

## Architecture des machines parallèles modernes


Ronan Keryell

Département Informatique ENST Bretagne

rk@enstb.org

14 février 2006

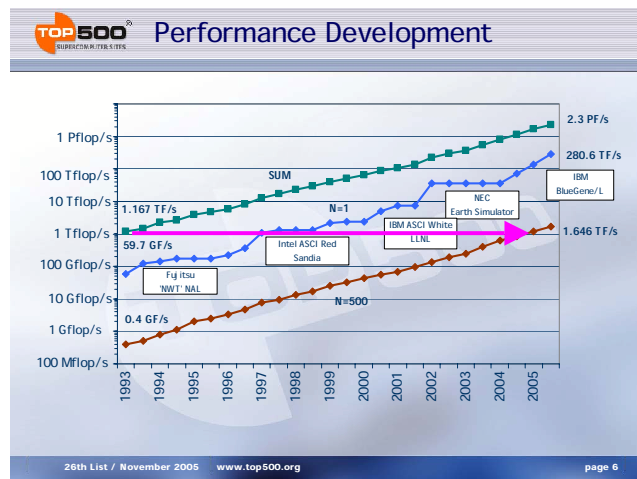
<http://top500.org>

- Liste 500 plus gros ordinateurs déclarés dans le monde depuis 1993
- Top 10 : crème de la crème
- Étalon : factorisation de matrice *LU* LINPACK
  - ▶ Plus de calculs que de communications
  - ▶ Cas d'école hyper régulier rarement rencontré dans la vraie vie
  - ▶  À considérer comme une puissance crête (efficace)



Architecture ordinateurs parallèles  
Département Informatique

• Introduction  
▶ Top 500



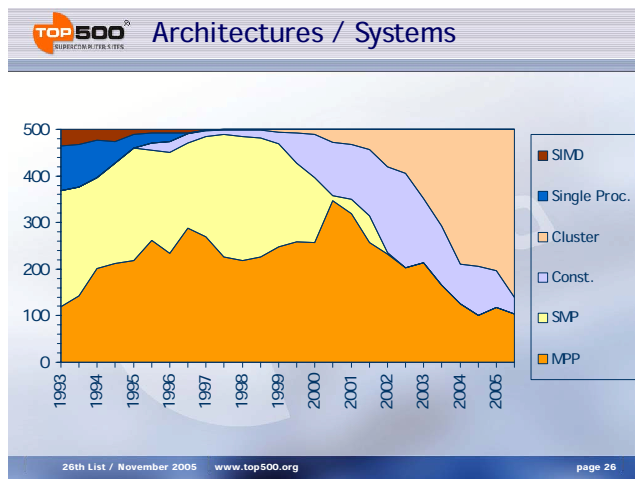
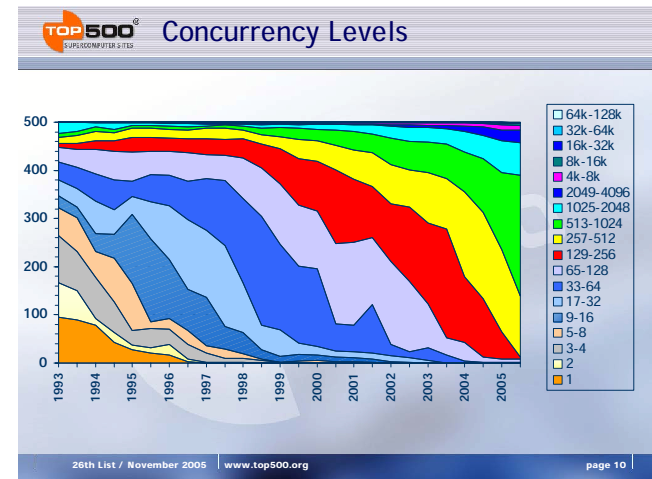
- IBM BlueGene/L au Lawrence Livermore National Laboratory du Département américain de l'énergie (DOE) : culmine à 280 TFLOPS ( $2,8 \cdot 10^{14}$  opérations flottantes par seconde) avec 131 072 processeurs
- IBM ASCI Purple dans même laboratoire et construit à base de systèmes *pSeries* 575 : 63 TFLOPS avec 10 240 processeurs
- SGI Columbia de la NASA/Ames : 51 TFLOPS
- 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 ;
- Japonais Earth Simulator de NEC, longtemps première place : relégué à la 7<sup>ème</sup> place avec ses « modestes »



35 TFLOPS

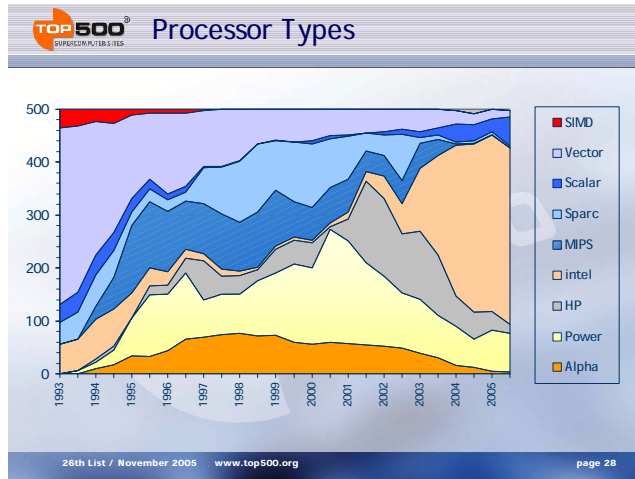
- Cray XT3 au Oak Ridge National Laboratory du DOE est 10<sup>ème</sup> avec 20 TFLOPS

Prédominance stratégique des USA... ☹



- Tendance à utilisation massive de processeurs standards
- Moins de processeurs vectoriels





- Processeurs vectoriels de 16 GFLOPS : 1 processeur scalaire + 4 processeurs vectoriels
- 8 processeurs par nœud
- 512 nœuds = 65 TFLOPS
- Processeur CMOS mono-chip 90 nm & 9 niveaux de cuivre
- 8 210 pattes dont 1 923 de signaux !
- Record du monde de 300 Go/s circuit-extérieur



### Composantes complémentaires

- Puissance brute des processeurs
- Débit et latence mémoire
- Débit et latence du réseau d'interconnexion
- Entrées-sorties



Depend de l'application visée ☺



- Processeurs vectoriels de 18 GFLOPS : 4 processeurs vectoriels
- 16 à 8 192  $\rightsquigarrow$  147 TFLOPS
- Mémoire partagée
- Réseau avec 16 tores 2D

<http://www.cray.com/products/x1e>



- ASCI Purple : 10 240 processeurs, N° 2 au Top 500 en 2005
- Version spéciale cluster (grappe)
- 8 Power5 1,9 GHz 64 bits ou 8 bi-cœur 1,5 GHz/lame
- 4 liaisons InfiniBand vers switches (TopSpin MPI...)
- 2 Ethernet 1 Gb/s
- AIX5L ou Linux



- Centaines de compteurs de performance pour comprendre ce qui se passe ☺



<http://www-03.ibm.com/servers/eserver/pseries/news/related/2004/m204>

- 64 bits
- Pipeline 15 étages
- 8 instructions/cycle
- SMT (*Simultaneous Multi-Threading*) à priorité pour remplir bulles du pipeline du Power4
- 120 registres physiques entiers + 120 flottants partagés par les 32+32 registres virtuels des 2 threads : renommage à la volée style

$$r3 = r1 + r2$$

$$r1 = r4 * r5$$

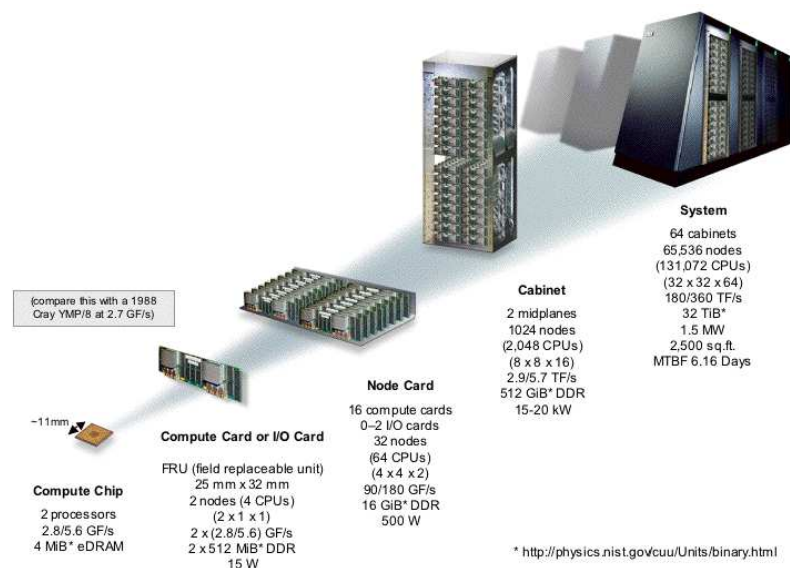
$$=: r3 = r1 + r2; \quad r1' = r4 * r5$$


<http://www.llnl.gov/asc/platforms/bluegenel/overview.html>

<http://www.llnl.gov/asc/platforms/bluegenel/arch.html>

- N° 1 au Top 500 en 2005 avec 131 072 processeurs
- Base de 2 PowerPC440 700 MHz avec 2 unités de calcul flottant
- 10× efficacité électrique par rapport aux pSeries
- Réseau tore 3D + arbre pour réductions/diffusions
  - ▶ Diamètre 64
  - ▶ Latence inter-nœud de 100 ns
  - ▶ 6,4 ns latence maximum
  - ▶ 175 Mo/s/lien assez faible



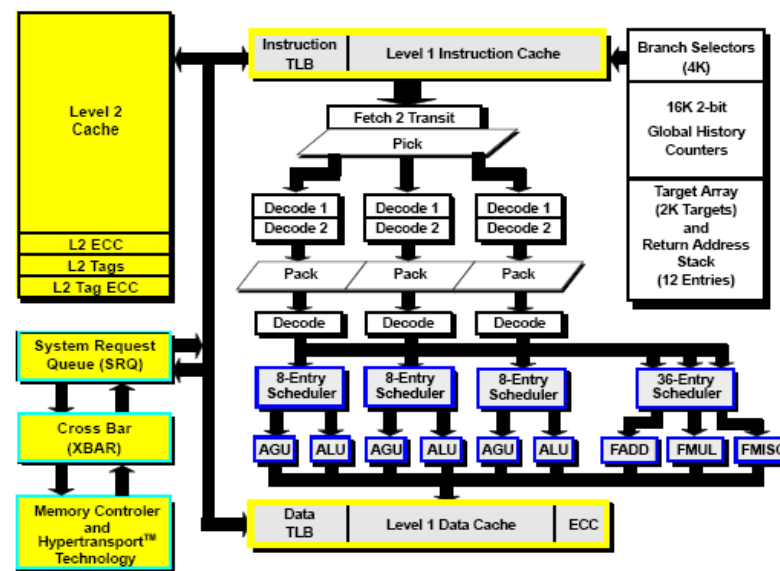


- Jeu d'instruction à la x86
- Mode 64 bits qui double aussi nombre de registres
- Superscalaire à exécution dans le désordre de 9 instructions/cycle
  - ▶ 3 instructions entières
  - ▶ 3 générations d'adresses
  - ▶ 3 calculs flottants (add, mul, mémoire)
- Interface par 3 canaux HyperTransport de 8 Go/s au monde extérieur (mais 0,1 x SX-8)

[http://www.amd.com/us-en/assets/content\\_type/white\\_papers\\_and\\_tech\\_d](http://www.amd.com/us-en/assets/content_type/white_papers_and_tech_d)



- MTBF de BlueGene/L : 6 jours...
- 1951 : simulateur temps réel Whirlwind de Jay FORRESTER & Bob EVERETT
  - ▶ 500 000 +/s, 50 000 x/s  
 $5,6 \cdot 10^{-9} \times \text{BlueGene/L} \odot \rightsquigarrow +51\%/\text{an en 54 ans}$
  - ▶ Mémoire à tores
  - ▶ Lampes  $\rightsquigarrow$  consomme \$32 000 de tubes/mois !  $\odot$
- $\rightsquigarrow$  Faire du check-pointing & redondance
- $\rightsquigarrow$  Projet ANR ARA SSIA SafeScale plus général prenant en compte attaques malicieuses dans grilles (ENSTB-IMAG-Paris 13-IRISA)



- Bi-cœur :  $10^9$  transistors...
- Pipeline et exécution *dans l'ordre* de 6 instructions/cycle
- 128 registres entiers, 128 registres flottants, 128 registres prédicats/conditions visibles
- Contrôle très fin de la micromachine avec VLIW EPIC
- Encore plus de stress sur le compilateur (et programmeurs ☹)
- Mais permet d'avoir de bonnes performances



## Instructions SIMD

22

- Applications multimédia : couleurs sur 8 bits, son sur 16 bits, modem...
- $\rightsquigarrow$  Adaptation des processeurs standards aux petites données
- Faire calcul sur petites données indépendantes plutôt qu'une grosse
- Instructions SSE3 traitent 128 bits de données comme du calcul SIMD ou vectoriel/cycle sur
  - ▶ 2 entiers ou flottants double précision 64 bits
  - ▶ 4 entiers ou flottants simple précision 32 bits
  - ▶ 8 entiers 16 bits
  - ▶ 16 entiers 8 bits



- 60 TFLOPS LINPACK au CEA/DAM
- 544 Bull Novascale 6160 avec 8 Intel Montecito double cœur
- 27 To de mémoire
- 64 serveurs d'E/S
- 1 Po de disques
- Linux & système de fichiers Lustre
- Réseau Quadrics



## Instructions SIMD

23

- Instructions
  - ▶ Calculs divers
  - ▶ Compactage-décompactage
  - ▶ Conversions de format
  - ▶ Comparaisons
- ⚠ Ne marche que pour données contiguës en mémoire ☹
- Pas de *scatter/gather* ☹

[http://www.amd.com/us-en/assets/content\\_type/white\\_papers\\_and\\_tech\\_d](http://www.amd.com/us-en/assets/content_type/white_papers_and_tech_d)





- Organisation depuis 1975 en tableau de bit avec commande selon une ligne (rangée) puis une colonne
- Adresse envoyée d'abord suivant la rangée (échantillonnée sur RAS (*Row Address Strobe*) puis la colonne (échantillonnée sur CAS (*Column Address Strobe*))
- Stockage dans un simple condensateur  $\rightsquigarrow$  nécessité de « rafraîchir » la donnée régulièrement (moins de 5 % du temps)
- Par commodité, mémoires vendues souvent sous forme de barrettes SIMM (*Single*) ou DIMM (*Dual Inline Memory Module*)
- Gain en vitesse :



- Mode de terminaison des lignes programmable et synchrone en multi-banc sur les écritures
- Taille de page (rangée) moitié par rapport au DDR : division consommation par 2 lors d'une commande ACTIVATE ( $\approx$ RAS)
- Réglage fin du pipeline des opérations
- DDR2-667 MHz (barrettes PC2-5400) 5-5-5 (latence CAS (CL), latence RAS (RCD), latence précharge (RP) en cycles) a une latence de 5 cycles CAS :  $3 \times 5 = 15$  ns
- Si accès aléatoire, latence totale = CL + RAS  $\rightsquigarrow$  10 cycles, 30 ns sur premier bit d'une page



- Grand débit interne disponible (parallélisme sur les rangées)
- Éviter un cycle de RAS si localité dans la même page : *Fast Page Mode*  $\rightsquigarrow$  se contente de lire dans le tampon de sortie des rangées
- Évite des verrous externes pour échantillonner les signaux  $\rightsquigarrow$  SDRAM (*Synchronous DRAM*) avec bascules D à l'intérieur
- Échantillonnage des signaux sur front montant et descendant  $\rightsquigarrow$  DDR (*Double Data Rate*)
- DDR2 pour des transferts de plus de 400 MHz, 256 Mb–4 Gb, diminution consommation
- 1,8 V



- Si accès aléatoires en permanence à la mémoire, temps de précharge en plus, 15 cycles, 45 ns
- Latence semblable à de la DDR mais débit double 45 ns pour Opteron 3 GHz = 1620 instructions ! ☺  

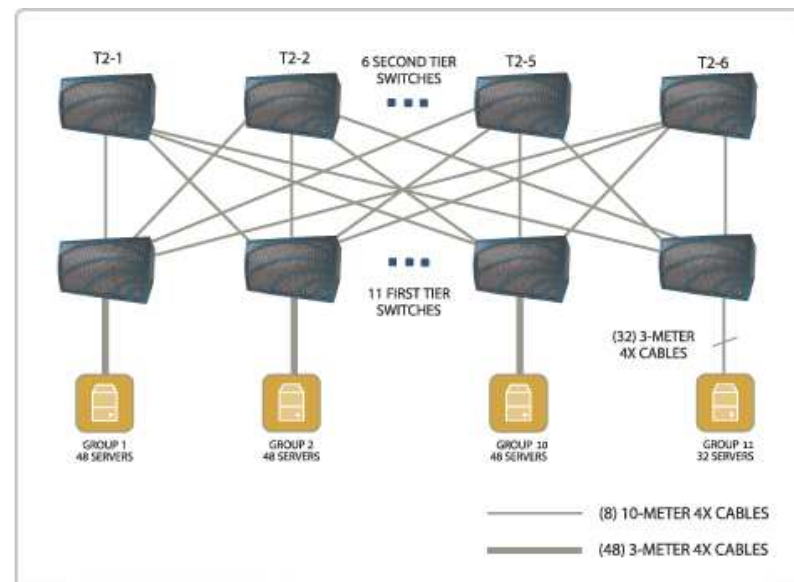
La localité est toujours importante !
- RAMBUS
  - Remplacer signaux mémoires DRAM classiques par des bus rapide à transactions éclatées (lecture/écriture, retour,...)
  - Interfaces plus chères
- Projets de recherche PIM (Processors In Memory) : *énorme* débit processeurs-mémoires si local  $\rightsquigarrow$  Revoir modèles de calcul/programmation ?



- Liaison à 10 Gb/s (4×)
- Latence 6  $\mu$ s
- Exemple de réseau de 512 nœuds  
<http://www.topspin.com> style *fat-tree*



- Elan4 QsNet II qui équipe Bull Tera-10 du CEA/DAM
- 912 Mo/s
- MPI sur Opteron
  - ▶ 1,5  $\mu$ s latence
  - ▶ <2  $\mu$ s barrière
- À la recherche du temps perdu...
  - ▶ 990 ns dans chipset processeur
  - ▶ 240 ns dans carte Elan
  - ▶ 218 ns dans câbles (vitesse de la lumière ☺)
  - ▶ 213 ns dans switches Elan
- <http://www.quadrics.com>
- ≈\$1 700/port en 2006

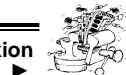


- 10 Gb/s en Gigabit Ethernet ou Myrinet
- Myrinet : protocole plus efficace qu'Ethernet (entêtes...) :  
9,8 Gb/s, 2  $\mu$ s de latence MPI
- <http://www.myri.com>





- Le réseau de base pour les masses !
- La solution du pauvre...
- 1 Gb/s mais latence assez élevée (couches protocolaires)
- Utilisé sur machines haut de gamme comme lien d'administration



- Processeurs généralistes : optimisés pour opérations courantes
- Certaines applications ne fonctionnent pas forcément très bien sur ces processeurs prédéfinis
- ↗ Pour dépasser inefficacité : rajout de circuits logiques reconfigurables (programmables)
- Réalisent matériellement algorithmes voulus
- Très efficace en bioinformatique ou traitement d'image



- Recherche débridée 1980-2000
- Dans la vraie vie actuelle : topologies simples à réaliser
  - ▶ Grilles 2D ou 3D
  - ▶ *Fat-tree* : réseaux multi-étage de routeurs favorisant localité
- Quelques constructeurs font encore du « sur mesure » (sur bus HyperTransport Opteron dans Cray XT3)



- 144 Opterons
- Réseau spécifique sur canaux HyperTransport
- Synchronisation matérielle
- Cartes accélératrices à base de FPGA Xilinx Virtex 4.



- Demande continue du grand public pour jeux vidéo toujours plus réalistes
  - ▶ Suréchantillonnage
  - ▶ Translucence
  - ▶ Modèles d'illumination globale
  - ▶ ...
- ↗ Cartes d'accélération graphiques extrêmement performantes
- Algorithmes en constante évolution ↗ Cartes graphiques ≡ véritables supercalculateurs
  - ▶ Beaucoup de mémoire



- Unix règne en maître absolu
- Toujours plus de puissance
  - ▶ Nombre de processeurs ↗
  - ▶ Superscalaires voire vectoriels, instructions SIMD, multi-cœurs (sauf portables)
  - ▶ Coprocesseurs graphiques, reconfigurables
  - ▶ Retour des machines virtuelles des années 70 : virtualisation des machines parallèles avec différents OS...
- ↗ Architectures de plus en plus hétérogènes
  - ▶ Outils automatiques peu efficaces généralement dans vraie vie



- ▶ Spécialisées mais néanmoins programmables avec compilateurs C ou C++
- ▶ Possible de faire travailler plusieurs cartes ensembles (technologie SLI de nVidia)
- Idée : utiliser pipelines de transformations géométriques pour calculs scientifiques



- ▶ Complexité pour le programmeur
- ▶ Diversité architecturale ↗ nivellement par le bas du modèle de programmation : passage de message (MPI ≡ assembleur du parallélisme)
- Comment rester proche puissances crêtes annoncées ?
- *Real Politik* : retour à compromis de modèles de programmation hétérogènes pour architecture hétérogène
  - ▶ Nœuds SMP programmés en OpenMP en interne (multi-thread pour les nuls ☺)
  - ▶ Interconnexion de ces nœuds programmés en MPI (passage de messages pour les nuls ☺)



- Pour programmeurs
  - ▶ Maîtriser complexité globale + complexité applications
  - ▶ Tolérer latence mémoire (NUMA) + réseaux (GRID)
- Architectes
  - ▶ Machines efficaces simplement
- Spécialistes en compilation
  - ▶ Créer chaînon manquant !
  - ▶ Fournir outils plus efficaces et de plus haut niveau



1 Titre . . . . .	0	11 IBM Power 5 . . . . .	13
2 Top 500 . . . . .	1	12 IBM BlueGene/L . . . . .	15
1 Introduction . . . . .	0	13 Tolérance aux pannes . . . . .	17
1 Top 500 . . . . .	0	14 AMD Opteron . . . . .	18
3 Le Top 10 . . . . .	3	15 Intel Itanium2 . . . . .	20
4 Parallélisme massif . . . . .	5	16 Bull Tera-10 . . . . .	21
5 Architectures . . . . .	6	17 Instructions SIMD . . . . .	22
6 Types de processeurs . . . . .	8	16 Jeux d'instruction SIMD	21
7 Performance globale . . . . .	9	18 Mémoire dynamiques (DRAM) . . . . .	24
8 Nec SX-8 . . . . .	10	17 Mémoire . . . . .	23
7 Architectures vectorielles . . . . .	9	19 Infiniband . . . . .	28
9 Cray X1E . . . . .	11	18 Réseaux d'interconnexion . . . . .	27
10 IBM p5-575 . . . . .	12	20 Quadrics . . . . .	30
9 Processeurs super-scalaires . . . . .	11	21 Myrinet . . . . .	31
		22 Ethernet . . . . .	



### Conférence à l'ENST Bretagne

- « Les grands moyens de simulation numérique du CEA »
- Hervé LOZACH
- Mercredi 8 mars 2006, 13h50–16h50



### Table des matières

23 Topologie . . . . .	33	27 Conclusion . . . . .	38
24 Systèmes reconfigurables . . . . .	34	26 Conclusion . . . . .	37
23 Les systèmes reconfigurables . . . . .	33	28 Défis futurs . . . . .	40
25 Cray XD1 . . . . .	35	29 Minute de publicité . . . . .	41
26 Cartes graphiques . . . . .	36	30 Table des matières . . . . .	42
25 Cartes d'accélération graphique . . . . .	35	31 Index . . . . .	43

