Use Case #8

AR-assisted emergency surgical care
[UMA Testbed]

Overview and Objectives

For the first-party experimentation of the 5G-EPICENTRE framework, ORAMA has developed a Unity-based[1] AR application, which facilitates first aid responders with critical information about an injured patient on the disaster site, using its novel Extended Reality (XR) authoring framework (MAGES) Software Development Kit (SDK). Using an AR headset equipped with this application, first aid responders can have access to step-by-step instructions for a specific surgical task or anatomical information (e.g., organs, vessels, bones, etc.), overlayed as deformable objects on top of the patient's body.

A figure showcasing the content of this application is shown below (Figure 1): A first aid responder, equipped with an AR-Head-Mounted Display (HMD), is depicted at a disaster site, where they located a patient. Utilizing the designed application, the responder gains access to step-by-step instructions for critical operations and is able to visualize the patient's veins, bones, and/or internal organs as deformable objects overlaid on the patient's body.

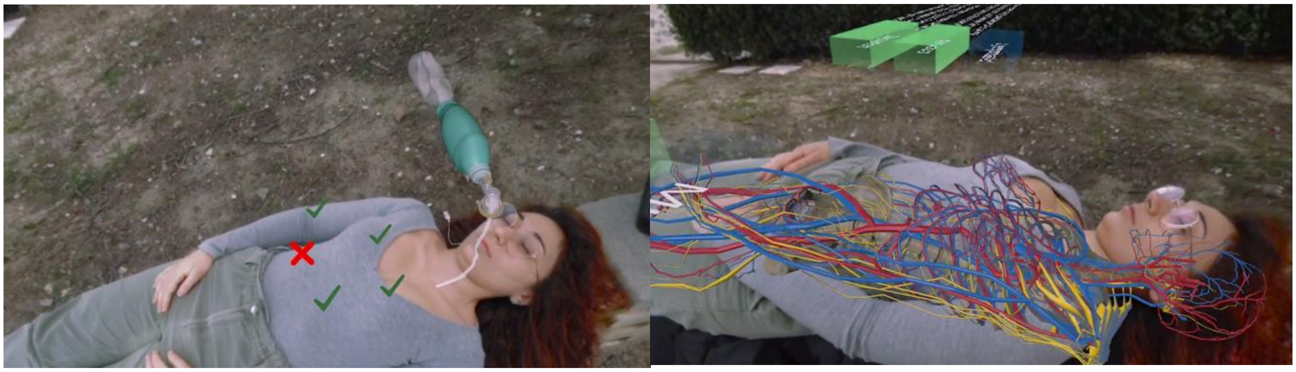

Figure 1 : Overlay on patient's body

The original application was modified to offload the process of scene rendering to a vertical application lying on the edge-cloud continuum, taking advantage of the available Central and Graphics Processing Unit (CPU & GPU) resources to perform the heaviest task in the pipeline, i.e., the scene generation and rendering. The output video stream will be transmitted over the 5G network to the AR headset, where a user-facing application was deployed to receive, decode and project it on the HMD's screen. The latter application is also used to track and send the headset's input (movement and controller triggers) to the edge-residing network application, so that the scene and game play is updated, based on the user's actions.

The objectives of the experimentation were to measure E2E network latency for video streaming from the edge resources to the HMD, the maximum aggregated bandwidth, and packet loss. Various experiments under different settings were conducted to optimize these metrics. In each experiment, in addition to recording these metrics, we also noted the qualitative result, i.e., the visual-driven QoE of the AR user. Additionally, we measured the HMD's energy consumption by recording its battery drain.

---

[1] www.unity.com

This project has received funding from the European Union's Horizon 2020
Innovation Action programme under Grant Agreement No 101016521.

Use Case Description

Surgeons should play a central role in disaster planning and management due to the overwhelming number of bodily injuries that are typically involved during most forms of disaster. In fact, various types of surgical procedures are performed by emergency medical teams after sudden-onset disasters, such as soft tissue wounds, orthopaedic traumas, abdominal surgeries, etc. [1] [2]. HMD-based Augmented Reality (AR), using state-of-the-art hardware such as the Magic Leap or the Microsoft HoloLens, have long been foreseen as a key enabler for clinicians in surgical use cases [3], especially for procedures performed outside of the operating room. In such condtions, monolithic HMD applications fail to maintain important factors such as user mobility, battery life, and Quality of Experience (QoE), leading to considering a distributed cloud/edge software architecture. Toward this end, 5G and cloud computing will be a central component in accelerating the process of remote rendering computations and image transfers to wearable AR devices.

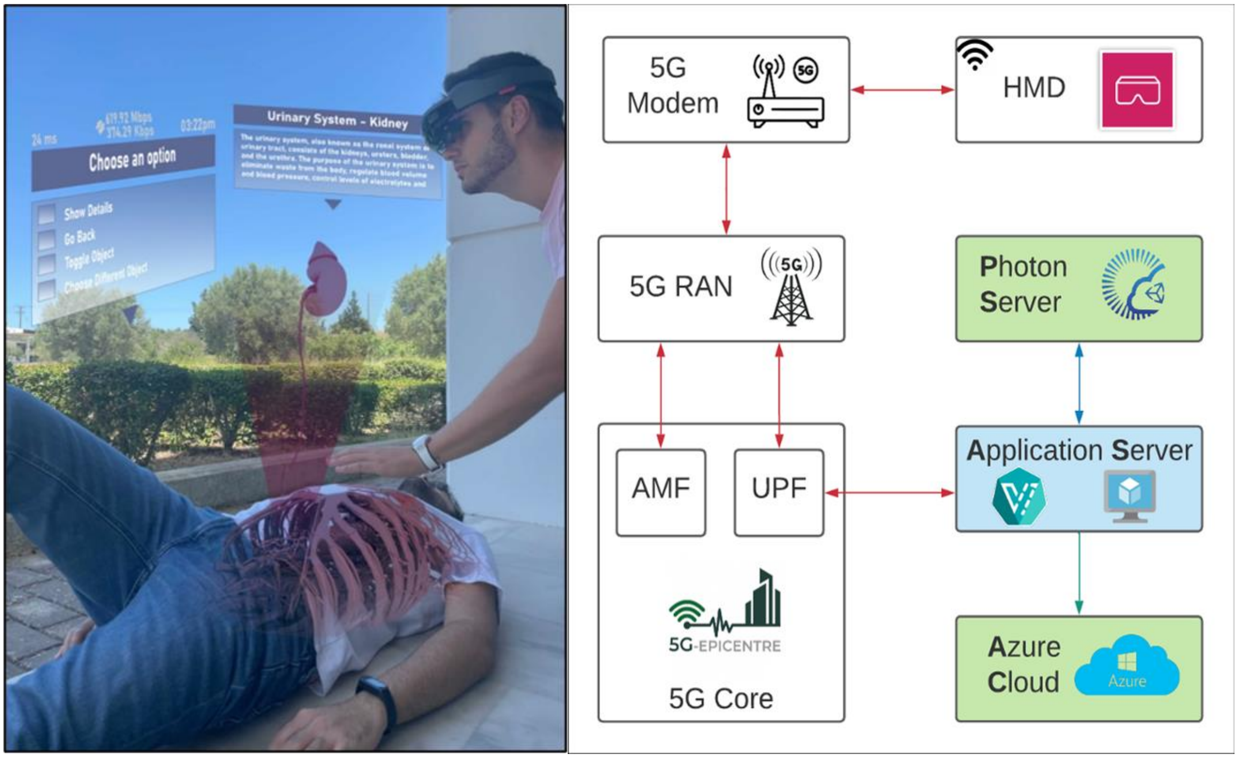

Figure 2: UC8 - The PPDR responder uses an AR HMD to see overlayed info and deformable objects on top of the patient. Envisioned example of UI layout and layout of the UC components.

Experiment Setup/Methodology/Deployment

UC8 opted to perform remote tests with UMA due to the availability of more advanced headsets at the testbed (Quest Pro). Additionally, UC8 deployed and tested the UC application using a standalone Windows machine instead of a Kubernetes VM (CTTC), while also exploring the higher bandwidth aggregation offered by the UMA testbed. Furthermore, conducting experiments in both testbeds allowed for cross-validation of the output results, providing a sanity check.

The network setup of the experiment with the UMA testbed and the HMD is depicted in Figure 3. Specifically, a VR HMD Quest Pro model (available at the UMA testbed for remote testing) was connected via WiFi with a 5G router, namely a Netgear Nighthawk M6 Pro, which had port-forwarding enabled for ports 9943 and 9944. The 5G router was connected via 5G to the 5GC, and eventually the Windows-11 machine, running the application server. Standard N2 and N6 interfaces were used for the 5G-Core communication.

An AR device available at that UMA testbed (namely the Quest Pro) was used to stream the video, during our remote testing sessions, using the UC8's streaming service, namely ALVR/ALXR2.

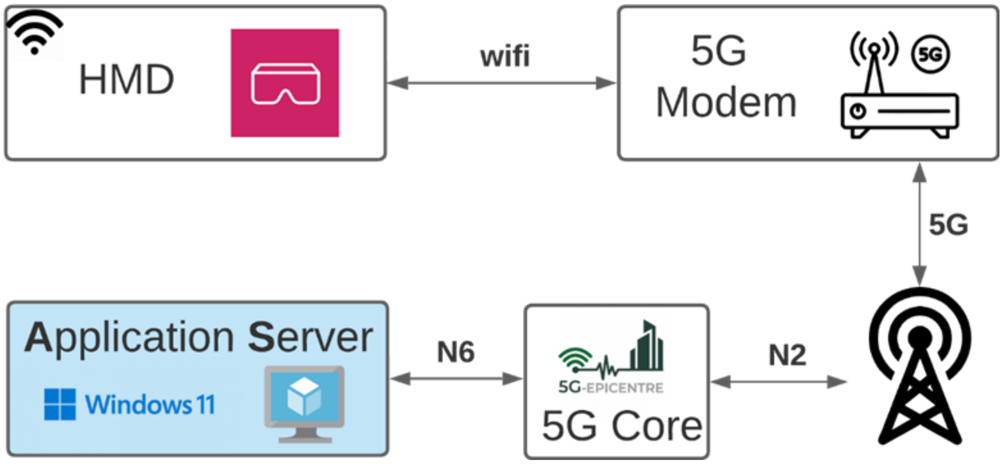

Figure 3 : Experiment setup for UC8 at UMA

Due to the absence of GPUs in UMA's cluster, a crucial prerequisite for UC8, a standalone Windows 11 machine was set up at the UMA premises with the following specifications:

- **CPU:** AMD Ryzen Threadripper PRO 3955WX 16-Cores 3.9GHz;
- **RAM:** 128 Gb DDR4 3200MHz;
- **GPU:** 2xGigabyte GeForce RTX 4080 16GB Gaming;
- **NETWORK:** 2 x 10 Gbit Ethernet; and
- **Disk Storage:** Nvme 1TB, HHD 6 TB.

---

[2] https://github.com/alvr-org/ALVR

Experiment Execution and Results

We have conducted a series of remote experiments of UC8 on the UMA testbed. Specifically, after UMA has connected the HMD with the Application server based on ORAMA's instructions, ORAMA has remotely connected to the Windows machine and properly initialized the remote rendering application, the streaming service and metrics broker. The latter was responsible for both documenting the metrics locally, as well as sending the metrics to the testbed analytics service, which in turn send the metrics to the 5G-EPICENTRE Analytics Engine (AE), developed by IST. After the initiation of the remote rendering, a user at UMA premises wearing the HMD was responsible of performing various actions, such as moving the HMD and its controllers and assessing the overall QoE, which was documented and is described below. The remote rendering service used includes a bitrate setting (in Mbps), which corresponds to the video bitrate that the encoder/streamer will try to achieve on average throughout the experiment, sometimes exceeding or undershooting it. Based on the UMA user feedback, ORAMA performed various experiments on different bitrate settings, aiming to unravel the optimal settings for network exploitation, as well as the user's QoE.

In all the experiments detailed below, video streaming predominantly utilized the UDP protocol. This choice arises from UDP being the most prevalent and suitable protocol for video streaming. Opting for TCP as an alternative would bring about extra latency attributable to the TCP handshake, involving continuous package delivery acknowledgments and the handling of multiple packets by the HMD.

All experiments were eventually summarized for the sake of clarity into three major categories: (1) a Network-stress experiment; (2) a QoE-driven experiment and (3) an energy-consumption experiment.

E1. Network-stress Experiment: In this experiment, our goal was to thoroughly investigate the capabilities of 5G networking when utilized with our developed application. The video streaming server embedded in the UC application was configured to operate at a high bitrate, specifically, 500 Mbps. Consequently, the edge-based application server endeavored to transmit a rendered video at this rate through the 5G network to the VR headset, which was tasked with decoding and projecting the content on its screen. The outcomes of this experiment are presented in Figure 4 and Figure 5.

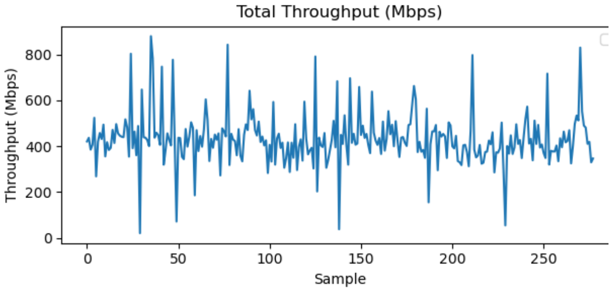

Figure 4: UC8.E1 experiment at UMA (Throughput)

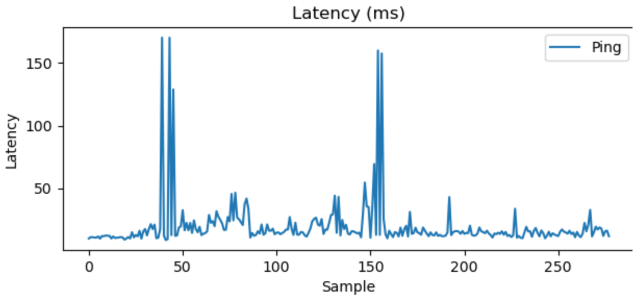

Figure 5: UC8.E1 experiment at UMA (Latency)

The findings in this particular scenario are elucidated as follows: a mean throughput of 431 Mbps was successfully transmitted over the 5G network, accompanied by a mean latency of 19.65 ms. It is noteworthy that the throughput consistently surpassed 400 Mbps throughout the experiment, predominantly hovering around 500 Mbps (albeit not consistently sustained), with fluctuations.

While this initial experiment achieved the KPIs envisioned for UC8, the user wearing the headset did not have an enjoyable AR experience. Issues such as video stuttering, minimal framerate (FPS), and subsequent feelings of dizziness led to a lack of immersion for the user. The strain on the headset's CPU/GPU to process incoming data led to decoding delays, causing video stuttering and HMD overheating, ultimately resulting in a non-existent QoE for the session.

E2. QoE-driven Experiment: In this configuration, we again aimed for optimal AR experience for the end-user, using a reduced bitrate of approximately 60Mbps, using the Quest Pro device. The metrics derived from this experimentation are illustrated in Figure 6 and Figure 7.

In these configurations, we accomplished a throughput averaging 66.8 Mbps (occasionally spiking at 383 Mbps) successfully transmitted over the 5G network, while the mean latency dropped to 10.8 ms, remarkably close to the KPI of 7 ms. In these settings, the experimenter reported a remarkable QoE, with a high framerate. The results of this experiment, depicted in Figure 6 and Figure 7, highlight the ability of delivering minimal latency using 5G networks.

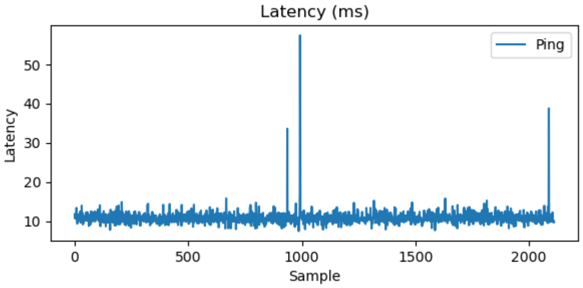

Figure 6: UC8.E2 experiment at UMA (Latency)

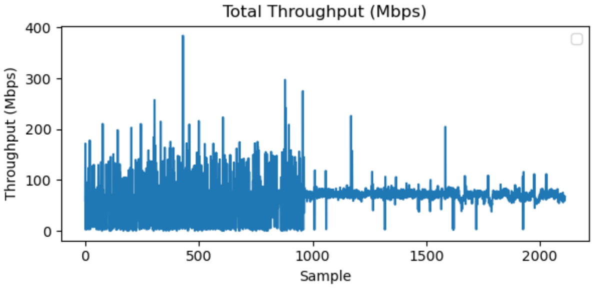

Figure 7: UC8.E2 experiment at UMA (Throughput)

E3. Energy-Consumption testing Experiment: In accordance with KPI3 of UC8, our objective was to assess the headset's battery consumption in both remote rendering (as in 5G-EPICENTRE) and local rendering (i.e., rendering and physics computations by the HMD) setups. This aimed to validate our claim that the HMD could experience up to a 30% reduction in battery drain when utilizing the remote rendering configuration. This reduction directly results from offloading the resource-intensive scene rendering process to the cloud-edge continuum, thereby replacing it with the simpler and lighter task of video reception and decoding.

For our experiments, we employed a Quest Pro AR headset and monitored its battery levels during both remote and local rendering. For the remote rendering, we opted for the QoE-driven configuration, *i.e.*, a bitrate of approximately 80Mbps, achieving 60-90 Mbps throughout the experiment. We used such a setting, as we also wanted to compare it with existing metrics ORAMA had taken from its local setup, achieving the same bitrate, and, indeed, they proved to be identical. Of course, a higher bitrate setting could affect the battery drain, potentially requiring higher battery consumption.

In order to record the battery level status, ORAMA tried to identify a proper way of automatically recording it and logging it. Despite testing multiple approaches, we were not able to identify a solution that was device agnostic and compatible with our sophisticated software and, most importantly, reliable. Therefore, in order to obtain the required battery metrics in a consistent way, we had users manually recording the HMD's battery status every 30 seconds. The conclusive outcomes, illustrated in Figure 8, confirm a 30% decrease in HMD battery consumption when employing the remote rendering setup. This reduction precisely corresponds to the remaining battery capacity of the HMD at the point when the local rendering setup exhausted its battery.

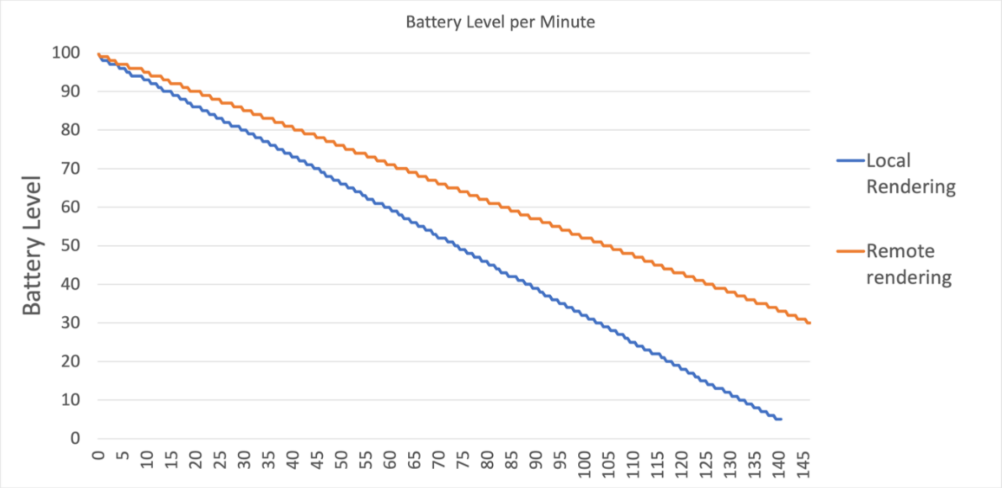

Figure 8: UC8 Battery drain tests – Experiment E3

The whole experimentation process described above resulted in measurement results shown in Table 1.

Table 1: UC8 performance at UMA

Experiment	Experiment Name	Target KPI	Performance
E1	Throughput Test	U ≤ 0.4 Gbps < A ≤ 0.7 Gbps < O	0.431 Gbps
E2	RTT Test	U ≥ 20ms > A ≥ 7ms > O	10.8 ms

The UC8 KPIs are shown in Table 2.

Table 2: UC8 KPIs

KPIs	Results expected	Experimentation results
UC 8.1	E2E latency less than 7ms	E2E latency measured at 10.8ms (on average)
UC 8.2	Maximum aggregated total system bandwidth of at least 0.7Gb/s	Maximum combined overall system bandwidth of at 0.431 Gbps (maintained), predominantly exceeding 0.5 Gbps (though not consistently sustained) most of the time.
UC 8.3	Decrease device energy consumption by at least 30%.	Achieved 30% decreased device energy consumption.

Overall evaluation

Concluding the experimental phase, UC8 has yielded commendable outcomes despite the ambitious KPIs. Notably, the remote rendering architecture successfully delivered a 30% extension in the battery life of the HMD. This enhancement implies that first aid responders can utilize the devices for prolonged durations on site before the battery depletes.

Although the RTT averaged 10.8ms, surpassing the anticipated KPI value of 7ms, it remains noteworthy. This achieved latency facilitates an immersive AR session with heightened QoE, devoid of video stuttering and featuring high video quality—a critical consideration in all AR applications.

In terms of throughput, we approached the KPI target of 0.7 Gbps (700 Mbps), achieving sustained rates exceeding 400 Mbps and occasionally surpassing 500 Mbps. This suggests the potential to surpass the limitations of 4G network architecture, if the respective AR application requires so.

Conclusions

In conclusion, we have managed to reach each individual KPI under various settings: minimal end-to-end latency, high bandwidth aggregation, and a 30% decrease in energy consumption on the HMD. Additionally, we have validated that 5G-EPICENTRE can provide a platform where Unity-based applications, which are notoriously difficult to containerize efficiently, can be deployed and run. We have successfully deployed and run such an AR application using VMs, designed for the PPDR community, providing results that validate its efficacy. Our results further illustrate that similar applications, requiring either minimal packet loss or high bandwidth aggregation, are supported by the 5G-EPICENTRE platform.

References

[1] C. A. Coventry, A. I. Vaska, A. J. Holland, D. J. Read, and R. Q. Ivers, "Surgical procedures performed by emergency medical teams in sudden- onset disasters: a systematic review," World journal of surgery, vol. 43, no. 5, pp. 1226–1231, 2019.

[2] T. Birrenbach, J. Zbinden, G. Papagiannakis, A. K. Exadaktylos, M. M˙uller, W. E. Hautz, and T. C. Sauter, "Effectiveness and utility of virtual reality simulation as an educational tool for safe performance of covid-19 diagnostics: Prospective, randomized pilot trial," JMIR Serious Games, vol. 9, no. 4, p. e29586, Oct 2021. [Online]. Available: https://games.jmir.org/2021/4/e29586

[3] P. Zikas, S. Kateros, N. Lydatakis, M. Kentros, E. Geronikolakis, M. Kamarianakis, G. Evangelou, I. Kartsonaki, A. Apostolou, T. Birrenbach et al., "Virtual reality medical training for covid-19 swab testing and proper handling of personal protective equipment: Development and usability," Frontiers in Virtual Reality, p. 175, 2022.

[4] G. Papagiannakis, P. Zikas, N. Lydatakis, S. Kateros, M. Kentros, E. Geronikolakis, M. Kamarianakis, I. Kartsonaki, and G. Evangelou, "Mages 3.0: Tying the knot of medical VR," in ACM SIGGRAPH 2020 Immersive Pavilion, 2020, pp. 1–2.

[5] G. Papagiannakis, N. Lydatakis, S. Kateros, S. Georgiou, and P. Zikas, "Transforming medical education and training with VR using MAGES," in SIGGRAPH Asia 2018 Posters, 2018, pp. 1–2.