Chap 1; But et définition du parallélisme

Définition

Le parallélisme est le fait d'avoir *plusieurs processeurs* (threads) qui *collaborent* à la solution d'un *même* problème #### But du parallélisme Les performances de calcul

Le parallélisme est un processus naturel de collaboration, aussi dans la vie de tous les jours.

Exemples

- La muraille de chine
- Calculs astronomiques au XIX siècle
- Calcul météo dans les années 1920
- ENIAC, 1 des premiers ordinateurs (plusieurs unités de calcul activables en même temps... Peu de succès à cause de la fiabilité matériel et de la difficulté de coordination)

En parallélisme, le programmeur est responsabe de coordonner tous les PE (Processing Element) C'est un peu comme un chef d'orchestre qui donne des partitions cohérentes à tous les musiciens.

Une difficulté avec le parallélisme est d'assurer une forte cohérence entre le modèle de programmation et le matériel.

Le modèle *séquentiel* de Von Newmann est un exemple d'un accord parfait entre matériel et logiciel. (Instructions -> CPU -> mémoire) On a donc une convergence entre le hardware et la façon de le programmer; suite d'instruction qui modifie les données. -> succès de l'architecture séquentielle durant de nombreuses années (car moins de problèmes ?) ##### Mais on veut toujours plus de performances ! (On est jamais satisfait et on demande toujours plus de ressources...) ##### Comment mesure-t-on les performances: flop/s (floating point or op per sec) ##### Actuellement: machines les plus puissantes: Exaflop: 10^18 -> 8.7 million de coeurs -> 20MW (Jet d'eau de Genève: 0.5 MW) (Suisse: 4ème jusqu'à 2020 course machine plus puissante) HPC: High Performance Computation

Limite de l'Architecture Von Neumann

- 1ère cause: chaleur dissipée limite la fréquence d'horloge -> les circuits fondent
- 2ème cause: séparation mémoire-CPU -> Il faut amener les données au CPU. C'est environ 10x plus cher en énergie qu'un calcul en virgule flottante, voire jusqu'à 1000x en fonction de la distance mémoire-CPU. ##### Cela s'appelle le *memory-wall*, ou le goulett d'étranglement de Von Neumann Cette limitation est illustrée par ce qu'on appelle le 'roofline

model' (but: relier la puissance de calcul R (Rate: flop/s) avec l'intensité arithmétique I [flop/byte], avec aussi la bande passante mémoire CPU Beta [byte/s]. Cela donne un graphe:

```
/  <- bandwith Beta
```

Rmax| /_____ |/.| CPU-bound and .:memory bound
|——-> R I

I grand: la même donnée est utilisée beaucoup de fois pour le calcul (voir cours pour plus de détails)

Où trouver des performances ?

Technologie: Processus physique utilisé pour traîté l'information

- Actuellement, les semi-conducteurs au silicium...
- Pendant longtemps, on a observé ce qu'on appelle la 'loi de Moore': les performances d'une machine double tous les 18 mois à coût identique (loi empirique). Mais on commence à en voir le déclin...
- Pendant longtemps, il y a eu beaucoup de lois exponentielles ### Limites
- Chaleur dissipée (p): p = v^2f avec v: voltage au bornes des transistors et f la fréquence d'horloge Si on augmente f, il faut donc diminuer le voltage v Cela peut se faire en miniaturisant les circuits. -> finesse du trait (delta: feature size ~ nm) de la photolithogravure -> effet tunnel (ordre des atomes -> physique quantique...) Mais si delta est trop petit, il y a des effets quantiques:
  - effet tunnel, perte de l'isolant.
  - le nombre d'électrons qui distingue l'état 1 de l'état 0 doit être assez grand >=1