Teraflop Workbench

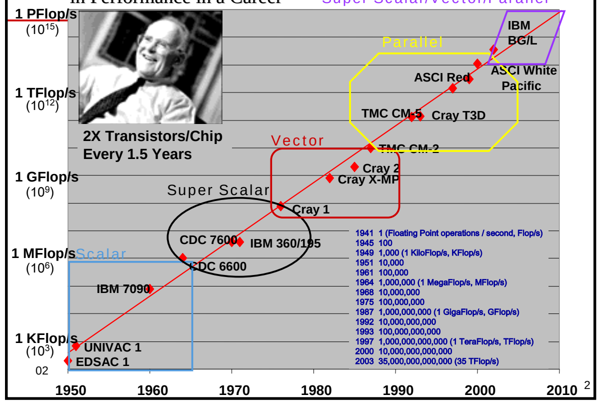
HLRS

An Overview of Supercomputers, Clusters and Grid

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory

3/18/2005

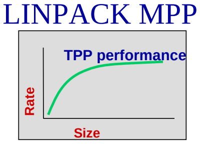
A Growth-Factor of a Billion in Performance in a Career



TOP500

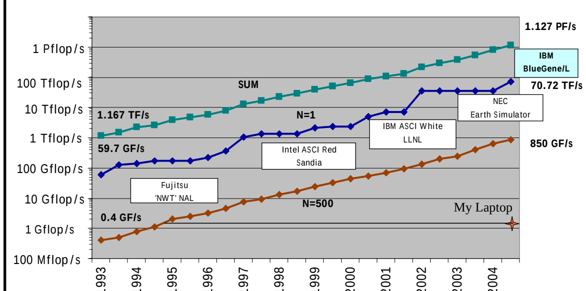
H. Meuer, H. Simon, E. Strohmaier, & JD

- 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 Mannheim, Germany in June
- All data available from **www.top500.org**

TOP500 Performance – November 2004



Vibrant Field for High Performance Computers

- Cray X1, XD1, XT3
- SGI Altix
- IBM Regatta
- IBM Blue Gene/L
- IBM eServer
- Sun
- HP
- Dawning
- Bull NovaScale
- Lanovo
- Fujitsu PrimePower
- Hitachi SR11000
- NEC SX-8
- Apple

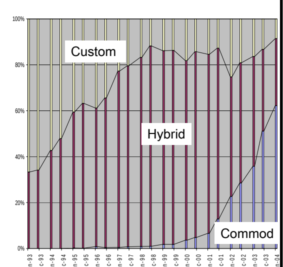
- Coming soon …
  - Cray BlackWidow
  - Galactic Computing
    - Steve Chen

Architecture/Systems Continuum

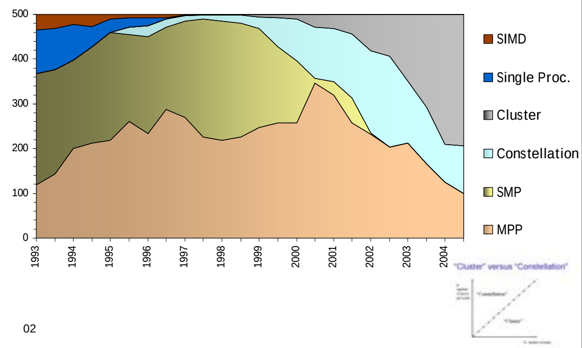
Tightly Coupled

- Custom processor with custom interconnect
  - Cray X1
  - NEC SX-8
  - IBM Regatta
  - IBM Blue Gene/L
- Commodity processor with custom interconnect
  - SGI Altix
    - Intel Itanium 2
  - Cray XT3 (Red Storm)
    - AMD Opteron
- Commodity processor with commodity interconnect
  - Clusters
    - Pentium, Itanium, Opteron, Alpha
    - GigE, Infiniband, Myrinet, Quadrics
  - NEC TX7
  - IBM eServer
  - Dawning

Loosely Coupled

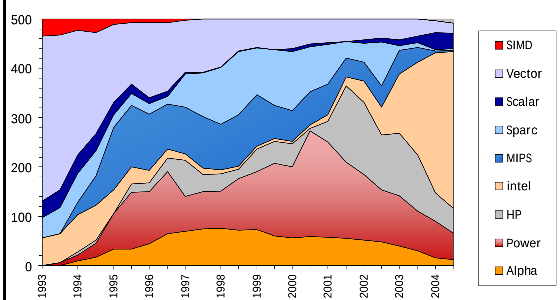


Architectures / Systems



02

Processor Types

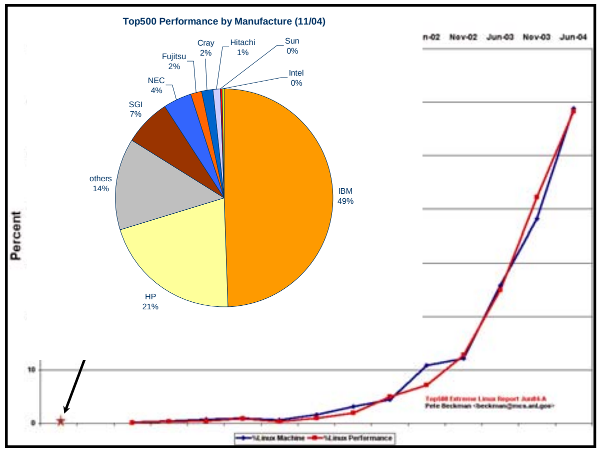


02

Commodity Processors

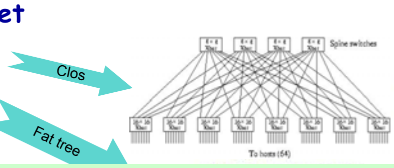
- Intel Pentium Nocona
  - 3.6 GHz, peak = 7.2 Gflop/s
  - Linpack 100 = 1.8 Gflop/s
  - Linpack 1000 = 3.1 Gflop/s
- AMD Opteron
  - 2.2 GHz, peak = 4.4 Gflop/s
  - Linpack 100 = 1.3 Gflop/s
  - Linpack 1000 = 3.1 Gflop/s
- Intel Itanium 2
  - 1.5 GHz, peak = 6 Gflop/s
  - Linpack 100 = 1.7 Gflop/s
  - Linpack 1000 = 5.4 Gflop/s
- HP PA RISC
- Sun UltraSPARC IV
- HP Alpha EV68
  - 1.25 GHz, 2.5 Gflop/s peak
- MIPS R16000

02



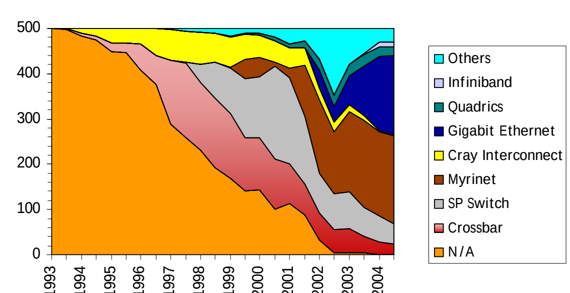
Commodity Interconnects

- Gig Ethernet
- Myrinet
- Infiniband
- QsNet



	Switch topology	Cost NIC	Cost Sw/node	Cost Node	MPI Lat (us) / 1-way MB/s / Bi-Dir MB/s
Gigabit Ethernet	Bus	$ 50	$ 50	$ 100	30 / 100 / 150
SCI	Torus	$1,600	$ 0	$1,600	5 / 300 / 400
QsNetII (R)	Fat Tree	$1,200	$1,700	$2,900	3 / 880 / 900
QsNetII (E)	Fat Tree	$1,000	$ 700	$1,700	3 / 880 / 900
Myrinet (D card)	Clos	$ 595	$ 400	$ 995	6.5 / 240 / 480
Myrinet (E card)	Clos	$ 995	$ 400	$1,395	6 / 450 / 900
IB 4x	Fat Tree	$1,000	$ 400	$1,400	6 / 820 / 790

Interconnects / Systems



02

24th List: The TOP10

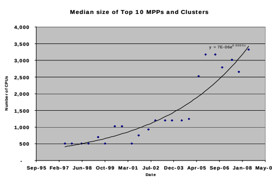
	Manufacturer	Computer	Rmax [TF/s]	Installation Site	Country	Year	#Proc
1	IBM	BlueGene/L β-System	70.72	DOE/IBM	USA	2004	32768
2	SGI	Columbia Altix, Infiniband	51.87	NASA Ames	USA	2004	10160
3	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	5120
4	IBM	MareNostrum BladeCenter JS20, Myrinet	20.53	Barcelona Supercomputer Center	Spain	2004	3564
5	CCD	Thunder Itanium2, Quadrics	19.94	Lawrence Livermore National Laboratory	USA	2004	4096
6	HP	ASCI Q AlphaServer SC, Quadrics	13.88	Los Alamos National Laboratory	USA	2002	8192
7	Self Made	X Apple XServe, Infiniband	12.25	Virginia Tech	USA	2004	2200
8	IBM/LLNL	BlueGene/L DD1 500 MHz	11.68	Lawrence Livermore National Laboratory	USA	2004	8192
9	IBM	pSeries 655	10.31	Naval Oceanographic Office	USA	2004	2944
10	Dell	Tungsten PowerEdge, Myrinet	9.82	NCSA	USA	2003	2500

399 system > 1 TFlop/s; 294 machines are clusters, top10 average 8K proc; 35 in Germany

How Big Is Big?

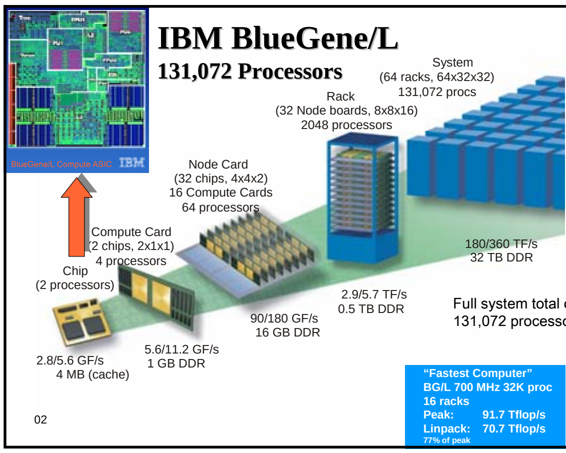
- Every 10X brings new challenges
  - 64 processors was once considered large
    - it hasn't been "large" for quite a while
  - 1024 processors is today's "medium" size
  - 8096 processors is today's "large"
    - we're struggling even here

- 100K processor systems
  - are in construction
  - we have fundamental challenges in dealing with machines of this size
  - … and little in the way of programming support



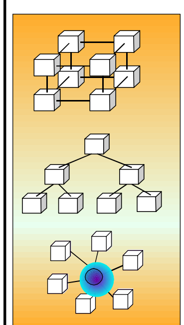
IBM BlueGene/L

131,072 Processors



"Fastest Computer"
BG/L 700 MHz 32K proc
16 racks
Peak: 91.7 Tflop/s
Linpack: 70.7 Tflop/s
77% of peak

BlueGene/L Interconnection Networks



3 Dimensional Torus
- Interconnects all compute nodes (65,536)
- Virtual cut-through hardware routing
- 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- 1 µs latency between nearest neighbors, 5 µs to the farthest
- 4 µs latency for one hop with MPI, 10 µs to the farthest
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 68TB/s total bandwidth

Global Tree
- Interconnects all compute and I/O nodes (1024)
- One-to-all broadcast functionality
- Reduction operations functionality
- 2.8 Gb/s of bandwidth per link
- Latency of one way tree traversal 2.5 µs
- ~23TB/s total binary tree bandwidth (64k machine)

Ethernet
- Incorporated into every node ASIC
- Active in the I/O nodes (1:64)
- All external comm. (file I/O, control, user interaction, etc.)

Low Latency Global Barrier and Interrupt
- Latency of round trip 1.3 µs

Control Network

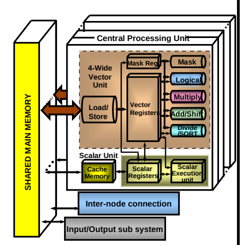
NASA Ames: SGI Altix Columbia 10,240 Processor System

- Architecture: Hybrid Technical Server Cluster
- Vendor: SGI based on Altix systems
- Deployment: Today
- Node:
  - 1.5 GHz Itanium-2 Processor
  - 512 procs/node (20 cabinets)
  - Dual FPU's / processor

- System:
  - 20 Altix NUMA systems @ 512 procs/node = 10240 procs
  - 320 cabinets (estimate 16 per node)
  - Peak: 61.4 Tflop/s ; LINPACK: 52 Tflop/s
- Interconnect:
  - FastNumaFlex (custom hypercube) within node
  - Infiniband between nodes
- Pluses:
  - Large and powerful DSM nodes
- Potential problems (Gotchas):
  - Power consumption - 100 kw per node (2 Mw total)

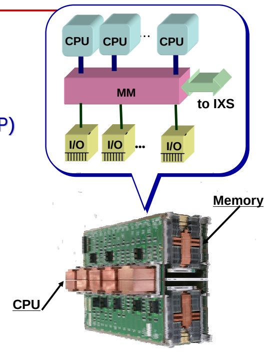
SX-8 Architecture

- Upward compatible to SX-5/SX-6
- Vector pipelines
  - 4 logical pipelines : 2GHz
  - 144KB vector register
  - Hardware support of SQRT instruction
- Scalar processor
  - 4 way superscalar RISC
- Main memory
  - 2 types of RAMs
    - DDR2-SDRAM : Large capacity 128GB/node
    - FCRAM : High-speed 64GB/node
- Multi node system
  - up to 512 nodes
  - 64 TFLOPS
- Enhanced I/O performance
  - Reduction of I/O overhead by adopting direct CPU control

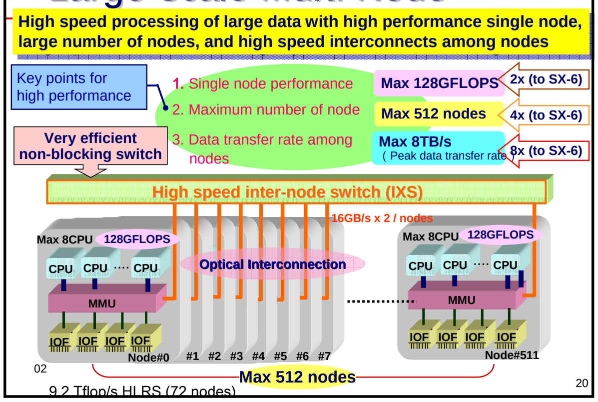


SX-8 Single Node Module

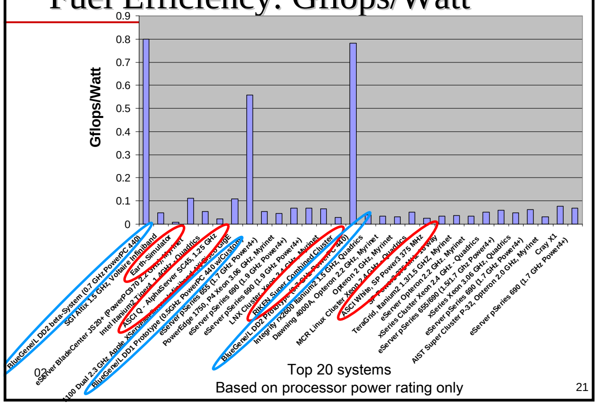
- Up to 8 CPUs/node
  - Peak Vector Performance(PVP):
    16 GFLOPS/CPU
    128 GFLOPS/node
- Symmetric multiprocessing (SMP)
- Large Capacity Memory
  - Up to 128GB
- Ultra-high memory bandwidth
  - 64GB/s per CPU
  - Total 512GB/s per node
- Large I/O throughput
  - 12.8GB/s per node



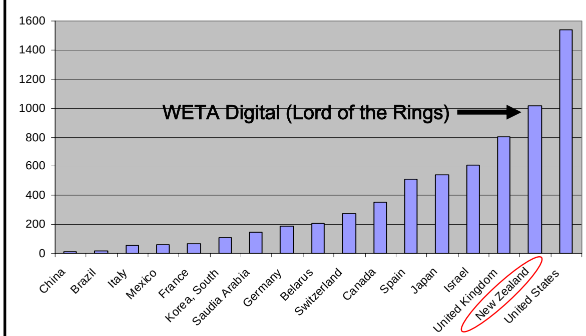
Large Scale Multi Node



Fuel Efficiency: Gflops/Watt



KFlop/s per Capita (Flops/Pop)



Important Metrics: Sustained Performance and Cost

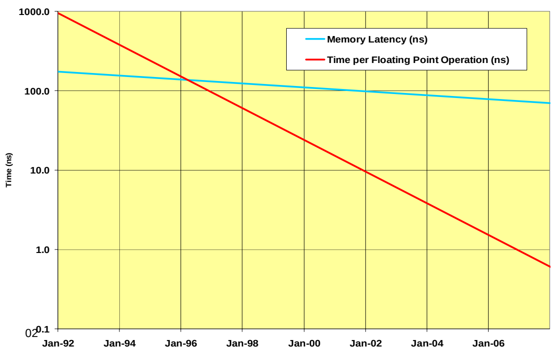
- **Commodity processors**
  - Optimized for commercial applications.
  - Meet the needs of most of the scientific computing market.
  - Provide the shortest time-to-solution and the highest sustained performance per unit cost for a broad range of applications that have significant spatial and temporal locality (good caches use).
- **Custom processors**
  - For bandwidth-intensive applications that do not cache well, custom processors are more cost effective
  - Hence offering better capacity on just those applications.

High Bandwidth vs Commodity Systems

- High bandwidth systems have traditionally been vector computers
  - Designed for scientific problems
  - Capability computing
- Commodity processors are designed for web servers and the home PC market
  (should be thankful that the manufactures keep the 64 bit fl pt)
  - Used for cluster based computers leveraging price point
- Scientific computing needs are different
  - Require a better balance between data movement and floating point operations. Results in greater efficiency.

	NEC SX-8 (NEC)	Cray X1 (Cray)	ASCI Q (HP EV68)	Intel (Dual Xeon)	VT Big Mac (Dual IBM PPC)
Year of Introduction	2005	2003	2002	2004	2003
Node Architecture	Vector	Vector	Alpha	Pentium	Power PC
Processor Cycle Time	2 GHz	800 MHz	1.25 GHz	3.6 GHz	2 GHz
Peak Speed per Processor	16 Gflop/s	12.8 Gflop/s	2.5 Gflop/s	7.2 Gflop/s	8 Gflop/s
Bytes/flop (main memory)	4	2.6	0.8	0.88	0.5

Commodity: Memory Latency and Flop Rate



02 25

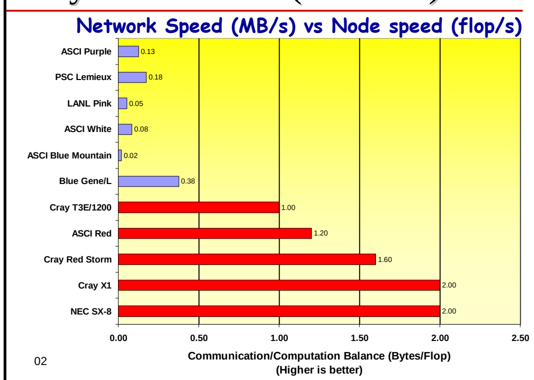
Commodity Processor Trends

	Annual increase	Typical value in 2004	Typical value in 2010	Typical value in 2020
Single-chip floating-point performance	59%	2 GFLOP/s	32 GFLOP/s	3300 GFLOP/s
Front-side bus bandwidth	23%	1 GWord/s = 0.5 word/flop	3.5 GWord/s = 0.11 word/flop	27 GWord/s = 0.008 word/flop
DRAM bandwidth	25%	100 MWord/s = 0.05 word/flop	380 MWord/s = 0.012 word/flop	3600 MWord/s =0.0011 word/flop
DRAM latency	(5.5%)	70 ns = 140 FP ops = 70 loads	50 ns = 1600 FP ops = 170 loads	28 ns = 94,000 FP ops = 780 loads

Source: *Getting Up to Speed: The Future of Supercomputing*, National Research Council, 222 pages, 2004, National Academies Press, Washington DC, ISBN 0-309-09502-6.

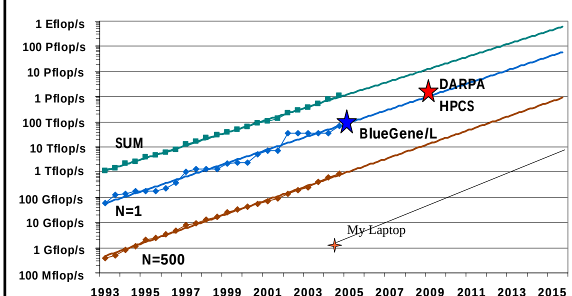
02 26

System Balance (Network)



02 27

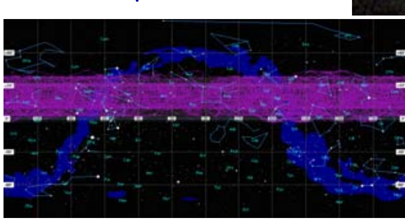
Performance Projection



02 28

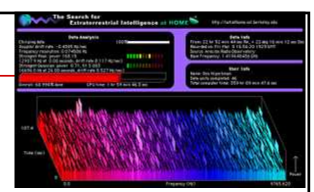
SETI@home: Global Distributed Computing

- Running on 500,000 PCs, ~1300 CPU Years per Day
  - 1.3M CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope



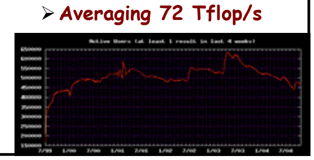
02 29

SETI@home



- Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
- When their computer is idle or being wasted this software will download ~ half a MB chunk of data for analysis. Performs about 3 Tflops for each client in 15 hours.
- The results of this analysis are sent back to the SETI team, combined with thousands of other participants.
- About 5M users

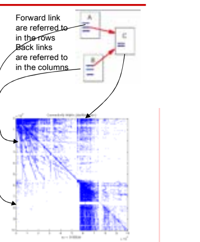
- **Largest distributed computation project in existence**
  - Averaging 72 Tflop/s



02

Google

- Google query attributes
  - 150M queries/day (2000/second)
  - 100 countries
  - 8.0B documents in the index
- Data centers
  - 100,000 Linux systems in data centers around the world
    - 15 TFlop/s and 1000 TB total capability
    - 40-80 1U/2U servers/cabinet
    - 100 MB Ethernet switches/cabinet with gigabit Ethernet uplink
  - growth from 4,000 systems (June 2000)
    - 18M queries then
- Performance and operation
  - simple reissue of failed commands to new servers
  - no performance debugging
    - problems are not reproducible



Eigenvalue problem; $Ax = \lambda x$
$n=8x10^9$
(see: MathWorks Cleve's Corner)

The matrix is the transition probability matrix of the Markov chain; $Ax = x$

Source: Monika Henzinger, Google & Cleve Moler

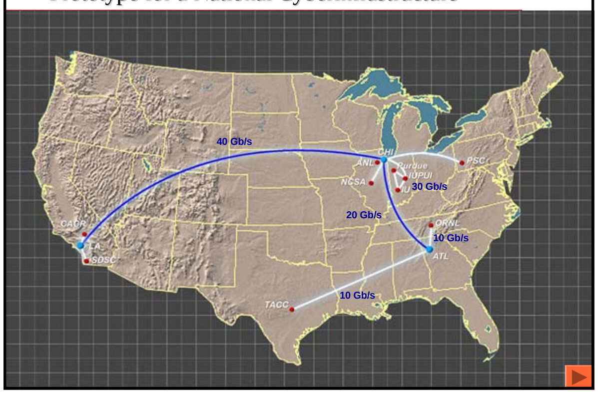
The Grid

- The Grid is about gathering resources …
  - run programs, access data, provide services, collaborate
- …To enable and exploit large scale sharing of resources
- Virtual organization
  - Loosely coordinated groups
- Provides for remote access of resources
  - Scalable
  - Secure
  - Reliable mechanisms for discovery and access
- In some ideal setting:
  - User submits work, infrastructure finds an execution target
  - Ideally you don't care where.

Science Grid Projects

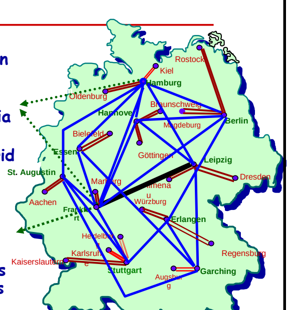
TeraGrid 2003

Prototype for a National Cyberinfrastructure



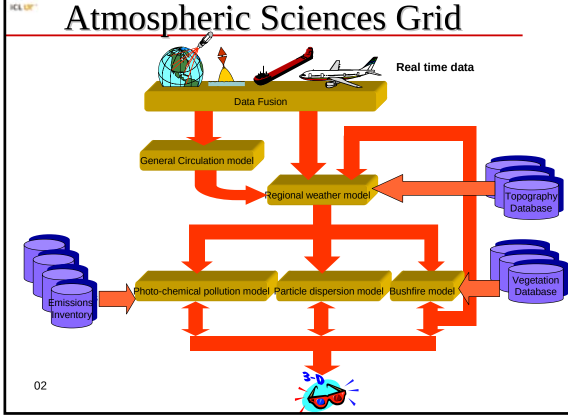
A German Grid Initiative: D-GRID

- Initially driven by the HGF centers and the DFN-Verein (2002)
- Meanwhile: More than 100 further partners in academia and industry
- Aim at a coordination of Grid activities in Germany
- Deployment of a new generation networking infrastructure (Example: Project VIOLA)
- Promotion of open standards for interfaces and protocols (GGF)

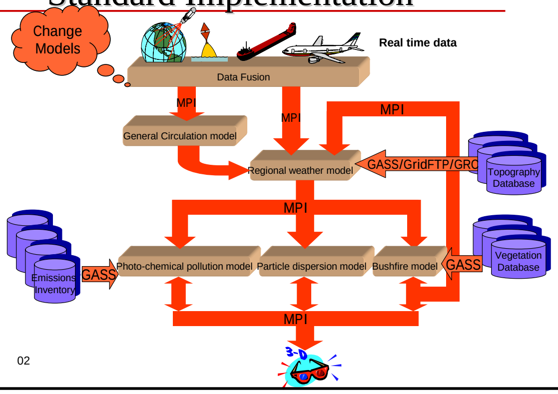


10 Gbit/s
2,4 Gbit/s
622 Mbit/s

Atmospheric Sciences Grid



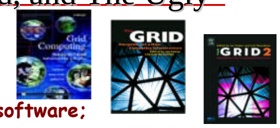
Standard Implementation



02 37

The Grid: The Good, The Bad, and The Ugly

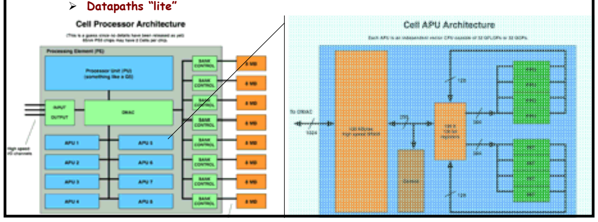
- **Good:**
  - Vision;
  - Community;
  - Developed functional software;
- **Bad:**
  - Oversold the grid concept;
  - Still too hard to use;
  - Solution in search of a problem;
  - Underestimated the technical difficulties;
  - Not enough of a scientific discipline;
- **Ugly:**
  - Authentication and security



02 38

PlayStation 3

- The PlayStation 3's CPU based on a chip codenamed "Cell"
- Each Cell contains 8 APUs.
  - An APU is a self contained vector processor which acts independently from the others.
  - 4 floating point units capable of 32 Gflop/s (8 Gflop/s each)
  - 256 Gflop/s peak!
  - 32 bit floating point only; not even IEEE
  - Datapaths "lite"



The Computing Continuum

- Each strikes a different balance
  - computation/communication coupling
- Implications for execution efficiency
- *Applications for diverse needs*
  - *computing is only one part of the story!*

02 40

Grids vs. Capability vs. Cluster Computing

- Not an "either/or" question
  - Each addresses different needs
  - Each are part of an integrated solution
- Grid strengths
  - Coupling necessarily distributed resources
    - instruments, software, hardware, archives, and people
  - Eliminating time and space barriers
    - remote resource access and capacity computing
  - Grids are not a cheap substitute for capability HPC
- Highest performance computing strengths
  - Supporting foundational computations
    - terascale and petascale "nation scale" problems
  - Engaging tightly coupled computations and teams
- Clusters
  - Low cost, group solution
  - Potential hidden costs
- Key is easy access to resources in a transparent way

02 41

Real Crisis With HPC Is With The Software

- Programming is stuck
  - Arguably hasn't changed since the 60's
- It's time for a change
  - Complexity is rising dramatically
    - highly parallel and distributed systems
      - From 10 to 100 to 1000 to 10000 to 100000 of processors!!
    - multidisciplinary applications
- A supercomputer application and software are usually much more long-lived than a hardware
  - Hardware life typically five years at most.
  - Fortran and C are the main programming models
- Software is a major cost component of modern technologies.
  - The tradition in HPC system procurement is to assume that the software is free.
- We don't have many great ideas about how to solve this problem.

02 42

Collaborators

- **TOP500**
  - **H. Meuer, Mannheim U**
  - **H. Simon, NERSC**
  - **E. Strohmaier, NERSC**