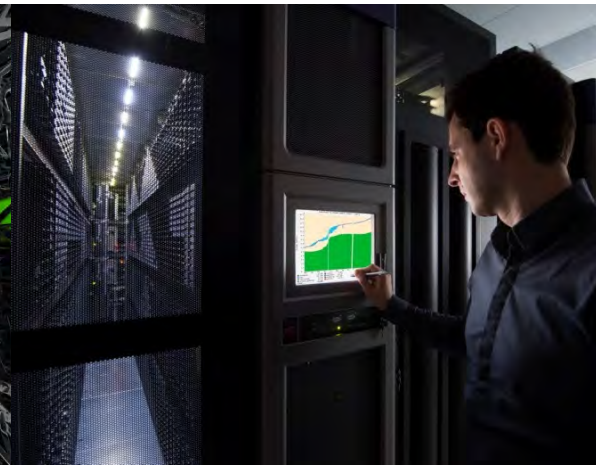
DE LA RECHERCHE À L'INDUSTRIE

cea

PANORAMA DES SUPERCALCULATEURS À L'ÈRE DU « MULTI-PETASCALE »



ASPROM

CALCULS PARALLELES ET APPLICATIONS

30 septembre 2014

Jean-Philippe Nominé
CEA/DIF/DSSI
jean-philippe.nomine@cea.fr

www.cea.fr

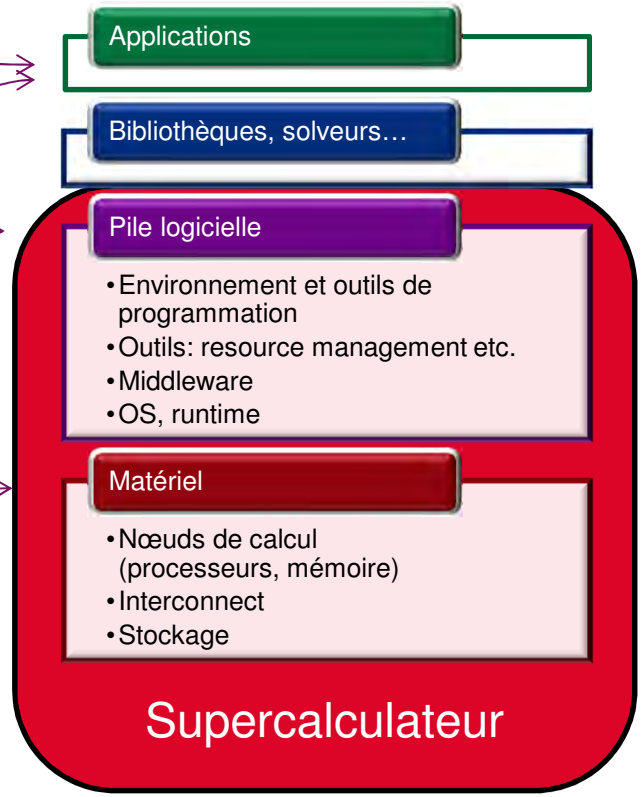
RÉSUMÉ

Le classement Top500 de **juin 2014** recense **55** supercalculateurs plus que « pétaflopiques » au sens de la performance crête – **37** au sens du benchmark LINPACK (HPL). Il est donc possible de prendre un peu de recul sur cette ère du « petascale » désormais bien entamée. Nous survolerons ainsi :

- les architectures de grands calculateurs d'aujourd'hui, dans ce '~Top50' mais aussi dans les segments d'usage plus généraux du HPC (High Performance Computing) à différentes échelles
- les tendances observables, ou pas, alors que l'horizon 2020, si fréquemment érigé depuis quelques années en « mur du son » de l'exascale, se rapproche doucement

PREAMBULE

- Le matériel passe (~5 ans),
  le logiciel reste (jusqu'à 30 ans…?).
- Le logiciel: le vrai patrimoine au travers des codes de calcul scientifiques et industriels - avec le savoir-faire des équipes des centres de calcul !
- Logiciel (stack; applications…)
  Voir les autres orateurs
- Nous allons parler un peu plus « architecture générale » - avec comme fil directeur les grands calculateurs 'Top500'



PETAFLOPS, PETASCALE etc.

- PETA 10^{15} million de milliards
- EXA 10^{18} milliard de milliards (PETA x 1000)

- « FLOP/S »
  renvoie à une puissance de calcul crête ou benchmarkée (HPL etc.)
- « SCALE »
  renvoie plutôt à l'idée d'un usage « soutenu » et plus banalisé

- Machine petascale ≠ application petascale….

DEFINITION: HPC

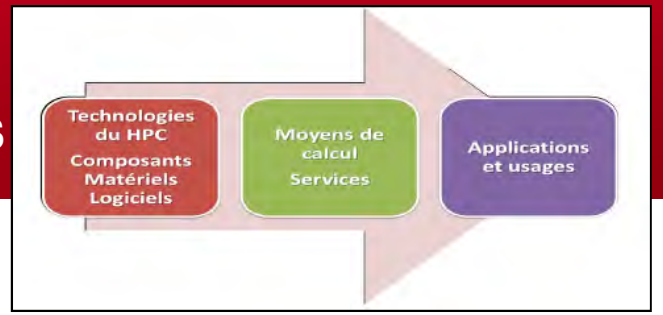
Toute activité de calcul :

- dépassant les capacités d'un système « standard » ou pur « commodity », afin de répondre à des impératifs de taille de problème ou de rapidité de traitement
- recourant au parallélisme et à l'intégration dense et généralement hiérarchisée de nombreuses unités de calcul (cœurs ; processeurs ; serveurs) et de mémoire à l'aide de réseaux d'interconnexion locaux rapides
- utilisant des composants électroniques standard et parfois plus spécifiques mais assemblés de manière spécialisée pour concourir aux objectifs ci-dessus, depuis des échelles modérées jusqu'aux plus extrêmes (petaflops aujourd'hui, exaflops dans quelques années)

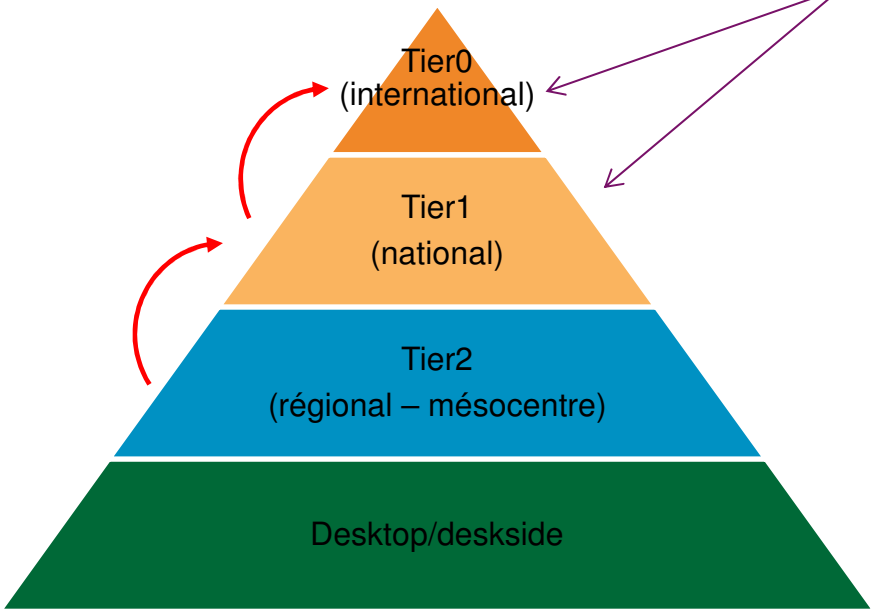
Le HPC peut être considéré comme la frontière (mouvante) de l'exploitation du calcul aux limites de ce qui peut se faire à un moment donné – « extreme computing ». Il se décline ensuite à différentes échelles, et tend à :

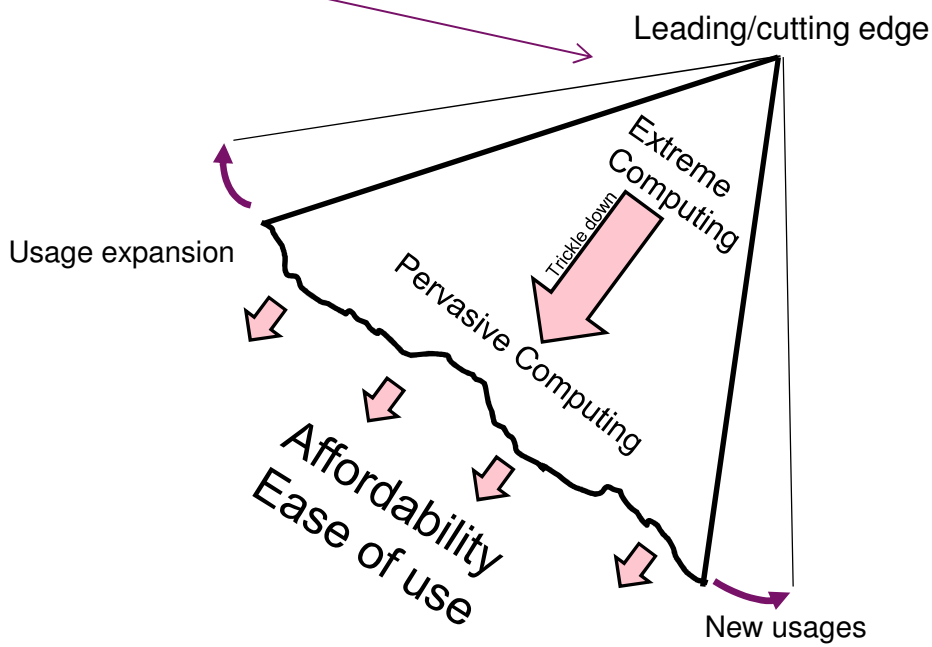
- pénétrer des domaines (segments de marché) et des usages de plus en plus diversifiés, en plus de la simulation numérique scientifique et industrielle habituellement considérée comme son terrain d'expression naturel, en mode 'capability' ou 'capacity' – tels que l'usage dans des grands instruments, dans des boucles de contrôle et de régulation, dans des dispositifs embarqués ou nomades, dans le traitement de grandes masses de données (big data).
- se prêter à des modes d'accès et d'utilisation également en cours de diversification : par exemple en nuage/à la demande (cloud computing), en interactif, en temps critique ou en temps réel

CHAINE DE VALEUR
TECHNOLOGIES - INFRASTRUCTURES - USAGES



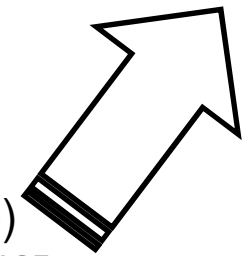
PETASCALE





Vision ‘infrastructure’
La pyramide des moyens
(ex. PRACE)

Vision ‘dynamique écosystème’
Des technologies aux usages
La flèche du HPC (ex. ETP4HPC)

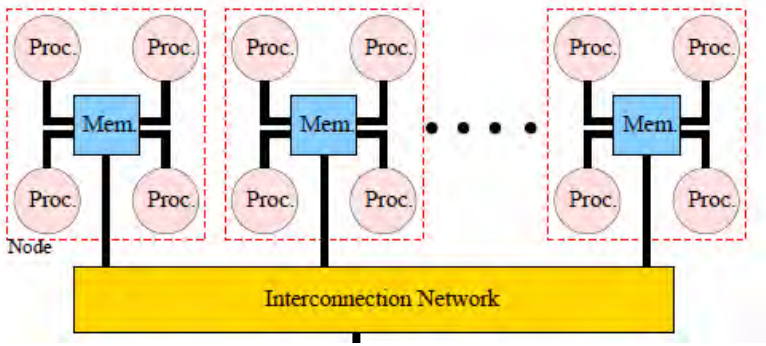


SUPERCALCULATEUR

Source/crédit illustrations et citation:
Overview of recent supercomputers
Aad J. van der Steen
EuroBen Foundation, SURFSara - The Netherlands
ajsteen1@euroben.nl
www.euroben.nl
http://www.euroben.nl/reports/overview13.pdf

Intra-nœud: mémoire partagée
Inter-nœud: mémoire distribuée

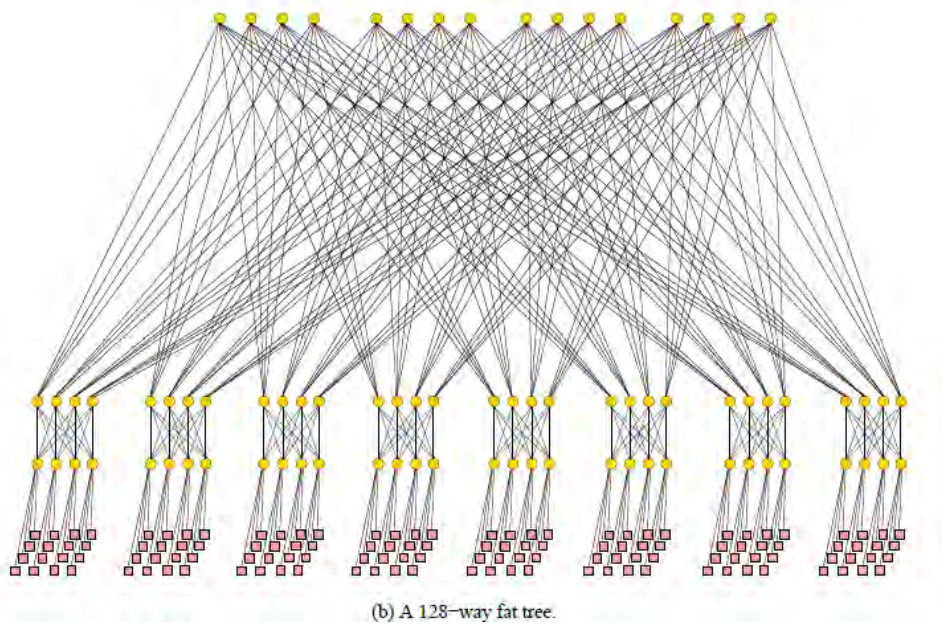
Unités de calcul
Interconnect
I/O & stockage



"The majority of systems, however, still look like minor variations on the same theme: clusters of Symmetric Multi-Processing (SMP) nodes are connected by a fast network. Culler et.al. [7] consider this as a natural architectural evolution. However, it may also be argued that economic pressure has steered the systems in this direction, thus letting the High Performance Computing world take advantage of the mainstream technology developed for the mass market."

Principe…

Exemple de réalisation -
en topologie « fat tree »
(autres: tore, mesh, hypercube…)


(b) A 128−way fat tree.

TOP 500

http://www.top500.org/lists/2014/06/

Table 2. Algorithms expected to play a key role within select scientific applications at the exascale, characterized according to a seven dwarfs classification

Opportunity	Application area	Structured grids	Unstructured grids	FFT	Dense linear algebra	Sparse linear algebra	Particles	Monte Carlo
Material science	Molecular physics			X	X		X	X
	Nanoscale science	X			X		X	X
Earth science	Climate	X	X	X		X	X	X
	Environment	X	X			X	X	X
Energy assurance	Combustion	X			X		X	
	Fusion	X	X	X	X	X	X	X
	Nuclear energy		X		X	X		
Fundamental science	Astrophysics	X	X		X	X	X	
	Nuclear physics				X			
	Accelerator physics		X			X		
	QCD	X						X
Engineering design	Aerodynamics	X	X		X	X		

FFT = Fast Fourier Transform
QCD = Quantum chromodynamics

Scientific Application Requirements for Leadership Computing at the Exascale, ORNL, December 2007 - ORNL/TM-2007/238

- Intérêt
  - Benchmark « simple » HP Linpack (algèbre linéaire dense)
  - Populaire
  - Long suivi historique
- Limites
  - La vraie vie n'est pas si 'algèbre linéaire dense' que cela… (pas que…)
  - Forte corrélation RPEAK/RMAX (% prévisible…)
  - Parties non visibles? Les non-inscrits…
    Cf aussi 'Boycott' NCSA BlueWaters
    http://www.ncsa.illinois.edu/news/stories/TOP500problem/ (W. Kramer: Top problems with the Top500)
- Discussion
  - Les prescripteurs et fournisseurs optimisent en général pour le Top500 ET un certain profil de production réelle…
  - Autres options de benchmarking
    - HPCG (HP Conjugate Gradient) – J. Dongarra et al.
    - Benchmarks synthétiques (fonctionnalités ciblées)
    - Mini/proxy apps
    - Green 500
    - Graph 500
    - …

HPCG @ ISC 14 – JUIN 2014

Site	Computer	Cores	HPL Rmax (Pflops)	HPL Rank	HPCG (Pflops)
NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.9	1	.580
RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx 8C + Custom	705,024	10.5	4	.427
DOE/OS Oak Ridge Nat Lab	Titan, Cray XK7 AMD 16C + Nvidia Kepler GPU 14C + Custom	560,640	17.6	2	.322
DOE/OS Argonne Nat Lab	Mira BlueGene/Q, Power BQC 16C 1.60GHz + Custom	786,432	8.59	5	.101#
Swiss CSCS	Piz Daint, Cray XC30, Xeon 8C + Nvidia Kepler 14C + Custom	115,984	6.27	6	.099
Leibniz Rechenzentrum	SuperMUC, Intel 8C + IB	147,456	2.90	12	.0833
CEA/TGCC-GENCI	Curie tine nodes Bullx B510 Intel Xeon 8C 2.7 GHz + IB	79,504	1.36	26	.0491
Exploration and Production Eni S.p.A.	HPC2, Intel Xeon 10C 2.8 GHz + Nvidia Kepler 14C + IB	62,640	3.00	11	.0489
DOE/OS L Berkeley Nat Lab	Edison Cray XC30, Intel Xeon 12C 2.4GHz + Custom	132,840	1.65	18	.0439 #
Texas Advanced Computing Center	Stampede, Dell Intel (8c) + Intel Xeon Phi (61c) + IB	78,848	.881*	7	.0161
Meteo France	Beaufix Bullx B710 Intel Xeon 12C 2.7 GHz + IB	24,192	.469 (.467*)	79	.0110
Meteo France	Prolix Bullx B710 Intel Xeon 2.7 GHz 12C + IB	23,760	.464 (.415*)	80	.00998
U of Toulouse	CALMIP Bullx DLC Intel Xeon 10C 2.8 GHz + IB	12,240	.255	184	.00725
Cambridge U	Wilkes, Intel Xeon 6C 2.6 GHz + Nvidia Kepler 14C + IB	3584	.240	201	.00385

https://software.sandia.gov/hpcg/html/index.html
http://www.sandia.gov/~maherou/docs/HPCG-Benchmark.pdf

TOP 500

http://www.top500.org/

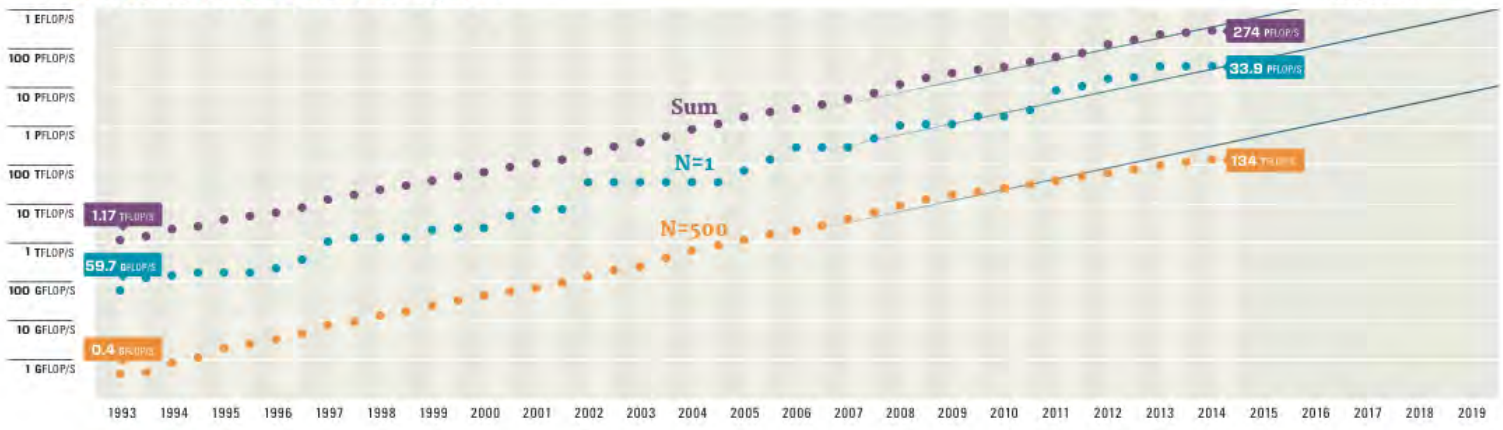
TOP500® JUNE 2014

PRESENTED BY ICL UT INNOVATIVE COMPUTING LABORATORY · BERKELEY LAB Lawrence Berkeley National Laboratory · PROMETEUS

FIND OUT MORE AT www.top500.org

	NAME	SPECS	SITE	COUNTRY	CORES	RMAX PFLOP/S	POWER MW
1	Tianhe-2 (Milkyway-2)	NUDT, Intel Ivy Bridge (12C, 2.2 GHz) & Xeon Phi (57C, 1.1 GHz), Custom interconnect	NSCC Guangzhou	China	3,120,000	33.9	17.8
2	Titan	Cray XK7, Operon 6274 (16C 2.2 GHz) + Nvidia Kepler GPU, Custom interconnect	DOE/SC/ORNL	USA	560,640	17.6	8.2
3	Sequoia	IBM BlueGene/Q, Power BQC (16C 1.60 GHz), Custom interconnect	DOE/NNSA/LLNL	USA	1,572,864	17.2	7.9
4	K computer	Fujitsu SPARC64 VIIIfx (8C, 2.0GHz), Custom interconnect	RIKEN AICS	Japan	705,024	10.5	12.7
5	Mira	IBM BlueGene/Q, Power BQC (16C, 1.60 GHz), Custom interconnect	DOE/SC/ANL	USA	786,432	8.59	3.95

PERFORMANCE DEVELOPMENT



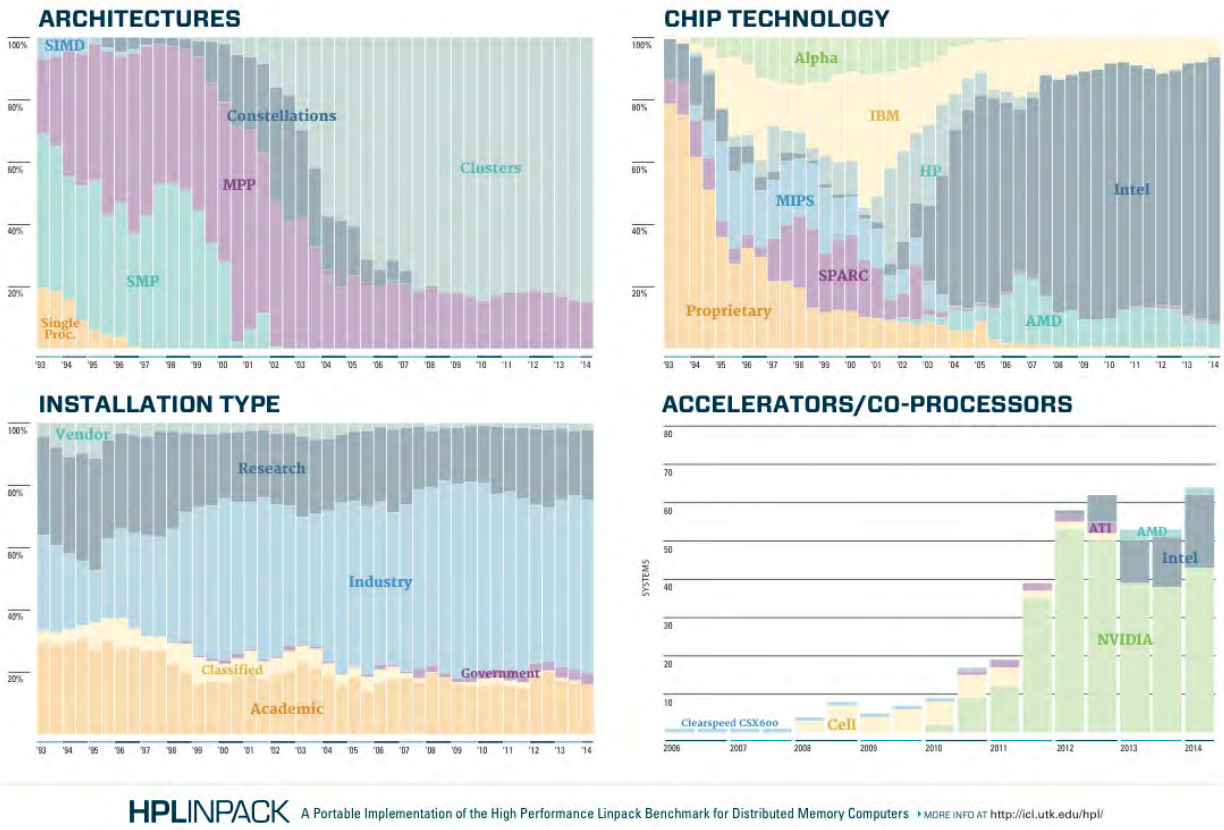
1. Hybride x86/MIC
2. Hybride x86/GPU
3. RISC
4. RISC
5. RISC

5x custom IC…

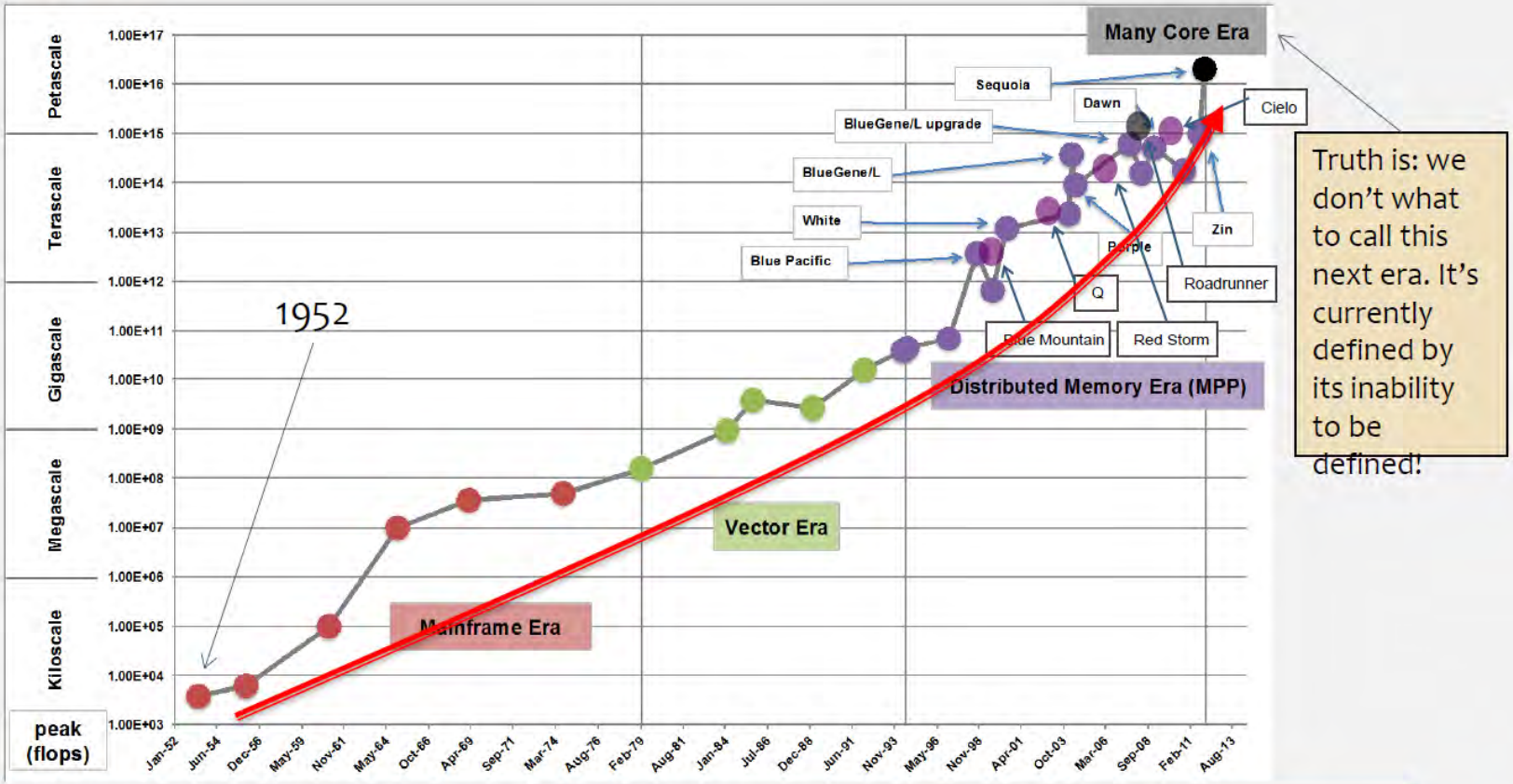
http://s.top500.org/static/lists/2014/06/top500_201406_poster.png

- In a cluster, each machine is largely independent of the others in terms of memory, disk, etc. They are interconnected using some variation on normal networking. The cluster exists mostly in the mind of the programmer and how s/he chooses to distribute the work.
  Ex: K (?), Bullx, iDataPlex…
- In a Massively Parallel Processor, there really is only one machine with thousands of CPUs tightly interconnected. MPPs have exotic memory architectures to allow extremely high speed exchange of intermediate results with neighboring processors.
  Ex: BG, Cray…
- Constellation - a cluster of large SMP nodes, where the number of processors per node is greater than the number of nodes.

  - Academic = au sens US (Univ.)
  - Research = p.ex. NSF ou DOE lab
  - Industry = souvent ‘anonyme’… comme ‘government’

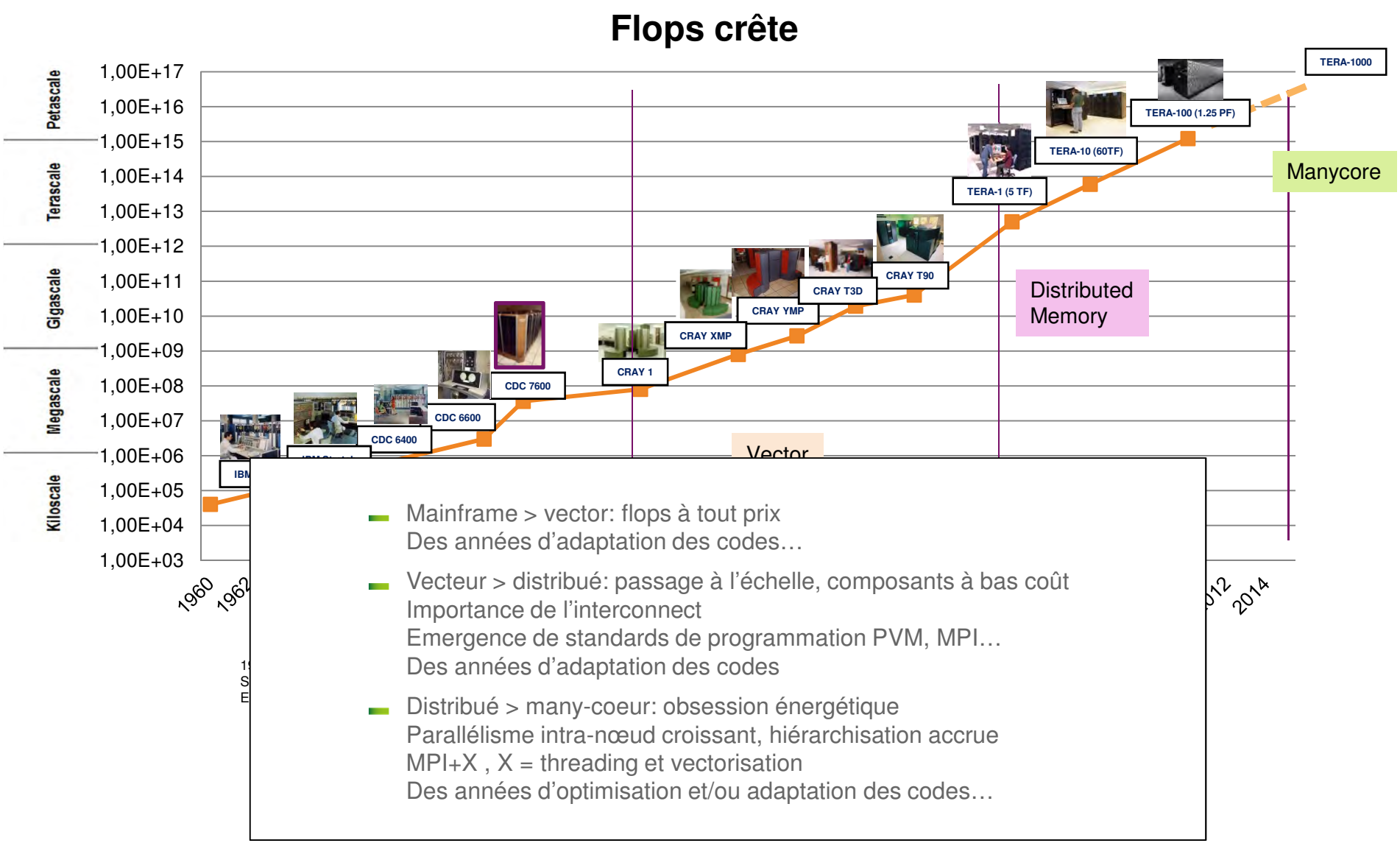


Advancements in (High Performance) Computing Have Occurred in Several Distinct "Eras"



Each of these eras define not so much a common hardware architecture, but a common programming model

CEA DAM (un échantillon…)



- Mainframe > vector: flops à tout prix
  Des années d'adaptation des codes…
- Vecteur > distribué: passage à l'échelle, composants à bas coût
  Importance de l'interconnect
  Emergence de standards de programmation PVM, MPI…
  Des années d'adaptation des codes
- Distribué > many-coeur: obsession énergétique
  Parallélisme intra-nœud croissant, hiérarchisation accrue
  MPI+X , X = threading et vectorisation
  Des années d'optimisation et/ou adaptation des codes…

L'ERE DU PETASCALE : ~TOP 50…

- 55 supercalculateurs plus que « pétaflopiques » au sens de la performance crête dans le TOP500
- 37 au sens du benchmark LINPACK (HPL) (liste de Juin 2014)

- Accelerated: use co-processors to handle part of the load (in red, 21 systems)
- Lightweight: use many low-power RISC processors (in green, 11 systems)
- Traditional: use only standard high-performance processors (in blue, 23 systems)

System	Site (Country)	Model (processor / accelerator)	LINPACK / peak (PFlop/s)
Tianhe-2	NSCC-GZ (China)	NUDT TH-IVB (Intel Xeon / Intel Xeon Phi)	33.86 / **54.90**
Titan	ORNL (USA)	Cray XK7 (AMD Opteron / NVIDIA Tesla)	17.59 / **27.11**
Sequoia	LLNL (USA)	IBM BlueGene/Q (IBM PowerPC)	17.17 / **20.13**
K Computer	RIKEN (Japan)	Fujitsu Cluster (Fujitsu SPARC64)	10.51 / **11.28**
Mira	ANL (USA)	IBM BlueGene/Q (IBM PowerPC)	8.59 / **10.07**
Piz Daint	CSCS (Switzerland)	Cray XC30 (Intel Xeon / NVIDIA Tesla)	6.27 / **7.79**
Stampede	TACC (USA)	Dell PowerEdge (Intel Xeon / Intel Xeon Phi)	5.17 / **8.52**
JUQUEEN	FZJ (Germany)	IBM BlueGene/Q (IBM PowerPC)	5.01 / **5.87**
Vulcan	LLNL (USA)	IBM BlueGene/Q (IBM PowerPC)	4.29 / **5.03**
Anonymous	Government (USA)	Cray XC30 (Intel Xeon)	3.14 / **4.88**
HPC2	Eni S.p.A. (Italy)	IBM iDataPlex (Intel Xeon / NVIDIA Tesla)	3.00 / **4.00**
SuperMUC	LRZ (Germany)	IBM iDataPlex (Intel Xeon)	2.90 / **3.19**
TSUBAME 2.5	GSIC (Japan)	NEC/HP ProLiant (Intel Xeon / NVIDIA Tesla)	2.79 / **5.74**
Tianhe-1A	NSCT (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	2.57 / **4.70**
Cascade	EMSL (USA)	Atipa Visione (Intel Xeon / Intel Xeon Phi)	2.54 / **3.39**
Pangea	Total EP (France)	SGI ICE X (Intel Xeon)	2.10 / **2.30**
Fermi	CINECA (Italy)	IBM BlueGene/Q (IBM PowerPC)	1.79 / **2.10**
Edison	NERSC (USA)	Cray XC30 (Intel Xeon)	1.65 / **2.57**
Anonymous	ECMWF (UK)	Cray XC30 (Intel Xeon)	1.55 / **1.80**
Anonymous	ECMWF (UK)	Cray XC30 (Intel Xeon)	1.55 / **1.80**
Pleiades	NASA (USA)	SGI ICE X (Intel Xeon)	1.54 / **2.11**
DARPA TS	IBM DE (USA)	IBM Power 775 (IBM POWER7)	1.52 / **1.94**
Blue Joule	STFC (UK)	IBM BlueGene/Q (IBM PowerPC)	1.43 / **1.68**
Spirit	AFRL (USA)	SGI ICE X (Intel Xeon)	1.42 / **1.53**
Archer	EPSRC (UK)	Cray XC30 (Intel Xeon)	1.37 / **1.65**
Curie TN	TGCC (France)	Bull B510 (Intel Xeon)	1.36 / **1.67**
Hydra TN	RZG-MPG (Germany)	IBM iDataPlex (Intel Xeon)	1.28 / **1.46**
Nebulae	NSCS (China)	Dawning TC3600 (Intel Xeon / NVIDIA Tesla)	1.27 / **2.98**
Yellowstone	NCAR (USA)	IBM iDataPlex (Intel Xeon)	1.26 / **1.50**
Helios	IFERC (Japan)	Bull B510 (Intel Xeon)	1.24 / **1.52**
Garnet	ERDC (USA)	Cray XE6 (AMD Opteron)	1.17 / **1.51**
Cielo	LANL (USA)	Cray XE6 (AMD Opteron)	1.11 / **1.37**
DiRAC	EPCC (UK)	IBM BlueGene/Q (IBM PowerPC)	1.07 / **1.26**
Hopper	NERSC (USA)	Cray XE6 (AMD Opteron)	1.05 / **1.29**
Tera-100	CEA (France)	Bull S6010/S6030 (Intel Xeon)	1.05 / **1.25**
Oakleaf-FX	SCD (Japan)	Fujitsu PRIMEHPC (Fujitsu SPARC64)	1.04 / **1.14**
Quartetto	RIIT-KU (Japan)	Hitachi/Fujitsu PRIMERGY (Intel Xeon / NVIDIA Tesla / Intel Xeon Phi)	1.02 / **1.50**
Raijin	NCI (Australia)	Fujitsu PRIMERGY (Intel Xeon)	0.98 / **1.11**
Conte	Purdue (USA)	HP ProLiant (Intel Xeon / Intel Xeon Phi)	0.96 / **1.34**
MareNostrum	BSC (Spain)	IBM iDataPlex (Intel Xeon)	0.93 / **1.02**
Lomonosov	RCC (Russia)	T-Platforms T-Blade (Intel Xeon / NVIDIA Tesla)	0.90 / **1.70**
Anonymous	RPI (USA)	IBM BlueGene/Q (IBM PowerPC)	0.89 / **1.05**
Hermit	HLRS (Germany)	Cray XE6 (AMD Opteron)	0.83 / **1.04**
Sunway BL	NSC (China)	Sunway Cluster (ShenWei SW1600)	0.80 / **1.07**
Tianhe-1A HS	NSCCH (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	0.77 / **1.34**
COMA	CCS-UT (Japan)	Cray CS300 (Intel Xeon / Intel Xeon Phi)	0.75 / **1.11**
Hydra AN	RZG-MPG (Germany)	IBM iDataPlex (Intel Xeon / NVIDIA Tesla)	0.71 / **1.01**
Big Red II	IU (USA)	Cray XK7 (AMD Opteron / NVIDIA Tesla)	0.60 / **1.00**
SANAM	KAUST (Saudi Arabia)	Adtech custom (Intel Xeon / AMD FirePro)	0.53 / **1.10**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon)	0.50 / **1.13**
Mole-8.5	IPE (China)	Tyan FT72-B7015 (Intel Xeon / NVIDIA Tesla)	0.50 / **1.01**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.46 / **1.14**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon)	0.43 / **1.09**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.42 / **1.08**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.29 / **1.05**

Source PRACE 2IP

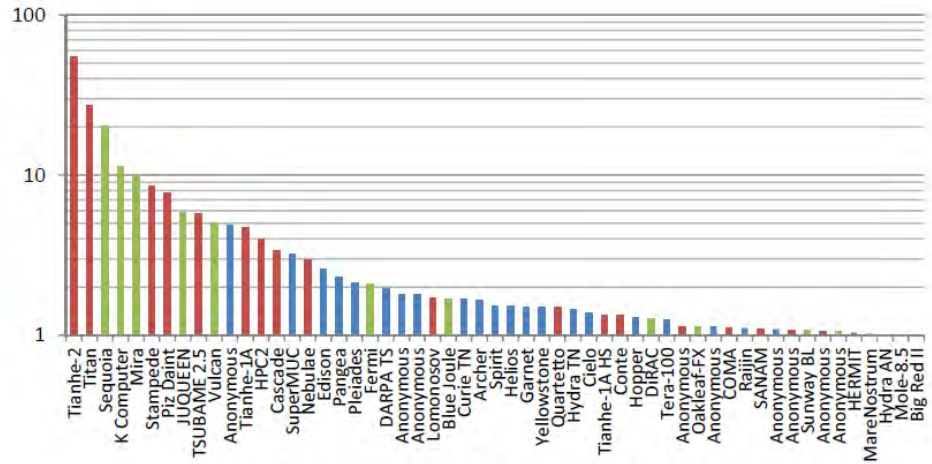


Figure 6: Peak performance of petascale systems (in PFlop/s)
(Red = accelerated, green = lightweight, blue = traditional)

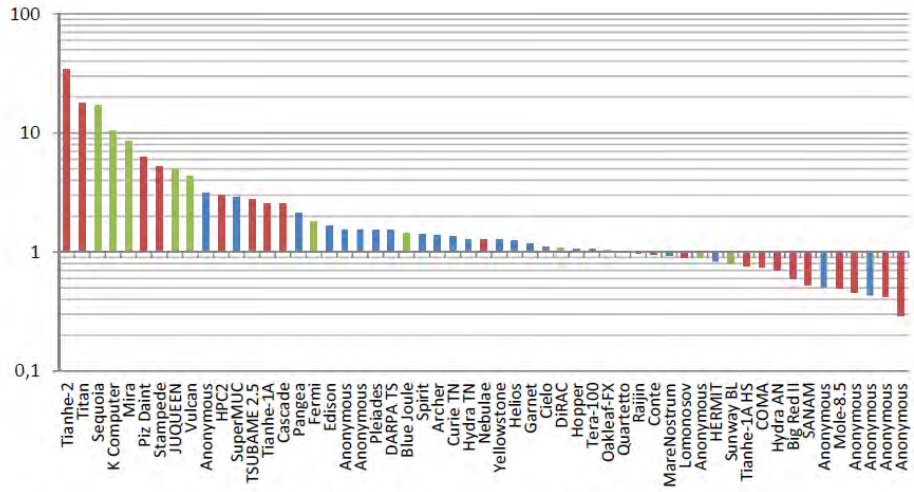
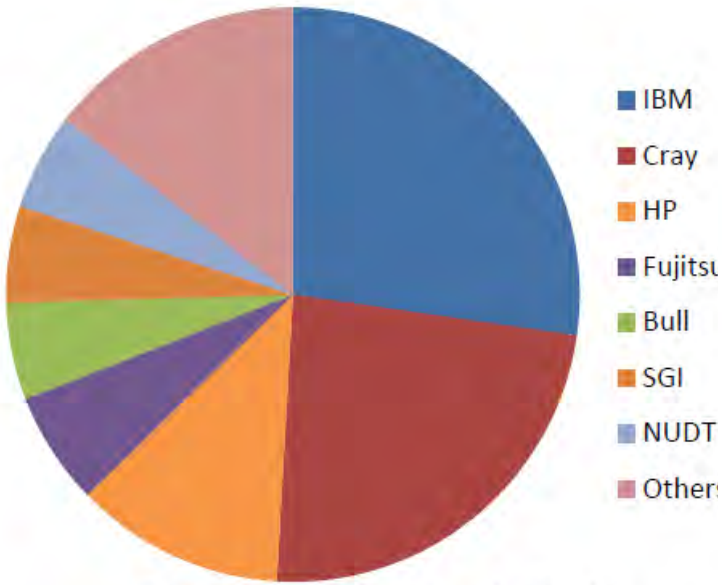


Figure 7: LINPACK performance of petascale systems (in PFlop/s)
(Red = accelerated, green = lightweight, blue = traditional)

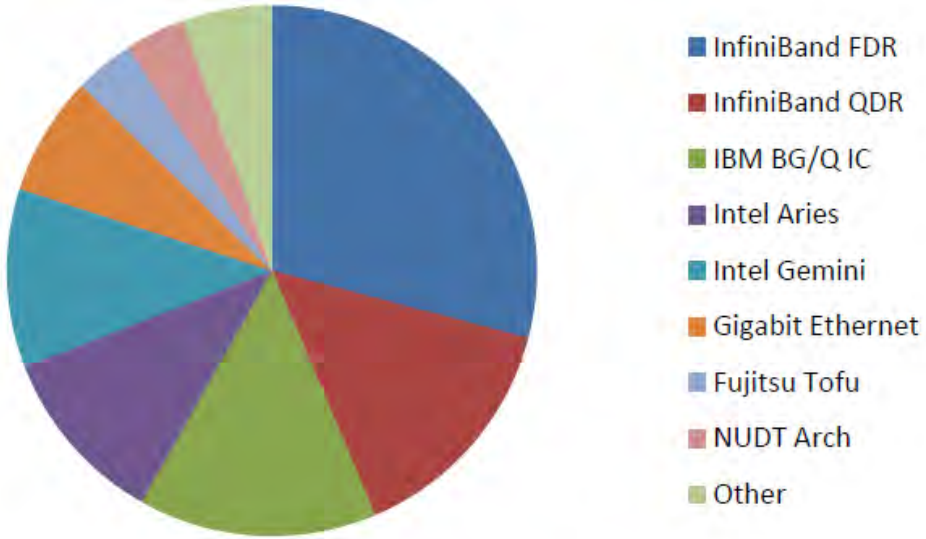
- Accelerated: use co-processors to handle part of the load (in red, 21 systems)
- Lightweight: use many low-power RISC processors (in green, 11 systems)
- Traditional: use only standard high-performance processors (in blue, 23 systems)

System	Site (Country)	Model (processor / accelerator)	LINPACK / peak (PFlop/s)
Tianhe-2	NSCC-GZ (China)	NUDT TH-IVB (Intel Xeon / Intel Xeon Phi)	33.86 / **54.90**
Titan	ORNL (USA)	Cray XK7 (AMD Opteron / NVIDIA Tesla)	17.59 / **27.11**
Sequoia	LLNL (USA)	IBM BlueGene/Q (IBM PowerPC)	17.17 / **20.13**
K Computer	RIKEN (Japan)	Fujitsu Cluster (Fujitsu SPARC64)	10.51 / **11.28**
Mira	ANL (USA)	IBM BlueGene/Q (IBM PowerPC)	8.59 / **10.07**
Piz Daint	CSCS (Switzerland)	Cray XC30 (Intel Xeon / NVIDIA Tesla)	6.27 / **7.79**
Stampede	TACC (USA)	Dell PowerEdge (Intel Xeon / Intel Xeon Phi)	5.17 / **8.52**
JUQUEEN	FZJ (Germany)	IBM BlueGene/Q (IBM PowerPC)	5.01 / **5.87**
Vulcan	LLNL (USA)	IBM BlueGene/Q (IBM PowerPC)	4.29 / **5.03**
Anonymous	Government (USA)	Cray XC30 (Intel Xeon)	3.14 / **4.88**
HPC2	Eni S.p.A. (Italy)	IBM iDataPlex (Intel Xeon / NVIDIA Tesla)	3.00 / **4.00**
SuperMUC	LRZ (Germany)	IBM iDataPlex (Intel Xeon)	2.90 / **3.19**
TSUBAME 2.5	GSIC (Japan)	NEC/HP ProLiant (Intel Xeon / NVIDIA Tesla)	2.79 / **5.74**
Tianhe-1A	NSCT (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	2.57 / **4.70**
Cascade	EMSL (USA)	Atipa Visione (Intel Xeon / Intel Xeon Phi)	2.54 / **3.39**
Pangea	Total EP (France)	SGI ICE X (Intel Xeon)	2.10 / **2.30**
Fermi	CINECA (Italy)	IBM BlueGene/Q (IBM PowerPC)	1.79 / **2.10**
Edison	NERSC (USA)	Cray XC30 (Intel Xeon)	1.65 / **2.57**
Anonymous	ECMWF (UK)	Cray XC30 (Intel Xeon)	1.55 / **1.80**
Anonymous	ECMWF (UK)	Cray XC30 (Intel Xeon)	1.55 / **1.80**
Pleiades	NASA (USA)	SGI ICE X (Intel Xeon)	1.54 / **2.11**
DARPA TS	IBM DE (USA)	IBM Power 775 (IBM POWER7)	1.52 / **1.94**
Blue Joule	STFC (UK)	IBM BlueGene/Q (IBM PowerPC)	1.43 / **1.68**
Spirit	AFRL (USA)	SGI ICE X (Intel Xeon)	1.42 / **1.53**
Archer	EPSRC (UK)	Cray XC30 (Intel Xeon)	1.37 / **1.65**
Curie TN	TGCC (France)	Bull B510 (Intel Xeon)	1.36 / **1.67**
Hydra TN	RZG-MPG (Germany)	IBM iDataPlex (Intel Xeon)	1.28 / **1.46**
Nebulae	NSCS (China)	Dawning TC3600 (Intel Xeon / NVIDIA Tesla)	1.27 / **2.98**
Yellowstone	NCAR (USA)	IBM iDataPlex (Intel Xeon)	1.26 / **1.50**
Helios	IFERC (Japan)	Bull B510 (Intel Xeon)	1.24 / **1.52**
Garnet	ERDC (USA)	Cray XE6 (AMD Opteron)	1.17 / **1.51**
Cielo	LANL (USA)	Cray XE6 (AMD Opteron)	1.11 / **1.37**
DiRAC	EPCC (UK)	IBM BlueGene/Q (IBM PowerPC)	1.07 / **1.26**
Hopper	NERSC (USA)	Cray XE6 (AMD Opteron)	1.05 / **1.29**
Tera-100	CEA (France)	Bull S6010/S6030 (Intel Xeon)	1.05 / **1.25**
Oakleaf-FX	SCD (Japan)	Fujitsu PRIMEHPC (Fujitsu SPARC64)	1.04 / **1.14**
Quartetto	RIIT-KU (Japan)	Hitachi/Fujitsu PRIMERGY (Intel Xeon / NVIDIA Tesla / Intel Xeon Phi)	1.02 / **1.50**
Raijin	NCI (Australia)	Fujitsu PRIMERGY (Intel Xeon)	0.98 / **1.11**
Conte	Purdue (USA)	HP ProLiant (Intel Xeon / Intel Xeon Phi)	0.96 / **1.34**
MareNostrum	BSC (Spain)	IBM iDataPlex (Intel Xeon)	0.93 / **1.02**
Lomonosov	RCC (Russia)	T-Platforms T-Blade (Intel Xeon / NVIDIA Tesla)	0.90 / **1.70**
Anonymous	RPI (USA)	IBM BlueGene/Q (IBM PowerPC)	0.89 / **1.05**
Hermit	HLRS (Germany)	Cray XE6 (AMD Opteron)	0.83 / **1.04**
Sunway BL	NSC (China)	Sunway Cluster (ShenWei SW1600)	0.80 / **1.07**
Tianhe-1A HS	NSCCH (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	0.77 / **1.34**
COMA	CCS-UT (Japan)	Cray CS300 (Intel Xeon / Intel Xeon Phi)	0.75 / **1.11**
Hydra AN	RZG-MPG (Germany)	IBM iDataPlex (Intel Xeon / NVIDIA Tesla)	0.71 / **1.01**
Big Red II	IU (USA)	Cray XK7 (AMD Opteron / NVIDIA Tesla)	0.60 / **1.00**
SANAM	KAUST (Saudi Arabia)	Adtech custom (Intel Xeon / AMD FirePro)	0.53 / **1.10**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon)	0.50 / **1.13**
Mole-8.5	IPE (China)	Tyan FT72-B7015 (Intel Xeon / NVIDIA Tesla)	0.50 / **1.01**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.46 / **1.14**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon)	0.43 / **1.09**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.42 / **1.08**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.29 / **1.05**

NB: EN NOMBRE DE SYSTEMES



Petascale systems by vendor



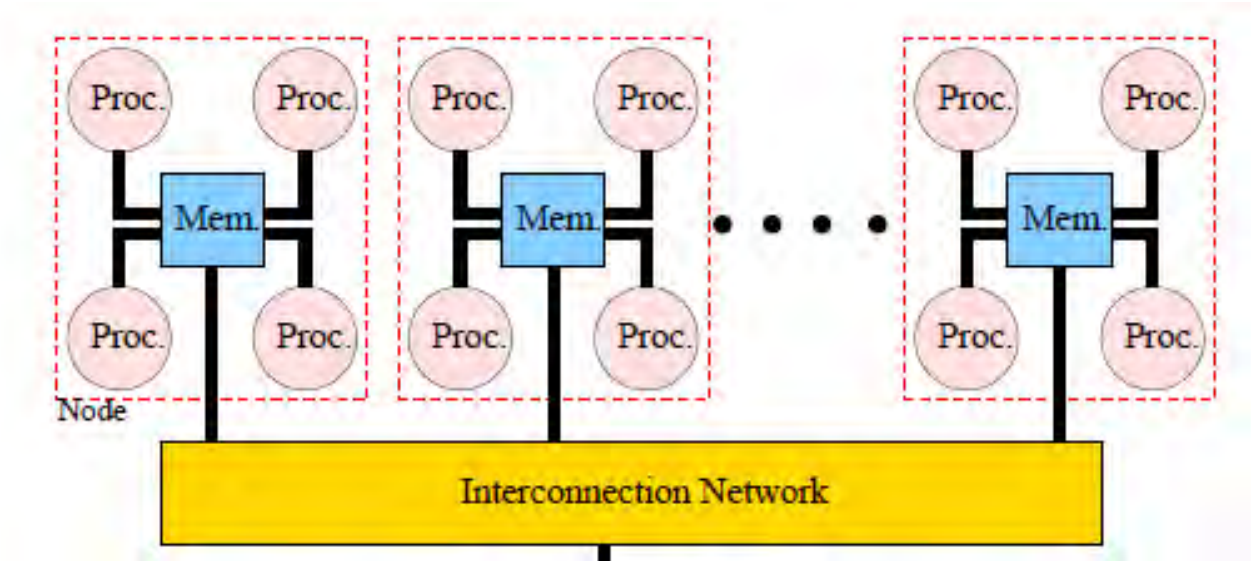
Petascale systems by interconnect

QUELQUES COMMENTAIRES

Rank	Name	Computer	Site	Manufacturer	Country	Year	Segment	Total Cores	Accelerator Co-Processor Cores	Rmax (PF)	Rpeak (PF)	Rmax/Rpeak	Power (MW)	Pflops/MW	Architecture	Processor	Processor Technology	Processor Speed (MHz)	Cores per Socket	Interconnect Family
1	Tianhe-2 (Mi	TH-IVB-FEP Cl	National Super C	NUDT	China	2013	Research	3120000	2736000	33,86	54,90	0,62	17,8	1,90	Cluster	Intel Xeon E5	Intel IvyBridge	2200	12,00	Custom
2	Titan	Cray XK7 , Opt	DOE/SC/Oak Ri	Cray Inc	USA	2012	Research	560640	261632	17,59	27,11	0,65	8,2	2,14	MPP	Opteron 6274	AMD x86_64	2200	16,00	Cray
3	Sequoia	BlueGene/Q, P	DOE/NNSA/LLN	IBM	USA	2011	Research	1572864	0	17,17	20,13	0,85	7,89	2,18	MPP	Power BQC 1	PowerPC	1600	16,00	Custom
4		K computer, SI	RIKEN Advance	Fujitsu	Japan	2011	Research	705024	0	10,51	11,28	0,93	12,66	0,83	Cluster	SPARC64 VI	Sparc	2000	8,00	Custom
5	Mira	BlueGene/Q, P	DOE/SC/Argonn	IBM	USA	2012	Research	786432	0	8,59	10,07	0,85	3,95	2,18	MPP	Power BQC 1	PowerPC	1600	16,00	Custom
6	Piz Daint	Cray XC30, Xe	Swiss National S	Cray Inc	Switzerland	2012	Research	115984	73808	6,27	7,79	0,81	2,33	2,70	MPP	Xeon E5-2670	Intel SandyB	2600	8,00	Custom
7	Stampede	PowerEdge C8	Texas Advanced	Dell	USA	2012	Academic	462462	366366	5,17	8,52	0,61	4,51	1,15	Cluster	Xeon E5-2680	Intel SandyB	2700	8,00	Infiniband
8	JUQUEEN	BlueGene/Q, P	Forschungszent	IBM	Germany	2012	Research	458752	0	5,01	5,87	0,85	2,30	2,18	MPP	Power BQC 1	PowerPC	1600	16,00	Custom
9	Vulcan	BlueGene/Q, P	DOE/NNSA/LLN	IBM	USA	2012	Research	393216	0	4,29	5,03	0,85	1,97	2,18	MPP	Power BQC 1	PowerPC	1600	16,00	Custom
10		Cray XC30, Inte	Government	Cray Inc	USA	2014	Government	225984	0	3,14	4,88	0,64			MPP	Intel Xeon E5	Intel IvyBridge	2700	12,00	Custom
11	HPC2	iDataPlex DX36	Exploration & P	IBM	Italy	2014	Industry	62640	36540	3,00	4,01	0,75	1,07	2,81	Cluster	Intel Xeon E5	Intel IvyBridge	2800	10,00	Infiniband
12	SuperMUC	iDataPlex DX36	Leibniz Rechenz	IBM	Germany	2012	Academic	147456	0	2,90	3,19	0,91	3,42	0,85	Cluster	Xeon E5-2680	Intel SandyB	2700	8,00	Infiniband
13	TSUBAME 2	Cluster Platform	GSIC Center, To	NEC/HP	Japan	2013	Academic	76032	59136	2,79	5,74	0,49	1,40	1,99	Cluster	Xeon X5670 6	Intel Nehalem	2930	6,00	Infiniband
14	Tianhe-1A	NUDT YH MPP	National Superc	NUDT	China	2010	Research	186368	100352	2,57	4,70	0,55	4,04	0,64	MPP	Xeon X5670 6	Intel Nehalem	2930	6,00	Propriet.
15	cascade	Atipa Visione I	DOE/SC/Pacific	Atipa Te	USA	2013	Research	194616	171720	2,54	3,39	0,75	1,38	1,83	Cluster	Xeon E5-2670	Intel SandyB	2600	8,00	Infiniband
16	Pangea	SGI ICE X, Xeo	Total Exploratio	SGI	France	2013	Industry	110400	0	2,10	2,30	0,91	2,12	0,99	Cluster	Xeon E5-2670	Intel SandyB	2600	8,00	Infiniband
17	Fermi	BlueGene/Q, P	CINECA	IBM	Italy	2012	Academic	163840	0	1,79	2,10	0,85	0,82	2,18	MPP	Power BQC 1	PowerPC	1600	16,00	Custom
18	Edison	Cray XC30, Inte	DOE/SC/LBNL/N	Cray Inc	USA	2014	Research	133824	0	1,65	2,57	0,64			MPP	Intel Xeon E5	Intel IvyBridge	2400	12,00	Custom
19		Cray XC30, Inte	ECMWF	Cray Inc	UK	2014	Research	83160	0	1,55	1,80	0,86			MPP	Intel Xeon E5	Intel IvyBridge	2700	12,00	Custom
20		Cray XC30, Inte	ECMWF	Cray Inc	UK	2014	Research	83160	0	1,55	1,80	0,86			MPP	Intel Xeon E5	Intel IvyBridge	2700	12,00	Custom
21	Pleiades	SGI ICE X, Inte	NASA/Ames Re	SGI	USA	2011	Research	96192	0	1,54	2,11	0,73	2,02	0,76	Cluster	Intel Xeon E5	Intel IvyBridge	2800	10,00	Infiniband
22		Power 775, PO	IBM Developmer	IBM	USA	2013	Vendor	63360	0	1,52	1,94	0,78	3,58	0,42	MPP	POWER7 8C	Power	3836	8,00	Custom
23	Blue Joule	BlueGene/Q, P	Science and Tec	IBM	UK	2012	Research	131072	0	1,43	1,68	0,85	0,65	2,18	MPP	Power BQC 1	PowerPC	1600	16,00	Custom
24	Spirit	SGI ICE X, Xeo	Air Force Resea	SGI	USA	2012	Government	73584	0	1,42	1,53	0,92	1,61	0,88	Cluster	Xeon E5-2670	Intel SandyB	2600	8,00	Infiniband
25	ARCHER	Cray XC30, Inte	EPSRC/Universi	Cray Inc	UK	2013	Research	76192	0	1,37	1,65	0,83			MPP	Intel Xeon E5	Intel IvyBridge	2700	12,00	Custom
26	Curie thin no	Bullx B510, Xe	CEA/TGCC-GEN	Bull SA	France	2012	Research	77184	0	1,36	1,67	0,82	2,25	0,60	Cluster	Xeon E5-2680	Intel SandyB	2700	8,00	Infiniband
27		iDataPlex DX36	Max-Planck-Ges	IBM	Germany	2013	Research	65320	0	1,28	1,46	0,88	1,25	1,02	Cluster	Intel Xeon E5	Intel IvyBridge	2800	10,00	Infiniband
28	Nebulae	Dawning TC360	National Superc	Dawning	China	2010	Research	120640	64960	1,27	2,98	0,43	2,58	0,49	Cluster	Xeon X5650 6	Intel Nehalem	2660	6,00	Infiniband
29	Yellowstone	iDataPlex DX36	NCAR (National	IBM	USA	2012	Research	72288	0	1,26	1,50	0,84	1,44	0,88	Cluster	Xeon E5-2670	Intel SandyB	2600	8,00	Infiniband
30	Helios	Bullx B510, Xe	International Fus	Bull SA	Japan	2011	Academic	70560	0	1,24	1,52	0,81	2,20	0,56	Cluster	Xeon E5-2680	Intel SandyB	2700	8,00	Infiniband
31	Garnet	Cray XE6, Opte	ERDC DSRC	Cray Inc	USA	2010	Research	150528	0	1,17	1,51	0,78			MPP	Opteron 16C	AMD x86_64	2500	16,00	Cray
32	Cielo	Cray XE6, Opte	DOE/NNSA/LAN	Cray Inc	USA	2011	Research	142272	0	1,11	1,37	0,81	3,98	0,28	MPP	Opteron 6136	AMD x86_64	2400	8,00	Cray
33	DiRAC	BlueGene/Q, P	University of Edi	IBM	UK	2012	Academic	98304	0	1,07	1,26	0,85	0,49	2,18	MPP	Power BQC 1	PowerPC	1600	16,00	Custom
34	Hopper	Cray XE6, Opte	DOE/SC/LBNL/N	Cray Inc	USA	2010	Research	153408	0	1,05	1,29	0,82	2,91	0,36	MPP	Opteron 6172	AMD x86_64	2100	12,00	Cray
35	Tera-100	Bull bullx super	Commissariat a	Bull SA	France	2010	Research	138368	0	1,05	1,25	0,84	4,59	0,23	Cluster	Xeon X7560 8	Intel Nehalem	2260	8,00	Infiniband
36	Oakleaf-FX	PRIMEHPC FX	Information Tech	Fujitsu	Japan	2012	Academic	76800	0	1,04	1,14	0,92	1,18	0,89	Cluster	SPARC64 IXf	Sparc	1848	16,00	Custom
37	QUARTETT(	HA8000-tc HT2	Research Institu	Hitachi/I	Japan	2013	Academic	222072	157128	1,02	1,50	0,68			Cluster	Xeon E5-2680	Intel SandyB	2700	8,00	Infiniband

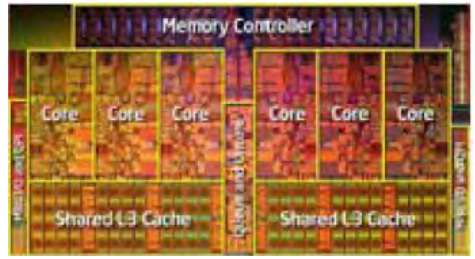
QUELQUES CONSIDERATIONS FINALES SUR LES « SOUS-SYSTÈMES »

Unités de calcul
Interconnect
I/O & stockage

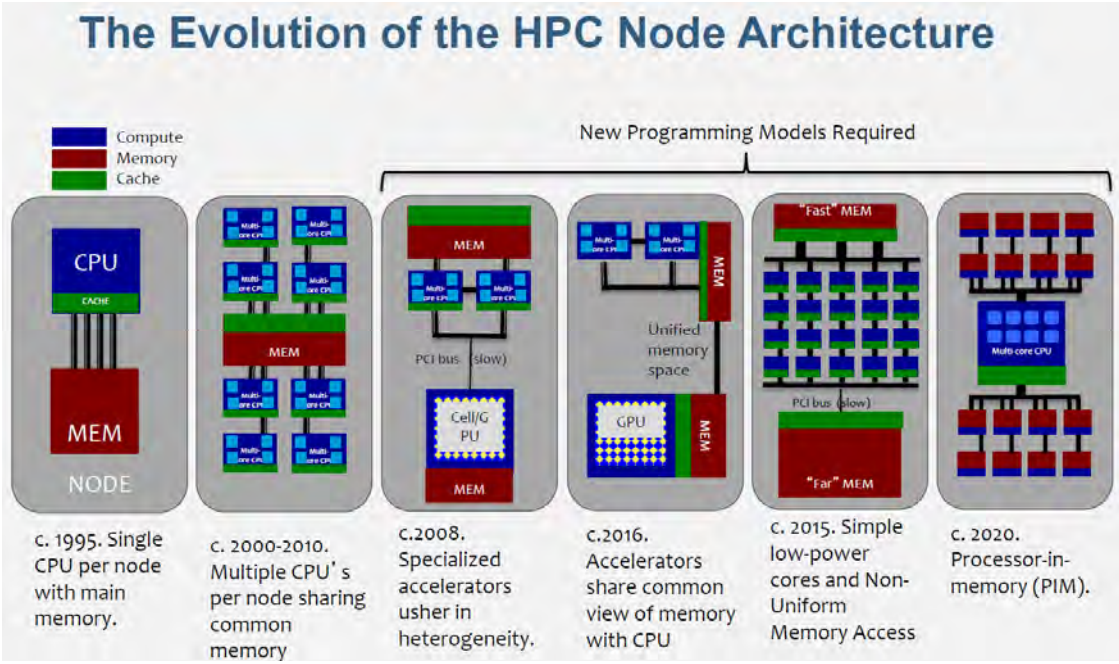


COMPUTE NODE

- Homogène ou hétérogène (hybride)
- Complexité de gestion mémoire
- Retour de la vengeance du vectoriel
  Unités / instructions spécialisées p.ex. SSE, AVX…
  Stream computing, SIMD sur GPU



Multicore ~16
Manycore > 32?



INTERCONNECT: TECHNOLOGIES, TOPOLOGIES

- Tendance: NIC (contrôleur réseau) sur processeur
- Tenir compte de la topologie dans les applications, dans le scheduling (placement)?
- Intérêt de minimiser les communications (data locality, communication avoiding algorithms…)

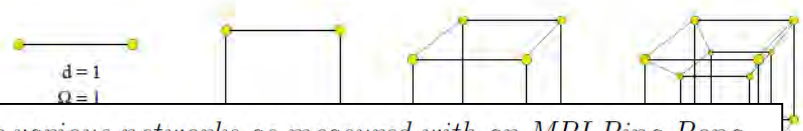
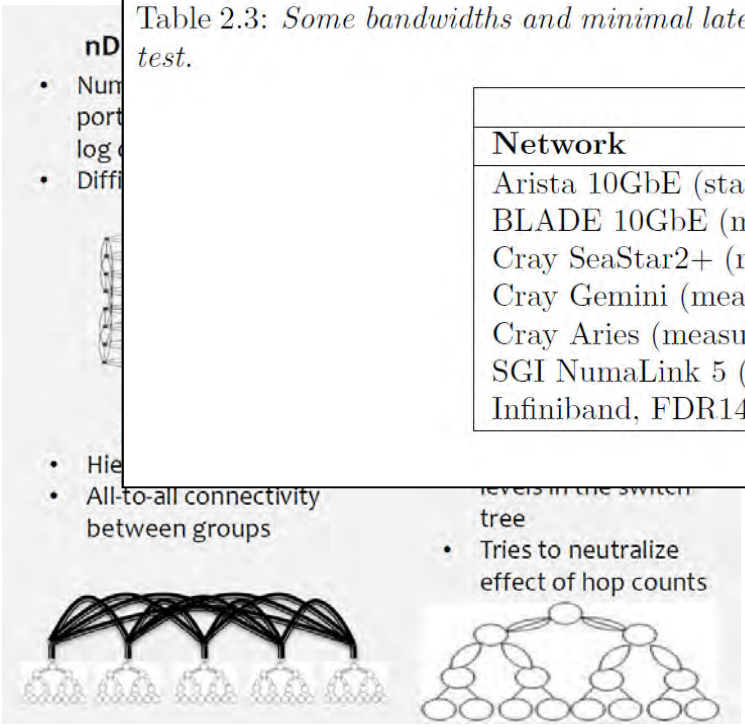
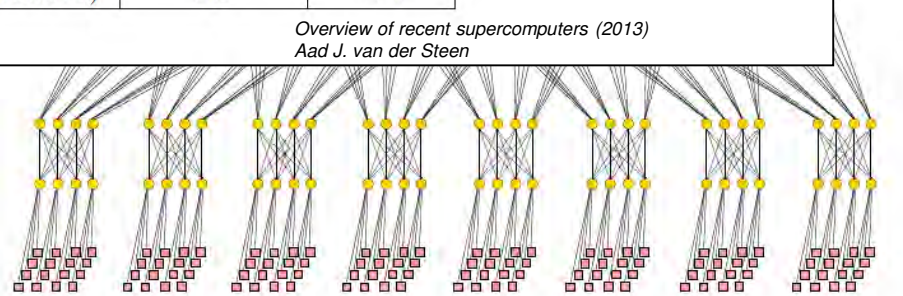


Table 2.3: *Some bandwidths and minimal latencies for various networks as measured with an MPI Ping-Pong test.*

	Bandwidth	Latency
Network	GB/s	μs
Arista 10GbE (stated)	1.2	4.0
BLADE 10GbE (measured)	1.0	4.0
Cray SeaStar2+ (measured)	6.0	4.5
Cray Gemini (measured)	6.1	1.0
Cray Aries (measured)	9.5	1.2
SGI NumaLink 5 (measured)	5.9	0.4
Infiniband, FDR14 (measured)	5.8	< 1

Overview of recent supercomputers (2013)
Aad J. van der Steen





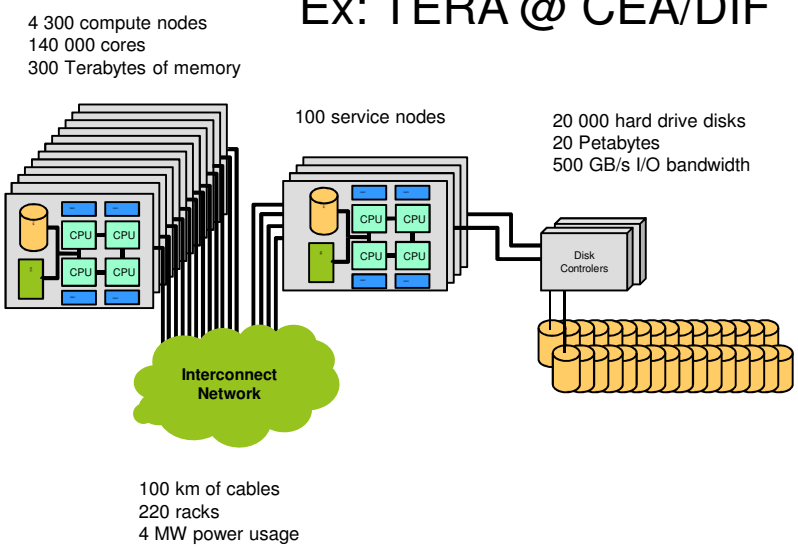
(b) A 128–way fat tree.

Figure 2.7: *Some often used networks for DM machine types.*

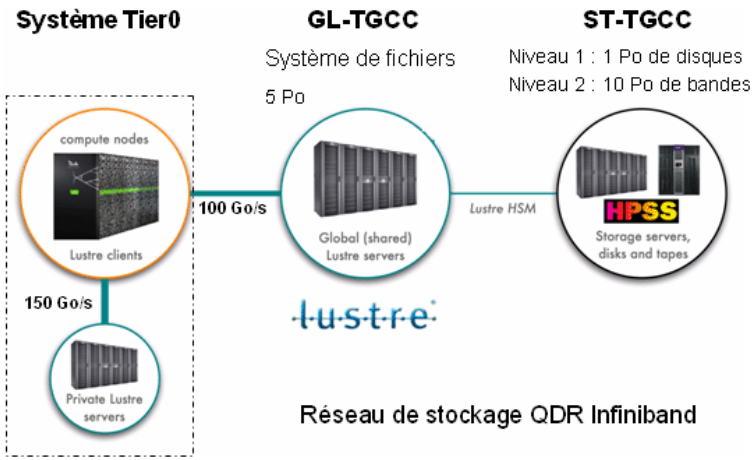
I/O ET STOCKAGE
EXASCALE = EXAFLOPS+EXABYTES

- Le HPC produit de plus en plus :
  - De données volumineuses plutôt ‘structurées’
  - De métadonnées associées à ces données et des données moins structurées et diversifiées (champ du Big ‘Data’)
- Inversement le ‘Big Data’ au sens large peut bénéficier d’approches HPC dans certains cas
  - ☞ Données qui pré-existent au traitement
- Importance croissante de l’équilibre compute-I/O-stockage
- L’interconnect joue un rôle dans la circulation multi-directionnelle des I/O

Ex: TERA @ CEA/DIF



Ex: CURIE @ TGCC



- Dedicated InfiniBand QDR storage network
- Lustre FS (GL-TGCC) shared between compute and post-processing resources
- Automatic migration from GL-TGCC to the hierarchical storage IBM/HPSS (ST-TGCC)

Organisations déjà ‘data centric’

CONCLUSION

- Ceci était une modeste introduction, faut-il conclure maintenant…???
- Le Petascale est bien là (2010-2014: il existe des technologies et solutions « compactes » désormais)
  Mais il n'est pas encore démocratisé en terme d'usage et de productivité
- L'incertitude économico/technologique concernant:
  la faisabilité,
  la viabilité, l'utilisabilité,
  la ou les formes,
  et la date d'arrivée de l'exascale
  oblige à se poser des questions de fond sur la manière d'aborder la programmation, la vie des applications, la formation et le développement des compétences HPC
- Se préparer dès maintenant est possible, quelques grands principes sont discernables, et cela peut bénéficier aussi aux applications actuelles sur les machines existantes…