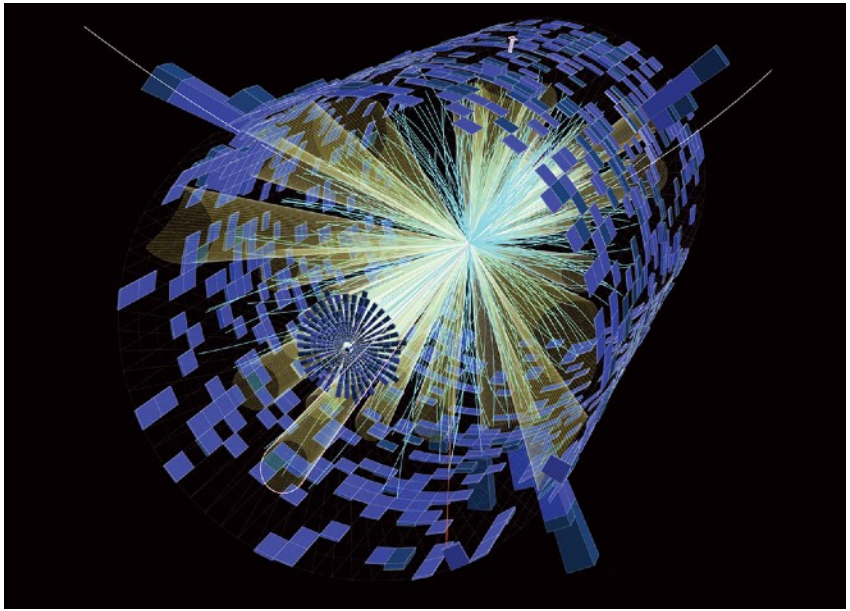
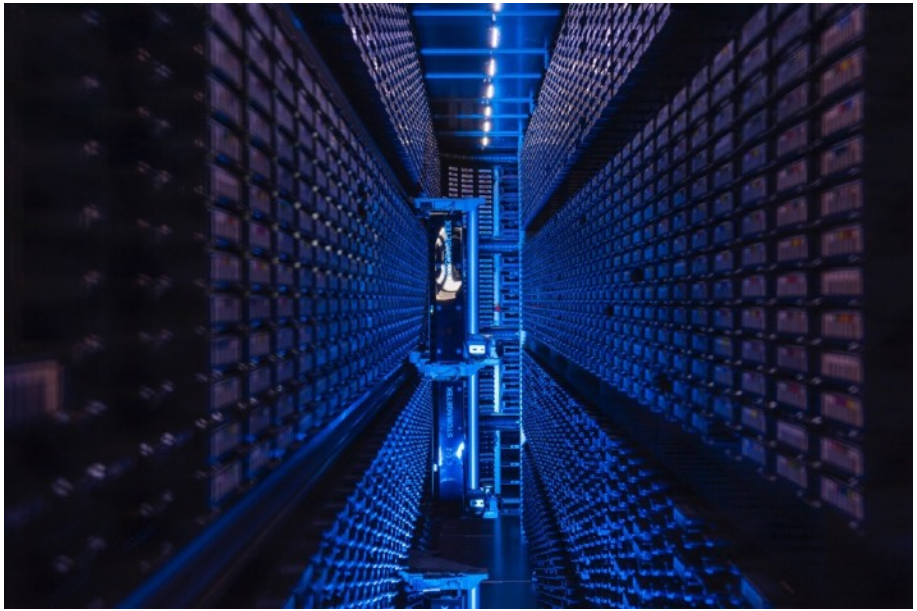
Institut national de physique nucléaire
et de physique des particules





Calcul et données à l'IN2P3

Sabine Crépé-Renaudin

Les défis

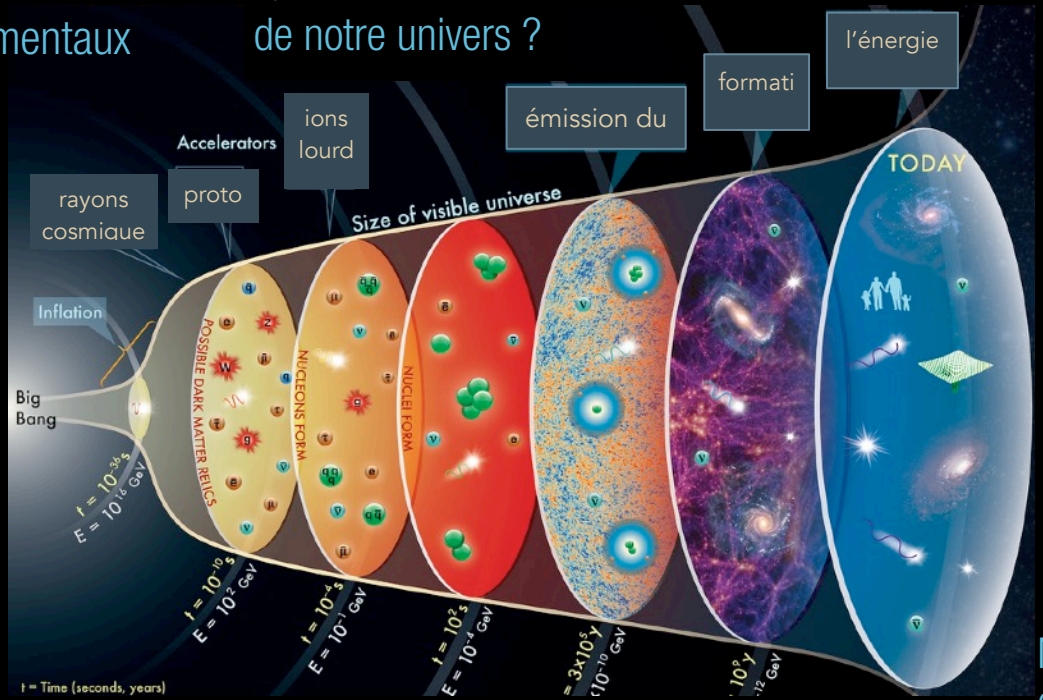
Des questions…

Quelles sont les éléments fondamentaux de la matière ?

Quelle est l'histoire de notre univers ?

Où et passé l'antimatière ?

Quelles sont les interactions fondamentales ?



Quelle est cette matière inconnue qui représente 85% de la matière de l'univers ?

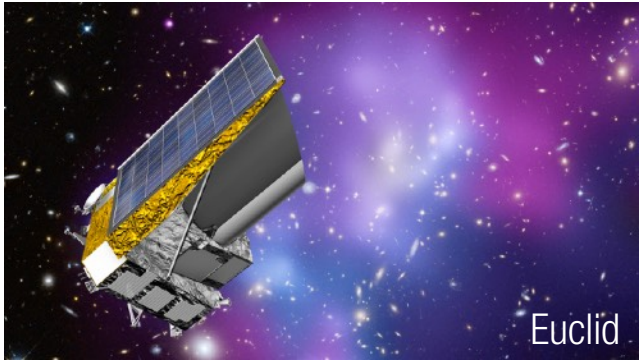
Pourquoi l'univers s'étend toujours plus vite ?

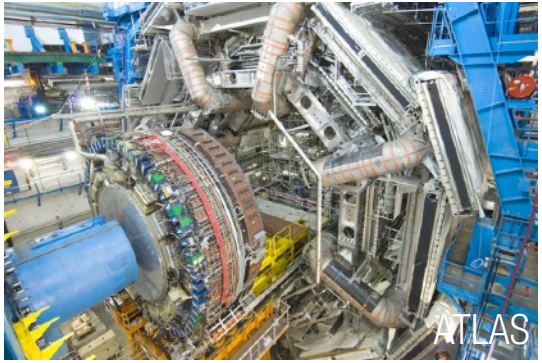
Notre univers est-il stable ? comment la masse vient aux particules ?

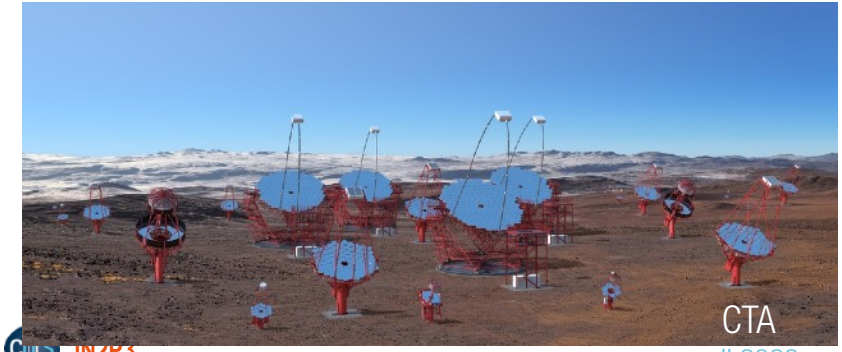
Quelles sont les propriétés de la matière nucléaire et son rôle de l'univers ?

De magnifiques projets scientifiques

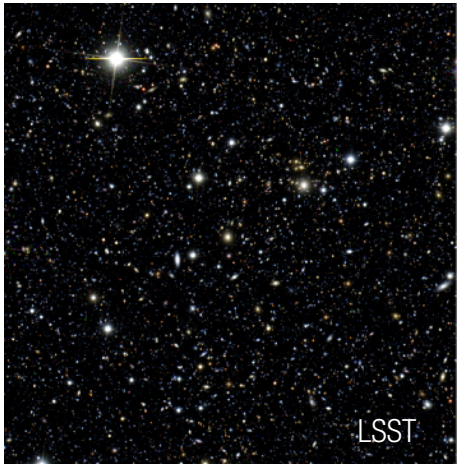
pour répondre à ces questions

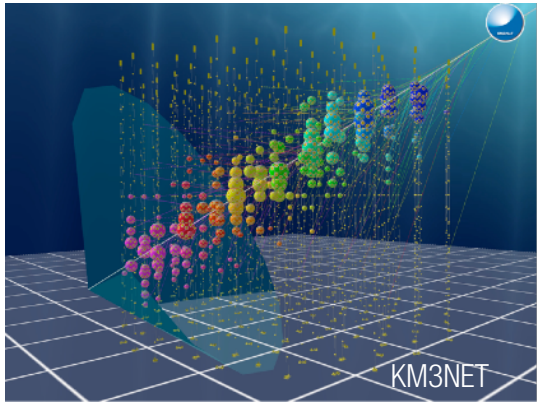


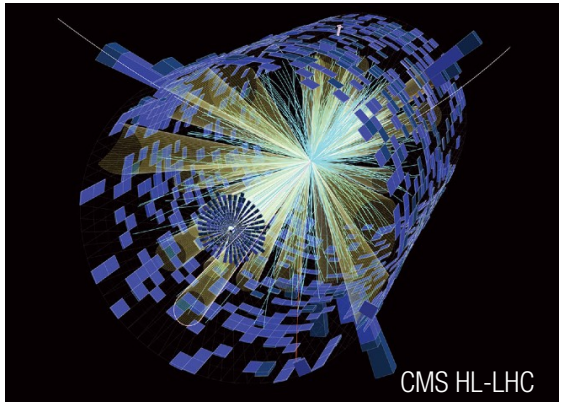


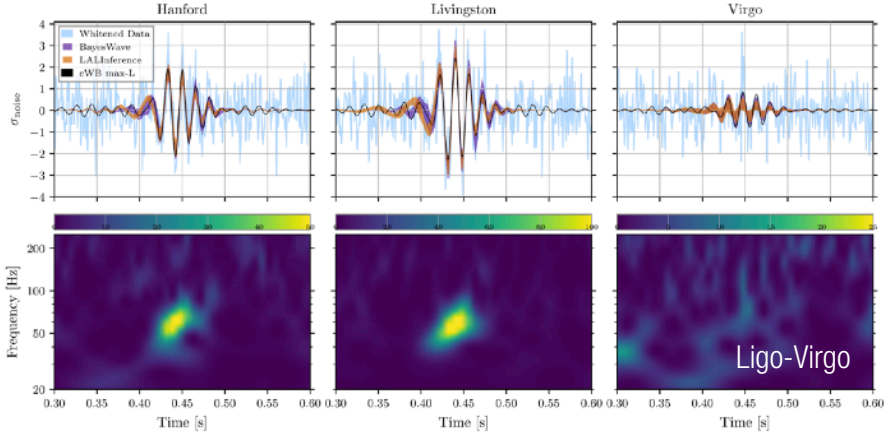


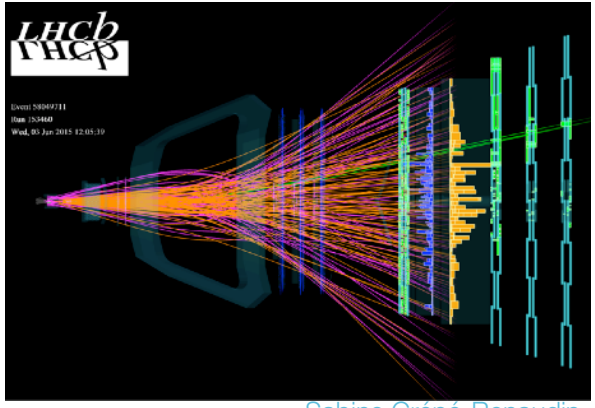
qui produisent des données complexes





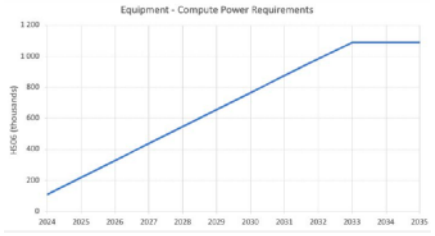






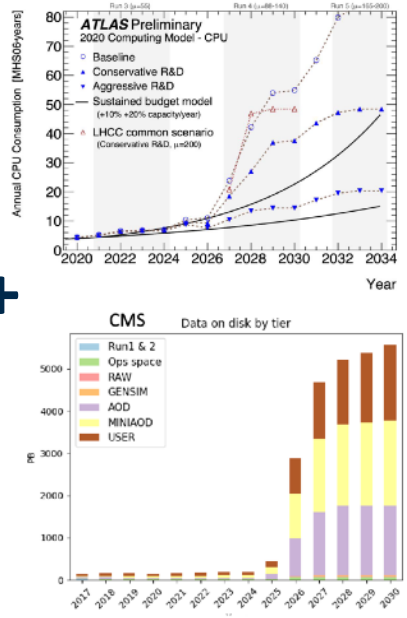
et qui sont très généreux en données

à traiter, stocker, déplacer, analyser



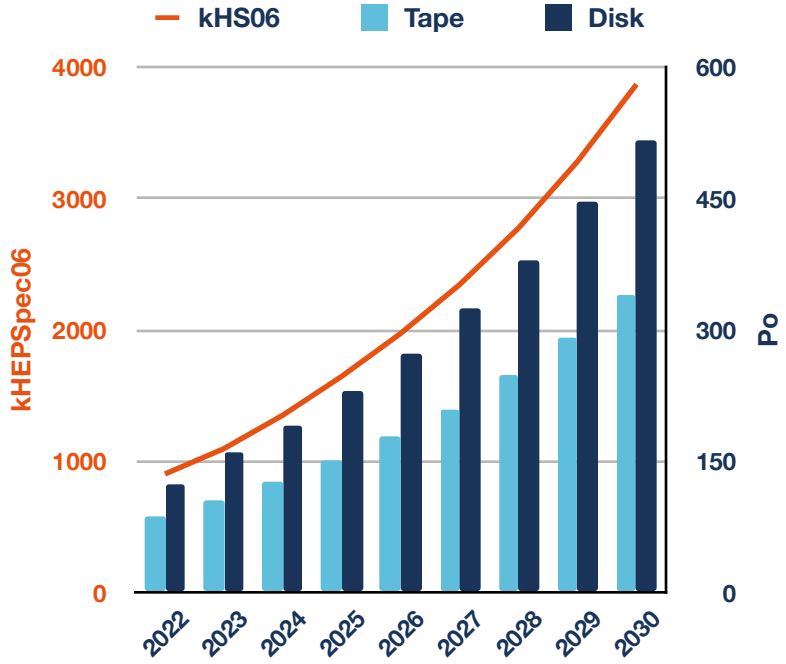
LSST

+



LHC

+ ... =



IN2P3 resource evolution projection

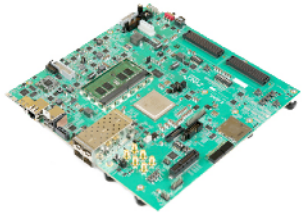
des technologies et des techniques

qui progressent et se diversifient : défis et opportunités

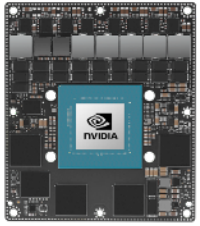
HTC

HPC

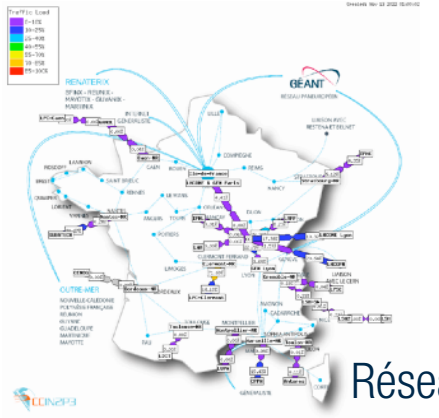
IA



GPU

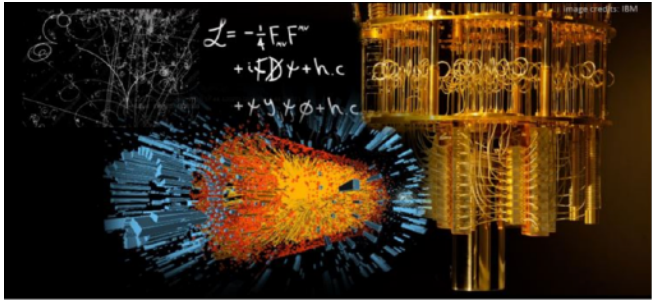


FPGA



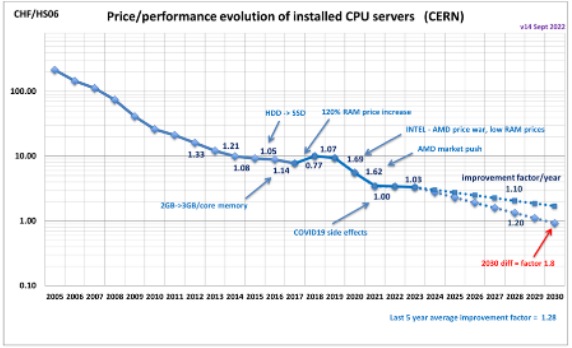
Réseaux

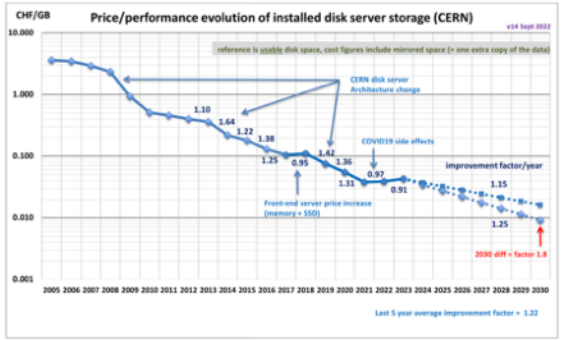
IQ



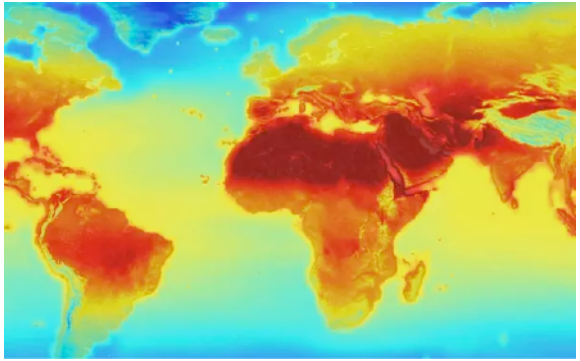
en tension

coût des matériels

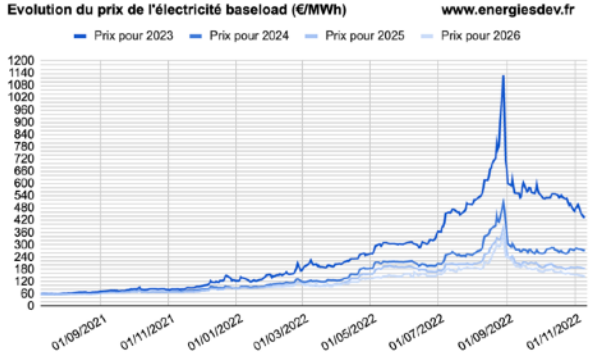




réchauffement climatique



coût des fluides



des ressources

Ressources humaines

Les nouvelles exigences de la Science Ouverte

Pourquoi ?

- recherche publique ➜ retour des résultats/données au public
  - d'ailleurs c'est dans le code de la recherche : « La recherche publique a pour objectifs le développement et le progrès de la recherche dans tous les domaines de la connaissance ; la valorisation des résultats de la recherche ; le partage et la diffusion des connaissances scientifiques ; le développement d'une capacité d'expertise ; la formation à la recherche et par la recherche et l'organisation de l'accès libre aux données de la recherche »
- reproductibilité des résultats scientifiques
- réutilisation des données par d'autres pour éviter la duplication du travail, et pour obtenir de nouveaux résultats scientifiques en croisant les données

Données ouvertes

- de bonnes pratiques pour la gestion des données sur tout le cycle

  ➜ collecte ➜ description/identifiant DOI ➜ référencement ➜ stockage ➜ traitement et analyse

  ➜ effacement ou archivage ➜ ouverture… ou pas
  - Aussi ouvert que possible aussi fermé que nécessaire : données sensibles (personnelles, sécurité, propriété intellectuelle/brevet…)
- déjà en application pour certaines expériences

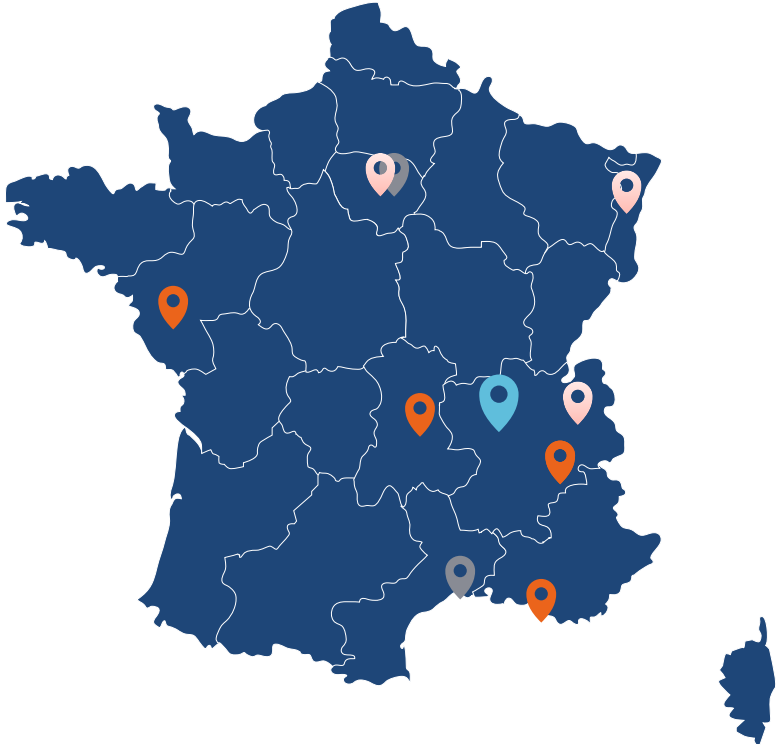
Implications

- à élargir à toutes les expériences et plateformes de l'IN2P3
- archivage : stockage sur le long terme + logiciel et possibilité de réutilisation
- ouverture : mettre à disposition les données intéressantes

Des infrastructures

➜ à la hauteur des défis à relever

Les infrastructures



Principaux centres de calcul et de stockage de l'IN2P3

- Centre national CC-IN2P3, Tier1 de WLCG
- Les centres régionaux : les Tier2s de WLCG, les plateformes de l'IN2P3 (aussi ouvertes vers les universités) : MUST, SCIGNE, Virtualdata et FACe

Infrastructures hors IN2P3

- Centres HPC: IDRIS (CNRS), TGCC (CEA), CINES (Universités)
- Renater: réseau national

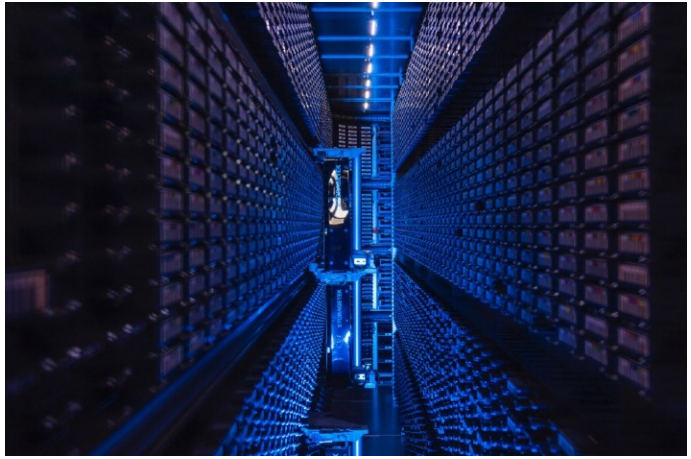
CC-IN2P3

Description et missions

- Infrastructure de recherche nationale pour nos thématiques de recherche (LHC/HL-LHC T1, LSST, Belle II, Euclid, JUNO, DUNE,…)
- Fournit du stockage (disque et bande) et des resources de calcul avec une architecture appropriée à nos besoins
  - Principalement HTC mais une part croissante de ressources GPU et HPC
- Fournit des services associés
  - connexion des sites IN2P3 à Renater
  - outils pour les développements logiciels et outils collaboratifs

CCIN2P3

- 2 salles informatiques : 1700 m²
- > 300 racks
- 600 kHS06 ~ 52 000 cores
- 150 PB disque et bandes
  - capacité : 340 PB
- 85 personnes
- Utilisateurs
  - 150 équipes
  - ~ 4000 utilisateurs actifs



Les plateformes

FACe : https://si-apc.pages.in2p3.fr/face-website/
Centre François Arago

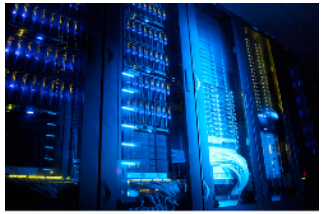
- Cluster HPC DANTE (multi Data ANalysis and compuTing Environment)
  - 652 coeurs, 42 To
- cloud FG
- Observatoire multi-messager (MMO), développement Euclid
- R&D sur les technologies collaboratives, formation

MUST : https://www.must-datacentre.fr/
Mésocentre de calcul et de stockage de l'USMB

- traitement et à l'archivage des données
- HTC 6100 coeurs, GPU, 20 Po, réseau 20Gb/s
- LHC du CERN (WLCG), CTA, LSST, HESS, …
- accompagnement des entreprises via le projet IDEFICS

SCIGNE : https://scigne.fr/
Scientific Cloud Infrastructure in Grand Est

- Calcul HTC, cloud IaaS / conteneur as a service et GPU à la demande
- Gestion et archivage de données
- 4000 coeurs, 3 Po, réseau 20Gb/s
- formations
- France Grilles, IFB, EGI et WLCG

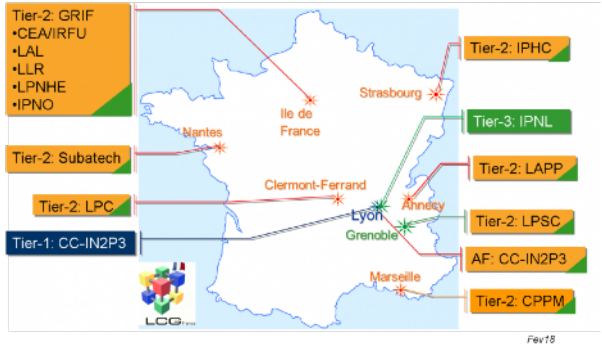


VirtualData

- HTC grille site GRIF
- Cloud Virtual Data : 10000 cœurs, 500 TB
- service Spark et JupyterHub => FINK, enseignement
- service multimessager GRANDMA

Les projets autour des infrastructures

Les projets autour des infrastructures

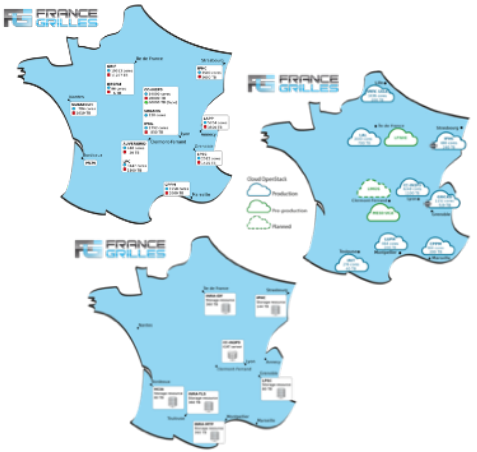


LCG France

- Composante française de l'infrastructure distribuée de WLCG pour le calcul LHC
- 8 sites distribués dans 12 laboratoires
  - T1 CC-IN2P3, 7 T2s => LPSC et Subatech vont s'arrêter
- > 10% des ressources WLCG: 700 kHS06 ~70 000 coeurs, 65 Po disque et 100 Po bande

France-Grille/France-Cloud

- GIS (CEA, CNRS, CPU, INRA, INRIA, INSERM, MESRI, RENATER) construit à partir de la grille LHC française
- France-Grille représente la France à EGI (European Grid Infrastructure) et contribue à son fonctionnement
- Fournir les services pour les besoins en stockage et calcul des données scientifiques sur des infrastructures distribuées comme la grille, le cloud ➜ services ouverts à toutes les communautés scientifique
- permet le partage d'expertise et offre du support pour différentes communautés scientifiques
- futur et rôle dans EOSC France en discussion



La construction de l'EOSC

EOSC ???

- Établir un partenariat entre tous les acteurs européens pour permettre à tous les chercheurs d'accéder aux données scientifiques, e-infrastructures et services
  - Fédérer les infrastructures et les services existants, favoriser la science ouverte et favoriser l'émergence d'un cloud européen
  - Document : Qu'est-ce que l'European Open Science Cloud ?

Structuration du paysage

- Création de l'EOSC Association AISBL et constitution des groupes de travail de l'association (plusieurs membres de l'IN2P3)
- Collège EOSC créé au ministère
- Création de groupes miroirs pour mieux influer sur l'écriture des appels à projets au ministère et au CNRS (retour sur les drafts 2023-24 via CNRS, ministère, EOSC, EGI, ESCAPE -clusters-)
- Info EOSC-France : : https://drive1.demo.renater.fr/index.php/s/47cbm4kYwfYBERf
- Mailing List info EOSC-France : https://groupes.renater.fr/sympa/info/eosc_france_info

Les projets EOSC à l'IN2P3

- EOSC-Pillar : Utiliser les initiatives nationales des membres européens pour construire un EOSC de la science ouverte et des données FAIR
  - LPC, CC-IN2P3, IPHC, IN2P3

- ESCAPE : The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures ➜ ESCAPE Open Collaboration Agreement à partir de 2023
  - dirigé par Giovanni Lamanna
  - LAPP, CC-IN2P3, CPPM

- EGI-ACE (Advanced Computing for EOSC) : construire la plateforme de calcul pur l'EOSC
  - CC-IN2P3, CPPM, IPHC

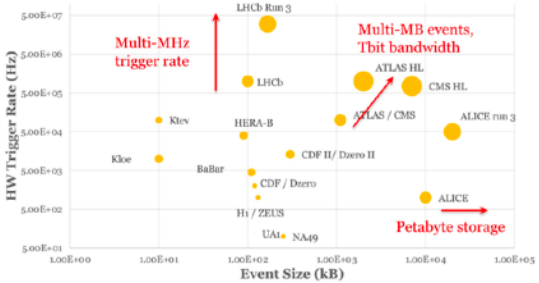
- EOSC-Future : définition de la vision stratégique pour EOSC
  - Connecter et intégrer les Infrastructures, les données et les services
  - LAPP, CC-IN2P3, CPPM

- Discussion en cours pour répondre à l'appel d'offre marché public pour les services coeurs de EOSC

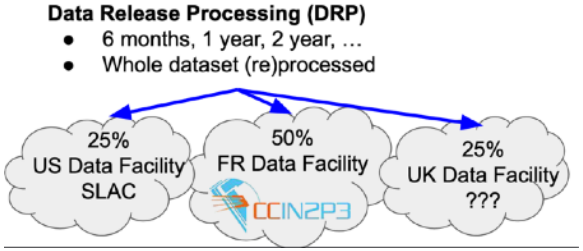
Les contributions aux expériences et les logiciels

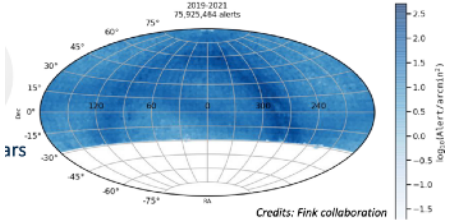
Des contributions majeures dans les expériences

- Physique des particules
  - LCG
  - Développement analyse en temps réel :
    - LHCb pour le run3
    - ATLAS et CMS pour le run4
  - contribution et préparation pour les expériences Belle II, DUNE en discussion



- Astroparticule et cosmologie
  - contribution majeure au calcul de LSST
  - contribution importante à CTA, Euclid
  - alertes (FINK), multimessagers





- Physique nucléaire
  - calcul ab initio, QCD sur réseau => HPC
  - contribution importante structuration et code pour les expériences de physique nucléaire (Agata, Nptools)
  - activité pionnière en informatique quantique pour les simulations de système complexe (n-corps) en physique nucléaire

- ....

Panorama* des logiciels

* non exhaustif

Cf CS juin 2022

Simulation

- **Geant4**
- **Geant4-DNA** (interaction entre les radiations ionisantes et le milieu biologique)
- **GATE** (imagerie médicale et application, basé sur Geant4)
- **nptool** (simulation et analyse des données pour la physique nucléaire à basse énergie (0-1 GeV)
- **Smilei** (code PIC massivement parallèle pour la simulation des plasmas)

Traitement des données et analyse

- **ACTS** (A Common Tracking Software) (HEP), code de reconstruction des traces indépendant des expériences
- **Gammapy** (astroparticle) python package for high-level γ-ray astronomy based on common data formats
- **AGATA** (nucléaire) : traitement des données issues du détecteur AGATA
- **KaliVeda** (nucléaire) : traitement des données issues des détecteurs INDRA et FAZIA
- **nptool** (nucléaire) : traitement des données pour les expériences basses énergies

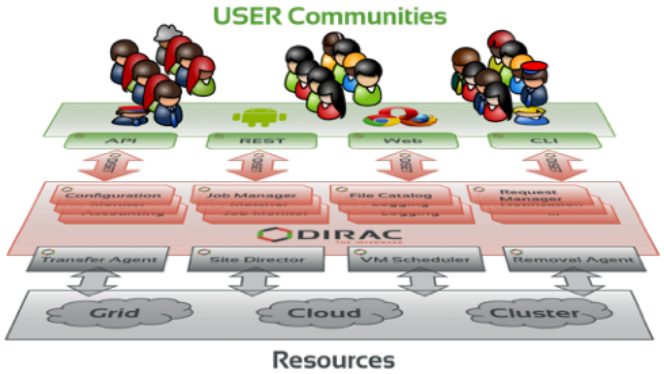
Gestion des tâches de calcul, des données et des métadonnées

- **DIRAC** : gestion des tâches de calcul et des données
- **AMI** : gestion des métadonnées

Les projets logiciel IN2P3

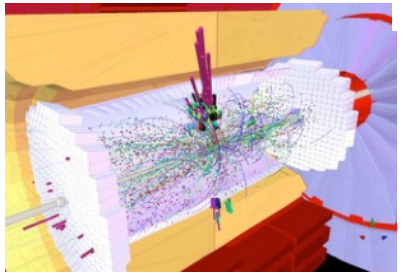
DIRAC

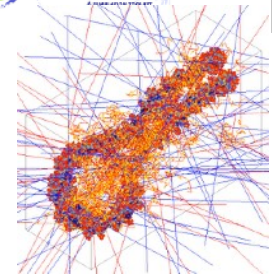
- Cadre logiciel pour les calculs distribués : « Intergiciel », interface unique entre les utilisateurs et les ressources de calcul et de stockage : Grilles, clouds, supercalculateurs, BOINC
- Solution complète pour les communautés scientifiques, gestion de calcul et de données
- Utilisé par LHCb, France-grille/EGI, ESCAPE, CTA, ILC, Belle II, Bes III, Pierre Auger, Juno, JINR, EISCAT, T2K, we-nmr, NA62



GEANT4

- Simulation des interactions des particules dans la matière largement utilisé dans toutes nos expériences
- Geant 4-DNA : modélisation des dommages biologiques induit par des radiations ionisantes à l'échelle de l'ADN





La recherche et les R&D

➜ indispensable pour relever les défis à venir

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, ...)

- Un poste de chercheur par an depuis 2020
  - CR de la section 06 en 2020 pour le CC-IN2P3 : 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
  - CR de la section 06 en 2021 : Vincent Reverdi a rejoint le LAPP dans le groupe LSST travaille sur la complexité logicielle
  - CR dans la section interdisciplinaire CID 55 en 2022 : Anja Butter rejoint le LPNHE, théorie-ATLAS : intelligence artificielle

Soutien aux projets de R&D => projets IN2P3

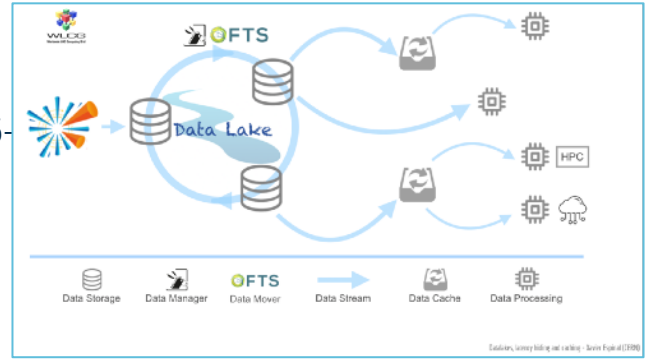
- 4 projets en cours : DOMA (se termine), Decalog, ML, QC2I
- encouragement à rejoindre ces projets et à en proposer d'autres (sobriété ?)
  - leçon des prospectives techniques : garder du temps pour ces activités
- encouragement à encadrer des étudiants (stages et thèses) et à publier [chercheurs et ingénieurs]
- nécessité de renforcer les liens à l'international => HSF, CERN, …

Les projets de R&D autour des infrastructures et des logiciels

DOMA-FR

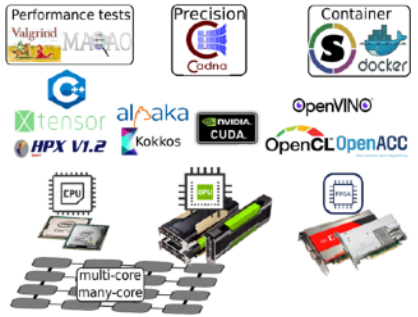
- DOMA = Data Organisation Management Access
- R&D sur les solutions de stockage et d'accès des données scientifiques à l'horizon 2025-HL-LHC et autres expériences
  - data lake et data carousel (utilisation des bandes en ligne)
  - se termine : développements déjà en production

➜ lien avec le projet ESCAPE



DECALOG : Reprises et ComputeOps

- Objectif : améliorer les performances à partir du développement d'un logiciel jusqu'à son déploiement
- Comparaison (performance, portabilité, productivité, précision) d'algorithmes et d'outils logiciels sur différents matériels (CPU, GPU, FPGA, …), y compris à travers des conteneurs (Docker, Singularity), nouvelles technologies d'administration
  - CSAN catalogue d'applications scientifiques open-source : https://gitlab.in2p3.fr/csan/csan/-/wikis/home
  - Calcul performant : https://reprises.in2p3.fr/

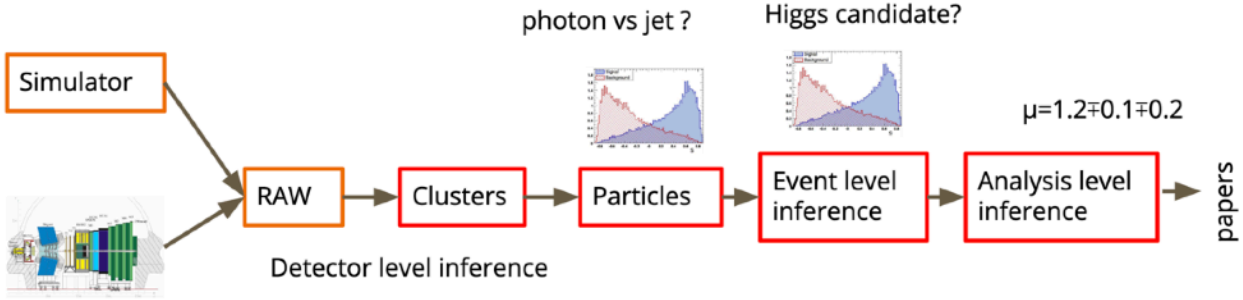


https://gitlab.in2p3.fr/CodeursIntensifs/DecaLog/-/wikis/home

Les projets de R&D autour des nouvelles technologies

Machine Learning : Compstat

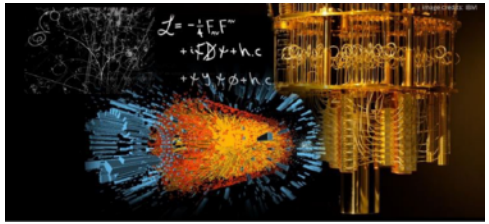
– Développement de l'Intelligence Artificielle sur tout le périmètre de l'IN2P3



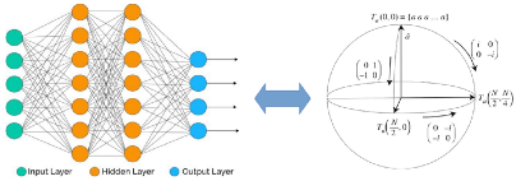
https://machine-learning.pages.in2p3.fr/

QC2I : Quantum Computing for the 2 infinites

- Préparer la transition vers les technologies de calculs scientifiques basées sur les processeurs quantiques
  - Apprendre à utiliser les algorithmes quantiques sur les plateformes actuelle et veille technologique sur les ordinateurs quantiques
  - Identifier des applications pilotes dans la physique IN2P3 et les utiliser comme jalons pour les ordinateurs quantiques et ainsi contribuer aux avancés dans le domaine de l'informatique quantique
  - beaucoup d'intérêt, encore peu de contributeurs



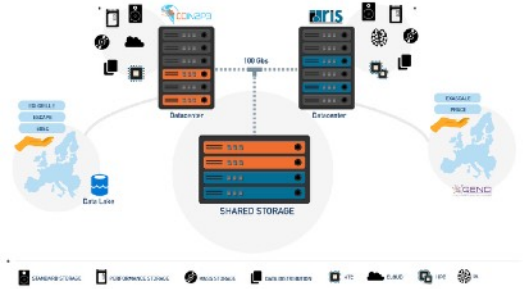
https://qc.pages.in2p3.fr



Les projets autour du HPC

Utilisation du HPC

- utilisé depuis longtemps à l'IN2P3 : QCD sur réseau, calcul ab initio phys nucléaire, simulation plasma
- devient de plus en plus important aussi dans LCG, intelligence artificielle
- utilisation freinée par les barrières d'accès liées au mode projet et à la sécurité des infra françaises (ZRR)



Projets

- Equipex FITS : collaboration entre le CC-IN2P3 et l'IDRIS + GENCI
  - augmentation capacités des sites avec des salles à l'état de l'art
  - développement de solutions pour permettre un accès transparent aux ressources HTC et HPC (stockage partagé, portail d'accès commun) pour les IR du CNRS

- Projet Exascale
  - Appels européen EuroHPC pour la mise en place de supercalculateurs HPC : France candidate pour l'appel en 2023-24
  - Phase de préparation pour répondre à l'appel et déterminer le design du calculateur (TGCC CEA)
    - sera de toute façon principalement composé de GPU
  - Projet Exascale = Groupe de travail DGRI GENCI et CEA CNRS INRIA Universités pour lister les applications qui pourraient utiliser un tel calculateur et évaluer les besoins pour porter les applications sur ces GPUs
  - une vingtaine d'applications proposées par les chercheurs/IT IN2P3 : QCD sur réseau, simulation plasma (Smilei), simulation ab initio, LHC (LHCb), etc
  - Résultat du GT plus complet que anticipé : large panorama des applications => volonté de garder le contact
  - pas de financement en vu pour le portage des codes, remonté des priorités auprès des organismes
  - Document résumant les contributions et évaluant les besoins publié : https://hal.archives-ouvertes.fr/hal-03736805

PEPR autour du numérique

PEPR ?

- PEPR = Programmes et équipements prioritaires de recherche du quatrième programme d'investissements d'avenir (PIA4)
- PEPR exploratoires (50M€) et PEPR des stratégies nationales (ex 150M€ Techno quantique)
- PEPR exploratoires proposés autour du calcul sont sur de la recherche en amont sur ces thématiques => pas au coeur de nos activités mais nous sommes sollicités pour certaines de nos activités spécifiques (grande masse de données, transferts de données I/O, traitement distribué des données) et pour des mises en applications, des preuves de concepts

PEPR Cloud

- PEPR exploratoire sur la recherche sur les clouds piloté par le CEA et INRIA avec CNRS, IMT et UDICE comme partenaires (56M€)
- Proposition thématique acceptée
- Projet défini : coordination brique applicative I/O orientée, contributions proposées IN2P3 dans la partie CNRS (CC, IPHC, IJCLab,CPPM)
- Coordination brique applicative I/O orientée, contributions proposées IN2P3 dans la partie CNRS (CC, IPHC, IJCLab, CPPM)
    - placement intelligent des données, reconfiguration des réseaux par les applications, interopérabilité des infrastructures de Cloud

PEPR NumPex HPC

- PEPR exploratoire Numérique haute performance pour l'exascale, portée conjointement par le CEA, CNRS et Inria (40 M€)
- Proposition accceptée, détail du projet en cours de définition
- Développements proposés par l'IN2P3 (équipe recherche CC : prise en compte des incertitudes ds les ordonnateurs, dev online LHCb Allen) et applications (LHCb, CMB,IA LHC hardware agnostic models, QCD sur réseau, SMILEI)

PEPR IA

- dans le cadre de la stratégie nationale pour l'IA, pilotage CEA, CNRS, INRIA (73 M€)
    - IA frugale, IA décentralisée et de confiance, fondements mathématiques, recrutements jeunes talents internationaux, industrie, start-up

Organisation,information

Organisation IN2P3

DAS Calcul et Données

- CC et les plateformes
- projets
- lien CNRS (DDOR)
- structures internationales
- IST

Parcours obligé pour les nouvelles expériences

- réfléchir au besoins informatiques, les définir en amont
- informer le DAS thématique et le DAS C&D
- discussion avec le CC et première évaluation des coûts
- prise en compte de ces coûts dans le financement et les discussions avec les partenaires
- décision des activités prises en charge et financées par l'IN2P3

Plateformes Calcul et Données

- CC-IN2P3
- FAce, MUST, SCIGNE, Virtual-Data
- T2s de WLCG

Projets Calcul et Données

Autour des infrastructures

LCG-France

FRANCE GRILLES

DIRAC THE INTERWARE

DOMA-FR

Autour du calcul et des logiciels

GEANT4 A SIMULATION TOOLKIT

Decalog

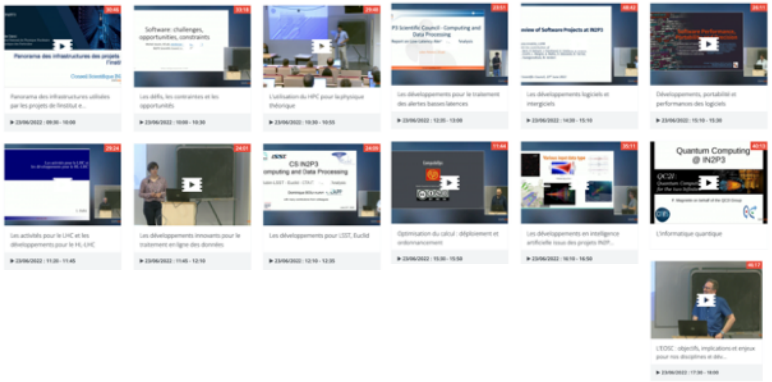
Machine Learning

QC2I
Quantum Computing

Rendez-vous

CS Calcul et Données juin 2022

- panorama des activités récentes à l'IN2P3 ➜ important travail très intéressant
  - site du CSI : https://www.in2p3.cnrs.fr/fr/le-conseil-scientifique-de-lin2p3
  - indico : https://indico.in2p3.fr/event/27438/
  - webcast : https://webcast.in2p3.fr/container/conseil-scientifique-de-l-in2p3-juin2022



Journées R&T

- Dernières journées : 17-19 octobre 2022 à Lyon ➜ https://indico.in2p3.fr/event/26475
- Objectifs : faire le point des avancées technologiques et les diffuser, susciter et organiser de larges débats techniques, coordonner, structurer les actions de R&T

➜ projets R&T informatiques participent

➜ un bon endroit pour initier des discussions entre technologies

Journées Informatiques

- https://indico.in2p3.fr/event/27495/

Rester informés

ML/IA

- Liste de diffusion : MACHINE-LEARNING-L@IN2P3.FR
- Workshop ML : https://indico.in2p3.fr/event/27507/
- Création d'un centre dédié à l'IA au CNRS : AISSAI
  - https://www.cnrs.fr/fr/le-centre-artificial-intelligence-science-science-artificial-intelligence-aissai
  - en préparation un semestre thématique pour l'IN2P3 discuté au workshop

HPC

- Liste de diffusion : HPC-INFO-L@IN2P3.FR
- suite projet Exascale ?

Développements autour des logiciels

- projet Decalog, HSF
  - listes internes au projet : reprises-l@in2p3.fr, computeops-l@in2p3.fr.
  - à utiliser plus largement ?
- des besoins d'échange de coordination ?
- des logiciels efficaces = un des meilleurs levier vers la sobriété !

Les projets européens

- Suivi des nouveaux appels : drafts sur Atrium, liste euro-computing-L
  - Au moins une personne par labo
  - des infos aussi sur les autres appels à projets
  - Note : souvent un seul projet retenu à la fin => création de consortium, nécessité de se préparer en amont

Conclusions

Pour conclure

Les données et leur traitement sont toujours au coeur de nos expériences et de notre travail scientifique

- depuis la préparation des expériences, en passant par leur exploitation et jusqu'à la publication
- nos besoins liés au traitement des données augmentent et deviennent de plus en plus complexes
  - avec des défis à relever : diversification des technologies, incertitude sur les coûts matériel/fluides/réseau, minimisation impact environnemental, attractivité
  - et aussi des opportunités : nouvelles idées, nouvelles technologies

➜ Nécessaire de réfléchir en amont dès la préparation des expériences

- quelles données ? stockées où ? sur quel support ? jusqu'à quand ?
- comment les traiter ? où les traiter ? quelles machines ? quelle puissance de calcul ? quel logiciel ? quelle efficacité ?

➜ Intégrer le traitement des données dès la conception de l'expérience

➜ Faire un DMP, contacter le DAS thématique et le DAS calcul et données puis le CC pour affiner les besoins techniques et chiffrer

Pour conclure

Nous avons des atouts pour relever les défis

- Vous !
- des infrastructures de qualité
- de fortes implications dans les projets informatiques des expériences
- nos projets de R&D

La R&D

- Les projets de R&D ouvriront la porte aux solutions de demain
  - nouvelles technologies, nouvelles méthodologies mais aussi des logiciels efficaces seront cruciaux

➜ Encouragement à participer aux projets calcul et données et aux développements R&D qui nous permettront de relever ces défis

➜ Encouragement à encadrer des thèses, des stages en informatique

➜ Activités à valoriser : diffusion, publication, renforcer les liens à l'international

des questions ?