Introduction

Module 1

INF8601 Systèmes informatiques parallèles

Michel Dagenais

École Polytechnique de Montréal
Département de génie informatique et génie logiciel

Sommaire

Introduction

Le cours

- Site Moodle
- Situation du cours
- Objectifs du cours et des laboratoires
- Déroulement du cours
- Evaluation
- Lectures
- Contenu du cours
- Horaire

Evolution technologique: 5 générations

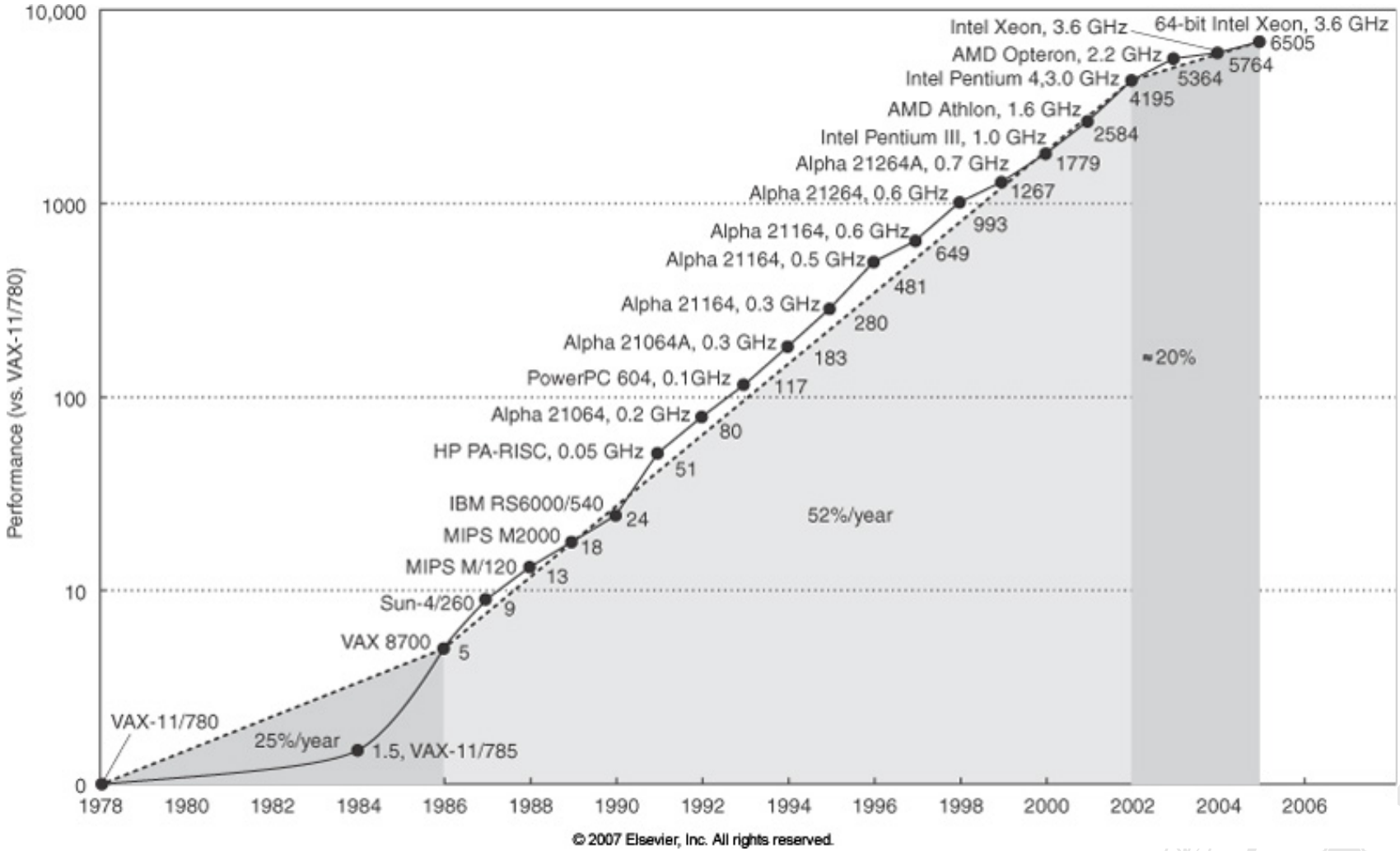
- 1940 - Lampes
- 1960 - Transistors
- 1970 - Circuits intégrés
- 1980 - Circuits LSI/VLSI
- 2000 - Ordinateurs de 5ème génération, parallèles, intelligents...

Introduction

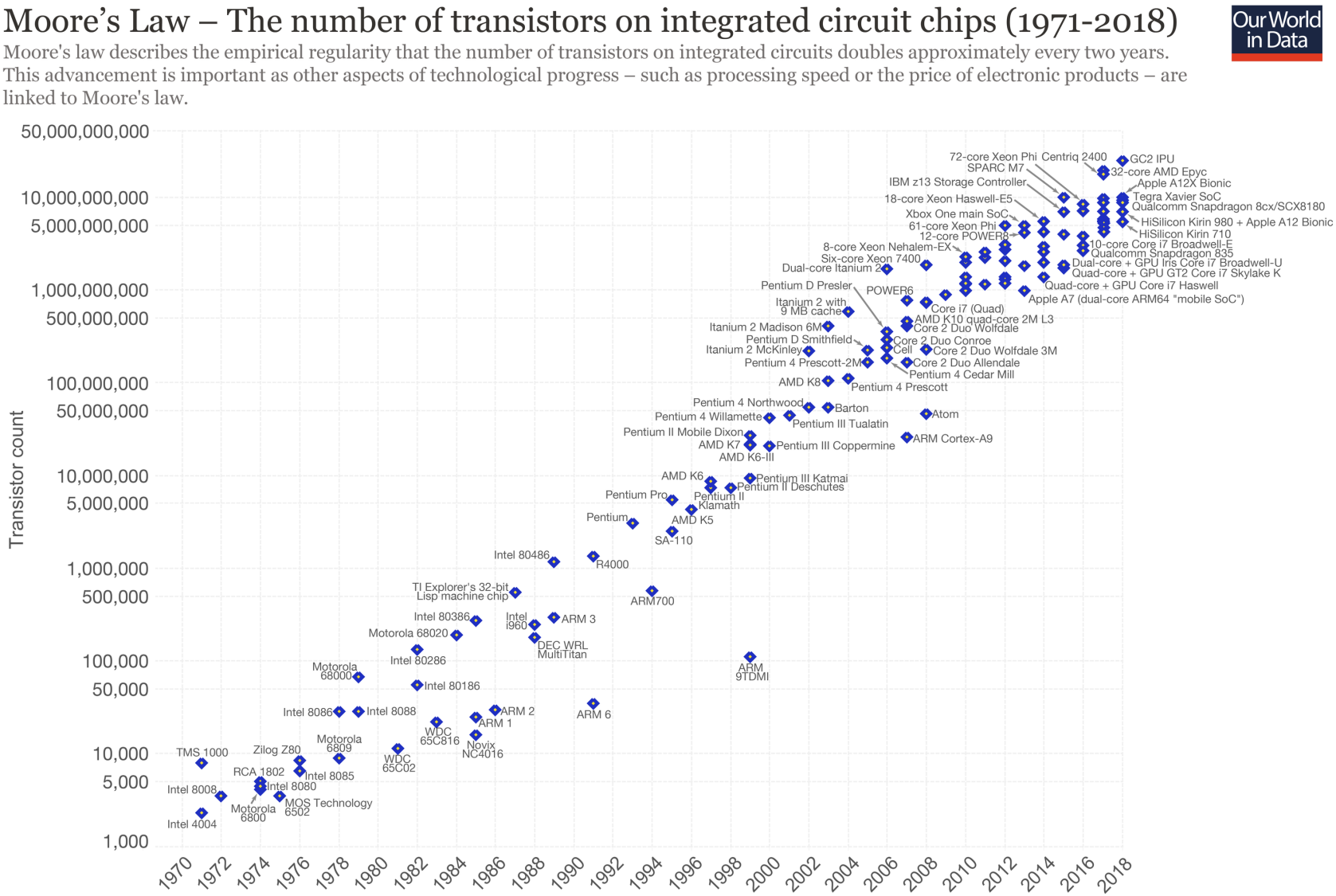
Evolution technologique

- Circuits logiques: 60% plus de transistors par an; la rapidité augmente presqu'en proportion de la densité;
- Mémoire vive: 60% plus dense à chaque année, vitesse 30% seulement par 10 ans;
- Mémoire disque: 50% plus dense à chaque année, vitesse 30% seulement en 10 ans.

Vitesse des processeurs



Vitesse des processeurs



Périphériques

- Interface réseau: 1Gbit/s, 12$;
- Disque: 6TB, 750MB/s, $150;
- Mémoire non volatile: 1TB, 750MB/s, $300;
- Rubans: LTO-5, 3TB, $2000;
- Carte graphique: GP100 Pascal, 16nm, 15300M transistors, 3584 coeurs;

Répertoires d'instruction

- IBM 360
- DEC VAX 11
- Motorola 68000
- Intel 80386
- Sun SPARC
- IBM PowerPC
- ARM
- Intel Itanium
- Intel 64 (AMD64)

Classes d'ordinateurs

- Super-ordinateur, entre 5 et 20M$, grande pièce dédiée;
- Serveur d'entreprise (mainframe), entre 1 et 5 M$, coin de pièce;
- Serveur départemental, occupe un coin de pièce, entre 50K$ et 1 M$;
- Poste de travail (workstation), dessus de bureau, 4K$ à 50K$;
- Micro-ordinateur, dessus de bureau ou creux de la main, 100 à 4K$;

Nouvelles classes d'ordinateur

- Le nuage, grappe d'ordinateurs flexible et efficace avec virtualisation, ressources à la demande et migration.
- Clients, de table, portatif, tablette, cellulaire.
- Bon rapport performance prix, importance de la consommation d'énergie, tout est un système parallèle, hétérogène en réseau!

Tendances

- Plafonnement de la vitesse par processeur (taille des transistors, voltage, dissipation de chaleur);
- Augmenter le nombre de processeurs, meilleure performance / prix, consommation flexible;
- La mémoire centrale et les disques n'augmentent pas de vitesse autant que le processeur;
- Client capable d'exécuter des jeux 3D et du vidéo HD;
- Complexité des systèmes: outils de monitoring, vulnérabilités nombreuses.

Historique des systèmes parallèles

- Burroughs D825, 1962, 4 processeurs;
- Honeywell Multics System, 1969, 8 processors;
- ILLIAC IV, 1965-1976, Vector machine, 256 units;
- Cray 1, 1976, Vector machine;
- Cray 2, 1985, 8 processors;
- Multi-processeurs (SGI, SUN, IBM);
- Grappes de calcul (Intel, IBM, HP...);
- Grappes de multi-processeurs hétérogènes (GPGPU);

Programmation des systèmes parallèles

- FORTRAN, compilateurs parallélisants, HPF;
- C, Parallel C, UPC, OpenMP (C et FORTRAN);
- PVM, Linda, MPI;
- CUDA, OpenCL

Introduction

Hiérarchie de mémoire

- De haut (vite) en bas (dense): registres, cache L1/L2/L3, mémoire vive (NUMA), disque SSD, disque magnétique;
- L'adresse se décompose en numéro de bloc et en adresse dans le bloc;
- Taux de succès: proportion du nombre d'accès complétés au niveau le plus haut sans interaction avec le niveau inférieur;
- Taux d'échec: (1-succès);
- Temps d'accès succès: temps d'accès lorsque c'est complété au niveau le plus haut;
- Pénalité d'échec: temps pour compléter la communication entre les deux niveaux et transmettre la donnée au niveau supérieur (CPU), (accès, transfert complété), outre le temps d'accès succès;
- Temps d'accès moyen = temps d'accès succès + taux d'échec * pénalité d'échec

Mémoire cache

- Blocs de 4 à 128 octets;
- Temps d'accès succès 1-4 cycles;
- Pénalité d'échec 8-32 cycles (accès 6-10), (transfert 2-22);
- Taux d'échec 1% à 20%;
- Dimension 1K à 512K (Intel Core I5, 32K I et 32K D);
- Correspondance directe, associative, par ensemble;
- Remplacement aléatoire ou LRU;
- Ecriture simultanée ou réécriture;

Mémoire centrale

- Pages de 4K, 2M ou 4M octets;
- Temps d'accès succès 10-100 cycles;
- Coût de l'échec 500000-6000000 cycles (quelques dizaines de milisecondes);
- Taux d'échec .00001%-.001%;
- Table de pages à 1, 2, 3 ou 4 niveaux, (associatif, remplacement LRU);
- Conversion adresse logique à adresse physique par matériel avec cache (TLB);
- Chargement des pages par le système d'exploitation;
- Une portion du système d'exploitation est toujours en mémoire centrale.

Le système d'exploitation

- Interface avec les périphériques (pilotes);
- Interface avec les processus (appels systèmes);
- Gestion de la mémoire (mémoire des processus, mémoire virtuelle, tampons E/S);
- Systèmes de fichiers;
- Contrôle des accès;
- Ordonnancement;
- Gestion des interruptions.

Taxonomie de Flynn

- Flux d'instructions et de données.
- SISD: ordinateur conventionnel.
- SIMD: une opération peut opérer sur plusieurs données parfois de manière conditionnelle (Connection Machine 65536 processeurs de 1 bit, Ordinateur vectoriel, GPGPU).
- MIMD: SMP, multi-core, many-core, grappe, grille, nuage...

Parallélisme

- La même opération sur des données différentes.
- Le même programme sur des données différentes.
- Différents programmes sur les mêmes données ou des données différentes.

Partage de données

- Mémoire partagée uniforme.
- Mémoire partagée répartie, accès non-uniforme (NUMA).
- Mémoire partagée répartie, par logiciel.
- Envoi de messages.

Fiabilité

- MTBF: Mean Time Between Failure. Temps moyen entre les pannes.
- MTTF: Mean Time to Failure. Temps moyen jusqu'à la panne.
- MTTR: Mean Time to Repair. Temps moyen pour réparer la panne.
- Fiabilité: F = MTBF= MTTF + MTTR
- Disponibilité: D = MTTF / (MTTF + MTTR)

Fiabilité

- Déterminer un taux de panne (failures in time), FIT = 1 / MTTF, exprimé en nombre de pannes par milliard d'heures (environ 114000 ans).
- Les taux de pannes s'additionnent si la probabilité de panne ne dépend pas de l'âge (accident aléatoire plutôt qu'usure).
- Exemple: MTTF total pour 10 disques (MTTF=1 Mh), 1 contrôleur (MTTF= 0,5 Mh), 1 alimentation (MTTF= 0,2 Mh), 1/(10/1Mh + 1/,5MH + 1/.2MH) = 58823 heures

Introduction

Calcul de disponibilité

- Probabilité qu'un composant soit fonctionnel au temps t: $\frac{MTTF}{MTTF+MTTR}$ (Mean Time to Fault / Repair).
- Probabilité p que deux événements indépendants, de probabilité p_1 et p_2, se produisent en même temps: $p = p_1 \times p_2$.
- Probabilité p que p_1 ou p_2 se produise: $p = p_1 + p_2 - p_1 \times p_2$
- Les probabilités de cas disjoints ($p_1 \times p_2 = 0$) peuvent s'additionner.
- Probabilité que k disques sur n soient fonctionnels ($n - k$ en panne), si la propabilité pour un disque est p_d.

$$p = \binom{n}{k} \times p_d^k \times (1 - p_d)^{n-k}$$

  - Sommation de k à n pour k disques ou plus fonctionnels.
  - Au moins 1 disque fonctionnel dans un miroir: $1 - p_d^2$

Probabilités: peut-on sommer

- Probabilité que deux événements indépendants de probabilité p1 et p2 se produisent au même moment: $p1 \times p2$. Par exemple, un ordinateur est fonctionnel si la carte mère (pm) et le disque (pd) sont fonctionnels. Il y a une panne sauf si les deux sont fonctionnels.

$$1 - p_m \times p_d$$

- Peut-on plutôt prendre la probabilité de panne de carte + la probabilité de panne de disque?

$$(1 - p_m) + (1 - p_d) \neq 1 - p_m \times p_d$$

Probabilités: sommer si disjoint

- Quelle est la différence? La panne de disque se produit indépendamment de la carte qui peut être fonctionnelle ou non. On compte donc deux fois le cas où le disque et la carte sont défectueux. Les deux éléments sommés se recoupent alors qu'il faut bien tout séparer. Si p1 et p2 sont grands, la partie sommée en double est petite et le résultat est incorrect mais très proche. Si p1 et p2 sont petits, le résultat dépasse largement 1!

$$P(p_1 \cup p_2) = p_1 + p_2 - p_1 \times p2 =$$

$$(1 - p_m) + (1 - p_d) - ((1 - p_m) \times (1 - p_d)) =$$

$$(1 - p_m) + (1 - p_d) - (1 - p_d - p_m + p_m \times p_d) = 1 - p_m \times p_d$$

Systèmes parfaitement tolérants aux pannes?

- Avec le bon niveau de redondance, on peut diminuer les probabilités de panne pratiquement à 0, sous certaines hypothèses.
- Cinq 9 (0.99999) donne 5.26 minutes par an de panne, six 9 donne 31.5 secondes par an.
- En pratique, les pannes sont plus fréquentes car certaines hypothèses ne tiennent pas, surtout sur les nouveaux systèmes.
- Pour les systèmes plus matures, on analyse les accidents et on apprend des erreurs (e.g., en aviation).

Disques en miroir

- Un disque tombe en panne environ une fois par 4 ans (1h/4ans = 0.00003). Avec 2 disques en miroir les probabilités de pannes simultanées sont à peu près nulles ($.00003^2$ = 1h/150000ans). Plus besoin de copies de sauvegarde?!?
- Lors d'une panne, il faut être alerté et effectuer la réparation rapidement pour revenir à une configuration redondante.
- Le matériel redondant doit être en bon état de marche.
- En fin de vie, les probabilités de panne augmentent rapidement. Deux vieux disques deviennent problématiques.
- Il y a parfois des lots de disques non fiables avec fin de vie prématurée. Plusieurs vont acheter des disques de lots / modèles différents.
- Vol, feu, foudre, surtension...

Introduction

Performance

- Temps écoulé.
- Temps d'utilisation (CPU, mémoire, disque...).
- Rendement, résultat/temps par rapport aux ressources utilisées.
- Accélération de performance A = performance avec amélioration / performance sans amélioration (ou temps sans amélioration / temps avec amélioration)
- Accélération partielle, (fraction f du traitement), et totale résultante, loi d'Amdahl. At = 1 / ((1-f) + f / Ap).

Performance parallèle

- Facteur d'accélération, S(p) = Temps séquentiel (optimal) / Temps avec p processeurs.
- Efficacité, E(p) = S(p) / p
- Temps de réponse optimal versus coût optimal...
- Si fraction parallélisable f accélérée de p, S(p) = 1 / ((1-f) + f/p), S(p) tend vers 1/(1-f) si p très grand.

Bancs d'essais

- Linpack (Top500.org).
- SPEC2006: 12 entiers (9 C, 3 C++), 17 point flottant (6 FORTRAN, 4 C++, 3 C, 4 C et FORTRAN).
- TPC: Transactions.

Top 500.org

- Ordinateurs les plus puissants au monde;
- Modélisation de phénomènes physiques (météo, turbines, fuselages, réactions chimiques/nucléaires, déchiffrement);
- Ordinateurs multi-coeurs, parallèles;
- Architecture? OS? Pays?