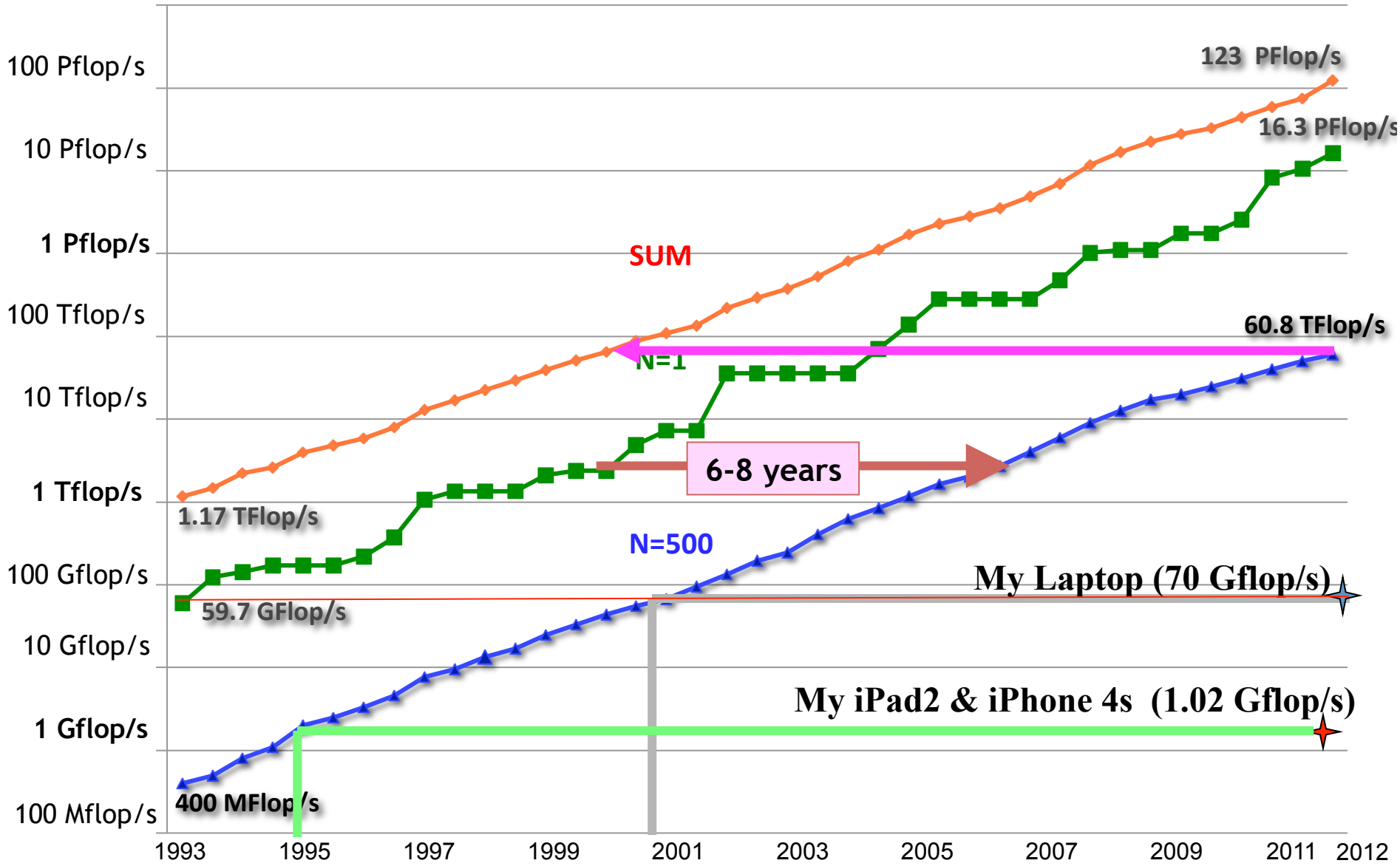
EXPERIENCE ON MULTI-PETAFLOPS ARCHITECTURES

Jack Dongarra
UTK/ORNL

Over Last 20 Years - Performance Development



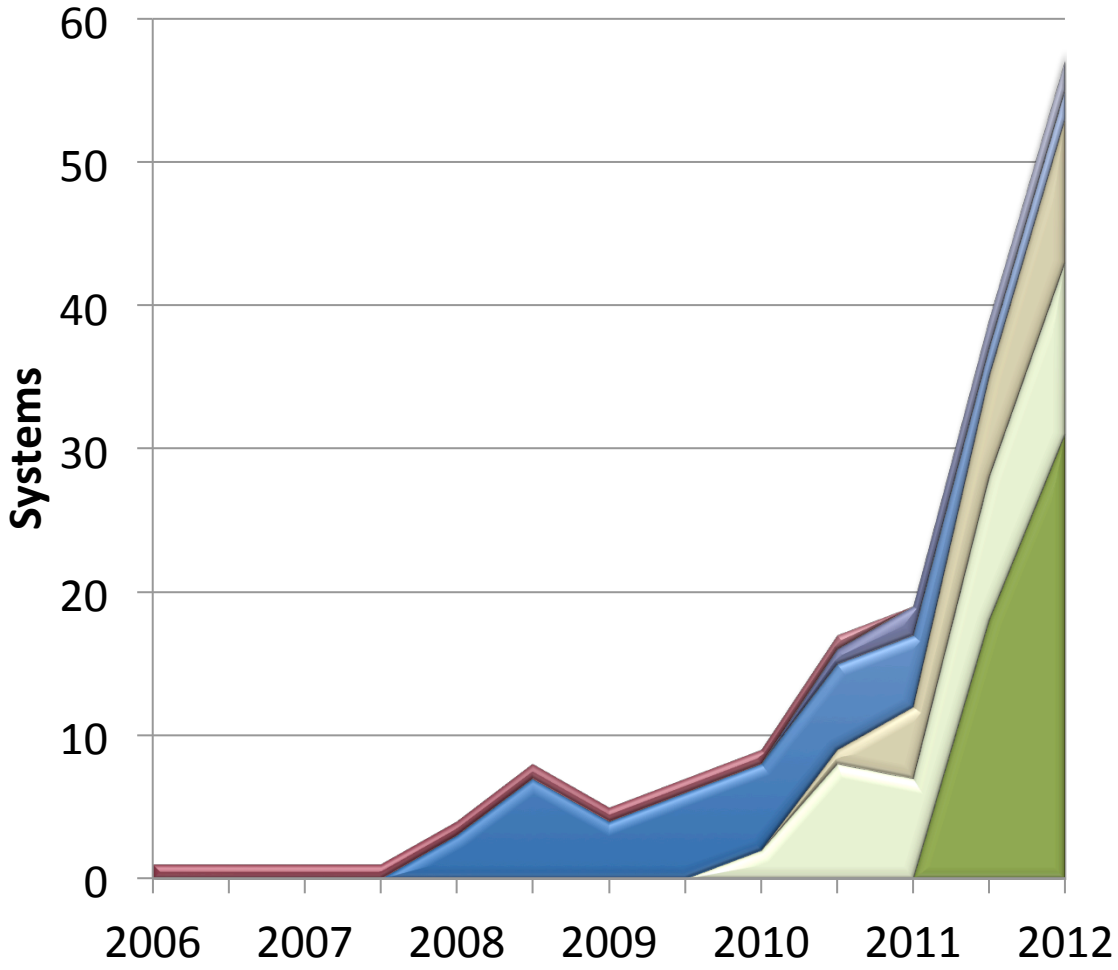
#1 System on the Top500 Over the Past 20 Years (15 machines in that club)

Top500 List	Computer	r_max (Gflop/s)	n_max	Hours
6/93 (1)	TMC CM-5/1024	60	52224	0.4
11/93 (1)	Fujitsu Numerical Wind Tunnel	124	31920	0.1
6/94 (1)	Intel XP/S140	143	55700	0.2
11/94 - 11/95 (3)	Fujitsu Numerical Wind Tunnel	170	42000	0.1
6/96 (1)	Hitachi SR2201/1024	220	138,240	2.2
11/96 (1)	Hitachi CP-PACS/2048	368	103,680	0.6
6/97 - 6/00 (7)	Intel ASCI Red	2379	362,880	3.7
11/00 - 11/01 (3)	IBM ASCI White, SP Power3 375 MHz	7226	518,096	3.6
6/02 - 6/04 (5)	NEC Earth-Simulator	35,860	1,000,000	5.2
11/04 - 11/07 (7)	IBM BlueGene/L	478,200	1,000,000	0.4
6/08 - 6/09 (3)	IBM Roadrunner -PowerXCell 8i 3.2 Ghz	1,105,000	2,329,599	2.1
11/09 - 6/10 (2)	Cray Jaguar - XT5-HE 2.6 GHz	1,759,000	5,474,272	17.3
11/10 (1)	NUDT Tianhe-1A, X5670 2.93Ghz NVIDIA	2,566,000	3,600,000	3.4
6/11 - 11/11 (2)	Fujitsu K computer, SPARC64 VIIIfx	10,510,000	11,870,208	29.5
6/12 (?)	IBM Sequoia BlueGene/Q	16,324,751	12,681,215	23.1

June 2012: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom	USA	1,572,864	16.3	81	8.6	1895
2	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + custom	Japan	705,024	10.5	93	12.7	830
3	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + custom	USA	786,432	8.16	81	3.95	2069
4	Leibniz Rechenzentrum	SuperMUC, Intel (8c) + IB	Germany	147,456	2.90	90*	3.52	823
5	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel (6c) + Nvidia GPU (14c) + custom	China	186,368	2.57	55	4.04	636
6	DOE / OS Oak Ridge Nat Lab	Jaguar, Cray AMD (16c) + custom	USA	298,592	1.94	74	5.14	377
7	CINECA	Fermi, BlueGene/Q (16c) + custom	Italy	163,840	1.73	82	.821	2099
8	Forschungszentrum Juelich (FZJ)	JuQUEEN, BlueGene/Q (16c) + custom	Germany	131,072	1.38	82	.657	2099
9	Commissariat a l'Energie Atomique (CEA)	Curie, Bull Intel (8c) + IB	France	77,184	1.36	82	2.25	604
10	Nat. Supercomputer Center in Shenzhen	Nebulea, Dawning Intel (6) + Nvidia GPU (14c) + IB	China	120,640	1.27	43	2.58	493
500	Energy Comp	IBM Cluster, Intel + IB	Italy	4096	.061	93*		

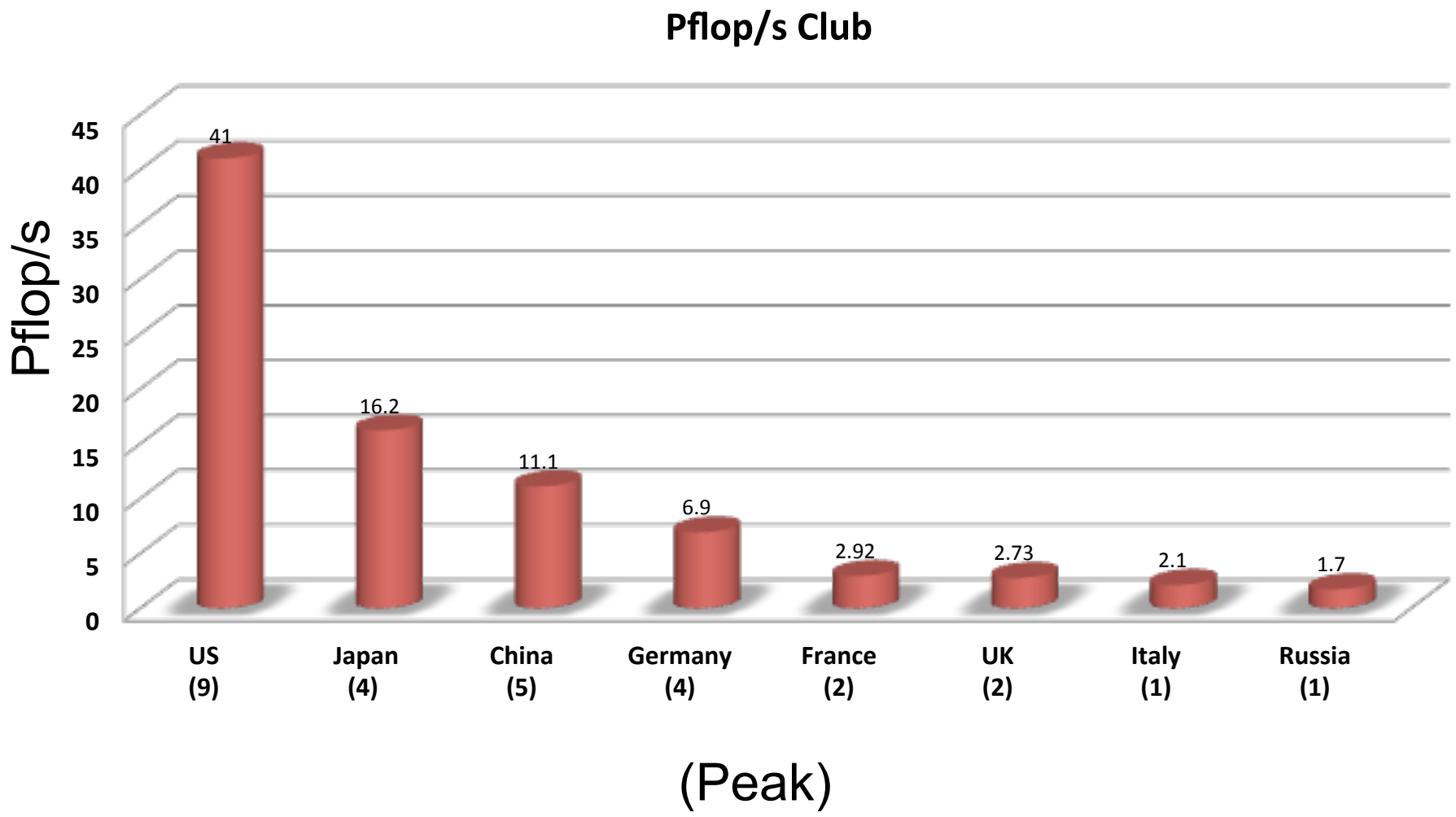
Accelerators (58 systems)



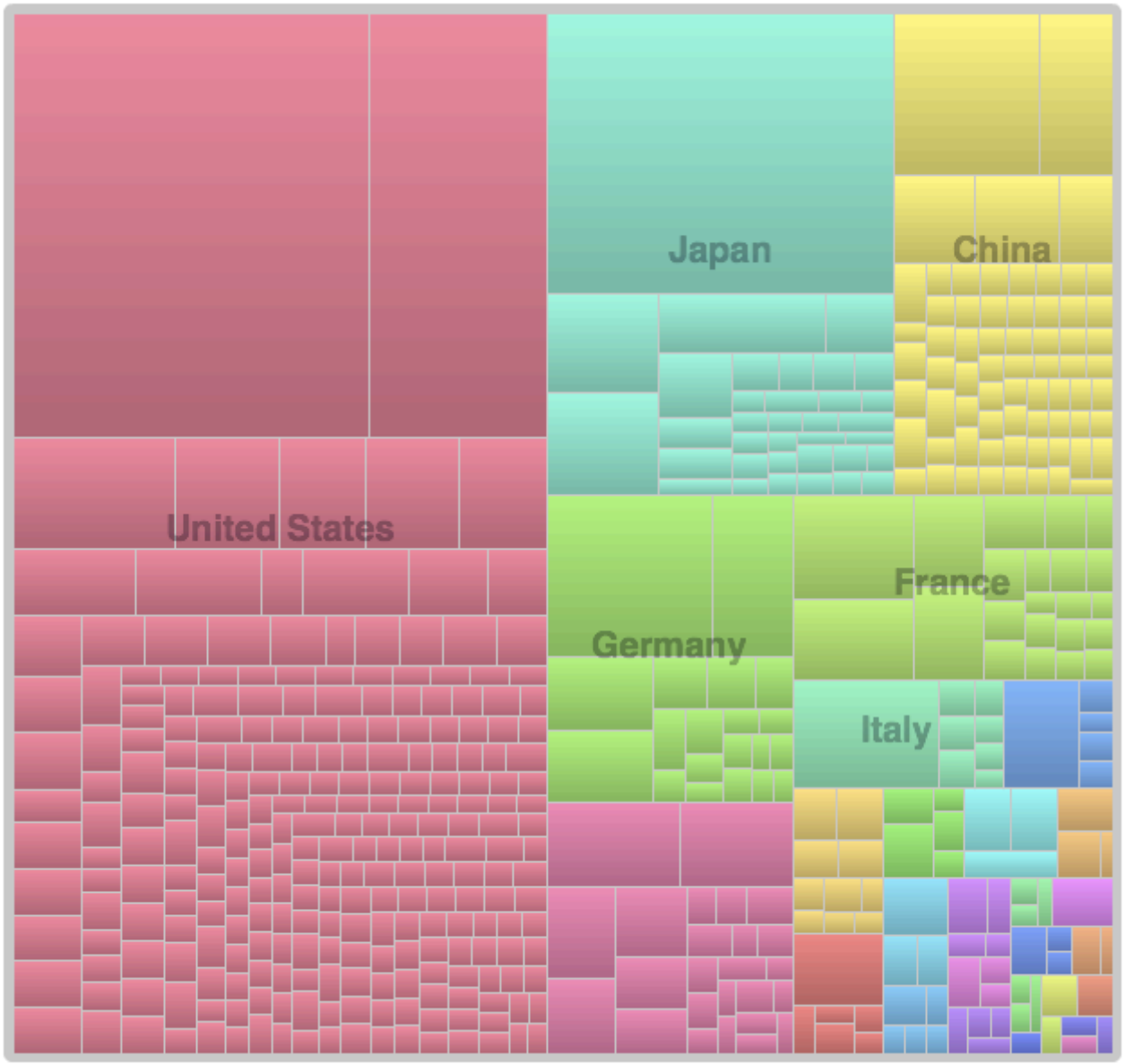
- Intel Xeon Phi (1)
- Clearspeed CSX600 (0)
- ATI GPU (2)
- IBM PowerXCell 8i (2)
- NVIDIA 2070 (10)
- NVIDIA 2050(12)
- NVIDIA 2090 (31)

27 US	2 Poland
7 China	1 Australia
4 Japan	1 Brazil
3 Russia	1 Canada
2 France	1 Singapore
2 Germany	1 Spain
2 India	1 Taiwan
2 Italy	1 UK

28 Systems at > Pflop/s (Peak)



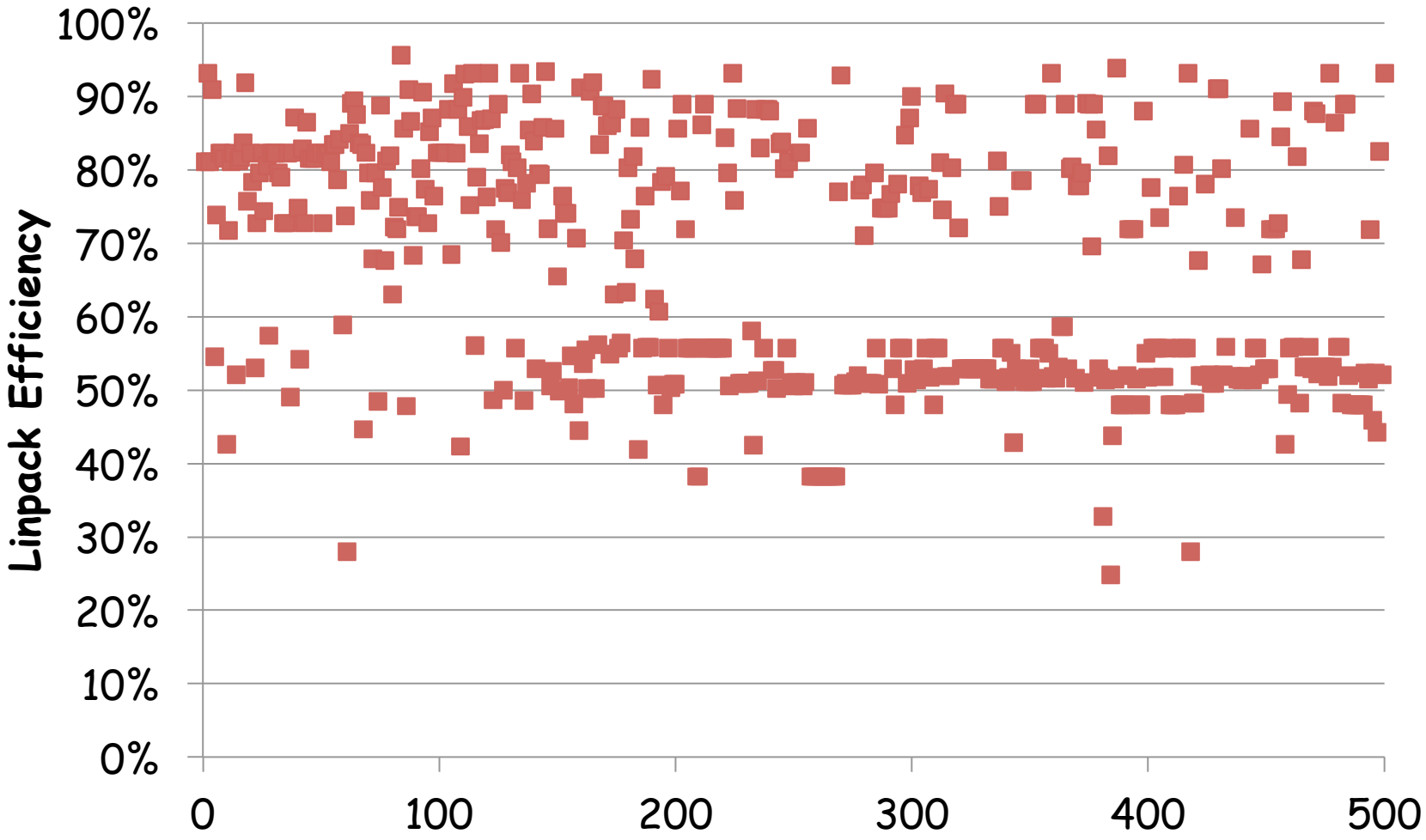
Countries Share



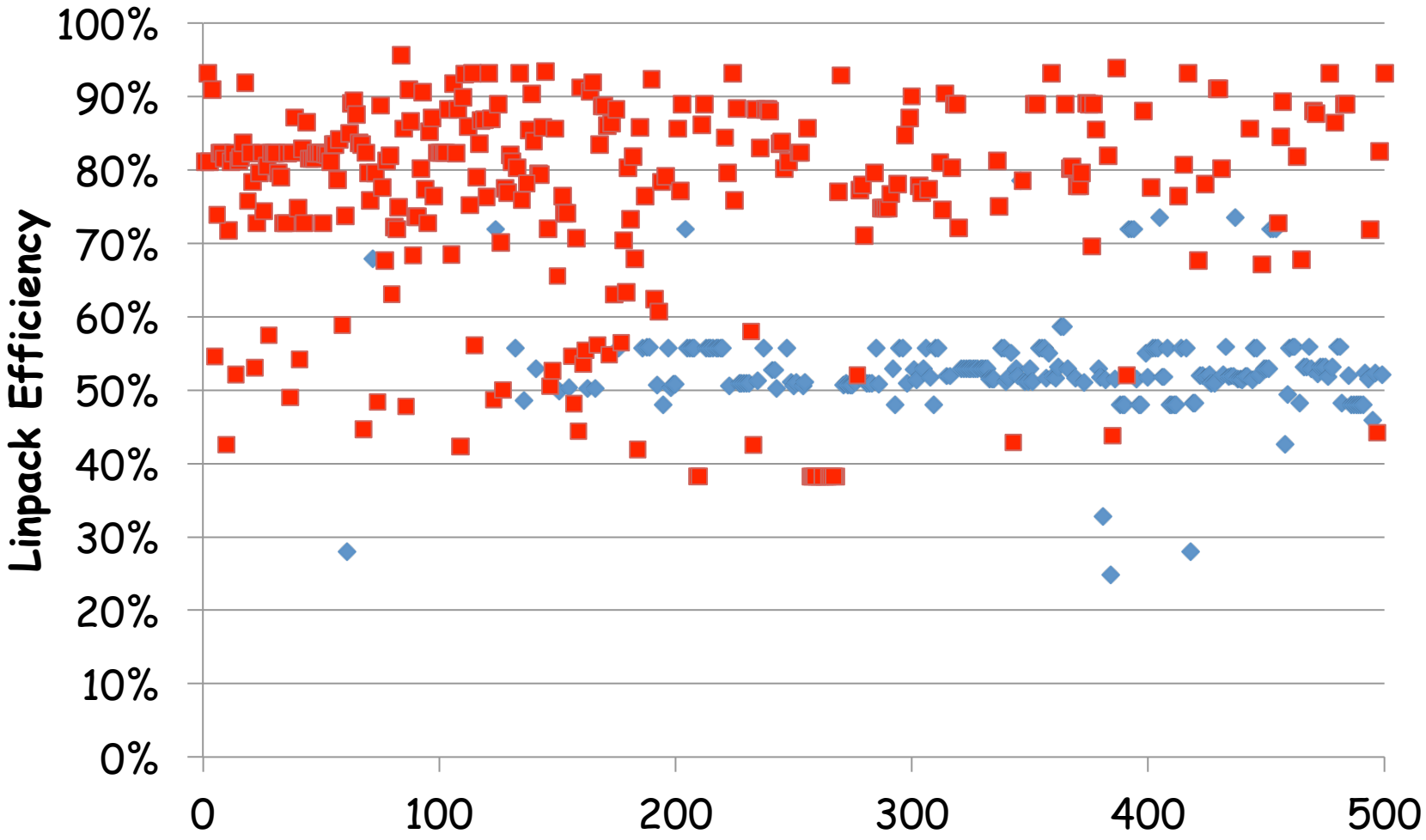
Absolute Counts

US:	252
China:	68
Japan:	35
UK:	25
France:	22
Germany:	20

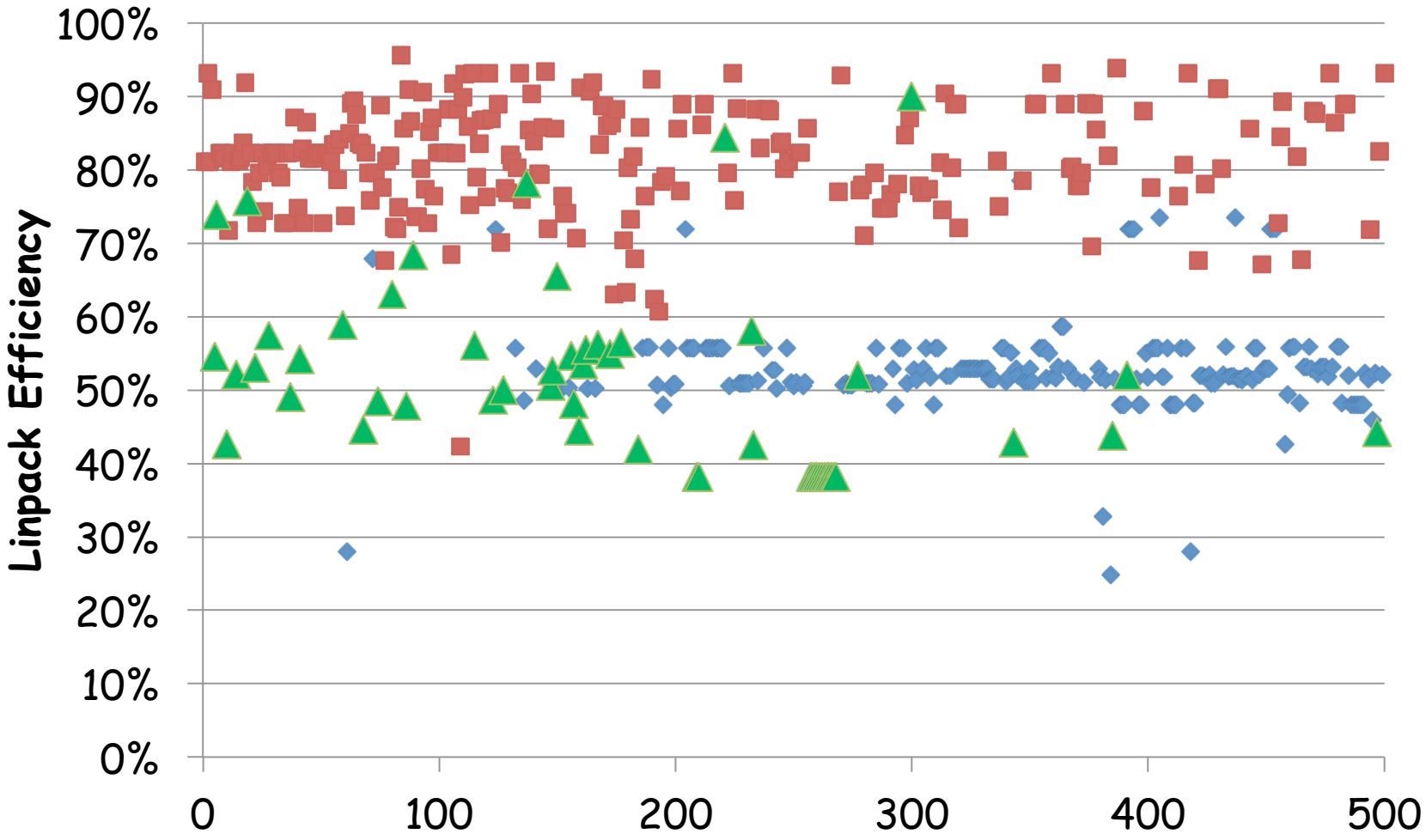
Linpack Efficiency



Linpack Efficiency



Linpack Efficiency



The High Cost of Data Movement

•Flop/s or percentage of peak flop/s become much less relevant

Approximate power costs (in picoJoules)

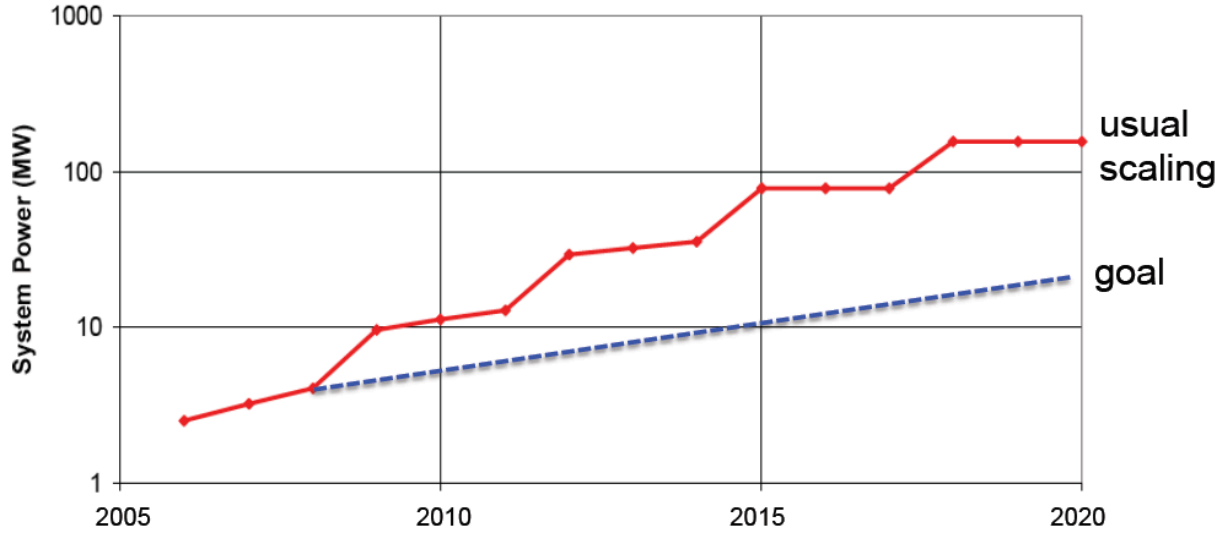
	2011
DP FMADD flop	100 pJ
DP DRAM read	4800 pJ
Local Interconnect	7500 pJ
Cross System	9000 pJ

Source: John Shalf, LBNL

•Algorithms & Software: minimize data movement; perform more work per unit data movement.

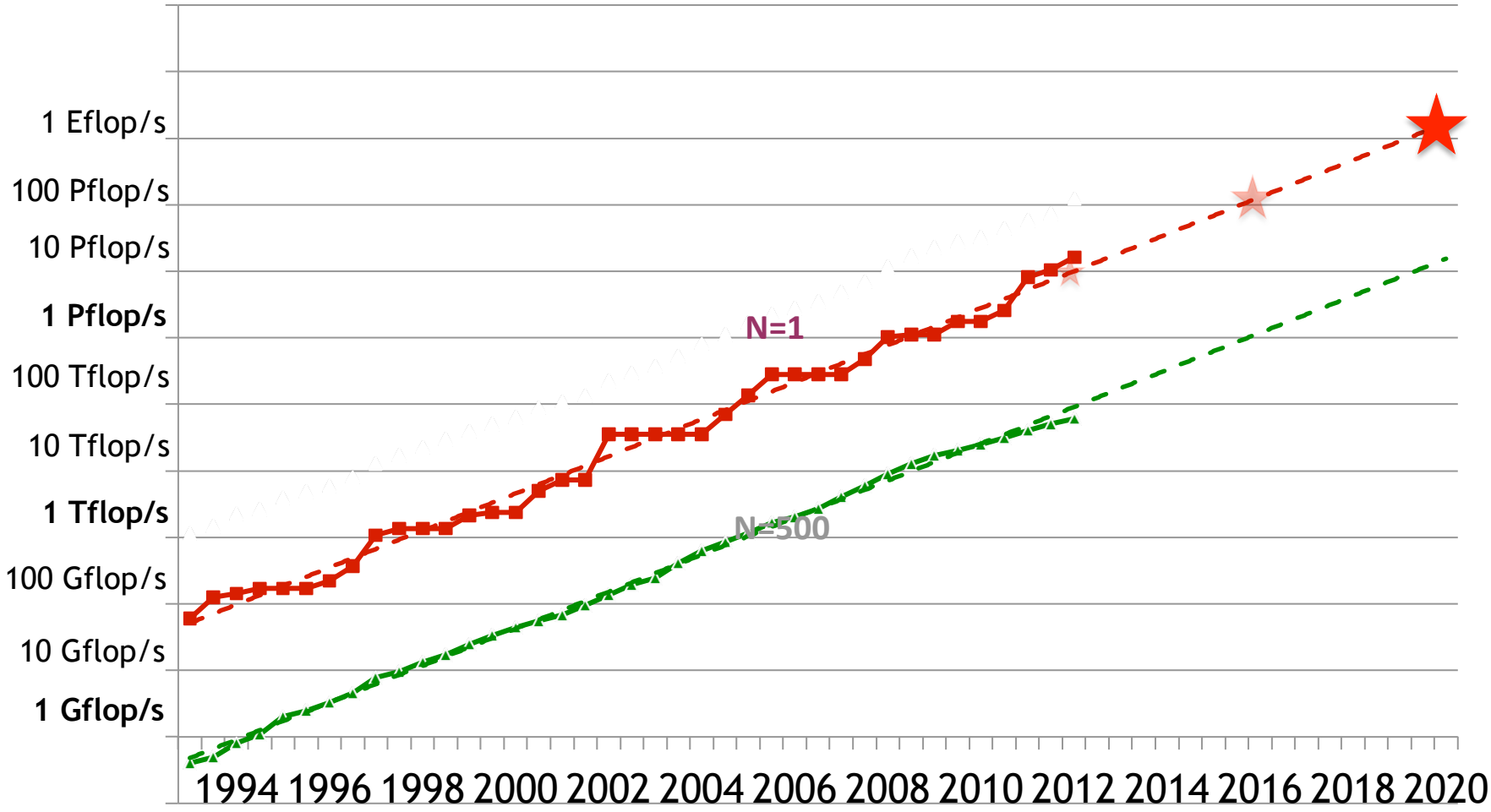
Energy Cost Challenge

- At ~$1M per MW energy costs are substantial
  - 10 Pflop/s in 2011 uses ~10 MWs
  - 1 Eflop/s in 2018 > 100 MWs



  - DOE Target: 1 Eflop/s in 2018 at 20 MWs

Performance Development in Top500



Potential System Architecture with a cap of $200M and 20MW

Systems	2012 BG/Q Computer
System peak	**20 Pflop/s**
Power	**8.6 MW**
System memory	1.6 PB (16*96*1024)
Node performance	205 GF/s (16*1.6GHz*8)
Node memory BW	42.6 GB/s
Node concurrency	64 Threads
Total Node Interconnect BW	20 GB/s
System size (nodes)	98,304 (96*1024)
Total concurrency	5.97 M
MTTI	4 days

Potential System Architecture with a cap of $200M and 20MW

Systems	2012 BG/Q Computer	2022	Difference Today & 2022
System peak	**20 Pflop/s**	**1 Eflop/s**	**O(100)**
Power	**8.6 MW**	**~20 MW**	
System memory	1.6 PB (16*96*1024)	32 - 64 PB	O(10)
Node performance	205 GF/s (16*1.6GHz*8)	1.2 or 15TF/s	O(10) – O(100)
Node memory BW	42.6 GB/s	2 - 4TB/s	O(1000)
Node concurrency	64 Threads	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	20 GB/s	200-400GB/s	O(10)
System size (nodes)	98,304 (96*1024)	O(100,000) or O(1M)	O(100) - O(1000)
Total concurrency	5.97 M	O(billion)	O(1,000)
MTTI	4 days	O(<1 day)	- O(10)

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.