ORganisation Associative du Parallélisme

BI-ORAP

Bulletin d'information
de l'ORganisation Associative du Parallélisme

Numéro 35 - Avril 2003

Sommaire

Editorial

La mise en service du "Earth Simulator" au Japon, avec un budget de l'ordre de 350 millions de dollars, dans le calendrier et avec les performances prévues, a montré que la voie de l'assemblage de composants sur étagère (COTS) n'était pas la seule solution dans cette course à la performance que se livrent les constructeurs d'une part, les Etats-Unis et le Japon d'autre part. Alors que les Etats-Unis ont surtout misé, à travers le programme ASCI en particulier, sur les clusters de PC ou de SMP légers utilisant des processeurs de grande diffusion, le Japon et NEC ont pris la tête du TOP500, en performance crête (40 Teraflops) mais surtout en performance réelle sur Linpack (près de 36 Teraflops), en s'appuyant sur des processeurs vectoriels.

Il est probable que cette « pôle position » est acquise pour de nombreux mois tant l'avance acquise est grande (cet ordinateur est cinq fois plus puissant que l'ordinateur le plus puissant du programme américain ASCI, ou équivalent à la performance cumulée des douze premiers ordinateurs américains du TOP500 de novembre 2002), et on a parlé de "Computenik" en référence à la "première" soviétique "Sputnik" dans le domaine des satellites, qui a conduit les Etats-Unis à mettre sur pied un véritable programme spatial. Leur domination étant mise à mal, les Etats-Unis ne peuvent pas rester inactifs dans ce domaine stratégique de la simulation et du calcul de très haute performance.

Quelles seront les réponses des constructeurs, éventuellement en accord et avec le soutien financier de leurs gouvernements ? Quelles seront les stratégies pour arriver les premiers, d'une part aux 100 Teraflops, d'autre part à un nouveau seuil symbolique : le Petaflops (million de milliards d'opérations pas seconde) ?

Le Petaflops, dont on a commencé à parler vers 1994 et qui semblait alors un fantasme d'informaticiens rêveurs, devient un objectif réaliste. De nombreux Forums ORAP[1] ont montré l'intérêt de la performance réelle des ordinateurs, en particulier le dernier Forum (mars 2003) avec des applications telles que la recherche et l'industrie nucléaire (projet CEA/EDF de nouvelle génération de codes de calcul pour l'électronucléaire), ou la prévision météorologique.

Multiplier le nombre de milliers de processeurs n'est certainement pas la seule réponse : la machine "Earth Simulator" l'a prouvé, des scientifiques le clament de plus en plus haut (lors d'un récent séminaire en Grande Bretagne, des chercheurs grands consommateurs de ressources de calcul ont rappelé qu'ils avaient besoin de « capability systems » mais que, par manque de budgets, ils n'avaient accès qu'à des « capacity systems »). Quelles voies choisir ?

BI-ORAP a proposé aux principaux compétiteurs de présenter leur approche technologique, leur stratégie vers le Petaflops. CRAY, HP et IBM ont accepté de proposer un article, qui a pu être révisé en collaboration avec BI-ORAP (sur la forme et la taille). SGI et NEC n'ont pas souhaité apporter leur contribution, ce que nous regrettons. Ces articles n'engagent bien entendu pas la responsabilité des constructeurs concernés ni celle de BI-ORAP.

1.La majorité des "transparents électroniques" utilisés par les présentateurs dans le cadre de ces forums sont disponibles sur le serveur Orap : http://www.irisa.fr/orap

Nous espérons pouvoir présenter, dans un prochain numéro, le point de vue de plusieurs personnalités de la communauté française des utilisateurs du calcul de très haute performance.

Jean-Loïc Delhaye

Dossier : Cray

Après plusieurs années de fortes tempêtes, Cray revient sur la scène des constructeurs de superordinateurs, en confirmant une approche originale. Si NEC continue sur la voie des processeurs vectoriels rapides, avec une bande passante mémoire importante, ce qui se traduit par une complexité de fabrication et donc des coûts élevés, Cray préfère avoir des noeuds de calcul de taille plus petite comportant 4 processeurs très rapides, et associer vectoriel et parallélisme en conservant une large bande passante et un réseau interne très performant.

Pourquoi des Petaflops ?

Question trop négligée tant la course au “Peak Rate” (ce dernier étant trop mis en avant, au détriment de la performance utile, sur de “vraies applications”) occulte souvent le fond du problème. Les grandes performances sont nécessaires, d’une part pour pouvoir traiter certains problèmes de très grande taille, d’autre part pour résoudre des problèmes avec des temps de réponse adaptés aux contraintes temporelles de certaines applications (prévision météorologique, alerte tremblement de terre, etc).

Quelques exemples, bien connus, de domaines applicatifs concernés :

- Industrie automobile : test complet de crash, y compris l’interaction entre l’airbag et le passager; intégration de la conception, de la fabrication et du test d’un nouveau véhicule ; aéro-acoustique ...
- Industrie aéronautique et spatiale : conception complète et intégrée d’un nouvel avion, ...
- Génomique, industrie pharmaceutique : conception de nouveaux médicaments ; interactions protéine-protéine, dynamique moléculaire...
- Médecine : modélisation complète et précise du coeur et de la circulation sanguine ...
- Sécurité : cryptographie ...
- Météorologie, climat, environnement : prévision météorologique précise et à moyen/long terme, changements climatiques, détection des tremblements de terre, ...
- Recherche pétrolière, énergie
- Défense : simulation de systèmes d’armes, ...

Quelle(s) réponse(s) ?

Il n’y a pas une solution unique pour résoudre un champ d’applications très diverses dans leurs structures et leurs contraintes, et dans des cadres économiques tout aussi divers. Les grilles de calcul, quelque soit leur forme, apportent une réponse satisfaisante techniquement et économiquement dans certains cas ; les clusters de SMP (de 2 à 64 processeurs par SMP) ont largement montré leur intérêt depuis de nombreuses années ; les systèmes massivement parallèles sont également intéressants dans des contextes plus spécifiques. Mais ces différentes approches n’apportent pas une réponse suffisante à certaines applications : l’accès à la mémoire et le pipeline sont la meilleure solution pour des applications qui manipulent des vecteurs, particulièrement quand ces vecteurs sont longs ; l’architecture vectorielle permet alors d’obtenir une performance réelle assez proche de la performance théorique de la machine, ce que la machine Earth Simulator a bien montré.

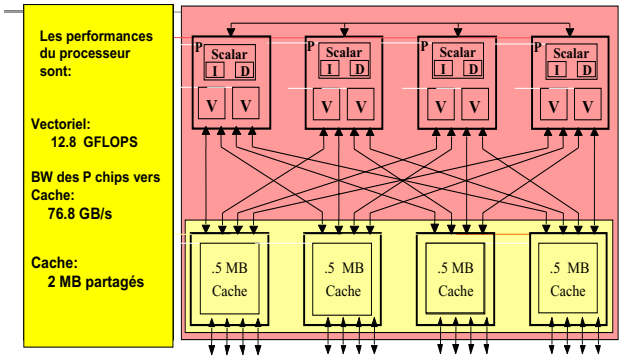
La puissance disponible ne cesse de croître, et ceci apparaît clairement dans les chiffres du TOP500. On peut donc penser, sans risque d’erreur, que le temps permettra d’arriver au Petaflops, du simple fait de l’évolution courante des technologies (processeurs, systèmes de communication interne, etc). Mais on peut aussi être plus volontariste et chercher des approches nouvelles qui permettent de “gagner du temps” et atteindre les 100 Teraflops réels, puis le Petaflops plus rapidement.

Le pari de Cray : proposer, dès 2010, un ordinateur ayant une performance efficace “soutenue” de 1 Petaflops. Ce pari se raccroche à l’histoire de la compagnie :

- 1990 : Cray YMP, première machine à fournir 1 Gigaflops efficace (application réelle)
- 2000 : Cray T3E, première machine à fournir 1 Teraflops efficace
- 2010 : 1 Petaflops efficace

Le Cray X1 : première étape

Le processeur du X1, dont le “fondeur” n’est autre qu’IBM, a une fréquence interne de 800 MHz. Il dispose de 4 processeurs élémentaires de calcul contenant 2 pipelines chacun, chaque pipeline pouvant délivrer deux résultats par cycle. Le processeur peut donc fournir 16 résultats par cycle soit 12,8 Gigaflops. On reste en technologie CMOS, l’innovation résidant dans le design du chip.

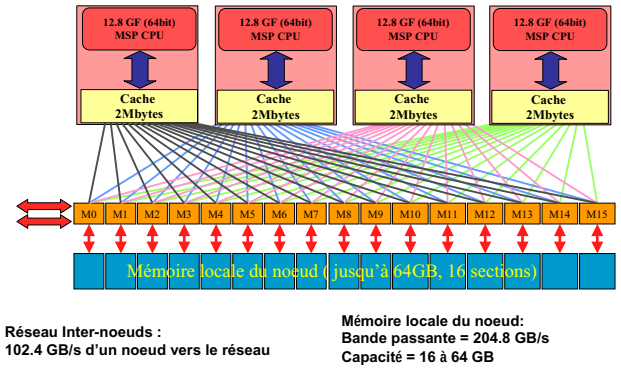


Le processeur

Chaque noeud du X1 comprend quatre processeurs, avec une mémoire locale qui peut atteindre 64 Go découpés en 16 sections. Le débit disponible entre le noeud et le réseau inter-noeuds est supérieur à 100 Go/s.

On dispose d’une interconnexion directe entre noeuds sur les systèmes jusqu’à huit noeuds ; au delà, on introduit des routeurs, le “coût” du passage par un routeur étant de 50 ns.

Le “packaging” compact permet d’avoir un faible encombrement au sol (un chassis, soit 800 Gflops sur moins de 3 m^2) et une faible consommation énergétique, et donc des coûts d’exploitation réduits.



Le noeud

La configuration maximale du X1 comprend 1024 noeuds (4096 processeurs) et permet d’atteindre 52 Teraflops théoriques.

Un système X1

Le logiciel : le système Cray X1 fonctionne en mode SSI (image système unique) ce qui simplifie la programmation d’une part, son exploitation d’autre part. Le système d’exploitation est Unicos, et il s’accompagne des outils d’administration, d’ordonnancement des travaux, etc.

Etat des lieux

Le Cray X1 a été annoncé en novembre 2002 et les premiers systèmes de production ont été livrés. La première “acceptation” d’un système de production X1 a été prononcée en avril par NCSI (Network Computing Services Inc.) pour le compte d’un centre de calcul de l’armée américaine (AHPCRC : Army High Performance Computing Research Center).

Parmi les premiers clients “non classifiés” :

- ARSC (Artic Region Supercomputing Center, Université de l’Alaska) qui a commandé un système à 128 processeurs (1,6 Teraflops) et 512 Go de mémoire centrale ; cette machine sera exploitée en parallèle avec un grand ensemble IBM basé sur des serveurs p655 ;
- ORNL (Oak Ridge National Laboratory, DoE) pour des applications de climatologie, de biologie, de matériaux, d’astrophysique, ... (256 processeurs). Les premiers résultats sur des codes réels (en particulier le “Parallel Ocean Program” de Los Alamos) utilisant les 16 premiers processeurs livrés ont donné d’excellents résultats d’après la presse américaine.

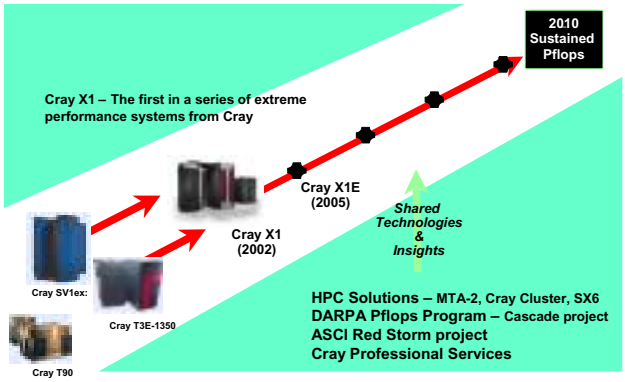
Il est prévu de livrer plus d'une vingtaine de "chassis" (un chassis peut contenir jusqu'à 16 noeuds) en 2003, dont huit à Oak Ridge.

Le X1 évoluera vers le X1e (fin 2004) qui, grâce à une plus forte intégration (8 processeurs par module contre 4 précédemment) et à une clock améliorée de 50%, aura une performance crête de 150 Teraflops en configuration maximale.

Sur le plan de l'entreprise, il a fallu un peu moins de deux ans (avril 2000 à début 2002) à Cray pour intégrer la fusion Tera-Cray, sur le plan des équipes comme sur le plan des produits. La nouvelle société comprend environ 850 personnes, dont 15 en France. Après des déficits en 2000 et 2001, Cray a réalisé un bénéfice de 5 M$ en 2002 (pour un CA de 155 M$) et les analystes prévoient pour 2003, un bénéfice de l'ordre de 20 M$ pour un CA de 220 M$.

La suite ...

Le schéma suivant présente la "roadmap" de Cray.



La gamme suivante, que nous appellerons "Y" (HPCwire cite le nom de code "Black Widow", dans un article du 28/2/03), pourrait sortir en 2006 avec une performance crête maximale de 300 Teraflops, et évoluer tous les 18 à 24 mois (Y2 à 500 Teraflops en 2008 ; Y3 à 1 Petaflops fin 2009).

La machine à 1 Petaflops réel ("soutenu") serait disponible en 2010 : il s'agit du projet "Cascade" réalisé avec le soutien de la Darpa américaine. Certains développements faits par Tera sur le multithreading seraient injectés dans ce projet.

Redstorm

Il s'agit d'une machine originale destinée aux Sandia National Laboratories, basée sur le processeur Opteron d'AMD.

Ce programme, d'un coût annoncé de 90 millions de dollars, aurait une puissance crête de 40 Teraflops obtenue par plus de 10000 processeurs cadensés à 2 GHz et 10 To de mémoire. Le système d'exploitation sera basé sur Linux et un système propriétaire de Sandia ("Catamount Operating System") sur les noeuds de calcul. La livraison de ce système commencerait mi-2004.

On est dans le scalaire massivement parallèle et on conserve un réseau d'interconnexion très puissant; cette technologie réseau spécifiquement développée par Cray pourrait d'ailleurs être partiellement utilisée dans la série "Y".

Redstorm donnera-il le jour à une gamme de machines commercialisées à destination des inconditionnels du scalaire, dont ceux du T3E ? On peut se poser la question !

Dossier : HP

Depuis l'apparition des premiers SuperCalculateurs "Gigaflopiques" puis "Teraflopiques" les besoins en calcul du monde académique comme industriel n'ont fait que croître et dépassent même les évolutions prévues par la loi de Moore. Nous observons depuis le début des années 90 une démocratisation des moyens de calcul rendue possible grâce à l'apparition des microprocesseurs RISC, la montée en puissance des microprocesseurs CISC et l'émergence du VLIW. Ces technologies ont permis de voir apparaître un nombre important de centres de calcul, qui aujourd'hui se développent et permettent d'adresser des besoins au jour le jour. Toutefois, un certain nombre de challenges scientifiques (étude de l'évolution de la Terre, maîtrise de l'Energie, décryptage du génome, simulation de phénomènes complexes ...) nécessitent la puissance de SuperCalculateurs qui n'existent toujours pas aujourd'hui. Les gouvernements américains et japonais ont ainsi créé des initiatives afin de supporter financièrement les besoins en recherche et développement pour résoudre ces grands challenges.

Le programme japonais basé sur la machine Earth Simulator a pour objectif de modéliser l'évolution de la terre. Le programme américain ASCI traite, quant à lui, le thème de la maîtrise de l'énergie.

HP est impliqué directement dans ces challenges et a décidé d'adresser ces programmes en fonc-

tion de la nature des enjeux. Notre démarche met en oeuvre des projets de recherches basés sur des partenariats qui aboutissent sur la mise au point de différents produits adaptés à chacun de ceux-ci.

Au début des années 90, le marché des SuperCalculateurs était essentiellement adressé par des systèmes SMP basés sur des technologies propriétaires, tant au niveau de leurs architectures que de leurs procédés de fabrication. L'émergence du CMOS dans un marché de production de masse, a permis de faire chuter le coût de fabrication des microprocesseurs RISC super scalaires. Leurs performances s'accroissant, les premiers microprocesseurs Alpha avec adressage sur 64 bits permettent alors d'envisager la construction de SuperCalculateurs (Digital, Cray).

Digital (puis Compaq) pénètre alors ce nouveau marché en introduisant les systèmes HPC160 et HPC320 (basés sur des noeuds quadri-processeurs Alpha). Ces deux systèmes représentent la première génération de SuperCalculateurs scalaires à base d'interconnexion de noeuds SMP. Le réseau haut débit alors employé développait un débit théorique de 80 Mo/s et permettait de relier jusqu'à 8 noeuds pour constituer une machine de 32 processeurs.

Conscient des avantages et inconvénients de cette technologie notamment au niveau de la scalabilité de la solution, nous avons candidaté en 1997 au programme ASCI pathforward[1] et notre projet a été retenu. L'objectif de ce programme étant l'étude de solutions d'interconnexion rapide pour système SMP, avec l'objectif d'aboutir à un réseau haut débit capable d'assembler un système d'une puissance crête théorique de 30 Tflops.

Ce contrat de recherche est à la base du développement des AlphaServerSC qui reposent sur l'interconnexion validée dans le projet ASCI pathforward. Un département spécifique s'est créé autour du calcul scientifique aux Etats-Unis à Marlboro, ainsi que des centres régionaux tels qu'Annecy en France pour le marché européen ou Kyoto (Japon) pour le marché asiatique accompagnés d'un important investissement en ressources humaines et matérielles à Galway en Irlande, pour le développement et le support logiciel des SuperCalculateurs HP.

Le projet SC nous a permis de nous structurer et de développer une première génération de SuperCalculateurs adoptés par la suite par un grand nombre de laboratoires de recherches et d'industriels sur le plan mondial ce qui nous a conduit à obtenir la place de n° 1 sur le marché du HPTC (Classements IDC 2001, 2002).

Parmi ces références, citons :

- Los Alamos National Laboratory (www.lanl.org)
- CEA (www.cea.fr)
- APAC (http://www.apac.edu.au/)
- PSC (www.psc.edu)
- Sanger Center
- Celera Genomics ...

AlphaServer SC au centre de calcul de Pittsburg

L'AlphaServerSC est devenu une des plates-formes de référence dans le domaine du calcul.

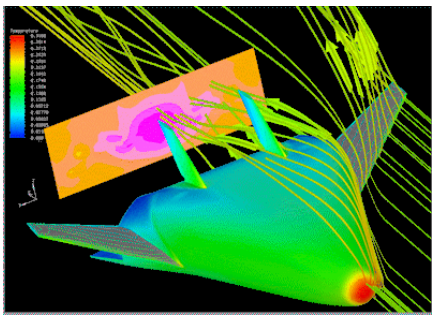
Nous observons aujourd'hui une nouvelle étape dans cette course à la puissance avec des besoins toujours plus grands qui nécessitent la conception de nouveaux systèmes avec un coût au Gigaflops toujours plus bas qui tend à marginaliser les architectures propriétaires. La montée en puissance du logiciel libre autour de Linux dans le domaine du calcul nous permet aussi d'améliorer nos coûts en recherche et développement en intégrant plus rapidement les nouvelles technologies et en externalisant une partie de notre recherche.

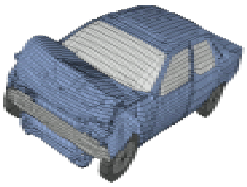
Fort de notre savoir-faire acquis avec le succès de l'AlphaServerSC, HP a lancé le nouveau programme XC (présenté lors du dernier salon SuperComputing 2002) basé sur des COTS (Component On The Shelf). Ce projet a pour objectif de réaliser la transition de la plate-forme SC vers un environnement autour de Linux basé sur des composants de masse (IA-32 ou IA-64).

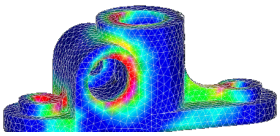
1.http://www.llnl.gov/asci-pathforward/

La conduite de ce projet est aussi basée sur différents contrats de collaboration entre des utilisateurs finaux du produit, HP et les éditeurs de logiciels. C'est pourquoi nous ciblons ce type de produits sur des marchés spécifiques afin de nous assurer de développer les bonnes fonctionnalités. Les marchés actuels sur lesquels nous nous concentrons sont les suivants :

- Biologie
- Oil and Gas
- Industrie
- Défense
- Simulation du climat
- Energie
- Education/Recherche







HP s'implique ainsi dans différents projets que sont les grands enjeux de la Recherche et de l'Industrie qui tireront parti du développement de ces technologies.

Parmi les principaux projets où nous avons une forte participation, citons :

- Le domaine des entrées/sorties avec les Global File Systems : LUSTRE (http://www.lustre.org)
- Le Portage et support de Linux en environnement Itanium SMP : Gelato (http://www.gelato.org)
- Le support à l'optimisation des codes avec les éditeurs de logiciels
- Les logiciels d'administration des Réseaux d'interconnexion à large bande passante et faible latence : (Quadrics, Myricom,...)
- Sécurité des systèmes informatiques

Chacun de ces projets représente des collaborations concrètes dans lesquelles HP s'implique. A ce titre plusieurs contrats de partenariat ont déjà été mis en oeuvre en France notamment avec le CEA, l'INRIA ou l'ESIEE[1].

Nous fournissons également des plates-formes de développement en partenariat, afin de mettre au point les technologies du futur qui permettront d'interconnecter entre eux les grands centres de calcul, afin de créer une communauté du calcul scientifique au sein d'une même grille (Inria Grenoble, PNNL USA,...).

D'une manière générale, nous ne pensons pas qu'il existe un ordinateur universel qui répondra à tous les grands challenges dans lesquels les scientifiques se sont lancés. Nous privilégions plutôt un dialogue fort avec nos clients, afin de réaliser au mieux, des systèmes informatiques qui correspondent à leurs besoins en terme de latence, de bande passante, de communication entre les processeurs, de localisation des données, de parallélisme ...

Jean-Marie Verdun et Philippe Devins (HP)

Dossier : IBM

C'est une véritable "vague bleue" qui a déferlé ces deux dernières années dans les grands centres académiques à travers l'Europe, en particulier en Allemagne, en Grande Bretagne et en France. Signe de ce dynamisme, le calcul haute performance a représenté en 2002 pour IBM quelques 550 millions de dollars de chiffres d'affaires dans le monde, ce qui en fait le numéro 1 mondial de la spécialité.

Les grandes références européennes et françaises

Parmi les installations récentes de supercalculateurs IBM à plusieurs Teraflops dans le secteur académique, citons : HLRN (4 Teraflops), Research Center Jülich (6 Teraflops), Postdam Institute for Climate Impact Research ou PIK (1 Teraflops), et IPP-Max Planck Society en Allemagne (3 Teraflops); CSCS Manno en Suisse (1 Teraflops), CSC en Finlande (2 Teraflops), HPCx (7 Teraflops), ECMWF (European Center for Middle-range Weather Forecast, 6 Teraflops), l'Université de Cambridge (2,5 Te-

1.http://www.cea.fr/fr/actualites/articles.asp?orig=com&annee=2002&id=271

raflops) et AWE (Atomic Weapon Establishment, (3 Teraflops) en Angleterre. Pour l'Europe du Sud, citons les centres Cineca (Bologne) et Caspur (Rome) en Italie, et le CEPBA (Barcelone) en Espagne.

Le cluster de 40 serveurs p690 destinés au consortium britannique HPCx rassemblés pour leur test à l'usine IBM de Ploughkeepsie aux Etats Unis

En France, l'IDRIS, centre national de calcul du CNRS, dispose d'un supercalculateur développant plus de 1,3 Teraflops à base de noeuds p690 (8 p690 32 processeurs soit une configuration totale de 256 processeurs POWER4 à 1,3 Ghz).

Le CINES, centre national de calcul des universités, vient d'acquérir deux p690 32 processeurs pour renforcer un RS6000/SP à 472 processeurs POWER3-II.

Le CNES à Toulouse, le pôle M3PEC de l'Université de Bordeaux I, le Centre de Ressources Informatiques de Haute Normandie (CRIHAN), l'Université des Sciences et Techniques de Lille, le Centre de Calcul Recherche et Réseau (CCR) de l'Université de Jussieu, Dassault Aviation, Beicip Franlab (filiale de l'IFP), Total Fina Elf à Pau, SNECMA Moteurs, Airbus à Toulouse, le SHOM ont fait l'acquisition de serveurs "Regatta" à base de POWER4.

Les raisons de ces succès tiennent à l'adéquation de l'offre IBM actuelle (entièrement renouvelée entre 2001 et 2002 d'ailleurs) aux besoins des grands centres de calcul : performance, pérennité, production, évolutivité.

Les deux futurs poids lourds du Top500 : ASCI Purple et Blue Gene/L

IBM s'apprête à concevoir les deux supercalculateurs les plus puissants du monde, soit quatre centre soixante-sept mille milliards d'opérations en virgule flottante à la seconde (467 Teraflops) pour un montant de 216 à 267 millions de dollars, et à reprendre ainsi la tête du TOP500.

Le premier des deux systèmes, nommé ASCI Purple, affichera une vitesse de calcul maximum de 100 Teraflops, contre moins de 40 pour le supercalculateur le plus puissant du monde, le NEC Earth Simulator 5120 japonais. Occupant tout un bâtiment en cours de construction en Californie, ce cluster de 196 unités pour un total de 12.544 microprocesseurs POWER5, 196 noeuds avec une bande passante mémoire totale de 156.000 Go/s, 50 To de mémoire et 2 petaoctets stockés sur disques. Il servira à simuler les explosions de bombes thermo-nucléaires.

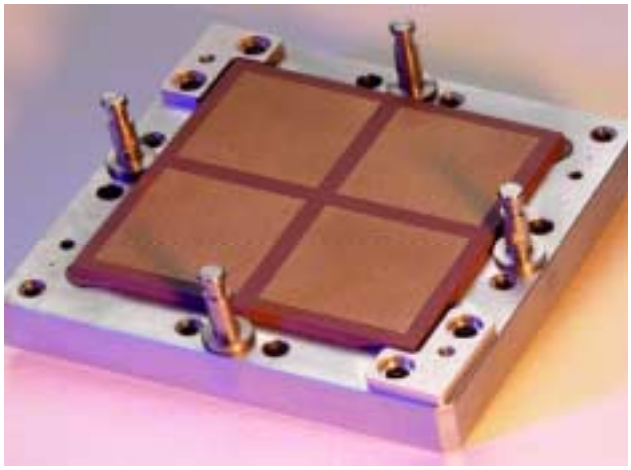
Le second supercalculateur, Blue Gene/L, utilisera dans sa version finale 130.000 processeurs exploitant Linux pour atteindre une vitesse maximum théorique de 367 Teraflops, quasiment 10 fois la performance du plus puissant supercalculateur actuel, et pas loin de cinquante mille fois celles du plus gros système installé en France il y a dix ans. Il traitera des données à la vitesse de 1 Tbit/s. Il sera utilisé pour des calculs portant sur des phénomènes intéressant les physiciens, tels que les écoulements turbulents ou ce qui se produit lors d'une explosion.

Ces deux systèmes combinés représentent une fois et demie la puissance de la liste actuelle complète du Top500.

Performance

Le design révolutionnaire de la technologie POWER4 préfigure la remarquable densité des serveurs de la famille "Regatta": 184 millions de transistors sur une puce bi-processeurs (dont la gravure ne dépasse pas 0,13 microns) avec cache L2 intégré de 1,5 MB. IBM a également développé une technologie d'interconnexion spécifique pour être à la hauteur des caractéristiques de débit du POWER4 : les MCM (Multi Chip Module).

Chaque MCM héberge 4 puces, créant ainsi un multiprocesseur symétrique, de 4 processeurs ("single core" pour une bande passante optimale, destinée à renforcer la tenue en charge ou aux applications sensibles à la bande passante) ou de 8 processeurs ("dual core" pour maximiser le nombre de processeurs), qui disposent d'un cache L3 et 128 MB. Le design "dual core" fait d'ailleurs des émules puisqu'il a été repris récemment par Intel pour son Itanium. Le "MCM" côtoie désormais le "Single Chip Module (SCM)", dont la granularité plus fine (blocs de puce bi processeur) le destine à habiller les serveurs milieu de gamme de 2 à 8 processeurs comme le p650.



Le Multi Chip Module (MCM), “brique de base” des serveurs IBM de la famille “Regatta”

Les 32 processeurs POWER4 à 1,7 Ghz du serveur p690, le modèle SMP haut de gamme du constructeur, développent une puissance de crête de 230 Gigaflops pour une surface au sol minime : pas plus 1,2 m2. Le serveur p655, modèle massivement parallèle (MPP) de la gamme, est même encore plus “concentré” : il peut quant à lui aligner sur une surface au sol équivalente jusqu’à 128 processeurs POWER4 à 1,5 Ghz, soit une puissance de crête de près de 800 Gigaflops !

A cet avantage s’ajoute celui de l’emploi de composants “dernier cri” permettant de réduire la dissipation calorifique, la consommation électrique et d’augmenter significativement les fréquences d’horloge : technologie cuivre (brevet industriel IBM vendu à ses concurrents d’ailleurs), isolants Silicon On Insulator ou Low K Dieletric pour les nouvelles générations, etc. C’est là tout l’intérêt d’une technologie en début de vie dont les capacités de progression sont devant elles.

Tableau 1 : Comparaison des caractéristiques POWER4/ POWER4+

Spécification	POWER4	POWER4+	Rapport
Taille (mm)	475	267	-44%
Consommation (Watts)	115	70	-40%
Cache L2	1,42	1,5	+5%
Queues L3	32	48	+44%
Contrôleurs mémoire (MHz)	433	566	+30%

L’après POWER4

Le processeur POWER5 fonctionne dans les laboratoires IBM depuis le mois de janvier de cette année. Les systèmes à base de cette technologie seront donc livrables comme annoncé dans le plan produit du constructeur à compter de la deuxième moitié de 2004. Les premiers systèmes, packagés sous la forme de noeuds SMP à 64 processeurs, seront livrés dans le cadre du projet “ASCI Purple”. Une deuxième génération de POWER5 fera son apparition en 2005 avec des fréquences d’horloge supérieures à 3 Ghz.

Le développement du processeur POWER6 est lui commencé depuis 2001. Le POWER6 devrait faire son apparition vers 2006 avec des fréquences d’horloge de l’ordre de 6 Ghz.

Ainsi IBM peut s’engager par contrat sans crainte de pénalités à la fourniture de puissance de calcul “à la demande” évoluant d’après un calendrier d’installation pré défini, grâce à ses roadmaps POWER4, POWER5, POWER6, qui habilleront les générations de serveurs se succédant tout au long du projet.

L’engagement pris par IBM dans des contrats titanesques signés avec le gouvernement américain, comme ASCI Purple et ses 100 TFlops, apportent également des garanties sur le devenir de la technologie à toute la communauté des utilisateurs.

Production

L’expérience d’IBM dans les très grands clusters de production comme ASCI White (un système RS6000/SP de 512 noeuds à 16 processeurs chacun soit 8192 processeurs en tout), resté n˚1 du Top500 pendant plusieurs années, lui confère un avantage décisif par rapport à ses concurrents. Ces supercalculateurs de taille exceptionnelle sont comparables à ce que la formule 1 est à l’automobile : un terreau où l’on invente et éprouve les technologies les plus avancées, qui se diffusent ensuite dans les secteurs plus traditionnels.

L’environnement logiciel “cluster” d’IBM, hérité des RS6000/SP, est par conséquent mature et robuste. L’administration est centralisée via le logiciel “Cluster System Management” (CSM) ; le système de fichier global et parallèle General Parallel File System permet de construire un espace disque performant et sécurisé ; la charge est répartie par Work Load Manager (WLM) et les ressources affectées via le soumissionnaire de tâches LoadLeveler. D’autant qu’avec la reconnaissance de Linux cet environne-

ment scientifique est commun aux plates-formes POWER4/AIX, Intel/Linux et même POWER4/Linux. Une aubaine lorsqu'on doit gérer un parc hétérogène de supercalculateurs ou que l'on s'intéresse au GRID.

Duel au sommet du 64 bit : POWER4 versus Itanium

Le basculement des technologies RISC 64 bits vers Intel 64 bits (Itanium) pose inévitablement des problèmes. Seul IBM garantit la compatibilité binaire entre ses générations de processeur POWER, qui utilisent des jeux d'instructions rigoureusement identiques. Le patrimoine des codes scientifiques, qui représente souvent des années d'investissements, est donc préservé, la rupture de technologie évitée, et la production des codes stabilisée.

L'offre logicielle autour de l'Itanium est encore très limitée, en particulier l'environnement de développement (compilateurs, bibliothèques scientifiques, analyseurs de performances, etc.). Le portage des codes sur cette nouvelle architecture 64 bits, en particulier des codes parallèles, sera difficile sans un environnement industrialisé.

D'autre part, les prix des serveurs à base d'Itanium sont encore loin d'égaler ceux des serveurs Intel 32 bits, dont ils sont pourtant les successeurs avérés, ce qui ne les différencie pas outre mesure des technologies RISC.

IBM prend également le pari du marché de commodité pour son POWER4, qui sera étendu à des plates-formes mono processeur exploitant Linux (Suse ou Redhat), reconnu par le géant d'Armonk comme un système d'entreprise dans lequel il investit massivement. Une stratégie qui lui permettra de faire baisser le prix de sa technologie POWER4 en assurant une grande diffusion.

L'arme secrète d'IBM sera sans aucun doute son serveur "Blade", une architecture supra-dense puisque qu'un même châssis de 7U pourra contenir 14 "lames" bi processeurs, soit 28 processeurs en Pentium IV DP Xeon 2,4 Ghz avec réseau Gigabit Ethernet et réseau d'administration intégré pour le calcul 32 bit, soit en POWERPC 970 à 3 Ghz (une version plus dense du POWER4) à 3 Ghz avec réseau Myrinet intégré pour le calcul 64 bit ! Le tout fonctionnant sous ... Linux bien sûr. Nul doute que l'annonce de ce produit fera l'effet d'une bombe ...

En conclusion : vers une nouvelle ère ? Le "calcul haute performance à la demande"

Pour les industriels ou les centres de recherche ayant des besoins irréguliers en puissance de calcul, IBM va commercialiser des services de calcul à la demande qui permettront à ses clients de se dispenser de l'achat, de l'exploitation et de la maintenance de supercalculateurs.

Si le marché réclamait depuis longtemps cette méthode de "paiement à l'usage", calculée en fonction de la capacité requise et de sa durée d'utilisation, sa mise en oeuvre semblait encore irréalisable il y a peu. Les clients disposent ainsi d'une ressource "virtualisée" à laquelle ils peuvent faire appel à volonté, et qui, en termes comptables, fait passer le calcul haute performance du poste budgétaire des coûts "fixes" vers celui des coûts "variables". Le strict nécessaire de puissance de calcul est prélevé, ce qui évite d'avoir à acquérir un supercalculateur dimensionné en fonction d'une puissance de crête ou de pics de charges dont ils n'ont pas forcément besoin dans leur production quotidienne.

En outre, certains secteurs d'activité, comme l'industrie du pétrole, l'industrie numérique et les sciences de la vie, ont besoin de supercalculateurs, mais uniquement à certaines étapes du cycle de développement de leurs produits. Le reste du temps, les supercalculateurs sont sous utilisés.

La première compagnie à avoir franchi le pas du calcul haute performance à la demande fourni par IBM est PGS Data Processing, une division de Petroleum Geo-Services, pour un projet d'imagerie sismique avancée dans les eaux profondes du Golf de Mexico.

Afin de fournir la capacité de traitement nécessaire pour le calcul haute performance à la demande, IBM va construire une grille de processeurs Intel et POWER. Cette grille sera constituée de centaines de systèmes pSeries p655, le serveur Unix surpuissant capable d'aligner 128 processeurs dans une seule armoire, et d'un énorme cluster linux basé sur des xSeries x335 et x345, serveurs multi processeurs Intel Xeon. Le premier site qui hébergera le matériel derrière le "calcul haute performance à la demande" sera à Poughkeepsie dans l'état de New York, suivi par d'autres à travers le monde qui seront à terme tous reliés entre eux.

Lionel Nouzarède (IBM France)

Bull annonce la gamme NovaScale

Le 21 mars, Bull a annoncé la gamme NovaScale, une nouvelle gamme de serveurs SMP multi-processeurs à base de processeurs Intel Itanium-2.

L'architecture des serveurs NovaScale a été conçue et développée par Bull en partenariat avec Intel dans le cadre du programme FAME (Flexible Architecture for Multiple Environments) et permettra d'intégrer les évolutions attendues des processeurs de la famille Itanium. Aujourd'hui, les serveurs NovaScale embarquent les processeurs 64-bit Itanium-2 (900 MHz ou 1 GHz au choix, et trois niveaux de cache intégrés sur la puce) ; ils sont prêts à accueillir les processeurs Itanium de 3ème génération ("Madison") dès l'été 2003. L'architecture FAME offre de multiples possibilités d'évolution : systèmes multiprocesseurs SMP jusqu'à 32 processeurs d'une part, assemblage en grappes de systèmes d'autre part (clustering).

Dans un premier temps, Bull annonce trois modèles :

- le serveur NovaScale 4040 : modèle compact intégrant jusqu'à 4 processeurs Itanium-2 ;
- les serveurs NovaScale 5080 et NovaScale 5160, basés sur une architecture parallèle et redondante, supportent respectivement jusqu'à 8 ou 16 processeurs.

Bull fournit le système d'exploitation 64 bits Linux standard, optimisé pour ses serveurs NovaScale, et propose un environnement logiciel de développement (compilateurs, bibliothèques, outils de mise au point et d'optimisation) et de production (administration, surveillance, ordonnancement, ...). En proposant plusieurs interconnexions (Gigabit Ethernet, Scali/SCI et Quadrics), Bull permet de choisir l'architecture de cluster en fonction de la granularité des applications envisagées.

http://www.bull.com/fr

Actualités Bi-Orap

➡ Lecture

"*Turbulence*" : un livre qui couvre largement les grands problèmes de la turbulence et traite notamment de la physique des phénomènes, de leur modélisation et de leur simulation. Les trois derniers chapitres synthétisent les techniques numériques et expérimentales actuelles. Ecrit par Christophe Bailly et Geneviève Comte-Bellot, ce livre est édité par CNRS Editions (Collection Sciences et techniques de l'ingénieur).

➡ Prix du CINES

Le prix du CINES (Centre Informatique de l'Enseignement Supérieur) a été remis le 31 mars, par Elisabeth Giacobino, Directrice de la recherche, à Paul Indelicato, directeur de recherche au CNRS et membre du laboratoire Kastler Brossel (ENS et Université Paris 6).

➡ Etats-Unis : "revitaliser" le HPC

Le gouvernement américain a décidé de développer son effort dans le domaine du calcul intensif pour répondre aux besoins croissants de la sécurité nationale, de la défense, des sciences fondamentales, etc. Une "Computing Revitalization Task Force" doit proposer des priorités pour le prochain budget fédéral.

http://www.itrd.gov/hecrtf-outreach/

➡ Inde : “Param Padma”

Le C-DAC (Centre for Development of Advanced Computing) a construit le Param Padma, ordinateur d’une perforpance crête de 1 Teraflops. Cette gamme est destinée à être commercialisée.

➡ NEC

La division “European Supercomputer Systems” (ESS) de NEC Allemagne est devenue une société indépendante : NEC High Performance Computing Europe (HPCE), dont le siège est à Düsseldorf (Allemagne). Cette société sera chargée de la vente et du support des clients des systèmes de haute performance NEC en Europe, en particulier les nouveaux serveurs équipés des processeurs Itanium-2 et les clusters.

http://www.hpce.nec.com

➡ Quadrics

Le cluster Linux “MCR”, installé au LLNL (Lawrence Livermore), qui était en cinquième position du TOP500, est passé en quatrième position après qu’une collaboration entre LLNL, l’Université du Texas à Austin et Quadrics ait permis d’optimiser la performance Linpack sur cette machine. La performance Linpack est passée de 5,7 à 7,63 Teraflops (pour un “peak” de 11 Teraflops, ce qui donne une efficacité de près de 70%). Un second cluster a été mis en service, dans une collaboration avec IBM et Quadrics : “ALC” a une performance Linpack de 6,6 Teraflops et entre en cinquième position dans le TOP500. Ces deux clusters utilisent des processseurs Intel Xeon à 2,4 GHz.

http://www.llnl.gov/linux/news_events.html

➡ SGI

SGI et PNNL (Pacific Northwest National Laboratory) ont lancé une collaboration visant à créer le premier ordinateur Linux à 128 processeurs avec une architecture mémoire globale partagée. Basé sur le serveur SGI Altix 3000, il devrait être livré cet été et sa performance crête serait de 768 Gigaflops.

➡ SUN

La récente annonce de la disponibilité des Sun Fire Superclusters, basés sur les plateformes Sun Fire 6800, Sun Fire 12000 et Sun Fire 15000, permet à Sun de proposer des systèmes de la classe du Teraflops (maximum : 2 TFlops).

Agenda

- 22 avril : **RAW 2003** : The 10th Reconfigurable Architectures Workshop (Nice)
- 22 avril : **HiCOMB 2003** : 2nd International Workshop on High Performance Computational Biology (Nice)
- 22 avril : **HIPS 2003** : 8th International Workshop on High-Level Parallel Programming Tools Nice)
- 22 au 26 avril : **IPDPS 2003** : International Parallel and Distributed Processing Symposium (Nice)
- 22 au 26 avril : **IWJAVAPDC 2003** : 5th International Workshop on Java for Parallel and Distributed Computing (Nice)
- 22 au 26 avril : **CAC’03** : Workshop on Communication Architectures for Clusters (Nice)
- 22 au 26 avril : **NIDISC’03** : The Sixth international Workshop on Nature Inspired Distributed Computing (Nice)
- 22 au 26 avril : **PDSECA’03** : The 4th Workshop on Parallel and Distributed Scientific and Engineering Computing with Applications (Nice)
- 26 avril : **WMPP’03** : Third Workshop on Massively Parallel Processing (Nice)
- 12 au 15 mai : **CCGRID 2003** : The third IEEE/ACM International Symposium on Cluster Computing and the Grid (Tokyo, Japon)
- 12 au 15 mai : **GP2PC 2003** : Global and Peer-to-Peer Computing on Large Scale Distributed Systems (Tokyo, Japon)
- 18 au 21 mai : **ICCSA 2003** : The 2003 International Conference on Computational Science and its Applications (Montreal, Canada)
- 19 au 22 mai : **ICDCS 2003** : The 23rd International Conference on Distributed Computing Systems (Providence, RI, Etats-Unis)
- 2 au 4 juin : **ICCS 2003** : 3rd International Conference on Computational Science (bilocalisée à Melbourne, Australie, et St. Petersbourg, Russie)
- 2 au 4 juin : Workshop on Java in Computer Science (Melbourne, Australie)
- 2 au 4 juin : Workshop Innovative Solutions for Grid Computing (Melbourne, Australie)
- 2 au 4 juin : **Terascale** : Terascale Performance Analysis Workshop (Melbourne, Australie)

- 9 au 11 juin : **ISCA 2003** : 30th Annual Symposium on Computer Architecture (San Diego, CA, Etats-Unis)
- 9 au 11 juin : **RSP 2003** : 14th IEEE International Workshop on Rapid System Prototyping (San Diego, CA, Etats-Unis)
- 10 au 14 juin : **SIGMETRICS 2003** : International Conference on Measurement and Modelling of Computer Systems (San Diego, CA, Etats-Unis)
- 11 au 13 juin : **LCTE**L'Inde**S'03** : ACM SIGPLAN Conference on Languages, Compilers, and Tools for Embedded Systems (San Diego, CA, Etats-Unis)
- 11 au 13 juin : **PPOPP'03** : The ACM SIGPLAN 2003 Symposium on Principles and Practice of Parallel Programming (San Diego, CA, Etats-Unis)
- 14 au 19 juin : **Euresco'03** : Advanced Environments and Tools for High Performance Computing (Albufeira, Portugal)
- 15 au 18 juin : **Arith 16** : 16th Symposium on Computer Arithmetic (Santagio de Compostela, Espagne)
- 16 au 17 juin : **HLPP 2003** : Second International Workshop on High Level Parallel Programming and Applications (Paris)
- 22 au 24 juin : **HPDC-12** : 12th IEEE International Symposium on High Performance Distributed Computing (Seattle, Etats-Unis)
- 22 au 24 juin : **GridForum** : Global GridForum 8 (Seattle, Etats-Unis)
- 23 au 26 juin : **ICS'03** : 17th International Conference on Supercomputing (San Francisco, Etats-Unis)
- 23 au 26 juin : **ClusterWorld** Conference & Expo (San Jose, Etats-Unis)
- 23 au 16 juin : **ERSA'03** : Engineering of Reconfigurable Systems and Algorithms (Las Vegas, Etats-Unis)
- 24 au 27 juin : **ISC'03** : The International Supercomputer Conference (Heidelberg, Allemagne)
- 27 au 30 juin : Second MIT Conference on Computational Fluid and Solid Mechanics (Cambridge,Ma, Etats-Unis)
- 25 au 29 août : Ecole d'été sur la construction d'applications réparties et les intergiciels (Autrans)
- 26 au 29 août : **EuroPar-03** : International Conference on Parallel and Distributed Computing (Klagenfurt, Autriche)
- 2 au 5 septembre : **Performance Tools** : 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation (Urbana, Il, Etats-Unis)
- 2 au 5 septembre : **ParCo 2003** : Parallel Computing 2003 Conference (Dresden, Allemagne)
- 15 au 21 novembre : **SC'2003** : Supercomputing Conference and Exhibition (Phoenix, Az, Etats-Unis)

Des informations complémentaires, en particulier les adresses http de ces manifestations, sont disponibles sur le serveur Web d'ORAP.

Appel à informations

Le contenu de BI-ORAP dépend, pour partie, de ses lecteurs ! N'hésitez pas à nous communiquer toute information concernant vos activités dans le domaine du calcul de haute performance : installations de matériel, expérimentations de nouvelles technologies, applications, organisation de manifestations, formations, etc.

Merci d'adresser ces informations au secrétariat d'ORAP ou directement à Delhaye@irisa.fr

HOISE - *Europe On-line Information Service*

PRIMEUR ! - *Advancing European Technology Frontiers*

http://www.hoise.com/primeur/

ORganisation Associative du Parallélisme
Structure de collaboration créée par le CEA, le CNRS et l'INRIA.

Secrétariat : chantal.letonqueze@irisa.fr
IRISA, campus de Beaulieu, 35042 Rennes cedex
Tél : 02.99.84.75.33, Fax : 02.99.84.74.99
http://www.irisa.fr/orap