Supercomputing

Anthony Scemama <scemama@irsamc.ups-tlse.fr>

Labratoire de Chimie et Physique Quantiques
IRSAMC (Toulouse)

Université Paul Sabatier

TOULOUSE III

LCPQ

cnrs

dépasser les frontières

Computers

Today (order of magnitude):

- 1 socket (x86 CPU @ 2.2-3.3 GHz, quad-core, hyperthreading)
- ~4-16 GB RAM
- 2TB HDD (7200 rpm) / SSD (~128GB)
- Graphics card : ATI Radeon, Nvidia GeForce
- Gigabit ethernet
- USB, Webcam, Sound card, etc

~600 euros

Computer designed for computation

Today (order of magnitude):

- 2 sockets (x86 CPU @ 2.8 GHz, 12 cores/socket, hyperthreading)
- 64-128 GB RAM
- Multiple 15k rpm or write-intensive SSD HDDs (RAID0)
- Gigabit ethernet
- Eventually an Accelerator (Nvidia Tesla/Pascal or Intel Xeon Phi)

~6k euros

Cluster

- Many computers designed for computation
- Compact (1-2U in rack) / machine
- Network switches
- Login server
- Batch queuing system (SLURM / PBS / SGE / LFS)
- Cheap Cooling system
- Requires a lot of electrical power (~10kW/rack)
- Eventually a Low-latency / High-bandwidth network (Infiniband or 10Gb ethernet)

>50k euros

Supercomputers

- Many computers designed for computation
- Very compact (<1U in rack) / machine
- Low-latency / High-bandwidth network (Infiniband or 10Gb ethernet)
- Network switches
- Parallel filesystem for scratch space (Lustre / BeeGFS / GPFS)
- Multiple login servers
- Batch queuing system (SLURM / PBS / SGE / LFS)
- Highly efficient cooling system (water)
- Requires a lot of electrical power (>100kW)

> 1M euros

Definitions

Top500

Rank of the 500 fastest supercomputers

Flop

Floating point operation

Flops/s

Number of Floating point operations per second

RPeak

Peak performance, max possible number of Flops/s

RMax

Real performance on the Linpack benchmark (dense eigenproblem)

SP

Single precision (32-bit floats)

DP

Double precision (64-bit floats)

FMA

Fused multiply-add (*a*x+b* in 1 instruction)

Number of hours

/month : 730, /year : 8760

Units

Kilo (K), Mega (M), Giga (G), Tera (T), Peta (P), Exa (E), ...

Example

RPeak of Xeon E5-2680 v4 : 14 cores, 2.4~GHz, 8 FMA (DP) per cycle.

14 x (4x2) x 2.4E+9 = 166.4E+9 flops/s = 268.8 Gflops/s (DP)

TOP 10 Sites for June 1996

For more information about the sites and systems in the list, click on the links or view the complete list.

Rank	Site	System	Cores	Rmax (GFlop/s)	Rpeak (GFlop/s)	Power (kW)
1	University of Tokyo Japan	SR2201/1024 Hitachi	1,024	220.4	307.2	
2	National Aerospace Laboratory of Japan Japan	Numerical Wind Tunnel Fujitsu	166	170.0	279.6	
3	Sandia National Laboratories United States	XP/S140 Intel	3,680	143.4	184.0	
4	DOE/SC/Oak Ridge National Laboratory United States	XP/S-MP 150 Intel	3,072	127.1	154.0	
5	Japan Atomic Energy Research Institute Japan	XP/S-MP 125 Intel	2,502	103.5	125.1	

TOP 10 Sites for November 2016

For more information about the sites and systems in the list, click on the links or view the complete list.

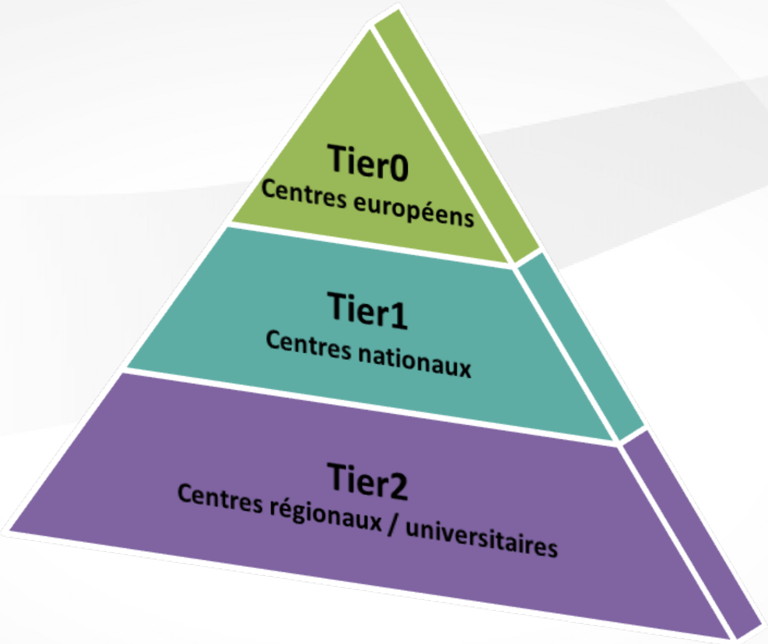
Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
5	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	14,014.7	27,880.7	3,939

Curie thin nodes (TGCC) : ranked 9th in 2012, 74th today

77184 cores, 1.7 PFlops/s, 2.132 MW

Mare Nostrum (Barcelona, rank 129): 48896 cores, 1.017 PFlops/s, 1.016 MW

Resources in France



- Europe (Tier0) : > 10M CPU hours
- National (Tier1) : ~ 1-10M CPU hours
- Regional (Tier2) : 100k-1M CPU hours

Europe (Tier0)

- PRACE : Partnership for Advanced Computing in Europe (25 countries)

System	Architecture	Site (Country)	Core Hours	Node Hours
Curie	Bull Bullx cluster	GENCI@CEA (FR)	**113 million**	**7 million**
Hazel Hen	Cray XC40 System	GCS@HLRS (DE)	**57 million**	**2.4 million**
Juqueen	IBM BlueGene/Q system	GCS@JSC (DE)	**350 million**	**4.6 million**
Marconi	Lenovo System	CINECA (IT)	**630 million (KNL)** **35 million (Broadwell)**	**9.2 million** **1 million**
MareNostrum	IBM System X iDataplex	BSC (ES)	**316 million**	**7 million**
Piz Daint	Cray XC30 System	CSCS (CH)	**476 million**	**7 million**
SuperMUC	IBM System X iDataplex/ Lenovo NextScale	GCS@LRZ (DE)	**30 million (SuperMUC 1)** **14 million (SuperMUC 2)**	**1.9 million** **0.5 million**

- EGI : European Grid Infrastructure

  730 000 cores for high throughput computing (1.6 million computing jobs per day)

  6 600 cores for Cloud computing

  Online storage / data transfer

- Both PRACE and EGI propose training

National (Tier 1)

GENCI : Grand Equipement National de Calcul Intensif

- IDRIS : Institut du développement et des ressources en informatique scientifique (Orsay)
    - Ada : 10624 x86 cores, 230 TFlops/s
    - Turing : IBM BlueGene/Q, 65536 cores, 836 TFlops/s
- CINES : Occigen (Montpellier)
    - Occigen : 50544 x86 cores, BullX, 2.1 PFlops/s
- TGCC : Curie (Bruyères-le-Chatel)
    - Curie Thin+Fat nodes : 92160 x86 cores, BullX, 2 PFlops/s

France Grilles :

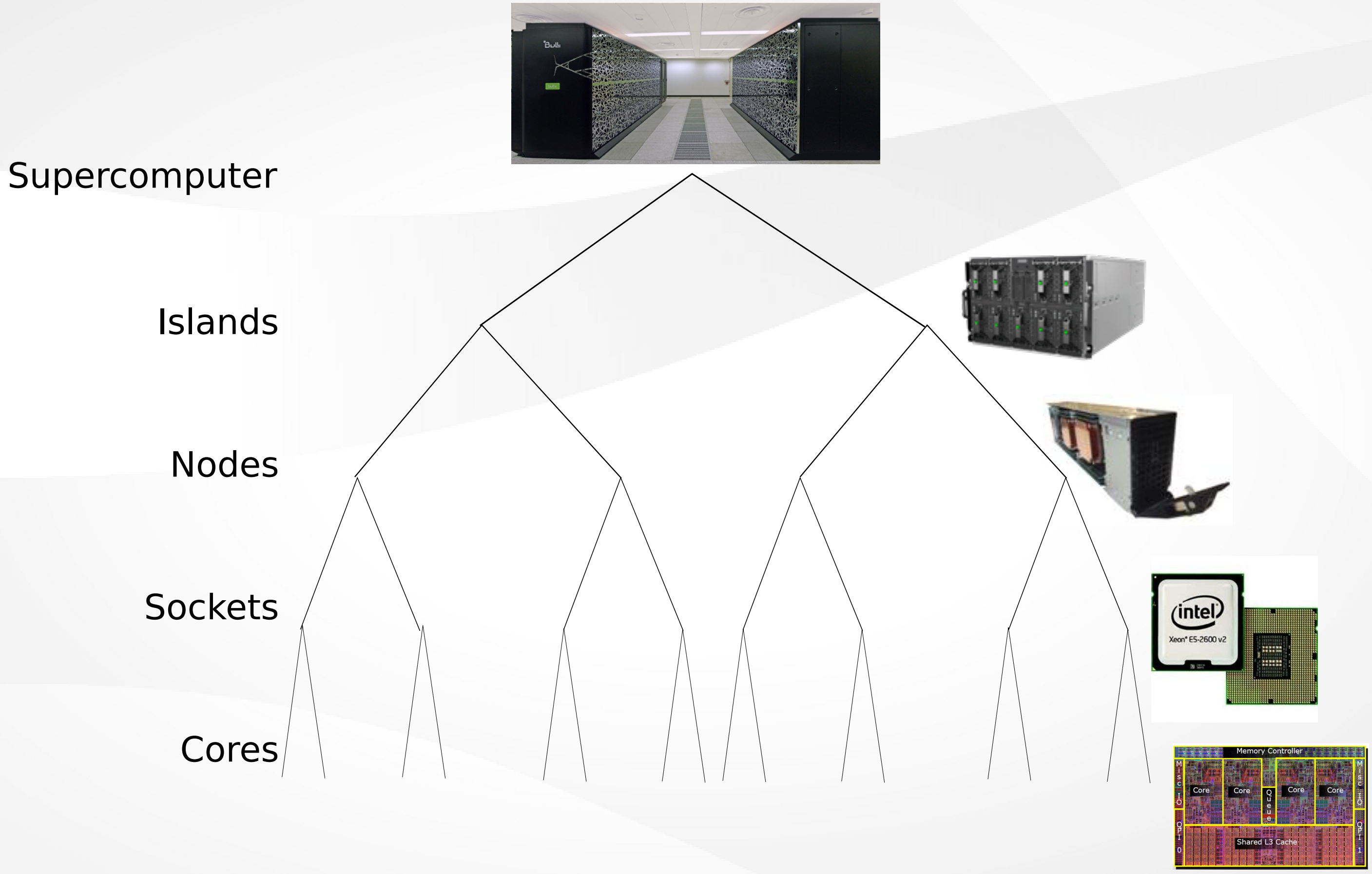
- 18 sites
- ~ 32000 cores (in 2015)

Regional (Tier 2)

CALMIP : Calcul en Midi-Pyrénées

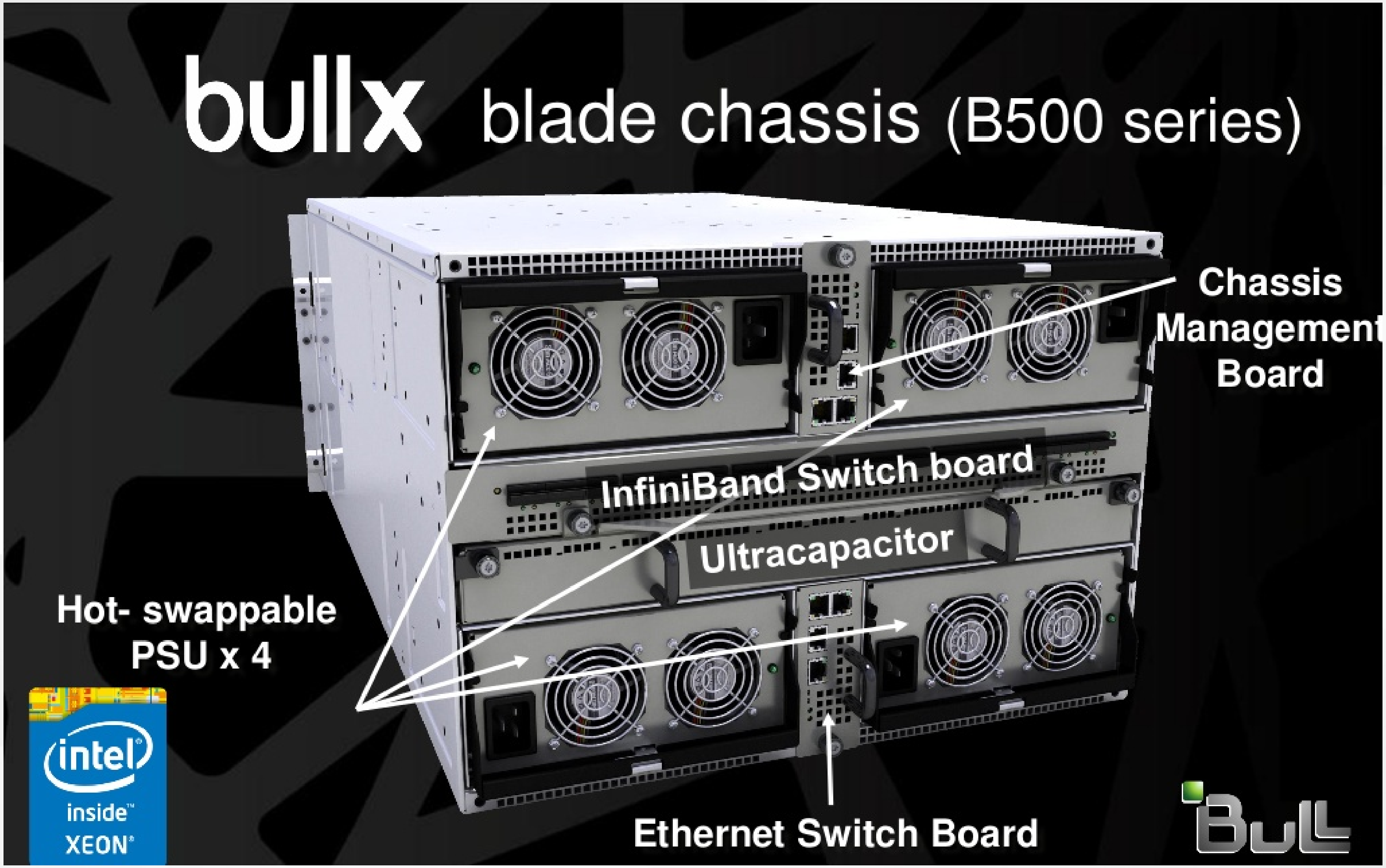
- Eos : 12240 cores, 255 TFlops/s (274 Tflops/s Peak)
- Mesca : 128 cores SMP, 3TB RAM

Architecture

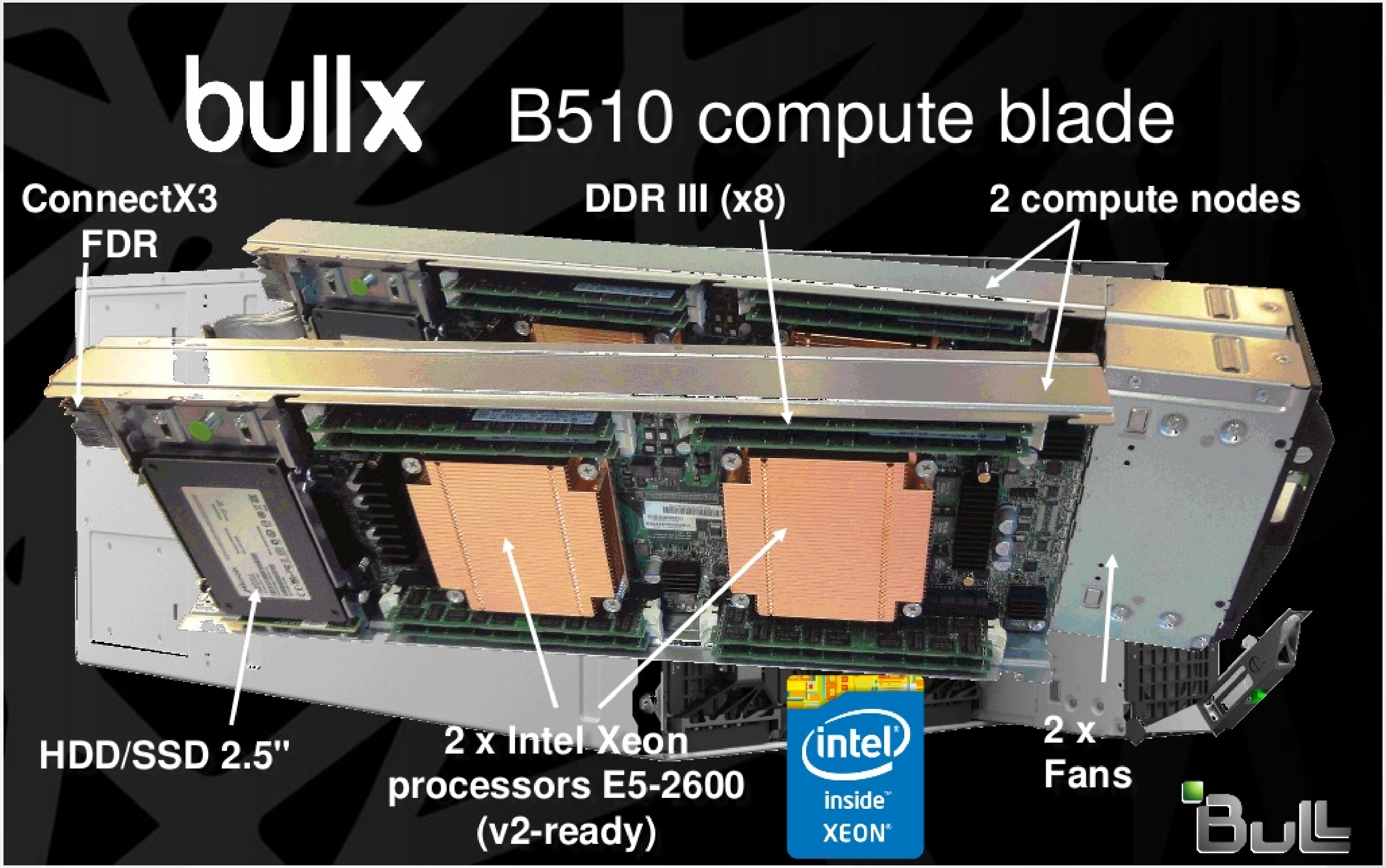


Chassis


bullx blade chassis (B500 series)
LCD unit
Fans
9 blades = 18 nodes
7U chassis
intel inside XEON
BULL

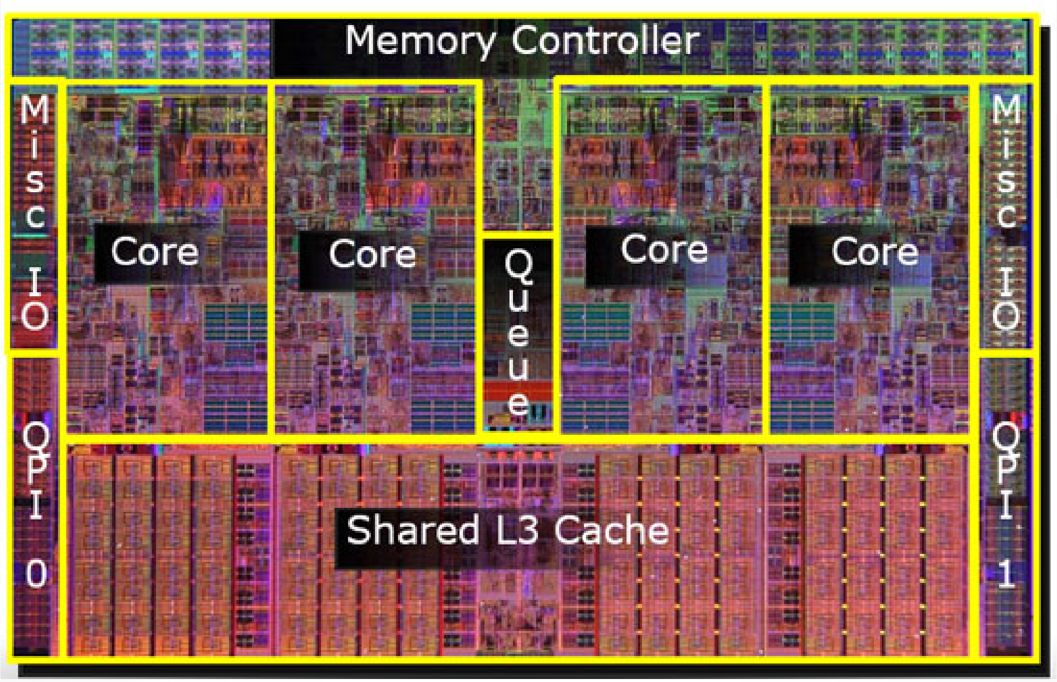

bullx blade chassis (B500 series)
Chassis
Management
Board
InfiniBand Switch board
Ultracapacitor
Hot- swappable
PSU x 4
intel
inside
XEON
Ethernet Switch Board
BULL

Node


bullx B510 compute blade
ConnectX3 FDR
DDR III (x8)
2 compute nodes
HDD/SSD 2.5"
2 x Intel Xeon processors E5-2600 (v2-ready)
intel inside XEON
2 x Fans
BULL

Socket

Core



Memory Controller
Misc IO
Core
Core
Queue
Core
Core
Misc IO
QPI 0
Shared L3 Cache
QPI 1

Why such an architecture?

Moore's law

- "The number of transistors doubles every two years".
- Often interpreted as "the computational power doubles every two years".
- Exponential law -> will fail.
- 14 nm semiconductors in 2017. The atomic limit is approaching.

Main Problem : energy

Dissipation : $D = F \times V^2 \times C$

- F : Frequency (~3 GHz)
- V : tension
- C : constant related to the size of semiconductors (nm)
- For the processor to be stable, the tension needs to be linear with the frequency: so $D = O(F^3)$ -> The frequency has to be kept around 3GHz.

- Solution 1 : double the number of Flops/s = double the number of processors.
- Solution 2 : Use accelerators (GPUs, Xeon Phi)

This requires to re-think programming -> Parallel programming is *unavoidable*.

Today's main problems

1. Energy, physical limits
2. Interconnect technology
3. Increase of computational power is faster than memory capacity : reduction of memory per CPU core
4. Latencies can't be reduced much more : moving data becomes very expensive
5. File systems are **extremely** slow
6. Supercomputers are well tuned for benchmarks (dense linear algebra), but not that much for scientific applications
7. Fault-tolerance

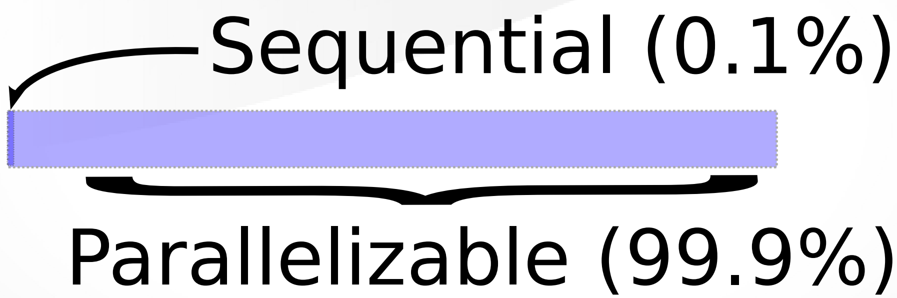
Practical solution: use less efficient/intelligent CPUs, but in a much larger number:

- Intel Atom CPUs (~2GHz, System On Chip, HP Moonshot) : >1500 x86 cores in 3U
- Intel Xeon Phi coprocessor (61 cores, 1GHz)
- Nvidia/ATI GPUs (~3584 cores, ~1GHz), with ARM host CPU
- Intel Knights Landing (72 cores, 288 threads, ~1.5GHz CPU socket)
- IBM OpenPower CPU accelerated with Nvidia GPUs

Amdahl's law

If P is the proportion of a program that can be made parallel, the maximum speedup that can be achieved is:

$$\frac{1}{(1-P)+P/n}$$



- Max speedup here : 1000x
- Scaling needs **all** the program to be fully parallelized (difficult)