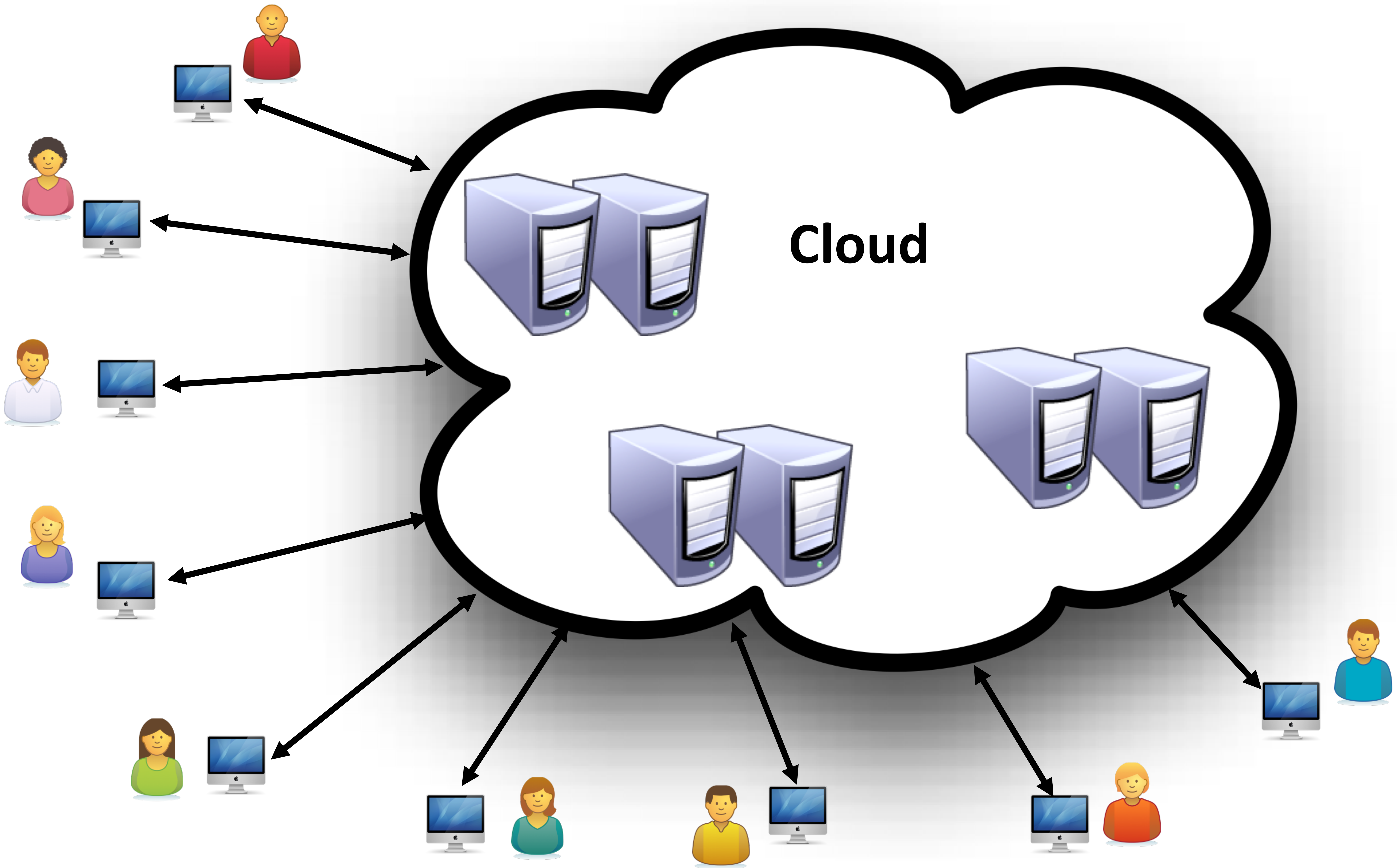
Munich Internet Research Retreat Raitenhaslach (MIR^3)

Pruning Edge Research with Latency Shears

Nitinder Mohan
Technical University Munich

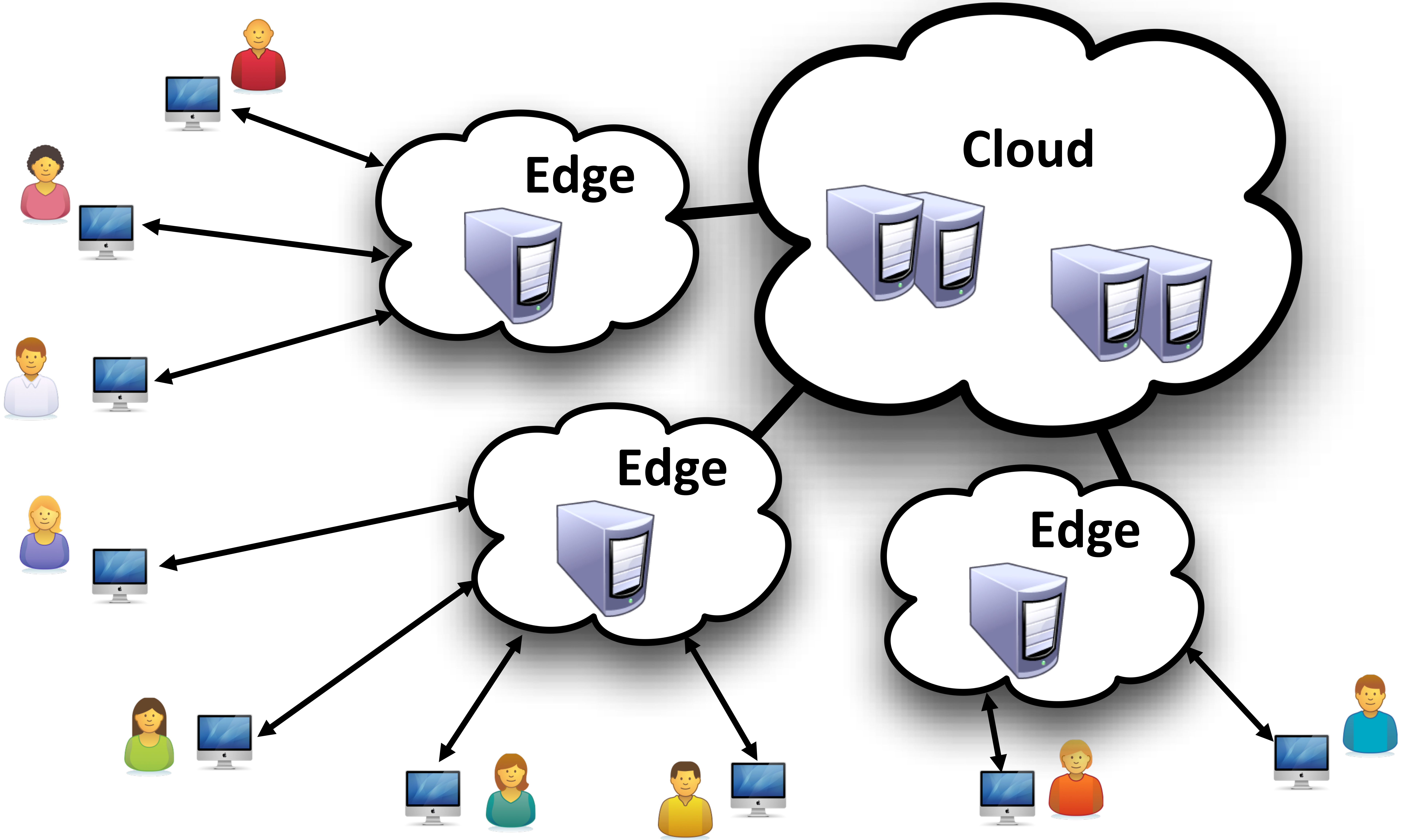
mohan@in.tum.de

Centralization to Clouds



Cloud

Decentralize to Edge

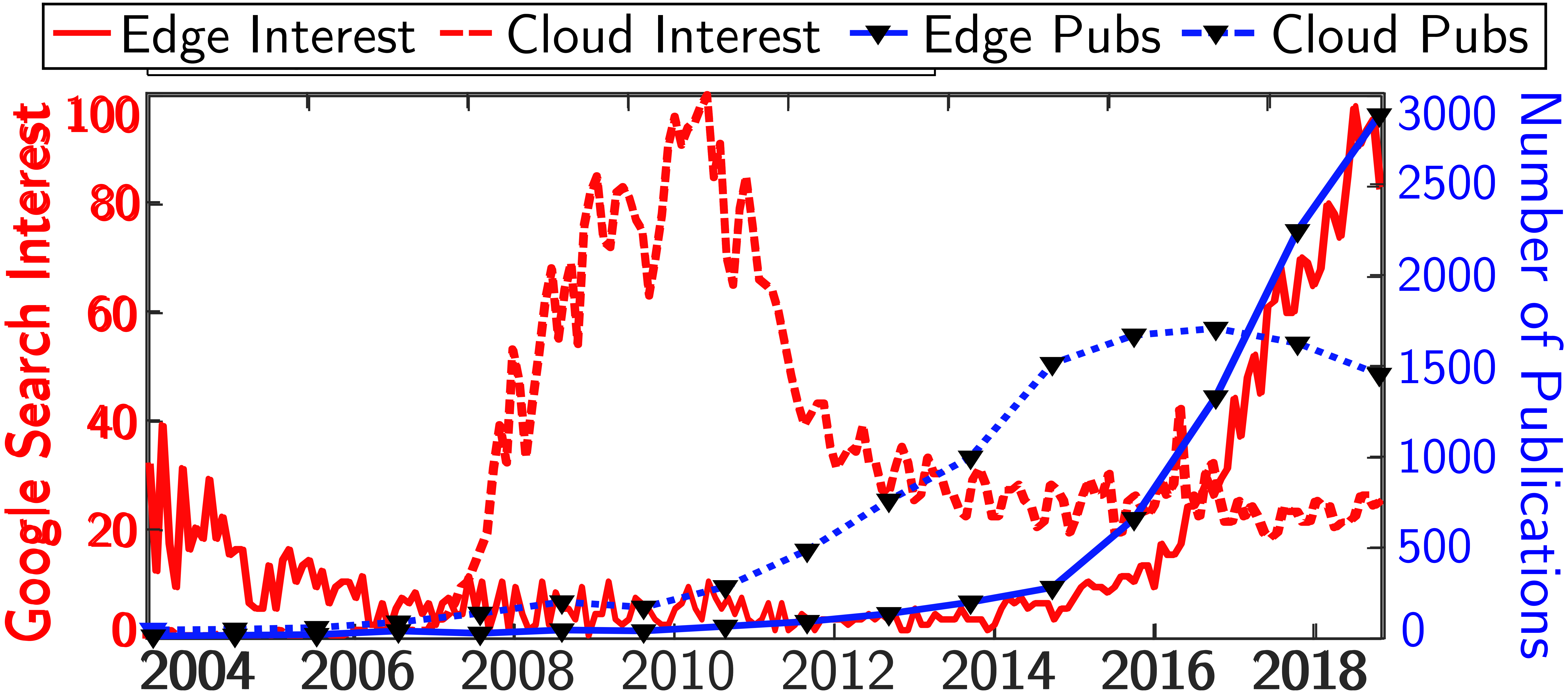


Why Edge?

“Claimed” Selling Points:

- Shorter latencies for clients
- Less network traffic towards the cloud
- Less processing at the cloud
- Better privacy via local processing
- …

It almost feels like hype!



In general public

and

research community

Why Edge?

"Claimed" Selling Points:

- Shorter latencies for clients
- Less network traffic towards the cloud
- Less processing at the cloud
- Better privacy via local processing
- …

Main proponents of Edge hype!

Are we ready for this tranformation?

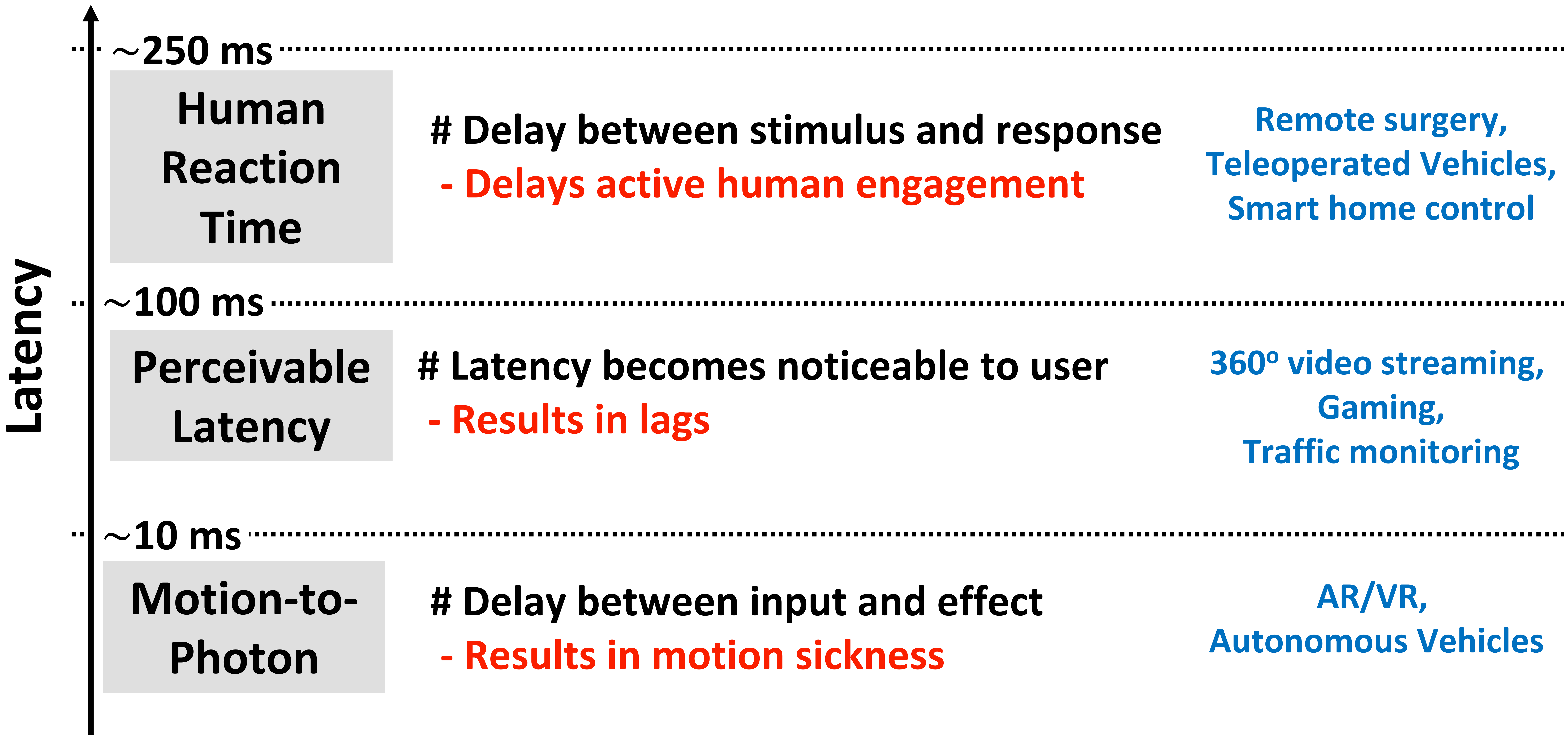
Cloud
Edge
Cloud
Edge
Edge

Is cloud at it's limit to support requirements of emerging applications?

Question 1

What are the latency requirements of ‘edge applications’?

‘Edge applications’ are driven by human limits!



Latency

~250 ms

Human Reaction Time

# Delay between stimulus and response

- Delays active human engagement

Remote surgery,
Teleoperated Vehicles,
Smart home control

~100 ms

Perceivable Latency

# Latency becomes noticeable to user

- Results in lags

360º video streaming,
Gaming,
Traffic monitoring

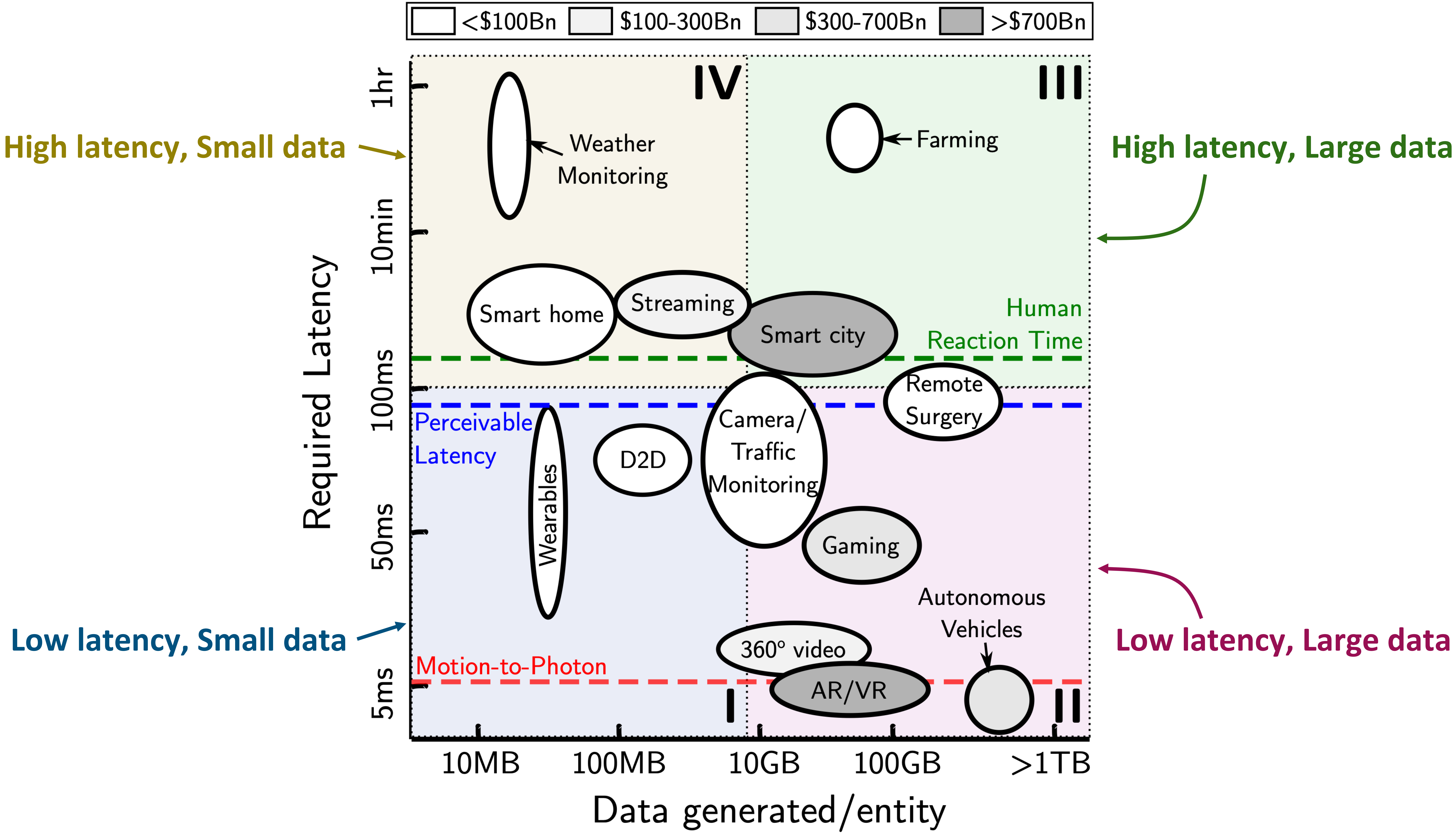
~10 ms

Motion-to-Photon

# Delay between input and effect

- Results in motion sickness

AR/VR,
Autonomous Vehicles


□ <$100Bn
□ $100-300Bn
□ $300-700Bn
■ >$700Bn
IV
III
I
II
High latency, Small data
High latency, Large data
Low latency, Small data
Low latency, Large data
Weather Monitoring
Farming
Smart home
Streaming
Smart city
Human Reaction Time
Remote Surgery
Perceivable Latency
Wearables
D2D
Camera/ Traffic Monitoring
Gaming
Autonomous Vehicles
360° video
AR/VR
Motion-to-Photon
Required Latency
1hr
10min
100ms
50ms
5ms
10MB
100MB
10GB
100GB
>1TB
Data generated/entity

Question 2

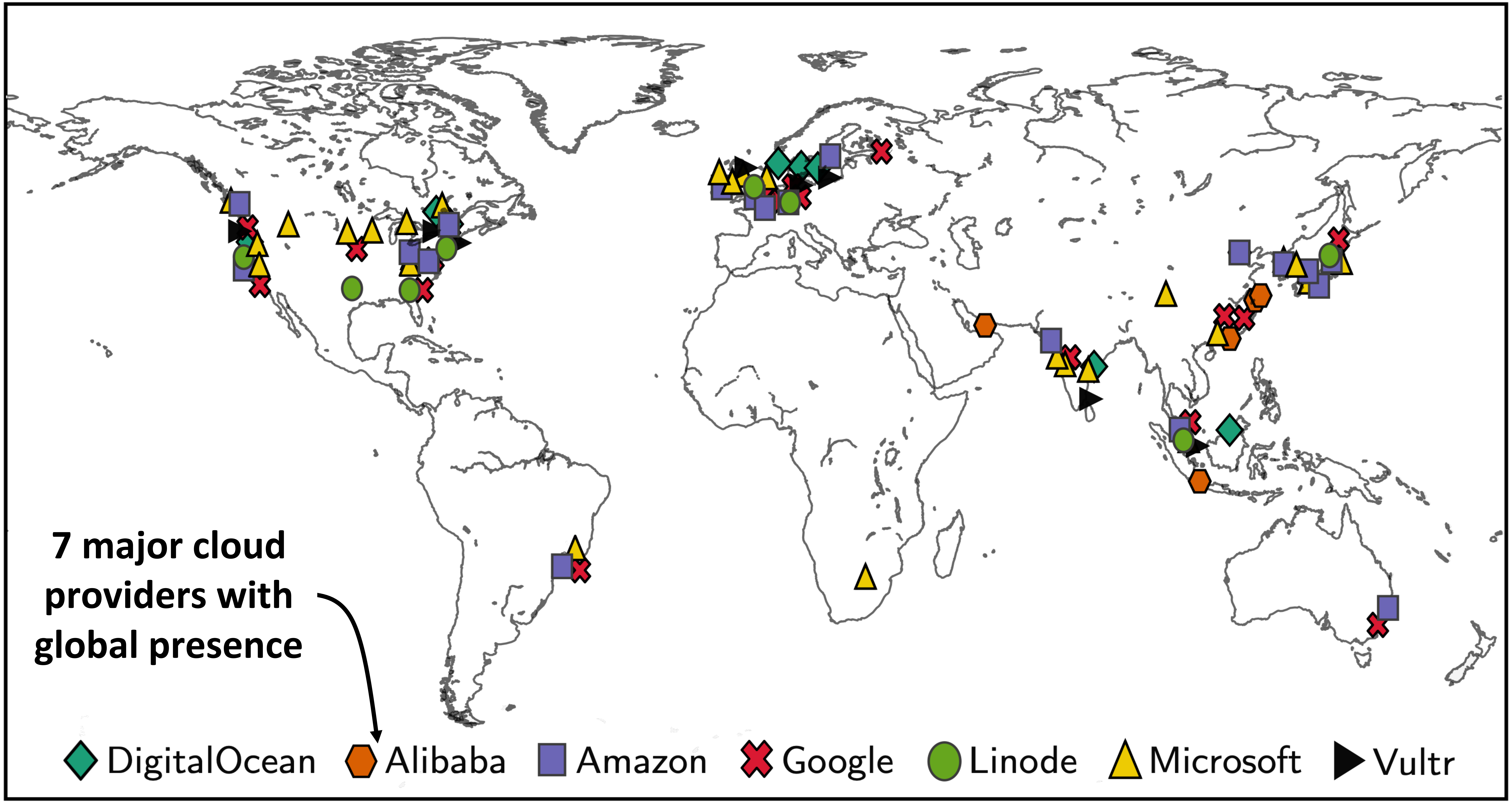
How far is the cloud?

Ongoing research since September 2019

Snapshot of work published in **ACM HotNets 2020**

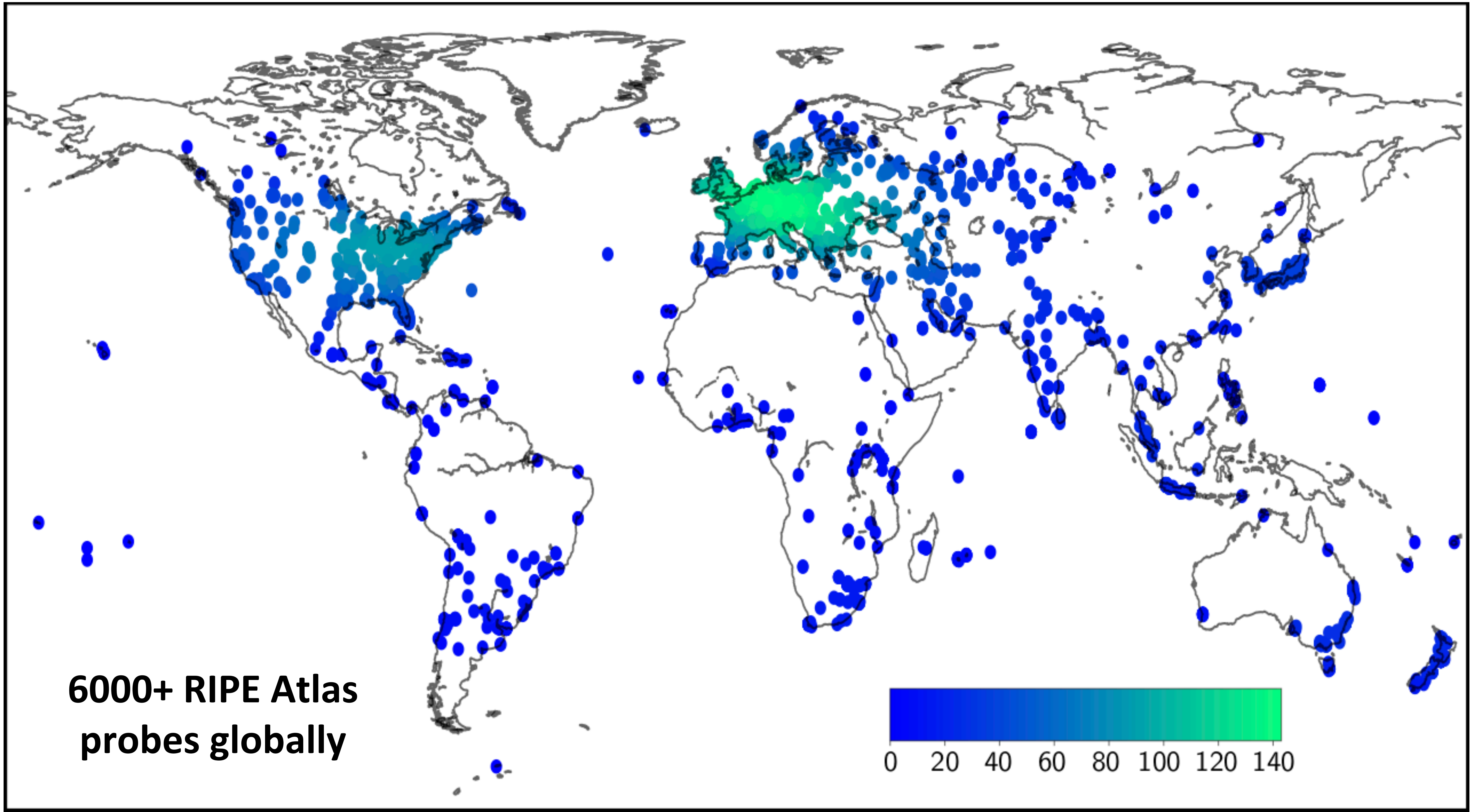
Pruning Edge Research with Latency Shears; Nitinder Mohan (Technical University of Munich), Lorenzo Corneo (Uppsala Universitet), Aleksandr Zavodovski (University of Helsinki), Suzan Bayhan (University of Twente), Walter Wong and Jussi Kangasharju (University of Helsinki)

Where Is the Cloud?



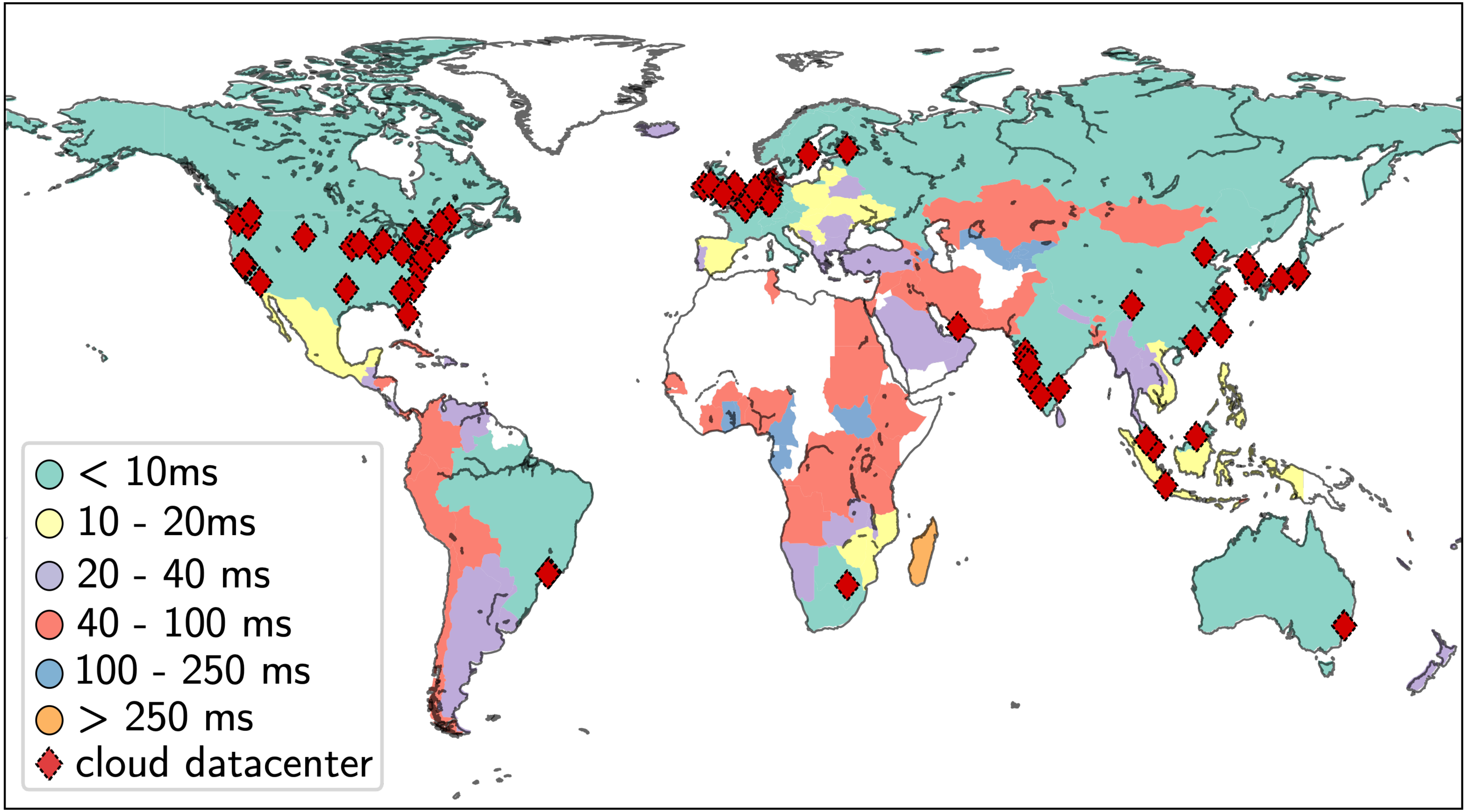
7 major cloud providers with global presence

Vantage Points

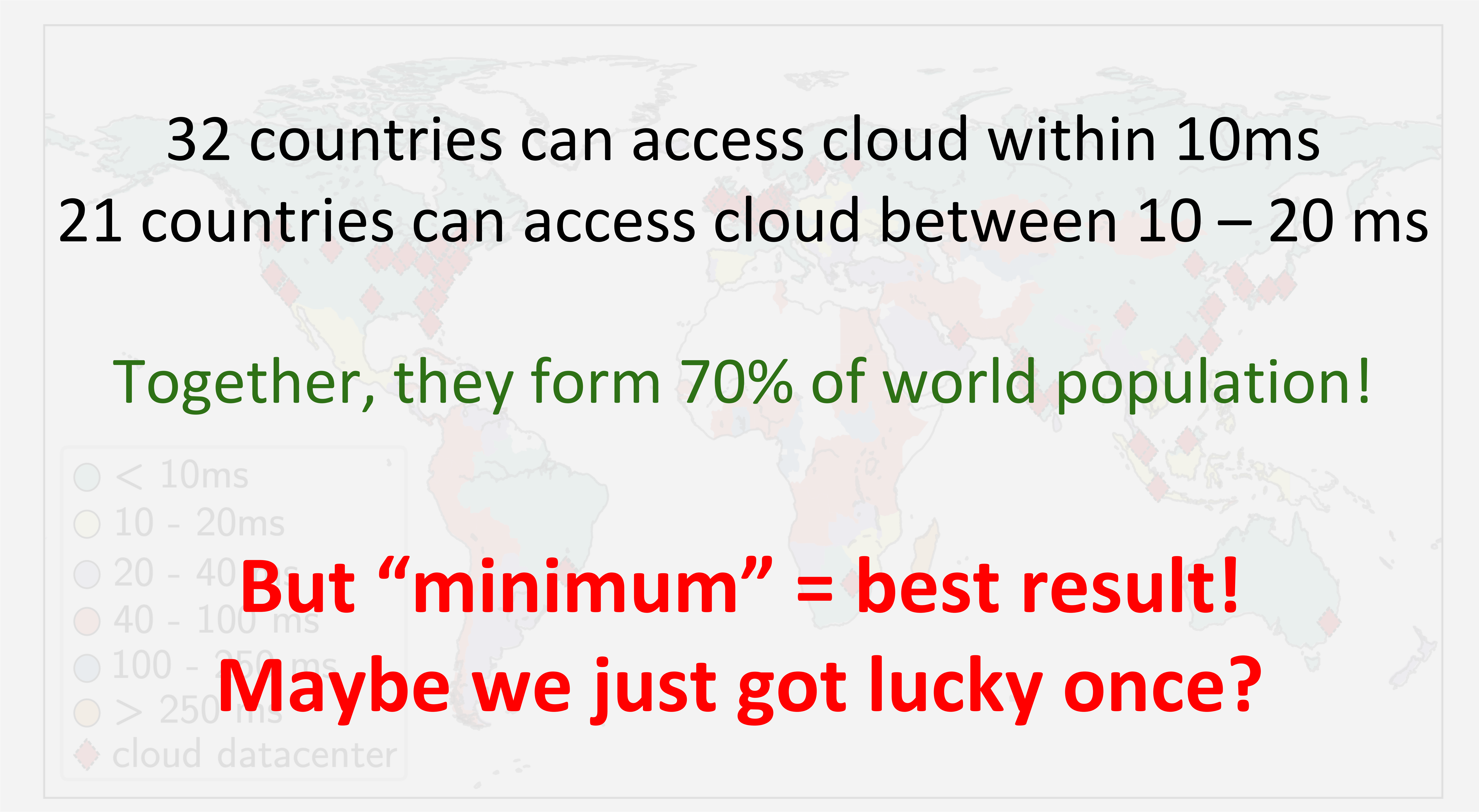


6000+ RIPE Atlas probes globally

What is the least possible latency to cloud?



What is the least possible latency to cloud?



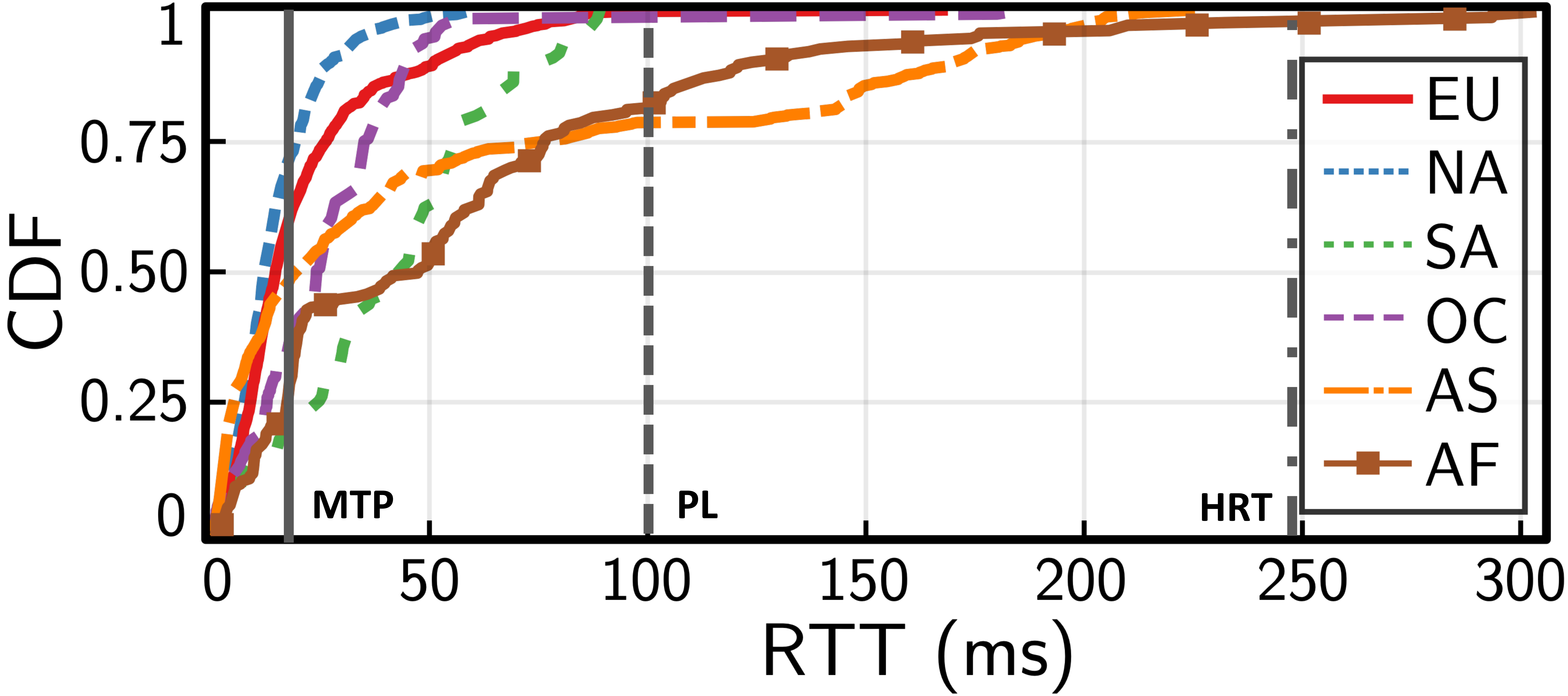
32 countries can access cloud within 10ms
21 countries can access cloud between 10 – 20 ms

Together, they form 70% of world population!

But "minimum" = best result!
Maybe we just got lucky once?

Actually, We Got Lucky More Than Once!

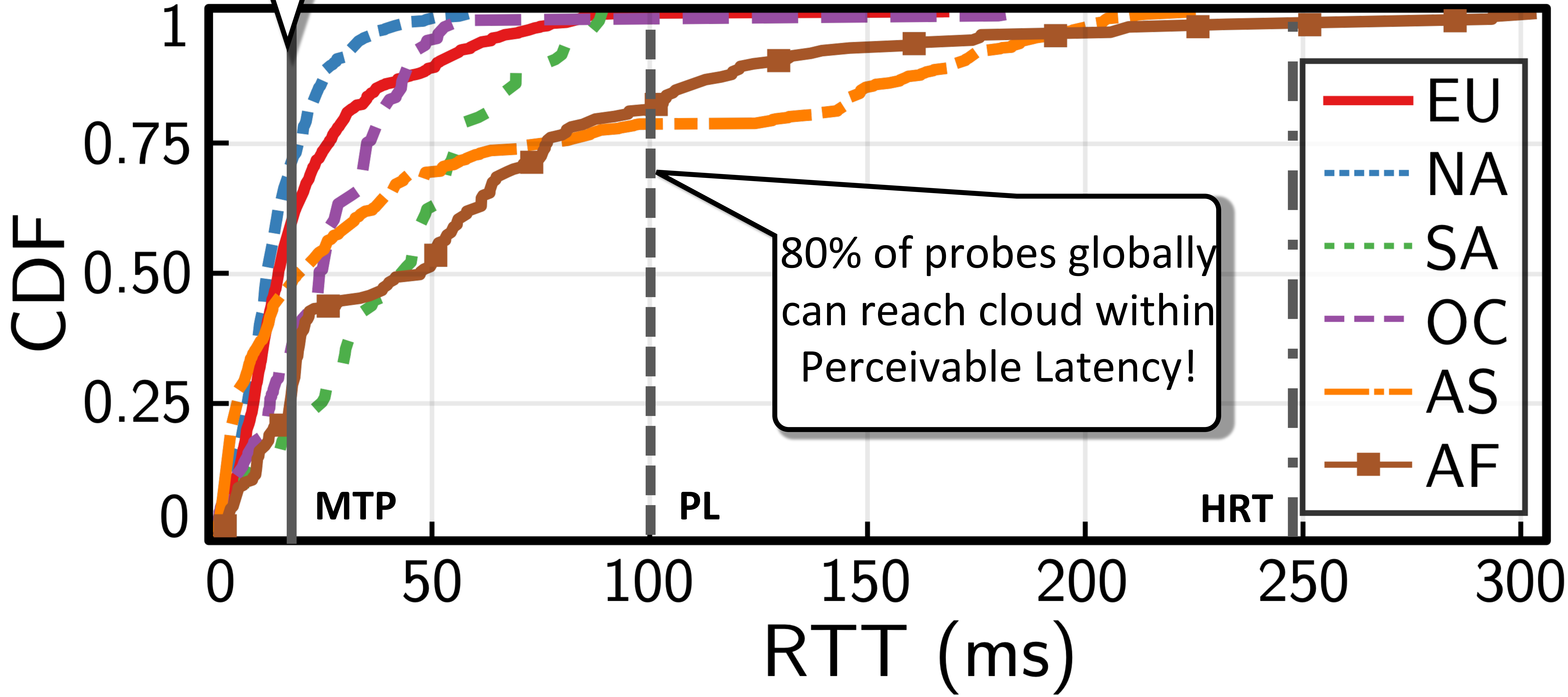
Minimum latencies for all 6000+ probes



Ve Got Lucky More Than Once!

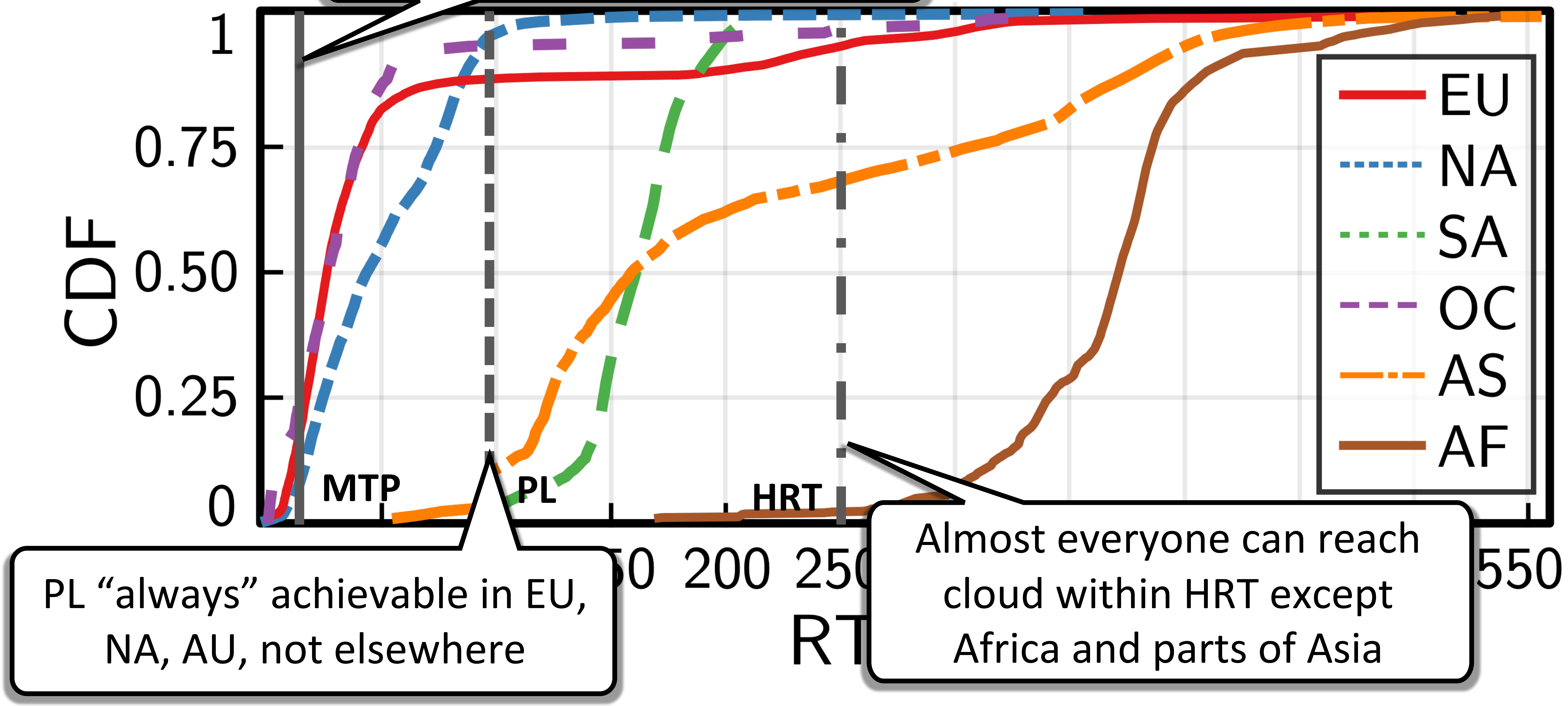
imum latencies for all 3000+ probes

Motion-to-Photon achievable for 70% of probes in EU, NA and 30% globally



80% of probes globally can reach cloud within Perceivable Latency!

ent Data

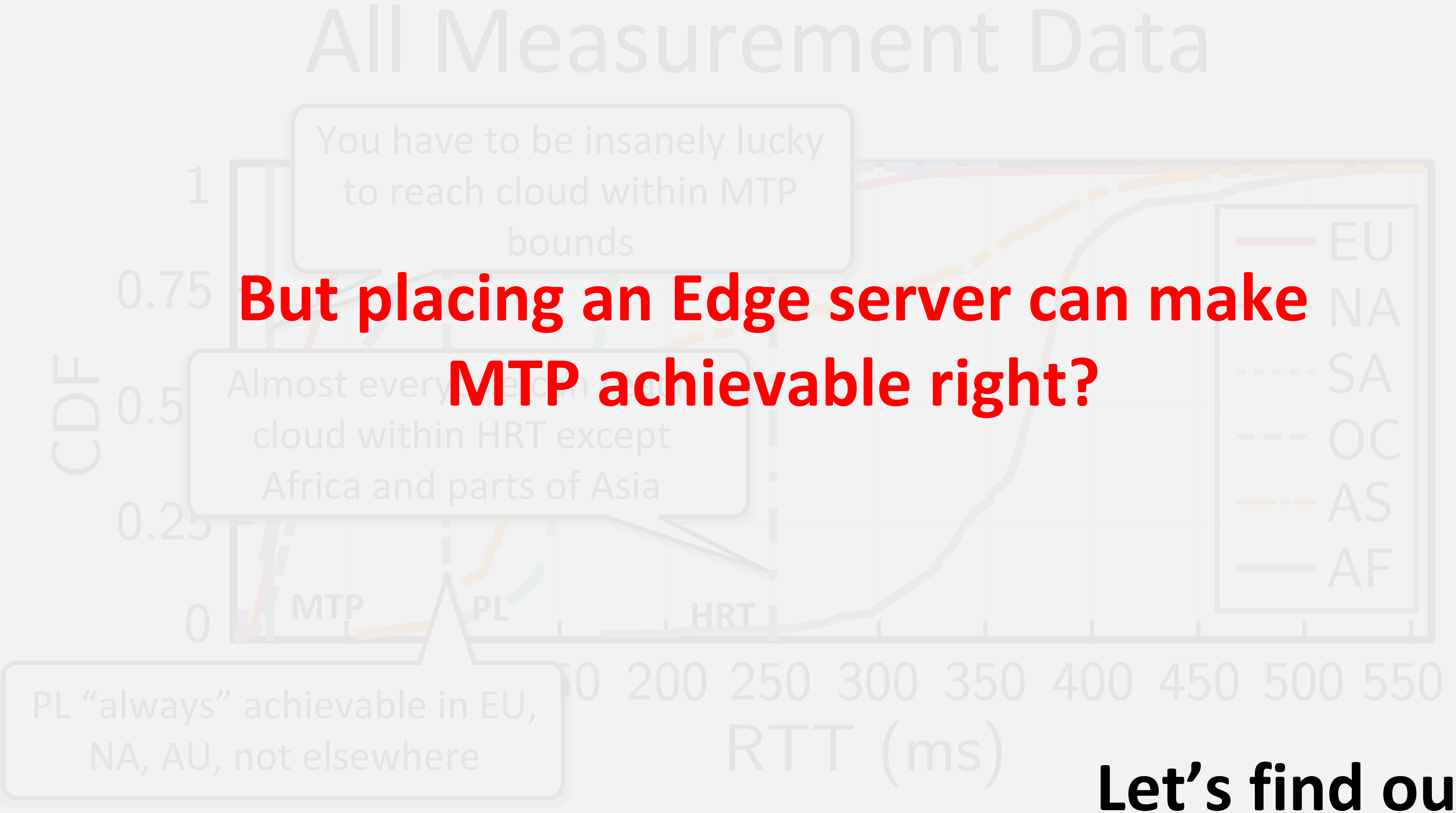


You have to be insanely lucky to reach cloud within MTP bounds

PL "always" achievable in EU, NA, AU, not elsewhere

Almost everyone can reach cloud within HRT except Africa and parts of Asia

All Measurement Data



You have to be insanely lucky to reach cloud within MTP bounds

Almost every[illegible] cloud within HRT except Africa and parts of Asia

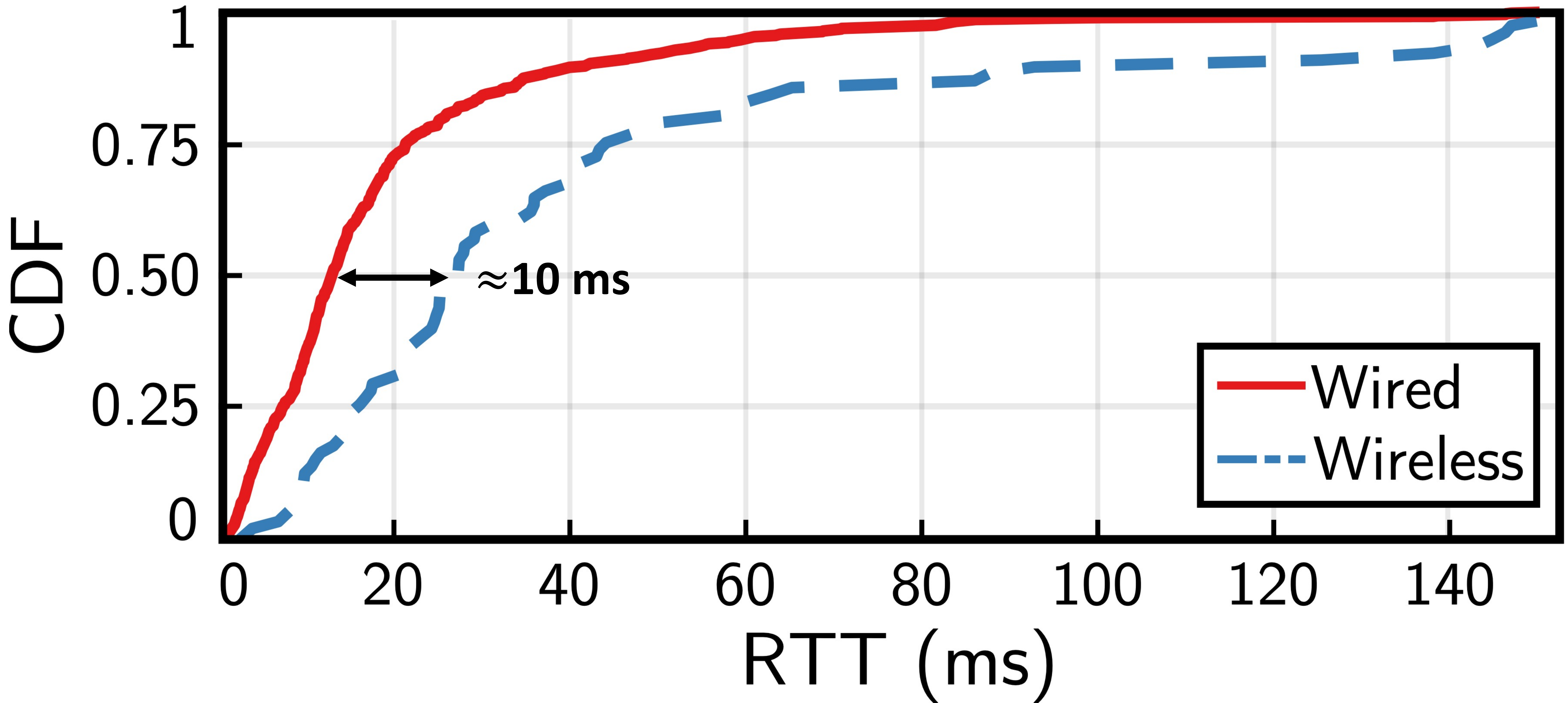
PL "always" achievable in EU, NA, AU, not elsewhere

But placing an Edge server can make MTP achievable right?

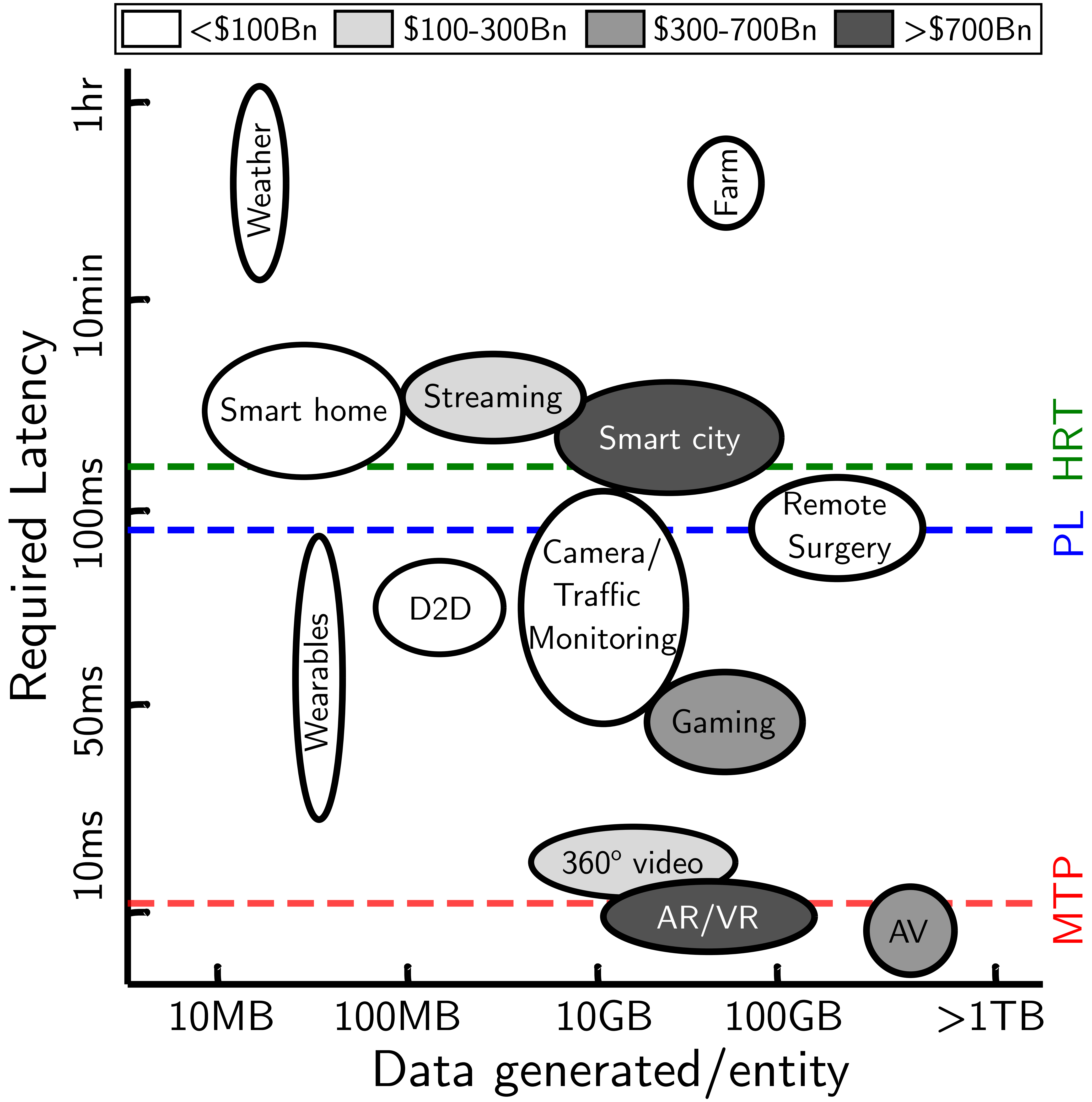
Let's find out…

Wired vs. Wireless

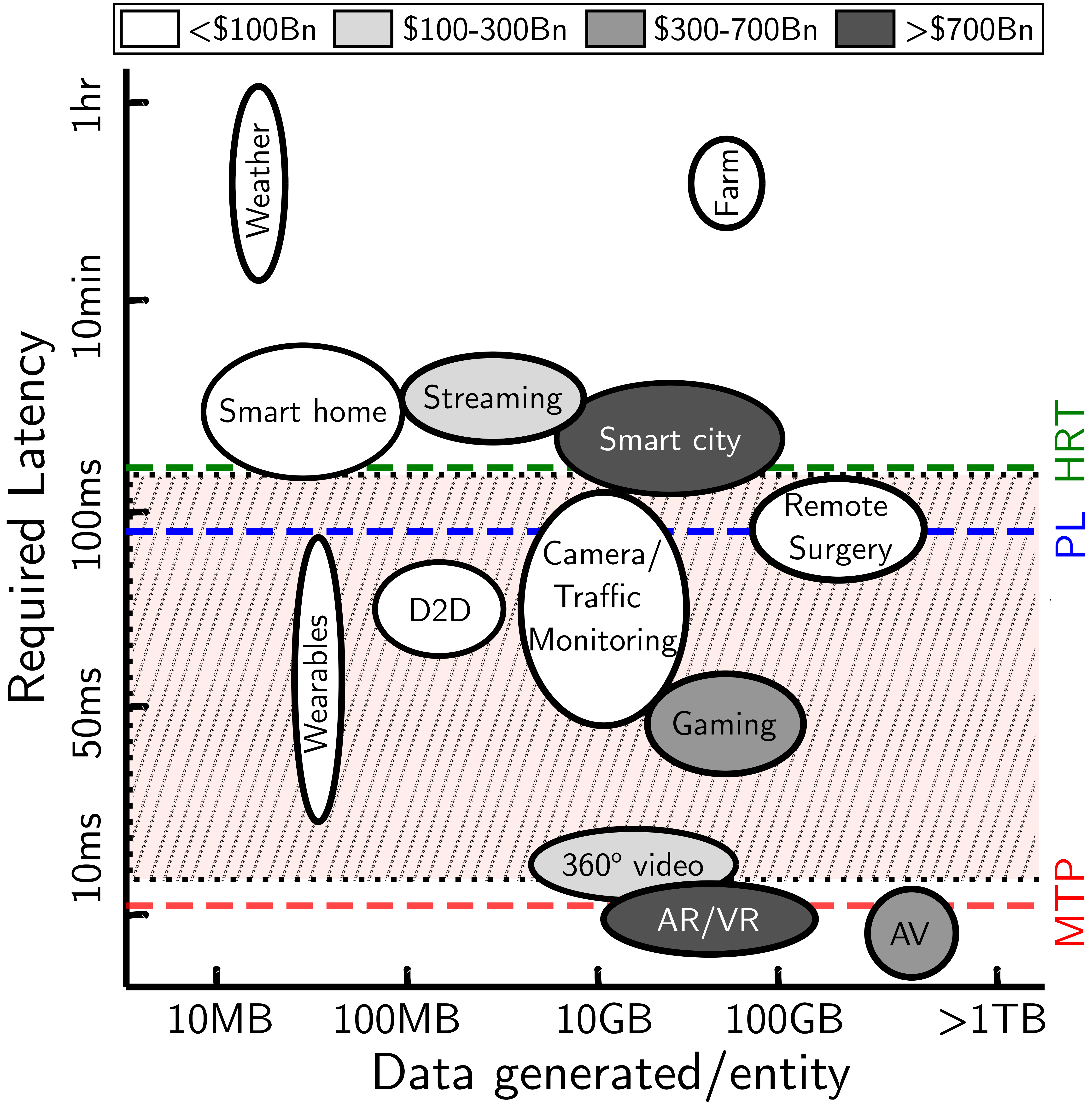
Minimum latencies from probes in **same location** connecting to **same datacenter** but via different last-mile



Revisiting Edge-Enabling Applications



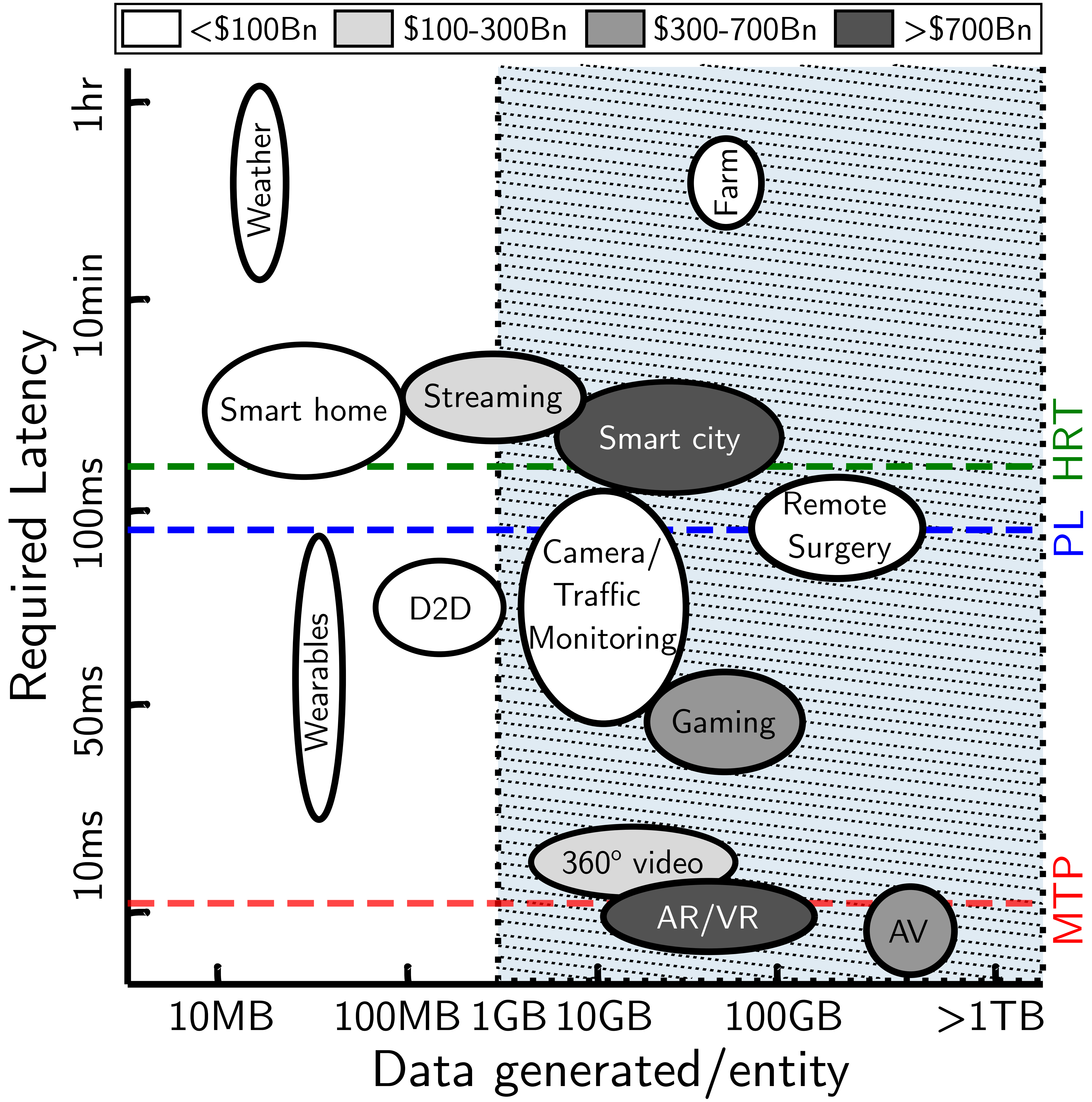
Revisiting Edge Applications



Latency Feasibility Zone

- Lower threshold is 10 ms limited by current wireless last-mile access performance
- Higher threshold is Human Reaction Time as current cloud deployment can easily support it

Revisiting Edge Applications



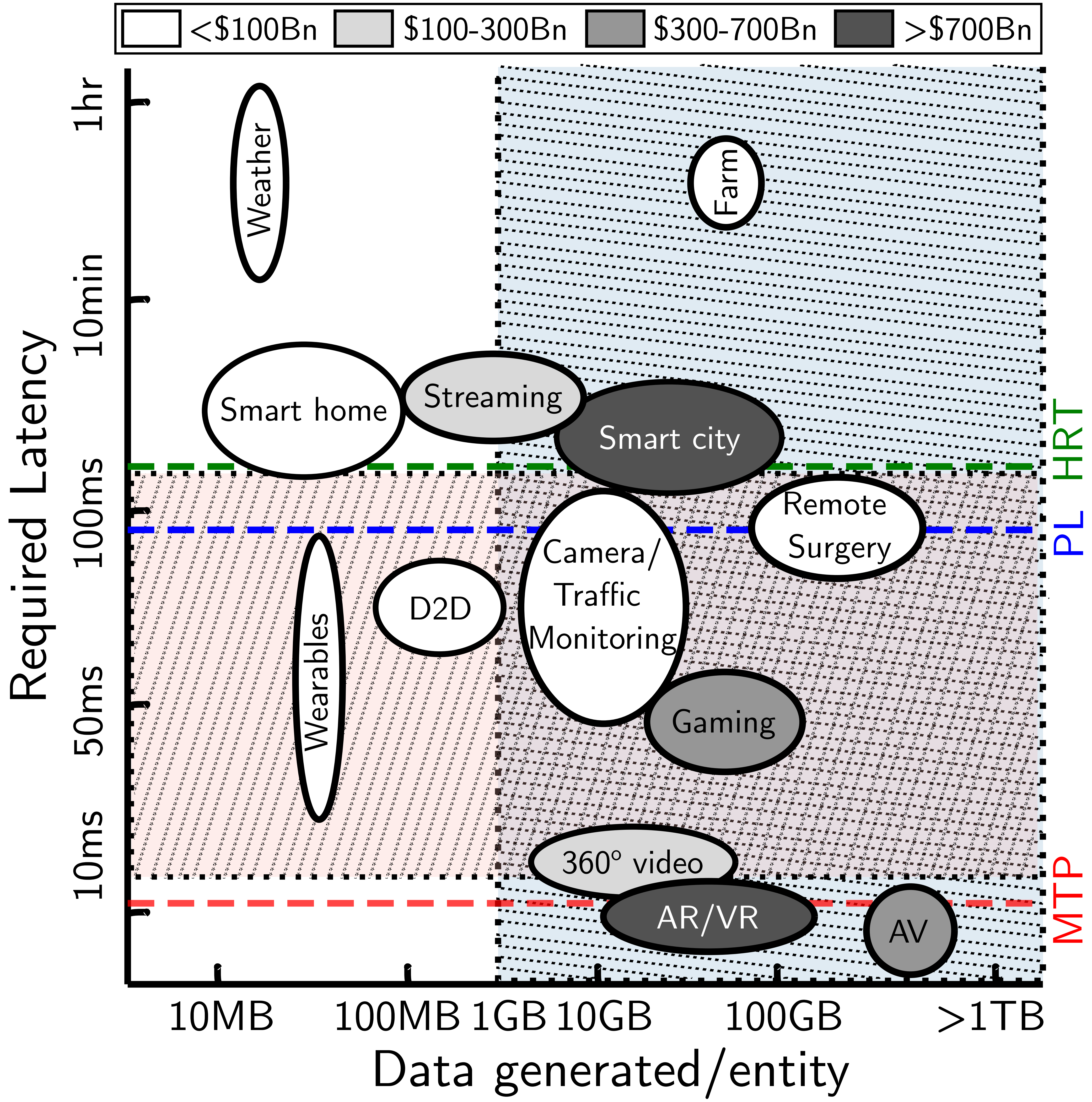
Bandwidth Feasibility Zone

Bandwidth aggregation gains for Edge doesn’t make sense for sensors producing small data volume

- Lower threshold is set at 1GB/sensor based on our measurements*
- Higher threshold is as much as possible

*Refer to our paper coming out soon

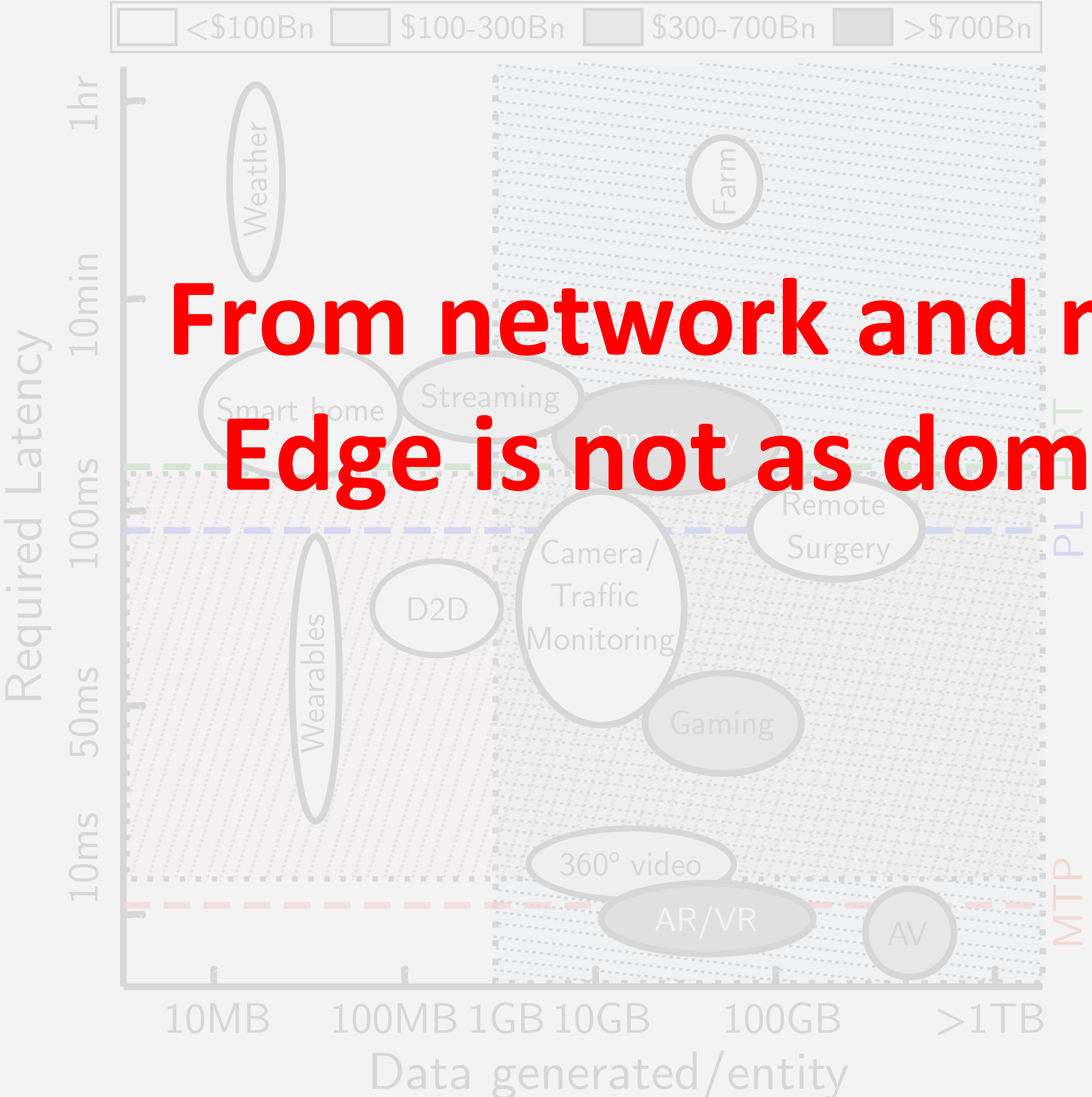
Revisiting Edge Applications



Edge Feasibility Zone

- Edge makes sense for only few applications
- Many hyped applications do not really benefit from edge
- Market share of “sweet spot” is relatively small

Revisiting Edge-Enabling Applications



Edge Feasibility Zone

- Edge makes sense for only few applications
- Many hyped applications do not really benefit from edge
- Market share of "sweet spot" is relatively small

From network and market perspective, Edge is not as dominant as assumed!

Is this a death knell for Edge computing?

Not really!

Proponents of Edge

Privacy via local processing

Trust and Security

Distributed AI

Bandwidth aggregation

▶ **The purpose of our work is to sway research perception away from "hype" around edge and towards areas where edge makes more sense**

Plugging in our **Limitations**

- More measurement platforms (e.g. Speedchecker) to remove platform biases and get more wireless connectivity perspective

- More cloud providers for more diversity in connectivity (Oracle, IBM, etc.)

- Measuring network performance of CDN-based cloud infrastructure (AWS Lambda)

Thank You!

mohan@in.tum.de

Backup

Research areas which can use some of that Edge “Hype”

Build consistent and faster cellular last-mile

- Last-mile is the biggest “bottleneck” for latency gains of Edge
- 5G promises 1 ms latency, same way 4G promised 10 ms at release
- *wink* at 6G technology makers

Focus attention on poorly-connected regions

- Majority of Africa and parts of Asia cannot access the cloud in reasonable latency
- However, these regions have a growing economy along with an already set-up cellular infrastructure
- Edge can do wonders here compared to already developed regions (US, Europe, etc.)

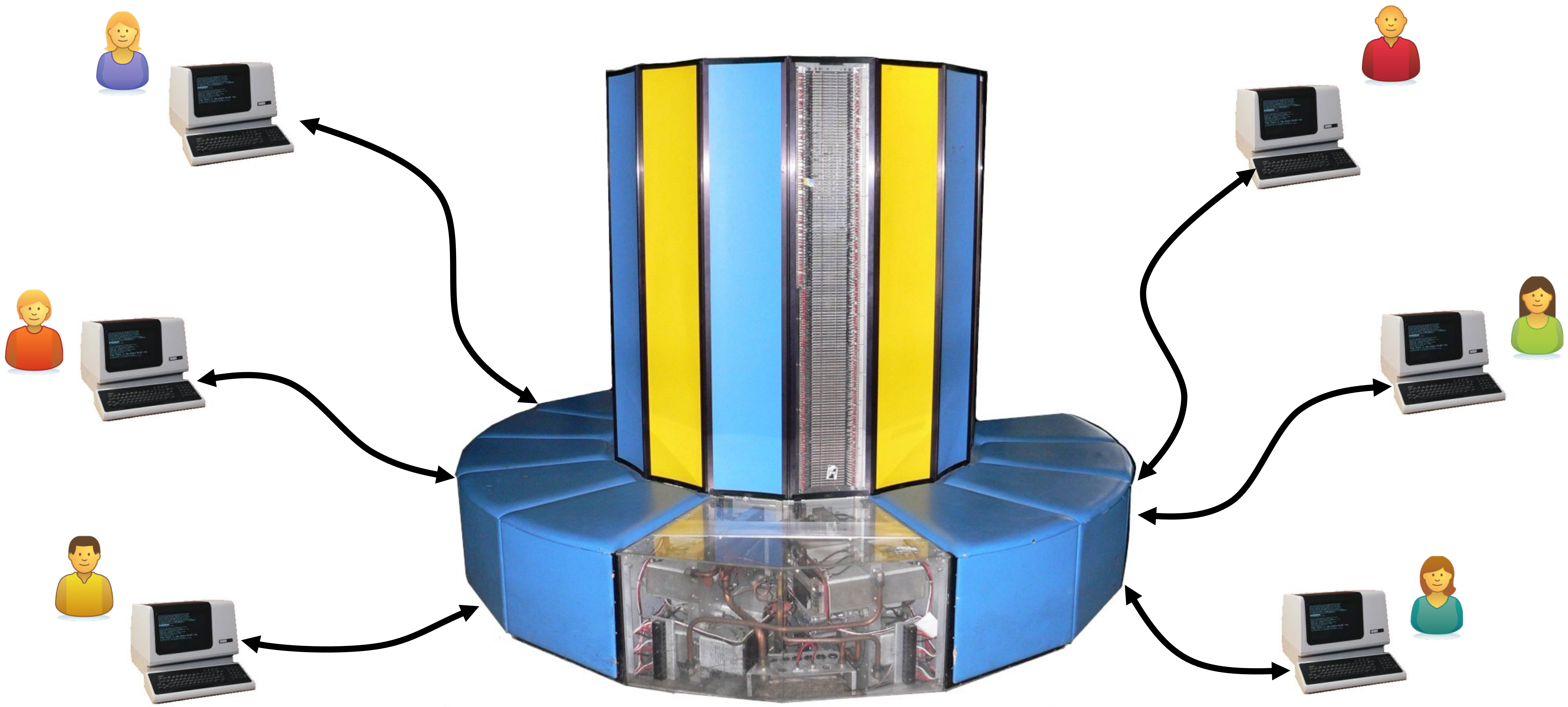
Cloud is a friend. Integrate it!

- Cloud providers are expanding their reach and improving the quality of their network
- Economies-of-scale is already working against Edge computing as deployment cost is high
- Seamless integration with existing cloud technologies and platforms adds some sense to Edge

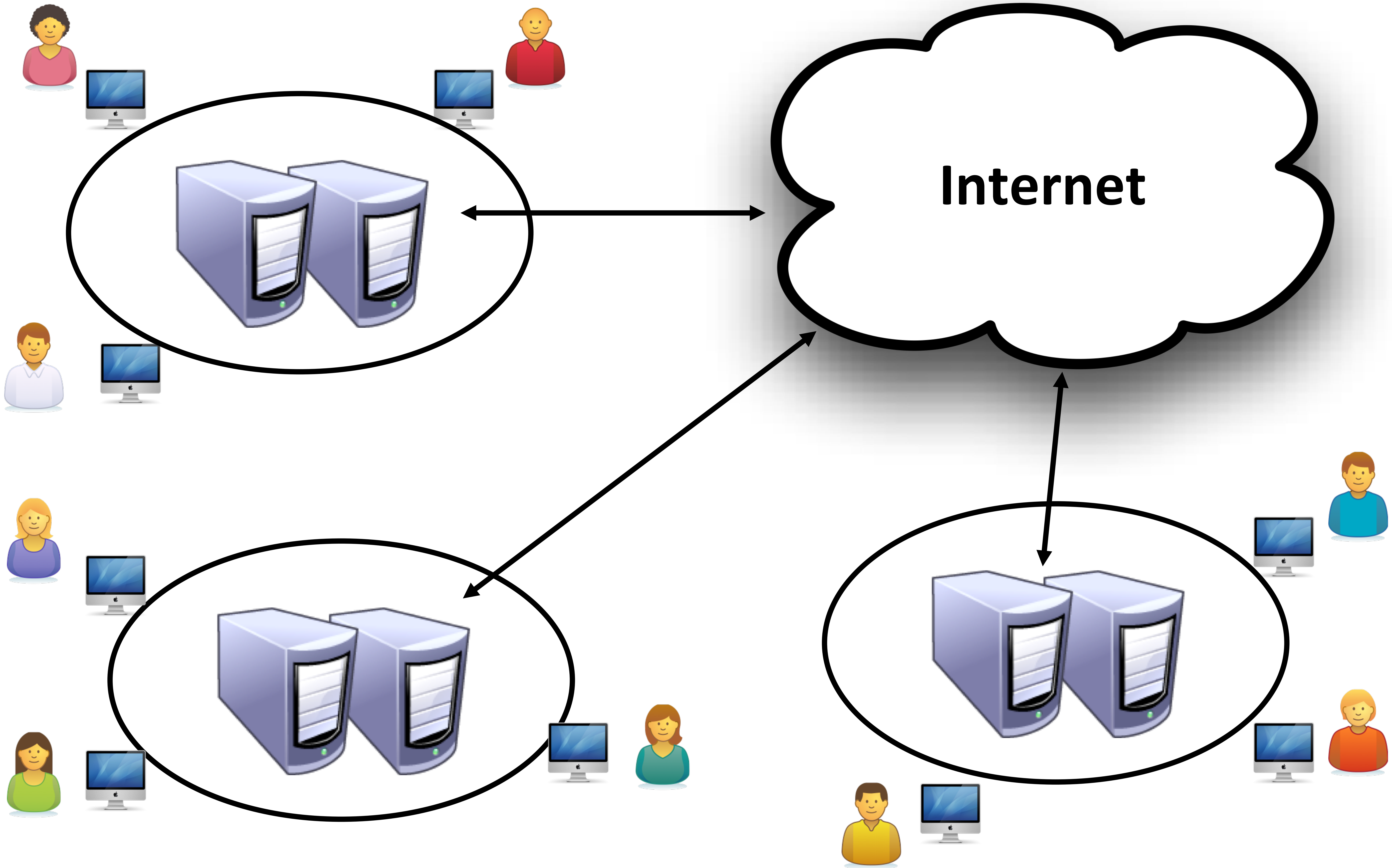
Outline

- Brief history of (edge) computing
- Does edge make sense from networking standpoint? Hype or Reality?
- Future?

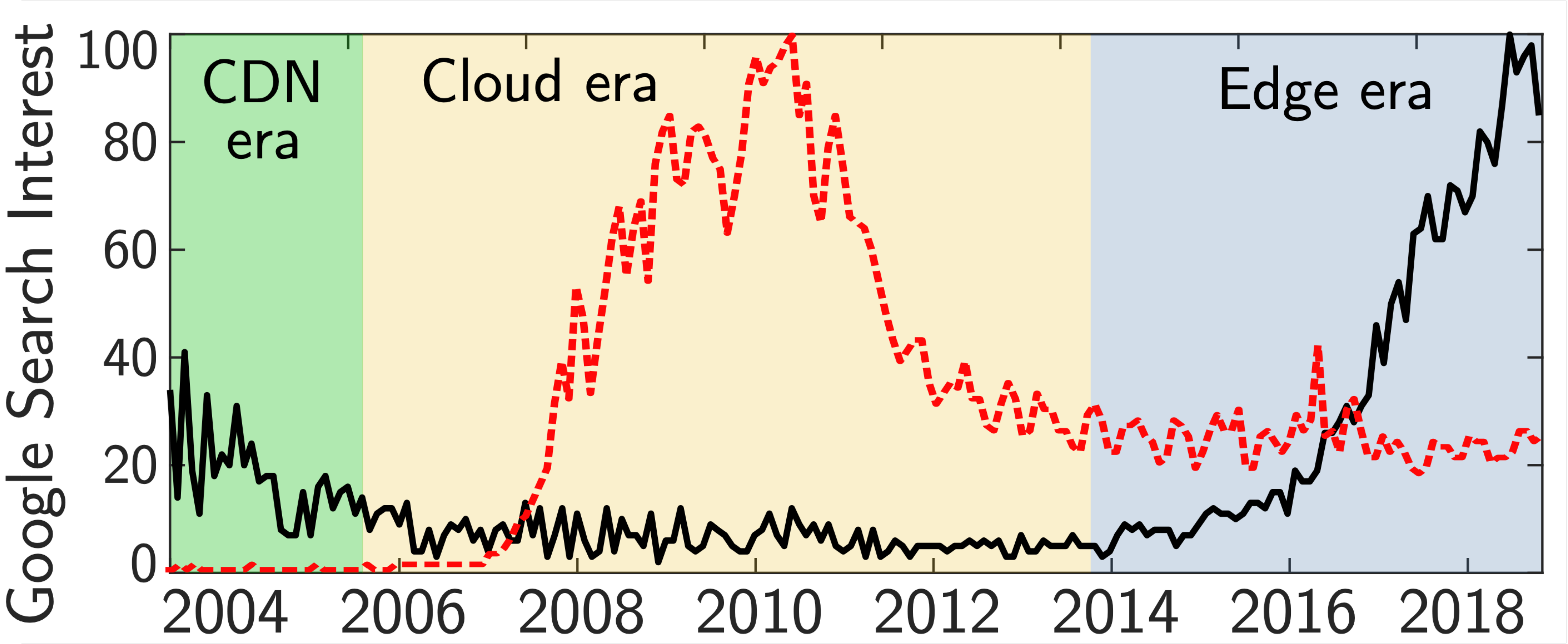
Mainframe Era

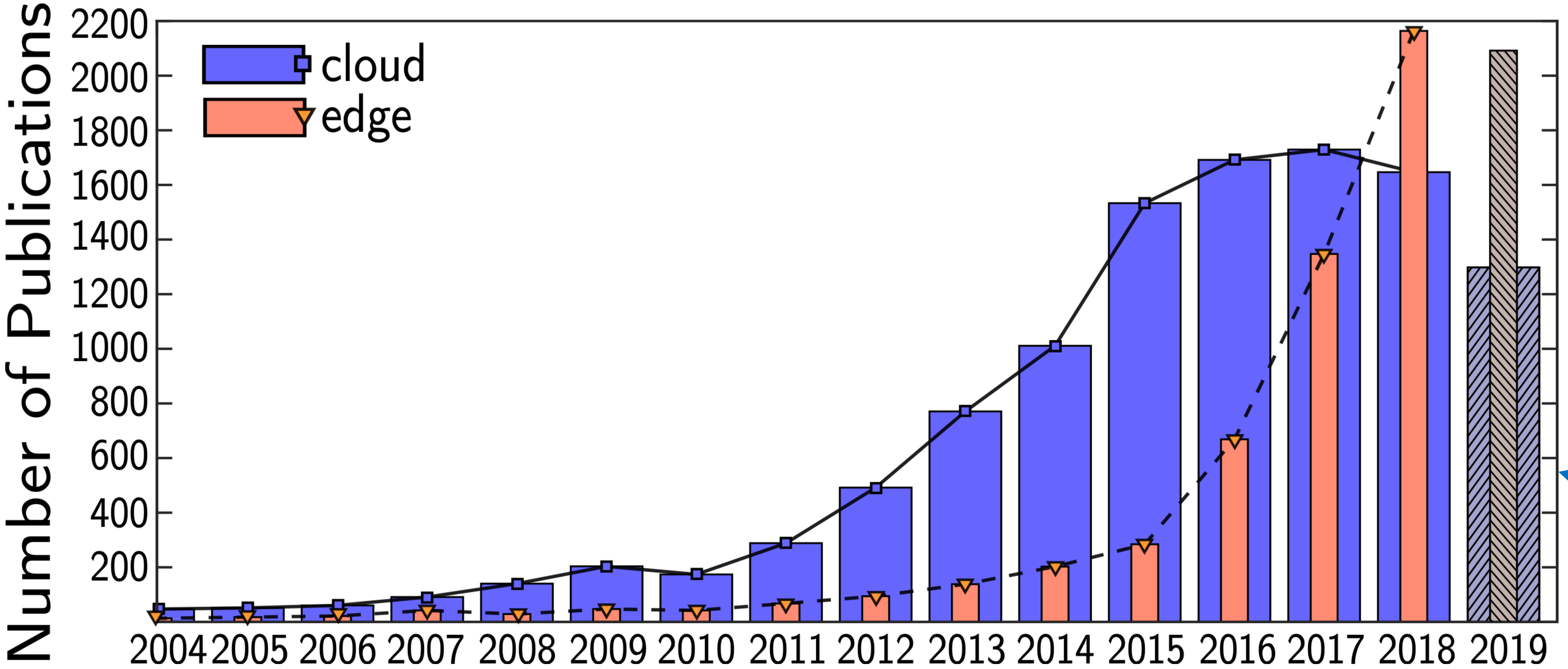


Back in 20th Century …



Atleast there is HYPE around it!





In general public
and
research community

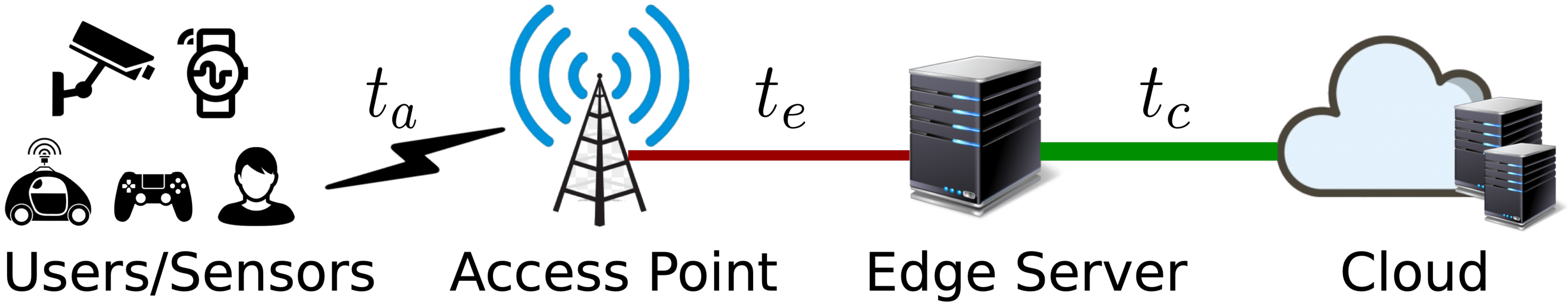
Why Edge?

Selling Points:

- Shorter latencies for clients
- Less network traffic towards the cloud
- Less processing at the cloud
- Better privacy via local processing
- …

Let's take a closer look from a network perspective

Network Perspective



Motion-to-Photon?

Perceivable Latency?

Human Response Time?

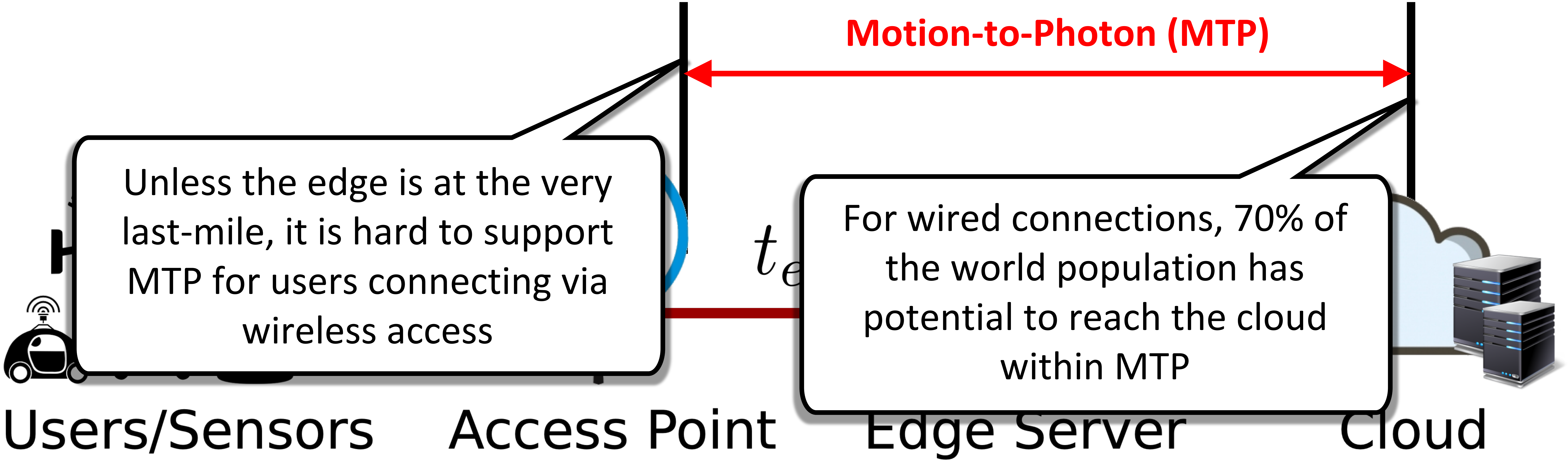
Where are the human limits in this figure?

Metrics

Tool	Metric	What does it mean?
ping	Latency	**How far is the cloud?** ✓
traceroute	Network Path	**Where to place the Edge?**

Measurements ongoing since September 2019

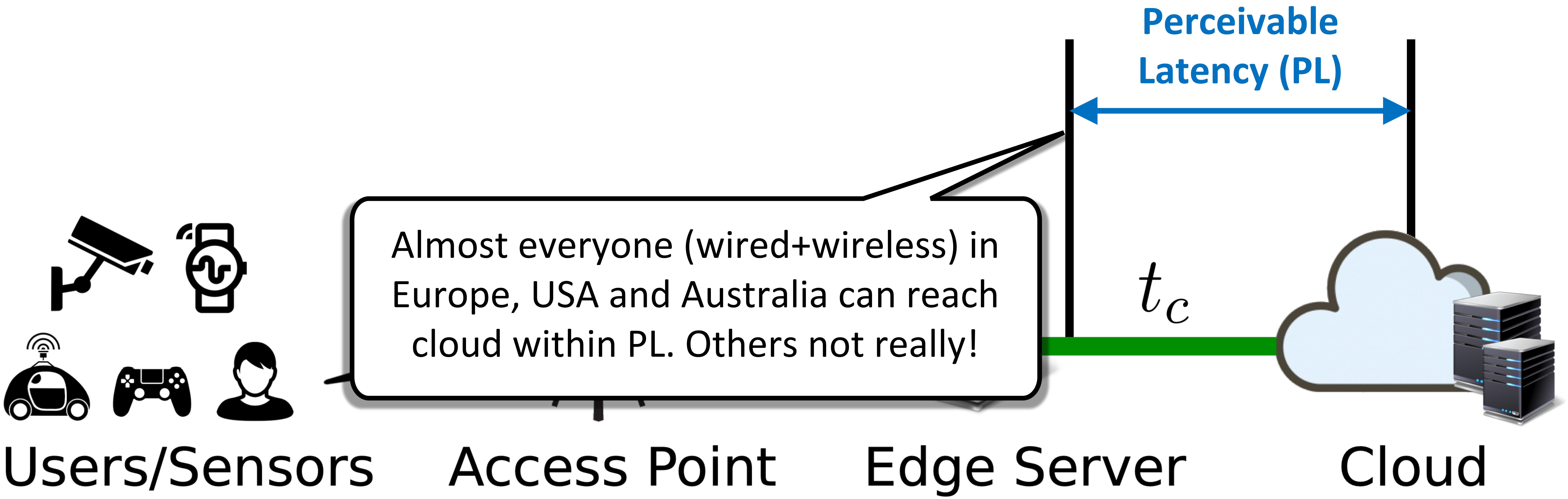
Putting Everything Together



Unless the edge is at the very last-mile, it is hard to support MTP for users connecting via wireless access

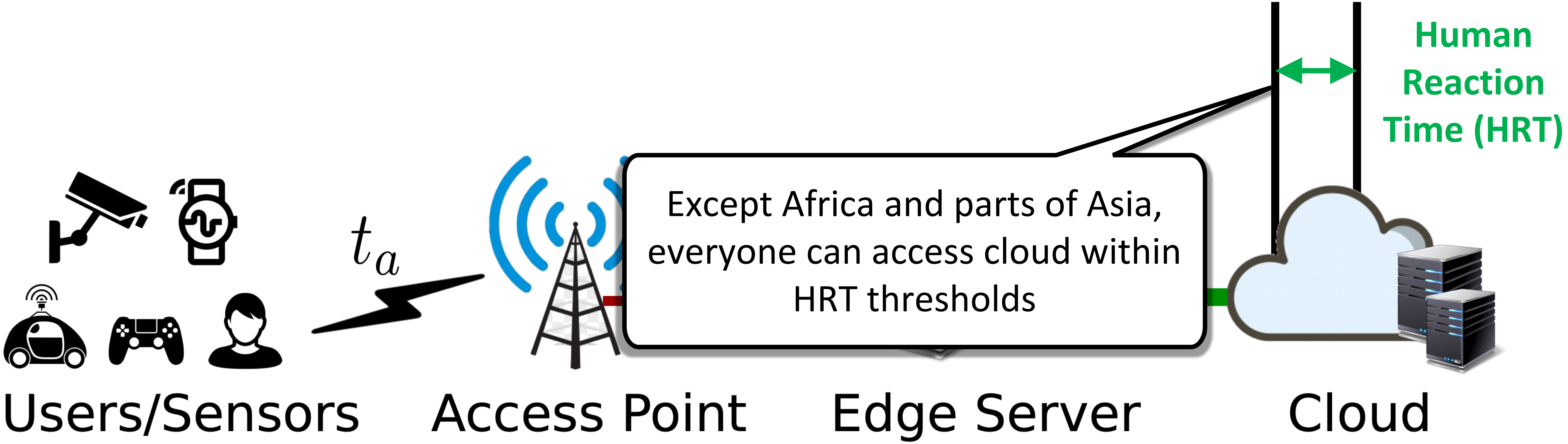
For wired connections, 70% of the world population has potential to reach the cloud within MTP

Putting Everything Together



Almost everyone (wired+wireless) in Europe, USA and Australia can reach cloud within PL. Others not really!

Putting Everything Together



Except Africa and parts of Asia, everyone can access cloud within HRT thresholds

Practicality of our Results

Our measurements provide transport-layer latencies

Application latency to cloud can be higher than what we report
e.g. AR/VR rendering time

Edge might be highly beneficial for applications lying at boundary of Edge Feasibility zone

Cloud deployment is also increasing

Cloud is increasing its reach to new locations in Asia, Africa and Europe

Specialized cloud deployments such as AWS Lambda@Edge, Azure Edge Zones, Google Edge clouds already working

Edge, What Is It Good For?

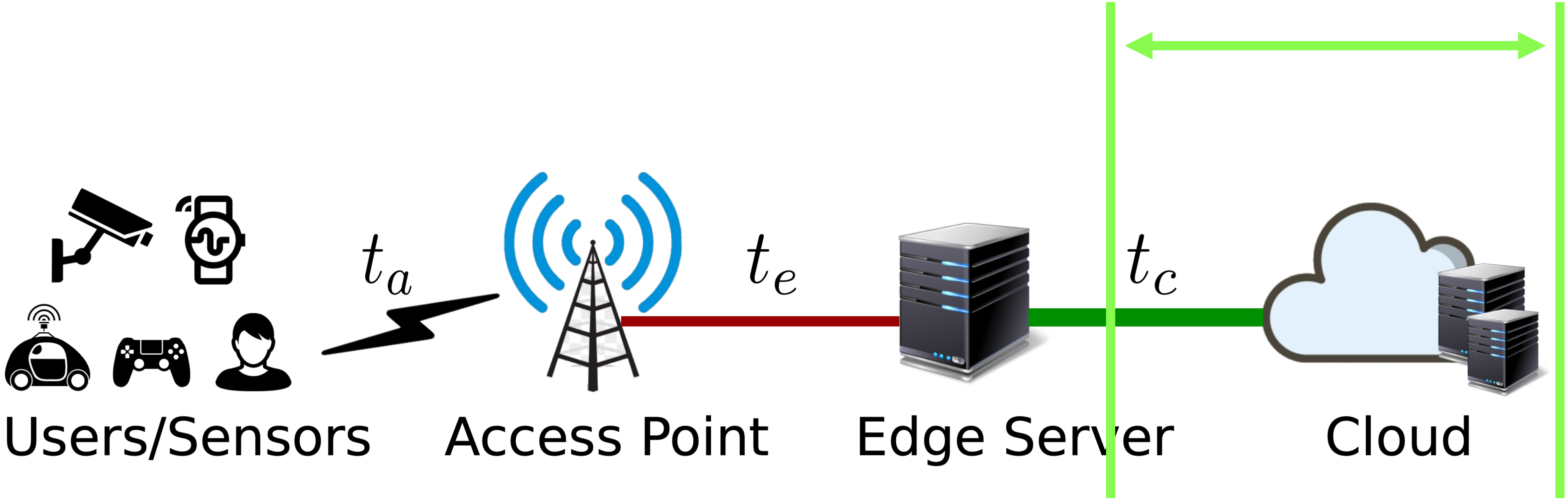
- ~~Shorter latencies for clients~~
- ~~Less traffic towards the clouds~~
- Less processing at the clouds
- Better privacy via local processing

What about privacy?

Other Considerations

- Privacy via local processing: Can be an advantage
- Trust and security: How many edge providers?
- Differences in processing power cloud vs. edge
- Cost of deployment of edge very large
- Performance advantage not useful for applications
- Makes sense in poorly-connected regions

What About 5G?



5G promises t_a + t_e of a few ms ➞ MTP feasible for edge!
LTE promised t_a ~ 10 ms ➞ Reality 50+ ms
Only time will tell if 5G delivers what it promises

What Have We Learned?

- Edge offers limited technical benefits
- “Limited” = Applications do not need them
- Need extensive deployment for better benefits
- Focus on areas which make sense
- 5G may or may not change the situation