RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

FRANCE 2030
PROGRAMME DE RECHERCHE
NUMÉRIQUE POUR L'EXASCALE

WP4 : Numerical Libraries

Marc Baboulin and Abdou Guermouche

Work Package Description

Work Package Objectives

Design Efficient and scalable numerical libraries

- Target linear and multilinear algebra (tensor) computations
- Extensive use of task-based programming models and runtime systems
- Speed-up computations by relying on mixed precision and approximate computing
- Explore strategies to enhance composability of our numerical libraries.
- Strong interaction with other Work Packages : WP3 (Runtime Systems), WP1 (Programming Models)

Targeted Software

- **Chameleon** - C++ interface / Scalability / extension to tensor computations
- **Composyx** - Exploring composability for high level linear algebra methods
- **Celeste** - Parallelism enhancement through the use of StarPU

Structure of the Work Package

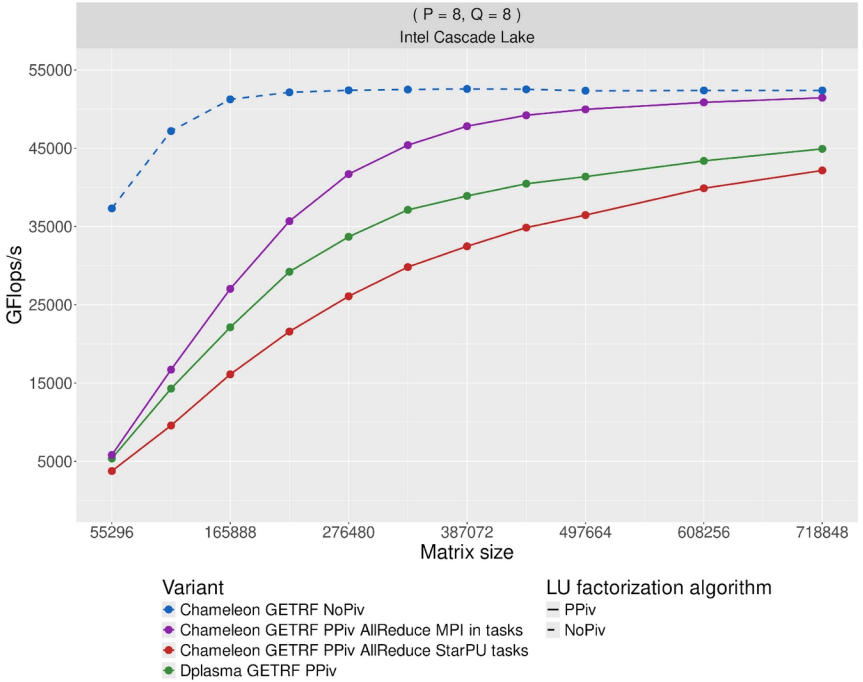
- Task 4.1 - Composability of numerical libraries (collab with WP1) - recruitments
- Task 4.2 - Towards scalable dense and sparse linear algebra using task-based programming models (collab. with WP3) - recruitments
- Task 4.3 - Efficient implementation of approximate computing algorithms (collab with ExaMA) - recruitments
- Task 4.4 - Sparse and dense tensor computations using task-based algorithms - recruitments
- Task 4.5 - Extension of Chameleon to small dimension tensors for large distributed systems with applications to deep neural networks - recruitments

4 PhD. students and 1 engineer already hired. Still needing to hire 1 engineer and 1 PhD. student

Achievements and Highlights

Chameleon

- Advanced Chameleon C++ standard interface (compliant with std::linalg)
  - Full interface support, Extension explored for Lapack-like operations, Asynchronism (ongoing)
- Advanced and scalable partial pivoting on top of task-based runtime systems (collab. with Eviden)
  - Use batching techniques to reduce the number of tasks submitted to the runtime system
  - Novel task-based reduction algorithms for distributed memory executions



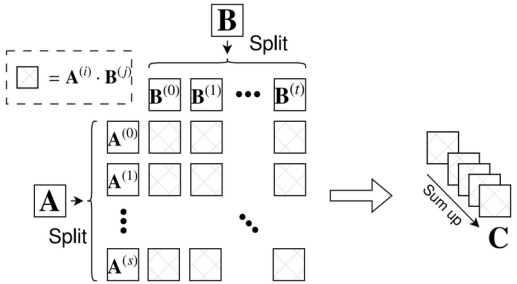
Performance of the Chameleon GETRF operation compared with DPLASMA and MKL on 32 intel Cascade Lake nodes (2 MPI processes per node having 18 cores each)

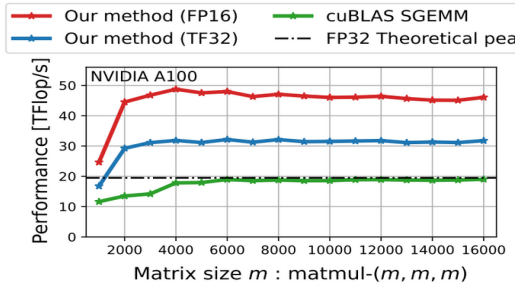
Precision Emulation

Joint work between ExaMA, Exa-SofT and NVIDIA

Double precision GPU units tend to become slower across generations

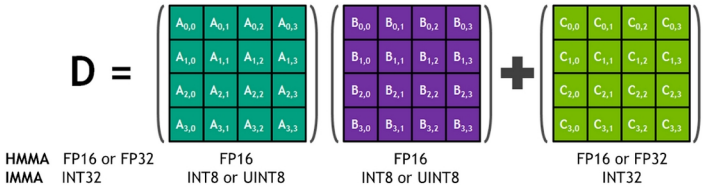
⟹ Emulating high precision with low precision hardware (or even integer units) (Ozaki et al. 2012, Fasi et al. 2023)





Same precision but better performance than FP32 since TF32 & FP16 units are much faster than FP32

(arXiv:2203.03341)

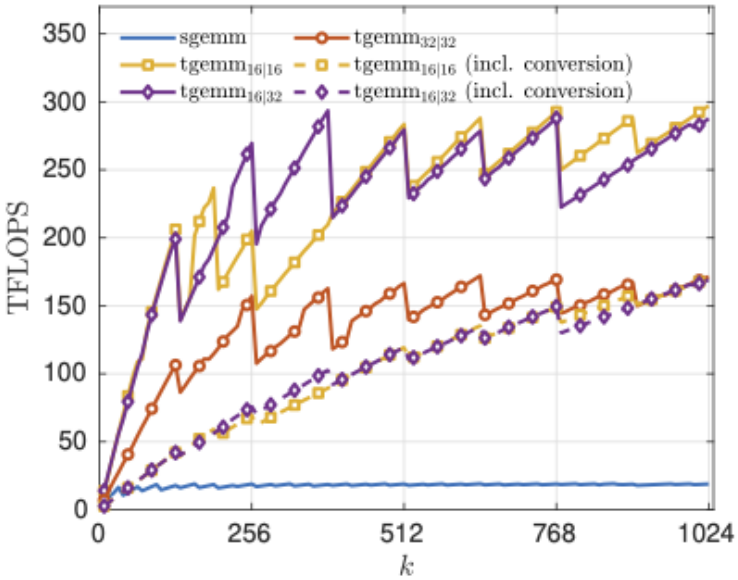


- 2024-2025: discussions with NVIDIA
- 12/09/2025: ExaMA-ExaSofT-NVIDIA meeting to finalize a formal collab

Approximate Computing

Low rank approximations

- Perform low rank approximation of matrices
- Combine mixed precision and randomization
- Leverage performance of GPU tensor cores (A100)
- Presented at Europar 2024



Tensor decompositions

- Perform low rank approximation of tensors and matrices
- Use mixed precision iterative refinement
- Accepted in SIAM SISC (2025)

Future Work

- Fruitful collaborations with industry (Nvidia, Eviden)
- Ongoing collaborations with other WP (1 and 3)
- Planning is so far respected

Next steps

- Exploratory work on T4.1 with WP1
- Intensify WP animation (talks on a monthly basis ?)
- Issues in recruiting (engineer, PhD student)
- Make software available for other Numpex teams
- Interaction with the ExaDI projects. Need for demonstrators with numerical libraries needs

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

FRANCE 2030

PROGRAMME DE RECHERCHE

NUMÉRIQUE POUR L'EXASCALE

Retrouvez toutes nos actualités

NumPEx