Institut national de physique nucléaire
et de physique des particules

Calcul et Données : nouvelles de l'IN2P3

Journées Informatiques 2020

19 novembre 2020

Sabine Crépé-Renaudin, DAS Calcul et Données à l'IN2P3

Introduction

Quelques informations générales

- Conséquences de la pandémie
- Prospectives
- Recherche en informatique
- Science ouverte

Les projets IN2P3

- Les projets en cours (hors projets européens, Cf présentation Volker)
- Nouveau projet

Calcul et Données en 2020

Un grand merci

- pour votre mobilisation pour que le télétravail se fasse dans les meilleures conditions possibles
  - vos efforts ont été appréciés par tous les personnels de l'IN2P3
- pour avoir maintenu en même temps le rythme et les performances de la plupart des activités informatiques

Impact de la pandémie sur les activités calcul et données

- Surcroit de travail pour mettre en place les outils pour le télétravail
- Impact très différents selon les activités
  - activités informatiques souvent adaptées au travail à distance
  - accès aux infrastructures a été finalement possible pour les urgences pendant les périodes de confinement, petit report des installations
  - activités collaboratives beaucoup plus impactées en particulier si discussions, développements en équipe nécessaires
    - report/annulation des réunions/ateliers/conférences, transformation en réunions en ligne
    - Impact plus important pour les nouveaux projets ou les nouveaux arrivants

Activités calcul et données pour la lutte contre le COVID

Contributions à Folding@home via la grille de calcul LHC

- Folding@home = simulation de la dynamique des protéines du virus qui permet de définir de nouvelles cibles pour des médicaments
- 5 à 10% des ressources des expériences, contribution WLCG au 30 septembre 2020 : 16 MWU
- Les sites français ont tous contribué à hauteur de leur poids ds WLCG
- Publication bioRxiv : Citizen Scientists Create an Exascale Computer to Combat COVID-19

Contribution DIRAC

- Accès aux sites OSG, US
- Organisation d'un canal prioritaires pour les jobs COVID
- Développement des étiquettes pour la comptabilité précises des contributions des sites à la recherche COVID

Portail d'imagerie virtuelle VIP : workflows spécifiques ont été mis en place (criblage virtuel basé sur AutoDock, labellisé EOSC-Pillar)

Contributions locales des labos via leurs contacts (IPHC, CPPM, CC…)

➜ Structuration du calcul scientifique via les organisations telles que EGI et, pour HEP, WLCG ainsi que la bonne connectivité des sites ont été une force pour nos discipline
➜ Importance du réseau humain
➜ Très forte réactivité de notre communauté

Prospectives nationales

Résumé des étapes précédentes

- Site web : https://prospectives2020.in2p3.fr/
- Constitution de 13 Groupes Thématiques (GT09 = Calcul, algorithmes et données)
- Collecte des contributions écrites de l'ensemble de la communauté
- Organisation de séminaires thématiques en région ;
- Rédaction d'un document de synthèse pour chaque GT (rapport du GT09)

Colloque de restitution à Giens

- prévu en octobre 2020 reporté au printemps 2021 à cause de la crise sanitaire
  - du mardi 29 mars au vendredi 2 avril 2021 à Giens https://indico.in2p3.fr/event/22028/overview
- période mise à profit pour travailler les aspects transverses
  - prochainement réunions entre groupes de travail

Prospectives Emplois & Compétences Techniques IN2p3 (PECTIN)

Rodolphe Clédassou
Rémi Cornat

Contexte

- diminution des effectifs IT (-1% / an en moyenne sur les 20 dernières années)

Enjeu majeur pour l'institut

- pérennité des compétences,
- adéquation des ressources et savoir-faire aux besoins scientifiques
- C'est aussi une opportunité pour évoluer en fonction des besoins et un facteur d'adaptation continu de l'institut et de ses unités au contexte scientifique

Objectif :

- Identifier nos forces et faiblesses techniques en vue d'anticiper une adaptation aux choix de l'institut

PECTIN

Une action dans la suite des prospectives scientifiques :

- Analyse des besoins des projets futurs identifiés par les prospectives scientifiques
- Mise en regard avec les dynamiques des technologies, des environnements, des populations, des métiers/spécialités
- Établissement d'un Référentiel « métiers / spécialités »
- Identification des partenariats, viviers, formations, permettant de répondre aux besoins
- Mise en place d'une projection des besoins

PECTIN : composition du groupe de travail

Nom	Qualité	Unité
Rodolphe Clédassou	DAT	IN2P3
Rémi Cornat	CdM / DT	LPNHE
Florence Ardellier-Desages	DT	APC
Thierry Ollivier	CdM	IP2I
Valérie Givaudan	Réseau RI3	IJCLab
Renaud Le Gac	PHY	CPPM
Bernard Genolini	Adj. DT	IJCLab
Martine Verdenelli	RA	IP2I
Laurent Gross	DT	IPHC
David Longuevergne	PHY	IJCLab
Magali Damoiseaux	BAP-F	CPPM
Philippe Laborie	CdM / RT	LPC Caen
Sandrine Pavy	CdM	LLR
Cyrille Thieffry	CdM & Responsable cellule SNR	IN2P3
Cyrille Berthe	Chef de Division Adj. Ops	GANIL

PECTIN Calendrier

Nature	Contenu	Echéance	Modalités
Définitions et plan de travail consolidés		**Oct-2020**	Travail interne du GT, communication vers unités.
Liste consolidée des spécialités	Cette liste fera l'inventaire des spécialités « classiques » et identifiera en plus les spécialités « rares ».	**Déc-2020**	Enquêtes terrain pilotées par GT avec accès aux CdS, réseaux…
Recensement des besoins nouveaux	Croisement avec les prospectives scientifiques, prise en compte feuilles de routes technologiques	**Fév-2021**	Enquêtes terrain pilotées par GT : accès aux CdS et CdP, aux réseaux, …
Document d'analyse	Besoins et spécialités associées, statistiques des populations par spécialité, projection à 5-10 ans des besoins : conséquences, nouvelles spécialités, évolutions technologiques, plans de recrutement, impact formations, …	**Juin-2021**	Consultation des laboratoires (notamment CdS et RT/DT) pour avoir les statistiques. Communautés externes
Référentiel des réalisations **Référentiel des spécialités**	Inventaire des réalisations techniques. Profil type de spécialité	**Sep-2021**	Enquêtes terrain pilotées par GT avec accès aux CdS, réseaux…
Documents de synthèse	Recommandations par spécialité et recommandations générales	**Nov-2021**	Travail interne du GT
Rapport Final	Consolidation dans un document de synthèse et bases de données (création de la base CC+ via un double stage RH+INFO ou une prestation à partir du printemps).	**Déc-2021**	Stages, prestations, CC-IN2P3, …

Les défis pour les prochaines années

Des demandes en informatique en forte croissance pour les prochaines années

- LHC Run3 + HL-LHC, Belle II, KM3NET, T2K…
- Vera-Rubin/LSST, Euclid…
- Utilisation croissante de l'intelligence artificielle

➜ Impact sur les ressources en calcul, stockage, réseau…

➜ Des nouvelles solutions à trouver

Une grande diversification des types ressources

- HTC, HPC, GPU, FPGA, …

Importance de développer de nouvelles solutions

- Proposer des projets R&T => peu de projets en cours
- Recherche en informatique

Recherche en informatique

Renforcement de la recherche en informatique à l'IN2P3 pour mieux préparer les défis à venir

- Équipe recherche au CC (Frédéric Suter)
  - Simulation d'applications et de systèmes distribués (https://simgrid.org) et de workflows scientifiques (https://wrench-project.org)
  - Ordonnancement batch et workflow
  - Évaluation de performances (HPC, workflows, ...)
  - Engagement des utilisateurs (càd comment collaborer de façon efficace avec d'autres communautés scientifiques pour leur apporter des solutions utilisables, utiles, et utilisées et pas uniquement des preuves de concepts ou des prototypes)

- Un poste CR de la section 06 en 2020 pour le CC-IN2P3
  - Affiché « Grandes masses de données et calcul haute performance en physique des hautes énergies »
  - Bertrand Simon a rejoint l'équipe de Frédéric Suter, il travaille sur l'élaboration et l'étude d'algorithmes principalement en ordonnancement mais aussi en structures de données

- Il devrait y avoir un poste CR de la section 06 au concours CNRS 2021
  - au LAPP dans le groupe LSST
  - « Recherche sur le traitement des très grandes masses de données et le calcul haute performance en cosmologie observationnelle »
  - Infos profil détaillé : Dominique Boutigny et Frédéric Suter

Sciences ouvertes

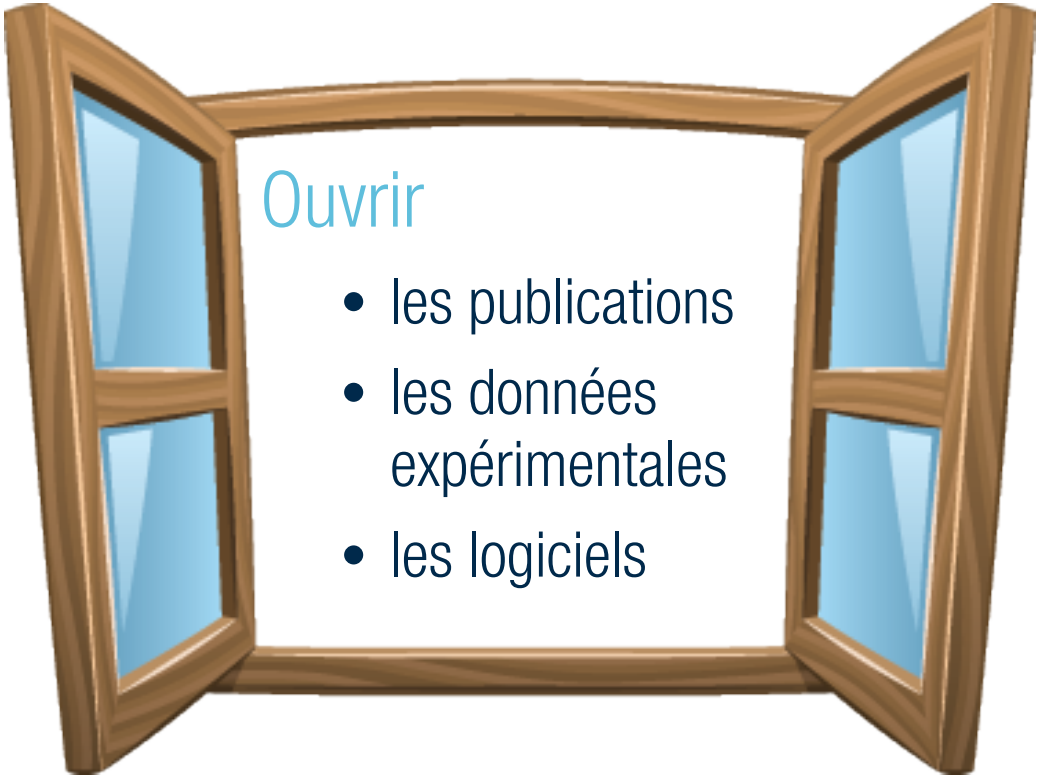
Contexte

- Plan national pour la science ouverte (juillet 2018)
- Feuille de route du CNRS pour la Science Ouverte (Novembre 2019)
- Contexte international: EOSC, RDA, GO FAIR, ...

Conséquences récentes

- Création de la DDOR : Direction des Données Ouvertes de la Recherche (novembre 2020)
  - Fusion de la DIST et de la mission MICADO sur les aspects calcul et données
- Officialisation du Plan Données de la Recherche du CNRS 2020
- 2ème journée Science Ouverte au CNRS

Sciences ouvertes



Tour d'horizon

Publications en accès libre

- ~82% des publications en accès libre
- Accords SCOAP3 prolongés pour 2 ans
- Moissonnage automatisé via Inspire-HEP et HAL
- HEPdata : données accompagnant les publis pour ré-interprétation résultats

Logiciels libres

- facilité par des outils comme Gitlab
- en nette augmentation
  - ex : Athena expérience ATLAS au LHC, NPTool en physique nucléaire
- Plan de gestion logiciel : projet PRESOFT développé au CC-IN2P3 et France-Grille disponible via DMP OPIDoR

Données ouvertes

- Astroparticule et cosmologie
  - tradition longue d'ouverture des données traitées après une période d'embargo
- Physique des particules
  - ouverture partielle, politique d'ouverture en cours de discussion
  - difficultés liées à la quantité et complexité des données
  - CERN open data portal
- Physique nucléaire, physique des accélérateurs, interdisciplinaire
  - plan d'ouverture en cours
  - données parfois confidentielles/sensibles

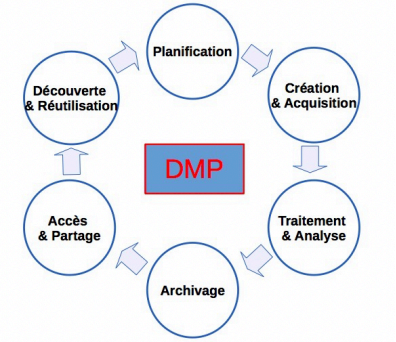
Sciences ouvertes pour les données à l'IN2P3

Sauvegarde des données

- Copie des données au CC-IN2P3 pour les expériences avec uniquement un stockage local

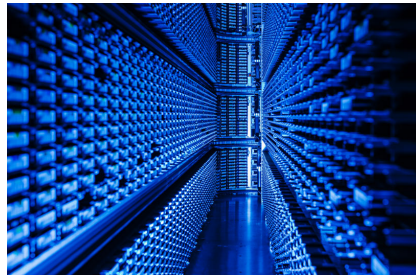
Définition d'un plan de gestion des données (DMP = Data Management Plan)

- disponible sur DMPOpidor/INIST et RDMO
- 123 questions organisées en 7 sections et 42 sous-sections
- 22 DMP remplis (~10%)
- Plan d'adoption en cours de discussion (forte recommandation pour les nouvelles expériences et les expériences en cours)



Étude de faisabilité de la mise en place d'un service d'archivage

- Archivage = stocker les données, les référencer, les préserver dans le temps et être capable de les relire
- Volumétrie estimée : ~10% des données du CC-IN2P3 sont à archiver (~ 10 PO)
- Collaboration sur les processus d'archivage avec les spécialistes IST à l'IN2P3 et hors IN2P3 (CINES, BNF)
- Stratégie de préservation via l'émulation
  - encapsulation des données et des logiciels + virtualisation
  - permet de conserver/reproduire l'environnement fonctionnel d'origine pour pouvoir continuer à l'exécuter à long terme
- Faisabilité technique de la mise en place d'un service d'archivage OAIS au CCIN2P3 vérifiée
- Solution trouvée pour la gestion de paquets AIP de grande taille (TiB ou PiB) avec la segmentation proposée par la norme CSIP
- Plan d'action proposé pour la mise en ouvre du service archivage OAIS à l'IN2P3



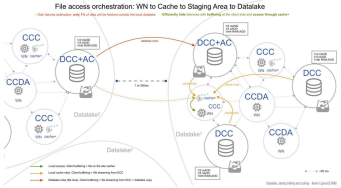
Les projets IN2P3* calcul et données

Autour des infrastructures

LCG-France

France-Grille

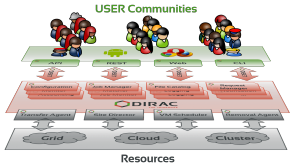
DOMA-FR



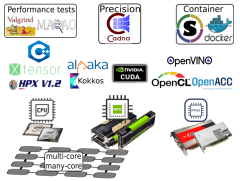
Autour du calcul et des logiciels

Geant4

DIRAC

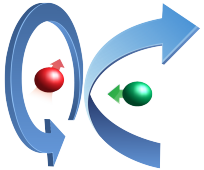


Decalog



Machine Learning

* hors projets européens



Nouveau projet : QC@IN2P3

En cours de validation

Objectif principal

- préparer la transition vers les technologies de calculs scientifiques basées sur les processeurs quantiques

Activités

- Apprendre à utiliser les algorithmes quantiques sur les plateformes actuelles
- Faire une veille technologique sur les ordinateurs quantiques
- Identifier des applications pilotes dans la physique IN2P3 et les simuler sur les plateformes actuelles.
- Utiliser ces cas comme jalons pour les ordinateurs quantiques et contribuer aux avancés dans le domaine de l'informatique quantique

Équipes

- Laboratoires impliqués : IJCLab, LLR, LPC
- Responsables scientifique et technique : Denis Lacroix (IJCLab) et Andrea Sartirana (LLR)
- ~2,5 ETP

Conclusion

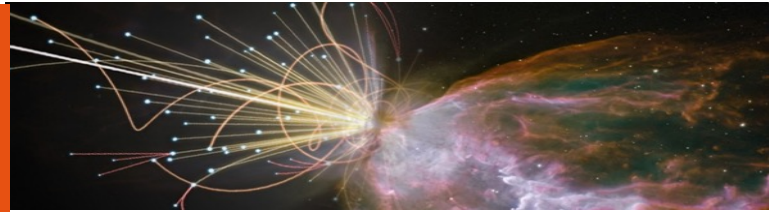
Activités Calcul et Données 2020

- Niveau d'activité très bon malgré la pandémie
  - Services et projets
- Équipes exemplaires pour faire face aux confinements
- Impact de la pandémie variable en fonction des activités, négatif sur les activités collaboratives

Préparer l'avenir

- Sciences ouvertes
- Actions de prospectives
- Défis : besoin nouvelles expériences, diversification des ressources
- Nouvelles solutions à inventer ➜ R&D, R&T

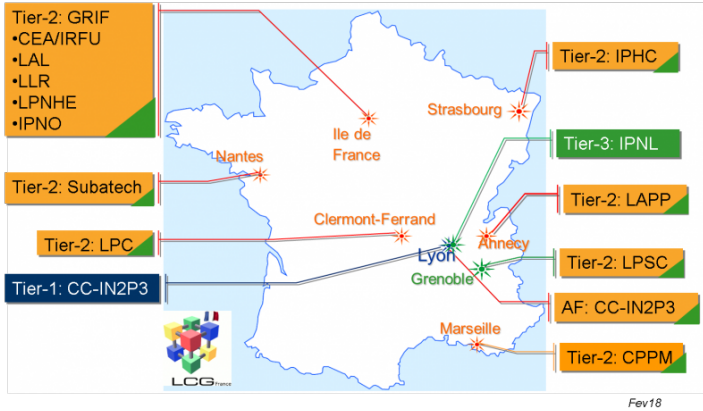
Pour en savoir plus sur les projets



LHC Computing Grid France

Organisation du calcul pour les expériences du LHC en France

- Très long terme : MoU signé en 2006 et a priori s'étend jusqu'à fin du LHC cad 2040.
- Objectif : fournir environ 10% des ressources de calcul mondiales
  - Tier 1 : gros centre de calcul avec stockage de masse : CC-IN2P3.
  - Tier 2 : CC-IN2P3, CPPM, IJCLAB, IPHC, IRFU, LAPP, LLR, LPC, LPNHE, LPSC, SUBATECH.
  - Tier 3 : IP2I.
- Ces ressources sont intégrées à l'infrastructure de la grille de calcul mondiale WLCG.



Équipes

- Responsables : Laurent Duflot et David Bouvet
- ~19 ETP

France Grilles

IN2P3

Calcul et données/Grille de calcul

- **Responsable scientifique :** Vincent Breton (LPC)
- **Laboratoires participants :** APC, CC-IN2P3, CPPM, IJCLab, IP2I, IPHC, LAPP, LPC, LPNHE, LLR, LPSC, LUPM, SUBATECH
- **Statut :** Infrastructure de recherche et GIS
- Projet financement principal par la France
- **Site web :** http://www.france-grilles.fr/

17 sites d'infrastructure en France — **29** laboratoires en France

1000 utilisateurs dynamiques — **100K** cœurs de calcul

50 péta-octets de ressources de stockage — **200** K€ de financement annuel

OBJECTIFS SCIENTIFIQUES :

France Grilles a pour objectif de fournir des services humains et numériques pour répondre aux besoins de traitement et de stockage des données scientifiques massives. Ces services numériques reposent sur une infrastructure informatique distribuée accessible par le biais de plusieurs technologies (grille de calcul, Cloud Computing, etc), et sont ouverts à l'ensemble des communautés scientifiques. France Grilles représente la France dans le Cloud européen fédéré EGI (www.egi.eu) et contribue avec les autres états membres à son fonctionnement.

CONTRIBUTIONS FRANÇAISES :

Les équipes France Grilles développent et administrent DIRAC, un outil IN2P3 utilisé à l'international pour la gestion des calculs distribués. Leur participation au est importante au niveau d'EGI, via le support aux utilisateurs, matériel et développement logiciel, ainsi qu'au travers de ses 12 sites qui contribuent à hauteur de 11 % aux ressources de type grille (HTC) de EGI. Elles forment des utilisateurs et administrateurs aux outils de gestion de données et de logiciels (IRODS, PRESOFT), et coordonnent des journées "Calcul et données".

MOYENS DÉPLOYÉS :

- Mise à disposition d'une infrastructure de niveau production (disponibilité > 99%) pour le traitement intensif des données
- Aide et accompagnement des administrateurs et des utilisateurs des services par le biais de formations et d'une documentation en ligne
- Développement d'outils pour faciliter l'utilisation optimale des services par les chercheurs (portail scientifique VIP, certificat robot, ...)
- Portail d'accès unique à l'ensemble des services

- **Responsable scientifique :** Vincent Breton (LPC)
- **Laboratoires participants :** APC, CC-IN2P3, CPPM, IJCLab, IP2I, IPHC, LAPP, LPC, LPNHE, LLR, LPSC, LUPM, SUBATECH
- **Statut :** Infrastructure de recherche et GIS
- Projet financement principal par la France
- **Site web :** http://www.france-grilles.fr/

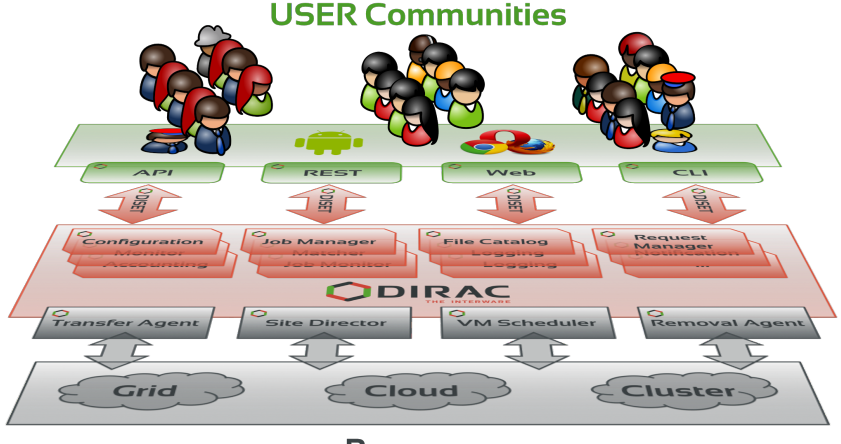
DIRAC

Cadre logiciel pour les calculs distribués : « Interware »

- Solution complète pour des communautés scientifiques
- Gestion de calcul et de données
- Interface unique entre les utilisateurs et les ressources de calcul et de stockage
- Grilles, clouds, supercalculateurs, BOINC

Équipe

- Responsable scientifique et technique : A.Tsaregorodtsev (CPPM), J.Bregeon (LPSC)
- Labos : CPPM, LPSC, LUPM, CC-IN2P3, IPHC
- ~3 ETP

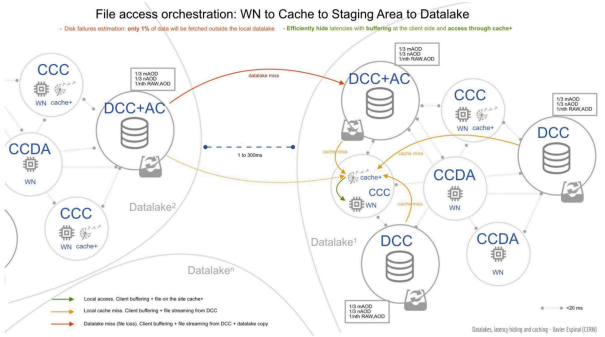


DOMA-FR

Composante Française des activités DOMA (Data Organization Management Access)

Objectifs :

- Etudier et valider les concepts et approches (techniques et opérationnelles) qui permettrons de mettre en œuvre les solutions de stockage des données scientifiques à l'horizon 2025-2027.
- Les activités suivies en France se veulent en phase avec les objectifs globaux de la collaboration mais aussi avec les contraintes. moyens et ambition de chacun des laboratoires impliqués.
- Les actions menés dans le cadre de ce projet sont motivées par l'objectifs HL-LHC mais se veulent plus génériques et ambitionnent de fournir au sens large, les futures solutions de stockage de données.



Équipe

- Responsable : Éric Fède
- CC-IN2P3, IPHC, IJCLAB(LAL), LLR, LPSC, CPPM, LPC, LAPP (participe effectivement aux activités sans être formellement membre du projet)
- ~ 1 ETP

Machine Learning/CompStat

Promouvoir le développement de l'utilisation de l'Intelligence Artificielle sur tout le périmètre de l'IN2P3.

- Par organisation de workshops et diverses actions pluri-disciplinaires
  - challenge TrackML
- promotion de thèse en co-direction Physique-ML

Équipe

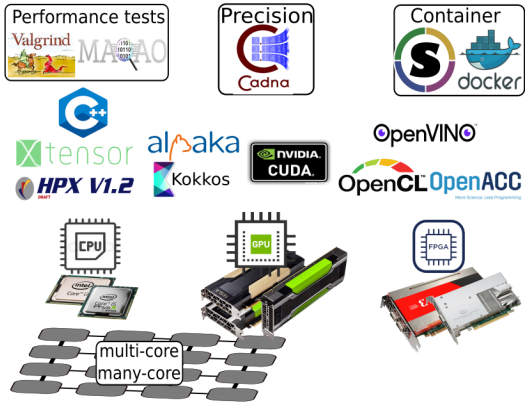
- Laboratoire : IJCLab, Clermont, Toulouse, mais aussi (sans ETP) LAPP, CPPM, LPSC, IPHC, APC, CC-IN2P3

DecaLog : 10 ans pour gagner un facteur 10

Le matériel de calcul se parallélise et se diversifie.

Du développement d'un logiciel à son déploiement, comment obtenir des performances durables ?

- Activité centrale : comparaison (performance, portabilité, productivité, précision) d'algorithmes et d'outils logiciels (OpenCL, OpenACC, Kokkos, SyCL, …) sur différents matériels (CPU, GPU, FPGA, …), y compris à travers des conteneurs (Docker, Singularity).
- En complément : échange d'expertises, développements spécifiques et adaptation des outils aux besoins propres de l'institut, recommandations aux chercheurs.
- Réalisations concrètes : publications, guides, ateliers et formations, partage de conteneurs, contributions à la prospective IN2P3.



Laboratoires impliqués

- APC, IJCLAB, IPHC, LAPP, LPC, LPNHE, LLR, LUPM, SUBATECH
- Responsable : David Chamont
- ~2,5-3 ETP

Geant4-core

Le projet regroupe les développeurs IN2P3 pour Geant4

Activités

- physique EM basse énergie, validations EM
- Géométrie
- Analyse, visualisation
- générateur conversion 5D
- réduction de variance, fast simulation
- et des tutoriels

Équipe

- Responsable : Marc Verderi
- Laboratoires CENBG, IJCLAB, LAPP, LLR
- ~1 ETP