Enabling AI at the Edge: design, security, performance, and runtime management

Prof. Danilo Ardagna

danilo.ardagna@polimi.it

AI SPRINT
AI IN SECURE PRIVACY-PRESERVING COMPUTING CONTINUUM

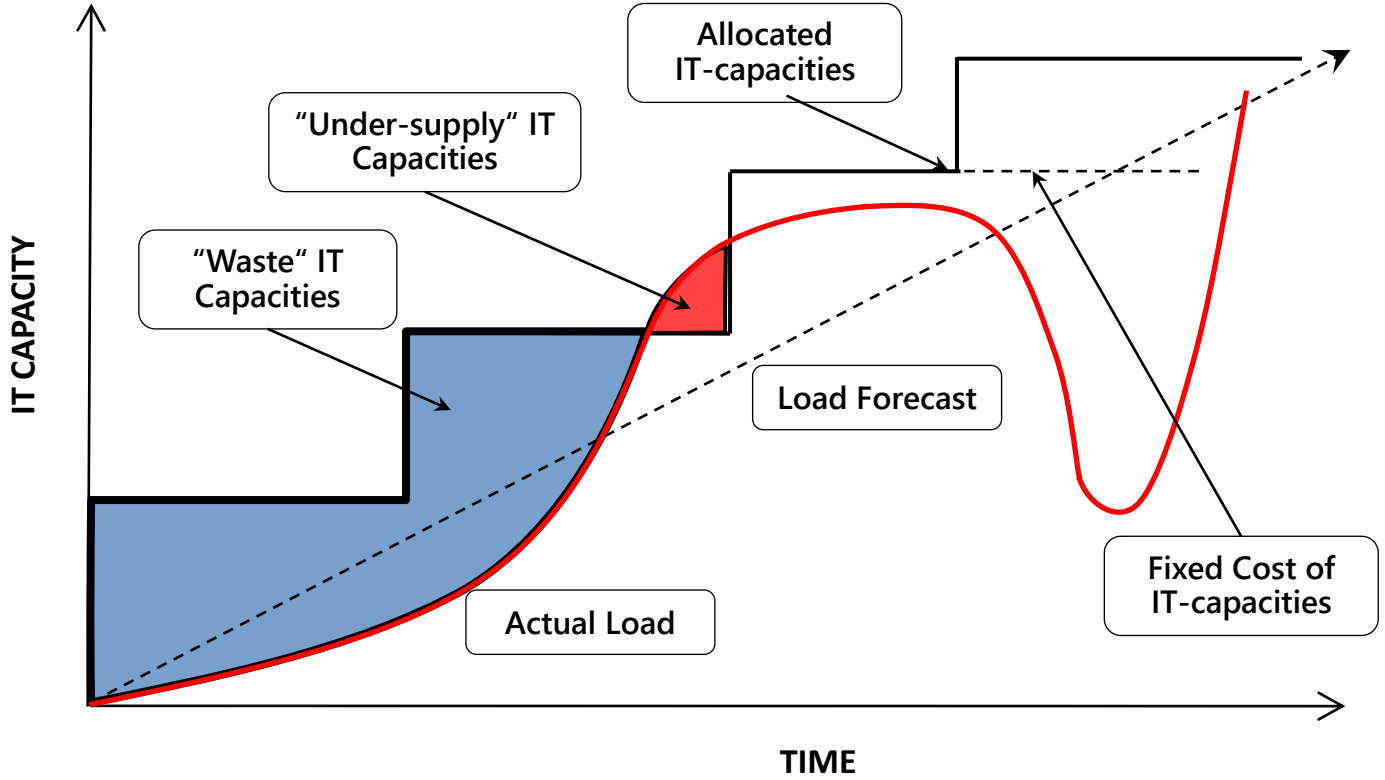
Content

- From Cloud to Edge Computing
- Why Edge AI?
- Course Overview

What is Cloud Computing?

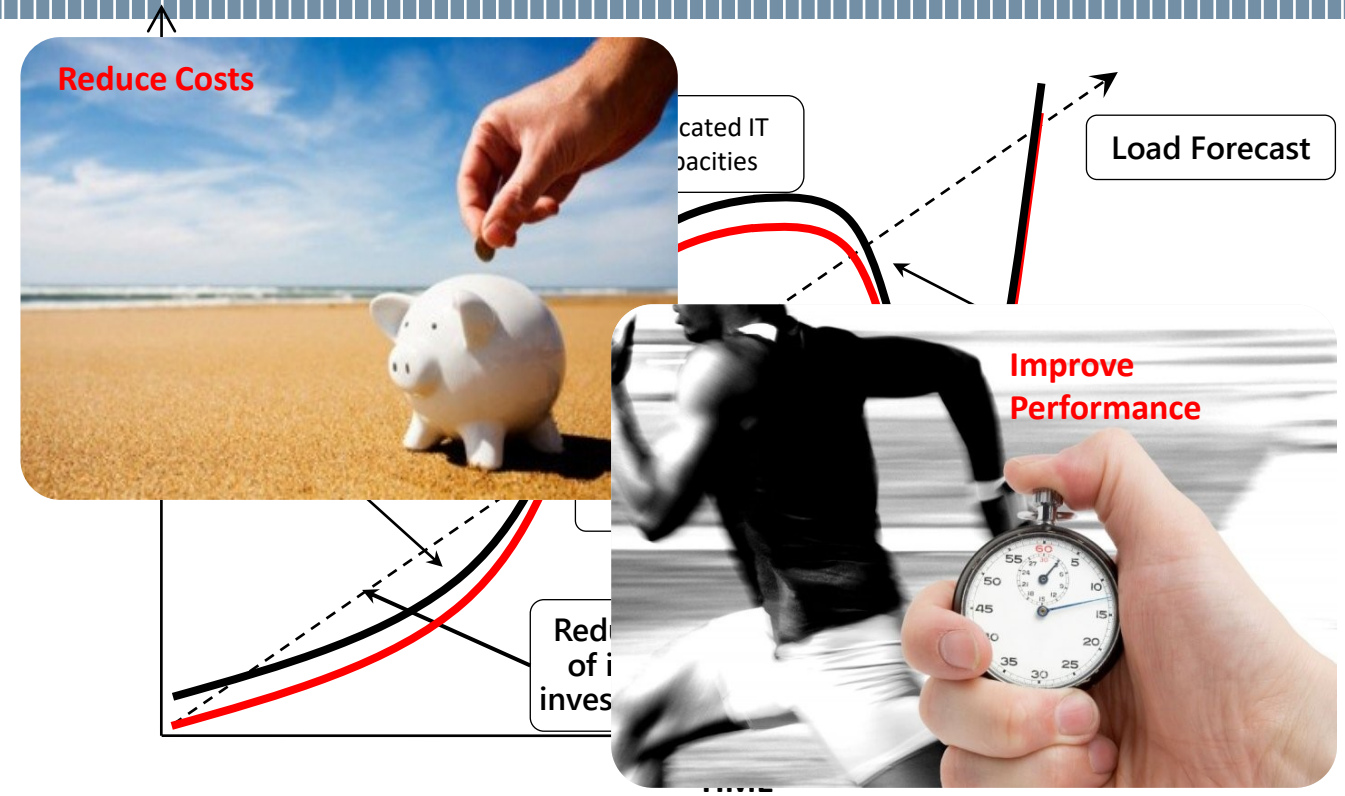
- A coherent, **large-scale**, **publicly accessible** collection of compute, storage, and networking **resources**
- Available via Web service calls **through the Internet**
- Short- or long-term access **on a pay per use basis**

Over-provisioning – Out of Cloud



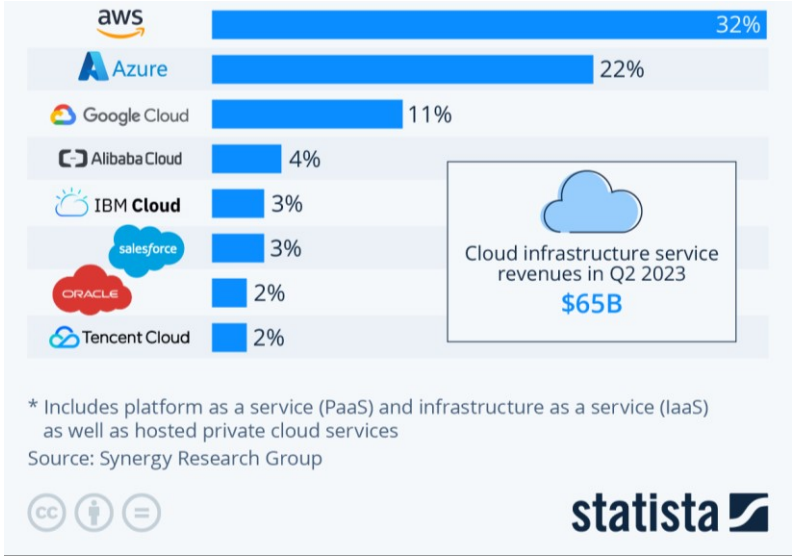
Courtesy of Microsoft

Cloud-provisioning



Courtesy of Microsoft

Cloud Computing Growth



- **Compound Annual Growth Rate (CAGR) 18-20%**
- **Increased emphasis on multicloud strategy**: According to Accenture, **93% of enterprises** have built up to a **multi-cloud** strategy
- **Increase adoption of hybrid cloud services**: Enterprises having their existing infrastructure are moving toward the adoption of cloud computing services and are willing to adopt the hybrid approach so that they can reap the benefits of on-premises and cloud services: According to Flexere 2021 State of the Cloud Report **87%** of enterprises have already adopted **hybrid cloud** strategies
- **Boosting the adoption of edge computing technology**: Most enterprises focus on edge computing as it minimizes delays, which is one of the major factors for any realtime application to perform efficiently. According to Cisco, **the number of devices** connected to IP networks is **more than three times** the **global population** in 2022

Source:
https://www.srgresearch.com/articles/q1-cloud-spending-grows-by-over-10-billion-from-2022-the-big-three-account-for-65-of-the-total

Edge Computing Motivations

Edge computing motivations

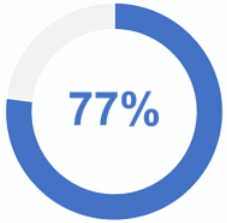
- When it comes to storage and computation of large scales of data, Cloud Computing is the de-facto solution

- With the massive growth in intelligent and mobile devices coupled with technologies like Internet of Things (IoT), V2X Communications, Augmented Reality (AR), the focus has shifted towards
  - gaining real-time responses
  - mobility
  - support for context-awareness

- Due to the delays induced on the WAN and location agnostic provisioning of resources on the cloud, there is a need to bring the features of the cloud closer to the consumer devices

IoT

- 41.6 billion IoT devices in the field by 2025
- These devices include machines, sensors and cameras as well as industrial tools
- The combination of IoT devices are expected to generate 79.4 zettabytes of data in 2025
- Approximately 23% of the devices will be located in Europe; 26% in China and 24% in North America



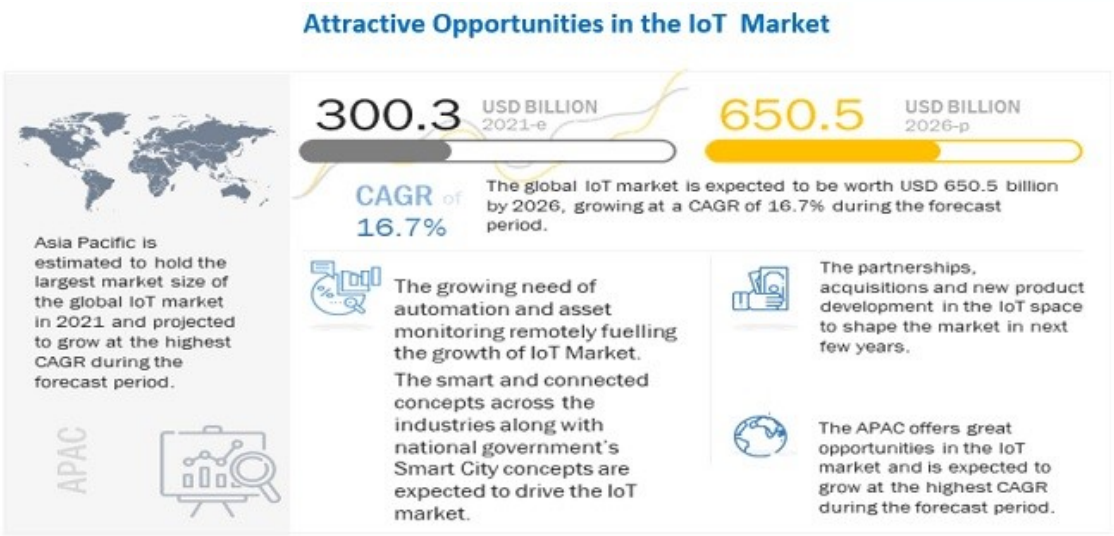
of European organizations plan to invest in IoT

European spending on IoT in 2022

Source: IDC

IoT

- 41.6 billior
- These dev
  and camer
- The combi
  to generat
- Approxima
  located in
  North Ame



Source: Investor Presentation, Secondary Literature, Expert Interviews, and MarketsandMarkets Analysis

EUR 176 bn

European spending on IoT in 2022

Source: IDC

Cloud limitations

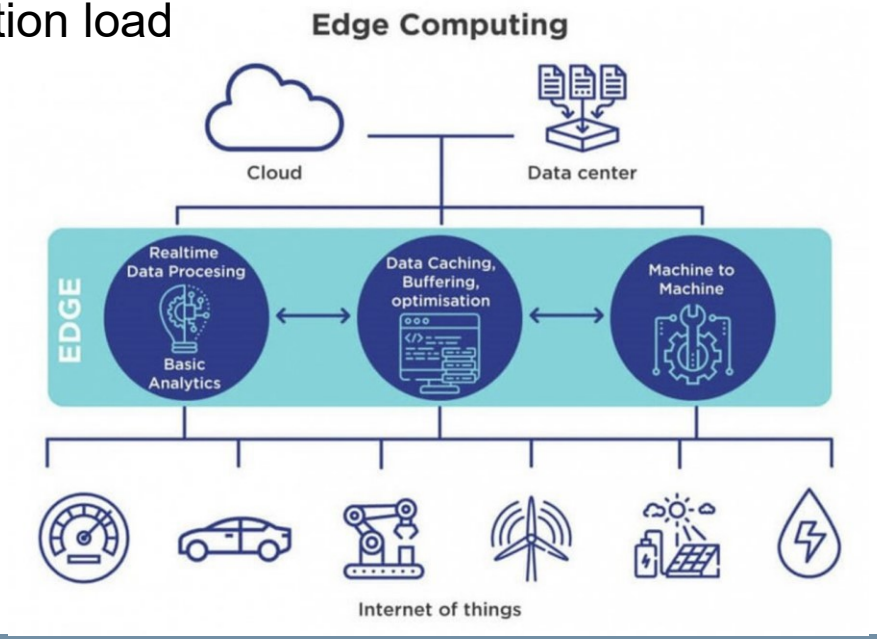
- Currently, the preferred deployment of application is based on cloud solutions, but:
  - A lot of devices produces a lot of data
  - Performance are on the shoulder of only clouds
  - Data transfer introduces latency

- Cloud is not enough:
  - Network dependent (**latency** and **continuous connectivity**)
  - Lack of Data sovereignty
  - Vendor lock-in problem

Edge computing

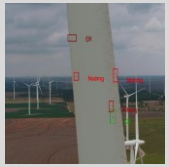
- Provides an intermediate layer between the end devices and the cloud

- Introducing ”Edge devices” the computation load at the data centers are reduced by handling some of the requests directed to the cloud locally:
  - reduced latency
  - allow real-time handling of a subset of requests
  - support mobility



An example: Maintenance and Inspection

Damage detection, classification and severity assessment



CLOUD BASED SOLUTION

Manual data upload to the cloud and cloud only data processing

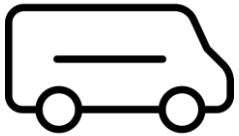
EDGE SOLUTION

Immediate data analysis on-site

Provide on-site data analysis capabilities for immediate data quality assessment and data volume reduction.

DATA ACQUISITION AND QUALITY ASSESSMENT

- Manual or auto flight UAV mission
- Data quality assessment with immediate feedback to the UAV operator
- Blade part detection



ON-SITE DATA PROCESSING

- Clearing house process
- Optional data compression (semantic segmentation)

CLOUD DATA PROCESSING

- Damage detection
- Damage and severity classification
- Data preparation for reporting

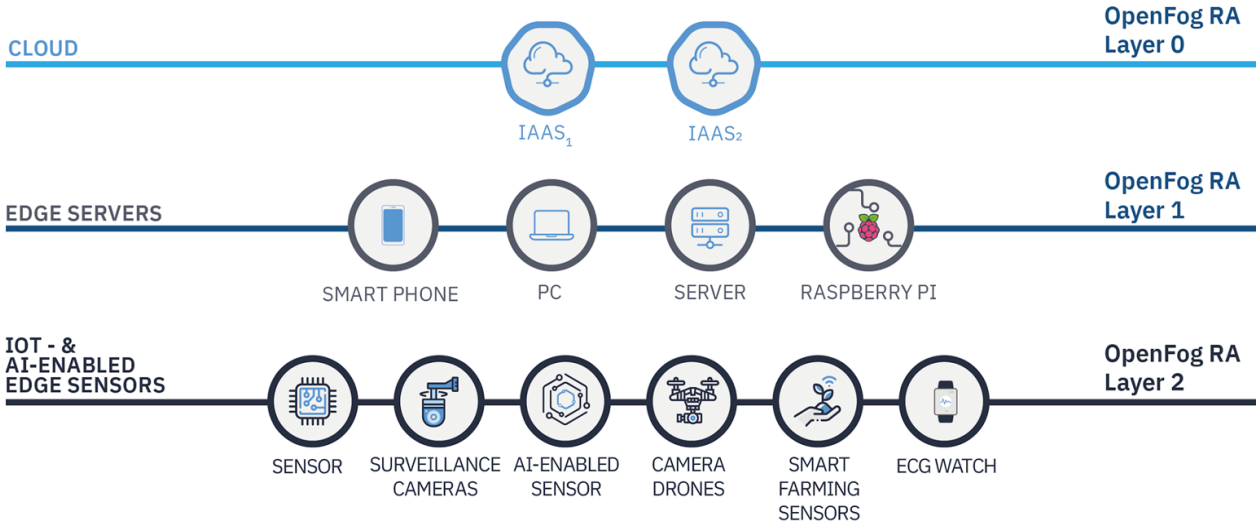
Edge computing benefits summary

- Edge computing makes the cloud truly distributed
- Delivers local storage, compute, and network services
- Moves core cloud services closer to the origin of data
- Mimics public cloud platform capabilities
- Reduces the latency by avoiding the roundtrip to the cloud

Why Edge AI?

Edge AI Global trends

- By 2026, AI worldwide market will approach $900 billion (CAGR 18.6%[1])
- AI needs resources at the edge of the network
- New challenges from the infrastructural perspective

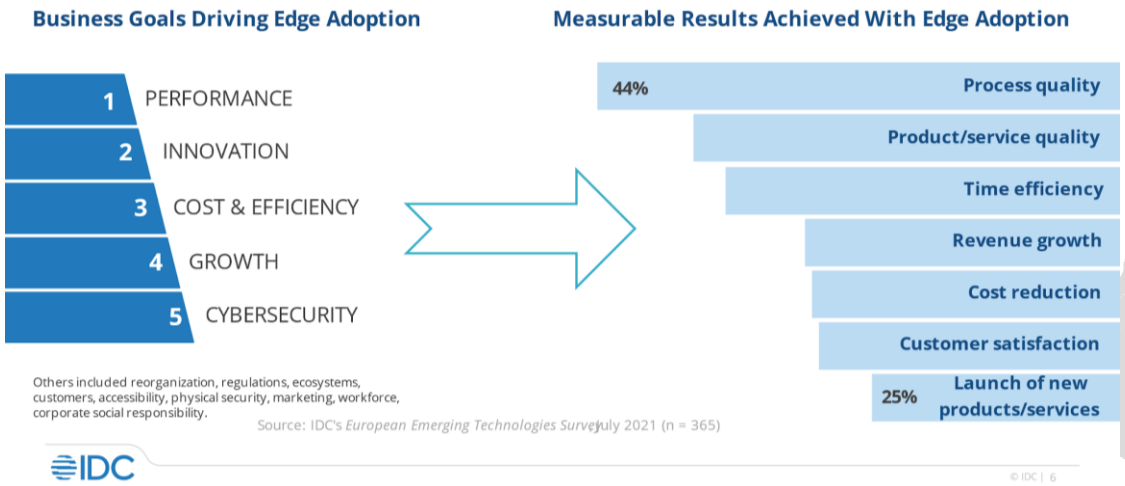


[1]IDC Semiannual Artificial Intelligence Tracker, July 2022

[2]IDC Worldwide Edge Spending Guide, August 2022

What about Europe and Which are the Challenges?

European organizations are already using **artificial intelligence** and **edge** technologies and this number is expected to **double** over the next two years



European **Enterprise edge spending** in Europe approaching

$33B

Cloud Services
(by 2026)
23%

Artificial Intelligence at the edge
(by 2026)
12%

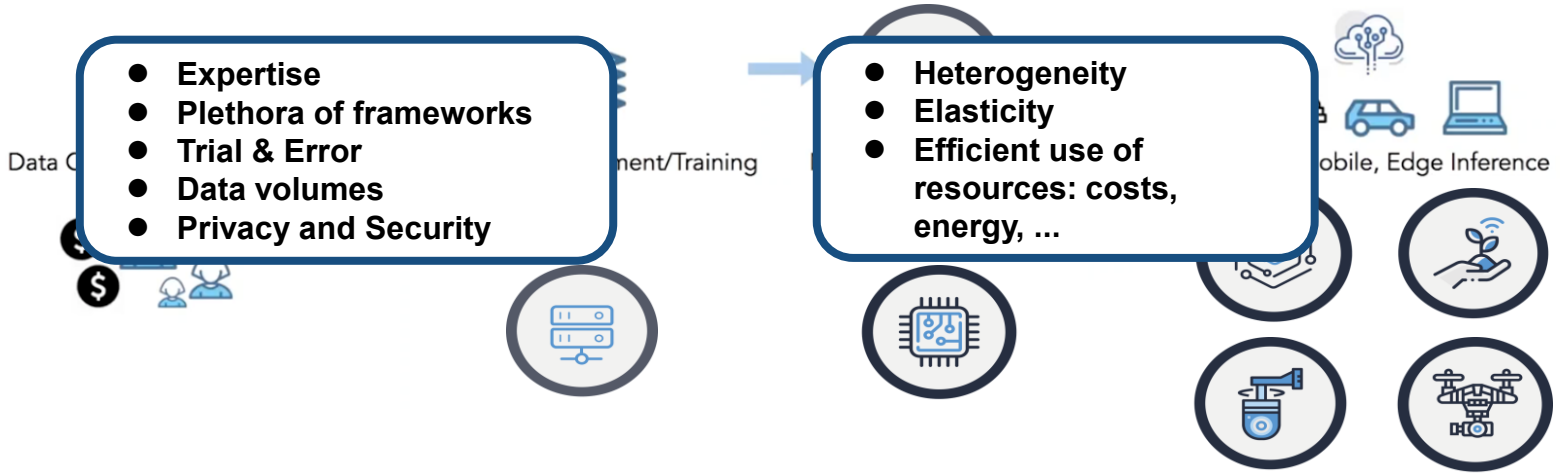
Source: IDC Research - IDC WW Edge Spending Guide - v1 2022; IDC European Emerging Technologies Survey, July 2021

Edge AI Challenges

Engineering hungry

Cloud hungry

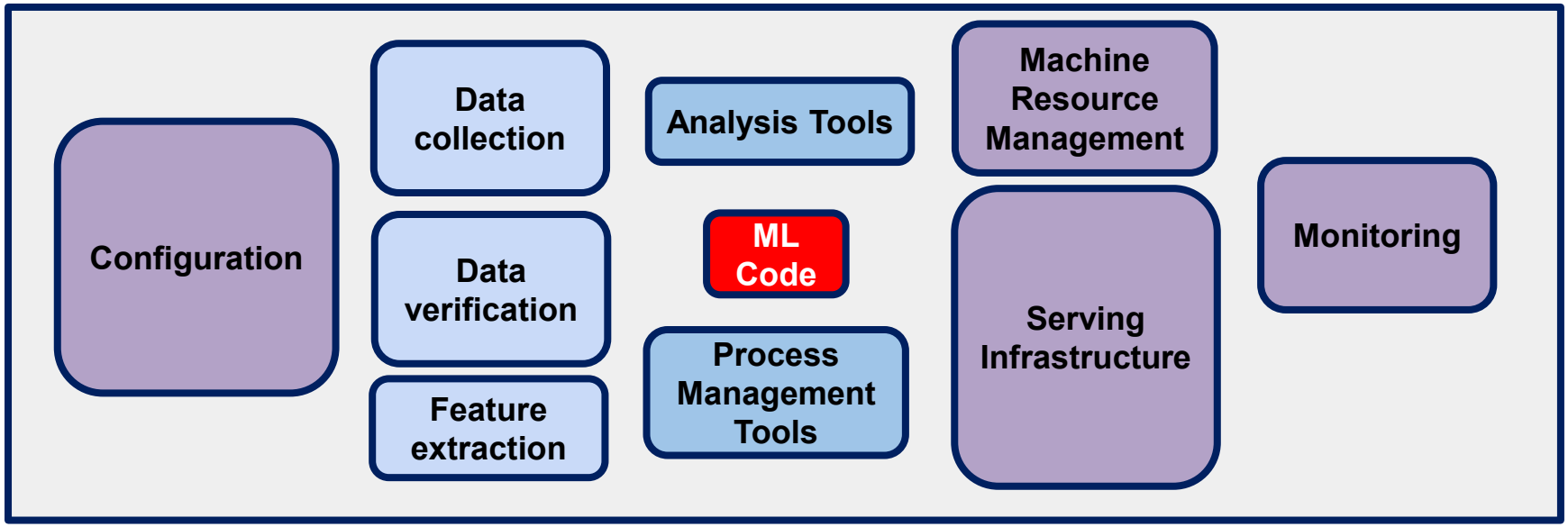
Silicon hungry



- Expertise
- Plethora of frameworks
- Trial & Error
- Data volumes
- Privacy and Security

- Heterogeneity
- Elasticity
- Efficient use of resources: costs, energy, ...

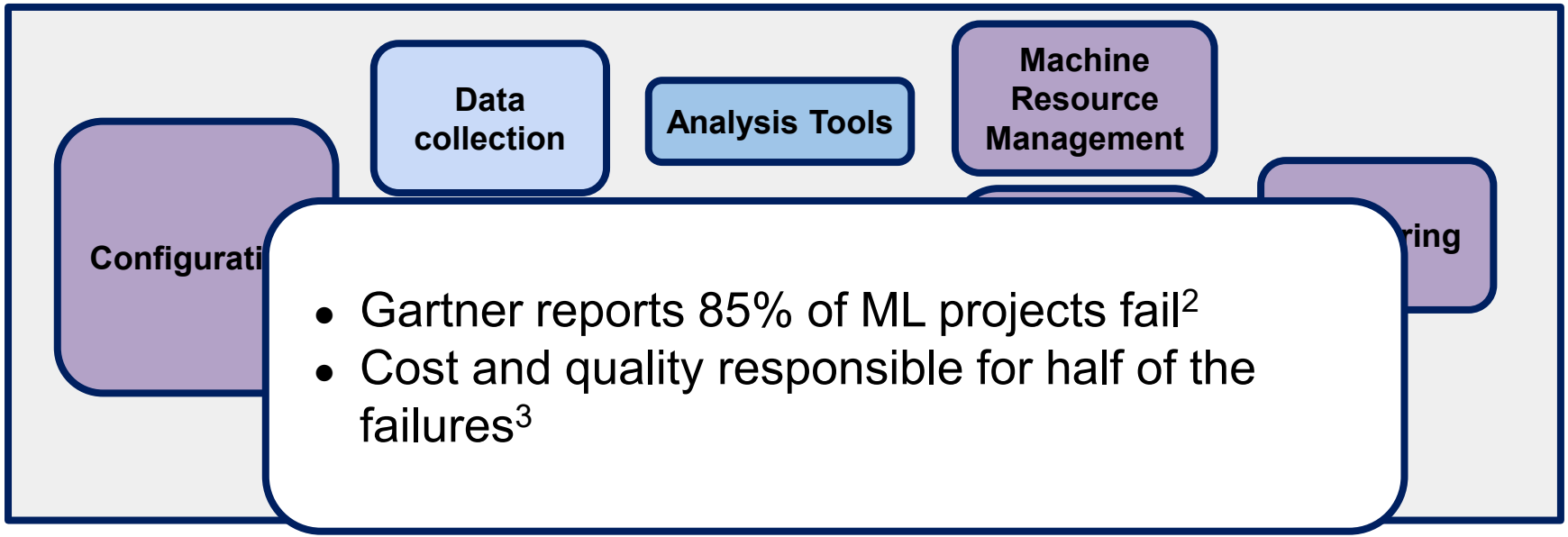
Hardest part of AI isn't AI



Only a small fraction of real world ML systems is composed of the ML code[1]

[1] Hidden Technical Debt in Machine Learning Systems, Google. NIPS 2015

Hardest part of AI isn't AI



Only a small fraction of real world ML systems is composed of the ML code[1]

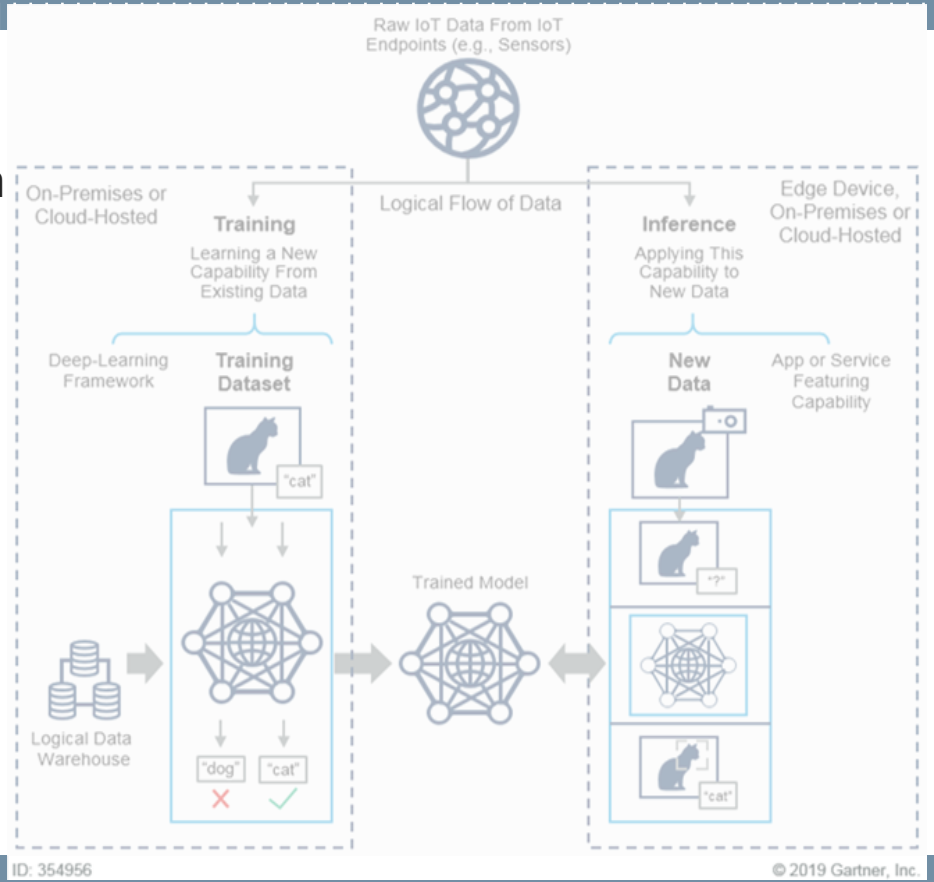
[1] Hidden Technical Debt in Machine Learning Systems, Google. NIPS 2015
[2] Why 85% of Machine Learning Projects Fail – How to Avoid This
[3] Machine Learning Engineering for the Real World, Databricks 2022

What problems the Industry faces?

- Lack of expertise
- Learning curves
- Cloud provider lock-in
- Fast evolving technologies
- Customization

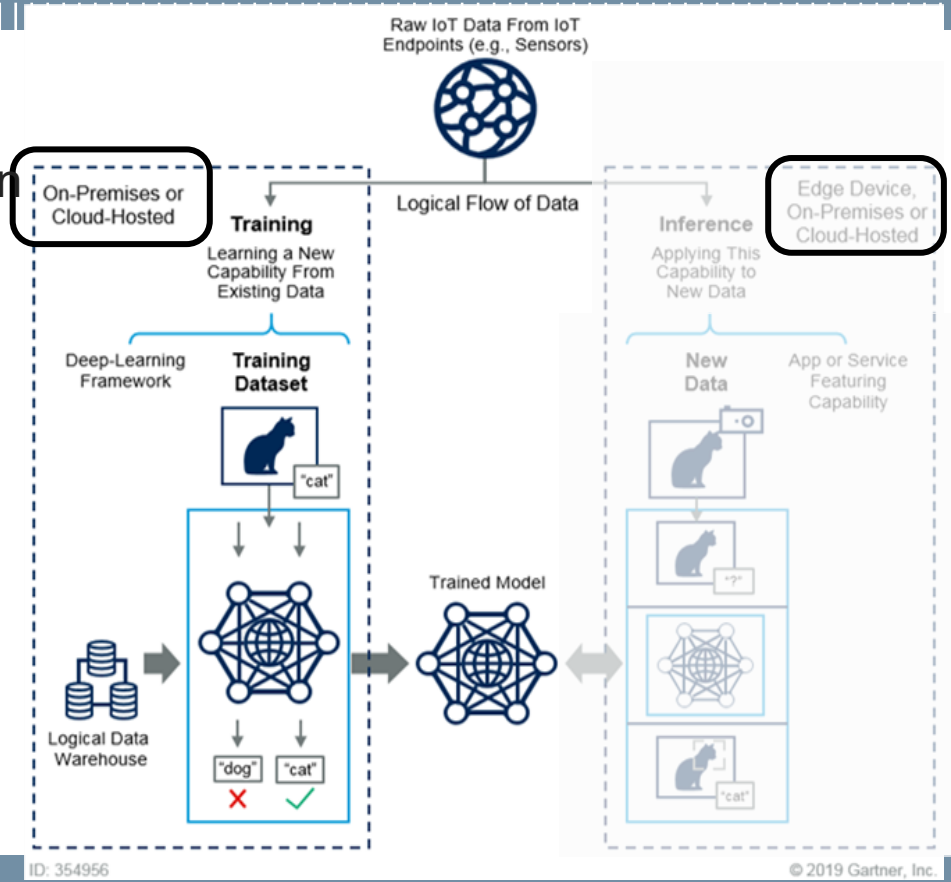
Edge AI the need for a novel approach

- Novel computing continua break tradition AI development paradigms



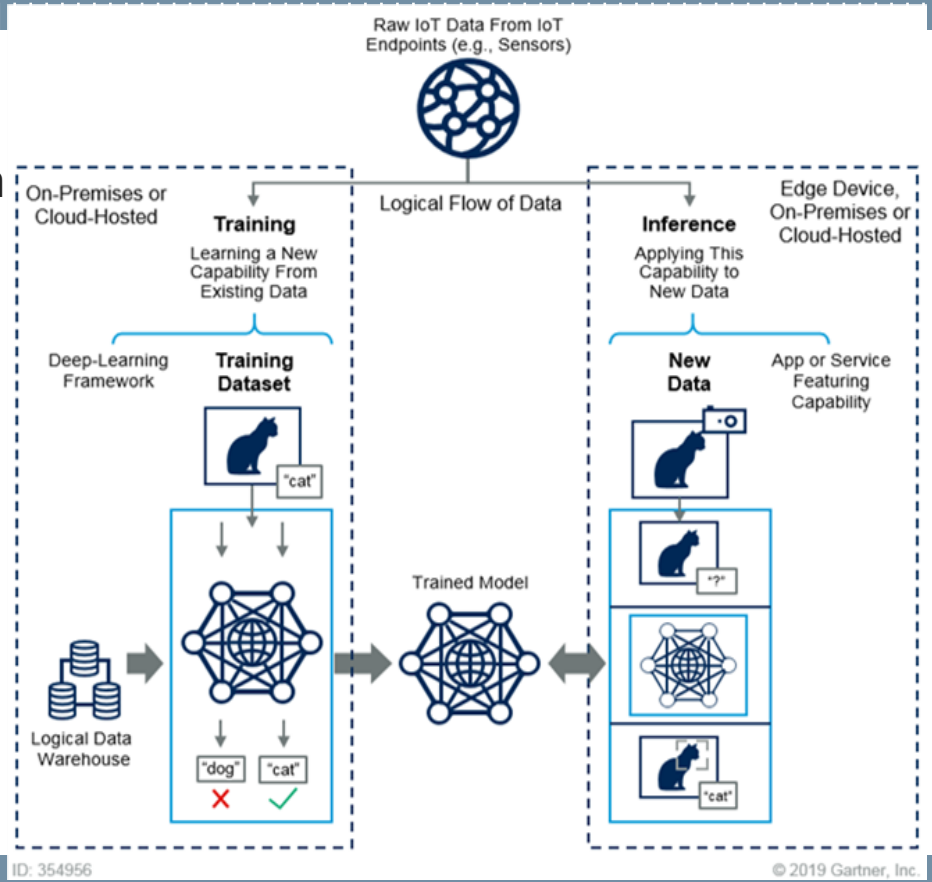
Edge AI the need for a novel approach

- Novel computing continua break tradition AI development paradigms
- AI development beyond the classic
  - Data from IoT
  - Train on the Cloud
  - Inference on the Edge / Cloud
- Engineering cannot be an afterthought

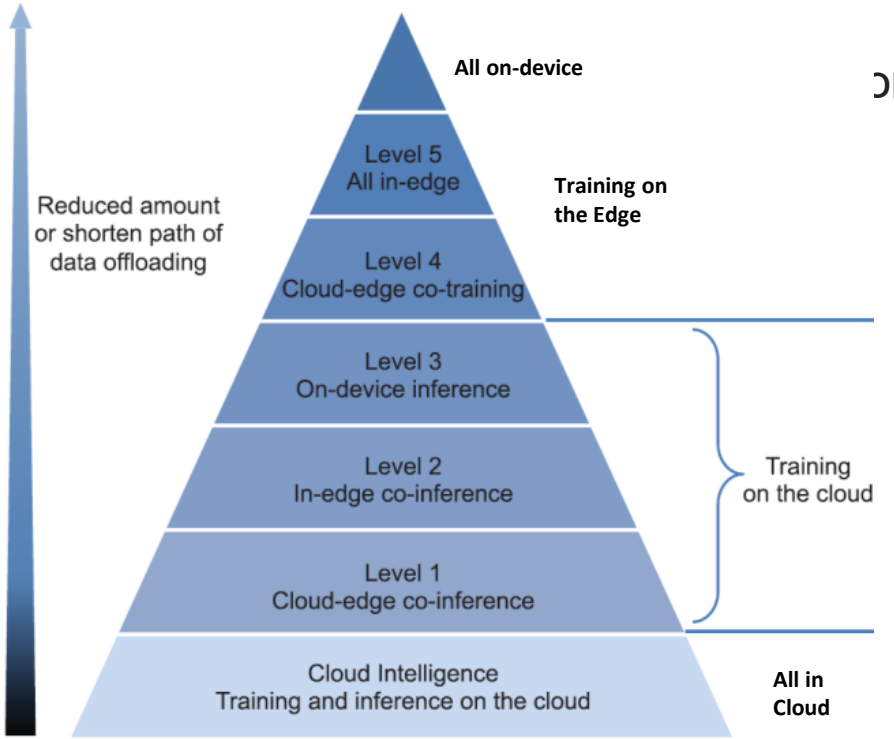


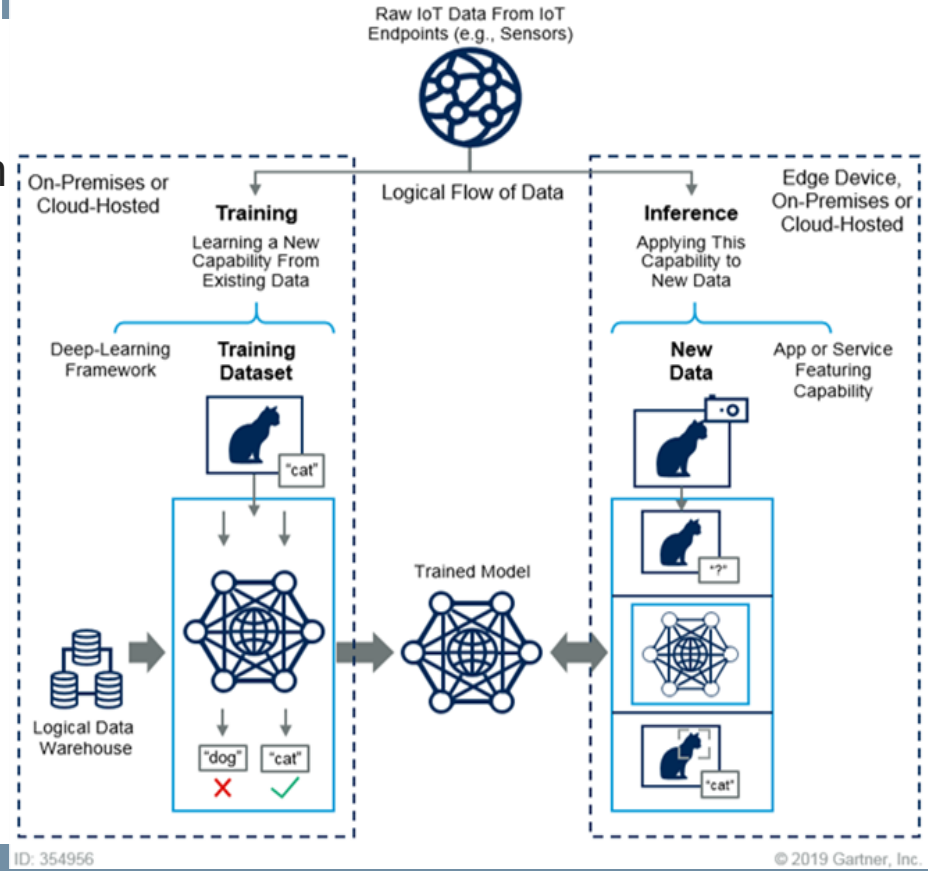
Edge AI the need for a novel approach

- Novel computing continua break tradition AI development paradigms
- AI development beyond the classic
  - Data from IoT
  - Train on the Cloud
  - Inference on the Edge / Cloud
- Engineering cannot be an afterthought



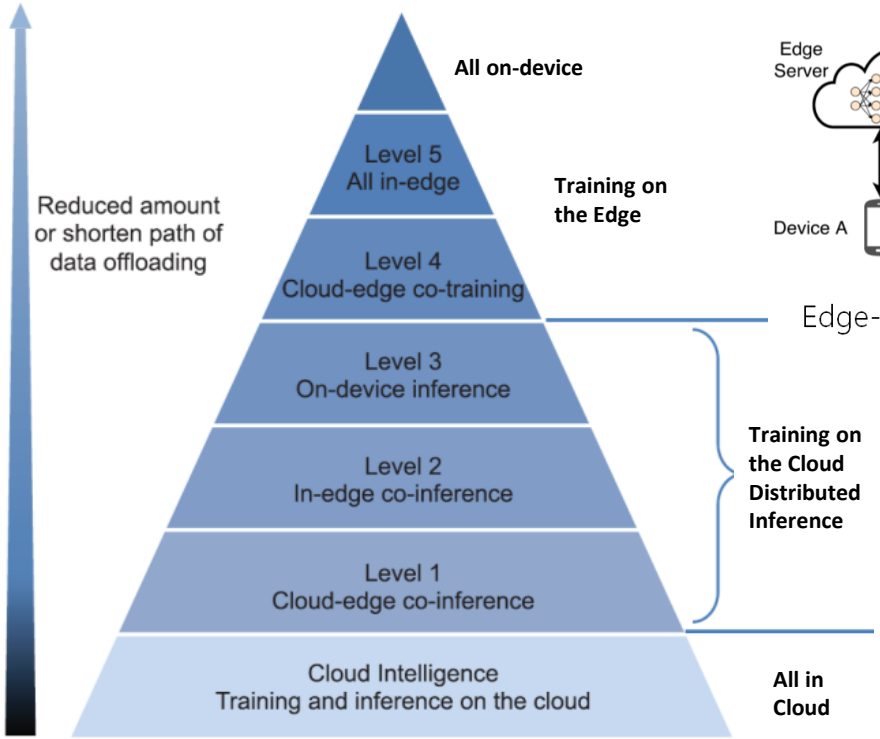
Edge AI the need for a novel approach

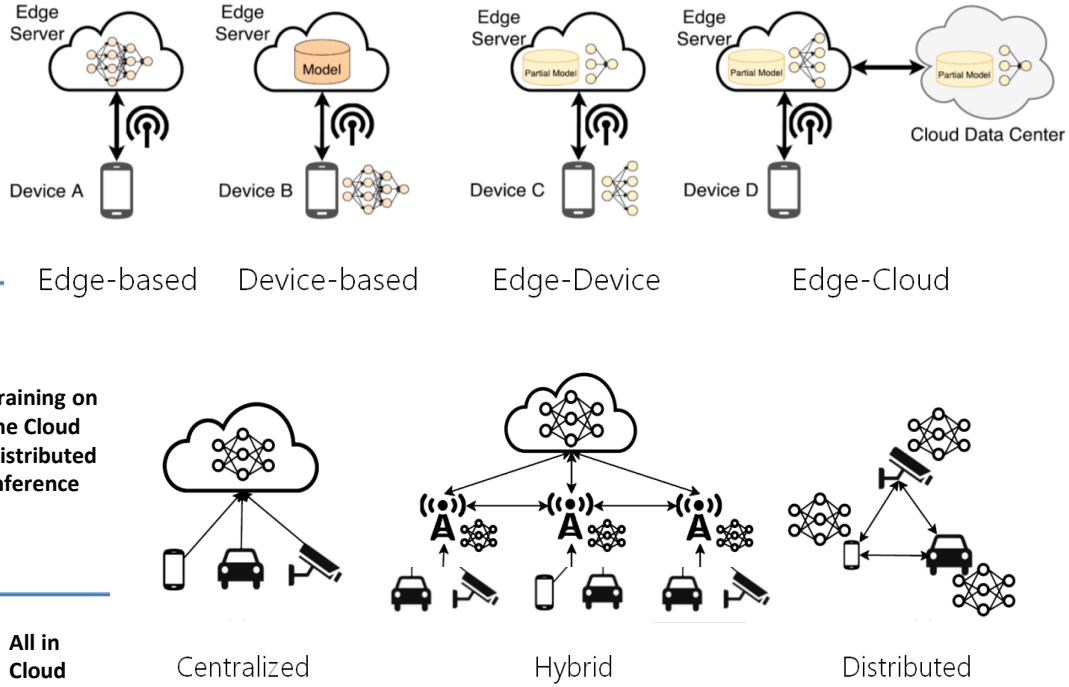




Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Junshan Zhang.
"Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing"

Edge AI the need for a novel approach





Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Junshan Zhang.
"Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing"

Course Overview

Course schedule

Day 1: Edge AI introduction

- Edge computing motivations
- AI@edge challenges
- Course organization
- AI Edge devices
- Moving (intelligent) processing from the Cloud to the edge: challenges & opportunities
- Computational and memory demand of AI solutions
- Tiny Machine Learning and Tiny Deep Learning

Day 2: Technologies for Edge AI

- Introduction to Cloud Computing
- Main IaaS Cloud services (Amazon EC2, Amazon S3)
- Infrastructure as Code (IaC) and DevOps
- Ansible, Docker, Kubernetes

Thanks for your attention…

…any questions?

danilo.ardagna@polimi.it