L'expérience de l'Observatoire de Grenoble dans la mise en place d'une architecture distribuée

-

Françoise Roch, roch@gag.observ-gr.fr

Rsumé. *L'Observatoire de Grenoble s'articule autour de 3 laboratoires physiquement distribués dans 3 bâtiments du Campus Universitaire de Grenoble. Ces trois laboratoires ont des besoins informatiques importants et très semblables. Une politique informatique coordonnée a permis de mettre en place une structure homogène, souple et permettant le traitement efficace de la quasi totalité des applications.*

1 Besoins et objectifs

Les applications traitées à l'Observatoire sont gourmandes tant au niveau puissance de calcul, qu'espace mémoire ou espace disque. Les modélisations numériques effectuées peuvent nécessiter plusieurs mois de calcul et plusieurs centaines de MO de mémoire. Par ailleurs, les applications de dépouillement de données mettent en jeu une masse importante de données.
Les objectifs principaux qui ont été visés lors de la mise en place de ce parc informatique sont de traiter des applications lourdes localement, disposer d'un service souple d'utilisation, exécuter nos applications sur des moyens adaptés et dédiés et utiliser des outils de calcul distribué pour développer des applications pour réseau de stations ou machines parallèles.

Nous nous sommes orientés vers une architecture répartie en réseau constituée d'une part, sur chaque laboratoire, d'une grappe de stations et de terminaux X qui assurent en priorité les besoins en calcul interactif et d'autre part, au niveau d'un centre commun, de stations très puissantes et très bien configurées (en terme d'espace mémoire et d'espace disque) assurant, au travers d'un système de batch, un service de calcul intensif. Cette séparation entre batch et interactif permet d'améliorer les temps de réponse en interactif, de dédier puissance de calcul et espace mémoire pour des calculs lourds, donc d'optimiser l'usage des calculateurs. L'architecture est homogène, ceci permet de partager une expertise commune entre les différents informaticiens de l'Observatoire.

2 L'architecture du site commun

L'utilisation efficace des stations et des disques distribués entre les différents sites nécessite un réseau performant et fiable.
En 1993, l'Observatoire et le CICG (Centre Interuniversitaire de Calcul de Grenoble) ont collaboré pour l'installation d'un réseau rapide FDDI reliant les serveurs principaux – IBM RS 6000 – des 3 pôles de l'Observatoire, le centre commun de calcul intensif et le CICG (fig. 1). Ce réseau nous permet d'obtenir, entre nos machines, des débits de l'ordre de 4 MO/s sur des applications

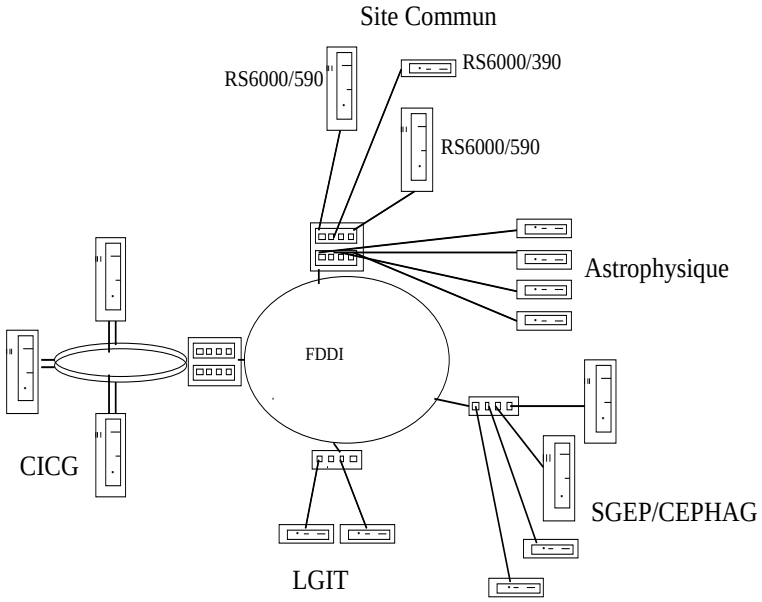

FIG. 1 – *Interconnexion FDDI des machines de l'Observatoire et du CICG*

telles que ftp (binaire). Ce réseau rapide nous permet en particulier de rapatrier sur les machines des laboratoires les résultats obtenus sur le site commun avec de bons débits.

3 Systèmes de fichiers

La plupart des transferts de données entre les machines de l'Observatoire se font au travers du système de fichiers NFS [3]. Nous avons testé d'autres systèmes, notamment DFS et PIOUS.

DCE et DFS. Au niveau de notre première expérience de configuration, DCE/DFS (version 1.3 IBM) [2] nous a paru être un environnement assez lourd au niveau administration système.
La version actuelle est limitative sur des réseaux de stations comportant plusieurs interfaces; dans notre cas, l'installation sur le réseau FDDI qui était défini comme un réseau indépendant dédié aux partages de données entre nos sites pour le calcul intensif, s'est avérée difficile et contraignante. Ces limitations sont directement liées aux aspects sécurité, particulièrement approfondis dans DCE.

Le système de fichiers DFS semble pour sa part très intéressant. NFS et DFS fournissent tous deux des mécanismes de "read-ahead" permettant d'anticiper des lectures et de réduire les côuts de transfert. Le mécanisme de DFS suit cependant une vraie logique de système de fichiers unique distribué. DFS permet des écritures asynchrones tout en intégrant la cohérence des données par un algorithme de jeton.

Le tableau 2 compare les performances en lecture et écriture entre NFS et DFS entre deux 590 munis de 512 et 256 MO de mémoire centrale, et sur des disques SEAGATE SD2400E, SCSI-I (les performances étant dépendantes de la configuration, ces mesures sont à considérer comme un ordre de grandeur de ce que l'on peut attendre sur ce type de plateforme). Les accès sont faits sur

débits en MO/s	écriture 100 MO	lecture 100 MO	écriture 500 MO	lecture 500 MO	écriture 1 GO	lecture 1 GO
débits NFS	0.6	6.25	0.55	2.34	–	–
débits JFS / DFS	1.96	4.76	1.8	2.48	1.88	2.47

FIG. 2 – *Entrées/sorties séquentielles avec les primitives read et write entre deux 590 sur FDDI.*

des partitions JFS (version 3.2.5 d'AIX) exportées NFS et DFS ; les lectures et écritures sont faites par blocs de 8 KO, et la taille du "chunk" DFS est de 64 KO.
Dans le cas de la lecture de 100 MO, le fichier est caché en mémoire, ce qui explique les débits supérieurs aux débits des disques. Par contre, pour 500 MO, la limitation provient du débit des disques et non du réseau. Les écritures asynchrones de DFS donnent des débits supérieurs à NFS, et comparables à ceux obtenus en JFS.

débits en MO/s	écriture 2*100MO	lecture 2*100MO	écriture 2*250MO	lecture 2*250MO	écriture 2*500MO	lecture 2*500MO
JFS / DFS	3.12	5.71	3.04	4.20	3.26	3.86

FIG. 3 – *Entrées/sorties parallèlles avec les primitives read et write entre deux 590 sur FDDI ; les 2 disques sont ici rattachés à la même CPU mais sur 2 contrôleurs SCSI 1 distincts.*

Nous avons indiqué dans le tableau 3 les performances que l'on peut obtenir en faisant cette fois-ci des "read" ou des "write" alternés sur 2 fichiers, par bloc de 8 KO. En raison des mécanismes de read-ahead et de write-behind, les entrées sorties se font en parallèle ; les performances sont d'autant meilleures que les disques sont rattachés à des contrôleurs distincts.

DFS semble donc très bien adapté à des machines puissantes sur un réseau rapide. Il est capable d'exploiter les performances de la machine et du réseau en lecture mais également en écriture.

Accès parallèles : DIO et PIOUS. Les bonnes performances que nous obtenons sur les accès parallèles (cf table3) grâce au read-ahead et au write-behind d'une part, et les limites que nous imposait le système de fichiers AIX sur la taille des fichiers et des systèmes de fichiers nous ont amené à développer un outil : DIO. DIO permet de travailler sur un fichier de taille supérieure à 2 GO qui est physiquement distribué sur plusieurs disques (de façon bloc-cyclique) mais qui est vu et accèdé par l'utilisateur comme un seul fichier. Cet outil utilise le système de fichiers natif et le système de fichier NFS. Les performances mesurées lorsque l'on travaille en local sur deux partitions peuvent être doublés par rapport à un accès séquentiel si les partitions sont sur deux disques distincts rattachés à deux contrôleurs SCSI distincts. En NFS, l'intérêt du logiciel est moindre, surtout au niveau des écritures en raison de leur caractère synchrone.

Nous avons par ailleurs testé l'outil PIOUS [1] basé sur l'environnement PVM. PIOUS est un système de fichiers parallèles qui peut être utilisé sur un réseau de machines hétérogène. Il permet en particulier de travailler en mode asynchrone sur un fichier physiquement distribué sur plusieurs systèmes interconnectés et donc d'additionner les débits (le fichier étant réparti de façon

nombre de segments	1 segment local	1 segment local, 1 segment distant	1 segment local, 2 segments distants
Ecriture en KO/s	2197	4302	4970
Lecture en KO/s	2064	3879	4413

FIG. 4 – *Entrées/Sorties avec PIOUS*

bloc-cyclique sur plusieurs segments).
Le tableau 4 indique les débits mesurés pour un fichier de 500 MO, la taille des blocs étant de 8 KO. Les débits obtenus sont ici aussi bons en écriture qu'en lecture et on tire parti des accès parallèles avec 2 et même 3 segments.

4 Calcul distribué

Le réseau de machines de l'Observatoire nous permet par ailleurs de développer des applications distribuées ou parallèles dans l'environnement PVM avec de très bons débits au niveau des communications. Certaines de ces applications sont alors portées sur des machines parallèles telles que la IBM-SP2 de l'IMAG. Des tests ont été réalisés, d'une part avec l'environnement PVM du domaine publique qui travaille au niveau TCP et UDP et d'autre part avec l' environnement PVMe développé par IBM pour FDDI et optimisé pour RS6000. Les débits mesurés sont inférieurs à 2 MO/s pour PVM domaine public et de l'ordre de 7-8 Mo/s pour PVMe. Ces résultats sont intéressants, sachant que les débits théoriques sur le switch de la machine SP2 d'IBM sont de l'ordre de 40 MO/s, difficilement atteints en pratique.

5 Conclusion

Le centre de calcul intensif de l'Observatoire de Grenoble, souple d'utilisation, est bien adapté à nos applications et permet des traitements qu'il serait particulièrement difficile de déporter vers d'autres centres, en raison des volumes de données mis en jeu et des visualisations rapides souvent nécessaires.

Il permet par ailleurs de développer des applications pour le calcul distribué ou le calcul parallèle qui peuvent ensuite être traitées sur des calculateurs nationaux.

Références

[1] Steven A. Moyer and V. S. Sunderam. A Parallel I/O System for High-Performance Distributed Computing. Technical Report CSTR-940101, Emory University, Atlanta, GA 30322, Dept. of Mathematics and Computer Science, january 1994. pious mathcs.emory.edu.

[2] Brice Muang-Khot. Interopérabilité entre DFS et NFS. *Les Cahiers d'AIX*, april 1994.

[3] Hal Stern. *Managing NFS and NIS*. O'Reilly , 1991.