Introduction

CS121 Parallel Computing

Fall 2023

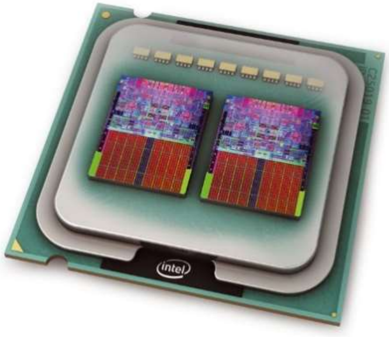
Course info

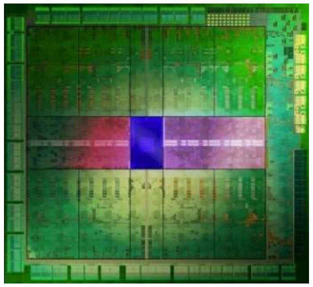
Instructor	Assoc Prof Rui FAN 范睿
Research	Parallel and distributed computing
Contact	fanrui@shanghaitech.edu.cn (English please)
Office hours	Thursdays 4:40-6pm, SIST 1A-504E
TA	范云潜, fanyq2022@shanghaitech.edu.cn
Recitation	TBA
Website	Blackboard and Piazza

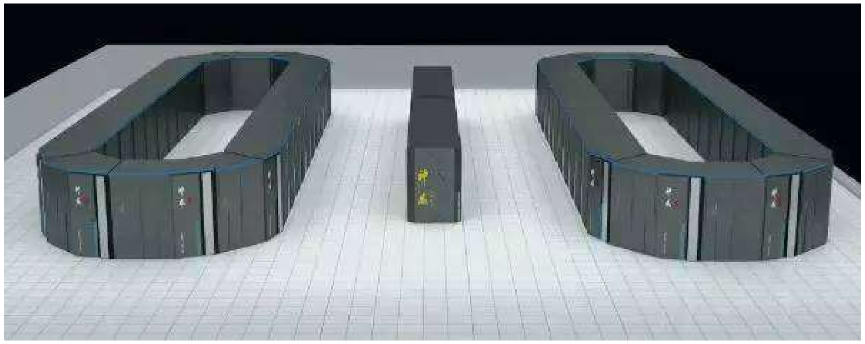
Problem sets	20%	▪ About once every 2 weeks
Labs	20%	▪ Solve problems using OpenMP and CUDA
Reading project	15% Teams of 2	▪ Find an interesting research paper from suggested reading list ▪ Tell me your paper by week 8 ▪ Submit a report after week 16
Programming project	15% Teams of 2	▪ Find an interesting problem and write an efficient parallel program for it ▪ Tell me your problem by week 8 ▪ Submit a report and give a 20 minute presentation in week 16
Midterm exam	10%	▪ In week 9
Final exam	20%	

Parallel computing: what and why

- Parallel computing studies how to use multiple computers together to solve a problem.
- Allows solving complicated problems faster.
  - Ideally, with *k* processors we can solve a problem *k* times faster.
  - Also more memory to solve larger problems, or same problem with more accuracy.
  - May be more fault tolerant; but also more prone to faults.
- Almost all modern computer systems are parallel.
  - Multicores, GPUs, cloud computing, etc.
- Parallel computing crucial for modern large scale applications, e.g. physical simulations, data mining, machine learning.





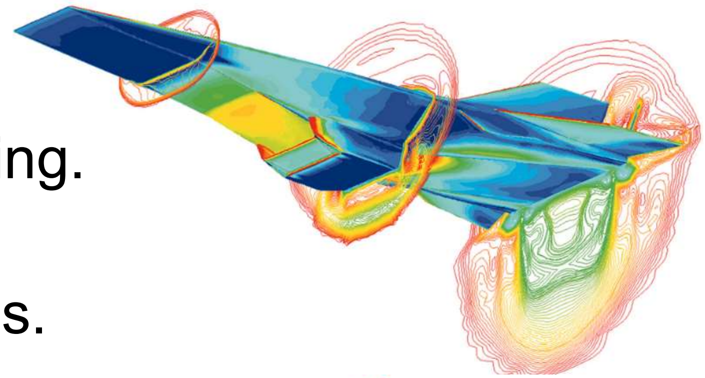
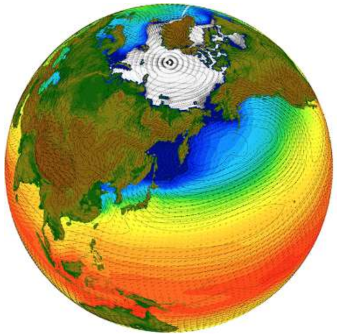
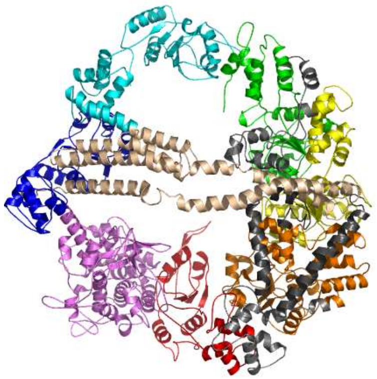
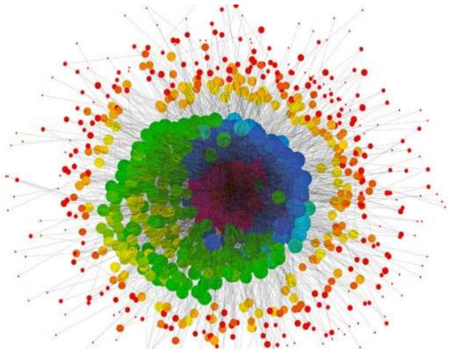


Course objectives

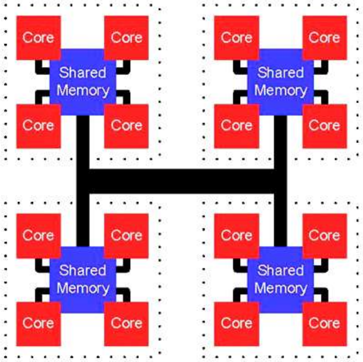
- To understand the concepts and techniques of parallel computing, and take advantage of the capabilities of modern systems.
  - Parallel hardware models and interaction with parallel software.
  - Power and limitations of parallelism.
  - Efficient parallel algorithms for important problems.

Applications

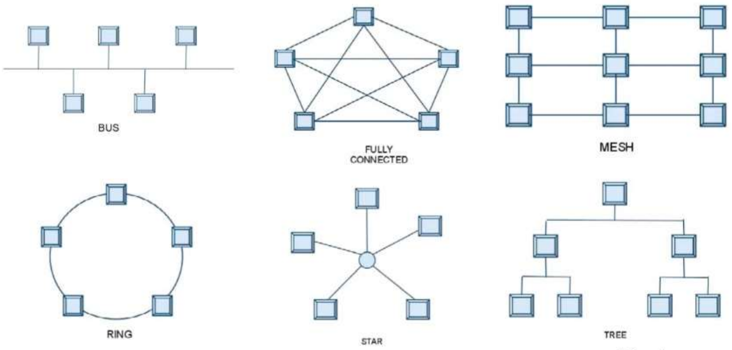
- Fluid dynamics, weather prediction, climate modeling.
- DNA, protein, drug structures and interactions.
- Quantum / atomic simulations, cosmological simulations.
- Cryptoanalysis.
- Big data analytics.
- Simulating financial and social behaviors.
- Machine learning and AI.
- Simulating the human brain.

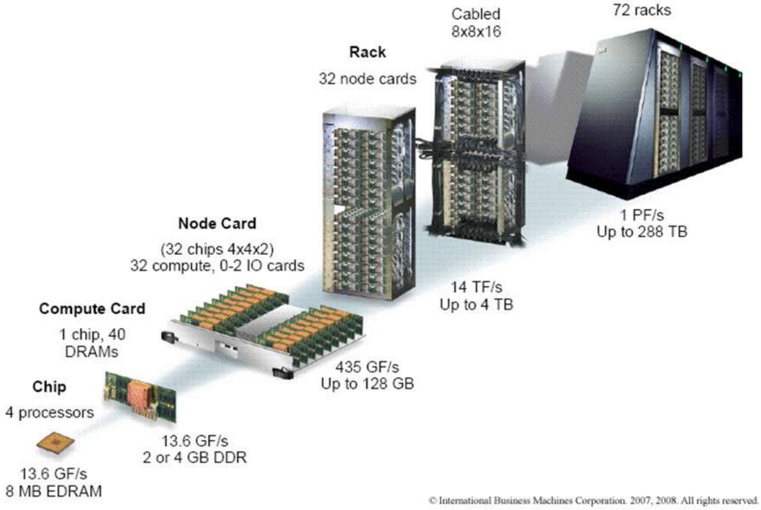





Parallel hardware



- Efficient parallel computing requires synergy between parallel hardware and software.
- Parallel system consists of multiple independent processors communicating over an interconnect.
- Unlike sequential (von Neumann) architecture, many parallel hardware designs.
  - Different types of processors (multicores, manycores, FPGA, etc.).
  - Heterogeneous designs combine multiple architectures, e.g. multicores and GPUs.
  - Different interconnect designs.
  - Communicate through shared memory, or message passing over network.
- Parallelism exists at many layers.
  - Instruction, core, chip, node, rack, etc.

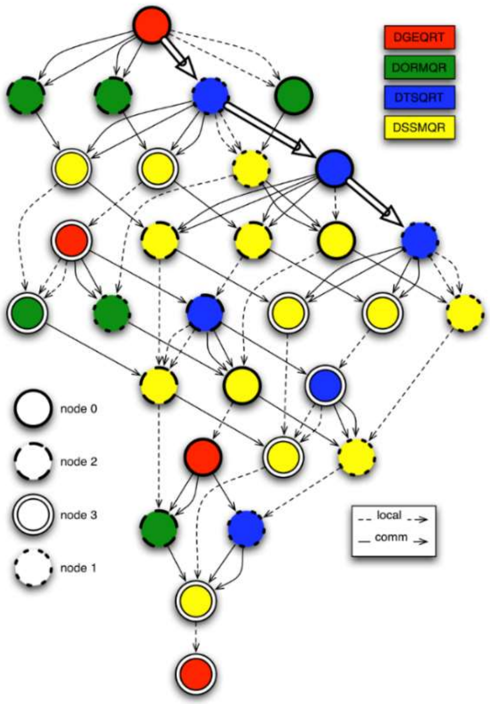




© International Business Machines Corporation. 2007, 2008. All rights reserved.

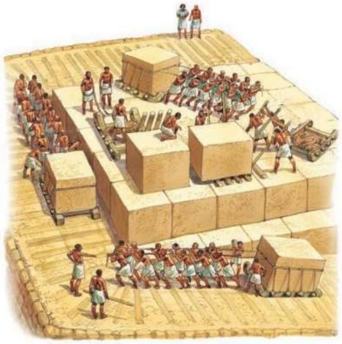
Parallel software

- Break a large problem into subproblems (tasks) that can be solved (somewhat) independently.
- OS and scheduler allocate tasks to different processors.
  - Respect dependencies between tasks.
- Parallel software must be matched to the hardware.
  - Similar amounts of concurrency in software and hardware.
  - Hardware must adequately handle software communication pattern.
  - No single hardware model suffices.
  - Parallel software is often not portable.
- PRAM model tries to abstract parallel hardware.
  - Useful for understanding inherent parallelism.
  - Unrealistically discounts cost of communication.



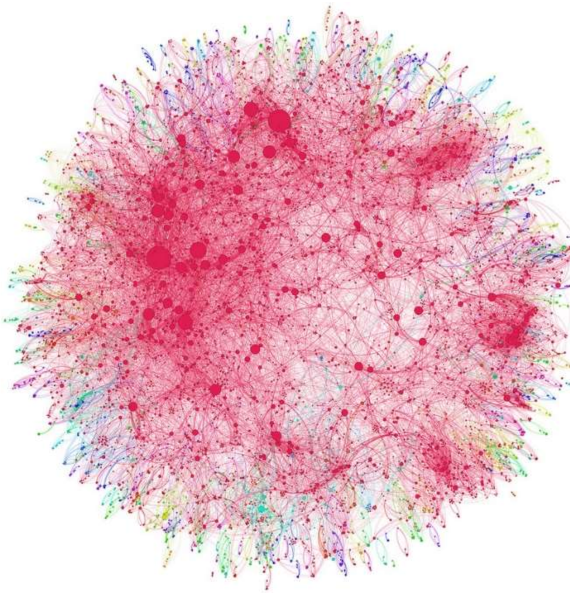
Challenges

- Harnessing power of the masses.
  - Easier said than done...
- Communication
  - Processors compute faster than they can communicate.
  - Problem gets worse as number of processors increases.
  - Main bottleneck to parallel computing.
- Synchronization
  - Tasks may interfere with each other, so can't be done at same time.
- Scheduling
  - Track and enforce dependencies.
  - Find good allocation of tasks to processors.
    - Data locality, heterogeneous processors
  - Maximize utilization and performance.



Challenges

- Structured vs unstructured
  - Structured problems can be solved with custom hardware.
  - Unstructured problems more general, but less efficient.
- Inherent limitations
  - Some problems are not (or don’t seem to be) parallelizable.
    - Ex Binary search, Dijkstra’s shortest paths algorithm.
  - Other problems require clever algorithms to become parallel.
    - Ex Fibonacci series ($a_n = a_{n-1} + a_{n-2}$).
- The human factor
  - Hard to keep track of concurrent events and dependencies.
  - Parallel algorithms are hard(er) to design and debug.



Course outline

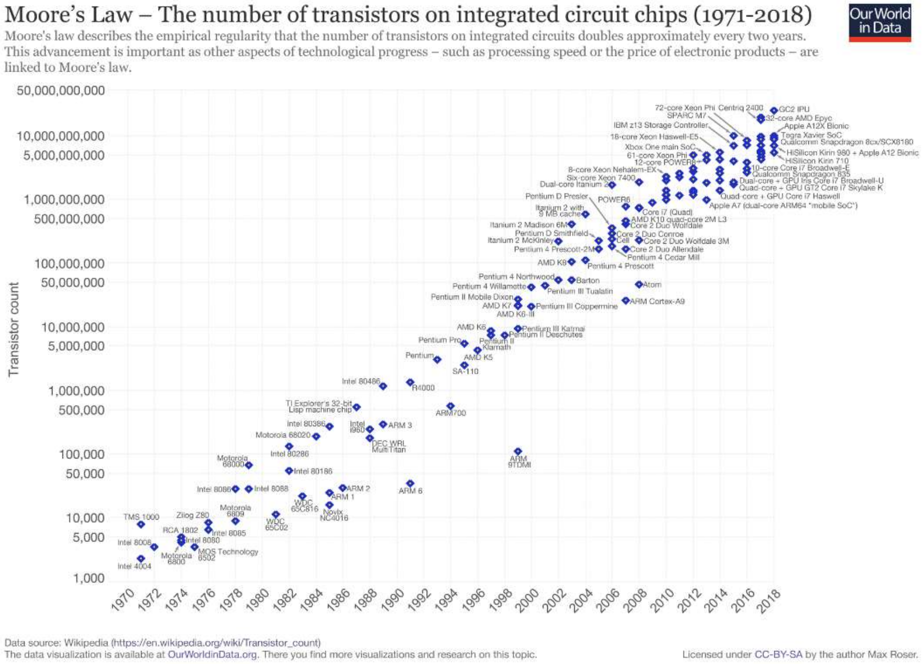
- Parallel architectures
  - Shared memory
  - Distributed memory
  - Manycore
- Parallel languages
  - OpenMP, MPI, CUDA, MapReduce
- Algorithm design techniques
  - Decomposition, load balancing, scheduling
- Parallel algorithms
  - Dense and sparse matrix algorithms, sorting, search, graph algorithms, PRAM algorithms, etc.

A brief history

- Research and theory started in the early 60's.
  - Cray-1 reached 160 MFLOPS in 1976.
- Commercially successful supercomputers (Cray, Thinking Machines, etc.) started in 1980's.
  - Used expensive custom processors.
- In 1990's massively parallel processors (MPPs) and clusters became dominant.
  - MPPs use commercial (OTS) processors with custom interconnects.
  - Clusters use OTS processors and interconnects running Linux.
    - Cheap, easy to build and relatively powerful.
    - Most data centers today are clusters.
- Fastest supercomputer today is Fujitsu Fugaku MPP.
  - Runs at 442 PFLOPS, about 3M times faster than a workstation.
- Apart from supercomputers, progress in parallel computing stalled in 1990's until mid 2000's.

Moore’s Law and parallel computing

- In 1965, Gordon Moore, co-founder of Intel, predicted transistor count would double every 18 months.
  - □ Held true for the last 50 years!
- Until mid 2000’s, this implied single processor performance doubled at same rate.
- This held back development of parallel computers, since in the time to develop one, single processor performance would improve dramatically.
- But since ca. 2005, parallel processing has become essential for taking advantage of Moore’s Law.



Moore’s Law – The number of transistors on integrated circuit chips (1971-2018)

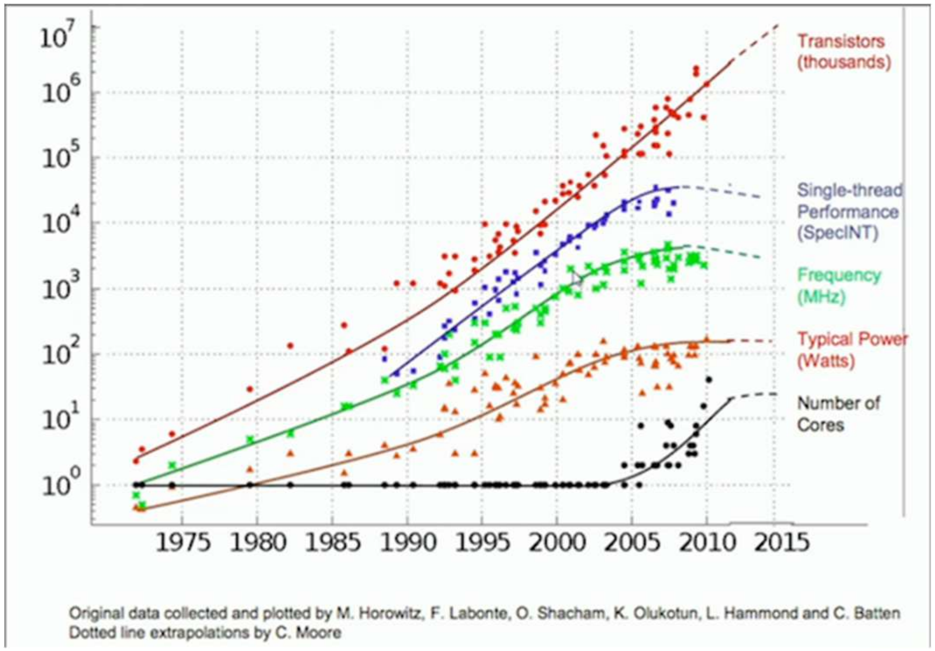
Moore’s law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore’s law.

Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Licensed under CC-BY-SA by the author Max Roser.

Moore’s Law and performance

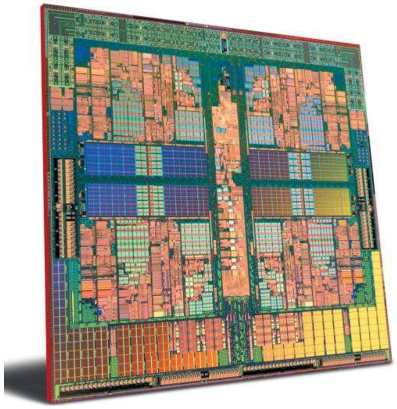
- Transistor properties, e.g. size and clock speed, do not scale equally.
- Higher single processor clock speeds is increasingly difficult to achieve.
  - Heat
  - Power consumption
  - Current leakage

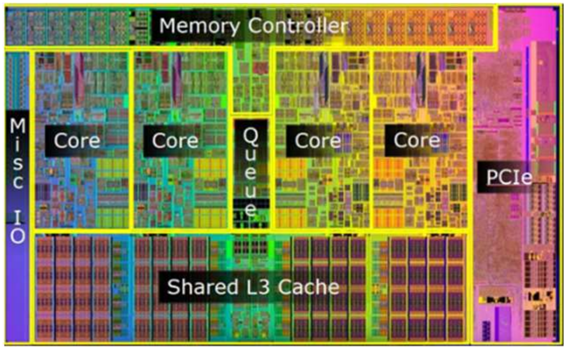


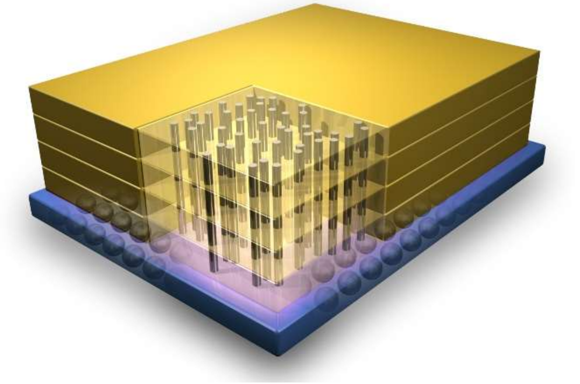
Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Moore’s Law revisited

- Multicore technology addresses (lack of) clock speed scaling.
  - Link multiple processing cores together on same chip.
  - More efficient to replace a single high speed processor with multiple slower processors.
  - Another approach is to stack chips in a 3D structure.
- Developing software for multicores has been harder than scaling hardware.
  - Software developers with parallel computing skills are in high demand.





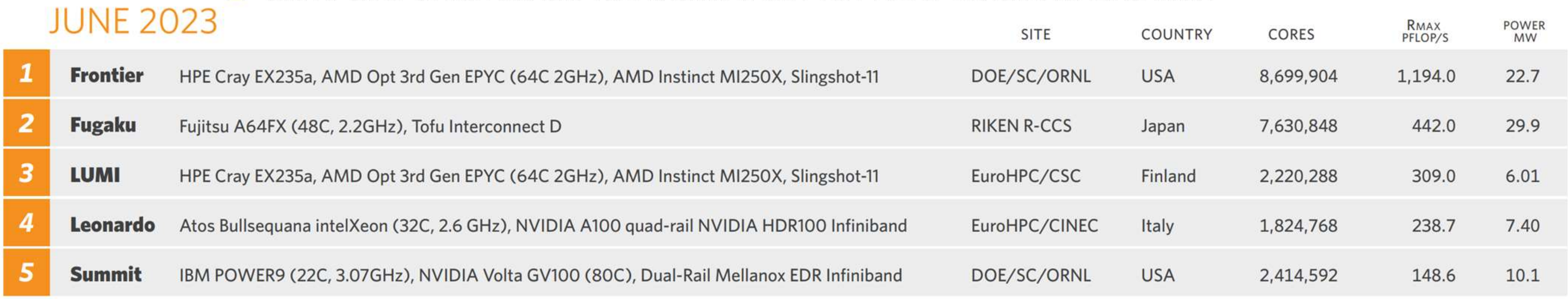


The state of the art

- Parallel computers today mainly based on four processor architectures.
  - Multicores
    - Small / moderate number (≤ 128) of fast, general purpose cores.
    - Ex AMD EPYC, Intel Xeon, IBM Power.
  - Manycores
    - Large number (10K's) of simple cores.
    - Ex Nvidia Ampere GPU, Intel Xeon Phi, Sunway SW26010Pro.
  - FPGA (field programmable gate arrays)
    - Reconfigurable hardware customized for specific problems.
  - ASIC (application specific integrated circuits)
    - Specially built hardware for specific problems.
    - Ex Google TPU, Graphcore IPU, IBM TrueNorth.
- In addition to processing speed, energy efficiency also increasing important.
  - Biggest datacenters consume over 100 MW of power, ~ 50K homes.
  - Biggest supercomputers consume ~ 20MW of power.
  - Best supercomputers achieve 50 GFLOPS / W.

Top 500 list

- Biannual ranking of fastest 500 supercomputers in the world.
  - □ Speed measured in floating point operations per second.
  - □ Uses high-performance LINPACK to solve a dense linear system Ax = b.
    - Compute intensive, but doesn't stress memory system.
    - May not represent performance on real-world problems



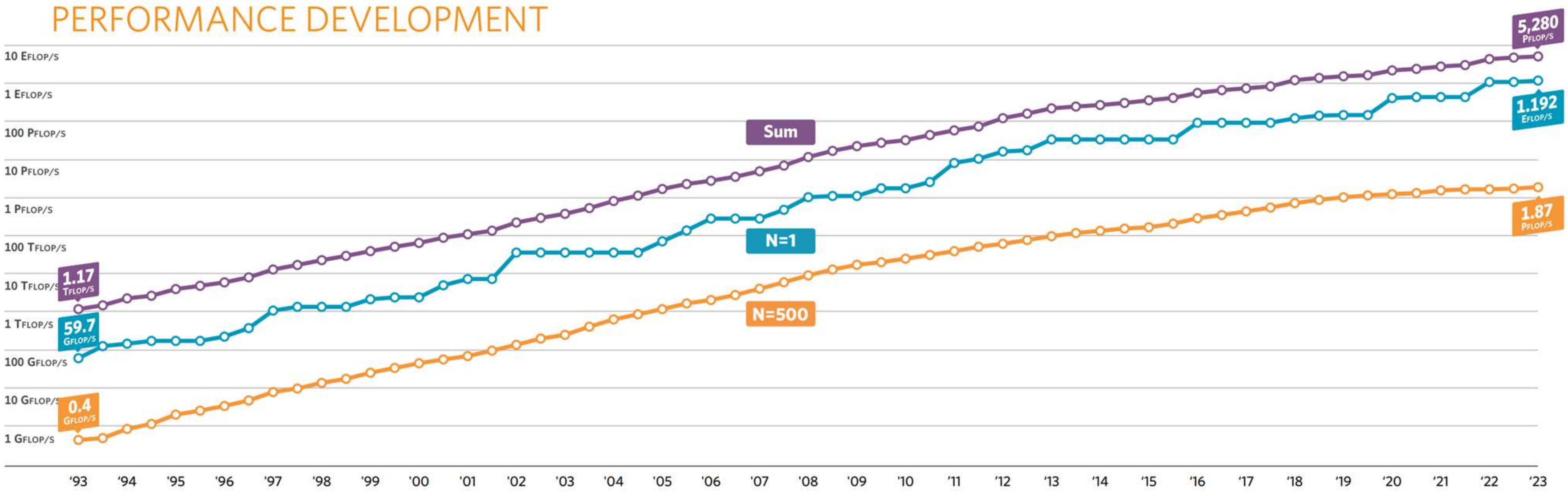
JUNE 2023

			SITE	COUNTRY	CORES	RMAX PFLOP/S	POWER MW
1	**Frontier**	HPE Cray EX235a, AMD Opt 3rd Gen EPYC (64C 2GHz), AMD Instinct MI250X, Slingshot-11	DOE/SC/ORNL	USA	8,699,904	1,194.0	22.7
2	**Fugaku**	Fujitsu A64FX (48C, 2.2GHz), Tofu Interconnect D	RIKEN R-CCS	Japan	7,630,848	442.0	29.9
3	**LUMI**	HPE Cray EX235a, AMD Opt 3rd Gen EPYC (64C 2GHz), AMD Instinct MI250X, Slingshot-11	EuroHPC/CSC	Finland	2,220,288	309.0	6.01
4	**Leonardo**	Atos Bullsequana intelXeon (32C, 2.6 GHz), NVIDIA A100 quad-rail NVIDIA HDR100 Infiniband	EuroHPC/CINEC	Italy	1,824,768	238.7	7.40
5	**Summit**	IBM POWER9 (22C, 3.07GHz), NVIDIA Volta GV100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/SC/ORNL	USA	2,414,592	148.6	10.1

Mega	**Giga**	**Tera**	**Peta**	**Exa**
10^6	10^9	10^{12}	10^{15}	10^{18}

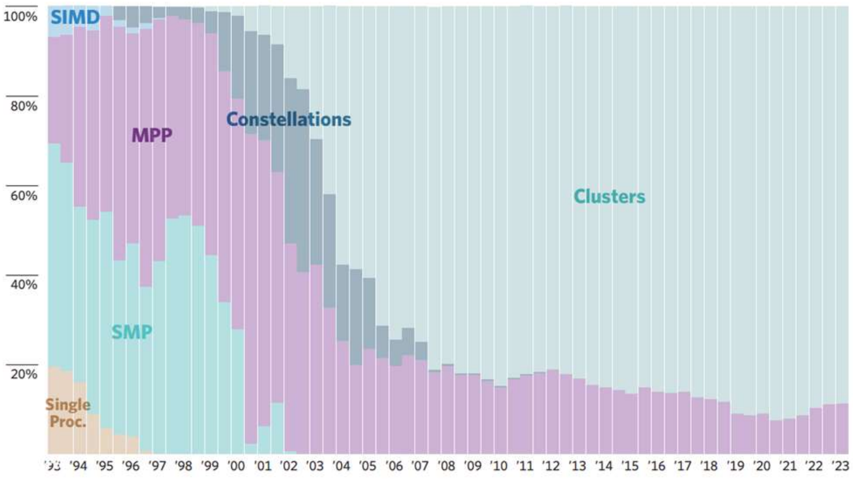
- □ For comparison, Intel multicore achieves ~100 GFLOPS / core, and GPU achieves ~50 TFLOPS / card.

Top 500 – Trends

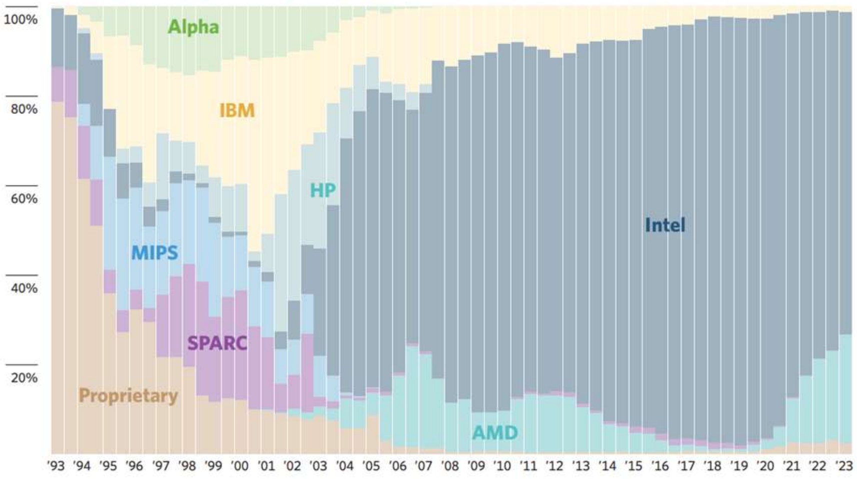


Top 500 – Architecture

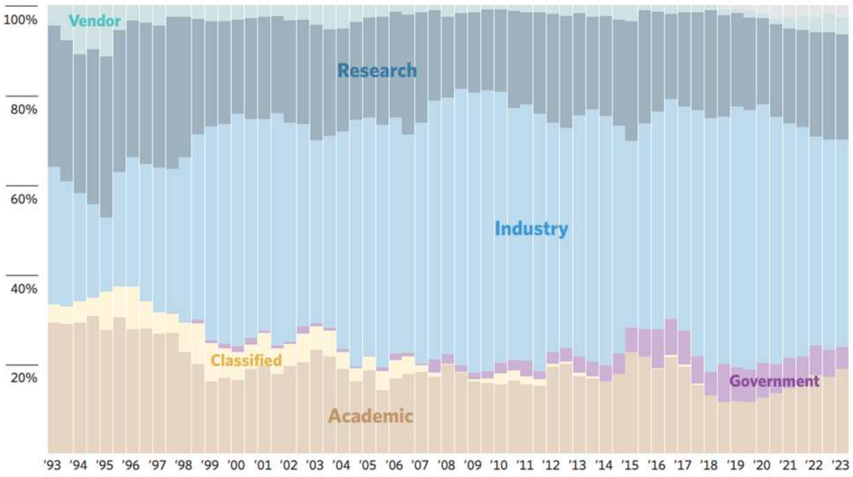
ARCHITECTURES



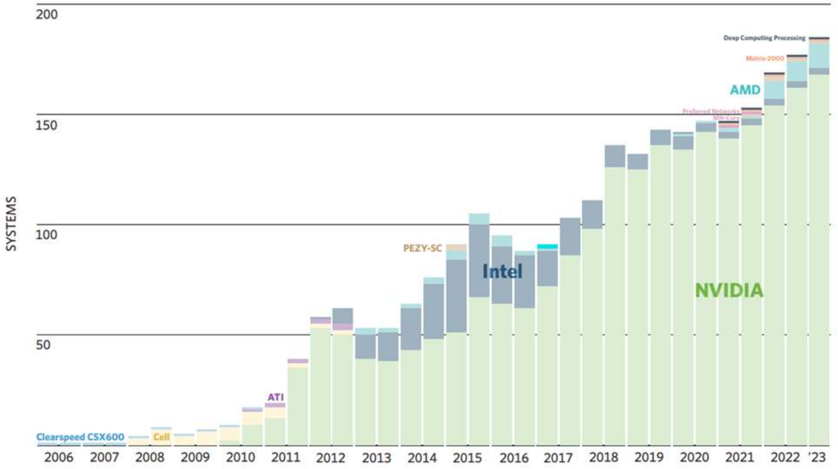
CHIP TECHNOLOGY



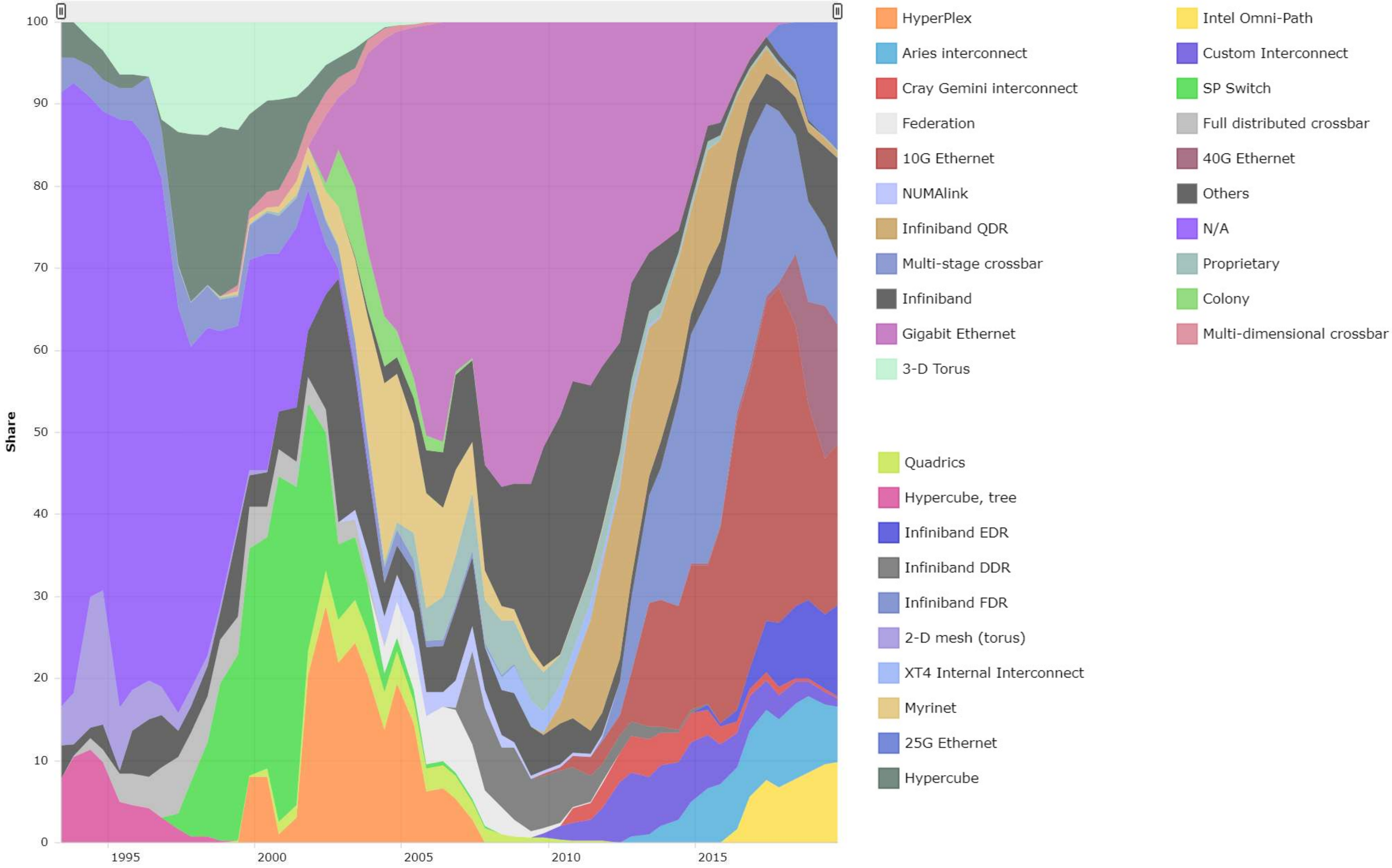
INSTALLATION TYPE



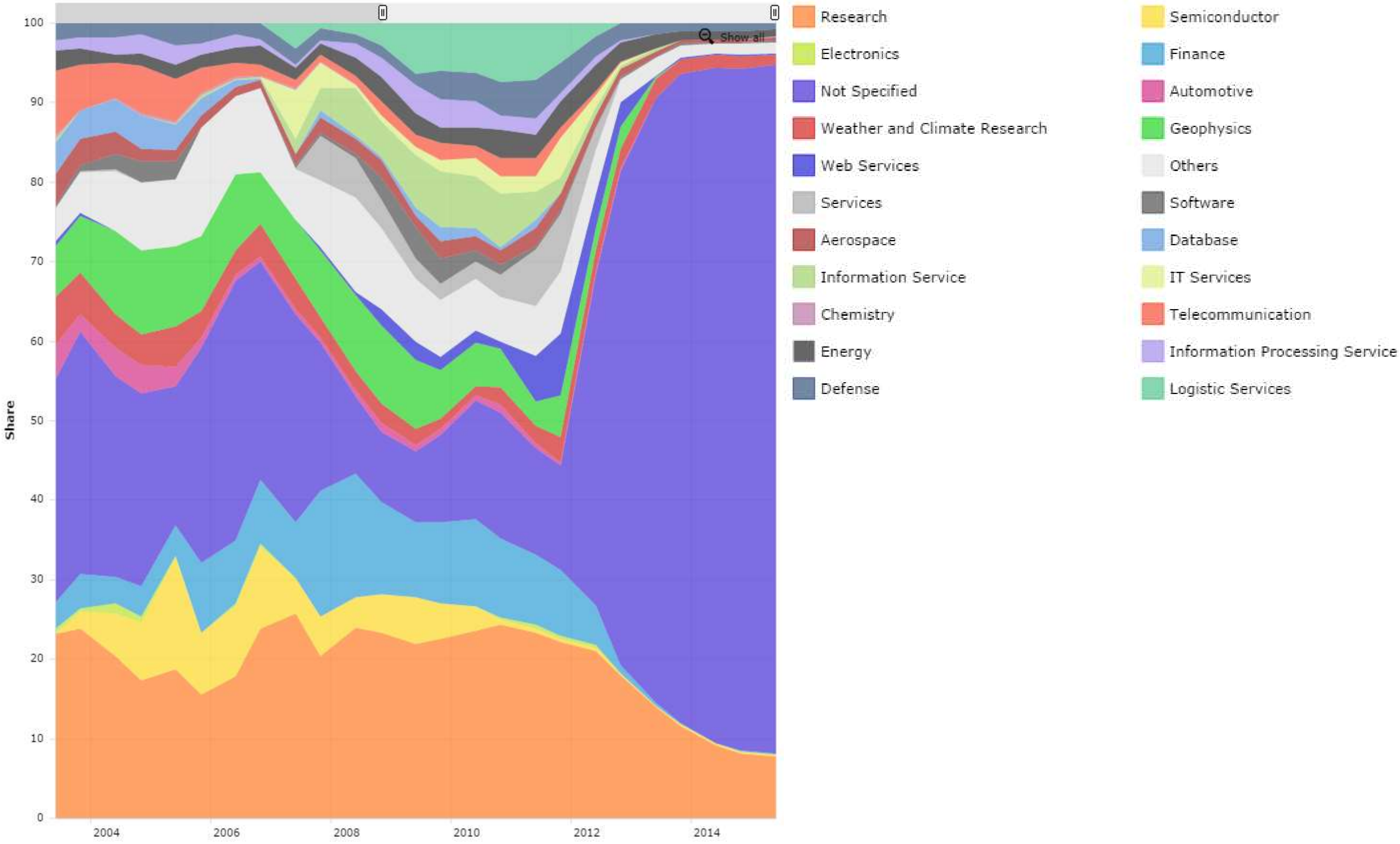
ACCELERATORS/CO-PROCESSORS



Top 500 – Interconnect



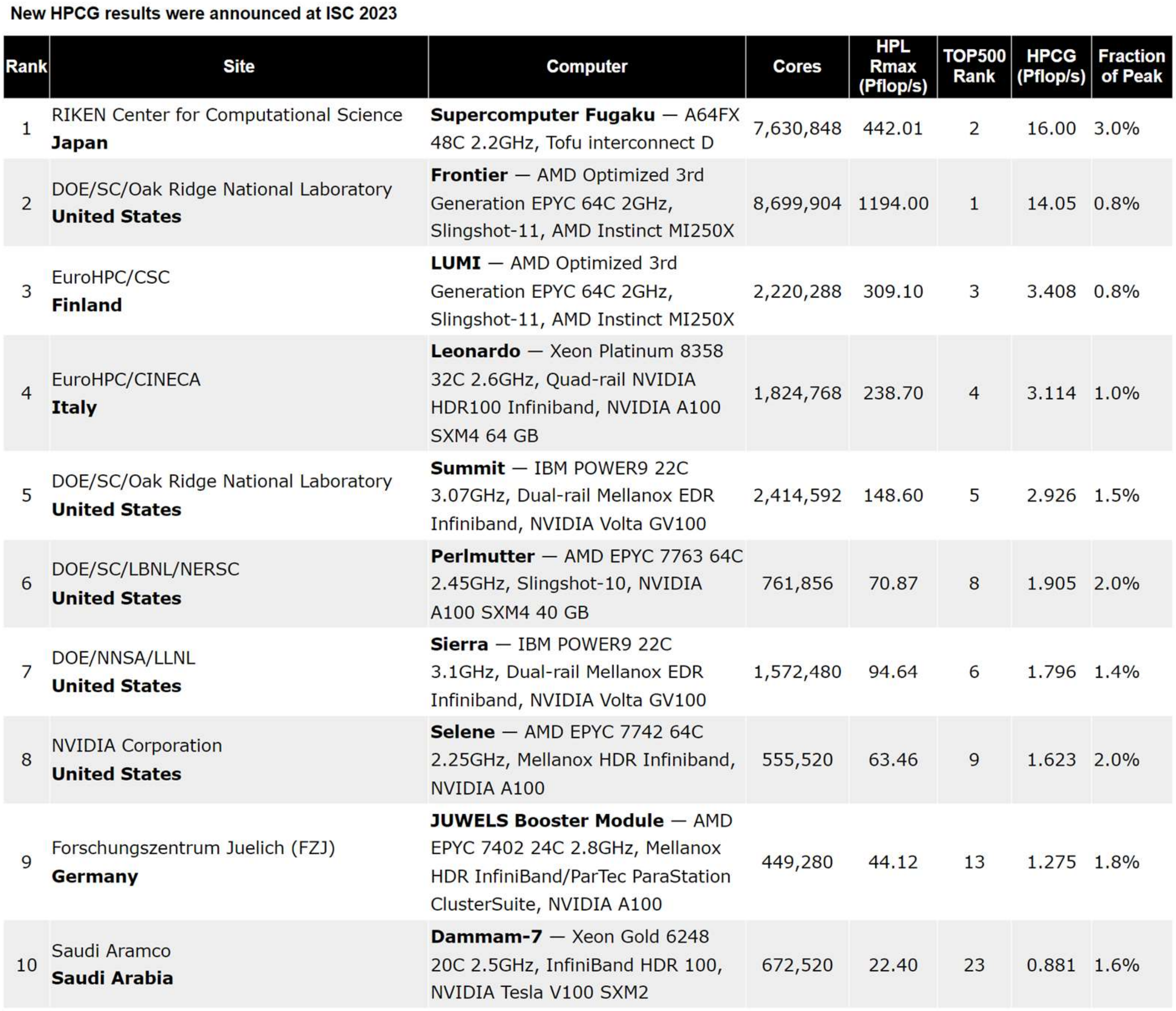
Top 500 – Applications



Other performance measures

- LINPACK does compute-intensive operations on structured dense matrices.
  - Uniform control flow, predictable and coalesced memory accesses.
  - Ideal for physical simulations.
- Data-intensive applications today have instruction divergence, branching and random memory accesses.
- New benchmarks give more complete performance picture
  - HPCG performs sparse matrix operations.
  - Graph500 performs breadth-first search.
- A computer's performance can differ dramatically depending on benchmark.

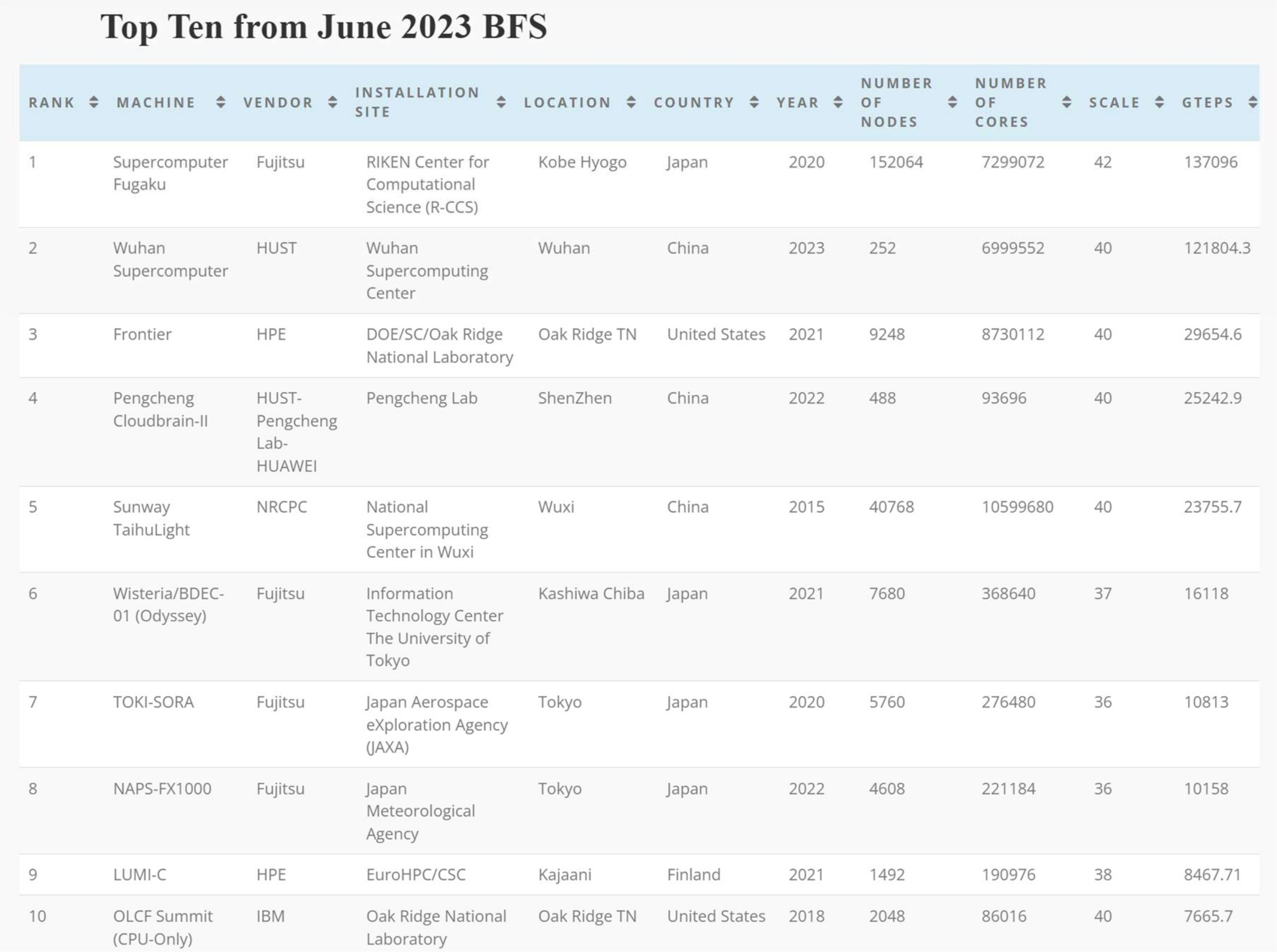
HPCG



New HPCG results were announced at ISC 2023

Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	RIKEN Center for Computational Science **Japan**	**Supercomputer Fugaku** — A64FX 48C 2.2GHz, Tofu interconnect D	7,630,848	442.01	2	16.00	3.0%
2	DOE/SC/Oak Ridge National Laboratory **United States**	**Frontier** — AMD Optimized 3rd Generation EPYC 64C 2GHz, Slingshot-11, AMD Instinct MI250X	8,699,904	1194.00	1	14.05	0.8%
3	EuroHPC/CSC **Finland**	**LUMI** — AMD Optimized 3rd Generation EPYC 64C 2GHz, Slingshot-11, AMD Instinct MI250X	2,220,288	309.10	3	3.408	0.8%
4	EuroHPC/CINECA **Italy**	**Leonardo** — Xeon Platinum 8358 32C 2.6GHz, Quad-rail NVIDIA HDR100 Infiniband, NVIDIA A100 SXM4 64 GB	1,824,768	238.70	4	3.114	1.0%
5	DOE/SC/Oak Ridge National Laboratory **United States**	**Summit** — IBM POWER9 22C 3.07GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100	2,414,592	148.60	5	2.926	1.5%
6	DOE/SC/LBNL/NERSC **United States**	**Perlmutter** — AMD EPYC 7763 64C 2.45GHz, Slingshot-10, NVIDIA A100 SXM4 40 GB	761,856	70.87	8	1.905	2.0%
7	DOE/NNSA/LLNL **United States**	**Sierra** — IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100	1,572,480	94.64	6	1.796	1.4%
8	NVIDIA Corporation **United States**	**Selene** — AMD EPYC 7742 64C 2.25GHz, Mellanox HDR Infiniband, NVIDIA A100	555,520	63.46	9	1.623	2.0%
9	Forschungszentrum Juelich (FZJ) **Germany**	**JUWELS Booster Module** — AMD EPYC 7402 24C 2.8GHz, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, NVIDIA A100	449,280	44.12	13	1.275	1.8%
10	Saudi Aramco **Saudi Arabia**	**Dammam-7** — Xeon Gold 6248 20C 2.5GHz, InfiniBand HDR 100, NVIDIA Tesla V100 SXM2	672,520	22.40	23	0.881	1.6%

Graph500



Top Ten from June 2023 BFS

RANK	MACHINE	VENDOR	INSTALLATION SITE	LOCATION	COUNTRY	YEAR	NUMBER OF NODES	NUMBER OF CORES	SCALE	GTEPS
1	Supercomputer Fugaku	Fujitsu	RIKEN Center for Computational Science (R-CCS)	Kobe Hyogo	Japan	2020	152064	7299072	42	137096
2	Wuhan Supercomputer	HUST	Wuhan Supercomputing Center	Wuhan	China	2023	252	6999552	40	121804.3
3	Frontier	HPE	DOE/SC/Oak Ridge National Laboratory	Oak Ridge TN	United States	2021	9248	8730112	40	29654.6
4	Pengcheng Cloudbrain-II	HUST-Pengcheng Lab-HUAWEI	Pengcheng Lab	ShenZhen	China	2022	488	93696	40	25242.9
5	Sunway TaihuLight	NRCPC	National Supercomputing Center in Wuxi	Wuxi	China	2015	40768	10599680	40	23755.7
6	Wisteria/BDEC-01 (Odyssey)	Fujitsu	Information Technology Center The University of Tokyo	Kashiwa Chiba	Japan	2021	7680	368640	37	16118
7	TOKI-SORA	Fujitsu	Japan Aerospace eXploration Agency (JAXA)	Tokyo	Japan	2020	5760	276480	36	10813
8	NAPS-FX1000	Fujitsu	Japan Meteorological Agency	Tokyo	Japan	2022	4608	221184	36	10158
9	LUMI-C	HPE	EuroHPC/CSC	Kajaani	Finland	2021	1492	190976	38	8467.71
10	OLCF Summit (CPU-Only)	IBM	Oak Ridge National Laboratory	Oak Ridge TN	United States	2018	2048	86016	40	7665.7

HPCL State Key Laboratory of High Performance Computing

Overview of Tianhe-2 (MilkyWay-2) Supercomputer

Yutong Lu

School of Computer Science, National University of Defense Technology;
State Key Laboratory of High Performance Computing, China
ytlu@nudt.edu.cn

国防科学技术大学
National University of Defense Technology

Motivation

Tianhe-2 (Milkyway-2) Supercomputer

Specification

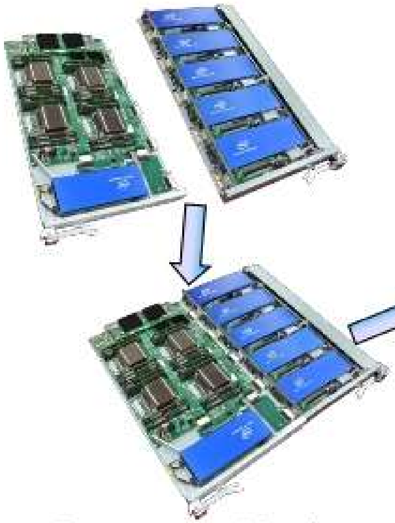
- **Hybrid Architecture**
  - **Xeon CPU & Xeon Phi**

Items	Configuration
Processors	32000 Intel Xeon CPUs + 48000 Xeon Phis + 4096 FT CPUs Peak performance is 54.9PFlops, HPL
Interconnect	Proprietary high-speed interconnection network TH Express-2
Memory	1.4PB in total
Storage	Global shared parallel storage system, 12.4PB
Cabinets	125+13+24=162 compute/communication/storage Cabinets
Power	17.8 MW (1902MFlops/W)
Cooling	Closed Air cooling system

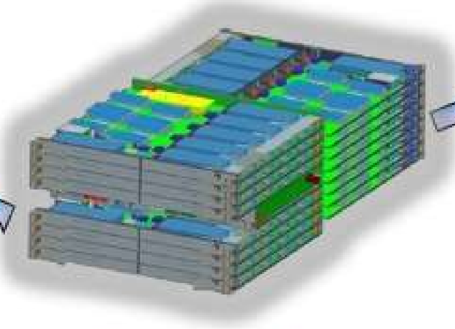
From Chips to Entire System

- ◆16000 compute nodes in total
- ◆Frame: 32 compute Nodes
- ◆Rack: 4 Compute Frames
- ◆Whole System: 125 Racks

System



Compute Node



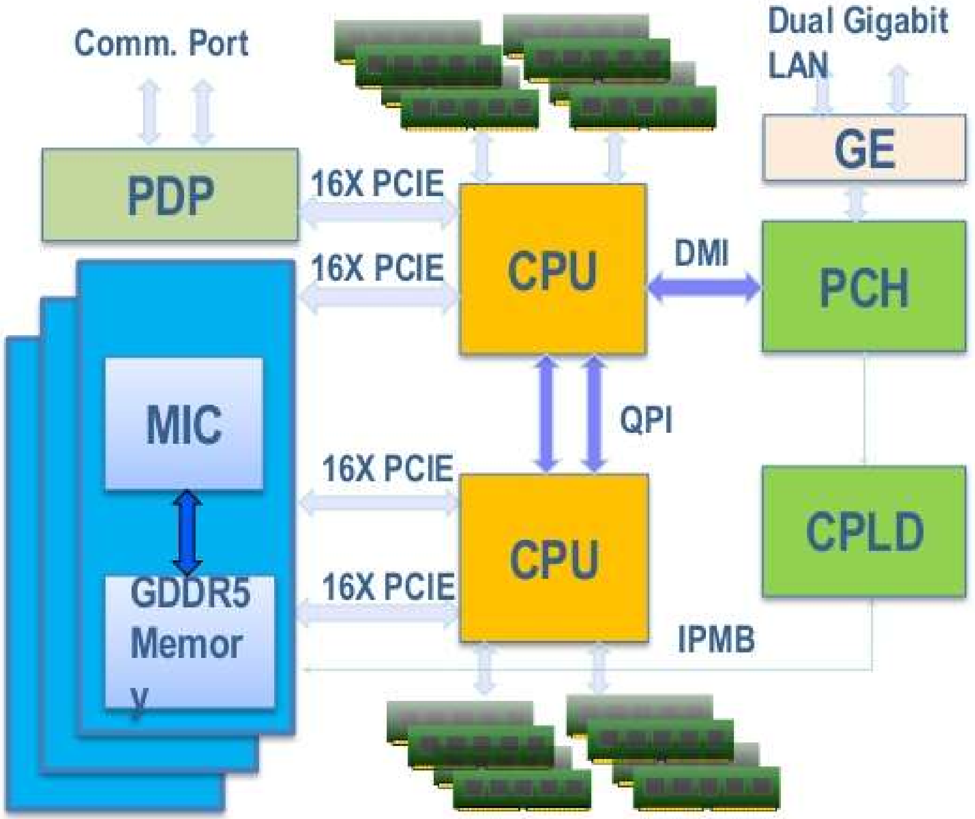
Compute Frame

Compute Rack

Compute Node

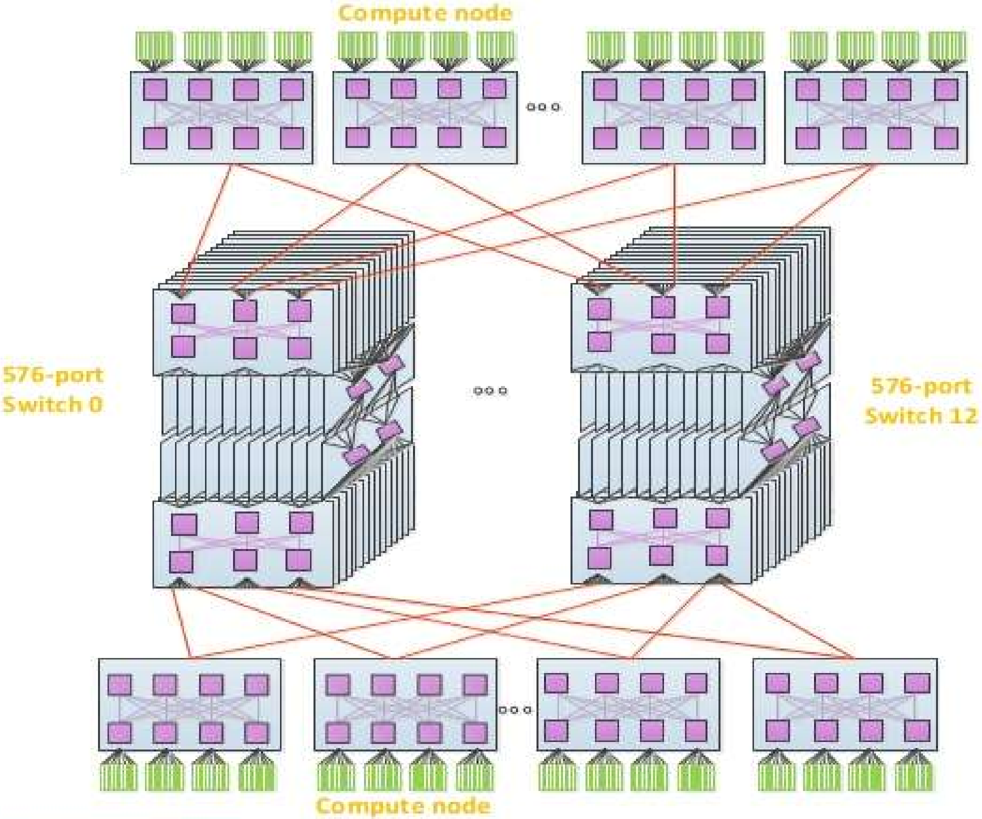
Neo-Heterogeneous Compute Node

- Similar ISA, different ALU
- 2 Intel Ivy Bridge CPU + 3 Intel Xeon Phi
- 16 Registered ECC DDR3 DIMMs, 64GB
- 3 PCI-E 3.0 with 16 lanes
- PDP Comm. Port
- Dual Gigabit LAN
- Peak Perf. : 3.432Tflops

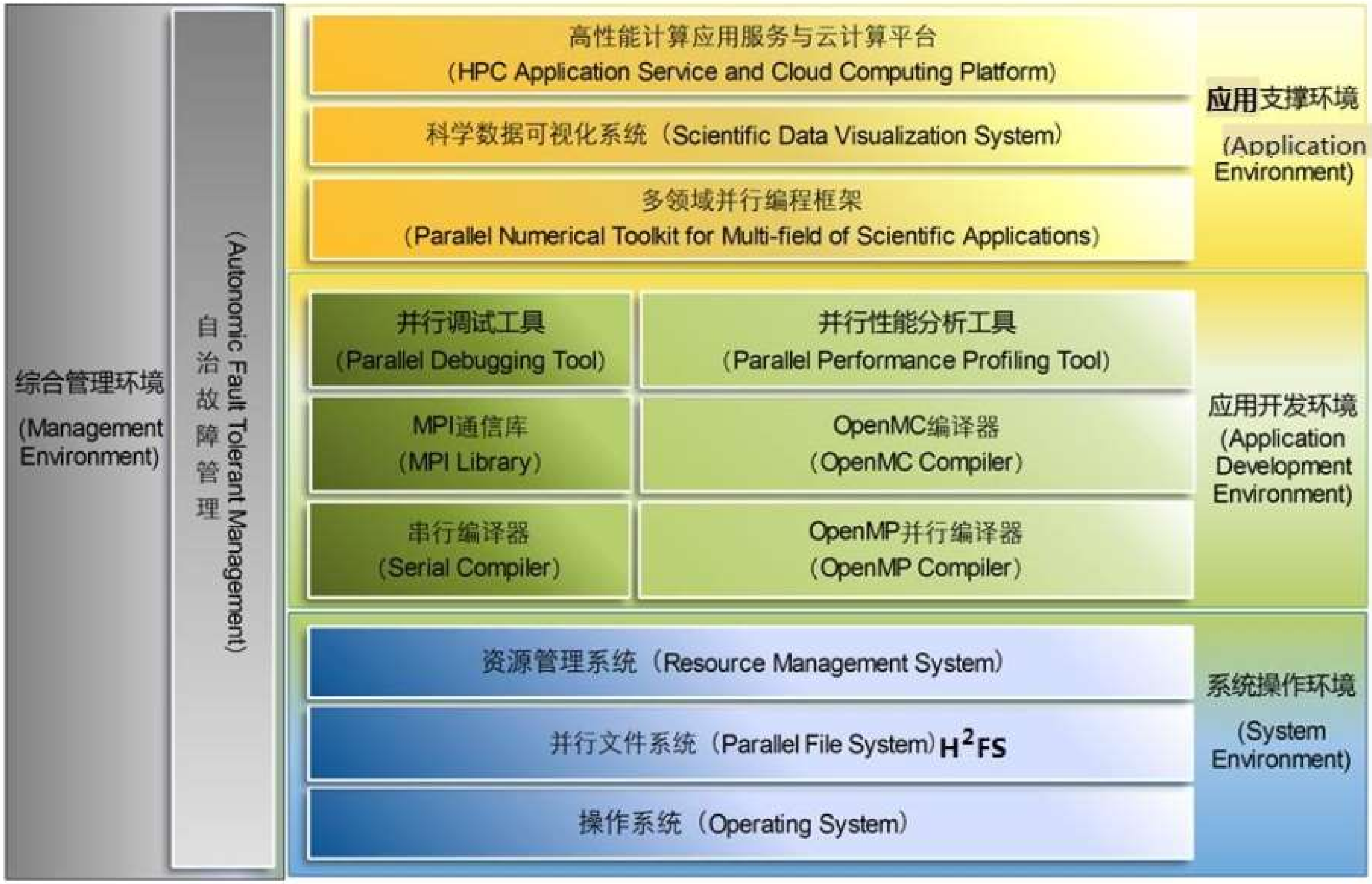


Interconnection network

- **TH Express-2 interconnection network**
  - Fat-tree topology using 13 576-port top level switches
  - Opto-electronic hybrid transport tech.
  - Proprietary network protocol
  - NRC +NIC



HPC Software stack



综合管理环境 (Management Environment)

自治故障管理 （Autonomic Fault Tolerant Management）

高性能计算应用服务与云计算平台 （HPC Application Service and Cloud Computing Platform）

科学数据可视化系统（Scientific Data Visualization System）

多领域并行编程框架 （Parallel Numerical Toolkit for Multi-field of Scientific Applications）

应用支撑环境 (Application Environment)

并行调试工具 （Parallel Debugging Tool）

并行性能分析工具 （Parallel Performance Profiling Tool）

MPI通信库 （MPI Library）

OpenMC编译器 （OpenMC Compiler）

串行编译器 （Serial Compiler）

OpenMP并行编译器 （OpenMP Compiler）

应用开发环境 (Application Development Environment)

资源管理系统（Resource Management System）

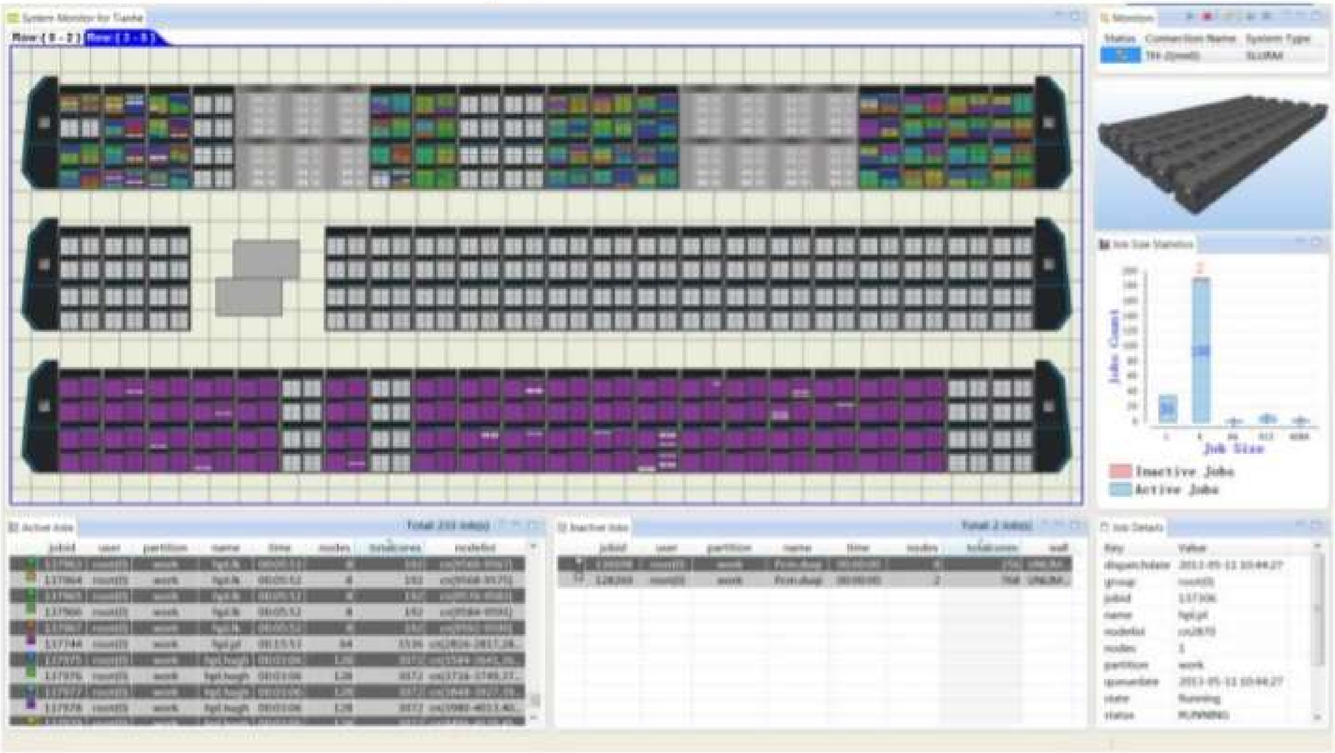
并行文件系统（Parallel File System）H^2FS

操作系统（Operating System）

系统操作环境 (System Environment)

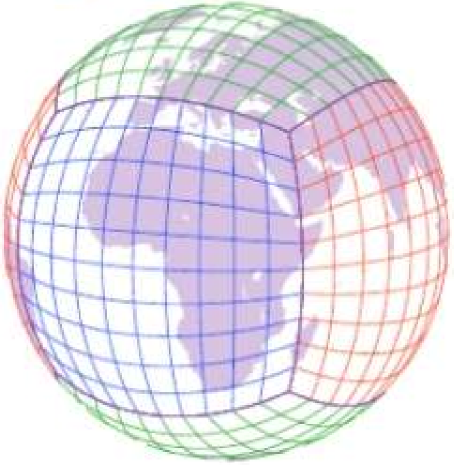
OS & RMS

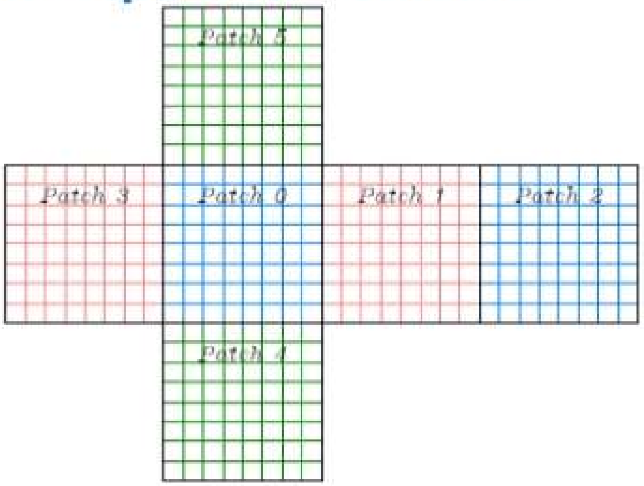
- **Operating System**
  - Kylin Linux
- **Resource manage system**
  - Power-aware resource allocation
  - Multiple custom schedule policies



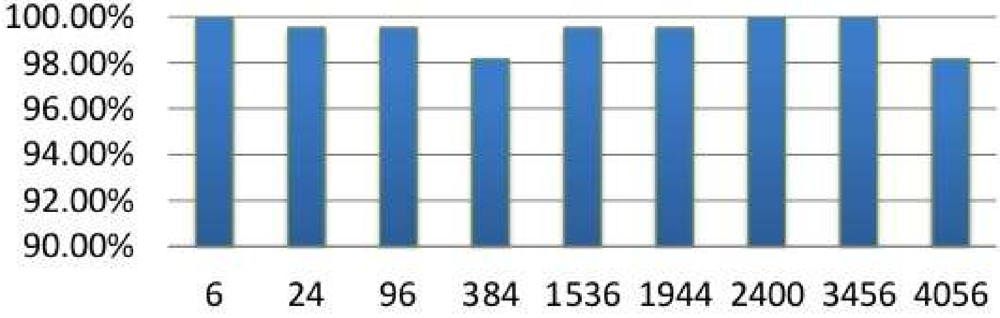
Application

- **Application of a global shallow water model: algorithms**
  - **Hierarchical data partition & communication on cubed-sphere**
  - **Balanced partition between CPU/MIC inside each node**
  - **Communication hiding algorithm based on "Pipe-flow" scheme**





- **Nearly ideal weak scaling on the Tianhe-2**
  - **Using up to 4,056 nodes (97,344 CPU cores + 693,576 MIC cores)**
  - **# of unknowns for the largest run: 200 billion**



Slides provided by Chao Yang from Institute of Software, Chinese Academy of Sciences

Course texts

- Course materials partly taken from the following texts.
  - But all topics covered by lecture slides.
- *Introduction to Parallel Computing*. Grama, Karypis, Kumar, Gupta. Pearson, 2003.
- *An Introduction to Parallel Programming*. Peter Pacheco. Morgan Kaufmann 2011.
- *Programming Massively Parallel Processors*. Kirk, Hwu. Morgan Kaufmann 2016.
- *CUDA by Example*. Sanders, Kandrot. Addison-Wesley 2010.