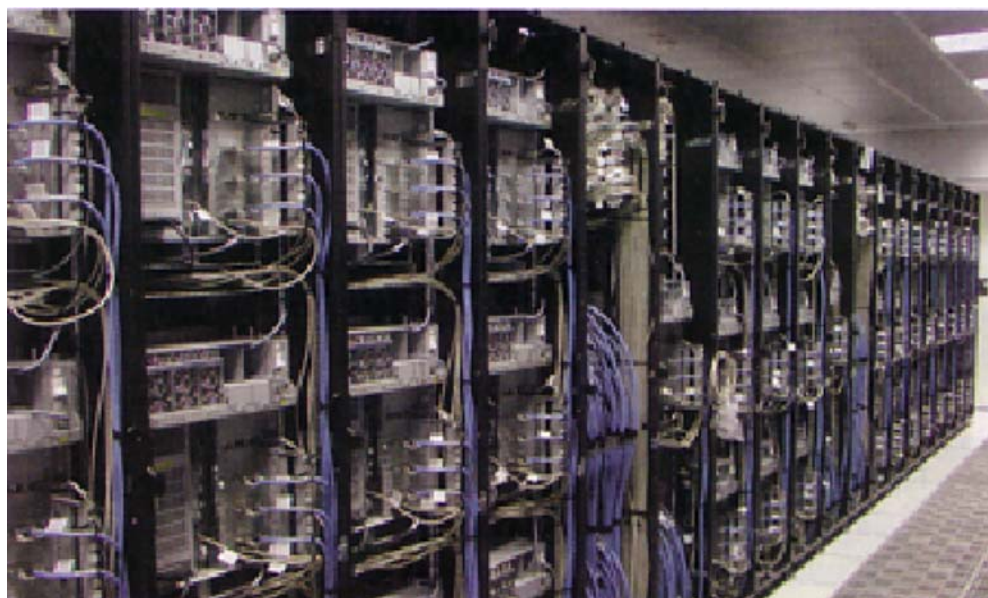


# Au delà de 50 mille milliards d'opérations par seconde !

**Jean Gonnord,**  
chef du projet simulation numérique et informatique au CEA/DAM Ile de France.

**Pierre Leca,**  
chef du département sciences de la simulation et de l'information au CEA-DAM Ile de France.

**François Robin,**  
adjoint au chef du département sciences de la simulation et de l'information au CEA-DAM Ile de France et responsable opérationnel du projet TERA-10.



**Avec plus de 50 mille milliards d'opérations par seconde, Tera-10 sera le plus puissant ordinateur d'Europe. Portrait du tout nouveau supercalculateur construit par Bull pour la direction des applications militaires du commissariat à l'énergie atomique**

Depuis 1998, date de la ratification par le Parlement du Traité d'interdiction complète des essais nucléaires, les nouvelles armes nucléaires françaises ne sont plus testées. Celles-ci évoluent pourtant... Pour s'assurer de leur pertinence et de leur sécurité, la simulation numérique a pris le relais des essais.

Les modèles prédictifs développés par les chercheurs de la direction des applications militaires (DAM) du CEA sont fondés sur des équations non linéaires dont la résolution passe par le calcul itératif sur des mailles, dont la finesse conditionne la fiabilité des prédictions.

En d'autres termes, ces modèles nécessitent de puissants ordinateurs. Les besoins en puissance de calcul, poussés également par l'amélioration des modèles, croissent à une vitesse vertigineuse : en gros, d'un facteur dix tous les quatre ans.

Cette cadence soutenue a conduit à installer un nouveau supercalculateur tous les quatre ans. Tera-1, installé par HP fin 2001, va ainsi laisser la place, d'ici la fin 2005, à Tera-10.

Cette machine, construite par Bull, devrait se hisser au niveau des premiers du Top 500 des supercalculateurs mondiaux.

## **Vectorielle ou parallèle ?**

Comme son prédécesseur, Tera-10 adopte une architecture en grappe de multiprocesseurs à mémoire partagée, ou « cluster de SMP\* ». Les processeurs, ces unités de base des calculateurs qui effectuent les calculs, sont en effet de deux types : scalaires et vectoriels. Les premiers exécutent des opérations simples portant sur des scalaires, des grandeurs entièrement définies par leur mesure, l'addition de deux nombres par exemple. Les seconds réalisent des opérations portant sur des vecteurs, grandeurs qui ne peuvent être définies que par plusieurs entités mathématiques.



**LES AUTEURS**  
devant une rangée  
d'armoires de la  
machine Tera.

© CEA

L'addition de deux vecteurs de cinq cents éléments est l'une d'elles.

Les processeurs vectoriels sont particulièrement adaptés aux calculs réguliers, très fréquents en simulation numérique : lors de l'exécution d'une opération de ce type, un processeur vectoriel peut fonctionner à une vitesse proche de sa performance maximale (dite de « crête\* »). La même opération avec un processeur scalaire exige en revanche de nombreuses opérations indépendantes (par composante des vecteurs) et ce, à une vitesse bien inférieure à la vitesse crête.

\* SMP, shared memory multiprocessor.

\* Puissance de crête, puissance maximale que le calculateur peut fournir pendant un bref instant.

Trois architectures permettent de mettre en parallèle des processeurs : les supercalculateurs vectoriels, les grappes de processeurs scalaires à mémoire partagée et les grappes de PC (l'ordinateur que chacun possède chez soi).

Peu coûteuse, cette dernière architecture est mal adaptée à des environnements où de nombreux utilisateurs font beaucoup de calculs très

gourmands en puissance machine et en mémoire.

Les supercalculateurs vectoriels s'avèrent très onéreux mais plus performants (sur du calcul scientifique, ils affichent des rendements pouvant dépasser 50 % des performances crêtes sur certaines applications scientifiques).

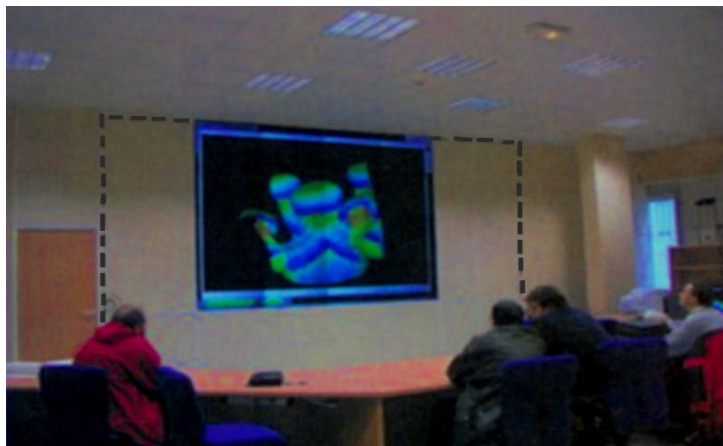
Les grappes de processeurs scalaires, processeurs qui équipent les machines grand public, sont certes moins performantes (5 % à 20 % des performances de crêtes), mais nettement moins chères. Elles offrent ainsi, et de loin, le meilleur rapport prix/performance pour les applications du CEA/DAM.

C'est la raison pour laquelle, dès Tera-1, nous avons délaissé les ordinateurs scientifiques classiques que nous utilisions jusque-là (des machines vectorielles CRAY) et opté pour une architecture de type cluster de SMP. L'unité de base du calculateur est constituée d'un ensemble de seize processeurs partageant une même mémoire. Ces unités (on parle de « nœuds ») sont assemblées en parallèle,

chacune effectuant une partie du calcul. Les nœuds sont interconnectés et partagent un espace disque commun. Certains d'entre eux (les contrôleurs d'entrée/sortie) sont dédiés au pilotage de l'espace disque. Ils permettent aux nœuds de calcul d'accéder à l'espace disque pour y déposer les résultats de leurs calculs ou y prélever les conditions initiales nécessaires aux calculs suivants.

L'expérience du CEA sur Tera-1 a permis d'identifier les forces et les limites de ce type de supercalculateurs. La première question était de savoir si cette architecture proposait une puissance de calcul suffisante. C'est le cas, à condition de disposer d'applications bien parallélisées. Quand une application tourne sur des centaines voire des milliers de processeurs, elle doit en effet être découpée en autant de petites tranches de calcul qu'il y a de processeurs.

**AVEC TERA-10, un mur d'image de 5,5 sur 3 mètres (en pointillé) va remplacer l'écran actuel.** © CEA



Ces tranches communiquent régulièrement pour que les résultats aux limites se recoupent, ce qui mange du temps de calcul. Cet échange doit donc être optimisé. Par ailleurs, l'application est fragmentée de manière à ce que chaque calcul sur chaque tranche se déroule sur une durée quasi identique.

Aucune tranche ne doit en effet freiner l'ensemble du calcul. Quand l'écriture des applications répond à ces critères de parallélisation, la performance des clusters de SMP est effectivement au rendez-vous.

Sa performance globale dépend également de la capacité de remplissage du supercalculateur.

Le calendrier des travaux à effectuer doit donc être correctement géré de manière à ce qu'un maximum de processeurs soient occupés en continu. Petits et gros travaux seront lancés simultanément.

La constitution d'une file d'attente optimisée, qui nécessite de prendre en compte la priorité relative des calculs, a permis d'atteindre sur Tera-1 un rendement de production supérieur à 80 %. Ce qui est assez exceptionnel pour de tels systèmes (en général inférieur ou égal à 70%).

### Inévitables pannes

L'architecture adoptée pour les calculateurs Tera présente cependant des limites. La plus importante est sans doute la charge représentée par les entrées-sorties vers un espace disque partagé du « cluster de SMP » de stockage. Seconde limite de ces très grands systèmes : le nombre élevé

d'éléments électroniques. Même s'ils sont individuellement très fiables, aucun n'est à l'abri d'une panne. Sur Tera-1, on en dénombre une vingtaine par mois. La plupart d'entre elles provoquent des retards de quelques heures seulement dans un calcul (qui dure souvent plusieurs centaines d'heures). Les pannes plus importantes - provoquant l'indisponibilité de la machine - sont heureusement assez rares : moins d'une par trimestre.

Dans ces conditions, il était essentiel de mettre en œuvre des stratégies de tolérance aux pannes et ce, à tous les niveaux. Ainsi, sur Tera-1 les éléments matériels critiques (le réseau d'interconnexions de la machine, une partie de l'espace disque partagé), indispensables au fonctionnement de l'ensemble, sont doublés. Les autres ne le sont pas, pour des raisons de coût. En cas de panne, le logiciel peut

alors être mis à contribution. Par exemple, si un nœud de calcul nous lâche, le calcul en cours d'exécution sur ce nœud est perdu. On le relance alors depuis un « point de reprise » écrit régulièrement par les applications qui tournent sur le calculateur. Bien évidemment, ce risque de panne nécessite la présence de systèmes de maintenance et de surveillance importants. Une quinzaine de personnes sont affectées à la surveillance au CEA, la maintenance étant assurée par le constructeur.

Testée sur Tera-1, l'architecture en « cluster de SMP » a été conservée sur Tera-10. Cela facilite grandement le transfert des logiciels de l'un à l'autre. Ceux-ci devront tout de même être optimisés, ce qui représente un lourd travail. Pour répondre à l'évolution actuelle et prévisible des codes de production et pour limiter le nombre de nœuds du cluster, le choix s'est porté sur une machine équipée de nœuds puissants (100 gigaflops/nœud), offrant une puissance soutenue\* réelle dix fois supérieure à ceux de Tera-1.

Ainsi, si l'architecture générale reste identique, les composants et leur nombre changent. Les 544 « nœuds de calcul » de Tera-10 sont équipés de 16 processeurs soit, en comptant les nœuds dédiés aux entrées-sorties, quelque 9000 processeurs. Si les performances de calcul de Tera-10 sont dix fois supérieures à celles de Tera-1, l'espace disque, lui, a été multiplié par vingt, afin de répondre aux besoins extrêmes dans le domaine de la quantité de résultats produits par les logiciels de simulation.

### \* Puissance

#### soutenue :

puissance moyenne que le calculateur fournit sur une application réelle.

### Stockage sur bandes

Du côté des logiciels « système », qui assurent le fonctionnement de la machine, le choix s'est porté sur des logiciels libres : Linux pour le système d'exploitation des serveurs et Lustre pour le logiciel de système de fichiers global parallèle. Des composants auxquels Bull ajoute des éléments et des encapsulations spécifiques pour optimiser les calculs.

Le passage de Tera-1 à Tera-10 implique également de nombreuses adaptations des infrastructures informatiques. En effet, pour permettre aux utilisateurs de tirer parti de cette nouvelle puissance, le supercalculateur doit être mis en réseau, relié à des systèmes de stockage de données, à des logiciels assurant l'interface et permettant de visualiser les données, etc. C'est ainsi que Tera-10 disposera d'un mur d'image de 5,5 sur 3 mètres. Il permettra aux utilisateurs de visualiser ensemble les résultats des simulations. Une fois Tera-10 en fonctionnement, Tera-1 sera démonté, libérant de l'espace pour Tera-100, qui verra le jour en 2009

Actuellement, la production journalière Tera-1 est de l'ordre de 3 à 5 téraoctets de données par jour. Avec Tera-10, ce volume augmentera de façon quasi proportionnelle avec l'évolution de la puissance de calcul. D'où la mise en place, en 2005 et 2006, d'équipements réseaux et de stockage de données plus performants. Les réseaux internes ont été consolidés, les systèmes de stockage de données revus à la hausse. Ce stockage est effectué à deux niveaux : les données de plus d'un an, qui ont le moins de chances d'être redemandées, sont gardées sur bandes magnétiques (le moyen le plus fiable et le plus économique pour stocker de gros volumes de données), conservées dans des silos. Les données les plus importantes sont dupliquées et placées dans des zones protégées. Pour le stockage à court terme, les bandes magnétiques seront remplacées par des disques durs magnétiques, dont le coût, porté par le marché du PC domestique, a énormément baissé. Dans ce cadre, un espace disque de 4 pétaoctets pour le premier niveau de stockage de

données sera installé pendant l'année 2006.

Pour accueillir Tera-1, un grand bâtiment (2 000 m<sup>2</sup> de salles machines) avait été construit. Il avait alors été conçu pour recevoir les différentes générations d'équipements de simulation (calculateurs et stockage) y compris Tera-100, prévue pour 2009. Toutefois, chaque nouvelle version nécessite quelques adaptations.

Ainsi, pour passer de Tera-1, qui consomme environ 0,6 MW, à Tera-10 qui va en absorber trois fois plus, il a été nécessaire d'installer un nouveau groupe froid de 2 MW, trois onduleurs (qui fournissent du courant alternatif à partir de batteries en cas de coupure électrique), sans oublier une extension du réseau de détection/extinction incendie et le câblage de la salle machine!

Une fois les codes de simulation utilisés pour la conception et la garantie de Tera-10 validés, Tera-1 sera arrêté et démonté. Cette opération est prévue à la mi-2006. L'espace libéré accueillera en 2009 Tera-100.

Avec Tera-10, les physiciens et concepteurs du CEA/DAM disposeront d'une capacité de simulation dix fois supérieure à celle fournies par Tera-1. Mais l'installation, puis les six mois d'optimisation d'un système de cette taille restent un défi.

Car, comme pour tout équipement exceptionnel, les technologies et les expériences humaines seront ici poussées à des niveaux rarement atteints. ■

J. G., P. L. et F. R.

*Cet article est issu du journal La Recherche, N°393, janvier 2006*

**Une fois Tera-10 en fonctionnement, Tera-1 sera démonté, libérant de l'espace pour Tera-100, qui verra le jour en 2009**

**TERA-10 consommant trois fois plus d'énergie que son prédécesseur, un groupe froid de 2 MW a été installé.**

© CEA

