Edge Computing – Always Just Beyond the Horizon?

The economics of edge AI don’t look great - or why edge computing may always be the future

Henning Schulzrinne - Columbia University

Edge computing – science vs. engineering

Premise: Edge computing is (largely) an economic issue

- Edge computing = **faster, cheaper, and better!**
- Constrained by performance or reliability ➔ implement in end system regardless of cost
  - or make system economically unviable
- Otherwise, shift computation where cheapest:
  - initial investment
  - load factor: most end user applications are intermittent or have variable computational needs
  - energy
  - operations (system management)
  - security costs

“Edge computing challenges: 5 Reasons why we still do not have large scale deployments” (STL Partners)

- The telco business case is still not fully defined
- The market dynamics and roles are still evolving
- Telcos are going through major architectural and organizational changes
    - “We are network plumbers that don’t want to be plumbers – and movie director didn’t work out so well”
- Building a telco edge node requires careful consideration
    - “Also, If deployed across their access network (at the far edge), telcos risk edge nodes which are costly, under-utilized, and unprofitable. If they only build a few edge nodes, benefits compared to regional edges provided by other players such as neutral host and data center providers may be relatively small.”
- The enterprise sector is also still exploring edge applications and deployments

https://stlpartners.com/articles/edge-computing/edge-computing-challenges/

The rise of edge computing (publications)

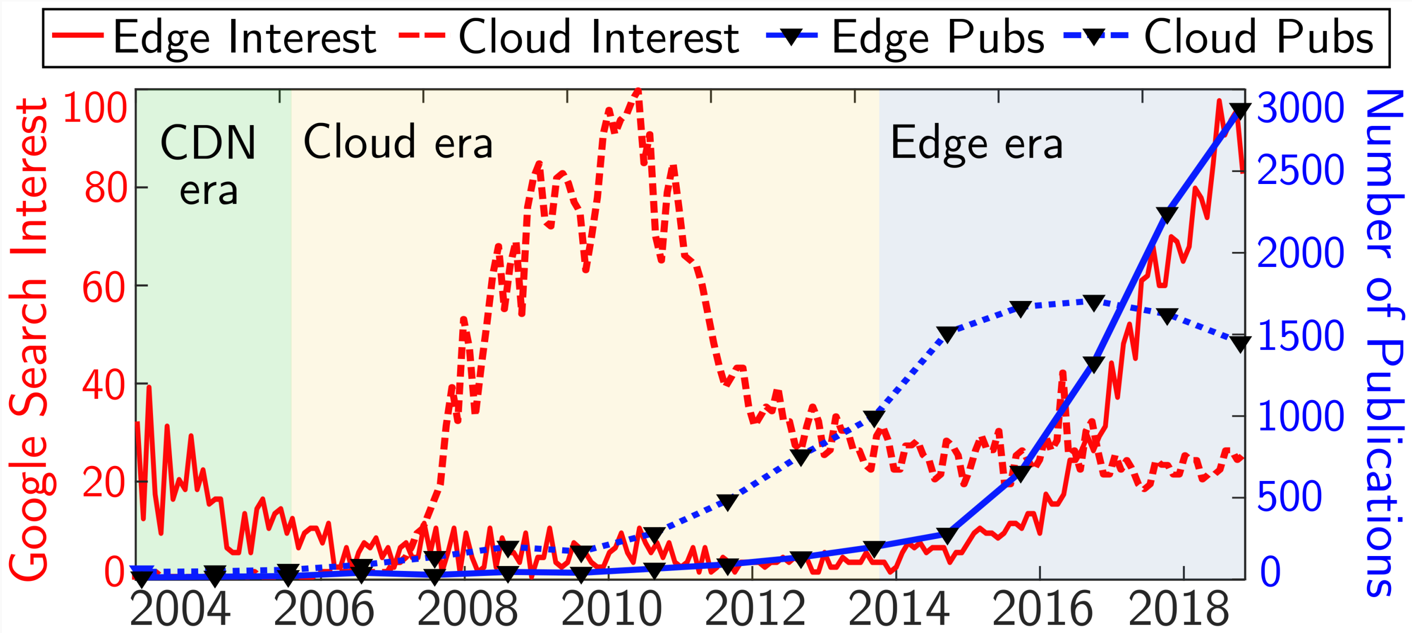
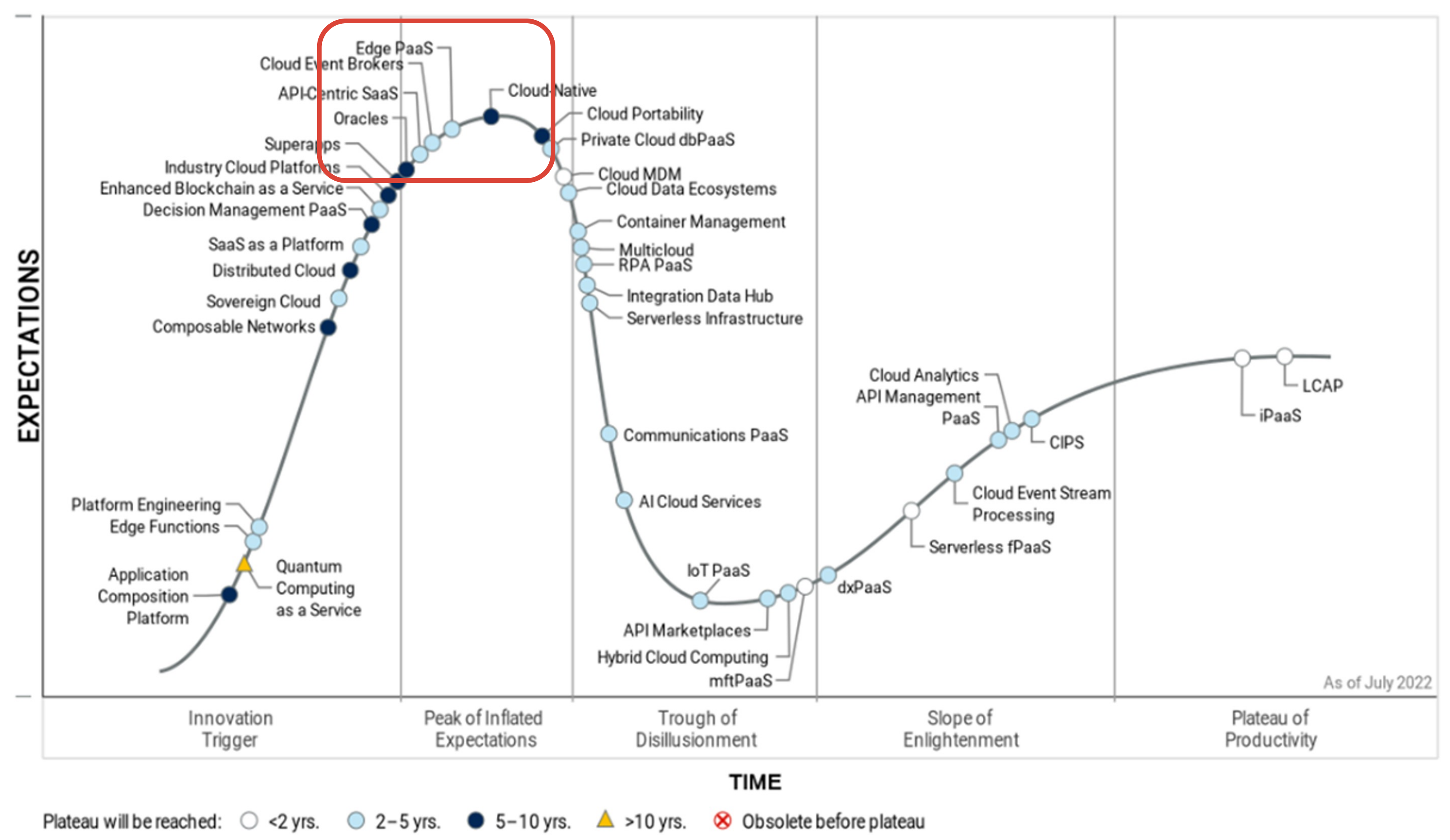


Figure 1: The popularity (in red) and publications (in blue) of keywords "edge computing" (in solid line) and "cloud computing" (in dashed line) in Google web searches and Google scholar respectively.

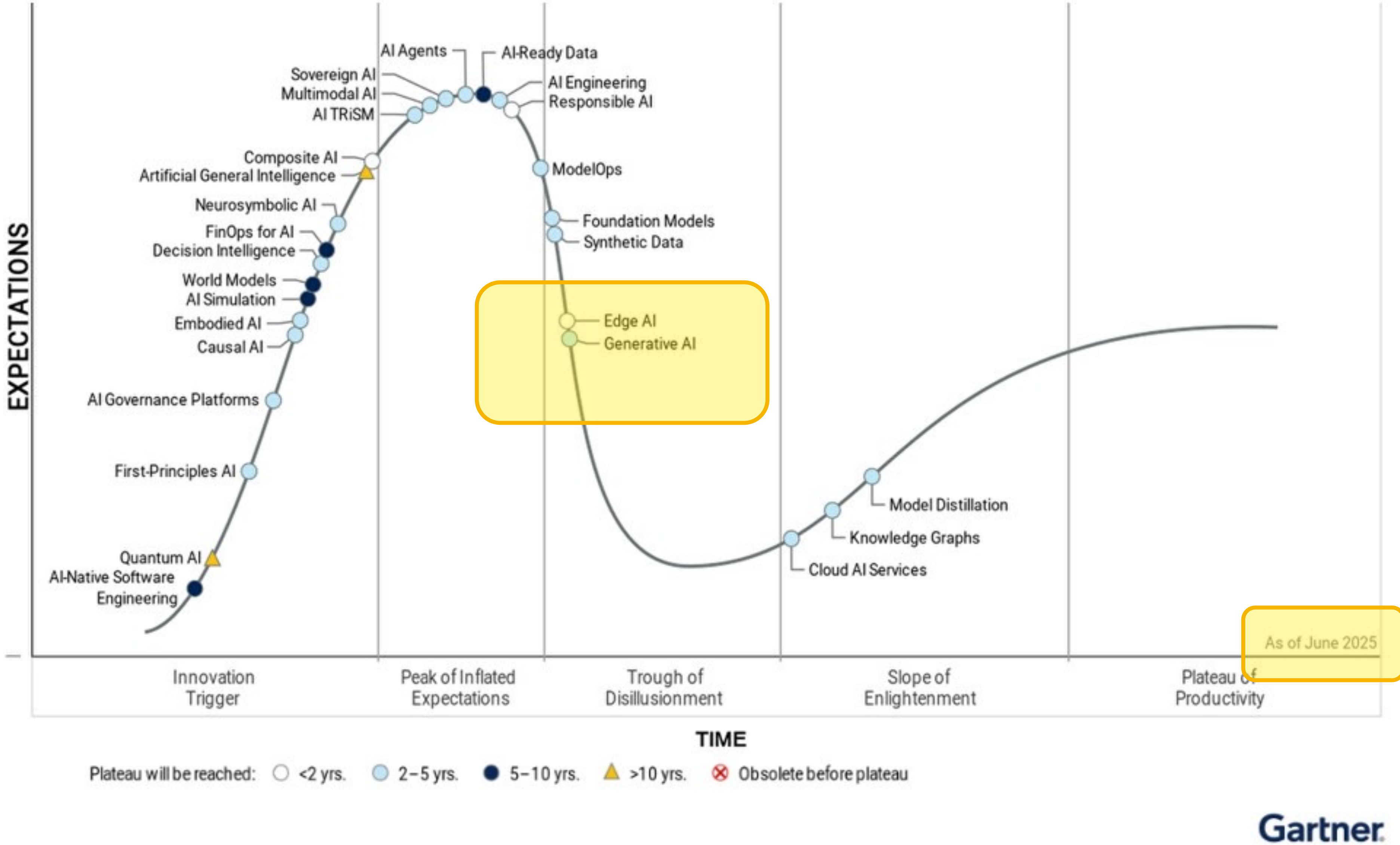
Mohan, et al., "Pruning Edge Research with Latency Shears," 2020.

Cloud hype cycle (2022)



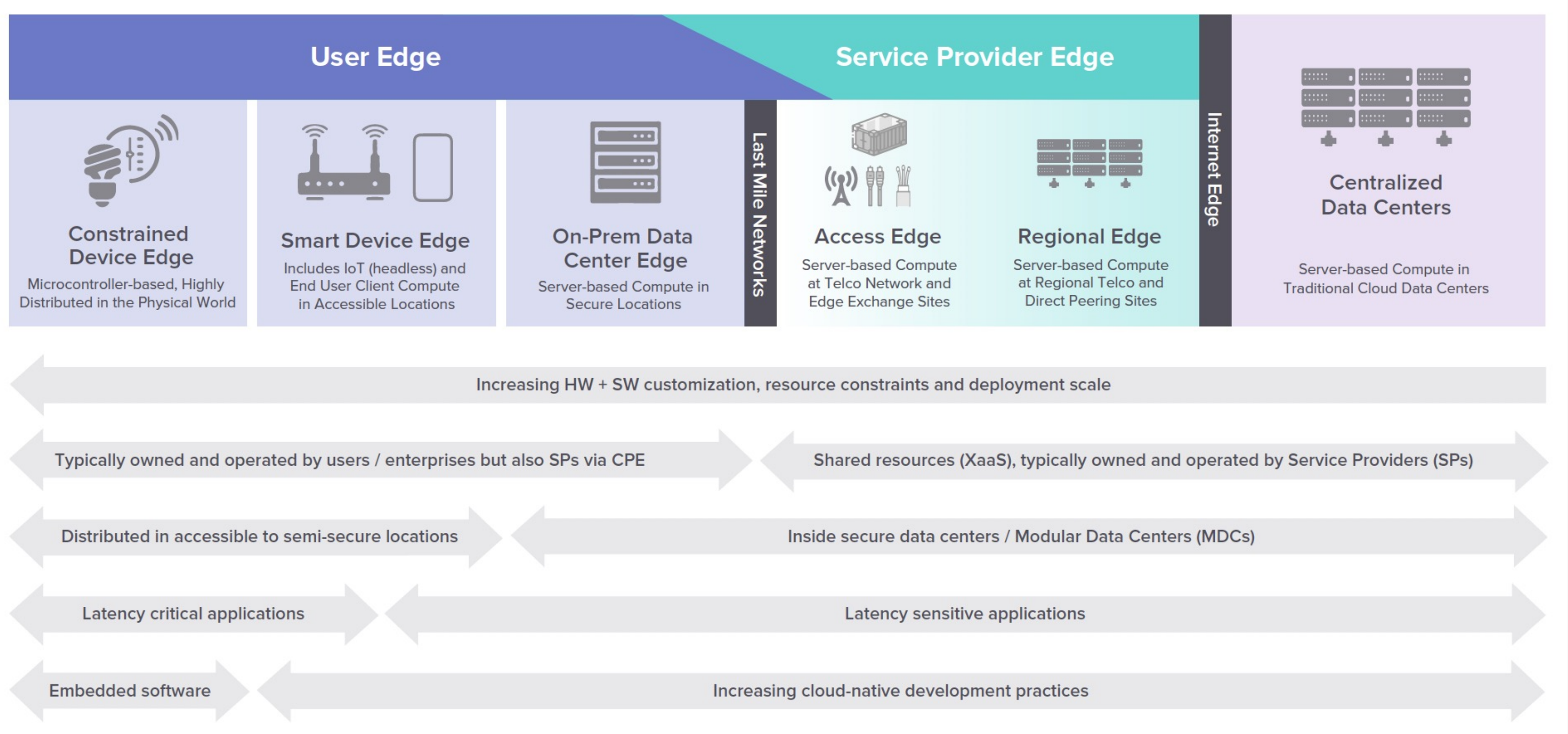
Gartner

Edge AI – we've made progress

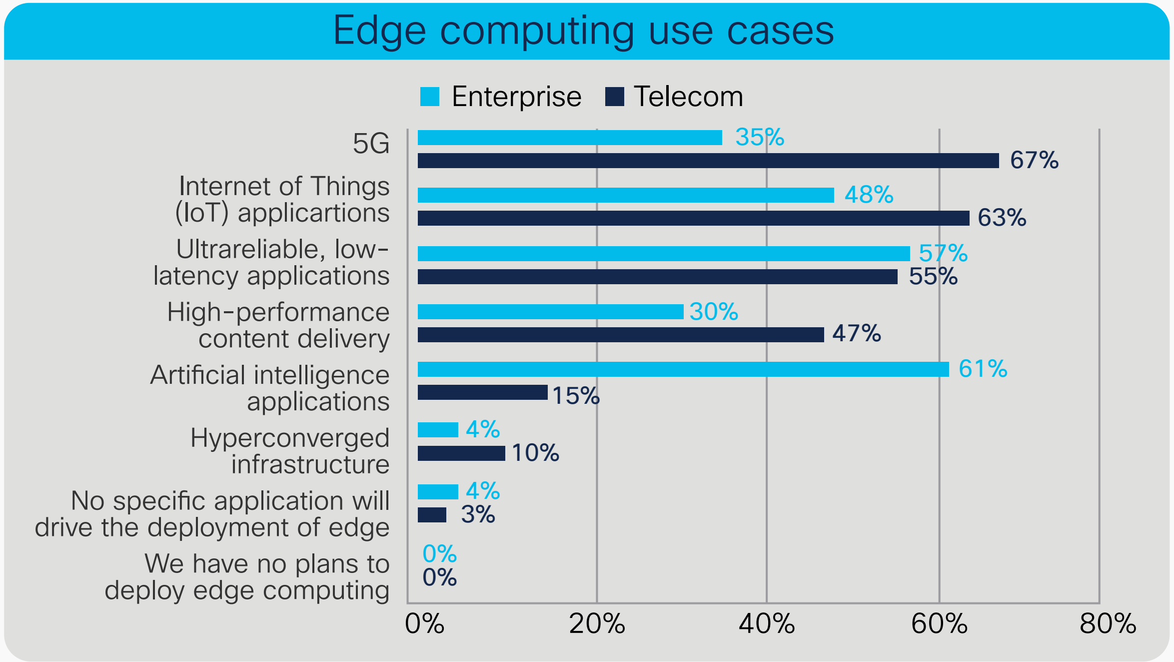


Edge computing contains multitudes

Edge-to-cloud spectrum



https://www.cloudmc.cloudops.com/blog/understanding-the-edge-to-cloud-spectrum

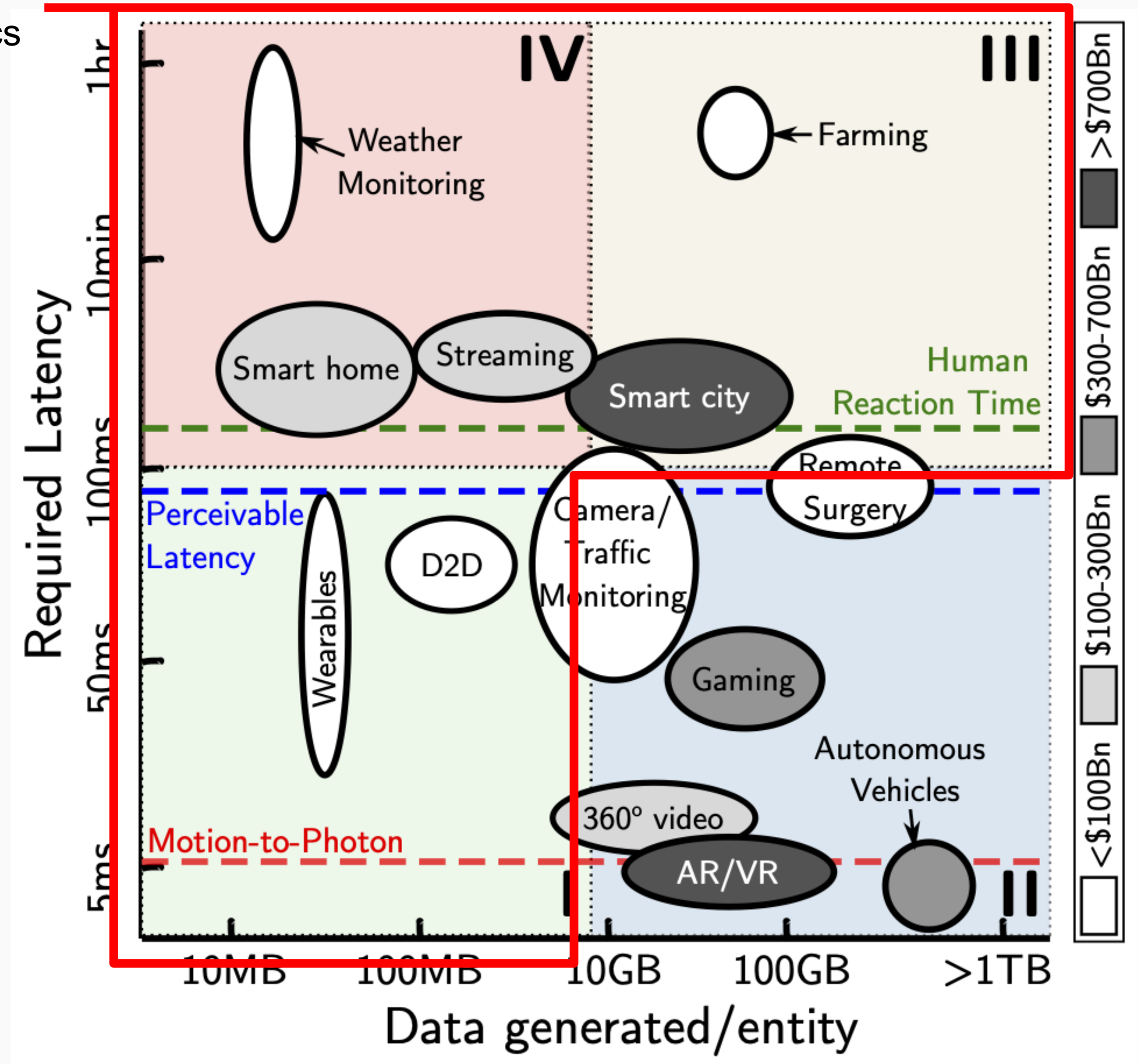


Source: Strategies for connecting the edge, heavy reading, September 2019

[Percent of respondents: N = 60 telecom, 23 enterprise]

Edge use cases

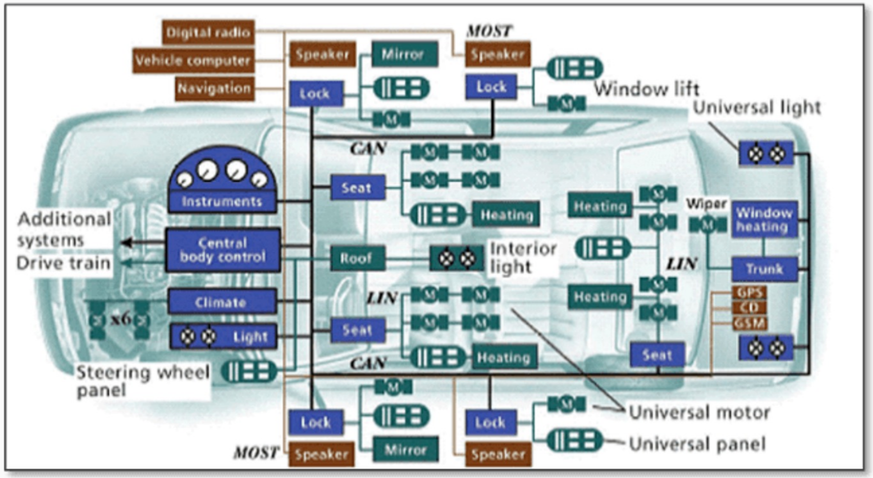
edge computing
optional - economics



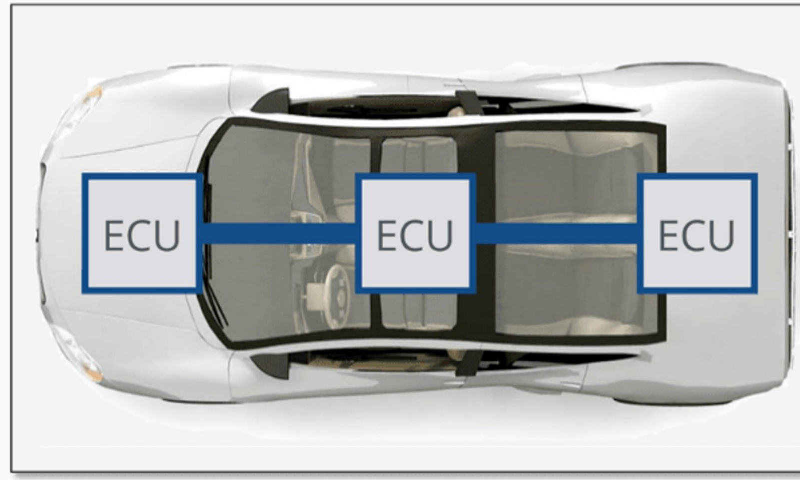
Mohan, et al., “Pruning Edge Research with Latency Shears,” 2020.

Easy and hard-edge computing

Conventional Architecture



Software Centric Approach



vs.

Edge provider A

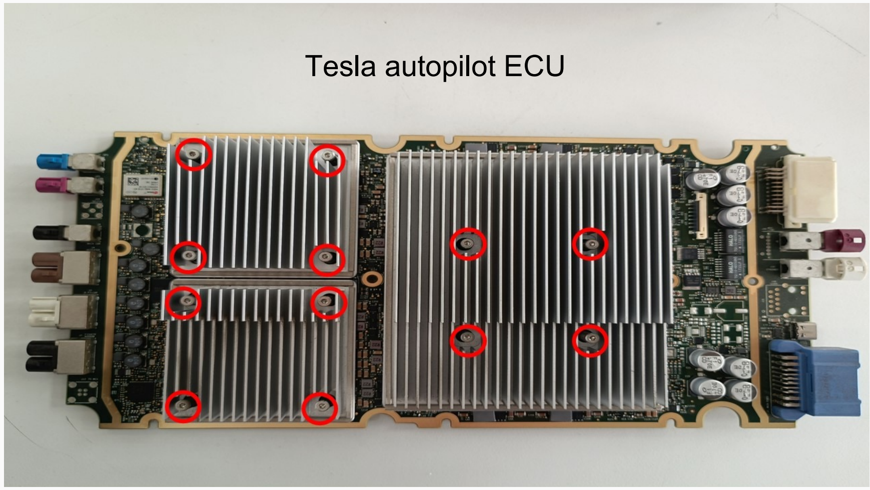
Cloud provider B

Edge provider C

vs.

API (vision, sensing, object recognition, …)

The end system itself: ns - μs



Advantages	Concerns	Applications
Ensures data privacy	Low utilization	Smartphone (camera, voice commands)
Low latency	Expensive GPU if only used for few applications	Autonomous vehicles Voice commands in vehicles
Fast local storage	Battery drain	Games
Works regardless of network		

Within 50 m: the home gateway: 4-5 ms

Advantages	Concerns	Applications
Ensures data privacy	No standards for management	Security camera object recognition
Low latency	Expensive GPU if only used for few applications	Privacy-respecting voice chat
Storage (NAS) can be co-located	Often provided by ISP → incentives lacking	

Within 1-2 km: macro cell sites: 1-5 ms

142,100 cell sites in US

URLLC

Maghsoudnia et al., HOTNETS’24

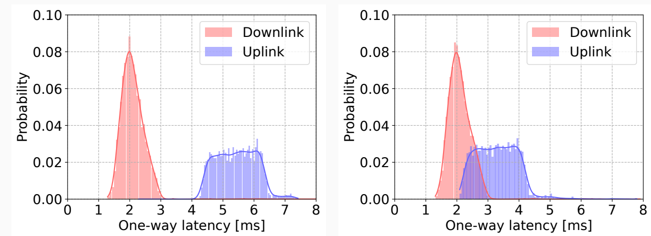
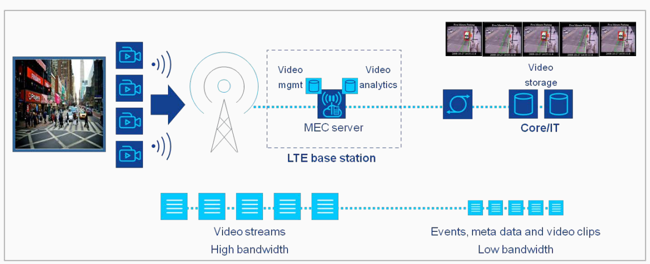
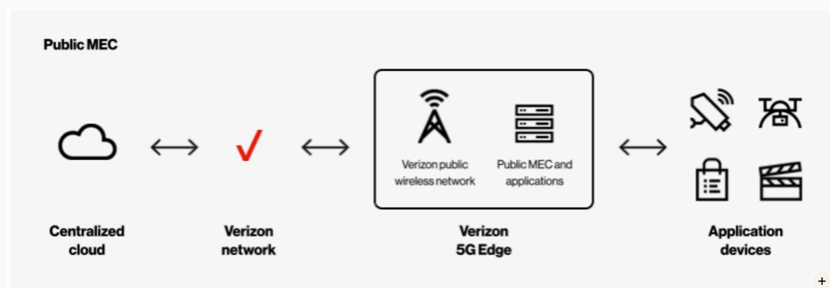


Figure 6: One-way latency: (a) grant-based and (b) grant-free.

Use Case 3: Video Analytics



Mobile-edge Computing (MEC) - 2015



Verizon MEC - 2024

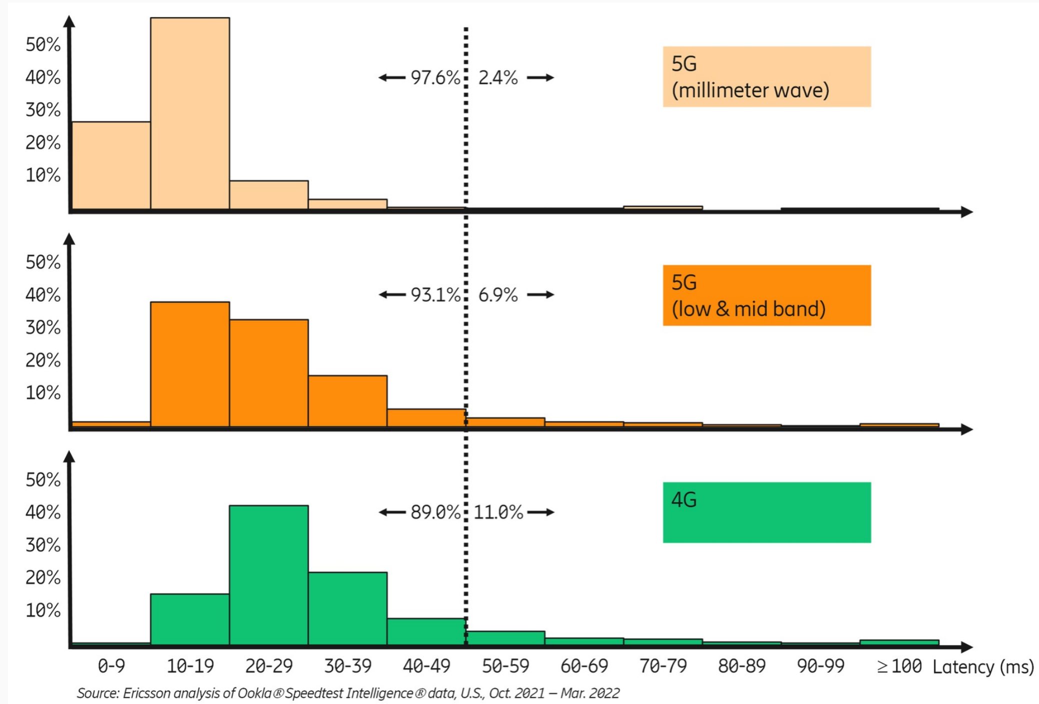
Verizon: 17 edge sites with AWS Wavelength and Vodafone which deployed around 4 sites

Cell sites

Advantages	Concerns	Applications
As close to edge as feasible	Remote – high cost for hands-on maintenance	?
Lowest latency	Expensive GPU if only used for few applications	
Fate sharing: likely will work if cell site does	Limited space & power (1-2 racks)	
	Limited use cases unless universal	

Within a few km: central offices (telephone exchanges) – 10 to 50 ms

5,600 British Telecom exchanges
25,000 US central offices



5G latency

The Openreach Network Today

Openreach's exchange footprint is a legacy of the original Public Switched Telephone Network (PSTN) rollout, which began over a century ago. Openreach leases space in c. **5,600 BT exchange buildings** providing services to CPs, who in turn then provide telephone and broadband services to their end customers.

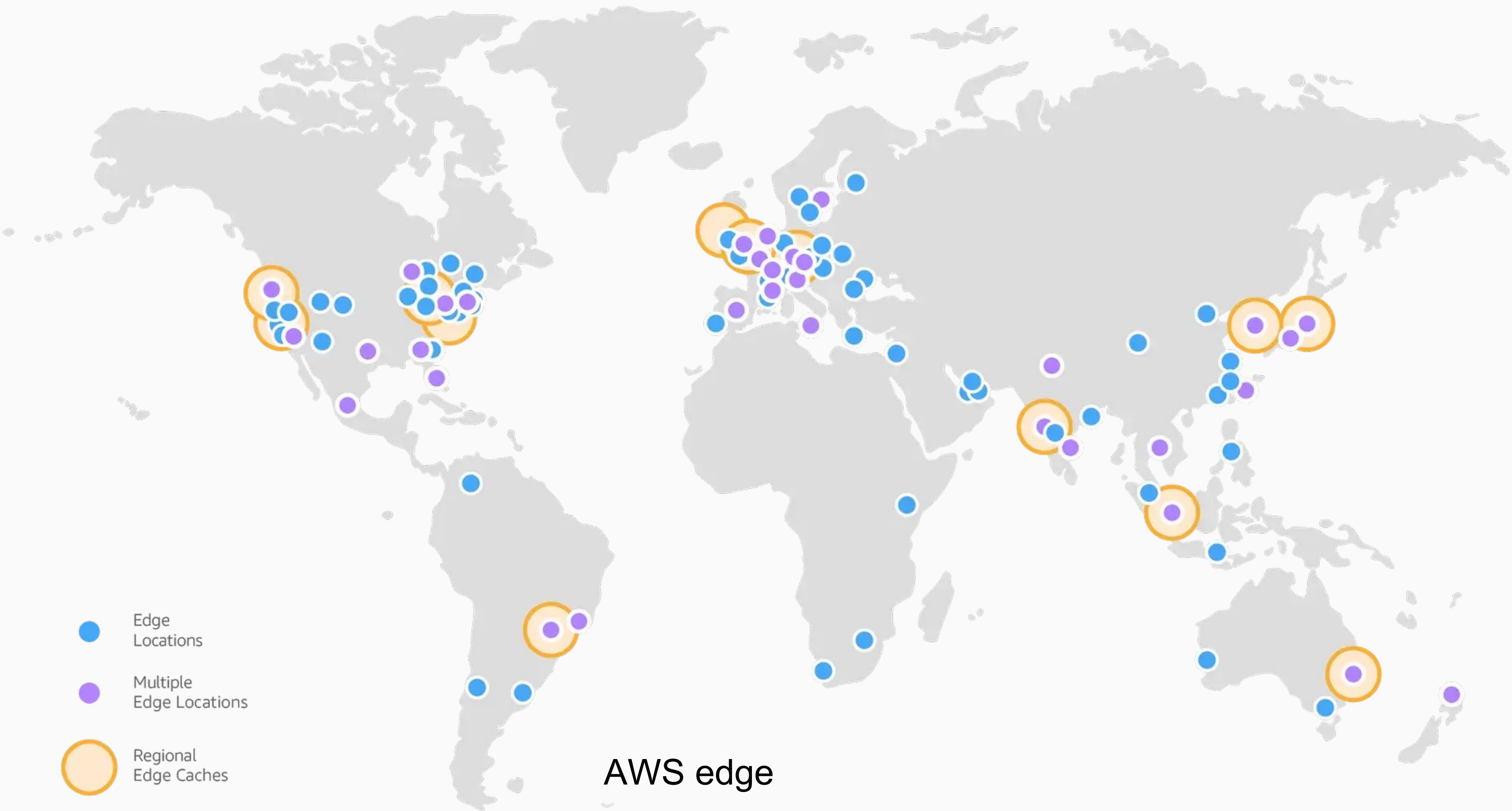
Today, we use c. **1,000 exchanges** - Openreach Handover Points (OHPs) – to provide nationwide coverage of fibre broadband services (FTTC, FTTP and Gfast). The remaining c. **4,600 exchanges** are used to provide ADSL broadband and "legacy" voice services (WLR, MPF and SMPF). Most are also used to provide Ethernet or other leased line services.

Central offices

Advantages	Concerns	Applications
Already has power (e.g., 375 kW)	Many will be decommissioned	?
Already owned by phone company		
May have fiber connectivity		

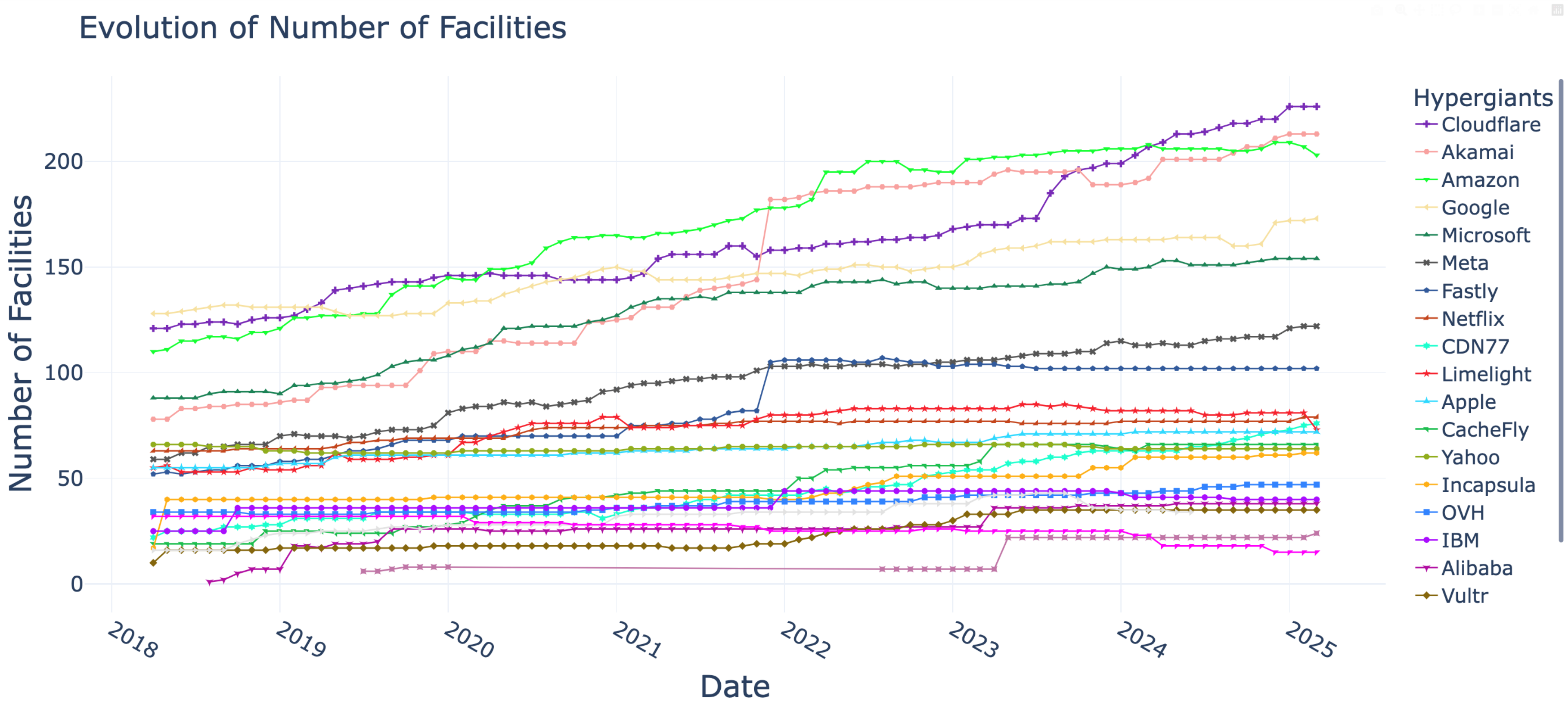
Edge computing is already boring reality

Cloud provider edge computing: AWS



AWS edge

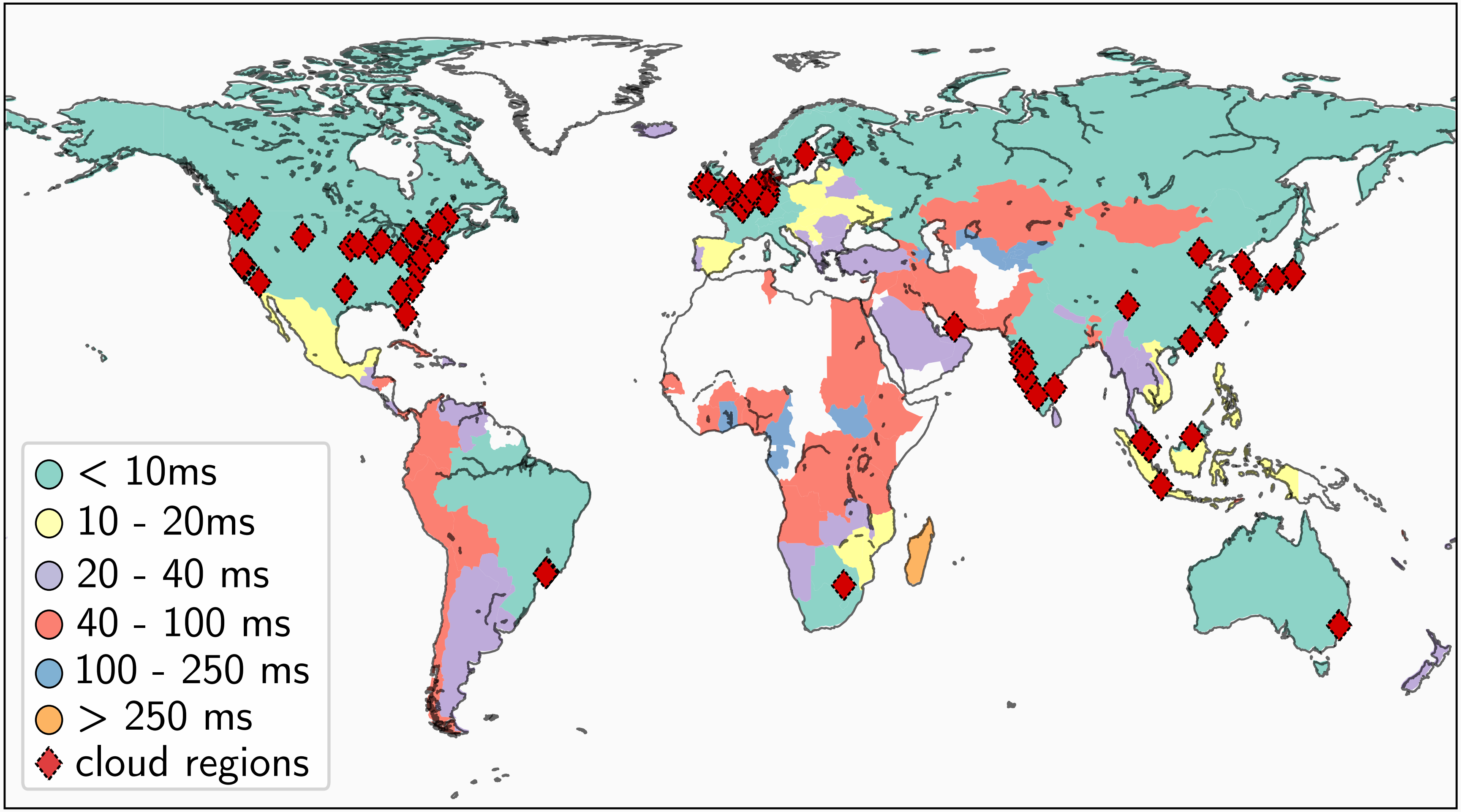
Points of presence



https://pulse.internetsociety.org/blog/visualizing-the-rise-of-hypergiants

Cloudflare

Most of the world is within 10-20 ms of a data center



note: best probe data in each country
does not include 43 AWS local zones

Cloud latency distribution

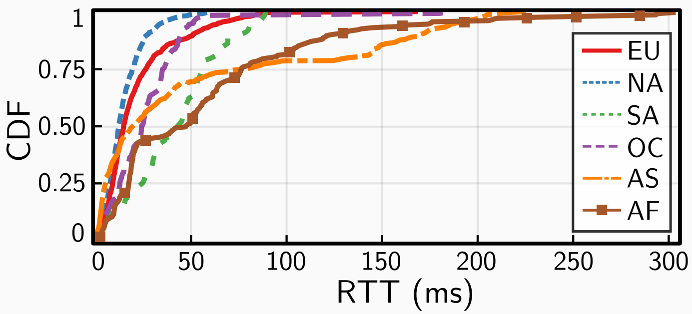


Figure 5: CDF of minimum RTT of all probes to nearest datacenter by continent.

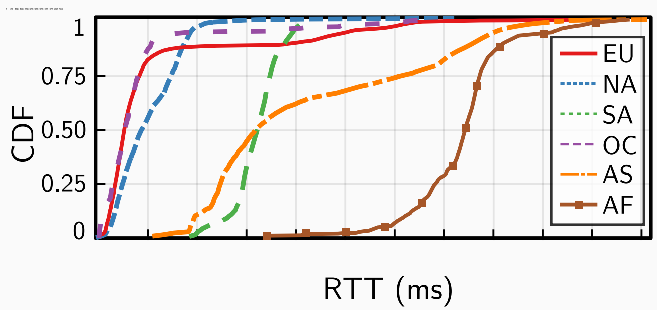


Figure 6: CDF of all ping measurements from all probes to their closest datacenter.

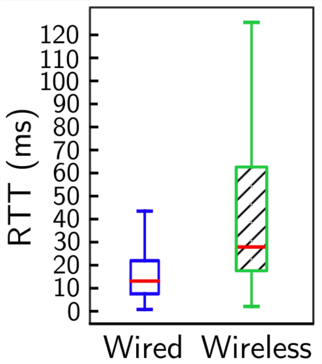
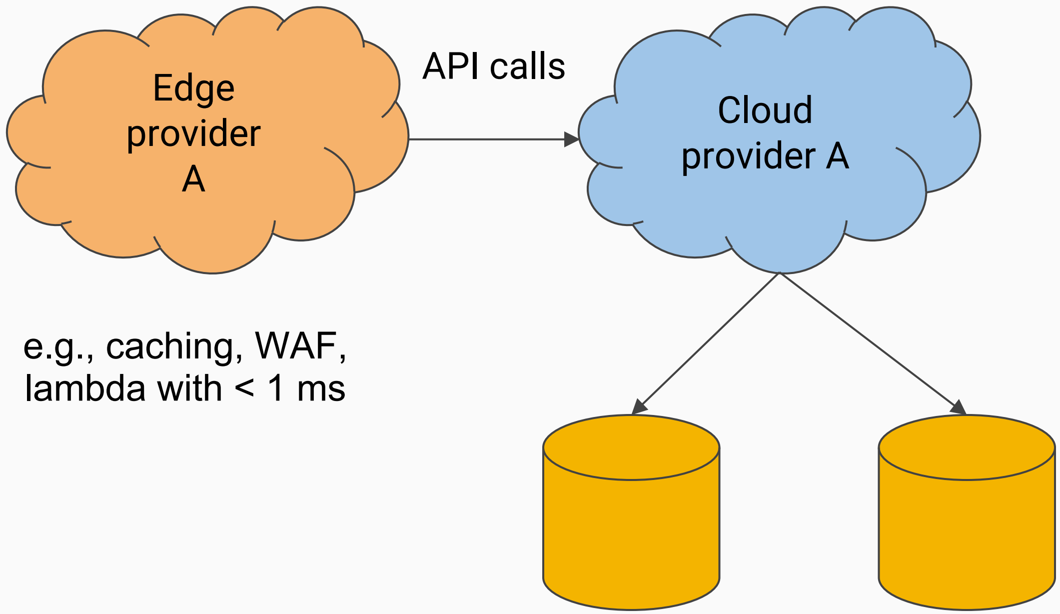


Figure 7: Wired vs. wireless access RTT.

Density conundrum

- Where there is a market, there's cloud infrastructure
  - US: about 80% live in urban areas
  - Europe: about 76%
- Low density means low utilization
- Example: 5G deployment in rural areas

Most common architecture



Dependability

Cheap vs. dependable

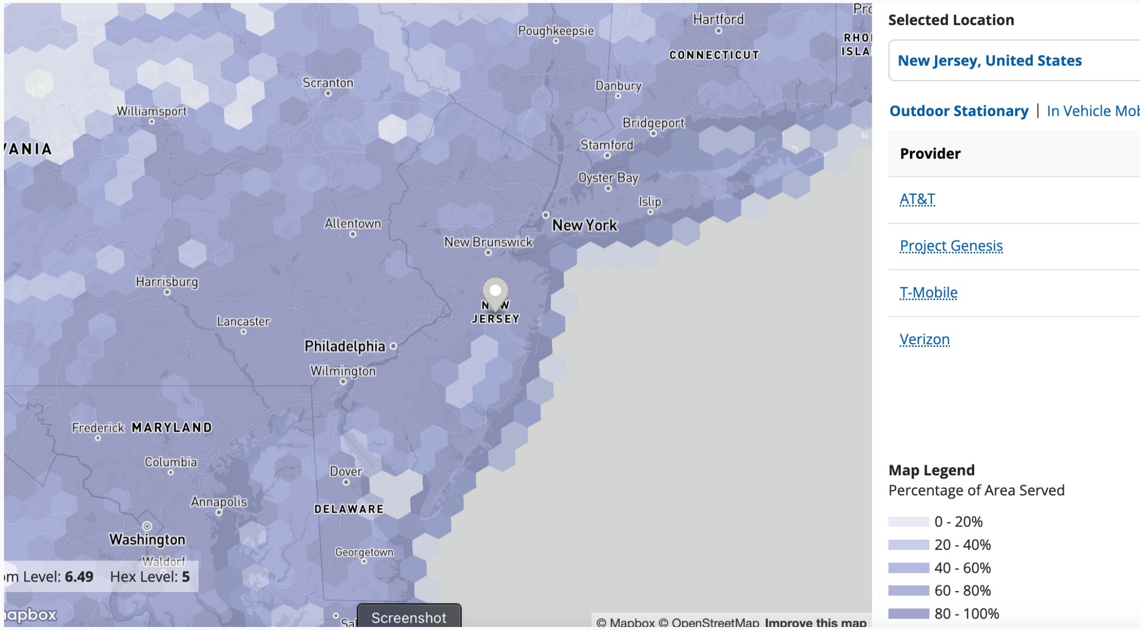
Cheap: don’t care about downtime or reachability

- roughly, batch computing
- e.g., training or analysis using datasets
- ⇒ re-use otherwise idle computing resources where power is “free” (e.g., excess solar or wind) ⇒ SETI@home (aliens don’t have GDPR rights)

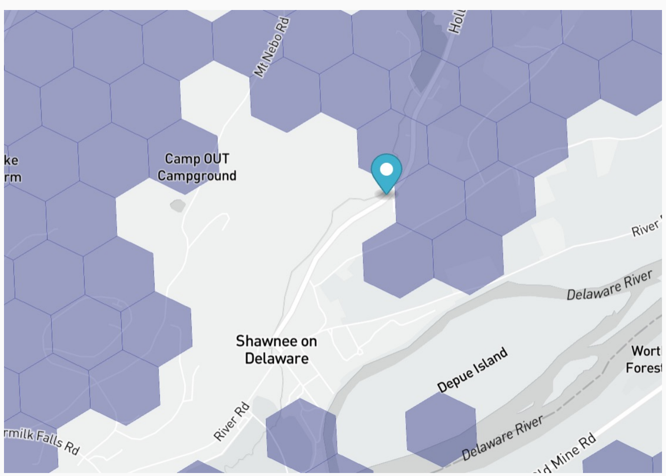
Dependable

- real-time needs (autonomous vehicles, consumer-facing interactions, real-time transaction monitoring)
- fate sharing with application desirable
- latency matters in a few cases → *closed loop (cyberphysical) systems*

Sudden loss of broadband pressure



270 road tunnels with a total length of 270 kilometers on federal highways in Germany. On rural, district and urban roads, there are 420 tunnels with a total length of over 350 kilometers.



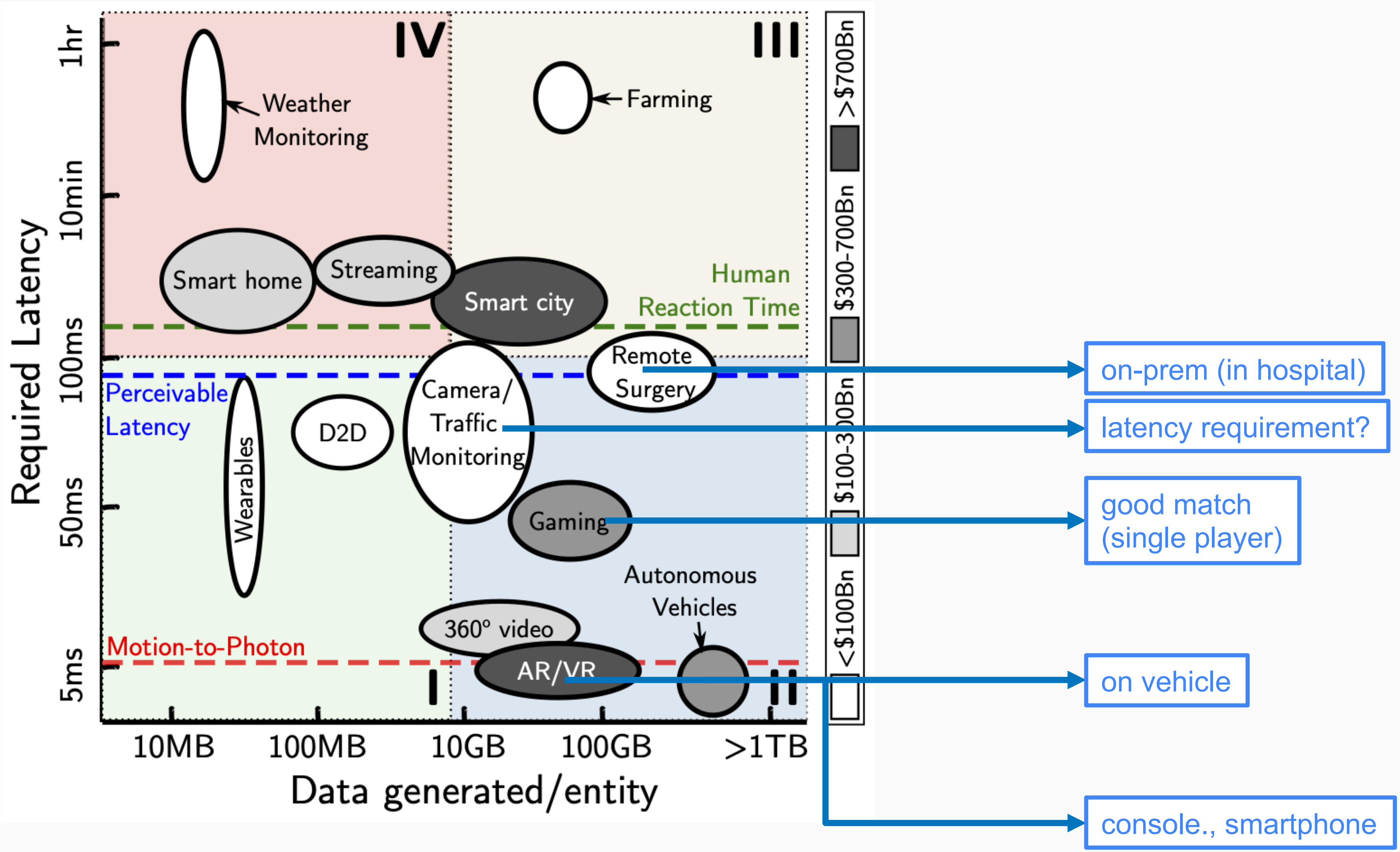
Arlinger Tunnel

https://www.heise.de/en/news/5G-Vodafone-wants-to-close-dead-spots-in-tunnels-with-innovative-antennas-10354500.html

Fate sharing

- **Fate-sharing** is an engineering design philosophy where related parts of a system are yoked together, so that they either fail together or not at all. (Wikipedia)
- “The fate-sharing model suggests that it is acceptable to lose the state information associated with an entity if, at the same time, the entity itself is lost.” (Clark, 1988) – about transport protocols
- New failure modes:
  - loss of (5G) connectivity
  - roaming capacity failure: no available suitable services (e.g., specific GPU)
  - latency failure zones – no edge infrastructure within permissible latency range

Edge use cases: dependability requirements



Mohan, et al., "Pruning Edge Research with Latency Shears," 2020.

An economic analysis

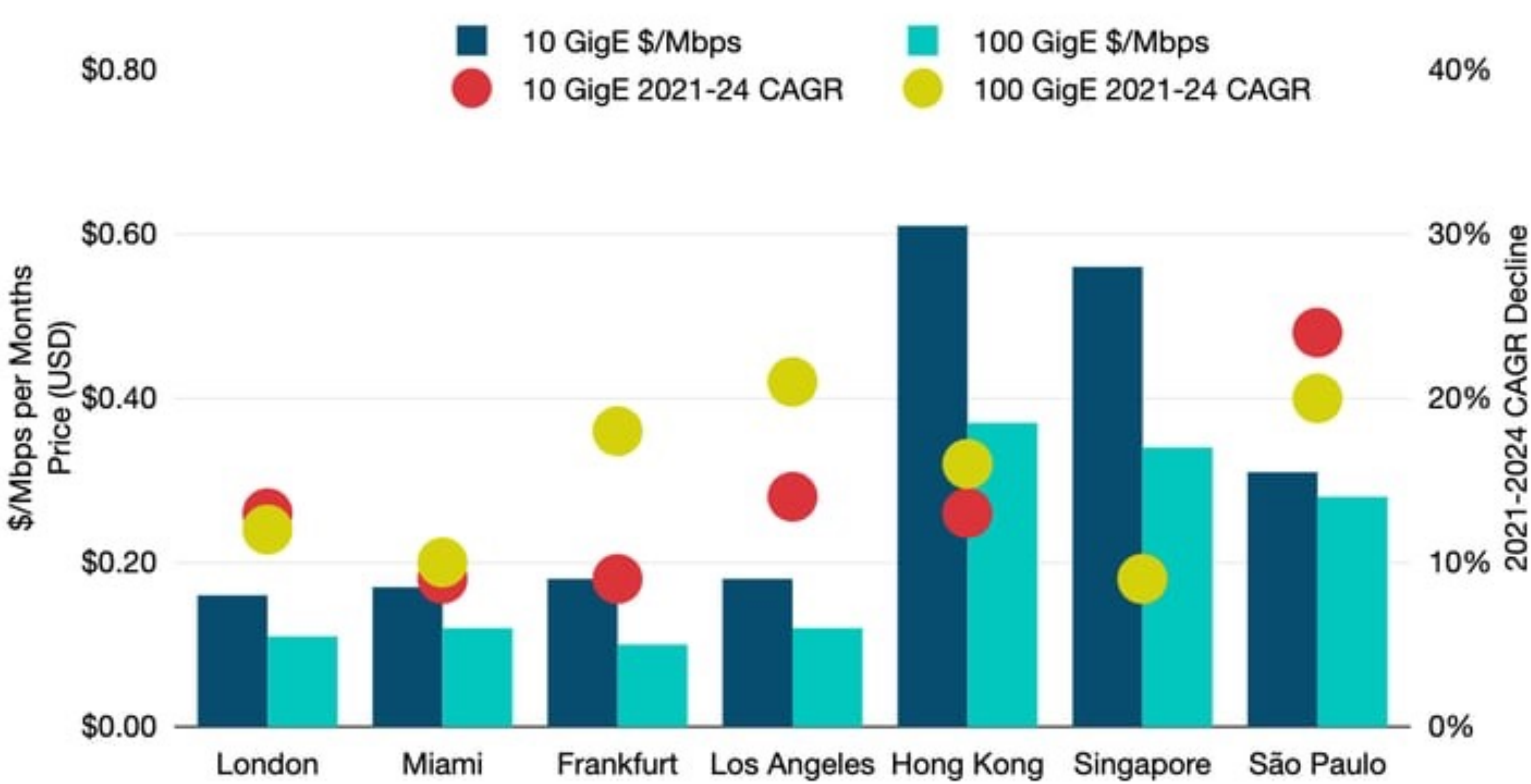
Costs of computing

- Capital cost: cloud server lifetime = 5 years (may be less for GPUs)
- Opex:
  - Electricity: about 60-70% of total operational cost (TOC) [BLS & Co.]
  - Network costs
  - Security cost (risk premium)
  - Development and DevOps costs (e.g., updates)
  - Failure risk (revenue, reputation, legal liability, …)

⇒ Which of these are **lower** for a multi-cloud & edge architecture?

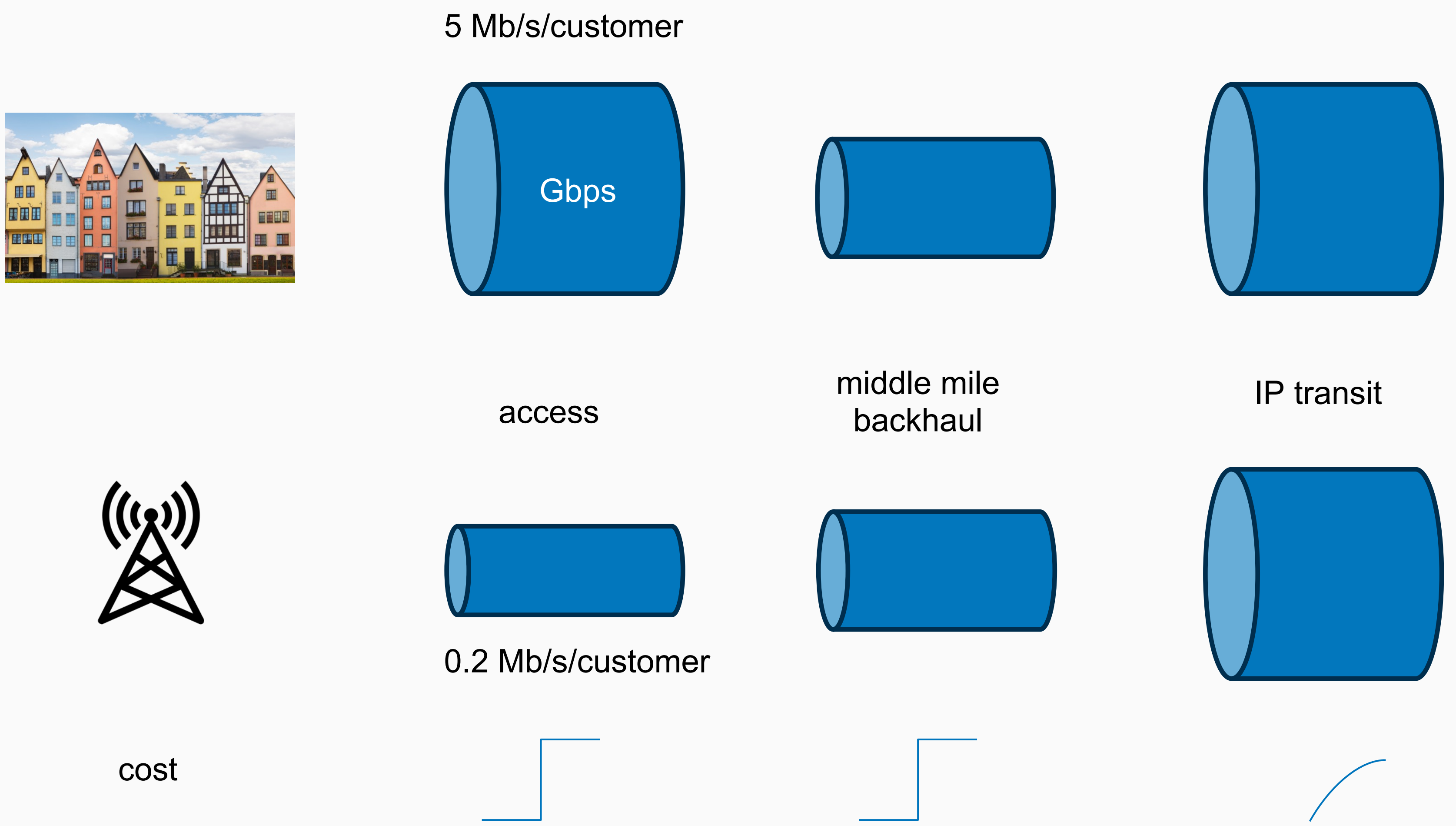
Bandwidth

IP transit costs



Telegeography

Bandwidth pipeline

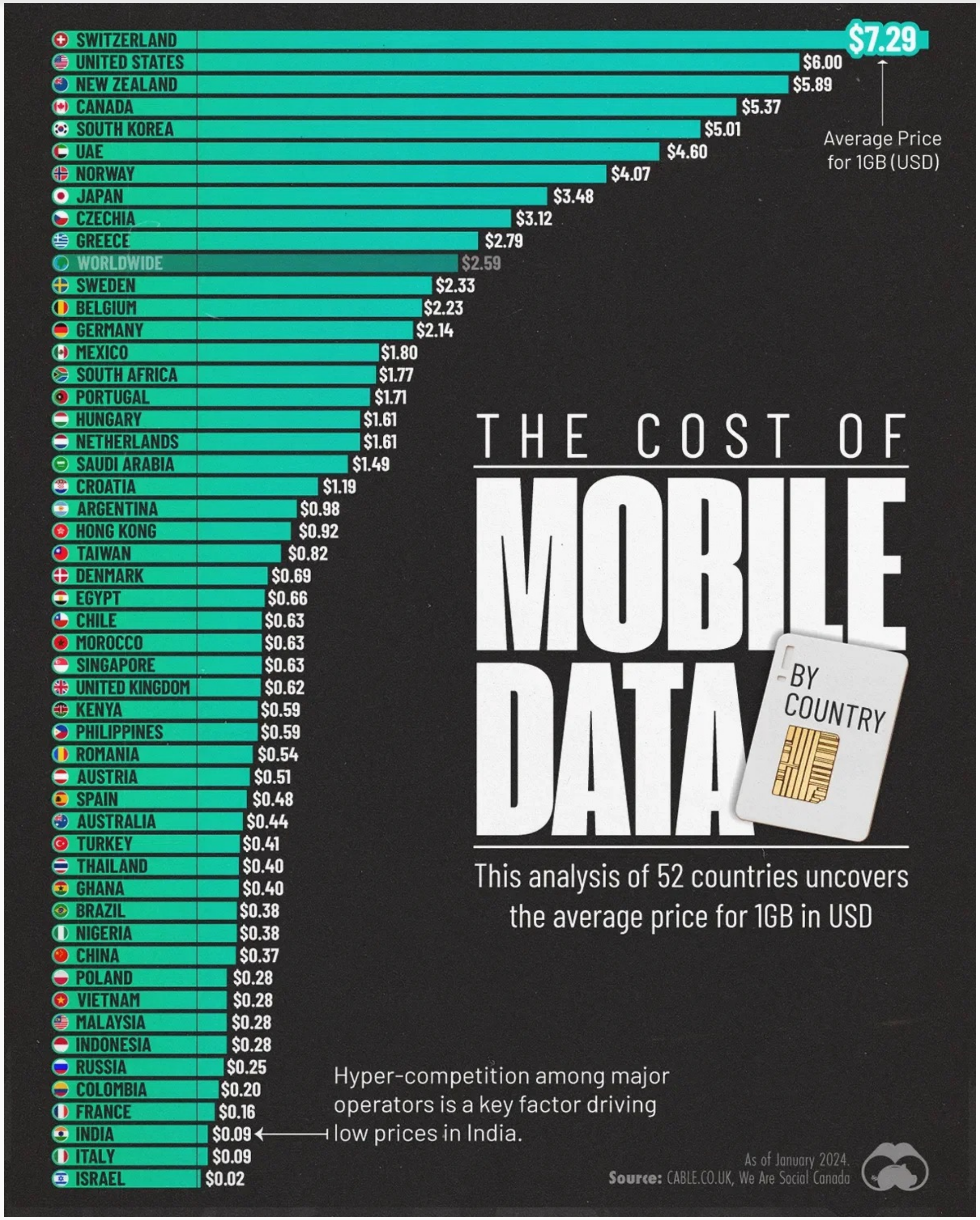


Bandwidth costs

Type	example	cost $/TB
IP transit	HE 40G, $2,000/month	$0.31 ($0.25/m/customer)
Home	$80, 1 TB average	$80
AWS inbound	$0	$0
AWS outbound	$0.09/GB for 10 TB to $0.05/GB	$90
European cellular roaming	€1.30/GB	€130

	CLOUDFRONT	AZURE	KEYCDN	FASTLY	GOOGLE
General					
Instant Setup	✓	✓	✓	✓	✓
Transparent Pricing	✓	✓	✓	✓	✓
Pay As You Go (PAYG)	✓	✓	✓	✓	✓
No Charges for Requests	✗	✓	✓	✗	✓
Bandwith Pricing US&EU	**$0.085**	**$0.087/GB**	**$0.04/GB** $49 min. / year	**$0.12/GB** $600 min. / year	**$0.08**
Points of Presence	100+	100+	60+	60+	100+

Consumer prices are driven by competition (and average consumption)



Cost of compute

FOTW #1356, August 19, 2024: Household Vehicles Were Parked 95% on a Typical Day in 2022

Household vehicles were driven an average of 64.6 minutes on a typical day in 2022 (including all trips made that day) and parked for the remainder of the time (95%).

Compute costs: rent vs. buy

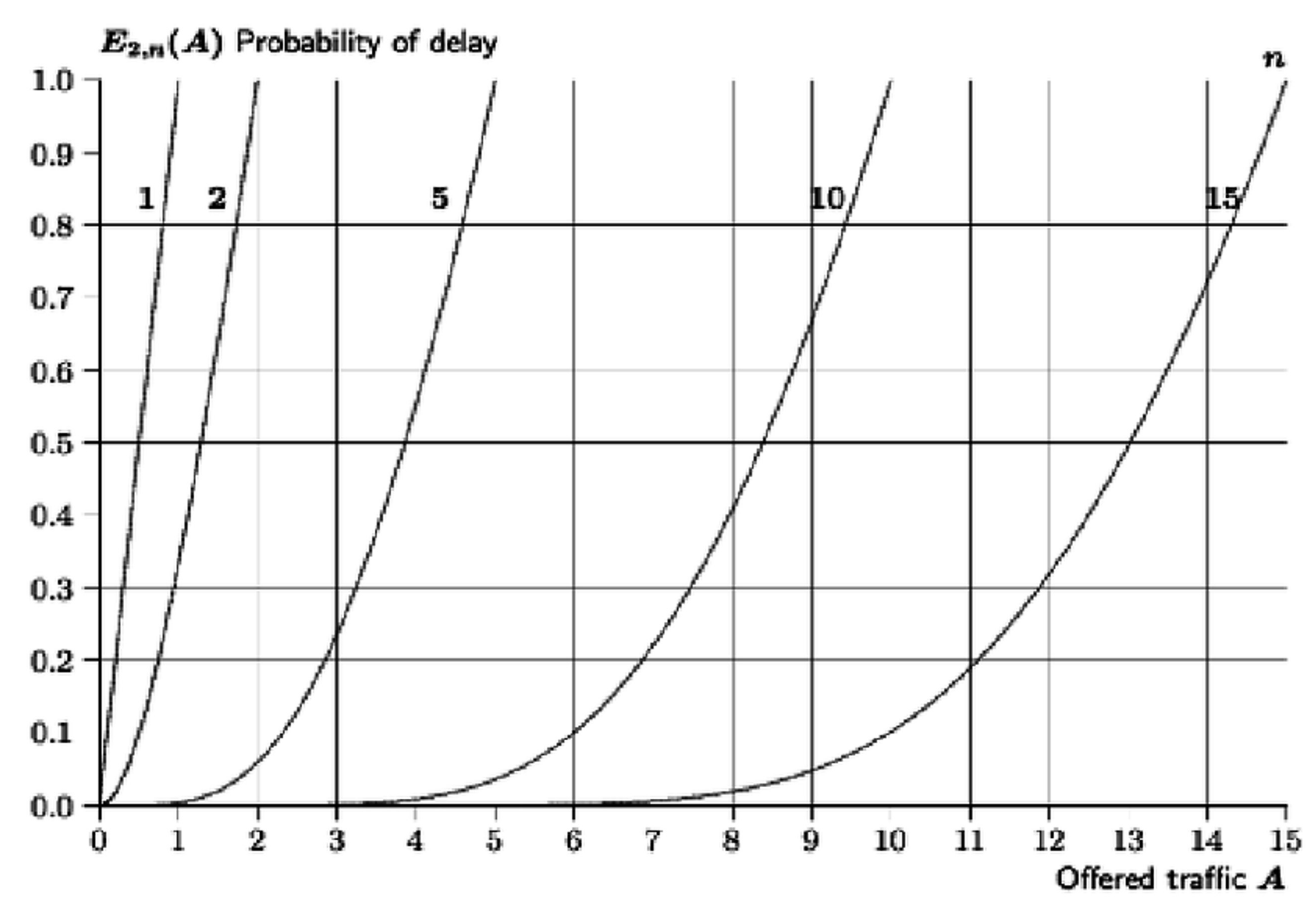
- NVIDIA H100: \$25k to \$31k
  - for 5 years: ~\$0.57/hour
- DGX H100 (8 GPUs): \$373k → \$8/hour
- Bare metal GPU: \$1.99/GPU/hour → \$16/hour
- AWS p5.48xlarge (8 GPUs): \$23/hour (3 year reserved)

Dell XE9680: \$320,499

On-Demand GPU Pricing

GPU Model	VRAM (GB)	Max pCPUs per GPU	Max RAM (GB) per GPU	Pricing Per Hour
NVIDIA H200 SXM	141	22	225	\$3.50
NVIDIA H100 SXM	80	24	240	\$2.40
NVIDIA H100 NVLink	80	31	180	\$1.95
NVIDIA H100	80	28	180	\$1.90

Erlang C does not favor small edge clouds



Cloud scaling vs. edge scaling

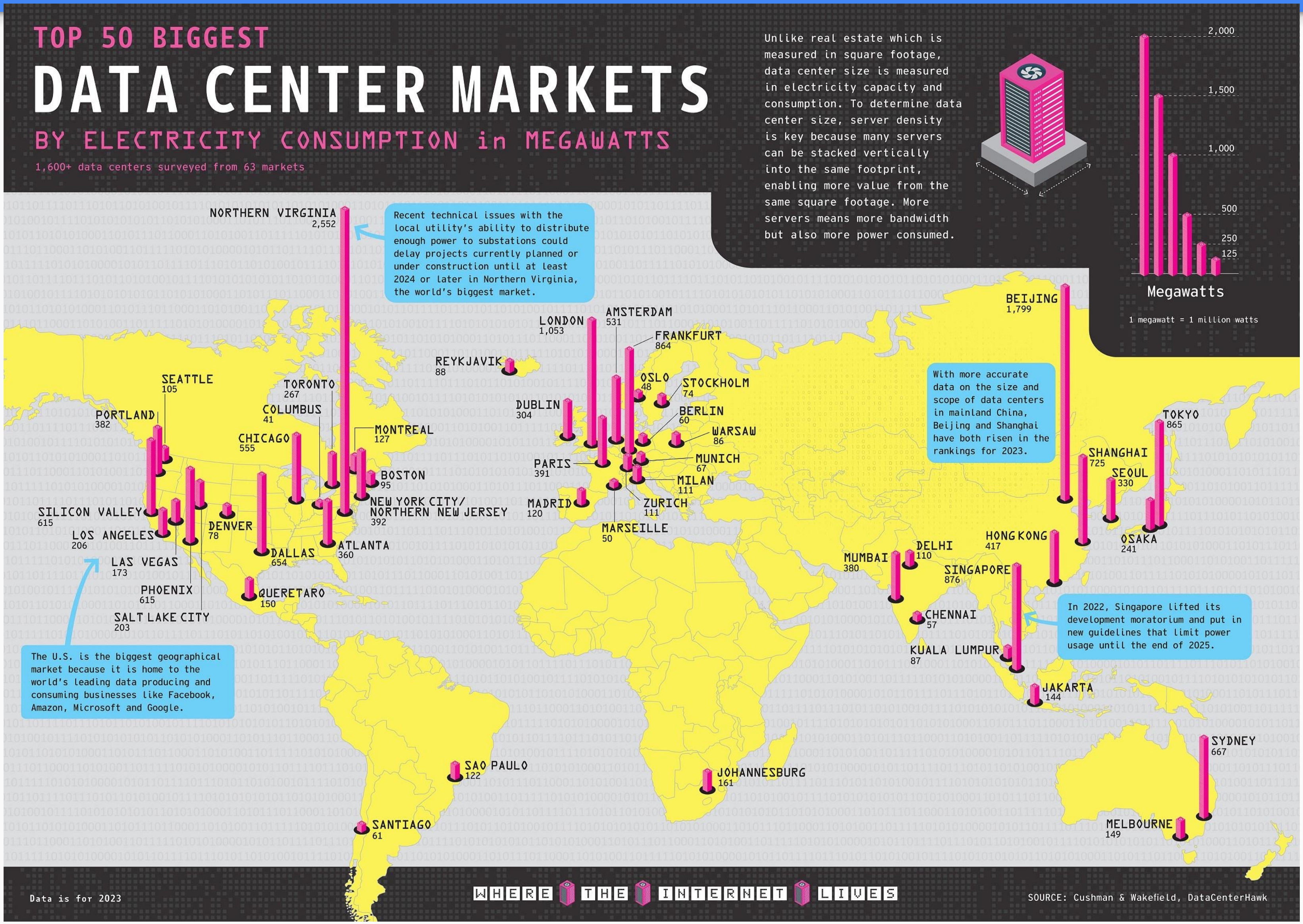
- Classical geographic scaling (and adoption) problem: to be useful, it needs to be everywhere (within a country or region)
  - and you
- Contrast: original cloud computing: start with one region and split demand
  - demand is also (mostly) fungible geographically
  - can afford to offer wide set of specialized compute services and platforms
  - e.g., AWS has 850 instance types
    - multiple generations
    - broad types with different memory and I/O trade-offs: general-purpose, compute-optimized, memory-optimized, accelerated, high-performance
    - multiple CPU (ARM, i86) and GPU types

Graceful degradation: The lesson of CDNs

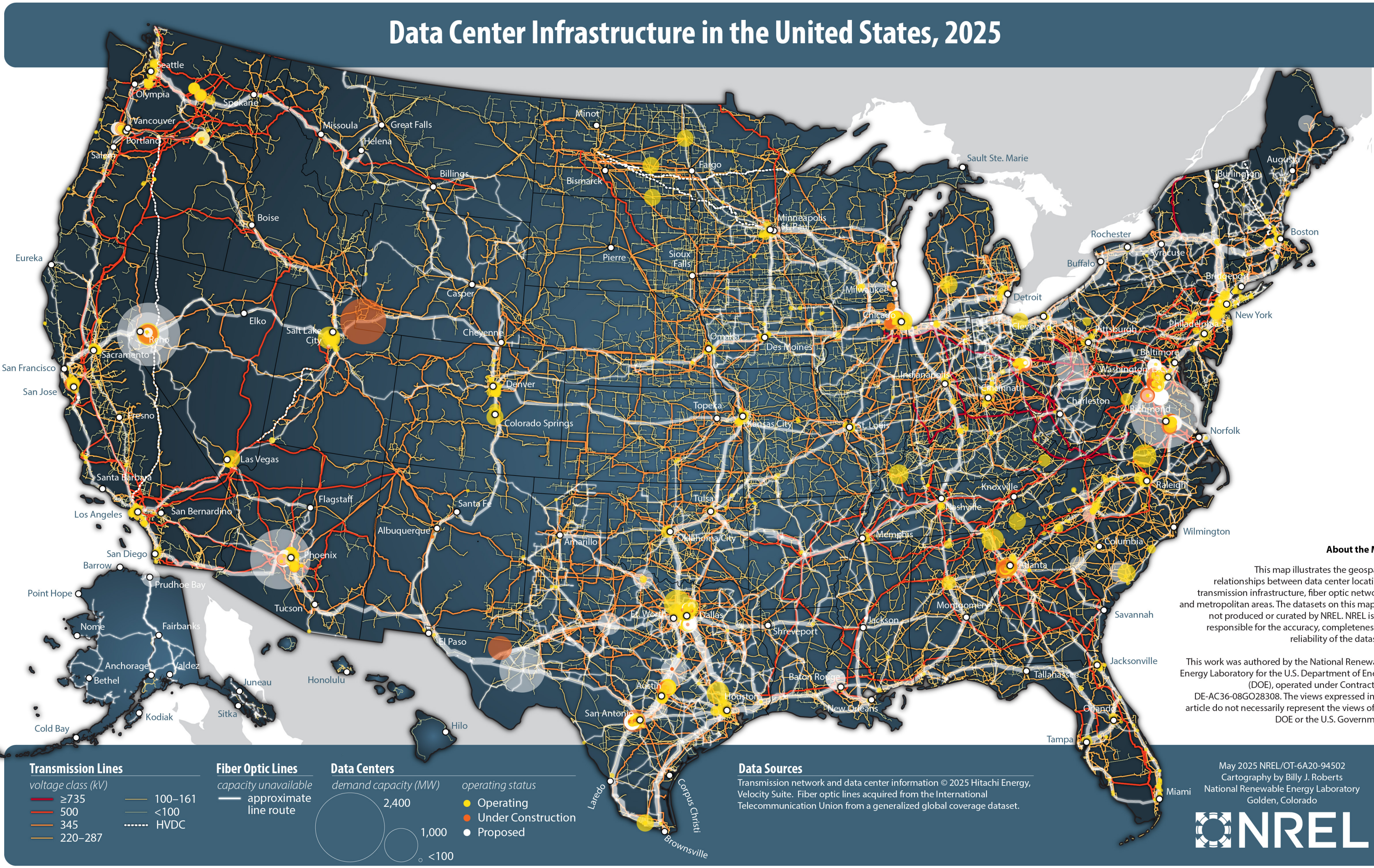
- Early edge computing: CDN
- Dominant cost: backhaul and interconnection
- Dominant cause of QoE variation: backhaul and interconnection
- Edge failures:
  - hardware & software: rare (minutes a year)
  - load (“health”)
  - no local cached copy: common
- Probabilistic performance model
  - slightly lower video quality (increasingly less so)
  - if only a small fraction fails at any point in time
    - ”Slashdot effect”
    - software upgrade
    - “reboot Manhattan”

Cost of energy

Data centers are defined by MW, not racks, CPUs or bandwidth



Data centers are driven by fiber and energy availability



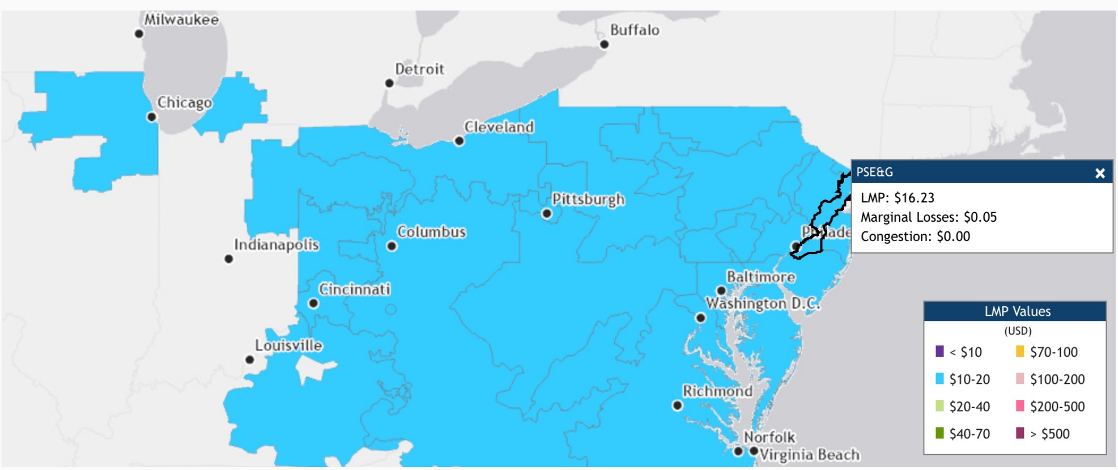
Energy costs drive data center locations

30-60% of a data center's operating expenses

"We've made savings of around 85%": Embracing green energy for data centers by migrating to Iceland

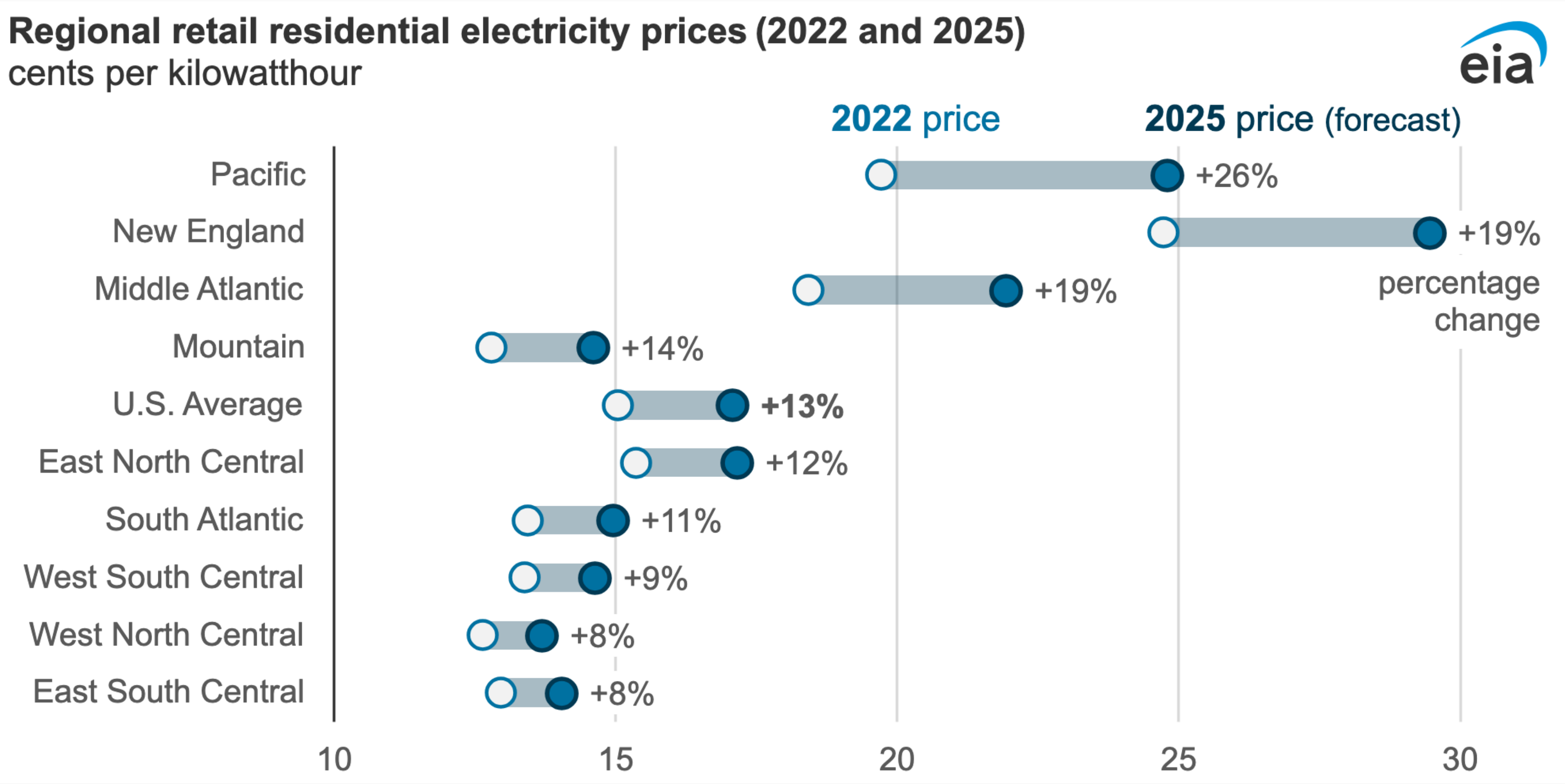
Shifting away from the UK helped Shearwater GeoServices adopt green energy for low-cost, clean seismic processing

$42/MWh

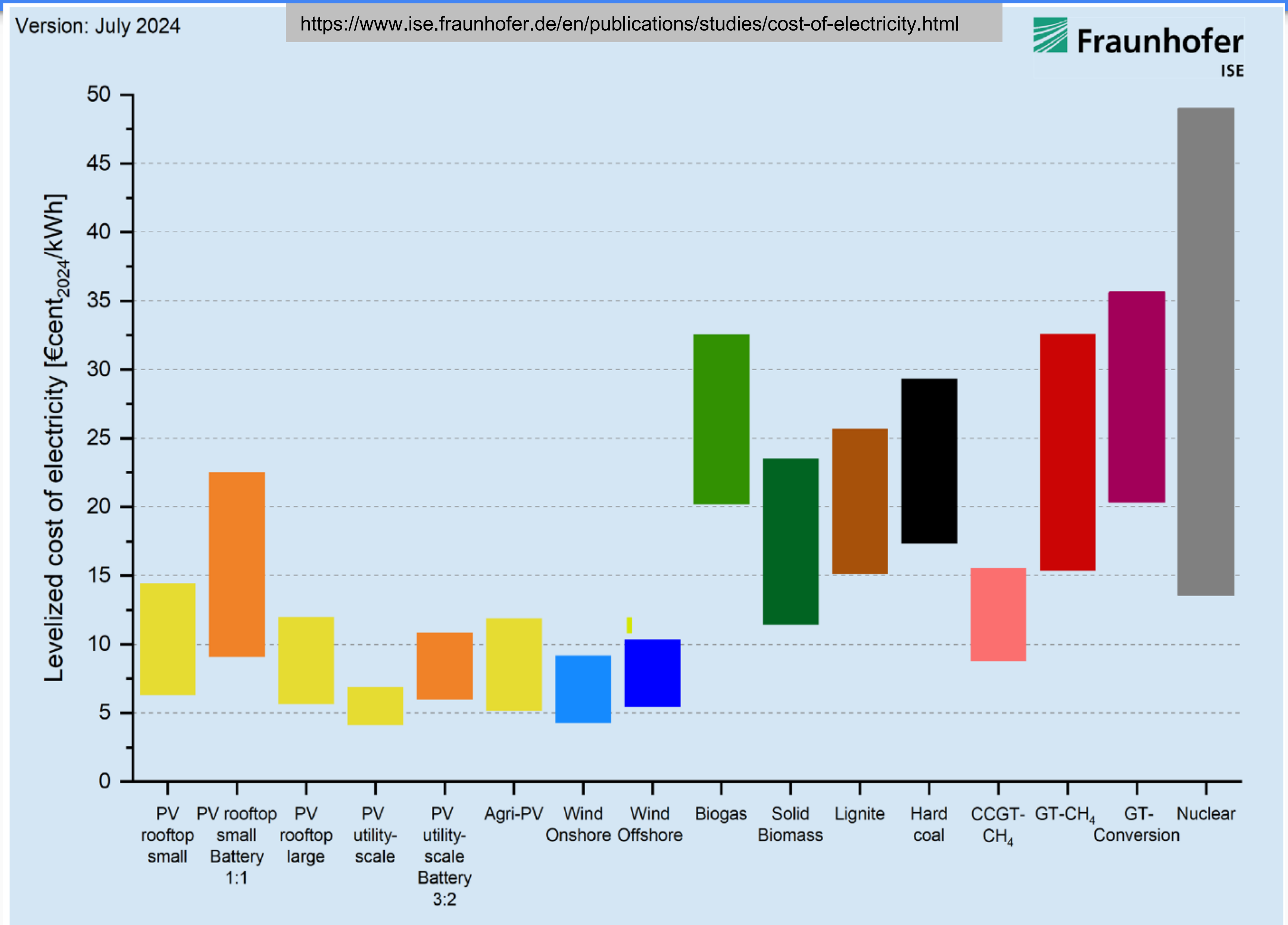


2024: $55.54 per MWh wholesale

Energy costs are rising



LCOE



Version: July 2024
https://www.ise.fraunhofer.de/en/publications/studies/cost-of-electricity.html
Fraunhofer ISE
Levelized cost of electricity [€cent$_{2024}$/kWh]
0
5
10
15
20
25
30
35
40
45
50
PV rooftop small
PV rooftop small Battery 1:1
PV rooftop large
PV utility-scale
PV utility-scale Battery 3:2
Agri-PV
Wind Onshore
Wind Offshore
Biogas
Solid Biomass
Lignite
Hard coal
CCGT-CH_4
GT-CH_4
GT-Conversion
Nuclear

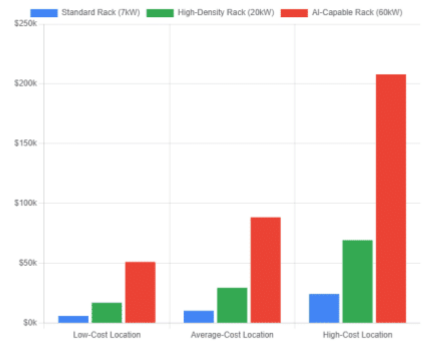
Good PUE is hard with smaller centers

The Bottom Line: Annual Power Costs

When rack density, PUE, and local electricity rates are combined, the annual cost to power a single rack can vary dramatically. The financial implications of deploying high-density AI infrastructure are staggering.

Annual Cost to Power One Rack (at Industry Average PUE of 1.58)

This chart compares the annual power cost for three rack types across different U.S. locations. Notice the exponential cost increase for AI-capable racks, especially in high-cost regions.



PUE: The Critical Efficiency Multiplier

Power Usage Effectiveness (PUE) measures how efficiently a data center uses energy. It's the ratio of total facility power to IT equipment power. A lower PUE means less energy is wasted on overhead like cooling, directly translating to lower costs.

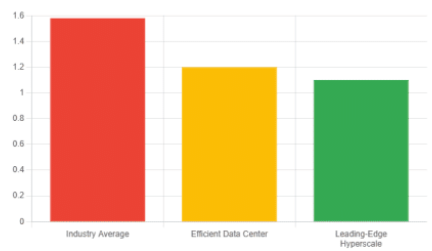
PUE Formula

$$\text{PUE} = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}}$$

An ideal PUE is 1.0, meaning zero waste. The industry average, however, shows significant room for improvement.

PUE Comparison: Average vs. Leading-Edge

Hyperscale operators like Google achieve remarkable efficiency, with a PUE approaching 1.10, while the industry average hovers around 1.58. Closing this gap is key to cost savings.



- liquid cooling
- hot-cold aisles
- server utilization

geographic variation in electricity rates: $0.0615/kWh in Iowa vs. $0.2496/kWh in Rhode Island

Small edge data centers may be efficient and effective in low-cost, rural areas where hyperscalers are far away

Trust, Security, Privacy, Complexity

Security

> Despite the many benefits, edge computing is full of challenges. For instance, decentralizing data processing brings security and privacy concerns. A friend who deployed edge systems on oil rigs had 10% of the edge computing devices stolen, along with data stored on the devices. It was encrypted, but what a huge wake-up call when systems can grow legs and walk away. That's never been a problem with the cloud.

https://www.infoworld.com/article/3709050/what-happened-to-edge-computing.html

- *Privacy:* does not need to be "edge" to be within same jurisdiction (for privacy purposes, e.g., EU)
- *Security:* not just physical security - do you know who has access to the management console?
- *Management credentials:* Containers and images store a lot of credentials (e.g., APIs, other cloud services).

Trust: The edge is also a black box

- Trust is given by enforceable commitments, not computing architecture
- Unless somebody is not just running their own system, but managing every program (i.e., a classic Unix sys admin)

Challenges for edge clouds or servers

- Hard if heterogeneous needs (e.g., ARM vs. i368; GPU models)
- Hard if different host operating systems or versions (e.g., Linux kernel)
- Transaction costs (set up accounts, payment, SLA enforcement)
- Run-time risk (even if “same,” small chance of peculiar failure)
  - hard to test and debug
- Failure impact (e.g., power outage - edge may not have UPS)
- Back-end latency: latency to cloud eco-system

Nomadic and roaming edge computing

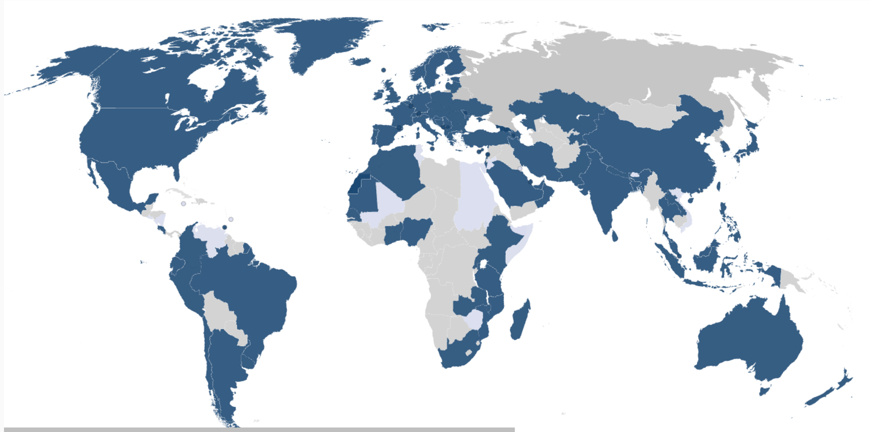
static, single provider	dynamic, single provider
static, multiple providers	dynamic, multiple providers

roaming (AAA)

need to create context (resources & data) on the fly
e.g., spin up containers
"nomadic" edge computing

The compute roaming problem

- Long-standing problem in wireless communications
    - Wi-Fi: free or eduroam
        - some attempts at paid roaming
    - Cellular roaming & eSIM
        - required regulatory intervention in EU
    - IoT via specialized MVNOs – model?



https://eduroam.org/where/

Compute roaming

- AAA – authentication, authorization, accounting
  - RADIUS and DIAMETER in eduroam and 5G
- No attempts in compute roaming
  - encourages hyperscaler economics
  - currently, requires setting up contractual and billing relationships
  - probably will require multi-cloud aggregators
  - real-time cost optimization?
- Roaming model for compute
  - authenticate via "home" network – ensures accountability
  - provides some customer protection against malicious providers (see Wi-Fi hotspots, IMSI catcher)
  - settlement mechanism – avoid retail billing by credit card
  - need temporary VPC extensions to remote edge provider
  - automated creation & deletion of resources

Compute is more than Docker containers

Almost none of these have standard APIs or definitions (e.g., functions)

Monitoring (CloudTrail)

Load balancing (ELB)

Local storage (EBS)

Identity management (IAM)

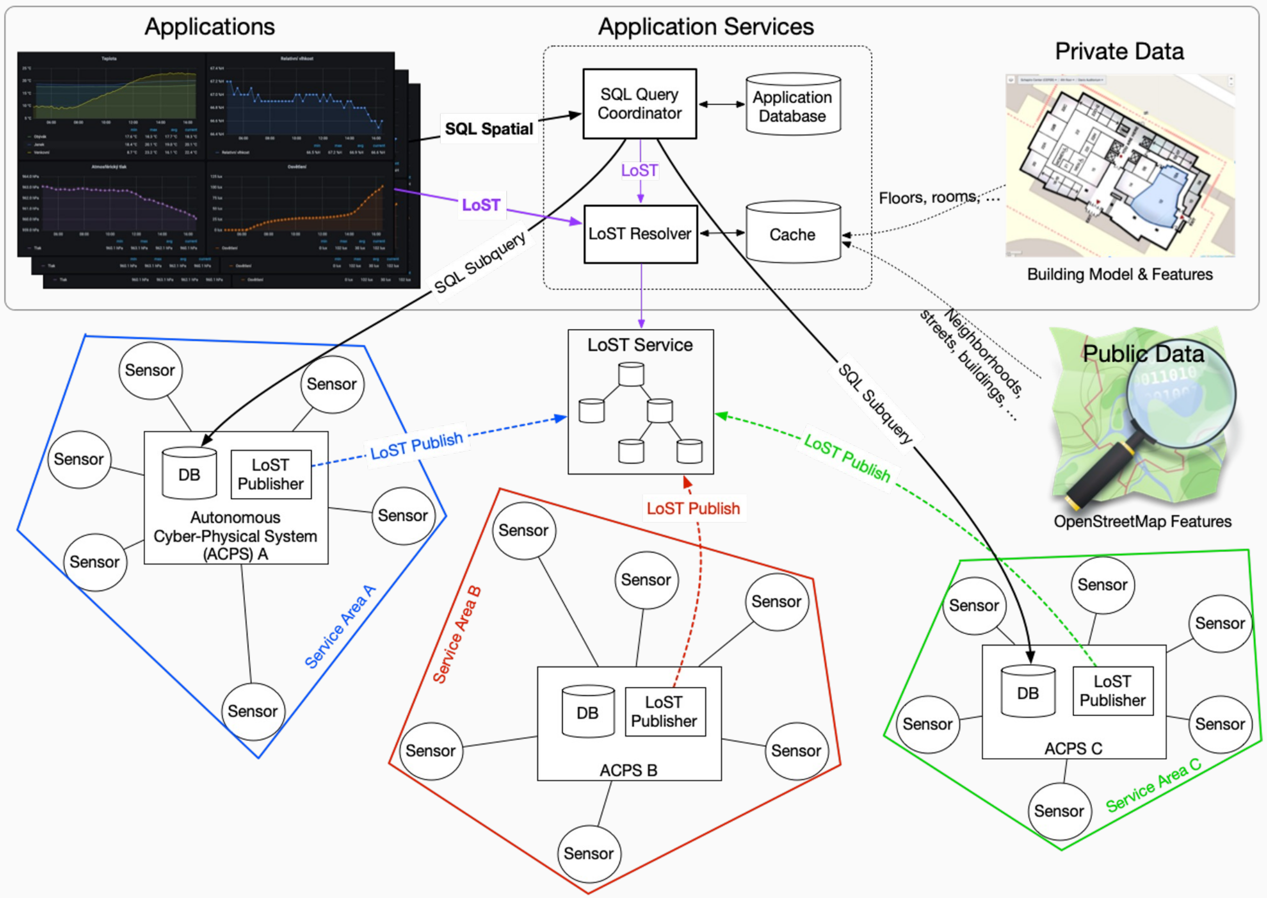
Virtual private cloud (VPC)

Databases (e.g., RAG)

with Jan Janak

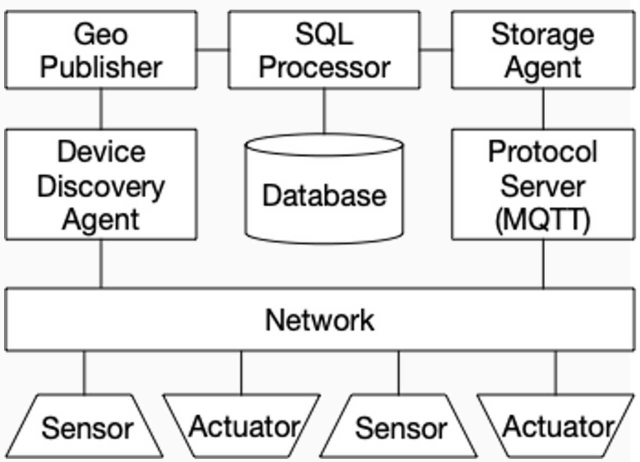
SenSQL & LoST – geographic databases and discovery

SenSQL System Architecture



Autonomous Cyber-Physical System (ACPS)

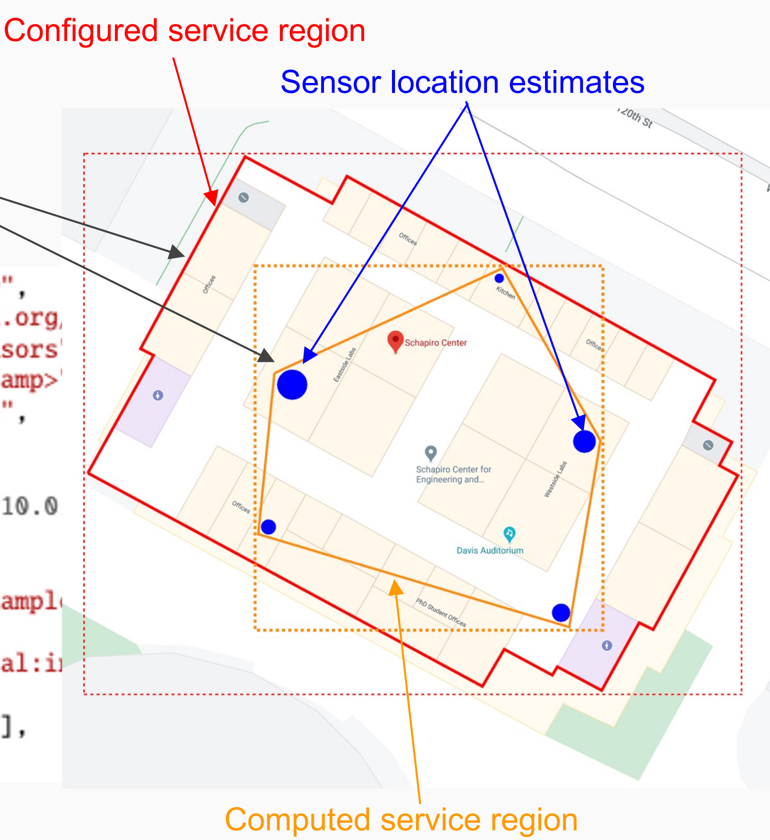
- Sensor & actuator infrastructure operated by a common administrative entity
- Geographically bounded by a **service region**
- Described by a **service descriptor**



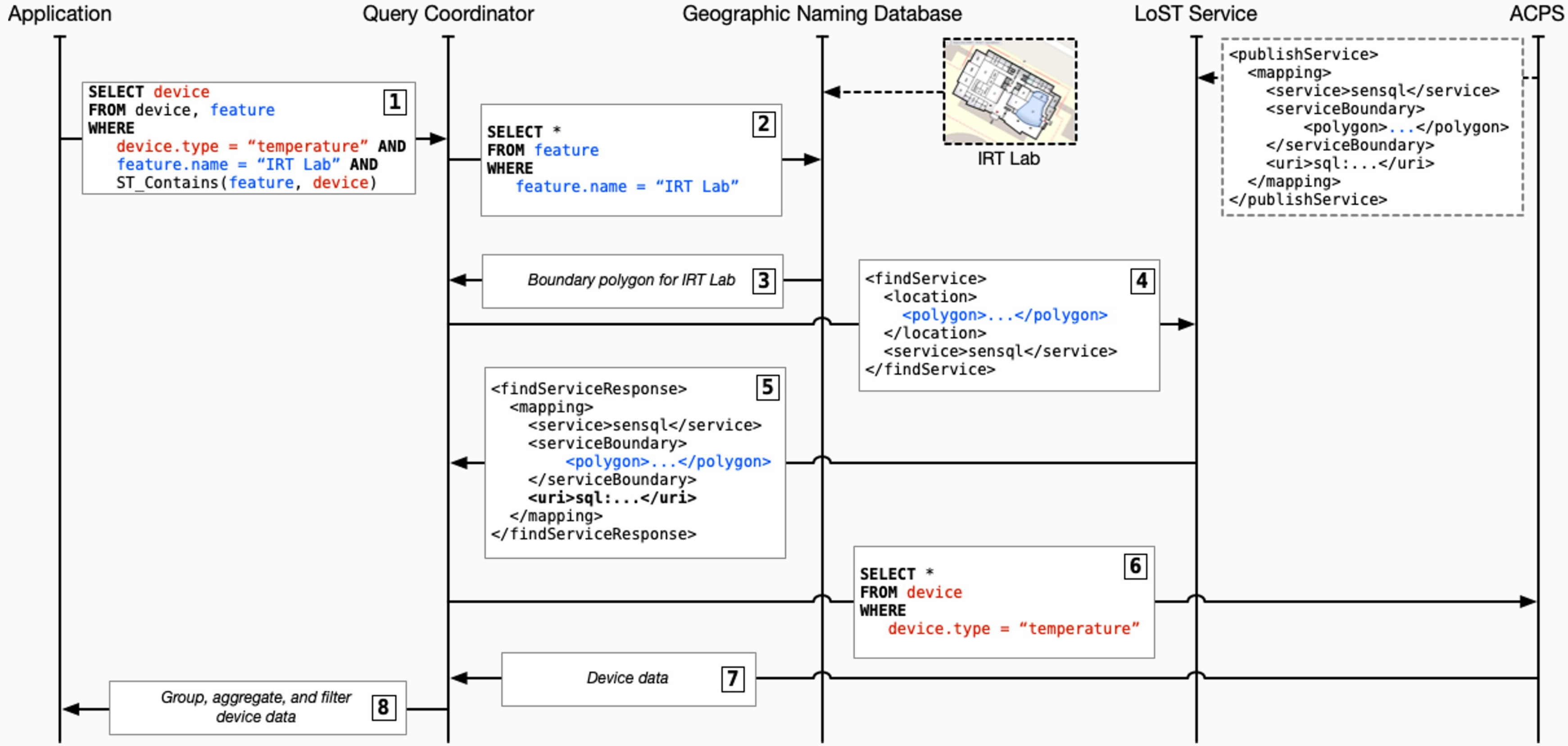
Internal architecture

```
{ "@type": "ServiceDescriptor",
  "@context": "https://schema.org,
  "displayName": "IRT Lab Sensors'
  "lastUpdated": "<ISO timestamp>'
  "expires": "<ISO timestamp>",
  "coverageRegion": {
      "type": "Polygon",
      "bbox": [-10.0, -10.0, 10.0
      "coordinates": [] },
  "private": false,
  "contact": "wss://as-123.exampl
  "sensorTypes": [
      "urn:sensor:environmental:i
      "urn:sensor:security",
      "urn:sensor:occupancy" ],
  "signature": {} }
```

Service descriptor



Original Approach (SQL Interface)

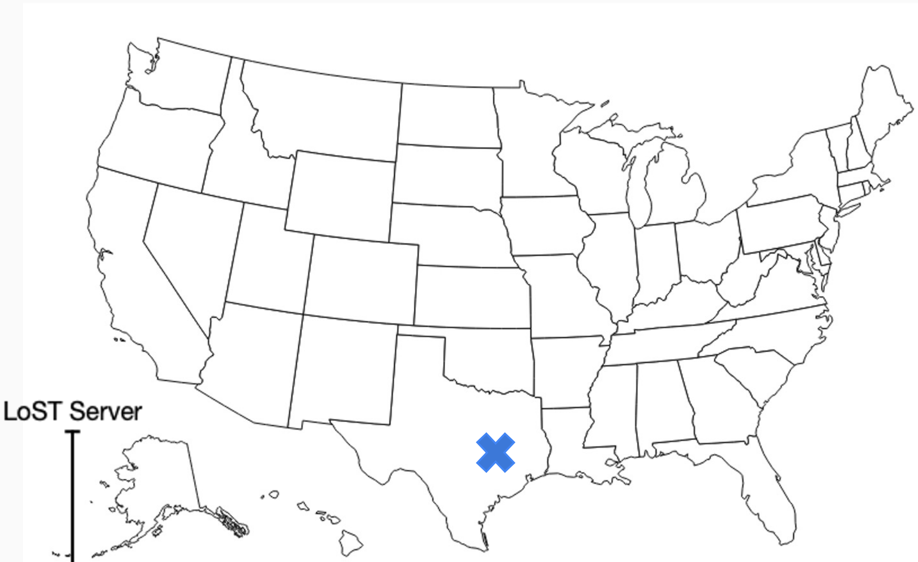


SQL Spatial Query Example

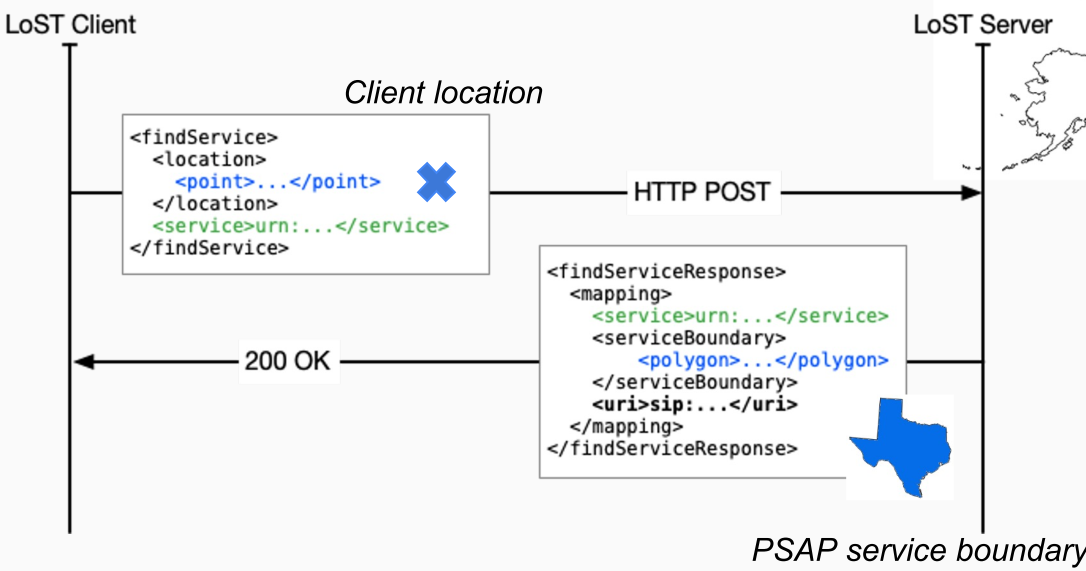
```
SELECT
    measurements.timestamp     AS ‘Time’,
    measurements.data::numeric AS ‘PM2.5'
FROM
    measurements, openstreetmap
WHERE
    ST_Contains(openstreetmap.bounds, measurements.location)
    AND openstreetmap.name     = 'Morningside Heights'
    AND measurements.quantity  = ‘PM2.5'
    AND measurements.timestamp > "2020-01-01'
ORDER BY timestamp
```

“Return the measurements ordered from January 1st, 2020 until now from all PM2.5 sensors within the Morningside Heights neighborhood.”

Background: LoST Protocol



911 Application: Lookup PSAP by client's geographic coordinates (longitude, latitude)



Pre-provisioned service boundary database

Summary & questions

- Edge computing is almost always the wrong answer
  - “Attractive nuisance” for challenge-seeking researchers
- Need clear statement of when and why edge (computing|AI) is superior - cannot assume result
  - otherwise, this is religion, not engineering
- Is this about economic decentralization, better function (how?) or lower cost (under what assumptions?)
- If provided by a single cloud provider, is there a challenge?
  - just standard load balancer & DNS-based request routing

What’s needed?

- **Resource** (not compute) discovery: sensors, APIs
- Compute costs reflecting current energy costs
  - see AWS spot instances for compute
- Resource provisioning (“roaming”) - API middlemen
  - see Twilio for VoIP and texting (and IoT universal e-SIMs)
- Generic identity and access management (beyond OAuth2 credentials)
  - see Okta and similar companies as IdS