



 POLITECNICO DI MILANO

Computing Infrastructures



The Datacenter as a Computer



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



HW Infrastructures:

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

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

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

SW Infrastructures:

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

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

Methods:

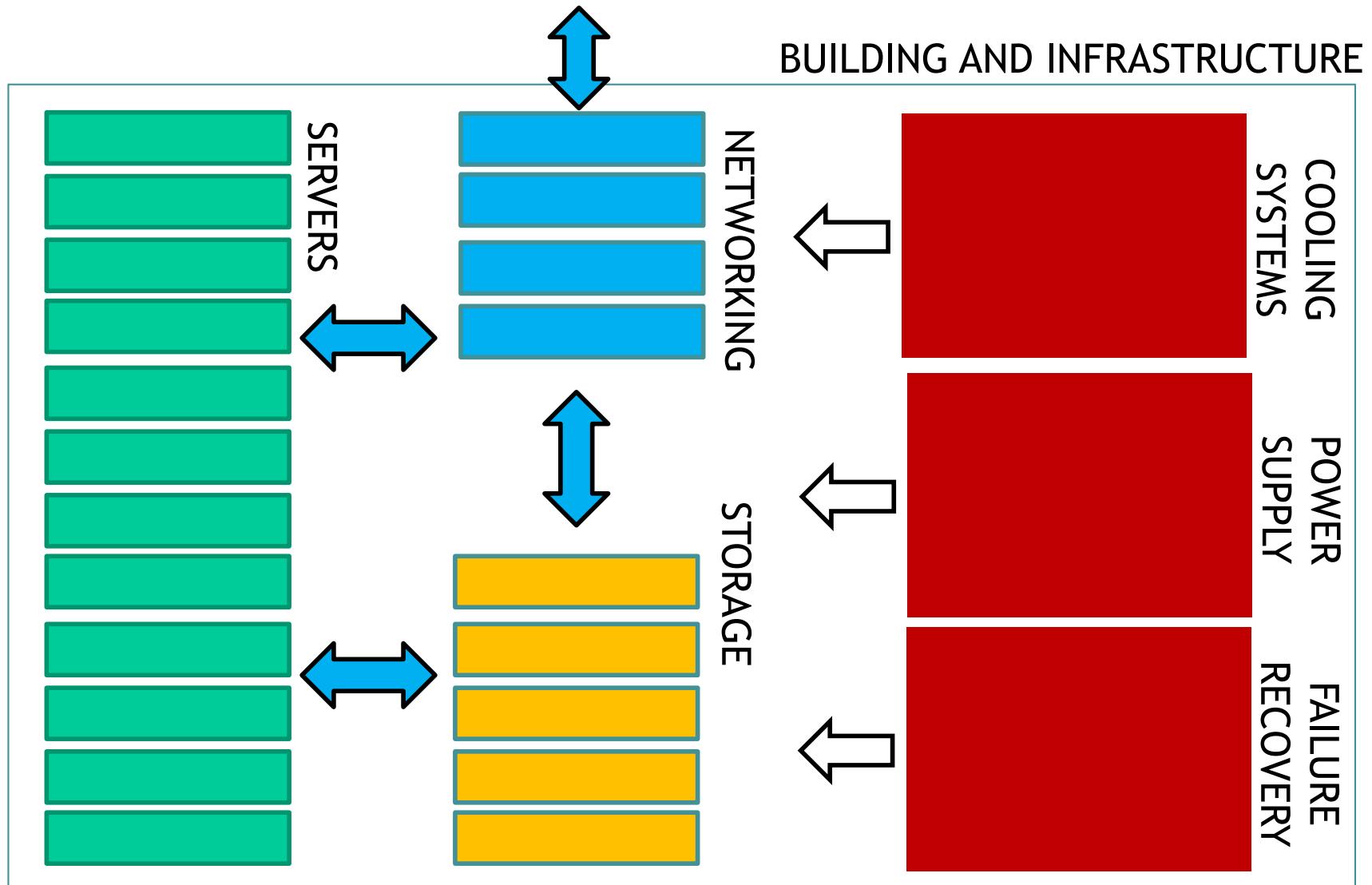
Reliability and availability of datacenters (definition, fundamental laws, RBDs)

Disk performance (Type, Performance, RAID)

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

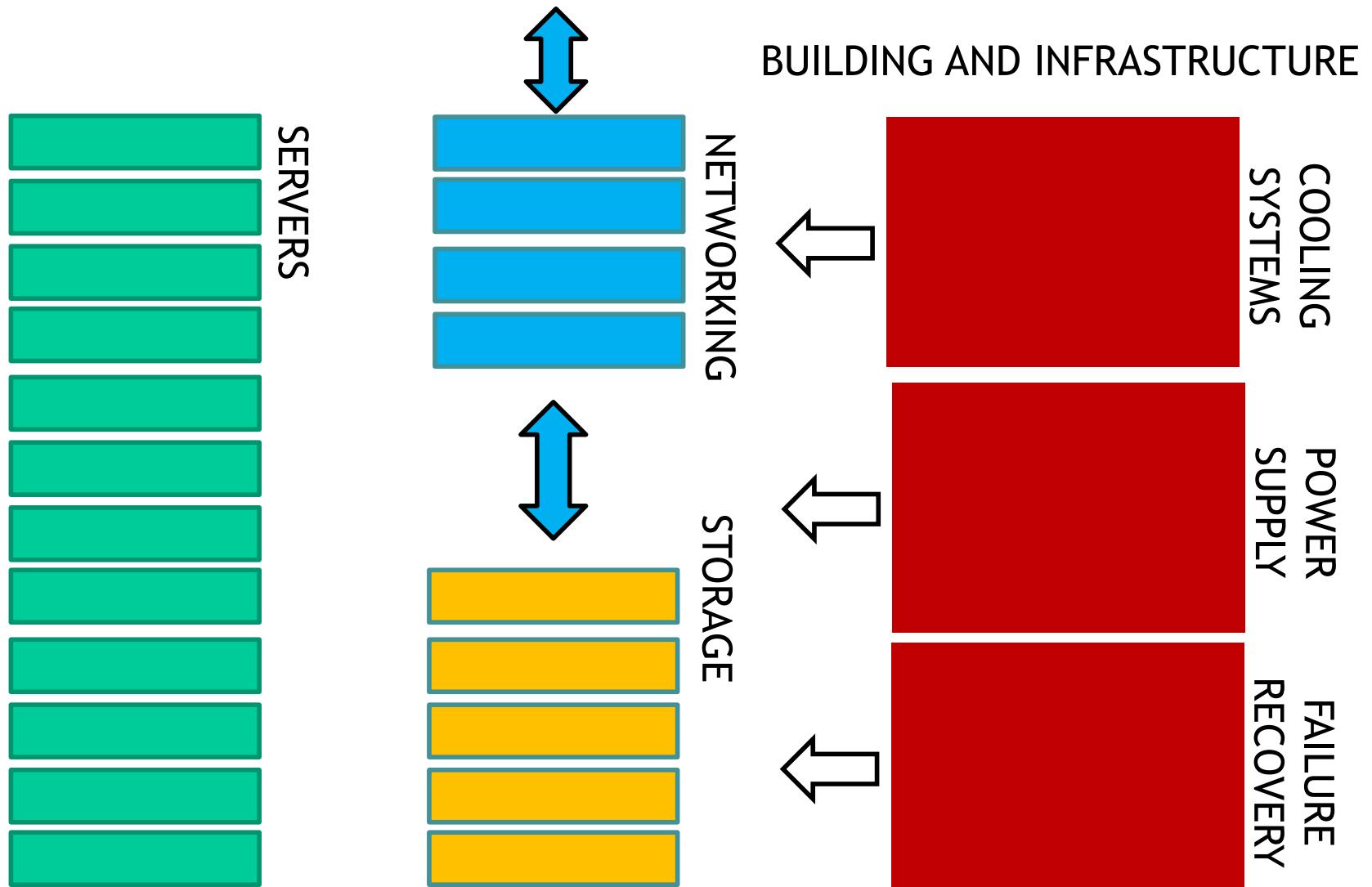


Architectural Overview of A Warehouse-scale Computer



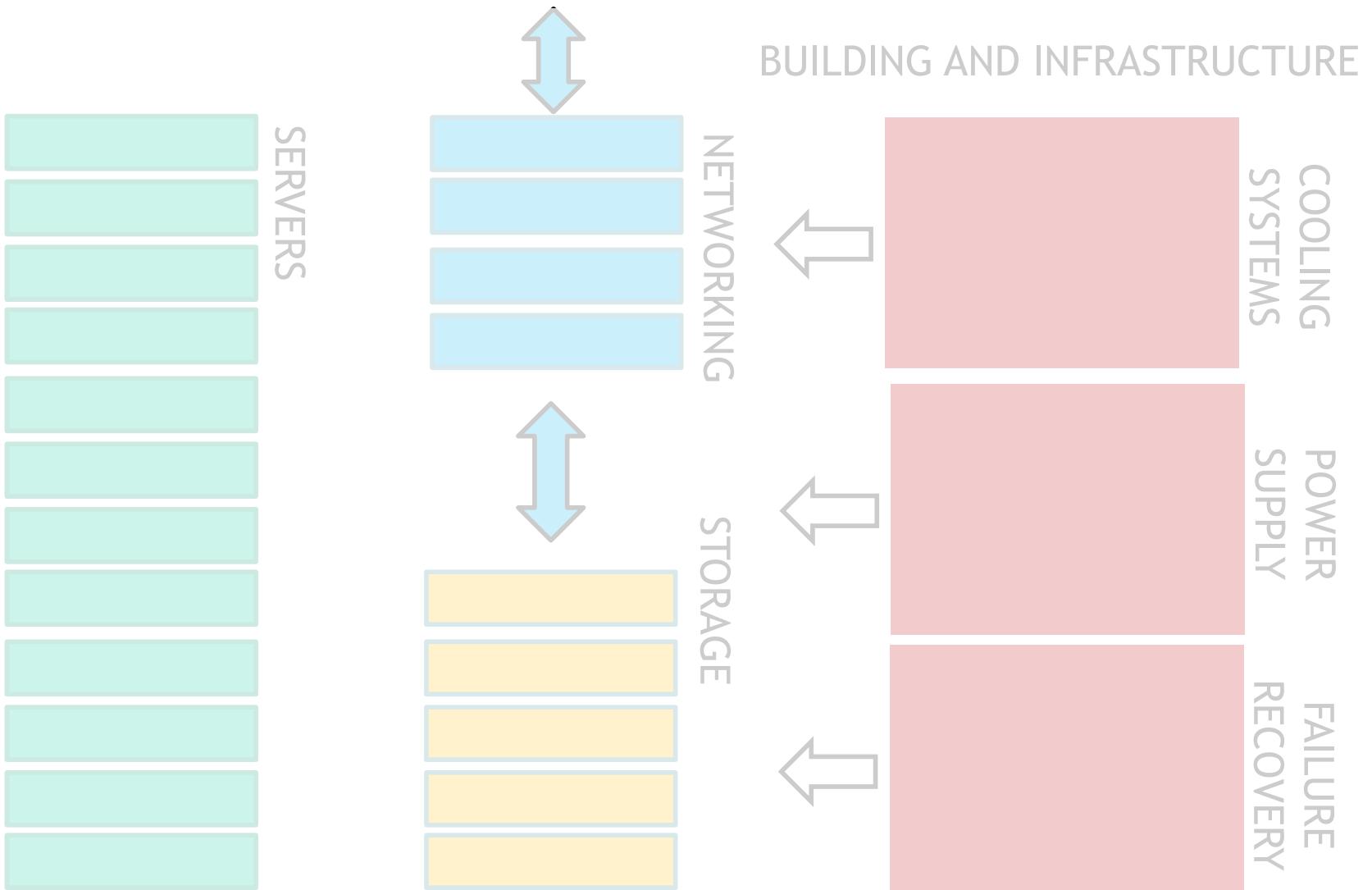


Node level: computation, storage and networking



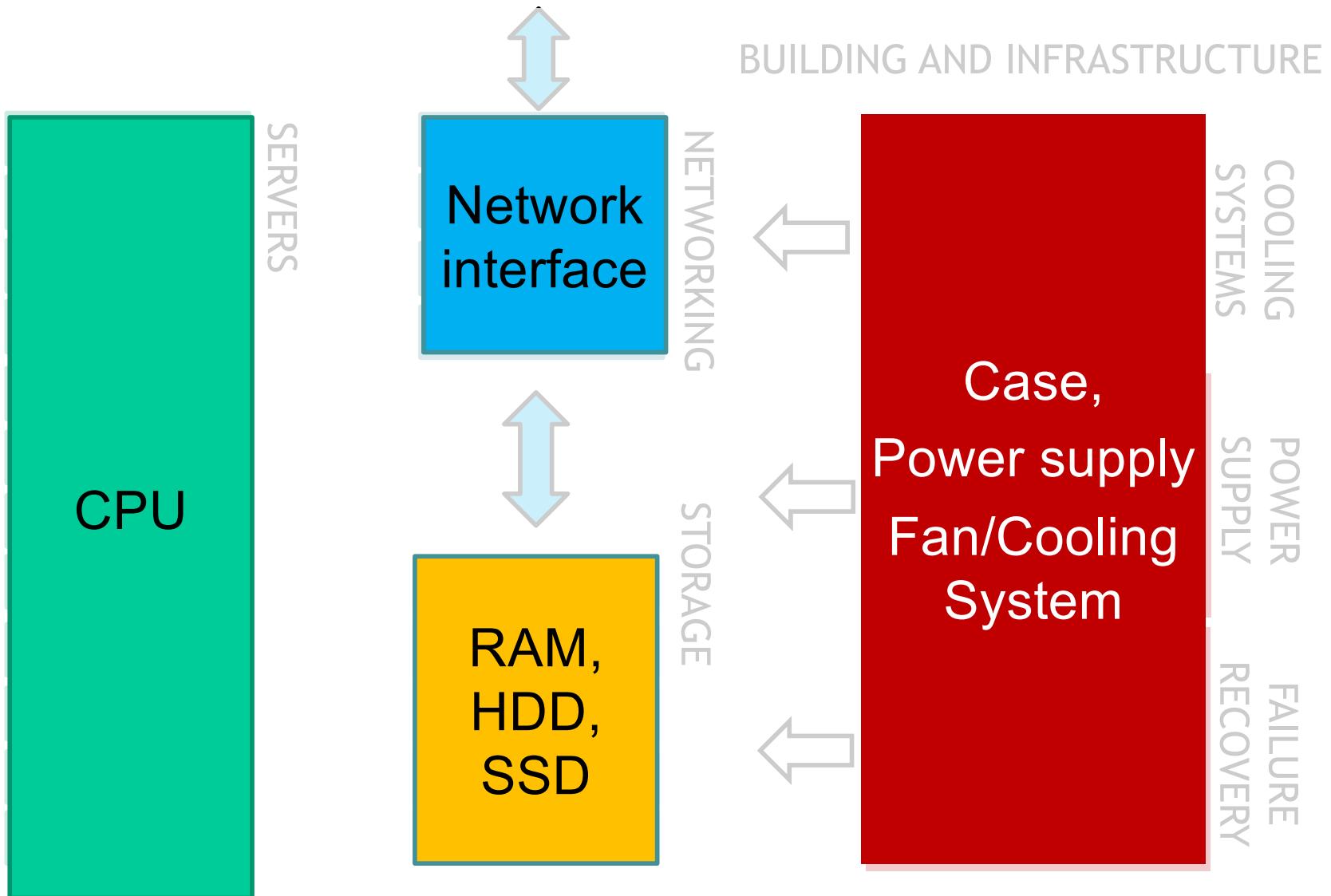


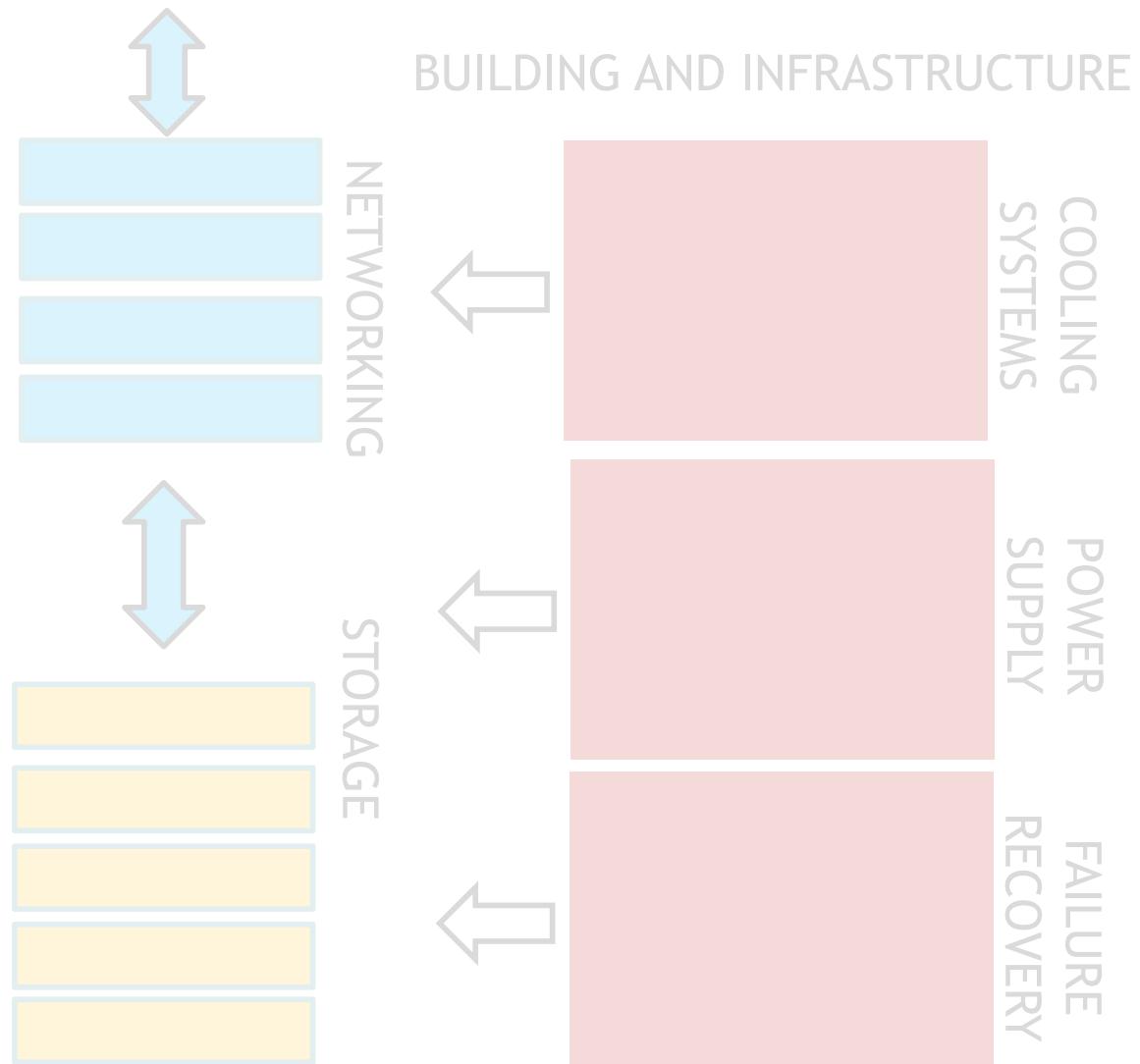
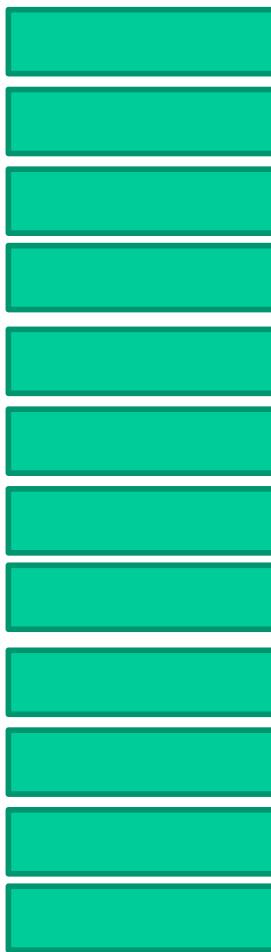
Node level: computation, storage and networking





Node level: computation, storage and networking







SERVERS: the main processing equipment



- Servers are like ordinary PC, usually more powerful, 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.



Unspecific.com



The motherboard



- The motherboard acts as the central hub, connecting all the crucial components of the server and enabling them to communicate and work together
- It 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)

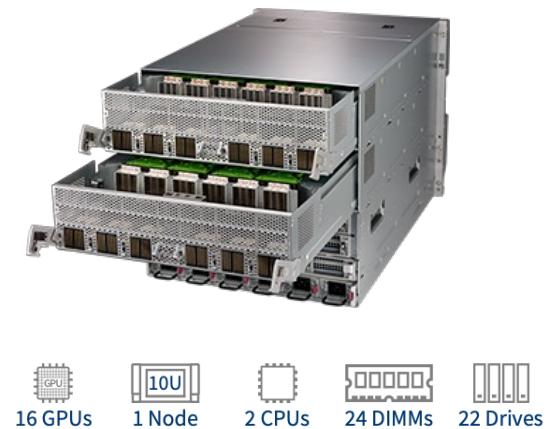
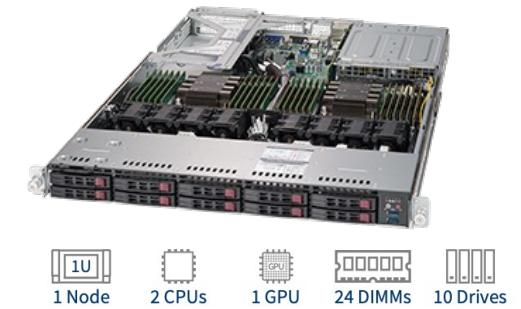
WSCs use a relatively homogeneous hardware and system software platform.



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, etc..
- ✓ Form factor:
 - From 1U to 10U
 - Tower





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



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



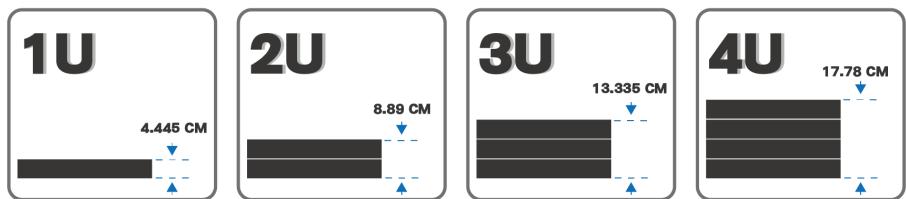
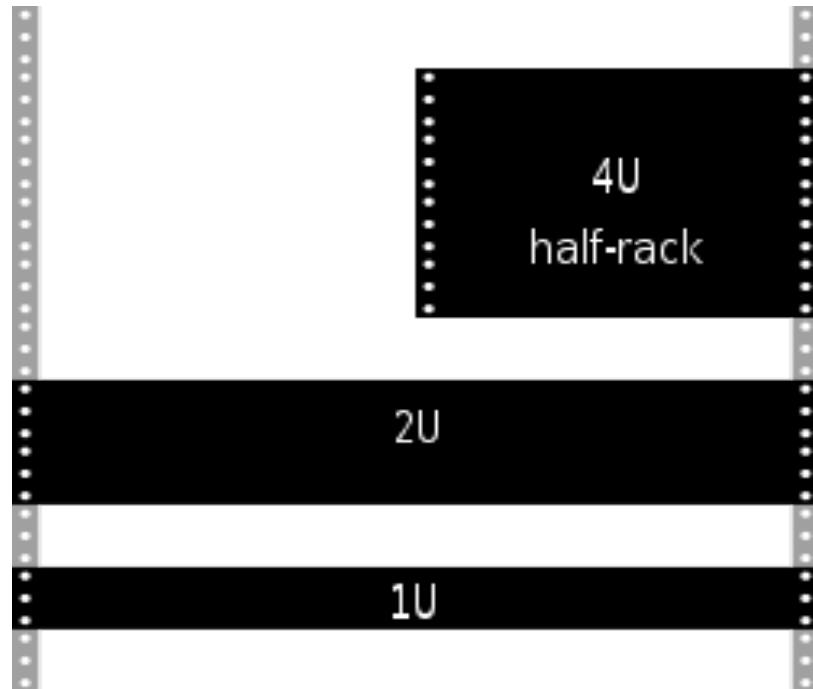
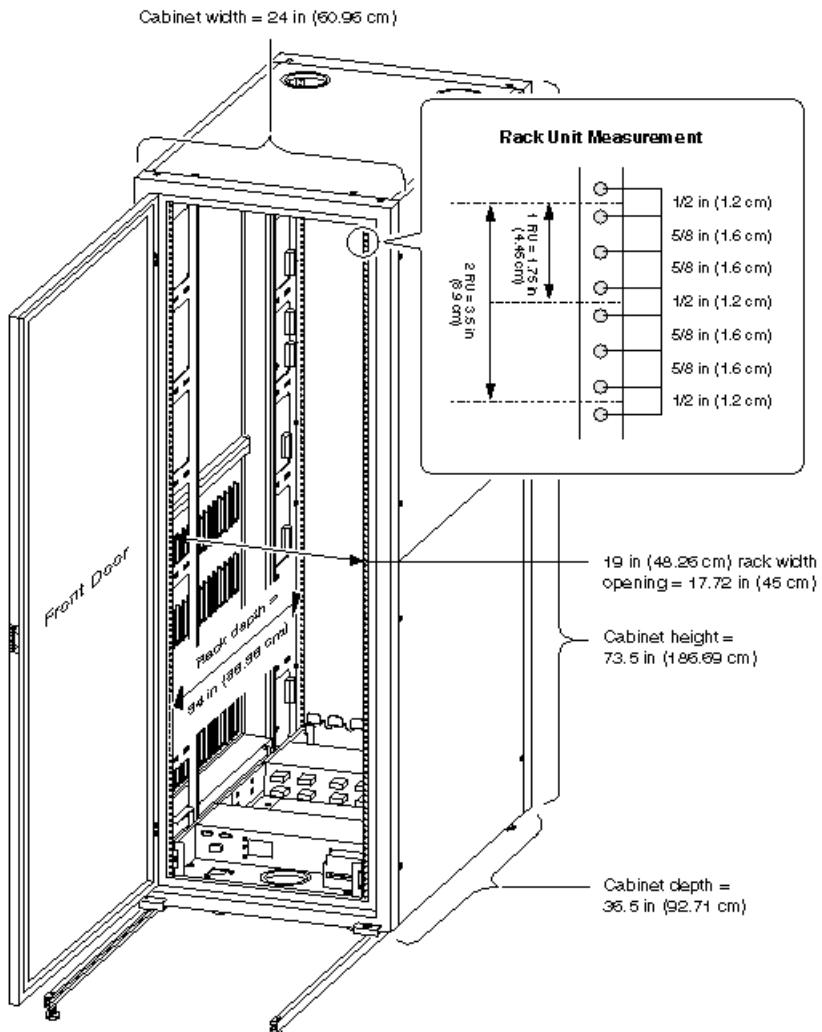
- The racks are used to store these rack servers
- Server racks are measured in rack units, or “U’s”.
- **1U is 44.45 mm (1.75 inches)**
- The advantage of using these racks is that it allows designers to stack up other electronic devices along with the servers.



Data-center racks



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





Rack 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.



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

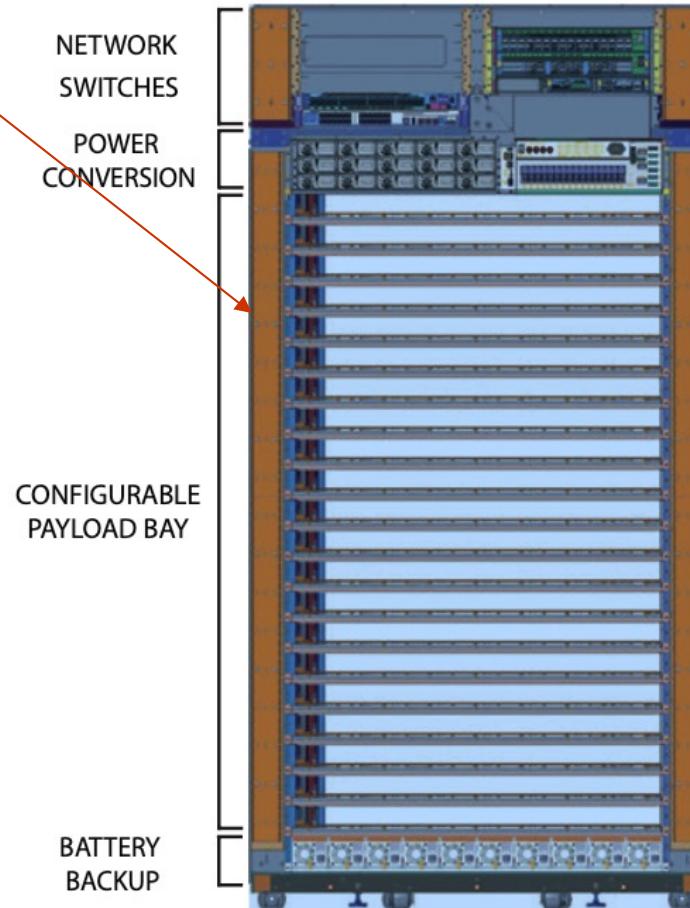


Image taken from “The Datacenter as a Computer», Barroso et al.



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

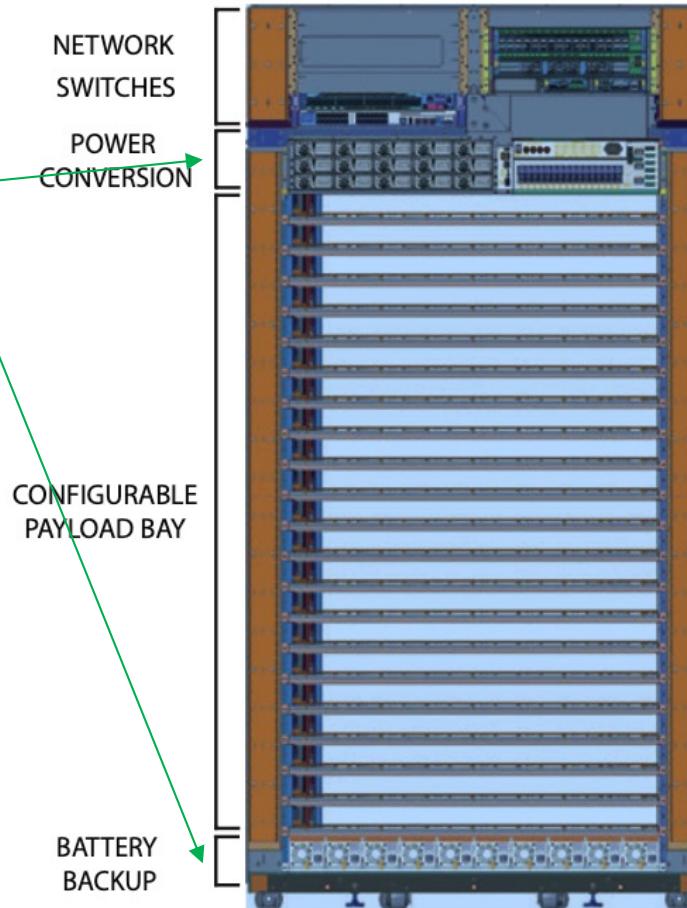


Image taken from “The Datacenter as a Computer», Barroso et al.



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

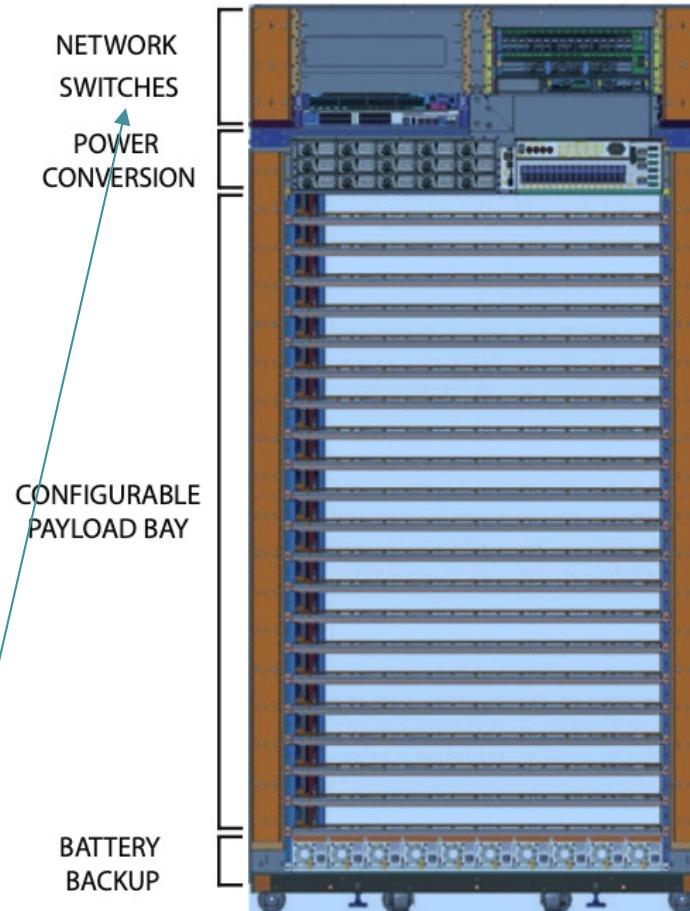


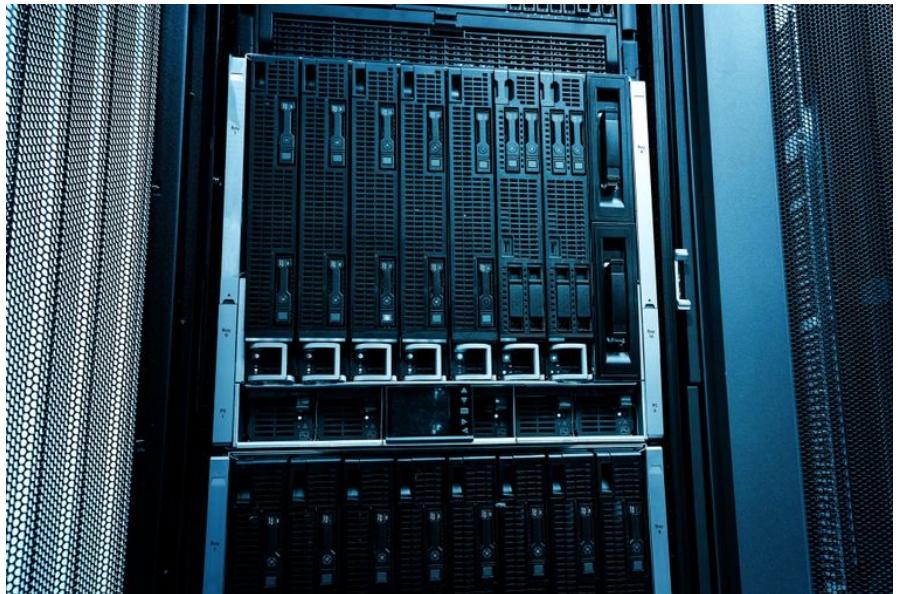
Image taken from “The Datacenter as a Computer», Barroso et al.



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. Blade servers offer **higher space efficiency** compared to traditional rack-mounted servers.
- ✓ **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:** Blade enclosures typically come with centralized management tools that allow administrators to easily monitor, configure, and update all blades from a single interface.
- ✓ **Load balancing, failover, scalability:** Uniform system, shared components (including network), simple addition/removal of servers



Blade servers: disadvantages



RACK MOUNT SERVERS



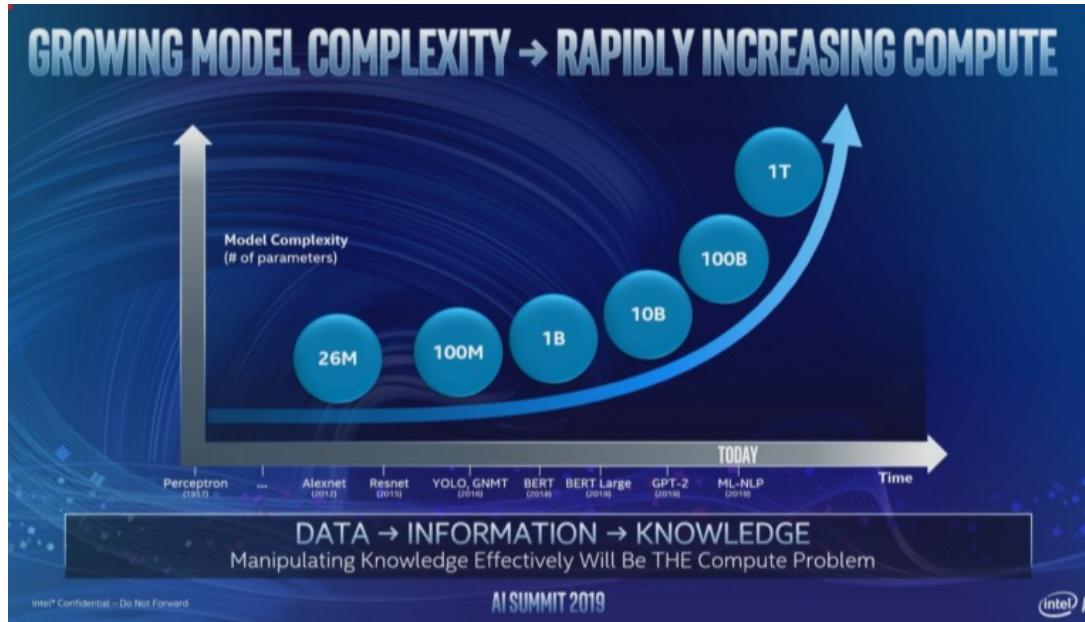
BLADE SERVERS

Cons

- ✓ **Expensive configuration and Higher initial cost:** Although upgrading the blade server is easy to handle and manage, the initial configuration or the setup requires more effort and higher initial investment;
- ✓ **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.
- ✓ **Cooling:** Blade servers come with high component density. Therefore, special accommodations have to be arranged for these servers to ensure they don't get overheated. Heating, ventilation, and air conditioning systems (HVAC) must be carefully managed and designed.



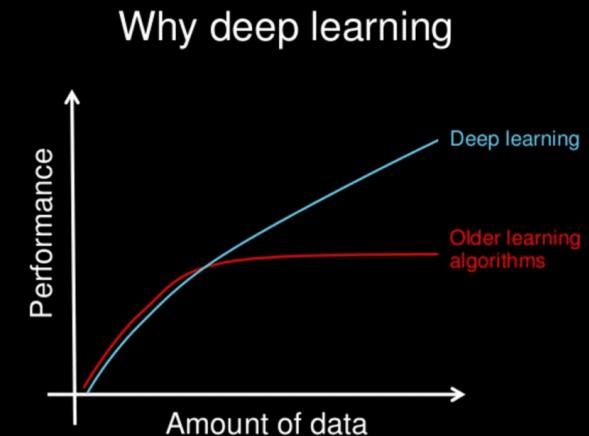
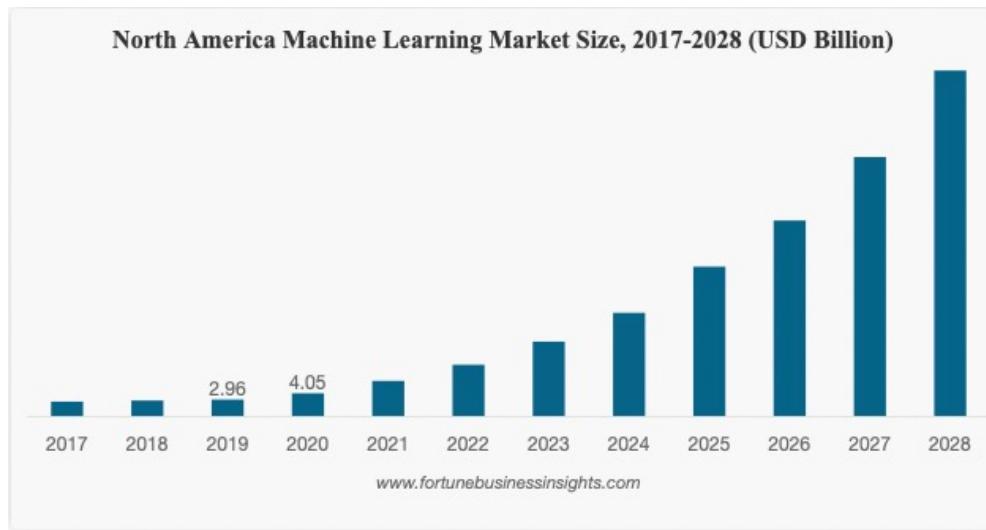
The need of hardware accelerators: ML-era



Complexity doubles every 3.5 months



18 months for Moore's Law



How do data science techniques scale with amount of data?



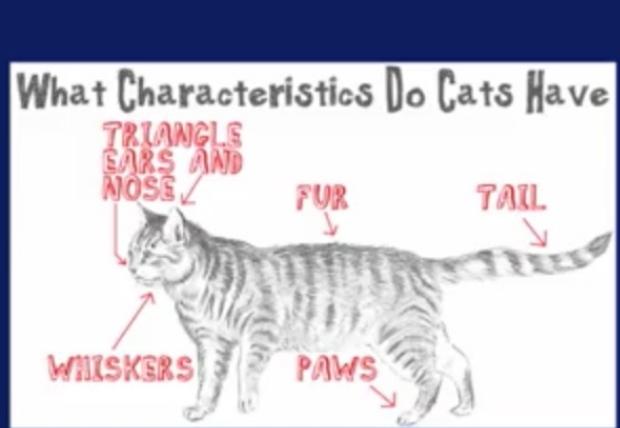
What is Machine Learning? (basic definition, this is not a ML course)



Learn from data through models

Learn with no explicit programming → learn from features

Discover hidden patterns





- 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**
 - 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
 - E.g Image classification:

$f(\text{apple}) = \text{“apple”}$

$f(\text{tomato}) = \text{“tomato”}$

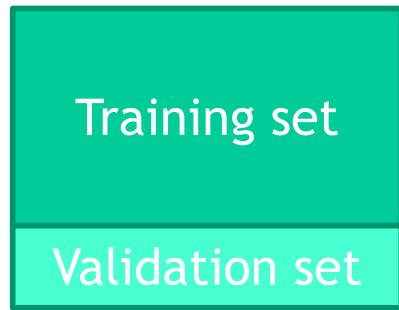
$f(\text{cow}) = \text{“cow”}$



The machine learning framework



Labeled Data



$$y = f(x)$$

Diagram illustrating the machine learning framework:

- output (red arrow pointing up from the bottom)
- prediction function (red arrow pointing up from the middle)
- Image feature (red arrow pointing right from the right side)

- **Training:** given a *training set* of labeled examples $\{(x_1, y_1), \dots, (x_N, y_N)\}$, 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 $y = f(x)$



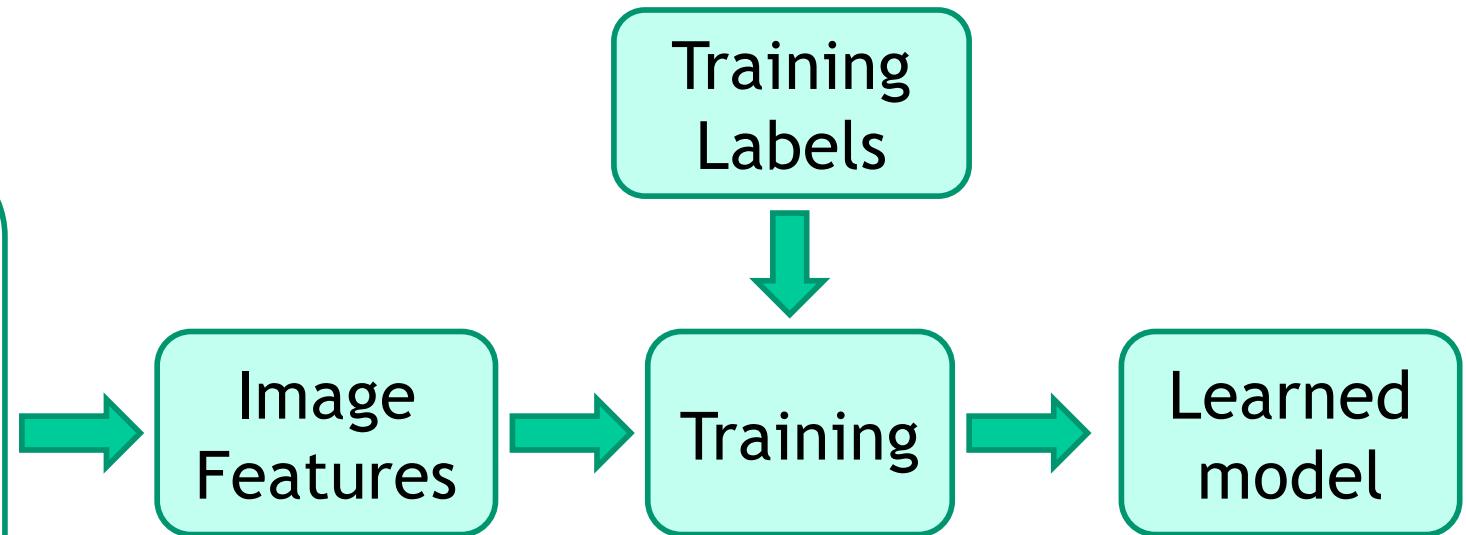


Steps

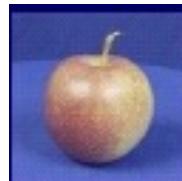


Training

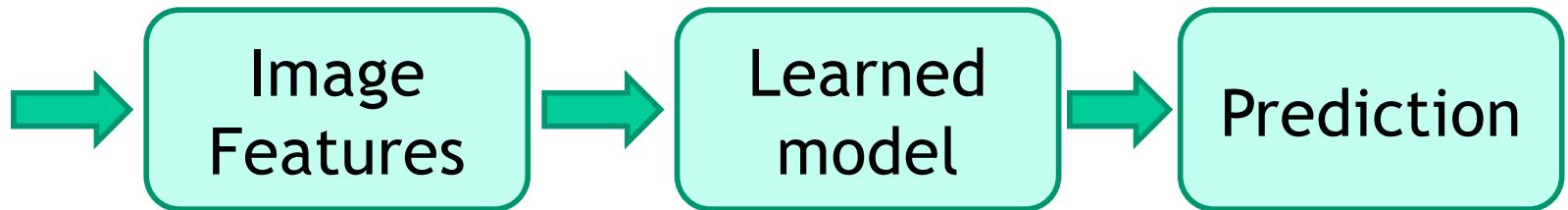
Training
Images



Testing



Test Image





Steps



Training

Training
Images



Image
Features

Training
Labels

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

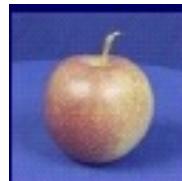


Image
Features

Learned
model

Prediction

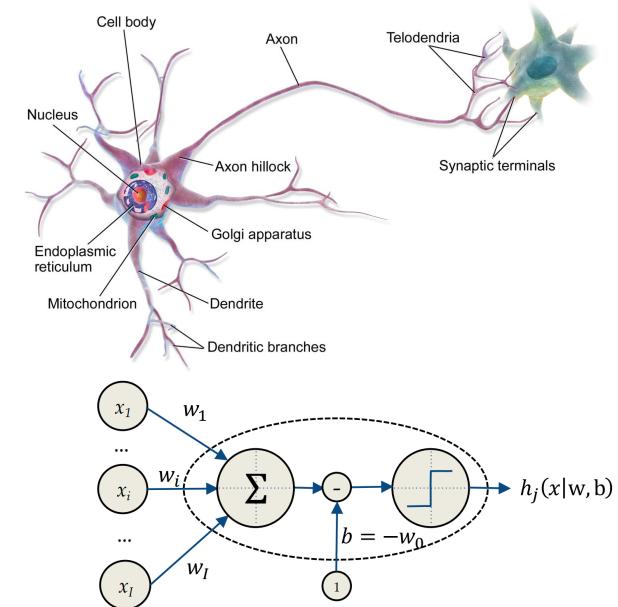
Test Image



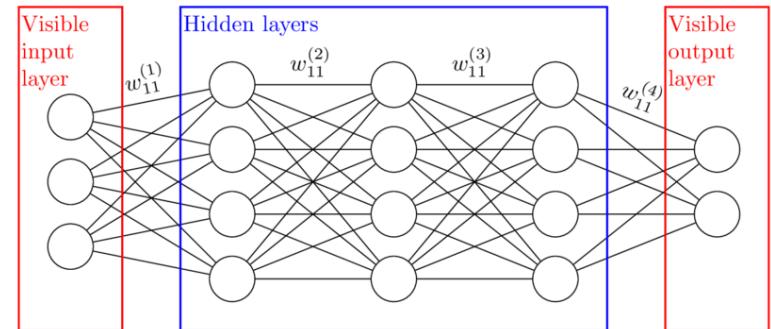
What is an Artificial Neural Network?



- 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 (2013) thanks to data availability, and computational power (!!!GPUs!!!)



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



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



How Neural Networks Learn?



- Learning process:
 - Neurons make decisions (activation functions)
 - Weights : connections between neurons are strengthened or weakened through training.
Randomly initialized.
- Training data
 - NN learn from historical data and examples
 - Labeled data are provided
- Backpropagation (Gradient descent, Chain rule)
 - Prediction on training data.
 - The difference (Loss) between the network's predictions and the actual data is calculated
 - Errors is used to adjust the model weights

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

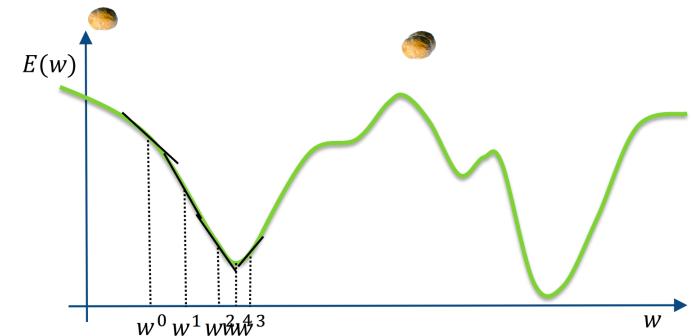
t_n : desired target

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



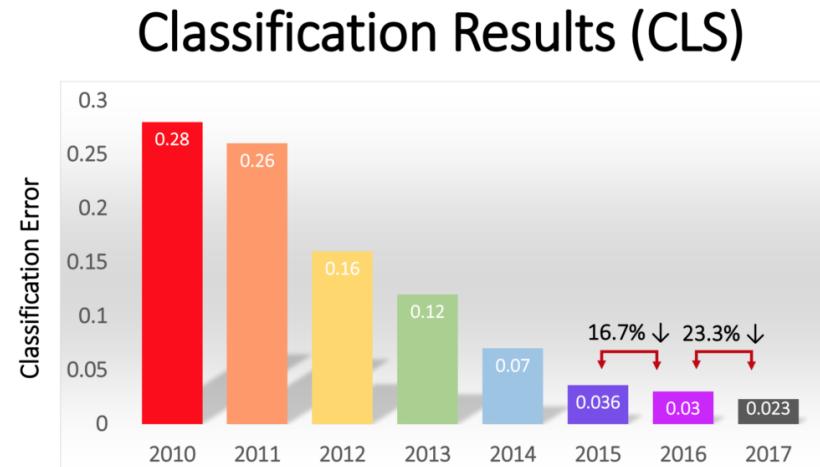
- Learning process:
 - Neurons make decisions (activation functions)
 - Weights : connections between neurons are strengthened or weakened through training.
Randomly initialized.
- Training data
 - NN learn from historical data and examples
 - Labeled data are provided
- Backpropagation (Gradient descent, Chain rule)
 - Prediction on training data.
 - The difference (Loss) between the network's predictions and the actual data is calculated
 - Errors is used to adjust the model weights

$$E = \sum_n^N (t_n - g(x_n|w))^2$$
$$w^{k+1} = w^k - \Delta \frac{\partial E(w)}{\partial w}$$





- 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:
 - GPUs
 - TPU (or ad hoc accelerators)
 - FPGAs

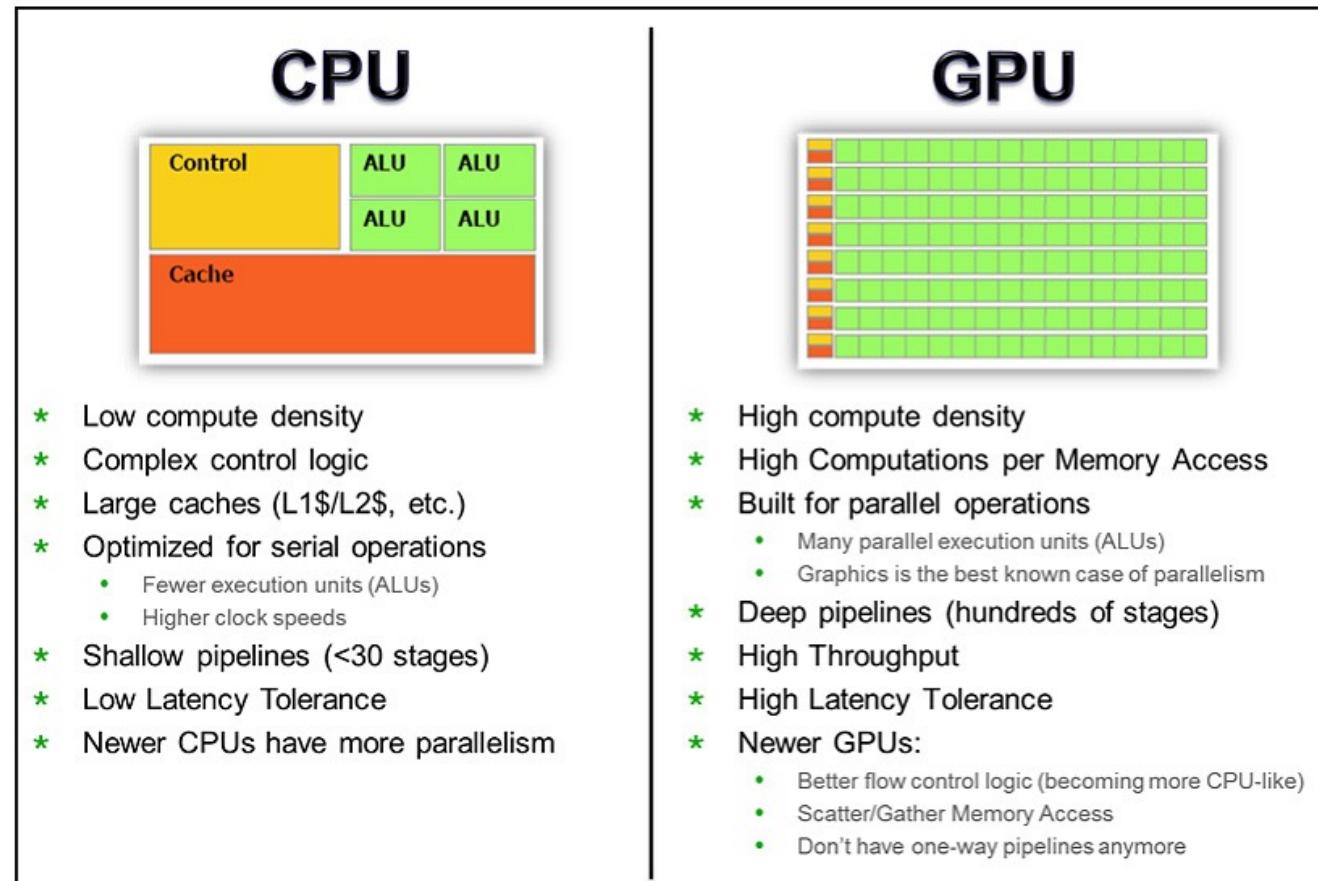




Graphical Processing Units (GPUs)

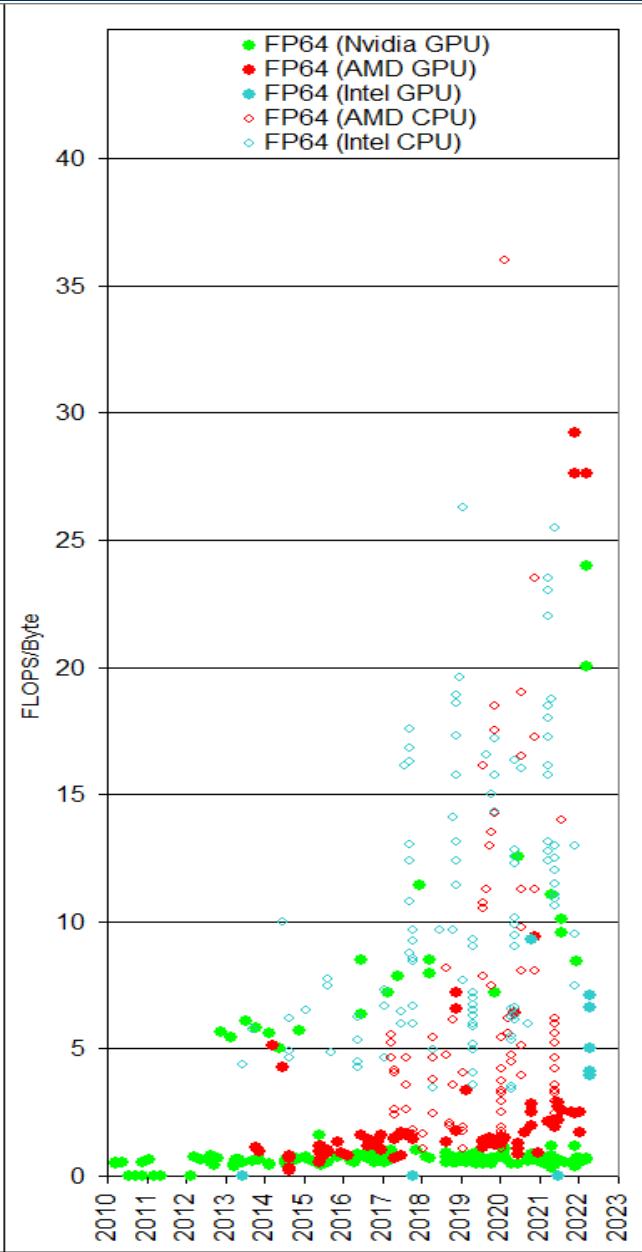
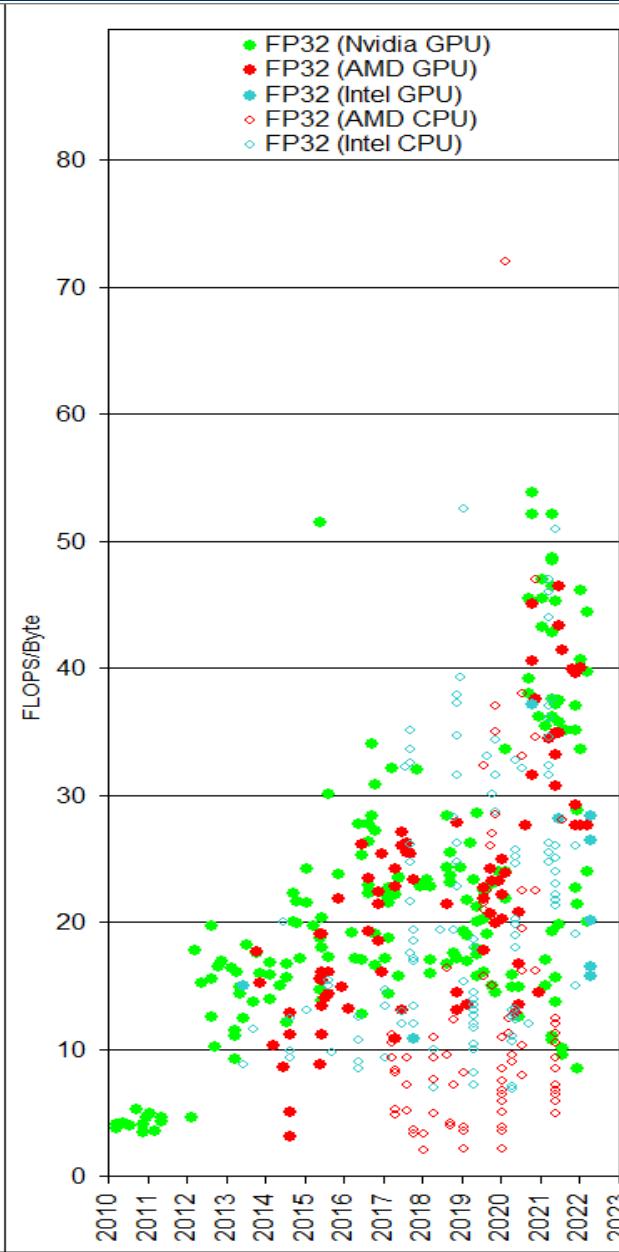
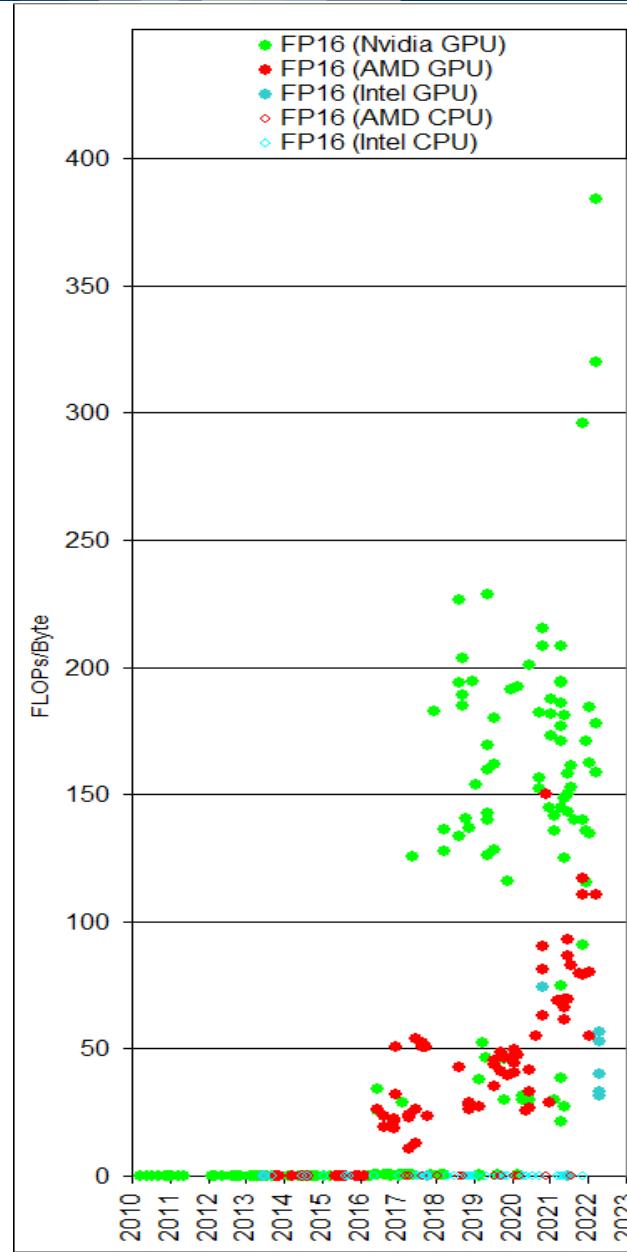


- Data-parallel computations: the same program is executed on many data elements in parallel
- The scientific codes are mapped onto the matrix operations.
- High-level languages (such as CUDA, OpenCL, OPENACC, OPENMP, SYCL) are required
- Up to 1000x faster than CPU



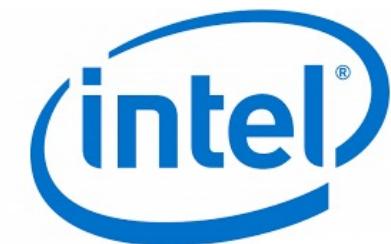
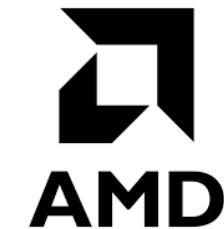


Not only one GPU type and vendor

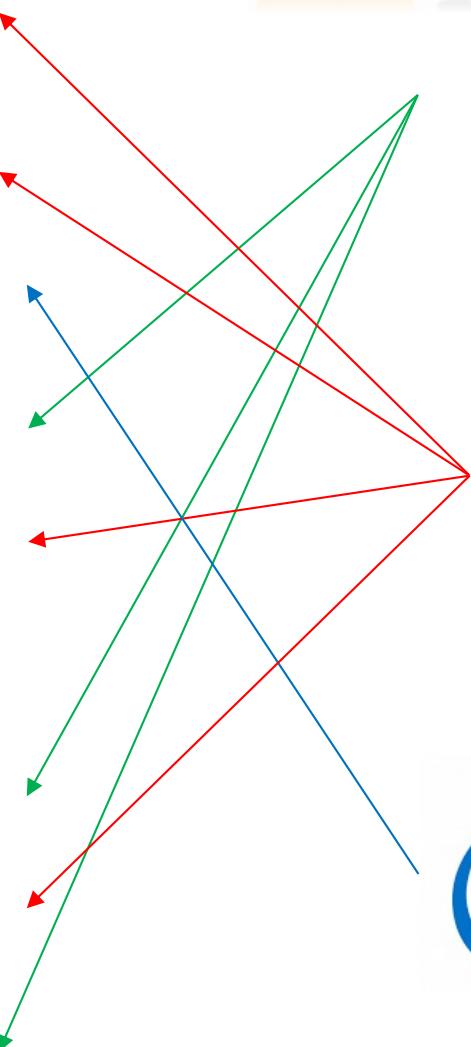




Not only one GPU type and vendor

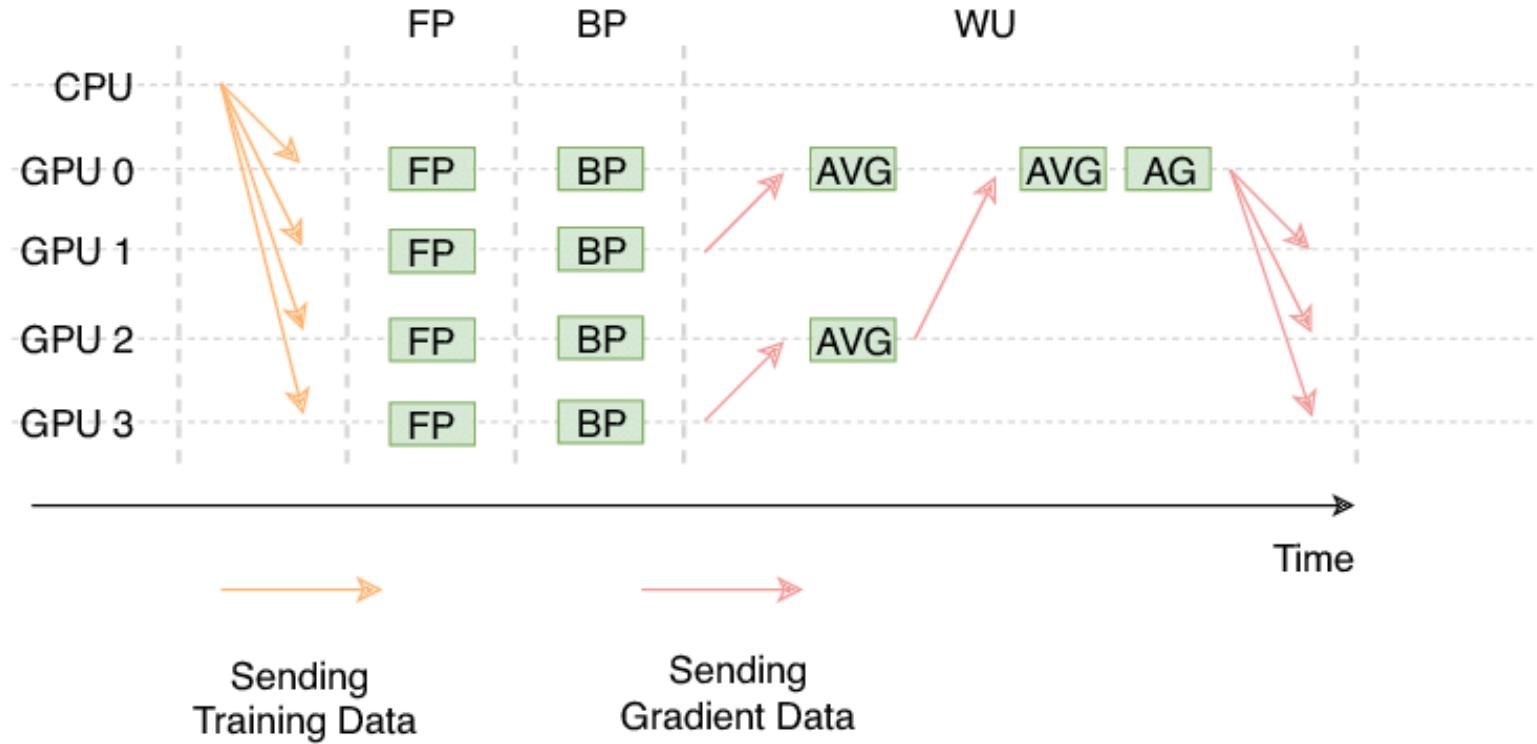


Rank	System	Cores	[PFlop/s]	[PFlop/s]	[kW]
1	El Capitan - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, 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





GPU: training a DNN on multiple GPUs



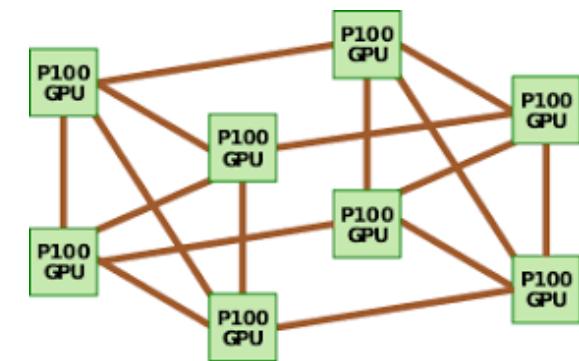
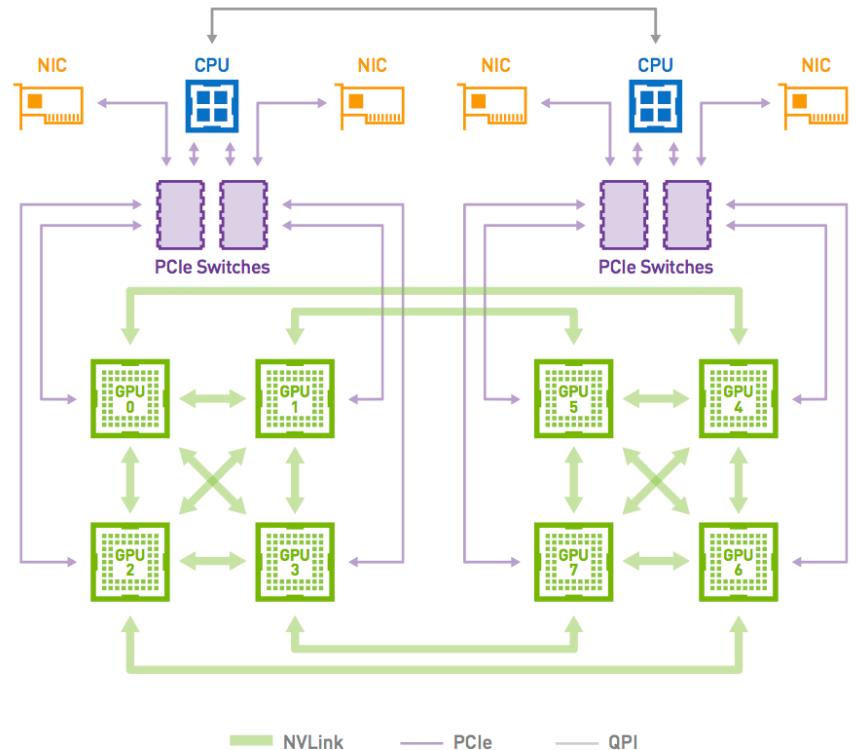
- The performance of such a synchronous system is limited by the slowest learner and slowest messages through the network.
- Since the communication phase is in the critical path, a high-performance network can enable fast reconciliation of parameters across learners



GPUs within the rack: PCI 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.

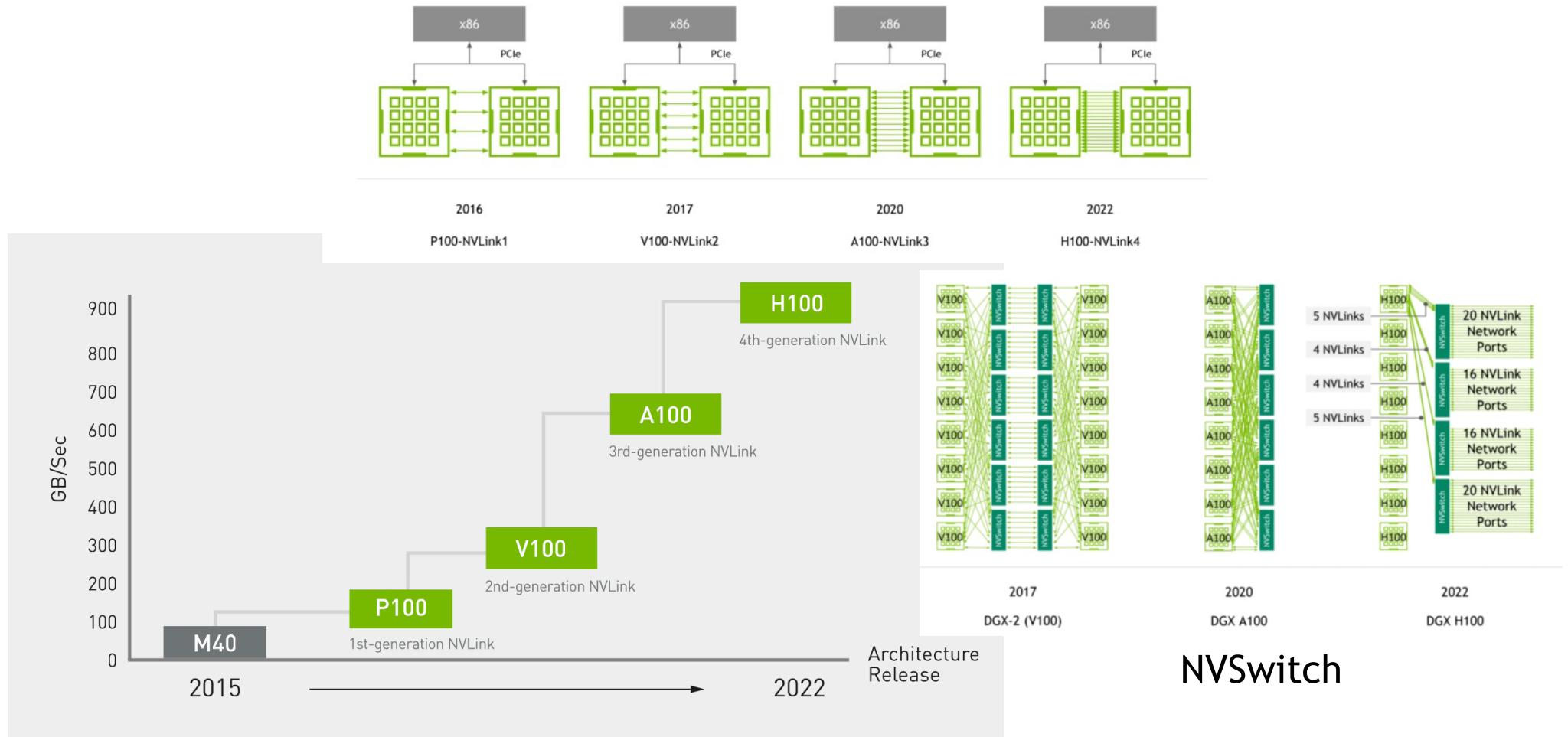




NVLINK evolution and NVSwitch

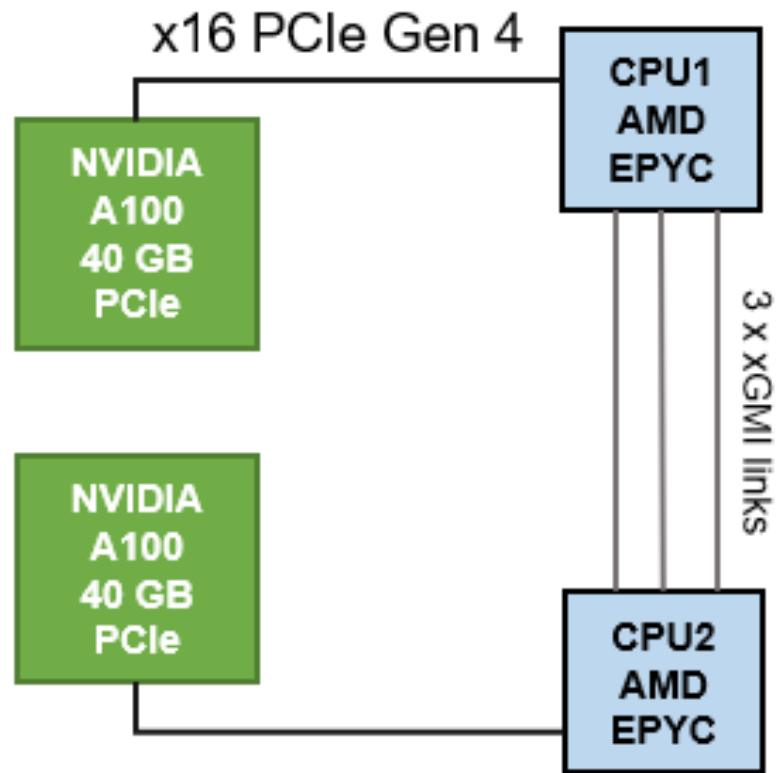
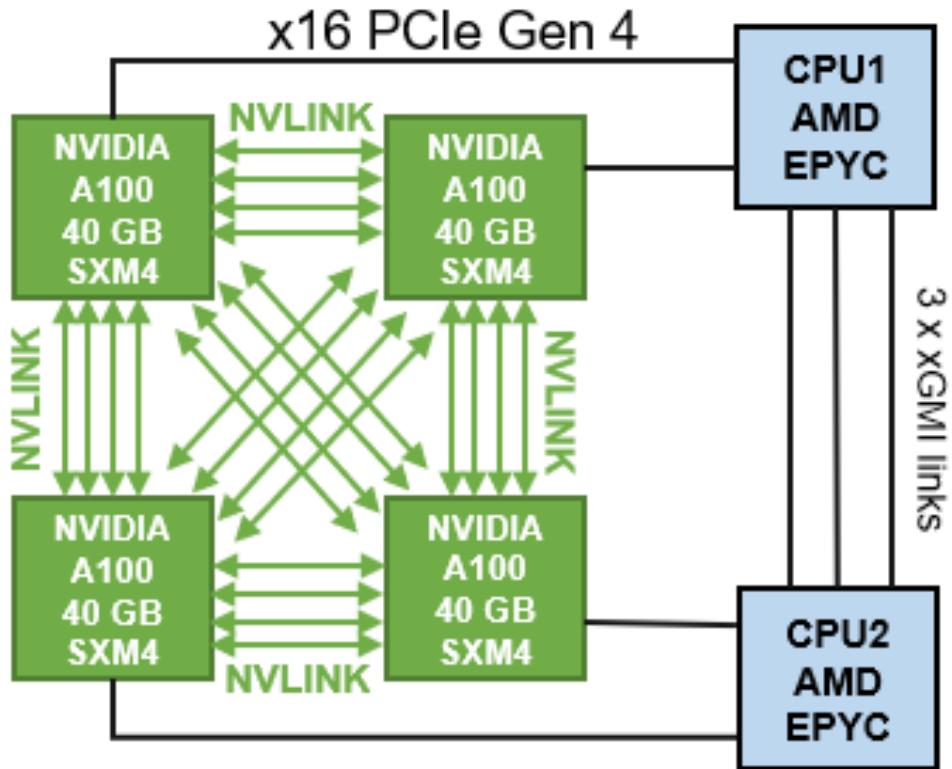


- Each NVLink lane supports a data rate of 50 Gb/s in each direction
- The total number of NVLink lanes increases from 6 lanes in the V100 GPU to 12 lanes in the A100 GPU and 18 for the H100 GPU





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



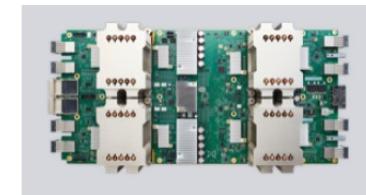
Tensor Processing Unit (TPU)



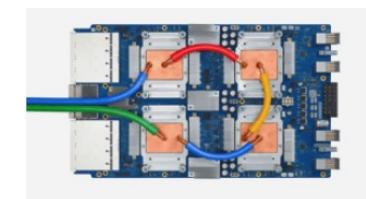
- While suited to ML, GPUs are still relatively general purpose devices
- In recent years designers further specialized them to ML-specific hardware
 - Custom-built integrated circuit developed specifically for machine learning and tailored for TensorFlow, Pytorch, or other ML frameworks
- Powering Google 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 - TPV5 focus both training and inference



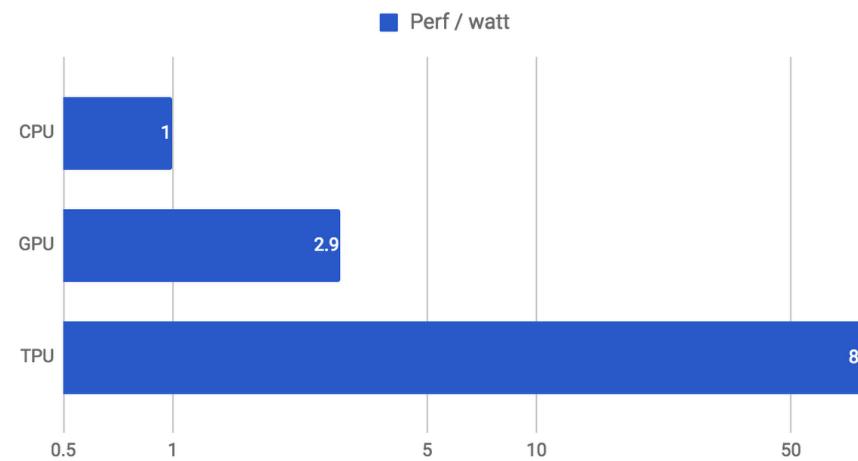
TPU V1



Cloud TPU v2

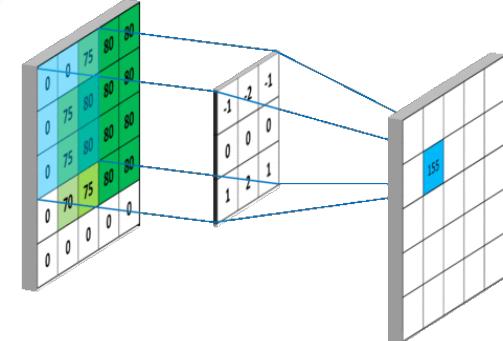
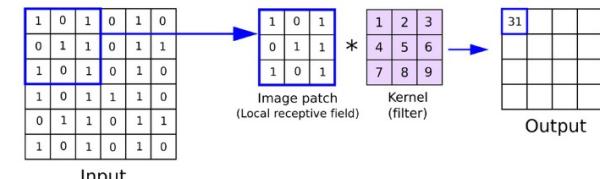
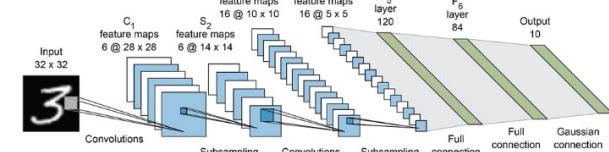
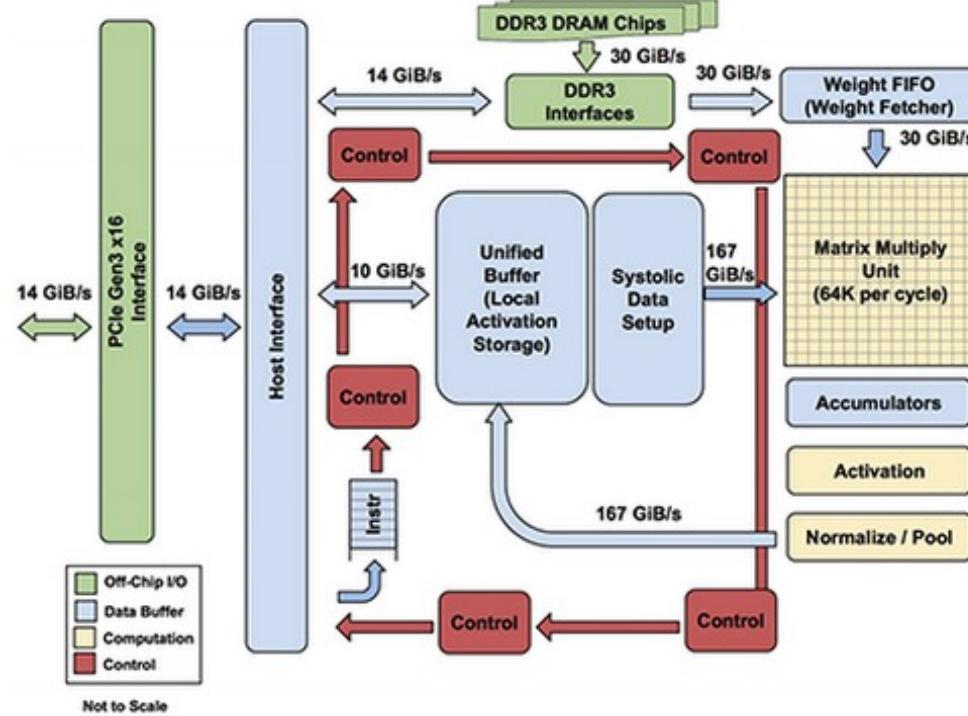


Cloud TPU v3





TPU Block Diagram



Inference

Training

THE CHAIN RULE STACKS

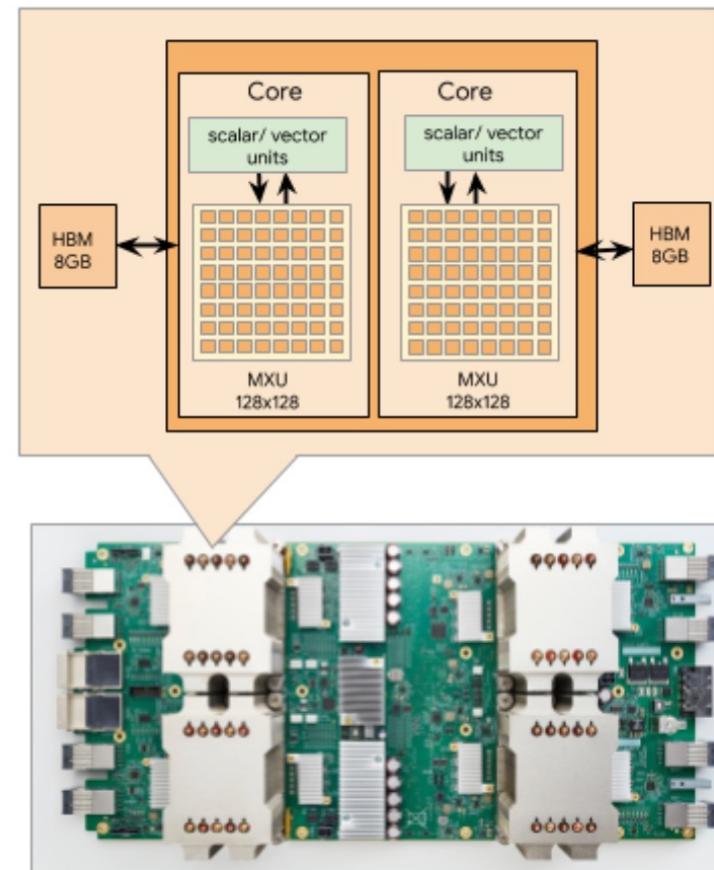
$$g(x) = a(b(c(x)))$$

3-layer composite function

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



- 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



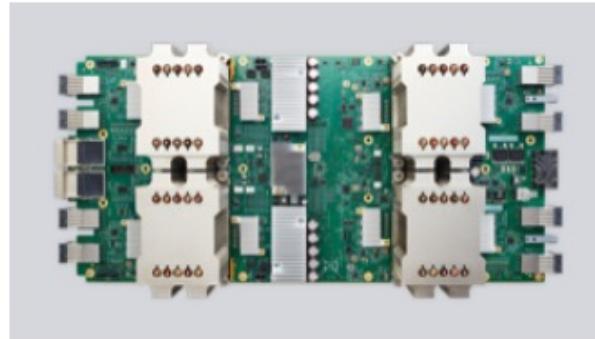
TPU v2 - 4 chips, 2 cores per chip



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 (beta)

11.5 petaflops

4 TB HBM

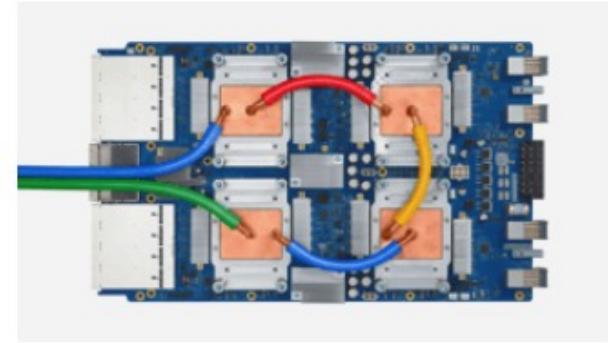
2-D toroidal mesh network



TPUv3 (liquid-cooled)



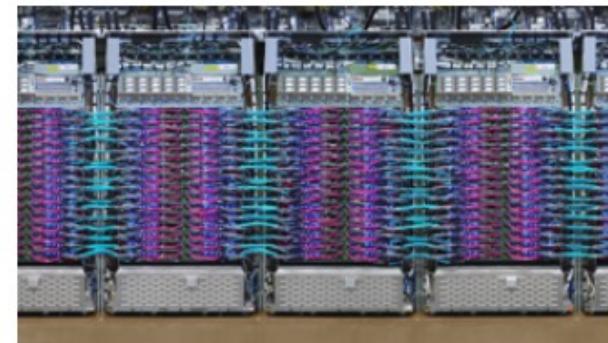
- TPUv3 is the first **liquid-cooled** accelerator in Google's data center.
- 2.5x faster than TPUv2
- Such supercomputing-class computational power supports:
 - new ML capabilities (e.g., AutoML)
 - rapid neural architecture search
- The v3 TPU Pod provides a maximum configuration of 256 devices for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

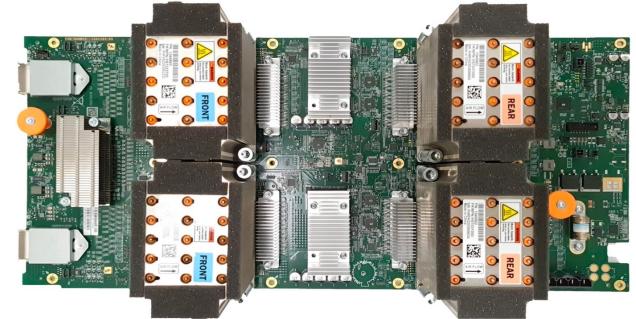
2-D toroidal mesh network



TPU v4-v5... 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)



Feature	TPU v1 (2016)	TPU v2 (2017)	TPU v3 (2018)	TPU v4 (2019)	TPU v5 (2021)
Process Technology	28nm	16nm	16nm	7nm	4nm
Memory per Core	8GB DDR	16GB HBM2	32GB HBM2	32GB HBM2	32GB HBM2e
Interconnect	Ring	Ring	Mesh	Mesh	Mesh
Pod Architecture	No	Yes	Yes	Yes	Yes
Applications	Inference	Training, inference	Training, inference	Training, inference	Training, inference
Performance (Compared to v1)	1x	15x	30x	40x	80x
Efficiency (Compared to v1)	1x	3x	6x	8x	10x



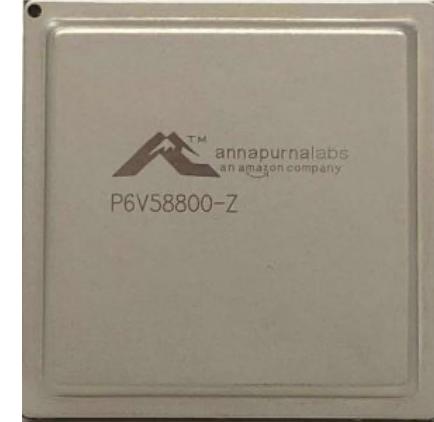
And what about AWS?



AWS Graviton4



AWS Trainium2



AWS Inferentia2

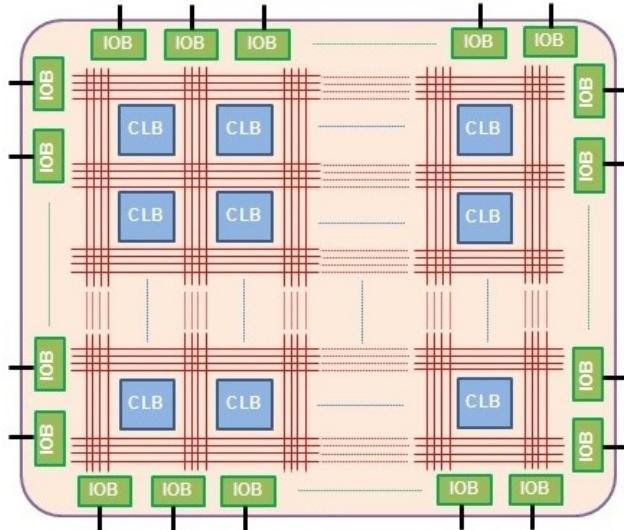
- Trainium2 optimized for LLMs
- Graviton4 based on ARM architecture, 30% more energy efficient for general AI tasks
- Inferentia2, for AI inference 2.3x higher throughput and up to 70% lower cost per inference than comparable EC2 instances.



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.



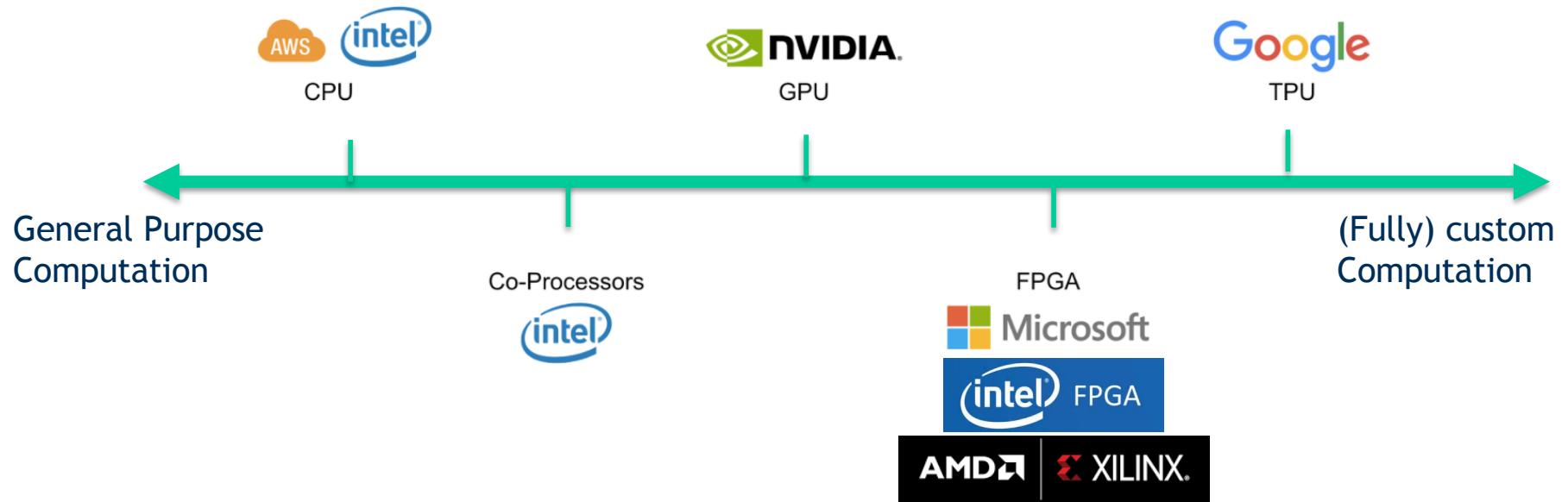
- Microsoft deployed FPGAs inside its Datacenters.
- Lower Carbon footprint since they are flexible, reconfiguration instead of reimplementation

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 comparison



	Advantages	Disadvantages
CPU	<ul style="list-style-type: none">• Easy to be programmed and support any programming framework.• Fast design space exploration and run your applications.	<ul style="list-style-type: none">• Suited only for simple AI models that do not take long to train and for small models with small training set
GPU	<ul style="list-style-type: none">• Ideal for applications in which data need to be processed in parallel like the pixels of images or videos.	<ul style="list-style-type: none">• Programmed in languages like CUDA and OpenCL and therefore provide limited flexibility compared to CPUs.
TPU	<ul style="list-style-type: none">• Very fast at performing dense vector and matrix computations and are specialized on running very fast program based on Tensorflow.	<ul style="list-style-type: none">• For applications and models based on the TensorFlow.• Lower flexibility compared to CPUs and GPUs
FPGA	<ul style="list-style-type: none">• Higher performance, lower cost and lower power consumption compared to other options like CPUs and GPU	<ul style="list-style-type: none">• Programmed using OpenCL and High-level Synthesis (HLS)• Limited flexibility compared to other platforms.



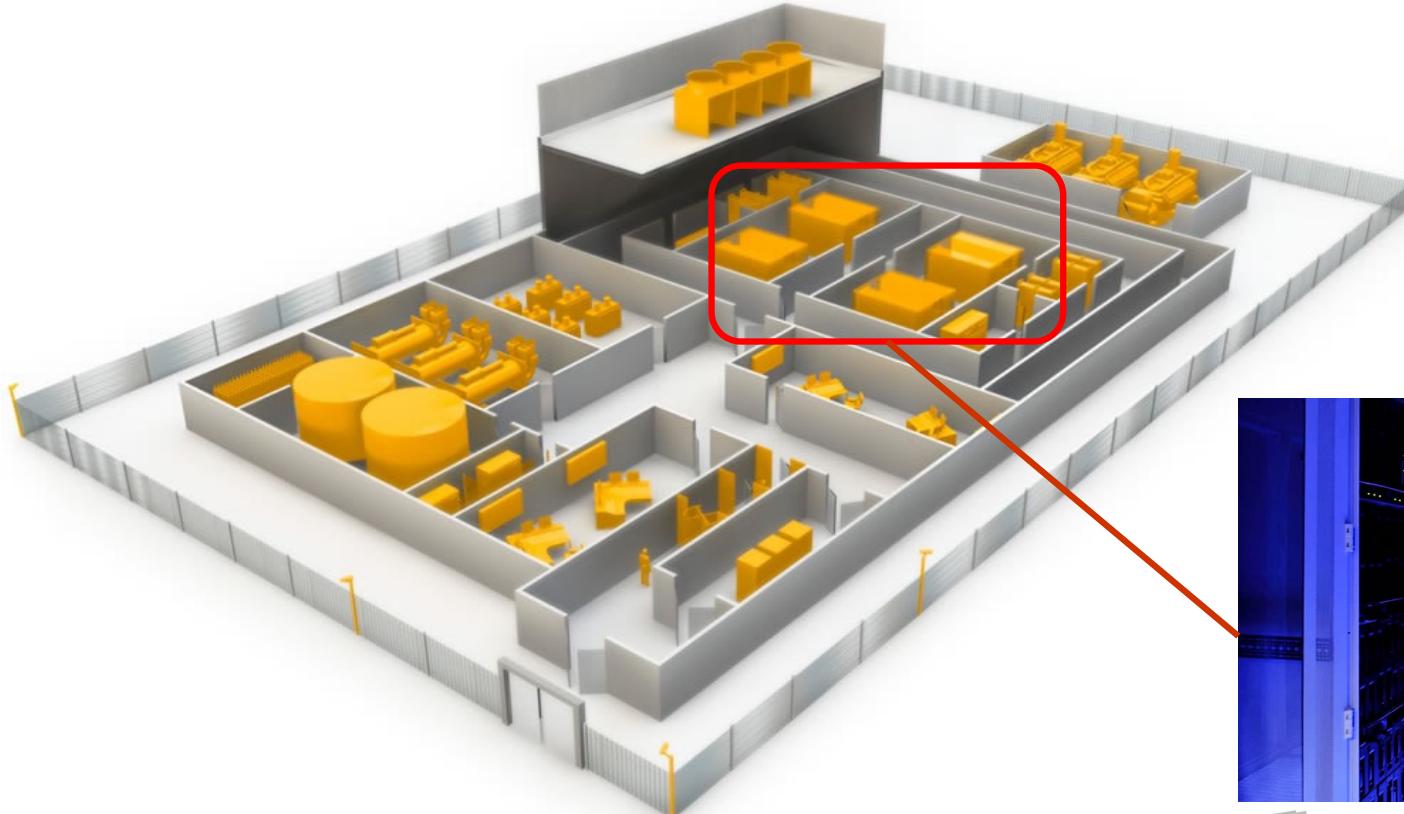
From the Rack to the Datacenter



Data-center architecture



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



Server corridors

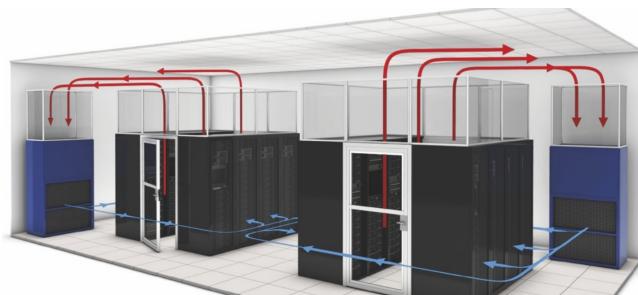
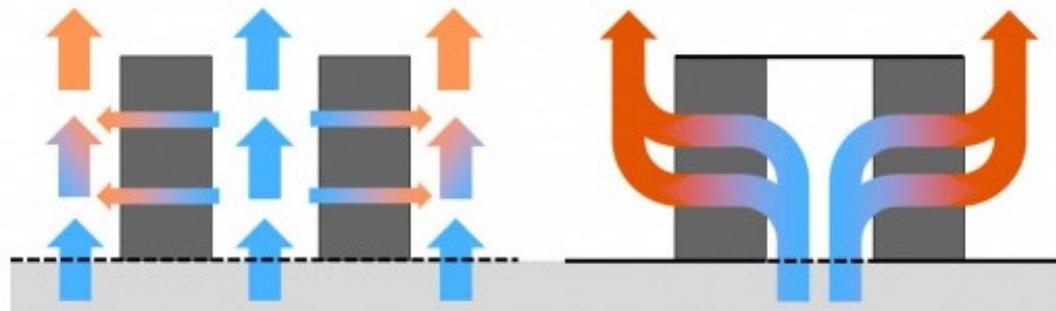




Data-center corridors



- 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

