POLITECNICO DI MILANO





The Datacenter as a Computer

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

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, failure recovery

B. **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
   - **Machine and deep learning-as-a-service**

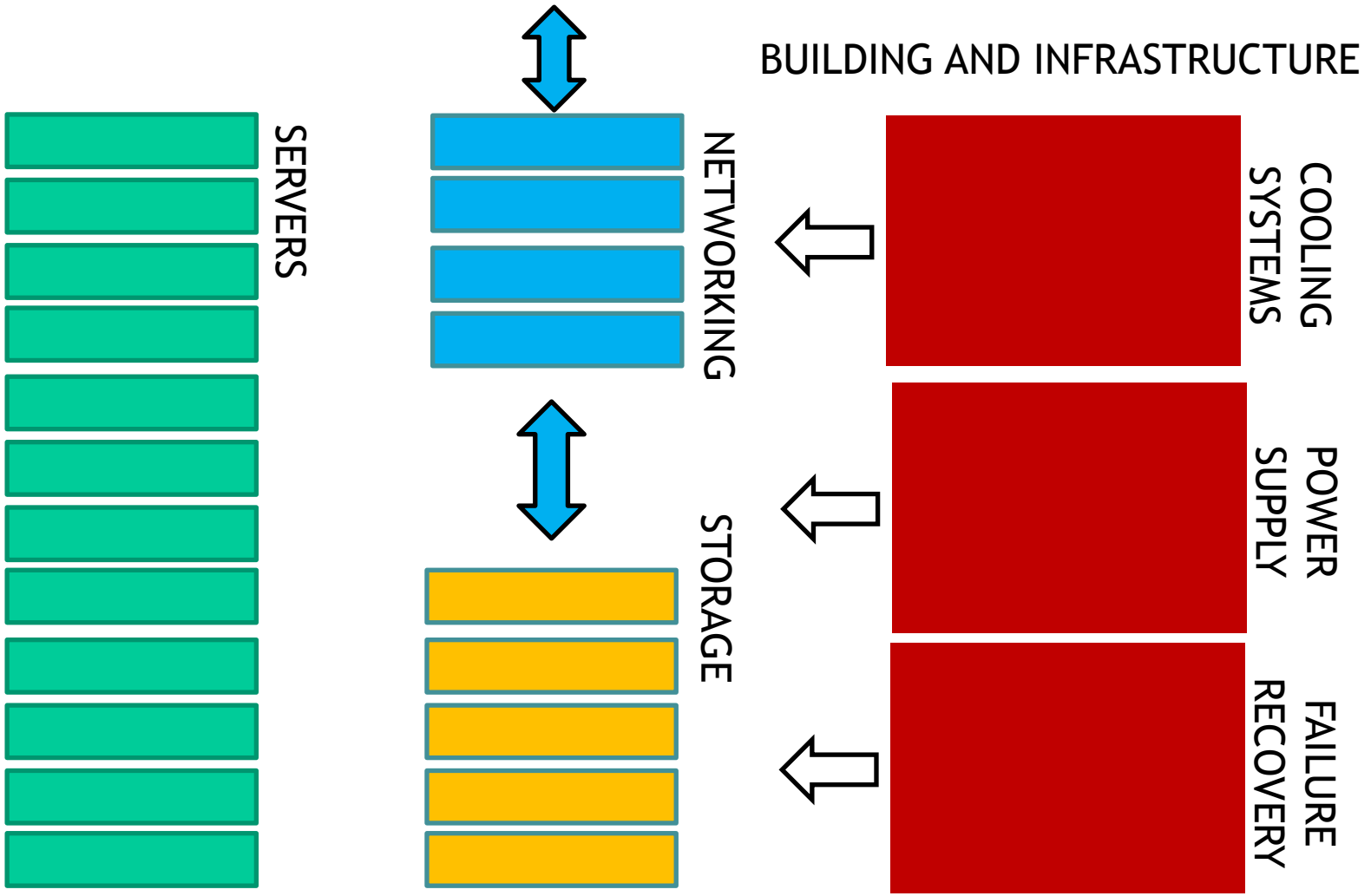
C. **Methods**:
   - **Reliability and availability of datacenters** (definition, fundamental laws, RBDs)
   - **Disk performance** (Type, Performance, RAID)
   - **Scalability and performance of datacenters** (definitions, fundamental laws, queuing network theory)

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

A. HW Infrastructures:

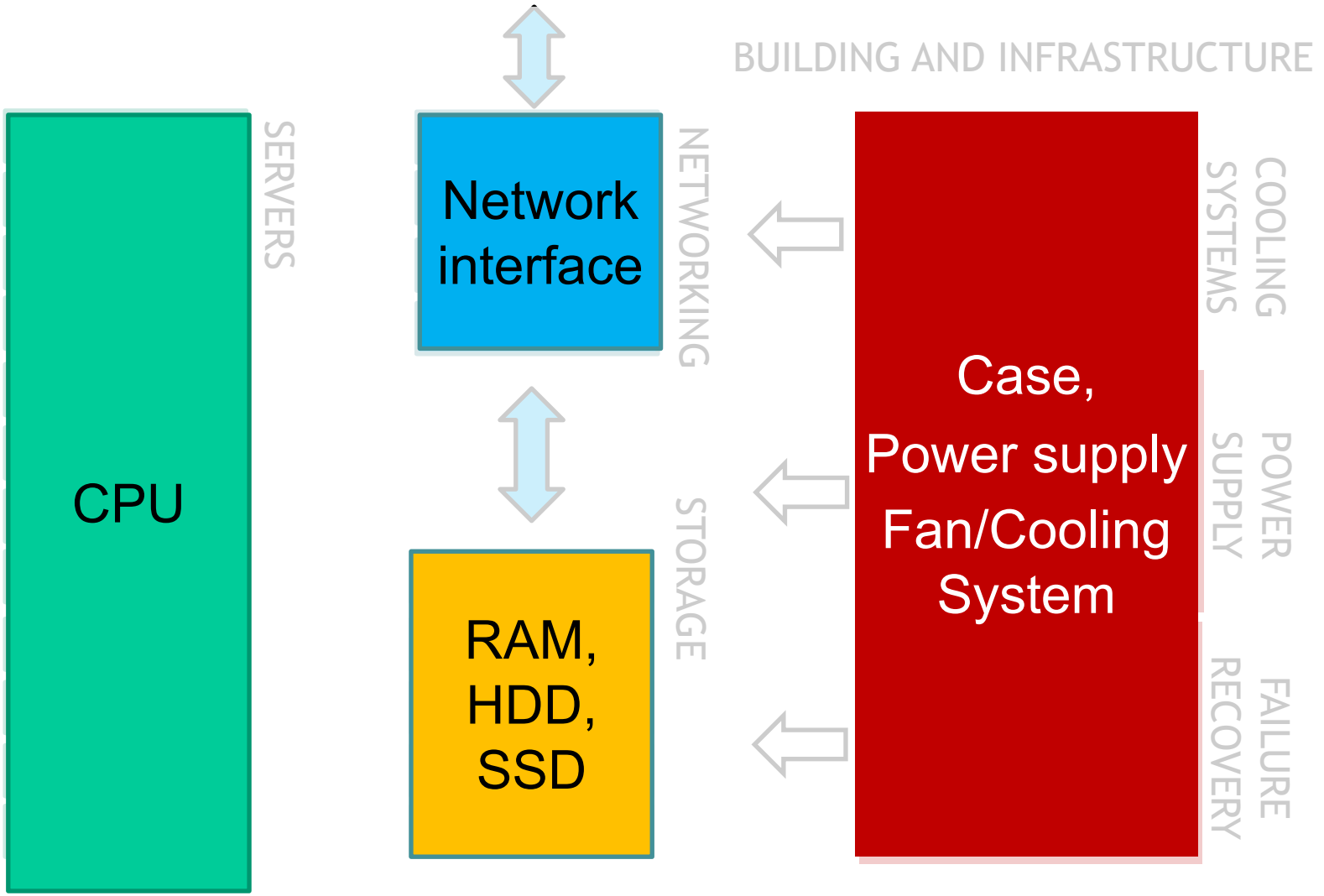
- **System-level**: Computing Infrastructures and Data Center Architectures, Rack/Structure;
- **Node-level**: Server (computation, HW accelerators), Storage (Type, technology), Networking (architecture and technology);
- **Building-level**: Cooling systems, power supply, failure recovery

Node level: computation, storage and networking

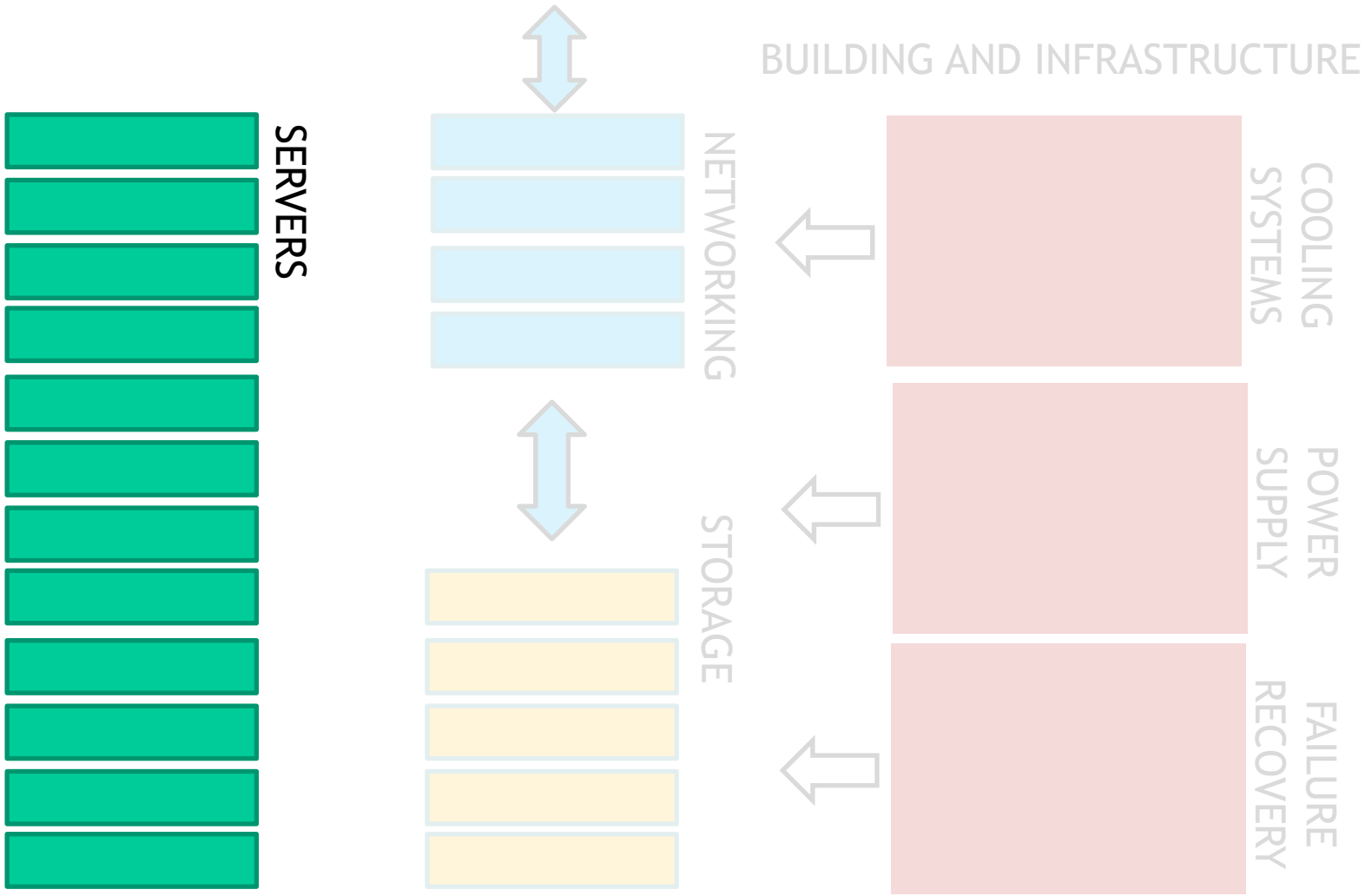


Node level: computation, storage and networking

Node level: computation, storage and networking



SERVERS


SERVERS
NETWORKING
STORAGE
BUILDING AND INFRASTRUCTURE
COOLING SYSTEMS
POWER SUPPLY
FAILURE RECOVERY

SERVER AND RACK OVERVIEW

- Servers hosted in individual shelves are the basic building blocks of WSCs. They are interconnected by hierarchies of networks, and supported by the shared power and cooling infrastructure.

A shelf storing the servers

SERVERS: the main processing equipment

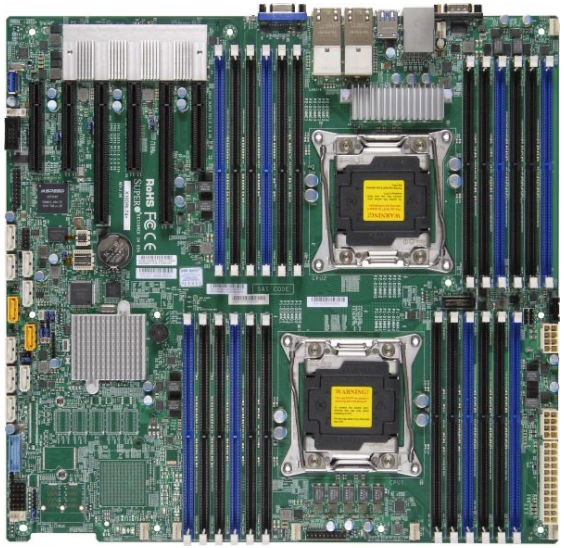
They are like ordinary PC, but with a form factor that allows to fit them into the shelves:

- Rack (1U or more)
- Blade enclosure format
- Tower

- Servers are usually built in a tray or blade enclosure format, housing
  - the motherboard,
  - chipset,
  - additional plug-in components.

The motherboard

- The motherboard provides sockets and plug-in slots to install CPUs, memory modules (DIMMs), local storage (such as Flash SSDs or HDDs), and network interface cards (NICs) to satisfy the range of resource requirements.

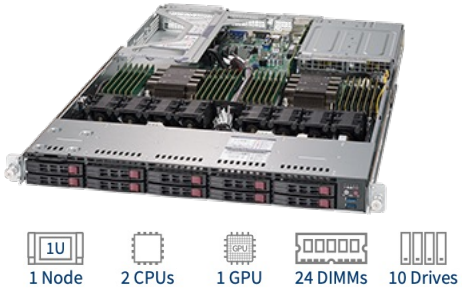


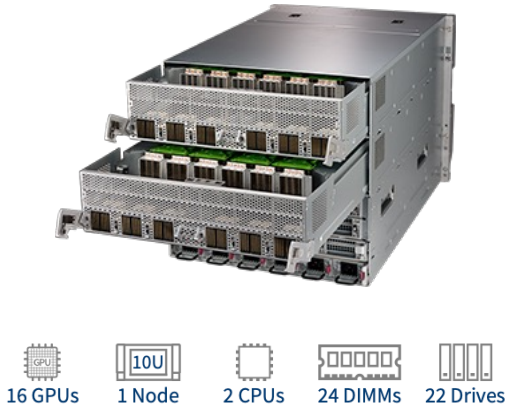
An example: Supermicro Motherboard X10DRi-T4+
Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 v4+/ v3 family; UPI up to 9.6GT/s; Intel® C612 chipset; Up to 3TB+ ECC 3DS LRDIMM, up to DDR4- 2400+MHz ; 24x DIMM slots; 2 PCI-E 3.0 x16, 3 PCI-E 3.0 x8, and 1 PCI-E 2.0 x4 (in x8) slot; Quad LAN w/ Intel® X540 10GBase-T; 10 SATA3 (6Gbps); RAID 0, 1, 5, 10; Integrated IPMI 2.0 and KVM with Dedicated LAN; 5 USB 3.0 (2 rear, 2 front panel, 1 Type-A) 4 USB 2.0 (2 rear, 2 front panel)

WSCs use a relatively homogeneous hardware and system software platform.

Chipset and additional components

- ✓ Number and type of CPUs:
  - From 1 to 8 CPU socket
  - Intel Xeon Family, AMD EPYC, etc.
- ✓ Available RAM:
  - From 2 to 192 DIMM Slots
- ✓ Locally attached disks:
  - From 1 to 24 Drive Bays
  - HDD or SSD (see specific lecture)
  - SAS (higher performance but more expensive) or SATA (for entry level servers)
- ✓ Other special purpose devices:
  - From 1 to 20 GPUs per node, or TPUs
  - NVIDIA Pascal, Volta, A100, etc.
- ✓ Form factor:
  - From 1U to 10U
  - Tower





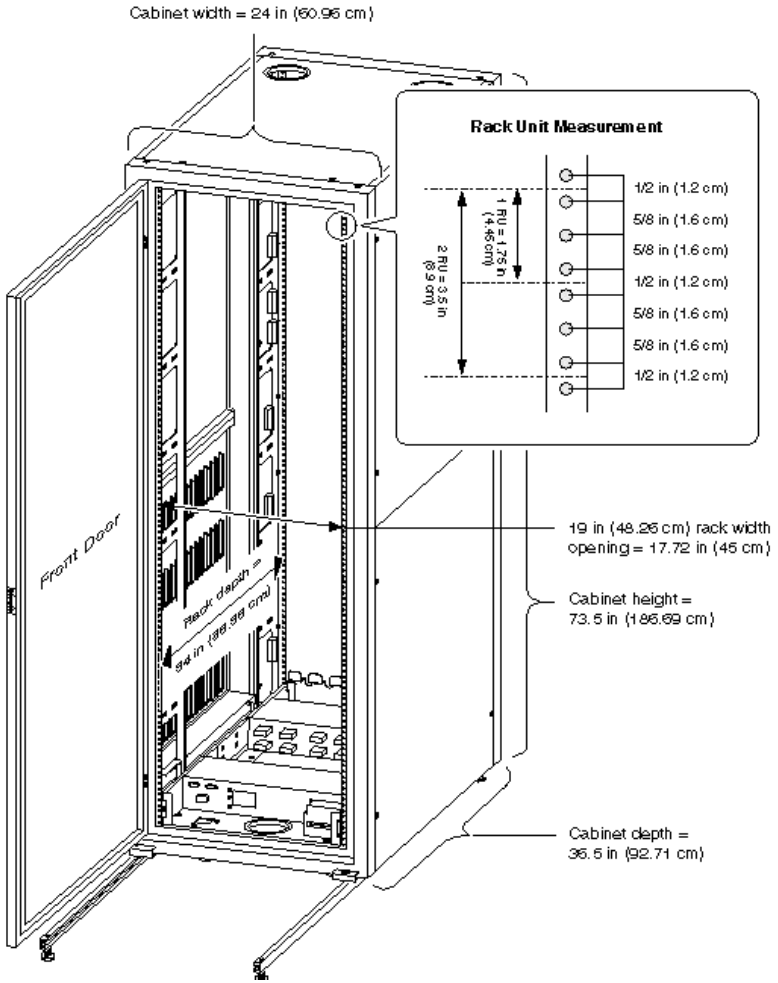
Rack servers

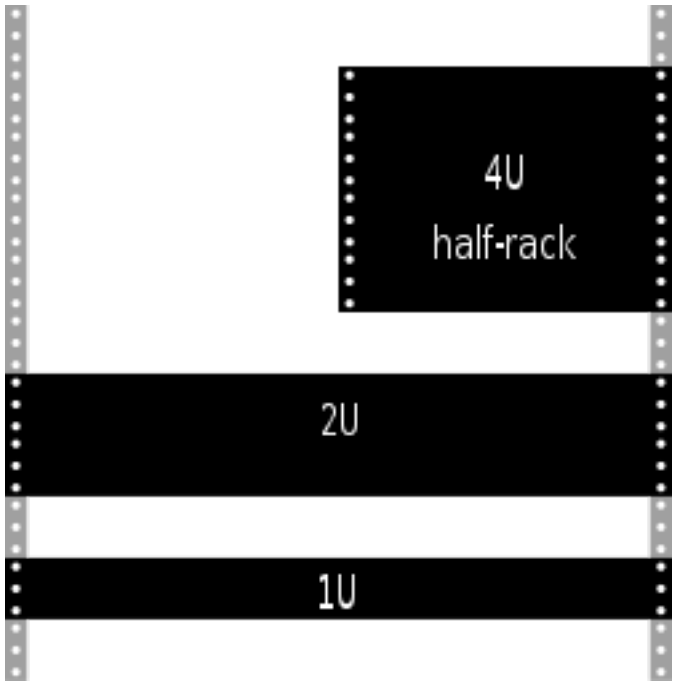
Racks are special shelves that accommodate all the IT equipment and allow their interconnection.

- The racks are used to store these rack servers
- Server racks are measured in rack units, or “U’s”.
- **1U is 44.45 mm** (1.75 inches)
- The advantage of using these racks is that it allows designers to stack up other electronic devices along with the servers.

Data-center racks

IT equipment must conform to specific sizes to fit into the rack shelves.





RACK is not only a physical structure

- The rack is the shelf that holds tens of servers together.
- Handle shared power infrastructure, including power delivery, battery backup, and power conversion
- The width and depth of racks vary across WSCs: some are classic 19-in wide, 48-in deep racks, while others can be wider or shallower.
- It is often convenient to connect the network cables at the top of the rack, such a rack-level switch is appropriately called a Top of Rack (TOR) switch

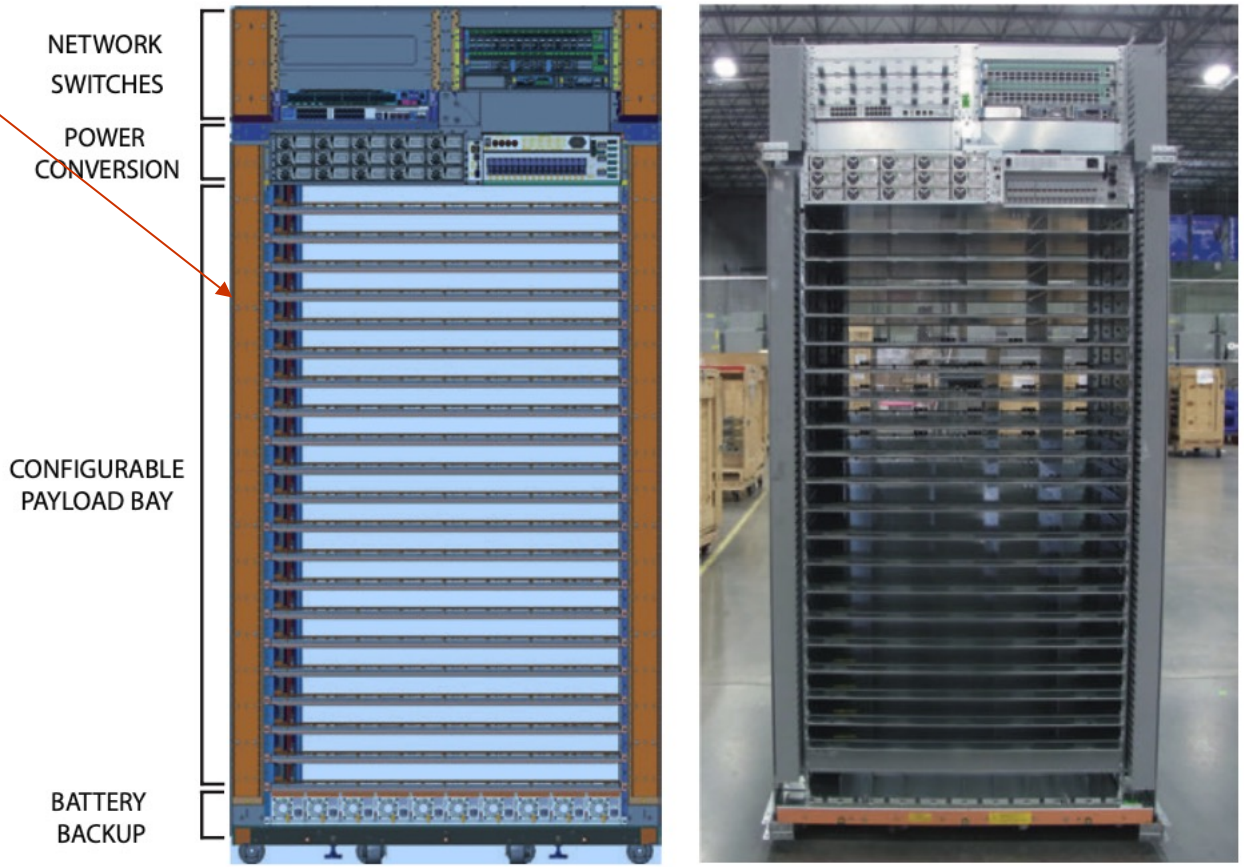


Image taken from **"The Datacenter as a Computer», Barroso et al.**

RACK is not only a physical structure

- The rack is the shelf that holds tens of servers together.
- Handle shared power infrastructure, including power delivery, battery backup, and power conversion
- The width and depth of racks vary across WSCs: some are classic 19-in wide, 48-in deep racks, while others can be wider or shallower.
- It is often convenient to connect the network cables at the top of the rack, such a rack-level switch is appropriately called a Top of Rack (TOR) switch

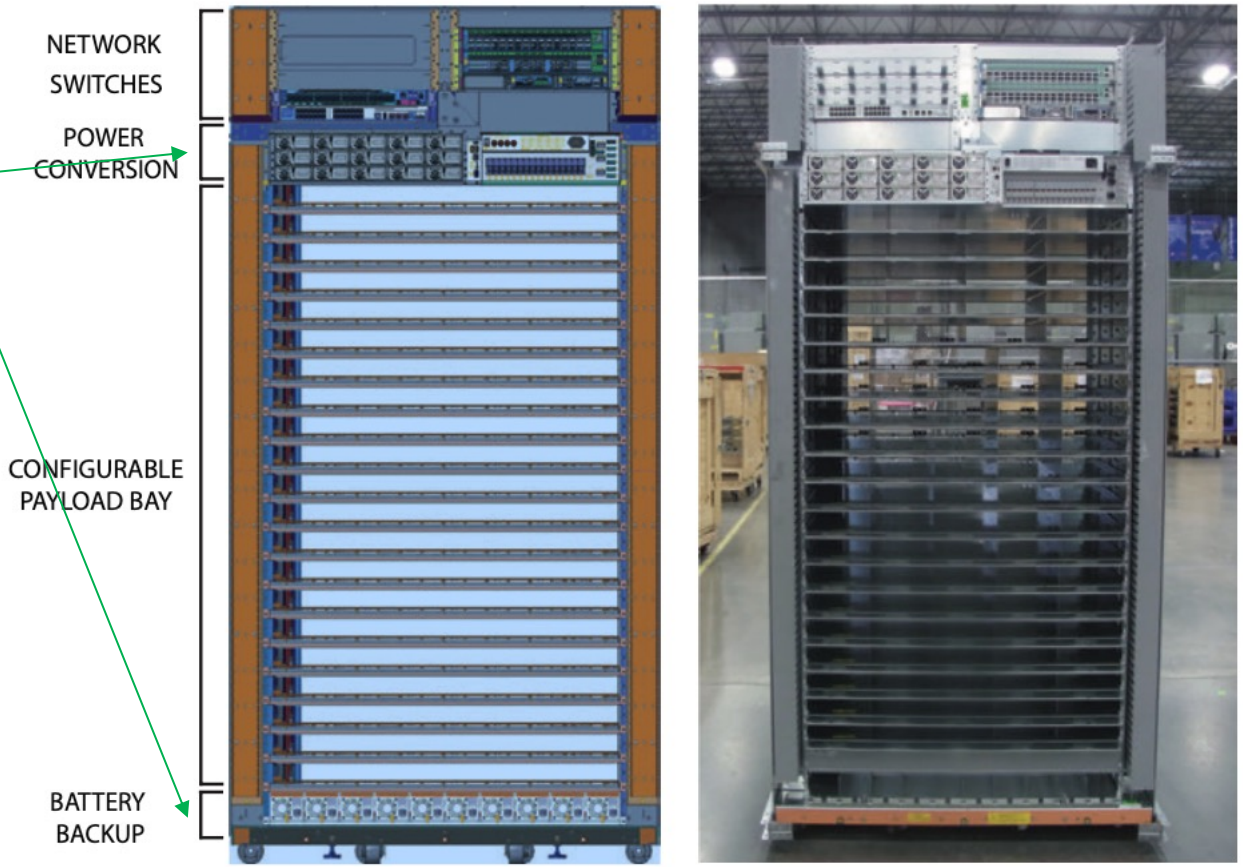


Image taken from **"The Datacenter as a Computer», Barroso et al.**

RACK is not only a physical structure

- The rack is the shelf that holds tens of servers together.
- Handle shared power infrastructure, including power delivery, battery backup, and power conversion
- The width and depth of racks vary across WSCs: some are classic 19-in wide, 48-in deep racks, while others can be wider or shallower.
- It is often convenient to connect the network cables at the top of the rack, such a rack-level switch is appropriately called a Top of Rack (TOR) switch

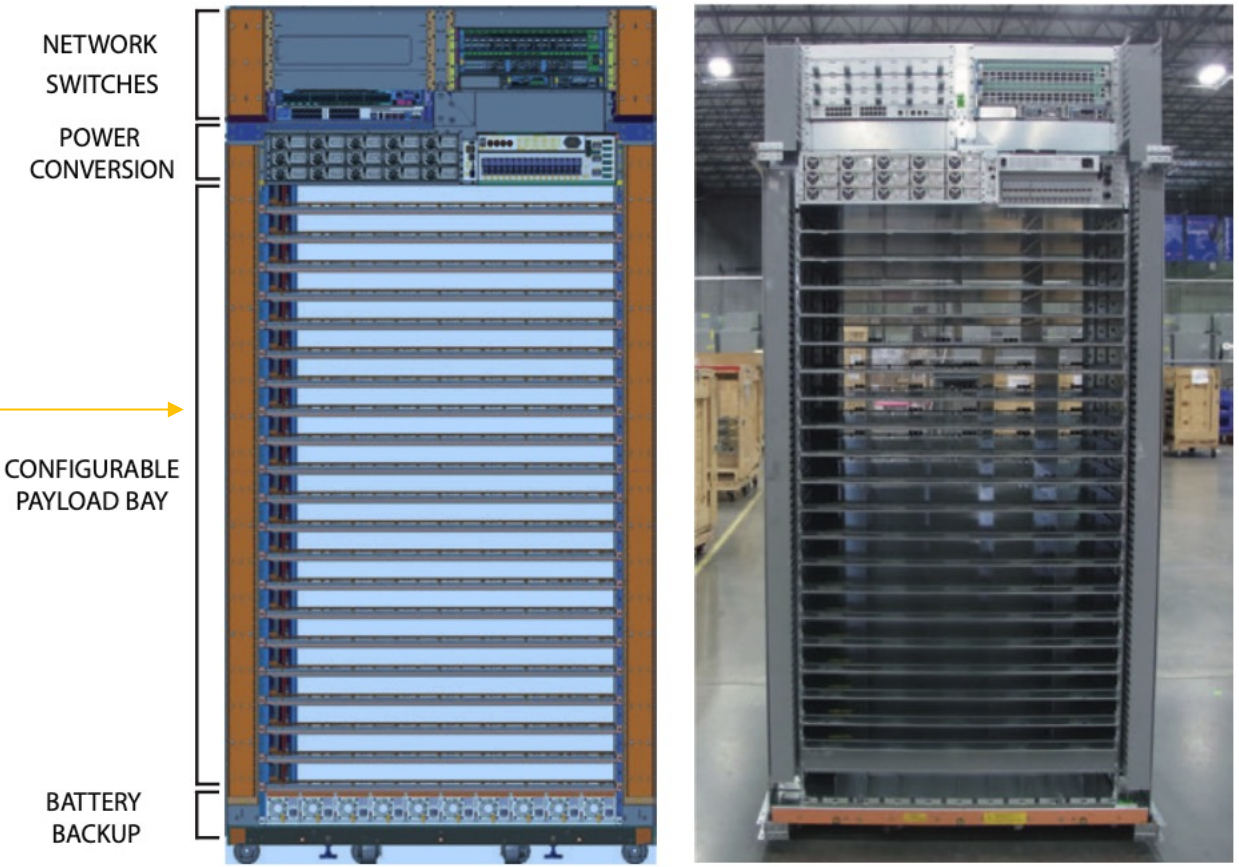


Image taken from **"The Datacenter as a Computer», Barroso et al.**

RACK is not only a physical structure

- The rack is the shelf that holds tens of servers together.
- Handle shared power infrastructure, including power delivery, battery backup, and power conversion
- The width and depth of racks vary across WSCs: some are classic 19-in wide, 48-in deep racks, while others can be wider or shallower.
- It is often convenient to connect the network cables at the top of the rack, such a rack-level switch is appropriately called a Top of Rack (TOR) switch

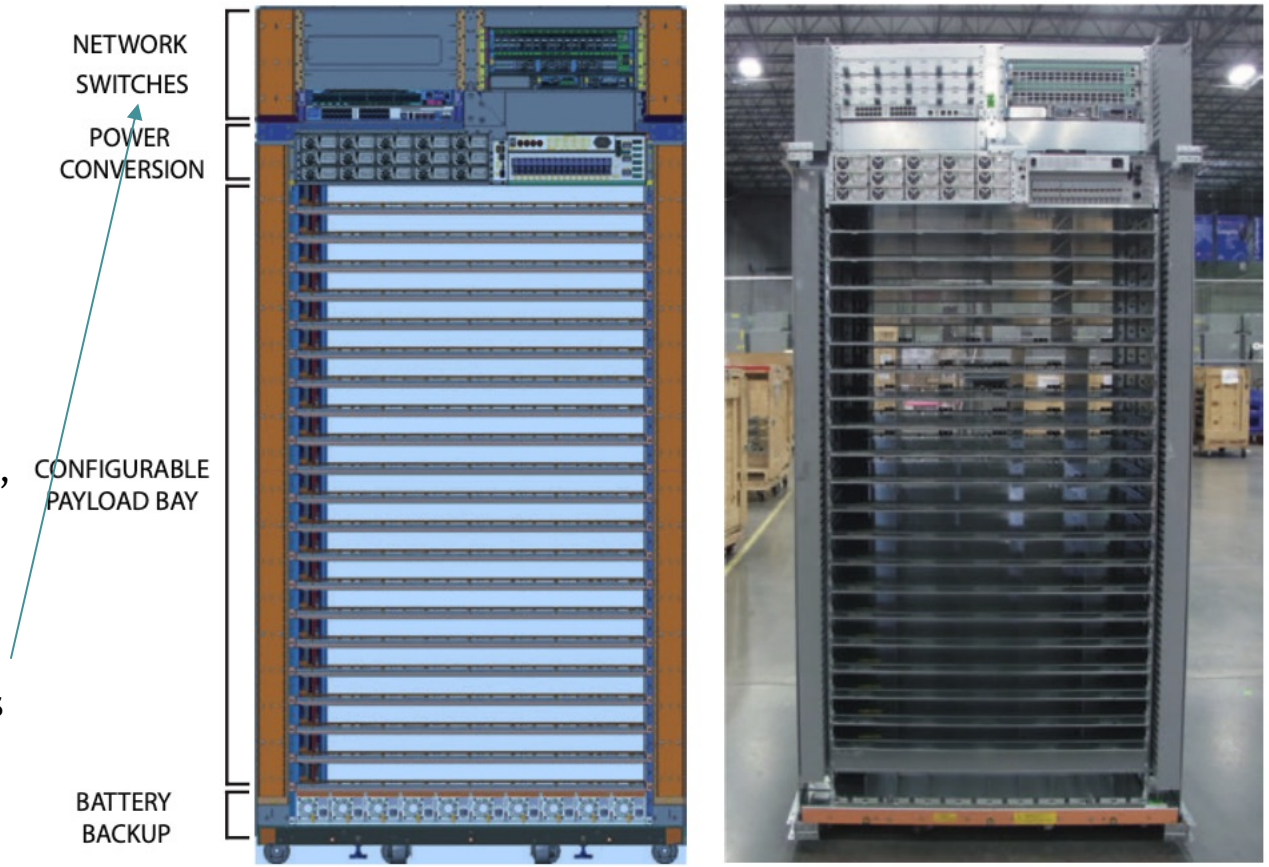


Image taken from **"The Datacenter as a Computer», Barroso et al.**

Rack (vs Tower) vs Blade

Tower Server

A tower server looks and feels much like a traditional tower PC

Pros

- ✓ **Scalability and ease of upgrade:** customized and upgraded based on necessity.
- ✓ **Cost-effective:** Tower servers are probably the cheapest of all kinds of servers
- ✓ **Cools easily:** Since a tower server has a low overall component density, it cools down easily.

Cons

- ✓ **Consumes a lot of space:** These servers are difficult to manage physically.
- ✓ **Provides a basic level of performance:** A tower server is ideal for small businesses that have a limited number of clients.
- ✓ **Complicated cable management**: Devices aren't easily routed together

Rack servers

A rack server is designed to be positioned in a bay, by vertically stacking servers one over the another along with other devices (storage units, cooling systems, network peripherals, batteries)

Pros

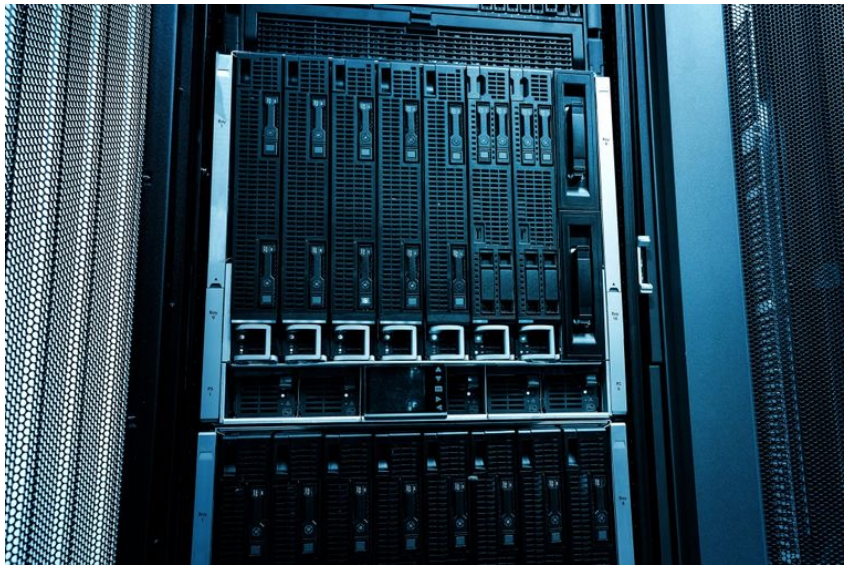
- ✓ **Failure containment:** very little effort to identify, remove, and replace a malfunctioning server with another.
- ✓ **Simplified cable management:** easy and efficient to organize cables.
- ✓ **Cost-effective:** Computing power and efficiency at relatively lower costs.

Cons

- ✓ **Power usage:** Needs of additional cooling systems due to their high overall component density, thus consuming more power.
- ✓ **Maintenance:** Since multiple devices are placed in racks together, maintaining them gets considerably tough with the increasing number of racks.

Blade servers

- Blade servers are the latest and the most advanced type of servers in the market.
- They can be termed as hybrid rack servers, in which servers are placed inside blade enclosures, forming a blade system.
- The biggest advantage of blade servers is that these servers are the smallest types of servers available at this time and are great for conserving space.



A blade system also meets the IEEE standard for rack units and each rack is measured in the units of “U’s”.

Blade servers: advantages

RACK MOUNT SERVERS

BLADE SERVERS

Pros

- ✓ **Load balancing and failover:** Thanks to its much simpler and slimmer infrastructure, load balancing among the servers and failover management tends to be much simpler.
- ✓ **Centralized management:** In a blade server, you can connect all the blades through a single interface, making the maintenance and monitoring easy.
- ✓ **Cabling:** Blade servers don't involve the cumbersome tasks of setting up cabling. Although you still might have to deal with the cabling, it is near to negligible when compared to tower and rack servers.
- ✓ **Size and form-factor:** They are the smallest and the most compact servers, requiring minimal physical space.

Blade servers: disadvantages

RACK MOUNT SERVERS

BLADE SERVERS

Cons

- ✓ **Expensive configuration:** Although upgrading the blade server is easy to handle and manage, the initial configuration or the setup might require heavy efforts in complex environments.
- ✓ **HVAC:** Blade servers are very powerful and come with high component density. Therefore, special accommodations have to be arranged for these servers in order to ensure they don't get overheated. Heating, ventilation, and air conditioning systems must be managed well in the case of blade servers.
- ✓ **Vendor Lock-In**: Blade servers typically require the use of the manufacturer's specific blades and enclosures, leading to vendor lock-in. This can limit flexibility and potentially increase costs in the long run.

An example of a server for WSCs

PowerEdge T640 Server (210-AMBC) — Q.tà: 1

Componenti

1	329-BEPB	PowerEdge T640 MLK Motherboard
2	338-BVJX	Intel Xeon Silver 4214R 2.4G, 12C/24T, 9.6GT/s, 16.5M Cache, Turbo, HT (100W) DDR4-2400
1	379-BDCO	Additional Processor Selected
1	379-BCSF	iDRAC,Factory Generated Password
1	379-BCQV	iDRAC Group Manager, Enabled
1	321-BCXN	Chassis with up to (16 + 16) x 2.5" SAS/ SATA Hard Drives, Single PERC, Rack Configuration
1	325-BCON	Dell EMC Logo Push Pin
1	343-BBFI	MOD,SHP MTL,T640
1	350-BBLG	Rack Quick Sync 2 (At-the-box-mgmt)
1	370-AAIP	Performance Optimized
1	370-AEVR	3200MT/s RDIMMs
12	370-AEVN	32GB RDIMM, 3200MT/s, Dual Rank
2	400-AXSF	960GB SSD SATA Read Intensive 6Gbps 512 2.5in Hot-plug AG Drive
4	400-AUTI	1.2TB 10K RPM SAS 12Gbps 512n 2.5in Hot-plug Hard Drive
1	405-AANW	PERC H730P+ Adapter RAID Controller, 2GB
2	412-AAJW	Standard Heat Sink for Less = 150W
1	429-ABBF	No Internal Optical Drive for x4 and x8 HDD Chassis
1	450-ADWM	Dual, Hot-plug, Redundant Power Supply (1+1), 1100W
2	450-AADY	C13 to C14, PDU Style, 10 AMP, 6.5 Feet (2m), Power Cord
1	461-AAEM	Trusted Platform Module 2.0
1	293-10049	Order Configuration Shipbox Label (Ship Date, Model, Processor Speed, HDD Size, RAM)
1	293-10025	Asset Tag - ProSupport (Website, barcode, Onboard MacAddress)
1	470-ACLI	GPU Cable for T640
1	389-DTCH	PowerEdge T640 CE,CCC,BIS Marking
1	490-BFRQ	NVIDIA Quadro RTX 6000 24 GB, 250W, Dual Slot, PCIe x16 Passive Cooled, Full Height GPU
1	542-BBCT	PE T640 On-Board LOM
1	750-AABF	Power Saving Dell Active Power Controller
1	780-BCDT	RAID 1 + Unconfigured RAID

An example of a server for WSCs

Software

1	384-BBSE	4 High Performance Mid Fans & 2 External FI Fans for T640
1	384-BBSQ	2 External Factory Install Fans, T640
1	619-ABVR	No Operating System
1	631-AACK	No Systems Documentation, No OpenManage DVD Kit

Servizi

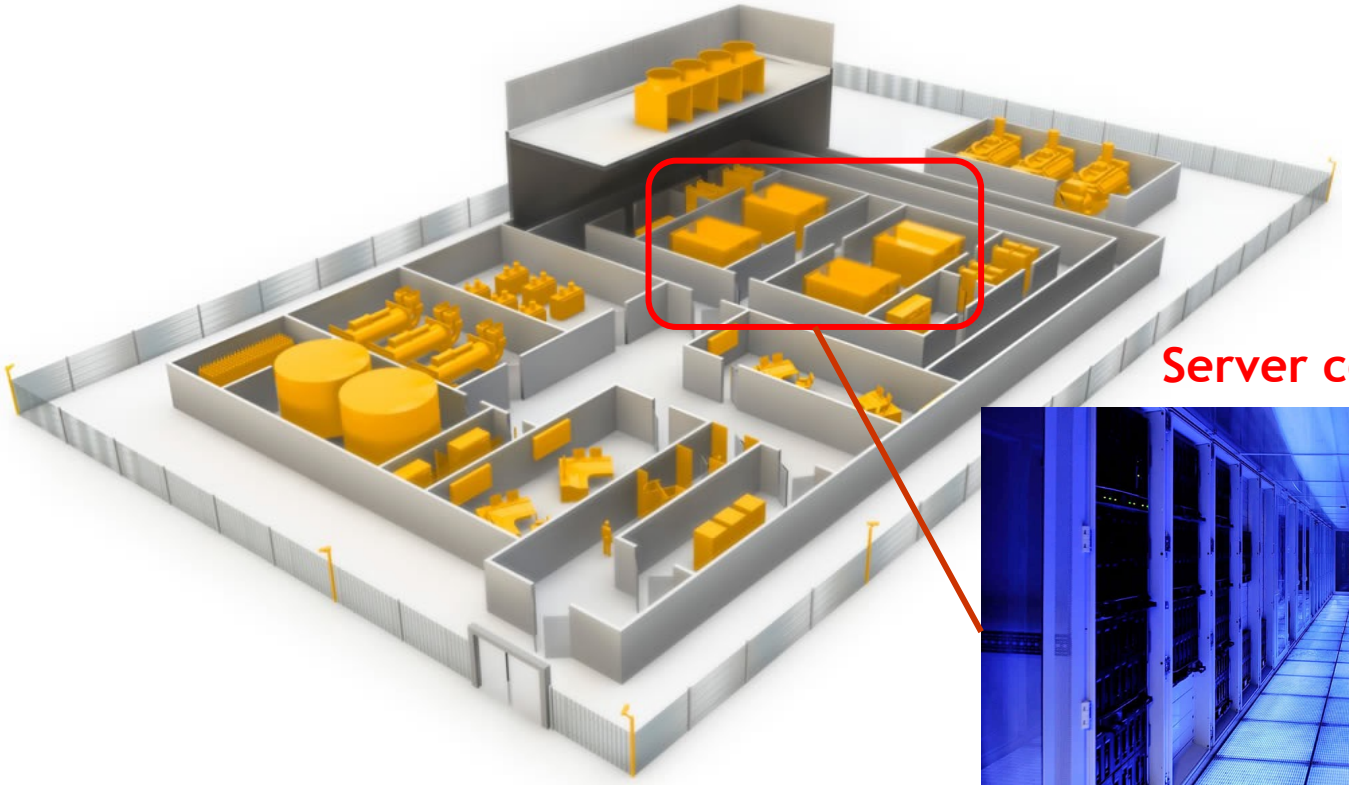
1	489-BBCW	GPU Installation Kit T640
1	683-19198	Basic Deployment Dell Server T Series
1	528-BIYY	OpenManage Enterprise Advanced
1	528-CIBI	iDRAC9 Datacenter 14G
1	709-BBIM	Next Business Day 36MONTHS
1	865-BBMY	ProSupport Next Business Day Onsite Service Initial, 36 Mese/i

Total Cost: 13.900,00€

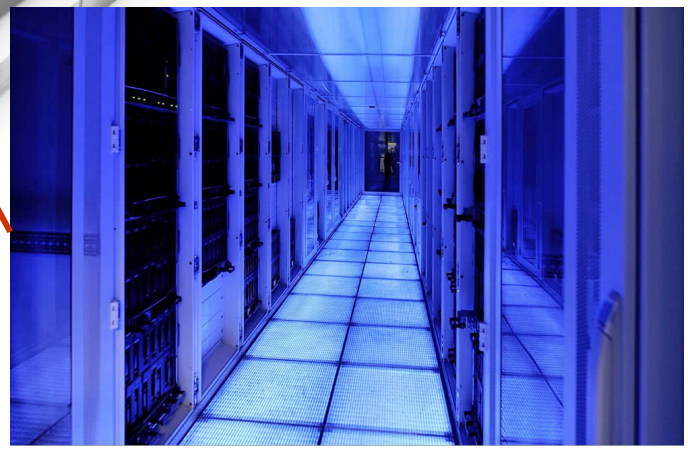
From the Rack to the Datacenter ….

Data-center architecture

The IT equipment is stored into corridors, and organized into racks.

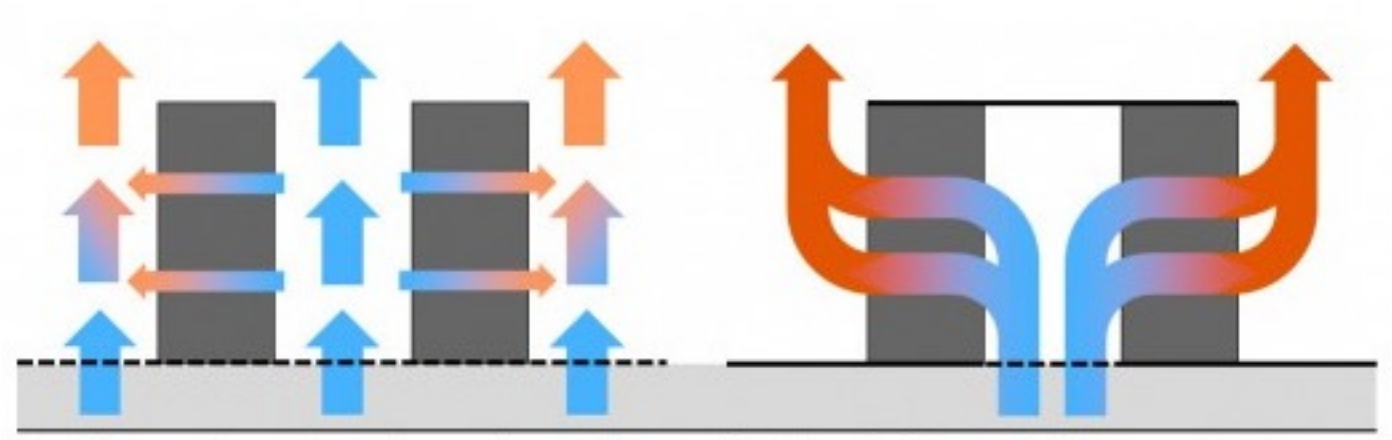


Server corridors



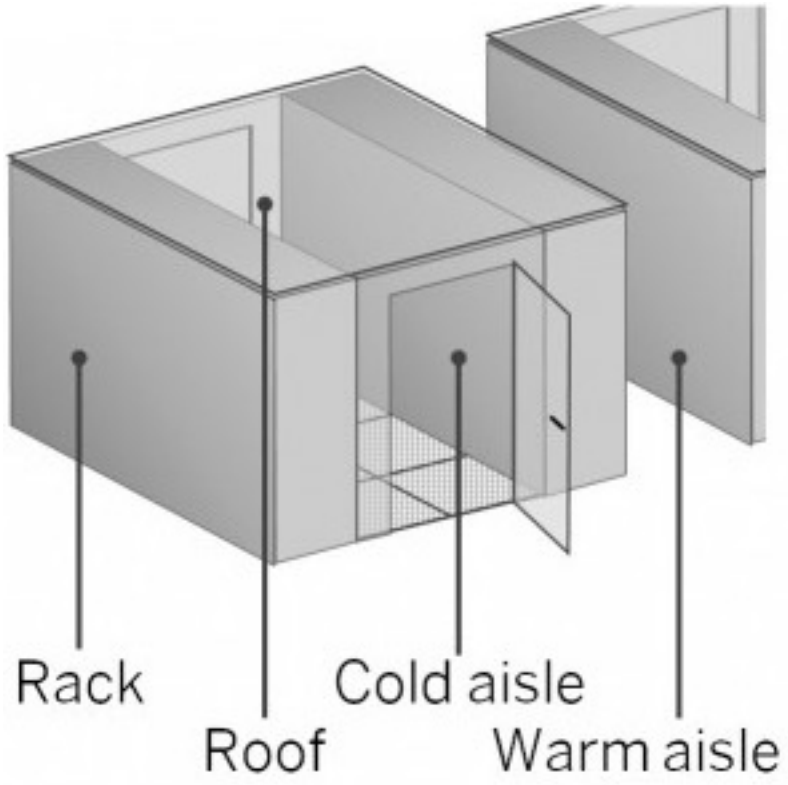
Data-center corridors

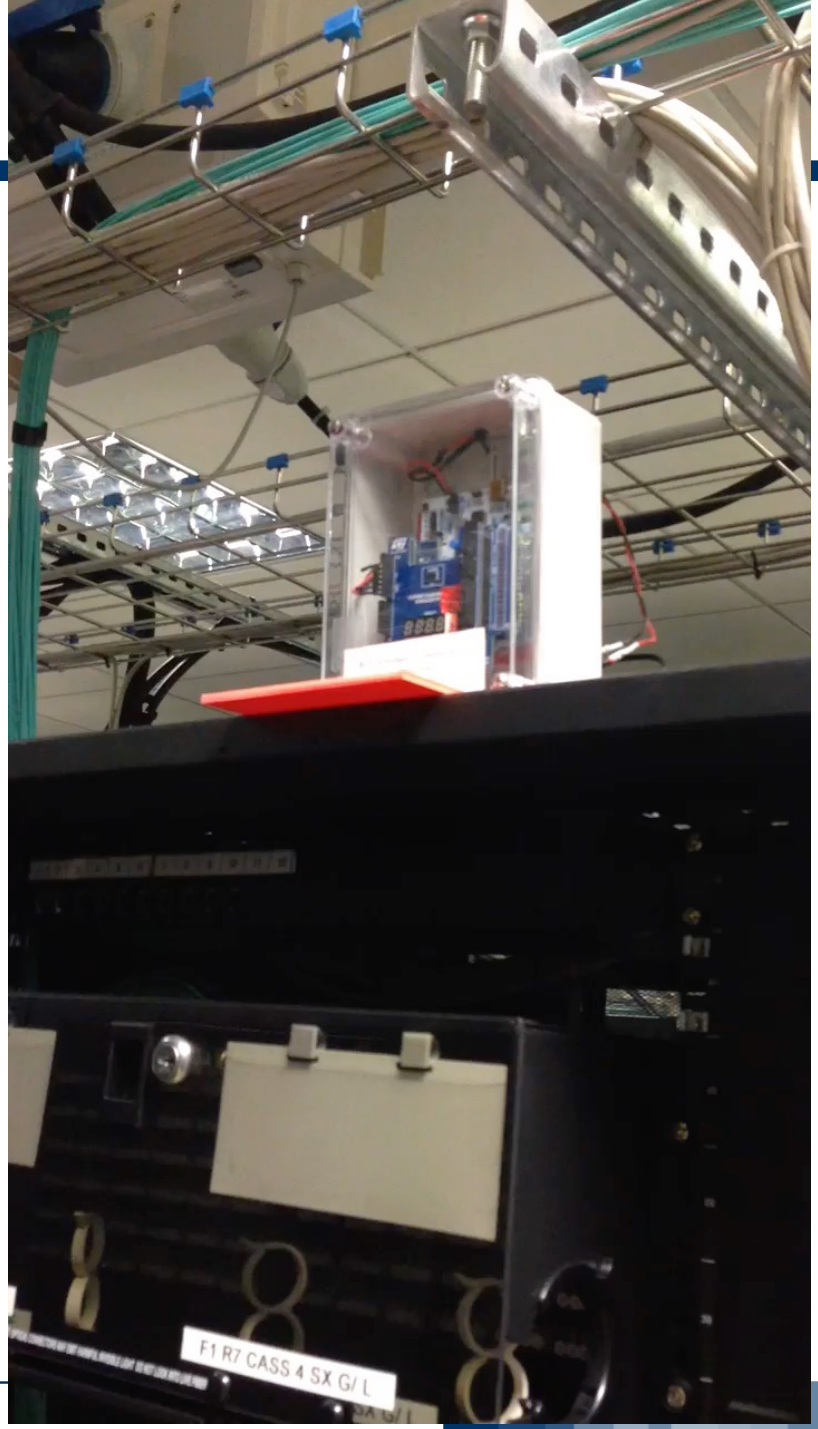
Cold air flows from the front (cool aisle), cools down the equipment, and leave the room from the back (warm aisle).



Data-center corridors

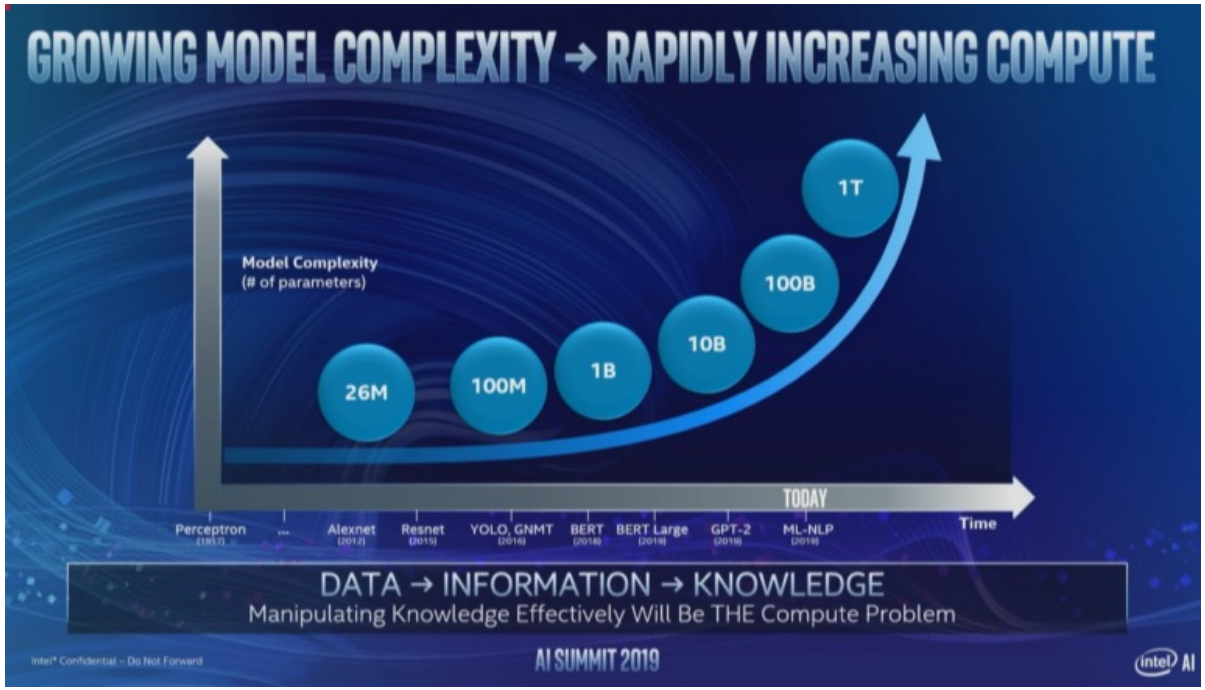
Corridors where servers are located are split into *cold aisle*, where the front panels of the equipment is reachable, and *warm aisle*, where the back connections are located.




F1 R7 CASS 4 SX G/ L

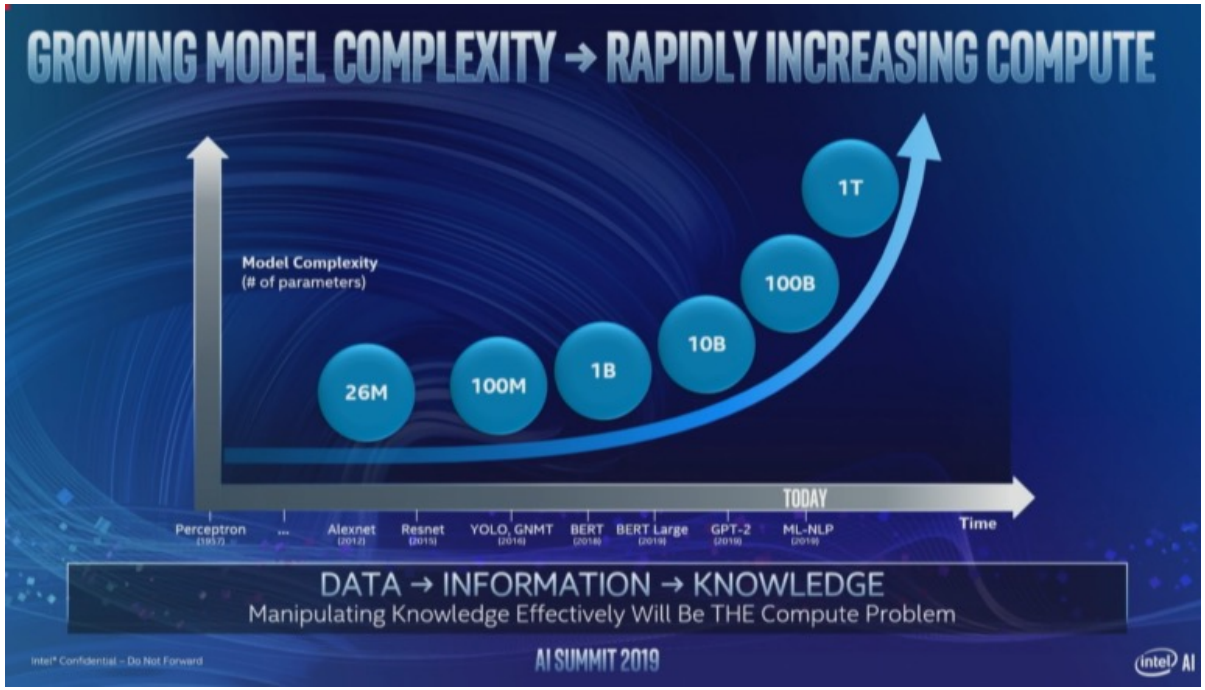

EMC
SYMMETRIX
VMAX

The need of hardware accelerators



Complexity doubles every 3.5 months

The need of hardware accelerators



Complexity doubles every 3.5 months

↓

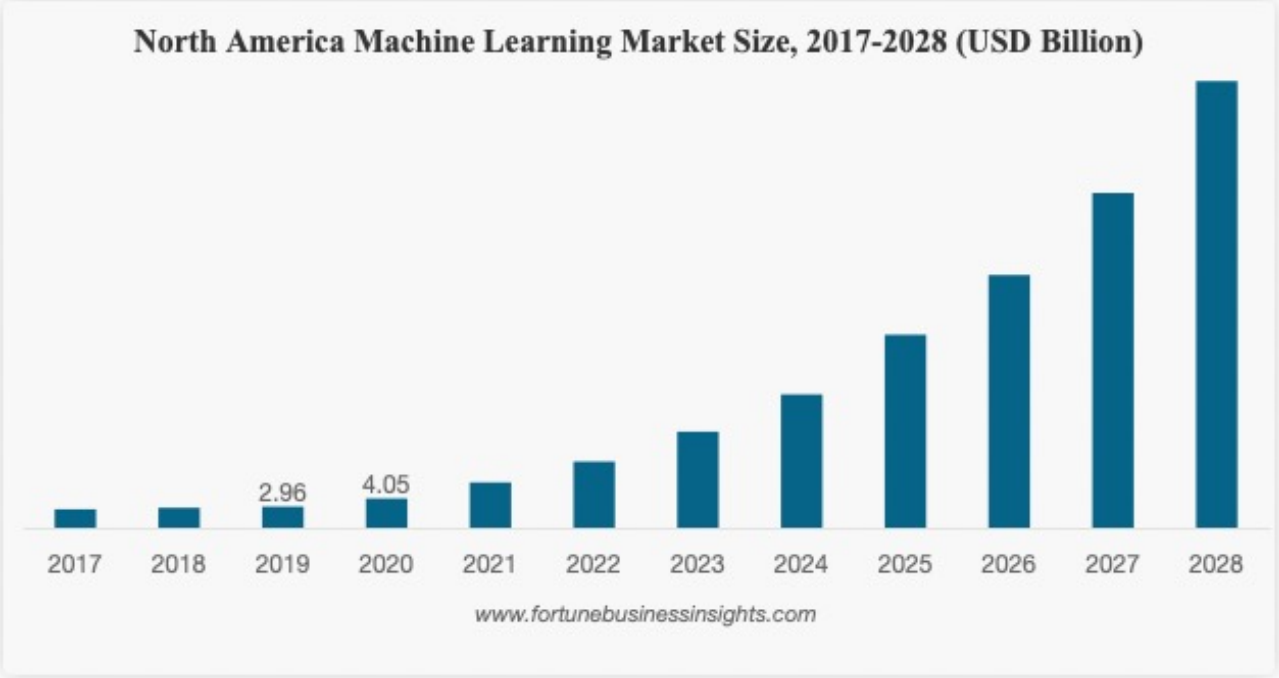
18-24 months for Moore's Law

↓

Nowadays Moore's Law capacity double every 4 or more years

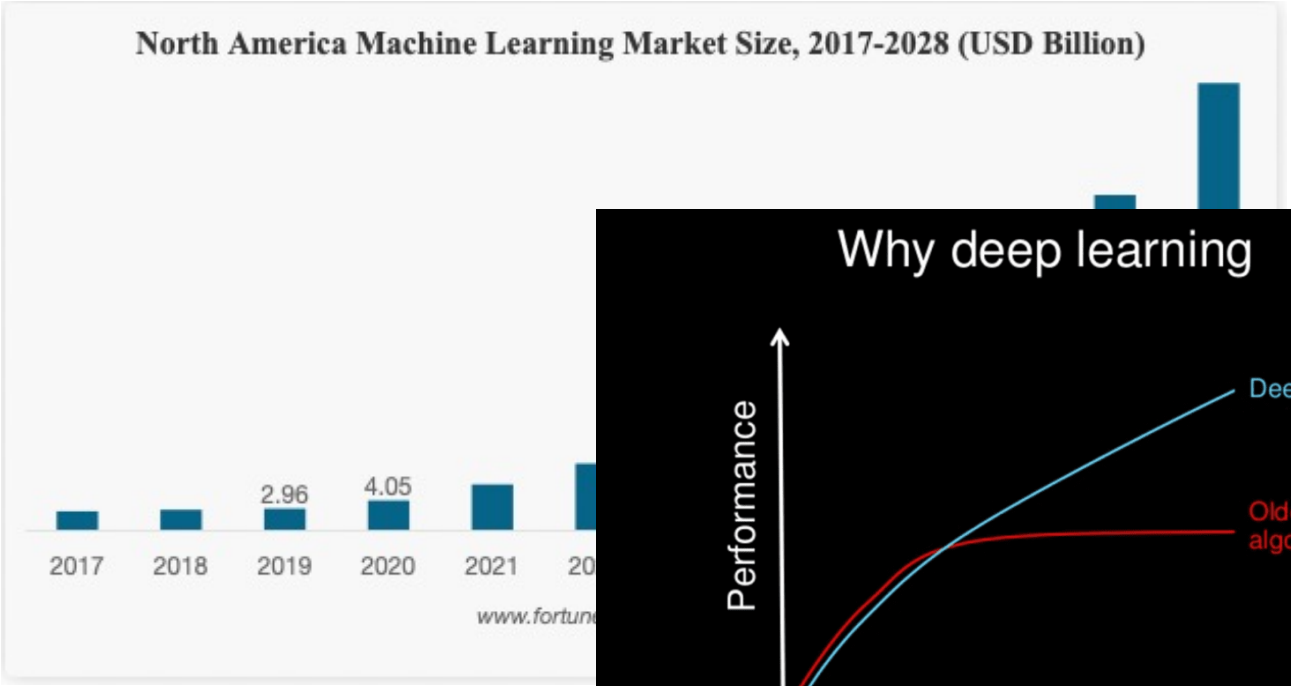
Machine Learning Market Growth

- The machine learning EU market reached a value of about $1.41 billion in 2020 and is expected to reach $8.81 billion by 2025



Machine Learning Market Growth

- The machine learning EU market reached a value of about $1.41 billion in 2020 and is expected to reach $8.81 billion by 2025



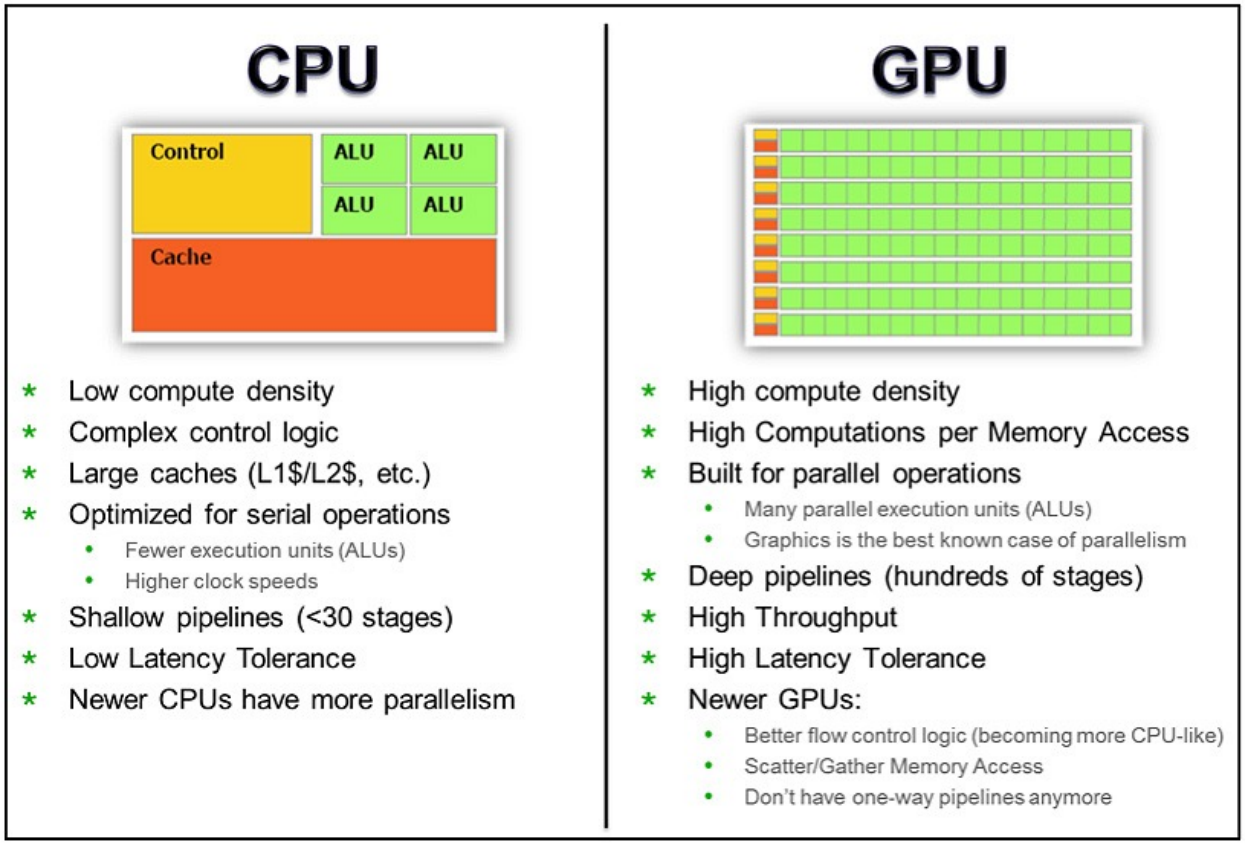
Slide by Andrew Ng

Hardware accelerators

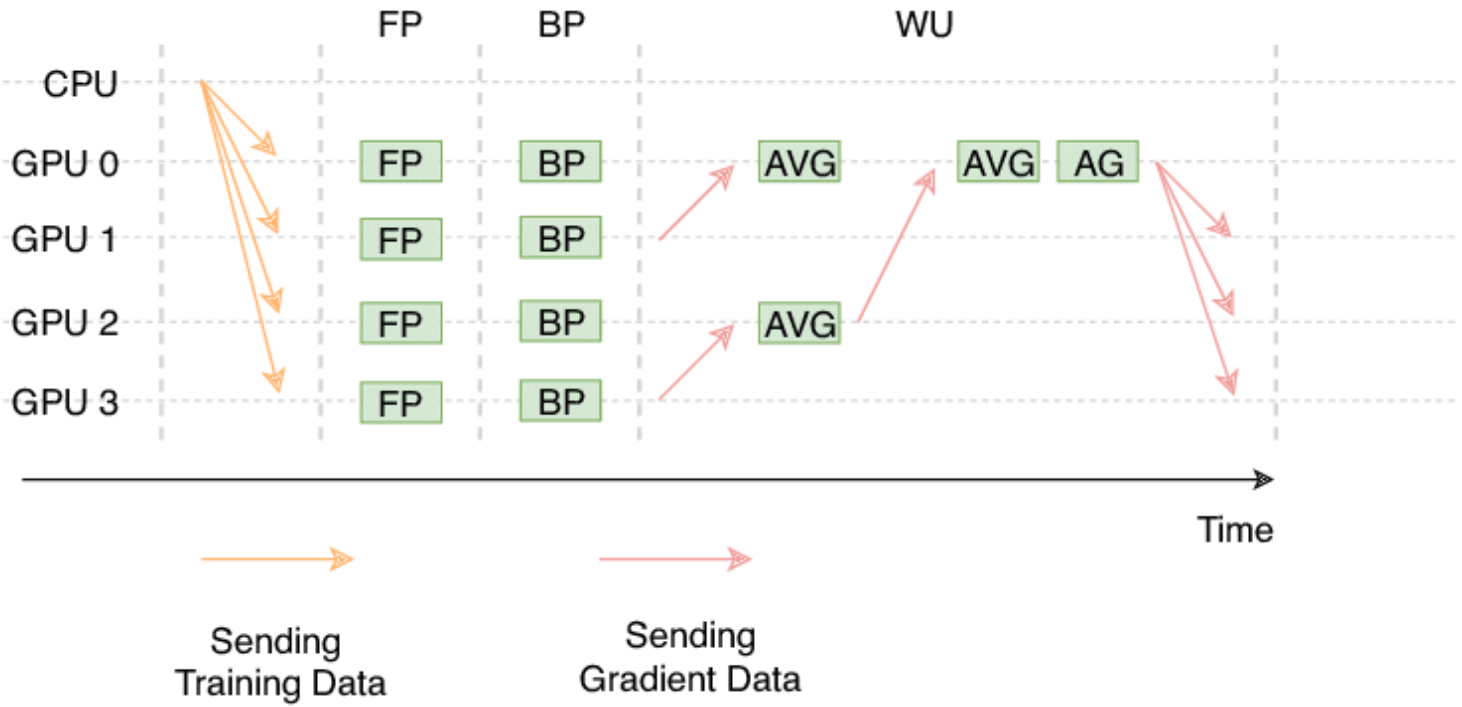
- Deep learning models began to appear and be widely adopted, enabling specialized hardware to power a broad spectrum of machine learning solutions.
- Since 2013, AI training compute requirements have doubled every 3.5 months (vs. 18–24 months expected from Moore's Law).
- To satisfy the growing compute needs for deep learning, WSCs deploy specialized accelerator hardware:
  - GPU
  - TPU
  - FPGA

Graphical Processing Units (GPU)

- **Data-parallel computations:** the same program is executed on many data elements in parallel
- The scientific codes are mapped onto the matrix operations
- High level languages (such as CUDA and OpenCL) are required
- Up to 1000x faster than CPU



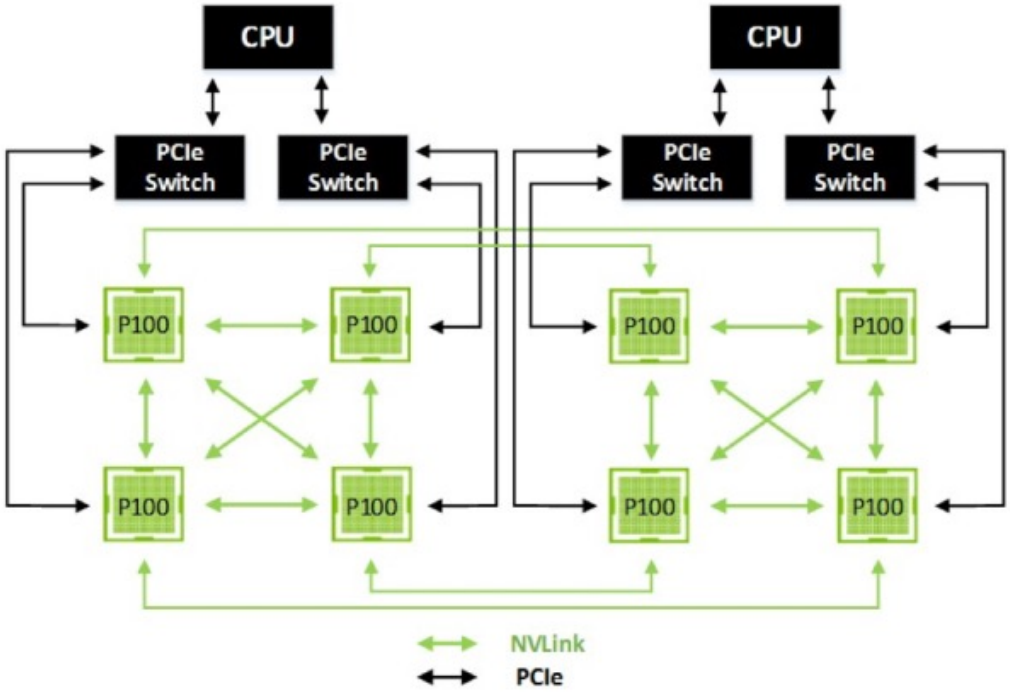
GPU: training a DNN on multiple GPUs



- The performance of such a synchronous system is limited by the slowest learner and slowest messages through the network
- Since the communication phase is in the critical path, a high performance network can enable fast reconciliation of parameters across learners

GPUs within the rack: PCIe AND NVlink

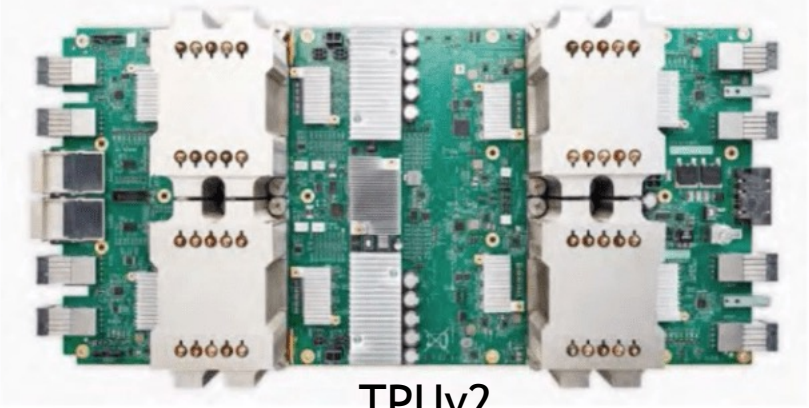
- GPUs are configured with a CPU host connected to a PCIe-attached accelerator tray with multiple GPUs
- GPUs within the tray are connected using high-bandwidth interconnects such as NVlink



Tensor Processing Unit (TPU)

- While suited to ML, **GPUs are still relatively general purpose devices**
- In recent years designers further specialized them to ML-specific hardware
  - **Custom-built integrated circuit** developed specifically for machine learning and tailored for **TensorFlow**
- Powering Google data centers since 2015 as well as CPUs and GPUs
- A **Tensor** is an n-dimensional matrix. This is the basic unit of operation in TensorFlow
- **TPUs are used for training and inference**
  - TPUv1 is an inference-focused accelerator connected to the host CPU through PCIe links.
  - Differently, TPUv2, TPV3, and TPV4 focus training and inference

TPUv1

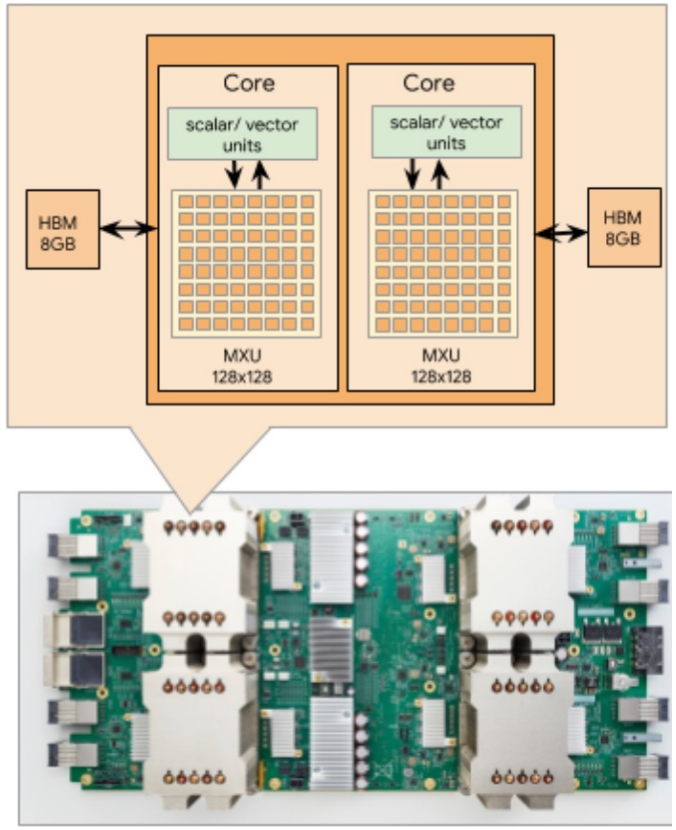


TPUv2

TPUv2

- Each Tensor core has an array for matrix computations (MXU) and a connection to high bandwidth memory (HBM) to store parameters and intermediate values during computation

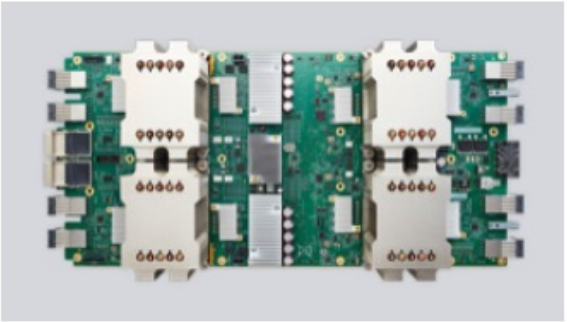
- TPU v2:
  - 8 GiB of HBM for each TPU core,
  - One MXU for each TPU core,
  - 4 chips, 2 cores per chip



TPU v2 - 4 chips, 2 cores per chip

TPUv2 in a Rack (Pod)

- In a rack multiple TPUv2 accelerator boards are connected through a custom high-bandwidth network to provide 11.5 petaflops of ML compute
- The high bandwidth network enables fast parameter reconciliation with well-controlled tail latencies
- Up to 512 total TPU cores and 4 TB of total memory in a TPU Pod (64 units)



Cloud TPU v2

180 teraflops

64 GB High Bandwidth Memory (HBM)

Cloud TPU v2 Pod

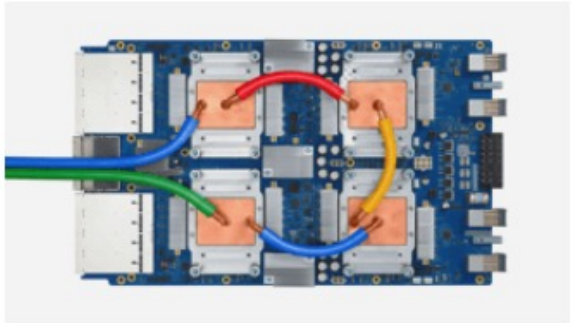
11.5 petaflops

4 TB HBM

2-D toroidal mesh network

TPUv3 (liquid-cooled)

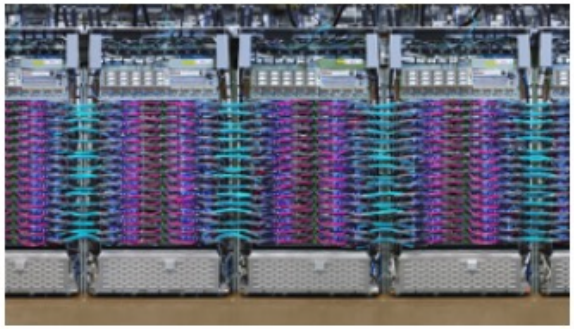
- TPUv3 is the first **liquid-cooled accelerator** in Google's data center
- 2.5x faster than TPUv2
- Such supercomputing-class computational power supports:
  - new ML capabilities (e.g., AutoML)
  - rapid neural architecture search
- The v3 TPU Pod provides a maximum configuration of 256 devices for a total 2048 TPU v3 cores, 100 petaflops and 32 TB of TPU memory



Cloud TPU v3

420 teraflops

128 GB HBM



Cloud TPU v3 Pod (beta)

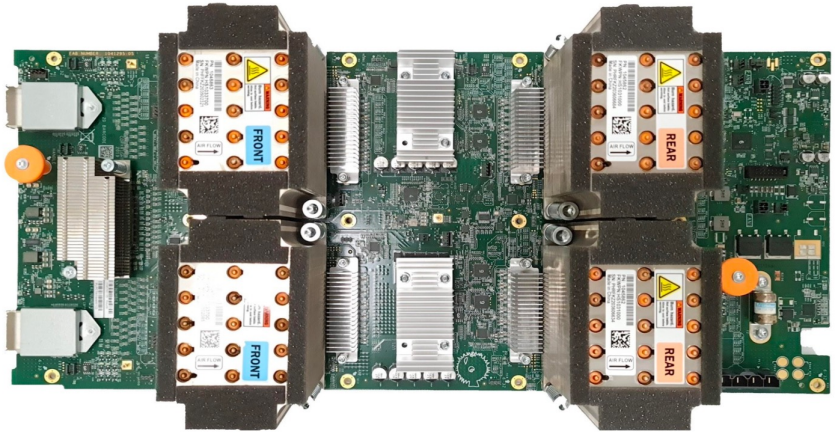
100+ petaflops

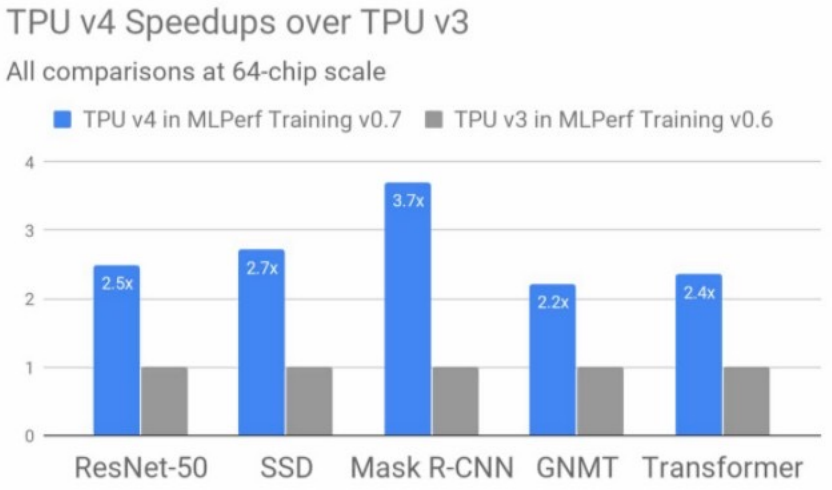
32 TB HBM

2-D toroidal mesh network

TPUv4

- TPUv4 announced June 2021, used to support Google services (not yet available as a cloud service)
- One v4 TPU pod includes 4096 devices
  - About 2.7x faster than TPUv3
  - Same computing capacity as 10 millions of laptops
- MLPerf on BERT:
  - 1.82 minutes with 256 TPUv4
  - 0.39 minutes with 4096 TPUv3
  - 0.81-minutes with 2048 Nvidia A100 (and 512 AMD Epyc 7742 CPU cores)

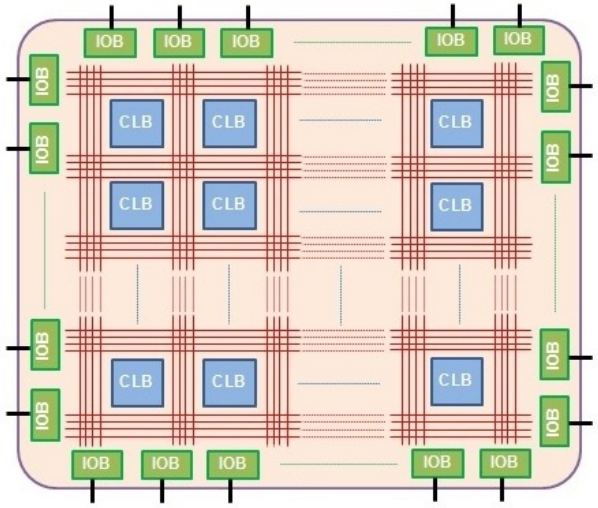




Field-Programmable Gate Array (FPGA)

46

- Array of logic gates that can be programmed (“configured”) in the field, i.e., by the user of the device as opposed to the people who designed it
- Array of carefully designed and interconnected digital subcircuits that efficiently implement common functions offering very high levels of flexibility. The digital subcircuits are called configurable logic blocks (CLBs)

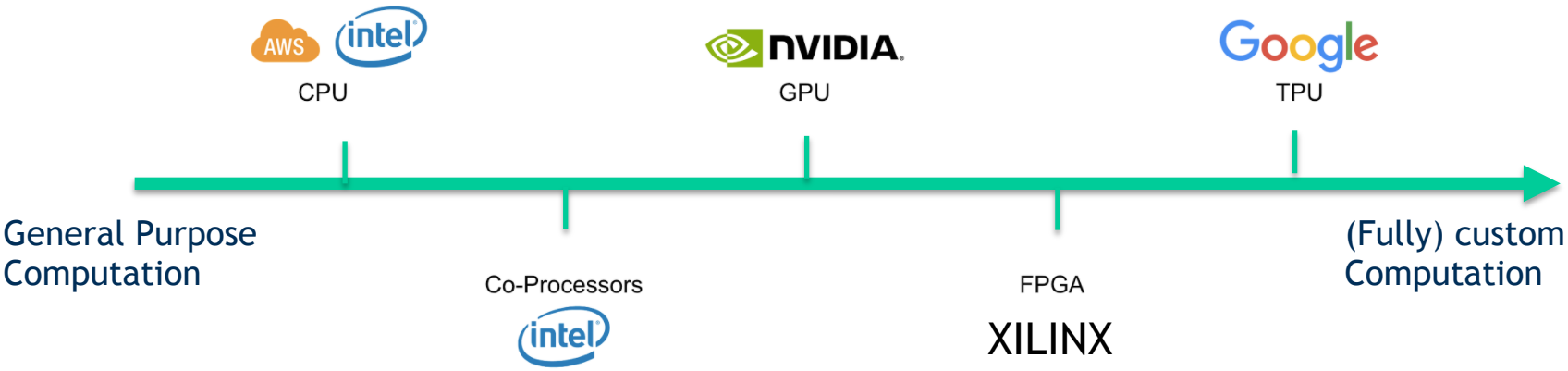


✓ VHDL and Verilog are hardware description languages (HDLs) languages that allow to “describe” hardware;

✓ HDL code is more like a schematic that uses text to introduce components and create interconnections.

GPU, TPU and FPGA: a technological comparison

47

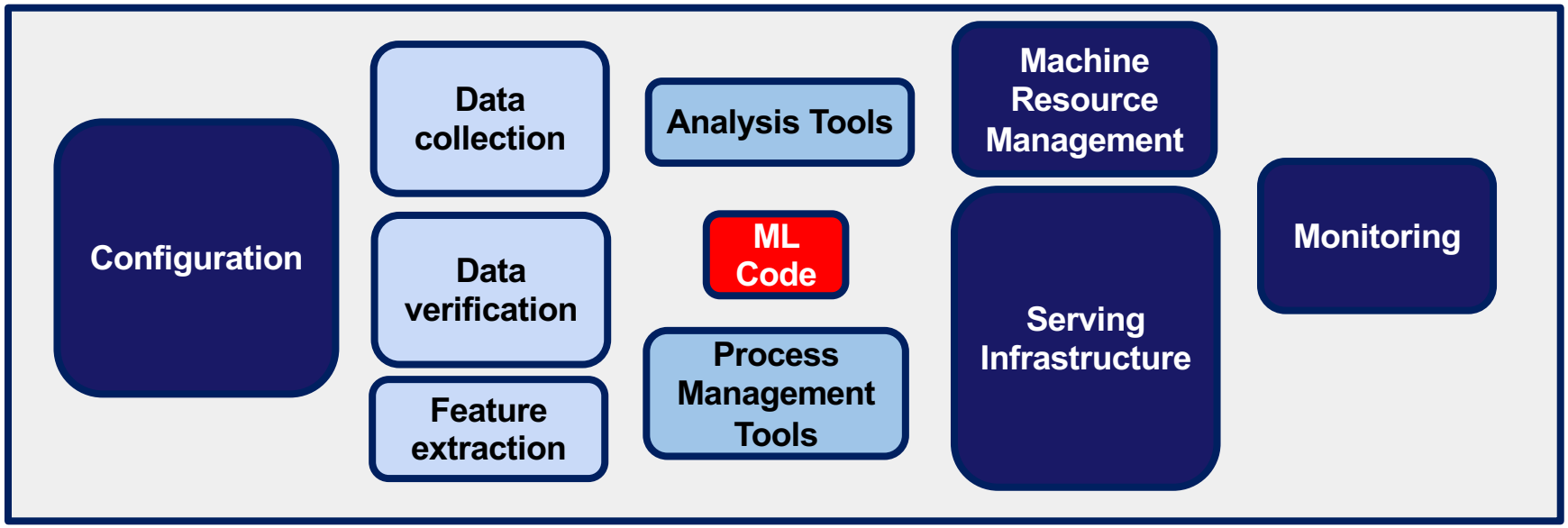


CPU, GPU, TPU and FPGA: an AI comparison

48

	Advantages	Disadvantages
CPU	• Easy to be programmed and support any programming framework. • Fast design space exploration and run your applications.	• Most suited for simple models that do not take long to train and for small models with small training set.
GPU	• Ideal for applications in which data need to be processed in parallel like the pixels of images or videos.	• Programmed in languages like CUDA and OpenCL and therefore provide limited flexibility compared to CPUs.
TPU	• Very fast at performing dense vector and matrix computations and are specialized on running very fast ML workloads	• For applications and models based on TensorFlow/PyTorch/ JAX • Lower flexibility compared to CPUs and GPUs
FPGA	• Higher performance, lower cost and lower power consumption compared to other options like CPUs and GPU	• Programmed using OpenCL and High-level Synthesis (HLS) . • Limited flexibility compared to other platforms.

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