QoS and QoE in 5G networks

Moving towards real-time applications

QSDG Workshop, Singapore, 19-21 August 2019

Dr. Jens Berger
Rhode & Schwarz Mobile Network Testing

Agenda

- From ‘always online’ towards ‘seamless connectivity’
- Live video and real-time streaming
- Online gaming as interactive challenge, AR/VR all over…
- What are the requirements to become ‘real-time interactive’?
- How to measure ‘interactivity’?
- How perception and QoE models will look like?

What happens behind the scenes technically?

- Simple Video Download is replaced by **DASH**
- HTTP almost disappeared and is replaced by **HTTPs**
- FTP almost disappeared
- TCP is more and more replaced by **QUIC**
- Today’s remaining UDP traffic could be replaced by a QUIC derivate to make it reliable

- New applications will create and use new types of protocols (e.g. AR, VR)

- Far most data traffic is handled by proprietary and encrypted protocols

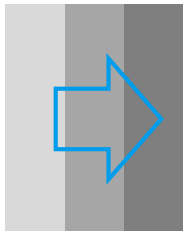
From ‘always online’ towards ‘seamless connectivity’

- Using mobile data services is not a question anymore today
- Almost each App or use case communicates with a host-server
- Smartphone users are always attached to the data network
- …it is a give and take. Connectivity is given, apps make use of it.

- There is a transition…

…because the networks are getting prepared for it

File-based, non-real-time web-services
- Retrieving a map
- Posting a message or a photo
- Browsing the internet
- Downloading a video



Continuous data-exchange, real-time applications
- Watching live video in real-time
- OTT (video-) telephony and conferencing
- On-line navigation
- On-line gaming

It’s enough to be always online
→ means ‘reachable’

Many shades of grey in between

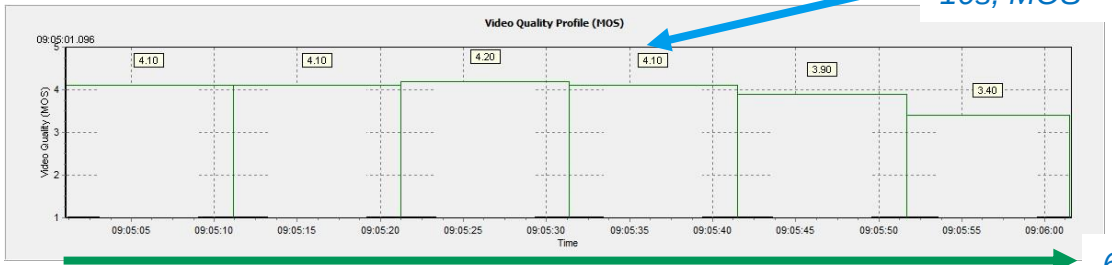
It requires more than ‘being online’:
→ seamless connectivity!

Let’s start simple: Video Streaming

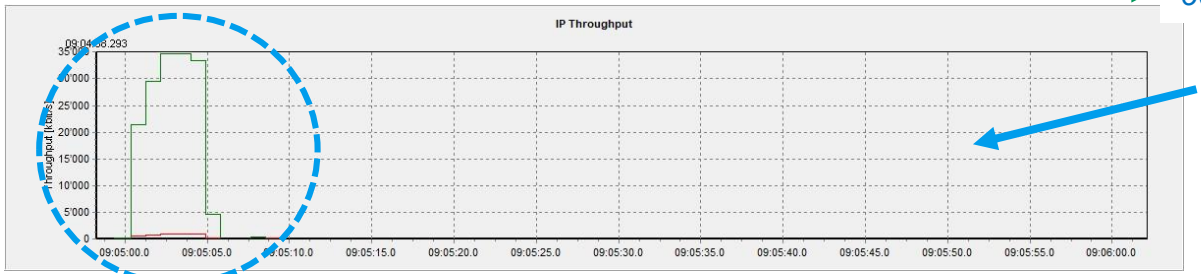
Video on demand
> 90% of all video traffic

- **Video on demand**, example video 60s length, 1080p HD
- YouTube™ v.13.14.55, excellent LTE coverage
- 1080p HD all time

Picture based Video Quality every 10s, MOS ~4.1 in good cases



60s video playout



IP traffic for 60s video is only a 5s peak ~30Mbit/s)

Let’s start simple: Video Streaming

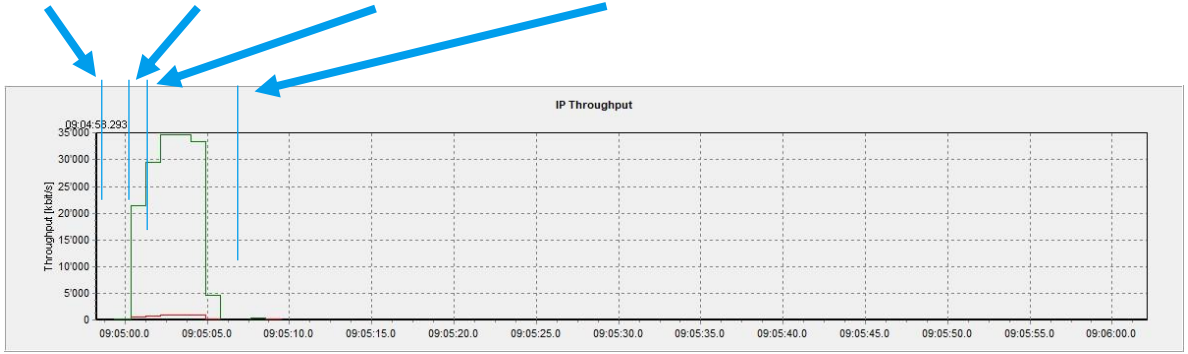
- **Video on demand**, example video 60s length, 1080p HD
- YouTube™ v.13.14.55, excellent LTE coverage

58.47s Start

59.81s → *00.42s* → *01.34s* → *06.46s*

1st contact ‘Google’ | Start Buffering | Start Displaying | End Download

29.05.2018 09:05:00.417	Youtube Service Player	Buffering started
29.05.2018 09:05:00.418	Youtube Service Player	Clip duration: 00:01:00
29.05.2018 09:05:01.096	Youtube Service Player	New resolution: 1920x1088
29.05.2018 09:05:01.329	Youtube Service Player	Buffering ended
29.05.2018 09:05:01.336	Youtube Service Player	Displaying



- ~2s to 1st video packet
- ~1s initial buffering
- ~6s complete download

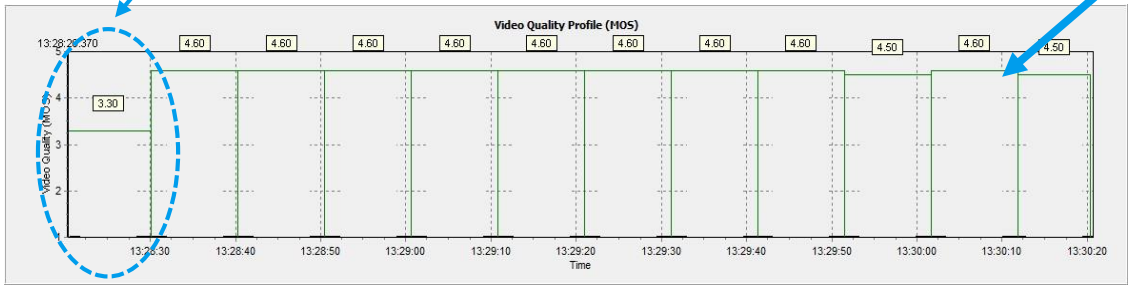
  → 60s displaying

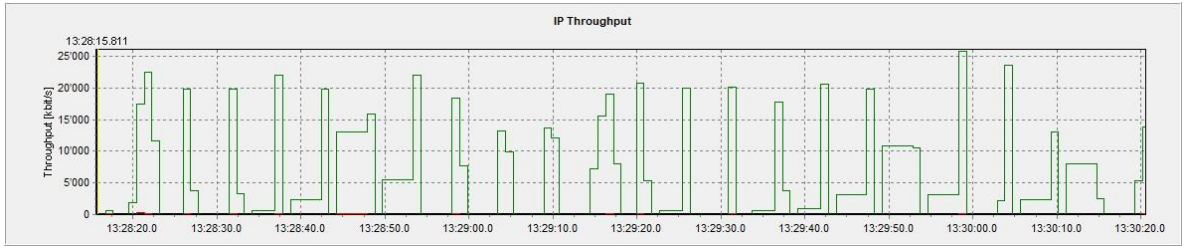
Let's go live: **Live** Video Streaming

- **Live video** (Sky News), 120s play length, 1080p HD*
- YouTube™ v.13.14.55, excellent LTE coverage

Initial 5s in 144p (to start fast)

1080p all time





Let's go live: **Live** Video Streaming

- **Live video** (Sky News), 120s play length, 1080p HD*
- YouTube™ v.13.14.55, excellent LTE coverage

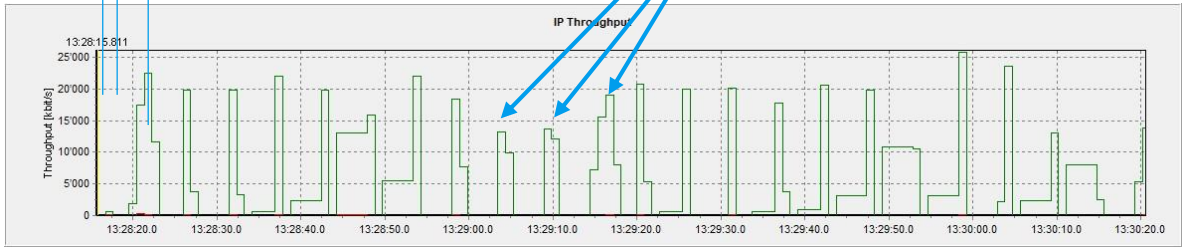
18.04.2018 13:28:17.228	Youtube Service Player	Live stream
18.04.2018 13:28:19.251	Youtube Service Player	Buffering started
18.04.2018 13:28:20.370	Youtube Service Player	New resolution: 256x144
18.04.2018 13:28:20.573	Youtube Service Player	Buffering ended
18.04.2018 13:28:20.574	Youtube Service Player	Displaying
18.04.2018 13:28:20.593	Youtube Service Player	Playing
18.04.2018 13:28:25.518	Youtube Service Player	New resolution: 1920x1088

16.81s → *19.25s* → *20.47s*

1st contact Google — Start Buffering — Start Displaying

This is not real time!

Chunks of video data every 5s



- ~2.5s to 1st video packet
- ~1.2s initial buffering

Examples for adaptive bitrate (YouTube™)

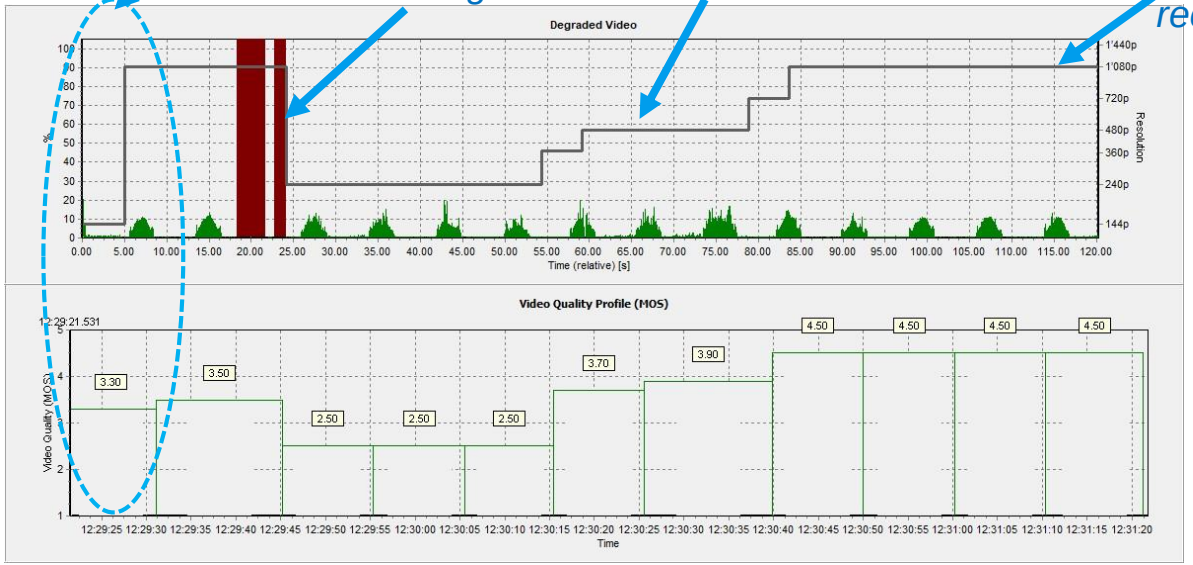
- **Live video** (Sky News), 120s play length
- YouTube™ v.13.14.55, bad coverage

Initial 5s in 144p

Freezing

Resolution

1080p after recovering



Examples for adaptive bitrate (YouTube™)

- **Live video** (Sky News), 120s play length
- YouTube v.13.14.55, bad 3G coverage

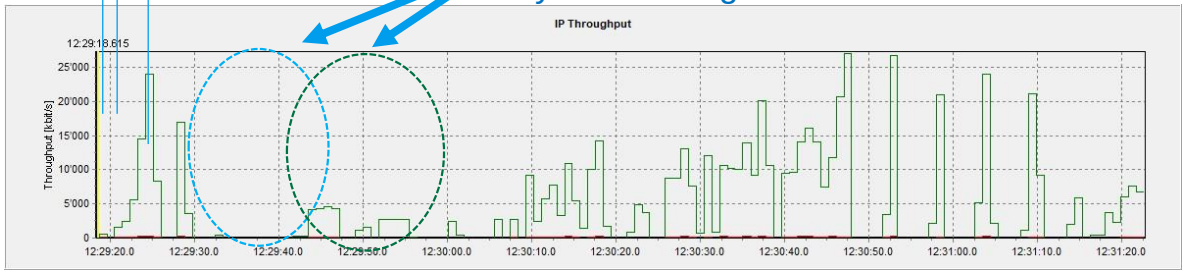
18.04.2018 12:29:18.912	Youtube Service Player	Live stream
18.04.2018 12:29:20.491	Youtube Service Player	Buffering started
18.04.2018 12:29:21.531	Youtube Service Player	New resolution: 256x144
18.04.2018 12:29:21.653	Youtube Service Player	Buffering ended
18.04.2018 12:29:21.654	Youtube Service Player	Displaying
18.04.2018 12:29:21.697	Youtube Service Player	Playing
18.04.2018 12:29:26.588	Youtube Service Player	New resolution: 1920x1088
18.04.2018 12:29:38.831	VMon	Intermediate MOS: 3.3
18.04.2018 12:29:39.997	Youtube Service Player	Rebuffering started
18.04.2018 12:29:43.274	Youtube Service Player	Rebuffering ended
18.04.2018 12:29:43.275	Youtube Service Player	Playing
18.04.2018 12:29:43.307	Youtube Service Player	Rebuffering started
18.04.2018 12:29:43.310	Youtube Service Player	Rebuffering ended
18.04.2018 12:29:43.312	Youtube Service Player	Playing
18.04.2018 12:29:43.330	Youtube Service Player	Rebuffering started
18.04.2018 12:29:43.332	Youtube Service Player	Rebuffering ended
18.04.2018 12:29:43.333	Youtube Service Player	Playing
18.04.2018 12:29:43.349	Youtube Service Player	Rebuffering started
18.04.2018 12:29:43.354	Youtube Service Player	Rebuffering ended
18.04.2018 12:29:43.354	Youtube Service Player	Playing
18.04.2018 12:29:43.357	Youtube Service Player	Rebuffering started
18.04.2018 12:29:43.376	Youtube Service Player	Rebuffering ended

18.95s → *20.49s* → *21.65s*

1st contact Google — Start Buffering — Start Displaying

This is not a seamless transmission!

Freezing followed by re-buffering

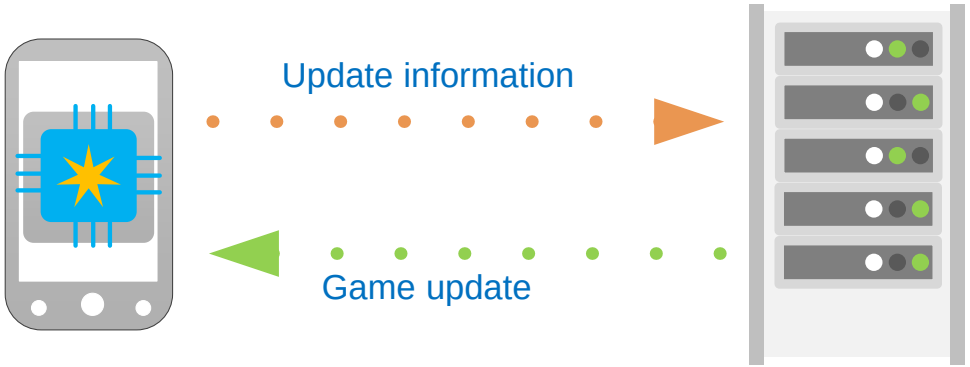


- ~1.5s to 1st video packet
- ~1.2s initial buffering

Online gaming – How real time is it?

- Live video is far a way from real-time today. How is the situation with online gaming?
- How does online gaming work today?
  - How ‘real time’, means how often information with the host server is exchanged?
  - Where the game is rendered, on the phone or on the server?

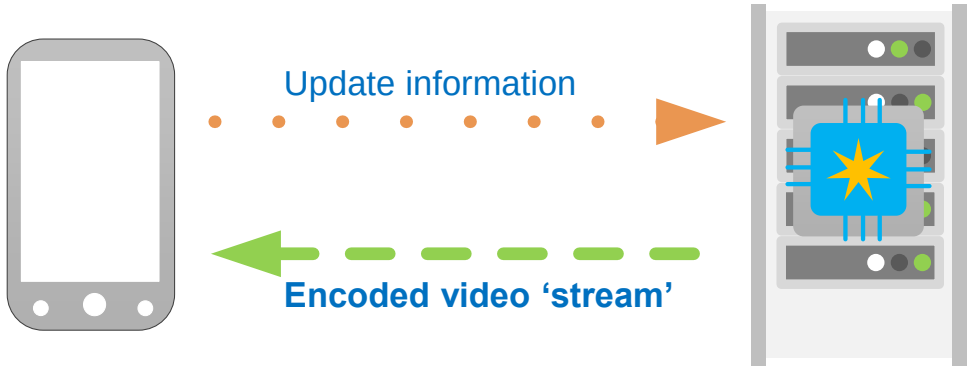
- Typical case
  - Device (player) sends position update
  - Server sends back game update
  - Device renders the graphics
  - → Low traffic in Up- and Downlink



Online gaming – How real time is it?

- Live video is far away from real-time today. How is the situation with online gaming?
- How does online gaming work today?
  - How ‘real time’, means how often information with the host server is exchanged?
  - Where the game is rendered, on the phone or on the server?

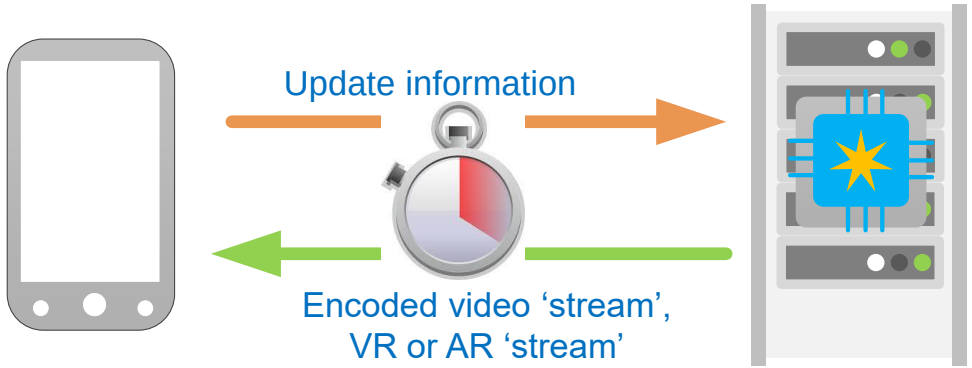
- Cloud gaming
  - Device (player) sends position update
  - Server renders the graphics
  - Server sends back graphics ‘as video’
  - → Considerable traffic in Downlink
  - → Platform independent



Online gaming – How real time is it?

- Live video is far a way from real-time today. How is the situation with online gaming?
- How does online gaming work today?
  - How ‘real time’, means how often information with the host server is exchanged?
  - Where the game is rendered, on the phone or on the server?

- Cloud gaming → **The Future**
  - Ultra short latency
  - Quasi continuous update information
  - Real-time interaction



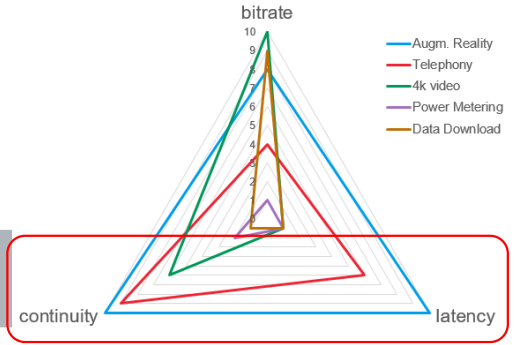
From ‘always online’ towards ‘seamless connectivity’

- Today’s mobile applications are still far away from being real-time interactive
- How to come to real-time interactivity?
  - Intelligence moved and is moving from network to device and server
  - Applications are adaptive to changing network conditions (e.g. variable bitrate for video streaming)

⇨ Applications will become <u>much more adaptive by ML / AI</u> (…no matter of 5G or not)

⇨ But: ML / AI will neither bring network latency down nor can avoid interruptions

⇨ Key for real-time interactive is: <u>Very short latency and seamless transmission of the underlying network</u>



How to measure QoE for tons of new apps and use cases?
How applications work behind the scenes?

- Most of today's media applications are based on existing (pre-compiled) libraries and frameworks
  - Toolboxes for video streaming applications, libraries for online gaming

⇨ Applications are using the same underlying techniques

⇨ The requirements/expectation on the individual use case are different

- This will be the same for VR/AT and similar
  - …many realizations will base on the same underlying libraries and may produce similar load-patterns in the network

⇨ There will be tons of individual use cases and Apps using e.g. AR/VR

⇨ How to measure QoE for all of them?

How to measure QoE for tons of new apps and use cases?

- An Application for ‘VR retail shopping’ and a ‘VR ego shooter’ may use the same techniques
- The user’s expectation and its experience (‘How tolerant I am in this particular case?’) is very different

- Today’s approach: Building an integrative objective model based on subjective tests
  - This is the today’s ‘integrative MOS approach’ as e.g. in ITU P.1203
  - It works for a limited set of applications (video streaming, posting a photo, browsing…)

⇨ Along with 5G we will face a vast number and more different applications and they will change rapidly
There is no way to build individual integrative models for each of it

Potential way out: **Change of Concept!**

An QoE model could be ‘a formula model’ based on a limited, identical set of QoS or technical KPIs.

How to measure QoE for tons of new apps and use cases?

- Let's go back to the integral QoE concept and its dimensions

Connectivity will be 'always' given, how to measure very rare 'fails'?

Instead chasing individual 'failed access', a **new concept** is a 'network stability score' ('probability to fail')

- Accessibility and sustainability
  - Do I have access to the 'service' at all and is it 'technically' kept?
- Waiting time for 'action' (task being started and/or completed)
  - How long the access takes (e.g. Call Setup Time, Video Access Time)
- How is the quality / experience during active use
  - How is 'quality' (e.g. video quality)
  - How is interaction, 'fluentness', response time,…

QoE of a use case or application

How to measure QoE for tons of new apps and use cases?

- Let's go back to the integral QoE concept and its dimensions

Access time will become very short.

Only minor influence on QoE in the future

- Accessibility and sustainability
  - Do I have access to the 'service' at all and is it 'technically' kept?
- Waiting time for 'action' (task being started and/or completed)
  - How long the access takes (e.g. Call Setup Time, Video Access Time)
- How is the quality / experience during active use
  - How is 'quality' (e.g. video quality)
  - How is interaction, 'fluentness', response time,…

QoE of a use case or application

How to measure QoE for tons of new apps and use cases?

- Let's go back to the integral QoE concept and its dimensions

- Accessibility and sustainability
  - Do I have access to the 'service' at all and is it 'technically' kept?
- Waiting time for 'action' (task being started and/or completed)
  - How long the access takes (e.g. Call Setup Time, Video Access Time)
- How is the quality / experience during active use
  - How is 'quality' (e.g. video quality)
  - How is interaction, 'fluentness', response time,…

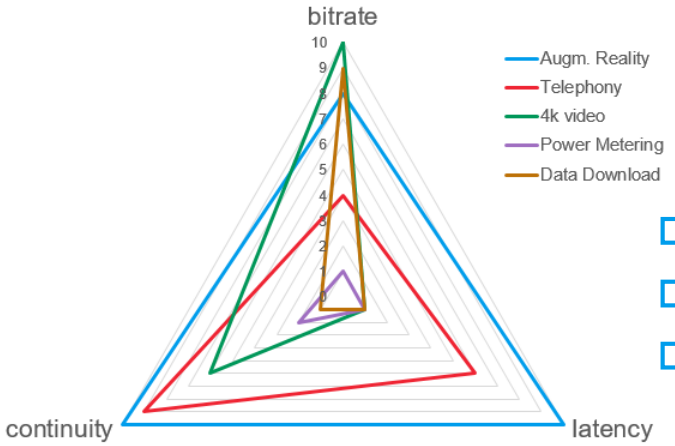
QoE of a use case or application

This is the key and the challenge at the same time!

Individual QoE models based on key QoS and technical parameters

What are the key parameters?

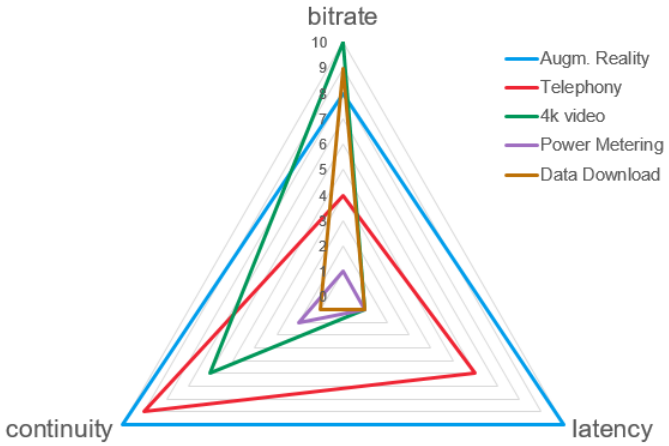
How to measure QoE for tons of new apps and use cases?
What are the key parameters?



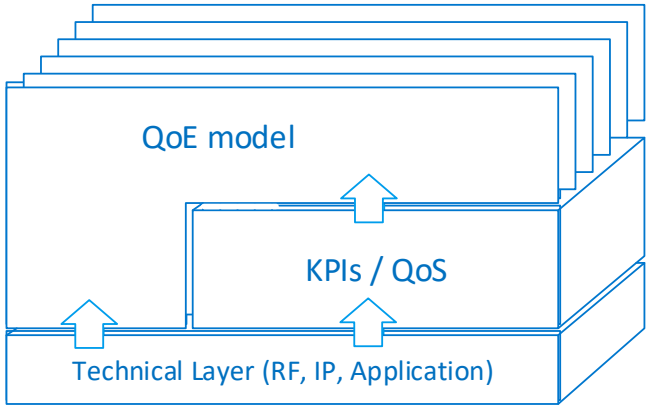
Good performance and QoE for an application and use case requires

- A minimum amount of data rate (transport capacity)
- A maximum response time (transport latency)
- A maximum length of interruptions or undercuts (transport continuity)

How to measure QoE for tons of new apps and use cases?



*There will not be **the** AR use case and **the** AR QoE model*



Individual applications and use cases have individual requirements

…and require individual QoE models…

…probably based on the same underlying measurement techniques

How to measure QoE for tons of new apps and use cases?
There are still some challenges on low layer…

- How to measure and to interpret ‘latency’ and ‘continuity’? It looks simple but it isn’t!
  - Measurement tools (smart phone devices!) must measure down to <1ms latencies with an accuracy of about 100µs. Today we are talking about latencies >>10ms…
  - How to define ‘continuity’? No interruptions at all? At which integration interval of packets? How to deal if just undershooting a target rate?
  - Use of archetype, generic test cases, individual apps or just plain technical measurements?

⇨ There are new measurement concepts and KPI definitions needed

⇨ ML and AI will help in realization of these new concepts but:

We have to feed this beast with the right cookies!

…accurate extraction of the right key parameters is key!

Summary

- Many of today's popular applications and use cases will continue under 5G
- New applications will become real-time interactive

⇨ Increasing bitrate and related KPIs are not sufficient anymore

- 'Service' accessibility and sustainability will move to 'always available', access time will be very short

⇨ Less influence on QoE of an application

Focus will move to quality in a given, running (active) use case

- Interactivity and continuity of transport become crucial for real-time applications under 5G

⇨ KPIs and QoS parameters for interactivity and continuity have to be developed and accepted

Evolving and new applications and use cases will require new QoE models and new QoE concepts


Thank you!
…and keep being real and interactive
ROHDE&SCHWARZ
Mobile Network Testing