DAIMOS:
Distributed AI Model training Optimization at Scale

Exa-MA Annual Meeting 2026

January 21st, 2026

Axe 1.2 : HPC pour l'apprentissage haute performance et une IA très grande échelle

Consortium : IRIT-INP, Inria Bordeaux, Sorbonne Université, Météo-France

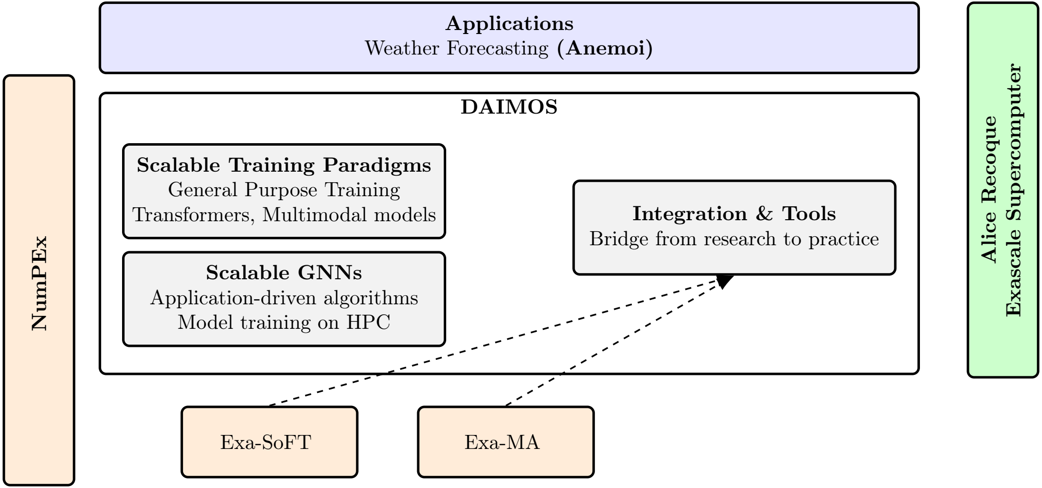
The Challenge

- **Explosion of training cost:** Modern AI models demand massive compute and memory; scaling is hard across heterogeneous, multi-GPU, multi-node systems.
- **Scientific aspect:** Training time, memory bottlenecks, and limited portability slow down innovation.
- **Where it hurts most:**
  - Parallelization techniques (data-, model-, tensor- parallelism) are understood individually but hard to co-optimize.
  - Graph and multi-modal models suffer from irregular structures, *communication* and *memory* bottlenecks.
- **Opportunity:** Harness Exascale HPC systems with smarter *algorithms* + *systems* co-design.
- **High-impact:** Weather forecasting is a flagship use case: frequent retraining on huge datasets; accuracy benefits from higher spatial resolutions.

Our vision

- **Goal:** Make distributed training of complex AI models **scalable, portable, and resource-efficient** on national HPC systems.
- **Main scientific objectives:**
  - Design efficient distributed training algorithms (including multilevel domain decomposition and alternative to back-propagation).
  - Tackle GNN & multi-modal training bottlenecks (memory, communication, scheduling).
  - Deliver a **modular, reusable software stack** for HPC environments.
  - Integration and evaluation of these contributions in a real-world application scenario, namely, GNN-based weather forecasting
- **Why it matters:** Beyond weather, benefits extend to bioinformatics, networks, climate modeling, and large scientific AI workloads.

DAIMOS – overview



Applications
Weather Forecasting **(Anemoi)**

DAIMOS

Scalable Training Paradigms
General Purpose Training
Transformers, Multimodal models

Scalable GNNs
Application-driven algorithms
Model training on HPC

Integration & Tools
Bridge from research to practice

NumPEx

Exa-SoFT

Exa-MA

Alice Recoque
Exascale Supercomputer

Scalable Training for Modern DL Architectures

Focus HPC for AI: advancing scalability of DL models

- Reduce memory, communication and synchronization bottlenecks
- Extend existing automated optimization tools to multimodal models
- Propose alternative to backpropagation

Contributors Sorbonne (E. Oyallon), Inria (J. Gusak)

Resources 1 PhD, 2 years postdoc

Graph Neural Networks (GNNs)

Focus Exploit specific structure of GNNs to improve scalability

- Load-balancing algorithms targeting the irregular structure of GNNs
- Multi-Level Domain Decomposition training algorithm for GNNs

Contributors Inria (L. Eyraud-Dubois), IRIT (A. Kopanicáková)

Resources 2 PhD

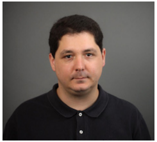
Integration and Reusable Tools

Focus Integration and evaluation of techniques

- New graph partitionner tailored for irregular GNN structures
- Adaptative communication
- Integration of algorithms in **Anemoi**
- Benchmark and demonstrators

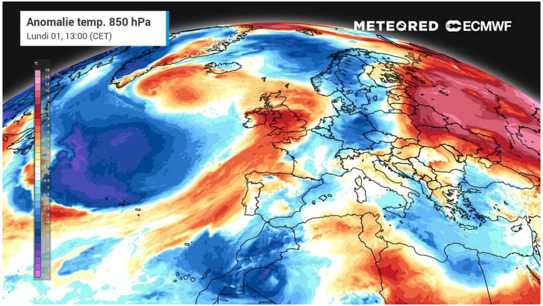
Contributors Meteo France (L. Raynaud), Inria (T. Herault), IRIT (J. Herrmann)

Resources 2 years postdoc, 1.5 year research engineer



DAIMOS : Distributed AI Model training Optimization at Scale

Julien Herrmann



Contact :
julien.herrmann@irit.fr