Chimie quantique *vs* machines parallèles

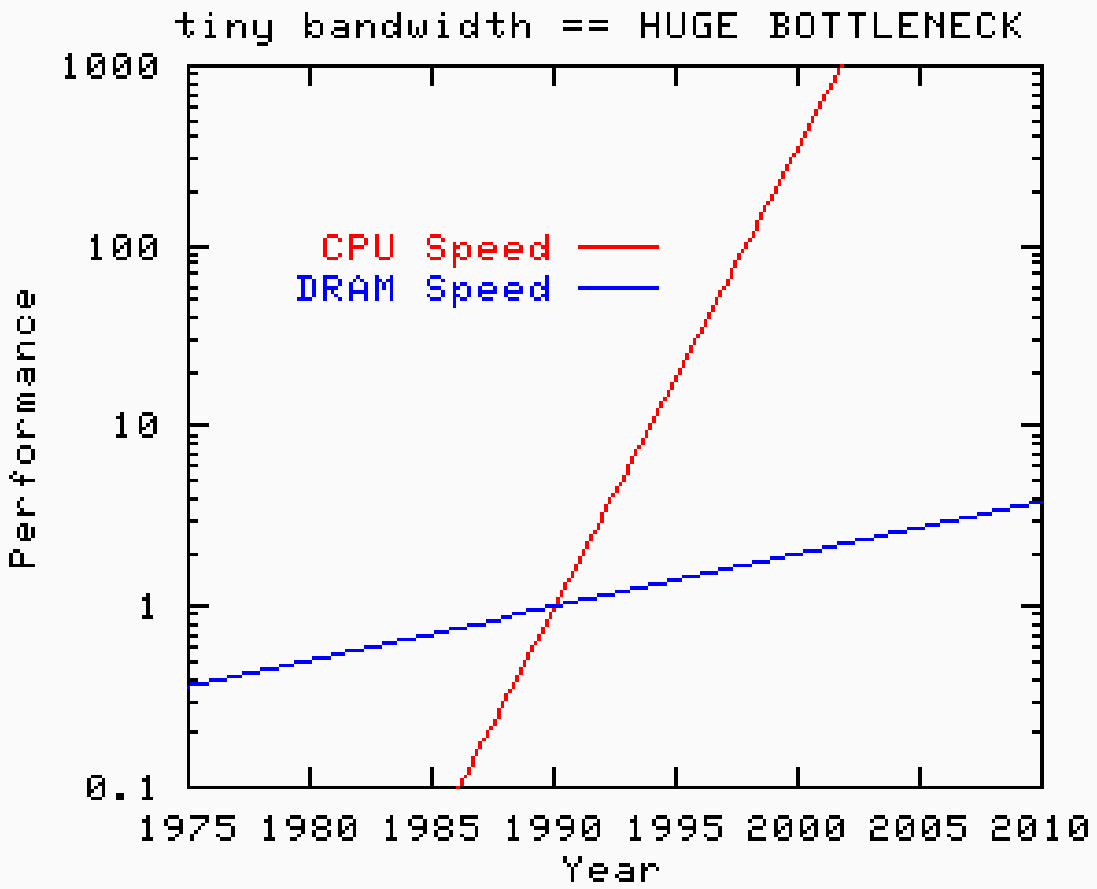
Anthony Scemama

11/10/2018

Lab. Chimie et Physique Quantiques, IRSAMC, UPS/CNRS, Toulouse

Problèmes liés aux machines modernes

Le “Mur” de la mémoire



- CPU : $\times 1.55$ / an
- Mémoire : $\times 1.10$ / an

Stream benchmark : https://www.cs.virginia.edu/stream

Importance des accès aux données

```
for (i=0 ; i<n ; i++)
  for (j=0 ; j<n ; j++)
    dist[i][j] = (X[0][i]-X[0][j])*(X[0][i]-X[0][j]) +
                 (X[1][i]-X[1][j])*(X[1][i]-X[1][j]) +
                 (X[2][i]-X[2][j])*(X[2][i]-X[2][j]) ;
```

- `n`=100 (78 KiB) : 1.3 cycles / `dist[i][j]`
- `n`=4000 (125 MiB) : 3.0 cycles / `dist[i][j]`
- `n`=40000 (12.2 GiB) : 4.7 cycles / `dist[i][j]`

Importance des accès aux données

```
for (i=0 ; i<n ; i++)
  for (j=0 ; j<i ; j++)   // 2 fois moins de flops
    dist[i][j] = (X[0][i]-X[0][j])*(X[0][i]-X[0][j]) +
                 (X[1][i]-X[1][j])*(X[1][i]-X[1][j]) +
                 (X[2][i]-X[2][j])*(X[2][i]-X[2][j]) ;
for (i=0 ; i<n ; i++)
  for (j=i+1 ; j<n ; j++)
    dist[i][j] = dist[j][i];
```

- n=100 : 1.9 cycles : $\times$1.46
- n=4000 : 18.9 cycles : $\times$6.3
- n=40000 : 35.3 cycles : $\times$7.5

Importance des accès aux données

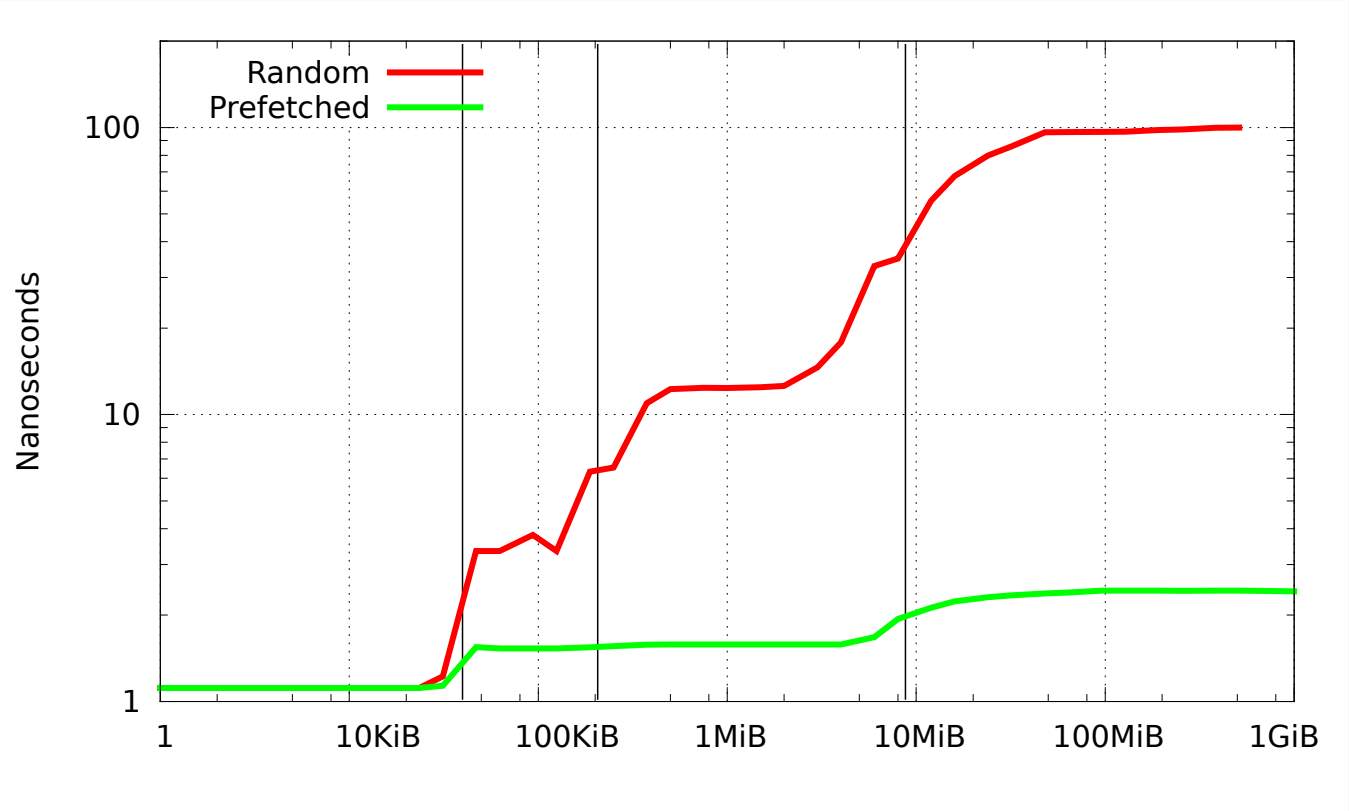
```
for (i=0 ; i<n ; i++)
  for (j=0 ; j<i ; j++)   // 2 fois moins de flops
    dist[i][j] = (X[0][i]-X[0][j])*(X[0][i]-X[0][j]) +
                 (X[1][i]-X[1][j])*(X[1][i]-X[1][j]) +
                 (X[2][i]-X[2][j])*(X[2][i]-X[2][j]) ;
for (i=0 ; i<n ; i++)
  for (j=i+1 ; j<n ; j++)
    dist[i][j] = dist[j][i];
```

- n=100 : 1.9 cycles : $\times$1.46
- n=4000 : 18.9 cycles : $\times$6.3
- n=40000 : 35.3 cycles : $\times$7.5

Flops are free!!!

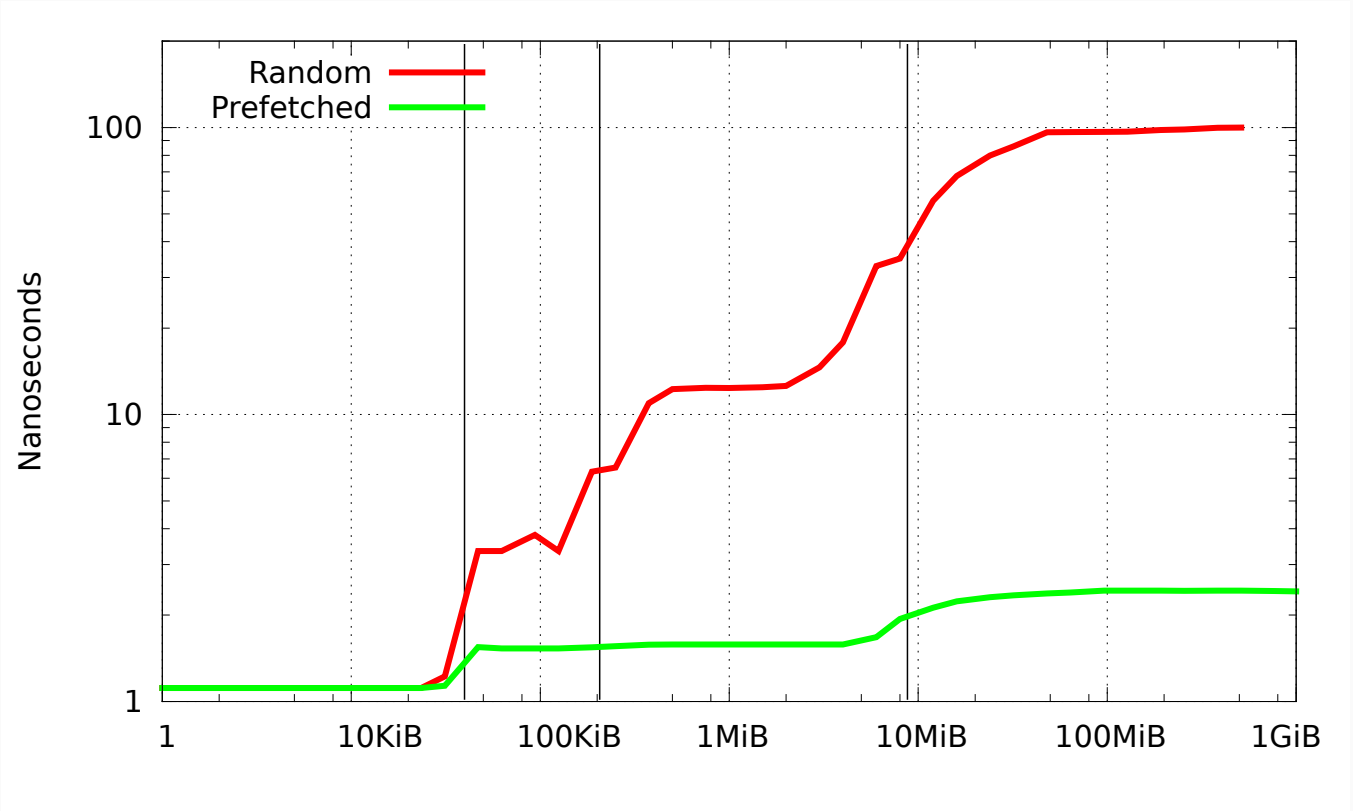
Latence (Nanosecondes)

Accès dans un tableau de taille croissante:



Latence (Nanosecondes)

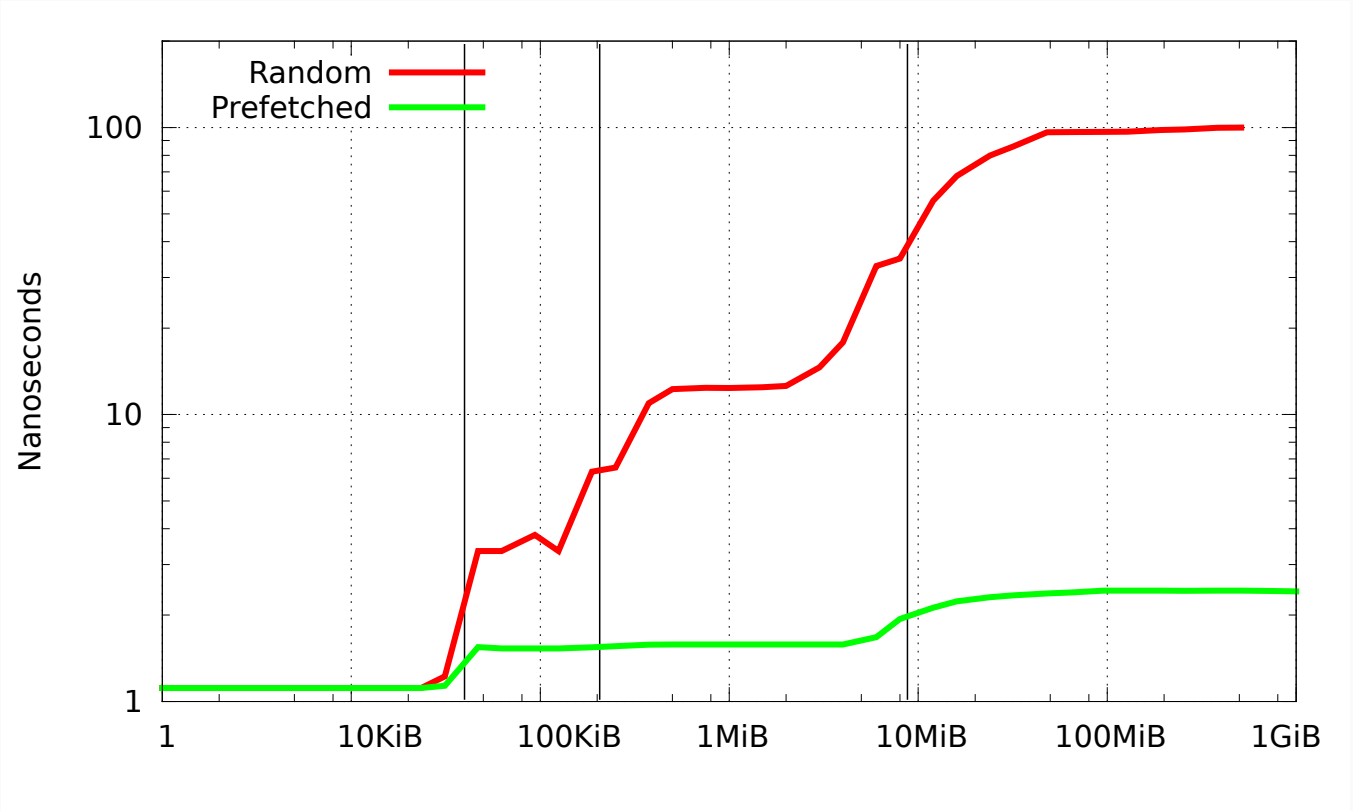
Accès dans un tableau de taille croissante:



2GHz ⇔ 0.5 ns/cycle AVX-512 ⇔ 32 flops/cycle (DP)

Latence (Nanosecondes)

Accès dans un tableau de taille croissante:



2GHz ⇔ 0.5 ns/cycle AVX-512 ⇔ 32 flops/cycle (DP)

Flops are free!!!

Latence

- Accès aléatoire à la mémoire très coûteux : 240-300 cycles $\Leftrightarrow$ 7 500–10 000 flops DP!
- Accès régulier à la mémoire : déclenche les prefetchers hardware pour masquer la latence

D'autres nombres

Mutex lock/unlock	100 ns
Infiniband (RDMA)	1 200 ns
Ethernet (TCP/IP)	50 000 ns
Disk seek (SSD)	50 000 ns
Disk seek (15k rpm)	2 000 000 ns

Bande passante

Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz

`a(i) = a(i) + b(i)*c(i)`

- 2 FMA vectoriels par cycle : `a = a + b*c`
  vecteurs de 8 flottants en DP (8×8 octets)

Bande passante

Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz

`a(i) = a(i) + b(i)*c(i)`

- 2 FMA vectoriels par cycle : `a = a + b*c`
  vecteurs de 8 flottants en DP (8×8 octets)
- 16 octets par flop

Bande passante

Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz

`a(i) = a(i) + b(i)*c(i)`

- 2 FMA vectoriels par cycle : `a = a + b*c`
  vecteurs de 8 flottants en DP (8×8 octets)
- 16 octets par flop
- Puissance crête
  2.3 GHz $\times$ ($2_{\text{FMA}} \times 2_{\text{vector units}} \times 8_{\text{vector size}}$ flops) $\times$ 18 coeurs = 1.325 TFlops/s (DP)

Bande passante

Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz

`a(i) = a(i) + b(i)*c(i)`

- 2 FMA vectoriels par cycle : `a = a + b*c`
  vecteurs de 8 flottants en DP (8×8 octets)
- 16 octets par flop
- Puissance crête
  2.3 GHz $\times$ ($2_{\text{FMA}} \times 2_{\text{vector units}} \times 8_{\text{vector size}}$ flops) $\times$ 18 coeurs = 1.325 TFlops/s (DP)
- Bande passante nécessaire :
  1.325 Tflops/s $\times$ 16 o/flops = 21.2 To/s

Bande passante

Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz

`a(i) = a(i) + b(i)*c(i)`

- 2 FMA vectoriels par cycle : `a = a + b*c`
  vecteurs de 8 flottants en DP (8×8 octets)
- 16 octets par flop
- Puissance crête
  2.3 GHz $\times$ ($2_{\text{FMA}} \times 2_{\text{vector units}} \times 8_{\text{vector size}}$ flops) $\times$ 18 coeurs = 1.325 TFlops/s (DP)
- Bande passante nécessaire :
  1.325 Tflops/s $\times$ 16 o/flops = 21.2 To/s
- Bande passante mémoire max
  6 canaux $\times$ 2666 MHz $\times$ 8 octets = 128 Go/s

Bande passante

Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz

`a(i) = a(i) + b(i)*c(i)`

- 2 FMA vectoriels par cycle : `a = a + b*c`
  vecteurs de 8 flottants en DP (8×8 octets)
- 16 octets par flop
- Puissance crête
  2.3 GHz $\times$ ($2_{\text{FMA}} \times 2_{\text{vector units}} \times 8_{\text{vector size}}$ flops) $\times$ 18 coeurs = 1.325 TFlops/s (DP)
- Bande passante nécessaire :
  1.325 Tflops/s $\times$ 16 o/flops = 21.2 To/s
- Bande passante mémoire max
  6 canaux $\times$ 2666 MHz $\times$ 8 octets = 128 Go/s

La bande passante mémoire est $\sim$ 165$\times$ trop faible!

Conclusions

1. La réduction du nombre de flops d'un algorithme ne conduit pas forcément à l'accélération d'un programme.
   Ex: Algorithmes "linear scaling" sont linéaires en stockage et en calcul, donc probablement memory-bound (pré-facteur 165×, *etc*).

Conclusions

1. La réduction du nombre de flops d'un algorithme ne conduit pas forcément à l'accélération d'un programme.
   Ex: Algorithmes "linear scaling" sont linéaires en stockage et en calcul, donc probablement memory-bound (pré-facteur 165×, *etc*).
2. La performance crête d'une machine n'est pas du tout représentative du temps de restitution d'une application scientifique moyenne : elle est de plus en plus difficile à atteindre.

Machines futures

Machine exaflopique (10^{18} flops/s) prévue en 2022

- Unité de mesure : flops/s $\rightarrow$ augmentation de la puissance crête

Machines futures

Machine exaflopique (10^{18} flops/s) prévue en 2022

- Unité de mesure : flops/s $\rightarrow$ augmentation de la puissance crête
- Encore plus de coeurs
- Un peu plus de RAM, donc moins de RAM par coeur

Machines futures

Machine exaflopique (10^{18} flops/s) prévue en 2022

- Unité de mesure : flops/s $\rightarrow$ augmentation de la puissance crête
- Encore plus de coeurs
- Un peu plus de RAM, donc moins de RAM par coeur
- La latence du réseau ne va pas beaucoup réduire
- La fréquence des CPUs va baisser pour réduire la consommation
- Généralisation des accélérateurs (GPU, Xeon Phi, FPGA) : hétérogénéité

Machines futures

Machine exaflopique (10^{18} flops/s) prévue en 2022

- Unité de mesure : flops/s $\rightarrow$ augmentation de la puissance crête
- Encore plus de coeurs
- Un peu plus de RAM, donc moins de RAM par coeur
- La latence du réseau ne va pas beaucoup réduire
- La fréquence des CPUs va baisser pour réduire la consommation
- Généralisation des accélérateurs (GPU, Xeon Phi, FPGA) : hétérogénéité

Ça s'annonce très mal si on ne fait rien...