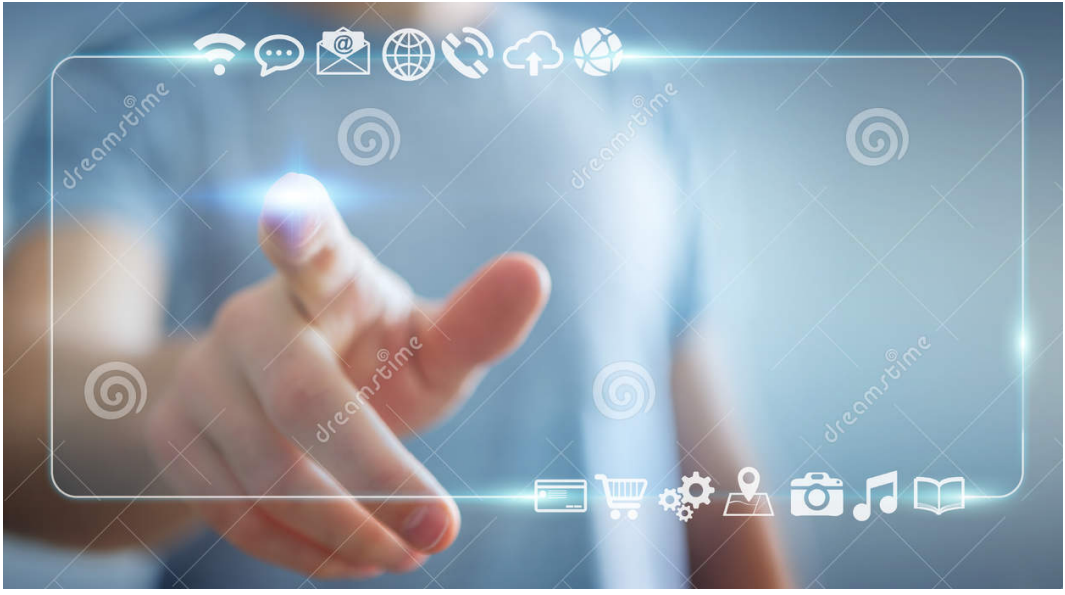
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Interactive Multimedia Applications in the Cloud

Gwendal SIMON

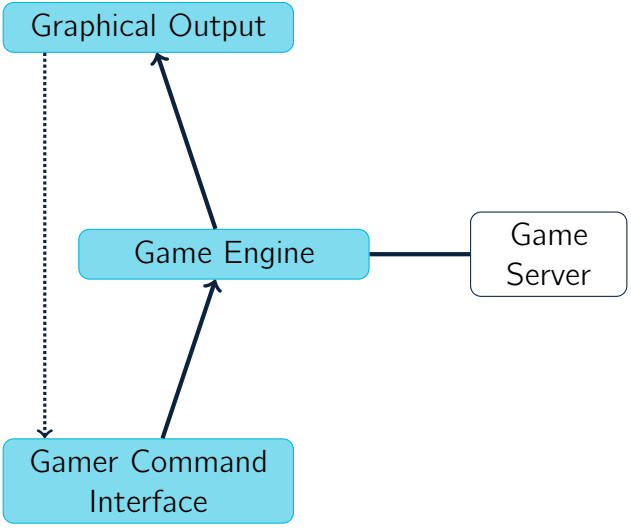


Interactivity Needs Low Latency

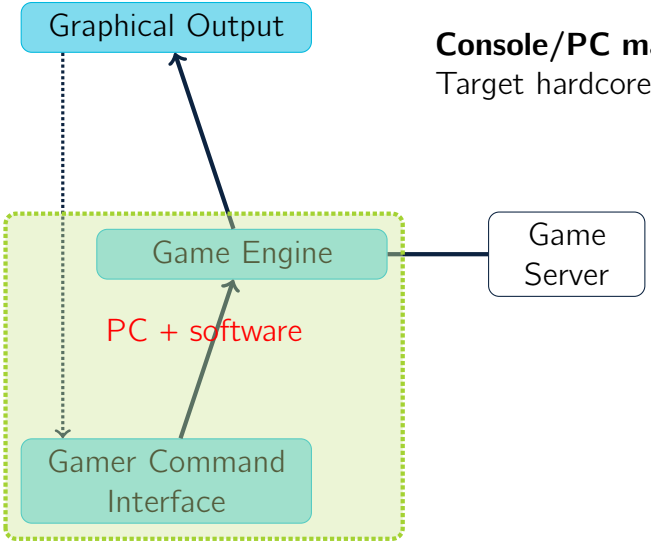
Strong requirements of 5G networks:
running cloud applications as if hosted in the device

Table of contents

Gaming modular architecture

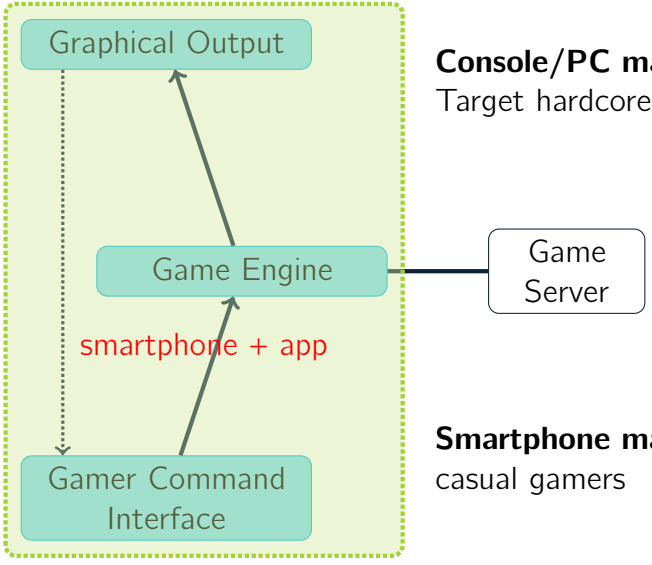


Gaming modular architecture



Console/PC market based on powerful hardware.
Target hardcore gamers

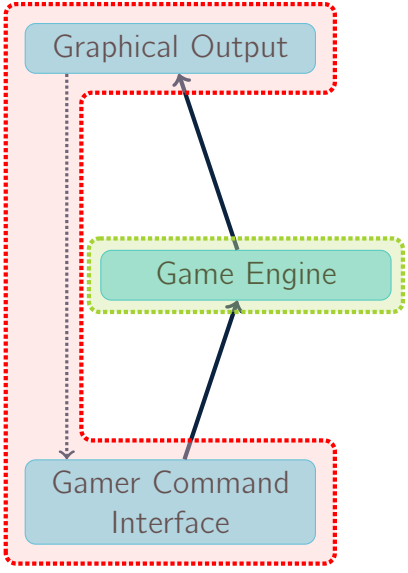
Gaming modular architecture



Console/PC market based on powerful hardware. Target hardcore gamers

Smartphone market based on app store. Target casual gamers

Gaming modular architecture

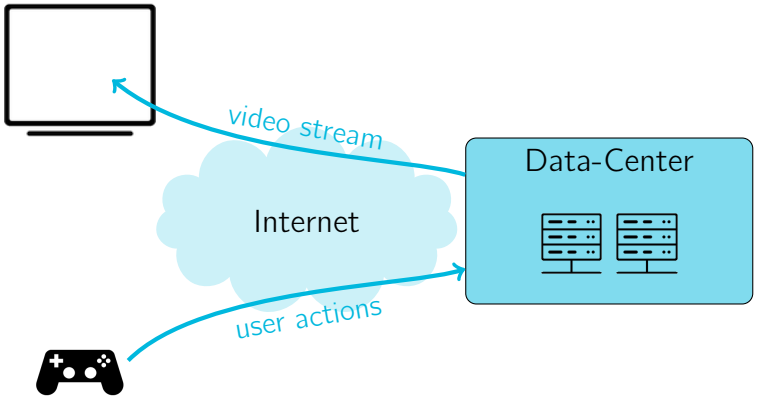


Console/PC market based on powerful hardware. Target hardcore gamers

Cloud gaming based on the cloud and lightweight devices. Target everybody

Smartphone market based on app store. Target casual gamers

Cloud gaming architecture in a nutshell



Why does it make sense?

For game developers:

- Better piracy control
- Support more client platforms
- Easier to upgrade
- New business offers (e.g. subscription-based)

For gamers:

- No more hardware
- Play games on thin clients
- Follow-me gaming

Business status

The pioneers have shut down:

- **OnLive** bankrupted in 2013. Too early, too bad quality
- **StreamMyGame** never took off

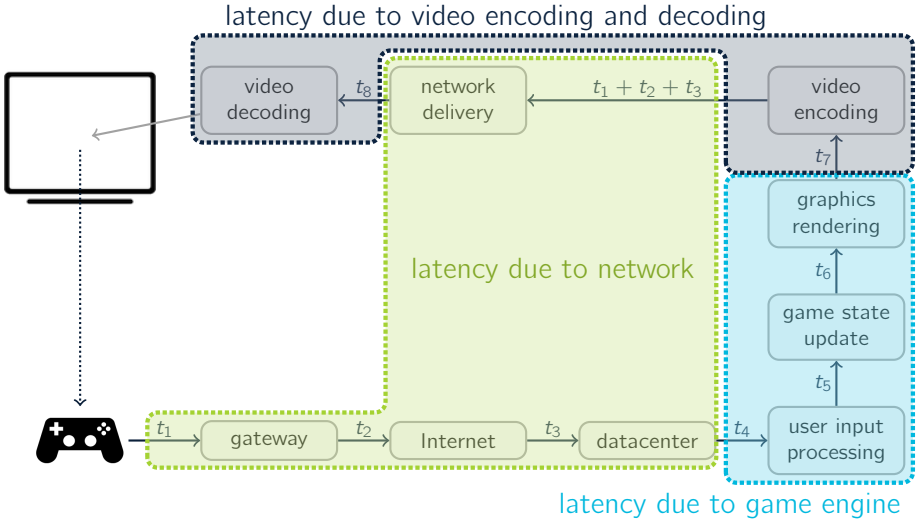
Last game consoles integrate cloud gaming solutions:

- Sony acquired **Gaikai** for PS4 (PS3 backward compatibility)
- Microsoft launched **XBox Live Cloud** a year later

New actors entered the market:

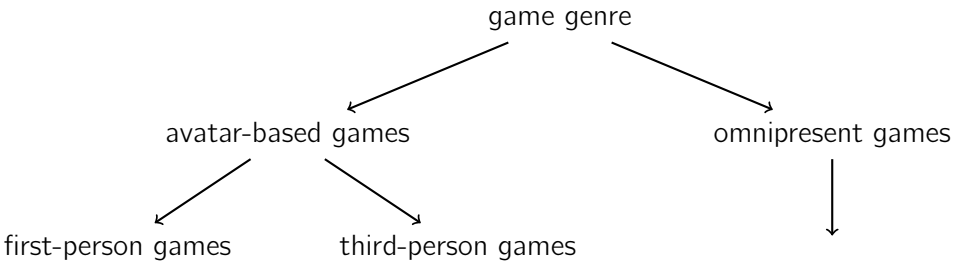
- NVidia released **Nvidia Grid** and **Shield**
- **Blacknut** in IMT Atlantique incubator

Cloud Gaming: The Chain of Latency



Acceptable Latency Target

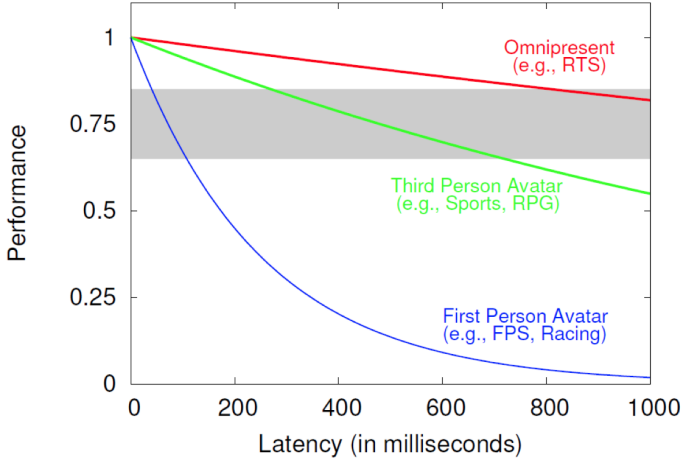
Game genre


game genre
avatar-based games
omnipresent games
first-person games
third-person games

Genre, latency, and performance

The **Claypool law** of latency acceptance:

genre	first-person	third-person	omnipresent
lag in ms	100	500	1000



Genre, latency, and performance

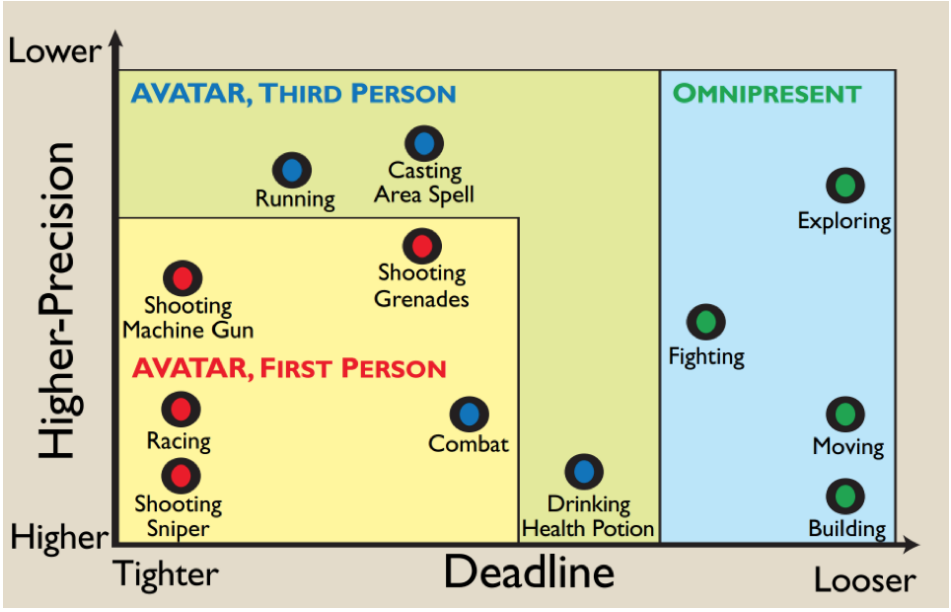
The **Claypool law** of latency acceptance:

genre	first-person	third-person	omnipresent
lag in ms	100	500	1000

Various other factors to consider:

- Limitations in the classification (*e.g.* StarCraft)
- User control games (more or less immersive)
- Gamer skill level (from casual to professional)

Game action taxonomy



Latency target

The shorter the better

The best range is from 100 to 150 ms

Up to 300 ms is enough for some games

Reality Check on Processing Time

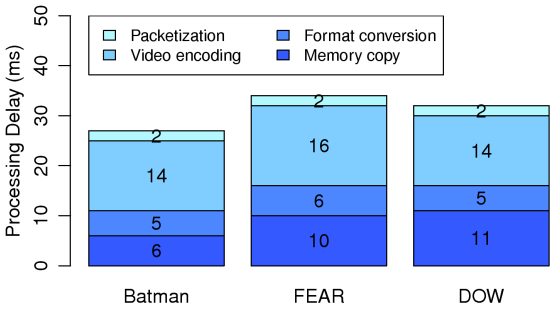
Gamers have **high expectations**:

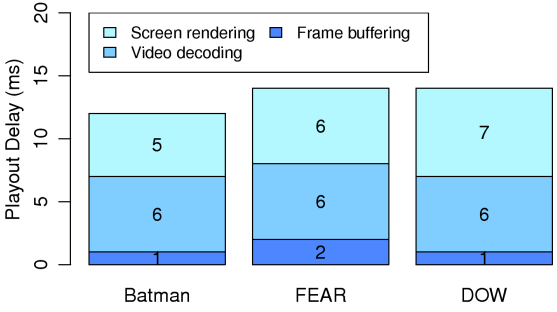
- High-resolution (1080p possible since PS3)
- High frame rate (60 fps is common)
- Complex, dynamic scenes

Screencasting technologies are booming:

- Capture, encode, and packetize
- *e.g.*, Open Broadcast Software

Measured in the **GamingAnywhere** platform:





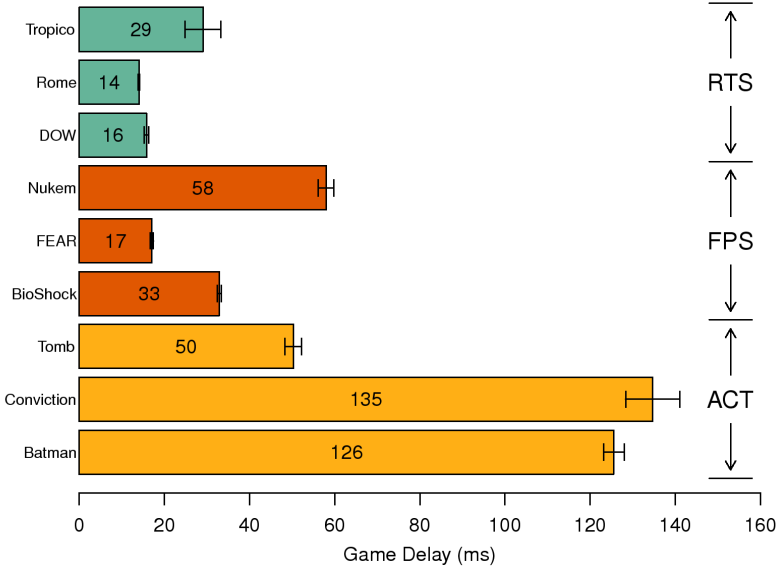
Measured in the **Rhizome** platform:

encoding	capture image	convert	encode frame	stream	total
720p	6.4 ms	4.4 ms	2.8 ms	0.2 ms	13.8 ms
1080p	11.0 ms	4.9 ms	3.0 ms	0.3 ms	19.2 ms

Use info from the game engine to encode video

- Reduce Motion Estimation (ME) using game object info and depth map
- Reduce encoding complexity using adaptive object encoding

Game engine latency



Virtualization on GPU

Cloud management is based on **Virtual Machines**

- Mature technologies to create, shut down, migrate
- Hardware consolidation: several VMs per machine

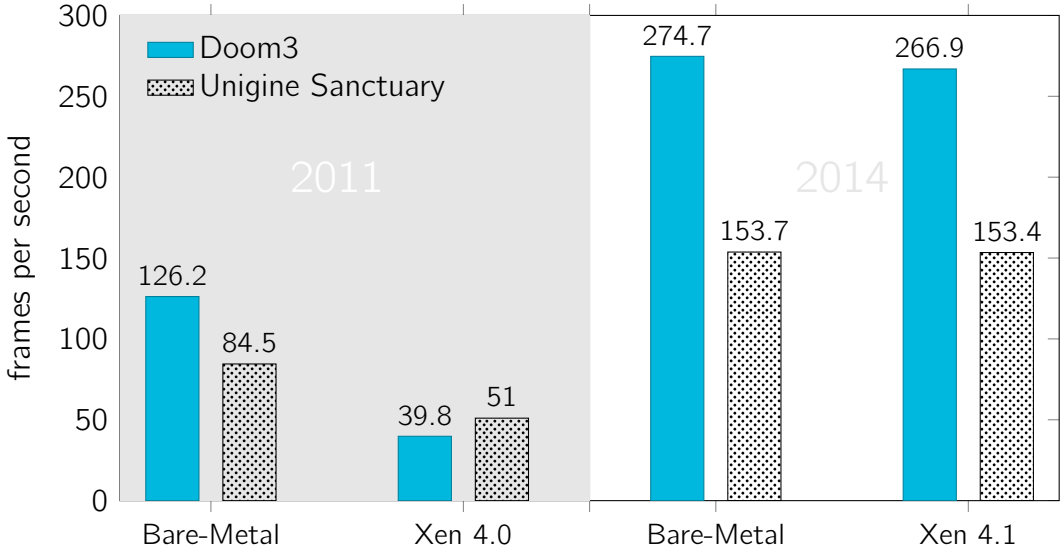
Poor performances on traditional GPU

- Degrades memory transfer (GPU bottleneck)
- Prevents Single-Instruction-Multiple-Data

Pass-through GPU enables better performances:

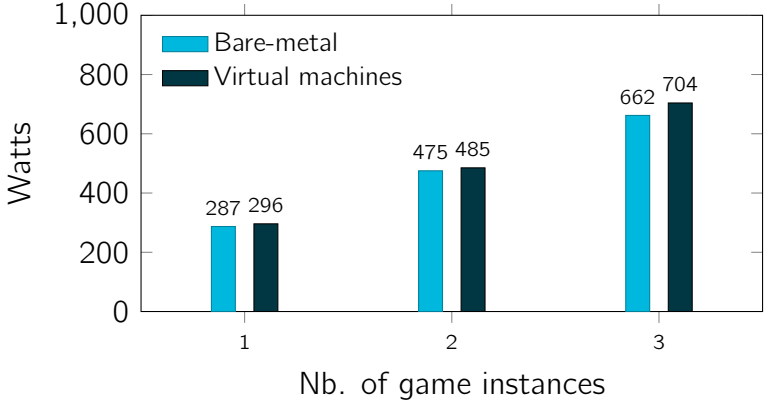
- One-to-one mapping between device and VM
- New technologies coming (*e.g.*, NVidia Grid)

Virtualization on GPU: performances



Game virtualization: consolidation

nb. of instances	Bare-metal (in fps)	VM (in fps)
1	30.2	29.2
2	30.0	28.9
3	29.8	28.9



Incompressible latency (without regards to network):

- Video encoding and decoding: 40 ms
- Game engine: from 15 to 50 ms

Network Latency for Cloud Platforms

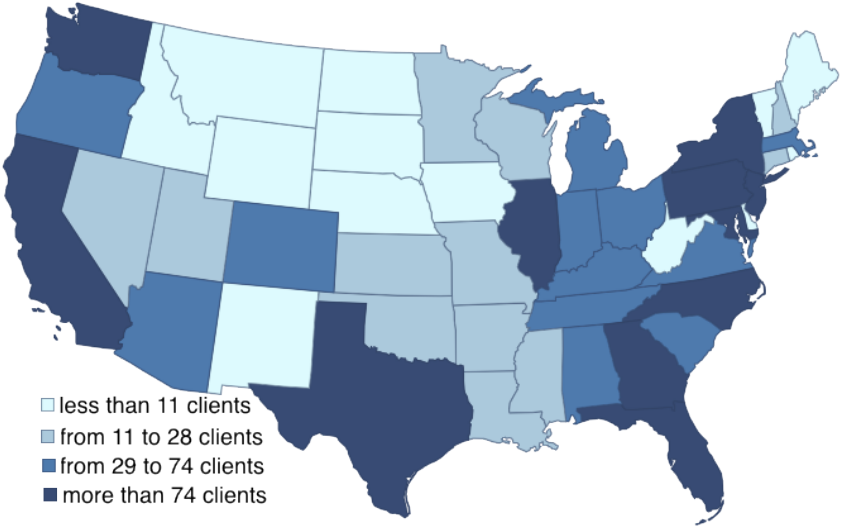
Our objectives

Can the current cloud support cloud gaming?

- Which percentage of users can play cloud games?
  - In excellent conditions (40 ms network latency)
  - In good conditions (80 ms network latency)

Our population sample

A modified BitTorrent crawler collected 2,504 IP addresses in Dec. 2012



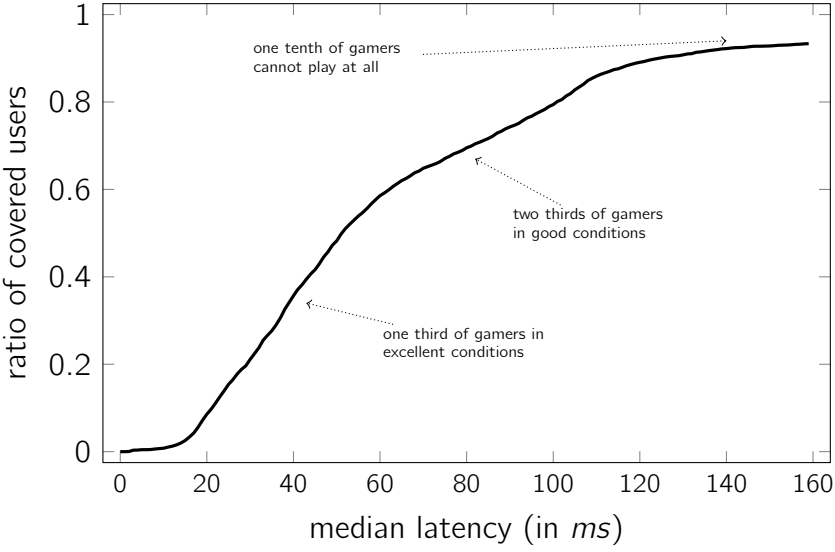
Latency to EC2: experiment

Amazon datacenters (Portland, Virginia, North Calif.)

For each client (IP address)

1. Measure latency from each data-center
2. Pick the shortest latency
3. Do it ten times

Latency to EC2: results



What if more datacenters are used

44 servers all over the US ("PlanetLabs" nodes hosted in universities)

For each client (IP address):

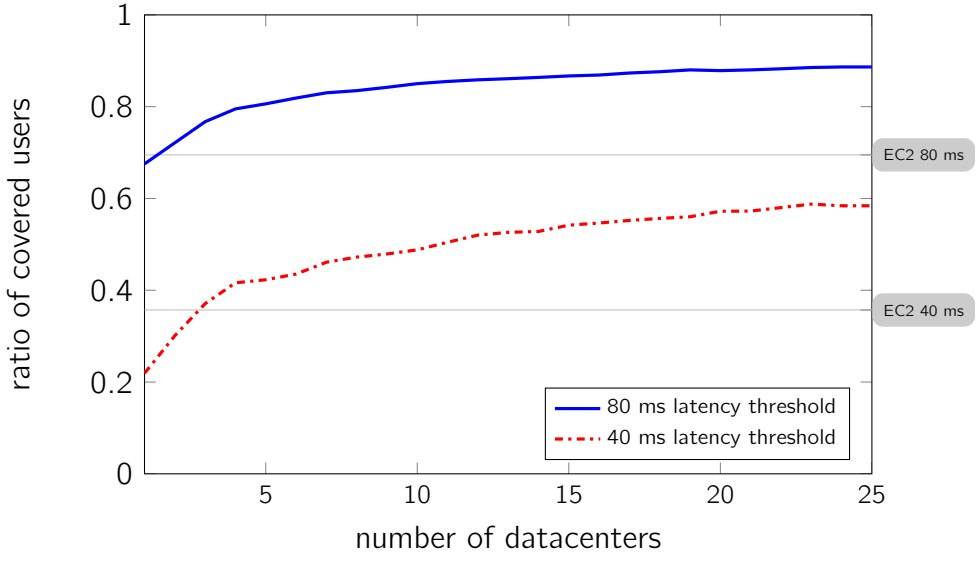
1. Measure latency from each PlanetLabs DC
2. Do it ten times

For each k-subset of datacenters:

1. Choose for each client the shortest latency
2. Sum all shortest latencies

Then pick the best k-subset of datacenters

More datacenters: results



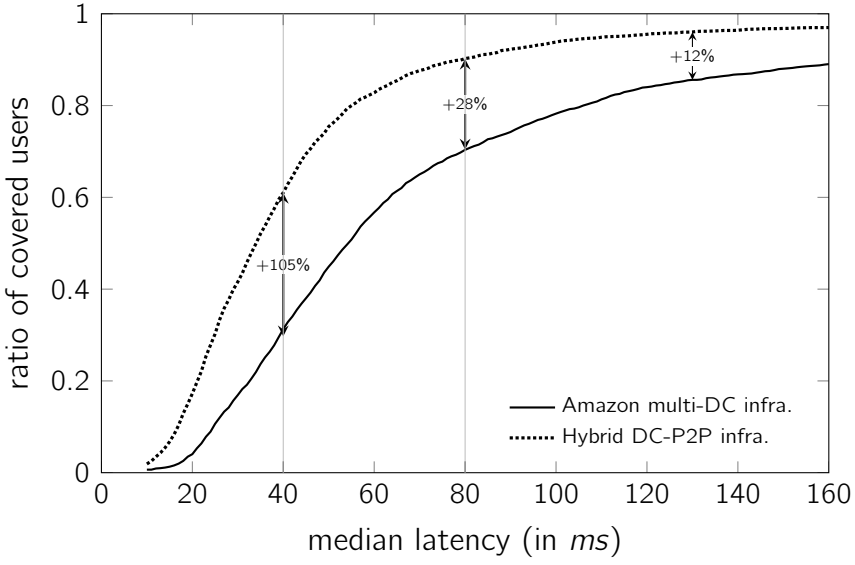
Out of the 2,504 clients:

- 1,000 are **edge servers**
- 1,504 are **clients**

For each client:

1. Measure latency to each Amazon datacenter
2. Measure latency to all edge-servers
3. Pick the shortest latency
4. Do it ten times

Edge cloud infrastructure: results



Network latency: summary

Not all users will be served

100 ms overall lag needs **a lot of** servers at the edges

Game placement on servers will be key

Table of contents

Simultaneous Releases of Key Technologies

omnidirectional cameras

panoramic videos

immersive devices

Classification of Expected Applications

Type	Examples	Implementation
360° scenes reacting to user command	gaming, 6DoF video, training simulation	game engine close to the end-users
passive navigable 360° scene (3DoF)	streaming, adult, sport, documentaries	Content Delivery Network (CDN)
live navigable 360° one-to-one call	emergencies, visiophone, online assistance	an integrated delivery chain including proxies

Today: Arcade Rooms

Requirements for 360-Degree Streaming

Bandwidth: bit-rate necessary to transport 360° multimedia flow

- 4k resolution in the viewport → 16k resolution for the full video
- at least 90 frames per second

⇒ minimum 500 Mbps?

Latency: delay between action and visible result

- head moves without visible change → motion sickness

⇒ maximum 10 ms!

Viewport-Adaptive Streaming

- **Existing (and considered) solutions**
  - Stream the whole 360-degree video

viewport	resolution	bandwidth
HD	6K	>80 Mbps
UHD	8K	>120 Mbps
4K	16K	>500 Mbps

- **Advantages**:
  - Interactive
  - DASH compliant today
- **Weaknesses**:
  - Bandwidth consumption

Viewport-Adaptive Streaming

- **Existing (and considered) solutions**
  - Stream only the user viewport

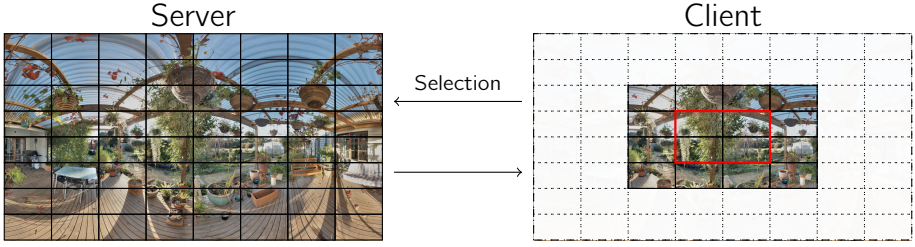
Action	Latency
Request to the server	RTT/2
Viewport extraction	5 ms
Re-encoding	10 ms
Transmission	RTT/2
Decoding	7 ms

- **Advantages**:
  - Bandwidth optimization

- **Weaknesses**:
  - Computation at the server

Viewport-Adaptive Streaming

- **Existing (and considered) solutions**
  - Split the 360-degree video into independent tiles selected by the client



- **Advantages**:
  - Good interactivity
  - Flexibility
  - Compliant with DASH SRD
- **Weaknesses**:
  - Client side processing
  - Storage at server
  - Compression efficiency

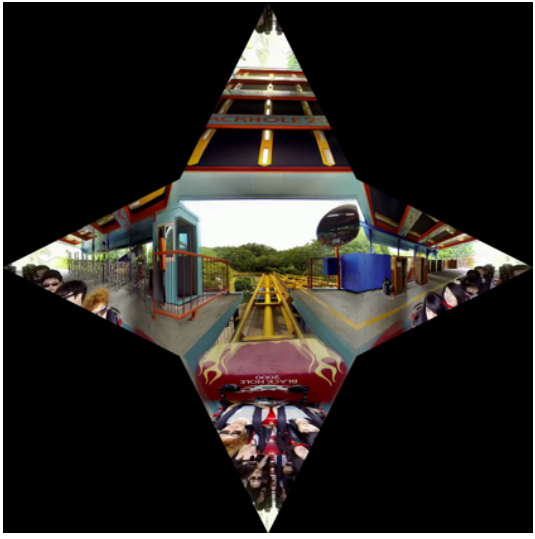
- **Our proposal**:
  - Introduce the notion of *Quality Emphasis Region* (QER)
  - Predict the head movements
  - Optimize video encoding by selecting the best QERs

Equirectangular

Equirectangular with 8 $\times$ 8 tiles packed with different resolutions

4 × 3 cubemap

Compact cubemap



Pyramid

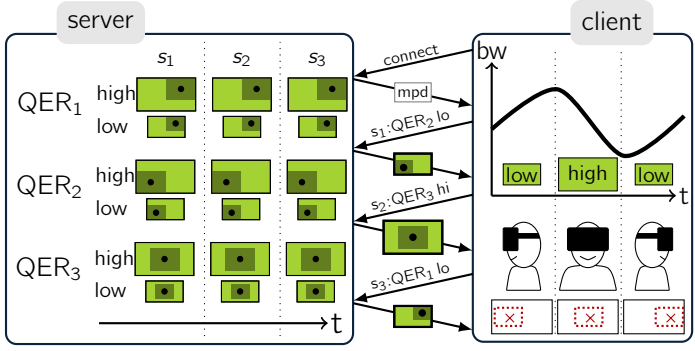
Compact pyramid

Standardization Efforts: OMAF at MPEG

Compact rhombicdodecahedron

Ongoing work under discussion:

- Signalling a Region of Interest (RoI)
- Sharing viewing statistics to enable accurate head movement prediction
- Interfacing with Server and Network-Assisted Delivery (SAND)



Standardization Efforts: Protocols

WebRTC at W3C: no specific action for 360° video yet

- Congestion control for high-bandwidth and low-latency videos

HTTP/2 at IETF: various improvements for low-latency videos

- Server push to anticipate requests from clients
- Multiplexing: priority and cancelling of video parts

MPTCP and **QUIC** at IETF: dealing with transport layer issues

- Cross-layer approaches to avoid useless retransmission

Table of contents

Support for multimedia applications

- Specific needs are addressed by specific solutions
- Requirement of high computing power

Support for interactive applications

- Response time depends on a chain of processes
- Anticipating behaviors and deal with priority