Exascale challenges

Christophe CALVIN

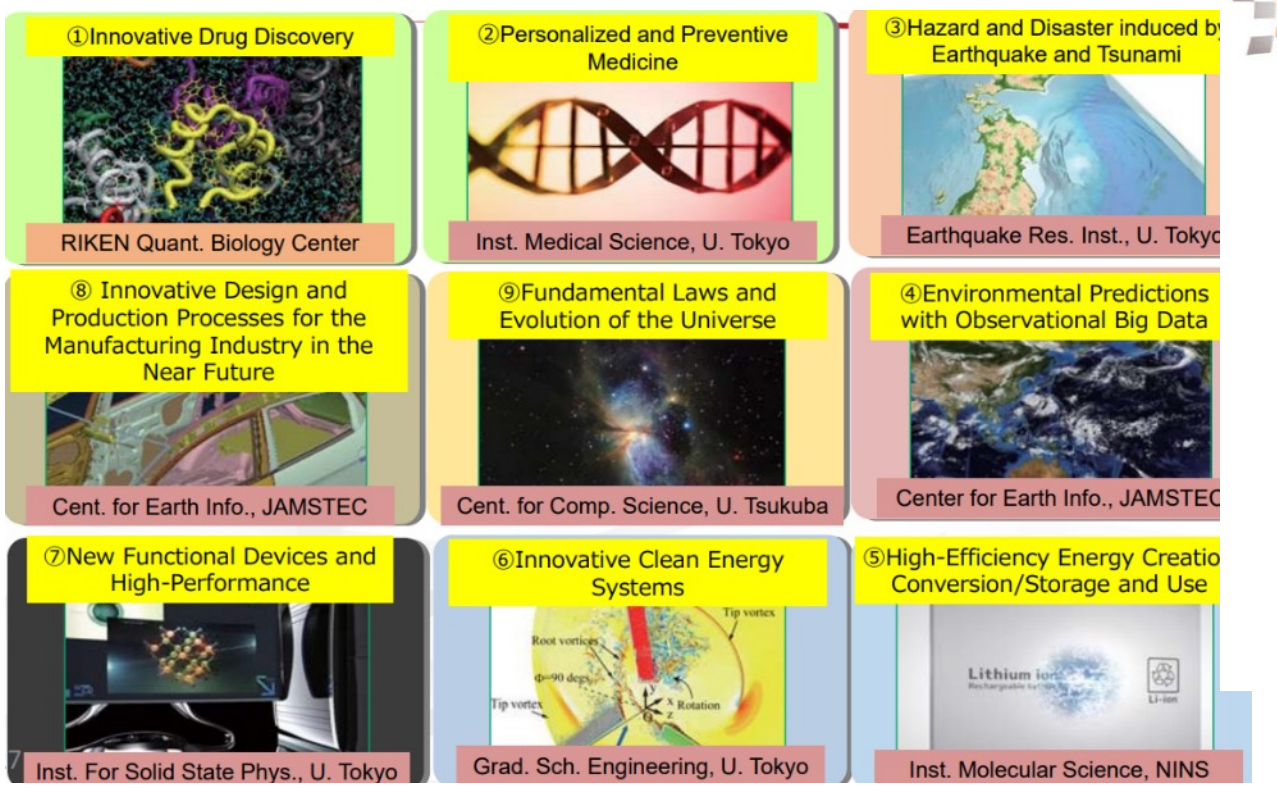
CEA/DRF

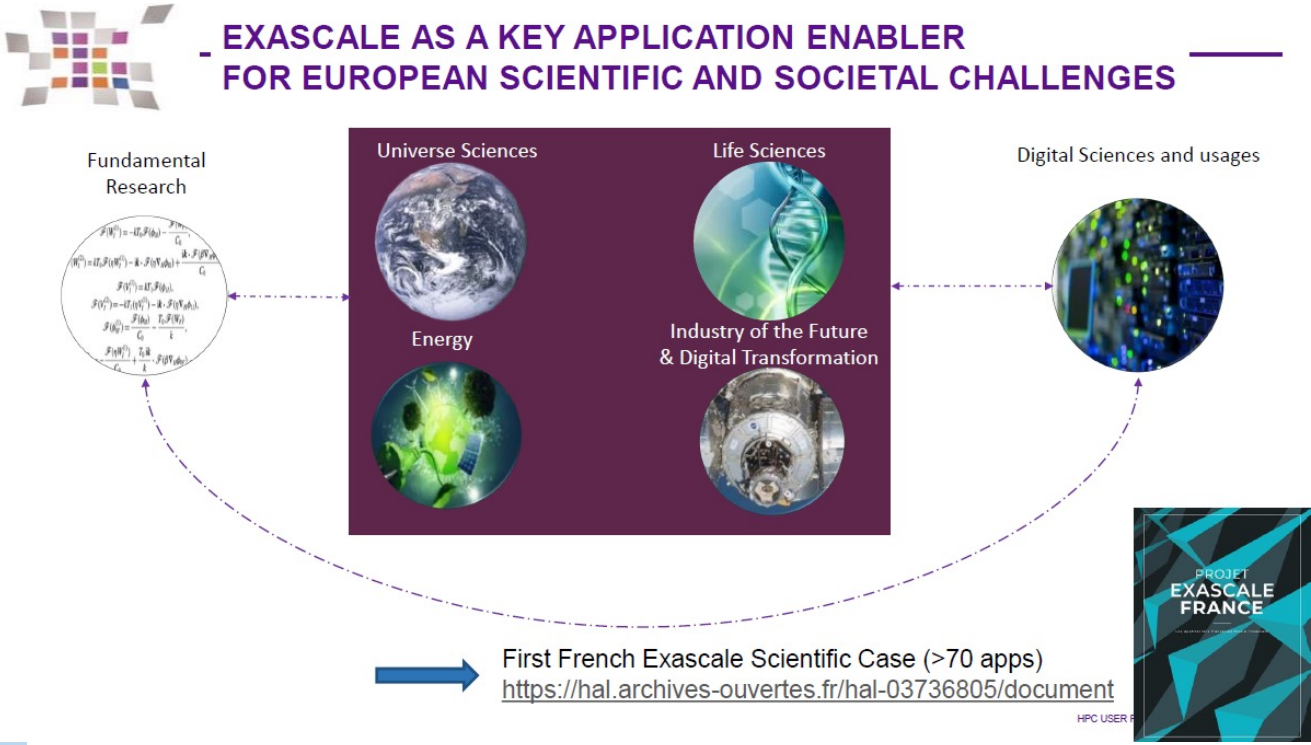
Deputy to the Director of Fundamental Research at the CEA in charge of HPC and simulation

What exascale? And for what?

Prefix (symbol E) which, placed in front of a unit, multiplies it by 10^{18}.

- Compute: Exaflops → 10^{18} floating points operations per seconds Eflops/s
- Storage: Exabyte → 10^{18} bytes



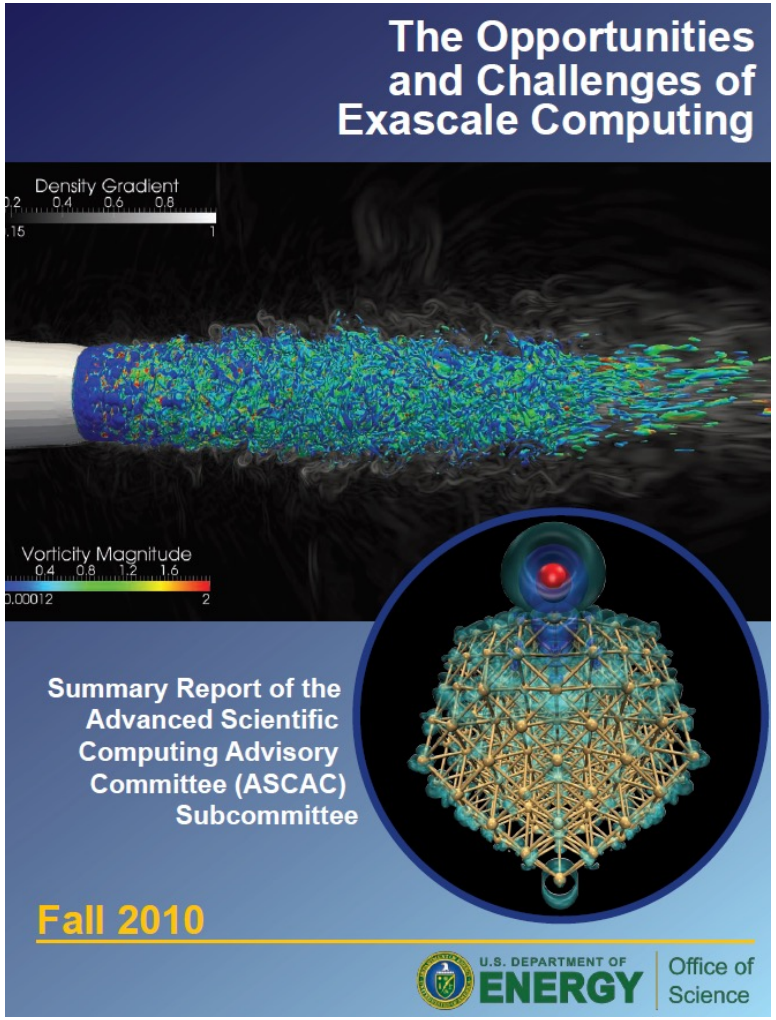


First French Exascale Scientific Case (>70 apps)
https://hal.archives-ouvertes.fr/hal-03736805/document

Technological challenges

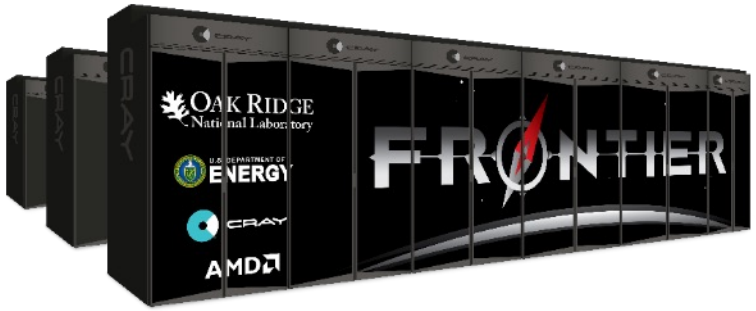
How to design and operate such a machine within a sustainable energy envelope?

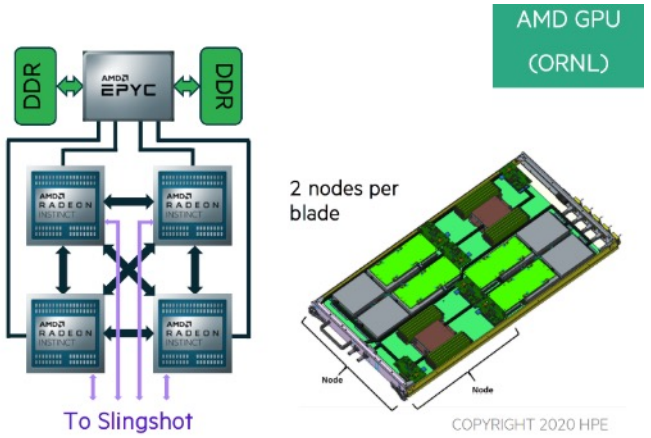
	2010	2018	Factor Change
System peak	2 Pf/s	1 Ef/s	500
Power	6 MW	20 MW	3
System Memory	0.3 PB	10 PB	33
Node Performance	0.125 Gf/s	10 Tf/s	80
Node Memory BW	25 GB/s	400 GB/s	16
Node Concurrency	12 cpus	1,000 cpus	83
Interconnect BW	1.5	50 GB/s	33
System Size (nodes)	20 K nodes	1 M nodes	50
Total Concurrency	225 K	1 B	4,444
Storage	15 PB	300 PB	20
Input/Output bandwidth	0.2 TB/s	20 TB/s	100



Technological challenges

	2010	2018	2022 (FRONTIER)
System peak	2 Pf/s	1 Ef/s	1.5 Ef/s
Power	6 MW	20 MW	22.7 MW
System Memory	0.3 PB	10 PB	37 PB
Node Performance	0.125 Gf/s	10 Tf/s	166 Tf/s
Node Memory BW	25 GB/s	400 GB/s	
Node Concurrency	12 cpus	1,000 cpus	64 CPU cores + 880 GPU cores
Interconnect BW	1.5	50 GB/s	100 GB/s
System Size (nodes)	20 K nodes	1 M nodes	9,472 nodes
Total Concurrency	225 K	1 B	9 B
Storage	15 PB	300 PB	500 / 1,000 PB
Input/Output bandwidth	0.2 TB/s	20 TB/s	5 / 10 TB/s





How to design and operate such a machine within a sustainable energy envelope? → **GPU**

TOP 500 List – Nov 2023

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE	8,699,904	1,194.00	1,679.82	22,703
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel	4,742,808	585.34	1,059.33	24,687
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft	1,123,200	561.20	846.84	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE	2,752,704	379.70	531.51	7,107
6	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN	1,824,768	238.70	304.47	7,404
7	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM	2,414,592	148.60	200.79	10,096
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 40C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR200, EVIDEN	680,960	138.20	265.57	2,560
9	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, Nvidia	485,888	121.40	188.65	
10	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox	1,572,480	94.64	125.71	7,438

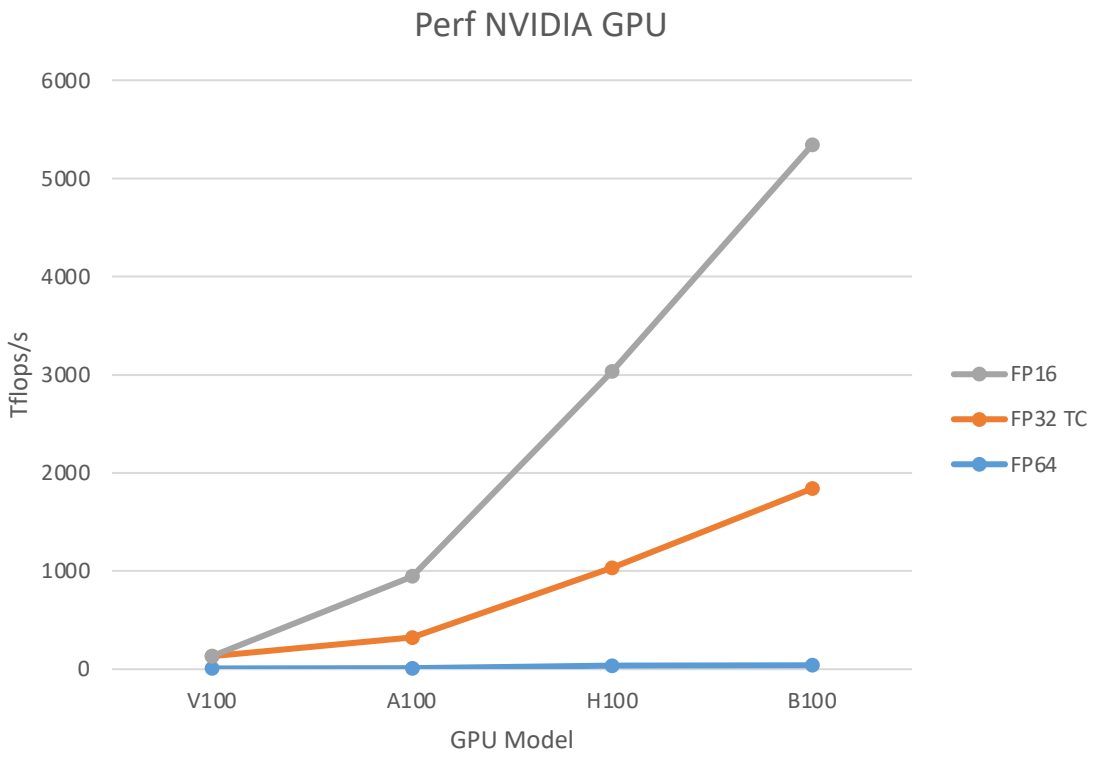
100% CPU

>90% of total performance is based on GPUs (NVIDIA and AMD)

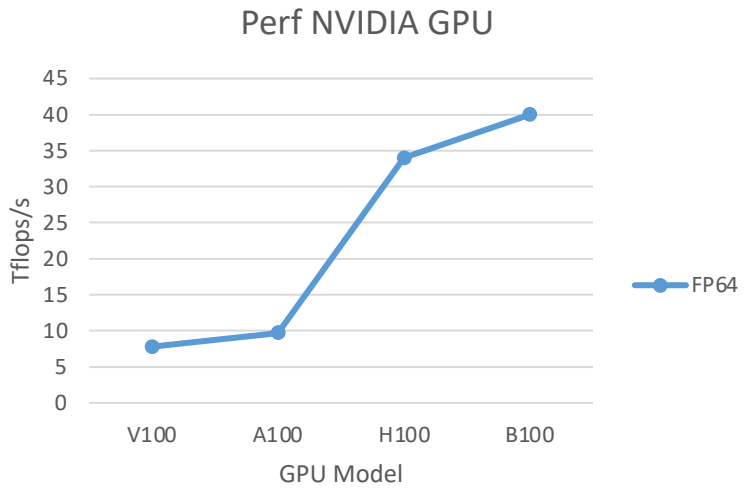
We talk about Peak and HPL perf. Not on real scientific applications

GPU from video games to AI

- Initially GPU have been designed for video games! → graphics
- Used for HPC (CUDA): excellent ratio Gflops/Watt
- Now GPU market: AI (and Gen AI) – FP64 is quite useless for AI workloads (Learning and inference)



- GPU: not especially designed for numerical simulation
- Not really easy to program – CUDA/OpenCL – explicit transfer from host→device
- Increase FP32/FP16/FP8/FP4 performances for AI and stagnation of FP64 perf

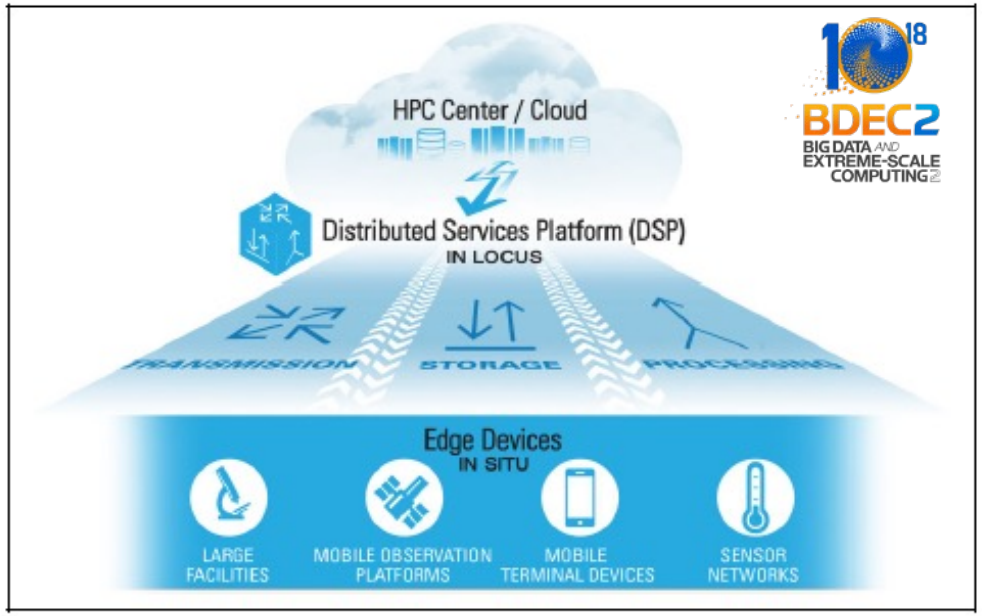


And what about usages?

Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry

The International Journal of High Performance Computing Applications
2018, Vol. 32(4) 435–479

DOI: 10.1177/1094342018778123
journals.sagepub.com/home/hpc
SAGE



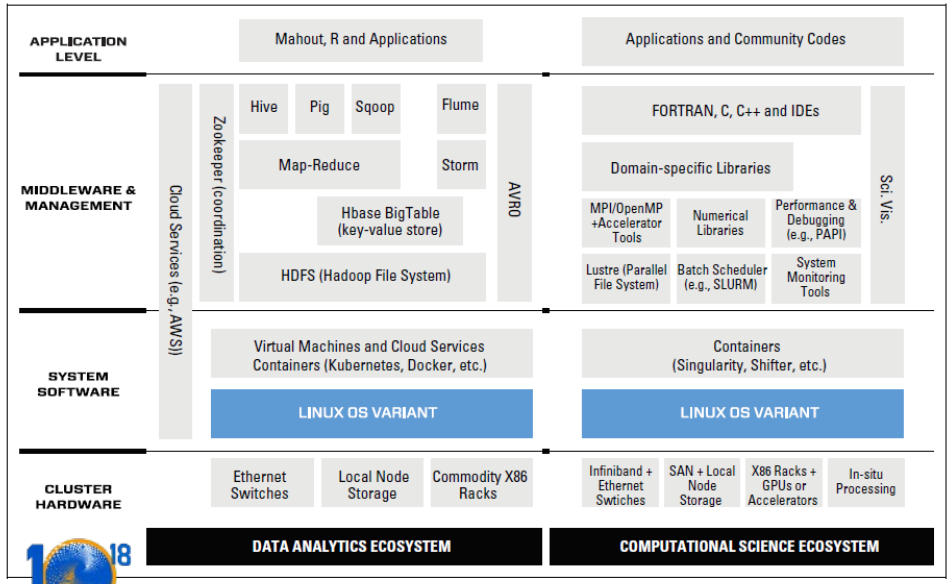
- Not only pure HPC (numerical simulation)
- More and more HPDA: treatment and analysis coming from large facilities and IoT
- Explosion of AI (and especially GenAI)

- *ChatGPT-3: 175 Billions of parameters - 3 months - 8 000 GPU Hopper – 15 MW mégawatts*
- *Chat GPT-4: 1 Trillion of parameters*

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

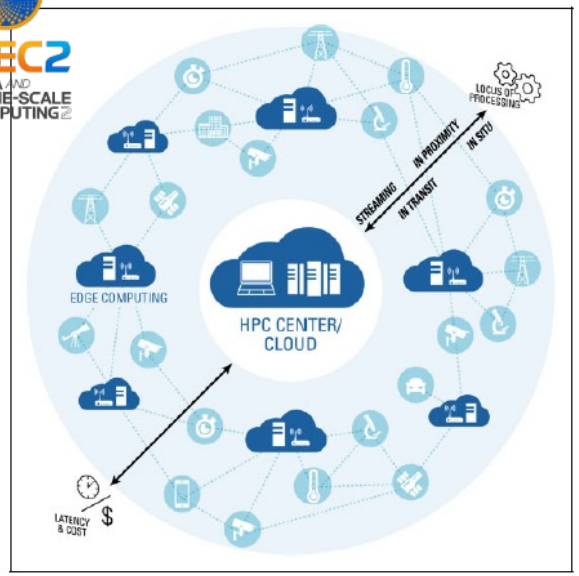
DataSets
Size of DataSets

What impact?



- SW stack: both for data and HPC
- FS: large data sets / small files / high performance (LUSTRE is not the only solution)
- Cloud access – HPC center is not “fortress” anymore
- HPC center not anymore in the center! A link in the chain

Exascale is not only a HW concern but need to rethink the usage of supercomputers, datacenter architecture and access to supercomputers



Exascale in Europe

EUROHPC, MAIN DIRECTIONS FOR #2 REGULATION (2021-2033)

2 main technological directions, 7B€ for the second phase

EuroHPC Joint Undertaking

➢ **Infrastructure deployment**

#EuroHPC Joint Undertaking

	2019 & 2020	2021	2022	2023	2024	2025	2026	2027
HPC Infrastructure	pre-exascale + petascale HPC systems	Several pre-exascale systems and exascale HPC systems				exascale and post-exascale HPC systems		
Quantum Infrastructure	Quantum simulators interfacing with HPC systems	1st generation of quantum computers + quantum simulators interfacing with HPC systems				2nd generation of quantum computers + quantum simulators		

1st Exascale CFEI published by EuroHPC on dec 2021
-> Selection of the German proposal (at FZJ) called JUPITER
2ème Exascale CFEI expected for end 2022

1sr CFEI for quantum computers from EuroHPC
-> EuroQCS-France application (with FR, GE, IE et RO) using photonics based solution integrated into HPCQS
-> results expected in October

➢ **Successive R&D calls for proposals including :**

- **an Exascale pilot** – Consortium EUPEX (lead Atos) EU PEX
- A quantum pilot – Consortium HPCQS ⟨HPC|QS⟩

HPC USER FORUM | 04/10/2022 | 7

LUMI: 540 Pflops

LEONARDO: 314 Pflops

MareNostrum5: 312 Pflops

PreExa

Meluxina: 18Pflops

Karolina: 13Pflops

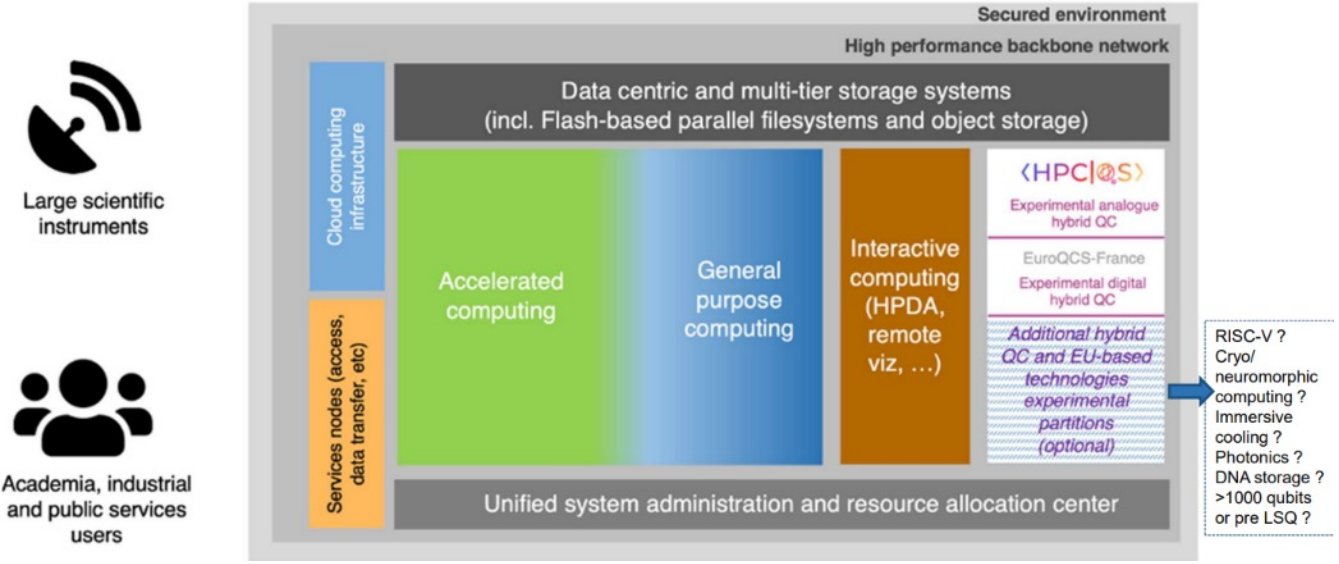
Discoverer: 6Pflops

Peta

Exascale in France: Jules Verne project

EXASCALE SYSTEM ARCHITECTURE OVERVIEW

Possible reference designs



Organization of the french application

- GENCI *Hosting Entity*
- CEA *Hosting Site*
- SURF (NL) as member of consortium

Name of the consortium/supercomputer : Jules Verne

Full TCO over 5 years : 542 € (50% EuroHPC, 50% consortium)

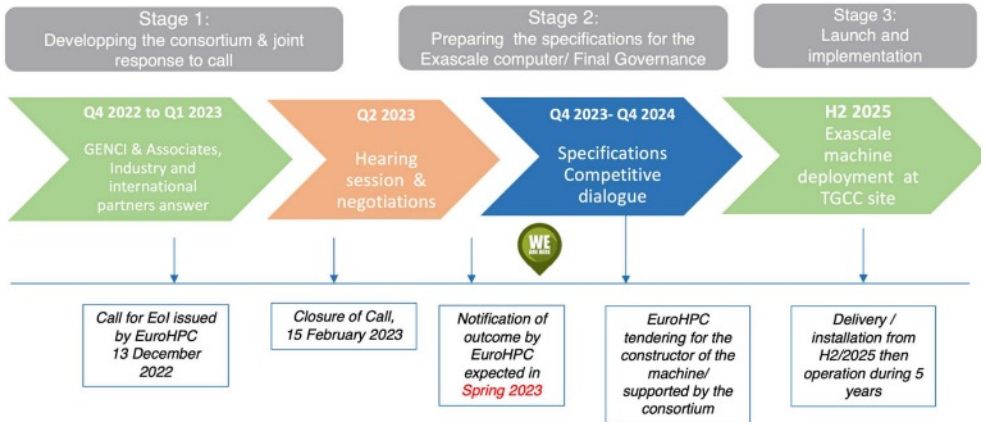
- French public contribution
- NL contribution

-Seeking more partners on the consortium to reach 300M€

- International partners
- French research institutions
- French industrial partners (end users)

JULES VERNE PROPOSAL – NEXT STEPS

From Call for Expression of Interest (CEI) to commissioning the Exascale supercomputer



Global performance targets

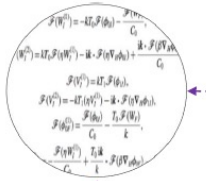
Sustained HPL performance = 1 Eflops
Composition : 60% accelerated nodes, 40% scalar nodes
but accelerated nodes will bring > 90% peak performance
>100 PB Flash/HDD and > 200 PB archive

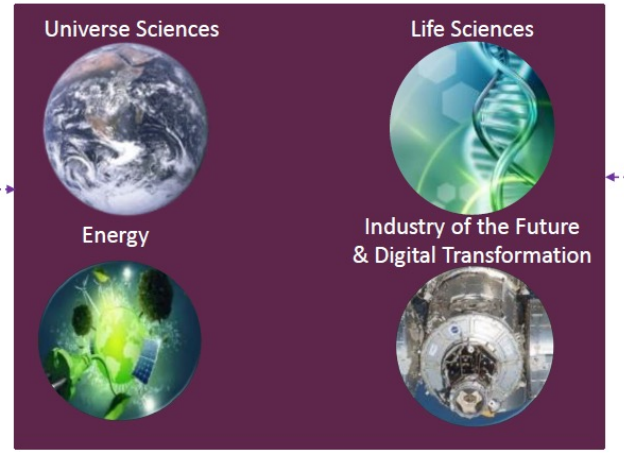
Estimated Total cost of ownership over 5 years ~ 500 M€
Power consumption < 20 MW

Exascale in France: support for application communities

EXASCALE AS A KEY APPLICATION ENABLER FOR EUROPEAN SCIENTIFIC AND SOCIETAL CHALLENGES

Fundamental Research



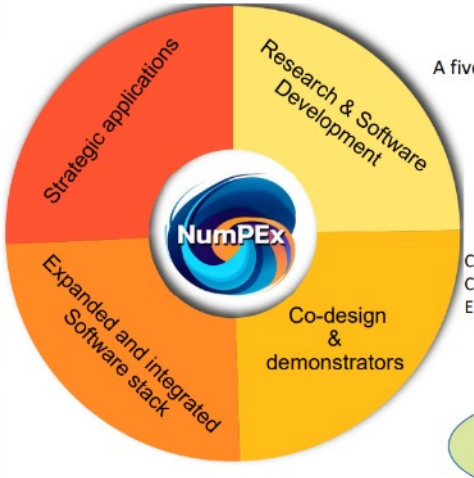


Digital Sciences and usages

First French Exascale Scientific Case (>70 apps)
https://hal.archives-ouvertes.fr/hal-03736805/document

Toward Exascale applications and usages : The NumPEx project

- Aggregate the French HPC/HPDA/IA community
- Contribute and accelerate the emergence of a European sovereign exascale software stack and strategic applications exascale capability in a coherent and multi-annuel framework
- Integrate and validate co-designed innovative methods, libraries and software stack with demonstrators of strategic applications.
- Accelerate science-driven and engineering-driven developers training and software productivity



5 Years
40,8 M€

A five-year, € 40,8 M budget

Core Research Institutions

Core Research Institutions: CNRS, CEA, INRIA, Universities Engineer schools, Industry

80 R&D teams
500 Researchers

12 PY for 5 years L3/L4 support funded by the project

Collaborations – National → International ecosystem

EU/JPN Collaboration
HPC/AI – Climate/Material/Biomedical

ADAC

Accelerated Data Analytics and Computing Institute

Int. collaboration on HPC/AI/HPDA/QC

NumPEx

PEPR – Software Stack

AIDAS

AI Data Analytics and Scalable Simulations

JÜLICH Forschungszentrum

European Lab for exascale
HPC/HPDA/AI/QC

RIKEN

Collaboration on HPC/AI/HPDA/QC

“ Thanks !

Questions?