Edge Computing

Lecture 03: Edge Systems: Architecture

Recap

- Evolution of computing paradigm
  - Dist. vs Cent.
  - Cloud View vs Edge View
- Virtualization
  - Virtual machine & containers
- Applications
- From design to deployment

Agenda

- The IoT Challenge
- Bandwidth, latency, throughput, pipeline
- Example system architectures
- Close-the-loop: sensing, compute, actuation

The IoT Challenge

- IoT devices have been viewed as ***simple nodes*** that ***collect data*** from sensors and then transmit it to a ***central location for processing***.
- Cloud computing -> Edge computing (BLERP)
  - Bandwidth, latency, economics, reliability, privacy
- But how much compute?

How much compute?

Autonomous Vehicle

Perception → Localization → Prediction → Reasoning → Planning

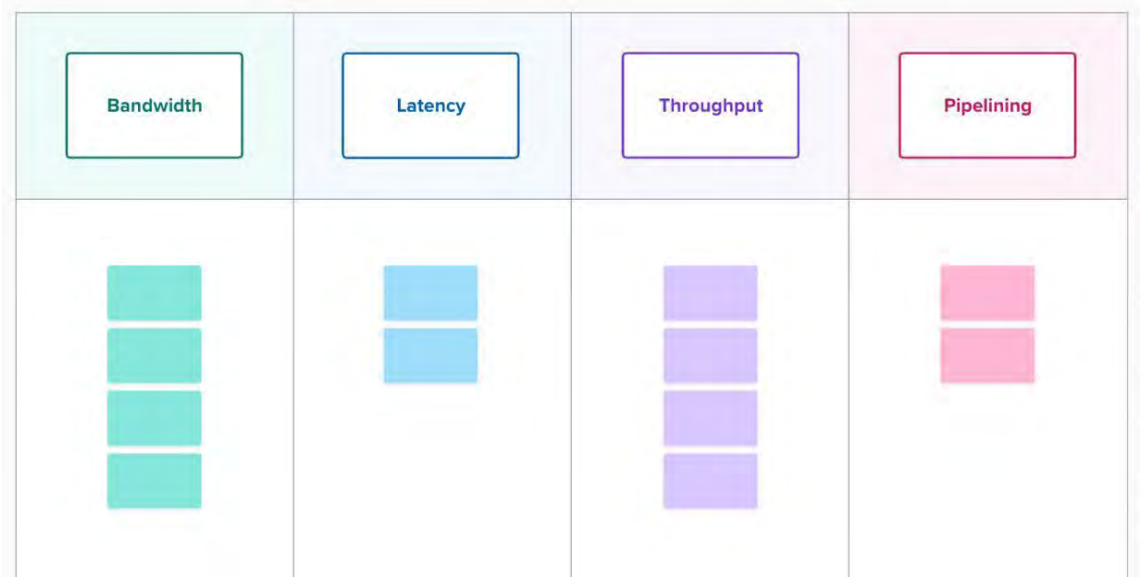
FPS! (Frame per Second)

- Let’s say the car is driving on a freeway at 65 mph (105 km/h).
- That means it’s going at 95 feet per second (~29 meters).
- Average stopping time at that speed is 120 feet.
- If the decision your car has to make is an immediate stop from the moment an object is detected, then the stopping distance will be:

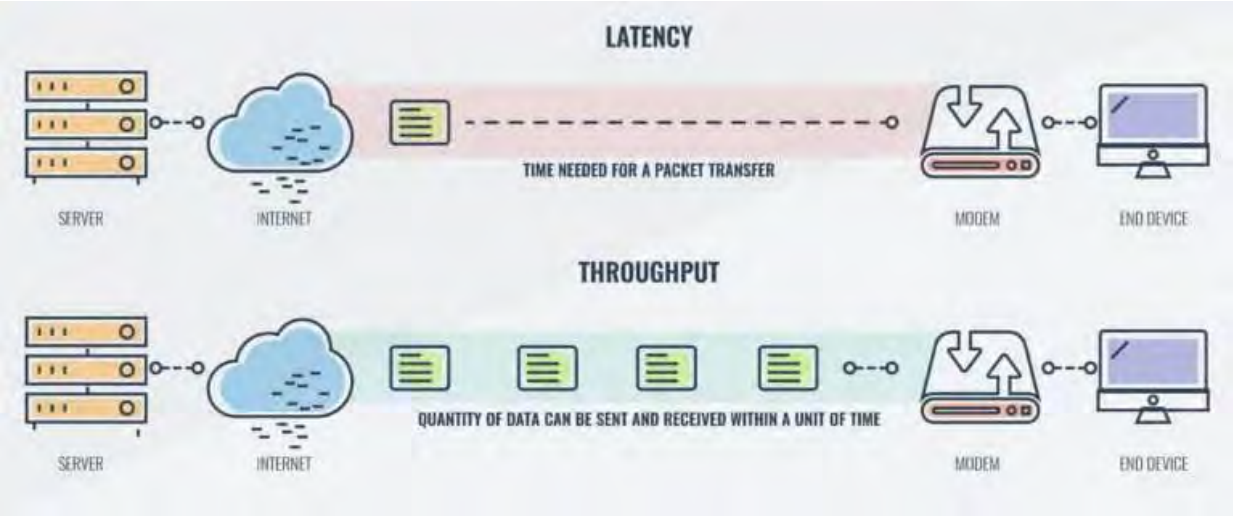
Frames per Second	Distance (feet / m)	Comparison
1	215 / 65 	Statue of liberty (w/o foundation)
5	139 / 42 	The Arc de Triomphe
15	126 / 38.4	Football field + a refrigerator
30	123 / 37.4	Football Field

Flash activity

- Quiz 00: What is Bandwidth, Latency, Throughput, Pipelining
  - Join Mural workspace: link
  - Open whiteboard: link

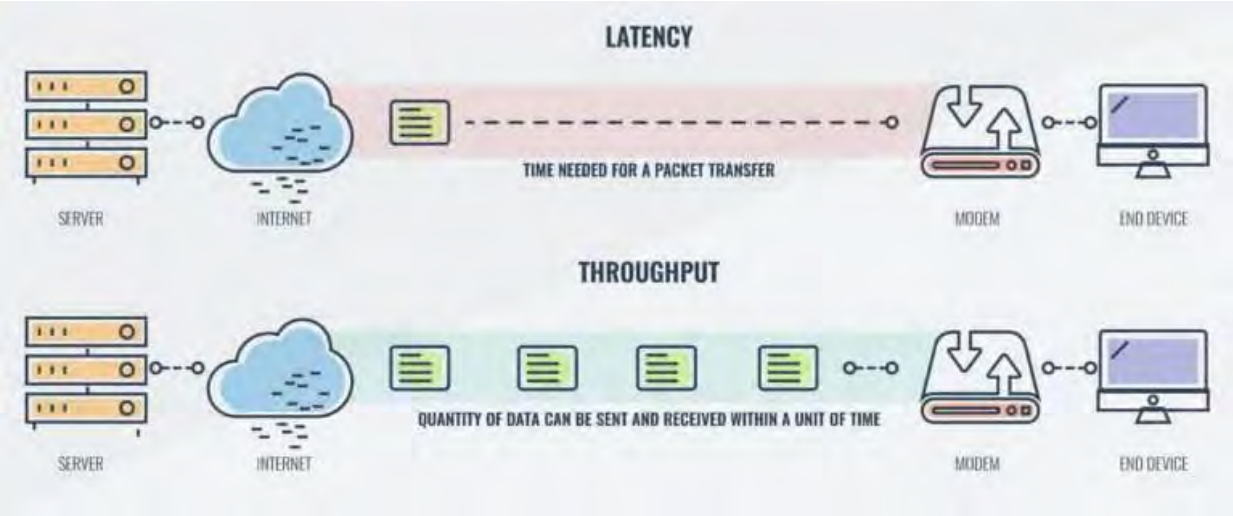


Latency vs Throughput



https://www.comparitech.com/net-admin/latency-vs-throughput/

Latency vs Throughput

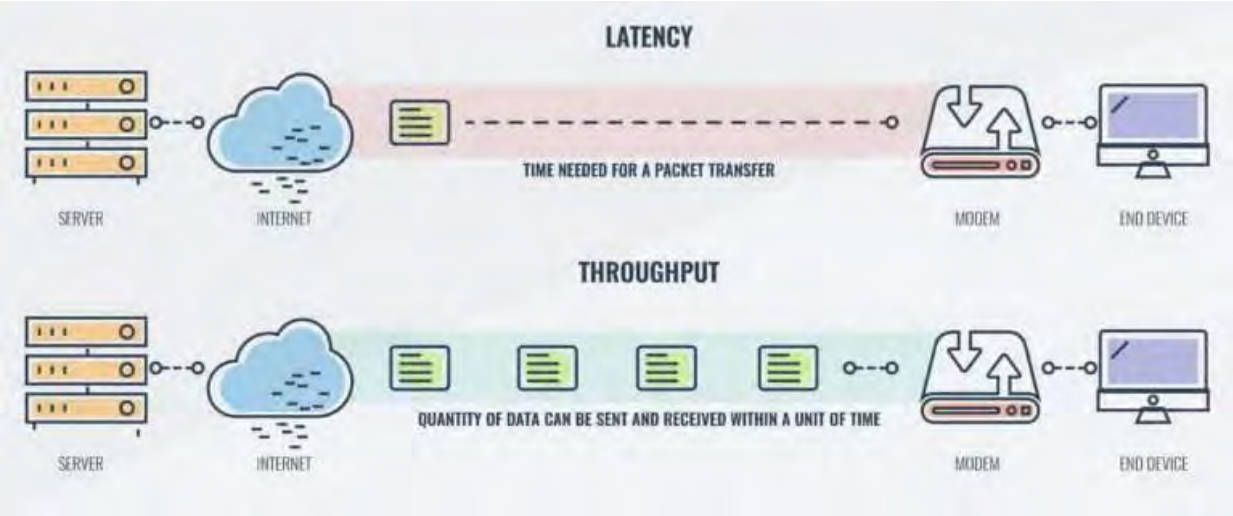


Questions

- Larger bandwidth == shorter latency?

https://www.comparitech.com/net-admin/latency-vs-throughput/

Latency vs Throughput

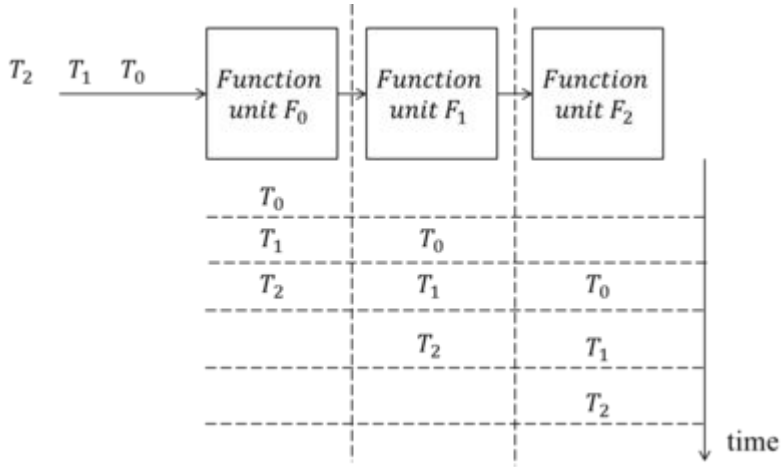


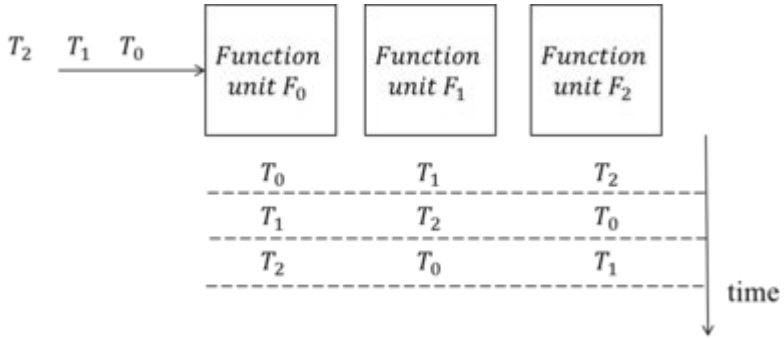
Questions

- Larger bandwidth == shorter latency?
- Larger bandwidth == larger throughput?

https://www.comparitech.com/net-admin/latency-vs-throughput/

Pipeline

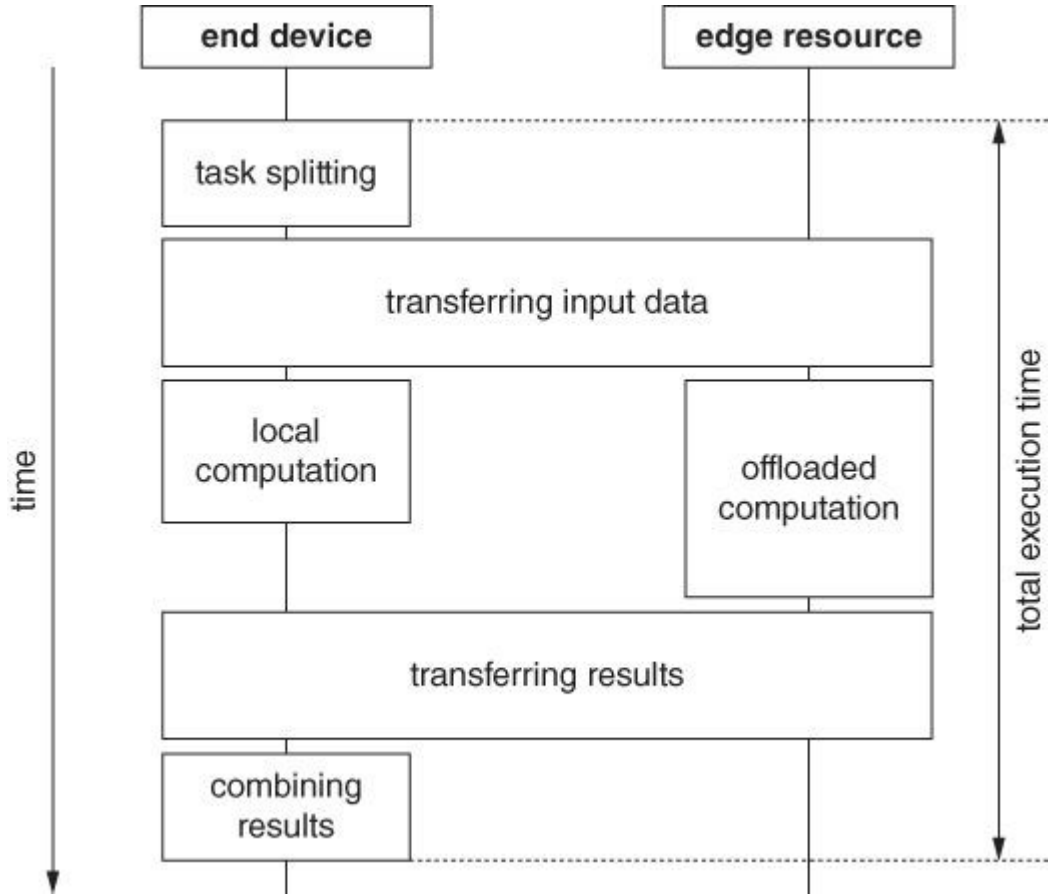




What is parallelizable? Why?

https://en.wikipedia.org/wiki/Pipelining_(DSP_implementation)

Execution Time



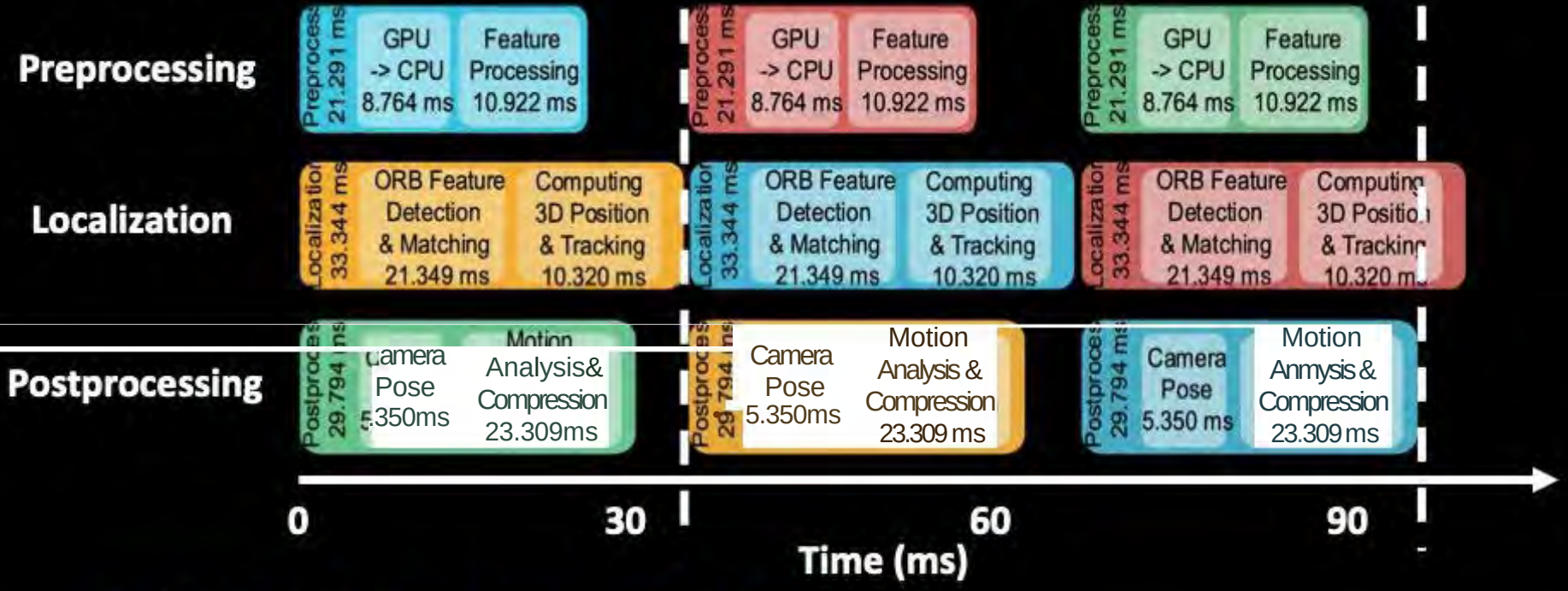
Example

Stage 1: Preprocessing **Stage 2: Localization** **Stage 3: Postprocessing**

18ms 33ms 28ms

- Optimal latency?
- Optimal FPS / throughput?

Latency and Frame Rate

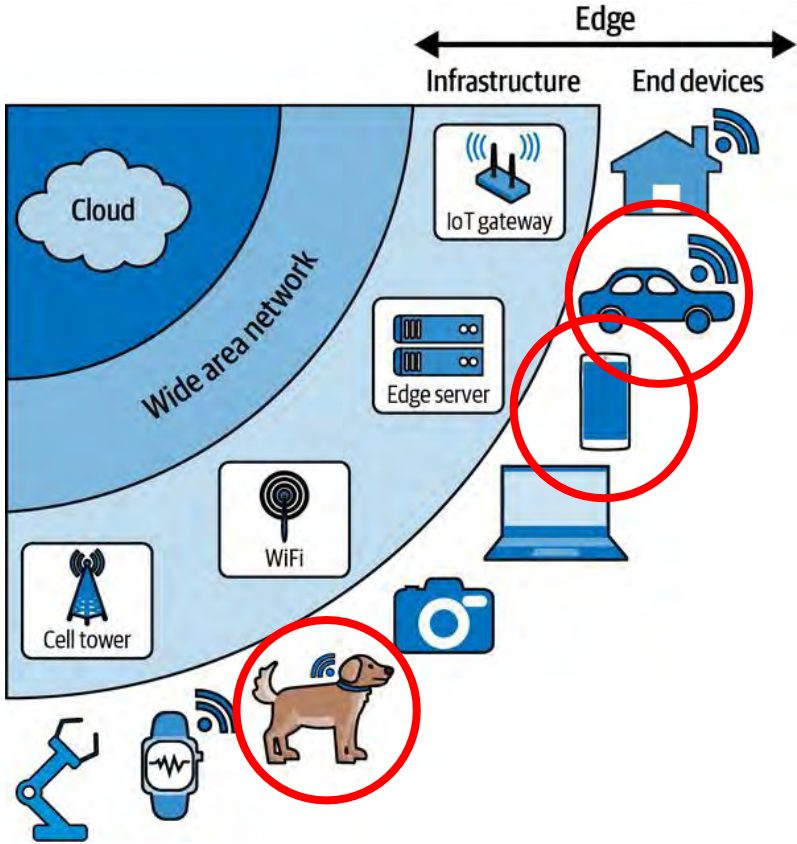


AVR induces 96 ms processing delay

AVR achieves 30 fps using a 3-stage pipeline

Edge Systems Architectures

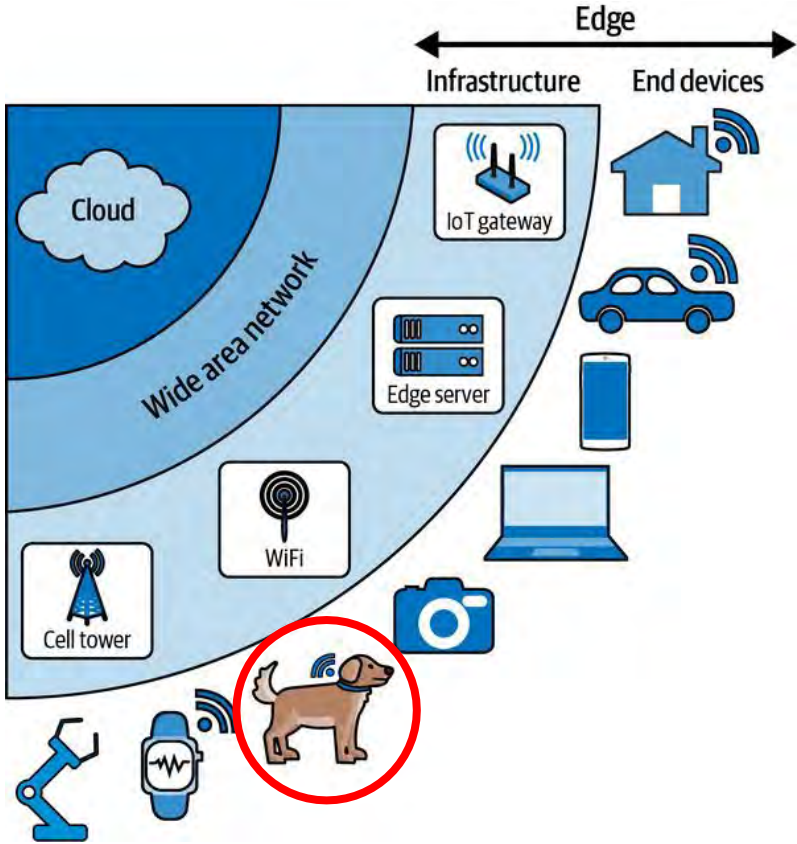
- Wildlife Monitoring
- Voice Assistant
- Self-driving Cars



Situnayake, D., & Plunkett, J. (2023). AI at the Edge. O'Reilly

Edge Systems Architectures

- Wildlife Monitoring



Situnayake, D., & Plunkett, J. (2023). AI at the Edge. O'Reilly

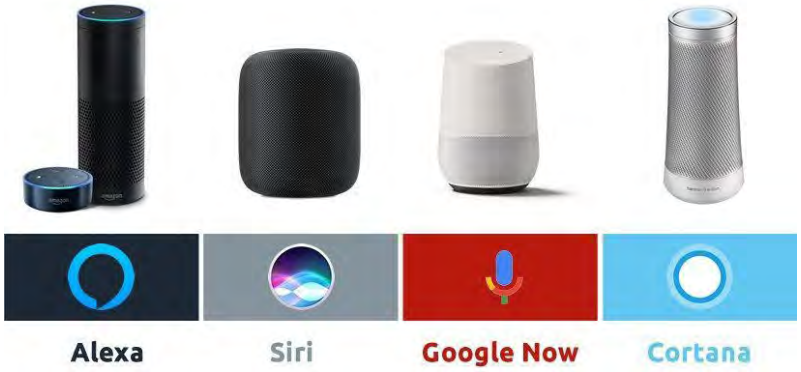
Edge Systems Architectures

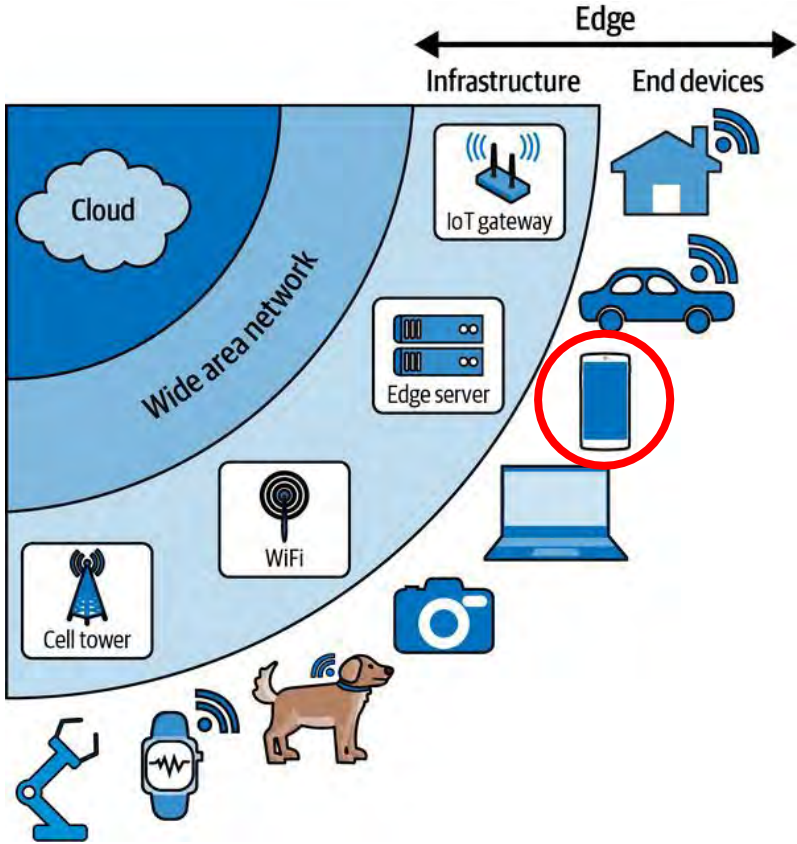
- Wildlife Monitoring
  - Periodic GPS beacons -> cloud

Device Compute	Bandwidth Available	Latency Requirement	Edge/Cloud Compute

Edge Systems Architectures

- Voice Assistant
  - “Hey Google” “Hey Siri” “Alexa” “Cortana”



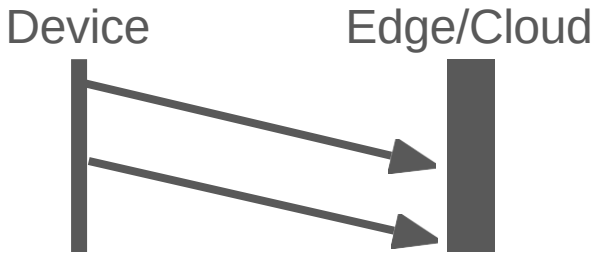


https://www.geeksfl.com/blog/best-voice-assistant/

Situnayake, D., & Plunkett, J. (2023). AI at the Edge. O'Reilly

Edge Systems Architectures

- Voice Assistant
  - Wake-up word -> device
  - Large language model -> cloud

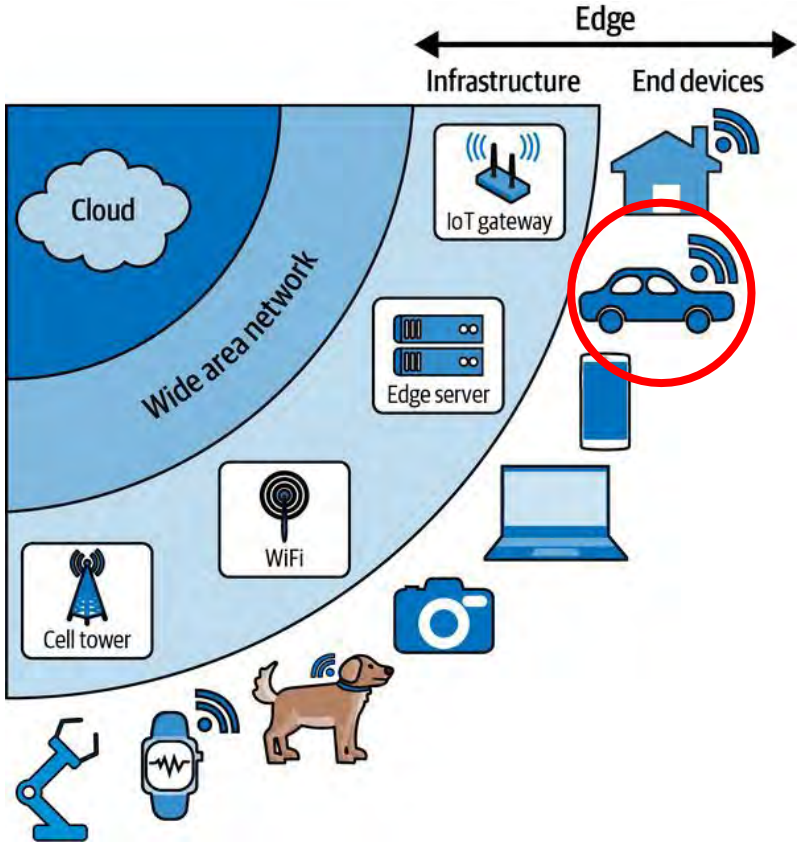


Device Compute	Bandwidth Available	Latency Requirement	Edge/Cloud Compute

https://www.geeksfl.com/blog/best-voice-assistant/

Edge Systems Architectures

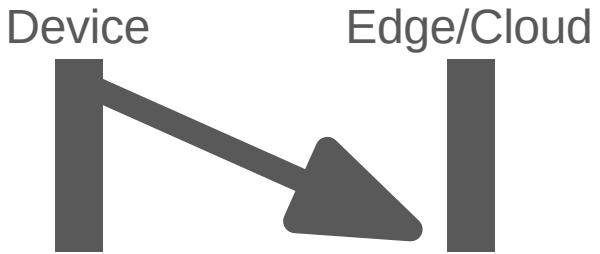
- Self-driving cars



Situnayake, D., & Plunkett, J. (2023). AI at the Edge. O'Reilly

Edge Systems Architectures

- Self-driving cars
  - Sensor data -> onboard compute



Device Compute	Bandwidth Available	Latency Requirement	Edge/Cloud Compute

Offloading

- Workload vs compute

Higher workload

More Compute

Device type	Low-frequency time series	High-frequency time series	Audio	Low-resolution image	High-resolution image	Video
Low-end MCU	Limited	Limited	None	None	None	None
High-end MCU	Full	Full	Full	Full	Limited	Limited
High-end MCU with accelera-tor	Full	Full	Full	Full	Full	Limited
DSP	Full	Full	Full	Full	Limited	Limited
SoC	Full	Full	Full	Full	Full	Full
SoC with accelera-tor	Full	Full	Full	Full	Full	Full
FPGA/ASIC	Full	Full	Full	Full	Full	Full
Edge server	Full	Full	Full	Full	Full	Full
Cloud	Full	Full	Full	Full	Full	Full

Close-the-loop: sensing, compute, actuation

- Moving Data to Compute vs Moving Compute to Data

Remote Operation for Self-driving Cars

Autonomous Vehicles

Perception

Localization

Prediction

Reasoning

Planning

Is it foggy?

Where is that traffic light in the map?

Are peds crossing?

Why is that police officer waving at me?

Shall I pass?

Resolution

Data Visualization

Data Visualization

Request Dispatch

Remote Monitoring

Remote Operator

Smart Home Actuation

- “Hey Google, turn off lights”
- “Hey Siri, close curtains”

Telesurgery

- Robot arm actuation

https://www.zdnet.com/article/surgery-digitized-telesurgery-becoming-a-reality/

Summary

- The IoT Challenge
- Bandwidth, latency, throughput, pipeline
- Example system architectures
- Close-the-loop: sensing, compute, actuation

Next Lectures

- Lab 1: profile performance, data for design optimization
- Optimization techniques
- Edge ML basics