NUMERO 58

BIORAP

BULLETIN D'INFORMATION ORAP

SOMMAIRE

Forum ORAP

Le **24ème Forum** aura lieu à Lille le 26 mars 2009. Il aura pour thème principal « *Les données et le calcul de haute performance* ».

Les informations détaillées seront disponibles sur le site ORAP, ainsi que le formulaire d'inscription ; nous vous rappelons que l'inscription est gratuite mais obligatoire (ceci afin de nous permettre d'organiser cette journée dans de bonnes conditions).

Contact : chantal.letonqueze@irisa.fr

Appels à proposition de l'ANR

L'Agence Nationale de la Recherche a publié les appels à projets[1] correspondant à sa programmation 2009.

On notera en particulier le programme COSINUS[2] (le programme Conception et Simulation vise à développer la conception et la simulation numérique, en s'appuyant, le cas échéant, sur l'utilisation du calcul intensif.), présenté dans le numéro 57 de BI-ORAP, pour lequel la date limite de soumission du dossier électronique est fixée au 19 février.

Cet appel a projets COSINUS comprend trois axes thématiques :

1. **Simulation et calcul intensif**, avec deux catégories : « Grands défis et passages à l'échelle pour les applications », « Outils et modèles de programmation ».
2. **Conception et optimisation**
3. **Environnements de simulation et masses de données**, avec deux catégories : « Pré-traitement, post-traitement, visualisation et interaction avec de grands volumes de données », « Simulation et modélisation des données ».

[1] http://www.agence-nationale-recherche.fr/AAPProjetsOuverts

[2] http://www-anr-ci.cea.fr/scripts/home/publigen/content/templates/show.asp?L=FR&P=141

CINES : inauguration de « Jade »

Le nouveau superordinateur du CINES (Centre Informatique National de l'Enseignement Supérieur, situé à Montpellier), surnommé « Jade », a été inauguré le 5 novembre par Valérie Pécresse, Ministre de l'Enseignement Supérieur et de la Recherche.

Photo : CINES

Rappelons qu'il s'agit d'un système SGI Altix ICE 8200EX (processeurs Xeon quadri-cœurs) dont la puissance crête est de 147 TeraFlops (128 TeraFlops sur Linpack), ce qui en fait l'ordinateur le plus puissant installé en France actuellement (et le met en 14ème position dans le monde).

La mise en place de ce supercalculateur fait suite à un appel d'offres international réalisé par GENCI (Grand équipement national de calcul intensif), chargé de l'achat, de la coordination et de la gestion des supercalculateurs destinés à la recherche en France.

L'inauguration a été précédée d'un colloque scientifique[3] qui a montré tout l'intérêt que ce supercalculateur avait pour la recherche, certaines équipes ayant déjà largement sollicité cette machine depuis sa mise en production en été 2008.

Europe

Programmes européens

La Commission européenne a publié le programme de travail pour le programme ICT et pour la période 2009-2010. Les trois prochains appels à projets (« Calls » 4, 5 et 6) s'appuieront sur ce texte[4].

La Commission a publié, le 19 novembre 2008, une série d'appels à projets[5], dont le **4ème appel à projets pour les TIC** (avec un budget indicatif de 801 millions d'euros). Ce « Call 4 », qui sera clos le 1er avril 2009, concerne en particulier l'objectif 3.6 : « Computing Systems ». Les documents et informations utiles sont disponibles sur le site européen Cordis : http://cordis.europa.eu/fp7/dc/index.cfm?fuseaction=UserSite.FP7DetailsCallPage&call_id=185

La base de données « CORDIS Projects » contient déjà 2163 projets financés au titre du 7ème PCRDT. Une mine d'informations, particulièrement pour tous ceux qui préparent des projets.

Nouvel appel du **programme "Capacités"**. L'appel (FP7-INFRASTRUCTURES-2009-1 [6]) concerne notamment le soutien aux e-infrastructures de technologies de l'information et de la communication (TIC) ainsi que le soutien au développement de politiques et à la mise en œuvre de programmes et des études, conférences et activités de coordination soutenant le développement politique dans le cadre d'une coopération internationale pour les e-infrastructures. Clôture de l'appel le 13 mars.

Institut européen de technologie

La première réunion du Comité directeur de l'EIT - Institut européen d'innovation et de technologie - s'est tenue le 24 novembre 2008 à Bratislava. Les documents relatifs à cette réunion sont en ligne sur le site de l'EIT[7].

Echos de Supercomputing 2008

Réunie à Austin (Texas) du 15 au 21 novembre, SC'08, « *The International Conference for High Performance Computing, Networking, Storage and Analysis* » était la 20ème édition de cette manifestation mondiale qui regroupe les acteurs du domaine du calcul de haute performance. Avec près de 11.000 participants, record absolu, elle a tenu ses promesses, tant au niveau de la conférence elle-même (programme technique, workshops, posters, etc) qu'au niveau de l'exposition associant entreprises et organismes de recherche.

Conférence

Un foisonnement de présentations était proposé aux participants à la conférence. 13 workshops (« node level parallelism », « HPC Finance », « Grid Computing Environments », ...) et 25 tutoriels d'initiation ou de perfectionnement (« Introduction to OpenMP », « Debugging Parallel and Distributed Applications », …) étaient organisés avant la conférence proprement dite (dimanche et lundi). Du mardi au vendredi matin, de nombreuses sessions en parallèle ont permis la présentation de 59 « technical papers » (sur 277 soumis), de tables rondes, de 54 Birds-of-a-Feather, de 55 posters, sans oublier les orateurs invités, et diverses manifestations de moindre importance.

Expositions

Les stands des industriels et des laboratoires de recherche étaient regroupés dans les grands halls du centre de congrès. Plus de 330 stands, de toutes dimensions, présentaient les derniers produits ou résultats dans les domaines couverts par la conférence : processeurs, systèmes, architectures, réseaux d'interconnexion, systèmes de stockage, logiciels, applications, etc. Forte présence, attendue, des constructeurs et des agences gouvernementales américaines (NASA, DoE, laboratoires nationaux, …) et des universités américaines. Forte présence éga-

[3] http://www.cines.fr/spip.php?article536
[4] ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/ict-wp-2009-10_en.pdf
[5] http://cordis.europa.eu/fp7/dc/index.cfm
[6] http://www.eurosfaire.prd.fr/news/consulter.php?id=2599
[7] http://ec.europa.eu/eit/seminars_en.htm

lement des asiatiques, particulièrement des japonais (32 stands, surtout académiques). Les européens font à peu près « jeu égal » avec les asiatiques, avec près de 40 stands de 13 pays européens, l'Allemagne, la France et le Royaume-Uni ayant la plus forte représentation. La Commission européenne est présente au travers de projets européens : Egee, EGI, Prace, XtreemOS.

Pour la France, trois entreprises exposent : BULL, et deux « jeunes pousses » : CAPS (qui développe et commercialise des logiciels innovants de mise au point pour la performance des applications des domaines du HPC et de l'embarqué) et GPU-Tech (développement d'outils pour l'utilisation des cartes graphiques dans le HPC). Sans oublier la « jeune pousse » KERLABS qui présentait en avant première, et en partenariat avec l'entreprise italienne SCREEN Group, un cluster de technologie avant-gardiste.

Photo : CEA

La recherche française est représentée par le CEA et l'INRIA. Le CEA présente plusieurs de ses applications, ainsi que ces deux grands projets que sont TERA100 (qui devrait voir le jour en été 2010) et l'extension des systèmes du CCRT (2009) en collaboration avec BULL et NVIDIA. Le stand de l'INRIA regroupe huit « pôles » de démonstration et de rencontre, tenus par des chercheurs de dix équipes-projets.

Photo : INRIA

Les réseaux disponibles, tant pour les exposants que pour les participants (accès Wifi facile et performant dans l'ensemble du centre de congrès), étaient d'une qualité remarquable.

TOP500

Les 500 systèmes les plus performants installés dans le monde en novembre 2008 ! Deux systèmes au dessus du PetaFlops (IBM au LANL, Cray à ORNL). Sept des systèmes du TOP10 sont installés dans des centres relevant du Département de l'Energie américain ! Le système le plus performant installé en dehors des Etats-Unis est une machine chinoise (Dawning 5000A) installée à Shanghai, classée à la 10ème place.

Après de nombreuses années de quasi-absence, la France retrouve une excellente position ; on notera en particulier trois systèmes classés entre les places 14 et 17 : le CINES (système SGI Altix), l'IDRIS (système IBM BG/P), Total Exploitation Production (système SGI Altix).

Une présentation plus détaillée de cette édition du TOP500 sera faite dans le prochain numéro de Bi-ORAP.

Que peut-on retenir de SC'08 ?

Pas d'annonce très marquante, pas de révolution ou d'innovation forte. On retrouve les thèmes dominants de 2007 : la course à la performance passe par les processeurs multi-cœurs mais aussi par les processeurs spécialisés (GPUs, Cell, etc) et donc par des architectures « hybrides » ; la bande passante et la latence sont souvent aujourd'hui des freins à la performance réelle ; la consommation énergétique est devenue un problème majeur et il faut tout faire pour la réduire ; la gestion des grands volumes de données est un enjeu qui se confirme.

Le parallélisme est inévitable, et la majorité des ordinateurs portables aujourd'hui dispose de processeurs multi-cœurs (sans nécessairement en tirer partie, le travail à réaliser sur les systèmes d'exploitation et sur les applications est considérable). De plus, l'ajout de cartes GPU, en particulier fournies par NVIDIA, permet de disposer de « superordinateurs personnels » à condition de savoir les utiliser dans de bonnes conditions ; l'arrivée d'environnements de développement, par exemple ceux commercialisés par CAPS Entreprise[8], va faciliter l'utilisation de ces cartes. La multiplication des séminaires de travail et des conférences sur le « multi-cœur », les GPU, etc. est un signe supplémentaire de l'importance de ce phénomène.

[8] http://www.caps-entreprise.com

Le passage du Petascale à l'Exascale ne se fera pas sans innovations et lourds développements.

La manifestation « Supercomputing » n'a cessé de se renforcer, et l'édition 2008 n'a pas failli à cette règle malgré les difficultés économiques actuelles.

Jean-Loïc Delhaye

SC'08 : du côté des constructeurs

L'exposition associée à la conférence SC'08 est un moment privilégié pour faire un tour d'horizon des projets des principaux constructeurs de systèmes HPC dans le monde. Les constructeurs sollicités ont répondu positivement, sauf HP, en proposant soit un entretien individuel soit de participer à une réunion d'information organisée avec des responsables de la société. Voici les principaux résultats de ces entretiens ou réunions.

Bull

Bull confirme la croissance retrouvée et met en avant deux secteurs : celui du calcul de haute performance et celui du stockage de données.

Bull a maintenant trois systèmes classés dans le TOP500 : ceux installés au CEA (processeurs Itanium, en position 48 et 63) et celui installé récemment à l'Université de Cardiff (processeurs Xeon, en position 171). ServiWare, filiale à 100% de Bull, a également un système classé au TOP500 (IFP, 498ème position). Par ailleurs, Bull va fournir au centre de calcul de Jülich un supercalculateur d'une puissance de 200 TeraFlops (processeurs Xeon) dans le cadre du projet JuRoPa.

Les systèmes HPC de Bull font partie de la gamme NovaScale et la majorité des offres utilisent des processeurs Xeon bi-sockets. Parallèlement, Bull a engagé d'importants développements autour des GPU de Nvidia.

Les principaux axes de développement sont les suivants :

- Le projet Tera 100, comprenant un contrat de développement avec le CEA
- Des lames « orientées HPC », c'est-à-dire dans lesquelles on aura enlevé tout ce qui est « inutile » pour ce type d'application.
- Un projet avec Nvidia visant à intégrer des GPU dans des lames.

Bull confirme sa montée en puissance dans le domaine du calcul de haute performance, ainsi que son partenariat stratégique avec Intel et attend beaucoup du processeur Nehalem. L'interconnexion de gros serveurs SMP (plusieurs dizaines de cœurs par SMP) est à la base de la stratégie de la compagnie.

Cray

Pete Ungaro, le PDG de Cray a annoncé que la société avait connu 50% de croissance en 2008 et qu'elle devrait devenir bénéficiaire rapidement. La moitié de son CA est fait en dehors des Etats-Unis, en particulier en Europe.

La ligne XT connait un grand succès, et Cray cite les systèmes « Hector » (EPSRC, Royaume Uni), le CSCS en Suisse, la météo danoise ; aux Etats-Unis, la machine « Red Storm » de Sandia a évolué vers des quadri-cœurs (284 TFlops), « Jaguar » à ORNL est au niveau du PetaFlops, etc. Plusieurs applications scientifiques ont dépassé le PetaFlops soutenu sur le système Jaguar (performance crête de 1.6 PetaFlops). Le dernier TOP500 liste 10 systèmes Cray parmi les 20 systèmes les plus puissants installés dans le monde, ce qui montre l'importance de cette compagnie dans le très haut de gamme.

Le nœud XT5 est un SMP 8 voies, avec 2 processeurs Opteron quadri-cœurs et une performance d'environ 70 GigaFlops. Une armoire contient 192 processeurs Opteron (soit actuellement 768 cœurs). Une configuration de 6 armoires XT5 contient 1112 processeurs pour 43 TeraFlops et une consommation énergétique inférieure à 250 kW (critère présent chez tous les constructeurs !). Une configuration de 144 armoires fournit une puissance de 1 PetaFlops.

Cray va prochainement annoncer officiellement le système XT5m, dérivé de la technologie XT, qui cible les configurations de 1 à 6 armoires. Un système XT5m de 6 armoires offrira une performance crête de l'ordre de 45 TeraFlops avec les processeurs Quadri-cœurs actuels. La consommation énergétique sera réduite et le refroidissement sera fait par air.

Quels seront les futurs systèmes massivement parallèles de Cray ? Les systèmes appelés « Baker » seront basés sur le nouveau processeur Opteron d'AMD (8 cœurs), avec un nouveau système d'interconnexion appelé « Gemini ». Un système de 80 armoires disposera de 15.024 cœurs et fournira une puissance crête de 1,44 PetaFlops pour une occupation au sol de 195 m^2. Les systèmes Baker évolueront en fonction des nouveaux processeurs AMD.

Steve Scott, CTO de Cray, a présenté la stratégie de Cray, en particulier dans le cadre du futur système « Cascade » que Cray doit cons-

truire dans le cadre du programme HPCS[9] de la DARPA. De nouveaux accords ont été signés avec AMD et Intel ; les processeurs AMD seront à la base des systèmes « Baker » tandis qu'un nouveau partenariat avec Intel se concentre sur de nouveaux processeurs destinés aux systèmes de très grande performance (« high-end »). C'est Intel qui fournira les processeurs de « Cascade » : les nœuds de calcul « Marble » basés sur le processeur Intel Xeon, et les noeuds de calcul « Granite » qui exploitera un « custom processor » très haute performance, avec de *nombreux* cœurs, des dispositifs vectoriels, etc. Cascade sera composé de lames « Marble » et de lames « Granite ». De nouveaux systèmes d'interconnexion sont en cours de développement, dont le routeur « Aries » capable de délivrer des dizaines de millions de messages par seconde avec une latence de l'ordre de la microseconde, le réseau « Dragonfly » destiné à la machine Cascade, etc. Les développements sur les logiciels sont destinés à contribuer à l'amélioration de la productivité, au sens large : compilateurs, outils de debogage, logiciels scientifiques, et le nouveau langage de programmation parallèle développé par Cray, « Chapel », dont une première version vient d'être annoncée ; Chapel sera disponible sur tout cluster.

IBM

IBM continue de dominer le TOP500 sur le paramètre performance mais la concurrence se renforce (IBM représente 38% de la performance installée en novembre 2008, contre 48% en juin 2008). Il reste le constructeur du système le plus puissant installé dans le monde : le « Road Runner » de Los Alamos.

La compagnie organise ses offres dans le HPC autour de quatre « familles » : les processeurs POWER, les systèmes Blue Gene, les accélérateurs basés sur le processeur CELL BE, les clusters Linux s'appuyant sur les processeurs multi-cœurs disponibles sur le marché.

Les processeurs POWER. Le processeur commercialisé actuellement est le POWER6, bi-cœurs. Il est à la base du p575, cheval de bataille d'IBM pour le HPC (IDRIS, SARA, etc), avec 32 cœurs dans un format 2U. Son successeur, le POWER7, sera disponible des 2010. Le POWER7 prendra le « virage du multi-cœurs » mais IBM a choisi de limiter le nombre de cœurs à 8, en privilégiant leur puissance (fréquence élevée, performance élevée par cœur).

IBM proposera 3 chips à base de POWER7 pour ses serveurs, visant des marchés différents ; le haut de gamme sera à la base de PERCS et de sa version qui doit être installée au NCSA (Université de l'Illinois) en 2011 : BLUE WATERS[10]. Rappelons que PERCS est la réponse d'IBM au programme HPCS[11] de la DARPA : il s'agit de fournir, dès 2010, un prototype de 1 PetaFlops démontrant la justesse des concepts retenus par le constructeur, puis de livrer, dès 2011, des systèmes pouvant fournir une puissance de 2 à 4 PetaFlops « soutenus ». Ces supercalculateurs sont des systèmes multi-cœurs homogènes. Avec plus de 500.000 cœurs, BLUE WATERS devrait fournir une puissance crête d'environ 10 PetaFlops.

Les accélérateurs et le CELL. IBM devrait annoncer prochainement une « stratégie accélérateur » (code : pNext) qui s'appuiera sur des standards de programmation comme OpenMP et OpenCL. L'architecture de ces processeurs sera différente (mais compatible) de celle du CELL. Quant au CELL lui-même, son évolution sera liée à celle du marché.

Blue Gene. La version Q du système Blue Gene sera disponible en 2011, avec une configuration maximale délivrant une puissance crête de 20 PetaFlops. Blue Gene/Q utilisera un nouveau processeur et un nouveau réseau, en cours de développement. Une armoire (rack) pourra contenir jusqu'à 512 nœuds (104 TeraFlops) avec refroidissement par air, ou 1024 nœuds (208 TeraFlops) avec refroidissement par eau. BG/Q améliorera d'un facteur 10 le rapport MFlops/Watt de BG/L.

Clusters Linux. IBM poursuit cette ligne, et le cluster installé à Barcelone reste une référence. L'offre iDataPlex est son « cheval de bataille » sur ce segment ; basé sur le processeur Xeon, il vise à réduire le « coût total d'exploitation » du système (surface au sol, consommation énergétique, etc). Les premières installations « recherche » en France : Institut d'Astrophysique de Paris (140 nœuds), IN2P3 à Villeurbanne, INRA à Toulouse).

NEC

NEC poursuit sa « stratégie vectorielle » sur la base de la gamme SX-9, tout en s'engageant dans les architectures « hybrides » incluant des accélérateurs et des GPU. Le processeur du SX-9 améliore celui du SX-8, en particulier par l'ajout d'une unité arithmétique et de plusieurs pipelines vectoriels. Le nœud SMP comprend 16 processeurs et fournit 1,6 TeraFlops de puissance crête. La performance crête de la configuration maximale (512 nœuds) est de 970 TeraFlops.

[9] Voir le numéro 50 de Bi-ORAP

[10] http://www.ncsa.uiuc.edu/BlueWaters/

[11] Voir Bi-Orap numéros 49 et 50

Les accords récents avec Nvidia ont permis d'ajouter 170 systèmes Tesla S1070 sur l'ordinateur « Tsubame » de l'Institut de Technologie de Tokyo (qui dispose aussi d'accélérateurs ClearSpeed). L'accord signé avec le HLRS (centre de calcul de Stuttgart) est aussi destiné à développer de nouvelles approches du « supercalcul hybride ».

NEC reste présent dans les clusters HPC, en particulier grâce à la gamme NEC LX.

Rappelons que NEC est très engagé dans le « Next-Generation Supercomputer Project », qui vise à redonner au Japon la première place dans le TOP500 et qui devrait avoir également des retombées importantes sur les futurs produits de la famille SX.

SGI

Le déclin de SGI semble enrayé, avec une forte croissance du chiffre d'affaire et des commandes (ces dernières ont cru de 78% entre le dernier trimestre 2007 et le dernier trimestre 2008. Un tiers du CA provient des universités et des centres nationaux, un autre tiers des entreprises, le dernier tiers provenant de clients divers, en particulier de l'industrie du cinéma d'animation.

Les succès de SGI sont particulièrement nets en Europe, avec le HLRN en Allemagne (25.000 cœurs répartis entre Berlin et Hanovre), le CINES (12.288 cœurs) et TOTAL (10.240 cœurs) en France.

Le « système phare » de SGI dans le HPC est l'Altix ICE 8200, basé sur le processeur Xeon d'Intel. La configuration la plus importante actuellement d'un tel système est celle de la NASA, avec 51.200 cœurs, la plaçant en 3ème position au TOP500 avec 487 TeraFlops

Les environnements Linux et les architectures hybrides dominent sur ces nouveaux grands systèmes. Les marchés en forte croissance concernent les domaines de l'image (dont le cinéma) et ceux liés à la gestion de très grands volumes de données « non structurées ».

Pour le CTO de SGI, l'accélération du nombre de cœurs par processeurs va se poursuivre (on se dirige vers des chips comportant 128 cœurs) et le million de cœurs sur un seul système n'est pas loin. Une difficulté majeure réside dans la « scalabilité » des logiciels. Les communications entre processeurs et mémoires sont un autre défi : la bande passante est un frein de plus en plus évident à la performance globale (remplacer le cuivre par des technologies optiques ?), de même que la latence.

Les processeurs Intel continueront d'être à la base des systèmes HPC de SGI. Les prochaines étapes intégreront les processeurs Itanium « Tukwila » (davantage de performance et de multithreading) et les processeurs Xeon « Nehalem » dotés d'une nouvelle micro-technologie. La technologie du silicium va également progresser, avec la « feuille de route » suivante : 32 nm en 2009, 22 nm en 2011, 16 nm en 2013.

La consommation énergétique et le refroidissement sont l'objet de travaux de recherche importants. Les technologies actuelles donnent des consommations de plusieurs centaines de MW pour un système exaflopique (qui pourrait voir le jour vers 2018), ce qui est inacceptable ! Il faut impérativement limiter à 100 MW la consommation d'un tel système (ce qui correspondrait à environ 100 KW pour un système pétaflopique).

Les responsables de la société sont confiants dans sa capacité à relever les défis qui sont devant elle. La prochaine étape sera la livraison du système appelé « Ultraviolet » qui doit être livré en 2009 à la NASA et qui utilisera le processeur Tukwila.

Sun

Les responsables de Sun confirment l'importance que la société accorde au calcul de haute performance, avec deux axes principaux : la virtualisation des services d'entreprise (ou « cloud computing ») et le HPC proprement dit. En automne 2008, Sun annonce 2,4 PetaFlops en cours de livraison et prévoit un chiffre d'affaires de 2 milliards de dollars pour les 18 mois à venir pour le seul HPC. Le système le plus puissant installé actuellement par Sun est le SunBlade x6420 du TACC (Université du Texas) avec 62.976 cœurs (processeurs Opteron), classé en 6ème position du TOP500 avec une puissance Linpack de 433 TeraFlops (579 TeraFlops crête).

Les deux « lignes » de produits, basées sur processeurs AMD d'une part, sur processeurs Intel d'autre part sont aussi confirmées, le choix dépendant surtout des applications du client concerné.

Les solutions innovantes pour le stockage de grands volumes de données sont toujours mises en avant. Le « flash storage » apparait, entre les disques et la mémoire RAM, et Sun parle de « stockage hybride » géré par un système de fichiers unique. Avec 2 TeraOctets sur 1U (80 « Flash DIMMs » de 24 GigaOctets chacun), ce nouveau niveau de stockage permet de ne transférer les données vers les disques que si elles doivent être réellement stockées et conservées. La possibilité de faire entre 560.000 et 800.000 écritures par seconde, et entre 2,1 et 3,2 millions de lectures par se-

conde, en mode aléatoire, est un atout pour les applications manipulant beaucoup de données.

ClusterVision

Il m'a semblé intéressant de mentionner ClusterVision, seule société européenne à afficher 5 systèmes dans le TOP500 : 2 en collaboration avec Dell (positions 200 et 234), 2 en collaboration avec IBM (positions 86 et 496) et 1 sans partenariat spécifique (position 459).

La société annonce 300 clusters installés, dont 95% en Europe. Les partenariats avec Dell d'une part, IBM d'autre part, sont très importants.

Le système d'exploitation proposé actuellement sur les clusters est soit « ClusterVision OS » (qui permet à l'administrateur système d'avoir la vision d'un système unique, et qui comprend de nombreux outils d'administration du cluster), soit Microsoft HPC Server 2008.

Jean-Loïc Delhaye

XtreemOS : première version du système d'exploitation

Le projet européen «XtreemOS» a pour objet de construire et promouvoir un système d'exploitation fondé sur Linux qui offre un support natif aux organisations virtuelles sur des grilles de nouvelle génération. Le consortium «XtreemOS» est composé de 19 partenaires industriels et académiques situés dans 8 pays différents (7 en Europe et la Chine). Dans le cadre de ce consortium, l'INRIA assure la coordination scientifique.

Une première version du système d'exploitation de« XtreemOS» vient d'être rendue publique[12].

Les fonctionnalités du système XtreemOS permettent :

- D'assurer la protection des données, des applications et des ressources dans un environnement distribué, administré de manière décentralisée par différentes institutions ;
- De s'auto-configurer face aux défaillances d'ordinateurs ou de liens du réseau et face à la dynamicité des grilles (une institution peut décider à tout moment d'ajouter ou de soustraire des ressources à une grille) ;
- D'assurer l'exécution fiable des applications distribuées en dépit des fréquentes reconfigurations et défaillances pouvant survenir dans une grille ;
- De gérer efficacement l'allocation des ressources et les accès aux données pour garantir de hautes performances aux applications.

Les premières expériences sont réalisées sur la plate-forme expérimentale Grid'5000.

Contact : christine.morin@inria.fr

[12] http://www.xtreemos.eu/

Strasbourg : nouvelle organisation du calcul scientifique

Introduction

Le premier janvier 2009 a marqué la naissance de l'Université de Strasbourg (UdS).

Issue de 5 établissements (les trois universités, l'IUFM de l'Académie de Strasbourg, et le Pôle Européen), elle rassemble 42 000 étudiants, dont 21% d'étudiants étrangers, 5 230 personnels, 39 composantes (unités de formation et de recherche, facultés, écoles, instituts) et 86 unités de recherche.

Au niveau informatique, la création de cette université s'accompagne de la création d'une Direction Informatique, dont le directeur est Pierre David, regroupant les 9 services informatiques des anciens établissements.

En particulier, les services de calcul scientifique sont dorénavant intégrés dans cette Direction Informatique sous la forme d'un Département d'Expertise pour la Recherche, présenté dans cet article.

1) Intégration au projet d'établissement

Le schéma directeur informatique et TIC prévoit que la Direction Informatique réponde aux besoins spécifiques des chercheurs et enseignants-chercheurs en matière :

- de calcul scientifique ;
- de modélisation ;
- de visualisation ;
- de parallélisation de code ;
- de statistique et d'analyse des données.

Chargé de répondre à ces besoins, le département « Expertise pour la Recherche » doit non seulement mettre à disposition des utilisateurs des ressources de calcul, mais également ac-

compagner les chercheurs dans la maîtrise des usages, des techniques et des outils correspondants. Pour l'Université de Strasbourg, ce département constitue le centre de compétences unique pour tous ces domaines. Il correspond ainsi au méso-centre de calcul de l'UdS.

Le département est doté d'un conseil scientifique, dont la composition sera déterminée au cours du premier semestre 2009.

2) Organisation

2.1) Intégration des services préexistants

Le Département Expertise pour la Recherche regroupe dans son périmètre :

- les activités du Centre d'Etudes du Calcul Parallèle et de la Visualisation de l'ancienne Université Louis Pasteur (ULP), déjà présentées dans Bi-Orap n°53 ;
- les activités de calcul scalaire intensif du Centre Universitaire Régional de Ressources Informatique (CURRI) de l'ULP.

Ces deux entités sont riches d'une expérience de la relation de proximité avec les utilisateurs, qui est au cœur des missions du nouveau Département.

La nouvelle Université de Strasbourg intégrant des domaines liés aux sciences humaines, le schéma directeur a tenu à étendre les attributions du département « Expertise pour la Recherche » au conseil en statistiques. Il s'agit là d'un domaine de compétences à structurer et à faire évoluer dans le courant de l'année 2009.

La restructuration de ces services assure une plus grande lisibilité des activités de support informatique à la recherche scientifique.

2.2) Membres de l'équipe

Romaric David, précédemment responsable technique du CECPV, a été nommé responsable du département « Expertise pour la Recherche ». L'équipe est également composée de Mehdi Amini (en provenance du CECPV) et de Michel Ringenbach (en provenance du CURRI).

2.3) Conseil scientifique

Le département « Expertise pour la Recherche » est doté d'un conseil scientifique propre. Le CS assurera le suivi et le pilotage des activités sous l'angle des grands domaines disciplinaires consommateurs de ressources de calcul. Une des questions que pourra traiter le conseil scientifique est le choix des projets de recherche auxquels participeront les ingénieurs du département.

3) Les plus pour les utilisateurs

Rappelons que le cœur de métier du département est l'accompagnement des utilisateurs (conseil, etc.) dans leur travail de recherche. Dans ce cadre, un des grands objectifs de cette restructuration est bien évidemment d'améliorer la qualité des services fournis aux utilisateurs.

Sur un terme plus rapproché, quels sont les changements sensibles ?

3.1) Un accès simplifié

Jusqu'à présent, l'accès aux différentes ressources de calcul se faisait par des canaux bien distincts selon le type de calculs (scalaires, parallèles). Un des premiers objectifs du département est l'unification des moyens d'accès aux ressources de calcul, afin de mettre en place une plate-forme d'informatique scientifique cohérente à l'échelle du site strasbourgeois.

3.2) Un accès assoupli

Cette remise à plat sera l'occasion également d'engager une réflexion globale avec les utilisateurs sur leurs besoins afin d'harmoniser les politiques d'exploitation. Cela permettra :

- de diriger naturellement certains utilisateurs vers le calcul parallèle, par la connaissance des applications. En effet, certaines applications ne sont pas utilisées en mode parallèle par simple habitude ;
- d'accueillir de nouvelles applications qui ne tournaient pas sur les ressources communes faute de profil adapté.

3.3) Un accès plus performant

La mise en commun des compétences de la Direction Informatique permettra aux utilisateurs de bénéficier d'un support technique et scientifique spécifique s'appuyant sur des ressources consolidées. En particulier, la collaboration avec le département « Exploitation » permettra de spécialiser le travail des ingénieurs du département « Expertise pour la Recherche ».

4) Conclusion

L'intégration des activités de calcul scientifique dans la Direction Informatique de l'Université de Strasbourg représente une étape importante de par le passage à l'échelle et la visibilité accrue qui en découlent. Les utilisateurs seront les grands gagnants de cette restructuration en raison de l'agrandissement de l'équipe. Le pilotage opéré par le conseil scientifique permettra de situer le travail dans le long terme.

Romaric David (david@unistra.fr)

Kerlabs : pour des clusters « made easy »

La société Kerlabs[13] est une spin-off de l'INRIA créée en octobre 2006 avec pour objectif d'industrialiser la technologie Kerrighed. Kerrighed est un système d'exploitation pour clusters dont le développement a débuté à l'INRIA en 1999. Depuis 2006, Kerrighed est développé principalement par la société Kerlabs et totalise à ce jour près de 30 hommes années de recherche et développement.

Kerrighed - un SMP virtuel

Kerrighed est un système d'exploitation distribué dérivé de Linux qui permet de créer un SMP virtuel sur un cluster de Pcs : un système à image unique ou Single System Image (SSI). Un cluster équipé de la technologie Kerrighed semble en tous points identiques à une machine SMP : l'utilisateur, l'administrateur et le programmeur ont l'illusion de travailler sur une unique machine de forte puissance. La complexité engendrée par la distribution des ressources d'un cluster est ainsi totalement masquée.

Kerrighed offre une interface identique à celle d'un système Linux et peut être intégré au sein de n'importe quelle distribution existante. Ainsi, toutes les applications existantes et fonctionnant sous Linux peuvent être exécutées sur Kerrighed sans aucune modification et alors tirer parti de l'ensemble des ressources du cluster.

Parmi les fonctionnalités de Kerrighed on peut citer l'équilibrage automatique de charge, permettant d'écouler des charges de calcul de manière optimum en plaçant et déplaçant au besoin les tâches sur les machines les moins chargées. Cet écoulement de charge peut être réalisé de manière interactive ou de manière programmée par l'intermédiaire d'un système de type « batch ». Le comportement de l'ordonnanceur global distribué peut de plus être configuré à chaud pour s'adapter aux besoins ou contraintes spécifiques des utilisateurs ou administrateurs.

De la même manière, il est possible de mettre en place des politiques d'utilisation de ressources par application, par utilisateur ou par groupe d'utilisateurs, le tout à l'échelle du cluster. On peut ainsi limiter l'accès à certaines ressources pour certains utilisateurs ou contrôler la quantité de ressources allouées à certaines applications.

[13] http://www.kerlabs.com

1 Téra de mémoire sous votre bureau

La dernière innovation de la société Kerlabs est la technologie « Terabox ». Cette technologie permet d'offrir un mini super-ordinateur à faible coût et pouvant s'installer sans contraintes sous un bureau.

Il s'agit en réalité d'un mini-cluster intégré dans un petit châssis et équipé de la dernière génération du système Kerrighed. Cette nouvelle version du système intègre un mécanisme inédit d'agrégation mémoire qui permet d'agglomérer les mémoires des différentes machines du cluster et de les mettre à disposition de toutes les tâches fonctionnant sur le système. Cinq machines disposant chacune de 200Go de mémoire offrent ainsi une capacité mémoire totale de 1 Téraoctet et d'un nombre total de cœurs pouvant aller jusqu'à 40. Le tout pour un tarif bien inférieur à celui d'un unique serveur avec 1To de mémoire.

Le secret de cette technologie réside dans l'exploitation de réseaux à haut débit (Infiniband, SCI ou encore Ethernet 10G). Le mécanisme d'agrégation mémoire de Kerrighed remplace le traditionnel « swap » sur disque par un système d'échange de mémoire avec les autres machines du cluster via le réseau. La mémoire des autres machines devient ainsi un nouveau niveau de cache, bien plus rapide que le traditionnel swap sur disque. On constate une amélioration de performance d'un facteur 20 à 50 dans la version actuelle du système et des performances bien supérieures sont attendues pour les prochaines versions du système. Pour s'en convaincre, il suffit de comparer le débit disque lors d'accès non séquentiels (quelques dizaines de Mo par seconde) au débit d'un réseau à haute performance (10 Gbits par seconde).

Vers des clusters « verts »

Kerlabs vient de débuter des travaux liés à la gestion de l'énergie dans les clusters dans le cadre du projet EcoGrappe de l'ANR Arpege. Ces travaux, réalisés en partenariat avec l'INRIA et EDF visent à limiter la consommation électrique d'un cluster en ajustant sa taille en fonction de sa charge d'utilisation. Des machines pourront ainsi être éteintes et allumées en fonction de la charge en déplaçant au besoin des applications en cours de fonctionnement pour regrouper les tâches sur un sous-ensemble de machines.

A noter enfin la participation de Kerlabs au projet PetaQCD dans le cadre de l'ANR COSINUS. Les travaux de Kerlabs au sein de ce projet viseront principalement les problématiques de

haute disponibilité et d'extensibilité du système Kerrighed pour les clusters de grande taille.

Renaud Lottiaux
contact@kerlabs.com

Lire - Participer

Lire

- Le numéro 9/08 de la lettre d'information de DEISA
  http://www.deisa.eu/news_press/newsletter/

Participer

- Journées « *Informatique Massivement Multiprocesseurs et Multicoeurs* ». Paris, 4 et 5 février 2009.
- Le centre de recherche INRIA de Grenoble organise la deuxième école d'hiver sur le thème « *Hot Topics in Distributed Computing* ». La Plagne, du 15 au 20 mars.
  http://sardes.inrialpes.fr/~quema/htdc2009
- Ecole thématique Archi09 : *Architectures des systèmes matériels enfouis et méthodes de conception associées*. Pleumeur-Bodou, 30 mars au 4 avril
  http://www.irisa.fr/archi09/
- La prochaine édition de la conférence conjointe RenPar, Sympa et CFSE est organisée par l'IRIT à Toulouse du 9 au 11 septembre 2009.
  - *RenPar'19: Rencontres francophones du Parallélisme*
  - *SympA'13 : Symposium en Architectures des machines*
  - *CFSE'7 : Conference Française sur les Systèmes d'Exploitations*

  http://www.irit.fr/Toulouse2009

NOUVELLES BREVES

➔ « Green IT » en France

La ministre de l'Economie, de l'Industrie et de l'Emploi, et le secrétaire d'Etat chargé de l'Industrie et de la Consommation, ont lancé un groupe de réflexion « Green IT » pour favoriser une utilisation éco-responsable des technologies de l'information et de la communication. Monsieur Michel Petit, président de la section scientifique et technique du Conseil général des technologies de l'information (CGTI) a été chargé de constituer ce groupe de réflexion.

➔ 13,5 TeraFlops à l'IAP

ServiWare a intégré et installé juste avant Noel un système de 13,5 Teraflops à l'Institut d'Astrophysique de Paris (http://www.iap.fr/) dans le cadre du projet Planck. Ce système réunit, par un commutateur InfiniBand non bloquant Voltaire, 140 serveurs IBM basés sur la technologie iDataplex qui permet d'atteindre une très haute densité pour une consommation d'électricité réduite et de moindre besoins en matière de climatisation.
Dans le cadre de ce projet, ServiWare met en œuvre le système de fichiers GPFS d'IBM qui accède un système de stockage IBM DCS 9900 (Data Direct Networks) de 120 To.
Un second ensemble de stockage, NFS, de plus de 200 To complète ce système de calcul qui va être mis au service du traitement des données du projet Planck Surveyor.

http://public.planck.fr/

➔ De EGEE à EGI

EGEE (Enabling Grids for E-sciencE) a permis de créer la plus importante grille pluridisciplinaire de calcul dans le monde. Le projet européen EGI_DS (European Grid Initiative Design Study) doit permettre de mettre en place une infrastructure réellement durable avec une excellente qualité de service. Cette grille sera formée des initiatives nationales (NGI : National Grid Initiatives) et d'une organisation légère, EGI, qui devra s'assurer de la cohérence entre les diverses NGI.

http://web.eu-egi.eu/

➔ PRACE

PRACE a tenu sa première conférence scientifique dans le cadre d'ICT 2008 à Lyon (26 novembre 2008). Elle était centrée sur les applications, les architectures et la formation nécessaire pour le calcul de très haute performance. Les présentations sont disponibles sur le site :

http://www.prace-project.eu/documents

➔ Se familiariser avec Chapel

Rappelons que Chapel est le langage de programmation parallèle développé par Cray dans le cadre du programme HPCS (High Productivity Computing Systems) financé par la DARPA américaine. L'objectif est d'améliorer la productivité des développeurs d'applications sur les grands systèmes de calcul intensif. La première version publique de ce langage est maintenant accessible à tous ceux qui s'intéressent à cette évolution des langages de programmation :

http://chapel.cs.washington.edu/

➔ Le CASC a 20 ans

Créée en 1989, la « Coalition for Academic Scientific Computation » (CASC) est une association américaine qui rassemble 57 institutions représentant les universités et centres de calcul les plus dynamiques dans le domaine du HPC. CASC entend faire la promotion du calcul de haute performance, fournir l'expertise dont peuvent avoir besoin les autorités fédérales, faciliter les échanges d'informations à l'intérieur de la communauté scientifique académique. A l'occasion de son 20ème anniversaire, CASC a publié une brochure.

http://www.casc.org/

➔ Arabie Saoudite

L'Arabie Saoudite a lancé, avec KAUST (King Abdullah University of Science and Technology), un projet très ambitieux dans le domaine de la recherche. Un partenariat avec IBM va permettre de créer un centre de calcul intensif de niveau mondial avec un système Blue Gene/P d'une performance crête de 222 TeraFlops, qui devrait évoluer vers le PetaFlops avant la fin de 2010. Ce système sera ouvert aux chercheurs de KAUST mais aussi d'autres universités dans le monde, à travers les réseaux Internet2 et Geant2. Plusieurs spécialistes mondiaux du HPC auraient déjà accepté de rejoindre KAUST pour plusieurs années.

➔ Chine

- Lenovo est le constructeur du Shenteng 7000, situé au rang 19 du TOP500, et installé au « Computer Network Information Center ». Avec 12216 cœurs Xeon, sa puissance crête est de 146 TeraFlops.

  http://www.cnnic.net.cn/en/index/
- Un superordinateur Dawning 5000A, construit en Chine, sera livré au SSC (Shanghai Supercomputer Center) en avril 2009. Equipé de processeurs AMD, il aura une performance crête supérieure à 200 TeraFlops.

  http://www.ssc.net.cn/en/

➔ Cray

- ORNL (Oak Ridge National Laboratory) a prononcé la réception de la nouvelle configuration de « Jaguar ». Ce système Cray XT a franchi la barre du PetaFlops en mode « soutenu » (et 1,64 PetaFlops « peak »).
- Cray commencera à livrer, début 2009, des XT5 équipés de processeurs AMD quadri-cœurs (« Shanghai »).

➔ IBM

- IBM a créé un « Green data center » dans son centre de support de Montpellier. Il entend en faire une vitrine pour le monde du savoir-faire d'IBM en matière d'informatique verte. (source : Les Echos, 12/12/08)

AGENDA

20 janvier 2009 – **DAMP 09** : Workshop on Declarative Aspects of Multicore Programming (Savannah, Georgia, Etats-Unis)

25 au 28 janvier 2009 – **HiPEAC 2009** : The 4th International Conference on High Performance and Embedded Architectures and Compilers (Paphos, Chypre)

25 au 28 janvier 2009 – **Rapido'09** : 1st Workshop on: Rapid Simulation and Performance Evaluation: Methods and Tools (Paphos, Chypre)

25 au 28 janvier 2009 - **INA-OCMC'09** : Third Workshop on Interconnection Network Architectures: On-Chip, Multi-Chip (Paphos, Chypre)

4 au 5 février 209 – **IMMM** : Journées « Informatique Massivement Multiprocesseur et Multi-coeur » (Paris, France)

18 au 20 février 2009 – **PDP 2009** : 17th International Conference on Parallel, Distributed and network-based Processing (Weimar, Allemagne)

18 au 20 février 2009 – **MSOP2P 2009** : 3rd International Workshop on Modeling, Simulation, and Optimization of Peer-to-peer Environments (Weimar, Allemagne)

1 au 6 mars 2009 – **ICONS 2009** : The Fourth International Conference on Systems (Guadeloupe, France)

2 au 6 mars 2009 – **SimuTools 2009** : Second International Conference on Simulation Tools and Techniques (Rome, Italie)

10 au 13 mars 2009 – **ARCS 2009** : 22nd International Conference on Architecture of Computing Systems (Delft, Pays-Bas)

15 au 20 mars 2009 – Ecole d'hivers sur le thème « *Hot Topics in Distributed Computing* ». La Plagne.

16 au 20 mars 2009 – **PerComm 2009** : IEEE International Conference on Pervasive Computing and Communications (Dallas, Etats-Unis)

25 au 26 mars 2009 – **MRSC 2009** : Many-core and Reconfigurable Supercomputing Conference (Berlin, Allemagne)

29 au 30 mars 2009 – **EGPGV'09** : Eurographics 2009 Symposium on Parallel Graphics and Visualization (Munich, Germany)

30 au 31 mars 2009 – **HotPar'09** : First USENIX Workshop on Hot Topics in Parallelism (Berkeley, Ca, Etats-Unis)

30 mars au 4 avril 2009 – **Archi'09** : Ecole thématique : Architectures des systèmes matériels enfouis et méthodes de conception associées (Pleumeur-Bodou)

1er au 4 avril 2009 – **ICST 2009** : Second International Conference on Software Testing, Verification, and Validation (Denver, Co, Etats-Unis)

19 au 21 avril 2009 – **ISPASS 2009** : International Symposium on Performance Analysis of Systems and Software (Boston, Ma, Etats-Unis)

20 au 24 avril 2009 – **MDD4RESS** : 4th International School on Model-Driven Development for Distributed, Realtime, Embedded Systems (Aussois, France)

23 au 24 avril 2009 – **SCOPES 2009** : 12th International Workshop on Software and Compilers for Embedded Systems (Nice, France)

26 au 28 avril 2009 – **ISPASS** : 2009 IEEE International Symposium on Performance Analysis of Systems and Software (Boston, Ma, Etats-Unis)

4 au 8 mai 2009 – **GPC'09** : 4th International Conference on Grid and Pervasive Computing (Genève, Suisse)

10 au 13 mai 2009 – **NOCS 2009** : The 3rd ACM/IEEE International Symposium on Networks-On-Chip (San Diego, Ca, Etats-Unis)

10 au 13 mai 2009 – **AICCSA 2009** : The seventh ACS/IEEE International Conference on Computer Systems and Applications (Rabat, Maroc)

11 au 13 mai 2009 – PRACE / DEISA Scientific Conference (Amsterdam, Pays-Bas)

18 au 20 mai 2009 – **Frontiers 2009** : 2009 International Conference on Computing Frontiers (Ischia, Italie)

18 au 20 mai 2009 – **MAW'09** : Workshop on Issues and Solutions for Memory Access on Cache Architectures: "Main memory: Fit for Manycore?" (Ischia, Italie)

18 au 21 mai 2009 – **CEC 2009** : Special Session on Parallel Bio-inspired Optimisation Methods: Algorithms and Applications (Trondheim, Norvège)

25 au 27 mai 2009 – **ICCS 2009** : International Conference on Computational Science (Baton Rouge, La, Etats-Unis)

25 au 27 mai 2009 – Workshop *on* "Using Emerging Parallel Architectures for Computational Science" (Baton Rouge, La, Etats-Unis)

25 au 27 mai 2009 – **PAPP 2009** : Sixth International Workshop on aPpplications of declArative and object-oriented Parallel Programming (Baton Rouge, La, Etats-Unis)

25 mai 2009 – **HPGC 2009** : The Sixth High-Performance Grid Computing Workshop (Rome, Italie)

25 mai 2009 – **HiCOMB 2009** : 28th IEEE International Workshop on High Performance Computational Biology (Rome, Italie)

25 mai 2009 – **HP-PAC** : Fifth IEEE Workshop on High-Performance, Power-Aware Computing (Rome, Italie)

25 au 29 mai 2009 – **IPDPS 2009** : 23rd IEEE International Symposium on Parallel & Distributed Processing (Rome, Italie)

25 au 29 mai 2009 – **LSPP 2009** : Workshop on Large-Scale Parallel Processing (Rome, Italie)

29 mai 2009 – **Hot-P2P 2009** : Sixth International Workshop on Hot Topics in Peer-to-Peer Systems (Rome, Italie)

29 mai 2009 – **PDCoF 2009** : The Second Workshop on Parallel and Distributed Computing in Finance (Computational Finance) (Rome, Italie)

29 mai 2009 – **Hot-P2P** : Sixth International workshop on Hot Topics in Peer-to-Peer Systems (Rome, Italie)

3 au 5 juin 2009 – **IWOMP 2009** : International Workshop on OpenMP (Dresde, Allemagne)

Les sites de ces manifestations sont accessibles depuis le serveur ORAP.

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
Irisa, campus de Beaulieu, 35042 Rennes
Tél : 02 99 84 75 33, fax : 02 99 84 74 99
chantal.letonqueze@irisa.fr
http://www.irisa.fr/orap