Refondation des laboratoires de la vallée d'Orsay: Thématique "Calcul scientifique, big data, informatique"

Effectif concerné: 27 ETP, 66 personnes, 1/3 ingénieurs; ne compte pas les personnels codant juste pour leurs analyses:
développement logiciel (ex: Geant 4) ou de techniques d'analyse de données (ex: Machine Learning)

Enjeux, axes principaux

Activité présente dans tous nos laboratoires, spectre très large:
– analyse d'images biomédicales
– traitement et visualisation d'imageur médical
– traitement et simulation des données d'expériences du LHC ou astrophysiques (LSST)
– calculs en QCD et prédictions des propriétés hadroniques
– étude des transitions de phase
– simulation de réacteurs et de scénarios nucléaires
– diffusion d'espèces chimiques dans le sol ou dans des matrices de stockage

Axes suivant lesquels nos laboratoires sont reconnus:
– traitement
– simulation
– analyse
– visualisation

Verrous communs à nos laboratoires:
– taille des données manipulées globalement (> 10 PO par an à LHC ou LSST)
– ressources de calcul limitées, algo performants à développer
– vectorisation et parallélisation des codes, problèmes parfois d'essence séquentiels, architecture informatique évoluant rapidement
– persistance: accès aux données sur une longue période, format d'écriture adapté et emploi du parallélisme
– visualisation de grands objets ou d'objets multidimensionnels
– pérennisation des codes écrits par les personnels (doctorants ou post-doctorants) amenés à quitter le système académique
– archivage et documentation des logiciels et formats de données pour demeurer utilisables sur une longue période

Positionnement scientifique, collaborations

Durée de vie des logiciels de plusieurs décennies pour des collaborations de plusieurs milliers de personnes, portabilité des codes; en biomédical, certification des logiciels, cahier des charges contraignant

Quelques collaborations françaises:
– Virtual Data (P2IO, CSNSM/IMNC/IAS/IPN/LAL/LLR/LPT/IRFU)
– GRIF: Virtual Data+LPNHE
– CDS 2 : Paris Saclay
– IN2P3 Master-Projet DecaLog; projets envisagés: conteneurs, langages dédiés et génération de code, précision et reproductibilité en contexte parallèle
– IN2P3 Master Projet MachineLearning en cours de définition
– SMURE (IPN, Subatech, LPSC): simulation de réacteur nucléaire

Réseaux d'experts: LoOPS (émanation francilienne de DevLog) et RI3 (IN2P3)

Rôle de premier plan dans des collaborations internationales:
– CERN : collaborations particulières avec la division IT ressource, SFT software
– HEP Software Foundation
– ALFA: ALICE et FAIR

Expériences internationales avec composante informatique forte à Orsay:
– ATLAS (LAL)
– LHCb (LAL)
– LSST (LAL)
– Planck (LPT, LAL)
– AGATA (CSNSM, IPN, Ganil)
– SVOM (LAL)

Projets logiciels de grande envergure, open source:
– Gate, Geant 4 pour le biomédical) (IMNC)
– Geant 4 (IPN, LAL)
– CLASS (IPN, Subatech, LPSC, Université du Wisconsin) : simulation de scenario électro-nucléaire

Objectifs à court, moyen et long terme

En lien avec les verrous technologiques:

- Assurer la performance des codes en utilisant toutes les possibilités des architectures informatiques, grande durée de vie des applications excluant les micro-optimisations: R&D en Domain Specific Language, génération automatique de codes, auto-optimisation; précision et reproductibilité des résultats,
- Entrées/sorties: goulets d'étranglement amoindrissant potentiellement les gains de performances
- Estimation des incertitudes: évaluer les différents types d'approximation et d'arrondis est crucial dans le cadre des approches "précision variable" pour augmenter la performance; techniques de calcul par intervalles ou de perturbation des arrondis pour tester la stabilité numérique pas assez développées dans nos laboratoires
- Nouveaux paradigmes d'analyse: de la grille, on passe à un standard du type "Spark" pour l'analyse distribuée des grandes masses de données; experts dans LSST (LAL)
- Machine Learning/Deep Learning: savoir-faire en plein essor, nous bénéficions de la structure Paris-Saclay "Center for Data Science" pour permettre à des experts d'horizons divers d'avancer concrètement, point fort dans les analyses LHC

Recrutement d'informaticiens répondant à ces objectifs, communs aux besoins de tous nos laboratoires, au-delà des spécificités des différentes collaborations

Organisation de la thématique

La structuration actuelle de nos laboratoires n'empêche pas les projets inter-labos de naître et d'avancer, mais ce sont souvent des projets à échelle nationale ou internationale. Difficulté pour les ingénieurs de s'émanciper car fortement impliqués dans les projets de laboratoires.

Quelle que soit la future structure, peu de conséquence dans l'immédiat sur le nombre de projets conduits ni sur les ressources dégagées.

Actuellement, dispersion des forces rendant insuffisante la capacíte de peser dans les décision au sein des collaborations internationales.

Formation et valorisation

Forte implication de nos laboratoires dans les formations en calcul scientifique dans les différents cursus universitaires locaux et l'organisation d'écoles thématiques spécifiques.

Besoin de développer des stages et des thèses en informatique appliquée, co-encadrés avec des laboratoires d'informatique de Paris-Saclay: formation par la recherche, expertise de la communauté de recherche en informatique.

Besoin de formation interne par échange d'expertise, recensement des savoirs-faire: écriture de tutos en ligne.

Développement des logiciels open-source, diffusion d'informations sur les licences open-source; spécificité de NDPITools (IMNC) qui a une double licence, open-source pour les milieux académiques et commerciale pour les autres (pas encore de vente à ce jour).