HPC challenges for the next years: the rising of heterogeneity and its impact on simulations

Luiz Angelo Steffenel

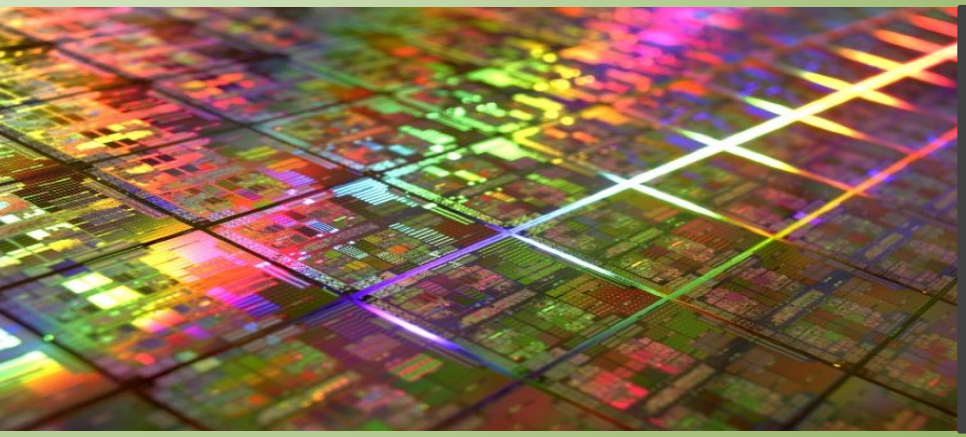
CReSTIC
crestic.univ-reims.fr

HPC Center ROMEO
Centre de Calcul Régional

MAISON DE LA SIMULATION
de Champagne-Ardenne

UNIVERSITE DE REIMS
CHAMPAGNE-ARDENNE

HPC challenges for the next years



The rising of heterogeneity and its impact on simulations

CECAM Workshop
Microscopic simulations: forecasting the next two decades
Toulouse, April 24-26 2019

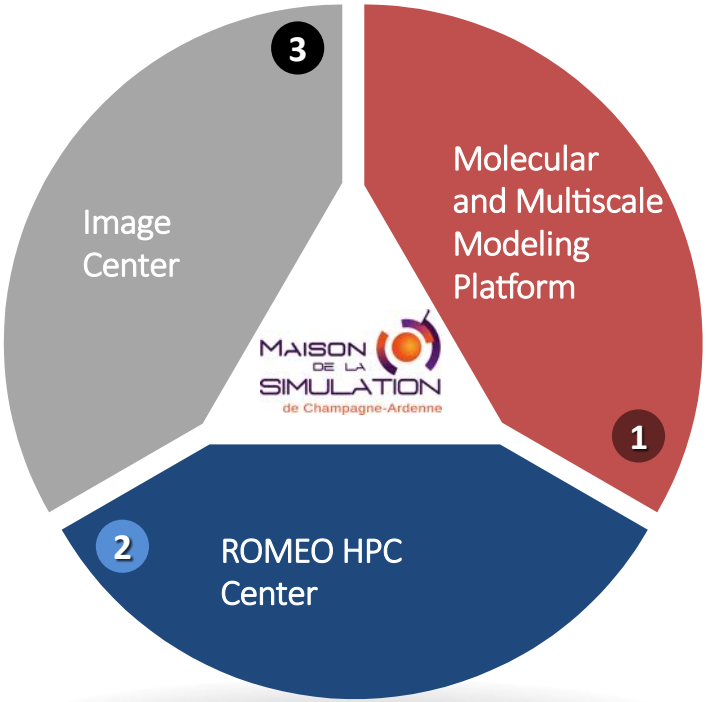
Luiz Angelo STEFFENEL
angelo.steffenel@univ-reims.fr

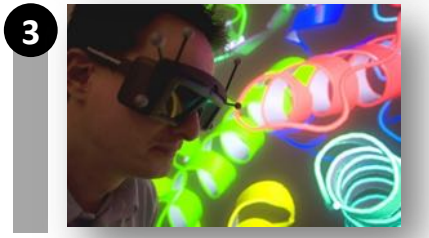
About my team

- Luiz Angelo Steffenel
  - Associate Professor, CReSTIC Laboratory
  - CASH Team (HPC, Autonomous computing, Heterogeneity)
- Our team has a long tradition on HPC
  - ROMEO supercomputing center
  - Part of MASCa

Maison de la Simulation Champagne-Ardenne

More than 15 years associating HPC and applied computing





UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

CReSTIC

HPC Center ROMEO
Centre de Calcul Régional

2013 - Biggest hybrid CPU/GPU cluster in France

- 270 TFlops
- 151th in Top500
- 5th in Green500

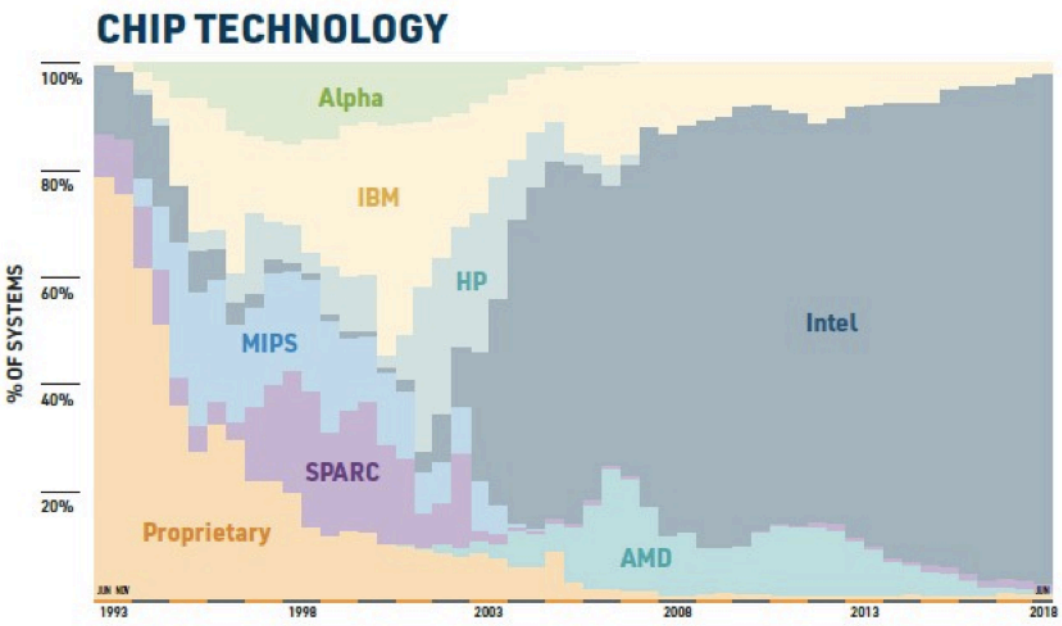
2018 – Biggest academic cluster in France

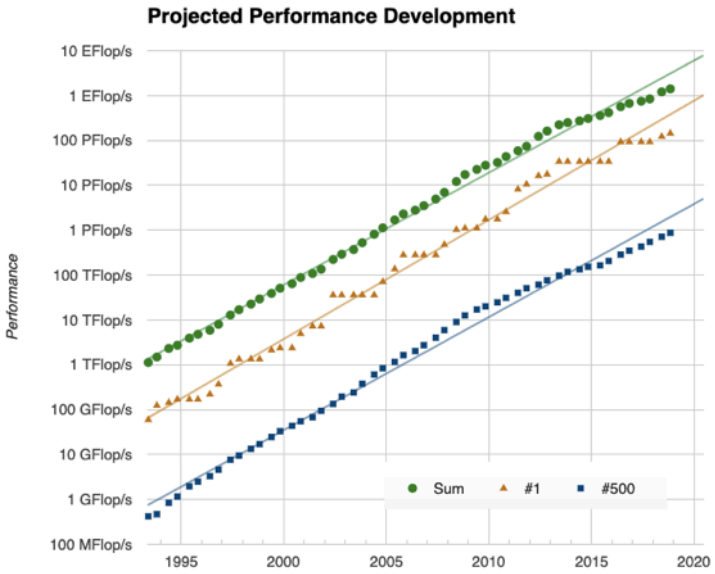
- 1022 Tflops
- 249th in Top500
- 20th in Green500

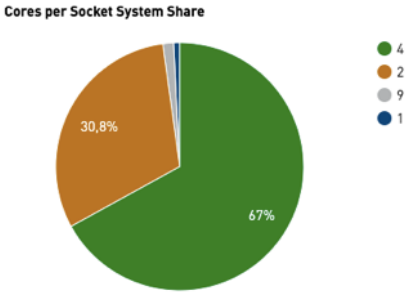
CReSTIC

Top 500 ranking over the time

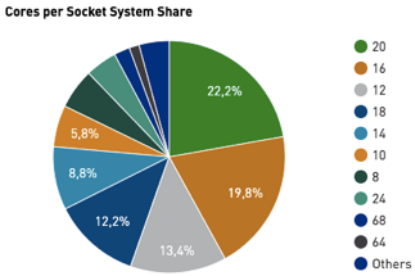
We are in a "calm" period







2008

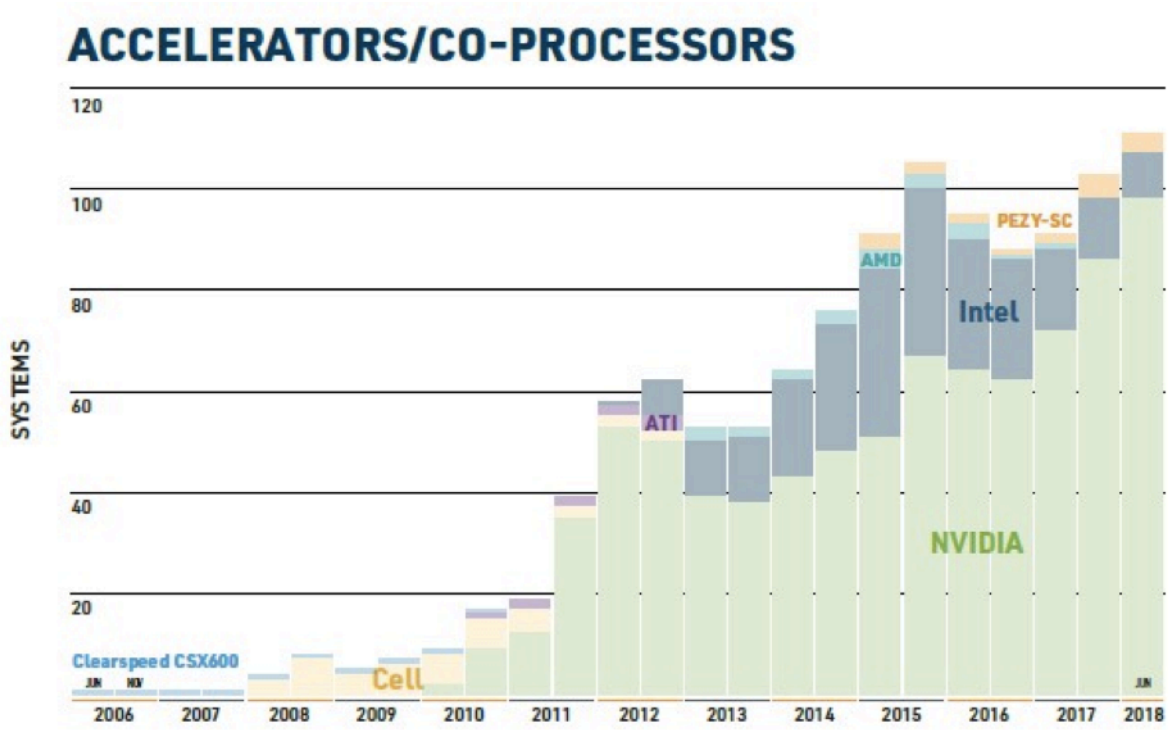


2018

Besides multicore, what is the biggest "innovation" since 2008?

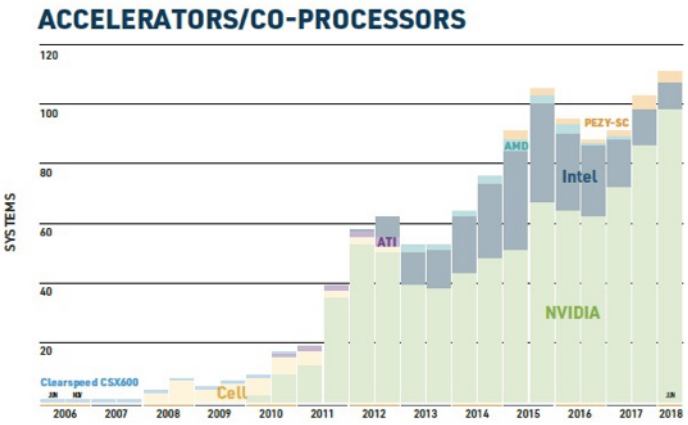
Hybrid architectures

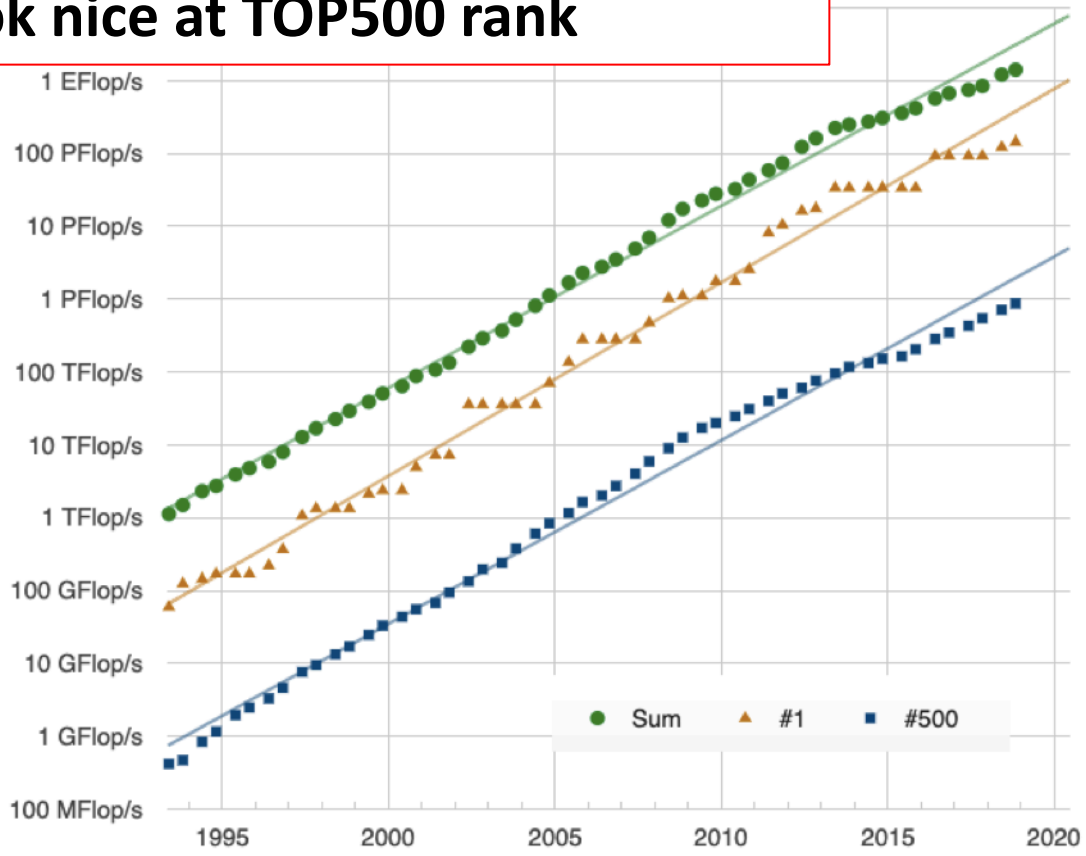
- Mix of CPUs and accelerators
  - GPUs (mostly NVIDIA)
  - Other accelerators (Xeon Phi)



TOP500 – Which is the impact of accelerators?

GPUs as a way to reduce overall costs and look nice at TOP500 rank





Accelerators can deliver extra FLOPS but they add an extra heterogeneity layer → harder to explore

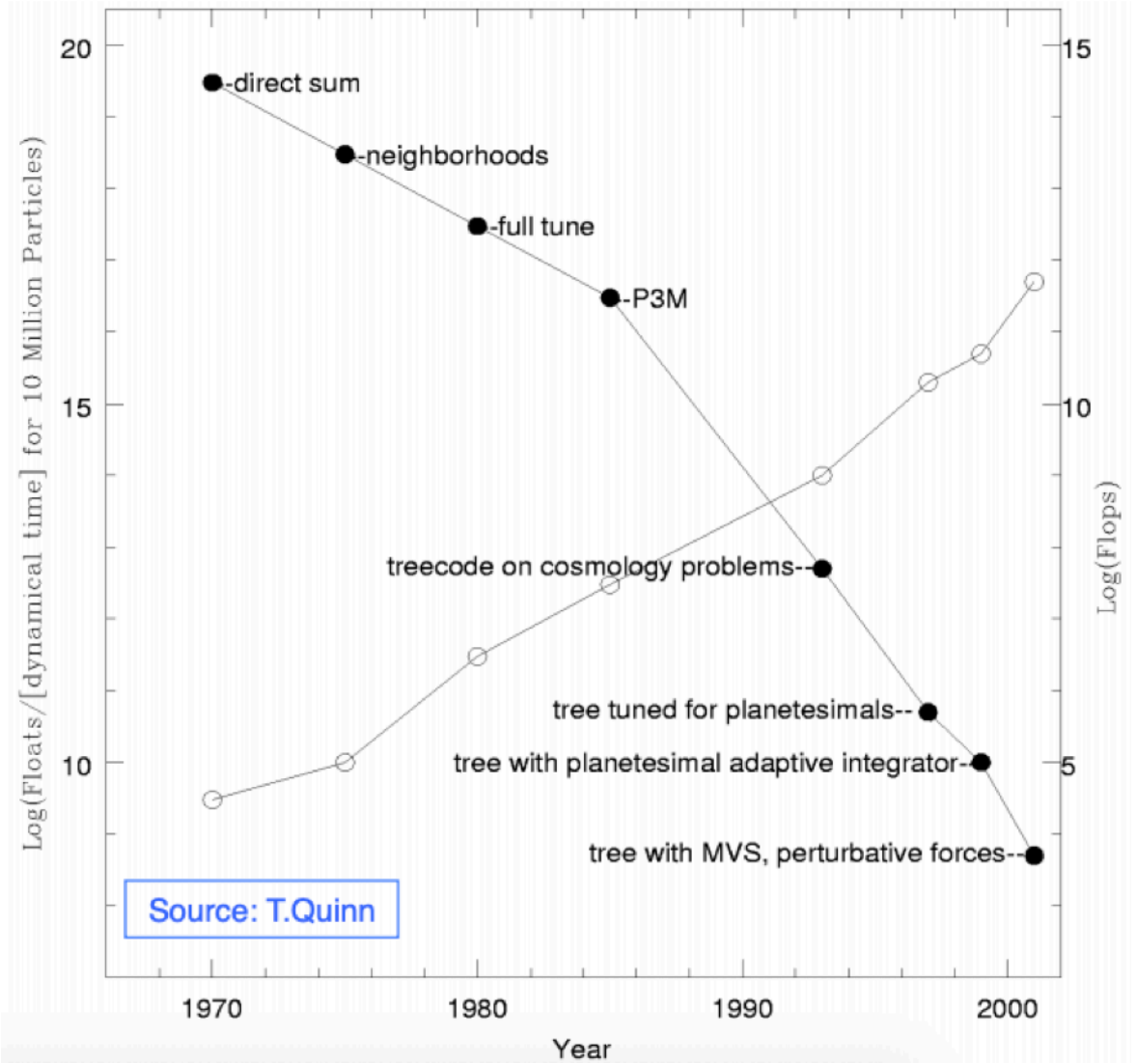
CReSTIC

How to extract more from the hardware?

- GPUs are **good tools**
  - Useful with specific code parts
- Some problems are **intrinsically hard**
  - Hardware evolution helps doing *faster*, but does not reduce complexity
  - Better results only come with additional software development
- Extra hardware = extra complexity

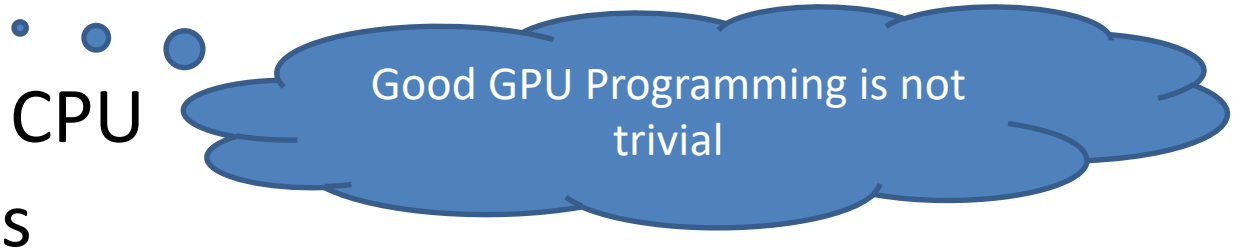
Example: n-body problems

- In 30 years
  - 10^7 hardware
  - 10^{10} software
- Our problem now is that hardware is much more complex
  - Software has to struggle to control it



The cost of Heterogeneity

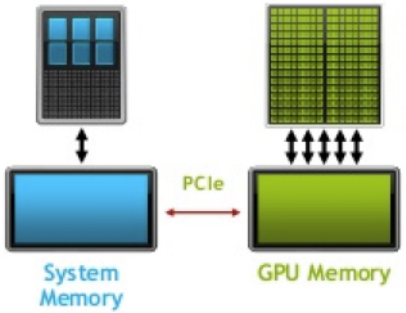
- Most of our programming models are 20+ years old (MPI, OpenMP, etc.)
  - Designed for homogeneous environments
    - Node-node, CPU-CPU, CPU-memory
- Current HPC has several layers
  - GPUs
  - Cores in a CPU
  - Multi CPUs
  - Multiple layers of memory (cache, RAM, etc.)
  - Interconnections



CReSTIC

So what are GPUs good for?

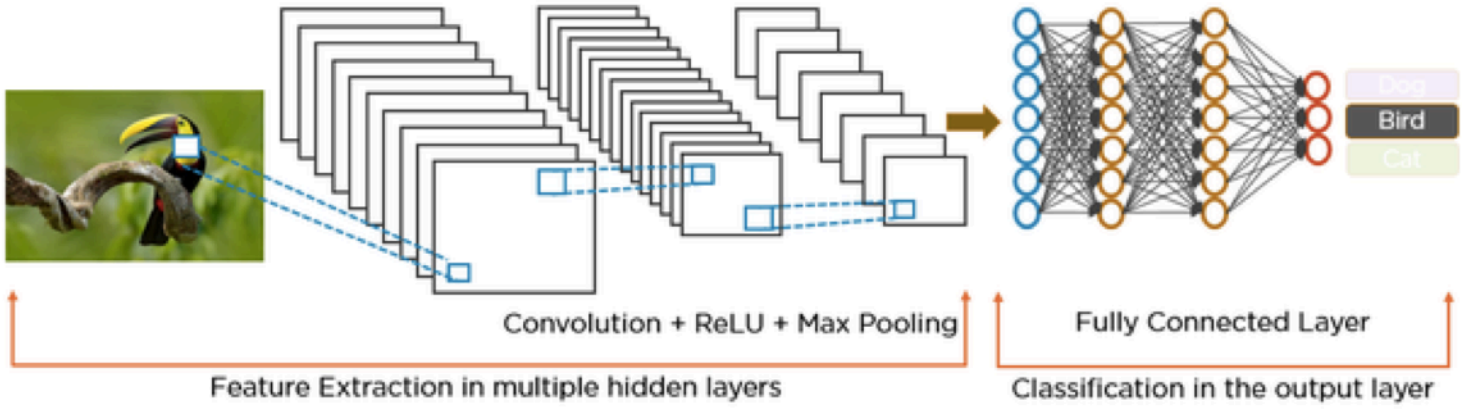
- As a "piece of hardware", GPUs are no more special than co-processors for i386/i486
- Early HPC developments with GPUs started by exploring their parallel processing capabilities (SIMT)
  - GROMACS ✅
  - Fluent ✅
  - OpenFOAM ✅
  - Autodock with GPU ❌
- Performance gains limited by memory and latency constraints
- Hard to code (CUDA, OpenCL, ...)



CReSTIC

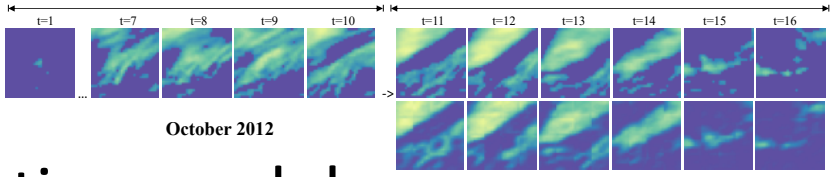
The revival of Neural Networks

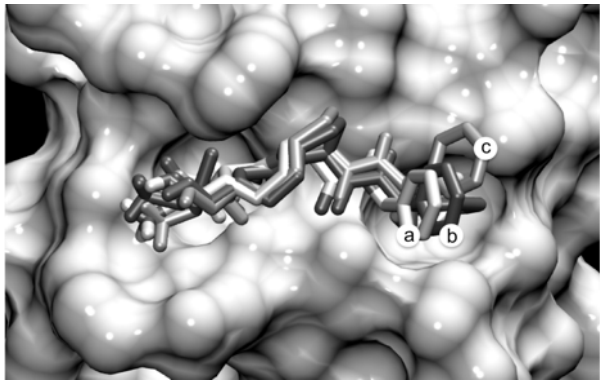
- GPUs are well-suited for the matrix/vector math involved in machine learning
  - Especially the famous ***Deep Learning***
  - Data is often provided as a matrix of pixels
    - Or matrices of n-dimensions called "tensors"
    - The work can be split in several parallel tasks
    - Data is kept in the GPU memory for a long time



Is AI the future of HPC?

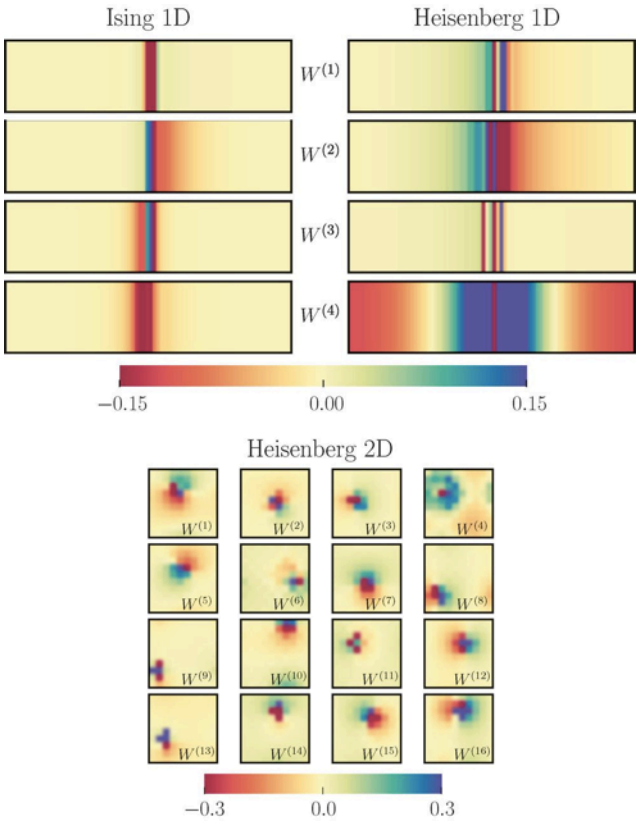
- Once again, it's a good tool, not the answer
- AI can help us to speed up simulations
- What AI can do for us?
  - **Unveil correlations**
    - Help improve the simulation models
    - Ex: meteorological models
  - **Identify/reproduce patterns**
    - Fill the gap between simulations steps
      - Ex: molecular docking





Ex: quantum many-body problem

- Microsoft and ETH project
- Use neural networks to represent the wave function and reduce the computing complexity
- AI does not replace the simulation models, just accelerate some steps



- https://science.sciencemag.org/content/355/6325/602

AI + Simulation = Synthesis Models

AI+HPC WORKFLOW FOR ENHANCEMENT MODELING

Using Simulation Data To Train AI- Fermilab NOvA

Simulation

4.7M simulated events

Simulation Apps.

FLUKA

CRY Cosmic-Ray

Genie

GEANT

DL Training

3.7M events for training

CVN-Trained for 1 week

DLT Frameworks

Caffe

Inference/Testing

1M events for Testing

Neutrino detection efficiency increased by 33%

1 NVIDIA

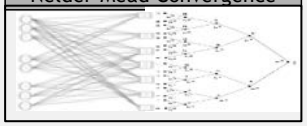
Source: NVIDIA

AI+HPC WORKFLOW FOR MODULATION

AI-led Experiment To Converge Faster-bose Einstien Condensate

Conventional Approach

Nelder-Mead Convergence



1. Slowly adjusts each parameter

2. BEC reached in 140 experiments

AI Synthesis Modeling

BEC

ML Training

1. Trained on 10 random experiments

AI-Led Convergence

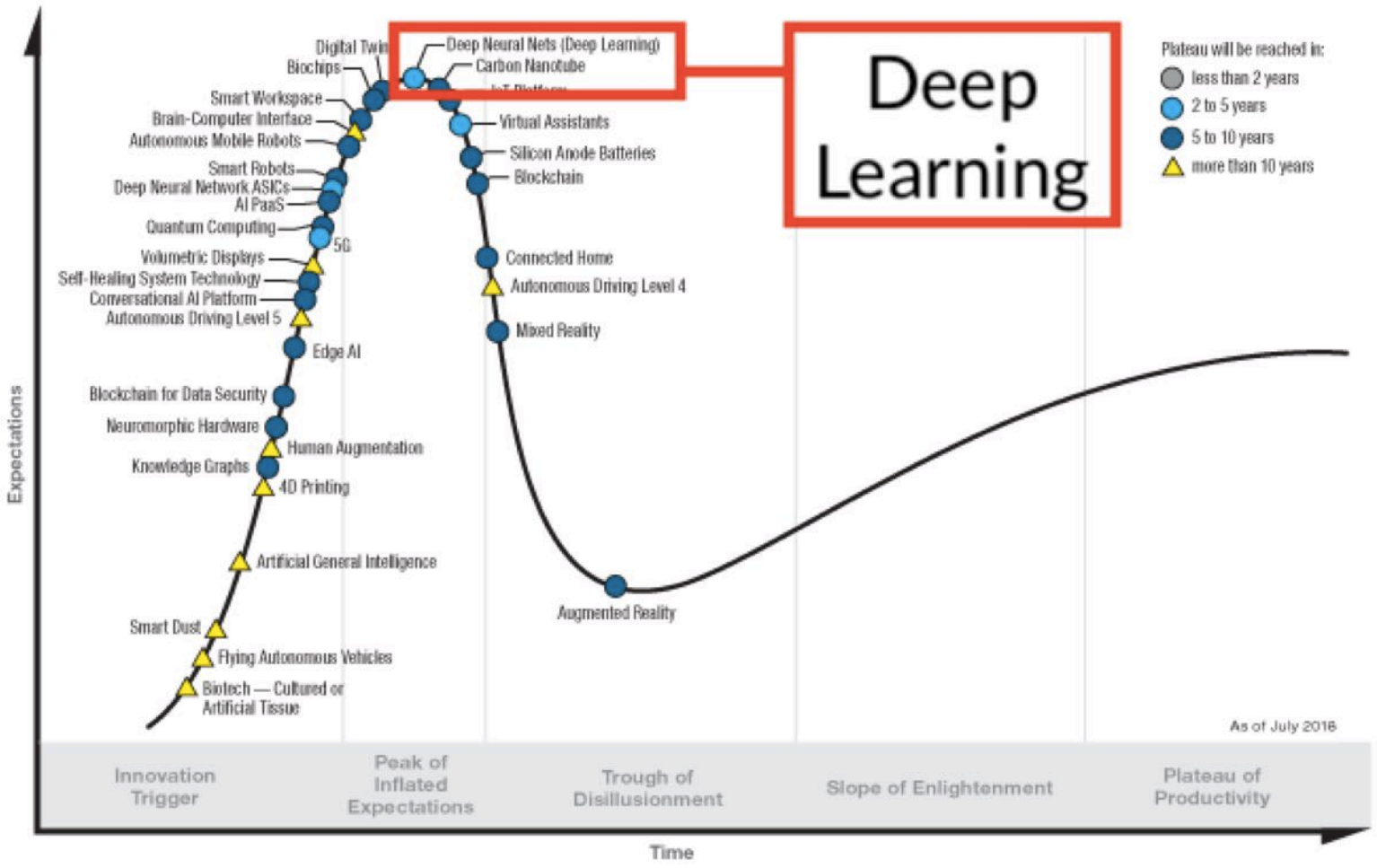
2. AI-Led experiments to converge on BEC

3. BEC reached in 10-12 experiments

2 NVIDIA

CReSTIC

Hype Cycle for Emerging Technologies, 2018

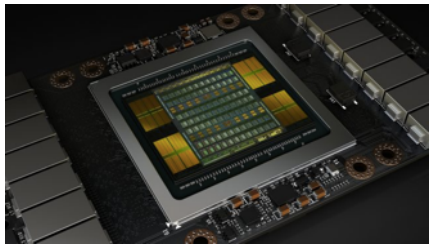


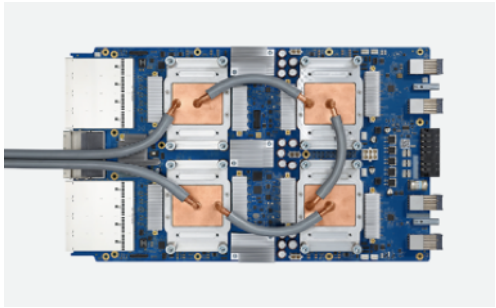
gartner.com/SmarterWithGartner

Gartner

Can we rely on GPUs for general computing?

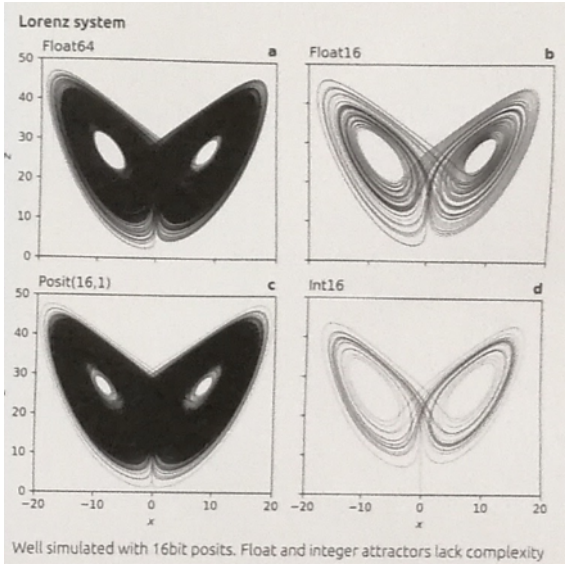
- Trends for NVIDIA/AMD
  - 7nm or less, energy constraints, interconnection speed, but...



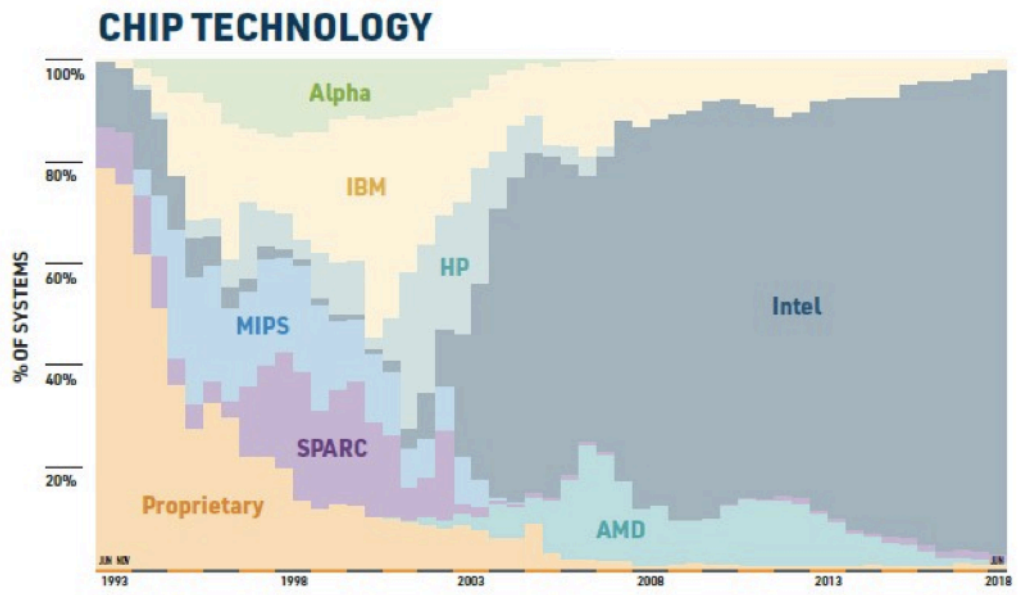


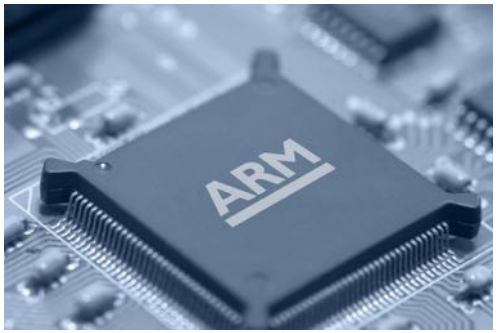
- More and more dedicated for AI
  - Ex: TPUs from Google
  - Autonomous cars (Tesla, etc)

- WARNING: all GPU development points towards **mixed precision**
  - **Faster,** acceptable precision
  - Not adapted for all problems

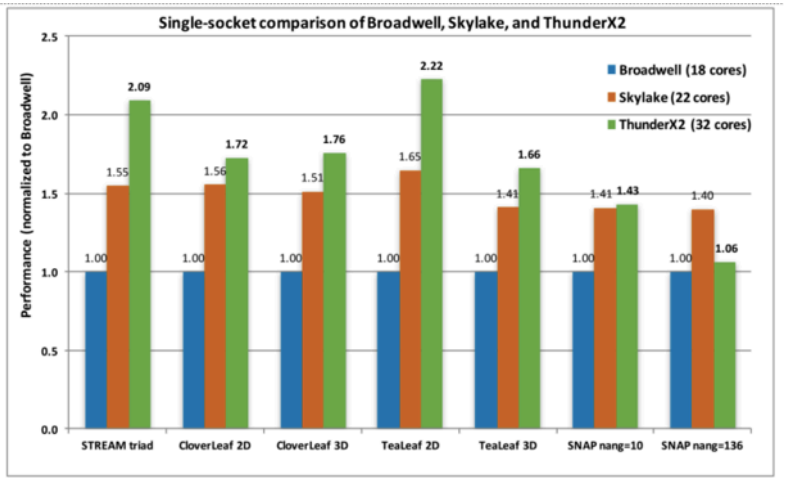


Even the CPUs are changing





- Arrival of ARM processors on the HPC market
  - Just an European dream?

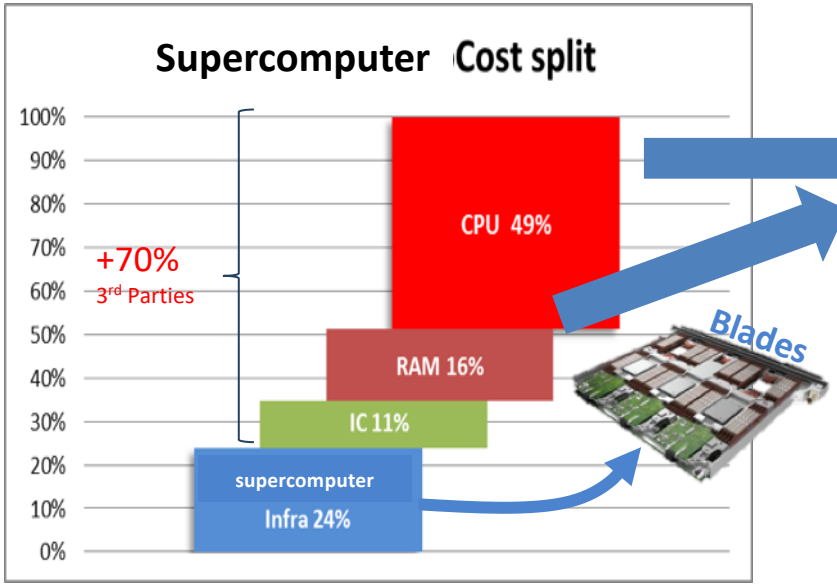


European processor Initiative (EPI)

Sovereignty

Independence from US components

Cost savings



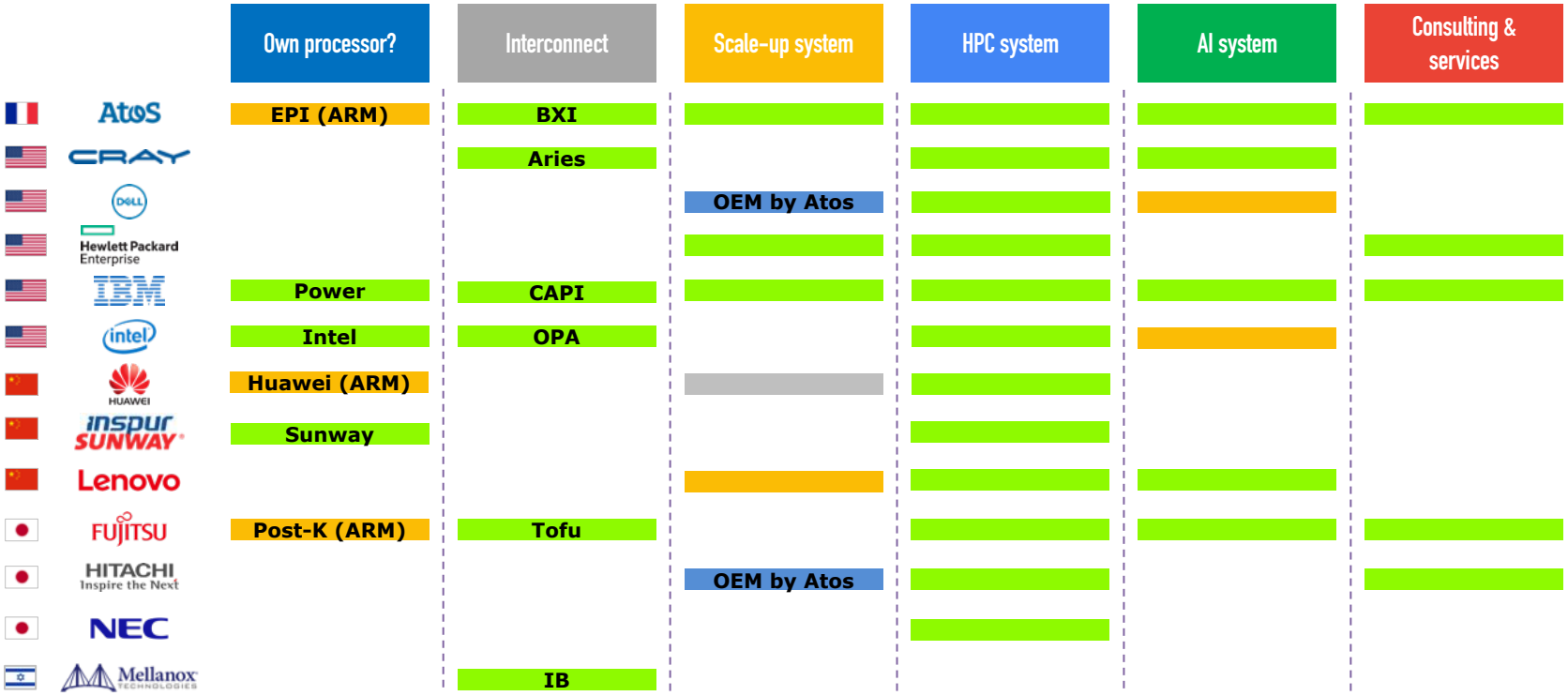
CPU + Memory = 65% of BOM

European Processor Initiative

Developing a pan-European supercomputing infrastructure
Public Members: 1 B€ ; EU Financial: 486 M€; Private partners 400 M€

CReSTIC

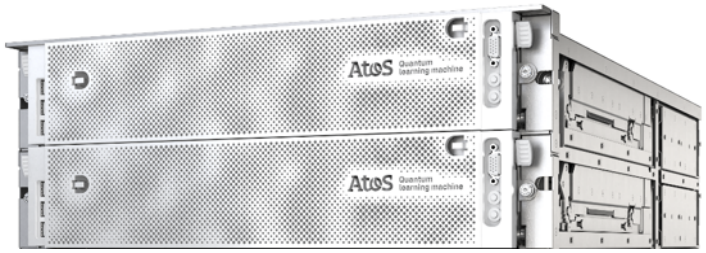
Europe can provide most of the elements



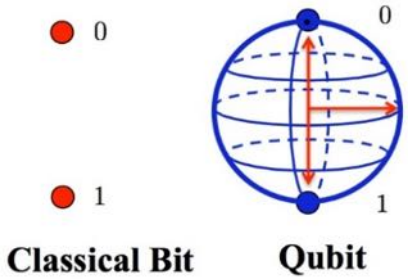
Lacks only a good GPU 😉

And what about Quantum Computing

- Potential to solve difficult problems
  - Classical bit VS Qubits
- Only a few "real" quantum computers
  - Mostly simulators
  - Ex: QLM (ATOS + partners)

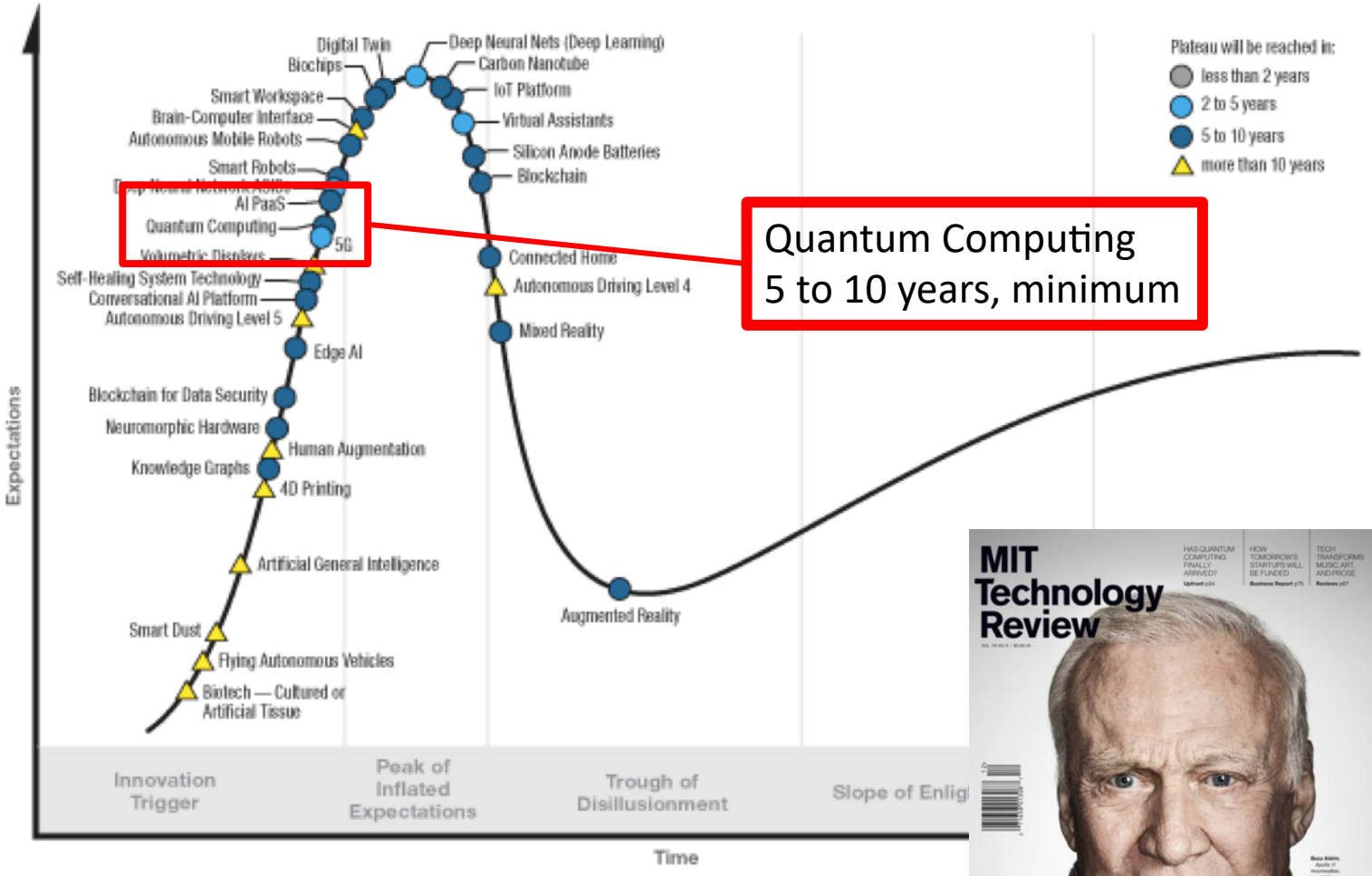


- Develop new algorithms
  - The "logic" is not the same
- Designing computing architectures
  - Many challenges on memory access, interconnection



CReSTIC

Hype Cycle for Emerging Technologies, 2018



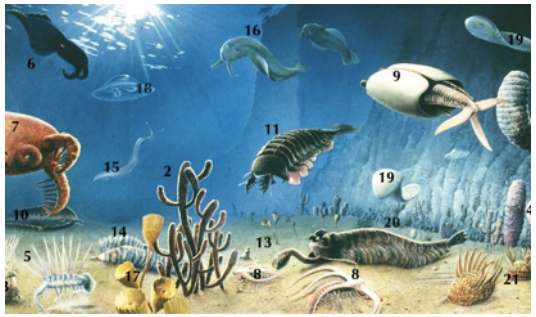
Quantum Computing
5 to 10 years, minimum

gartner.com/SmarterWithGartner

Source: Gartner (August 2018)

Some Conclusions and Forethoughts

- After a calm period, HPC is facing a new "Cambrian explosion" due to hardware heterogeneity
- HPC software is still bound to 2000's methods → not enough!!!
- GPUs have driven developers towards a risky path
  - Architecture-dependent
  - Low-level programming
  - This has a price
- AI is not the "holy grail"
  - Neither Quantum Computing
- **The next years will be agitated!**



CReSTIC

Thanks!