NumPEx

Numérique haute performance pour l'Exascale

NumPEx Kick-off meeting - 26-27-28 Juin 2023

Atelier thématique - IA

https://tinyurl.com/numpex-ia

Intelligence Artificielle

Contexte:

HPC for IA:
Optimisation de runtime:
- Modèles massifs basés sur des accélérateurs.
- Parallélisme pour les données, les modèles.

IA for HPC:
Besoin d'outils d'aide à la décision pour:
- Traiter les données les parties intéressantes,
- Piloter les simulations
- Accélérer les calcules

Problématique(s):

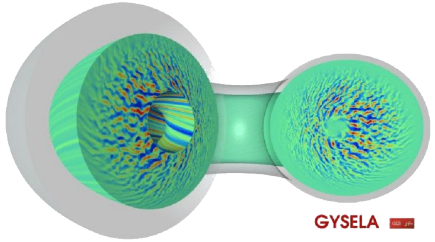
- Comment avoir une approche cross-domaine?
- Quels outils (librairies, opérations)?
- Quelles taches, quelles entrées, quels algorithmes?
- Quels workflows pour le pilotages?
- Comment gérer les données?

"Tout le monde fait de l'IA mais pas forcément d'expertise."

Texte

Code

Image



Complex data from physical science

SKAO

Enjeux:

Organisation

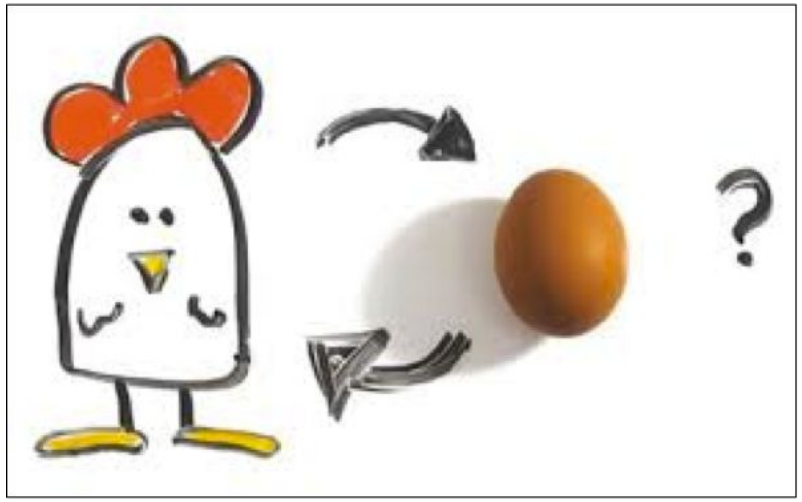
- Identifier les usages convergent du HPC et de l'IA
- Avoir une approche commune/fédérée
- Développer une stack permettant l'exploration (I/O des modèles, db, …)

Impact

- Modèles qui impactent la science (modèle fondateur, LLM): *demande grande ressources*.
- Comment utiliser les outils IA pour impacter NumPEx: *interoperabilité (jax/pytorch), aide au design (approche type copilot), …*

Identifier les outils adaptés

- Modèle de parallelisme pour l'IA: *parallelisme des données, des modèles, …*
- Approche pour accelerer les codes: *modèles réduits, surrogate model, …*
- Utiliser les GNNs pour les processings de graph/mesh



Plan d'action (quoi, quand/agenda, comment, qui, combien):

- Création d'un groupe de travail sur l'IA:
  - Identification de patterns spécifiques (HPC/IA),
  - Identification de technologies et de gaps,
  - Identification d'experts IA compatible HPC pour aider le pilotage.

- Organisation de workshops transverses sur l'IA:
  - HPC pour l'IA, avec applicatifs de l'IA (*e.g.* LLM/Bloom),
  - IA pour HPC avec experts de technologies IA,
  - Avec un temps dédié à la mise en commun de vocabulaire.

- Développement de briques de bases:
  - Modèle ML à grande échelle (PC1),
  - Liens HPC->IA sur briques existantes (PC2 *e.g.* starPU/joblib),
  - Liens IA->HPC sur briques existantes (PC3 *e.g.* compression w/ scikit-learn, model training w/pytorch on HPC outputs).