An Edge Cloud Architecture for Latency-Critical Interactive Applications

01 The problem

The emergence of holographic and head-mounted displays is fostering the emergence of a new generation of “hyper-interactive” applications whose response times remain below the limits of human perception: augmented reality applications blending virtual and real worlds, massive multi-player alternate-reality games etc. Such technologies are now available using fixed machines connected to high-speed wired networks. Currently, however, delivering this quality-of-experience remains impossible using mobile devices connected via mobile phone networks.

Interactive multi-user applications usually rely on intermediate cloud servers to mediate the inter-user interaction. However, current mobile networks exhibit network latencies in the order of 50-150 ms between the device and any cloud. Such latencies make it impossible to create smooth interactions with the end user. To enable an “instantaneous” feeling, augmented reality applications require that end-to-end latencies should remain below 20 ms.

02 The solution

- Mobile edge cloud physically deployed close to the users (e.g., in a shopping street)
- Client devices establish direct Wi-Fi connections to the cloud nodes
- Large number of small and cheap cloud nodes

Virtualization

- ARM7 CPUs do support virtualization, but deploying multiple VMs per node turned out to be infeasible due to limited installed memory.
- We rather use LXC containers which allow a much more conservative memory management.

OpenStack

Deploying OpenStack is a challenge:

- OpenStack was not designed with this type of hardware in mind...
- OpenStack services require considerable resources: authentication, virtual machine image storage, etc.

Solution: relieve the edge nodes as much as possible

- Setup the resource-hungry centralized services in a centralized powerful machine
- Edge nodes execute only the latency-sensitive application deployment mechanisms
- This speeds up container deployment time from 15-20 minutes to 1-2 minutes at most

Application Deployment

- Elastic application deployment using ConPaaS [1].
- Cloud instance migration to follow end users’ mobility: we plan to use a solution similar to [2].

03 The demo

Let’s deploy a video game server (Cube 2 Sauerbraten) in two different clouds:

- A powerful external commercial cloud provider
- Our mobile edge cloud running on Raspberry Pis.

In both cases, we start two game instances and connect them to the game server running in the cloud.

When one client shoots the other, we can hear:

- The gunfire sound heard by the shooting client
- The gunfire sound heard by the second client

How does the client-cloud latency affect the gaming experience?

Is a Raspberry Pi powerful enough to run a game server inside a cloud container instance?

**References**