RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030
PROGRAMME
DE RECHERCHE
NUMÉRIQUE
POUR L'EXASCALE

WP5 : Optimization

Prof. El-Ghazali TALBI

Objectives and tasks

Objectives

- Solving large-scale optimization problems (decision variables, many-objectives, expensive objectives, big data) using Exascale optimization algorithms
- Inverse, continuous, and discrete optimization problems

Tasks

1. **Exascale discrete and continuous optimization**
   - Exact optimization (Branch and bound, tree search)
   - Heuristic optimization (Computational intelligence)
2. **Exascale surrogate-based and Bayesian optimization**
   - Parallel coupling of surrogates, optimization and sampling
3. **Exascale shape optimization**
   - Involving multi-physics models
4. **Exascale optimization for AutoML** (Optimization of deep/large ML models)

Main Partners

- Inria Bonus
- Unistra

- 2 Phds
- Still to hire
  - 1 Engineer
  - 1 Postdoc

Progress made

1. Discrete optimization

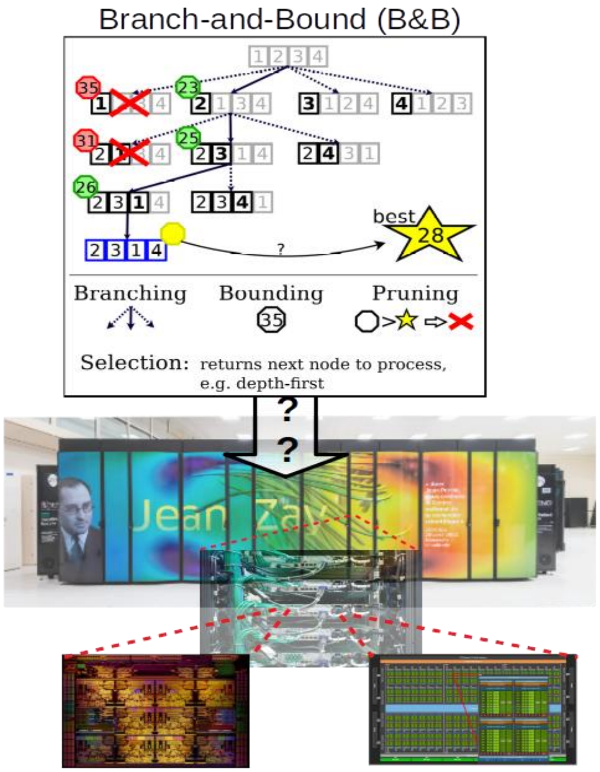
- **Scaling Branch and Bound (B&B) algorithms**
- Design of ultra-scale B&B dealing with both …
  - Search tree and supercomputer characteristics: **large, heterogeneous, unreliable**

➔ Scalable and unified data structures

➔ Work Stealing-based load balancing at different levels

➔ Handling of **GPU**-based heterogeneity at the intra-node level

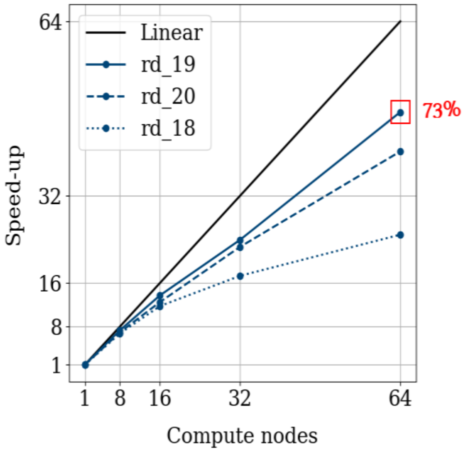
- Investigation of MPI+X *vs.* PGAS (Chapel)
- Solving hard pending discrete benchmarks (*e.g.* scheduling)



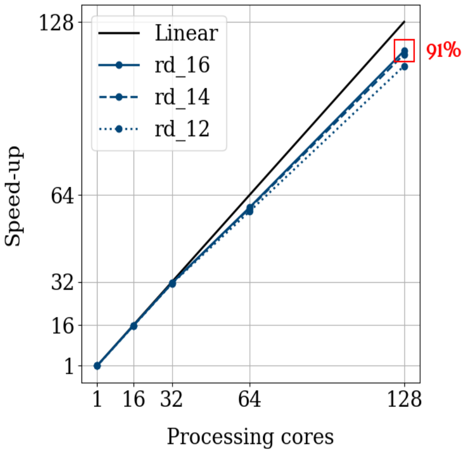
1. Discrete optimization

- Experimentation on ultra-scale supercomputers (CPU/GPU)
  - Frontier, LUMI, Perlmutter, TGCC/Irene, MeluXina, ...
  - Up to 51 200 CPU cores and 1 024 GPUs
- High scalability

Intra-node



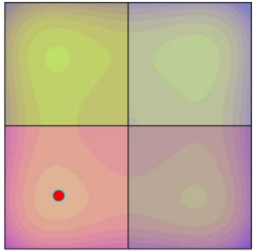
Inter-node

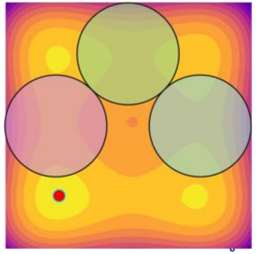


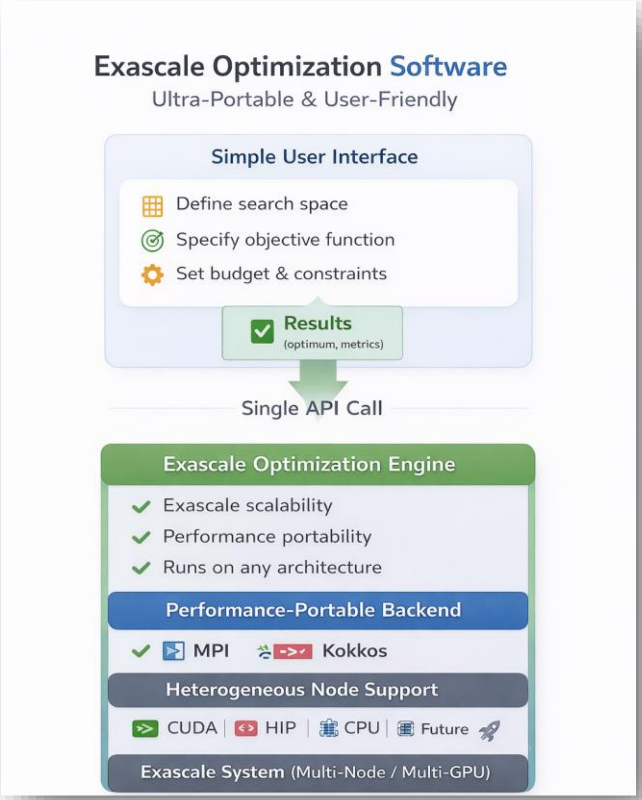
1. Continuous optimization

New Innovative Algorithms: Decomposition-based methods

- Fractal-based decomposition
- Exascale scalability
- Software framework Zellij
  - Tree search (non regular, dynamic)
- MPI + Kokkos
- Portability
- Multi-node, multi-GPUs

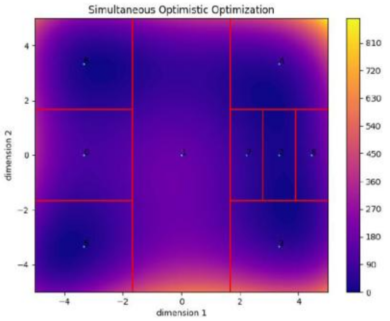


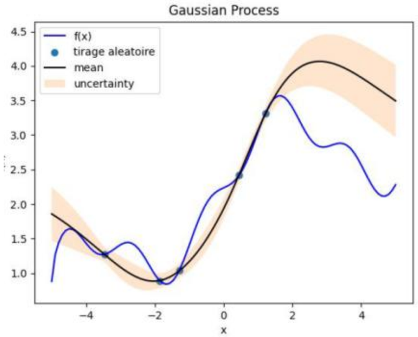




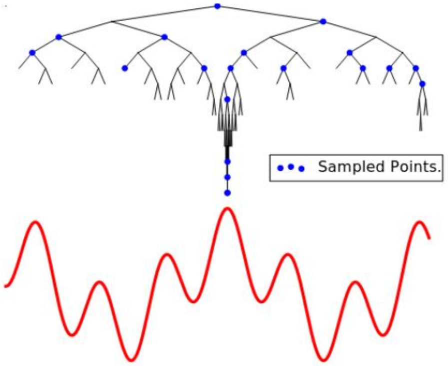
2. Bayesian and surrogate-assisted optimization

- Fractal based Bayesian Optimization
  - Exploration based on Fractals decomposition
  - Scoring based on acquisition functions on Fractals





- Surrogate-assisted for Fractal optimization
  - Any Supervised ML model can be used : GP, …



3. Inverse shape optimization

- Development of a **complete mathematical framework**
  - PDE-constrained shape optimization (Poisson problem)
  - Shape deformation via **volume-preserving ODE flows**.

- **Key message:** A constrained shape optimization problem is reformulated into a parametric optimization framework suitable for large-scale computation.

- **Software development & HPC-oriented design**
  - **Open-source software package** (GeSONN)
  - **HPC-driven design choices** : Multiple independent PDE solvers, Gradient evaluations, neural network training on GPUs

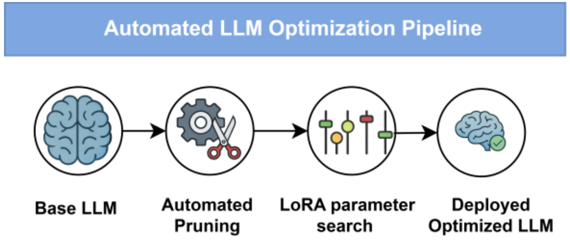
- Numerical scalability & computational challenges

4. AutoML optimization

- AutoML
  - Deep neural networks, spiking neural networks, LLMs
  - Hyper-parameter optimization, neural architecture search of DNN, SNNs
  - Fine tuning, Pruning of LLMs

- Target optimization problems
  - Big mixed optimization problem (i.e. very expensive objective function)
  - Variable-size search space
  - Multi-objective: Accuracy, energy consumption, complexity, …

- Optimization algorithms
  - Evolutionary algorithms
  - Bayesian optimization
  - Ultra-scale Fractal-based Bayesian optimization



Scientific highlights

1. Fractal-based Bayesian Optimization

- Exascale ready

- **Superlinear speedup** thanks to the distributed multiple surrogate approach peeking at 10 nodes with 463.1%

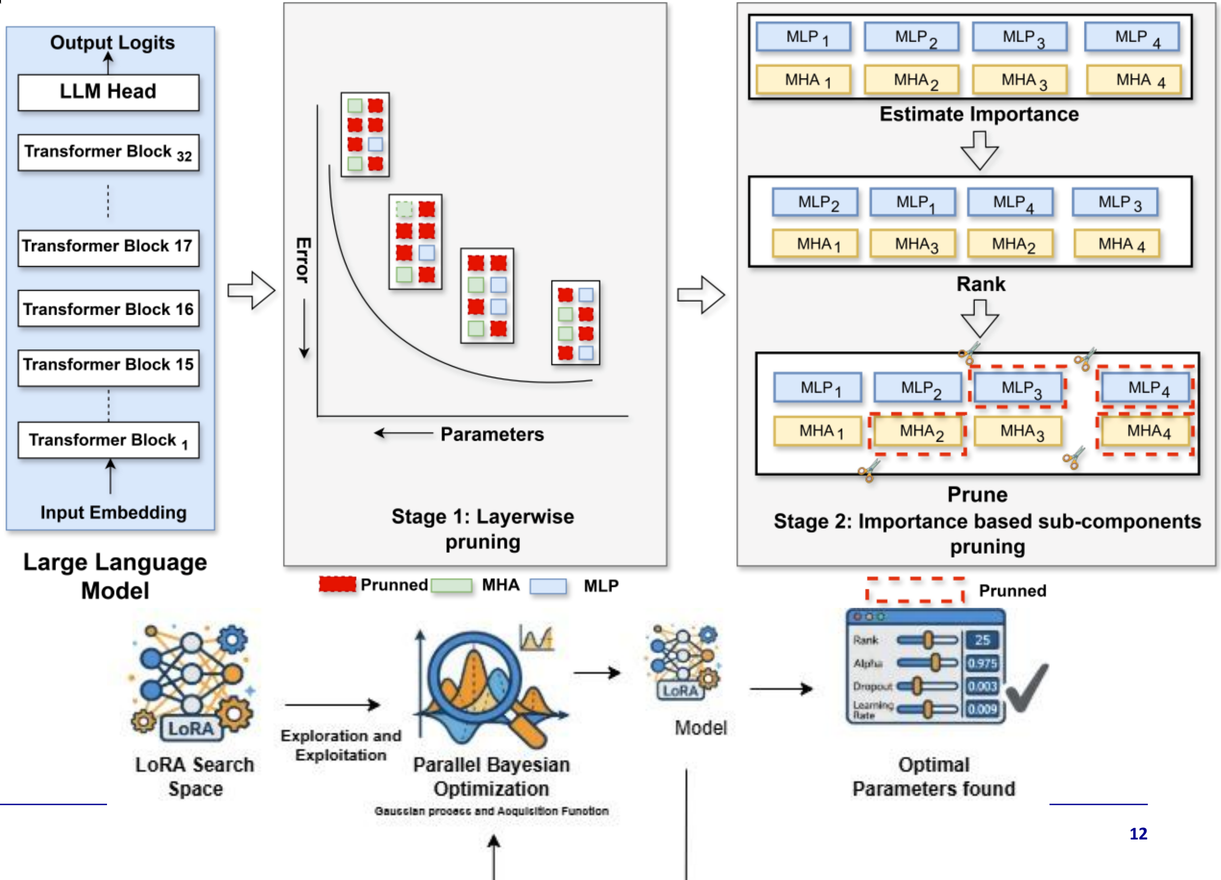
- Up to **84.4% strong scaling** efficiency at 1000 Lumi-G nodes – 8000 GPUs

Nodes	Speed up	Scaling Efficency
1	1.0×	100.0%
10	46.3×	463.1%
100	227.8×	227.7%

Nodes	Speed up	Scaling Efficency
100	1.0×	100.0%
200	3.2×	159.7%
1000	8.5×	84.4%

2. Hierarchical Two-Stage Pruning for Efficient LLMs

- A two-stage strategy for **LLM pruning & fine tuning**
- Hierarchical decomposition of the search space
- **Multi-objective Bayesian algorithm** to find Pareto optimal architectures
  - Accuracy
  - Complexity



Pruning and fine tuning of LLMs results

- Evaluation of the **Llama-2.7b** (6,74B paramètres) pruned with 30% sparsity ratio on multiple benchmarks
- Assessment on multiple benchmarks: ARC-Easy, ARC-Challenge, PIQA
- Proposed approach shows overall better performance
- Experimental Setup: The results are evaluated on same pruning ratios

Method	ARC-Easy	ARC-Challenge	PIQA	Average
SliceGPT [1]	51.77	31.23	63.55	48.85
LLM Surgeon [2]	60.72	36.69	63.55	53.65
DISP-LLM [3]	60.10	37.03	73.72	59.90
AMP [4]	64.31	39.85	74.21	59.45
2SSP [5]	52.65	27.39	70.29	50.11
Proposed (Ours)	**73.23**	**40.52**	**76.93**	**63.56**

Next steps

1. Optimization 2. Bayesian Optimization

- **Fault-tolerance**
- **Application to Variational Quantum Eigen Solver problems**
  - **Classical VQE**
    - Sequential Gradient algorithm on classical computer
    - Quantum evaluation of solutions
  - **Our new approach**
    - Fractal algorithm on Exascale computer
  - **Exascale-ready preliminary results**
    - Completely black-box
    - Efficiency at 1000 nodes of Lumi-G corresponding to 8000 GPUs
    - Aggregate theoretical peak 383 PFLOP/s in double precision

Classical computer (Gradient) ↔ Quantum computer (Evaluation)

Exascale computer (Fractal) ↔ Quantum computer (Evaluation)

3. Inverse optimization

- **Reach constraints in shape optimization**
- **Scaling up with HPC**
  - Extension to realistic 3D problems: complex geometries, complex topologies
  - Full exploitation of HPC resources: Distributed neural network training, Hybrid CPU/GPU workflows.
- **Long-term perspective**
  - Generic framework for: PDE-constrained optimization and shape optimization under geometric constraints.
  - Potential applications:
    - Mechanics,
    - Heat transfer
    - Engineering design

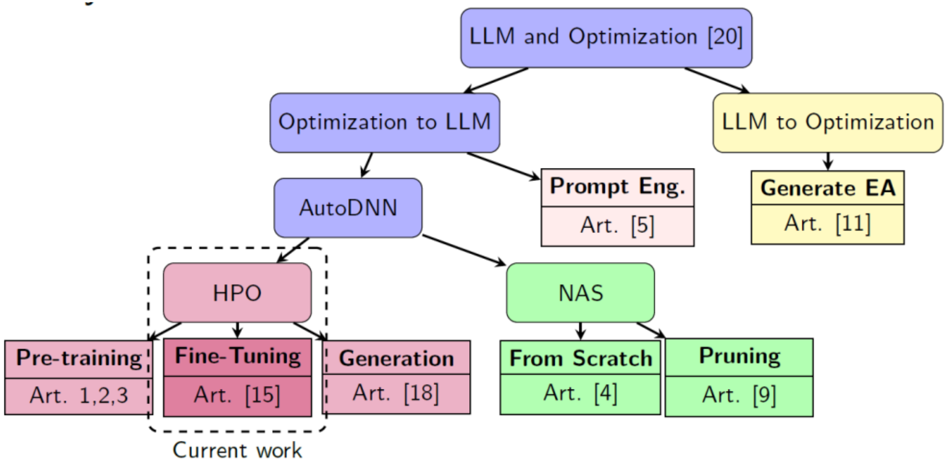
- Clear positioning at the interface of **Mathematical Analysis – SciML – High-Performance Computing.**

4. Exascale AutoML

- **Application of Exascale Parallel Bayesian Optimization**

- **LLMs**
  - Fine tuning
  - Pruning

- **Spiking Neural Networks**
  - Hyper-parameter optimization
  - Architecture Search



RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030

PROGRAMME
DE RECHERCHE
NUMÉRIQUE
POUR L'EXASCALE

Retrouvez toutes nos actualités

NumPEx