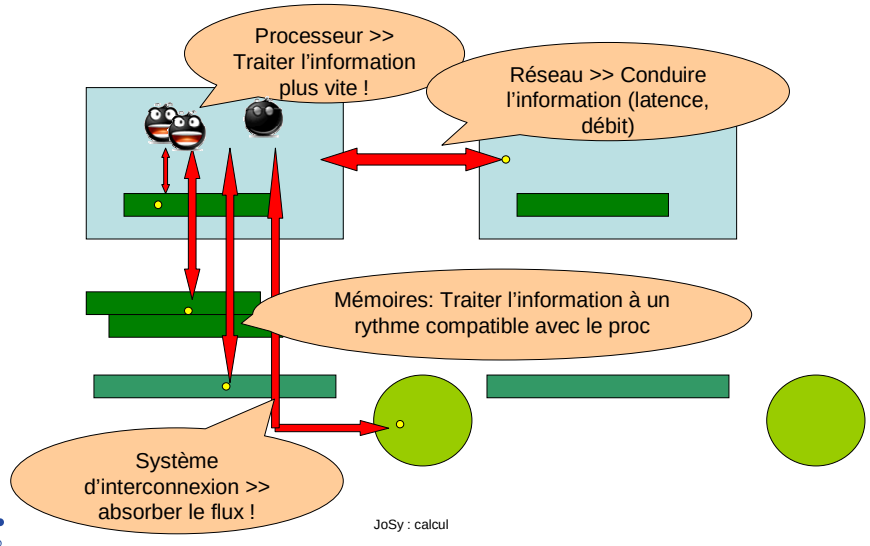
Architectures matérielles

Quelques éléments (Août 2007)

Françoise BERTHOUD, LPMMC, CNRS

Une architecture équilibrée pour ses besoins



Comment caractérise-t-on ?

- **Processeur** :
  - Architecture, jeux d'instructions
  - Nombre de CPU
  - Hz = Nombre d'impulsions par seconde
  - MIPS (Millions d'instructions par seconde) = Fréquence d'horloge / CPI (nombre moyen de cycles d'horloge nécessaire à l'exécution d'une instruction sur un microprocesseur)
- **Front Side Bus** : Hz
- **Registres** :
  - Nombre
  - Taille (32, 64, 128 bits …)
- **Mémoire** :
  - Fréquence (MHz)
  - Taille et son implantation (par rapport au CPU)
  - Les caractéristiques techniques de son lien avec le proc (largeur en bit, fréquence en MHz)
  - Latence (mesurée en ns = qqs dizaines de cycles)
  - Taille des lignes de cache (bit)
- **Réseaux d'interconnexion interne ou externe** (bus, réseau, …) :
  - Latence (en ns)
  - Largeur (en bit)
  - Fréquence (en MHz)

Exemple : Pour le calcul de la bande passante mémoire locale sur un nœud XEON :
-BP au niveau de l'interface processeurs (FSB) : sur une carte bi-sockets Xeon 5xxx, on a 2 FSB 64bits à 1333MHZ, soit 2*1333*8=21.3GB/s
-BP au niveau de l'interface mémoire : la mémoire est adressée via 4 bus cadencés à 667MHz (si DDR-667). La largeur du bus est de 72 bits, dont 8 bits de contrôle, soit 64bits de données. Ce qui donne pour le calcul : 4*667*8=21.3GB/s

Les processeurs - rappels

1 CPU = plusieurs unités fonctionnelles :
-une unité de gestion des bus
-une unité d'instruction (control unit) : lit les données, les décode et les envoie à
-une unité d'exécution :
- >= une unité arithmétique et logique (**UAL**) (fonctions basiques entiers et op logiques)
- >= une unité de virgule flottante (FPU) : calculs complexes non entiers : différents types de **FPU** (certaines font * ou +, d'autres **FMA** (a*b +c) …)
- etc.

Les FPU travaillent a partir des registres, si registres 128 bits ➔ Performance crete peut dépendre de la taille des registres et est fonction de la précision (SP, DP)

Phases d'exécution d'une instruction :

LI (lecture dans cache), DI (décodage et recherche opérandes), EX (exécution) ADD, SUB etc, MEM (accès mémoire, écriture ou chargement), ER (écriture de la valeur calculée dans les registres)

Les cpu : vers une optimisation de leur utilisation : les jeux d'instruction

Le mode SIMD permet d'appliquer la même instruction simultanément à plusieurs données (stockées dans un registre) pour produire plusieurs résultats. Exemples de jeux d'instructions SIMD :

- MMX : permettent d'accélérer certaines opérations répétitives dans des domaines tels que le traitement de l'image 2D, du son et des communications.
- SSE (SSE2,SSE3, SSE4 en 2008) : par ex instructions pour additionner et multiplier plusieurs valeurs stockées dans un seul registre (registre SSE)
- 3DNow

Processeur CELL (performance crete (registres 128 bits, SP) :

4 (SP SIMD) * 2 (FMA) * 8 SPUs *3.2 GHz = 204.8 GFlops / socket (en SP)

Les cpu : vers une optimisation de leur utilisation

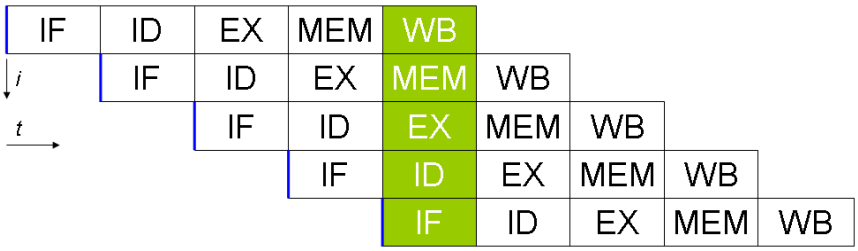
- **Augmenter la fréquence d'horloge** (limites techniques, Augmenter l'intégration en diminuant la taille des transistors. Ceci augmente la vitesse de transfert entre les différentes parties, solution coûteuse)
  - ➔ Nb de cycles d'horloge pour interruptions, « cache miss » ou mauvaise prédiction de branchement augmentent

D'après Intel, le passage au 45 nm permettra de doubler le nombre de transistors dans une puce par rapport à une gravure en 65 nm, mais aussi d'augmenter de 20 % la fréquence de fonctionnement de la puce, tout en réduisant de 30 % sa consommation.

- Permettre **l'exécution simultanée de plusieurs instructions** :
  - Instruction Level Parallelism : pipelining, superscalabilité, architecture VLIW et EPIC
  - Thread Level Parallelism : multithreading et SMT.

Les cpu : vers une optimisation de leur utilisation : IPL

Pipelining (pas de duplication de composants hardware) : Le pipelining consiste à exécuter simultanément des étapes différentes d' opérations différentes et indépendantes

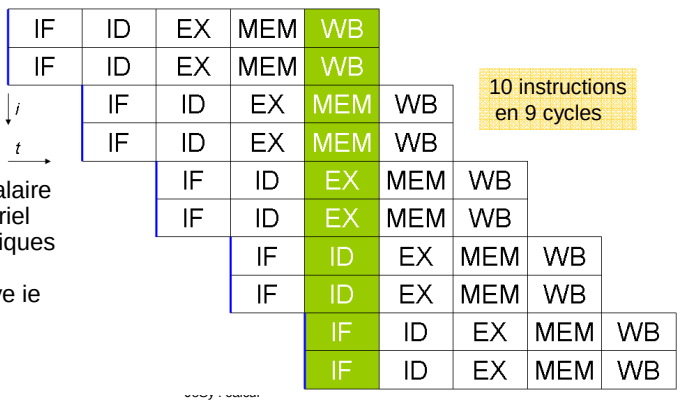


5 instructions en parallèle en 9 cycles (25 cycles en séquentiel)

➔ permet de multiplier le débit avec lequel les instructions sont exécutées par le processeur.

Les cpu : vers une optimisation de leur utilisation : IPL

Processeur superscalaire : Exécution en parallèle (duplication de composants hardware (unités d'exécution, UAL, FPU, FMA, ..), coûteux) :



10 instructions en 9 cycles

calculateur superscalaire (gestion par le matériel des instructions statiques ou dynamiques, exécution spéculative ie avt contrôle de dépendances)

Les cpu : vers une optimisation de leur utilisation : IPL

Les architectures VLIW (Very Long Instruction Word) et EPIC (Explicitly Parallel Instruction Set Computing) sont des architectures dont le jeu d'instructions permet d'exprimer le parallélisme entre opérations.

VILW : parallélisme entièrement géré par le compilateur qui produit un code spécifique à une implémentation de l'architecture (ordonancement statique). (par exemple le processeur Trimedia)

➔On gagne en performance (pas de contrôle), mais la compatibilité binaire en générations successives en difficile ou impossible.

EPIC :parallélisme exprimé de manière indépendante de la mise en oeuvre du processeur. les instructions // sont déterminées par le compilateur. (ex IA 64)

➔ l'effort d'optimisation repose sur le compilateur, qui a la charge d'organiser statiquement les dépendances inter-instructions.

Exemple : Sur un Itanium, l'organisation d'un mot est la suivante : 3 instructions de 41 bits chacune, et un template de 5 bits qui détaille les dépendances inter-instructions (et éventuellement par rapports aux mots précédents/suivants), soit 128 bits (3 x 41 + 5).

Les cpu : vers une optimisation de leur utilisation : TPL

Améliorer le remplissage du flot d'instructions du processeur : Multithreading, SMT (Simultaneous Multithreading), hyperthreading :

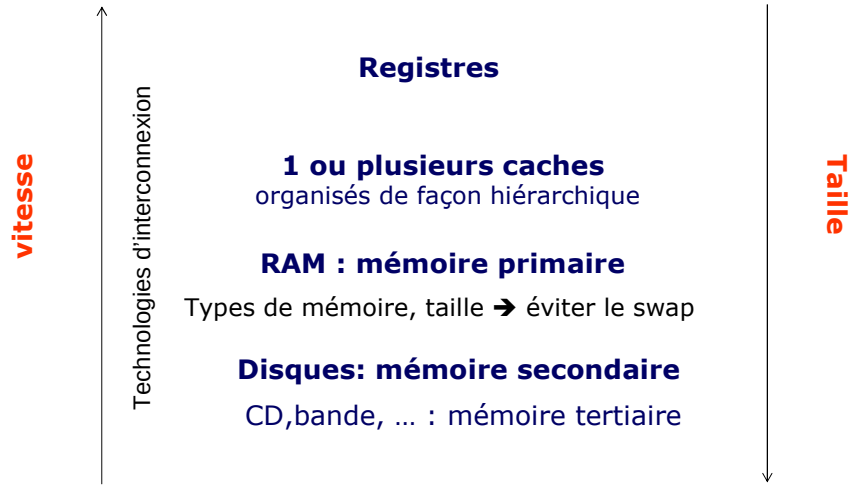
- Le multithreading, définir plusieurs processeurs logiques au sein d'un processeur physique. Le système reconnaît deux processeurs physiques et se comporte en système multitâche en envoyant deux threads simultanés.
- Le SMT : exploite le pipelining, registres et cache entre les threads (même programme ou non)
- L'hyperthreading = SMT (intel)

Exemples : Si deux threads peuvent se partager le pipeline, on parle de SMT à deux voies (comme pour l'Hyperthreading d'Intel), de SMT à 4 voies pour 4 threads (comme pour le DEC Alpha EV8).
Le POWER 5 d'IBM intègre un SMT deux voies complexe, puisqu'il peut attribuer des priorités aux threads et activer/désactiver le SMT de manière dynamique pour les cas où la méthode n'augmente pas les performances.

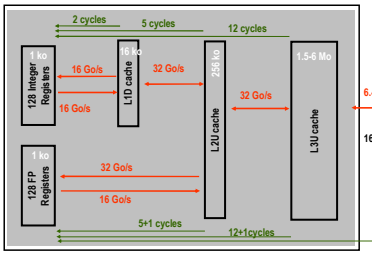
Les processeurs multi cœurs

- Unités de calcul homogènes :
  - AMD : opteron quadri core (2 FPU / cœur)
  - Intel : xeon : quadri core (2 FMA / cœur)
  - Intel : itanium : dual core (2 FMA / cœur)
  - Sun : ultrasparc II : octo core (1 FPU / cœur)
  - IBM : power 6 : dual core (2 FMA / cœur)
- Unités de calcul hétérogènes :
  - AMD : fusion (multicoeur CPU/GPU)
  - IBM : cell : octo core (1 power PC (système) et 8 SPU (calcul))
  - Accélérateurs matériels : FPGA (*field-programmable gate array*, réseau de portes programmables in-situ), processeurs graphiques (NVIDIA) .. ➔ mais problèmes 64 bits et communication avec l'hôte (BP)

Mémoire hiérarchique

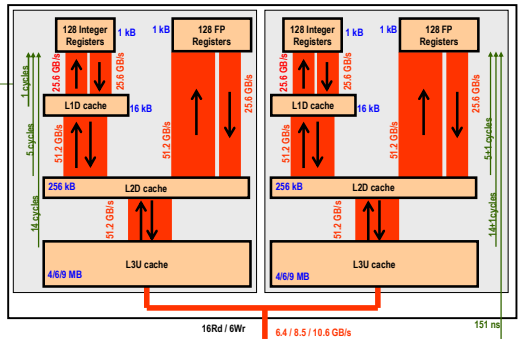


Hiérarchisation des mémoires : exemple l'itanium II (madison)



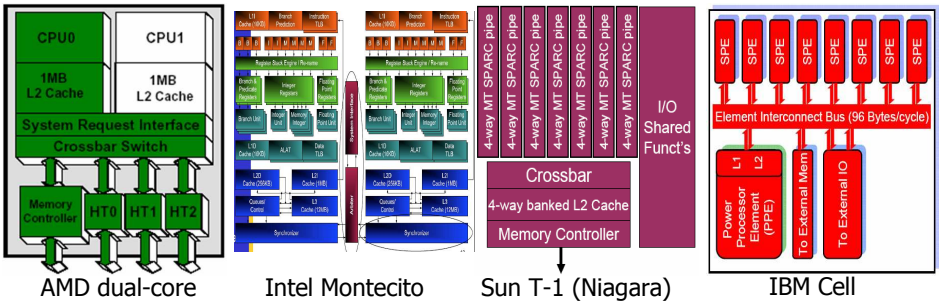
Itanium™2 Madison Data Flow (1.3 - 1.5 GHz)

Itanium™2 Montecito@1.6GHz Data Flow



JoSy : calcul

Mémoires hiérarchiques : exemples d'implémentation

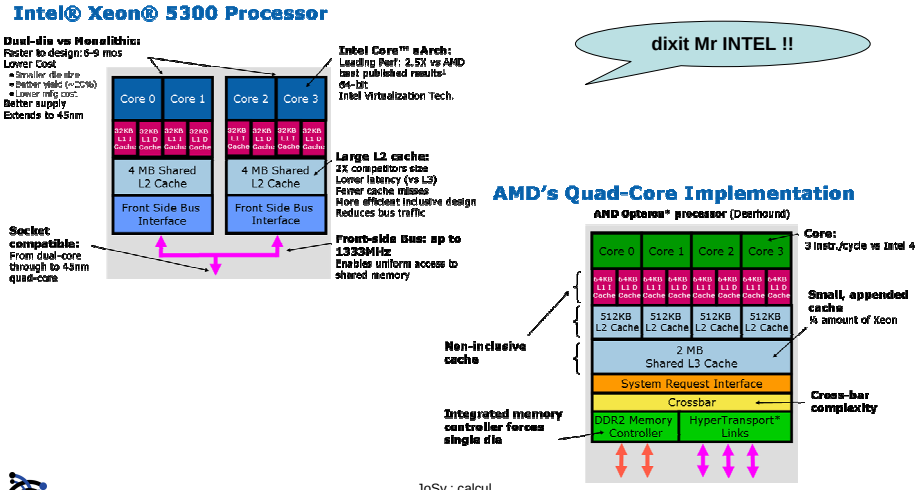


AMD dual-core | Intel Montecito | Sun T-1 (Niagara) | IBM Cell

- AMD dual-core :2 coeurs, mémoire partagée, pas de cache commun
- Intel Montecito : Mémoire partagée, pas de cache partagé, SMT (2)
- Sun T-1 : Cache partagé, 8 coeurs, MT (4)
- IBM Cell : Mémoire distribuée, 32 bits FP, 8 coeurs

JoSy : calcul

Mémoires hiérarchiques : exemples d'implémentation



mémoires

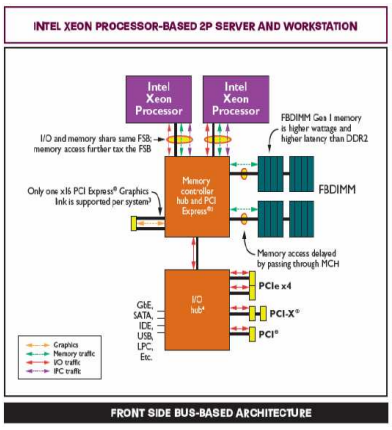
- DDR2 (667 ou 800 MHz), DDR3 (opteron)
  Le successeur de la DDR2 devrait conserver le même format (240 pin), améliorer la bande passante (jusqu'à 10.6 GB/s) et la consommation énergétique, au prix d'une latence plus grande.
- Fully-Buffered DIMM (FBD) (intel) : 533 ou 667 Mhz, Bande passante : 17 ou 21 Go/s en lecture , 8.5 ou 10.5 Go/s en écriture
  - permet des accès simultanés Lecture /écriture à haute vitesse indépendamment du nombre de slots et de la capacité installée.
  - DDR2 (ou DDR3) qui intègre un Advanced Memory Buffer (AMB). L'AMB sert d'intermédiaire entre le contrôleur mémoire et la mémoire et permet d'augmenter la fiabilité de la transmission

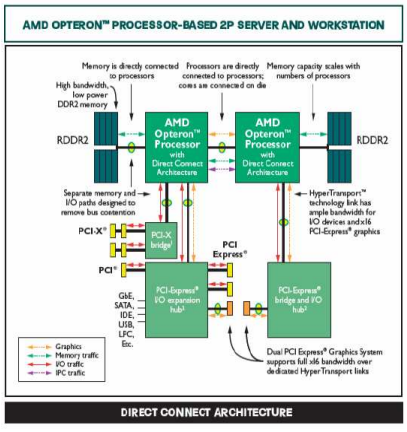
Les différents modes d'interconnexion

- PCI-X
- Bus série full duplex : PCI-Express
- Bus série/parallele (interconnexion de processeurs) : Hypertransport (AMD), CSI (Common System Interface) (INTEL)
- Bus propriétaires (cray, sgi)

Limite de la bande passante sur PCI-X: 64 Bit x 133 MHz = 1056 MB/s
Limite de la bande passante sur PCI-Express (V1 : 250 MB/s par canal ; en V2 : 500 MB/s par canal):
2X : 1 GB/s (soit 8 Gb/s)
4X : 2 GB/s (soit 16 Gb/s)
Limite de la bande passante sur hypertransport : (V3 en 2,6 GHz) 20.8 GB/s
(cf http://en.wikipedia.org/wiki/List_of_device_bandwidths)

Bus d'interconnexion : exemples d'implémentation





Disques : contrôleur / bus

- SCSI Ultra 320: 2.5 Gbps (320 Mo/s mais problèmes récurrents de signaux sur cette dernière évolution du protocole). Limité à l'attachement direct.

  > 180 à 200 Mo/s (si PCI-X 64 bits/100 MHz) (en PCI 100 Mo/s - crete 132 Mo/s).

- SAS (Serial SCSI) : 3 Gbps par voie, généralement disponible en 2 ou 4 voies soit 6 Gbps ou 12 Gbps (remplace le SCSI, corrige ses problèmes de stabilité, et permet d'aller plus loin en terme de performance et de longueur de cable). Limité à l'attachement direct.

  sur bus PCI-X 64 bits 100 (crete 800 Mo/s) ou 133 MHz (crete 1064 Mo/s). Si 12 disques 890 Mo/s soutenus sur un lien (en bus PCI-E on reussit avec plus de disques a atteindre 1 Go/s).

- FC: dernière évolution 4 Gbps. Offre la plus grande souplesse en terme d'interconnexion: SAN ou attachement direct.

  sur un lien, 400 Mo/s. (env 1 Go/s sur une carte multiport)

Disques : performances

- Lecture séquentielle :
  - 7.2K RPM SATA : Max 70-75 MB/s
  - 15K RPM SAS : Max 130 MB/s.
  - 10K RPM FC : Max 90 MB/s.
  - 15K RPM FC : Max 130 MB/s.
- Lecture aléatoire :
  - SATA : < 5Mo/s
  - autres : entre 15 et 20 Mo/s

Technologies réseau

Technology	**Vendor**	**MPI latency** usec, short msg	**Bandwidth per link** (unidirectional, MB/s)
NUMAlink 4	SGI	1	3200
QsNet II	Quadrics	Entre 1.3 et 2	900
Infiniband (si PCI-E 8X (4X : 10 Gb/s)	Ex : Pathscale infinipath	1.3	953
High Performance Switch	IBM	5	1000
Myrinet XP2	Myricom	5.7	749
Ethernet 10Gb		Env 10	862
Ethernet 1Gb		>40	50 à 100

En conclusion

- Attention aux « goulots d'étranglement » (processeurs multi core) (FSB, Mémoire, bus, réseaux, disques)
- Importance des caches
- Performances fortement « code dépendant » et charge de la machine

Les autres critères importants pour une machine de calcul :
- Fiabilité (MTBF global) matériel et stabilité system/logiciel
- Énergie
- Portage applications

Attention au coût réel de la solution !

LA question récurrente : « que faire à prix constant ? »
-plus de noeuds ?
-plus de cœurs par processeur ?
-plus de processeurs par noeuds ?
-meilleur réseau d'interconnexion ?

Pas de réponse standard !

Références et remerciements

- www.Intel.com
- www.amd.com
- en.wikipedia.org
- www.transtec.com
- Mathrice.org
- Ciment.ujf-grenoble.fr
- www.ybet.be
- http://www.guideinformatique.com/fiche-normes_disques-328.html
- « Architectures et Systèmes des Calculateurs Paralleles » cours de François PELLEGRINI (ENSEIRB)

Remerciements : François Bodin (IRISA), Françoise Roch (observatoire de Grenoble), Bruno Leconte (SGI), Jacques Rolland (Caliseo)