Une « fermette » de PCs

■ Noël GIRAUD, giraud@ipnl.in2p3.fr
Institut de Physique Nucléaire de Lyon

■ Christophe MARTIN, cmartin@ipnl.in2p3.fr
Institut de Physique Nucléaire de Lyon

L'équipe de théorie de l'Institut de Physique Nucléaire de Lyon a régulièrement besoin de ressources informatiques importantes pour des calculs de structure nucléaire. Ces calculs sont effectués sur le Cray de l'Idris, mais pour effectuer des tests, ils avaient obtenu un budget pour l'achat d'une station locale. Dans cette enveloppe budgétaire, une solution à base de PCs sous Linux a été chiffrée, puis retenue.

■ Introduction

Les théoriciens de l'Institut de Physique Nucléaire de Lyon ont une longue tradition de gros consommateurs de ressources informatiques. Ils ont longtemps été clients du Centre de Calcul de l'In2p3, particulièrement à l'époque où celui-ci était doté d'un ?mainframe? équipé de processeurs vectoriels. Après l'arrêt de cette facilité, lié au passage de ce centre à des fermes de stations sous Unix, certains se sont orientés vers l'utilisation du Cray de l'Idris. Ce centre de calcul correspond parfaitement à leurs besoins, tant du point de vue de la puissance de calcul que des moyens de stockage des résultats. Cependant, pour la mise au point des codes, le temps de retour d'une soumission est parfois long quand il s'agit de petits calculs.

Ils ont donc obtenu, dans le cadre d'une « AP », une enveloppe budgétaire d'environ 100 KF pour l'achat d'une station de travail locale ainsi que de l'environnement logiciel (compilateurs et librairies) permettant de préparer ces travaux, et éventuellement d'utiliser les temps morts, forcément importants, pour effectuer localement des calculs complets. La solution envisagée initialement correspondait à une station IBM de la famille RS6000 avec le compilateur Fortran du constructeur et la librairie ESSL.

■ Les contraintes

Pour trouver une configuration répondant aux besoins, les contraintes suivantes doivent être prises en compte :

- une puissance CPU importante. Bien qu'initialement il ne s'agisse que de travaux de test, des calculs qui demandent plusieurs minutes sur un Cray peuvent prendre facilement près d'une heure sur une machine moins performante. Le but étant de réduire le temps d'attente entre soumission et retour du résultat, il faut arriver à un ratio de puissance de l'ordre de 5 pour en retirer un bénéfice ;
- une mémoire importante. Les programmes concernés manipulent constamment des matrices de taille assez conséquente, l'espace mémoire requis se mesure en dizaines de Mega-octets et peut facilement atteindre 100 ou 200 Mega-octets ;
- un compilateur Fortran compatible avec celui du Cray, en particulier au niveau des options. Les travaux de test exécutés doivent pouvoir être soumis sur le Cray sans modification. La qualité du compilateur doit d 'autre part être suffisante pour que les résultats obtenus soient fiables ;
- une librairie mathématique. Les calculs concernés font régulièrement appel à des opérations courantes d'analyse numérique (inversions de matrices en particulier) et une librairie performante et fiable est nécessaire ;
- un système de batch. Plusieurs personnes doivent pouvoir accéder à ce système et y effectuer des calculs volumineux et répétitifs. Pour éviter un engorgement, particulièrement au niveau de la mémoire, un système d'ordonnancement des travaux est nécessaire ;
- un espace disque important. Des résultats intermédiaires volumineux (fonctions d'ondes) doivent pouvoir être stockés entre l'exécution des programmes successifs, un volume de quelques dizaines de Giga-octets doit donc être prévu. Par contre, une sauvegarde automatique de cet espace n'est pas vitale, ces données pouvant être assez facilement reconstituées.

■ La solution station de travail

La solution initialement envisagée consistait en une station de travail de type IBM RS6000 et les logiciels (système d'exploitation, compilateur Fortran et librairie mathématique) de ce constructeur. En dehors de la « tranquillité » que ce genre de choix procure (garantie contractuelle d'un service), il présente cependant quelques inconvénients :

- pour la somme envisagée, on doit se limiter à un ou deux CPUs,
- les compilateurs et librairies sont d'un prix élevé,
- les extensions mémoire et disques sont de même fort coûteux.

Une étude auprès d'autres constructeurs (en particulier Digital, qui s'appelait encore ainsi à cette époque et dont les processeurs offraient des performances séduisantes) nous ont montré que ce genre de solutions restait d'un prix élevé. De même, les diverses offres promotionnelles sur certaines configurations comportent toute une solution graphique aussi coûteuse qu'inutile pour notre propos.

■ La solution « PC sous NT »

Pour obtenir le plus de puissance de calcul au moindre prix, il est naturel de s'intéresser à la plate-forme « PC » qui en raison des volumes produits, de la standardisation des composants et de la concurrence féroce que se livrent les divers fournisseurs, atteignent des coûts hors de porté des autres solutions. De plus, depuis l'apparition de l'architecture du Pentium Pro, de nombreux tests ont montré que ces machines pouvaient parfaitement concurrencer, en termes de puissance de calcul, les autres architectures.

Malheureusement, la tendance des vendeurs à proposer ces plates-formes avec un système d'exploitation imposé, qu'il s'agisse de Windows 95/98 ou de sa version soit disant professionnelle Windows NT limite quelque peu l'enthousiasme. Ce système présente en effet un certain nombre d'inconvénients qui le rendent impropre à l'usage que nous envisagions :

- instabilité chronique, particulièrement en ce qui concerne les « fuites » de mémoire rendant l'exploitation continue de ces machines illusoire,
- système cryptique dont la configuration n'est possible que par son interface graphique, l'accès aux fichiers de configuration étant quasiment impossible,
- manque des outils qui font la richesse des divers Unix, à commencer par un langage de script,
- maintenance délicate du fait de l'impossibilité de se connecter à distance.

■ La solution « PC sous Linux »

Si l'on arrive à se débarrasser de Windows NT, il reste pour utiliser une plate-forme « PC » le choix parmi les diverses saveurs d'Unix dans le domaine public. Ayant une certaine expérience avec Linux et connaissant la qualité de certaines de ses distributions, c'est ce système que nous avons retenu. Cette solution présente de nombreux avantages :

- vu le prix des machines, on peut envisager une configuration bien plus importante. On peut en particulier envisager une véritable petite « ferme » avec un serveur et des machines spécialisées pour le calcul situées sur un réseau isolé,
- le système Linux est maintenant dans une phase de maturité qui le rendent tout aussi stable que les systèmes commerciaux et facile à maintenir, l'accès aux sources étant probablement la meilleure des garanties,
- la mémoire et les disques de ces machines ont atteint des prix qui rendent possible des configurations importantes pour un coût raisonnable,
- des compilateurs et des librairies mathématiques bien connus sont disponibles sur cette plate-forme à des prix extrêmement compétitifs.

Cependant, il ne faut pas ignorer que ce genre de solution n'est pas sans inconvénients :

- avoir de nombreuses unités centrales est fort pratique pour le calcul, mais pose rapidement un problème d'encombrement évident,
- de même, les Pentiums sont de véritables radiateurs et posent un problème de refroidissement non négligeable quand on les empile. Le bruit dégagé par les unités de disques est aussi à prendre en compte et rend le voisinage de l'engin peu sympathique. Dans notre cas, la présence d'une ancienne salle machine, rescapée de la glorieuse épopée des Vax à l'Institut, fut providentielle !

- le nombre de machines rend nécessaire un système de batch évolué pour utiliser efficacement les divers CPUs de manière transparente pour les utilisateurs.
- ce genre de montage a un aspect « Mécano » évident et on peut raisonnablement se poser la question de sa pérennité. Pour ce qui concerne la maintenance, la standardisation des composants permet d'envisager l'avenir après la période sous garantie avec une certaine sérénité, le nombre de machines permettant un certain cannibalisme. En ce qui concerne l'évolution, tant Linux que l'architecture IA32 semblent loin d'avoir atteint leurs limites.

■ La configuration retenue

La configuration comporte :

- un serveur relié au réseau général de l'Institut et au réseau isolé de la ferme. Il est la seule machine visible des utilisateurs et la seule sur laquelle ils peuvent se connecter. Ils offrent les services suivants,
  - les divers clients X11,
  - les disques,
  - les services NIS,
  - le contrôle du batch,
  - les compilateurs,
- des ?workers? chargés de l'exécution des programmes. Ces exécutions sont contrôlées par le système de batch.

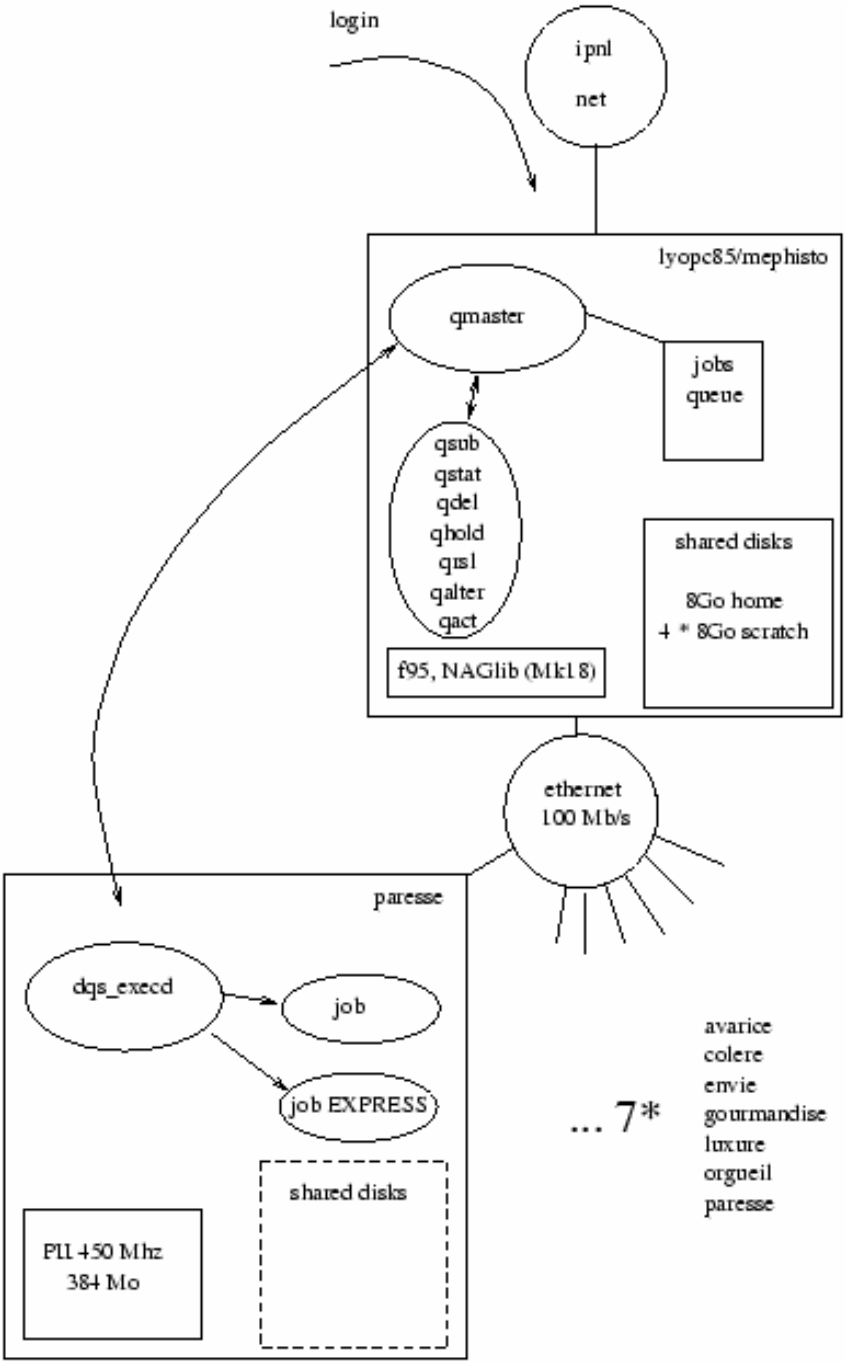
Du point de vue de l'utilisateur, le serveur se comporte comme une machine usuelle, offrant un environnement X11 (XDMCP, outils graphiques et logiciels de traitement et de visualisation de données développés au Cern tels que Paw ou Root), des compilateurs et des librairies permettant la mise au point des codes. Les exécutions gourmandes sont expédiées au batch qui s'effectue exclusivement sur les ?workers? de manière transparente.

■ Les machines

Les machines ont été choisies chez un assembleur de manière à obtenir des configurations dépouillées des éléments inutiles : système d'exploitation, écran et lecteur de CD. Par contre, il est difficile de se passer des claviers qui se révèlent bien encombrants pour ce genre d'usage !

Les configurations sont les suivantes :

- pour le serveur, un Pentium II à 350 MHz avec 256 Mo de mémoire, 1 disque IDE de 4 Go, 5 disques SCSI Fast UltraWide de 8 Go chacun avec une carte Adaptec et deux cartes Fast Ethernet,
- pour les 7 ?workers? des Pentium II à 450 Mhz avec 384 Mo de mémoire, un disque local IDE de 6Go et une carte Fast Ethernet,
- un hub 8 voies Fast Ethernet constitue le réseau privé de ces machines.



■ Le système de batch

Sans parler des systèmes commerciaux tel Load Leveler et autres, la prolifération de telles fermes de calculs dans les grands centres de calcul scientifique a entraîné le développement de systèmes de batch plus ou moins gratuits. Parmi ceux-ci, nous avons envisagé l'utilisation de :

- une évolution de NQS (Network Queuing System) développée au Cern,
- le système BQS (Batch Queueing System) développé au Centre de Calcul de l'In2p3 pour ses besoins propres,
- DQS (Distributed Queueing System) développé à Florida State University.

Nous avons finalement retenu ce dernier pour différentes raisons :

- son extrême simplicité d'installation, de configuration et de maintenance,
- le peu de ressources qu'il nécessite (un seul démon par machine) et son comportement très réactif,

- son interface utilisateur simple et standard (qsub, qstat, etc.),
- sa robustesse dans les situations critiques,
- la possibilité d'effectuer la répartition des travaux en fonction de ressources demandées à la soumission et d'avoir des queues dont l'exécution est suspensive pour les autres.

Au vu de la nature des calculs et du public de cette ferme, une configuration extrêmement simple a été choisie, avec sur chaque machine deux queues d'exécution, une, dite « normale » sans limite et une, dite ? express? qui suspend l 'exécution d'un éventuel job dans l'autre. Cette configuration permet la soumission massive de travaux de production sans différer l'exécution de calculs courts.

La répartition des travaux se fait suivants les méthodes habituelles de ces systèmes, offrant une répartition équitable des ressources et empêchant une monopolisation par un utilisateur.

Il est à noter que DQS permet, par son système de gestion des ressources et l'exécution préliminaire de scripts d'initialisation, l'exécution de jobs parallèles sur un groupe de processeurs communiquant par PVM ou MPI. Cette facilité n'étant pas demandée par nos utilisateurs actuels n'a pas été utilisée.

■ Les compilateurs et les librairies

Notre choix s'est porté sur les logiciels de Nag pour les raisons suivantes :

- compatibilité avec le compilateur Cray, ce qui était une condition impérative,
- qualité du code généré avec en particulier une optimisation correcte,
- nombreuses options permettant aussi bien de compiler du code fossile que du Fortran 90,
- prix très intéressant de la version Linux sur PC pour les sites académiques !
- en ce qui concerne la librairie, ses nombreuses routines et sa réputation ont emporté la décision.

■ Le réseau

Les machines sont reliées sur un « hub » Fast-Ethernet dans un réseau non routé. Le seul protocole qui y circule est bien sur tcp/ip. La seule machine accessible de l'extérieur est le serveur qui est relié au réseau général de l'Institut. Sur ce réseau, le serveur est client nis et ntp. Sur le réseau privé, le serveur diffuse les ?maps? nis et ntp. Il exporte son espace disque en nfs et contrôle le système de batch (dqs). Les ?workers? sont eux clients nis, ntp et dqs, montent les disques nfs du serveur.

Aucun service ne transite par le serveur à une exception près : le courrier électronique. Les ?workers? en ont besoin pour signaler les différentes phases d'avancement des travaux : démarrage, fin ou condition d'erreur. L'agent de transport utilisé est smail, les ?workers? se contentants de transmettre au serveur qui relaie les courriers sur le serveur de l'Institut (sous l'origine user@ipnl.in2p3.fr). En ce qui concerne le service de nom, seul des fichiers hosts sont utilisés.

Pour l'accès à la ferme, l'ensemble de la map nis est propagée mais seuls les comptes ayant un foyer peuvent se connecter. Autoriser (ou fermer) l'accès se limite donc à créer (ou renommer) un répertoire.

■ Quelques détails

L'installation de la ferme a bénéficié de l'expérience acquise lors de la mise en place d'une salle de TP d'informatique basée sur des PCs sous Linux.

La configuration des machines a été préparée sur un PC « normal » (c'est à dire avec un écran, un lecteur CDROM, etc.). Une fois cette installation prête, elle a été copiée sur chaque disque des machines. Celles-ci sont ensuite personnalisées par un script. Quelques outils facilitent la maintenance de la ferme : propagation automatique de fichiers de configuration ou d'installation, installation sur l'ensemble des machines, gestion des utilisateurs, comptabilité du batch, etc.

Un écran de récupération est gardé à proximité de la ferme pour les quelques cas où le réseau ne peut pas être utilisé.

En ce qui concerne les sauvegardes, elles sont pour l'instant effectuées par copie sur les serveurs de l'Institut. En cas de besoin, nous envisageons d'utiliser le produit Networker déjà installé sur ces serveurs. En ce qui concerne l'espace de 40 Go utilisé pour les résultats intermédiaires, une solution de migration sur cartouches DLT à été étudiée et sera mise en place en cas de besoin.

La distribution utilisée est une Debian HAMM.

■ Conclusion : bilan d'une année

La première constatation est tout simplement : cette solution fonctionne et donne satisfaction. Les utilisateurs ont réellement l'impression d'utiliser un machine « normale » dotée des services habituels avec la possibilité de soumettre à l'exécution des quantité impressionnantes de calculs. Le fait que les fichiers de sortie des travaux soient placés dans le répertoire de soumission permet un suivi simple de l'évolution du travail.

Une des difficultés réside dans la puissance disponible avec ce genre de solution : les physiciens étaient habitués à soumettre un travail, à récupérer le résultat, à l'analyser, avec un taux de soumission de l'ordre de quelques travaux par jour. La solution proposée permet de traiter en un week-end plusieurs centaines de travaux de ce type ! Il est évident dans ces conditions que le système ne peut pas être utilisé à l'optimum de ses possibilités. La fonction créant le besoin, de nouvelles applications sont rapidement apparues. La « fermette » est actuellement utilisée pour des calculs de modélisation de diffusion qui semblaient irréalistes avant : une production qui aurait demandé environ un mois de calcul sur un Cray est en cours de préparation !

Un aspect essentiel à la réussite de ce genre de projet est bien sur la rédaction d'une documentation claire et de la tenue de séances de formation, le temps que l'on y consacre se retrouve largement à l'usage !

Parmi les aspects très positifs, il faut aussi noter le peu de maintenance nécessité par cette « fermette », sans passer sous silence les petits soucis inévitables avec ce genre de montages : quelques barrettes mémoire défectueuses, des tiroirs SCSI présentant des problèmes de connectique, etc.

Pour conclure, une batterie de PCs sous Linux, dotée d'un bon compilateur et d'un système de batch peut rendre de grands services dans un laboratoire scientifique pour un prix très compétitif.

■ Remerciements

Merci à Jacques pour avoir « osé », à Eric et à Stéphane pour avoir montré que cela pouvait marcher, à Nadège et à Dany pour avoir montré que cela servait !