# *Refondation GT Calcul scientifique, big data, informatique*

## *1. Nature de la thématique: enjeux scientifiques, axes principaux, projets associés, effectif «publiant»...*

C'est une thématique qui a historiquement occupé une place importante dans nos laboratoires : pour ne citer que cet exemple compréhensible par tous, le déploiement des expériences de physique sur le site du CERN s'est accompagné d'un vaste effort de mise au point des outils numériques nécessaires à l'analyse des données expérimentales, effort auquel notre communauté n'est pas restée étrangère.

Au delà de ce coeur thématique et historique, cette thématique est aujourd'hui présente dans tous nos laboratoires et rattachée à un spectre très large d'activités scientifiques. Nous développons en effet des codes pour :
- analyse d'images biomédicales
- traitement et visualisation d'imageur médical
- traitement et simulation des données d'expériences du LHC ou astrophysiques (LSST),

le tout représentant un volume de données avoisinant la dizaine de PO par an
- 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

Nous pouvons relever quelques axes suivant lesquels nos laboratoires peuvent mettre en avant leur expertise :
- traitement
- simulation
- analyse
- visualisation

Nous remarquons que les différentes équipes de recherche ont pour beaucoup à faire face à des verrous similaires pour l'aspect numérique de leurs travaux :
- taille des données : dizaines de PO par an pour les grandes expériences comme LHC ou LSST dans leur globalité. Par ailleurs, il y a aussi un seuil à une dizaine de GO où les données ne peuvent plus être chargées en mémoire
- ressources de calcul limitées : développements algorithmiques indispensables pour atténuer au maximum cette contrainte extérieure
- vectorisation et parallélisation des codes : un gisement important de performance , mais difficile à utiliser avec des problèmes parfois séquentiels "par nature, et d'autant plus que

les architectures informatiques évoluent beaucoup plus vite que la durée de vie de moyenne des codes (au moins un facteur 3).

- persistance : lié au besoin de traiter des grandes masses de données, la performance de l'accès aux données en lecture ou écriture est devenue critique. C'est un problème qui intègre à la fois la question du format des données et des techniques permettant d'écrire un fichier en parallèle sans (ou avec peu de) sérialisation.
- visualisation : représentation de grands objets ou d'objets multidimensionnels (d>3)

Nous estimons que ces équipes vont également rencontrer des difficultés communes :

- pérennisation des codes écrits par les (post) doctorants : d'ailleurs nous voyons là un enjeu de pratique éditoriale, la reproductibilité des articles, et un enjeu de formation des doctorants pour ceux qui quittent la discipline.
- archivage à long terme : autant ce n'est pas un problème technique pour conserver les données mais il s'agit plutôt de l'archivage et de la documentation des logiciels et des formats de données pour qu'ils restent utilisables dans le futur.

Enfin nous relevons un enjeu pour lequel nous n'avons pas toutes les cartes en main, puisque cela dépendra en grande partie de décisions prises au niveau des EPST :

- nécessité à terme de rendre publics les données scientifiques et, peut-être, les codes ; nous nous interrogeons cependant sur l'intérêt pour le grand public d'une telle mesure concernant nos champs de recherche

Nous évaluons à une soixantaine de chercheurs et ITA publiant le résultat de travaux ou des logiciels ayant fait appel à l'expertise mise en avant dans cette thématique.

## *2. Contexte: Positionnement scientifique actuel local/national/international : spécificités/originalités/impact/limites/absences... dans les activités/collaborations /projets. Collaborations existantes entre les cinq labos.*

Nos domaines de recherche ont naturellement conduit les équipes de nos laboratoires à se structurer pour mener à bien les recherches. Le travail est conduit dans des collaborations à plusieurs échelles: bassin d'enseignement supérieur de l'Ile-de-France Sud, Ile-de-France, et aux échelles nationale et internationale. La durée de vie d'un logiciel est de plusieurs décennies pour les collaborations rassemblant plusieurs milliers de scientifiques. Cela explique le souci de la portabilité des codes, sachant que la nature des architectures informatiques sur lesquelles ils s'exécutent peut connaître une évolution radicale, comme par exemple les GPU. Dans le domaine du biomédical, les produits informatiques doivent avoir une validation par les autorités compétentes, ce qui impose un cahier des charges très bien cadré.

Les collaborations françaises les plus emblématiques pour le périmètre de la thématique sont :

- 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
- CEPS, sous l'égide de la Maison de la Simulation, projet non financé, mais créateur d'une dynamique par le groupe de travail Paris-Saclay "Simulation/Modèlisation" : Paris Saclay
- SMURE (IPN, Subatech, LPSC) : simulation de réacteur nucléaire
- collaboration embryonnaire avec le LRI autour des murs d'image

Dans plusieurs de ces collaborations, en particulier dans les projets IN2P3, nos laboratoires jouent un rôle moteur et collaborent entre eux sur des activités transverses aux expériences. Nos laboratoires s'insèrent dans des réseau d'experts, citons en particulier LoOPS, l'instance francilienne de DevLog, qui regroupe tous les établissements d'enseignement supérieur, et RI3, qui regroupe les informaticiens de l'IN2P3 et de l'Irfu.

Nos équipes ont un rôle de premier plan dans les collaborations internationales :

- CERN : collaborations particulières avec la division IT ressource, SFT software
- HEP Software Foundation
- ALFA : framework pour ALICE et FAIR

En outre de nombreuses expériences internationales ont une composante informatique forte dans la vallée d'Orsay :

- ATLAS (LAL)
- LHCb (LAL)
- LSST (LAL)
- Planck (LPT, LAL)
- AGATA (CSNSM, IPN, Ganil)
- SVOM (LAL)

Nous participons à des projets logiciels de grande envergure, tous open source :

- Gate, specialization de Geant4 pour le biomédical) (IMNC)
- Geant4 (IPN, LAL)
- CLASS (IPN, Subatech, LPSC, Université du Wisconsin) : simulation de scenario électro-nucléaire

Les équipes des différents laboratoires ont accès à d'importantes ressources mutualisées de calcul scientifique, citons notamment

- GENCI : IDRIS (plateau de Saclay), TGCC (CEA, nœud français de PRACE), CINES Montpellier)
- CC-IN2P3 (Lyon)

## *3. Objectifs: quelle ambition et quel impact dans la thématique à court (2-3 ans), moyen et long terme (~15 ans). Axes nouveaux possibles/repositionnement. Décliner le positionnement « idéal » sur les axes principaux, les projets associés. Identifier les profils des compétences scientifiques et techniques à associer.*

Au-delà des domaines d'applications différents couverts par nos laboratoires (physique nucléaire, physique des hautes énergies, astro-particules, cosmologie, théorie, santé, énergie nucléaire...) et des différentes catégories d'applications (simulation, reconstruction, analyse, visualisation…), les principaux défis sont liés aux verrous technologiques énoncés précédemment.

- Performance en utilisant toutes les possibilités des architectures modernes (parallélisation, vectorisation…) tout en prenant en compte la durée de vie de nos applications qui exclut dans la plupart des cas la possibilité d'une micro-optimisation pour une architecture particulière. Pour cela, il faut combiner plusieurs approches, encore au stade de la R&D pour certaines d'entre elle, telle que les Domain Specific Language (DSL), la génération automatique de code, l'auto-optimisation pour lesquelles une collaboration avec la recherche en informatique, et plus largement la communauté HPC, est nécessaire. Lié à cette problématique de la performance durable, se pose la question de la gestion de la précision et de la reproductibilité des résultats, un autre problème important pour la validation des applications.
- Entrées/sorties : fortement lié à la problématique précédente, les entrées/sorties deviennent des goulets d'étranglement qui annulent les gains de performance obtenus par la parallélisation. Des solutions commencent à être implémentées dans certains frameworks mais des solutions plus génériques sont souhaitables.
- Estimation des incertitudes : il s'agit d'une problématique au coeur de nos problématiques scientifiques dont l'analyse in fine suppose une bonne estimation des systématiques. Les différents types d'approximation et d'arrondis à l'oeuvre dans une application informatique sont une source d'incertitude dont la maîtrise peut s'avérer critique dans certains cas. Cette problématique se voit renforcer par les approches de type "précision variable" pour augmenter la performance des applications. Plusieurs techniques (e.g. calcul par intervalles, perturbation d'arrondi pour la stabilité numérique) se développent qui sont encore assez peu maîtrisés dans nos laboratoires.
- Nouveaux paradigmes d'analyse : la physique des hautes énergies aujourd'hui mais les autres disciplines très prochaines reposent fortement sur des ressources de computing distribuées pour l'analyse des grandes masses de données. Initialement bâtie sur des solutions spécifiques (telle la grille), elles doivent être revisitées pour s'appuyer sur les solutions qui ont émergées ces dernières années dans le monde du Big Data, tel Spark devenu un standard pour l'implémentation du paradigme

MapReduce qui est au coeur de l'analyse distribuée. Au-delà d'une nouvelle plateforme, c'est un reengineering important des applications qui est nécessaire, basé sur une approche plus "fonctionnelle". Des compétences commencent à être développées dans nos communautés, en particulier au LAL dans le cadre de l'expérience LSST.

- Machine Learning/Deep Learning : ces techniques d'analyses sont en pleine expansion et démontre une capacité à jouer un rôle important dans un nombre croissant de nos applications, loin des "terrains de jeu" d'origine du ML. Nos laboratoires sont dans un contexte exceptionnel avec Paris Saclay du fait du Center for Data Science (coordinateur : B. Kegl, LAL) qui regroupe plusieurs centaines de personnes d'horizons scientifiques différents. Nos laboratoires ont donc le potentiel d'être un lien entre les experts du ML/DL et nos communautés, à l'image du rôle que joue le LAL pour les expériences LHC.

L'ensemble des développements logiciels de nos laboratoires ne se résument pas à ces problématiques et il est important de maintenir le potentiel de développement de nos laboratoires pour les différentes expériences et projets dans lesquels ils sont impliqués. Dans le même temps, il semble important aussi que les futurs recrutements pour les besoins en développement informatique contribuent à renforcer le potentiel de nos laboratoires sur ces différents sujets (qui ont des liens entre eux comme mentionnés plus haut). En effet, les collaborations responsables des différents projets ont besoin que certains laboratoires puissent fournir des experts sur ces sujets et l'importance numérique du potentiel présents dans l'ensemble de nos laboratoires nous crée une forme d'obligation d'être l'un d'eux. Une difficulté potentielle liée à ces recrutements est que dans de nombreux cas, ils exigent une personne de niveau IR et plus rarement IE.

Une autre caractéristique de l'informatique des projets dans lesquels nous sommes impliqués est l'importance des contributions fournies par les doctorants et postdocs (dont la formation initiale n'était souvent pas centrée autour de l'informatique et du développement). Afin de pouvoir capitaliser sur ces contributions et d'assurer leur pérennité, il est nécessaire d'organiser la formation des nouveaux venus aux bases du génie logiciel et des méthodes utilisés dans nos collaborations.

## *4. Organisation de la thématique: Déploiement et organisation pour atteindre les objectifs, dans le périmètre des cinq laboratoires et liens externes (à la fois en local, mais aussi avec les autres grands centres pour cette thématique).*

La structure actuelle des laboratoires n'a pas empêché des collaborations entre nos laboratoires, mêm si les personnes participent principalement aux projets de leurs laboratoires. Il y a actuellement peu de collaboration entre laboratoires dans le cadre des projets. Le contexte des collaborations dans lesquelles nous sommes impliquées est souvent national ou international. En même temps, comme indiqué ci-dessus, les

problématiques auxquelles nous sommes confrontés dans les différents projets sont largement commune et la capacité à mieux partager l'expertise dispersée dans nos laboratoires (souvent une personne par labo) est parfois rendu difficile, surtout pour les IT, par la priorité à donner aux projets du laboratoire.

L'ensemble des personnes impliquées dans ces activités sont engagés sur des projets. Quelque soit les évolutions de structure qui pourraient être décidées, le nombre de projets à soutenir resteraient identique en première approximation et cela ne dégagerait vraisemblablement pas de ressources à court terme, même si cela peut avoir un impact sur la collaboration entre les personnes.

Une des limites de la structuration actuelle est la difficulté, du fait de la fragmentation des forces, d'avoir un poids dans les collaborations internationales, malgré le nombre de personnes impliquées dans nos laboratoires. Une éventuelle évolution de structure devrait permettre de renforcer cette capacité à peser sur les choix de nos projets.

## *5. Formation et valorisation : prospectives liées à l'enseignement et projection du thème dans le domaine de la valorisation*

L'ensemble de nos laboratoires sont fortement impliquées dans la formation autour du calcul scientifique dans les différents cursus universitaires locaux ainsi que par l'organisation d'écoles thématiques spécifiques. Les principales implications actuelles sont:

- LAL: M2 NPAC génie logiciel/collaboration , analyse/reconstruction
- IPN: M1 Nuclear Energy
- IPN Ecole Geant4 : Ecole doctorale PHENIICS, IN2P3, CNRS Entreprises
- IMNC L3 traitement d'image,
- LAL Ecole IN2P3 Parallèlisme
- LAL Licence Pro API : base de données
- LAL Formation permanente DR5 python/C++
- LPT L3/M1 physique langage C
- IPN à completer
- IMNC: M2 «Systèmes biologiques et concepts physiques» , cours «Projets de
- physique numérique»
- IMNC: L2 physique : cours «Algorithmique et langage de programmation» et «Simulations numériques»
- LAL : M2 Machine Learning (à préciser)

Il conviendrait de développer les stages et même les thèses d'informatique appliquée, éventuellement co-encadrées avec des laboratoires d'informatique de Paris-Saclay : c'est à la fois une façon de contribuer à la formation par la recherche et de bénéficier des expertises de la communauté recherche en informatique.

En plus de la formation d'étudiants ou des écoles, il y a un besoin de formation interne par échange d'expertise, au fil de l'eau, qui nécessiterait un recensement des expertises locales. A cette fin, une approche possible est le développement de tutos en ligne type Jupyter Notebook (qu'il s'agisse de développement logiciel, machine learning, ou traitement statistique avancé), pouvant être utilisés et réutilisés aussi bien dans un cadre individuel que dans une session de formation/hackathon, et pouvant favoriser l'échange d'expertise (rayonnement possible à l'IN2P3 voire dans les collaborations).

Concernant la valorisation, la problématique autour du logiciel est assez spécifique dans la mesure où la quasi-totalité des développements de nos laboratoires sont fait en open-source et ne font pas l'objet d'une valorisation financière de type prestations de service autour des logiciels développés. Toutefois, il est sans doute important de diffuser l'information sur les problématiques de licence open-source (qui ne sont pas une absence de licence). Une exception connue dans nos développements est NDPITools (IMNC) qui a une double licence open-source pour la communauté académique et commerciale pour les autres (mais pas de licence commerciale vendue à ce jour).

## *6. Eléments statistiques. Identification RH et indicateurs factuels de l'évolution des forces /moyens (en interaction avec le pole RH et finances).*

L'effort concerné par ce groupe thématique a été recensé dans les différents laboratoires. Il comprend des ingénieurs et des chercheurs qui ne sont pas forcément affecté sur un projet clairement identifié comme contribuant au calcul scientifique.

Les conditions aux limites ont été définies ainsi: il s'agit du développement logiciel ou de technique d'analyse de données (e.g. Machine Learning)

- pour une équipe ou pour une collaboration
- pour l'exploitation d'un instrument (mais pas du logiciel DAQ/en-ligne)
- un logiciel open source public (avec des utilisateurs, pas juste un repository github)
- développement d'un logiciel "personnel" suffisamment complexe avec des techniques de programmation avancées

En clair, un chercheur qui écrit un code aussi long soit-il, tournant sur une grille de calcul pour remplir des histogrammes et les ajuster ne serait pas concerné. N'est pas concerné celui faisant tourner un code de simulation développé par d'autres en changeant quelques paramètres. La gestion d'infrastructures n'est pas non plus concernée.

Les non-permanents et doctorants ont été comptés.

| Activité | ITA (ETP/personnes) | Chercheurs (ETP/personnes) | Grand Total |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Visualisation | IMNC 0.1/1 LAL 0.5/1 | IMNC 0.1/1 LPT 0, LAL 0, IPN 0.2/2 | 0.9/5 |
| Code d'Analyse | IMNC 0.5/1 LAL 2./4 | IMNC 0.2/1 LPT 0.35/3, LAL 6.8/14, IPN 0.2/2 | 10.05/25 |
| Optimisation de code/Vectorisation | IMNC 0.1/1 LAL 2./3. | IMNC 0.05/1 LPT 1.15/7, LAL 0.2/1, IPN 0.3/1 | 3.8/14 |
| Modélisation/simulation | IMNC 0.5/1 LAL 1/1 | IMNC 0.1/1 LPT 1.15/7, LAL 0.5/2, IPN 6.5/13 | 9.75/25 |
| Base de données | IMNC 0. LAL 2.6/4 | IPN 0 | 2.6/4 |
| Totaux | IMNC 1.2/4, LAL 8.1/13 = 9.3/17 | IMNC 0.45/4, IPN 7.2/18 LAL 7.5/17 LPT 2.65/10 = 17.8/49 | 27.1/66 |

En ce qui concerne l'évolution prévisible des effectifs à 5 ans :

- LPT -0.5ETP modèlisation.
- LAL : pas significative à 5 ans, mais oui à 10 ans
- IMNC pas d'évolution
- IPNO : +1ETP dans les 5 ans
- CSNSM