Introduction

CS121 Parallel Computing

Spring 2020

Administrivia

- Instructor Assoc Prof Rui FAN 范睿
  fanrui@shanghaitech.edu.cn / 131-2259-1828
- Online “flipped classroom” format
  - □ Lecture notes and videos uploaded to course Blackboard 4 days before class.
  - □ Students need to study the online lecture material before class.
  - □ Online classes held weekly on Mondays & Wednesdays at 3-4:40pm.
  - □ During class students and the lecturer will interact and discuss the lecture content.
    - Additional discussions on course Piazza.
  - □ Classes done using the teleconferencing software Zoom.

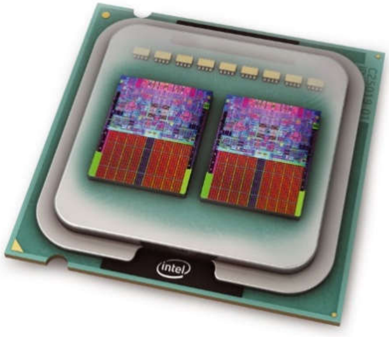
  1. Download and install Zoom from https://zoom.us/client/latest/ZoomInstaller.exe.
  2. At the start of class, log into Zoom and join meeting ID 364-929-4960.
  3. Students will see live content from the instructor’s computer.
- TA Wang Leshan 王乐山
  wangleshan1996@outlook.com
- Recitation Format and time TBD.

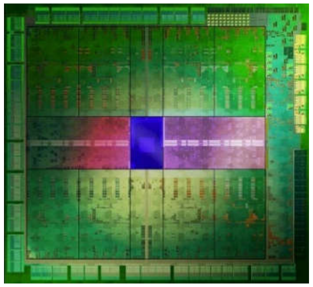
Course assessment

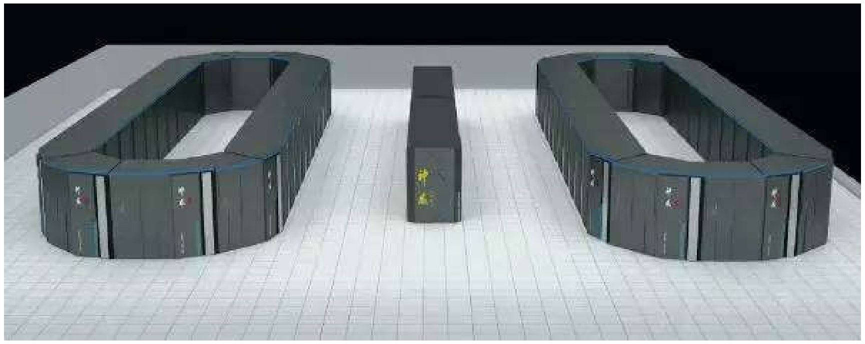
Problem sets	20%	
Labs	20%	▪ Solve problems using OpenMP and CUDA
Reading project	20% Teams of 2	▪ Find an interesting research paper from the suggested reading list. ▪ Tell me your paper by week 8 (April 24). ▪ Submit a report and 20 minute video about the paper by end of week 16 (June 19).
Course project	20% Teams of 2	▪ Find an interesting problem and write an efficient parallel program for it. ▪ Tell me your problem by week 10 (May 8). ▪ Submit a report and 20 minute video about your project by July 3.
Final exam	20%	
Class participation	May move borderline grades up half a letter grade (e.g. B+ to A-)	Actively participate in online class discussions.

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 minining, machine learning.





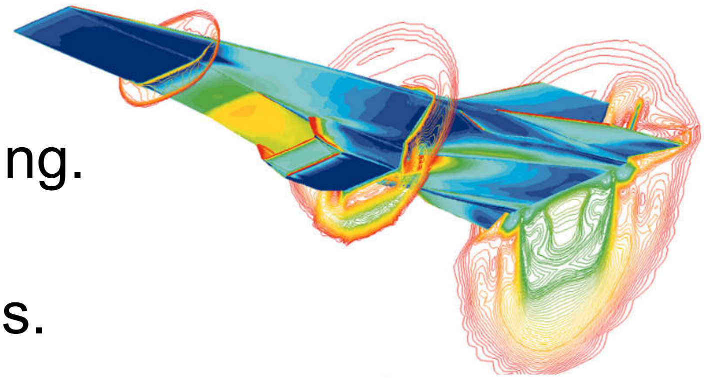


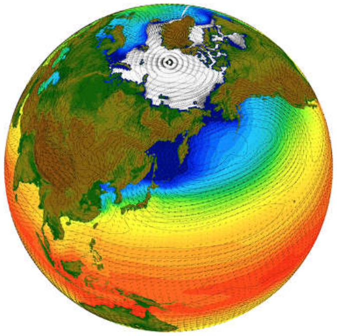
Course objectives

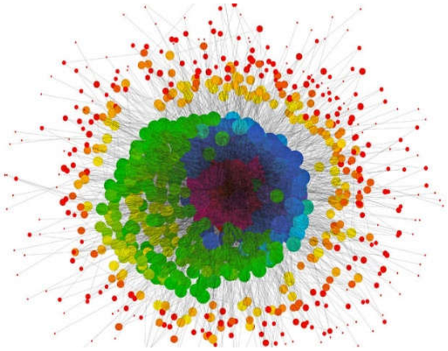
- To understand the concepts and techniques of parallel computing, and take advantage of the capabilities of modern systems.
  - Parallel hardware models and interaction with parallel software.
  - Power and limitations of parallelism.
  - Efficient parallel algorithms for important problems.

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.

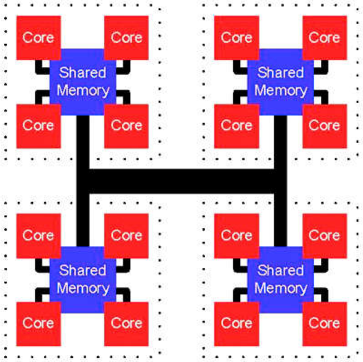


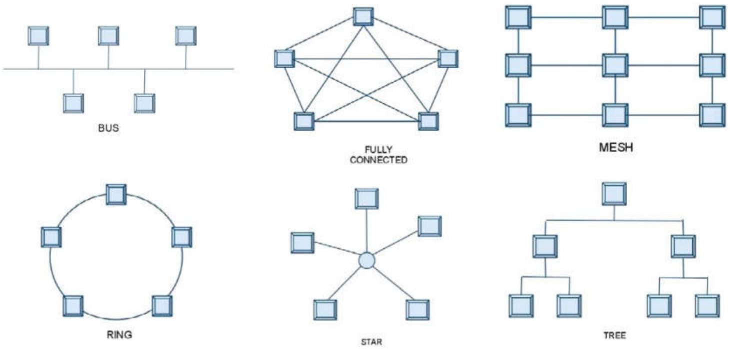


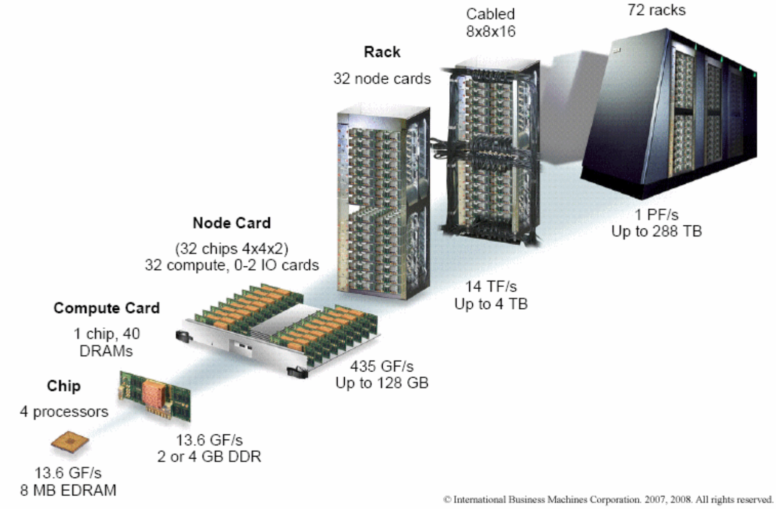


Parallel hardware

- 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.



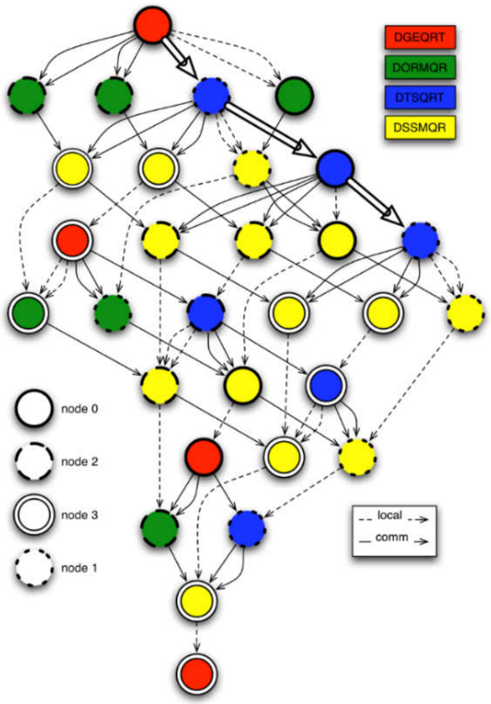




© International Business Machines Corporation. 2007, 2008. All rights reserved.

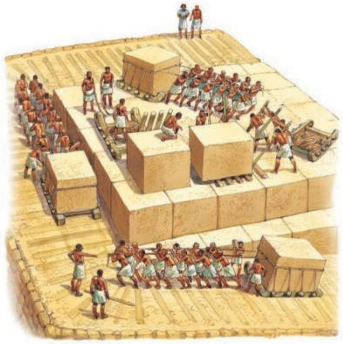
Parallel software

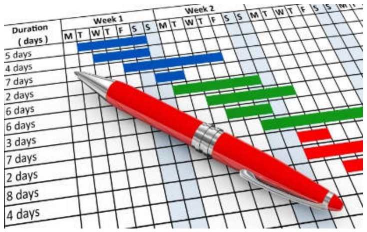
- Break a large problem into subproblems (tasks) that can be solved (somewhat) independently.
- OS and scheduler allocate tasks to different processors.
  - Respect dependencies between tasks.
- Parallel software must be matched to the hardware.
  - Similar amounts of concurrency in software and hardware.
  - 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.



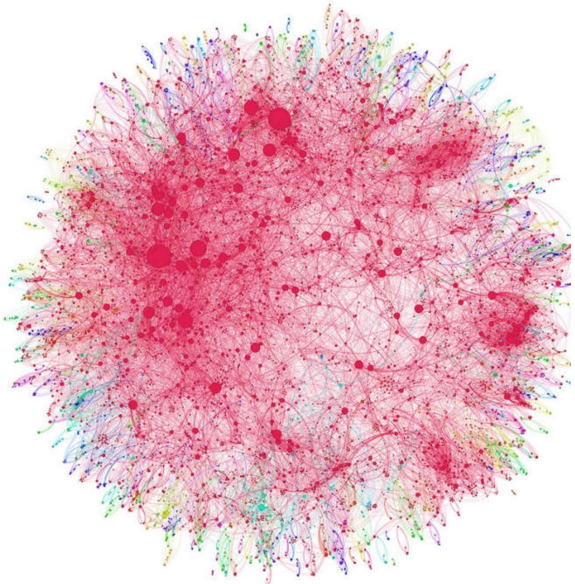
Challenges

- 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 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

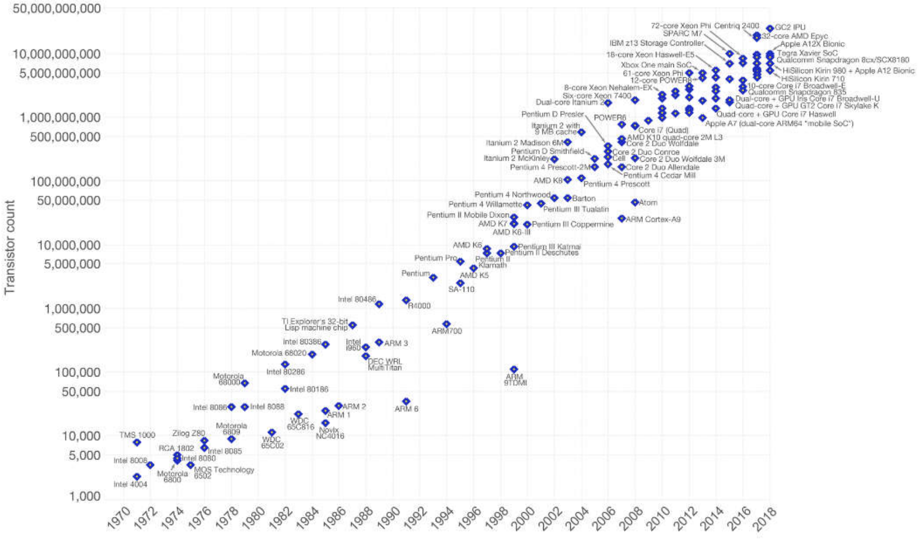
- Research and theory started in the early 60’s.
  - Cray-1 reached 160 MFLOPS in 1976.
- Commercially successful supercomputers (Cray, Thinking Machines, etc.) started in 1980’s.
  - Used expensive custom processors.
- In 1990’s massively parallel processors (MPPs) and clusters became dominant.
  - 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 IBM Summit MPP.
  - Runs at 148 PFLOPS, about 1M 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 to taking advantage of Moore’s Law.

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

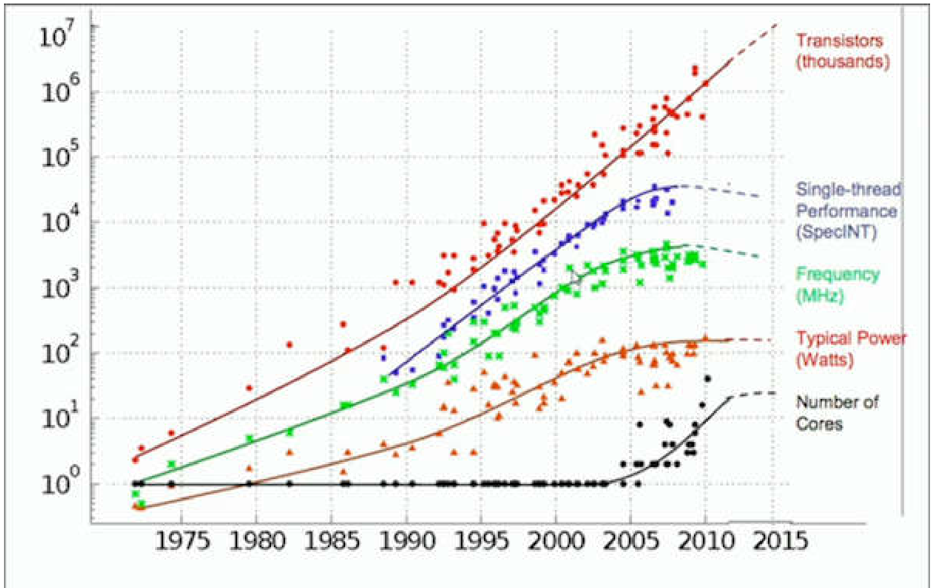
Moore’s law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore’s law.



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.

Moore’s Law and performance

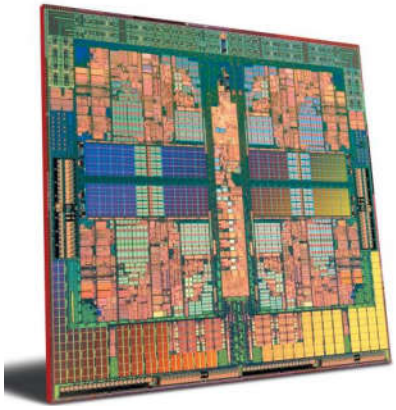
- Transistor properties, e.g. size and clock speed, do not scale equally.
- Higher single processor clock speeds is increasingly difficult to achieve.
  - □ Heat
  - □ Power consumption
  - □ Current leakage

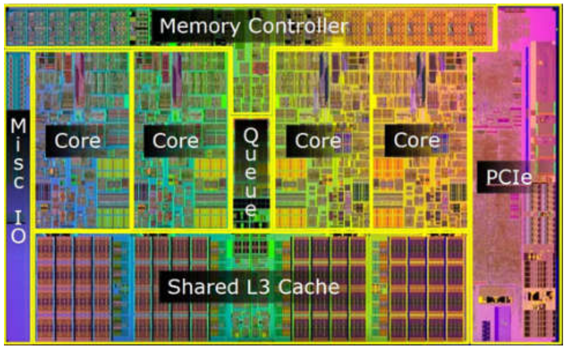


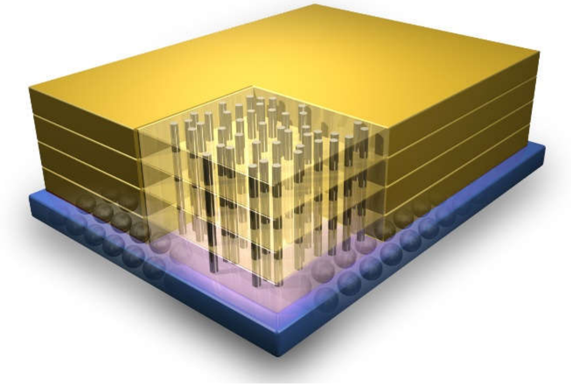
Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

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.





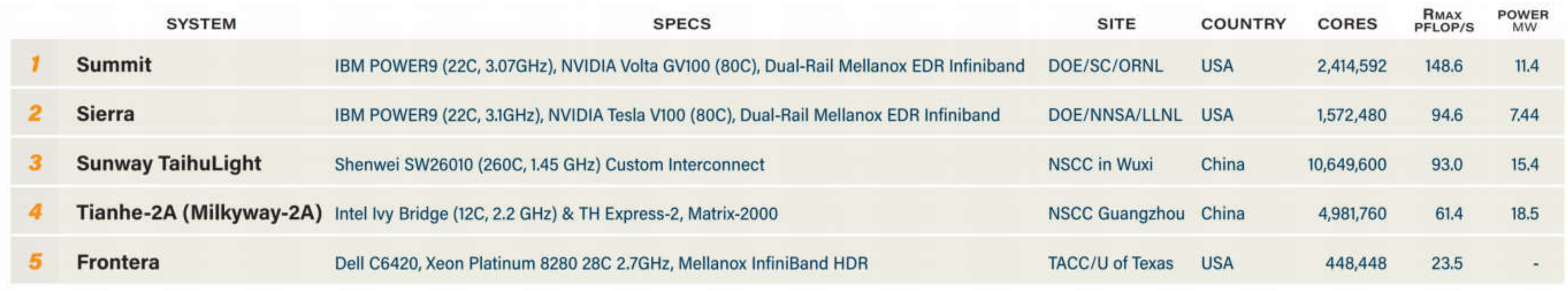


The state of the art

- Parallel computers today mainly based on four processor architectures.
  - Multicores
    - Small / moderate number (≤ 64) of fast, general purpose cores.
    - Ex Intel Xeon, IBM Power, Sun SPARC.
  - Manycores
    - Large number (1000's) of simple cores.
    - Ex Nvidia Pascal GPU, Intel Xeon Phi, Sunway SW26010.
  - 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, Apple Neural Engine, 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.
  - Goal is a supercomputer achieving 50 GFLOPS / W.
    - Current supercomputers achieve 1-6 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.
  - Latest list from November 2019.

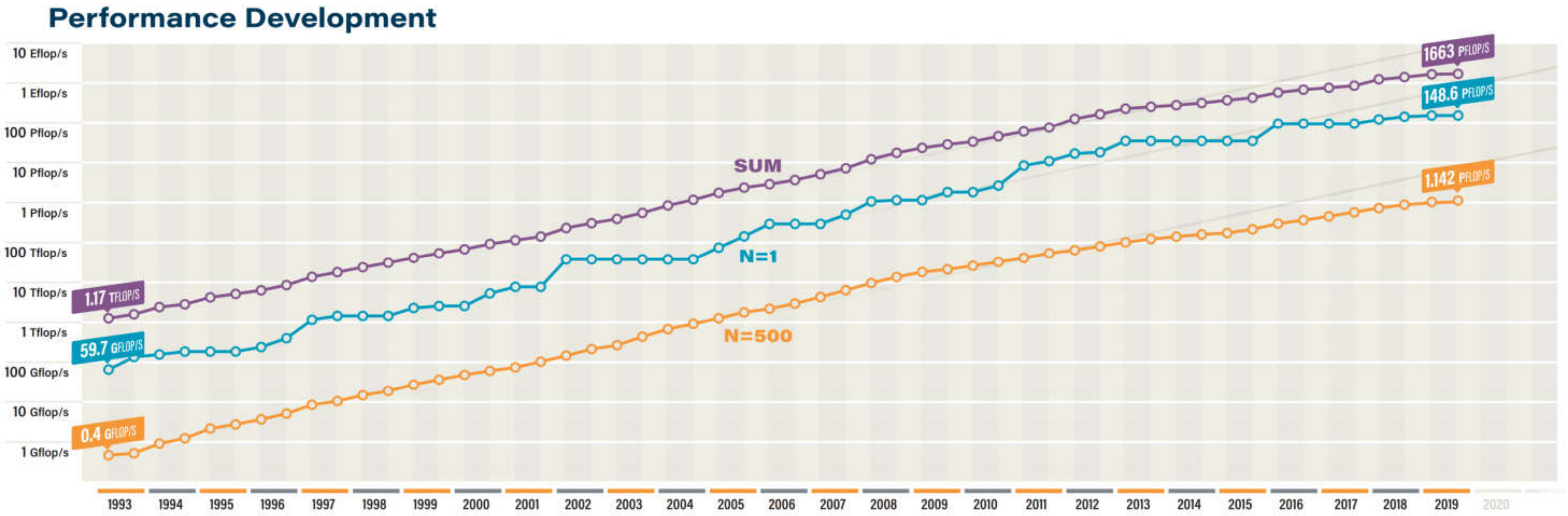


	SYSTEM	SPECS	SITE	COUNTRY	CORES	RMAX PFLOP/S	POWER MW
1	Summit	IBM POWER9 (22C, 3.07GHz), NVIDIA Volta GV100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/SC/ORNL	USA	2,414,592	148.6	11.4
2	Sierra	IBM POWER9 (22C, 3.1GHz), NVIDIA Tesla V100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/NNSA/LLNL	USA	1,572,480	94.6	7.44
3	Sunway TaihuLight	Shenwei SW26010 (260C, 1.45 GHz) Custom Interconnect	NSCC in Wuxi	China	10,649,600	93.0	15.4
4	Tianhe-2A (Milkyway-2A)	Intel Ivy Bridge (12C, 2.2 GHz) & TH Express-2, Matrix-2000	NSCC Guangzhou	China	4,981,760	61.4	18.5
5	Frontera	Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR	TACC/U of Texas	USA	448,448	23.5	-

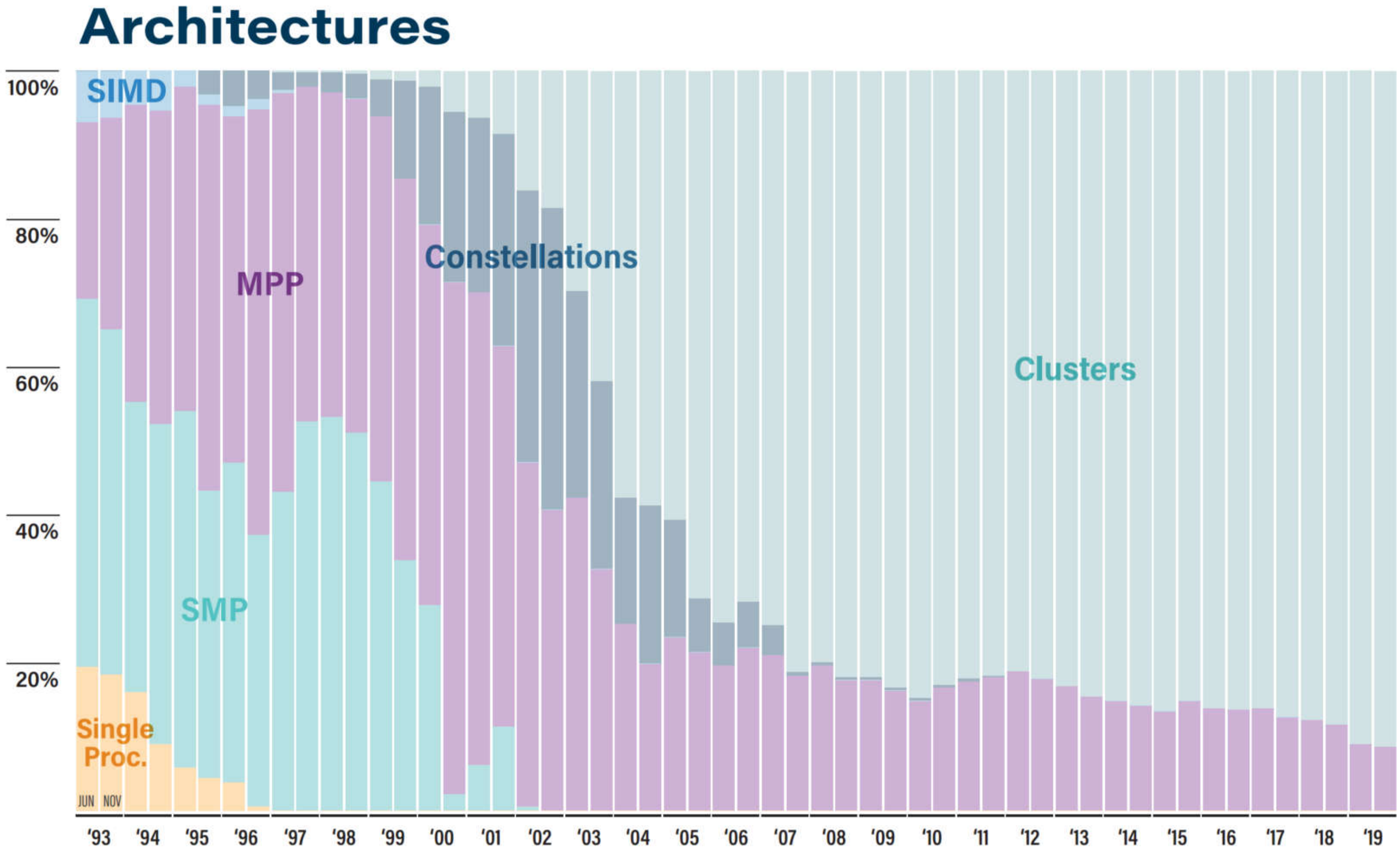
Mega	Giga	Tera	Peta	Exa
10^6	10^9	10^{12}	10^{15}	10^{18}

- For comparison, Intel multicore achieves ~50 GFLOPS / core, and GPU achieves ~ 10 TFLOPS / board.

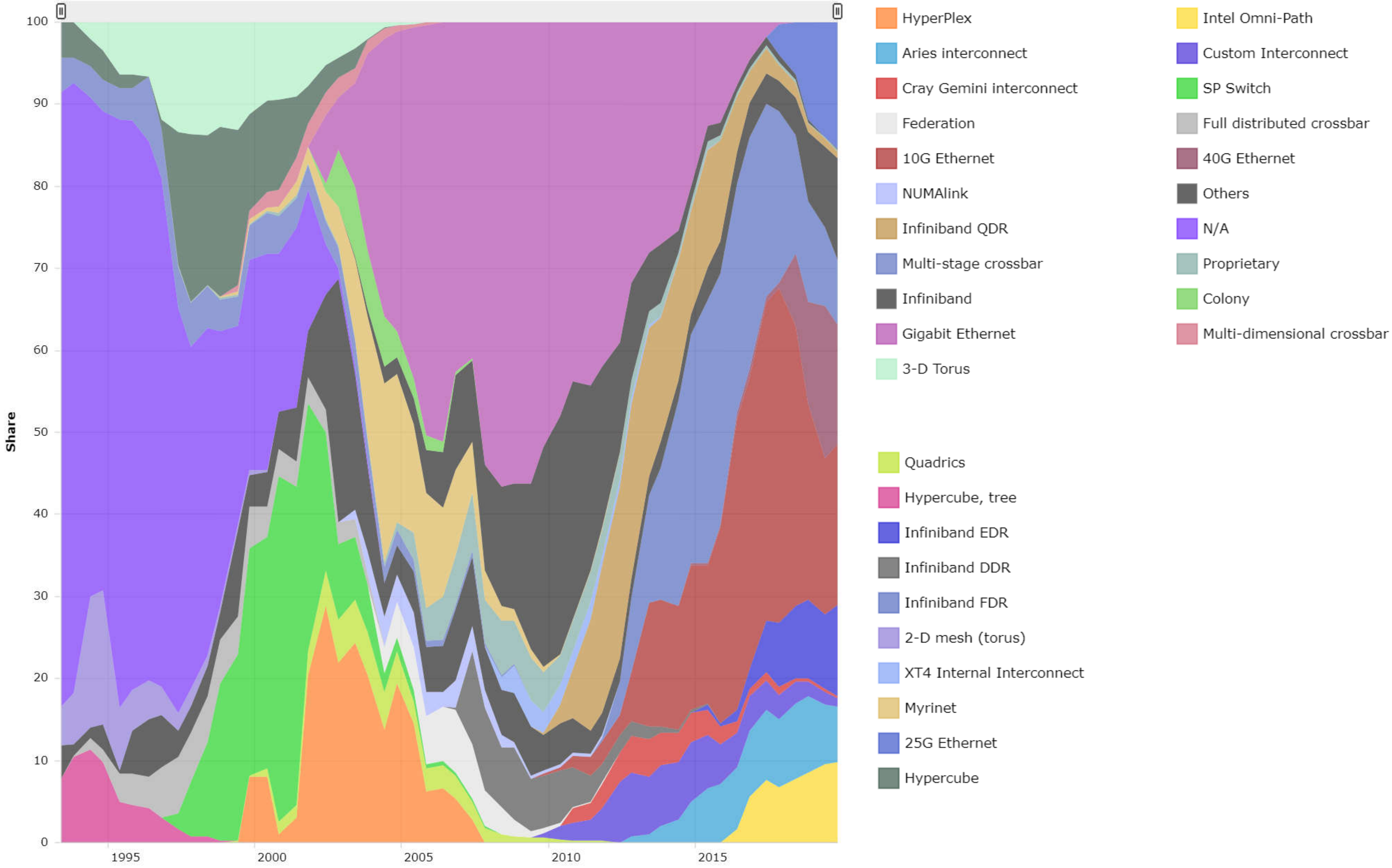
Top 500 – Trends



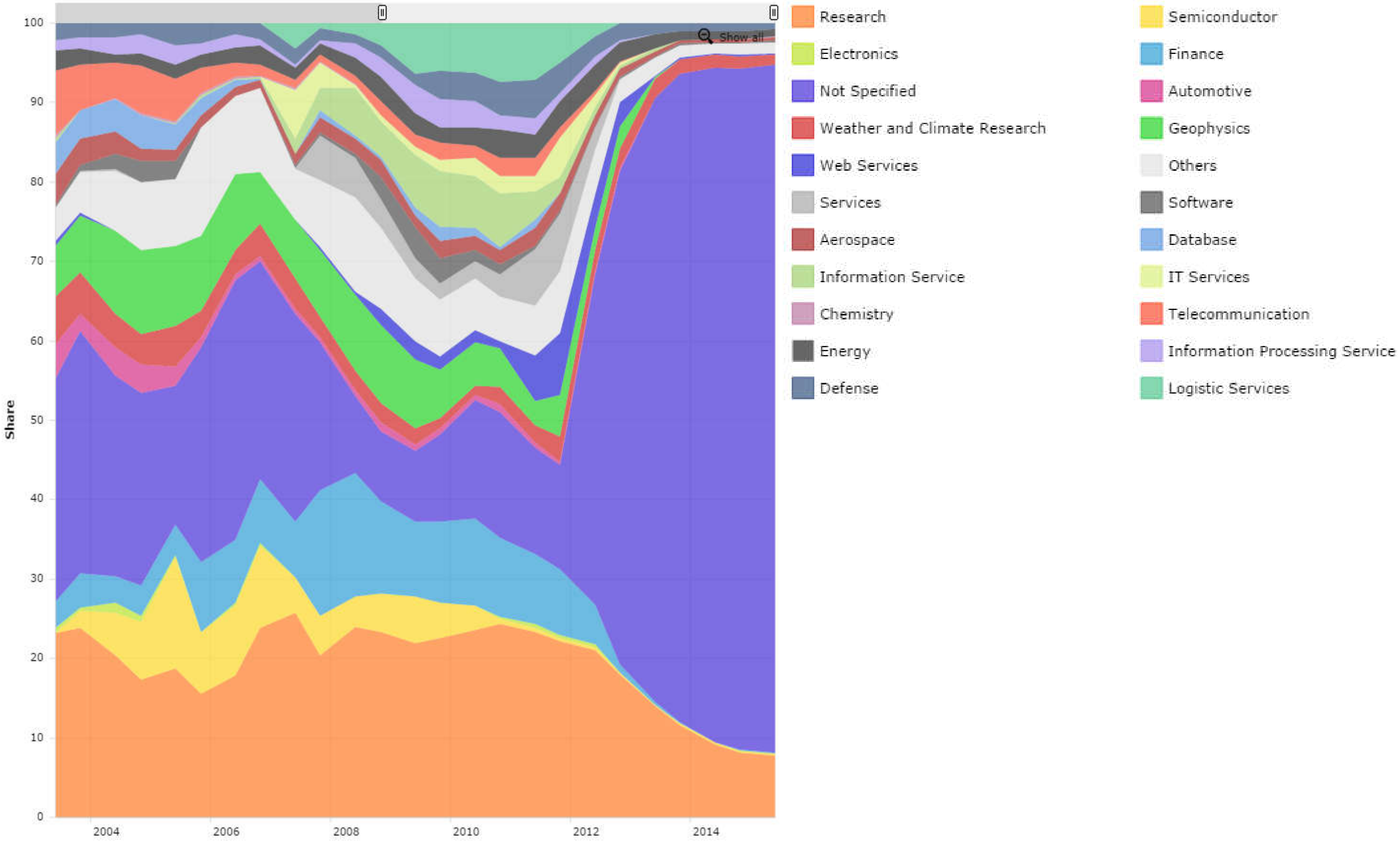
Top 500 – Architecture



Top 500 – Interconnect



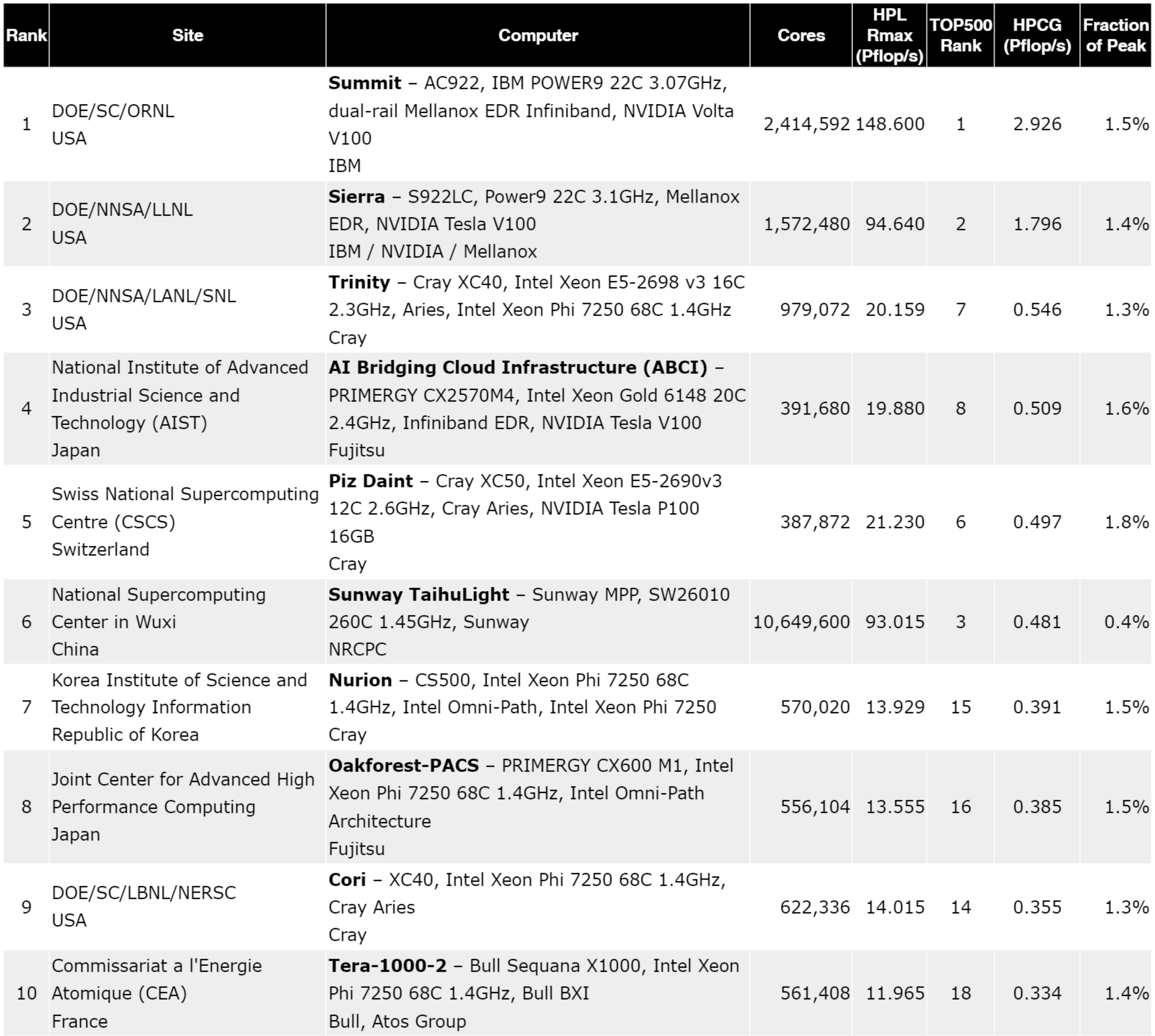
Top 500 – Applications



Other performance measures

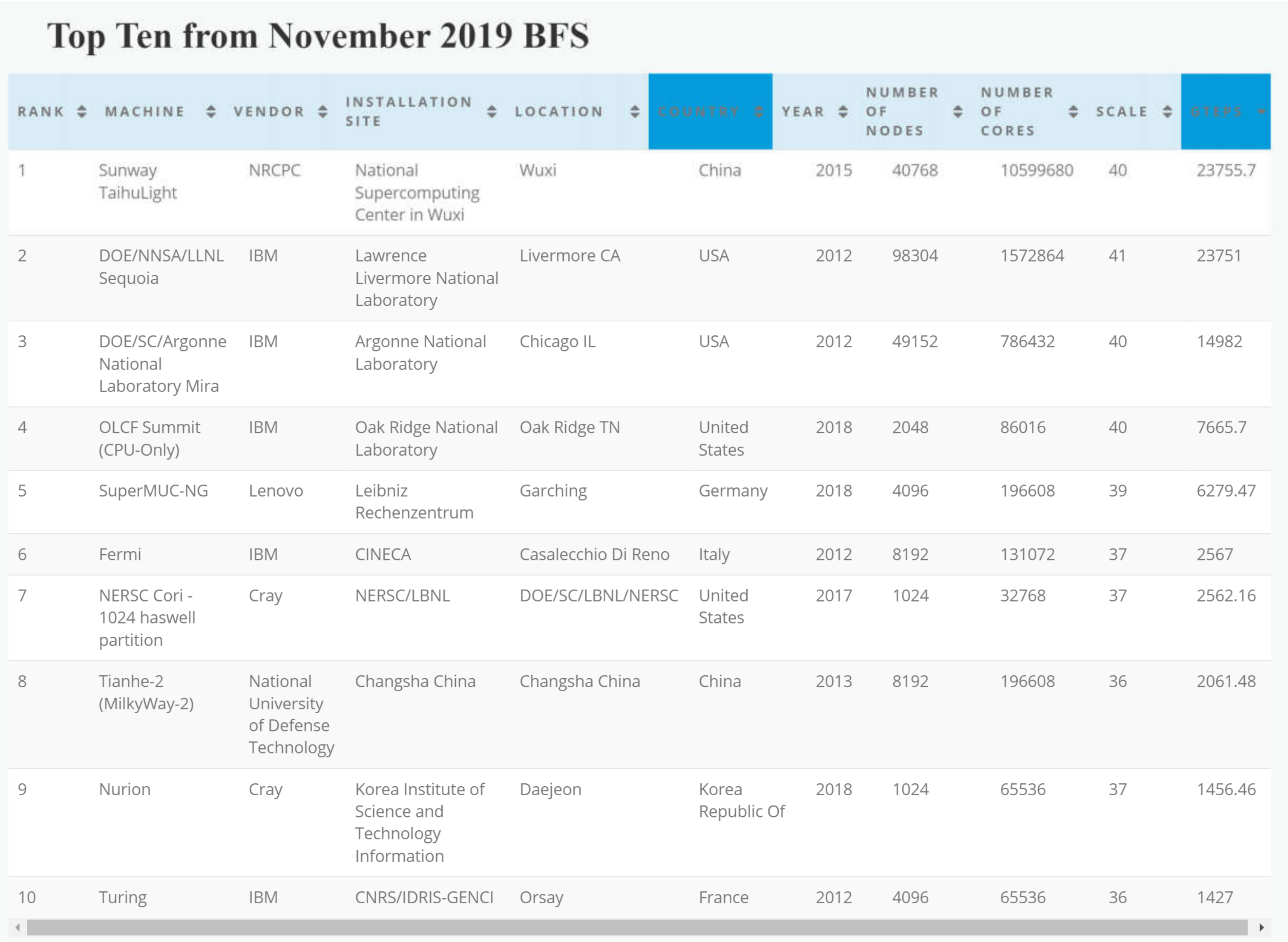
- LINPACK does compute-intensive operations on structured dense matrices.
  - Uniform control flow, predictable and coalesced memory accesses.
  - Ideal for physical simulations.
- Data-intensive applications today have instruction divergence, branching and random memory accesses.
- New benchmarks give 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



Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	DOE/SC/ORNL USA	**Summit** – AC922, IBM POWER9 22C 3.07GHz, dual-rail Mellanox EDR Infiniband, NVIDIA Volta V100 IBM	2,414,592	148.600	1	2.926	1.5%
2	DOE/NNSA/LLNL USA	**Sierra** – S922LC, Power9 22C 3.1GHz, Mellanox EDR, NVIDIA Tesla V100 IBM / NVIDIA / Mellanox	1,572,480	94.640	2	1.796	1.4%
3	DOE/NNSA/LANL/SNL USA	**Trinity** – Cray XC40, Intel Xeon E5-2698 v3 16C 2.3GHz, Aries, Intel Xeon Phi 7250 68C 1.4GHz Cray	979,072	20.159	7	0.546	1.3%
4	National Institute of Advanced Industrial Science and Technology (AIST) Japan	**AI Bridging Cloud Infrastructure (ABCI)** – PRIMERGY CX2570M4, Intel Xeon Gold 6148 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 Fujitsu	391,680	19.880	8	0.509	1.6%
5	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** – Cray XC50, Intel Xeon E5-2690v3 12C 2.6GHz, Cray Aries, NVIDIA Tesla P100 16GB Cray	387,872	21.230	6	0.497	1.8%
6	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** – Sunway MPP, SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93.015	3	0.481	0.4%
7	Korea Institute of Science and Technology Information Republic of Korea	**Nurion** – CS500, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path, Intel Xeon Phi 7250 Cray	570,020	13.929	15	0.391	1.5%
8	Joint Center for Advanced High Performance Computing Japan	**Oakforest-PACS** – PRIMERGY CX600 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Architecture Fujitsu	556,104	13.555	16	0.385	1.5%
9	DOE/SC/LBNL/NERSC USA	**Cori** – XC40, Intel Xeon Phi 7250 68C 1.4GHz, Cray Aries Cray	622,336	14.015	14	0.355	1.3%
10	Commissariat a l'Energie Atomique (CEA) France	**Tera-1000-2** – Bull Sequana X1000, Intel Xeon Phi 7250 68C 1.4GHz, Bull BXI Bull, Atos Group	561,408	11.965	18	0.334	1.4%

Graph500



Top Ten from November 2019 BFS

RANK	MACHINE	VENDOR	INSTALLATION SITE	LOCATION	COUNTRY	YEAR	NUMBER OF NODES	NUMBER OF CORES	SCALE	GTEPS
1	Sunway TaihuLight	NRCPC	National Supercomputing Center in Wuxi	Wuxi	China	2015	40768	10599680	40	23755.7
2	DOE/NNSA/LLNL Sequoia	IBM	Lawrence Livermore National Laboratory	Livermore CA	USA	2012	98304	1572864	41	23751
3	DOE/SC/Argonne National Laboratory Mira	IBM	Argonne National Laboratory	Chicago IL	USA	2012	49152	786432	40	14982
4	OLCF Summit (CPU-Only)	IBM	Oak Ridge National Laboratory	Oak Ridge TN	United States	2018	2048	86016	40	7665.7
5	SuperMUC-NG	Lenovo	Leibniz Rechenzentrum	Garching	Germany	2018	4096	196608	39	6279.47
6	Fermi	IBM	CINECA	Casalecchio Di Reno	Italy	2012	8192	131072	37	2567
7	NERSC Cori - 1024 haswell partition	Cray	NERSC/LBNL	DOE/SC/LBNL/NERSC	United States	2017	1024	32768	37	2562.16
8	Tianhe-2 (MilkyWay-2)	National University of Defense Technology	Changsha China	Changsha China	China	2013	8192	196608	36	2061.48
9	Nurion	Cray	Korea Institute of Science and Technology Information	Daejeon	Korea Republic Of	2018	1024	65536	37	1456.46
10	Turing	IBM	CNRS/IDRIS-GENCI	Orsay	France	2012	4096	65536	36	1427

HPCL State Key Laboratory of High Performance Computing

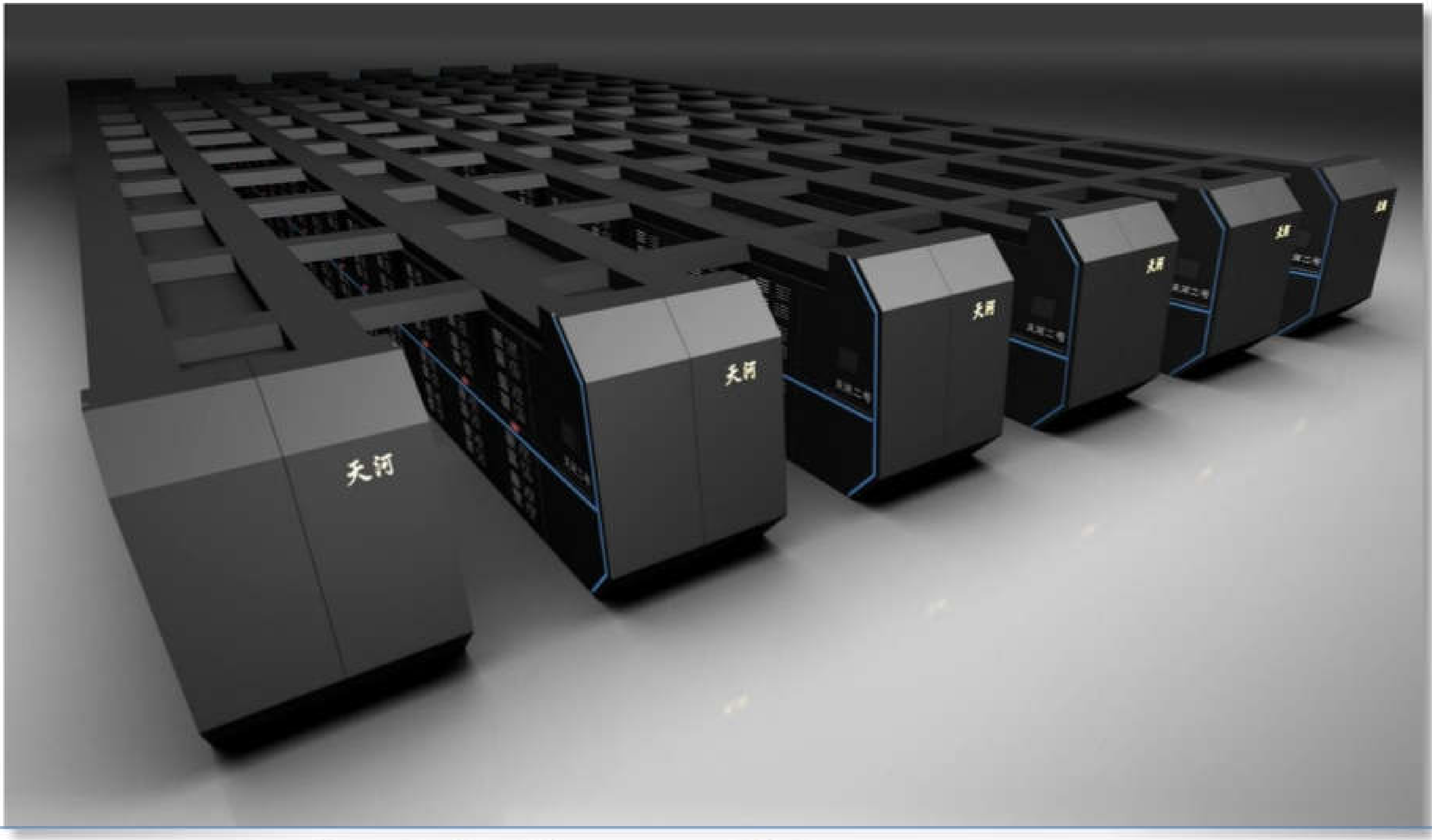
Overview of Tianhe-2 (MilkyWay-2) Supercomputer

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

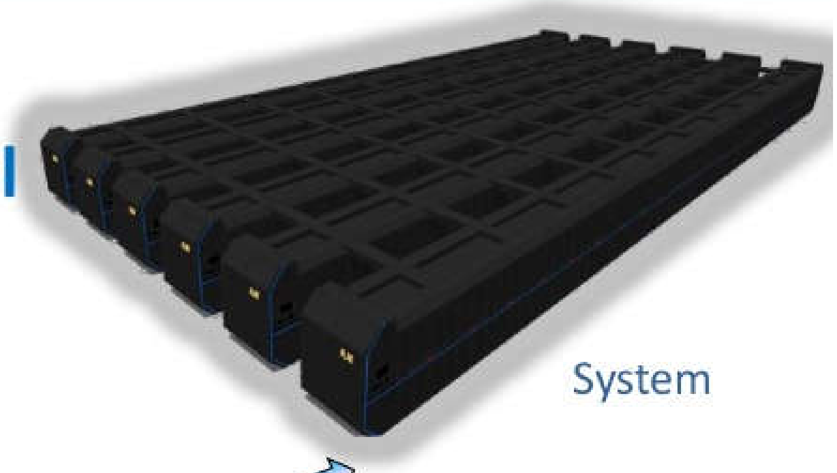
Specification

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

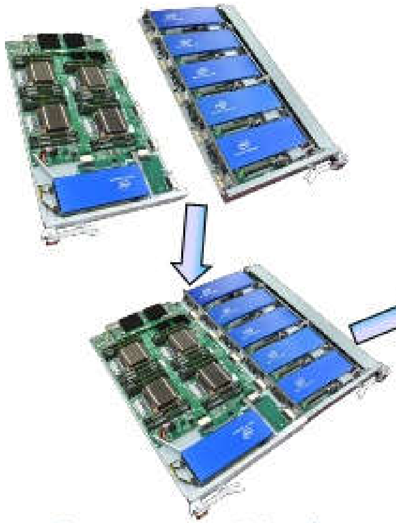
Items	Configuration
Processors	32000 Intel Xeon CPUs + 48000 Xeon Phis + 4096 FT CPUs Peak performance is 54.9PFlops, HPL
Interconnect	Proprietary high-speed interconnection network TH Express-2
Memory	1.4PB in total
Storage	Global shared parallel storage system, 12.4PB
Cabinets	125+13+24=162 compute/communication/storage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

From Chips to Entire System

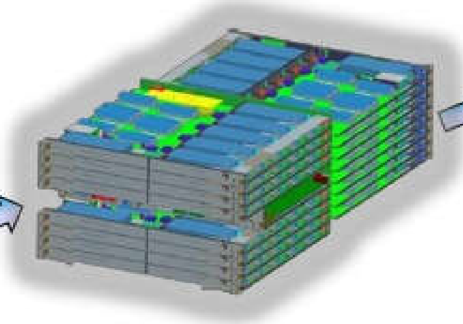
- ◆16000 compute nodes in total
- ◆Frame: 32 compute Nodes
- ◆Rack: 4 Compute Frames
- ◆Whole System: 125 Racks



System



Compute Node

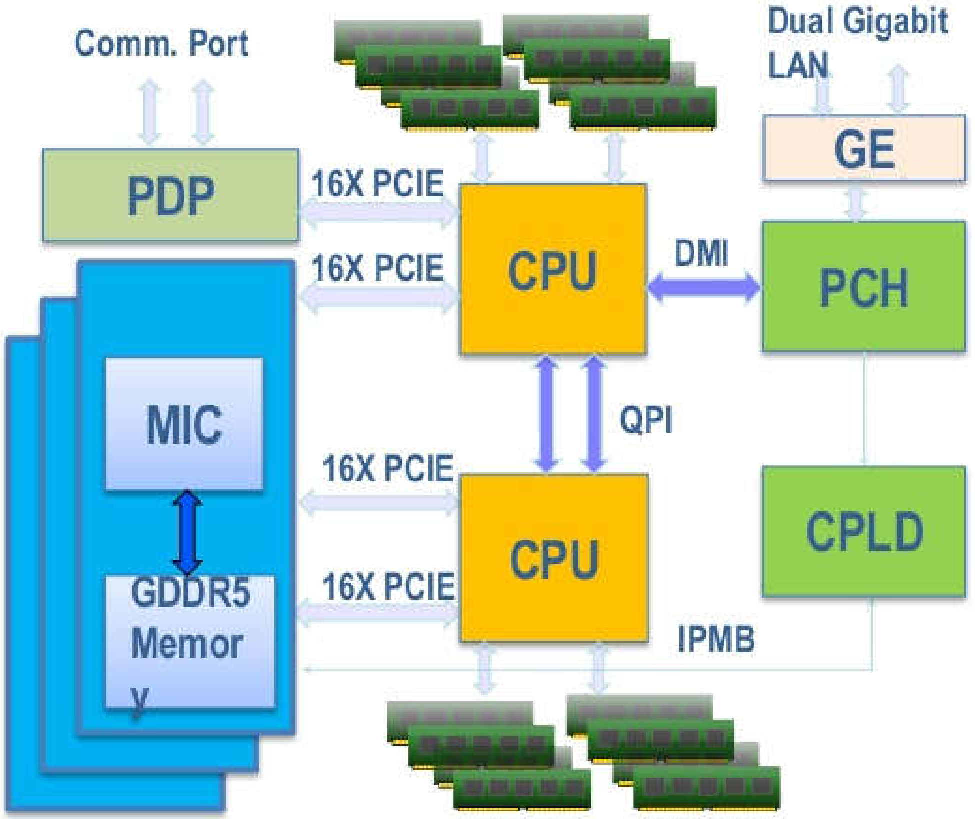


Compute Frame

Compute Rack

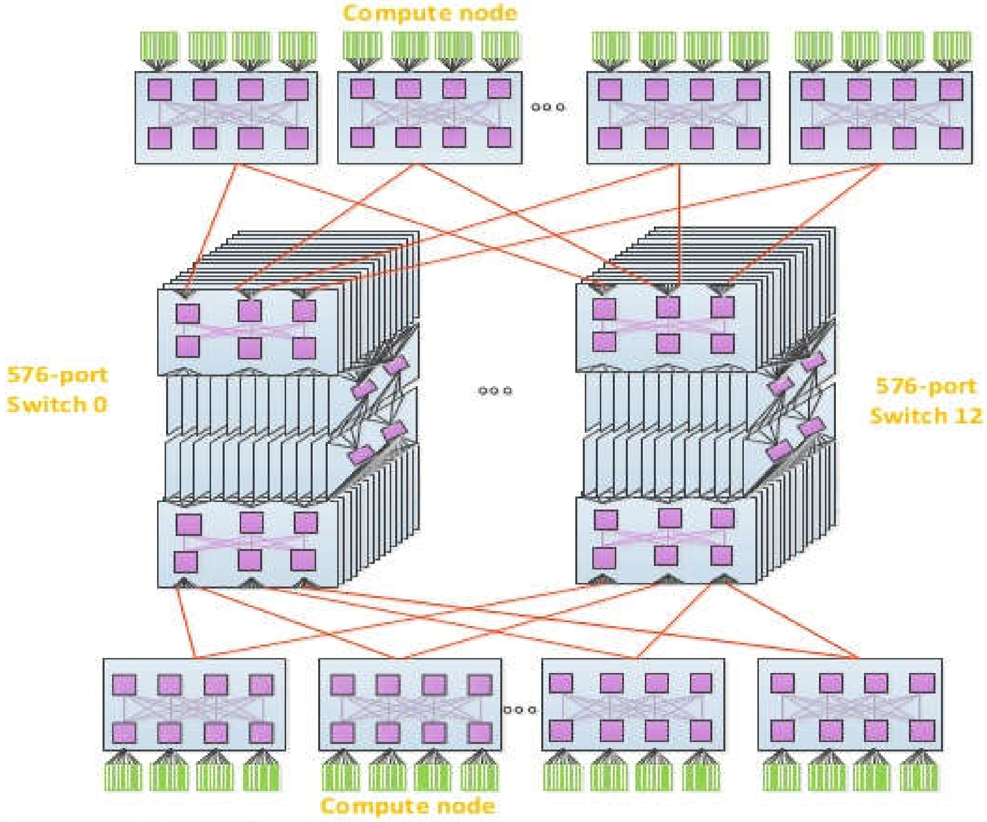
Compute Node

- 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



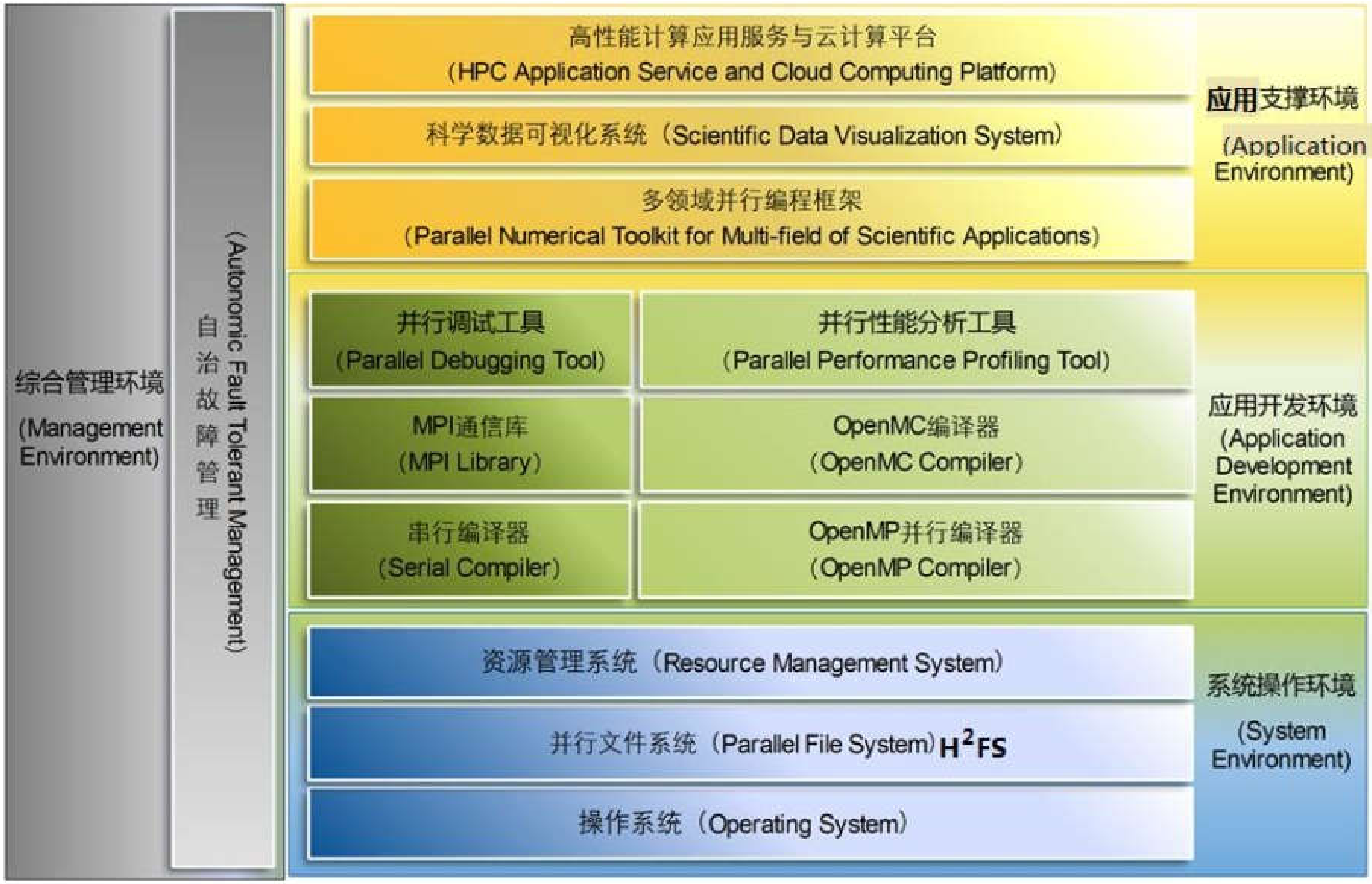
Interconnection network

- **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



HPCL

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）

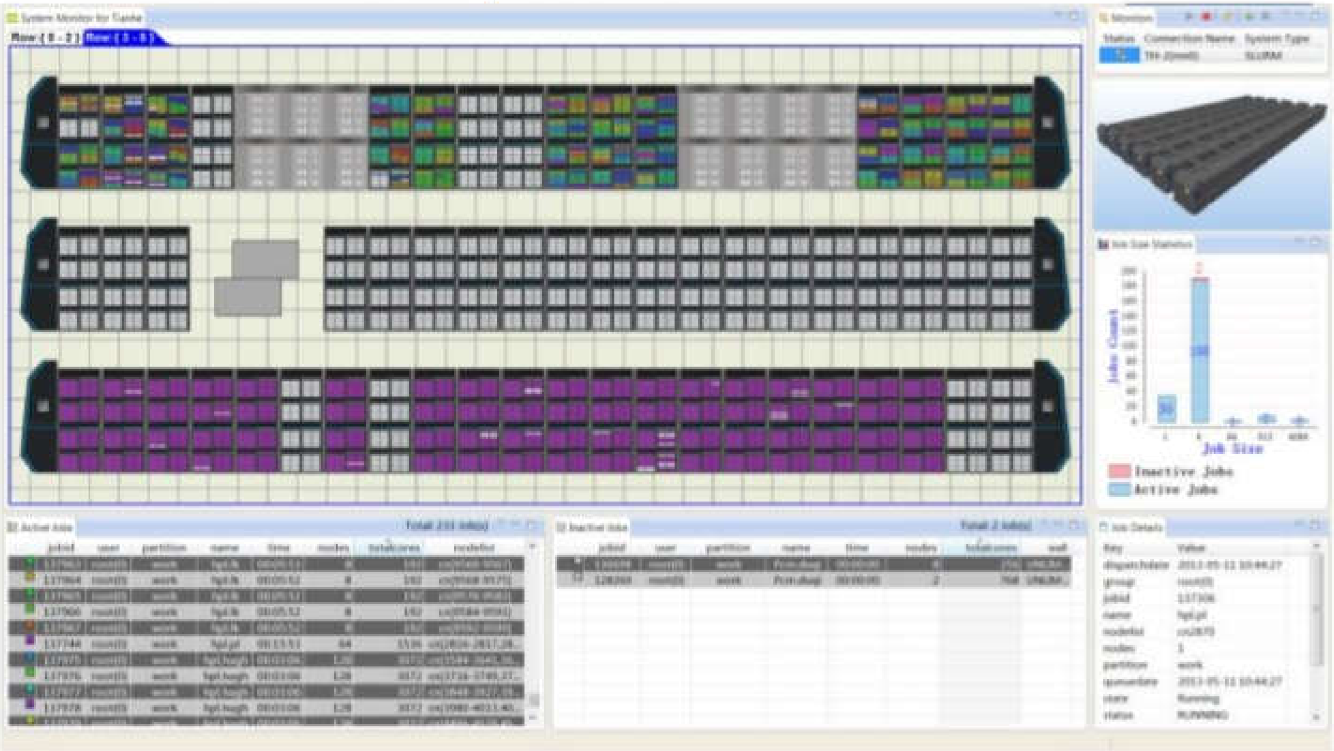
并行文件系统（Parallel File System）H^2FS

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**



Active Jobs — Total 233 Job(s)

	jobid	user	partition	name	time	nodes	totalcores	nodelist
	137963	root(0)	work	hpl.lk	00:05:12	8	192	[illegible]
	137964	root(0)	work	hpl.lk	00:05:12	8	192	cn[9568-9575]
	137965	root(0)	work	hpl.lk	00:05:12	8	192	[illegible]
	137966	root(0)	work	hpl.lk	00:05:12	8	192	cn[9584-9591]
	137967	root(0)	work	hpl.lk	00:05:12	8	192	[illegible]
	137744	root(0)	work	hpl.pl	00:13:51	64	1536	cn[2816-2817,28...
	137975	root(0)	work	hpl.hugh	00:03:06	128	3072	[illegible]
	137976	root(0)	work	hpl.hugh	00:03:06	128	3072	cn[3716-3749,27...
	137977	root(0)	work	hpl.hugh	00:03:06	128	3072	[illegible]
	137978	root(0)	work	hpl.hugh	00:03:06	128	3072	cn[3980-4013,40...

Inactive Jobs — Total 2 Job(s)

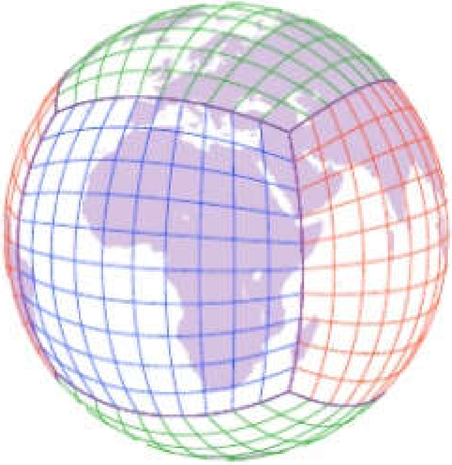
jobid	user	partition	name	time	nodes	totalcores	wall
[illegible]	root(0)	work	Pcm.dual	00:00:00	[illegible]	[illegible]	UNLIM...
[illegible]	root(0)	work	Pcm.dual	00:00:00	2	768	UNLIM...

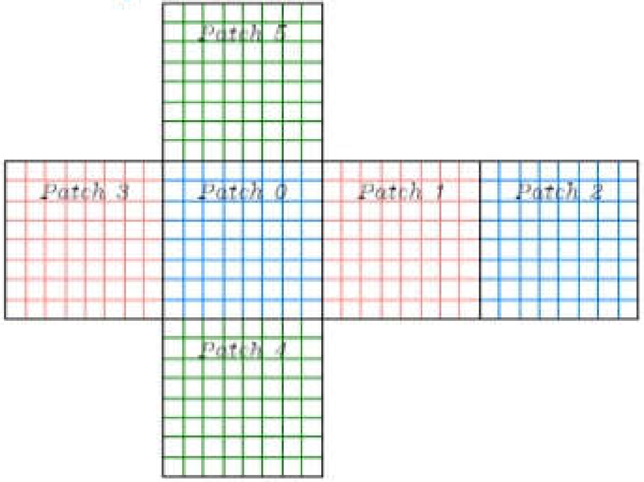
Job Details

Key	Value
dispatchdate	2013-05-11 10:44:27
group	root(0)
jobid	137306
name	hpl.pl
nodelist	cn2873
nodes	1
partition	work
queuedate	2013-05-11 10:44:27
state	Running
status	RUNNING

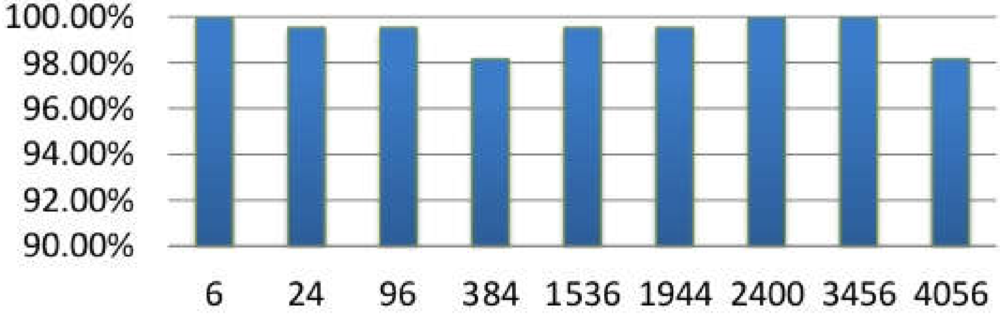
Application

- **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

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

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.