Ressources de calcul parallèle au CC-IN2P3

Ecole IN2P3 d'informatique ï Parallélisme sur matériel hétérogène

27 mai 2016

Plan

- Le Centre de Calcul de l'IN2P3
- Histoire du calcul parallèle au CC-IN2P3
- Pourquoi développer le calcul parallèle au CC-IN2P3 ?
- Quelles ressources pour quels calculs ?
  - Matériel
  - Logiciel
  - Accès et support
- Bonus : Optimisation de codes parallèles
  - CoE POP

Le CC-IN2P3

Institut National de Physique Nucléaire et de Physique des Particules - IN2P3

Mission principale : déployer et opérer les moyens informatiques nécessaires à la mise en î uvre de la politique scientifique de lôIN2P3

Centre de calcul dédié fait partie dôun réseau mondial de grands centres de calcul pour la physique des hautes énergies

Calcul et stockage pour la physique des 2 infinis :
- physique nucléaire
- physique des hautes énergies
- astro-particules

Le CC-IN2P3 est une Unité de Service et de Recherche du CNRS fonctionnant en partenariat avec le CEA / DRF / Irfu

Le CC-IN2P3 en quelques chiffres

- Fonctionnement 24/24, 7/7, 365/365
- 2 000 m2 de bureaux ~100 personnes (hébergés inclus)
- 4 000 m2 de locaux techniques
- 2 salles informatiques de 850 m2

- Hébergement de points de présence opérateurs réseau nationaux (Renater) et régionaux (Lyres, Amplivia…)

Capacités actuelles (mai 2016)

CPU

30 856 vcores - 312 kHS06

Stockage Disque (dCache, iRods, xRootd, GPFSé)

Performance standard = ~ **20 Po**
Haute performance = ~**2 Po**

Bandes

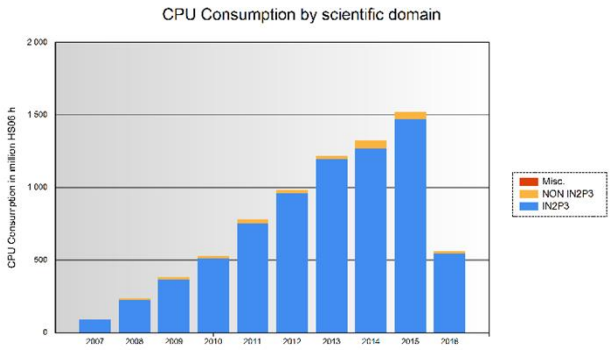
Volume stocké sur bandes : **31 PB** (capacité de **340 Po**)

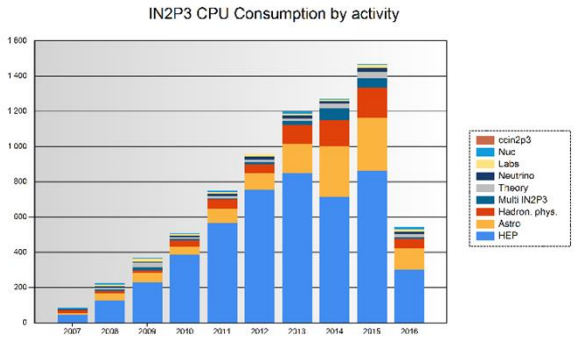
Sauvegarde *(TSM)*

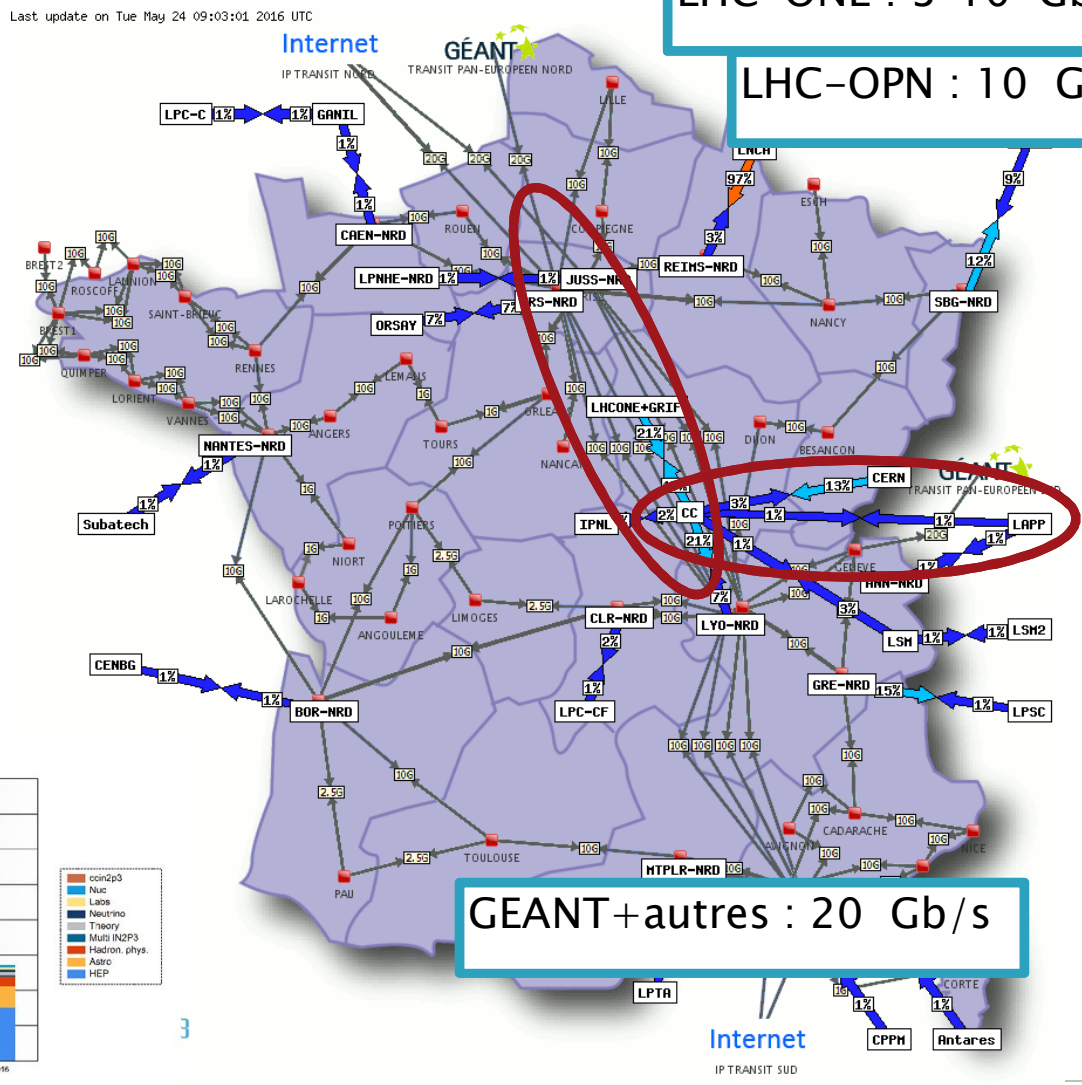
Volume stocké : ~ 1,1 **Po** sur 5 Po possible

Utilisateurs

75 (+12) groupes actifs (dont 22 non « IN2P3 »)
Consommation CPU : **97 % IN2P3** / 3% autres

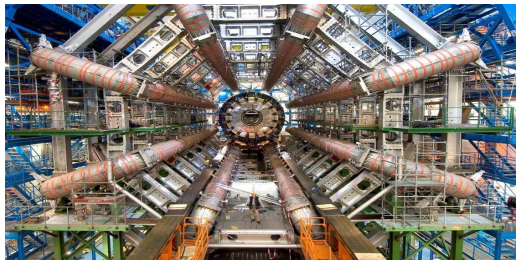


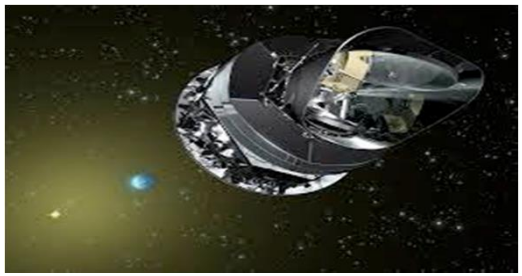


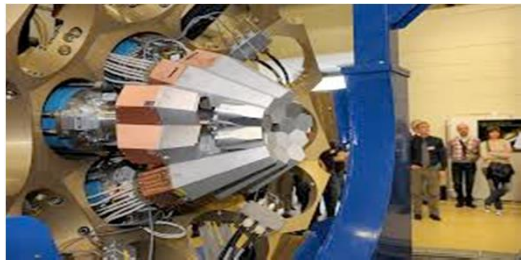


75 groupes actifs au CC-IN2P3

- **Physique des particules**
  - Modèle standard et au delà : ATLAS, CMS, D0, H1, é
  - Violation de symétries : LHCb, Babar, é
- **Physique des astroparticules**
  - Neutrinos : DOUBLE-CHOOZ, NEMO, OPERA, é
  - Cosmologie : PLANCK, EDELWEISS, SNLS, SNF, é
  - Rayons cosmiques : AMS, ANTARES, AUGER, FERMI, HESS
  - Ondes gravitationnelles : VIRGO
- **Physique Hadronique / Nucléaire**
  - Plasma quark-gluon : ALICE, Phenix, é
  - Structure nucléaire : AGATA, INDRA, é
  - Radiobiologie, Imagerie, é : HADRONTHERAPIE, é
- **Physique théorique**
  - QCD (interaction forte), NANTHEO (HEP), ...



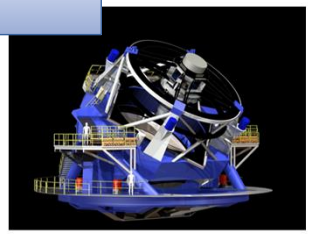


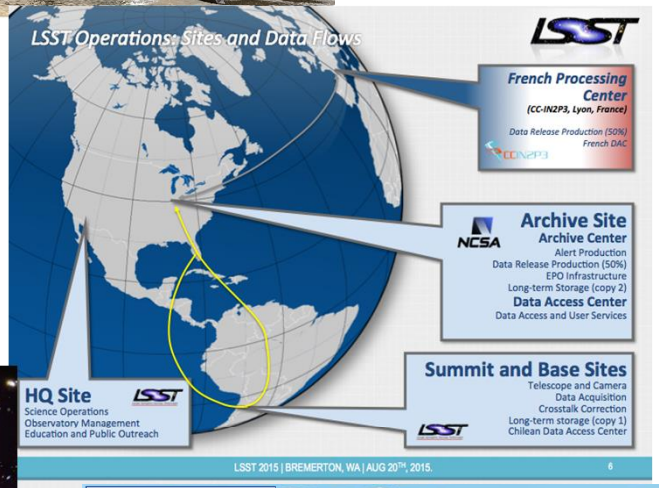


Estimation des ressources nécessaires en 2030

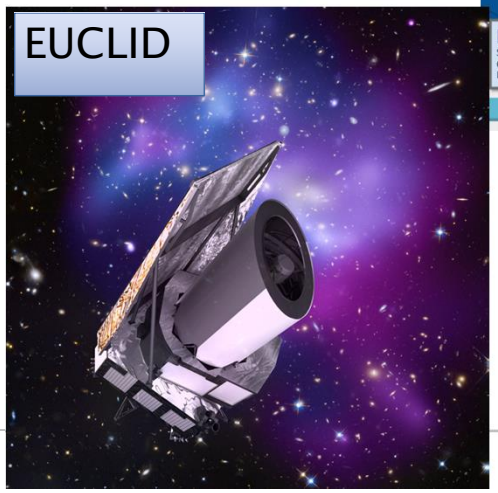
HL-LHC

LSST

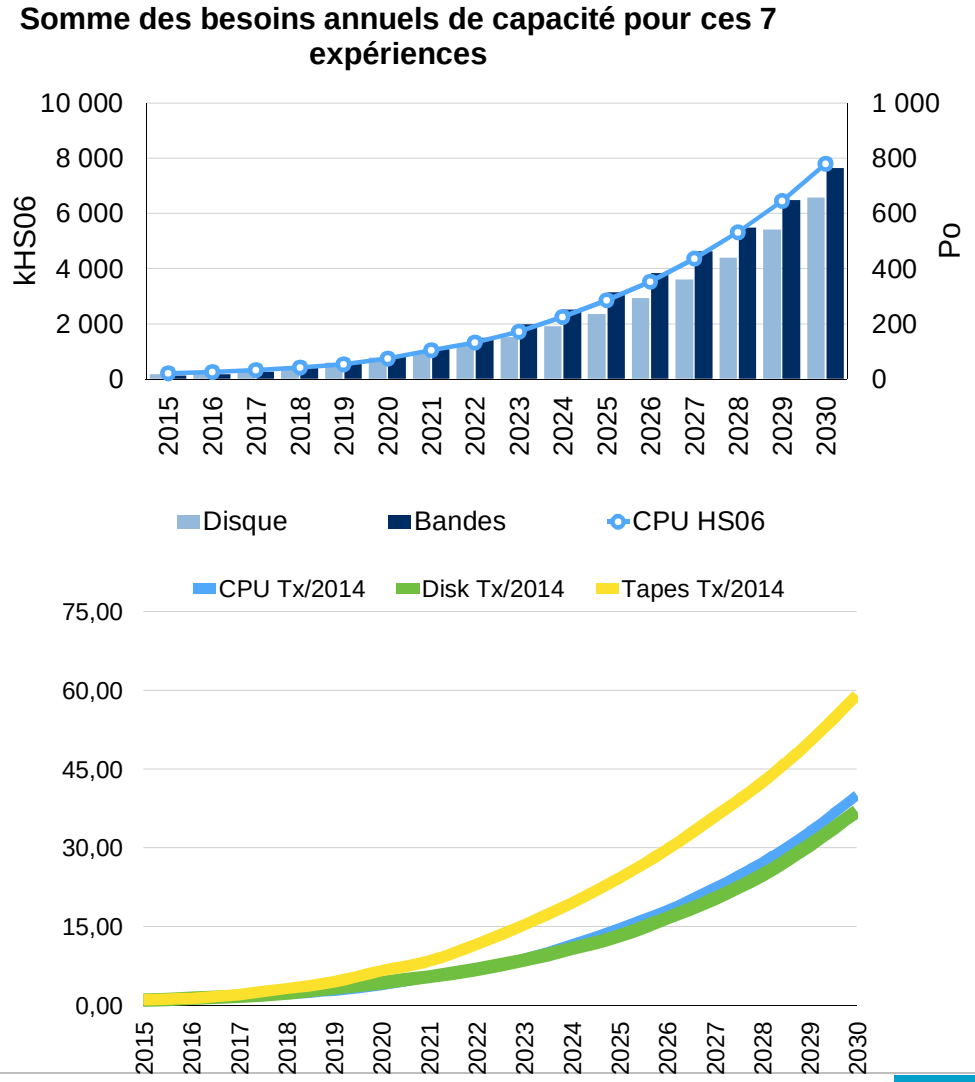




EUCLID

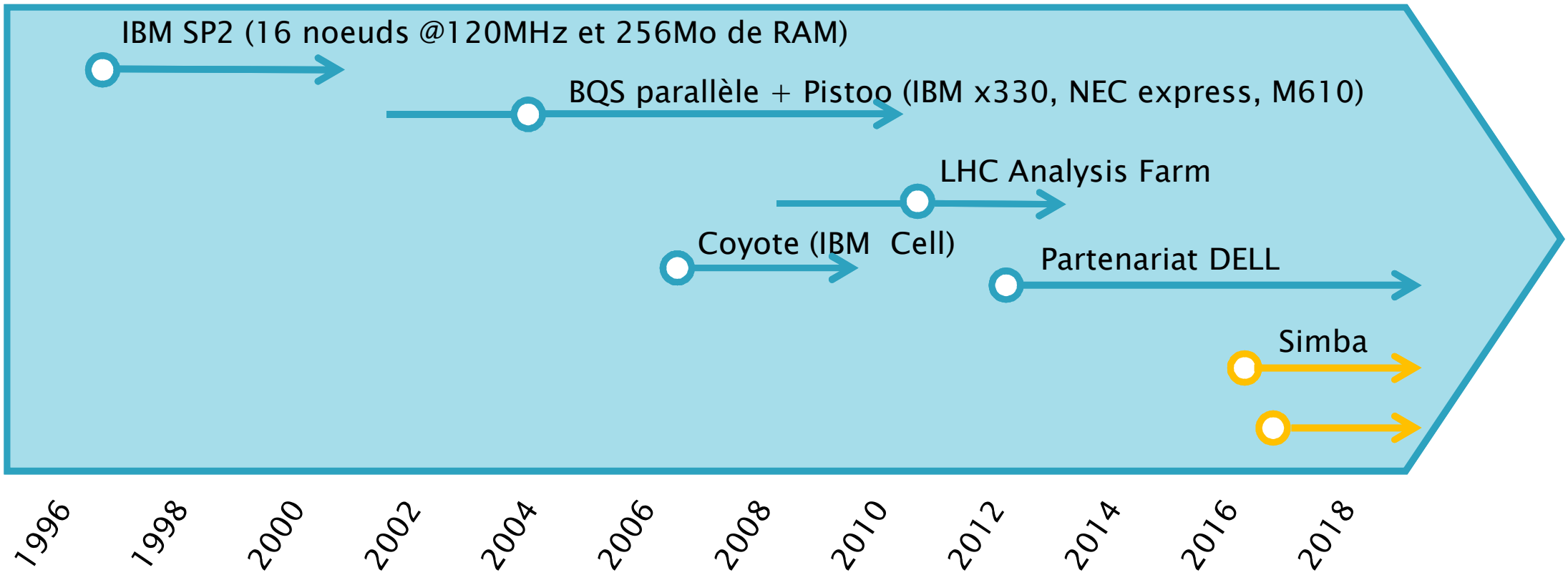


CTA



Histoire du calcul parallèle au CC-IN2P3

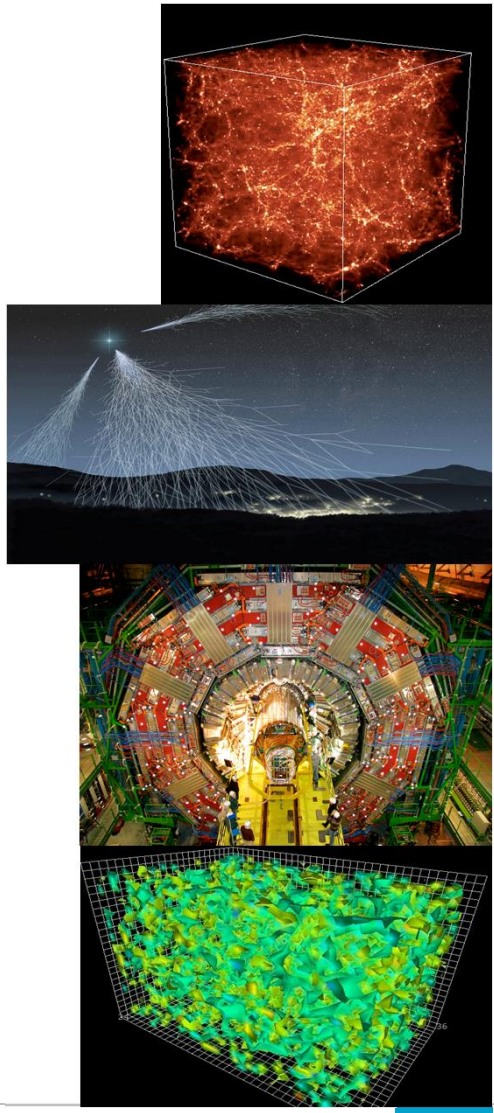
Histoire du calcul parallèle au CC-IN2P3



Pourquoi développer le calcul parallèle au CC-IN2P3 ?

Pourquoi développer le calcul parallèle au CC-IN2P3 ?

- Parce les utilisateurs le demandent !
  - Gros besoins mémoire → jobs multicoeurs
    - CMS, ATLAS, PLANCK, ...
  - Accélération des temps d'exécution → jobs MPI
    - AMS, EUCLID, ...
  - Sensibilité à la latence réseau → InfiniBand
    - QCD, N-Body, ...
  - Vectorisation du Calcul → GPUs
    - Traitement du signal (Virgo)
    - Traitement d'images (LSST)
    - Propagation de particules (Auger, CTA, KM3NET)
  - Hybridation des calculs



Mais les utilisateurs se débrouillent ...

Qui travaille sur quoi à l'IN2P3

- *Multi-coeurs*
  - Cadriciels Pp (Geant4,GaudiHive) : IPNO, LAL, LLR, LAPP, LPC-Clermont, CENBG.
  - Online, multi-tâches ADA : CSNSM, IPNO.
- *Vecteurs et accélérateurs*
  - Médical, simulation et imagerie sur GPU : CPPM, IPHC, IMNC.
  - Pp, tracking accéléré : IPNL, LAL, LLR (OpenCL), IPNO.
  - Pp, déclenchement haut-niveau accéléré : LAL
  - Astro, traitement d'images : LAPP, LLR.
  - GPU et Phi divers : IPN, LPC Caen, LPSC, SUBATECH, CENBG, LPC Clermont…
- *Multi-noeuds (MPI, HPC)*
  - Astro, cosmo : APC, CPPM, LAPP, LUPM.
  - Théorie, LQCD : LAL, CENBG.
  - Nucléaire : IPNO, LPSC, IPNL.
  - Accélération par laser : LLR.
  - Physique des réacteurs : LPC Caen, LPSC.
  - Neutrinos : SUBATECH.

2014/02/12 David Chamont & Gilles Grasseau

Qui calcule où à l'IN2P3

- *Supercentres*
  - TGCC/CCRT : APC.
  - IDRIS : CENBG, IPHC, LPSC.
  - CINES : APC LUPM.
  - CC "parallele" : SUBATECH…
  - NERSC : APC.
- *Mésocentres*
  - MCIA (Bordeaux) : CENBG (Groupe théorique)
  - Université d'Aix-Marseille : CPPM
  - MUST (Savoie) : LAPP
  - CIMENT : LPSC
  - HPC@LR (Montpellier) : LUPM.
- *Laboratoires*
  - Cluster MPI : APC, IPHC, IPNL, IPNO, LLR, LPC Caen, LPSC, LUPM, SUBATECH.
  - Accélérateurs (GPU,PHI) : CENBG, CPPM, IPNL ?, IPNO, LLR, LPC Caen.
  - Machine à grande mémoire partagée : LPNHE.

2014/02/12 David Chamont & Gilles Grasseau

Quel avenir parallèle pour le CC-IN2P3 ?

- Remplir son rôle de ... Centre de Calcul de l'IN2P3
  - Offrir les ressources demandées par les expériences
  - Assurer le support expert aux utilisateurs

- Mais pour du HPC (en plus du HTC « classique)
  - Sans chercher à devenir un centre HPC national
  - Etre l'équivalent d'un mésocentre (pour l'IN2P3)

- Développer synergies et passerelles avec l'IDRIS
  - Transfert d'expertises
  - Transfert de charge

Quelles ressources pour quels calculs ?

Cluster MPI ï Confguration matérielle

- 16 machines (DELL C6230) → 512 coeurs physiques
  - 2 x Intel Xeon E5-2698 v3 (@2.3GHz) → 32 coeurs / noeud

Hyperthreading désactivé

  - 8 x 16Go RDIMM (@2133MT/s) → 4Go par coeur
  - 2 x 1To SATA (@7200 rpm) → 2To par noeud

- 1 Switch Infiniband Mellanox IS5030 → IB QDR 40Gb/s
  - 36 ports QDR → 32 Gb/s utiles

- **Type de calculs: Jobs parallèle MPI fortement communiquants**

Cluster MPI ï Configuration logicielle

- Distribution Linux → CentOS 7
  - Configuré pour éliminer les latences autant que possible
- Runtime(s) MPI
  - OpenMPI 1.10 (avec support VERBS) → Support InfiniBand natif
  - MPICH 3.0 (avec IPoIB)
- Compilateur → GCC/G++ 4.8
- Outils supplémentaires dans /usr/local
  - Compilateur intel, gcc plus récent, bibliothèques, ...
  - Communs à toutes les ressources de calcul du CC-IN2P3
  - http://cc.in2p3.fr/docenligne/150

Cluster MPI ï Accès et utilisation

- Accès restreint
  - Se rapprocher du czar et du support CC-IN2P3

- Soumission similaire à celle sur la ferme généraliste
  - Via UGE en précisant lôutilisation dôun engine parallèle
  - -l os=cl7 -pe openmpi <number_of_cores> -q <queue_name>
  - -l os=cl7 -pe mpich2 <number_of_cores> -q <queue_name>

- Quelques variables dôenvironnement à fixer
  - http://cc.in2p3.fr/docenligne/969#paralleljobs pour les détails

Cluster GPU ï Configuration matérielle

- 10 machines (DELL C4130)
  - 2 x Intel Xeon E5-2640 (@2.6GHz) → 16 coeurs / noeud (32 threads)
  - 8 x 16 Go RDIMM (@2133MT/s) → 4 Go par thread
  - 1 x 400 Go SSD
  - **2 x NVIDIA Tesla K80** → **40 GPUs et 1.28 To RAM**
    - **4 992 coeurs CUDA**
    - **24 Go GDDR5**
    - **480 Go/s de bande passante globale**
- Interconnection 1Gb/s (public) + InfiniBand (swicth QDR cluster MPI)
- Disponibilité
  - Installation : été 2016
  - Production : automne 2016
- **Type de calculs: Calcul vectoriel**

Et OpenMP alors ?

- Déja disponible et utilisé sur la ferme généraliste
  - Jobs multicoeurs
  - Option UGE: -pe multicores <number_of_cores> -q <queue_name>
- Accès restreint
  - Se rapprocher du czar et du support CC-IN2P3
- Remarque
  - Nouvelles ressources parallèles → Jobs hybrides MPI/OpenMP

Optimisation de codes parallèles

POP CoE

- A Centre of Excellence
  - On Performance Optimisation and Productivity
  - Promoting best practices in parallel programming
- Providing Services
  - Precise understanding of application and system behaviour
  - Suggestion/support on how to refactor code in the most productive way
- Horizontal
  - Transversal across application areas, platforms, scales
- **For (your?)** academic AND industrial codes and users !

2

Partners

- **Who?**
  - BSC (coordinator), ES
  - HLRS, DE
  - JSC, DE
  - NAG, UK
  - RWTH Aachen, IT Center, DE
  - TERATEC, FR

- **A team with**
  - Excellence in performance tools and tuning
  - Excellence in programming models and practices
  - Research and development background AND proven commitment in application to real academic and industrial use cases

Tools

- **Install and use already available monitoring and analysis technology**
  - Analysis and predictive capabilities
  - Delivering insight
    - With extreme detail
    - Up to extreme scale
- **Open-source toolsets**
  - Extrae + Paraver
  - Score-P + Cube + Scalasca/TAU/Vampir
  - Dimemas, Extra-P
  - SimGrid
- **Commercial toolsets**

  (if available at customer site)
  - Intel tools
  - Cray tools
  - Allinea tools

5

Services provided by the CoE

? Parallel Application Performance Audit

- Primary service
- Identify performance issues of customer code (at customer site)
- Small effort (< 1 month)

! Parallel Application Performance Plan

- Follow-up on the audit service
- Identifies the root causes of the issues found and qualifies and quantifies approaches to address them
- Longer effort (1-3 months)

✓ Proof-of-Concept

- Experiments and mock-up tests for customer codes
- Kernel extraction, parallelisation, mini-apps experiments to show effect of proposed optimisations
- 6 months effort

7

Contact us !!

- If you have the feeling you are not getting the performance you expected
- If you are not sure whether it is a problem of your application, the system, ...
- If you want an external view and recommendations on suggested refactoring efforts
- If you would like some help on how to best restructure your code

POP Coordination

Prof. Jesus Labarta, Judit Gimenez
Barcelona Supercomputing Center (BSC)
Email: pop@bsc.es
URL: http://www.pop-coe.eu

9

Merci de votre attention
Questions ?