Impacts organisationnels et techniques d'un noeud de grille dans un laboratoire du CNRS

Denis PUGNÈRE
Institut de Physique Nucléaire de Lyon
Domaine Scientifique de la Doua, Bât DIRAC
4, rue FERMI, F-69622 Villeurbanne cedex

Guillaume BAULIEU
Institut de Physique Nucléaire de Lyon
Domaine Scientifique de la Doua, Bât DIRAC
4, rue FERMI, F-69622 Villeurbanne cedex

Yoan GIRAUD
Institut de Physique Nucléaire de Lyon
Domaine Scientifique de la Doua, Bât DIRAC
4, rue FERMI, F-69622 Villeurbanne cedex

Résumé

Nous allons explorer dans cet article les implications qui découlent de l'installation et de l'exploitation d'un nœud de grille dans un laboratoire du CNRS de taille moyenne. Un nœud de grille est une variante particulière d'un cluster de calcul. Son architecture technique est complexe car composée d'éléments multiples qui, dans leur ensemble, constituent un plateau technique significatif à l'échelle d'un laboratoire. Les impacts liés à son exploitation sont aussi bien d'ordre technique, que financier, humain ou organisationnel. Plus précisément : Quels sont les implications pour le service informatique du laboratoire ? Quels sont les besoins en formation ? Quelles sont les compétences à acquérir ? Quelles sont les contraintes en terme de disponibilité du personnel vis-à-vis du nœud de grille ? Quels sont les coûts de fonctionnement ? Quels sont les coûts cachés ? Mais aussi, quels sont les bénéfices pour le laboratoire, pour les équipes de recherche, ou encore pour le service qui maintient ce nœud de grille ? Sur la base de notre expérience, nous apporterons quelques réponses à ces questions.

Mots clefs

Cluster ; grille de calcul ; organisation ; EGEE ; LCG ; coût ; infrastructure ; formation ; exploitation ; support aux chercheurs

1 Introduction

À coté des grands centres de calcul français (CC-IN2P3, CEA, CINES, IDRIS), certains laboratoires de recherche français (CNRS, universitaire, grandes écoles, EPST) se sont regroupés pour créer des structures de type mésocentres[1] (31 sites), mais il existe également un certain nombre de laboratoires qui, pour leur propre besoins, se sont dotés de moyens de calcul locaux de tailles très variées. Certains de ces laboratoires participent aux différents grands projets de calcul que sont les grilles de recherche comme GRID5000[2] (18 laboratoires) ainsi que les grilles de production EGEE[3] (15 laboratoires) et LCG[4] (12 laboratoires).

Une étude de février 2009 portant sur le calcul dans les laboratoires de recherche ayant pour objet de réaliser « un constat sur les pratiques et l'utilisation actuelle et future des ressources liées au calcul scientifique, en particulier par rapport aux options d'externalisation » nous éclaire sur les différents aspects, les moyens et sur les particularités du calcul au sein des laboratoires.

À notre connaissance, aucune publication n'a abordé les changements ou évolutions à la fois humains, techniques mais aussi financiers que constituent l'installation d'un cluster de calcul au sein de petites structures.

---

[1]Mésocentre :Un ensemble de moyens humains, de ressources matérielles et logicielles issus de plusieurs entités (EPST, Universités, Industriels). Liste des mésocentres réalisée par le Groupe Calcul : http://calcul.math.cnrs.fr/spip.php?rubrique6

[2]GRID5000 : liste des laboratoires participant au projet grille de recherche : https://www.grid5000.fr/mediawiki/index.php/Grid5000:Laboratories

[3]EGEE : Enabling Grids for E-sciencE, Liste des laboratoires participant au projet EGEE : http://goc.grid.sinica.edu.tw/gstat//France.html (GSTAT)

[4]LCG France : Liste des laboratoires français participant au projet LHC Computing Grid : http://lcg.in2p3.fr/wiki/index.php/French_Sites

Nous allons explorer dans cet article les implications humaines, techniques mais aussi organisationnelles qui découlent de l'installation et de l'exploitation d'un cluster de calcul dans un laboratoire du CNRS de taille moyenne.

1.1 Contexte

Le contexte de cette étude se fait dans un laboratoire, l'Institut de Physique Nucléaire de Lyon, où travaillent environ 220 personnes, effectif réparti entre chercheurs, ingénieurs et techniciens. L'IPNL est une Unité Mixte de Recherche agissant sous la double tutelle de l'Université Claude Bernard Lyon I[5] et de l'Institut National de Physique Nucléaire et de Physique des Particules[6] du CNRS. Les activités de l'IPNL visent à étudier les propriétés des composants subatomiques de la matière ainsi que leurs interactions. Laboratoire essentiellement de physique expérimentale, ses thématiques de recherche sont variées puisqu'elles concernent la physique des particules et des astroparticules, la matière nucléaire et les interactions ions/agrégats-matière. Certaines des équipes de recherche de l'IPNL sont impliquées dans les expériences internationales Alice et CMS du LHC[7], l'accélérateur de particules du CERN[8].

1.2 Le renouvellement du cluster de calcul

Lors du renouvellement de la ferme de calcul[1] de l'IPNL, nous nous sommes posés la question du type de cluster à implémenter pour répondre aux besoins locaux des équipes de recherche. Nous avions plusieurs possibilités : mettre en oeuvre un cluster classique tel que l'on en retrouve dans différents laboratoires ou centres de calcul : Cette configuration ne pouvait répondre qu'à une partie seulement des besoins de certaines équipes de recherche du laboratoire, en particulier les équipes de recherche impliquées dans l'analyse des données issues du LHC ne pourraient pas utiliser ce cluster. Nous avions aussi la possibilité de mettre en oeuvre un noeud de grille qui pouvait répondre également à certains besoins, mais pas tous.

Un noeud de grille est une variante d'un cluster de calcul dont la particularité est de s'insérer dans une infrastructure de plus grande échelle : La grille. Le fonctionnement de l'ensemble des éléments du noeud est géré par un intergiciel. Il s'agit d'un ensemble de packages implémentant des services standardisés. Ces services sont vus comme une interface entre le noeud local et les utilisateurs, mais aussi entre les autres noeuds distants implémentant le même intergiciel.

Nous avons opté pour implémenter un noeud de grille. Il a été tout naturel dans notre cas d'utiliser l'intergiciel Glite[9]. Cet outil a été développé dans le projet européen EGEE afin de mettre en commun les ressources de calcul des différents centres en Europe, il est également utilisé par le projet WLCG[10] dans le cadre de l'exploitation des données issues des 4 expériences du LHC (Alice, Atlas, CMS et LHCb). Le laboratoire a donc décidé de devenir noeud de grille EGEE et Tier-3[11] LCG en 2006.

L'intergiciel Glite définit plusieurs éléments distincts nécessaires pour le fonctionnement d'un cluster :

−Les machines de service : gestionnaire de tâches, serveurs de stockage, système d'information...

−Les machines de calcul (*Workers-Nodes*).

On commence à s'apercevoir que pour le démarrage d'un noeud de grille quelconque de type *Glite* nous avons besoin au minimum d'un *CE*, d'un *BDII*, d'une *UI*, d'un *SE* et d'au moins un *worker-node* (avec un seul *worker-node*, ce n'est pas encore un cluster !). Les autres services ne sont pas obligatoires, mais très souvent nécessaires pour l'exploitation du noeud : le serveur de catalogue, les serveurs (ou services) nécessaires pour le monitoring, un autre pour le déploiement des machines de calcul, un dépôt pour les différentes versions des fichiers de configuration...

1.3 L'impact du choix

Les implications de l'implémentation d'un noeud de grille ne sont pas neutres dans un laboratoire. Les impacts sont aussi bien d'ordre technique, mais aussi d'ordre financier, humain et organisationnel. Plus précisément : Quels sont les impacts sur le service informatique d'un projet de cluster de calcul ?

−En terme de ressources humaines : Quels sont les besoins en formation ? Quelles sont les compétences à acquérir ? Sur les contraintes organisationnelles : Quelles sont les contraintes en terme de disponibilité du personnel vis-à-vis du noeud ?

[5]UCBL : http://www.univ-lyon1.fr

[6]IN2P3 : http://www.in2p3.fr

[7]LHC : Large Hadron Collider : http://www.lhc-france.fr

[8] CERN : Organisation Européenne pour la recherche nucléaire : http://www.cern.ch

[9]gLite Middleware : http://glite.web.cern.ch

[10]Worldwide LHC Computing Grid : http://lcg.web.cern.ch/LCG

[11]Le Tier-3 de l'IPNL est décrit sur http://www.ipnl.in2p3.fr/spip.php?rubrique114

−Sur les aspects financiers : Quels sont les coûts d'investissement d'un cluster de calcul ? Quels sont les coûts de fonctionnement ? Quels sont les coûts cachés ?

−Sur les aspects techniques : Quelle infrastructure est nécessaire pour accueillir un noeud de grille ? Le réseau, la climatisation et l'installation électrique seront-il suffisants pour permettre l'extension du noeud de grille ? Quels nouveaux services seront nécessaires pour l'exploitation du noeud ?

Sur la base de notre propre expérience, nous allons aborder ces différents aspects.

Afin d'alléger le texte de ce document, à partir de maintenant, le noeud de grille de l'IPNL sera appelé *Tier3*.

2 Ressources humaines

À l'IPNL, le service informatique est actuellement composé de 13 personnes. L'équipe « grille » comprends 4 personnes qui ne travaillent pas à plein temps sur le Tier3. Nous estimons que leur contribution totale à cette activité est d'environ 1,4 ETP[12] sur une année.

Le service informatique d'un laboratoire étant généralement assez sollicité pour différentes tâches (maintenance de parc, administration des serveurs...), il est clair que l'installation d'un noeud de grille induit une charge de travail non négligeable, celle-ci est répartie en plusieurs tâches :

−La gestion administrative du projet (achats, appels d'offre, préparation, supervision de travaux...),

−L'installation ou la mise à niveau de l'infrastructure (salle serveurs, alimentation électrique, climatisation, racks, réseau...)

−L'acquisition de nouvelles connaissances (suivi de formations, veille technologique...),

−l'installation et le déploiement du cluster (installation matérielle et déploiement logiciel),

−la mise à jour de l'intergiciel Glite,

−l'exploitation quotidienne (administration, installation de nouvelles applications, modification des configurations),

−la résolution des problèmes divers (défaillances matérielles, problèmes espace disque, résolution de bugs, instabilités...),

−participation aux réunions et conférences des projets liés à la grille.

2.1 Formations

Dans le déploiement d'un cluster, les équipes informatiques peuvent avoir le besoin d'acquérir des compétences complémentaires, celles-ci n'étant pas forcément nécessaires dans l'exploitation quotidienne d'un parc informatique.

Les grilles de calcul sont un concept particulier qui met en oeuvre un grand nombre de techniques souvent complexes, de haut niveau et faisant appel à un éventail de compétences dont le spectre est très large : connaissances dans les systèmes d'exploitation de type Linux (leur fonctionnement, les librairies systèmes, l'optimisation, leurs limites), en programmation de scripts (*shells*, *PERL*, *Python*), en réseau haut débit (limites des protocoles, optimisation de buffers réseaux, agrégation de liens...), en filtrage réseau (ports réseau, services de grille), en sécurité informatique (gestion des certificats, authentification, gestion et recherche dans les logs), en stockage (architectures de stockage de grande capacité, optimisation de l'accès au stockage, systèmes de fichiers), calcul distribué (*Multi-threading*, MPI, *OpenMP*...). Très souvent, des compétences en déploiement d'un grand nombre de machines, en monitoring de services, en gestion de systèmes de « batches » ou encore en ordonnancement des tâches dans le cluster sont également nécessaires.

[12]ETP : équivalent temps plein : somme des contributions de l'ensemble des personnes impliquées dans un projet.

Au delà des compétences acquises, l'exploitation quotidienne d'un cluster amène à supporter de nouvelles applications pouvant avoir des profils très hétérogènes. Certaines nécessiteront un système de fichier capable d'être efficace même avec plusieurs millions de fichiers et pour d'autres il faudra une architecture de stockage capable de soutenir de bonnes performances en entrées sorties sur plusieurs centaines de connexions.

Ces compétences nécessitent un besoin en formation assez important. Nous trouvons assez régulièrement des formations à destination des utilisateurs de la grille, par contre, les formations des administrateurs de noeuds de grille sont plus rares. Quand ces formations existent, elles traitent seulement qu'une partie des compétences nécessaires.

Dans le métier de l'administration système et réseau, il est assez courant de voir les personnels s'auto-former par de la veille « technologique » en s'abonnant par exemple aux listes de diffusion traitant des sujets techniques, ou alors en consultant des sites web spécialisés, de forums ou bien lors d'échanges directs avec les collègues.

Les projets EGEE et LCG organisent des réunions qui nous permettent d'interagir avec les administrateurs des autres sites et de partager nos expériences respectives (implémentation d'un service, difficultés rencontrées, etc...). Ces rencontres sont très bénéfiques, elles nous procurent un gain de temps important dans la prise en main des nouveaux services du noeud de grille.

Par ailleurs, les mécanismes mis en oeuvre évoluent, certains sont abandonnés, d'autres les remplacent. Nous sommes sans arrêt confrontés (quasiment à chaque réunion) à de nouveaux acronymes apparaissant régulièrement. Ces acronymes dont la quantité croit quasiment aussi vite que le nombre de *CPU* disponibles sont un frein à la compréhension de ces mécanismes.

De plus, pour assurer la surveillance, la prise en charge des incidents, la notification des périodes de maintenance de notre noeud et comme tous les autres sites EGEE ou WLCG, nous utilisons aussi les outils mis à disposition de la communauté comme *GSTAT*[13], *APEL*[14], *GRIDVIEW*[15] pour la comptabilité (*accounting*), le *CIC PORTAL*[16], *GGUS*[17]... pour l'information sur le site ou pour la notification des périodes de maintenance et les *dashboard*[18] *CMS*, *Alice* ou *MonALISA*[19] pour la visualisation de l'état du site par les expériences LHC. Une longue période d'apprentissage est nécessaire pour s'approprier ces différents outils.

Il en résulte que la quantité d'informations à assimiler est très importante, d'autant plus qu'il n'est pas toujours aisé de savoir où la trouver. Il est très difficile pour une personne seule de prendre en charge la totalité des aspects liés à l'exploitation d'un noeud de grille ; pour que cela soit possible, il faudrait qu'elle maîtrise le fonctionnement approfondi des systèmes d'exploitation, des réseaux, des mécanismes de sécurité, des calculs d'expériences de physique lancés sur la grille, du stockage, du déploiement, du monitoring et bien d'autres encore ce qui relève de la science fiction pour la plupart d'entre nous.

2.2 Disponibilité et contraintes

Participer aux projets de grilles EGEE et LCG induit la participation aux nombreuses conférences (EGEE, LCG-france), réunions (hebdomadaires), visio-conférences (bi-hebdomadaires ou mensuelles), groupes de travail (sécurité, monitoring...) ce qui, comme nous l'avons vu précédemment, est très enrichissant d'un point de vue technique, mais est également très chrono-phage.

Nous essayons dans la mesure du possible de traiter les différents problèmes à plusieurs personnes : cela nous permet d'intervenir sur l'exploitation du Tier3 en l'absence de l'un d'entre nous. Ainsi par exemple sur les quatre personnes de l'équipe grille, deux interviennent sur la configuration du réseau, trois suivent la configuration et le déploiement des machines, trois suivent également la configuration du stockage et deux suivent la configuration du site en interface avec les expériences *CMS* et *Alice*.

Dans le projet LCG, les noeuds de grille doivent passer une étape de « certification » avant d'entrer en production. Une fois en production, le site est tenu de respecter un certain nombre d'engagements en terme de disponibilité et de fiabilité. Même si juridiquement aucune contrepartie ne sera exigée en cas de non respect des engagements, une fois « certifié », il n'est plus envisageable que le site sorte de production à cause d'un manque de disponibilité ou de fiabilité des services.

---

[13] http://goc.grid.sinica.edu.tw/gstat/IN2P3-IPNL

[14] APEL (Accounting Processor for Event Logs) : http://www3.egee.cesga.es/gridsite/accounting/CESGA/egee_view.html

[15] GRIDVIEW : http://gridview.cern.ch

[16] CIC PORTAL (Communication Interface for Central Operations) : http://cic.gridops.org

[17] GGUS : Global Grid User Support : https://gus.fzk.de

[18] http://dashboard.cern.ch

[19] http://pcalimonitor.cern.ch

Nous avons ici une contrainte en terme de disponibilité et de fiabilité de service qui repose sur le service informatique du laboratoire, celle-ci se traduit par une organisation qui consiste à ce que le site soit surveillé en permanence par un membre du service, même pendant les périodes de congés. Cela implique que les membres de l'équipe ne prennent pas leurs vacances en même temps ou qu'un de ses membre soit toujours à même d'intervenir rapidement au laboratoire en cas de problème technique. Actuellement, cette organisation repose sur la bonne volonté de l'équipe.

D'autre part, comme les grilles de calcul sont par nature distribuées, les noeuds de grille sont inter-dépendants les uns des autres. Pour que son propre fonctionnement soit optimal, il est courant qu'un site nécessite que deux ou trois autres sites soient parfaitement opérationnel. Dans notre cas, pour que notre noeud de grille soit opérationnel, nous utilisons (et sommes donc dépendants) des services répartis sur les sites suivants : CC-IN2P3, GRIF et LAPP.

3 Aspects financiers

Alors qu'il n'est généralement pas très difficile d'estimer la taille initiale d'un cluster, il est nettement plus difficile de prévoir son évolution : quelle sera sa taille dans 1 an ou dans 4 ans ? Quelle sera l'évolution des besoins des équipes de recherche ? D'après notre expérience, la plupart des équipes de recherches n'ont qu'une très vague idée de leurs besoins en matière de calculs et de stockage. Il arrive qu'une équipe n'ayant initialement exprimé aucun besoin particulier revienne quelques mois après en demandant plusieurs To de stockage afin d'héberger de nouvelles données.

Répondre aux besoins exprimés par les équipes de recherche revient d'une part à prendre en compte les besoins exprimés en fonction de la technologie à notre disposition, mais aussi à estimer la capacité nécessaire pour répondre à ces besoins en fonction des évolutions technologiques des futurs processeurs ou d'évolution des capacités de stockage dans 1 ou 4 ans.

Le projet ne se limite pas à acheter un nombre de serveurs de calcul fournissant la puissance de calcul visée. Nous l'avons vu dans le choix du type de cluster (noeud de grille dans notre cas), un certain nombre de serveurs sont nécessaires pour que le noeud soit fonctionnel, mais ce n'est pas tout ; il faut aussi prendre en compte toute l'infrastructure nécessaire pour héberger le noeud. Dès lors, quels sont les coûts réellement associés à l'installation d'un noeud de grille ?

Le modèle de coût utilisé par le projet LCG-France[20] est basé sur le prix d'achat en € / *KSI2K* pour la *CPU* (en fonction de la puissance *CPU* normalisée selon la suite de benchmark *SPEC-CPU*[21] ou selon la normalisation WLCG[22]) et en € / Go utile pour le stockage (c'est l'espace de stockage disponible qui est pris en compte après mise en place du *RAID* éventuel et de la perte due au système de fichiers). Le modèle de coûts prends également en compte certains composants annexes comme les cartes d'alimentation, le rack, les interfaces du commutateur réseau correspondant, le serveur de disques pour le stockage , la garantie de 3 ans pour le *CPU* et de 4 ans pour le disque. Par contre, le modèle de coûts ne prends pas en compte les machines de service de grille pourtant nécessaires au fonctionnement du site, ni l'installation électrique, frigorifique ni les coûts opérationnels tels que les fluides, électricité.

Par exemple, estimer le coût d'un serveur 1U (*worker-node*) disposant actuellement de 2 *CPU Quad-cores* reviendrait à prendre en compte le coût :

- de 1/32eme d'un *rack* 32U,

- de 1 alimentation électrique (généralement un *PDU* partagé entre plusieurs *worker-nodes*),

- d'une (ou 2) connexion(s) réseau gigabit,

- de la bande passante minimum garantie par *worker-node* jusqu'au stockage. Par exemple : pour garantir un débit de 500Mb/s par *worker-node*, il faudrait prévoir une bande passante de 10Gb/s pour 20 *worker-nodes*,

- du pourcentage de l'espace de stockage associé : Dans les expériences du LHC, les spécifications du rapport entre la puissance de calcul et l'espace de stockage (*KSI2K* / To) est généralement de l'ordre de 3 ou 4. Par exemple pour une puissance de calcul de 100 *KSI2K* et un rapport de 4, l'optimum est de mettre à disposition du cluster un volume de stockage de : 100 / 4 = 25 To

À ceci, il faut ajouter l'installation nécessaire pour :

[20] Modèle de coûts LCG-France pour les Tier-2 : https://edms.in2p3.fr/document/I-013175/1

[21] SPEC-CPU : http://www.spec.org/benchmarks.html

[22] HEP-SPEC06 Benchmark : https://hepix.caspur.it/benchmarks

−fournir un courant secouru propre au cluster (coût d'un onduleur ou de son extension),

−dissiper la chaleur produite par ces *worker-nodes* (coût d'une climatisation ou de son extension),

En matière d'investissement, nous arrivons donc au bilan suivant :

−installation ou extension de la salle serveur : faux plancher éventuel, alarme, protection incendie, sécurisation,

−installation ou extension du système électrique : onduleur, cheminement des câbles jusqu'au local serveurs, boitier(s) électrique(s) dans la salle serveurs, départs électriques vers les baies,

−installation ou extension d'une climatisation (quelle que soit la technologie utilisée : *free-cooling*, détente directe...),

−achat des baies,

−serveurs de service : généralement au moins 5 (ordonnanceur, *serveur de stockage*, *serveur d'accès interactif*, *système d'information*, monitoring, déploiement...),

−les commutateurs réseau en prenant en compte le rapport les uplinks nécessaires pour l'accès au stockage : un à deux commutateurs par baie puis un commutateur de concentration,

−les *worker-nodes*,

−le stockage : *SAN* (baie de stockage et serveurs d'entrées / sorties), *DAS* (serveurs de disques)...

Nous devons également compter les coûts récurrents d'exploitation tels que : les contrats de maintenance (serveurs, climatisation, onduleur, stockage), la facture électrique et les différents coûts de renouvellement du matériel (remplacement des batteries de l'onduleur, renouvellement des serveurs et du stockage...). Les coûts de renouvellement peuvent être estimés en fonction du différentiel du coût d'achat d'une certaine puissance de calcul (ou d'une certaine quantité de stockage) d'une année sur l'autre. Ainsi on peut par exemple estimer que le coût d'achat du matériel de même capacité sera 30% moins cher pour l'année N+1.

Si nous prenons le coût d'un *worker-node* seul (par exemple 2500€) mais également l'ensemble des coûts environnés (infrastructure nécessaire pour que ce *worker-node* rende un service optimal), les coûts récurrents d'exploitation et enfin la main d'oeuvre pour mettre en place et maintenir le service, il apparaît clairement que le coût d'achat du *worker-node* ne représente qu'une petite partie du coût total de possession de ce *worker-node*.

Au delà des considérations de coûts d'achat ou de maintenance, un certain nombre de considérations techniques entrent en compte dans l'exploitation d'un noeud de grille, nous allons détailler ces considérations dans la suite de ce document.

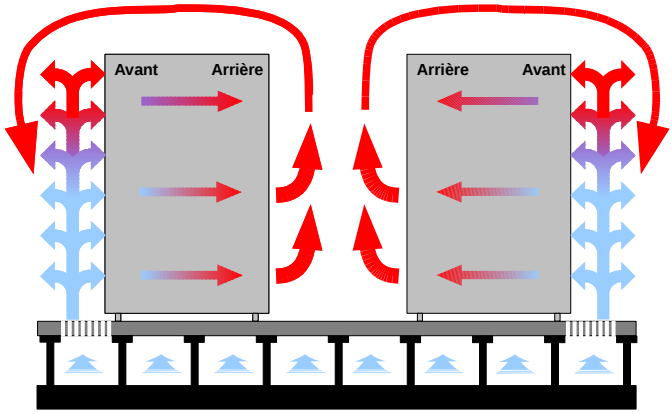
4 Aspects techniques

L'installation d'un noeud de grille au sein d'une infrastructure informatique implique souvent un accroissement important du nombre de machines à gérer. Il est essentiel d'évaluer l'infrastructure informatique nécessaire pour installer ces nouvelles machines et supporter les éventuelles évolutions. Ces questions sont souvent considérées à tord comme secondaires par rapport à l'achat du matériel informatique, les problèmes étant ensuite découverts au fur et à mesure de l'évolution de l'installation. Le but de cette partie est de soulever différents points auxquels il est important de réfléchir en amont.

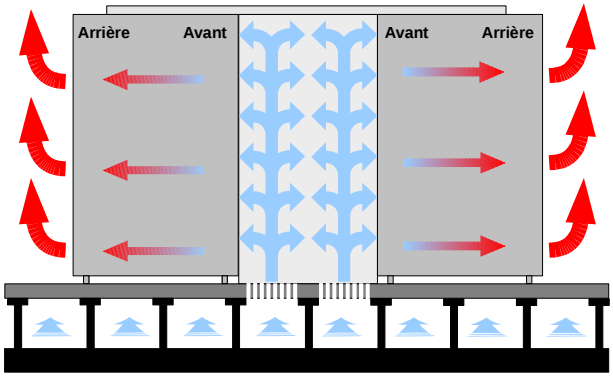
4.1 Climatisation

La présence de plusieurs dizaines/centaines de machines dans une pièce dégage une quantité de chaleur pouvant être considérable, un système de climatisation est donc indispensable. Celui-ci doit être dimensionné de façon à pouvoir dissiper la chaleur produite par les machines installées au démarrage du site mais également dans le futur. Il est souvent difficile, voire impossible, de faire évoluer un système de climatisation pour en augmenter la puissance, il faut bien souvent remplacer l'intégralité du système. Un sous-dimensionnement initial se traduit alors par une facture non négligeable lorsque l'on veut augmenter la taille du site.

La configuration de la salle serveur joue un rôle vis à vis de l'efficacité de la climatisation. Il peut être possible d'optimiser son fonctionnement par une étude des circuits d'air chaud/froid. La disposition des baies, l'existence de faux planchers ou faux plafonds, la création de couloirs chauds ou froids éventuellement cloisonnés sont autant d'éléments permettant d'améliorer le refroidissement des machines. Dans notre cas, nous avions initialement disposés les baies en deux rangés dos à dos (cf Dessin 1 page 7), avec une sortie d'air froid en façade créant ainsi deux couloirs froids en avant et un couloir chaud en arrière des baies. Après avis de plusieurs experts frigoristes il nous a été conseillé de retourner toutes nos baies de manière à n'avoir plus qu'un seul couloir froid isolé du restant de la pièce par des plaques de poly-carbonate (cf Dessin 2 page 7). Cette méthode optimise la circulation de l'air froid en l'obligeant à passer à travers les machines avant d'être repris dans le circuit de la climatisation. D'apparence anodine, cette opération nécessite tout de même l'arrêt complet du site ainsi qu'un re-câblage électrique et réseau, le genre d'intervention dont on se passe volontiers.



Dessin 1: Avant retournement des baies (plusieurs couloirs froids)



Dessin 2: Après retournement des baies (un seul couloir froid confiné)

De plus, le comportement parfois capricieux de ce genre d'installation (alarmes haute pression, basse pression, haute température, givrage du système...) nécessite une surveillance sérieuse, d'autant plus qu'un arrêt de fonctionnement de la climatisation peut avoir des conséquences dramatiques (dans notre cas une augmentation de la température pouvant aller jusqu'à deux degrés par minute dans la salle serveur!). Ceci nous a amené à installer un coupe-circuit électrique permettant de protéger l'infrastructure au delà d'une certaine température.

Il est important de se poser ces questions dès le début de la réflexion sur la création du site car suivant la technologie utilisée, des conduits devront être aménagés afin d'acheminer l'air ou l'eau aux systèmes de climatisation. Ceci va influer sur l'ampleur et les coûts des travaux mais aussi sur le choix de l'emplacement de la salle serveur (proximité d'accès à l'eau brute, conduits pour accéder au toit...).

4.2 Installation électrique

L'installation du matériel sensible et couteux que représente un noeud de grille nécessite une installation électrique saine. Afin de la garantir, un onduleur semble indispensable. Celui-ci peut également permettre, pour peu de disposer de suffisamment de batteries, de palier à une petite coupure de courant ou de permettre le démarrage d'un éventuel groupe électrogène. Tout comme la climatisation, il va falloir calibrer ce système au démarrage du site avec les mêmes contraintes : l'onduleur initial ne pourra évoluer que jusqu'à une certaine limite. Au delà un remplacement et donc un investissement conséquent seront nécessaires.

À titre d'exemple, notre consommation électrique est passée de 10KVA à 45KVA avec l'installation du Tier3. Les contraintes exercées sur l'alimentation électrique ne sont pas les mêmes, il est donc important d'analyser toute la chaine depuis l'alimentation principale jusqu'à la salle serveur afin de l'isoler et de garder le contrôle sur les niveaux de consommation. Il suffit par ailleurs d'un seul disjoncteur mal calibré dans un petit boitier électrique au fond d'un sous-sol oublié pour couper l'alimentation d'une salle flambant neuve...

4.3 Salle serveurs

La pièce accueillant le noeud de grille doit non seulement permettre l'installation du matériel initial et sa mise en opération mais également permettre l'évolution du site. L'idéal est évidemment de pouvoir installer le matériel dans une pièce capable d'accueillir toutes les évolutions futures. Il est difficile de réaliser des travaux (comme la destruction d'une cloison pour agrandir la pièce) alors que des machines sont en production, de même le déménagement dans une nouvelle salle nécessite des efforts considérables.

Les points vus précédemment peuvent également orienter le choix de la pièce : il faudra y faire arriver l'alimentation en eau (ou air) de la climatisation ainsi que l'alimentation électrique. L'accès à ces réseaux doit donc être le plus simple possible pour limiter les coûts et les risques de panne. L'aspect pratique ne doit pas non plus être oublié : des baies de machines devront être entrées ou sorties de cette pièce. Un baie pleine ayant une masse approximative de 700/800 kg, est-il facile de l'acheminer jusqu'à son emplacement final? Si la pièce est en étage y a-t-il un monte charge? Une marche ou un faux plancher à l'entrée de la salle? Une baie chargée sur un transpalette peut-elle passer par la porte?

Le poids total des baies a également sont importance. Si une ou deux baies ne posent pas problème, la charge au sol devient très conséquente lorsque le nombre de baies augmente. Le choix d'une pièce en étage peut alors devenir problématique.

Les choix concernant le système électrique, la climatisation et la salle serveurs sont essentiels car ils ont un impact direct sur les couts d'installation, la facilité d'exploitation, la fiabilité et la capacité d'évolution du système. Les exigences de chaque système peuvent être contradictoires, il est donc important de définir des priorités pour prendre les décisions les plus adaptées.

4.4 Réseau

Outre les machines de calcul et de stockage, il est important de ne pas négliger l'importance de l'infrastructure réseau. Selon le type de calculs exécutés sur le site, cet aspect peut devenir essentiel. Il est fréquent de voir des machines n'utiliser qu'un faible pourcentage de la puissance processeur car ceux-ci sont en attente de données arrivant trop lentement par le réseau. Dans notre cas, un coeur de réseau à 1Gb/s s'est avéré insuffisant, nous contraignant à évoluer vers un coeur de réseau à 10Gb/s nettement plus couteux .

4.5 Installation et maintenance des machines

Le facteur d'échelle qu'apporte souvent l'installation d'un noeud de grille impose une approche différente dans la gestion de l'installation et de la configuration des machines. Il n'est pas vraiment envisageable d'installer des dizaines, voire des centaines de machines une par une puis recommencer à chaque mise à jour... La mise en place d'un système d'automatisation devient incontournable.

A titre d'exemple, les deux outils les plus utilisés au sein du projet LCG sont *YAIM*[23] et *Quattor*[24]. Ayant opté pour *Quattor*, nous allons détailler un peu plus cette solution.

Quattor permet de définir des profils de machines contenant l'ensemble de la configuration : partitionnement des disques, système d'exploitation, liste des logiciels à installer, configuration de ces logiciels... Le profil décrit l'état final dans lequel doit se trouver le serveur. Installer une machine de calcul revient donc à écrire son profil. Ce travail n'est pas négligeable mais n'est à réaliser que pour la première machine : pour toutes les suivantes il suffira de faire correspondre un nom, une adresse *MAC* et le profil en question. L'installation à proprement parler revient ensuite à … allumer la machine. Celle-ci se connectera alors au serveur *Quattor* qui contient les profils, systèmes d'exploitation et logiciels à installer. La séquence d'installation se déroule alors automatiquement :

- Démarrage *PXE*, installation du système d'exploitation
- Re-démarrage
- Installation des logiciels et configuration
- Re-démarrage
- La machine peut être mise en production

Il est important de remarquer qu'une machine gérée par *Quattor* ne peut être administrée manuellement par la suite. Toute modification de la configuration (nouveau logiciel installé, changement d'un fichier de configuration...) effectuée sans en informer le serveur *Quattor* sera supprimée par celui-ci. La seule démarche possible est de modifier le profil de la machine sur le serveur puis de déployer la modification. Si ce protocole est contraignant, il permet cependant de garantir la cohérence du parc de machine. En effet la modification de la configuration d'une machine en particulier est à proscrire : il est très compliqué de garantir le fonctionnement ou de diagnostiquer un problème au sein d'un parc de machine dont la configuration est hétérogène. Il est essentiel qu'une modification apportée à une machine le soit à toutes les machines semblables. De plus *Quattor* garde une version de chaque configuration déployée, ce qui permet à tout moment de revenir à une configuration précédente.

Cet outil permet un gain en efficacité important dans la gestion du site mais nécessite un temps d'apprentissage relativement long pour le maitriser.

[23] http://yaim.info/

[24] http://www.quattor.org/

4.6 Monitoring du site

Outre l'installation et la maintenance des machines, il est également important de surveiller leur fonctionnement et de détecter au plus vite les dysfonctionnements éventuels. L'accroissement important du nombre de machines ainsi que les différents systèmes impliqués (onduleur, climatisation...) nécessite un effort de surveillance accru. Généralement tous ces systèmes peuvent être connectés au réseau: il est donc possible de mettre en place un système centralisé pour vérifier le fonctionnement et gérer les alertes. Les problèmes détectés peuvent aller d'un disque plein sur une machine de calcul, compromettant les travaux en cours sur celle-ci à une panne du système de climatisation compromettant l'intégrité physique de la salle serveur. Les alertes envoyées doivent donc être adaptées en fonction de la gravité de la situation : du simple mail de notification à l'arrêt automatique de machines en passant par l'envoi de *SMS*. Au delà de l'aspect technique de la surveillance et des alertes, c'est également l'aspect organisationnel et humain qu'il faut prévoir: qui doit intervenir et dans quel délai?

5 Conclusion

En retour des investissements importants que nécessite l'installation et l'exploitation d'un noeud de grille, le laboratoire peut espérer diverses retombées bénéfiques.

D'un point de vue infrastructure informatique, la plupart des modifications apportées pour le noeud de grille bénéficient directement à l'infrastructure du laboratoire. En tout premier lieu la réhabilitation, voire la construction de la salle serveur permet d'héberger les machines de service du laboratoire dans de bonnes conditions d'exploitation et de maintenance. De la même façon, le système de surveillance mis en place pour la grille peut être utilisé pour les serveurs du laboratoire. De façon générale, le noeud de grille impose la mise en place d'une infrastructure et d'outils d'administration performants qui, faute de temps et de moyens, ne sont pas toujours mis en place auparavant.

Pour les équipes de recherche du laboratoire, la présence d'un noeud de grille et du personnel capable de l'exploiter facilite largement l'accès à la grille. Cela permet tout d'abord la formation des étudiants en thèse du laboratoire à l'utilisation de ce type d'outils. De la même façon, il est beaucoup plus simple pour le personnel d'utiliser la grille lorsque les utilisateurs peuvent s'adresser à un correspondant local pour la formation ou la résolution des problèmes rencontrés.

La participation à un projet grille d'envergure nationale ou internationale permet également de développer les relations entre les services informatique des différents laboratoires participants de part les nombreuses réunions ou visio-conférences organisées. C'est également un moyen pour le laboratoire d'accroître sa visibilité et d'affirmer son investissement dans les collaborations scientifiques auxquelles il prend part.

6 Bibliographie

[1] : Noël GIRAUD, Christophe MARTIN, Une « fermette » de PCs, *Actes du congrès JRES1999*, http://1999.jres.org/articles/giraud-ahp-02-final.pdf