www.cnrs.fr

Bilan des activités du COCIN : Une vision de l'écosystème du calcul (et des données) en France

M. Daydé

Dr du Comité d'Orientation pour le Calcul Intensif (COCIN) au CNRS

Délégué Scientifique INS2I en charge HPC / Grille / Cloud

Rôle et missions du COCIN

- Créé en Décembre 2010
- Réflexion collective sur les besoins, la structuration et les évolutions en calcul intensif au CNRS
- Prospective sur les besoins des différentes communautés, proposition de maintenance et de développement coordonné des moyens / ressources liées au calcul intensif, en particulier pour l'IDRIS.
- Dix personnalités scientifiques désignées par chacun des instituts du CNRS plus le Directeur de l'IDRIS.
- Le président et directeur désignés par le Président du CNRS

Composition / Fonctionnement /Production

- Président : Ph. Baptiste (INS2I) parti au Ministère jusqu'en Avril 2013, Michel Bidoit nommé en Mai
- Dr : M. Daydé (INS2I)
- Membres : S. Bosi (INSHS), D. Girou (IDRIS), Ph. Helluy (INSMI), L. Lellouch (INP), P.-E. Macchi (IN2P3), Th. Meinnel (INSB), C. Pouchan (INC), G. Bornette (INEE), D. Veynante (INSIS), JP Vilotte (INSU)
- Invités : V. Breton (IDGC), O. Porte (DSI)
- *Livre Blanc sur le Calcul Intensif au CNRS fin 2012*
- *Propositions pour une nouvelle stratégie du calcul et des données au CNRS que le COCIN doit mettre en œuvre*
- Enquête sur l'Informatique en appui à la recherche en cours de finalisation

Livre Blanc sur le Calcul Intensif au CNRS

- Lancé en Avril 2011 sur la base d'un questionnaire
- Contribution CNRS à la réflexion lancée par GENCI sur les besoins des communautés et sa stratégie pour 2012-2016 (renouvellement IDRIS en 2012)
- Panorama du calcul intensif dans les Instituts : pratiques, besoins, verrous, cartographie, aspects recherche, …
- Certains des instituts (e.g. INC, INP) mènent régulièrement ce type de travail
- Objectif : travail finalisé fin 2012 avec pour objectif de mener des réactualisations régulières

Constats

- CNRS embrasse tout l'écosystème du calcul intensif : PRACE, IDRIS, CC IN2P3, Institut des Grilles, recherche, interaction avec les méso-centres (Gdr Calcul, UMR), … et toute la diversité des besoins / pratiques au CNRS (HPC, grille, cloud, …) et caractère interdisciplinaire
- Favoriser les échanges entre communautés HPC / grilles / Cloud
- Synergie avec les acteurs du HPC (CEA, GENCI, INRIA, universités, ...)
- Promotion du calcul intensif à tous les niveaux : centres nationaux et méso-centres
- Emergence rapide d'un fort besoin en matière de stockage / traitement sur les grands volumes de données

Le calcul : un enjeu stratégique

- Calcul intensif au cœur des grandes avancées de la recherche scientifique:
  - Génome humain, découverte potentielle du boson de Higgs, évolution du climat, risques naturels, pollution atmosphérique, environnement…
- De nombreux autres défis scientifiques :
  - Structure de l'univers, astrophysique, neuroscience, combustion, sismologie, climat, biologie et recherche médicale, matériaux, ….
- Enjeu stratégique de compétitivité et d'attractivité internationale: multiples champs disciplinaires; importantes retombées socio-économiques

Calcul Intensif

- Plus possible de dissocier le calcul haute performance de l'analyse et valorisation des masses de données issues des :
  - simulations numériques, en climat, fluides turbulents (combustion, fusion, astrophysique)…;
  - grands instruments, i.e., LHC, ITER, LSST, LOFAR, plateformes génomiques … ;
  - grands systèmes d'observation au sol, i.e., sismologie et géodésie (RESIF) … et dans l'espace (Euclid, WFIRST, GAIA, imagerie et interférométrie)…
- Compétitivité scientifique : adosser aux infrastructures et ressources informatiques un environnement d'expertise pluri et inter disciplinaire pour les valoriser et les exploiter (e.g. *USA, Japon, Allemagne, UK)*
- Calcul intensif pas uniquement guidé par les avancées technologiques

Big Data motivations (1)

- Accumulation de données issues des capteurs, communications, stockage pour business, science, gouvernements, société, ….
- Google, Yahoo!, Microsoft, … ont créé une nouvelle activité économique en récupérant des informations libres de droit sur le Web et en les présentant aux utilisateurs de façon exploitable
- Les moteurs de recherche ont transformé notre façon d'accéder à l'information

Big Data motivations (2) : sciences

Large Synoptic Survey Telescope (LSST):

- installé au Chili, enregistre 30 10^{12} octets d'images par jour soit 2 *Sloan Digital Sky Surveys* par jour (basé *sur un télescope de* 2.5m installé à Apache Point Observatory, New Mexico,!
- Objectif : origines de l'univers

Large Hadron Collider (LHC):

- Accélérateur de particules pour comprendre la structure de l'univers
- Va engendrer 60 téraoctets de données par jour soit 15 pétaoctets annuellement

Evolutions

- Course vers *l'Exascale* au niveau international avec de multiples initiatives EU (PRACE, EESI), Japon (machine K), USA (DOE, NSF), Chine, ….: préparer les communautés scientifiques, développer nouvelles méthodes et outils face à ces grands défis
- L'Exascale est la prochaine frontière :
  - Exaflops (10^{18} opérations flottantes par seconde)
  - Exaoctets (10^{18} octets) et même Zetaoctets, volumes des données à analyser, exploiter/visualiser
- En France et en EU :
  - Pilotage tiré par les enjeux stratégiques et la nécessité de ne pas décrocher en terme de puissance installée
  - Il faut renforcer l'argumentation autour de défis scientifiques

La course aux performances

- US, EU, Chine, Japon et Russie ont tous annoncé qu'ils auraient des systèmes exascales vers 2020
- Objectifs en terme de performance aux USA :
  - 100 PFlops 2016-2017
  - 1 Exaflops en 2018-2020
- Contraintes
  - Utiliser des technologies sur étagère ou au moins viables commercialement
  - < 20 MW
- Calculateurs suffisamment généralistes

Actuellement top500 :

1. Titan, Cray XK7, DOE Oak Ridge : 17,59 Pflops avec 560,640 cœurs, 8,209 MW, peak 27,112 Pflops, Opteron + NVIDIA
2. Sequoia, BlueGene/Q, DOE, LLNL : 16,324 Pflops avec 1,572,864 cœurs, 7,890 MW, peak 20,132 Pflops
3. "K Computer" au Japon (10.51 Pétaflop/s sur le Linpack benchmark avec 705,024 SPARC64 cœurs, 12,7 MW)
4. Mais USA > 50% des systèmes installés

Caractéristiques des systèmes

- Consommation actuelle : 130 W par chip
- Bande passante mémoire augmente moins vite due fréquence CPU
- #coeurs :
  - 10^4 en 2000 => proche de 10^7en 2012
- #sockets :
  - 10^3 vers 2005 => 10^5 en 2012
- 100 millions de threads sur un système exaflops
- Avec la technologie actuelle 475 MW pour un LINPACK à l'exaflops mais uniquement 2% énergie pour le calcul :
  - 50% accès stockage données et déplacement des données
  - 50% : cache, mémoire virtuelle, …

Conséquences

- Evolution technologique + consommation énergétique croissante + évolution requise des codes + compétences + support + impact sur l'organisation de la recherche => ***augmentation des coûts***
- Calcul intensif: pas uniquement problème de ressources mais un ***changement de paradigme*** dans la recherche scientifique
- Nouvelle ***approche holistique*** autour des défis scientifiques:
  - Inter/pluridisciplinarité et plus d'interactions entre informatique, mathématiques et autres disciplines,
  - Nouvelles méthodes et algorithmes: défis logiciels
  - Infrastructures de calcul et de données *en synergie* avec celles des grands instruments, plateformes expérimentales et des systèmes d'observation

Financements côté TGIR CNRS

- Budget TGIR 2012 environ 140 M€
- Budget calcul de l'ordre de 20 M€ pour IDRIS, GENCI, CC IN2P3, RENATER, IDGC / France GRILLES
- *Est-ce suffisant eut égard aux enjeux scientifiques ?*
- *Situation assez similaire au niveau national*
- Personnel: ~ 30-40% du budget d'un centre
- Engagements nationaux et internationaux pour des centres d'envergure comme le CC IN2P3 ou l'IDRIS
- Aspects données / stockage pas encore suffisamment pris en compte (e.g. Tiers-1)
- Problématique du financement de l'évolution des infrastructures des centres

Ecosystème riche au sein du CNRS

- Centres de calcul d'envergure nationale :
  - ***CC IN2P3*** Tier-1 WLCG : traitement et analyse des données issues des grands instruments en physique des hautes énergies (CERN), ouverture vers biologie, Astronomie & Astrophysique (données du LSST). Compétence / expertise reconnues en technologies Grilles et Cloud
  - ***IDRIS*** Tier-1 GENCI: HPC et analyse des masses de données produites par les simulations. Expertise, qualité de support aux projets pluridisciplinaires et formation, uniques et reconnues internationalement.
- Institut des Grilles et du Cloud (coordonne France Grilles), Maison de la Simulation, Mission pour l'interdisciplinarité (MASTODONS), lien avec les méso-centres …
- Autres éléments d'envergure nationale du CNRS : Observatoires des Sciences de l'Univers (OSU), les grandes plateformes génomiques …
- Besoin d'une d'une stratégie globale autour de ces initiatives

Vision du calcul intensif au CNRS

- Calcul intensif: un *grand instrument scientifique pluridisciplinaire*, catalyseur de nouvelles connaissances scientifiques

- Exploitation et valorisation scientifique dépendent de la capacité à:
  - l'insérer au sein des pôles de recherche et d'expertise associant recherche informatique, analyse numérique aux autres disciplines autour de grands enjeux scientifiques ;
  - évoluer en phase avec les pratiques de la recherche ;
  - Promouvoir l'appropriation du calcul intensif par les communautés encore trop peu utilisatrices des moyens nationaux et européens de calcul intensif, alors qu'elles pourraient en tirer profit pour aborder des sujets ambitieux nouveaux, au meilleur niveau mondial

Avec une stratégie globale

- Structurante ambitieuse, holistique et incitative.
- Meilleure coordination des initiatives locales
- *Qui a un coût* !!
- Un plan financier programmé dans le temps à l'échelle des enjeux scientifiques et des efforts de structuration nécessaires
- Spécificité pluri/interdisciplinaire du calcul intensif: modèle de financement différent des grands instruments disciplinaires
- Aussi bien en interne CNRS que au niveau national

Recommandations

- Mieux articuler les initiatives CNRS
- Dépasser modèle actuel de pyramide de ressources calcul et données
- Structurer les besoins en calcul, traitement de données, … autour de grands défis scientifiques: en leur associant infrastructures de calcul et de données, des équipes de recherche inter/pluridisciplinaires.
  - ➢ Intégrés aux grands instruments, plateformes expérimentales et systèmes d'observation
  - ➢ Renforcer les synergies avec les acteurs du domaine CEA, INRIA, GENCI, et pôles de recherche et d'enseignement universitaires (méso-centres, LABEX, IRT, …)
- Reconnaissance des chercheurs et ingénieurs, à l'interface domaine disciplinaires / maths / informatique / ….

Eléments stratégiques

- Aller vers une stratégie holistique du calcul intensif au CNRS incitative, structurante et dimensionnante
  - Renforcer synergies entre les diverses initiatives au sein du CNRS mais plus largement au niveau national
  - Prendre en compte les enjeux pluri / interdisciplinaires des défis scientifiques du calcul intensif et du *data-intensive*
  - Développer des synergies avec grands instruments, plateformes expérimentales nationales, systèmes d'observation et infrastructures nationales de données (e.g. Observatoires des Sciences de l'Univers, …).
- Réfléchir à l'émergence de pôles de Recherche Scientifique et Technologique en Calcul et Données (RSTCD) agrégeant infrastructures calcul et données, équipes de recherche autour de grands défis scientifiques, s'appuyant sur le contexte local

Enquête sur l'informatique en appui à la recherche

- Menée par le CCIS rattaché au COCIN : *Comité de Coordination et de pilotage de l'Informatique en Soutien à la recherche rattaché aux besoins scientifiques au CNRS*
- *Informatique scientifique* : calcul numérique, développement scientifique, Base de données, IHM et/ou interface web scientifique, traitements et/ou analyses de données, traitements et/ou analyses statistiques, traitements et/ou valorisations de données d'observation, pérennisation des données scientifiques, instrumentation, acquisition de données et/ou de signal, architecture et infrastructure pour le calcul haute performance, etc.
- *Besoins / enjeux / défis liés à l'IS dans l'ensemble des instituts du CNRS*
  - Calcul / simulation / modélisation mais surtout énorme enjeu autour des données signalé dans la plupart des instituts
  - Ressources humaines + compétences

Enquête CCIS (suite)

- Risques les plus cités :
  - L'arrêt ou ralentissement des projets + perte de qualité scientifique
  - La perte de compétitivité internationale
  - La perte de données scientifiques
- Points récurrents mentionnés :
  - Perte de compétitivité internationale si l'IS ne suit pas les évolutions des besoins (baisse des publications, baisse pertinence et exhaustivité des résultats scientifiques, décalage par rapport aux avancées industrielles, complique la participation à de grands projets internationaux,
  - Préoccupations autour de la valorisation des données qui constituent un véritable patrimoine (données expérimentales ou issues de grands instruments ou de simulations, ethnographiques, études cliniques, expériences métrologiques, analyses NanoSims, vielles revues, …) càd utilisation, accessibilité, conservation et exploitation, rapatriement, bases de données, maintenance des logiciels, archivage (pérenne ou non)
  - Inquiétude sur un fonctionnement reposant sur des CDD avec une compétence susceptible de disparaître
  - L'évolution des moyens de calcul requiert des compétences toujours plus importantes et le développement des codes devient lui aussi de plus en plus coûteux

Conclusion

- Le CNRS a confortésa vision de l'écosystème des calculs et des données
- Il s'est doté d'une stratégie qui s'appuie sur des grands défis scientifiques (pôles RSTCD) qui doivent émerger d'une concertation avec tous les acteurs d'un site : universités, écoles, IDEX, LABEX, IRT, organismes de recherche CEA, INRIA, …), méso-centres, GENCI, ….