Edge Computing

Lecture 02: Edge Computing & Applications

Agenda

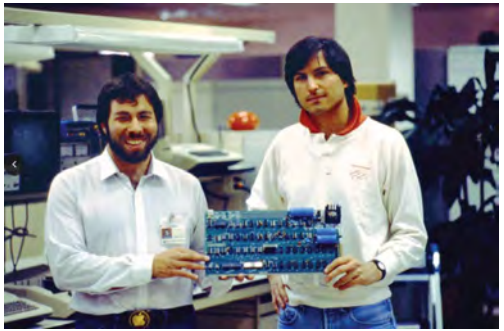
- Edge computing evolution
- Virtualization & containers
- Application case study
- Edge AI App: design to deployment

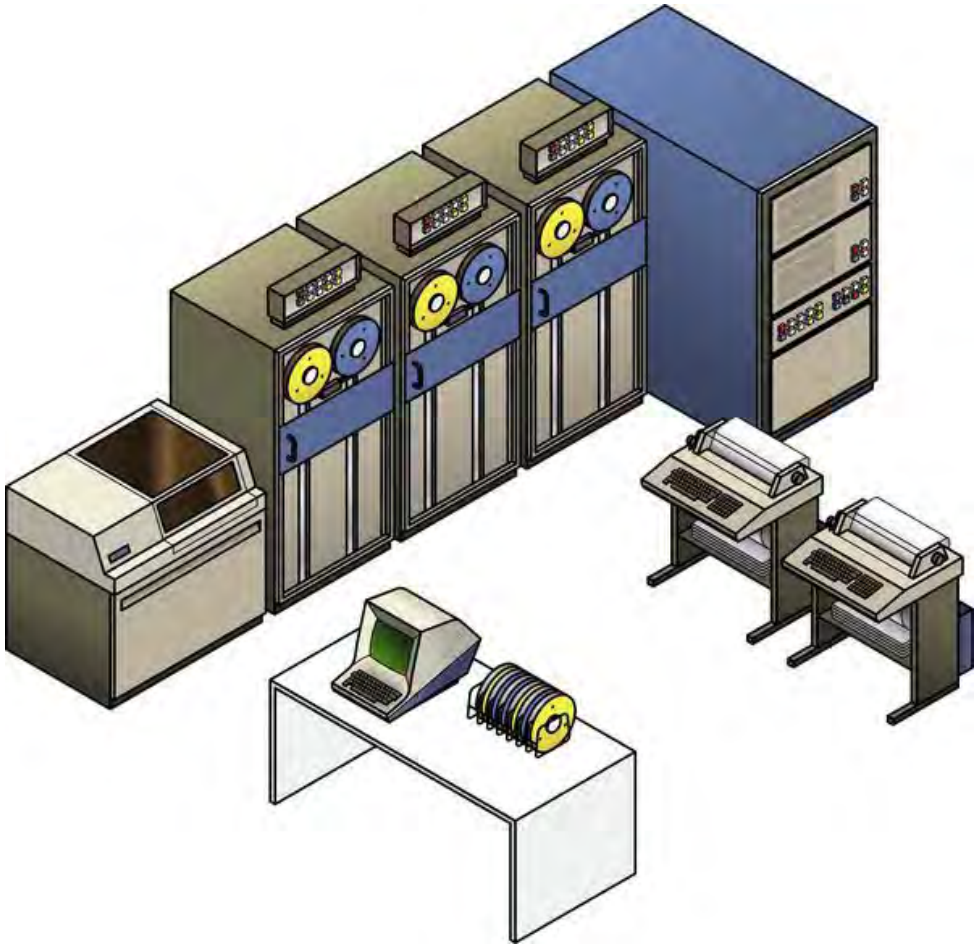
Edge Computing

- Is a paradigm (model, standard, pattern), subset of distributed computing, that pushes computational logic towards wherever data is generated.

The Evolution (Push & Pull)

- Gigantic computer (Central)
  - E.g. IBM mainframe (1964), $5B
- Terminal
  - Rendering letters
- PCs (Dist.)
  - Standalone computation
  - Apple I (1976)
  - Steve Wozniak & Steve Jobs

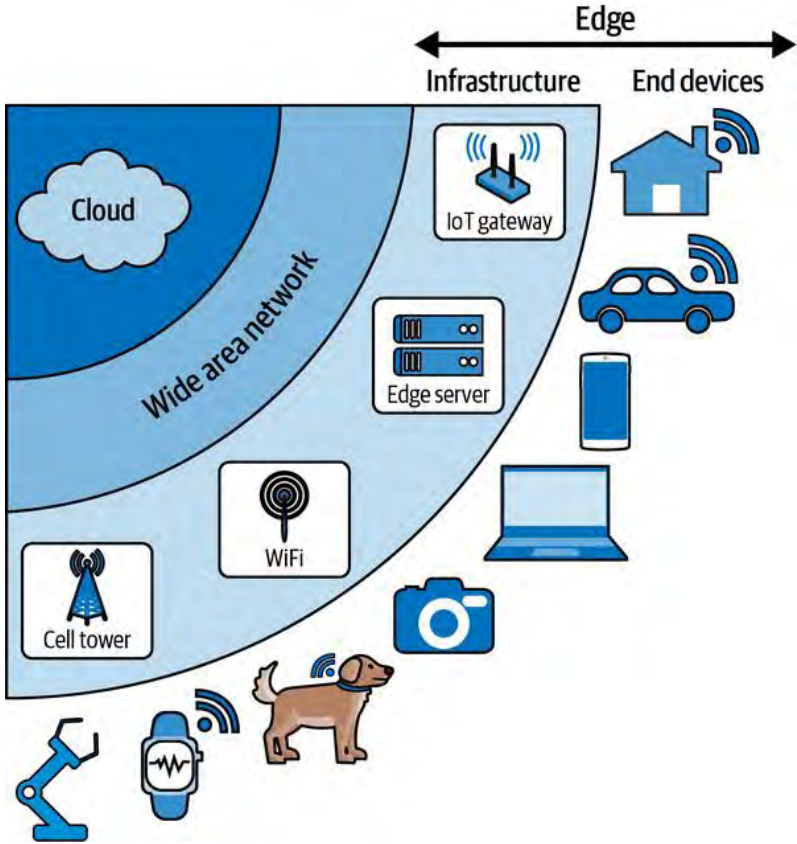




Situnayake, D., & Plunkett, J. (2023). AI at the Edge. O'Reilly

The Evolution (Push & Pull)

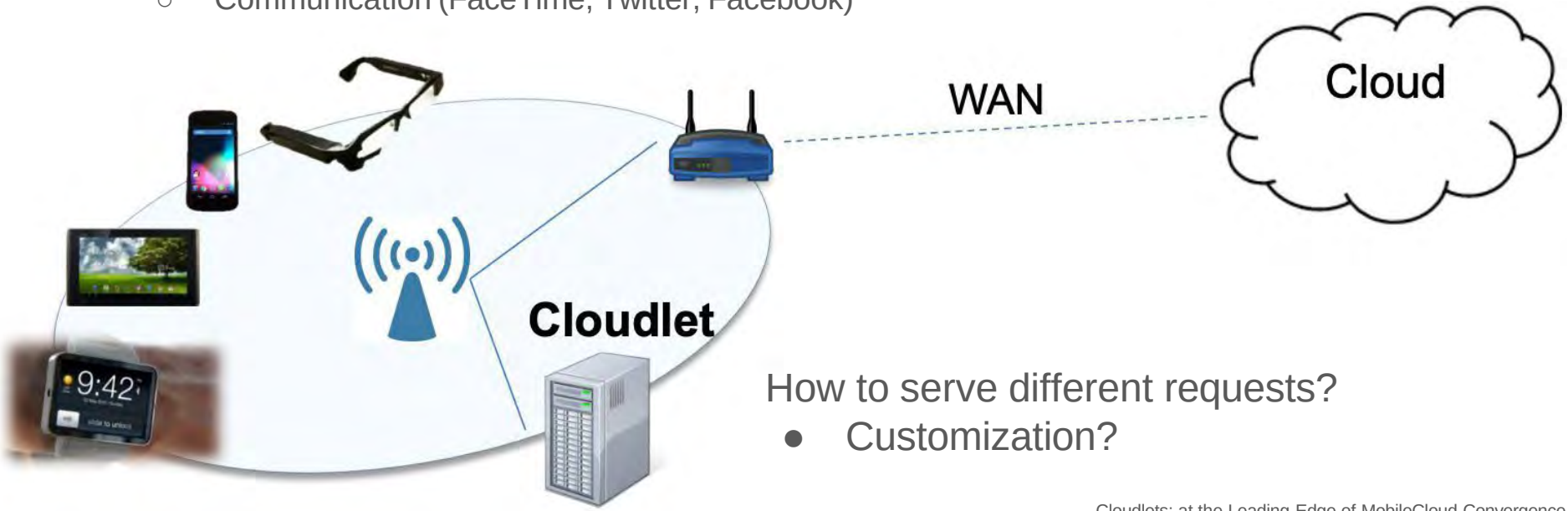
- Cloud computing (2000, Central)
  - Growth of Internet (since 1983)
  - Web applications, remote servers
  - Data Centers, Content Distribution Network (CDN)
  - Everything tethered to the cloud
    - Network is indispensable
- Edge computing (Dist.)
  - Cloudlet (CMU, 2009), for CDN
  - Fog computing (Cisco, 2012)
  - Mobile Edge Computing (ETSI, 2014)
  - IoT, 12.2 B connected devices as of 2021
- What's next? (Cent.?)



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

Cloud View

- Cloud Provider
  - Data (YouTube, NetFlix)
  - Web App (Amazon, Gmail)
  - Communication (FaceTime, Twitter, Facebook)

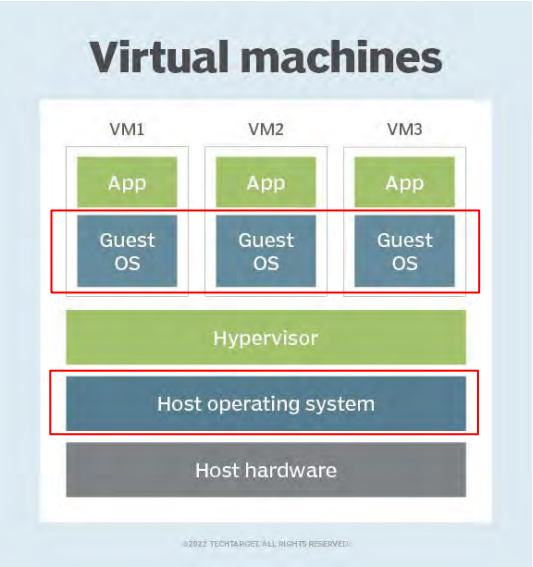


How to serve different requests?

- Customization?

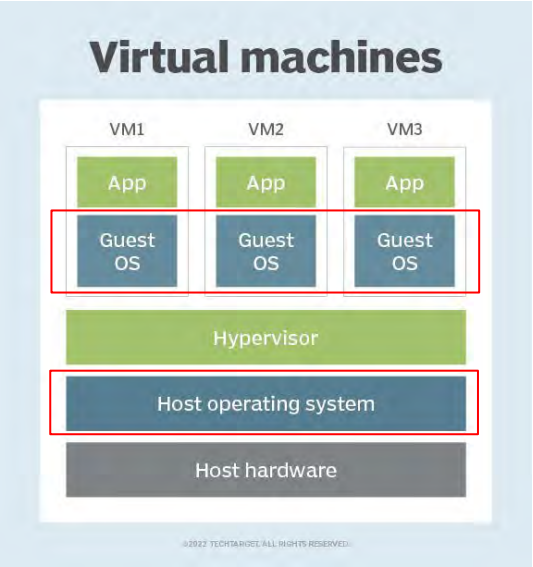
Customized Cloudlet: Virtualization

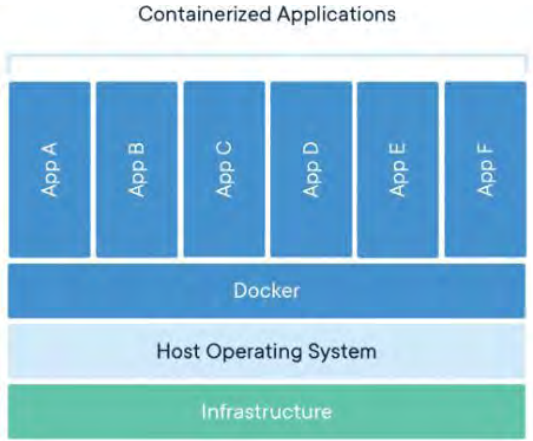
- Virtualization
  - Same physical machines
  - Dynamically provision the resources (CPU, RAM, etc.) upon usage
  - Hardware transparent to users
- Virtual machines
  - A software system to emulate one independent physical machine
  - e.g. VirtualBox



Customized Cloudlet: Virtualization

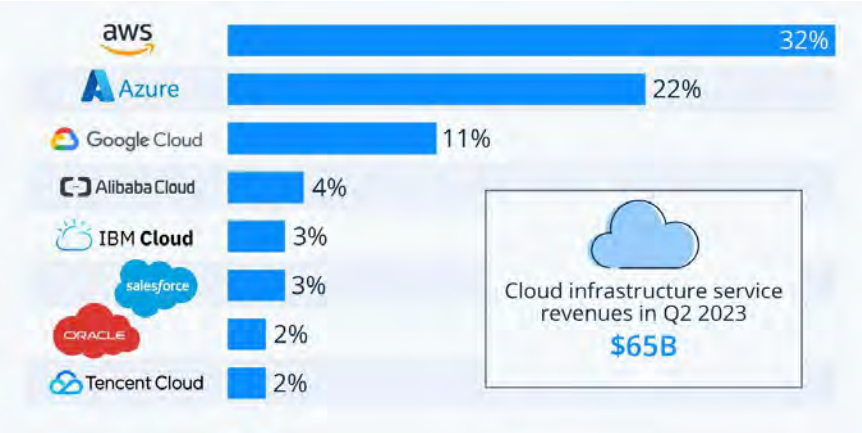
- Virtualization
  - Same physical machines
  - Dynamically provision the resources (CPU, RAM, etc.) upon usage
  - Hardware transparent to users
- Virtual machines
  - A software system to emulate one independent physical machine
  - e.g. VirtualBox, Microsoft Azure
- Container
  - A software package of an application and its dependencies
  - e.g. Docker, Kubernetes





Business Model: Getting closer to users

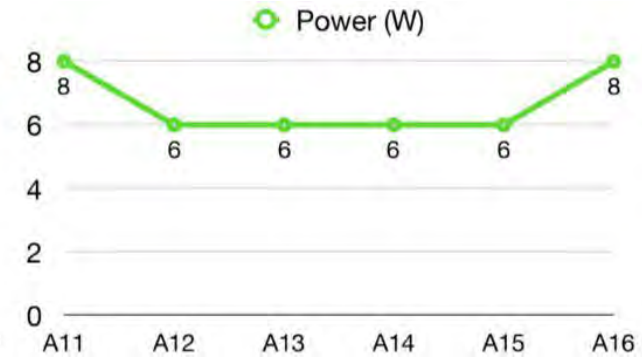
- Cloud Providers
  - Microsoft, Azure IoT Edge
  - Amazon, CloudFront
  - Google
- Internet Service Providers (ISPs)
  - AT&T
  - Verizon
  - T-Mobile

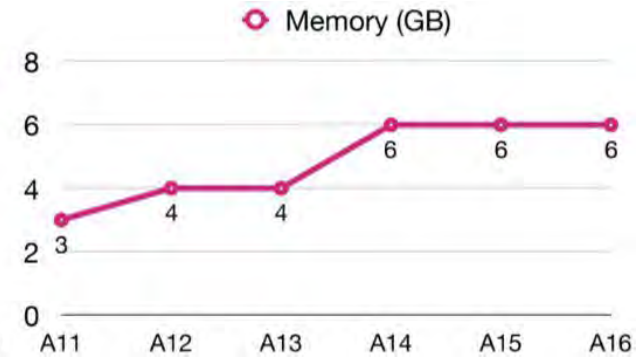


https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/

Edge View

- Growing local computation
  - Provided with enough compute, local is the fastest
- Apple Neural Engine



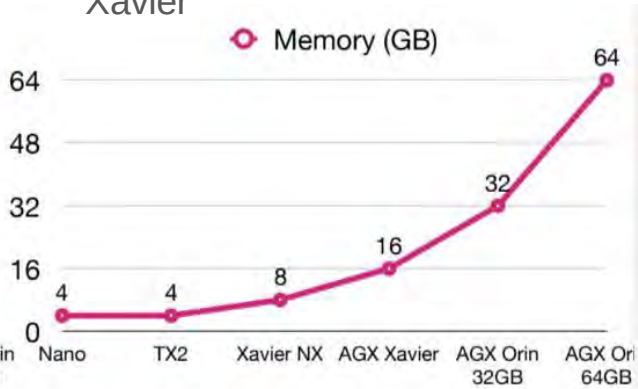


https://nanoreview.net/en & Song, Han, EfficientML.ai

Edge View

- Nvidia Jetson
  - Custom boards vs Dev boards

Nano Xavier AGX Xavier Orin



https://connecttech.com/jetson/jetson-module-comparison/ & Song, Han, EfficientML.ai

Edge View

- Google Tensor Processing Unit (TPU)
  - Equipped since Pixel 6

https://en.wikipedia.org/wiki/Tensor_Processing_Unit & Song, Han, EfficientML.ai

Edge View

- Efficient ML / Tiny ML
  - Quantization
  - Pruning
  - Network architecture search
  - Explainable AI (XAI) for feature selection
- Offloading / Slicing
  - Lazy execution (on / off switch on edge)
  - Model / network / compute slicing

Cent. vs Dist.: Drivers

- 1st Gen Computer -> PC
- PC -> Cloud computing
- Cloud computing -> Edge computing

Cent. vs Dist.: Drivers

- 1st Gen Computer -> PC
- PC -> Cloud computing
- Cloud computing -> Edge computing (BLERP)
  - Bandwidth
  - Latency
  - Economics
  - Reliability
  - Privacy

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

Responsible Design & AI Ethics

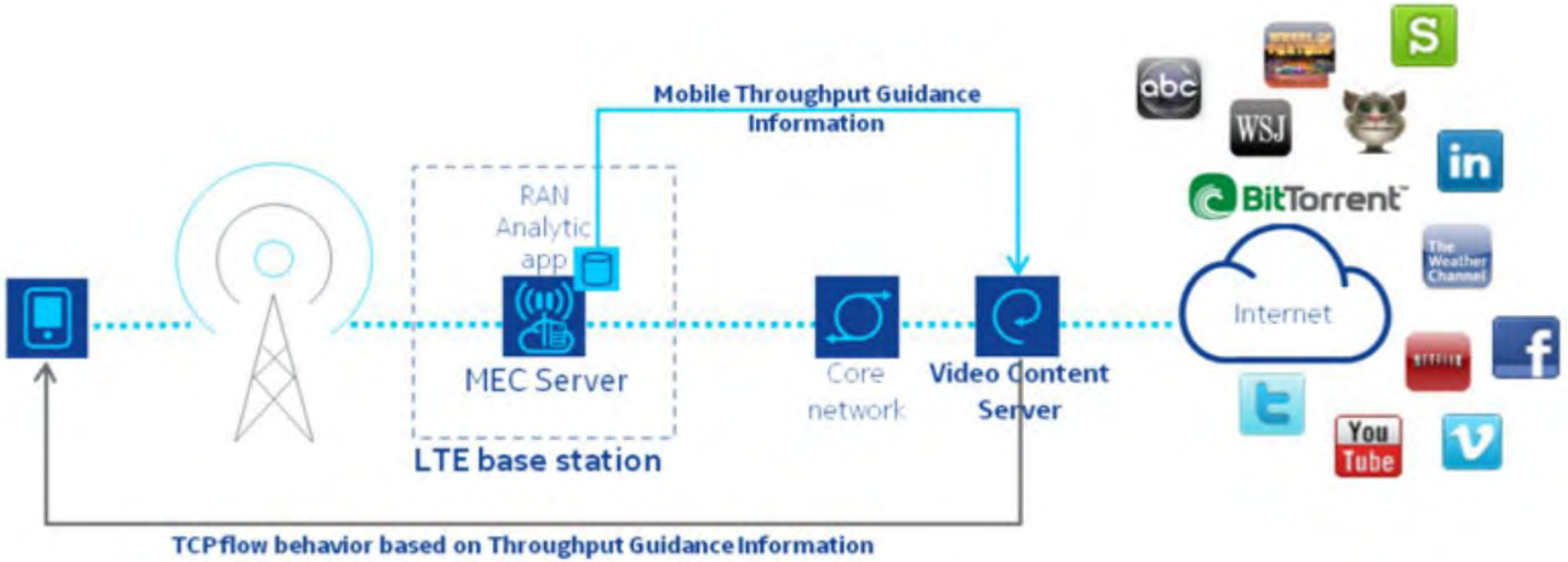
- Tesla, May 2016, March 2018
- Uber, Mar 2018

Sep 2023, Austin, Texas


viralhog

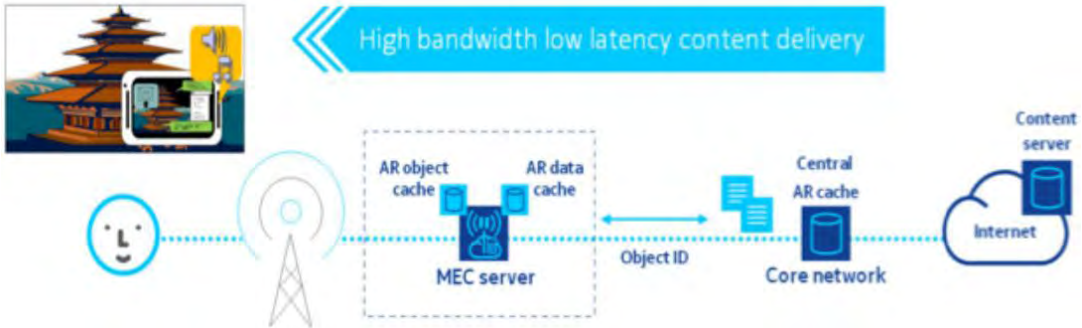
Edge Computing Applications

- Video Streaming
  - Where is the content?
  - How to speed up the streaming?



Edge Computing Applications

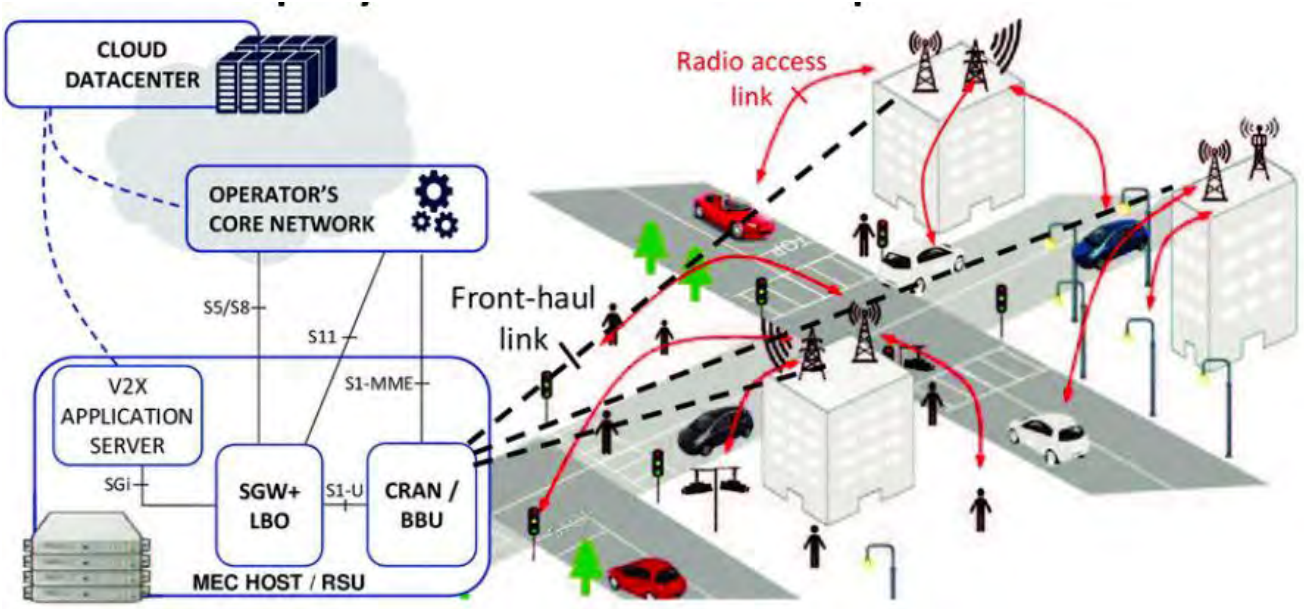
- AR/VR
  - How is content delivered?
  - Where does 3D rendering happen?



https://lh3.googleusercontent.com/p/AF1QipMa0drbYS-V30IcmHJJz5gP2oMMDjBcssFFEUnX=s680-w680-h510

Edge Computing Applications

- Transportation
  - How can edge computing help?



Edge Computing Applications (Project Ideas)

Type of Applications

- Keeping Track of Objects
- Understanding and Controlling Systems
- Understanding People and Living Things
- Transforming Signals

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

Keep Track of Things

- BLERP

Bandwidth	Y
Latency	N
Economics	Y
Reliability	Y
Privacy	N

Use case	Key sensors
Monitoring shipments using smart packaging to detect damage during transit	Accelerometer, vibration, GPS, temperature, humidity
Counting products on store shelves using embedded cameras, so items can be restocked before they run out	Vision
Analyzing the movement of plastic waste in the ocean so it can be cleaned up	Vision
Identifying and tracking obstacles at sea to help ships avoid collisions	Radar
Locating buried natural resources using geophysical sensors	Electromagnetic, acoustic

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

Understanding & Controlling Systems

- BLERP

Bandwidth	Y
Latency	Y
Economics	Y
Reliability	Y
Privacy	N

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

Human-centered Apps

- BLERP

Bandwidth	Y
Latency	Y
Economics	Y
Reliability	Y
Privacy	Y

Use case	Key sensors
Alerting workers in a dangerous environment when they are missing protective equipment	Vision
Understanding human gestures to control a video game	Vision, accelerometer, radar
Identifying when an intensive care patient's health is deteriorating and notifying a medical attendant	Biosignals, medical equipment
Recognizing when a thief has broken into a home and alerting the authorities	Vision, audio, accelerometer, magnetic sensors
Categorizing physical activities using sensors in a smart watch	Accelerometer, GPS, heart rate
Recognizing a user's voice commands and controlling an appliance	Audio
Counting the number of people who are waiting at a bus stop	Vision
Warning a driver when they are falling asleep at the wheel of a car	Vision

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

Apps Involving Living Things

- BLERP

Bandwidth	Y
Latency	Y
Economics	Y
Reliability	Y
Privacy	Y

Use case	Key sensors
Informing researchers when wildlife of interest is spotted by a remote trail camera	Vision, audio
Diagnosing crop diseases in a remote rural location with no cellphone coverage	Vision, volatile organic compound
Recognizing sounds made by marine mammals to track their movements and understand their behavior	Acoustic
Warning villagers of an approaching elephant so they can avoid human-animal conflict	Thermal imaging, vision
Categorizing farm animal behavior using a smart collar to understand health	Accelerometer
Cooking food to perfection by monitoring and controlling a sensor-equipped kitchen appliance	Vision, temperature, volatile organic compounda

Transforming Signals

- BLERP

Bandwidth	Y
Latency	Y
Economics	Y
Reliability	Y
Privacy	Y

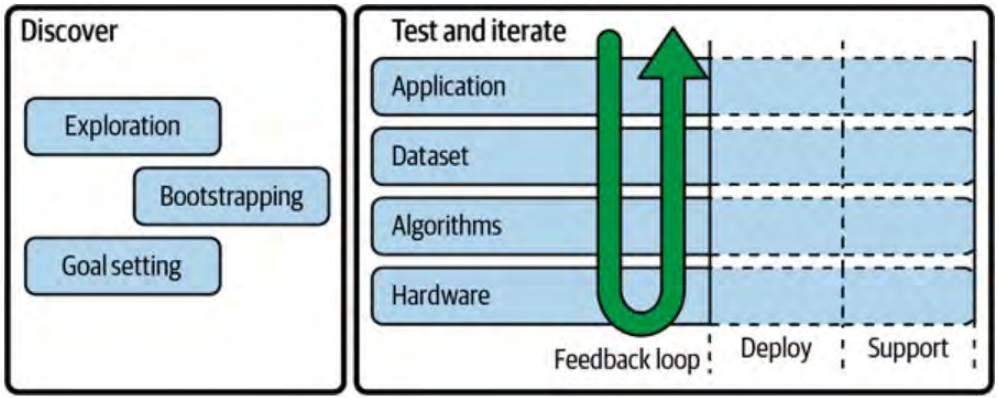
Use case	Signal type
Filtering background noise to improve call quality on a cell phone	Audio
Removing noise from photographs captured with a smartphone camera	Vision
Generating music to accompany a musician during practice	Audio
Blurring the background of a video stream during a remote work meeting	Vision
Generating realistic human speech from text	Audio
Translating one written language into another using a smartphone camera	Vision, text
Upsampling low-resolution audio so that it sounds better to the human ear	Audio
Compressing video using deep learning so that it can be transmitted via a low bandwidth connection	Video
Creating a spoken representation of a visual scene for visually impaired people	Audio
Transcribing a spoken conversation into text for convenience of note-taking	Audio
Using data from cheap sensors to simulate the output of an expensive one	Time series

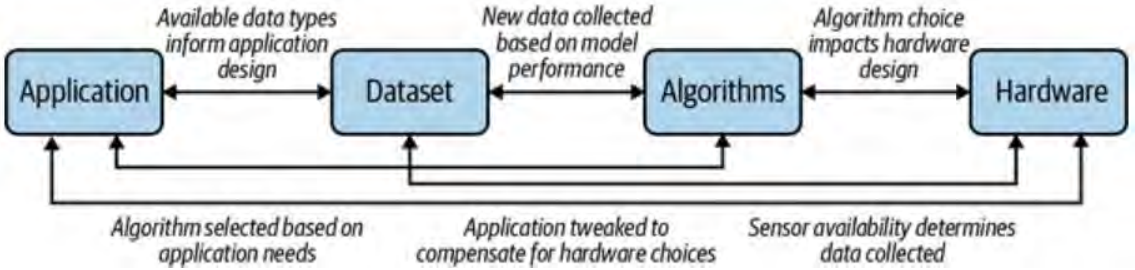
Edge AI App: Design to Deployment

- Finding the problem to solve (what is the pain point?)
- Building a dataset
- Design an application
- *Iteration, iteration, iteration*
- Testing, deployment, and monitoring

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

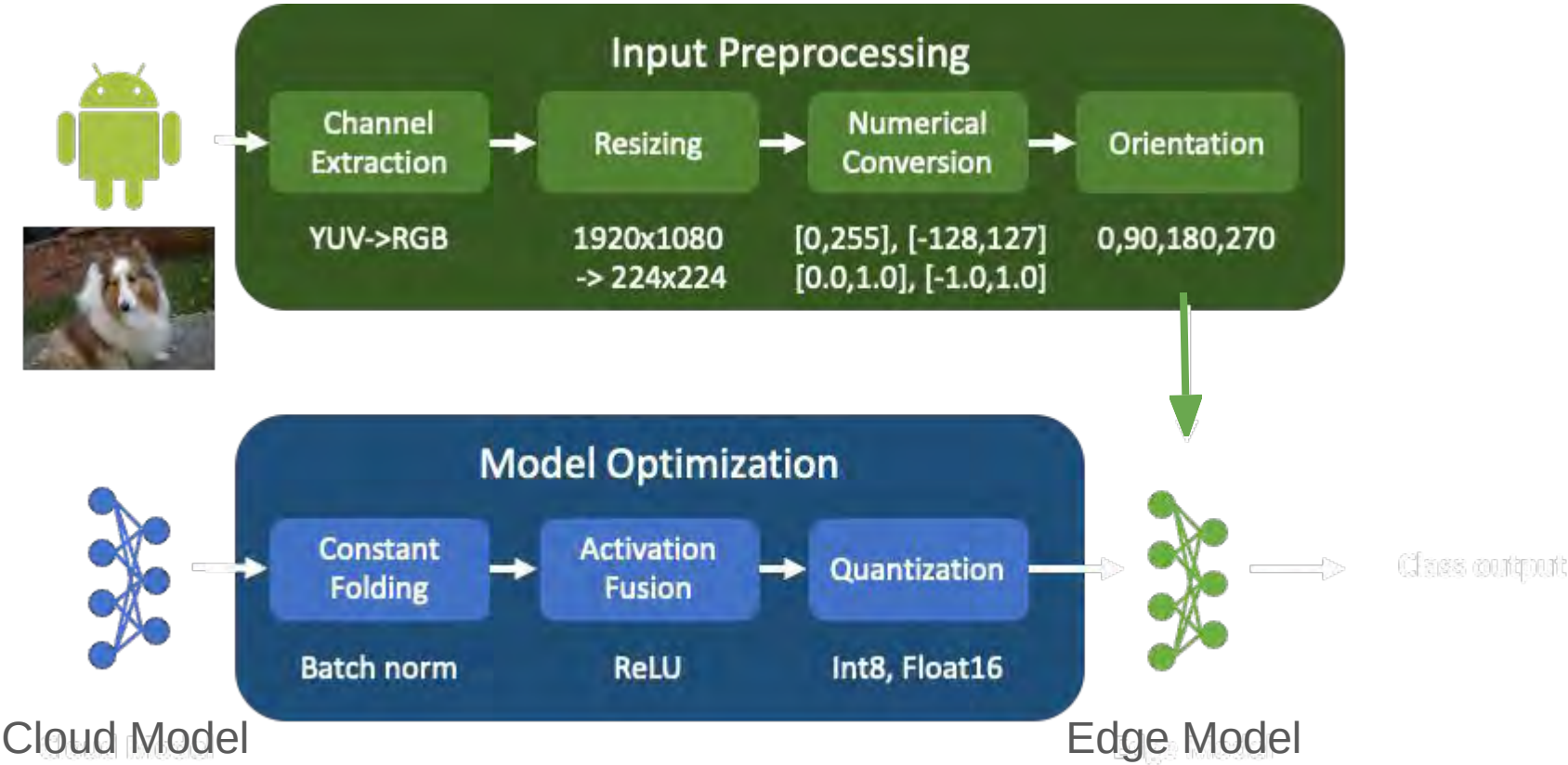
Edge AI Workflow: Intertwined





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

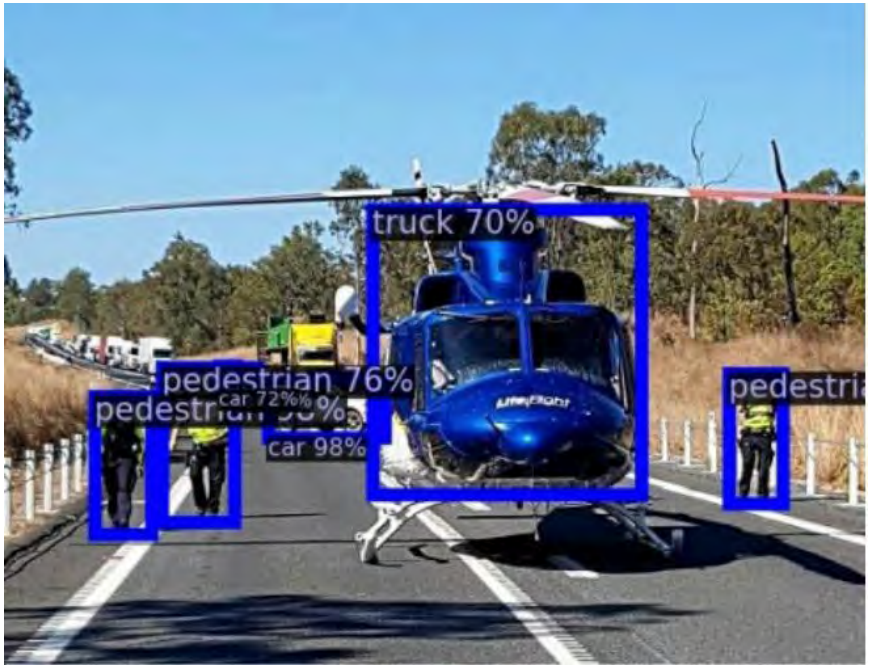
Deployment Example: Image Classifier on Android



Deployment Example: Azure VMs on AT&T

Monitoring and Support

Monitoring and Support



- Design a monitoring approach
- Mining the failure mode
- Resolve the failure online
  - Remote operation
- Continual learning

Summary

- Evolution of computing paradigm
- Cloud vs Edge View
- Virtualization
- Applications
- From design to deployment

Next Lecture

- Edge systems architectures
- Lab 1: performance profiling
- Design and optimization