ESA's 2025 Big Data Foundations for Earth Observation Training Course

Addressing Data Processing Challenges: Scaling Computational Resources With High Performance Computing

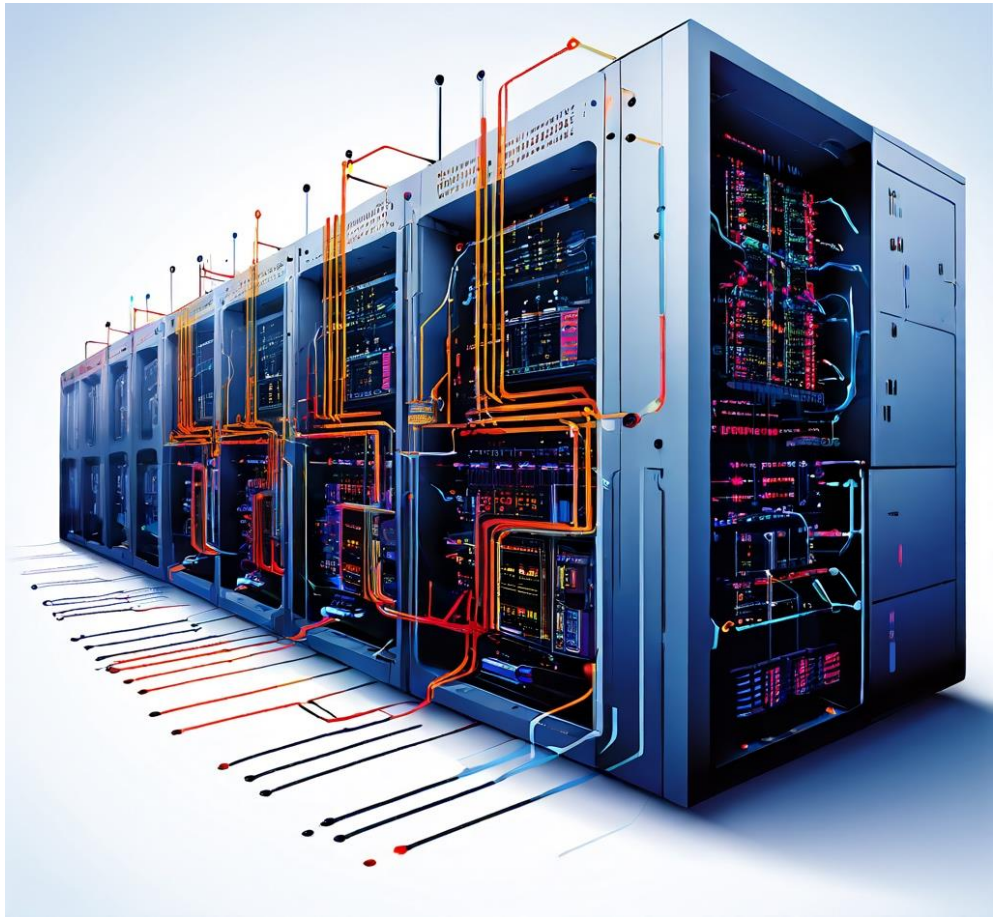
Sean Quin
ESA
23/09/25

European Commission | researchLatvia Value Through Knowledge | Ministry of Education and Science Republic of Latvia | esa academy

Current Challenges

- **Increase in EO Data Sets**: Satellite data volumes are expected to grow tenfold in the coming years.
- **Computationally Intensive Processing:** Converting raw data into actionable insights now requires significantly more time and resources.
- **Accessibility and Usability**: Ensuring that stakeholders can easily access, interpret, and apply this data is becoming a growing challenge.
- To name a few…

What is HPC?

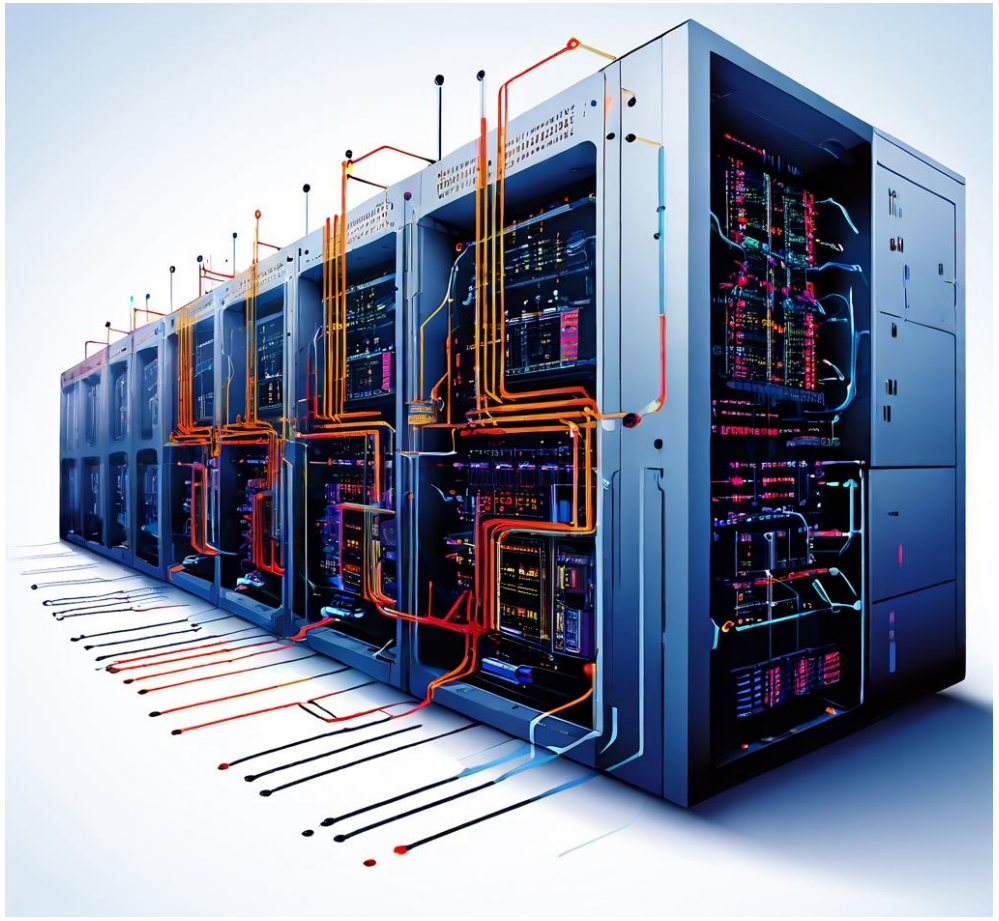


High Performance Computing (HPC) uses powerful computer systems to solve complex problems that regular computers can't handle.

An HPC system, like a supercomputer or computer cluster, is made up of many individual computers working together. They're connected by a fast network, so they can share data quickly.

Instead of relying on one super-powerful computer, HPC systems use lots of computers at once. This teamwork lets them tackle big tasks that regular computers couldn't manage on their own.

What is HPC?

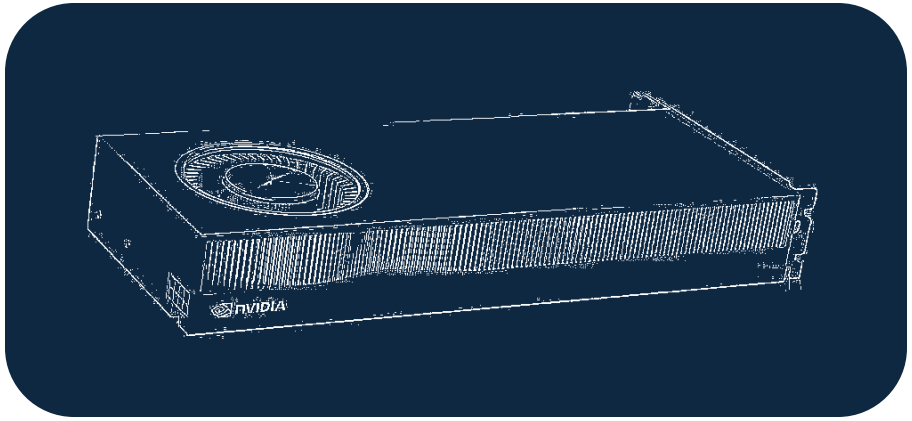


Parallel Processing: HPC systems split big problems into smaller parts and solve them simultaneously across many computers, speeding up calculations dramatically.

Specialised Hardware: HPC systems use advanced components like high-performance processors, Graphical/AI Processing Units, and fast storage to handle specific tasks more efficiently.

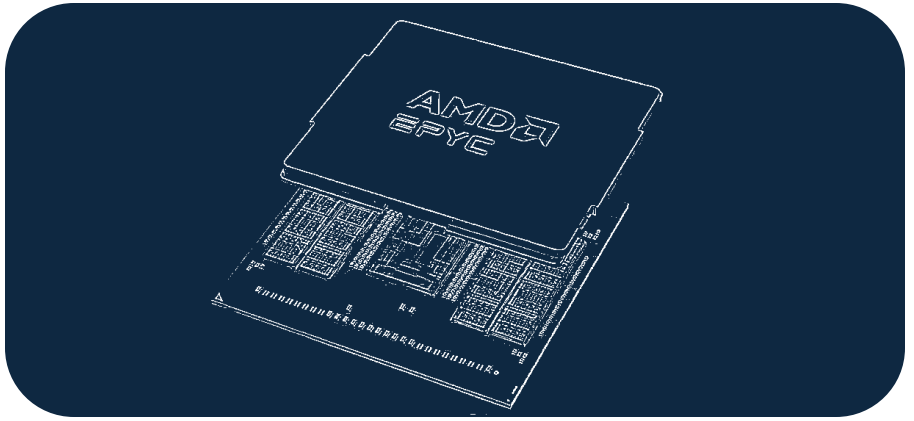
HPC is used across fields like scientific research, engineering, weather prediction, finance, and more. It enables quick solutions to complex problems that would take years on regular computers.

HPC Hardware



GPU

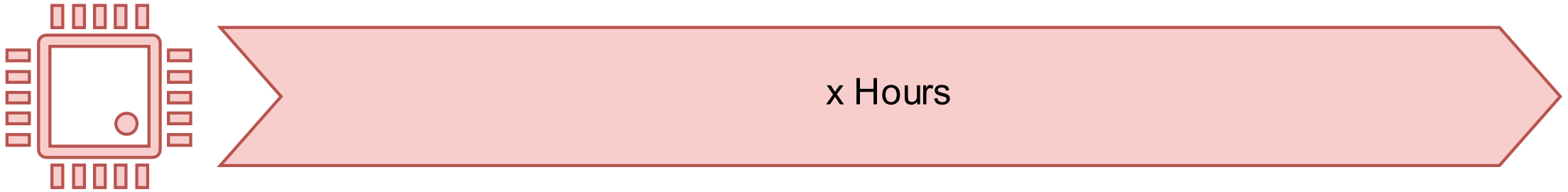
- High level of parallelism
- Tasks such as AI/ML model training, data mining operations, high-res graphics rendering, etc.
- Less flexible than CPUs in terms of tasks



CPU

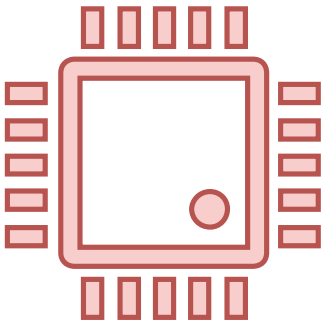
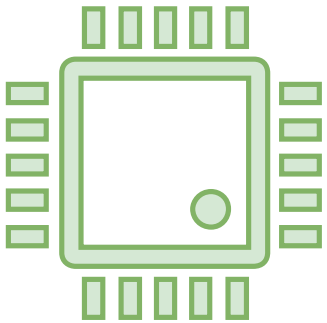
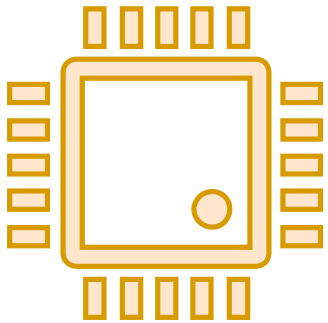
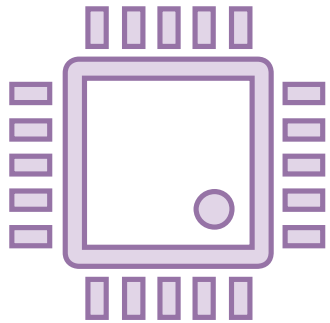
- Can handle a variety of general tasks
- More precise than GPUs in FLOP
- Used for a wider range of tasks such as simulations, I/O, etc.

Parallel Execution

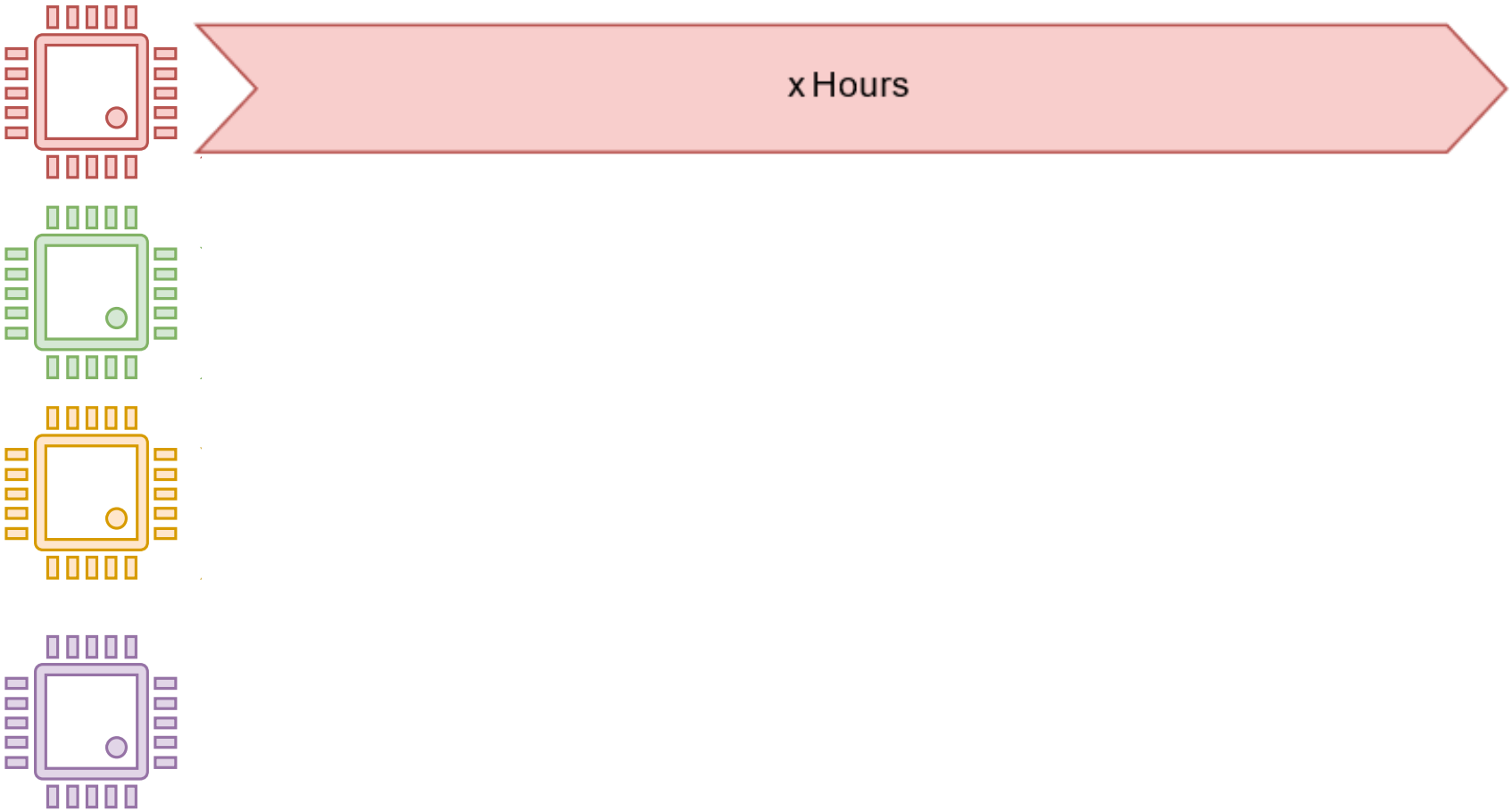


Execution Time

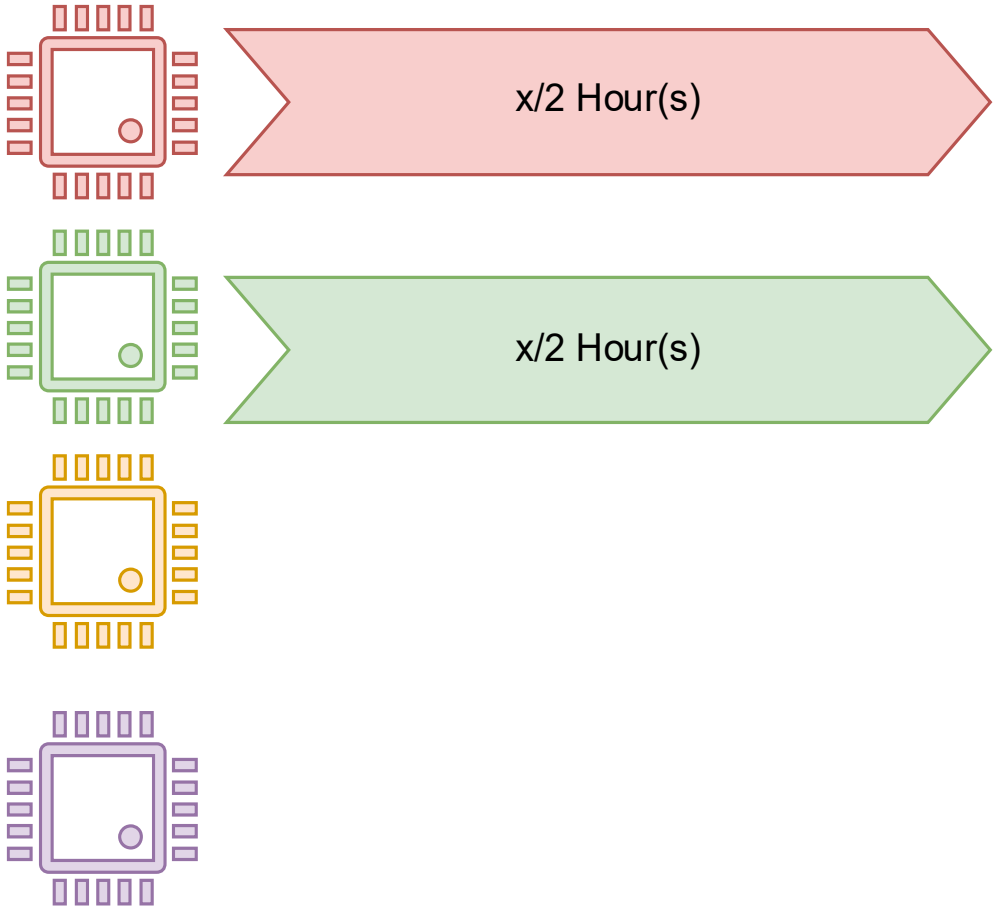
Parallel Execution

Parallel Execution

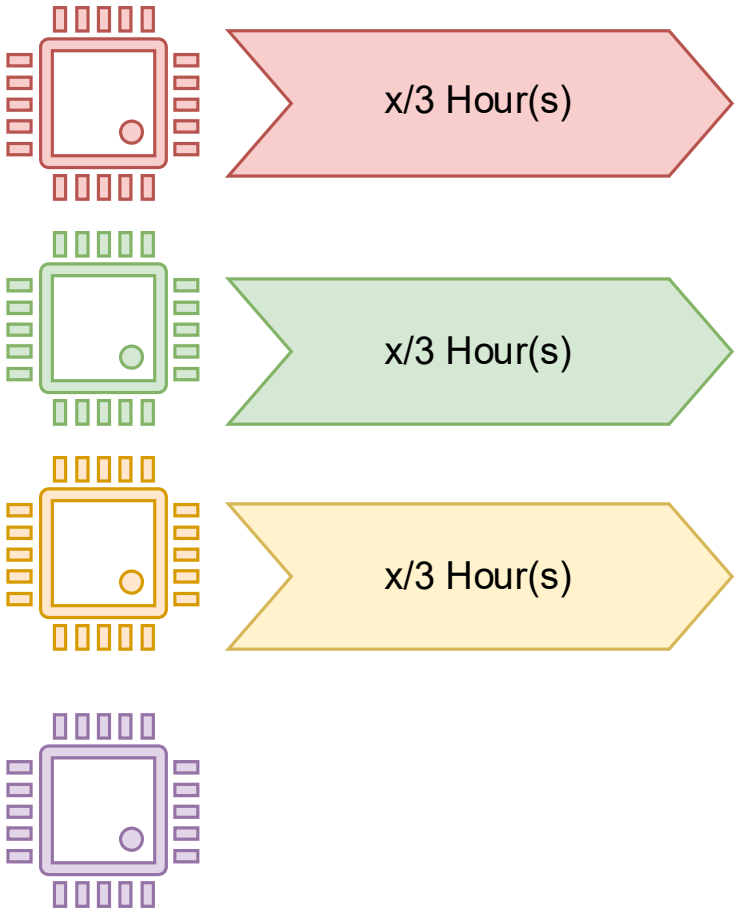


Parallel Execution



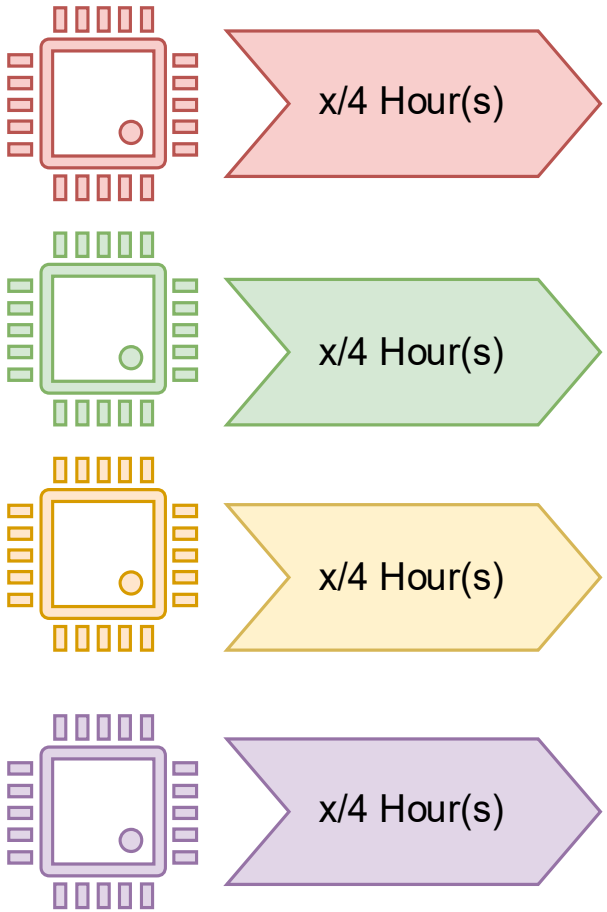
Execution Time

Parallel Execution



Execution Time

Parallel Execution

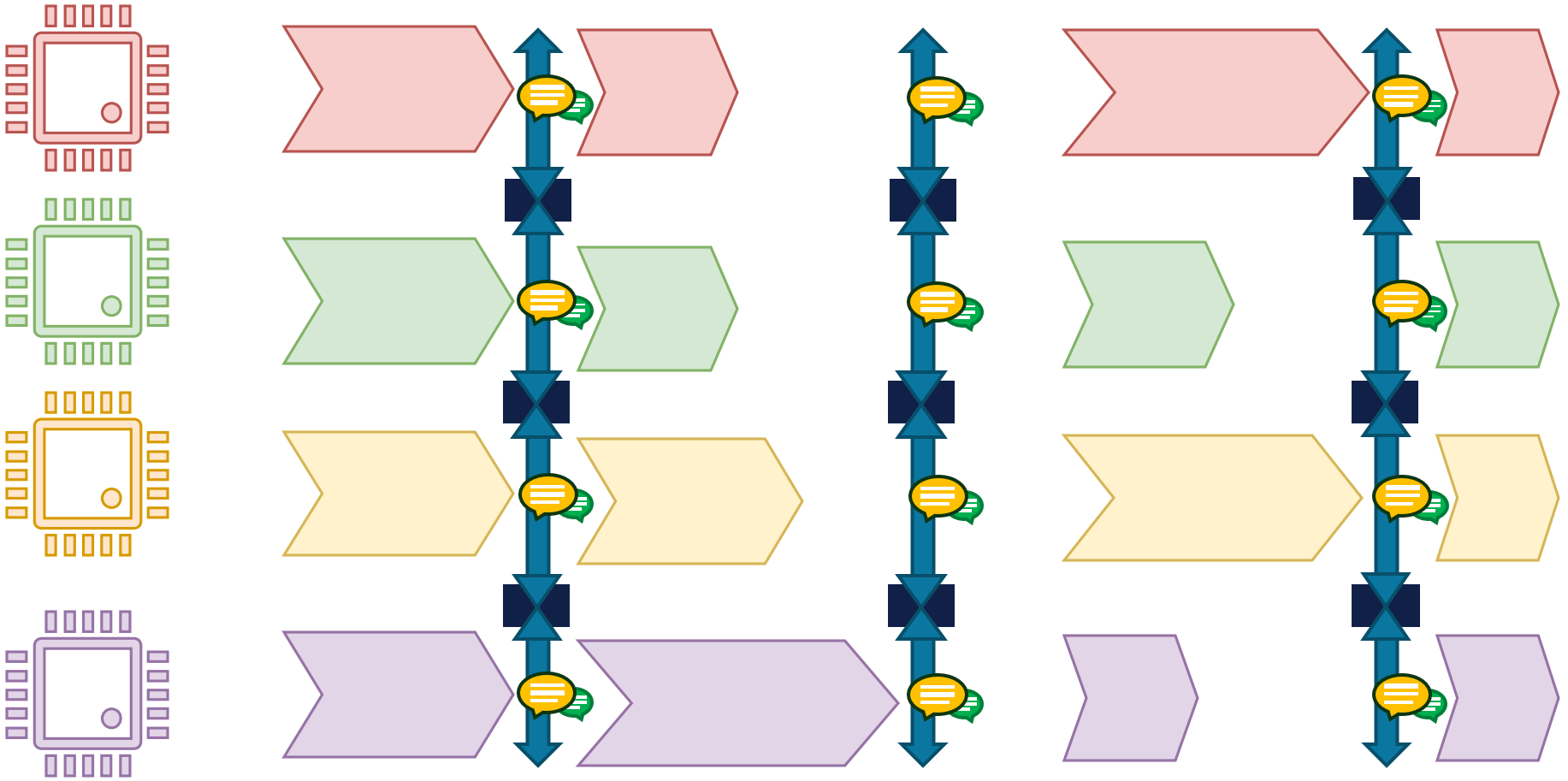


Execution Time

Parallel Execution

Credits: Rosie Wilson

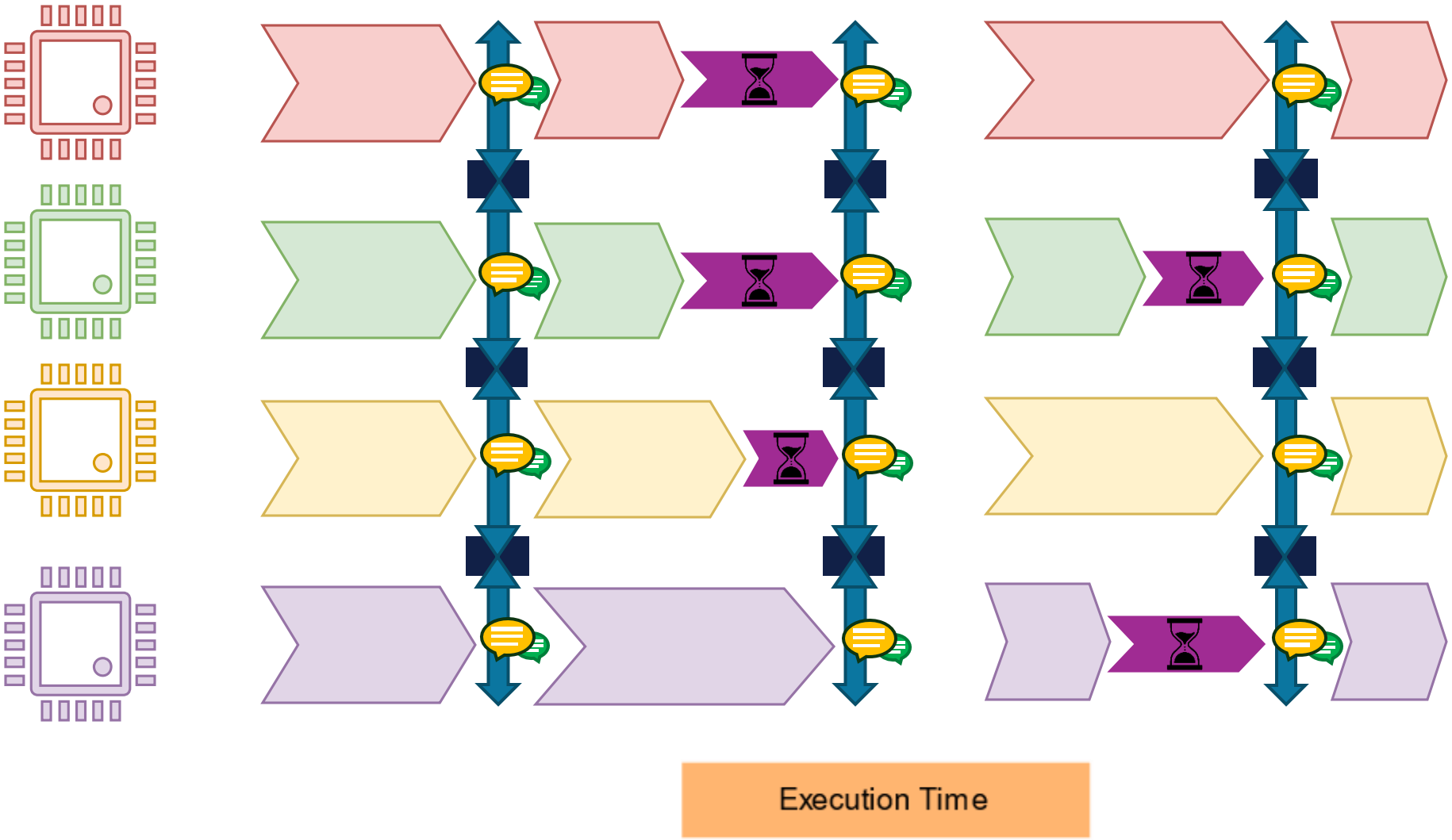
Parallel Execution



Communication between CPUs

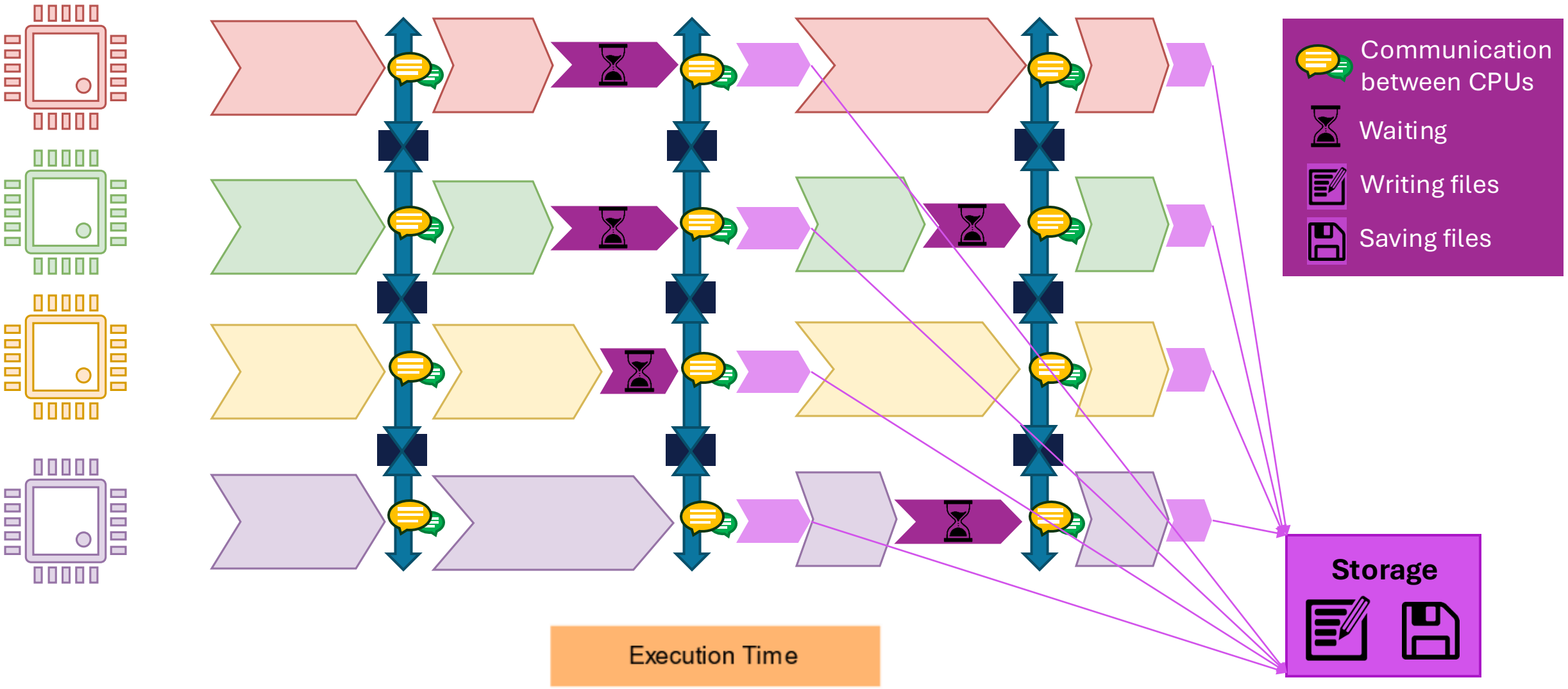
Credits: Rosie Wilson

Parallel Execution



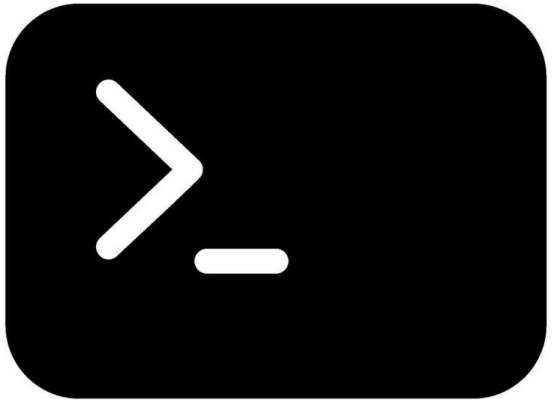
Credits: Rosie Wilson

Parallel Execution



Credits: Rosie Wilson

Challenges with using HPC systems



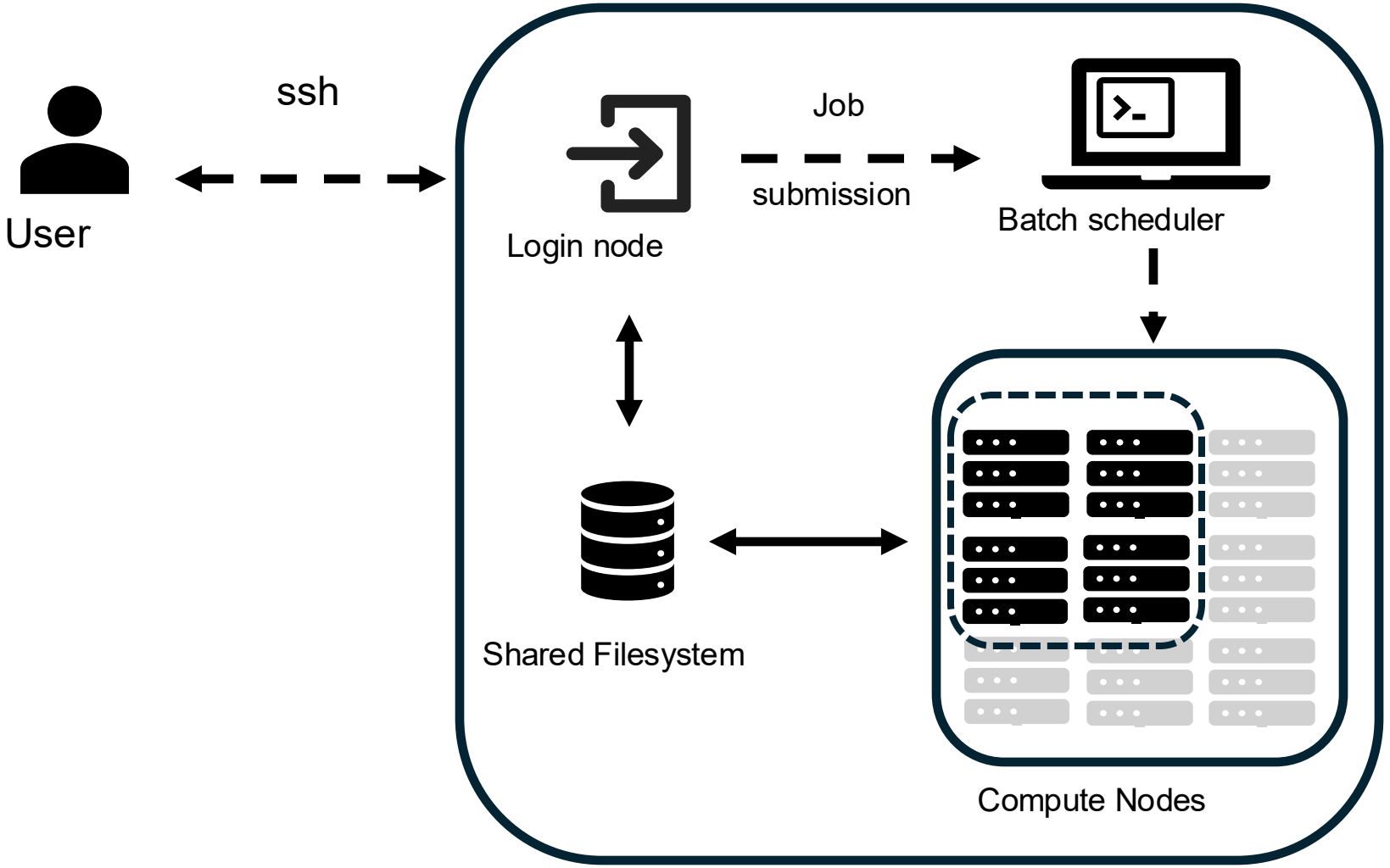
Usability

- Often CLI Based
- Often Linux based environment
- Requires familiarity with schedulers and module based environments
- Learning curve may be high for some
- **Transfer of data sets is still an ongoing challenge**

Application and Environment

- Porting of application ready environment can be tricky (not out of the box)
- May require adjustment of code for parallel execution
- Familiarity of the intersection between hardware and software is required for effective use of systems

Challenges with using HPC systems
Workflow



Challenges with using HPC systems
Example Pansharpening

- 4 bands
- 6 meters resolution

- 1 band
- 1.5 meters resolution

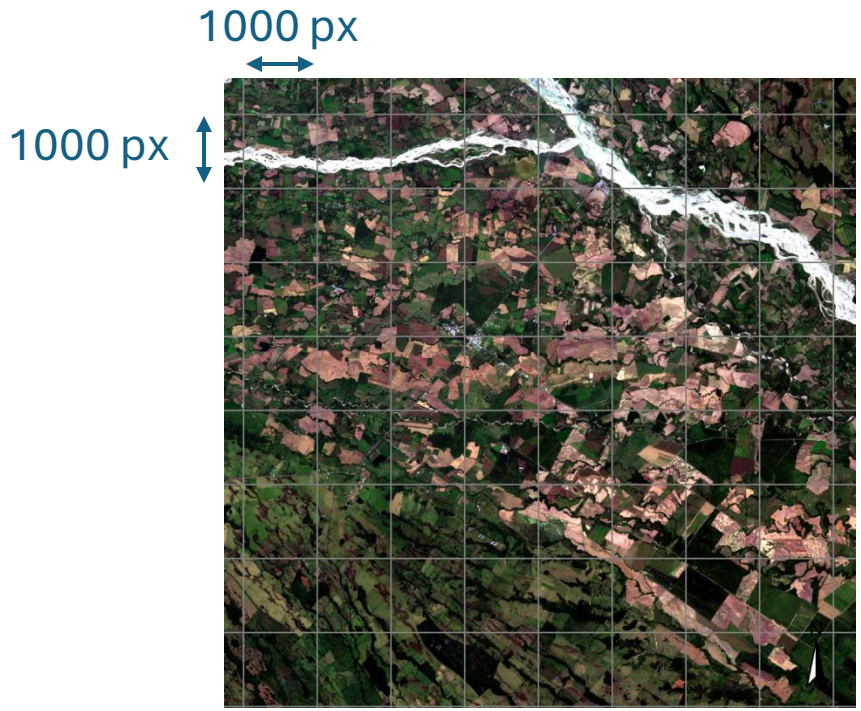
- 4 bands
- 1.5 meters resolution

Challenges with using HPC systems
Example Pansharpening

To use effectively an HPC, you need parallelization

= split a big problem in many small problems

➔ Easy for satellite image processing!



Example of libraries:

- Dask
- Orfeo Toolbox
- Custom code

Challenges with using HPC systems
Example Pansharpening

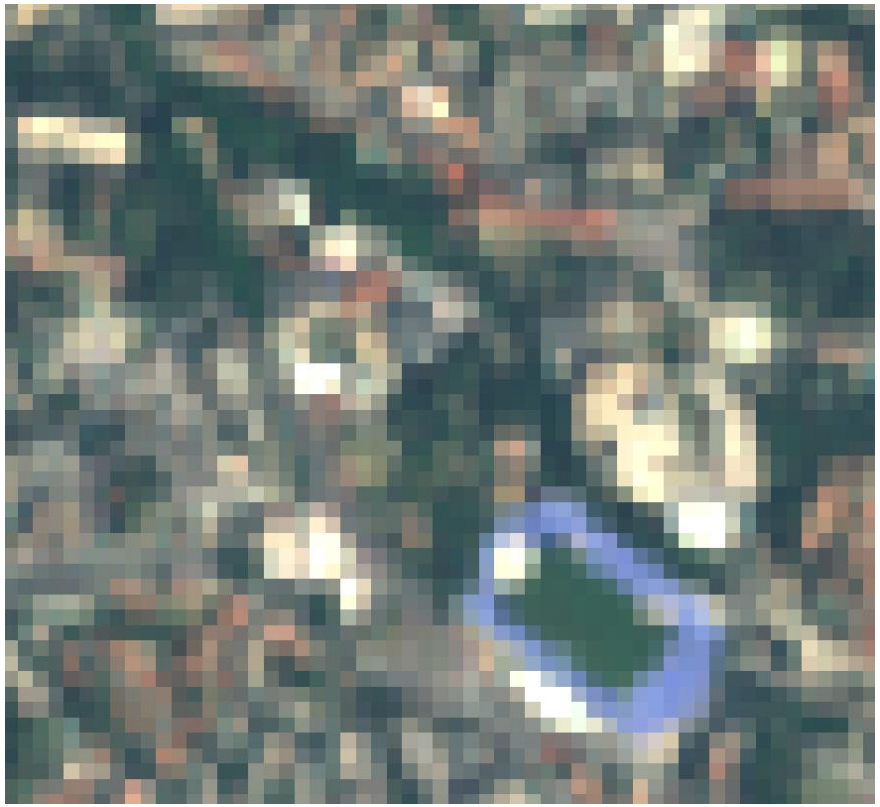
```
cp /lustre/projects/hpc-prj-9001/LPS25/CPU/submit_cpu.pbs .

qsub submit_cpu.pbs
```

Content of *submit_cpu.pbs*

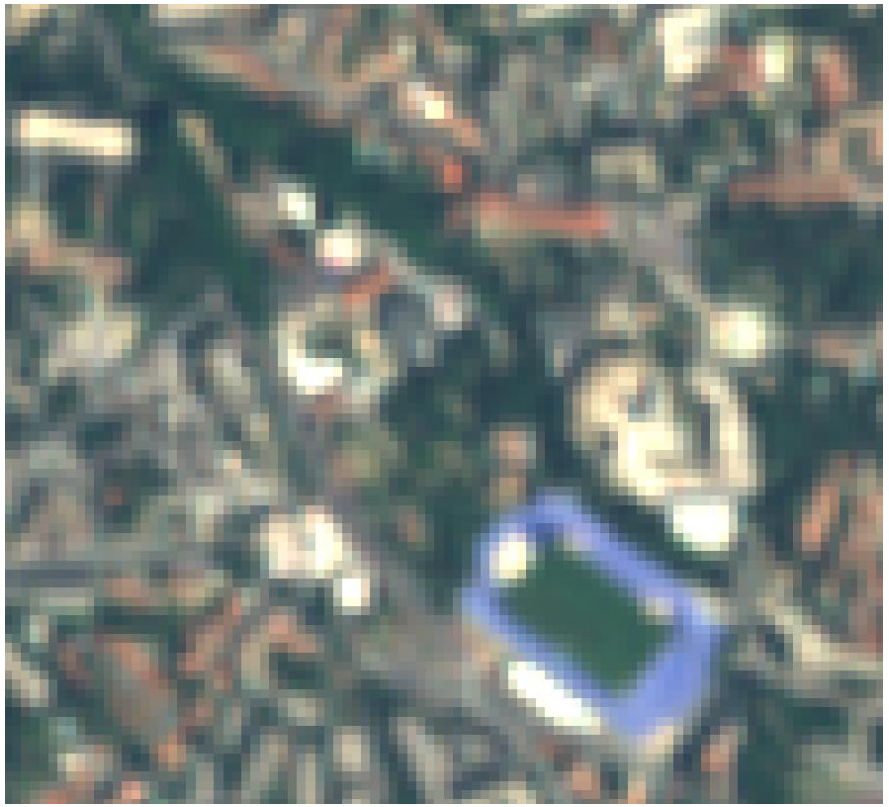
```
# Resource Specification
# Selects 3 ”chunks” (here 3 cpu nodes, with 192 cores each so 576 cores)
#PBS -l select=3:ncpus=192:mpiprocs=192
#PBS -l walltime=00:30:00
#PBS -q cpu
#PBS -N workshop-job

# Environment loading
module load cray-mpich craypals craype-x86-genoa
source /lustre/projects/hpc-prj-9001/miniconda3/etc/profile.d/conda.sh
conda init
conda activate gdal

# Execution
mpirun –n 192 python /lustre/projects/hpc-prj-9001/LPS25/CPU/process_images_mpi.py
```

Challenges with using HPC systems
Example Resolution Improvement

SEN2VENμS : Sentinel 2 to VenμS



10 meters resolution



5 meters resolution

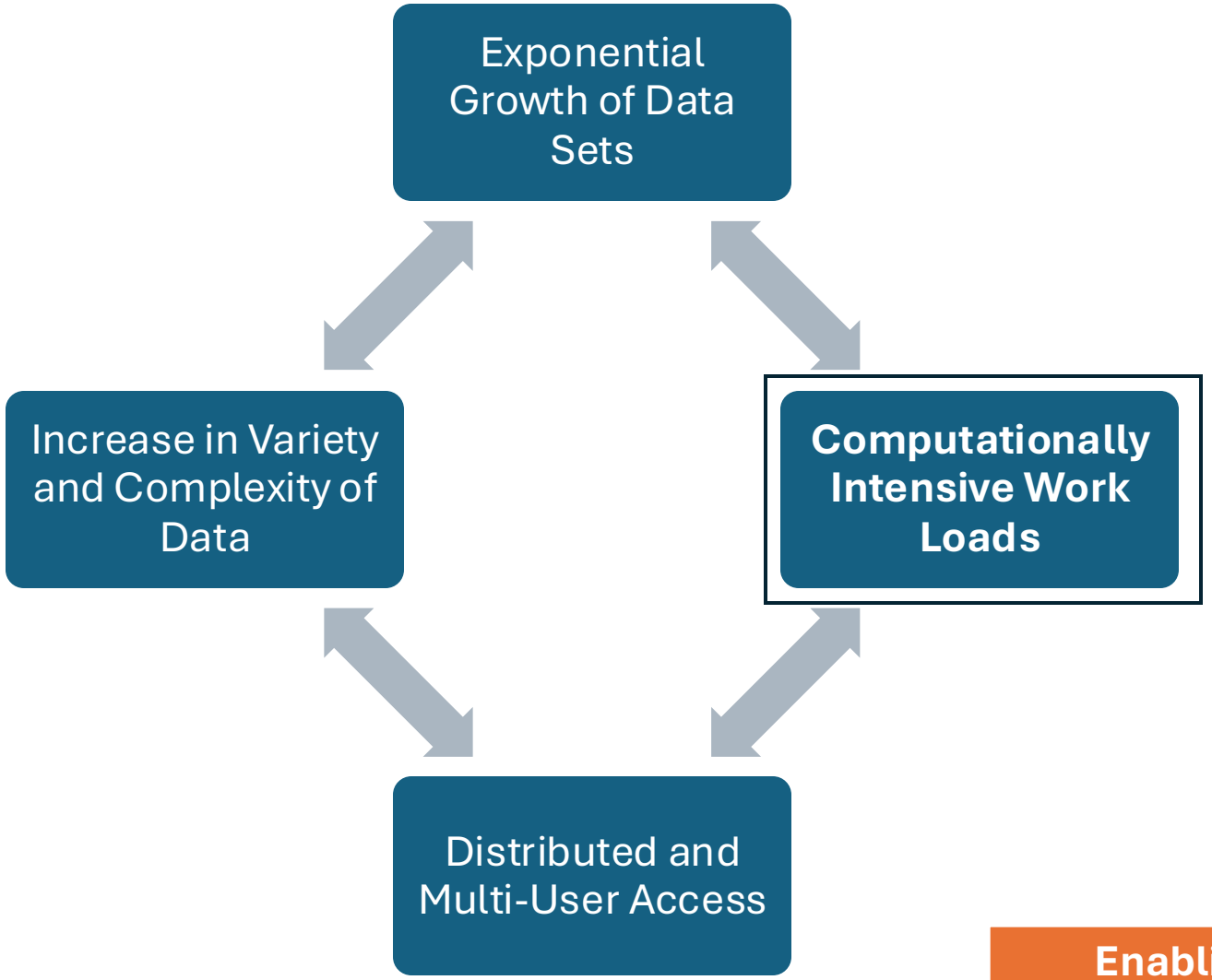
Challenges with using HPC systems
Example Resolution Improvement

```
cp /lustre/projects/hpc-prj-9001/LPS25/GPU/submit_train_gpu.pbs .

qsub submit_train_gpu.pbs
```

Content of *submit_train_gpu.pbs*

```
# Selects 4 "chunk" (here with 4 GPUs per node so 16 GPUs)
#PBS -l select=4:ngpus=4:ncpus=64:mem=650gb
#PBS -l walltime=00:20:00
#PBS -q gpu

export NCCL_SOCKET_IFNAME=ib0
source /lustre/projects/hpc-prj-9001/miniconda3/etc/profile.d/conda.sh
conda init
conda activate sen2venus

python train.py
```

Advances in Technology: Hybrid Cloud / HPC
EO Data Challenge



Credits: Cineca

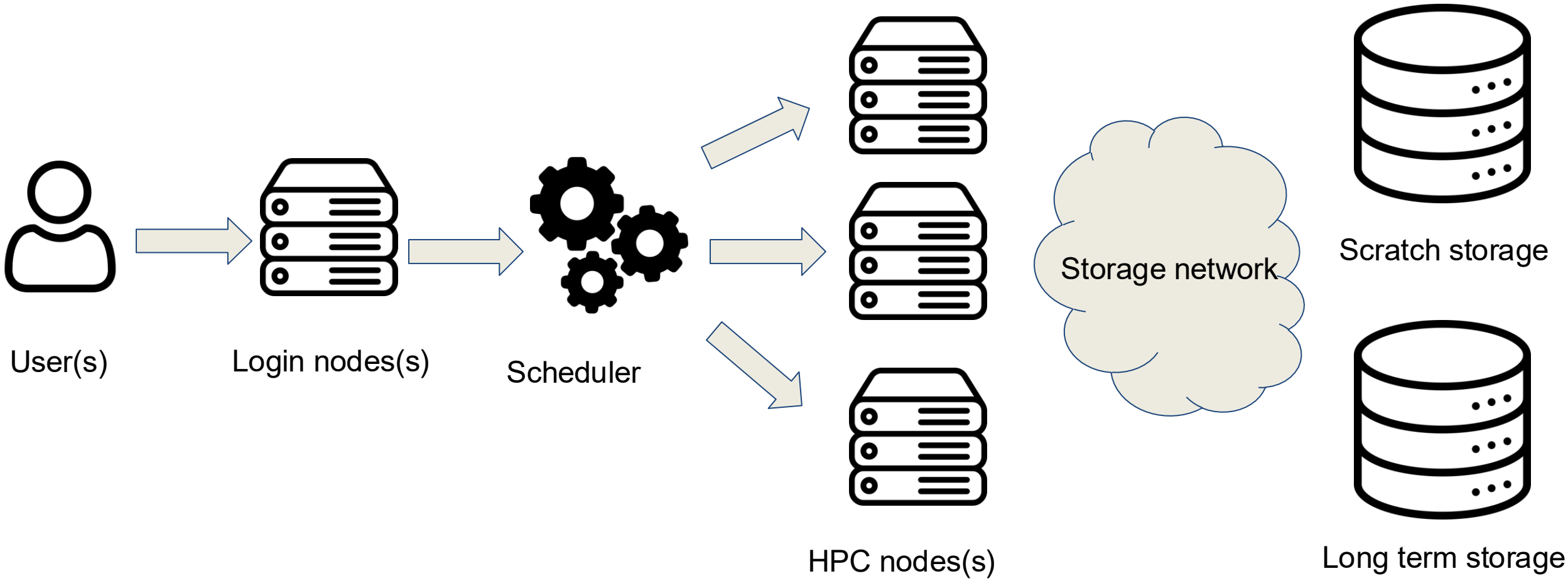
- Scalable on demand resources (**cloud**) and optimized large-scale computing solutions (**HPC**) are required.

- I.E. :
  - EO Data processing into models (**HPC**)
  - On demand pre-processing of EO data (**Cloud**)
  - Global deforestation mapping via the processing of Sentinel-2 and Landsat data in archives spanning over a decade.
    → Cloud for regional analysis.
    → HPC for continental processing

Enabling users from a cloud platform to undergo ML workloads on an HPC system

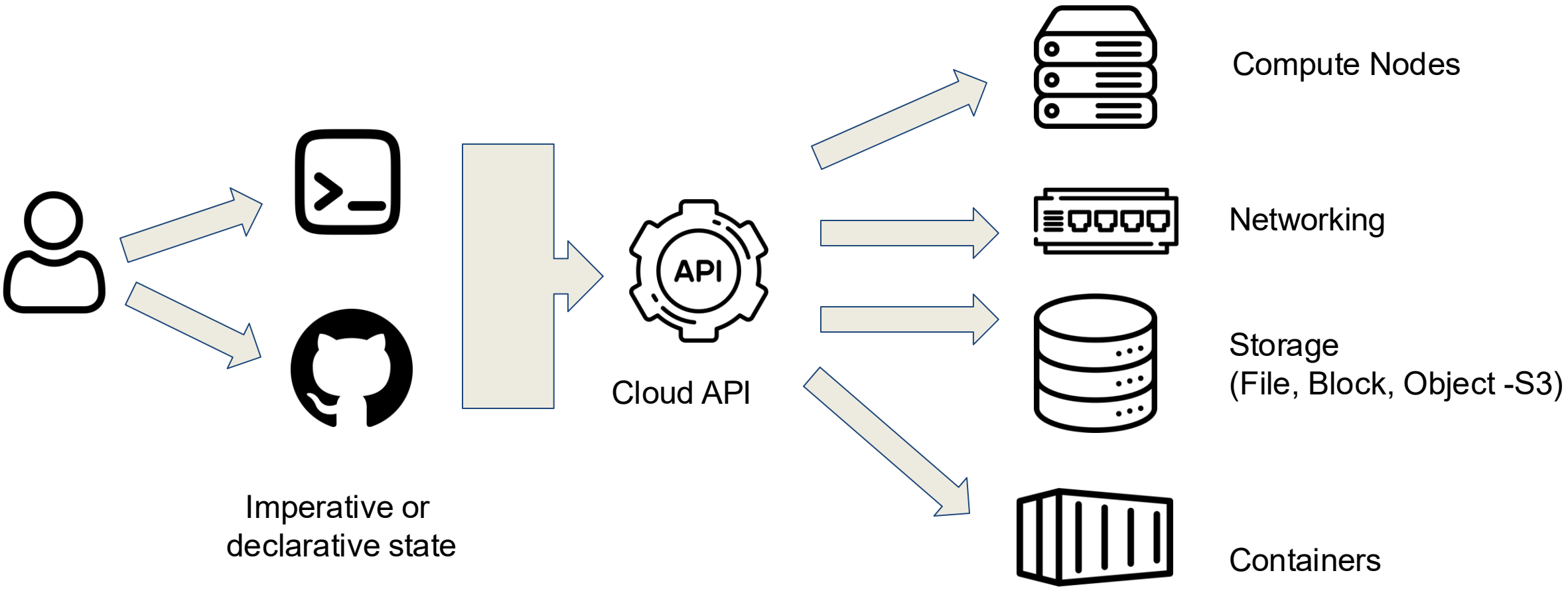
Advances in Technology: Hybrid Cloud / HPC

(Very) Generic HPC System Workflow



Credits: ECMWF

Advances in Technology: Hybrid Cloud / HPC
(Very) Generic Cloud System Workflow



Credits: ECMWF

Advances in Technology: Hybrid Cloud / HPC

Aspect	HPC & Cloud
Parallel Processing	Use multiple compute resources working simultaneously for large or complex tasks
Virtualization	Can use virtual machines or container technologies for workload isolation
Concepts of Distributed systems	Utilize distributed memory, compute, and storage resources
Presource Scheduler & Management	Schedule and manage jobs or workloads across available compute resources

Advances in Technology: Hybrid Cloud / HPC

Aspect	High-Performance Computing (HPC)	Cloud Computing
Purpose	Designed for intensive, tightly-coupled computational tasks	Designed for scalable, on-demand IT services
Infrastructure	Dedicated, specialized supercomputers or on-premise clusters and cloud based (Azure, Amazon etc)	Virtualized, scalable infrastructure shared across multiple users via the internet
Performance	Optimized for low-latency, high-speed interconnects (like InfiniBand), and parallel processing	Performance varies by instance type; often has higher network latency, many recent technologies converging to HPC interconnect speeds).
Resource Management	Centralized job schedulers (e.g., SLURM, PBS) managing compute jobs on dedicated resources	Elastic, on-demand provisioning via orchestration platforms (e.g., K8)
Scalability	Limited by physical hardware; scaling often requires hardware upgrades	Virtually unlimited scalability both horizontal and vertical on demand
Cost Model	High upfront capital expenditure (CapEx) for hardware and infrastructure	Operational expenditure (OpEx) model — pay-as- you-go or reserved pricing
Software Environment	Often customized, tightly integrated with hardware, OS, and libraries	Standardized VM images, containers, or platform services, customizable per deployment
Data Locality	Data typically co-located with compute resources for high- speed access	Data may be remote; storage and compute can be decoupled

Advances in Technology: Hybrid Cloud / HPC
Why it is needed

- Different Strengths for Different Needs: While HPC focuses on performance, the Cloud offers flexible, API-driven, and scalable services for modern applications.

- Hybrid Workflows are Becoming the Norm: Pipelines utilizing both HPC (e.g, model training) and Cloud (e.g, analytics, sharing, visualization) Infrastructure

- Innovation meets Compliance: HPC systems provide controlled, secure, and policy-compliant environments. Cloud platforms enable rapid innovation through DevOps, automation, and modern tooling.

What is Space HPC?

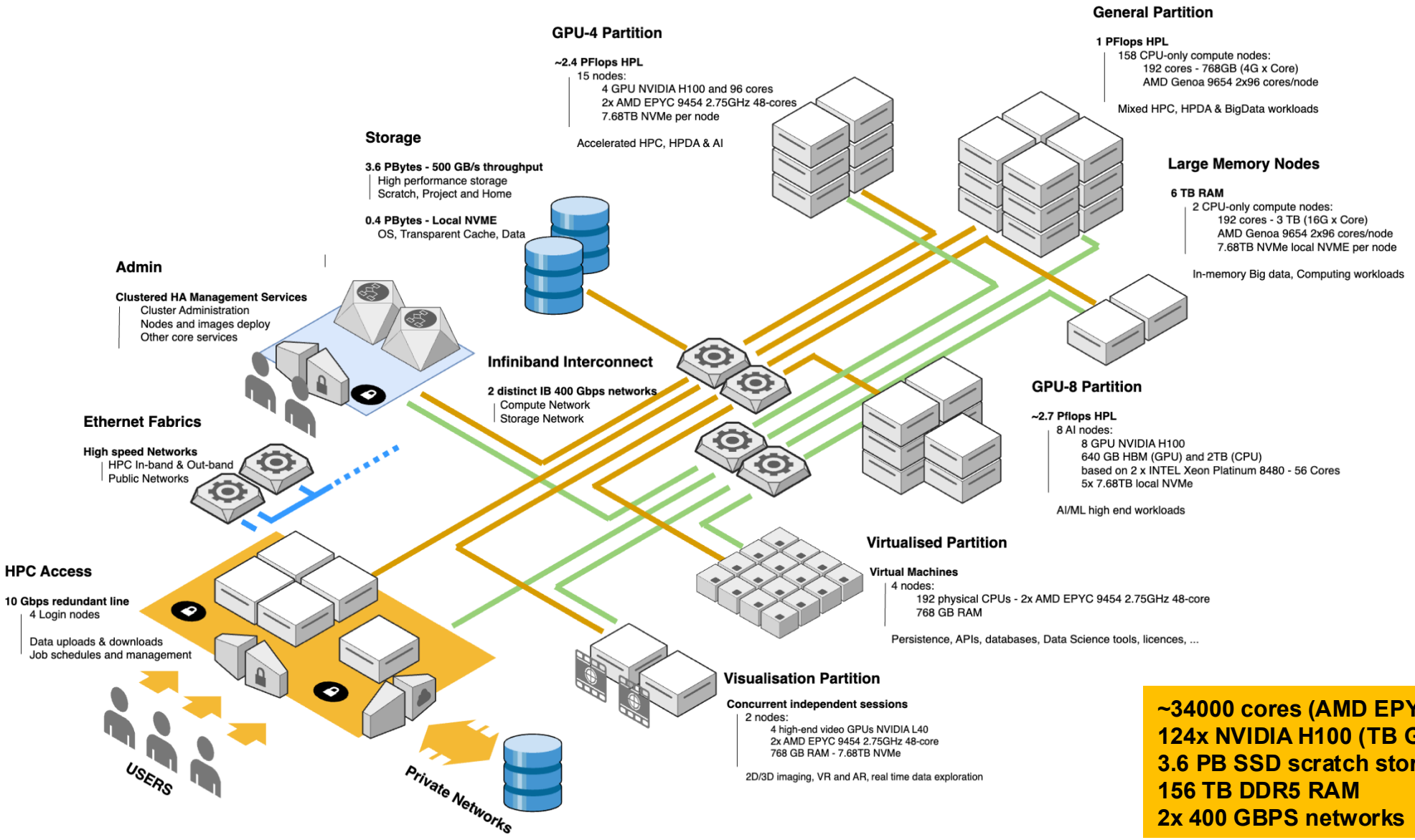
Space HPC at ESA ESRIN

The ESA Space HPC is technology demonstrator, purpose-built for research and development teams aiming to **push the boundaries of innovation in the space industry**.

It was built as a modular system at **ESA ESRIN, Italy**, using the best hardware available at the time.

The use cases for Space HPC range from running **complex simulations, processing large amounts of data to training machine learning models, artificial intelligence (AI) and much more.**

What is Space HPC?



~34000 cores (AMD EPYC)
124x NVIDIA H100 (TB GPU memory)
3.6 PB SSD scratch storage
156 TB DDR5 RAM
2x 400 GBPS networks

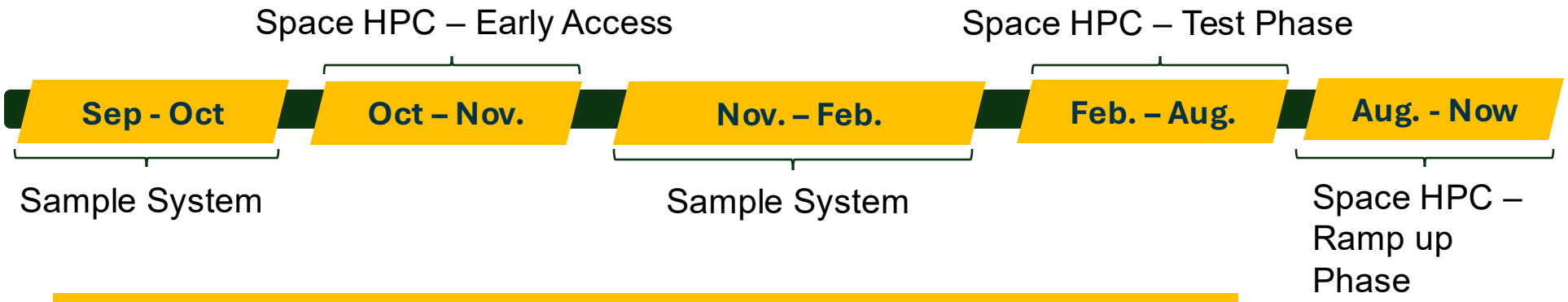
What is Space HPC?

Space HPC was ranked **261st in the TOP500 supercomputers** and **59th in the GREEN500**. (www.top500.org)

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
59	261	**Space HPC** - HPE Cray XD670/XD665, EPYC 9454 / Xeon 8480L, NVIDIA H100 SXM5 80GB, Infiniband NDR200, RHEL 8.10, HPE European Space Agency (ESRIN) Italy	16,368	4.47	106	42.163

Our experiences with the Space HPC so far

We have been testing and refining the system with our users:



Providing them insights into future usage:

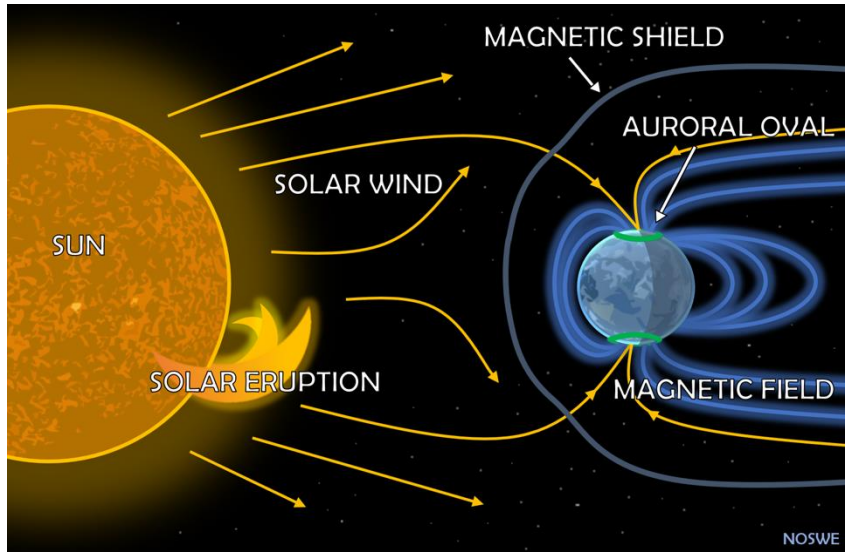
Implicit Neural Network for Greenland Glacier, Peter Naylor:
"Before the SpaceHPC we were limited to the moderate sized data set…"

Mars Elevation Mapping, Pablo Gomez:
"We are reliant on having access to a high-performance computing environment for this particular task…"

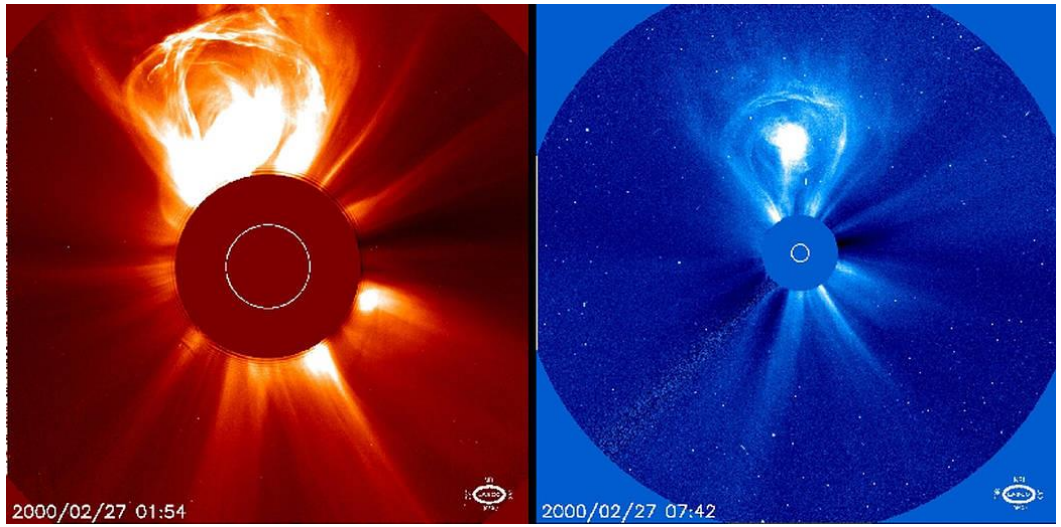
Space Weather Simulation, Jorge Amaya:
"We will be able to provide much faster forecasting and more accurate forecasting…"

Our experiences with the Space HPC so far
Use Case Example – Space Weather Modelling

- Semi-empirical models of the Sun's corona and the heliosphere between 0.1-2 AU.
- EUHFORIA: widely-used model in Python.
- Icarus: faster model in Fortran, using grid stretching and AMR (Adaptive Mesh Refinement) to save CPU time.
- CMEs injected -> we can forecast the arrival time and solar wind parameters at Earth (density, velocity)
- Important because: If we can run the models fast enough after (remotely) detecting that a colossal CME has erupted from the Sun, we can predict how much time is available to prepare for its effects: power grid shutdown, satellite navigation loss, grounded polar flights, etc.



Credit: Norwegian Centre for Space Weather / Tromsø Geophysical Observatory



Credit: SOHO ESA & NASA

Credits: Jorge Amaya, Rosie Wilson

Our experiences with the Space HPC so far

Virtual Machine

8 CPUs

Grenoble test system

Limit: 2 nodes = 384 CPUs

Beta testing of full HPC system

Limit: 64 nodes = 12288 CPUs

~ 2 Days

~ 10 hr

~ 7 min

*** Take it with a grain of salt**
*** Results from running 6 hours of simulation time**

NOAA Space Weather Prediction Center

Credits: Jorge Amaya, Rosie Wilson

Thank you for your time!
Any Questions?