ORganisation Associative du Parallélisme

BI-ORAP

Bulletin d'information
de l'ORganisation Associative du Parallélisme

Numéro 25 - octobre 2000

Sommaire

11ème Forum Orap

Le 11ème Forum de l'ORAP aura lieu le **27 mars 2001** dans les locaux d'EDF à Clamart.

Deux thèmes seront abordés :

- Les "petits clusters"
- Les applications "non scientifiques"

Le programme sera diffusé à la fin de cette année.

Stratégie d'évolution des équipements de simulation du CEA/DAM : le projet TERA

Jean Gonnord

Commissariat à l'Energie Atomique
Direction des Applications Militaires

Le Programme Simulation

En février 1996, après avoir réaffirmé que la Dissuasion Nucléaire constituera pour longtemps l'élément essentiel de notre défense, le Président de la République concluait la dernière série de tests effectués dans le Pacifique en demandant au CEA/DAM de relever le défi de la Simulation.

Le Programme Simulation a ainsi pour objectif de garantir la sûreté et la fiabilité des armes nucléaires françaises. Il permet d'assurer la crédibilité de celles-ci. La loi de programmation militaire prévoit 25 milliards de Francs pour ce programme majeur, qui mobilisera au CEA/DAM entre 800 et 1000 ingénieurs pendant 10 ans.

Le Programme Simulation comprend trois grands volets : l'amélioration des modèles physiques, la simulation numérique et la validation expérimentale. Les simulations numériques, qui mettent en oeuvre ces modèles physiques améliorés, sont validées de deux manières : d'une part par rapport aux tests nucléaires passés, d'autre part par rapport à des expériences de physique. Ces dernières sont effectuées à l'aide de deux grands outils expérimentaux du CEA/DAM que sont la machine radiographique AIRIX et le futur laser Mégajoule.

La qualité des simulations numériques repose tout à la fois sur la modélisation fine des phénomènes physiques concourant au fonctionnement d'une arme, la prise en compte de leurs interactions, la description détaillée des objets impliquant le passage à des calculs tridimensionnels. Ces exigences impliquent une augmentation importante des capacités de calcul et l'amélioration des algorithmes numériques. Si l'on veut conserver des temps de réponse raisonnables pour un calcul (résultat le lendemain), un accroissement d'un facteur 10000 sur la capacité de production actuellement disponible est nécessaire. Cela se traduit par une puissance soutenue de 100 Tflops.

Corrélativement, un accroissement équivalent sera nécessaire en capacité mémoire (100 à 200 To) et en moyens de stockage (100 Po).

Enfin, la manipulation et l'interprétation de l'énorme quantité de résultats produite par la simulation suppose des moyens dimensionnés en conséquence, en particulier : des logiciels adaptés, des réseaux rapides, des stations de travail performantes et des équipements avancés de visualisation.

L'ensemble de ces considérations ont conduit à mettre en place le projet TERA.

Le projet TERA

L'objectif de ce projet est de fournir au Programme Simulation les équipements de simulation numérique dont il a besoin. Le projet TERA comprend, en ce qui concerne les calculateurs, trois phases principales : 1 Tflops soutenu en 2001, 10 Tflops soutenus en 2005, 100 Tflops soutenus en 2009. Il couvre quatre domaines étroitement liés : les supercalculateurs, les systèmes de stockage, l'infrastructure nécessaire à l'installation de ces équipements et enfin le portage des codes et logiciels de simulation existants d'une plate-forme à la suivante.

L'architecture retenue pour les supercalculateurs est de type "Cluster de SMP". Cette architecture, adéquate aux besoins du projet TERA, bénéficie d'un fort soutien de l'industrie informatique. Ainsi, sa pérennité devrait permettre de conserver la même architecture pendant toute la durée du projet. C'est sur cette architecture unique que seront développés les futurs systèmes de codes qui assureront la pérennité et la garantie des armes nucléaires en 2009.

Le choix de l'architecture "Cluster de SMP"

L'architecture "Cluster de SMP", choisie dès 1996, utilise un parallélisme à deux niveaux :

- un premier niveau à mémoire partagée (SMP), pouvant être programmé en particulier à l'aide de directives (OpenMP ...),
- un deuxième niveau en réseau rapide, utilisant le modèle de programmation pas échange de messages.

Le choix de l'architecture "Cluster de SMP", osé pour certains en 1996, avait été, dès le départ complété par une condition technique restrictive : il ne serait considéré comme validé que si l'on obtenait sur un microprocesseur de SMP, pour nos propres codes de production, une puissance au moins égale à 1/10 de celle d'un processeur du CRAY T90.

Ceci a conduit à acheter en 1998 deux machines de test, qui ont chacune fait l'objet en 1999 d'une mise à jour avec les dernières technologies disponibles :

- l'une IBM, avec 5 noeuds comprenant chacun 8 microprocesseurs Power3 à 255 MHz et interconnectés par un switch SPS (réseau interne des IBM/SP),
- l'autre COMPAQ (initialement DIGITAL) avec 4 noeuds comprenant chacun 4 microprocesseurs EV67 à 667 MHz et interconnectés par un réseau développé par la société européenne QUADRICS.

Le portage de l'ensemble de l'environnement logiciel, réalisé entièrement en juin 1998, a permis de démarrer celui des codes de production avec des premiers résultats très encourageants.

Lors du lancement de l'appel d'offre pour la première machine de production T1 (1 Tflops soutenu) mi 1999, toutes les extrapolations annonçaient pour fin 2000 des ratios de l'ordre de 1/2 à 1 équivalent processeur T90 par microprocesseur et plus personne ne doutait de la validité du choix "Cluster de SMP". Aujourd'hui, l'ensemble du portage des codes de production et des logiciels d'environnement associés est terminé. La parallélisation de certains d'entre eux reste néanmoins à faire.

Le lancement de la première phase du projet TERA

Une fois acquise la validation de l'architecture de type "Cluster de SMP", il était possible de lancer, début 1999, la première phase opérationnelle du projet, décomposée en quatre sous-tâches :

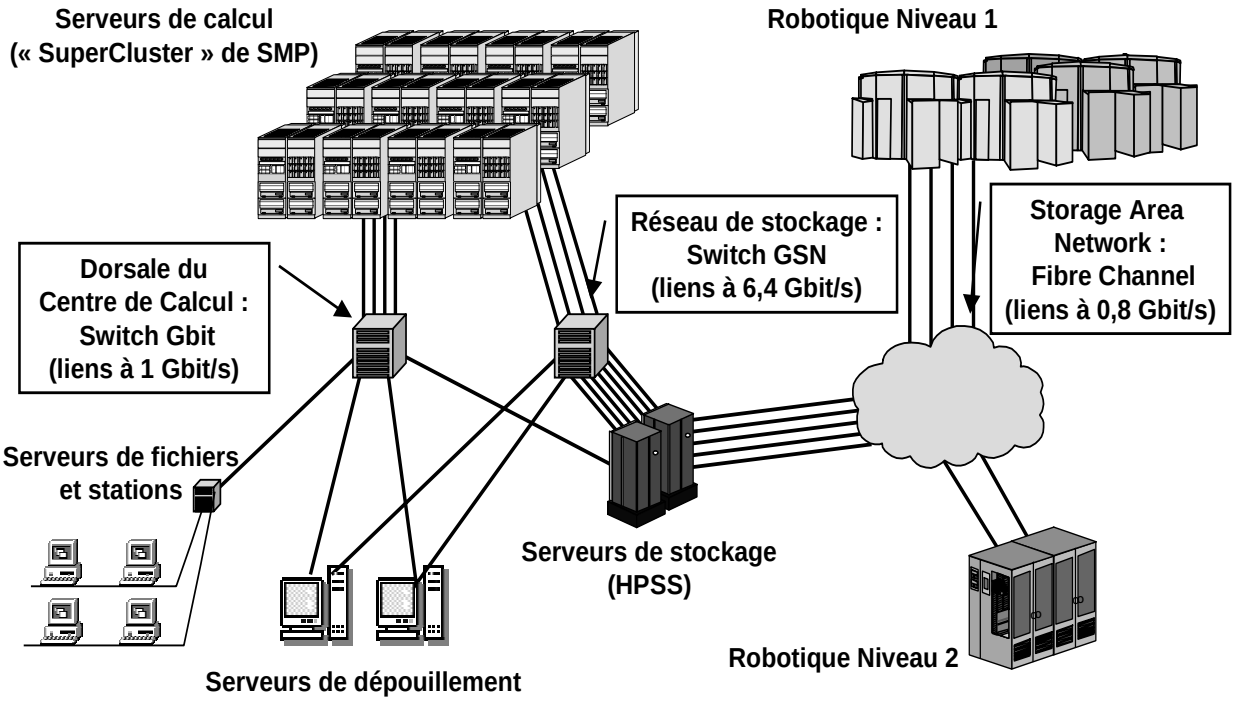
1. La réalisation des benchmarks pour l'appel d'offre.
2. La définition de l'architecture de stockage, l'achat et l'implémentation des matériels.
3. La réalisation des infrastructures nécessaires.
4. Enfin, la rédaction du cahier des charges, le lancement de l'appel d'offre, l'analyse des réponses, la proposition du meilleur choix et l'installation d'un serveur de calcul opérationnel début 2002.

Nouvelle architecture du centre de calcul

Le passage d'une architecture intégrée, systèmes CRAY assurant à la fois le service calcul et le service stockage des données (via le logiciel DMF), à une architecture ouverte séparant calcul et stockage a été étudié et validé en parallèlle avec l'expérimentation "Cluster de SMP". Elle conduit à une nouvelle architecture du Centre de Calcul, qui est illustrée dans la figure de la page suivante.

Cette architecture est caractérisée par une totale symétrie entre la partie "serveur de calcul" et la partie "serveur de données", basée sur l'utilisation du logiciel HPSS. Compte tenu de l'énorme quantité de données à stocker, deux niveaux de stockage bande ont été prévus, le premier supportant une année de production sur des médias rapides, le second contenant l'archivage de l'ensemble des données sur des médias plus lents mais possédant des capacités plus importantes.

Future architecture du centre de calcul



Infrastructures

Enfin, on ne saurait éviter de parler des infrastructures requises pour installer de tels équipements. Les “Clusters de SMP” mettent en oeuvre des technologies électroniques peu intégrées refroidies à l’air. Cela se traduit par un encombrement au sol, une puissance électrique consommée et une dissipation de chaleur dans l’ambiance, forts différents de ceux connus à l’époque des machines CRAY.

Notre machine 1 Tflops occupera une surface au sol de 750 m^2 (extensible à 1250 m^2 pour les étapes suivantes), demandera des faux planchers à 1,2 m pour faire passer les nombreux câbles d'alimentation et d'interconnexion, consommera entre 2 et 4 Mégawatts, naturellement maintenus et qu'il faudra évacuer dans l'air ... Il s'est donc avéré indispensable de lancer la construction d'une nouvelle salle machine respectant toutes les contraintes en matière de sécurité des personnes, des données et des biens.

La photo de la page précédente montre une vue du chantier sur le site de DAM-île de France à Bruyères-le-Châtel. L'excavation de 9m de profondeur a la surface d'un terrain de football (60m x 60m). Pour l'anecdote, cette excavation correspond au déblaiement de 40000 m^3, soit 4000 camions de 10m^3 pendant 3 semaines à raison d'un camion toutes les 4 mn !

Choix de la machine TERA

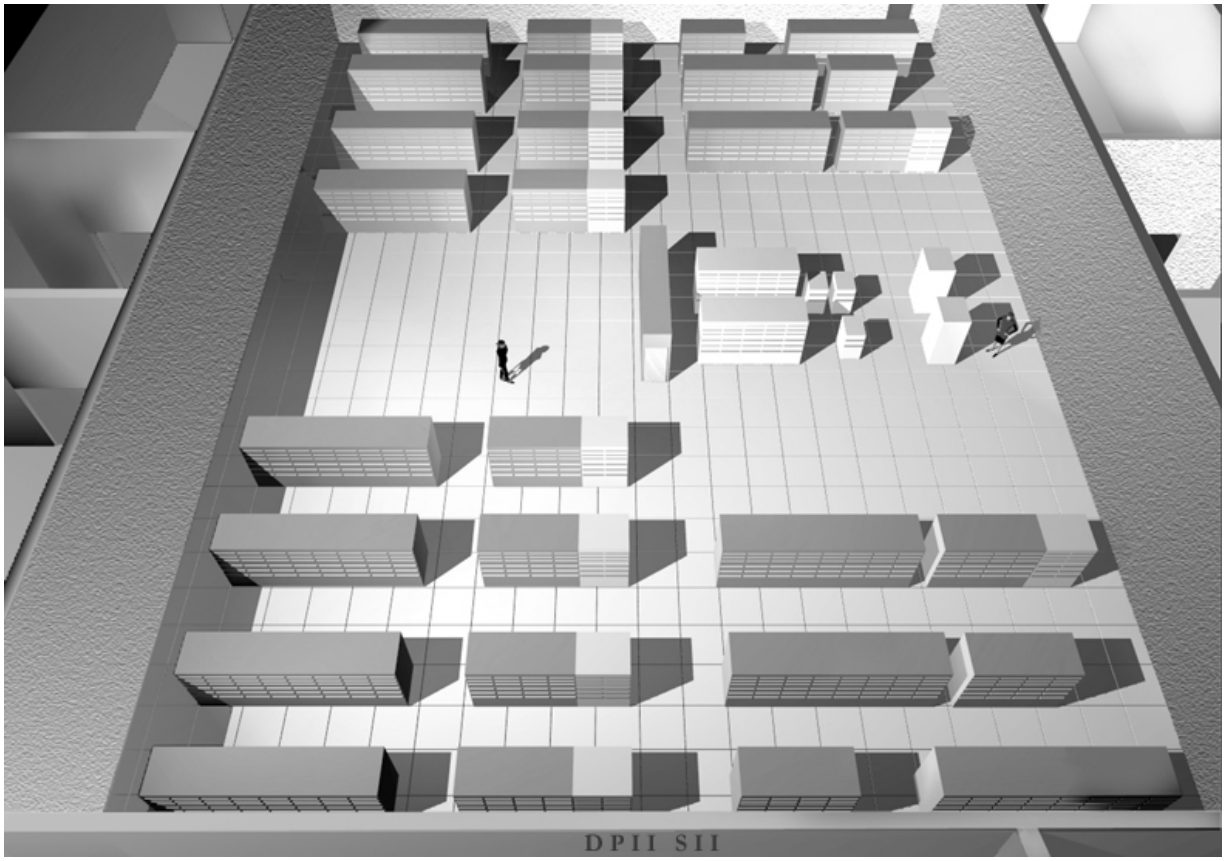
Afin de préparer, rédiger le cahier des charges, développer les benchmarks, lancer l'appel d'offre, dépouiller les résultats, adapter les offres aux besoins et contraintes du CEA/DAM et finalement proposer un choix à la Maîtrise d'Ouvrage, le CEA/DAM a mis en place, pendant une période d'un an, une équipe de 8 personnes.

Dans le but de simplifier le travail de préparation des réponses des constructeurs, d'obtenir des engagements clairs et précis et de simplifier le dépouillement de l'appel d'offre, une grille de 238 critères requis (comportant 76 résultats de benchmarks), de 60 critères souhaités et de 34 questions a été établie. La méthode de notation permettant d'aboutir au choix final a été également définie au préalable.

Ceci a permis de conduire cette opération dans les délais. Ainsi, le Directeur des Applications Militaires a pu signer le contrat avec COMPAQ le 21 février 2000.

Le choix de COMPAQ, issu des recommandations de l'équipe technique de dépouillement a été étudié par la Maîtrise d'Ouvrage. Celle-ci a, en outre, tenu compte dans sa décision des différents risques (techniques, économiques et commerciaux) attenants. Différentes dispositions ont été prises au niveau contractuel pour limiter, autant que possible, les conséquences de ces risques.

Le schéma ci-dessous est une représentation virtuelle de la salle machine. Celle-ci sera en réalité divisée en trois zones pour des raisons de sécurité incendie.



La machine COMPAQ retenue constitue :

- Une étape vers la solution permettant d'atteindre les 100 Tflops en 2009. COMPAQ intègre en effet les développements permettant de palier les défauts d'extensibilité constatés sur d'autres systèmes. En particulier, les excellentes valeurs prévues pour les débits mémoire, les débits et latences réseau, permettent d'envisager de bons rendements sur des simulations de grandes tailles.
- Une machine de production assurant les calculs requis par les programmes armes et qui seront partiellement effectués à partir des codes 2D vectoriels portés sur un processeur. COMPAQ, par la puissance de son processeur Alpha, est de loin le meilleur pour répondre à cet objectif.

Trois machines (initiale, intermédiaire et finale) sont prévues dans le cadre du contrat, respectivement en avril 2000, fin 2000 et fin 2001.

1. Une machine initiale de 35 Gflops
2. Une machine intermédiaire de 0,5 Tflops formée de 74 noeuds ES40 de 4 processeurs Alpha EV68 reliés par un réseau QUADRICS et comportant 5 unités de stockage
3. La machine finale, sur laquelle sera démontré le Tflops soutenu, d'au moins 5 Tflops crête, disposant d'une mémoire de 2,5 To, constituée de 42 serveurs de 64 processeurs EV7 reliés par un double réseau d'interconnexion QUADRICS. Elle comportera 56 unités de stockage, reliés par un SAN, pour un espace disque utile supérieur à 50 To.

Un certain nombre d'options : doublement mémoire, augmentation du nombre de plans d'interconnexion, 1 Tflops supplémentaire ont été chiffrées et seront éventuellement levées après la recette de la machine intermédiaire.

Conclusions

Pour relever le défi de la Simulation, après la décision prise par la France de l'arrêt définitif et total des essais, la Direction des Applications Militaires du Commissariat à l'Energie Atomique a décidé de se doter du plus puissant complexe de Simulation Numérique jamais réalisé en Europe.

La première étape de ce Projet est maintennat dans sa phase opérationnelle. Elle se traduira par la mise en service, fin 2001, d'une machine permettant à nos applications d'atteindre et dépasser le Tflops. Un environnement matériel et logiciel cohérent avec ce niveau de puissance permettra d'en tirer le meilleur parti en production.

Cet investissement considérable de la Nation, à côté de l'appareil de radiographie AIRIX et du futur laser Mégajoule, démontre tout à la fois l'engagement de la France pour la dissuasion nucléaire et le souhait du CEA/DAM de jouer un rôle majeur au sein de la communauté HPC.

On notera enfin, comme l'annonçait le Directeur des Applications Militaires, Jacques Bouchard, le 23 février dernier, que la machine TERA pourra, en complément de son usage principal et à l'instar du laser Mégajoule, être mise à la disposition de la Communauté Scientifique.

Data Mining

Super Select™, un moteur de requêtes de Hautes Performances

Bernard Nivelet, Bull

L'informatique décisionnelle rencontre des difficultés pour mettre en oeuvre le "Data Mining" sur des données stockées dans le "Data Warehouse", car la puissance informatique nécessaire à un traitement performant des requêtes a considérablement cru avec les augmentations conjuguées des volumes de données et de la complexité des stratégies à mettre en oeuvre. C'est une préoccupation majeure des organisations qui placent le client au centre de leur stratégie.

Les requêtes statistiques ont ceci de particulier qu'elles exigent le balayage de l'ensemble des tables sur lesquelles elles portent, et que le nombre de lignes de ces tables se compte (au moins) en millions. Il est clair que les mécanismes standard des moteurs relationnels, conçus pour l'accès à des champs, sont inadaptés pour les exécuter dans des temps raisonnables. Pour espérer atteindre les plus hautes performances, il faut être capable de travailler sur les colonnes après une transformation leur donnant la représentation interne très régulière des tableaux numériques : alors les processeurs vectoriels pourront délivrer toute leur puissance.

Or, comme les statistiques sont effectuées sur des données historiques, la période de rafraîchissement est au moins la journée. Il en résulte que les valeurs contenues dans les tables constituent un ensemble stable pendant la période d'exploitation et il est possible de définir un codage stable des données transformant ces valeurs en nombres : les colonnes numériques sont conservées telles quelles, les colonnes alphanumériques sont codées en remplaçant

chaque prédicat par son rang dans la nomenclature ordonnée des valeurs présentes dans la colonne. Ce codage, évidemment bijectif, respecte l'ordre : il permet de traduire les requêtes en transposant les comparaisons sur critères usuels et de décoder chaque résultat pour le fournir en clair.

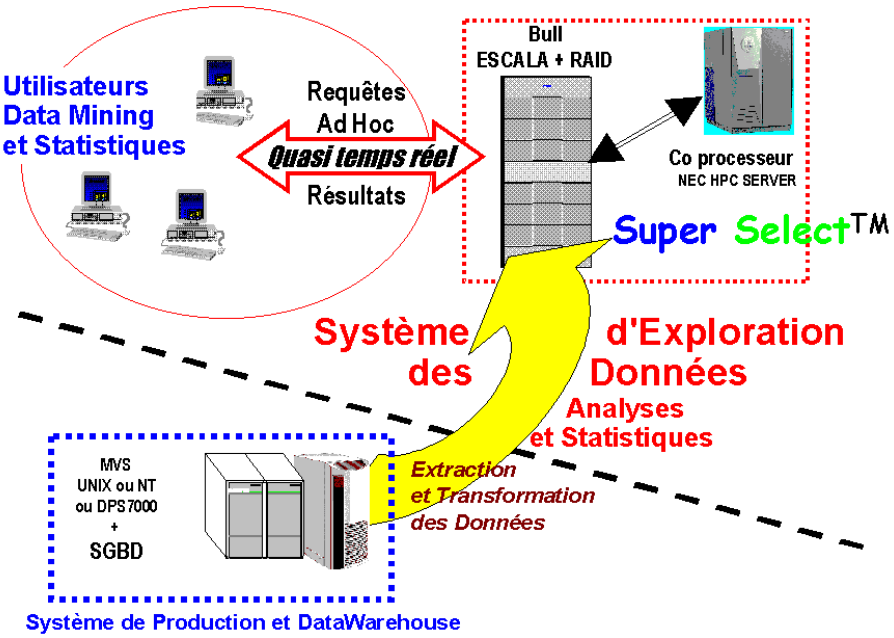
Ce procédé présente par ailleurs quelques avantages importants. D'abord, contrairement aux moteurs multidimensionnels, le codage est indépendant des axes de recherche, ce qui garantit les mêmes performances quelques soient les champs présents dans la base ; ensuite, il compresse les données, ce qui limite les besoins en mémoire ; enfin, le procédé permet de réaliser d'une façon naturelle une fonction très utile aux statisticiens et absente de SQL : définir une colonne virtuelle de la base construite avec des valeurs prédéfinies lorsque certains critères sont satisfaits : c'est la fonction DEFINIR, proposée en extension à SQL.

Cette technologie, brevetée par Bull, est mise en oeuvre dans Super Select™, qui utilise les Serveurs NEC de haute performance (dérivés des supercalculateurs vectoriels) comme coprocesseurs derrière un système Bull Escala sous AIX.

Les performances sont spectaculaires, comme le montrent les exemples ci-dessous :

- La base de 2,5 Go ayant servi aux tests compte neuf tables jointes sur deux niveaux et la table de faits comporte plus de 7 000 000 de lignes. Son codage a pris 55 minutes.
- La première requête exécute un WHERE portant sur trois colonnes, deux DEFINIR et un GROUP BY; il en a résulté la sélection de 12 711 lignes et le calcul de 209 classes en 2,7 secondes.
- La deuxième requête exécute un WHERE portant sur deux colonnes, un DEFINIR et un GROUP BY ; il en a résulté la sélection de 1 172 216 lignes et le calcul de 1823 classes en 7,9 secondes.
- Une mesure comparée avec Oracle 8.0.5 (la base étant entièrement en mémoire, situation exceptionnelle) sur EPC 2400 avec 12 processeurs PowerPC 360 MHz a mis en évidence un rapport 59 en faveur de Super Select™ équipé d'un seul processeur SX-5S.

Une solution Super Select™ du type de celle représentée ci-dessous est proposée par Bull à un prix d'entrée de l'ordre de 900 KEuro.



Les Applications du Calcul Parallèle

Ce numéro spécial de la revue **Calculateurs Parallèles** (volume 11 - n°3, 2000) a été réalisé sous la direction de Bernard Philippe (IRISA) avec le concours du bureau et du conseil scientifique d'ORAP.

Il comprend trois parties : Applications du calcul parallèle ; Algorithmes parallèles ; Outils pour le programmation parallèle.

Cet ouvrage de 168 pages est édité par Hermès (ISBN 2-7462-0141-0).

Actualités Bi-Orap

➡ Textes des tutorials sur les clusters et le “Grid”

Les textes des “tutorials” donnés dans le cadre de la Conférence de Mannheim (13 et 14 juin 2000) sont disponibles sur le Web :

http://www.supercomp.de/programm/tutorium/13_06_00.htm
http://www.supercomp.de/programm/tutorium/14_06_00.htm

➡ Compaq

- Compaq et le Pittsburgh Supercomputing Center vont collaborer pour mettre à la disposition de la communauté scientifique, dès 2001, un ordinateur d’une puissance de l’ordre de 6 Tflops. Ce système sera basé sur 2728 processeurs Alpha. La NSF a dégagé un budget de 45 M$ pour ce projet.

  http://www.psc.edu/publicinfo/tcs

- Compaq a livré 1300 ordinateurs au Sandia National Lab pour compléter les clusters Linux dans le cadre du programme Cplant (ce programme “Computational Plant” avait été présenté au cours du 8ème Forum ORAP). Sandia dispose mainenant de 2600 ordinateurs Compaq regroupés en plusieurs clusters.

  http://www.cs.sandia.gov/cplant

- Le Département de l’Energie (DOE) a retenu Compaq pour construire la prochaine machine ASCI : “ASCI Q”. La performance crête de ce système, destiné au laboratoire de Los Alamos et qui doit être opérationnel en 2002, sera de 30 Tflops. Une évolution vers un système de 100 Tflops (processeurs Alpha EV7 et EV8) pourrait être réalisée en 2004.

➡ Cray

- Phillips Petroleum a acheté un système Cray T3E-1350 ayant 136 processeurs cadensés à 675 MHz.

➡ Hewlett Packard

- L’Université du Kentucky remplace son superordinateur HP Exemplar SPP par un cluster de serveurs HP composé de 12 serveurs N-4000. Chacun de ces noeuds dispose de 8 processeurs PA-8500 cadensés à 440 MHz. Ce cluster se situe environ en 200ème position dans le Top500.

  http://hpc.uky.edu

- HP a annoncé la gamme de serveurs Unix Superdome. Un serveur peut avoir 16, 32 ou 64 processeurs PA-8600 cadensés à 550 MHz. La mise en cluster permet de disposer d’un nombre maximum de 256 processeurs.

http://www.unixservers.hp.com/highend/superdome/

➡ IBM

- Le centre de recherche sur le climat de Postdam a commandé un système RS/6000 SP. Cette machine évoluera en 2002 vers le Tflops, grâce à la technologie Power4.
- L’Université de Floride (FSU) a acheté un IBM SP/6000 SP avec 680 processeurs. Sa puissance crête est de 2,5 Tflops

➡ NEC

- L’Université d’Osaka a commandé un SX-5/128M8 d’une puissance théorique supérieure à 1,2 Tflops..

➡ SGI

- NASA Ames a commandé deux systèmes 3800 (architecture NUMAflex) de 512 processeurs. Ces systèmes seront reliés pour former un ordinateur de 1024 processeurs à mémoire partagée.
- L’Université des Sciences et Technologie de Norvège a commandé à SGI deux systèmes Origin2000 avec des évolutions prévues en 2001 et 2002 pour atteindre une puissance crête de 1 Tflops.
- L’US Navy a commandé deux systèmes de la nouvelle famille 3000 : un système 3800 avec 128 processeurs et un système 3800 avec 512 processeurs. De son côté, ARL (US Army Research Lab) a commandé deux systèmes de cette même famille : l’un avec 512 l’autre avec 256 processeurs.

➡ Formation à MPI sur le Web

Le NCSA (National Computational Science Alliance) aux Etats-Unis propose, sur le Web, des formations pour le calcul de haute performance. Une formation à MPI (Message Passing Interface) vient d’être mise en place.

Ces cours utilisent WebCT, un outil développé par l’Université de British Columbia.

http://webct.ncsa.uiuc.edu:8900/

➡ Grid Physics Network

Les universités de Floride et de Chicago vont

piloter un ambitieux projet financé à hauteur de 11,9 millions de dollars par la NSF dans le cadre de son nouveau programme "*Information Technology Research*". Il s'agit d'étudier la mise en place des bases d'une "grille" ("GRID") d'une puissance sans précédent. L'objectif est de donner aux scientifiques un outil pour interpréter des volumes considérables de données qui seront générés par les grandes expérimentations de physique et d'astronomie à travers le monde.

GriPhyN implique une douzaine d'institutions aux Etats-Unis. Les ressources mises à disposition des chercheurs apparaîtront comme un unique système de calcul et de stockage d'informations.

La mise en oeuvre complète du projet demanderait environ 70 millions de dollars.

Le CERN (Genève) doit participer à ce projet.

http://www.griphyn.org

➡ **EPCC propose un "programme visiteurs"**

Dans le cadre du programme européen TRACS, le centre de calcul parallèle d'Edinburgh (EPCC) accueille des chercheurs pour participer à des projets liés au calcul de haute performance.

http://www.epcc.ed.ac.uk/tracs/

➡ **l'EPFL inaugure le Swiss-T1**

L'Ecole Polytechnique Fédérale de Lausanne a inauguré son superordinateur Swiss-T1 développé par l'EPFL en collaboration avec Compaq et la société suisse SCS (Supercomputing Systems). Cette machine, qui comprend 70 processeurs Alpha, avait été présentée lors du 8ème Forum ORAP.

http://capawww.epfl.ch

Agenda

- 11 au 14 octobre : **SGI 2000** : First Worldwide SGI User's Conference (Krakow, Pologne)
- 12 au 13 octobre : **EuroPAM 2000** : International Conference and Exhibition on Virtual Prototyping by Numerical Simulation (Nantes)
- 12 au 14 octobre : 3rd **Extreme Linux Workshop** (Atlanta, Etats-Unis)
- 15 au 19 octobre : **PACT'2000** : International Conference on Parallel Architectures and Compilation Techniques (Philadelphie, Etats-Unis)
- 16 au 18 octobre : **SRDS 2000** : 19th IEEE Symposium on Reliable Distributed Systems (Nurnbert, Allemagne)
- 19 octobre : Simulation et Petaflops (IPN, Orsay)
- 19 au 20 octobre : **HUG** : Fourth Annual High Performance Fortran User Group (Tokyo, Japon)
- 26 octobre : journée CINES-IDRIS (Montpellier)
- 30 octobre au 3 novembre : **NGN'2000** : Next Generation Networks (Washington, Etats-Unis)
- 4 au 10 novembre : **SC 2000** : Supercomputing 2000 (Dallas, Tx, Etats-Unis)
- 28 novembre au 2 décembre : **Cluster 2000** : International Conference on Cluster Computing (Chemnitz, Allemagne)
- 7 au 9 février 20001 : **Euro Micro-PDP2001** : 9th Euromicro Workshop on Parallel and Distributed Processing (Mantova, Italie)

Des informations complémentaires, en particulier les adresses http de ces manifestations, sont disponibles sur le serveur WWW d'ORAP. Contactez le secrétariat d'ORAP si vous ne disposez pas de l'accès vers le serveur Web.

Appel à informations

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

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

HOISE - *Europe On-line Information Service*

PRIMEUR ! - *Advancing European Technology Frontiers*

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

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

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