POLITECNICO DI MILANO

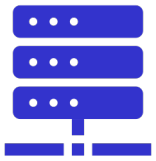
Computing Infrastructures





The Datacenter as a Computer

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, failure recovery

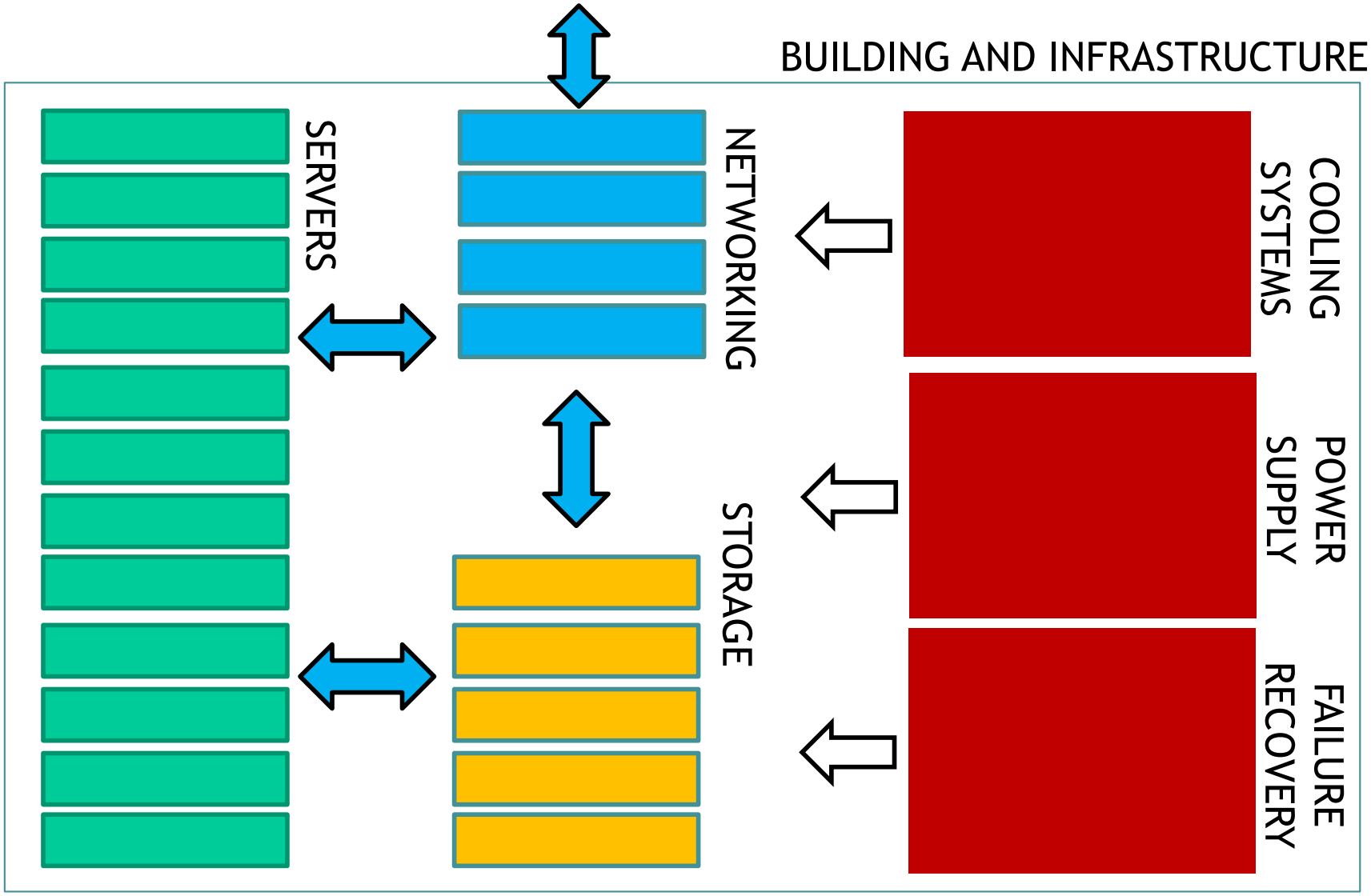
B. SW Infrastructures:

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

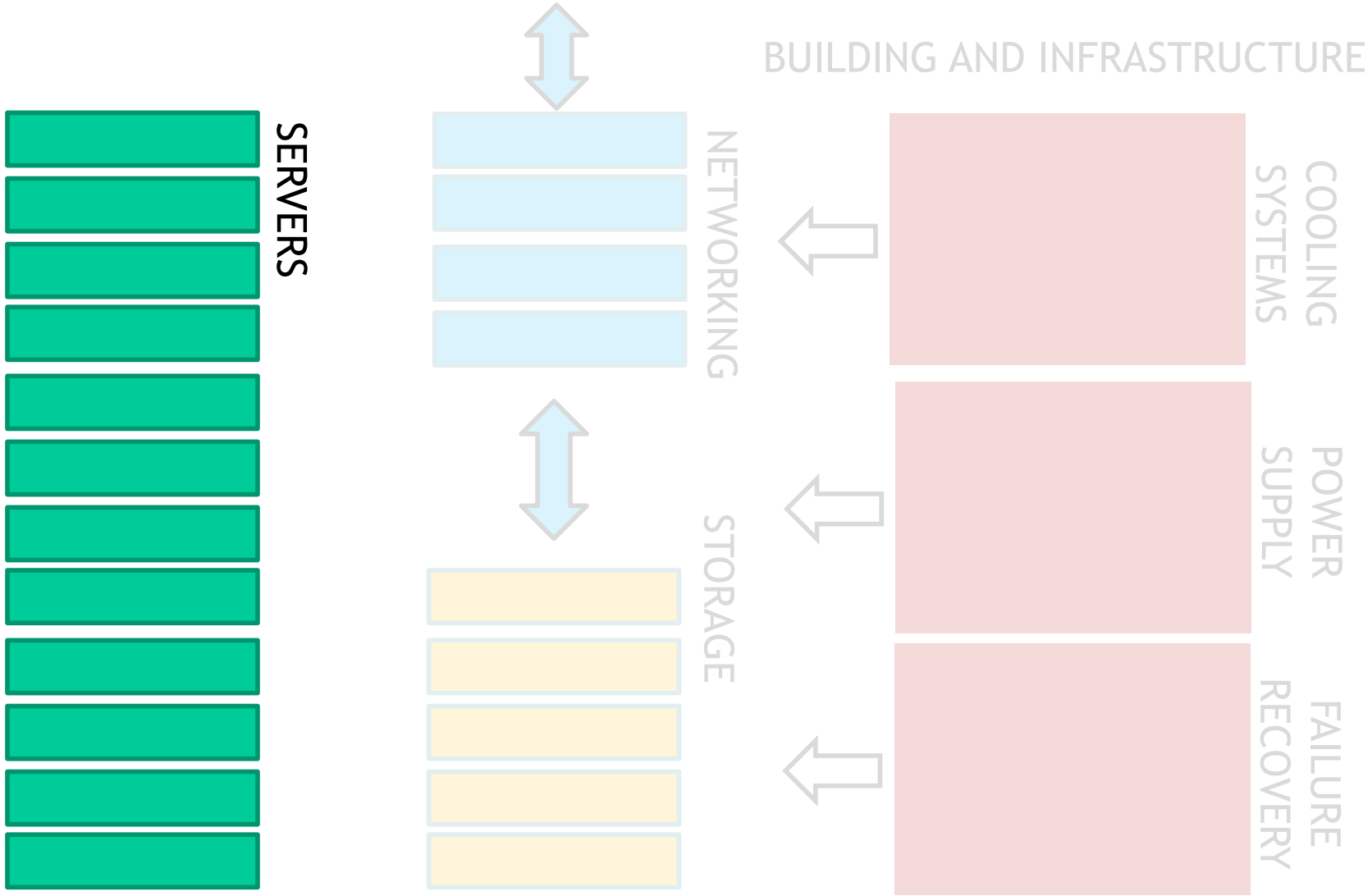
C. Methods:

- **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)

Architectural Overview of A Warehouse-scale Computer



SERVERS

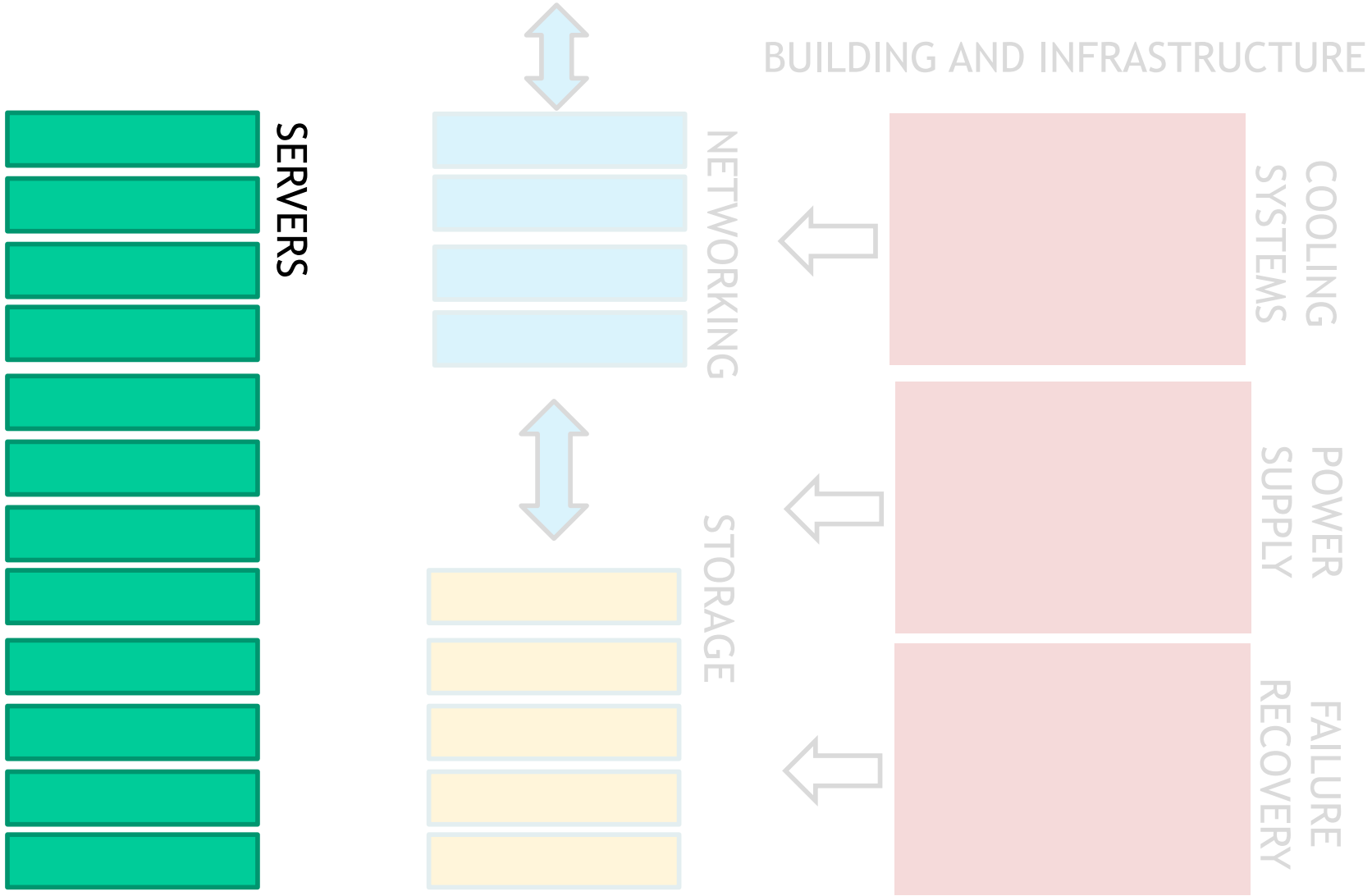

BUILDING AND INFRASTRUCTURE
SERVERS
NETWORKING
STORAGE
COOLING SYSTEMS
POWER SUPPLY
FAILURE RECOVERY

SERVER AND RACK OVERVIEW

- Servers are the main processing equipment
  - Different types according to CPUs, RAM, local storage, accelerators, and form factor
- Servers are hosted in individual shelves and are the basic building blocks of DCs and WSCs
- They are interconnected by hierarchies of networks, and supported by the shared power and cooling infrastructure.

Shelves storing the servers

SERVERS


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

SERVERS: the main processing equipment

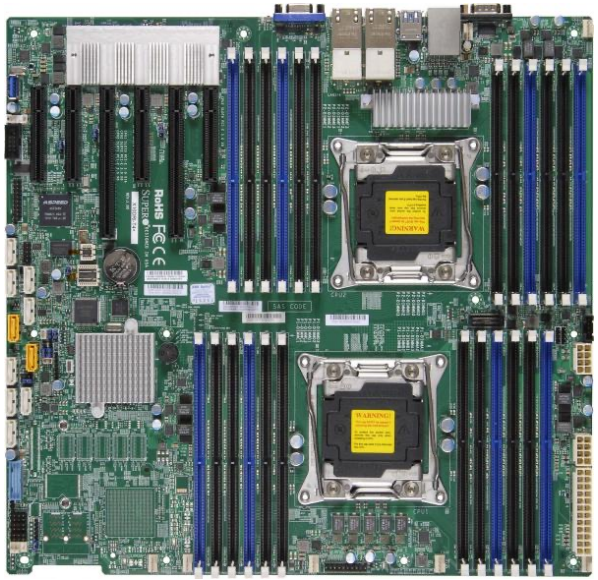
They are like ordinary PC, but with a form factor that allows to fit them into the shelves:

- 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.

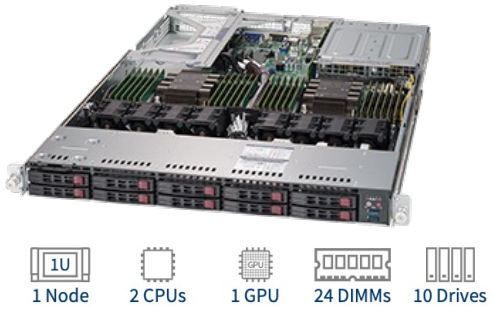


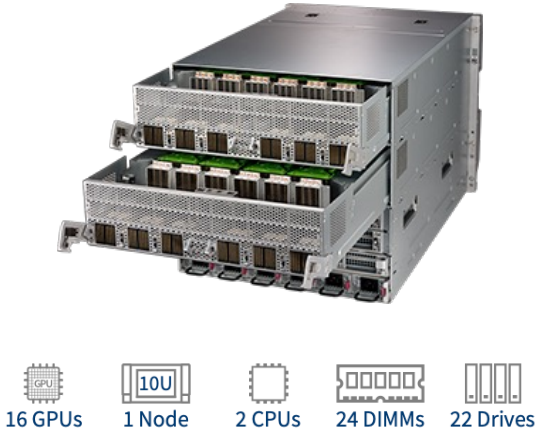
An example: Supermicro Motherboard X10DRi-T4+
Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 v4+/ v3 family; UPI up to 9.6GT/s; Intel® C612 chipset; Up to 3TB+ ECC 3DS LRDIMM, up to DDR4- 2400+MHz ; 24x DIMM slots; 2 PCI-E 3.0 x16, 3 PCI-E 3.0 x8, and 1 PCI-E 2.0 x4 (in x8) slot; Quad 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 Volta, A100, H100, B100, etc.
- ✓ Form factor:
  - From 1U to 10U
  - Tower





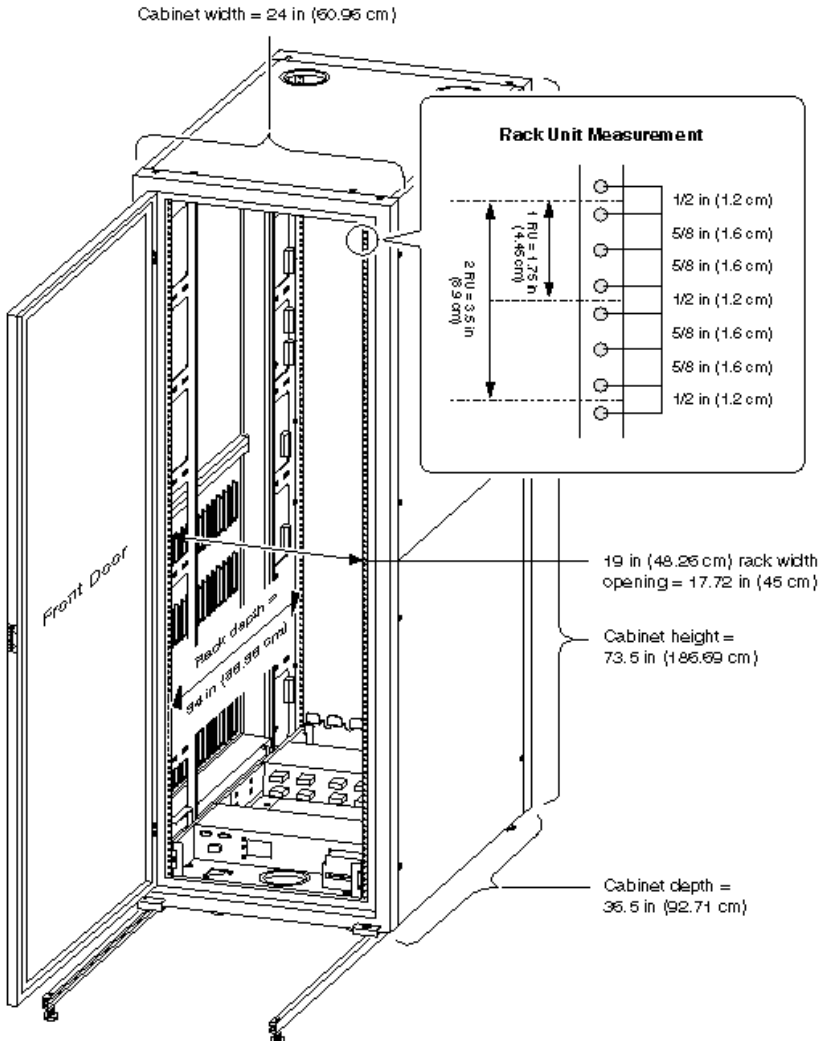
Rack servers

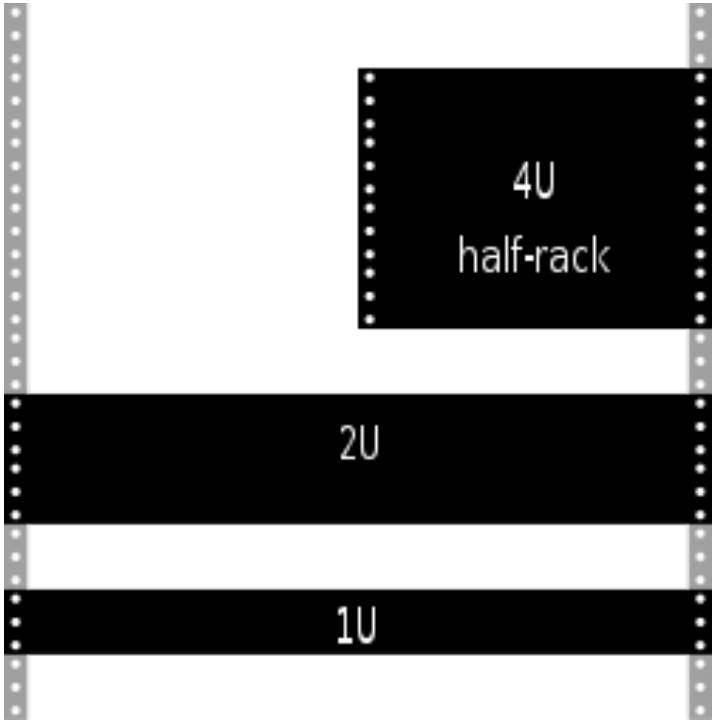
Racks are special shelves that accommodate all the IT equipment and allow their interconnection.

- 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

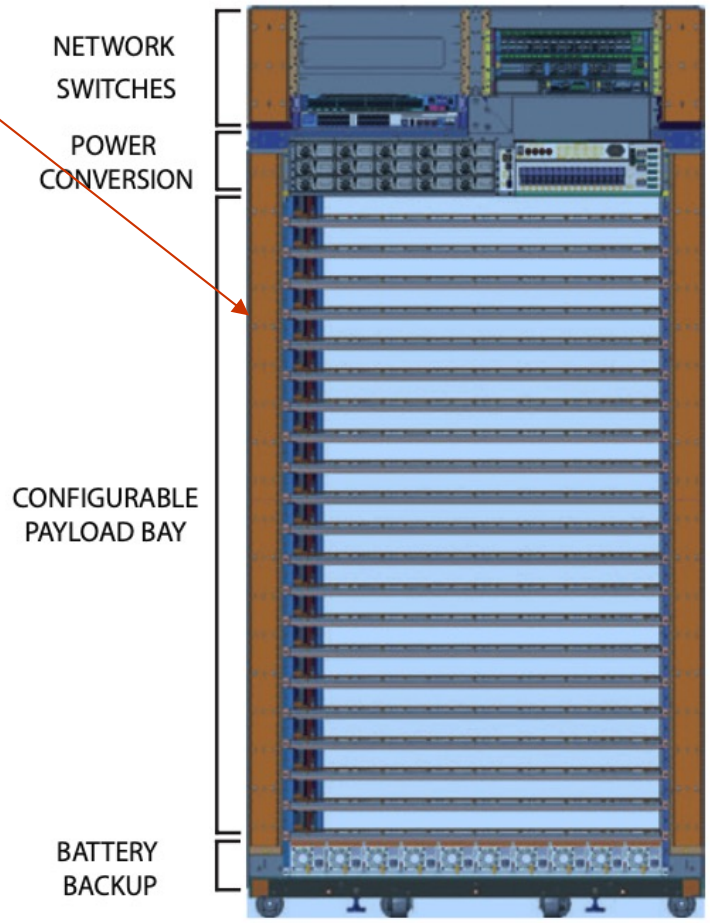


Image taken from **"The Datacenter as a Computer», Barroso et al.**

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

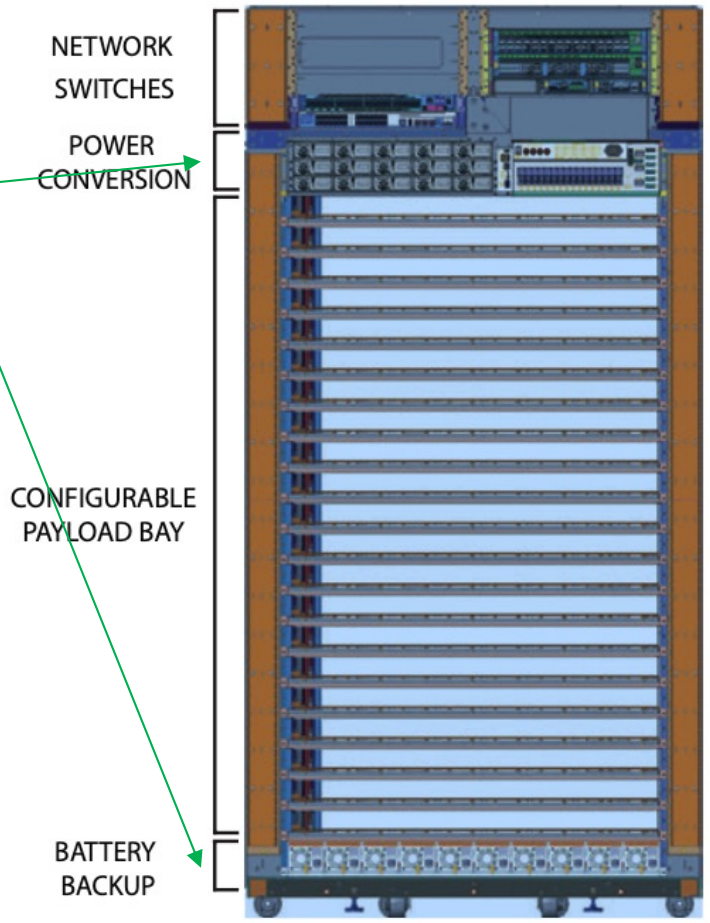


Image taken from **"The Datacenter as a Computer»**, **Barroso et al.**

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

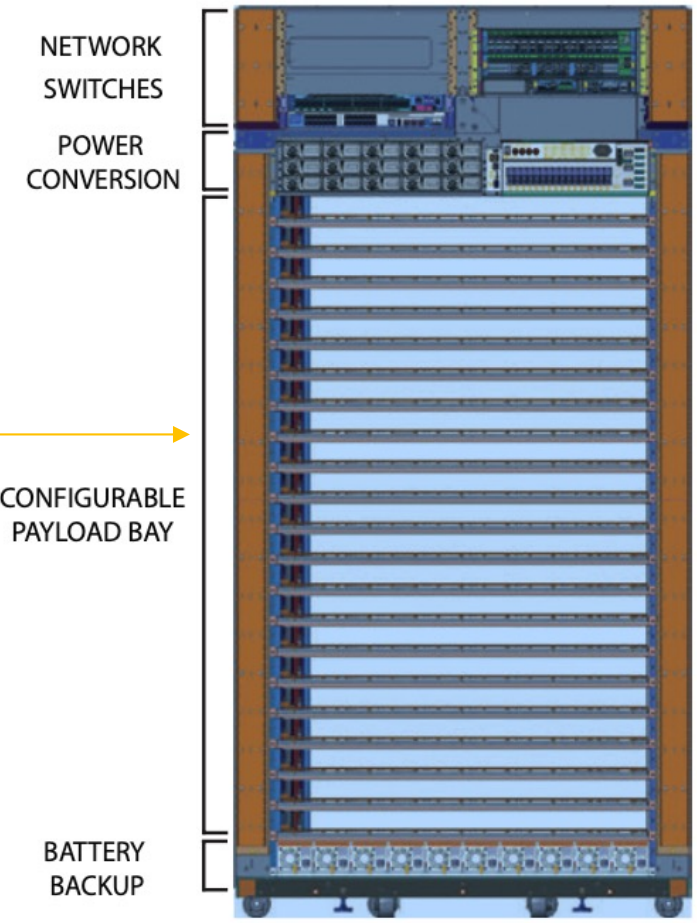


Image taken from **"The Datacenter as a Computer», Barroso et al.**

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

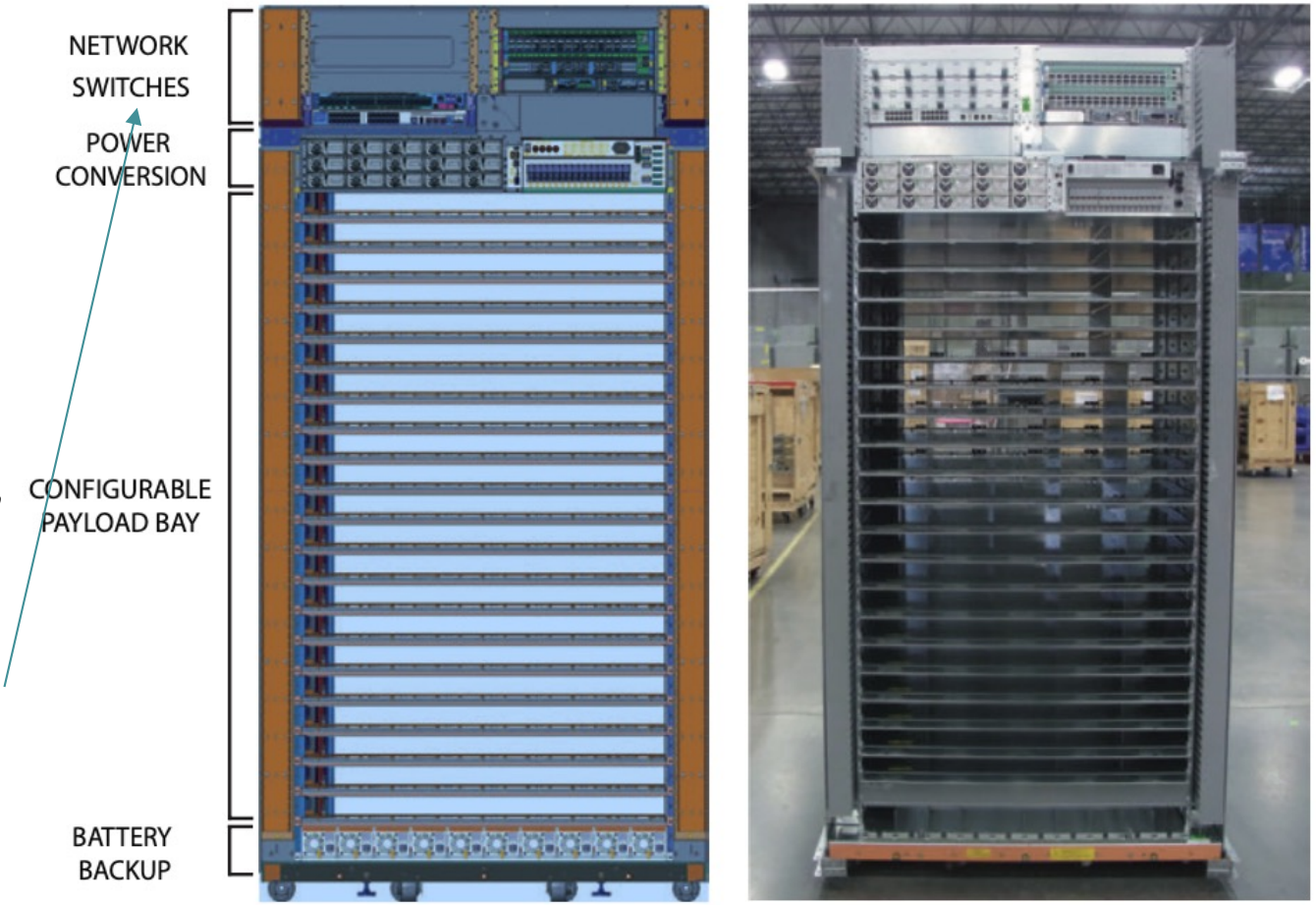


Image taken from **"The Datacenter as a Computer»**, **Barroso et al.**

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

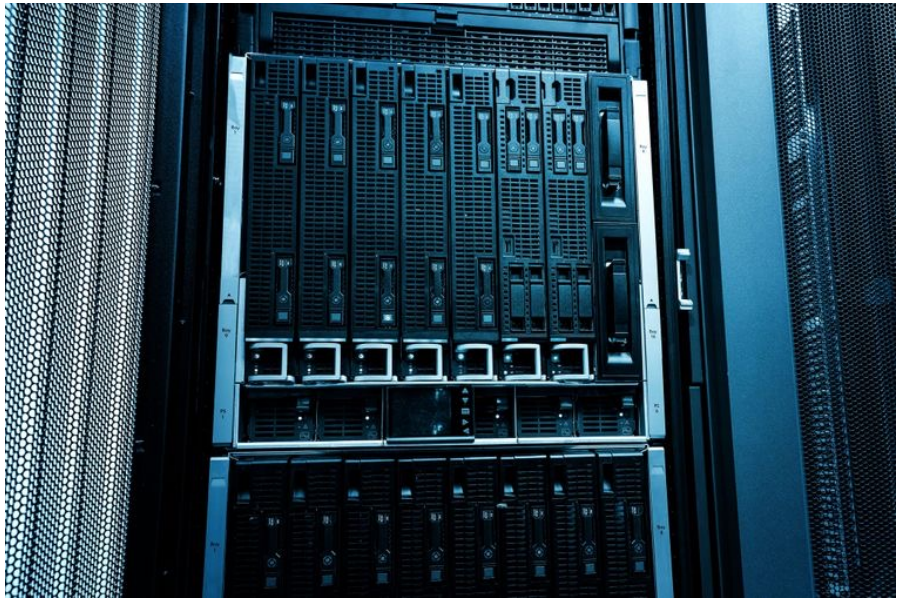
- ✓ **Failure containment:** very little effort to identify, remove, and replace a malfunctioning server with another.
- ✓ **Simplified cable management:** easy and efficient to organize cables.
- ✓ **Cost-effective:** Computing power and efficiency at relatively lower costs.

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

- ✓ **Size and form-factor:** They are the smallest and the most compact servers, requiring minimal physical space.
- ✓ **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.
- ✓ **Centralized management:** In a blade server, you can connect all the blades through a single interface, making the maintenance and monitoring easy.
- ✓ **Load balancing, failover, scalability:** Uniform system, shared components (including network), simple addition/removal of servers

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.
- ✓ **Vendor Lock-In**: Blade servers typically require the use of the manufacturer specific blades and enclosures, leading to vendorlock-in. This can limit flexibility and potentially increase costs in the long run.
- ✓ **Cooling:** 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.

An example of a server for WSCs

Descrizione	Prezzo Unit.	Q.tà	Prezzo Tot.
RB120: Server 1U Dual Socket Intel GPU – 2 bays SAS/SATA + 2 internal			

1 x 1U – 2 x SAS/SATA – Ridondante 2000W
1U Rackmount Black Chassis. 2000W Redundant Power Supplies. 43mm (H) x 437mm (W) x 894mm (D). N. 2 Hot-swap 2.5" SAS/SATA drive bays, n. 2 Internal 2.5" drive bays.

1 x Dual Xeon Scalable – C621 – Server GPU
Proprietary Motherboard. Intel® C621 chipset. Dual Socket P (FCLGA3647). Support up to 165W TDP. N. 12 DIMM Slots supported Memory Types: 2666/2400/2133MHz RDIMM, LRDIMM and 3DS ECC LRDIMM modules. Optimal memory configuration: Six memory channels per CPU.

An example of a server for WSCs

2 x Xeon 8-Core 4110 2,1Ghz 11MB
Intel® Xeon® Silver 4110 Processor. 8Cores. 16Threads. FCLGA3647 Socket.11MB L3. 2,1Ghz Base Frequency. 85W max. TDP. DDR4-2400 Memory type.

6 x DDR4-2666 Reg. ECC 16 GB module
Full brand memory, tested and certified by manufacturer for thorough compatibility with proposed system. The real operating speed depends on the processor's model and on the number of the installed modules. Better performances are achieved through a proper channel configuration.

1 x Intel C621 SATA III 4 ports #

1 x SEAGATE 2TB 2,5" SATA III 7.200RPM
Seagate Enterprise Capacity hard disk drive. Form factor: 2,5''. Capacity: 2TB. Interface: 512N SATA 6Gb/s. Buffer: 128MB. Rotational Speed: 7200RPM. Max. Sustained Transfer Rate (MB/s): 136MB/. 2Million-hour MTBF.

1 x Intel S4500 240GB 2,5" SSD SATA III
Intel® SSD DC S4500 Series. Sequential Read (up to): 500MB/s. Sequential Write (up to): 190MB/s. Random Read (100% Span): 69000 IOPS. Random Write (100% Span): 16000 IOPS. Endurance Rating (Lifetime Writes): 0.62 PBW.

1 x Backplane 2 bays SAS/SATA

An example of a server for WSCs

1 x Backplane 2 bays SAS/SATA

1 x Internal Cage SAS/SATA 2 dischi

1 x BMC Integrated Aspeed AST2500 #
AST2500 embedded on chipset graphic card integrated on the motherboard.

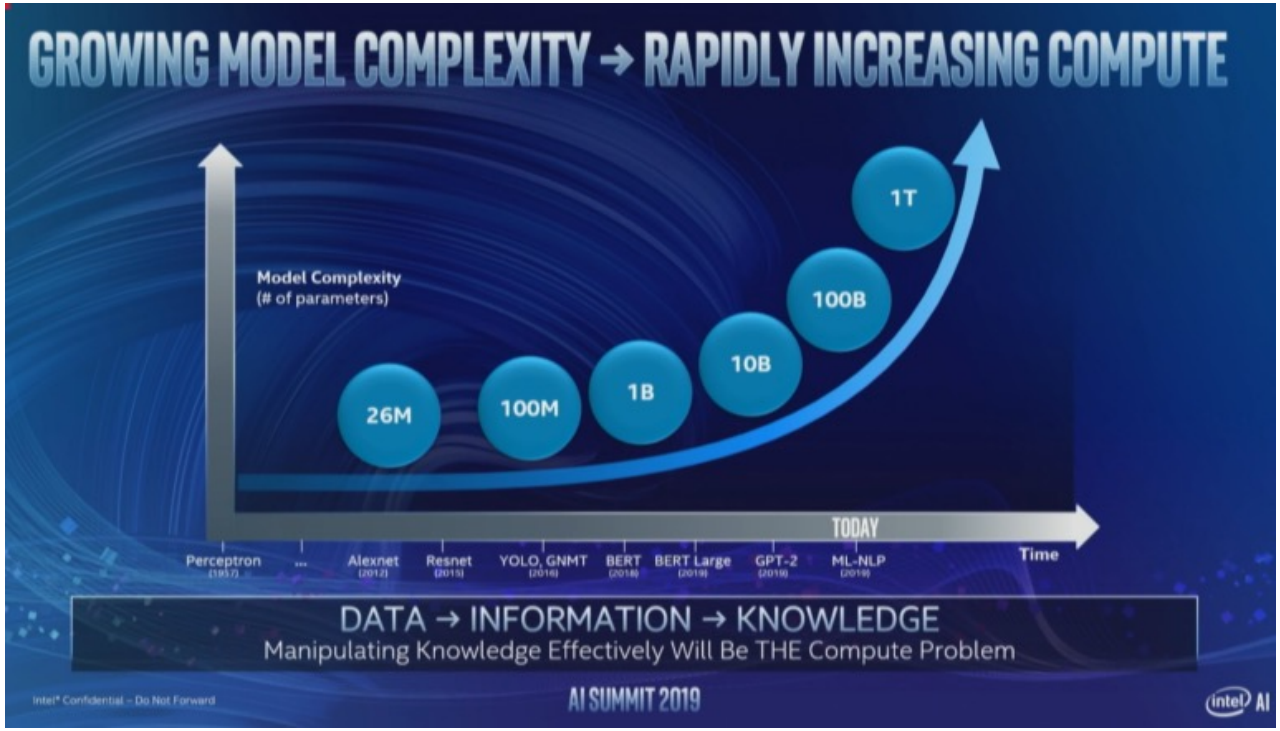
Please Note: this graphic card will not work if you install another graphic card.

1 x 24784 NVIDIA P100 16GB

1 x Aspeed AST2500 10/100 #
Dedicated LAN 10/100Mb/s for management (IPMI).

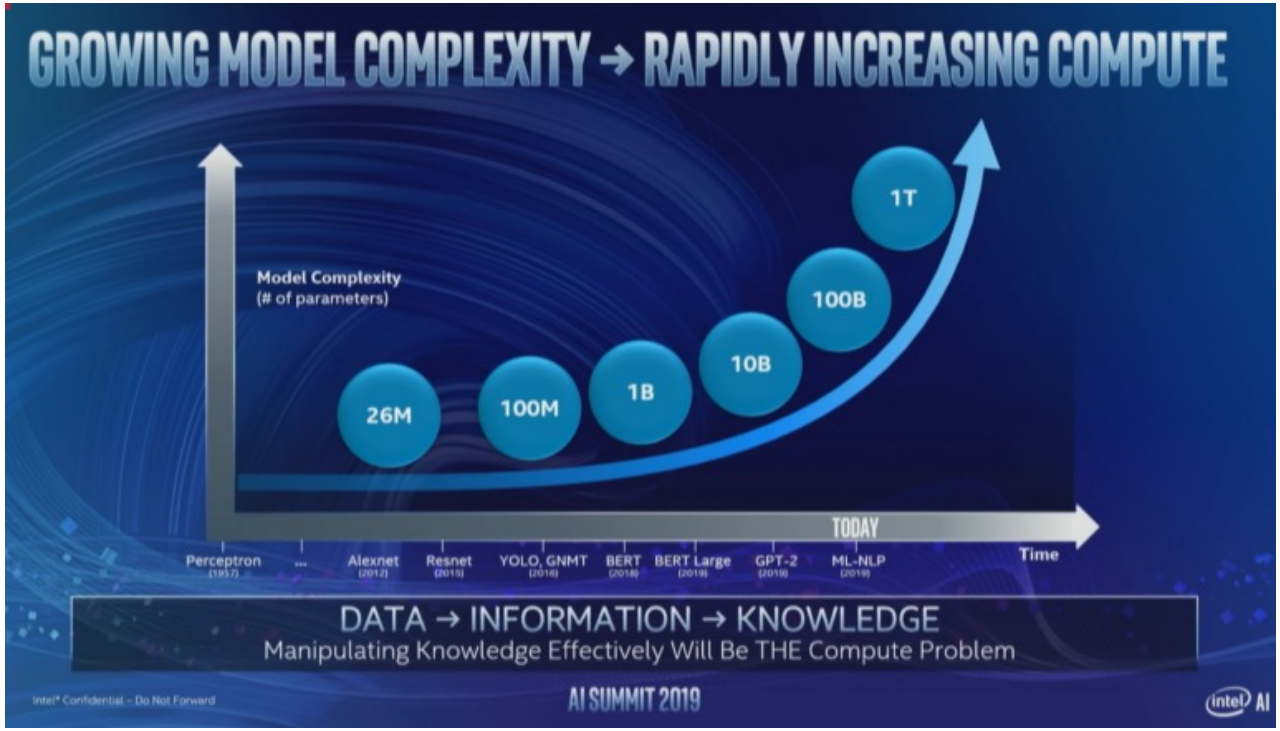
2 x Intel 10 Gigabit 10GBase-T #
Intel 10 Gigabit X540 Network Adapter, RJ45 10GBase/T interface. Integrated on the mother board.

The need of hardware accelerators



Complexity doubles every 3.5 months

The need of hardware accelerators



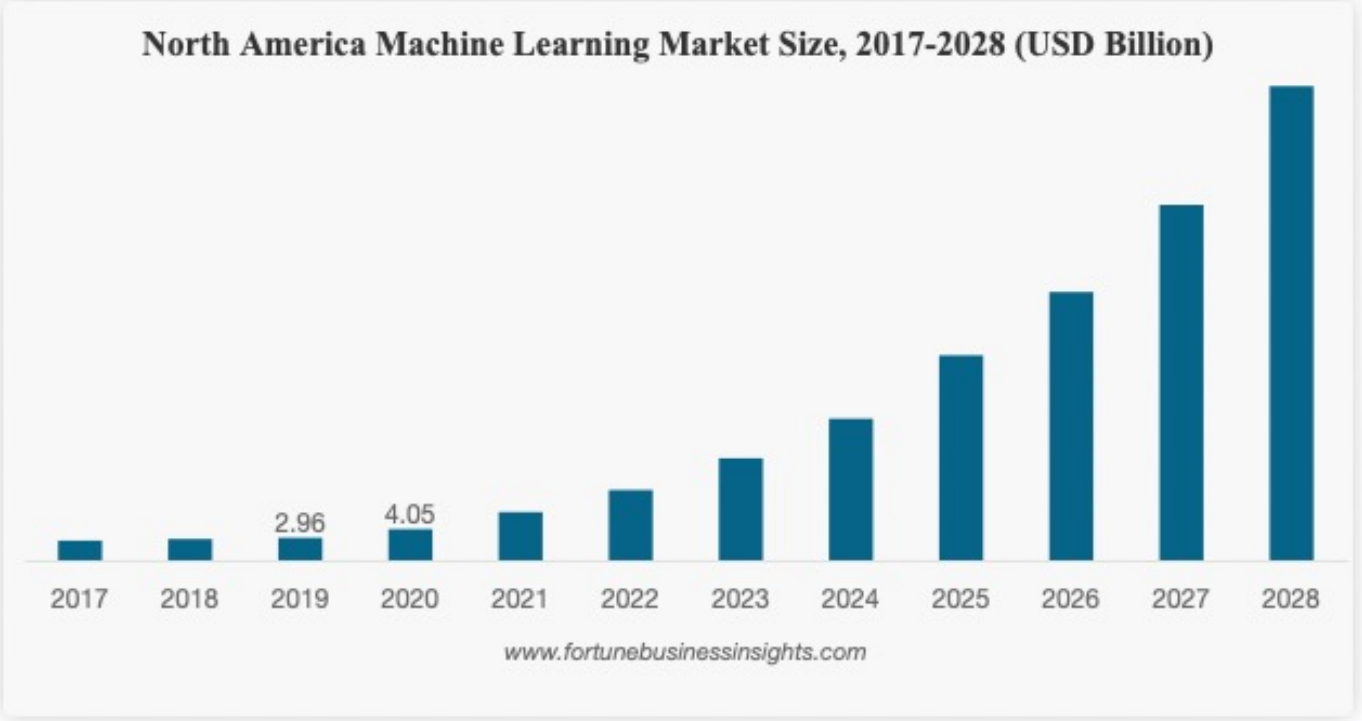
Complexity doubles every 3.5 months

18-24 months for Moore's Law

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

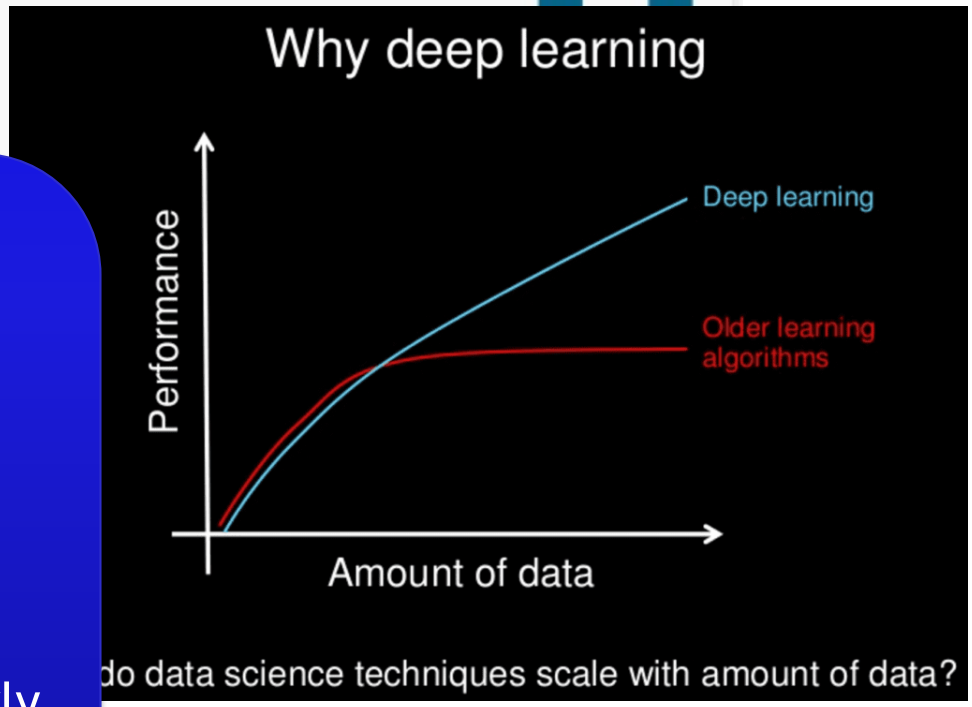
Machine Learning Market Growth

- 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

Three observations by Sam Haltman

1. The intelligence of an AI model roughly equals the log of the resources used to train and run it
2. The cost to use a given level of AI falls about 10x every 12 months
3. The socioeconomic value of linearly increasing intelligence is super-exponential in nature

What is Machine Learning?
(basic definition, this is not a ML course)

Learn from data through models

Learn with no explicit programming ➡ learn from features

Discover hidden patterns

What do we mean by learning?

Given:

- a data set **D**
- a task **T**
- a performance measure **M**

A computer system is said to learn from **D** to perform the task **T** if after learning the system performance on **T** improves as measured by **M**

In other words, the learned model helps the system to perform T better as compared to no learning

Machine learning (supervised learning)

- Humans learn from **past experiences**
- A computer does not have “experiences”
  - A computer system learns from **data**, which represent some “**past experiences**” of an application domain

- Goal: learn a **target function** that can be used to **predict** the values of
  - a discrete class attribute, e.g., cat or not-cat, approve or not-approved, and high-risk or low risk (discrete world)
  - a continuous value, e.g., flight delays, cash at a bank branch/ATM (continuous setting)

The machine learning framework (classification)

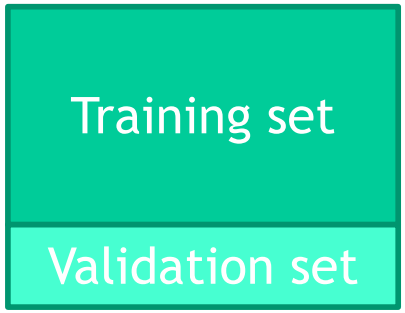
Apply a prediction function to a feature representation of the image to get the desired output:

f() = “apple”

f() = “tomato”

f() = “cow”

The machine learning framework



$$y = f(x)$$

output — prediction function — Image feature

Training: given a *training set* of labeled examples $\{(x_1,y_1), \ldots, (x_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

Testing: apply f to an unseen *test example* x and output the predicted value $y = f(x)$

Steps

Training

Training Images

Training Labels

Image Features → Training → Learned model

Testing

Test Image

Image Features → Learned model → Prediction

Steps

Training

Training Images

Training Labels

Image Features

Training

Learned model

Once you are happy with the testing accuracy, the ML model is put in production to run **inference** (same as the testing stage)

Testing

Test Image

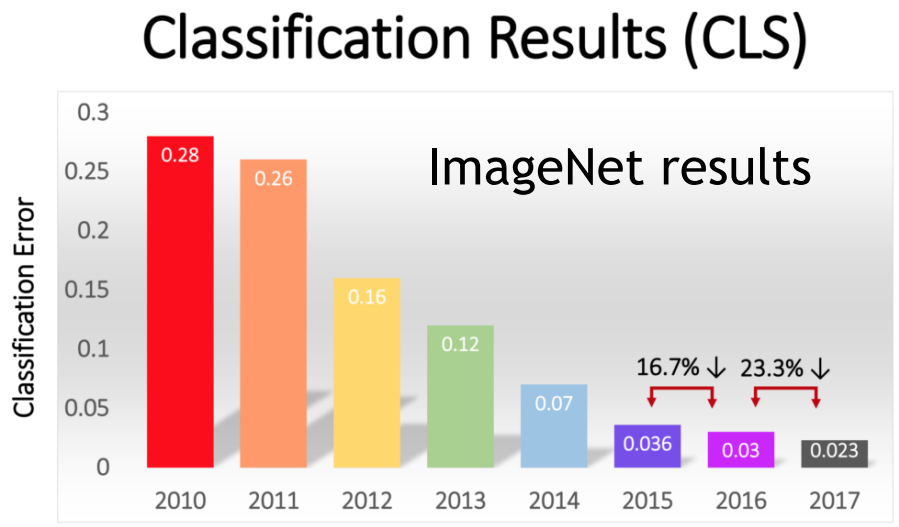
Image Features

Learned model

Prediction

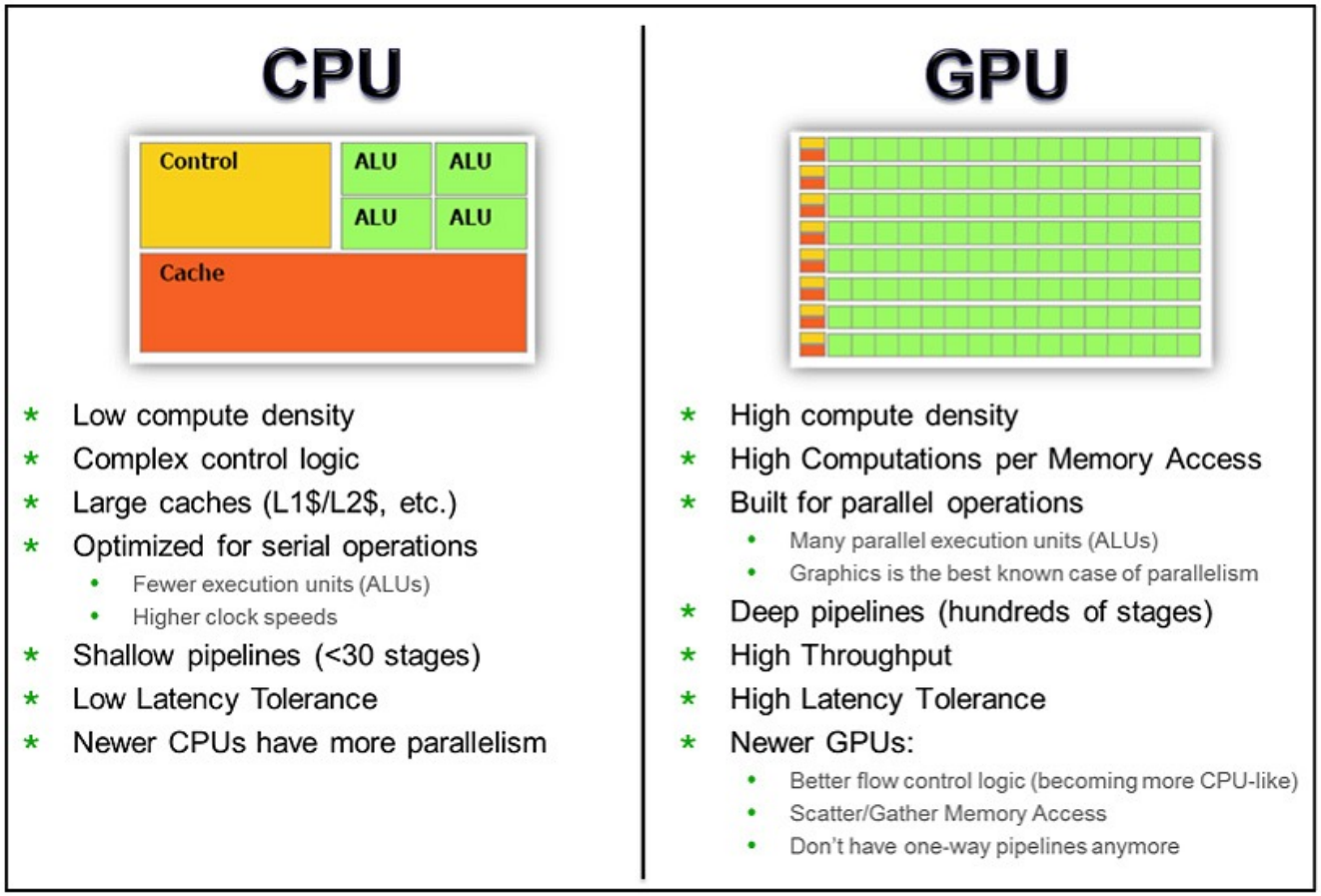
Hardware accelerators

- 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

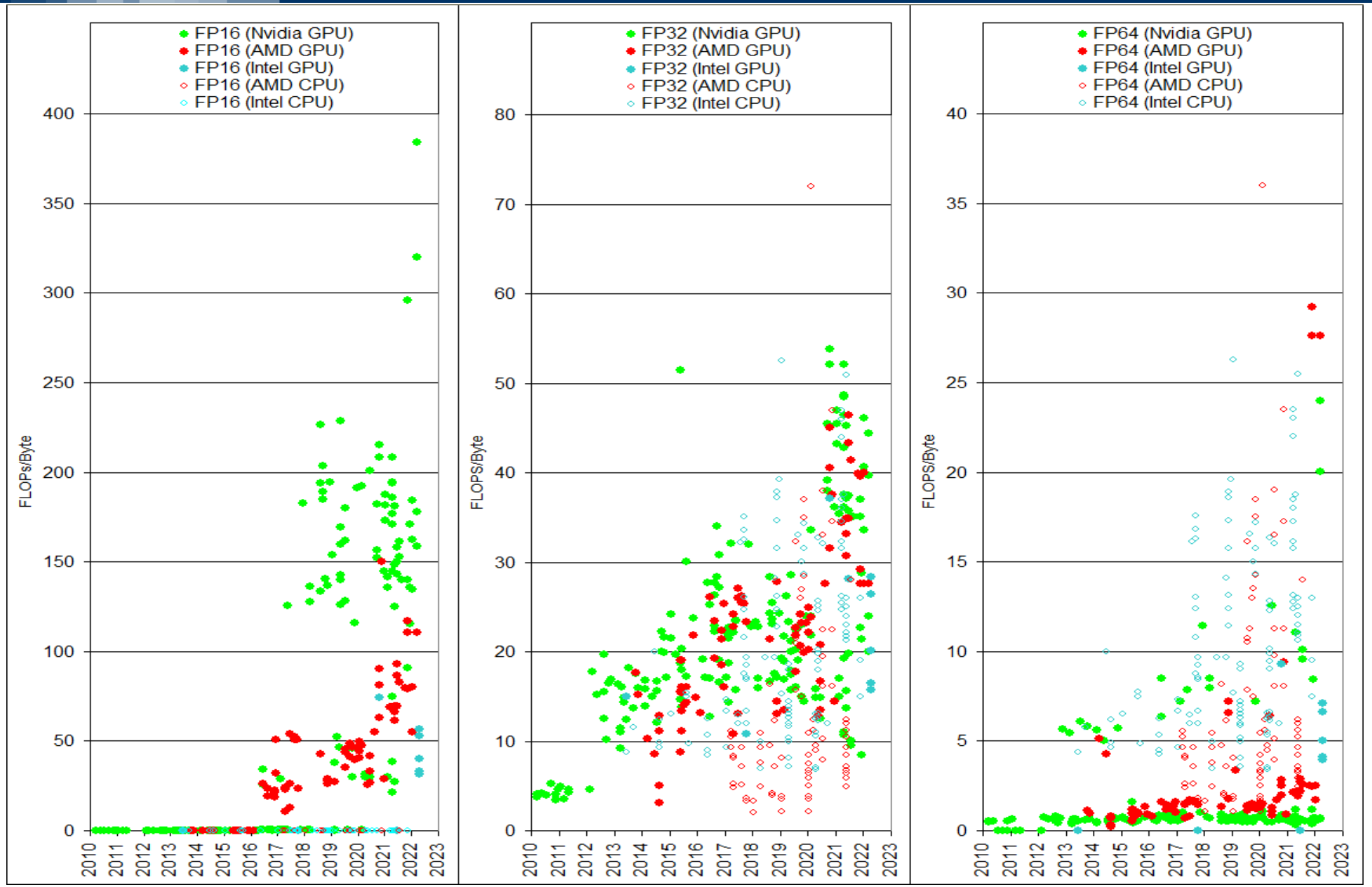


Graphical Processing Units (GPU)

- **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

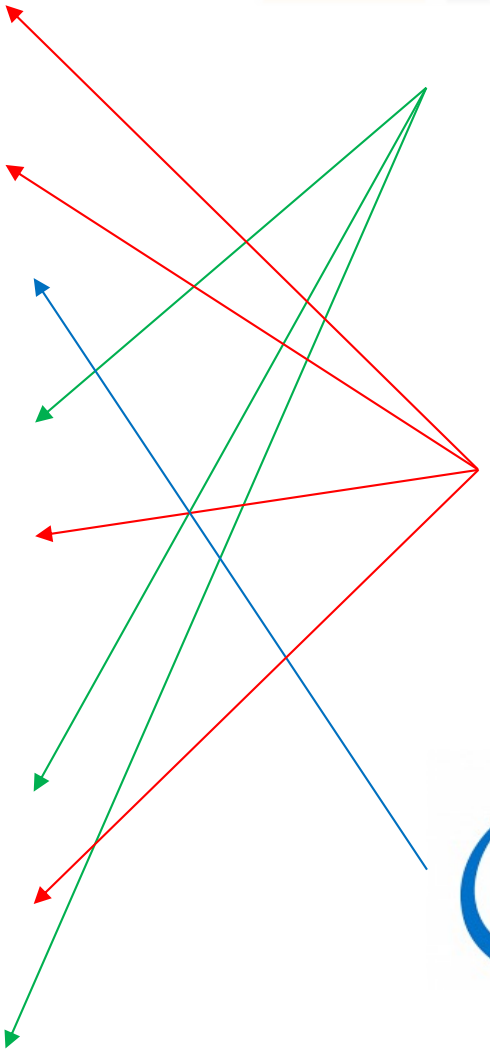


Not only one GPU type and vendor

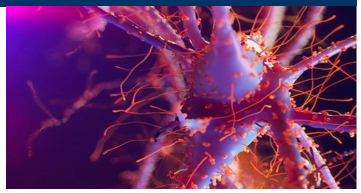


Not only one GPU type and vendor

Rank	System	Cores	(PFlop/s)	(PFlop/s)	(kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
5	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461
6	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
7	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90	574.84	7,124
8	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107
9	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	241.20	306.31	7,494



What is a Neural Network?



Definition:

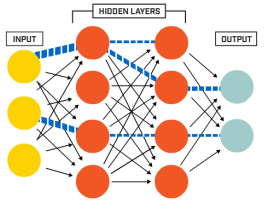
- A computational model inspired by the human brain (perceptron)
- Consists of interconnected nodes (neurons) organized in layers to process and analyze data
- Used to learn data representation from data (learn features and the classifier/regressor)

Brief history

- Neural networks have a rich history dating back to the 1940s
- Notable developments in the 1980s
- Resurgence in recent years (Among 10 Breakthrough in 2013[1] : data availability, computational power)

https://www.technologyreview.com/10-breakthrough-technologies/2013/

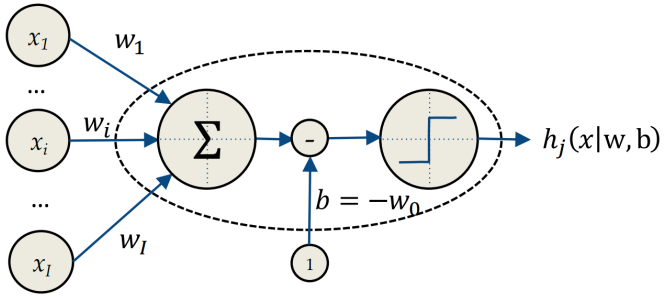
Anatomy of a Neural Network (1)

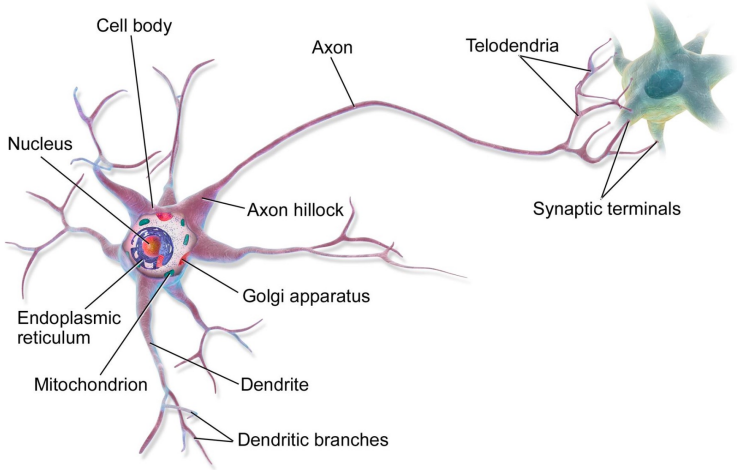


Brains Neurons:

- Information is transmitted through chemical mechanisms
- Dendrites collect charges from synapses, both Inhibitory and Excitatory
- Cumulates charge is released (neuron fires) once a Threshold is passed

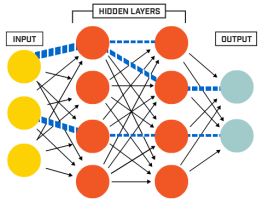
Artificial Neurons





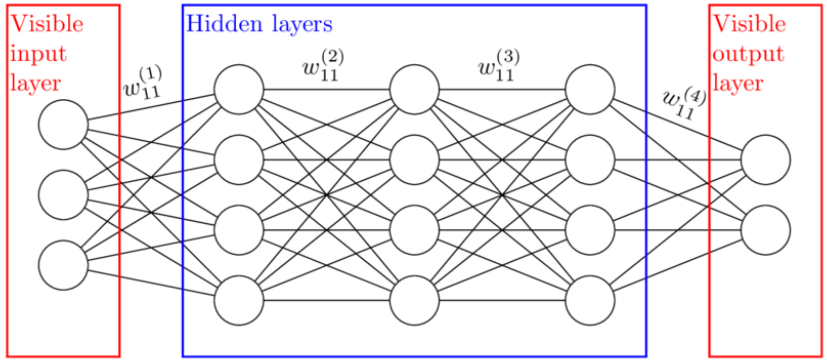
$$h_j(x|\mathrm{w},\mathrm{b}) = \mathrm{h_j}\left(\Sigma_{i=1}^{I} w_i \cdot x_i - b\right) = \mathrm{h_j}\left(\Sigma_{i=0}^{I} w_i \cdot x_i\right) = h_j(w^T x)$$

Anatomy of a Neural Network (2)



Neural Networks:

- Input layer: Where data is introduced
- Hidden Layers: Intermediate layers that process data
- Output layers: Provide the final results

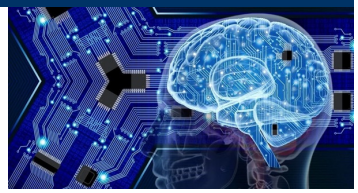


Considerations:

- Layers are connected through weights
- Activation functions must be differentiable
- The output of a neuron only depends on the previous

Non-linear model characterized by the number of neurons, activation functions, and the values of the weights.

How Neural Networks Learn?

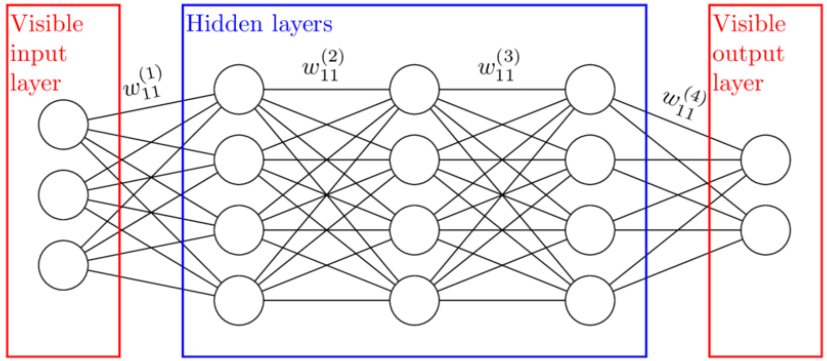


Learning process:

- Neurons make decisions (activation functions)
- Weights : connections between neurons are strengthened or weakened through training

Training data

- NN learn from historical data and examples



Backpropagation (Gradient descent, Chain rule)

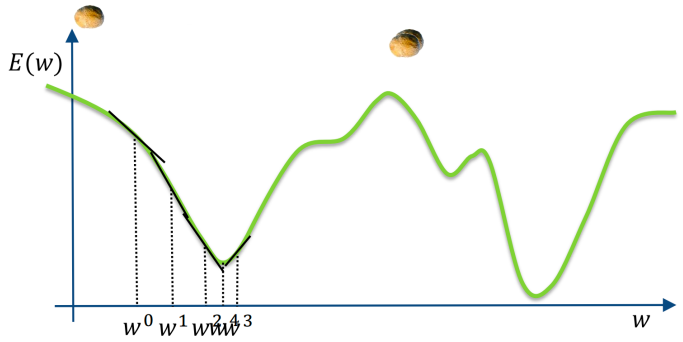
- Errors are calculated and used to adjust the model

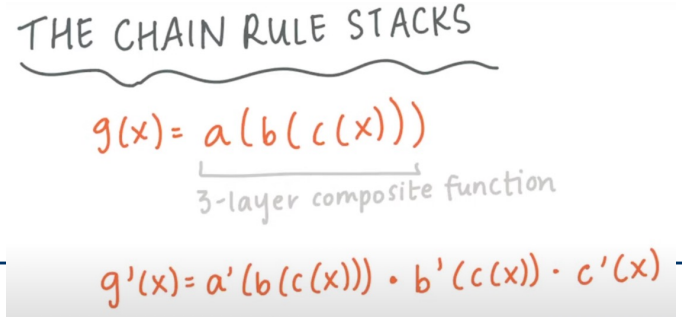
$$E = \sum_n^N \left(t_n - g(x_n|w)\right)^2$$

$$w^{k+1} = w^k - \Delta \frac{\partial E(w)}{\partial w}$$

t_n: *desired target*

$g(x_n|w)$: *learned from data*





Types and Applications of Neural Networks

Types of NN

- Feed Forward Neural Network
- Convolutional Neural Networks
- Recurrent Neural Networks
- Transformers (diffusion models, MoE, …)

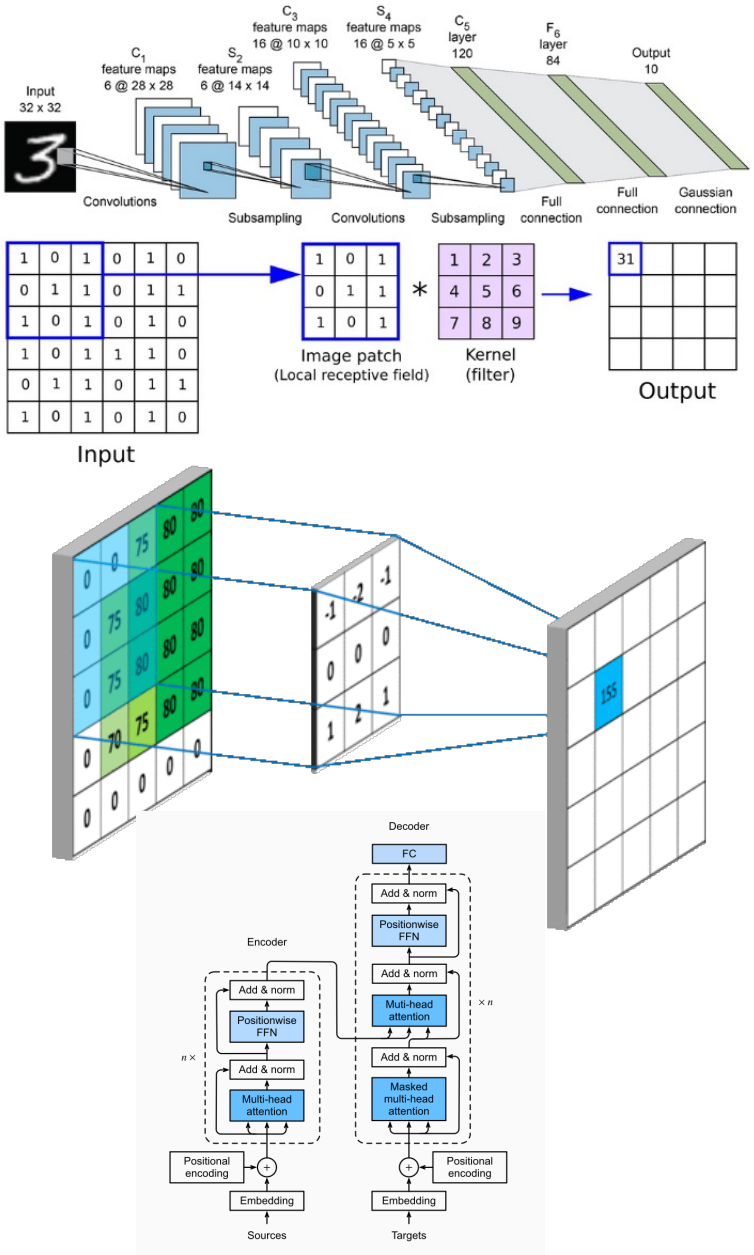
Application areas:

- Image recognition (FaceNet, YOLO, …)
  - Facial recognition, Object detection, Instance segmentation
- Natural language processing (BERT, GPT,…)
  - Chatbot, Sentiment Analysis, Chat-GPT, Speech-to-text

Potential for innovation and future developments:

- Rapid growth of NN applications: Generative AI

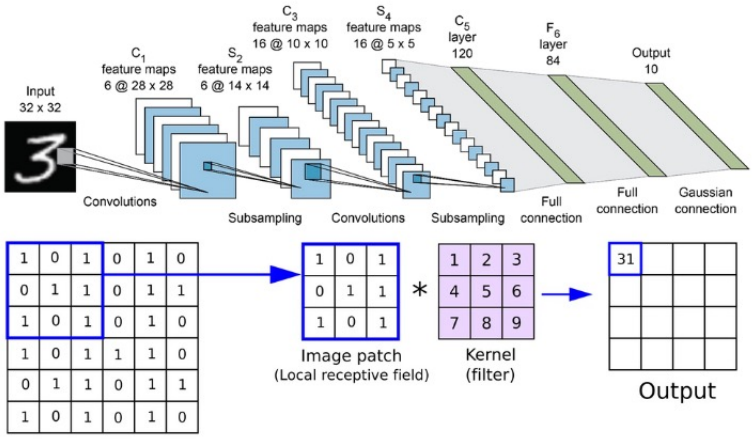
Covers many areas: (Social media, Aerospace, E-commerce, finances,…)



Types and Applications of Neural Networks

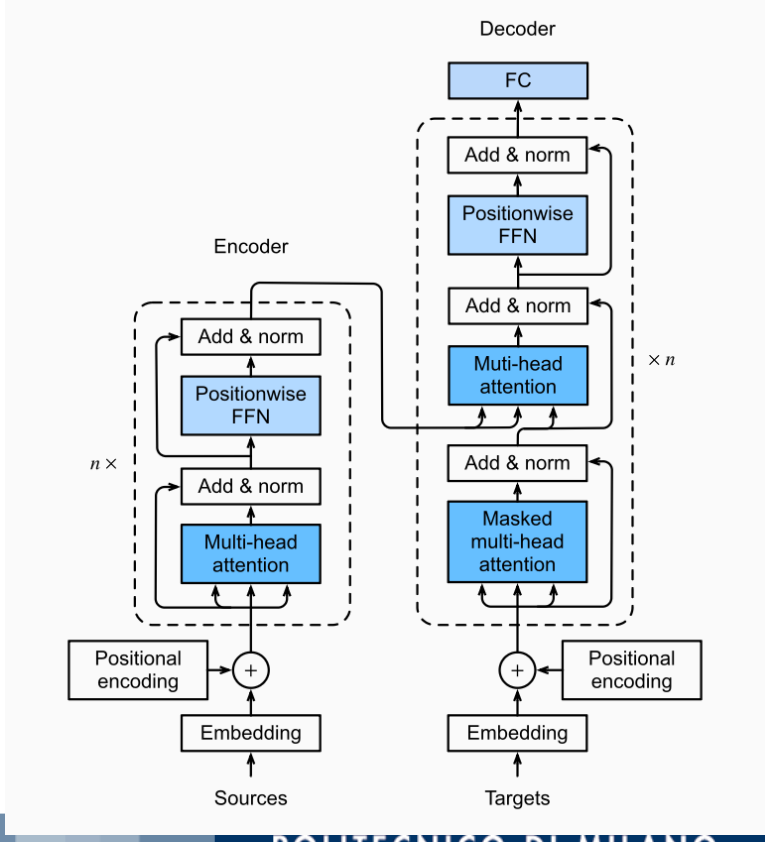
Types of NN

- Feed Forward Neural Network
- Convolutional Neural Networks
- Recurrent Neural Networks
- Transformers (diffusion models, MoE, …)



Application areas:

- Image recognition (FaceNet, YOLO, …)
  - Facial recognition, Object detection, Instance segmentation
- Natural language processing (BERT, GPT,…)
  - Chatbot, Sentiment Analysis, Chat-GPT, Speech-to-text

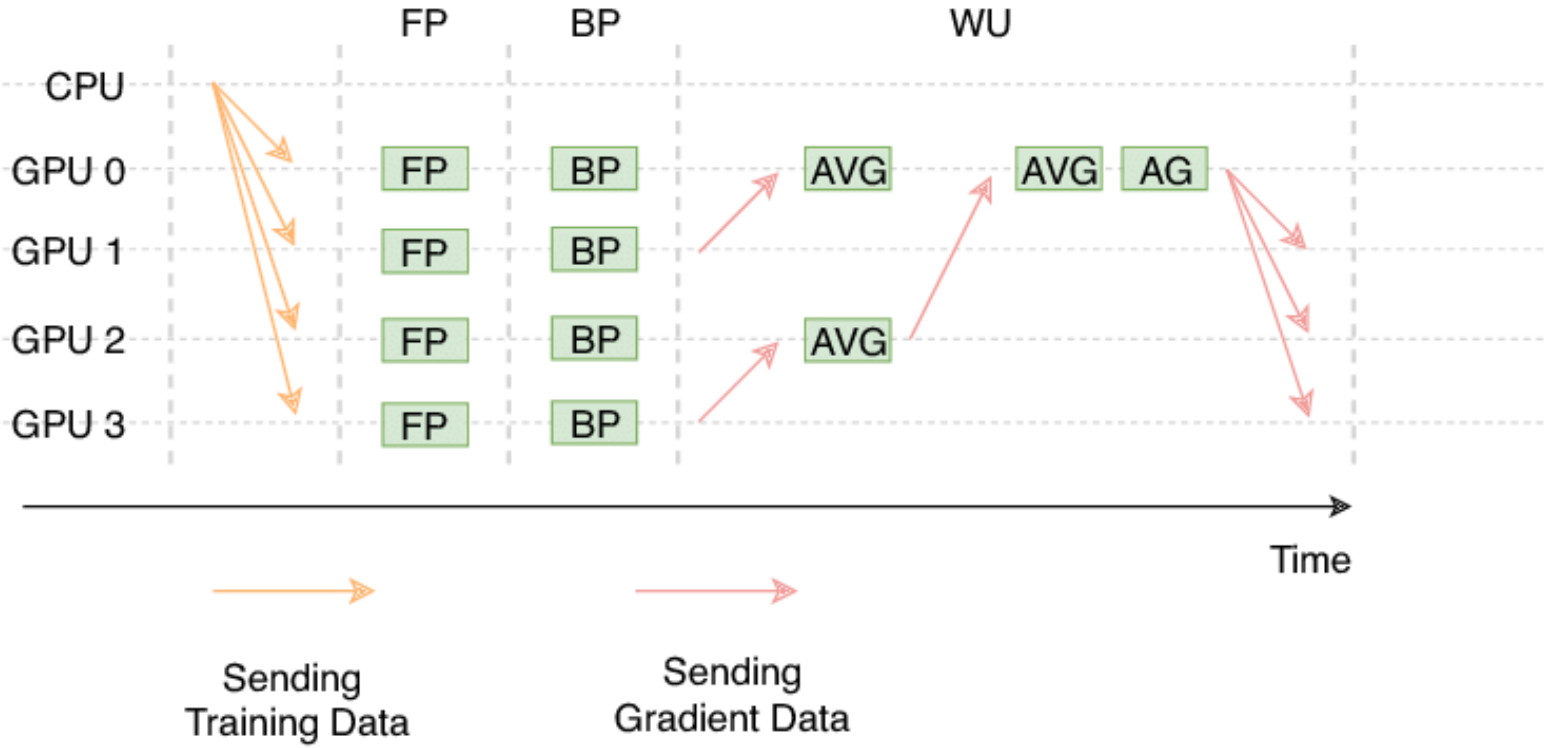


Potential for innovation and future developments:

- Rapid growth of NN applications: Generative AI

Covers many areas: (Social media, Aerospace, E-commerce, finances,…)

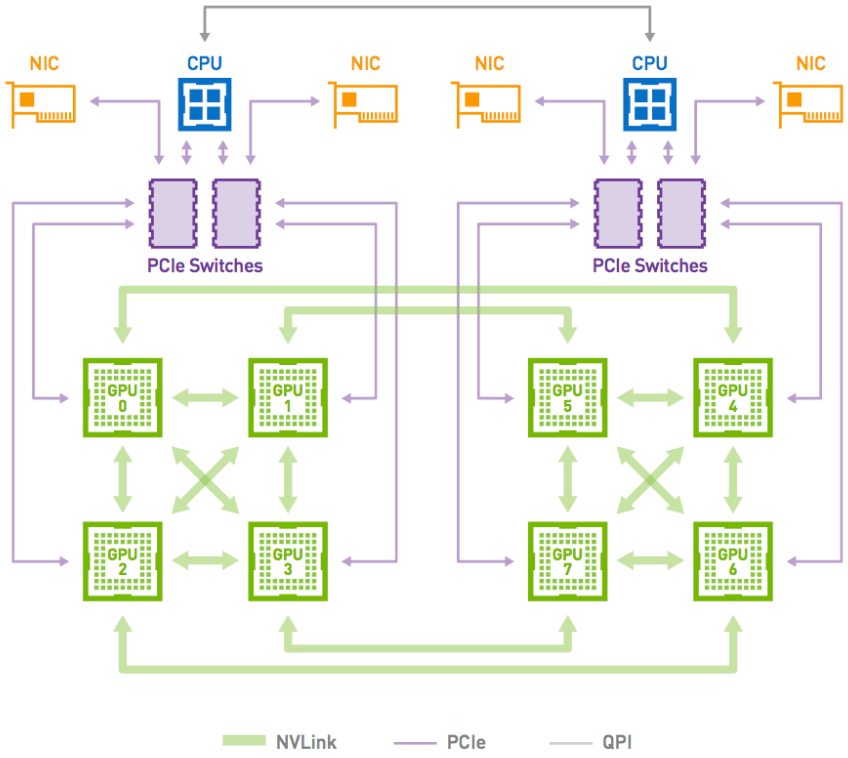
GPU: training a DNN on multiple GPUs

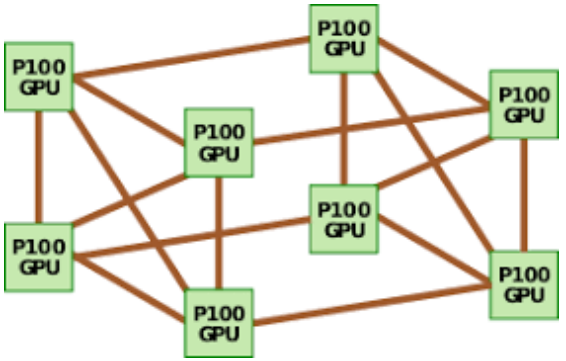


- The performance of such a synchronous system is limited by the slowest learner and slowest messages through the network
- Since the communication phase is in the critical path, a high performance network can enable fast reconciliation of parameters across learners

GPUs within the rack: PCIe AND NVlink

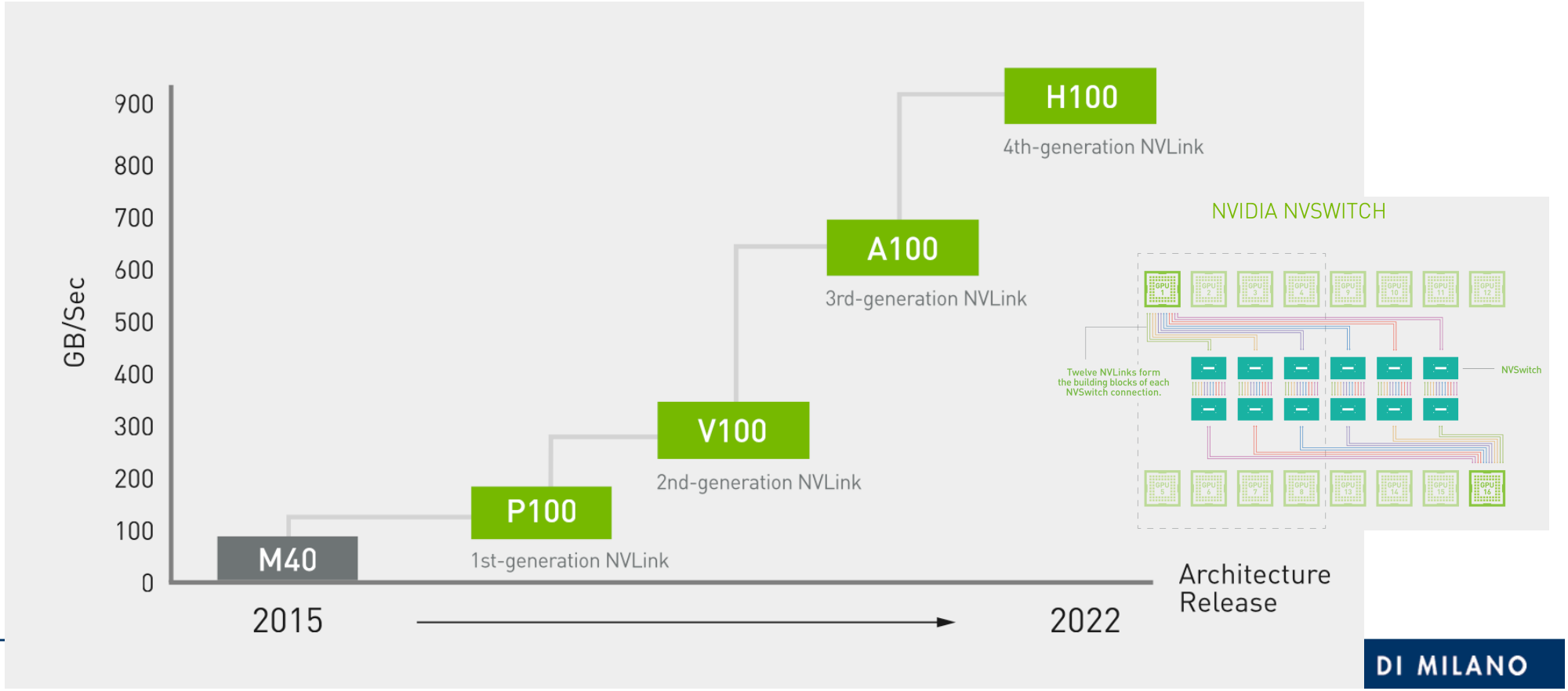
- GPUs are configured with a CPU host connected to a PCIe-attached accelerator tray with multiple GPUs
- GPUs within the tray are connected using high-bandwidth interconnects such as NVlink



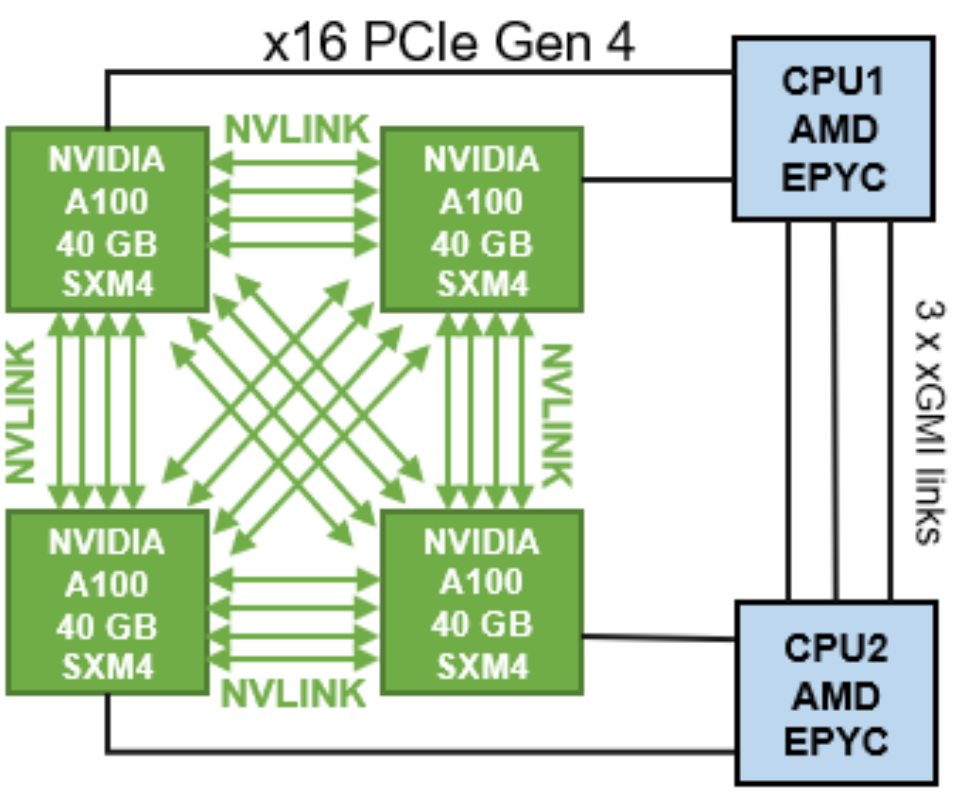


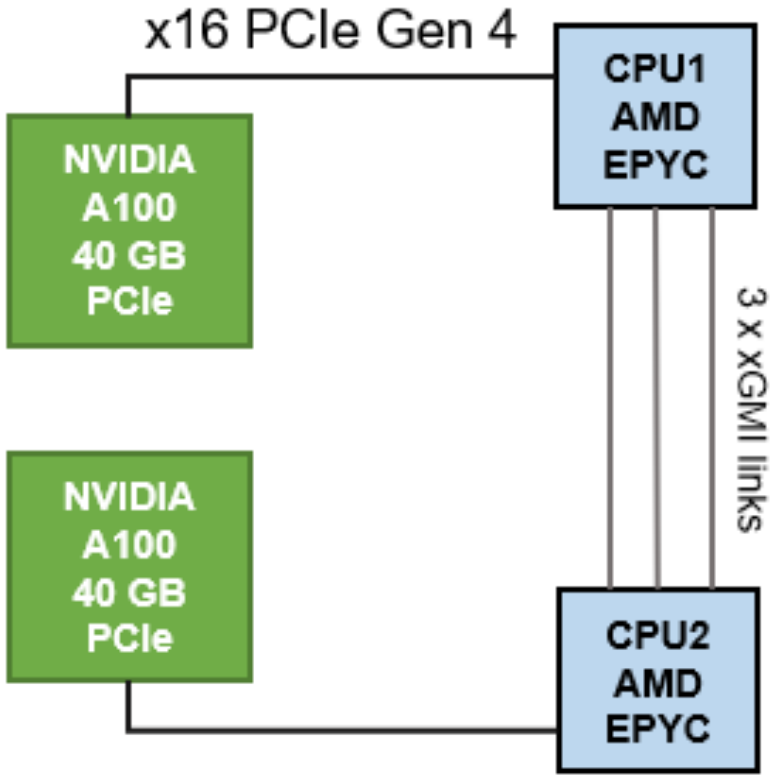
NVLINK evolution and NVSwitch

- In the A100 GPU, each NVLink lane supports a data rate of 50 x 4 Gb/s in each direction
- The total number of NVLink lanes increases from six lanes in the V100 GPU to 12 lanes in the A100 GPU, yielding 600 GB/s total
- With H100 each GPU can have up to 18 lanes and we reach 900 GB/s total



Examples of DC Servers

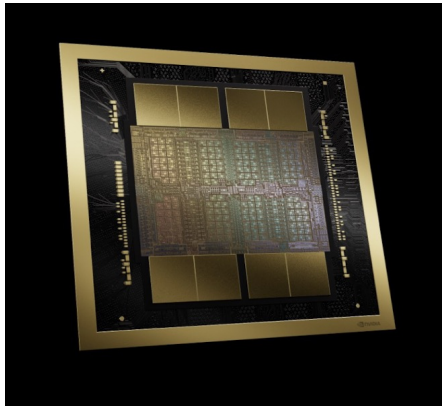




Server	PowerEdge XE8545	PowerEdge R7525
Processor	Dual AMD EPYC 7713, 64C, 2.8 GHz	
Memory	512 GB (16 x 32 GB @ 3200 MT/s)	1024 GB (16 x 64 GB @ 3200 MT/s)
Height of system	4U	2U
GPUs	4 x NVIDIA A100 SXM4 40 GB	2 x NVIDIA A100 PCIe 40 GB

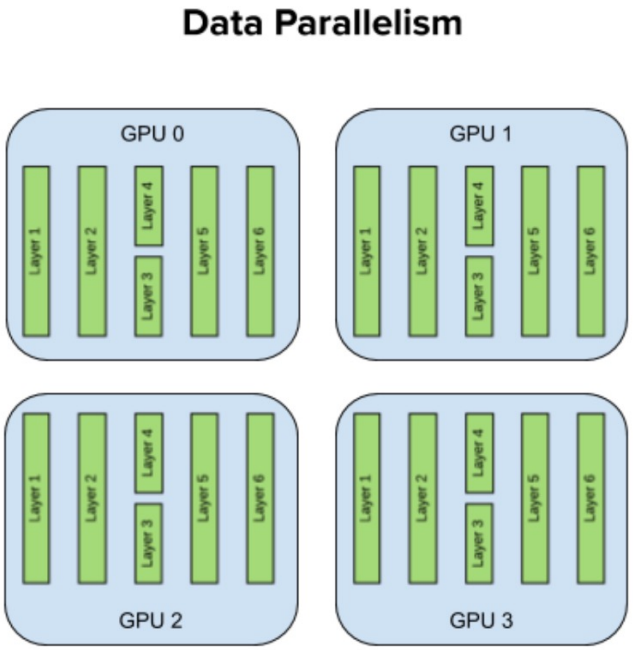
NVIDIA Blackwell Architecture B200 GPU Overview

- Target: AI and HPC workloads
- Performance:
  - Up to 72 petaFLOPS (FP16) for training and 144 petaFLOPS (FP16) for inference
  - Equipped with 192GB HBM3e memory per system, delivering up to 8TB/s memory bandwidth
  - Features 1.8TB/s NVLink bandwidth, enabling high-speed communication between GPUs
  - Reduces energy consumption by up to 25x compared to Hopper GPUs for large-scale AI workloads (maximum of ~14.3kW for the entire system)
- Applications:
  - Accelerates training and inference of large-scale AI models, including multi-trillion parameter language models
  - Generative AI, scientific simulations, and real-time applications like speech recognition and robotics



	GB200 NVL72	HGX B200	HGX B100
Blackwell GPUs	72	8	8
FP4 Tensor Core	1,440 petaFLOPS	144 petaFLOPS	112 petaFLOPS
FP8/FP6/INT8	720 petaFLOPS	72 petaFLOPS	56 petaFLOPS
Fast Memory	Up to 30 TB	up to 1.5 TB	Up to 1.5TB
Aggregate Memory Bandwidth	Up to 600 TB/s	Up to 64 TB/s	Up to 64 TB/s
Aggregate NVLink Bandwidth	130 TB/s	14.4 TB/s	14.4 TB/s
CPU Cores	2592 Arm Neoverse V2 cores	-	-
Per GPU Specifications			
FP4 Tensor Core	20 petaFLOPS	18 petaFLOPS	14 petaFLOPS
FP8/FP6 Tensor Core	10 petaFLOPS	9 petaFLOPS	7 petaFLOPS
INT8 Tensor Core	10 petaOPS	9 petaOPS	7 petaOPs
FP16/BF16 Tensor Core	5 petaFLOPS	4.5 petaFLOPS	3.5 petaFLOPS
TF32 Tensor Core	2.5 petaFLOPS	2.2 petaFLOPS	1.8 petaFLOPS
FP64 Tensor Core	45 teraFLOPS	40 teraFLOPS	30 teraFLOPS
GPU memory | Bandwidth	Up to 192 GB HBM3e | Up to 8 TB/s		
Multi-Instance GPU (MIG)	7		
Decompression Engine	Yes		
Decoders	2x 7 NVDEC 2x 7 NVJPEG		
Power	Configurable up to 1,200W	Configurable up to 1,000W	Configurable up to 700W
Interconnect	5th Generation NVLink: 1.8TB/s PCIe Gen6: 256GB/s		
Server options	NVIDIA GB200 NVL72 partner and NVIDIA-Certified Systems with 72 GPUs	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs	NVIDIA HGX B100 partner and NVIDIA-Certified Systems with 8 GPUs

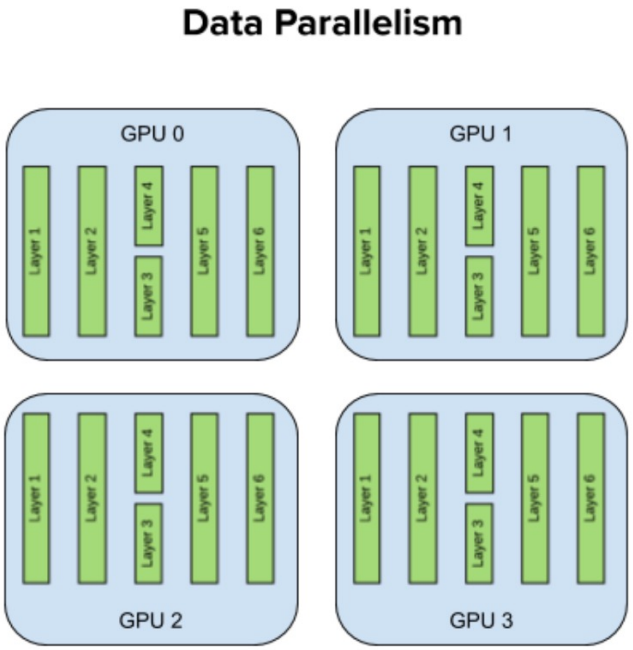
GPU: training a DNN on multiple GPUs

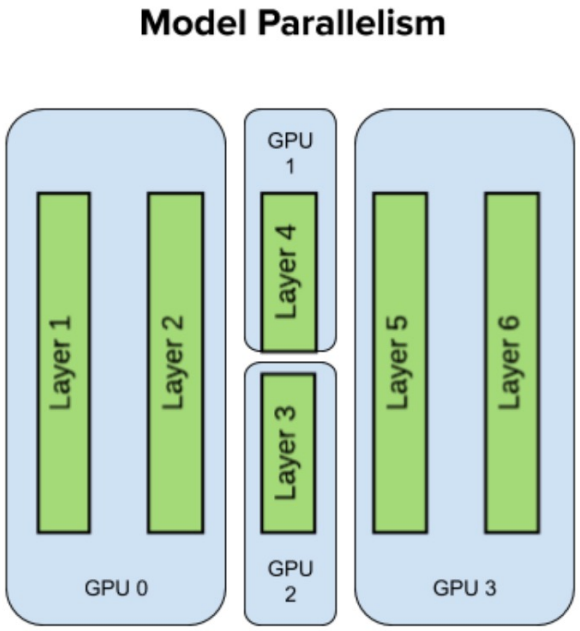


CNN, RNN, …

- each GPU up to 80GB
- GPU memory used to store mainly batches of the training set

GPU: training a DNN on multiple GPUs





CNN, RNN, …

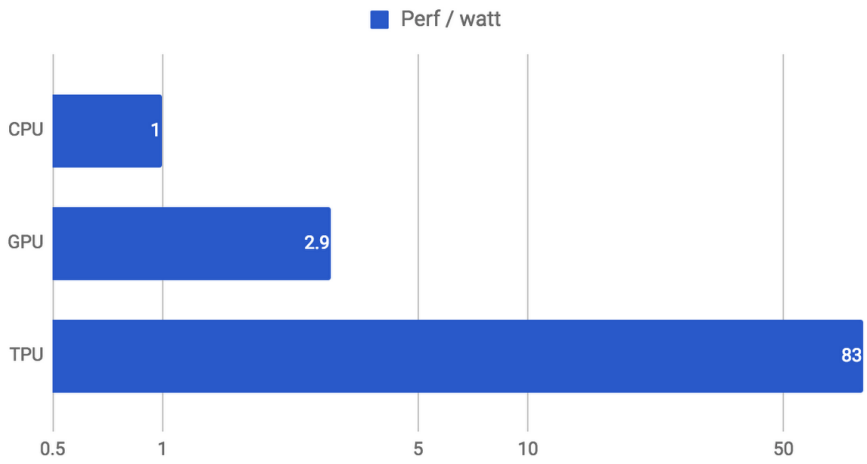
- each GPU up to 80GB
- GPU memory used to store mainly batches of the training set

LLMs

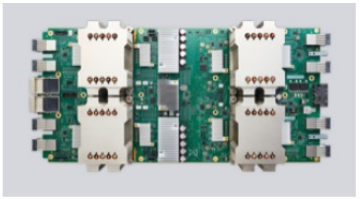
- GPT4 estimate 6.8 TB

Tensor Processing Unit (TPU)

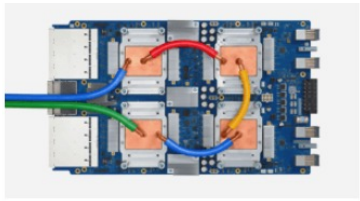
- While suited to ML, **GPUs are still relatively general purpose devices**
- In recent years designers further specialized them to ML-specific hardware
  - **Custom-built integrated circuit** developed specifically for machine learning and tailored for **TensorFlow, Pytorch,** or other ML frameworks
- Powering Google data centers since 2015 as well as CPUs and GPUs
- A **Tensor** is an n-dimensional matrix
- **TPUs are used for training and inference**
  - TPUv1 is an inference-focused accelerator connected to the host CPU through PCIe links
  - Differently, TPUv2 – TPV6 focus both training and inference



TPU V1

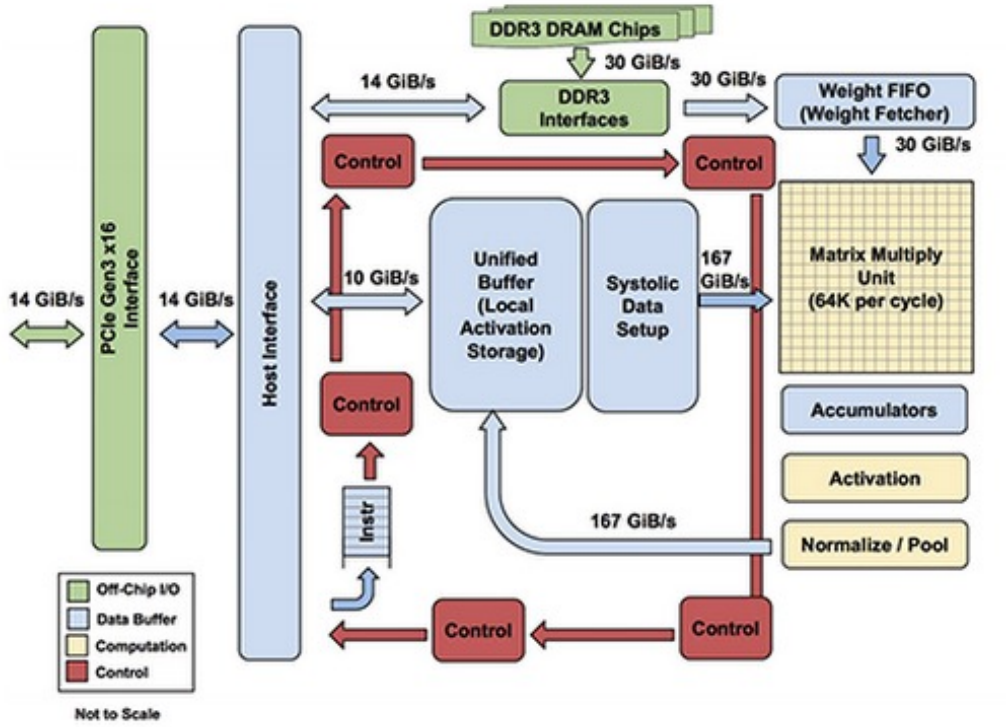


Cloud TPU v2



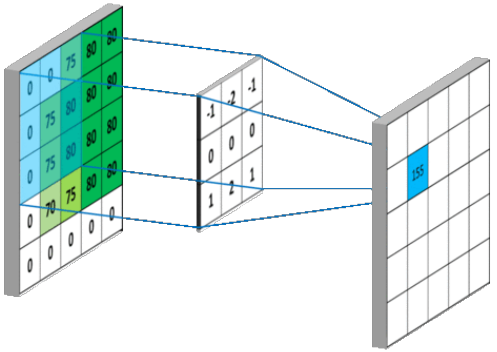
Cloud TPU v3

TPU Block Diagram



Inference

Training



THE CHAIN RULE STACKS

$$g(x) = a(b(c(x)))$$

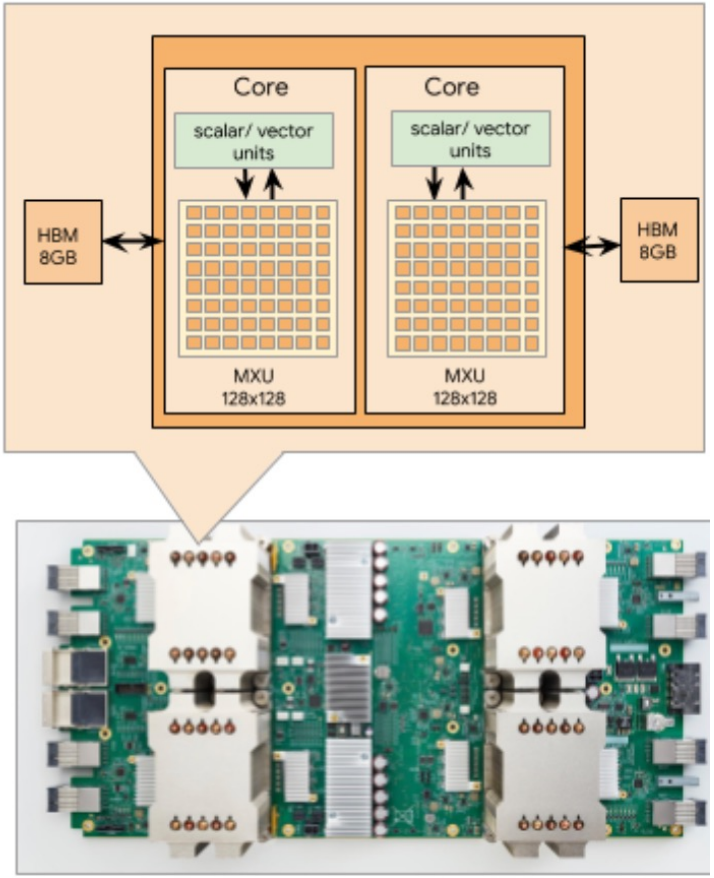
3-layer composite function

$$g'(x) = a'(b(c(x))) \cdot b'(c(x)) \cdot c'(x)$$

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

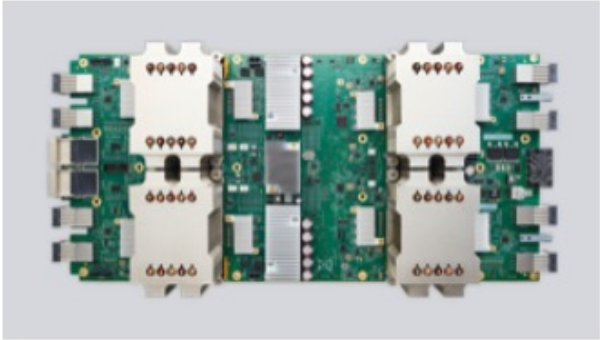
- TPU v2:
  - 8 GiB of HBM for each TPU core
  - One MXU for each TPU core
  - 4 chips, 2 cores per chip



TPU v2 - 4 chips, 2 cores per chip

TPUv2 in a Rack (Pod) - NEW

- 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 (beta)

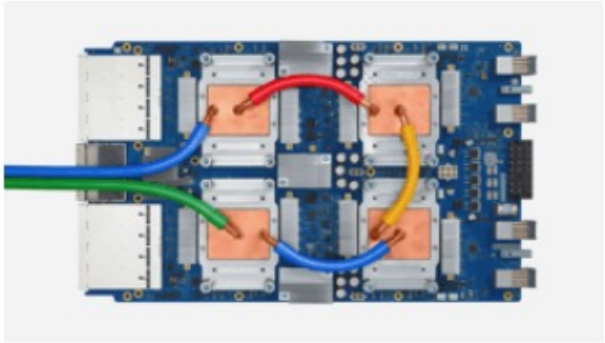
11.5 petaflops

4 TB HBM

2-D toroidal mesh network

TPUv3 (liquid-cooled) - NEW

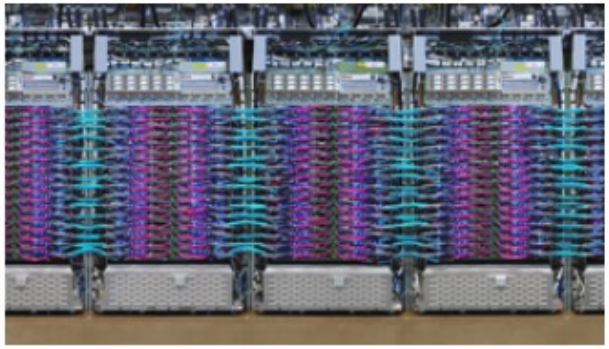
- TPUv3 is the first **liquid-cooled accelerator** in Google's data center.
- 2.5x faster than TPUv2
- Such supercomputing-class computational power supports:
  - new ML capabilities (e.g., AutoML)
  - rapid neural architecture search
- The v3 TPU Pod provides a maximum configuration of 256 devices for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

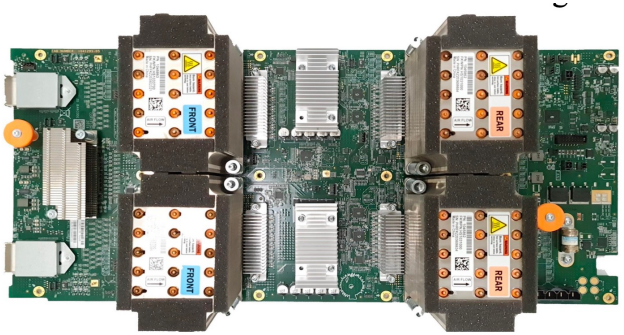
100+ petaflops

32 TB HBM

2-D toroidal mesh network

TPU v4-v6…

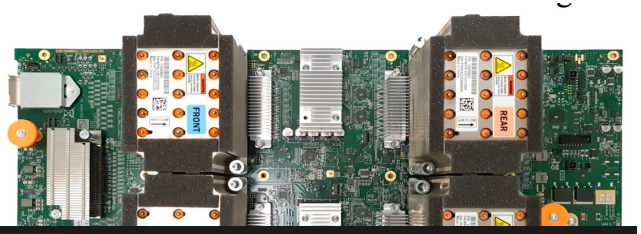
- **TPUv4**
  - Announced in 2021
  - One v4 TPU pod includes 4096 devices (2x w.r.t. v3)



- **TPUv5**
  - Announced in 2023
  - The first one available in non-US Datacenters
  - Two different versions
    - **V5e**: «cost-efficient» AI accelerator with PODs that scale up to 256 devices
    - **V5p**: designed to push more FLOPS and scale to even larger clusters (8K devices)
- **TPUv6**
  - Announced in 2024
  - Support modern AI workloads (transformers, text-to-image, LLM fine-tuning and serving)

TPU v4-v6…

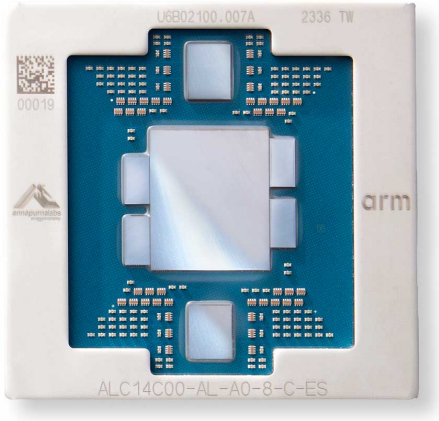
- **TPUv4**
  - Announced in 2021
  - One v4 TPU pod includes 4096 devices



Feature	TPU v1	TPU v2	TPU v3	TPU v4	TPU v5e	TPU v5p	TPU v6e (Trillium)
Year Introduced	2015	2017	2018	2021	2023	2023	2024
Process Node	28 nm	16 nm	16 nm	7 nm	Unstated	Unstated	Unstated
Peak Compute	23 TOPS	45 TOPS	123 TOPS	275 TOPS	197 TFLOPs (BF16)	459 TFLOPs (BF16)	918 TFLOPs (BF16)
					393 TOPS (INT8)	918 TOPS (INT8)	1836 TOPS (INT8)
Memory Capacity	8 GB DDR3	16 GB HBM	32 GB HBM	32 GB HBM	16 GB HBM	95 GB HBM	32 GB HBM
Memory Bandwidth	34 GB/s	600 GB/s	900 GB/s	1200 GB/s	819 GB/s	2765 GB/s	1640 GB/s
Interconnect Bandwidth	N/A	N/A	N/A	~7.12 Tbps	~1.6 Tbps	~4.8 Tbps	~3.58 Tbps
Pod Size	N/A	Multi-chip Pods	Multi-chip Pods	Up to 1024 chips	Up to 256 chips	Up to 8960 chips	Up to 256 chips
TDP (Power Consumption)	~75 W	~280 W	~220 W	~170 W	Not Listed	Not Listed	Not Listed
Topology	N/A	N/A	N/A	3D Torus	N/A	Reconfigurable Links	N/A

- **TPUv6**
  - Announced in 2024
  - Support modern AI workloads (transformers, text-to-image, LLM fine-tuning and serving)

And what about AWS?



AWS Graviton4

AWS Trainium2

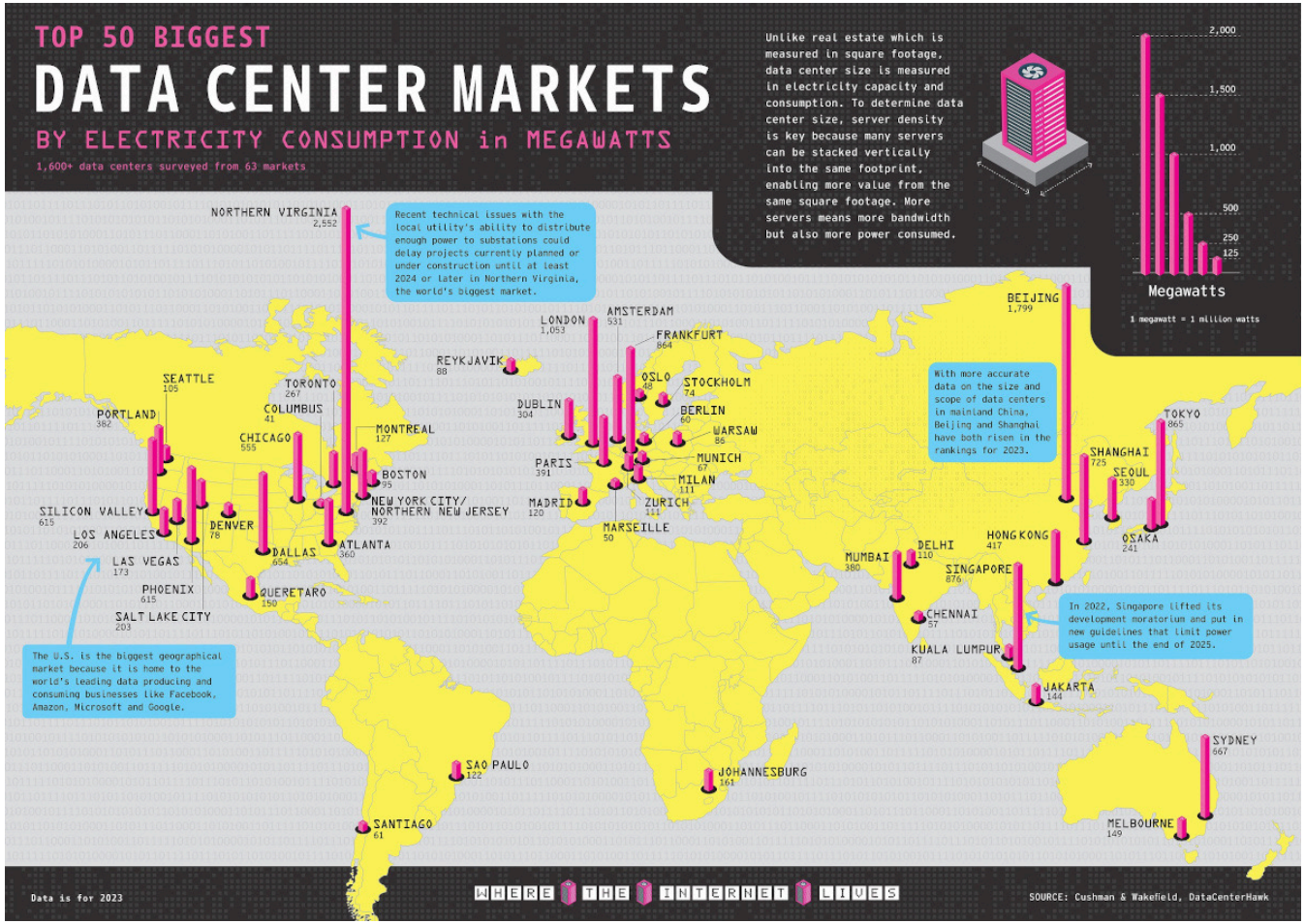
AWS Inferentia2

- Graviton4 based on ARM architecture, 30% more energy efficient for general AI tasks
- Trainium2 optimized for LLMs training, 50% lower training costs than comparable Amazon EC2 instances
- Inferentia2, for AI inference 2.3x higher throughput and up to 70% lower cost per inference than comparable EC2 instances

TPUs vs Trainium

	TPU	Trainium2
Architecture	Custom ASIC	Modified NVIDIA Ampere GPU
Cores	(>?) 2048	512
Memory	95GB HBM	640GB GDDR6
Performance	Up to 926 teraflops	Up to 450 teraflops
Efficiency	Up to 20-30x better than CPUs	Up to 10x better than CPUs

AI impact on DCs Energy consumption



https://www.visualcapitalist.com/cp/top-data-center-markets/

AI impact on DCs Energy consumption

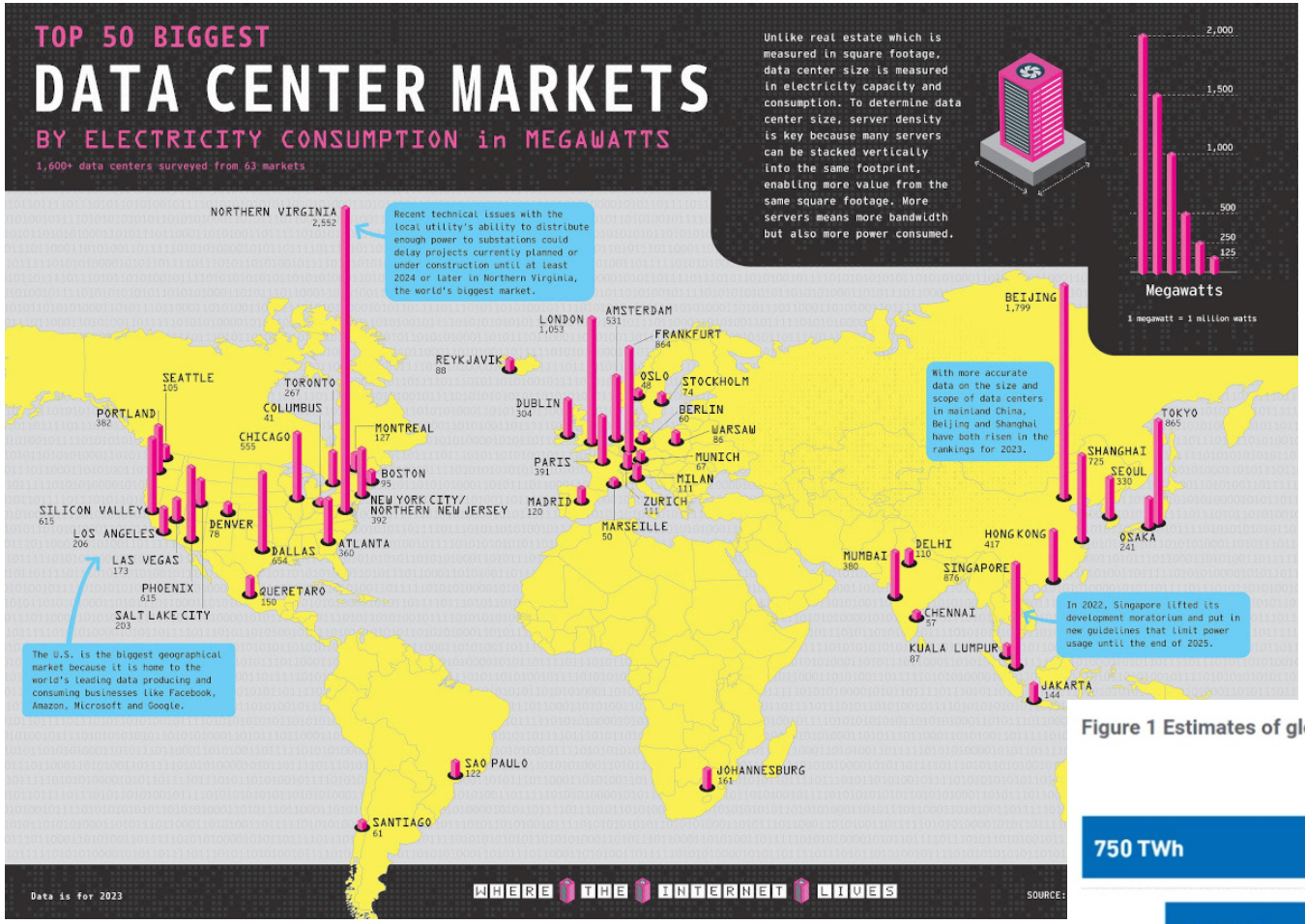
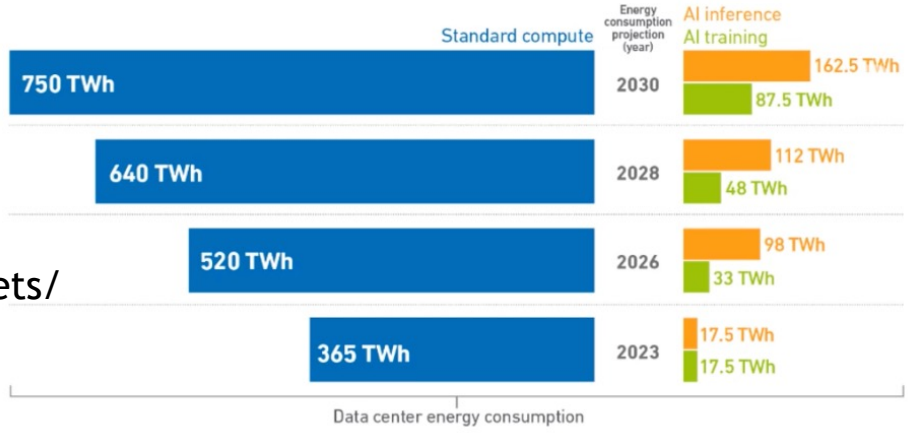


Figure 1 Estimates of global data center energy consumption 2023 to 2030



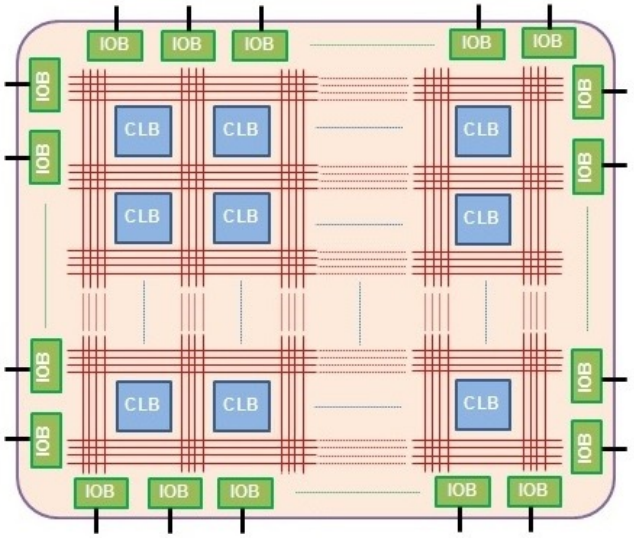
DELOITTE INSIGHTS, ELECTRIC POWER RESEARCH INSTITUTE, GOLDMAN SACHS, MOODY RATINGS, LAWRENCE BERKELEY NATIONAL LABORATORY, SCHNEIDER, INTERNATIONAL ENERGY AGENCY, S&P GLOBAL

uptime INTELLIGENCE

https://www.visualcapitalist.com/cp/top-data-center-markets/

Field-Programmable Gate Array (FPGA)

- Programmable HW device -> Custom Logic
- Array of logic gates that can be programmed (“configured”) in the field, 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.

- While not a replacement for traditional processors, FPGAs serve as a complementary technology, offering potential performance and efficiency improvements for specific data center workloads.

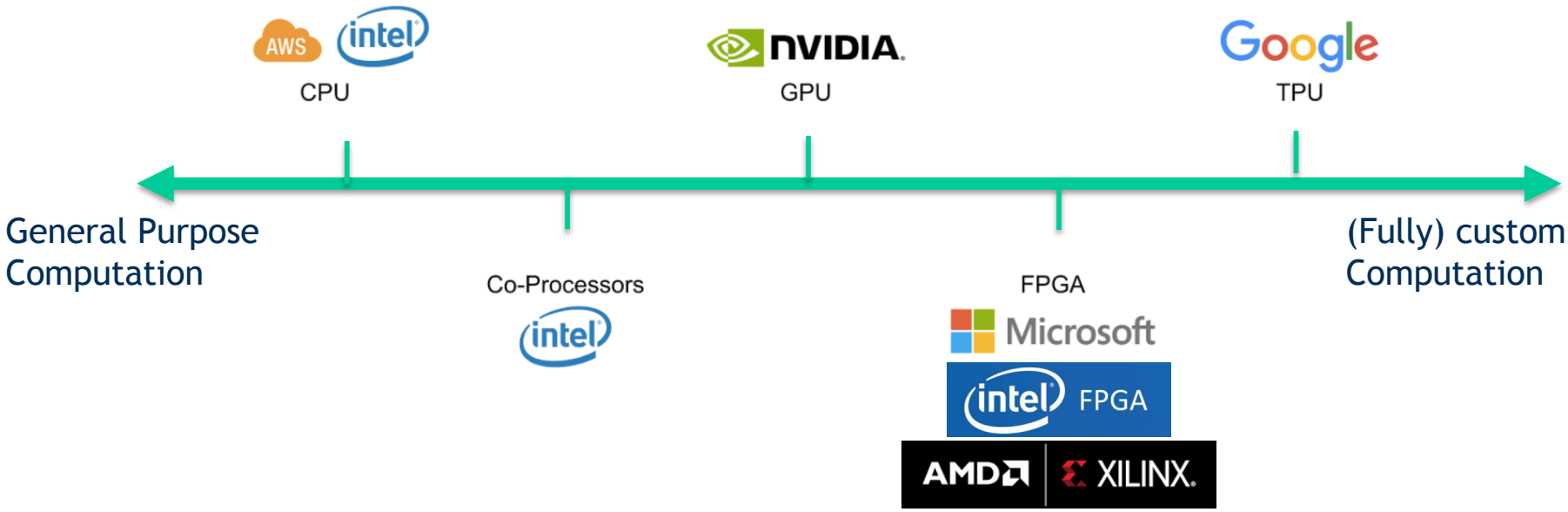
FPGA in Datacenters

- Microsoft deployed FPGAs inside its Datacenters
- Lower Carbon footprint since they are flexible, reconfiguration instead of reimplementation

FPGA Applications in Data Centers:

- **Network acceleration:** FPGAs can offload specific network processing tasks from CPUs, improving overall network performance and reducing CPU workload
- **Security acceleration:** Encryption, decryption, and other security-related tasks can be accelerated using FPGAs, enhancing data center security while maintaining performance
- **Data analytics:** FPGAs can be used to accelerate specific algorithms used in data analytics workloads, leading to faster data processing and analysis
- **Machine learning:** FPGAs can be configured to implement specific machine learning algorithms efficiently, potentially offering performance advantages for specialized tasks

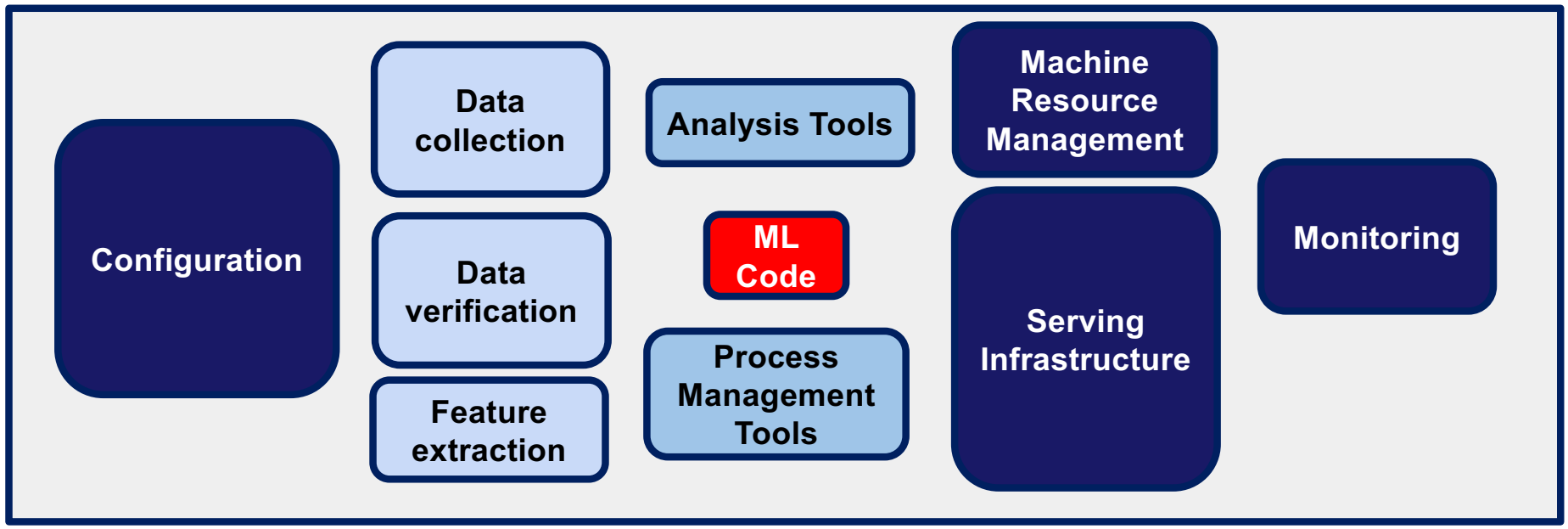
GPU, TPU and FPGA: a technological comparison



CPU, GPU, TPU and FPGA: an AI comparison

	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. The role of the Computing Infr.



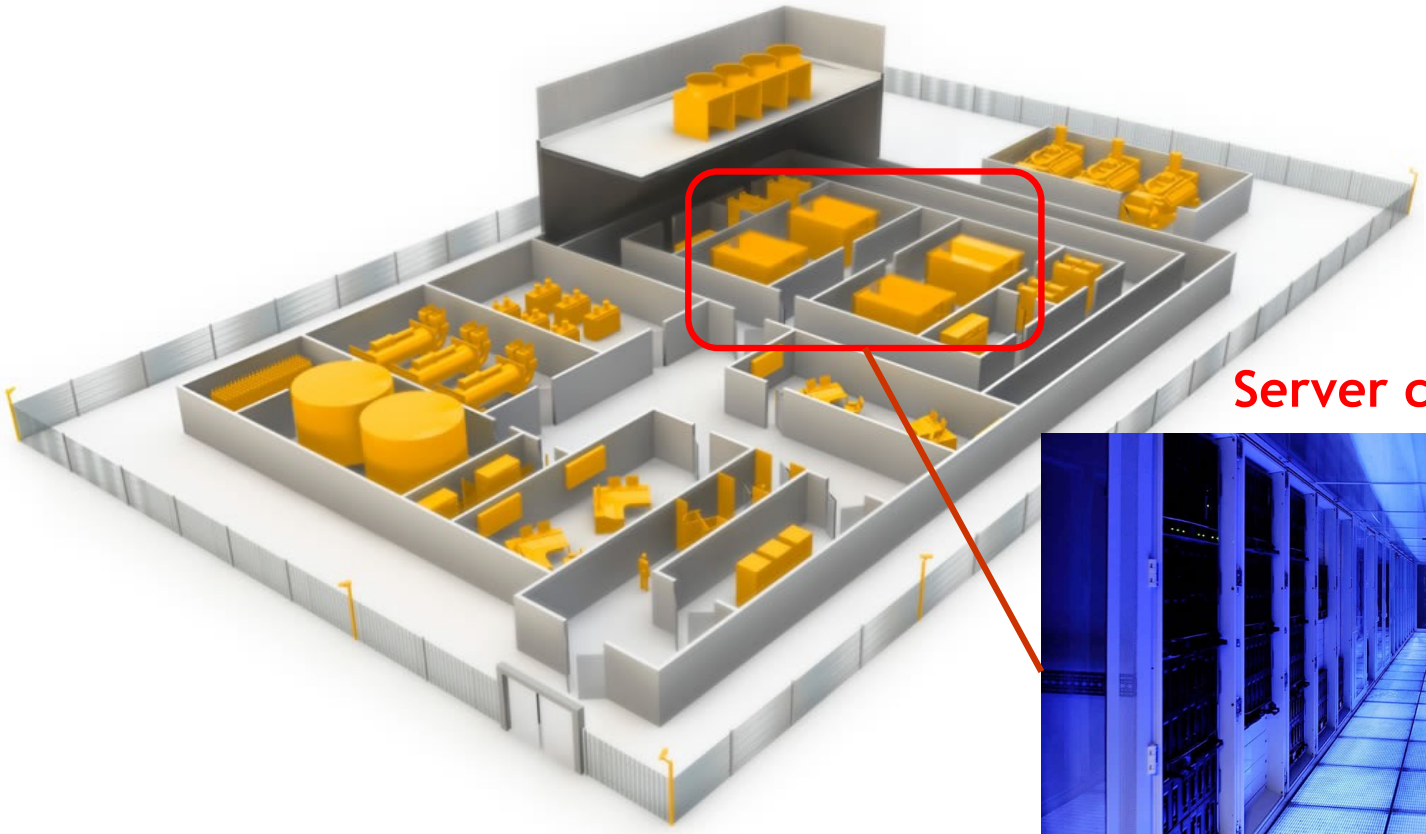
Only a small fraction of real world ML systems is composed of the ML code[1]

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

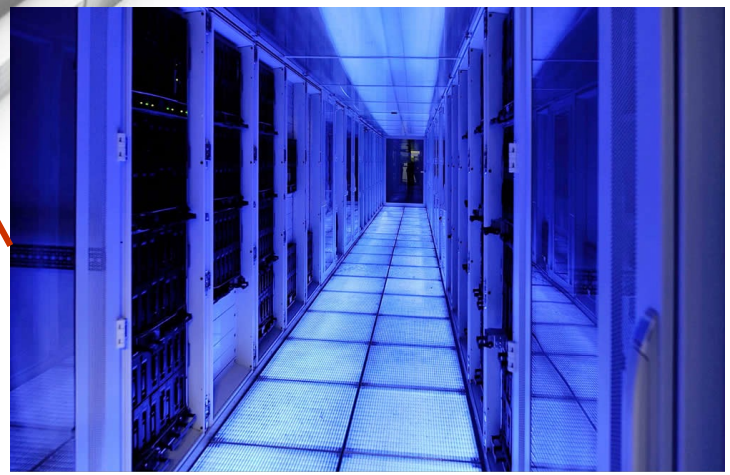
From the Rack to the Datacenter ….

Data-center architecture

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

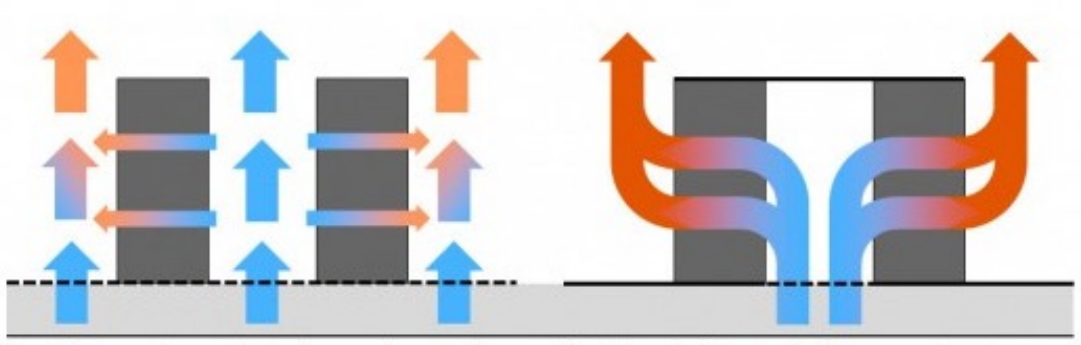


Server corridors

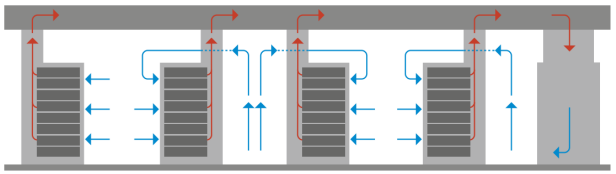


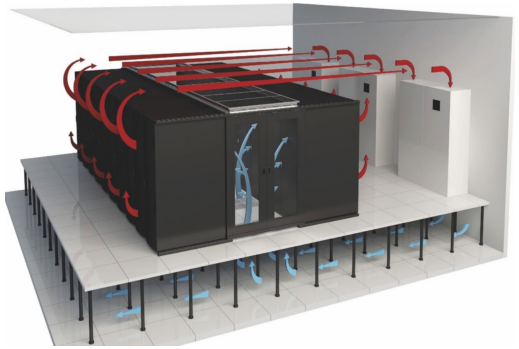
Data-center corridors

- Server Racks are **NEVER BACK-to-BACK**
- 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
- Cold air flows from the front (cool aisle), cools down the equipment, and leave the room from the back (warm aisle)



Not Unique Solution






F1 R7 CASS 4 SX G/ L


EMC
SYMMETRIX
VMAX