Compte-rendu de la réunion C4 du 07 Décembre 2018 au CINES

Participants C4 :	Participants CINES :
- Sébastien Theetten (CT1) : Excusé	- Boris Dintrans (directeur du CINES)
- Julien Bodart (CT2A) : Visio	- Eric Boyer (responsable DCI depuis 01/09/2018)
- Florent Duchaine (CT2B) : Visio	- Jean-Christophe Penalva (DCI)
- Alain Miniussi (CT4) : Excusé	- Mathieu Cloirec (DCI)
- Virginie Grandgirard (CT5) : Visio	- Romain Couturat (DCI)
- Rachel Nuter (CT5) : Visio	- Philippe Prat (DCI)
- Michel Kern (CT6) : Visio	- Gaétan Péris (DCI)
- Nicolas Floquet (CT7) : Présent	- Hervé Toureille (DS2I)
- Sébastien Le Roux (CT8) : Excusé	- Emilie Boulard (DCI)
	- Nicole Audiffren (DCI)
Participants GENCI :	- Jérôme Chapelle (DS2I)
- Jean-Philippe Proux	- Marie Galez (DS2I)
- Delphine Theodorou	- Gérard Vernou (DS2I)

Ordre du jour :

- Gestion automatisée du /scratch (Emilie Boulard)
- Présentation et premiers retours sur la solution Visu / Prépost (Jean-Christophe Penalva et Mathieu Cloirec)
- Architecture Datacentrique du CINES (Hervé Toureille)
- Projet énergie/métrologie sur O2 (Eric Boyer et Jérôme Chapelle)

Sujet de discussion :

- Nouveaux usages, services et accès (Jupyter Notebook, intégration continue, checkpoint restart)
- Sécurité : Mise en service de SELinux
- Quels besoins en formation et accompagnement ?

Comme chaque fois la journée a débuté par une discussion d'une heure entre membres du C4 où nous avons fait le point sur les questions et commentaires des utilisateurs. La journée s'est ensuite poursuivie par une réunion C4/CINES/GENCI de 11h00 à 16h00. Vous trouverez ci-dessous un bref compte-rendu des points qu'il nous semble importants de souligner. Pour plus de détails n'hésitez pas à consulter les transparents ci-joints présentés par le CINES et GENCI lors de la réunion.

Points marquants GENCI (présentation Jean-Philippe Proux et Delphine Theodorou)**:**

1) **Actualités sur centres des calculs nationaux**
   - ✓ **TGCC** : La machine Joliot-Curie est en production depuis cet été. Elle est équipée dans cette première phase de nœuds KNL et Skylake pour une puissance crête de 9 Pflops/s. Elle devrait être étendue à 20 Pflops/s en 2020. L'appel d'offre est en cours pour cette extension. Mais attention il devra s'agir d'une architecture innovante pour pouvoir prétendre à un financement à hauteur de 35% pour l'acquisition et le coût de la première année de fonctionnement supporté par l'Europe.

- ✓ **Idris** : Le renouvellement de la machine Turing est engagé. La nouvelle machine, qui dans une première phase aura une puissance crête de 14Pflops/s, devrait être accessible aux utilisateurs à l'été 2019. Il s'agira d'un mix de technologies Intel « Cascade Lake » et Nvidia GPU Volta. Ce choix technologique s'inscrit dans la volonté annoncée par le plan Intelligence Artificielle national d'offrir à la communauté IA des moyens de calcul nationaux conséquents. Cette nouvelle machine sera la plus importante machine convergée (HPC/IA) en Europe.

2) Actualités cellule de veille technologique

Pour rappel, cette cellule de veille technologique coordonnée par GENCI a pour objectif d'anticiper l'arrivée des futures machines (pré)exascales. A l'heure actuelle 2 prototypes sont disponibles : (i) la plateforme Frioul au CINES à base de nœuds KNL et (ii) la plateforme Ouessant à l'IDRIS à base de nœuds OpenPower où une pile logicielle IA est disponible. Une nouvelle plateforme à base de lames ARM est en cours d'installation au TGCC. 10 lames devraient être disponibles d'ici la fin de l'année. Deux plateformes (Frioul et Ouessant) sont accessibles via une demande DARI d'accès préparatoire (www.edari.fr) le troisième prototype sera ouvert à tous ultérieurement.

3) Plan Intelligence Artificielle

- ✓ Le plan IA s'inscrit dans une stratégie nationale voulue par le président de la république de propulser la France parmi les champions mondiaux de l'Intelligence Artificielle. Pour plus d'info vous pouvez consulter le document disponible à l'adresse http://www.genci.fr/fr/node/965.
- ✓ Dans ce contexte, GENCI porte la responsabilité d'acheter un supercalculateur dédié pour faciliter l'accès au calcul pour l'ensemble de cette communauté de recherche. Comme dit précédemment il s'agira de la prochaine machine IDRIS.
- ✓ Avec l'arrivée de cette nouvelle machine la puissance GPU actuellement disponible en France sera doublée. Le choix d'une grande partition GPU a pour premier but de palier au besoin de la communauté IA. Le second objectif est d'aider les utilisateurs HPC à porter leurs codes sur GPU. Notamment les codes de chimie qui pourraient être très adaptés. Dans ce cadre, une partie du budget est dédiée au portage de 6 applications phares avec 6 ingénieurs dédiés.
- ✓ Pour répondre aux besoins des chercheurs en IA et être capable de rivaliser avec des acteurs tels que Google, le fonctionnement des attributions d'heures sur cette future machine sera en partie différent du fonctionnement DARI standard :
    - ✓ Il y aura 1 partie DARI (dédiée au HPC standard) et 1 partie dynamique (dédiée à l'IA). La frontière entre les 2 sera adaptable si les ressources sont non utilisées côté IA ou HPC. Dans le cas de contraintes fortes des 2 côtés la répartition se fera à 50/50.
    - ✓ L'accès à la partie dynamique (uniquement GPU) sera possible via un dossier plus léger qu'un dossier DARI. L'appel à projet sera ouvert en permanence et sera soumis à l'accord du directeur de centre de l'IDRIS, avec pour les dossiers trop volumineux en nombre d'heures la possibilité de faire appel à des experts. Le projet aura une durée d'un an.
    - ✓ Pour pouvoir débuter les projets IA dès acceptation, une procédure de pré-enregistrement est prévue afin d'anticiper la partie sécurité. Cette procédure sera en essai pendant 2 ans et pourrait se généraliser à l'ensemble du DARI si elle s'avère efficace et si les ressources le permettent.

4) Bilan des ressources GENCI

- ✓ Concernant l'appel A5 (Nov 2018-Oct 2019), la trop forte demande en nombre d'heures sur Joliot-Curie a nécessité beaucoup de travail au niveau des CTs pour transférer des demandes utilisateurs vers les autres calculateurs. Il faut éviter que cela se reproduise pour l'appel A6.
- ✓ L'appel à projets A6 aura lieu du 8 janvier au 7 février 2019 pour un démarrage des projets le 2 mai 2019.
- ✓ Pour info un séminaire sur les besoins d'accompagnement à court et moyen terme des communautés utilisatrices sera organisé par GENCI le 17 janvier 2019. Il devrait regrouper les présidents de CT, les responsables de centres et des supports utilisateurs ainsi que les représentants des 3 COMUT.

5) Infrastructure européenne (PRACE)

- ✓ En 2018, les calculateurs PRACE ont offert une puissance de 70 à 110 Pflops/s.
- ✓ Dans PRACE1, les scientifiques français étaient les premiers en nombre de projets et en présence dans les projets et les deuxièmes bénéficiaires en nombre d'heures de calcul avec 2 milliards d'heures (soit 20% des heures totales proposées).
- ✓ Dans PRACE2 (calls 14 à 17), 38 projets français ont obtenu 1.3 milliard d'heures. Il est important de signaler que 70% des projets français proposés ont été retenus.
- ✓ Pour rappel, il y a 2 appels à projets PRACE par an (www.prace-ri.eu).

Points marquants CINES :

1) Gestion automatisée du /scratch (présentation d'Emilie Boulard)

- ✓ Comme déjà discuté lors du dernier C4 le CINES envisage de mettre en place une politique d'effacement systématique des fichiers non accédés depuis plus de 60 jours. A l'heure actuelle ces fichiers représentent 2.6 Poctets, soit la moitié de la capacité du SCRATCH.
- ✓ Pour rappel, une surcharge du SCRATCH peut avoir un impact néfaste pour l'ensemble des utilisateurs. De plus, les disques de stockage des futures machines seront basés sur de la mémoire SSD donc à accès plus rapide mais beaucoup plus chère. Pour cause de coût, le stockage sera donc réduit par rapport aux calculateurs actuels. Par exemple, le CINES envisage pour sa future machine un SCRATCH de l'ordre de 1 Poctet (soit 5 fois moins qu'actuellement). Une politique de meilleure gestion des fichiers est donc essentielle.
- ✓ Cette politique devrait être mise en œuvre au CINES au premier trimestre 2019. Par contre cette gestion automatique ne sera déployée qu'à partir du moment où le CINES aura finalisé le développement d'outils pour : (i) visualiser l'état du SCRATCH et les fichiers susceptibles d'être automatiquement supprimés et (ii) aider au transfert des données entre le SCRATCH et le STORE.

  Le C4 insiste sur le fait que ces outils sont essentiels avant le déploiement de la suppression automatique des fichiers.

  C4: Sera-t-il possible d'épargner la suppression complète d'un répertoire à partir du moment où plusieurs fichiers de ce répertoire ont été accédés ?

  CINES : Non, il n'y aura pas possibilité d'un tel raffinement.

  C4: A-t-on à l'heure actuelle intérêt de travailler sur le STORE ?

CINES : La bande passante du STORE étant de 50Go/s, il est déconseillé de travailler directement sur le STORE. Cette zone doit rester une zone de stockage. Il manque actuellement un WORKDIR. Cette zone sera présente lors de l'arrivée du nouveau stockage (plus de 5M€ d'investissements sont prévus pour 2019 à cet effet).

Question utilisateur : Y aura-t-il une commande particulière (type mfret sur Ergon à l'IDRIS) pour faire savoir au système qu'on veut prolonger l'existence de certains fichiers sur le scratch?
CINES : La commande 'mfret' est disponible à l'IDRIS pour agir sur du stockage pérenne, et n'est pas disponible sur les espaces de manœuvre type /scratch sur TURING et ADA, où la suppression de fichiers après la fin des travaux est quasi immédiate (et devrait être portée à quelques jours). Cette commande n'est pas envisagée pour la gestion des fichiers sur le scratch.

2) **Présentation et premiers retours sur la solution Visu** (présentation de Jean-Christophe Penalva et Mathieu Cloirec)
   - ✓ Les 4 nœuds de visualisation tant attendus sont enfin opérationnels.
   - ✓ Vizalloc : est un sbatch qui facilite beaucoup de choses à l'utilisateur.

   Le C4 fait part du retour de plusieurs utilisateurs très satisfaits de la solution proposée ainsi que du support apporté pour la mise en œuvre.

   Question utilisateur : La possibilité de session interactive au CINES était proposée à partir de 2 méthodes ; session interactives VNC ou NoMachine. On a surtout testé la visu à partir d'un client VNC. Il semble que la méthode NoMachine ne soit plus possible; elle ne figure pas dans les pages WEB du cines. Est-ce-que l'option NoMachine a été abandonnée, si oui pourquoi ?
   CINES : NoMachine n'a pas été abandonné mais ne fonctionne pas pour le moment. Normalement à terme, il y a espoir que cette solution marche mieux que VNC.

   Question utilisateur : L'un des problèmes rencontrés a été le fait que la session de Visu se termine brutalement à l'issue de l'allocation, sans possibilité de la prolonger. Est-ce-qu'une solution est envisagée ?
   CINES : Grâce à vizalloc on peut demander des sessions allant jusqu'à 6h00. Par contre pour le moment pas de solution trouvée pour prolonger une session en cours.
   C4 : Est-ce qu'il ne serait pas possible de rajouter au moins un message d'alerte comme au TGCC pour signaler que la session va bientôt terminer ?
   CINES : Si, on peut tout à fait regarder une première solution dans ce sens, par exemple par une ouverture d'une boite de dialogue.

3) **Pré/Post-traitement**
   - ✓ 1 nœud large mémoire (3TB de RAM accessibles en NUMA) a été ajouté à Occigen (accessible via ses frontaux).
   - ✓ Ce nœud est encore actuellement en test car il reste des problèmes d'allocation SLURM. Il faut pour le moment utiliser l'intégralité du nœud soit les 224 cœurs. Ce nœud devrait être à terme partageable mais il faut débloquer un problème d'allocation des GPUs avec SLURM.

- ✓ Un formulaire d'information est disponible (https://www.cines.fr/form-mail-svp). Le plus simple est que les personnes intéressées contactent directement SVP pour discuter de l'utilisation optimale de ce nœud.

4) **Architecture Datacentrique du CINES : Renouvellement, nouvelles technologies et nouveaux usages, évolution de l'architecture, retour sur le questionnaire** (présentation d'Hervé Toureille).
   - ✓ Un groupe de travail existe au niveau des centres pour réfléchir à :
     - ✓ L'uniformisation des stockages (nom, politique, …).
     - ✓ Un login unique : Au TGCC il existe déjà un login unique associé à plusieurs projets. Le CINES envisage cette solution pour la prochaine machine. Le transfert des données entre tiers 1 : amélioration du CCFR.
   - ✓ Résultats de l'enquête sur les besoins en stockage et services de diffusion des données réalisée auprès des utilisateurs du CINES
     - ✓ Le directeur du CINES tient à remercier le C4 pour son relais auprès des utilisateurs et tous les utilisateurs ayant répondu à l'enquête (121 projets sur 428 ont répondu au questionnaire). Vous trouverez plus de détails sur cette volumétrie en consultant les transparents présentés par le CINES joints. Cette enquête va permettre de mieux dimensionner l'achat des futurs moyens de stockage et de mieux cibler les besoins utilisateurs. Le CINES prévoit de se doter d'un espace de stockage performant WORKDIR de 10 Poctets.
     - ✓ Concernant la question sur « Avez-vous un Data Management Plan (DMP)?» seulement 6 sur 121 réponses déclarent que oui. Pour rappel un DMP est un outil pour encadrer les données produites, la manière de les diffuser et celle de les conserver. Il faudra très certainement bientôt présenter un DMP lors de la soumission d'une demande d'heures DARI pour tous les gros projets. Avec ces compétences en archivage pérenne le CINES est tout à fait prêt à proposer des formations sur les DMP.
     - ✓ Seulement 16.5% des utilisateurs souhaitent utiliser des technologies d'accès aux données de type cloud.
     - ✓ Moins de 10% utilisent les technologies BIG DATA, HPDA ou IA. Par contre plus de 60% souhaitent être informés ou formés à ces techniques. Ce type de formation est tout à fait envisageable au CINES en faisant attention de se coordonner avec les autres formations déjà existantes via PATC (PRACE Advanced Training Center).

5) **Projet énergie/métrologie sur O2** (Eric Boyer et Jérôme Chapelle)
   - ✓ Le CINES a hébergé un système pilote ATOS dans la dernière phase du projet PRACE-PCP depuis 2017 qui a permis le développement d'un monitoring fin des coûts énergétiques de l'ensemble de la chaine (réseau, calculs, management, refroidissement, …).
   - ✓ Dans un contexte où la facture électrique du CINES (actuellement de 1.5M€) augmentera de 30% l'an prochain, cette tendance générale impose que l'on se pose réellement la question à tous les niveaux de solutions pour réduire les coûts énergétiques :
     - ✓ Concernant la partie calcul, chaque job a un profil énergétique particulier et il faut travailler aux développements d'algorithmes numériques énergétiquement les moins coûteux possibles.
     - ✓ Il faut garder en tête que transporter des données coûte cher.

- ✓ La course au passage à l'échelle doit être raisonnée.
- ✓ Le pilotage des éléments du calculateurs doivent être étudiés pour dégager des économies d'énergie avec impact raisonnable sur les performances.

6) Réponses aux questions utilisateurs restantes

a) *Concernant les heures non consommées et qui ne le seront pas d'ici la fin de l'allocation : faut-il faire une démarche auprès du CINES (ou GENCI) pour les restituer ?*

GENCI : Il s'agit d'une information intéressante pour les centres car elle peut permettre aux directeurs d'avoir plus de flexibilité pour les demandes au fil de l'eau. La possibilité de restituer les heures via le DARI est à l'étude. Ceci permettrait à l'utilisateur d'être moins pénalisé lors de l'expertise pour le renouvellement d'un projet.

CINES: A l'heure actuelle un mail à SVP peut permettre de restituer les heures qui ne seront de consommées ou de pas faire l'objet de régulation mensuelle. Même si une anticipation de dérégulation de fin de période est plus difficile à gérer qu'en début.

b) *Le directeur du CINES avait promis en 2017 à la communauté CP2K et CPMD d'organiser une journée de présentations et discussions pour rétablir un lien de confiance avec le CINES et surtout permettre d'avancer ensemble avec plus de coordinations. Quels sont les retours de cette réunion ?*

CINES : Un workshop chimie devait initialement avoir lieu en mai. Il a été décalé en septembre pour des raisons de grèves SNCF. Les retours des participants sont très positifs. Des codes chimies ont montré de très bon passage à l'échelle. Les communautés CP2K et CPMD n'étaient pas les plus représentés. Concernant ces codes l'équipe support a travaillé et continue de travailler avec les utilisateurs. Des référents par logiciels ont été identifiés. Il semble que la plupart des utilisateurs aient trouvé une version stable pour tourner.

c) *J'utilise beaucoup l'interface web https://reser.cines.fr pour suivre l'évolution du projet que je gère et je regrette que dans la version actuelle on ne voit plus la courbe théorique du nombre d'heures qu'il faudrait avoir dépensé au temps t. Plus embêtant, l'affichage de l'utilisation du "store" sont maintenant "cachés" (il faut cliquer sur le nom de chaque membre du projet pour voir l'info)... c'était bien mieux avant quand tout était sous les yeux dans le tableau.*

CINES : L'interface a évoluée pour être compatible à une lecture sur tablette ou smartphone. L'affichage par défaut est donc maintenant plus minimaliste et des options d'affichage supplémentaires sont disponibles. Par contre le fait que la courbe théorique du nombre d'heures qu'il faudrait avoir consommées est une erreur. Cette erreur va être corrigée. Le CINES insiste sur le fait qu'il ne faut pas hésiter à envoyer un mail à SVP pour faire remonter ce type de problème et ne pas hésiter non plus à proposer des améliorations.

d) *Pourquoi avoir activé hyper-threading sur les nœuds de calcul ?*

CINES : L'hyper-threading est depuis toujours activé sur OCCIGEN. Lorsqu'il ne donne pas de bons résultats il s'agit souvent d'un mauvais placement sur les nœuds. Il existe une FAQ sur le binding (https://www.cines.fr/calcul/faq-calcul-intensif/ → jobs hybrides MIP-Open-MP) sur le site du CINES. Ne pas hésiter à contacter SVP pour se faire aider.

e) *Une utilisatrice a remarqué qu'un même run (même binaire, même configuration, même script de soumission) pouvait prendre un facteur 2 d'une instance à l'autre et n'a à ce jour reçu aucune explication à ce phénomène. Y compris lorsqu'elle a pu signaler un cas ou les deux job étaient en train de s'exécuter. Au-delà du problème de performance, cela l'oblige à prévoir une marge de sécurité d'un facteur 2.*
CINES : Ce n'est pas normal et il ne faut en aucun cas prévoir une telle marge de sécurité. Pas d'autres retours dans ce sens. Contacter SVP.

7) Divers

a) Représentants au C4 : Avec l'arrivée de l'IA à l'IDRIS la communauté du comité CT10 devrait s'accroitre. Elle n'a à l'heure actuelle pas de représentant au C4. Il serait important que cette communauté soit représentée. Pour rappel les membres du C4 sont élus pour 3 ans et ceux du C4 actuel le sont depuis décembre 2016. Un appel à candidature pour un titulaire et un suppléant va être lancé pour palier à ce manque avant les prochaines élections qui auront donc lieu en Décembre 2019. Nous lancerons également cet appel pour le CT3 qui n'a pas non plus de représentant actuellement.

b) Sécurité : Mise en service de SELinux qui permet une sécurisation au niveau du noyau LINUX. Cela permet notamment de référencer les utilisateurs ayant la possibilité de faire des appels systèmes. Ce service (déjà en place au TGCC) ne sera déployé que sur les nœuds de login et de visu car un coût CPU trop important pour être mis en œuvre sur les nœuds de calcul.
Attention, la mise en œuvre ne devrait pas avoir d'impact pour les utilisateurs mais merci de faire remonter à SVP si vous voyez apparaitre un message du type « permission denied ».

c) Nouveaux usages, services et accès (Jupyter Notebook, intégration continue, checkpoint restart) :
- Les jupyter Notebook sont de plus en plus utilisés. La possibilité d'offrir le service au CINES est à l'étude mais il reste d'importants problèmes de sécurité.
- Il existe également des problèmes de sécurité pour le déploiement d'outils d'intégration continue tels que JENKINS. Les équipes admin-support des 3 centres tiers 1 se réunissent une fois par trimestre. Il serait bien de mettre l'intégration continue à l'ordre du jour de la prochaine réunion (action CINES)

d) Les besoins en formation qui ressortent actuellement seraient les suivants: (i) Big Data, (ii) Intégration continue et (iii) visualisation 3D parallèle. Le CINES est preneur de toutes autres propositions.

Actions CINES à prévoir d'ici la prochaine réunion:
- Corriger le bug d'affichage sur https://reser.cines.fr
- Prévoir un mail pour appel à candidature pour titulaires et suppléants C4 pour les comités CT3 et CT10.
- Proposer l'intégration continue comme ordre du jour à la prochaine réunion inter-centres des équipes admin-support.

Prochaine réunion : le 22 mars 2019 (en présentielle au CINES)