N° 9 - Juin 2008

La Lettre de l'IDRIS

Institut du Développement et des Ressources en Informatique Scientifique

www.idris.fr

Édito

Comme vous le savez déjà, le parc de supercalculateurs de l'IDRIS se renforce cette année par l'installation d'une plate-forme de calcul scalaire intégrée d'une puissance crête de 207 téraflops, constituée d'une composante massivement parallèle, le système Blue Gene, et d'une composante SMP Power6 qui est la suite logique de l'architecture Power4 en exploitation depuis quelques années. Je souhaiterais évoquer ici la manière dont nous voyons cette plate-forme de calcul repousser les frontières de la recherche scientifique en France.

Nous avons tous grandi et vieilli depuis plus de vingt ans sous l'emprise de la célèbre « Loi de Moore » : les processeurs de calcul doublaient leur puissance tous les dix-huit mois (croissance exponentielle). L'évolution naturelle des technologies de fabrication des puces permettait de réduire les tailles et d'augmenter les fréquences naturelles de fonctionnement de manière soutenue, pour aboutir à ce résultat. Mais, pour des raisons diverses (dont la génération de chaleur), les vitesses de fonctionnement des puces ne peuvent pas continuer d'augmenter de la même manière que par le passé. Nous arrivons aujourd'hui à la fin de la «Loi de Moore». C'est la raison pour laquelle les fabricants des microprocesseurs n'augmentent plus les fréquences de fonctionnement mais, en revanche, introduisent dans un «processeur» plusieurs unités de calcul et de traitement de l'information autonomes – appelées «cœurs» – pouvant travailler en parallèle. Les processeurs à deux ou quatre cœurs sont aujourd'hui monnaie courante. Vers 2015, n'importe quel PC portable sera équipé d'un processeur incorporant entre 128 et 256 cœurs. C'est-à-dire que ce sera une machine parallèle à part entière.

La conclusion est que le calcul parallèle et le traitement parallèle de l'information deviennent alors absolument incontournables pour « penser pétaflops ». La non moins célèbre «loi d'Amhdal» – qui établit les limites de l'augmentation des performances émanant du parallélisme – remplace la Loi de Moore. Dans ce contexte, la volonté du CNRS a été d'installer des systèmes qui poussent à l'extrême le parallélisme dans deux directions complémentaires : le parallélisme à mémoire distribuée (parallélisme massif) et le parallélisme à mémoire partagée à l'intérieur d'un nœud de calcul.

Le système Blue Gene, avec ses multiples réseaux d'interconnexion très performants et adaptés chacun à un type différent de communications (point à point, collectives, barrières, …) est la plate-forme idéale pour pousser l'extensibilité des codes bien au-delà des limites actuelles et aboutir à des applications qui s'exécutent sur des dizaines de milliers de processeurs. D'autre part, la multiplication des cœurs à l'intérieur d'une puce ouvre un boulevard pour le calcul parallèle à mémoire partagée dont la pertinence ne cessera de croître. Le système Power6, grâce à une nouvelle fonctionnalité architecturale appelée *symultaneous multi-threading,* permet de gérer avec efficacité des applications parallèles de type OpenMP qui engagent jusqu'à 64 threads (alors que pour la plupart des clusters SMP cette limite se situe aujourd'hui à 8 threads).

Il nous semble évident que le très long terme ne s'inscrit pas dans le droit fil de l'évolution des systèmes informatiques telle que nous l'avons connue jusqu'à présent, et qu'un certain nombre de paradigmes doit nécessairement évoluer. Mais il nous semble aussi que les nouveaux systèmes informatiques de l'IDRIS sont en plein accord avec les exigences et les expectatives actuelles du calcul scientifique de haute performance.

Victor Alessandrini

Sommaire

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

La Blue Gene redonne des couleurs à la chromodynamique quantique sur réseau en France

Par Philippe Boucaud et Olivier Pène, LPT/CNRS-MP (Orsay), Mariane Brinet et Jaume Carbonell, LPSC/CNRS-IN2P3 (Grenoble), Pierre Guichon, SPhN/CEA-IRFU (Saclay).

Les objets qui nous entourent, y compris les organismes vivants, sont tous faits d'atomes, c'est-à-dire un noyau de charge électrique positive autour duquel gravitent les électrons. Le noyau, qui porte pratiquement toute la masse de l'atome, est constitué de protons et de neutrons liés par une interaction forte, dont la théorie a été élaborée au début des années 1970. C'est la chromodynamique quantique (QCD pour "Quantum Chromo-Dynamics") dont certains des fondateurs ont été récompensés en 2004 par le prix Nobel de Physique (http ://nobelprize.org/nobel_prizes/physics/laurats/2004/index.html). Cette théorie, dont les particules élémentaires sont les quarks et les gluons, a été vérifiée en détail par des expériences avec les accélérateurs de très haute énergie. Elle doit aussi expliquer la cohésion des noyaux ainsi que la structure des protons et des neutrons, c'est-à-dire l'essentiel de la matière visible de l'univers. Son domaine d'application est même beaucoup plus vaste puisqu'elle contrôle la structure et les interactions de tous les hadrons : protons, neutrons, hypérons, pions, kaons, etc… Quand on sait que les hypérons, particules étranges et instables sur terre, constituent une fraction importante du cœur des étoiles à neutrons ou que les premières microsecondes de l'Univers ont vu le plasma de quarks et de gluons se transformer en hadrons, en omettant bien d'autres domaines essentiels, on ne s'étonnera pas que la QCD soit l'objet d'un immense effort théorique pour en comprendre tous les ressorts.

La QCD stipule que tous les hadrons sont composés de quarks et de gluons. Elle ne compte que sept paramètres : une masse pour chacun des 6 quarks et une constante de couplage qui règle l'intensité de l'interaction forte. La théorie permet d'interpréter un nombre immense de phénomènes physiques à partir de peu de paramètres et d'un formalisme mathématique bien défini et très compact. C'est l'une des théories physiques les plus élégantes de l'histoire des sciences.

Toute médaille a son revers et dans le cas de cette splendide théorie, il se manifeste par l'extraordinaire difficulté des calculs de QCD. Cette complexité est le prix à payer pour une théorie qui est capable de confiner les quarks et les gluons dans les hadrons, ce qui explique pourquoi ces particules ne peuvent pas être observées. La seule méthode de calcul systématique et rigoureuse est la chromodynamique quantique sur réseau (LQCD pour "lattice QCD").

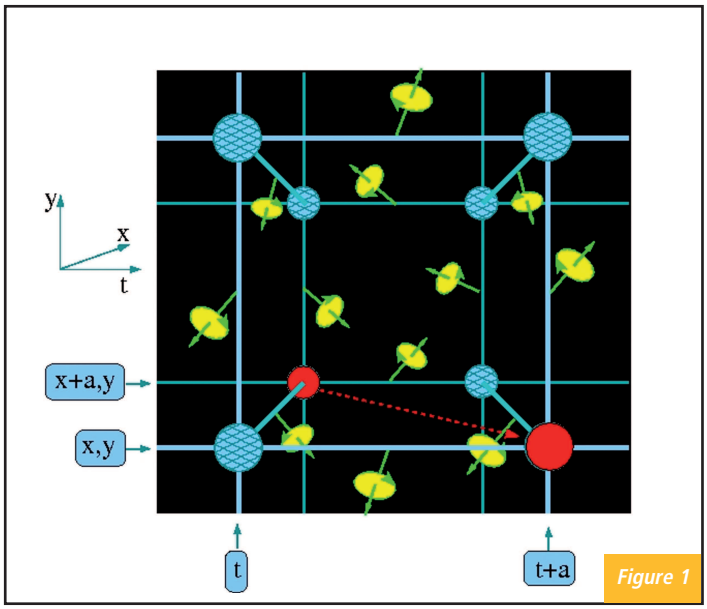


Figure 1

Figure 1 : *les matrices unitaires dans l'espace de couleur (gluons) schématisées par les toupies coloriées sont associées à chaque lien. Chacune interagit avec ses proches voisines selon leurs orientations relatives, un peu comme dans le modèle XY de la physique statistique. La figure montre 1 quark (rouge) qui saute de la position (x+a,y) au temps t à la position (x,y) au temps (t+a). Les autres sites de quarks sont vides dans cet exemple.*

Cette méthode, inspirée de la Physique statistique, consiste à représenter l'espace-temps par une grille quadridimensionnelle dont la longueur est typiquement de 3 à 5 fm (1 fm=10^{-15} m) et dont la maille mesure moins de 0,1 fm. Les gluons y sont représentés par des matrices unitaires 3x3 associées à chaque lien du réseau et les quarks sautent de site en site (voir figure 1). Les calculs, basés sur des méthodes stochastiques, sont très longs, en particulier quand ils prennent en compte l'apparition et la disparition de paires de quarks et d'anti-quarks virtuels. Ces fluctuations quantiques sont cruciales pour respecter des symétries fondamentales de la théorie mais le temps de calcul pour les traiter croît quand la masse des quarks diminue. Or la nature possède deux quarks particulièrement légers, le "u" et le "d", qui constituent l'essentiel des protons et neutrons.

De nos jours, un calcul pour un seul jeu de paramètres (masses des quarks, volume du réseau, longueur de la maille) dure des mois sur un ordinateur téraflopique, et pourtant les masses des quarks que l'on peut simuler sont encore trop grandes (voir figure 2). On anticipe ainsi la nécessité du pétaflop pour parvenir à des calculs véritablement réalistes.

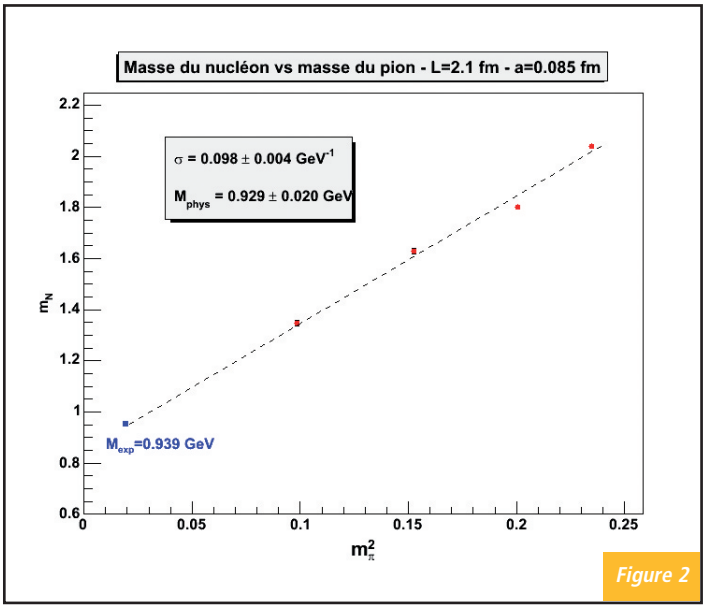


Figure 2

Figure 2 : *masse du nucléon (en GeV/c²) obtenue dans les simulations de LQCD en fonction de la masse carrée du pion $m^2\pi$, (en GeV²/c⁴) proportionnelle à la masse du quark u. La masse du nucléon a été corrigée d'un terme cubique $m^3\pi$ bien connu. Les calculs (en rouge) sont extrapolés jusqu'à la valeur physique de la masse du pion, pour donner une masse du nucléon de 929±20 MeV. Le point bleu représente la valeur expérimentale m = 939 MeV (non inclus dans l'extrapolation). La pente de la droite est le terme σ, paramètre essentiel en physique nucléaire, qui joue également un rôle en cosmologie en relation avec la matière noire.*

La LQCD existe depuis les années 1980. Très vite les physiciens ont réalisé que les ordinateurs commerciaux étaient trop chers pour permettre des calculs à un coût raisonnable. Au milieu des années 80, des projets d'ordinateurs dédiés sont apparus, dont un en Italie et un autre aux USA à l'université de Columbia. L'ordinateur italien s'appelait APE et notre groupe du LPT/CNRS à Orsay, avec un groupe de l'IRISA/INRIA à Rennes, a contribué, avec nos collègues italiens et allemands, physiciens et informaticiens, à la recherche et développement de la quatrième génération, l'"apeNEXT" qui est maintenant en activité. Dans le même temps, des collègues américains et britanniques développaient avec l'aide d'IBM la QCDOC *(QCD on chip)* selon une architecture assez semblable à celle de apeNEXT.

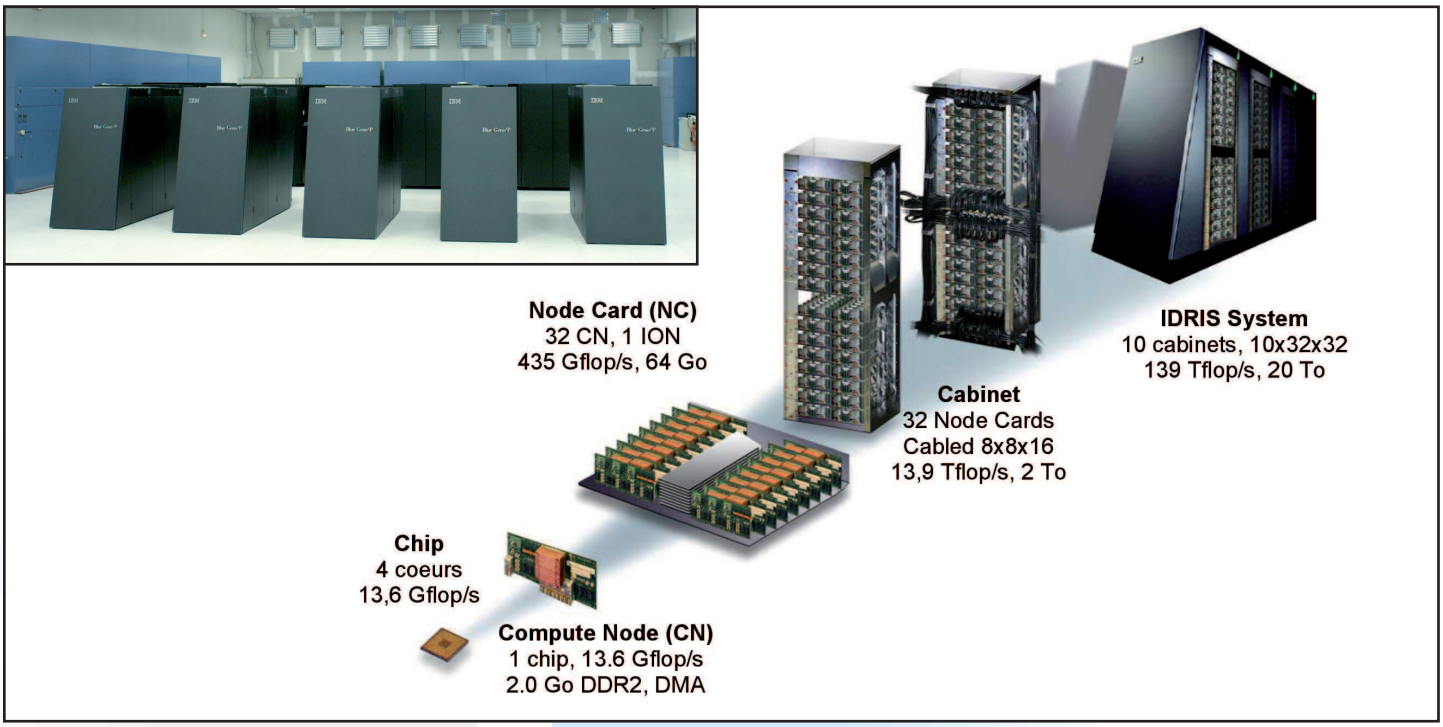
L'entreprise IBM a trouvé cette architecture intéressante. En partant de la QCDOC et avec la participation de certains de ses concepteurs, elle a développé la ligne Blue Gene. Un examen de l'architecture des Blue Gene révèle sa parenté avec les machines dédiées à la QCD : vitesse d'horloge modérée, réseau multi-dimensionnel de nœuds de calcul en grille, équilibre de la vitesse de calcul avec la vitesse d'accès à la mémoire et la vitesse des échanges entre nœuds. Les ordinateurs de ce type ont une faible consommation électrique et une faible occupation du sol. Le coût du teraflop est faible. Au vu de cette histoire, on ne sera pas surpris que nous ayons accueilli avec une joie non dissimulée l'annonce de l'achat, par le CNRS, d'une Blue Gene/P d'une puissance qui remet la France à niveau dans le domaine du calcul intensif.

Décrivons brièvement qui nous sommes et quel est notre projet. Trois laboratoires français, le LPT/CNRS-MP d'Orsay, le LPSC/CNRS-IN2P3 de Grenoble et le SPhN/CEA-IRFU de Saclay, collaborent avec des collègues allemands, italiens, britanniques, espagnols, chypriotes, suisses et néerlandais au sein de *l'European Twisted Mass Collaboration* (ETMC) (http ://www-zeuthen.desy.de/~kjansen/etmc/) sur un vaste programme de recherche basé sur une nouvelle façon de décrire les quarks. Cette collaboration a utilisé plusieurs ordinateurs européens, dont les apeNEXT, les Blue Gene/L de Jülich et de Groningen, la Mare Nostrum de Barcelone et la Bull Platine du CCRT. Nos groupes français ont contribué à la mesure des propriétés des nucléons avec un succès encourageant (voir figure 2).

Jusqu'à présent nous avons pris en compte seulement les fluctuations quantiques des quarks virtuels les plus légers "u" et "d". Nous entreprenons maintenant le calcul en considérant aussi les quarks "s" et "c" car nous suspectons que le quark « s » en particulier pourrait jouer un rôle important dans la structure du proton. Nous avons déjà une expérience de la Blue Gene/L grâce à nos collègues étrangers et aussi grâce à un test autorisé amicalement par EDF sur leur Blue Gene/L. Nous avons admiré l'efficacité de cet ordinateur. Nous n'avons pas encore d'expérience directe sur la Blue Gene/P mais les publications nous promettent une puissance effective environ cinq fois supérieure à celle de la Blue Gene/L. Nous espérons avoir bientôt la possibilité de faire directement les essais sur la nouvelle Blue Gene/P de l'IDRIS. Dans la compétition amicale entre nations lancées dans l'exploration de QCD sur réseau, l'accès à la Blue Gene/P sera un atout considérable pour les équipes françaises.

Babel, l'IBM Blue Gene/P de l'IDRIS

La nouvelle plate-forme de calcul scalaire de l'IDRIS est composée de deux architectures complémentaires. La première est une architecture massivement parallèle de type IBM Blue Gene/P (BG/P) d'une puissance nominale de 139 Tflop/s. Constituée de 10 cabinets, elle comporte 40 960 cœurs de calcul qui se partagent une mémoire de 20 To. La seconde est une architecture généraliste SMP de type IBM Power6 IH (P6IH) refroidie par eau, évolution de la plate-forme IBM Power4 Zahir actuellement en exploitation à l'IDRIS. D'une puissance nominale de 68 Tflop/s et dotée de 18 To de mémoire, elle est constituée de 112 serveurs répartis en 8 cabinets. Chaque serveur est un nœud SMP de 32 processeurs. 84 de ces serveurs embarquent chacun 128 Go de mémoire partagée alors que les 28 serveurs restant en embarquent le double, soit 256 Go. Ces deux architectures se partageront un système de fichiers global de haute performance (GPFS) de 800 To. En complément, un système de fichiers local de 400 To sera accessible depuis la P6IH. Dans la suite de cet article, nous nous intéresserons plus particulièrement à la BG/P, la machine P6IH sera elle traitée en détail dans un prochain numéro de la lettre de l'IDRIS.



Architecture de la Blue Gene/P (BG/P)

Le processeur PowerPC450

La brique de base de la machine BG/P est un processeur quadri-cœur SMP PowerPC 450, appelé dans la suite nœud de calcul ou « compute node » (CN). Cadencé à 850 MHz, il ne dissipe que de l'ordre de 33 W. La mémoire de 2 Go est partagée par les quatre cœurs, avec la flexibilité de pouvoir associer la totalité des 2 Go à un seul processus MPI. Chaque cœur possède un cache L1 de 32 Ko pour les instructions, un cache L1 de 32 Ko pour les données et un cache L2 de 2 Ko servant essentiellement à la gestion des streams et du prefetch. Le cache L3 de 8 Mo est partagé par les quatre cœurs du nœud. Les latences respectives des caches L1, L2, L3 et de la RAM DDR sont de respectivement 3, 11, 50 et 104 cycles. La bande passante mémoire théorique est de 13,6 Go/s pour le CN. Avec deux unités flottantes couplées double-précision de type FMA *(Fused Multiply Add)* par cœur, celui-ci peut délivrer jusqu'à 4 résultats par cycle, soit 3,4 Gflop/s (13,6 Gflop/s par CN). Il est à noter que le CN intègre physiquement les différentes interfaces réseaux ainsi qu'un moteur DMA (ce qui permet notamment de faire efficacement du recouvrement calcul/communication).

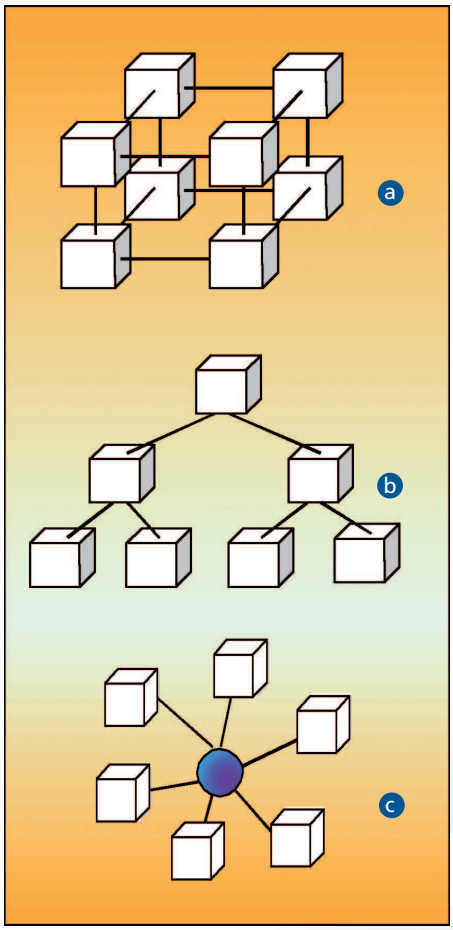
Design de l'architecture BG/P

Partons de la brique de base, le nœud de calcul (CN). Sur une carte Node Card (NC) sont assemblés 32 CN ainsi qu'éventuellement un nœud d'entrée-sorties IO Node (ION). Ces cartes (NC) sont interconnectées par groupe de 16 pour former un Midplane (MP). Enfin, deux Midplanes remplissent physiquement un cabinet (rack). Le système BG/P installé à l'IDRIS comprend 10 cabinets avec 16 ION par cabinet. Le nombre de cœurs d'un système BG/P est calculé de la façon suivante :

Nombre de cœurs = (nombre de cabinets) x (nombre de MP par cabinet) x (nombre de NC par MP) x (nombre de CN par NC) x (nombre de cœurs par CN) = 10 x 2 x 16 x 32 x 4 = 40960 cœurs.

Les cinq réseaux d'interconnexion de la BG/P

La Blue Gene/P est constituée de pas moins de cinq réseaux d'interconnexion, chacun étant optimisé pour un usage précis :



• **Réseau torique 3D (a) :** dans ce réseau, chaque nœud est physiquement relié à chacun de ses six voisins. D'usage généraliste, ce réseau est principalement utilisé pour les communications point à point. La latence varie de 64 ns pour un nœud voisin à 3 µs dans le cas le plus défavorable. Le débit est de 425 Mo/s pour chacun des 6 liens dans chaque direction, soit un débit total cumulé par nœud de 5,1 Go/s. À noter que ce réseau peut aussi être utilisé pour les communications collectives, au choix de l'utilisateur en fonction des caractéristiques de l'application et de la charge du réseau collectif. La plus petite entité constituant un vrai tore 3D physiquement câblé dans la machine est le MP, avec une répartition de 8 CN sur chacune des trois directions (i.e. tore 3D 8x8x8, soit 512 CN).

• **Réseau collectif (b) :** c'est un réseau haut débit *one-to-all* avec une topologie en arbre optimale pour les communications collectives de type *broadcast*/réduction et pour le transfert des entrées-sorties vers les nœuds d'IO. Chaque nœud possède 3 liens à 850 Mo/s par direction, soit une bande passante totale bidirectionnelle de 5,1 Go/s pour une latence inférieure à 2 µs.

• **Réseau d'interruption (c) :** c'est un réseau asynchrone séparé. Il a en charge le transfert de signaux rapides pour les interruptions et la synchronisation des processeurs. Les spécifications de ce réseau permettent d'obtenir d'excellentes performances de synchronisation (barrière globale) avec une latence MPI approximative de l'ordre de 2,5 µs pour 10 cabinets.

• **Réseau 10 Gigabit Ethernet :** ce réseau sert à interconnecter les ION avec les composants extérieurs (système de fichiers, station de contrôle). Les nœuds de calcul ne sont pas interconnectés à ce réseau.

• **Réseau de contrôle :** interface JTAG (1 Gb/s Ethernet). Ce réseau est utilisé pour le démarrage, le débogage et la surveillance des CN et des ION. Il permet aussi d'accéder aux compteurs de performance ou de surveillance du système sans aucune perturbation par rapport aux calculs en cours.

Mode d'éxecution et environnement logiciel

Pour l'utilisateur, l'accès à la BG/P se fait via une frontale et il n'est pas possible de se connecter directement sur les nœuds de calcul. Schématiquement, l'utilisateur soumet un travail depuis la frontale, ce travail consistant en une application MPI dont les différentes tâches sont distribuées sur les nœuds de calcul. La machine BG/P fonctionnant en mode dédié, pour chaque travail une partition d'exécution (composée d'au moins un ION et de plusieurs CN) est créée puis affectée à l'utilisateur pendant toute la durée d'exécution de son travail. Du fait de l'architecture de la machine, la plus petite partition qui puisse être réservée par le système pour un utilisateur est constituée d'un ION et de 64 CN. Même si l'utilisateur peut toujours n'en utiliser qu'une fraction, il lui sera comptabilisé les ressources réservées. Afin d'optimiser les performances, le système limite le nombre de processus MPI et le nombre de threads à une entité par cœur au maximum. Les CN peuvent être utilisés suivant trois modes d'exécution distincts :

• SMP : un nœud de calcul est vu comme un nœud SMP quadri processeurs. Il n'y a alors qu'un seul processus MPI par CN, processus qui peut accéder aux 2 Go de mémoire. Si on n'utilise pas de *multithreading*, seul un cœur physique est réellement utilisé sur les quatre que comporte le nœud. Pour utiliser les trois autres cœurs, il faut mettre en œuvre une parallélisation hybride de son application de type MPI entre les CN et *OpenMP/Pthreads* à l'intérieur du CN. Dans ce cas, on peut utiliser un maximum de 4 *threads* par processus MPI.

• DUAL : un nœud de calcul est vu comme 2 SMP biprocesseurs. Il y a alors 2 processus MPI par CN, chacun ayant accès à 1 Go de mémoire. Si le code n'est pas *multithreadé*, seul 2 cœurs sont utilisés sur les quatre disponibles. Pour utiliser les deux autres cœurs, il faut mettre en œuvre une parallélisation hybride de son application avec 2 *threads* par processus MPI.

• VN : un nœud de calcul est vu comme 4 monoprocesseurs indépendants, chacun n'ayant accès qu'à un quart de la mémoire disponible sur le CN. Le modèle de programmation est alors purement MPI (pas de possibilité de *multithreading*). Il y a alors 4 processus MPI par CN, chacun ayant accès à 512 Mo de mémoire.

Il est à noter que, sur cette machine, il n'existe pas de mécanisme de gestion de page virtuelle de manière à optimiser les performances (en particulier les accès aléatoires en mémoire) et que la mémoire est gérée suivant un modèle 32 bits (les entiers 64 bits étant toutefois supportés).

Suite page 6

Babel, l'IBM Blue Gene/P de l'IDRIS (suite)

Sur la frontale, le système d'exploitation est un système SUSE Linux (SLES version 10) qui offre à l'utilisateur les fonctionnalités suivantes : la compilation croisée pour la BG/P, la préparation, la soumission et le suivi des travaux. Le système d'exploitation sur la BG/P est lui aussi de type Linux. Sur les nœuds de calcul fonctionne un micronoyau propriétaire (CNK : *Compute Node Kernel*) supportant un environnement d'exécution complet : programmation *multithreading* (les appels systèmes sont à la norme POSIX), programmation par passage de messages (MPI), débogage, diagnostics, fonctions SMP, etc. Sur les nœuds d'IO tourne un vrai système Linux complet. Le gestionnaire de travaux, commun à la BG/P et à la P6IH, est LoadLeveler. Il est connu des utilisateurs de l'IDRIS puisque déjà en service sur la machine Power4 Zahir. Les compilateurs disponibles sur BG/P sont la suite de compilateurs du GNU (version courante gcc 4.1.1) ainsi que les compilateurs IBM XL en version V9.0 pour le compilateur XL C/C++ et en version V11.1 pour le compilateur XL Fortran.

Quelques éléments de performance de la BG/P

Les premiers résultats obtenus sur la BG/P sont évidemment partiels mais très intéressants. Suivant les types d'accès mémoire, les débits mémoire-CPU soutenus comparés à ceux de la machine Power4 Zahir (P655) sont les suivants :

- Accès contigus en mémoire : 2,2 Go/s, soit deux fois moins performant que Zahir (4,4 Go/s).
- Accès non contigus en mémoire avec un pas constant : 104 Mo/s, soit 1,6 fois plus performant que Zahir (64 Mo/s).
- Accès aléatoires en mémoire : 104 Mo/s, soit 3,5 fois plus performant que Zahir (30 Mo/s) !
- Sur 4096 cœurs (1 cabinet BG/P), on obtient sur le test *HPCC EP-STREAM Triad* un débit de 2,23 Go/s, soit une bande passante soutenue cumulée mémoire-CPU de plus de 9 To/s !

Concernant les performances du réseau d'interconnexion, sur un cabinet (4096 cœurs), on obtient pour le *HPCC RandomRing Bandwidth* une valeur de 20 Mo/s et 5,31 µs pour le test *RandomRing Latency*. Au niveau applicatif, plusieurs codes aux caractéristiques variées (langages Fortran90, C ou C++, paradigme de parallélisation de type MPI ou hybride MPI+OpenMP/Pthreads) ont été exécutés sur des configurations allant de 128 cœurs à plus de 16 000 cœurs. La performance moyenne comparée à celle de Zahir sur ce même panel de codes est de 0,75 (le facteur varie de 0,42 pour le moins bon des codes à 1,7 pour le meilleur). Pour certains codes, un simple travail d'adaptation et d'optimisation aux spécificités de l'architecture engendre des gains importants de performance. Enfin, les différents tests ont montré une extensibilité parfaite de l'architecture BG/P au sein d'un cabinet pour les codes qui s'y prêtent.

Finalement, à la vue des résultats déjà obtenus, cette architecture novatrice et très bien équilibrée semble plus généraliste qu'il n'y paraît. Première architecture massivement parallèle accessible à la recherche académique française, elle doit être le moyen pour la communauté scientifique de franchir le pas et de relever les défis du développement, du déploiement et de l'exploitation des applications à fort parallélisme massif (*Petascaling*).

DEISA2 s'engage pour un écosystème HPC européen

Communiqué de presse (21 mai 2008)

Distributed European Infrastructure for Supercomputing Applications

Un contrat supplémentaire de trois ans a été attribué par la Commission européenne dans le 7e programme cadre (EU FP7) au consortium DEISA en charge de l'Infrastructure Distribuée Européenne pour les Applications de Calcul Intensif. À partir du 1er mai 2008, le projet FP7 DEISA2 est engagé dans l'évolution de l'infrastructure HPC européenne vers un écosystème HPC intégré.

Dans le cadre du FP7, le consortium DEISA continue de promouvoir et développer l'infrastructure de calcul de haute performance ainsi que ses services à travers le projet DEISA2 financé pour 3 ans à partir de mai 2008. Les activités et services de type mise en œuvre d'applications, exploitation et technologies sont maintenus et développés car indispensables au support effectif des sciences computationnelles dans le domaine du HPC. Le modèle de fourniture de services sera étendu du support de type mono-projet simple au support de communautés européennes virtuelles. Des collaborations seront mises en place avec de nouvelles initiatives européennes et internationales. La coopération avec le projet PRACE qui prépare l'installation d'un nombre limité de supercalculateurs de classe Tier-0 en Europe est d'une importance stratégique. Les rôles et objectifs clés seront de délivrer une solution opérationnelle clé en mains pour un écosystème HPC européen persistant, comme suggéré par ESFRI. L'écosystème intègrera les centres nationaux Tier-1 et les nouveaux centres Tier-0.

DEISA : les réalisations sur lesquelles s'appuyer

Au printemps 2002, l'idée émergea de pallier à la dispersion des ressources de calcul intensif en Europe à la fois en termes de disponibilité des systèmes et de compétences nécessaires pour un support efficace au calcul intensif. La mise en place d'une Infrastructure Distribuée Européenne pour les Applications de Calcul Intensif fut proposée. En mai 2004 le projet DEISA a été lancé comme une initiative d'infrastructure intégrée de l'EU FP6 par huit centres européens leaders dans le calcul intensif. En 2006, DEISA a été rejoint par trois autres centres. Grâce aux efforts communs, DEISA a atteint rapidement un niveau de production de qualité pour le support d'applications de *capability computing* d'avant-garde au service de la communauté

scientifique européenne. DEISA a aussi contribué à la sensibilisation au besoin d'une infrastructure HPC persistante comme recommandée dans le rapport ESFRI de 2006. L'Initiative DEISA pour le Calcul Extrême (DECI), lancée en 2005, avec des appels annuels, a permis le support à des projets européens de calcul intensif extrêmes pendant les trois dernières années. Pour tous ces projets, les architectures HPC les plus puissantes et les plus appropriées disponibles en Europe ont pu être offertes. À ce jour, des scientifiques de 15 pays différents, avec des collaborateurs de 4 autres continents, en ont bénéficié.

DEISA2 : l'essentiel

Dans DEISA2, l'initiative DECI continue mais ces activités orientées mono-projet vont être étendues de manière qualitative vers le support persistant de communautés scientifiques virtuelles. DEISA2 leur fournira une plate-forme de calcul, offrant une intégration via des services distribués et des applications web, aussi bien qu'une gestion du stockage des données. L'accent sera mis sur la collaboration avec les projets d'infrastructures de recherche établis par le rapport ESFRI et les projets de grilles et de HPC européens. Cette activité renforcera les relations vers les autres centres HPC européens ainsi que les centres HPC internationaux leaders et les projets HPC mondiaux. Pour supporter les communautés scientifiques internationales par delà les frontières des politiques existantes, DEISA2 participe à l'évaluation et à l'implémentation des standards pour l'interopérabilité. La prise en charge de l'administration de l'infrastructure et du support de son utilisation efficace est la tâche des trois activités de service opérations, technologies et applications, qui sont complétées par deux activités communes de recherches.

Les membres de DEISA et les nouveaux partenaires associés

En plus des onze membres de DEISA appartenant à sept pays, BSC/Espagne, CINECA/Italie, CSC/Finlande, ECMWF/Royaume Uni, EPCC/Royaume Uni, FZJ/Allemagne, HLRS/Allemagne, IDRIS/France, LRZ/Allemagne, RZG/Allemagne et SARA/Pays-Bas, les trois centres CSCS/Suisse, KTH/Suède et JSCC/Russie ont rejoint le projet DEISA2 comme partenaires associés.

Plus d'informations : www.deisa.eu
Contact : Hermann Lederer (Lederer@rzg.mpg.de) et Stefan Heinzel (Heinzel@rzg.mpg.de) *Rechenzentrum Garching der Max-Planck-Gesellschaft* (RZG).

Remerciements :
Le consortium DEISA remercie la Commission européenne pour son soutien au travers des contrats FP6 RI-508830 et RI-031513, ainsi que du contrat FP7 RI-222919.

4è Symposium DEISA, les 28 et 29 mai 2008 à Edimbourg

Après Paris, Bologne et Munich, le symposium annuel du projet européen DEISA a eu lieu les 28 et 29 mai derniers à Edimbourg en Écosse. Le premier jour a été consacré à des interventions d'acteurs clés du domaine des infrastructures HPC. Lors de la seconde journée, des utilisateurs DEISA ont fait part à la fois de leur expérience de l'infrastructure DEISA et des résultats obtenus.

Les présentations, résumés et biographies des orateurs sont disponibles sur le site web DEISA :
www.deisa.eu/news_events/symposium/Edinburgh2008

Le prochain symposium annuel DEISA se tiendra à Amsterdam (Pays-Bas) du 11 au 13 mai 2009.

Journée Blue Gene/P du 8 avril 2008

Afin d'accompagner la mise en service du système Blue Gene/P, l'IDRIS a organisé en collaboration avec IBM France une journée d'information adressée à l'ensemble de la communauté scientifique nationale. Cette journée a eu lieu au siège du CNRS à Paris le 8 avril 2008. Différents thèmes y ont été abordés :

Description du système Blue Gene

- BG/P hardware : Jim Sexton, IBM Watson Research Centre, USA
- BG/P software : Jim Sexton, IBM Watson Research Centre, USA
- BG/P programmation : Jim Sexton, IBM Watson Research Centre, USA
- Amber-NAMD : Carlos Sosa, IBM Life Sciences, USA

Applications sur Blue Gene/P

- MPI-Blast : Carlos Sosa, IBM Life Sciences, USA
- QCD : Stefan Krieg, Rechenzentrum Julich, Allemagne
- GADGET : Clausion dalla Vecchia, University of Leiden, The Netherlands
- AVBP : Gabriel Staffelbach, CERFACS, France

Les séminaires de l'IDRIS

L'IDRIS organise de nouveau, à partir de juin 2008, des séminaires sur le thème du Calcul de haute performance dans ses locaux à Orsay. Les deux premiers ont eu lieu :

- jeudi 12 juin 2008 : *Current HPC architectures and a little bit beyond* par Aad van der Steen (université d'Utrech, Pays-Bas)
- jeudi 26 juin 2008 : *Recent trends in high performance microprocessor architecture* par William Jalby (LRC IT@CA, CEA/DAM et université de Versailles-Saint Quentin).

Les suivants seront annoncés après l'été sur le site web de l'IDRIS :
www.idris.fr > L'IDRIS > Les séminaires de l'IDRIS

Informations

Calendrier des formations IDRIS programmées d'ici fin 2008

Introduction générale à l'IDRIS	Intro	07/10/2008
Fortran de base	F95-1	30/09-02/10/2008
Fortran : apports des normes 90/95	F95-2	02-04/12/2008
Le Langage C	C	20-24/10/2008
Unix : utilisation	Unix-u	08-09/10/2008
Calcul parallèle : MPI-1	MPI-1	13-15/10/2008
Calcul parallèle : MPI-2	MPI-2	08-09/12/2008
Calcul parallèle : OpenMP	OpenMP	18-19/11/2008

Ces dates sont communiquées à titre d'information et sont susceptibles d'être mises à jour. Pour une information récente et plus complète, veuillez consulter le site web consacré aux cours donnés à l'IDRIS :

https ://cours.idris.fr

ou bien la rubrique « Cours de l'IDRIS » de notre serveur web :

http ://www.idris.fr

A partir de ce serveur, vous pouvez :
- consulter le catalogue des formations dispensées à l'IDRIS,
- vous inscrire en ligne au cours souhaité (si vous appartenez au CNRS ou à l'éducation nationale), après vous être créé un compte,
- accéder aux supports de cours.

Nous vous rappelons que les formations IDRIS sont gratuites pour les personnes appartenant au CNRS ou à l'éducation nationale. Elles sont aussi accessibles au personnel d'entreprises publiques ou privées via CNRSFormation Entreprises. Les conditions d'inscription sont dans ce cas consultables sur le site web :

http ://cnrsformation.cnrs-gif.fr

Demande d'abonnement

Si vous êtes utilisateur de l'IDRIS, vous recevez La Lettre systématiquement.
Sinon, envoyez-nous vos coordonnées postales par messagerie à : **La-Lettre@idris.fr**

Directeur de la publication : Victor Alessandrini
Rédacteur en chef : Thierry Goldmann
Rédactrice-adjointe : Geneviève Morvan
Comité de rédaction : Sylvie Brel, Serge Fayolle, Denis Girou, Pierre-François Lavallée
Conception graphique, réalisation, impression : CVS - www.cvs.com@wanadoo.fr

IDRIS - Institut du Développement et des Ressources en Informatique Scientifique
BP 167, Bâtiment 506, 91403 ORSAY Cedex - Fax : +33 (0)1 69 85 37 75 - www.idris.fr
Secrétariat : + 33 (0)1 69 85 85 05 - secretariat@idris.fr
Support utilisateurs : +33 (0)1 69 35 85 55 - assist@idris.fr