Introduction to devices for edge computing

Dipartimento
di Ingegneria Chimica,
dei Materiali e della
Produzione Industriale
Università degli Studi
di Napoli Federico II

Speaker: Eng. Giulio Mattera

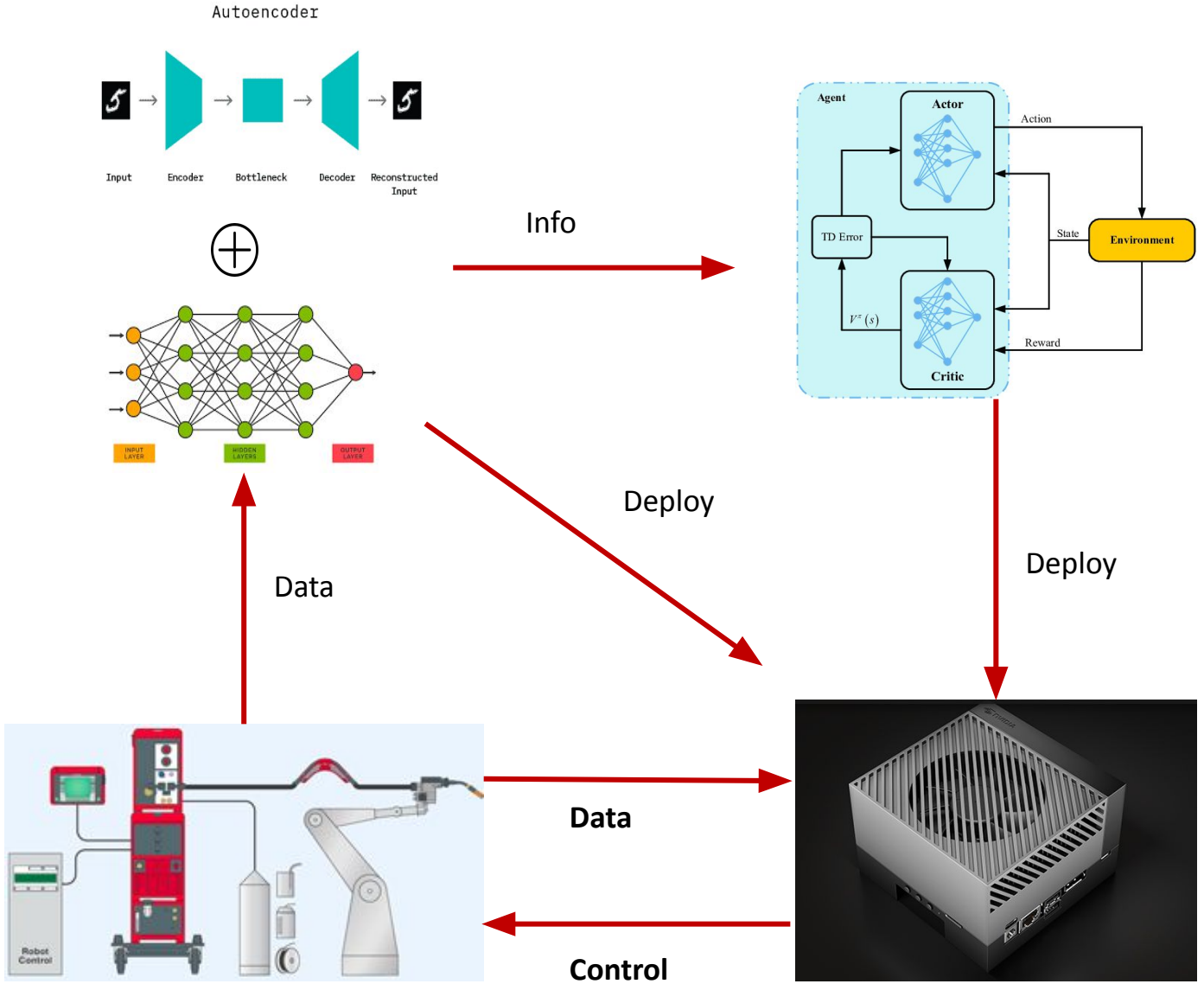
mail:giulio.mattera@unina.it

Dipartimento
di Ingegneria Chimica,
dei Materiali e della
Produzione Industriale
Università degli Studi
di Napoli Federico II

About myself

Giulio Mattera

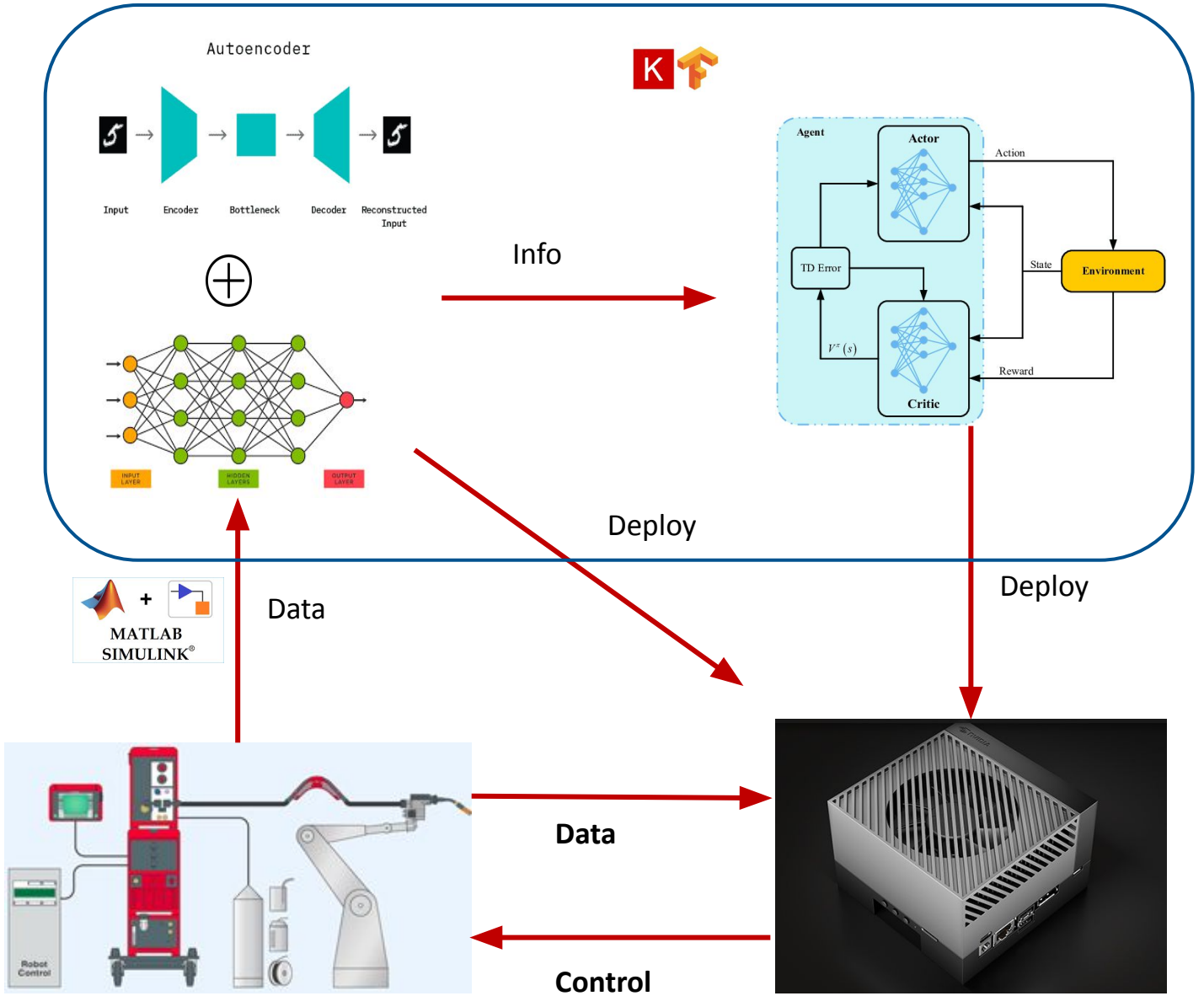
- PhD student at DICMAPI
- **Research topic**: Intelligent Robotics for WAAM
- **Contact**: giulio.mattera@unina.it



About myself

Giulio Mattera

- PhD student at DICMAPI
- **Research topic**: Intelligent Robotics for WAAM
- **Contact**: giulio.mattera@unina.it



Principal research topics

The **M**anufacturing with **A**dvanced **R**obotics **S**ystems (**M.A.R.S.**) research group born in the winter of 2019 from an idea of Prof. Luigi Nele, with the aim to shake up aeronautical industry and facilitate the transition to Industry 4.0.

The principal research topics are:

- Smart manufacturing
  - Advanced Additive Manufacturing with robotics systems
  - Smart welding
  - Smart machining
  - Digital twin for estimation of **RUL** (**R**emaining **U**seful **L**ife)
  - Smart factories
- Fabrication of components with non conventional materials

Introduction to devices for edge computing

Dipartimento
di Ingegneria Chimica,
dei Materiali e della
Produzione Industriale
Università degli Studi
di Napoli Federico II

Speaker: Eng. Giulio Mattera

mail:giulio.mattera@unina.it

Dipartimento
di Ingegneria Chimica,
dei Materiali e della
Produzione Industriale
Università degli Studi
di Napoli Federico II

Recap: Industry 4.0

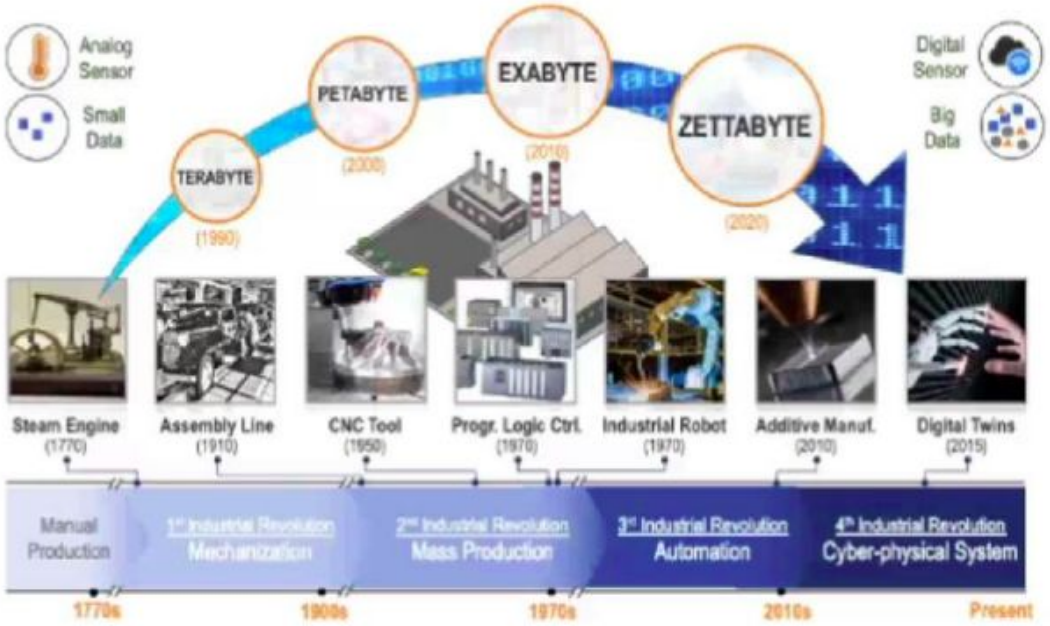
- "*Strong customization of products under conditions of highly flexibilized (mass-) production* "
- Green systems
- High customization
- High production
- Smart factories
- KETs
  - Modular and flexible systems
  - IoT
  - Cloud computing
  - Edge computing
  - Cyber security
  - Communication protocols
  - Cyber Physical System
  - Augmented reality
  - Artificial Intelligence

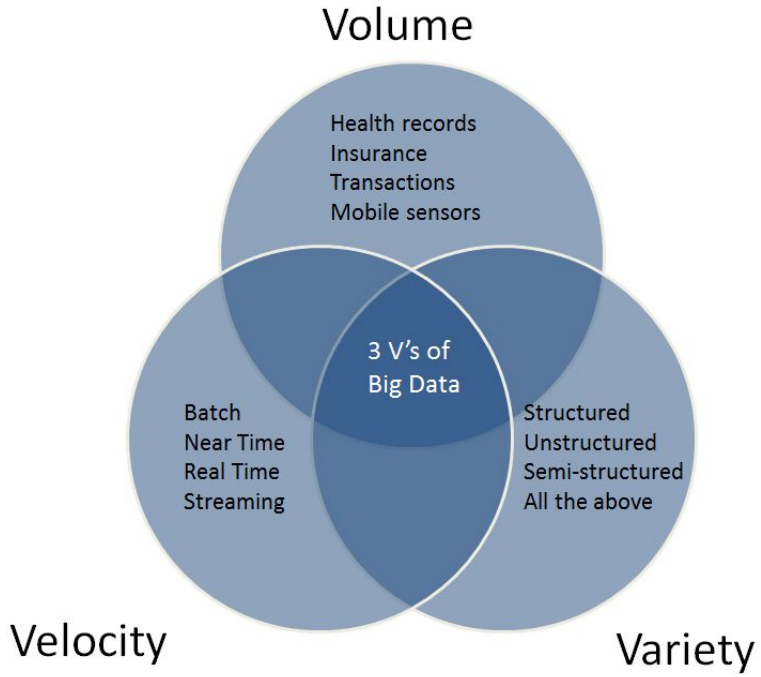
Recap: Industry 4.0

- "*Strong customization of products under conditions of highly flexibilized (mass-) production* "
- Green systems
- High customization
- High production
- Smart factories
- KETs
  - Modular and flexible systems
  - **IoT (Big data)**
  - Cloud computing
  - **Edge computing**
  - Cyber security
  - Communication protocols
  - **Cyber Physical System**
  - Augmented reality
  - Artificial Intelligence

Daily recap

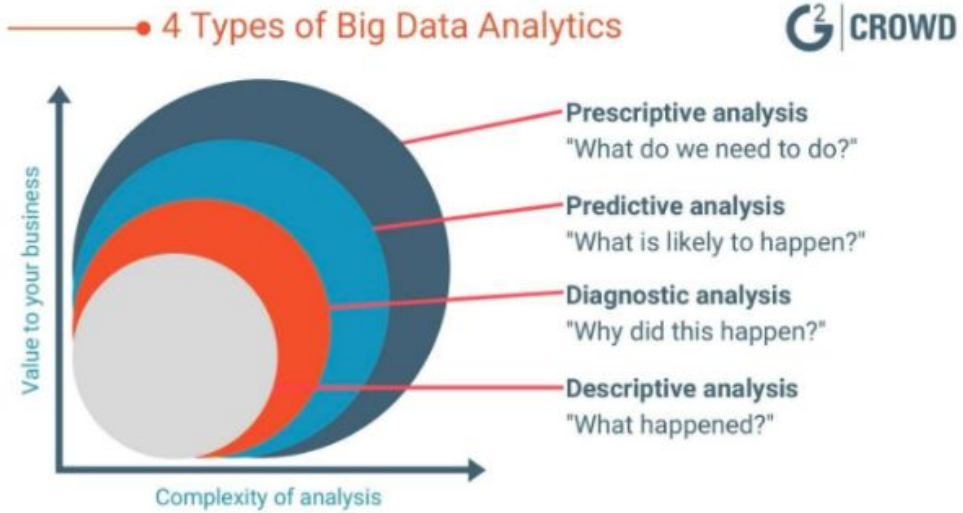
Recap : Big data



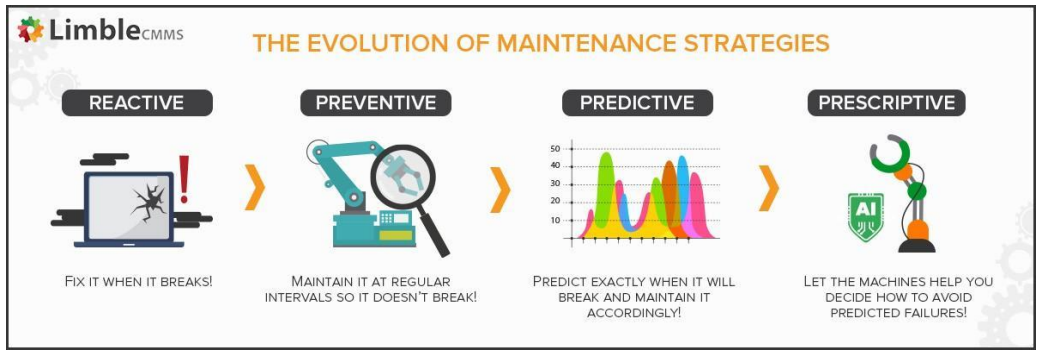


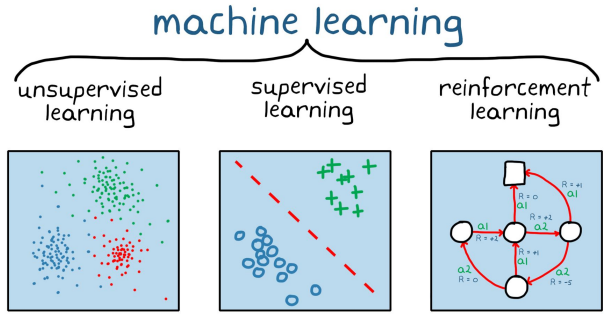
https://www.coforge.com/salesforce/blog/data-analytics/understanding-the-3-vs-of-big-data-volume-velocity-and-variety/

Recap : Big data

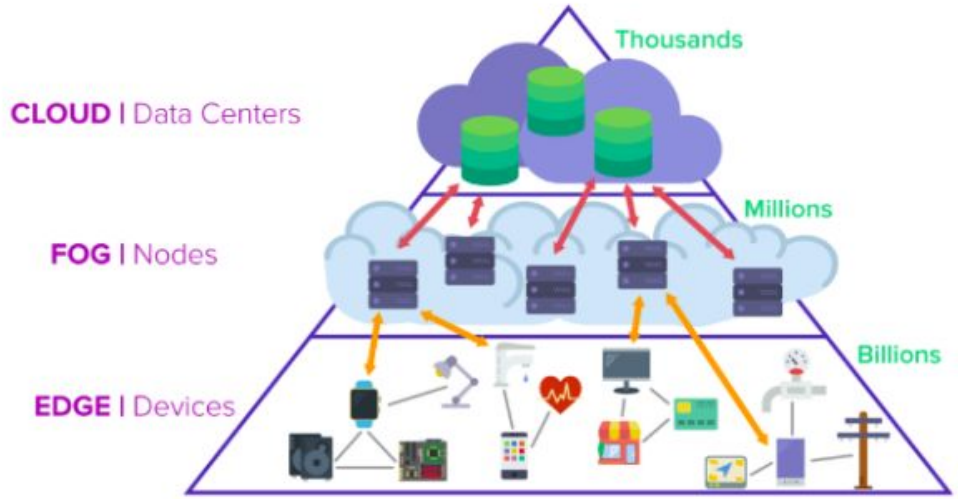


- **Descriptive analysis**: the set of tools oriented to describe the current and past situation of business processes and/or functional areas (e.g. Business Intelligence)
- **Predictive Analytics**: advanced tools that perform data analysis to answer questions about what might happen in the future (e.g. Predictive Maintenance)
- **Automated Analytics**: tools capable of autonomously implementing the proposed action according to the result of the analysis performed (e.g. Machine learning)





Recap : Cloud/Edge computing



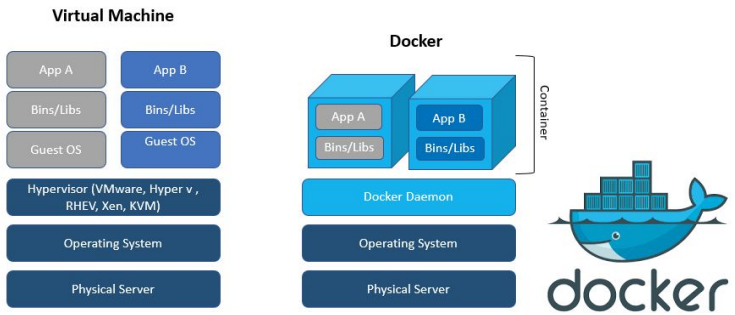
To overcome the limitations of the cloud, edge computing devices are increasingly used:

- Reduced latency (best for AI SW)
- Reduces data storage costs by sending only the data you need

amazon web services

Microsoft Azure

Google Cloud Platform



docker

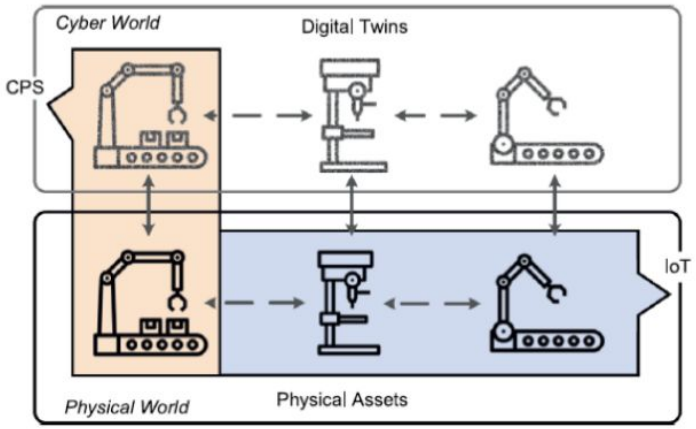
M.A.R.S. LABORATORY

DI C Ma PI Dipartimento di Ingegneria Chimica, dei Materiali e della Produzione Industriale Università degli Studi di Napoli Federico II

https://www.power-solutions.com/industry-trends/fog-computing-and-edge-computing-what-you-need-to-know/

https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/

Recap : Cyber Physical Production Systems



A cyberphysical system is a computer system capable of interacting continuously with the physical system in which it operates. The system is composed from physical elements endowed each one with computational ability:

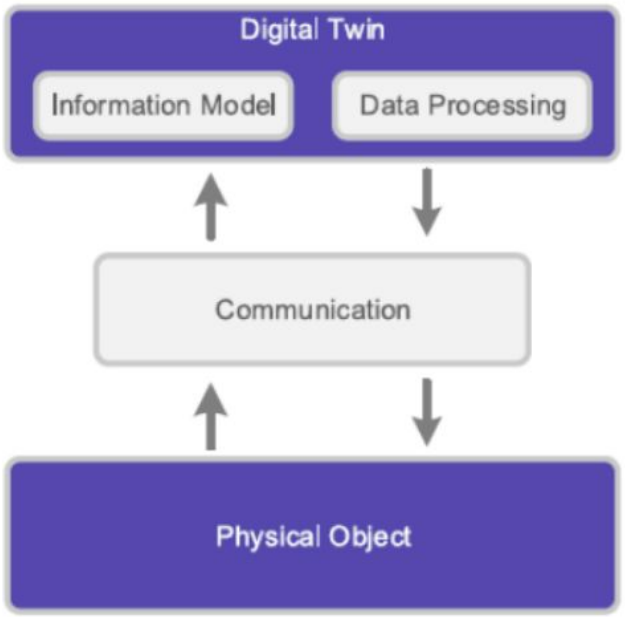
- Computational capacity
- Communication
- Capacity of control.

Devices that through communication protocol interacts directly and dynamically with the real world that surrounds it. At the base of the system, the single element is the embedded device

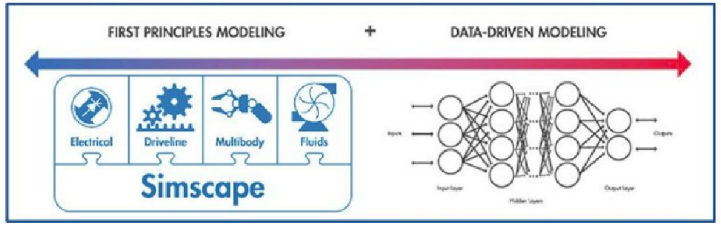
CPPSs consist of autonomous and cooperative subsystems of a manufacturing system, from processes to machines to production and logistics networks. The three characteristics that a CPPS must possess are:

- Intelligence (smartness), the various CPPS are able to acquire information from their surroundings and act autonomously.
- Connection, the ability to establish and use connections with other elements of the system-including humans-for cooperation and collaboration,
- Reactivity to internal and external changes

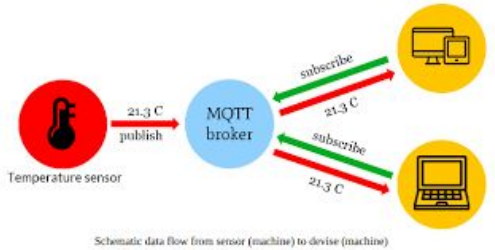
Recap : Digital twin



- An ultra-realistic simulation with high scalability, using the best available physical models, sensor data, and historical data for mirroring one or more real systems
- Simulation, which focuses on what-if scenarios, is one of the features of a DT, others are evolving in real time with the physical world, allowing:
  - monitoring
  - control
  - diagnostics and forecasting
  - All in a decentralized unit.

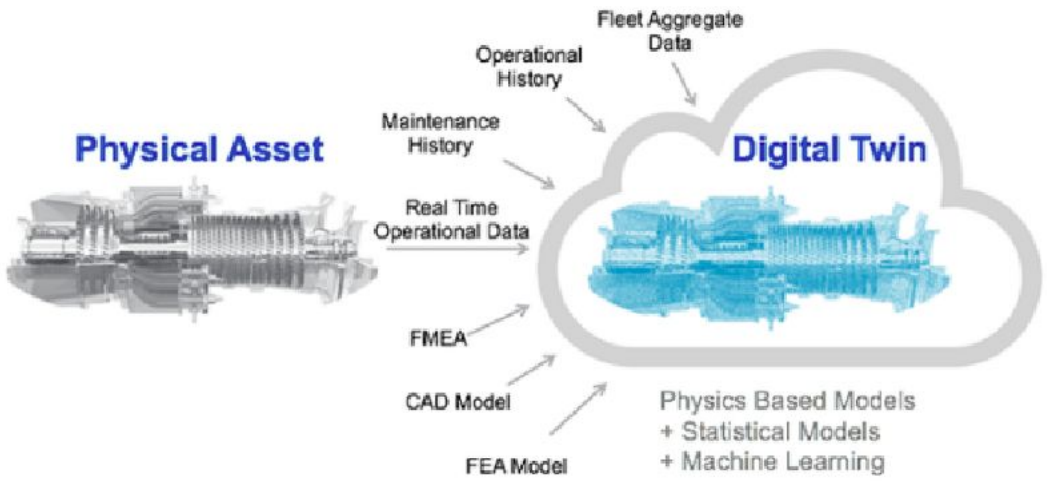


- **Information model**
- **Two-way communication mechanism with physical counterpart**
- **Data Processing Module**



Schematic data flow from sensor (machine) to device (machine)

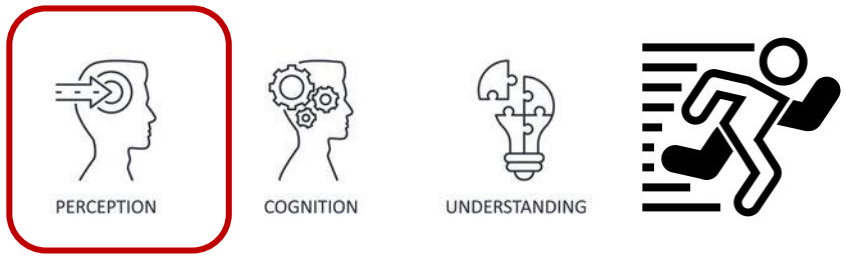
Recap : Digital twin



Constituent Elements:

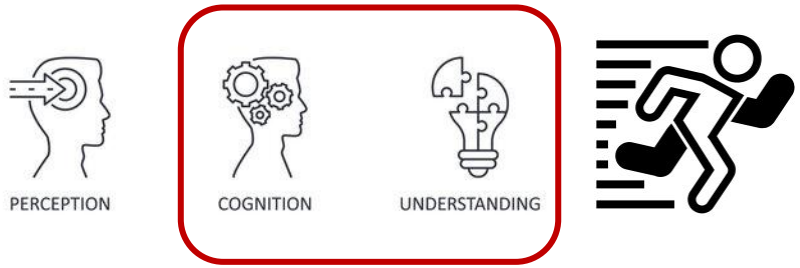
- **Sensors** that allow the Digital Twin to capture operational and environmental data.
- **Actuators** that operate on the production process in order to optimize it;
- **Synthetic data**: come from the digital world and are nothing more than aggregations of multiple pieces of information received via sensors from the physical world, such as data-driven models, state observers, inverse dynamic models, etc.
- **Techniques for processing**, such as machine learning techniques.

Recap : Digital twin



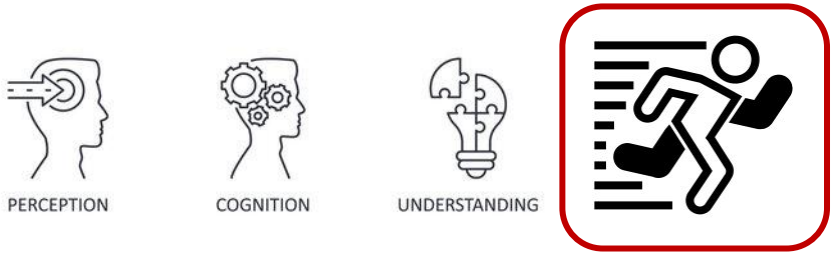
- **See (perception module)**
- Think
- Do

Recap : Digital twin



- **See (perception module)**
- **Think (data analysis in cognition module)**
- Do

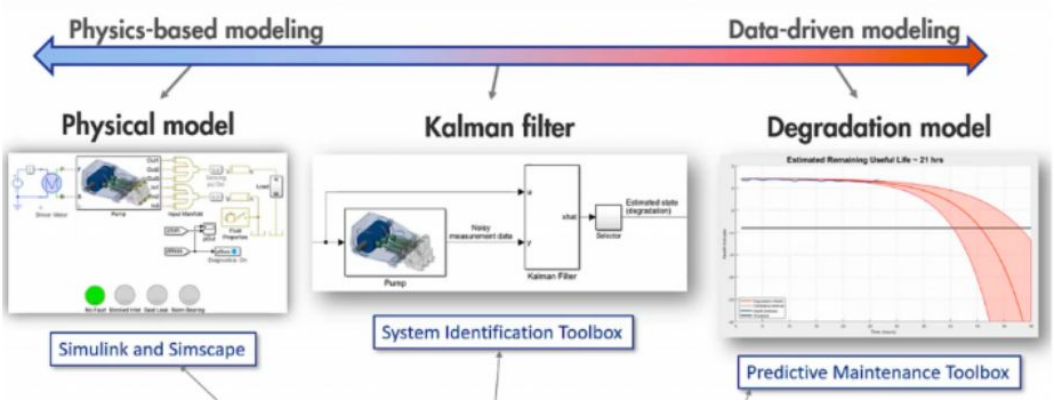
Recap : Digital twin



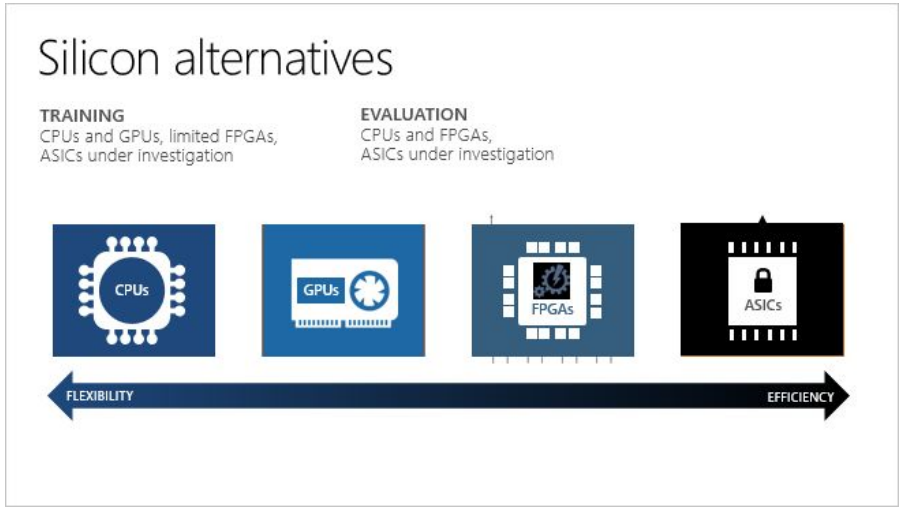
- See (perception module)
- Think (data analysis in cognition module)
- Do (action module)

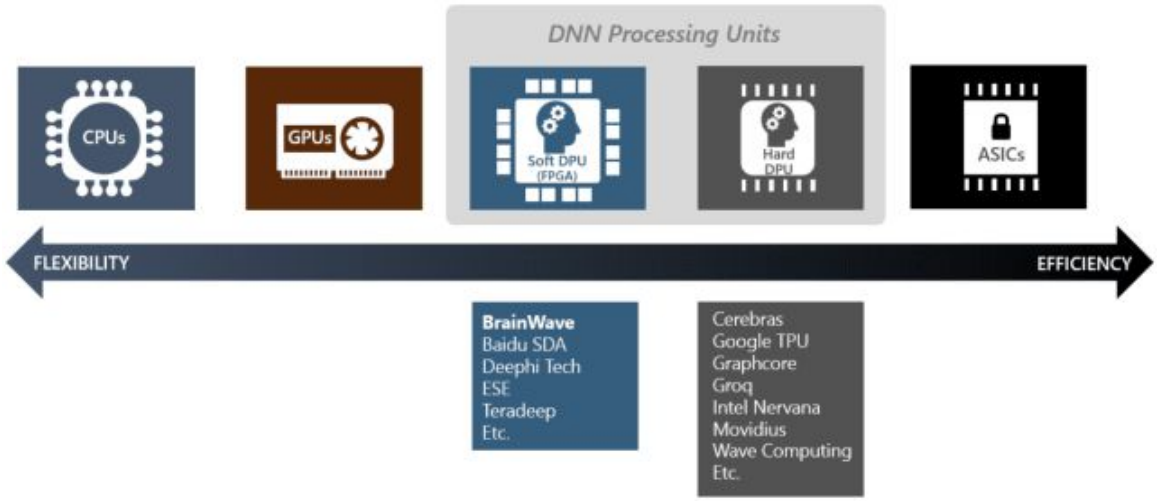
Recap : Digital twin

- From a system perspective, it remains a challenge to integrate various models with different dimensions, different spatial and temporal scales.
- Lack of effective tools for 5D modeling
- More complex models suffer from latency issues
- Intensive use of data-driven models has made it possible to develop models even in the absence of extensive system knowledge and high inference rates, however the scalability and flexibility of these models is not guaranteed
- The most innovative solutions make use of gray-models



Computing devices:





https://github.com/gopala-kr/a-week-in-wild-ai/blob/master/01-ai-accelerators/README.md

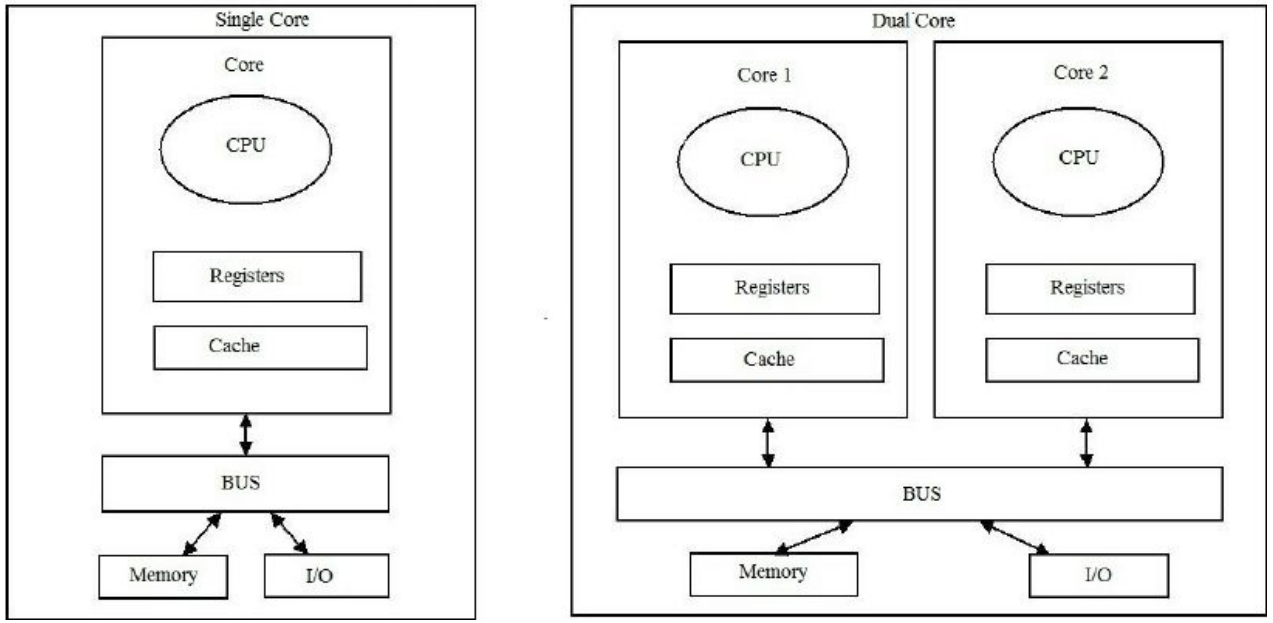
CPU (I)

- The CPU (Central Processing Unit) is the main component of a computer
  - A single-core processor is a microprocessor with a single core on a chip, running a single thread at any one time
    - At the hardware level, a computer executes sequences of individual instructions
    - Each instruction tells the computer to add, subtract, multiply….

CPU (II)

- The CPU (Central Processing Unit) is the main component of a computer
  - A single-core processor is a microprocessor with a single core on a chip, running a single thread at any one time
    - At the hardware level, a computer executes sequences of individual instructions
    - Each instruction tells the computer to add, subtract, multiply….

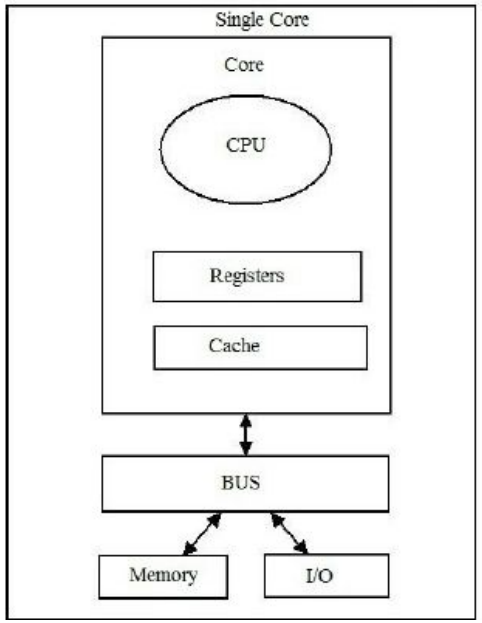
- The shift from single- core to multi-core architectures allowed to integrate more complex SW-C and **parallelism**

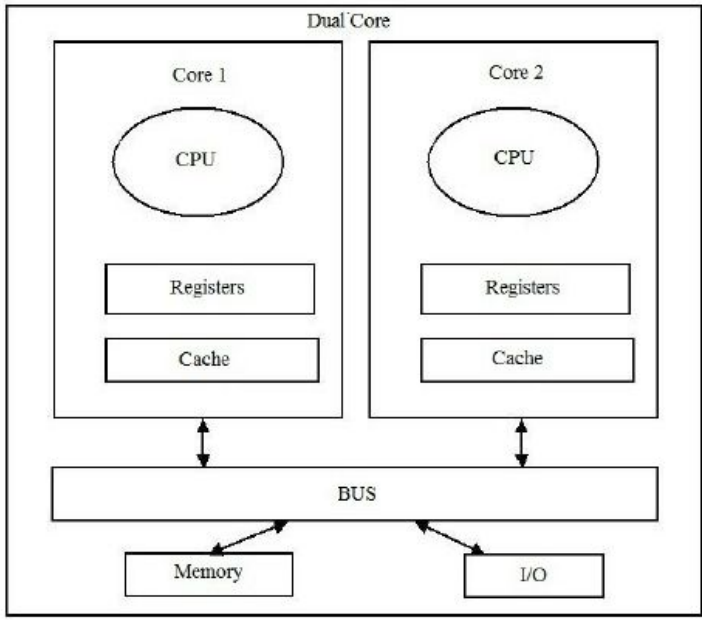


CPU (II)

- The CPU (Central Processing Unit) is the main component of a computer
  - A single-core processor is a microprocessor with a single core on a chip, running a single thread at any one time
    - At the hardware level, a computer executes sequences of individual instructions
    - Each instruction tells the computer to add, subtract, multiply….

- The shift from single- core to multi-core architectures allowed to integrate more complex SW-C and **parallelism**





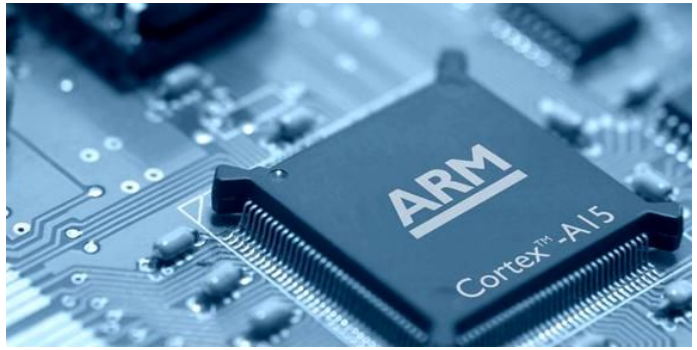
- Not all codes can be parallelized
  - **parallelization** is simply running different parts of the computer code simaltaneously.
  - **race condition :** when the end result of executing a series of processes depends on the timing or sequence with which they are executed. With parallel computing you must manage the timing or sequence through atomic operations. An atomic operation consists of an execution operation that is logically indivisible, i.e., if no other operation can begin before the first one is finished, and therefore there can be no interleaving (messy process followed by a race condition).

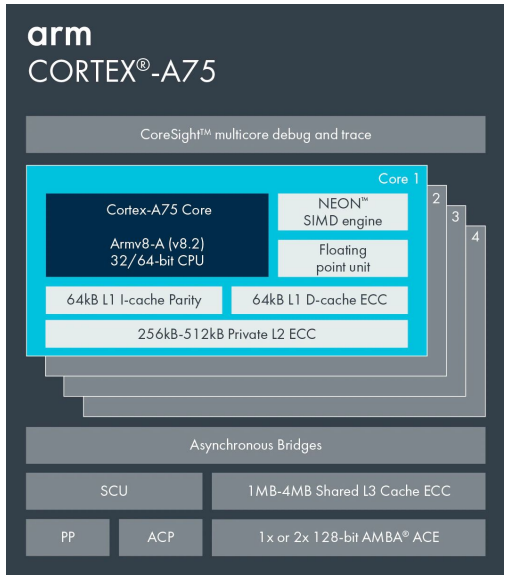
https://gist.github.com/peterhurford/3e2f88f549afc390c204

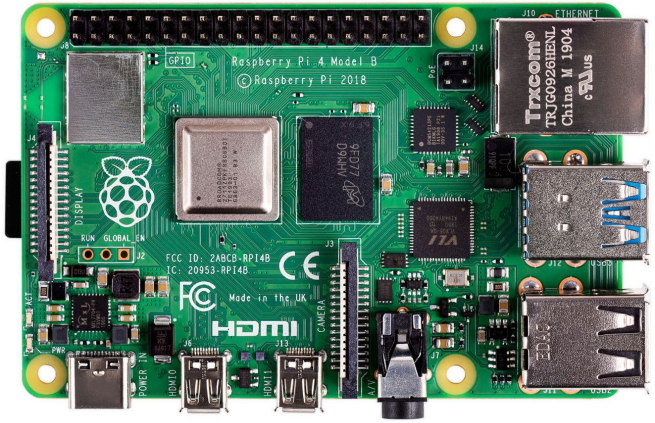
CPU (III)

- A microcontroller (MCU or µC) is essentially a single-chip which incorporates a microprocessor
  - Is a small computer on a single metal-oxide-semiconductor (MOS) integrated circuit (IC) chip

- A microcontroller contains one or more CPUs (processor cores) along with memory and programmable input/output peripherals
  - Some microcontrollers execute code directly from the flash memory to execute code fast after powering up
  - Microcontrollers are generally used to execute small purpose-built programs in automatically controlled products and devices, such as automobile engine control systems, implantable medical devices, remote controls….





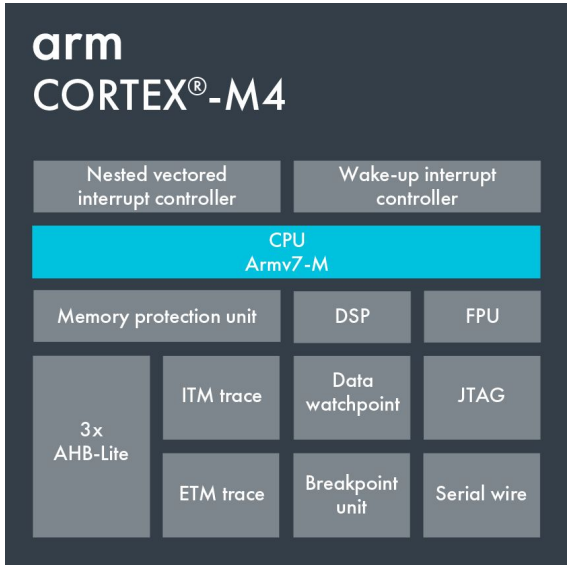


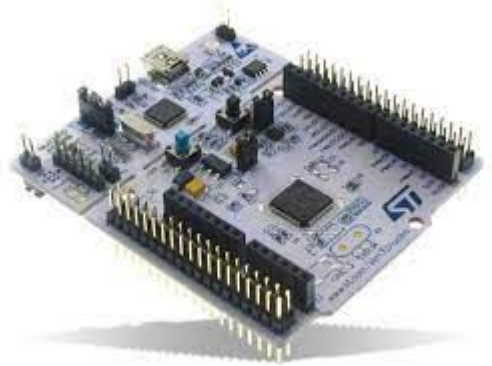
https://www.alphr.com/features/390064/arm-vs-intel-processors-what-s-the-difference/

CPU (IV)

- A microcontroller (MCU or µC) is essentially a single-chip which incorporates a microprocessor
  - Is a small computer on a single metal-oxide-semiconductor (MOS) integrated circuit (IC) chip

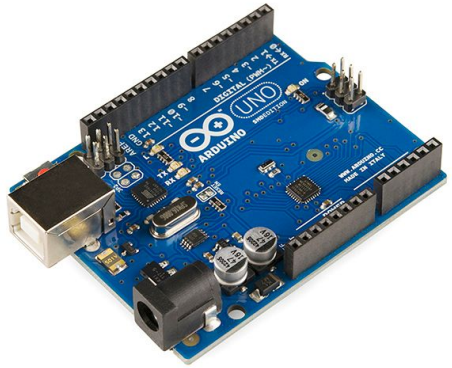
- A microcontroller contains one or more CPUs (processor cores) along with memory and programmable input/output peripherals
  - Some microcontrollers execute code directly from the flash memory to execute code fast after powering up
  - Microcontrollers are generally used to execute small purpose-built programs in automatically controlled products and devices, such as automobile engine control systems, implantable medical devices, remote controls….





STM32 Nucleo F401RE (512k) 24-48 MHz

CPU ARM Cortex M4F (single-core)



32 KB (ATmega328P) 16 Mhz

CPU (V)

Single-Board Computer vs. Microcontroller

Rough Specifications

	Microcontrollers	Single-Board Computers
Processor Speed	~50 Mhz	1 Ghz +
Onboard Storage	64 Kb	Flash, SD cards ~Gb
Memory (RAM)	64 Kb	~ 1 Gb
Power Consumption	12 mA (2200 mAh battery -> 183hr)	500 mA+ (2200 mAh battery-> 4.4 hr)
Reboot Time	<1 sec	~ Multiple seconds
Other Features		Operating system Extendable Storage Network Connection

beagleboard.org

ARDUINO

Raspberry Pi

life.augmented

NVIDIA JETSON

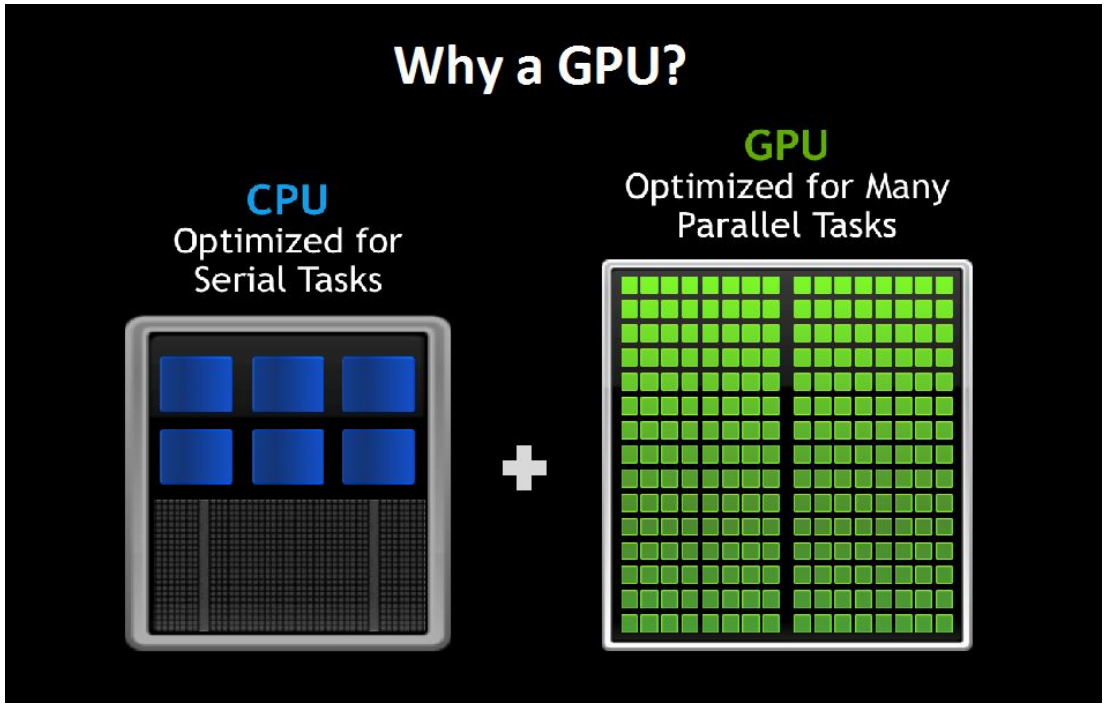
Texas Instruments

https://all-bout-secret.blogspot.com/2016/10/different-about-microcontroller-and.html

NVIDIA : GPUs (I)

CPU vs GPU

CPU	GPU
Central Processing Unit	Graphics Processing Unit
Several cores	Many cores
Low latency	High throughput
Good for serial processing	Good for parallel processing
Can do a handful of operations at once	Can do thousands of operations at once



http://gpu.di.unimi.it/cuda.html

https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html

https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/

NVIDIA : GPUs (II)

TFLOPs = Tera Floating Point Operation for second

Nano

Module Technical Specifications	
GPU	NVIDIA Maxwell™ architecture with 128 NVIDIA CUDA® cores 0.5 TFLOPs (FP16)
CPU	Quad-core ARM® Cortex®-A57 MPCore processor
Memory	4 GB 64-bit LPDDR4 1600MHz - 25.6 GB/s

TX2

Specifiche tecniche	
Prestazioni IA	1,33 TFLOPS
GPU	Architettura NVIDIA Pascal™ con 256 core NVIDIA® CUDA®
CPU	CPU dual-core NVIDIA Denver 2 64-bit e processore quad-core Arm® Cortex®-A57 MPCore
Memoria	LPDDR4 4 GB 128-bit 51.2 GB/s

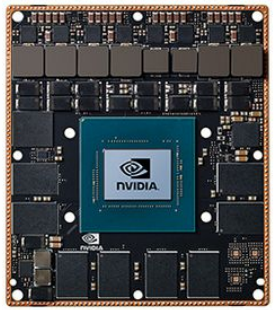
https://www.nvidia.com/it-it/autonomous-machines/embedded-systems/

NVIDIA :GPUs (III)

	Serie Jetson Xavier NX		Serie Jetson AGX Xavier			Serie Jetson Orin NX		Serie Jetson AGX Orin	
	Jetson Xavier NX 16 GB	Jetson Xavier NX	Jetson AGX Xavier 64 GB	Jetson AGX Xavier	Jetson AGX Xavier Industrial	Jetson Orin NX 8 GB	Jetson Orin NX 16 GB	Jetson AGX Orin 32 GB	Jetson AGX Orin 64 GB
Prestazioni IA	21 TOPS		32 TOPS		30 TOPS	70 TOPS	100 TOPS	200 TOPS	275 TOPS
GPU	GPU NVIDIA Volta™ a 384 core con 48 Tensor Core		GPU NVIDIA Volta a 512 core con 64 Tensor Core			GPU NVIDIA Ampere a 1024 core con 32 Tensor Core		GPU NVIDIA Ampere a 1792 core con 56 Tensor Core	GPU NVIDIA Ampere a 2048 core con 64 Tensor Core
CPU	CPU NVIDIA Carmel Arm®v8.2 64-bit a 6 core 6 MB L2 + 4 MB L3		CPU NVIDIA Carmel Arm®v8.2 64-bit a 8 core 8 MB L2 + 4 MB L3			CPU Arm® Cortex®-A78AE 8-core v8.2 64-bit 1,5 MB L2 + 4 MB L3	CPU Arm® Cortex®-A78AE 6-core v8.2 64-bit 2 MB L2 + 4 MB L3	CPU Arm® Cortex®-A78AE 8-core v8.2 64-bit 2 MB L2 + 4 MB L3	CPU Arm® Cortex®-A78AE 12-core v8.2 64-bit 3 MB L2 + 6 MB L3

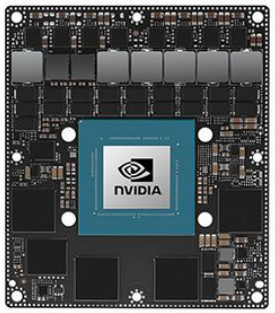
https://www.nvidia.com/it-it/autonomous-machines/embedded-systems/jetson-orin/

NVIDIA :GPUs (IV)



Jetson AGX Xavier 64 GB
Matrice densa 32 INT8 TOPS
Da 10 W a 30 W
$ 1299 (1KU+)

Jetson AGX Xavier 32 GB
Matrice densa 32 INT8 TOPS
Da 10 W a 30 W
$ 899 (1KU+)

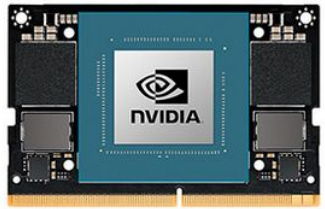


Jetson AGX Orin 64 GB
Matrice sparsa 275 TOPS | Matrice densa 138 INT8
Da 15 W a 60 W
$ 1599 (1KU+)

Jetson AGX Orin 32 GB
Matrice sparsa 200 TOPS | Matrice densa 100 INT8 TOPS
Da 15 W a 40 W
$ 899 (1KU+)

Jetson Xavier NX 16 GB
Matrice densa 21 INT8 TOPS
Da 10 W a 20 W
$ 499 (1KU+)

Jetson Xavier NX 8 GB
Matrice densa 21 INT8 TOPS
Da 10 W a 20 W
$ 399 (1KU+)



Jetson Orin NX 16 GB
Matrice sparsa 100 |Matrice densa 50, INT8 TOPS
Da 10 W a 25 W
$ 599 (1KU+)

Jetson Orin NX 8 GB
Matrice sparsa 70 |Matrice densa 35, INT8 TOPS
Da 10 W a 20 W
$ 399 (1KU+)

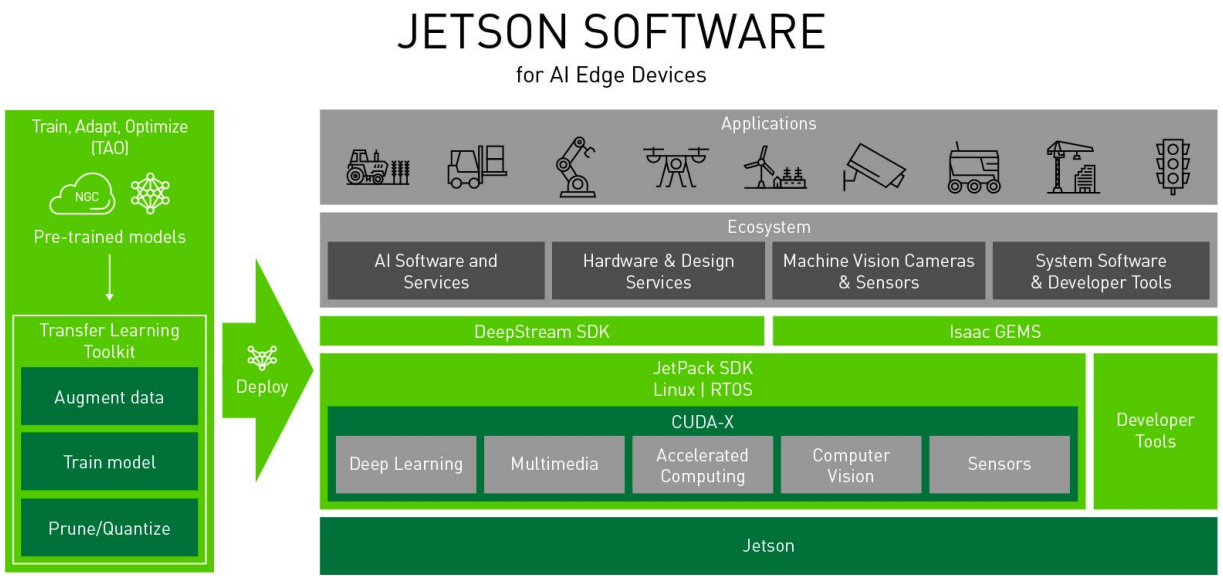
https://www.nvidia.com/it-it/autonomous-machines/embedded-systems/jetson-orin/

NVIDIA :GPUs (V)

Model	Jetson Nano		Jetson TX2 series		Jetson Xavier NX		Jetson AGX Xavier	
	FPS (limited latency)	FPS (max throughput)	FPS (limited latency)	FPS (max throughput)	FPS (limited latency)	FPS (max throughput)	FPS (limited latency)	FPS (max throughput)
Inception V4 (299x299)	11*	13	24*	32	320	405	528	704
VGG-19 (224x224)	10*	12	23*	29	67*	313	276	432
Super Resolution (481x321)	15*	15	33*	33	164	166	281	302
Unet (256x256)	17*	17	39*	39	166	166	240	251
OpenPose (256x456)	15*	15	34*	35	238	271	439	484
Tiny YOLO V3 (416x416)	48*	49	107	112	607	618	1100	1127
ResNet-50 (224x224)	37*	47	84	112	824	1100	1946	2109
SSD Mobilenet-V1 (300x300)	43*	48	92	109	909	1058	1602	1919
SSSD Resnet34 (1200x1200)	1	1	3	2	29	29	55	55

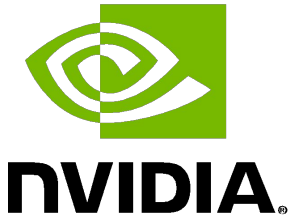
https://developer.nvidia.com/embedded/jetson-benchmarks

NVIDIA :GPUs (VI)



https://developer.nvidia.com/embedded/linux-tegra

GPU computing by NVIDIA DLI

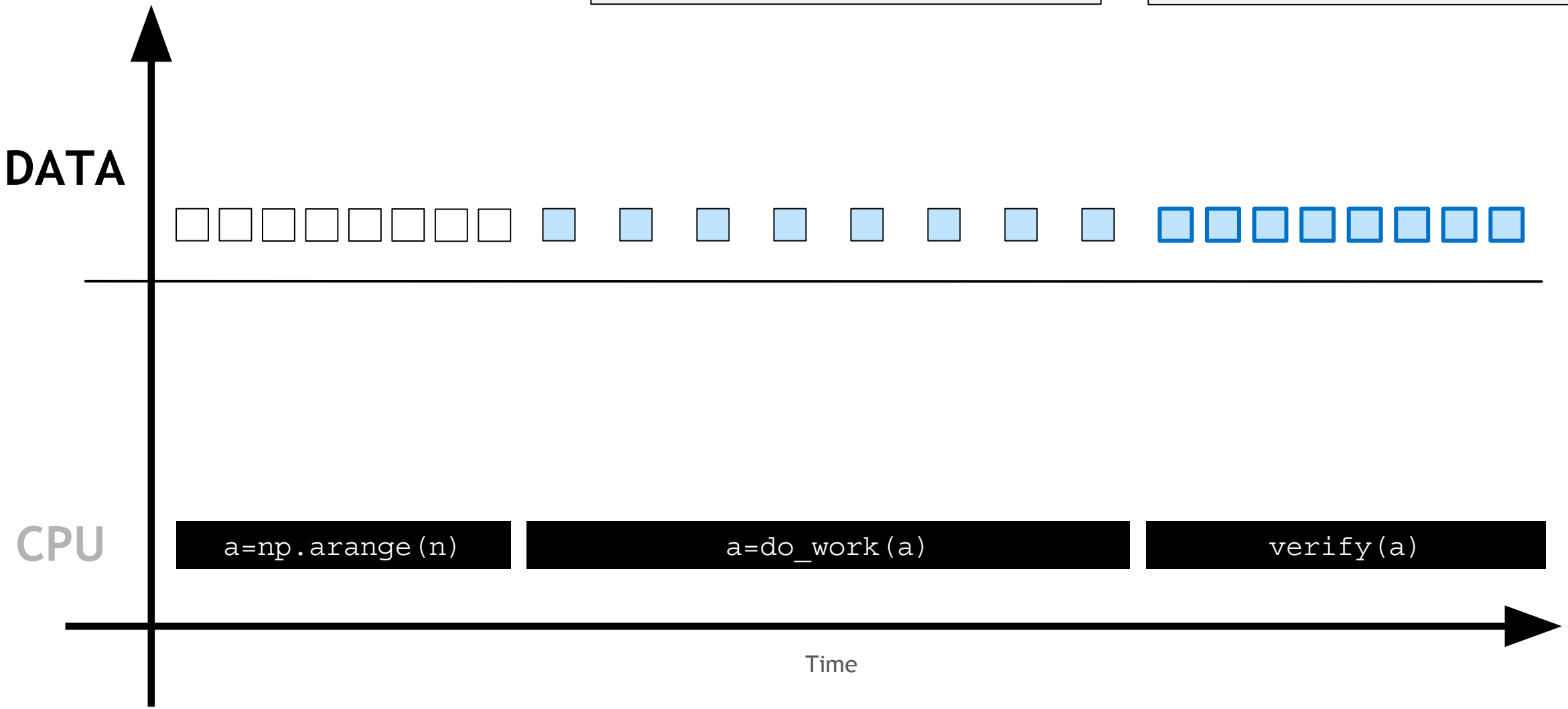


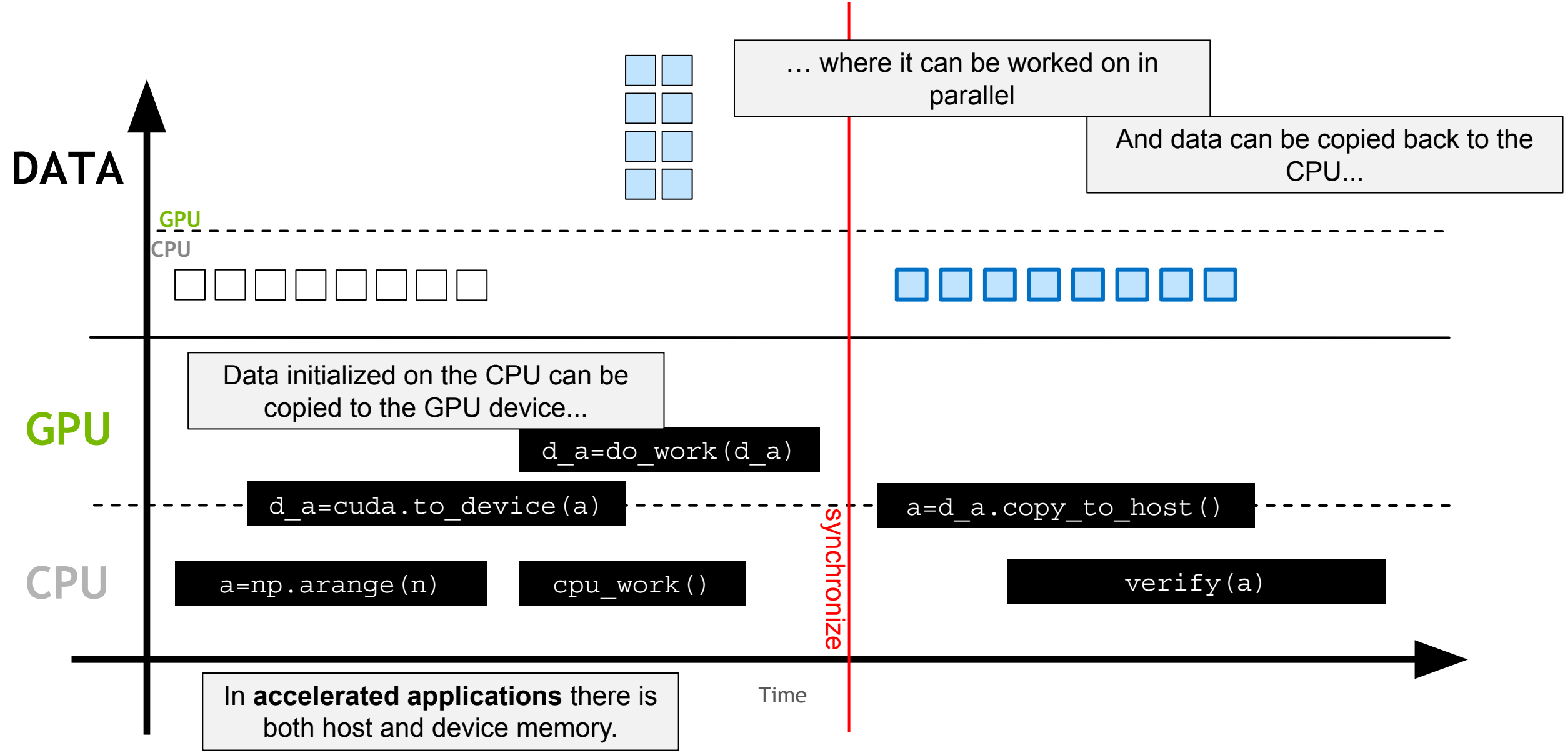
DEEP
LEARNING
INSTITUTE

www.nvidia.com/dli

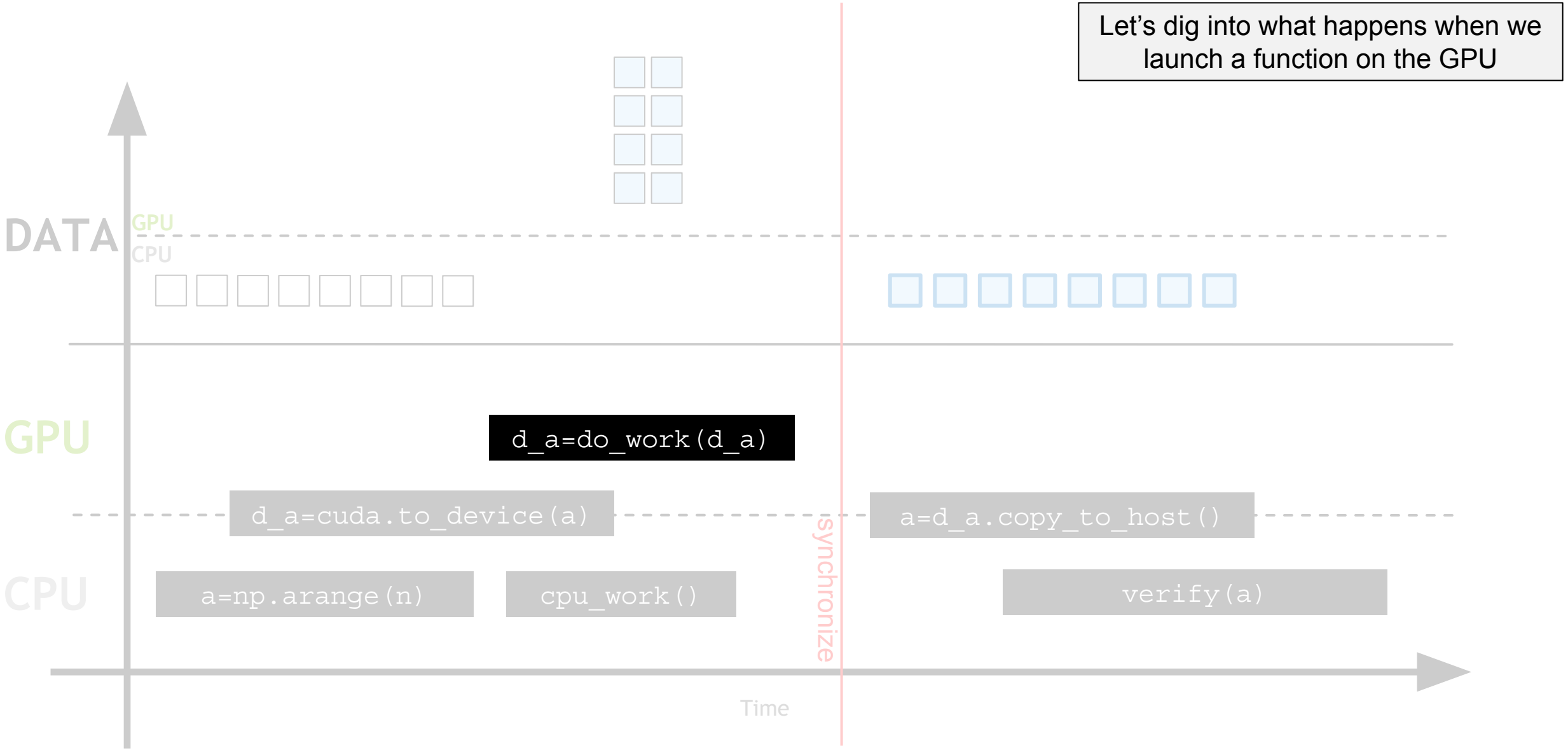
In **CPU-only applications** data is allocated on the CPU

…and all work is performed serially on the CPU



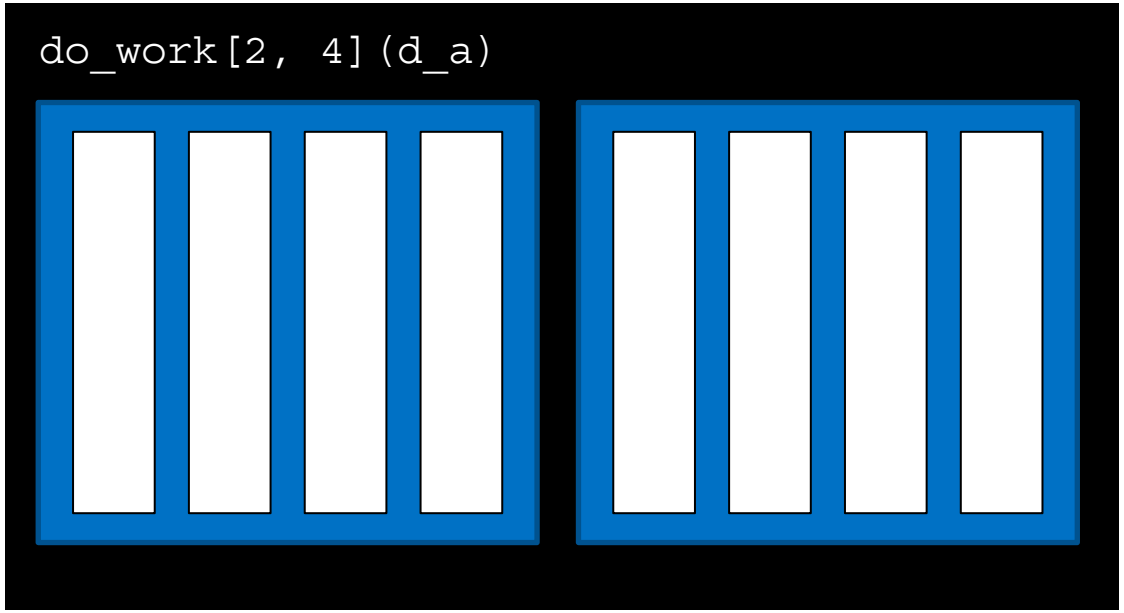


Let's dig into what happens when we launch a function on the GPU



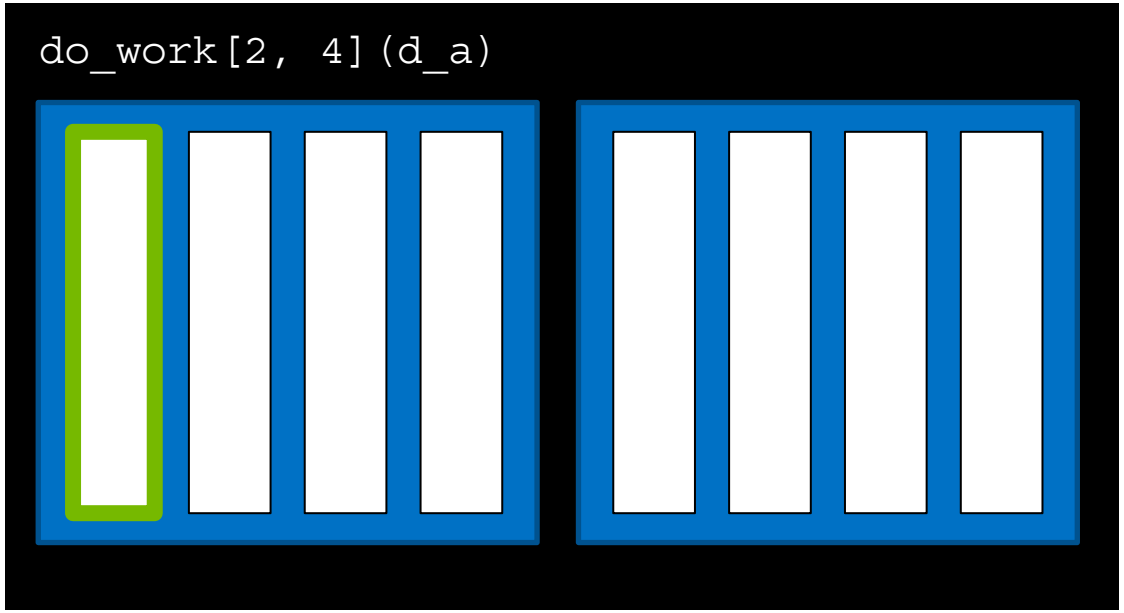
GPUs do work in parallel

GPU



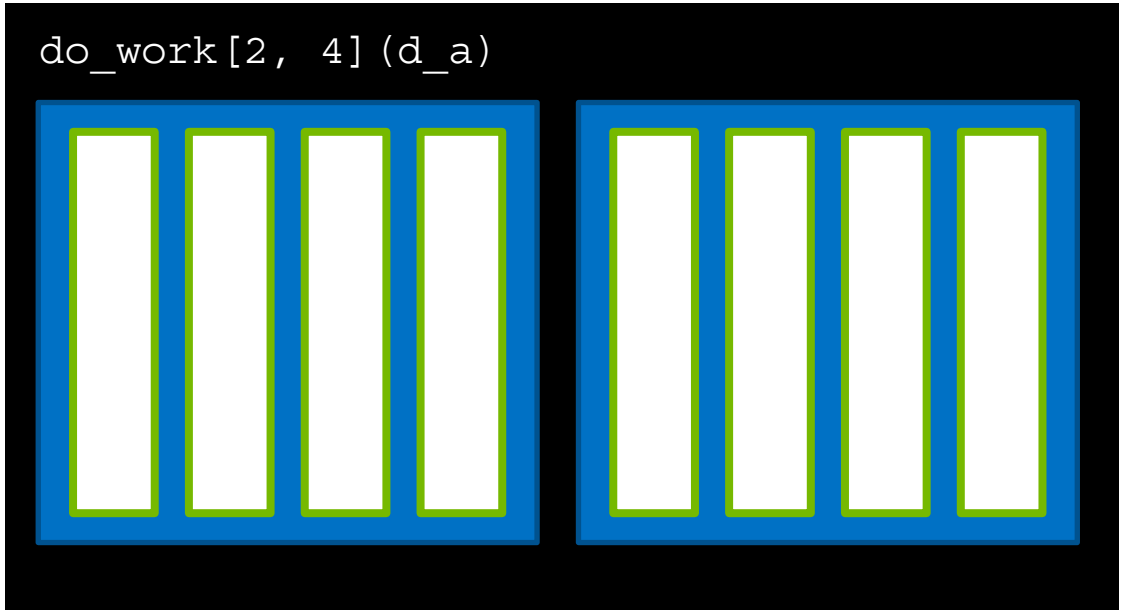
GPU work is done in a **thread**

GPU



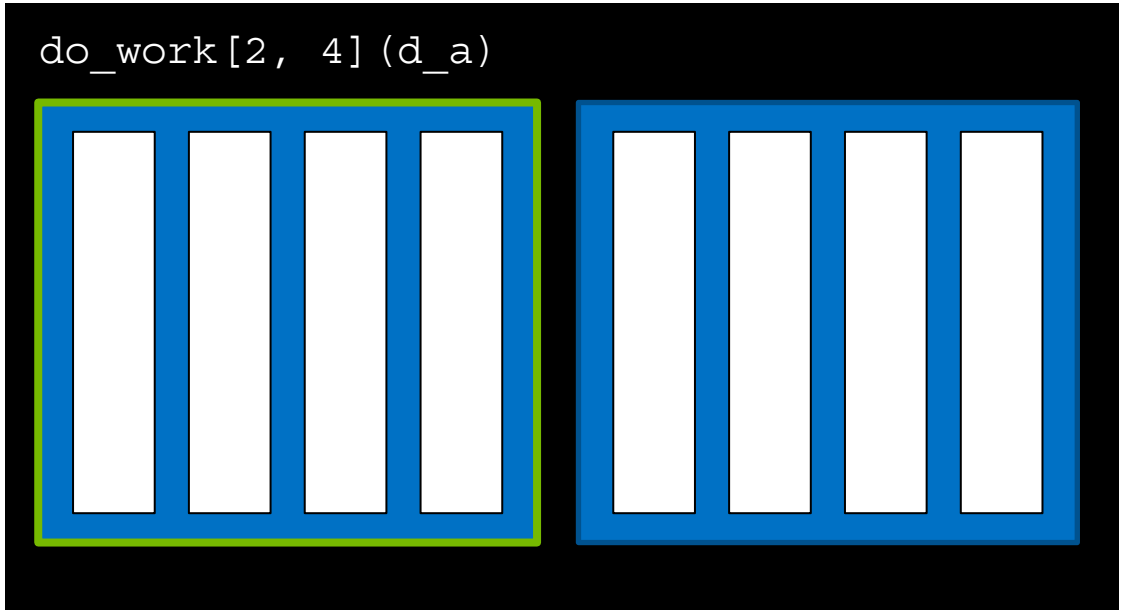
Many threads run in parallel

GPU



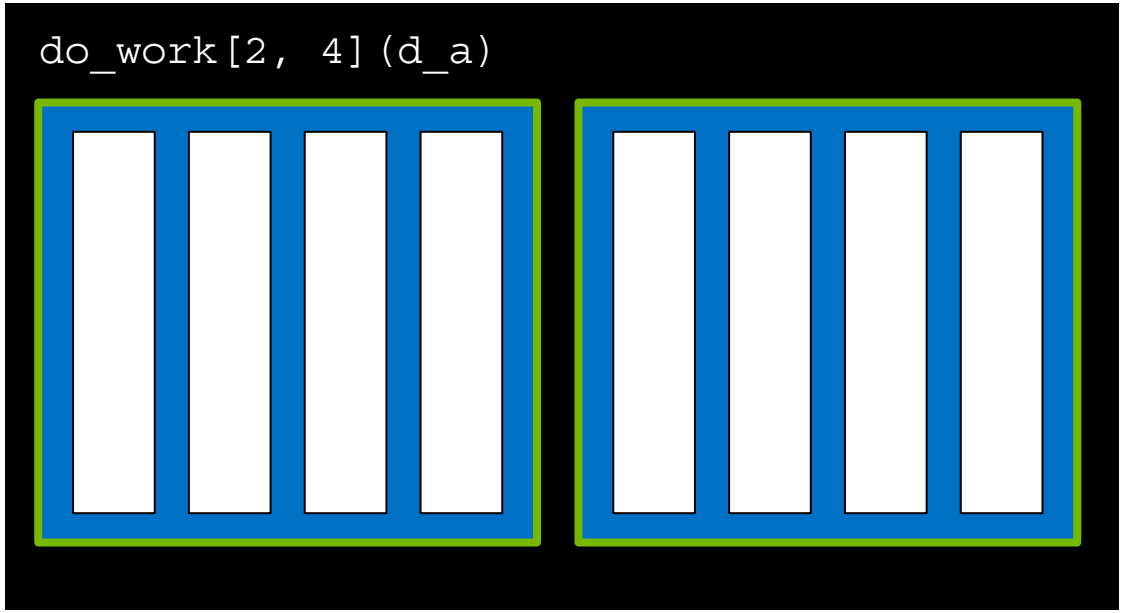
A collection of threads is a **block**

GPU



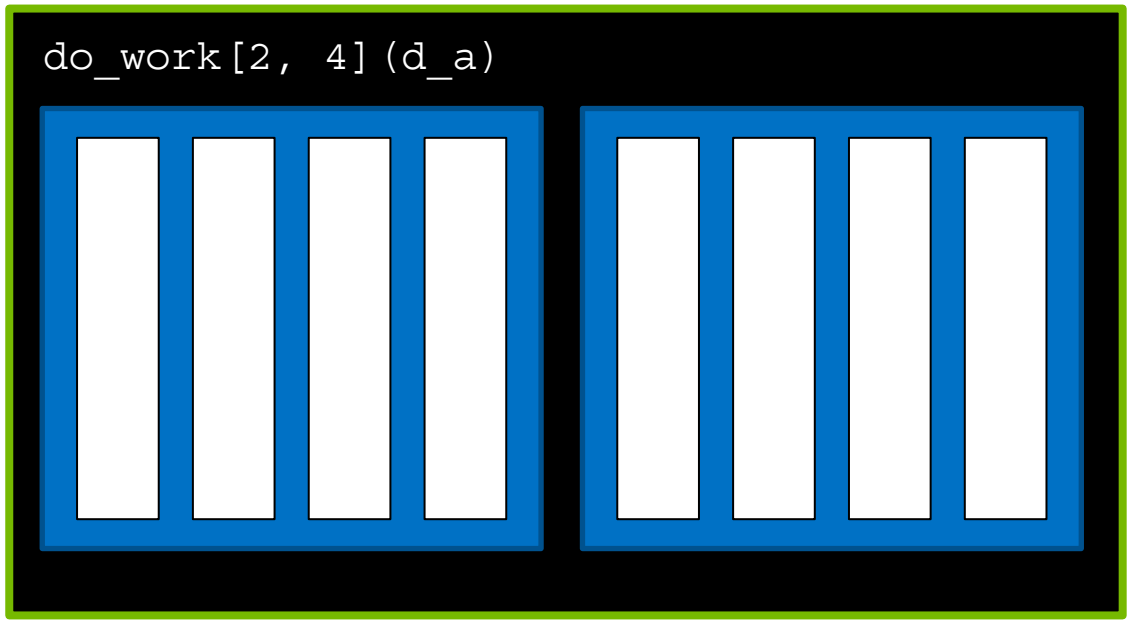
There can be many blocks

GPU



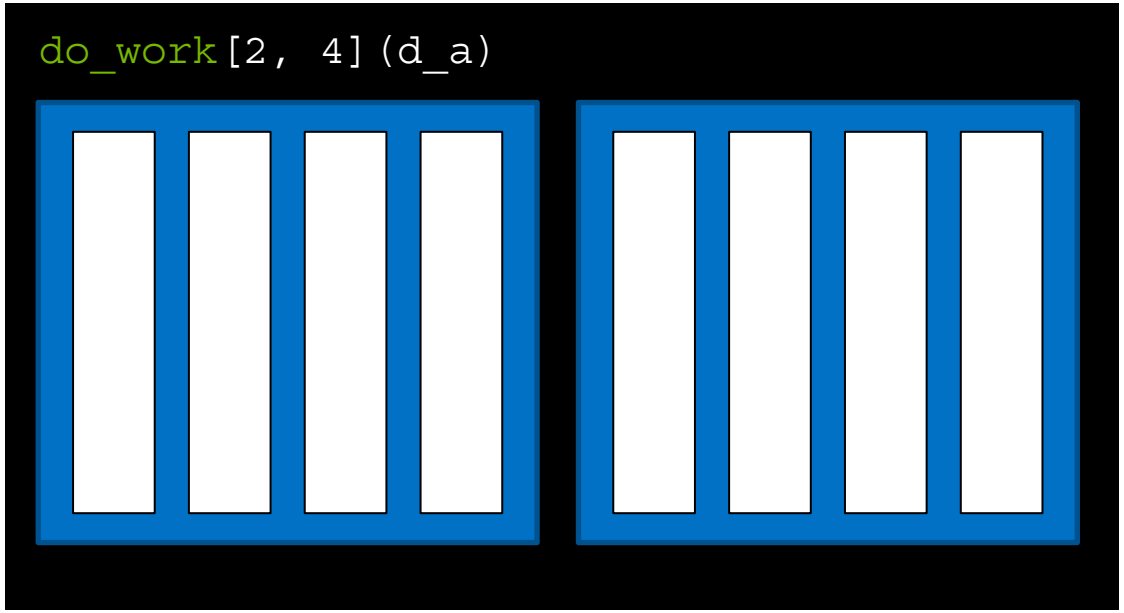
A collection of blocks associated with a given kernel launch is a **grid**

GPU



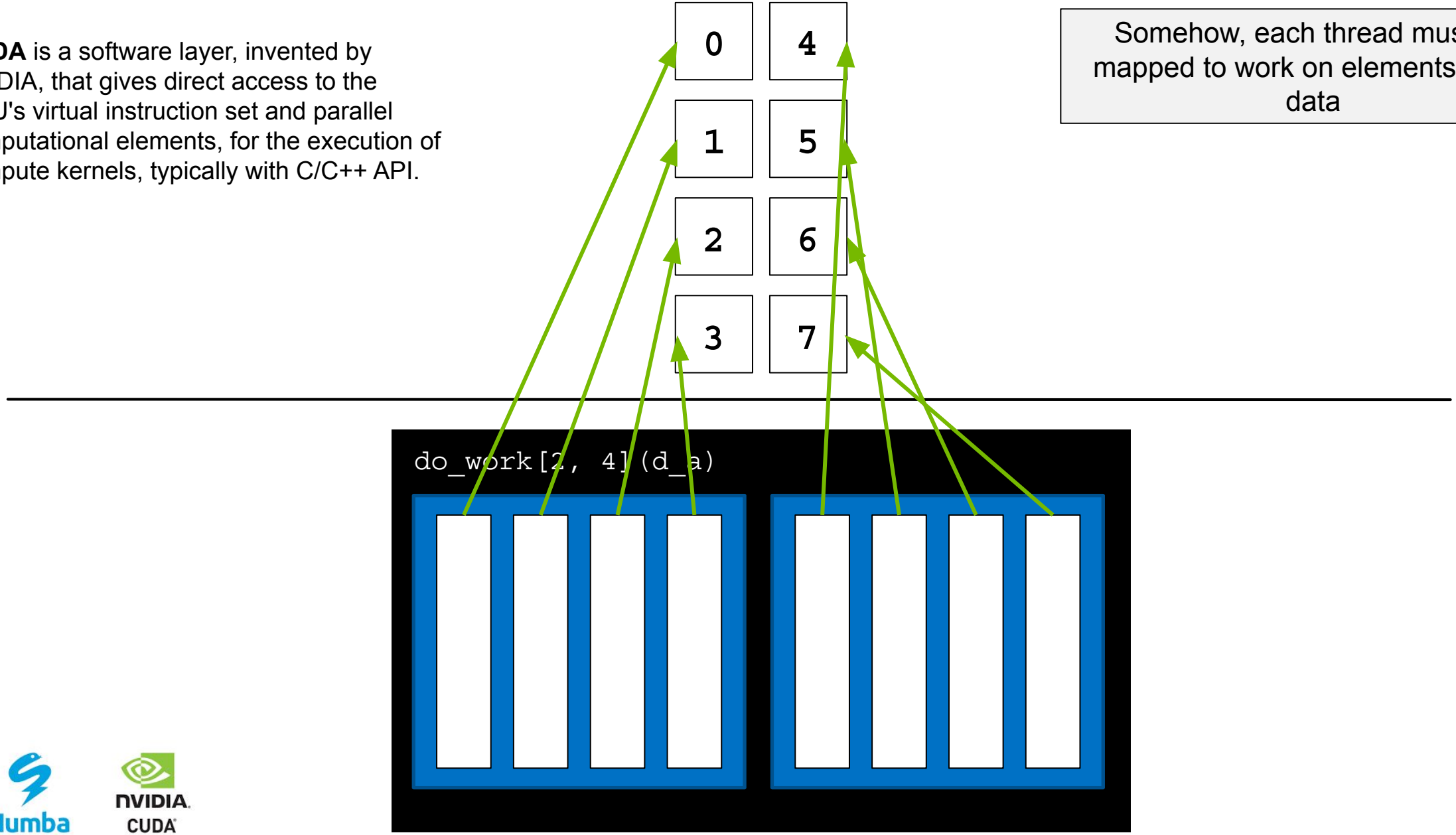
GPU functions are called **kernels**

GPU



CUDA is a software layer, invented by NVIDIA, that gives direct access to the GPU's virtual instruction set and parallel computational elements, for the execution of compute kernels, typically with C/C++ API.

Somehow, each thread must be mapped to work on elements in the data

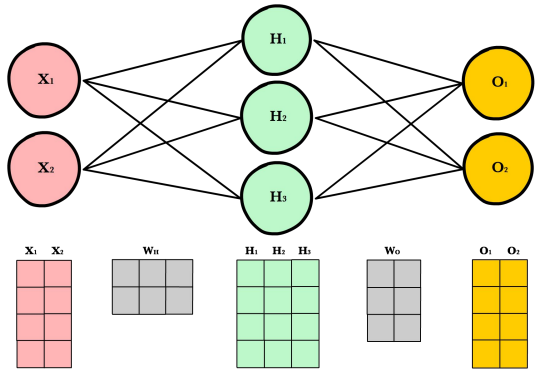


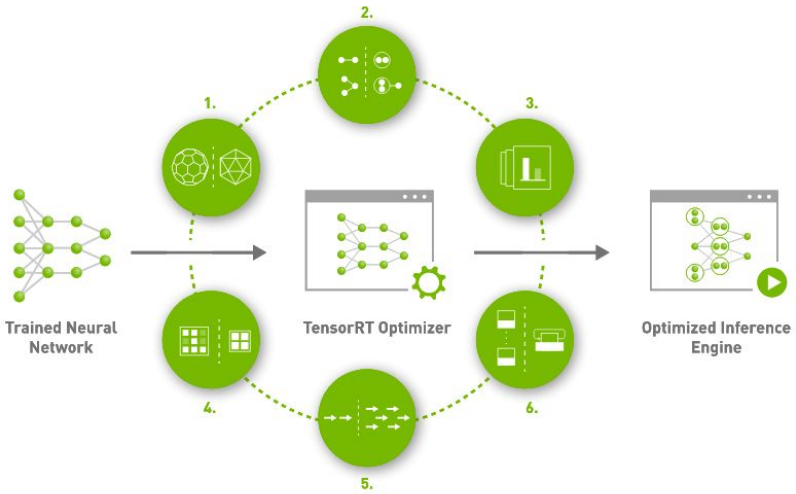
Numba

NVIDIA CUDA

https://developer.nvidia.com/how-to-cuda-python

TensorRT : Accelerating DNN on GPU





1. Reduced Precision

Maximizes throughput with FP16 or INT8 by quantizing models while preserving accuracy

2. Layer and Tensor Fusion

Optimizes use of GPU memory and bandwidth by fusing nodes in a kernel

3. Kernel Auto-Tuning

Selects best data layers and algorithms based on the target GPU platform

4. Dynamic Tensor Memory

Minimizes memory footprint and reuses memory for tensors efficiently

5. Multi-Stream Execution

Uses a scalable design to process multiple input streams in parallel

6. Time Fusion

Optimizes recurrent neural networks over time steps with dynamically generated kernels

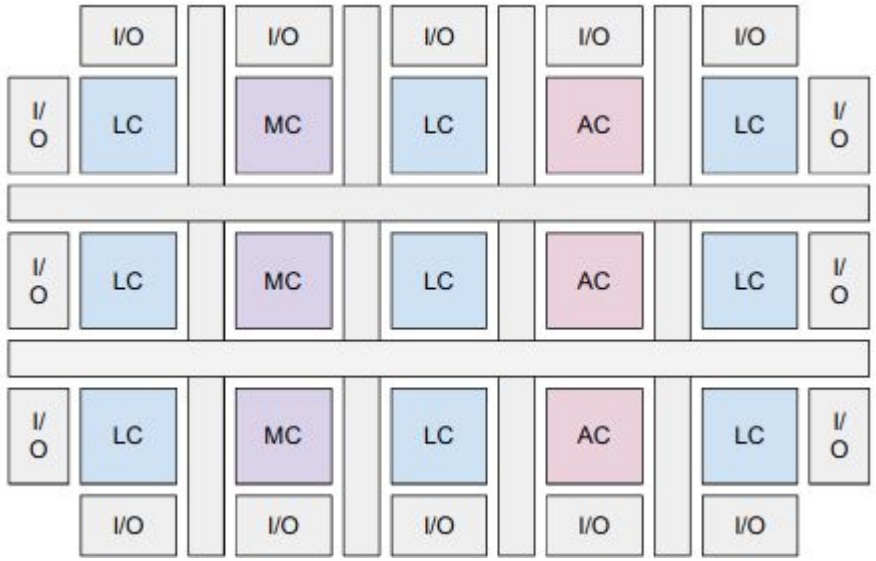
https://developer.nvidia.com/tensorrt

GPUs vs FPGAs (I)

Field-programmable gate arrays (FPGA) are integrated circuits designed to be configured after manufacturing, via HDL, for implementing arbitrary logic functions in hardware.

Besides Logic Cells (LC), FPGA also includes special-purpose cells:

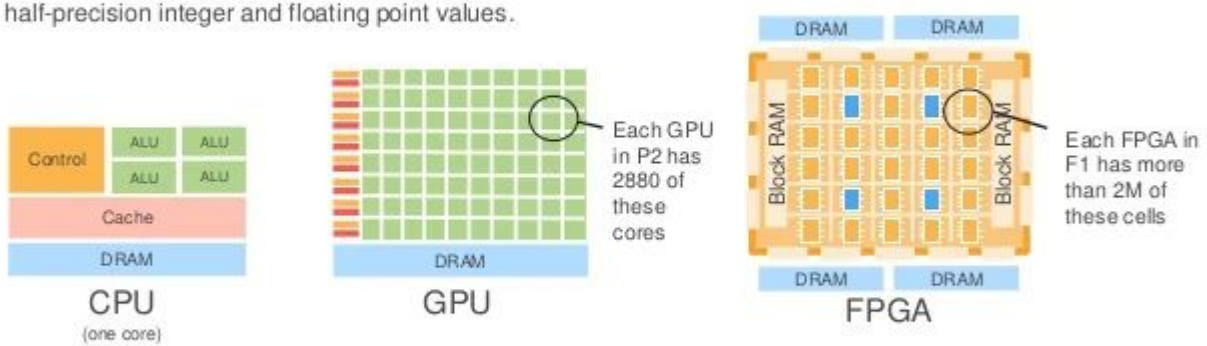
- Digital signal processing (DSPs) / arithmetic cells (AC);
- Dense memory cells (MC) (e.g., BRAMs).
- Input / output (I/O) cells.



GPUs vs FPGAs (II)

Parallel Processing in GPUs and FPGAs

A GPU is effective at processing the same set of operations in parallel – single instruction, multiple data (SIMD). A GPU has a well-defined instruction-set, and fixed word sizes – for example single, double, or half-precision integer and floating point values.



An FPGA is effective at processing the same or different operations in parallel – multiple instructions, multiple data (MIMD). An FPGA does not have a predefined instruction-set, or a fixed data width.

amazon web services | Webinars

Great performance with high throughput and low latency: FPGAs can inherently provide low latency as well as deterministic latency for real-time applications like video streaming, transcription, and action recognition by directly ingesting video into the FPGA, bypassing a CPU. Designers can build a neural network from the ground up and structure the FPGA to best suit the model.

Excellent value and cost: FPGAs can be reprogrammed for different functionalities and data types, making them one of the most cost-effective hardware options available. Furthermore, FPGAs can be used for more than just AI. By integrating additional capabilities onto the same chip, designers can save on cost and board space. FPGAs have long product life cycles, so hardware designs based on FPGAs can have a long product life, measured in years or decades. This characteristic makes them ideal for use in industrial defense, medical, and automotive markets.

Low power consumption: With FPGAs, designers can fine-tune the hardware to the application, helping meet power efficiency requirements. FPGAs can also accommodate multiple functions, delivering more energy efficiency from the chip. It's possible to use a portion of an FPGA for a function, rather than the entire chip, allowing the FPGA to host multiple functions in parallel.

https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/fpga-gpu.html

GPUs vs FPGAs (III)

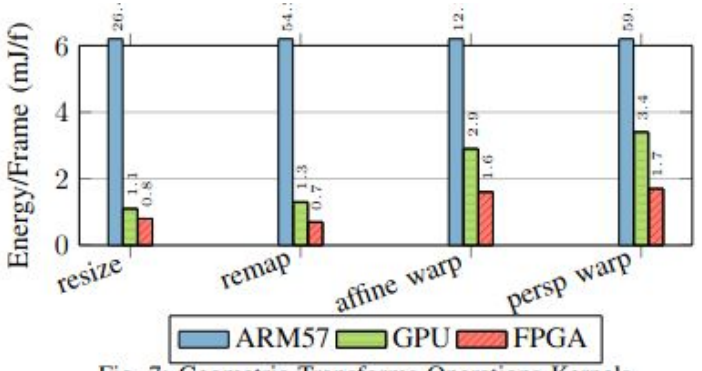


Fig. 7: Geometric Transforms Operations Kernels

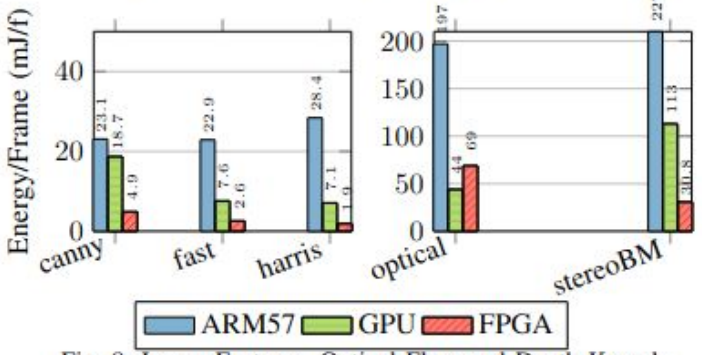


Fig. 8: Image Features, Optical Flow and Depth Kernels

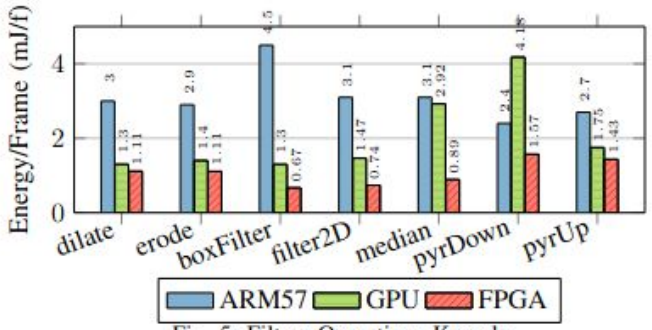


Fig. 5: Filters Operations Kernels

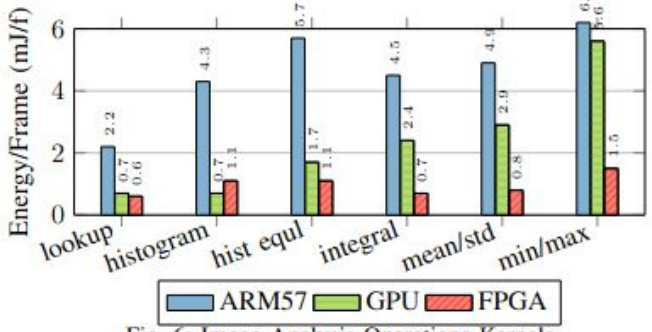


Fig. 6: Image Analysis Operations Kernels

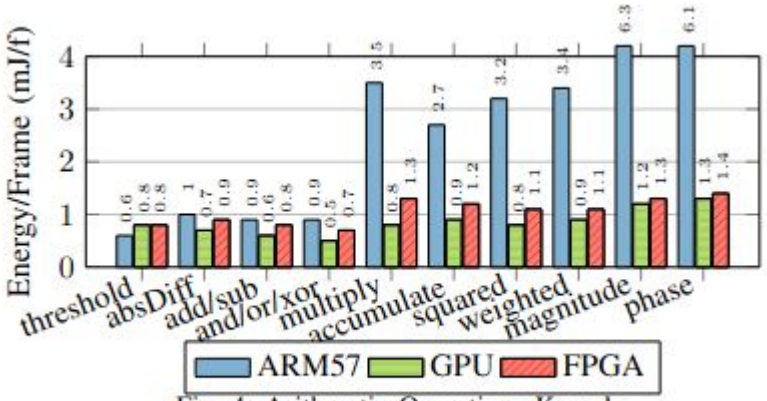


Fig. 4: Arithmetic Operations Kernels

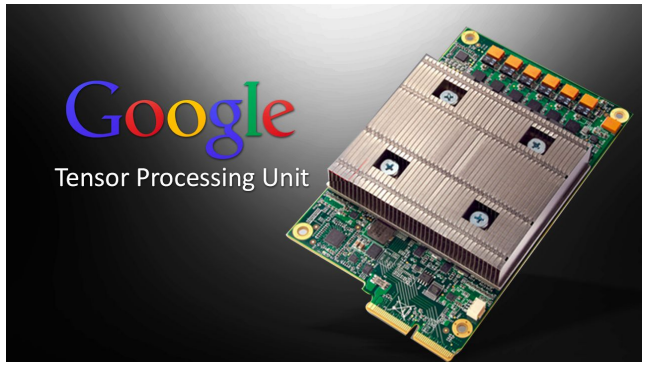
https://arxiv.org/pdf/1906.11879.pdf

NPU

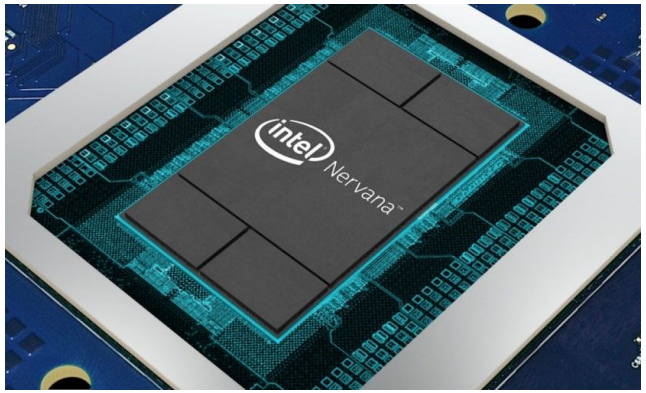
- A neural processing unit (NPU) is a microprocessor that specializes in the acceleration of machine learning algorithms
- Is studied and developed for a specific purpose (unlike general CPUs)

https://en.wikipedia.org/wiki/AI_accelerator

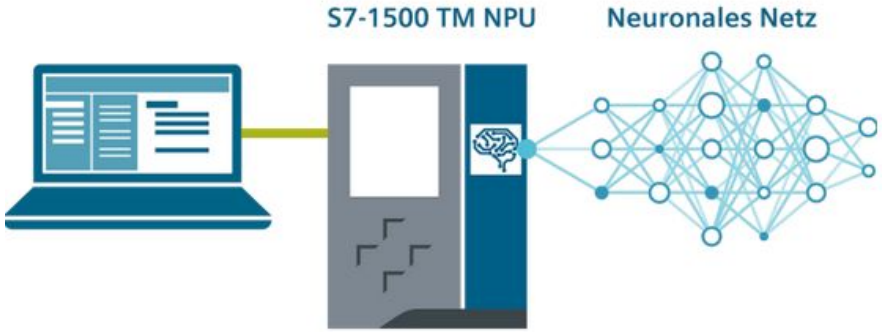
https://www.intel.com/content/www/us/en/products/details/processors/movidius-vpu.html



https://cloud.google.com/tpu



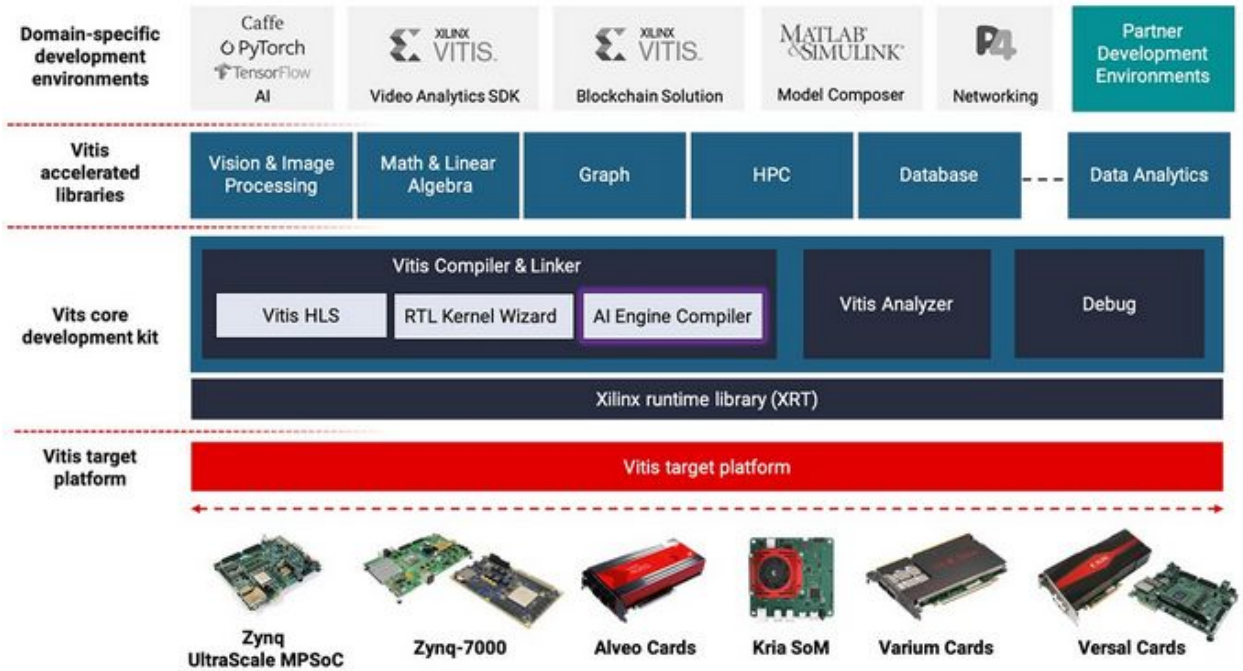
https://siliconangle.com/2020/02/02/intel-dumps-nervana-neural-network-processors-habanas-ai-chips/



https://new.siemens.com/fr/fr/produits/automatisation-entrainements/systemes-automatisation/industrial/io-systems/artificial-intelligence.html

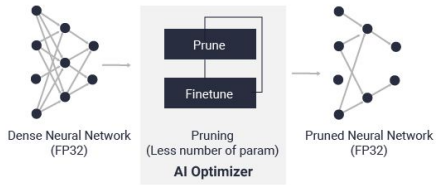
https://new.siemens.com/global/en/products/automation/systems/industrial/plc/simatic-s7-1500/simatic-s7-1500-tm-npu.html

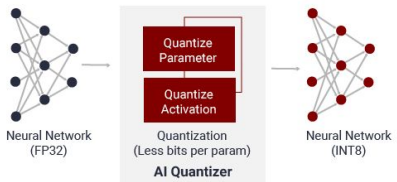
Xilinx :FPGAs



AI Optimizer

With world-leading model compression technology, we can reduce model complexity by 5x to 50x with minimal accuracy impact. Deep Compression takes the performance of your AI inference to the next level.



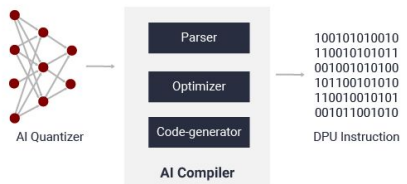


AI Quantizer

By converting the 32-bit floating-point weights and activations to fixed-point like INT8, the AI Quantizer can reduce the computing complexity without losing prediction accuracy. The fixed-point network model requires less memory bandwidth, thus providing faster speed and higher power efficiency than the floating-point model.

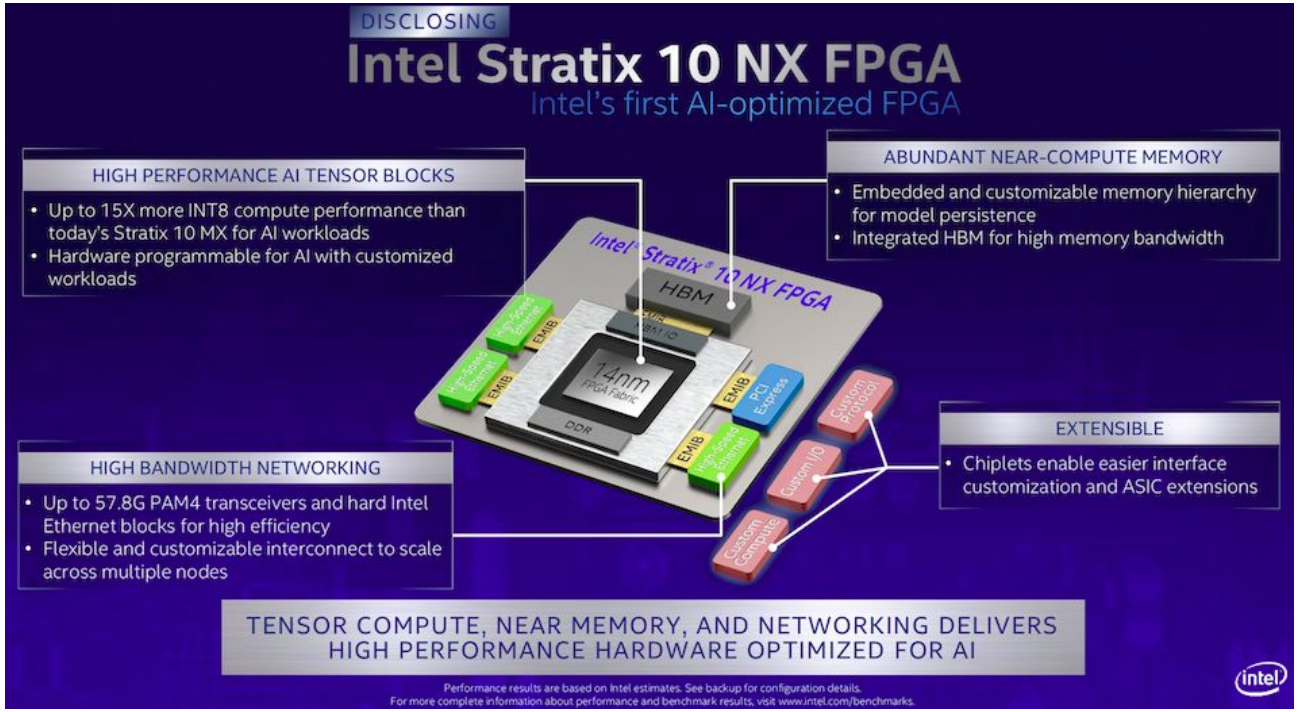
AI Compiler

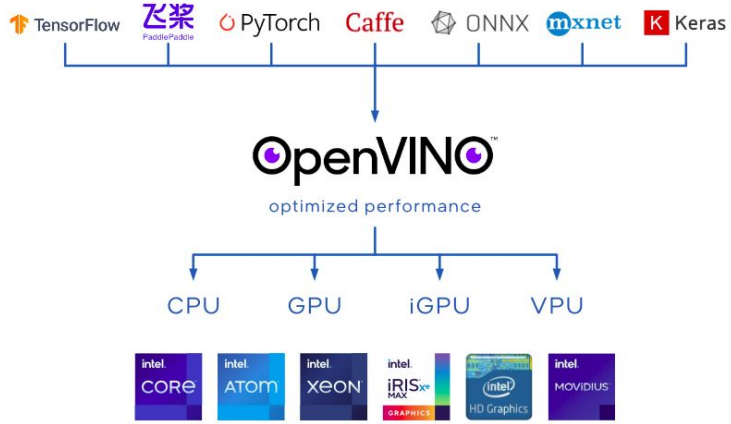
Maps the AI model to a high-efficient instruction set and data flow. Also performs sophisticated optimizations such as layer fusion, instruction scheduling, and reuses on-chip memory as much as possible.



https://www.xilinx.com/products/design-tools/vitis/vitis-ai.html
https://www.xilinx.com/products/design-tools.html

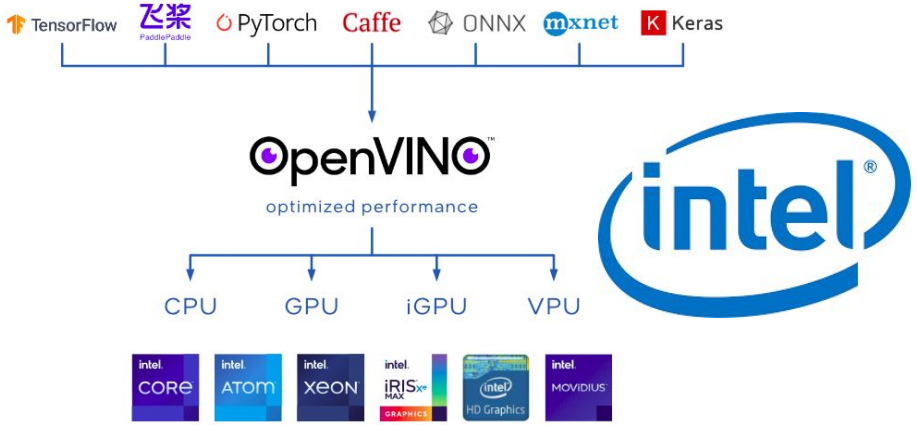
Intel :FPGAs

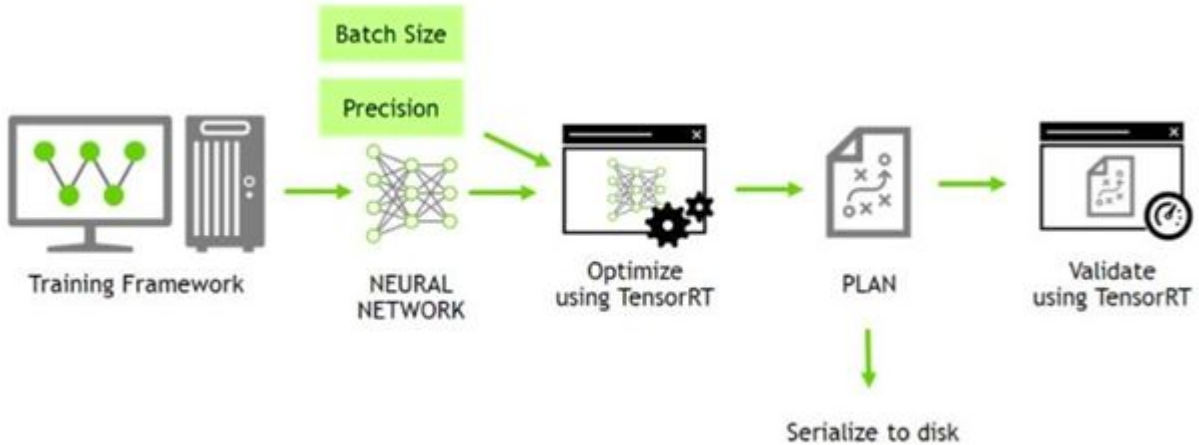


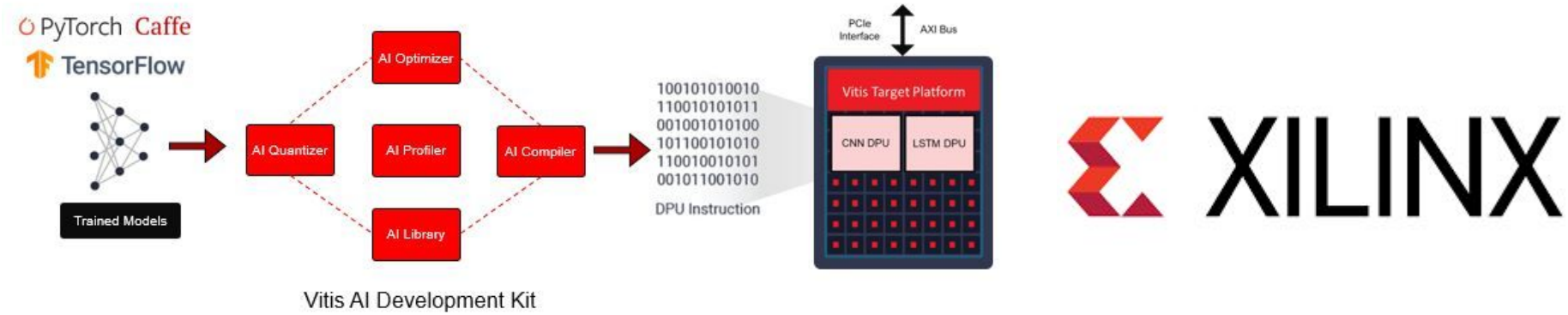


https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/solutions.html

Recap







https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/solutions.html