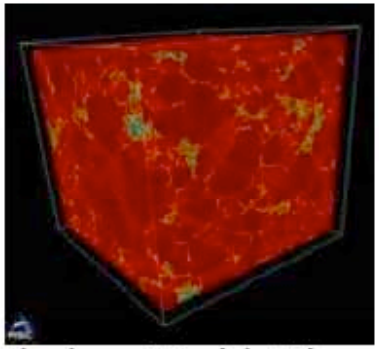
Fiberoptic Interconnect Opportunities in Supercomputers & High End Servers

NEFC FiberFest
May 11, 2009

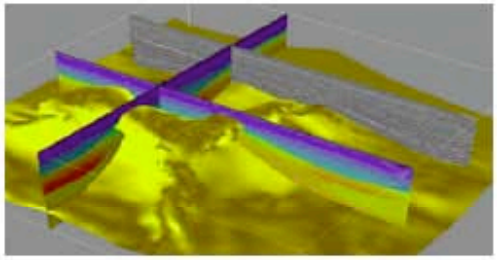
Alan Benner, bennera@us.ibm.com
IBM Corp.

Supercomputers are used to gain more insight into complex systems

- Improve understanding – **significantly larger scale, more complex and higher resolution models; new science applications**
- Multiscale and multiphysics – **From atoms to mega-structures; coupled applications**
- Shorter time to solution – **Answers from months to minutes**



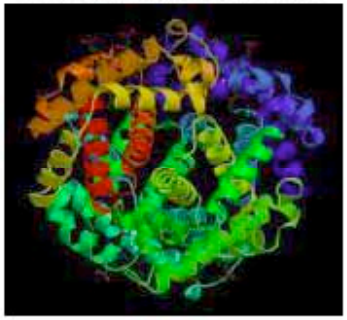
Physics – Materials Science
Molecular Dynamics



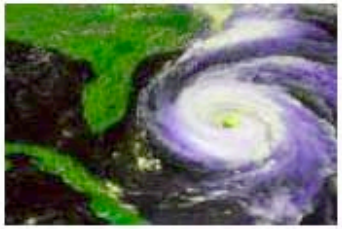
Geophysical Data Processing
Upstream Petroleum

Biological
Modeling – Brain Science

Life Sciences: In-Silico
Trials, Drug Discovery

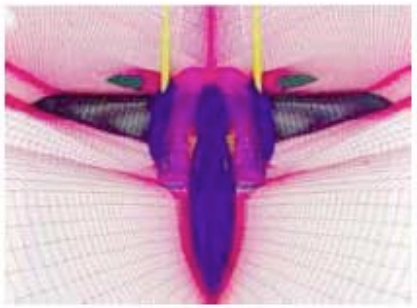


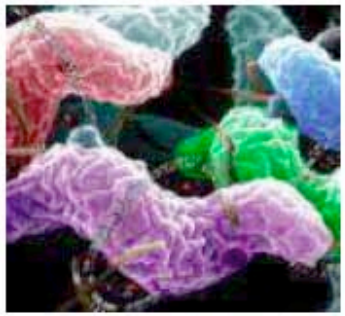
Financial Modeling
Streaming Data Analysis



Environment and Climate Modeling

Computational Fluid Dynamics





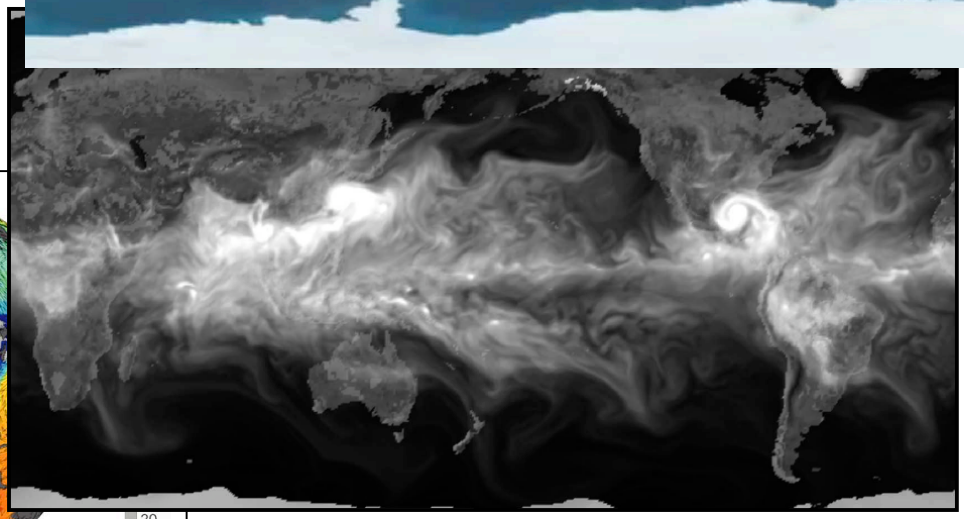
Life Sciences: Sequencing

Supercomputer performance is improving steadily

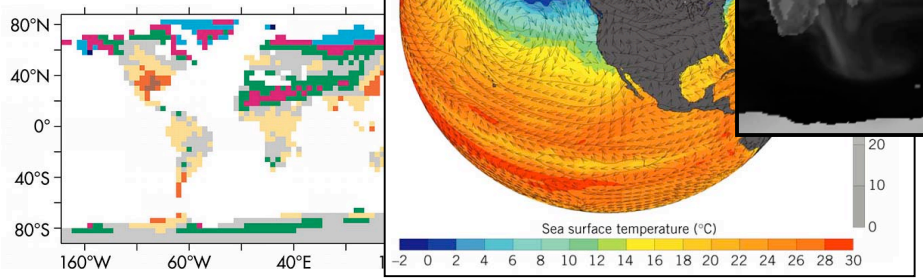
An example:

Weather Simulation

~2009



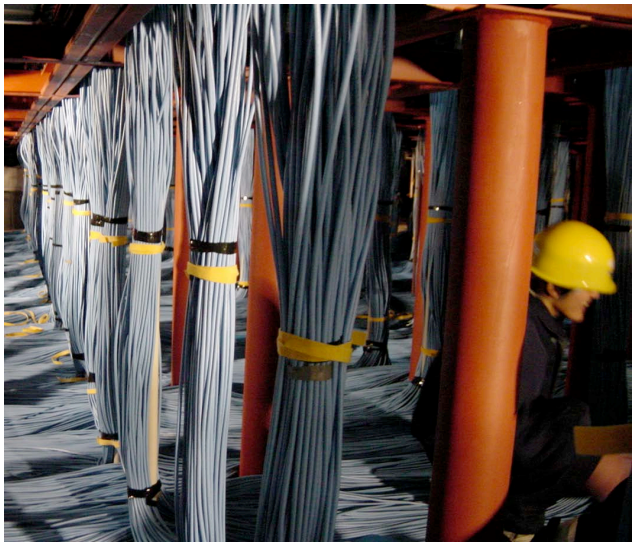
~2005

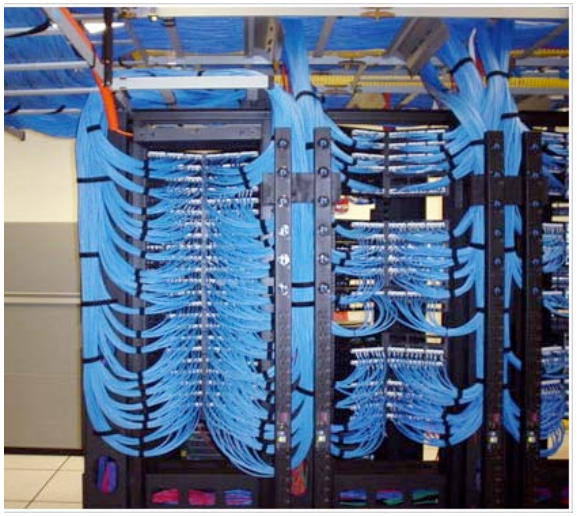


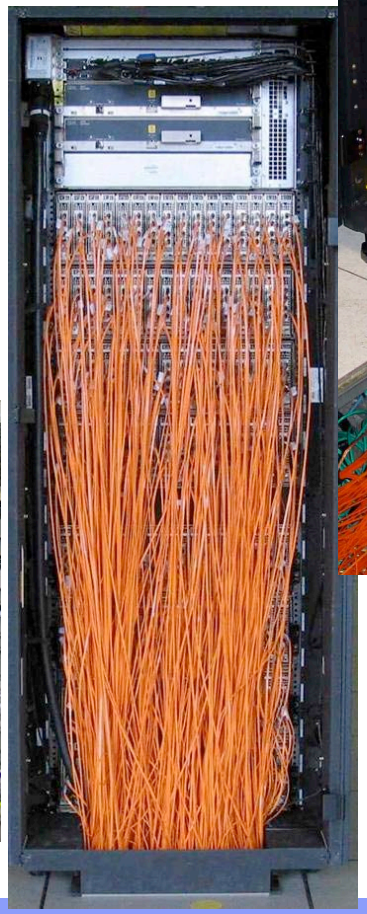
~1995 ~2000

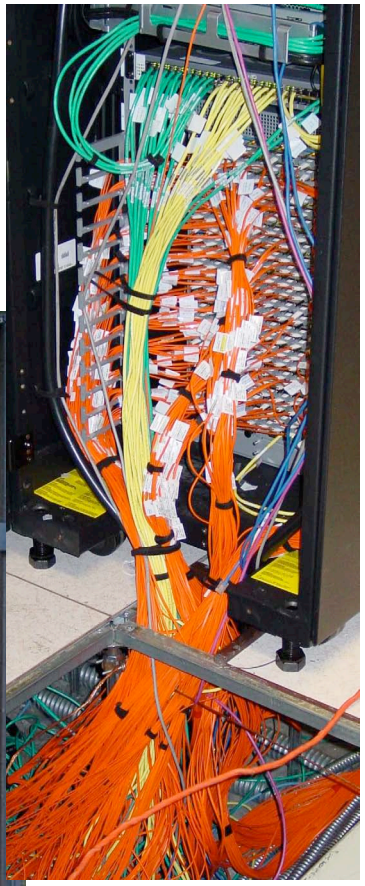
Supercomputers = CPUs + DIMMs + Power + Cooling + Interconnect

Supercomputers & High-End Servers have many many cables. A few real-world scenarios:









Top500 list - Summary

- **www.top500.org: world's fastest machines at doing linear algebra benchmark**
- **11/08 includes 2 machines at >1 PFLOP/s. Top 6 machines together add to >4 PFLOPs.**
- **3.12 million total CPU cores in all 500 systems – roughly $1.5B to $3B worth**
  - ƒA significant slice of the whole IT hardware market

32nd List: The TOP10

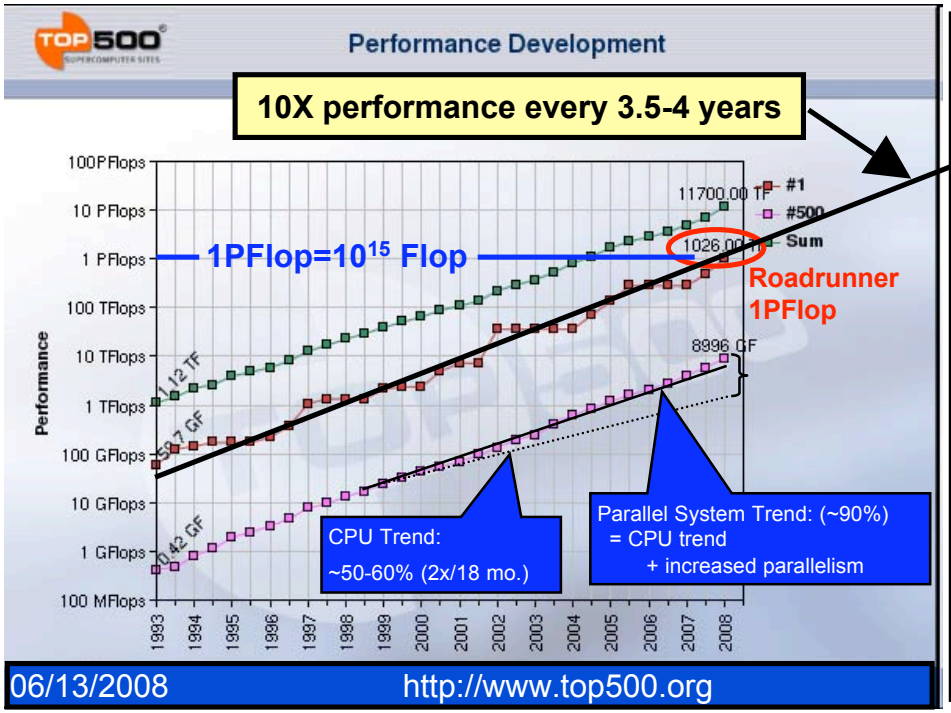
Rank	Site	Manufacturer	Computer	Country	Cores	Rmax [Tflops]	Power [MW]
1	DOE/NNSA/LANL	IBM	Roadrunner - BladeCenter QS22/LS21	USA	129600	1105.0	2.48
2	Oak Ridge National Laboratory	Cray Inc.	Jaguar - Cray XT5 QC 2.3 GHz	USA	150152	1059.0	6.95
3	NASA/Ames Research Center/NAS	SGI	Pleiades - SGI Altix ICE 8200EX	USA	51200	487.0	2.09
4	DOE/NNSA/LLNL	IBM	eServer Blue Gene Solution	USA	212992	478.2	2.32
5	Argonne National Laboratory	IBM	Blue Gene/P Solution	USA	163840	450.3	1.26
6	Texas Advanced Computing Center/ Univ. of Texas	Sun	Ranger - SunBlade x6420	USA	62976	433.2	2.0
7	NERSC/LBNL	Cray Inc.	Franklin - Cray XT4	USA	38642	266.3	1.15
8	Oak Ridge National Laboratory	Cray Inc.	Jaguar - Cray XT4	USA	30976	205.0	1.58
9	NNSA/Sandia National Laboratories	Cray Inc.	Red Storm - XT3/4	USA	38208	204.2	2.5
10	Shanghai Supercomputer Center	Dawning	Dawning 5000A, Windows HPC 2008	China	30720	180.6	

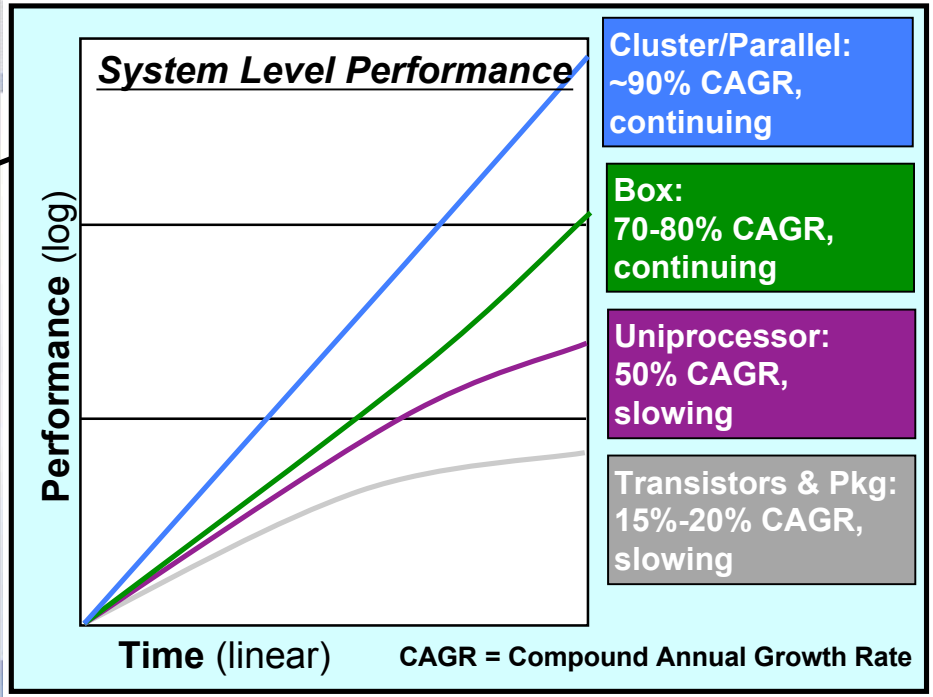
TOP500 SUPERCOMPUTER SITES

H. Meuer, E. Strohmaier, J. Dongarra, H. Simon

Top500.org – SC08 BOF Presentation

Bandwidth Must Increase to Sustain System Performance

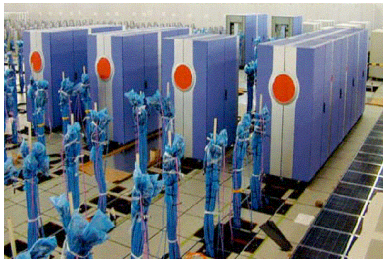




- **Moore's Law (at the system performance level) no longer comes just from improvements at the chip level**
  - ƒ Parallel System performance increasingly comes from high-level interconnection of increasingly-parallel chips & boxes
- **BW requirements must scale with System Performance, ~1B/FLOP (memory + network)**
- **Requires exponential increases in communication bandwidth at all levels of the system**
  - ƒ Inter-rack, backplane, card, chip,..

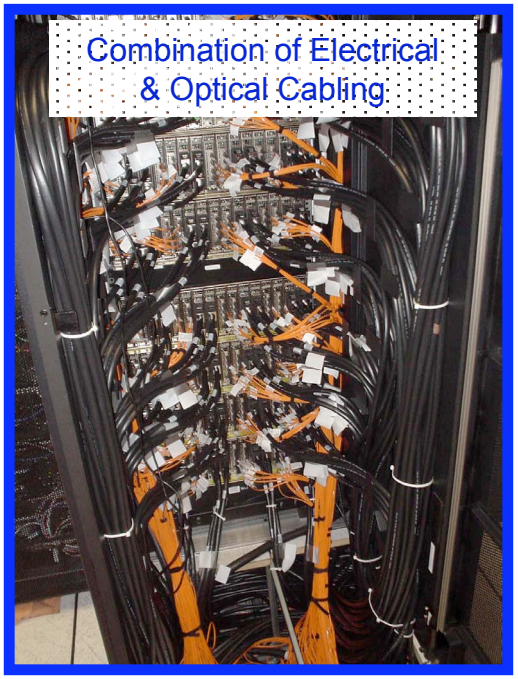
Evolution of Rack-to-Rack Optics in Supercomputers

2002



NEC Earth Simulator
- **no optics**

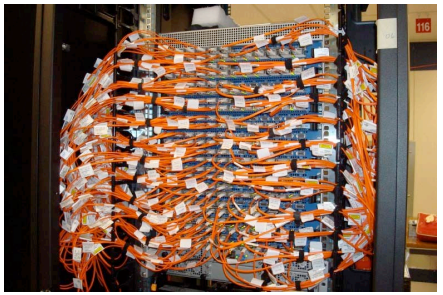
2005



IBM Federation Switch for ASCI Purple (LLNL)
- Copper for short (≤10 m) links, Optical for (20-40m)
- ~3000 parallel links, 12+12@2Gb/s/channel each

2008: 1PF/s

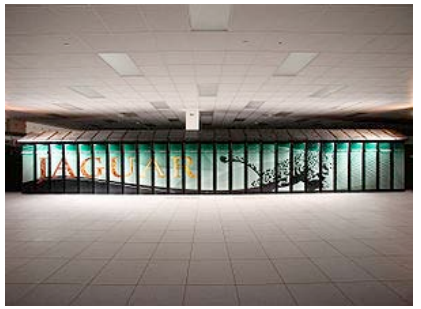
IBM Roadrunner (LLNL)



*http://www.lanl.gov/roadrunner/

- InfiniBand, (4+4)x5 Gb/s
- 55 miles of Active Optical Cables (AOCs)

Cray Jaguar(ORNL)



*http://www.nccs.gov/jaguar/

- InfiniBand
- 3 miles of Optical Cables
- Longest = 60m

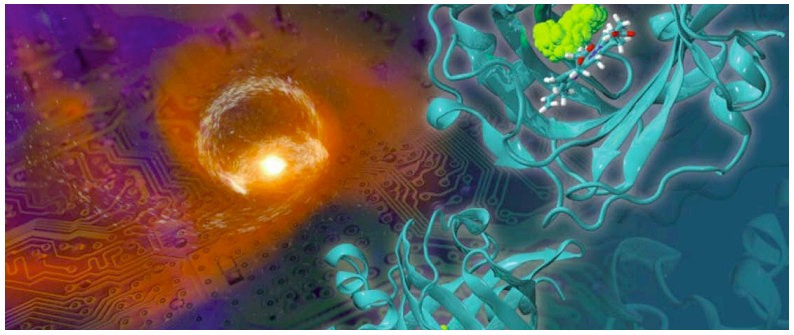
- **Future directions for optical cables:**
  - Lower cost (well below $1 per Gb/s)
  - Higher bitrates: 10-20 Gb/s per channel
  - More optics as BWs increase
  - Smaller footprint for O/E modules
  - Move optics closer to logic

Next Steps, 2010-2013: Practical Petascale Blue Waters System

- **Target: #1 productivity supercomputer in 2011:**
  - ƒ1-2 PetaFLOP/s Sustained (~10 PF Peak)
    - n"PetaFLOP"= 1015 Floating-point Ops/sec
- **Example Statistics:**
  - ƒMore than 200,000 cores.
  - ƒMore than 1 petabyte of memory.
  - ƒMore than 10 petabytes of user disk storage
  - ƒHalf an exabyte of archival storage.
  - ƒUp to 400 Gbps external connectivity.
- **Uses: Modeling Very Complex Systems**
  - ƒCells, Organs, and Organisms
  - ƒHurricanes, (incl. storm surge, ..)
  - ƒGalaxy formation in early universe
  - ƒEffect of Sun's corona on Earth's ionosphere
  - ƒDesign: Aircraft, Jet engines, motors, fusion,
  - ƒAtom-level New materials design
  - ƒ……
- **Reference: www.ncsa.uiuc.edu/BlueWaters/**

An architectural rendering of the new Illinois Petascale Computing Facility that will house Blue Waters.

Petascale Computing Facility



Next steps: 2013-2019: Exascale Systems

- **Roadmap to the Exascale-size systems is challenging (!), but clearly possible**
  - ƒSee: 2008 DARPA workgroup report "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems " by P. Kogge et al.,
    - nAvailable at http://www.nd.edu/~kogge/reports.html

- **A few selected system parameters:**

Year	Peak Performance	Machine Cost	Total Power Consumption
2008	1PF	$150M	2.5MW
2012	10PF	$225M	5MW
2016	100PF	$340M	10MW
2020	1000PF (1EF)	$500M	20MW

- **A few technical parameters:**

Year	Peak Performance	(Bidi) Optical Bandwidth	Optics Power Consumption	Optics Cost
2008	1PF	0.012PB/s ($1.2\%10^5$Gb/s)	0.012MW	$2.4M
2012	10PF	1PB/s (10^7Gb/s)	0.5MW	$22M
2016	100PF	20PB/sec ($2\%10^8$Gb/s)	2MW	$68M
2020	1000PF (1EF)	400PB/sec ($4\%10^9$Gb/s)	8MW	$200M

- **#1 Single biggest challenge: Power-efficient data transfer**

Summary and Take-Home points

- **Optics will play an increasing role in supercomputers as they approach the Exascale**
  - ƒ Parallel optical interconnects are fast replacing copper cables today
    - n Low cost ($1/Gb/s) is critical to wider adoption, including optical backplane circa 2012
    - n Single wavelength multimode VCSEL-based links appear to be lowest cost and lowest power ‡ Readily extensible to 10-20 Gb/s, perhaps 5mW/Gb/s
    - n Position optics near logic for largest benefits
- **If cost can be further lowered, optically-enabled circuit cards based on polymer waveguides will be deployed, circa 2012-2016**
  - ƒ Optical board manufacturing "ecosystem" needs to evolve
  - ƒ Work today is single wavelength multimode VCSEL-based, could migrate to CWDM
  - ƒ Or even singlemode DWDM with cheap Si photonics
- **Optics directly on the chip for on- and off-chip global interconnects is a future possibility**
  - ƒ Drive is power savings for communications
  - ƒ Still at an early stage, basic building blocks being developed
  - ƒ Needs to be approached from a systems level, not individual devices

- **Optical interconnect for supercomputers and other high-end systems volumes could be growing at ~~10x / 4 years, if the cost per Gb/s can be reduced by 3x/4 years at the same time.**