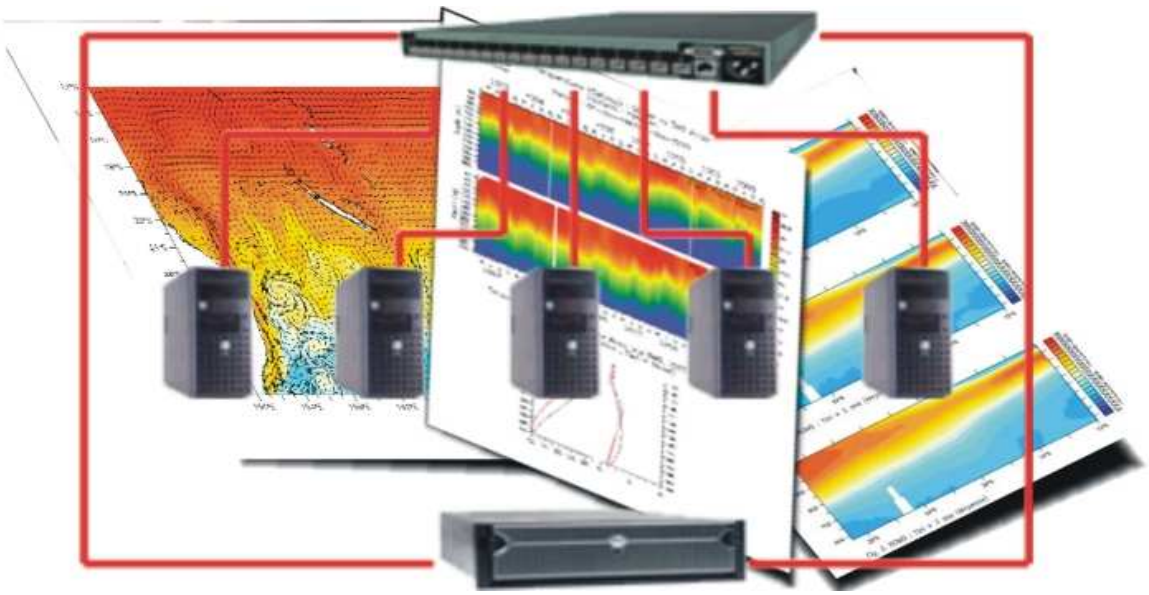
PROJET SPIRALE 2008

Institut de recherche pour le développement

CLUSTER DE CALCUL DU CENTRE DE NOUMEA

UPGRADE DU CLUSTER DE NOUMEA

RESEAU RAPIDE DE COMMUNICATION



DEMANDE DES UNITES DE RECHERCHE 103, 65 ET 182 DU CENTRE DE NOUMEA, NOVEMBRE 2008

Groupe de travail

Jérôme Lefèvre, resp. Informatique UMR 65 / UR 103 (jlefevre@noumea.ird.nc)

Michel Ménézo, resp. SIL Nouméa (menezo@noumea.ird.nc)

Patrick Marchesiello, resp. Scientifique UMR 65 (Patrick.Marchesiello@noumea.ird.nc)

Pascal Douillet, resp. Scientifique UR 103 (douillet@noumea.ird.nc)

Christophe Menkès, resp. Scientifique UR 182 (menkes@noumea.ird.nc)

SOMMAIRE

1. Objet et contexte 2
2. Présentation du cluster du Centre de Nouméa 3
Espace de stockage 3
Environnement Logiciel 4
Les composants et les perspectives d'upgrade 4
3. Activités associées au cluster et portage de l'expérience Cluster vers les pays du Sud 6
4. Objet de la demande de financement 8
Equipement en réseau InfiniBand 8
Les perspectives de parallélisation hybride des codes de calcul 10
5. Visibilité du cluster de Nouméa à l'échéance 2010 12
6. Synthèse de la demande 13
Bibliographie 14

ANNEXES

ANNEXE A Proforma Alinéos : Passage réseau en technologie InfiniBand

ANNEXE B White Paper: Connectivity in the Multi-core Environment, Mellanox Technologie 2007

ANNEXE C Note technique sur l'optimisation du réseau de communication Gigabit du cluster de Nouméa

Liste des tableaux

Tableau 1 : Performances des principaux réseaux de communication employés dans les fermes de PC 8

Tableau 2. Synthèse de la demande de participation 13

Liste des figures

Figure 1 Synoptique de l'architecture installée au centre de Nouméa 3

Figure 2 Illustration de l'overhead des processeurs d'un nœud de calcul sur une configuration ROMS employant successivement 1, 2, 3 et 4 nœuds (4 CPUs/nœud). 9

Figure 3 Evolution du prix de l'équipement en réseau rapide (InfiniBand et/ou Myrinet) pour 10 noeuds (switch + adaptateurs). 10

1. Objet et contexte

Les équipes scientifiques des UMR LEGOS (IRD-UR65), CAMELIA (IRD-UR103) et LOCEAN (IRD-UR182) basées à Nouméa bénéficient depuis le début de l'année 2005 d'un cluster[1] de calcul à base de processeurs Opteron, dont le financement de départ a été apporté par la DSI au titre de l'opération SPIRALES.

Mis en production en janvier 2005, le cluster et son environnement logiciel constituent une plateforme d'investigations numériques adaptée aux activités de modélisation océanique et atmosphérique régionale conduites par les 3 équipes :

- *Etude des processus et répercussion de la variabilité climatique dans le Pacifique Sud-Ouest*
- *Relation thonidès et Environnement dans le Pacifique Sud-Ouest*
- *Risques et Aléas Cycloniques en relation avec le climat*
- *Modélisation côtière et lagonaire*
- *Océanographie opérationnelle dans le cadre du chantier Prévimer*
- *Couplage régional Océan/Atmosphère*

En outre, cette plateforme de calcul évolutive permet d'accompagner les perspectives de recherche exprimées par les scientifiques recourant à la modélisation. L'architecture en grappe offre en effet l'avantage d'augmenter la puissance de calcul en ajoutant de nouveaux serveurs de calcul au parc existant. C'est cette piste qui a été explorée et mis en œuvre avec l'aide de la DSI dans le cadre de l'opération Spirales 2006. Les fonds alloués ont permis de doubler le nombre de machines afin de répondre aux besoins des nouveaux utilisateurs.

A la fin de l'année 2007, ce sont les CPUs qui ont fait l'objet d'un upgrade. La capacité de calcul est passée de 20 à 34 processeurs depuis le remplacement des anciens CPUs par des Opteron 285 bicoeur. Cet upgrade a été réalisé avec la participation financière des 3 unités de Recherche impliquées dans le programme. La capacité de stockage du cluster a également été augmentée en 2007 avec l'ajout de 4 TB, puis 7 TB d'espace disque en raid 5.

Le présent projet Spirales 2008 s'attache à upgrader le réseau de communication, devenu le principal goulet d'étranglement du cluster. Non envisageable au début de l'année 2005, car de technologie trop récente et trop onéreuse à l'époque, l'équipement du cluster de Nouméa d'un réseau rapide basé sur la technologie **InfiniBand**, permettra d'optimiser l'utilisation des processeurs multi-cœur et redonner une nouvelle jeunesse au cluster de Nouméa, prolongeant du même coup sa durée de vie.

[1] Cluster : désigne une grappe d'ordinateurs connectés entre eux. Chaque machine est un noeud du cluster, l'ensemble est considéré comme une seule et unique machine pouvant concurrencer la puissance des supercalculateurs pour un coût moindre

2. Présentation du cluster du Centre de Nouméa

L'architecture de calcul installée au centre de Nouméa, sous la supervision de Michel Ménézo et Jérôme Lefèvre, se compose de 10 serveurs bi-processeurs à base d'Opteron 285 double cœur et d'Opteron 246 monocoeur.

Le réseau de communication inter-noeuds repose sur la technologie Gigabit Ethernet, la plus répandue car la moins onéreuse, mais en contrepartie la moins performante des réseaux de communication. Des tentatives d'optimisation de notre réseau de communication Gigabit Ethernet ont été menées (cf Annexe C). Elles ont permis d'améliorer les performances de notre réseau, mais elles ne seront jamais à la hauteur des autres réseaux rapides, InfiniBand ou Myrinet.

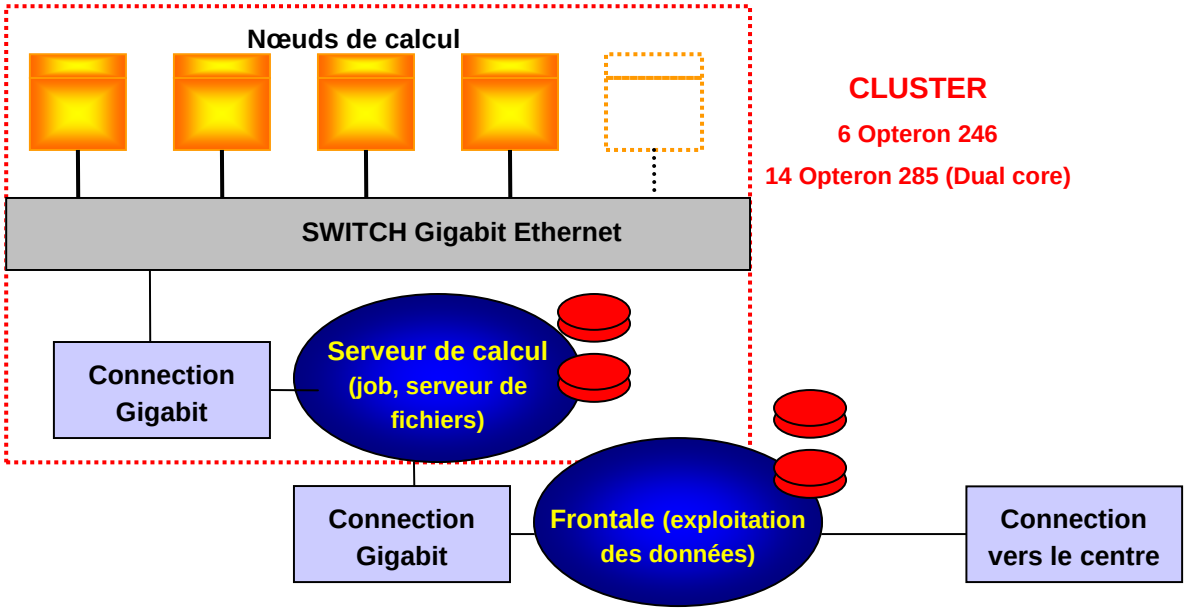


Figure 1 Synoptique de l'architecture installée au centre de Nouméa

Espace de stockage

Les activités de modélisation associées au cluster sont gourmandes en espace de stockage. Régulièrement, elles demandent d'augmenter l'espace disque. Nous avons opté pour l'emploi de disques SATA 750 GB montés en RAID5 sur carte contrôleur Raid et de disques Externe Lacie de 2 TB, pour les espaces temporaires.

Le système de calcul dispose localement d'un espace de stockage temporaire de 1.2 TB, dont 600 GB accessible sous le système de fichier parallélisé PVFS2, offrant de meilleures performances que le système NFS, ce dernier tendant rapidement à saturer à mesure que le nombre d'accès en I/O des nœuds de calcul augmente.

Outre une frontale disposant de 7.1 TB d'espace de stockage, un nouveau serveur a été récemment mis en service disposant de 7 TB d'espace supplémentaire. L'investissement a été réalisé dans le cadre du projet ANR ''Cyclone et Climat'' (UMR65/UR182).

Environnement Logiciel

Sur le plan logiciel, le cluster est installé sur une base FEDORA Core 3 64 bits. Le système d'administration du cluster provient du kit de déploiement de clustering OSCAR 4.2 (cf. http://oscar.openclustergroup.org).

L'environnement logiciel est tourné vers l'exploitation d'applications scientifiques parallélisées, modèles numériques Océan et Atmosphère pour l'essentiel (ROMS, MARS, WRF et FVCOM). A l'environnement de développement GNU de la distribution Fedora, a été ajouté d'autres softs :

- Les compilateurs Fortran Intel et PGI
- Les librairies de communication MPI : Lam, openMPI, MPI Intel
- Le Gestionnaire de job Torque/PBS
- L'outil de monitoring Ganglia
- Les librairies netCDF 3.6
- Divers logiciels (Matlab, ferret, NCO ...)

Les composants et les perspectives d'upgrade

Le processeur Opteron à la base de notre cluster a été retenu en raison de son haut niveau de mémoire cache et son excellent ratio coût/performance. En outre, ce modèle de processeur équipe la plupart des solutions en serveurs lames (DELL, SUN, HP etc) du marché HPC (High Performance Computing).

Avec la baisse des prix des processeurs double cœur de la gamme opteron 2xx, l'upgrade des processeurs mono coeur par des processeurs double coeur est devenu pertinent. Cette solution est avantageuse car elle ne nécessite pas de serveur supplémentaire et n'augmente que très peu la consommation d'énergie, tout en doublant la capacité de calcul.

Cet upgrade est également rendu possible par la compatibilité des cartes mères Tyan S2885 avec les processeurs double coeur. Ces derniers mois, 7 de nos 10 machines de calcul ont fait l'objet d'un upgrade après acquisition de 14 processeurs Opteron 285 double coeur (cadence 2,6 Ghz). Les Unités de Recherche 182, 65 et 103 ont participé à leur achat. Au moyen d'une configuration ROMS utilisée pour nos benchmarks (grille Safe, 228 X 192 X 32), nous avons noté des temps de calcul de 1,96 seconde/iteration dans le cas du processeur 246 (2 Ghz) contre 1,42 seconde/iteration dans le cas du nouveau processeur 285 (2.6 Ghz). Outre ce gain de performance, il faut ajouter le cpu supplémentaire qu'apporte chaque processeur double-coeur.

D'une capacité initiale de 20 processeurs, le cluster affiche désormais 34 processeurs. Cette capacité pourra être portée à 40 au cours de l'année 2008, en effectuant le remplacement des 6 anciens CPUs restant par des modèles Opteron 285 double coeur.

D’autre part le chipset AMD-8131 intégré sur la carte mère Tyan S2885 offre une perspective d’interconnexion haut débit et rapide au moyen de 2 connecteurs PCI-X 64 bits 133 Mhz (bande passante disponible par slot : 6,4 GBits/s). Ces slots autorisent l'usage d'adaptateurs pour réseaux rapides, tels Myrinet, InfiniBand ...

La technologie Gigabit Ethernet a été initialement retenue pour le réseau, afin de ne pas grever le coût initial de l'équipement (l'équipement en InfiniBand ou Myrinet du cluster en 2004 aurait impliqué un doublement du coût total du cluster à l’époque). Cette possibilité d’upgrade avait été soulevée lors de la précédente demande Spirales 2006. Elle fait l’objet de la présente demande Spirales 2008.

3. Activités associées au cluster et portage de l'expérience Cluster vers les pays du Sud

Le cluster dispose d'un site web dédié, dans lequel le visiteur trouvera des informations sur la configuration matérielle et les chantiers de modélisation actuellement conduits en appui sur le cluster : www.ird.nc/UR65/ROMS

Cet équipement a permis d'augmenter de manière significative l'activité modélisation au Centre de Nouméa.

Dans le cadre des activités de modélisation en océanographie opérationnelle, le Centre de Nouméa s'inscrit en partenariat avec le GIE Mercator/Océan pour la validation et l'utilisation des sorties du modèle opérationnel Français sur la région du Pacifique Sud-Ouest et des passerelles sont lancées avec le système opérationnel Australien BlueLINK.

La plateforme de calcul contribue également au développement d'une nouvelle collaboration technologique entre l'IFREMER et l'IRD sur le thème de la prévision océanique régionale sur la ZEE de la Nouvelle-Calédonie. Le site http://prevision.ird.nc/ en cours de développement brosse une revue des moyens mis en œuvre dans ce chantier et va s'efforcer de diffuser les résultats des simulations opérationnelles ROMS, MARS et WRF sur la région.

D'autre part, les simulations réalisées avec le modèle atmosphérique WRF et les modèles couplées ROMS/PISCES au moyen du cluster ont également permis d'établir de nouveaux projets scientifiques dans le cadre des appels à projets ANR. Ces projets permettront d'accueillir, dès 2008, 2 nouveaux Posts-Docs en modélisation. Dans ce cadre, des simulations longues et mettant en oeuvre de grosses grilles de calculs seront réalisées.

L'expérience Cluster de Nouméa a valeur d'investissement dans la définition et le développement de calculateurs scientifiques qui soient à la fois performants et adaptés aux moyens de la recherche dans les pays du Sud. Dans ces pays, les alternatives du type cluster demeurent en effet les seules accessibles, tant sur le plan financier que sur le plan de la maîtrise hardware/software. Les développements récents vont dans le bon sens, notamment la distribution libre des systèmes d'exploitation (Linux), des bibliothèques (Netcdf, MPICH, LAM …), des langages de programmation (Fortran, C++), ainsi que des codes de calcul. Il en va de même pour l'accès aux composants informatiques de qualité entrant dans la réalisation d'une ferme de PC, dont les coûts ont fortement chuté ces dernières années.

L'IRD (UMR LEGOS et UR 97-ECO-up) en association avec l'INRIA participe activement au développement et la distribution du codes de calcul océanique ROMS (http://www.brest.ird.fr/Roms_tools/ et https://gforge.inria.fr/projects/romsagrif/), et des outils et modèles d'applications pluridisciplinaires qui l'entourent. Ce code a été porté sur de multiples plateformes notamment en Afrique, en Amérique du Sud et dans le Pacifique Sud, avec le souci de la performance (incluant parallélisation et raffinement de maillage AGRIF) mais aussi de l'accessibilité et du coût. CAMELIA (UR 103) de son coté porte actuellement le

projet d’exporter son expertise en Environnement Côtier et modélisation lagonaire à d’autres pays du Sud, en particulier la zone Caraïbe (Cuba et Mexique).

Nous pensons que ce souci de combiner performance et accessibilité doit s’étendre à présent à tout le système de calcul incluant hardware et software pour que les pays du Sud soient réellement en mesure de produire du calcul scientifique de haut niveau. Sur la base de l'expérience Cluster de Nouméa, l’IMARPE (Instituto del Mar del Peru), Lima, Perou, a récemment acquis un cluster composé de stations DELL. Cet équipement est en cours de montage et reçoit le soutien technique du Centre de Nouméa pour le déploiement du système d'exploitation et l'environnement logiciel de calcul ainsi que la formation de l’ingénieur chargé de son administration (Augusto Ingunza sera accueilli à l'IRD Nouméa pour une formation sur le thème "Montage et Administration Cluster" du 15 janvier au 15 Février 2008).

Un nouveau projet est prévu dans le cadre de la coopération entre CAMELIA et l’Universidad Autonoma Metropolitana au Mexique. L’équipement à base de 4 serveurs DELL et réseau InfiniBand sera implémenté pour les applications parallélisées MARS, ROMS et WRF (porteur du projet : Pascal Douillet). Le montage de ce cluster et le choix du réseau rapide de communication s’appuient en grande partie sur l’expérience de Nouméa.

Fin 2006, un autre projet mené par l’équipe Thetis (Olivier Maury, UR109 Sète) a également bénéficié de l’expérience de Nouméa pour la définition des composants de son cluster à base de processeur Xeon bi-coeur.

4. Objet de la demande de financement

La demande de participation financière porte sur l'optimisation et l'amélioration des performances de calcul du cluster du centre de Nouméa, en dotant le cluster **d'un réseau rapide de communication**.

Equipement en réseau InfiniBand

Pour la communication inter-noeuds, notre cluster est actuellement équipé d'un réseau reposant sur la technologie Gigabit Ethernet (GigE) et emploi le protocole TCP/IP pour l'échange des messages. Comme le montre le tableau 1, les performances du GigE sont très en deçà de celles procurées par les réseaux rapides InfiniBand ou Myrinet généralement employés dans les fermes de calcul HPC.

Performance	InfiniBand	Myrinet	Gigabit Ethernet
Switch Latency	160 ns	200 ns	10,000 ns
End-to-End Latency (8 byte packet)	7.6 us*	8 us	60 us
End-to-End Latency (1K byte packet)	13 us	22 us	80 us
Throughput	822 MB/sec	250 MB/sec	100 MB/sec
CPU Overhead	3%	6%	80%

* Planned software/firmware releases will further reduce latency. Data provided by: Prof Dhabaleswar K. Panda, Ohio State University, Jan. 2003

Tableau 1 : Performances des principaux réseaux de communication employés dans les fermes de PC

De part ses temps de latence et sa faible bande passante, le réseau GigE devient rapidement le principal goulet d'étranglement pour les applications parallélisées intensives en échange de données et employant des messages de grandes tailles.

D'autre part, la communication reposant sur le protocole TCP/IP, celui-ci induit des consommations CPU lourdes et des pertes de temps en raison des copies systématiques dans une mémoire tampon : le socket buffer. Les interruptions fréquentes du processeur, la perte de cycles CPU réduisent la performance globale du système de calcul.

Nous rencontrons actuellement ce problème sur notre cluster pour nos grosses configurations ROMS-PISCES et également avec le modèle MARS de l'IFREMER, dont la stratégie de parallélisation employée implique l'échange de paquets de grande taille, ce qui est très pénalisant avec le réseau GigE, dont la performance reste satisfaisante seulement

dans la gamme des paquets de petites taille (< 1000 Bytes). La Figure 2 illustre l'overhead[2] des CPUs à mesure que le partionnement de la grille de calcul augmente, impliquant une augmentation de l'échange des données. Pour une configuration mettant à contribution 2 nœuds de calcul, à raison de 4 processeurs par nœud, les processeurs affichent chacun 30 % de perte de cycle CPU à cause des interruptions liées à la gestion des communications. A partir de 3 nœuds, la perte atteint 60 %, ce qui engendre une dégradation de la scalabilité[3] du code ROMS sur notre cluster équipé en réseau GigE. Le code MARS montre le même comportement, mais la limite de scalabilité du code est plus rapidement atteinte à cause de la taille des messages plus grande (Cf Annexe C).

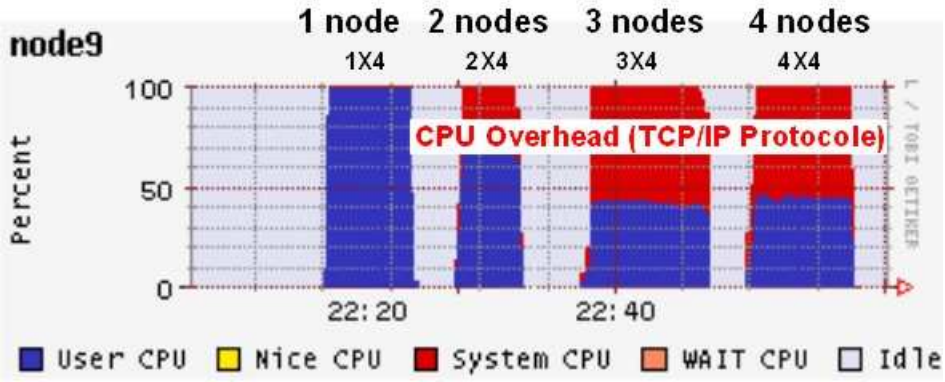


Figure 2 Illustration de l'overhead des processeurs d'un nœud de calcul sur une configuration ROMS employant successivement 1, 2, 3 et 4 nœuds (4 CPUs/nœud).

Avec l'arrivée des stations multi-coeurs, la demande en entrée/sortie est très significativement augmentée et le protocole TCP/IP devient très vite pénalisant. La solution est de passer par un réseau plus rapide, n'employant pas le protocole TCP/IP et ne sollicitant pas le processeur, ce dernier restant entièrement dédié au calcul. C'est ce que procurent comme avantage les réseaux basés sur la technologie InfiniBand ou Myrinet.

L'annexe B donne les détails techniques de la technologie InfiniBand et les performances qui sont obtenues dans le cas de l'application CFD Fluent, très voisine de nos applications parallélisées. L'annexe C présente une comparaison de la performance du code ROMS sur le nouveau cluster du Cap Town, pour les deux réseaux Gigabit et InfiniBand.

Comme le montre la figure 3, le prix de l'équipement en technologie InfiniBand a fortement chuté depuis la mise en service de notre cluster début 2005, ce qui rend cet investissement aujourd'hui pertinent, en plus de la réponse qu'il apporte pour les configurations à base de stations multi-cœurs.

---

[2] Overhead est le temps passé (perdu) à effectuer une commutation entre deux processus en temps partagé. C'est le temps nécessaire au commutateur pour sauvegarder un contexte et en recharger un autre.

[3] Scalabilité est la capacité d'un système à évoluer en cas de montée en charge

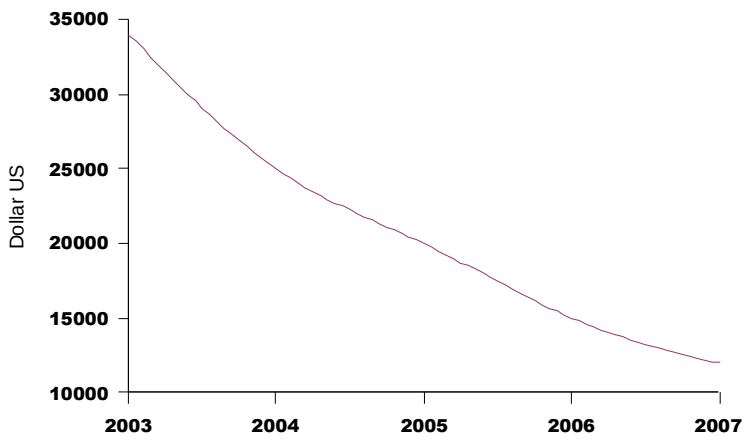


Figure 3 Evolution du prix de l'équipement en réseau rapide (InfiniBand et/ou Myrinet) pour 10 noeuds (switch + adaptateurs).

(Informations obtenues depuis les forums www.beowulf.org et l'intégrateur Alinéos)

Nous retenons la technologie InfiniBand pour le passage en mode rapide de notre réseau de communication. Cette technologie est la plus répandue et la moins onéreuse parmi les communications à hautes performances. De plus, elle est supportée dans les différentes librairies de communication MPI que nous employons actuellement dans nos applications parallélisées.

Montant de l'upgrade de notre réseau de communication en technologie InfiniBand, incluant adaptateur PCI-X, câbles et switch 24 ports (Hors Taxe) : **8790 Euros**

Voir Annexe A, Proposition ALINEOS[4]

Les perspectives de parallélisation hybride des codes de calcul

Les grappes de PC multi-coeurs offrent la perspective de tirer profit de l'approche hybride dans la parallélisation des applications. La programmation hybride associe 2 modes de programmation parallèle : le mode multi-thread OpenMP (utilisant la mémoire partagée des processeurs) et le mode multi-processus MPI (utilisant la mémoire distribuée). Contrairement à la parallélisation muti-processus (tout MPI), chaque nœud de calcul héberge les différents threads OpenMP d'un même processus MPI avec l'approche Hybride.

L'approche hybride permettrait une meilleure efficacité du code, en recherchant à optimiser les caractéristiques des machines (mémoires caches des processeurs, mémoires partagées), en déchargeant le réseau du flot de données échangées entre les noeuds et en

[4] Des demandes de proforma ont été déposées auprès de Dataswift (France) et auprès de la société SYSCOM (représentant IBM, Nouvelle-Calédonie). Mais aucune réponse nous a été retournée.

réduisant les attentes liées à la synchronisation entre les processus. La contrepartie est une programmation qui peut s'avérer lourde pour rendre la parallélisation hybride efficace. Les performances peuvent variées selon la qualité de la programmation. D'autre part, les architectures de calcul haute performance dotées de processeurs et réseau rapides tendent à effacer les gains que procurent l'approche hybride.

Les codes MARS3D et WRF sont prêts à être employés en mode hybride. Cependant, la performance de ces 2 codes en mode hybride est moins bonne comparée au mode MPI pur, après test sur notre cluster.

Dans le cas du code ROMS, le gain de la parallélisation hybride peut s'avérer payante, dans la mesure où la programmation OpenMP du code a été pensée dès l'origine pour permettre une très bonne scalabilité sur les supercalculateurs à mémoire partagée, en tirant parti de leur grande mémoire cache.

En partenariat avec l'INRIA (Laurent Debreu) et UCLA (Alexander Shchepetkin), nous menons des investigations visant à rendre la parallélisation hybride opérationnelle dans ROMS. Les premiers tests sont encourageants. Ce travail permettra d'optimiser l'utilisation du code ROMS sur les clusters de calcul multi-coeurs, notamment à Nouméa, Lima et Mexico.

5. Visibilité du cluster de Nouméa à l'échéance 2010

En 2010, certains composants du cluster seront âgés de 5 ans et seront dépassés au regard de l'évolution technologique du matériel informatique. Toutefois, en raison de la qualité des éléments achetés (qualité Serveur), des efforts d'upgrade des composants consentis par les utilisateurs (CPUs, Mémoire…), nous croyons à la pérennité et à la réponse satisfaisante du Cluster actuel au regard des besoins en terme de modélisation pour les 3 prochaines années.

La modernisation du réseau de communication au moyen de la technologie InfiniBand permettra assurément d'améliorer significativement la performance globale du système de calcul pour les grosses configurations.

Au-delà, et face à une éventuelle demande en puissance de calcul supplémentaire, laquelle ne se profile pas pour le moment, la stratégie d'évolution du cluster reposera sur l'ajout de nouveaux serveurs de calcul multi-cœurs.

6. Synthèse de la demande

OPERATION	DETAIL	BUDGET	PART. UR	PART. SPIRALES 2008
I. Equipement en réseau rapide technologie InfiniBand	10 adapteurs PCI-X 4x InfiniBand (SFF-8470) - 2 ports 10 câbles 4x InfiniBand 1 switch 24 ports InfiniBand	8 790 €	A hauteur des frais de taxes (soit 30% environ)	8790 €
		DEMANDE SPIRALE 2008 2007 : Cluster de Calcul UR103, UMR65, UR182 Nouméa		**11 929 €**

La demande de participation financière **s'élève à 11 929 Euros, dont 8 790 Euros pour la part Spirales 2008**

Tableau 2. Synthèse de la demande de participation

Bibliographie

Documents techniques / Sites de référence cluster

Site d'Oscar : http://oscar.openclustergroup.org

Cluster Computing, Architectures, Operating Systems, Parallel Processing & Programming Languages, Richard S. Morrison (Document pdf)

Beowulf HOWTO : http://www.ixus.net/howto.php

Beowulf Tutorial : Building a Beowulf System
http://www.cacr.caltech.edu/beowulf/tutorial/building.html

Conception d'un système à haute performance, CETMEF :
http://www.cetmef.equipement.gouv.fr/projets/transversaux/cluster/

ORganisation Associative du Parallélisme (ORAP) : http://www.irisa.fr/orap

Veille technologique en architecture de calcul haute performance : Site de Clusterbuilder.org : http://www.clusterbuilder.org

Cluster de calcul du Centre de Nouméa : Extension de la capacité de calcul et développement d'outils logiciels. Demande des UR 103 et 65 du Centre de Nouméa, ''Spirales 2006'', Janvier 2006

Cluster de calcul pour données océanographiques à Nouméa, ''Spirales 2004'', Juin 2004.

Technologie InfiniBand :

InfiniBand: The Next Step in High Performance Computing, A Voltaire White Paper, February 2003
(http://www.hitachi-hitec.com/jyouhou/network/@gif/voltaire_20hpc.pdf)

Cluster Interconnects: The Whole Shebang
http://www.clustermonkey.net//content/view/124/33/

White Paper: Connectivity in the Multi-core Environment, Mellanox Technologie, 2007 (Document pdf, cf Annexe B)

ANNEXE A Proforma Alinéos : Passage réseau en technologie InfiniBand

Devis n° 5414 du 05/11/2007

Votre contact commercial : Fabien DEVILAINE
Mail : info@alineos.com – Tel. (+33) 1 64 78 57 65

solutions linux haute performance

IRD NOUMEA
A l'attention de Monsieur LEFEVRE

Evolution du réseau de calcul de votre cluster :

Configuration matérielle		
Item	**Caractéristiques**	**Quantité**
Réseau Infiniband 10Gb/s	**Fourniture d'un switch 24 ports SDR Infiniband 4X (10 Gb/s)**	**1**
Réseau Infiniband 10Gb/s_HBA	**Carte Dual 10Gb/s 4X InfiniBand Ports – PCI-X** *Câble d'interconnexion Infiniband – Longueur 5 mètres - inclus*	**10**

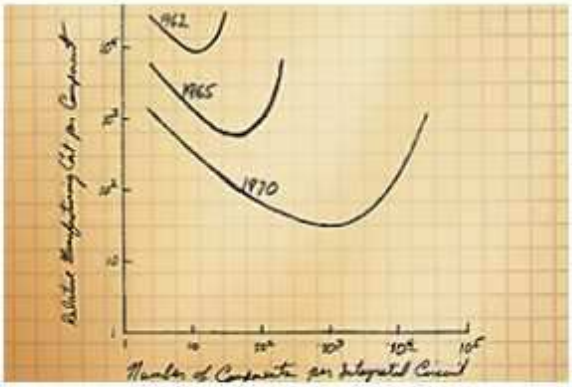
Garantie
La garantie générale (pièces, main d'œuvre et logicielle) est de **2 ans** pour l'ensemble des matériels. Elle comprend : L'échange des composants physiques défectueux, la mise en œuvre de corrections destinées à traiter l'anomalie ou l'incident ou tout autre intervention de maintenance.

Mise en service et exploitation
Total € HT : 7 990,00 *Frais de transport UPS Express Saver € HT + 800,00*

ANNEXE B White Paper: Connectivity in the Multi-core Environment, Mellanox Technologie 2007

"In 1978, a commercial flight between New York and Paris cost around $900 and took seven hours. If the principles of Moore's Law had been applied to the airline industry the way they have to the semiconductor industry since 1978, that flight would now cost about a penny and take less than one second." (Source: Intel)

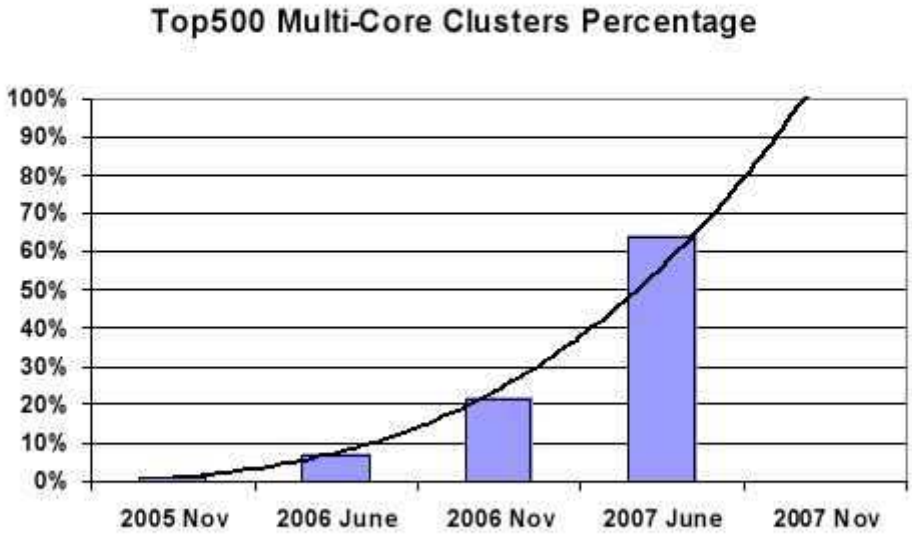
In 1965, Gordon Moore predicted that the number of transistors that could be integrated into a single silicon chip would approximately double about every two years. For more than forty years Intel has been transforming that law into reality. The increase in transistor density enables more transistors on a single chip and therefore increases in the CPU performance. However, it is not the only factor driving the CPU performance, as the increase of the CPU clock frequency, a bi-product of the transistor density was an important factor in the overall performance improvement.



Gordon Moore's original graph from 1965

High-performance computations are rapidly becoming a critical tool for conducting research, creative activity, and economic development. In order to provide intense computing platforms and still maintain the historic rates of performance and price/performance improvements, more execution cores are being integrated into each CPU. With multiple cores executing simultaneously, CPU clock frequency can be reduced in order to contain heat generation, while still increasing total system performance. This mega-trend is one of three trends that are shaping

the technical computing market – clusters, multi-core environments, and high-performance industry standard interconnects.



Connecting multi-core platforms

Efficient data transfer between clustered compute nodes is critical for balanced system performance. In a balanced system, the overall performance is equal to or greater than the sum of its components, while in a non-balanced system, the performance is less than the sum. The challenge of achieving balanced performance becomes more evident in multi-core environments. A multi-core environment introduces high demands on the cluster interconnect and the interconnect needs to be able to handle multiple I/O streams simultaneously.

By providing low-latency, high-bandwidth and extremely low CPU overhead, InfiniBand is emerging as the most deployed high-speed interconnect, replacing proprietary or low-performance solutions. In a multi-core environment, it is essential to avoid interconnect protocol processing in the CPU cores. In order to maximize the overall compute cluster efficiency and to allow performance-hungry applications to efficiently utilize the CPU's core resources, a fully hardware transport-offload solution is needed. Furthermore, un-necessary overhead on the CPU cores reduces the ability of balanced computing between the various cores, leading to higher degradation in real application performance.

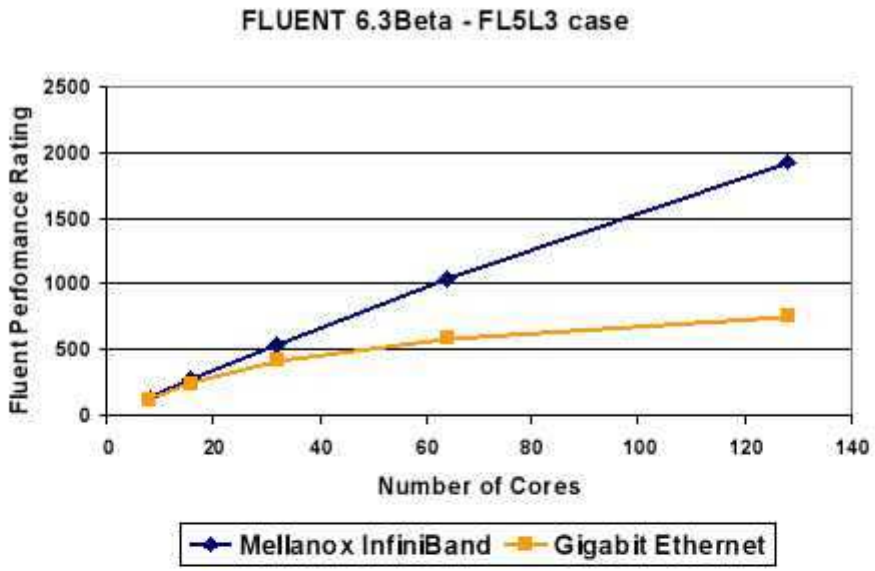
Interconnect flexibility is another requirement for multi-core systems. As various cores can perform different tasks, it is necessary to provide remote direct memory access (RDMA) along with the traditional semantics of Send/Receive. RDMA and Send/Receive in the same network provides the user with a variety of tools that are crucial for achieving the best application performance and the ability to utilize the same network for multiple tasks, such as compute, storage and management.

Mellanox InfiniBand provides both the flexibility and a full hardware transport-offload implementation. Transport-offload capabilities enable various applications and software interfaces, such as Message Passing Interface (MPI) to use overlapping of CPU computations with the interconnect communication cycles to reduce run time of MPI-based applications and to

increase application performance. Mellanox's hardware implementation also provides quality of service (QoS), so different I/O streams could be served as required by the application.

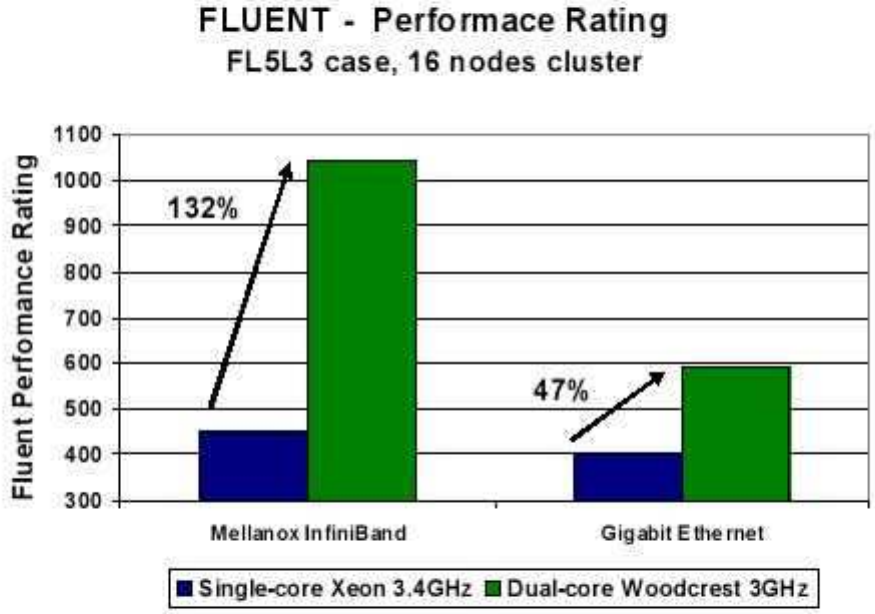
Applications demand a high-speed interconnect

Computational Fluid Dynamics (CFD) is one of the branches of fluid mechanics that uses numerical methods and algorithms to solve and analyze problems that involve fluid flows. At the core of any CFD calculation is a computational grid, used to divide the solution domain into thousands or millions of elements where the problem variables are computed and stored. FLUENT, a leading commercial software provider for solving fluid flow problems, implemented flexible parallel processing capabilities in order to effectively utilize the multi-core environments. Dynamic load balancing automatically detects and analyzes parallel performance and adjusts the distribution of computational cells among the processors and the server nodes. The following chart compares Mellanox InfiniBand and Gigabit Ethernet using FLUENT FL5L benchmark, on Intel dual-core Xeon 3GHz 5100 series (code name Woodcrest) servers.



Mellanox InfiniBand delivers superior performance than Gigabit Ethernet - up to 155% higher performance on 128 CPU cores - due to InfiniBand's proven efficiency and super-linear scaling capabilities.

In order to determine the importance of the interconnect architecture for multi-core environments, the same benchmark was used to compare between single-core Xeon 3.4GHz and dual-core Xeon 5100 series 3GHz (Woodcrest). In both cases, InfiniBand shows higher performance, but the difference between Mellanox InfiniBand and Gigabit Ethernet increases on the multi-core setting. In order to meet the requirements of each CPU core, multi-core servers demand higher I/O throughput from the interconnect solution. InfiniBand is proved to provide the aggregate CPU cores demands, while Gigabit Ethernet fails to do so.



As dual-core environments introduce higher I/O requirement than single-core ones and quad-core environments will increase that demand, a high throughput interconnect with low CPU overhead is vital in order to maintain high CPU and application efficiency.

Multi-core environments increase the demand for I/O throughput, low-latency, low CPU overhead, flexibility and high-efficiency in order to maintain a balanced system and to achieve high application performance and scaling. Low-performance interconnect solutions, or lack of native hardware support, will result in degraded system performance. Mellanox high-speed InfiniBand meets the multi-core system requirements and provides a balanced compute solution with Intel multi-core technology.

ANNEXE C Note technique sur l’optimisation du réseau de communication Gigabit du cluster de Nouméa

Nouméa, le 17 avril 2007

De : Jérôme Lefèvre (IRD - UR65, LEGOS Nouméa)
avec la participation de Pierrick Penven (IRD - UR097 ECO-UP)

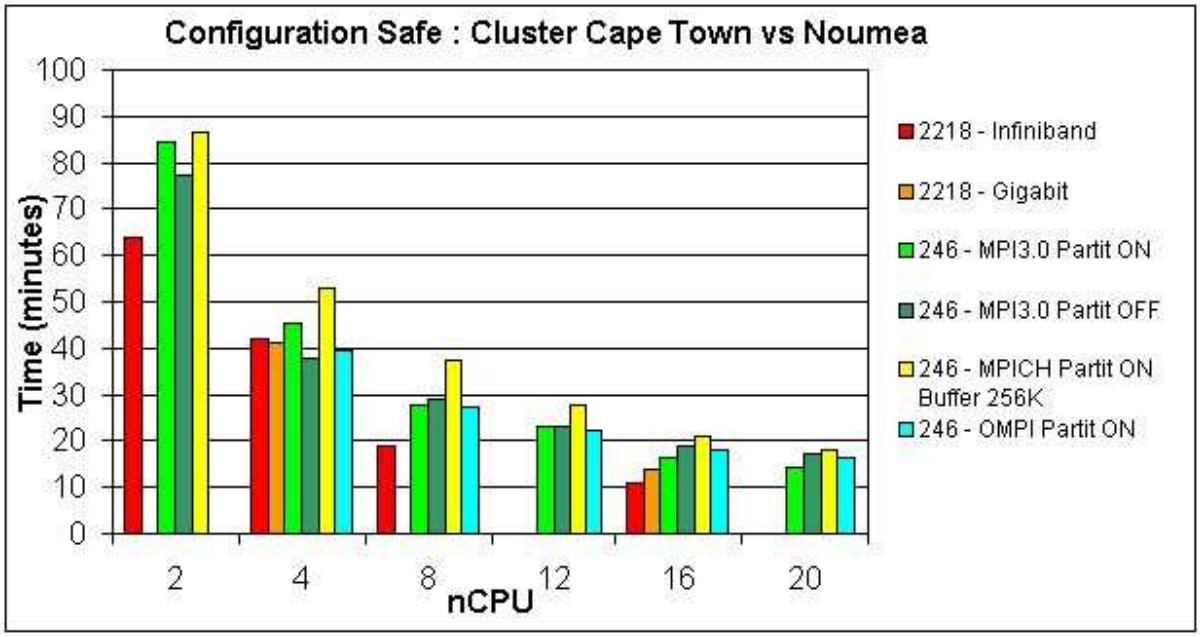
Chers Collègues,

Je profite d'un test comparatif entre le cluster de Nouméa et celui de Cape Town (grâce à Pierrick Penven) pour vous communiquer quelques observations et renseignements.

- Comparaison des performances de ROMS sur Cluster Nouméa vs Cluster Cape Town

Voici les résultats sur notre cluster de la configuration Safe(1) de Pierrick pour différents scénarios : 2,4,8,12,16 et 20 CPUs

(1) Safe : Configuration ROMS sur l'Afrique du Sud, grille 228 X 192 X 32, 1200 itérations pour un mois de simulation



En rouge et orange les temps de calcul obtenus sur le cluster de Cape Town (opteron 2218 Dual core, 2.6 Ghz, seconde génération), réseau InfiniBand et Gigabit.
Les autres motifs correspondent à nos simulations sur notre cluster à base d'opteron 246 Monocore (2.0 Ghz première génération), réseau Gigabit.

Quelques commentaires :

Evidemment, les performances sont meilleures dans le cas du cluster de Cape Town, dont les processeurs sont plus véloces : 17% de mieux sur 2 CPUs et l'écart se creuse avec le partitionnement : 33% plus rapide à 16 CPU mais ceci grâce à l'apport du réseau InfiniBand.

Sur le cluster de Cape Town, on peut noter un gain de 20% en recourant au protocole RDMA plutôt qu'au protocole TCP. Pour mémoire, l'InfiniBand repose sur le "Remote direct Memory access" qui offre

l'avantage de s'affranchir des tampons systèmes en s'adressant directement à l'espace de données de l'application utilisatrice.

Néanmoins, la différente atteint seulement 15% si l'on compare les 2 clusters à réseau égal (TCP/Gigabit) pour un découpage à 16 CPU.

Pour la configuration 4 CPUs, les 4 cores d'un seul noeud sont sollicités sur le cluster de Cape Town, alors que sur le cluster à base de 246, ce sont 4 CPUs monocoeur repartis sur 2 noeuds de calcul qui sont sollicités. La communication s'effectue par protocole TCP sur le cluster de Nouméa en empruntant le réseau de communication Gigabit Ethernet. On remarque une dégradation des perfs pour l'architecture 2218 lorsque les 4 cores du nœud de calcul 2 ways sont sollicités, par rapport au cluster de Nouméa.

Quelques explications, pour certaines récemment rencontrées sur le cluster de Nouméa, peuvent être avancées pour expliquer cette dégradation des performances :

- Un partage du bus mémoire devenu pénalisant sur cette architecture dual-core pour la configuration testée (mais peu probable)
- Une installation non optimisée de la librairie MPI avec la couche RDMA pour l'architecture InfiniBand
- l'absence d'optimisation à la compilation du code ROMS
- Une migration des processus d'un Cpu à l'autre, qui peut significativement ralentir l'exécution du code (possible sur les plateformes AMD avec les librairies MPICH et LAM, rectifié dans OpenMPI et MPI Intel)
- L'activation de l'option Powernow dans le bios et un mauvais réglage de l'option CPU Frequency Scaling dans le noyau linux 2.6 (laissé par défaut en mode PowerSafe et non en mode Performance par exemple)

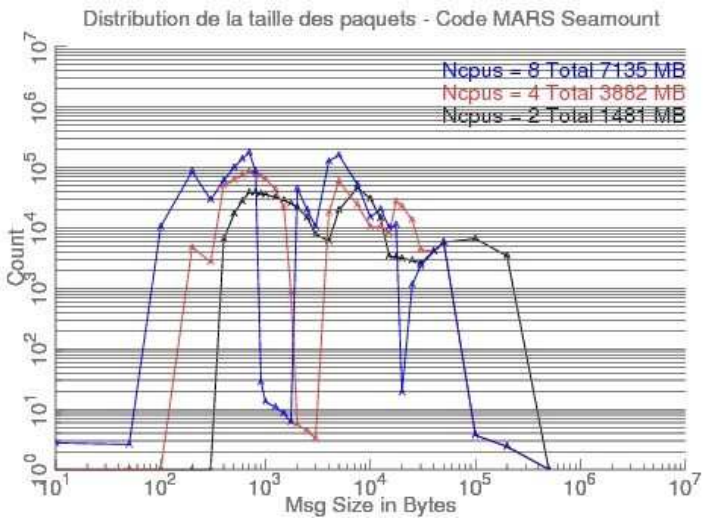
Si l'on regarde maintenant les performances du code à architecture égale (Cluster Nouméa) mais pour des couches de communication MPI variables, on peut établir le classement suivant des librairies MPI par ordre décroissant de performance :
1 - MPI Intel 3.0
2 - OpenMPI
3 - MPICH

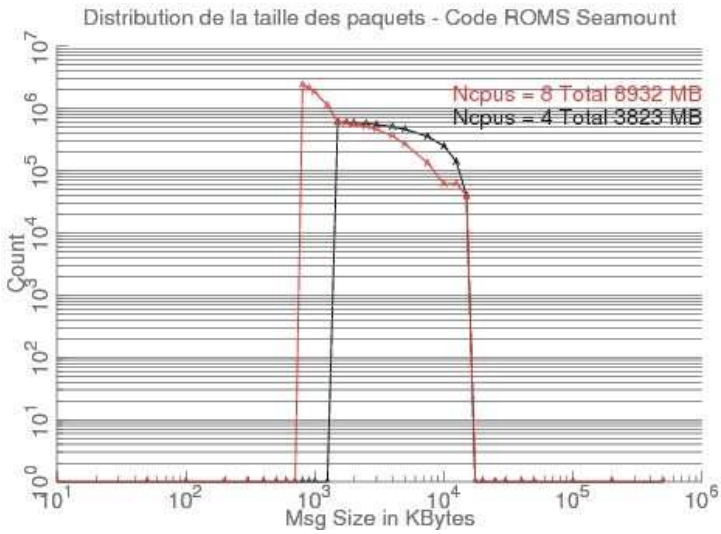
Si l'on regarde l'effet du partitionnement des fichiers pour l'optimisation des I/O, celui-ci devient significatif à partir d'une configuration à 16 noeuds ! Ce qui n'est pas un résultat attendu (On devrait s'attendre à une amélioration des performances dès 4 CPUs).

- Optimisation du réseau de communication Gigabit du Cluster Nouméa

J'ai récemment cherché à optimiser les performances du réseau TCP/Gigabit du cluster de Nouméa.
Ceci est motivé par les mauvaises performances obtenues sur notre cluster avec le code MARS de l'Ifremer. En effet, le code MARS montre une mauvaise performance et scalabilité sur notre cluster au-delà de 4 CPUs. Pourtant, sur le cluster Nymphéa qui emploi un réseau InfiniBand, la scalabilité du code MARS est bonne (Source Tina Odaka).

Pour expliquer en partie ce résultat, je retiens que la communication du code MARS emploie des messages de grandes tailles (>> 10000 bytes), contrairement au code ROMS dont la taille des messages n'excède pas 20000 bytes (voir graphes ci-dessous) à configuration égale.





Dans ces conditions, les avantages procurés par le réseau InfiniBand (temps de latence 10 à 100 fois plus réduits, bande passante 10 fois supérieure) sont évidents pour le code MARS.

Le réseau du cluster de Nouméa présente les caractéristiques suivantes :
NIC : Intel Pro/1000 MT Desktop RJ45 PW LA8391 MT - chipset 82541 sur PCI slot 32-bit 33MHz
Switch : 3COM superstack 3812
Cable : ethernet level 5
MTU : 1500

L'optimisation du réseau TCP/Gigabit a été améliorée en procédant aux interventions suivantes (classé par ordre décroissant d'impact) :

- Upgrade du driver e1000 de nos cartes réseau Intel : passage de version 6.3.9 à 7.4.35
- Réglage de InterruptThrottleRate sur 1 (Mode dynamique)
- Ouverture des buffers TCP réglés par défaut par le système :

    net.core.rmem_max = 10000000

    net.core.wmem_max = 10000000

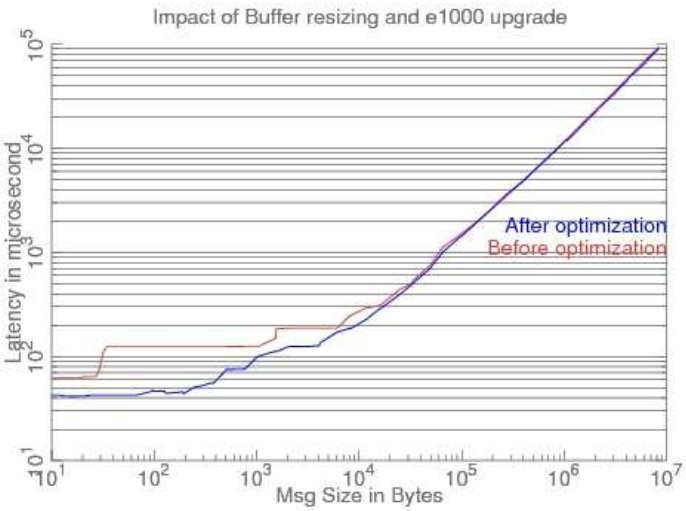
    net.core.rmem_default = 10000000

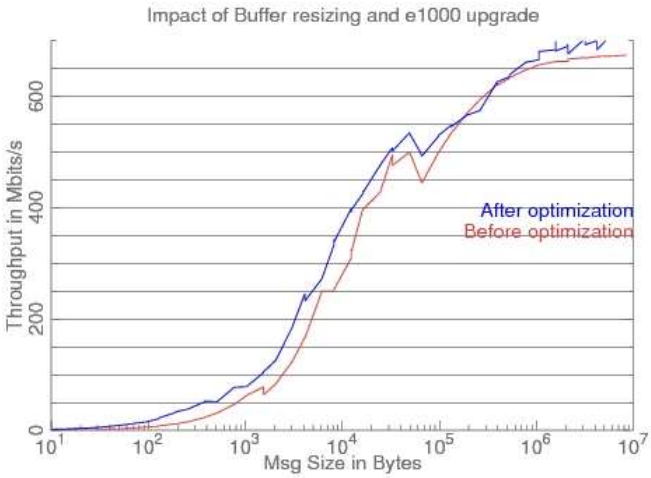
    net.core.wmem_default = 10000000

    net.ipv4.tcp_rmem = 10000000 10000000 10000000

    net.ipv4.tcp_wmem = 10000000 10000000 10000000

Les graphiques ci-dessous illustrent la réponse du réseau pour les couches TCP+MPI avant et après les interventions. La réponse du réseau de communication est dorénavant plus conforme avec celle croisée sur les forums de l'Internet (tests effectués avec NetPipe 3.6).





Quelques Liens utiles :

http://www.clustermonkey.net/content/view/124/34/ "Cluster Interconnect : The Whole Shebang"

http://www.clustermonkey.net/content/view/45/33/

http://www-didc.lbl.gov/TCP-tuning/linux.html

http://agenda.clustermonkey.net/index.php/Tuning_Intel_e1000_NICs

http://www.cita.utoronto.ca/mediawiki/index.php/McKenzie_Networking

http://support.intel.com/support/network/sb/cs-012904.htm