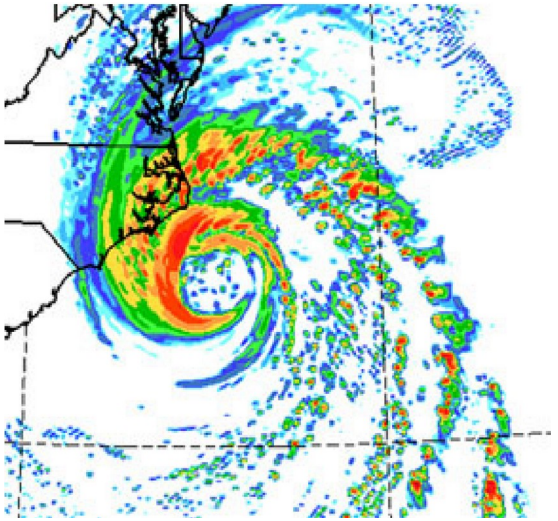
Systèmes Hautes Performances
Introduction

Elisabeth Brunet

CSC5001 - Septembre 2020

Calcul et simulation scientifiques

- **Essentiels pour l'innovation scientifique et industrielle**
- **Nombreux domaines d'application**
  - Météorologie, astrophysique, nanosciences, etc.
  - Automobile, aéronautique, 7ème art, défense, etc.
- **Simulation nécessaire lorsque les problèmes sont …**
  - ...trop complexes
  - ...trop massifs
  - …trop chers
  - …trop dangereux
  - ...prédictifs



Calcul et simulation scientifiques

- **Beaucoup de calcul**
- **Manipulation de masses de données**
- **Contraintes de temps**

- **Conséquences**
  - Accroissement des ressources de calcul
  - Parallélisation des problèmes

→ Pour aller toujours plus vite

→ mais surtout traiter des problèmes toujours plus gros

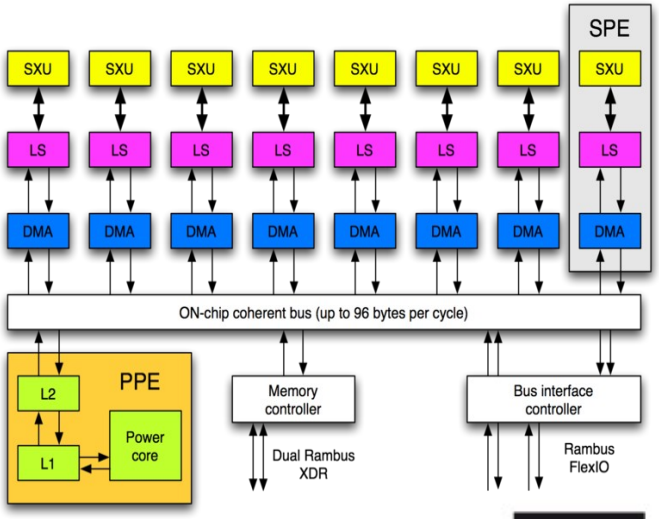
Optimisation des plates-formes

- **Capacités des composants tirées au maximum**

- **Diversité des processeurs**
  - Architectures grand public boostées
  - Architectures spécialisées
    - Processeurs Alpha, MIPS, Cray, Power, Sparc, NEC, etc.
      - Jeux d'instructions
      - Mode de fonctionnement :
        - Processeur vectoriel
          - Exécution de la même instruction sur chaque entrée d'un tableau
          - Processeurs Cray, unité de calcul AltiVec, instructions SSE
        - Processeur Cell

Processeur Cell

- **Processeur conçu par IBM, Sony et Toshiba**
- **Février 2005**
- **Modèle d'architecture totalement différent**
  - 1 processeur *maître* PPE de type PowerPC
  - 8 co-processeurs SPE vectoriels
  - Bus interne d'interconnexion EIB
- **Performances maximales théoriques**
  - 230,4 GFLOPS en simple précision
  - Initialement pour le multimédia (PS3, etc.) mais détourné par le HPC
- **Extrêmement difficile à programmer**
  - Abandonné en 2009



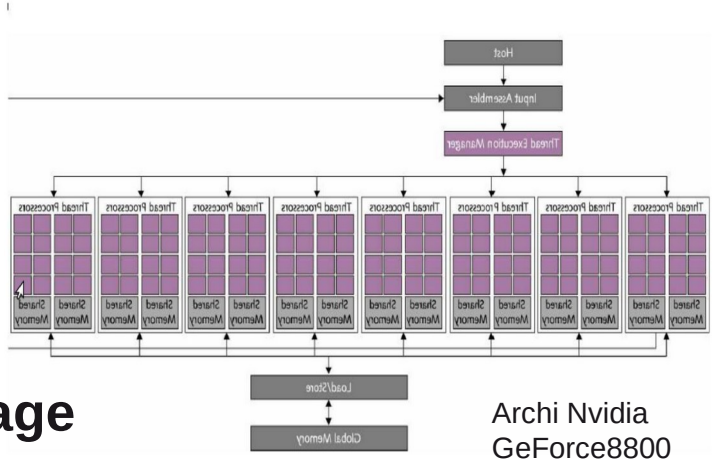
Optimisation des plates-formes

- **Capacités des composants tirées au maximum**

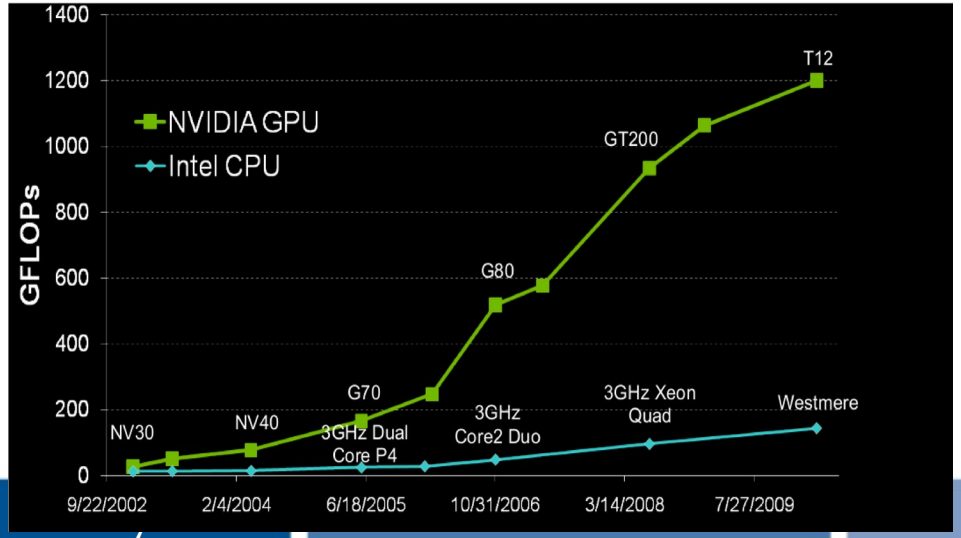
- **Diversité des processeurs**
  - Architectures grand public boostées
  - Architectures spécialisées
  - Adjonction de co-processeurs
    - FPGA : circuit logique programmable
    - GPU : Graphics Processing Unit

GPU Computing

- **1 GPU = centaines de cœurs limités**
  - pas d'allocation dynamique de mémoire
  - pas de pile, donc pas de récursion
- **Conçu pour le calcul 3D en synthèse d'image**
  - API 3D : OpenGL, DirectX
- **Orientation calcul parallèle**
  - Nvidia : Architecture Tesla(2006), Fermi(2009) / API CUDA
  - AMD : Archi RadeonHD / API ATI Stream SDK
  - API unifiée GPU+CPU : OpenCL (2008)



Archi Nvidia GeForce8800



- **Résurrection de la loi de Moore**
  - Projet Larabee (non abouti)
- **Traitement vectoriel distant**
  - Adapté au calcul «naturellement parallèle»

  Transfert de données via PICe
  1 *noyau* exécuté par tous les threads
  Rapatriement des résultats via PCIe
  - ➔ Interactions CPU/GPU lentes

TELECOM SudParis

Optimisation des plates-formes

- **Capacités des composants tirées au maximum**

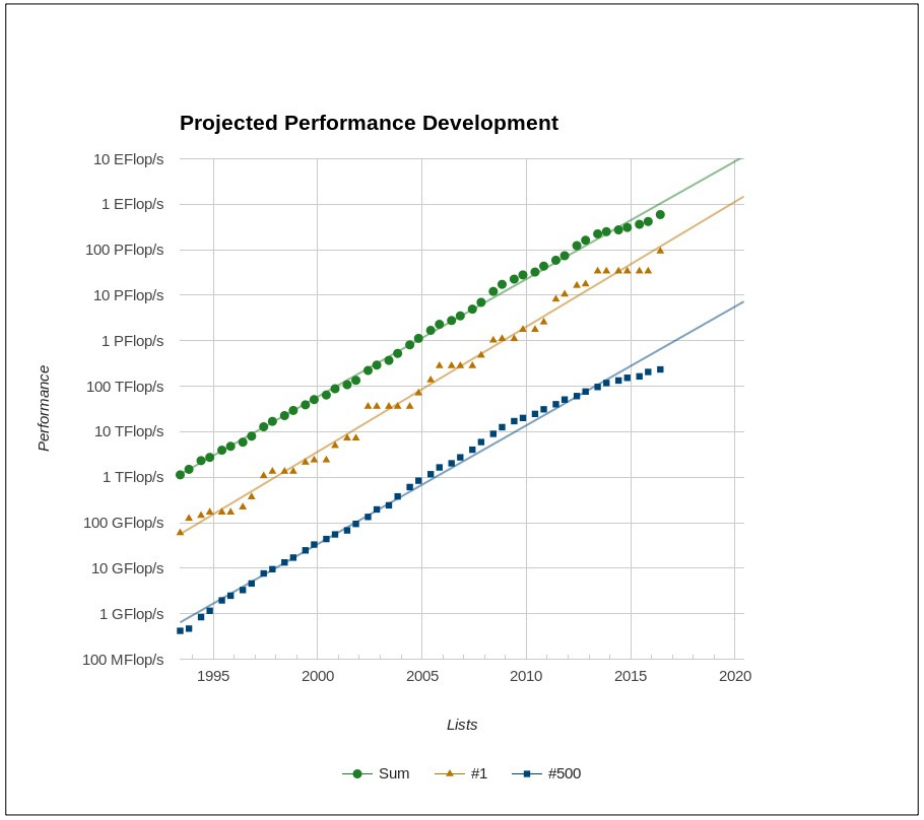
- **Diversité des processeurs**
  - Architectures grand public boostées
  - Architectures spécialisées
  - Adjonction de co-processeurs
  - Agrégation massive de ressources
    - Réseaux haute performance
      - InfiniBand, Myrinet, 10G-Ethernet, etc.
    - Topologies
      - Supercalculateurs : plate-formes propriétaires
      - Grappes : machines dites *grand public* interconnectées par un réseau point-à-point
      - Grilles : supercalculateurs/grappes interconnectés par un réseau longue distance

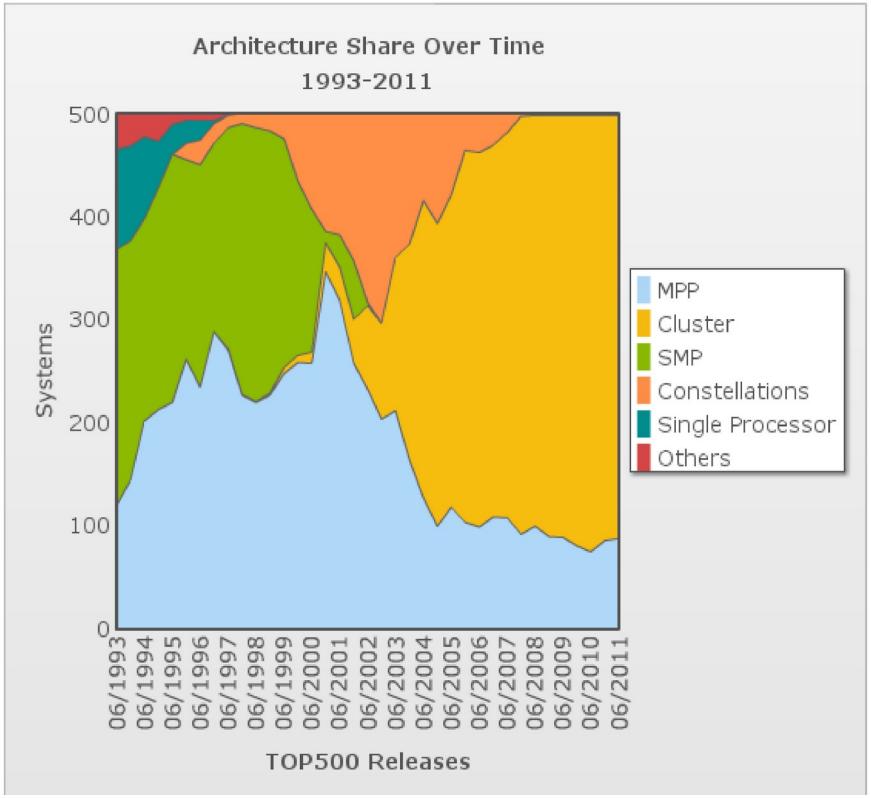
Bilan des ressources de calcul dans le HPC

- **Architectures hybrides**
- **Composants tous à la pointe de la technologie**
- **Explosion de la puissance de calcul**
  - 1 nœud surpasse une grappe de mono-processeurs
  - Échelle actuelle : le petascale, voire l'exascale
  - Recensement dans le Top500

Top500 - www.top500.org

- **Classement bi-annuel des 500 machines les plus puissantes du monde**





Top500 de juin 2020

Rmax et Rpeak en Tflop/s
Power en kW

1 **Fugaku –** A64FX48C 2.2GHz, Tofu Interconnect – Fujitsu RIKEN Center of Computational Science (Japan)
#coeurs = 7,299,072 Rmax = 415,530 Rpeak=513,854 Power=28,335

2 **Summit** – IBM POWER9, NVIDIA Volta GV100 – Oak Ridge National Lab (USA) → **#1 en 2018**
coeurs=2,414,592 Rmax=148,600 RPeak=200,794 Power=10,096

3 **Sierra** – IBM POWER9, NVIDIA Volta GV100 – Lawrence Livermore National Lab (USA)
#coeurs=1,572,480 RMax=94,640 RPeak=125,712 Power=7,438

4 **Sunway TaihuLight** – Sunway MPP (Chine)
#coeurs=10,649,600 RMax=93,014 RPeak=125,435 Power=15,371 → **#1 en 2016**

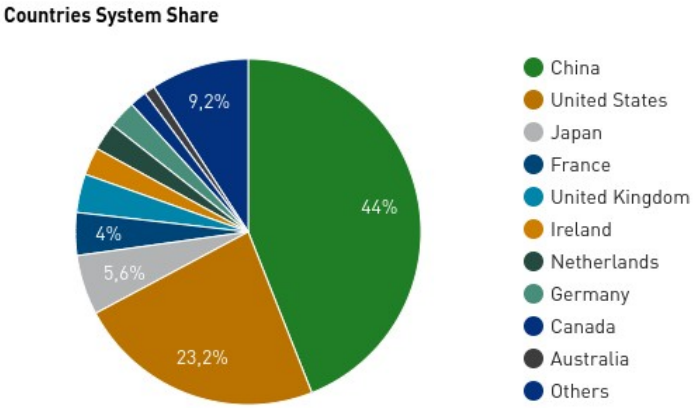
5 **Tianhe-2A** – Intel Xeon E5-2692v2, Matrix 2000 – National Super Computer Center in Guangzhou (Chine)
#coeurs=4,981,760 RMax=61,444 RPeak=100,678 Power=18,482 → **#1 en 2013**

500 Graham Huawei X6800 V3, Xeon E5-2683 v4 16C 2.1GHz, Infiniband EDR/FDR, NVIDIA Tesla P100 (Canada)
#coeurs=5,1200 RMax=1,228 Rpeak=2,641 Power =546 → **#96 en 2017**

Top 5 français de juin 2020

Rmax et Rpeak en Tflop/s
Power en kW

- **15. PANGEA III** – IBM POWER9, NVIDIA Volta GV100 – Total Exploration Production
  #coeurs=291 024 Rmax=17860 Rpeak=25025 Power=1367
- **20. Terra-1000** – Bull Sequana X1000 – Xeon Phi 7250 – CEA
  #coeurs=561 408 Rmax=11965 Rpeak=23396 Power=3178
- **29 . Belenos** - Bull Sequana XH2000 , AMD EPYC 7742 64C 2.25GHz, Mellanox HDR100, Atos Meteo France
  #coeurs=294,912 Rmax=7,683.4 Rpeak=10,469.4 Power=1,655
- **34 . JOLIOT-CURIE ROME**- Bull Sequana X1000 – Xeon Platinium – CEA/TGCC GENCI
  #coeurs=197,120 Rmax=6,988 Rpeak=12,039.4 Power=1,436
- **49 . Pangea** - SGI ICE X, Xeon E5-2670 8C 2.600GHz – Total Exploration Production → #**11 en 2016**
  #coeurs=220 800 Rmax=5283 Rpeak=6712 Power=4150
- **54. Jean Zay –** HPE SGI, Xeon gold6248 20C 2.5GHz, Nvidia Tesla V100 – CNRS / IDRIS-GENCI
  #coeurs=93 960 Rmax=4478 Rpeak=7345 Power= ?



Green 500 en 2020

- **Meilleur ratio performance/consommation d'énergie**

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	393	**MN-3** - MN-Core Server, Xeon 8260M 24C 2.4GHz, MN-Core, RoCEv2/MN-Core DirectConnect, Preferred Networks Preferred Networks Japan	2,080	1,621.1	77	21.108
2	7	**Selene** - DGX A100 SuperPOD, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	272,800	27,580.0	1,344	20.518
3	468	**NA-1** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz, PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	1,271,040	1,303.2	80	18.433
4	204	**A64FX prototype** - Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D, Fujitsu Fujitsu Numazu Plant Japan	36,864	1,999.5	118	16.876
5	26	**AiMOS** - IBM Power System AC922, IBM POWER9 20C 3.45GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM Rensselaer Polytechnic Institute Center for Computational Innovations (CCI) United States	130,000	8,339.0	512	16.285

- ***La plupart sont sont accelerator-based.***
- ***Trend of aggregating many low-power processors tops the Green500***

Points critiques des applications HPC

En terme de mise en œuvre :

- **Problèmes classiques exacerbés**
- **Utilisation des ressources matérielles**
- **Distribution des données**
- **Diffusion des résultats**
  - Opérations collectives (*alltoall*, *broadcast*, réduction, etc.)
- **Problèmes liés à la taille des plate-formes**
  - Tolérance aux fautes, etc.

En terme d'efficacité :

- **Localité des données**

- **Granularité des données**
  - Équilibrage de charge coûteux

Assistance aux applications HPC

- **Interfaces d'abstraction du matériel**
  - Exemples : OpenCL, PVM, MPI, etc.
  - Pour garantir la portabilité logicielle

Assistance aux applications HPC

- **Supports d'exécution**
  - Architectures multiprocesseurs
    - Gestion du placement des fils d'exécution
    - Gestion du placement des données
    - Ordonnanceur de threads, Cilk, etc.
  - Architectures distribuées :
    - Distribution des données
    - Gestion des transferts de données
    - MPI, RPC, DSM , etc
  - Architectures hétérogènes
    - Équilibrage de charge
    - StarSs, Intel Ct, StarPU, etc.

Pour garantir la portabilité des performances!

Assistance aux applications HPC

- **Bibliothèques de plus haut niveau**
  - Résolution intégrée de problèmes *classiques*
    - FFT, algèbre linéaire (BLAS, etc.), résolution systèmes linéaires, etc.
- **Utilitaires**
  - Analyse de performance (PAPI, Perf, Vampir, Scalasca, etc.)
  - Détection de bugs (Valgrind, gdb, etc.)
  - Tolérance aux pannes, aux fautes
  - .....
- **Middlewares**
  - Chaque brique développée de manière isolée
  - Couplage de code
  - Arbitrage des accès aux ressources physiques

Métriques

- **Performances dépendantes de plusieurs facteurs**
  - Fraction de l'application parallélisable
  - Qualité de l'ordonnancement sur les ressources de calcul
  - Surcoût introduit par la version parallèle

- **Speedup :** mesure l'accélération entre les versions parallèle et séquentielle
  - Sp = Tseq /Tp , où

    Sp = speedup sur P processeurs

    Tp = temps de la version parallèle sur P processeurs

    Tseq = temps de la version séquentielle sur 1 processeur
  - Objectif ultime : Sp = P

- **Loi d'Amdahl :** borne d'accélération en fonction de la qualité de parallélisation
  - R = 1 /((1−S)+ S/P) , où

    S = proportion de code parallélisé, P = le nombre de processeurs
  - L'accélération est bornée à 1/S − > ajout de processeurs rapidement inutile