NUMERO 66

BIORAP

BULLETIN D'INFORMATION ORAP

SOMMAIRE

Forums ORAP

Le Bureau de l'ORAP, en concertation avec la présidente du conseil scientifique, avait décidé d'annuler le forum qui devait avoir lieu le 14 octobre, cette décision étant motivée par le désistement d'un grand nombre de participants suite à l'annonce de forts mouvements de grève.

Ce forum devrait avoir lieu les 23 et 24 mars 2011, avec un programme qui reprendrait en grande partie celui qui était prévu pour le 14 octobre.

Des informations détaillées seront diffusées prochainement.

Modélisation sismologique haute performance sur un calculateur parallèle équipé de cartes GPU

Dimitri Komatitsch
Université de Pau, CNRS, INRIA et IUF

1. INTRODUCTION

La partie sismologique de mon travail de recherche a été effectuée en collaboration avec Lev P. Vinnik (Académie des Sciences de Russie, Moscou, Russie) et Sébastien Chevrot (CNRS, Observatoire Midi-Pyrénées, Toulouse). La couche géologique D" (prononcer "D seconde"), qui est la couche de la terre située au dessus de l'interface noyau-manteau et qui a une épaisseur moyenne d'approximativement 150 kilomètres, est une région complexe de la terre et fait l'objet de nombreuses études en sismologie, basées sur la propagation des ondes sismiques, qui sont notre moyen principal d'étudier la structure profonde de la terre. Elle est non seulement latéralement hétérogène mais également anisotrope, au moins dans certaines régions (Panning & Romanowicz 2004). Dans les données sismiques enregistrées autour du globe à la suite de gros tremblements de terre, la séparation (le "splitting") des ondes SV et SH dans les données est classiquement interprétée par les sismologues comme un signe de la présence d'anisotropie dans la couche D". Mais ici (voir aussi Komatitsch et al. (2010c)), en utilisant des modèles numériques isotropes de la Terre mais possédant un manteau inférieur hétérogène, nous examinons les propriétés relatives des ondes de cisaillement diffractées SVdiff et SHdiff et nous montrons clairement pour la première fois qu'elles peuvent être prises par erreur pour une indication de la présence d'anisotropie dans cette couche de la Terre.

Comme nous allons manipuler des maillages de grande taille et simuler un grand nombre de pas de temps pour chaque modèle géophysique étudié, nous allons recourir au calcul hybride utilisant de nombreuses cartes graphiques GPU (Graphic Processing Unit) en parallèle à l'aide de passages de messages MPI (Message Passing Interface) non bloquants. En effet, au cours de ces dernières années, les processeurs graphiques (GPUs) ont rapidement pris de l'importance en tant qu'architecture viable pour effectuer des calculs scientifiques. Cette partie de mes recherches a été effectuée en collaboration avec Gordon Erlebacher (Department of Scientific Computing, Florida State University, USA), Dominik Göddeke (Institut für Angewandte Mathematik, TU Dortmund, Germany), et David Michéa (INRIA Magique3D, Pau Bordeaux Sud-Ouest, France).

2. METHODE DES ELEMENTS SPECTRAUX SUR UN CLUSTER DE CARTES GRAPHIQUES GPU

Nous utilisons la méthode des éléments spectraux (Spectral Element Method - SEM) pour simuler numériquement la propagation des ondes sismiques résultant de tremblements de terre ou d'expériences d'acquisition sismique active dans l'industrie pétrolière (Tromp et al. (2008)). Au cours de la dernière décennie, en collaboration avec plusieurs collègues, j'ai été à l'origine du développement de SPECFEM3D, un logiciel qui effectue la simulation numérique tridimensionnelle de la propagation des ondes sismiques en utilisant cette méthode. Nous allons devoir modifier significativement notre logiciel SPECFEM3D pour le faire fonctionner avec une accélération très élevée sur un cluster de GPUs (Komatitsch et al. (2009, 2010a,b); Michéa & Komatitsch (2010)).

Les deux problèmes clés à résoudre sont 1) la minimisation des parties séquentielles du code pour éviter les effets de la loi d'Amdahl et 2) le recouvrement des communications MPI par du calcul. Nous employons NVIDIA CUDA pour notre implémentation. L'implémentation comprend trois étapes implémentées sous la forme de kernels CUDA séparés. La première étape met à jour le vecteur déplacement global en se basant sur sa valeur au pas de temps précédent, la deuxième étape effectue le calcul des forces élastiques et l'assemblage de ces forces au sein du maillage d'éléments finis, et la dernière étape calcule le vecteur d'accélération global. Dans la deuxième étape, un coloriage de maillage est nécessaire afin d'extraire un maximum de parallélisme (figure 1).

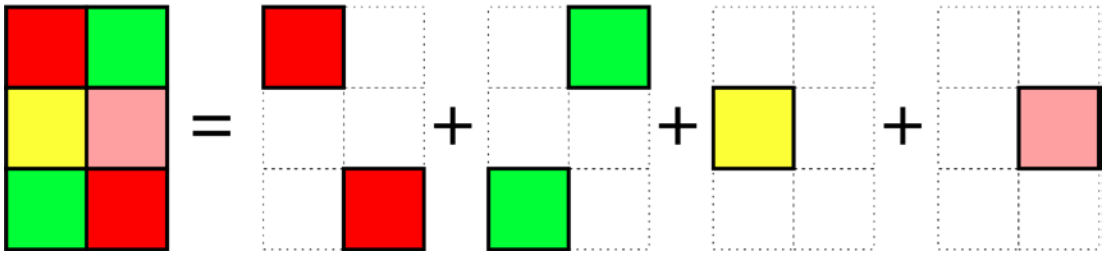

Figure 1. Coloriage de maillage : un maillage d'éléments connectés peut toujours être décomposé en sous-ensembles d'éléments disjoints

3. RESULTATS SUR UN CLUSTER DE GPUS

La machine que nous utilisons est le cluster 'Titane' de 48 Teslas S1070 du CCRT/CEA/GENCI à Bruyères-le-Châtel ; chaque Tesla S1070 possède quatre GPUs GT200 et deux bus PCI Express-2 (donc deux GPUs partagent un bus PCI Express-2). Pour les tests de scalabilité, nous utilisons des tranches contenant chacune 446,080 éléments spectraux. Chaque tranche contient approximativement 29.6 millions de points de grille uniques, c'est-à-dire 88.8 millions de degrés de liberté, ce qui correspond à 3.6 GB de mémoire utilisés sur chaque GPU (sur les 4 GB disponibles). La plus grande taille de problème possible, en utilisant chacun des 192 GPUs du cluster, est donc de 17 milliards d'inconnues.

La Figure 2 montre le temps moyen écoulé par pas de temps de l'algorithme SEM pour des simulations sur 4 à 192 GPUs (c'est-à-dire sur l'ensemble de la machine), par pas de quatre GPUs. Le weak scaling est quasiment parfait. Les petites fluctuations que nous observons sont sur l'ordre de 2 – 3 %. Nous répétons cette expérience en utilisant seulement un GPU par nœud, et par conséquent, nous pouvons seulement aller jusqu'à 96 GPUs en maintenant la charge par GPU constante. Les fluctuations sont maintenant entièrement supprimées, ce qui prouve que toutes ces fluctuations sont provoquées par le partage du bus PCIe dans chaque demi-Tesla S1070. Cependant, la durée totale du calcul est en moyenne seulement 3% plus rapide quand les bus PCIe ne sont pas partagés, ce qui signifie que le partage du bus PCIe impliqué par la structure des Tesla S1070 n'est pas un goulot d'étranglement important. Ceci démontre de façon claire que le recouvrement des communications MPI non bloquantes et des transferts PCIe par des calculs sur les GPUs est excellent dans notre application.

Pour mesurer le facteur d'accélération (speedup), nous répétons l'expérience de weak scaling avec deux configurations différentes sur les CPUs. Dans la première, nous assignons chaque tranche de 3.6 gigaoctets à un cœur de CPU, et nous assignons deux tranches de maillage à chaque CPU. Dans la seconde, nous coupons chaque tranche en deux moitiés et assignons quatre de ces plus petites tranches aux quatre cœurs de chaque CPU. Cette seconde série d'expériences nécessite seulement la moitié des nœuds de calcul de la première car une plus grande quantité de mémoire est

disponible pour chaque CPU complet que pour chaque carte GPU.

La Figure 2 (à droite) montre les mesures de weak scaling que nous obtenons. Les fluctuations sont plus grandes dans le cas des CPUs (parce que le temps écoulé est plus long), mais la quantité relative de bruit dans les mesures est la même que dans le cas de calculs purement GPU. La configuration qui emploie quatre cœurs de CPU par nœud pour calculer quatre tranches de taille moitié est 1.6 fois plus rapide que celle utilisant seulement deux cœurs de CPU avec des tranches de maillage complètes. Nous n'observons pas un facteur d'accélération idéal de 2 en raison du partage des ressources entre les cœurs. Quand nous combinons les mesures, nous pouvons en déduire un speedup moyen de 12.9 d'un GPU d'un Tesla S1070 par rapport à quatre cœurs d'un CPU Nehalem pour notre application, et d'un facteur d'environ 20.6 par rapport au cas où l'on utilise seulement deux cœurs de chaque CPU Nehalem.

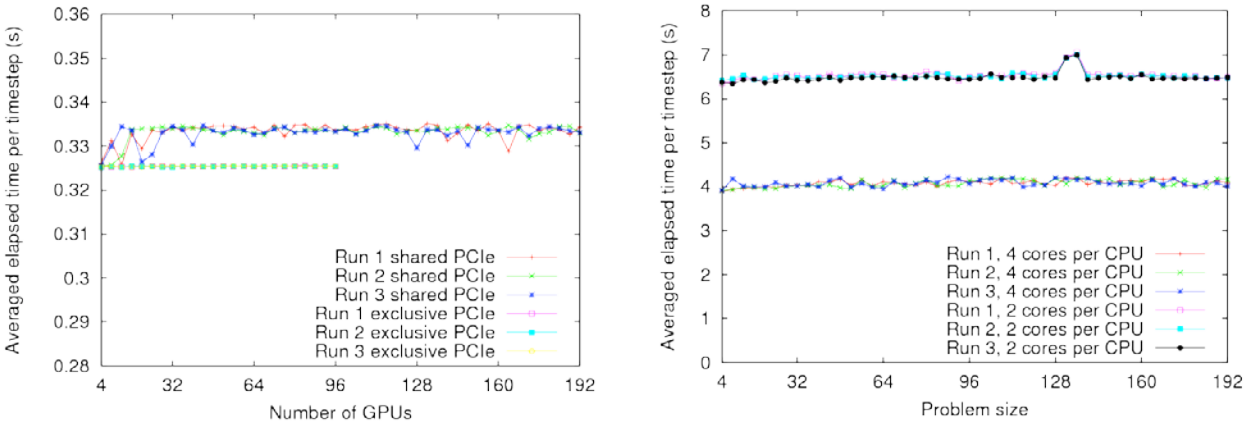

Figure 2. (A gauche) Weak scalability mesurée en utilisant les GPUs, avec ou sans partage du bus PCIe. (A droite) Weak scalability obtenue en utilisant seulement des cœurs de CPU.

4. MODELISATION DES ONDES SHDIFF ET SVDIFF A TRES HAUTE RESOLUTION SUR UN TRES GROS CLUSTER DE CPUS

Les plus gros calculs que nous pouvons effectuer actuellement pour des études paramétriques (c'est-à-dire en comparant les résultats obtenus pour différents modèles) peuvent atteindre une période sismique minimum de typiquement 5 secondes (Komatitsch et al. (2008)), c'est-à-dire une fréquence d'un cinquième de Hertz, pour l'ensemble du globe terrestre élastique 3D.

Dans cette étude, nous souhaitons que nos simulations soient précises au moins jusqu'à la même période (5 secondes). Nous maillons donc l'ensemble de la terre en utilisant 639.995.904 éléments spectraux hexaédriques et nous employons des fonctions de base polynomiales de degré 4 pour discrétiser le champ d'ondes à l'intérieur de chaque élément spectral. Ceci correspond à approximativement 42.2 milliards de points de grille dans tout le maillage et 115.1 milliards de degrés de liberté devant être calculés à chaque itération de la boucle de temps de l'algorithme SEM.

Nous effectuons les calculs en parallèle sur 6144 cœurs de processeurs du supercalculateur 'Jade' de CINES/GENCI à Montpellier en divisant le maillage en 6144 tranches ayant le même nombre d'éléments. La période dominante du spectre d'amplitude de nos sismogrammes synthétiques des composantes du vecteur vitesse est $\pi\sqrt{6.5} = 8$ s, et elles ont une énergie significative jusqu'à une période minimum d'environ 5 secondes.

La Figure 3a montre les sismogrammes synthétiques pour le modèle standard en utilisant les amplitudes vraies, et la Figure 3b montre les mêmes enregistrements mais avec des amplitudes normalisées. La décroissance d'amplitude de l'onde SVdiff en fonction de la distance épicentrale est grande par rapport à celle de SHdiff, et à une distance de 105° l'amplitude de SHdiff devient plusieurs fois plus grande que celle de SVdiff (Figure 3a). Cependant, avec des amplitudes normalisées, on observe clairement SVdiff jusqu'à une distance de 120° (Figure 3b). Les sismogrammes normalisés de la Figure 3b montrent un retard de SVdiff par rapport à SHdiff qui augmente avec la distance de 0.0 s à une distance de 94° jusqu'à 2.4 s à une distance de 120°. La différence relative de lenteur est ~0.1 s/°.

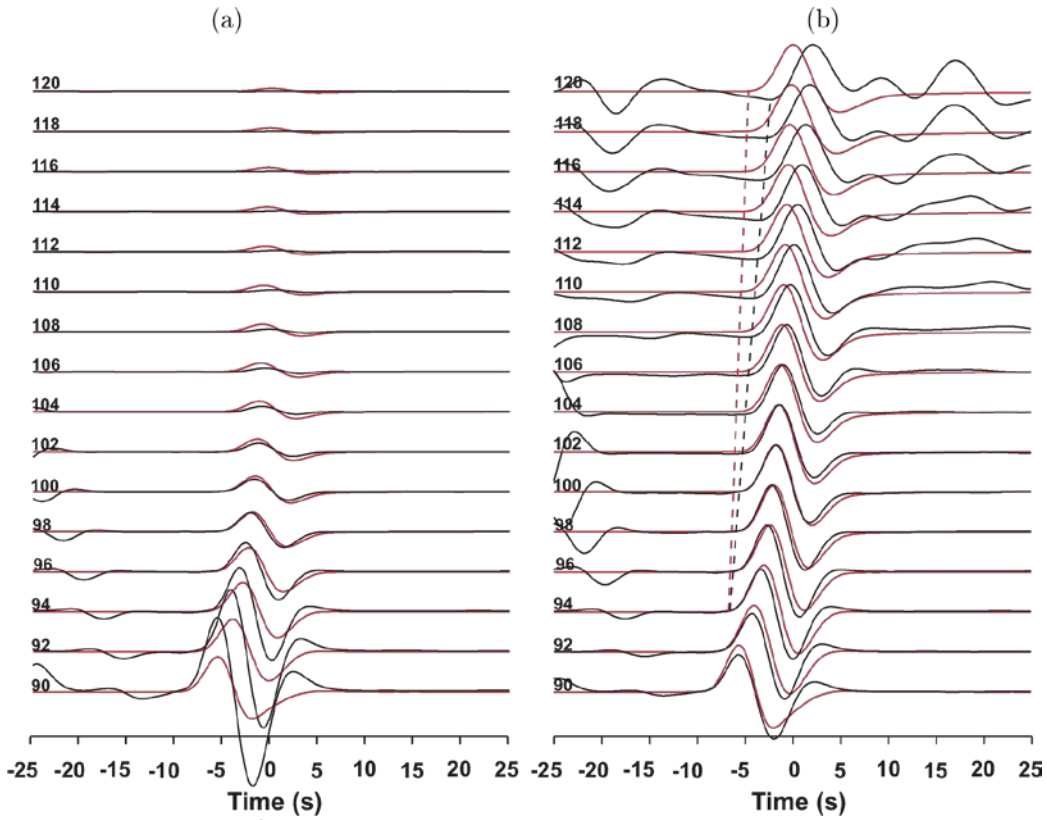

Figure 3. Sismogrammes synthétiques des composantes SH (rouge) and SV (noir) du vecteur vitesse le long de l'équateur de la Terre pour le modèle géologique IASP91 : avec les amplitudes vraies (a) et des amplitudes normalisées individuellement (b). Les nombres sur le côté gauche représentent la distance épicentrale en degrés. Les temps d'arrivée sont montrés avec une lenteur de réduction de 8.3 s/°. Les lignes pointillées rouge et noire indiquent les arrivées de SHdiff et de SVdiff, respectivement. Elles représentent le décalage (le "splitting") des ondes sismiques de cisaillement SV et SH que nous observons clairement dans nos calculs pour un tel modèle isotrope, alors que la communauté sismologique interprétait jusqu'alors de tels décalages comme un signe de la présence d'anisotropie significative dans le modèle. Ici nous montrons grâce au calcul parallèle haute performance qu'un tel phénomène est également possible dans un modèle de terre isotrope.

REMERCIEMENTS

Les calculs ont été effectués sur le cluster de GPUs « Titane » BULL Novascale R422 du CCRT/CEA/GENCI à Bruyères-le-Châtel, avec le soutien de Christine Ménaché, Edouard Audit, Jean-Noël Richet, Gilles Wiber, Julien Derouillat, Laurent Nguyen et Pierre Bonneau ainsi que Stéphane Requena de GENCI, et également sur la machine « Jade » SGI du CINES/GENCI de Montpellier, avec le soutien d'Eric Boyer, Francis Daumas et Gérard Gil. Je remercie David Michéa et Jesus Labarta pour leur aide pour optimiser notre code parallèle classique d'éléments spectraux.

REFERENCES

Komatitsch, D., Labarta, J., & Michéa, D., 2008. A simulation of seismic wave propagation at high resolution in the inner core of the Earth on 2166 processors of MareNostrum, Lecture Notes in Computer Science, 5336, 364–377.

Komatitsch, D., Michéa, D., & Erlebacher, G., 2009. Porting a high-order finite-element earthquake modeling application to NVIDIA graphics cards using CUDA, Journal of Parallel and Distributed Computing, 69(5), 451–460.

Komatitsch, D., Erlebacher, G., Göddeke, D., & Michéa, D., 2010a. High-order finite-element seismic wave propagation modeling with MPI on a large GPU cluster, J. Comput. Phys., 229(20), 7692–7714.

Komatitsch, D., Göddeke, D., Erlebacher, G., & Michéa, D., 2010b. Modeling the propagation of elastic waves using spectral elements on a cluster of 192 GPUs, Computer Science Research and Development, 25(1-2), 75–82.

Komatitsch, D., Vinnik, L. P., & Chevrot, S., 2010c. SHdiff/SVdiff splitting in an isotropic Earth, J. Geophys. Res., 115(B7), B07312.

Michéa, D. & Komatitsch, D., 2010. Accelerating a 3D finite-difference wave propagation code using GPU graphics cards, Geophys. J. Int., 182(1), 389–402.

Panning, M. & Romanowicz, B., 2004. Inferences on flow at the base of Earth's mantle based on seismic anisotropy, Science, 303, 351–353.

Tromp, J., Komatitsch, D., & Liu, Q., 2008. Spectral-element and adjoint methods in seismology, Communications in Computational Physics, 3(1), 1–32.

NDLR : ces travaux ont été récompensés par le premier prix du « Prix Bull Joseph Fourier 2010 ».

CINES 2010 : Jade double sa puissance de calcul

Les chercheurs lancent leurs grands défis scientifiques

267 Teraflop/s, 23040 cœurs, 700 Teraoctets en ligne - voilà résumée la carte d'identité de la nouvelle machine Jade hébergée au CINES pour le compte du GENCI au début de l'année 2010, caractéristiques qui lui ont valu d'être classée 18ème machine mondiale, 3ème européenne et 1ère française, au TOP500 du mois de juin 2010.

De début avril à fin août, une campagne de grands défis a permis à quelques projets d'intérêt scientifique majeur d'accéder à la machine dans des conditions privilégiées tout en validant, au plan technique et au plan contractuel, la capacité de la configuration à soutenir dans la durée une charge extrême. La fin de cette campagne a été marquée par la tenue le 1er octobre d'une journée scientifique où chaque équipe grand défi a pu exposer les bénéfices qu'elle a retiré de la nouvelle configuration étendue et les avancées scientifiques réalisées ou promises.

Le premier octobre dernier la communauté des utilisateurs du CINES s'est retrouvée pour achever un cycle de grands défis initiés grâce à la mise en production en février de l'extension de JADE, financée par GENCI. Ces défis, qui avaient débuté en avril, ont été couronnés par une journée scientifique du CINES. Originalité de cette session, l'ouverture des grands défis à des partenaires européens de PRACE (Allemands et Italiens) et à la recherche privée (VEOLIA).

Pour mémoire, la première tranche de cette machine avait été installée en 2008. Suite à cette installation, avait eu lieu une première série de grands défis, suivie du passage en production. Le même protocole a été mis en place pour la seconde tranche : livraison et installation début 2010, passage des grands défis sur l'extension et enfin mise à disposition de l'ensemble de la configuration Jade à la communauté au premier juillet. Avec cette seconde tranche, la machine offre toute la puissance nécessaire pour répondre temporairement aux besoins des chercheurs et préparer certains codes au passage à l'échelle supérieure.

La première tranche qui dispose de 12288 coeurs en technologie Harpertown, est à présent complétée par une seconde partie de 10752 cœurs en technologie Nehalem, soit un nombre total de 23040 cœurs. L'ensemble est constitué de 2880 nœuds, disposant chacun de deux processeurs quadri-cœurs, et offre plus de 91 To de mémoire utile. L'ajout de la seconde tranche ne se limite donc pas à une augmentation du nombre de processeurs. La machine a également subi un profond remaniement. Ces changements ont notamment porté sur les processeurs (leur type, leur vitesse), la quantité de mémoire par cœur, mais aussi sur la partie réseau d'interconnexion. Rappelons que cette machine dispose d'une topologie réseau tout à fait originale. Au lieu d'être traditionnellement comme dans les grands clusters conventionnels, en forme d'arbre (fat-tree), la machine Jade profite d'une architecture innovante sous la forme d'un hypercube de dimension 9. Cette topologie particulière apporte des avantages significatifs dans le domaine de la gestion des flux de communications (moins de points de concentration, donc moins de maillons faibles). Par ailleurs, le calcul ne peut aujourd'hui se concevoir sans espace de stockage adéquat. Cette partie a donc elle aussi, été redimensionnée: encore plus d'espace, et évidemment, encore plus de débit pour les écritures et les lectures sur disques.

Avec l'ouverture en production de cette extension, le CINES est maintenant capable de produire plus de 200 millions d'heures de calcul par an. Cela paraît énorme, mais la charge de la machine indique que les besoins ne cessent d'augmenter d'un mois sur l'autre.

Au cours de la journée scientifique du 1er octobre, les résultats des 13 grands défis ont été présentés. Cela a donné lieu à des échanges nombreux, parfois animés mais toujours courtois et intéressants. Les chercheurs qui ont eu la chance de participer à ces grands défis ont produit des articles qui seront prochainement diffusés dans un numéro spécial de la "Gazette du CINES". Durant ces présentations, les intervenants nous ont démontré la raison de leur intérêt pour l'accès réservé à cette machine mais aussi les résultats novateurs qu'ils ont su en tirer. Pour certains, il s'agissait d'avoir accès à une très grande quantité de processeurs, pour d'autres, c'était d'avantage l'utilisation d'un réseau alliant une grosse bande passante à une architecture totalement différente. Une majorité de ces chercheurs ont d'ailleurs « pris goût » à cette machine et ils ont, en grande majorité, déposé leur candidature pour avoir des heures attribuées via le DARI.

Le but poursuivi par ces grands défis était double : d'une part permettre à des chercheurs d'accéder à des ressources hors normes, mais aussi placer progressivement la machine dans son état de production. Ces programmes grands défis ont ainsi permis de détecter des problèmes qu'il aurait été difficile de déceler

dans le cadre d'une production normale. Tout le monde s'en est donc trouvé gagnant.

La journée du 1er octobre s'est déroulée en deux temps, d'abord les présentations des résultats, suivies d'un débat en fin d'après-midi. Les présentations ont couvert l'ensemble de l'échelle des grandeurs physiques : une plongée vertigineuse vers l'infiniment petit, avec l'exploration du monde quantique, puis vers l'infiniment grand avec des études astrophysiques. Le monde à l'échelle humaine était aussi largement présent, avec les domaines de la biologie moléculaires, l'étude des cellules, leur comportement vis à vis d'un courant électrique destiné à faciliter l'absorption de produit pharmaceutique. Une place avait également été faite à des sujets d'apparence plus terre à terre, mais d'un intérêt et d'une utilité indiscutables au plan pratique, comme l'étude de la mécanique des fluides dans des stations d'épuration. Après la présentation des résultats, la journée s'est conclue par un débat entre les utilisateurs et les personnels du CINES. Ce fut une occasion unique de confronter les attentes des utilisateurs avec les points de vue des personnels en charge de la gestion au quotidien de l'environnement de calcul. Le but pour tous est d'arriver à un taux de satisfaction maximum.

Depuis le premier juillet, l'extension a été banalisée en production dans la configuration de JADE. A cette occasion, des « benchmarks », dont le fameux Linpack, ont été exécutés sur la totalité de la machine, permettant à celle-ci d'apparaître dans le classement mondial des machines de calcul. Le TOP 500 qui classe les 500 plus grosses machines de calcul publiques la situe au 18ème rang mondial, 3ème rang européen et 1er rang français à la date de juin 2010.

Le classement n'est rien sans une bonne efficacité ainsi qu'un taux élevé de disponibilité de la machine. Le but premier n'est pas en soi d'être placé au plus haut dans le classement, mais bien d'offrir un environnement de qualité en étant toujours à l'écoute des besoins des chercheurs. De ce point de vue la machine Jade affiche un excellent « rendement ». Elle dépasse très régulièrement le taux de 92% d'activité. Sa file d'attente serait capable de remplir plus de dix fois la capacité de la machine. Nous pouvons aussi mesurer le succès que rencontre la machine au taux de remplissage de son espace de stockage. Celui-ci atteint plus de 80% et ne cesse de croître. Bref, la nature a horreur du vide, et les chercheurs aussi.

Le CINES veille depuis toujours à placer la qualité du service rendu dans ses priorités. Quand les ressources sont présentes, ce qui est aujourd'hui le cas, les projets peuvent progresser à la juste mesure de leurs besoins. Le CINES est, et restera, présent pour accompagner toute la communauté, et il assiste confiant à l'augmentation régulière de la charge.

Jean-Christophe Penalva

Nouvelles de PRACE

Dans le cadre du premier appel à propositions régulier de PRACE, neuf projets de recherche (deux allemands, deux espagnols, un français, un hongrois, un néerlandais, un portugais et un britannique) se sont vus attribuer au total 362,8 millions d'heures de calcul sur la machine JUGENE, IBM BlueGene/P, hébergée par Jülich et qui est le premier supercalculateur pétaflopique de l'infrastructure de recherche PRACE. Le projet français est conduit par Jérémie Bec, de l'Observatoire de la Côte d'Azur, le titre de ce projet étant « *Droplet growth by coalescence in turbulent clouds : kinetics, fluctuations, and universality* ». Il s'est vu attribuer 50 millions d'heures de calcul (cœur).

Le deuxième appel à propositions[1] concerne le supercalculateur JUGENE et le cluster Bullx CURIE hébergé par le CEA (et financé par GENCI) à Bruyères-Le-Châtel (France). La date limite de soumission des propositions est fixée au 11 janvier 2011.
Informations détaillées sur http://www.genci.fr/

Europe : étude IDC

IDC a publié le rapport final sur l'étude qui lui avait été commandée par la Commission européenne (voir le numéro 63 de Bi-ORAP) qui a besoin de recommandations pour définir, pour la première fois, sa stratégie dans le domaine du calcul de haute performance. Les domaines concernés incluent le temps et le climat, les énergies durables et propres, la conception dans l'aéronautique et l'automobile, les sciences de la vie, la physique des particules, le « *cloud computing* », les nouveaux matériaux, etc. Ce rapport[2] est intitulé « *A Strategic Agenda for European Leadership in Supercomputing : HPC 2020* ».

IDC constate que l'investissement de l'Europe dans le HPC, par rapport au reste du monde, a

[1] Cet appel a été diffusé dès novembre sur la liste électronique Orap
[2] http://www.hpcuserforum.com/EU/

diminué de 34% à 25%. Ce déclin est inquiétant car des études précédentes de IDC montrent une forte corrélation entre le HPC et la compétitivité scientifique et industrielle. Sur la base du PIB, l'Europe n'investit que 55 cents pour un euro investi aux Etats-Unis.

Le rapport recommande que l'Europe lance ces actions prioritaires:

- « *Expand the number, size, and access to HPC resources across the UE* ».
- « *Create a set of HPC exascale development lab/testbed centers* ».
- « *Attract more students into scientific, engineering and HPC fields, and to attract more experts from around the world to join EU scientific collaborations* ».
- « *Invest in developing next-generation exascale software* ».
- « *Target a few strategic application areas for global leadership* ».

PRACE est sans aucun doute un exemple d'action qui va dans ce sens !

Conférence Supercomputing 2010

L'édition 2010 de la conférence Supercomputing (*conference on high performance computing, networking, storage and analysis*), SC'10, a eu lieu à la Nouvelle Orléans du 13 au19 novembre. Les deux premiers jours étaient consacrés à des tutoriels et à des workshops.

Comme chaque année, les records de participation ont été battus (plus 10% d'inscrits au « Technical Program » : 51 présentations techniques, sélectionnées parmi 253 papiers soumis. Sans oublier les « Panel discussions », « Birds of a Feather », posters, etc.)

L'exposition a également battu des records : plus de 320 exposants sur près de 12.000 m^2.

Les défis posés par l'*extreme computing* sont bien évidemment l'objet de débats : calcul hybride, tolérance aux pannes, consommation énergétique. Les problèmes posés par la « programmabilité » des applications sur les architectures hétérogènes sont abordés par de nombreuses personnes. Le développement des environnements de programmation PLASMA (Parallel Linear Algebra for Scalable Multi-core Architectures) et MAGMA (Matrix Algebra on GPU and Multicore Architecture), piloté par l'université du Kentucky, en collaboration avec d'autres centres de recherche dont l'INRIA, tente de répondre à une partie de ces défis.

Expo recherche

Comme toujours, les stands des universités (Illinois, Californie, Texas, …) et des laboratoires nationaux américains (Argonne, Los Alamos, Nasa, NOAA, Sandia, …) dominent.

L'Europe était présente, avec en particulier l'Allemagne (HLRS, le Fraunhofer, le centre de calcul de Juelich, …), le centre de calcul de Barcelone, l'université d'Edimbourg, l'INFN italien, la recherche des Pays-Bas, l'université de Vienne, etc. PRACE disposait également d'un stand.

La France était représentée par le CEA qui mettait en avant le succès de Tera100, et par l'INRIA qui proposait 12 démonstrations scientifiques s'articulant autour de deux grands axes : la recherche d'outils et d'architectures au service du HPC et les recherches axées sur des domaines applicatifs du HPC.

Le stand CEA. Photo : jld, INRIA

Le stand INRIA. Photo : jld, INRIA

Du côté des constructeurs

Voici les principales informations (hors informations confidentielles) résultant de réunions ou d'entretiens directs que j'ai pu avoir avec les principaux constructeurs engagés dans le HPC. Je remercie vivement ces constructeurs et je précise que ces résumés ne les engagent en aucune façon.

On verra que Sun n'apparaît pas dans cette liste, ne disposant pas de stand dans cette exposition ; le (petit) stand d'Oracle (qui a racheté Sun en 2010) mentionnait Sun et StorageTek dans une approche globale « *Hardware and software to work together* ».

AMD

La série Opteron 6000 destinée aux serveurs comprend le processeur « Magny-cours » (8 ou 12 cœurs) et le processeur Interlagos (12 ou 16 cœurs). Notons que des processeurs Opteron équipent 6 des systèmes du nouveau TOP10.

Une nouvelle gamme de processeurs basée sur la technologie baptisée « Bulldozer », ayant une finesse de gravure de 32 nm, va voir le jour. Cette technologie va permettre de construire des processeurs à partir de « modules ». Bulldozer est composé de 1 à 8 modules, chaque module contenant deux cœurs dédiés aux entiers, et un seul cœur dédié aux flottants (AMD compte donc sur ses GPU pour la performance en virgule flottante).

La stratégie d'AMD dans le calcul hybride s'appelle « Fusion » : il s'agit de prendre le meilleur d'un CPU et le meilleur d'un GPU pour en faire la synthèse sur le plan intégration. Les premiers APU (Accelerated Processing Unit) « Fusion » devraient être disponibles avant l'été 2011. Les modèles destinés au HPC auraient 4 cœurs associés à 1 GPU.

Bull

Bull renforce sa position dans le TOP500, d'une part en prenant la sixième position mondiale (et la première en Europe) grâce à la machine Tera100 installée au CEA/DAM, d'autre part en enregistrant 9 systèmes dans cette liste. Bull annonce une croissance de 50% de son chiffre d'affaire dans le secteur « Extrême computing » entre 2009 et 2010 .

Soulignons que Tera100 est n°1 en terme d'efficacité (84%, rapport entre la performance Linpack et la performance théorique), devant le Cray XT5 « Jaguar » de Oak Ridge.

Cette position sera encore confortée avec l'installation, fin 2011, de la seconde phase du système « Curie » financé par GENCI et destiné au CEA (voir l'annonce dans le précédent numéro de Bi-Orap), qui devrait avoir une performance crête supérieure à 1,5 PetaFlops.

La gamme « Extreme computing » comprend :

- Les bullx R (rackable) à 1 ou 2 U
- Les bullx B (blades), avec des lames à 2 processeurs Intel ou des doubles lames intégrant des GPU
- Les bullx S (supernodes) quadri sockets, disposant de 128 coeurs

Cette gamme suit l'évolution des processeurs Intel (Westmere, Sandybridge au second semestre 2011, …) et des GPU Nvidia.

Dans le domaine des logiciels, Bull a annoncé, pendant la conférence SC10, la « bullx supercomputer suite » qui offre une solution globale permettant aux clients de déployer et d'administrer facilement leurs clusters HPC.

Cray

Cray prend la deuxième place du TOP500 en terme de performance installée. Le XE6, annoncé il y a un an, a fait un début remarquable avec plus de 5 PetaFlops installés ou commandés (NOAA, NERSC, Los Alamos, etc.).

Il sera suivi en 2012 par le système « Cascade » développé dans le cadre du programme HPCS [3] , disposant d'un nouveau système d'interconnexion baptisé « Aries » et comprenant, entre autres, des lames d'accélérateurs. Puis viendrait « Shasta » vers 2015 avec l'interconnect « Pisces ».

La consommation énergétique et la programmation des applications sont au centre des réflexions de Cray sur la route vers l'exaflops. Le développement d'un nouveau nœud, appelé « Glacier », associant 8 à 16 cœurs x86 et plusieurs centaines d'accélérateurs fait partie des réponses à cette réflexion. Sur le plan de la programmation, Cray s'oriente vers un modèle de programmation unique, intégrant des directives permettant de disposer de codes portables sur des architectures différentes multicœurs, avec ou sans accélérateur.

Enfin, le langage de programmation Chapel [4] sera disponible avec la sortie de Cascade.

IBM

La stratégie d'IBM se décline en trois niveaux :

- L'exascale, correspondant aux grands défis ; la réponse est dans des technologies innovantes telles que Blue Gene.
- Le HPC, stratégique pour la compagnie ; la réponse est dans les technologies x86 et Power.
- Le « mainstream » avec des produits standards (Power et x86).

Le Power7, annoncé début 2010, dispose d'un packaging dépendant des performances attendues. Une évolution, avec amélioration des performances, devrait être disponible en 2011. Dans le cadre du programme PERCS, IBM fournira des « super nœuds » à 1024 cœurs (et 512.000 cœurs pour le système complet). Le système Blue Waters doit être installé au NCSA

[3] Voir Bi-ORAP n° 49 et n°50

[4] http://chapel.cray.com/

en 2011, avec une performance crête de 10 PetaFlops.

Un nouveau processeur, Power8, devrait sortir en 2012.

Les systèmes Blue Gene ont vu leur performance multipliée par 1000 en 10 ans. Leur architecture (haut niveau de parallélisme, avec une fréquence limitée à 1,6 GHz, permet d'obtenir une excellente performance énergétique (le prototype BG/Q arrive en première position dans le dernier classement de Green500, liste qui évalue le rapport performance par watt consommé, loin devant les autres systèmes). Blue Gene/Q dispose de processeurs à 8 ou 16 cœurs PowerPC. La machine « Sequoia » devrait être installée au Los Alamos National Lab dans le cadre du programme ASC (Advanced simulation and computing) fin 2011, avec 1,6 million de cœurs et une performance crête de 20 PetaFlops.

HP

Le calcul de haute performance représente 10% du chiffre d'affaire « serveurs » de HP. La position de HP dans le TOP500 reflète son objectif : aller prioritairement vers les machines de niveau « Tier 2 » (mésocentres, dans le monde académique) plutôt que vers les machines « Tier 0 » ou « Tier 1 ». Sa présence est forte dans les secteurs industriels et bancaires (certains clients ne souhaitent d'ailleurs pas apparaître dans le liste Top500, pour des raisons de confidentialité et de concurrence).

Les serveurs ProLiant SL390s G7 forment le fer de lance de HP dans le HPC. Ils peuvent intégrer des GPU. Début 2011 sortira un serveur équipé de 8 GPU par carte mère.

Ceci n'empêche pas HP de créer un centre de compétence sur l'exascale, travaillant à la fois sur la performance, la consommation énergétique et le prix. Quatre axes de travail :

- Un interconnect optique
- Des nœuds multi-cœurs (vers 256 cœurs, avec Intel ou un autre fournisseur)
- Un environnement logiciel global (de l'OS aux outils de gestion)
- Le développement des applications.

NEC

NEC HPCE (High Performance Computing Systems Europe) est une division européenne indépendante de NEC Corporation dédiée au calcul de haute performance et qui fournit des équipements, des solutions et des services. NEC est à ce jour le seul constructeur à proposer des solutions scalaires et des solutions vectorielles.

Contrairement à ce qui a pu être dit ou écrit, NEC n'a pas abandonné le vectoriel (gamme SX), considérant que c'est une des réponses au défi de la croissance de la performance tout en conservant une consommation énergétique acceptable. Le successeur du SX9 actuel, surnommé « NGV : next generation vector », est en préparation et devrait être disponible en 2014. Utilisant des chips multi-cœurs, la performance d'un chip serait supérieure à 256 GigaFlops, avec une bande passante mémoire supérieure à 250 Go/s. La partie vectorielle de chaque cœur utilisera une nouvelle architecture, en particulier au niveau des mémoires caches. La partie scalaire est également améliorée, sans toutefois donner plus de détails, ainsi que l'efficacité énergétique.

La gamme LX est le nom des solutions scalaires. Le haut de gamme est le LX 4000, serveur à base de lames utilisant des processeurs Westmere ou MagnyCours et une interconnexion Infiniband et un système de gestion de fichiers parallèle LXFS (Lustre). La gamme SMP Express 5800/A1080a évolue et permet de disposer jusqu'à 8 sockets 64 cœurs Intel Nehalem EX et 2 To de mémoire à plat. et de GPU. Comme d'autres constructeurs, NEC propose également des solutions de type containers.

Afin de faciliter les architectures hybrides, NEC propose depuis peu la librairie NEC MPI LX, version scalaire de la librairie MPI SX. Deux clients européens testent actuellement des installations hybrides vectorielles, scalaires et GPU. Ils permettent ainsi de consolider la stratégie « hybride » de NEC.

Signalons enfin que NEC est également un « intégrateur », son plus beau succès récent étant la deuxième génération de la machine Tsubame (n°4 dans le TOP500), le matériel ayant été essentiellement fourni par HP, La définition de l'architecture, l'installation et la maintenance étant de la responsabilité de NEC.

SGI

Les deux gammes destinées à la haute performance sont :

- la gamme Altix UV à architecture à mémoire partagée (lames Intel Xeon, jusqu'à 2048 cœurs et 16 To de mémoire) particulièrement destinée aux applications adaptées à ce type d'architecture, fonctionnant sous Linux ou sous Windows. PSC (Pittsburg Supercomputing Center) dispose de 2 machines de ce type (2048 c/ 16 To).
- la gamme Altix ICE 8400 pouvant accueillir jusqu'à 1536 cœurs dans un seul rack (processeurs AMD 6100 ou Intel Xeon 5600), ainsi que des GPU avec le choix de la topo-

logie (hypercube, fat-tree, …) en fonction des besoins du client. La performance disponible sur un seul rack peut atteindre 15,36 TeraFlops. La configuration va de 16 à 131.072 nœuds (plus d'un million de cœurs).

Le calcul hybride prend une place importante, pour des raisons de performance et de consommation énergétique, que ce soit avec des accélérateurs « many-core » (processeur Tilera, jusqu'à 100 cœurs, bientôt 200) avec un environnement de programmation classique (C++, …) ou avec des GPU pour des codes massivement parallèles, avec un environnement de programmation plus complexe (OpenCL). SGI a annoncé, pendant cette conférence, le serveur PrismXL dédié au calcul Hybrid (nom de code « Mojo »). Il supporte tous types d'accélérateurs de calcul en simple et double précision (Nvidia, AMD/ATI, Tilera) permettant d'atteindre la performance record de 200 TeraFlops avec un seul châssis. La barrière du PetaFlops est ainsi dépassée avec seulement 5 châssis.

De plus SGI met à la disposition des utilisateurs et administrateurs deux suites logicielles : SGI Management Suite et SGI Performance Suite. Elles permettent de gérer très simplement de petits clusters de quelques centaines de cœurs jusqu'à des configurations de plusieurs dizaines de milliers de cœurs.

Jean-Loïc Delhaye

36ème TOP500

La conférence Supercomputing est le cadre de la publication de l'édition d'automne du TOP500[5], liste des 500 systèmes les plus performants installés dans le monde, performance mesurée sur le benchmark « Linpack ».

L'événement était sans aucun doute l'arrivée d'un système chinois, Tianhe-1A, à la première place de cette liste, position convoitée par tous les grands constructeurs. La Chine est également présente à la troisième place avec Nebulae, et totalise 41 systèmes dans cette liste, arrivant en deuxième position derrière les Etats-Unis.

Quelques faits marquants :

- Sept systèmes ont une performance Linpack supérieure au PetaFlops.
- IBM continue de dominer le TOP500 (en nombre et en performance), Cray étant maintenant n°2 pour la performance cumulée.
- HP (137) et IBM (136) ont placé 273 des 281 systèmes installés dans l'industrie ou la finance.
- 398 des 500 systèmes utilisent des processeurs Intel, 57 des processeurs AMD.
- 17 systèmes utilisent des accélérateurs Nvidia, Cell, ATI)
- Des processeurs quadri-cœurs sont utilisés dans 73% des systèmes, 19% utilisent des processeurs ayant au moins 6 cœurs.

La position relative des principaux constructeurs, en nombre de systèmes et en performance Linpack, est maintenant la suivante :

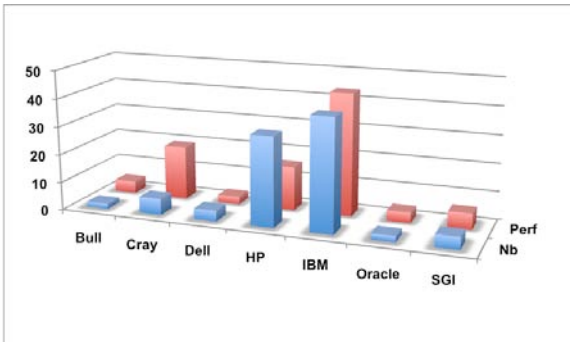

Figure : position relative des constructeurs

La position relative des pays ayant plus de 15 systèmes dans le TOP500 est la suivante :

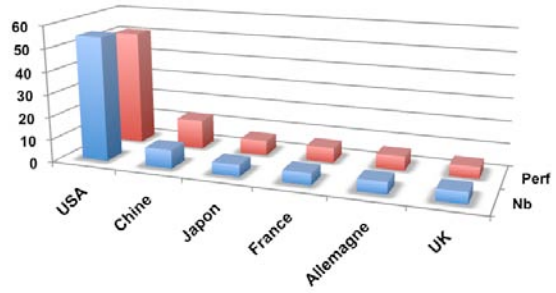

Figure : position relative des principaux pays

TOP10

Voici la liste des 10 systèmes les plus performants installés dans le monde (Rmax en TeraFlops), d'après la liste TOP500.

Système	Localisation	Rmax
Tianhe-1A : NUDT TH MPP, X5670, GPU Nvidia	Centre de calcul de Tianjin (Chine)	2566
Jaguar : Cray XT5-HE, Opteron	ORNL, USA	1759
Nebulae : Dawning TC3600 Blade	Centre de calcul de Shenzhen (Chine)	1271
Tsubame : NEC/HP	Titech, Tokyo	1192
Hopper : Cray XE6	NERSC, USA	1054
Tera100, Bull	CEA-DAM, France	1050

[5] http://www.top500.org

Roadrunner : IBM BladeCenter	NNSA/LANL, USA	1042
Kraken : Cray XT5	NICS, USA	831
Jugene : IBM BG/P	FZJ, Allemagne	826
Cielo : Cray XE6	NNSA/LANL, USA	817

L'efficacité (rapport entre performance Linpack et performance théorique) est un élément intéressant. Sept des systèmes ont une efficacité supérieure à 75%, la première place revenant à Tera100 (Bull) avec une efficacité de 84%. Les machines chinoises et japonaises ont une efficacité inférieure à 55% ; l'utilisation massive de GPU est sans doute une explication.

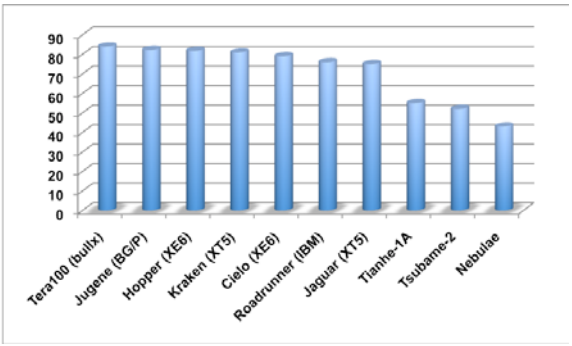

Figure : performance des systèmes du TOP10

France

La France arrive donc en quatrième position, derrière le Japon et devant l'Allemagne (même nombre de systèmes dans chacun de ces pays, mais performance inférieure).

Voici les dix systèmes les plus puissants :

N°	Système	Localisation	Rmax
6	Bull bullx supernode	CEA-DAM	1050
27	SGI Altix ICE 8200EX	CINES/GENCI, Montpellier	238
36	HP Cluster Platform 3000	Gouvernement	180
37	IBM iDataPlex	EDF R&D	169
55	IBM Blue Gene/P	CNRS IDRIS Orsay	119
61	Bull Novascale R422-E2	CEA CCRT	108
65	SGI Altic ICE 8200EX	TOTAL	106
76	IBM Blue Gene/P	EDF R&D	95
83	HP Cluster Platform 3000	Industrie	89
85	Bull bullx supernode	Bull	87

Jean-Loïc Delhaye

NOUVELLES BREVES

➔ Exascale Technology and Computing Institute

Le DoE américain a créé un institut dédié à la résolution des problèmes liés aux systèmes exaflopiques. L'ETCi travaillera avec des chercheurs et des industriels du monde entier ; il est dirigé par Pete Beckman qu'ORAP a eu le plaisir d'accueillir dans ses forums.

➔ CAPS entreprise

- Après l'ouverture d'un bureau à Shanghai en avril dernier, CAPS poursuit son développement à l'international avec la création d'une filiale américaine basée à Santa Clara (CA).
- La version 2.4 de HMPP a été officiellement présentée à SC'10. Rappelons que HMPP est une solution de programmation basée sur des directives, qui permet de simplifier l'utilisation des accélérateurs matériels mais aussi de conserver la portabilité des codes.

➔ Cray

- L'université de Stuttgart a commandé un Cray XE6 et un système de la génération « Cascade » destinés au HLRS. Le XE6 sera opérationnel en 2011, le système Cascade le sera en 2013. Le montant du contrat serait supérieur à 45 millions d'euros.
- Le centre brésilien de prévision météorologique et de recherche sur le climat a passé commande d'un Cray XT6 d'une performance crête de 244 TeraFlops.
- Cray a annoncé, pendant SC10, le XE6m, équipé de processeurs AMD Opteron et de l'interconnect Gemini. La compagnie vise à étendre son marché vers des clients n'ayant pas besoin du très haut de gamme, avec un prix démarrant à 500.000 dollars.

➔ NEC

L'ICM, centre international de recherche sur le cerveau, se dote d'une solution de stockage innovante constituée du logiciel Active Circle et de matériel NEC pour gérer le stockage de plusieurs pétaoctets de données scientifiques. Ces données, essentiellement des données de mesures, des images et des résultats d'analyse, bénéficient d'un stockage virtualisé, auto-sécurisé et extensible.

➔ SGI

- Le HLRN, en Allemagne, vient de mettre en production plusieurs systèmes Altix ICE et Altix UV, sur deux sites : Berlin et Hanovre. Chaque site dispose de 1280 lames ICE et 152 lames UV avec 400 To de stockage sous Lustre.
- Biogemma et IBCP (Institut de biologie et chimie des protéines, à Lyon) ont acquis des Altix UV récemment pour la partie « mémoire partagée » sur des applications en biotechnologies.
- La configuration du système Pleiades de la NASA a été augmentée de 8 racks ce qui porte sa performance Linpack à 840 TeraFlops.

AGENDA

24 au 26 janvier 2011 - **HiPEAC'11** : 6th International Conference on High-Performance and Embedded Architectures and Compilers (Heraklion, Crete, Grèce)

9 au 11 février 2011 - **PDP 2011** : The 19th Euromicro International Conference on Parallel, Distributed and Network-Based Computing (Ayia Napa, Chypre)

9 au 11 février 2011 - **MSOP2P 2011** : 5th International Workshop on Modeling, Simulation, and Optimization of Peer-to-peer Environments (Ayia Napa, Chypre)

12 février 2011 – **SAW-2** : 2nd workshop on SoC Architectures, Accelerators and Workloads (San Antonio, Tx, Etats-Unis)

12 au 16 février 2011 - **HPCA 17** : The 17th International Symposium on High-Performance Computer Architecture (San Antonio, TX, Etats-Unis)

12 au 16 février 2011 - **PPoPP 2011** : 16th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (San Antonio, TX, Etats-Unis)

13 février 2011 – **HipHaC** : 2nd international workshop on New Frontiers in High-performance and Hardware-aware Computing (San Antonio, Tx, Etats-Unis)

22 au 25 février 2011 - **ARCS 2011** : 24th International Conference on Architecture of Computing Systems (focus on many-cores architectures) (Lake Como, Italie)

8 au 11 mars 2011 – **LCI 2011** : 11th LCI International Conference on High-Performance Clustered Computing (Pittsburgh, Etats-Unis)

2 au 6 avril 2011 - **CGO 2011** : The Ninth ACM/IEEE International Symposium on Code Generation and Optimization (Chamonix, France)

10 au 12 avril 2011 - **ISPASS 2011** : International Symposium on Performance Analysis of Systems and Software (Austin, TX, Etats-Unis)

11 au 13 avril 2011 – **MRSC 2011** : 4th International Many-cores and Reconfigurable Supercomputing Conference (Bristol, UK)

12 au 15 avril 2011 - **PARENG 2011** : The Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering (Ajaccio, France)

12 au 15 avril 2011 - **PARENG 2011 - S02** : Special Session "High Performance Green Computing" (Ajaccio, France)

13 au 14 April 2011 - **DEISA-PRACE 2011** : DEISA PRACE Symposium 2011 (Helsinki, Finlande)

1 au 4 Mai 2011 - **NOCS 2011** : Fifth ACM/IEEE International Symposium on Networks-on-Chip (Pittsburgh, PA, Etats-unis)

3 au 5 Mai 2011 - **CF'11** : 2011 ACM International Conference on Computing Frontiers (Ischia, Italie)

10 au 12 mai 2011 – **NOTERE** : 11ème conférence internationale sur les nouvelles théories de la répartition (Paris, France)

10 au 13 Mai 2011 - **Renpar'20** : Rencontres francophones du parallélisme (Saint Malo, France)

16 au 17 Mai 2011 - **ValueTools'11** : 5th Intl. Conference on Performance Evaluation Methodologies and Tools (Paris, France)

16 Mai 2011 - **HPPAC'11** : Seventh IEEE Workshop on High-Performance, Power-Aware Computing (Anchorage, Alaska, Etats-Unis)

16 Mai 2011 - **CASS 2011** : 1st Workshop on Communication Architecture for Scalable Systems (Anchorage, Alaska, Etats-Unis)

16 au 17 Mai 2011 - **RAW 2011** : 18th Reconfigurable Architectures Workshop (Anchorage, Alaska, Etats-Unis)

16 au 20 Mai 2011 - **IPDPS 2011** : 25th IEEE International Parallel & Distributed Processing Symposium (Anchorage, Alaska, Etats-Unis)

16 au 20 Mai 2011 - **PCO'11** : Workshop : new trends in Parallel Computing and Optimization (Anchorage, Alaska, Etats-Unis)

16 au 20 Mai 2011 - **NIDISC'11** : The 14th International Workshop on Nature Inspired Distributed Computing (Anchorage, Alaska, Etats-Unis)

20 May 2011 - **LSPP'11** : Workshop on Large-Scale Parallel Processing (Anchorage, Alaska, Etats-Unis)

25 May 2011 - **SNAPI 2011** : 7th IEEE International Workshop on Storage Network Architecture and Parallel I/Os (Denver, CO, Etats-Unis)

Les sites de ces manifestations sont accessibles sur le serveur ORAP (rubrique Agenda).

Si vous souhaitez communiquer des informations sur vos activités dans le domaine du calcul de haute performance, contactez directement Jean-Loic.Delhaye@inria.fr

Les numéros de BI-ORAP sont disponibles en format pdf sur le site Web d'ORAP.

ORAP est partenaire de

Europe on-line Information Service
http://www.hoise.com/primeur

ORAP
Structure de collaboration créée par
le CEA, le CNRS et l'INRIA

Secrétariat : Chantal Le Tonquèze
INRIA, campus de Beaulieu, 35042 Rennes
Tél : 02 99 84 75 33, fax : 02 99 84 74 99
chantal.le_tonqueze@inria.fr
http://www.irisa.fr/orap