L'ère du calcul exascale

une opportunité à saisir pour l'astrophysique

Geoffroy Lesur (IPAG Grenoble)
CNRS researcher
President of the Astrophysics & Geophysics committee (CT4) @ GENCI

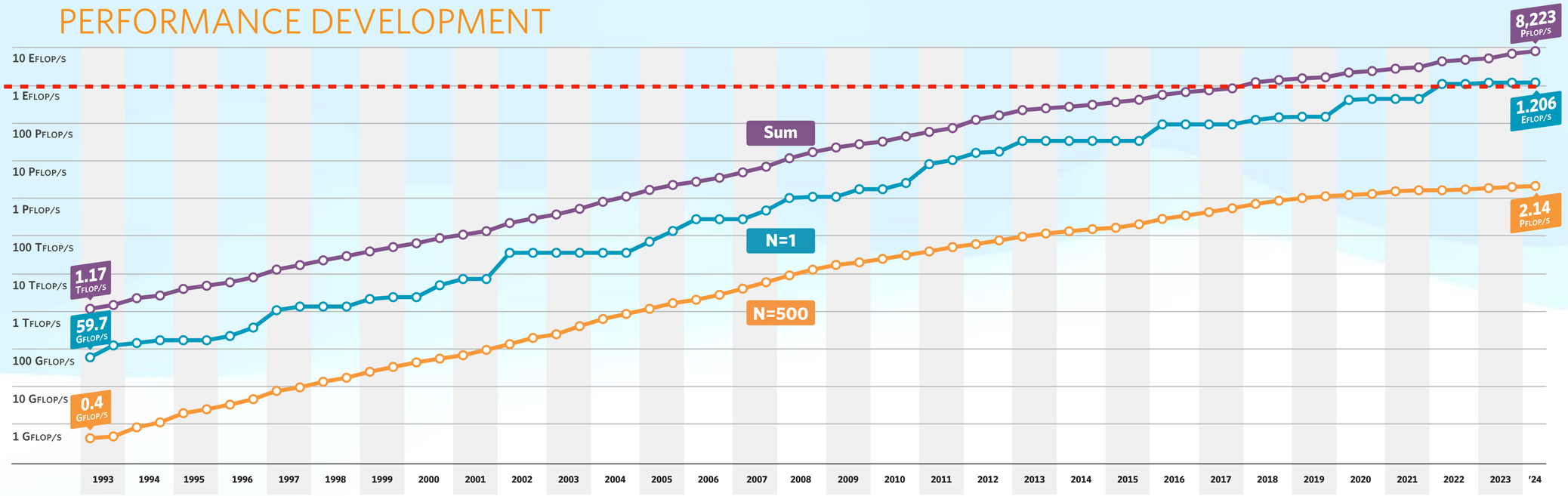
Part I: Exascale supercomputers

1 Exaflop = 10^{18} floating point operation per second

Top #500 super computers in the world

TOP 500 The List.

PERFORMANCE DEVELOPMENT



MAY 2024

			SITE	COUNTRY	CORES	RMAX PFLOP/S	POWER MW
1	**Frontier**	HPE Cray EX235a, AMD Opt 3rd Gen EPYC (64C 2GHz), AMD Instinct MI250X, Slingshot-11	DOE/SC/ORNL	USA	8,699,904	1,206.0	22.7
2	**Aurora**	HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 (52C 2.4GHz), Intel Data Center GPU Max, Slingshot-11	DOE/SC/ANL	USA	9,264,128	1,012.0	38.7
3	**Eagle**	Microsoft NDv5, Xeon Platinum 8480C (48C 2GHz), NVIDIA H100, NVIDIA Infiniband NDR	Microsoft Azure	USA	1,123,200	561.2	
4	**Fugaku**	Fujitsu A64FX (48C, 2.2GHz), Tofu Interconnect D	RIKEN R-CCS	Japan	7,630,848	442.0	29.9
5	**LUMI**	HPE Cray EX235a, AMD Opt 3rd Gen EPYC (64C 2GHz), AMD Instinct MI250X, Slingshot-11	EuroHPC/CSC	Finland	2,220,288	379.7	6.01

Towards Exascale Machines

An electrical power issue

=?

- The current average power consumption of a x86 processor (your laptop!) is ~2Gflop/s/W
- an Exaflop machine would need about 500 MW of power=1 french nuclear reactor
- The « socially acceptable » power consumption is 30 MW → reduce the energy footprint by a factor 6

Accelerated clusters (GPU-like) are mandatory

Energy efficiency

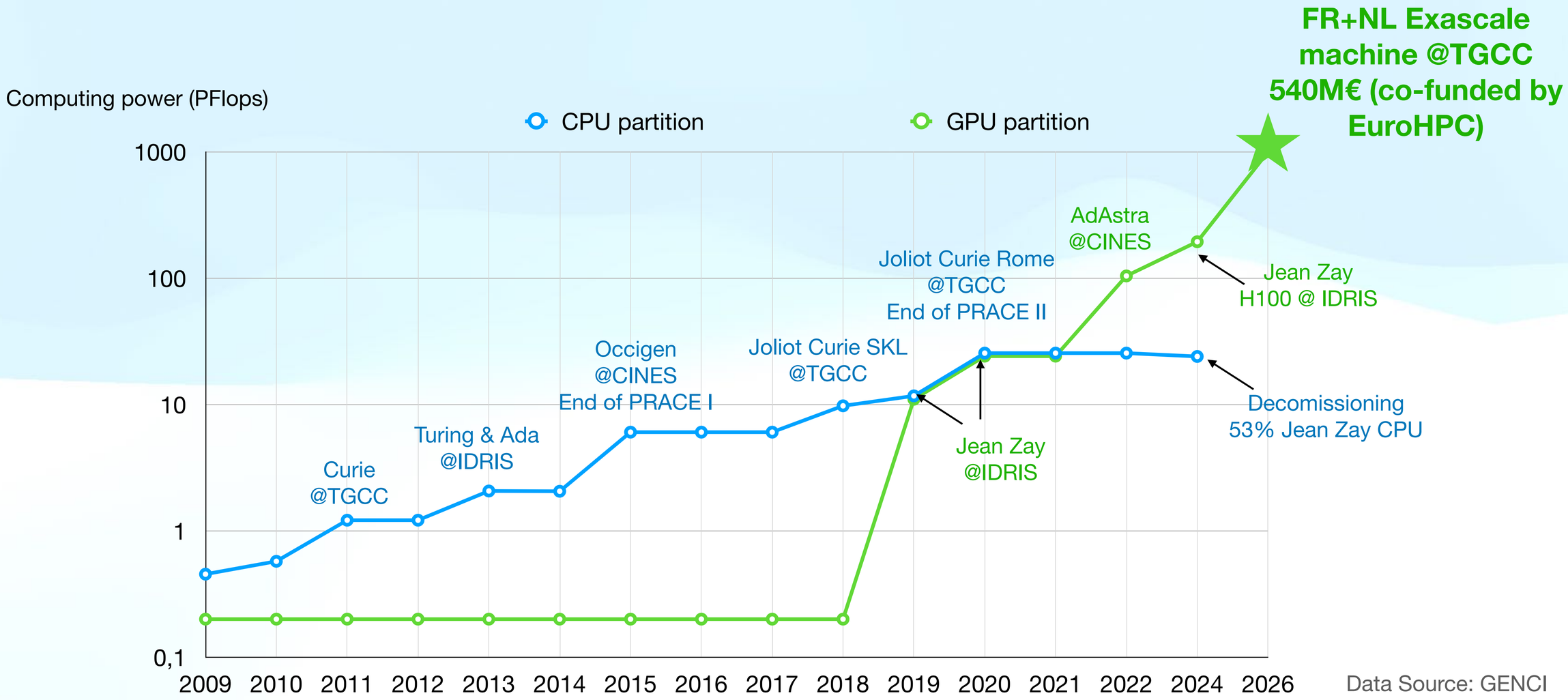
The green 500 list

- all of the « green » cluster are accelerated (Efficiency> 50 GFlops/W)
- The most efficient clusters are not the largest (small clusters also benefit from acceleration!)

The World Top #6 clusters in energy efficiency (November 2023)

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	293	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	44	**Frontier TDS** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684
3	17	**Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	921	58.021
4	25	**Setonix – GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Pawsey Supercomputing Centre, Kensington, Western Australia Australia	181,248	27.16	477	56.983
5	92	**Dardel GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE KTH - Royal Institute of Technology Sweden	52,864	8.26	146	56.491
6	8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 40C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR200, EVIDEN EuroHPC/BSC	680,960	138.20	2,560	53.984

High Performance Computing in France



Data Source: GENCI

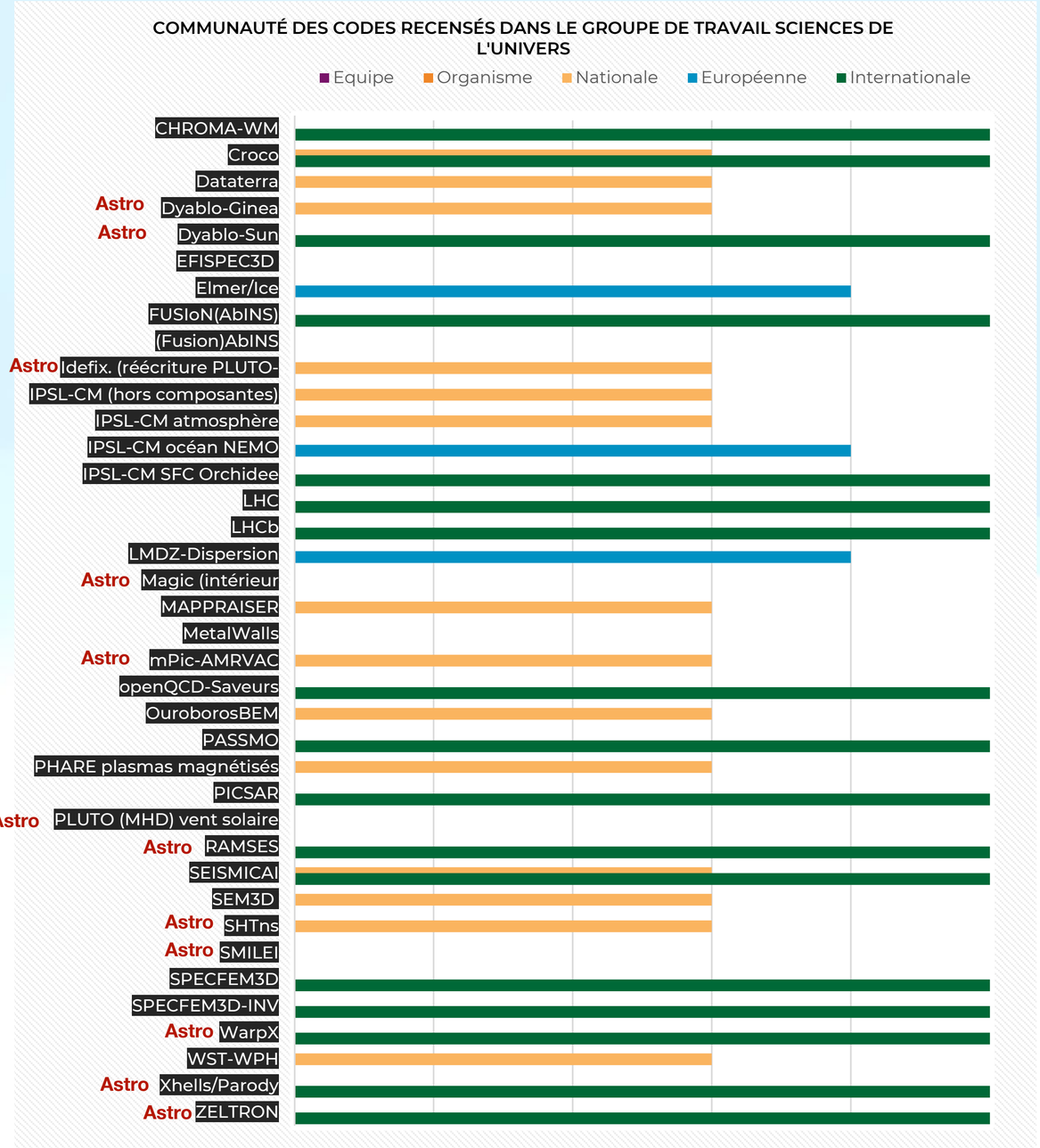
GPU-based HPC clusters are now the norm
Installed CPU power slowly declines

Part II: A new frontier for numerical modeling

PROJET

EXASCALE FRANCE

Les applications françaises face à l'exascale



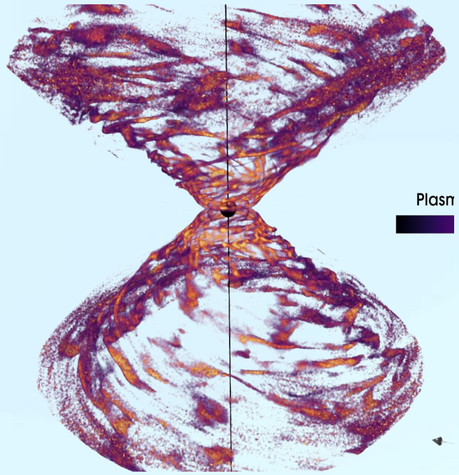
https://hal.science/hal-03736805

Exascale applications

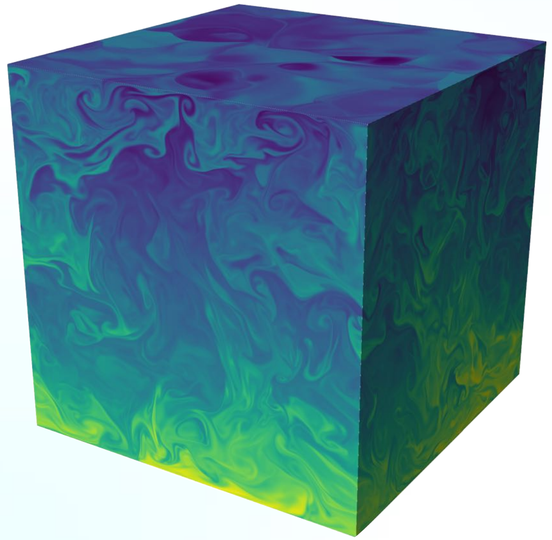
Machine learning	HPC	High throughput computing	Complex workflow
Physics driven - Code emulation - Multi-scale modeling - Empirical models from simulation data **Data driven:** - Analyse (classification) - Data assimilation - Uncertainity quantification	- Simulation of complex phenomenon, multi-physics models - High spatial resolution, multi-scale exploration (zoom) - Uncertainities (ensemble simulation) - Energy profiling, low energy footprint simulation	- Operational processing chain - Real time data processing (e.g. satellite data)	- Scalable processing chain, HPC/HPDA coupling - Task graph parallelism

Some astrophysical applications

Multi-scale physics

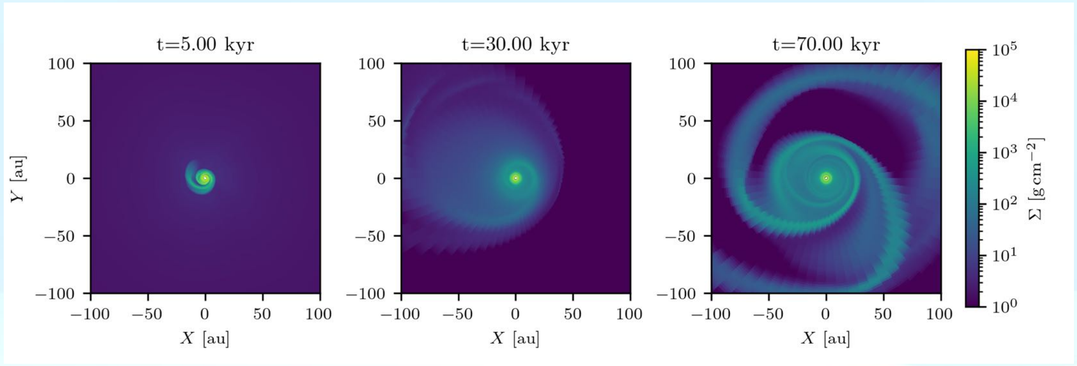


Reconnection sheets in black hole magnetospheres [El Mellah+ 2023]



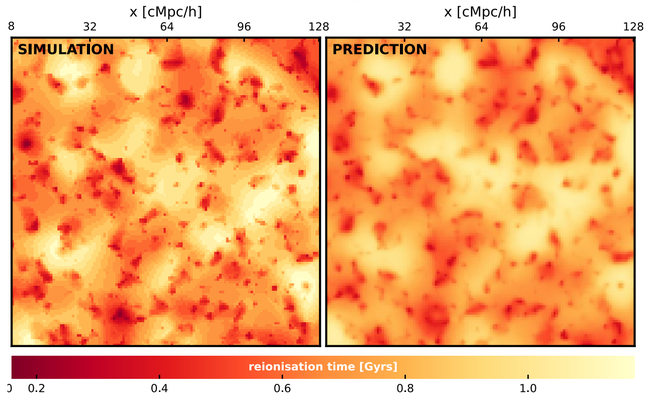
Convection in exoplanet atmospheres [Daley-Yates+ 2021]

Secular evolution of astrophysical objects



Disc formation & evolution in YSOs [Mauxion+2024]

Simulation & Machine learning convergence

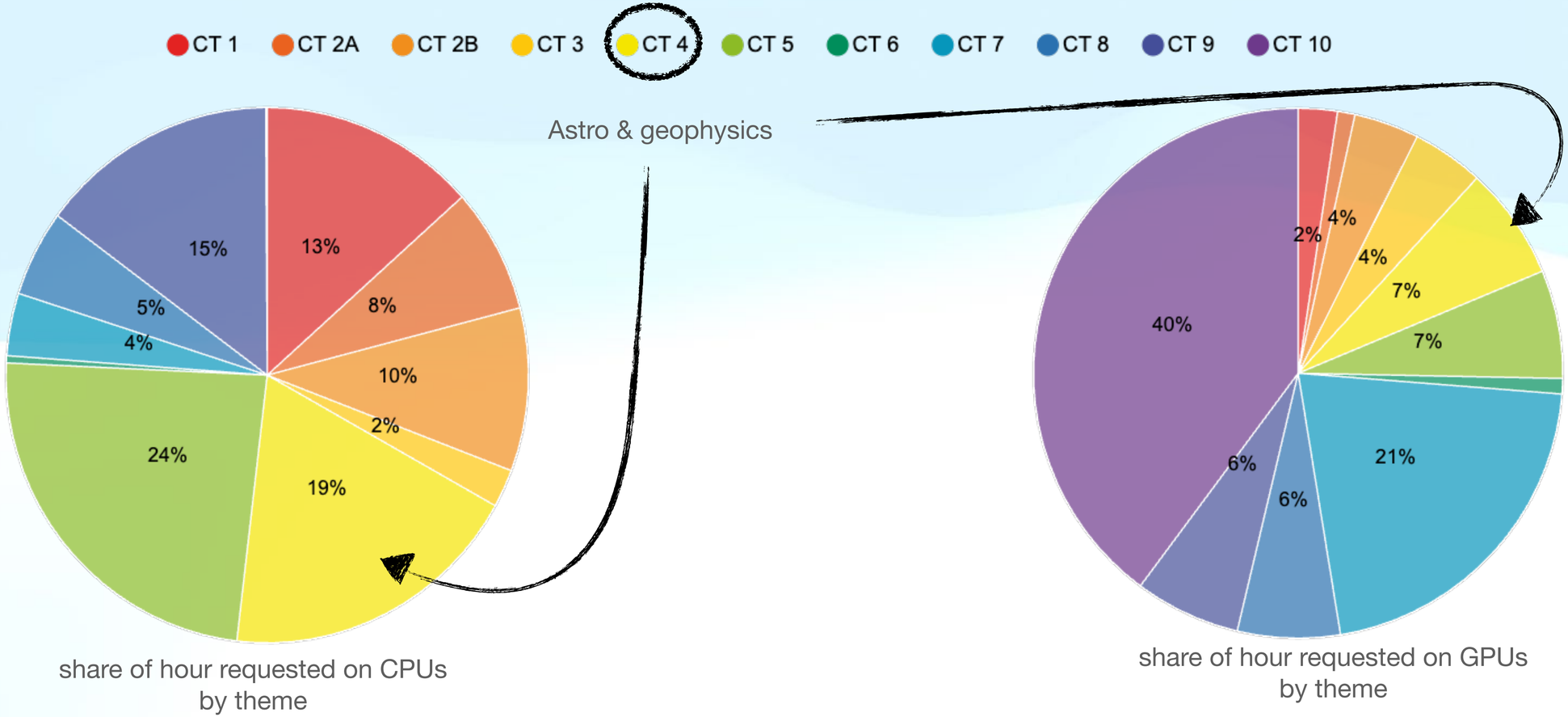


Reconstruction of the reionisation field through machine learning [Hiegel+2023]

Part III: Exploiting exascale computer

A not-so-easy move

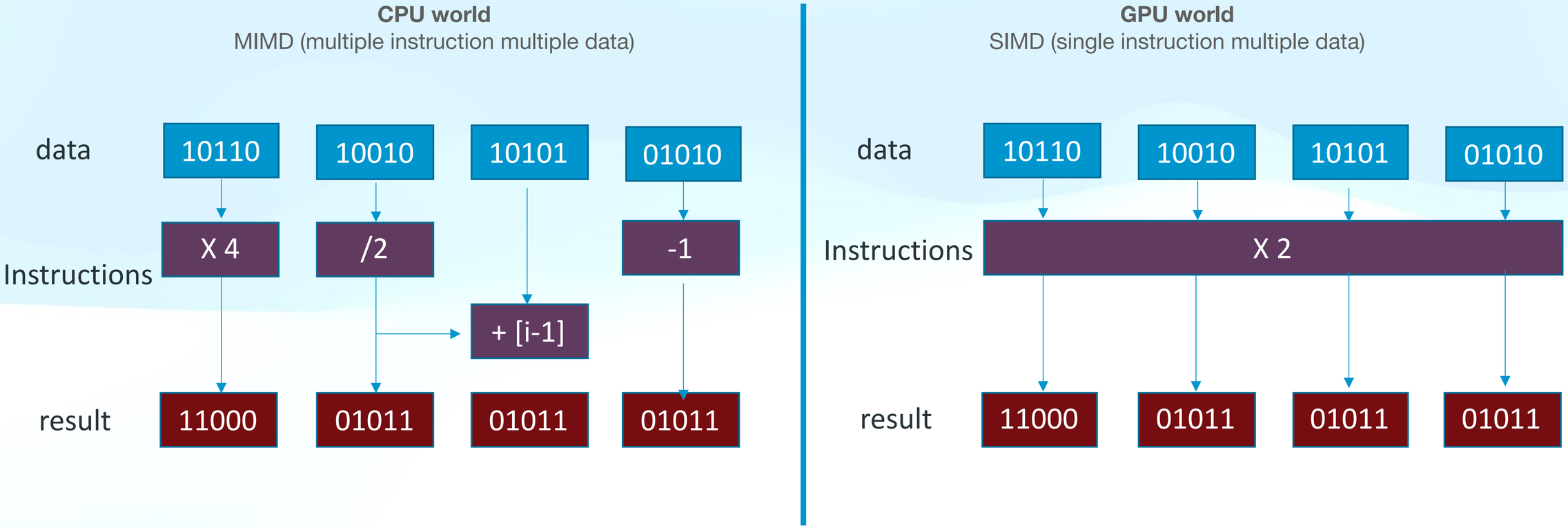
Share of astrophysics & Geophysics on national clusters



The share of GPU usage in astro & geophysics is 3 times less on GPUs than on CPUs. Our community is not ready!

Source: GENCI

Moving towards accelerated machines



Moving to accelerated machines requires a complete rethinking of algorithms and memory access patterns

« Running on GPUs is simple,
just add `-fxxx` when you compile »

Heard during a high energy astro workshop

« Running on GPUs is simple, just add `-fxxx` when you compile »

Heard during a high energy astro workshop

✔ your code will (hopefully) compile

✘ your code will not be faster

your code will probably be less energetically efficient

« Running on GPU is simple,
just add –fx… when you compile »

WRONG

Heard during a high energy astro workshop

- ✓ your code will (hopefully) compile
- ✗ your code will not be faster
-  your code will probably be less energetically efficient

The need for accelerated codes

- Not possible to directly use one's favorite code [Ramses, Pluto, AMR VAC, Flash] on an accelerated machine (SIMD!)

➡ One needs to port codes, or to write them from scratch (in both cases, years of development)

  - Diversity: each manufacturer tend to promote its own technology (e.g. CUDA@NVidia, HIP@AMD, SYSCL@INTEL…)

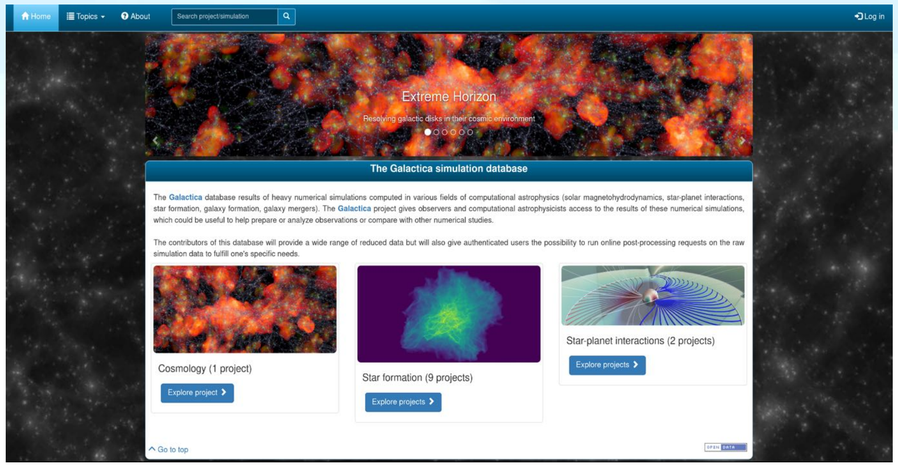
  - Resilience: failure rate can be as high as 1 failure/hour on exascale machines. Codes must be *resilient* to node failures.

MHD@EXASCALE

Fast & efficient simulations for astrophysics

FRANCE 2030 PROGRAMME DE RECHERCHE Origins

- **Develop three French MHD codes: Dyablo, Idefix & Shamrock, targeting accelerated machines based on modern C++**
  - Why 3? Reproducibility (different algorithms), Robustness, testbed of different technological choices
  - Avoid the unmaintainable “one code to rule them all”
- **Builds up on existing prototypes from CEA and CNRS**
  - Minimize risk (prototypes have already been validated against standard test suites)
  - Maximize scientific impact (shorter development to get scientific applications out)
- **Simulation data sharing with the community though Galactica**
  - Databases disseminated across local meso-centers (Tier 2)
  - Simplify the re-usability of published simulation data
- **FAIR principle (Findability, Accessibility, Interoperability, and Reuse).**
  - Codes publicly released under Open Source licenses



The Galactica web server
http://www.galactica-simulations.eu

MHD@EXASCALE

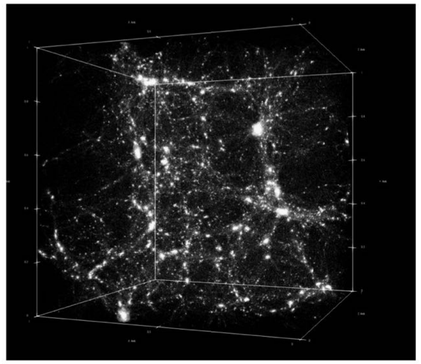
Fast & efficient simulations for astrophysics

FRANCE 2030 PROGRAMME DE RECHERCHE Origins

The 3 code prototypes

Dyablo

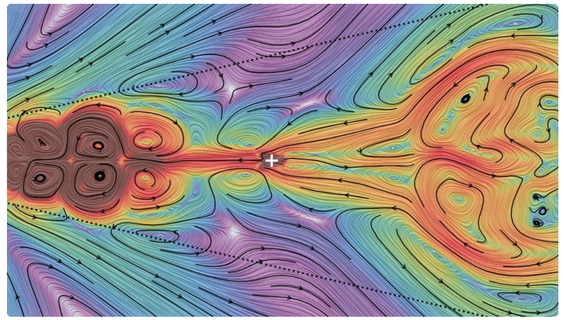
- Block-based adaptative mesh refinement
- Kokkos-based performance portability
- Hydro+particles



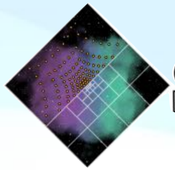
Large-scale structure formation

iDEFiX

- Stretched curvilinear grids
- Kokkos-based performance portability
- Constrained transport MHD
- In production on national centers
- First science publications are out

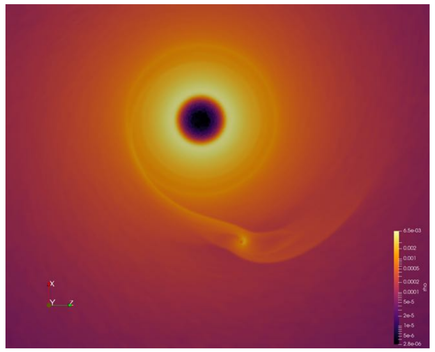


Disk winds



Shamrock

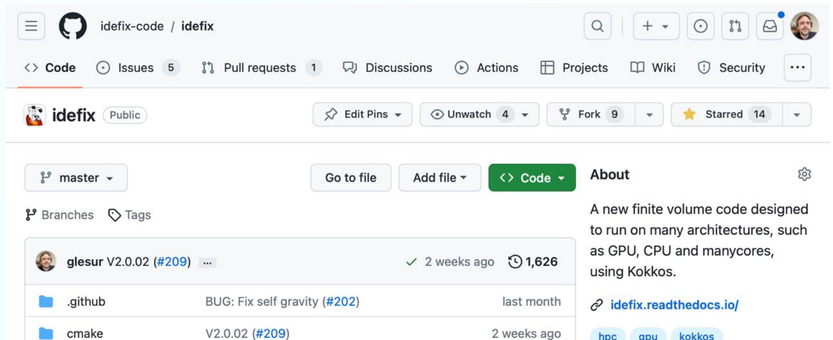
- Meshless smoothed particle hydrodynamics
- SYSCL-based performance portability
- Hydrodynamics



Planet-disk interaction

Idefix

- Idefix is technically a new code (written from scratch), finite-volume Godunov algorithm in C++ 17, relying on Kokkos meta-programming objects.
  - 1st, 2nd or 3rd order in time (using Runge-Kutta TVD time integrators)
  - 1st, 2nd, 3rd, 4th order spatial reconstruction
  - Constrained transport for the MHD module
- Inputs, outputs and data structures are very similar to Pluto: simplified setup portability
- Strong encapsulation: each physical module/algorithm is a C++ object with its own constructor/destructor
- Method paper : Lesur+2023, A&A, 677, A9
- Online documentation: idefix.readthedocs.io
- Code under CECILL license, available on GitHub: github.com/idefix-code/idefix



Feature	Status
HD & MHD	✔
Multiple Riemann solvers (Lax, HLL, HLLC/D, Roe)	✔
Geometry (cartesian, cylindrical, spherical, polar)	✔
Non-ideal MHD (Ohmic, Ambipolar, Hall)	✔
MPI, MPI+OpenMP, MPI+Cuda, MPI+HIP	✔
Checkpointing & restart	✔
Super time-stepping (RKL scheme: viscosity, ambipolar diffusion, Ohmic resistivity)	✔
Orbital advection (FARGO)	✔
Radiative transfer	✘
Dust (particle approach)	(not yet public) ✔
Dust (fluid approach)	✔
Self gravity	✔

Planet-disc-MHD wind interaction

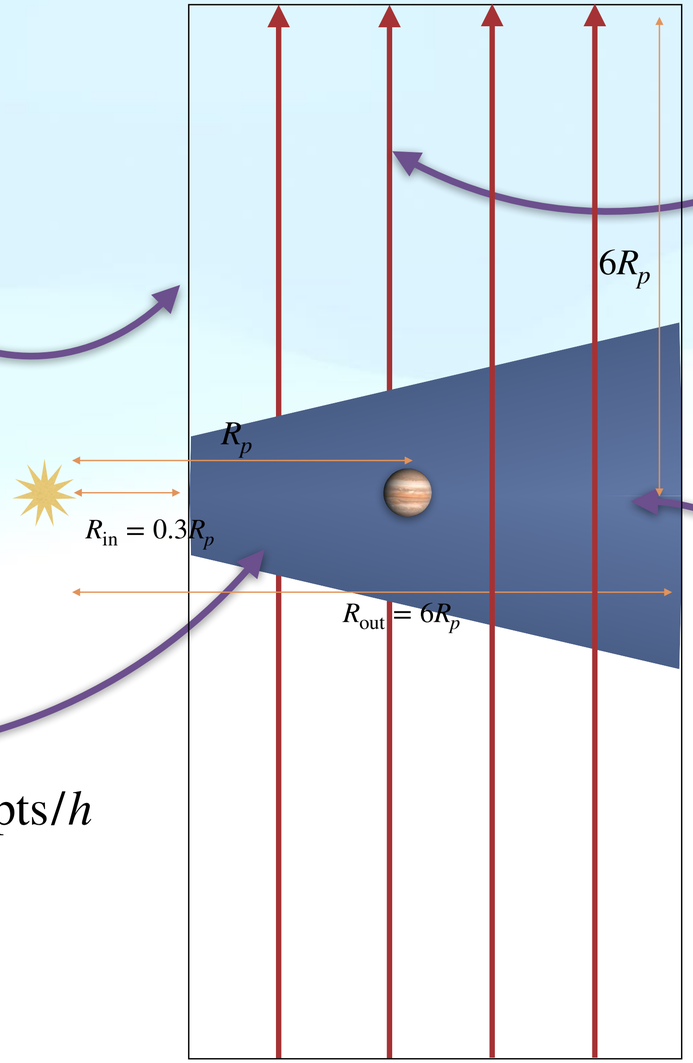
3D cylindrical grid,
Idefix code [Lesur+2023],
static mesh refinement in the disc

$(N_R, N_\phi, N_z) = (640, 2048, 384)$

Locally isothermal model.

Disc thickness $h/r = 0.05$

Resolution at planet location: $16\,\mathrm{pts}/h$

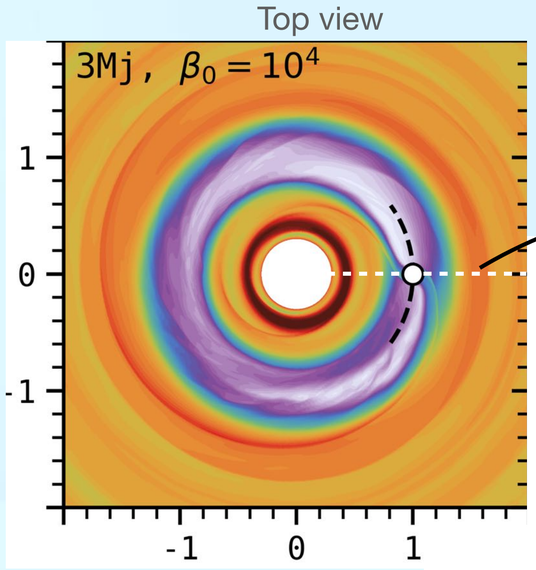


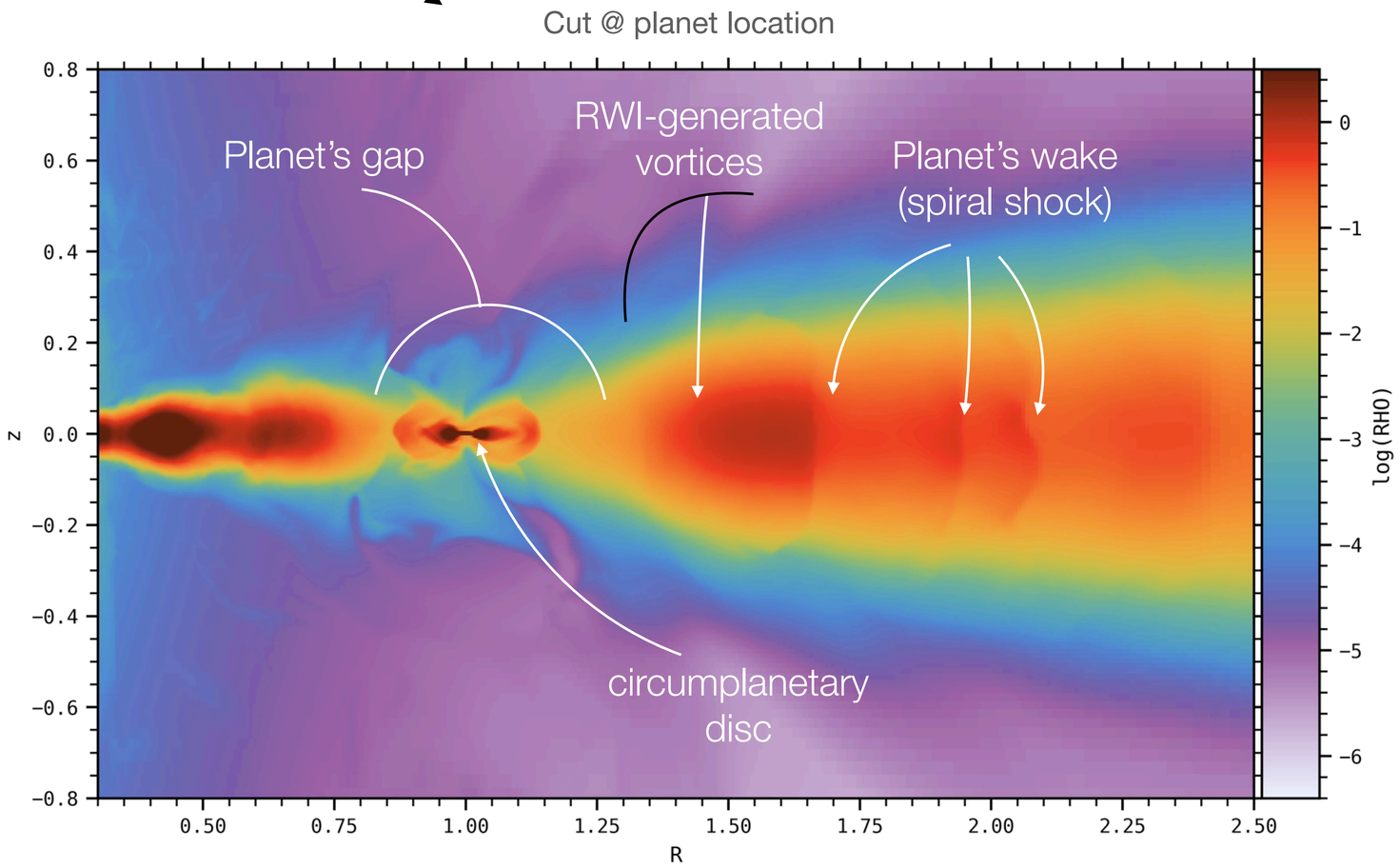
Poloidal magnetic field threading the disc

Prescribed ambipolar diffusion vertical profile.
Runge-Kutta-Legendre scheme to accelerate the integration of diffusive terms

Jean Zay GPU (IDRIS), 128x Nvidia V100

Planet-disc-wind interaction in action

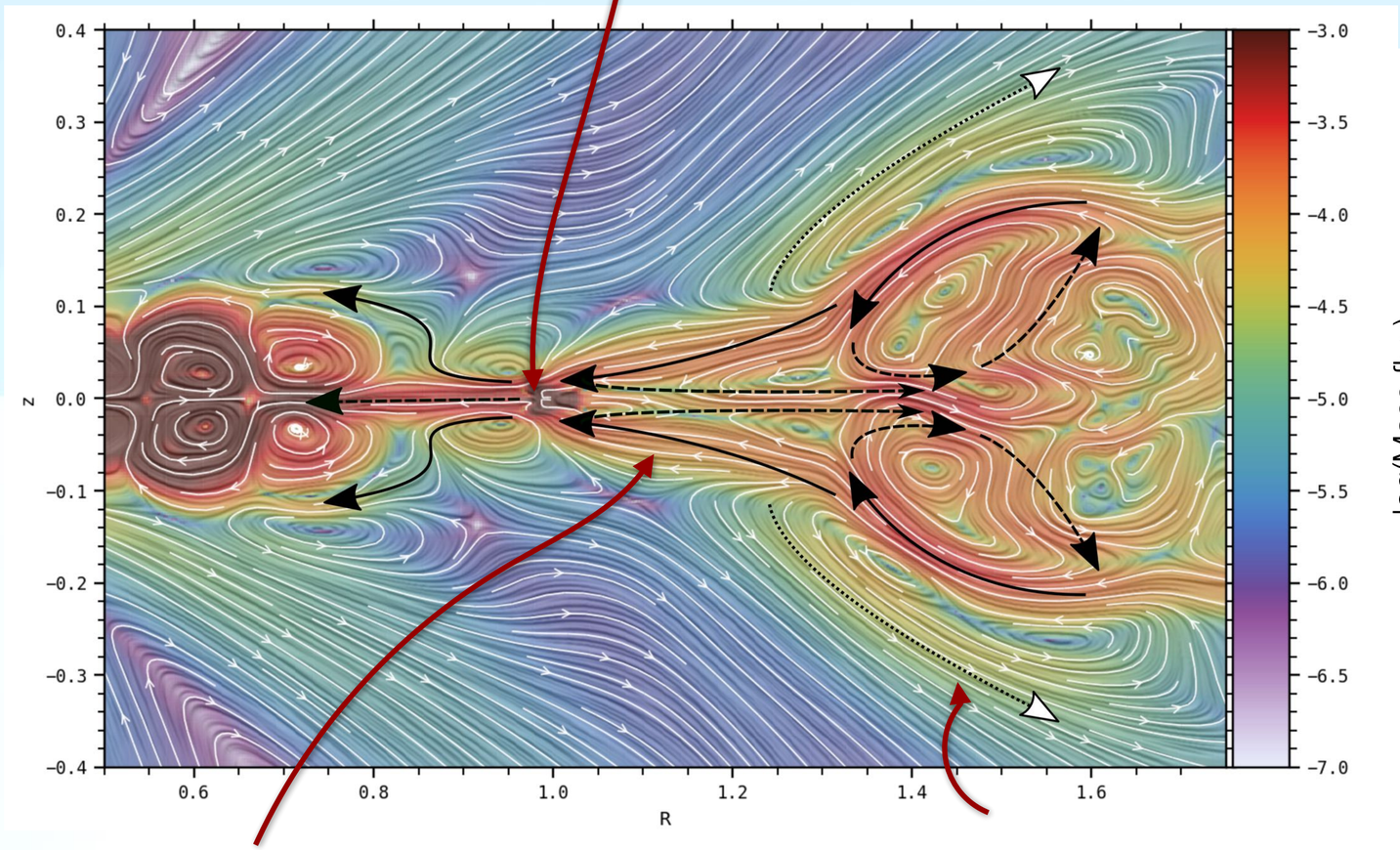




[Wafflard-Fernandez & Lesur 2023]

3D Circulation around a planet

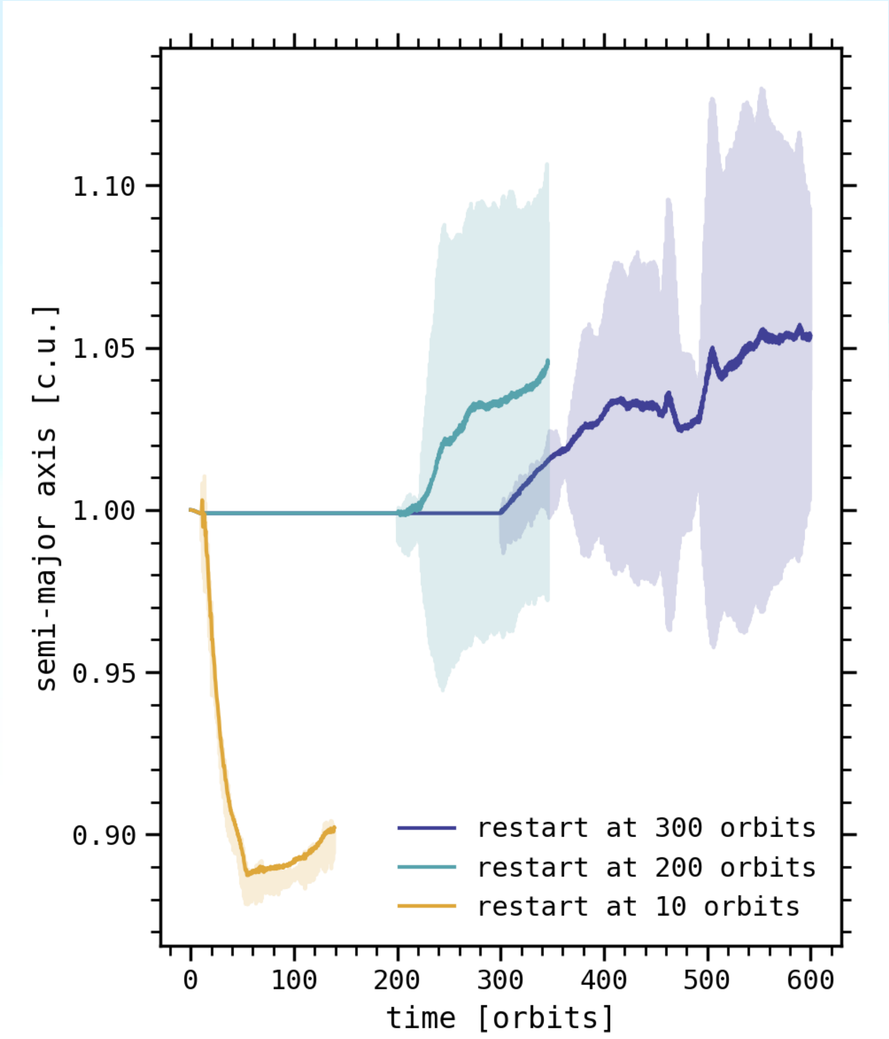
3x  weak 



Accretion streamer « through » the gap

Wind « plume »

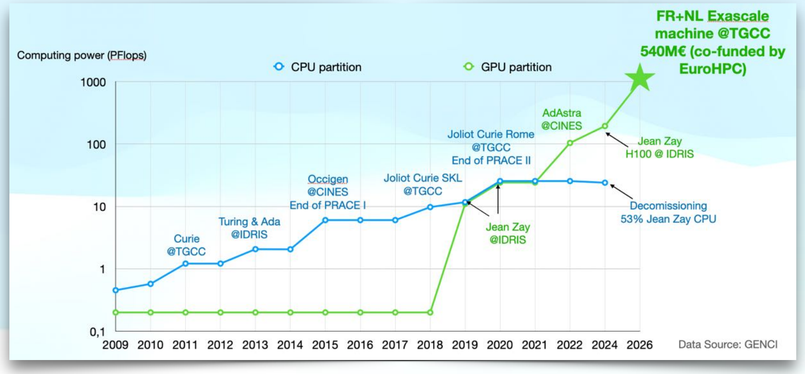
Outwards giant planet migration



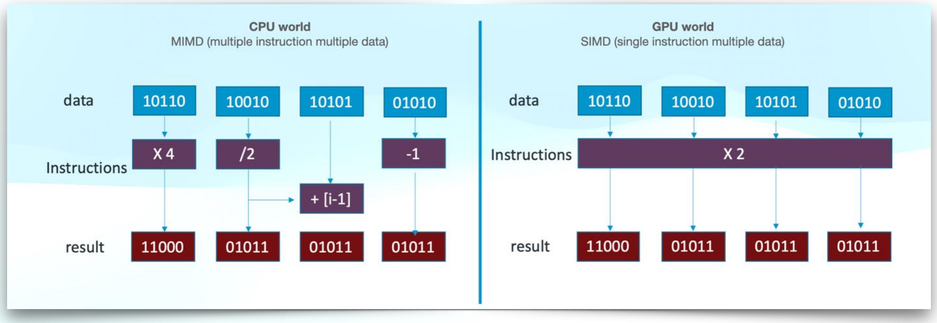
Wafflard-Fernandez + 2024, in prep

Take away messages

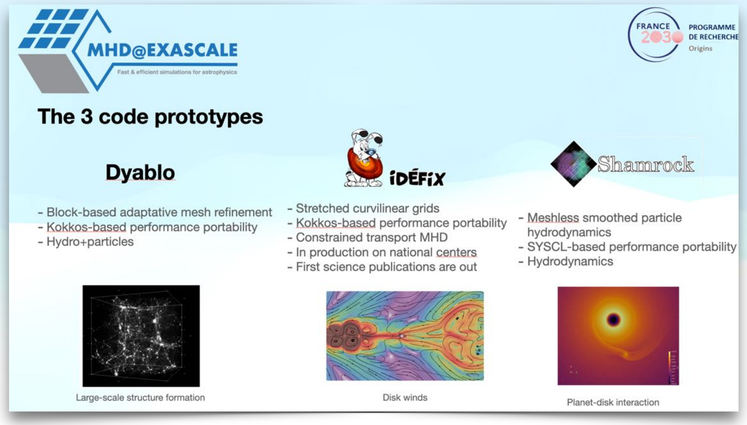
- Future French & European clusters (including Exascale machines) will be GPU-based



- Exploiting these accelerated cluster requires a full rewrite of existing codes



- Several projects have started to write/port existing codes (PEPR Origins, PEPR Numpex). These are short-term projects, which don't address the question of maintenance, formation and deployment on future architectures



- IDEFIX is one such Exascale-ready community code. Open source, multi-physics, tested on all available clusters in France on up to 4000 GPUs