The Global Race for Exascale High Performance Computing

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Outline

- Overview of High Performance Computing
- Directions for the Future

State of Supercomputing in 2017

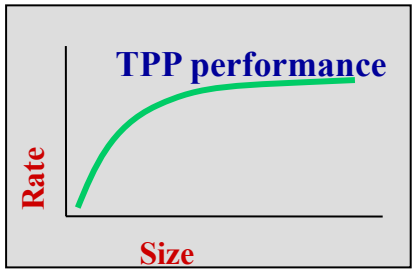
- Pflops (> 10^{15} Flop/s) computing fully established with 117 computer systems.
- Three technology architecture or "swim lanes" are thriving.
  - Commodity (e.g. Intel)
  - Commodity + accelerator (e.g. GPUs) (88 systems)
  - Lightweight cores (e.g. IBM BG, ARM, Intel's Knights Landing, ShenWei 26010, PEZY SC2)
- Interest in supercomputing is now worldwide, and growing in many new markets (~50% of Top500 computers are in industry).
- Exascale (10^{18} Flop/s) projects exist in many countries and regions.
- Intel processors largest share, 92%, followed by AMD, 1%.

The Top500 List

H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

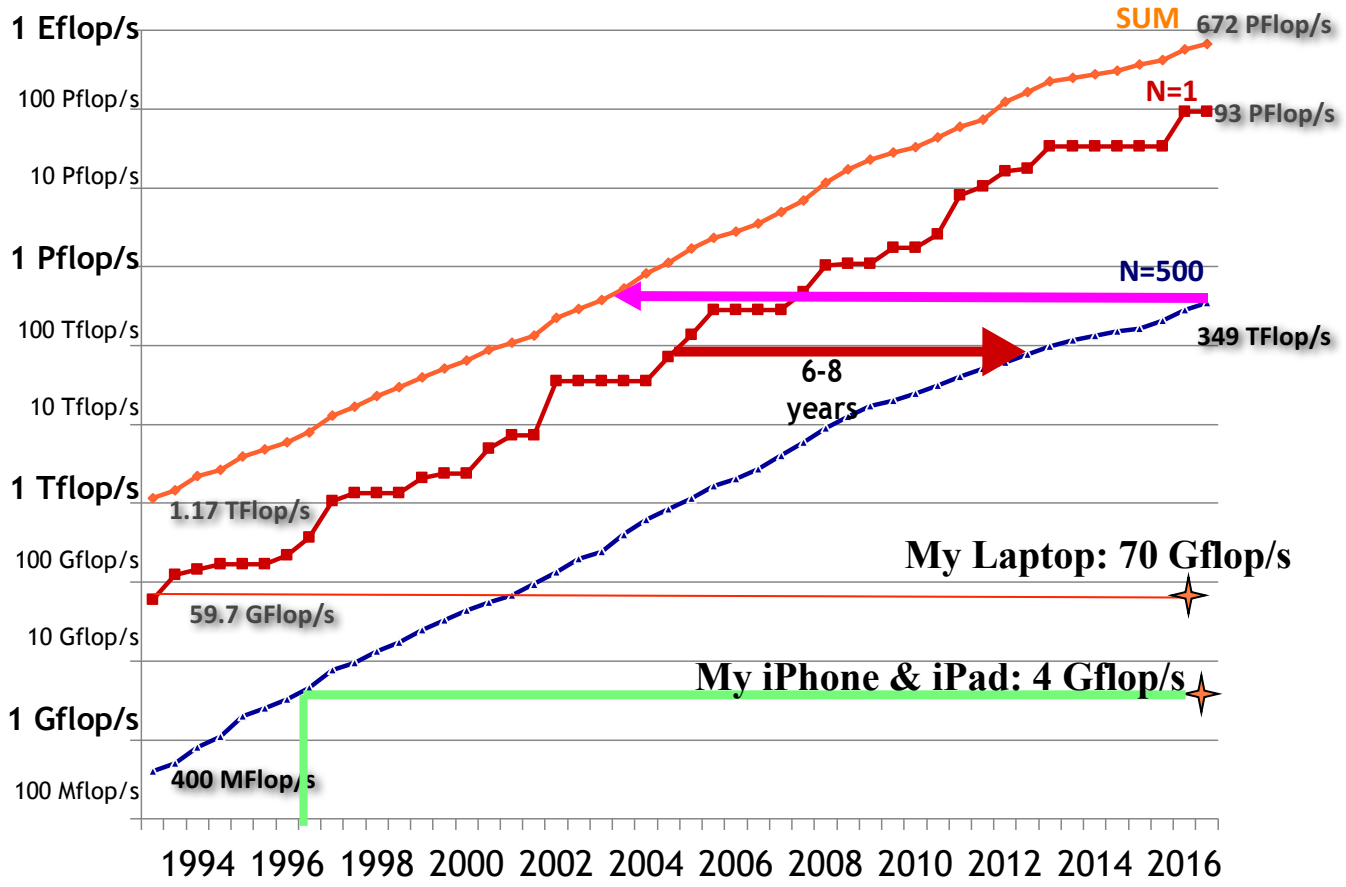
$Ax=b$, *dense problem*



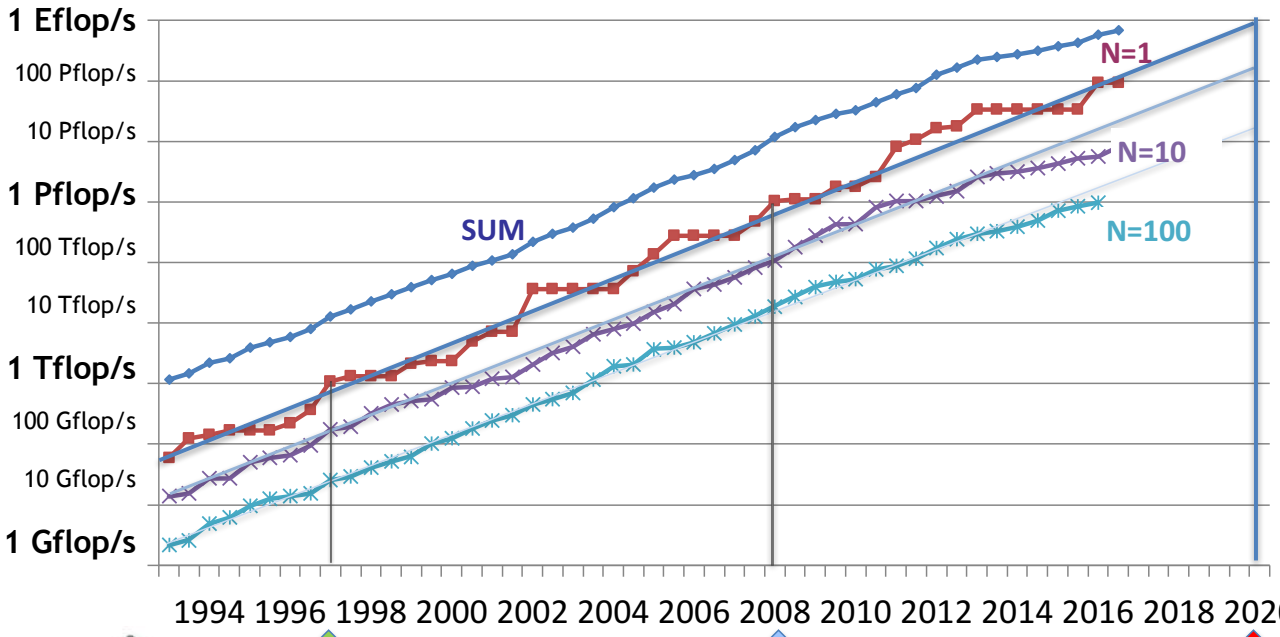
- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June

- All data available from **www.top500.org**

Performance Development of HPC over the Last 24 Years from the Top500



PERFORMANCE DEVELOPMENT



Today: SW 26010 & Intel KNL
~ 3 Tflop/s

Tflops (10^{12})
Achieved
ASCI Red
Sandia NL

Today: 1 cabinet
of TaihuLight
~ 3 Pflop/s

Pflops (10^{15})
Achieved
RoadRunner
Los Alamos NL

Eflops (10^{18})
Achieved?

China says 2020
U.S. says 2021

November 2016: The TOP 10 Systems

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	National Super Computer Center in Wuxi	Sunway TaihuLight, SW26010 (260C) + Custom	China	10,649,000	93.0	74	15.4	6.04
2	National Super Computer Center in Guangzhou	Tianhe-2 NUDT, Xeon (12C) + IntelXeon Phi (57C) + Custom	China	3,120,000	33.9	62	17.8	1.91
3	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7, AMD (16C) + Nvidia Kepler GPU (14C) + Custom	USA	560,640	17.6	65	8.21	2.14
		+ Omni-Path						
7	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8C) + Custom	Japan	705,024	10.5	93	12.7	.827
8	Swiss CSCS	Piz Daint, Cray XC50, Xeon (12C) + Nvidia P100(56C) + Custom	Swiss	206,720	9.78	61	1.31	7.45
9	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16C) + Custom	USA	786,432	8.59	85	3.95	2.07
10	DOE / NNSA / Los Alamos & Sandia	Trinity, Cray XC40,Xeon (16C) + Custom	USA	301,056	8.10	80	4.23	1.92
500	Internet company	Inspur Intel (8C) + Nnvidia	China	5440	.286	71		

TaihuLight is 5.2 X Performance of the ORNL'sTitan

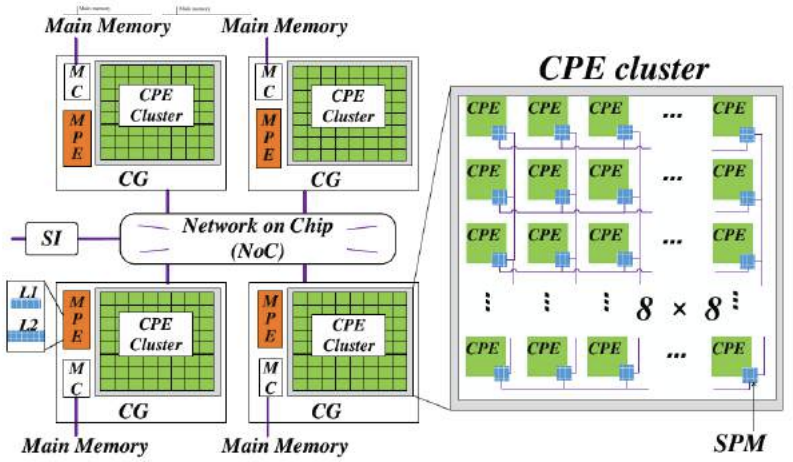
TaihuLight is 69% Sum of All EU's 100 Systems

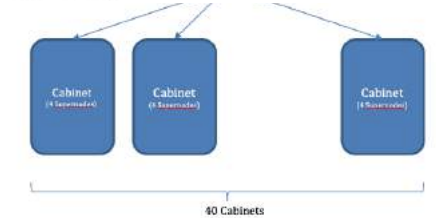
China’s First Homegrown Many-core Processor

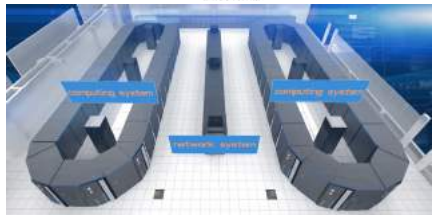
- ShenWei SW26010 Processor
- Vendor: Shanghai High Performance IC Design Center
- Supported by National Science and Technology Major Project (NMP): Core Electronic Devices, High-end Generic Chips, and Basic Software
- 28 nm technology
- 260 Cores
- 3 Tflop/s peak

Sunway TaihuLight http://bit.ly/sunway-2016

- SW26010 processor
- Chinese design, fab, and ISA
- 1.45 GHz
- Node = 260 Cores (1 socket)
  - 4 – core groups
    - 64 CPE, No cache, 64 KB scratchpad/CPE
    - 1 MPE w/32 KB L1 dcache & 256KB L2 cache
  - 32 GB memory total, 136.5 GB/s
  - ~3 Tflop/s, (22 flops/byte)
- Cabinet = 1024 nodes
  - 4 supernodes=32 boards(4 cards/b(2 no
  - ~3.14 Pflop/s
- 40 Cabinets in system
  - 40,960 nodes total
  - 125 Pflop/s total peak
- 10,649,600 cores total
- 1.31 PB of primary memory (DDR3)
- 93 Pflop/s for HPL Benchmark, 74% peak
- 15.3 MWatts, water cooled
  - 6.07 Gflop/s per Watt
- 1.8B RMBs ~ $280M, (building, hw, apps, sw, …)







Gordon Bell Award

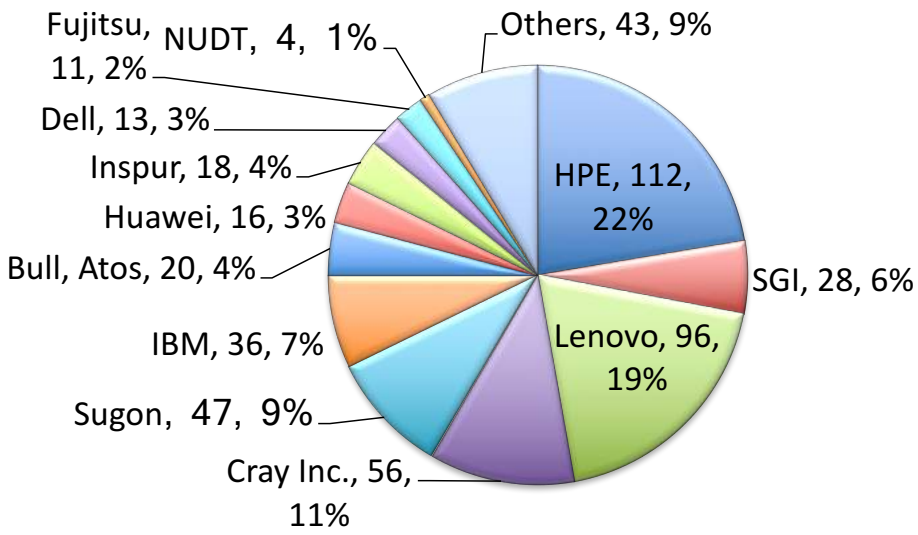
- Since 1987 the ACM's Gordon Bell Prize is awarded at the ACM/IEEE Supercomputing Conference (SC) to recognize outstanding achievement in high-performance computing.
- The purpose of the award is to track the progress of parallel computing, with emphasis on rewarding innovation in applying HPC to applications.
- Financial support of the $10,000 award is provided by Gordon Bell, a pioneer in high-performance and parallel computing.

- Authors' mark their SC paper as a possible Gordon Bell Prize competitor.
- Gordon Bell Committee reviews the papers and selects 6 papers as finalists for the competition.
- Presentations are made at SC and a winner is chosen.

Gordon Bell Award 6 Finalists at SC16 in November

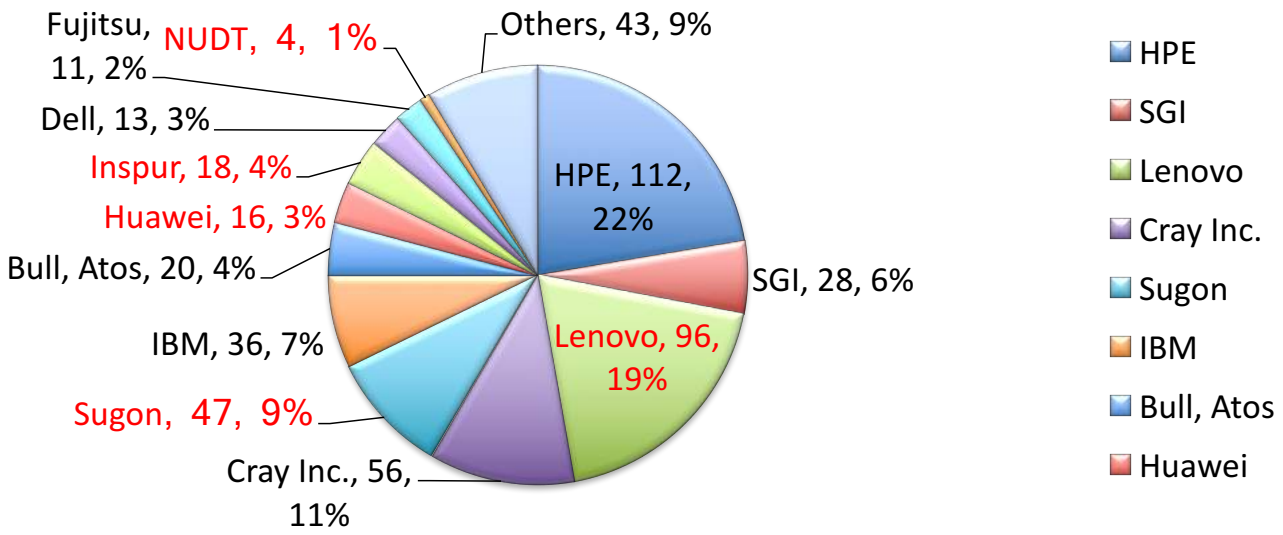
- "**Modeling Dilute Solutions Using First-Principles Molecular Dynamics: Computing More than a Million Atoms with Over a Million Cores**,"
  - Lawrence-Livermore National Laboratory (Calif.)
- "**Towards Green Aviation with Python at Petascale**,"
  - Imperial College London (England)
- "**Simulations of Below-Ground Dynamics of Fungi: 1.184 Pflops Attained by Automated Generation and Autotuning of Temporal Blocking Codes**,"
  - RIKEN (Japan), Chiba University (Japan), Kobe University (Japan) and Fujitsu Ltd. (Japan)
- "**Extreme-Scale Phase Field Simulations of Coarsening Dynamics on the Sunway Taihulight Supercomputer**,"
  - Chinese Academy of Sciences, the University of South Carolina, Columbia University (New York), the National Research Center of Parallel Computer Engineering and Technology (China) and the National Supercomputing Center in Wuxi (China)
- "**A Highly Effective Global Surface Wave Numerical Simulation with Ultra-High Resolution**,"
  - First Institute of Oceanography (China), National Research Center of Parallel Computer Engineering and Technology (China) and Tsinghua University (China)
- "**10M-Core Scalable Fully-Implicit Solver for Nonhydrostatic Atmospheric Dynamics**,"
  - Chinese Academy of Sciences, Tsinghua University (China), the National Research Center of Parallel Computer Engineering and Technology (China) and Beijing Normal University (China)

VENDORS / SYSTEM SHARE



# of systems, % of 500

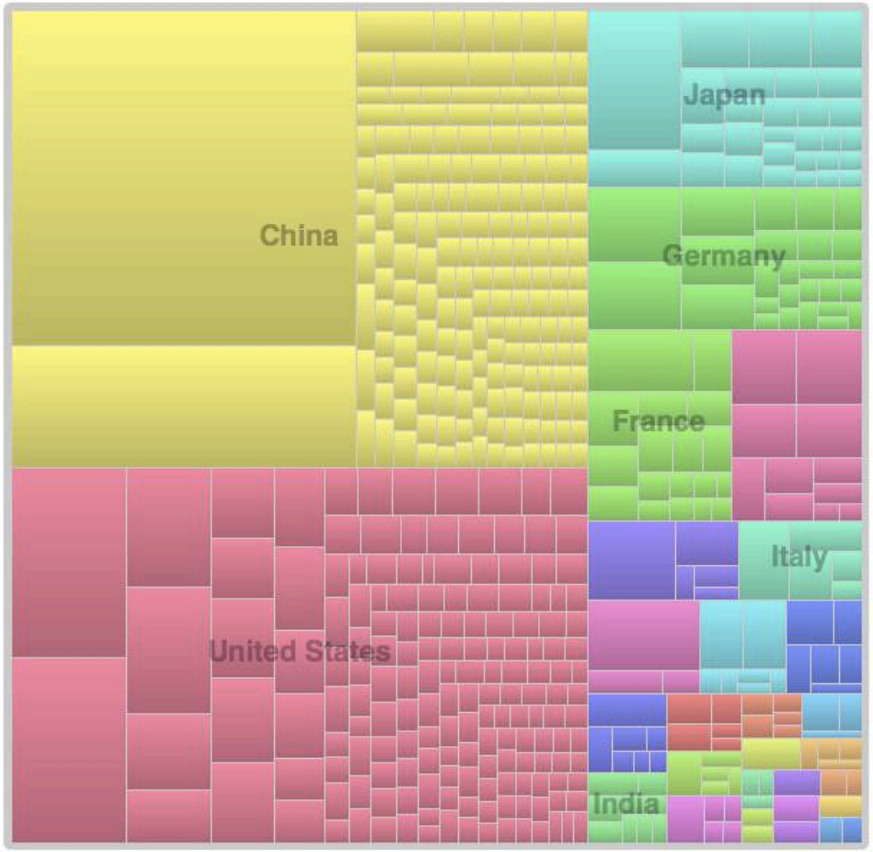
VENDORS / SYSTEM SHARE

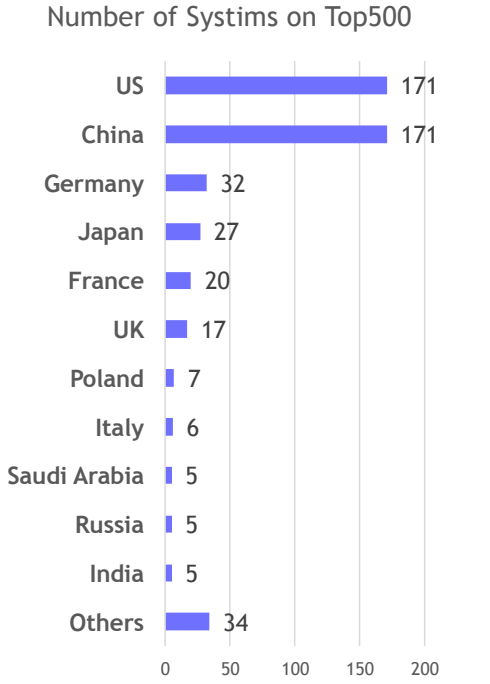


36% of the top 500 systems are from Chinese vendors

# of systems, % of 500

Countries Share of Top500



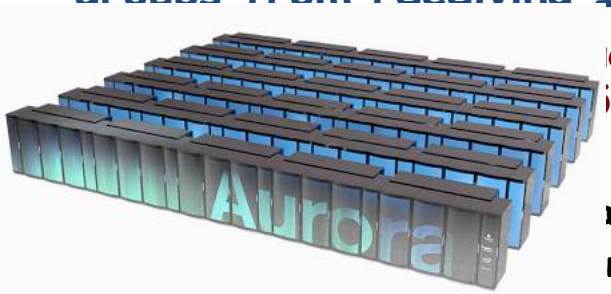


China has 1/3 of the systems, while the number of systems in the US has fallen to the lowest point since the TOP500 list was created.

Each rectangle represents one of the Top500 computers, area of rectangle reflects its performance.

Recent Developments

- US DOE planning to deploy O(100) Pflop/s systems for 2017-2018 - $525M hardware
- Oak Ridge Lab and Lawrence Livermore Lab to receive IBM and Nvidia based systems
- Argonne Lab to receive Intel based system
- After this Exascale systems
- US Dept of Commerce is a
  groups from receiving In
  cle th the
  .



  - ngz
  - njin
  - National University for Def
  - National SC Center Changs

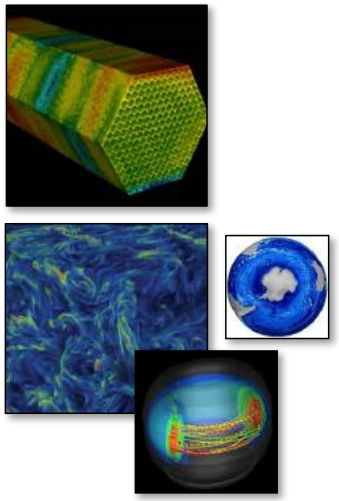
Toward Exascale

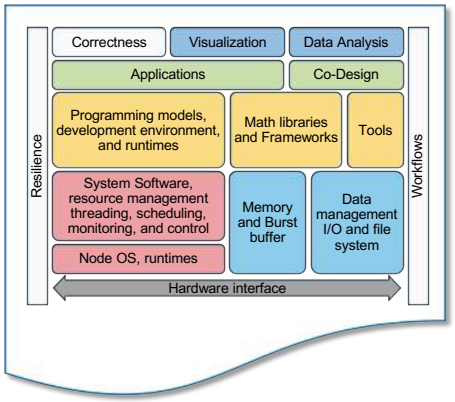
- China plans for Exascale 2020
  - Three separate developments in HPC; "Anything but from the US"
  - Wuxi
    - ShenWei O(100) Pflops all Chinese, 2017
  - National University for Defense Technology
    - Tianhe-2A O(100) Pflops will be Chinese ARM processor + accelerator, 2017
  - Sugon - CAS ICT
    - X86 based; collaboration with AMD
- US DOE - Exascale Computing Program – 7 Year Program
  - Initial exascale system based on *advanced architecture* and delivered in 2021
  - Enable *capable exascale* systems, based on ECP R&D, delivered in 2022 and deployed in 2023

Exascale
- 50X the performance of today's 20 PF systems
- 20-30 MW power
- ≤ 1 perceived fault /week
- SW to support broad range of apps

DOE Exascale Computing Program has formulated a holistic approach that uses co-design and integration to achieve capable exascale

Application Development	Software Technology	Hardware Technology	Exascale Systems
Science and mission applications	Scalable and productive software stack	Hardware technology elements	Integrated exascale supercomputers





ECP's work encompasses applications, system software, hardware technologies and architectures, and workforce development

The DOE ECP Plan of Record

- A **7-year project** that follows the *holistic/co-design* approach, which runs through 2023 (including 12 months schedule contingency)
  - To meet the ECP goals
- Enable an **initial exascale system** based on ***advanced architecture*** and delivered in **2021**
- Enable ***capable exascale* systems**, based on ECP R&D, delivered in **2022 and deployed in 2023** as part of an DOE facility upgrade
- Acquisition of the exascale systems is outside of the ECP scope, will be carried out by DOE facilities

Funding for ECP Application, Co-design Center, and Software Project

NEWS RELEASE

The Exascale Computing Project Awards $34 Million for Software Development

OAK RIDGE, Tenn., Nov. 10, 2016 – The Department of Energy's Exascale Com (ECP) today announced the selection of 35 software development proposals research and academic organizations.

The awards for the first year of funding total $34 million and cover many com software stack for exascale systems, including programming models and run mathematical libraries and frameworks, tools, lower-level system software, and I/O, as well as in situ visualization and data analysis.

NEWS RELEASE

The Exascale Computing Project Announces $48 Million to Establish Four Exascale Co-Design Centers

OAK RIDGE, Tenn., Nov. 11, 2016 – The Department of Energy's Exascale Computing Project (ECP) today announced that it has selected four co-design centers as part of a 4 year, $48 million funding award. The first year is funded at $12 million, and is to be allocated evenly among the four award recipients.

The ECP is responsible for the planning, execution, and delivery of technologies necessary for a capable exascale ecosystem to support the nation's exascale imperative including software, and early testbed platforms.

NEWS RELEASE

For Immediate Distribution

The Exascale Computing Project (ECP) Announces $39.8 million in First-Round Application Development Award

OAK RIDGE, Tenn., Sept. 07, 2016 – The Department of Energy's Exascale Computi (ECP) today announced its first round of funding with the selection of 15 applicatio development proposals for full funding and seven proposals for seed funding, repr teams from 45 research and academic organizations.

The awards, totaling $39.8 million, target advanced modeling and simulation solut specific challenges supporting key DOE missions in science, clean energy and natio as well as collaborations such as the Precision Medicine Initiative with the Nation Health's National Cancer Institute.

EXASCALE COMPUTING PROJECT

Exascale Race/Technologies

IDC-Projected Exascale Dates and Suppliers

U.S.

- Sustained ES: 2023
- Peak ES: 2021
- Vendors: U.S.
- Processors: U.S.
- Initiatives: NSCI/ECP
- Cost: $300-500M per system, plus heavy R&D investments

EU

- Sustained ES: 2023-24
- Peak ES: 2021
- Vendors: U.S., Europe
- Processors: U.S., ARM
- Initiatives: PRACE, ETP4HPC
- Cost: $300-$350 per system, plus heavy R&D investments

China

- Sustained ES: 2023
- Peak ES: 2020
- Vendors: Chinese
- Processors: Chinese (plus U.S.?)
- 13th 5-Year Plan
- Cost: $350-500M per system, plus heavy R&D

Japan

- Sustained ES: 2023-24
- Peak ES: Not planned
- Vendors: Japanese
- Processors: Japanese
- Cost: $600-850M, this includes both 1 system and the R&D costs…will also do many smaller size systems

IDC-Projected Exascale Investment Levels (In Addition to System Purchases)

U.S.

- $1 to $2 billion a year in R&D (including NRE)
- Investments by both governments & vendors
- Plans are to purchase multiple exascale systems

EU

- About 5 billion euros in total
- Investments in multiple exascale and pre-exascale systems
- Investments mostly by country governments with a little from the EU

China

- Over $1billion a year in R&D
- Investments by both governments & vendors
- Plans are to purchases multiple exascale systems each year
- Already investing in 3 pre-exascale systems by 2017/18

Japan

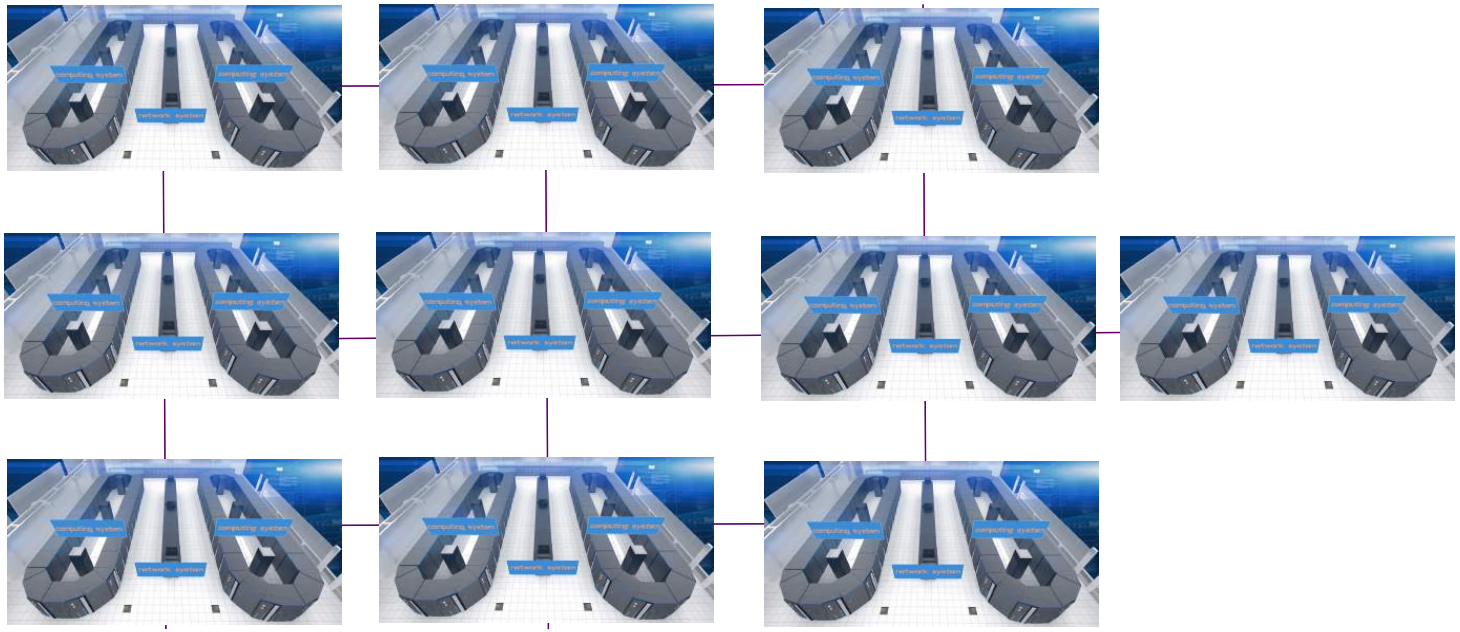
- Planned investment of just over $1billion* (over 5 years) for both the R&D and purchase of 1 exascale system
- To be followed by a number of smaller systems ~$100M to $150M each
- Creating a new processor and a new software environment

07

** Note that this includes both the system and R&D*

We Can Build an Exascale System Today?

Connect together 10 Sunway TaihuLight systems

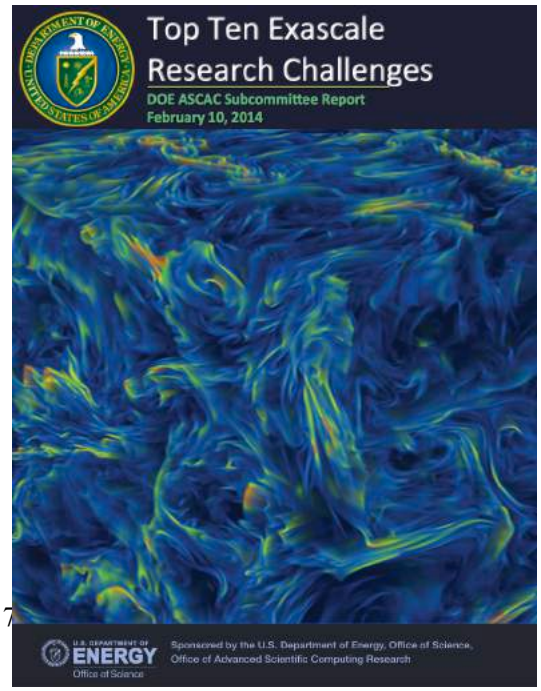


Require 150 MW of power, programming for 100 M threads, and $2.7B price tag

Top 10 Challenges to Exascale

U.S. Department of Energy report identified ten research challenges

(Google "Top 10 Challenges to Exascale")



ASCAC Subcommittee for the Top Ten Exascale Research Challenges

Subcommittee Chair
Robert Lucas (University of Southern California, Information Sciences Institute)

Subcommittee Members
James Ang (Sandia National Laboratories)
Keren Bergman (Columbia University)
Shekhar Borkar (Intel)
William Carlson (Institute for Defense Analyses)
Laura Carrington (UC, San Diego)
George Chiu (IBM)
Robert Colwell (DARPA)
William Dally (NVIDIA)
Jack Dongarra (U. Tennessee)
Al Geist (ORNL)
Gary Grider (LANL)
Rud Haring (IBM)
Jeffrey Hittinger (LLNL)
Adolfy Hoisie (PNLL)
Dean Klein (Micron)
Peter Kogge (U. Notre Dame)
Richard Lethin (Reservoir Labs)
Vivek Sarkar (Rice U.)
Robert Schreiber (Hewlett Packard)
John Shalf (LBNL)
Thomas Sterling (Indiana U.)
Rick Stevens (ANL)

Top 10 Challenges to Exascale

3 Hardware, 4 Software, 3 Algorithms/Math Related

- **Energy efficiency:**
  - Creating more energy efficient circuit, power, and cooling technologies.
- **Interconnect technology:**
  - Increasing the performance and energy efficiency of data movement.
- **Memory Technology:**
  - Integrating advanced memory technologies to improve both capacity and bandwidth.
- **Scalable System Software:**
  - Developing scalable system software that is power and resilience aware.
- **Programming systems:**
  - Inventing new programming environments that express massive parallelism, data locality, and resilience
- **Data management:**
  - Creating data management software that can handle the volume, velocity and diversity of data that is anticipated.
- **Scientific productivity:**
  - Increasing the productivity of computational scientists with new software engineering tools and environments.
- **Exascale Algorithms:**
  - Reformulating science problems and refactoring their solution algorithms for exascale systems.
- **Algorithms for discovery, design, and decision:**
  - Facilitating mathematical optimization and uncertainty quantification for exascale discovery, design, and decision making.
- **Resilience and correctness:**
  - Ensuring correct scientific computation in face of faults, reproducibility, and algorithm verification challenges.

Confessions of an Accidental Benchmarker

- Appendix B of the Linpack Users' Guide
  - Designed to help users extrapolate execution Linpack software package
- First benchmark report from 1977;

UNIT = 10**6 TIME/(1/3 100**3 + 100**2)

Facility	TIME N=100 secs.	UNIT micro-secs.	Computer	Type	Compiler
NCAR	.049	0.14	CRAY-1	S	CFT, Assembly BLAS
LASL	.148	0.43	CDC 7600	S	FTN, Assembly BLAS
NCAR	.192	0.56	CRAY-1	S	CFT
LASL	.210	0.61	CDC 7600	S	FTN
Argonne	.297	0.86	IBM 370/195	D	H
NCAR	.359	1.05	CDC 7600	S	Local
Argonne	.388	1.33	IBM 3033	D	H
NASA Langley	.489	1.42	CDC Cyber 175	S	FTN
U. Ill. Urbana	.506	1.47	CDC Cyber 175	S	Ext. 4.6
LLL	.554	1.61	CDC 7600	S	CHAT, No optimize
SLAC	.579	1.69	IBM 370/168	D	H Ext., Fast mult.
Michigan	.631	1.84	Amdahl 470/V6	D	H
Toronto	.890	2.59	IBM 370/165	D	H Ext., Fast mult.
Northwestern	1.44	4.20	CDC 6600	S	FTN
Texas	1.93*	5.63	CDC 6600	S	RUN
China Lake	1.95*	5.69	Univac 1110	S	V
Yale	2.59	7.53	DEC KL-20	S	F20
Bell Labs	3.46	10.1	Honeywell 6080	S	Y
Wisconsin	3.49	10.1	Univac 1110	S	V
Iowa State	3.54	10.2	Itel AS/5 mod3	D	H
U. Ill. Chicago	4.10	11.9	IBM 370/158	D	G1
Purdue	5.69	16.6	CDC 6500	S	FUN
U. C. San Diego	13.1	38.2	Burroughs 6700	S	H
Yale	17.1*	49.9	DEC KA-10	S	F40

* TIME(100) = (100/75)**3 SGEFA(75) + (100/75)**2 SGESL(75)

Many Other Benchmarks

- TOP500
- Green 500
- Graph 500
- Sustained Petascale Performance
- HPC Challenge
- Perfect
- ParkBench
- SPEC-hpc
- Big Data Top100
- Livermore Loops
- EuroBen
- NAS Parallel Benchmarks
- Genesis
- RAPS
- SHOC
- LAMMPS
- Dhrystone
- Whetstone
- I/O Benchmarks
- WRF
- Yellowstone
- Roofline
- Neptune

LINPACK Benchmark
High Performance Linpack (HPL)

- Is a **widely recognized** and discussed metric for ranking high performance computing systems
- When HPL gained prominence as a performance metric in the early 1990s there **was a strong correlation between its predictions of system rankings and the ranking that full-scale applications would realize**.
- **Computer system vendors pursued designs that would increase their HPL performance**, which would in turn improve overall application performance.
- Today HPL remains **valuable as a measure of historical trends**, and as a stress test, especially for leadership class systems that are pushing the boundaries of current technology.

Peak Performance - Per Core

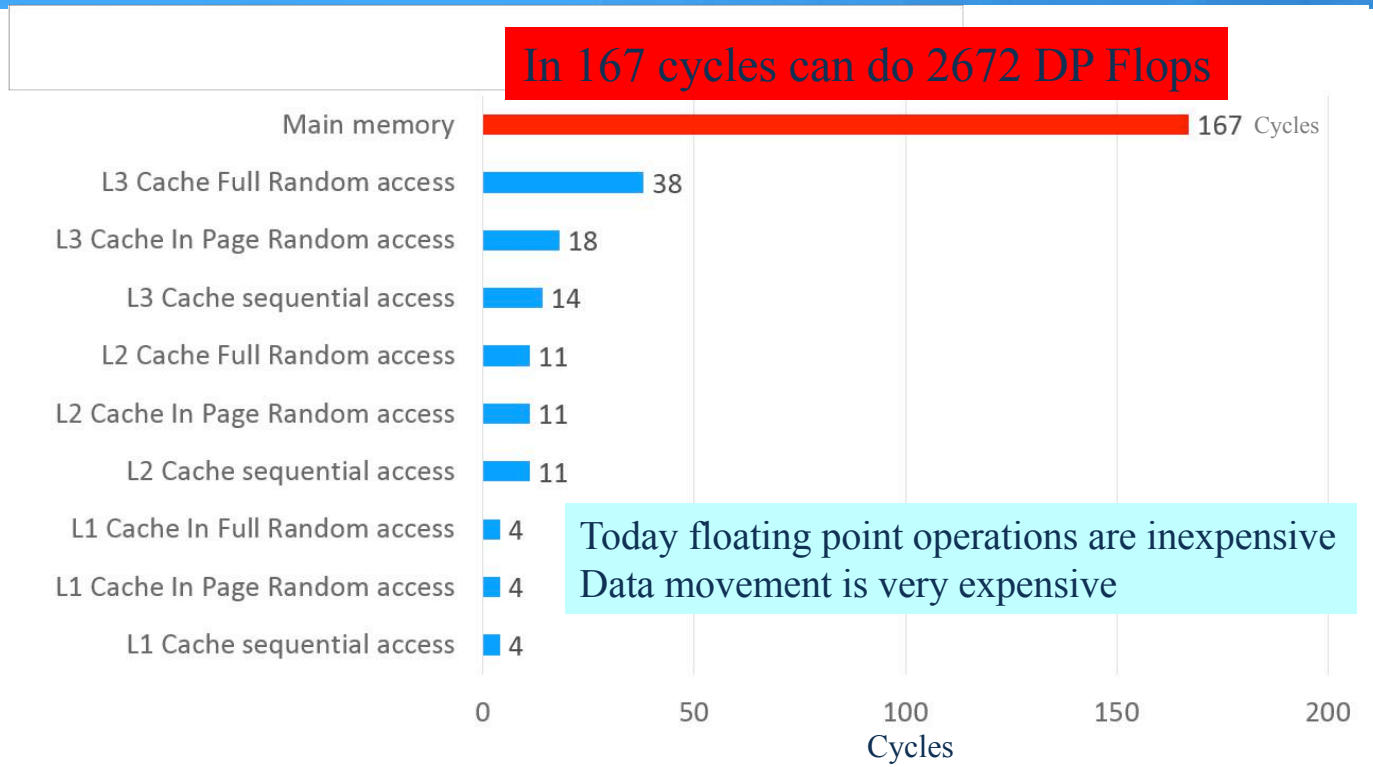
$$\text{FLOPS} = \text{cores} \times \text{clock} \times \frac{\text{FLOPs}}{\text{cycle}}$$

Floating point operations per cycle per core

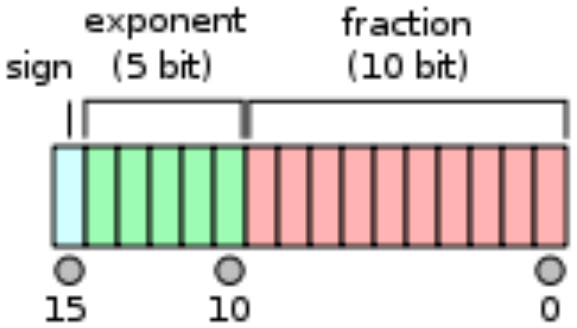
- Most of the recent computers have FMA (Fused multiple add): (i.e. x ←x + y*z in one cycle)
- Intel Xeon earlier models and AMD Opteron have SSE2
  - 2 flops/cycle DP & 4 flops/cycle SP
- Intel Xeon Nehalem ('09) & Westmere ('10) have SSE4
  - 4 flops/cycle DP & 8 flops/cycle SP
- Intel Xeon Sandy Bridge('11) & Ivy Bridge ('12) have AVX
  - 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 (server) AVX 512
  - 32 flops/cycle DP & 64 flops/cycle SP
  - Knight's Landing

We are here

CPU Access Latencies in Clock Cycles



IEEE 754 Half Precision (16-bit) Fl Pt Standard



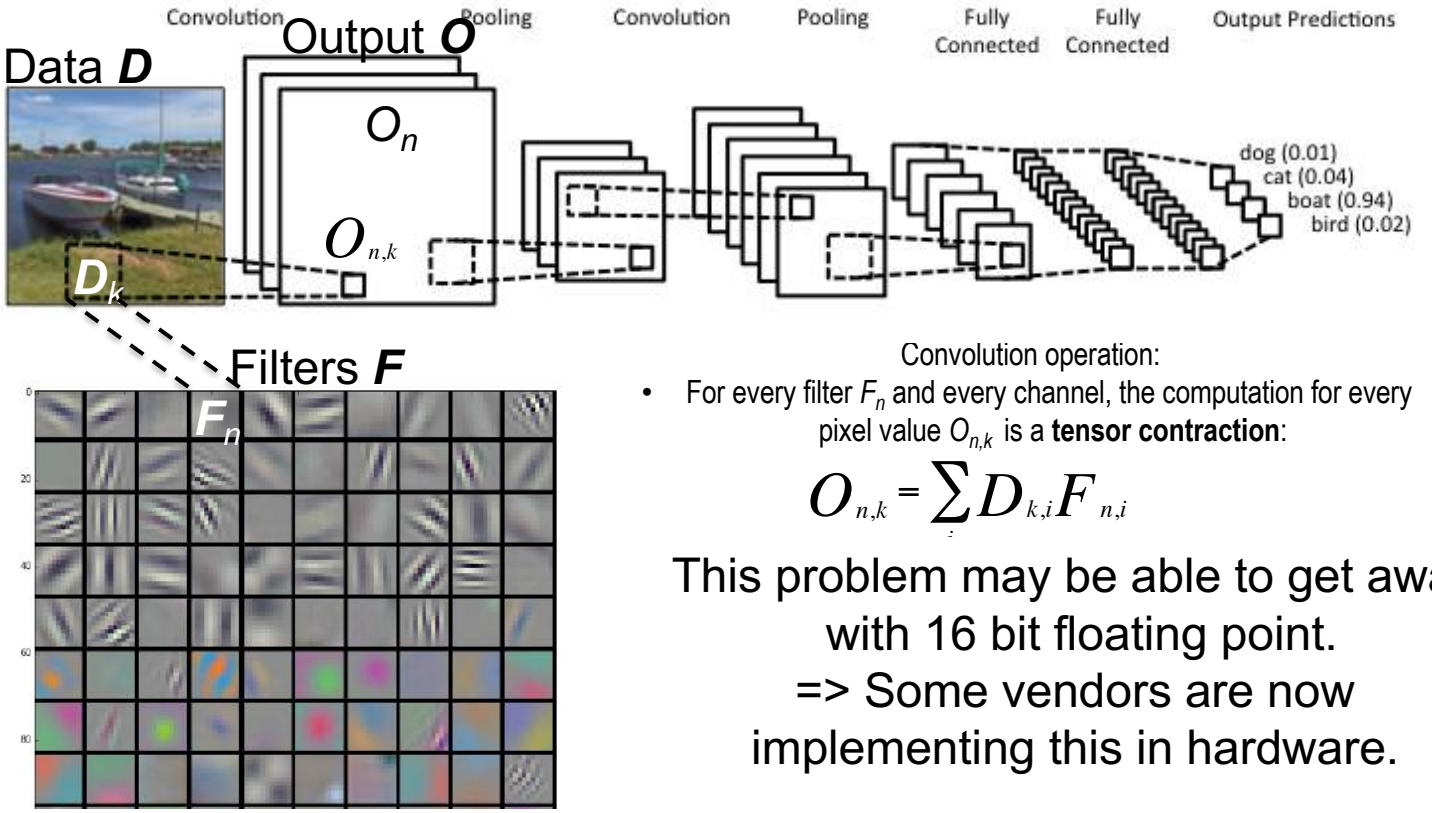
AMD Radeon Instinct			
	Instinct MI6	Instinct MI8	Instinct MI25
Memory Type	16GB GDDR5	4GB HBM	"High Bandwidth Cache and Controller"
Memory Bandwidth	224GB/sec	512GB/sec	?
Single Precision (FP32)	5.7 TFLOPS	8.2 TFLOPS	12.5 TFLOPS
Half Precision (FP16)	5.7 TFLOPS	8.2 TFLOPS	25 TFLOPS
TDP	<150W	<175W	<300W
Cooling	Passive	Passive (SFF)	Passive
GPU	Polaris 10	Fiji	Vega
Manufacturing Process	GloFo 14nm	TSMC 28nm	?

Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK110 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	NA	NA	NA	8
Tensor Cores / GPU	NA	NA	NA	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1455 MHz
Peak FP32 TFLOP/s*	5.04	6.8	10.6	15
Peak FP64 TFLOP/s*	1.68	2.1	5.3	7.5
Peak Tensor Core TFLOP/s*	NA	NA	NA	120

Machine Learning

Need of **Batched and/or Tensor contraction** routines in **machine learning** –

e.g., Convolutional Neural Networks (CNNs) used in computer vision
Key computation is convolution of Filter Fi (feature detector) and input image D (data):



Convolution operation:

- For every filter F_n and every channel, the computation for every pixel value $O_{n,k}$ is a **tensor contraction**:

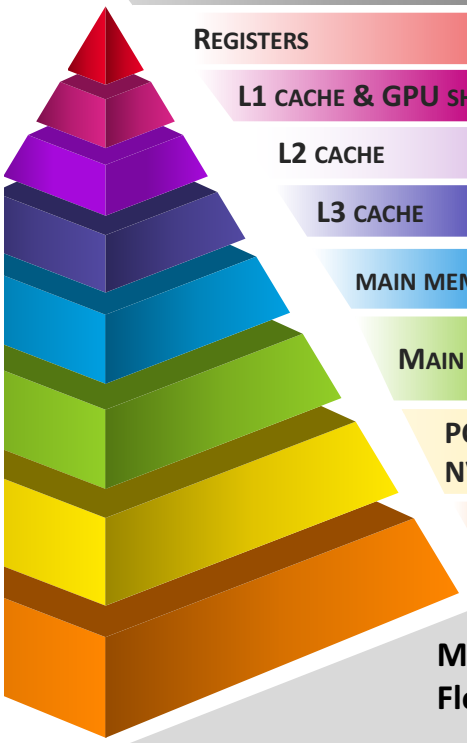
$$O_{n,k} = \sum_i D_{k,i} F_{n,i}$$

This problem may be able to get away with 16 bit floating point.
=> Some vendors are now implementing this in hardware.

The Problem with the LINPACK Benchmark

- HPL performance of computer systems are **no longer so strongly correlated to real application performance**, especially for the broad set of HPC applications governed by partial differential equations.

- **Designing a system for good HPL performance can actually lead to design choices that are wrong** for the real application mix, or add unnecessary components or complexity to the system.



Memory hierarchies	Intel Haswell E5-2650 v3	Intel KNL 7250 DDR5|MCDRAM	ARM Cortex A57	Nvidia P100	Nvidia V100
	10 cores 368 Gflop/s 105 Watts	68 cores 2662 Gflop/s 215 Watts	4 cores 32 Gflop/s 7 Watts	56 SM 64 cores 4700 Gflop/s 250 Watts	80 SM 64 cores 7500 Gflop/s 300 Watts
REGISTERS	16/core AVX2	32/core AVX-512	32/core	256 KB/SM	256 KB/SM
L1 CACHE & GPU SHARED MEMORY	32 KB/core	32 KB/core	32 KB/core	64 KB/SM	96 KB/SM
L2 CACHE	256 KB/core	1024 KB/2cores	2 MB	4 MB	6 MB
L3 CACHE	25 MB	0...16 GB	N/A	N/A	N/A
MAIN MEMORY	64 GB	384 |16 GB	4 GB	16 GB	16 GB
MAIN MEMORY BW	68 GB/s 5.4 flops/byte	115 | 421 GB/s 23 | 6 Flops/byte	26 GB/s 1.2 flops/byte	720 GB/s 6.5 flops/byte	900 GB/s 8.3 flops/byte
PCI EXPRESS GEN3X16 NVLINK	16 GB/s 23 flops/byte	16 GB/s 166 flops/byte	16 GB/s 2 flops/byte	16 GB/s 294 flops/byte	300 GB/s (NVL) 25 flops/byte
INTERCONNECT INFINIBAND EDR	12 GB/s 30 flops/byte	12 GB/s 221 flops/byte	12 GB/s 2.6 flops/byte	12 GB/s 392 flops/byte	12 GB/s 625 flops/byte

Memory hierarchies for different type of architectures
Flops per byte transfer (all flop rates for 64 bit operands)

Many Problems in Computational Science Involve Solving PDEs; Large Sparse Linear Systems

Given a PDE, e.g.:

$$-\Delta u + \vec{c} \bullet \nabla u + \gamma u \equiv Pu = f$$

over some domain (where P denotes the differential operator)

Discretization (e.g., Galerkin equations)

$$\text{Find } u_h = \sum_{i=1}^{n} \Phi_i x_i$$

$$\sum_{i=1}^{n} (P u_h, \Phi_j) = (f, \Phi_i) \text{ for } \forall \Phi_j$$

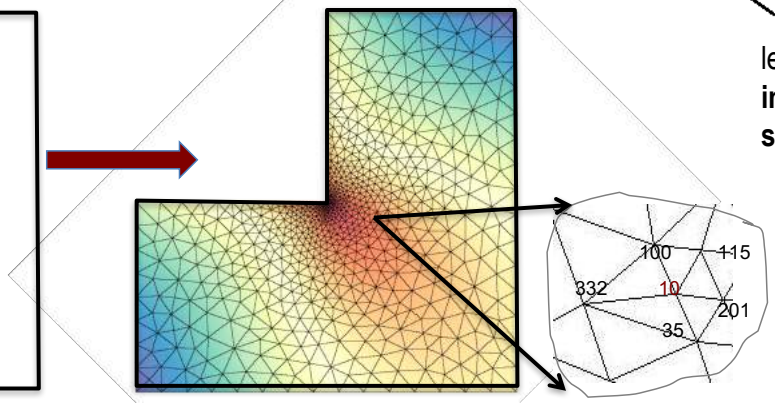
$$\Leftrightarrow \sum_{i=1}^{n} \underbrace{(P\Phi_i, \Phi_j)}_{a_{ji}} x_i = \underbrace{(f, \Phi_i)}_{b_j}$$

$\Leftrightarrow$ Sparse Linear System

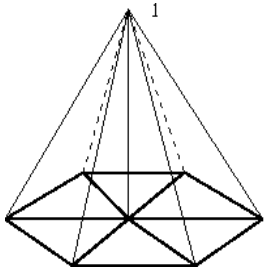
$$\mathbf{A x = b}$$

\+

boundary conditions



Basis functions Φ_j are often with local support, e.g.,



leading **to local interactions** & hence **sparse matrices**, e.g.,

row 10 in this case will have only 6 non-zeroes:

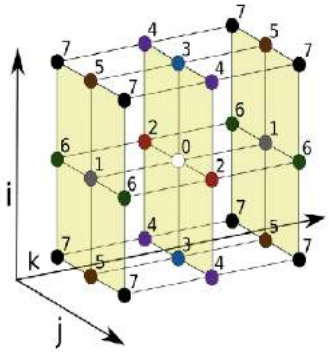
$a_{10,10}$, $a_{10,332}$, $a_{10,100}$, $a_{10,115}$, $a_{10,201}$, $a_{10,35}$

Modeling Diffusion Fluid Flow

hpcg-benchmark.org

HPCG

- High Performance Conjugate Gradients (HPCG).
- Solves *Ax=b, A* large, sparse, *b* known, *x* computed.
- An optimized implementation of PCG contains essential computational and communication patterns that are prevalent in a variety of methods for discretization and numerical solution of PDEs
- Synthetic discretized 3D PDE (FEM, FVM, FDM).
- Sparse matrix:
  - 27 nonzeros/row interior.
  - 8 – 18 on boundary.
  - Symmetric positive definite.

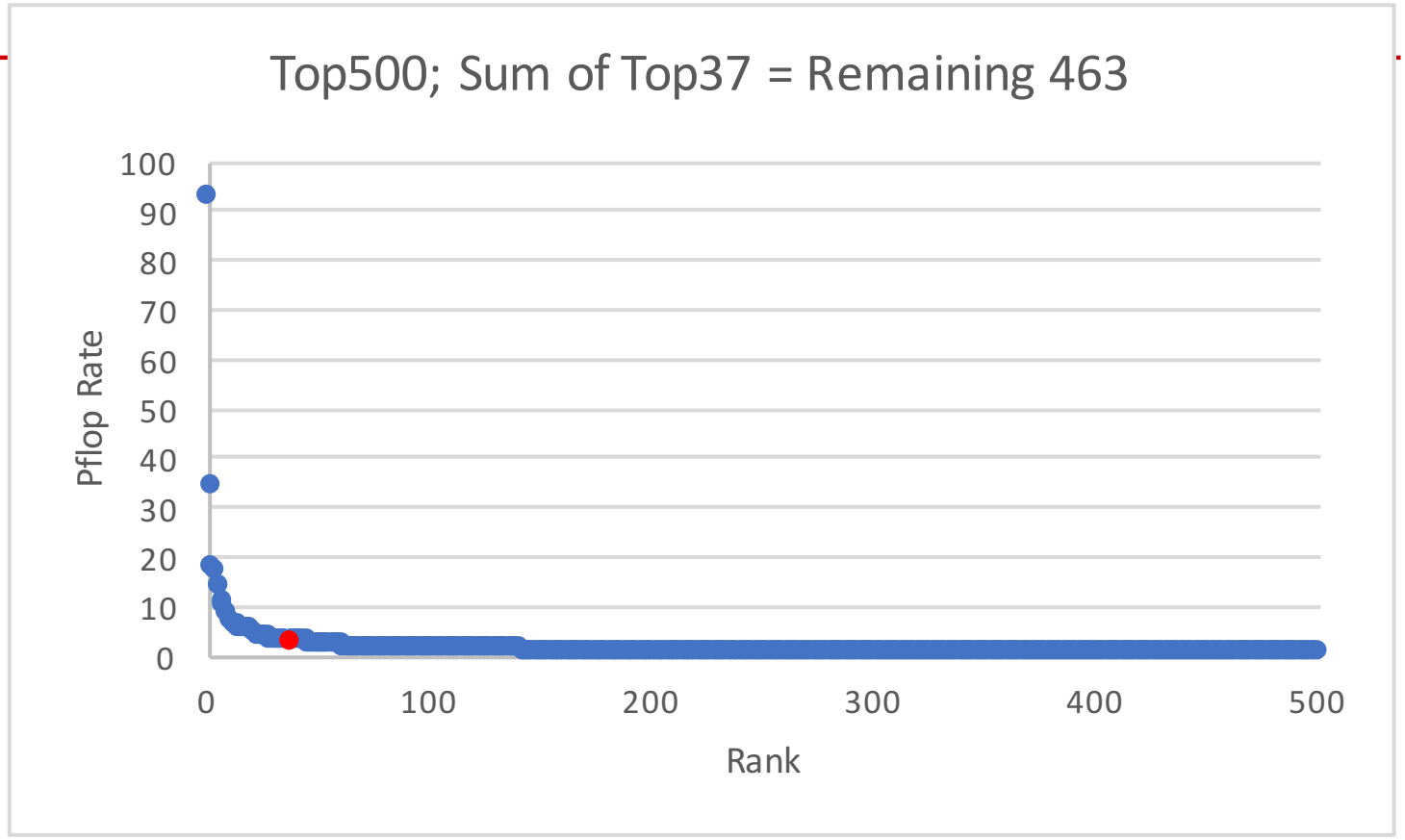


27-point stencil operator

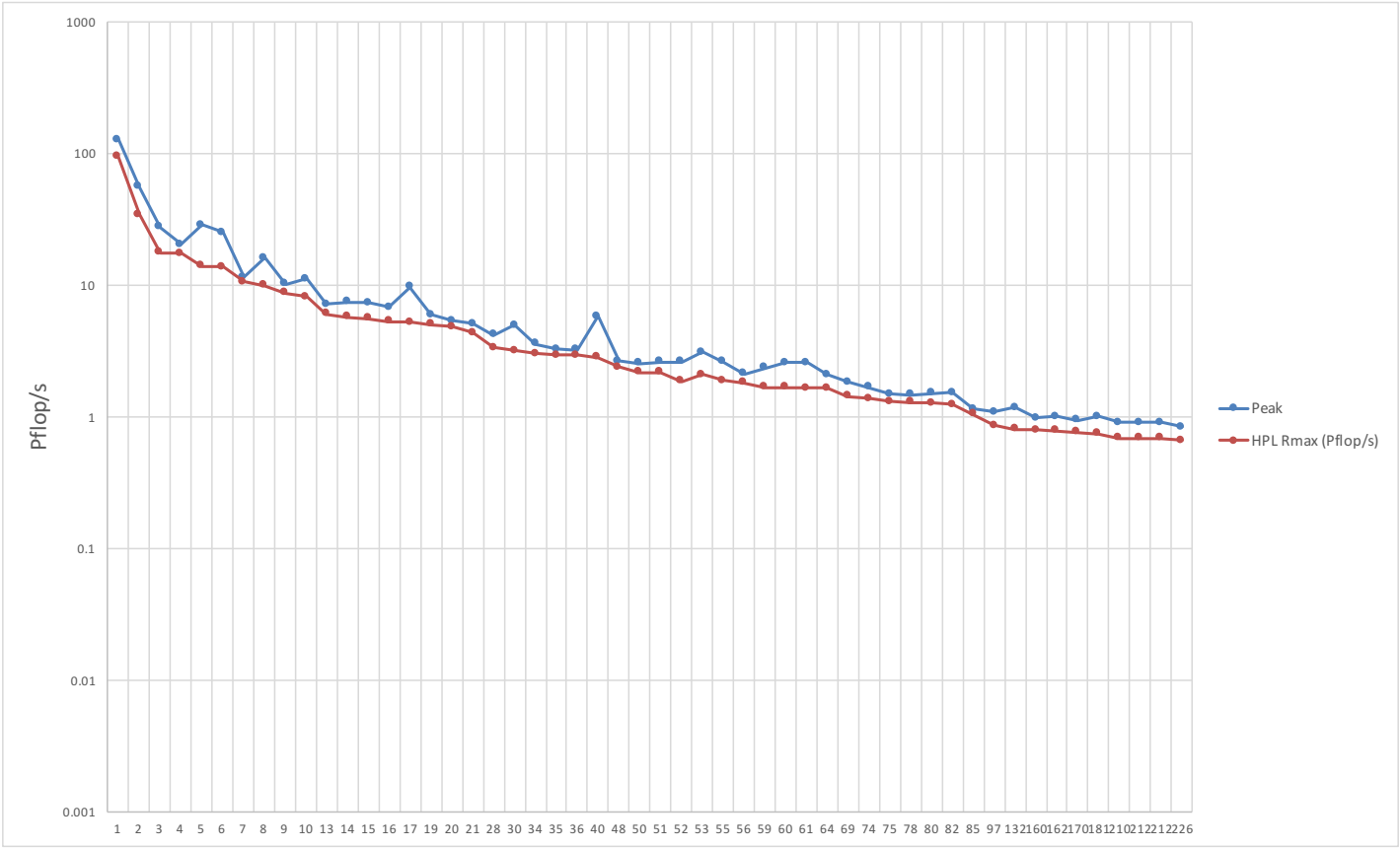
- Patterns:
  - Dense and sparse computations.
  - Dense and sparse collectives.
  - Multi-scale execution of kernels via MG (truncated) V cycle.
  - Data-driven parallelism (unstructured sparse triangular solves).
- Strong verification (via spectral properties of PCG).

HPCG Results, Nov 2016, 1-10

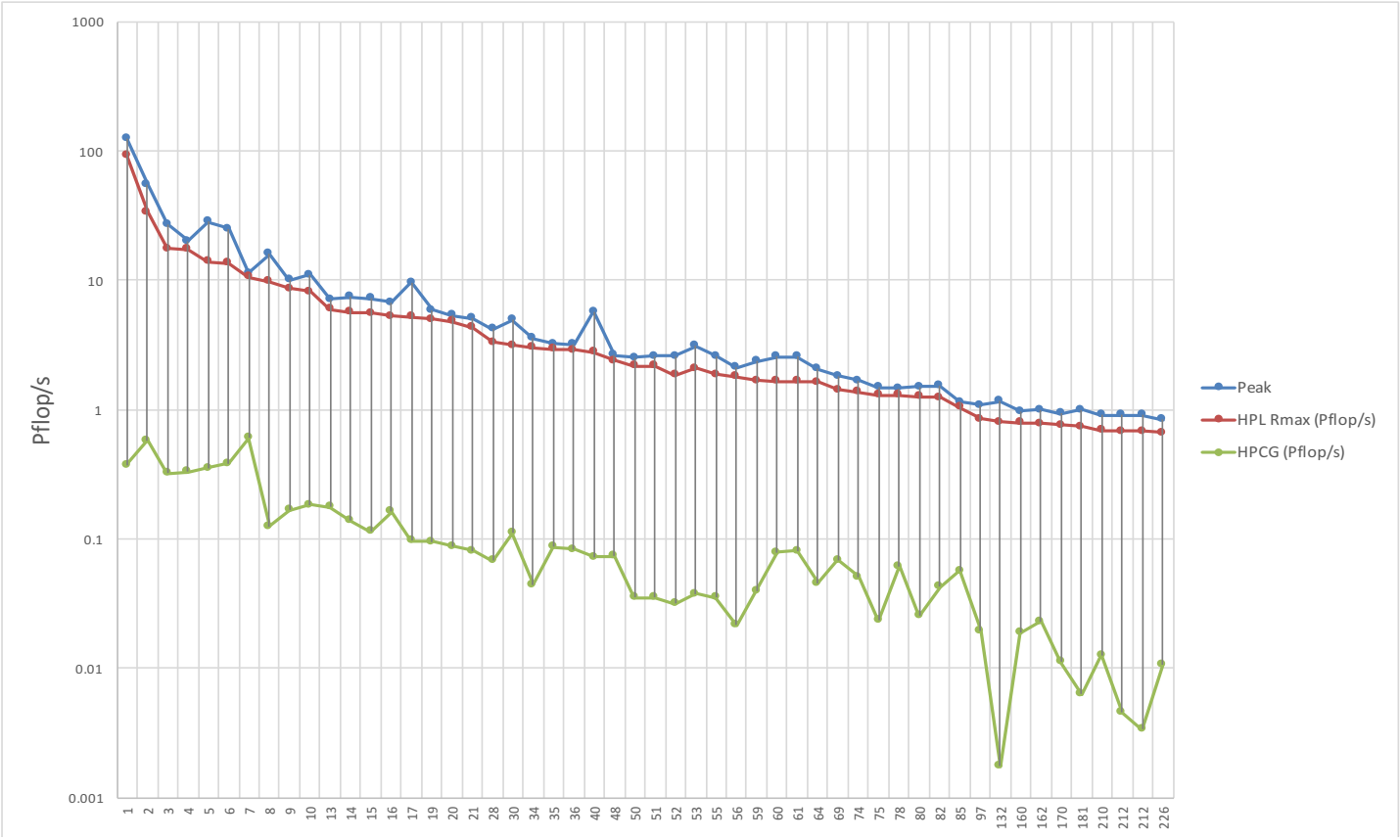
#	Site	Computer	Cores	HPL Pflops	HPCG Pflops	% of Peak
1	RIKEN Advanced Institute for Computational Science	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	705,024	10.5	0.603	5.3%
2	NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.8	0.580	1.1%
3	Joint Center for Advanced HPC, Japan	Oakforest-PACS – PRIMERGY CX600 M1, Intel Xeon Phi	557,056	24.9	0.385	2.8%
4	National Supercomputing Center in Wuxi, China	Sunway TaihuLight – Sunway MPP, SW26010	10,649,600	93.0	0.3712	0.3%
5	DOE/SC/LBNL/NERSC USA	Cori – XC40, Intel Xeon Phi Cray	632,400	13.8	0.355	1.3%
6	DOE/NNSA/LLNL USA	Sequoia – IBM BlueGene/Q, IBM	1,572,864	17.1	0.330	1.6%
7	DOE/SC/Oak Ridge Nat Lab	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x	560,640	17.5	0.322	1.2%
8	DOE/NNSA/LANL/SNL	Trinity - Cray XC40, Intel E5-2698v3, Aries custom	301,056	8.10	0.182	1.6%
9	NASA / Mountain View	Pleiades - SGI ICE X, Intel E5-2680, E5-2680V2, E5-2680V3, Infiniband FDR	243,008	5.90	0.175	2.5%
10	DOE/SC/Argonne National Laboratory	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom	786,432	8.58	0.167	1.7%


Top500; Sum of Top37 = Remaining 463
Pflop Rate
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
Rank

Comparison Peak, HPL, & HPCG



Comparison Peak, HPL, & HPCG



Critical Issues at Peta & Exascale for Algorithm and Software Design

- Synchronization-reducing algorithms
  - Break Fork-Join model
- Communication-reducing algorithms
  - Use methods which have lower bound on communication
- Mixed precision methods
  - 2x speed of ops and 2x speed for data movement
- Autotuning
  - Today's machines are too complicated, build "smarts" into software to adapt to the hardware
- Fault resilient algorithms
  - Implement algorithms that can recover from failures/bit flips
- Reproducibility of results
  - Today we can't guarantee this. We understand the issues, but some of our "colleagues" have a hard time with this.