Evolution des technologies de calcul intensif vers les systèmes multi-cœurs et accélérateurs

Marc Mendez-Bermond

Expert solutions HPC

Programme

- Contexte
- Technologies
- Evolutions

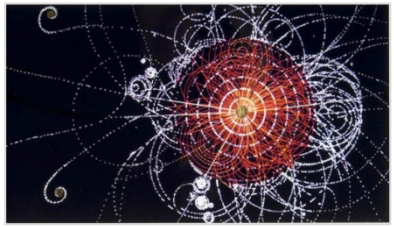
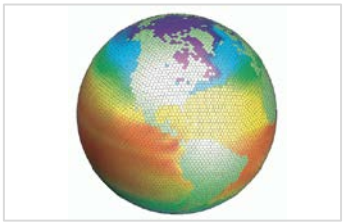
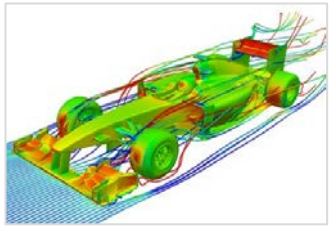
Contexte

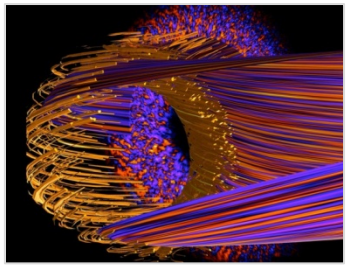
Principes du HPC

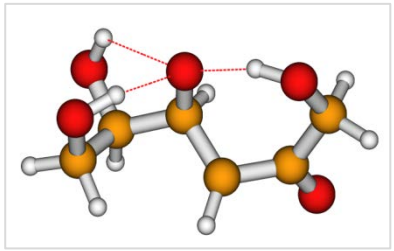
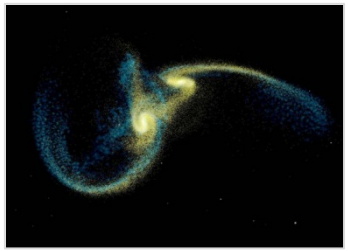
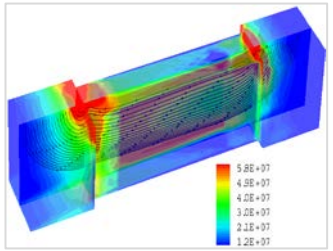
- **Objet**
  - Simuler les phénomènes pour s'affranchir du coût, de la complexité ou de la période des systèmes étudiés
  - Assurer un traitement massif de données
- **Historique**
  - Des systèmes mainframe aux grappes de calcul
  - Loi de Moore, de Amdahl et quelques autres
  - Révolution multi-cœur
  - Système hybrides à coprocesseurs
- **Simulation/prototypage numérique**
  - 3ème discipline après la théorie et l'expérimentation
  - Ne pas oublier le traitement massif de données !

Applications HPC

- Large spectre
- Profils variés
- Sujets cruciaux
- Besoins croissants de puissance

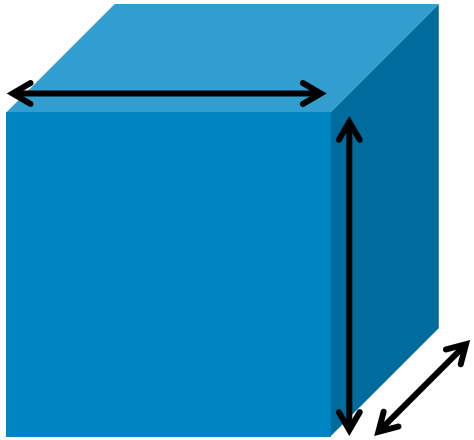


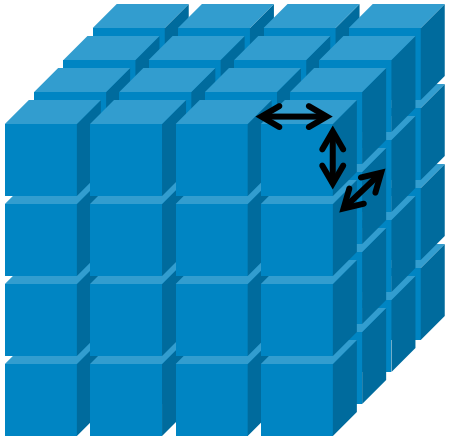
Accélération par parallélisme

- Décomposition en **sous-domaines** selon le phénomène simulé (3D ici)
- **Motif de communications** entre les sous-domaines spécifique

- Permet de soulager :
  - Stress processeur : multiplier leur nombre plutôt que leur performance
  - Quantité de mémoire par système : systèmes plus économiques
- Inconvénients :
  - Travail à fournir pour sortir d'un motif séquentiel
  - Optimisation pour le passage à l'échelle en fonction des objectifs visés
  - Infrastructure logicielle et matérielle complexifiée



1 problème de taille N
Tséq. = X * Y * Z



N problèmes de taille 1
T// = x * y * z + séq. global

Grandes étapes

- **Ere SMP – système SSI NUMA, mainframes**
  - Nombre de processeurs toujours croissant (max. 2048)
  - Complexité des systèmes
  - Technologies propriétaires
- **Ere Beowolf – grappes de calcul x86**
  - Nombre de systèmes grandissant (> 10000)
  - Utilisabilité et complexité d'administration
  - Taux de pannes dus aux nombre de composants
  - Technologies ouvertes
- **Ere hybride – grappes de calcul x86 + autre chose …**
  - Toutes les contraintes du Beowolf
  - Stress du gestionnaire de ressources
  - Difficultés de programmation

Contraintes

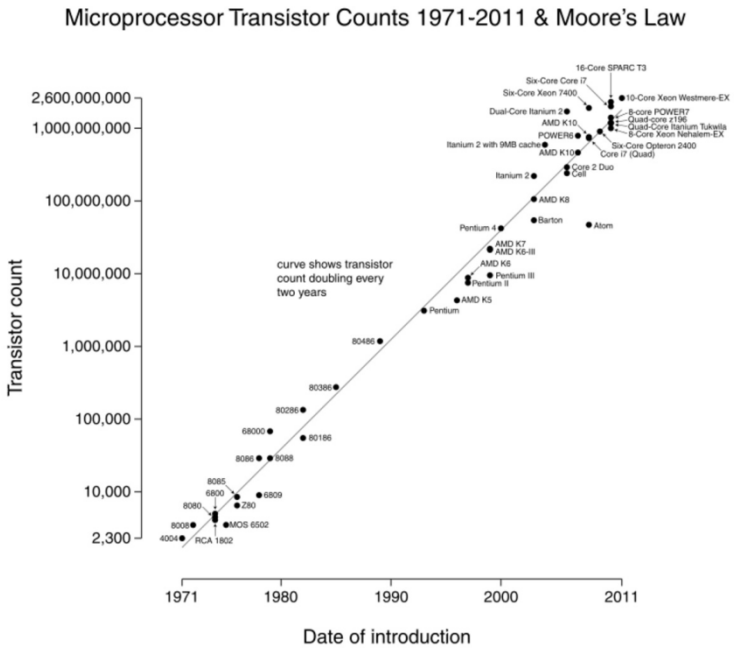
- **Clé du progrès**
  - Adoption soutenue
  - Analyses de gigantesques quantités de données
  - Rapport GES consommés / GES économisés > 10 !
- **Evolutions des technologies existantes**
  - Processeurs, stockage, réseaux
  - Efficacité énergétique en objectif principal
- **Systèmes massivement parallèles**
  - Rupture technologique au bénéfice du rapport perf/conso
  - Nécessité d'adapter l'application et son environnement
  - Evolution planifiée de ces technologies
- **L'outil HPC**
  - Systèmes clé en main ou accompagnement sur le long terme
  - Systèmes dynamiques orientés services : Cloud Computing

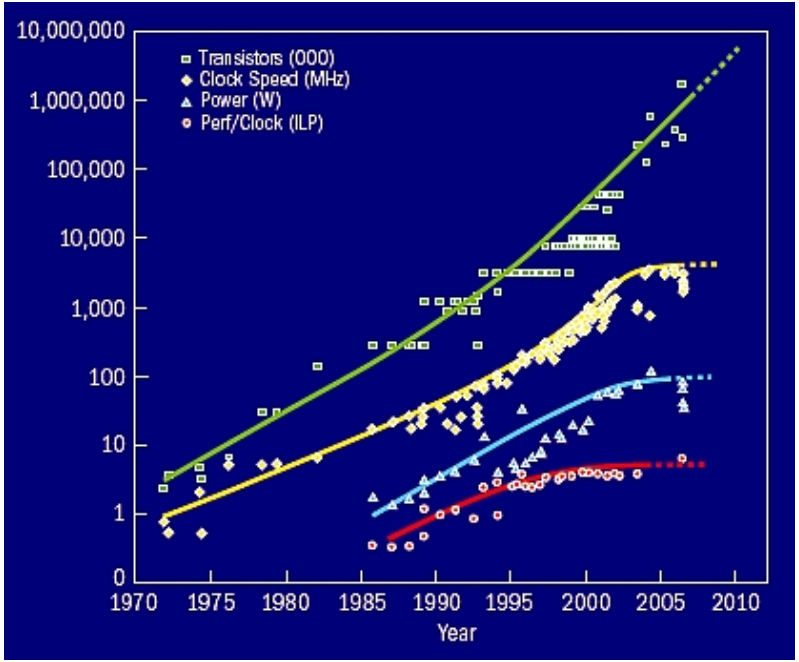
Consommations/dissipations

Une course au MegaWatt

- **L'échelle des systèmes devient ingérable**
  - ✓ > 100000 nœuds de calcul
  - ✓ ~1000000 cœurs de calcul
  - ✓ Rpeak : 27 PFlop/s (Titan)
- **Spécialisation (~)= efficacité**
  - ✓ Haute efficacité: 1 GFlop/s/W (K Supercomputer)
  - ✓ Très haute efficacité : 2 GFlop/s/W (BG/Q)
  - ✓ Ultra haute efficacité : 6 GFlop/s/W (architecture de calcul spécifique)
- **Pour 1 EFlop/s**
  - ✓ Technologies actuelles : 1000 MW
  - ✓ Loi de Moore : 200-300 MW
  - ✓ Cible réaliste : 50 MW max.
- **Dell/TACC Stampede**
  - ✓ Xeon + Xeon Phi
  - ✓ ~1.5 GFlop/s/W réels !

Progrès des processeurs





1. Moore avait raison …
2. Mais sa loi ne définissait que le nombre de fonctions du CI !

Consommation des circuits intégrés

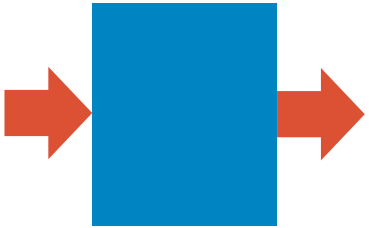
- **$P = a \cdot C \cdot V^2 \cdot f$**
  - a : facteur d'activité, C : capacitance, V : tension de travail, f : fréquence d'opération
- **$P = k \cdot V^3$**
- **$P = x \cdot f^3$**

L'évolution de la puissance consommée évolue au cube de la fréquence !

Optimisation de la consommation

Base

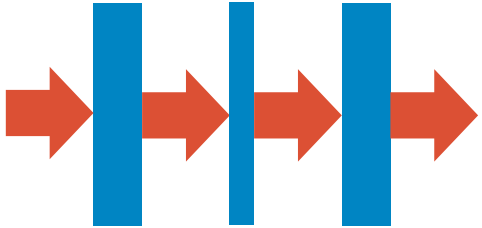
C, V, f



P

Pipeline

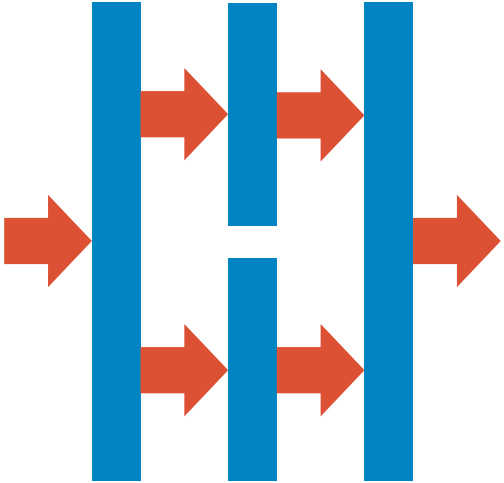
1.2 x C, 0.6 x V, f



0.432 x P

Parallèle

2.2 x C, 0.6 x V, 0.5 x f



0.396 x P

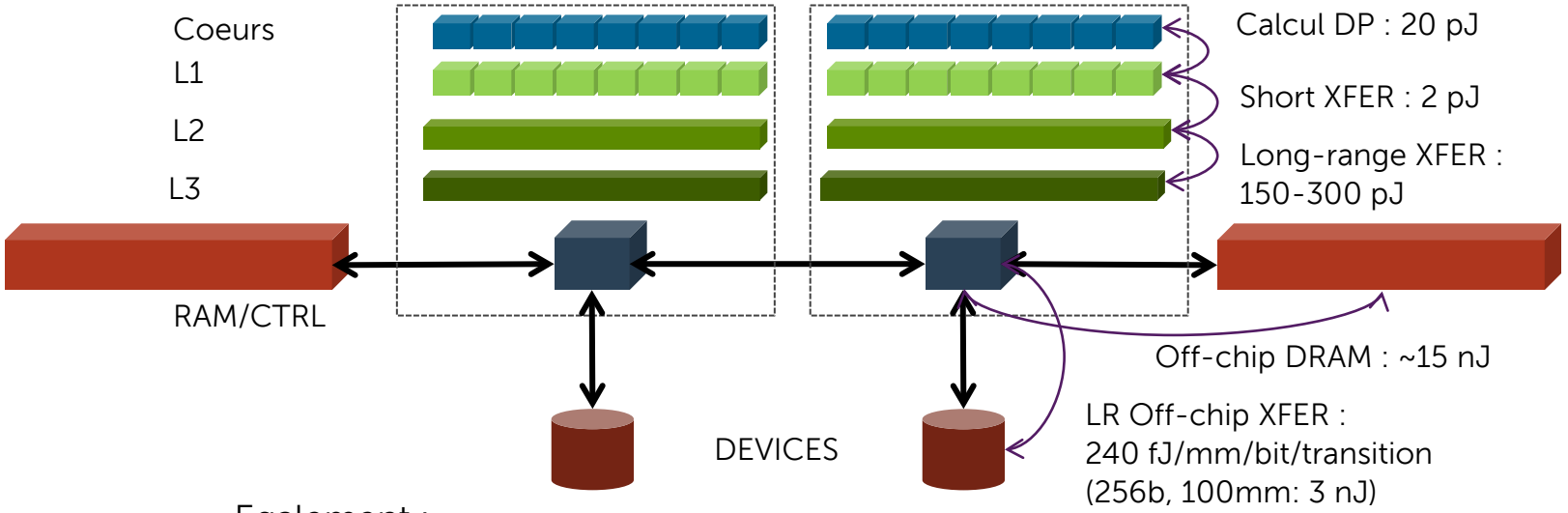
Les surfaces augmentent avec la complexité, ce qui induit des fuites supplémentaires (courant statique).

Bien entendu, les combinaisons sont valables.

Source : A. Tisserand, CNRS-IRISA-CAIRN. Introduction to Power Consumption in Digital Integrated Circuits – 2010.

Topologies des nœuds de calcul

Considérations énergétiques

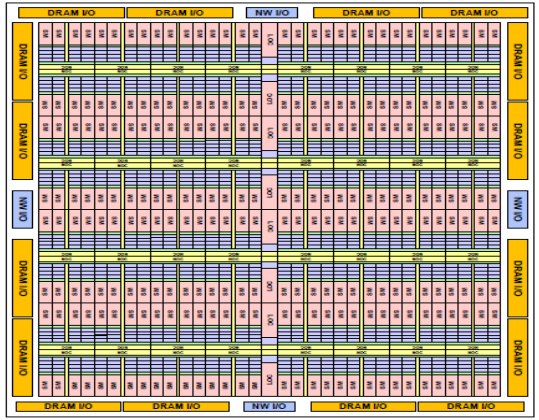


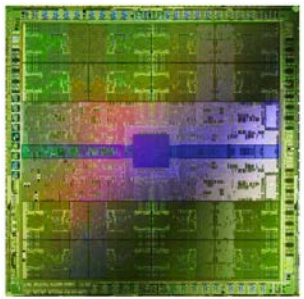
Egalement :

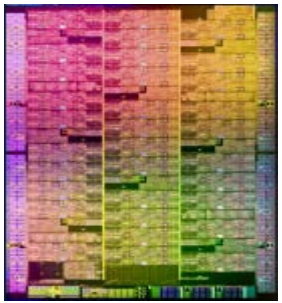
- L'ordonnanceur d'une unité de calcul peut aujourd'hui consommer 2nJ pour une opération de 25pJ
- L'équilibre entre les différents débits, latences et capacités est primordial pour maintenir l'efficacité de l'ensemble !

Topologies des CI

Localisation des échanges et transferts







Chip Echelon (17x17mm)

- Localité des caches
- Interfaces RAM/NW proches des cœurs
- Accès courtes distances
- Structure symétrique

- Concept minimaliste : maximum de puissance de calcul sur un minimum de surface pour MoBo totalement spécifique

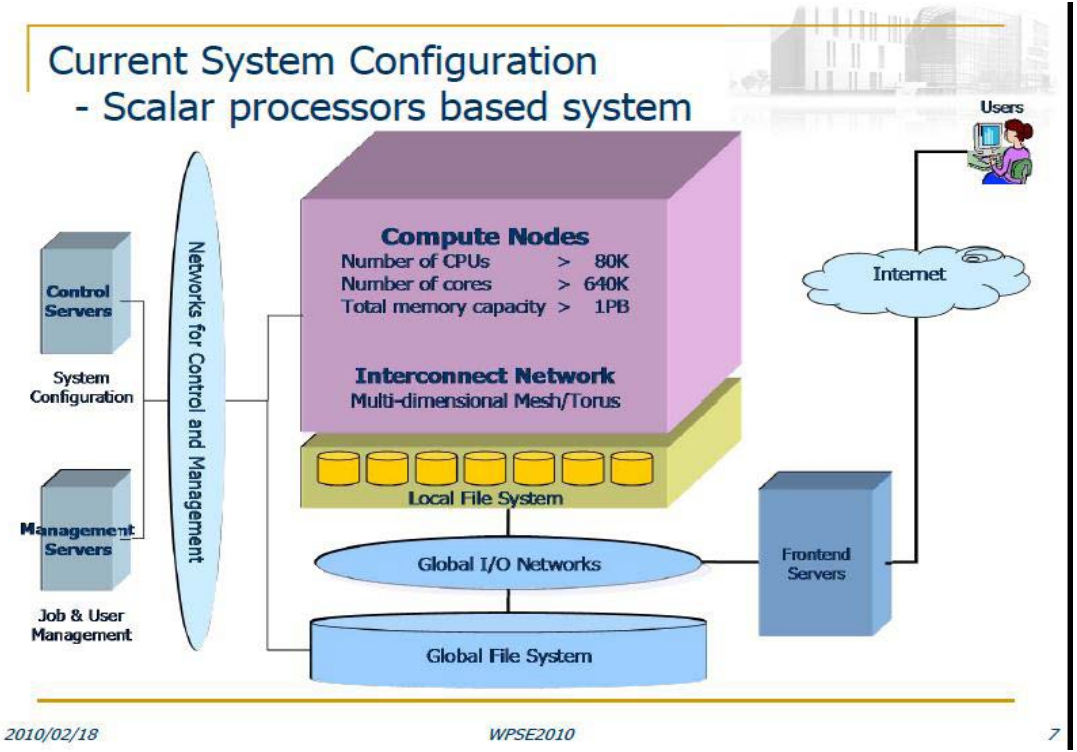
Chip MIC & Fermi

- Applications de la plupart des principes de localité
- Besoin d'exprimer/gérer la localité

Topologie globale (calcul + I/O)

Super Calculateur K, Kobe, Japon

- Les grands calculateurs sont hiérarchisés
- Localisation des données
- Optimisation des transferts de données
- Adaptation à la charge globale et locale



Résumé topologie/architecture

- **Système HPC actuel**
  - Hautement hiérarchique (caches, réseaux, architecture)
  - Non-transparent
  - Faiblement contrôlable
- **Trajectoire vers les systèmes exaflopiques**
  - Fortement contrainte par la consommation électrique et le refroidissement
  - Améliorer la localité des données : cache, mémoire, stockage …
  - Intégrer de manière optimale la pile d'exécution et l'architecture matérielle
  - Optimiser les technologies matérielles
  - Adopter les modèles de programmations adaptés
- **Position de Dell**
  - Intégrateur de technologies ouvertes et x86
  - Expertise calcul/calculateurs
  - Collaborations technologiques

Retour sur terre …

Processeurs Intel

4S

Intel Xeon E5-4600

- Plateforme 4 CPU 4-8 cœurs
- Jusqu'à 20 Mo cache L3 (2.5 Mo/cœur)
- 4 canaux mémoire 1600 MT/s max.
- 2x QPI 6.4-8.0 GT/s
- 40 liens PCIe3

2S

Intel Xeon E5-2600

- Plateforme 2 CPU 4-8 cœurs
- Caractéristiques identiques au E5-4600

2S

Intel Xeon E5-2400

- Plateforme 2 CPU 4-8 cœurs
- 3 canaux mémoire - 1x QPI – 24x PCIe3

Processeurs AMD

AMD Opteron 6300
- Plateforme 4 CPU 4-16 cœurs
- Jusqu'à 16 Mo cache L3 (1 Mo/cœur)
- 4 canaux mémoire 1600 MT/s max.
- 4x HT 6.4 GT/s

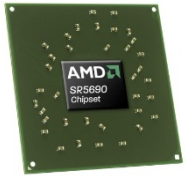
4S

AMD Opteron 4300
- Plateforme 4 CPU 4-8 cœurs
- Jusqu'à 8 Mo cache L3 (1 Mo/cœur)
- 2 canaux mémoire 1600 MT/s max.
- 2x HT 6.4 GT/s

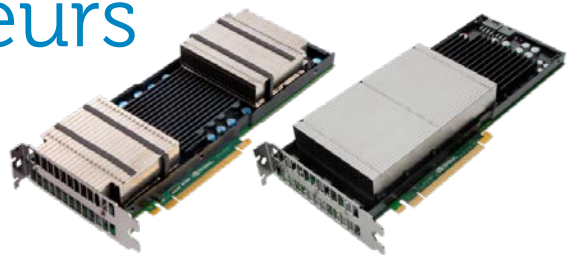
2S

AMD I/O
- SB : AMD SP5100 - USB/SATA
- IOH : SR5650, SR5670, SR5690 - 22-30-42 PCIe

I/O



Accélérateurs



NVIDIA « Kepler »

- **Caractéristiques générales**
  - 28nm – architecture Kepler
  - Mémoire GDDR5
  - PCIe G2 16x
- **NVIDIA Kepler K10 – simple précision**
  - 2x 1536 coeurs @ 0.745 GHz
  - 2x 8 canaux @ 5.0 GT/s
  - 2x 2.29 TFlop/s SP
- **NVIDIA Kepler K20 – double précision**
  - 2496 coeurs @ 0.705 GHz
  - 12 canaux @ 5.2 GT/s
  - 1.17 TFlop/s DP – 3.5 TFlop/s SP

Intel Xeon Phi

- **Caractéristiques générales**
  - 22nm – architecture MIC x86_64
  - PCIe G2 16x
  - 16 c. RAM GDDR5 - 320-352 Go/s
  - 1.02-1.22 TF/s DP, 2.02-2.44 TF/s SP
- **Intel Xeon Phi 5110P**
  - 60 coeurs @ 1.053 GHz - 225W
  - RAM 5.0 GT/s
- **Intel Xeon Phi 7120P**
  - 61 coeurs @ 1.1 GHz - 300 W
  - RAM 5.5 GT/s

TACC : 10 PFlop/s

- 6400 nœuds E5-2600 + 6400 Intel Xeon Phi ~7120P
- > 40 kW / rack, 5 MW
- > 1 M de threads !
- OpenMP/MPI sur Xeon **OU** Xeon Phi (2 et 7 PFlop/s)
- MPI sur Xeon **ET** OpenMP pour déporter sur Xeon Phi (10 PFlop/s)
- Retour d'expérience :
  - Portage sur Intel Xeon Phi : presque trop facile !
  - Optimisation sur Xeon Phi => meilleures performances sur Xeon !
- Avant tout :
  - **Paralléliser** - plusieurs threads/cœur Xeon Phi (mini 2, 4HT)
  - **Vectoriser** - pour bénéficier des performances de l'unité AVX (256b)

Conclusion

Conclusion

Limites technologiques des processeurs et architectures

- Contraintes thermiques
- Organisation des systèmes

Evolutions

- Plateformes, réseaux
- Accélérateurs
- Logiciels

Préparez le terrain !

- Vectorisez !
- Parallélisez !

Merci !

marc_mendez_bermond@dell.com