POLITECNICO DI MILANO

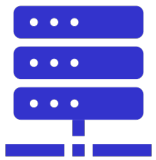
Computing Infrastructures





Overview of Computing Infrastructures

The topics of the course: what are we going to see today?



HW Infrastructures:

System-level: Computing Infrastructures and Data Center Architectures, Rack/Structure;

Node-level: Server (computation, HW accelerators), Storage (Type, technology), Networking (architecture and technology);

Building-level: Cooling systems, power supply, failure recovery

SW Infrastructures:

Virtualization: Process/System VM, Virtualization Mechanisms (Hypervisor, Para/Full virtualization)

Computing Architectures: Cloud Computing (types, characteristics), Edge/Fog Computing, X-as-a service

Methods:

Reliability and availability of datacenters (definition, fundamental laws, RBDs)

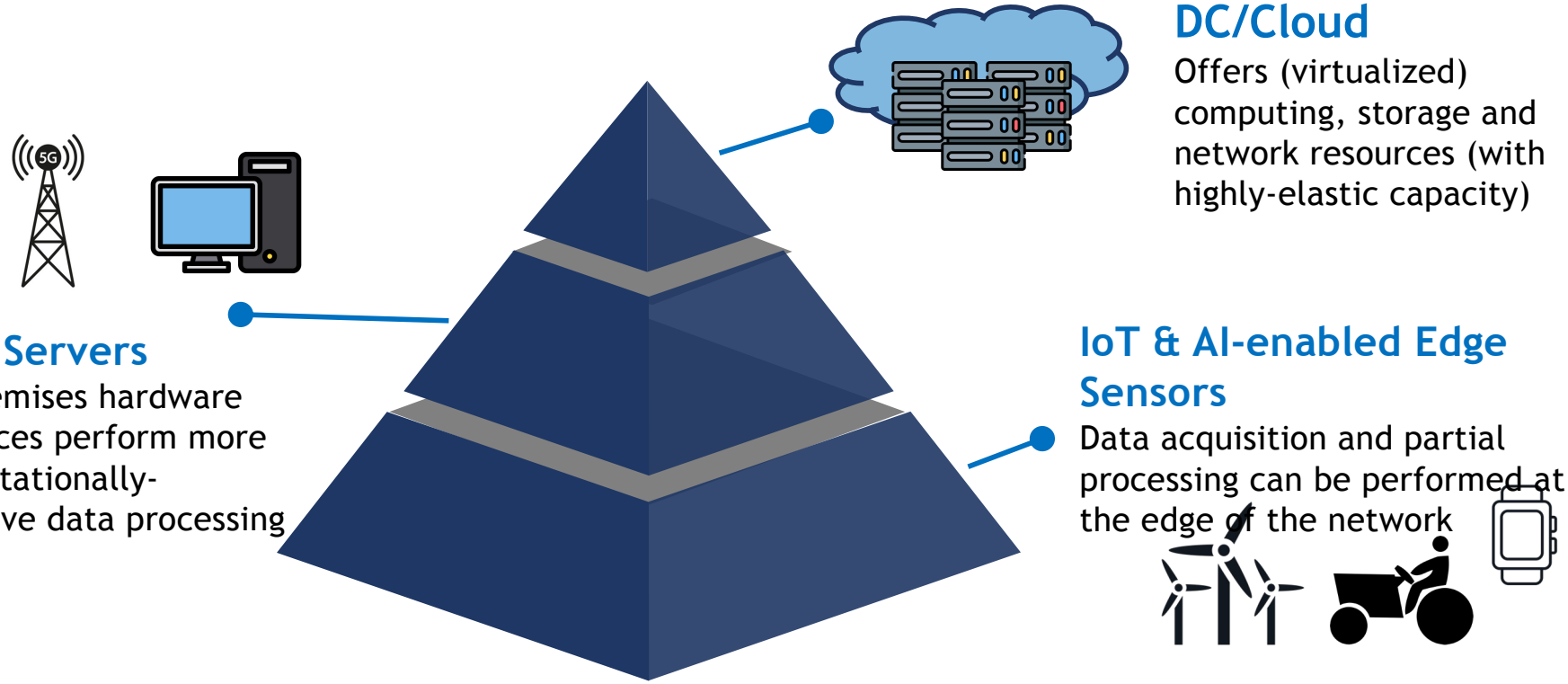
Disk performance (Type, Performance, RAID)

Scalability and performance of datacenters (definitions, fundamental laws, queuing network theory)

What is a computing infrastructure?

Technological infrastructure that provides hardware and software for computation to other systems and services

Not a unique Computing Infrastructure



DC/Cloud

Offers (virtualized) computing, storage and network resources (with highly-elastic capacity)

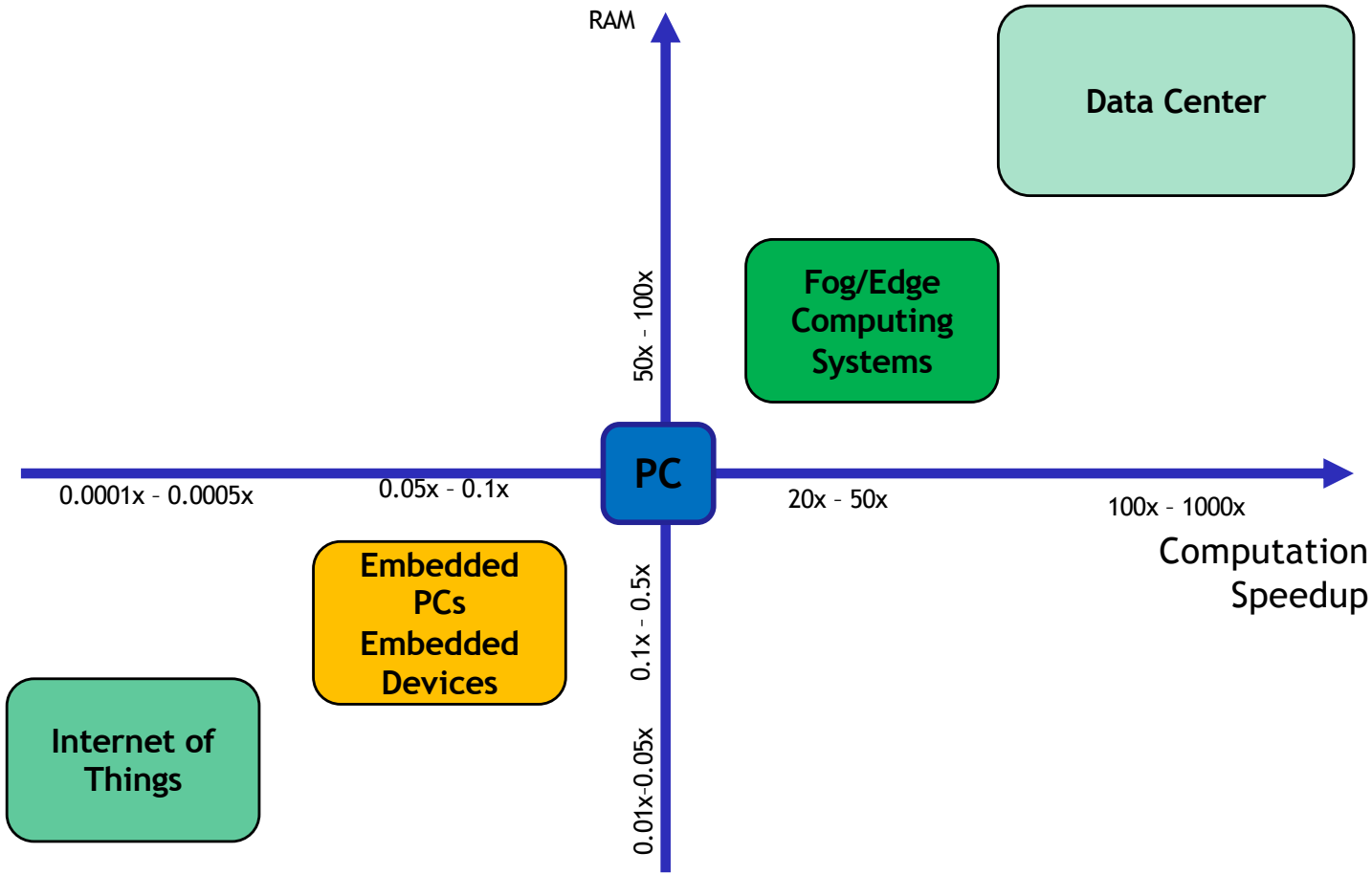
Edge Servers

On-premises hardware resources perform more computationally-intensive data processing

IoT & AI-enabled Edge Sensors

Data acquisition and partial processing can be performed at the edge of the network

Examples of Computing Infrastructures



Computing Continuum

- **Definition:** The computing continuum models a distributed environment incorporating endpoints, edge, and cloud processing elements.
- **Data Growth from Connectivity:** Connected devices exponentially increase data generation, necessitating innovative solutions for efficient processing and analysis.
- **Challenges for Centralized Systems:** Traditional centralized data centers struggle with scalability and latency (real-time processing), hindering performance in resource-intensive applications.

The Computing Continuum

IoT/Edge — Fog — HPC/Cloud/Instrument

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	Adafruit Trinket	Particle.io Boron	Array of Things	Linux Box	Co-located Blades	1000-node cluster	Datacenter
Memory	0.5K	256K	8GB	32GB	256G	32TB	16PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10GigE	40GigE	N*100GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M




Count = 10^9
Size = 10^1

Count = 10^1
Size = 10^9

Computing Continuum

Endpoints include sensors and IoT devices, crucial for **real-time data collection in distributed environments.**

Edge/Fog infrastructure utilizes local servers and gateways to process data near its source, **enhancing responsiveness.**

Data centers provide remote, **large-scale processing** capabilities essential for handling **vast data volumes.**

The Computing Continuum

IoT/Edge → Fog → HPC/Cloud/Instrument

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	Adafruit Trinket	Particle.io Boron	Array of Things	Linux Box	Co-located Blades	1000-node cluster	Datacenter
Memory	0.5K	256K	8GB	32GB	256G	32TB	16PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10GigE	40GigE	N*100GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M

Count = 10^9
Size = 10^1




Count = 10^1
Size = 10^9

Challenges and Technological Innovations

Challenges for computing Continuum

- **Decentralized Processing Shift** optimizeing local intelligence, reducing latency and enhancing efficiency across networks.
- **Hybrid Architecture Integration** merging edge and cloud solutions, maximizing resource utilization and ensuring flexibility in data management.
- **IoT Data Volume** necessitating innovative distributed strategies for effective data handling.

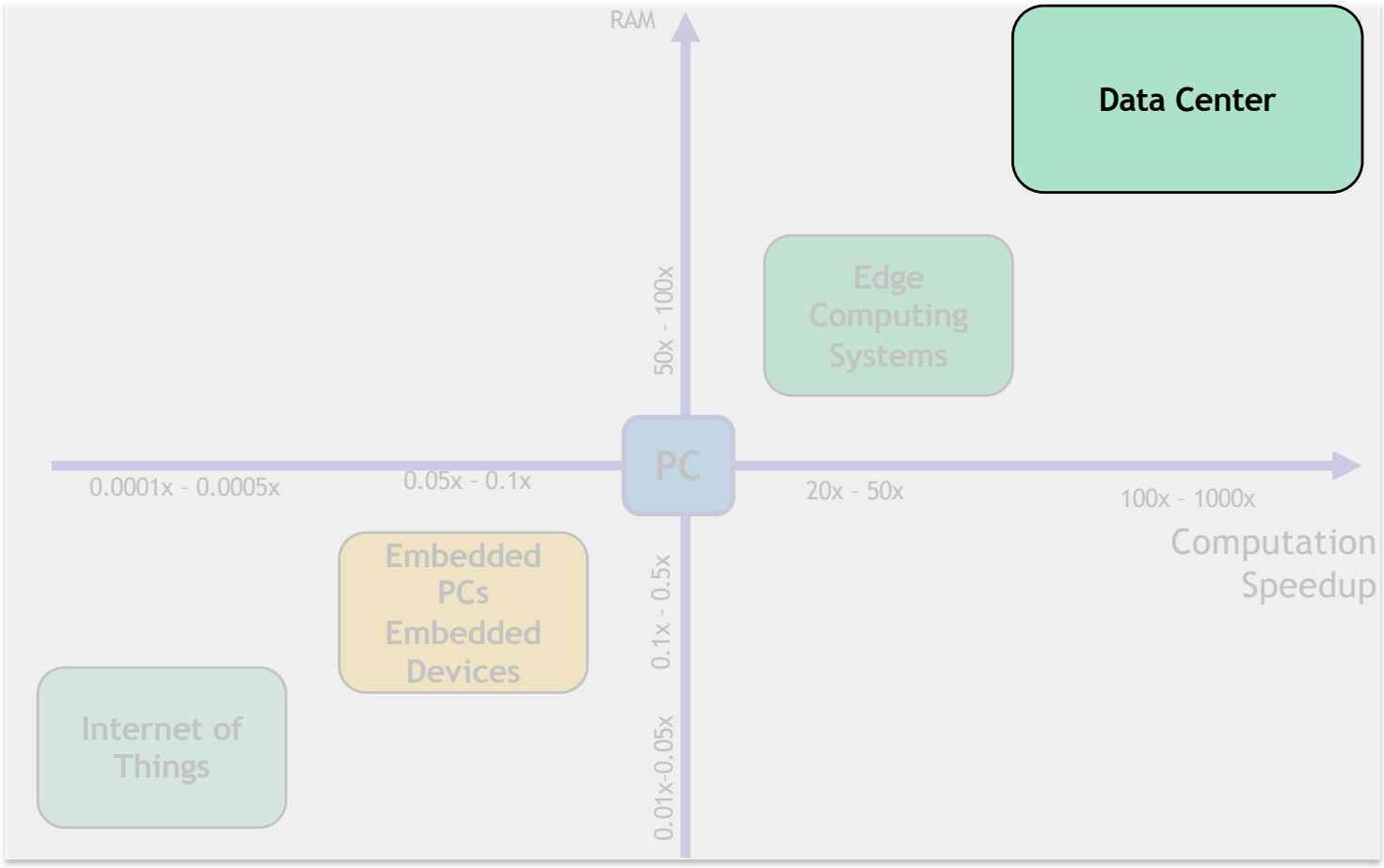
Technological Innovations Driving the Continuum

- **IoT Innovations:** Recent IoT advancements improve data collection, facilitating seamless integration within the computing continuum framework significantly.
- **Edge AI Capabilities:** Deploying edge AI enhances processing capabilities, allowing real-time analytics and decision-making closer to data sources.
- **5G Network Impact:** 5G networks drastically reduce latency, enabling faster data transfer and connectivity for numerous devices simultaneously.

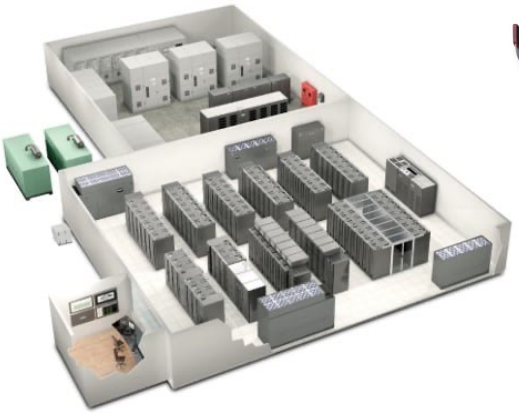
Real World Examples:

- Industrial IoT (e.g. machine monitoring), Smart City management (e.g. traffic management), Connected Healthcare (e.g. patient monitoring)

Examples of Computing Infrastructures



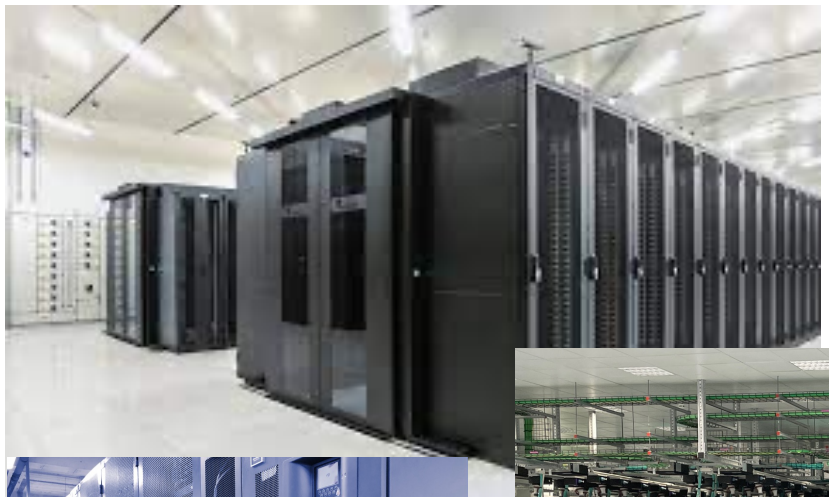
Data Centers: a technological perspective



Server for Processing

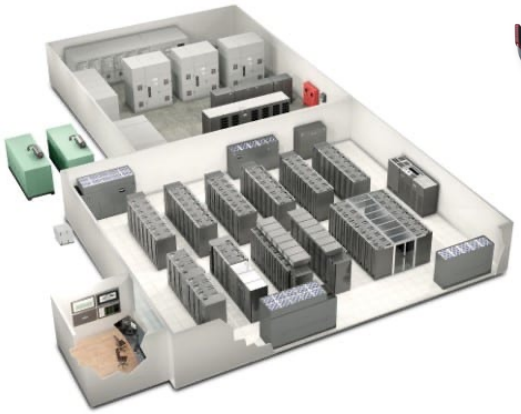
Server for Storage

Server for Communication



DataCloud
Analytics for Knowledge
POLITECNICO MILANO 1863

Data Centers: a technological perspective

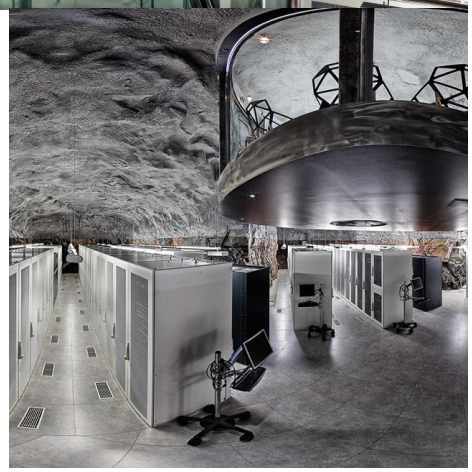


Server for Processing

Server for Storage

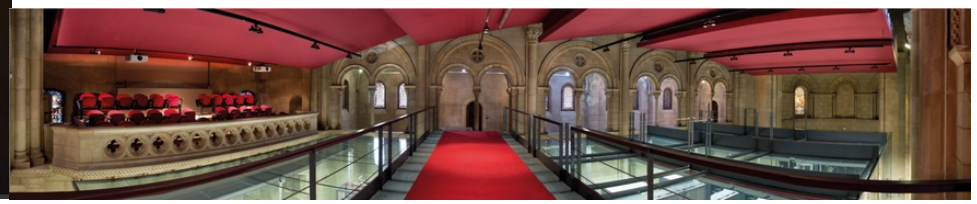
Server for Communication

The Pionen White Mountains is a Swedish data center. This center is located in Stockholm.



Nice Book: The art of the data center – Douglas Alger, 2012

Barcellona Supercomputing Center Torre Girona Chapel



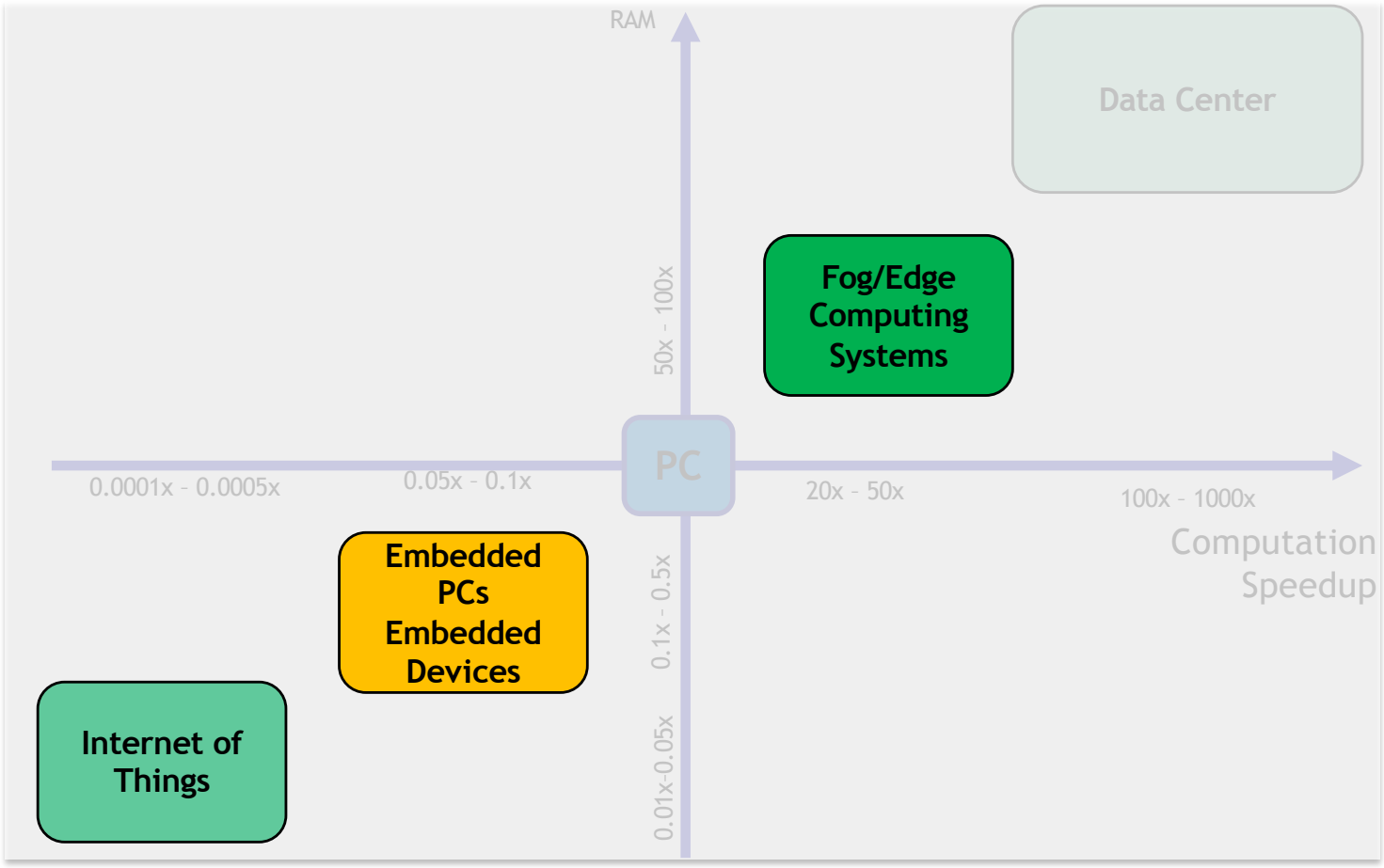
… Not always nice places

Data Centers: advantages and disadvantages

- ✓ Lower IT costs
- ✓ High performance
- ✓ Instant software updates
- ✓ “Unlimited” storage capacity
- ✓ Increased data reliability
- ✓ Universal data access
- ✓ Device Independence

- ➢ Require a constant Internet connection
- ➢ Do not work well with low-speed connections
- ➢ Hardware Features might be limited
- ➢ Privacy and security issues
- ➢ High power Consumption
- ➢ Latency in taking decision

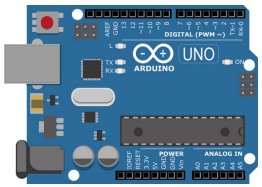
Edge Computing, PC Embedded and IoT



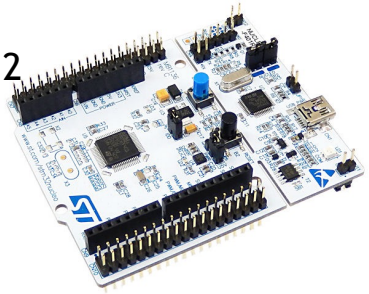
Internet-of-Things

- An Internet of Things (IoT) device is any everyday object embedded with sensors, software, and internet connectivity.
- This allows to collect and exchange data with other devices and systems, typically over the internet, with limited need of process ad store data

Arduino

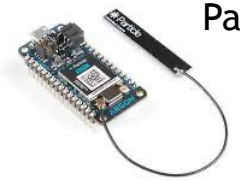


STM32



ESP32

Particle Argon



	Arduino Nano	Nano ESP32-S3	Nano 33 IoT	Nano 33 BLE	Nano 33 BLE Sense Rev2	Nano RP2040 Connect
Microcontroller	ATmega328	ESP32-S3	SAMD21 Arm® Cortex®-M0+ / u-blox® NINA-W102	nRF52840 / u-blox® NINA-B306	nRF5284 / u-blox® NINA-B306	Raspberry Pi RP2040 / u-blox® NINA-W102
USB connector	Mini-B USB	USB-C®	Micro USB	Micro USB	Micro USB	Micro USB
I/O voltage	5 V	3.3 V	3.3 V	3.3 V	3.3 V	3.3 V
Input range	7-12 V	5-21 V	7-12 V	7-12 V	7-12 V	5-21 V
Clock speed	16 MHz	up to 240 MHz	48 MHz	64 MHz	64 MHz	133 MHz
SRAM	2 kB	512 kB	256 kB	256 kB	256 kB	264 kB
Flash	32 kB	16 MB	1 MB	1 MB	1 MB	16 MB

- ✓ Highly Pervasive
- ✓ Wireless connection
- ✓ Battery Powered
- ✓ Low costs
- ✓ Sensing and actuating

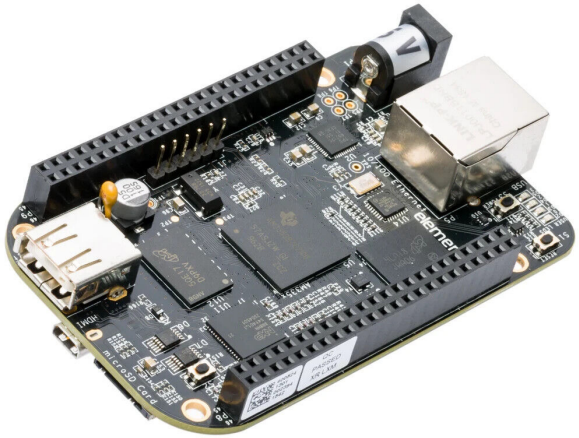
- ➢ Low computing ability
- ➢ Constraints on energy
- ➢ Constraints on memory (RAM/FLASH)
- ➢ Difficulties in programming

Embedded PCs

Raspberry Pi 5 8GB Quad-Core ARMA76 (64 Bits - 2,4 GHz)

Nvidia Jetson Orin™ Nano 8 GB 6 x 1.5 GHz

- ✓ Pervasive computing
- ✓ High performance unit
- ✓ Availability of development boards
- ✓ Programmed as PCs
- ✓ Large community



BeagleBone Black - CPU ARM Cortex A8 1GHz - 512MB RAM

Google Coral Dev Board, 1 GB RAM

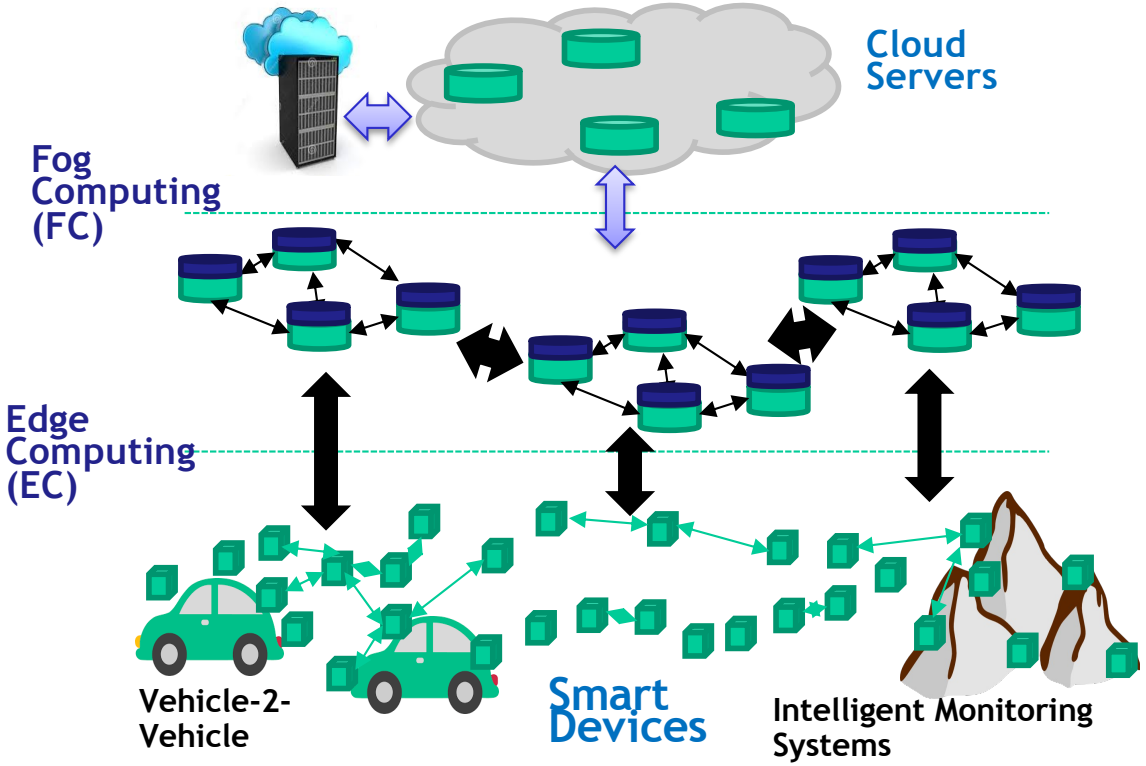
- ➢ Pretty high power consumption
- ➢ (Some) HW design has to be done

Edge/Fog Computing Systems

The key *difference between fog computing* and edge computing is associated with the location **where the data is processed**

- In **edge computing**, the data is processed closest to the sensors
  - *component producing the data*
- In **fog computing**, the computing is moved to processors linked to a local area network (IoT gateway).

Edge computing places the intelligence in the connected devices themselves, whereas, fog computing puts it in the local area network



Feature	Edge Computing	Fog Computing
Location	Directly on device or nearby device	Intermediary devices between edge and cloud
Processing Power	Limited due to device constraints, sending data to a central server for analysis	More powerful than edge devices. However, sending data to a central server for analysis
Primary Function	Real-time decision-making, low latency. However, central server analyzing combined data and sending only relevant information further	Pre-process and aggregate data, reduce bandwidth usage. However, central server analyzing combined data and sending only relevant information further
Advantages	Low latency, reduced reliance on cloud, security for sensitive data	Bandwidth efficiency, lower cloud costs, complex analysis capabilities
Disadvantages	Limited processing power, single device focus	Increased complexity, additional infrastructure cost

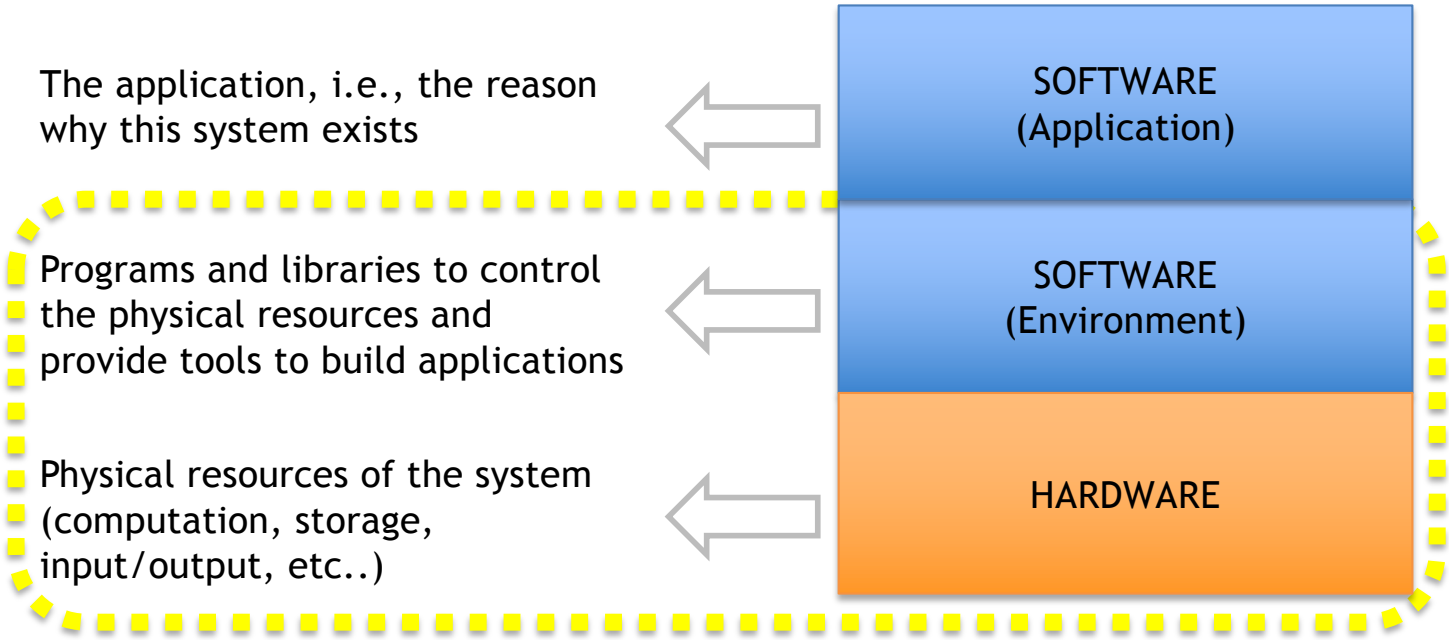
Fog/Edge Examples

MODELS	WORKLOADS	RACK UNITS	PROCESSOR	MAX MEMORY	MAX STORAGE	GPU
 PowerEdge XE9680 Open 3D viewer	AI and machine learning, DL large data set training, HPC, CRISP, and healthcare	6U	2 x 5th Generation Intel® Xeon® Scalable processors	32 x 128 GB DDR5 (4 TB)	16 x E3.S or 8 x 2.5" SAS/SATA/NVMe (122.88 TB)	8 x (NVIDIA A100, H100, H200, or AMD MI300X)
 PowerEdge XE9640 Open 3D viewer	AI, machine learning, DL training modeling and simulation, HPC, and healthcare	2U	2 x 5th Generation Intel® Xeon® Scalable processors	32 x 128 GB DDR5 (4 TB)	4 x 2.5" NVMe (61.44 TB)	4 x NVIDIA H100
 PowerEdge XE8640 Open 3D viewer	AI, machine learning, DL medium data set training,HPC, finance, and research	4U	2 x 5th Generation Intel® Xeon® Scalable processors	32 x 128 GB DDR5 (4 TB)	8 x E3.S or 8 x 2.5" SAS/SATA/NVMe (122.88 TB)	4 x NVIDIA H100
 PowerEdge R760xa Open 3D viewer	AI, machine learning, DL training and inferencing, HPC, and render farms and virtualisation	2U	2 x 5th Generation Intel® Xeon® Scalable processors	32 x 256 GB DDR5 (8 TB)	8 x 2.5" SAS/SATA/NVMe (122.88 TB)	4 x 400W (DW) or 12 x 75W (SW)

- ✓ High computational capacity
- ✓ Distributed computing
- ✓ Privacy and security
- ✓ Reduced Latency in making a decision

- ➢ Require a power connection
- ➢ Require connection with the Cloud

An IT perspective for Computing Infrastructures



An IT perspective for Computing Infrastructures