AV&VR challenges to Network and transport technology

Lin Han
Future Network Lab. USA
Agenda

• Network based AR&VR in the future
• Requirements for network, bandwidth, latency and burst
• Problems in current technologies
• Prospective solutions
The dawn of new era

AR&VR integrated with Internet and Cloud

Localized VR

Network based VR

Cloud VR Service + Local Rendering

Cloud VR Service + Cloud Rendering

http://www.trendhunter.com/trends/backpack-computer
Applications bring up technology revolution

VR live broadcast

NextVR focuses on the VR live broadcast operation (e.g. NBA, Boxing, presidential debate).
- It gained $3.5 million investment from Comcast & Time Warner.

VR online Games

Oculus focus on development of VR online game.
- It has been acquired by Facebook for $2 billion.

VR online shopping

Alibaba is popularizing the concept of Buy+, which adopts the VR technology to provide an interactive 3D online shopping environment for the customers.
Technologies ready?

Cloud VR service + local rendering
- **Network**
  - Ultra-short latency
- **Pro**
  - Less burden to network
- **Con**
  - Powerful local machine
  - Local software upgrading for different applications
  - Less flexibility for APP

Cloud VR service + cloud rendering
- **Network**
  - Ultra-short latency and Ultra-high throughput
- **Pro**
  - Simple local machine
  - More flexibility for APP
- **Con**
  - More challenges to network

Common requirements
- **Sensor**
  - Action detection/capturing
  - Action simulation
- **Computing**
  - Scenario synthesis
  - Rendering
  - Color sampling
  - Coding/decoding
- **Display**
  - Super high resolution
  - Super fast response time
- **Power Supplier**
  - Power consumption
  - Heat Dissipation

Network based AR&VR in the future

Cloud VR service + local rendering
- **Pro**
  - Less burden to network
- **Con**
  - Powerful local machine
  - Local software upgrading for different applications
  - Less flexibility for APP

Network based AR&VR in the future

Cloud VR service + cloud rendering
- **Pro**
  - Simple local machine
  - More flexibility for APP
- **Con**
  - More challenges to network
Encoding and Bitrate

- Normal Encoder setting
  - Size
  - FPS
  - Color, sampling
  - Profile
  - Bitrate mode (CBR, VBR)
  - Pass number
  - Bitrate, maximum bitrate (VBR)
  - GOP distance
  - B-frame
  - ….

No exact formula for bitrate

\[
PSNRY[dB] = 10 \cdot \log_{10} \left( \frac{(2^b - 1)^2}{W \cdot H \sum_{h=1}^{W} \sum_{w=1}^{H} (I_1(w, h) - I_2(w, h))^2} \right).
\]

https://www.sri.com/sites/default/files/publications/3_07_h264_format_bitrate_quality_tradeoff_study.pdf

J.R. Ohm et al., “Comparison of the Coding Efficiency of Video Coding Standards—Including High Efficiency Video Coding (HEVC),” IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 22, NO. 12, DECEMBER 2012
Math for Bit rate, Peak bit rate, Burst (remote rendering)

Bit rate = W * H * FPS * Rank * 0.07  
Adobe, "H.264 Primer", 2016

W: Pixel number in horizontal direction  
H: Pixel number in Vertical direction  
FPS: Frame per second  
Rank: Motion rank,  
- it can be Low motion, Medium motion or High motion.  
- Low motion = 1, Medium motion = 2, High motion = 4.  
- Low motion: video that has minimal movement  
- Medium motion: video that has some degree of movement  
- High motion: video that has a lot movements and movement is unpredictable.

Average Bit rate = T * W * H * S * d * FPS / Cv  
Bit rate for I-frame = T * W * H * S * d * FPS / Cj  
Burst size = T * W * H * S * d / Cj  
Burst time = 1/FPS

T: Type of video, 1 for 2D, 2 for 3D  
W: Pixel number in horizontal direction  
H: Pixel number in Vertical direction  
S: scale factor,  
1 for YUV400, 1.5 for YUV420, 2 for YUV422, 3 for YUV444  
d: Color depth bits  
FPS: Frame per second  
Cv: Average Compression ratio for video.  
Cj: Compression ratio for I-frame

---

<table>
<thead>
<tr>
<th>Entry-level VR</th>
<th>Advanced VR</th>
<th>Ultimate VR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>4K 2D Video</td>
<td>12K 2D Video</td>
</tr>
<tr>
<td>Resolution W*H 360 degree video</td>
<td>3840*1920</td>
<td>11520*5760</td>
</tr>
<tr>
<td>HMD Resolution/ view angle</td>
<td>060*060/ 90</td>
<td>3840*3840/ 120</td>
</tr>
<tr>
<td>FPS</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>Mean Bit rate</td>
<td>22Mbps</td>
<td>398Mbps</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cj</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Peak bit rate</td>
<td>132Mbps</td>
<td>1.9Gbps</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burst size</td>
<td>553K byte</td>
<td>4.15M Byte</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burst time</td>
<td>33ms</td>
<td>17ms</td>
</tr>
<tr>
<td>Infor Ratio of HMD/Whole Video</td>
<td>0.125</td>
<td>0.222</td>
</tr>
</tbody>
</table>
MTP Latency – Motion-to-Photon Latency

【1】**Motion-to-Photon latency** is the time needed for a user movement to be fully reflected on a display screen. (http://www.chioka.in/what-is-motion-to-photon-latency/)

【2】**Motion-to-photon latency** also known as the End-to-end latency is the delay between the movement of the user's head and the change of the VR device's display reflecting the user's movement. As soon as the user's head moves, the VR scenery should match the movement. The more delay (latency) between these 2 actions, the more unrealistic the VR world seems. To make the VR world realistic, VR systems want low latency of <20ms and even really low latency of <7ms (http://xinreality.com/wiki/Motion-to-photon_latency)

However, still some difference

<table>
<thead>
<tr>
<th>MTP (from the movement of user header to …)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oculus, Atman Binstock the screen starts the display of new image</td>
<td><a href="https://developer.oculus.com/blog/optimizing-vr-graphics-with-late-latching/">https://developer.oculus.com/blog/optimizing-vr-graphics-with-late-latching/</a></td>
</tr>
<tr>
<td>Oculus, John Carmack the screen finishes the display of new image</td>
<td><a href="https://www.twentymilliseconds.com/post/latency/mitigation-strategies">https://www.twentymilliseconds.com/post/latency/mitigation-strategies</a></td>
</tr>
<tr>
<td>Valve, Alex Vlachos the screen finishes the display of new image</td>
<td>Advanced VR Rendering, GDC 2015</td>
</tr>
<tr>
<td>Qualcomm the screen center starts the display of new image</td>
<td><a href="http://www.qualcomm.cn/news/blog-2016-07-20">http://www.qualcomm.cn/news/blog-2016-07-20</a></td>
</tr>
</tbody>
</table>
Network Latency is Bigger Challenge

- The maximum latency for network device = 5~7 ms – Propagation Delay(200km/ms)
  User to server distance 500km (round propagation delay= 5ms),
  The maximum queuing latency accumulated on all device (one direction) < 1ms
- Local rendering does not reduce the latency requirement

Latency of Localized VR APP: ~18ms

~1ms
T1: Action capture

~5ms
T2: ROI rendering

~10ms
T4: Display refresh

~2ms
T3: Local transmit

Network based AR/VR APP has 4 more processing than localized APP, Total Latency >> 20ms

Total latency of network based VR APP by current technologies: >>20ms

~1ms
T1: Action capture

~11ms
T3: GOP framing/streaming

~10ms
T6: Display

~5ms
T4: network transport

~5ms
T5: Terminal decoding

~1ms
T2: ROI encoding

~110ms – 1s
Nanjing to Beijing: 1000KM
Measured average latency 112ms

Network latency must be < 5~7ms

Major optimization for processing time in future VR:
- Action capture ~1 ms
- Display refreshing ~0.01ms (AMOLED screen, dynamic refreshing, TimeWap)
- Server coding ~2 ms (HW parallel coding)
- Streaming re-order ~5 ms
- Terminal decoding ~5 ms
- Network transport ~5 ms – 7 ms

Latency in future networks must be about 5~7 ms, considering the technology advances in future
Summary

- Network needs to support “Ultra-short latency and/or Ultra-high throughput”
- Network quality greatly impacts AR&VR User Experience
- Much lower tolerance of packet loss or delay than watching HD TV.
- Compromise of network quality may cause CyberSickness
Current Internet

Hierarchy architecture, different level of aggregation

- The client to server physical distance is too long
- Too many admin domains and devices
  - Increase Queuing delay
  - Reduce throughput
  - Increases OPEX
Can Current Technologies Support Ultra-short Latency and/or Ultra-high Throughput?

Wired Access

- Because of the complex situations in network, it is usually hard for transport protocols to judge whether congestion has happened and obtain high throughput continuously.
- TCP fairness forces all flows share the bandwidth equally, it does not allow an application to obtain more bandwidth or shorter latency than others.
- Maximum throughput and Minimum Latency are conflicting targets, cannot be achieved without flow based processing.

Wi-Fi

- Attenuation and interference in WIFI scenario will make the network status change rapidly, which requires more sensitive transport protocols to reduce the transmission latency.
- Home WIFI access speed is also limited after taking signal quality into account, and employs different rate selection algorithms.

LTE

- Cellular wireless networks usually experience rapidly varying link rates and occasional long outages. For example, in LTE-A today, just the air interface link of an ETE internet path itself is typically much more than 5-7ms.
- Experiments[1] has shown that the transport protocols in use hardly achieve both high throughput and low delay simultaneously.

TCP Throughput ≤ \( \min(BW, \frac{1}{RTT \times \sqrt{3}} \times MSS, \frac{1.22}{RTT \times \sqrt{3}}) \)

[1] NSDI 2013, Keith et.al.
Can Current Transport Support
Ultra-short Latency and/or Ultra-high Throughput?

• Many variations for TCP
  » Place to change: Host only, Network only, or Host plus network
    » Host only: TCP-reno, TCP-vegas, TCP-cubic, TCP-compound, TIMELY, BBR, PCC
    » Network only: PIE(no ECN), CoDel(no ECN), FQ-CoDel(no ECN)
    » Host plus network: DCTCP, PIE(with ECN), CoDel(with ECN), FQ-CoDel(with ECN), XCP, RCP, PERC
  » Reactive or Proactive
    » Proactive: PERC
    » Reactive: Others
  » Allowed bandwidth detection: Congestion based, performance based, rate based, calculation based
    » Congestion based: TCP-reno, TCP-cubic, TCP-vegas, PIE, TIMELY, DCTCP, CoDel, FQ-CoDel
    » Performance (rate, RTT, loss) based: PCC, BBR
    » Rate based: XCP, RCP,
    » Calculation based: PERC
  » Congestion detection: packet loss, RTT, packet loss+RTT, Delay on router, Q depth
    » Packet loss: TCP-reno, TCP-cubic
    » RTT: TCP-vegas, TIMELY
    » Packet loss+RTT: TCP-compound
    » Delay on router: PIE, CoDel, FQ-CoDel,
    » Q depth: DCTCP,
  » Rate detection: Implicit or Explicit rate
    » Explicit: XCP, RCP, PERC, BBR, PCC
    » Implicit: Others

Current TCP problems
• Fairness forces all flows share the bandwidth equally, it does not allow a application to obtain more bandwidth or shorter latency than others instantaneously
• Maximum throughput and Minimum Latency are conflicting targets, cannot be achieved without flow based processing

Current UDP problems
• UDP without congestion control is worse than TCP in four targets
• UDP with congestion control (like QUIC) behaves same as TCP

Root cause
• IP only provides best effort service
• Current QoS tech is not for end-user APP
Prospective solution: From interface to protocol to architecture

<table>
<thead>
<tr>
<th>Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>• New radio technologies, higher bandwidth and lower latency</td>
</tr>
<tr>
<td>• Heterogeneous access</td>
</tr>
</tbody>
</table>

| NG RAN |
|• NFV and slicing based 5G architecture |
|• Flat internet, less hierarchy |
|• More distributed DC, closer to user, MEC |

| NG EPC and internet |
|• User driven flow based quality assurance for transport |
|• Simpler and more efficient congestion control |
|• Better utilization, convergence and fairness |
|• Coexists with traditional transport |

### Current research
- ETSI: 3GPP, NGP (Next Generation Protocol)
- IRTF: iccrg (Internet Congestion Control)