Introduction to Edge AI

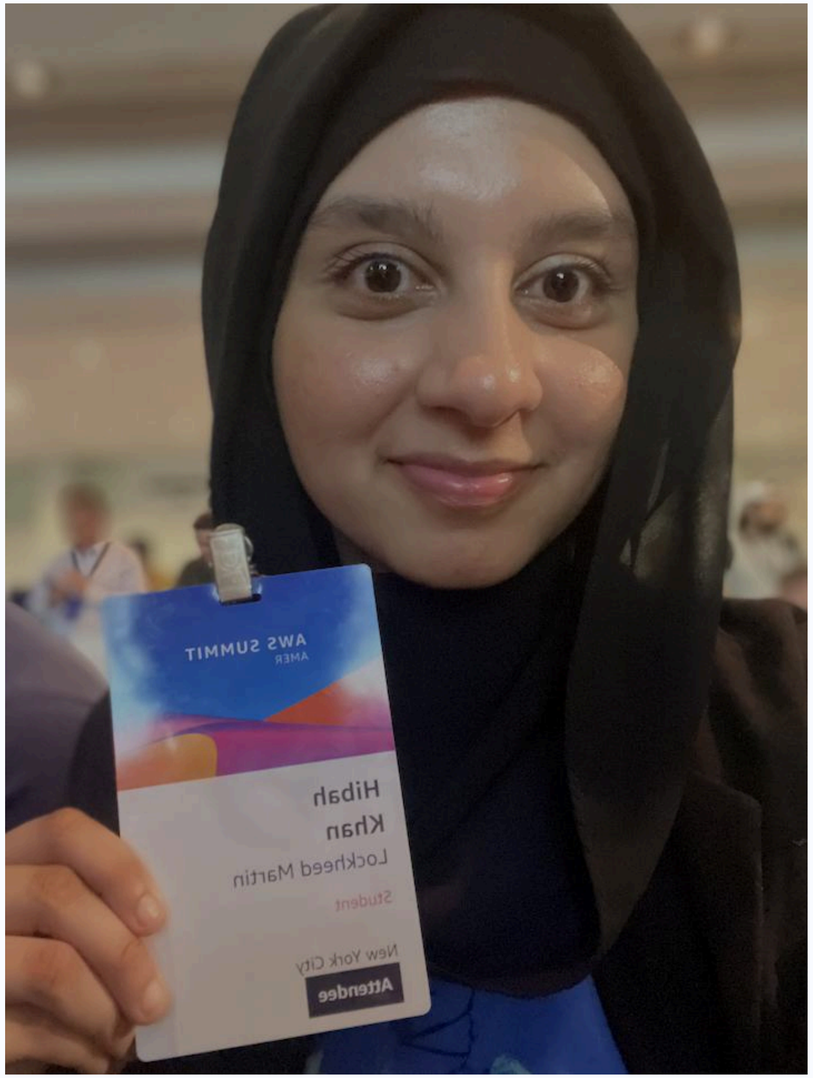
Hibah Khan

January 2025

About Me

Presenter: Hibah Khan
Major: Electrical Engineering at Penn State

Interests: Public Speaking, Programming, surfing the web about new AI concepts!



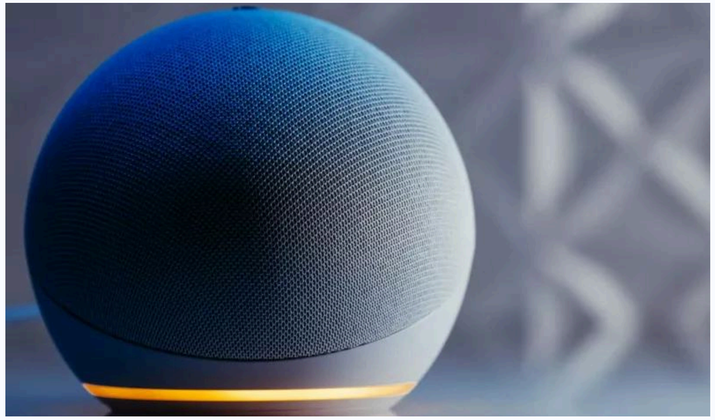
Agenda

1. Edge AI vs. Cloud AI
2. Traditional Use Cases
3. Paradigm Shift – Cloud, Compute, Chips
4. Hands-On Demo
5. Conclusion

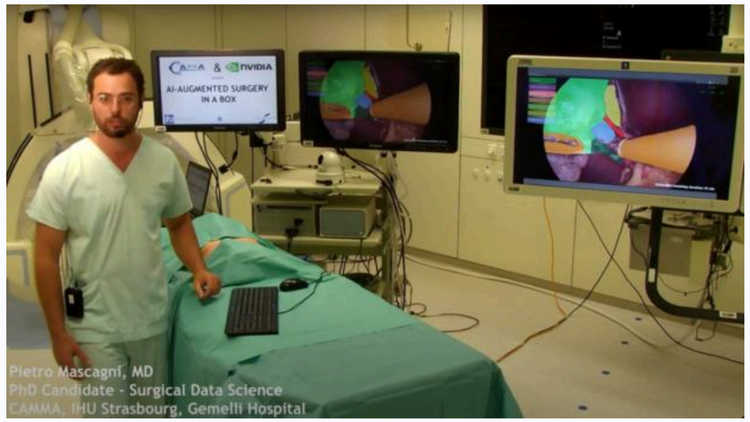
What is Edge AI?

Enables **real-time data processing** and analysis without **reliance on cloud infrastructure**

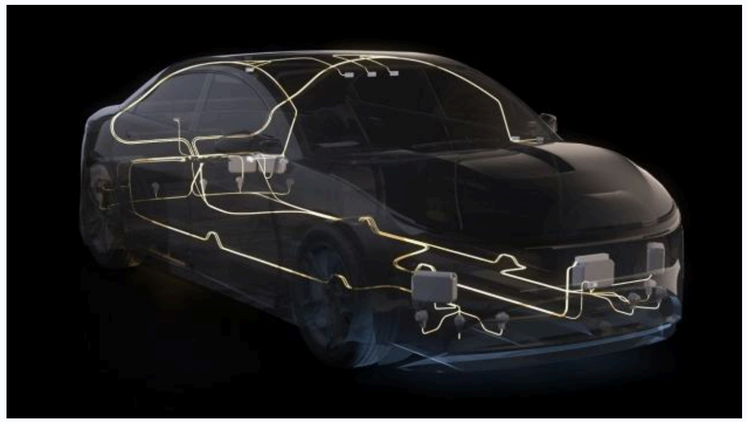
Many use cases...Alexa AI, surgical operations, AV systems, agricultural substitutions, factory operations, etc.



Amazon's Alexa traditional commands are processed instantly on edge



Live applications to help surgical workflow



NVIDIA'S DRIVE Platform uses SoCs that run complex AI models on the vehicle

Edge AI vs Cloud AI

Key Differences

1. Computing Power

Cloud AI can leverage the power of virtual compute resources – think CPUs and GPUs – data centers. Cloud AI provides greater computational power than edge. Edge relies solely on local resources.

2. Latency

The power of fast processing significantly lowers the latency. Latency includes the time and resources needed for data transfer; Cloud AI relies on remote servers and data centers for processing.

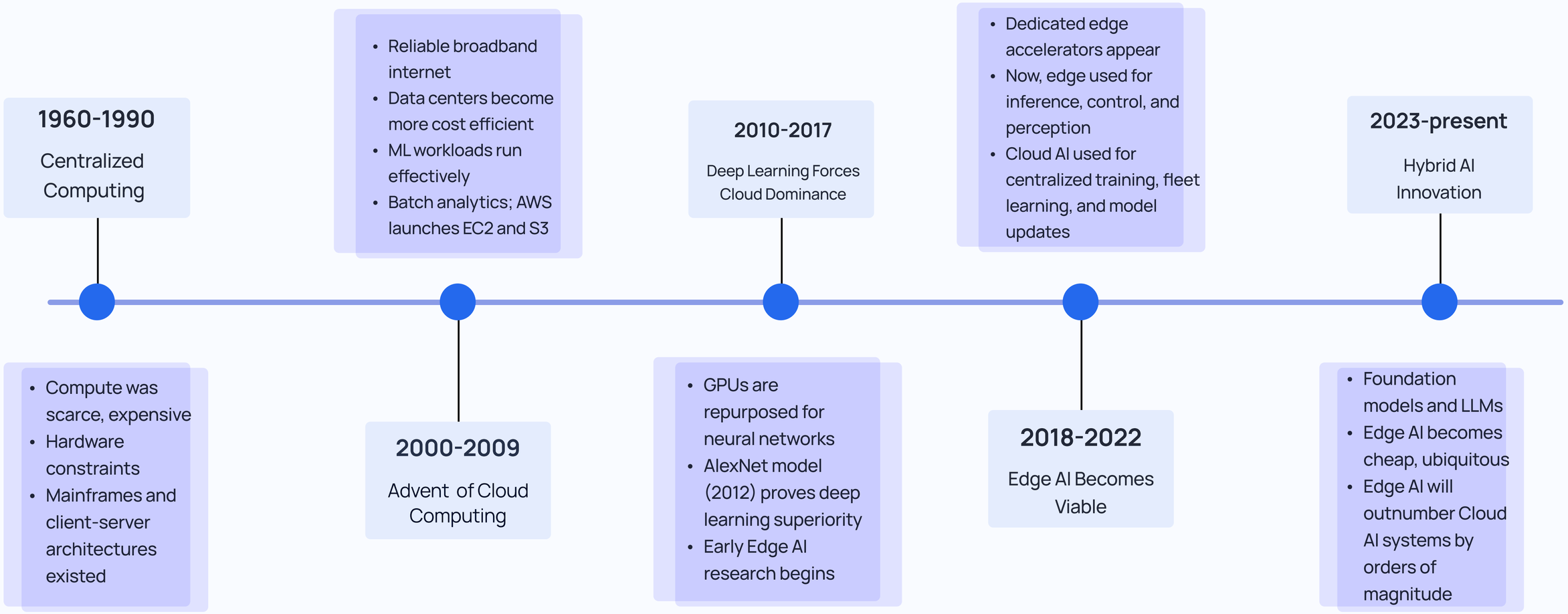
3. Network Bandwidth

Edge AI is considered low bandwidth because it processes data in a box. Cloud AI is considered high bandwidth because it requires a network for data transmission to remote servers.

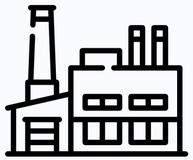
4. Security

Edge AI manages sensitive data locally, on a device where it is gathered, stored, and processed. External data on Cloud AI is transferred between core and local more frequently, making it more vulnerable.

History: Development of Cloud AI -> Edge AI



Edge AI Market Size 2025 - 2030 (USD Billions)



Example: Smart Quality Expansion in a Factory

Scenario: A company makes automotive parts wants to conduct quality inspection

1. Cloud AI

How it Works:

1. Machine captures high-resolution images of parts.
2. Images sent over from network to the cloud
3. Cloud AI models analyze images for defects
4. Results then sent back to the factory

Benefits:

- Easy model updates and centralized learning
- Large compute power allows for advanced analytics

Constraints:

- Latency, bandwidth costs, connectivity issues, privacy and compliance

2. Edge AI

How it Works:

1. Cameras with built-in AI accelerators or edge servers deployed next to production lines
2. Cameras capture image, and model recognizes it in seconds
3. Defective parts are flagged instantly
4. Edge device triggers:
    a. Immediate line stop
    b. Robotic diverter
    c. Alert to operator

Improvements:

- Real-time decisions
- Lower costs
- Cloud-independent operations
- Data privacy

Hands-on Demo: Real-Time Product Objection Detection

Project Overview

Goal: In three categories, identify and label ”skincare”, “makeup” and “perfume”categories

Steps:

1. Preprocess and augment: Make sure model input is consistent with pretrained weights
2. Choose model and training plan: Select a lightweight model MobileNetV3
3. Train and evaluate: Check for training loss, accuracy, and validation accuracy
4. Improve as needed

Connect with me on Linkedin!