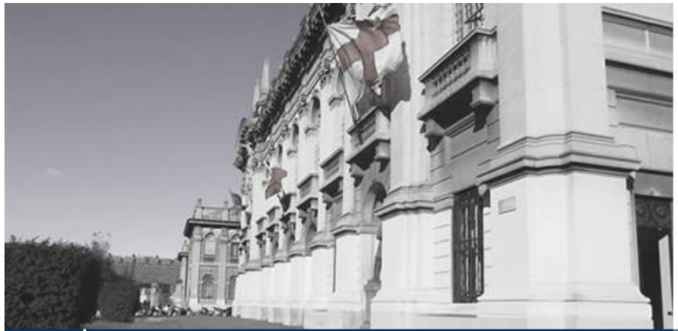


Computing Infrastructures

POLITECNICO DI MILANO

1) Computing Infrastructures

Prof. Danilo Ardagna

Credits: Prof. Manuel Roveri

The topics of the course

A. 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, fa

Feb and mid March

B. SW Infrastructures:

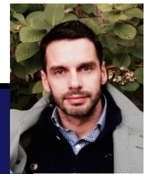
- **Virtualization**: Process/System VM, Virtualization Mechanisms (Hypervisor, Para/Full virtualization, Containers)
- **Computing Architectures**: Cloud Computing (types, characteristic X-as-a service, Edge/Fog Computing
- **Machine and deep learning-as-a-service**

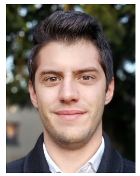
End of May

C. Methods:

- **Reliability and availability of datacenters** (definiti laws, RBDs)
- **Disk performance** (Type, Performance, RAID)
- **Scalability and performance of datacenters** (de fundamental laws, queuing network theory)

Mid March to April



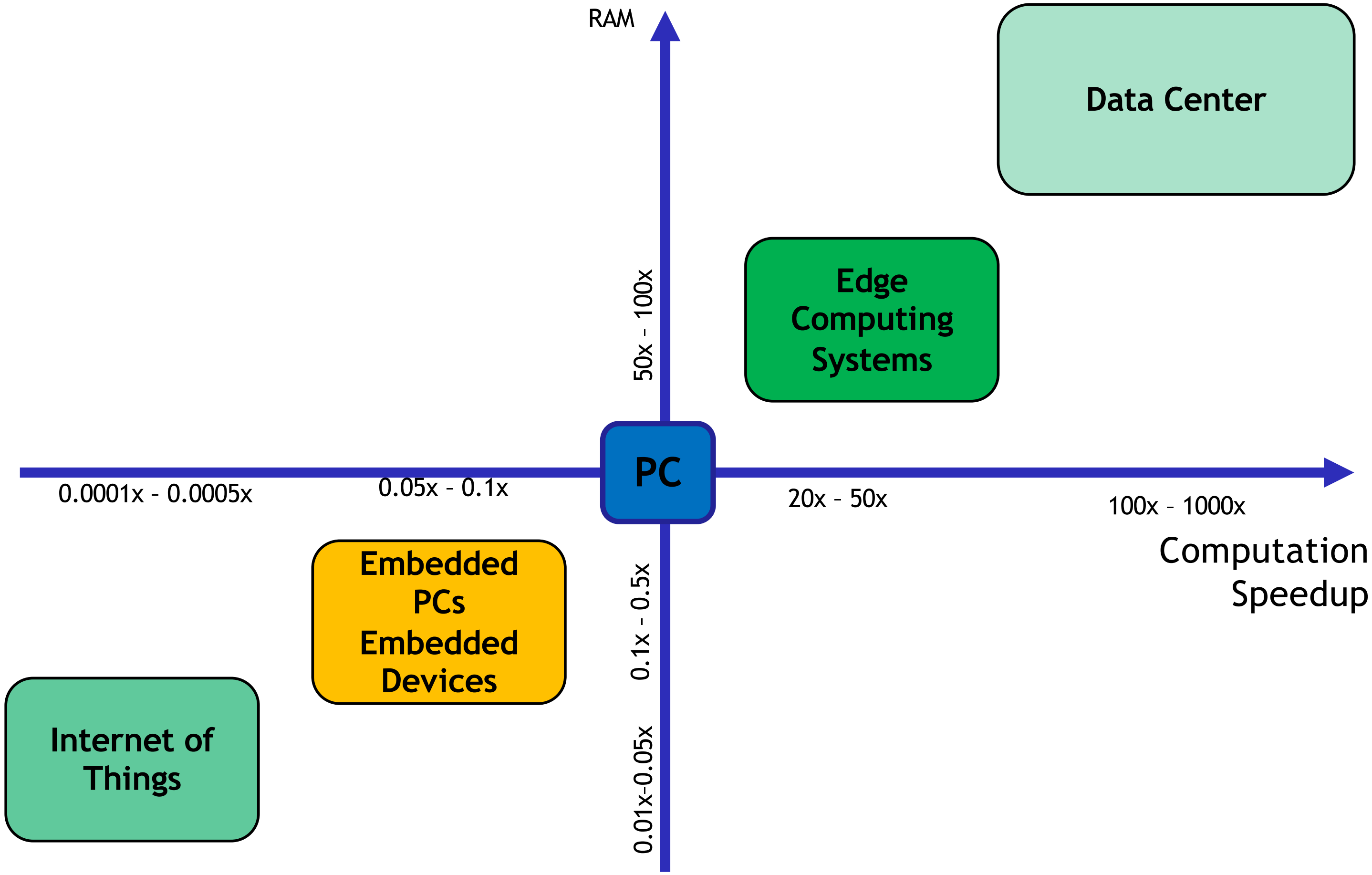


Mid April to mid May

What is a computing infrastructure?

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

Examples of Computing Infrastructures



Examples of Computing Infrastructures

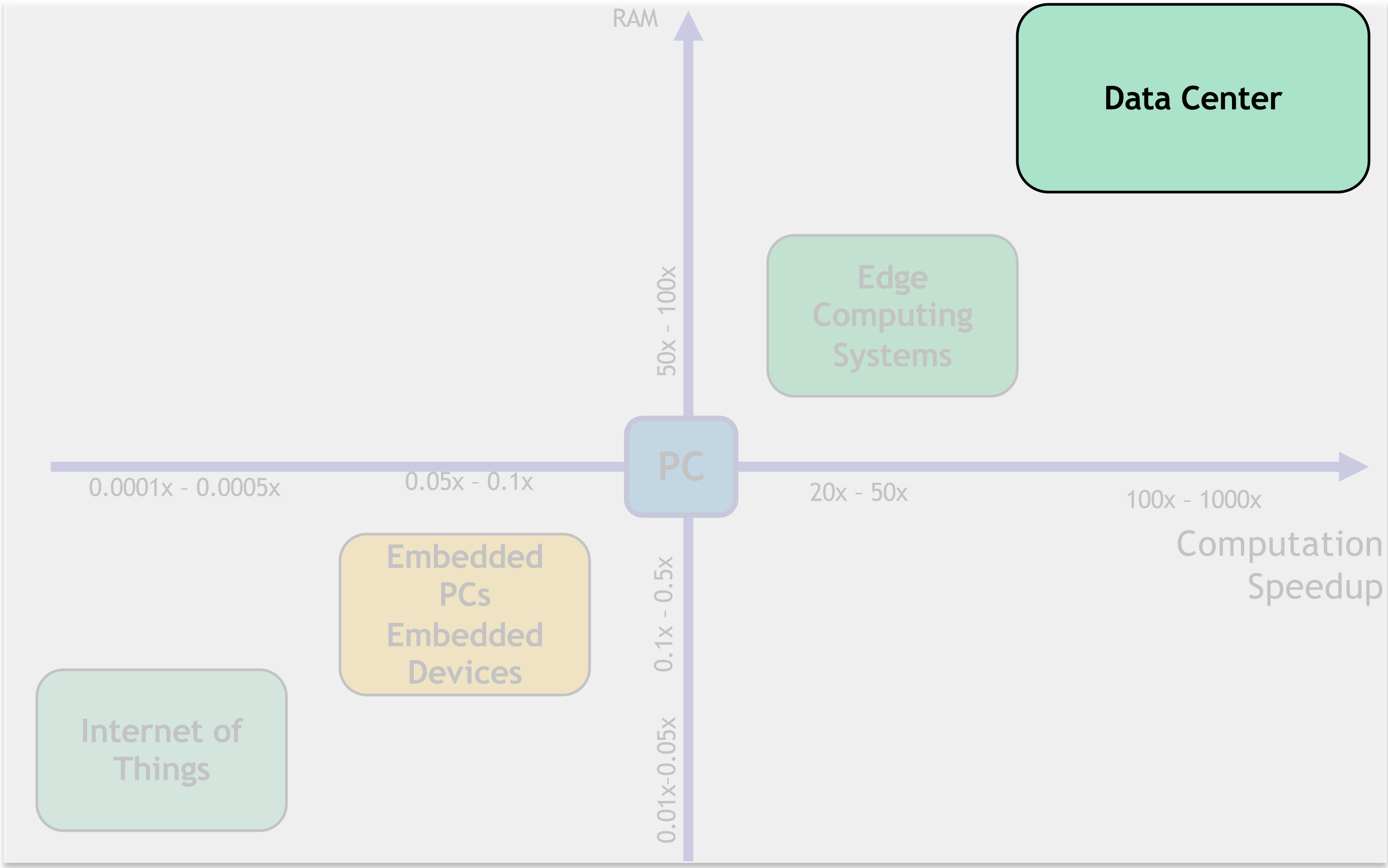
IoT/Edge — Fog — HPC/Cloud/Instrument

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

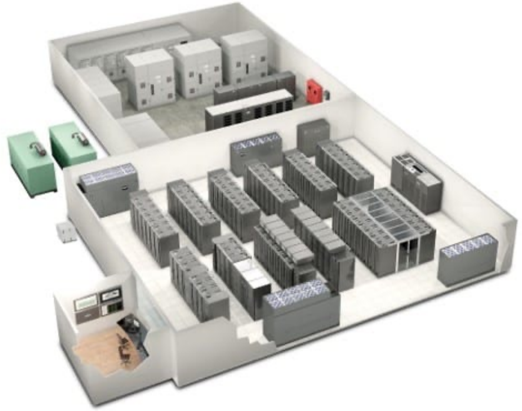
Count = 10^9
Size = 10^1

Count = 10^1
Size = 10^9

Examples of Computing Infrastructures



Data Centers: a technological perspective

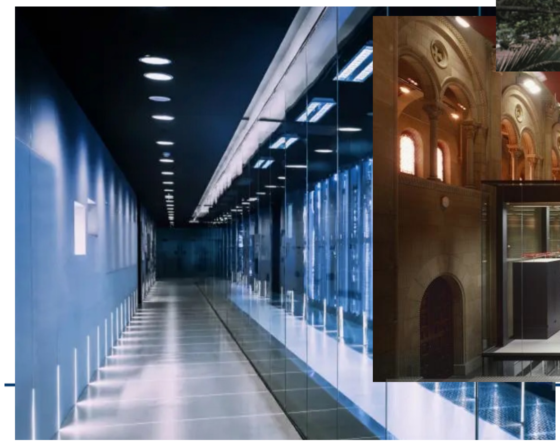


Server for Processing

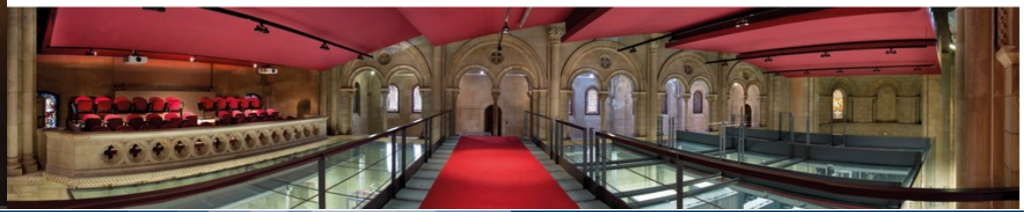
Server for Storage

Server for Communication

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



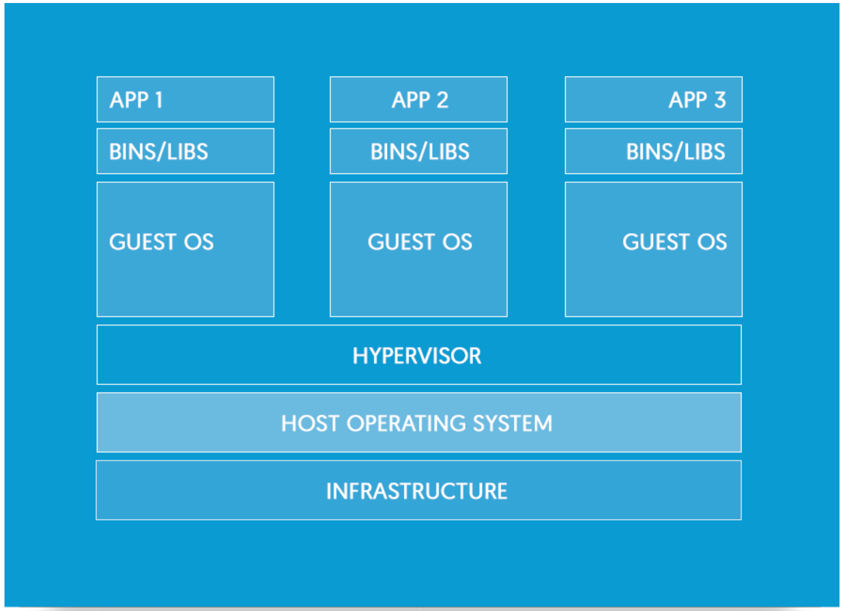
Barcellona
Supercomputing Center



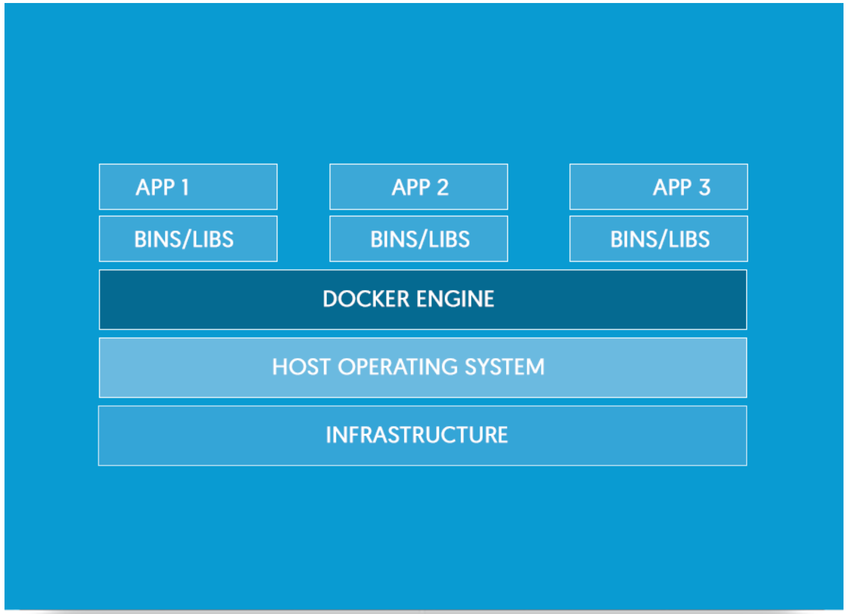
… Not always nice places

Virtual Machines and Containers

VMMs
Provide the full stack (OS, LIB, APP).
Applications depend on guest OS.



Containers
Applications are packaged with all their dependencies into a standardized unit for software development/deployment.



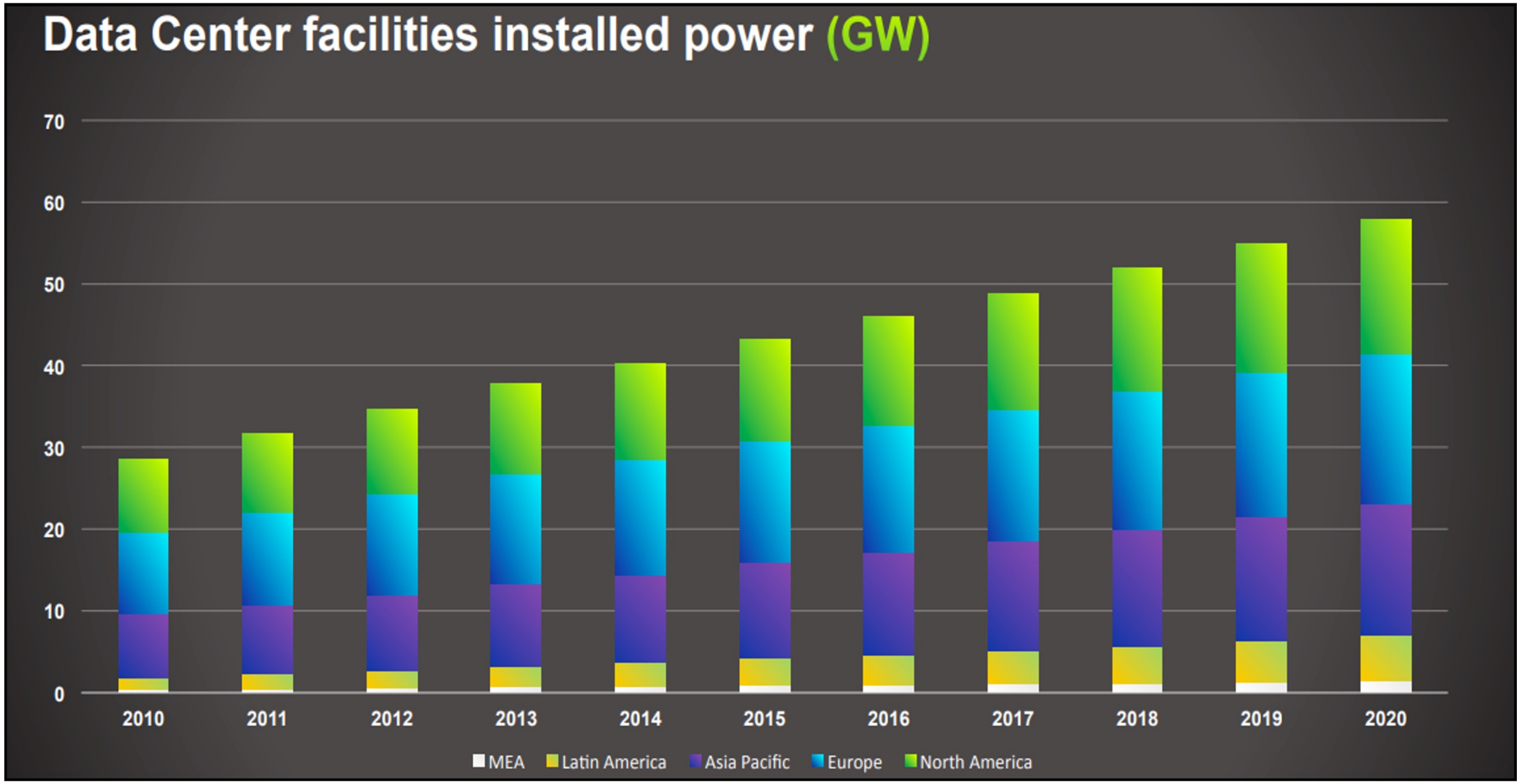
Data Centers: advantages and disadvantages

- ✓ Lower IT costs
- ✓ High performance
- ✓ Instant software updates
- ✓ "Unlimited" storage capacity
- ✓ Increased data reliability
- ✓ Universal document 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
- ➢ <u>Latency in making decision</u>

Data Centers: advantages and disadvantages

INSTALLED POWER



Data Centers: advantages and disadvantages

WATER

Waterlogged

A midsize data center uses roughly as much water as about 100 acres of almond trees or three average hospitals, and more than two 18-hole golf courses.

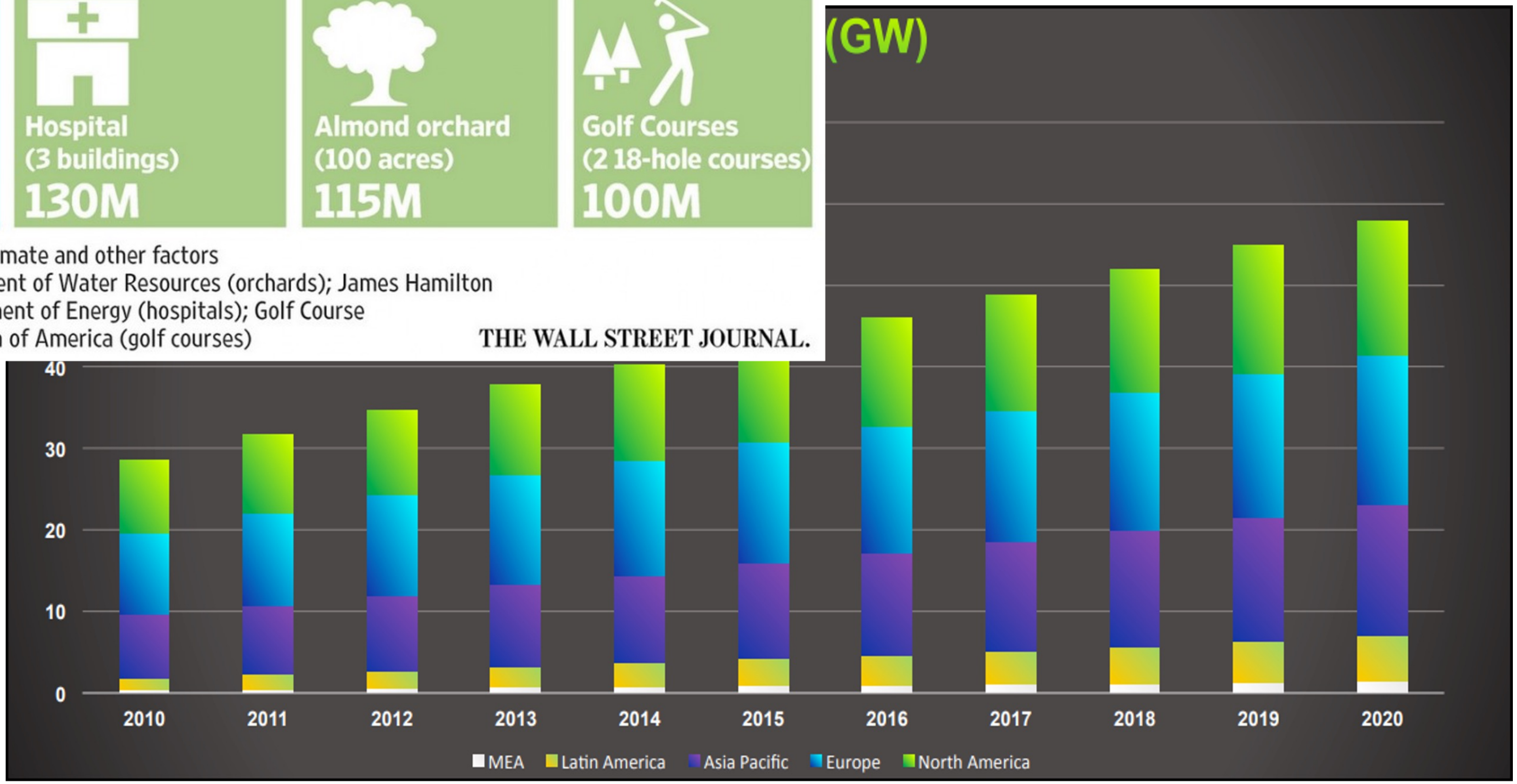
Approximate annual water usage, in gallons*

Data Center (15 megawatt)	Hospital (3 buildings)	Almond orchard (100 acres)	Golf Courses (2 18-hole courses)
130M	130M	115M	100M

*Use varies depending on climate and other factors
Sources: California Department of Water Resources (orchards); James Hamilton (data centers); U.S. Department of Energy (hospitals); Golf Course Superintendents Association of America (golf courses)

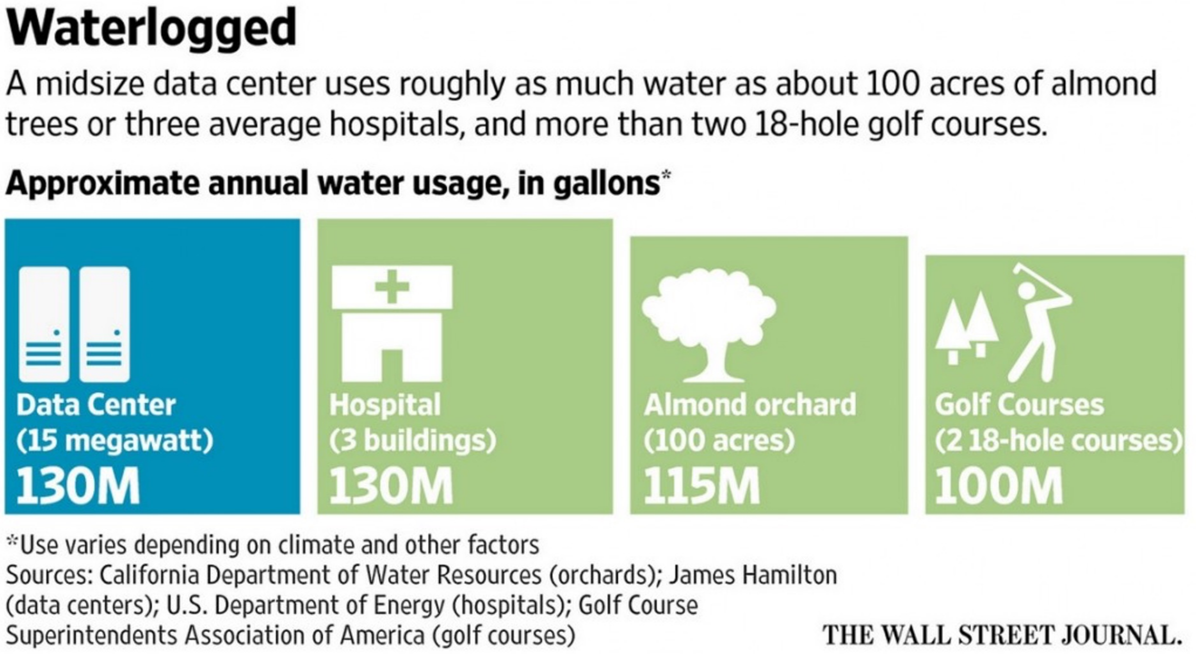
THE WALL STREET JOURNAL.

ED POWER

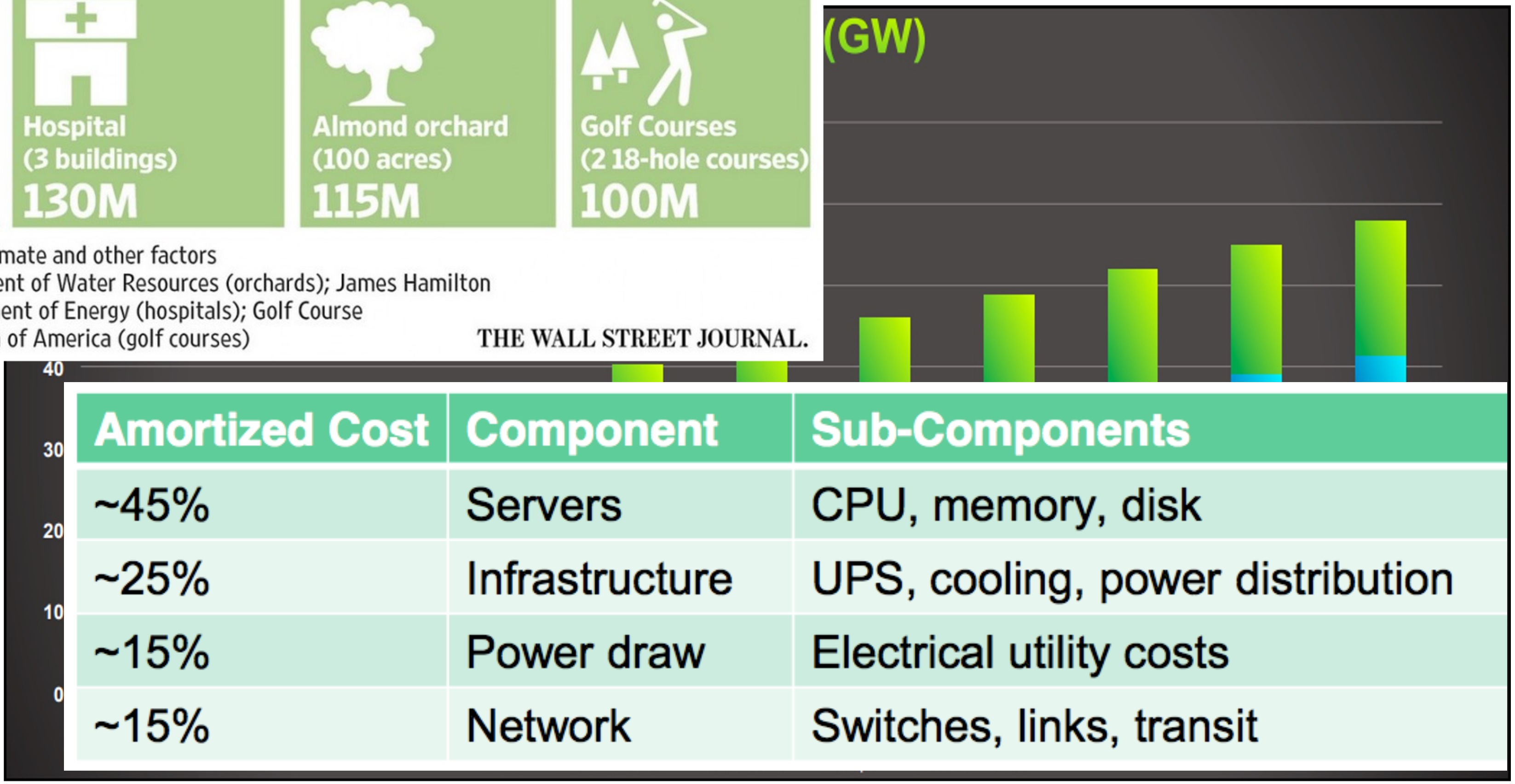


Data Centers: advantages and disadvantages

WATER

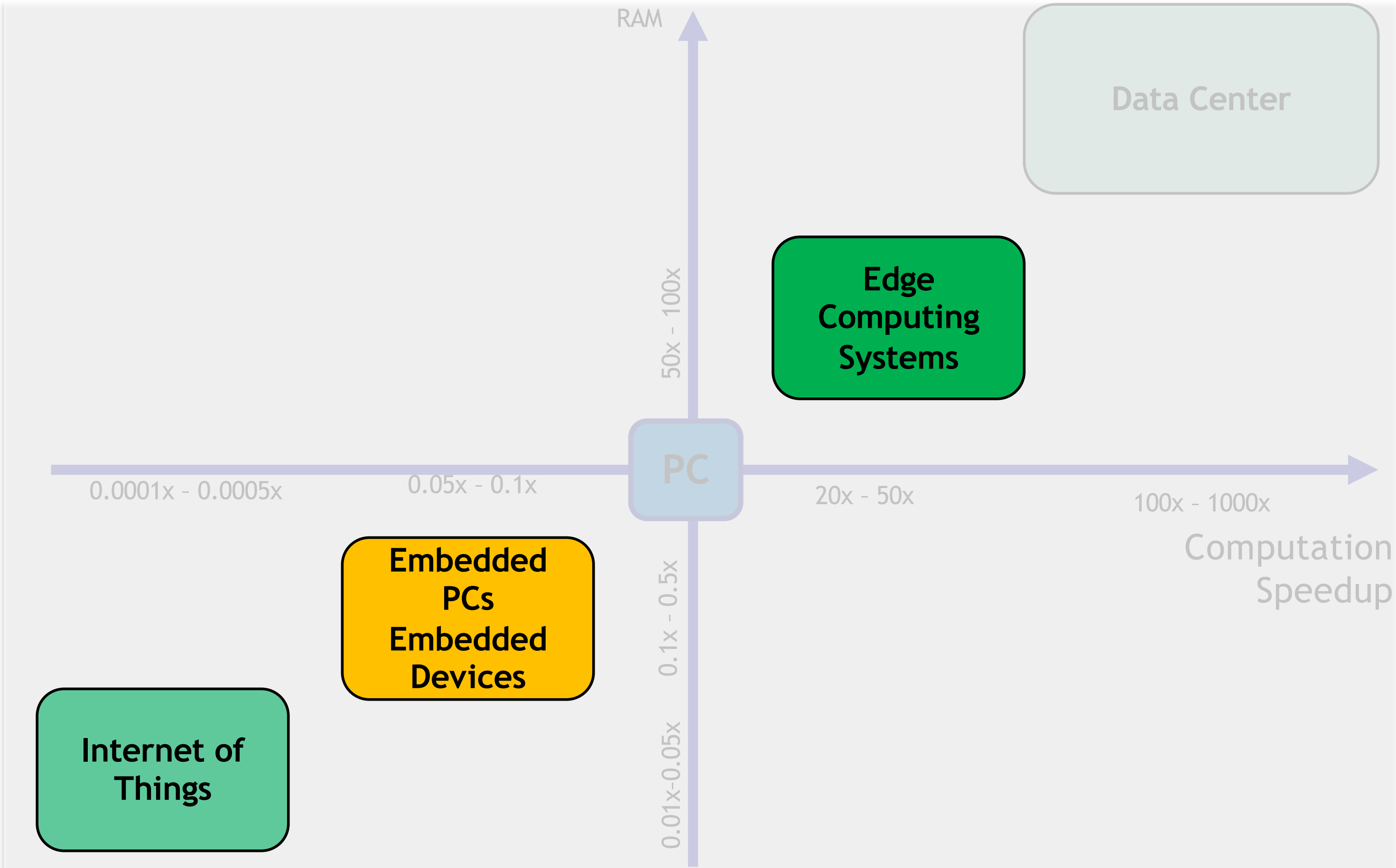


ED POWER

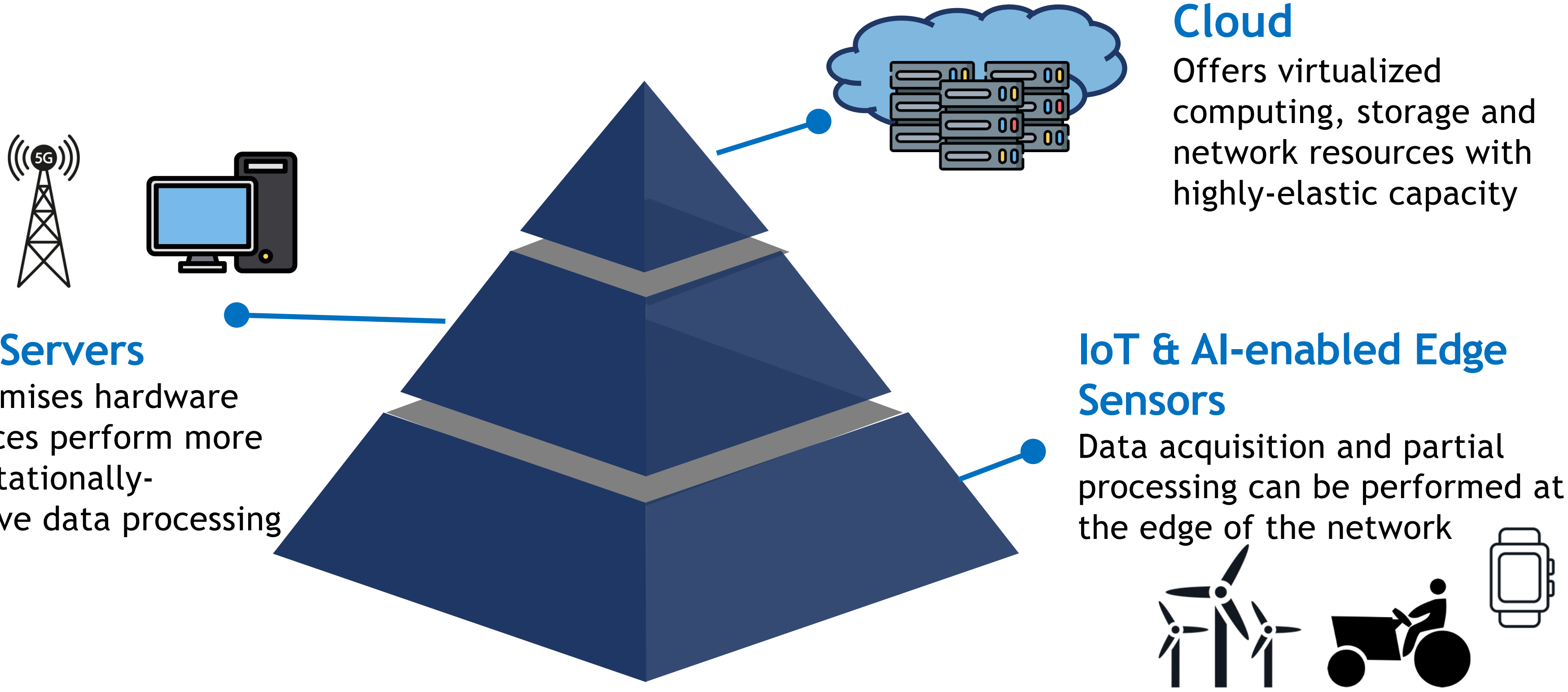


Amortized Cost	Component	Sub-Components
~45%	Servers	CPU, memory, disk
~25%	Infrastructure	UPS, cooling, power distribution
~15%	Power draw	Electrical utility costs
~15%	Network	Switches, links, transit

Edge Computing, PC Embedded and IoT



Edge/Fog Computing Systems

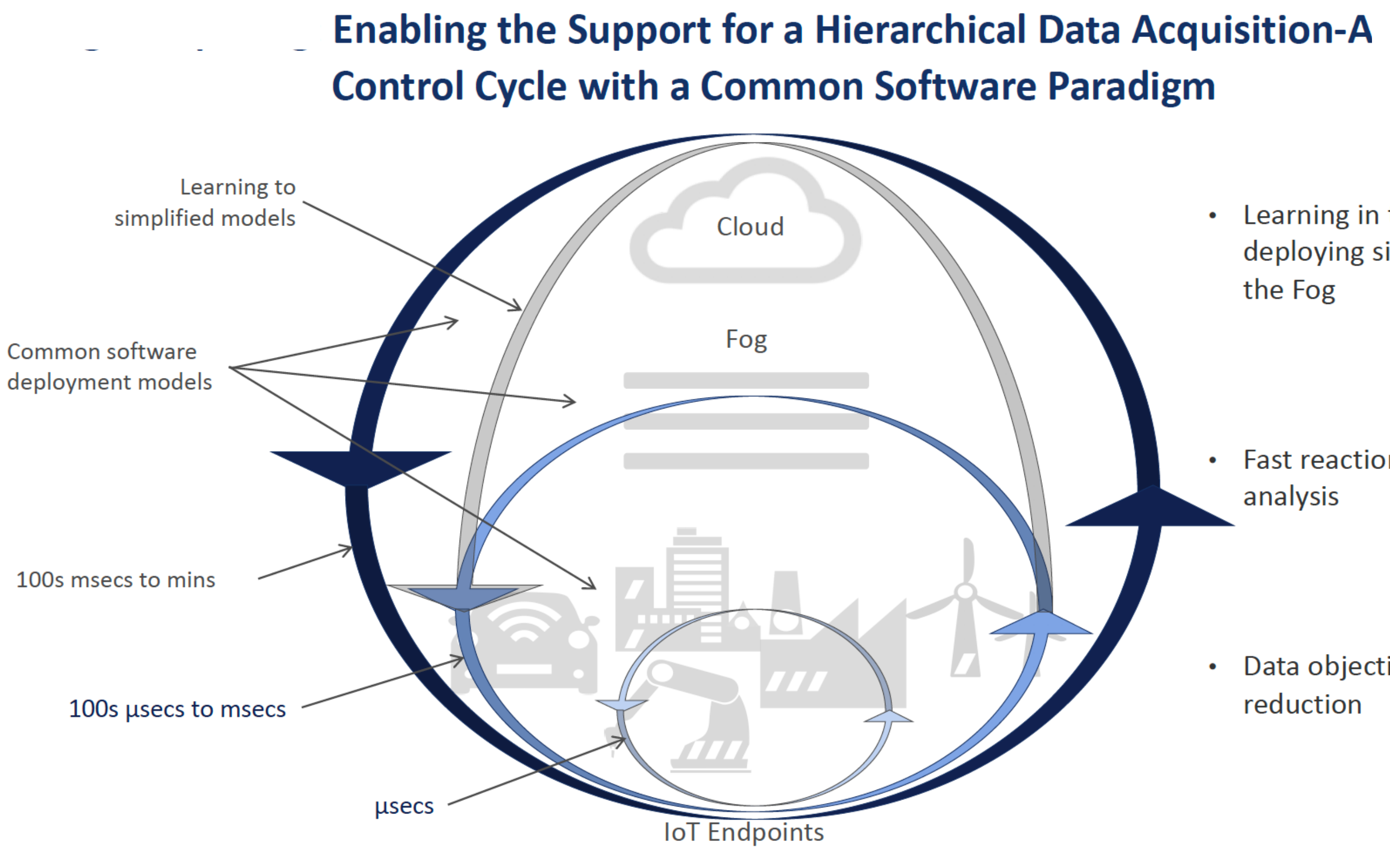


Cloud
Offers virtualized computing, storage and network resources with highly-elastic capacity

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

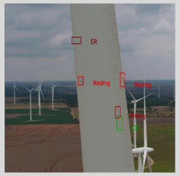
An example: Edge computing in manufacturing



- Learning in the Cloud and deploying simpler models in the Fog
- Fast reaction to rich local analysis
- Data objectization and reduction

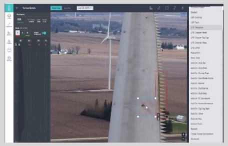
An example: Maintenance and Inspection

Damage detection, classification and severity assessment



CLOUD BASED SOLUTION

Manual data upload to the cloud and cloud only data processing



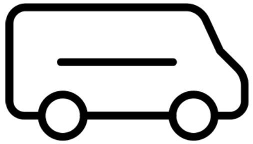
EDGE SOLUTION

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



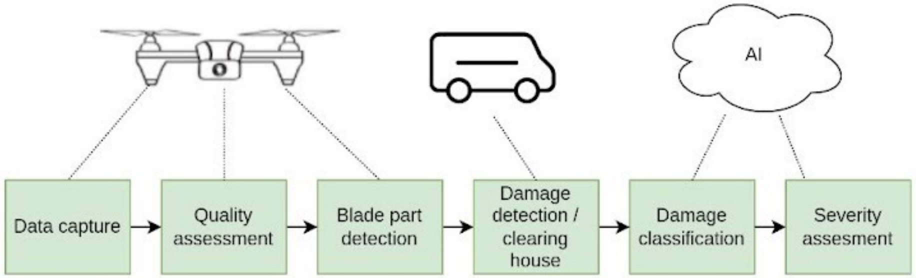
ON-SITE DATA PROCESSING

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

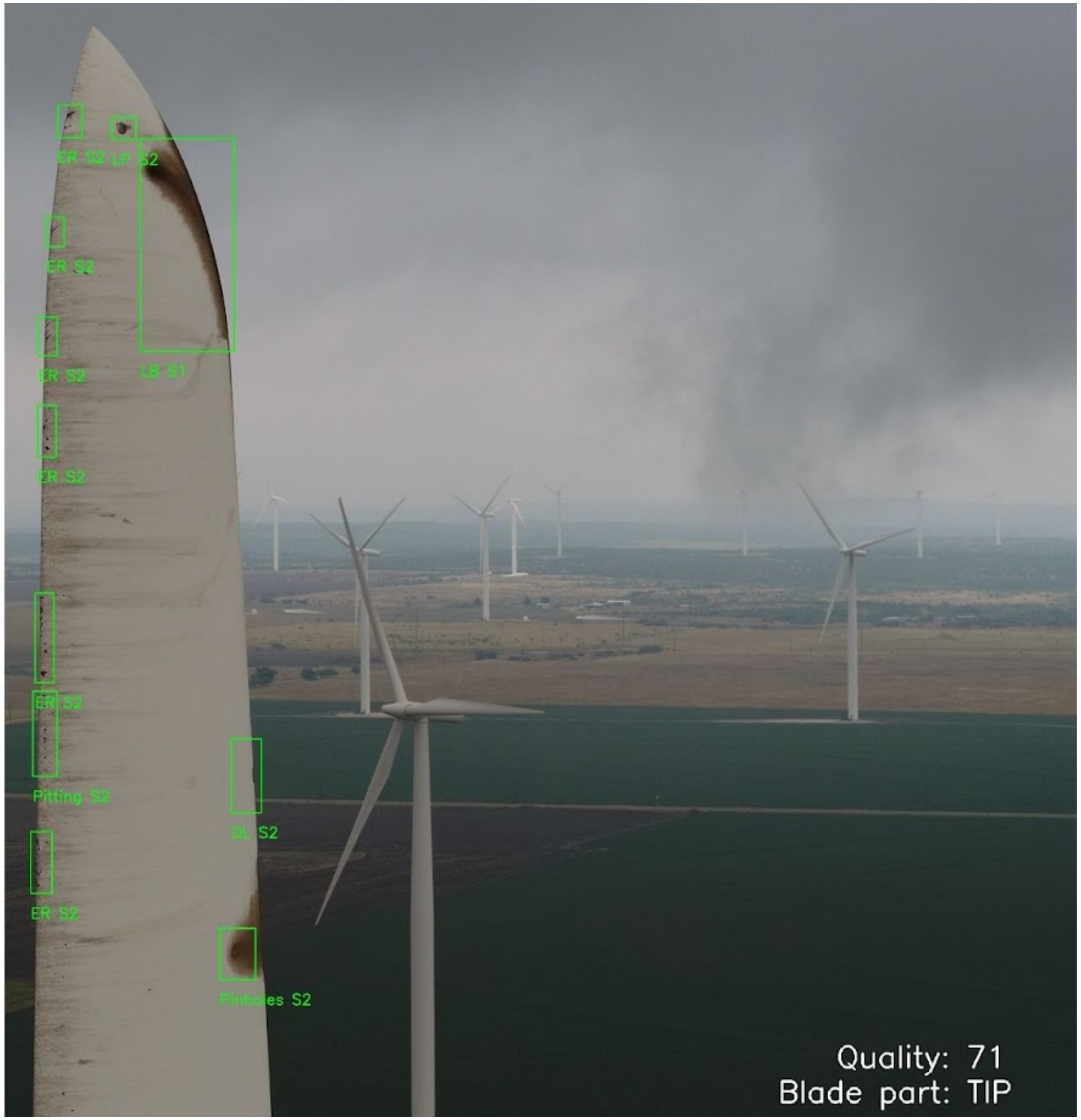
CLOUD DATA PROCESSING

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

An example: Maintenance and Inspection




```
{
    "sha256": "a20aa9ef34bea503ea9ffaa3e3132cccc8d08b3dd89a9768b83d6e56fac7c822",
    "size": {
        "depth": 3,
        "height": 2736,
        "width": 3653
    },
    "quality": "71",
    "element": "TIP",
    "object": [
        {
            "bndbox": {
                "xmax": 1961,
                "xmin": 1773,
                "ymax": 1199,
                "ymin": 772
            },
            "name": "LB",
            "severity": "S1"
        },
        {
            "bndbox": {
                "xmax": 2004,
                "xmin": 1932,
                "ymax": 2460,
                "ymin": 2357
            },
            "name": "Pinholes",
            "severity": "S2"
        },
        {
            "bndbox": {
                "xmax": 2016,
                "xmin": 1956,
                "ymax": 2125,
                "ymin": 1977
            },
            "name": "DL",
            "severity": "S2"
        },
        {
            "bndbox": {
                "xmax": 1764,
                "xmin": 1716,
                "ymax": 774,
                "ymin": 729
            },
            "name": "LP",
            "severity": "S2"
        },
        {
            "bndbox": {
```

Edge/Fog Computing Systems

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)

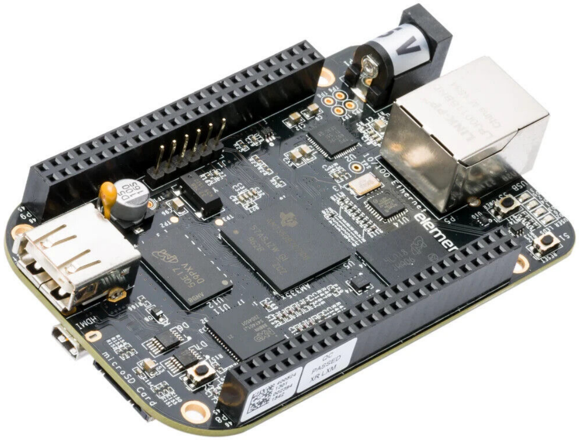
- ✓ High computational capacity
- ✓ Distributed computing
- ✓ Privacy and security
- ✓ Reduced Latency in making a decision

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

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



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

Google Coral Dev Board, 1 GB RAM

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

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

Internet-of-Things

	Arduino Nano	Nano ESP32-S3	Nano 33 IoT	Nano 33 BLE	Nano 33 BLE Sense Rev2	Nano RP2040 Connect	Nano Every	Nano Matter
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	ATMega4809	Silicon Labs MGM24
USB connector	Mini-B USB	USB-C®	Micro USB	Micro USB	Micro USB	Micro USB	Micro USB	USB-C®
I/O voltage	5 V	3.3 V	3.3 V	3.3 V	3.3 V	3.3 V	5 V	3.3 V
Input range	7-12 V	5-21 V	7-12 V	7-12 V	7-12 V	5-21 V	7-12 V	5-21 V
Clock speed	16 MHz	up to 240 MHz	48 MHz	64 MHz	64 MHz	133 MHz	16 MHz	78 MHz
SRAM	2 kB	512 kB	256 kB	256 kB	256 kB	264 kB	6 kB	256 kB
Flash	32 kB	16 MB	1 MB	1 MB	1 MB	16 MB	48 kB	1536 kB

- ✓ 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

An IT perspective for Computing Infrastructures

The application, i.e., the reason why this system exists — SOFTWARE (Application)

Programs and libraries to control the physical resources and provide tools to build applications — SOFTWARE (Environment)

Physical resources of the system (computation, storage, input/output, etc..) — HARDWARE

An IT perspective for Computing Infrastructures