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

Marc Mendez-Bermond

Expert solutions HPC

Programme

- Contexte
- Technologies
- Evolutions

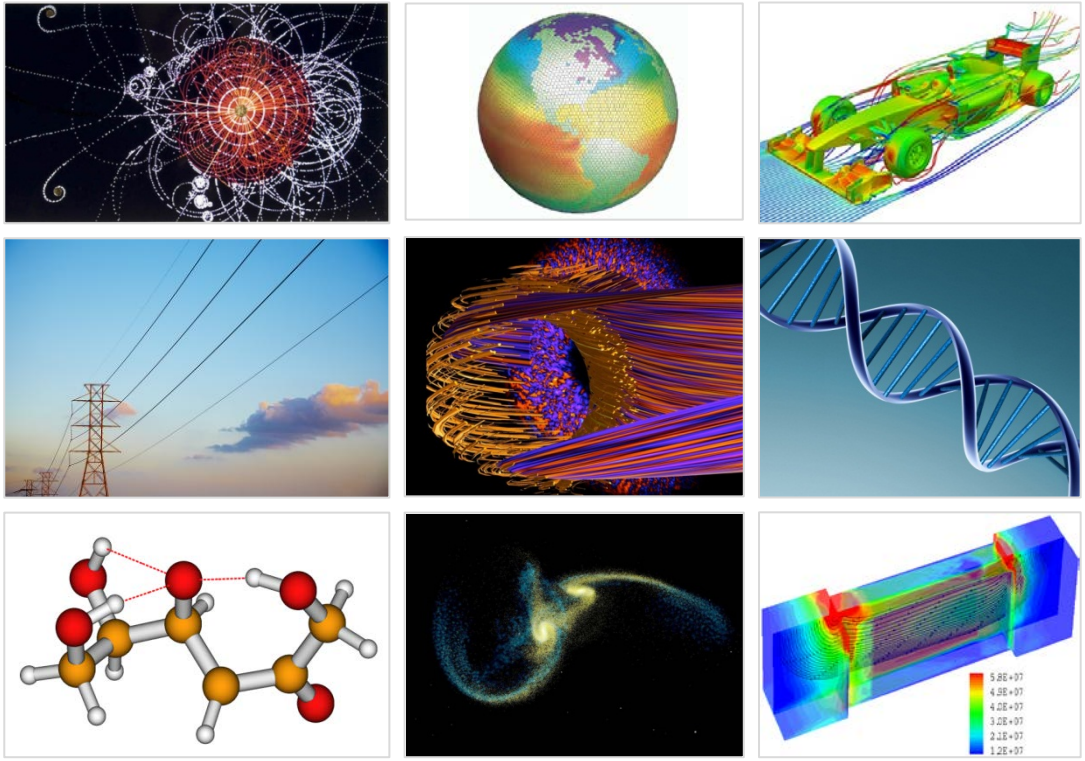
Contexte

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 d'applications
- Besoins croissants de puissance:
  - Augmentation des tailles des problèmes
  - Augmentation de la précision
  - Ajout de physique additionnelle
- Adoption soutenue 8% / an
- Rapport GES économisés / consommés > 10

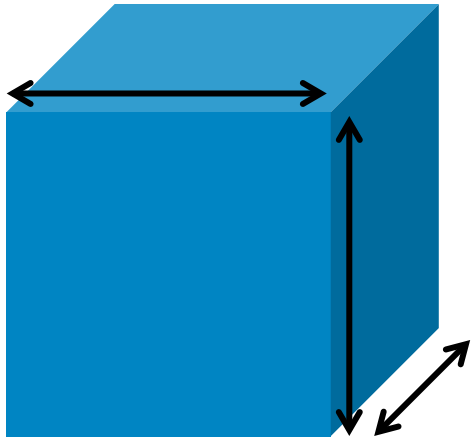


Objectif Exaflop = 10^18 flop/s !

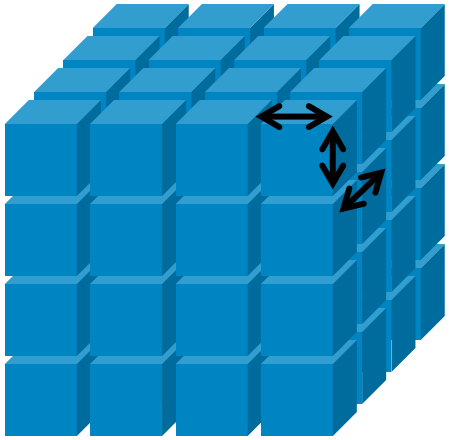
Accélération par parallélisme

- 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

- **SMP – système NUMA, vectoriels, mainframes (1960-2000)**
  - Nombre de processeurs limité (max. 2048)
  - Complexes et propriétaires

- **Beowolf – grappes de calcul x86 (2000-...)**
  - Limites moins franches et très éloignées
  - Technologies ouvertes
  - Meilleurs rapports performance / couts

- **Hybride – grappes x86 + accélérateur (2006-...)**
  - Idem Beowolf
  - Difficultés de programmation

Evolutions des systèmes

Evolution naturelle

Les composants progressent de par les **évolutions des procédés de fabrication**.

Aucun effort sur l'application.

Evolution contrainte par le marché/besoin

Pousser les solutions aux limites permises par la technologie.

Pousser la scalabilité de l'application existante.

Evolution par saut technologique

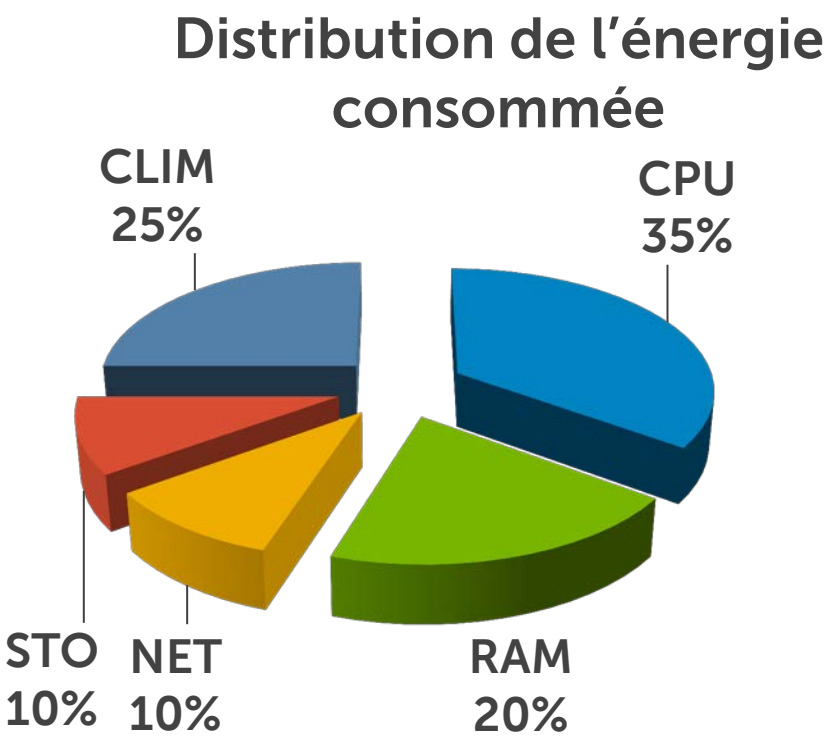
Revoir les concepts, technologies et solutions pour progresser d'un **ordre de magnitude**.

Changer de paradigme de programmation ou **spécialiser** la solution.

Limites et ordres de grandeur

	Composants	Système 1PF/s - 2013	Système 1EF/s - 2013	Système 1EF/s - 2018
Calcul	CPU : 100 GF/s ACC : 1 TF/s	1 PF/s Cœurs > 10^6	1 EF/s Cœurs > 10^9	1 EF/s Cœurs > 10^8
Energie	CPU : 100 W ACC : 200 W	1 MW	500 MW	50-80 MW
Réseau	HCA : ~50 Gb/s	100k ports	10-100M ports	1-10M ports
Efficacité		1 GFlop/s/W	2 GFlop/s/W	> 10 GFlop/s/W

Cibles d'évolution



- CPU
- RAM
- NET
- STO
- CLIM

Accélérateurs, vectorisation, threading, gravure

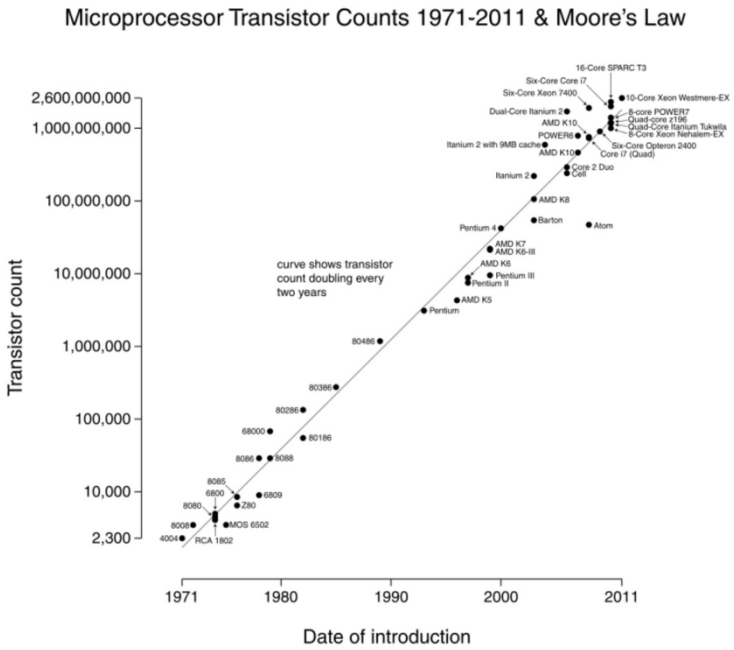
DDR4, RAM 3D, connexions optiques

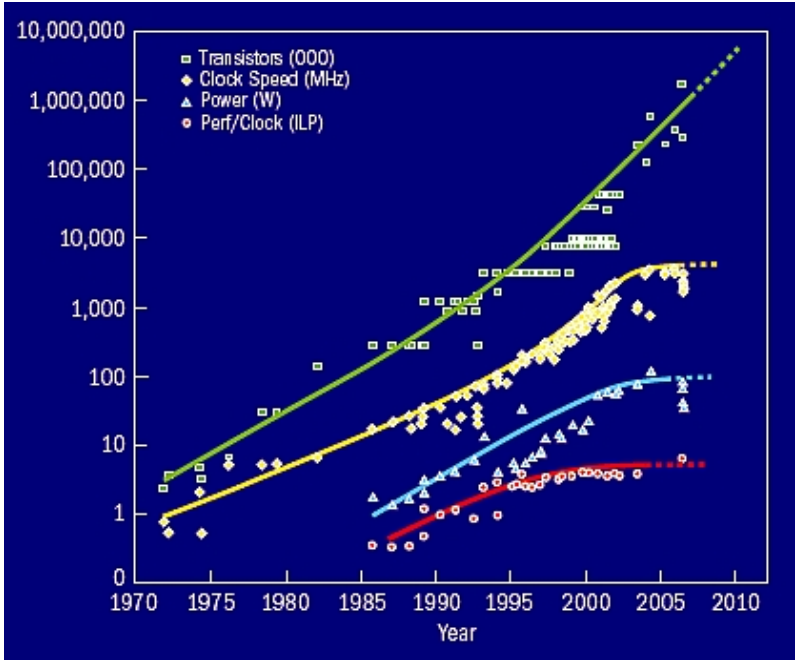
Topologies, connexions optiques, optimisation commutation

Technologie HDD/Si, contrôleurs, hiérarchisation

Climatisation ou ventilation, air ou eau, allées, confinements, conteneurs ...

Progrès des processeurs





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

Consommation des circuits intégrés

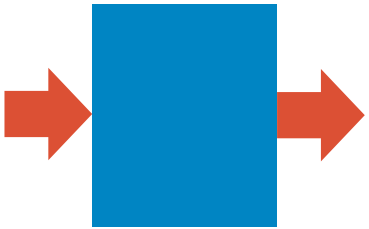
- **$P = a \cdot C \cdot V^2 \cdot f$**
  - a : facteur d'activité, C : capacitance, V : tension de travail, f : fréquence d'opération
- **$P = k \cdot V^3$**
- **$P = x \cdot f^3$**

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

Structures des unités de calcul

Base

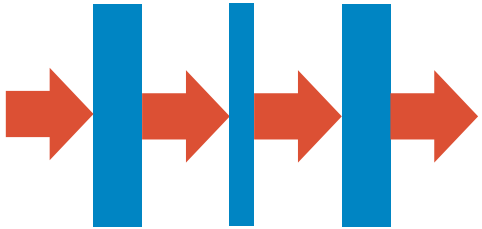
C, V, f



P

Pipeline

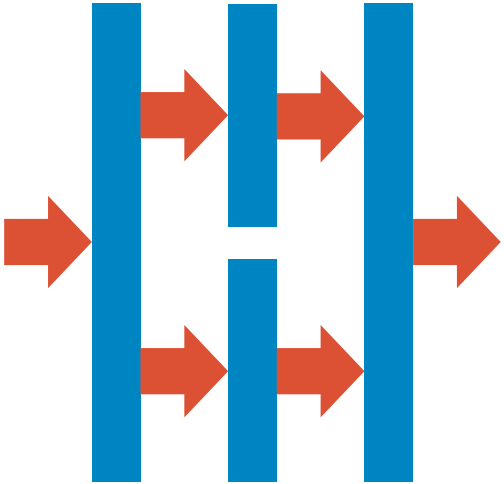
1.2 x C, 0.6 x V, f



0.432 x P

Parallèle

2.2 x C, 0.6 x V, 0.5 x f



0.396 x P

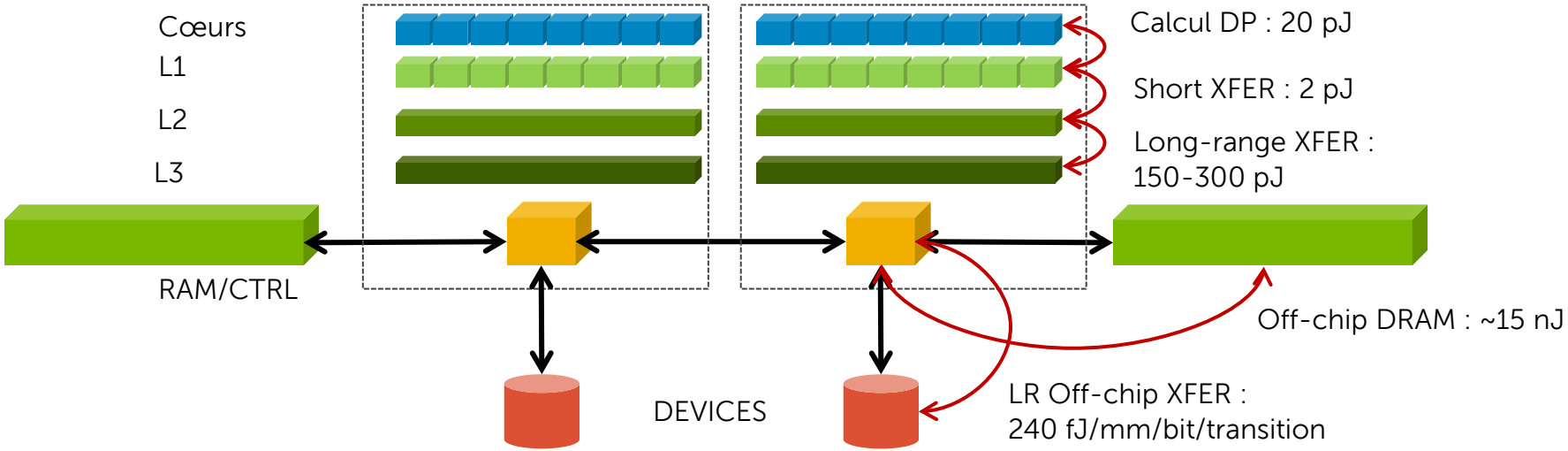
Les surfaces augmentent avec la complexité, ce qui induit des fuites supplémentaires (courant statique).

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

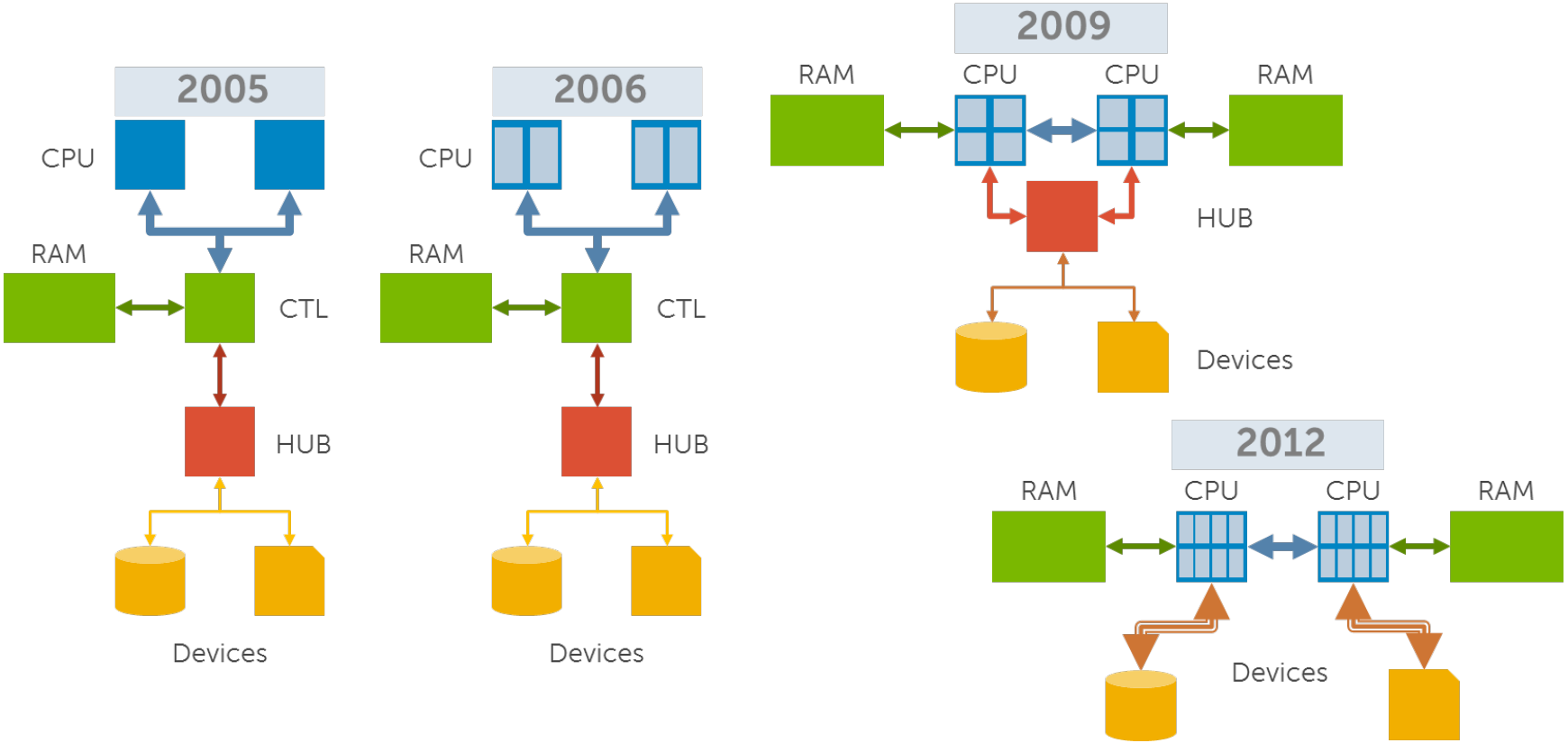
Considérations énergétiques



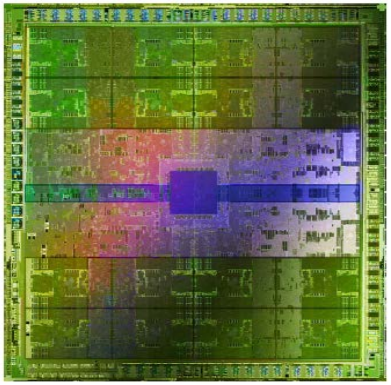
Egalement :

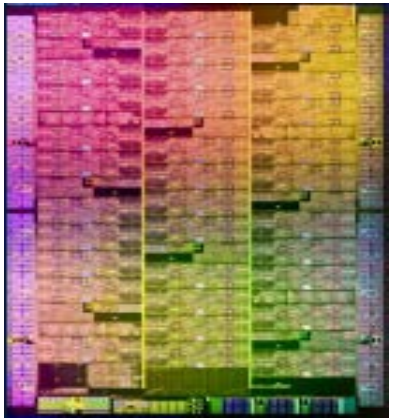
- L'ordonnanceur d'une unité de calcul peut aujourd'hui consommer 2nJ pour une opération de 25pJ
- L'équilibre entre les différents débits, latences et capacités est primordial pour maintenir l'efficacité de l'ensemble !

Evolution des plateformes



Organisation des CI





Principes généraux

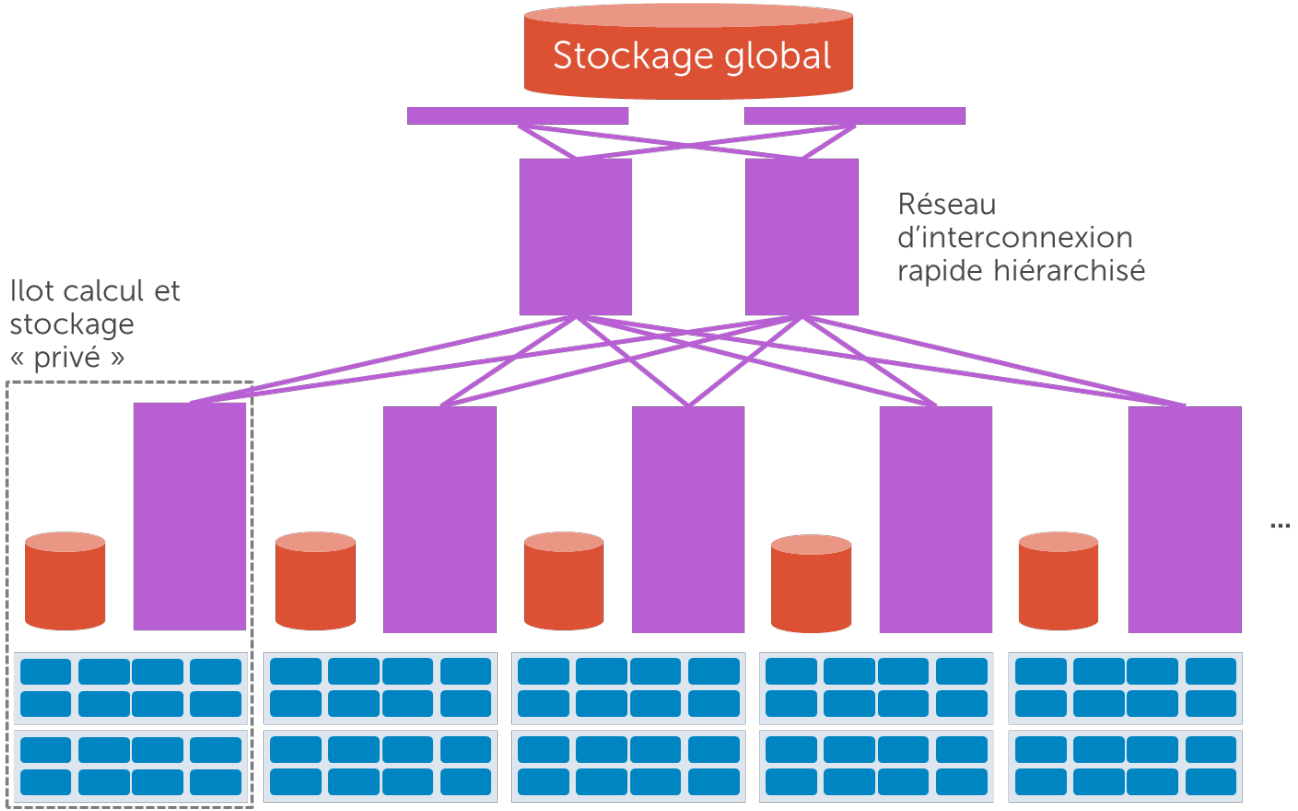
- Localité des caches
- Interfaces RAM/IO/NET proches des cœurs
- Optimisation des distances
- Structure symétrique

Chip Phi & Kepler

- Applications des principes de localité avec
  - 7B transistors (Kepler)
  - 5B transistors (Phi)

Topologie globale (calcul + I/O)

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

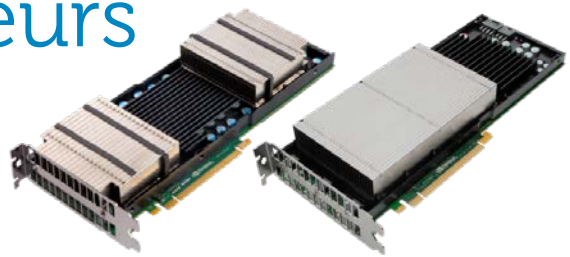


Résumé topologie/architecture

- **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

Retour sur terre …

Accélérateurs



NVIDIA « Kepler »

- **Caractéristiques générales**
  - 28nm – architecture Kepler
  - Mémoire GDDR5
  - PCIe G2 16x
- **NVIDIA Kepler K10 – simple précision**
  - 2x 1536 cœurs @ 0.745 GHz
  - 2x 8 canaux @ 5.0 GT/s
  - 2x 2.29 TFlop/s SP – 250W
- **NVIDIA Kepler K20 – double précision**
  - 2496/832 cœurs @ 0.705 GHz
  - 12 canaux @ 5.2 GT/s
  - 1.17 TFlop/s DP - 300W

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 cœurs @ 1.053 GHz - 225W
  - RAM 5.0 GT/s
- **Intel Xeon Phi 7120P**
  - 61 cœurs @ 1.1 GHz - 300 W
  - RAM 5.5 GT/s

Dell PowerEdge et accélérateurs

Dell PowerEdge C8220x
- 4-5 nœuds 2S E5-2600
- 2 accélérateurs par nœud
- Liens PCIe 16x III dédiés

Phi
Kepler

Dell PowerEdge C6220 & C410x
- 4 nœuds 2S E5-2600
- Jusqu'à 8 accélérateurs par nœud
- Châssis externe PE-C410x

Phi
Kepler

Dell PowerEdge R720
- Serveur 2S E5-2600
- 2 accélérateurs par nœud
- Liens PCIe 16x III dédiés

Phi
Kepler

Systèmes de calcul accéléré et de visualisation

Dell PowerEdge T620

- Station de travail 2S E5-2600
- Jusqu'à 4 accélérateurs
- Kit de mise en armoire

Dell Precision T7600/5600/3600

- Stations de travail 1-2S E5-2600
- 1-2 accélérateurs 300W

Matrice serveurs de calcul accélérés

Modèle	GPU	Processeurs	RAM	Commentaires
C8220x	2x NVIDIA Tesla K10/K20 2x Intel Xeon Phi 5110/7120P	2S E5-2600	16 DIMMs/ 256Go	2 emplacements internes 300W
C6220	8x NVIDIA Tesla K10/K20 8x Intel Xeon Phi 5110P	2S E5-2600	16 DIMMs / 256Go	Châssis externe PE-C410x (16 emplacements 300 W / 8 ports PCIe HIC)
C6145	8x NVIDIA Tesla K10/K20 8x Intel Xeon Phi 5110P	4S Opteron 6300	32 DIMMs / 512Go	Châssis externe PE-C410x (16 emplacements 300 W / 8 ports PCIe HIC)
R720	2x NVIDIA Tesla K10/K20/K20x 2x NVIDIA Tesla Grid K1/K2 2x NVIDIA Quadro K2000/K4000 2x Intel Xeon Phi 5110/7120P 2x AMD FirePro S7000/S9000	2S E5-2600	24 DIMMs / 768Go	2 emplacements internes 300W
T620	4x NVIDIA Tesla K20 4x NVIDIA Tesla K4000 4x Intel Xeon Phi 3120A 4x AMD FirePro W7000	2S E5-2600	24 DIMMs / 768Go	4 emplacements internes Refroidissement actif

TEXAS ADVANCED COMPUTING CENTER

Powering Discoveries That Change The World

TOP500 NOVEMBER 2012

- **$27.5M funded by US National Science Foundation**
- **Built in partnership with Dell and Intel**
- **In production January 2013**
- ~9-10 PetaFLOPS peak (20% CPU, 80% co-processor)
- > 6 PetaFLOPS rMax (Linpack) when in full production.
- 272 TB RAM
- Intel Xeon Phil (MIC) co-processors
- 56 Gigabit FDR Infiniband
- 14 petabytes of storage with 150 GBps Lustre file system
- 5 Mwatts Power

5	Forschungszentrum Juelich (FZJ)	Germany	**JUQUEEN** - BlueGene/Q, Power 16C 1.600GHz, Custom Interconn IBM
6	Leibniz Rechenzentrum	Germany	**SuperMUC** - iDataPlex DX360M4 E5-2680 8C 2.70GHz, Infiniband IBM
7	Texas Advanced Computing Center/Univ. of Texas	United States	**Stampede** - PowerEdge C8220, 2680 8C 2.700GHz, Infiniband FD Xeon Phi Dell
8	National Supercomputing		**Tianhe-1A** - NUDT YH MPP, Xe

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 !
  - Optimiser 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 AVX2

Conclusion

Conclusion

Limites technologiques des processeurs et architectures pour les très grands systèmes

- Contraintes thermiques
- Organisation des systèmes

Evolutions se diluant vers les systèmes plus modestes

- Plateformes, réseaux
- Accélérateurs
- Logiciels

Préparez le terrain !

- Vectorisez !
- Parallélisez !

Merci !

marc_mendez_bermond@dell.com