NAS

Long Beach 2016

11th IEEE INTERNATIONAL CONFERENCE ON NETWORKING, ARCHITECTURE AND STORAGE

IEEE NAS 2016

Long Beach, CA, USA August 8-10 2016

An Overview of HPC and the Changing Rules at Exascale

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Outline

- Overview of High Performance Computing
- Look at some of the adjustments that are needed with Extreme Computing

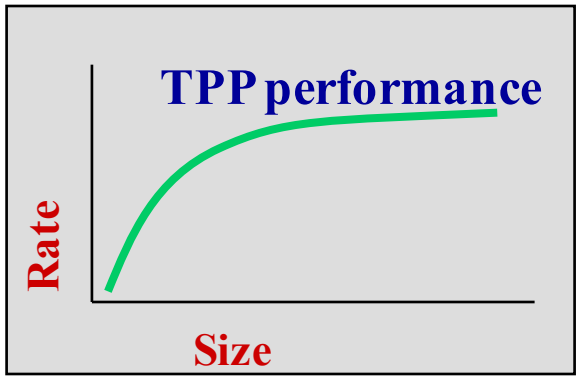
State of Supercomputing Today

- Pflops (> 10^{15} Flop/s) computing fully established with 95 systems.
- Three technology architecture possibilities or "swim lanes" are thriving.
  - Commodity (e.g. Intel)
  - Commodity + accelerator (e.g. GPUs) (93 systems)
  - Special purpose lightweight cores (e.g. ShenWei, ARM, Intel's Knights Landing)
- Interest in supercomputing is now worldwide, and growing in many new markets (around 50% of Top500 computers are used in industry).
- Exascale (10^{18} Flop/s) projects exist in many countries and regions.
- Intel processors have largest share, 91% followed by AMD, 3%.

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

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

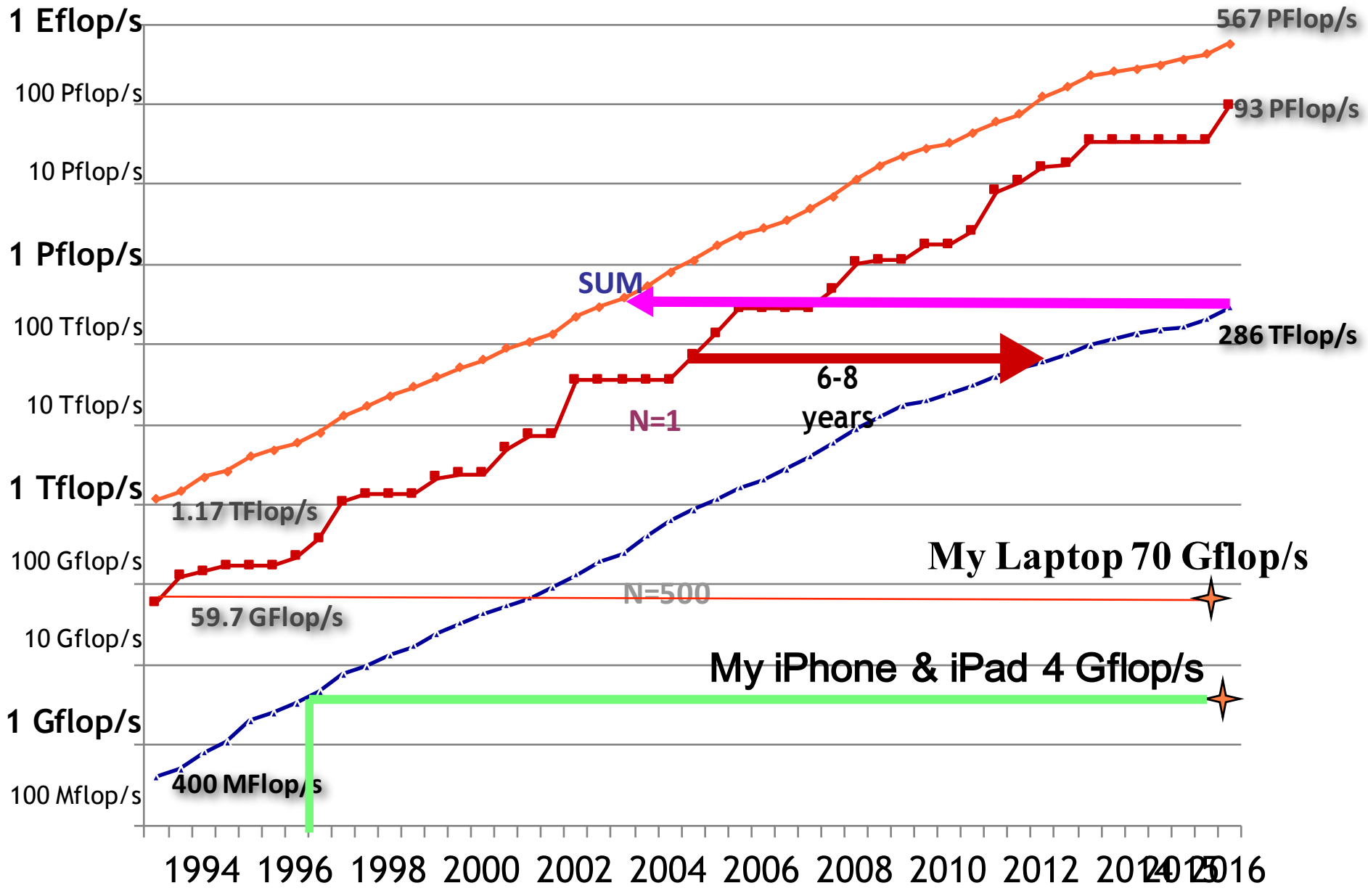
$Ax=b$, *dense problem*



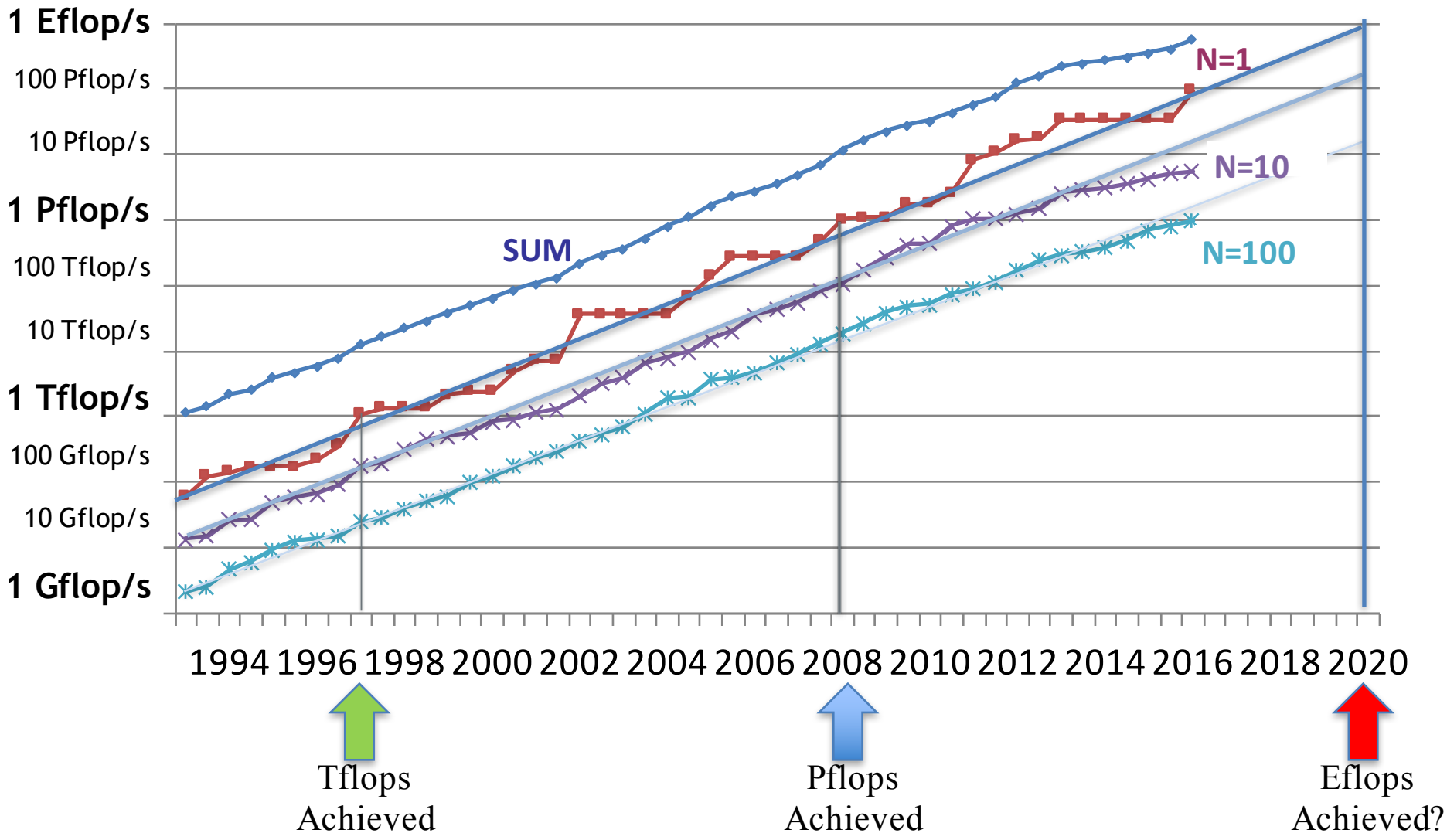
- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June

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

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



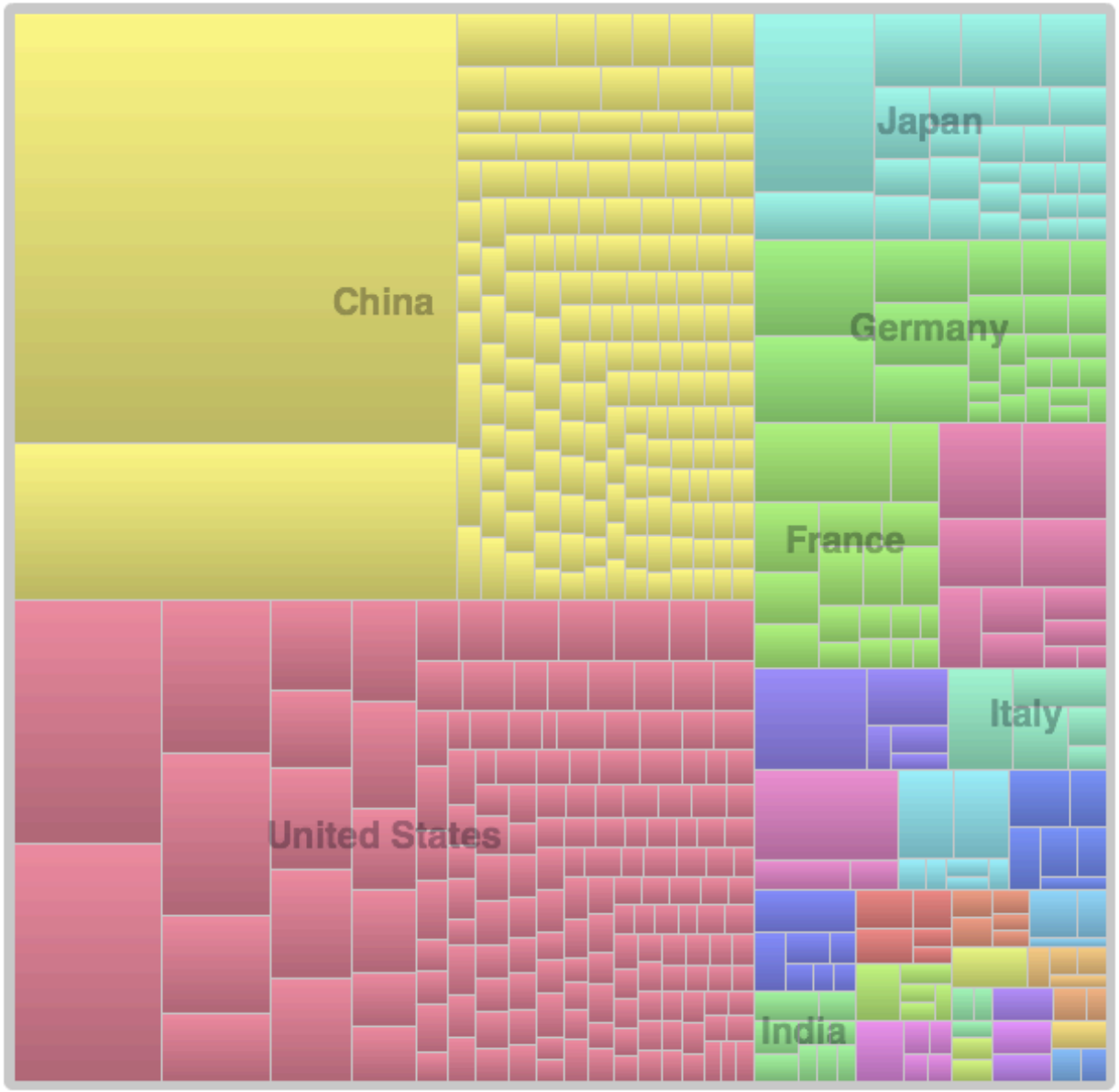
PERFORMANCE DEVELOPMENT



June 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
4	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16C) + custom	USA	1,572,864	17.2	85	7.89	2.18
5	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8C) + Custom	Japan	705,024	10.5	93	12.7	.827
6	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16C) + Custom	USA	786,432	8.16	85	3.95	2.07
7	DOE / NNSA / Los Alamos & Sandia	Trinity, Cray XC40,Xeon (16C) + Custom	USA	301,056	8.10	80	4.23	1.92
8	Swiss CSCS	Piz Daint, Cray XC30, Xeon (8C) + Nvidia Kepler (14c) + Custom	Swiss	115,984	6.27	81	2.33	2.69
9	HLRS Stuttgart	Hazel Hen, Cray XC40, Xeon (12C) + Custom	Germany	185,088	5.64	76	3.62	1.56
10	KAUST	Shaheen II, Cray XC40, Xeon (16C) + Custom	Saudi Arabia	196,608	5.54	77	2.83	1.96
500	Internet company	Inspur Intel (8C) + Nnvidia	China	5440	.286	71		

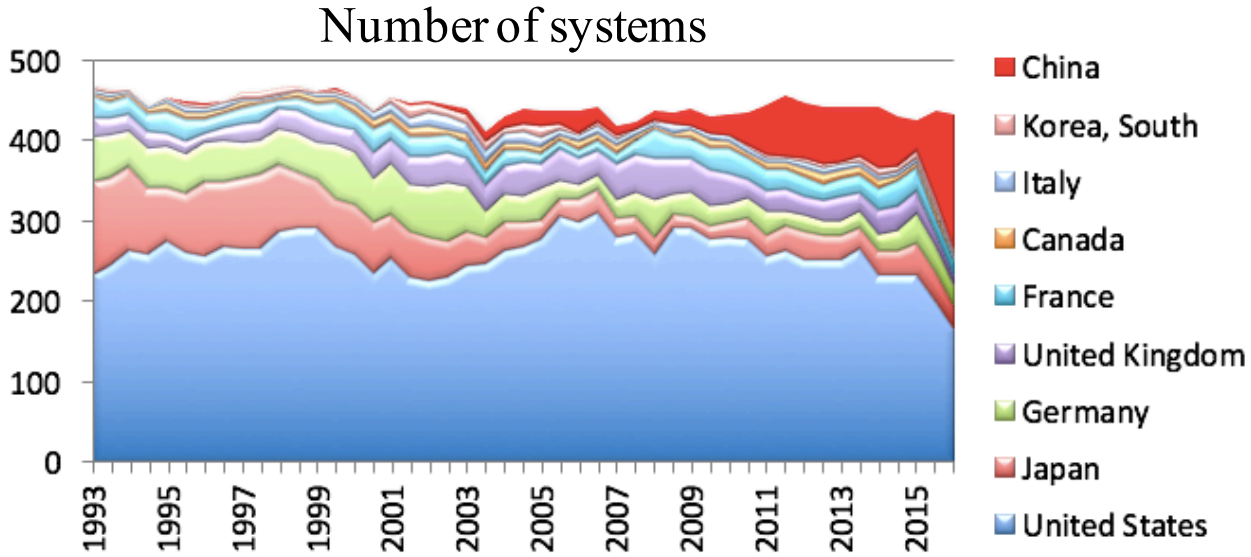
Countries Share

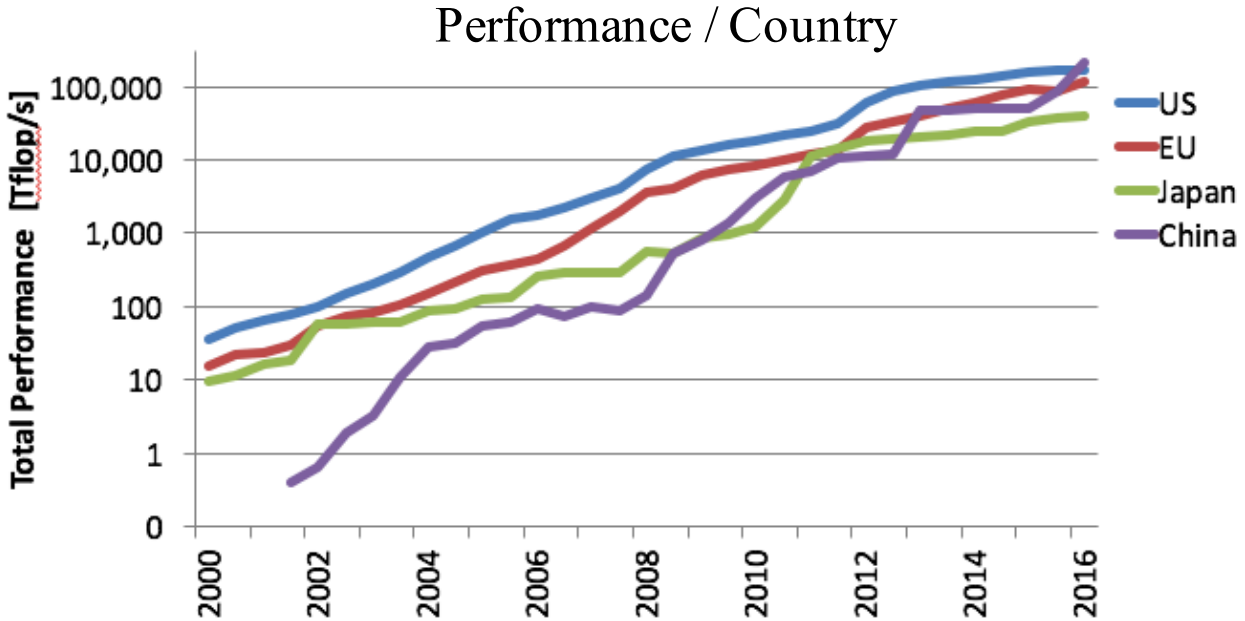


COUNTRY	NUMBER OF SUPERCOMPUTERS
China	**167**
United States	165
Japan	29
Germany	26
France	18
Britain	12
India	9
Russia	7
South Korea	7
Poland	6
other	54

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.

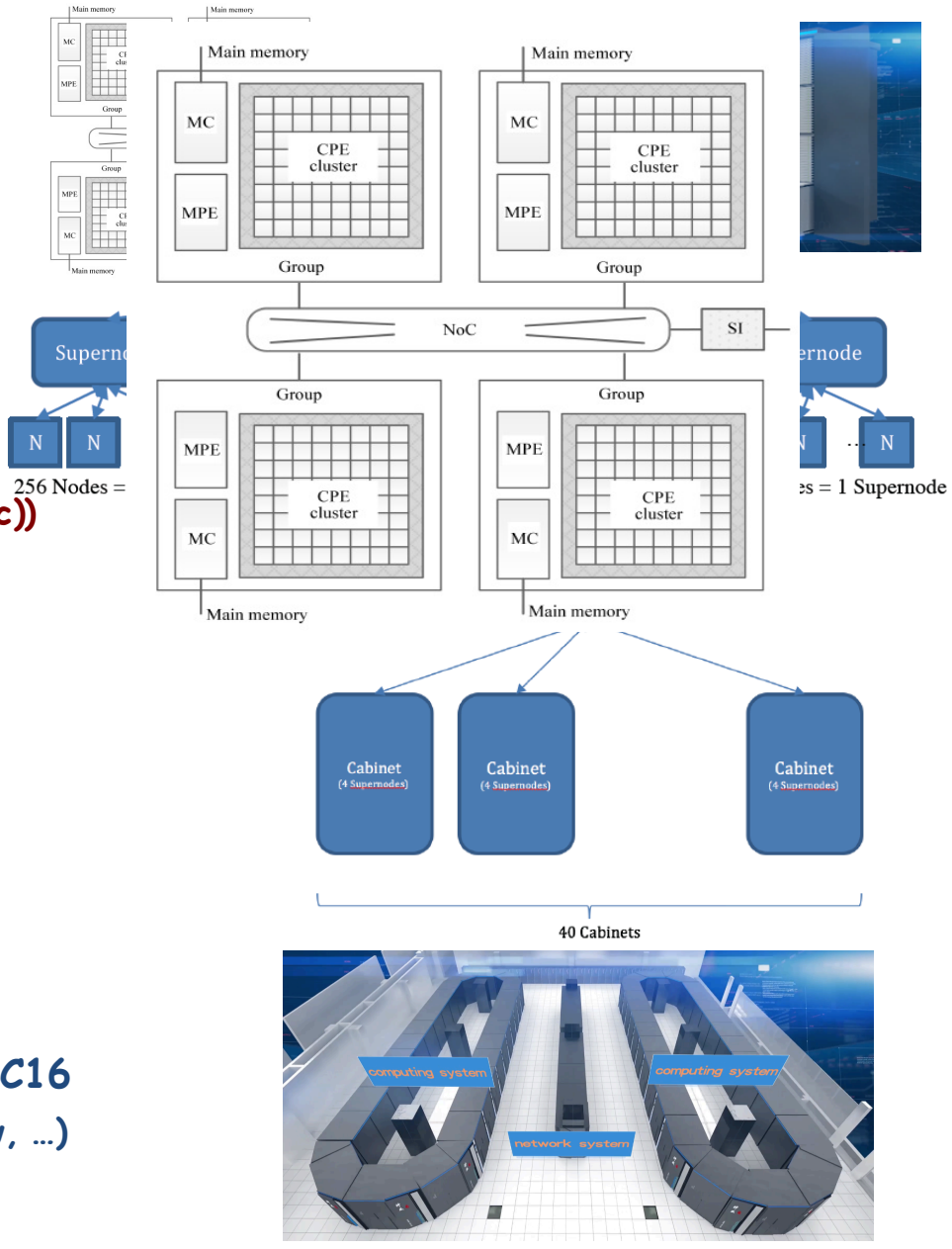
Countries Share





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/CG
    - 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 node/c))
  - ~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 HPL, 74% peak
- 0.32 Pflop/s HPCG, 0.3% peak
- 15.3 MW, water cooled
  - 6.07 Gflop/s per Watt
- 3 of the 6 finalists Gordon Bell Award@SC16
- 1.8B RMBs ~ $270M, (building, hw, apps, sw, …)



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

hpcg-benchmark.org

HPCG Snapshot

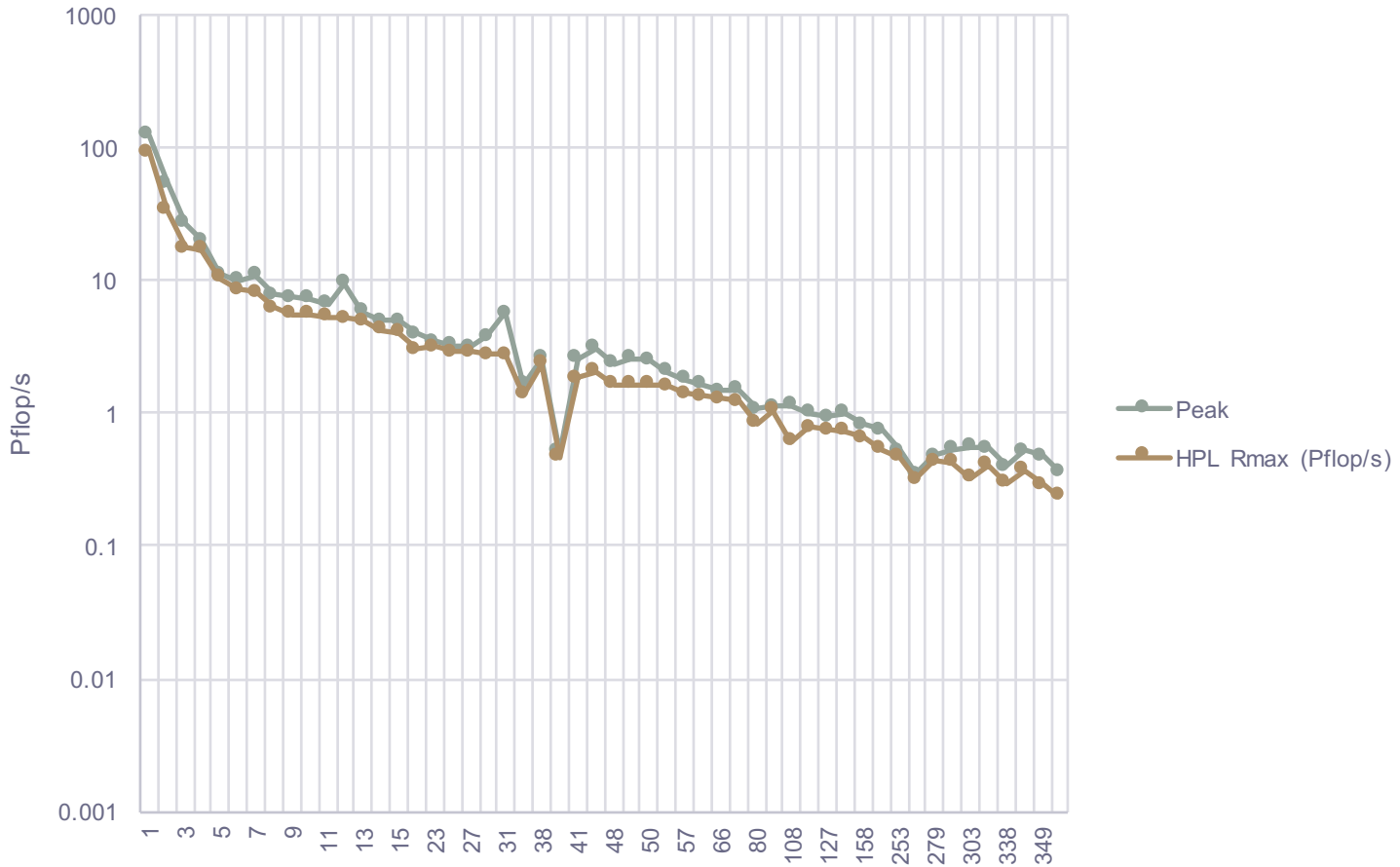
- High Performance Conjugate Gradients (HPCG).
- Solves $Ax=b$, A large, sparse, b known, x computed.
- An optimized implementation of PCG contains essential computational and communication patterns that are prevalent in a variety of methods for discretization and numerical solution of PDEs

- 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).

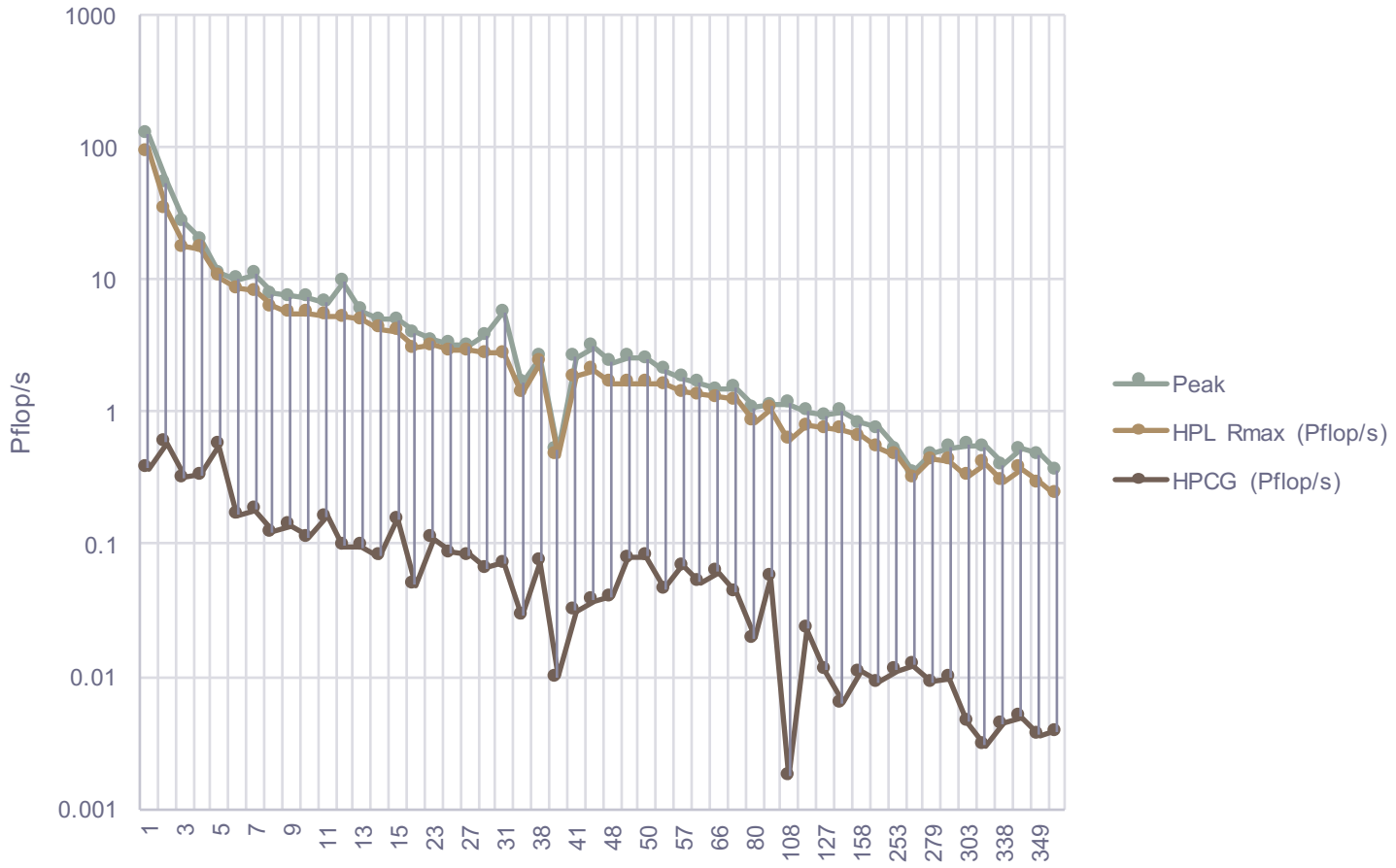
HPCG with 80 Entries

Rank (HPL)	Site	Computer	Cores	Rmax	HPCG	HPCG / HPL	% of Peak
1 (2)	NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.86	0.580	1.7%	1.1%
2 (5)	RIKEN AICS	K computer, SPARC64 VIIIfx 2.0GHz, custom	705,024	10.51	0.554	5.3%	4.9%
3 (1)	NCSS / Wuxi	Sunway TaihuLight -- SW26010, Sunway	10,649,600	93.01	0.371	0.4%	0.3%
4 (4)	DOE NNSA / LLNL	Sequoia - IBM BlueGene/Q + custom	1,572,864	17.17	0.330	1.9%	1.6%
5 (3)	DOE SC / ORNL	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, custom, NVIDIA K20x	560,640	17.59	0.322	1.8%	1.2%
6 (7)	DOE NNSA / LANL& SNL	Trinity - Cray XC40, Intel E5-2698v3, + custom	301,056	8.10	0.182	2.3%	1.6%
7 (6)	DOE SC / ANL	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, + Custom	786,432	8.58	0.167	1.9%	1.7%
8 (11)	TOTAL	Pangea -- Intel Xeon E5-2670, Ifb FDR	218592	5.28	0.162	3.1%	2.4%
9 (15)	NASA / Mountain View	Pleiades - SGI ICE X, Intel E5-2680, E5-2680V2, E5-2680V3 + Ifb	185,344	4.08	0.155	3.8%	3.1%
10 (9)	HLRS / U of Stuttgart	Hazel Hen - Cray XC40, Intel E5-2680v3, + custom	185,088	5.64	0.138	2.4%	1.9%

Bookends: Peak, HPL, and HPCG



Bookends: Peak, HPL, and HPCG



Apps Running on Sunway TaihuLight

Table 4 Summary of the major applications on the Sunway TaihuLight, compared with similar applications on other large-scale systems

Category	System	Application summary	Scale of run	Performance
Non-linear solver	Sunway TaihuLight	A fully-implicit nonhydrostatic dynamic for cloud-resolving atmospheric simulation	131072 MPEs and 8388608 CPEs	1.5 PFlops
	Sequoia	An implicit solver for complex PDEs in highly heterogeneous flow in Earth's mantle [3]	1572864 cores	687 TFlops
Molecular dynamics	Sunway TaihuLight	Atomic simulation of silicon nanowires	131072 MPEs and 8388608 CPEs	14.7 PFlops
	Tianhe-1A	Molecular dynamics simulation of crystalline silicon [20]	7168 GPUs (3211264 CUDA cores)	1.87 PFlops
Phase-field simulation	Sunway TaihuLight	Coarsening dynamics based on Cahn-Hilliard equation with degenerated mobility	131072 MPEs and 8388608 CPEs	39.678 PFlops
	Tsubame 2.0	Dendritic solidification [6]	16000 CPU cores and 4000 GPUs (1792000 CUDA cores)	1.017 PFlops

07

Peak Performance - Per Core

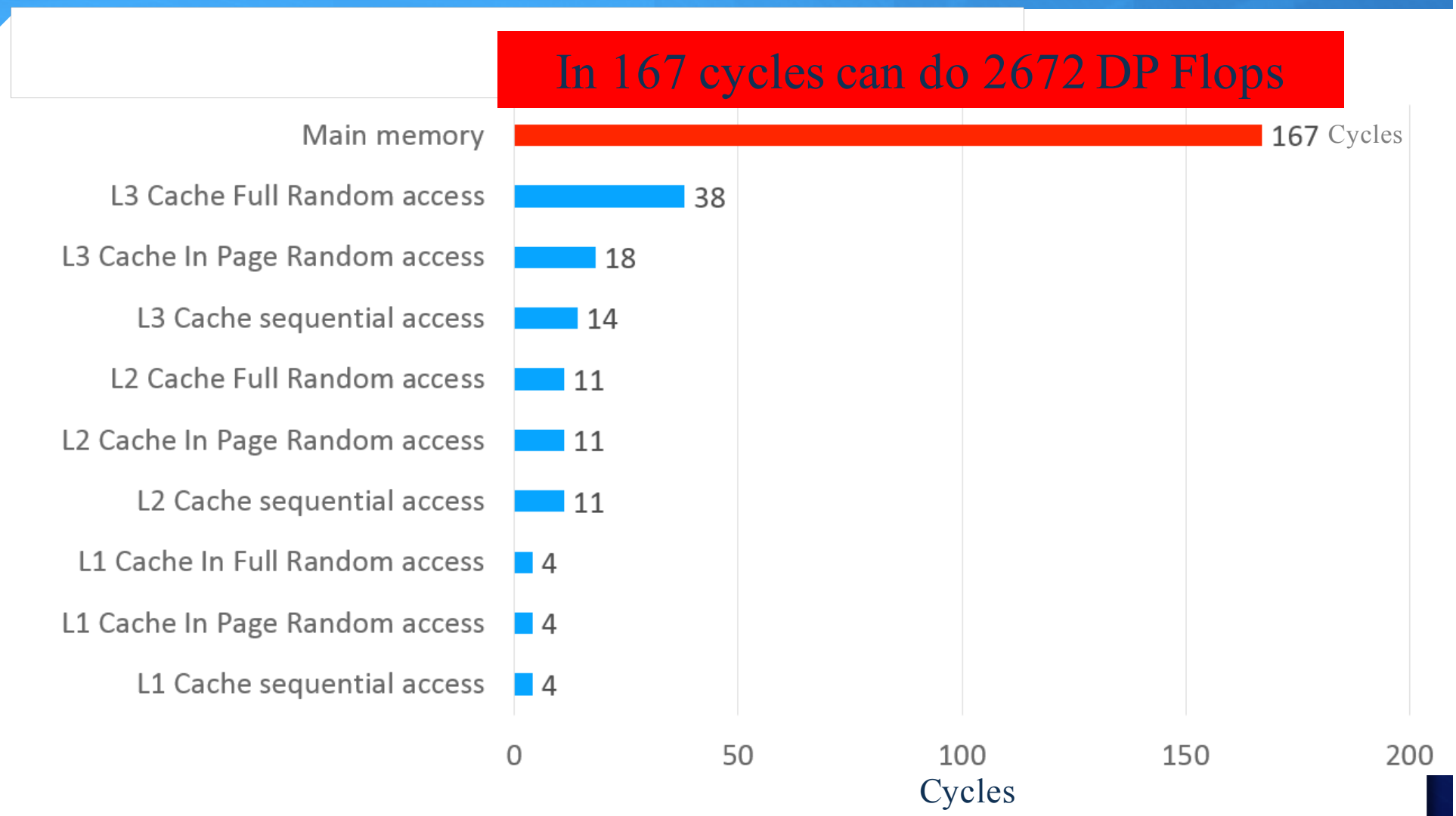
$$\text{FLOPS} = \text{cores} \times \text{clock} \times \frac{\text{FLOPs}}{\text{cycle}}$$

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 (almost)

CPU Access Latencies in Clock Cycles



Classical Analysis of Algorithms May Not be Valid

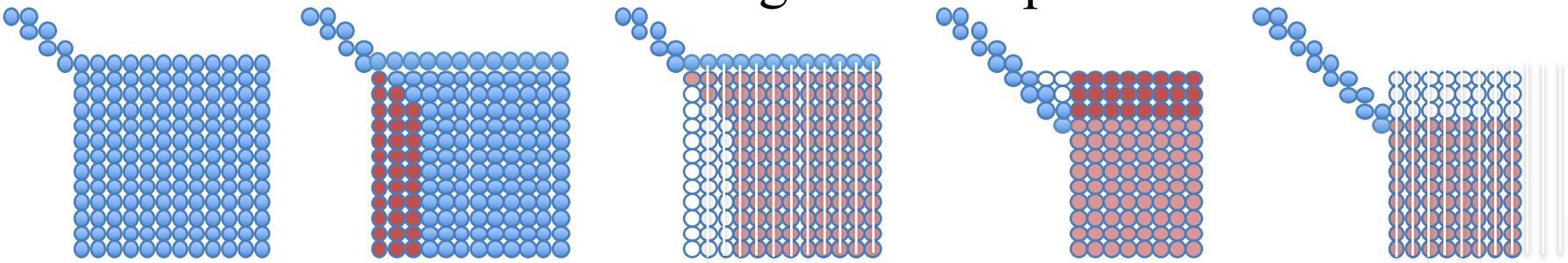
- Processors over provisioned for floating point arithmetic
- Data movement extremely expensive
- Operation count is not a good indicator of the time to solve a problem.
- Algorithms that do more ops may actually take less time.

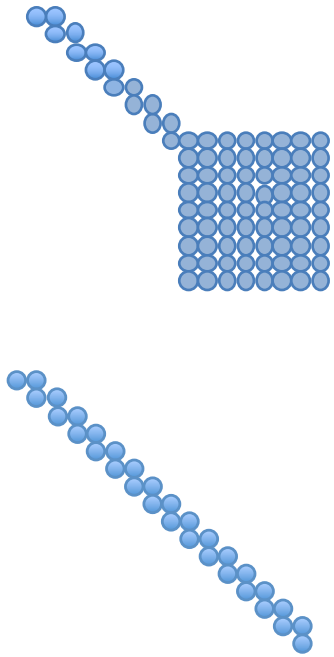
Singular Value Decomposition

LAPACK Version 1991

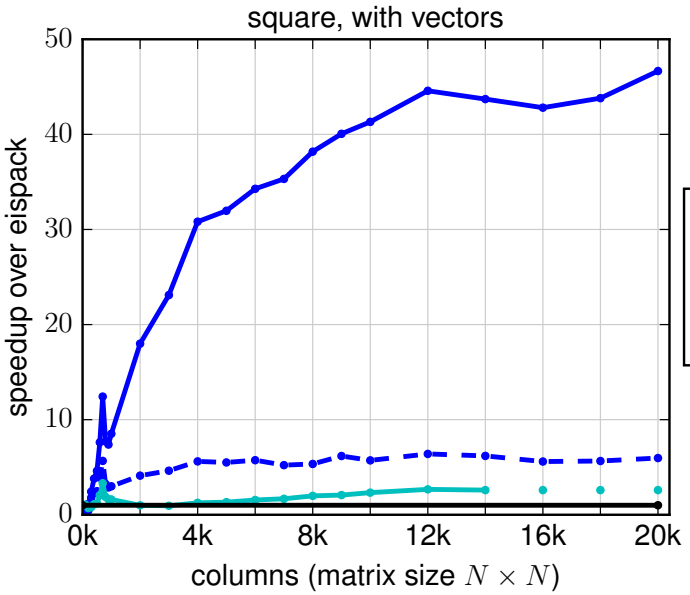
Level 1, 2, & 3 BLAS

First Stage 8/3 n^3 Ops





3 Generations of software compared

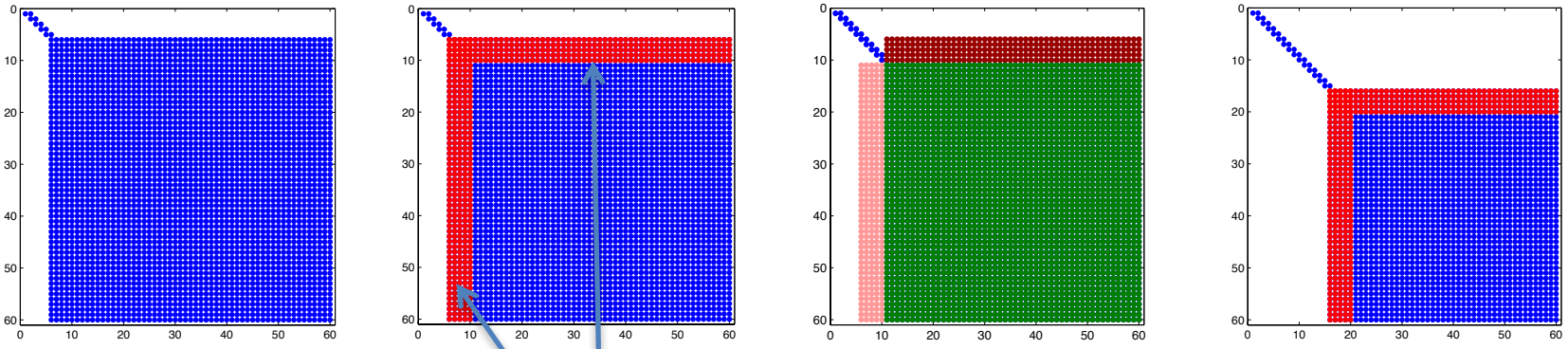


LAPACK QR (BLAS in ||, 16 cores)
LAPACK QR (using1 core)(1991)
LINPACK QR (1979)
EISPACK QR (1975)

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



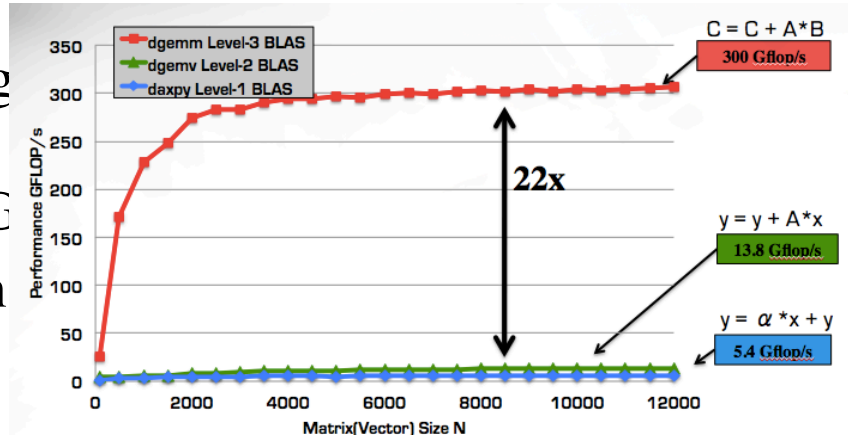
factor panel k then update ➔ **factor panel k+1**

Requires 2 GEMVs

Q*A*P^H

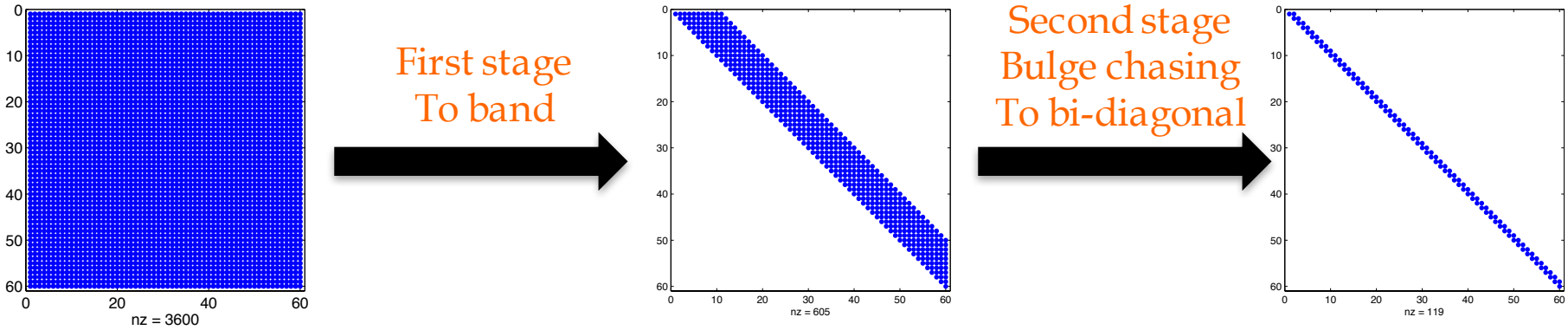
✸ Characteristics

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



16 cores Intel Sandy Bridge, 2.6 GHz, 20 MB shared L3 cache.
The theoretical peak per core double precision is 20.4 Gflop/s per core.
Compiled with icc and using MKL 2015.3.187

Recent Work on 2-Stage Algorithm

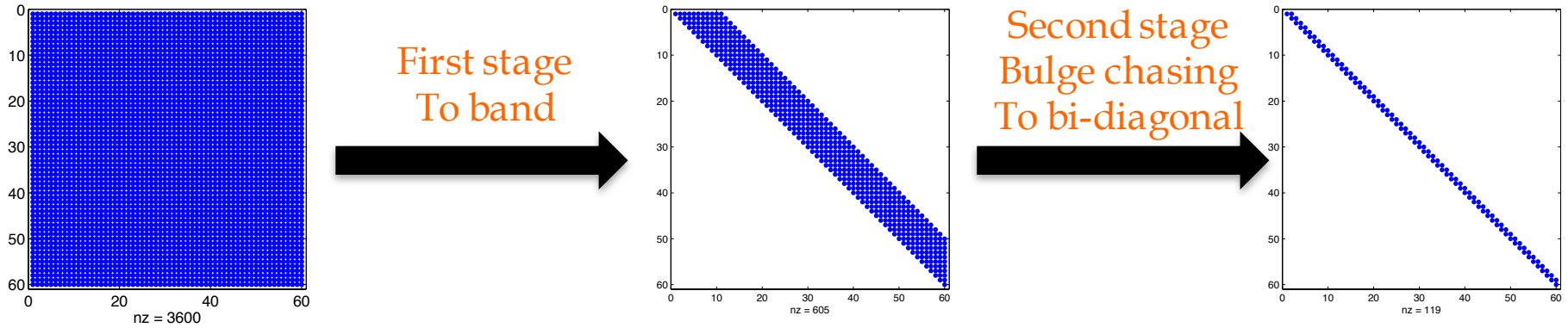


✸ Characteristics

- **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



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

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

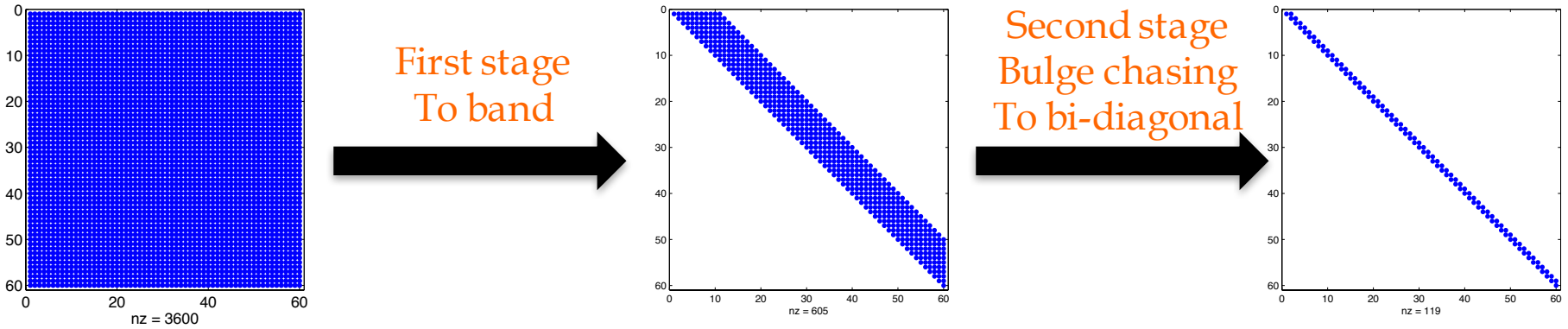
$$\approx \frac{10}{3}n^3 + \frac{10n_b}{3}n^2 + \frac{2n_b}{3}n^3$$

$$\approx \frac{10}{3}n^3(\text{gemm})_{\text{first stage}}$$

$$\text{flops} = 6\times n_b \times n^2(\text{gemv})_{\text{second stage}}$$

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

Recent work on developing new 2-stage algorithm

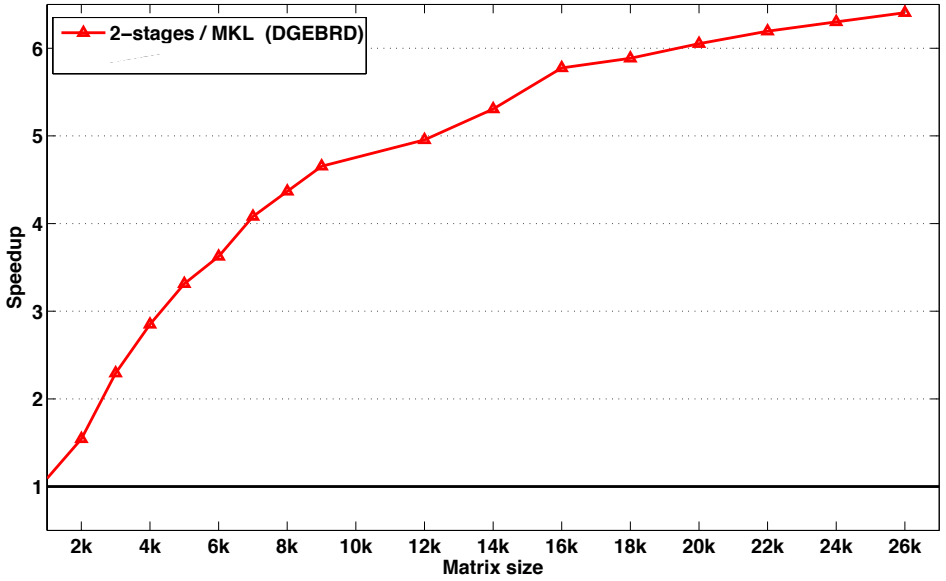


$$\text{speedup} = \frac{\text{time of one-stage}}{\text{time of two-stage}}$$

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

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

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



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

Umeå University
INRIA
Rutherford Appleton Laboratory
University of Manchester