GPU Virtualization, 5G & MEC: Making Cloud XR a Reality

James Li

Principal, Orange Next Reality

(XR, HPC, Embedded AI)

Email: james.li@orange.com

John Benko

Principal, 5G/Emerging Wireless

Email: john.benko@orange.com

Orange Silicon Valley

Outline

Intro

What/Why are we doing on XR

Cloud XR architecture & requirements

Building a quick end-to-end Cloud XR POC

Latency & data-rate

Cellular Networks (LTE/5G/MEC/Slicing)

Tests & Conclusion

Orange Intro at a glance

www.orange.com

#51 in the 2017 global brand ranking

269 million
customers worldwide (30 June 2017)

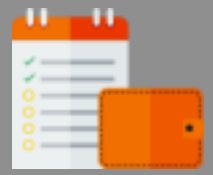
3.3 million
fibre customers

450,000 kms
of undersea cables (enough to go around the earth 10 times!)

29 million
Orange Money customers in 17 countries

705 million
Euros invested in research and innovation

6,844
patents in our R&D portfolio



4G
in 18 countries

+22 million
fans on Facebook

+3 million
followers on Twitter

+1.5 million
visits on orange.com each month

Orange Silicon Valley Intro

How we work

- Orange's Bay Area presence
- Innovation Center
- 52 people
- 20+ years in the Silicon Valley with strong R&D background
- Focused primarily on innovations 2-5 years beyond current roadmap
- Our job is to bring Silicon Valley innovation to 269 million customers in 29 countries.

www.orangesv.com

- Key Roles:

Innovate

Engage

Invest

Our network

Orange invests and contributes to the development of high performance global networks:

Orange ranked best network 2G/3G/4G in France by Arcep for the 6th consecutive year in 2016.

Orange is actively preparing for the arrival of 5G

- Orange to Launch 5G in 17 European cities in 2019
- Conduct end-to-end 5G trials in France with Ericsson

- 5G FWA trials in Romania with Samsung and Cisco
- Partner with Nokia and Kathrein for smart 4G/5G antenna design
- Orange Spain testing 5G technology in 7 cities in the country, and to deploy in 3-4 cities in 2019

Our Vision

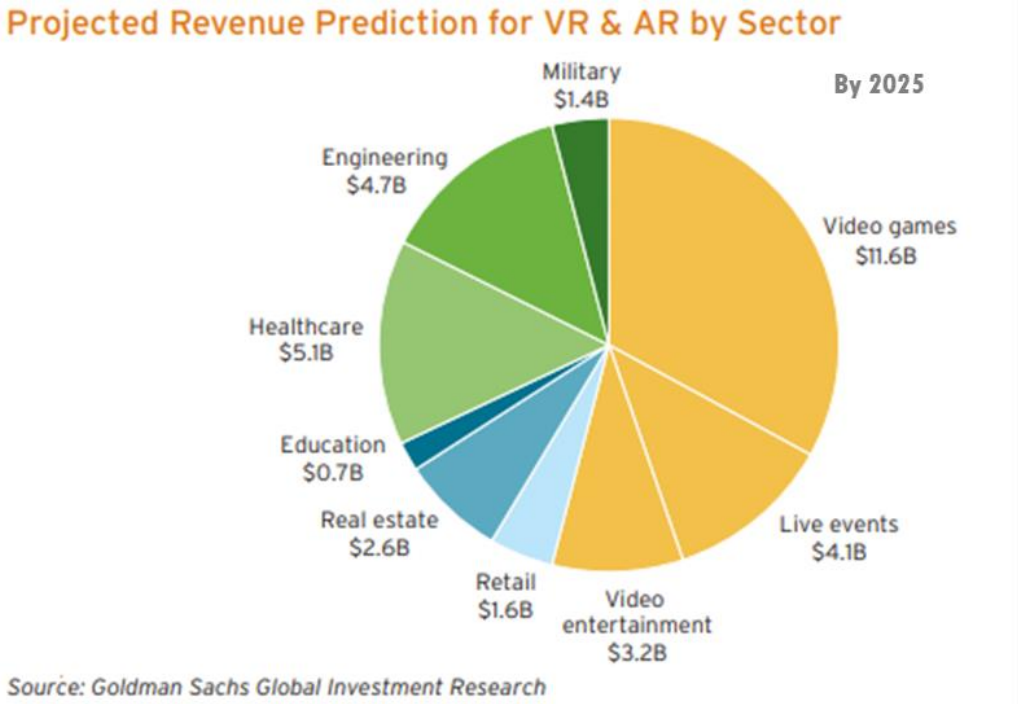
1. **Why Orange is doing XR (AR/VR)?**

   - it is the fourth wave of computing platform

   - it is one of the top 5G use cases

   - it is cross-industry enabler, colliding with 5G, HPC, AI, IoT, Blockchain

   - It will also impact different industry verticals beyond gaming & entertainment



Orange Next Reality

is created to explore XR business opportunities for impacted industries and to help transform enterprises

Our Vision

2. What XR experience do we expect to deliver to users?

High Quality Content & Graphics

Mixed Reality (XR)

Multi-User

Portable/Untethered

Seamless & Responsive

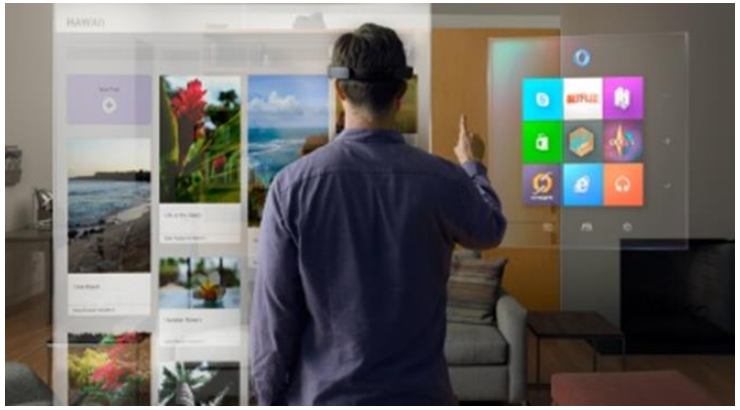


Image credit: Microsoft Hololens, Nvidia Holodeck

Our Vision

3. Current High-End XR Challenges

- No Solution for XR that delivers all:
  - Untethered experience
  - Support streaming to multiple users
  - Support for ultra-high resolutions (4K->8K)
  - While keeping it cost effective

5G is rolling out to address the demand, but we need to assess data-driven apps, content, services and computing & network architecture, with the follow guidelines:

- Demand on connectivity
- Service availability
- Commercial use cases
- User experience
- Scalability

5G Cloud XR

XR is a transformative technology which will revolutionize the consumption of content in both the consumer and enterprise sectors.

XR requires low latency, high data bandwidth, large storage and massive computing capabilities.

Telco will play an important role delivering 5G, Edge Computing and the Cloud.

Key Question:

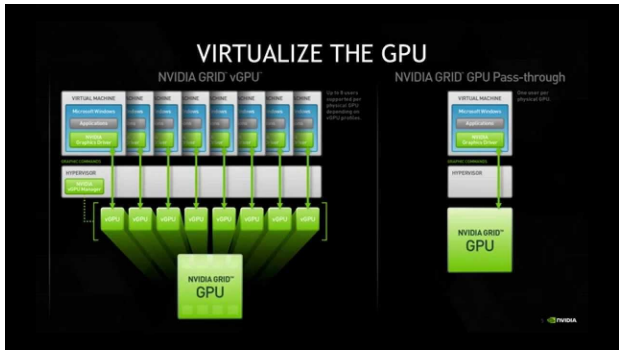
Can 5G, Edge Compute and the (GPU) Cloud make Cloud XR a reality…..?

Cloud XR is about Streaming

1. Streaming XR require 3 major components
   a) Server Requirements – efficient rendering, compression and streaming frames
   b) Network Requirements
      i. Low Latency – motion-to-photon, fast response from headset to avoid motion sickness
      ii. Bandwidth (based on quality, resolution, refresh-rate, compression, etc.)
   c) Client Requirements – lightweight rendering and de-compression (related to FOV and resolution)

Let's Build An End to End Minimum Viable Solution for Cloud XR

High-End Wireless XR Solution Architecture



Cloud

XR Content

Remote GPU

Telco Edge / Distributed

Edge Node

Edge Compute (Rendering/AI/ML for optimization)

5G

Consumer Edge & Devices

End User Computing

5G Network & Light Weight Compute

Building An End to End Link for Wireless Streaming XR

- Server
  - Rendering
  - Streaming
  - Encoding/Compressing

- ISP Network
  - Fixed Network(Fiber, DSL, Cable, etc.)
  - Wireless Core (LTE, 5G)

Simulated 5G

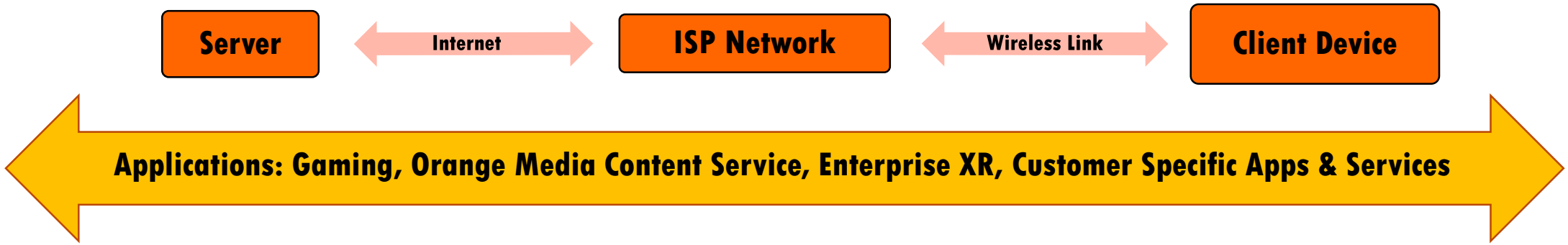
Orange 5G network (European Cities)

Telecom partners' 5G network/testbed/ equipment

- Client XR Device
  - Decoding/ uncompressing
  - Local Rendering

Orange Silicon Valley

Minimum Viable End-To-End Cloud XR POC @ Orange Silicon Valley



Quick GPU Virtualization Solution: Proxmox

10Gbps Link

AC/AD WiFi Router

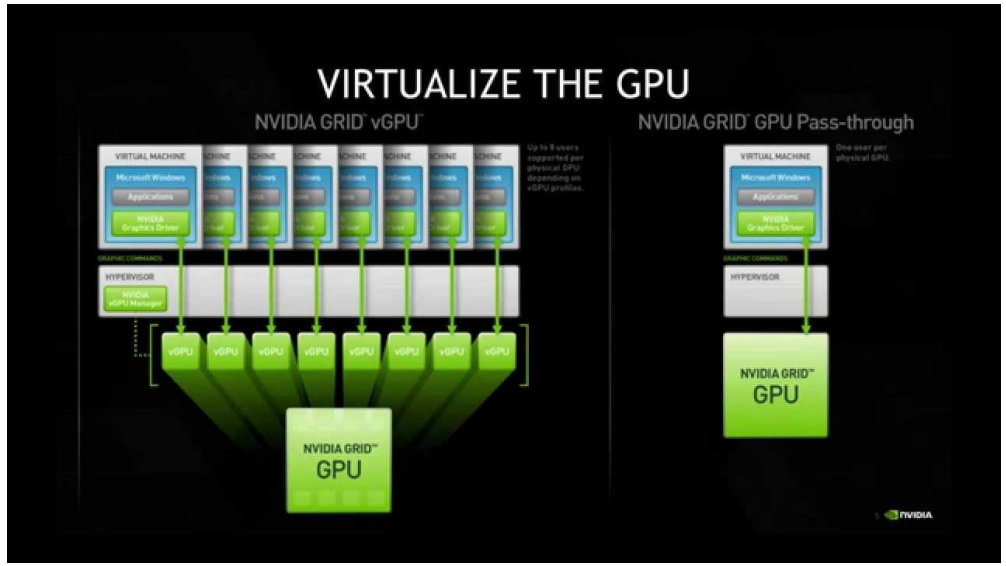
Simulated 5G 1+Gbps

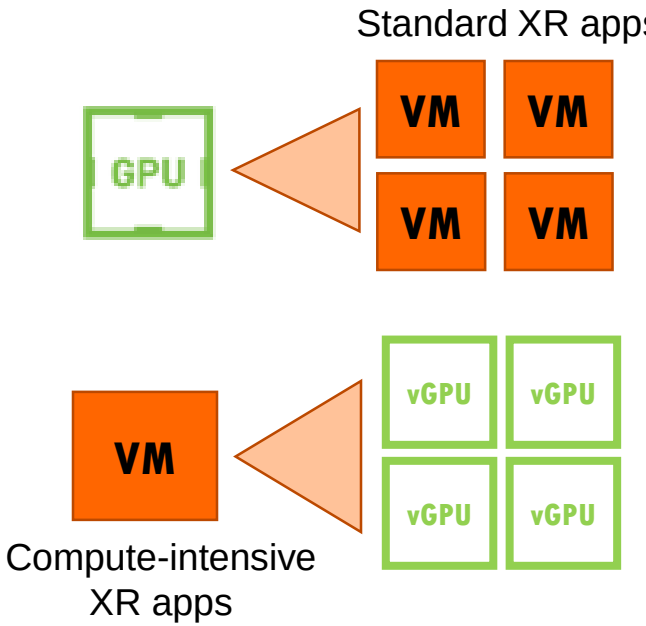
Wireless XR End Points

Mobile XR -> High-End XR

Orange Silicon Valley

Leveraging on the Power of GPU Virtualization





Advantage:
GPU Virtualization can be applied in cloud datacenter, **and at the edge**

Next-Gen RTX Server + GeForce Now

A great step-forward toward high quality, low-latency Cloud XR streaming

- High Service availability
- High Scalability / Multi-User
- Cinematic-quality graphics
- User experience
- Commercial use cases

Orange Silicon Valley

Latency + Data-rate (BW) Requirements

“Motion-to-Photon” Latency (round-trip end-to-end)

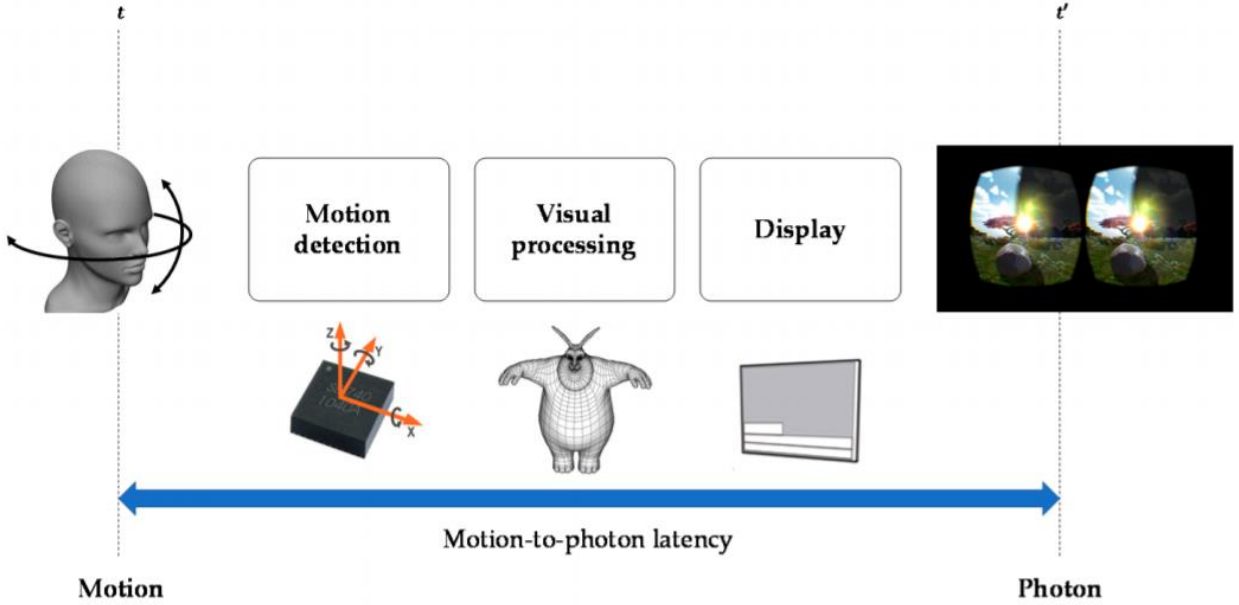
- < 20 msec to give user a realistic & comfortable experience

“Motion-to-Sound” Latency

- ~20 msec

Data-rate

- Highly Content/Display Dependent!
- Resolution – ex. SD, HD, 2k, 4k, 8k, …
- VR (ex. HD to HTC Vive ~ 6 Gbps)
- Compression (lossy – pixelization), Foveation
- Basic AR (heads-up display)
- XR



Orange Silicon Valley

Cellular Network Architecture & Mobile Edge Compute (MEC)

4G/LTE (Long Term Evolution)

Versions: (R8-R11) - LTE-A, LTE-A Pro, etc.

BW = 5 MHz – (4x20 MHz) CA, LAA

Data-rates 5-100+ Mbps

Latency = 25-50 ms, Jitter = 5 - 50 ms

5G

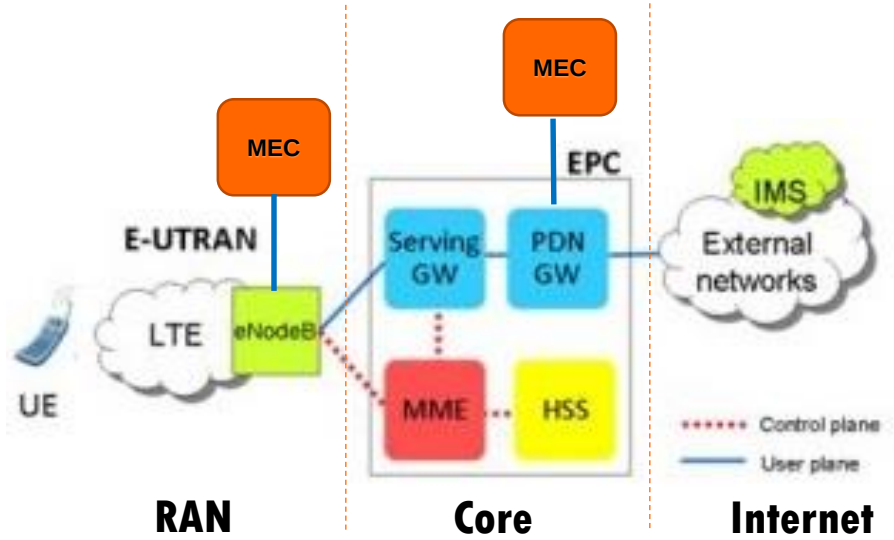
Versions: NSA / SA (Rel 15), Rel 16 (uRLLC)

BW : 5 - 800 MHz, <6 GHz, mmWave

Mobile Edge Computing (LTE, 5G)

i. Core (Mobile Core Network)

ii. RAN (Radio Access Network)



5G Applications and Requirements (ITU-2020)

Data rates

- 10 Gbps (Peak)
- 1 Gbps (Office)
- ~100 Mbps (DL), 50 Mbps (UL) – Dense Urban

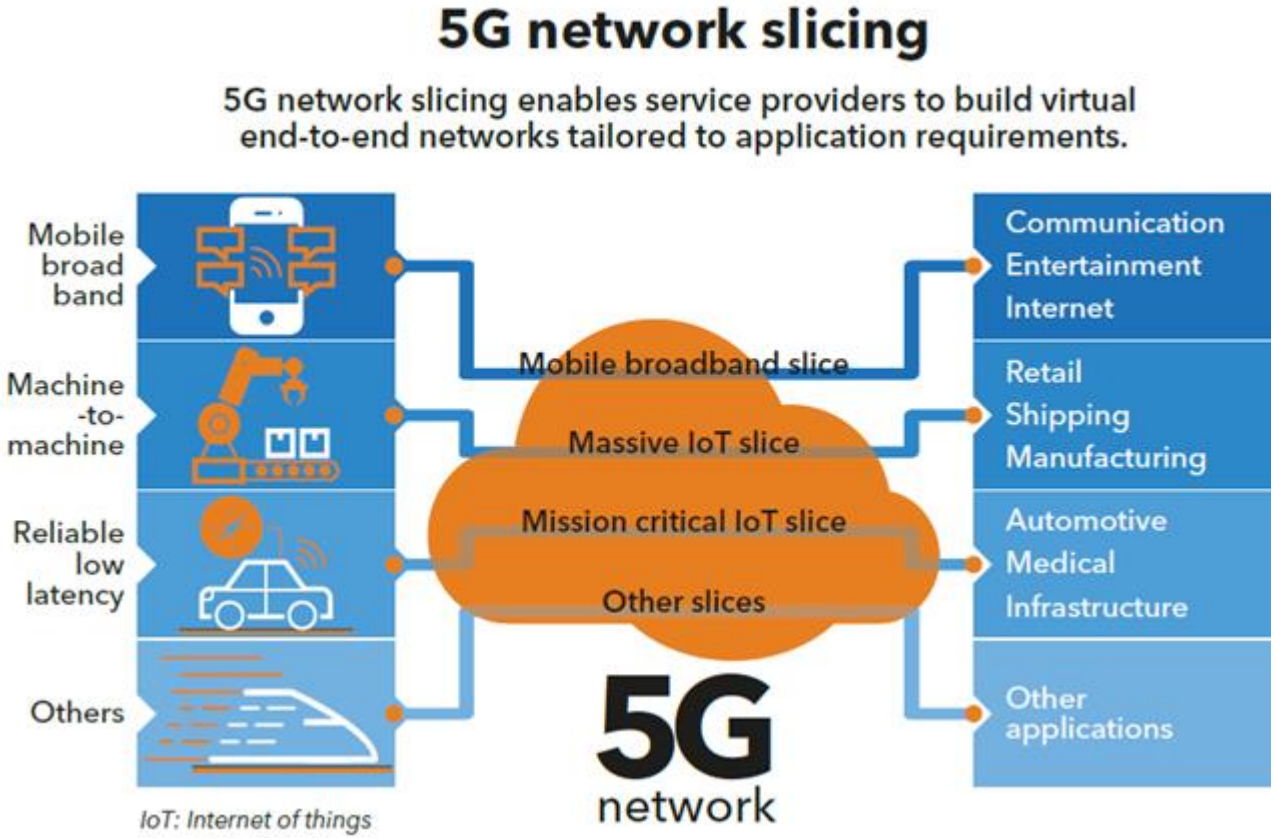
Latency

- 1-10 ms (E2E)

Connections

- ~ 1 million per square-km

5G Network Slicing (ITU)



XR Slices

- Gaming/entertainment
- Enterprise Use case (remote assistance)
- Remote Surgery (strict QoS/reliability requirements)

5G Network Slicing for VR (GSMA Requirements*)

Weak-Interactive VR slice

- Passive experience
- No direct interactivity with virtual environment
- Users can select observation point & location

	Entry-Level VR (8K 2D/3D) (0-2 years)	Advanced VR (12K 3D) (2-5 years)	Ultimate VR (24K 3D) (5-10 years)
Data rate	40Mbps(2D), 63Mbps(3D)	340 Mbps	2.34 Gbps
Typical RTT	30 ms (2D), 20 ms (3D)	20 ms	10 ms
Packet loss	2.40E-5	1.00E-6	1.00E-6

Strong-Interactive VR slice

- Direct interactivity w/ virtual environment responding in real-time
- Support for Motion-to-Photon latency of 5-15ms
- Up to 120 fps or more

	Entry-Level VR (8K 2D/3D)	Advanced VR (12K 3D)	Ultimate VR (24K 3D)
Data rate	120Mbps （2D） 200Mbps （3D）	1.40Gbps	3.36Gbps
Typical RTT	10ms	5ms	5ms
Packet loss	1.00E-6	1.00E-6	1.00E-6

* GSMA Network Slicing Use Case Requirements, April 2018

Cloud XR: Motion to Photon Flow and Timing

Flow of Events

1. User/HMD moves
2. Detected and packet goes out thru antenna
3. Packet goes thru network to server (uplink)
4. Cloud Server receives packet
5. Image rendering, coding/compression, streaming
6. Frames sent back to Client Device through network (downlink)
7. Client device receives frame thru wireless interface
8. Client processing (Decoding/Decompression, rendering on HMD)

Timing/Latency

1. Start
2. Minimal
3. Depending on network (Wireless Link, ISP Network, Internet)

   -Wireless Link: Ex. Cellular, WiFi - 802.11ac/ax/ad/ay
5. Rendering(<5ms), Encoding/Compression… (HW/SW dependent)
6. Similar as uplink? Not necessarily…
7. Minimal
8. HW/SW dependent (<3ms)

Cloud XR: End-to-End Latency Calculation

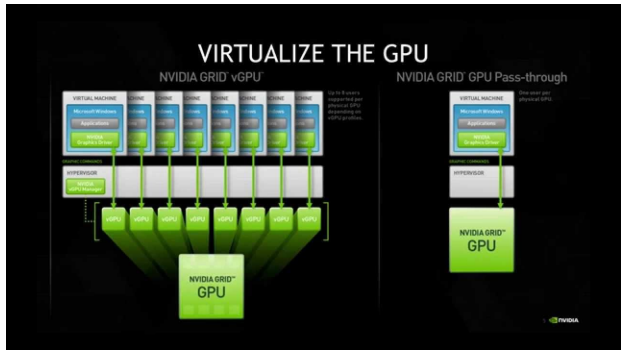
$$Latency = t_{network\ uplink} + \underbrace{t_{render} + t_{coding} + t_{streaming}}_{\text{Server}} + t_{network\ downlink} + \underbrace{t_{decode} + t_{render}}_{\text{Client}}$$

What if Latency > 20 ms? Is there any hope?

- Is 20 ms a hard limit?
- Decrease frame rate (relaxes latency constraints)
- Motion prediction
- Asynchronous space warp = Over-rendering(server) + motion smoothing(client)

Test-bed Setup

1. Cloud VR experience on HTC Vive using Intel Compute Stick
2. Cloud VR experience on HTC Focus



AC WiFi

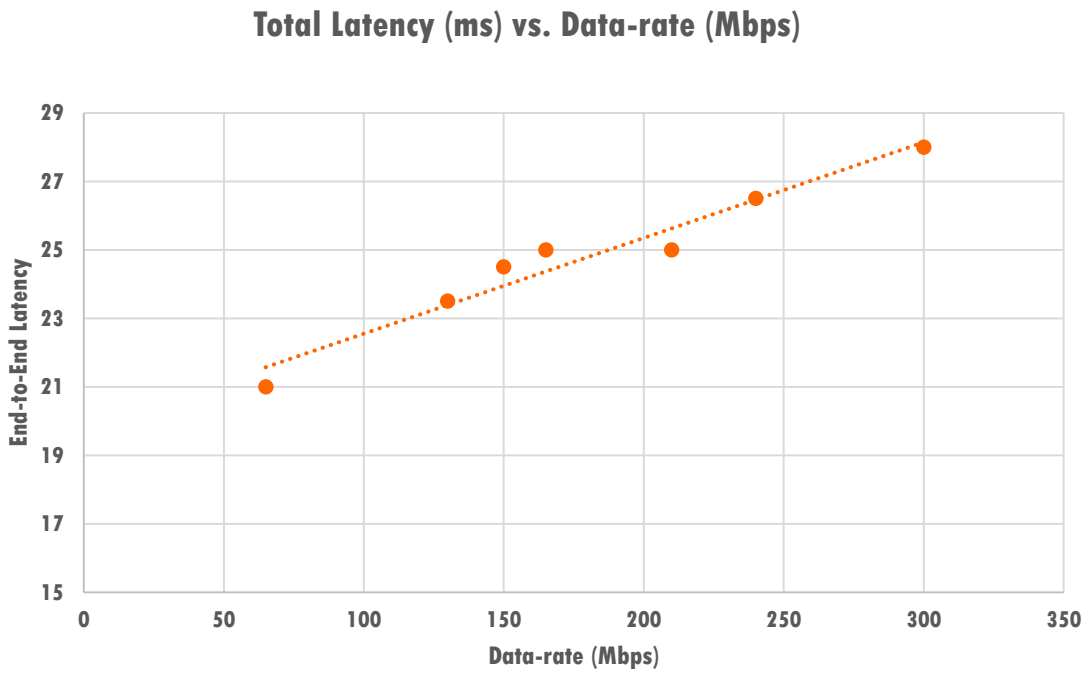
10Gbps Link

Simulated 5G:
1Gbps

Orange Silicon Valley

Results

1. Cloud VR experience on HTC Vive using Intel Compute Stick (QuarkVR JPEG encoding)



- 90 fps
- Latency >20 ms, and quality is good
- 11 ms prediction
- Rendering on GPU server 4-9 ms
- Wi-Fi 802.11ac (3-4 ms – uplink+downlink)
- Streaming works well w/ additional 10 ms latency
    - System could support 5G (R16)

2. Cloud VR experience on HTC Focus (QuarkVR H.264 encoding)

- 75 fps, Data-rate: 37 Mbps, Latency: 50 ms!

Conclusion + Next Steps

Key Question: Can Cloud GPU, 5G and Edge Compute make Cloud XR a reality?

Answer: We think so!

But, it will take work to integrate all the pieces to make a great end user experience.

- Client Devices/HMD/AR glasses?
- Network
- Content (HD 3D, etc.)

Future work:

- Expand the test-bed (RTX server, real 5G network)
- Assess more applications for applicable use cases
- We are open for innovative solutions/partners that can help accelerate 5G-enabled Cloud XR

Thank You

James Li

Principal, Orange Next Reality

(AR/VR, HPC, Embedded AI)

Email: james.li@orange.com

John Benko

Principal, 5G/Emerging Wireless

Email: john.benko@orange.com

To Know More About Us:

http://www.orange.com

http://www.orangesv.com

See Cloud XR in action!

QuarkVR

Our partner QuarkVR is showcasing Cloud XR demo at 6:00PM, March 20th, 2019 (Wednesday)

Location: VR Theater

orange™

Orange Silicon Valley