Linear solvers on modern computing architectures: exploiting GPU acceleration and mixed precision

CINES, June 25, 2025

13:30-14:00 Visit of the Adastra supercomputer

14:00-14:30 Mumps Tech
MUMPS: MUltifrontal Massively Parallel Solver for the direct solution of sparse linear equations

14:30-15:15 Gabriel HAUTREUX (CINES, France)
Adastra: an exascale architecture for national research in AI and HPC

15:15-15:45 Coffee Break

15:45-16:35 Thierry GAUTIER and Pierre-Etienne POLET (Inria-LIP, ENS Lyon, France)
On the Use of APU Architectures in MUMPS / XKBlas

16:35-17:00 Théo MARY (CNRS-LIP6, Sorbonne University, France)
Mixed Precision Algorithms in Numerical Linear Algebra

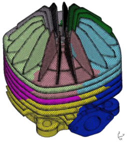
17:00-17:25 Antoine JEGO (LIP6, Sorbonne University, France)
BLAS-based Block Memory Accessors with Applications to Mixed-Precision Sparse Direct Solvers

MUMPS: MUltifrontal Massively Parallel Solver for the direct solution of sparse linear equations

MUMPS group

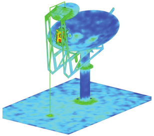
Workshop CINES, June 25, 2025

Sparse direct solvers

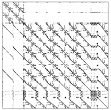


Code Aster (EDF)

Wide range of applications
(e.g. structural analysis, geoscience, electromagnetism, circuit simulation, finite element and optimization ...)



FEKO-EM (Altair)



$\Rightarrow$ Solve $\mathbf{AX} = \mathbf{B}$, with $\mathbf{A}$ a sparse matrix $\Rightarrow$

critical step in HPC simulations

Sparse direct linear solvers

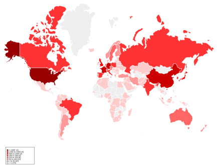
Factor $\mathbf{A} = \mathbf{LU}$; Solve: $\mathbf{LY} = \mathbf{B}$, then $\mathbf{UX} = \mathbf{Y}$

Method of choice for its accuracy and robustness

The MUMPS solver – http://mumps-solver.org

- Free software package (≈ 700 research citations per year, google scholar)
- Fed by the research (15 theses)

Map of the download requests



- First public version: March 2000
- Latest release: MUMPS 5.8.0, May 2025
- License: CeCILL-C
- User community (3 software requests/day)

MUMPS Solver has been awarded in July 2024 by the European Mathematical Society (EMS) and the European Consortium for Mathematics in Industry (ECMI),

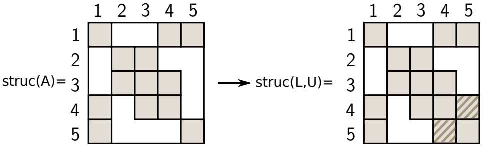
the Lanczos Prize for Mathematical Software

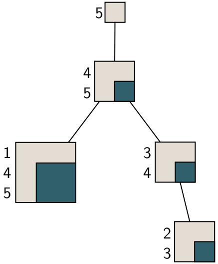
From sparse matrix to dense kernels: the multifrontal approach

Solution of $\mathbf{AX} = \mathbf{B}$ performed in 3 phases:

($\mathbf{A}$ $n \times n$ sparse matrix with NZ non-zeros)

1. **analysis**, on the graph of $\mathbf{A}$
   - build ordering (METIS, SCOTCH, parMETIS, pt-SCOTCH, ...)
   - prepare factorization, build elimination tree





2. **numerical factorization**, decompose $\mathbf{A} = \mathbf{LU}$
   - work on dense matrices following elimination tree
   - stability relies on numerical pivoting
3. **solve**, forward and backward substitutions $\mathbf{LY} = \mathbf{B}$, $\mathbf{UX} = \mathbf{Y}$

Outline

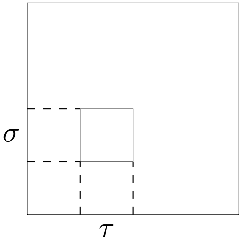
Data sparsity and mixed precision

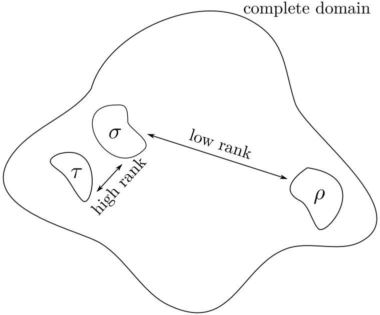
Computer driven algorithms

Performance illustration and concluding remarks

Data sparsity

In some applications the frontal matrices exhibit low-rank blocks

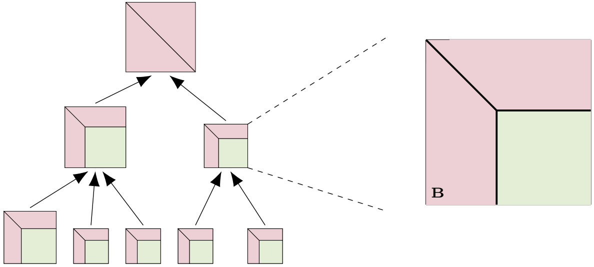


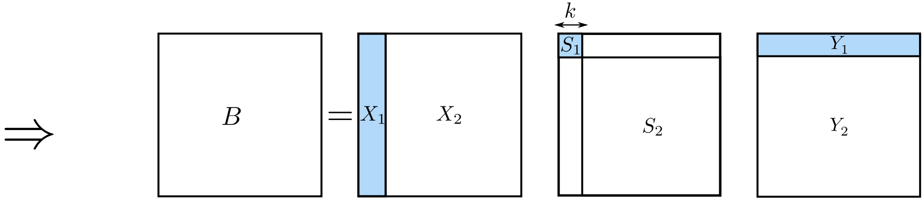


A block B represents the interaction
between two subdomains σ and τ.
Small diameter and far away $\Rightarrow$ low numerical rank.

$\Rightarrow$ Many representations: Recursive $\mathcal{H}, \mathcal{H}^2$, HSS, HODLR, BLR ...

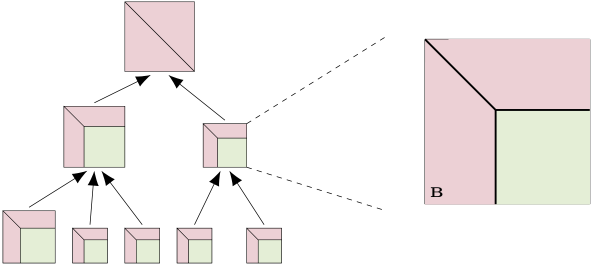
Block Low-Rank Multifrontal feature: principle



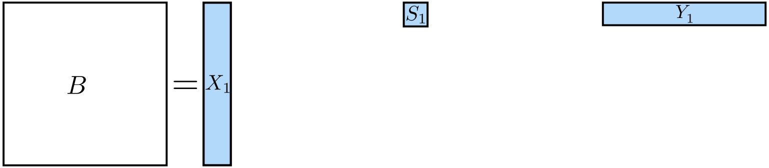


Singular value decomposition (SVD) of each block B
$\Rightarrow B = X_1 S_1 Y_1 + X_2 S_2 Y_2$

Block Low-Rank Multifrontal feature: principle



$\Rightarrow$



rank $k(\varepsilon)$: $B = X_1 S_1 Y_1 \cancel{+ X_2 S_2 Y_2}$ $\qquad \|E\|_2 = \|X_2 S_2 Y_2\|_2 = \sigma_{k+1} \leq \varepsilon$

Block Low-Rank (BLR) main features and properties

- BLR is based on a flat 2D block partitioning, compatible with features of a general solver

> P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, and C. Weisbecker. *"Improving Multifrontal Methods by Means of Block Low-Rank Representations"*. In: SIAM SISC (2015).

Block Low-Rank (BLR) main features and properties

- BLR is based on a flat 2D block partitioning, compatible with features of a general solver

> P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, and C. Weisbecker. *"Improving Multifrontal Methods by Means of Block Low-Rank Representations"*. In: SIAM SISC (2015).

- BLR reduces asymptotic complexity:

  Complexity reduction (3D Poisson, $n = N \times N \times N$ mesh, BLR rank bound in $O(1)$):

  $O(n^2) \rightarrow O(n^{4/3})$ flops
  $O(n^{4/3}) \rightarrow O(nlog\ n)$ memory

> P. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. Mary. *"On the Complexity of the Block Low-Rank Multifrontal Factorization"*. In: SIAM SISC (2017).

Block Low-Rank (BLR) main features and properties

- BLR is based on a flat 2D block partitioning, compatible with features of a general solver

  P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, and C. Weisbecker. *"Improving Multifrontal Methods by Means of Block Low-Rank Representations"*. In: SIAM SISC (2015).

- BLR reduces asymptotic complexity:

  Complexity reduction (3D Poisson, $n = N \times N \times N$ mesh, BLR rank bound in $O(1)$):

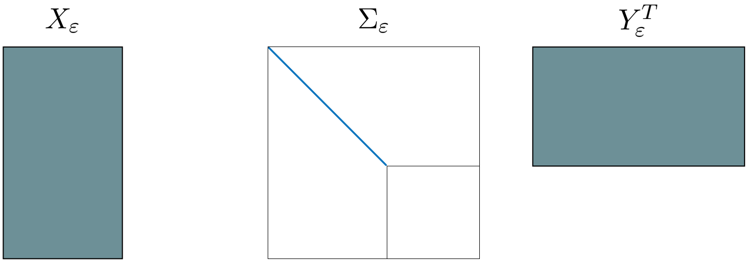
  $O(n^2) \rightarrow O(n^{4/3})$ flops
  $O(n^{4/3}) \rightarrow O(nlog\ n)$ memory

  P. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. Mary. *"On the Complexity of the Block Low-Rank Multifrontal Factorization"*. In: SIAM SISC (2017).

- BLR is backward stable

  N. Higham and T. Mary. *"Solving Block Low-Rank Linear Systems by LU Factorization is Numerically Stable"*. In: IMA J. Numer. Anal.(2021).

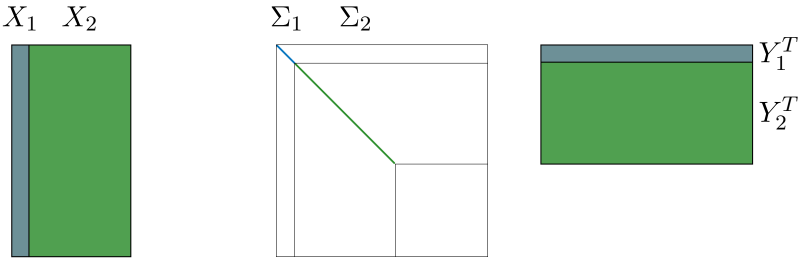
Mixed precision Block Low-Rank approximation



Truncated SVD

- $B = \sum_{k=1}^{r} x_k \sigma_k y_k^T$, with r such that
- $\|B - X_\varepsilon \Sigma_\varepsilon Y_\varepsilon^T\| \le \varepsilon \|A\|$

Mixed precision Block Low-Rank approximation



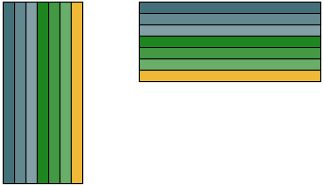
Truncated SVD with 2-precision formats (fp64, fp32)

- The idea: convert X_2 and Y_2 to single precision (fp32)
- Criterion for storing columns x_i and y_i in precision fp32: $\sigma_i \leq \frac{\varepsilon}{u_s}\|A\|$, with $u_s = 6 \times 10^{-8}$
- $\|B - X_1\Sigma_1 Y_1^T - X_2\Sigma_2 Y_2^T\| \lesssim 3\,\varepsilon\,\|A\|$

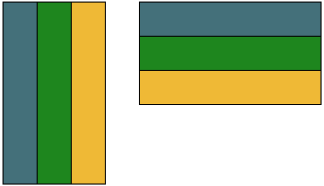
Mixed BLR: dissociate storage and compute precisions

Exploiting precisions for computations other than fp64 and fp32 is hardware dependent **but** mathematical theory applies to any number of precisions[1]

P. Amestoy, O. Boiteau, A. Buttari, M. Gerest, F. Jézéquel et al.. *"Mixed Precision Low Rank Approximations and their Application to Block Low Rank LU Factorization"*. In: Journal of Numerical Analysis (2022) .



Storage precisions:
large number, arbitrary format

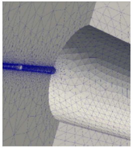


Compute precisions:
small number, available in hardware

[1] work supported by EDF, PhD grant of M. Gerest

Mixed precision Block Low-Rank approximation: results

- thmgaz (thermo-hydro-mechanics) matrix ($n = 5M$)
  - Factor size (Full-Rank): **141 GigaBytes**
  - BLR: $\varepsilon = 10^{-10}$

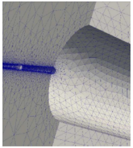


(from code_aster)

Olympe computer (CALMIP), 2MPI×18threads					
	Factor size (GigaBytes)	Total memory (GigaBytes)	Factorization time (sec)	Solve time (sec)	Backward error
fp64 BLR	**103**	132	61	1.7	4×10^{-14}

Mixed precision Block Low-Rank approximation: results

- thmgaz (thermo-hydro-mechanics) matrix ($n = 5M$)
  - Factor size (Full-Rank): **141 GigaBytes**
  - BLR: $\varepsilon = 10^{-10}$
  - Mixed BLR: 2/7 precisions for LU storage, and
    2 precisions during solve computation

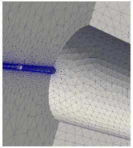


(from code_aster)

	Olympe computer (CALMIP), 2MPI×18threads				
	Factor size (GigaBytes)	Total memory (GigaBytes)	Factorization time (sec)	Solve time (sec)	Backward error
fp64 BLR	**103**	132	61	1.7	4×10^{-14}
Mixed BLR(2)	**80**	120	68	1.9	5×10^{-14}

Mixed precision Block Low-Rank approximation: results

- thmgaz (thermo-hydro-mechanics) matrix ($n = 5M$)
  - Factor size (Full-Rank): **141 GigaBytes**
  - BLR: $\varepsilon = 10^{-10}$
  - Mixed BLR: 2/7 precisions for LU storage, and
    2 precisions during solve computation



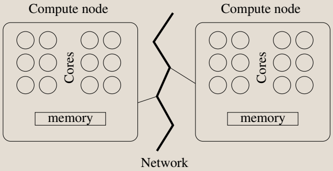
(from code_aster)

Olympe computer (CALMIP), 2MPI×18threads					
	Factor size (GigaBytes)	Total memory (GigaBytes)	Factorization time (sec)	Solve time (sec)	Backward error
fp64 BLR	**103**	132	61	1.7	4×10^{-14}
Mixed BLR(2)	**80**	120	68	1.9	5×10^{-14}
Mixed BLR(7)	**67**	111	68	2.1	5×10^{-14}

$\Rightarrow$ significant memory gains considering 7 precisions w.r.t. 2 precisions

Outline

Managing parallelism



Many cores sharing memory per compute node

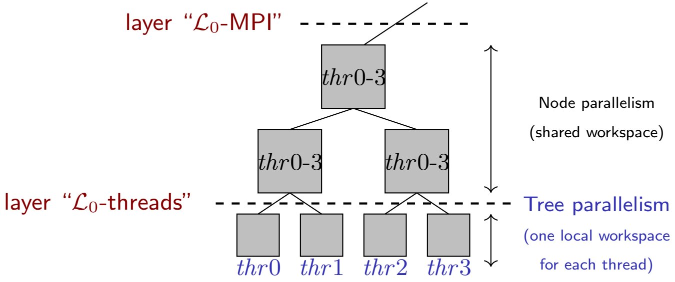
Hybrid parallelization

- Distributed memory parallelism (MPI based) combined to
- shared memory parallelism (multithreading):
  - use of multithreaded BLAS
  - OpenMP directives
  - multithreading between independent tasks

Hybrid parallelism

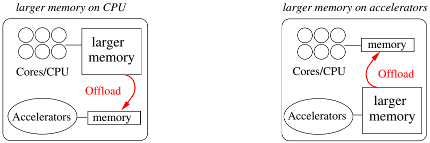
Strategy for hybrid parallelization (case of multiple threads per MPI process):

- under "$\mathcal{L}_0$-MPI": **one** MPI process per subtree (to limit communication)
- **one** thread per subtree



Exploit accelerators

Types of compute nodes with accelerators



- Larger memory on CPU: offload from CPU to GPU, use runtime libraries for BLAS on GPU:
  - cublasXt: provided by Nvidia
  - `XKBlas`: collaboration with Inria-ENS Lyon, also supports AMD GPU
  - efficiency relies on exploiting both CPU and GPU
- Larger memory on accelerator, most data and related computing on GPU
- Unified memory should enable to use the best of CPUs and GPUs

Offload from CPU to GPU

External libraries can take care of tiling, allocation/memory management on GPU, CPU $\leftrightarrow$ GPU data transfers

- cublasXt: provided by NVIDIA
- XKBlas: collaboration with T. Gautier[2] (LIP laboratory, ENS Lyon)

Offload approach

if Arithmetic Intensity of frontal matrix "large enough" (AI_Threshold) **then**
  Adjust blocking; asynchronous memory pinning
  Wrap `GEMM/TRSM` to call cublasXt or XKBlas
else
  Standard multicore processing of frontal matrix
end if

AI_Threshold depends on cublasXt or XKBlas, GPU type, CPU cores

[2] Gautier et al., A Runtime System for [...] on Heterogeneous Architectures [. . .], IPDPS 2013

Outline

Data sparsity and mixed precision

Computer driven algorithms

Performance illustration and concluding remarks

Experiments on Adastra (CINES), CPU partition (AMD-Genoa)

Adastra CPU partition (536 nodes)

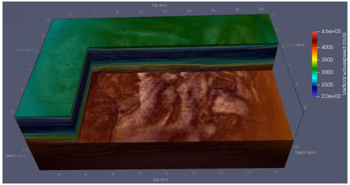
- 192 AMD cores: bi-procs AMD GENOA with 96 cores each (4th Gen AMD EPYC 9654, 2.4GHz)
- 768 GBytes memory/node

Applications in Seismic imaging

- Two Adastra CPU CINES Grand Challenge projects in seismic imaging:
  - *"Large scale modelisation of harmonic waves based on high order polynomials Hybridizable Discountinous Galerkin (HDG) method"*
    led by Makutu team (Inria-TotalEnergies), (45 millions CPU hours).
  - *"MUMPS4FWI (Full Waveform Inversion using MUMPS direct solver)"*
    led by WIND project (UMR Géoazur, Sophia Antipolis, France), (37 millions CPU hours).

Efficiency of a large simulation in Full-Waveform Inversion [3]

- Adastra MUMPS4FWI project led by WIND team
- Application: Gorgon Model, reservoir 23km x 11km x 6.5km
- Single precision complex matrix, 531 Million dofs
- Single complex flops for one LU factorization:
  Full-Rank: 2.6×10^{18}; BLR ($\varepsilon_{BLR} = 10^{-5}$): 0.5×10^{18};

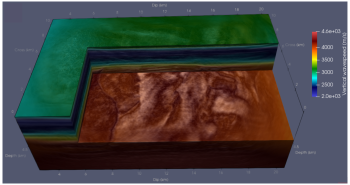


(25-Hz Gorgon FWI velocity model)

[3] Work presented at EAGE 2024 conference *"Pushing the limits of 3D frequency-domain FWI with the 2015/2016 OBN Gorgon dataset"*

Efficiency of a large simulation in Full-Waveform Inversion [3]

- Adastra MUMPS4FWI project led by WIND team
- Application: Gorgon Model, reservoir 23km x 11km x 6.5km
- Single precision complex matrix, 531 Million dofs
- Single complex flops for one LU factorization:
  Full-Rank: 2.6×10^{18}; BLR ($\varepsilon_{BLR} = 10^{-5}$): 0.5×10^{18};



(25-Hz Gorgon FWI velocity model)

Performance analysis

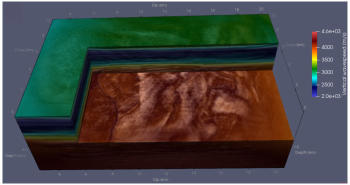
Theoretical peak performance

- Simulation performed on 48 000 cores
  (500 MPI $\times$ 96 threads/MPI)
- Peak perf.: 3686 TFlops/s (single real flops)
  (500$\times$96$\times$2.4GHz$\times$(2 (single real)$\times$ 16 flops/cycle)

[3] Work presented at EAGE 2024 conference *"Pushing the limits of 3D frequency-domain FWI with the 2015/2016 OBN Gorgon dataset"*

Efficiency of a large simulation in Full-Waveform Inversion [3]

- Adastra MUMPS4FWI project led by WIND team
- Application: Gorgon Model, reservoir 23km x 11km x 6.5km
- Single precision complex matrix, 531 Million dofs
- Single complex flops for one LU factorization:
  Full-Rank: 2.6×10^{18}; BLR ($\varepsilon_{BLR} = 10^{-5}$): 0.5×10^{18};



(25-Hz Gorgon FWI velocity model)

Performance analysis

Theoretical peak performance

- Simulation performed on 48 000 cores
  (500 MPI $\times$ 96 threads/MPI)
- Peak perf.: 3686 TFlops/s (single real flops)
  (500 $\times$ 96 $\times$ 2.4GHz $\times$ (2 (single real) $\times$ 16 flops/cycle)

Effective performance

- BLR flops $= 2 \times 10^{18}$ single real flops
  $((2+6)/2 \times 0.5 \times 10^{18})$
- Time for factorization: 5946 sec
- Effective performance: 336.4 TFlops/s
  $(2 \times 10^{18}/5946)$
  $\rightarrow$ 9% of the peak (w.r.t effective BLR flops)

[3] Work presented at EAGE 2024 conference *"Pushing the limits of 3D frequency-domain FWI with the 2015/2016 OBN Gorgon dataset"*

Modeling of time-harmonic waves with HDG method

- Context: Adastra high order polynomials HDG method, Makutu team (Inria-TotalEnergies)
- Application: Helmholtz equation, polynomials orders: 3-8
- Complex matrix, 1050 Million dofs, storage(A)=1.5 TBytes; order($G(A)$)= 84 M
- Full-Rank (FR) cost: flops for one LU factorization= 1.2×10^{17};
  estimated storage for LU factors= 13 TBytes

48 000 cores (1000 MPI $\times$ 48 threads/MPI); BLR with $\varepsilon_{BLR} = 10^{-7}$; FR: fp32;
Mixed precision BLR: 3 precisions (32bits, 24bits, 16bits) for storage

LU size (TBytes)			Flops		Time BLR + Mixed (sec)			Scaled Resid.
FR	BLR	+mixed	FR	BLR+mixed	Analysis	Facto	Solve	BLR+mixed
13	7	5	1.2×10^{17}	1.9×10^{16}	550	1384	22	$\approx \times 10^{-5}$

in practice: hundreds to thousands of Solve steps

Concluding remarks

Linear algebra is at the heart of numerical simulation;
computer architecture evolution strongly influences our algorithms
and need to be anticipated

- Architecture of exascale computers need to be analysed/understood
  → see talk of Gabriel HAUTREUX (CINES, France),
  Adastra: an exascale architecture for national research in AI and HPC
- Accelerators plays an important role in computer evolution
  → see talk of Thierry GAUTIER and Pierre-Etienne POLET (Inria-LIP, ENS Lyon, France),
  On the Use of APU Architectures in MUMPS / XKBlas
- Low precision storage and computation is a promising research axis for large applications:
  → see talks of Théo MARY (CNRS-LIP6, Sorbonne University, France)
  Mixed Precision Algorithms in Numerical Linear Algebra
  → and of Antoine JEGO (LIP6, Sorbonne University, France)
  BLAS-based Block Memory Accessors with Applications to Mixed-Precision Sparse Direct Solvers

Workshop on Approximate Computing in NLA, 7-10 Oct 2025

- Location: Sorbonne University (4, place Jussieu), Paris
- Dates:
  - Habilitation defense of T. Mary: 7 Oct afternoon
  - Workshop: 8-10 Oct
- Program available online! 54 talks and posters on mixed precision, low-rank approximations, randomization, emulation, direct and iterative solvers, preconditioners, multigrid, tensors, ...
- Registration is free but mandatory, limited number of seats remaining!

https://approxcomputing.sciencesconf.org/

Experimental environment

- CALMIP center of Toulouse (grant number P0989):

  `Olympe nodes`

  - CPU node: Two Intel 18-cores Skylake 6140 @2.3 GHz (Peak/core=73.6 GF/s, Peak/node=2.6 TFlops/s FP64), 192 GB memory per node
  - GPU node: Two Intel 18-cores Skylake 6140 @2.3 GHz (Peak/core=73.6 GF/s, Peak/node=2.6 TFlops/s FP64), 384 GB memory per node, 4 GP-GPU Nvidia Volta (V100 - 7.8 TFlops/s FP64)
- GENCI-CINES, ADASTRA supercomputer: HPE Cray EX235a
  - 61.6 PFlops/s peak, 46 PFlops/s (Linpack); 50 GFlops/Watt
  - Partition with accelerated nodes (338 nodes):
    - accelerated nodes based on AMD Optimized 3rd Generation EPYC 64C 2.0 GHz, 512 GB on four AMD Instinct MI250X GPU, 256 GB on CPU
  - Partition with CPU nodes (536 nodes):
    - 192 AMD cores: bi-procs AMD GENOA with 96 cores each (4th Gen AMD EPYC 9654, 2.4GHz)
    - 768 GBytes memory/node
- Nvidia GraceHopper
  - 72 core ARM @ 3.0Ghz; DDR 480GB @ 384 GB/s
  - GPU: Nvidia H100 (34 Tflops/s FP64); HBM 96GB @ 4000 GB/s; comms: 450GB/s full duplex