Le prospettive per il Calcolo

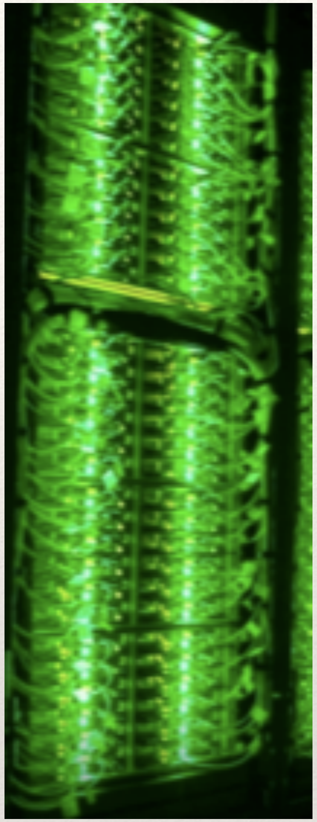
D. Bonacorsi (University of Bologna)

Roma - 6 Settembre 2018

Credits: all the huge work by software/computing experts in all INFN CSNs, C3S members, computing Tiers admin and service experts.
Special thanks on input in the slides to A. Zoccoli, D. Lucchesi, G. Maron, T. Boccali, C. Grandi, P. Vicini, D. Menasce

Outline / Executive Summary

Four main pillars.



Multi-architecture computing

Storage

(and **Network**)

Training

R&D

(medium- and long-term)

Setting the scene

Scientific computing has been <u>enabling the HEP program</u>

- until today, software and computing have <u>NOT</u> been a limiting factor for Physics!

For this to continue to be true, we need to face the challenges ahead of us:

- The *usual* challenge: computing remains a significant cost driver
  - → measure-optimise-measure cycles, Computing models undergo “adiabatic” (more or less) evolutions
- The *new* challenge(s): ramp-up in global resource needs in the next decade(s)
  - e.g. HL-LHC, theory, astro-particle - in addition, new experiments
  - → further/deeper optimisations, evaluation and adoption of new (even “disruptive”) paradigms

Uncertainties in quantitative definition of needs, and specification of computing environments

Scale of the HEP challenges

Disclaimer: not a complete list, and only experimental physics here

HL-LHC

- 10x trigger rate, 6x event complexity, plus detector complexity: >60x resources needs (main concern is disk)

SKA

- aims at collecting ~300 PB/yr
- major challenge on software, computing, data movement

LSST

- aims at collecting ~50 PB/yr
- same as for SKA

VIRGO-LIGO, multi-messenger astronomy

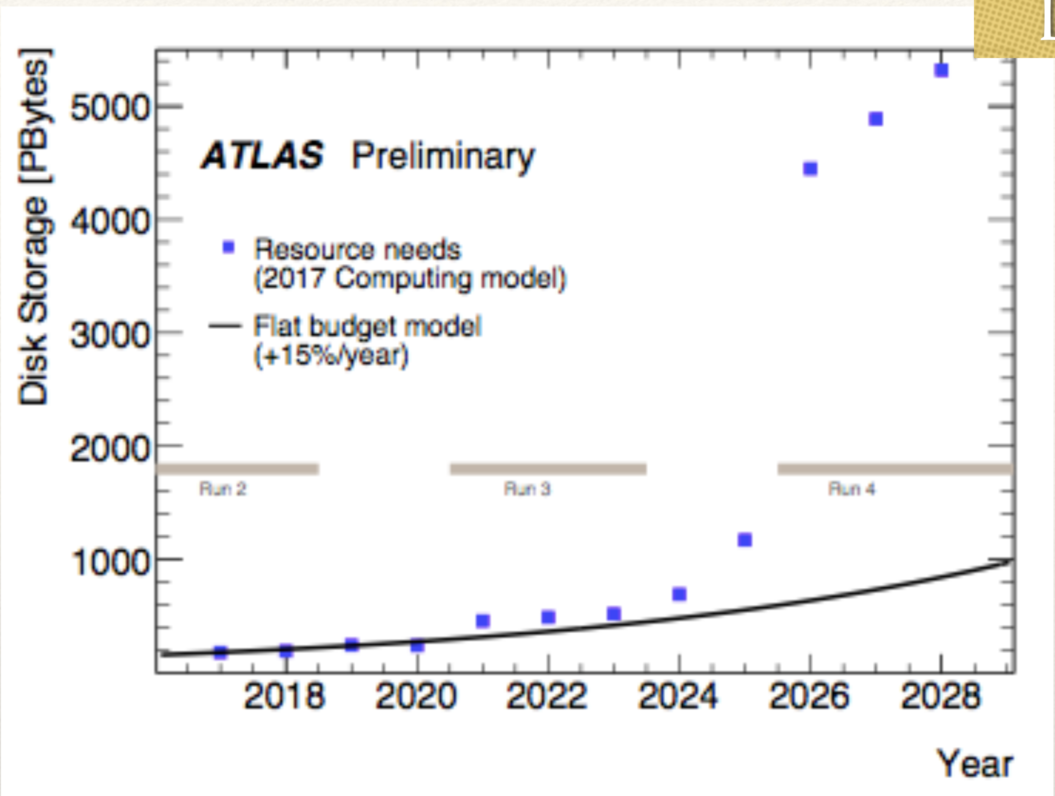
- processing velocity is a challenge, more than data volume

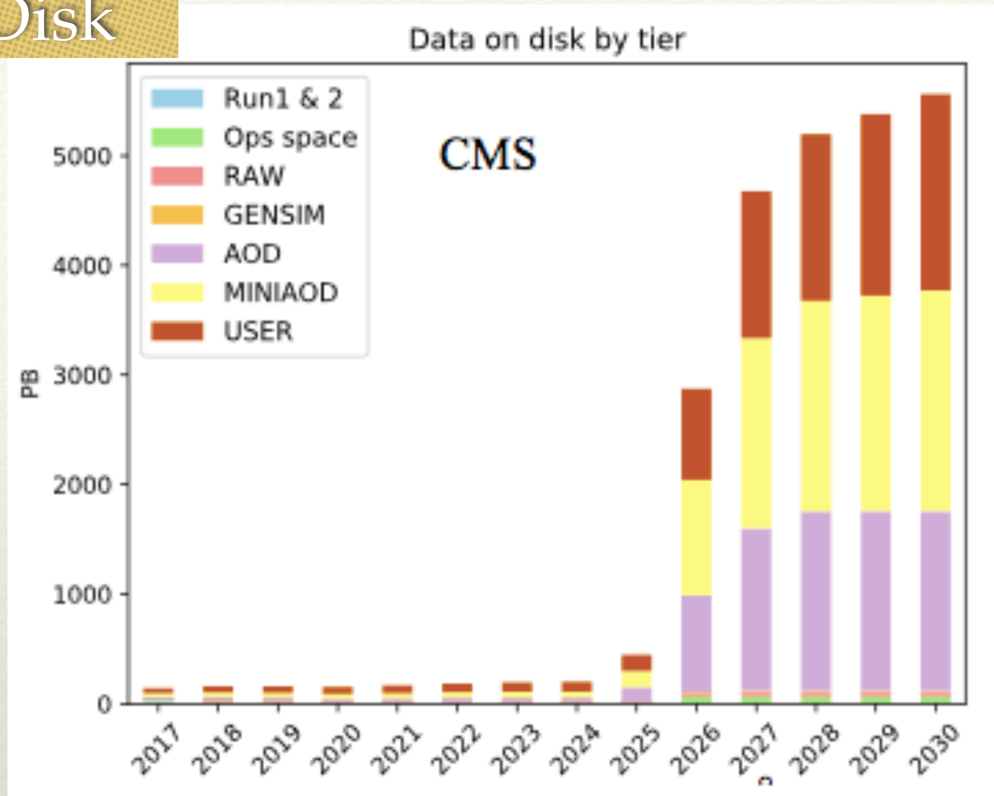
NOTE: Not only more events, but also high-granularity detectors. *HEP will pay the price of not having computing costs folded in since the design phase.*

- e.g. DUNE Liquid Argon TPC at ~150 EB/yr, impossibly large, to be reduced to <50 PB/yr. Similar considerations may be applied to selected HL-LHC upgraded detector design concepts

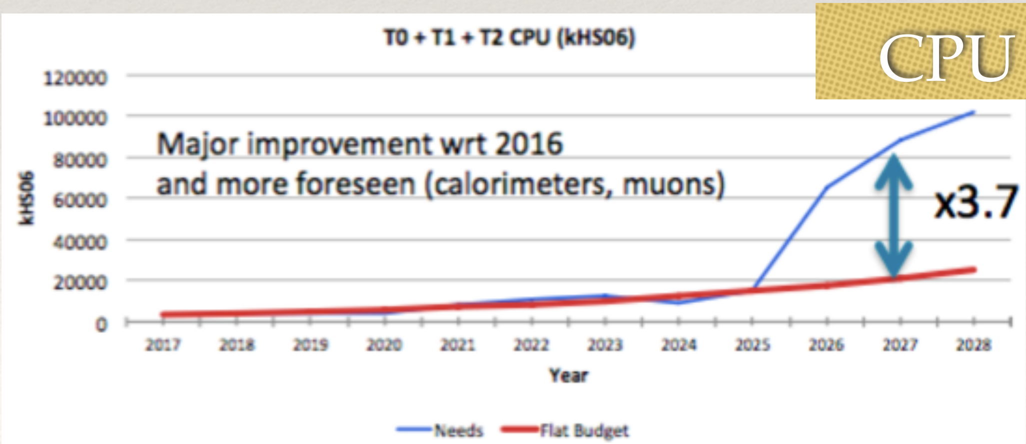
E.g. ATLAS and CMS towards HL-LHC

Disk





CPU



E.g. WLCG CPU cycles below ~4x what is achievable with the technological improvement alone

C3S working group on resources evolution

On the computing resources side, a starting point to look further is the work recently done on a C3S mandate:

- Focus: "***Valutazione delle necessità delle risorse di calcolo INFN***"
  - time span: next 8 years
  - Group: Bonacorsi (chair), Cosmai, Giagu, Cirrone, Piano, Punturo
  - Mandate on April 30th. Talk(s) delivered on June 28th, July 18th. Report delivered on July 18th (linked below)

Report:

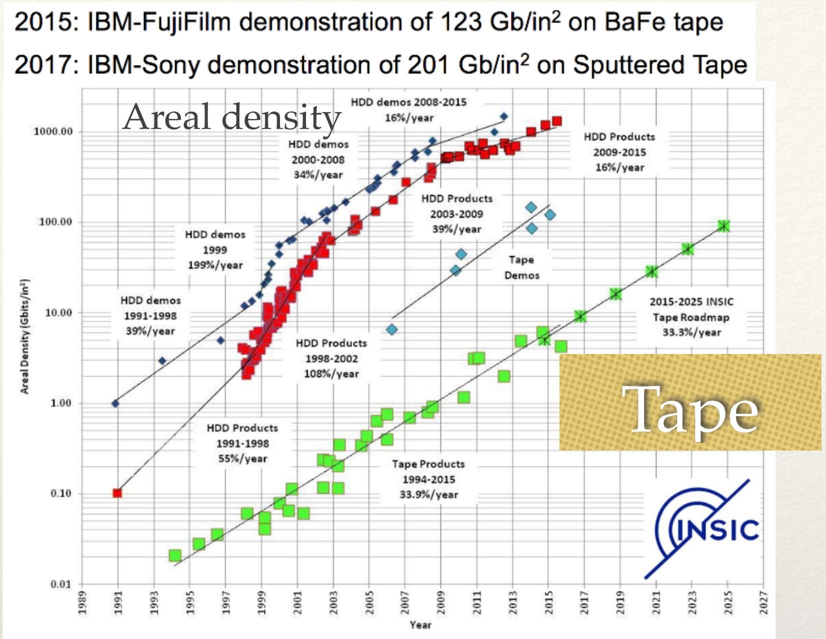
- bit.ly/C3S-GruppoValutazioneRisorse (56 pages)

All details for each CSN are in the report

- Additional important input from the C3S working group on technology tracking (plus current and future WLCG/CERN technology tracking initiatives)

In the following, highlights on common points across CSNs

Technology evolution and projected shortfalls



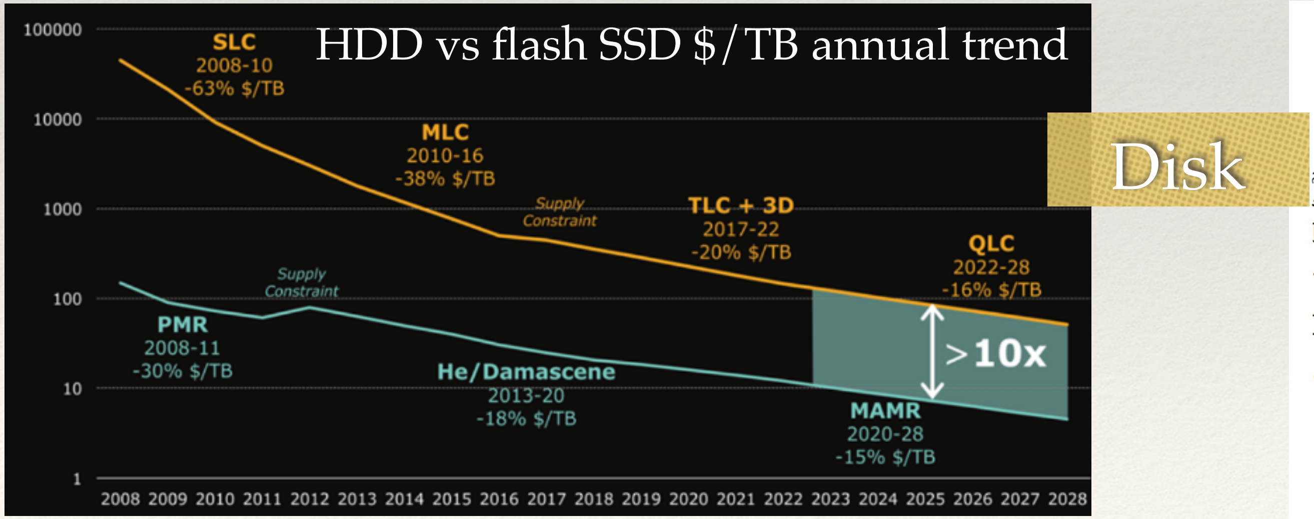
Market trends indicate **slow price/perf improvements** in both compute and storage

- ~10-20%/yr for compute; ~20% for storage

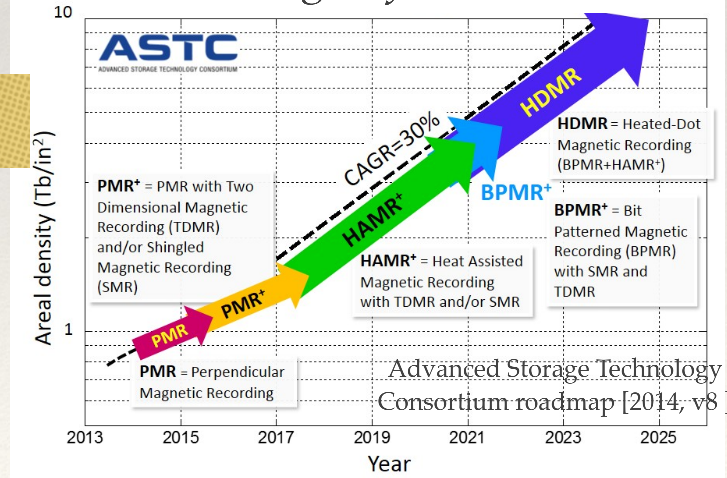
Projected shortfalls assuming constant budgets remain high

- No technology breakthrough or competitive market pressure on the horizon

SSD not replacing HDD any time soon.



*Composite roadmap of different *MR technologies for HDD*



CPU: evolution of the processing landscape

Trends towards **many-core architectures**, and **GPU**s (and **FPGA**s.. and TPUs?)

Easy to use them? No! Requires **massive code rewriting**.

- code development and maintenance specific to each processor generation, i.e. hard to migrate to a new architecture, to concurrency-based programming models, and/or to embrace a high heterogeneity of resources
- a big experiment typically has a code base of 5-10M lines of code, written in last 15-20 years..

On the other hand, **code writing could be perhaps focussed on most computationally expensive parts** (e.g. sim, or pattern reco) and evaluate new approaches/algorithms - *most important e.g. for new or upgraded detectors*

- highly parallelised code that could e.g. run on GPUs or TPUs
- application of Machine/Deep Learning at large

more later

Developers/users must have platforms to try code out. A benefit is hence identified in **investing in medium-size {GPU/FPGA}-based resources**

more later

- testbeds for code prototyping, DL model training, integration and pre-production scale tests
- one (or few) locations only, to profit from investments and scale up, focussed manpower. Access granted to remote developers, larger users base and distributed know-how.
  - local small-size GPU resources have different R&D goals, still important and to be supported!

(Re)-rise of Machine/Deep Learning → HEP applications

A zoo of supervised/unsupervised machine learning algorithms, plus evolutions towards deep (many hidden layers) artificial neural networks

More on ML algorithms in HEP

BDTs/**ANN**s typically used to classify particles and events

- they are also used for regression, e.g. to obtain the best estimate of particle's energy based on the measurements from several detectors

ANNs being used for a while in HEP, then.. → rise of **DNN**s

- particularly promising when there is a large amount of data and features, as well as symmetries and complex non-linear dependencies between inputs and outputs

Dif

Plu

CCR -

Summary

The **use of ML** is becoming **ubiquitous in HEP**

- a rapidly evolving approach in HEP to characterising and describing data with the potential to radically change how data is reduced and analysed

Applications domain varies:

- Some will qualitatively (directly) improve the *physics reach of datasets*. Others will allow more efficient use of computing resources, thus (indirectly) extending the *physics reach of experiments*

DL is starting to make a visible impact in HEP

- firstly, with HEP problems that are closely related to those commonly solved using DL

Collaboration with **CS** and synergy with the **world-class ML community** is vital for HEP, and a challenge in itself for both sides!

- HEP has interesting features from a CS perspective (sparse data, irregular detector geometries, heterogeneous information, systematics, ..)
- HEP should be open to other communities, and improve in how to formulate problems in a way CS can understand and be attracted to

CCR - Rimini, June 2018 34 D. Bonacorsi

CCR, Rimini, June 2018

- bit.ly/CCR-Rimini-Jun18

REVIEW **nature**

https://doi.org/10.1038/s41586-018-0361-2

Machine learning at the energy and intensity frontiers of particle physics

Alexander Radovic[1]*, Mike Williams[2]*, David Rousseau[3], Michael Kagan[4], Daniele Bonacorsi[5,6], Alexander Himmel[7], Adam Aurisano[8], Kazuhiro Terao[4] & Taritree Wongjirad[9]

Our knowledge of the fundamental particles of nature and their interactions is summarized by the standard model of particle physics. Advancing our understanding in this field has required experiments that operate at ever higher energies and intensities, which produce extremely large and information-rich data samples. The use of machine-learning techniques is revolutionizing how we interpret these data samples, greatly increasing the discovery potential of present and future experiments. Here we summarize the challenges and opportunities that come with the use of machine learning at the frontiers of particle physics.

The standard model of particle physics is supported by an abundance of experimental evidence, yet we know that it cannot be a complete theory of nature because, for example, it cannot incorporate gravity or explain dark matter. Furthermore, many properties of known particles, including neutrinos and the Higgs boson, have not yet been determined experimentally, and the way in which the emergent properties of complex systems of fundamental particles arise from the

Big data at the LHC

The sensor arrays of the LHC experiments produce data at a rate of about one petabyte per second. Even after drastic data reduction by the custom-built electronics used to readout the sensor arrays, which involves zero suppression of the sparse data streams and the use of various custom compression algorithms, the data rates are still too large to store the data indefinitely—as much as 50 terabytes per second,

Recent HEP review work on **Nature** (Aug 2nd, 2018)

- bit.ly/ML-DBonacorsi

A glance to a GPU/FPGA testbed [credits to Piero Vicini]

(more later when we talk about HPC - High Performance Computing)

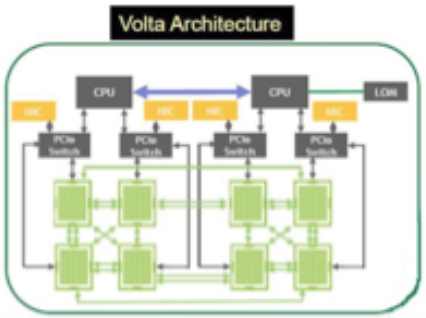
Supercomputing centres are dominated by **hybrid systems**

- based on the CPU+accelerators paradigm (large fraction of computing is delivered by accelerators)

E.g. a strong user is the INFN theoretical physics community: not all its code is (yet) fully optimised to exploit all parallelism levels of these architectures. Same (worse?) for most HEP/ astrophysics experiments code

- we need to learn how to make effective use of such computing platforms

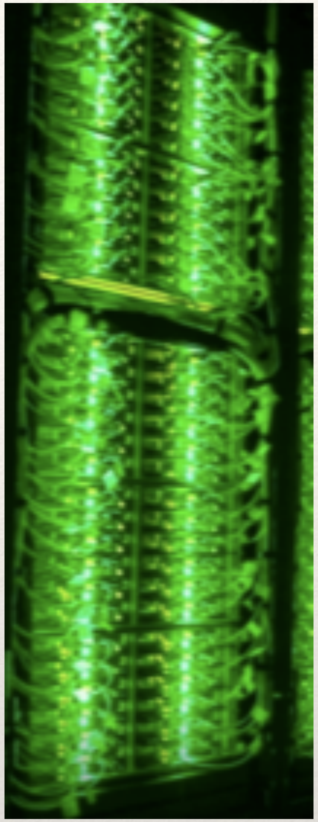
4029GP-TVRT



INFN launched a new initiative aiming to:

- build a testbed through procurement of small/medium state-of-the-art systems for training, test and benchmarking activities (NOT for production).
  - **Now** (Sept 2018): a small SuperMicro Cluster with 2x 4029GP-TVRT and Mellanox IB EDR switch
    - X86 CPU system accelerated by 8x Nvidia Volta GPUs interconnected on board with Nvlink proprietary technology (for latency saving and bandwidth improvement)
  - **End 2018** (1Q2019): an X86-based platform accelerated by FPGA systems (Xilinx VU+ and/or Intel PAC) based
- execute exploration and training activities for efficient programming of existing and up-coming hybrid systems

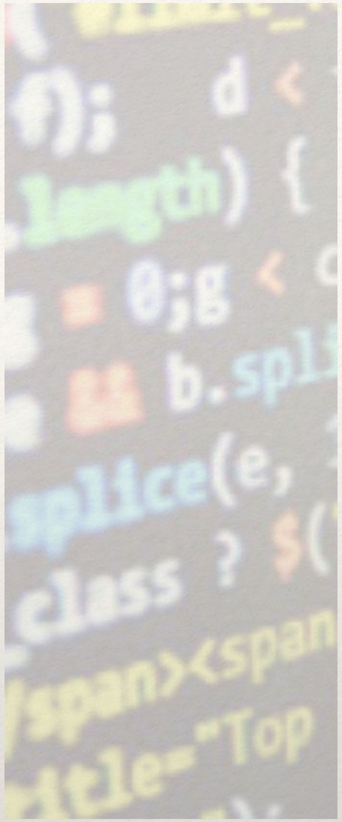
This is the kind of testbed-level investments that might make a difference



Multi-architecture computing

Storage

(and Network)



Training

R&D

(medium- and long-term)

HTC vs HPC

HTC (**High Throughput Computing**) is a computing approach that aims to make available a large number of computers to quickly accomplish tasks that are easily broken up into smaller, independent components

- distributed computing, cloud computing, compute servers

HPC (**High Performance Computing**) aims at building hardware and software that are focused on peak computing capability (i.e. speed) and extremely fast interconnectedness, rather than on the number of simultaneous tasks that can be accomplished

- high-end computing, supercomputing, world-class computing

As of today, **HEP computing is** (dominantly) **HTC-based**

- in-house custom built computing centres, interconnected by Grid middleware
- all key components and services developed / tested / deployed / used in operations over the last ~2 decades

What about HEP using HPC, instead?

The HPC worldwide reality (neglecting HEP for a while..)

HPC is here to stay. Worldwide race for leadership in HPC systems driven by the need to address societal and scientific grand challenges more effectively, and for strategy reasons

- climate evolution forecasting, early detection and treatment of diseases, human brain studies, preventing and managing large-scale catastrophes, need of industry to innovate products and services, ..

HPC will expand further, towards higher performances

- multiple processors connected by fast network to achieve higher performances than single processors: O(100k) processors and O(10 PFlops)

- HPC systems get ~1000x more powerful every decade. "Exascale" resources expected by 2020-25

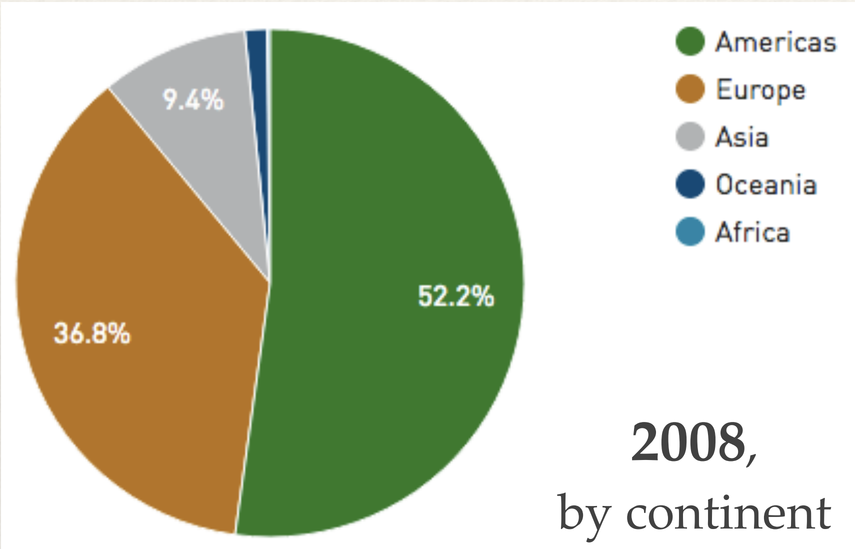
Feel the scale: a typical high-end HPC system (e.g. CINECA Marconi) deploys a computing power comparable to the entire world-wide HEP processing scale!

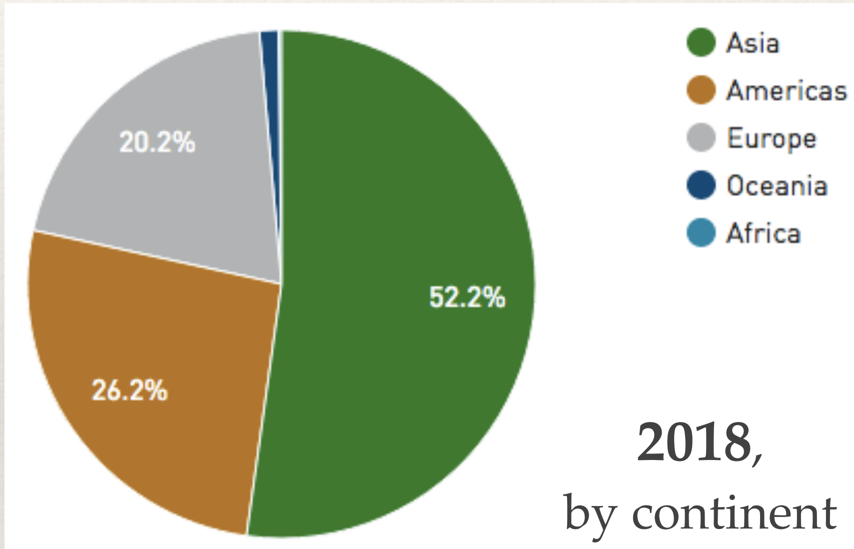
Connection to EU strategy: macroeconomics and market of HPC

Returns on investment in HPC are extremely high. Companies (and countries) that invest the most in HPC lead in science and economic success

- Advances in HPC (new computing technologies, software, energy efficiency, storage applications, ..) feed into the broader ICT industry and consumer mass market within ~5 yrs from high-end HPC.

Not every continent is running at the same speed: **EU positioning is fragile.**





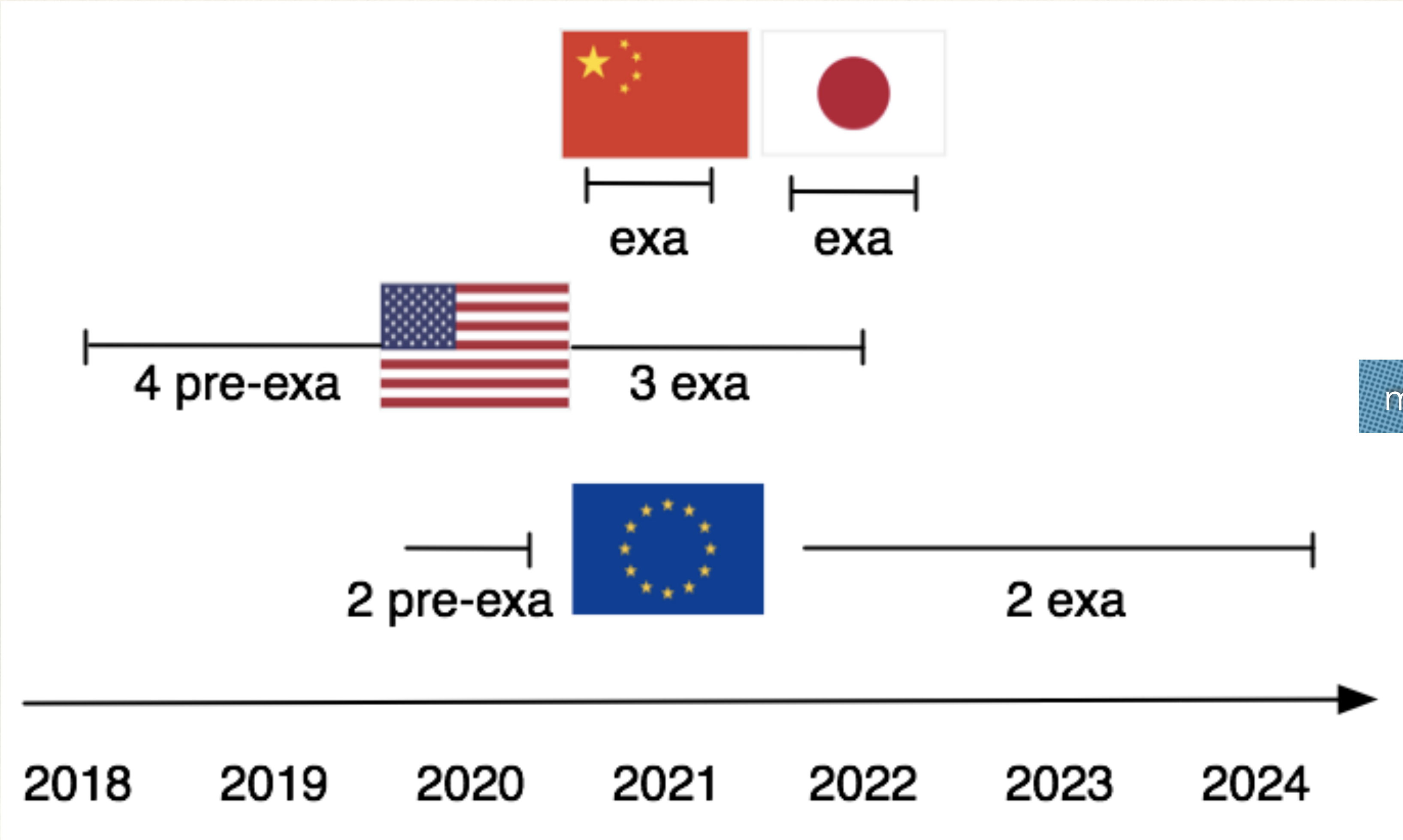
Note: the legenda is different between the two plots.. *[*] top500.org*

Strategic nature of HPC for EU:

- crucial asset for the EU's innovation capacity, and EU funding opportunities on HPC ahead of us, targeted to improve the EU leadership positioning on HPC
  - avoid scientists (and know-how) relocation. Radical innovations in technologies needed to meet the "exascale" challenge will be opportunities to EU industrial and academic players to reposition themselves in the field

Pre-exascale and Exascale (projections)

Global picture of worldwide HPC exascale and pre-exascale plans:



more later

eurohpc.eu

EuroHPC

European HPC Initiative: a joint collaboration between European countries and the EU about **developing and supporting exascale supercomputing by 2022/2023**

- declaration signed in March 2017 by 7 countries [*]; so far, 22 countries [**] have joined

Plan 2018-2020 (~1.5B€ - 30% from EC)

- build 2 pre-exascale machines by 2020(21?), ~100-200 PFlops
- build 2 (4?) petascale machines

Plan 2021-2026 (proposed 2.7B€, out of 9.2B€ for digital infra)

- build 2 exascale machines by 2024 (25?), ~1 EFlops

Italy participation in EuroHPC is strategic

- we are doing it right: one of the two pre-exa machine will arrive to Italy

More info: EuroHPC Joint Undertaking (bit.ly/EuroHPC-JU)

[*] France, Germany, Italy, Luxembourg, Netherlands, Portugal, Spain

[**] countries above (March 2017), plus: Belgium, Slovenia, Bulgaria, Switzerland, Greece, Croatia, Czech Republic, Cyprus, Poland, Lithuania, Austria, Finland, Sweden

ETP4HPC

Importance of HPC for EU was recognised by the EC in one of its Communication [*], which also stipulated the creation of **ETP4HPC**, a European Technology Platform in the area of HPC

- an industry-led think tank, composed of EU HPC technology stakeholders (technology vendors, research centres, end users)
- Goal: align the efforts of the EU HPC technology providers and facilitate the emergence of a globally competitive HPC industry in Europe. Method: define research priorities and action plans in the area of HPC technology provision
- ETP4HPC issues and maintains a **Strategic Research Agenda** [**] as a mechanism to help the EC define the contents of the HPC Technology Work Programmes - thus acting as the "one voice" of the EU HPC industry in relations with the EC and national authorities.
    - latest is SRA 3 [***], issued in December 2017

[*] "High Performance Computing: Europe's place in a Global Race" (EC COM(2012) 45, Feb 2012)
[**] http://www.etp4hpc.eu/sra.html
[***] http://www.etp4hpc.eu/pujades/files/SRA%203.pdf (123 pages)

What about HEP and HPC?

2.1
THE VALUE OF HPC

2.1.1
HPC as a Scientific Tool

Scientists from throughout Europe increasingly rely on HPC resources to carry out advanced research in nearly all disciplines. European scientists play a vital role in HPC-enabled scientific endeavours of global importance, including, for example, CERN (European Organisation for Nuclear Research), IPCC (Intergovernmental Panel on Climate Change), ITER (fusion energy research collaboration), and the newer Square Kilometre Array (SKA) initiative. The PRACE Scientific Case for HPC in Europe 2012 – 2020 [PRACE] lists the important scientific fields where progress is impossible without the use of HPC.

Finally, in the field of High Energy Physics and QCD, the goal by 2021 and beyond is to perform most lattice calculations of hadronic systems at or near the physical pion mass, with lattices representing physical volumes of (4 fm)3 and larger. To achieve robust signals from these types of calculations, the scale and of the problem must be increased by at least a 1000-fold compared to today's calculations, and most likely larger.

From SRA 3, page 16

HPC as scientific tool recognised as strategical for Europe competitiveness by ETP4HPC.

Expansion of future EU HPC centres to more data processing needs is envisioned.

But:

- most computing-intensive communities are participating in EuroHPC (**HEP** left out for now)
- **HEP** requirements are not mentioned in the SRA (apart from lattice QCD)

For EU strategy: how to ensure HEP presence in future <u>design</u> of facilities in the EuroHPC landscape

HEP towards HPC: *opportunities*

As said, **HEP relies heavily on HTC**

Enormous work over the years to **evolve HEP Computing models**.

- more flexibility: e.g. less reliance on data locality, more reliance on high-performance networking, smaller list of requirements in general
- large help by system virtualisation: despite not predominant, Cloud access to resources is a reality for most medium-large experiments

But… concerns about long-term sustainability of the HTC-based HEP-specific infrastructure and middleware.

A few points leaning **in favour of HPC**:

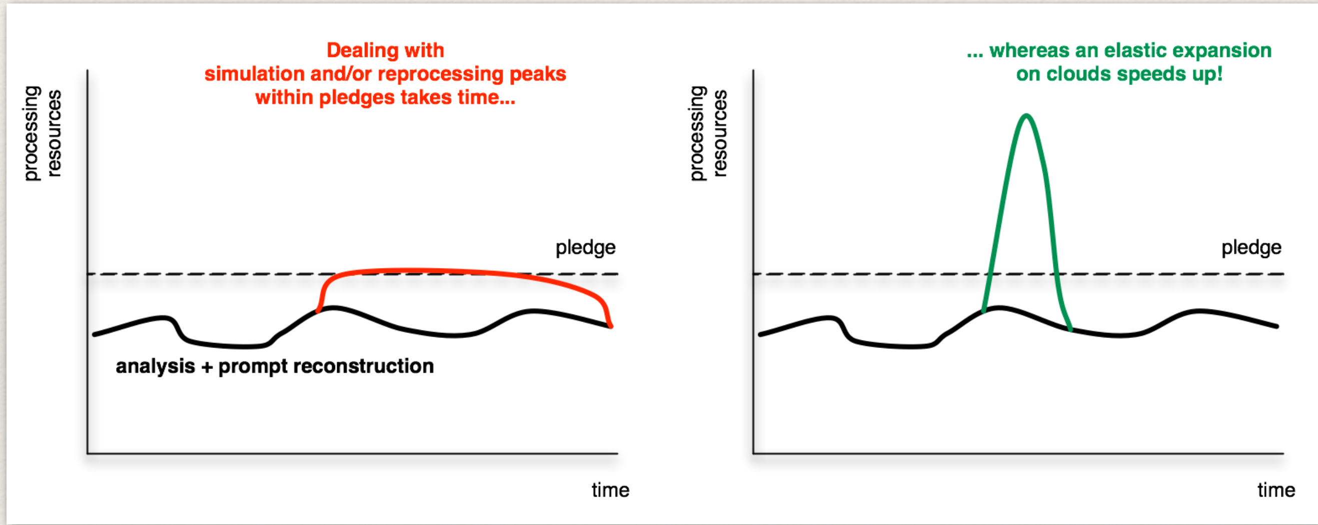
- **resource size**: large research grants from HPC centres to HEP use-cases
- **cutting edge technology**: high-end HPC systems have latest greatest technologies
  - current HTC systems for HEP have been built with a performance vs price trade-off to be economically affordable
- **access to new funding opportunities**: HPC funded via specific EU programs (e.g. PRACE)
  - HTC for HEP funded with standard budget of regional Research Institutions, with extrapolated needs in next decade(s) that go well beyond sustainable budget levels
- **strategic impact on competitiveness (beyond science)**: HPC perceived as more strategical than HTC
  - regional FAs push against deployment of two - HTC and HPC - infrastructures coexisting in parallel

Cloud(s)

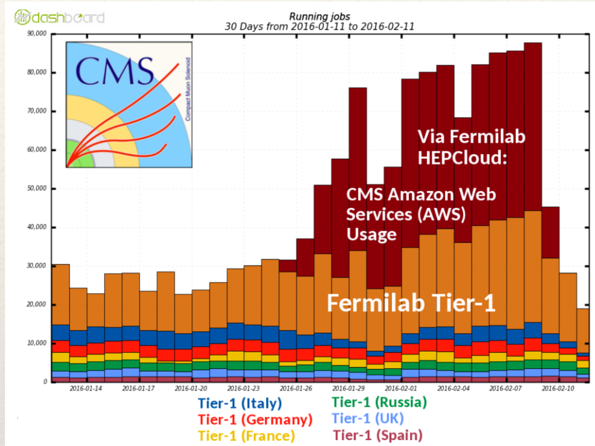
Disclaimer: examples from CMS, but similar successful efforts by other experiments too

Can we (technical) access not-owned resources on-demand?

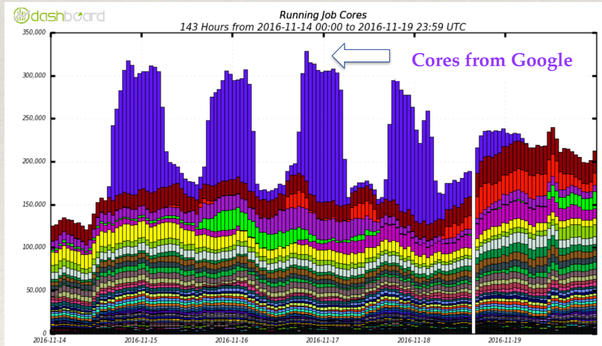
- Yes! Cloud computing offers demonstrated success stories



Amazon *(AWS)*



Google *(GCP)*



"So, let's do it on HPC systems and we are done!". **Is it so simple?**

HEP towards HPC: *challenges*

The use of HPC systems by HEP is far from trivial, as they are custom-built, and tailored around use-cases largely different from HEP ones

- e.g. material science, molecular dynamics, simulation of complex systems, ..

Where **HPC** mostly differ from HEP-**HTC** (as it is today):

- **HPC** facility design driver is the ability to demonstrate **best performance on standard benchmarks** (in order to be as high as possible on the official HPC ranking [1]).
- **HPC** systems must have **highly performant node-to-node interconnectedness** needed for large scale MPI tasks. in HEP-**HTC**, node-to-node connectivity is scarcely relevant
- **HPC** systems have **scarce local scratch disk** on the single nodes. HEP-**HTC** needs large on-node scratch areas, crucial for intermediate stage-out of computation results
- **HPC** systems are **not meant for accessing data hosted outside the facility**. HEP workflows are typically data-intensive, and global WAN connectivity is needed to access also remote datasets: thus, HEP-**HTC** design deploys large-capacity storage systems "close" to the computing farms
- **HPC storage optimisation criteria** are for high speed and low latency, not for overall size. HEP-**HTC** storage size is highly relevant, and a cost vs benefit tradeoff applies (e.g. less use of SSD w.r.t HDD)
- **HPC** systems include **accelerators** (GPGPU, FPGA, ..) to boost total performance (and hence global ranking). In HEP-**HTC** their use is relatively marginal
- **HPC processor architectures** just aim at a high global ranking. HEP-**HTC** infrastructure is almost entirely based on cost-affordable Intel x86_64 architecture, which HEP software stack is designed and optimised for

Viable strategy for HEP towards "more HPC" [1/2]

Q: What can our community do to put current and future HEP experiments/ groups in the condition to exploit any high-end HPC facility that would eventually be deployed in the next decade(s)?

The CPU architecture choice may not be an issue in the longer-term

- assuming we will have CPUs with a supported Linux(-like) OS and a performant C++ compiler (well, better if GCC, better if C++17 recent extensions are supported, ..)

Efficient exploitation of on-board accelerators, instead, requires **massive work**

- they are anyway supposed to be contributing to a _large_ part of the overall computing power on any such HPC systems: **HEP expanding to HPC with a HEP software unable to exploit accelerators is just not an option**.
  - even more crucial, in view of the "**co-design**" principle towards the exascale environment!
- GPGPUs', or FPGAs', utilisation pattern diverge from standard C++ multithreaded programming: **nearly full code rewriting is needed for their efficient exploitation**
- Programmers need to learn how to rewrite the code. But it is experimental physicists who wrote most of the code! Skills and R&D needed are not trivial. Large payoff, but **a strong investment is needed in training on modern software development** and user education

More.. (next)

Viable strategy for HEP towards "more HPC" [2/2]

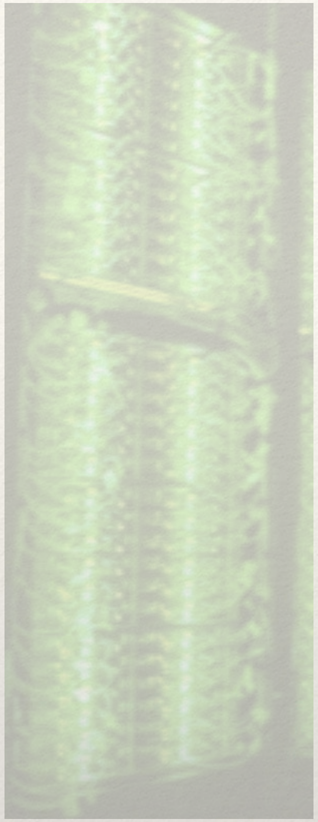
As of today, if a future generic HPC facility gets deployed somewhere and following "traditional" criteria (top500 ranking, usual users communities), the only chance for HEP workflows to run on it, is to <u>hope that its specific facility happens to have a flexible and general purpose design</u>

- we need to hope that design decisions would eventually happen not to be against us

So, for **EU strategy**: **HEP stakeholders need to sit at the same table with HPC future facilities designers *<u>in the design phase</u>***, well in advance w.r.t construction phase

Additionally, future HPC systems are experiencing a changing users base: we should **work in synergy with other e-Sciences** and not apply this pressure alone

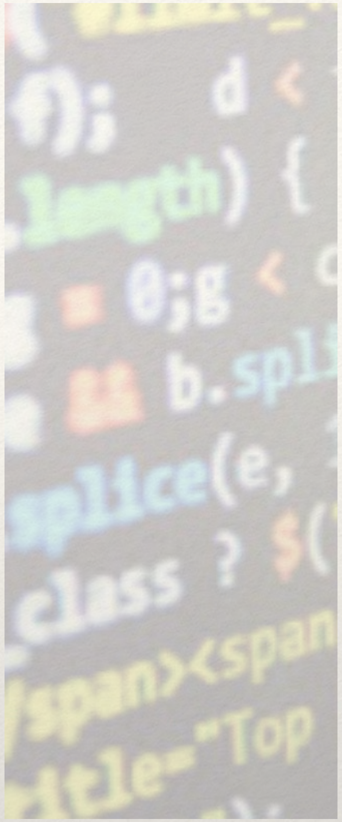
- system openness and large data access capability are nowadays requirements also from others disciplines, and HPC is less close than before to the users who originally drove the realisation of past decades' HPC facilities
    - Genomics, Medical Physics, Big Data mining and analytics (plus Machine/Deep Learning usage)



Multi-architecture computing

Storage

(and **Network**)



Training

R&D

(medium- and long-term)

A “data lake” strategy [1/2]

Storage and data management remain big challenges for HEP in the next decades

- 1) main cost driver of computing; 2) resource needs will grow; 3) heterogeneous systems across activities

No magic here: at least, we can push on economy scale models!

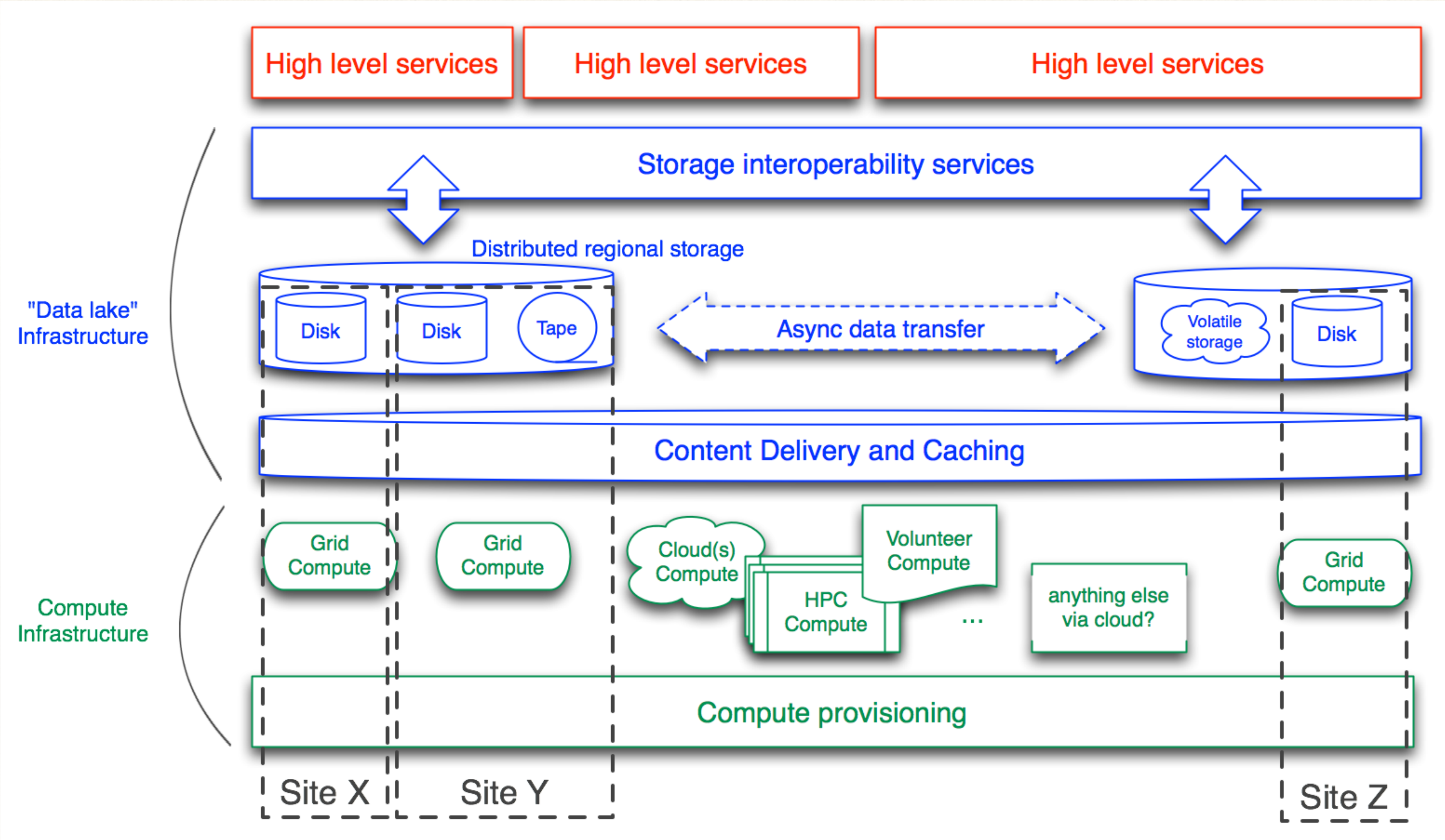
Storage consolidation with a (**federated**) **data centre concept** is gaining thrust in the EU international discussions

- geographically distributed storage centres (with any implemented storage solution) to be operated and accessed as a single **logical** entity

Technical and strategical points of a data lake - towards the **EU strategy**:

- overall cost reduction for storage (less needed replicas, thus smaller storage capacity - plus reduced maintenance costs, and better management of storage and network in the long-term, whilst maintaining a relatively high CPU efficiency)
- the vision of the centrality of who hosts “the data” retains your key player role in the global infrastructure and give you a voice in its future evolutions (strategically crucial towards exa-scale computing)

A “data lake” strategy [2/2]



Data lake on pair with a national INFN cloud (requests orchestrator, on-demand CPU provisioning, instantiations of clusters, PaaS, cloud bursting to other clouds (public or commercial), ..)

Networking

Redundant high-perf networks are best friend of computing models evolution.

High-throughput network infrastructure based on DCI (Data Center Interconnect) links, a technology used in the recent past, is very promising:

- e.g. across-Tier storage resource management, homogeneously via the same batch system
- e.g. cloud bursting mode, i.e. elastically extending the T1 farm on external cloud providers to absorb peaks of CPU requests
- e.g. CNAF-CINECA optical DCI connection, in collaboration with GARR

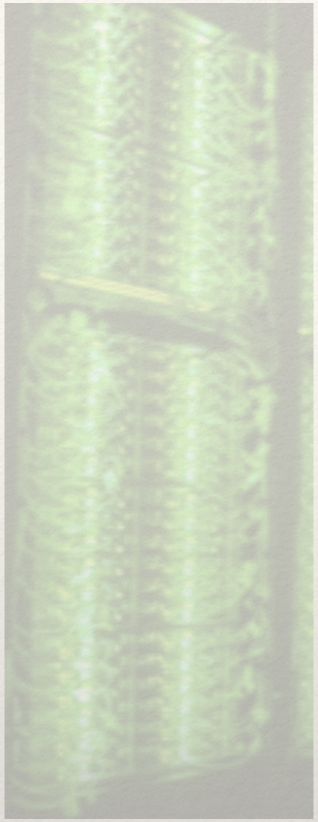
The "data lake" scenario is a potential use-case for a DCI approach

- "software defined WAN" implementation - being experimented with GARR - can grant access to data stored into the data lake

Price to pay: a necessarily improved WAN connectivity.

But, strategically: <u>isn't probably needed anyway?</u> **Network over-provisioning** is a very clever tactical choice to enable future innovative approaches

- strategic to continue deep technical investment in experimentation with GARR, in actively joining the HEPIX NFV working group activities (on R&D with SDN and NFV), etc



Multi-architecture computing

Storage

(and **Network**)

Training

R&D

(medium- and long-term)

Training and skills

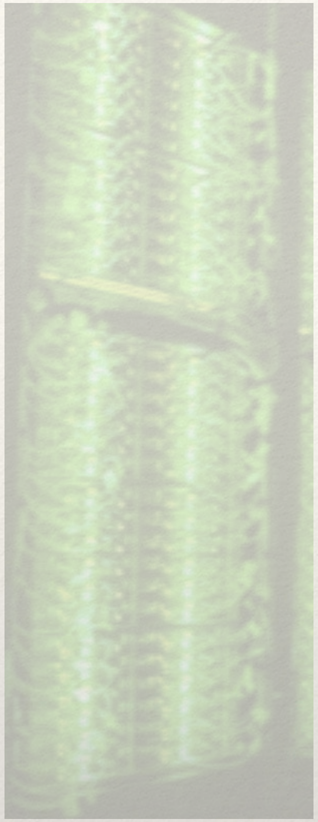
The success of most of all this depends on **skilled manpower**.

Key objectives:

- adequate and pervasive training paths (mainly for young collaborators, but not only!)
- acquires competencies we largely lack, e.g. advanced computing and software skills to exploit all processing architectures, heterogeneous computing, new computing paradigms, ML/DL/, ..
- knowledge transition, i.e. from experts to non-experts
- **career recognition** to avoid skilled personnel haemorrhage out of HEP (industry, data science companies, ..)

International, HEP-wide coordinated training efforts (logistically and financially supported) are the only chance to build something that stays

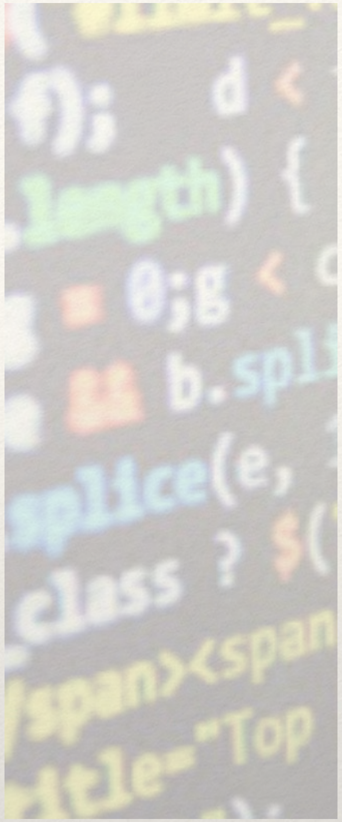
- **EU strategy**: support HEP-focussed training projects/networks EU proposals



Multi-architecture computing

Storage

(and **Network**)



Training

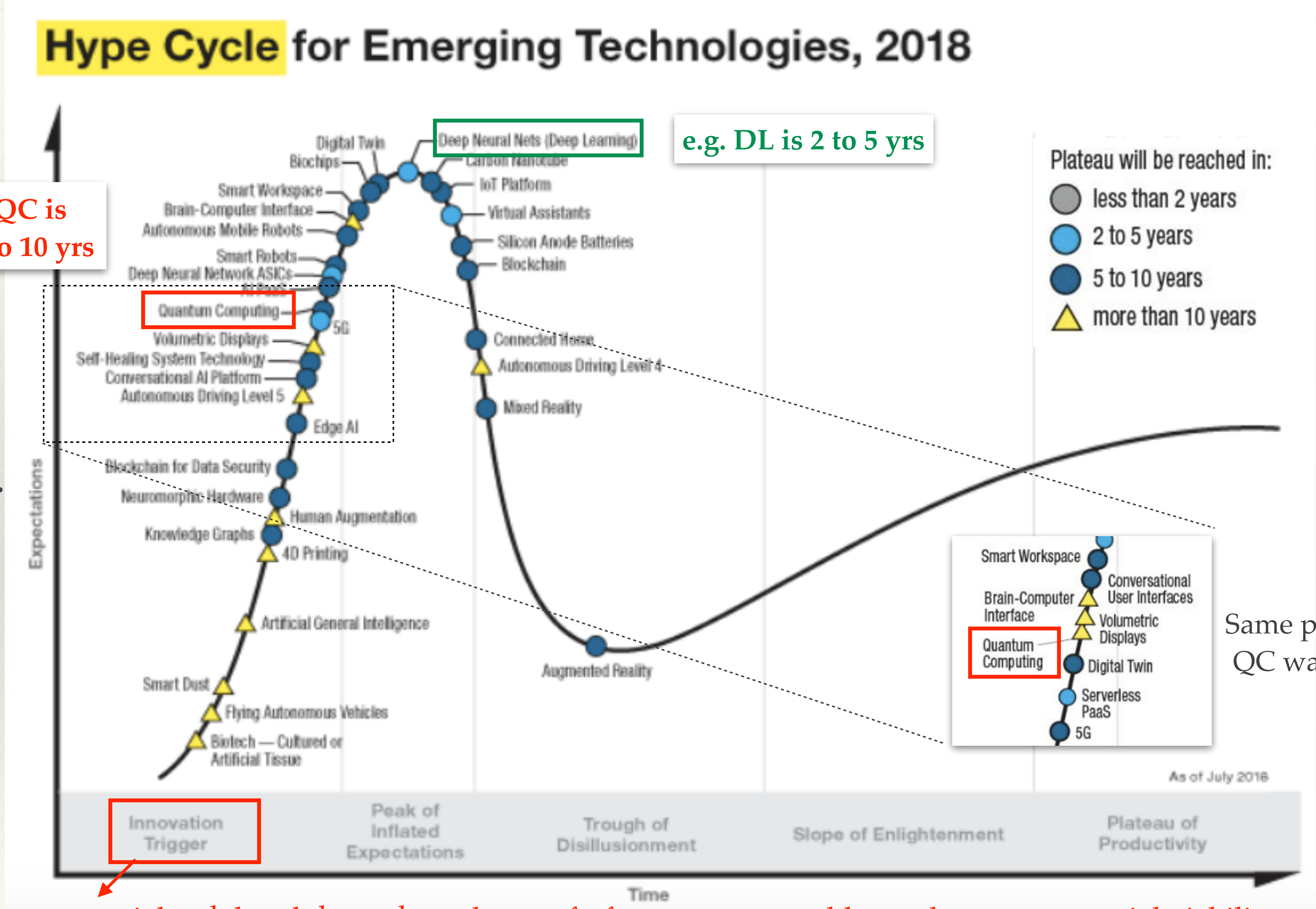
R&D

(medium- and long-term)

Previous slides → discussed medium-term R&D for processing and storage

Next slides → will discuss a long-term R&D: **Quantum Computing**

QC is getting closer?



QC is 5 to 10 yrs

e.g. DL is 2 to 5 yrs

Same plot in 2017: QC was >10 yrs..

potential tech breakthrough, early proof-of-concept, no usable products, commercial viability unproven

Quantum Information Science (**QIS**) and Quantum Computing (**QC**) systems

Some problems are just too complex, even for today's fastest supercomputers

- classical computing (CC) systems may take <u>months</u> or even <u>years</u> to run through a series of permutations, making it impractical to attempt
- quantum computing (QC) systems may take just <u>days!</u>

Huge potential of QC to be a disruptive technology for accelerated computing. Question is: **will it be real, or just hype?**

So far, QIS has been focussing on:

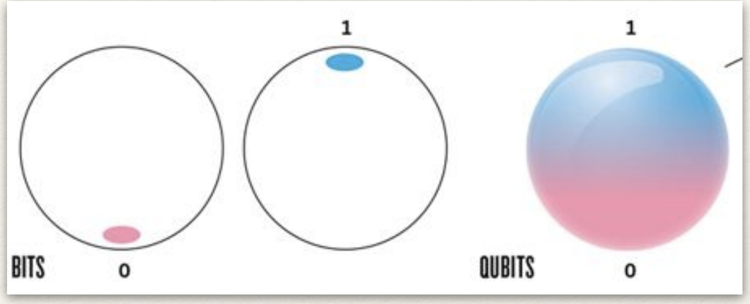
- can we gain some advantage by storing, transmitting and processing information encoded in systems that exhibit unique quantum properties? Today it is understood that the answer in principle is yes. And many research groups started working to build QC systems

Today, QIS unanswered questions are:

- which technology, if any, will ultimately prove successful? when might this happen? how might QC become a practical platform for computing? at large or mostly for ML/DL/AI, data science, cryptography, or.. ?

Quite a big set of open questions.

bits vs qubits   [1/2]



Bits vs qubits. QC is a type of non-CC based on the quantum state of subatomic particles

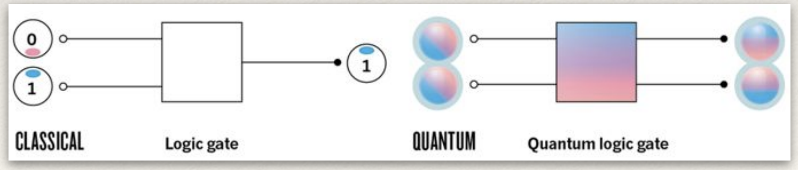
- CC operates in binary bits (0,1).
  QC operates in "qubits"
- a qubit can represent 0 or 1 or a superposition of both partly 0 and partly 1
    - 10 bits can encode one number in the [0,1024[interval; 10 qubits can encode all 1024 numbers simultaneously

More qubits. Create, manipulate, join.

- Create qubits from several physics systems with distinct quantum states.
- Manipulate these systems (with e.g. lasers or microwaves) to create quantum superpositions of states.
- Join many qubits together and you can encode a huge amount of information.

(Q-)Operations. Exploit superposition of states.

- CC uses logic gates to get (0,1) as output. QC takes the entire superposition state of all input qubits and a quantum logic gate transforms it into another superposition state that that encodes all outputs.

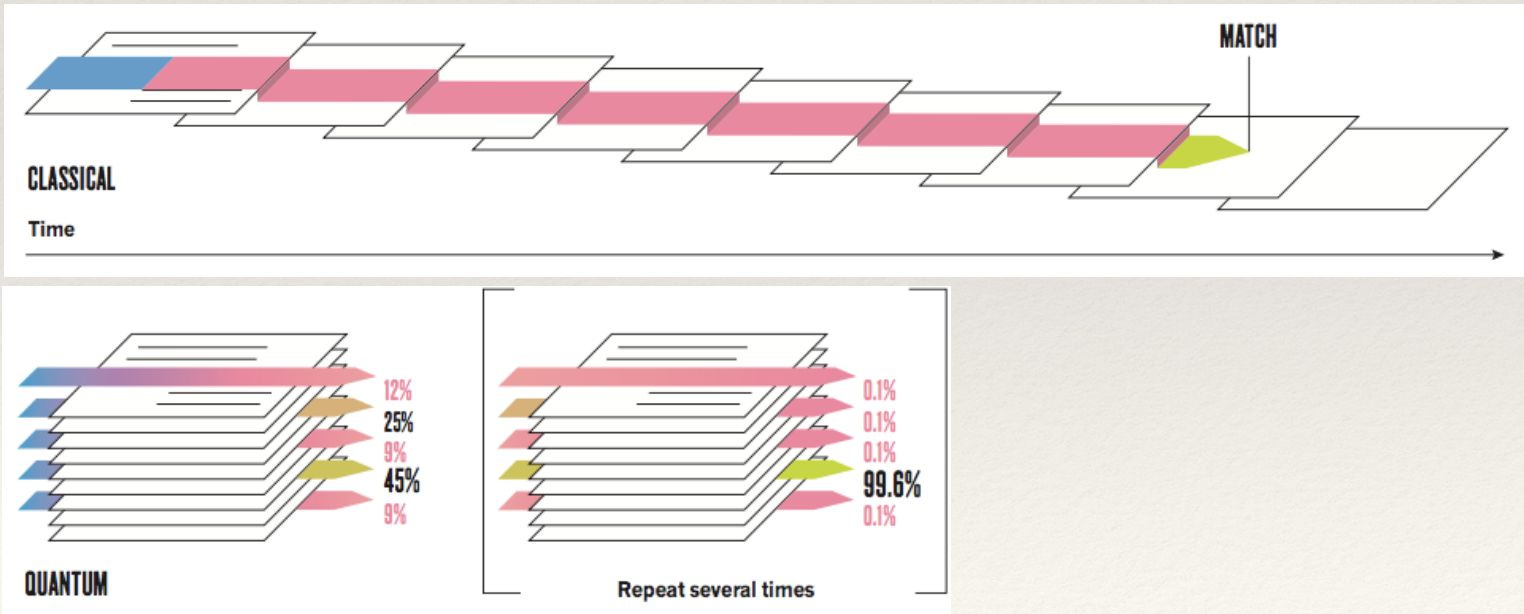


- criticality: protect the QC against perturbations to avoid unwanted changes to the quantum states, which lead to errors or loss of quantum superposition

bits vs qubits [2/2]

<u>(Q-)Algorithms</u>. Exploit the parallelism afforded by the superposition of states.

- All possibilities are analysed at the same time, instead of individually
- A Q-algo (Grover's) gives each possibility a probability of being 'right'. Perform several iterations (much quicker than classic searching anyway), and the cumulative probability of the target possibility will be higher than the others.
- The larger the DB, the bigger the advantage.



Q-challenges

Maintaining coherence

- Superposition is the key, but **maintaining coherent superposition** of quantum state is challenging, e.g. external disturbances - but at the same time you need to control the qubits to perform operations

Scaling up in qubits

- Added qubits must enter into a state of superposition with the other qubits. They are entangled, i.e. they influence each others in ways classical particles can't. **More qubits** make the overall QC system **more fragile**

Error correction

- CC systems have good error control, enabling robust outcomes despite imperfect components. Similar fault tolerance levels and **error correction capabilities in QC systems are not (yet) in place**: they may require adding ancillary "correcting" qubits surrounding existing one, which increase fragility and makes scaling up even more difficult

What are QC systems good for?

QC will most probably not replace CC, but may **excel in tasks that would remain prohibitive also for most powerful HPC systems**

- searching through huge DBs
- encryption, cryptography, secure communication schemes
    - e.g. via finding prime factors of large numbers
- calculate the behaviour of other quantum systems
    - e.g. detailed understanding of chemistry of molecules (requires knowing QM of all their electrons)
    - e.g. find optimal configuration of a folded protein
- optimisation problems on complex systems, minimisations, etc
    - applications in ML/DL/AI

In general, QC approaches are superb (and unique) on **accurate calculations of the properties of complex systems** with by far too many interacting elements to be dealt with by CC systems

Where are we with QC?

Conceivement in the <u>**1980**'s</u>

- it might be possible to construct computers based on the laws of quantum physics instead of on classical physics ("*if you want to make a simulation of Nature, you'd better make it quantum mechanical*", R.Feynman, 1981)

Excitement in the <u>**1990**'s</u>

- Shor's algorithm (1994) on factorisation (a general purpose QC could be used to efficiently factor large numbers). Hughes (1997) on cryptography with QC with trapped ions. Grover's algorithm (1996) on search. Zoo of quantum algos.
  - algebraic and number theoretical algos (→ cryptography), oracular algos (→ optimization, ML), approximation and simulation algos (→ quantum physics and chemistry)

Research for possible quantum hardware in <u>late **2000**'s and early **2010**'s</u>

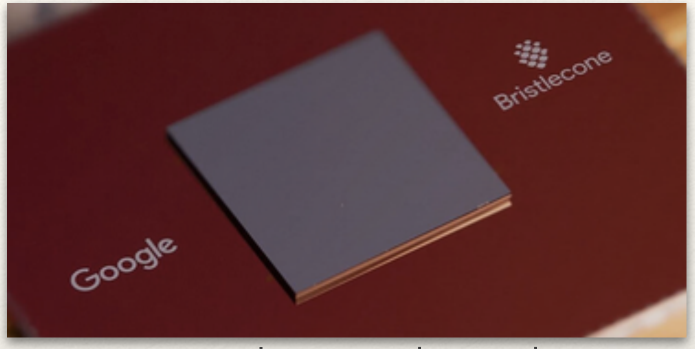
- Studies about isolating, manipulating and measuring elements that might form the basis of a QC, either single quantum entities (atoms, electrons, photons) or artificial systems that display QM behaviour (semiconductor structures or miniature electronic circuits)
  - ion trap, NMR, NV center, quantum dot, linear optical, superconducting, …

<u>Today</u>: 1) **proof of principle that QC could work;** 2) **superconductivity** seems one of the most promising quantum hw approach; 3) increased awareness of the **tremendous scale of the challenges** ahead!

Commercial QC efforts [1/3]

As of today, few companies have announced they have produced QC systems in the 5-72 qubits range

Google



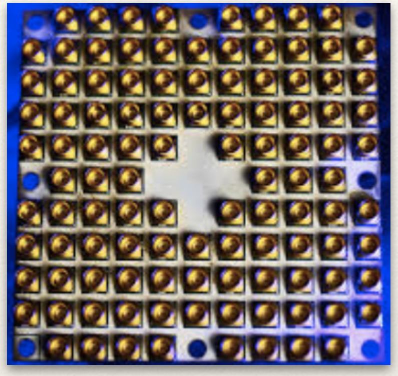
- "Quantum" effort in Google AI. "**Bristlecone**" is Google's newest **72-qubit** quantum processor. Superconducting qubits with chip-based scalable architecture, targeting two-qubit gate error < 0.5%.
- Focus on accelerating tasks for AI, quantum simulations, quantum neural networks, qubit metrology (a quantum supremacy experiment targeting two-qubit loss below 0.2% - critical for error correction), quantum assisted optimisation (i.e. hybrid quantum-classical solvers for approximate optimisation)

IBM

- **IBM Q**, an industry-first initiative to build commercially available universal QC systems for business and science, based on their **prototype commercial Q processor**. They use a fixed-frequency superconducting transmon qubit, a Josephson-junction-based one that is insensitive to charge noise (no tunable qubit to minimize sensitivity to external B fluctuations). Devices on silicon wafers with superconducting metals such as niobium and aluminum; refrigeration (^{3}He, ^{4}He) at 15 mK.
- Large effort in education, training, cloud-based simulation platform open to users

Commercial QC efforts [2/3]

Intel



- Collaboration with **QuTech**, their quantum research partner. "**Tangle Lake**" is their **49-qubits** superconducting quantum processor, currently under test at low temperature. It features 108 RF connectors on a 3x3 inches surface to carry microwave signals to operate the qubits (Josephson-junction-based, in gold).

Rigetti

- a full-stack QC company based in Berkeley. They open "**Forest**", a quantum developer environment, plus training etc. Oak Ridge National Labs performed the **first ever cloud-based nuclear simulation (deuteron) run** on a Rigetti QC system
- focus on quantum simulation, quantum ML, complex optimisation

IonQ

- ~unique in using ion traps

Commercial QC efforts [3/3]

D-Wave

- 2010: D-Wave One (**256-qubits**).
  2013: D-Wave Two (**512-qubits**).
  2015: D-Wave 2X (**1000+ qubits**).
  2017: D-Wave 2000Q (**2000 qubits**, and advanced control features). 2000Q also available via their quantum cloud. Quantum annealing machine. All superconducting technology.
  - Shielded to 50k times less than Earth's B; high vacuum (pressure is 10 billion times lower than atmospheric pressure), refrigerated at 15 mK, 200 I/O and control lines to the chip, chip consumption <25 kW. Producing and delivering QC hw and sw to Lockheed Martin, Google, NASA, Los Alamos National Laboratory, Oak Ridge National Labs, Volkswagen, ..
- Large set of quantum applications: optimisation, ML, sampling / Monte Carlo, pattern recognition, anomaly detection, cyber security, image analysis, financial analysis, bioinformatics, cancer research, traffic flow, manufacturing processes, internet advertising placement

For sure, more (growing) stakeholders that I have not quoted..

A note on Q-Metrics

If you think of QC in terms of a strategy for the future, do not stand in the headlights of the "number of qubits" metric..

Having more qubits is not necessarily the right metric. Number of gates that can be applied before loosing quantum coherence is (currently) the limiting factor for most applications

- range is from "few" to "thousands", and not all gates are the same..

A concept is emerging, as of the number of "**logical qubits**" incorporating error correction

- this seems to appropriately capture the complexity and be a more decent metric
- estimates indicate that **O(1000) qubits per logical qubit are required**

HEP positioning about QC

A **"QC for HEP" workshop at CERN** in November 2018

- https://indico.cern.ch/event/719844/

Work going on in HEP labs. *One example*: Fermilab has active work on:

- QC for FNAL science
  - HEPCloud will extend to QC. Ongoing testbed efforts in collaboration with Google. Focus on optimisation (proposed work on finding approx solutions to combinatorial optimisation problems), ML (exploratory project leveraging a D-Wave annealer for astrophysics - chosen because low enough in dimensionality: star/galaxy classification, anomaly detection, using autoencoders) and quantum simulation
- HEP technology for QC
  - R&D to improve qubit coherence, cold instrumentation electronics for QIS, ..
- Quantum technology for HEP exps
  - R&D applied to HEP detectors (new interferometric sensors, quantum imaging, quantum metrology)
- Quantum Networking
  - R&D on quantum communication channels (in collaboration with Caltech)

QC in the HEP strategy

A complicated (long-term) business.

Such a disruptive and "innovation trigger"-level technology, requires a **constant technical-level follow-up** as well as **caution in effort investments**

Gains are potentially huge. A lot will come from the evolution of the commercial sector on QC. And if a breakthrough happens, our community should be prepared to exploit it.

Possible approach:

- not thinking of a QC facility in Italy. **Cloud-based access to QC** (e.g. Google, IBM, D-Wave) might be an option in our computing landscape. Participation to **national and EU R&D projects on QC** would be strategical to get prepared.
- **follow the Q-technolog(ies) trends, with special attention to potential HEP-specific benefits**. In case HEP advantage happens to be demonstrated, we must be technically able to quickly exploit any usable (cloud access) resource that might eventually be offered
- **develop a culture to stably follow new promising technologies**, and open training programs as appropriate. It might be impossible to recover if you start late

Conclusions / Discussion time

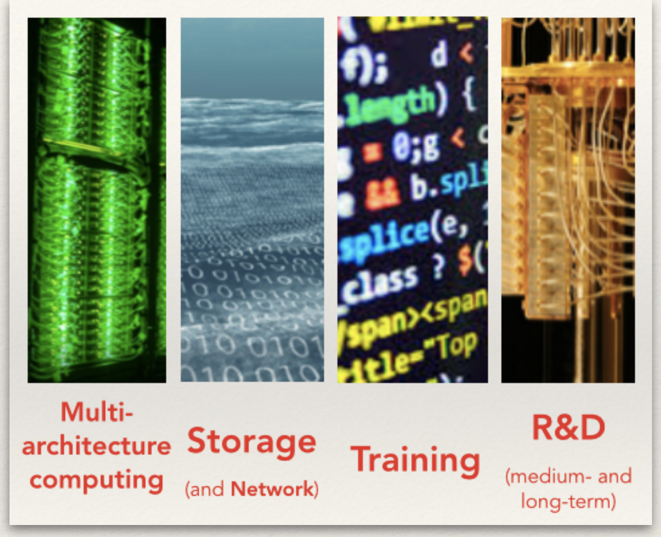
①

- code rewriting challenge
- HEP work with HPC in *design* phase
- cloud / virtualisation as heterogeneity enabler

②

- distributed storage - "data lake" model as strategic
- network over-provisioning as engine for innovation

① ② ③ ④



③

- international-level training
- horizontal knowledge transmission
- career recognition (highly strategic)

④

- medium-term R&D on GPUs, FPGAs, .. and cruciality of testbed resources
- long-term R&D (e.g. on QC) with focus on HEP benefits