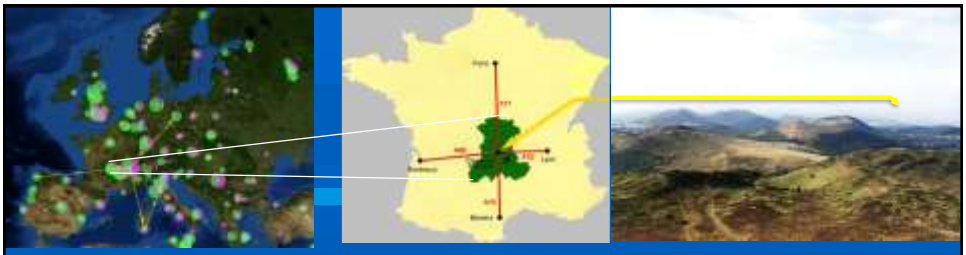


Introduction to High Performance Computing

& Distributed Computing

David Hill - UMR CNRS 6158
Blaise Pascal University

10 years ago…
A revolution in the HPC Wordl
Vidéo – SC 2008 – NVIDIA Tesla Supercomputer

_IDM & SIM 1 - Tesla Personal Supercomputer

2

What you will see in the HPC course

- Enough background in HPC to tackle with intensive computing in many domains
  - Master Students coming from Computer Science, Applied Mathematics (and some from Biology !)
  - ISIMA students (regular & international track.)
- Content (mainly lectures + few labs)

For more programming oriented stuff see the ISIMA lectures:

- MPI, OpenMP, pthreads, Grid Computing, Hybrid computing…

How will you be assessed ?

- Class & Lab participation
- A small quiz and/or a short academic
  - Literature review :
    - 3 pages (computer scientists)
    - 6 to 10 pages (if biologists)
- Final exam
    - With technical & code questions for computer scientists.

Part I
A bit of History dealing with supercomputers…

5

History of Super Computing

- The notion of parallel processing can be traced to a tablet dated around 100 BC.
  - Tablet has 3 calculating positions capable of operating simultaneously.
  - From this we can infer that:
    - They were aimed at
      - "speed" and/or
      - "reliability".

Motivating Factor

- The human brain consists of a large number of neural cells (more than a billion) that process information.
- Each cell works like a simple processor and only the massive interaction between all cells and their parallel processing makes the brain's abilities possible.
  - Individual neuron response speed is slow (ms)
  - Aggregated speed with which complex calculations carried out by (billions of) neurons demonstrate feasibility of parallel processing.

Why Parallel Processing?

- Computation requirements are ever increasing:
  - simulations, scientific prediction (earthquake), distributed databases, weather forecasting), Internet Search engines, e-commerce, Internet service applications, Data Center applications, Finance (investment risk analysis), Oil Exploration, Mining, etc.
- Silicon based (sequential) architectures reaching their limits in processing capabilities (clock speed) as they are constrained by:
  - the speed of light vs thermodynamics
  - Quantum computing ?

Speed for computing FLOPS : an old performance unit…

- FLOPS : stands for floating-point operations per second
- It is a measure of computer performance, useful in fields of scientific calculations.
- 1 GF = 1 billion flops
- 1 EF = 1 billon billion flops

$$\text{FLOPS} = \text{sockets} \times \frac{\text{cores}}{\text{socket}} \times \text{clock} \times \frac{\text{FLOPs}}{\text{cycle}}$$

Name	FLOPS
kiloFLOPS	10^3
megaFLOPS	10^6
gigaFLOPS	10^9
teraFLOPS	10^{12}
petaFLOPS	10^{15}
exaFLOPS	10^{18}
zettaFLOPS	10^{21}
yottaFLOPS	10^{24}

- PC CPUs can produce 4 FLOPs per clock cycle (16 to 32 for server CPUs). An octo-core PC at 2.5 GHz has a theoretical performance of 80 billion FLOPS = 80 GFLOPS.

9

A bit of more recent History…

- Supercomputing is about pushing out the leading edge of computer speed and performance.
- Like in Formula 1 :
  "what you learn is important…
  …for regular cars"

http://www.cio.com.au/article/132504/brief_history_supercomputers/

10

Supercomputing applications…

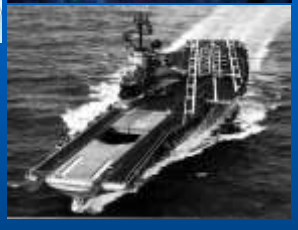
Supercomputers have been used for :

- **Weather forecasting, Satellite image analysis**
- **Fluid dynamics (such as modeling air flow around airplanes or automobiles)**
- **Simulations of particle physics, astrophysics,… with vast numbers of variables and equations that have to be solved or integrated numerically through an almost incomprehensible number of steps, or probabilistically by Monte Carlo sampling**
- **For many stochastic applications (more than 50% supercomputing usage)**

11

A piece of History : Earth Simulator 1

Ranked #1 for 5 contest [2002 2005]



NEC produced the Earth Simulator in 2002. It used vector technology, and went from 32 Tflops to approx 40 Tflops
Between 2002 and 2005.
Now at 139 PF with fast interconnect

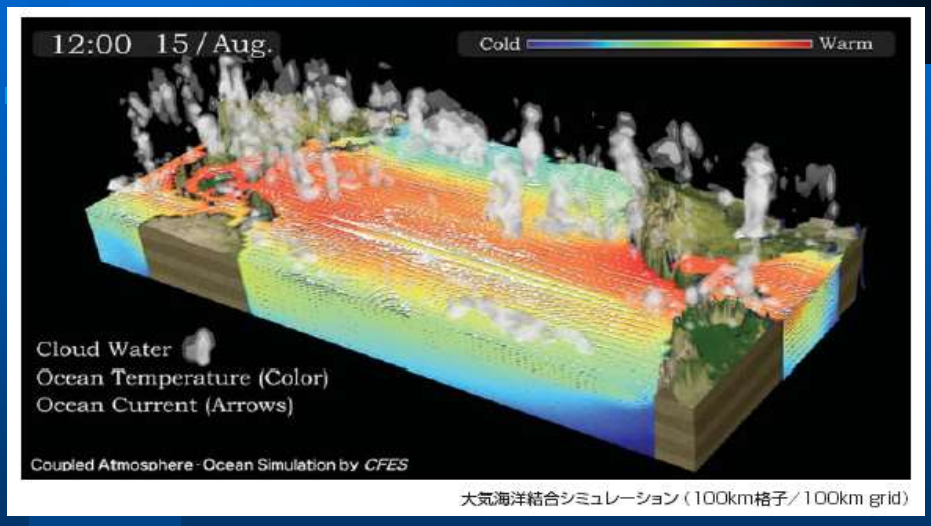
A dedicated building

- 5 years later IBM's Blue Gene/L achieved about 200 TFLOPS.
- It consumes 15 times less power per computation
- 50 times smaller than Earth Simulator 1

ES-2 (139 TF) ISIMA inside…

14

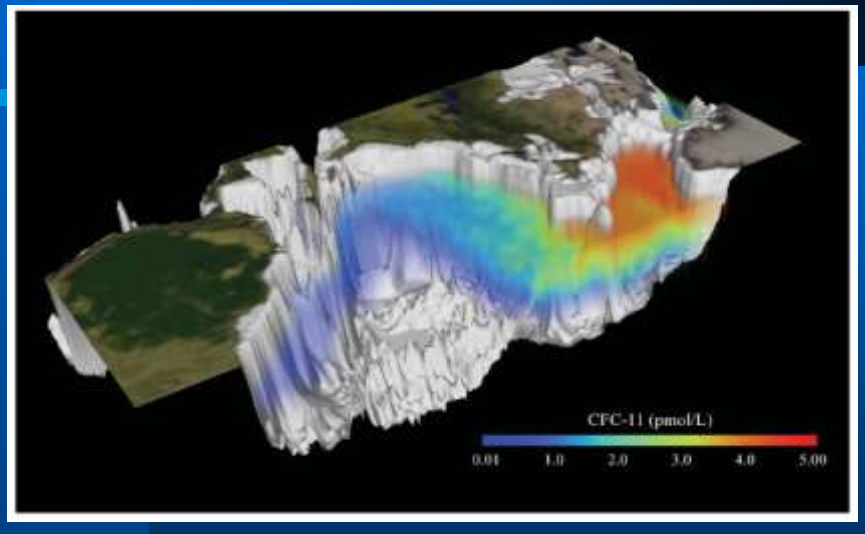
Ex. App 1 : Ocean - Atmosphere Interactions



The ocean warms the atmosphere, this generates winds which drive the ocean currents.

15

Ex. App. 2 : Deep current simulation in the Atlantic

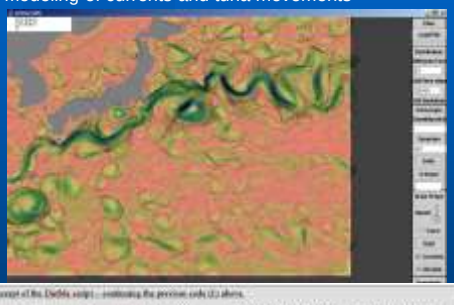


Freon gas is transported from the north pole to the south by a deep littoral current

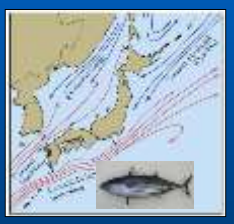
16

Ex. App 3 : Simulation of tuna moves around Japan
Stochastic simulations Data from the Earth-Simulator Center

Modeling of currents and tuna movements



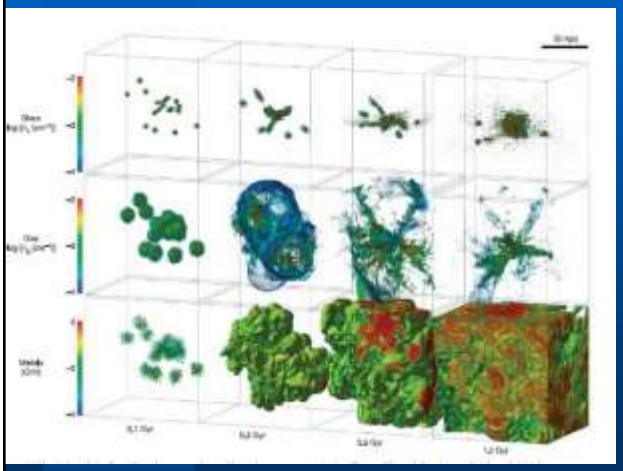
Collaboration with Dr. Komatsu
Ocean Research Institute
Univ. De Tokyo



A single tuna can be sold : 60 000€ !
A particular grease « toro »

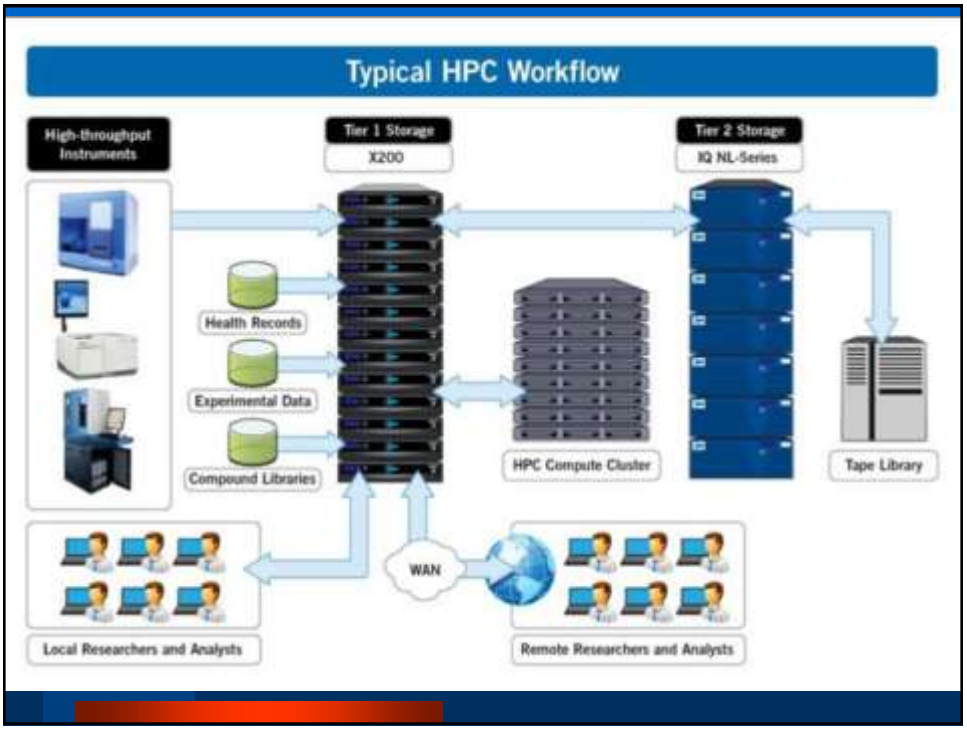
Preserve species – optimized fishing – fluctuations of hot current : *Kuroshio*

Ex. App. 4 : Simulations of galaxies



Sept. 27th 2012,
Two astronomers have made one of the largest ever conducted simulations in astrophysics to model the growth of galaxies.
Masao Mori (LA Univ.) and Masayuki Umemura (Tsukuba Univ.) are able to simulate galaxy evolution since 300 million years after the Big Bang until today.
Their results show that galaxies may have evolved much faster than previously thought.

18



Top 500 ranking for Supercomputers…

- A benchmark and ranking for the world's fastest super computers

http://www.top500.org/

TOP 500

- Since 1993 a list computers ranked by their performance on the LINPACK Benchmark.

20

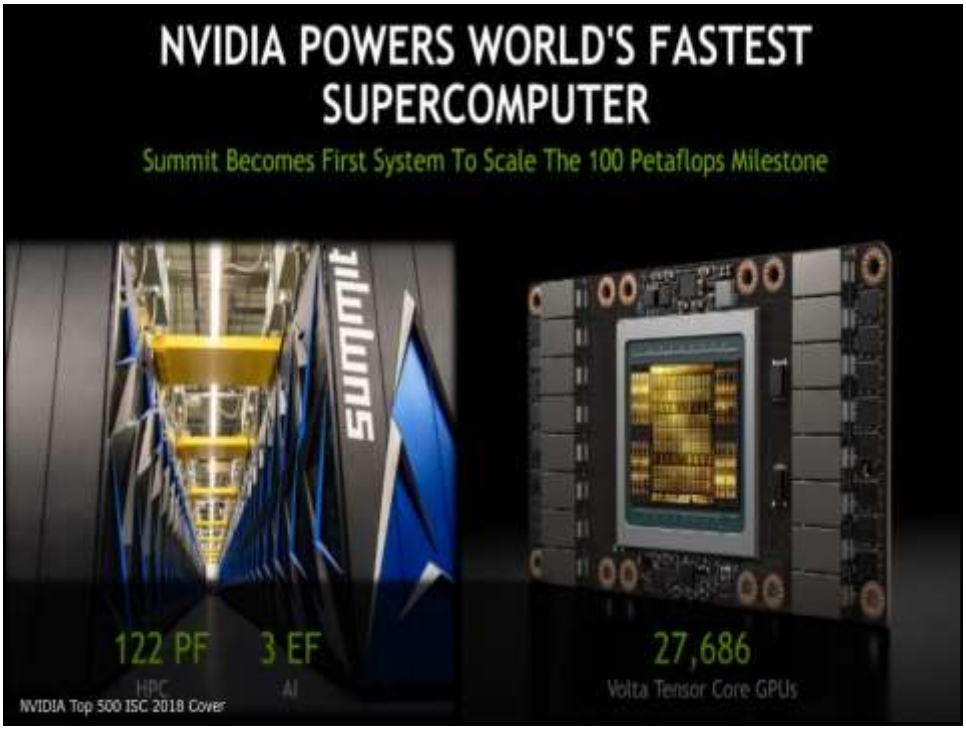
HPC performance evaluation

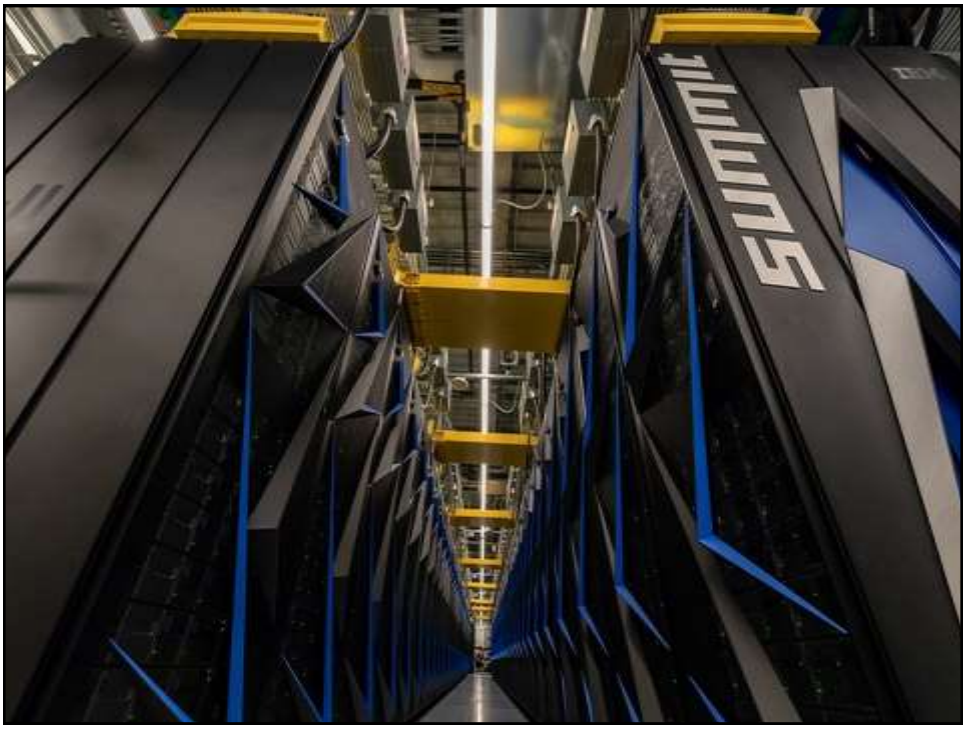
- In high-performance computing, Rmax and Rpeak are scores used to rank supercomputers based on their performance using the LINPACK Benchmark.
- the Rpeak score describes its theoretical peak performance
- A system's Rmax score describes its maximal achieved performance;

21

Top 500 – current list (June 2018)

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,282,544	122,300.0	187,659.3	8,806
2	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
3	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	1,572,480	71,610.0	119,193.6	
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	National Institute of Advanced Industrial Science and Technology (AIST) Japan	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR	391,680	19,880.0	32,576.6	1,649


NVIDIA POWERS WORLD'S FASTEST SUPERCOMPUTER
Summit Becomes First System To Scale The 100 Petaflops Milestone
summit
122 PF
HPC
3 EF
AI
27,686
Volta Tensor Core GPUs
NVIDIA Top 500 ISC 2018 Cover


summit

Previous ranking – Nov. 2017

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-2A** - 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	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	361,760	19,590.0	25,326.3	2,272
4	Japan Agency for Marine-Earth Science and Technology Japan	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz ExaScaler	19,860,000	19,135.8	28,192.0	1,350
5	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

Recent History…

CHINA'S TIANHE-2 SUPERCOMPUTER MAINTAINS TOP SPOT ON LIST OF WORLD'S TOP500 SUPERCOMPUTERS

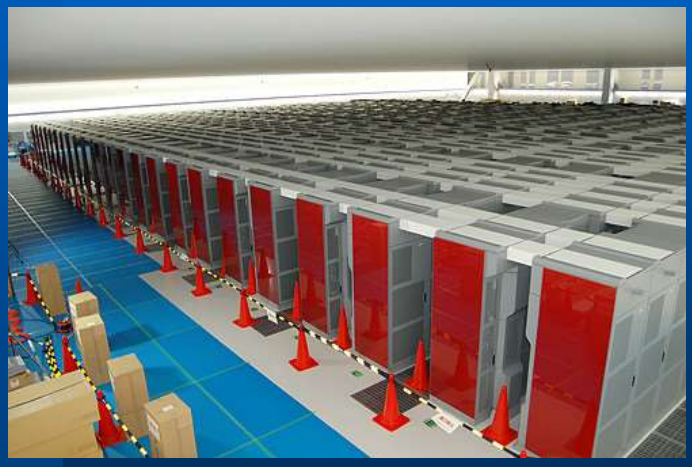
Feature Article

Posted 2 days, 18 hours ago

For the sixth consecutive time, Tianhe-2, a supercomputer developed by China's National University of Defense Technology, has retained its position as the world's No. 1 system, according to the 46th edition of the twice-yearly TOP500 list of the world's most powerful supercomputers.

The world's fastest computer is still China's Sunway TaihuLight, according to the latest TOP500 list released on Monday. **(4 times)**

2011 Lead for 2 benchmarks K - Super computer (Japan)



27

Some details…

- **10.51 Petaflops**

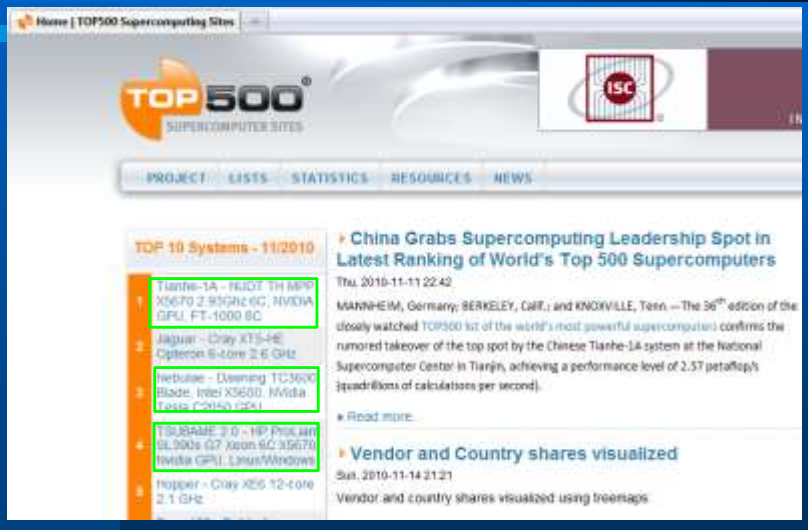
Meaning (10 510 000 of billions of floating points operations per second)

- **88128 processors by Fujitsu**
- **Riken Labs in Kobe – Japan**
- **Usage : climate research, extreme meteorological event prevention, medicine…).**

28

Tianhe-1A : 4.7 Petaflops - Nov. 2010

Introduction of Hybrid HPC (CPU/GPU)

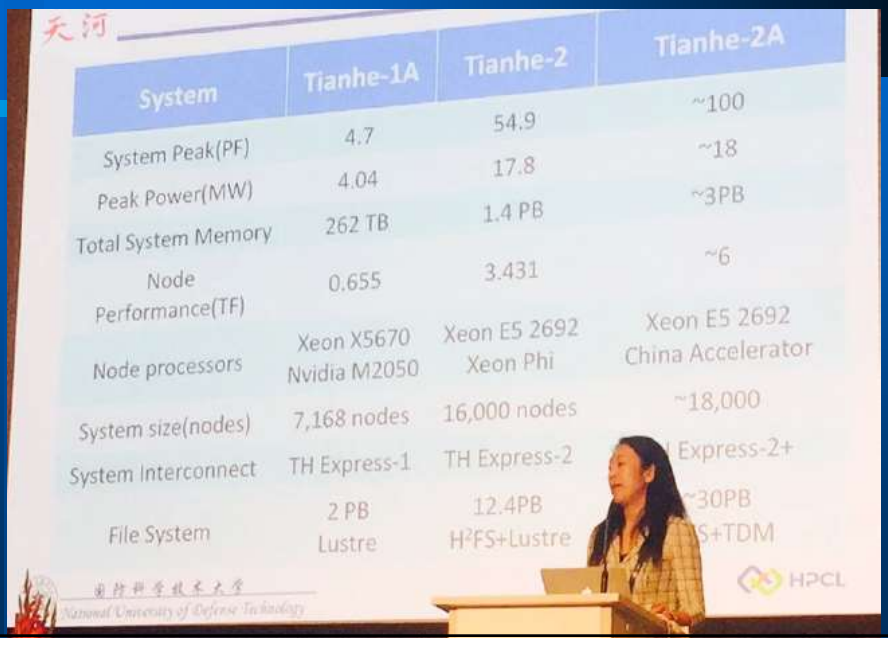


Some details…

- Linux Operating System
- 186368 processors
- 229376 GB or RAM (229 TBytes)
- Hybrid nodes
  - Intel EM64T Xeon X56xx (Westmere-EP) 2930 MHz (11.72 GFlops)
  - NVIDIA Fermi GPUs

30

Progess of Tianhe Systems



System	Tianhe-1A	Tianhe-2	Tianhe-2A
System Peak(PF)	4.7	54.9	~100
Peak Power(MW)	4.04	17.8	~18
Total System Memory	262 TB	1.4 PB	~3PB
Node Performance(TF)	0.655	3.431	~6
Node processors	Xeon X5670 Nvidia M2050	Xeon E5 2692 Xeon Phi	Xeon E5 2692 China Accelerator
System size(nodes)	7,168 nodes	16,000 nodes	~18,000
System Interconnect	TH Express-1	TH Express-2	Express-2+
File System	2 PB Lustre	12.4PB H²FS+Lustre	~30PB S+TDM

Part II
How do we
build supercomputers ?

32

The first supercomputer

- The first machine generally referred to as a supercomputer (though not officially designated as one), the IBM Naval Ordonance Research Calculator, was used at Columbia University from 1954 to 1963 to calculate missile trajectories.
- It predated microprocessors, had a clock speed of 1 microsecond and was able to perform about 15,000 operations per second (+, -, / , *).

33

Seymour Cray… The beginning of "SC"' architecture…

- The beginning of supercomputers is closely associated with Seymour Cray
- He designed the first officially designated "supercomputers" for Control Data in the late 1960s.
- His first design, the CDC 6600, had a pipelined scalar architecture and used the RISC instruction set that his team developed.
- In this architecture, a single CPU overlaps fetching, decoding and executing instructions to process one full instruction each clock cycle.
- Evolution of this machine included multi-processors

34

From multiprocessors to vector processors

- In 1972 Cray started his own company, Cray Research he abandoned the multiprocessor architecture in favour of vector processing (to unrolling "for" "do" loops)
- Using a CDC 6600, the European Centre for Medium-Range Weather Forecasts (ECMWF) produced a 10-day forecast in 12 days!
- Using one of Cray Research's first products, the Cray 1-A, the ECMWF was able to produce a 10-day forecast in five hours.

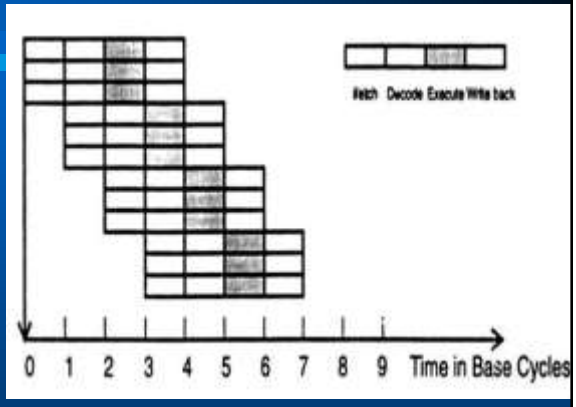
Scalar, Vector & superscalar

- "The simplest processors are scalar processors. Each instruction executed by a scalar processor typically manipulates one or two data items at a time.
- By contrast, each instruction executed by a vector processor operates simultaneously on many data items.
- An analogy is the difference between scalar and vector arithmetic.
- A superscalar processor is sort of a mixture of the two. Each instruction processes one data item, but there are multiple redundant functional units within each CPU thus multiple instructions can be processing separate data items concurrently."

Cray T3e Alpha board

Superscalar pipeline

- "early superscalar" CPUs had two ALUs and a single FPU,
- A modern design can include four ALUs, two FPUs, and two vector units.



- If the dispatcher is ineffective at keeping all the units fed with instructions, the performance of the system will suffer."

Wikipaedia

National Security & protectionism

- In their early history, the production and use supercomputers was carefully controlled, since they were used in critical nuclear weapons research.
- They were also a source of national pride, symbolic of technical leadership.
- Antidumping legislation was brought to bear against the importation of Japanese supercomputers in the US
- It was revoked in 1998

38

National Security

- The first Bush administration (1990) defined supercomputers as being able to perform more than 195 Millions of Theoretical Operations per Second (MTOPS).
- Anyway by 1997, ordinary microprocessors for PCs were capable of over 450 MTOPS.
- Technologists continued to increase the performances of massive parallel supercomputers.
- Peripheral speeds had increased so that I/O was no longer a bottleneck.
- High-speed communications made distributed and parallel designs possible.

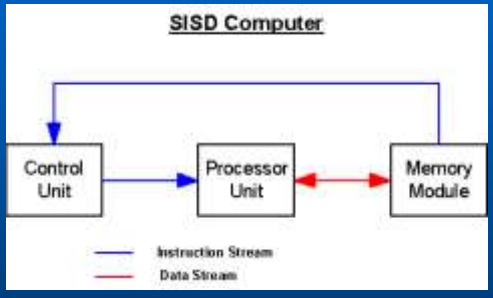
39

Part III
Parallel Architectures and processing elements

Taxonomy of Parallel Architectures

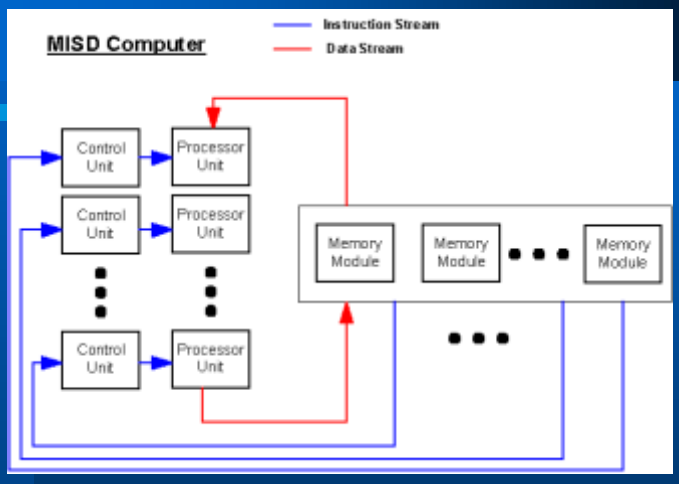
- Flynn proposed a classification of computer systems based on a number of instruction and data streams that can be processed simultaneously.
- They are:
  - SISD (Single Instruction and Single Data)
    - Conventional (old) computers
  - SIMD (Single Instruction and Multiple Data)
    - Data parallel, vector computing machines
  - MISD (Multiple Instruction and Single Data ??!)
    - Systolic arrays
  - MIMD (Multiple Instruction and Multiple Data)
    - General purpose machine

SISD : A Conventional 'old' Computer



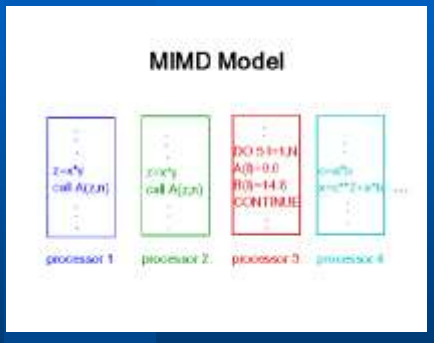
- The Speed is limited by the rate at which computers can transfer information internally.
- Ex: Old PCs and Workstations

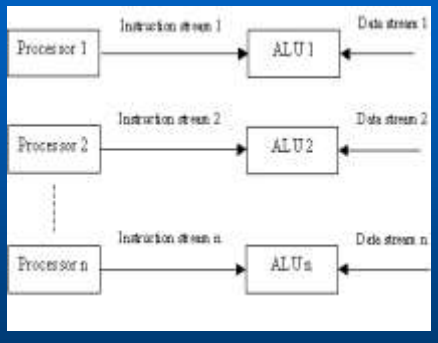
The MISD Architecture



- More of an intellectual exercise than a practical configuration.
- Few built in labs, but commercially not available

MIMD Architecture





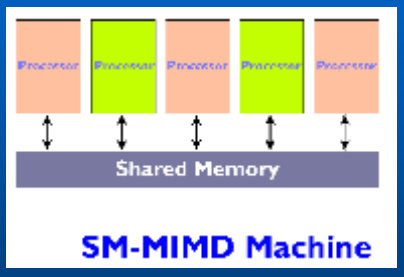
Unlike SISD, MISD, MIMD computer works asynchronously.

1- Shared memory (tightly coupled) MIMD e.g., Multicore

2 - Distributed memory (loosely coupled) MIMD

Shared Memory MIMD

Communication:
Source PE (Processing Element) writes data to a Shared Memory and the destination PE retrieves it



Conventional OSes of SISD can be adapted (up to a certain Scale)

Limitation : reliability & expandability.

A memory component or any processor failure affects the whole system.

Increase of processors leads to memory contention.

Silicon graphics supercomputers allows Manycore machines

Symmetric MultiProcessing (SMP) (1/3)

- Symmetric multiprocessing involves a multiprocessor computer hardware architecture where two or more identical processors are connected to a single shared main memory and are controlled by a single OS instance.
- Most common multiprocessor systems today use an SMP architecture.
- In the case of multi-core processors, the SMP architecture applies to the cores, treating them as separate processors.

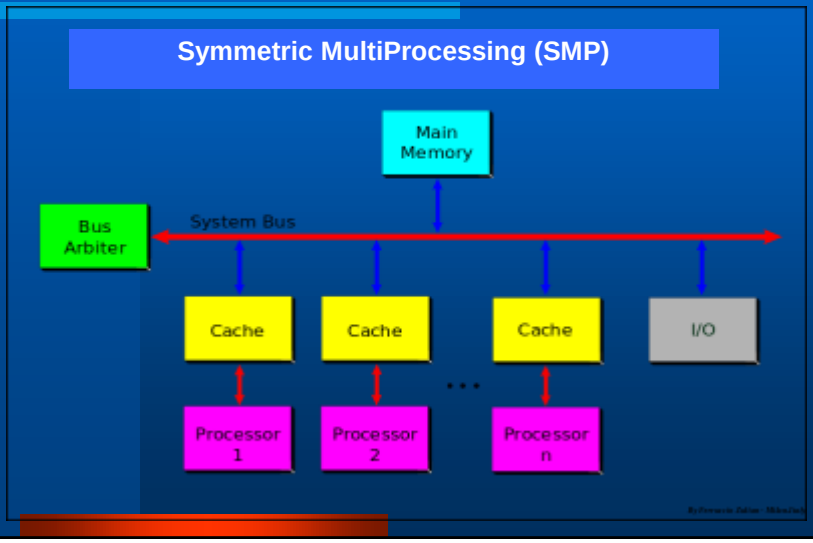
46

Symmetric MultiProcessing (SMP) (2/3)

- SMP systems are tightly coupled multiprocessor systems with a pool of homogeneous processors running independently
- Each processor executing different programs and working on different data and with capability of sharing common resources (memory, I/O device, interrupt system and so on) and connected using a system bus or a crossbar.

47

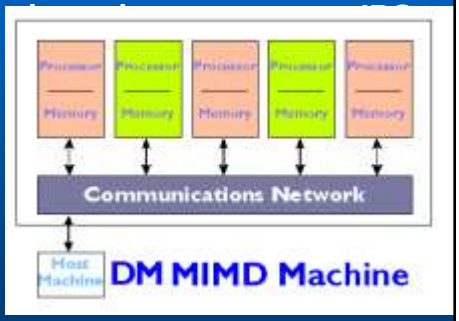
Symmetric MultiProcessing (SMP) (3/3)



Distributed Memory MIMD

- Communication
  (Inter-Process Communication)
  via High Speed Network
- Network can be configured to meet different topologies :

  Tree, Mesh, Cube, Torus, etc.



- Unlike Shared MIMD
  - ➔ easily/ readily expandable
  - ➔ Highly reliable (any CPU failure does not affect the whole system)
  - ➔ Cluster and grid computing architectures are of this kind

Shared Nothing MIMD

MPP (massively parallel processing) (1/2)

- Massively Parallel Processing) is the coordinated processing of a program by multiple processors that work on different parts of the program, with each processor using its own operating system and memory.
- MPP processors communicate using some messaging interface.
- Any "interconnect" which can arrange a data paths allows messages to be sent between processors.
- The setup for MPP is complicated, requiring thought about how to partition a common database among processors and how to assign work among the processors and how to communicate.

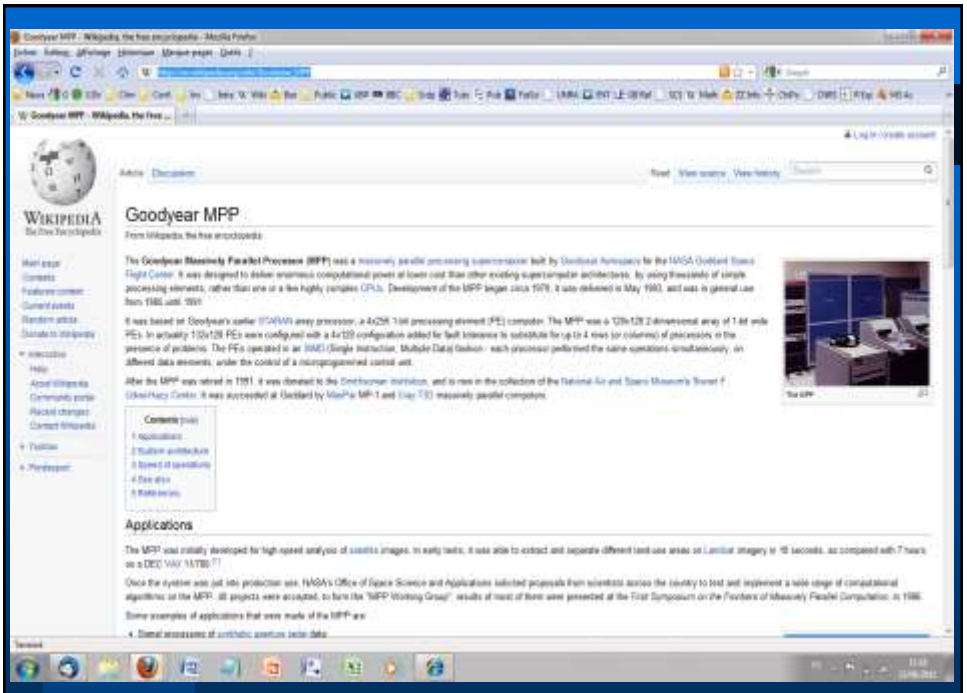
50

MPP (massively parallel processing) (2/2)

- An MPP system is also known as a :
  - "loosely coupled"

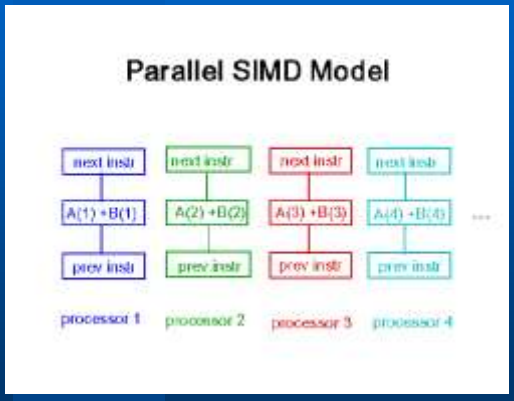
  or
  - "shared nothing" system.
- An MPP system is considered better than a symmetrically parallel system (SMP) for applications that
  - allow a number of databases to be searched in parallel (see Map/Reduce and other related approaches)
- These include decision support system and data warehouse applications.

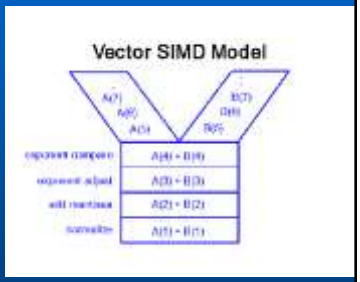
51



SIMD Architecture

$C_i = A_i * B_i$





Ex: Initial CRAY machine vector processing,

Thinking machine cm* and Intel MMX, SSEn, and now AVXn

SPMD – common parallelism

- SPMD (single program, multiple data) is a technique employed to achieve parallelism it is a subcategory of MIMD with a behavior close to SIMD
- SPMD vs SIMD:
  - In SPMD, multiple autonomous processors simultaneously execute the same program at independent points, rather than just a lockstep that SIMD imposes on different data.
  - With SPMD, tasks can be executed on general purpose CPUs;
  - SIMD requires vector processors to manipulate data streams. The two are not mutually exclusive.

54

Part IV
HPC evolution
Electronics behind

Moore's law and Dennard scaling

- **Gordon Moore (Intel co-founder) – number of devices per chip doubles every 18 months (Electronics magazine 1965) => 2X transistors every 1.5 years**
- **Moore's secret : Dennard et al. Scaling – IEEE JSSC 1974**

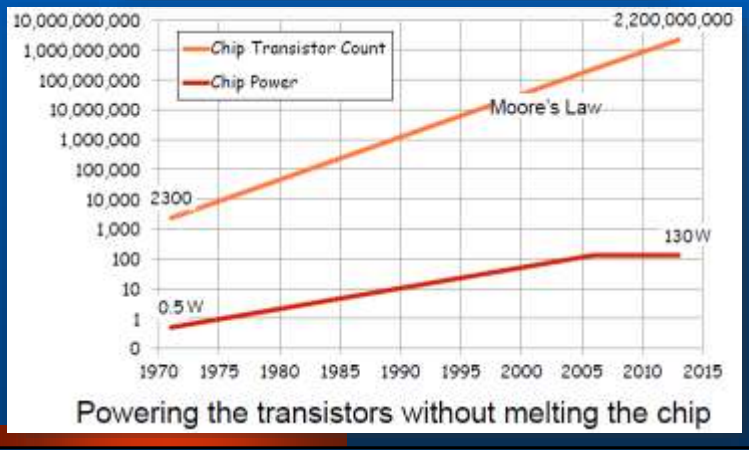
Dennard Scaling :
- Decrease feature size by a factor of λ and decrease voltage by a factor of λ ; then
- # transistors increase by λ^2
- Clock speed increases by λ
- **Energy consumption does not change**

2x transistor count
40% faster
50% more efficient

56

Dennard scaling is over since 10 years

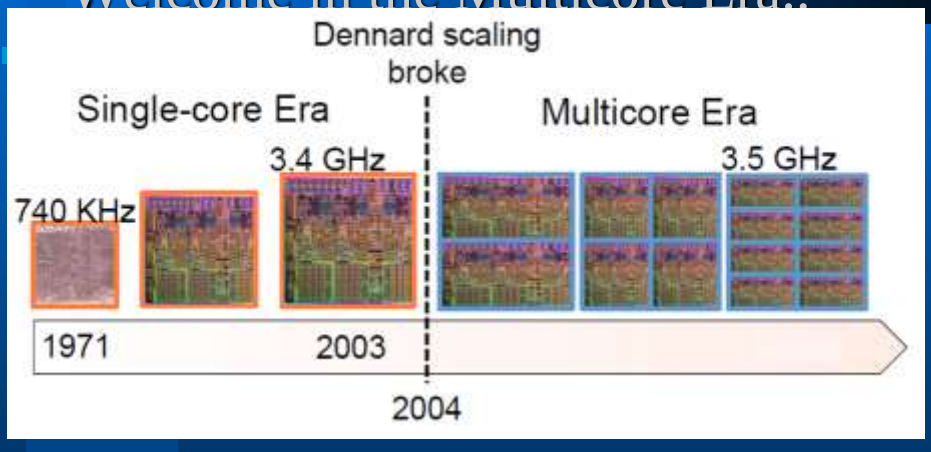
Breakdown is the result of small feature sizes, current leakage poses greater challenges, and also causes the chip to heat up



Powering the transistors without melting the chip

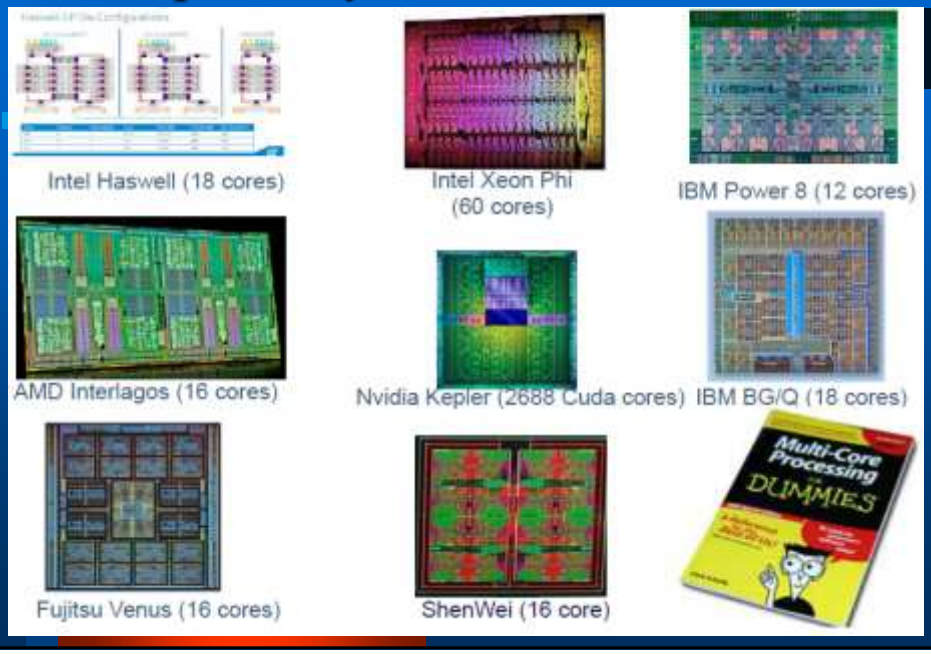
57

Welcome in the Multicore Era..



58

All top 500 systems use multicores



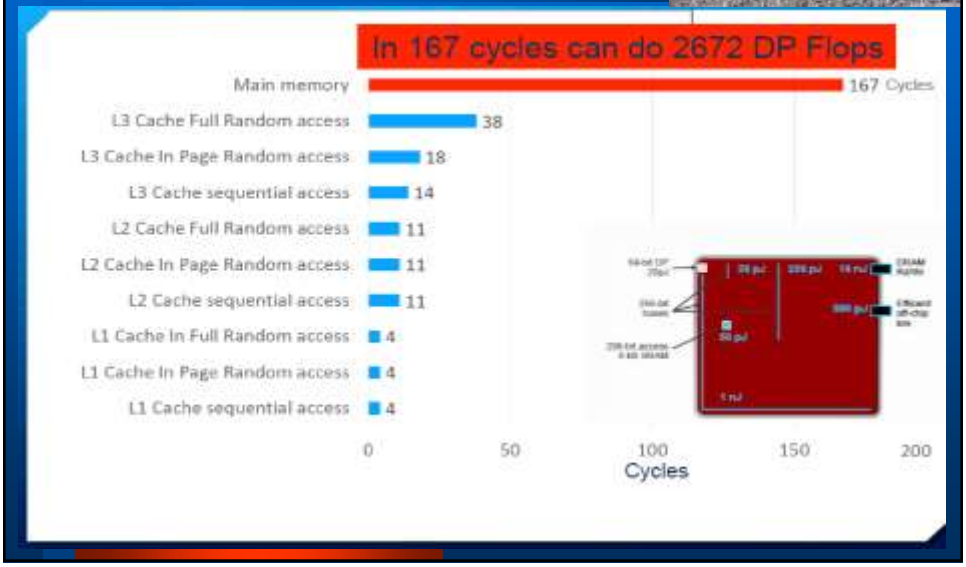
Peak performance per core

Floating point operations per cycle per core

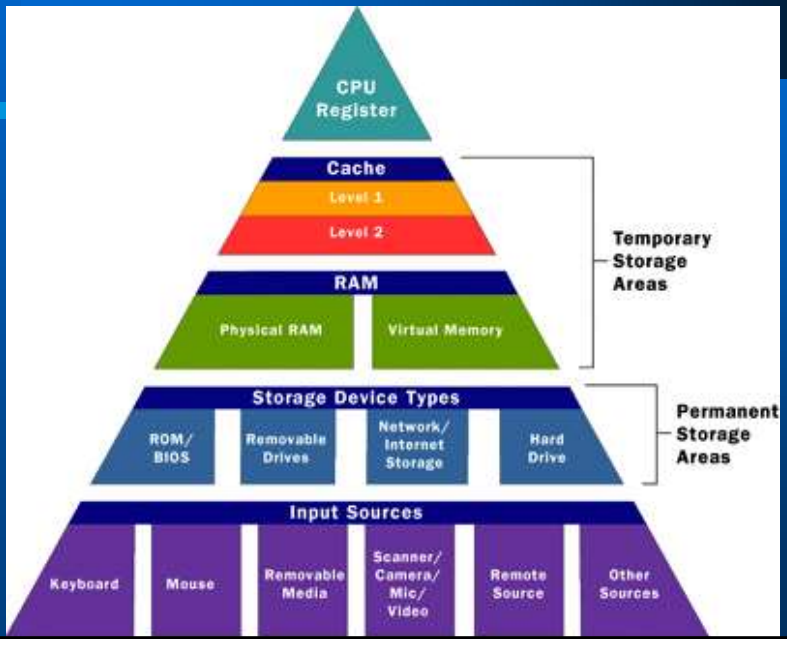
- Most of the recent computers have FMA (Fused multiple add): (i.e. x ←x + y*z in one cycle)
- Intel Xeon earlier models and AMD Opteron have SSE2
  - 2 flops/cycle DP & 4 flops/cycle SP
- Intel Xeon Nehalem ('09) & Westmere ('10) have AVX
  - 4 flops/cycle DP & 8 flops/cycle SP
- Intel Xeon Sandy Bridge('11) & Ivy Bridge ('12) have AVX & AVX2
  - 8 flops/cycle DP & 16 flops/cycle SP
- Intel Xeon Haswell ('13) & (Broadwell ('14)) AVX2
  - 16 flops/cycle DP & 32 flops/cycle SP
  - Xeon Phi (per core) is at 16 flops/cycle DP & 32 flops/cycle SP
- Intel Xeon Skylake & Kabylake
  - 32 flops/cycle DP & 64 flops/cycle SP

We are here

Memory Access Latency
« The memory wall »

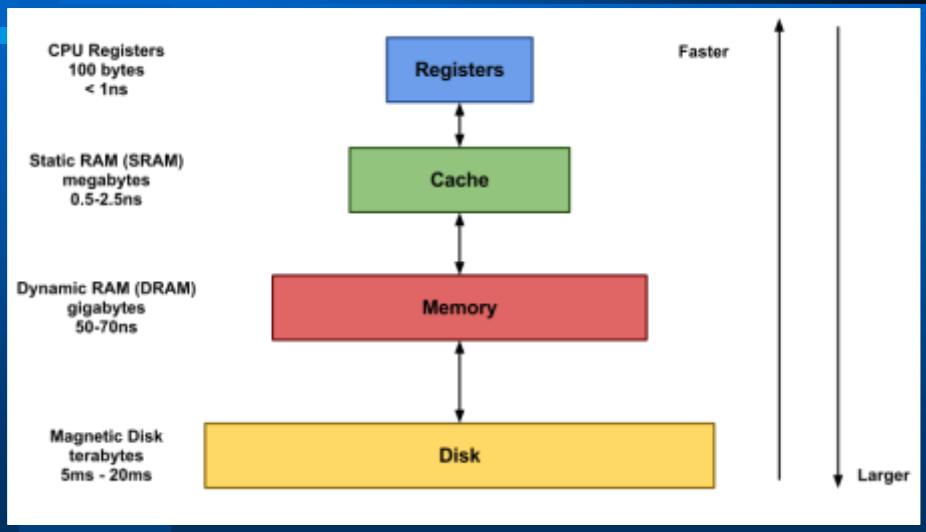


Memory Hierarchy (1/2)

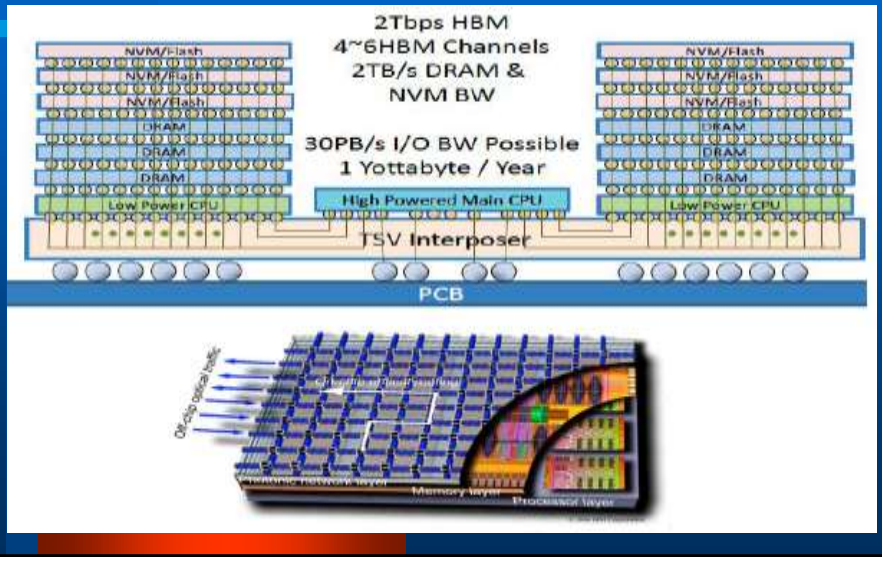


62

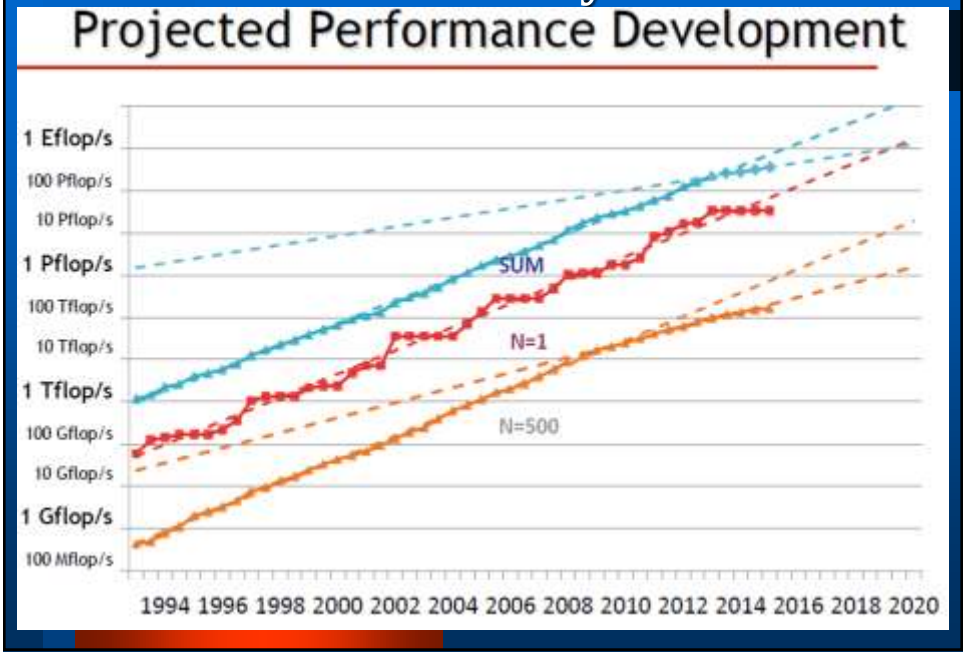
Memory Hierarchy (2/2)
Speed scale



Next generation 3D design with integrated photonics (HP,IBM,Intel…)



Where will we be in 2 years ?



Big data and data science… (since 2001)

- **CERN – HPC**
  European Computing Grid
- **Finding : Higgs' Boson (Introduced in 1964)**
- **Discovery by LHC - ATLAS experiment in 2012 (4/7)**
- **Confirms current Model for Standard Physics**