Moscow University Supercomputing Center

Алгоритмического и программного Проблемы при переходе на Exascale

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Moscow University Supercomputing Center

Algorithmic and Software Challenges when Moving Towards Exascale

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Key Message

- **Exascale has been discussed in numerous workshops, conferences, planning meetings for about five years.**
- **Exascale projects have been started in the US and many other countries and regions.**
- **Progress has been made, but key challenges to exascale remain.**

State of Supercomputing in 2013

- Pflops computing fully established with 26 machines.
- Three technology "swim lanes" or architecture possibilities are thriving.
- Interest in supercomputing is now worldwide, and growing in many new markets (over 50% of Top500 computers are in industry).
- Exascale projects exist in many countries and regions.
- Rapid growth predicted by IDC for the next three years.

June 2013: The TOP10

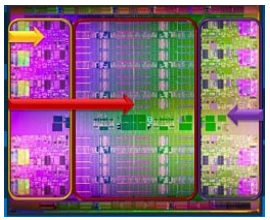
Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	National University of Defense Technology	Tianhe-2 NUDT, Xeon 12C 2.2GHz + IntelXeon Phi (57c) + Custom	China	3,120,000	33.9	62	18	1902
2	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7 (16C) + Nvidia Kepler GPU (14c) + Custom	USA	560,640	17.6	65	8.3	2143
3	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom	USA	1,572,864	17.1	85	7.9	2177
4	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + Custom	Japan	705,024	10.5	93	13	830
5	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + Custom	USA	786,432	8.58	85	3.9	2177
6	Texas Advanced Computing Center	Stampede, Dell Intel (8c) + Intel Xeon Phi (61c) + IB	USA	204,900	5.16	61	4.5	1146
7	Forschungszentrum Juelich (FZJ)	JuQUEEN, BlueGene/Q, Power BQC 16C 1.6GHz+Custom	Germany	458,752	5.01	85	2.3	2177
8	DOE / NNSA L Livermore Nat Lab	Vulcan, BlueGene/Q, Power BQC 16C 1.6GHz+Custom	USA	393,216	4.29	85	2.0	2177
9	Leibniz Rechenzentrum	SuperMUC, Intel (8c) + IB	Germany	147,456	2.90	91*	3.4	846
10	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel (6c) + Nvidia Fermi GPU (14c) + Custom	China	186,368	2.57	55	4.0	635
500	Web Company	HP Cluster	USA	17,904	.096	50		

Russian Systems on Top500

Rank	Name	Computer	Site	Manufacturer	Total Cores	Accelor Cores	Rmax (Tflop/s)	% of Peak	Accelerator/ Co-Processor
31	Lomonosov	T-Platforms T-Blade2/1.1, Xeon X5570/ X5670/E5630 2.93/2.53 GHz, Nvidia 2070 GPU, PowerXCell 8i Inf QDR	Moscow State University - Research Computing Center	T-Platforms	78,660	29,820	901	53	NVIDIA 2070
72	MVS-10P	RSC Tornado, Xeon E5-2690 8C 2.900GHz, Ind FDR, Intel Xeon Phi SE10X	Joint Supercomputer Center	RSC Group	28,704	25,376	375	72	Intel Xeon Phi
217		Cluster Platform 3000 BL460c Gen8, Xeon E5-2660 8C 2.200GHz, GEnet	IT Services Provider	Hewlett-Packard	18,032	0	160	51	
249	RSC Tornado SUSU	RSC Tornado, Xeon X5680 6C 3.330GHz, Inf QDR, Intel Xeon Phi SE10X	South Ural State University	RSC Group	14,016	11,712	146	62	Intel Xeon Phi
355	MVS-100K	Cluster Platform 3000 BL460c/BL 2x220/ SL390, Xeon E5450/5365/X5675 4C 3.000GHz, InF DDR, NVIDIA 2090	Joint Supercomputer Center	Hewlett-Packard	13,004	2432	119	53	NVIDIA 2090
428	Uran	ClusterPlatform SL390s/ SL270s, Xeon X5675 6C 3.060GHz, InF QDR, NVIDIA 2090	IMM UrOAN	Hewlett-Packard	5900	5312	105	46	NVIDIA 2090
464		Cluster Platform 3000 BL 2x220, Xeon E5450 4C 3.000GHz, Inf QDR	Kurchatov Institute Moscow	Hewlett-Packard	10,304	0	101	82	
471	SKIF Aurora	SKIF Aurora Platform - Intel Xeon X5680, Infiniband QDR	South Ural State University	RSC Group	8832	0	100	86	

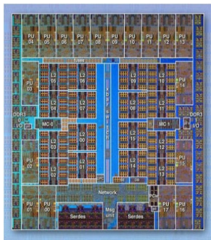
Three Possible Architecture Paths

- **Multicore:** Maintain complex cores, and replicate (x86, SPARC, Power7) [#4 and 9]



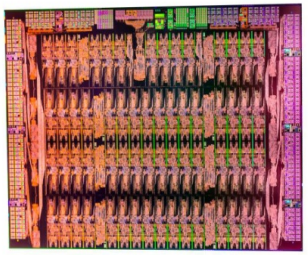
Intel Xeon E7
(10 cores)

- **Manycore/Embedded:** Use many simpler, low power cores from embedded (BlueGene, future ARM) [#3, 5, 7, and 8]



IBM BlueGene/Q
(16 +2 cores)

- **GPU/Coprocessor/Accelerator**: Use highly specialized processors from graphics market space (NVidia Fermi, Intel Xeon Phi, AMD) [# 1, 2, 6, and 10]



Intel Xeon Phi
(60 cores)

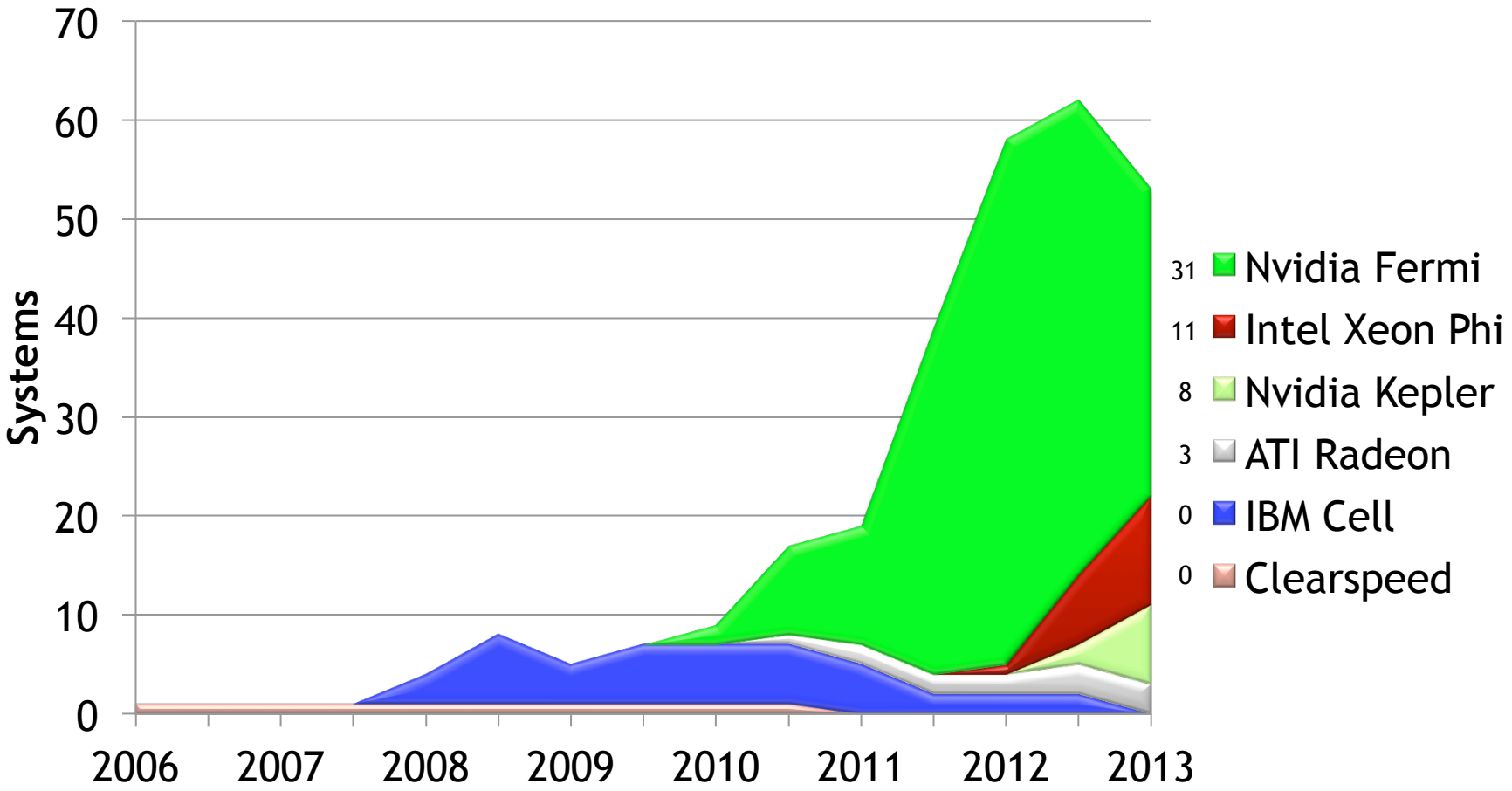
ICL UT

Petaflops Club

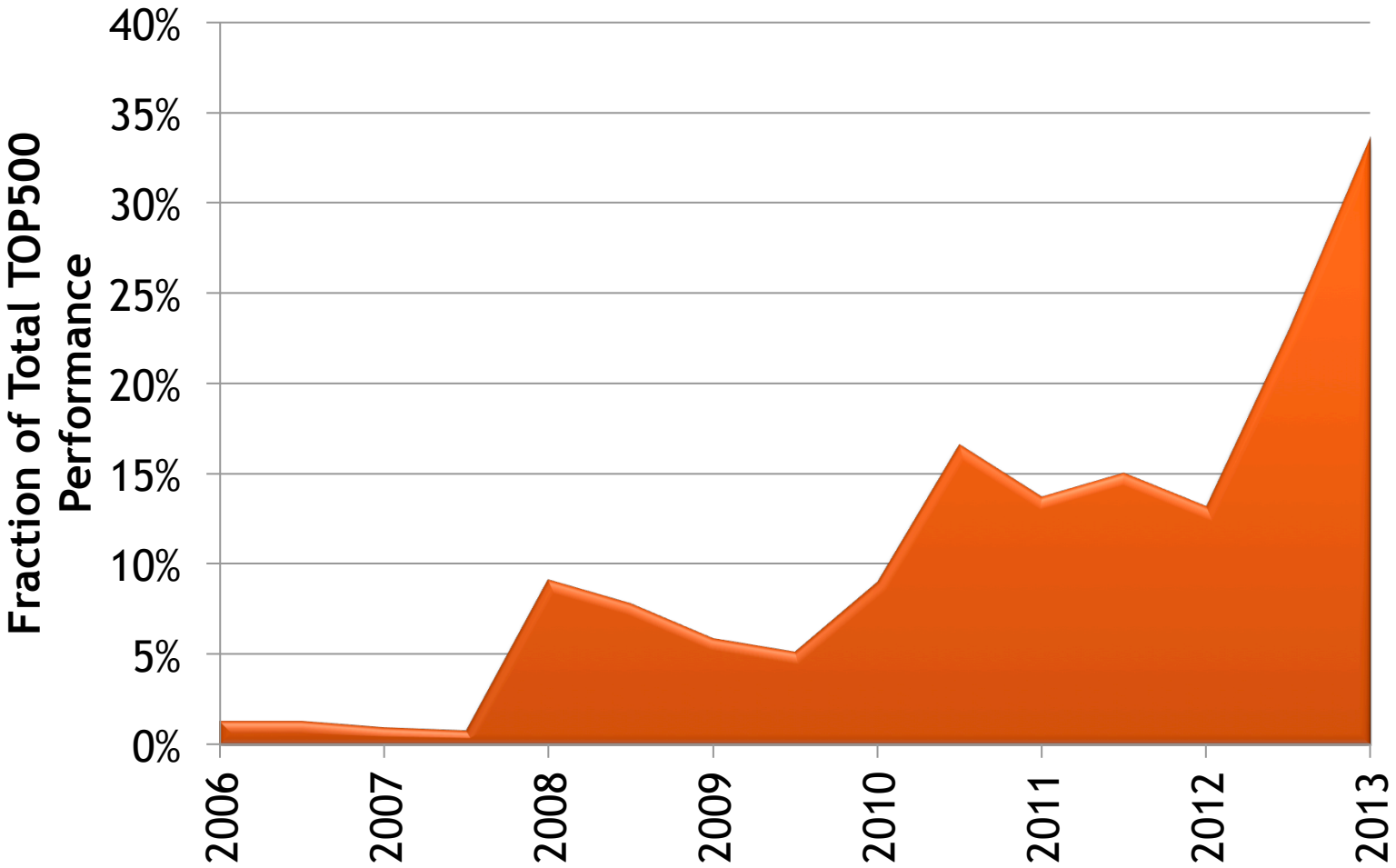
Name	Rmax Linpack# Pflops	Country	11 4 2 3 2 3 1
Tianhe-2 (MilkyWay-2)	33.9	China	NUDT: Hybrid Intel/Intel/Custom
Titan	17.6	US	Cray: Hybrid AMD/Nvidia/Custom
Sequoia	17.2	US	**IBM: BG-Q/Custom**
K Computer	10.5	Japan	Fujitsu: Sparc/Custom
Mira	8.59	US	**IBM: BG-Q/Custom**
Stampede	5.17	US	Dell: Hybrid/Intel/Intel/IB
JUQUEEN	5.01	Germany	**IBM: BG-Q/Custom**
Vulcan	4.29	US	**IBM: BG-Q/Custom**
SuperMUC	2.90	Germany	IBM: Intel/IB
Tianhe-1A	2.57	China	NUDT: Hybrid Intel/Nvidia/Custom
Pangea	2.10	France	Bull: Intel/IB
Fermi	1.79	Italy	**IBM: BG-Q/Custom**
DARPA Trial Subset	1.52	US	IBM: Intel/IB
Spirit	1.42	US	SGI: Intel/IB
Curie thin nodes	1.36	France	Bull: Intel/IB
Nebulae	1.27	China	Dawning: Hybrid Intel/Nvidia/IB
Yellowstone	1.26	US	**IBM: BG-Q/Custom**
Blue Joule	1.25	UK	**IBM: BG-Q/Custom**
Pleiades	1.24	US	SGI Intel/IB
Helios	1.24	Japan	Bull: Intel/IB
TSUBAME 2.0	1.19	Japan	NEC/HP: Hybrid Intel/Nvidia/IB
Cielo	1.11	US	Cray: AMD/Custom
DiRAC	1.07	K	**IBM: BG-Q/Custom**
Hopper	1.05	US	Cray: AMD/Custom
Tera-100	1.05	France	Bull: Intel/IB
Oakleaf-FX	1.04	Japan	Fujitsu: Sparc/Custom

6 Hybrid Architectures
8 IBM BG/Q
15 Custom X
11 Infiniband X
9 Look like “clusters”

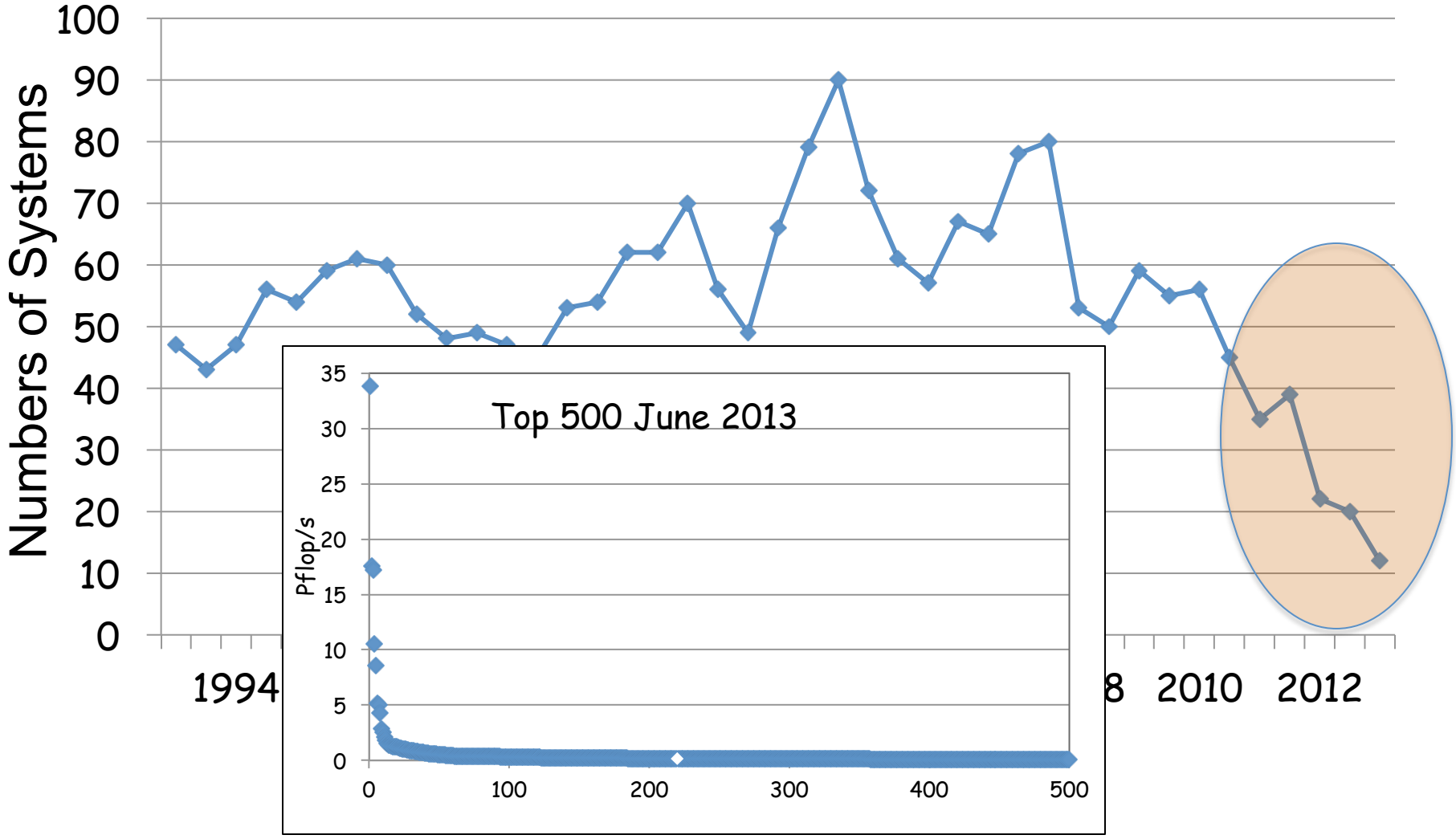
Hybrid/Accelerators (53 Systems)



Top500 Performance Share of Accelerators



For the Top 500: Rank at which Half of Total Performance is Accumulated



Commodity plus Accelerator Today

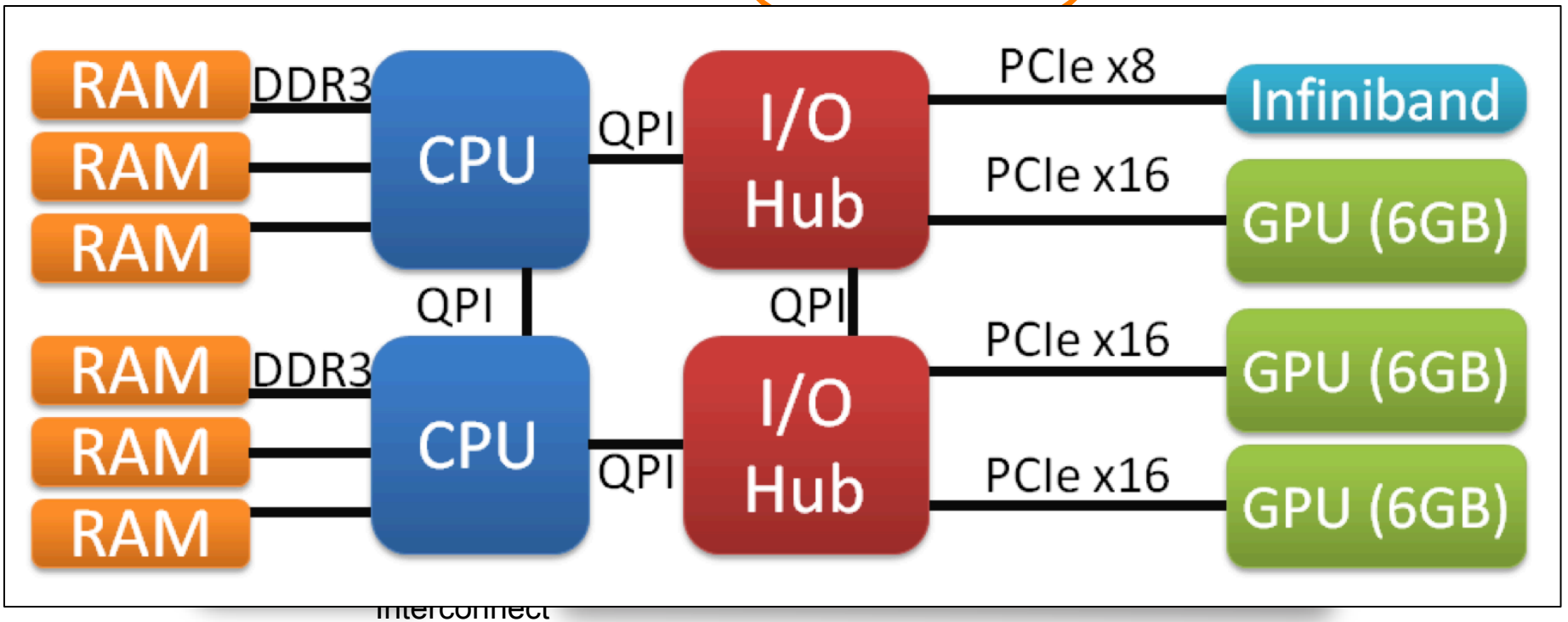
Commodity

Intel Xeon
8 cores
3 GHz
8*4 ops/cycle
96 Gflop/s (DP)

Accelerator (GPU)

Nvidia K20X “Kepler”
2688 “Cuda cores”
.732 GHz
2688*2/3 ops/cycle
1.31 Tflop/s (DP)

192 Cuda cores/SMX
2688 “Cuda cores”



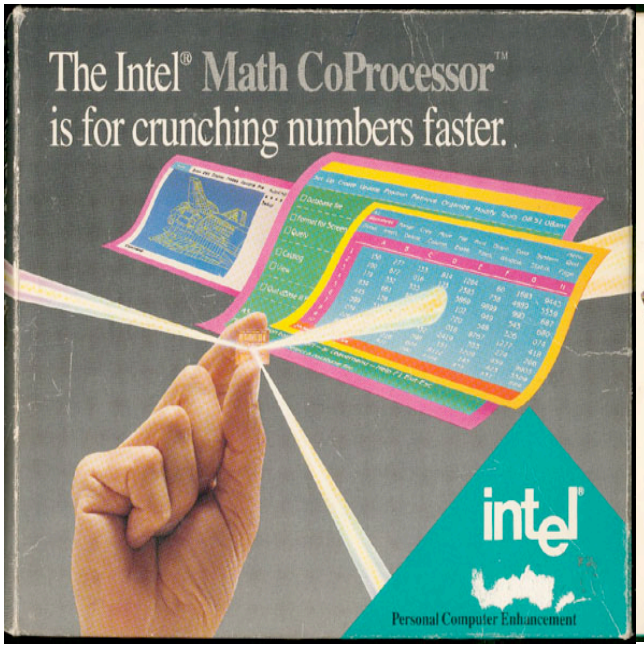
Interconnect
PCI-X 16 lane
64 Gb/s (8 GB/s)
1 GW/s

We Have Seen This Before

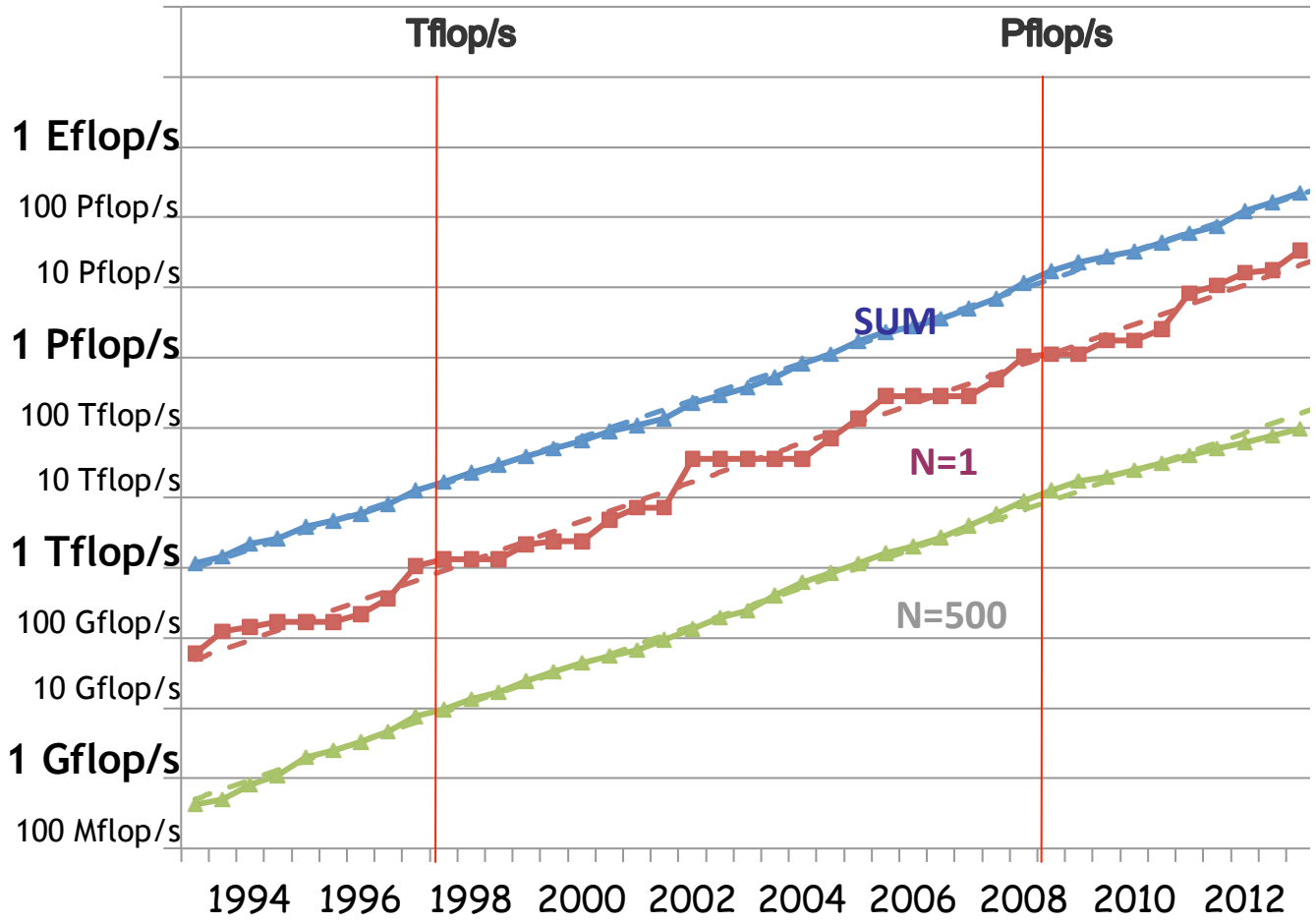
- Floating Point Systems FPS-164/ MAX Supercomputer (1976)
- Intel Math Co-processor (1980)
- Weitek Math Co-processor (1981)

1976

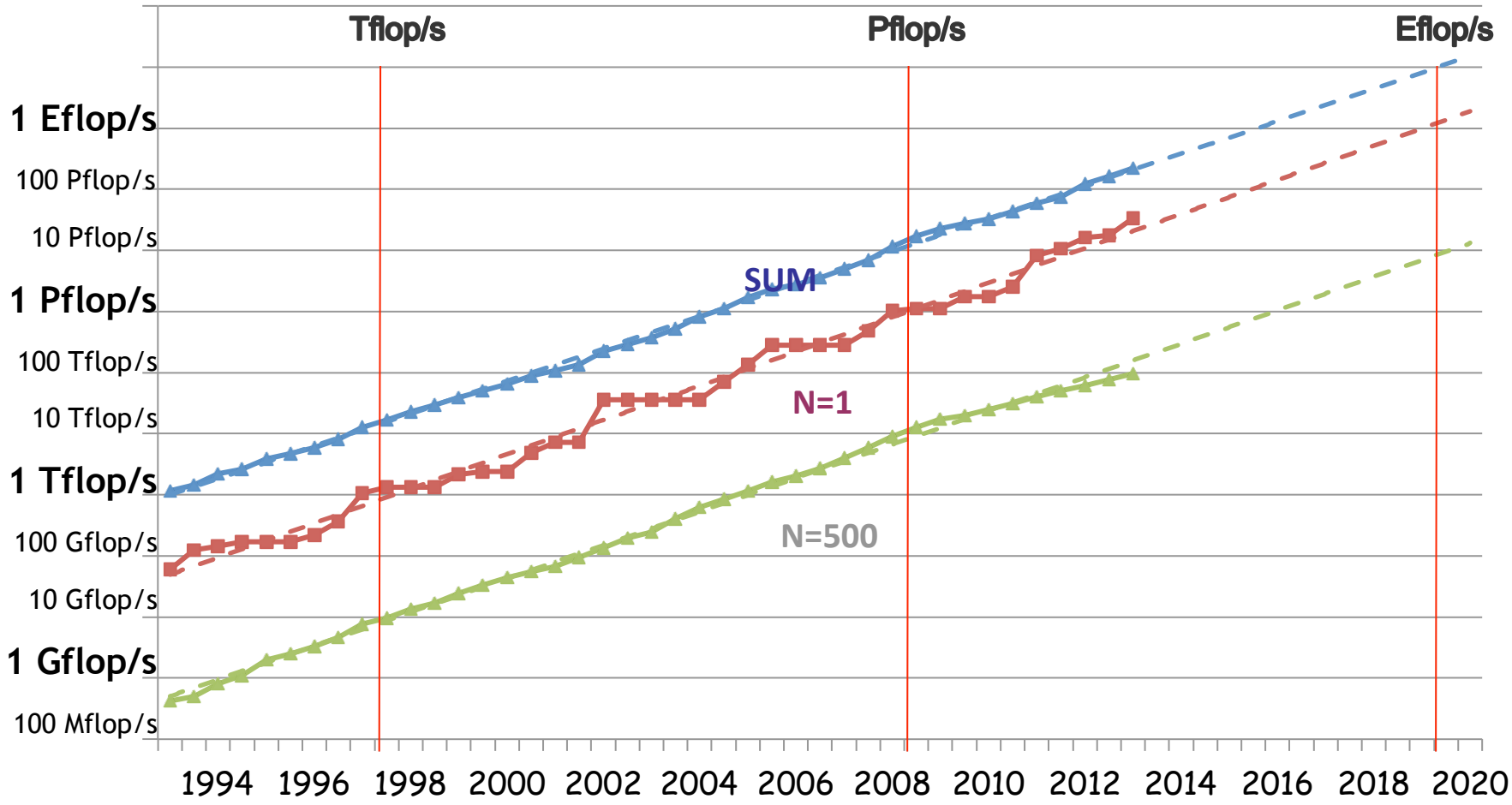
1980



TOP500 Editions (41 so far, 20 years)



TOP500 Editions (53 edition, 26 years)



Today's #1 System

Systems	2013 Tianhe-2
System peak	**55 Pflop/s**
Power	**18 MW** **(3 Gflops/W)**
System memory	1.4 PB (1.024 PB CPU + .384 PB CoP)
Node performance	3.43 TF/s (.4 CPU +3 CoP)
Node concurrency	24 cores CPU + 171 cores CoP
Node Interconnect BW	6.36 GB/s
System size (nodes)	16,000
Total concurrency	3.12 M 12.48M threads (4/core)
MTTF	Few / day

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

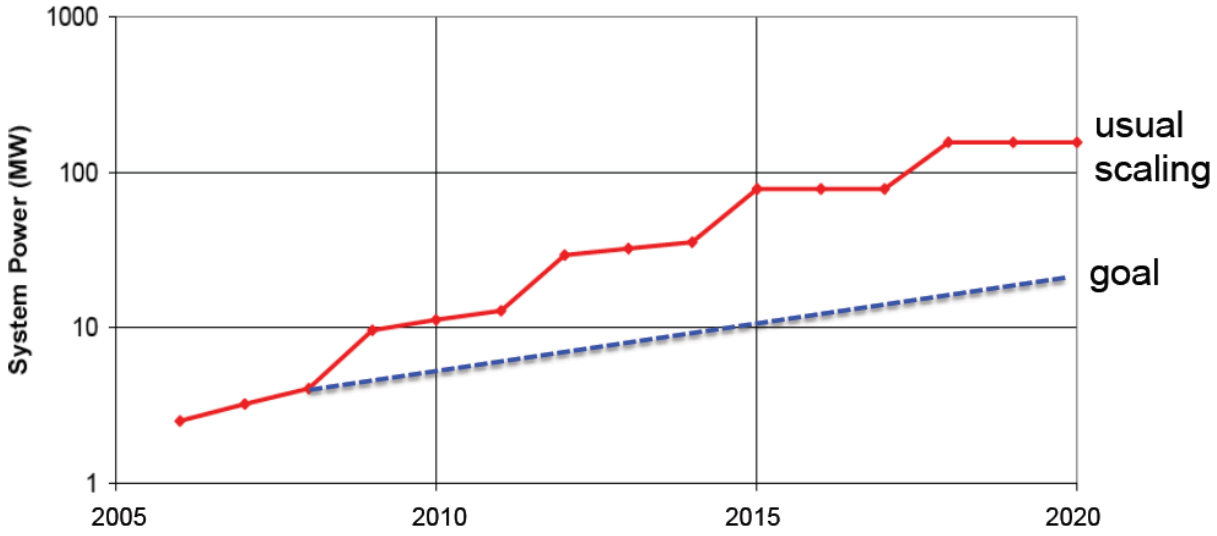
Systems	2013 Tianhe-2
System peak	**55 Pflop/s**
Power	**18 MW** **(3 Gflops/W)**
System memory	1.4 PB (1.024 PB CPU + .384 PB CoP)
Node performance	3.43 TF/s (.4 CPU +3 CoP)
Node concurrency	24 cores CPU + 171 cores CoP
Node Interconnect BW	6.36 GB/s
System size (nodes)	16,000
Total concurrency	3.12 M 12.48M threads (4/core)
MTTF	Few / day

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

Systems	2013 Tianhe-2	2020-2022	Difference Today & Exa
System peak	**55 Pflop/s**	**1 Eflop/s**	**~20x**
Power	**18 MW** **(3 Gflops/W)**	**~20 MW** **(50 Gflops/W)**	**O(1)** **~15x**
System memory	1.4 PB (1.024 PB CPU + .384 PB CoP)	32 - 64 PB	~50x
Node performance	3.43 TF/s (.4 CPU +3 CoP)	1.2 or 15TF/s	O(1)
Node concurrency	24 cores CPU + 171 cores CoP	O(1k) or 10k	~5x - ~50x
Node Interconnect BW	6.36 GB/s	200-400GB/s	~40x
System size (nodes)	16,000	O(100,000) or O(1M)	~6x - ~60x
Total concurrency	3.12 M 12.48M threads (4/core)	O(billion)	~100x
MTTF	Few / day	Many / day	O(?)

Energy Cost Challenge

- At ~$1M per MW energy costs are substantial
  - 10 Pflop/s in 2011 uses ~10 MWs
  - 1 Eflop/s in 2020 > 100 MWs



  - DOE Target: 1 Eflop/s around 2020-2022 at 20 MWs

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

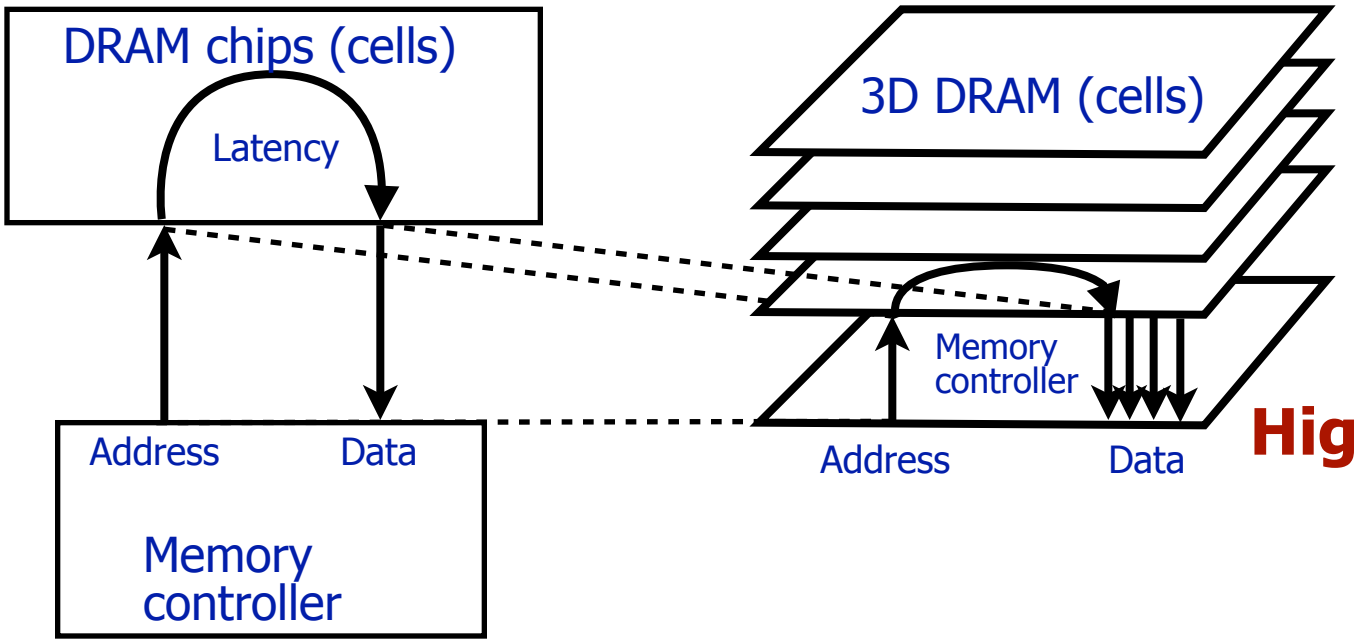
Source: John Shalf, LBNL

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

Evolution Over the Last 30 Years

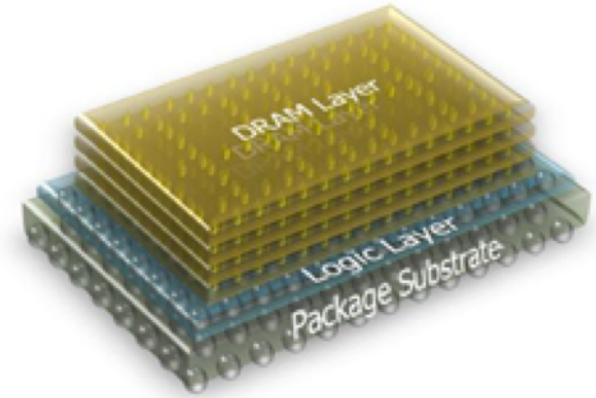
- Initially, commodity PCs where decentralized systems
- As chip manufacturing process shank to less than a micron, they started to integrate features on-die:
  - 1989: FPU (Intel 80486DX)
  - 1999: SRAM (Intel Pentium III)
  - 2009: GPU (AMD Fusion)
  - 2016: DRAM on chip (3D stacking)

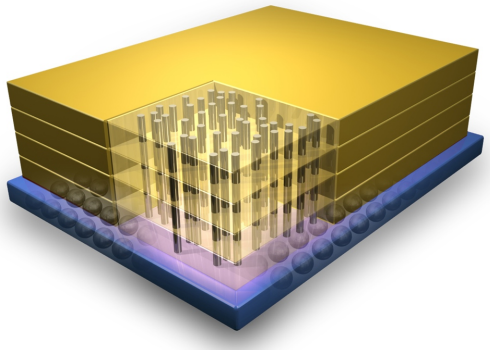
Future Systems May Be Composed of Different Kinds of Cores

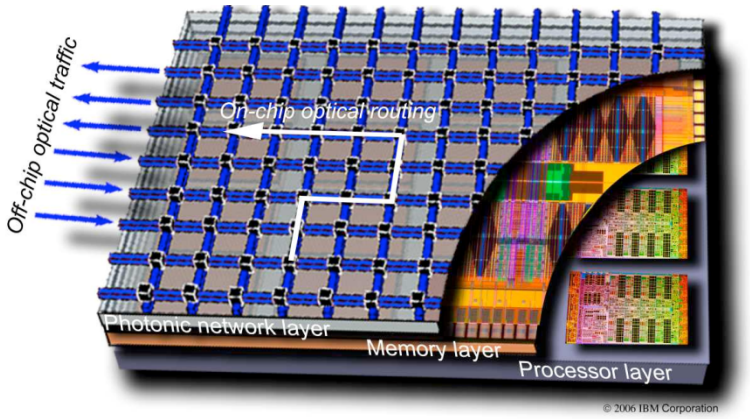


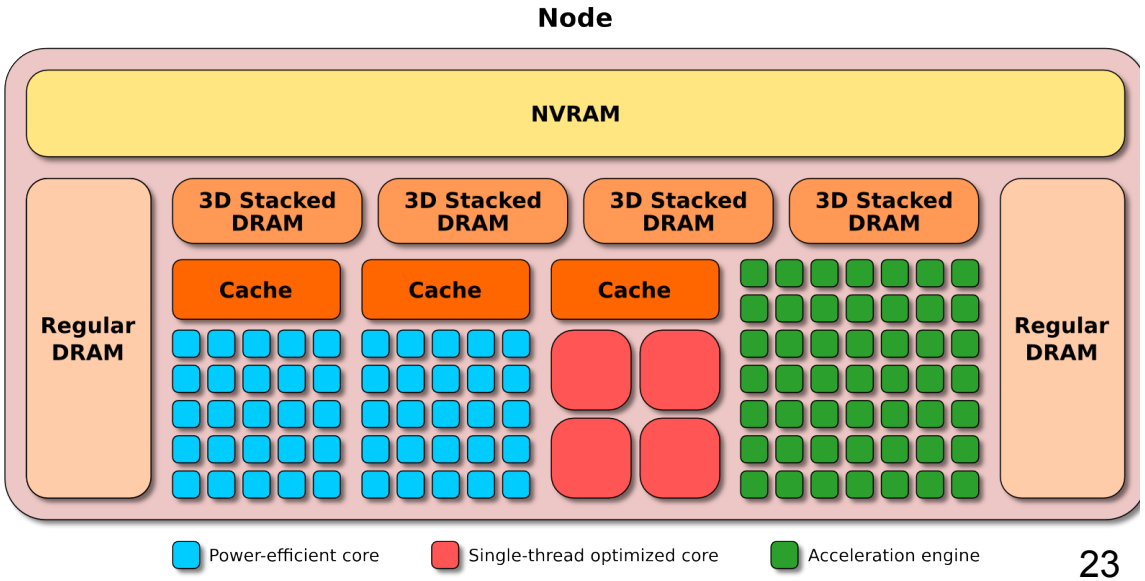
Lower latency

Higher bandwidth






Off-chip optical traffic
On-chip optical routing
Photonic network layer
Memory layer
Processor layer
© 2006 IBM Corporation


Node
NVRAM
Regular DRAM
3D Stacked DRAM
3D Stacked DRAM
3D Stacked DRAM
3D Stacked DRAM
Cache
Cache
Cache
Regular DRAM
Power-efficient core
Single-thread optimized core
Acceleration engine

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
  - Cache aware
- 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.

Summary

- Major Challenges are ahead for extreme computing
  - Parallelism $O(10^9)$
    - Programming issues
  - Hybrid
    - Peak and HPL may be very misleading
    - No where near close to peak for most apps
  - Fault Tolerance
    - Today Sequoia BG/Q node failure rate is 1.25 failures/ day
  - Power
    - 50 Gflops/w (today at 2 Gflops/w)
- We will need completely new approaches and technologies to reach the Exascale level

Collaborators / Software / Support

- **PLASMA**
  http://icl.cs.utk.edu/plasma/
- **MAGMA**
  http://icl.cs.utk.edu/magma/
- **Quark (RT for Shared Memory)**
  http://icl.cs.utk.edu/quark/
- **PaRSEC**(Parallel Runtime Scheduling and Execution Control)
  http://icl.cs.utk.edu/parsec/

- Collaborating partners
  University of Tennessee, Knoxville
  University of California, Berkeley
  University of Colorado, Denver

MAGMA **PLASMA**