Foundations of Edge AI

Lecture02

Background of Edge Computing

Lanyu (Lori) Xu

Email: lxu@oakland.edu
Homepage: https://lori930.github.io/
Office: EC 524

Lecture Plan

Today we will:

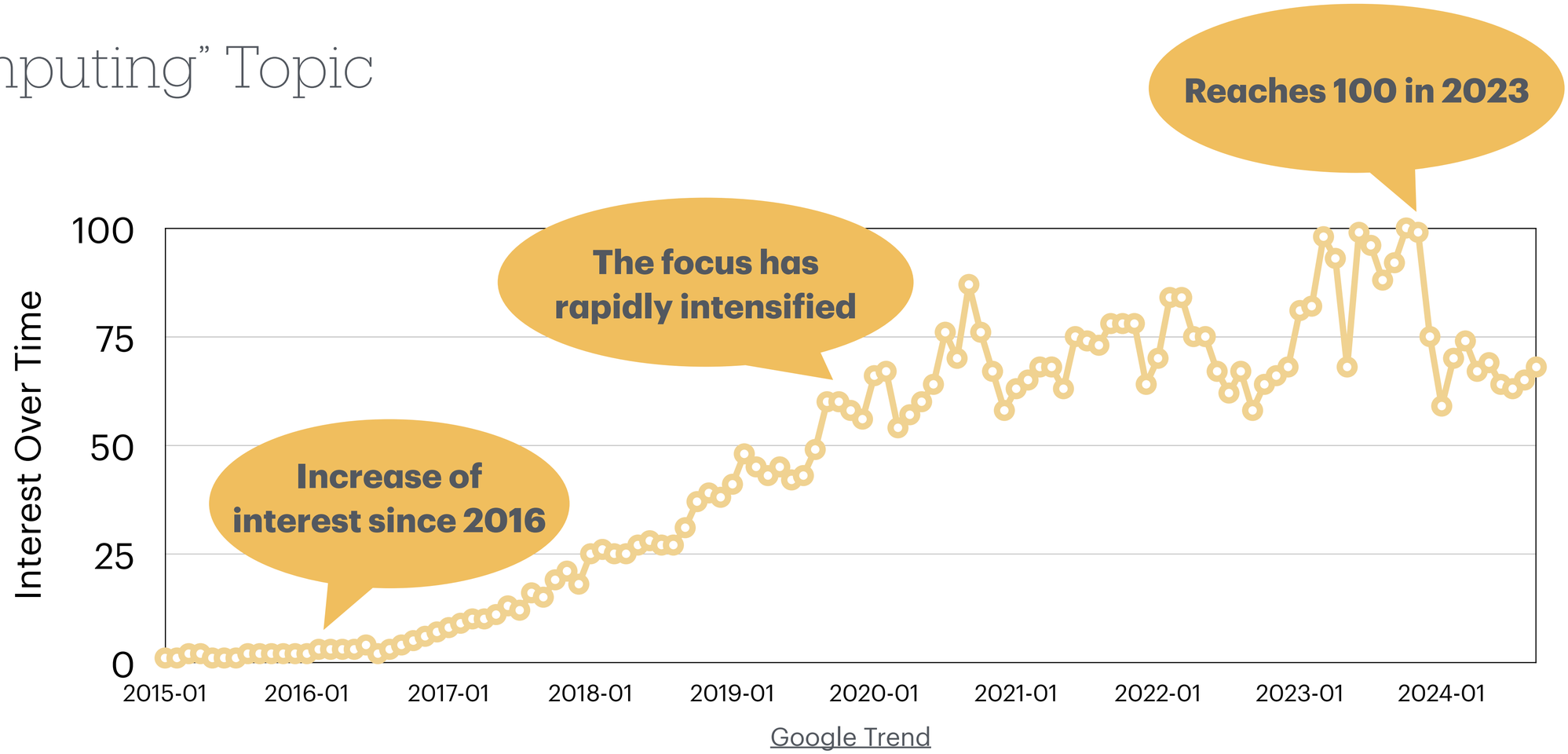
- Introduce Edge Computing
  - Definition and Motivation
  - Development History
- Introduce Cloud Computing
  - Definition
  - Characteristics
- Compare Cloud Computing and Edge Computing

Google Trends

“Edge Computing” Topic

Numbers represent **search interest** for the given region and time.

A value of **100 is the peak** popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.



What is Edge Computing?

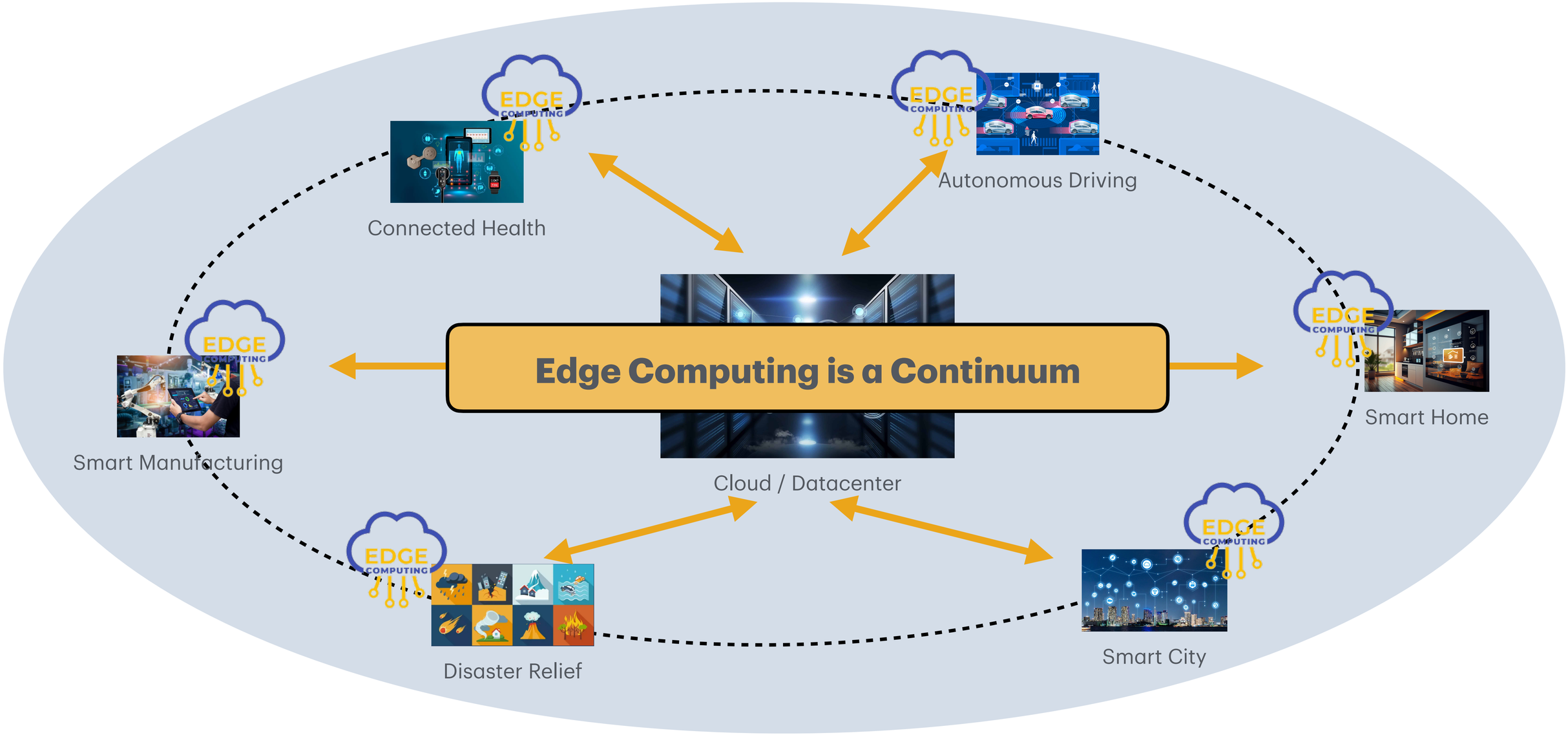
Perform computation at the “edge of the network”.

Ok, but what is the “edge of the network”?

Let’s see the answer first.

- The term “edge” is a relative concept
- Any computing, storage, and network resources along the path from the data source to the cloud data center

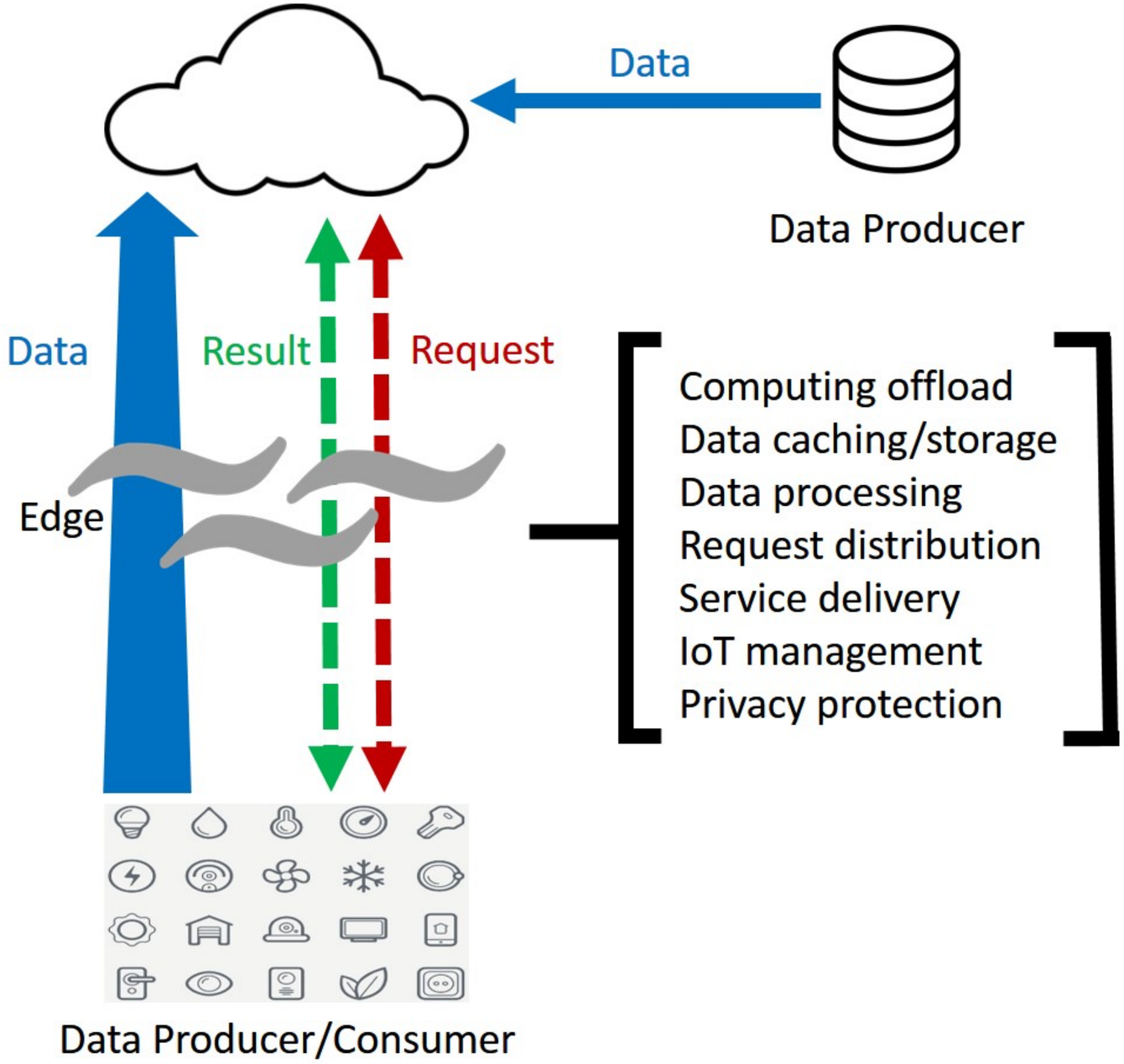
Edge: Along the Path from the Data Source to the Cloud



Computing Paradigm

Two-way computing streams

- Things are not only **data consumers** but also **data producers**.
- Computation can be performed at the edge of the network
  - On **downstream data** on behalf of cloud service
  - & **upstream data** on behalf of IoT services



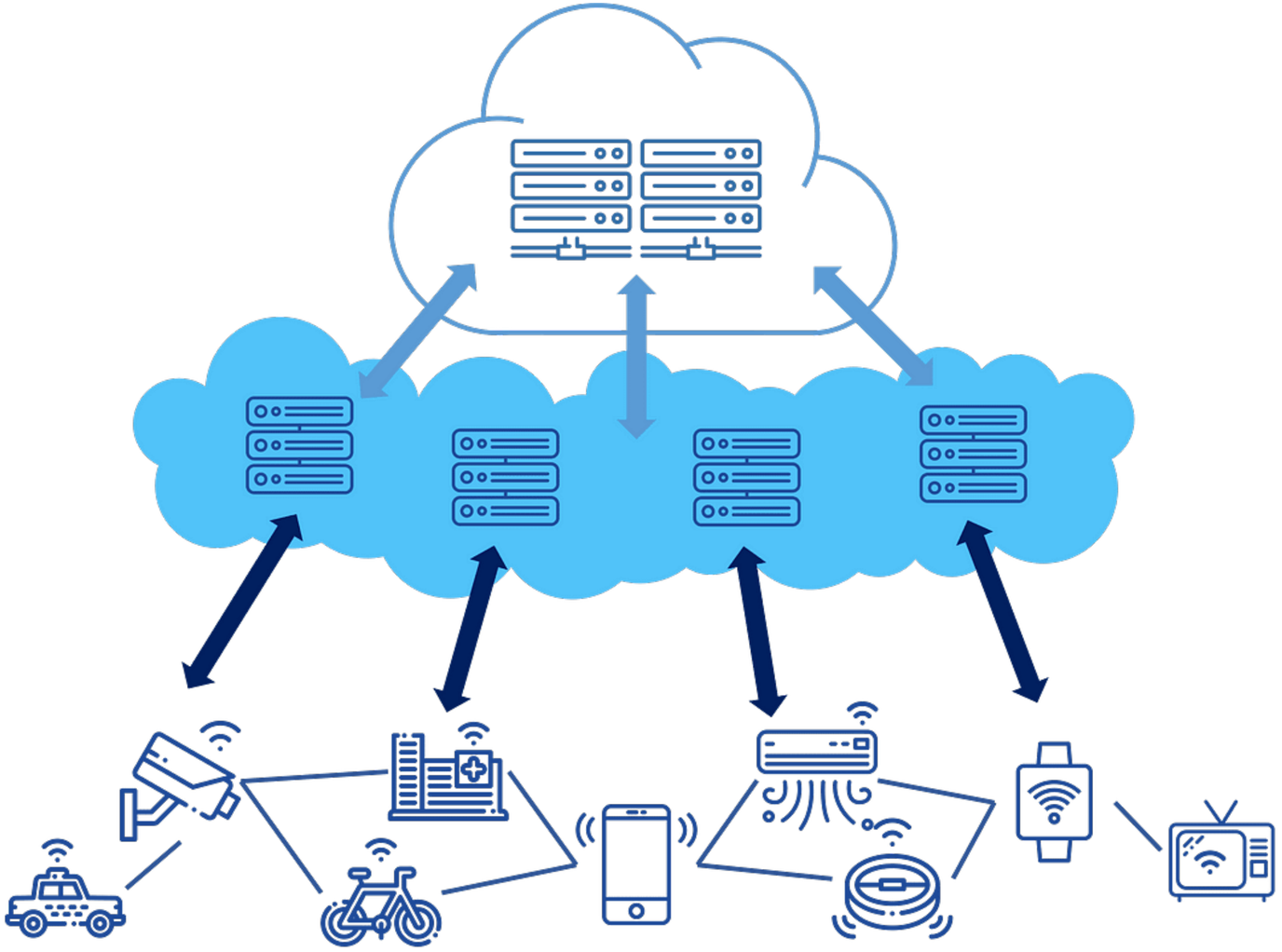
Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE internet of things journal, 3(5), 637-646.

Examples

Cloud

Edge nodes

Edge devices



The Core Concept of Edge Computing

Computing should be closer to the data source and near to the user.

Data Source

Cloud / Datacenter

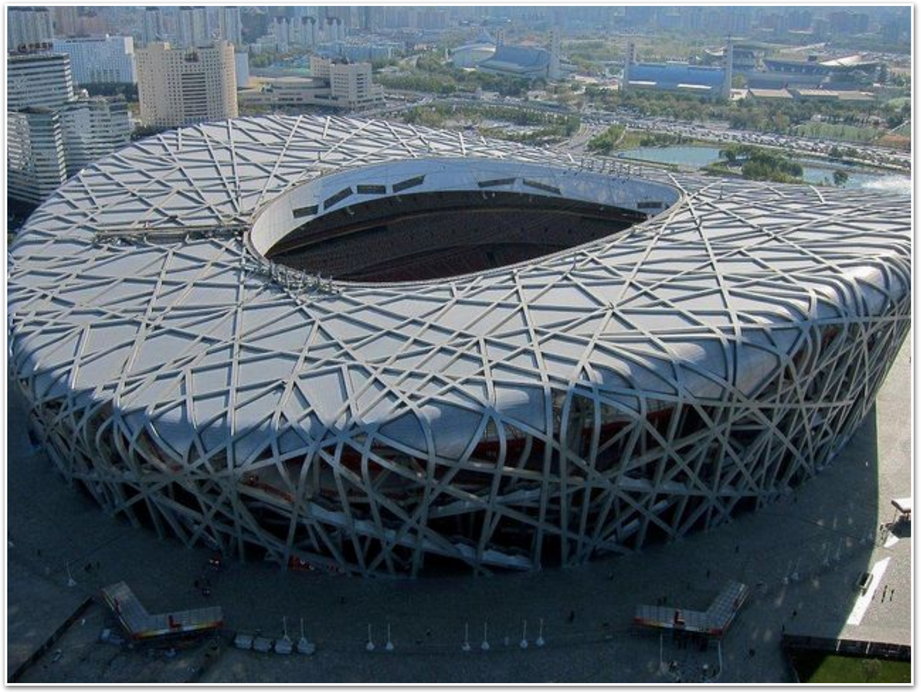
Being “Closer”

1. Shorter **network** distance / reduced scale of the network
   - Unstable factors (e.g., bandwidth, latency, jitter) are easier to control and improve
2. Proximity in **physical** space / location
   - Personalized services can be offered to the user

- **Network distance** and **spatial distance** are **not** necessarily **correlated**
- Applications can choose the appropriate computing nodes based on their specific needs

From A Biomimetic Perspective

We can draw edge computing as an analogy like this



Bird's Nest Stadium (Beijing, China)

From A Biomimetic Perspective

- The cloud data center is akin to the human brain

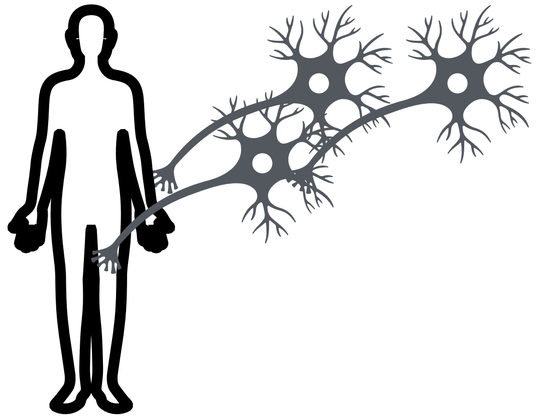
Cloud / Datacenter

=

- Edge devices resemble the human body's peripheral nerves

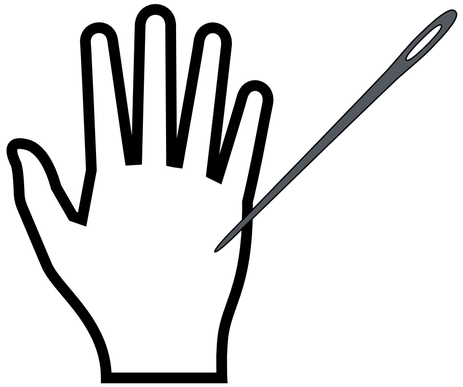
Edge

=



From A Biomimetic Perspective

- When a needle pricks the hand
  - The immediate reaction is to withdraw the hand
  - Before the brain even registers the prick action



Hand retraction
A direct, unconditioned reflex processed by the peripheral nerves

👍👍👍

- Accelerating response times
- Minimizing potential harm
- Allowing the brain to focus on more complex, high-level tasks

Back to Edge Computing

- 🤔 Relying on cloud computing to serve as the “brain” for every device?
  - 🤦‍♀️
- 👉 Edge computing equips devices with their own “brains”
  - More autonomous and immediate responses

What Retail Industry Says

The Future of Edge Computing (IBM Technology)

What Autonomous Vehicle Practitioners View

Exploring edge computing in autonomotive (Canonical Ubuntu)

Development History of Edge Computing

Data Explosion Era

- **Social media platforms**
  - Every minute, millions of photos are uploaded, and billions of likes and comments are made on platforms.

Data Explosion Era

- **Internet of Things (IoT)**
  - Billions of devices worldwide, including smartphones, home appliances, vehicles, and industrial equipment, are connected to the Internet, generate and exchange data.

Data Explosion Era

- **E-commerce Transactions**
  - Online shopping platforms process millions of transactions daily, generating vast amounts of data on consumer preferences, purchase history, and supply chain logistics.

Data Explosion Era

- **Healthcare sector**
  - Wearable devices track health metrics like heart rate, sleep patterns, physical activiy, generating massive data.

How Wearable Technology is Impacting our Lives

Data Explosion Era

- **Financial services**
  - Financial institutions handle vast quantities of data related to transactions, market trends, and customer profiles.

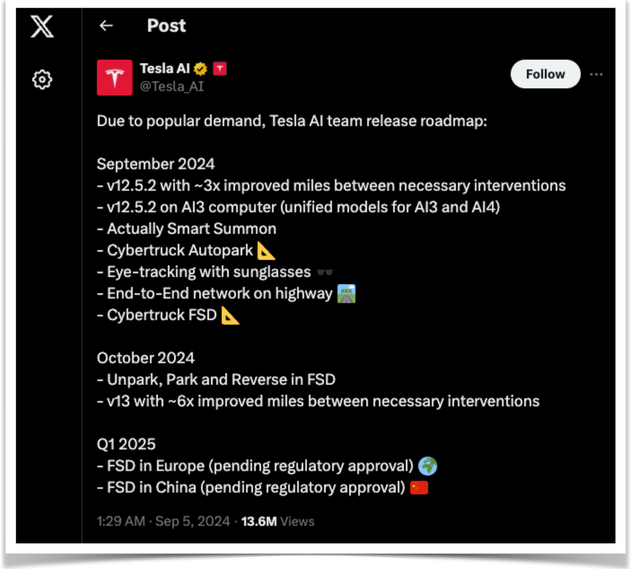
Data Explosion Era

- **Smart cities**
  - Smart infrastructures gather data on traffic patterns, energy usage, and public safety.

Global Distributed Trust Management

Data Explosion Era

- **Automotive industry**
  - Modern vehicles are equipped with numerous sensors, generating terabytes of data about vehicle performance, road conditions, and driving patterns.



Tesla AI X Account

uber to shift to autonomous taxis as wayve and cruise's self-driving AI cars join its platform

News

Baidu's driverless robotaxi service Apollo Go on the road in Wuhan, Hubei province, China on February 24, 2023. Josh Arslan/Reuters

Editor's Note: Sign up for CNN's Meanwhile in China newsletter which explores what you need to know about the country's rise and how it impacts the world.

Hong Kong (CNN) — In China, it's possible to travel six miles in a driverless taxi for just about 50 cents.

Apollo, News

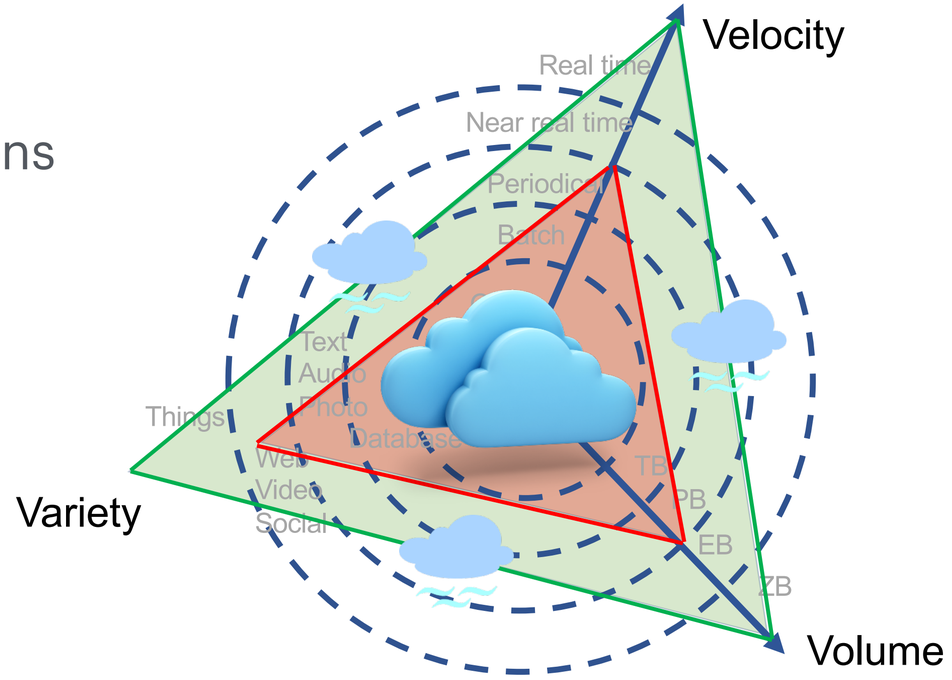
Data Explosion Era

- **Cybersecurity**
  - Organizations and security software continuously monitor network activity to detect and prevent security threats.

The growing need for cybersecurity

Data Explosion Era

- Social media platforms
- Internet of Things (IoT)
- E-commerce Transactions
- Healthcare sector
- Financial services
- Smart cities
- Automotive industry
- Cybersecurity



Data Transmission
Data Computation
Data Storage

Traditional Solutions

Shifting computational tasks from data centers to the network edge

- Distributed database
- Peer-to-Peer (P2P)
- Content Delivery Network (CDN)

🧐

Data Transmission
Data Computation
Data Storage

Traditional Solutions

Shifting computational tasks from data centers to the network edge

- **Distributed database**
- Peer-to-Peer (P2P)
- Content Delivery Network (CDN)

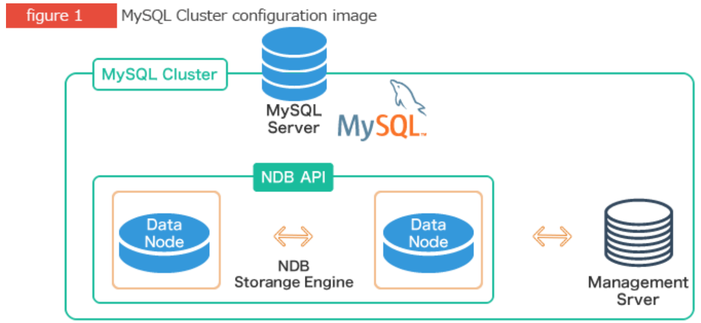
🧐

Data Transmission
Data Computation
Data Storage

Distributed Database

Based on the structure of database systems

- **Homogeneous** systems
  - **Identical** software and hardware
  - Share a **single**-access interface
- **Heterogeneous** systems
  - Hardware, operating systems, database management systems, and data models **vary**
  - e.g., Apache Cassandra + PostgreSQL, MongoDB + MySQL, Microsoft SQL Server + Oracle DB



MySQL Cluster

Distributed Database

Based on the type of data

- SQL (relational)
  - MySQL, PostgreSQL, Oracle, Microsoft SQL Server
- **NoSQL (non-relational)**
  - MongoDB (JSON-like document), HBase (wide-column), Redis (key-value)
- **NewSQL (relational)**
  - Google Spanner (former BigTable), CockroachDB, VoltDB, NuoDB

Distributed Database vs. Ede Computing

- Facilitate **data storage** and sharing in big data environments
- With less emphasis on the heterogeneous computing and storage capacities of the devices they operate on

Distributed database is a cornerstone technology in big data processing

- Store data on edge devices
- High privacy, reliability, and availability
- Recognizing the heterogeneity of endpoint architectures and the need for supporting various application services

Traditional Solutions

Shifting computational tasks from data centers to the network edge

- Distributed database
- **Peer-to-Peer (P2P)**
- Content Delivery Network (CDN)

🧐

Data Transmission
Data Computation
Data Storage

Peer-to-Peer (P2P) Computing

- Participants (peers) share resources and communicate directly with each other **w/o centralized control**.
- Closely related to edge computing
- Represents an earlier technology for migrating computing to the network edge
  - Initially used for file transfer
- The term “P2P” was first proposed in the Year 2000.

For file sharing


For cryptocurrency

For distributed web


Initially P2P VoIP

For video hosting

Peer-to-Peer (P2P) Computing

- Evolved into **a significant subfield of distributed systems**
- Related **core research topics**
  - Decentralization
  - Scalability maximization
  - Tolerance to the loss of high-level nodes
  - Prevention of malicious activities

P2P Computing vs. Ede Computing

- *Structure*
  - **Decentralized network architecture** where each node (peer) acts both as a client and a server
- *Usage*
  - Resource sharing
- *Characteristics*
  - Decentralization: allowing the network to function w/o central servers

- *Structure*
  - **Distributed computing** moves computing tasks from centralized data centers to the edge of the network, closer to the data source or user
- *Usage*
  - IoT, autonomous vehicles, smart cities, ...
- *Characteristics*
  - Bring **computing capabilities** to the location where data is generated
  - Fast responsiveness

Traditional Solutions

Shifting computational tasks from data centers to the network edge

- Distributed database
- Peer-to-Peer (P2P)
- **Content Delivery Network (CDN)**

What is Content Delivery Network (Tech Arkit)

Historical evolution of edge computing

🧐

Data Transmission
Data Computation
Data Storage

Content Delivery Networks (CDNs)

- Aim to reduce data download latency from remote sites
- Accelerate content delivery by deploying cache servers at the network edge
- Prominent companies like **Amazon** and **Akamai** have developed mature CDN technologies
  - Proposed by Akamai in the Year 1998

CDNs vs. Edge Computing

- In the **early stage** of edge computing
  - The “edge” was limited to CDN cache servers
- The “edge” in **modern edge computing**
  - Any computing, storage, and network resources situated along the path from the data source to the cloud computing center
  - Edge computing places a greater emphasis on computational capabilities

Cloud Computing

What is Cloud Computing?

- A way of providing services that allow applications to access and use **computational**, **network**, and **storage resources** through the Internet.

Cloud Computing Services

X as a Service (XaaS)

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Database as a Service (DBaaS)
- Artificial Intelligence as a Service (AIaaS)
- Function as a Service (FaaS)

Cloud Computing Services

X as a Service (XaaS)

- **Infrastructure as a Service (IaaS)**
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Database as a Service (DBaaS)
- Artificial Intelligence as a Service (AIaaS)
- Function as a Service (FaaS)

Definition

- Provide **virtualized computing resources** over the internet
- Businesses: rent or lease servers for computing and storage in the cloud

Example

- Amazon Web Services (AWS) Elastic Compute Cloud (EC2)
- Google Cloud Compute Engine
- Microsoft Azure Virtual Machines

Cloud Computing Services

X as a Service (XaaS)

- Infrastructure as a Service (IaaS)
- **Platform as a Service (PaaS)**
- Software as a Service (SaaS)
- Database as a Service (DBaaS)
- Artificial Intelligence as a Service (AIaaS)
- Function as a Service (FaaS)

Definition

- Provide a **platform** allowing customers to **develop, run, and manage applications**
- Without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app
- Delivered over the web: a suite of tools for application development and deployment

Example

- Google App Engine
- Microsoft Azure App service
- IBM Cloud Foundry
- Salesforce Platform (force.com)

Cloud Computing Services

X as a Service (XaaS)

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- **Software as a Service (SaaS)**
- Database as a Service (DBaaS)
- Artificial Intelligence as a Service (AIaaS)
- Function as a Service (FaaS)

Definition

- Provide **software applications over the Internet**, on a subscription basis
- Eliminates the need for organizations to install and run applications on their own computers or in their data centers
- Eliminates the expense of hardware acquisition, provisioning, maintenance, software licensing, installation, and support

Example

- Google Workspace
- Microsoft 365
- Slack, Discord
- Dropbox
- Zoom

Cloud Computing Services

X as a Service (XaaS)

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- **Database as a Service (DBaaS)**
- Artificial Intelligence as a Service (AIaaS)
- Function as a Service (FaaS)

Definition

- Provide **a fully managed database** including backup, updates, performance optimization
- Without needing to handle underlying infrastructure, maintenance, or scaling

Example

- Amazon Relational Database Service (RDS)
- Google Cloud SQL
- Microsoft Azure SQL Database
- MongoDB Atlas

Cloud Computing Services

X as a Service (XaaS)

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Database as a Service (DBaaS)
- **Artificial Intelligence as a Service (AIaaS)**
- Function as a Service (FaaS)

Definition

- Provide **pre-built AI tools and infrastructure**, simplifies access to advanced AI technologies, allowing for scalable and cost-effective AI deployment
- Businesses: integrate AI capabilities (ML, NLP, CV) into their applications w/o building or maintaining the underlying AI systems

Example

- Google Cloud Ai Plotform
- Microsoft Azure Cognitive Services
- AWS AI Services (Rekognition, Polly, Lex)
- IBM Watson

Cloud Computing Services

X as a Service (XaaS)

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Database as a Service (DBaaS)
- Artificial Intelligence as a Service (AIaaS)
- **Function as a Service (FaaS)**

Definition

- Provide **serverless** models that allow developers to execute individual functions or pieces of code in response to events
- Automatically scales based on demand, charging only for the compute time used during execution

Example

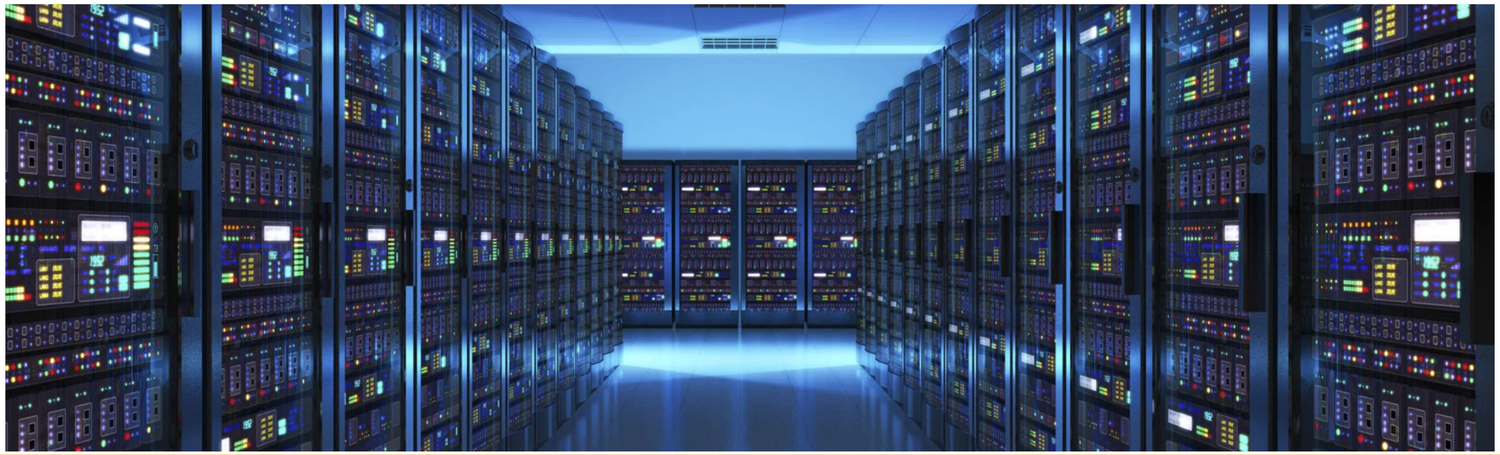
- AWS Lambda
- Google Cloud Functions
- Microsoft Azure Functions
- IBM Cloud Functions


XaaS

Characteristics of Cloud Computing

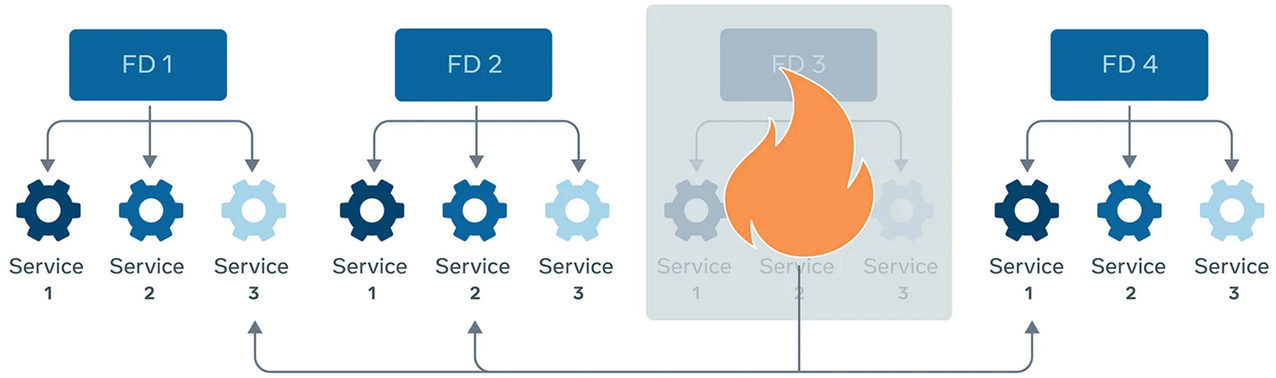
Large Scale of Central Servers

- The number of servers **varies** depending on their size, purpose, and design.
- **Hyperscale** data centers
  - tens of thousands - hundreds of thousands
- **Smaller** facilities
  - A few hundred - a few thousand



High Reliability

- **Distributed** server cluster design
  - If one server fails, the others can take over, maintaining the platform's overall reliability
  - Avoid losing data or service



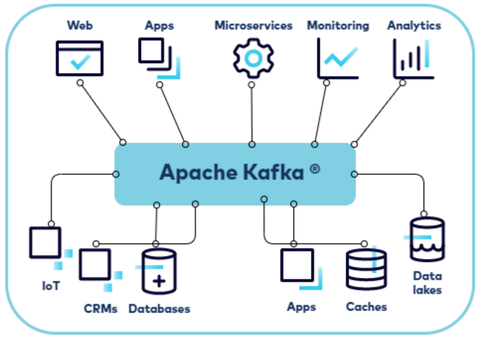
Fault tolerance

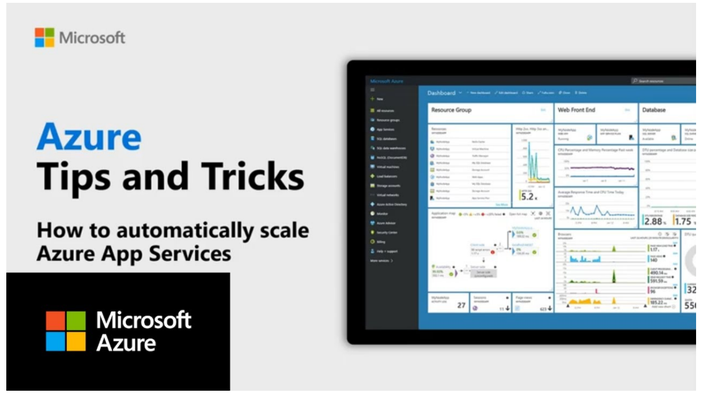
Scalability

- **Dynamically allocate** or release resources based on specific user needs
  - If a user needs more resources, the cloud can quickly allocate corresponding resources
  - Otherwise, the cloud will release the taken resources

on-call

Walmart Cloud Native Platform (WCNP)

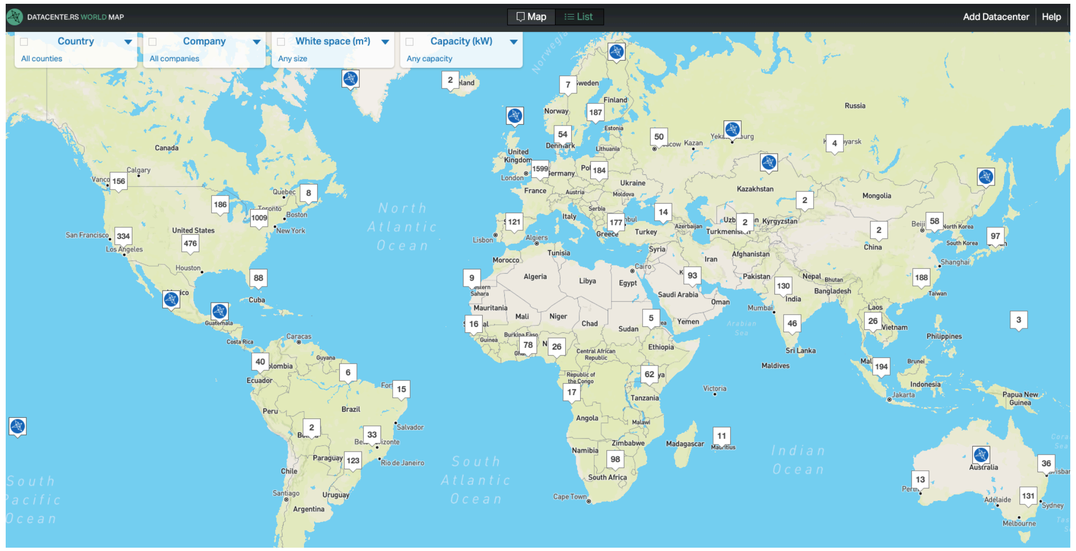




How to automatically scale Azure App Services

Virtualization

- Combine resources from different locations into a **logically unified pool**
  - Hides the underlying complexity and heterogeneity of the hardware
  - Offers a unified resource management and deployment manner



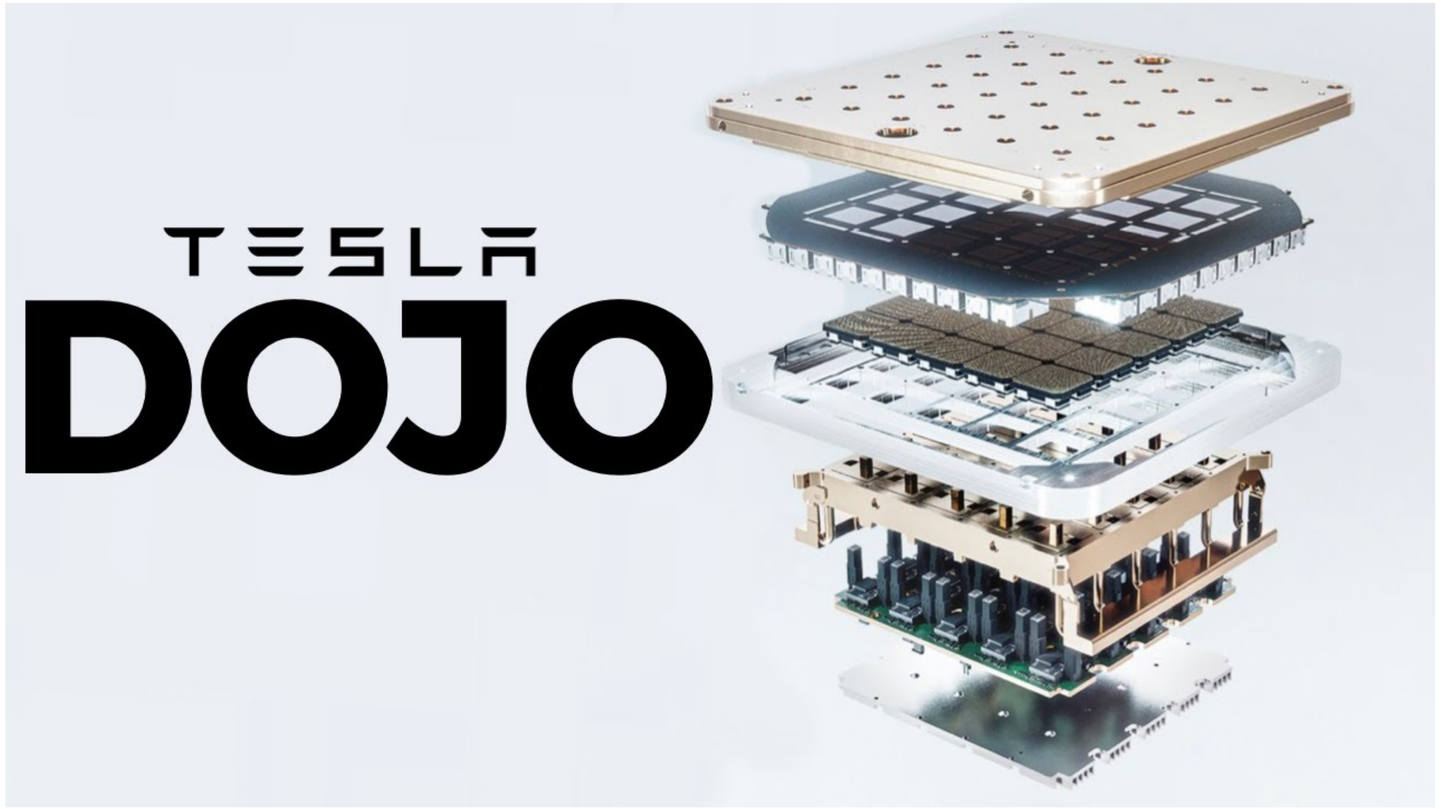
https://datacente.rs/

Drawbacks?

Does the cloud data center itself have any drawbacks?

💰💰💰💰💰💰

Tesla's supercomputer for computer vision video processing and recognition



Training tile in Tesla Dojo

- Millions of the most advanced GPU chips from Nvidia
  - A cost not just companies but even some small countries could not afford.
- These chips might NOT be available for purchase
- Annual AI training
  - ~ $2 billion USD for operation and maintenance

Google 2024 Environmental Report

https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf

WATER

Global operational water	Unit	2019	2020	2021	2022	2023
Water consumption	Million gallons	3,412.4	3,748.9	4,561.8	5,564.7	6,352.0 ✓
Water discharge	Million gallons	1,748.3	1,939.8	1,734.8	2,034.9	2,301.3 ✓
Water withdrawal	Million gallons	5,160.7	5,688.7	6,296.6	7,599.6	8,653.3 ✓

Data centers total	**Million gallons**	**7,657.2** ✓	**1,556.6** ✓	**6,100.6** ✓
Potable water		5,984.6		
Non-potable water		917.5		
Reclaimed wastewater		755.1		

Enough for 274K people to drink during their whole lives.
* 3 liters/day, to 80 years old.

Datacenters in Arizona

- Close to CA
- Less cost (utility, land, ...)
- Friendly tax policies

City of Mesa Obligations

Water

- The City proposes to make an initial 1,120 AFY of water supply available for the project, and should they meet the Performance Milestones, they will have the ability to reach 4,480 AFY of supply.
- Project Red Hawk will also have sufficient emergency back-up supplies for water, which includes utilizing on-site storage, existing wells on their site, acquiring their own water rights & supplies, and transferring long-term storage credits to the City.

City of Mesa Obligations

Water

- The City proposes to make an initial 1,120 AFY of water supply available for the project, and should they meet the Performance Milestones, they will have the ability to reach 4,480 AFY of supply.
- Project Red Hawk will also have sufficient emergency back-up supplies for water, which includes utilizing on-site storage, existing wells on their site, acquiring their own water rights & supplies, and transferring long-term storage credits to the City.

mesa·az
ECONOMIC DEVELOPMENT

CBS News

Google's water use is soaring in The Dalles, records show, with two more data centers to come

Updated: Feb. 22, 2023, 10:17 a.m. | Published: Dec. 17, 2022, 7:04 a.m.

Steam rises above the cooling towers in The Dalles data center in Oregon. These plumes of water vapor create a mist at dusk. Google photo

Oregonlive

Google's data centers used 355 million gallons of The Dalles' water last year, **29% of the city's total water consumption.**

Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

Pengfei Li
UC Riverside

Jianyi Yang
UC Riverside

Mohammad A. Islam
UT Arlington

Shaolei Ren[1]
UC Riverside

Abstract

The growing carbon footprint of artificial intelligence (AI) models, especially large ones such as GPT-3, has been undergoing public scrutiny. Unfortunately, however, the equally important and enormous water (withdrawal and consumption) footprint of AI models has remained under the radar. For example, training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly evaporate **700,000 liters** of clean freshwater, but such information has been kept a secret. More critically, the global AI demand may be accountable for 4.2 – 6.6 billion cubic meters of water withdrawal in 2027, which is more than the total annual water withdrawal of 4 – 6 Denmark or half of the United Kingdom. This is very concerning, as freshwater scarcity has become one of the most pressing challenges shared by all of us in the wake of the rapidly growing population, depleting water resources, and aging water infrastructures. To respond to the global water challenges, AI models can, and also must, take social responsibility and lead by example by addressing their own water footprint. In this paper, we provide a principled methodology to estimate the water footprint of AI models, and also discuss the unique spatial-temporal diversities of AI models' runtime water efficiency. Finally, we highlight the necessity of holistically addressing water footprint along with carbon footprint to enable truly sustainable AI.

	Topic	Target	Unit	2022	2023	Target year
Net-zero carbon	**Achieve net-zero emissions across all of our operations and value chain by 2030**					
	Carbon-free energy	Run on 24/7 carbon-free energy on every grid where we operate by 2030	% global average carbon-free energy	64%	64%	2030
	Carbon reduction	Reduce 50% of our combined Scope 1, 2 (market-based), and 3 absolute emissions (compared to our 2019 base year) by 2030 [181]	% reduction since 2019	30% increase	48% increase	2030
Water stewardship	**Replenish more water than we consume and help improve water quality and ecosystem health in the communities where we operate**					
	Water replenishment	Replenish 120% of the freshwater volume we consume, on average, across our offices and data centers by 2030	% freshwater replenished	6%	18%	2030

Google 2024 Environment Report

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. arXiv preprint arXiv:2304.03271.

Cloud Computing vs. Edge Computing

	Edge Computing	Cloud Computing
Target Application	Mobile applications and IoT	General Internet applications
Server Node Location	Edge of the networks (e.g., gateways, WiFi access points, cellular base stations)	Data centers
Server & Client Communication	Wireless local area network, 4G/5G/6G, etc.	Wide area network
Number of served devices	Billions	Millions
Service Type	Services based on local information	Services based on global information

Benefits

Cloud Computing

- **Lower upfront cost**
  - Get applications to market quickly
- **Flexible pricing**
  - Pay-as-you-go
- **Limitless compute on demand**
  - React and adapt to changing demands instantly
- **Simplified IT management**
  - Users can access IT support and focus on the business's core needs
- **Easy updates**
  - Get the latest hardware, software, and services
- **Reliability**
  - Data backup, disaster recovery, business continuity
- **Save time**
  - No more configuring private servers and networks

Edge Computing

- **Lower latency**
  - Reduced data travel.
  - Benefit fully autonomous vehicles, augmented reality, etc.
- **Reduced cost**
  - LAN: higher bandwidth and storage, lower cost
  - Less data needs to travel to the cloud
- **Model accuracy**
  - Lower the data size to fed into a cloud model when bandwidth is low
  - Data feedback loops can improve edge AI model accuracy
- **Wider reach**
  - No need for internet access. Extends the range of computing to previously inaccessible or remote locations
- **Data sovereignty**
  - Reduced exposure to cybersecurity attacks, more compliance with data laws

Edge Computing is to Complement Cloud Computing

- Edge computing provides a better computing platform for **mobile devices, IoT, and more**.
- Edge computing is NOT intended to replace cloud computing but rather to **complement** and **extend** cloud computing.

Data Source

- **Robust computational power**
- **Vast storage support**

- **Real-time processing**
- **Privacy protection**
- **Reduced energy/bandwidth consumption**

Cloud / Datacenter

References

- Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE internet of things journal, 3(5), 637-646.

- Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. arXiv preprint arXiv:2304.03271.