www.cnrs.fr

Bilan des activités du COCIN : informatique scientifique et évolution des infrastructures de calcul et de données

M. Bidoit

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

Directeur CNRS / INS2I

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 + Directeur de l'IDRIS + 4 ingénieurs experts
- Le président et directeur désignés par le Président du CNRS

Missions et production du COCIN

Créé en Décembre 2010 : réflexion collective sur les besoins, la structuration et les évolutions de l'écosystème du calcul intensif et des données scientifiques au CNRS

- Composition : 1 représentant / institut + DU IDRIS + 4 experts

Michel Bidoit (INS2I) : Président	S. Bosi (INSHS)	**M. Daydé (INS2I) : Directeur**
D. Girou (IDRIS)	F. Godeferd (INSIS)	Ph. Helluy (INSMI)
S. Lamarre (INEE)	L. Lellouch (INP)	P.-E. Macchi (IN2P3)
JC. Michalski (INSB)	C. Pouchan (INC)	JP Vilotte (INSU)

Experts : D. Bascle (INC), F. Berthou (INP), M. Libes (INSU), V. Miele (INEE)

Invités : V. Breton (IDGC), O. Porte (DSI), JP Proux (GENCI)

- Production :

-Livre Blanc sur le Calcul Intensif au CNRS fin 2012

-Propositions pour une nouvelle stratégie du calcul et des données au CNRS (Décembre 2013)

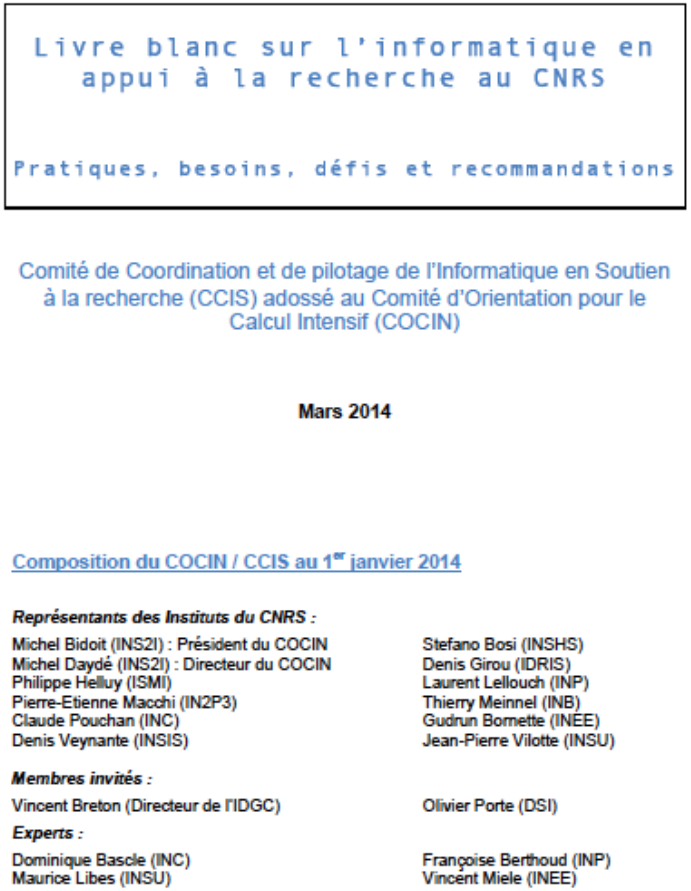
-Livre blanc sur l'Informatique en appui à la recherche (Mars 2014)

-Evolution des coûts des infrastructures pour le calcul intensif et le traitement des données à grande échelle (étude lancée 2014)

Enquête SUR L’Informatique Scientifique

Informatique en appui à la recherche au CNRS (IS)

- *Définition* : calcul scientifique ; développement logiciel ; bases de données ; interface homme-machine et interface web ; traitement, analyse et pérennisation des données scientifiques ; instrumentation et acquisition de données, architecture et infrastructure pour l'informatique scientifique
- Enquête nov. 2012 – janv. 2013 pour collecter données issues de l'ensemble des Instituts de Recherche du CNRS au travers de leurs unités
- Taux de retour environ 25 % des unités, satisfaisant mais panel des unités ± représentatif selon les instituts



Livre blanc sur l'informatique en appui à la recherche au CNRS

Pratiques, besoins, défis et recommandations

Comité de Coordination et de pilotage de l'Informatique en Soutien à la recherche (CCIS) adossé au Comité d'Orientation pour le Calcul Intensif (COCIN)

Mars 2014

Composition du COCIN / CCIS au 1er janvier 2014

Représentants des Instituts du CNRS :

Michel Bidoit (INS2I) : Président du COCIN
Michel Daydé (INS2I) : Directeur du COCIN
Philippe Helluy (ISMI)
Pierre-Etienne Macchi (IN2P3)
Claude Pouchan (INC)
Denis Veynante (INSIS)
Stefano Bosi (INSHS)
Denis Girou (IDRIS)
Laurent Lellouch (INP)
Thierry Meinnel (INB)
Gudrun Bornette (INEE)
Jean-Pierre Vilotte (INSU)

Membres invités :

Vincent Breton (Directeur de l'IDGC)
Olivier Porte (DSI)

Experts :

Dominique Bascle (INC)
Maurice Libes (INSU)
Françoise Berthoud (INP)
Vincent Miele (INEE)

Synthèse

- IS essentielle : lien direct avec la qualité de la production scientifique
- Communautés scientifiques du CNRS très impliquées dans tout ou partie des spécialités en IS
- Vision partagée au sein des instituts qu'ils aient une forte tradition en IS (INSU, IN2P3...) ou des besoins plus récents (INEE, INSHS...)
- Personnel technique est au cœur de l'IS
- L'implication des chercheurs inversement corrélée au niveau de soutien en ITA / BIATSS
- Pratiques en IS très différentes selon les instituts e.g. instituts avec fortes compétences en calcul intensif ou données

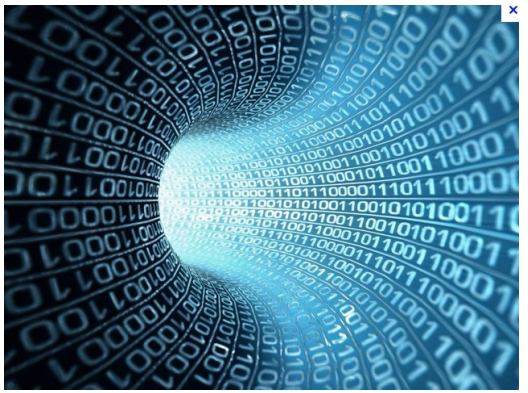
Constat majeur : manque de compétences et/ou de personnels sur les spécialités liées au traitement des données au sens large

Recommandations

- CNRS doit faire face aux défis liés au référencement, au traitement et l'analyse des **données** : **renforcer** et **coordonner les moyens à allouer** et la **diffusion des compétences** existantes à ce jour dans de nombreuses équipes.
- Favoriser le **partage et le transfert de compétences** IS entre instituts : la **formation** doit être l'élément central et récurrent
- Veiller à la réponse aux besoins exprimés, particulièrement dans les secteurs où la demande émerge : **maintien ou création de postes en IS** et explorer des approches complémentaires (**échanges de compétences inter-instituts, …)**
- IS doit être reconnue comme un véritable **pôle stratégique** au sein du CNRS
- Le CNRS doit conduire une réflexion autour de l'évolution des besoins, des métiers et des réponses à apporter en associant les divers acteurs (DSI, DIST, experts métiers) et les instituts dans paysage multi-organismes.

EVOLUTION DES Infrastructures de calcul et de données

Données scientifiques & HPC : des enjeux stratégiques



- Modélisation et simulation : 3ème pilier de la science après la théorie et l'expérimentation
- L'exploitation des données (« Big Data ») est maintenant considérée comme le 4ème pilier de la science

- 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 / analyse de données

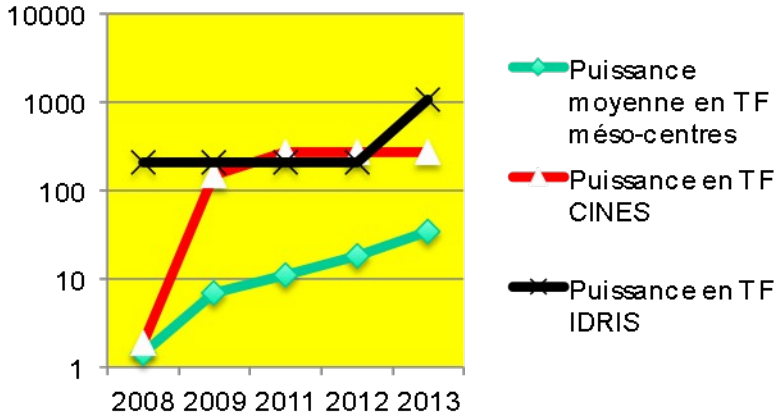
Large Synoptic Survey Telescope (LSST): Chili, 30 10^{12} octets d'images / jour (*télescope* 2.5m, Apache Point Observatory, New Mexico)

Large Hadron Collider (LHC): 60 To / jour soit 15 Po / an

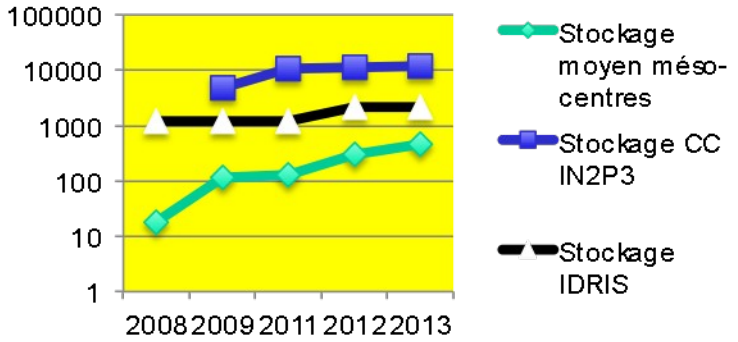
- Ne plus dissocier HPC de l'analyse et valorisation des masses de données issues des simulations numériques (climat, fluides turbulents,…), grands instruments (, LHC, ITER, LSST, LOFAR, plateformes génomiques …) et grands systèmes d'observation au sol (i.e., sismologie et géodésie : RESIF) et dans l'espace (Euclid, WFIRST, GAIA, imagerie et interférométrie)…
- *Calcul intensif pas uniquement problème de ressources mais un* ***changement de paradigme*** *dans la recherche scientifique :*
  - Plus d'inter/pluridisciplinarité (informatique, maths et autres disciplines),
  - Vision holistique des Infrastructures calcul / données / grands instruments / plateformes expérimentales / systèmes d'observation

Evolution des Tiers-1 et Tiers-2 2009-2014 (données Groupe Calcul et EcoInfo)

- Performances de calcul multipliées par **30**,
- Volumes de stockage multipliés par **25**,
- Facture électrique augmentation d'un facteur allant de **1,25** au CC IN2P3 à **1,5** à l'IDRIS,
- **Avec un coût d'investissement stable** (qui n'a pas toujours permis de satisfaire la demande des utilisateurs) **et un volume de maintenance informatique et d'ETP décroissant.**
- **ETPT** entre ½ **et** ¾ des dépenses de fonctionnement
- Estimation consommation électrique : cœur Intel environ 21W et 1 Teraoctets 18W



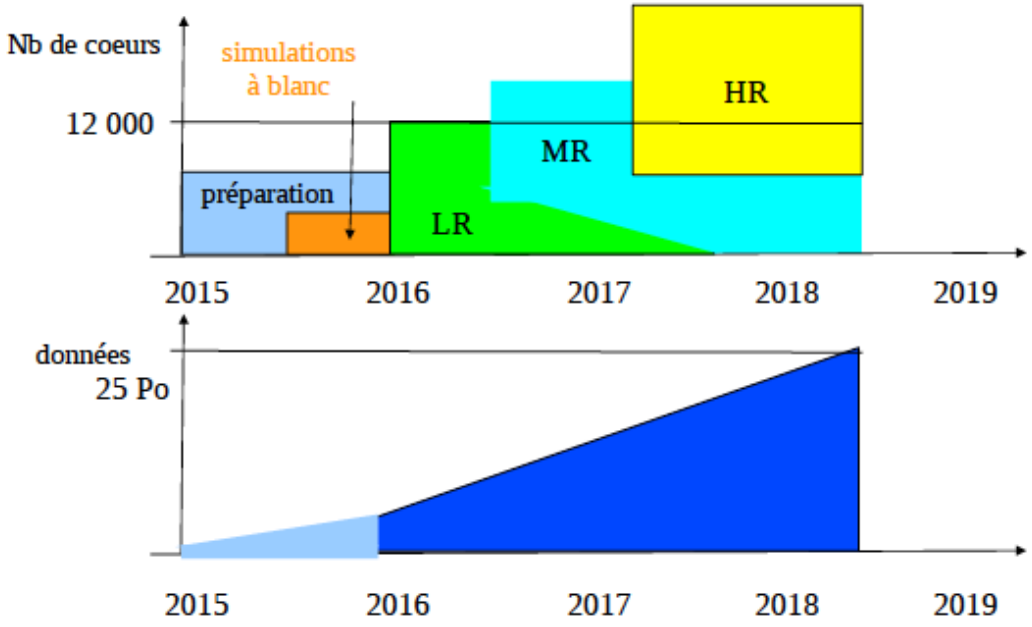
Evolution des performances en TFlops (TF) de deux Tiers-1 (CINES et IDRIS) et en moyenne sur l'ensemble des méso-centres



Evolution des volumes de stockage sur disques en TOctets (TB) sur le CC IN2P3 et IDRIS et en moyenne sur l'ensemble des méso-centres

Infrastructures de données

- Communautés scientifiques avec des besoins / compétences bien établies :
  - Sciences de l'univers (OSU, observatoires virtuels, ….)
  - Physique des hautes énergies (grille WLCG)
  - Biologie (RENABI, France Génomique, IFB)
  - …
- Reste des besoins immenses plus ou moins émergents et un besoin de structuration nationale et au niveau des sites



Demandes IPSL pour CMIP6 2015-2018 (JL Dufresne) : 100 millions d'heures pendant 3 ans (12,000 cœurs par an), 25 Po stockage

Conséquences : coûts croissants

- Calcul intensif :
  - Coordination au moins au sommet de la pyramide du calcul (Tiers-0 / Tiers-1 / Tiers-2, EQUIP@MESO,…)
  - Evolution technologique + consommation énergétique croissante + adaptation des codes + compétences + support
- Données :
  - Explosion des besoins et des demandes non coordonnées (CPER) même si certaines communautés sont structurées (e.g. physique des hautes énergies, sciences de l'univers, bio, …)
  - Conforter compétences + support
- Impact sur l'organisation de la recherche

Rationaliser déploiement des infrastructures / coordonner les demandes

Stratégie nationale / de site autour de défis scientifiques et maîtrise des coûts !!!

Maîtrise des coûts (suite)

- Autres pistes permettant de maîtriser les coûts suggérées par le Groupe ***EcoInfo*** :
  - **Maîtriser les coûts électriques** en installant une métrologie systématique
  - **Eviter les installations surdimensionnées** et opter pour de la modularité
  - Innover en terme de systèmes de froid
  - Contraindre à l'achat de matériel **efficient énergétiquement,** ce qui implique pour avoir des gains significatifs **mutualisation des infrastructures en particulier au niveau des Tiers-3** : e.g. économie de 60% sur le coût électrique avec un hébergement plus efficace (réduction du PUE de 2 à 1,3) qui induit économie de **500 K€ / an** pour les 2000 serveurs de calcul et de données achetés sur les marchés CNRS entre 2003 et 2013.
- Etude mutualisation du stockage à l'Université Grenoble-I :
  - Recensement de 44 locaux climatisés
  - Coût fluides estimé à 700 K€ (1/3 consommation énergétique de l'établissement)
  - Besoins de stockage croissant de 1,2 Po en 2012 à 3,4 Po en 2018 (sans compter le stockage nécessaire au calcul intensif).

Conclusion

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

- Besoins calcul / stockage **en forte croissance** d'où l'**explosion des demandes au niveau du CPER**

- F**acteur majeur** de la dérive des coûts informatiques : **foisonnement d'infrastructures de calcul et de données au niveau local** (i.e. Tiers-3) aggravé par **l'augmentation des demandes non-coordonnées + morcellement** et de **désorganisation** des **infrastructures de données**

- *Stratégie du CNRS : coordination / rationalisation des investissements aux niveaux site / national autour de défis scientifiques avec l'ensemble des acteurs concernés*