Foundations of Edge AI

Lecture03

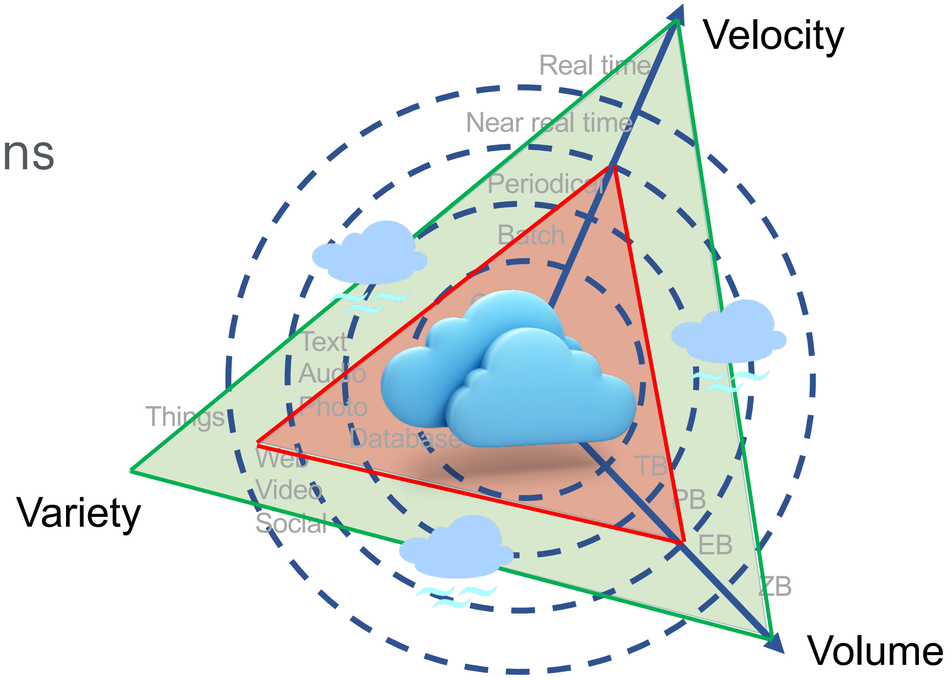
Edge Computing Architecture

Lanyu (Lori) Xu

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

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

Lecture Plan

Today we will:

- Introduce Edge Computing Architecture
  - Different layers
  - Capabilities of Edge Infrastructure
- Discuss Computing Models
  - Independent Model
  - Collaborative Model
- Talk about Lab 1

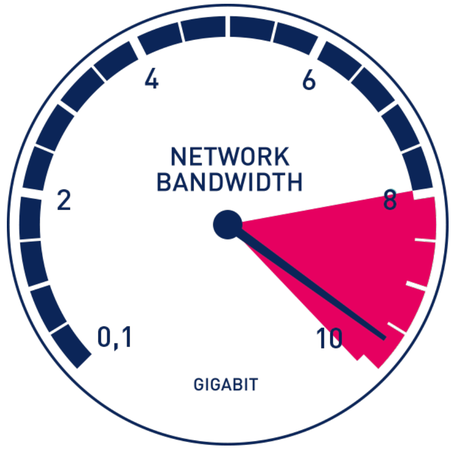
Use one word to describe the structural difference between Cloud Computing and Edge Computing

Distributed.

Centralized modes rely on powerful data centers.

How is edge infrastructure uniquely constructed to meet demands?

What are the Demands/Goals?



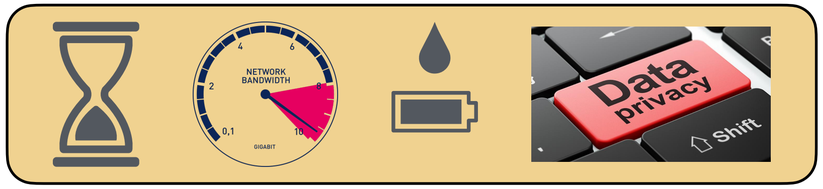
Enhancing real-time capability

Diminishing bandwidth demand

Mitigating energy consumption

Safeguarding data security and privacy

Different Layers of Edge



Cloud

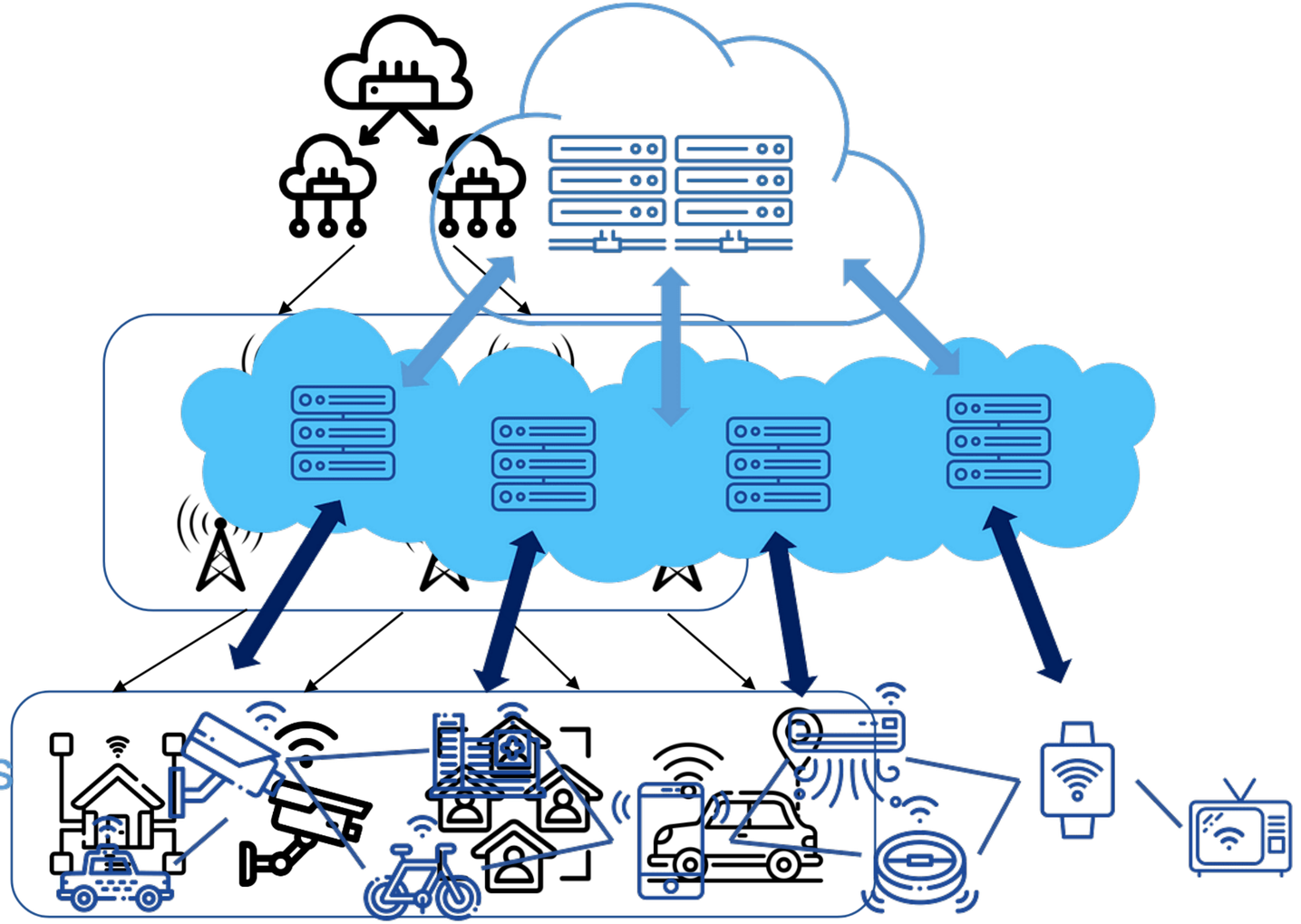
Data Center Edge

Edge nodes

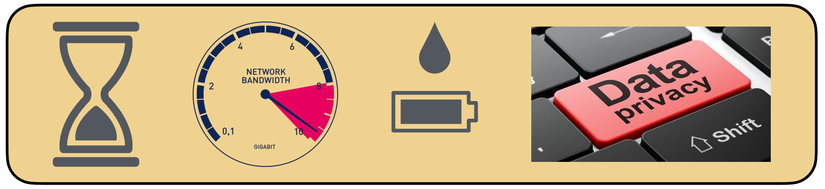
Network Edge

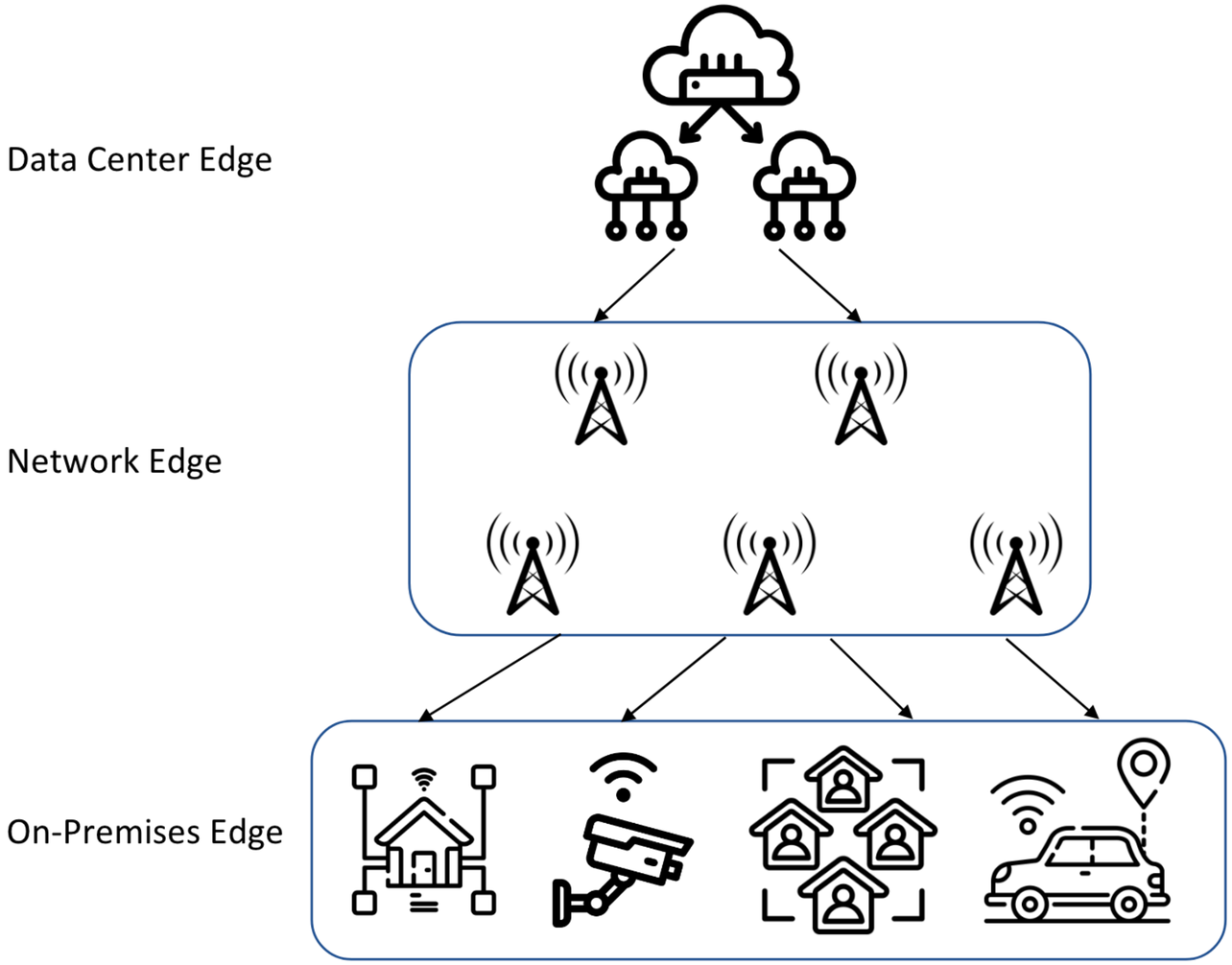
Edge devices

On-Premises Edge



Different Layers of Edge

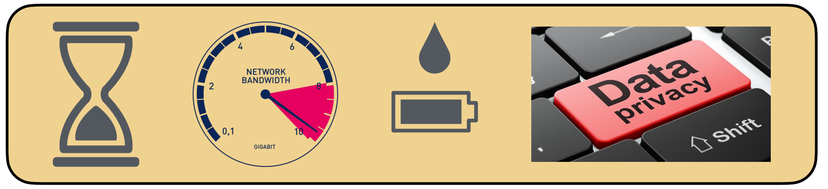


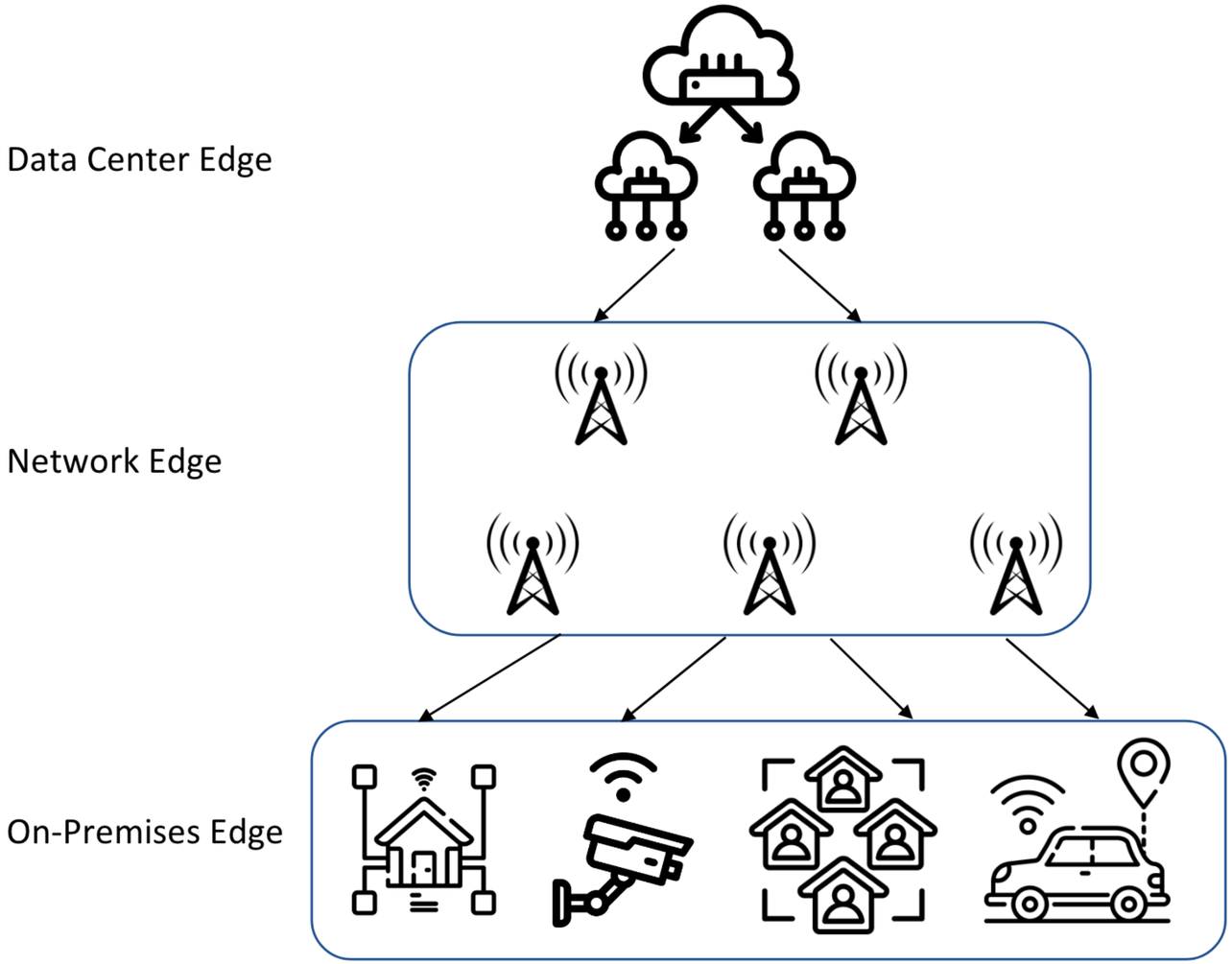


On-Premises Edge (Edge Device)

- Process at the physical location of the user/data source.
- Real-time processing and rapid response (e.g., collision prevention)

Different Layers of Edge

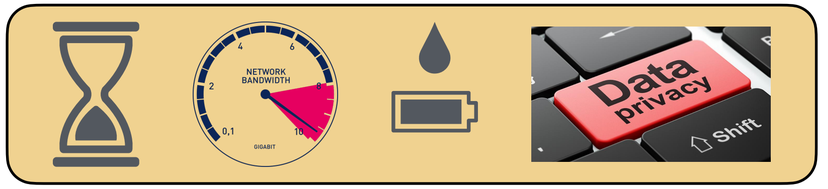




Network Edge (Edge Node)

- Expand the reach of computing by leveraging the infrastructure of telecommunication operators.
- Enable high-quality service delivery (video streaming, online gaming)
- Optimize mobile network and accelerate content delivery.
- Local servers, routers, on-site gateways, base station,...

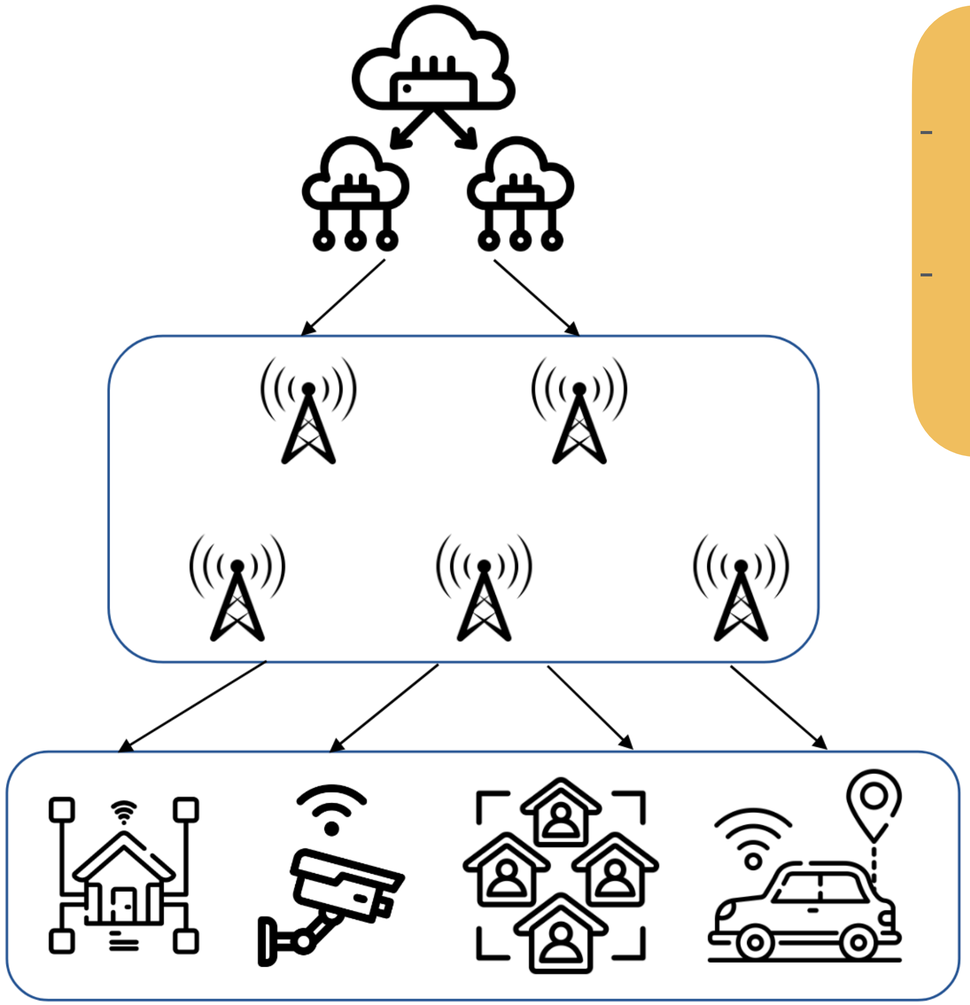
Different Layers of Edge



Data Center Edge

Network Edge

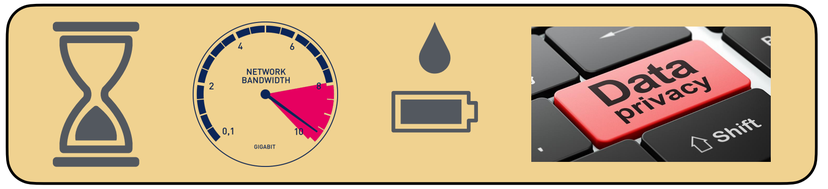
On-Premises Edge

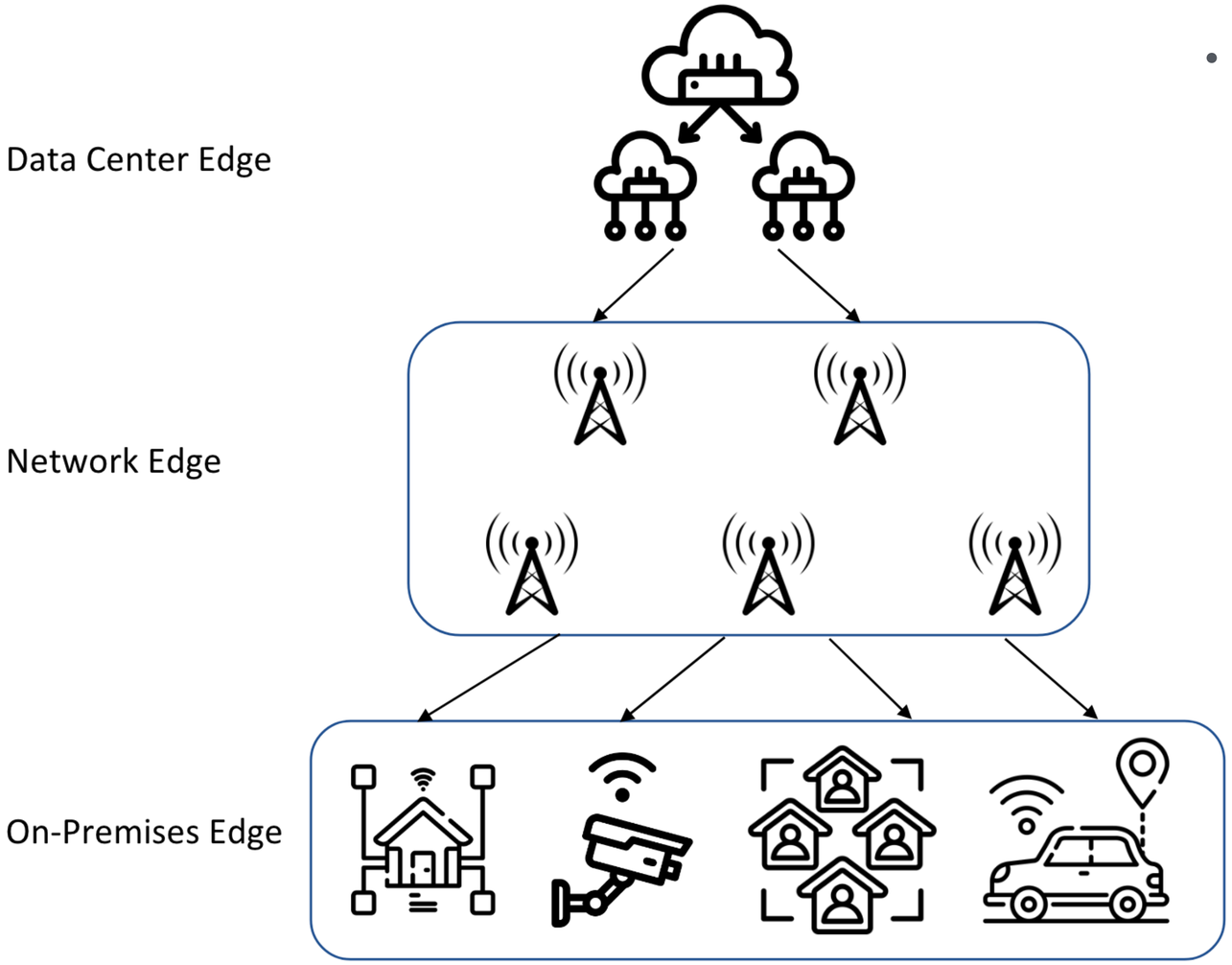


Data Center Edge

- Provide cloud-like computing power but are geographically distributed.
- Regional data centers for content delivery, gaming servers, smart city infrastructure, …

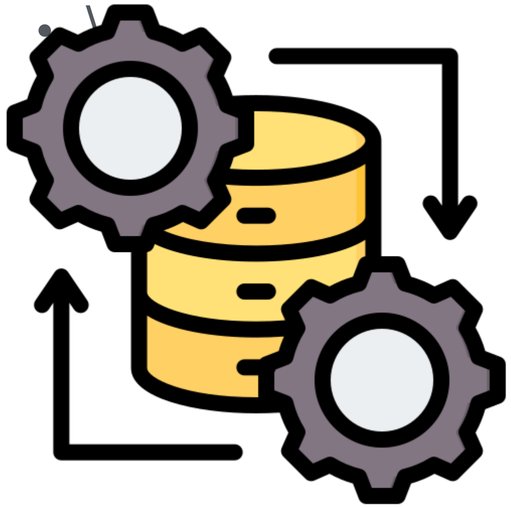
Different Layers of Edge



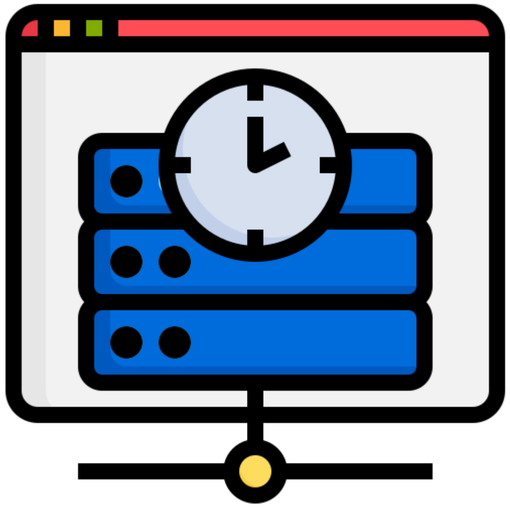


- The divisions between layers are **blurry and flexible**
  - **Overlapping** functions
    - e.g., data processing, storage, communication; different scales
  - **Dynamic** resource allocation
    - Based on current needs
  - **Distributed** nature
    - Functions can occur at multiple points along the continuum
  - **Context-dependent** design

Core Capabilities of Edge Infrastructure



Data processing

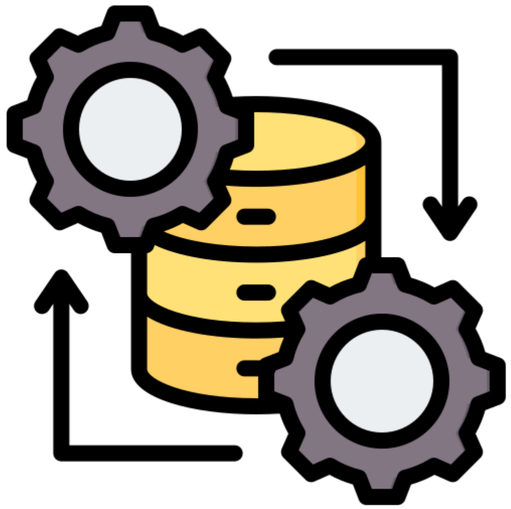


Cache and storage

Communication

Content delivery

Core Capabilities of Edge Infrastructure

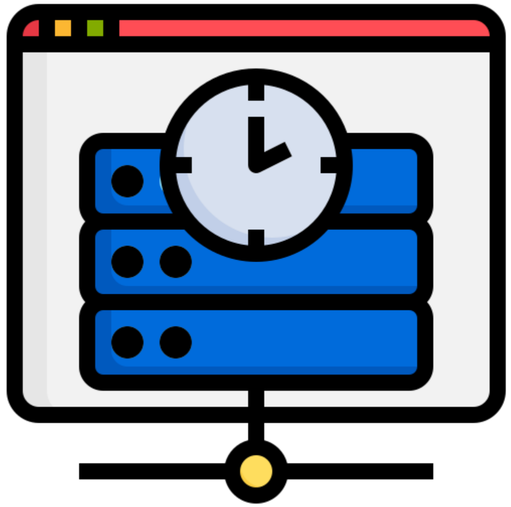


Data processing

1. **Collection** — Gather data from various sources
2. **Filtering** — Select relevant data points based on predefined criteria
3. **Cleansing** — Remove corrupt or irrelevant data to maintain data integrity
4. **Transformation** — Adjust data structure/format for downstream applications
5. **Aggregation** — Summarize data into a more manageable and insightful form for analysis

Core Capabilities of Edge Infrastructure

"Cache and storage is the foundational capability that supports the data lifecycle within edge infrastructure."



Cache and storage

- **Temporary storage (cache)**
  - Quick access and retrieval of frequently used data
- **Permanent storage**
  - Historical data is vital for trend analysis, long-term planning, or compliance with data retention policies
- **Design considerations**
  - Data durability, accessibility, security
  - Optimize distributed file systems, object storage, and databases for efficient data storage, management, and retrieval

Zhao, Y., Zhang, W., Zhou, L., & Cao, W. (2021). A survey on caching in mobile edge computing. Wireless Communications and Mobile Computing, 2021(1), 5565648.

Core Capabilities of Edge Infrastructure

Communication

- **Low-latency** communication by
  - Reduced travel distance
  - Technology support like 5G, Wi-Fi, and fiber networks
- **Improved performance** by
  - Enable local data processing
  - Improve bandwidth efficiency
  - Increase network scalability

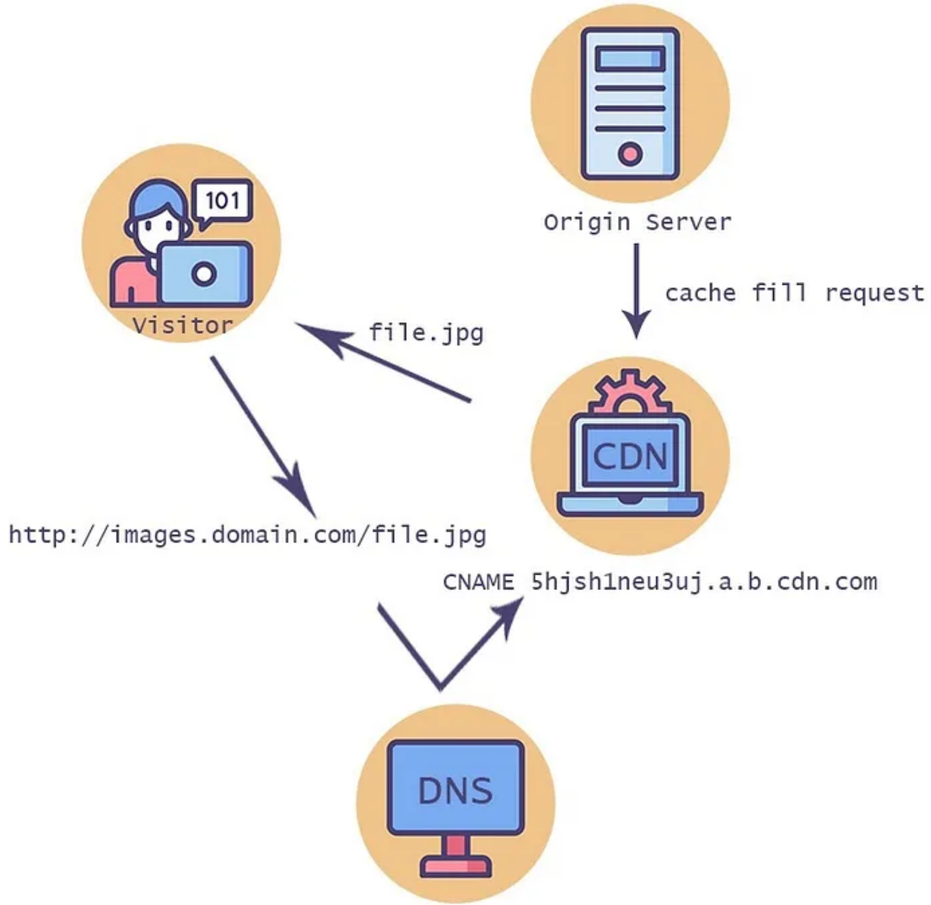
Core Capabilities of Edge Infrastructure

CDN represents a specialized application of edge infrastructure capabilities.

Content delivery

- Optimize the content delivery by replicating across multiple edge nodes.
  - Reduce latency and ensure high availability, even during peak traffic periods.
- **More than content caching**
  - Content routing, load balancing, dynamic content optimization, ...

Content Delivery Network (CDN)



How the Cloud and CDN Architecture Works for Netflix

NETFLIX

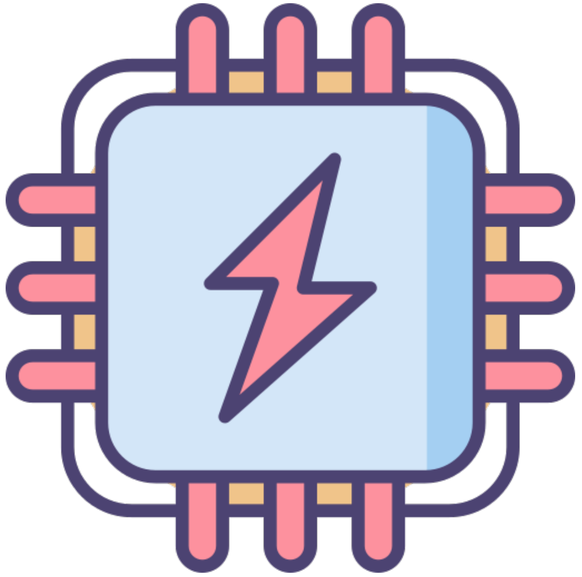
YouTube

GitHub

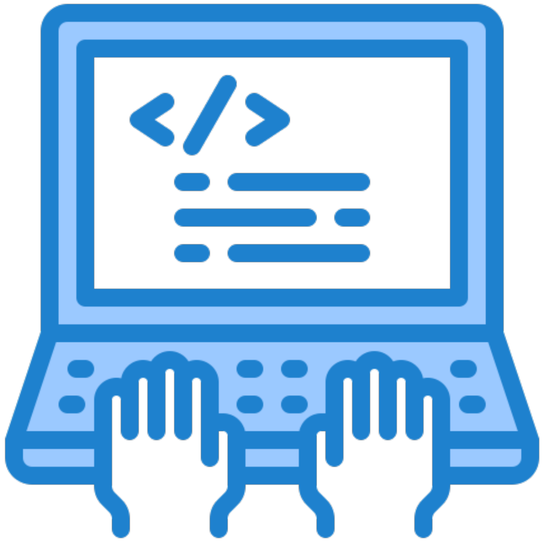
twitch

Spotify®

What are the challenges in deploying Edge AI on Edge Infrastructure?



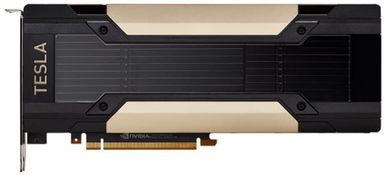
Computing Ability



Programmability

Cloud AI Hardware

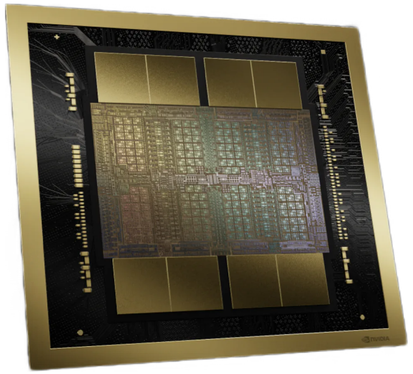
P100 (2016)



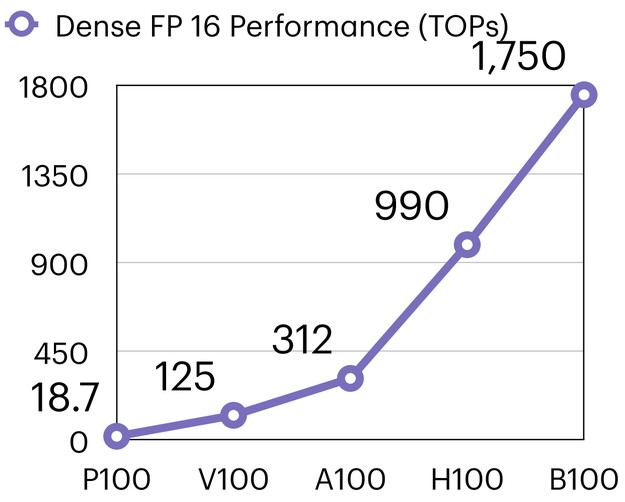
V100 (2017)

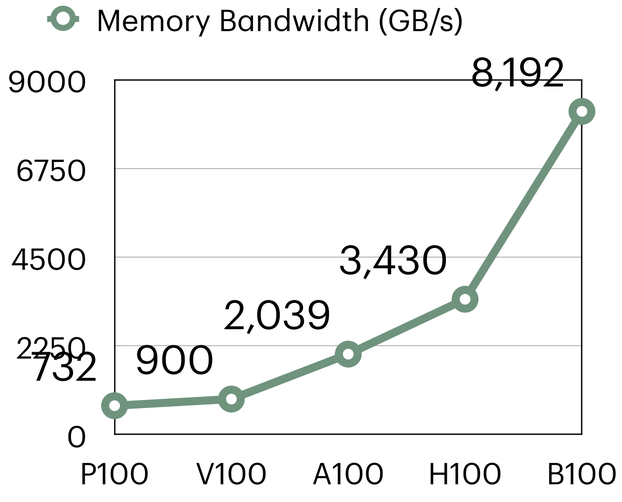
A100 (2020)

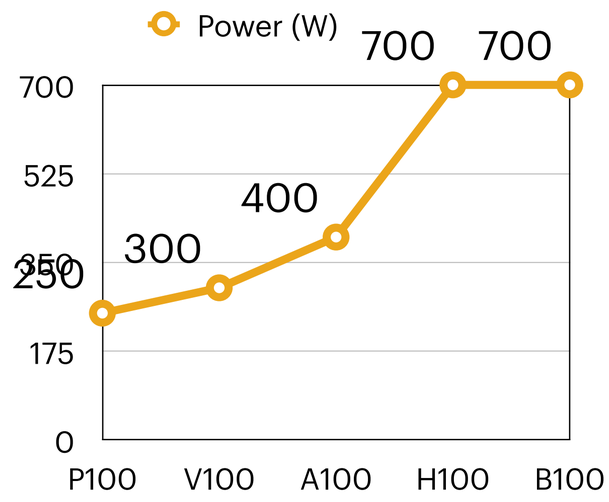
H100 (2022)

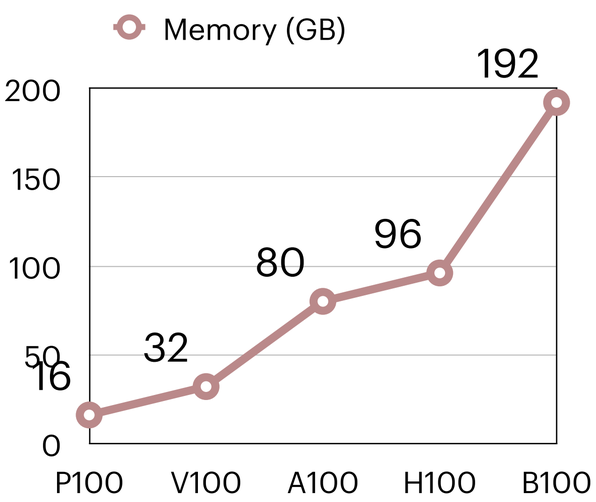


B100 (2024)





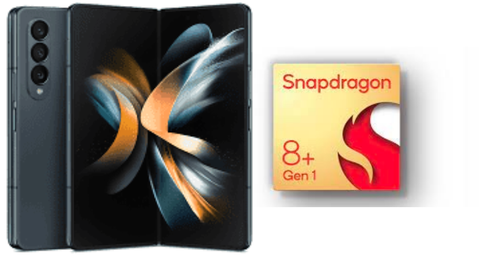


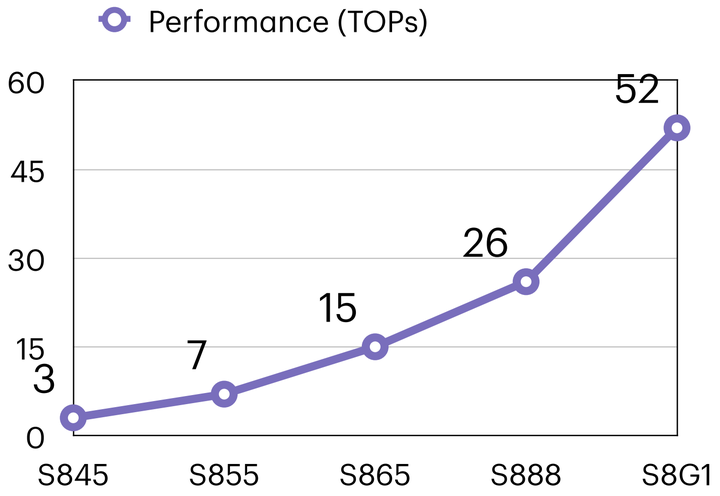


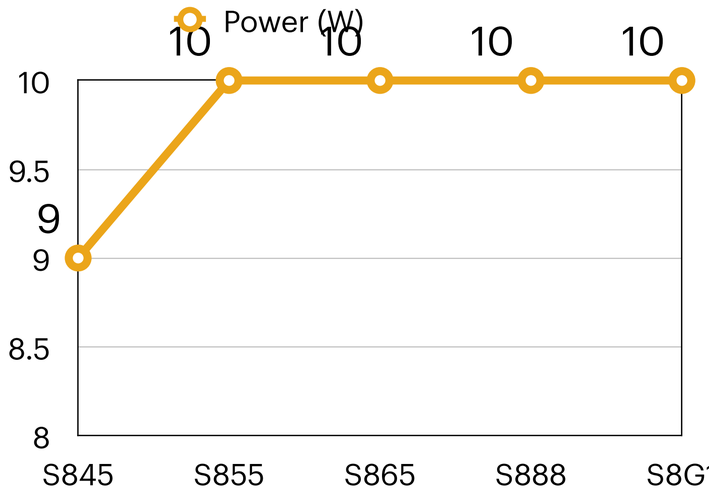
Edge AI Hardware

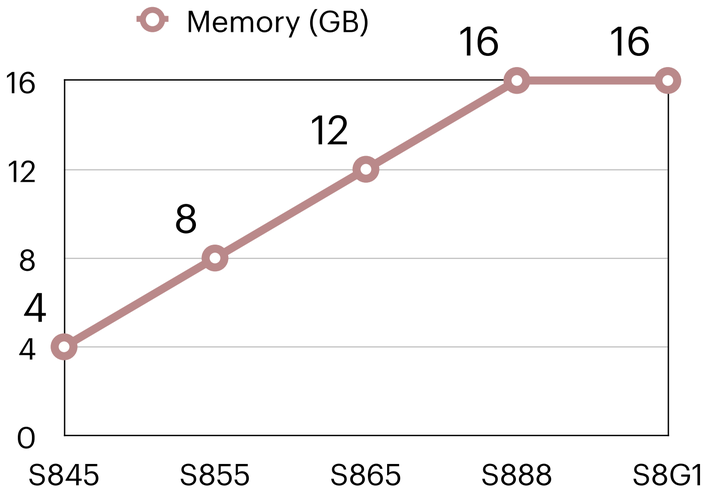
Qualcomm Hexagon DSP

- Qualcomm Hexagon is a family of digital signal processor (DSP) products by Qualcomm. It is designed to deliver performance with low power over a variety of applications.





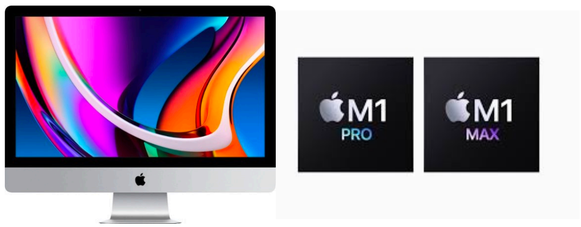


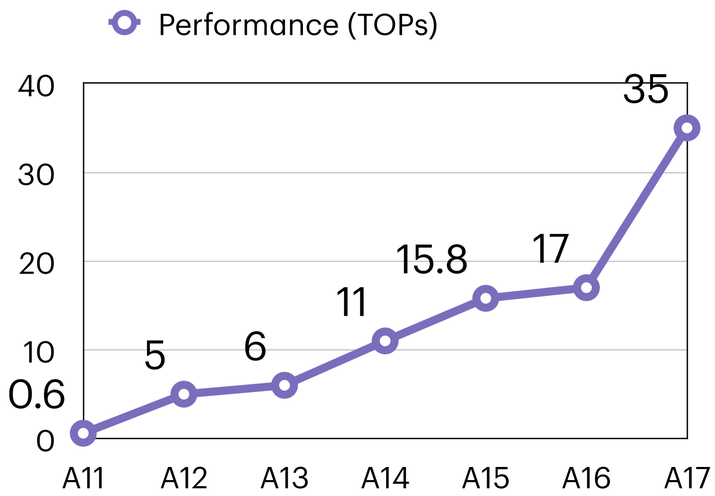


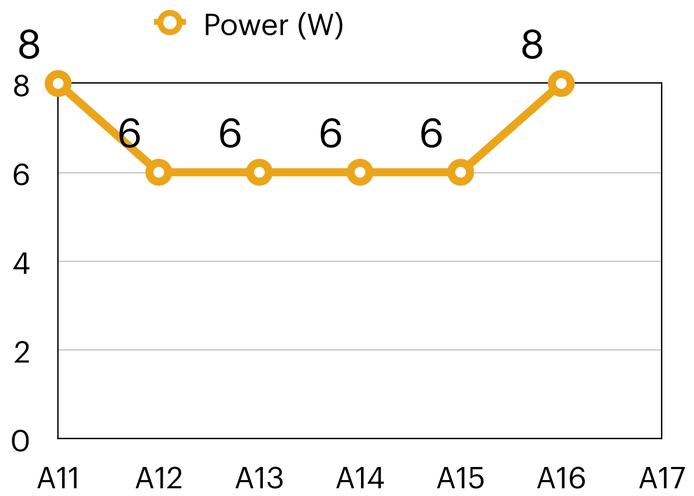
Edge AI Hardware

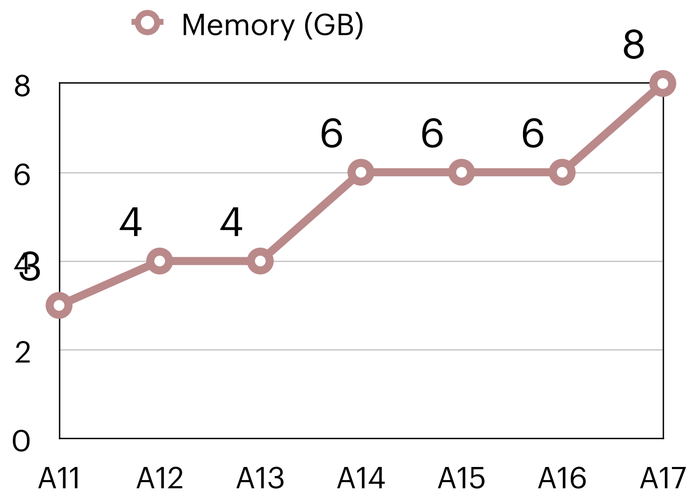
Apple Neural Engine

- The Apple Neural Engine (ANE) is an energy-efficient and high-throughput engine for ML inference on Apple silicon.







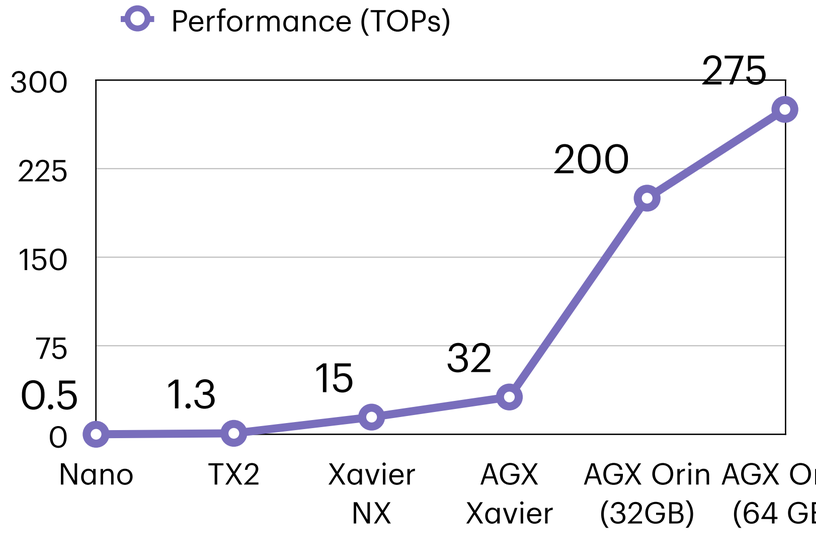


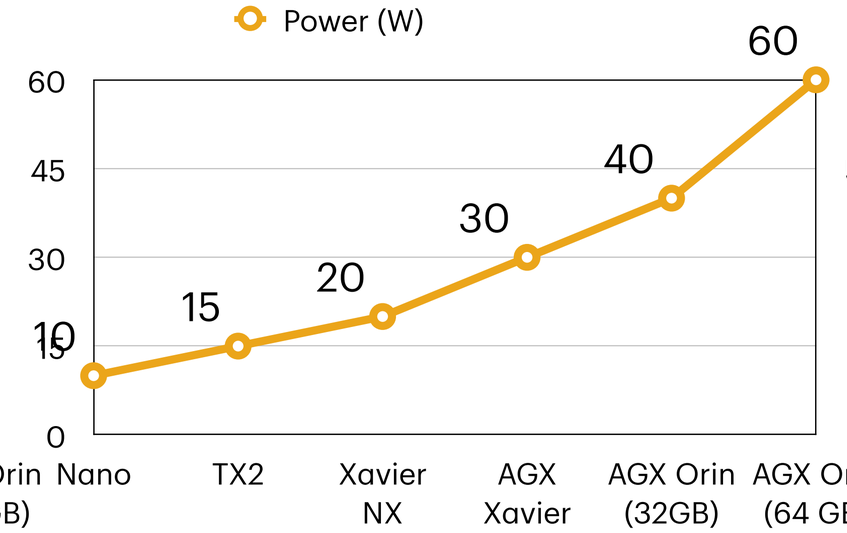
Edge AI Hardware

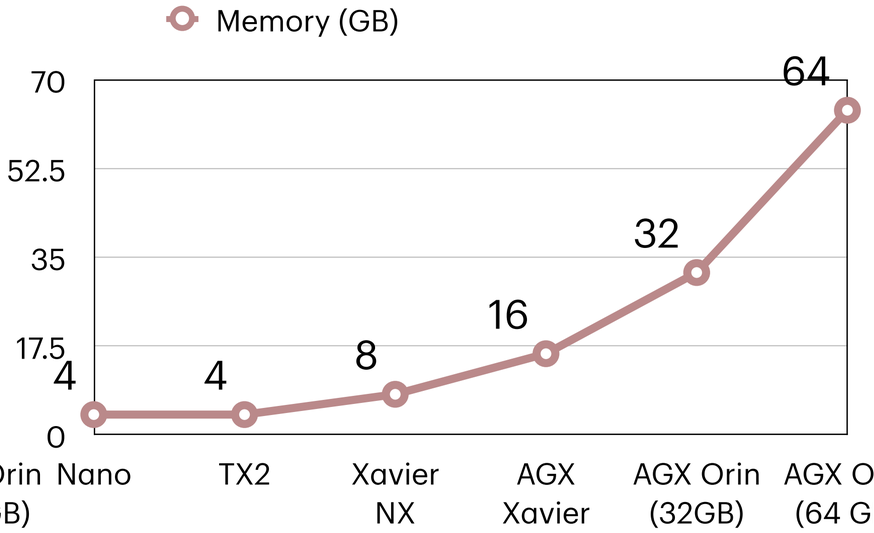
Nvidia Jetson

- NVIDIA Jetson is a complete System on Module (SOM) that includes a GPU, CPU, memory, power management, high-speed interfaces, and more.



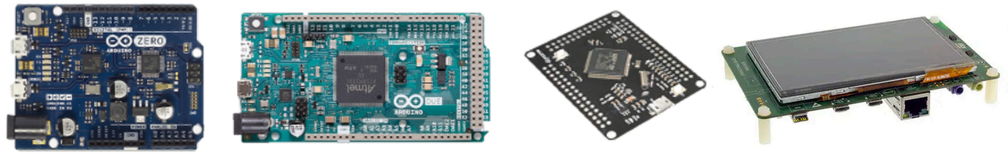


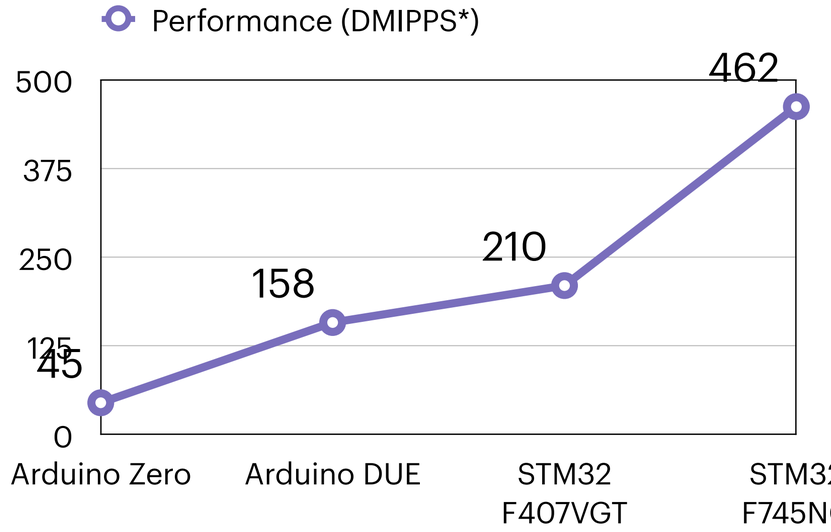


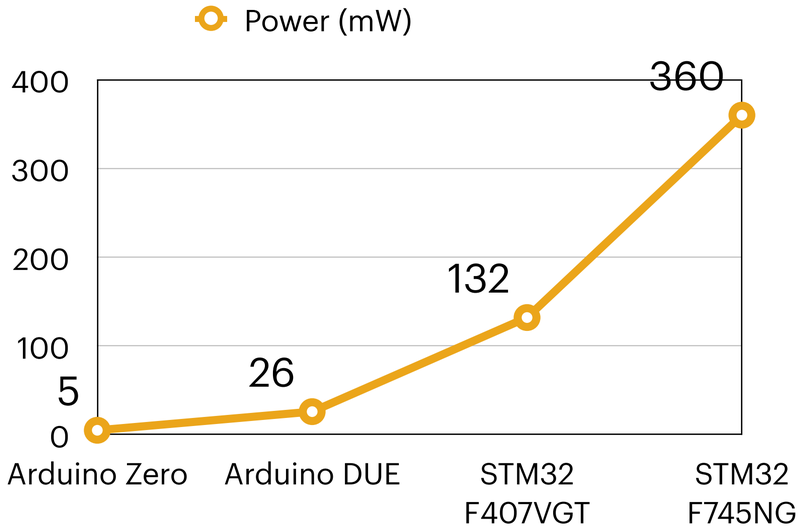
Edge AI Hardware

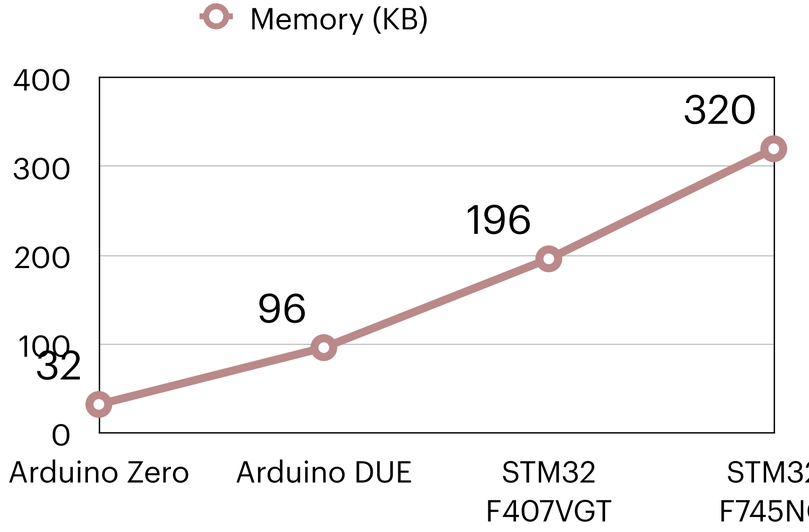
Microcontrollers

- A microcontroller is a compact integrated circuit designed for embedded systems. A typical microcontroller includes a processor, memory and input/output (I/O) peripherals on a single chip.





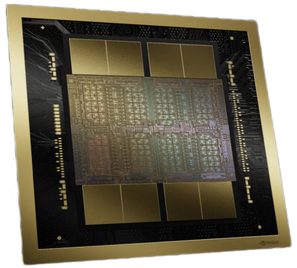




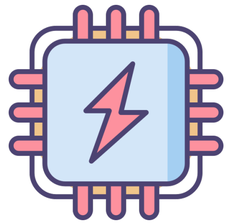
* Dhrystone Million Instructions Per Second (DMIPS) is an index for integer computation

Edge AI Hardware

Edge AI devices still have hug gap to cloud processors



	Cloud AI	Mobile AI	Tiny AI
Memory (Activation)	80GB	4GB	320KB
Storage (Weights)	~TB/PB	256GB	1MB



Computing Ability

Programmability

Edge Computing Models

Mobile/Multi-access Edge Computing

Mobile -> Multi-access Edge Computing

- Standardized by the European Telecommunications Standards Institute (ETSI)
- Integrates with the mobile network infrastructure, leveraging the existing cellular network
  - Ideal for telecom operators to offer value-added services and optimize network performance
  - Enable low-latency, high-bandwidth applications, especially with the rise of 5G

Hu, Y. C., Patel, M., Sabella, D., Sprecher, N., & Young, V. (2015). Mobile edge computing—A key technology towards 5G. ETSI white paper, 11(11), 1-16.

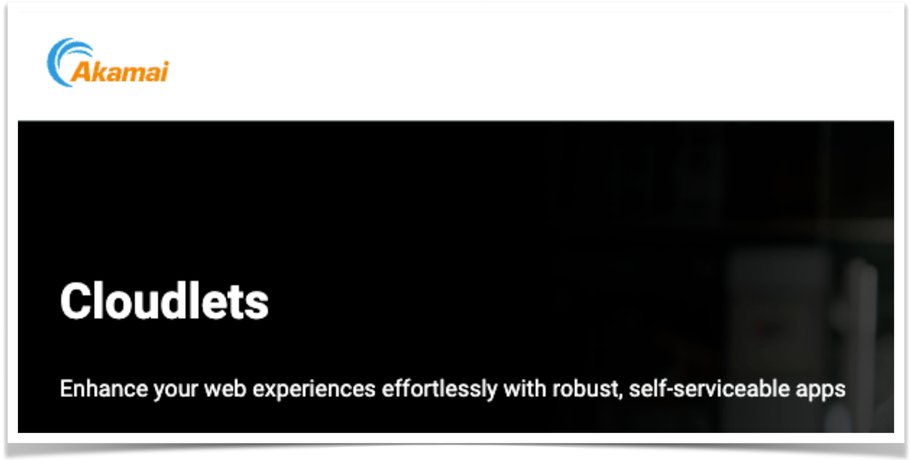
Mobile/Multi-access Edge Computing

Mobile -> Multi-access Edge Computing

Verison 5G Edge with Public MEC

Cloudlet Computing

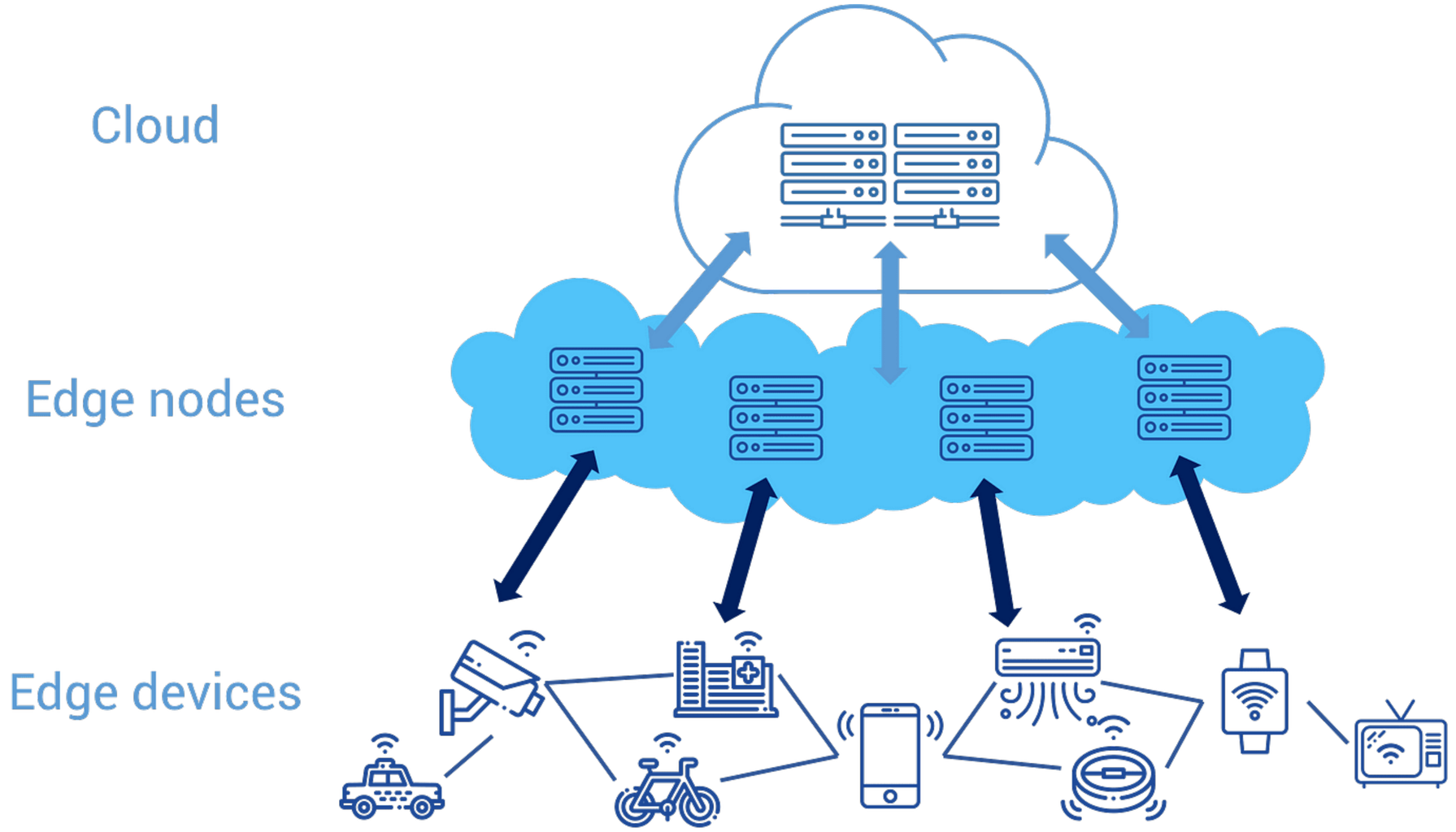
- Providing localized, powerful computing resources near mobile devices
- Serve as small-scale data centers
  - Offloading computation-intensive tasks from mobile devices to enhance performance and extend battery life



Akamai Cloudlet

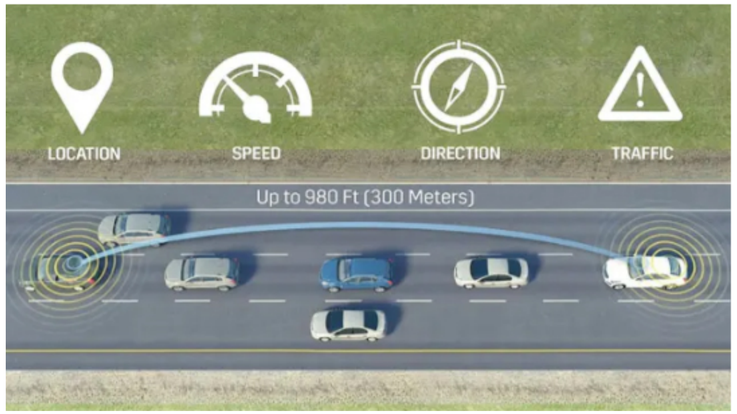
Satyanarayanan, M. (2013, June). Cloudlets: at the leading edge of cloud-mobile convergence. In Proceedings of the 9th international ACM Sigsoft conference on Quality of software architectures (pp. 1-2).

Collaborative Models



Edge-Edge Collaboration

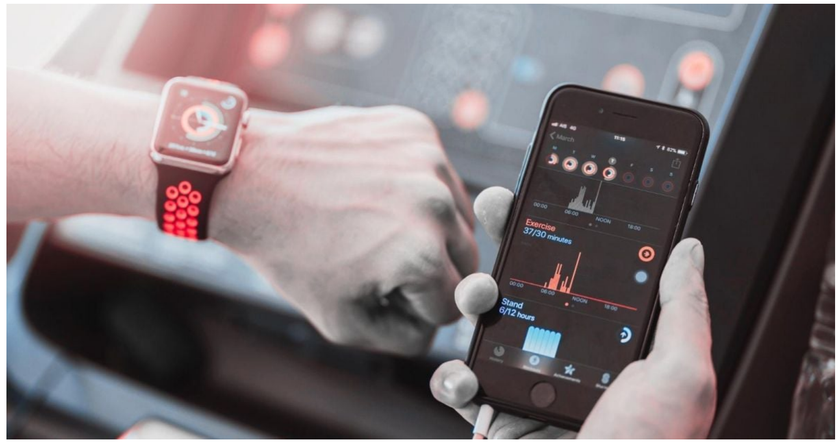
- **Direct** interaction and coordination to enhance performance, reliability, and scalability.
  - Edge nodes communicate with each other to share data, balance workloads, and provide redundancy.
  - Efficient resource allocation and management enabled by **decentralized** protocols and frameworks.



V2V

Edge-Device Collaboration

- The pivotal way to improve device performance and **extend battery life**
  - Computation-intensive tasks are offloaded to nearby edge nodes.



Edge-Cloud Collaboration

- A balanced approach to data processing and storage.
  - Cloud resources can handle peak loads and extend data storage
  - Edge performs critical, time-sensitive computations

Amazon Go Store (A device-cloud collaboration)

- Edge devices (cameras, sensors) track items and make decisions (charging the customer's account).
- The cloud handles more complex/centralized tasks, like inventory management, data analysis, and user behavior tracking.

Cloud-Edge-Device Collaboration

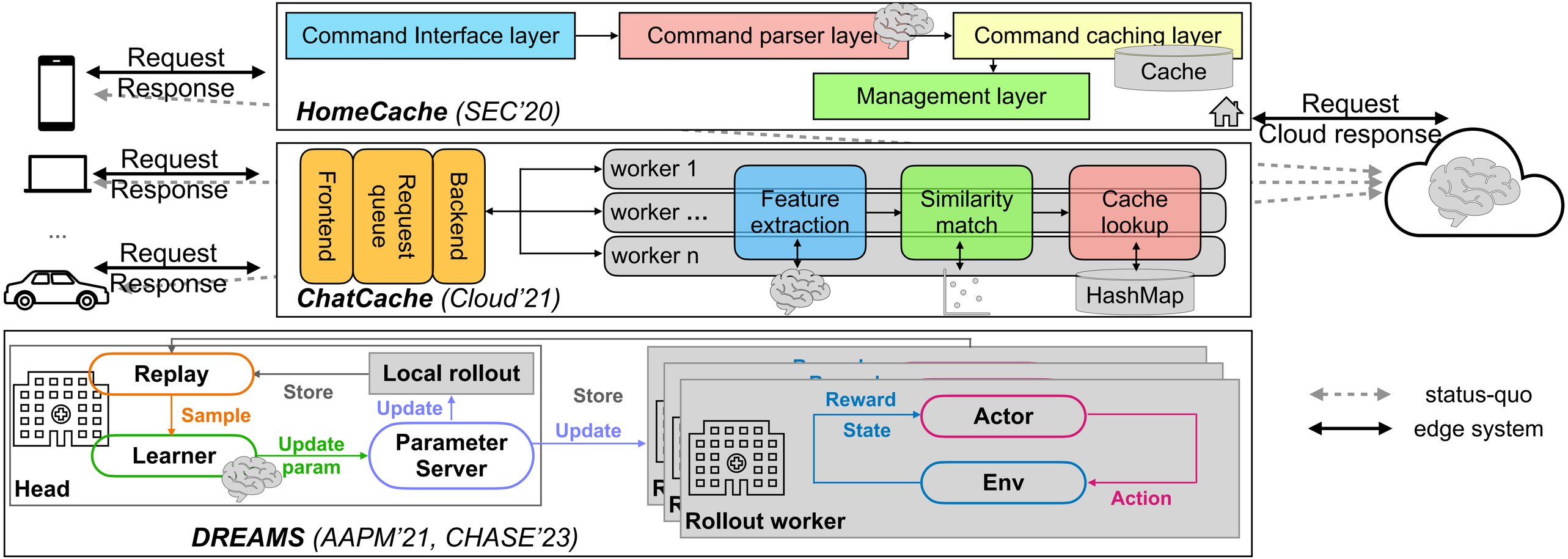
- A holistic approach where tasks are distributed across the cloud, edge nodes, and edge devices based on their computational requirements and latency sensitivity.
  - Cloud handles large-scale data processing and long-term storage
  - Edge nodes perform intermediate processing and provide low-latency response
  - Edge devices focus on data collection and immediate user interactions

Choose the Right Model.

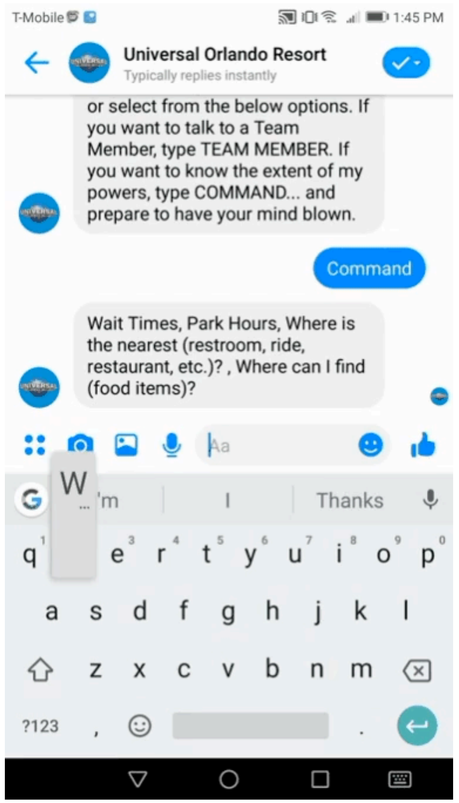
Depends on the needs.

Too high-level? 🤷‍♀️
I know.

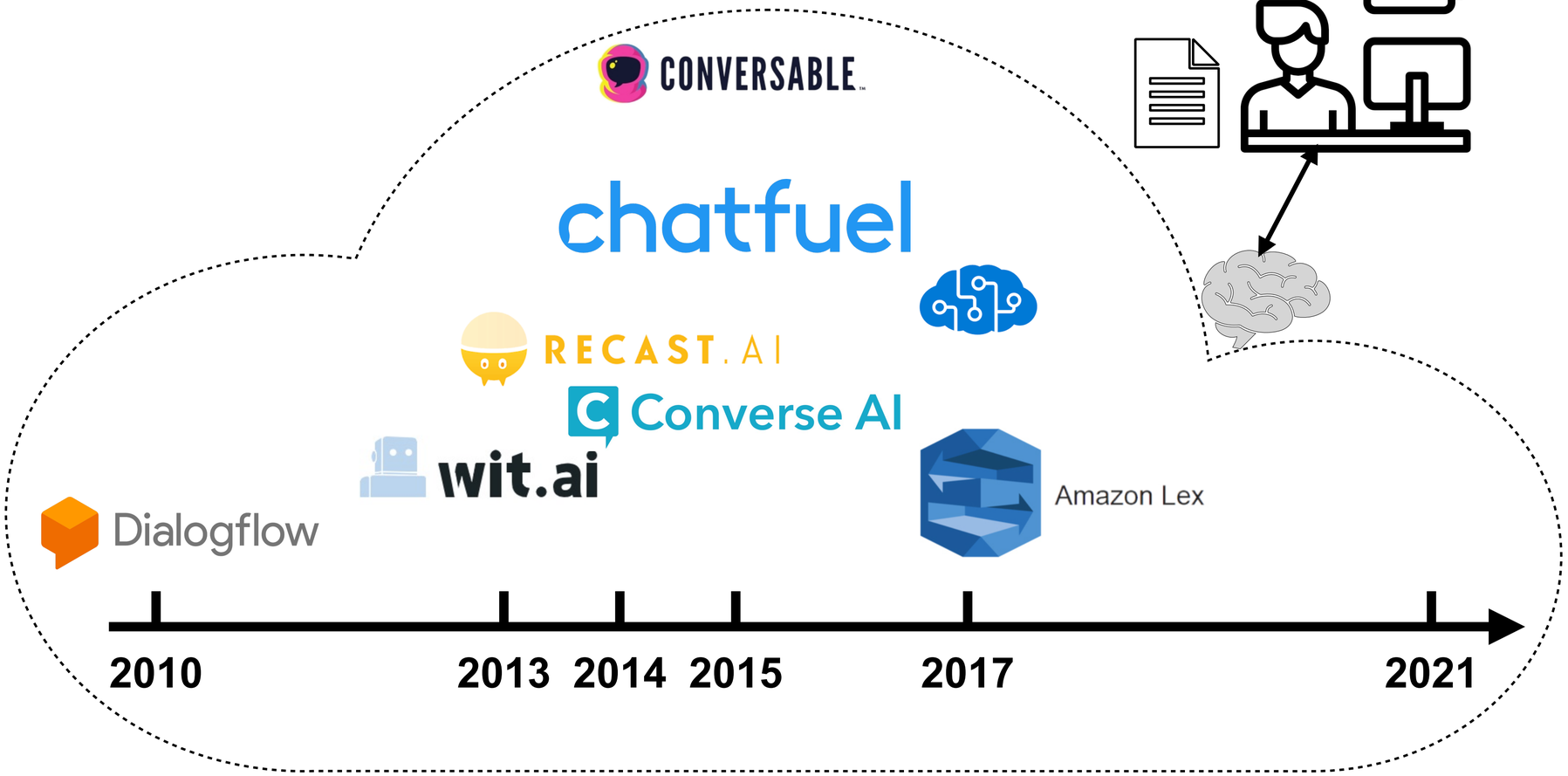
Edge Systems Support for Intelligent Service: Cloud-Edge Collaboration



Background



Picture credit: https://landbot.io/blog/conversational-interfaces-explained/



- BERT was the best NLP model (2018)

Smart Home Assistant

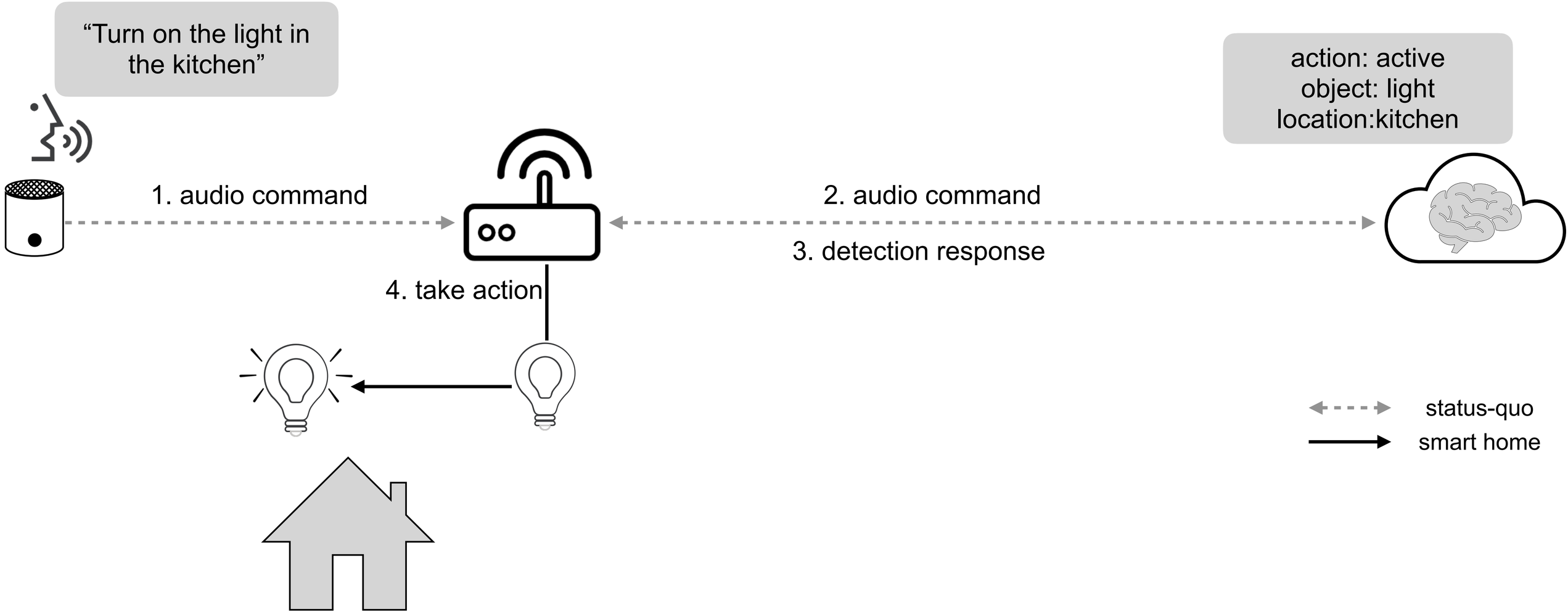
- A typical **spatial-temporal locality**

Cache!

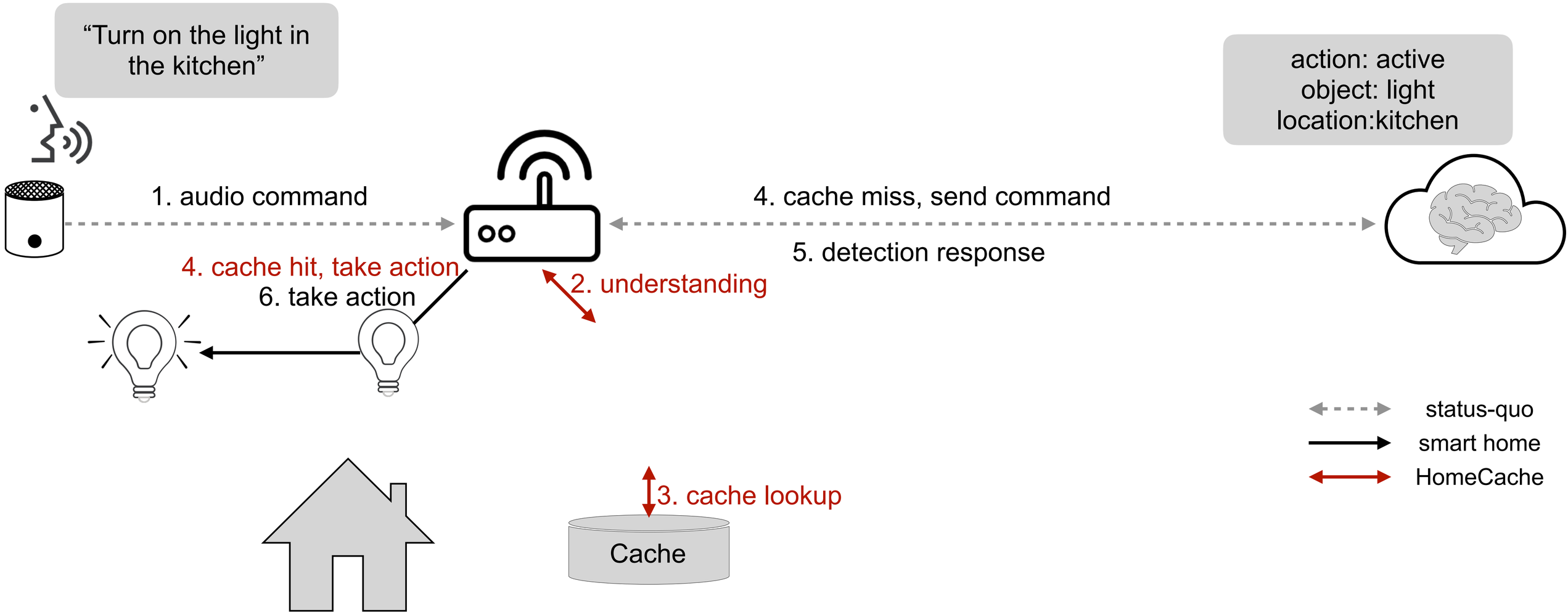
- A privacy-prioritized place
- Status-quo was cloud-based
- Edge can bring lower latency, more stable performance

From Cloud

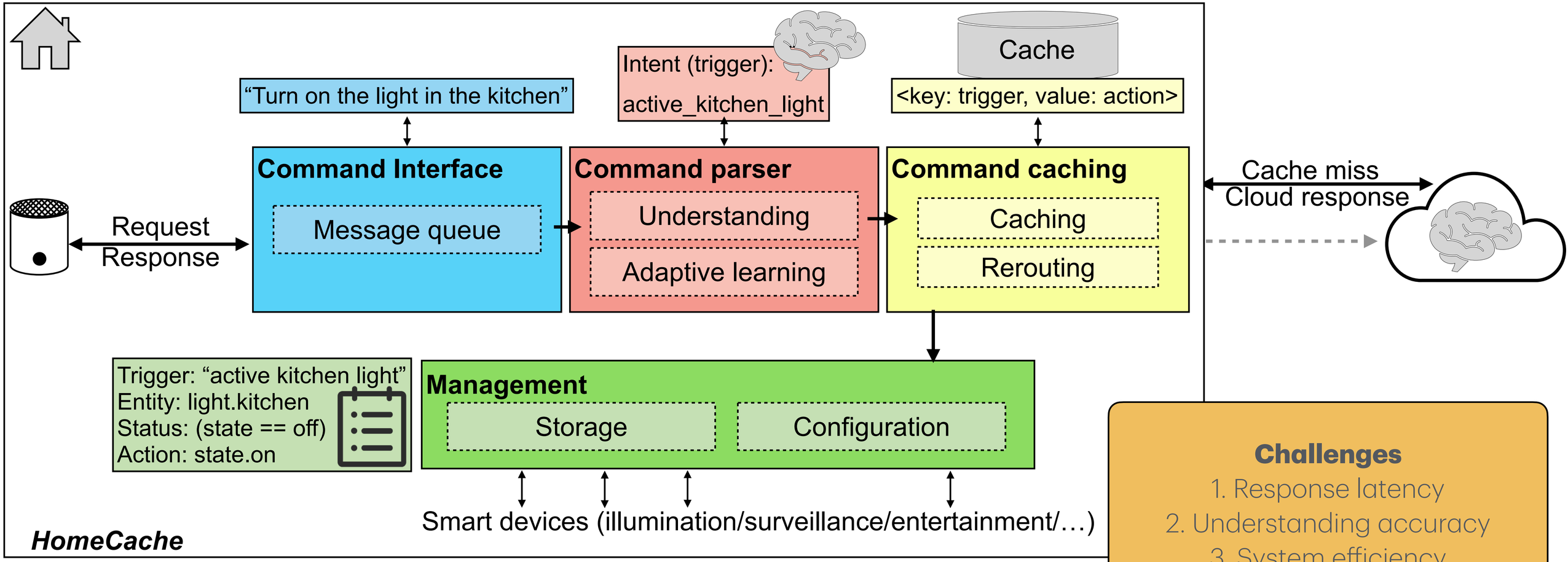
	Turn	on	the	light	in	the	kitchen
Slot	B-activate	I-activate	O	B-object	O	O	B-location
Intent	Active_kitchen_light						



To Cloud + Edge



A caching framework for home-based voice assistant systems



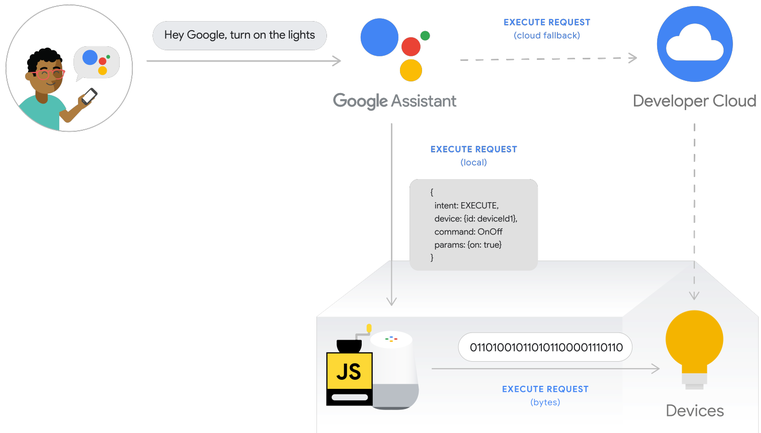
Challenges

1. Response latency
2. Understanding accuracy
3. System efficiency

Xu, L., Iyengar, A., & Shi, W. (2020, November). CHA: A caching framework for home-based voice assistant systems. In 2020 IEEE/ACM Symposium on Edge Computing (SEC) (pp. 293-306). IEEE.

Future of Smart Home? Maybe!

- Google Assistant Local Home SDK
  - Home gets SYNC response from cloud => local triggered
  - Cloud uses IDENTIFY intent to label if the device is locally controllable
  - Device ID in IDENTIFY ? local : cloud
  - User triggers an action with a local fulfillment path => EXECUTE or QUERY intent



Google Developer Center - Local Fulfillment

From Exact Duplication To Semantic Duplication

- Precise/fuzzy redundancy: **syntactic same/similar** with **semantic duplication**

- **Semantic redundancy**: **syntactic differences** with **semantic duplication**

What is the waiting time for rides?
Rides' waiting time?
How long I need to wait for rides?

Louder, please.
Turn sound up.

- We found that by extracting **semantic features**, we can find a **threshold** to identify different statements (voice/text) with the same semantic meaning.

Semantic Redundancy Lookup

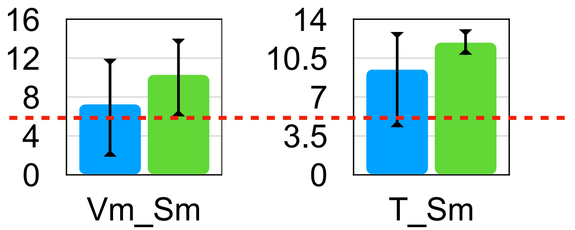
Challenges

1. High-dimensional similarity search
2. Dominant class selection
3. Lightweight model design

Semantic Feature

ANN Search

Threshold Filter

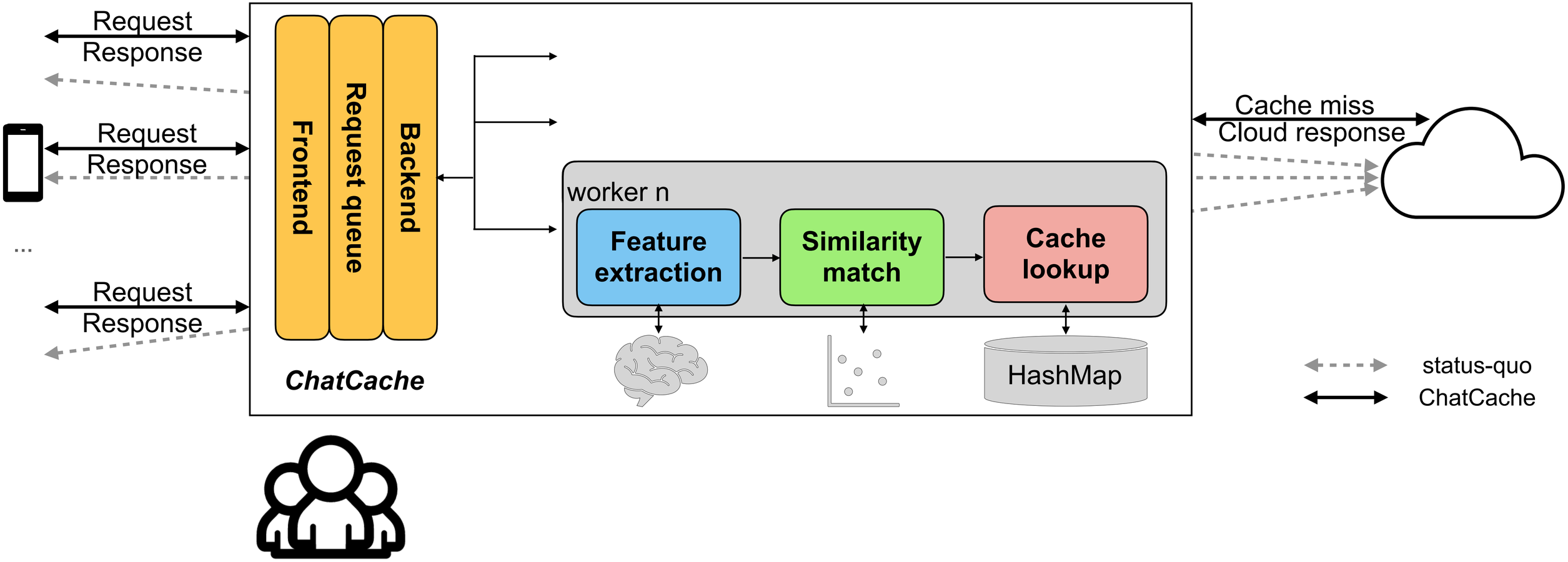


Class Dominate

Cache Item
[vector, ID, class, fulfillment, ...]

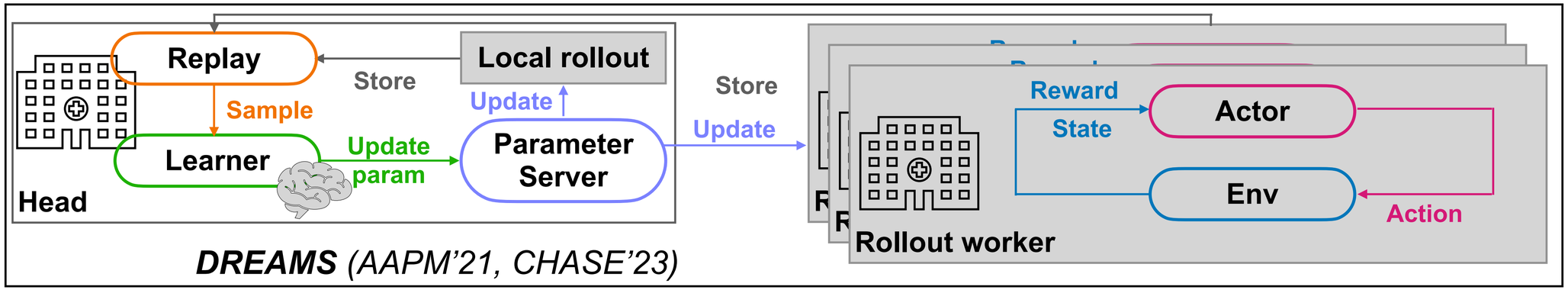
Feature extraction → **Similarity match** → **Cache lookup**

A Hierarchical Semantic Redundancy Cache System for Conversational Services at Edge

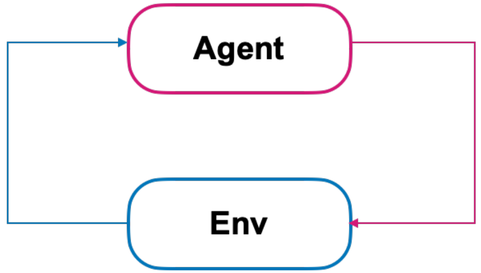


Xu, L., Iyengar, A., & Shi, W. (2021, September). ChatCache: A Hierarchical Semantic Redundancy Cache System for Conversational Services at Edge. In 2021 IEEE 14th International Conference on Cloud Computing (CLOUD) (pp. 85-95). IEEE.

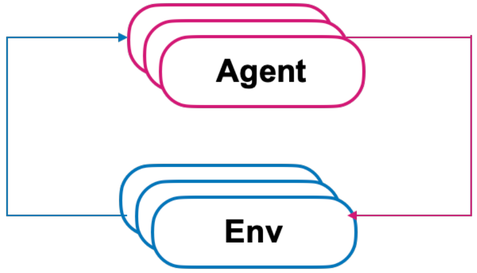
From Cloud+Edge To Edge+Edge



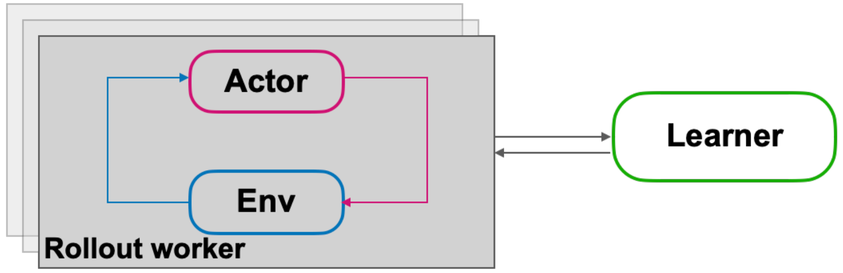
1. Decouple and abstract



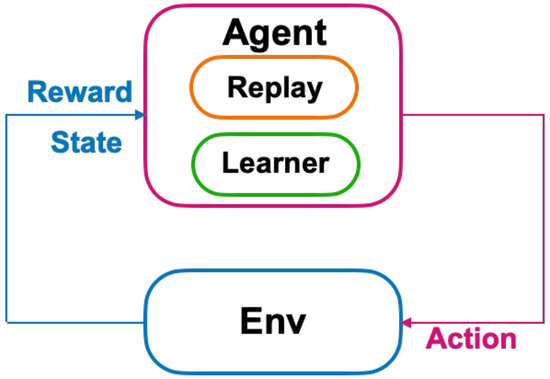
Standard DRL

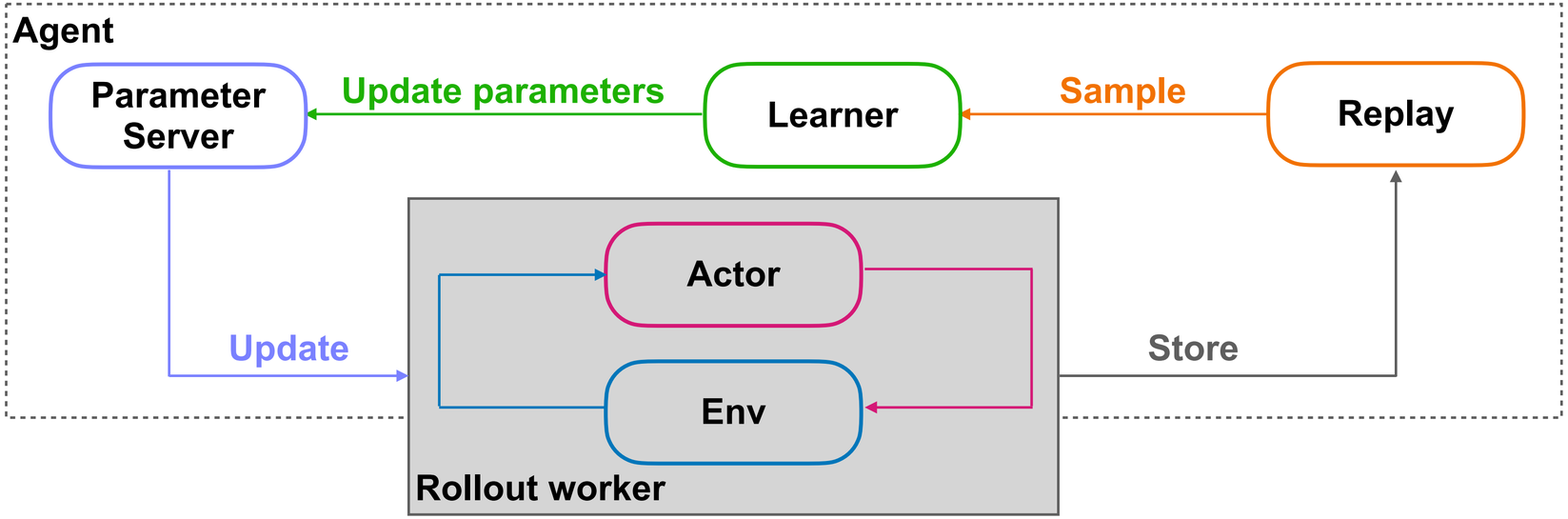


Distributed DRL

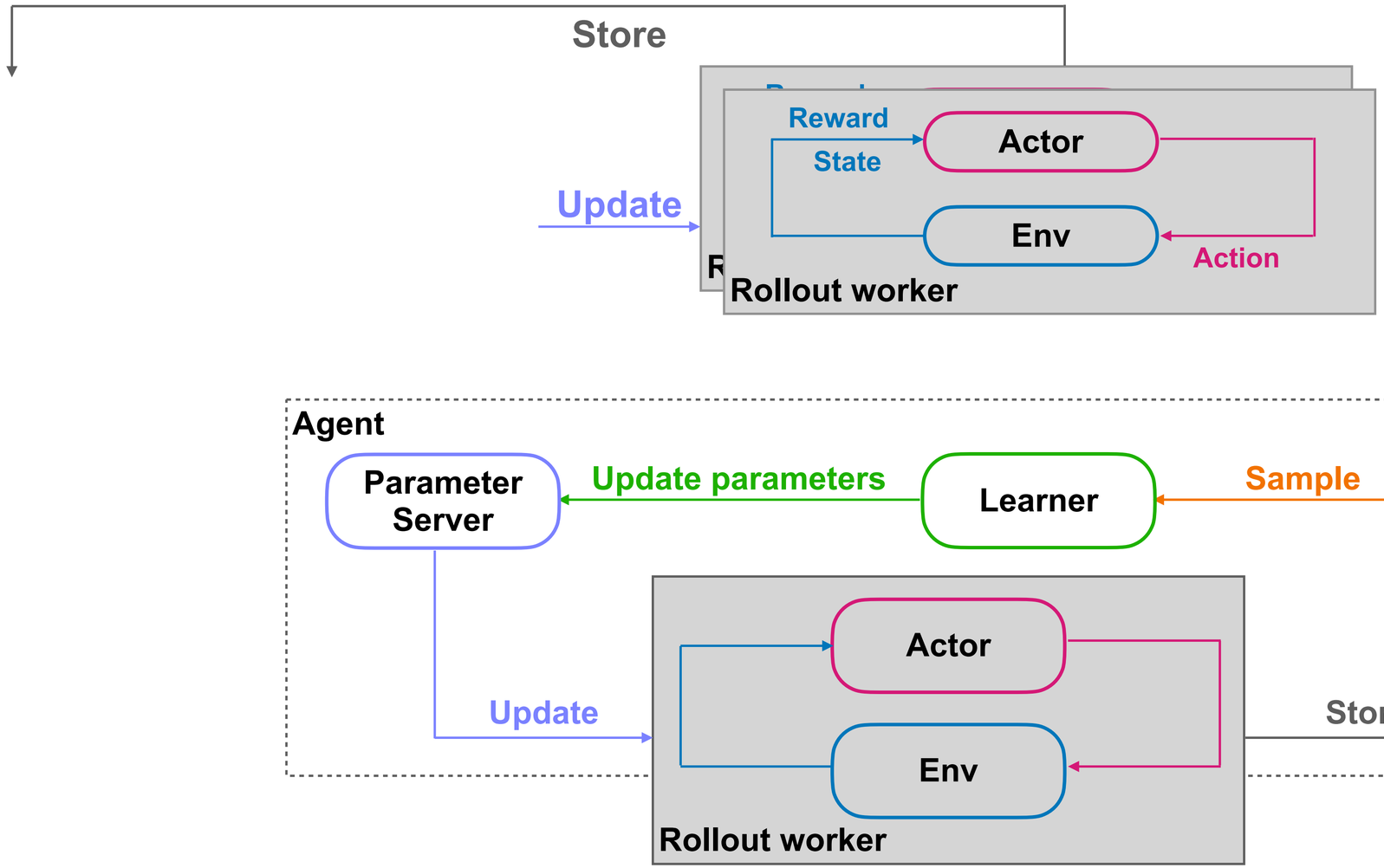


Implementation of distributed DRL

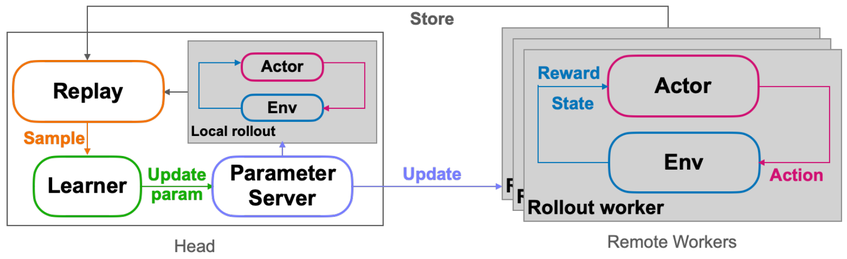


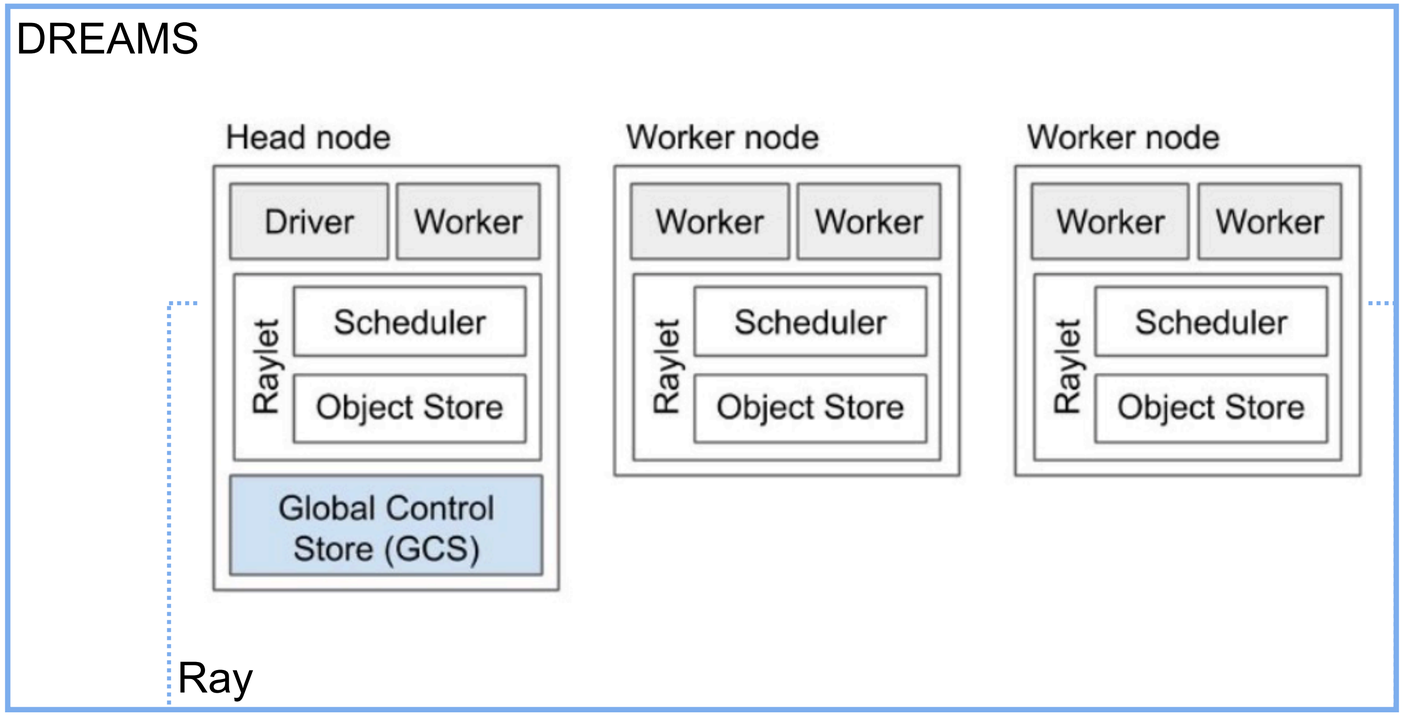


2. Encapsulate and distribute



System Implementation

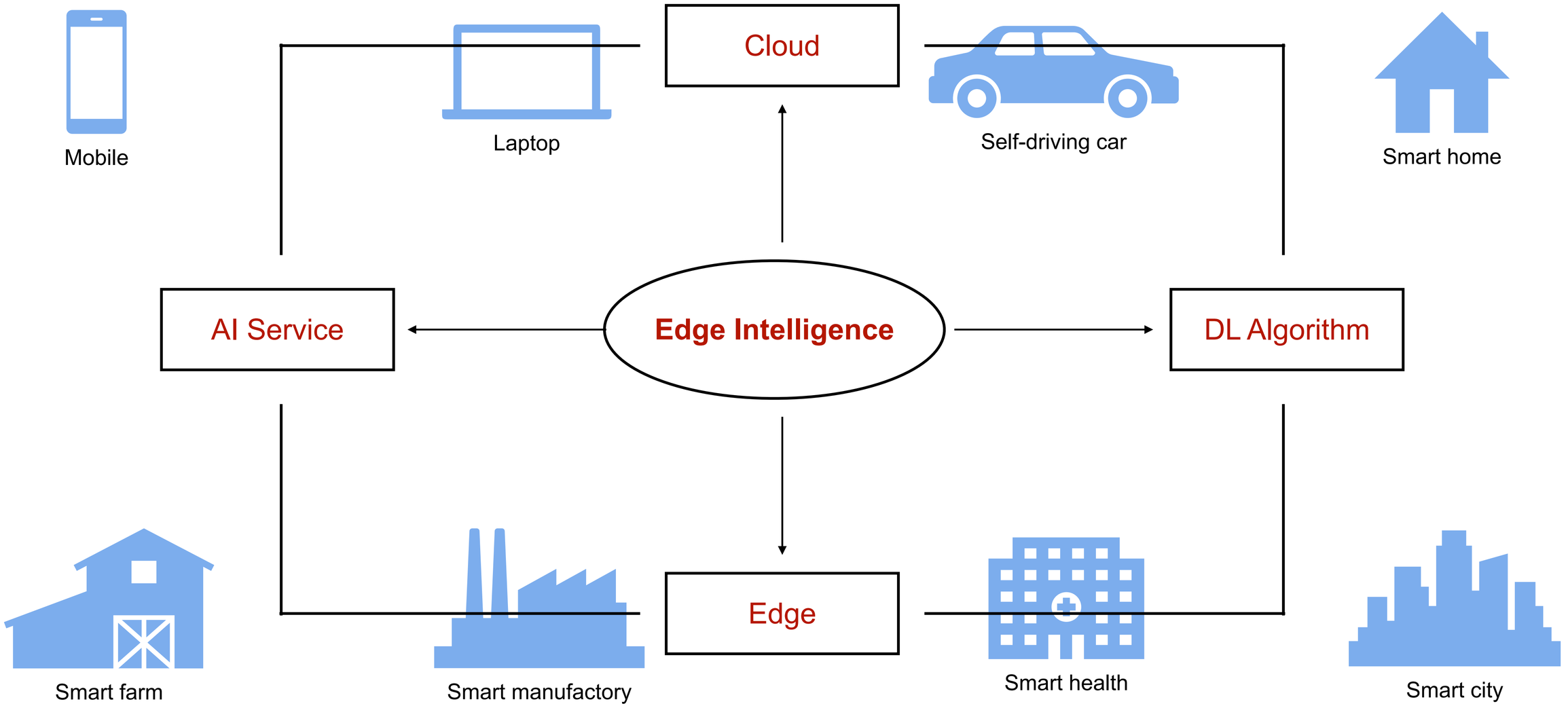




Multi-Insitutional Collaboration

- **DREAMS vs Federated Tumor Segmentation (FeTS)**
  - DREAMS: Self-supervised, privacy-preserved, single learner with multiple workers.
  - FeTS: Supervised, single aggregator with multiple collaborators.
- **Pros**
  - No data exchange/transmission physically.
  - Avoid tedious privacy contract signing.
  - Contribution from different institutions to train a better, more general model.
- **Cons**
  - Waiting time comes from synchronization (buckets effect).
  - Complex system design and schedule.

Perelman SCHOOL OF MEDICINE UNIVERSITY of PENNSYLVANIA


Mobile
Laptop
Cloud
Self-driving car
Smart home
AI Service
Edge Intelligence
DL Algorithm
Edge
Smart farm
Smart manufactory
Smart health
Smart city

Lab 1

Selections will open during the next lecture (09/17).

First-come, first-served.

References

- Zhao, Y., Zhang, W., Zhou, L., & Cao, W. (2021). A survey on caching in mobile edge computing. Wireless Communications and Mobile Computing, 2021(1), 5565648.

- Hu, Y. C., Patel, M., Sabella, D., Sprecher, N., & Young, V. (2015). Mobile edge computing—A key technology towards 5G. ETSI white paper, 11(11), 1-16.

- Satyanarayanan, M. (2013, June). Cloudlets: at the leading edge of cloud-mobile convergence. In Proceedings of the 9th international ACM Sigsoft conference on Quality of software architectures (pp. 1-2).

- Xu, L., Iyengar, A., & Shi, W. (2020, November). CHA: A caching framework for home-based voice assistant systems. In 2020 IEEE/ACM Symposium on Edge Computing (SEC) (pp. 293-306). IEEE.

- Xu, L., Iyengar, A., & Shi, W. (2021, September). ChatCache: A Hierarchical Semantic Redundancy Cache System for Conversational Services at Edge. In 2021 IEEE 14th International Conference on Cloud Computing (CLOUD) (pp. 85-95). IEEE.

- https://hanlab.mit.edu/courses/2024-fall-65940