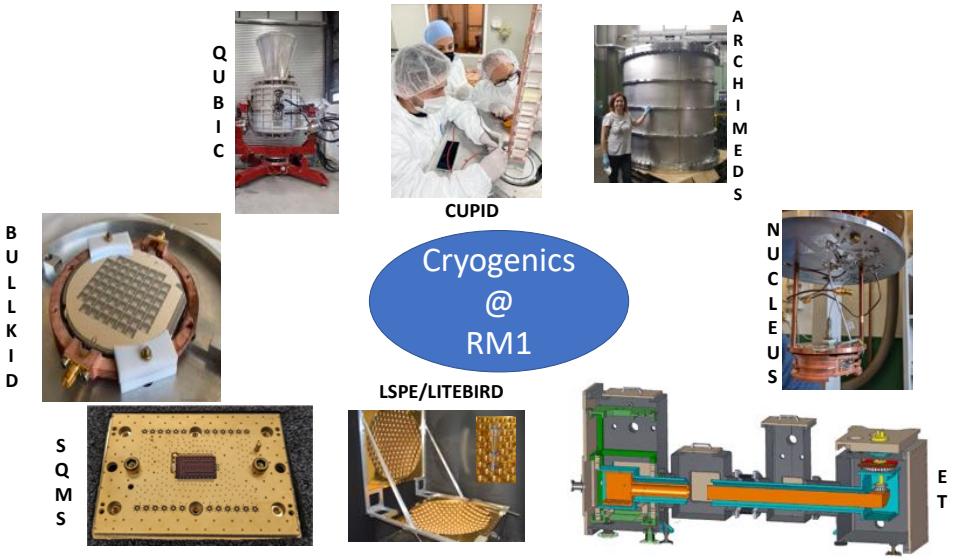
# Tecnologie del calcolo e attività INFN

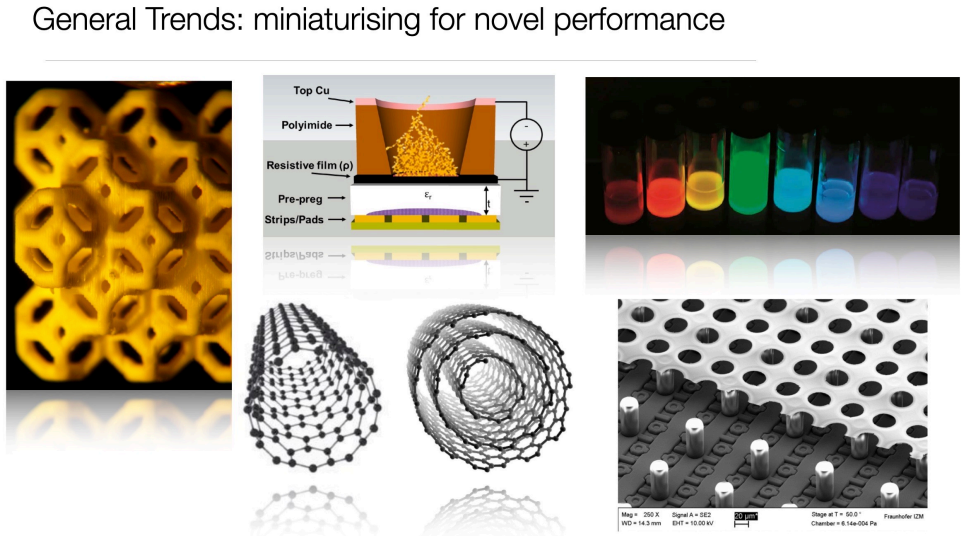
Piero Vicini

Retreat Sezione di Roma - Assisi 2022

# Disclaimer

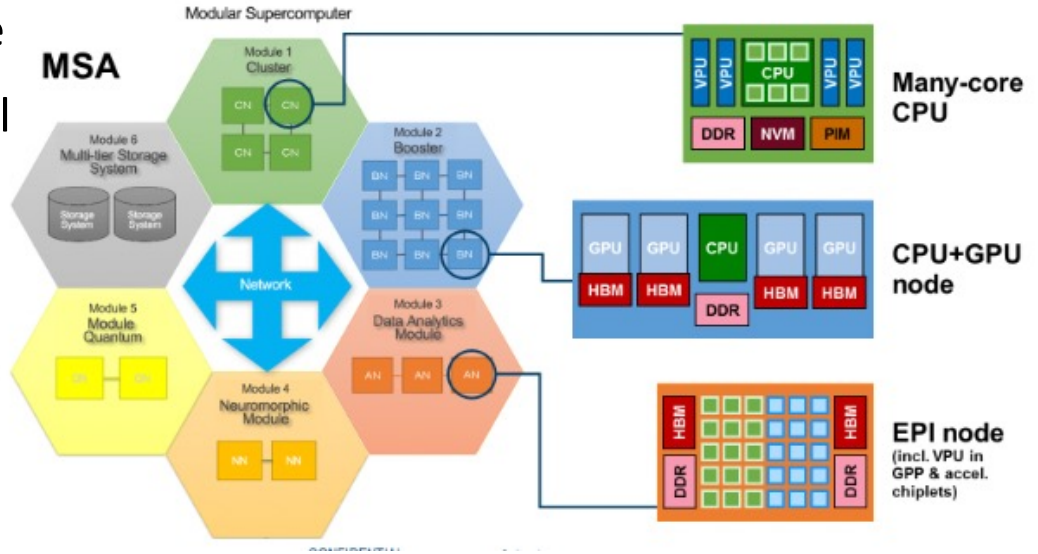
- Un talk sulla “Tecnologia” (brivido lungo la schiena....)
- Per fortuna molti topics correlati (almeno dal punto di vista applicativo) in altri talks
- In questa sessione due contributi alle nostre tecnologie di basso livello (Criogenia e Detector) di interesse per CSN5 e in generale per INFN
  - Criogenia (A. Cruciani)
  - Rivelatori innovativi (D. Pinci)
- Mi mettero’ nella mia comfort zone e ovviamente non copriro’ tutto lo spettro delle tecnologie del calcolo (promettenti e meno) e dell’infrastruttura correlata. Evitero’ di fare un elenco sterile di tutte (e sono tante) le attivita’ di ricerca tecnologica della sezione.
- Provero’ a darvi un mio punto di vista (parziale) su alcune evoluzioni di interesse a breve e medio termine utili per le nostre ricerche



# Specializzazione, eterogeneità e convergenza

- Sistema eterogeneo modulare (Low level Computer continuum...)
  - Aggregazione di moduli diversi ognuno specializzato per task computazionali differenti
  - Puo' valere ad ogni scala di sistema
    - In particolare MSA Modular Supercomputer Architeture
  - Ingredienti base CPU, acceleratori computazionali (GPU, DPU, FPGA), network, componenti programmabili per implemetazione di acceleratori per specific task computazionali (FPGA/GPU/ASIC per ML o data analytics), programming models, OS integrato, real time schedulers, storage
  - Target Data Centers i.e. cloud per HPC, Data analytics, IA ed in futuro QC?
- Open issues:
  - Network: esiste one architecture fits for all???
  - Interfaccia ai singoli moduli computazionali
    - Eterogenea per definizione ma omogenea per garantire l'integrazione
  - I diversi moduli hanno differente maturita' tecnologica, differente complessita', differente peculiarita' dei dati, differente caratteristica delle interfacce (tipologia e timing) etc...
  - Orchestrazione e modello di programmazione

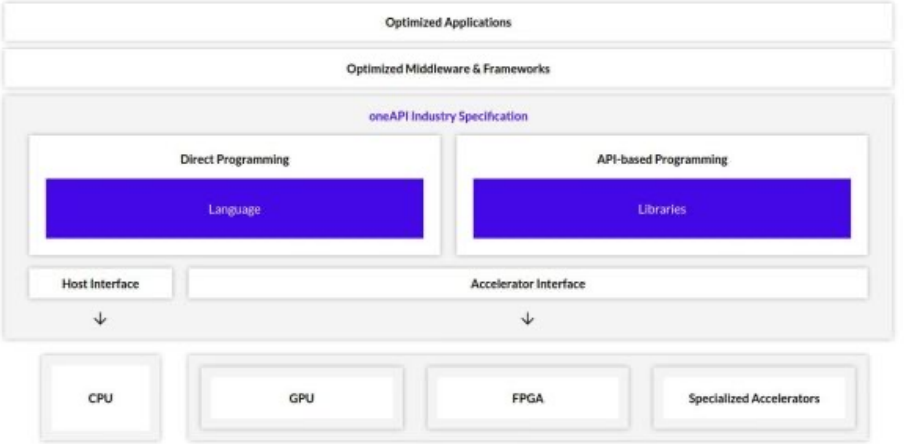


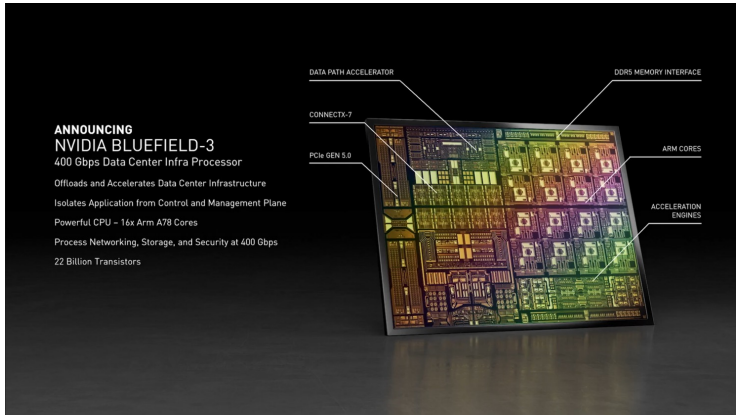

*Se l'obiettivo e' il sistema integrato ad alta efficenza computazionale e low power siamo ancora lontani e serve ancora ricerca...*

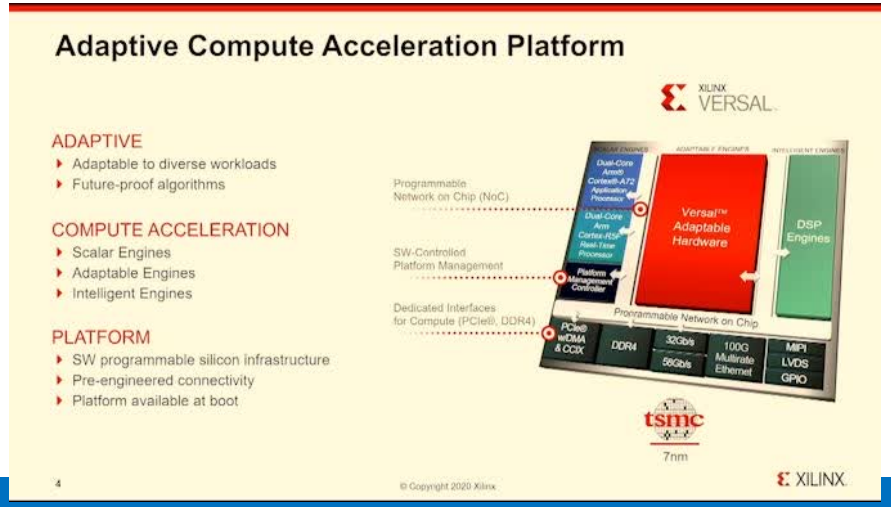
# Panorama commerciale

Nel frattempo i (big) players commerciali si muovono:

- 2016: INTEL(CPU) acquisisce ALTERA(FPGA) per 17 B$!!!
  - realizzare “convergenza CPU+FPGA” i.e. integrazione di FPGA (componenti programmabili) e CPU a “die” level
  - stato ongoing dal punto di vista hardware
  - release di un framework integrato di programmazione (almeno per FPGA SoC+CPU per GPU annunciato) INTEL oneAPI basato su linguaggio DPC++ (Data Parallel C++)
  - e’ open → supporto per GPU in progress
  - Release di Ponte Vecchio XE-based GPU orientata a HPC/AI https://www.nextplatform.com/2021/08/24/intels-ponte-vecchio-gpu-better-not-be-a-bridge-too-far/
- 2020: NVIDIA (GPU) acquisisce Mellanox (Infiniband Network per HPC) per 7 B$
  - integrare GPU e Network per realizzare mesh scalabili di GPU a bassa latenza
  - stato ongoing ma per adesso business units separate
  - interessanti SoC eterogenei (ex BlueField) per accelerazione on the edge
  - *issues: non piu’ competitors indipendenti per reti interconnessione HPC*
- 2022: AMD (CPU) acquisisce Xilinx (FPGA) per 35 B$ con
  - obiettivo simile a INTEL/Altera
  - mettere a sistema con CPU e GPU, le varie architetture di SoC programmabili e specializzate per HPC, ML
  - *non piu’ competitori indipendenti providers di FPGA*



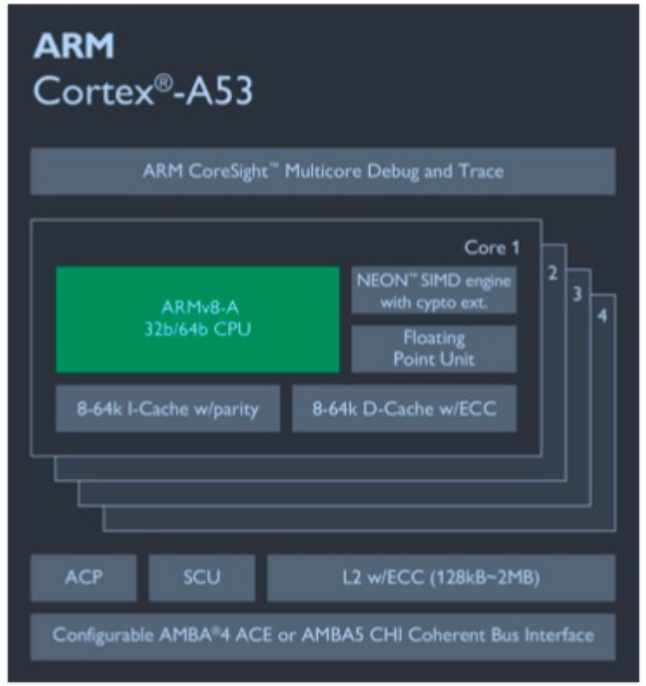



# Ingredienti: CPU low power (ARM)

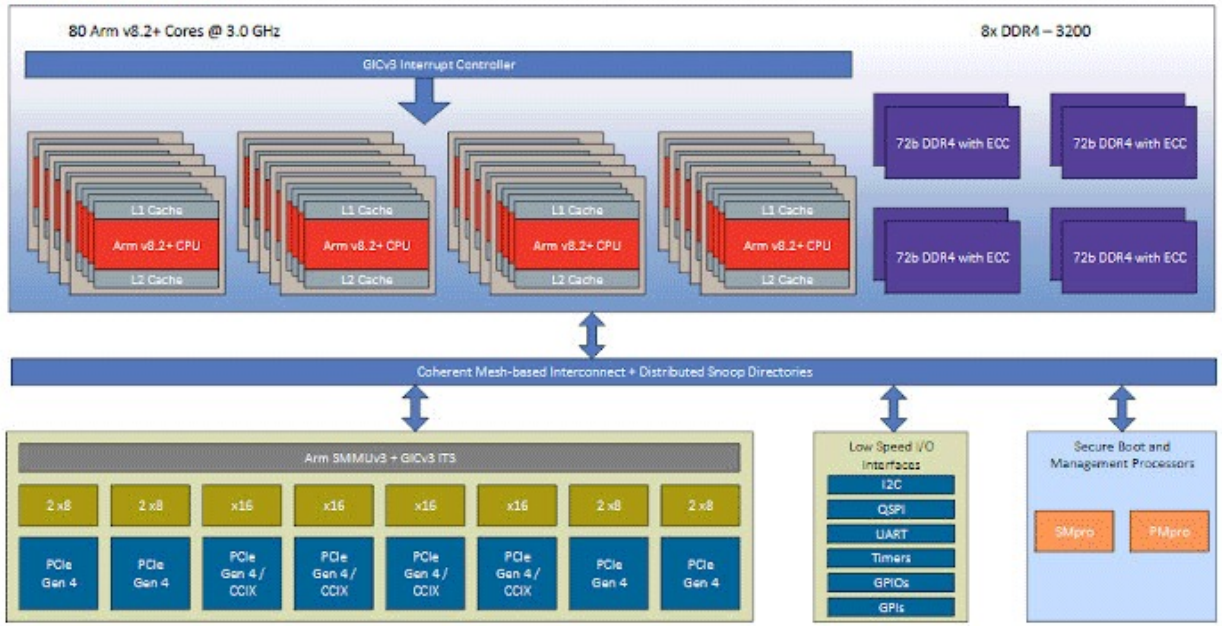
Beyond the x86 mainstream: Architetture ARM

- e' stato per lungo tempo produttore Europeo di CPU specializzate per processori embedded (leader di mercato)
- business model: vendono license non prodotti
- architetture a 32b/64b utilizzate per server e microserver, multi-core integrati in FPGA
- low power → alto numero di cores
- costi contenuti
- cavium, amcc alcuni esperimenti di integrazione
- ARM-64 based ma completamente customized
  - APPLE M1-M1X- ULTRA CPU fino a 20 core (fino a 16 ad alte prestazioni, 4 ad alta efficienza), una GPU fino a 64 core, un Neural Engine (NPU) con 32 core e una memoria interna fino a 128GB.
  - Ampere (startup US per server ARM-based) multicore → 64-80
  - progetti EU: Mont Blanc, EuroServer, ExaNeSt...




| *Integer Performance* | Ampere Altra | AMD Epyc 7742 | AMD Epyc 7702 | Intel 8280 SP | Intel Xeon SP 8276 | Intel Xeon SP 6238R |
|---|---|---|---|---|---|---|
| Cores | 80 | 64 | 64 | 28 | 28 | 28 |
| Clock Speed | 3.3 GHz | 2.25 GHz | 2.0 GHz | 2.7 GHz | 2.2 GHz | 2.2 GHz |
| SPECrate 2017 Integer Base | - | 667 | 593 | 342 | 296 | 287 |
| SPECrate 2017 Integer Base (GCC) | 579 | *557* | *495* | *260* | *225* | *218* |
| Performance/CPU | 290 | 278 | 248 | 130 | 112 | 109 |
| Performance/Core | 3.62 | 4.35 | 3.87 | 4.64 | 4.02 | 3.90 |
| Watts | 205 | 225 | 200 | 205 | 165 | 165 |
| Performance/Watt | 1.41 | 1.24 | 1.24 | 0.63 | 0.68 | 0.66 |
| Watts/Core | 2.56 | 3.52 | 3.13 | 7.32 | 5.89 | 5.89 |
| CPU Price | *$5,800* | $6,950 | $6,450 | $10,009 | $8,719 | $2,612 |
| Price/Performance | $20.03 | $24.96 | $26.05 | $77.02 | $77.52 | $23.95 |

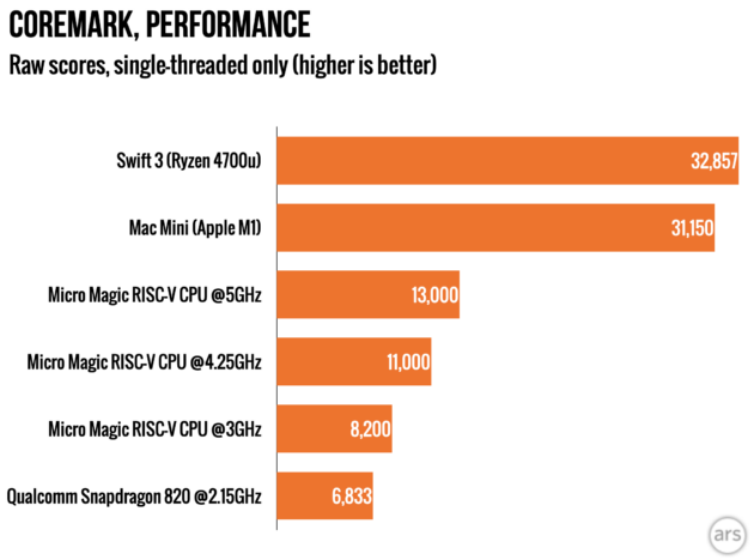
*Altra Block Diagram*

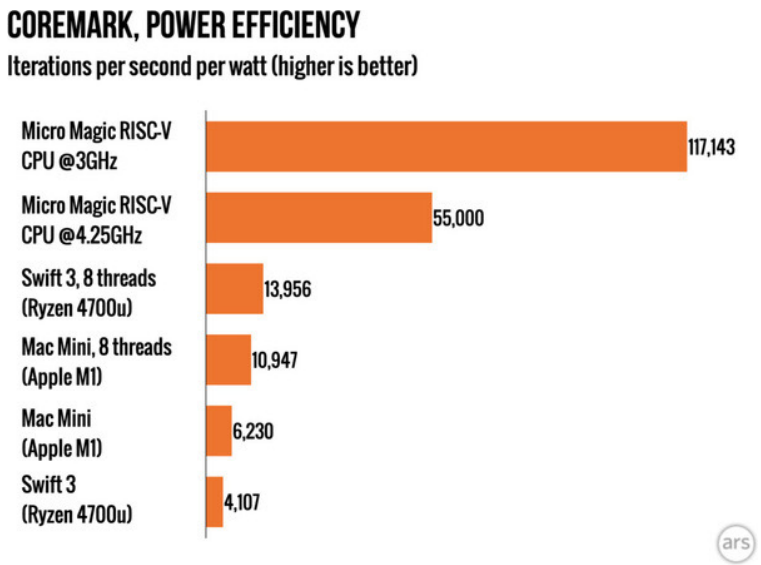


# Ingredienti: CPU low power (RISC-V)

Beyond the x86 mainstream: RISC-V

- NON e' un'architettura di CPU MA un insieme di istruzioni "open source"
  - ridurre la complessita' HW del core e il suo power consumption e aumentarne l'efficenza computazionale
- Una lunga storia (Berkeley 1981→)
- Oggi 5th generazione supportato da una fondazione RISC-V international (*https://riscv.org/)* a cui aderiscono i principali produttori di CPU
  - 2K+ PARTNERS, tra cui IBM, Intel, Google, Samsung, Nvidia...
- Supporto per CPU, many-cores acceleratori, ML, uControllori, HPC,...



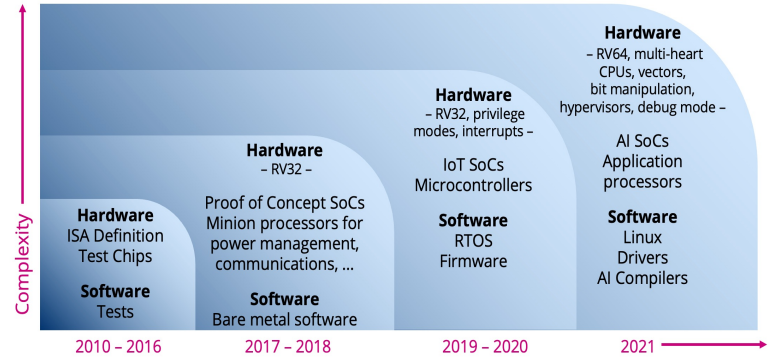


## Disruptive **Technology**

| Barriers | Legacy ISA | RISC-V ISA |
|---|---|---|
| Complexity | 1500+ base instructions Incremental ISA | 47 base instructions Modular ISA |
| Design freedom | $$$ – Limited | Free – Unlimited |
| License and Royalty fees | $$$ | Free |
| Design ecosystem | Moderate | Growing rapidly. Numerous extensions, open & proprietary cores |
| Software ecosystem | Extensive | Growing rapidly |

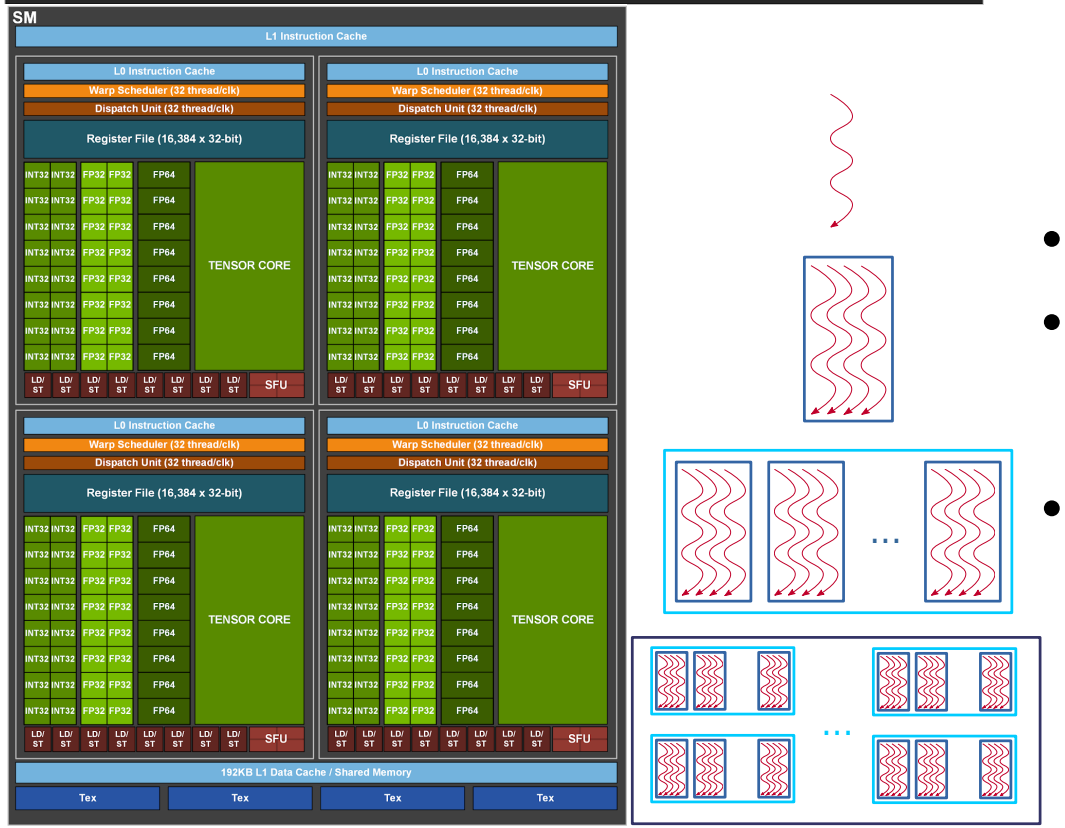
RISC-V

## Industry innovation on RISC-V




RISC-V

# Ingredienti: GPU

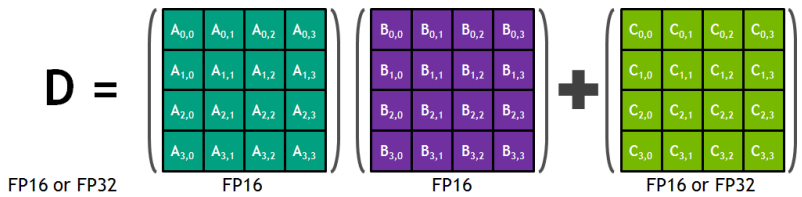



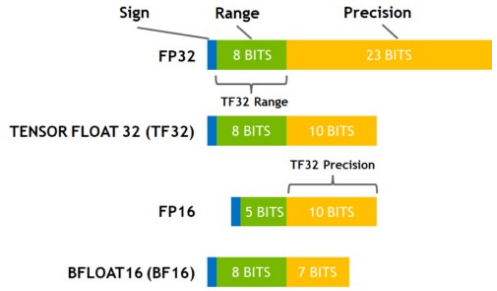



- In origine processori specializzati per la grafica
- GPU sono altamente multithreaded e fanno uso intensivo del parallelismo per ottenere alte prestazioni (molte istruzioni SIMD)
  - esecuzione di molti threads (fino a $10^3$...) in parallelo distribuiti su molti cores elementary ($10^3$) di calcolo
  - non necessaria la cache per mascherare la latenza di accesso alla memoria → molto computing meno memoria
  - Uso di graphic memory “larga“ ($10^2$ bit) e “veloce” (N∗Ghz per bit line
- Tanta tecnologia allo stato dell’arte:
- Linguaggi di programmazione standard (DirectX, OpenGL, OpenCL) o proprietary (NVidia Compute Unified Device Architecture (CUDA))
- Evoluzione verso sistemi scalabili ottimizzati (anche) per AI
  - Sistemi a scala estrema HGX essenzialmente dedicati al training efficenti di reti deep
  - Integrazione CPU+GPU (Grace)

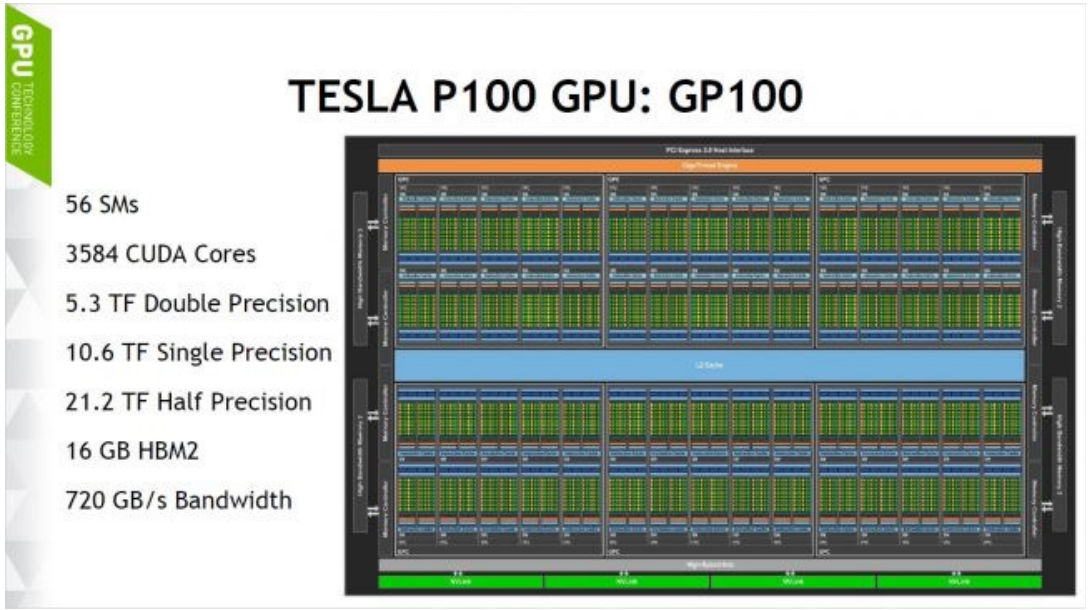
# GPU: tecnologie allo stato dell'arte

Tensor core con supporto per dati a precisione ridotta/ottimizzata



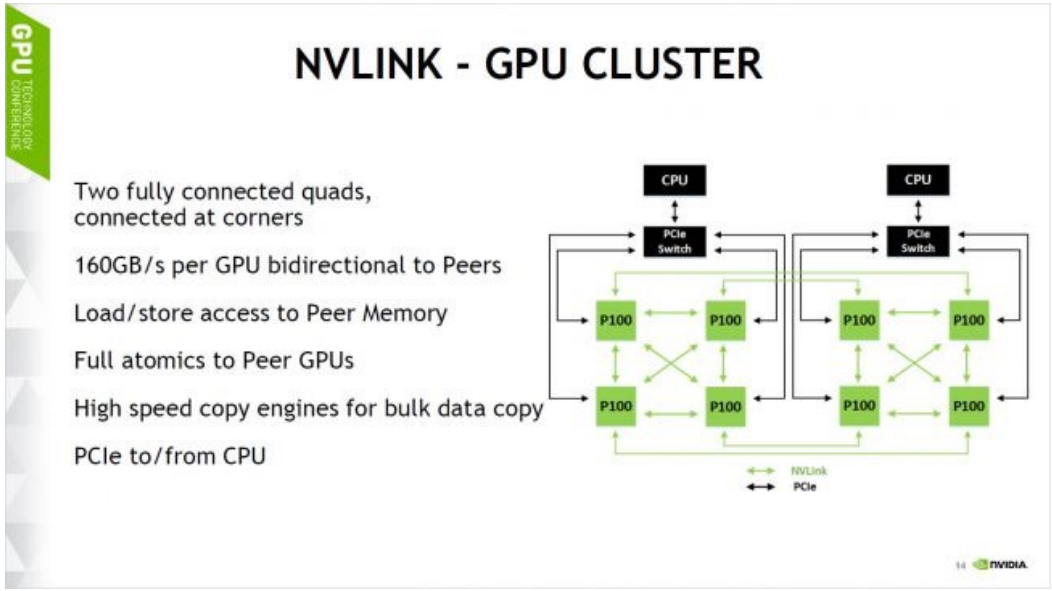




Network integrata per connessione peer-to-peer ad alta banda e bassa latenza



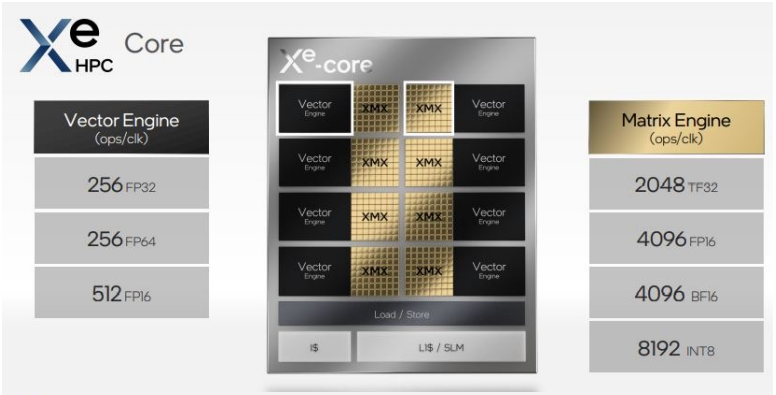


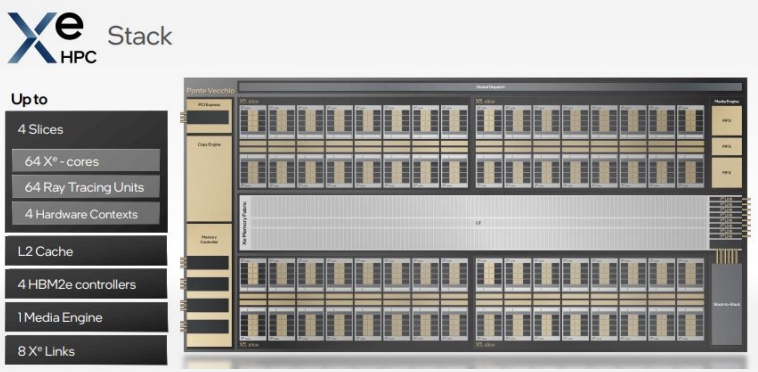
Network integrata per connessione peer-to-peer ad alta banda e bassa latenza

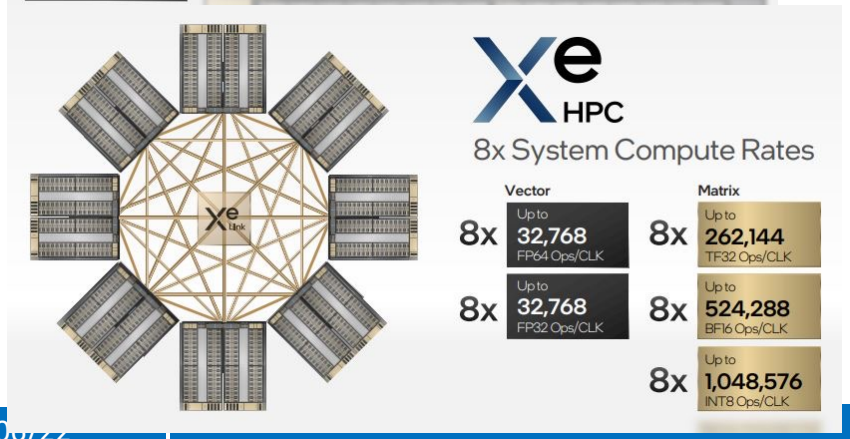


# GPU: non solo NVidia

- INTEL Ponte Vecchio (release in 2022)

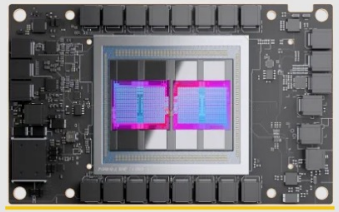
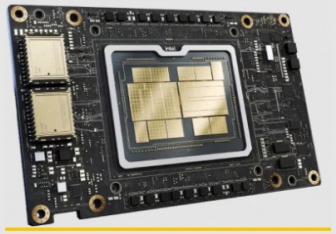




**2022 HPC GPUs**

| VideoCardz.com | NVIDIA H100 SXM | AMD Instinct MI250X OAM | Intel Ponte Vecchio OAM |
|---|---|---|---|
| Picture |  |  |  |
| GPU | **GH100** | Aldebaran (MCM) | Ponte Vecchio (MCM) |
| Transistors | **80B** | 58.2B | 100B |
| Die Size | **814 mm²** | 2x ~790 mm² | 2x 640 mm² |
| Architecture | **Hopper** | CDNA2 | Xe-HPC |
| Fabrication Node | **TSMC N4** | TSMC N6 | Intel 7, TSMC N5/N7 |
| GPU Clusters | **132 (SMs)** | 220 (CUs) | 128 Xe-Cores |
| FP32 Cores | **16896 CUDAs** | 14080 Stream Proc. | 32768 (FP32) |
| L2 Cache | **50MB** | 32MB | 408 MB |
| Tensor/Matrix Cores | **528** | 2x 440 | 128 |
| Memory Bus | **5120-bit** | 8192-bit | 8192-bit (?) |
| Memory Size | **80 GB HBM3** | 128GB HBM2e | 128GB HBM2e (?) |
| TDP | **700W** | 560W | ~600W |
| Interface/Form Factor | **SXM5/PCIe Gen5** | OAM/PCIe Gen5 | OAM/PCIe Gen5 |
| Launch Year | **2022** | 2021 | 2022 |

# Ingredienti: FPGA

FPGA: componente programmabile per flessibilita' ed efficenza energetica e riduzione del "time-to-market"

- dalle PAL/GAL agli attuali system-on-chip da miliardi di gates
- Un esempio: Xilinx Ultrascale+
  - → 14nm/16nm FinFET node (prossima generazione 7nm) diverse 10x Btransistors
  - → Multiple (4->8) ARM Cores (a53/57) @1.5GHz
  - → 128 transceiver 32-56 Gbps per interconnessione chip-to-chip o via backplane
  - → 9M system logic cells (up to 1GHz)
  - → Up to 16GB in-package HBM DRAM (~500GB/s) e 500Mb memory
  - → molti standard industriali implementati come IP core hardware: ETH100g o 200g, PCIExpress gen3/4 x16...
  - → 38 TOPs (22 TeraMACs) DSP computing performance
  - → IP specializzate per ML inference

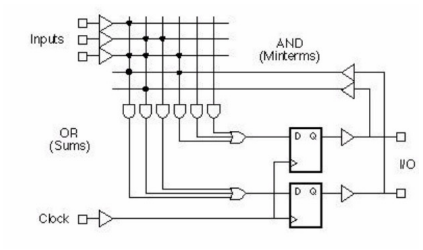
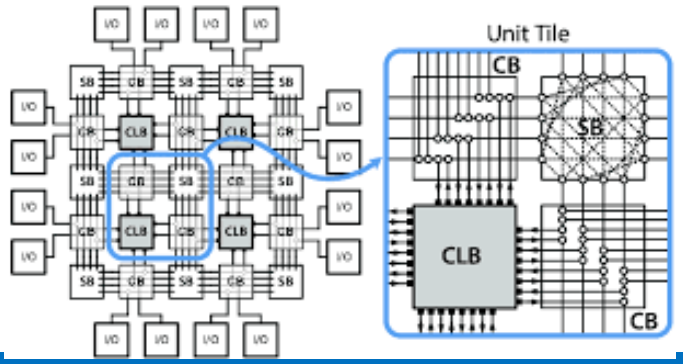
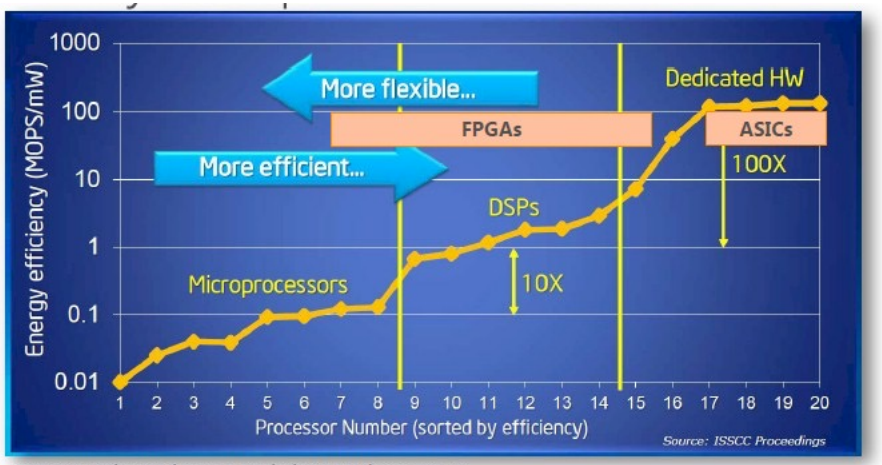
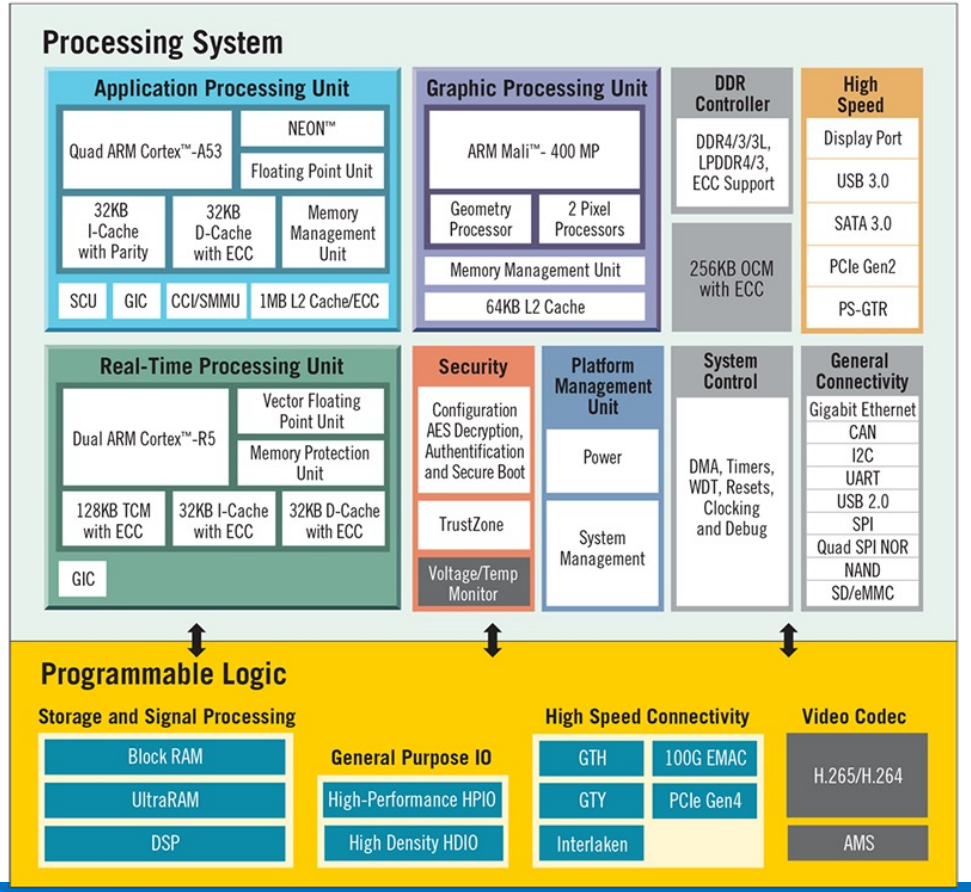


Figure 3 PAL Architecture



Source: Bob Broderson, Berkeley Wireless group



# FPGA Programming

- in origine sintesi a mano... poi strumenti di disegno “grafico” dello schema (Schematic design) e sintetizzatori che mappano disegno di alto livello in una netlist hardware (dipendente dal particolare componente)
- con l’aumento della complessita’ → linguaggi astratti di descrizione circuitale HDL Hardware Description Language (VHDL, Verilog) + compilatori + sintesi
  - ostici per softwaristi, astratti dall’HW ma non troppo
- necessita’ di rendere ancora piu’ astratto il design → HLS (High Level Synthesis) dal C++ (o simili)
  - Modello ”task parallelism” che sfrutta il parallelismo intrinseco della FPGA
  - Tools di sviluppo Xilinx VITIS HLS, INTEL oneAPI (FPGA HLS),...
- Grazie a questi higher-level programming tools → “democratizzazione” dell’uso delle FPGA
  - ridotto time–to-design (dalla descrizione in linguaggio standard di alto livello al firmware di programmazione della FPGA)
  - non e’ necessaria conoscenza profonda dell’HW e dell’architettura del componente; un “semplice” application specialist puo’ realizzare un design FPGA ottimizzato per quella particolare applicazione
  - implementazione di nuovi strumenti di alto livello per mapping automatico e specializzato
    - OmPSS- FPGA (BSC) per programmare sistemi paralleli basati su FPGA partendo da applicazioni MPI-based
    - HLS4ML (CERN) generazione automatica di firmware FPGA per task di ML

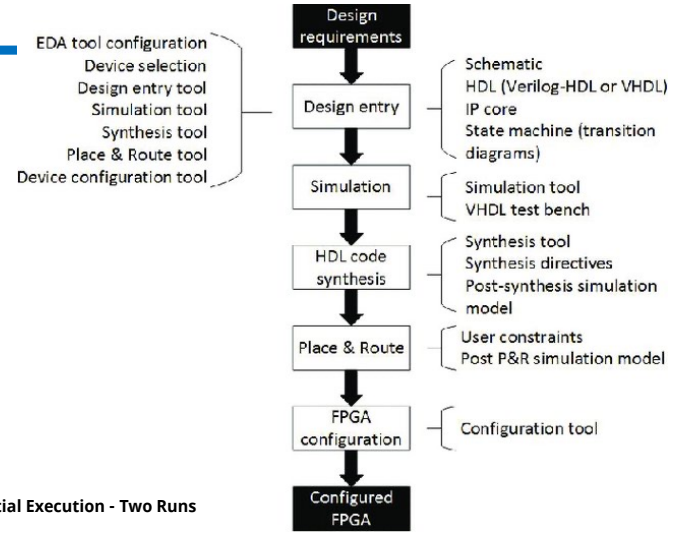



*Figure 5:* **Sequential Execution - Two Runs**

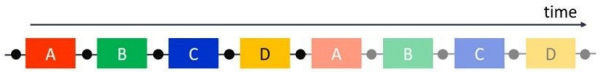



*Figure 6:* **Task Parallelism within a Run**

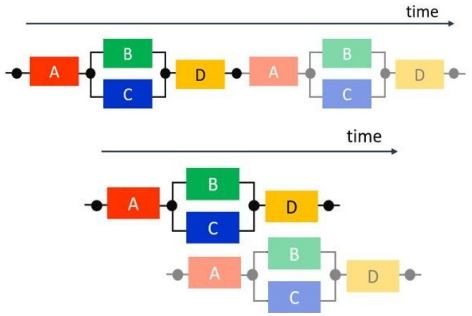
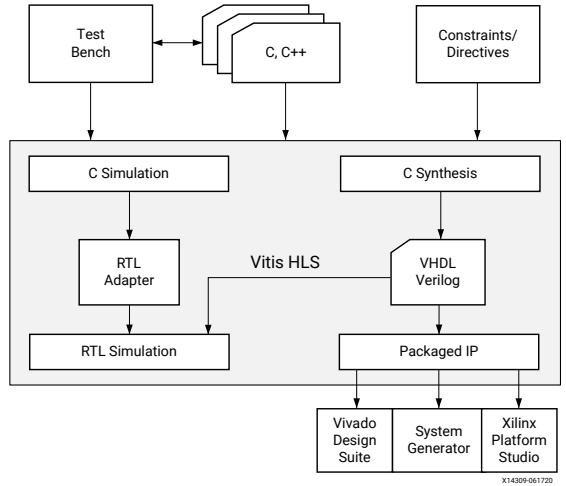


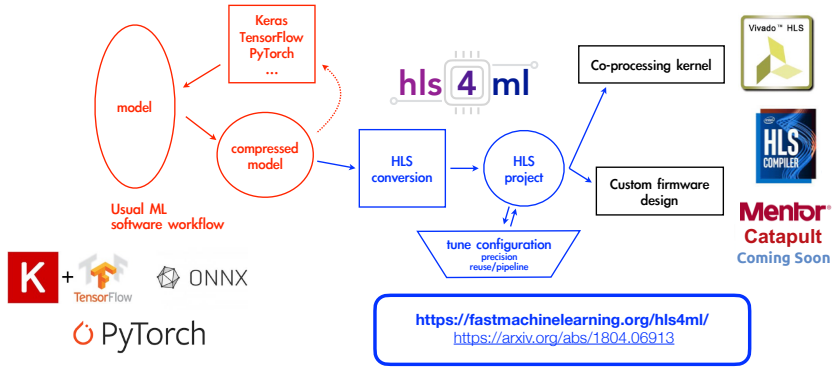

*Figure 18:* **Vitis HLS Design Flow**



# FPGA Programming (per ML): HLS4

## HLS4ML: the idea

- HLS4ML aims to be this automatic tool
  - reads as input models trained on standard DeepLearning libraries
  - comes with implementation of common ingredients (layers, activation functions, etc)
  - Uses HLS softwares to provide a firmware implementation of a given network
  - Could also be used to create co-processing kernels for HLT environments

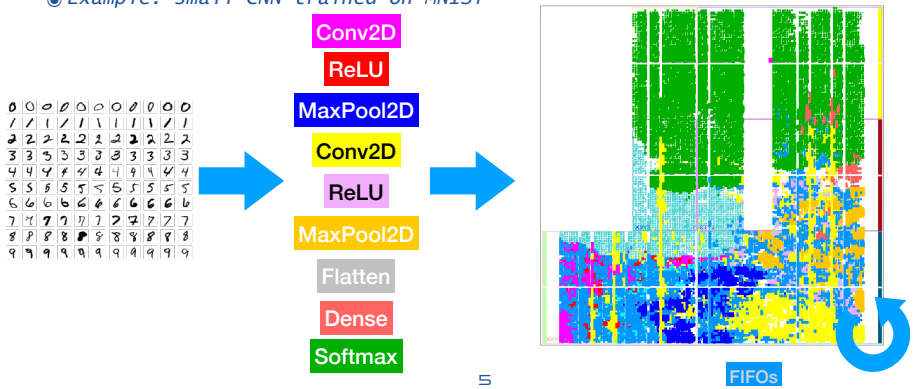



4

- Supporto (a vari livelli di maturita') per
  - pruning
  - compression
  - quantizzazione
  - parallelizzazione
  - Graph Nets
  - "Knowledge distillation" (teacher-student model)

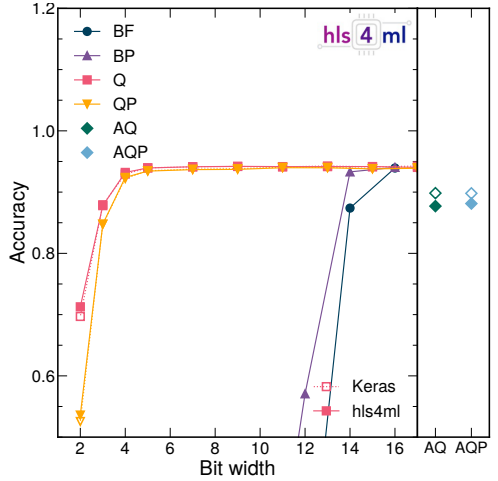
## HLS4ML: the implementation

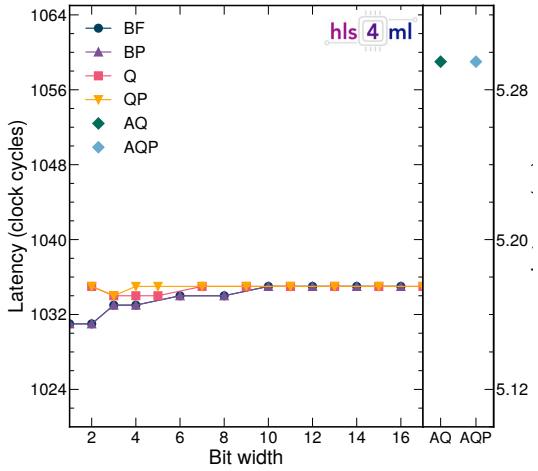
- Dataflow architecture: each layer is an independent compute unit
  - With tunable parallelism and quantization
- Fully on-chip: NN must fit within available FPGA resources (pynq-z2 floorplan shown)
  - Example: small CNN trained on MNIST




5

## Fast CNN inference on FPGAs



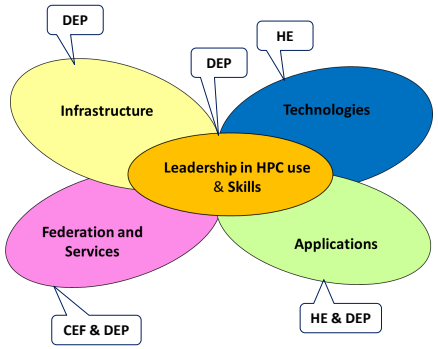




**Execution time reduced to 5 µsec to basically no accuracy loss down to 6 bits**

M. Pierini (WS AI INFN 2022) https://agenda.infn.it/event/29907/contributions/163448/attachments/90265/121584/AI%40INFN.pdf

# Finanziamenti EU

## Proposal for EuroHPC JU - The 5 pillars of activity




***Pillar 1: Infrastructure***

- ***Exascale*** *and* ***post exascale*** *supercomputers, National supercomputers,* ***Industrial-grade*** *supercomputers*
- ***Quantum Computers*** *(standalone or in hybridisation)*

***Pillar 2: Federation of supercomputing services***

- ***Interconnecting*** *all supercomputers and the Union's common European data spaces via terabit networks [CEF]*
- ***Federation*** *and secure service provisioning of supercomputing service and data infrastructures [DEP]*

***Pillar 3 and 4: Technologies & Applications***

- *R&D on new HPC technologies and architectures and their integration in supercomputing systems*
- *Advanced industrial, scientific and public sector applications*

***Pillar 5: Leadership in use and Skills***

- *Wide use mainly for civilian applications + for EU strategic initiatives (Destination Earth, digital human, etc.)*
- *HPC Skills → Education, Training*

## Infrastructure Investment Plan for 2021-2027

Aim: Provide a world-class supercomputing and data infrastructure for the EU

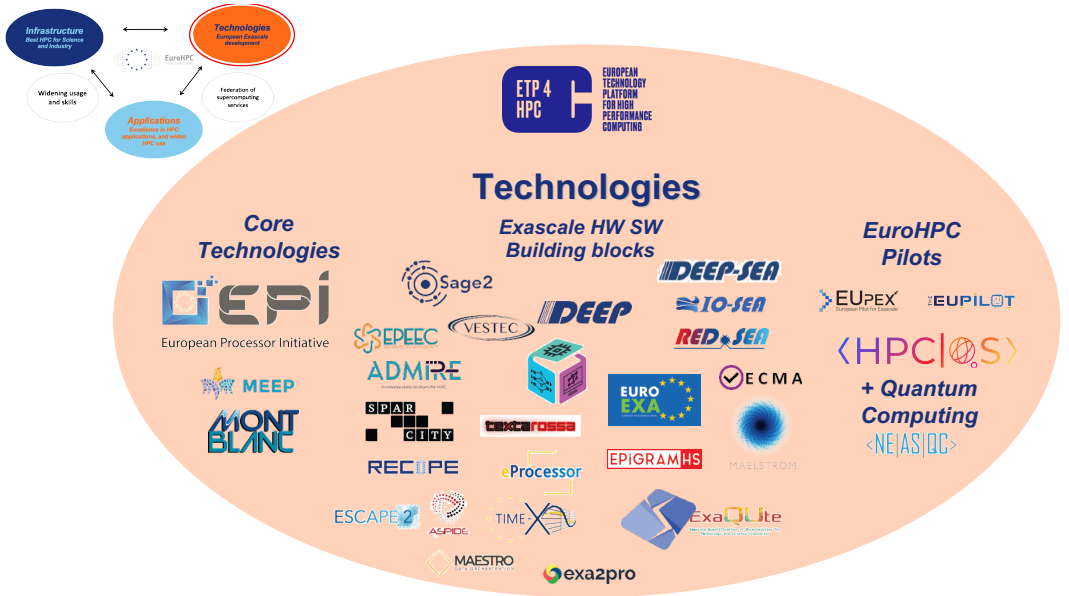
**More than 1.5 B€ of investments (EU budget)**

DRAFT PROPOSAL

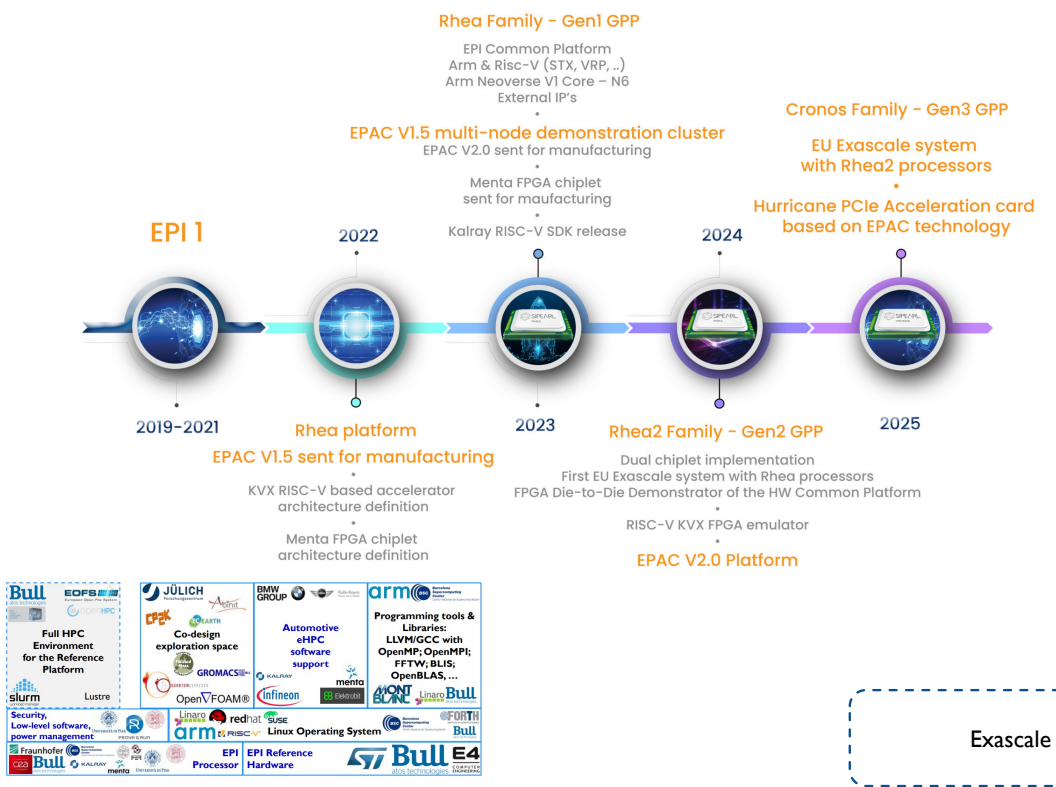
| | 2019 & 2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 | 2027 |
|---|---|---|---|---|---|---|---|---|
| **HPC Infrastructure** | 3 pre-exascale and 5 petascale HPC systems | Several pre-exascale systems and 2 exascale HPC systems | | | | One or more exascale and post-exascale HPC systems | | |
| **Quantum Infrastructure** | Quantum simulators interfacing with HPC systems | First generation of quantum computers | | Quantum simulators interfacing with HPC systems | | Second generation of quantum computers | | |

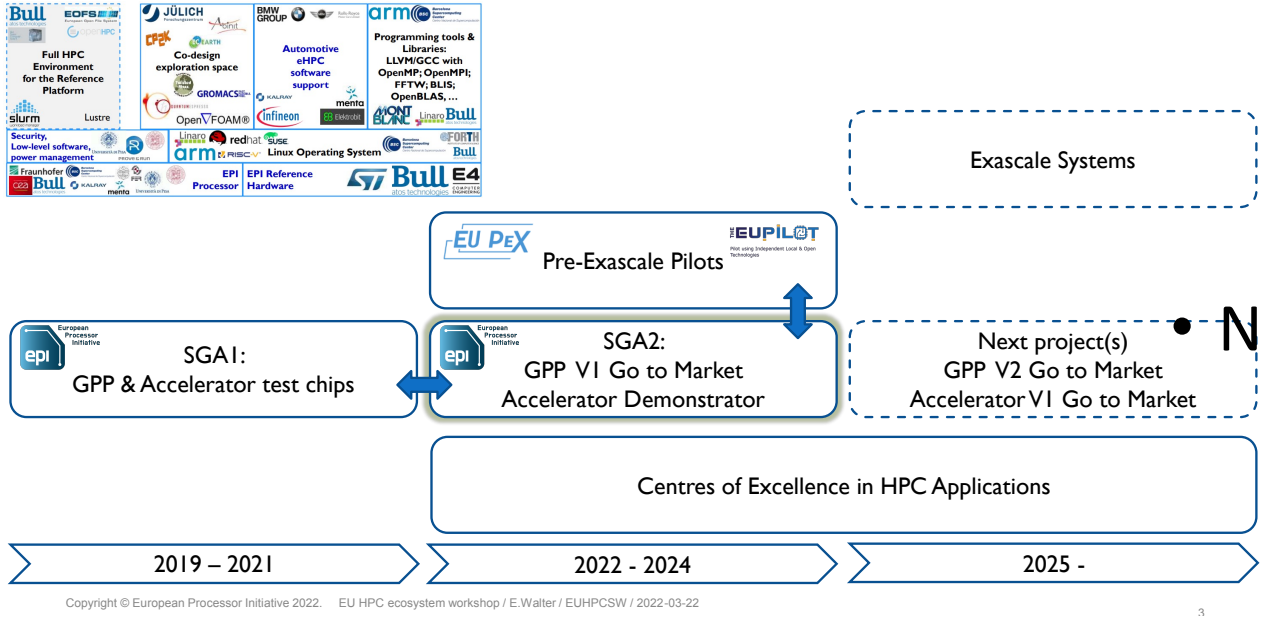
## Potential synergy with other major initiatives

- ***The common European Data Spaces***
  (rollout of common European data spaces in crucial economic sectors and domains of public interest, looking at data governance and practical arrangements - for AI purposes and applications)
- ***AI and Cybersecurity***
- ***The Quantum Technologies Flagship (Horizon-Europe)***
  (link to quantum computing)
- ***Joint Undertaking on Key Digital Technologies***
  (for edge computing, AI/neuromorphic chips, processor technologies)
- ***The European Open Science Cloud (EOSC, Horizon-Europe)***
- ***Destination Earth initiative (Digital Europe and Horizon-Europe)***
  (digital twin of the Earth for climate modelling and crisis management)



# EPI (European Processor Initiative)



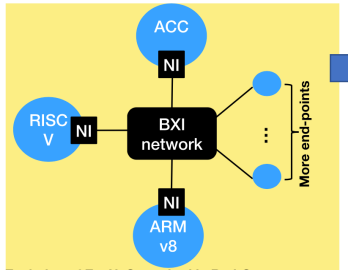

- European Processor Initiative: EU (EuroHPC JU) funds per provare a chiudere il "gap" con US e Japan
  - Investimento pubblico/privato per un totale aspettato a regime di circa 1-5 BEuro
    - multi step project basato su ARM e poi RISC-V per CPU e acceleratori
    - R&D accademico/industriale iniziale (SGA1 e SGA2 ~200MEuro) e trasferimento tecnologico (SiPearl)
- E' sufficente questo investimento?
  - NO per un prodotto competitivo in termini di mass production
  - necessario per garantire una forma di presidio tecnologico e per permettere lo sviluppo di nuove idee, architetture, hardware e software, nuovi campi applicativi
- Nel frattempo…
  - BSC partnership con INTEL per next gen processor (RISC-V)
    - https://www.techradar.com/news/**will-intel-abandon-x86-for-risc-v-for-its-next-gen-supercomputing-chips**

# EuroHPC JU R&D: progetto RED-SEA (Network) (A.Biagioni, P. Vicini

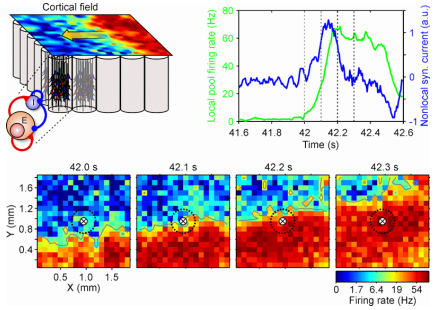
## The four pillars of RED-SEA research

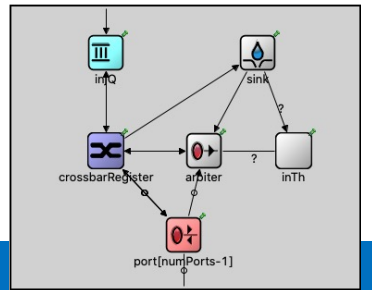
| | | |
|---|---|---|
| Architecture, co-design and performance | Optimizing the fit with the other EuroHPC projects and with the EPI processors | INFN |
| High-performance Ethernet | Development of a high-performance, low-latency, seamless bridge with Ethernet | Atos |
| Efficient Network Resource management | Including congestion management and Quality-of-Service targets while sharing the platform across application and users | Universitat Politècnica de València |
| Endpoint functions and reliability | End-to-end enhancements to network services - from programming models to reliability & security and to in-network compute | FORTH |


Project start: 01/04/2021
Project duration: 36 months
Project budget: 8 M€ (INFN 700k€)
)




END-POINT: INFN APEnetX




Integrazione della Network Interface (NI) con **RISC-V e ARMv8 cores (EPI) , piattaforma EU di HPC Network (Atos BXI)e con acceleratori FPGA e GPU**

- NEST (Spiking NN simulator) come benchmark e co-design application
  - Sviluppo di network IP per ottimizzazione Spiking NN simulator
- APEnet+ network simulators a larga scala
- Funzioni di network routing assistite da tecniche di ML

# EuroHPC JU R&D: progetto TextaRossa (A. Lonardo, P. Vicini)

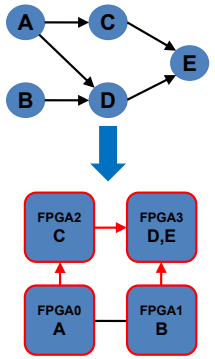
Obiettivi principali

- Energy Efficiency
- Sustained Performance delle applicazioni
- Integrazione di acceleratori riconfigurabili (FPGA)
- Sviluppo di IP
  - comunicazione, mixed precision AI, security, power monitoring,...
- Rilascio di nuove piattaforme (IDV)

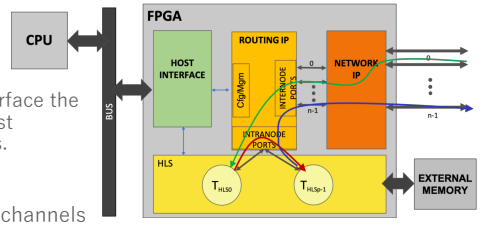

11 partners from 5 countries:
ENEA, Fraunhofer, INRIA, ATOS, E4, BSC, PSNC, INFN, CNR, IN QUATTRO, CINI (Politecnico di Milano, Università di Torino, Università di Pisa),
LTP: Universitat Politecnica de Catalunya (UPC), Université de Bordeaux.

## INFN Contribution to WP2/WP4: APEIRON

- Goal: offer hardware and software support for the execution on a system of **multiple interconnected FPGAs of applications developed according to a dataflow programming model**
- Map the directed graph of tasks on the distributed FPGA system and offer runtime support for the execution.
- Allow users with **no (or little) experience in hardware design tools** to develop their applications on such distributed FPGA-based platforms
  - Tasks are implemented **in C++ using High Level Synthesis tools (Vitis).**
  - Simple **Send/Receive** C++ communication API.




## INFN in WP2: IPs for low-latency FPGA commun.

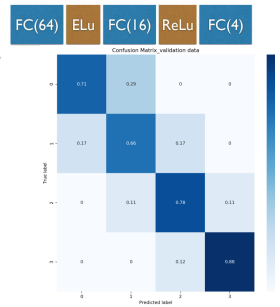


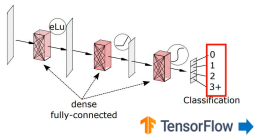

- **Host Interface IP:** Interface the FPGA logic with the host through the system bus.
  - PCI Express Gen3 → Gen4
- **Network IP:** Network channels and **Application-dependent I/O**
  - APElink 32 Gbps → 64/100 Gbps
  - UDP/IP over 10-25 GbE → 40/100 GbE
- **Routing IP**
  - Routing of intra-node and inter-node messages between processing tasks on FPGA.
- Implemented as incremental development on APEnet IPs over XILINX platforms.
- **Deliverable D2.5**
  - Intermediate database at M18
  - Deployed in the IDVs (WP5) at M30

## RAIDER Rings detection - Dense model on FPGA

**Fully Connected**

- **Input: 64 hits per event**
- **Architecture: 3 fully connected layers**
- **Output: 4 classes (0, 1, 2, 3+ rings per event)**
- **Qkeras, quantization aware training:**
  - ~75% average accuracy with low resource usage: LUT 14%, DSP 2%, BRAM 0% (VCU118)
- **Latency: 22 cycles @ 150MHz**
- **Initiation Interval (II): 8 cycles**

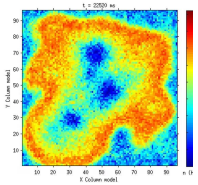



TensorFlow → hls4ml → VIVADO HLS → IP integration in LOTP+ infrastructure

https://fastmachinelearning.org/hls4ml/

## Nest GPU (as NEST on GPU)



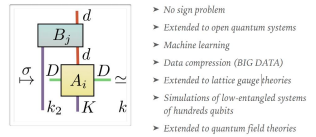
- The engine driving the neural simulations is the Nest GPU code which is C++ with CUDA extensions and is production-ready
- The Python script detailing the experimental protocol is ready – a 1000ms simulation of dynamics of one hemisphere of cortex of mouse brain with a realistic connectome inferred from data obtained with optical imaging methods on anesthetized mice – and will be run by the Nest GPU engine on the reference platform.
- As soon as the GPU-equipped is available, the simulation is ready to be benchmarked comparable with the same experiment on CPU-only engine (NEST).
- The specific KPI are:
  - Time-to-solution: Simulated-milliseconds-per-second
  - energy-to-solution: Synaptic UPdates per second (SUPs) per Watt

## High Energy Physics high-level software tools

- For simulation, reconstruction (i.e. the transformation of detector signals to physics objects), data analysis
- Initial focus will be on the reconstruction software of the CMS experiment
  - Efforts are on-going to investigate parallelism and heterogeneous computing (CPU, GPU, possibly FPGA), based on TBB, CUDA, SYCL/OneAPI, Cupla/Alpaka, Vitis HLS, ...
  - Some solutions are already in production, but investigation continues
- We have identified two software components, for particle tracking and calorimeter clustering
- Two directions of work
  - Use of GPUs and FPGAs via SYCL
  - Remote offloading of computation to specialized nodes
- Activity just started, due to delays in recruiting

## Tensor Network Methods

TENSOR NETWORK ALGORITHMS

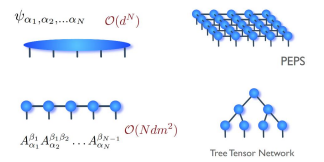


- State of the art in 1D (poly effort)
- No sign problem
- Extended to open quantum systems
- Machine learning
- Data compression (BIG DATA)
- Extended to lattice gauge theories
- Simulations of low-entangled systems of hundreds qubits
- Extended to quantum field theories

S. Montangero "Introduction to Tensor Network Methods", Springer (2019)

U. Schollwock, RMP (2005) A. Cichocki, ECM (2013) I. Glasser, et al. PRX (2018)

TENSOR NETWORKS STATES




Tensor networks states are a compressed description of the system tunable between mean field and exact

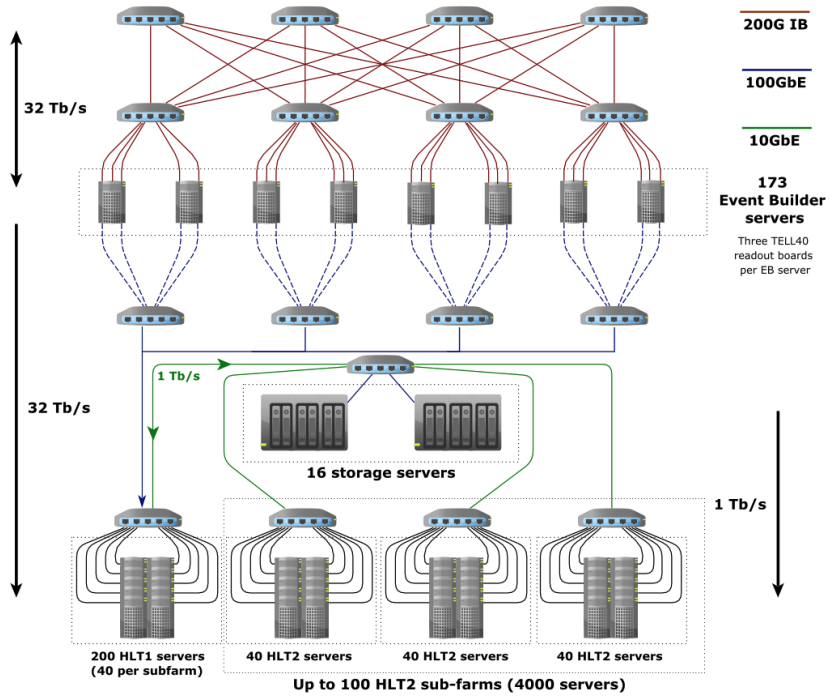
Tensor network are state of the art methods for the simulation of many-body quantum systems, to understand complex quantum phenomena and to benchmark, verify and guide the developments of emerging quantum technologies (computers, simulations, sensors and communication).

Interpolation between mean field theory and exact description, faithful compression of the exponentially large many-body wave function.

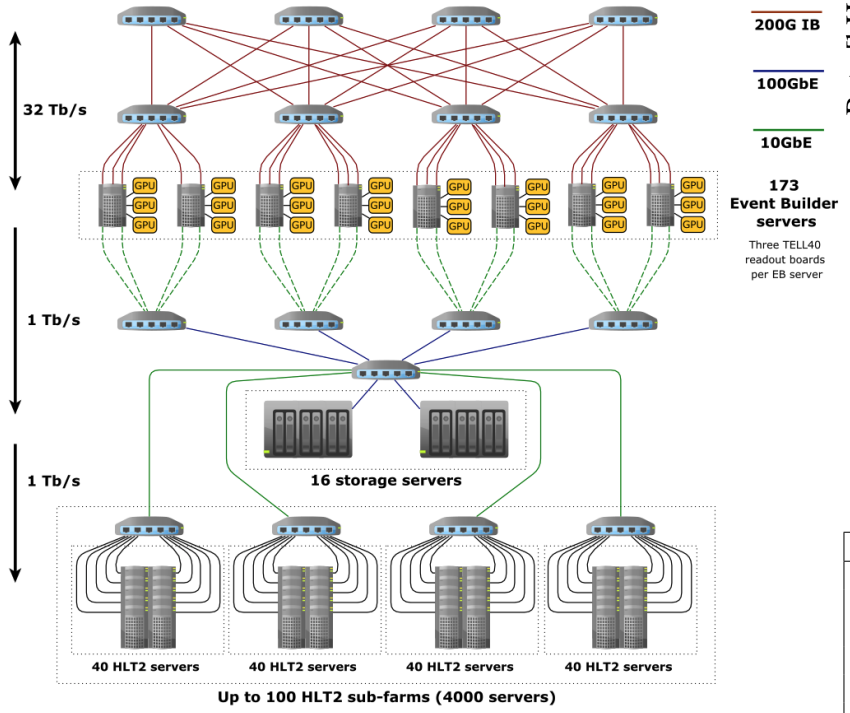
# Calcolo eterogeneo e HEP/GW/Simulazioni SNN

- GPU necessarie per la tipologia di carico computazionale aspettato per il prossimo futuro
  - SIMD (parallele), power effective, cost effective, ad alta densita'...
  - Qualsiasi sistema HPC prevede acceleratori GPU-based
- Esistono varie applicazioni ad alte prestazioni in HEP che ne motivano l'uso ma la valutazione di integrazione nei sistemi di DAQ, EB e analisi e' partita con un certo ritardo
  - Event selection ma anche simulazione, ricostruzione etc
- In generale imparare ad usarle efficacemente e' un'opportunita' per agganciarsi al treno del mainstream tecnologico dell'HPC e parassitizzare le risorse HPC dei data center
- Esempi di use cases (una lista parziale) :
  - LHCB-HLT1
  - CMS (Patatrak)
  - ALICE O2
  - NA62 (GPU-Rich)
  - GPU in GWA
  - GPU nel simulatore di HBP(Spiking Neural Network)
- FPGA usate principalmente per on-line low level trigger e DAQ
  - Nuove applicazioni tipicamente legate al ML

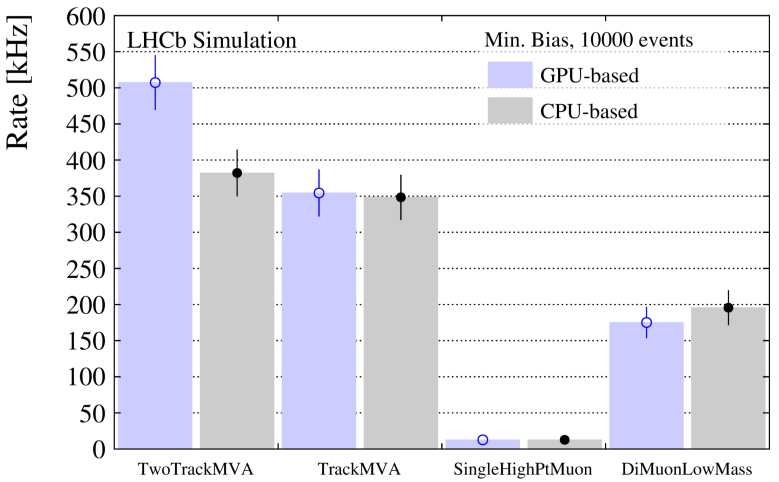
# GPU in HEP Use case: LHCB-HLT1




CPU-only




Hybrid (real-time analysis)




| Item | CPU-only | hybrid | Difference |
|---|---|---|---|
| Event Builder nodes | 1000 | 1000 | 0 |
| HLT1 network | 275 | 25 | 250 |
| HLT1 compute | 450 | 125 | 325 |
| Storage for 1 MHz output | 575 | 575 | 0 |
| **Sub-total** | 2300 | 1725 | 575 |
| Storage add. cost 2 MHz output | 575 | 575 | 0 |
| **Total** | 2875 | 2300 | 575 |

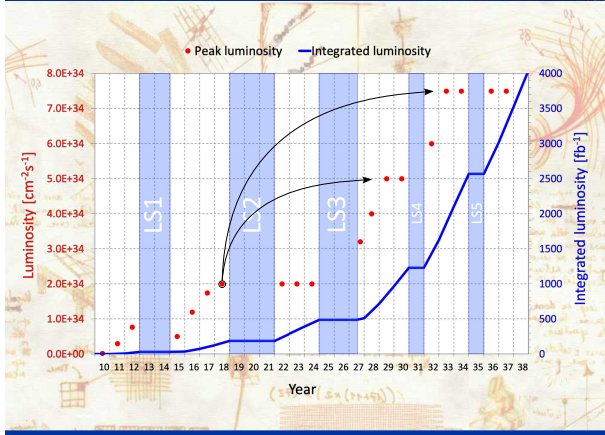
Costi

- ”lower” perfomance networks (EB→HLT1)
- no server dedicati per HLT1 (HLT1 su GPU)
- Stessa *physics efficency* (forse meglio) → decisione su adozione puo’ basarsi SOLO su valutazioni di costo/densita’

*Aaij, R., et al. A Comparison of CPU and GPU Implementations for the LHCb Experiment Run 3 Trigger. Comput Softw Big Sci **6**, 1 (2022)*
https://cds.cern.ch/record/2766501/files/2105.04031.pdf

# GPU in HEP Use case: CMS Patatrack

## High Luminosity LHC




CMS Phase 2 upgrades
- high granularity endcap calorimeter
- new silicon tracker and muon detectors
- tracking and particle flow at Level 1 trigger
- improved electronics for barrel calorimeter

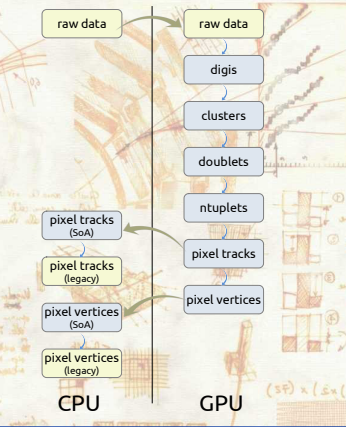
Run 4 (vs 2018)
- 2.5x LHC luminosity and pileup
- 5x CMS Level 1 Trigger rate

Run 5+ (vs 2018)
- ~4x LHC luminosity and pileup
- 7.5x CMS Level 1 Trigger rate

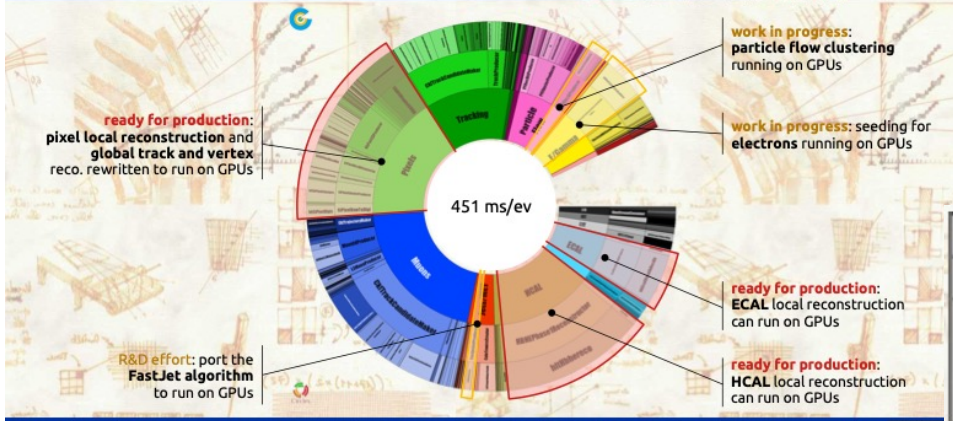


## the Patatrack pixel reconstruction

- the overall approach
  - reconstruct pixel-based tracks and vertices on the GPU
  - leverage existing support for threads and on-demand reconstruction
  - minimise data transfer
- the full workflow
  - copy the raw data to the GPU
  - run multiple kernels to perform the various steps
    - decode the raw data
    - cluster the pixel hits
    - form hit doublets
    - form hit ntuplets (triplets or quadruplets) with a Cellular Automaton algorithm
    - clean up duplicates
  - take advantage of the GPU computing power to improve the physics
    - fit the track parameters (Riemann fit, broken line fit) and apply quality cuts
    - reconstruct vertices
  - copy only the final results back to the host (optimised SoA format)
    - convert to legacy format if requested

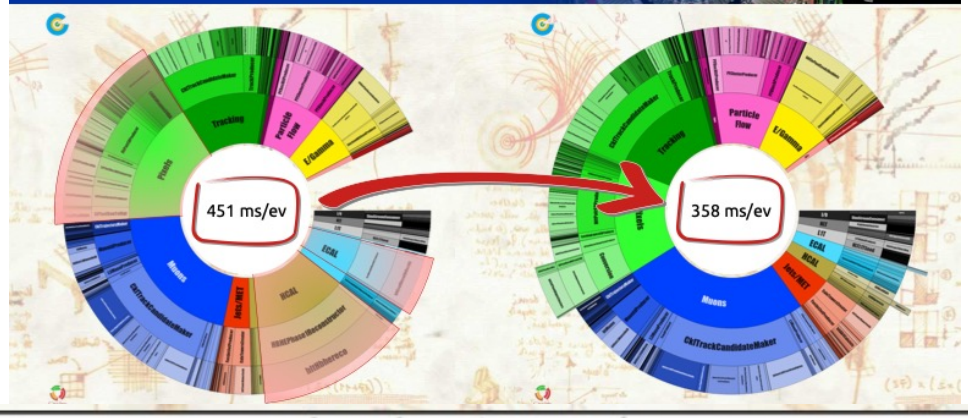



## online reconstruction on GPUs ?






## online reconstruction on GPUs !




| Configuration | throughput in events/s | | | | |
|---|---|---|---|---|---|
| | Triplets CPU | Triplets GPU | Quadruplets CPU | Quadruplets GPU | CMS 2018 |
| no copy | 611 | 870 | 892 | 1386 | 476 |
| copy, no conv. | — | 867 | — | 1372 | — |
| conversion | 585 | 861 | 855 | 1352 | — |

- pixel reconstruction performance on a single GPU vs a dual-socket CPU system

*A. Bocci 2020- https://indico.hep.caltech.edu/event/883/attachments/648/824/A._Bocci_-_Towards_a_heterogeneous_computing_farm_for_the_CMS_High_Level_Trigger.pdf*

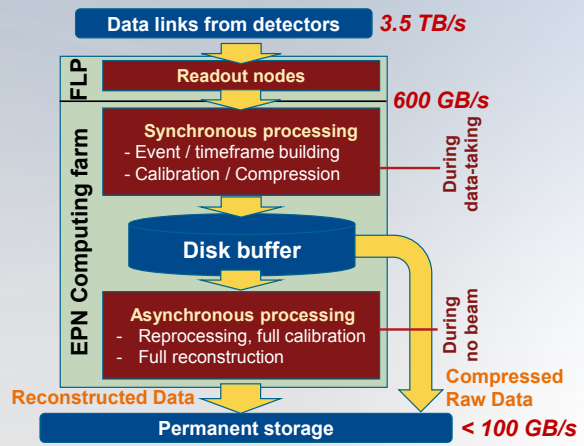
# GPU in HEP Use case: ALICE O2

## Run 3 data processing in a nutshell

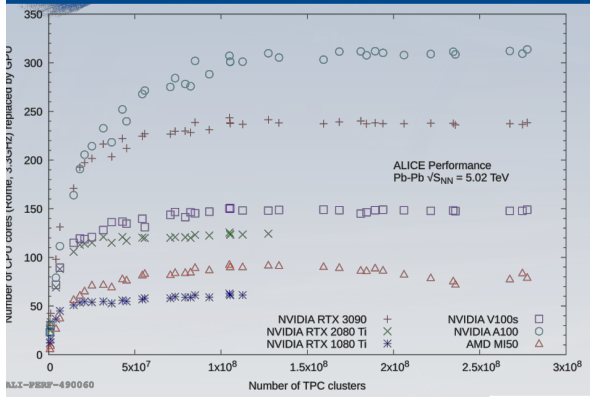
- **Synchronous processing** while there is beam in the LHC and raw data is recorded.
  - **99%** of compute time spent for **TPC**.
  - No trigger
    - All data is compressed and stored on a **disk buffer**.
- **Asynchronous reprocessing** when the EPN farm is not fully used for synchronous processing (No beam, pp data taking, ...).
  - More detectors with significant computing contribution.
- Following up 2 scenarios:

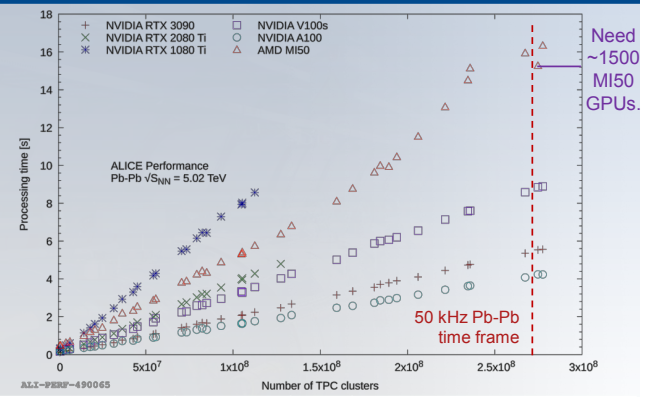
**Baseline solution (available today):**
- Mandatory for synchronous processing
- Most of sync. reco on GPU

**Optimistic solution:**
- Achieve best GPU usage in async phase
- Run most of tracking + X on GPU




## GPU Performance (standalone benchmark)




- **MI50 GPU** replaces **~80 Rome cores** in **synchronous** reconstruction.
  - Includes **TPC clusterization**, which is **not optimized** for the CPU!
- **~55 CPU cores** in **asynchronous** reconstruction (more realistic comparison).

| GPU Model | Performance | | GPU Model | Performance |
|---|---|---|---|---|
| NVIDIA RTX 2080 Ti | 100.0% | | NVIDIA V100s | 122.7% |
| NVIDIA Quadro RTX 6000 (active) | 105.8% | | NVIDIA RTX 3090 | 187.3% |
| NVIDIA Quadro RTX 6000 (passive) | 96.1% | | NVIDIA T4 | 59.3% |
| NVIDIA RTX 2080 | 83.5% | | AMD MI50 | 67.8% |
| NVIDIA GTX 1080 | 60.1% | | AMD Radeon 7 | 71,2% |

## Summary

- **ALICE will record 50 kHz Pb-Pb minimum bias collision data in Run 3 without trigger.**
  - **Continuous TPC readout, time frames** of 10 – 20 ms instead of events.
- **Full online data processing on GPUs.**
  - Computing farm consists of 250 servers, with 8 AMD MI50 GPUs, 2 32-core Rome CPUs, and 512 GB RAM each.
  - Currently 230 servers are sufficient for processing 50 kHz Pb-Pb (peak load).
  - MI50 GPU replaces ~55 CPU cores.
- **All GPU software written in generic way, can run on different GPUs and on the CPU.**
- **Processing farm used for synchronous (online) and asynchronous (periods without beam) processing.**
  - **Full baseline scenario** with **synchronous GPU processing ready**.
  - Planning to use GPUs as much as possible also in **asynchronous processing.**
    - **In the optimistic scenario, we will be able to offload ~95% of the workload to the GPU**.

*D.Rohr @ CHEP2021 https://indico.cern.ch/event/948465/contributions/4324179/attachments/2245852/3808987/2021-05-18%20CHEP2021.pdf*

# GPU in HEP Use case: NA62 GPU-RICH

NaNet: Design and implementation of a family of FPGA-based PCIe Network Interface Cards :

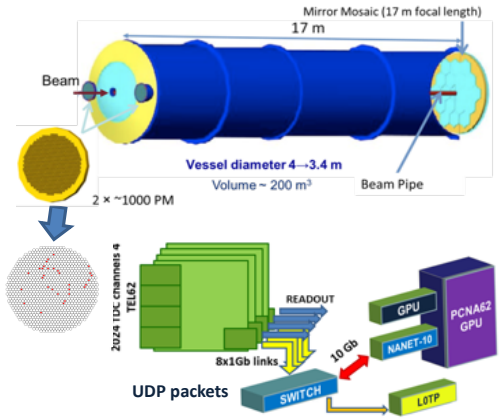
- Bridging the front-end electronics and the software trigger computing nodes.
- Supporting multiple link technologies and network protocols.
- Enabling a low and stable communication latency.
- Having a high bandwidth.
- Processing data streams from detectors on the fly (data compression/decompression and re-formatting, coalescing of event fragments, …).
- Optimizing data transfers with GPU accelerators.

NA62 KM3NeT NA62

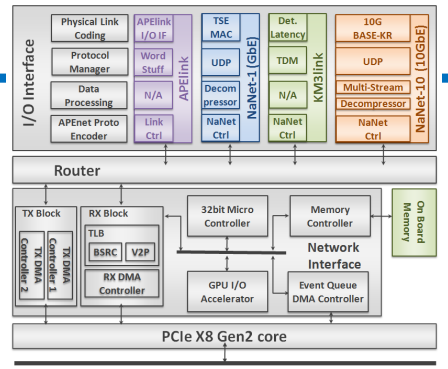
| | NaNet-1 | NaNet³ | NaNet-10 | NaNet-40 |
|---|---|---|---|---|
| Year | Q3 - 2013 | Q1 - 2015 | Q2 - 2016 | Q3 - 2019 |
| Device Family | Altera Stratix IV | Altera Stratix V | Altera Stratix V | Altera Stratix V |
| Channel Technology | 1 GbE | KM3link | 10 GbE | 40 GbE |
| Transmission Protocol | UDP | TDM | UDP | UDP |
| Number of Channel | 1 | 4 | 4* | 2 |
| PCIe | Gen2 x8 | Gen2 x8 | Gen3 x8** | Gen3 x8 |
| SoC | NO | NO | NO | NO |
| High Level Synthesis | NO | NO | NO | YES |
| nVIDIA GPUDirect RDMA | YES | YES | YES | YES |
| Real-time Processing | Decomp. | Decomp. | Decomp. Merger | ? |

*https://apegate.roma1.infn.it/?page_id=821*

## GPU-RICH overview

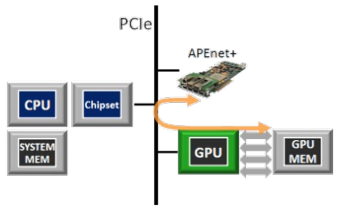



## NaNet architecture

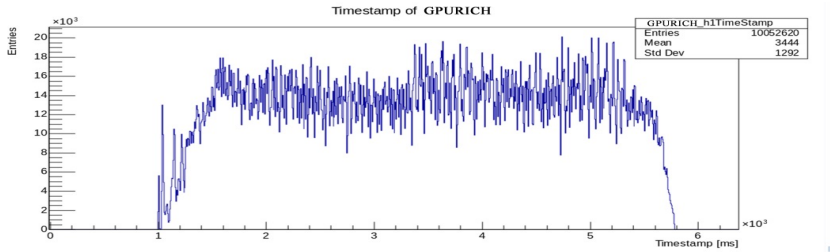



CPU+GPU

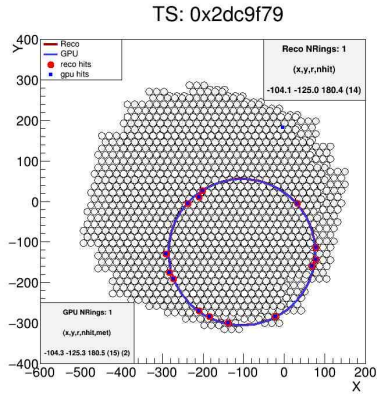
## FP7-ICT EURETILE APEnet+ (2012)



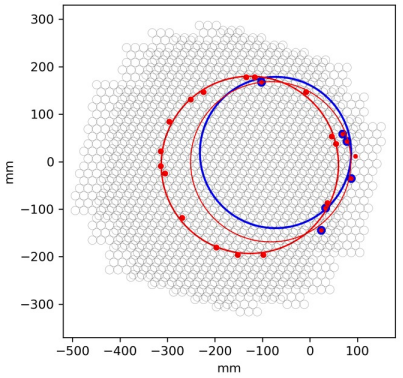

## GPU-RICH generated primitives (late Oct 2018)



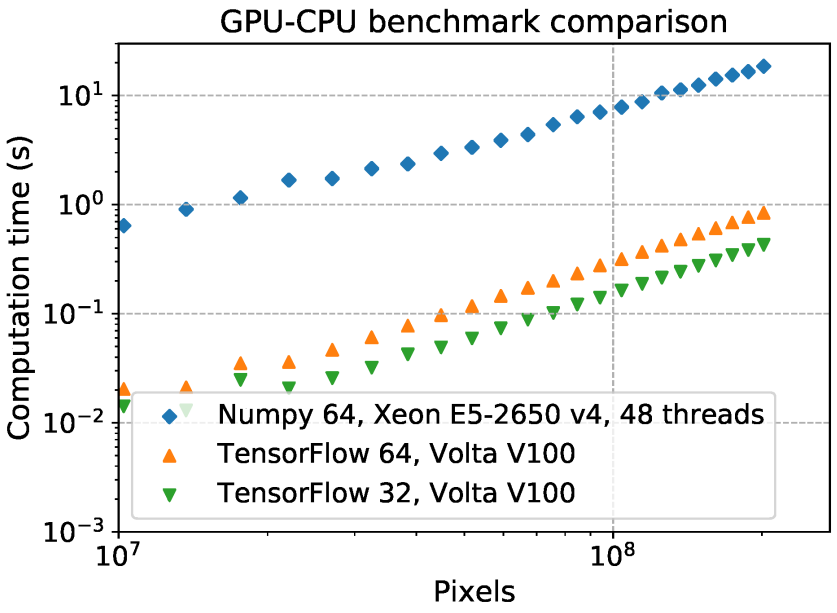

## GPU 1 ring == Reco 1 ring

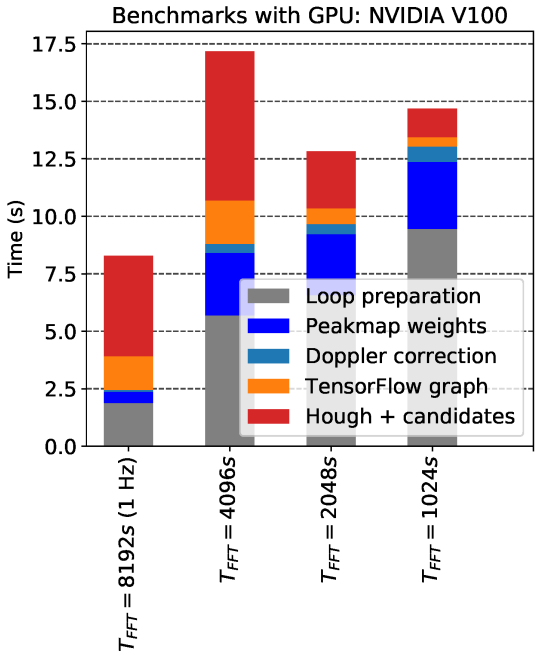


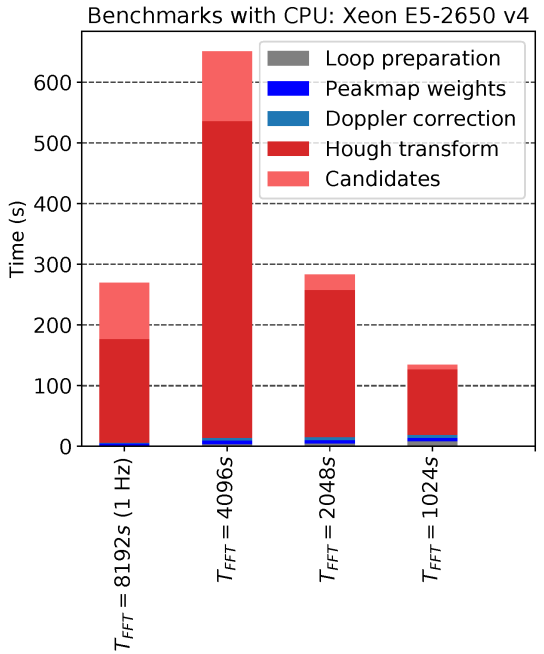

## GPU 1 rings Reco 2 rings

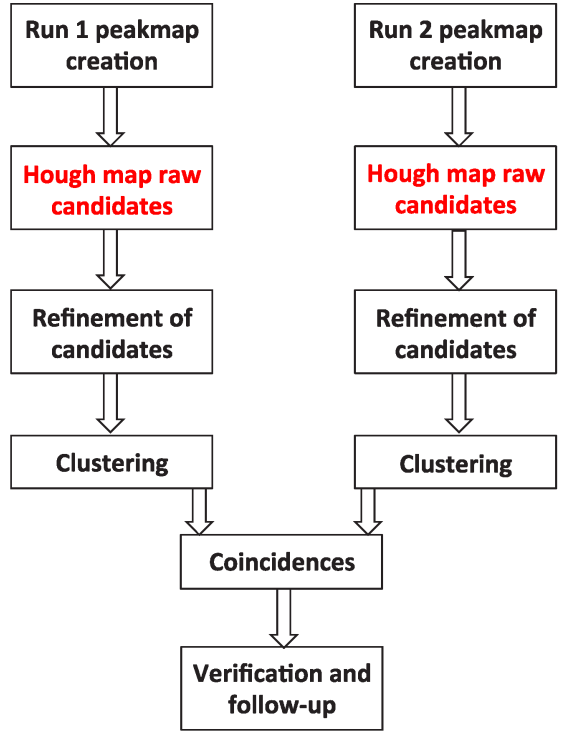


# GPU in GW Use case: Continuous gravitational Waves (CWs)



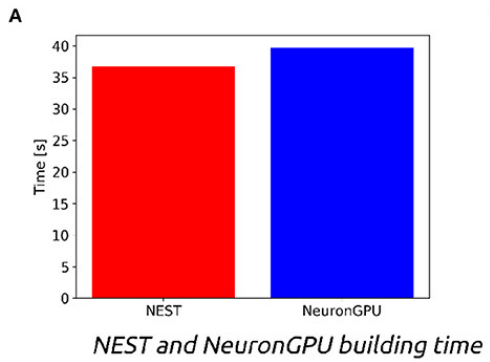

- Implementazione su GPU di FrequencyHough transform
- Utilizzo di TensorFlow (non la parte dedicata all'AI) come framework di programmazione efficente per calcolo scientifico su GPU
- Un fattore 10 di speed-up rispetto ad una CPU multi-core di taglia comparabile

*La Rosa, I.; Astone, et al "Continuous Gravitational-Wave Data Analysis with General Purpose Computing on Graphic Processing Units." Universe* ***2021****, 7, 218. https://doi.org/10.3390/universe7070218pdf*

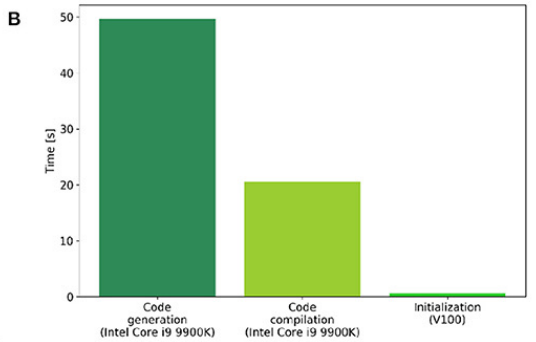
# GPU in Human Brain Project Use case: NEST-GPU

- EBRAIN: EU infrastructure per la comunita' di neuroscienziati (HBP)
- NEST e' "IL" simulator per spiking networks in HBP
  - si focalizza sulla dinamica e sulla struttura di un sistema neurale (quasi) biologico senza descriverne l'esatta morfologia del neurone
  - Configurabile per differenti modelli e scale della rete. Sviluppato in origine per CPU, poi CPU cluster.
  - Simulazione di modelli ma in principio anche NN engine per robotica
- Neuron-GPU (ora NEST-GPU) e' una libreria GPU-MPI per simulazioni a larga scala di spiking networks
  - circa 2 ordini di grandezza meglio a parita' di network size per simulazioni real-time

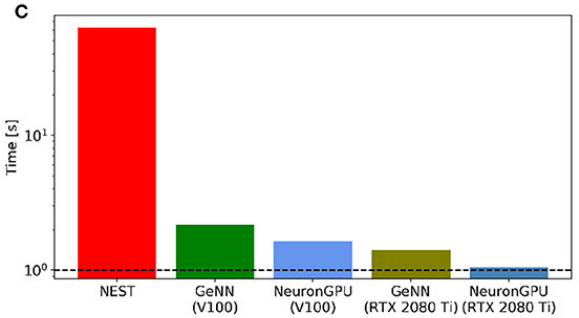




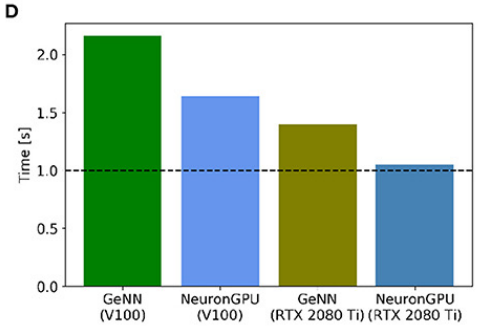

NEST and NeuronGPU building time




GeNN building time




NEST, GeNN and NeuronGPU simulation time



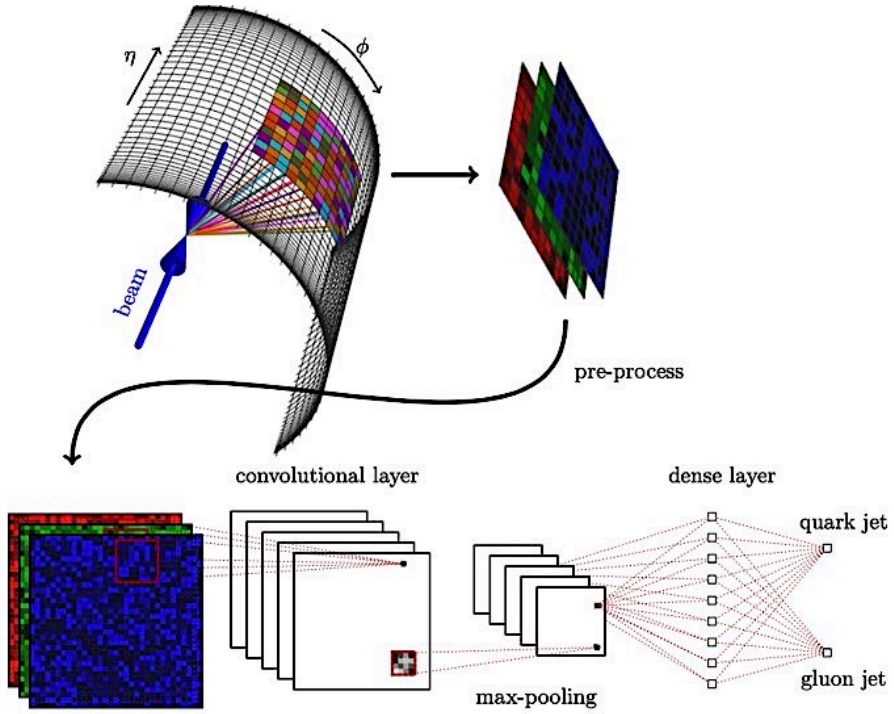

GeNN and NeuronGPU simulation time

Golosio,B et al Front. Comput. Neurosci., 17 February 2021 | https://doi.org/10.3389/fncom.2021.627620

# MACHINE LEARNING IN ONE SLIDE (credit S.Giagu)

- rough definition: a set of computational methods able to learn how to solve specific problems based on the experience, e.g. based on a set of examples of the problem and, possibly, a set of inductive priors (invariances of the problem, etc.)
  - can produce abstract and powerful representation of the input data (deep representation learning)
  - highly flexible in learning the set of rules that map these representations to the target of the task ($\hat{y} = \hat{f}_w(x)$)
  - underlying mathematical formulation closely tied to statistics, calculus of variations, approximation theory, and optimal control theory

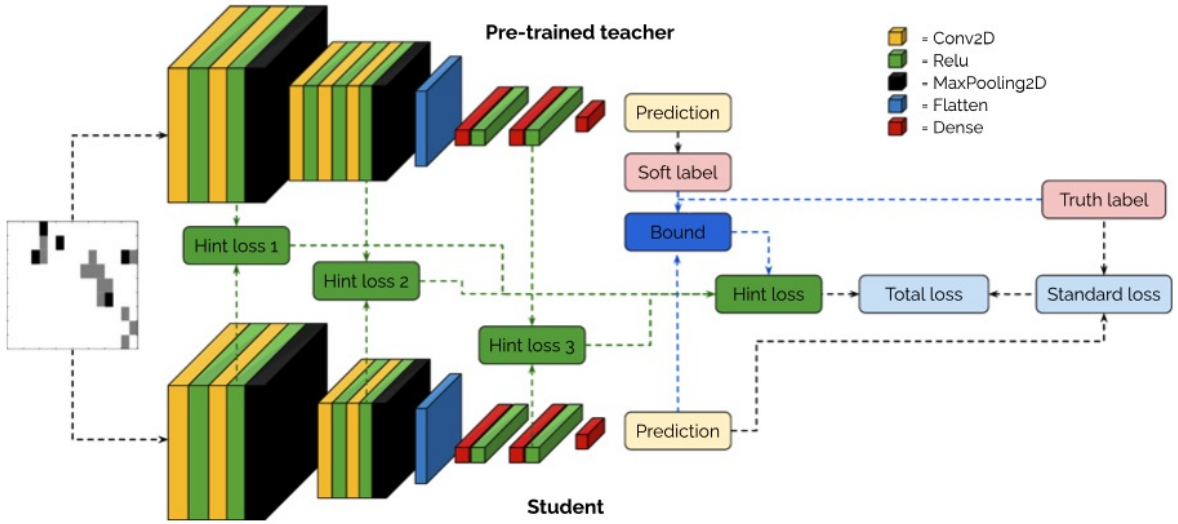
- deep learning: based on differentiable neural networks trained through gradient-based optimisation
- pervasive in AI applications in the last decade, thanks to the availability of large data sets, advances in computing (GPUs), and ANN techniques …



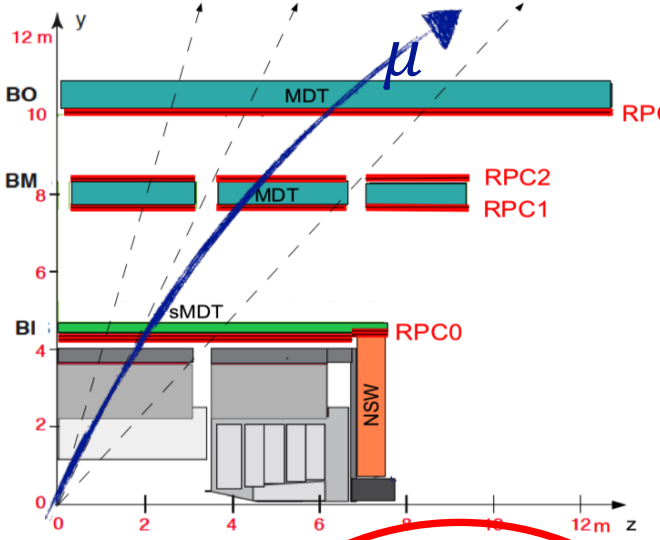
# DEEP NEURAL NETWORK FOR REAL-TIME TRIGGERS IN HEP (credit S.Giagu)

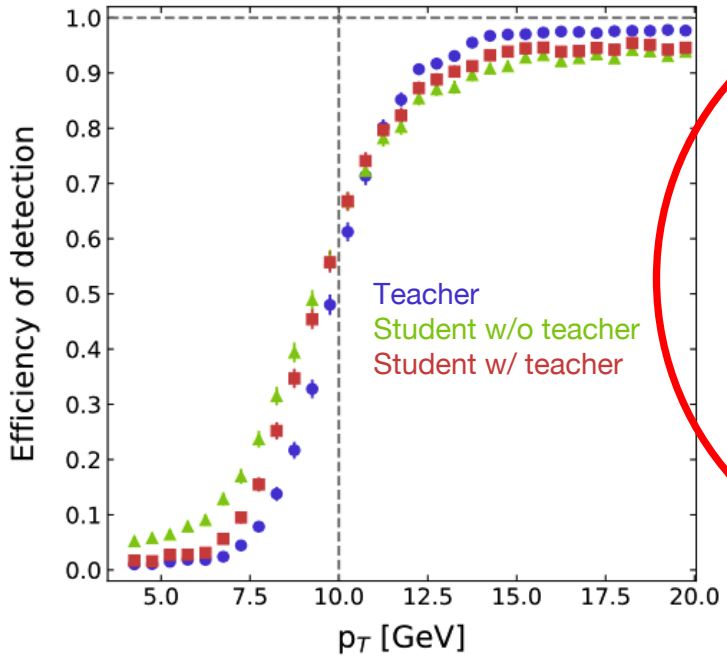
Goal: using FPGA to accurately reconstruct the momentum and angle of the muon track from the ATLAS RPC detector hit information **in less than 400ns** (x1000 faster than fastest AI models on CPUs and GPUs)

Strategy: multi-stage **AI model compression** and simplification based on **aggressive quantisation** and **knowledge transfer techniques** to avoid degradation of physics performances




transfer knowledge learned by a larger neural network pre-trained for the same task to a smaller and quantised (4-bits per activations and weights) model




Inference time
Xilinx Ultrascale+ XCV13P

- Teacher: 5 ms (V100 GPU)
- Student: 84 ns (FPGA)

*S. Francescato, S.Giagu, F. Riti, G.Russo, L.Sabetta, F.Tortonesi, Eur. Phys. J. C (2021) 81:969*

# ML-RELATED HEP ACTIVITIES IN ROME (credit S. Giagu)

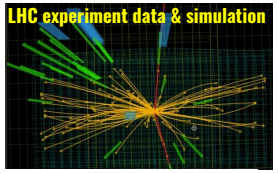
- several experimental activities for offline (reconstruction/simulation, analysis) and online (hw and sw triggers, detector readout, monitoring, …)
  - **ATLAS:**
    - ultra-fast CNN for the L0 RPC muon trigger at HL-LHC
    - fast DNN on GPU/FPGA accelerators for the muon High Level Trigger of HL-LHC
    - DNN for forward muon tracking with the NSW detector
    - Graph Neural Networks and geometrical deep learning for tau identification, particle flow, …
    - Applications of DNN for long lived particle identification, cosmic and non collisional backgrounds, flavor tagging, analysis optimization, ...
  - **CMS: applicazioni AI in ambito analisi dati**
  - **FCCee/CepC: Dynamic Graph Neural Network for tau and jet identification and particle flow with the IDEA dual readout calorimeter (S.Giagu, L.Torresi, M.Di Filippo)**
  - **MEG-2:** Drift chamber tracking with Graph Neural Networks
  - **NA62:** Partial Particle ID with the NA62 RICH (use-case for the APEIRON project)
- activities more focused on interpretation/statistic/theory aspects:
  - **The DNNLikelihood:** enhancing likelihood distribution with Deep Learning (L. Silvestrini et al)
- activities related to Quantum Computing applications in HEP:
  - **CERN-INFN-IBM QC_NPHEPGW:** Quantum Machine Learning for Event Classification and Simulation in Nuclear Physics, High-Energy Physics and Gravitational Wave experiments (P.Astone, C.Palomba, S.Giagu, S.Bordoni)

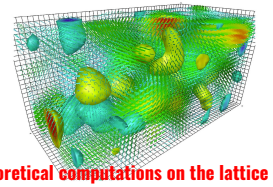
# Iniziativa ML-INFN

## Machine Learning Technologies for INFN

Most of the experiments and initiatives produce, analyse or process digital data.

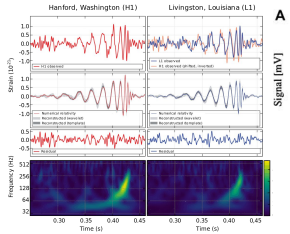
**Enthusiasm on the modern data processing technologies!**



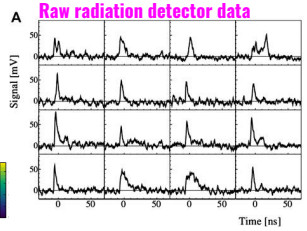



Theoretical computations on the lattice

Gravitational wave detection




Raw radiation detector data





Lucio Anderlini Istituto Nazionale di Fisica Nucleare – Sezione di Firenze May 2022 AI@INFN - Artificial Intelligence at INFN 2

- Disseminazione del ML in ambito INFN
- Infrastruttura on cloud
- Hackaton, tutorial, raccolta e conservazione documentazione
- Roma is in...

## Federated baremetal resources

1× SuperMicro + 1× E4 servers:

- 1 TB RAM
- 64-128 CPU cores
- 36 TB local storage (NVMe)
- 8× **Tesla T4** GPUs
- 5× **RTX 5000** GPUs
- 1× **A30** GPU
- 10 GbE connection to CNAF resources

**Federated to CNAF OpenStack and INFN Cloud**

**Storage solutions**

Storage from CERN experiments can be mounted with NFS from the Tier-1 storage

Hypervisors integrated to Ceph to manage persistent virtual volumes accessed from the VM with POSIX

Lucio Anderlini Istituto Nazionale di Fisica Nucleare – Sezione di Firenze May 2022 AI@INFN - Artificial Intelligence at INFN 8

## Lowering The potential barriers with ML_INFN

Employing machine learning techniques often requires:

- specialized hardware and software setup

WP1: provide a centrally maintained cloud-based infrastructure for interactive and batch ML fast prototyping, with access to modern GPU hardware and systems tuned for ML performance

- specific training to identify tools and learning resources

WP2: organize national training events for INFN users (Machine Learning hackathons)

- a community of experts providing support to research use cases

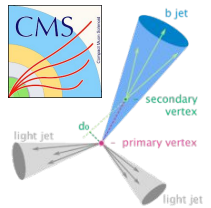
WP3: provide and organize example applications in a knowledge base

Lucio Anderlini Istituto Nazionale di Fisica Nucleare – Sezione di Firenze May 2022 AI@INFN - Artificial Intelligence at INFN 4

## Hackathon use cases: *10 groups, one tutor per group*

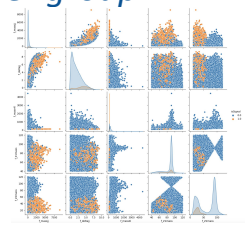
**Jet b-tagging at CMS**

Recurrent Neural Networks with LSTM



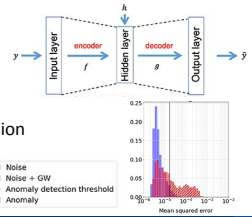

**Higgs searches at CMS**

Deep Neural Networks and Advanced Keras



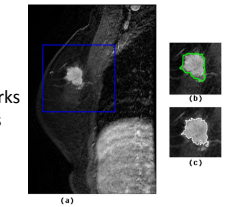
**Gravitational Waves with Virgo**

Autoencoders, anomaly detection and compression




**Segmentation of CT scans**

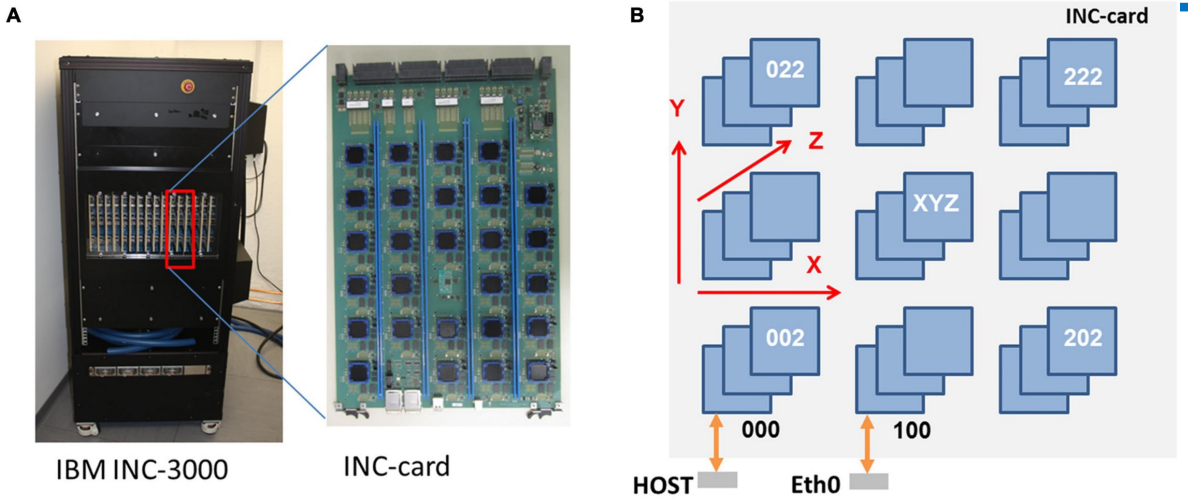
Convolutional Neural Networks Handling 2D and 3D datasets



Lucio Anderlini Istituto Nazionale di Fisica Nucleare – Sezione di Firenze May 2022 AI@INFN - Artificial Intelligence at INFN 12

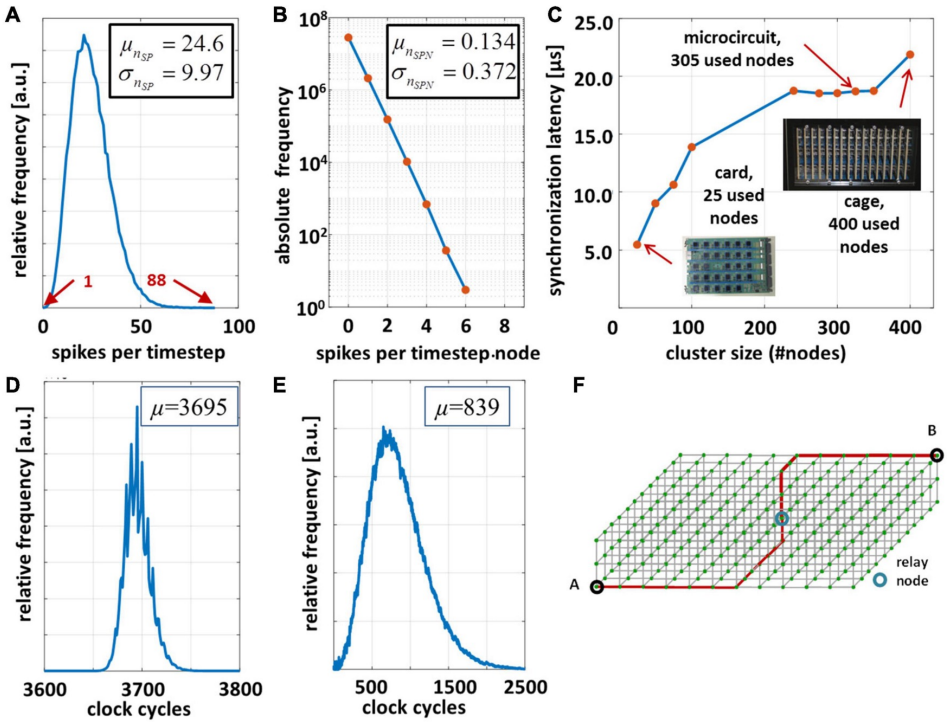
L. Anderlini (WS AI INFN 2022)
https://agenda.infn.it/event/29907/contributions/163440/attachments/90261/121581/ML_INFN%20initiative%20-%20AI%20Workshop%20Bologna%20%281%29.pdf

# Spiking Neural Network on FPGA: IBM INC-300



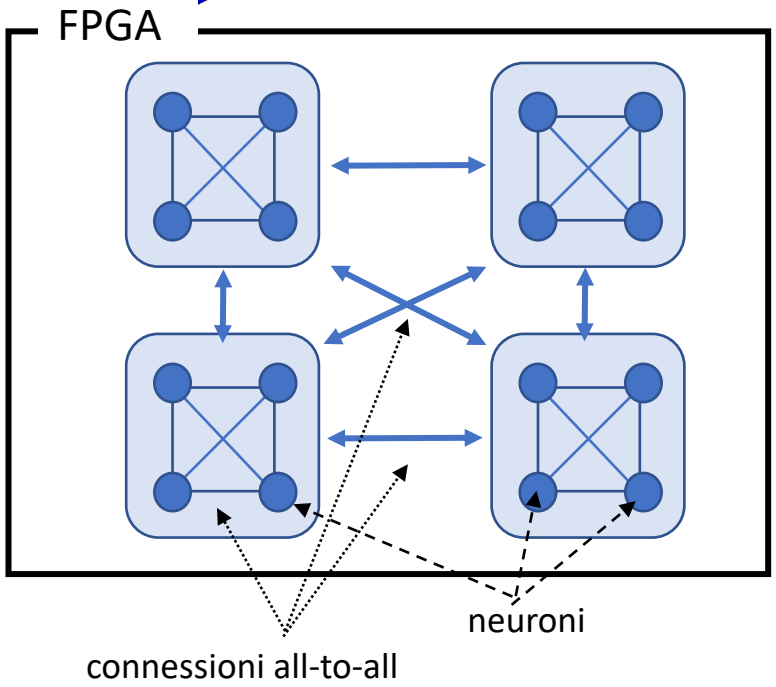

IBM Neural Supercomputer (Narayanan et al., 2020) (INC300)

- basato su FPGA Xilinx SoC (Zynq)
- struttura 3D di interconnessione tra i vari moduli con partizione di circa 30 moduli per board e 20 moduli per rack
  - Nets use case: 305 CN ognuno con circa 256 neuroni con 4k syn per neurone
- Performance: 3700 cicli @150 MHz per 100uS di update fisico → 4,5x real time biologico




- https://www.frontiersin.org/articles/10.3389/fnins.2021.728460/full#B4

# ML Spiking Neural Network-based su FPGA (credit M. Martinelli – C. De Luca)



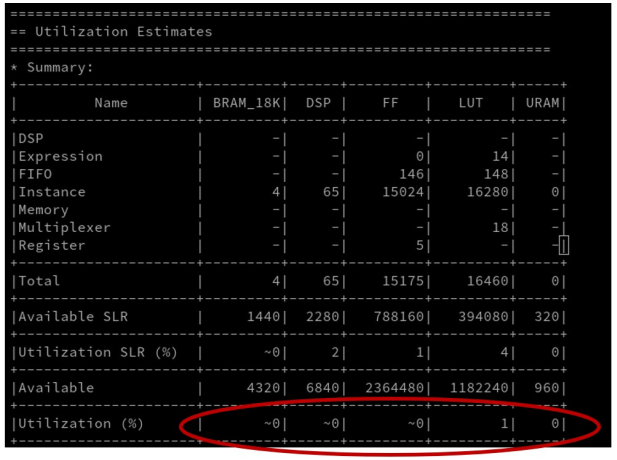

Modello: rete ricorrente **spiking** allenata per riprodurre pattern spazio-temporali complessi [1]

Ipotesi di architettura:

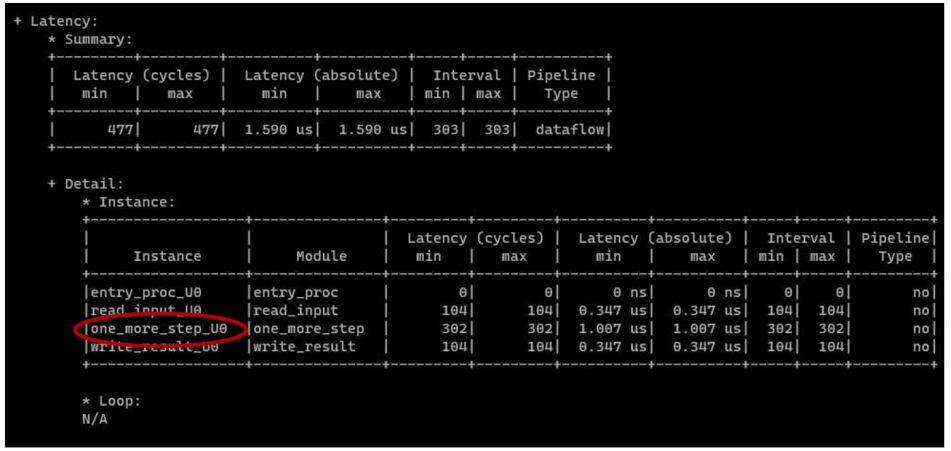
- diversi neuroni connessi **all-to-all** divisi in "blocchetti" (testati fino a 70 neuroni)
- Codice sviluppato in HLS su singola FPGA (Xilinx Alveo U200)
- Connessione HOST-DEVICE basata su BUS PCI (Core PCI Xilinx XDMA)

Ipotesi di **scalabilità (next steps)**:

- diversi "blocchetti" istanziati su singola FPGA
- diverse FPGA connesse insieme
- sviluppo dell'infrastruttura di collegamento intra-FPGA e inter-FPGA (e.g. astrazione con dispatcher aggregator per semplificare la scalabilità)



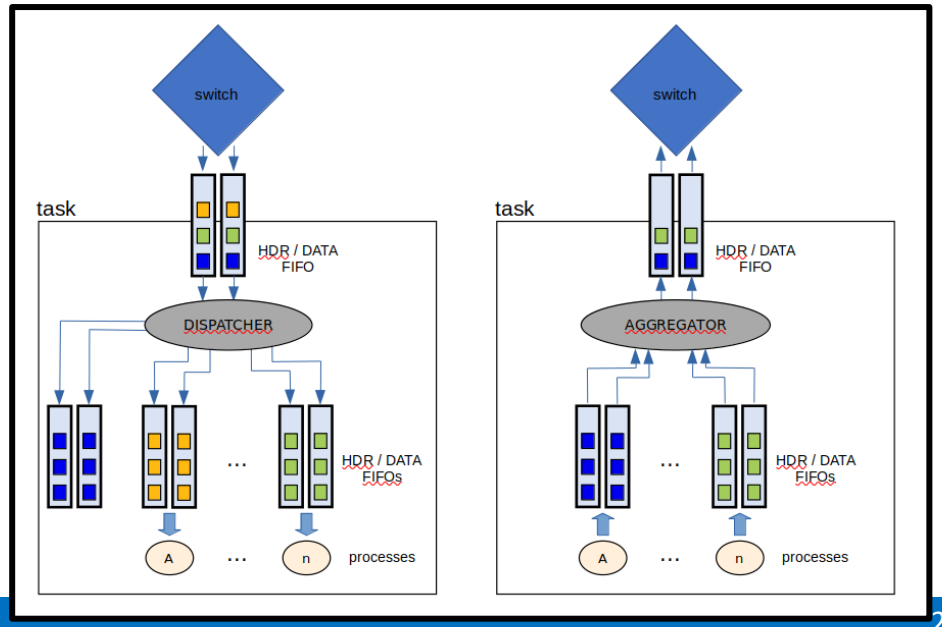
```
== Utilization Estimates
* Summary:
+---------------------+---------+------+---------+---------+-----+
|         Name        | BRAM_18K| DSP  |    FF   |   LUT   | URAM|
+---------------------+---------+------+---------+---------+-----+
|DSP                  |        -|     -|        -|        -|    -|
|Expression           |        -|     -|        0|       14|    -|
|FIFO                 |        -|     -|      146|      148|    -|
|Instance             |        4|    65|    15024|    16280|    0|
|Memory               |        -|     -|        -|        -|    -|
|Multiplexer          |        -|     -|        -|       18|    -|
|Register             |        -|     -|        5|        -|    -|
+---------------------+---------+------+---------+---------+-----+
|Total                |        4|    65|    15175|    16460|    0|
+---------------------+---------+------+---------+---------+-----+
|Available SLR        |     1440|  2280|   788160|   394080|  320|
+---------------------+---------+------+---------+---------+-----+
|Utilization SLR (%)  |       ~0|     2|        1|        4|    0|
+---------------------+---------+------+---------+---------+-----+
|Available            |     4320|  6840|  2364480|  1182240|  960|
+---------------------+---------+------+---------+---------+-----+
|Utilization (%)      |       ~0|    ~0|       ~0|        1|    0|
+---------------------+---------+------+---------+---------+-----+
```



```
+ Latency:
    * Summary:
    +---------+---------+----------+----------+-----+-----+----------+
    | Latency (cycles) | Latency (absolute) | Interval | Pipeline |
    |   min   |   max   |    min   |    max   | min | max |   Type   |
    +---------+---------+----------+----------+-----+-----+----------+
    |      477|      477| 1.590 us | 1.590 us |  303|  303| dataflow |
    +---------+---------+----------+----------+-----+-----+----------+

    + Detail:
        * Instance:
        +-----------------+---------------+---------+---------+----------+----------+-----+-----+----------+
        |                 |               | Latency (cycles) | Latency (absolute) | Interval | Pipeline |
        |     Instance    |     Module    |   min   |   max   |    min   |    max   | min | max |   Type   |
        +-----------------+---------------+---------+---------+----------+----------+-----+-----+----------+
        |entry_proc_U0    |entry_proc     |        0|        0|     0 ns |     0 ns |    0|    0|        no|
        |read_input_U0    |read_input     |      104|      104| 0.347 us | 0.347 us |  104|  104|        no|
        |one_more_step_U0 |one_more_step  |      302|      302| 1.007 us | 1.007 us |  302|  302|        no|
        |write_result_U0  |write_result   |      104|      104| 0.347 us | 0.347 us |  104|  104|        no|
        +-----------------+---------------+---------+---------+----------+----------+-----+-----+----------+

        * Loop:
        N/A
```




[1] Muratore et al 2022, https://doi.org/10.1371/journal.pone.0247014

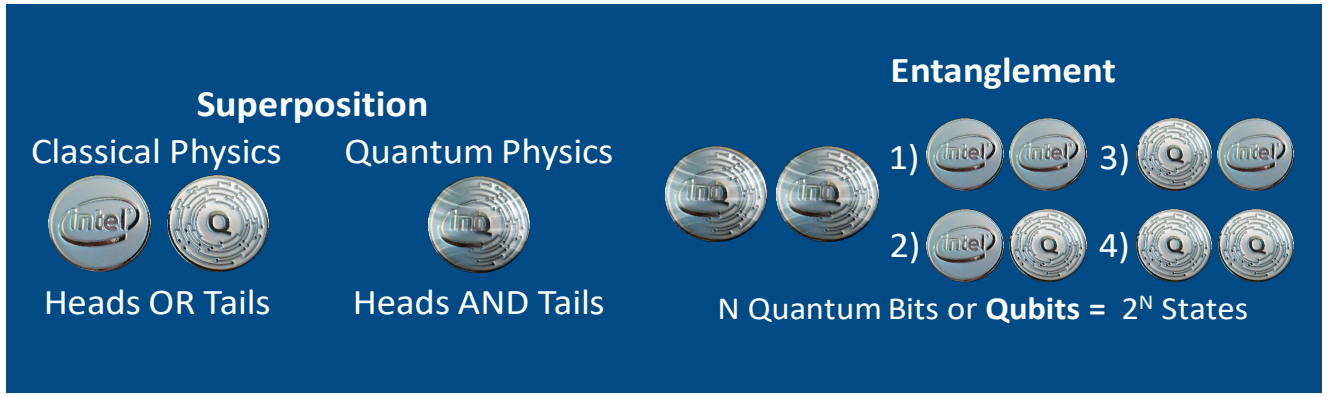
# Quantum Computing for dummies....

- Quantum computer (QC): una macchina che esegue calcoli sfruttando le leggi della meccanica quantistica (Principio di sovrapposizione e entanglement quantistico)
- Sulle spalle dei giganti…
  - Feynman (1982) prima proposta di un calcolatore basato sulla meccanica quantistica
  - Deutsch (1985) universalita' dei circuiti quantistici (Quantum Turing Machine)
  - Primi algoritmi quantistici per risoluzione di problemi  Shor (1994) fattorizzazione di grandi numeri in tempo polinomiale e Grover (1997): quantum search
- Qbit (elemento minimo di informazione) rappresentabile con la sovrapposizione di due stati **|0>** and **|1>**.

  → $|\psi\rangle = \alpha_1|0\rangle + \alpha_2|1\rangle$ dove $|\alpha_1|^2 + |\alpha_2|^2 = 1$
- Un possibile stato di un *registro a 3 qbit*

  → $|\psi\rangle = 1/\sqrt{8}\ |000\rangle + \ldots + 1/\sqrt{8}\ |111\rangle$ ed in generale un registro a N qubit codifica $2^N$ stati "simultanemente".

Applicando la misura la sovrapposizione di stati collassa in uno degli stati classici in funzione dei valori delle ampiezza di probabilita'.




How to use a Quantum Computer

Applications of Quantum Computers

**Conventional Computer:**

0000
0001
0002
0003
0004
0005
…

Max. 10.000 Steps

SuperMUC-NG:
26.900.000.000.000.000 Flop/s

**Quantum Computer:**

14 Qubits
$2^{14}$ = 16385 States

➔ 1 Step

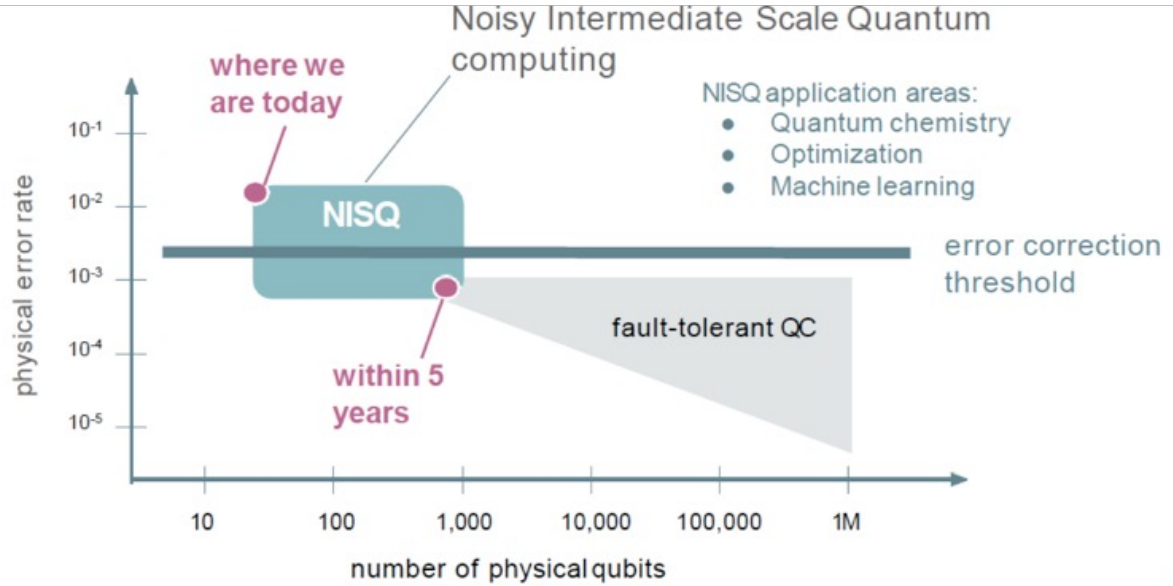
Speedup: 10.000



https://www.amazon.de/dp/B075392B4T/ref=pe_3044161_189395811_TE_SCE_dp_1

# Quantum Computing for dummies....

- QC Analogici: controllo collettivo dei qbits
  - HW: quantum simulator e quantum annealer (D-Wave)
  - SW: evoluzione in tempo naturale del sistema quantistico preparato in uno stato noto

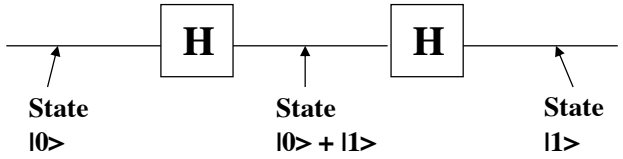



"Quantum computing in the NISQ era and beyond" Preskill, 2018 https://arxiv.org/abs/1801.00862

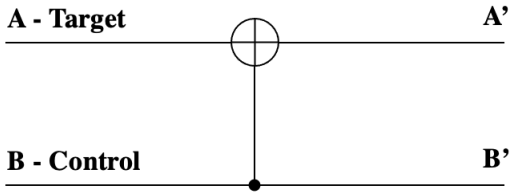
- QC digitali: controllo individuale del singolo qbit
  - HW: sistema basato su gates quantistici (IBM-Q, Google, AQT, ...)
  - SW: sequenze di 1-qbit o 2-qbit gates ordinate a formare circuiti quantistici

# Quantum Computing for dummies....

- Quantum gates: simili ai gates classici ma devono essere reversibili
  - lo stato di input deve poter essere derivato dallo stato di output
- Hadamard gate per costruire la sovrapposizione di stati

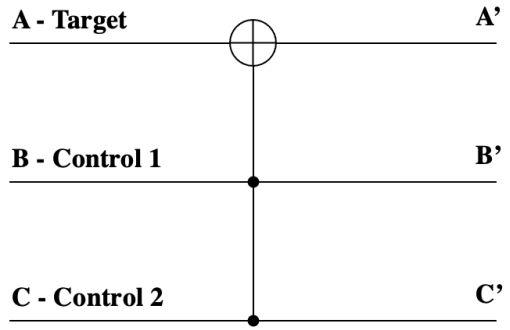



- Controlled NOT (C-Not): come lo Xor ma con bit ancillare per reversibilita'




| Input | | Output | |
|---|---|---|---|
| A | B | A' | B' |
| 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 1 |

- Controlled-controlled NOT (C–C-Not): operatore a tre qbits **universale**
  - Se i due controls sono 1 → A' = NOT(A)
  - Se A = 1 → A' = not(B*C) NAND

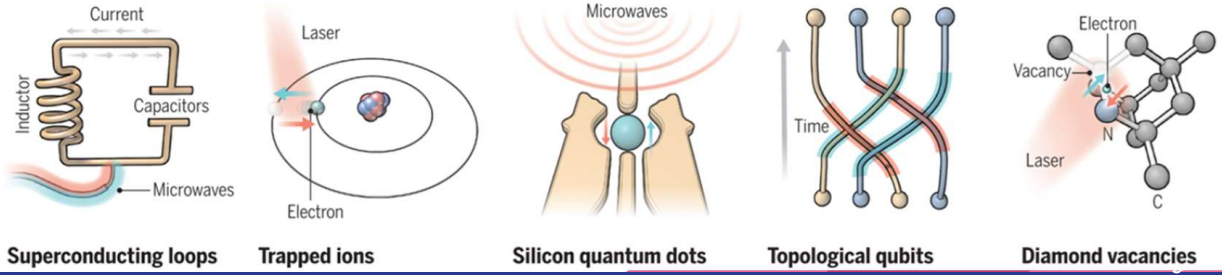



| Input | | | Output | | |
|---|---|---|---|---|---|
| A | B | C | A' | B' | C' |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 1 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 | 1 | 1 |

| Input | | | Output | | |
|---|---|---|---|---|---|
| A | B | C | A' | B' | C' |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 1 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 | 1 | 1 |

# Tecnologia del Qbit

- E' difficile fare un'overview esaustiva, molte tecnologie a differente grado di maturita' caratteristiche d'integrazione scalabilita' etc..
  - Qubits stato solido superconduttivi (con giunzioni Josephson)
  - Trapped ION (transizioni atomiche controllate da campi elettromagnetici)
  - Atomi neutri in optical lattice
  - Quantum dots
  - Nitrogen-Vacancy (N-V) in diamante
  - Qubit topologici
  - Fotoni




Non-exhaustive technology examples from: Popkins, 2016, Science, DOI: 10.1126/science.354.6316.1090

- Ad oggi loop superconduttivi, trapped ions and Rydberg atoms sembrano essere le tecnologie + mature
- Difficile fare previsioni sulla scalabilita': sistemi ancora troppo piccoli (decine di qbits) e nessuno fault tolerant

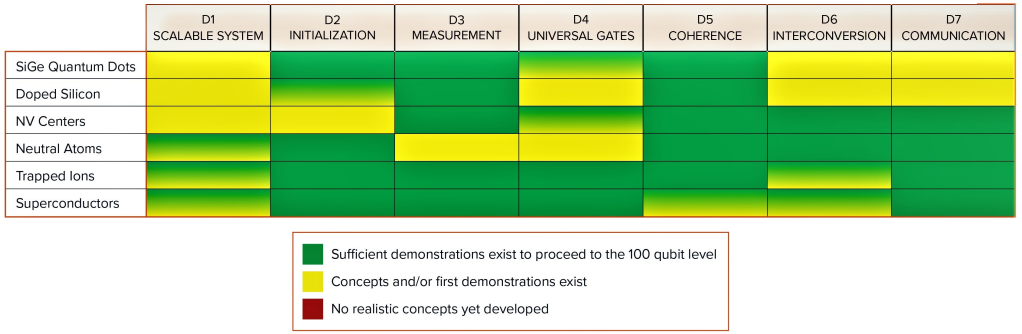
*Table CEQIP-22 Gate-Based Quantum Computing Status Summary*

| *Qubit type* | *Quantum volume [627]* | *Qubit count* | *Qubit connectivity* | *2-qubit gate depth* | *Quantum teleportation* | *Qubit function* | *System scalability* |
|---|---|---|---|---|---|---|---|
| Superconducting | 64 | 53 | 3.25 | 667 | yes | fair | fair |
| Trapped ion | 512 | 11 | 10 | > 100 000 | yes | fair | fair |
| Quantum dot | – | 4 | 1 | 104 | – | poor–fair | fair–good |
| Photonic | – | 4 | | | yes | poor | fair |

2-qubit gate depth: ratio of coherence time divided by 2-qubit gate time ($T_2^*/t_{2q}$)

DIVINCENZO CRITERIA

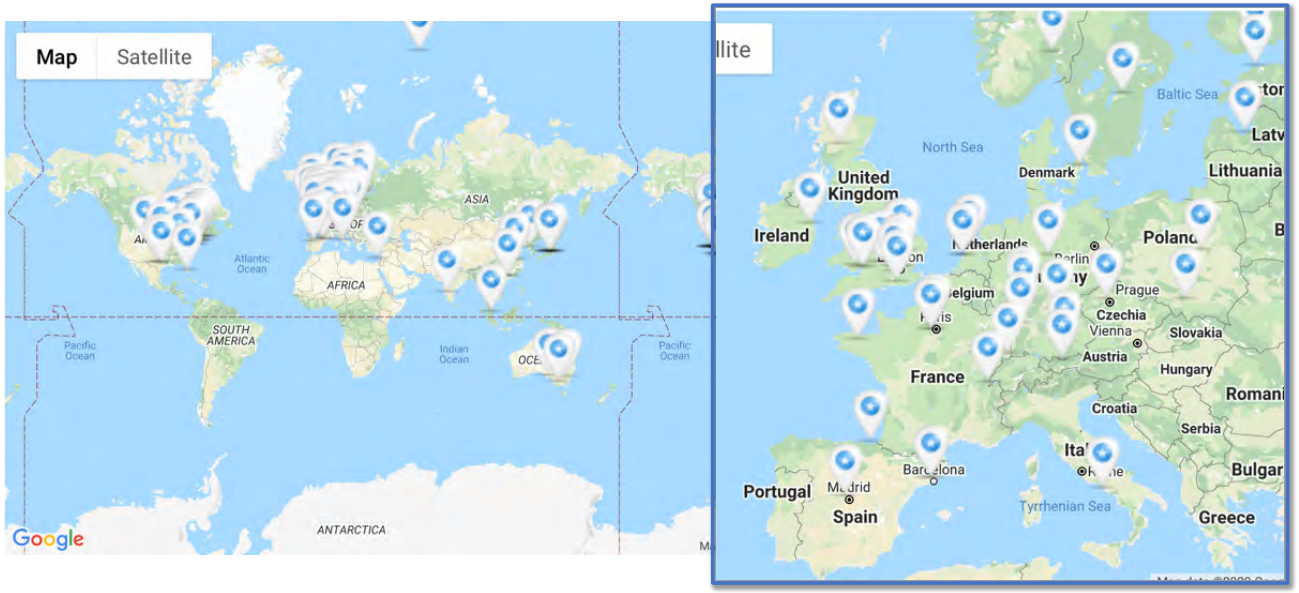
| REQUIREMENTS FOR THE PHYSICAL IMPLEMENTATION OF QUANTUM COMPUTATION | |
|---|---|
| D1: Scalable qubits | Scalable physical system of well-defined, characterized qubits |
| D2: Initialization | Prepare a simple, fiducial input state |
| D3: Measurement | Measure the qubit state |
| D4: Universal gate set | Perform a universal set of gate operations with high fidelity |
| D5: Coherence | Robustly represent quantum information (long coherence times) |
| REQUIREMENTS FOR ROUTING QUANTUM INFORMATION | |
| D6: Interconversion | Ability to interconvert stationary and flying qubits |
| D7: Communication | Ability to transmit flying qubits faithfully between two locations |

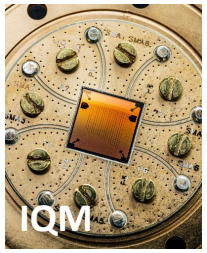


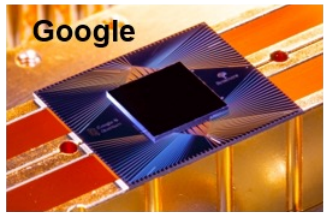

- Alcuni challenges tecnologici futuri:
  - Migliorare il qbit:
    - setup veloce, tempi (molto) lunghi di coerenza, ridurre la sensibilita' all'ambiente (noise tolerant)
  - Migliorare error correction
    - Richiede numero grande di qbits → aumento sostanziale del tempo di coerenza collettivo, etc
  - Meccanismi per incrementare la scalabilita' a grandi numeri
  - "Easy to use"
    - denso, funzionamento a $T_{ambiente}$ , form factor tipico dell'elettronica, low(er) cost,…

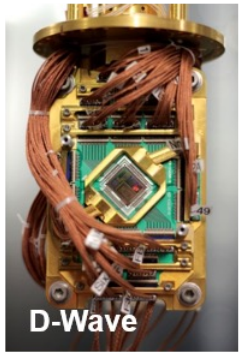
# QC providers

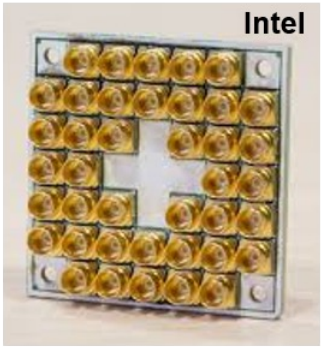
- Una pletora di iniziative industriali finalizzate alla costruzione di QC
  - IBM Q
  - INTEL
  - Google
  - D-Wave
- Ma anche un'esplosione di start-up QC-oriented
  - Hardware e controllo, programming, applicazioni,...

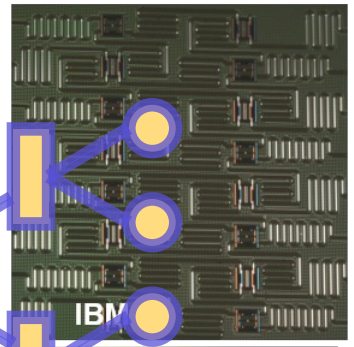


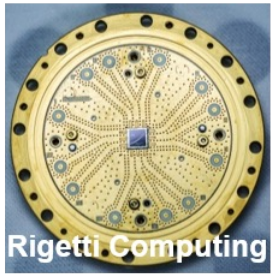



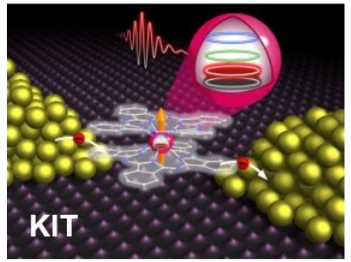



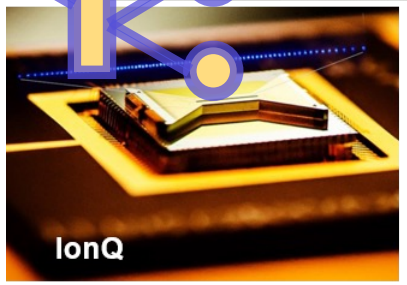



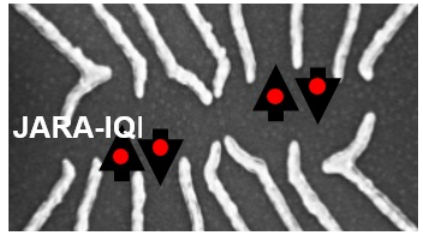



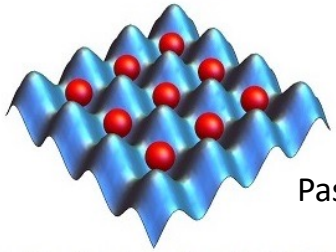



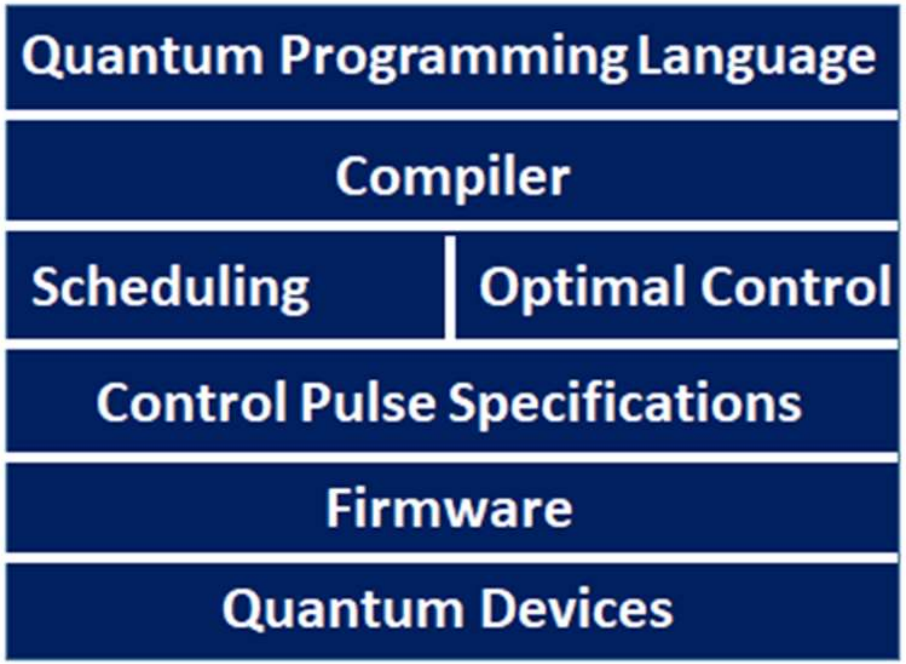



Pasqal

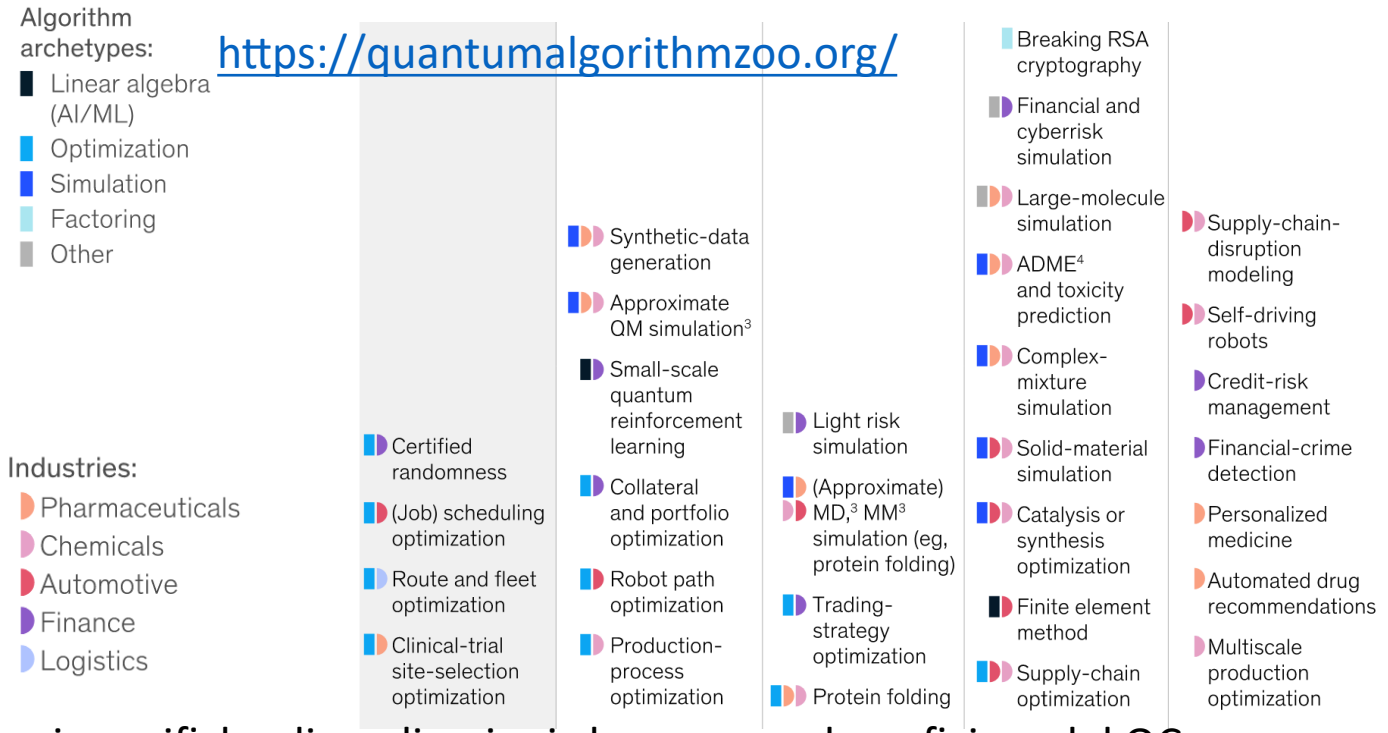
https://quantumzeitgeist.com/interactive-map-of-quantum-computing-companies-from-around-the-globe/

**Quantum computers and simulators (QCS)**

# QC software stack e applicazioni




- Stack softwares sono proprietari dei fornitori di QC
  - C'e molto margine di miglioramento
  - Dovrebbe convergere verso uno standard




Classi specifiche di applicazioni che possono beneficiare del QC

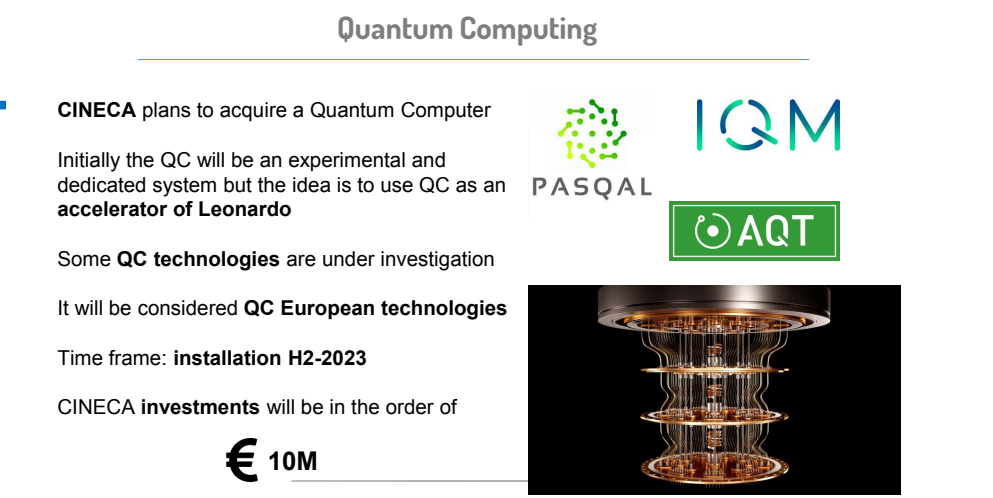
- *Hybrid Quantum Algorithms*
  - Sezione classica e accelerazione quantistica
  - Specifici problem di ottimizzazione Es: Variational Quantum Eigensolver (VQE) o Quantum Approximate Optimisation Algorithm (QAOA)
  - Quantum chemistry
  - Problemi di ottimizzazione esprimibili con una funzione di costo che deve essere minimizzata
  - Quantum Machine Learning

# QC: possibile contributo INFN

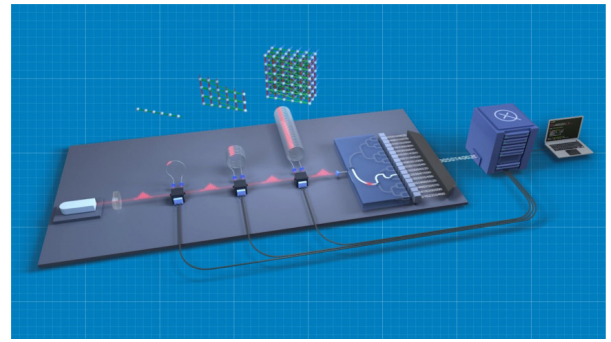
- Ha senso per INFN puntare sull'hardware dei QC? Probabilmente no...
  - Molti lo stanno gia' facendo e il gap tecnologico e' difficilmente recuperabile
- INFN puo' dare un contributo fondamentale alle tecnologie di base
  - criogenia, misure di precisione in ambiente "noisy", controllo fine di sistemi distribuiti, etc…
  - Call tematiche CSN5
- Contribuire alla generazione di nuovi algoritmi e nuovi campi di applicazione
  - HEP (almeno) e' un campo applicativo di grande interesse per QC (ottimizzazione, ML, LQCD, many body…) e viceversa
    - CERN openlab QTI,
    - TNM per LQCD (Montangero),
    - Superconducting Quantum Materials and Systems Center (SQMS, INFN@Fermilab)...
    - Iniziativa Quantum della CSN4 (https://web.infn.it/CSN4/IS/Linea4/QUANTUM/QUANTUM.html)
- Contribuire all'interfaccia tra QC e outer world
  - Interfaccia con HPC per MSA: controllo e I/O di Sistema basato su FPGA
  - Network scalabile per clustering di sistemi QC di piccolo taglia
- Ad oggi non e' chiaro (i.e. non esiste) un percorso di sviluppo tecnologico che garantisca la scalabilita' di un QC ai $10^6$ qbits
  - Mantenere il presidio tecnologico cercando di operare su tutte le piattaforme disponibili per essere pronti ad utilizzare efficentemente IL QC...

# Accesso a sistemi QC

- QC in cloud: IBM-Q o Google
  - standard access e/o accesso prioritario a IBM-Q con accordo CERN-INFN
- Cineca: acquisizione di un sistema medio/piccolo trapped ions (a disosizione del PNRR CN spoke quantum?)
- QCSC Padova (Quantum Computing and Simulation Center) nuova iniziativa con cofin INFN
  - Scuole, formazione e accesso a testbed QC commerciali (1 anno)
  - Acquisizione di un sistema a piccola/media scala (O(10-50) qbits) (2 anno →) per ricerca tecnologica e  algoritmica e produzione
    - https://800anniunipd.it/event/padova-e-la-nuova-sfida-del-computer-quantistico/
- Accordo Amazon INFN per uso computer BOREALIS XANADU in cloud (Amazon Bracket o Xanadu Cloud)
  - fault tolerant processore fotonico programmabile a temperatura ambiente...
    - Madsen, L.S., Laudenbach, F., Askarani, M.F. *et al.* Quantum computational advantage with a programmable photonic processor. *Nature* **606,** 75–81 (2022).
    - https://www.xanadu.ai/products/borealis
  - Quantum Computing advantage...
- Spoke 10 del nuovo ISCS (PNRR)
  - https://agenda.infn.it/event/30202/contributions/168463/attachments/91484/124226/Bozzi_20220526_QC_CCR.pdf







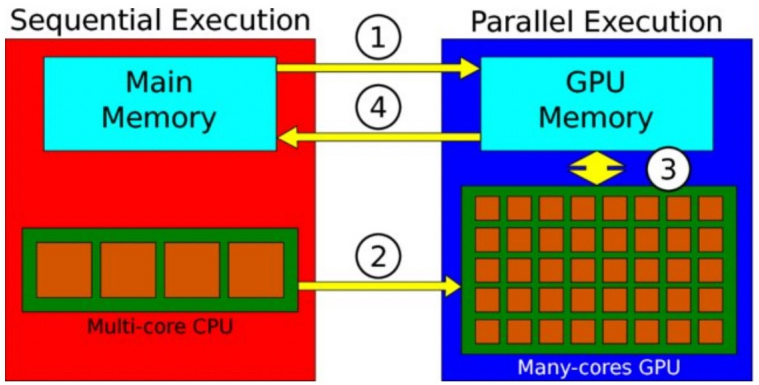
# Considerazioni finali

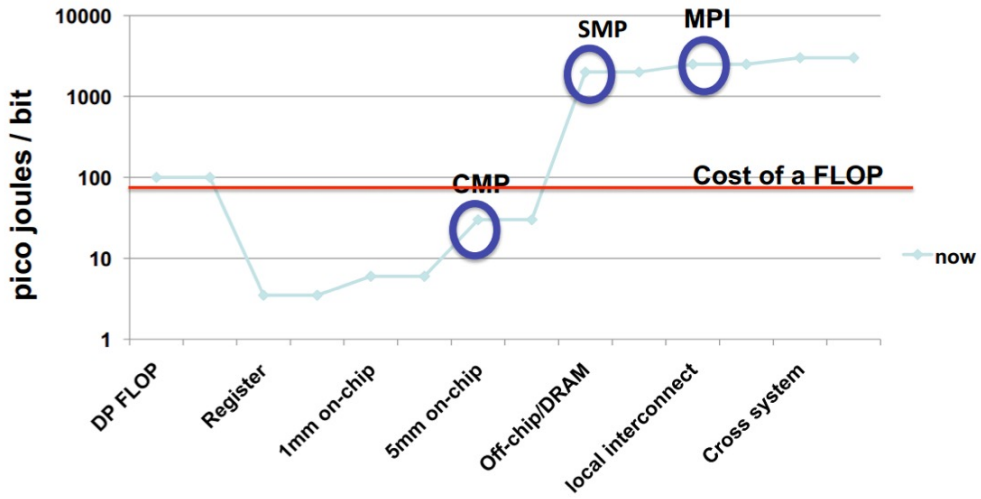
- In generale ricerca tecnologica e' (anche) fondazionale per la ricerca di base sperimentale oltre a rappresentare un'opportunita' di visibilita', innovazione e finanziamento
- Le risorse umane sono il vero valore aggiunto
  - Non e' sufficente essere *fisici puri* e avere competenze di analisi dati per costruire Esperimenti, Acceleratori, Detector
  - esperienza trasversale in tecnologie di «basso» livello (elettronica, informatica, software di sistema e applicativo, meccanica, etcc) e capacita' di lavorare all'interfaccia tra la fisica e la tecnologia
- Non sono (piu') competenze a cui abbiamo accesso facilmente e in abbondanza o che possiamo comprare sul mercato
  - Bisogna prevedere occasioni di formazione e promozione specifiche all'interno dei nostri corsi di laurea
  - Dobbiamo tornare ad essere competitivi con l'industria valorizzando le persone coinvolte
- INFN deve tornare ad essere un porto sicuro per chi vuole (continuare a) fare ricerca tecnologica. Il Management dovrebbe:
  - riconoscere la fondamentale importanza di queste attivita' per le strategie dell'Ente
  - studiare (collegialmente) ed implementare meccanismi correttivi prima che sia troppo tardi...
  - migliorare il processo di comunicazione
  - governarle con visione a lungo termine evitando la gestione un po' approssimativa delle varie opportunita' a breve termine (GRID, PNRR...)

# BACKUP SLIDES

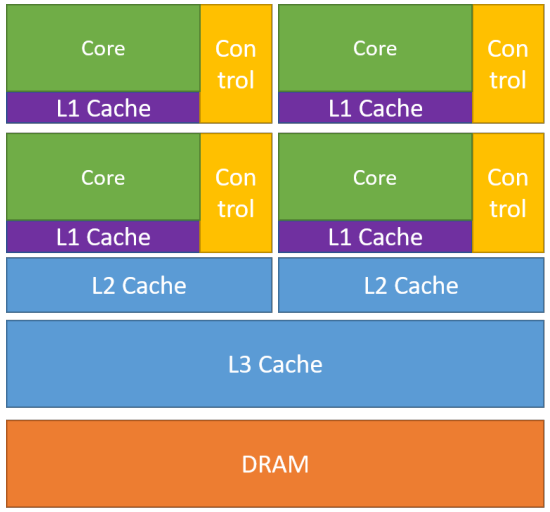
# Intro: problemi aperti

- Alcuni problemi aperti per le attuali tecnologie del calcolo
  - come massimizzare l'efficenza computazionale
    - non esiste una architettura di calcolo "one fits for all"
    - ogni applicazione o classe di applicazioni ha specificita' differenti (fp vs int, compute bounded vs memory bounded...)
    - Il corrente modello eterogeneo e accelerato "CPU+Acceleratore" (FPGA/GPU/???) e' sufficentemente scalabile e prospetticamente sostenibile?
  - power: minimizzare il rapporto W/operazioni e incrementare la densita' del sistema
    - vale a tutte le scale dal supercomputer dato il cap di potenza per sito (N*10MW) fino al sistema embedded alimetato a batteria
  - architettura di memoria ad accesso a bassa latenza ma larga abbastanza e network a bassa latenza ed alto throughput...
  - per non parlare di:
    - software e modelii di programmazione scalabili
    - resilienza e fault tolerance
    - sostenibilita' economica
    - ...

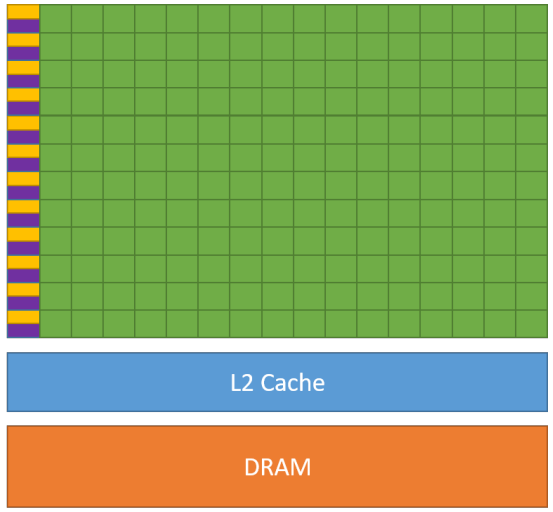


# CPU vs GPU

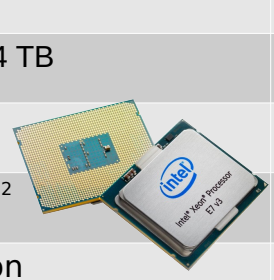


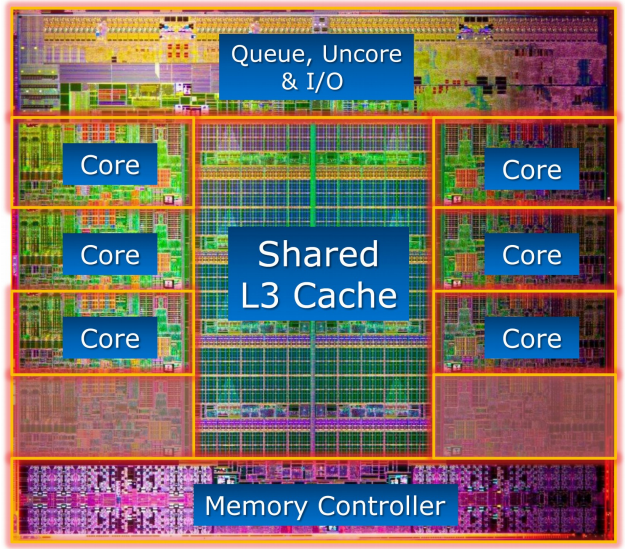

CPU



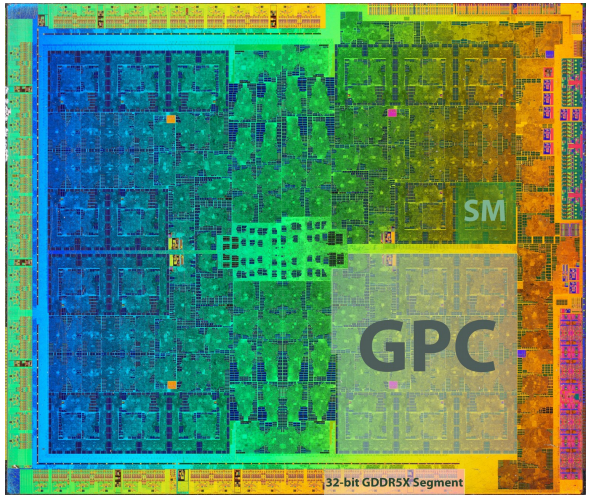

GPU

| | Intel Core E7-8890 v3 | GeForce GTX 1080 |
|---|---|---|
| Core count | 18 cores / 36 threads | 20 SMs / 2560 cores |
| Frequency | 2.5 GHz | 1.6 GHz |
| Peak Compute Performance | 1.8 GFLOPs | 8873 GFLOPs |
| Memory bandwidth | Max. 102 GB/s | 320 GB/s |
| Memory capacity | Max. 1.54 TB | 8 GB |
| Technology | 22 nm | 16 nm |
| Die size | 662 mm$^2$ | 314 mm$^2$ |
| Transistor count | 5.6 billion | 7.2 billion |
| Model | Minimize latency | Hide latency through parallelism |



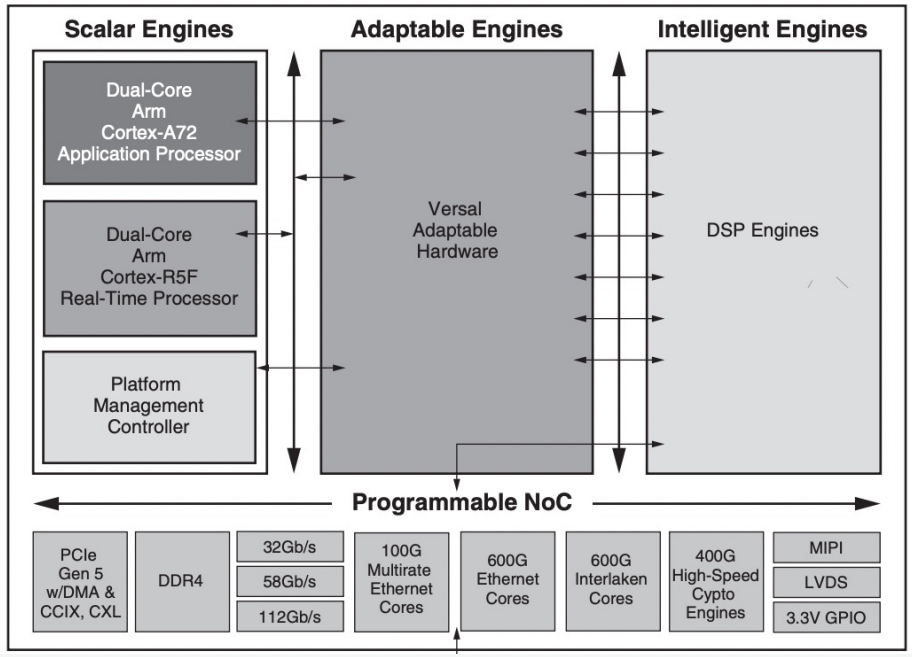


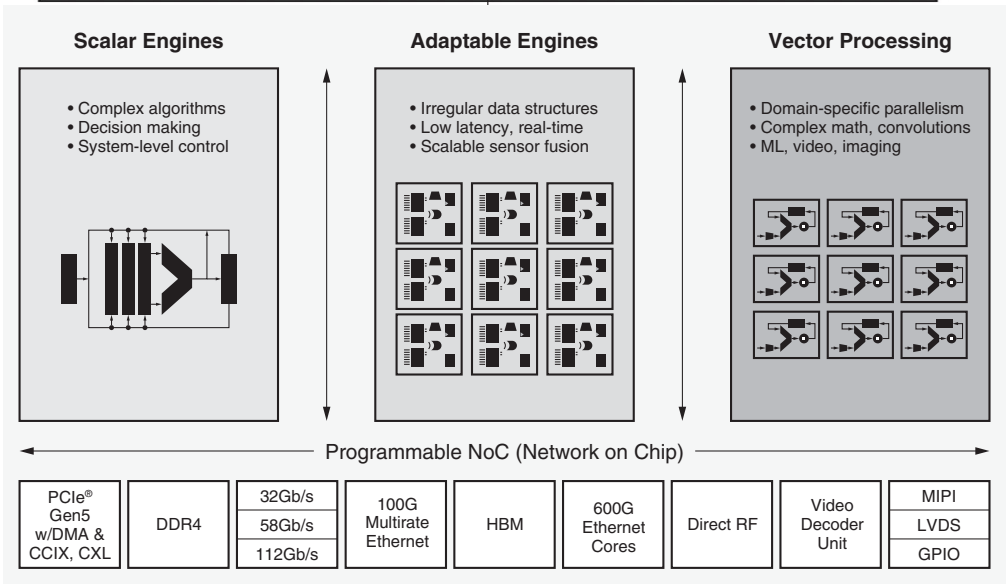

http://images.anandtech.com/reviews/cpu/intel/SNBE/Core_I7_LGA_2011_Die.jpg




http://wccftech.com/nvidia-gtx-1080-gp104-die-shot/

# Next Gen FPGA




Nuova tendenza per gestire al meglio la complessità

- Da libreria di core IP e gates programmabili, strutture di I/O specializzate interconnesse da una matrice di connessioni → un sistema di uP e IP hardware indipendenti interconnessi da una network on chip (NoC) evoluta ad alte prestazioni
  - Non si programma la matrice di interconnessione ma si costruisce il sistema assemblando le IP (lego mode...)
- Strutture differenti (Scalar Engines, Adaptable Engines, Intelligent Engines) per implementazione efficente di blocchi computazionali specializzati per task computazionali diversi (eterogenei)
- il tutto a 7nm di processo...
- Core market per questo tipo di FPGA:
  - networks (software defined network)
  - ML inference (per efficenza e specializzazione) con uso di AI Tensor Block specializzati
  - Accelerazione computazionale per data analytics e calcolo scientifico
- Problema:
  - una struttura eterogenea complessa e configurabile va programmata al meglio
  - la complessita' hardware/software/sistemistica deve essere astratta per permettere all'utente (medio) di poterla sfruttare in maniera efficente

# Finanziamenti EU

## The EuroHPC Joint Undertaking today

**Mission:** Establish an integrated world-class supercomputing & data infrastructure and support a highly competitive and innovative HPC and Big Data ecosystem

https://eurohpc-ju.europa.eu/

**The Joint Undertaking is located in Luxembourg**

- **32 Participating States (26 MS + 6 associated countries) + EU + 2 Private Members (ETP4HPC & BDVA)**
- **Launched in October 2018**
- **Budget from this MFF (2019-2020): ~1.5 B€**
  - 536 M€ from EU
  - Matching funds from MS for JU activities
  - In-kind contributions from the Private Members


## EuroHPC JU: Investments in 2019-2020

**World-class supercomputing Infrastructure: 830 M€ investments**

***3 Pre-exascale supercomputers (150 - 370 Petaflops) – Top 5***

- *Installation in 2021 in Kajaani (FI), Bologna (IT), Barcelona (ES)*
- *EuroHPC JU is the owner, co-ownership with 17 Participating States (PS)*
- ***Investment: ~€650 million, 50% EU - 50% PS***

***5 Petascale supercomputers (6-12 petaflops) – Top 50***

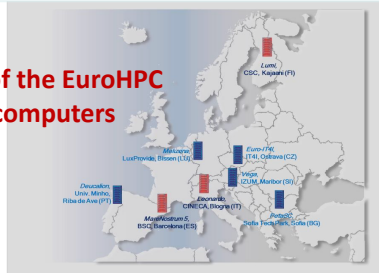
One tender is still work in progress

- *Under approval: installation end 2020 – early 2021 in Bissen (LU), Ostrava (CZ), Maribor (SI), Guimarães (PT), Sofia (BG)*
- *MS are the owners; EuroHPC JU owns 35%*
- ***Investment: ~€180 million, 35% EU - 65% PS***

**HPC ecosystem: 360 M€ R&I investments**

- Building a European processor technology (for HPC) and exascale pilot systems
- 32 HPC competence centres promoting innovation and training activities in all PS
- Innovative industrial applications
- Actions for SMEs

**Map of the EuroHPC supercomputers**




## Proposal for EuroHPC JU – New mission – New budget – New duration

***Mission***

*By 2027: develop, deploy, extend and maintain in the Union* ***a world leading federated, secure and hyper-connected supercomputing, quantum computing, service and data infrastructure ecosystem****; support the production of innovative and competitive* ***supercomputing systems based on a supply chain that will ensure components, technologies and knowledge*** *limiting the risk of disruptions and the development of a* ***wide range of applications optimized for these systems****; and,* ***widen the use*** *of this supercomputing infrastructure to a large number of public and private users, and support the development of* ***key skills for European science and industry****.*

*Expected Budget (EU, Participating States, Private Members):*

*EUR 8 billion [not in the Regulation], for the period 2021-2033*

## Budget of the EuroHPC JU 2021-2027

**+/-EUR 8 billion**

***EU contribution***

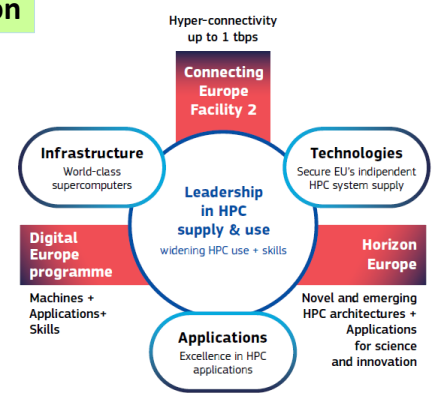
- *Horizon Europe: TBD*
- *Digital Europe Programme: up to [2.4 B€]*
- *Connecting Europe-2 Facility : up to [200 M€]*

***Contributions from the PS***

- *At least equal to the Union contribution*

***Contribution from the Private Members***

- *For R&I, mainly in kind: TBD*




***Administrative Costs***

- *Share of total costs: EU: 45%, PS: 45%, Private Partners: 10% (in cash)*