Visite et autopsie de Tera 10, le méga supercalculateur européen

Le centre de la

Direction des applications militaires du CEA de Bruyères-le-Chatel héberge le complexe de calcul scientifique, qui constitue le principal pôle de simulation de la recherche militaire française. On y développe en particulier les simulations de l'arme nucléaire. Ne cherchez pas le bâtiment des supercalculateurs Tera, il est enterré sous un terre-plein qui était auparavant un terrain de foot. Tera 10 a nécessité une infrastructure gigantesque: elle occupe une surface au sol de 800 m². Les structures sont réparties sur plusieurs niveaux. 2.000 m² sont consacrés aux machines, serveurs, stockage et visualisation, et 2.000 m² aux servitudes. La consommation électrique est proche des 2 mégawatts ! Passons sur les alignements de serveurs – 270 baies – les 90 kilomètres de câbles, les 12 silos de stockages sur bandes, la salle de visualisation et son mur d'images de 5×3 mètres en 14 millions de pixels. C'est le système de refroidissement, chargé de maintenir une température constante de 23 degrés, qui impressionne. Le CEA a fait le choix d'un système basé sur une centrale à eau glacée. L'eau circule dans le faux plancher d'une hauteur de 1,2 mètre. Elle permet de produire de l'air froid pulsé vers la face avant des alignements de serveurs. La reprise de l'air chaud s'effectue par extraction à l'arrière des serveurs vers le faux plafond. Enfin, le CEA a encore fait un choix original pour la sécurité incendie. Pas d'émission de gaz, comme c'est le cas classiquement dans une salle blanche. La méthode est jugée trop agressive, le gaz sensé éteindre l'incendie est agressif sur l'humain et corrosif sur le matériel. Le CEA a choisi la brumisation, la projection de minuscules gouttes d'eau en pluie. Pas de danger pour l'humain et un risque de dégradation matérielle limité. La détection des excès de chaleur s'effectue au plafond, mais le système de brumisation ne se déclenchera que sur la zone en danger, et pas sur l'ensemble des salles. **L'autopsie du Tera 10 à découvrir ci dessous, après les photos.**

Autopsie du Tera 10: 9.000 processeurs Itanium 2…

Tera 10 est le puissant supercalculateur européen jamais construit. Il regroupe 602 serveurs, 9000 processeurs Itanium 2 'Montecito', 30 tera-octets en RAM et 1 peta-octets d'espaces disque. Avec l'abandon des essais en réel des bombes nucléaires, la France s'est dotée du plus puissant supercalculateur d'Europe (il devrait occuper la quatrième place au Top 500 mondial et la première pour les calculateurs sous Linux), principalement destiné à la simulation pour la sûreté et la fiabilité de l'arme nucléaire. Bull, unique constructeur européen, a remporté l'appel d'offre. Tera 10 est un puissant **cluster de 602 serveurs Bull NovaScale**. 544 serveurs 'n?uds' de calcul disposent chacun de 16 processeurs Intel Itanium 2, répartis dans 270 baies. 56 serveurs sont consacrés aux entrées-sorties et 2 serveurs à l'administration. Le **processeur Intel Itanium 2 'Montecito'** fait son apparition pour la première fois en application. Il n'est pas encore commercialisé par le fondeur ! Sur le calcul, 4352 processeurs dual core fournissent 8704 c?urs. Pour une puissance de calcul de plus de 50 Teraflops (50.000 milliards d'opérations par seconde). Un processeur Montecito réunit 1,7 milliard de transistors. Cette composition technologique fait appel au concept des 'composants sur étagères', des Cots ou '*component off the shelf*). Une stratégie qui abandonne les processeurs spécifiques et spécialisés ? le CEA était jusqu'à présent équipé de calculateurs Cray ? au profit de

composants standards fabriqués en très grande série, sans doute moins performants, mais 100 fois moins chers. Pour répondre aux contraintes imposées par l'appel d'offre du CEA, Bull a adapté l'architecture FAME de ses serveurs NovaScale. Les échanges entre les blocs de processeurs mémoire (QBB) sont assurés via le **chip FSS** (*Fame Scalability Switch*), développé par le constructeur. FSS assure la cohérence globale de la mémoire et des caches, et synchronise l'ensemble des échanges. Chaque serveur comporte plusieurs FSS qui lui permettent d'obtenir, outre des débits importants, une très grande bande passante répondant aux besoins d'entrées-sorties. La **mémoire centrale** est de 30 tera octets distribués, à raison de 48 à 128 Go par n?ud. Pour le CEA, c'est la puissance de calcul qui est la priorité, pas le volume de mémoire RAM disponible. La partie **réseau** a été confiée à la société anglo-italienne Quadrics. Ce choix d'une technologie haut débit répond à deux objectifs du CEA : un temps de latence de 3µs (micro seconde) et une capacité d'échange de 650 Go/s, ou la valeur de 160 millions de pages de texte à la seconde. La **bande passante** est de 100 Go/s, soit l'équivalent de 200.000 films en streaming vidéo diffusés en simultané. Indispensable, le **stockage** atteint le Péta octet, ou 1 million de milliards d'octets directement accessibles, sur 7800 disques SATA. Cette capacité est équivalente à 30 fois la capacité de la *Très Grande Bibliothèque*, ou 250 milliards de pages de texte, ou encore 250.000 films au format DVD. La multiplication des disques n'est pas seulement liée à la capacité de stockage. En exploitant simultanément plusieurs disques, elle permet d'obtenir un débit supérieur à celui d'un seul disque. Sur une simulation où la lecture et l'écriture des données peut atteindre 100 Go par seconde avec Tera 10, le débit d'un disque ne dépasse pas en revanche les 40 Mo/s ! L'autonomie de Tera 10 en cas d'accident ou de coupure est de 15 minutes, à peine le temps de sauvegarder ses données par un backup sur bandes. Enfin, le choix du CEA s'est porté sur le logiciel libre. Un choix dicté tout d'abord par l'ouverture, pour partager les développements avec d'autres laboratoires. Mais aussi par la pérennité avec Linux, disponible sur un grand nombre de plateformes et avec une durée de vie qui s'annonce longue. Sans oublier les coûts réduits. Bull s'est engagé lui aussi sur l'open source. Il a modifié le code du noyau de Linux à partir d'une distribution Red Hat, afin de l'adapter à sa technologie HPC. Il a adopté une bibliothèque de communication et le système de fichier global et parallèle Lustre. Et enfin développé un 'cluster management' pour simplifier la gestion de l'ensemble. L'ensemble reste open source. Tera 10 pourra traiter quotidiennement entre 10 et 30 tera octets de données? Cette puissance a un prix, 50 millions d'euros. Auxquels il faudra ajouter le coût de 60 ingénieurs pour la maintenance du système, 100 ingénieurs pour les développements et 800 ingénieurs pour son utilisation.