National Institute of Standards and Technology
U.S. Department of Commerce

An Overview of High Performance Computing and Benchmark Changes for the Future

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory

State of Supercomputing in 2016

- Pflops (> 10^{15} Flop/s) computing fully established with 117 systems.
- Three technology architecture possibilities or "swim lanes" are thriving.
  - Commodity (e.g. Intel)
  - Commodity + accelerator (e.g. GPUs) (88 systems)
  - Lightweight cores (e.g. IBM BG, ARM, Knights Landing)
- 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%.

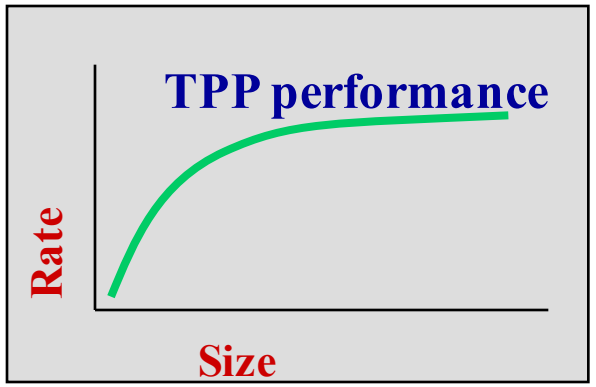
ICL UT

TOP500 supercomputer

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

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

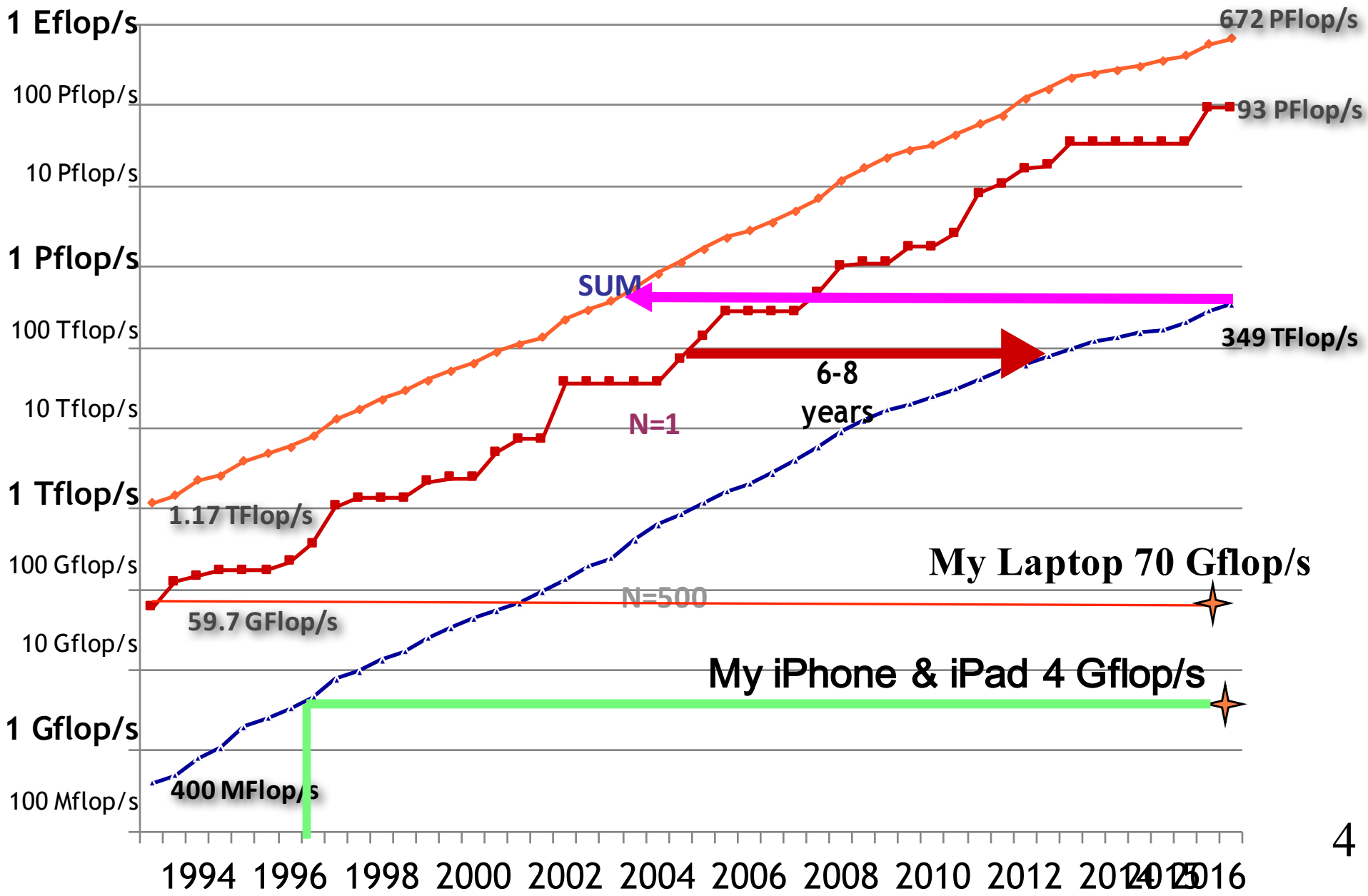
$Ax=b$, *dense problem*



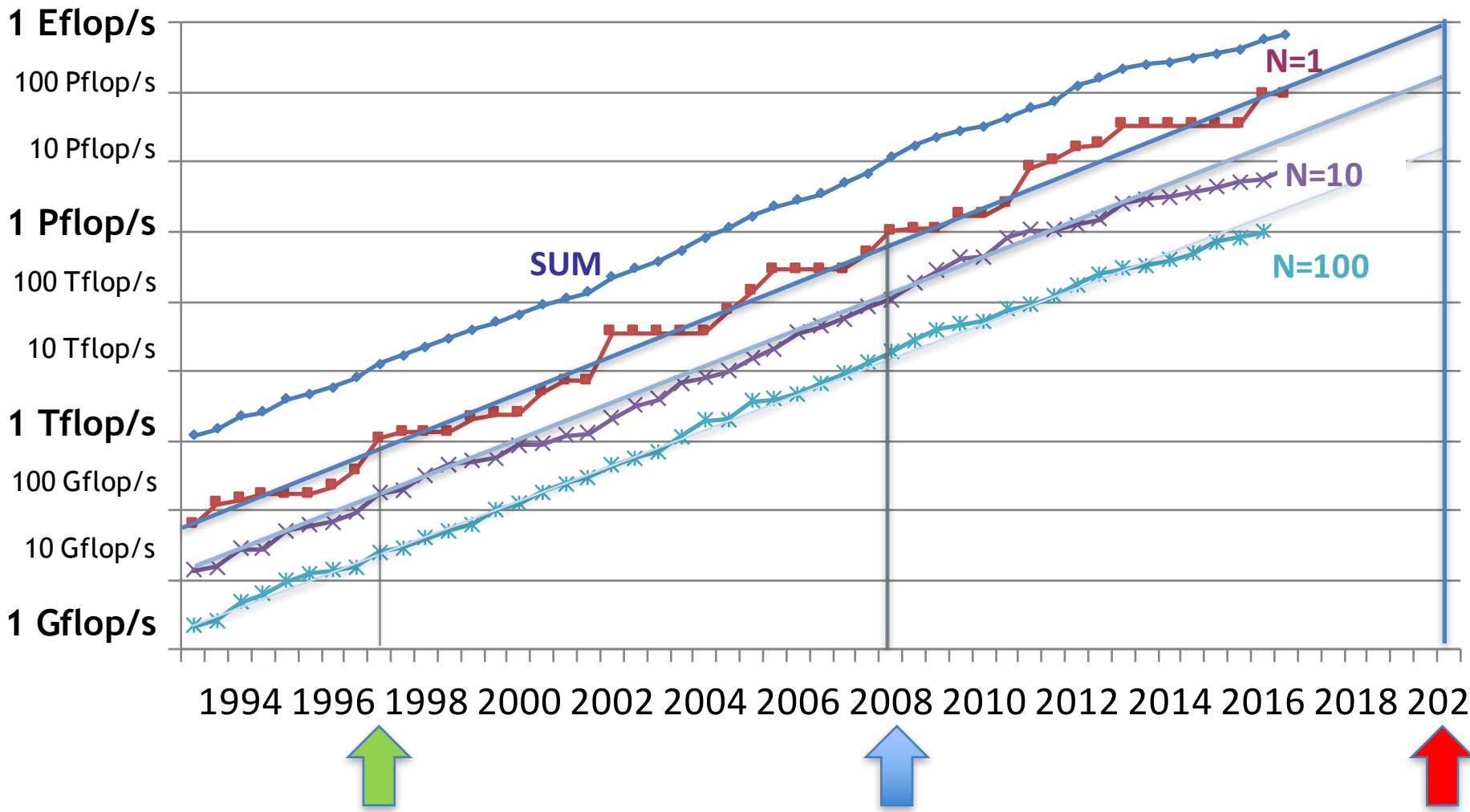
- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June

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

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



PERFORMANCE DEVELOPMENT



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
							89	2.18
6	Advanced HPC	Primergy CX1640, Xeon Phi (68C) + Omni-Path	Japan	558,144	13.6	54	2.72	4.98
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 Titan

TaihuLight is 1.1 X Sum of All DOE Systems

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 Exaflops
- US Dept of Commerce is a
  groups from receiving In
  cle th the
  .

ngz
njin
Def
- National SC Center Changs

Since the Dept of Commerce Action …

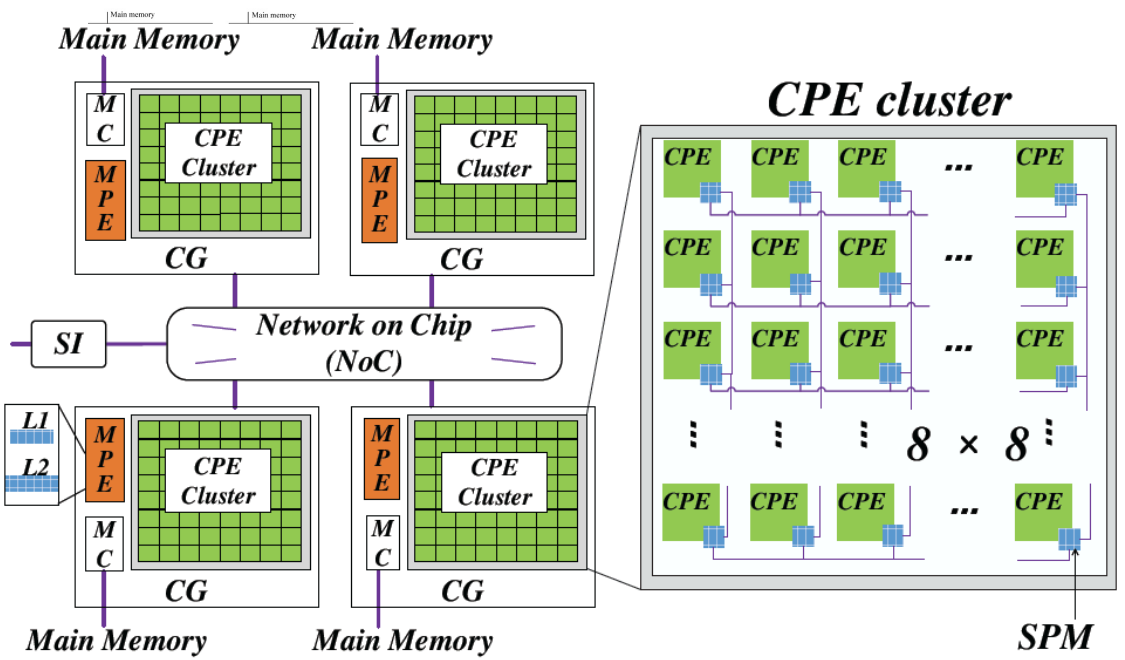
- Expanded focus on Chinese made HW and SW
  - "Anything but from the US"
- Three separate developments in HPC
  - Wuxi
    - ShenWei O(100) Pflops all Chinese, June 2016
  - NUDT
    - Tianhe-2A O(100) Pflops will be Chinese ARM + accelerator, 2017
  - Sugon - CAS ICT
    - AMD? new processors
- In the latest "5 Year Plan"
  - Govt push to build out a domestic HPC ecosystem.
  - Exascale system, will not use any US chips

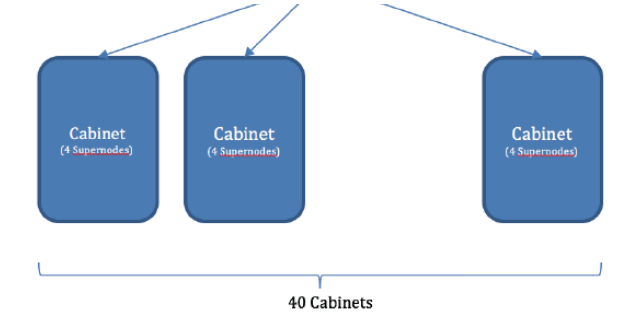
SW26010 Processor

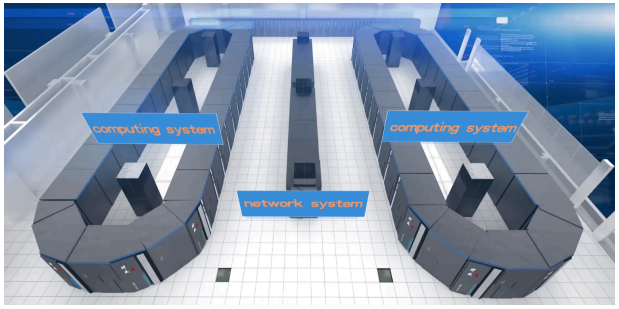
- China's first homegrown many-core processor
  - Vendor: Shanghai High Performance IC Design Center
  - Supported by National Science and Technology Major Project (NMP): Core Electronic Devices, High-end Generic Chips, and Basic Software
  - 28 nm technology
  - 260 Cores
  - 3 Tflop/s peak

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, 74% peak
- 0.32 Pflop/s for HPCG, 0.3% peak
- 15.3 MW, water cooled
  - 6.07 Gflop/s per Watt
- 1.8B RMBs ~ $280M, (building, hw, apps, sw, …)







Gordon Bell Award

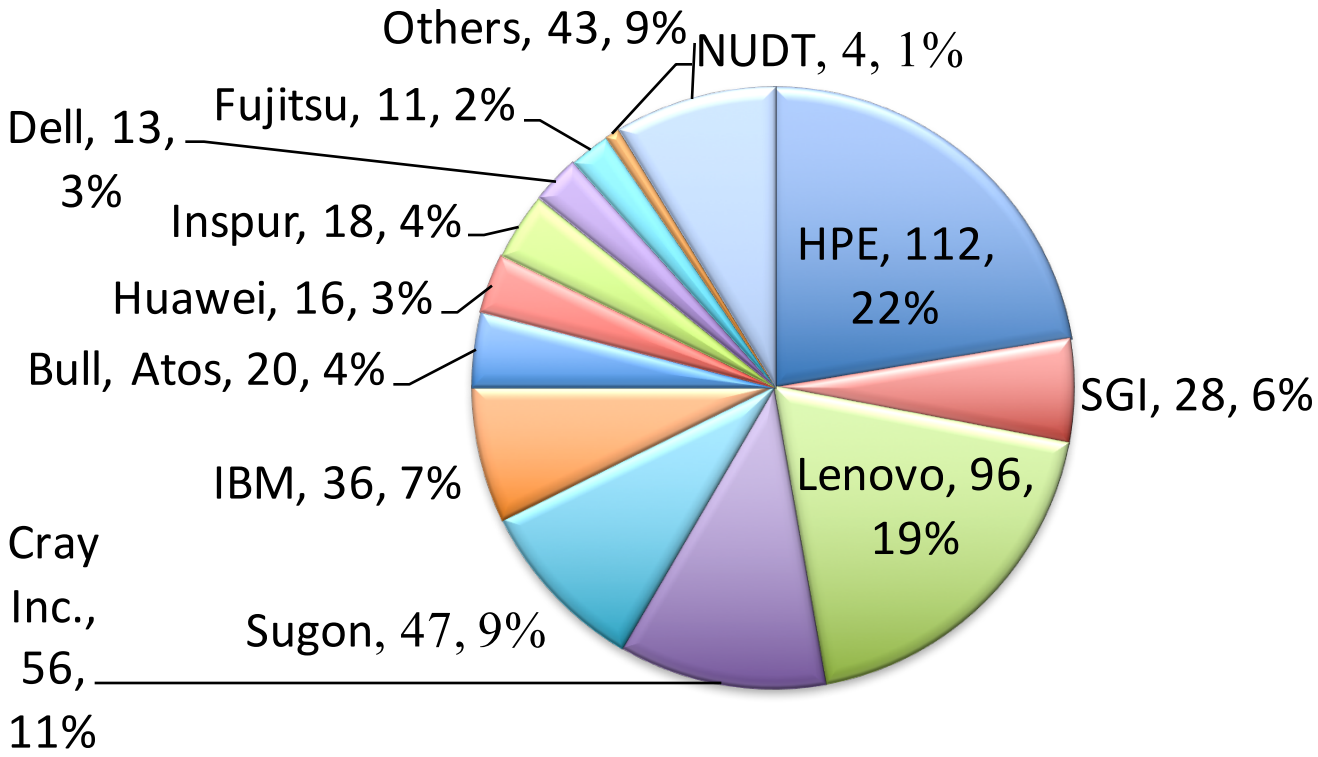
- Since 1987 the Gordon Bell Prize is awarded at the SC conference to recognize outstanding achievement in high-performance computing.
- The purpose of the award is to track the progress of parallel computing, with emphasis on rewarding innovation in applying HPC to applications.
- Financial support 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 for the competition.
- Presentations are made at SC and a winner is chosen.

Gordon Bell Award Finalists at SC16

- **"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)

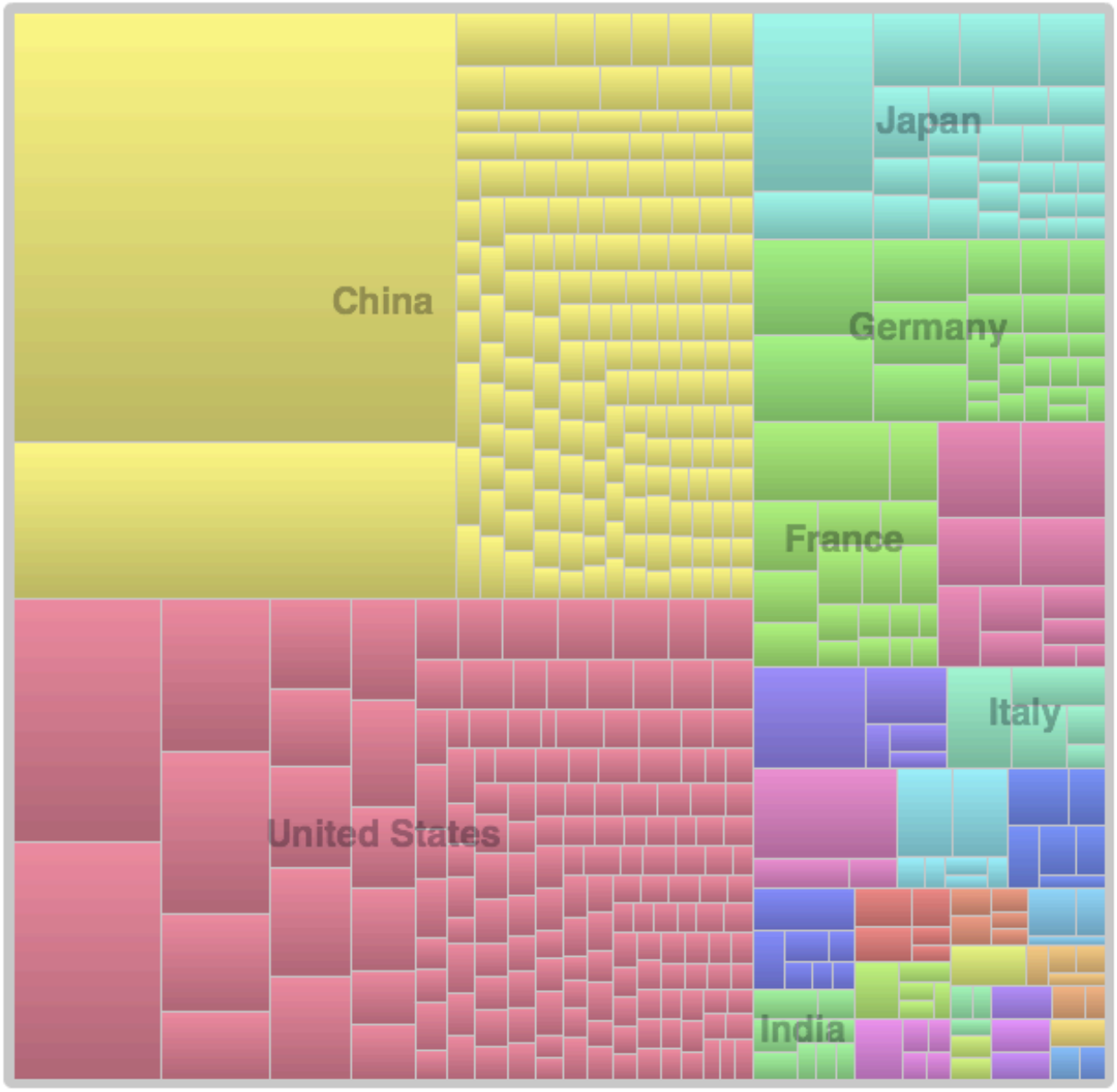
VENDORS / SYSTEM SHARE

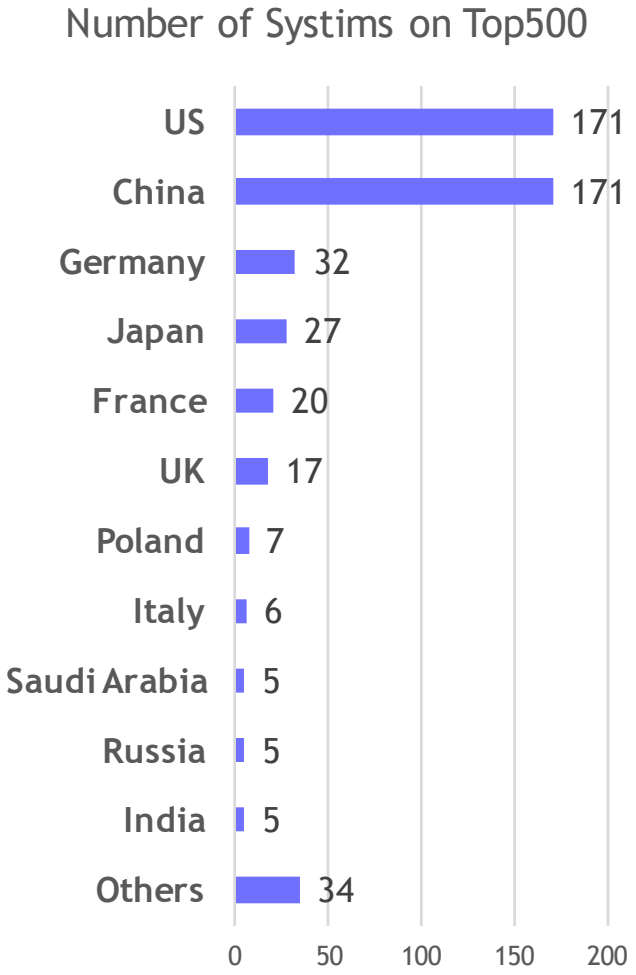


# of systems, % of 500

36% of the Vendors are from China

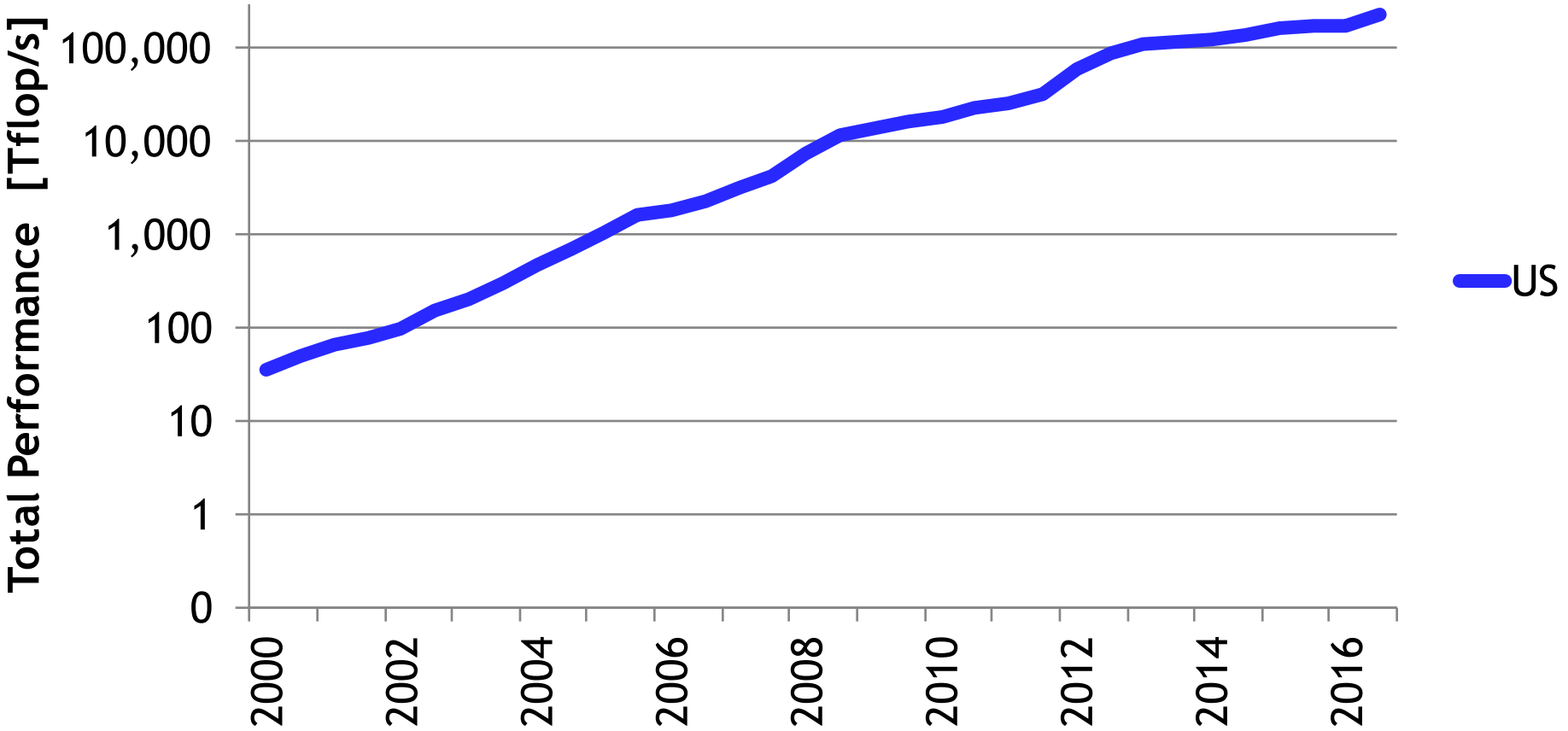
Countries Share



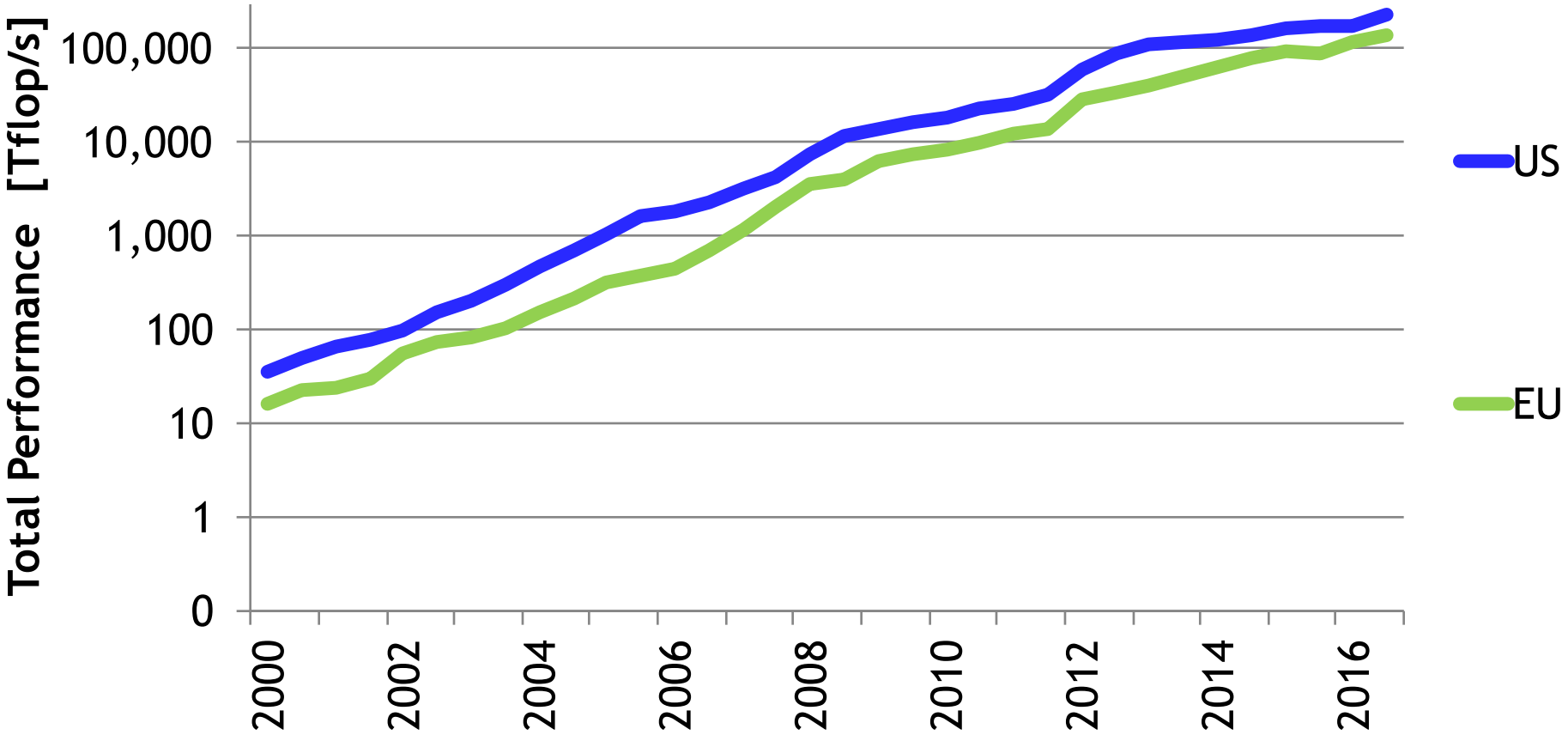


China has 1/3 of the systems, while the number of systems in the US has fallen to the lowest point since the TOP500 list was created.

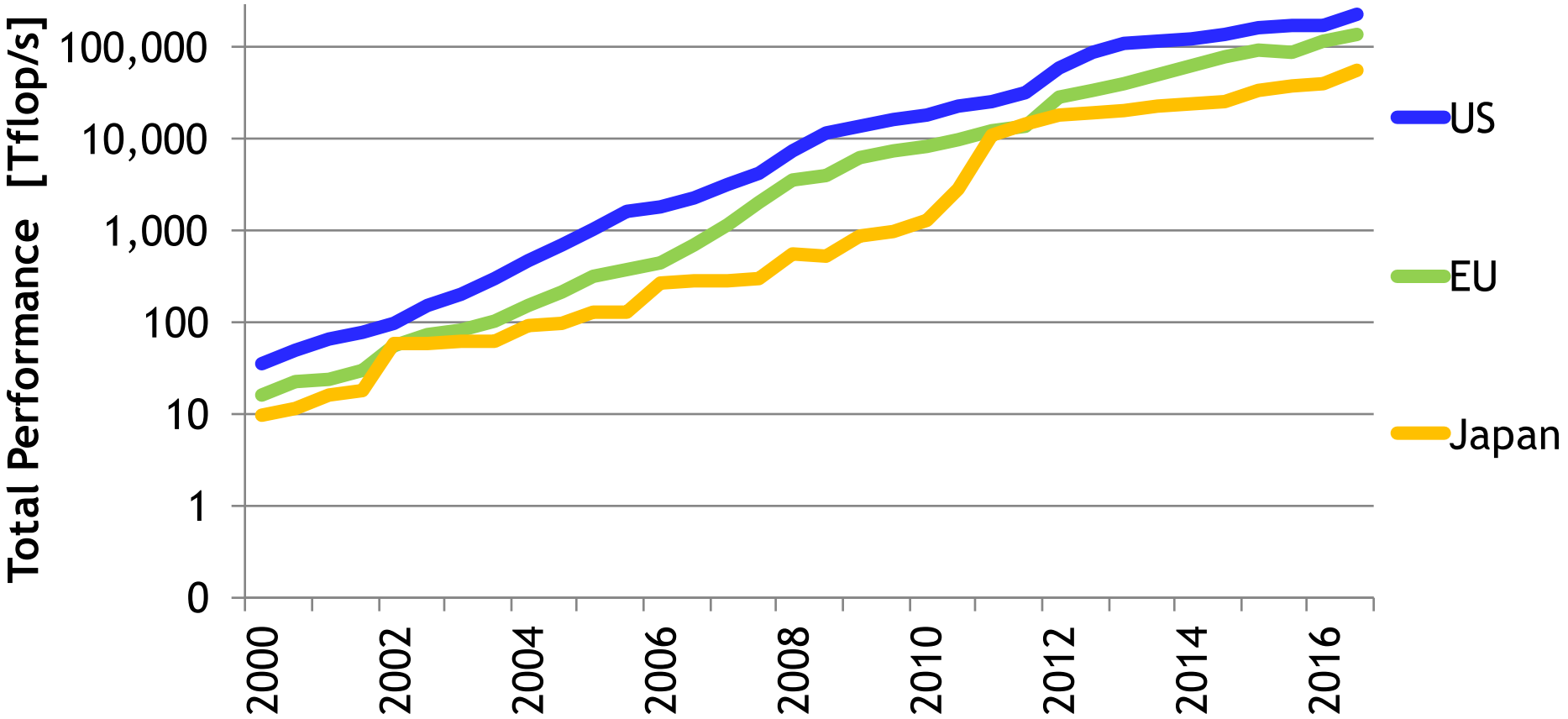
Performance of Countries



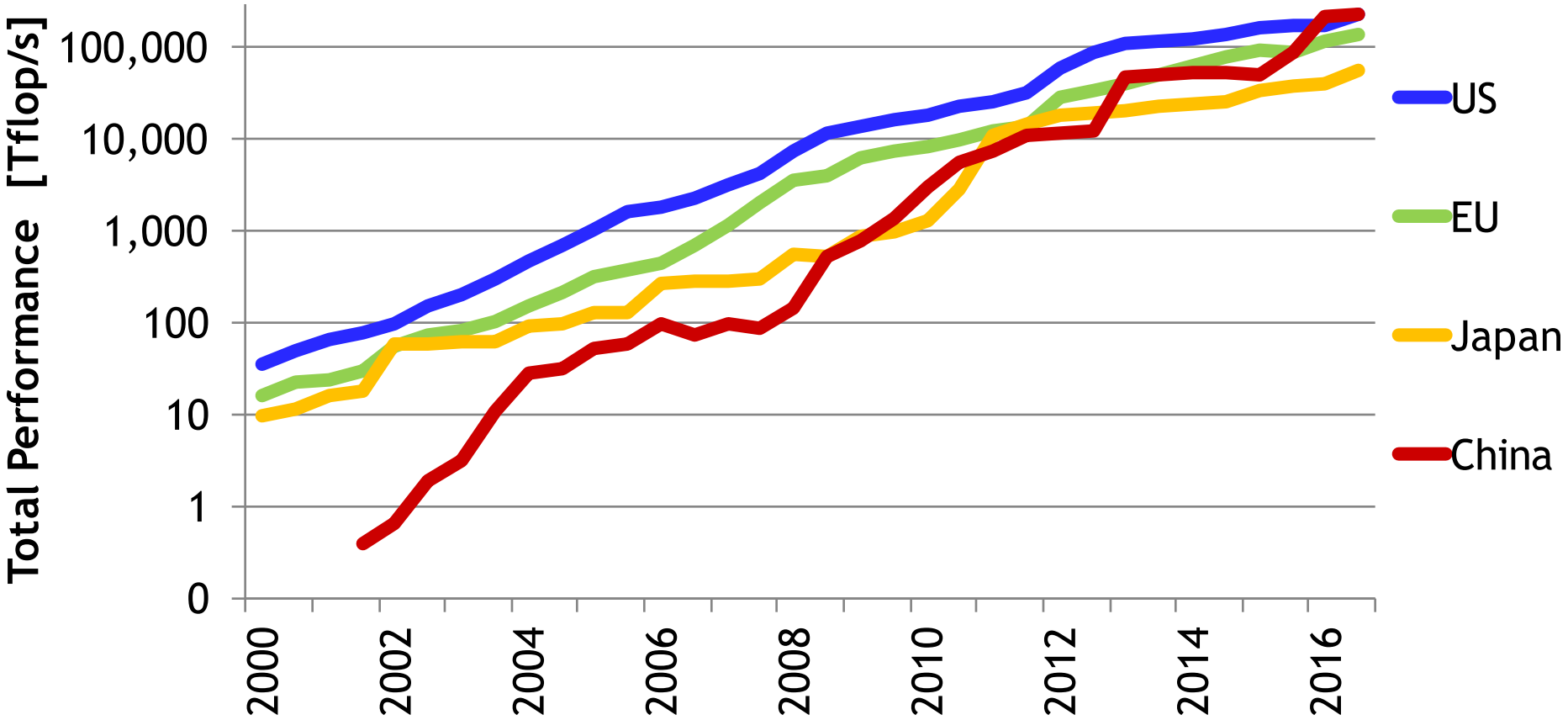
Performance of Countries



Performance of Countries

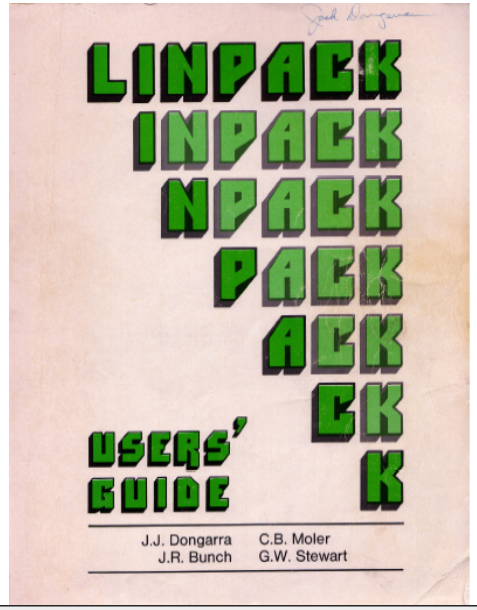


Performance of Countries



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;
  - Cray 1 to DEC PDP-10



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

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.

The Problem

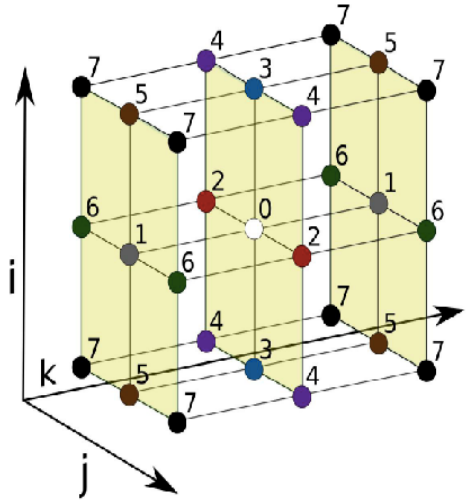
- HPL performance of computer systems are **no longer so strongly correlated to real application performance**, especially for the broad set of HPC applications governed by partial differential equations.

- **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.

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	Rmax Pflops	HPCG Pflops	HPCG /HPL	% of Peak
1	RIKEN Advanced Institute for Computational Science	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	705,024	10.5	0.603	5.7%	5.3%
2	NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.8	0.580	1.7%	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%	2.8%
4	National Supercomputing Center in Wuxi, China	Sunway TaihuLight – Sunway MPP, SW26010	10,649,600	93.0	0.3712	0.4%	0.3%
5	DOE/SC/LBNL/NERSC USA	Cori – XC40, Intel Xeon Phi Cray	632,400	13.8	0.355	2.6%	1.3%
6	DOE/NNSA/LLNL USA	Sequoia – IBM BlueGene/Q, IBM	1,572,864	17.1	0.330	1.9%	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	1.8%	1.2%
8	DOE/NNSA/LANL/SNL	Trinity - Cray XC40, Intel E5-2698v3, Aries custom	301,056	8.10	0.182	2.3%	1.6%
9	NASA / Mountain View	Pleiades - SGI ICE X, Intel E5-2680, E5-2680V2, E5-2680V3, Infiniband FDR	243,008	5.9	0.175	2.9%	2.5%
10	DOE/SC/Argonne National Laboratory	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom	786,432	8.58	0.167	1.9%	1.7%

Peak Performance - Per Core

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

Floating point operations per cycle per core

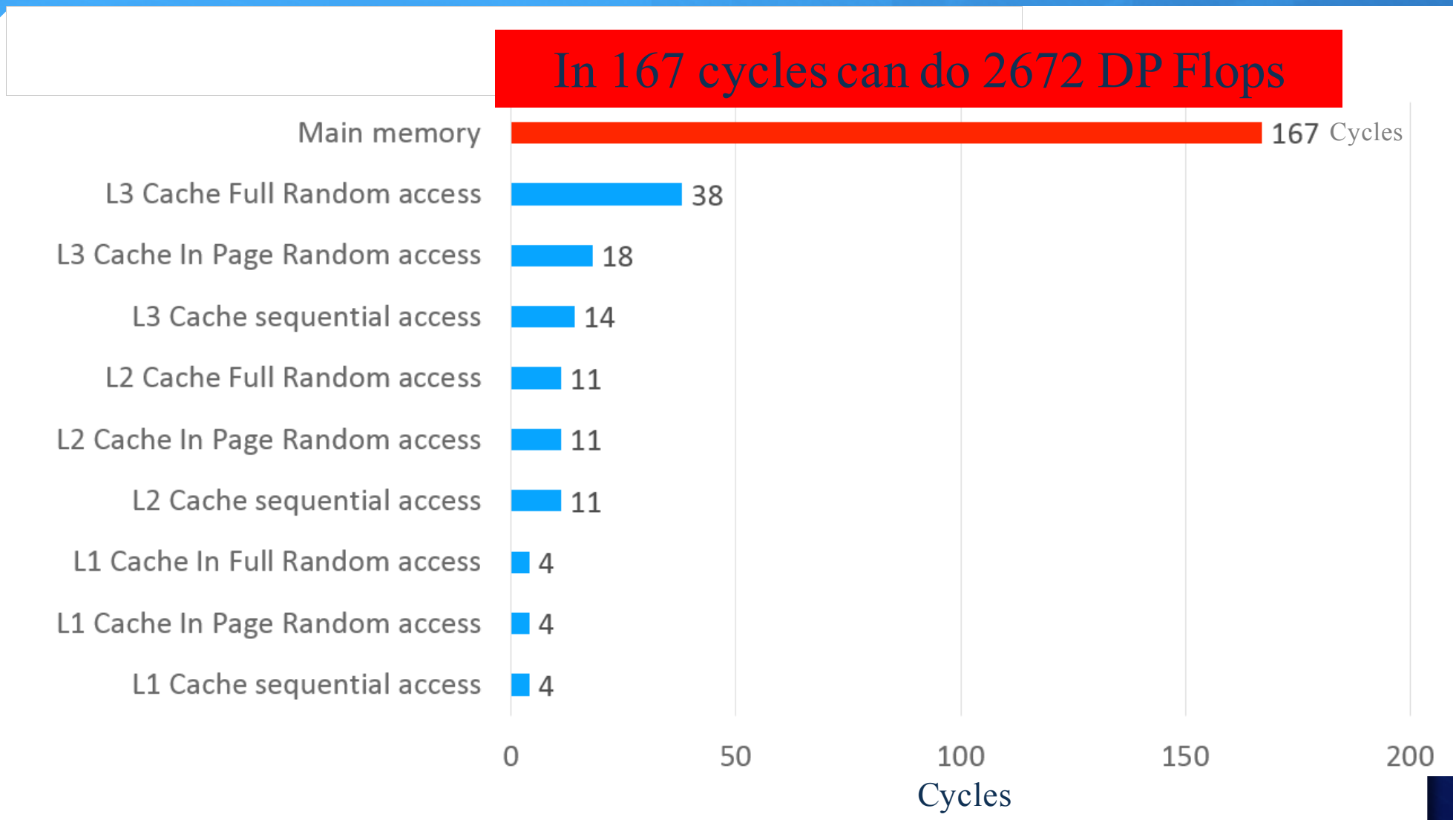
- Most of the recent computers have FMA (Fused multiple add): (i.e. x ←x + y*z in one cycle)
- Intel Xeon earlier models and AMD Opteron have SSE2
  - 2 flops/cycle DP & 4 flops/cycle SP
- Intel Xeon Nehalem ('09) & Westmere ('10) have SSE4
  - 4 flops/cycle DP & 8 flops/cycle SP
- Intel Xeon Sandy Bridge('11)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

We are here (almost)

- Intel Xeon Skylake (server) AVX 512
  - 32 flops/cycle DP & 64 flops/cycle SP
  - Knight's Landing

CPU Access Latencies in Clock Cycles

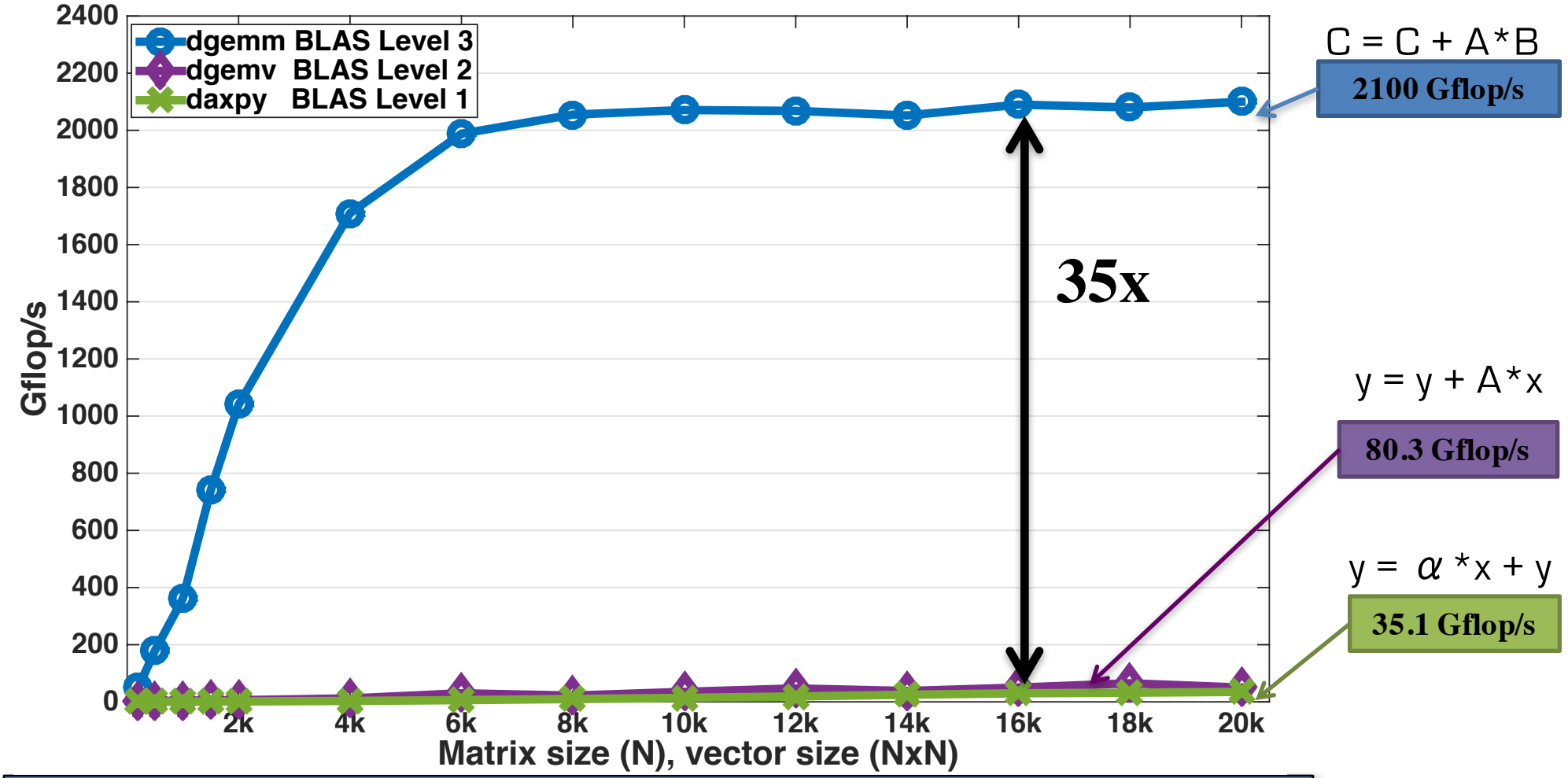


Classical Analysis of Algorithms May Not be Valid

- Processors over provisioned for floating point arithmetic
- Data movement extremely expensive
- Operation count is not a good indicator of the time to solve a problem.
- Algorithms that do more ops may actually take less time.

Level 1, 2 and 3 BLAS

68 cores **Intel Xeon Phi KNL**, 1.3 GHz, Peak DP = 2662 Gflop/s



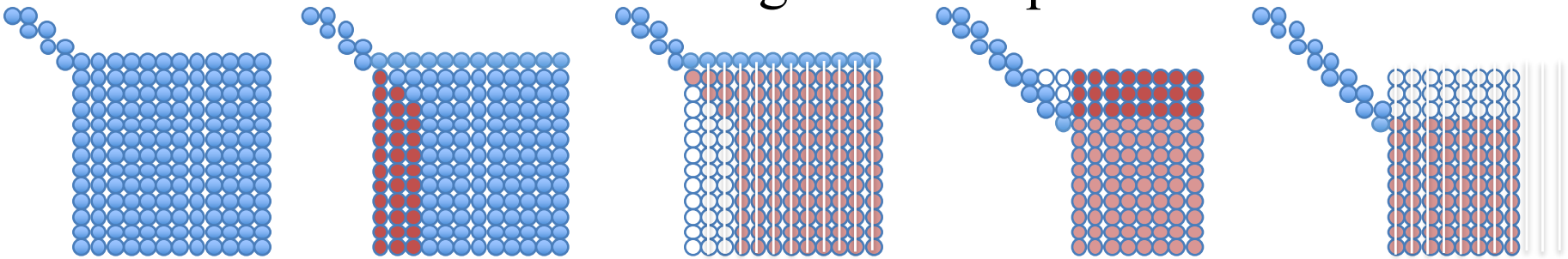
68 cores Intel Xeon Phi KNL, 1.3 GHz
The theoretical peak double precision is 2662 Gflop/s
Compiled with icc and using Intel MKL 2017b1 20160506

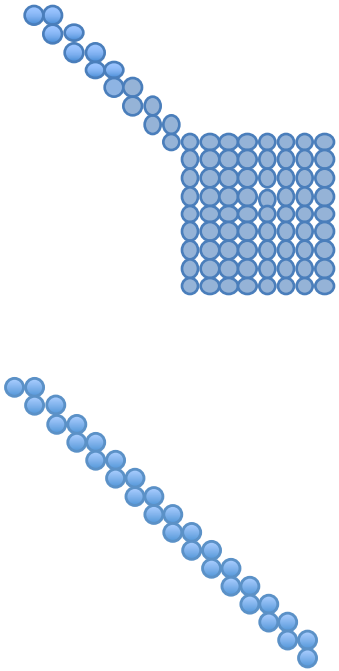
Singular Value Decomposition

LAPACK Version 1991

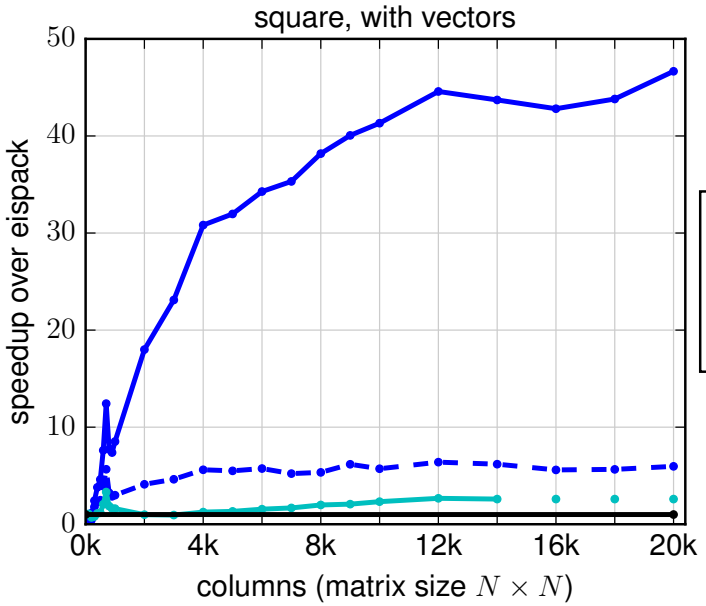
Level 1, 2, & 3 BLAS

First Stage 8/3 n^3 Ops





3 Generations of software compared

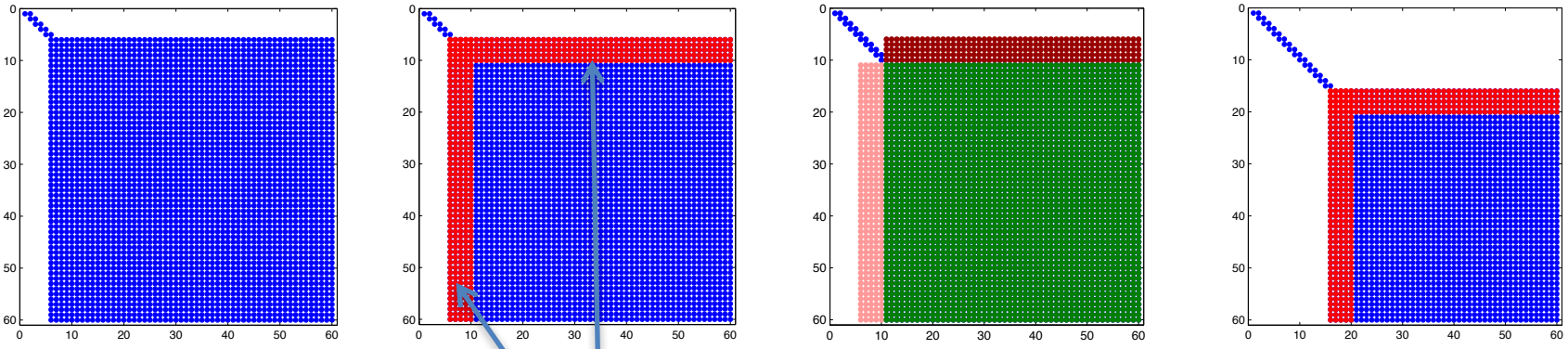


- LAPACK QR (BLAS in ||, 16 cores)
- LAPACK QR (using1 core)(1991)
- LINPACK QR (1979)
- EISPACK QR (1975)

QR refers to the QR algorithm for computing the eigenvalues

Dual socket – 8 core
Intel Sandy Bridge 2.6 GHz
(8 Flops per core per cycle)

Bottleneck in the Bidiagonalization
The Standard Bidiagonal Reduction: xGEBRD
Two Steps: Factor Panel & Update Tailing Matrix



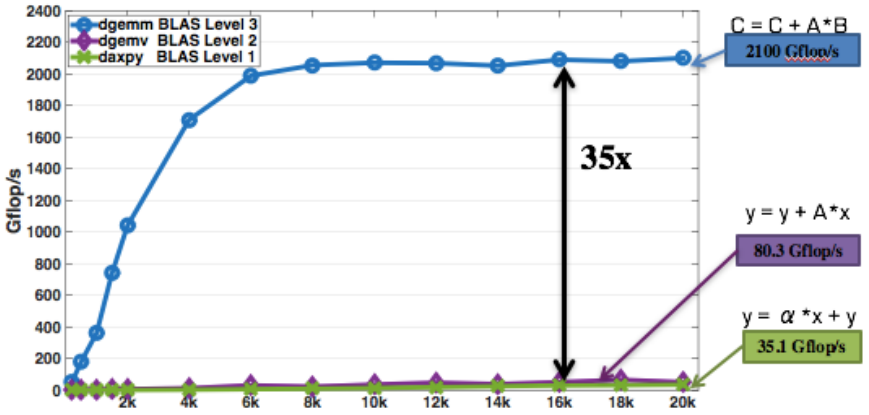
factor panel k then update ➔ **factor panel k+1**

Requires 2 GEMVs

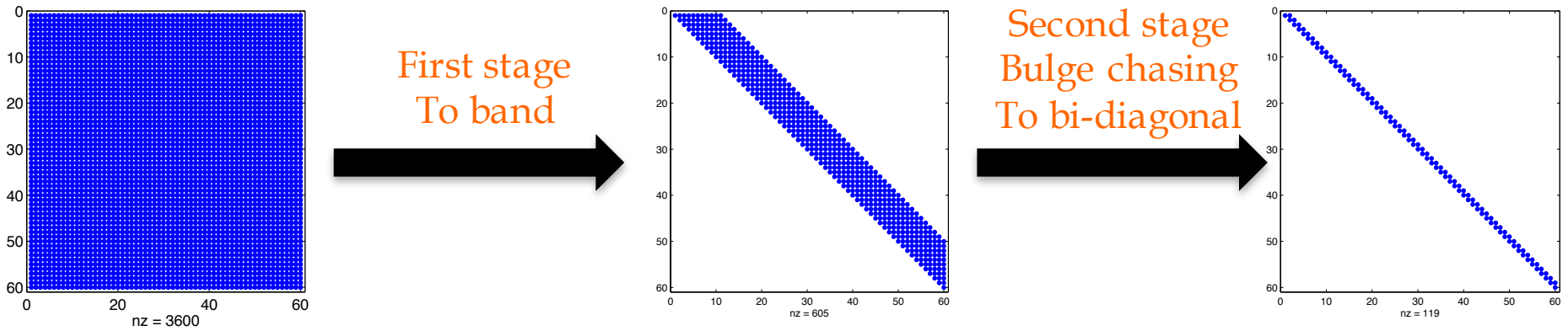
Q*A*P^H

✸ Characteristics

- Total cost $8n^3/3$, (reduction to bi-diagonal)
- Too many Level 2 BLAS operations
- 4/3 n^3 from GEMV and 4/3 n^3 from (
- Performance limited to 2* performa
- ➔**Memory bound algorithm.**



Recent Work on 2-Stage Algorithm

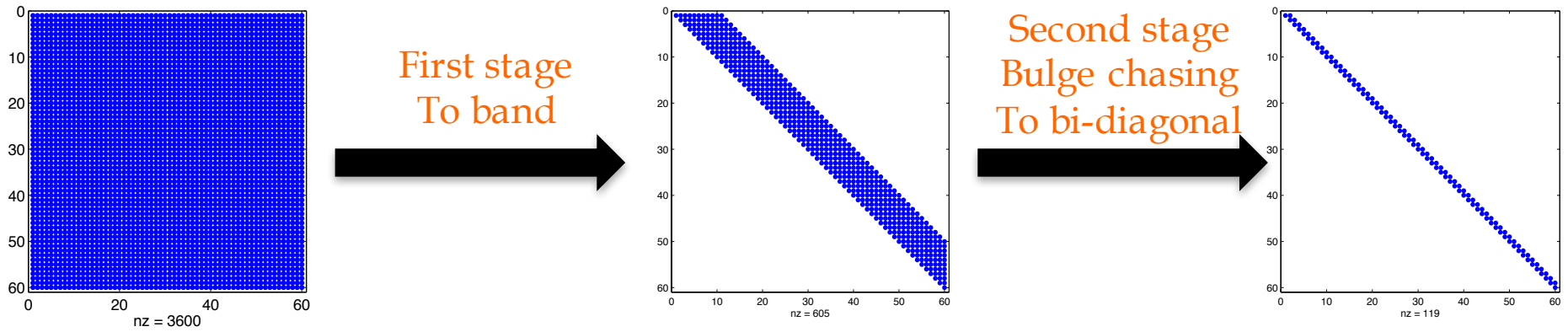


✸ Characteristics

- **Stage 1:**
  - Fully Level 3 BLAS
  - Dataflow Asynchronous execution

- **Stage 2:**
  - Level "BLAS-1.5"
  - Asynchronous execution
  - Cache friendly kernel (reduced communication)

Recent work on developing new 2-stage algorithm



$$\text{flops} \approx \sum_{s=1}^{\frac{n-n_b}{n_b}} 2n_b^3 + (nt-s)3n_b^3 + (nt-s)\frac{10}{3}n_b^3 + (nt-s)\times(nt-s)5n_b^3$$

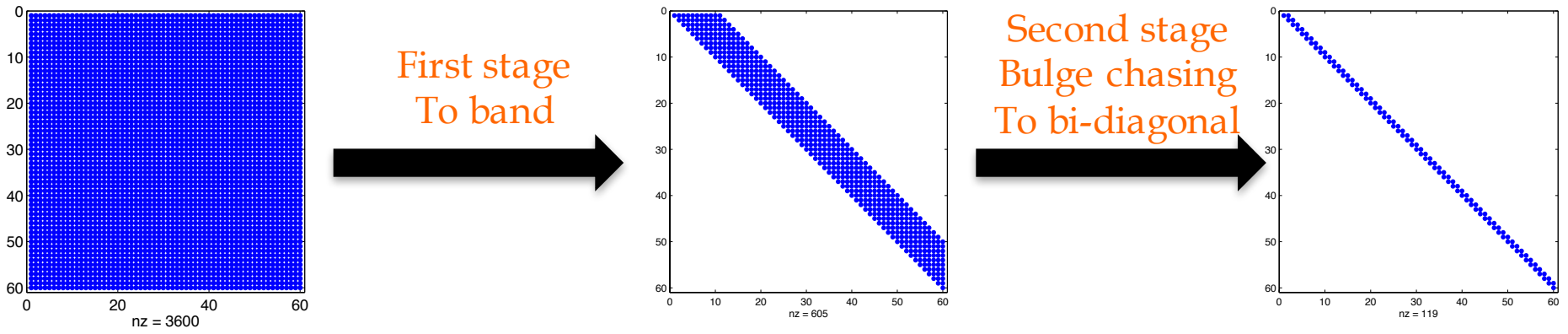
$$+ \sum_{s=1}^{\frac{n-n_b}{n_b}} 2n_b^3 + (nt-s-1)3n_b^3 + (nt-s-1)\frac{10}{3}n_b^3 + (nt-s)\times(nt-s-1)5n_b^3$$

$$\approx \frac{10}{3}n^3 + \frac{10n_b}{3}n^2 + \frac{2n_b}{3}n^3$$

$$\approx \frac{10}{3}n^3(\text{gemm})_{\text{first stage}} \qquad \text{flops} = 6\times n_b \times n^2(\text{gemv})_{\text{second stage}}$$

More Flops, original did 8/3 n^3
25% More flops

Recent work on developing new 2-stage algorithm

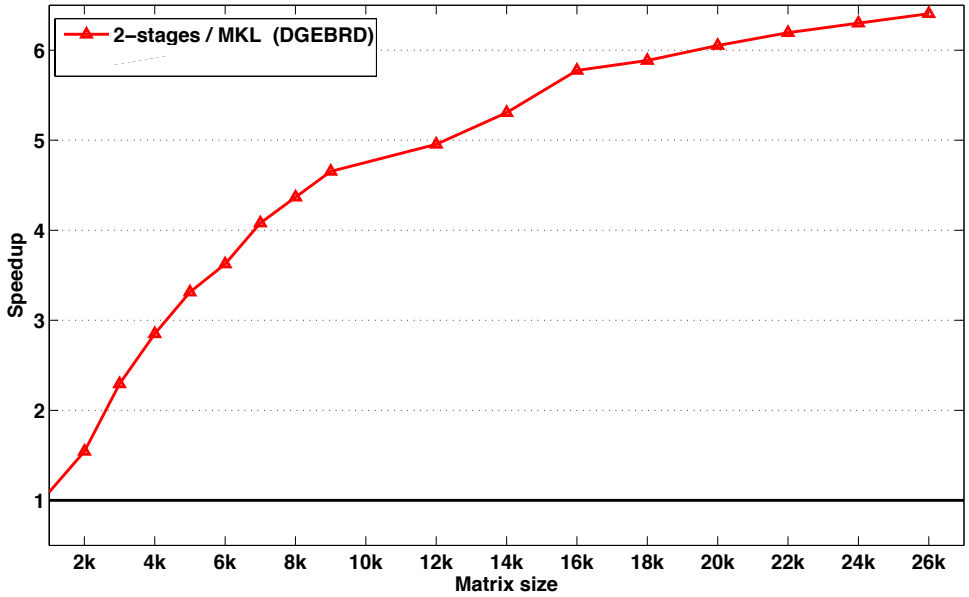


$$\text{speedup} = \frac{\text{time of one-stage}}{\text{time of two-stage}}$$

$$= \frac{4n^3/3P_{gemv} + 4n^3/3P_{gemm}}{10n^3/3P_{gemm} + 6n_b n^2/P_{gemv}}$$

$$\Longrightarrow \frac{84}{70} \leq \text{Speedup} \leq \frac{84}{15}$$

$$\Longrightarrow 1.8 \leq \text{Speedup} \leq 7$$



16 Sandy Bridge cores 2.6 GHz

if P_{gemm} is about 22x P_{gemv} and $120 \leq n_b \leq 240$.

25% More flops and 1.8 – 6 times faster

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.

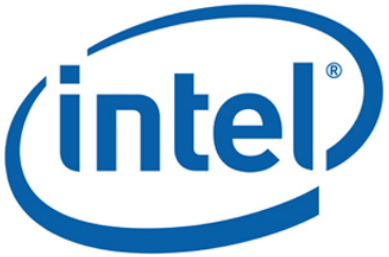
Collaborators and Support

MAGMA team

http://icl.cs.utk.edu/magma

PLASMA team

http://icl.cs.utk.edu/plasma



Collaborating partners

University of Tennessee, Knoxville
Lawrence Livermore National Laboratory, Livermore, CA
University of California, Berkeley
University of Colorado, Denver
INRIA, France (StarPU team)
KAUST, Saudi Arabia

U.S. DEPARTMENT OF ENERGY

Umeå University

INRIA

Rutherford Appleton Laboratory

University of Manchester