Architecture des machines parallèles modernes

Ronan Keryell
rk@enstb.org
Département Informatique, ENST Bretagne

14 février 2006

1 Introduction

Depuis 1993, la majorité des ordinateurs les plus puissants sont référencés dans le « Top 500 » [TOP] et la crème dans le « Top 10 ». Ils sont étalonnés selon le programme de factorisation de matrice *LU* LINPACK.

Parmi ces 10 premiers, on remarquera :

- l'IBM BlueGene/L au Lawrence Livermore National Laboratory du Département américain de l'énergie (DOE) qui culmine à 280 TFLOPS ($2{,}8.10^{14}$ opérations flottantes par seconde) sur le test du LIPACK avec ses 131 072 processeurs ;
- l'ordinateur ASCI Purple dans le même laboratoire et construit aussi par IBM mais à base de systèmes *pSeries* 575 qui atteint 63 TFLOPS avec 10 240 processeurs ;
- l'ordinateur Comlumbia de la NASA/Ames construit par SGI avec ses 51 TFLOPS ;
- les 2 ordinateurs des Sandia National Laboratories encore du DOE, une grappe à base de PowerEdge de Dell et un Cray XT3 à base d'Opteron ;
- le japonais Earth Simulator de NEC qui a occupé la première place pendant plusieurs années est relégué à la 7ème place avec ses « modestes » 35 TFLOPS ;
- le Cray XT3 au Oak Ridge National Laboratory du DOE est 10ème avec 20 TFLOPS.

Ces ordinateurs de course, même s'ils ne sont pas représentatifs d'une part de ce qu'un laboratoire standard peut se permettre, ni d'autre part de vraies applications plus demandeuses en ressources matérielles, sont intéressants car ils montrent les différentes technologies mises en œuvre et celles qui se démocratiseront dans un futur proche (figure 1). Il faut aussi noter la prédominance des USA en terme de puissance de calcul installée mais aussi en terme de construction, ce qui peut inquiéter devant l'importance stratégique de ces moyens de calcul dans l'industrie, la recherche et la défense.

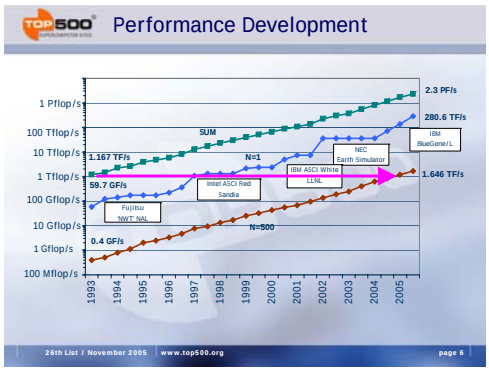

FIG. 1 – Évolution des performances du Top 500.

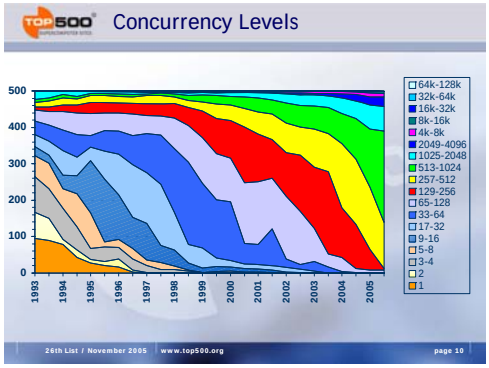

FIG. 2 – Évolution du parallélisme dans le Top 500.

Les technologies employées sont principalement basées sur l'utilisation massive de processeurs scalaires souvent agrégés sous forme

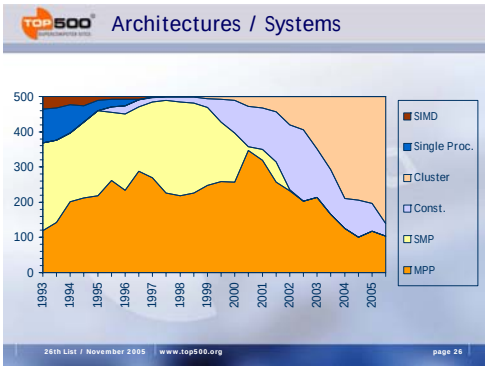

FIG. 3 – Évolution des architectures du Top 500.

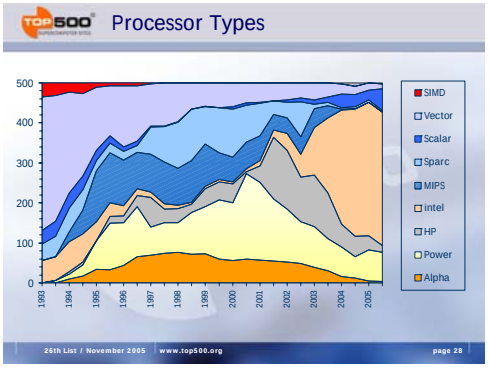

FIG. 4 – Évolution des types de processeurs des architectures du Top 500.

de grappe (figure 2) pour atteindre des performances élevées contrairement à l'hégémonie des ordinateurs vectoriels par le passé (figure 3).

La performance globale d'un ordinateur parallèle va s'appuyer sur ses trois composants [HPC] :

- la puissance brute des processeurs ;
- le débit et la latence mémoire ;
- le débit et la latence du réseau d'interconnexion.

2 Les processeurs modernes

Les processeurs standards ont bénéficié de l'économie de masse et des formidables investissements opérés et sont de plus en plus utilisés dans les ordinateurs à haute performance (figure 4) par rapport aux machines vectorielles style Nec SX-8 [NEC] ou Cray X1E [Cra].

Les performances des processeurs sont obtenues en utilisant de nombreuses unités d'exécution fonctionnant en parallèle. Afin d'alimenter en données ces unités, le (relativement) faible débit mémoire est compensé par des mémoires caches internes aux processeurs de plus en plus importantes. Pour dépasser ce parallélisme d'unités fonctionnelles, on assiste à la démocratisation des processeurs multi-cœurs qui partagent le même circuit intégré.

Afin d'adapter au mieux les opérations à la taille des opérandes, un mode « multimédia » ou vectoriel a été rajouté pour traiter, en une instruction sur 2 gros opérandes, de nombreuses opérations identiques sur de petits opérandes rangés dans les gros. Ce sont par exemple les modes SSE (*Streaming* SIMD *Extension*) d'Intel [Ita].

3 Réseaux d'interconnexion

La démocratisation d'Ethernet 1 Gb/s et l'apparition de la version 10 Gb/s fait que de nombreuses machines utilisent cette technologie bon marché même si la latence est rédhibitoire et le débit faible pour des applications haut de gamme.

Pour plus de performance, des constructeurs rajoutent un système d'interconnexion dédié au plus proche des processeurs (cas de la *Direct Connected Processor (DCP) Architecture* du Cray XD1 ou SeaStar du Cray XT3 sur le bus HyperTransport des Opteron [Cra]) pour ne pas avoir la latence des bus PCI.

Avec des performances intermédiaires, de nombreux fabricants spécialisés dans les réseaux de calculateurs proposent des solutions (Myrinet [Myr], Quadrics [Qua] qui équipe la dernière machine Bull Tera-10 du CEA, InfiniBand...) permettant d'avoir des performances raisonnables de type 10 Gb/s, même avec les bus d'ordinateurs bon marché PCI.

Au niveau des topologies d'interconnexion, on revient à des choses simples, de type processeurs sur grille 2D ou 3D ou pour de plus grosses machines de *fat trees*, sortes d'hypercubes écrasés qui offrent un plus faible diamètre que les grilles.

4 Les systèmes reconfigurables

Les processeurs généralistes sont optimisés pour les opérations courantes et certaines applications ne fonctionnent pas forcément très bien sur ces processeurs prédéfinis. Pour dépasser cette inefficacité, certains ont eu l'idée d'utiliser des circuits logiques reconfigurables permettant de réaliser directement matériellement les algorithmes voulus. Bien que coûteuse, certaines applications telles que la bioinformatique ou le traitement d'image peuvent être accélérées de manière importante.

On peut noter dans cette catégorie le Cray XD1 [Cra] avec ses cartes accélératrices à base de FPGA Xilinx Virtex 4.

5 Cartes d'accélération graphique

La demande continue du grand public pour les jeux vidéo toujours plus réalistes fait qu'on trouve actuellement sur le marché des cartes d'accélérations graphiques extrêmement performantes permettant d'améliorer la qualité des images : suréchantillonnage, transluscence, modèles d'illumination globale, etc.

Pour permettre de s'adapter à tous ces algorithmes en constante évolution, les cartes graphiques sont devenues de véritables supercalculateurs dotés de beaucoup de mémoire et qui sont certes spécialisés mais néanmoins programmables avec des compilateurs C. Pour plus de puissance, on peut faire travailler plusieurs cartes ensembles (technologie SLI de nVidia [nVI]).

Bien que plutôt adaptées aux applications graphiques (pipelines de transformations géométriques...), elles peuvent être utilisées pour accélérer des morceaux d'application, typiquement lorsqu'on manipule massivement des opérandes de taille réduite.

6 Conclusion

Paradoxalement, le monde de l'informatique haute performance ne va pas encore vers une simplification pour les programmeurs. Pour atteindre des performances élevées, les ordinateurs utilisent de plus en plus de techniques pour exploiter le parallélisme à tous les niveaux. Atteindre de telles performances se fait au prix d'une synergie subtile entre différents niveaux hétérogènes (parallélisme à l'intérieur des processeurs avec plusieurs cœurs, eux-mêmes avec jeux d'instructions vectorielles, mémoires à accès non uniforme selon les processeurs, mémoires caches, topologie réseau, etc) qui a vite fait de plomber les performances si elle n'est pas bien exploitée.

Malheureusement, les outils logiciels développés par le passé se sont montrés décevants et sont de plus en plus distanciés par cette hétérogénéité galopante, ce qui explique que les programmeurs reviennent à des outils plus basiques mais plus stables et plus portables.

Les défis futurs sont donc

- pour les programmeurs, d'arriver à maîtriser cette complexité globale en plus de celle des applications et d'être de plus en plus tolérants à la latence que l'on peut rencontrer dans les vastes grilles de calcul ;
- pour les architectes, d'arriver à fournir des machines efficaces simplement ;
- pour les spécialistes de la compilation, de fournir des outils plus efficaces et de plus haut niveau que la programmation par passage de messages.

Références

[Cra] « Cray – the supercomputing company ». `http://www.cray.com`.

[HPC] « HPC Challenge Benchmark Results - Systems for Kiviat Chart ». `http://icl.cs.utk.edu/hpcc/hpcc_results_kiviat.cgi`.

[Ita] « Intel Itanium 2 Processor Product Information ». `http://www.intel.com/products/processor/itanium2`.

[Myr] « Myricom ». `http://www.myri.com`.

[NEC] « NEC SX-8 Scalable Vector Supercomputer ». `http://www.hpce.nec.com/sx-8.0.html`.

[nVI] « nVIDIA ». `http://www.nvidia.com`.

[Qua] « Quadrics ». `http://www.quadrics.com`.

[TOP] « TOP500 Supercomputer Sites ». `http://top500.org`.