Comment mutualiser sans centraliser ?

Romaric David
Université de Strasbourg, Direction Informatique
7 Rue de l'Université 67000 STRASBOURG

Mehdi Amini
Université de Strasbourg, Direction Informatique
7 Rue de l'Université 67000 STRASBOURG

Résumé

Qui n'a pas rêvé de réutiliser la puissance de calcul des ordinateurs récents qui équipent les salles d'enseignement de nos universités ? À l'occasion d'un calendrier favorable et à des fins d'expérimentation, nous avons pendant un an réutilisé des PC présents dans des salles d'enseignement comme nœuds de calcul d'un cluster existant sur un autre site de notre université.

Nous présentons dans cet article les choix techniques effectués afin de simplifier l'infrastructure logicielle nécessaire, l'utilisation que nous avons pu faire de ces ressources et les enseignements de l'expérimentation.

En particulier, nous avons mis à profit ce travail pour tester et valider dans notre contexte et en grandeur réelle des outils du monde du calcul intensif et de la gestion des clusters (Torque, Maui). Grâce à cela, nous avons pu préparer la fin de vie de logiciels commerciaux et la migration de configurations existantes.

Enfin, nous présenterons les enseignements de cette expérience, qui a permis de valider l'utilisation de ressources de calcul réparties sur plusieurs sites, aussi bien logiciellement que matériellement. Ce mode géographiquement distribué constitue désormais la clef de voûte de notre architecture de cluster et nous permet de répartir les contraintes d'infrastructure des salles machines dédiées au calcul.

Mots clefs

Mutualisation, Cluster, Calcul, Infrastructure

1 Introduction

Ce travail a pour cadre le méso-centre de l'Université de Strasbourg (UdS), intégré au Département Expertise pour la Recherche de la Direction Informatique de l'Université de Strasbourg, créée au 1er Janvier 2009. Le Département Expertise pour la Recherche a pour mission l'accompagnement scientifique et technique à l'utilisation de ressources de calcul intensif. De plus, le département met des clusters de calcul à la disposition des chercheurs rattachés à l'UdS. Ces clusters sont utilisés en mode production.

Dans [1], nous présentions une démarche de mutualisation s'appuyant sur des ressources centralisées géographiquement. Ce précédent travail, portant sur les ressources de production, avait permis de consolider la puissance de calcul proposée par le méso-centre. La mutualisation s'appuyait sur des politiques d'exploitation innovantes permettant d'offrir un niveau de service satisfaisant aux financeurs de ces ressources.

Dans une optique d'expérimentation et à la faveur d'une collaboration informelle entre l'UFR de Mathématiques et la Faculté des Sciences de la Vie de l'Université Louis Pasteur de Strasbourg[1], nous avons utilisé jusqu'à 80 postes de travail comme nœuds de calcul additionnels. Partageant les données et les comptes utilisateurs, ces machines étaient accessibles depuis le frontal de soumission habituel des utilisateurs.

[1]Au 1er janvier 2009, l'ULP a fait partie des établissements intégrant la nouvelle Université de Strasbourg

2 Motivations du projet

2.1 Le matériel en place

En Janvier 2008, au démarrage du projet, les ressources de calcul parallèle accessibles à la communauté recherche de l'Université se composaient de (année d'acquisition entre parenthèses) :

- 30 serveurs IA-64, bi-processeurs Itanium 2/1.3 Ghz, 8 Go de RAM par machine (2003) ;
- 32 serveurs x86-64, bi-processeurs Opteron/2.4 Ghz, 4 Go de RAM par machine (2005-2007) ;
- 17 serveurs x86-64, mono-processeur dual-cœurs Athlon/2.4 Ghz, 2 Go de RAM par machine (2006).

Regroupés dans une salle machine unique (désignée par *site principal* dans la suite de l'article), ces 3 clusters sont utilisés à 80% de leurs temps, malgré l'âge des processeurs pour certains ou la faible quantité de mémoire vive pour d'autres.

La convergence des processeurs utilisés en calcul scientifique vers des processeurs de type x86-64 rend les machines que l'on peut trouver dans les salles d'enseignement relativement proches (hors facteur de forme) des serveurs de calcul. L'hétérogénéité matérielle généralement constatée entre différentes salles d'enseignement modère cette proximité, les clusters de calcul étant constitués d'un grand nombre de machines identiques.

Or, en Janvier 2008, il est apparu que 80 machines homogènes et récentes (Intel Core 2 Duo/2Ghz, 2 Go de RAM par machine) étaient présentes dans un même bâtiment et qu'il serait possible d'intervenir sur les systèmes d'exploitation (OS) déployés sur ces postes. Dès lors, comment ne pas penser à fédérer ces machines en cluster ?

Nous avons alors dressé l'inventaire des contraintes à prendre en compte pour ce projet.

2.2 Intégration de ressources extérieures : contraintes

2.2.1 Du coté infrastructure de calcul

Le projet d'intégration des machines de salles de ressources informatiques aux nœuds des clusters existants devait satisfaire les contraintes suivantes :

- coût matériel et logiciel réduit. Par exemple, il n'est pas imaginable d'investir dans des réseaux haut-débit dédiés calcul ;
- coût humain limité. Par exemple, les systèmes d'exploitation en place ne doivent pas être modifiés de fond en comble pour la simple réalisation de la maquette.

2.2.2 Du coté salles de ressources

- priorité aux enseignements : pas de machine qui « rame ». Les tâches de calcul doivent pouvoir être interrompues en début de séance de TP. Volontairement, nous n'avons pas prévu de mécanisme de reprise de ces calculs. Nous expliquerons dans la suite les motivations de ce choix ;
- automatisation de la mise à disposition des machines pour le calcul ;
- pas de déménagement des machines !
- mêmes contraintes de coût que précédemment.

2.2.3 Contraintes transversales

- accès à l'espace de stockage déjà existant ;
- accès à l'annuaire utilisateur sans modification de celui-ci.

3 Réalisation

De nombreux projets de fédération de ressources de calcul (CiGri [2], ComputeMode [3]) existent. À la date de réalisation de ce travail, CiGri et ComputeMode auraient nécessité l'installation d'un serveur dédié, ce qui était inenvisageable dans notre contexte. En effet, nous avons estimé que les coûts humains liés à l'acquisition de compétences sur l'un ou l'autre de ces produits allaient à l'encontre des contraintes évoquées plus haut. De plus, le but était de rester volontairement proche des configurations et des outils déjà en place.

Nous souhaitions en premier lieu relier les machines d'enseignement au réseau de nos clusters du site principal.

3.1 Mise en réseau des machines

Sur le site principal, les clusters sont construits autour de réseaux locaux dédiés, avec une plage d'adresses par réseau local. Un routeur dédié au calcul sert de point d'entrée unique (route statique définie sur le réseau métropolitain) pour ces 3 réseaux. Les données des utilisateurs sont partagées par NFS sans autre identification que l'adresse IP source. Les utilisateurs sont stockés dans un annuaire simple (NIS à la date de réalisation du projet).

Pour intégrer les machines hors site, nous avons choisi d'ajouter un nouveau réseau local et une nouvelle plage d'adresses. Ce réseau a été autorisé à procéder aux montages NFS.

Comme le nouveau réseau local doit être à cheval sur plusieurs sites, nous avons demandé la création d'un VLAN transversal au réseau Osiris, le réseau métropolitain enseignement/recherche de Strasbourg. Ce VLAN relie le bâtiment abritant physiquement les salles de TP à la salle machine hébergeant les ressources de calcul *historiques*, à 6 kms à vol d'oiseau. Le sous réseau IP utilisant ce VLAN est comme les réseaux in-situ routé statiquement par le routeur dédié au calcul.

À ce stade de l'opération, n'importe quelle machine de la salle de ressources pourrait avoir accès à ce VLAN, ce qui n'est pas souhaitable pour des raisons de sécurité élémentaires :

– accès à l'annuaire d'authentification ;
– accès aux fichiers exportés par NFS.

C'est pourquoi nous avons mis en place une authentification 802.1X sur les commutateurs auxquels sont reliées les machines de la salle de ressources. Ainsi, pour accéder au VLAN calcul, une machine doit s'authentifier. En l'absence d'authentification, la machine rejoindra le VLAN enseignement. Ceci permet de s'assurer qu'aucune machine n'aboutit, par hasard, sur le VLAN calcul. Le schéma du réseau est représenté Figure 1.

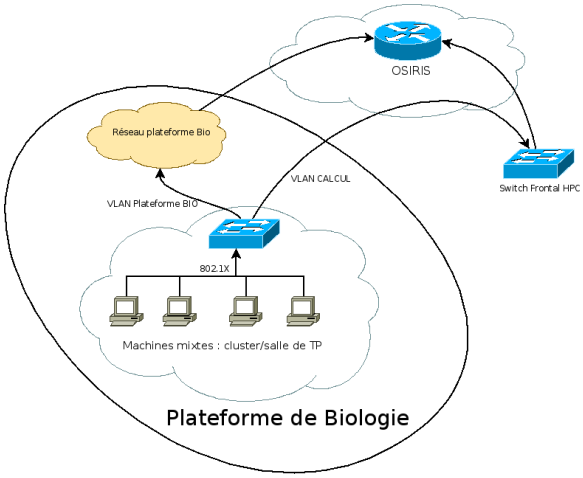

Figure 1 - VLAN transversal authentifié

Les paramètres d'authentification sont figés dans la configuration de xsupplicant, utilisé pour la mise en œuvre de 802.1X. Une fois authentifiée, la machine a accès au serveur NFS et au serveur d'annuaire. Nous conservons ainsi une authentification simple, sans certificats utilisateurs ou re-mapping d'UID, comme ce que l'on trouve dans les grilles de production comme LCG[2].

Une fois les machines d'enseignement reliées au bon réseau, il reste à leur fournir tous les outils nécessaires au calcul.

3.2 Aspects logiciels

Nous avons installé sur une nouvelle partition un autre système sur les machines d'enseignement. Trois images systèmes cohabitent ainsi :

– Une image Windows (Enseignement) ;
– Une image Linux (Enseignement) ;
– Une image Linux (Calcul). Cette image est très proche de l'image utilisée sur l'un de nos clusters (à la version de noyau près).

Les images *Enseignement* démarrent uniquement en réponse à une sollicitation de l'utilisateur. Sans sollicitation de sa part, la machine démarre en mode calcul. Une fois amorcée, un écran dissuasif comme celui de la Figure 2 s'affiche, attirant l'attention de l'utilisateur et l'invitant à redémarrer la machine :

[2]LCG : Large Hadron Collider (LHC) Computing Grid, grille construite pour traiter les données du Grand Collisionneur de Hadrons, installé au CERN à Genève.

Figure 2 - Écran dissuasif

Après la mise en place des configurations réseau, les systèmes d'exploitation *calcul* ont été installés sur les 80 machines par clonage. Les outils systèmes et calcul sont identiques sur le site principal et sur le site Enseignement, assurant une *compatibilité des exécutables* déjà compilés par les utilisateurs.

Un des éléments importants des systèmes de gestion des clusters est fondamentalement différent par rapport à l'existant : le **gestionnaire de ressources et l'ordonnanceur**. Le gestionnaire de ressources est le logiciel chargé de démarrer et de contrôler des travaux en environnement *calcul*. Il obéit à l'ordonnanceur, qui choisit quels travaux démarrer en fonction d'une politique d'exploitation donnée. L'association de ces deux logiciels forme un **gestionnaire de files d'attente**.

Afin de satisfaire la contrainte du coût logiciel faible, nous avons testé et installé plusieurs gestionnaires de files d'attente libres et gratuits en lieu et place de la solution propriétaire (LSF de Platform Computing) utilisée jusqu'à présent. Si cela impose aux utilisateurs de modifier leurs scripts de soumission de travaux, la puissance de calcul potentiellement accessible et l'intérêt pour nous de maquetter de nouveaux gestionnaires de files d'attente nous ont convaincus de procéder au changement.

Nous avons ainsi testé sur cette configuration :

- Sun Grid Engine, gestionnaire de ressources et ordonnanceur ;
- Slurm (Simple Linux Utility for Ressource Management), gestionnaire de ressources ;
- Torque, gestionnaire de ressources ;
- MAUI, ordonnanceur.

Lors du test de ces outils, nous avons tout particulièrement vérifié l'intégration d'OpenMPI, la souche de MPI utilisée sur nos machines de calcul. Nous avons vérifié que les gestionnaires de ressources étaient capables de :

- démarrer une application MPI sur les nœuds choisis ;
- prendre le contrôle de l'application, c'est-à-dire arrêter et redémarrer les processus. L'arrêt des processus se fait par envoi de signaux SIGTSTP et SIGSTOP, le redémarrage par SIGCONT. Cette fonctionnalité est utilisée pour implémenter la préemption d'une application par une autre. Pour une application parallèle sur un cluster, le problème est surtout de s'assurer que *tous* les processus de l'application recevront les signaux en question. En général, les gestionnaires de ressources s'assurent de ceci en mettant en place des « wrappers » autour de rsh ou de ssh afin d'enregistrer les PID des processus démarrés sur les machines distantes. Avec les versions testées, seul Sun Grid Engine a échoué au test.

4 Bilan

Nous présentons le bilan de cette expérience sous trois aspects : logiciel, matériel et utilisation des ressources. En premier lieu, sans centraliser autre chose que des procédures d'administration, notons que nous avons pu mutualiser des ressources.

4.1 Aspect logiciel

Le choix de conserver l'image système existante était pertinent, puisqu'il a permis d'éviter une recompilation des applications de calcul spécifique à cette nouvelle plate-forme.

Nous avons acquis de l'expérience sur des gestionnaires de files d'attente sur une configuration réelle. Avec ces différents logiciels, nous avons pu mettre en œuvre des politiques d'exploitation prenant en compte des priorités spécifiques liées à notre fonctionnement *mutualisé*. Ces politiques étaient auparavant mises en œuvre via des logiciels commerciaux.

Valider les logiciels libres dans le cadre de notre gestion des ressources de calcul constitue le principal bénéfice de notre travail.

Ainsi, en dehors bien évidemment du temps nécessaire à la réalisation du projet (quelques jours.homme), aucun coût logiciel n'est à signaler.

4.2 Aspect Matériel

Le seul coût matériel à noter est l'acquisition d'un commutateur Gigabit d'Infrastructure pour le site principal, permettant de diffuser les VLAN avec plus de souplesse.

4.3 Utilisation calcul

Le cluster mis en place n'est pas forcément très performant pour des applications parallèles très communicantes en raison de la simplicité du réseau d'interconnexion des machines de la salle de ressources – réseau Gigabit Ethernet –. Pour l'application PMEMD [4], nous avons observé un ralentissement au delà de 4 processeurs, ralentissement qui ne se présente pas sur les machines disposant de réseaux dédiés.

Ce cluster peut donc servir pour des applications soit sérielles soit parallèles et peu communicantes.

En raison de l'utilisation des machines en journée pour l'enseignement, nous pouvons prévoir qu'au pire, des créneaux de 12h sans interruption seront disponibles la nuit. Il s'agit là d'un niveau de service *plus faible* que ce que nous pouvons offrir sur le site principal (créneaux de 3 jours garantis). Néanmoins, compte tenu du nombre de machines disponibles et de l'utilisation modérée des salles de ressources, nous pouvons espérer qu'une partie des travaux soumis se termineront correctement dans le temps imparti (cf 2.2.2).

Ce cluster a néanmoins pu être utilisé comme puissance de calcul de secours. En effet, lors d'une panne de ressources de calcul situées sur un 3ème site, un utilisateur a témoigné de son impatience et de la gène qu'il ressentait dans le déroulement de ses travaux. Nous avons pu lui proposer d'utiliser les machines de la salle de ressources. Il nous a suffit de démarrer le bon gestionnaire de ressources et les travaux bloqués ont pu s'exécuter.

Enfin, durant l'été 2009, ces machines ont été utilisées intensivement par des chercheurs en mécanique des fluides ayant besoin de réaliser des études de cas séquentielles (parameter steering).

Le passage en production de ces ressources ne s'est pas fait en raison :

- d'incertitudes quant à l'affichage politique de ces ressources ;
- d'incertitudes quant à la pérennisation des procédures d'administration des machines (3 OS = complexité de maintenance, changement d'administrateur système sur le site, simplification globale de la gestion des salles de ressources décidée politiquement) ;
- de l'arrivée en nombre de nouvelles ressources de calcul au milieu du projet.

Néanmoins ces nouvelles ressources ont bénéficié directement des enseignements logiciels mentionnés ci-dessus, comme nous allons le voir dans la suite de l'article.

5 Les mêmes idées... en production !

5.1 Genèse de l'opération

En Septembre 2008, nous avons été mandatés pour acquérir plusieurs clusters de calcul mutualisés. Les participants à cette opération sont :

- l'Institut de Chimie de Strasbourg (CNRS) ;
- l'INSA de Strasbourg ;
- l'Université de Strasbourg.

Mise en place en Février 2009 et inaugurée en Juillet 2009, la configuration acquise est composée de 64 serveurs bi-processeurs quadri-cœurs (Opteron Shangaï / 2.7 Ghz).

Le coût global de 165 k€ se décompose en :

- 17 k€ de l'UdS pour la partie stockage ;
- 34 k€ de l'UdS pour la partie calcul ;
- 57 k€ du CNRS ;
- 46 k€ de l'INSA ;
- 11 k€ du projet Européen Euforia pour les serveurs frontaux.

5.2 Matériel et réseau

Vu la taille de l'ensemble et les caractéristiques de la salle machine du site principal, il n'était pas possible d'y installer ce nouveau cluster. En particulier, les climatiseurs nécessaires à la production des 50 kw de puissance froid n'auraient pu rentrer dans la salle.

Forts de l'expérience multi-sites acquise précédemment, nous avons proposé l'installation de ces nouveaux clusters dans une nouvelle salle machine, située sur un autre site. Nous avons mis en place un VLAN, sans 802.1X cette fois-ci. Comme le précédent, ce VLAN permet de faciliter l'intégration du nouveau cluster à la configuration existante.

Au vu du nombre de machines présentes sur ce nouveau site, il a été décidé de le transformer en site principal ce qui a impliqué le déplacement de la baie de disques, afin de rapprocher les données du lieu d'utilisation où se trouvent la majorité des cœurs. Le serveur d'annuaire se trouve également sur ce site, tout comme le serveur de licences.

Nous avons depuis créé un nouveau VLAN transversal afin de procéder à une renumérotation IP des nœuds de calcul nous permettant d'homogénéiser l'administration système de nos serveurs.

Sur ce nouveau site, les machines sont intégralement administrables à distance, de l'allumage jusqu'à l'extinction, en passant par le suivi de la séquence de démarrage.

5.3 Logiciels

L'utilisation de gestionnaires de files d'attente libres, Torque + Maui, a permis d'économiser 17 k€ sur le coût global mentionné plus haut. Torque a été retenu plutôt que Slurm parce que plus répandu dans la communauté, y compris localement sur notre campus. Si Slurm paraissait séduisant par sa bonne prise en compte des processeurs multi-cœurs dans le placement des processus sur les machines, nous pensons qu'il aurait compliqué inutilement la tâche des utilisateurs ayant accès à plusieurs clusters. De plus, le jeu de commandes *batch* était trop restreint et peu intuitif d'après notre expérience.

Sur ces nouvelles ressources, les politiques d'exploitation reposent sur :

- un partage des ressources entre contributeurs au prorata du pourcentage de financement des machines. À la date de rédaction de l'article, les contributeurs peuvent soumettre des travaux de durée infinie ;
- un partage des ressources avec le reste de la communauté en offrant des créneaux pour des travaux courts (jusqu'à 24h), non préemptés.

La mise en place de ces logiciels découle directement de l'expérience acquise.

6 Conclusion

Notre expérience montre qu'il est possible de mutualiser des ressources de calcul et de les regrouper logiquement sans les rassembler géographiquement. Cela peut s'avérer intéressant pour lever certains freins psychologiques décrits dans [1]. Néanmoins, le mode de déploiement des images doit être simplifié. Nous pensons que notre configuration à base de VLAN peut s'appliquer si des ressources importantes sont disponibles sur un site dont on maîtrise l'infrastructure réseau. De même, en production, les procédures d'administration à distance doivent être fiables (la distance est l'ennemie de l'administrateur système).

Bibliographie

[1] Romaric David, Mutualisation des ressources de calcul à l'Université Louis Pasteur : Un bilan Dans *Actes du congrès JRES2007*, 3(10) : 133-139, Strasbourg, Novembre 2007.

[2] Logiciel CiGri : http://cigri.imag.fr/

[3] Logiciel ComputeMode http://computemode.imag.fr/mediawiki/index.php/ComputeMode_Grid_Manager

[4] Application PMEMD, Particle Mesh Ewald Molecular Dynamics, http://www.hpcx.ac.uk/research/chemistry/pmemd.html