Media Operations Use Case for an Augmented Reality Application on Edge Computing Infrastructure

draft-ietf-mops-ar-use-case-10

Renan Krishna, Akbar Rahman

MOPS WG IETF 116 Yokohama March 2023

Draft's Table of Contents

Sections 5.1 and 5.2 have been updated

Table of Contents

New Update

New Update

Section 5.1 Update: XR Workload Characteristics

Application	Throughput Required
Image and Workflow Downloading	1 Mbps
Video Conferencing	2 Mbps
3D Model and Data Visualization	2 to 20 Mbps
Two way Telepresence	5 to 25 Mbps
Current-Gen 360 degree video (4K)	10 to 50 Mbps
Next-Gen 360 degree video (8K, 90+ FPS, HDR, Stereoscopic)	50 to 200 Mbps
6DoF Video or Point Cloud	200 to 1000 Mbps

Table 1: Throughput of some XR Applications

- As seen from the table [METRICS_1], an XR application such as our use case transmits a larger amount of data per unit time as compared to traditional video applications:
  - As a result, issues arising out of heavy tailed parameters such as long-range dependent traffic [METRICS_2], self-similar traffic [METRICS_3], would be experienced at time scales of milliseconds and microseconds rather than hours or seconds.
  - Additionally, burstiness at the time scale of tens of milliseconds due to multi-fractal spectrum of traffic will be experienced [METRICS_4].

Section 5.1 Update: Operational Consequences of XR Workload Characteristics

- The operational consequences of XR traffic having characteristics such as long-range dependency, and self-similarity is that the edge servers to which multiple XR devices are connected wirelessly could face **long bursts of traffic**.
- In addition, multi-fractal spectrum burstiness at the scale of milli-seconds could induce **jitter** contributing to motion sickness.
- The operators of edge servers will need to run a **'managed edge cloud service**' [METRICS_5] to deal with the above problems.
  - Functionalities that such a managed edge cloud service could operationally provide include dynamic placement of XR servers, mobility support and energy management [METRICS_6].
  - Providing Edge server support for the techniques being developed at the DETNET and RAW Working Groups at the IETF could guarantee performance of XR applications.

Section 5.2 Update: XR Performance Metrics

Application	Expected End-To-End Latency	Expected Data Latency	Possible Implementations/ Examples
AR-based remote surgery with uncompressed 4K (3840x2160 pixels) 120 fps HDR 10-bit real time video stream	Less than 750 microseconds	Greater than 30 Gbps	World's first remote surgery over 5G
Mobile AR based remote assistance with uncompressed 4K (1920x1080 pixels) 120 fps HDR 10-bit real-time video stream	Less than 10 milliseconds	Greater than 7.5 Gbps	Assisting maintenance technicians, Industry 4.0 remote maintenance, remote assistance in robotics industry
Indoor and localized outdoor navigation	Less than 20 milliseconds	50 to 200 Mbps	Theme Parks, Shopping Malls, Archaeological Sites, Museum guidance
Cloud-based Mobile AR applications	Less than 50 milliseconds	50 to 100 Mbps	Google Live View, AR-enhanced Google Translate

Table 2: Traffic Performance Metrics of Selected XR Applications

- The adjoining Table 2 [METRICS_6] shows a taxonomy of applications with their associated expected latencies and bandwidths. Our use case requires an RTT of 20ms at most and preferably between 7-15ms as discussed earlier. The required bandwidth for our use case as discussed in section 5.2 is 200 Mbps-1000 Mbps.

Section 5.2 Update: Operational Consequences of XR Performance Metrics

- Since our use case envisages multiple users running the XR applications on their devices, and connected to an edge server that is closest to them, these latency and bandwidth connections **will grow linearly with the number of users**.
  - The operators should match the network provisioning to the maximum number of tourists that can be supported by a link to an edge server.

Some additional changes in the draft for WG's consideration

- Distinguish between the response times required by the XR applications such as our use case and end-to-end latency metric used by network operators.
- Update the numbers as per the above distinction including stating that the latency and throughput numbers reflect the current technology i.e. 2018-till present.
- With these additional changes does the WG think the document will be ready for WGLC?

References

[METRICS_1] ABI Research, "Augmented and Virtual Reality: The first Wave of Killer Apps.", https://gsacom.com/paper/augmented-virtual-reality-first-wave-5g-killer-apps-qualcomm-abi-research/, 2017.

[METRICS_2] Paxon, V. and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modelling.", In IEEE/ACM Transactions on Networking, pp. 226-244., 1995.

[METRICS_3] Willinger, W., Taqqu, M.S., Sherman, R., and D.V. Wilson,"Self-Similarity Through High Variability: Statistical Analysis and Ethernet LAN Traffic at Source Level.", In IEEE/ACM Transactions on Networking, pp. 71-86., 1997.

[METRICS_4] Gilbert, A.C., "Multiscale Analysis and Data Networks.",In Applied and Computational Harmonic Analysis, pp. 185-202., 2001.

[METRICS_5] Beyer, B., Jones, C., Petoff, J., and N.R. Murphy, "Site Reliability Engineering: How Google Runs Production Systems.", O'Reilly Media, Inc., 2016.

[METRICS_6] Siriwardhana, Y., Porambage, P., Liyanage, M., and M.Ylianttila, "A survey on mobile augmented reality with 5G mobile edge computing: architectures, applications, and technical aspects.", In IEEE Communications Surveys and Tutorials, Vol 23, No. 2, 2021.