Le calcul intensif face aux défis de l'Exascale et du Big Data

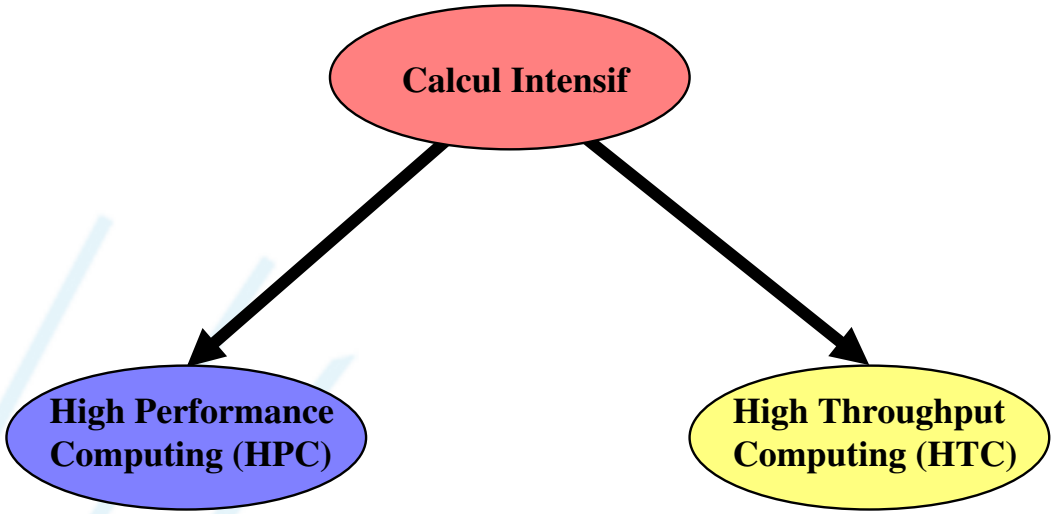
Journées Prospectives 2013

F. Suter

23 mai 2013

Calcul intensif : HPC vs. HTC

Au cœur des grandes avancées de la recherche scientifique



- 1 job sur N CPUs
- Optimiser les I/O
- Supercalculateur
- Ex : Modèle climatique

- N jobs sur 1 CPU
- Maximiser le débit
- Ferme de calcul
- Ex : Recherche du Higgs

Des organisations similaires

HPC

- Mésocentres
- Centres nationaux
  - CINES, IDRIS
- Structuration Nationale
  - GENCI
- Structuration Européenne
  - PRACE

HTC

- Sites Grilles
- Centre national
  - CC-IN2P3
- Structuration Nationale
  - France Grilles
- Structuration Européenne
  - EGI

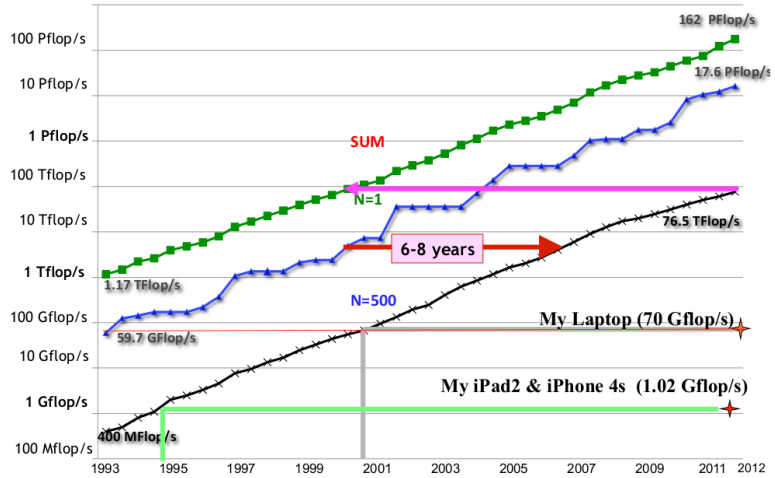
Mais des différences fondamentales

- Applications cibles
- Communautés d'utilisateurs
- Technologies
- Domaines d'expertise

HPC : la course à la performance

- Top 500
  - Classement mondial
  - Tous les 6 mois
  - x1000 tous les 11 ans
- Top 3 actuel
  - Titan, Cray XK7, DOE Oak Ridge
    - 17.59 Pflops pour 8,21 MW
    - 560 640 cœurs (AMD + NVIDIA)
  - Sequoia, BlueGene/Q, DOE, LLNL
    - 16,324 Pflops pour 7,89 MW
    - 1 572 864 cœurs
  - K Computer, Riken, Kobe
    - 10.51 Pflops pour 12.7 MW
    - 705 024 cœurs (SPARC64)

Courtesy : J. Dongarra

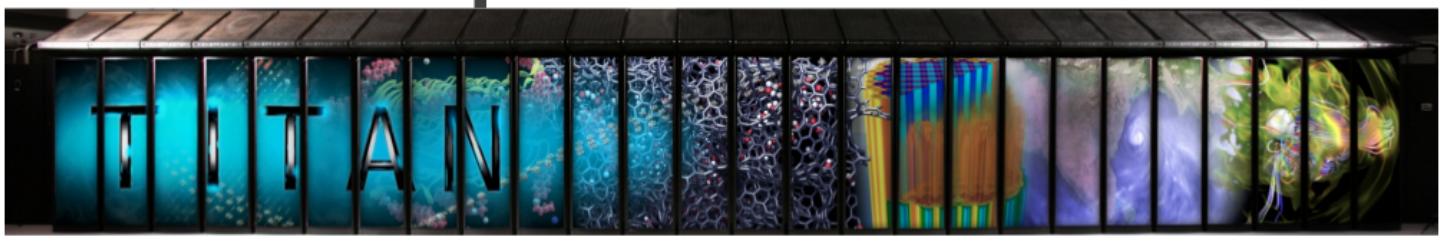


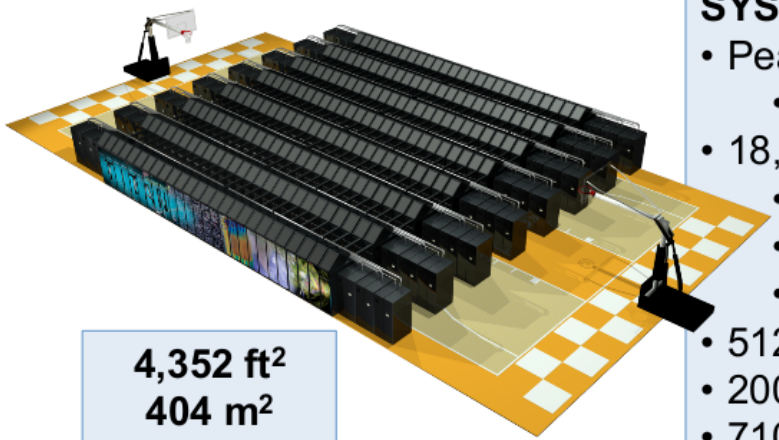
http://www.top5000.org

Megaflops = 10^6 opérations flottantes / seconde, Gigaflops = 10^9, Teraflops = 10^{12}, Petaflops = 10^{15}, Exaflops = 10^{18}.

And the winner is . . .

ORNL's "Titan" Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla processors





4,352 ft²
404 m²

SYSTEM SPECIFICATIONS:

- Peak performance of 27 PF
  - 24.5 Pflop/s GPU + 2.6 Pflop/s AMD
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - 14-Core NVIDIA Tesla "K20x" GPU
  - 32 GB + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 9 MW peak power

11 OLCF | 20

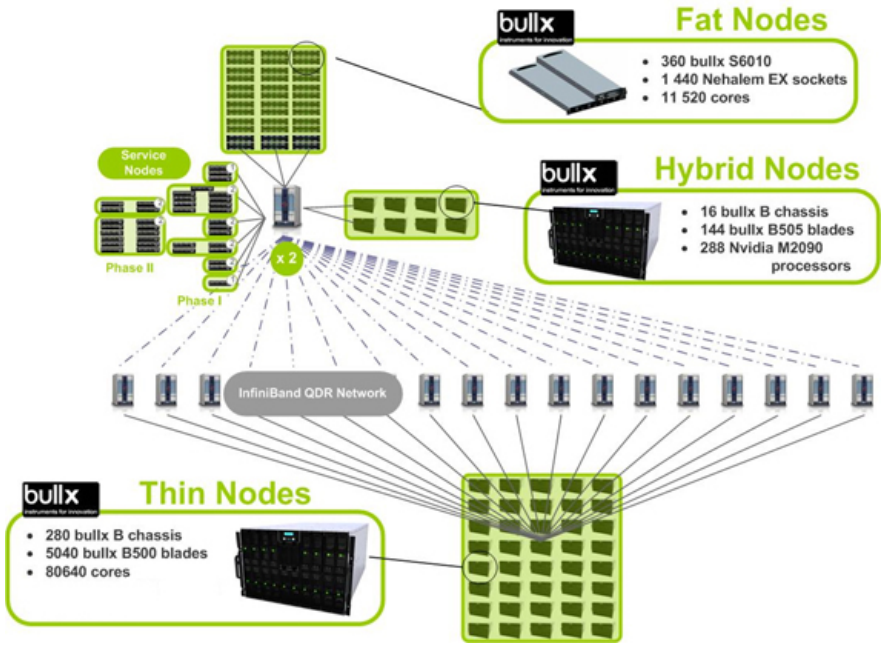
OAK RIDGE National Laboratory

Et en France ?

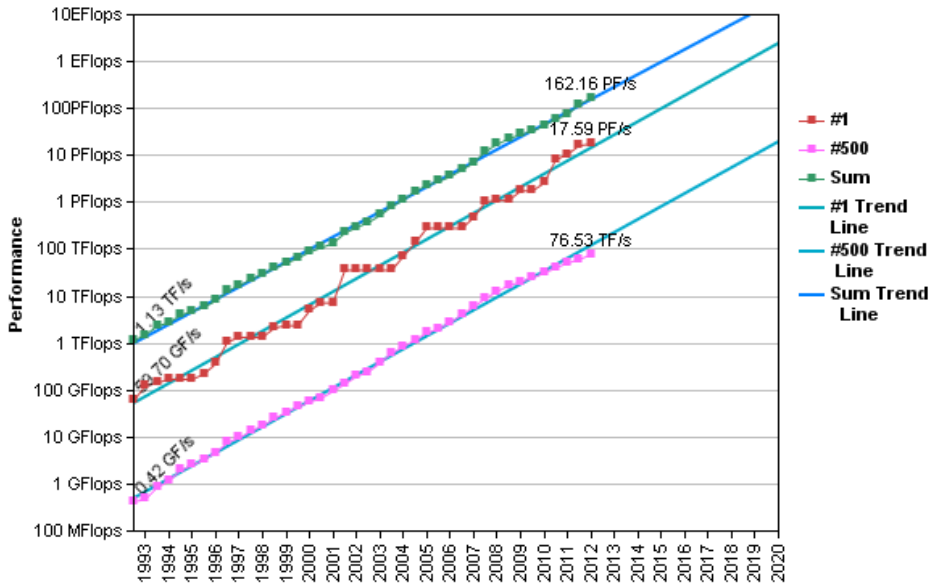
Machine Curie du TGCC

- 11ème position
- 1.36 Pflops pour 2.25 MW
- 77 184 cœurs

© CEA/Cadam



Vers l'Exascale (et au delà ?)



- Systèmes exascales en US, EU, Chine, Japon et Russie vers 2020
- Problème
  - Technologies actuelles $\Rightarrow$ 1 Exaflops = 475MW !
    - Et seulement 2% pour le calcul
    - $\sim$ 50% :accès stockage données et déplacement des données
    - $\sim$ 50% : cache, mémoire virtuelle, ...

Objectifs et moyens

- Aux USA (via le DOE)
  - Mais partagés par tous
- Performances
  - 2016-2017 $\Rightarrow$ 100 PFlops
  - 2018-2020$\Rightarrow$ 1 Exaflops
- Contraintes
  - Utiliser des technologies sur étagère ou au moins viables commercialement
  - Comsommation $<$ 20 MW
  - Calculateurs suffisamment généralistes
- Moyens
  - Projet mondial sur l'Exascale financé par le G8
  - Le premier du genre
  - Regroupe tous les grands acteurs du HPC

Une machine exascale ?

- A moins de 200 millions de dollars
- Et consommant moins de 20MW

Systems	2012 Titan Computer	2022	Difference Today & 2022
System peak	27 Pflop/s	1 Eflop/s	O(100)
Power	8.3 MW (2 Gflops/W)	~20 MW (50 Gflops/W)	O(10)
System memory	710 TB (38*18688)	32 - 64 PB	O(100)
Node performance	1,452 GF/s (1311+141)	1.2 or 15TF/s	O(10)
Node memory BW	232 GB/s (52+180)	2 - 4TB/s	O(10)
Node concurrency	16 cores CPU 2688 CUDA cores	O(1k) or 10k	O(100) - O(10)
Total Node Interconnect BW	8 GB/s	200-400GB/s	O(100)
System size (nodes)	18,688	O(100,000) or O(1M)	O(10) - O(100)
Total concurrency	50 M	O(billion)	O(100)
MTTF	?? unknown	O(<1 day)	O(?)

Courtesy : J. Dongarra

En quoi c'est dur ?

- CPU : finie la loi de moore
  - Place aux GPU and ManyCores ⇒ Hétérogénéité
- Réseau : machines de plus en plus éloignées
  - Minimiser la bissection ⇒ réseaux complexes (Tores 6D, . . .)
- Mémoire et Stockage
  - Facteur d'échelle ⇒ Assurer la cohérence
- Énergie
  - Fortement contrainte ⇒ Optimisation de partout
- Parallélisme massif
  - des millions de CPU ⇒ Algorithmes
- Pannes
  - De plus en plus fréquentes ⇒ Robustesse
- Programmation
  - Tout change ⇒ Comment former la relève ?

Quel rapport avec le CC-IN2P3 ?

- Hardware : HPC pousse à l'amélioration
  - Multi-cœurs et GPU
  - Réseau Gigabit
  - Grandes capacités mémoires (partagées ou non)
  - $\Rightarrow$ Fini toujours par devenir standard
- Software : s'adapter aux changements
  - Conserver la même façon de programmer $\Rightarrow$ sous-exploitation
  - Nouvelles applications $\Rightarrow$ plus de parallélisme
  - Scale down toujours possible
- Humanware : voir loin
  - Poser des contraintes
  - Anticiper les besoins
  - Chercher la bonne solution ($\neq$ la plus facile)

Mais . . . et le Big Data ?

- Le Big Data : pour quoi ? pour qui ?
  - Autant de définitions/solutions que de domaines

Mais . . . et le Big Data ?

- Le Big Data : pour quoi ? pour qui ?
  - Autant de définitions/solutions que de domaines
- Big Data en HEP
  - Beaucoup de fichiers indépendants (ou presque)
  - Solution : Grille de calcul/données

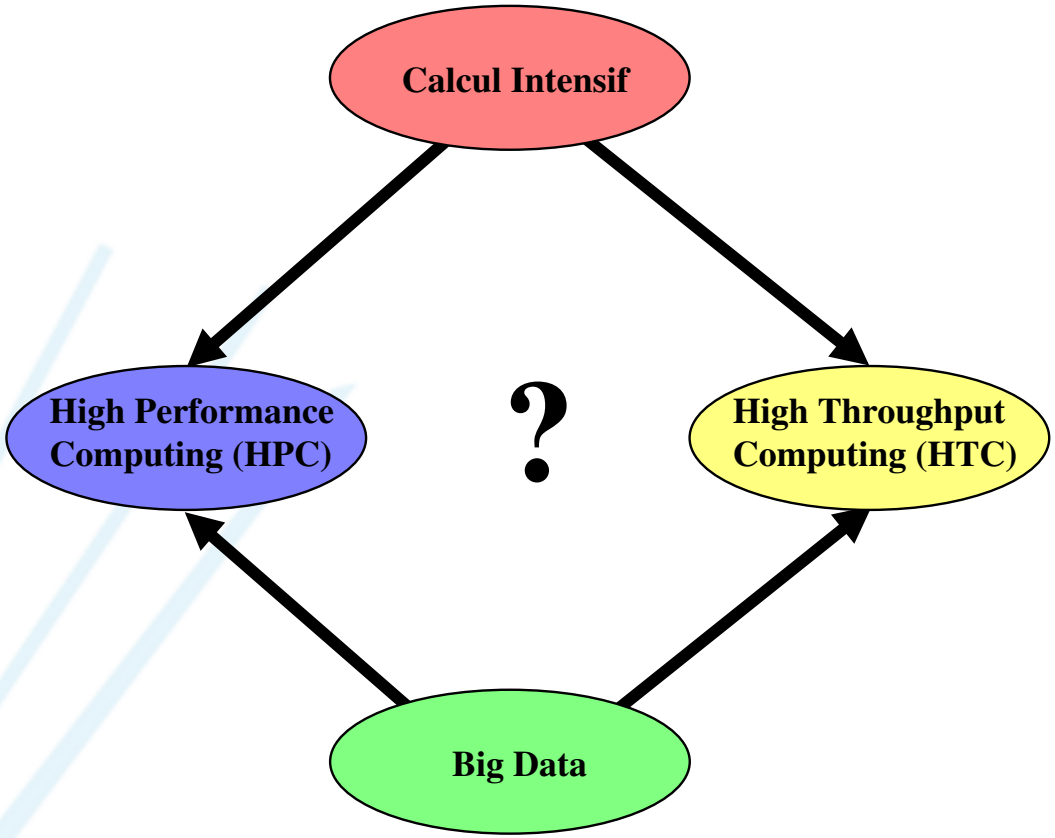
Mais . . . et le Big Data ?

- Le Big Data : pour quoi ? pour qui ?
  - Autant de définitions/solutions que de domaines
- Big Data en HEP
  - Beaucoup de fichiers indépendants (ou presque)
  - Solution : Grille de calcul/données
- Big Data en HPC
  - Gros jobs parallèles : milliers de cœurs $\Rightarrow$ grosse mémoire distribuée
  - Big Data $\Rightarrow$ gros traitements en *in-core*
  - Solution : I/O haute performance (disque et réseau)

Mais . . . et le Big Data ?

- Le Big Data : pour quoi ? pour qui ?
  - Autant de définitions/solutions que de domaines
- Big Data en HEP
  - Beaucoup de fichiers indépendants (ou presque)
  - Solution : Grille de calcul/données
- Big Data en HPC
  - Gros jobs parallèles : milliers de cœurs ⇒ grosse mémoire distribuée
  - Big Data ⇒ gros traitements en *in-core*
  - Solution : I/O haute performance (disque et réseau)
- Big Data Analystics (implicite pour beaucoup de monde)
  - indexation du web, fouille de données, analyse de logs, machine learning, analyse financière, . . .
  - Solution : Map-Reduce

Calcul intensif et Big Data



Calcul intensif et Big Data