Future of High Performance Computing

Thom H. Dunning, Jr.

National Center for Supercomputing Applications

Institute for Advanced Computing Applications and Technologies

Department of Chemistry

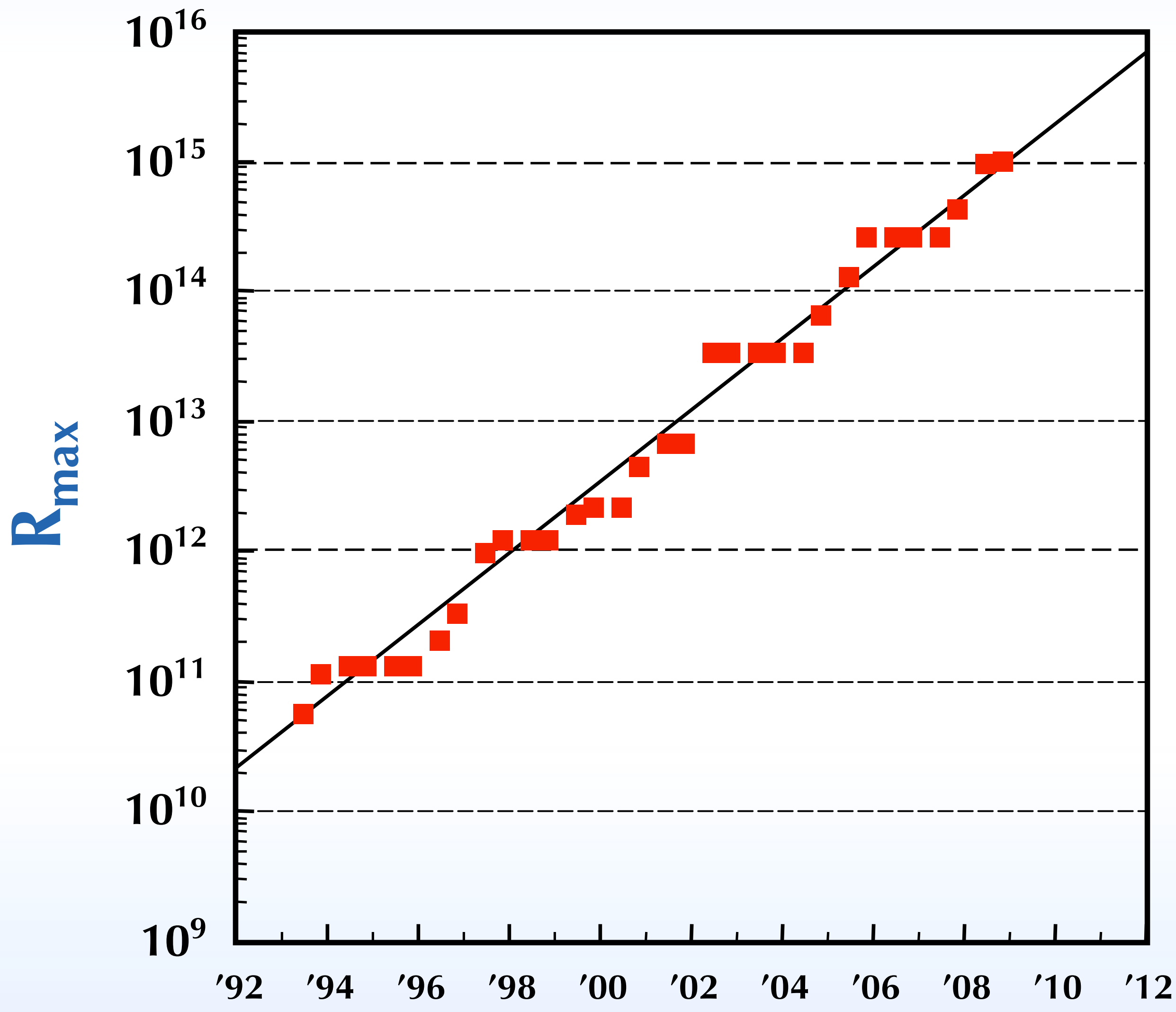
Outline of Presentation

- **Progress in High Performance Computing**
- **Directions in Computing Technology**
  - Multi-core and many-core chips
  - Memory subsystem
  - Communications subsystem
- **Era of Petascale Computing**
  - Science @ Petascale
  - Petascale Computing Systems
  - Blue Waters Petascale Computing System
- **Challenges of Petascale Computing**

Progress in
High Performance Computing

Progress in High Performance Computing

Relentless Increase in Performance



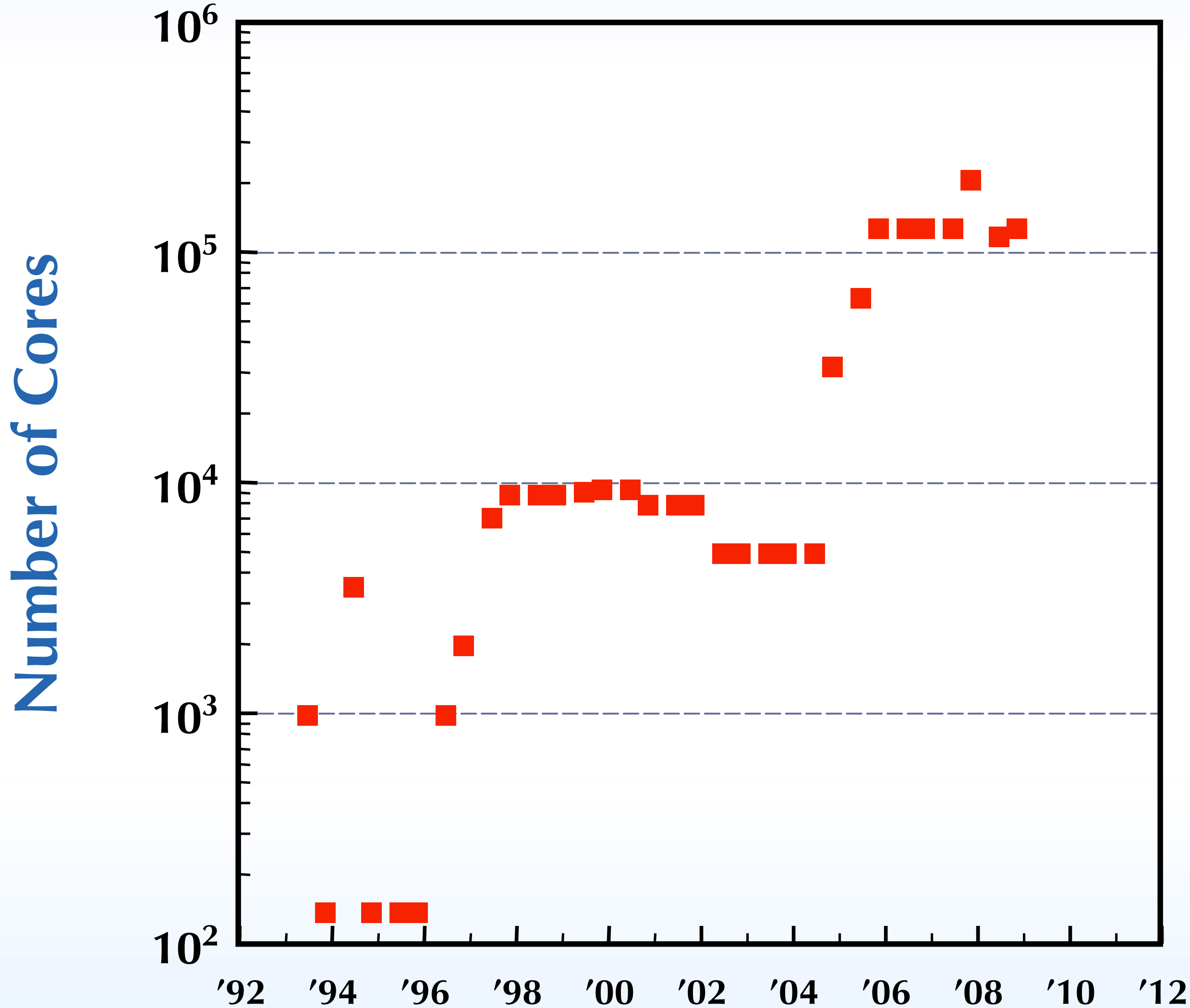
Top 500: #1

1 GF: late 1980s

1 TF: 1997

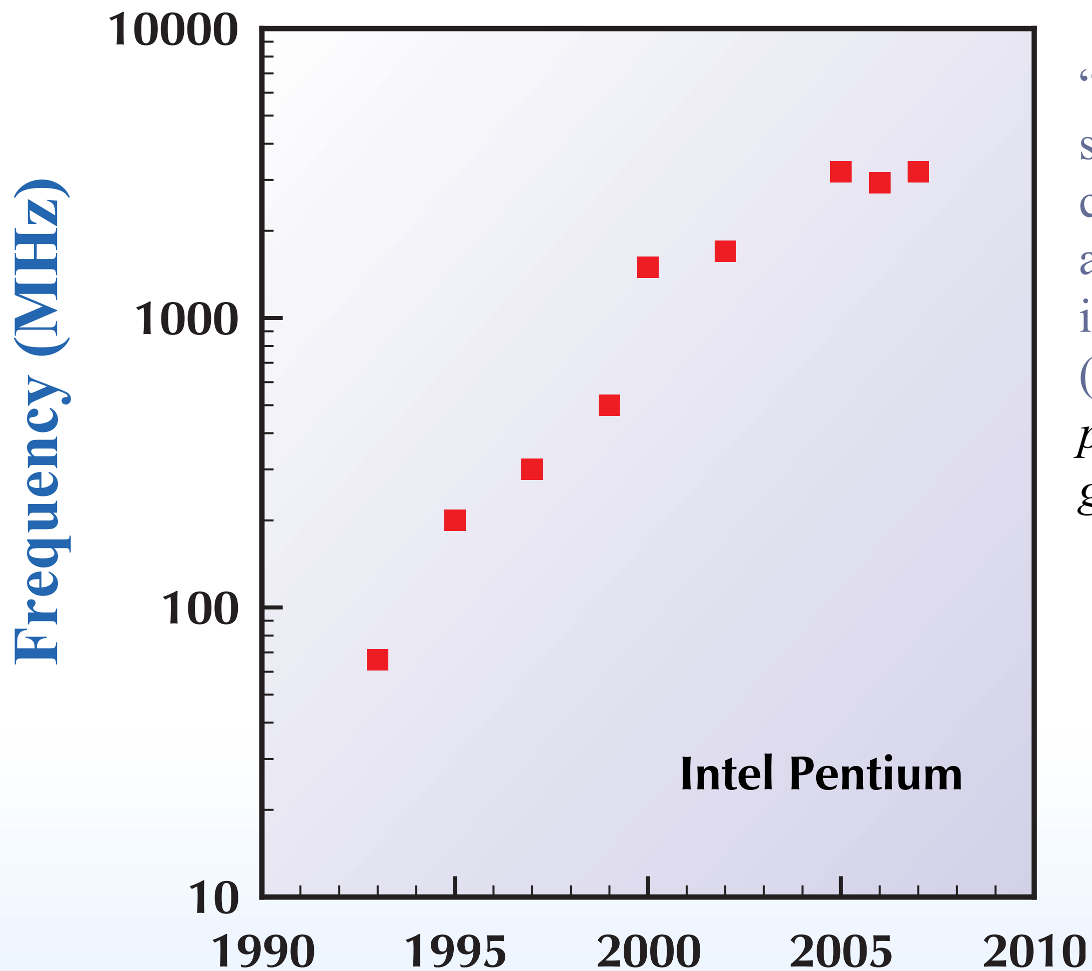
1 PF: 2008

Relentless Increase in Number of Cores



Directions in Computing Technologies

Increasing Clock Frequency & Performance



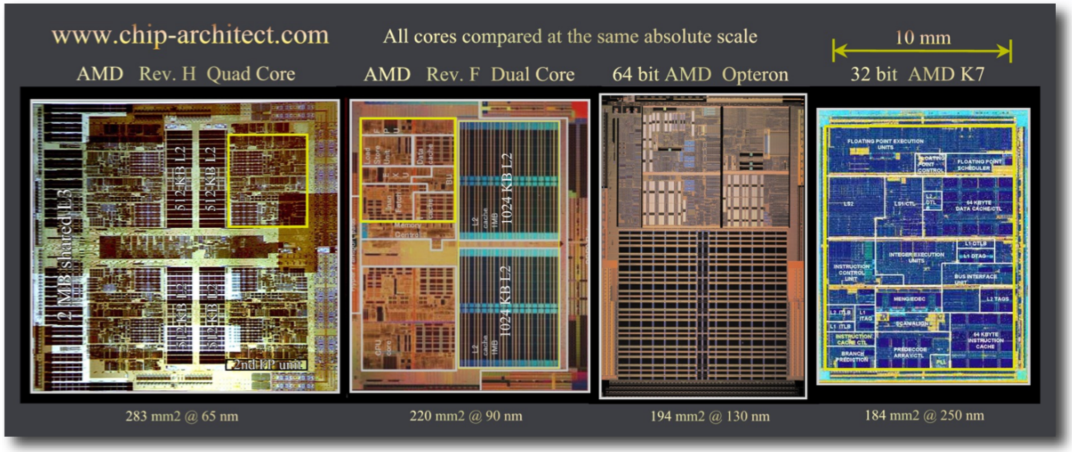
"In the past, performance scaling in conventional single-core processors has been accomplished largely through increases in clock frequency (*accounting for roughly 80 percent of the performance gains to date*)."

Platform 2015
S. Y. Borkar *et al.*, 2006
Intel Corporation

Directions in Computing Technologies

From Uni-core to Multi-core Processors

AMD
Uni-, Dual-, Quad-core, Processors

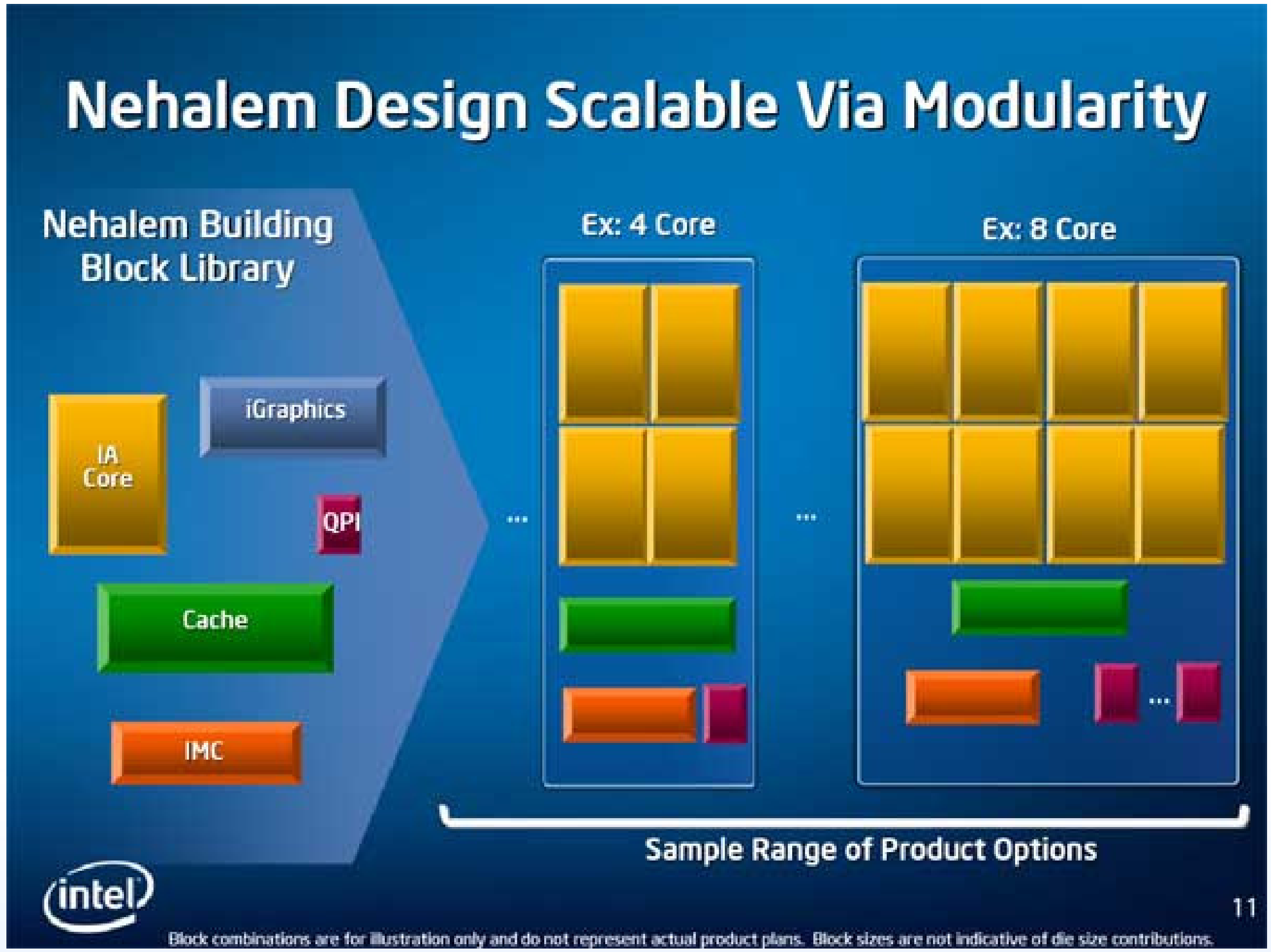


Intel
Multi-core Performance

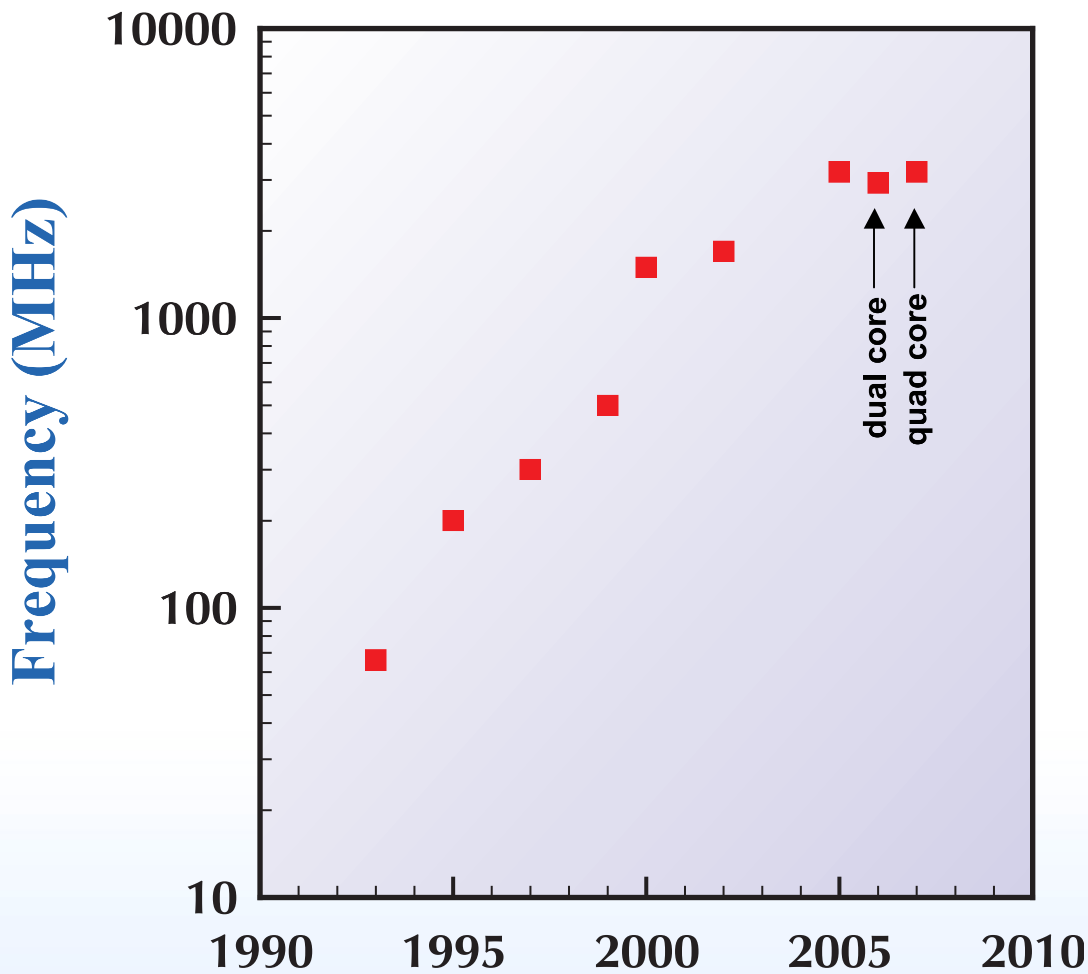
Multi-core 2009: Intel's Nehalem

Nehalem

- Modular
- Up to 8 cores
- 3 levels of cache
- Integrated memory controller
- Multiple QuickPath Interconnects



Switch to Multicore Chips

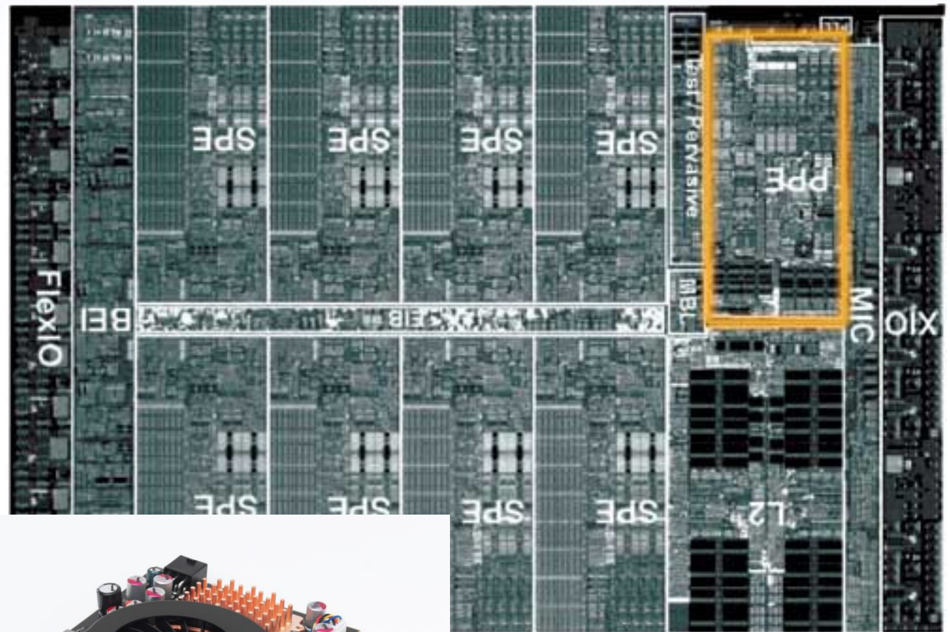


"For the next several years the only way to obtain significant increases in performance will be through increasing use of parallelism:

- 8× in 2009
- 16× in 2011
- 32× in 2013
- *etc.*

On to Many-core Chips

NVIDIA Telsa
(240 cores)

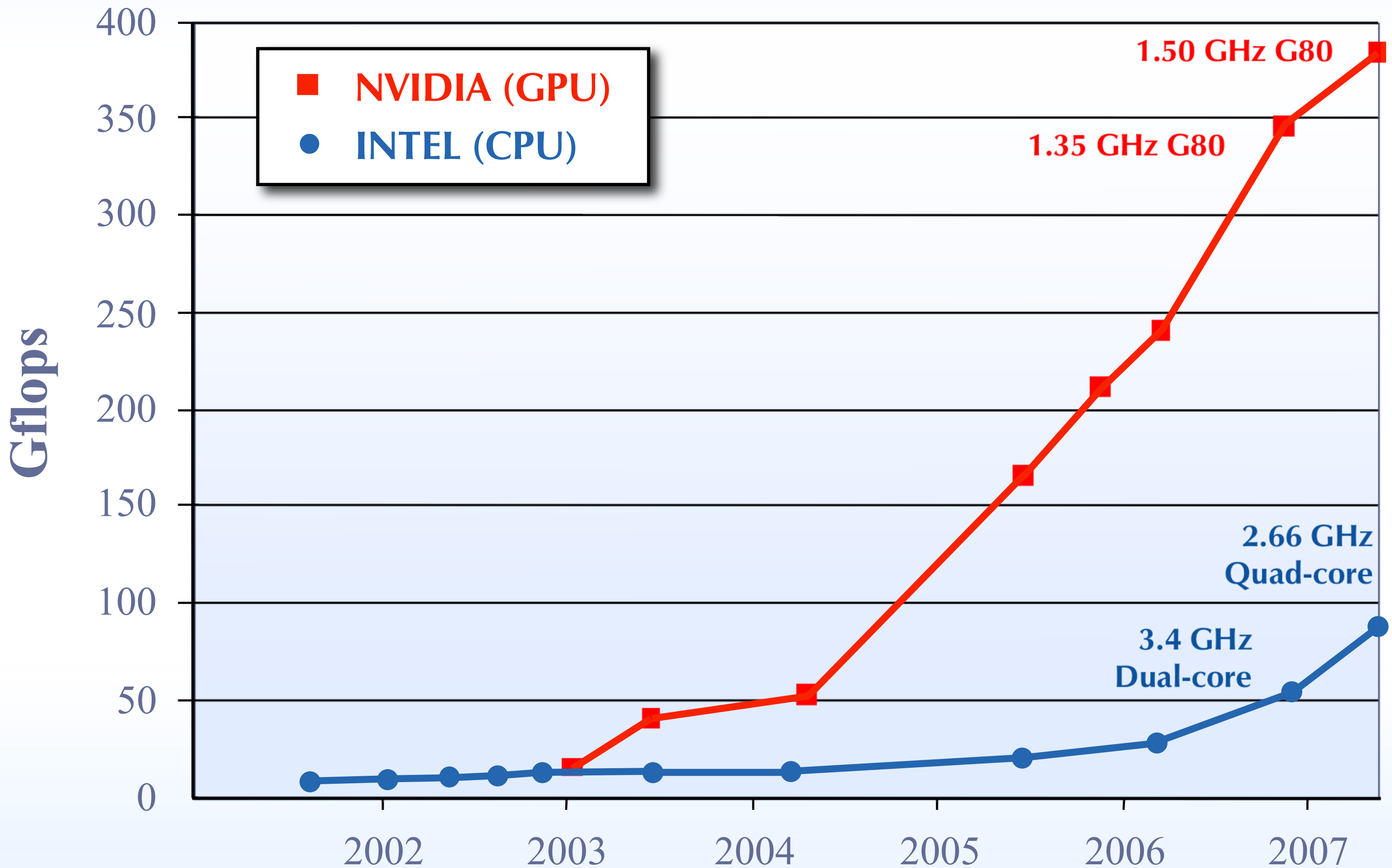


IBM Cell
(1+8 cores)

AMD Firestream
(800 cores)

Directions in Computing Technologies

New Technologies for HPC



Courtesy of John Owens (UCSD) & Ian Buck (NVIDIA)

NVIDIA: Selected Benchmarks

Application	Description	Kernel X	App X
H.264	SPEC '06 version, change in guess vector	20.2	1.5
LBM	SPEC '06 version, change to single precision and print fewer reports	12.5	12.3
FEM	Finite element modeling, simulation of 3D graded materials	11.0	10.1
RPES	Rys polynomial equation solver, 2-electron repulsion integrals	210.0	79.4
PNS	Petri net simulation of a distributed system	24.0	23.7
LINPACK	Single-precision implementation of saxpy, used in Gaussian elimination routine	19.4	11.8
TRACF	Two Point Angular Correlation Function	60.2	21.6
FDTD	Finite-difference time domain analysis of 2D electromagnetic wave propagation	10.5	1.2
MRI-Q	Computing a matrix Q, a scanner's configuration in MRI reconstruction	457.0	431.0

* *For GeForce 8800 @ 346 GF (SP), W-m. Hwu et al., 2007*

Benchmarks: Direct SCF Calculations*

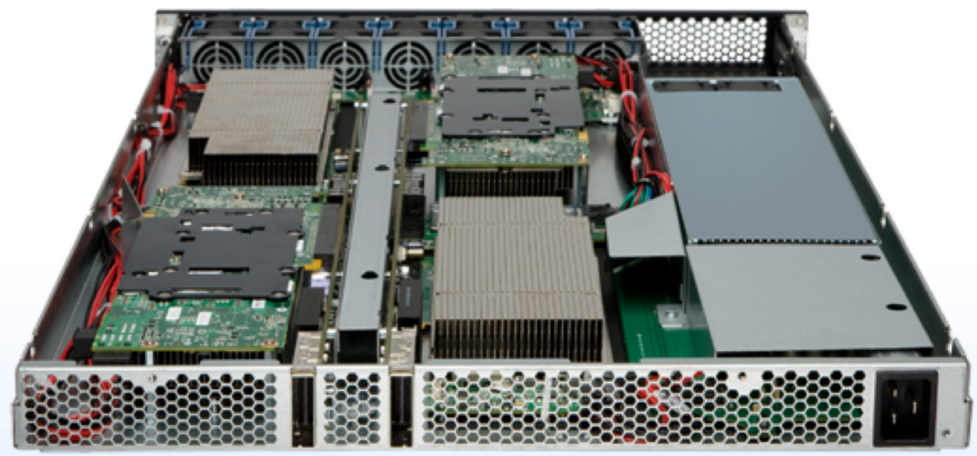
	Time/Iter (s)		Energy		
Molecule	**GPU**	**CPU****	**GPU**	**CPU****	**Speedup**
Caffeine	0.16	4.1	-1605.91827	-1605.91825	**25**
Cholesterol	1.36	67.4	-3898.82189	-3898.82189	**50**
Buckyball	7.32	279.4	-10709.0757	-10709.0839	**40**
Taxol	4.91	269.2	-12560.6830	-12560.6828	**55**
Valinomycin	8.44	691.2	-20351.9813	-20351.9904	**80**

Differences due to use of 32-bit precision,
will be eliminated in 64-bit version of INVIDIA chip

* *GeForce 8800 @ 346 GF (SP), I. Ufimtsev and T. Martinez, CiSE* **10**, 26-34 (2008).
** *Using GAMESS on AMD Opteron 175 CPU.*

NVIDIA: Tesla S1070

- 4 Tesla T10s
- Frequency: 1.44 GHz
- 960 cores (240/T10)
- Performance
  - SP: 4.14 TF
  - DP: 0.34 TF
- 16 GB memory (4/T10)
- 408 GB/s memory bandwidth (104/T10)
- CUDA programming environment



Memory Subsystem

- **Memory Wall**
  - Limitation on computation speed cause by the growing disparity between processor speed and memory latency and bandwidth
  - From 1986 to 2000, processor speed increased at an annual rate of 55%, while memory speed improved by only 10% per year
- **Memory Subsystem**
  - Caches
    - Two to three levels: L1–L3
    - On chip (*faster*) and off chip (*slower*)
  - Main memory
    - DDR2: 3.2–8.5 GB/s
    - DDR3: 6.4–12.8 GB/s
- **Issue**
  - Memory latency and bandwidth limitations make it difficult to achieve major fraction of peak performance of chip

Communications Subsystem

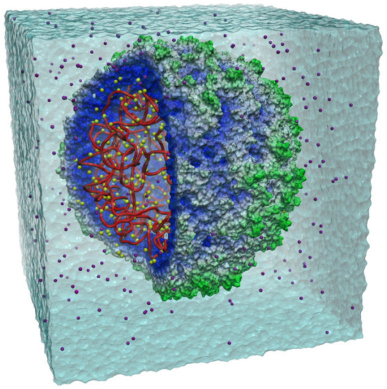
- **Communications Fabric**
  - Infiniband
    - Standard for HPC systems
    - Used in TACC's Ranger (Sun) system
  - SeaStar2+: Cray's proprietary interconnect
  - IBM working on next generation (proprietary) interconnect
- **Issue**
  - Latency and bandwidth limitations make it difficult to scale science and engineering applications to large numbers of processors

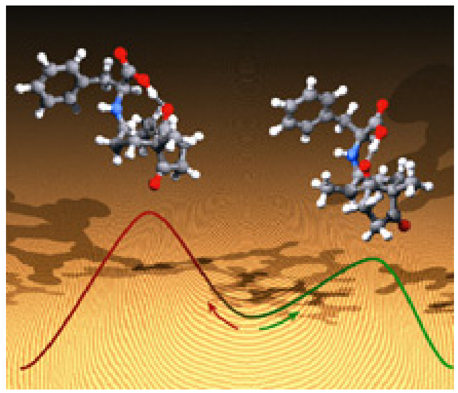
Era of Petascale Computing

Science @ Petascale

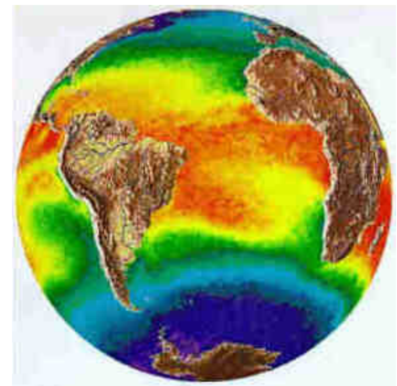
Petascale computing will enable advances in a broad range of science and engineering disciplines:

Molecular Science

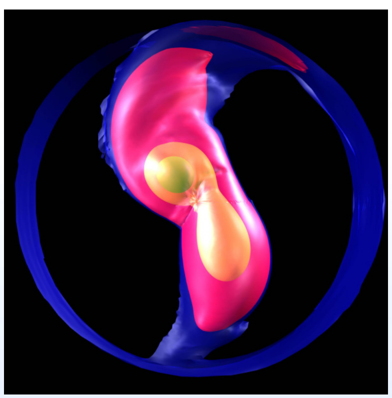




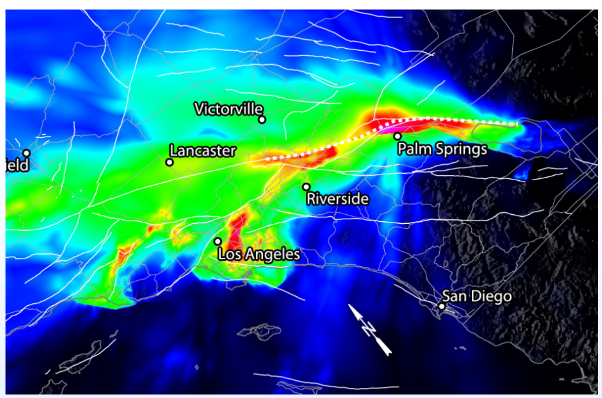
Weather & Climate Forecasting



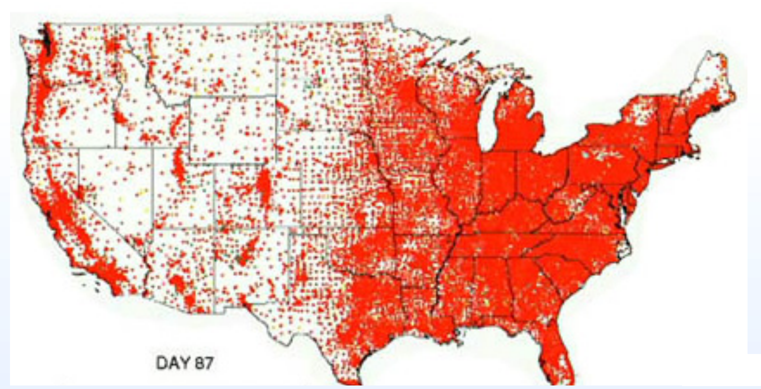
Astronomy



Earth Science

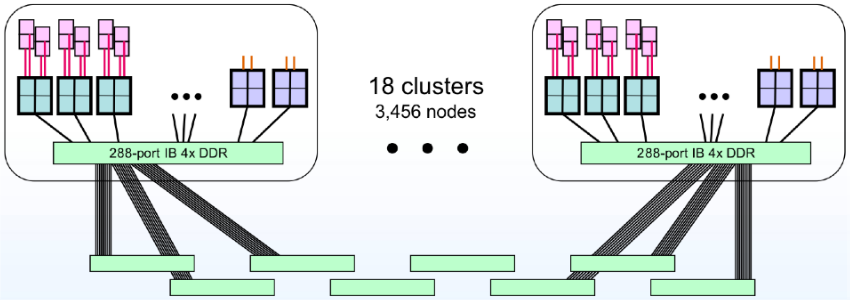


Health



LANL Roadrunner Computer System

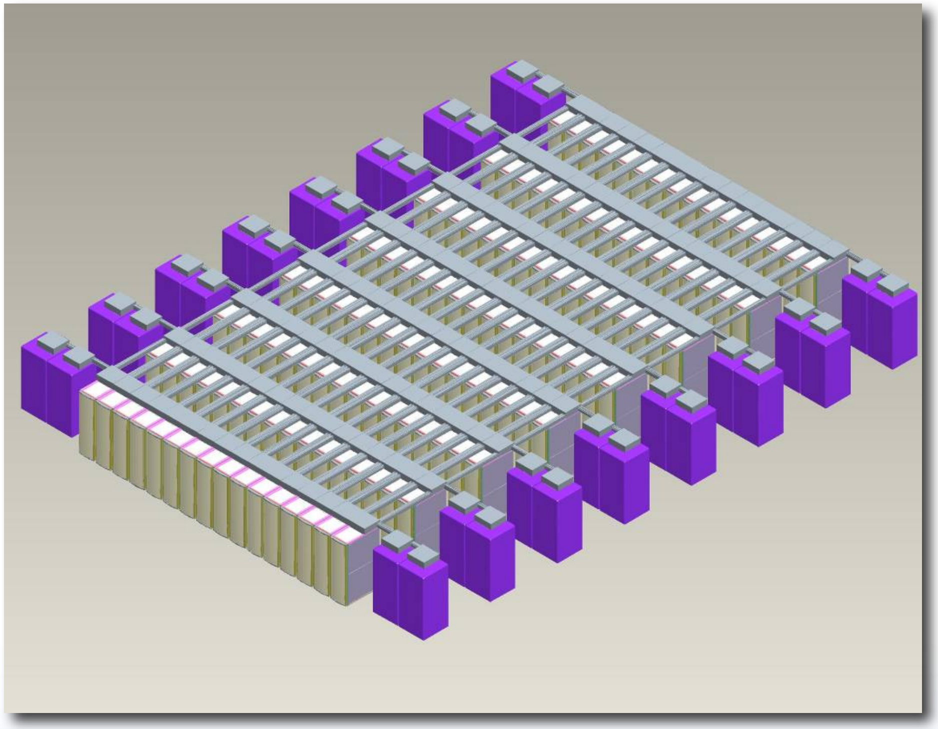
- **Computing resources**
  - 12,960 IBM PowerXCell 8i accelerators (116,640 cores)
  - 6,480 AMD dual-core Opterons (12,960 cores)
  - 1.46 PF peak
  - 1.1 Petaflop/s Linpack
- **Memory**
  - 52 TB (accelerators)
  - 104 TB total
- **Electrical power**
  - 3.9 MW (maximum)
  - ≥ 250 Megaflops/Watt
- **Floor space**
  - 296 racks, 6800 ft^2



IBM Roadrunner Petascale System

ORNL Jaguar Computer System

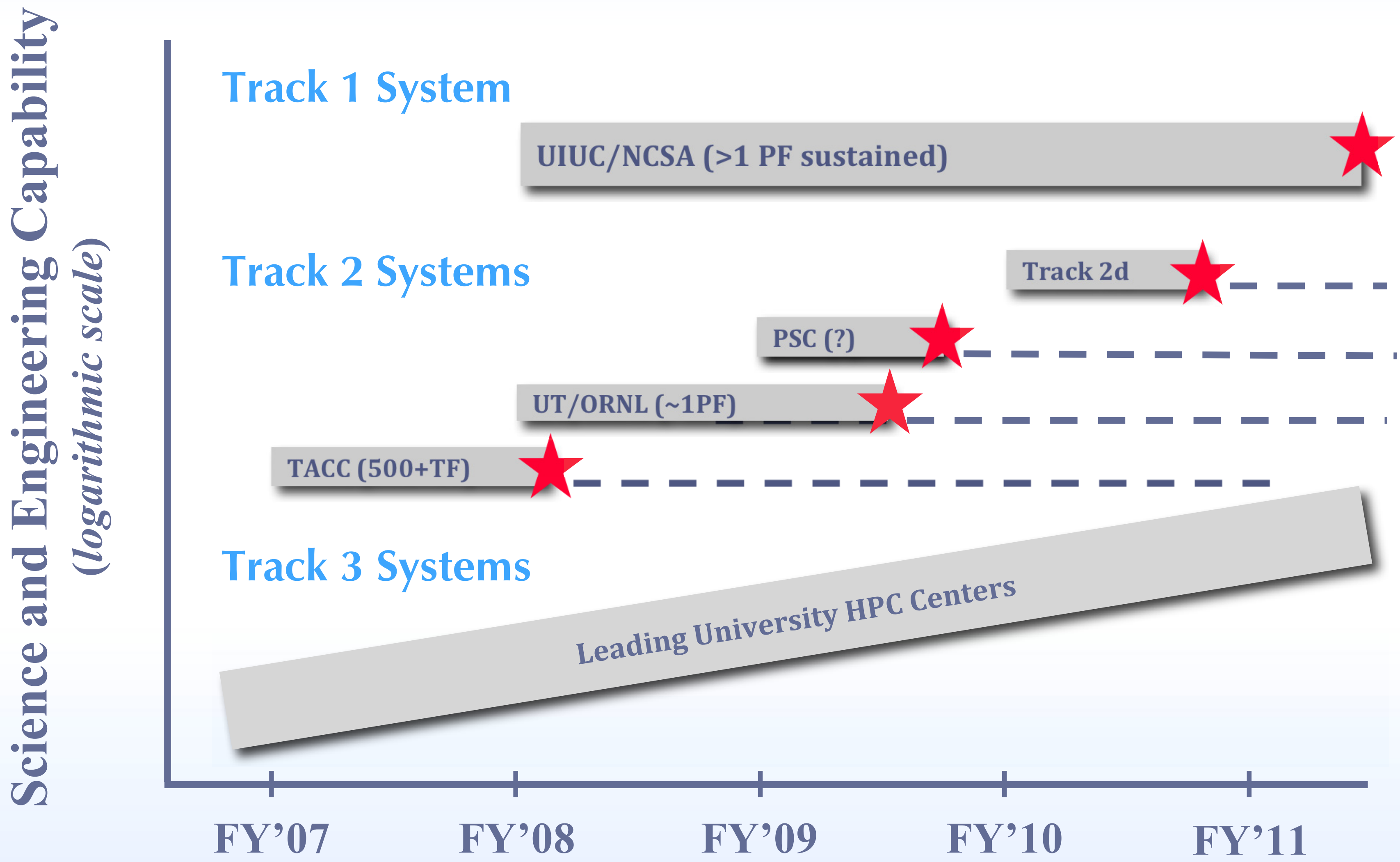
- **Computing resources**
  - 37,544 AMD quad-core Opterons
  - 150,176 cores
  - 1.38 PF peak
  - 1.06 Petaflop/s Linpack
- **Memory**
  - 300 TB
- **I/O Storage and Bandwidth**
  - 10 PB
  - 240 GB/s
- **Interconnect Bandwidth**
  - 374 TB/s
- **Floor space**
  - 4400 ft^2



Cray Jaguar (XT5) Petascale System

Era of Petascale Computing

NSF's Strategy for High-end Computing



NSF's Track 2 Computing Systems

	TACC	UT-ORNL	PSC
System Attribute	**Ranger**	**Kraken**	
Status	*Operational*	*In progress*	*In progress*
Vendor	Sun	Cray	SGI
Processor	AMD	AMD	Intel
Peak Performance (TF)	579	~1,000	
Number Cores/Chip	4	6	
Number Processor Cores	62,976	~100,000	~100,000
Amount Memory (TB)	123	~100	~100
Amount Disk Storage (PB)	1.73	3.3	
External Bandwidth (Gbps)	10	20	

Blue Waters
Petascale Computing System

Goals for Blue Waters

- **Maximize Core Performance**

  … minimize number of cores needed for a given level of performance as well as lessen the impact of sections of code with limited scalability

- **Maximize Application Scalability**

  … low latency, high-bandwidth communications fabric

- **Solve Memory-intensive Problems**

  … large amount of memory

  … low latency, high-bandwidth memory subsystem

- **Solve Data-intensive Problems**

  … high-bandwidth I/O subsystem

  … large quantity of on-line disk, large quantity of archival storage

- **Provide Reliable Operation**

  … maximize system integration

  … mainframe reliability, availability, serviceability (RAS) technologies

Blue Waters Computing System

System Attribute	Abe	Blue Waters*
Vendor	Dell	IBM
Processor	Intel Xeon 5300	IBM Power7
Peak Performance (PF)	0.090	
Sustained Performance (PF)	0.005	≥1
Number of Cores/Chip	4	
Number of Processor Cores	9,600	>200,000
Amount of Memory (PB)	0.0144	>0.8
Amount of Disk Storage (PB)	0.1	>10
Amount of Archival Storage (PB)	5	>500
External Bandwidth (Gbps)	40	100-400

* *Reference petascale computing system (no accelerators).*

Blue Waters Petascale Computing System

Blue Waters Project

Blue Waters Project
Work Breakdown Structure

- **WBS 1.0** Blue Waters System
  - **WBS 1.1** IBM HW & SW
  - **WBS 1.2** Hardware (non-IBM)
  - **WBS 1.3** Software (non-IBM)
  - **WBS 1.4** Integ., Test & Trans
- **WBS 2.0** Applications/Research
  - **WBS 2.1** Petascale Apps
  - **WBS 2.2** Petascale Research
- **WBS 3.0** **ETAO**
  - **WBS 3.1** Undergrad. Ed.
  - **WBS 3.2** Graduate Ed.
  - **WBS 3.3** Training
  - **WBS 3.4** Broadening Part.
- **WBS 4.0** **Industrial Engagement**
  - **WBS 4.1** IPiPE
  - **WBS 4.2** ISV Forum
  - **WBS 4.3** Workforce Dev.
- **WBS 5.0** **Facilities**
  - **WBS 5.1** PCF
  - **WBS 5.2** Trans. From ACB
- **WBS 6.0** **Project Management**
  - **WBS 6.1** Tracking & Controls
  - **WBS 6.2** Finan., Contracts, HR
  - **WBS 6.3** Risk Management
  - **WBS 6.4** Config. Management
  - **WBS 6.5** Reporting & Reviews
  - **WBS 6.6** Communications
  - **WBS 6.7** Cybersecurity
  - **WBS 6.8** MIS Tools

***Note:** this is not an organizational chart*

Blue Waters Petascale Computing System

Great Lakes Consortium for Petascale Computation

Goal: Facilitate the widespread and effective use of petascale computing to address frontier research questions in science, technology and engineering at research, educational and industrial organizations across the region and nation.

Charter Members

Argonne National Laboratory
Fermi National Accelerator Laboratory
Illinois Math and Science Academy
Illinois Wesleyan University
Indiana University*
Iowa State University
Illinois Mathematics and Science Academy
Krell Institute, Inc.
Los Alamos National Laboratory
Louisiana State University
Michigan State University*
Northwestern University*
Parkland Community College
Pennsylvania State University*
Purdue University*
The Ohio State University*
Shiloh Community Unit School District #1
Shodor Education Foundation, Inc.
SURA – 60 plus universities
University of Chicago*
University of Illinois at Chicago*
University of Illinois at Urbana-Champaign*
University of Iowa*
University of Michigan*
University of Minnesota*
University of North Carolina–Chapel Hill
University of Wisconsin–Madison*
Wayne City High School

** CIC universities*

Petascale Computing Facility

Partners

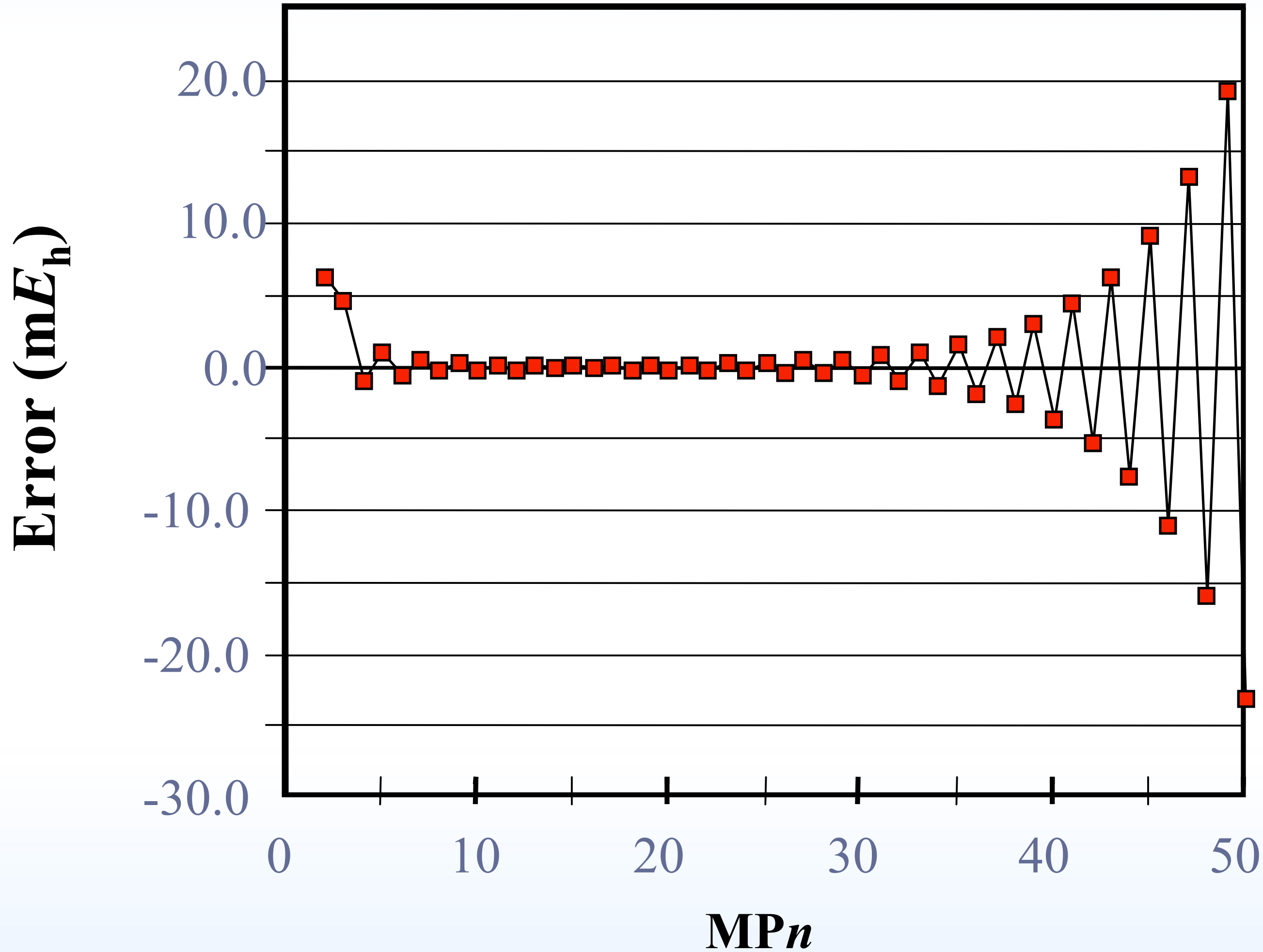
EYP MCF/
Gensler
IBM
Yahoo!

- **Modern Data Center**
  - 90,000+ ft^2 total
  - 20,000 ft^2 machine room
- **Energy Efficiency**
  - LEED certified (silver)
  - Efficient cooling system

www.ncsa.uiuc.edu/BlueWaters

Challenges in Petascale Computing

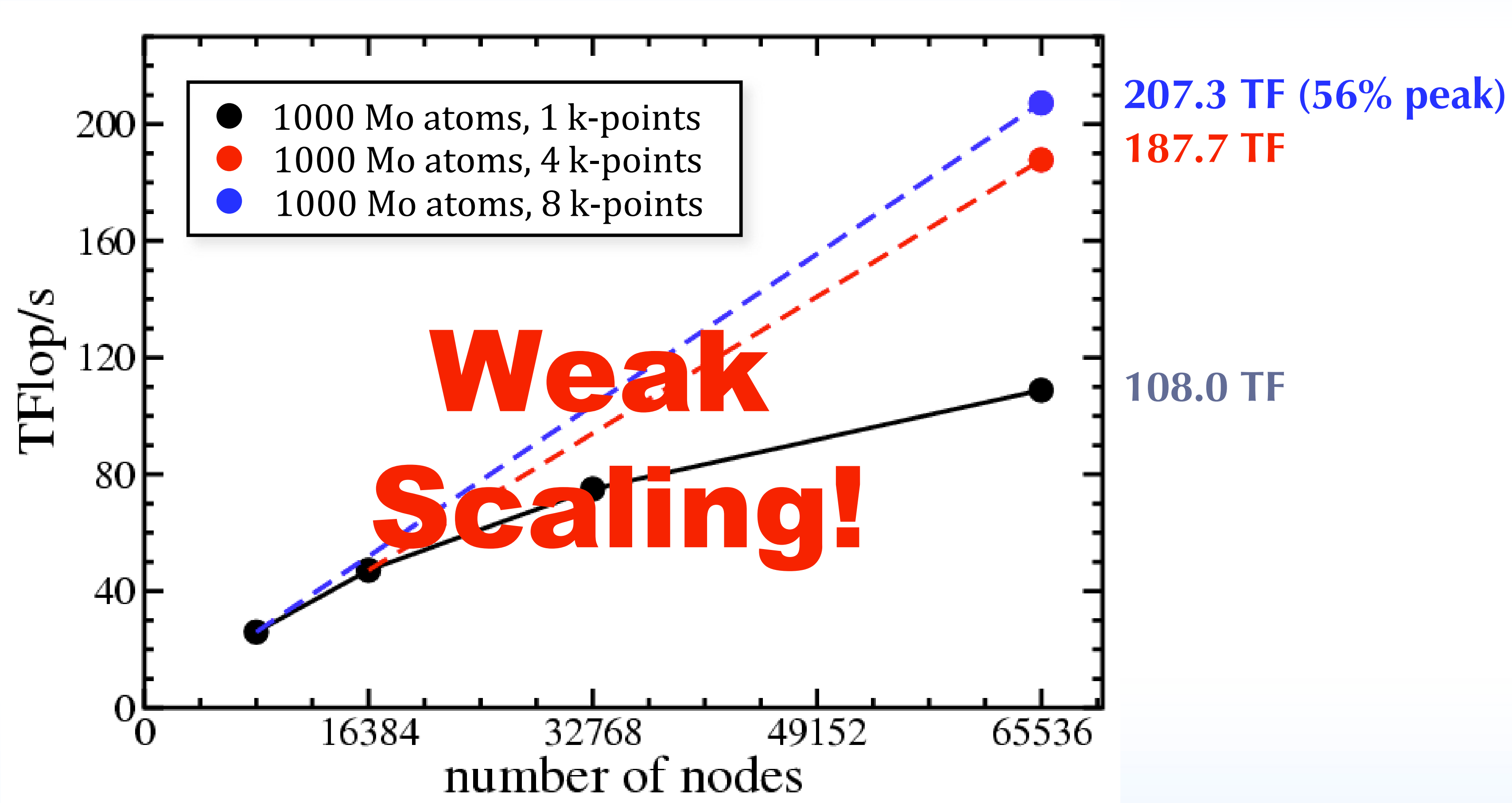
Accuracy of Computational Models



Perturbation expansion not converging!

Energy of HF(aug-cc-pVDZ): Olsen et al., J. Chem. Phys. ***105****, 5082 (1996)*

Scalability of Algorithms



(F. Gygi *et al.*, "Large-Scale Electronic Structure Calculations of High-Z Metals on Blue Gene/L Platform," Proceedings of Supercomputing, 2006)

More Challenges

- **Programming Models and Languages**

  … *will MPI be adequate*
  - PGAS (partitioned global address space) programming model
  - Universal parallel C (UPC), Co-array Fortran (CAF)
- **New Computing Technologies**

  … *new/revised algorithms will be needed*
  - Multicore and many-core chips
  - Heterogeneous multicore/many-core chips
- **Enhanced Reliability**

  … *need to minimize impact of/ride through failure*
  - Systems level (e.g., virtualization)
  - Applications level

Questions?