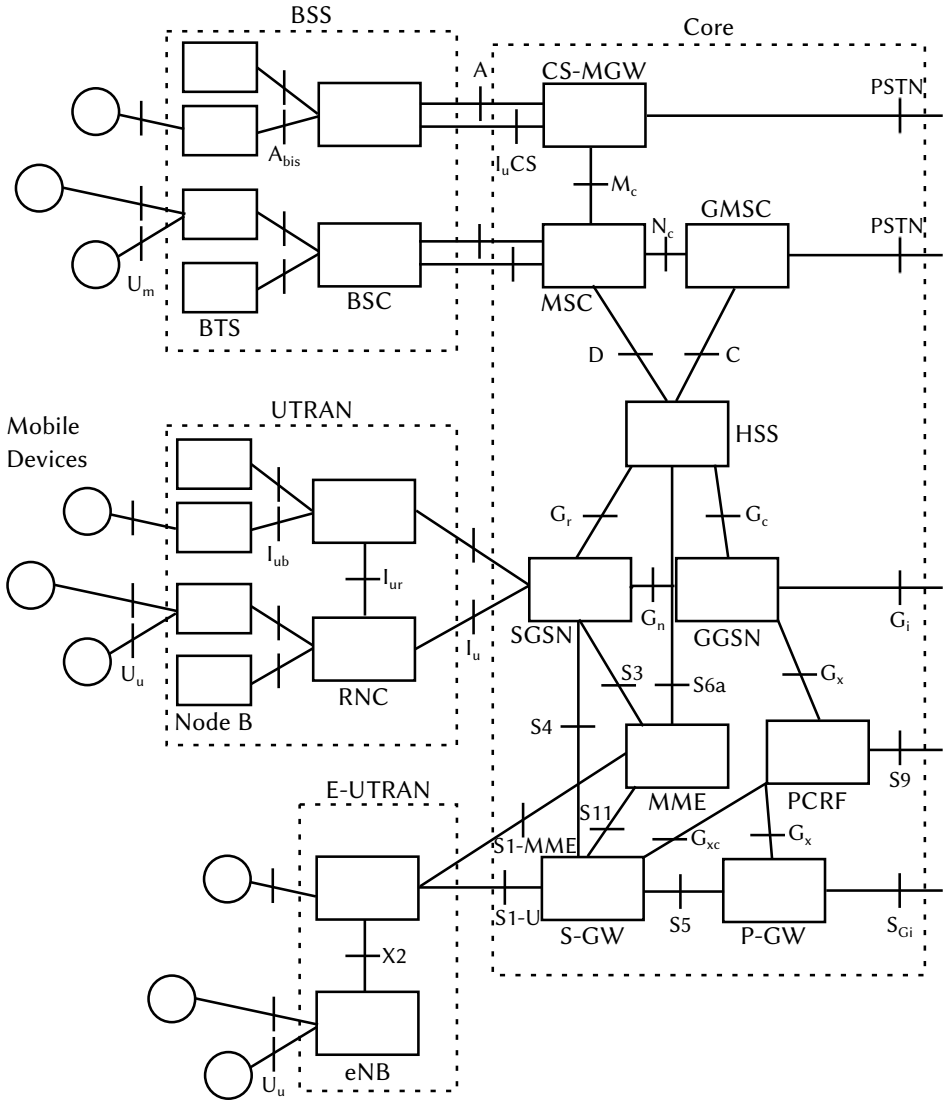


Figure: Overview of a combined CS/PS GSM/UMTS/LTE architecture.

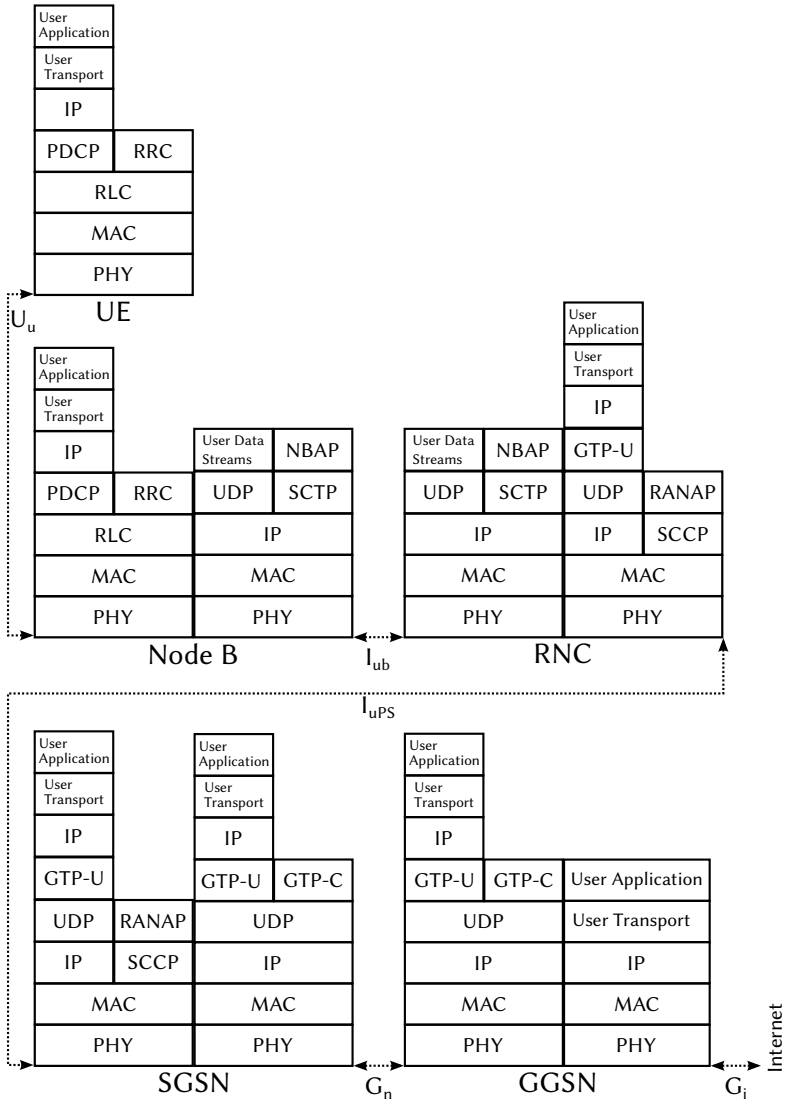


Figure: Simplified control plane and user plane IP-based protocol stacks on the user traffic path through the mobile network.

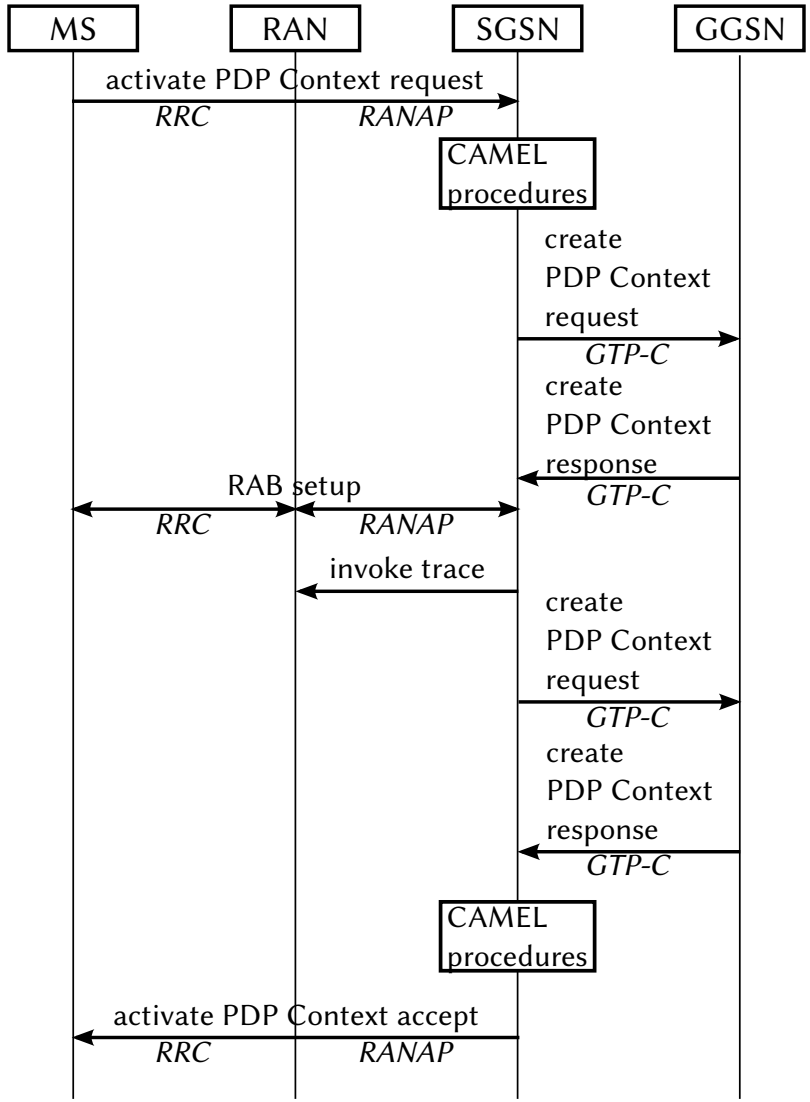


Figure: PDP Context activation procedure signaling interaction diagram for UMTS, including involved signaling protocols.

	Bits							
Octets	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
1	Version			1	0	E	S	PN
2	Message Type							
3	Length							
4								
5	Tunnel Endpoint Identifier							
6								
7								
8								
9	Sequence Number							
10								
11	N-PDU							
12	Next Extension Header Type							

Figure: General 12 *B* GTP header format.

Table: All IE in a Create PDP Context request and size thereof for IPv4 network and user traffic only. The denoted sizes exclude the first message type byte.

IE	Presence	Size
IMSI	cond.	8 *B*
RAI	opt.	6 *B*
Recovery	opt.	1 *B*
Selection mode	cond.	1 *B*
TEID Data I	mand.	4 *B*
TEID Control Plane	cond.	4 *B*
NSAPI	mand.	1 *B*
Linked NSAPI	cond.	1 *B*
Charging Characteristics	cond.	2 *B*
Trace Reference	opt.	2 *B*
Trace Type	opt.	2 *B*
End User Address	cond.	8 *B*
APN	cond.	max 102 *B*
PCO	opt.	max 255 *B*
SGSN signaling address	mand.	6 *B*
SGSN user traffic address	mand.	6 *B*
MSISDN	cond.	max 17 *B*
QoS Profile	mand.	max 257 *B*
TFT	cond.	max 257 *B*
Trigger Id	opt.	var.
OMC Identity	opt.	var.
Common Flags	opt.	3 *B*
APN Restriction	opt.	3 *B*
RAT	opt.	3 *B*
User Location Information	opt.	10 *B*
MS Time Zone	opt.	4 *B*
IMEI (SV)	cond.	10 *B*
CAMEL Charging Information Container	opt.	var.
Additional Trace Info	opt.	11 *B*
Correlation-ID	opt.	3 *B*
Evolved Allocation Retention Priority I	opt.	3 *B*
Extended Common Flags	opt.	3 *B*
User CSG Information	opt.	10 *B*
APN-AMBR	opt.	11 *B*
Signaling Priority Indication	opt.	3 *B*
Private Extension	opt.	var.

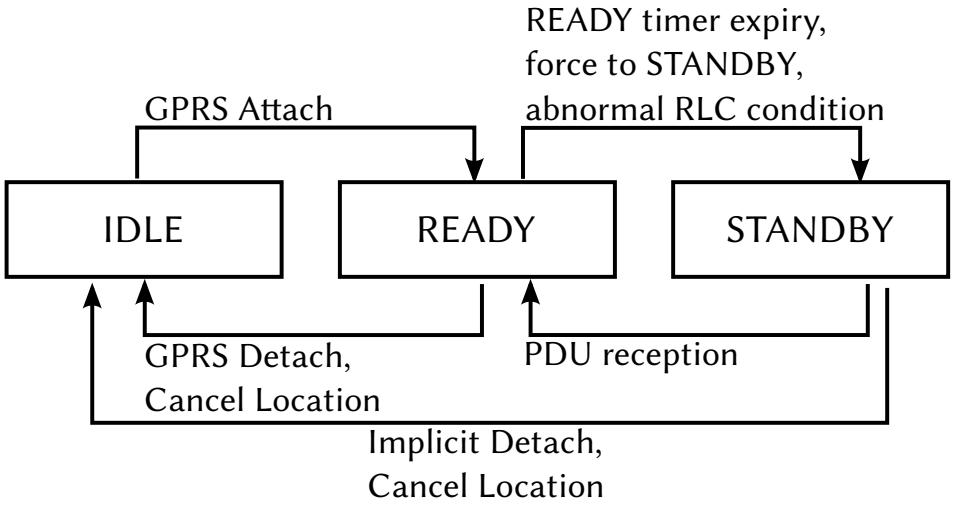


Figure: SGSN MM state machines for 2G radio access as defined in [3GP13, Section 6.1].

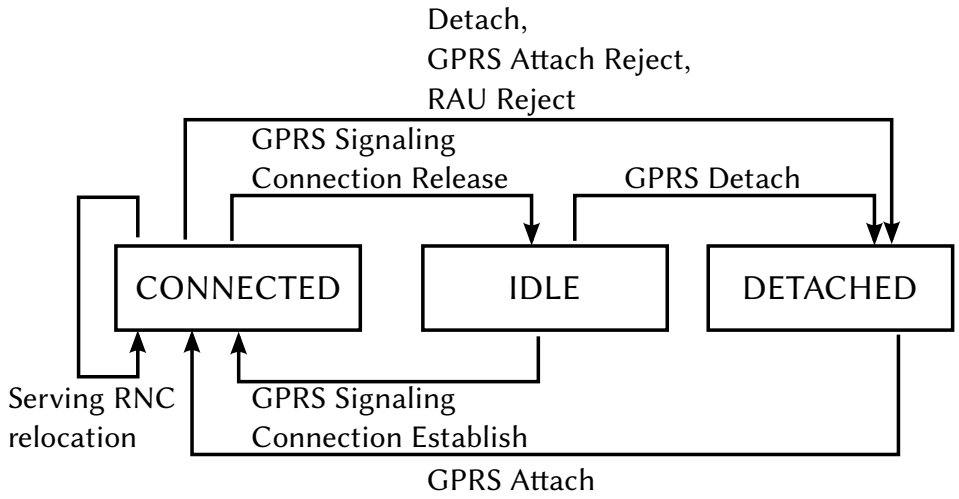


Figure: SGSN MM state machines for 3G radio access as defined in [3GP13, Section 6.1].

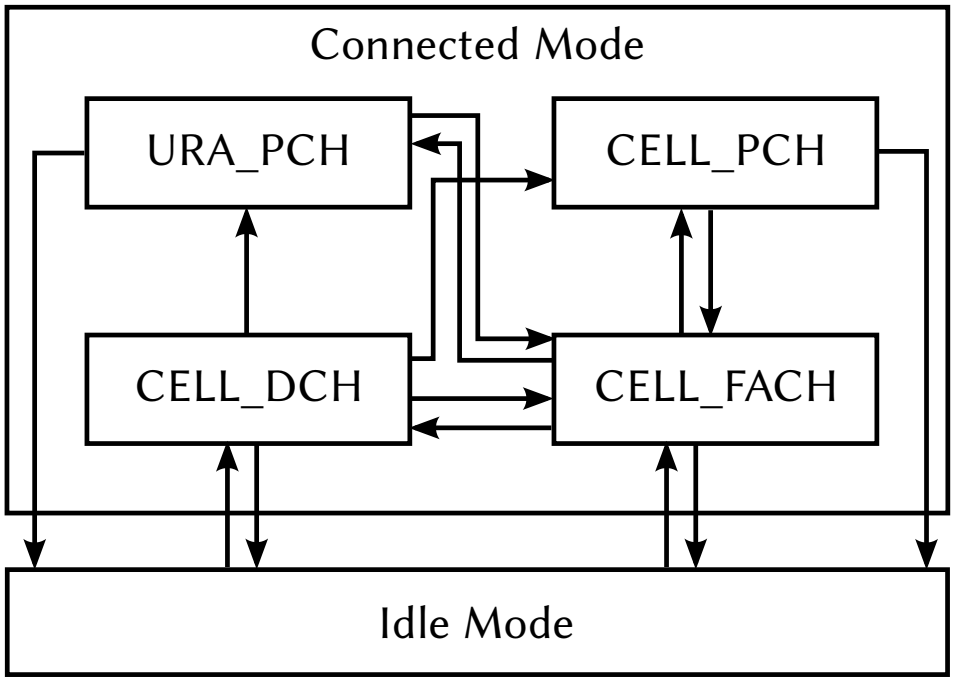


Figure: RRC State Model as per [3GP12, Section 7.1].

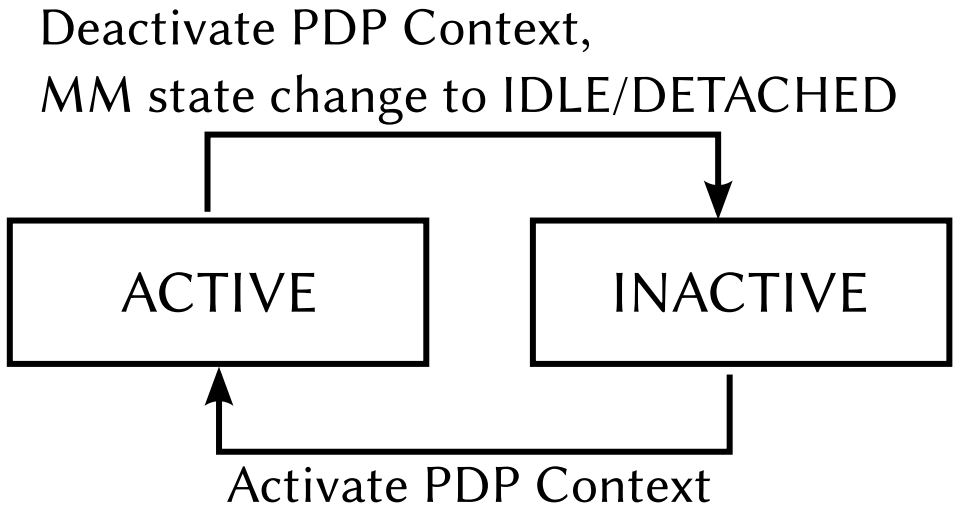


Figure: PDP State Model defined in [3GP13, Section 9].

Table: Relative TAC statistics.

Type	**Relative number of devices with an entry in the TAC dataset**
Total number of flows	99.72 %
Ratio of total traffic	99.97 %
Total number of tunnels	87.57 %
Total number of GTP signaling messages	90.95 %
Number of distinct UEs	80.93 %

Table: Relative device-discriminated traffic statistics extracted from the dataset.

	Flows	Traffic	Tunnels	GTP message pairs	Devices
By device type					
Smartphones	20.58%	12.81%	60.31%	75.99%	37.97%
Regular phones	0.26%	0.37%	5.40%	0.94%	9.25%
3G dongles	66.55%	75.12%	12.71%	9.53%	25.10%
By OS					
Android	10.82%	6.48%	14.33%	43.33%	14.01%
iOS	7.22%	4.47%	18.91%	20.35%	7.94%
Symbian	1.02%	1.09%	21.17%	4.51%	12.97%
Blackberry OS	0.07%	0.10%	2.17%	2.60%	1.48%

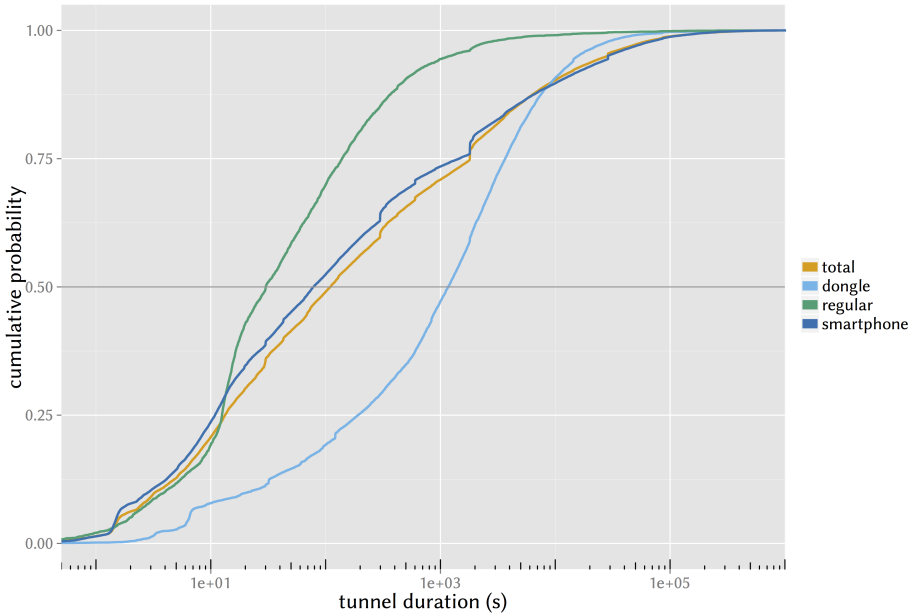


Figure: Tunnel duration distribution, separated for 3G dongles, smartphones and regular phones with medians at 115 s (total), 31 s (regular), 82 s (smartphone), and 1207 s (dongle).

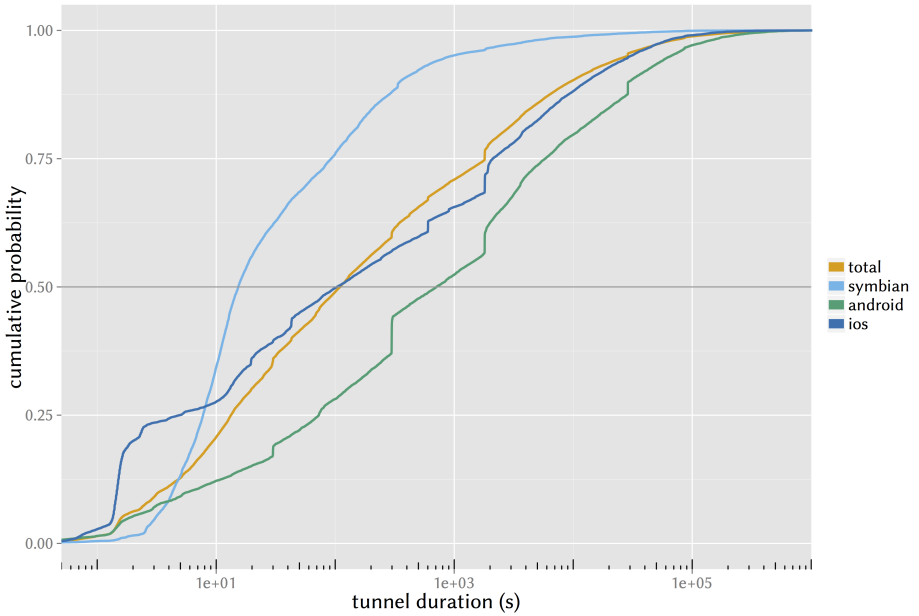


Figure: Tunnel duration CDF, separated for select OSs; Medians at 115 s (total), 15.5 s (symbian), 104 s (iOS), and 765 s (android).

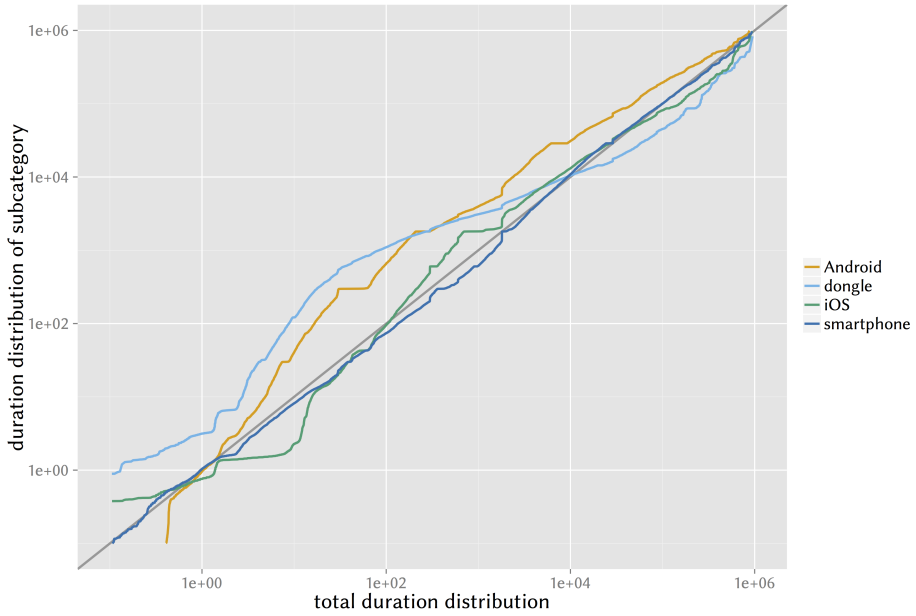


Figure: Q-Q Plots of the tunnel duration distributions in comparison to device classification categories.

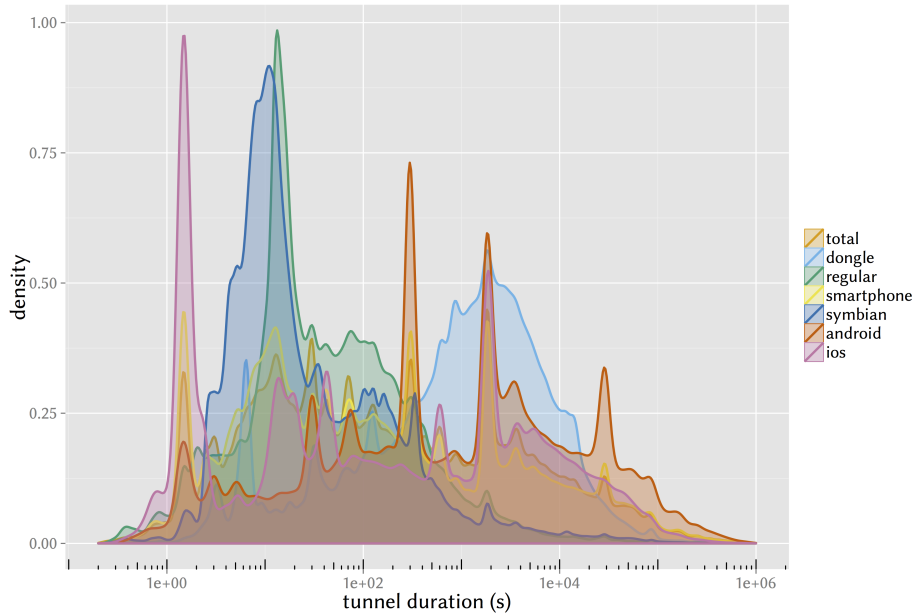


Figure: Logscale density plot of the tunnel duration with all classifications.

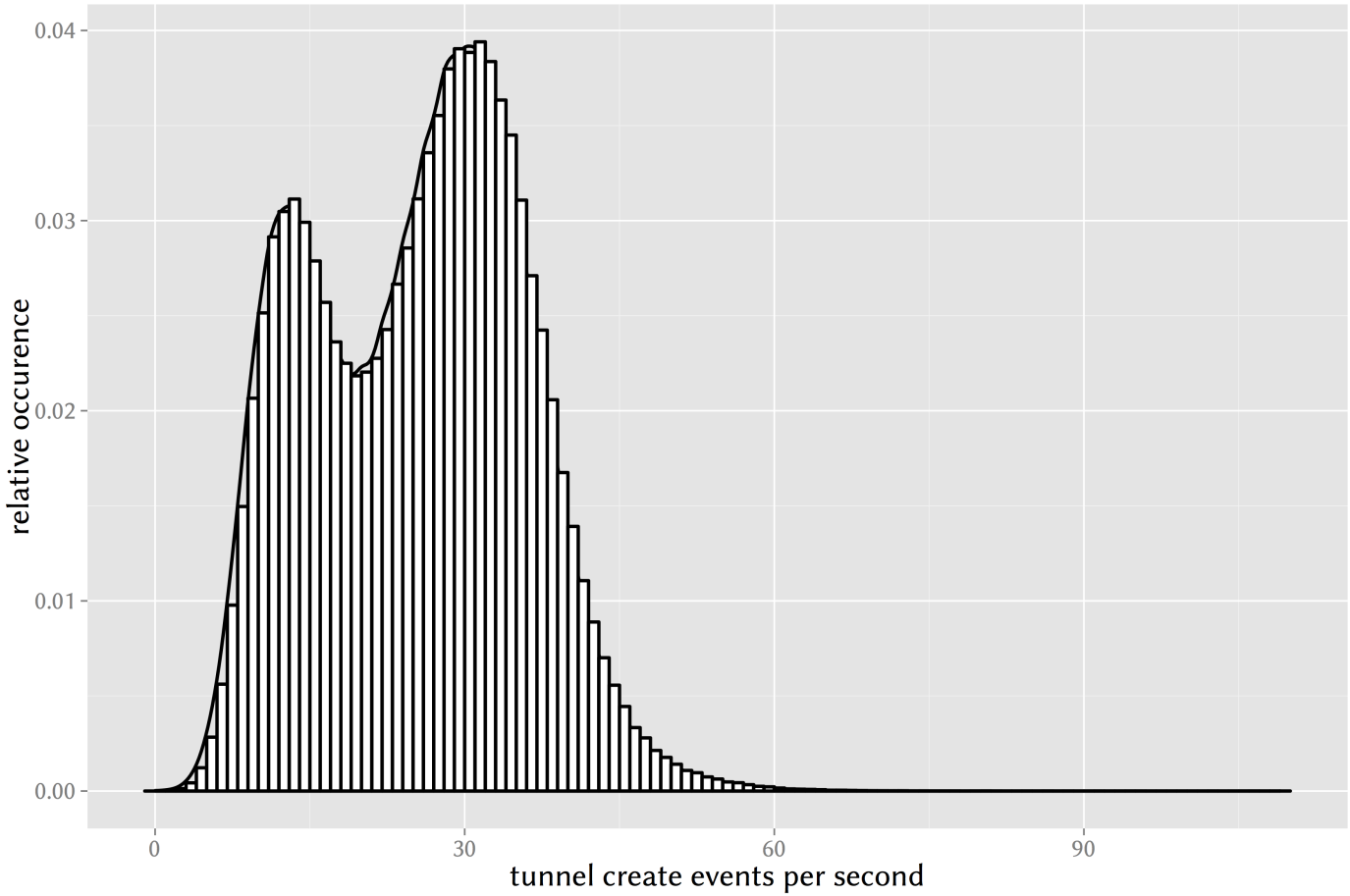


Figure: Density plot of tunnel arrival rate.

ECDF of the tunnel IAT in seconds by time of day

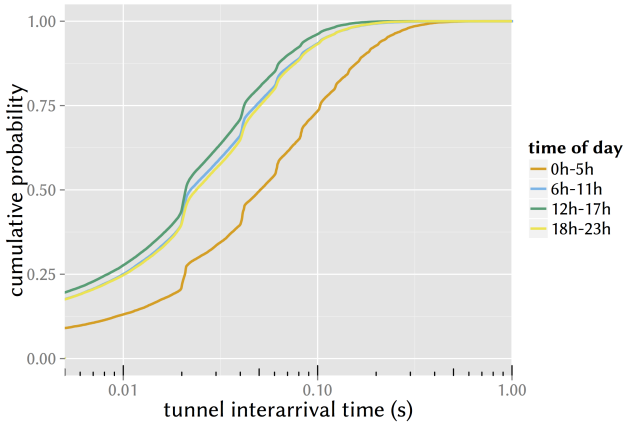


Figure: All tunnel requests.

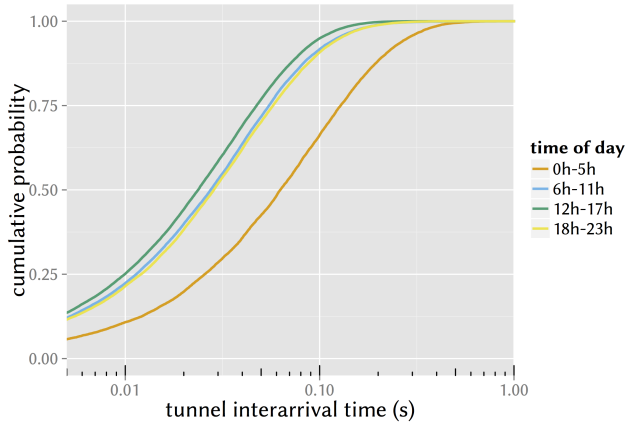


Figure: Only tunnels with data flows.

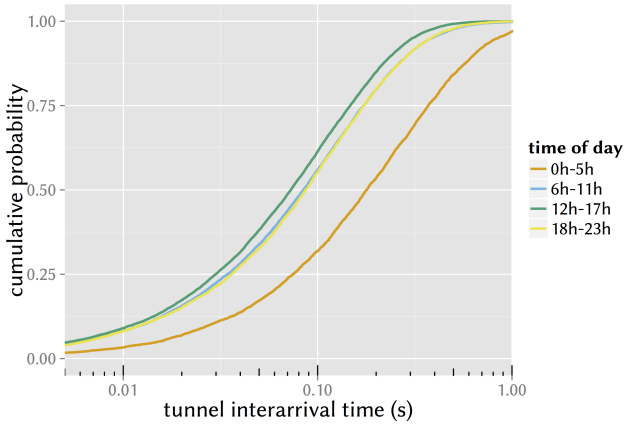


Figure: Tunnels with GPRS data flows.

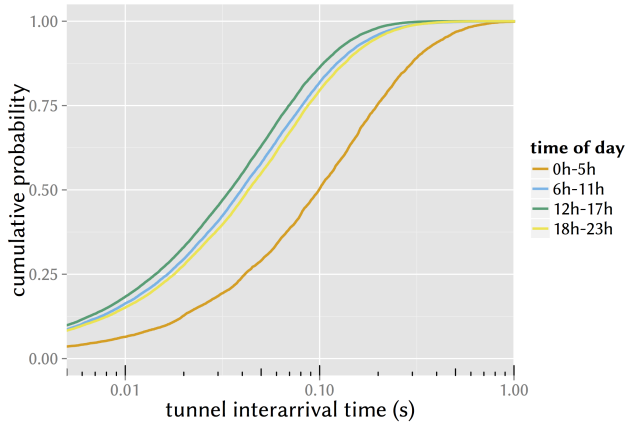


Figure: Tunnels with UMTS data flows.

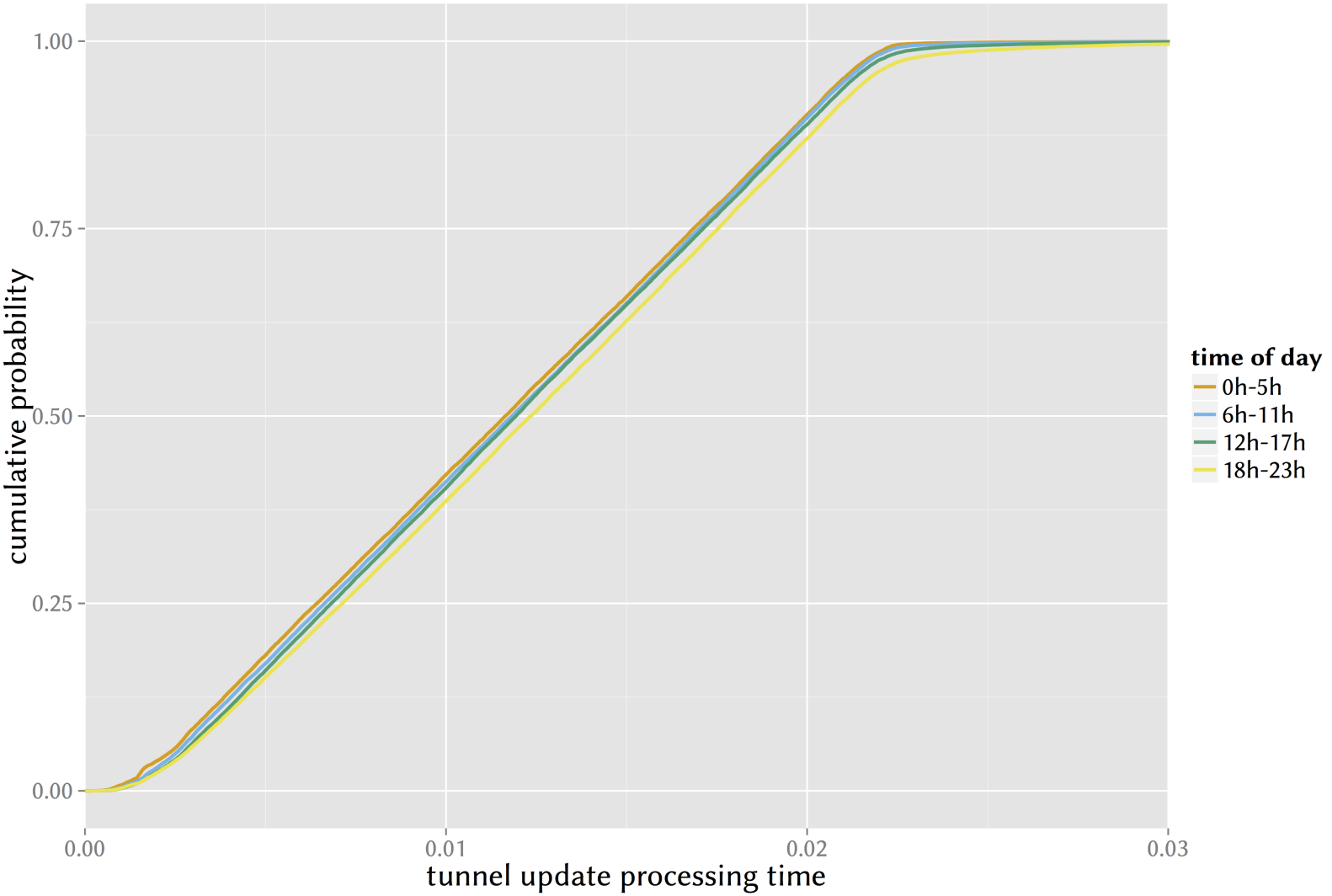


Figure: ECDFs of the time in seconds it takes a GGSN to process a GTP update event, separately plotted for four time slots each day.

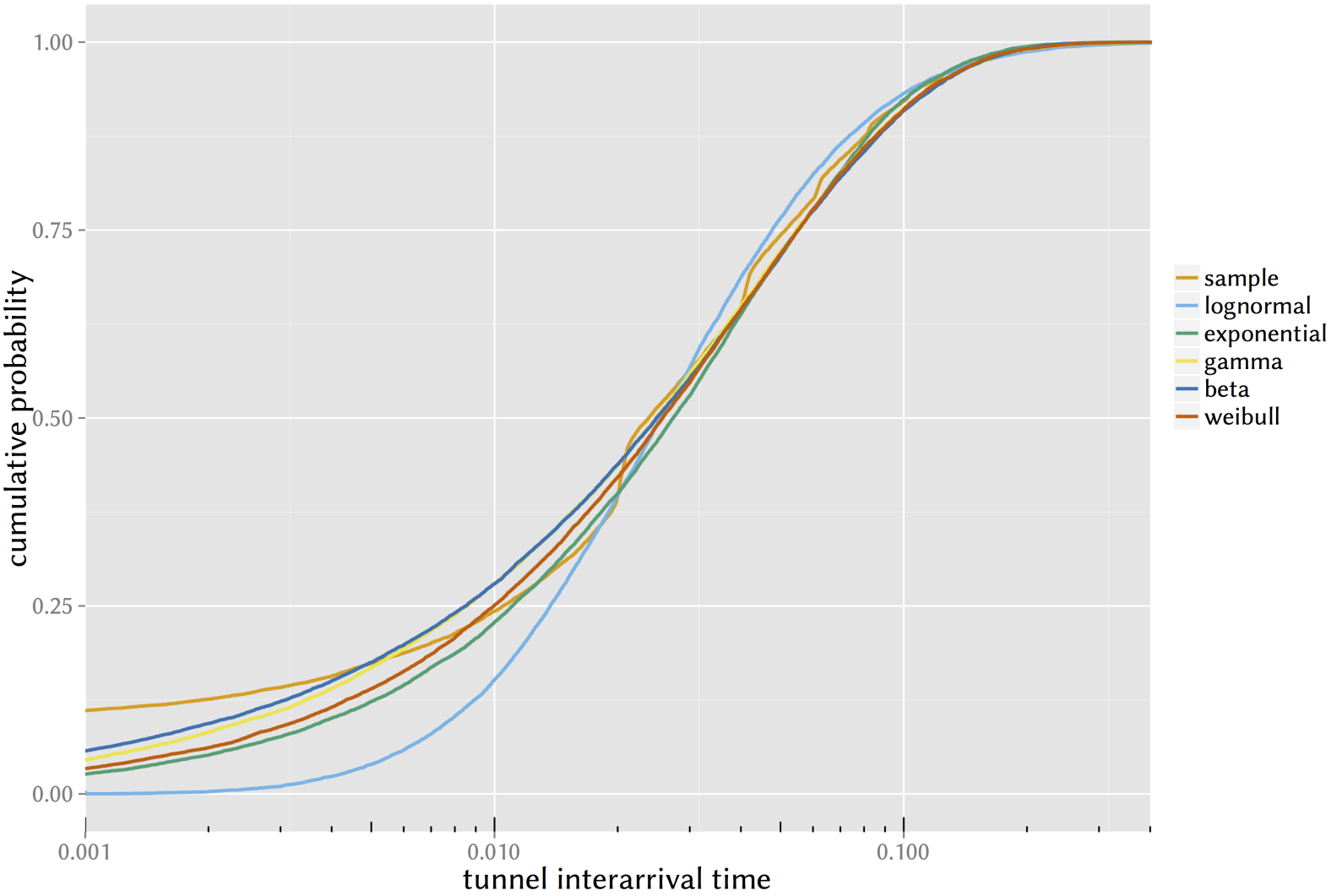
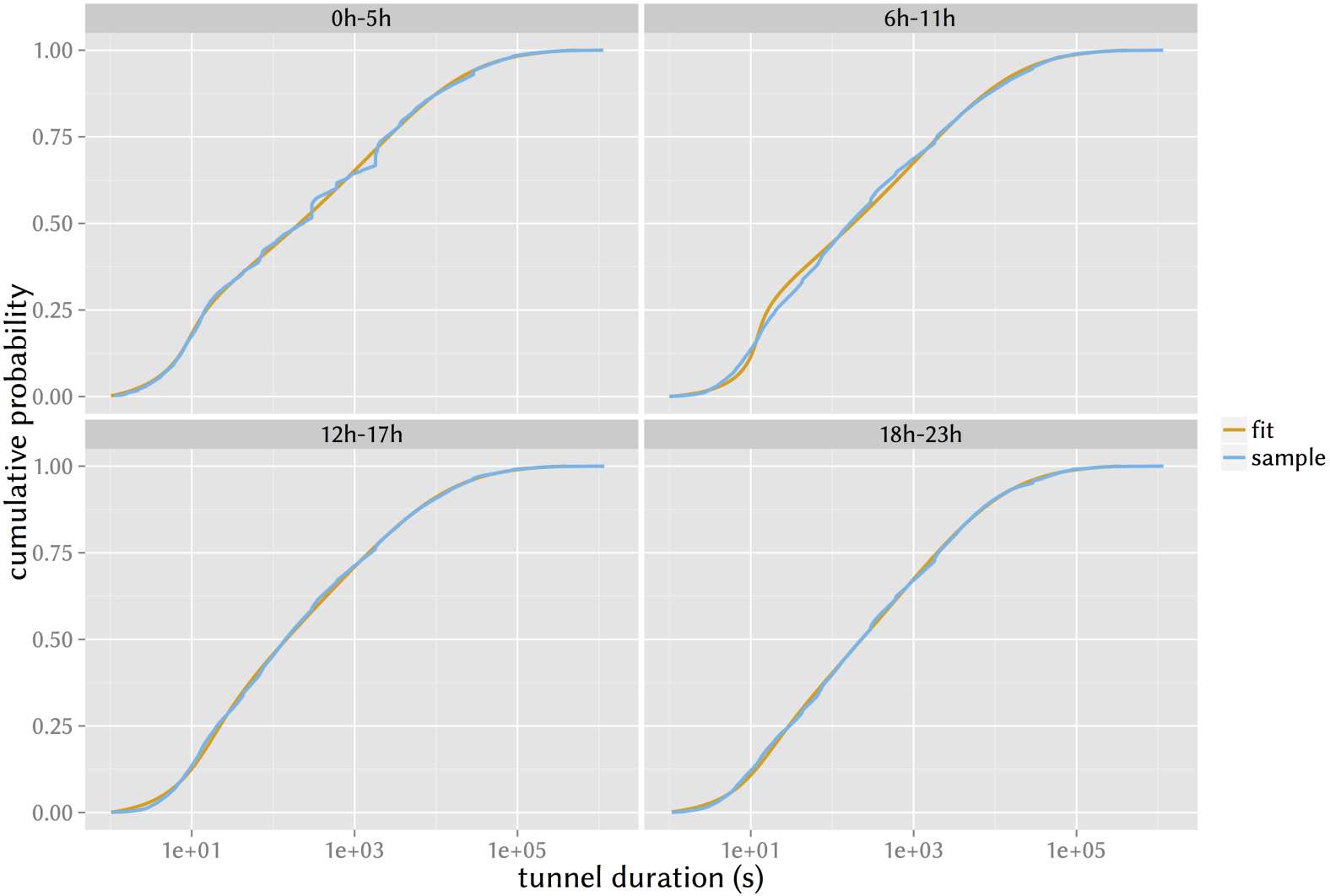


Figure: Sampled inter-arrival time CDF and fitted theoretical distributions.

Exponential Arrival Process Fits



Serving Time Rational Functions Fit

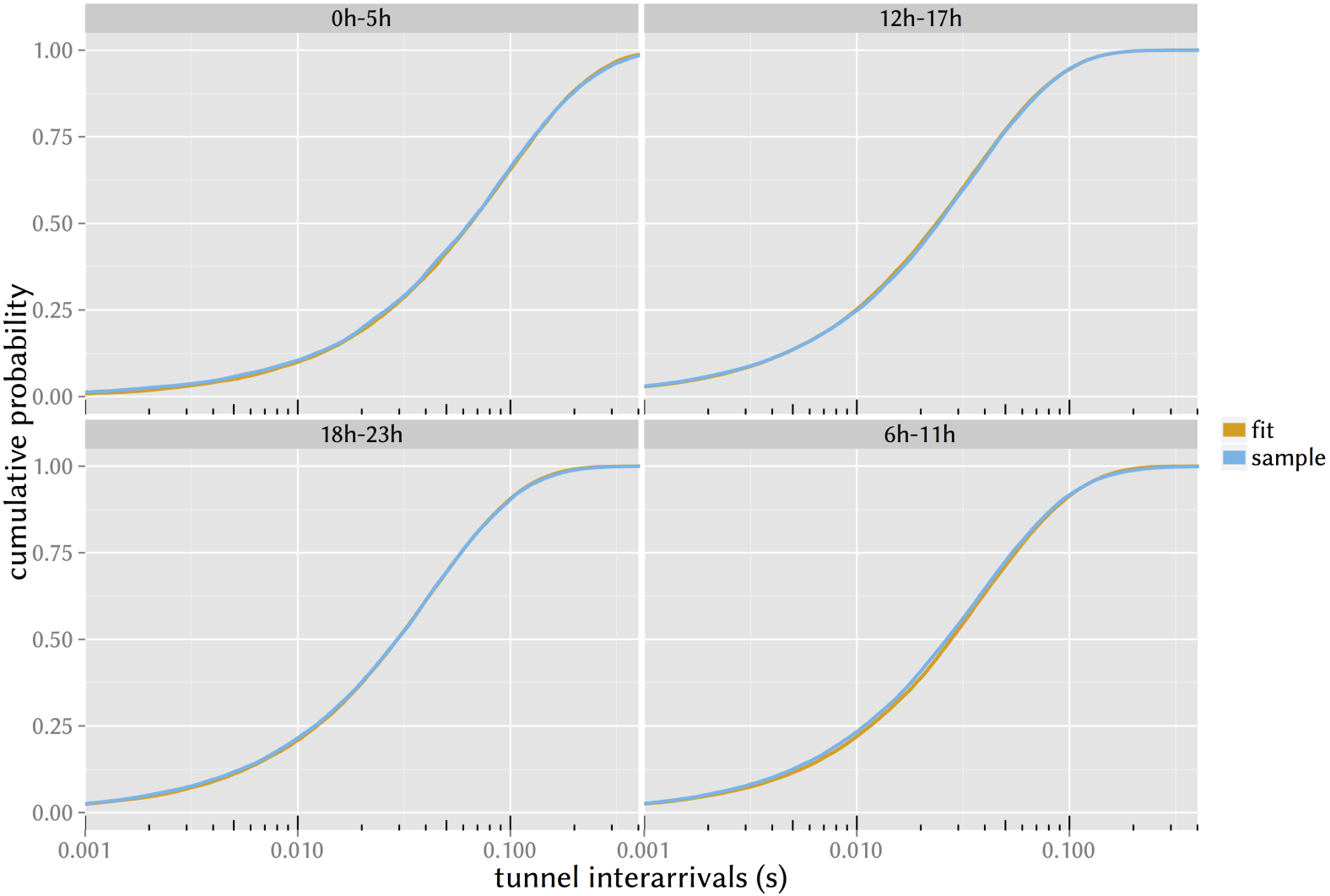


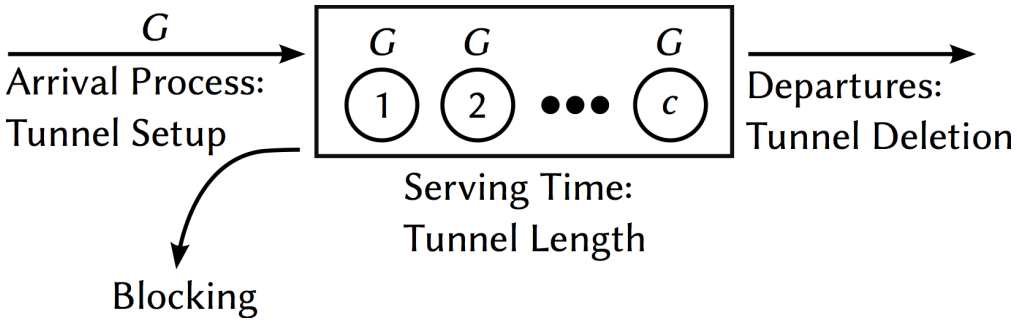
Table: Parameters for the exponentially distributed inter-arrival times and corresponding Pearson correlation coefficients.

Time of Day	λ	$R_{arrival}$
0h-5h	10.67477	0.995
6h-11h	24.53298	0.992
12h-17h	29.2504	0.993
18h-23h	23.49983	0.986

Table: Inverse rational functions fitted to the ECDFs of the tunnel duration by time of day and correlation coefficients of the fit.

Time of Day	**Inverse Serving Time CDF Representation**	R_{dur}
0h-5h	$0.919208 - 60.6136y - 3498.78y^3 - \frac{110.707y+2289.94y^3}{y-1.00469}$	0.999
6h-11h	$1 + 117.484y - 368.643y^2 - \frac{1720.13y^4}{y-1.0041}$	0.999
12h-17h	$0.952566 + 69.4907y + \frac{81146.1y^3+1.08572\times10^6y^5}{805-802.01y}$	0.999
18h-23h	$0.911924 + 82.0562y - \frac{2936.93y^4}{1.94468y-1.9532}$	0.999

Queuing Models



Described by Kendall's Notation $A/S/c/q$

- Distribution of the arrival process A
- Distribution of the serving time S
- Number of Servers c
- Queue Length q
  - $q = \infty$ no loss will occur
  - 0 loss/blocking system, no queue
- Evaluate
  - Average queue length and server occupation
  - Blocking probability

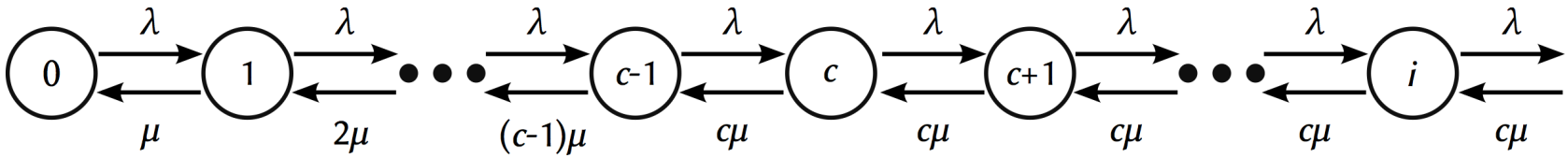


Figure: $M/M/c/\infty$ Markov chain model.

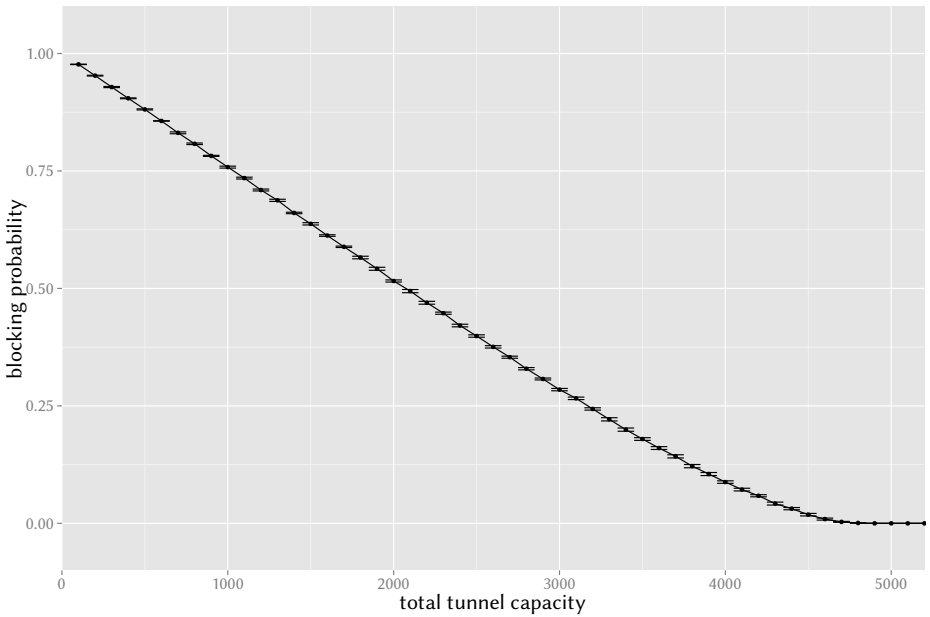


Figure: Impact of the number of supported parallel tunnels on the blocking probability for the traditional GGSN model.

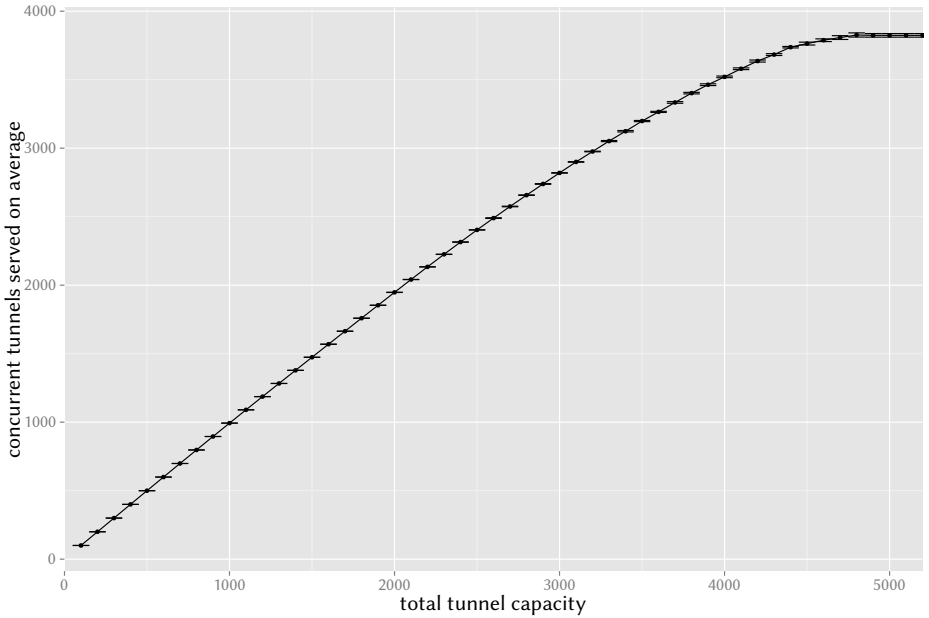


Figure: Mean number of tunnels concurrently served by the GGSN for incrementally increasing capacity.

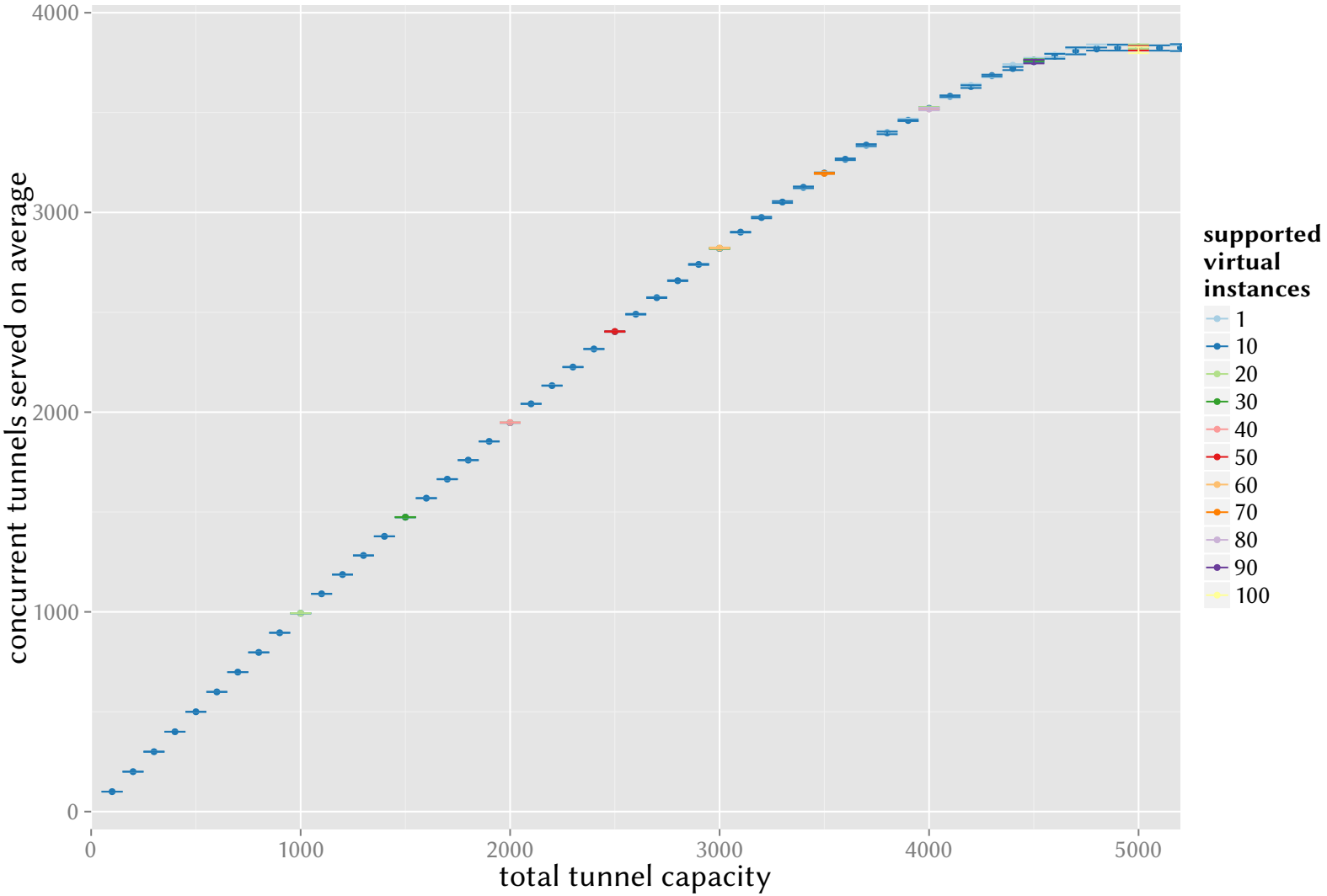


Figure: Comparison of the mean tunnel capacity usage of the individual virtual instance configurations.

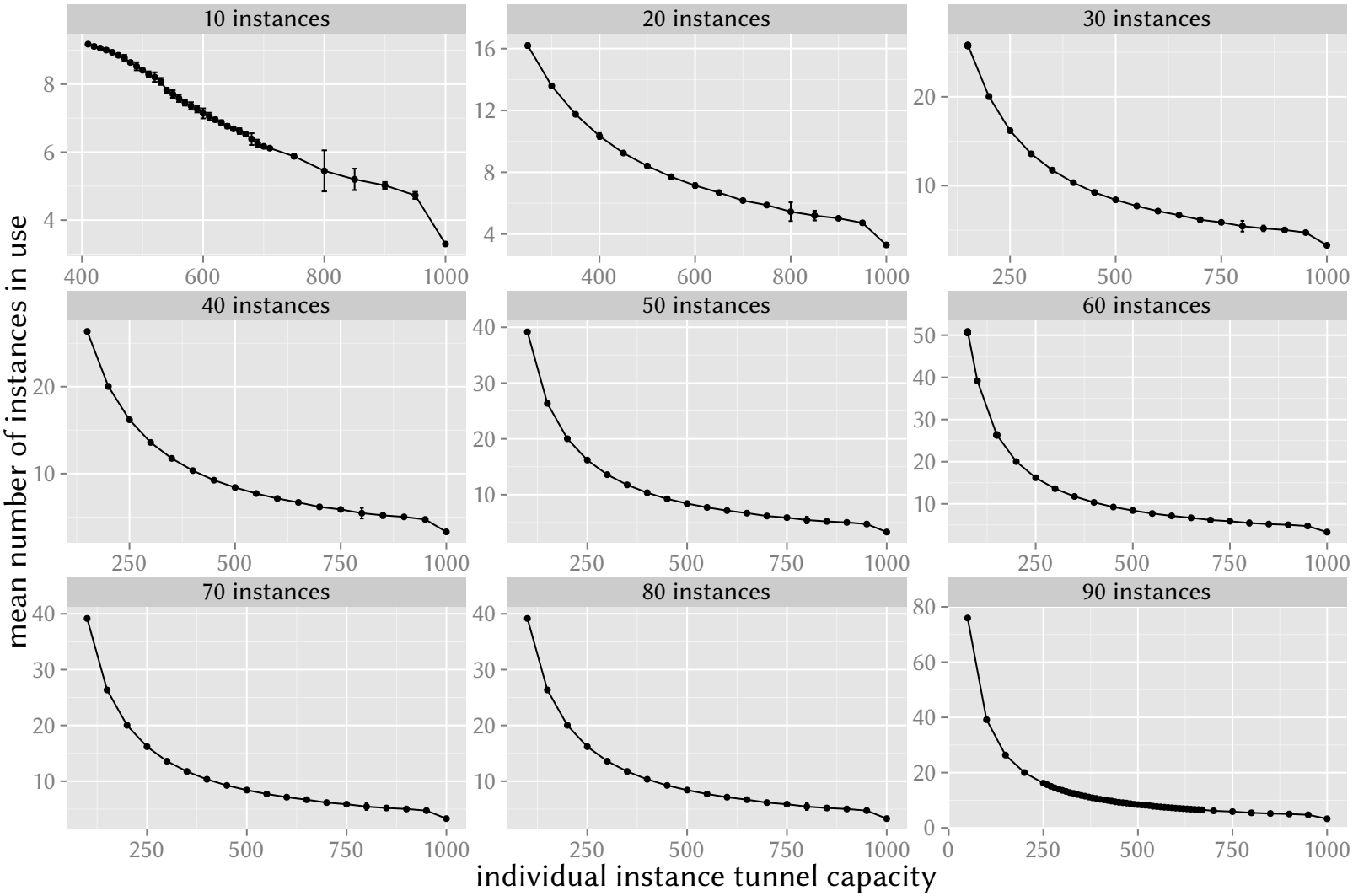


Figure: Mean instance usage of various virtualization configurations. A higher number of total instances results in a finer granularity of scaling and energy efficiency as more instances can be kept shut down.

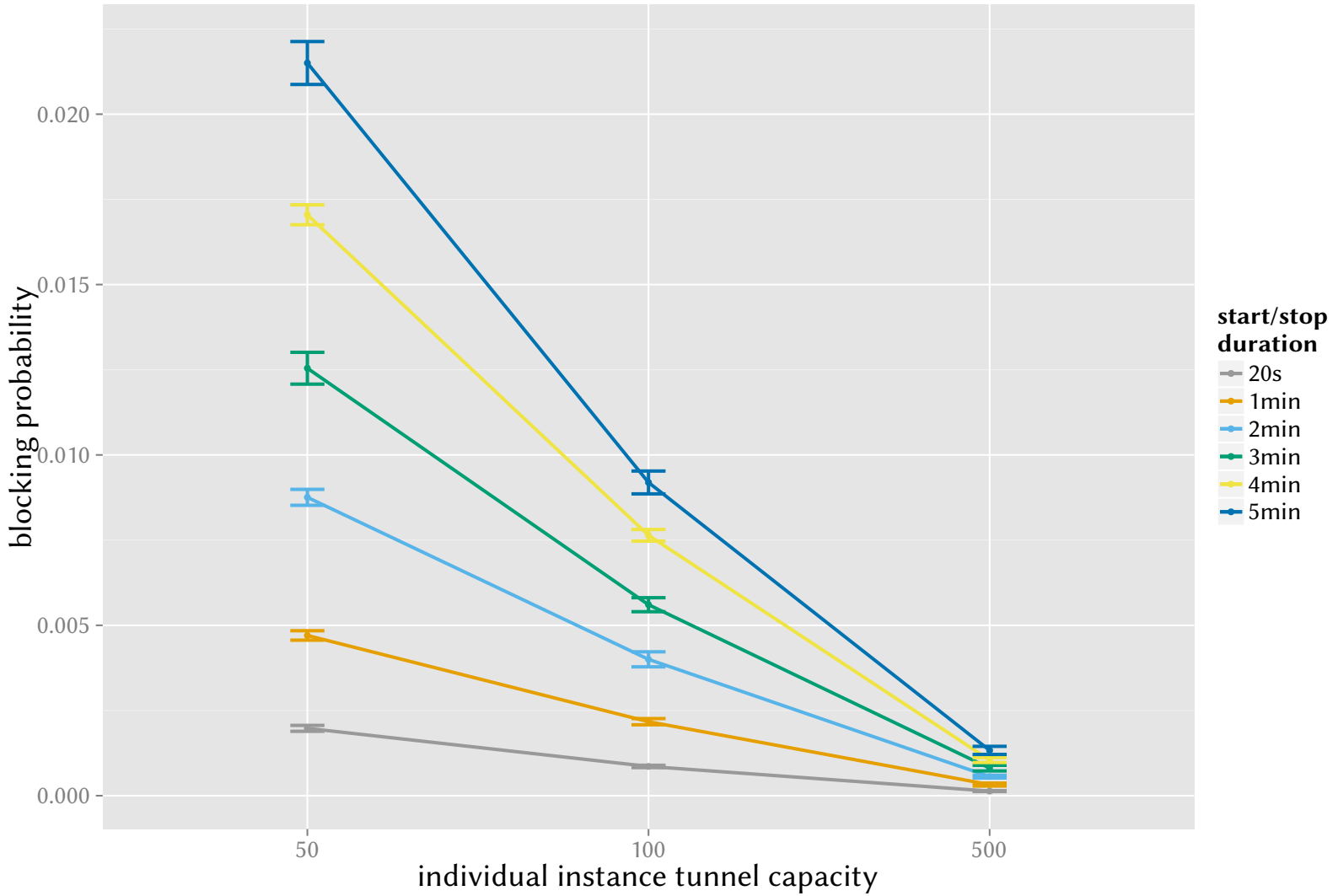


Figure: Influence of start up and shut down time on blocking probability with regard to different numbers of instances.

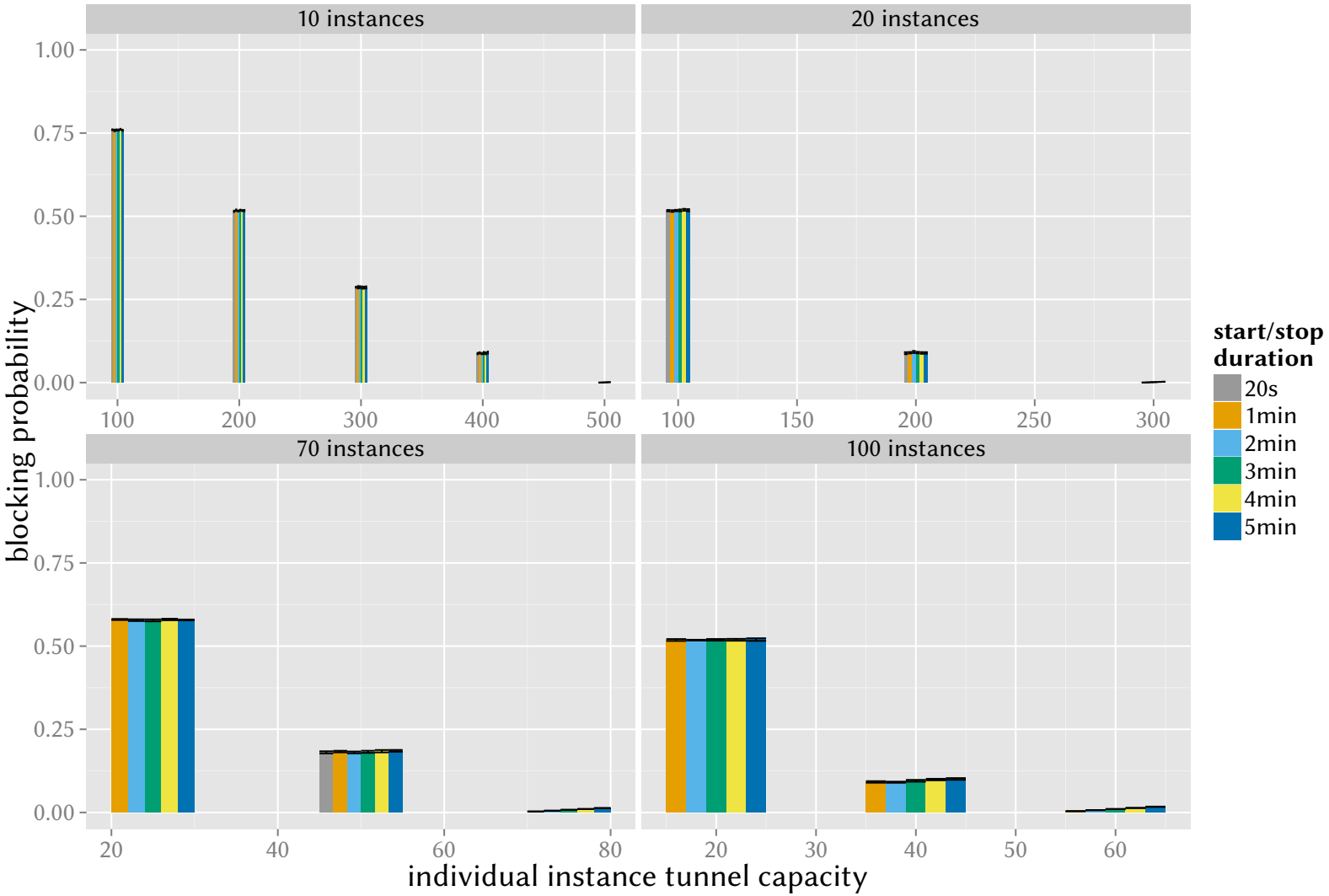
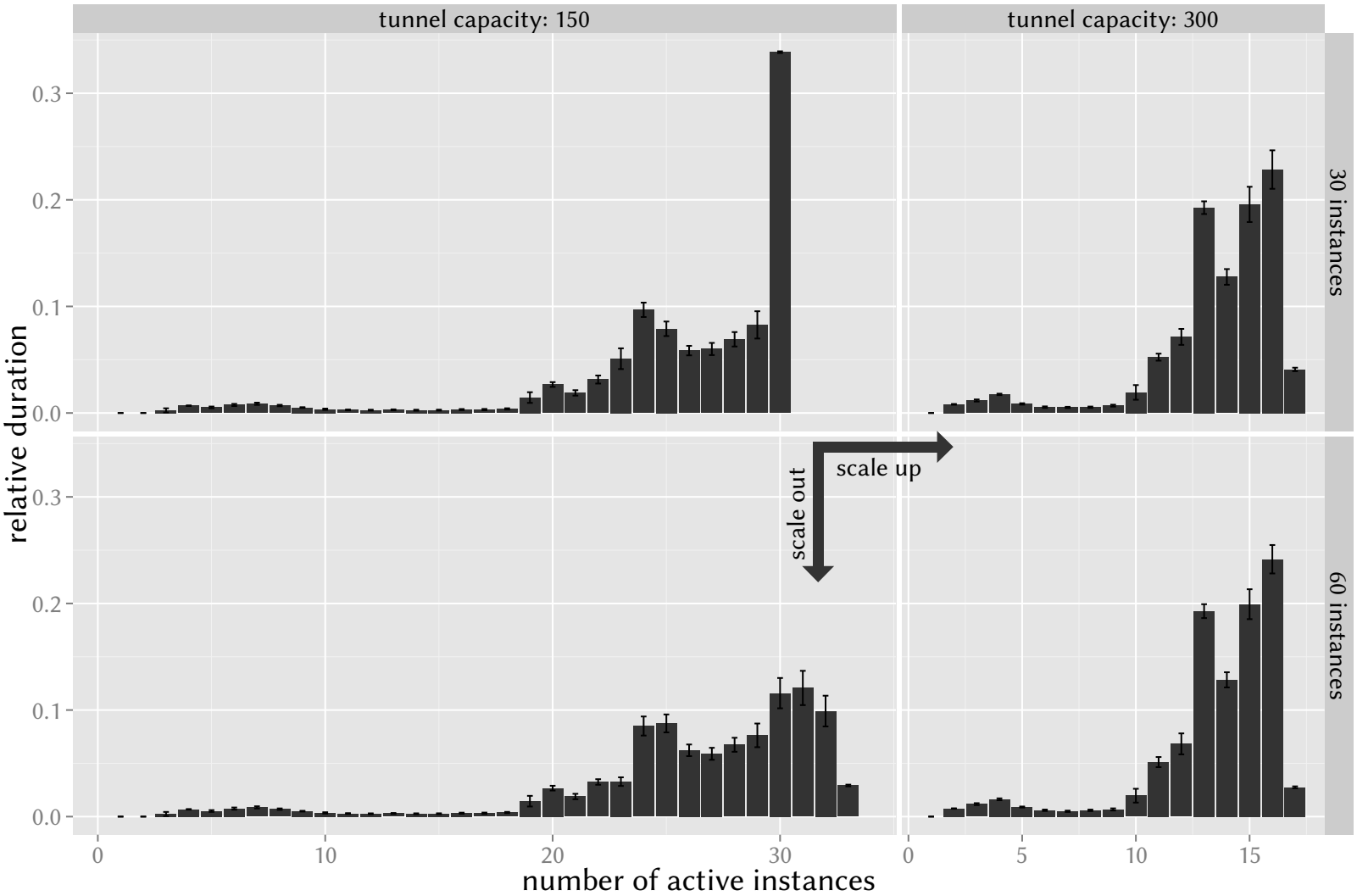


Figure: Influence of the boot and shutdown time on the blocking probability.

Scaling Up or Out with a Virtualized GGSN



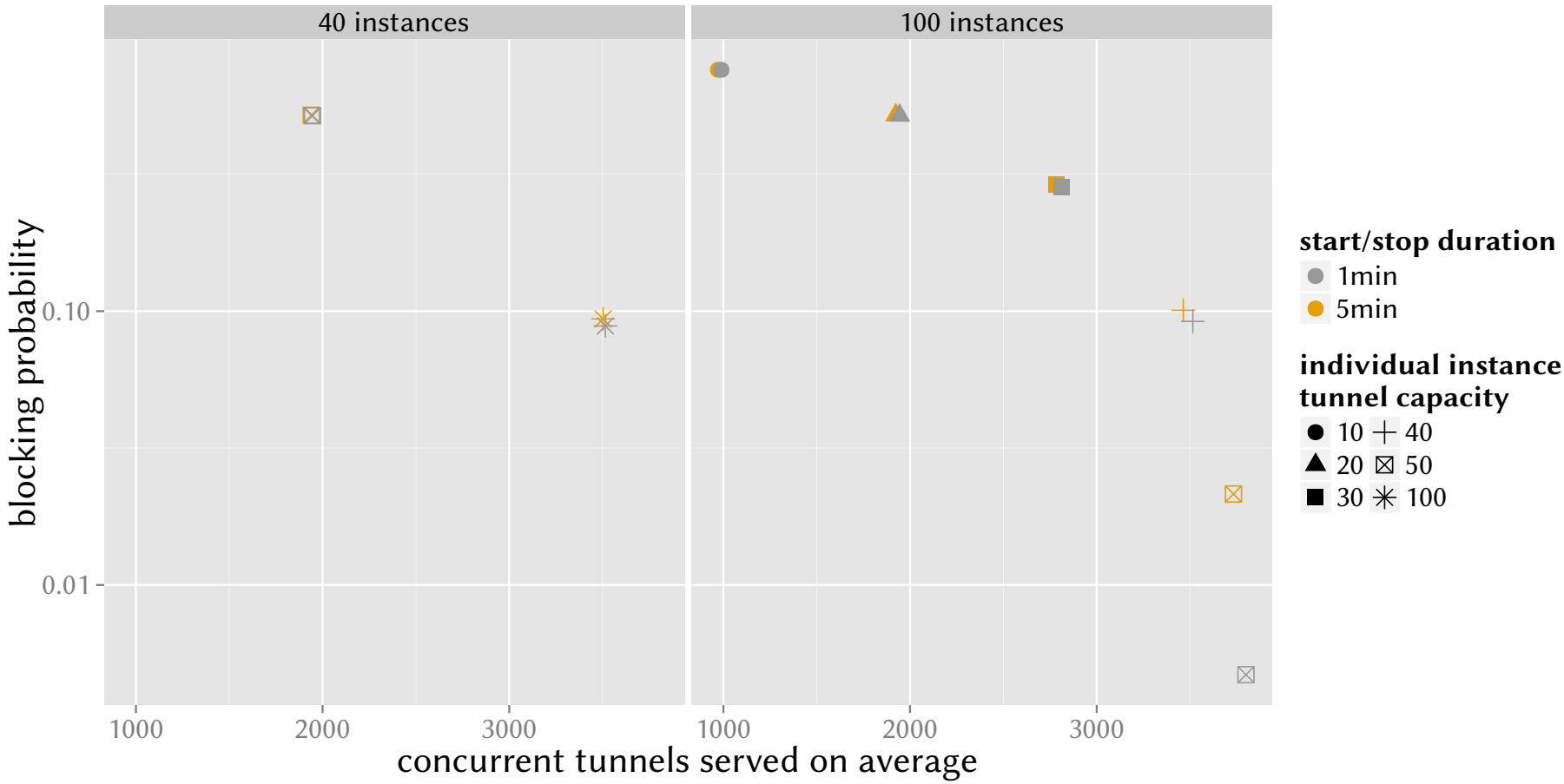
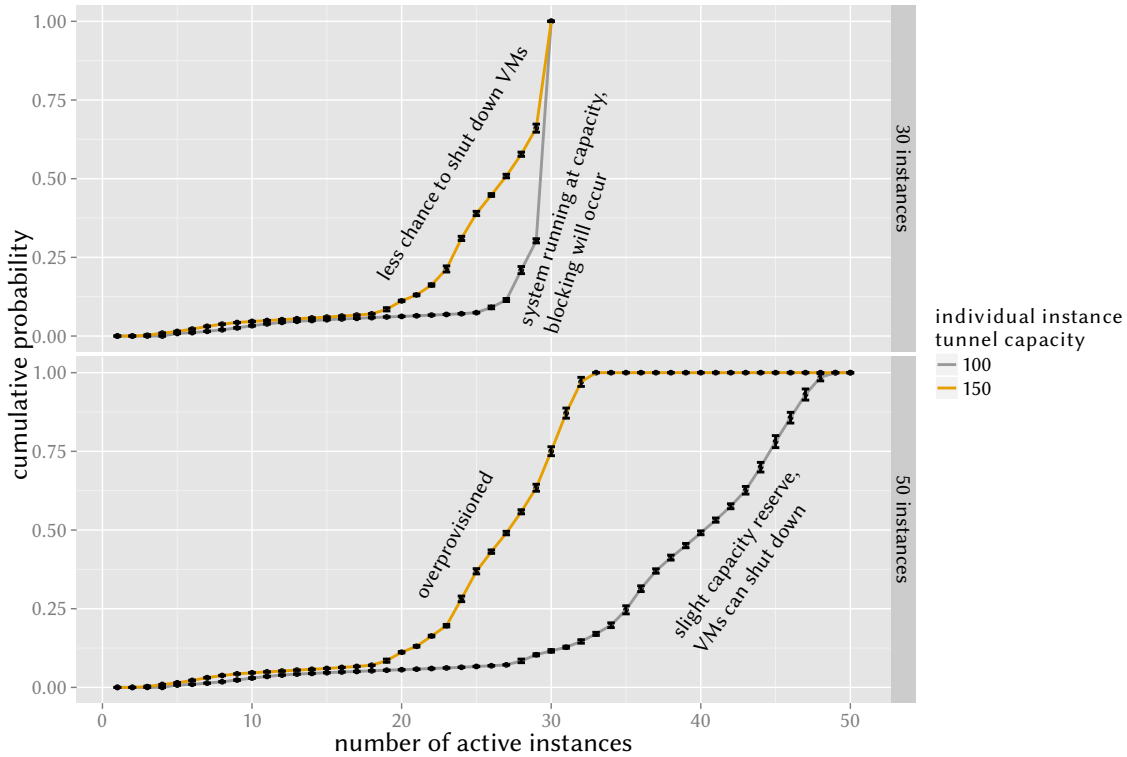


Figure: Trade-off between blocking probability and mean resource utilization with regard to maximum number of instances, instance tunnel capacity, and start and stop time.

Virtualized GGSN Impact



- Monolithic and virtualized GGSN scale equally with supported tunnels
- Scaling VM model by VM count instead of size has minimal impact on p_B
- Unused instances can be shut down for increased energy efficiency compared to monolithic model

Table: Effect sizes of the simulation parameters based on a one-way ANOVA.

	$F-ratio$	$p-value$	η^2	ω^2
Blocking probability				
Individual instance tunnel capacity	104	< 0.001	0.468	0.463
Number of instances	9.29	< 0.001	0.056	0.050
Start/stop duration	0.21	0.931	< 0.001	0.002
Total tunnel capacity	317257	< 0.001	0.999	0.999
Mean number of tunnels				
Individual instance tunnel capacity	105.7	< 0.001	0.472	0.467
Number of instances	9.39	< 0.001	0.056	0.050
Start/stop duration	0.25	0.912	< 0.001	0.002
Total tunnel capacity	365753	< 0.001	0.999	0.999

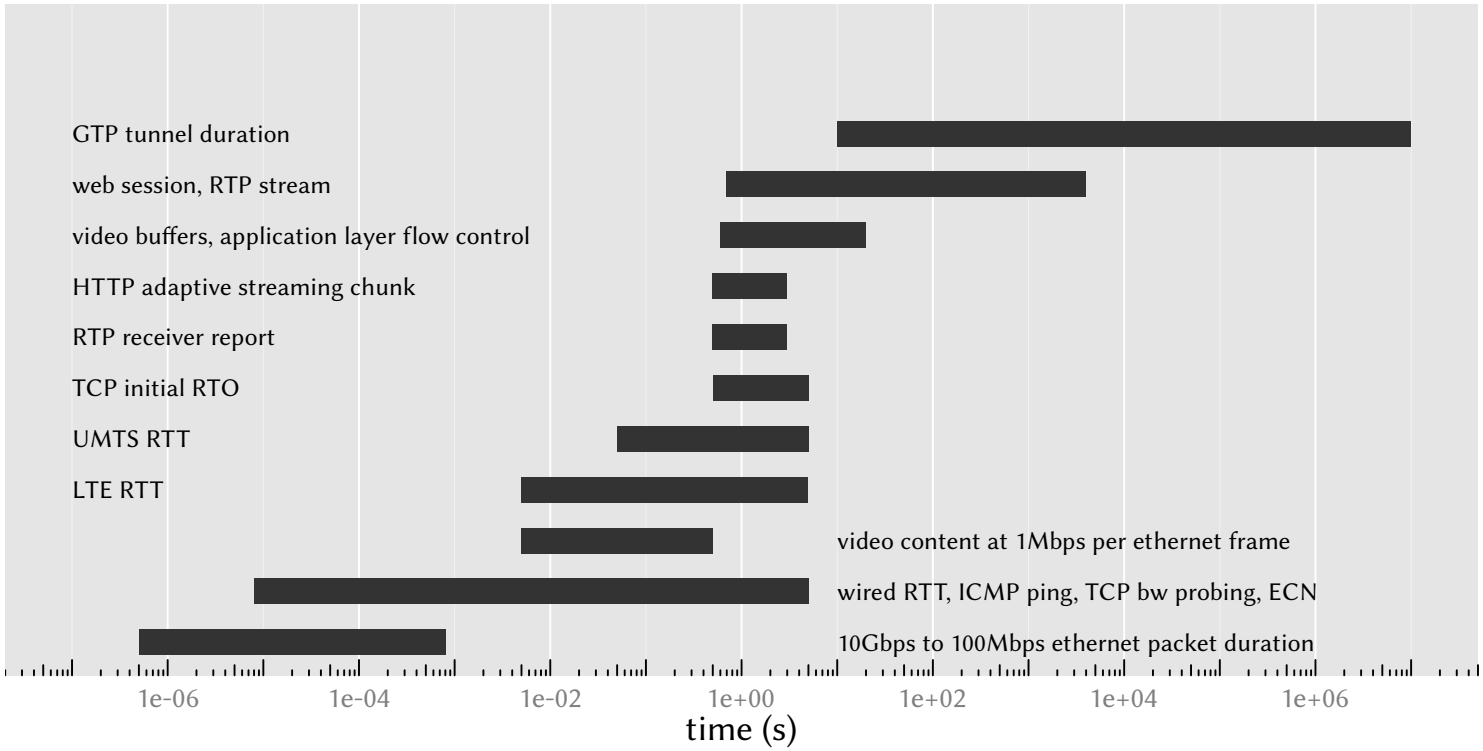


Figure: Approximate discernible time scales the networking stack protocols operate on in each layer.

Mobile Streaming Interactions

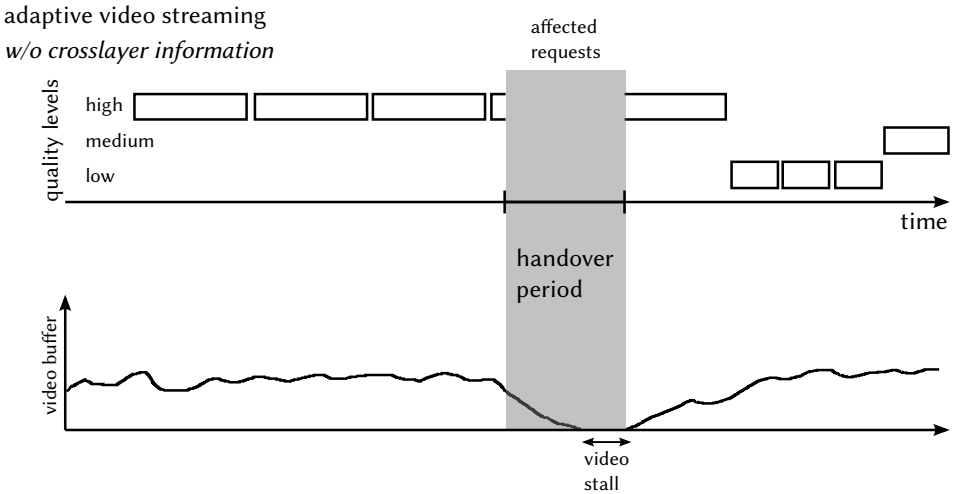
Use Case: Leveraging Crosslayer Information

- Make streaming apps aware of mobile net state to deal with influences
- Crosslayer information exchange model to pass baseband info to applications and include info in their decision making
- Predict future events through other context information sources, e.g. handover events through GPS data and neighboring cells

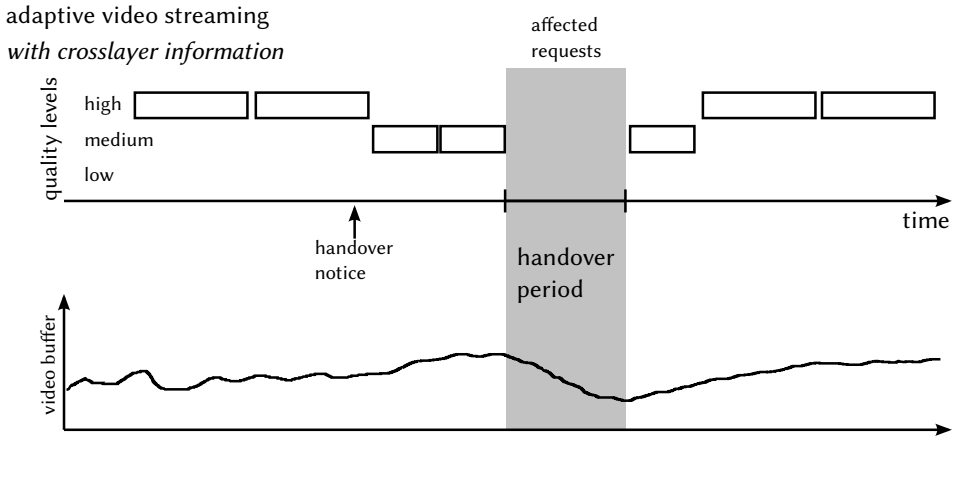
Mobile Streaming Interactions

Use Case: Leveraging Crosslayer Information

- Make streaming apps aware of mobile net state to deal with influences
- Crosslayer information exchange model to pass baseband info to applications and include info in their decision making
- Predict future events through other context information sources, e.g. handover events through GPS data and neighboring cells



Stalling occurs without handover hinting



Stalling can be prevented by proactively filling the playback buffer

TCP Changes in the Linux Kernel

Change	Related Work	Kernel	Date
BIC as default congestion avoidance algorithm (from Reno)		2.6.8	August 2004
CUBIC as default congestion avoidance algorithm	[HRX08]	2.6.19	November 2006
New TCP Slow Start: HyStart	[HR11]	2.6.29	March 2009
Multipath TCP	[RFC6824]	external	2011
TCP User Timeout	[RFC5482]	2.6.37	January 2011
Initial Receive Window $10\,MSS$	[RFC6982]	2.6.38	March 2011
Initial Congestion Window $10\,MSS$	[RFC6982]	2.6.39	May 2011
$1\,s$ initial RTO (from $3\,s$)	[RFC6298]	3.1	October 2011
Changes to sstresh and CWND bevahior	[RFC5681]	3.1	October 2011
TCP Proportional Rate Reduction	[RFC6937]	3.2	January 2012
Byte queue limits and TCP buffer limits		3.3	March 2012
CoDel AQM	[NJ14]	3.5	July 2012
TCP Early Retransmit	[RFC5827]	3.5	July 2012
TCP small queues		3.6	September 2012
TCP Fast Open (client side)	[Che+14]	3.6	September 2012
TCP Fast Open (server side)		3.7	December 2012
TCP tail loss probe		3.10	June 2013
TCP Forward RTO-Recovery	[RFC5682]	3.10	June 2013
Low latency network polling		3.11	September 2013
Improved RTO calculation and handling of reordering		3.12	November 2013
TCP Fast Open enabled by default		3.13	January 2014
TCP auto corking		3.14	March 2014
PIE AQM		3.14	March 2014
TCP Fast Open over IPv6		3.16	2014
LISP	[RFC6830]	3.16	2014

Leveraging Cross-Layer Information

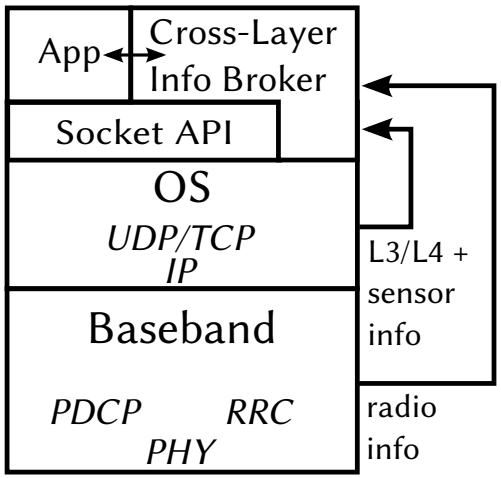
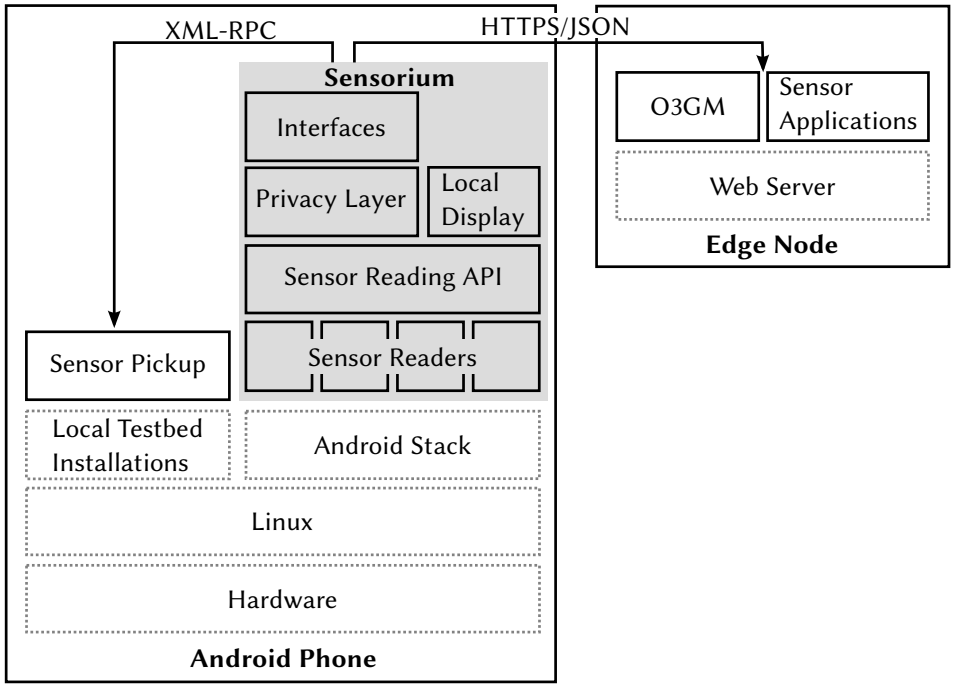


Figure: Model and architecture of the proposed cross-layer information exchange.

Active Measurement Testbed with Metadata

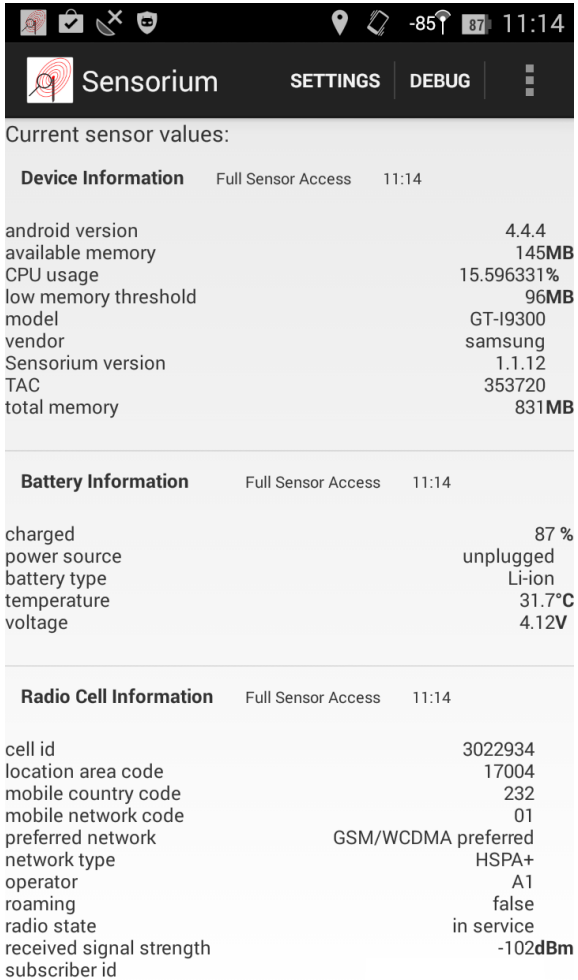
Using Sensorium

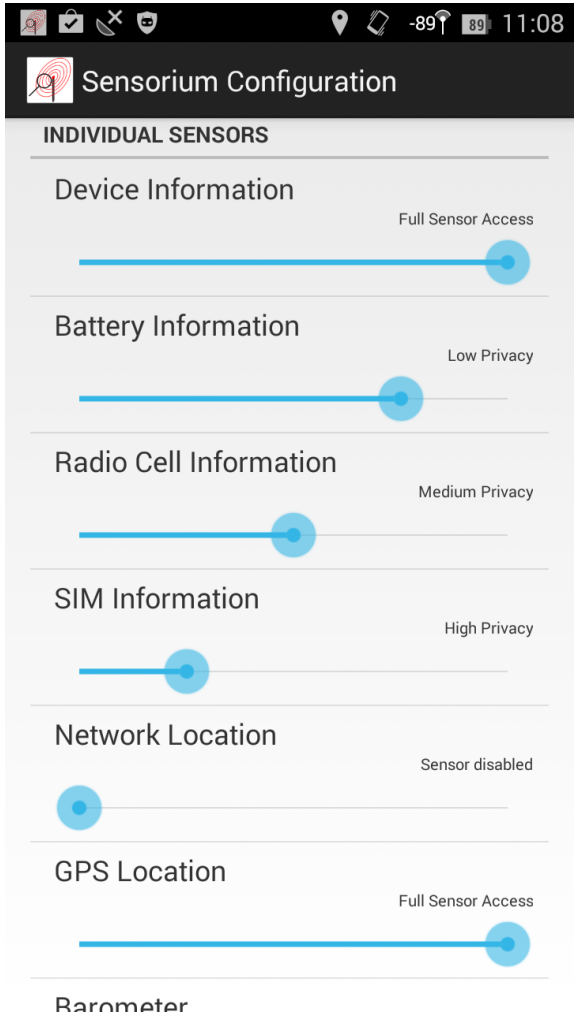
- Difficulty to conduct realistic mobile measurement campaigns
- Importance of additional metadata in mobile measurements (location, connectivity, baseband state)
- Privacy issues, let the user control data access and granularity
- Provide sensor-reading and collection framework for smartphones to conduct ones own campaign on “foreign” devices



- Demo campaign **Open 3G Map**: Collected 3G coverage data and aggregate to map overlay

Sensorium


-85 87 11:14
Sensorium
SETTINGS
DEBUG
Current sensor values:
Device Information
Full Sensor Access
11:14
android version 4.4.4
available memory 145MB
CPU usage 15.596331%
low memory threshold 96MB
model GT-I9300
vendor samsung
Sensorium version 1.1.12
TAC 353720
total memory 831MB
Battery Information
Full Sensor Access
11:14
charged 87 %
power source unplugged
battery type Li-ion
temperature 31.7°C
voltage 4.12V
Radio Cell Information
Full Sensor Access
11:14
cell id 3022934
location area code 17004
mobile country code 232
mobile network code 01
preferred network GSM/WCDMA preferred
network type HSPA+
operator A1
roaming false
radio state in service
received signal strength -102dBm
subscriber id


-89 89 11:08
Sensorium Configuration
INDIVIDUAL SENSORS
Device Information
Full Sensor Access
Battery Information
Low Privacy
Radio Cell Information
Medium Privacy
SIM Information
High Privacy
Network Location
Sensor disabled
GPS Location
Full Sensor Access
Barometer

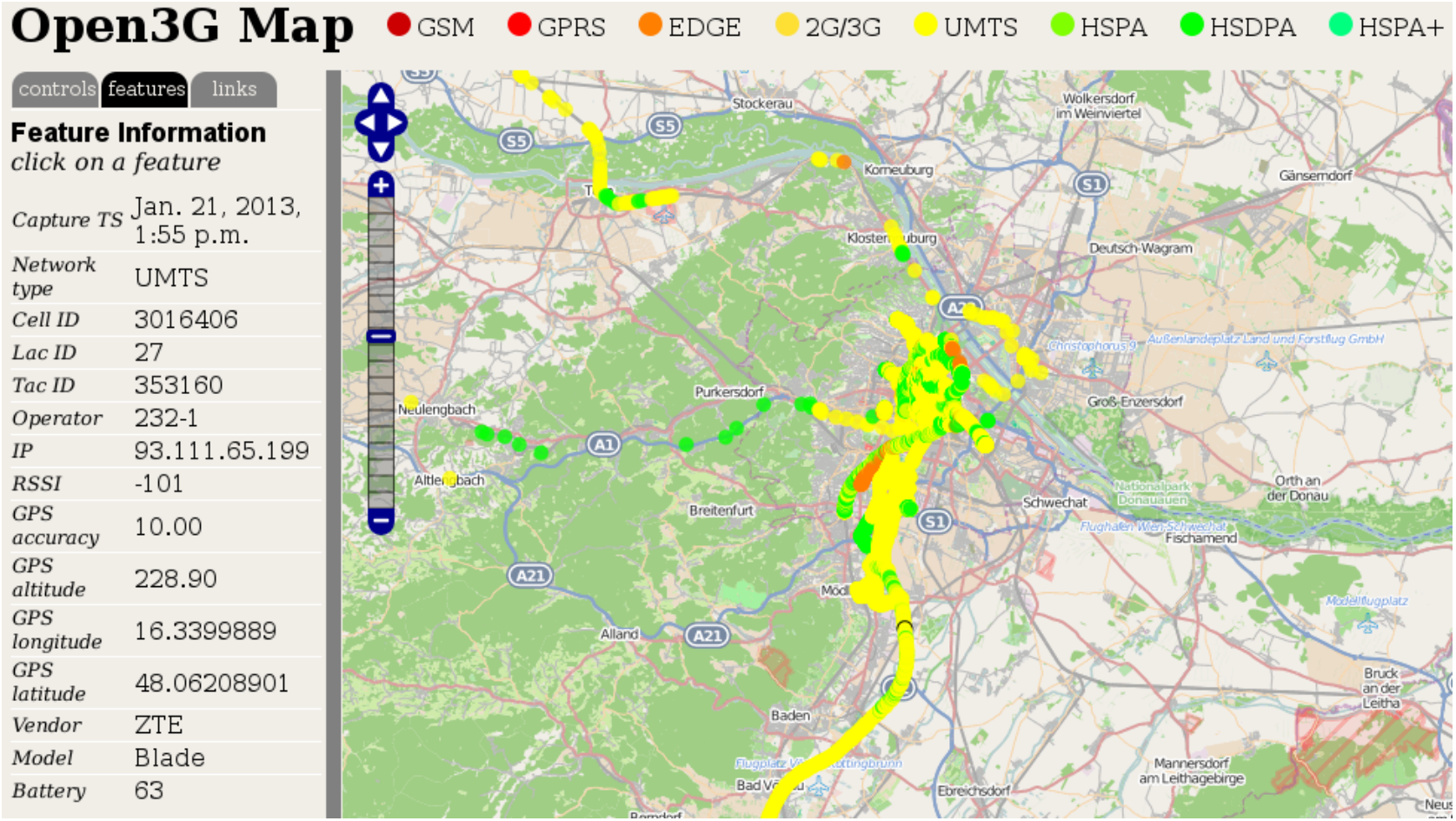


Figure: The O3GM web page, displaying a 3G coverage measurements layer with data collected by Sensorium on top of the OpenStreetMap base layer.

Reliable Streaming

Categorization and Measuring

Streaming categorization criteria

- Video source (Live, VOD)
- Adaptivity, e.g.
  - Implicit through unreliable transmission (RTP)
  - Explicit through multiple quality levels (adaptive streaming, e.g. DASH)
- Location of streaming control
  - Server-side: push-based (RTP)
  - Client-side: pull-based (HTTP)
- Transport/application protocol and transmission pattern

Reliable Streaming

Categorization and Measuring

Streaming categorization criteria

- Video source (Live, VOD)
- Adaptivity, e.g.
  - Implicit through unreliable transmission (RTP)
  - Explicit through multiple quality levels (adaptive streaming, e.g. DASH)
- Location of streaming control
  - Server-side: push-based (RTP)
  - Client-side: pull-based (HTTP)
- Transport/application protocol and transmission pattern

Measurement Approach

- Transmission process has no influence on the image quality (reliability)
- Evaluate streaming solely on the basis of buffering behavior
- Number and duration of stalls as metric for non-adaptive streaming

Mobile Streaming Simulation

Simulator Survey

- Much easier to conduct than active measurements, but often with limited fidelity
- Conducted survey for the suitability of existing mobile streaming simulators
  - Most only concerned with radio interface simulation, not core or signaling
  - Needs to have a fully functioning up-to-date network stack
  - None freely available and up-to-date for 3G
  - Candidates for LTE: Omnet++ and ns-3
  - Both provide rudimentary core and tunneling but no control plane
  - Attributable to specification complexity, control plane in particular
- ns-3 chosen as basis for simulation framework, with limitations in mind
- Can be additionally used as a network emulator

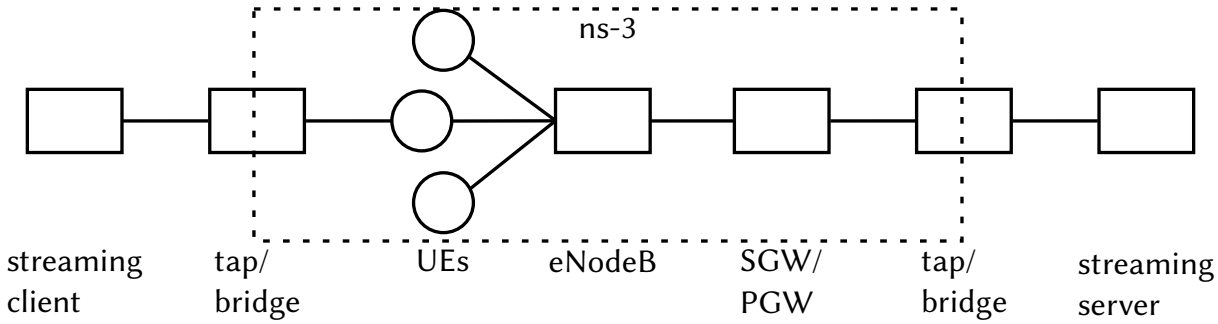
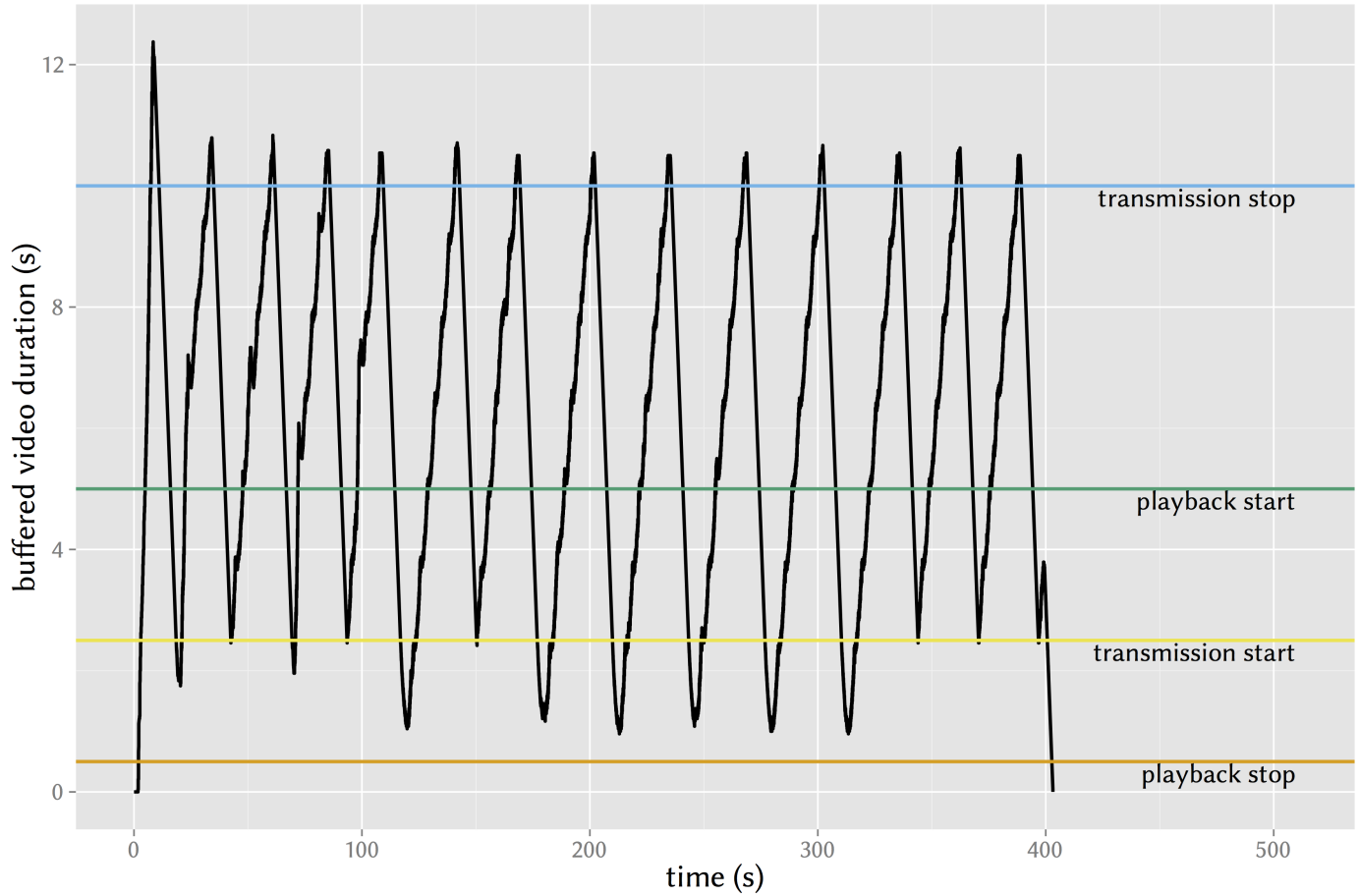


Figure: Future testbed iteration: hybrid of ns-3 LTE simulation and actual or emulated streaming client and server bridged to it.

Mobile Streaming Strategies

Example Four Threshold Strategy



Limits for playback start/stop, transmission start/stop

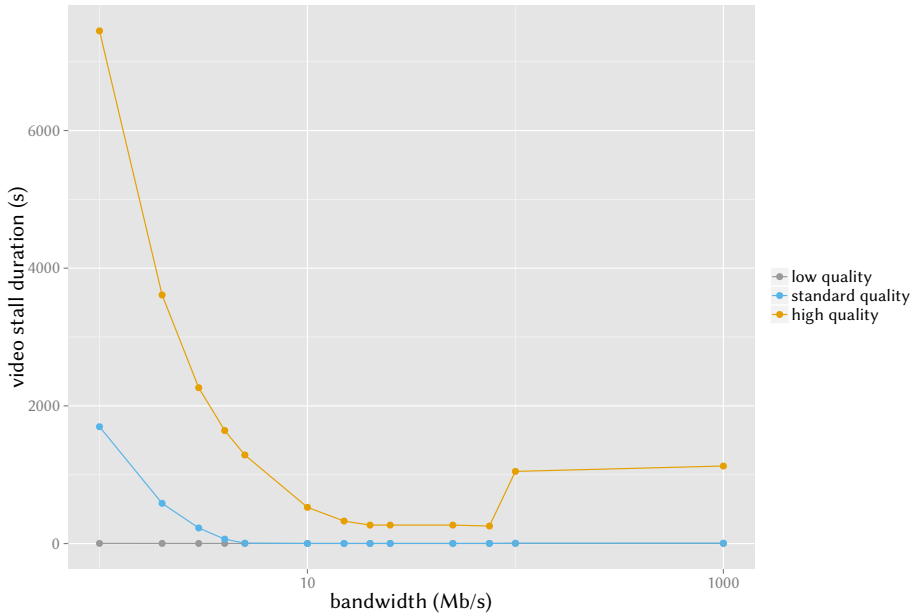


Figure: Relative stalling duration of the simulated reliable streaming player under limited Internet bandwidth.

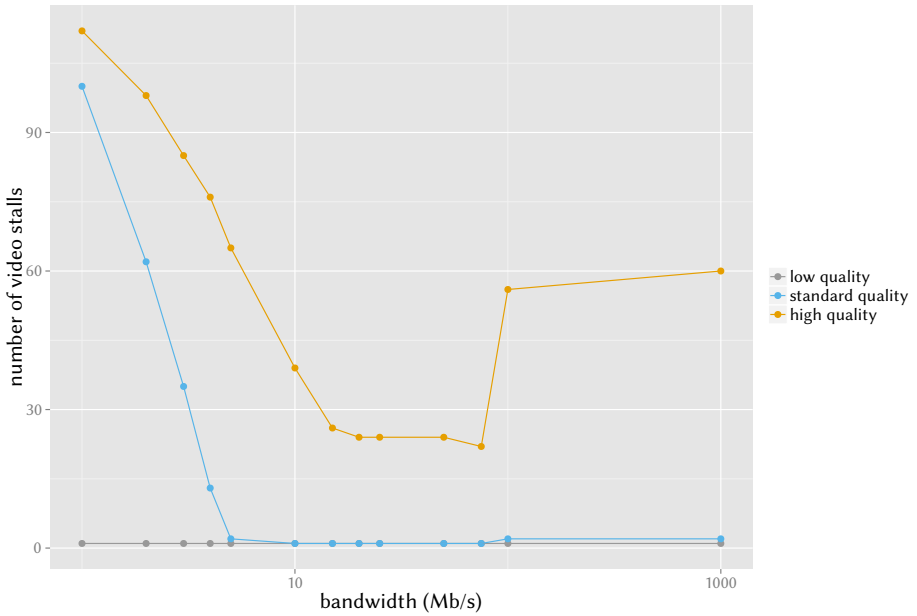


Figure: Number of stalling events of the simulated reliable streaming player under limited Internet bandwidth.

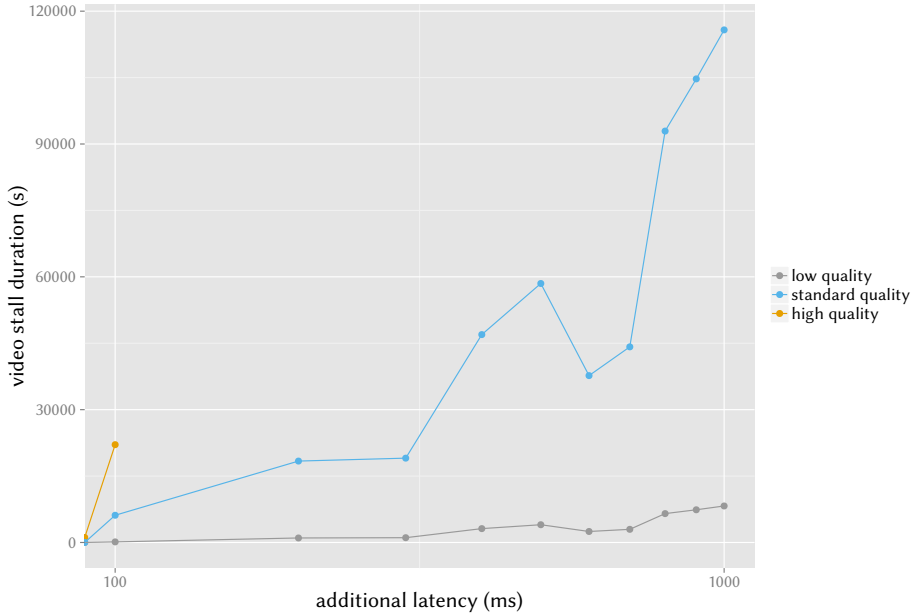


Figure: Relative stalling duration of the simulated reliable streaming player under increased Internet latency.

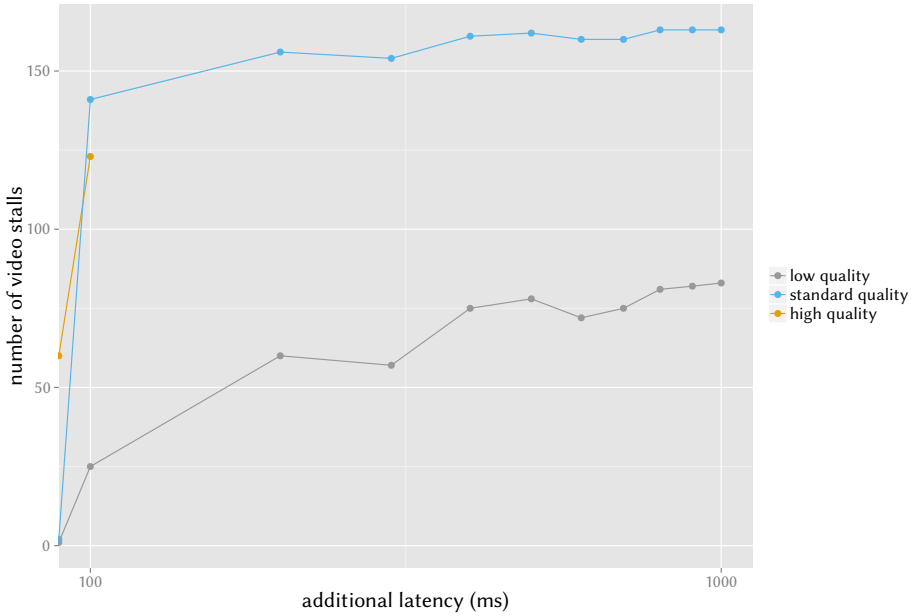


Figure: Number of stalling events of the simulated reliable streaming player under increased Internet latency.