Propositions pour l'architecture pour un cluster mutualisé entre CIMENT et Grid'5000

Pierre Neyron, Bruno Bzeznik, Lucas Nussbaum

Propositions pour l'architecture pour un cluster mutualisé entre CIMENT et Grid'5000

Pierre Neyron[1], Bruno Bzeznik[2], Lucas Nussbaum[3]
Janvier 2017

Contents

[1] pierre.neyron@imag.fr, LIG/CNRS, membre du comité d'architecte Grid'5000, responsable technique du site Grid'5000 de Grenoble, responsable technique de la communauté Informatique Distribuée de CIMENT
[2] bruno.bzeznik@univ-grenoble-alpes.fr, GRICAD/Université Grenoble Alpes, directeur technique de CIMENT, directeur du pôle Calcul de Gricad
[3] lucas.nussbaum@loria.fr, LORIA/Université de Lorraine, directeur technique de Grid'5000 et responsable scientifique du site Grid'5000 de Nancy

1 Introduction

Ce document est une étude pour la mise en exploitation d'une machine de type HPC-DA[4] opérée et utilisée à la fois dans la plateforme Grid'5000[5] et dans le mésocentre de l'Université Grenoble-Alpes: CIMENT [6]. Ce montage commun doit permettre d'optimiser les collaborations et donc les coûts (d'achat et d'opérations), comme cela est encouragé par les tutelles (CNRS, Inria, universités) notamment, tout en répondant aux besoins des utilisateurs. Transversal de par les communautés utilisatrices cibles (Grid'5000 et CIMENT), ce projet vise à répondre aux objectifs des différents acteurs, grâce d'une part aux solutions flexibles que les 2 plateformes développent depuis de nombreuses années, mais également par des compromis pertinents.

Les auteurs[1, 2, 3] de ce document sont membres de la communauté des chercheurs en informatique distribuée, et membres de l'équipe de pilotage technique d'une des plateformes ou des deux. Ils ont également une expertise forte sur la question des infrastructures HPC dites de production[7] et plus généralement sur les problématiques des Datacentre.

Tout d'abord ce projet à un objectif double pour la recherche dans le domaine informatique distribuée rassemblée autour de Grid'5000:

1. avoir un cluster dimensionnant pour la validation expérimentale à l'échelle de travaux de recherche sur les domaines HPC-DA[4]

2. avoir un cluster répondant aux besoins pour supporter un champ d'expérimentation large en informatique distribuée.

Pour 1, cela encourage le montage d'une cluster commun avec les communautés d'utilisateurs d'autres domaines scientifiques, qui ont des besoins de calcul qualifiés de "production" [7], et donc typiquement le mésocentre CIMENT.

Pour 2, cela se traduit par un degré de conformité aux spécifications d'un cluster Grid'5000 à maximiser, notamment pour permettre des expérimentations qui débordent du cadre HPC classique. Par exemple:

[4] High Performance Computing & Data Analytics
[5] `http://www.grid5000.fr`
[6] `http://ciment.ujf-grenoble.fr/`
[7] Production: on veut le résultat le plus rapidement possible, la méthode est secondaire ; Grid'5000: on travaille au contraire sur la méthode, les résultats éventuellement produits ne sont pas une fin en eux-mêmes.

1. déploiement de solutions de type Cloud Computing et Big Data pour l'étude de la convergence avec le HPC;
2. interconnexion de Grid'5000 pour des expériences d'envergure nationales et extra-nationales (avec d'autres testbeds en Europe: FIRE ou aux USA: Chameleon Cloud) et impliquant le site grenoblois;
3. reconfiguration, contrôle bas niveau et traçabilité de la plateforme fournis comme service aux utilisateurs.

1 et 2 font de la fonctionnalité de reconfiguration du réseau un élément clé (outil kavlan de Grid'5000). 3 nécessite une solutions très avancée de contrôle des informations sur la plate-forme (la référence API[8] de Grid'5000) en plus des outils d'instrumentation des systèmes (kadeploy notamment). Cette instrumentation implique un grande dynamicité de la configuration de la plateforme et donc des solutions pour en assurer la robustesse (g5k-check).

Pour CIMENT d'autre part, le premier objectif est bien-sûr l'accès à de nouvelles ressources de calcul dans le cadre de la collaboration historiques et transversale entre domaines scientifiques. Mais un second objectif important est que les fonctionnalités Grid'5000 mises en oeuvre pour les expérimentations sur le déploiement de solutions de types Cloud Computing ou Big-Data puissent également faire l'objet d'un transfert vers les usages de la communauté CIMENT. En effet, que cela soit le déploiement d'applicatifs HPC via des machines virtuelles ou des containers, ou la convergence entre calculs et données pour certaines applications, ce sont des besoins émergents. Ainsi la communauté CIMENT est intéressée par les fonctionnalités Grid'5000 en tant que telles, pour des travaux d'expérimentation sur les nouveaux outils utiles pour les traitements intensifs de calculs ou de données.

2 Contexte du projet

Historiquement, le site grenoblois à depuis de nombreuses années travaillé sur cette convergence entre les plateformes expérimentales pour l'informatiques distribuée et les plateformes HPC de production tous domaines scientifiques confondus. Les informaticiens impliqués dans CIMENT ont apportés une expertise et des outils (OAR, Cigri, Colmet) ainsi que des plateformes, certes parfois restées marginales ou exotiques, dans CIMENT. Ils sont également des architectes de la plateforme Grid'5000 depuis ses débuts.

2.1 Historique des convergences d'infrastructures entre Grid'5000 et CIMENT

2.1.1 Idpot (2004)

Cluster prototype de Grid'5000, mis a disposition de la communauté CIMENT également.

2.1.2 Icare (2006)

Cluster sous Solaris de l'OSUG, dont il était prévu qu'il soit intégrable à Grid'5000 grâce à kadeploy pour des expériences. Projet non abouti.

2.1.3 Digitalis 2048 (2008)

Cluster genepi, edel et adonis. Achat commun entre CIMENT et Grid'5000 (CPER/Inria). Entre 2008 et 2012, les clusters étaient accessibles par CIMENT via Cigri. Par ailleurs les utilisateurs CIMENT pouvaient accéder aux machines via leurs comptes CIMENT. Depuis Grid'5000 à cependant beaucoup évolué:

- Centralisation des configurations au niveau national (puppet)
- Instrumentation beaucoup plus évoluée (kavlan, reference API)
- Nouvelle gestion des comptes centralisée

En conséquence, le mécanisme simple de l'époque ne répond plus au besoin aujourd'hui.

[8] Inventaire de la plateforme consultable par programme, qui fournit les fonctionnalités d'introspection et de traçabilité et permet l'instrumentation et la vérification des composants. Cet outil est primordial pour des expériences scientifiques de qualité (reproductibilité, etc.)

2.1.4 Kinovis (2014)

Kinovis est une plateforme d'acquisition de 68 caméras connectées à un cluster de 17 machines (2x Intel Xeon 8 coeurs, Infiniband QDR). Les acquisitions vidéos n'utilisant le cluster que de l'ordre de 20% du temps, il a été question de permettre au cluster de basculer en mode CIMENT et/ou Grid'5000 (mais avec un fonctionnement diskless) le reste du temps. Un prototype à été développé mais malheureusement abandonné suite au départ de l'ingénieur Inria qui aurait du prendre en charge l'exploitation (formation et support utilisateur) et faute de remplaçant. Finalement, ce cluster ne fait partie ni de Grid'5000 ni de CIMENT aujourd'hui.

2.2 Site Grid'5000 actuel

Le site Grid'5000 grenoblois (plateforme digitalis: `http://digitalis.imag.fr`) actuel est composé:

- Des clusters de digitalis 2048:
  - genepi: 34 noeuds, 2x Intel Xeons 4 cores, Infiniband DDR (2008)
  - edel: 72 noeuds, 2x Intel Xeons 4 cores, Infiniband QDR (2010)
  - adonis: 10 noeuds, 2x Intel Xeon 4 cores + 2 GPU, Infiniband QDR (2010)
- De machines singulières:
  - idfreeze: machine quad-CPU AMD, 48 coeurs (2011)
  - idgraf: machine bi-CPU Intel + 8 GPU (2011)
  - idphix: machine équipée d'un Xeon Phi KNC (2013)
  - idbool: machine 12-CPU AMD Numascale 192 coeurs (2014)
  - idarm-1&2: 2 machines ARM64[9] (2015)
  - idkat: machines quad-CPU Intel 48 coeurs (2015)
  - idcin-1&2: 2 machines bi-CPU, 28 coeurs + 3 GPU (2015)
- D'un coeurs de réseau de 2005:
  - Extrem Network Aspen 8800
  - Connexion directe vers le reste de Grid'5000

L'ensemble est actuellement hébergé dans la salle des machines expérimentales (F212) du centre Inria Grenoble Rhône-Alpes (Montbonnot).

Le montage du nouveau cluster Grid'5000+CIMENT sera l'occasion de déménager le site Grid'5000 sur le campus (conformément à la stratégie générale d'hébergement des infrastructures informatiques académiques). Pour ce faire un nouveau routeur de site Grid'5000 doit être installé et la fibre Renater dédiée pour Grid'5000 devra arriver dans la salle IMAG du datacentre (un lien 10GE fibre IMAG - Inria devra être disponible pour la période de migration, il est déjà opérationnel entre Inria et SIMSU). Il est probable que l'ensemble des machines actuelles ne sera pas déménagé sur le campus: certaines machines pourront être arrêtées mais les machines singulières les plus récentes et/ou pertinentes pour la recherche seront déménagées afin de permettre une fermeture de la salle Inria à terme notamment.

Une machine assez récente pour l'hébergement des services et données est disponible: digsed, machine Dell R730XD achetée fin 2014, équipée de 2 CPU octocœurs, 64GB de RAM, d'un stockage de 32TB en Raid 6 et de d'ethernet 10G. L'acquisition d'une deuxième machines pour héberger les services pourra être nécessaire (Grid'5000 utilise 2 machines de services par site, pour la tolérance au panne et faciliter les maintenances).

3 Hébergement

3.1 Herbergement dans "le" datacentre mutualisé UGA, dans la salle du batiment IMAG:

- La gestion de l'hebergement est sous la responsabilité de l'UMS Gricad (pôle Datacentre, responsable: Gabrielle Feltin) et du groupe de travail datacentre (ct-datacenter@univ-grenoble-alpes.fr)

[9] ARM Juno, boitier non rackable

- La salle IMAG fournit un hébergement généraliste (pas d'arrivée d'eau), pour une capacité totale de 300kW
- C'est le partenaire privé du PPP (dit le "groupement") qui est propriétaire du bâtiment et des armoires et PDUs
- 5 rangées d'armoires en allées chaudes/allées froides
- 1 baie haute densité en bout de chaque allée: 30kW max
- autres baies 10kW max possible, et moyenne à 5kW max sur l'ensemble
- Les armoires et PDU sont les suivants:
  - Baies non-HD:
    * Armoire: APC modèle AR2580, 42U, 80cm de large `http://www.apc.com/shop/us/en/products/NetShelter-SV-42U-800mm-Wide-x-1200mm-Deep-Enclosure-with-Sides-Black/P-AR2580`
    * PDU: AP 8681 APC, zeroU, 16A, manageable `http://www.apc.com/shop/fr/fr/products/Rack-PDU-2G-Metered-by-Outlet-with-Switching-ZeroU-110kW-230V-21-C13-3-C19/P-AP8681`
    * Grid'5000 opère déjà des PDU APC, modèles: AP8659, AP7953, AP7851 et AP8659.
  - Baies HD:
    * Armoire: APC modèle AR2580, 42U, 80cm de large `http://www.apc.com/shop/us/en/products/NetShelter-SV-42U-800mm-Wide-x-1200mm-Deep-Enclosure-with-Sides-Black/P-AR2580`
    * PDU: `http://www.lichtchef.de/verbindungstechnik/bachmann/180447/bach329.800/bachmann/pdu-steckdosenleiste-bluenet-managed-3-phasig-network-19-1he-einzeln-schaltbar`

3.2 Infrastructure réseau unifiée pour tous les salles du datacentre: ACI Cisco

- Cette infrastructure est opérée par le groupe de travail Spring et la DSI UGA
- A priori il ne sera pas possible d'utiliser le réseau 172.16.0.0/16 dans des fonctions de routage virtualisées dans l'ACI, car cette plage est utilisée en interne pour la gestion de l'infrastructure. Un vlan sans IP pourrait par contre être compatible avec des équipements externes exploitant cette plage (routeur de site Grid'5000, noeuds).
- A priori il n'est pas possible de modifier les affectations de vlan sur les ports des équipements de l'ACI programmatiquement (serait nécessaire pour kavlan sans switch dédiés): même si techniquement c'est éventuellement possible, Spring ne propose pour pas cette fonctionnalité.
- Cett infrastructure étant nouvelle et encore en cours d'évolution, ces points pourront être rediscutés.

4 Intervenants sur les plateformes

4.1 Utilisateurs

- Grid'5000: communauté nationale expérimentions en informatique distribuée (HPC, Cloud Computing, Big Data, Réseau)
- CIMENT: communauté académique grenobloise

4.2 Technique

4.2.1 CIMENT:

- Gricad/Calcul: Bruno Bzeznik, Romain Cavagna, Laure Tavard
- Laboratoires: GRICAD coopère avec des ingénieurs des laboratoires impliqués dans CIMENT pour l'administration des machines et le support utilisateur

4.2.2 Grid'5000:

- Le pôle support de équipe technique nationale est composé d'environ 8 ETP ingénieurs (principalement des jeunes ingénieurs en CDD), 2 de ces ingénieurs sont hébergés à Grenoble (au LIG)
- Grid'5000 est portée par le LIG et Inria sur le site grenoblois. Au LIG, Pierre Neyron est l'ingenieurs responsable technique pour le site

4.3 Pilotage

4.3.1 CIMENT:

- Comité de pilotage sous la responsable scientifique d'Emmanuel Chaljub
- Le directeur technique est Bruno Bzeznik
- La communauté informatique de CIMENT est répartie sur 2 pôles du fait des contraintes géographiques notamment:
  - Pour le pôle "Informatique Distribué", Olivier Richard est le responsable scientifique et Pierre Neyron est le responsable technique
  - Pour le pôle INRIA, Frédéric Desprez est le responsable scientifique et Jean-François Scariot le responsable technique
- Par ailleurs CIMENT est maintenant structuré comme le pôle Calcul de l'UMS Gricad, dont Bruno Bzeznik est le responsable.

4.3.2 Grid'5000:

- Grid'5000 est structuré en un Groupement d'Intérêt Scientifique
- Les arbitrages concernant la plateforme (par ex: accords d'exceptions pour des usages ponctuels spéciaux) sont pris par le bureau du GIS assisté par le comité des responsables de sites.
- Le directeur scientifique du GIS est Frédéric Desprez
- Le directeur technique du GIS est David Margery (Rennes), cessant sa place a Lucas Nussbaum (Nancy) à partir de 2017
- Le responsable scientifique pour le site de Grenoble est Olivier Richard, suppléé par Pierre Neyron
- Le responsable technique local pour le suivi du site grenoblois est Pierre Neyron

4.3.3 Hébergement:

- Datacentre: Gricad, ct-datacenter@univ-grenoble-alpes.fr
- Réseau: Groupe de travail Spring / DSI UGA

5 Architecture matérielle du nouveau cluster

La configuration matérielle visée dans un premier temps pour le cluster devrait rentrer dans une enveloppe de l'ordre de 500 a 600K€, tout en permettant des extensions futures. L'objectif est le montage d'un cluster HPC orienté mouvement de données, équipé notamment de stockages performants (SSD, NVMe burst buffers) et d'un réseau d'interconnexion rapide 100 Gbps. La configuration matérielle actuelle telle que proposée par le fournisseur (Dell, marché Matinfo 3 lot 4) est la suivante:

- 120 noeuds Intel Xeon
  - 2x Xeon E5 2630 v4 (10 Coeurs)
  - 128GB RAM
  - 1TB HDD sata
  - 400GB SSD sata

- 8 noeuds Intel Xeon "Burst Buffers"
  - 2x Xeon E4 2630 v4 (10 coeurs)
  - 128GB RAM
  - 1TB HDD sata
  - 2TB SSD NVME
- Réseau rapide Intel Omnipath
  - 6 switchs pour une évolution future jusqu'à 192 noeuds.

Il faut ajouter à cette configuration les matériels nécessaires pour l'interconnexion avec les autres équipements des 2 plateformes:

- d'une part les besoins spécifiques de Grid'5000 (routeur de site pour l'interconnection nationale dédiée, switchs configurables pour kavlan, hyperviseurs Xen pour les services) ;
- d'autre part une rationalisation des coûts (achâts et opérations) en utilisant la solution réseau prédéployée dans le datacentre cible par l'UGA (Cisco ACI).

6 Scénarios d'exploitation

Plusieurs scénarios d'exploitation du cluster permettant de faire cohabiter des usages de type production et de type expérimentation en informatique distribuée peuvent être considérés[10]. Certains scénarios sont déjà mis en place sur d'autres sites Grid'5000 (Nancy[11] et Sophia[12]), mais jusque là aucun site Grid'5000 n'a encore été couplé avec un mésocentre et une communauté scientifique aussi large que celle de CIMENT [13]. Nous présentons ici un ensemble de scénarios sur la base de combinaisons de solutions techniques afin de concrétiser un fonctionnement global. Les scénarios 1 et 2 sont dérivés des fonctionnements déjà expérimentés dans Grid'5000: le scénario 1 est proche de l'architecture mise en place à Nancy et le scénario 2 est proche de celle mise en place a Sophia. Le scénario 3 quant à lui est dérivé de l'architecture d'un cluster CIMENT classique. Pour chacun des scénarios, nous proposons également des variantes ou adaptations envisageables pour notre projet. Ces propositions de scénarios ne sont bien sûr pas figées et ne prétendent pas répondre à toutes les questions: l'objectif est de proposer une base concrète pour des discussions, permettant ainsi de converger vers une solution technique réelle.

6.1 Description des scénarios

6.1.1 Scénario 1:

- L'integralité du cluster est instrumentée par Grid'5000 (ethernet pour kavlan sur tous les ports)
- L'integralité des services Grid'5000 est deployée (puppet)
- Cluster entièrement administré à la Grid'5000, avec le système d'exploitation Grid'5000
- Gricad/Calcul participe à l'administration au moins sur les aspects locaux
- Gricad/Caclul fournit la pile HPC, et participe aux efforts pour la rendre compatible/disponible sur l'ensemble de Grid'5000
- Les comptes G5K et CIMENT sont disponibles sur frontales et environnement par défaut des noeuds. 2 frontales (1 par communauté) peuvent être mises à disposition pour isoler les usages sur ces machines, et éventuellement pour simplifier l'authentification (domain par défaut si utilisation de SSSD)

[10] Usages et utilisateurs de Grid'5000: stratégie pour l'accès aux ressources - `https://hal.inria.fr/hal-01294910`

[11] Pour référence sur la configuration mise en place à Nancy:

- `https://www.grid5000.fr/mediawiki/index.php/Nancy:Production`
- `https://www.grid5000.fr/mediawiki/index.php/TechTeam:Production_Clusters_Design_Doc`

[12] Pour référence sur la configuration mise en place à Sophia:

- `https://www.grid5000.fr/mediawiki/index.php/TechTeam:CT-113-ClusterSophiaProd`

[13] les couplages de Nancy et Sophia sont internes au centre Inria/laboratoire d'informatique

- Le conflit d'uid/gid et d'username/groupname CIMENT et Grid'5000 est corrigé, de manière à ce qu'une même machine puisse authentifier les comptes des 2 communautés
- 1 seul serveur OAR: gestion à grain fin de la cohabitation des jobs Grid'5000 et CIMENT
  - Ressources partitionnées Grid'5000/CIMENT au prorata des financements
  - Chaque job, à tout moment, va sur une partition ou l'autre suivant la communauté à laquelle appartient l'utilisateur
  - Une queue/un scheduler pour G5K (charte G5K, réservation à l'avance favorisées) qui utilise la partion de ressources Grid'5000
  - Une queue/un scheduler pour CIMENT (charte CIMENT, batch privilegié, fairsharing par projet) qui utilise la partition de ressources CIMENT
  - Des jobs "challenge" autorisés sur demande du comité de pilotage seulement, qui peuvent utiliser l'intégralité des ressources du cluster. Concernant l'ordonnancement de ces jobs, différentes options seront possibles, suivant la priorité que l'on souhaitera leurs donner (utilisation de la queue CIMENT pour ne pas pénaliser l'usage de production classique par exemple).
  - Une campagne de jobs best-efforts pourra également accèder à toutes les ressources facilement
  - Les quelques noeuds en avance de phase technologique (noeuds burst-buffer) sont positionnées dans la partition Grid'5000.

6.1.2 Scénario 1.1:

Scénario 1 +:

- L'OS par défaut sur les noeuds n'est pas le même sur la partition Grid'5000 et sur la partition CIMENT:
  - L'unification des comptes n'est plus nécessaire sur les noeuds et les frontales de la communauté, mais le reste sur le serveur OAR
  - L'OS CIMENT peut être différent de celui de Grid'5000, et est donc non contraint aux choix techniques (distribution Linux) et aux évolutions de ce dernier
  - L'OS CIMENT peut intégrer des tunings HPC sans contraindre l'OS Grid'5000
  - L'OS CIMENT peut éventuellement tourner sur un VLAN différent de l'OS Grid'5000 (cas ou une seule interface 10GE est cablée, et nécessite de pouvoir programmer l'ACI si kavlan n'est pas sur un réseau dédié)
- Les jobs challenges sont nécessairement des jobs deploy afin d'uniformiser le système d'exploitation (configuration des comptes, montage NFS, VLAN ?)
- Une campagne de jobs best-effort non-deploy ne pourra pas utiliser l'ensemble de la plateforme avec un environnement de calcul identique
- Pas de mutualisation des efforts pour un système d'exploitation commun: la pile HPC CIMENT n'est pas nécessairement integrée dans l'environnement Grid'5000.

6.1.3 Scenario 1.2:

Scénario 1 (système d'exploitation Grid'5000 + pile HPC CIMENT identique sur tous les noeuds / 1 seul OAR) mais:

- Pas de partition: les utilisateurs CIMENT et Grid'5000 sont en compétition sur l'ensemble des ressources avec une charte d'usage commune:
  - Les utilisateurs Grid'5000 peuvent faire des réservations à l'avance
  - Les utilisateurs CIMENT peuvent faire des jobs avec un walltime jusqu'à X jours
  - Les chartes Grid'5000 et CIMENT nécessitent donc d'être fortement adaptées pour proposer un compromis. La charte résultante restera a priori spécifique au cluster (différente des autres sites Grid'5000 ou autres clusters CIMENT)
  - Les jobs Grid'5000 et CIMENT sont gérés dans un même queue/avec un même scheduler (hors best-effort) afin de ne pas donner de priorité à une communauté ou l'autre.

6.1.4 Scenario 2:

- L'integralité du cluster est instrumentée par Grid'5000 (ethernet pour kavlan sur tous les ports)
- L'integralité des services Grid'5000 est deployé (puppet)
- Les ressources du cluster basculent entre 2 modes d'exploitation, suivant un calendrier:
  - Mode Grid'5000: OS par défaut de Grid'5000 avec comptes utilisateurs Grid'5000 uniquement, sans adaptation pour CIMENT. Pile d'instrumentation Grid'5000 complète disponible.
  - Mode CIMENT avec l'OS et la pile HPC CIMENT
- 2 serveurs OAR, 1 pour CIMENT et 1 pour Grid'5000: gestion à gros grain de la répartition du temps d'utilisations Grid'5000 et CIMENT:
  - Bascule pour passer des noeuds d'un mode de fonctionnement à l'autre (réservation en cosystem et redéploiement du cluster dans l'environement de l'autre mode)
  - Faut-t'il reconfigurer le réseau suivant le mode de fonctionnement pour isoler les usages ? (comme à Sophia ? Ethernet et réseau rapide ?)
- Calendrier d'affectation des ressources à un mode fonctionnement défini par le comité de pilotage.
- La bascule se fait pour des périodes longues (bi-hebdomadaire?) et décorrélées des jobs des utilisateurs (ce ne sont pas les jobs utilisateurs qui déclanchent la bascule, il peut y avoir des disponibilités inexploitées car pas dans le bon mode de fonctionnement)
- Très peu de collaboration nécessaire entre Gricad/Calcul et Grid'5000

6.1.5 Scenario 2.1:

Scénario 2 +:

- Les ressources du cluster sont partitionnées: une partition Grid'5000, une partition CIMENT
- Chaque partition fonctionne avec le mode de fonctionnement correspondant
- Pas de bascule calendaire
- Un "mode challenge" permet de basculer l'ensemble des ressources dans un même mode de fonctionnement, en attente de soumission de jobs utilisateurs.
- Ce mode "challenge" est activé pour des périodes définies par le comité de pilotage, typiquement pour plusieurs semaines.
- Les quelques noeuds en avance de phase technologique (noeuds burst-buffer) sont positionnées dans la partition Grid'5000.

6.1.6 Scenario 2.2:

Scénario 2.1 (partition de la machine pour faire un cluster G5K et un cluster CIMENT), mais:

- Les noeuds CIMENT ne sont pas instrumentés (pas de réseau 10GE/kavlan)
- Pas de bascule des noeuds de CIMENT vers Grid'5000
- Il reste par contre possible de passer les noeuds Grid'5000 dans CIMENT: Typiquement, l'utilisateur à un compte CIMENT et Grid'5000 et se débrouille pour réserver les noeuds en même temps dans les 2 differents OAR, à l'aide d'une réservation à l'avance coté Grid'5000 par exemple.

6.1.7 Scenario 3:

- Le cluster est installé avec la pile CIMENT
- On maximise le nombre de noeuds au détriment de l'équipement nécessaire pour l'instrumentation Grid'5000 (pas de réseau supportant kavlan, certains noeuds n'ont éventuellement même pas d'Ethernet 10G cablé ?)
- Seule une sous partie de l'instrumentation Grid'5000 est mise en place (kadeploy ? Quoi d'autre ?).
- Les services Grid'5000 authentiques n'ont pas besoin d'être installés (installation CIMENT, pas de puppet Grid'5000)
- Les utilisateurs Grid'5000 peuvent demander des comptes et ont un accès privilégié aux ressources en rapport avec l'investissement Inria/CPER (via paramètrage du fairesharing + réservation à l'avance ?)
- Investissement de l'equipe technique Grid'5000 sur le cluster minimal

6.2 Analyse des scénarios

Ces analyses sont celles des auteurs, elles sont ouvertes à discussion.

6.2.1 Scénario 1: Un seul cluster intégré au plus près de Grid'5000

C'est le scénario le plus fin en terme d'optimisation de l'usage du cluster et d'agilité pour mobiliser toutes les ressources, avec la pile expérimental complète de Grid'5000 et en même temps la pile HPC Ciment. C'est le scénario qui implique la collaboration la plus forte entre les équipes techniques des 2 plateformes.

C'est aussi le scénario pour lequel il y a le plus de challenges techniques pour les plateformes:

- Pour Grid'5000, la gestion du cluster doit être homogène avec le reste de Grid'5000 afin de pouvoir être supportée par l'équipe technique et d'offrir au utilisateurs Grid'5000 un environnement semblable au reste de la plateforme (sans spécificités majeures). Si des évolutions doivent être apportées à l'architecture Grid'5000, elles devront être considérées d'un point de vue national.
- Pour CIMENT, il faut limiter les divergences par rapport au reste de CIMENT en terme d'usage (interface utilisateur et charte d'accès au ressources), et permettre un accès performant vers les services externes au cluster (stockage, autres clusters).

1. Avantages:
   - Permet d'avoir un cluster uniforme
   - Permet des expériences exploitant l'intergralité du cluster avec la pile expérimentale Grid'5000
   - Permet des calculs sur l'intégralité du cluster avec la pile HPC CIMENT
2. Inconvénients:
   - Nécessité d'identifier les spécificités locales nécessaires, de les implémenter puis de les maintenir dans le temps
     - Environnement logiciel par défaut: est-ce qu'un environnement Grid'5000 standard sur lequel on ajouterait Nix/modules serait suffisant ?
     - Quid de la gestion de comptes ? Est-ce que la création (à la main) de comptes Grid'5000 est suffisante ? Jusqu'où veut-on aller ?
     - Quelle politique de réservation de ressources ? Est-ce que la charte Grid'5000 convient ? Faut-il utiliser uniquement une queue production ? Ou une autre politique de réservation ? Comment préserver les usages des deux communautés ?
     - Quelles ressources humaines pour ce travail spécifique et sa maintenance dans le temps ?
     - Comment faire de l'accounting (stats d'utilisation, reporting) pour deux communautés à la fois ?

6.2.2 Scénario 1.1:

Par rapport au scénario 1, on supprime la contrainte sur l'environnement système ce qui permet de déplacer des problèmes, mais globalement le challenge technique reste toujours élevé.

6.2.3 Scénario 1.2:

Il semble très difficile de concilier les usages CIMENT et Grid'5000 dans une même charte et une même politique de scheduling. Ce scénario qui est éventuellement la solution la plus naïve sera donc a priori pas retenu.

6.2.4 Scénario 2: Des noeuds qui basculent entre deux clusters

On a un pool de noeuds, qui peuvent basculer dynamiquement d'un cluster à un autre selon une politique d'allocation. c'est le scénario utilisé à Sophia avec une alternance une semaine sur deux (d'autres "compromis" sont possibles, par exemple: 20% G5K / 80% CIMENT en journée, l'inverse la nuit, par exemple, ce qui amène à la variante 2.1).

1. Avantages:
    - C'est une approche plus conservatrice par rapport à l'existant que le scenario 1
    - Un fonctionnement similaire est déjà en place à Sophia
2. Inconvénients:
    - Gestion des ressources à gros grain:
        - Du fait des walltimes eventuellement importants, cela va générer des trous dans l'ordonnancement
        - Problème de corrélation du calendrier avec les deadlines pour les soumissions de papiers ?
    - Cela revient à séparer de manière rigide les communautés. Dans une optique de mutualisation, ce n'est pas une très bonne solution.
    - A Sophia, cela fonctionne avec KaVLAN, donc c'est d'autant plus important d'avoir KaVLAN fonctionnel (10GE + Omnipath ?)

6.2.5 Scénario 2.1:

Cette variante du scénario 2 avec un partitionnement des ressources semble plus intéressante que la bascule calendaire (typiquement bimensuel) contraignante par rapport aux walltimes CIMENT.

6.2.6 Scénario 2.2: Chacun pour soi

On sépare les financements et on fait un cluster par communauté d'utilisateurs. Le cluster Grid'5000 est pleinement intégré à Grid'5000, sans aucune différence avec les autres clusters.

1. Avantages:
    - C'est encore plus simple
    - Pas besoin d'équiper tout le cluster avec l'instrumentation Grid'5000
2. Inconvénients:
    - Cela revient à ne rien mutualiser en dehors de l'hébergement
    - La plateforme Grid'5000 resultantes est de taille inferieure aux espérences
    - Il reste la question des spécificités imposées par l'hébergement dans le DC UGA (réseau – KaVLAN possible ?; PDU – Kwapi possible ?).

6.2.7 Scénario 3: cluster opéré avec la pile CIMENT

1. Avantages:
   - Permet d'avoir un cluster uniforme
   - Permet des expériences exploitant l'intergralité du cluster, dans le respect de la charte CIMENT: quelle sera cette charte ?
2. Inconvénients:
   - Ne permet pas tout le panel d'experimentation normalement supporté par Grid'5000.
   - Le cluster ne fait pas effectivement partie de "Grid'5000": l'équipe technique Grid'5000 n'intervient pas sinon de manière détournée en fournissant des environnements kadeploy par exemple. Les utilisateurs ne retrouvent pas l'environnement Grid'5000 habituel.
   - Probablement inacceptable pour Grid'5000, car:
     - Pour les utilisateurs, la plate-forme serait très différente des autres, alors que les autres sites de Grid5000 sont tous identiques sur le plan logiciel pour faciliter les expériences multi-sites et le portage d'expériences entre sites;
     - L'efficacité du travail de l'équipe technique repose sur l'homogénéité de la configuration de chaque site pour faciliter la gestion (notamment le déploiement ou les évolutions des services spécifiques). Le fait de gérer plusieurs environnements cibles aurait un coût très important.

7 Développements sur les challenges techniques

Cette section a pour objectif de développer les challenges techniques et de proposer des solutions pour la faisabilité des scénarios. Nous nous intéressons en particulier au cas du scénario 1 car c'est celui qui propose l'approche la plus fine pour un l'inter-opération du cluster sur les 2 plateformes, que cela soit en terme d'utilisation et de coopération des équipes techniques.

7.1 Administration de la plate-forme et support utilisateur

Dans l'hypothèse ou le cluster est co-administré par les équipes Gricad/Calcul et Grid'5000 selon le scénario 1, le fonctionnement suivant est envisageable:

- L'équipe Grid'5000 ne traitera pas des particularités du site grenoblois sauf si elles peuvent être généralisées (intégrées) au niveau national. Elle fournira cependant une grosse partie de l'effort sur la plateforme (bénéfice de l'effort national), notamment pour fournir un environnement système robuste malgré le degré de reconfiguration fort.
- L'équipe Gricad/Calcul prendra en charge la pile logicielle HPC et en particulier les développements nécessaires pour répondre aux besoins des utilisateurs CIMENT. Cette pile logicielle est commune avec les autres clusters de CIMENT. L'environnement logiciel des noeuds étant le même sur l'ensemble du cluster, cette pile logicielle sera également disponible dans la partition Grid'5000. Une question ouverte consiste à savoir si cette pile logicielle pourrait également être utilisée sur d'autres sites Grid'5000.

Les 2 communautés utiliseront des partitions distinctes du cluster (sauf jobs "challenge") mais elles seront également bien identifiée par le comptes utilisateurs (bases de gestion des comptes qui resteront distinctes). Il sera donc aisé d'aiguiller les utilisateurs vers l'équipe support pertinente en fonction de la communauté d'appartenance: CIMENT ou Grid'5000. Seuls les utilisateurs informaticiens ayant un usage de type la production seraient dans une situation équivoque: ils devront avoir un compte sur chaque plateforme pour accèder à la bonne partition du cluster (ainsi qu'aux autres équipements de chaque plateforme), et pour identifier facilement leurs usages.

Les responsabilités des équipes techniques des 2 plateformes pourraient s'organiser ainsi:

7.1.1 Equipe technique Grid5000:

- Fourniture du système de base du cluster
  - Système d'exploitation généraliste (Quid du tuning HPC ?)
  - Installation de la pile logicielle réseau rapide: Omnipath/Infiniband
  - Installation de la pile logicielle GPU/Accélérateurs (pas utile sur ce cluster)
- Fourniture d'environnements alternatifs
- Administration de la pile outils grid'5000 intégré à l'OS:
  - kadeploy
  - OAR
  - kavlan
  - g5k-check
  - sudo-g5k
  - autre?
- Administration des services Grid'5000 et de leur infrastructure de déploiement (puppet)
- Gestion des comptes Grid'5000
- Support utilisateur Grid'5000 + aiguillage vers le support CIMENT si pertinent.

7.1.2 Equipe Gricad/Calcul:

- Fourniture de l'environnement HPC (Nix ou modules)
- Participation à la configuration de OAR
- Gestion des comptes CIMENT
- Support utilisateurs CIMENT/production
- Participation au support global, notamment si cela concerne les utilisateurs CIMENT (participation aux reunions PS et CT pour 1 personne de Gricad/Calcul au moins)

7.2 Gestion des comptes

Les 2 plateformes gardent leurs propres systèmes de gestion des utilisateurs, et notamment la base d'authentification de compte unix. Pour un système d'exploitation commun pour les 2 plateformes, il est donc nécessaire de mettre en place un mécanisme de réunion des 2 ensembles de comptes. Les problèmes techniques pour les comptes unix sont les suivants:

1. Les 2 plateformes doivent utiliser des uid et des gid distincts (ce n'est pas le cas actuellement)
2. Les 2 plateformes doivent faire en sorte d'utiliser des identifiants username/groupname distincts (sans garantie actuellement)
3. L'OS doit être capable d'inter-opérer avec les 2 gestions de comptes à la fois
4. Le montage des répertoires home doit être homogène avec le reste de la plateforme (au moins pour Grid'5000, quid CIMENT ?)

Les configurations des uid/gid et le montage home des 2 plateformes sont actuellement comme suit:

- Pour Grid5000:
  - uid à partir de 10000 et plus grand uid actuel égal à 20057.
  - gid par défaut pour les utilisateurs (groupe users) positionné à 8000. D'autres gid sont utilisés pour des groupes "fonctionnels" et utilisent la plage [9000-10000[. Enfin des groupes "structurels" (typiquement pour les sites) utilisent quelques gid répartis toutes les dixaines de milliers à partir de 10000, le plus grand gid utilisé étant 21001.

  - home pour tous les utilisateurs à plat dans /home

- Pour CIMENT:
  - uid à partir 10000, et plus grand uid actuel égal à 10769
  - gid actuellement entre 10000 et 10316. Ces gid sont utilisés pour identifier les pôles, laboratoires et projets.
  - home pour tous les utilisateurs à plat dans /home

Une proposition serait donc de travailler à une renumérotation des uid/git de CIMENT vers dans la plage 5000+ ou 30000+ par exemple. Le choix d'un travail coté CIMENT plutôt que Grid'5000 se justifierait par le fait que la répartition des uid/gid est plus simple sur CIMENT, et que Grid'5000 donnant l'accès root à ces utilisateurs, il est beaucoup plus difficile de maîtriser l'impact d'un tel changement. Concernant les home directories, il faudra étudier l'impact d'un changement de chemin par rapport aux autres clusters dans CIMENT, en espérant que cela soit moins problématique que pour Grid'5000. Une fois ce problème de conflit d'uid/gid réglé, plusieurs solutions techniques seront envisageables pour mettre en place une authentification unique sur les machines:

7.2.1 Utilisation de SSSD

SSSD permet une gestion de compte qui mixte plusieurs annuaire LDAP au niveau du système d'authentification, ajoutant un suffixe de domaine pour chaque communauté aux usernames et goupnames. Un seul domain peut être défini comme "par défaut" mais avec une frontale par communauté et éventuellement des noeuds dans un environnement spécifique pour chaque partition (scénario 1.1), chaque communauté pourrait avoir le bon domaine par défaut sur la machine. Les domaines seraient donc peut intrusifs dans l'interface utilisateur, même si à gérer coté OAR serveur.

- Nécessite uniquement des uid/gid distincts
- Solution poussée par redhat pour l'interface Linux/LDAP
- L'utilisation de plusieurs domain implique un changement dans l'affichage des usernames et groupnames (suffixe @domain)
- Il est possible de paramétrer le chemin vers le home directory des utilisateurs fourni par LDAP.

7.2.2 Utilisation de branches LDAP greffée

- Nécessite de bien avoir uid/gid mais aussi username/groupname distincts entre Grid'5000 et CIMENT.
- Support peu robuste dans le passé avec openldap ?
- Quel base DN utiliser ? dc=grid5000,dc=fr est il requis pour le fonctionnement de la plateforme ? idem dc=ci-ra,dc=org ?

Avec une telle solution opérationnelle, l'interface de compte proposée au utilisateur changerait moins qu'avec SSSD, car pas de domaine suffixé.

7.2.3 Utilisation d'un mécanisme d'import des comptes CIMENT dans la base de comptes Grid'5000

- Nécessite également que les uid/gid et username/groupname de CIMENT ne soient pas utilisées par Grid'5000 ou bien si d'autres identifiants compatibles avec Grid'5000 sont utilisés, de mettre en place d'un mapping pour les autres ressources de CIMENT: CIMENT réalise déjà ce genre de manipulation pour l'accès et le montage NFS de la solution de stockage universitaire (SUMMER) [14].
- Nécessite un développement conjoint Grid'5000 et CIMENT, par exemple:

[14] L'utilisation de gid supplémentaires peut apporter une solution technique pour permettre à un même utilisateurs ayant des comptes d'uid différents d'avoir accès à ses fichiers

  - CIMENT: mise en place d'un mécanisme d'export des informations et mise en place du mécanisme de mapping des identifiants pour les ressources externes au cluster
  - Grid'5000: import dans le système de gestion des comptes (UMS) des comptes issus de CIMENT et identification de ceux-ci (avec un groupe ?). En retour, report dans CIMENT des username/groupname et uid/gid utilisés pour la mise en place effective du mapping pour chaque nouvel utilisateur

- Ces comptes "CIMENT" dans UMS devenant des compte Grid'5000 nationaux, il faudra spécifier leur usage en dehors de Grenoble: restreindre l'accès sur les autres sites, ou juste la soumission des jobs sauf cas execptionel (mais dans ce cas le compte deviendrait un compte Grid'5000 à part entière ?), ou alors juste bien expliquer dans la charte donnée au utilisateurs CIMENT ?

- Ciment utilise le projet associé aux jobs OAR: ces projets sont définit dans la gestion de compte CIMENT (Perseus) et utilisés pour le fairsharing (pas utilisé par Grid'5000, donc pas de conflit ?) et dans les admissions rules (à prendre en compte pour la queue de CIMENT)

7.3 Système d'exploitation

- Système d'exploitation par défaut fourni par Grid'5000: Debian jessie actuellement, stretch à moyen terme: quelle gestion des évolutions vis-à-vis de CIMENT ?

- Autres système d'exploitation mis a disposition par kadeploy, et éventuellement d'autres solutions logicielles futures de Grid'5000

- Reboot très frequents sur les noeuds Grid'5000

- Dans le cas du scénario 1.1, pour lequel la partition CIMENT aurait un environnement par défaut (utilisé pour les jobs non deploy) différent de la partition Grid'5000, il faudrait étudier comment préserver un mécanisme de vérification des ressources. En effet Grid'5000 utilise l'outil g5k-check au niveau de son environnement par défaut pour vérifier la conformité des noeuds avec les informations enregistrées dans l'inventaire de la plateforme (reference API), notamment pour s'assurer que les ressources éventuellement affectées aux jobs deploy sont saines. Avec le mécanisme de job challenge, il faudrait donc avoir en particulier cette assurance sur les ressources de la partition CIMENT.

7.4 Environnement HPC

- Gestion des licences ? Comment sécuriser les accès aux jetons de licenses CIMENT ?

- Tuning HPC pour CIMENT ? (Hyperthreading, C-State/P-State/Turbo, option de gestion mémoires/numa pour HPC, etc)

7.5 Réseau et accès aux ressources externes au cluster

Grid'5000 est une plateforme nationale disposant d'un réseau national dédié, d'un routage intersite cohérent et de passerelle d'accès de et vers Internet. Pour permettre au site Grid'5000 de Grenoble d'être opéré par l'équipe technique nationale, il est nécessaire de limiter autant que possible les particularités de site. Par ailleurs, le cluster doit être connecté aux autres équipement CIMENT: stockage Irods, BeeGFS, autres clusters, et être accessible de manière performante depuis les laboratoires. Ainsi, un compromis au niveau de la configuration du réseau doit être trouvé entre la configuration conforme au plan d'adressage Grid'5000 national (Golden Rules[15]) et l'adressage CIMENT (IP publiques et routage UGA).

L'hébergement disponsant d'une infratstructure réseau préinstallée avec l'ACI Cisco, il est important de rationaliser les coût et efforts en l'utilisant au maximum. L'architecture de l'ACI consiste en 2 niveaux de switchs:

- 2 spines interconnectés vers N leaves par 2 liens 40GE pour chacun (équilibrage de charge ou simple redondance ?)

- Leaves 48 ports 10GE (sfp+) vers l'exterieur de l'ACI: machines dans les différentes salles du datacentre + machines externes (labo, accès renater, etc)

[15] `https://www.grid5000.fr/mediawiki/index.php/Grid5000:Network`

Les contraintes des plateformes sont les suivantes:

- Grid'5000:
  - Requiert la capacité de reconfigurer le réseau Ethernet du cluster (et Omnipath ?)
  - Requiert un accès fin au équipements pour les mesures.
  - Requiert de se conformer à son plan d'adressage IP national[15] (réseau privé 172.16.x.y), à sont routage (passerelle par défaut des noeuds vers les autres sites Grid'5000 ou VPN de secours) et à sa politique de sécurité (machines d'accès nationnales, NAT national en sortie)
- CIMENT
  - Requiert une connectivitée performante vers ses autres équipements (éviter tout routage superflu): stockage BeeGFS et cluster Luke notamment
  - Requiert une connectivitée performante vers les laboratoires
  - Fonctionne habituellement avec des IP publiques sur ses noeuds qui ont accès à internet via le réseau UGA

Scénarios possibles:

7.5.1 Installation de switchs dédiés pour Grid'5000 + routeur Grid'5000

- Implique des coûts supplémentaires à l'achat et à la maintenance
- Upling 40GE vers un switch leaf de l'ACI (4x 10GE coté swith ACI)

7.5.2 Utilisation de l'ACI Cisco avec paramétrage des VLANs via son interface programmatique par kavlan

- Pour l'instant Spring/DSI UGA ne permet pas de programmer ainsi l'ACI
- Impossible d'utiliser un routeur virtuel de l'ACI pour le routeur de site Grid'5000 (problème des IP 172.16.x.y), il faudra dans tous les cas faire l'acquisition d'un équipement routeur de site Grid'5000 dédié (à rediscuter car Cisco semble avoir infirmé le conflit sur les IP 172.16.x.y)

7.5.3 Pas de kavlan

- Limitations pour les expériences
- Divergence par rapport aux autres site Grid'5000
- Limitations pour le mécanisme de bascule du scénario 2
- Routeur de site Grid'5000 dédié, idem ?

8 Conclusion

La mise en place d'un cluster pertinent pour l'expérimentation sur l'informatique distribuée est un objectif pour les 2 communautés: CIMENT et Grid'5000. Ainsi la configuration qui sera mise en place permettra, nous l'espérons, une convergence importante entre les usages. On notera bien sûr que Grid'5000 reste une infrastructure recherche pour la communauté informatique et demande des connaissances techniques très pointues pour son utilisation. Ce projet encouragera et nécessitera donc les collaborations fortes entre les 2 communautés.

Du point de vue de l'ingénierie, ce projet vise une solution technique qui pousse un cran plus loin la coopération sur les infrastructures scientifiques dites "pour l'expérimentation" et "de production". Nous espérons que ce document permettra d'éviter autant d'écueils que possible dans la mise en place aussi réelle que maîtrisée de ce cluster commun CIMENT et Grid'5000.