

Network QoE Benchmarking

Luis Prado



Agenda

- 01** Intro to QoE benchmarking methods
- 02** Testing telemeeting and social media services QoE and QoS using native Android OTT applications
- 03** XR/VR emulated one way traffic testing
- 04** ITU-T SG12 Y.Testbed: Framework showcase and Example QoE evaluation

Agenda

- 01** Intro to QoE benchmarking methods
- 02 Testing telemeeting and social media services QoE and QoS using native Android OTT applications
- 03 XR/VR emulated one way traffic testing
- 04 Test Bed Framework for Mobile Application QoS and QoE Evaluation (ITU-T SG12 Y.Testped)

Challenges

MONITORING

- How to keep mobile subscribers satisfied with the service they are paying for?
- How the issues in QoE can be identified before customers do?
- How to reduce mean time to repair, when the issues exist?



Why Network Benchmarking?

- ✓ Know where you excel
- ✓ Know where you could improve
- ✓ Close gaps on competition
- ✓ Help focus engineering on specific problems and issues
- ✓ **Use data for marketing**
- ✓ Keep customers, improve Quality of Experience (QoE)
- ✓ To win Mobile Experience Awards
- ✓ National SLA targets

1. Voice Calls
 2. Latency
 3. Data Throughput (DL & UL)
 4. OTT applications (WhatsApp, Facebook, YouTube, etc.)
- Coverage – Accessibility – Retainability**



Network Benchmarking Compares Experience Quality Across Networks



**Compare customer
QoE across mobile
networks**



**Perform network
benchmarking
campaign**



**Measure service
coverage, accessibility,
retainability and integrity**



**Summarize results to
reports,
NPS 3.0 score and
interactive dashboards**

Challenges when Comparing networks QoE using different device models

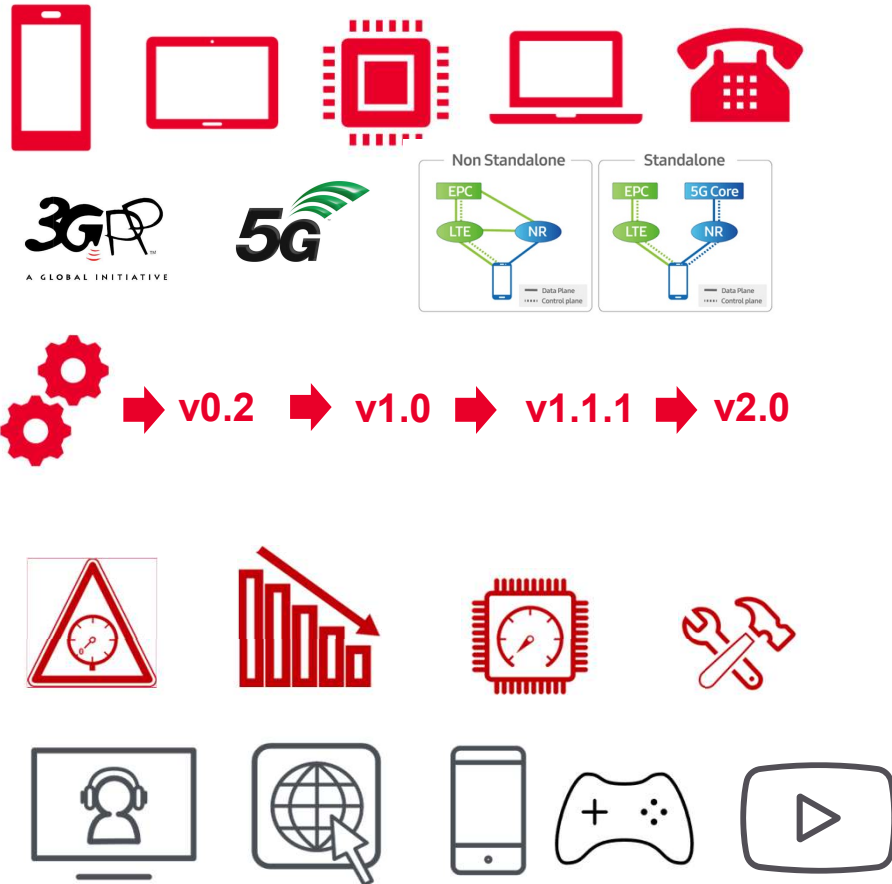
Constant new device announcements

Constant network equipment vendor feature updates

Flood of s/w and h/w version updates for devices, network equipment and configurations/settings

Performance issues seen when used actively in production network

Services with OTT apps



Active Testing and Air Interface Full Stack Logging



- One-way/two-way delay, loss, jitter
- ICMP, UDP, TCP ping
- UDP and TCP throughput
- COS qualification

Data Performance



- Web based app availability
- Cloud access assurance
- Business apps: Office365, Citrix etc.

Cloud/SaaS Availability



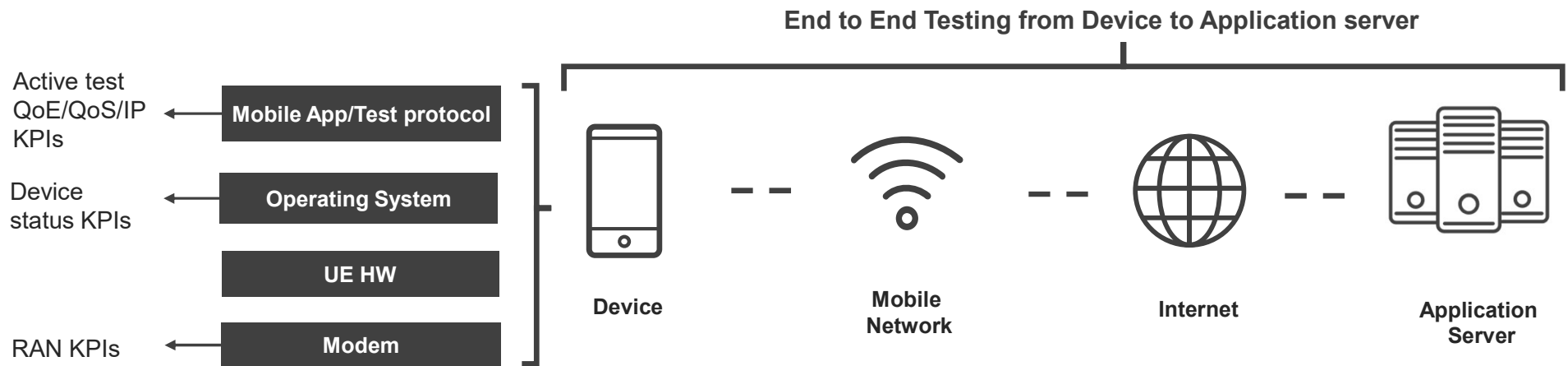
- VoIP multiple codecs
- Teams, Zoom, Skype
- Video, Netflix, YouTube etc.

QoS verification with Test Protocol/ Synthetic traffic



- Accessibility, retainability
- Voice quality, video quality

QoE verification with Real Mobile Apps



Measure Service Coverage, Accessibility, Retainability and Integrity for QoE



Social media and messaging

Activity success rate
Service access time
Transfer success rate



Telephony

Success ratio
Drop ratio
MOS
Setup time
90th percentile of call setup time



Data testing

Transfer success ratio DL and UL
Throughput
Success ratio
10th percentile of low throughput
90th percentile of high throughput



Video Streaming

Access time
Success ratio
Video quality
MOS
10th percentile MOS



Browsing

Success ratio
Average duration
Activity duration



E-conferencing and e-gaming

Interactivity
Latency
Packet loss
Jitter

Importance of OTT Apps

- Users prefer to use native, dedicated apps to access content instead of browsing
- 2.6 Million apps in Google Play store in June 2022
- Apps are continuously changing, update releases monthly or even weekly



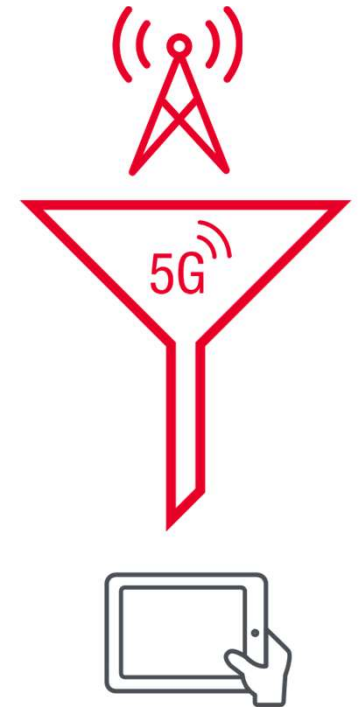
Adaptive Applications

- Many applications are adaptive
- Adaptive apps try to cope network speed and connection quality changes with their dynamic behavior
- Advanced native apps like YouTube player or Microsoft Teams are performing many advanced and dynamic tricks like bandwidth savings, optimizing packet size and changing codecs to adopt to data pipe



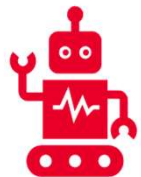
Network Capacity Shaping

- Operators often use traffic profiling and network capacity shaping (scaling) for each user, based on the applications that each consumer is using
- Operators can limit the bandwidth based on application type to save network capacity to secure QoS for other users too
- It works so that the cellular network detects the app in use, then tunes and limits the data pipe accordingly



What Value Does Keysight Nemo ATA Provide?

- Enables test automation for end user applications which have been traditionally difficult to control and automate
- Supports also apps without API
- Measures as KPIs what user can see on the screen, those can then be correlated with L1-L3 KPIs
- Scripts are designed to be device universal, across wide suite of devices, as long as the script supports the application version



ATA: Screen Control and Monitoring
Real customer experience



Screen	Real Application	Adaptation to NW	NW traffic shaping	Independent QoS for DL and UL	Asymmetric DL/UL traffic	Adaptation to App updates
Always	Always	Audio codec and audio frame adapting to available bandwidth	QoS is impacted by network traffic	OTT Voice is independent for uplink and downlink	OTT apps have asymmetric traffic profile in UL and DL	App behavior may change over new app versions

Agenda

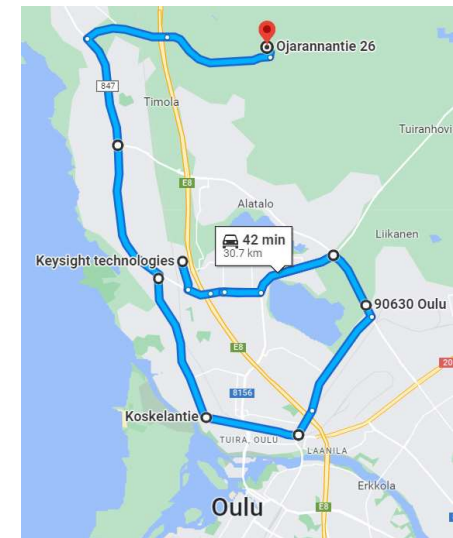
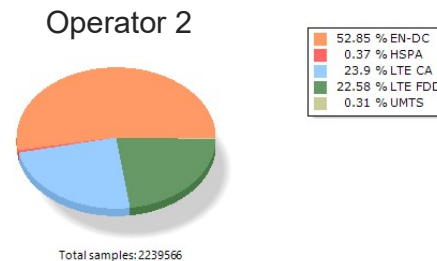
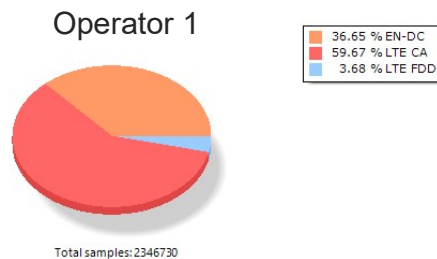
- 01 Intro to QoE benchmarking methods
- 02** Testing telemeeting and social media services QoE and QoS using native Android OTT applications
- 03 XR/VR emulated one way traffic testing
- 04 ITU-T SG12 Y.Testbed: Framework showcase and Example QoE evaluation

QoS vs. QoE

Throughput comparison

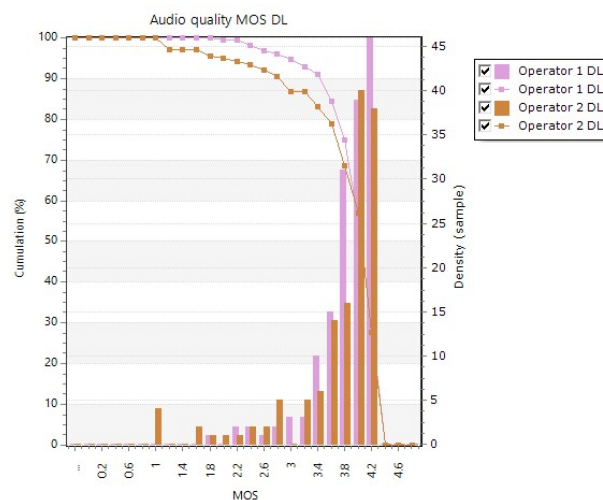
- Keysight Nemo Outdoor SW with Application Test Automation option
- Tests are run on a Samsung SM-S906B device (S22+ 5G)
- DL bulk data transfer + MS Teams audio quality (POLQA v3)
- 2 Finnish operators

	Operator 1	Operator 2	Comparison
Average DL Application Throughput	68.9 Mbps	109.4 Mbps	+59%
Max DL Application Throughput	221 Mbps	792 Mbps	+258%
Average MS Teams Start Delay	0.65 s	0.45 s	-31%



QoS vs. QoE

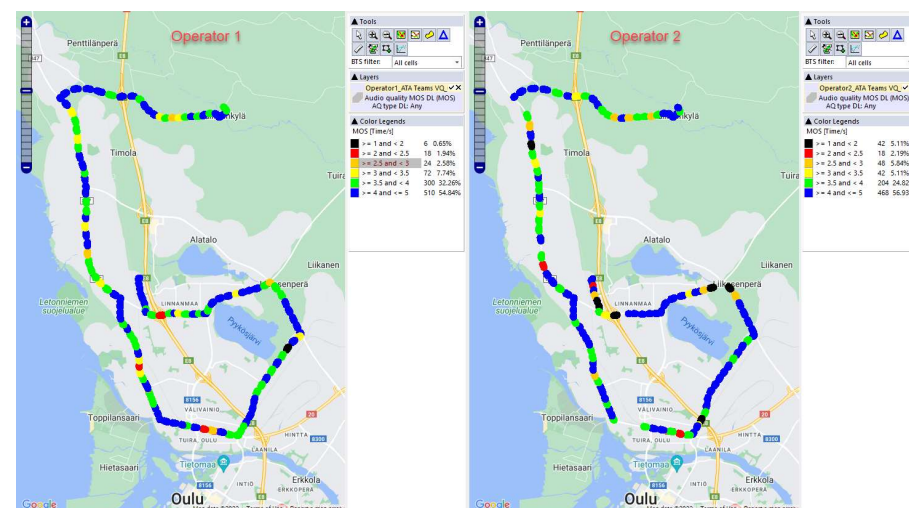
MS Teams Audio Quality



	Operator 1	Operator 2	Comparison
Average DL Application Throughput	68.9 Mbps	109.4 Mbps	+59%
Max DL Application Throughput	221 Mbps	792 Mbps	+258%
Average MS Teams Start Delay	0.65 s	0.45 s	-31%
Average DL MOS POLQA v3	3.9	3.7	-5%
Average UL MOS POLQA v3	4.1	3.8	-8%
Conference Start Success Rate	100%	100%	0%

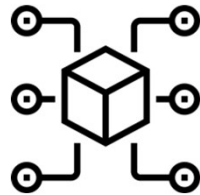
User Experience tells a different story!

- Operator 1: better audio quality (QoE)
- Operator 2: faster network (QoS)



QoS vs. QoE

QoS



- Close to the network
- Typically not computationally demanding
- Defined on specific application settings
- Objective, reasonably well defined

QoE



- Close to the end user
- Can be computationally demanding
- Defined on classes of applications
- Subjective and hard to define and model

QoE Benchmarking measurements made in Europe

Using Keysight Nemo Benchmarking Solutions



30 000 000
SAMPLES



Samsung S22
DEVICES



European
measurements



WhatsApp
MS Teams
YouTube
Fast.com
APPLICATIONS



Nemo Outdoor SW +
Backpack Pro VQ MOS



April 2023

DATA COLLECTION
TIME PERIOD

Nemo Backpack Pro



Free OTA reception for all 5G devices

- Unblocked RF reception and transmission



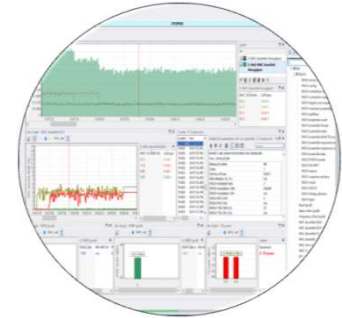
Ease to use and lightweight

- Ergonomic structure
- Hot swap batteries
- Wireless access
- Carbon fiber structure



Measure everything simultaneously

- No need to visit the same location twice
- Up to 18 UE



Nemo Outdoor SW

- The one tool for all needs
- 1st to market with new terminals and technologies
- Standard test methodology
- OTT App testing
- Reliable LAN connected system
- Modular and Scalable device management

Overview

- Free OTA RF reception to all measurement phones
- No obstacles blocking the RF OTA reception of transmission
- Enables scientific QoE measurements
- High-end carbon fiber design
- Lightweight
- Up to 18 UE
- Hot swap batteries



Features

- Fabrics has zippers for fast and easy access for all UE's
- Waterproof and durable fabrics
- Innovative construction
- Reliable LAN connected system without PC USB resource issues



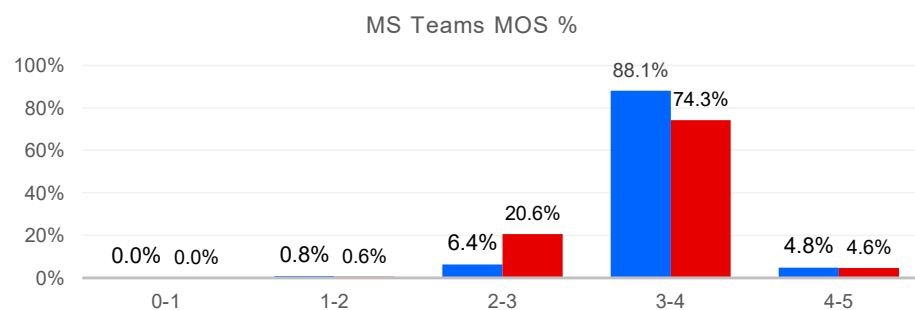
MS Teams Application

Call Setup Time and MOS

Latest available MS Teams application in Google Play was used.

Operator 1 comes out on top, achieving better results than Operator 2. **Concluding 92,9% of samples scoring between 3-5 MOS versus 78,9% at Operator 2 mobile network.**

Call Setup Time in the Operator 2's mobile network is slightly better between than in Operator 1, and in both cases VoLTE is performing much better than MS Teams.



Application	Operator 1	Operator 2
MS Teams Success Ratio %	100	100
MS Teams Call Setup Time sec (Avg)	8,78	8,42
MS Teams Call Setup Time sec (Min-Max)	8,47 – 9,42	8,12 – 9,39
MS Teams Audio Quality MOS (Avg)	3.61	3.43
MS Teams Audio Quality MOS (Min-Max)	1.03 – 4.11	1.67 – 4.15

MS Teams application doesn't provide a technical mechanism to secure RAN neither transport and CORE network resources to secure a success call, but **voice quality was a bit better in MS Teams than VoLTE calls even with that limitation.**

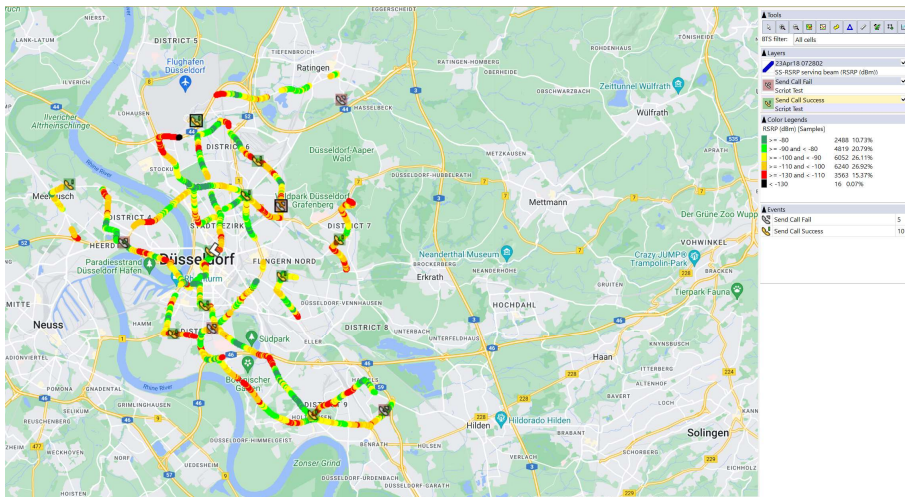
- Send Call Success and Send Call Fail over NR SS-RSRP and LTE RSRP. Packet technology can be seen below. Some MS Teams Calls are failing due to bad NR SS-RSRP/LTE RSRP values.



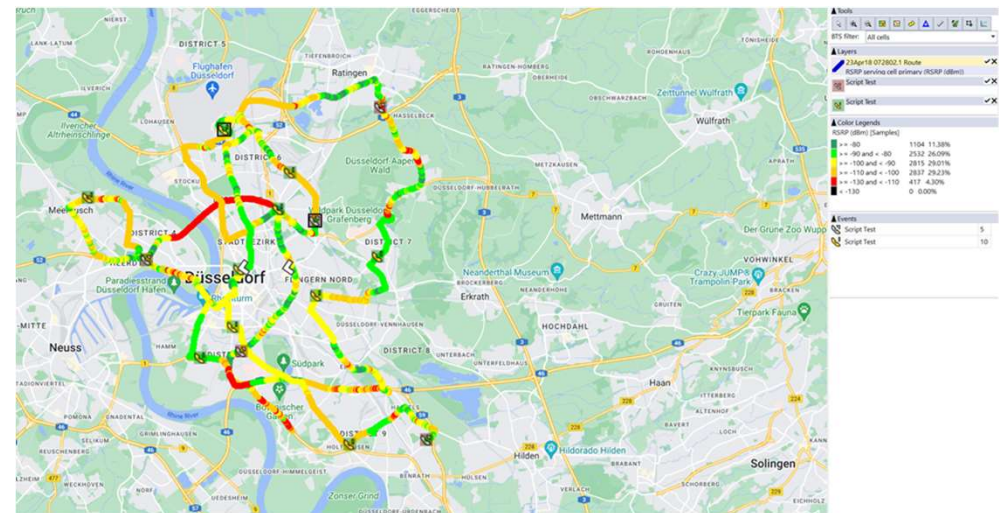
Example of Troubleshooting Microsoft Teams

- Send Call Success and Send Call Fail over NR SS-RSRP and LTE RSRP. Packet technology can be seen below. Some MS Teams Calls are failing due to bad NR SS-RSRP/LTE RSRP values.

NR SS-RSRP:

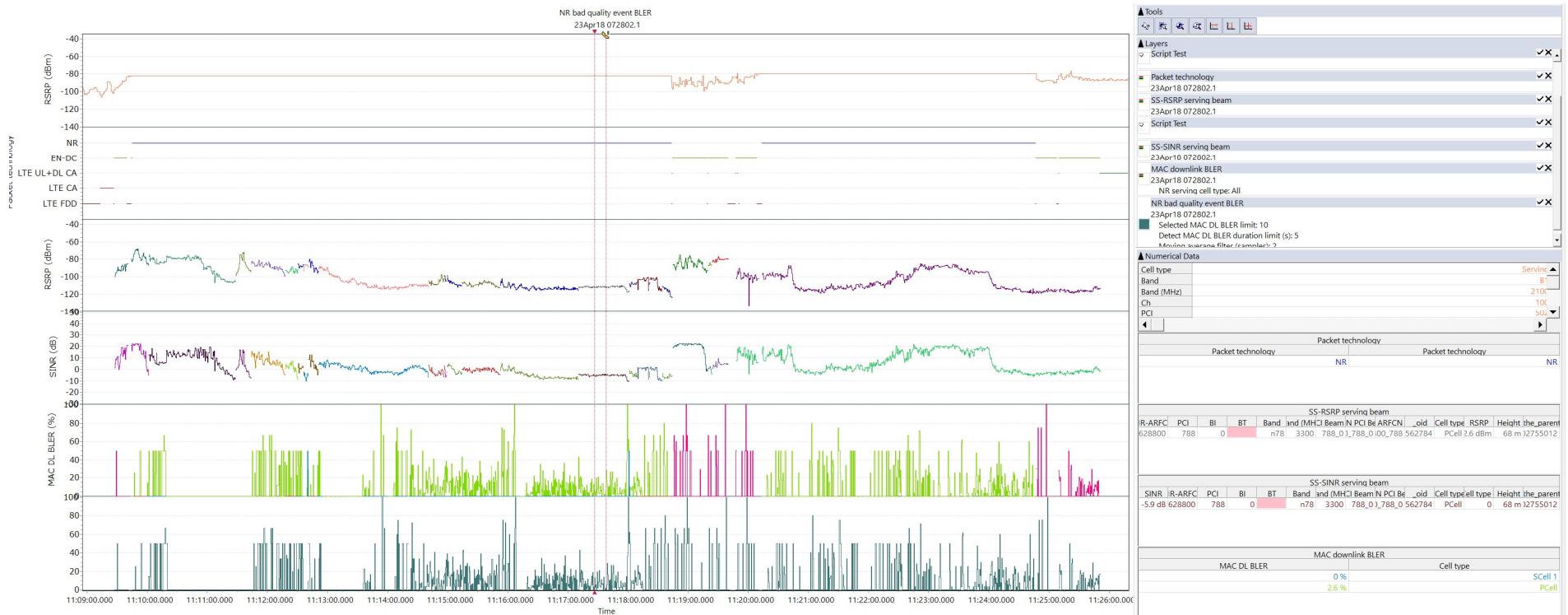


LTE RSRP:



Example of Troubleshooting Microsoft Teams

- Specific Call Analysis in NR, high MAC BLER%, low SINR with Analyze 5G bad Analysis event being triggered at the time of the MS Teams Call Send that has Fail Status



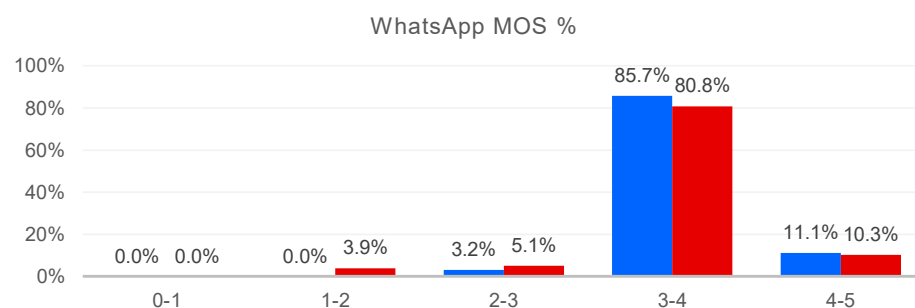
WhatsApp Application

Call Setup Time and MOS

Latest available WhatsApp application in Google Play was used.

Operator 1 comes out on top, achieving better results than Operator 2. **Concluding 96.8% of samples scoring between 3-5 MOS versus 91.1% in Operator 2 mobile network.**

Call Setup Time in the Operator 2's mobile network is slightly better between than in Operator 1, and in both cases VoLTE is performing much better than WhatsApp.






Application	Operator 1	Operator 2
WhatsApp Success Ratio %	100	100
WhatsApp Call Setup Time sec (Avg)	3,74	3,42
WhatsApp Call Setup Time sec (Min-Max)	3,49 – 4,01	3,32 – 4,10
WhatsApp Audio Quality MOS (Avg)	3,73	3,67
WhatsApp Audio Quality MOS (Min-Max)	2,86 – 4,08	1,48 – 4,14

WhatsApp application doesn't provide a technical mechanism to secure RAN neither transport and CORE network resources to secure a success call, but **voice quality was a bit better in WhatsApp than VoLTE calls even with that limitation.**

Benchmarking results

MNO = Mobile Network Operator

		VoLTE			
VoLTE		MNO_1	MNO_2	MNO_3	MNO_4
	CSSR%	100.00%	100.00%	100.00%	100.00%
	Call Continuity SR%	100.00%	98.96%	100.00%	100.00%
		WhatsApp			
		MNO_1	MNO_2	MNO_3	MNO_4
	CSSR%	100.00%	100.00%	100.00%	100.00%
	Call Continuity SR%	100.00%	100.00%	100.00%	100.00%
		Microsoft Teams Calls			
		MNO_1	MNO_2	MNO_3	MNO_4
	CSSR%	100.00%	33.33%	100.00%	90.00%
	Call Continuity SR%	100.00%	100.00%	100.00%	80.00%
		Microsoft Teams Meetings			
		MNO_1	MNO_2	MNO_3	MNO_4
	CSSR%	100.00%	100.00%	100.00%	100.00%
	Call Continuity SR%	100.00%	98.21%	90.91%	100.00%

MNO 1 - Customer experience summary

Application	KPI	Bad	Poor	Fair	Good	Excellent				
VoLTE	Success Rate (%)	80	80	85	85	90	90	95	95	100
										100
	Call Setup Time (sec)	> 10	10	7	7	4	4	2	2	0
								2.47		
	MOS POLQA v3	< 2,5	2,5	3	3	3,5	3,5	4,2	4,2	5
								3.60		
WhatsApp	Success Rate (%)	80	80	85	85	90	90	95	95	100
										100
	Call Setup Time (sec)	> 10	10	7	7	4	4	2	2	0
								3.42		
	MOS POLQA v3	< 2,5	2,5	3	3	3,5	3,5	4,2	4,2	5
								3.67		
MS Teams	Success Rate (%)	80	80	85	85	90	90	95	95	100
								90.53		100
	Call Setup Time (sec)	> 10	10	7	7	4	4	2	2	0
				8.42						
	MOS POLQA v3	< 2,5	2,5	3	3	3,5	3,5	4,2	4,2	5
								3.43		
YouTube	Success Rate (%)	80	80	85	85	90	90	95	95	100
										100
	Start Time (sec)	> 10	10	5	5	3	3	1	< 1	
								2.47		
	Resolution (Pixel)	240	360	480	480	720	720	1.8		

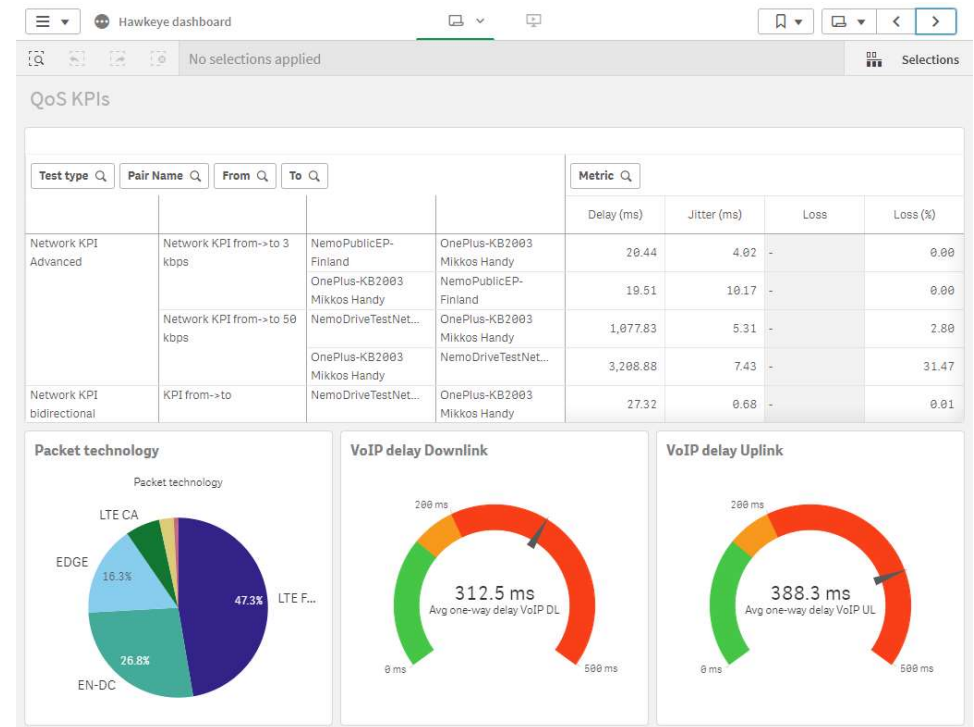
Agenda

- 01 Intro to QoE benchmarking methods
- 02 Testing telemeeting and social media services QoE and QoS using native Android OTT applications
- 03 XR/VR emulated one way traffic testing**
- 04 ITU-T SG12 Y.Testbed: Framework showcase and Example QoE evaluation

One-way Latency Measurements with Hawkeye

Monitor Critical SLA Metrics for New 5G Services

- Low latency is a key driver for 5G deployments
 - Smart factory
 - Vehicle to vehicle
 - Augmented reality
- Measure one-way latency (upstream downstream)
- Correlate within depth radio analytics
- Ensure you meet your SLA 24/7 for critical applications



Real traffic is one-way, QoS testing should be as well

- To overcome the limitations of TWAMP and other two-way test protocols, a one-way test protocol should be used that allows a different traffic pattern to be sent uplink and downlink, or to send a downlink-only or uplink-only synthetic traffic pattern.
- Also, the test protocol should allow one-way latency, packet error rate, and jitter to be measured for uplink and downlink transmission directions independently.

Test endpoint (UE)
Measures DL latency,
DL PER, DL jitter



Uplink traffic pattern send



Downlink traffic pattern send

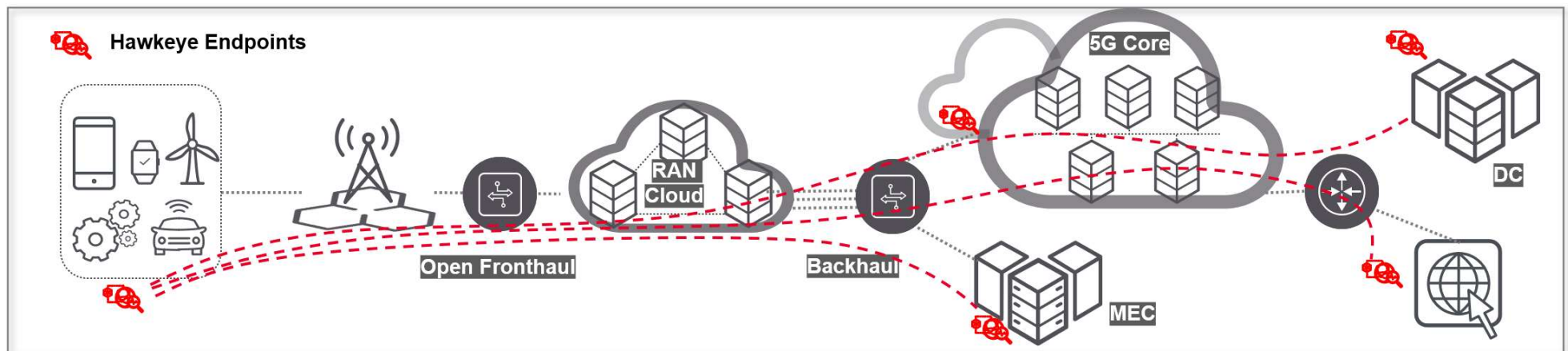


Test endpoint (server)
Measures UL latency, UL
PER, UL jitter



One-way latency measurements with Keysight Nemo Hawkeye

- Keysight's Nemo Hawkeye is an active testing solution that supports one-way QoS testing with configurable traffic patterns.
- Hawkeye has a library of off-the-shelf traffic profiles for many popular applications, including Netflix, Zoom meeting, XR real time video, VoIP with different audio codecs, video, and Microsoft Teams.
- Hawkeye includes automatic time synchronization between the test endpoints with millisecond-level accuracy



One-way latency measurements for Zoom e-conferencing

- In the time series graph, we are monitoring the uplink QoS with uplink latency and uplink packet error rate
- KPIs along with the key radio KPIs. The recommended QoS thresholds for Zoom are one-way latency less than 150 ms and packet error rate less than 2%.
- At the beginning of the time series data, latency is around 40-60 ms and PER is 0% for uplink. In other words, the mobile network is delivering sufficient
- QoS. Then the coverage and the quality of the mobile connection starts declining gradually. The uplink radio connection starts to reach its coverage limit, which can be seen in the uplink transmitted power hitting the maximum of the device.
- This shows up first in the uplink delay with occasional delay spikes higher than 150 ms and eventually also in the packet error rate.

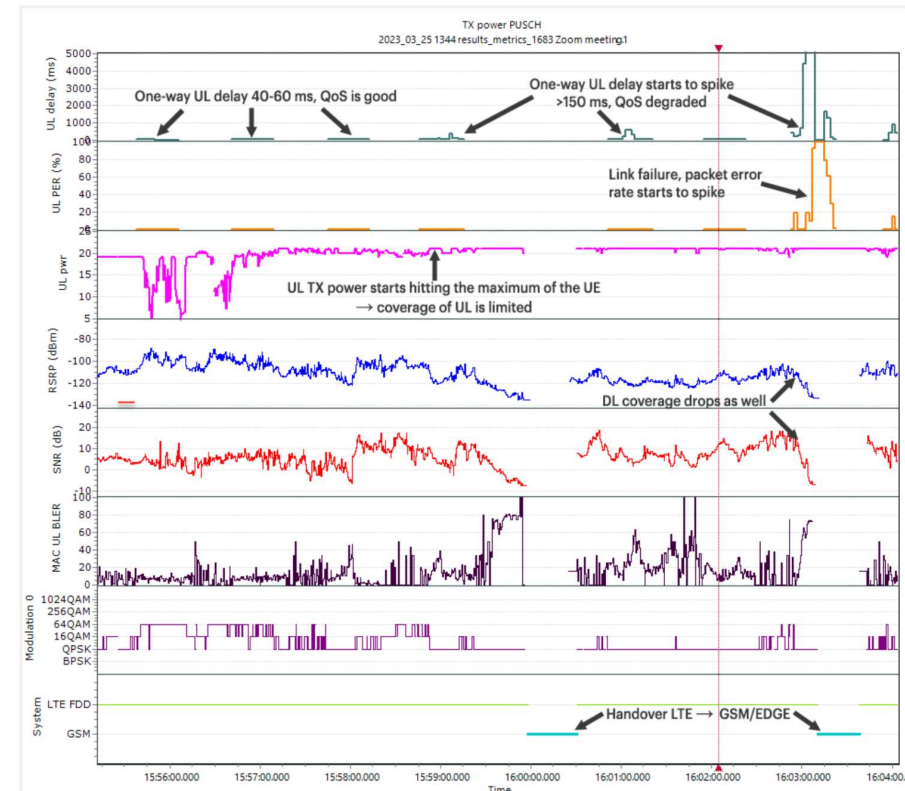


Figure: Time series view of Hawkeye Zoom meeting QoS test results over LTE network with Nemo Outdoor measured radio data

Agenda

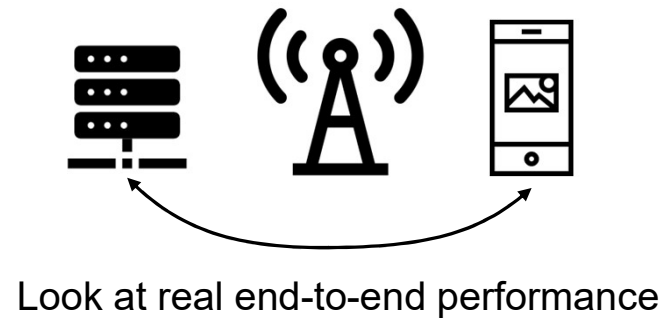
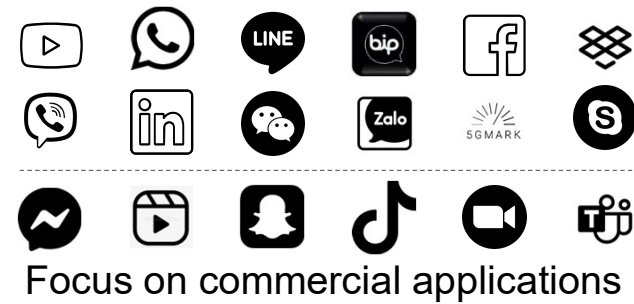
- 01 Intro to QoE benchmarking methods
- 02 Testing telemeeting and social media services QoE and QoS using native Android OTT applications
- 03 XR/VR emulated one way traffic testing
- 04** ITU-T SG12 Y.Testbed: Framework showcase and Example QoE evaluation

Y.Testbed

The Reason To Exist



Look coherently at all the 5G use cases

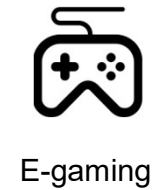


Test use cases in multiple NGMN scenarios

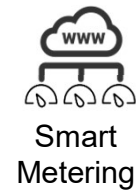
Slide 36

ML1 [@German Madueno] is there anything you want to add?
Mattia Lecci, 7/1/2022

The Process Illustrated



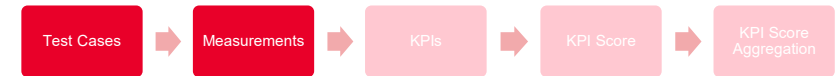
...



Measurement Process

Test Case: Content Streaming

Execute user flow in the App, while collecting Network and App KPIs



Measurement points

- Application starts
- Application closed
- Video requested
- Video starts
- Video stops
- Freeze start
- Freeze end
- Resolution/encoding change
- ...

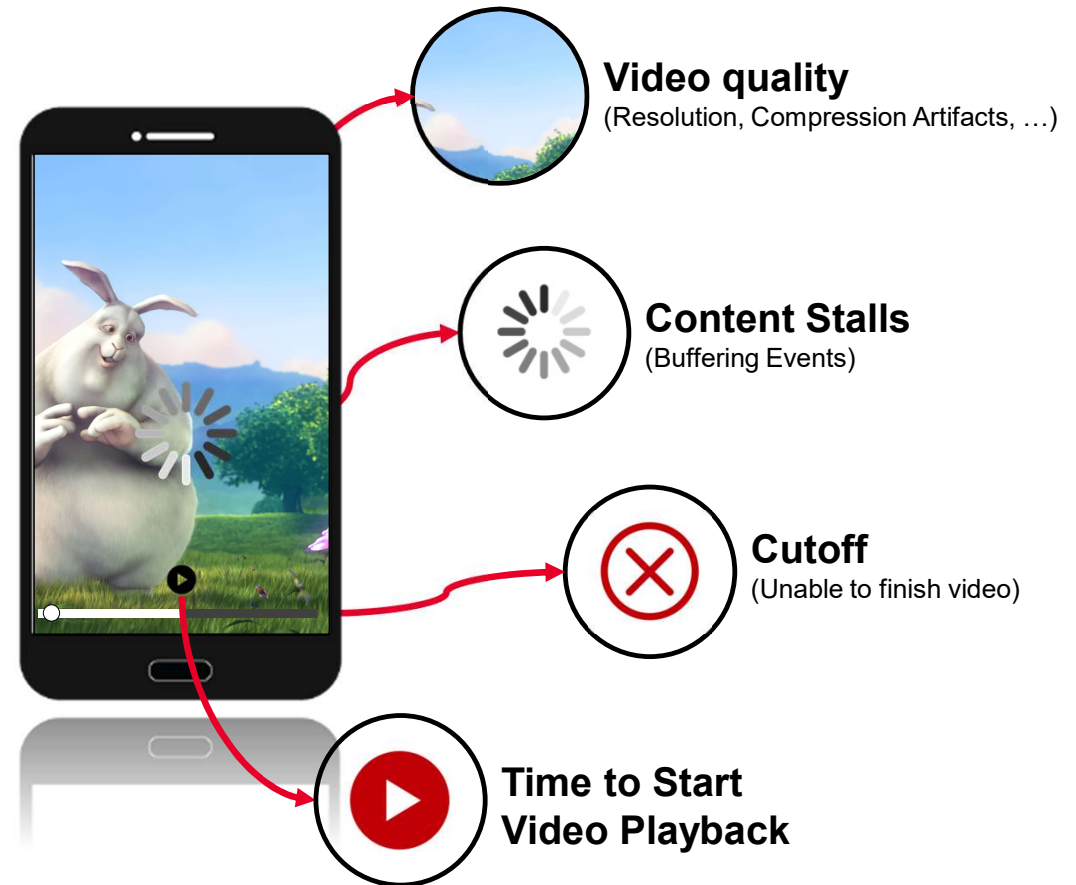
Measurements can be passive (external tools) or active (made by the application itself)

Key Performance Indicators

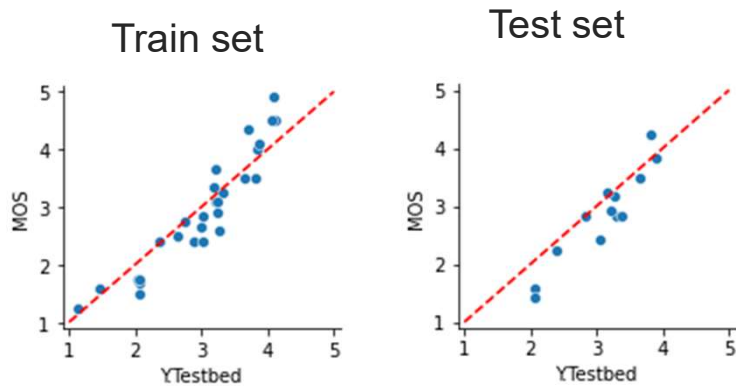
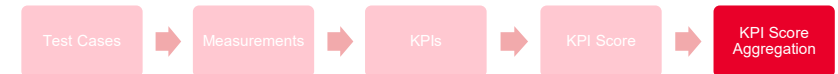
Test Case: Content Streaming

- Estimate the QoE based on the real application events.
- The KPIs are normalized and weighted according to human panel data

KPI	MOS: 1	MOS: 5
Time to load first frame (s)	9.82	0
Playback cutoff (bool)	1	0
Content stall ratio (%)	62.7%	0
Video vertical resolution (px)	119	1297



Performance



Y.Testbed

- ⚠ KPIs are not always straightforward to obtain
- ✓ The model operates directly on user-level features
- ✓ Good performance
- ✓ Good generalization performance

	Metric range	Train set	Test set
RMSE	[0, 4]	0.373	0.381
Pearson R	[-1, 1]	0.938	0.942
Spearman R	[-1, 1]	0.934	0.860



Conclusions

Y.Testbed Framework

QoS vs. QoE:

- Network metrics can be misleading
- Modern applications dynamically adapt to the network: hard to model
- User-level KPIs hardly change among applications of the same type

Y.Testbed:

- Single framework for 5G/6G services
- Solid academic and industrial background
- Adopted by NGMN
- Preliminary results are promising

Next steps:

- Contribute to Y.Testbed work item
- Extend content streaming use case
 - More comprehensive dataset
 - Larger human panel
- Include more use cases
 - Telemeeting
 - Gaming
 - VR/AR
 - ...



Thank you

Luis.prado@keysight.com