The 14th Symposium on Computer Architecture and High Performance Computing

Vitoria/ES-Brazil-October 28-30, 2002

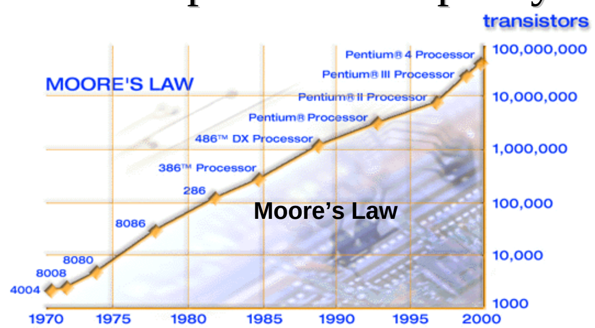
SBAC-PAD 2002

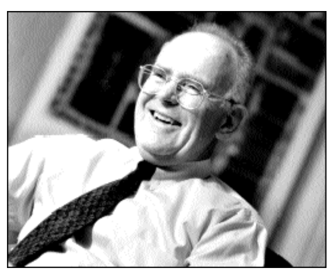
High Performance Computing, Computational Grid, and Numerical Libraries

Jack Dongarra
Innovative Computing Lab
University of Tennessee
http://www.cs.utk.edu/~dongarra/

1

Technology Trends: Microprocessor Capacity





2X transistors/Chip Every 1.5 years

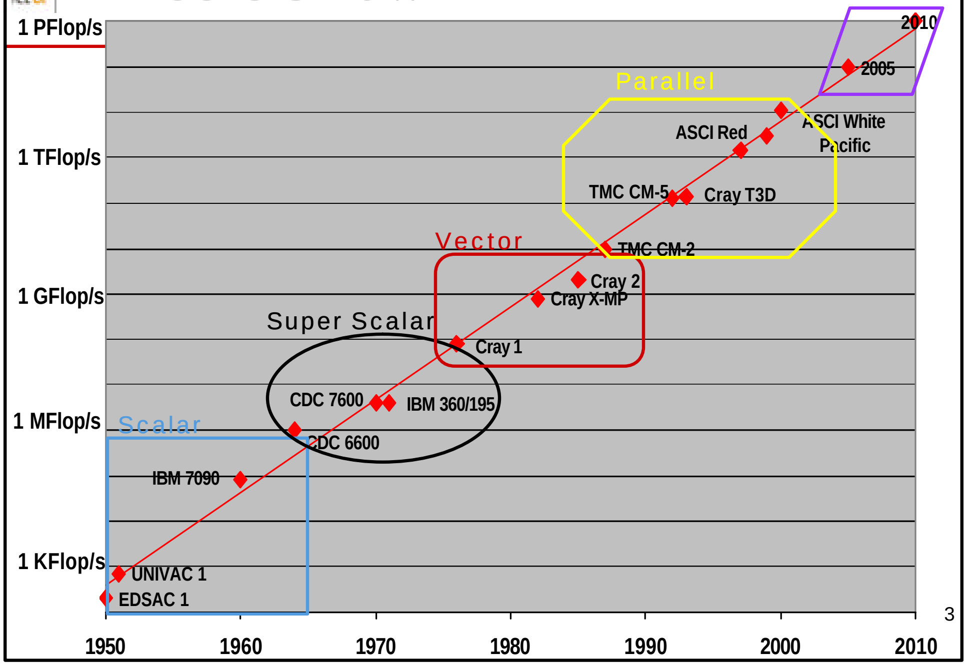
Called "Moore's Law"

Microprocessors have become smaller, denser, and more powerful.
Not just processors, bandwidth, storage, etc

Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

2

Moore's Law



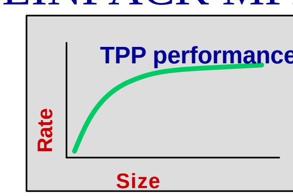
3

TOP500 SUPERCOMPUTER

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

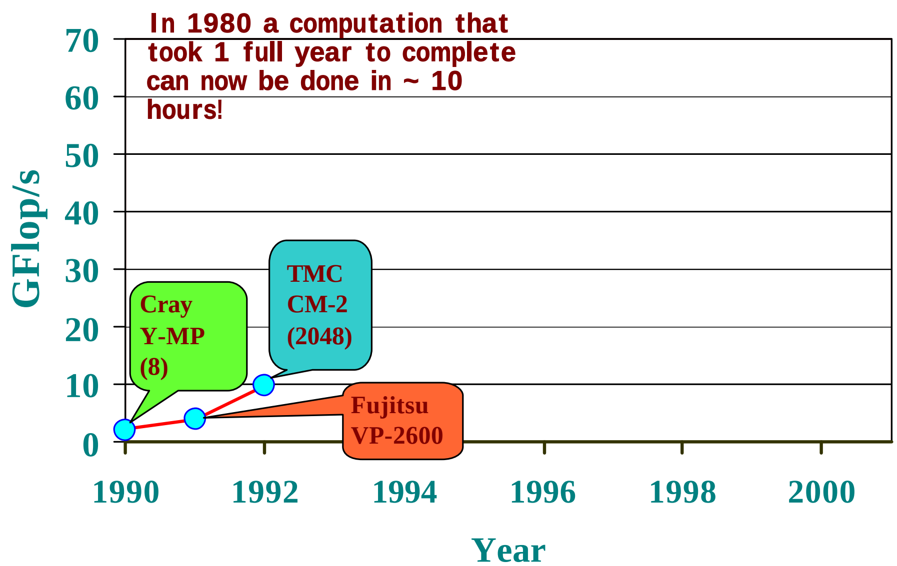
  SC'xy in the States in November

  Meeting in Mannheim, Germany in June

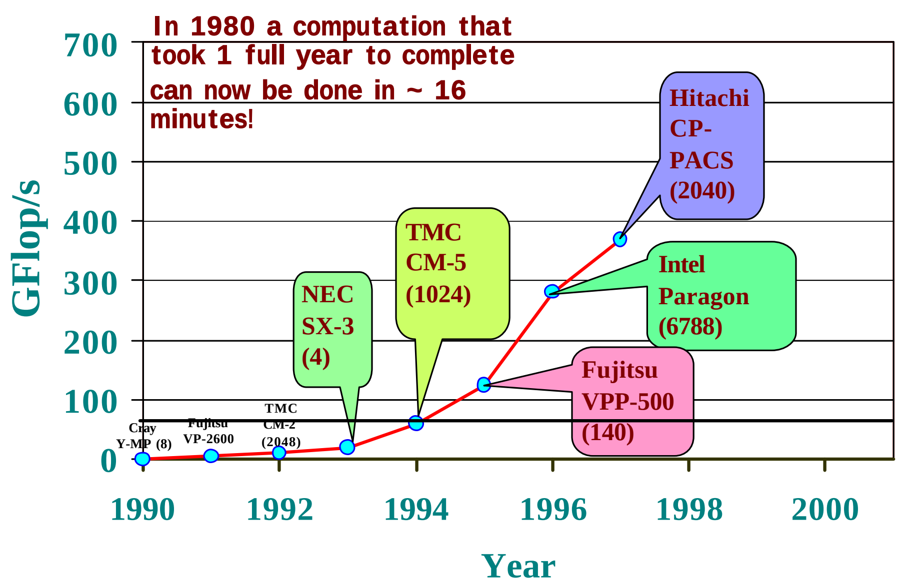
- All data available from **www.top500.org**

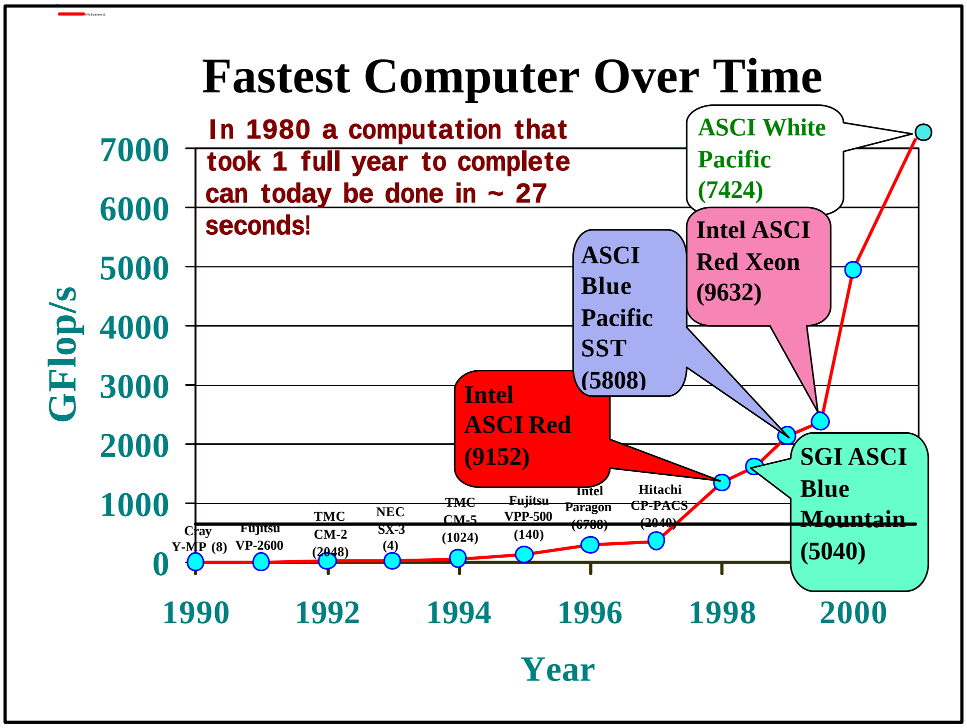
4

Fastest Computer Over Time

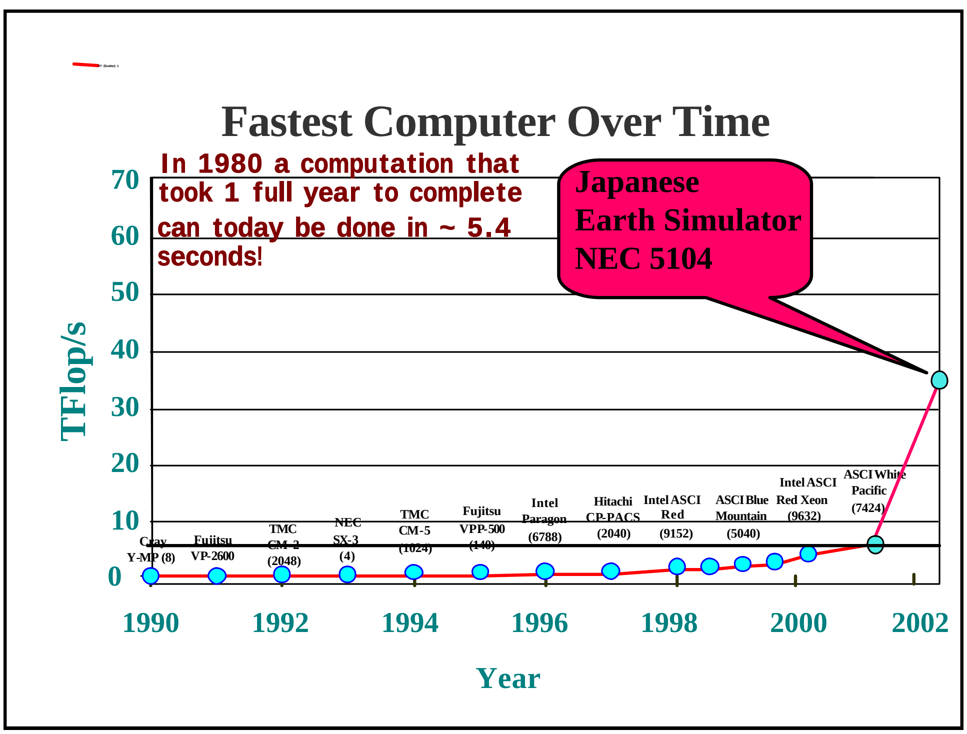


Fastest Computer Over Time





Fastest Computer Over Time
In 1980 a computation that took 1 full year to complete can today be done in ~ 27 seconds!
GFlop/s
7000
6000
5000
4000
3000
2000
1000
0
1990
1992
1994
1996
1998
2000
Year
Cray Y-MP (8)
Fujitsu VP-2600
TMC CM-2 (2048)
NEC SX-3 (4)
TMC CM-5 (1024)
Fujitsu VPP-500 (140)
Intel Paragon (6788)
Hitachi CP-PACS (2040)
Intel ASCI Red (9152)
ASCI Blue Pacific SST (5808)
SGI ASCI Blue Mountain (5040)
Intel ASCI Red Xeon (9632)
ASCI White Pacific (7424)



Fastest Computer Over Time
In 1980 a computation that took 1 full year to complete can today be done in ~ 5.4 seconds!
Japanese Earth Simulator NEC 5104
TFlop/s
70
60
50
40
30
20
10
0
1990
1992
1994
1996
1998
2000
2002
Year
Cray Y-MP (8)
Fujitsu VP-2600
TMC CM-2 (2048)
NEC SX-3 (4)
TMC CM-5 (1024)
Fujitsu VPP-500 (140)
Intel Paragon (6788)
Hitachi CP-PACS (2040)
Intel ASCI Red (9152)
ASCI Blue Mountain (5040)
Intel ASCI Red Xeon (9632)
ASCI White Pacific (7424)

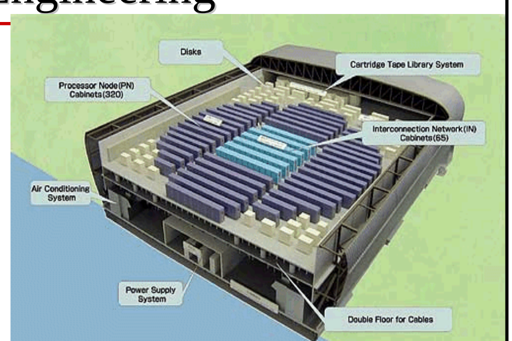
Machines at the Top of the List

Year	Computer	Measured Gflop/s	Factor ? from Pervious Year	Theoretical Peak Gflop/s	Factor ? from Pervious Year	Number of Processors	Efficiency
2002	Earth Simulator Computer, NEC	35860	5.0	40960	3.7	5120	88%
2001	ASCI White-Pacific, IBM SP Power 3	7226	1.5	11136	1.0	7424	65%
2000	ASCI White-Pacific, IBM SP Power 3	4938	2.1	11136	3.5	7424	44%
1999	ASCI Red Intel Pentium II Xeon core	2379	1.1	3207	0.8	9632	74%
1998	ASCI Blue-Pacific SST, IBM SP 604E	2144	1.6	3868	2.1	5808	55%
1997	Intel ASCI Option Red (200 MHz Pentium Pro)	1338	3.6	1830	3.0	9152	73%
1996	Hitachi CP-PACS	368.2	1.3	614	1.8	2048	60%
1995	Intel Paragon XP/S MP	281.1	1	338	1.0	6768	83%
1994	Intel Paragon XP/S MP	281.1	2.3	338	1.4	6768	83%
1993	Fujitsu NWT	124.5		236		140	53%

9

A Tour d'Force in Engineering

- Homogeneous, Centralized, Proprietary, Expensive!
- Target Application: CFD-Weather, Climate, Earthquakes
- 640 NEC SX/6 Nodes (mod)
  - 5120 CPUs which have vector ops
- 40TeraFlops (peak)
- $250-$500 million for things in building
- Footprint of 4 tennis courts
- 7 MWatts
  - Say 10 cent/KWhr - $16.8K/day = $6M/year!
- Expect to be on top of Top500 until 60-100 TFlop ASCI machine arrives
- For the Top500 (June 2002)
  - Equivalent ~ 1/6 S Top 500
  - Performance of ESC > S Next Top 12 Computers
  - S of all the DOE computers = 27.5 TFlop/s
  - Performance of ESC > All the DOE + DOD machines (37.2 TFlop/s)

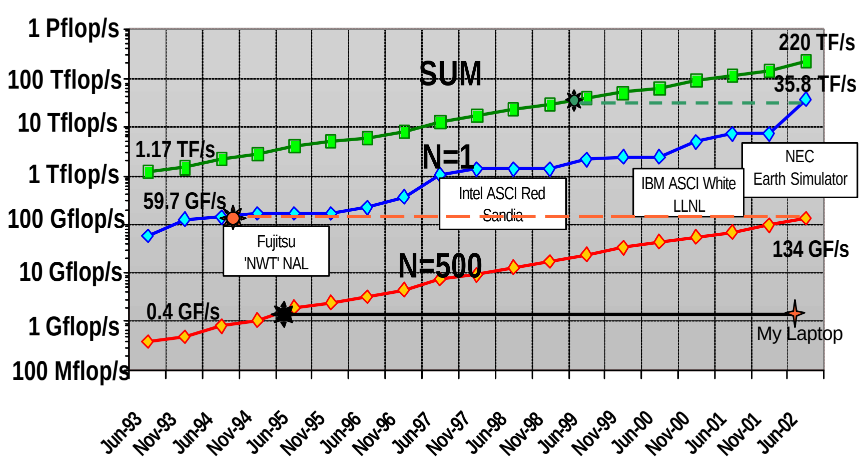


Top10 of the Top500

Rank	Manufacturer	Computer	R_{max} [TF/s]	Installation Site	Country	Year	Area of Installation	# Proc
1	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	Research	5120
2	IBM	ASCI White SP Power3	7.23	Lawrence Livermore National Laboratory	USA	2000	Research	8192
3	HP	AlphaServer SC ES45 1 GHz	4.46	Pittsburgh Supercomputing Center	USA	2001	Academic	3016
4	HP	AlphaServer SC ES45 1 GHz	3.98	Commissariat a l'Energie Atomique (CEA)	France	2001	Research	2560
5	IBM	SP Power3 375 MHz	3.05	NERSC/LBNL	USA	2001	Research	3328
6	HP	AlphaServer SC ES45 1 GHz	2.92	Los Alamos National Laboratory	USA	2002	Research	2048
7	Intel	ASCI Red	2.38	Sandia National Laboratory	USA	1999	Research	9632
8	IBM	pSeries 690 1.3 GHz	2.31	Oak Ridge National Laboratory	USA	2002	Research	864
9	IBM	ASCI Blue Pacific SST, IBM SP 604e	2.14	Lawrence Livermore National Laboratory	USA	1999	Research	5808
10	IBM	pSeries 690 1.3 Ghz	2.00	IBM/US Army Reseach Lab (ARL)	USA	2002	Vendor	768

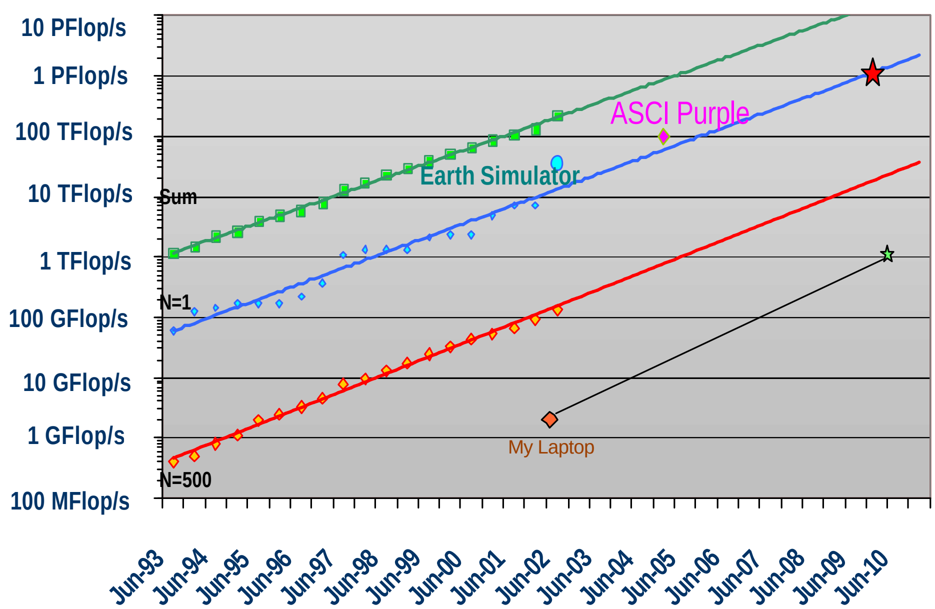
11

TOP500 - Performance

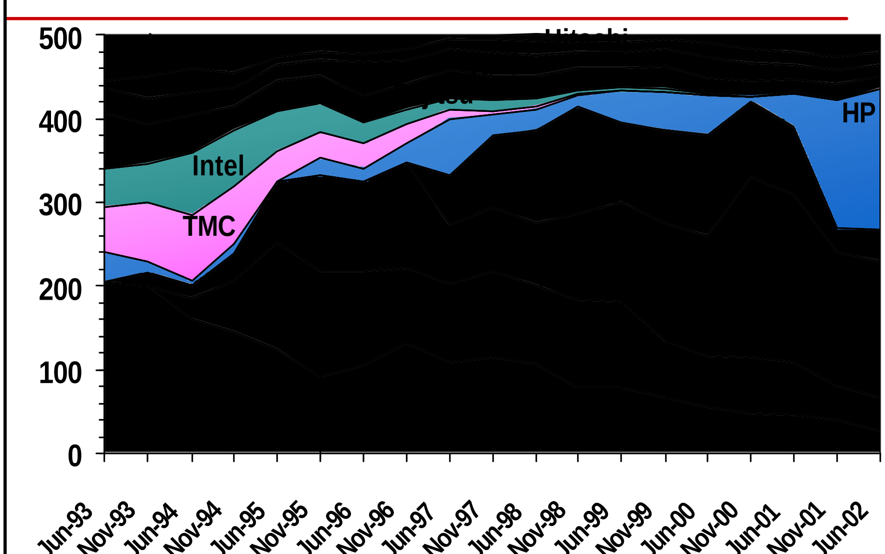


12

Performance Extrapolation



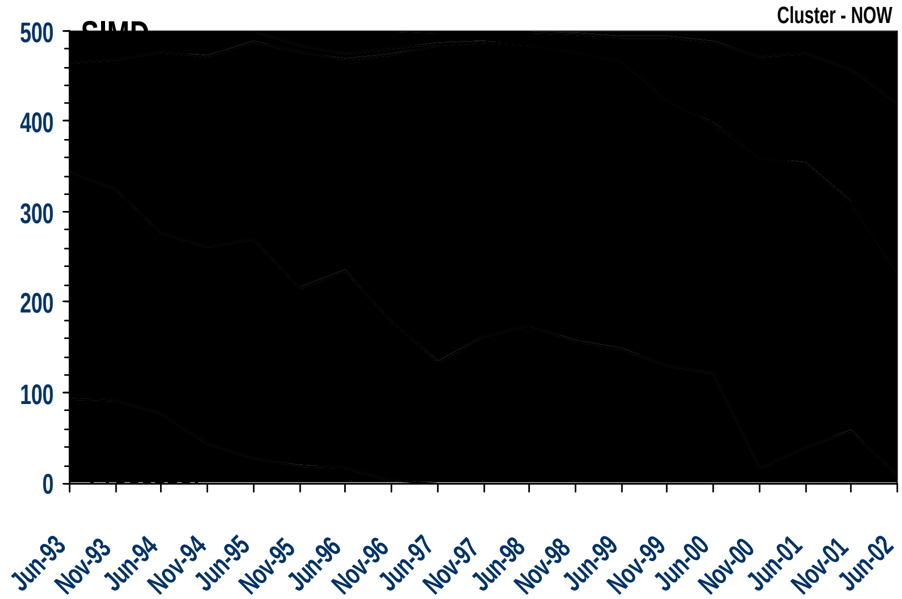
Manufacturers



14

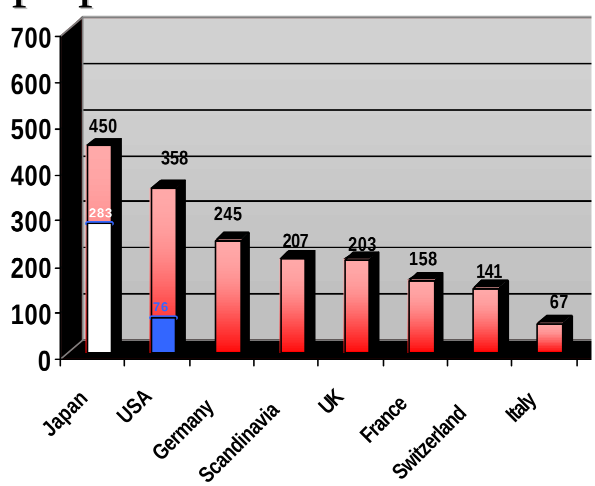
HP 168 (12 < 100), IBM 164 (47 < 100)

Architectures



Constellation: # of p/n ≥ n

Kflops per Inhabitant



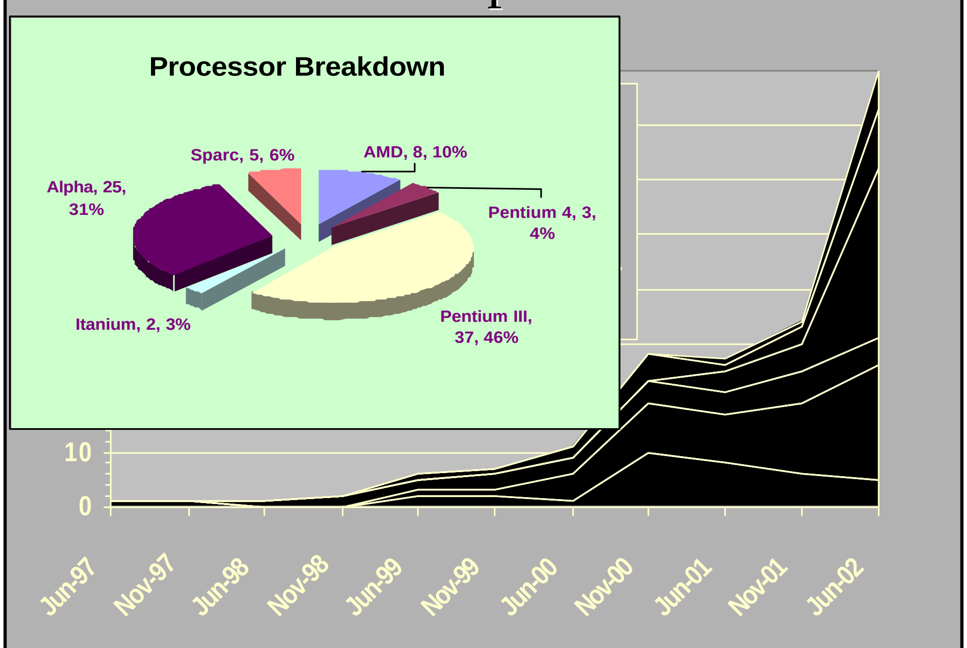
White is ES contribution and Blue is ASCI contribution

80 Clusters on the Top500

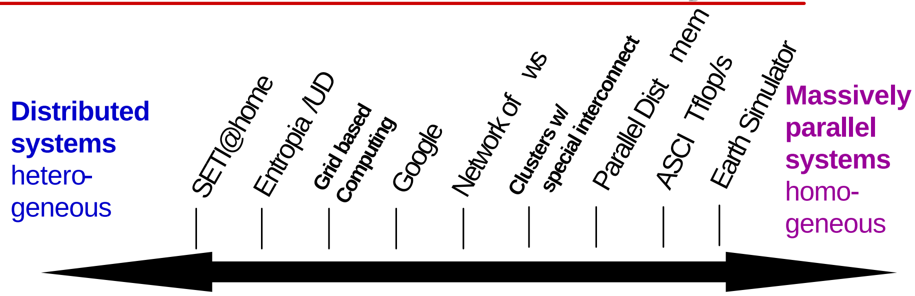
- A total of 42 Intel based and 8 AMD based PC clusters are in the TOP500.
  - 31 of these Intel based cluster are IBM Netfinity systems delivered by IBM.
- A substantial part of these are installed at industrial customers especially in the oil-industry.
  - Including 5 Sun and 5 Alpha based clusters and 21 HP AlphaServer.
- 14 of these clusters are labeled as 'Self-Made'.

17

Cluster on the Top500



Distributed and Parallel Systems



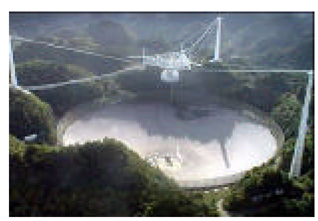
- Gather (unused) resources
- Steal cycles
- System SW manages resources
- System SW adds value
- 10% - 20% overhead is OK
- Resources drive applications
- Time to completion is not critical
- Time-shared
- SETI@home
  - ~ 400,000 machines
  - Averaging 40 Tflop/s

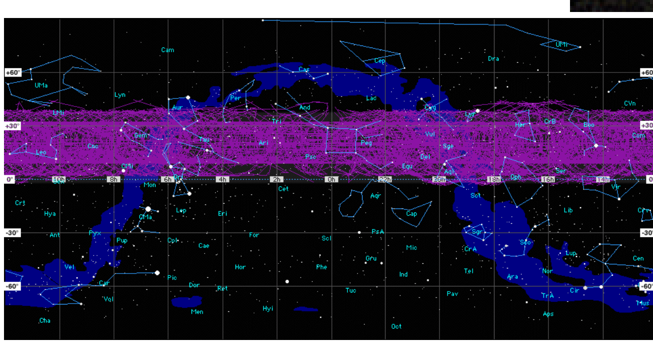
- Bounded set of resources
- Apps grow to consume all cycles
- Application manages resources
- System SW gets in the way
- 5% overhead is maximum
- Apps drive purchase of equipment
- Real-time constraints
- Space-shared
- Earth Simulator
  - 5000 processors
  - Averaging 35 Tflop/s

19

SETI@home: Global Distributed Computing

- Running on 500,000 PCs, ~1000 CPU Years per Day
  - 485,821 CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope

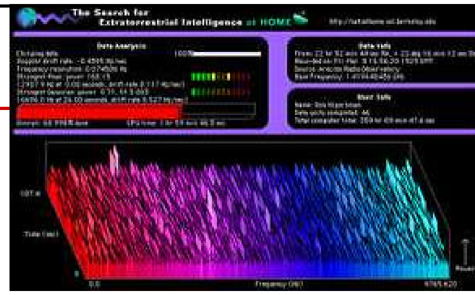




20

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 a 300 kilobyte 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.



- Largest distributed computation project in existence
  - 2500 machines today
  - Averaging 40 Tflop/s
- Today a number of companies trying this for profit.

Grid Computing - from ET toAnthrax

CNN.com / SCI-TECH

PCs tapped to help fight anthrax

January 22, 2002 Posted: 12:16 PM EST (1716 GMT)

SAN JOSE, California (AP) -- A coalition of scientists and technology companies is asking people around the world to use their computers' extra processing power to help search for a cure for anthrax.

The project follows similar efforts to use "distributed computing" to hunt for extraterrestrial life and a cure for cancer. It is being launched Tuesday to help Oxford University researchers find ways to treat anthrax that can no longer be treated by antibiotics.

The project is based on the premise that the average personal computer uses between 13 percent and 18 percent of its processing power at any given time. It employs "peer-to-peer" technology, in which millions of computers can share files over the Internet.

Participants download a screen-saver that runs whenever their computers have resources to spare, and uses that power to perform computations for the project. When the user connects to the Internet, the computer sends data back to a central hub and gets another assignment.

The company that designed the program, United Devices Inc. of Austin, Texas, promises that no personal information on participants' PCs can be compromised while they take part.

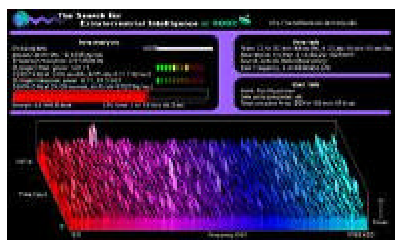
If the project attracts more than 100,000 participants, it can give researchers more computational power than the world's 10 best supercomputers combined, said United Devices spokesman Andy Prince.

With enough participants, the project would provide researchers 10 times more power than the world's best supercomputer, said Graham Richards, the Oxford professor leading the study.

"The screen-saver doesn't cost you anything, and at least you're taking part in something, adding your bit," he said.

Intel, Microsoft involved

Scientists have discovered that the anthrax toxin is made up of three proteins that



UNITED DEVICES™

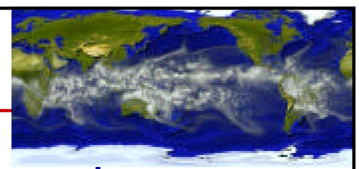
Google

- Google query attributes
  - 150M queries/day (2000/second)
  - 3B documents in the index
- Data centers
  - 15,000 Linux systems in 6 data centers
    - 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

Source: Monika Henzinger, Google

23

In the past: Isolation
Motivation for Grid Computing

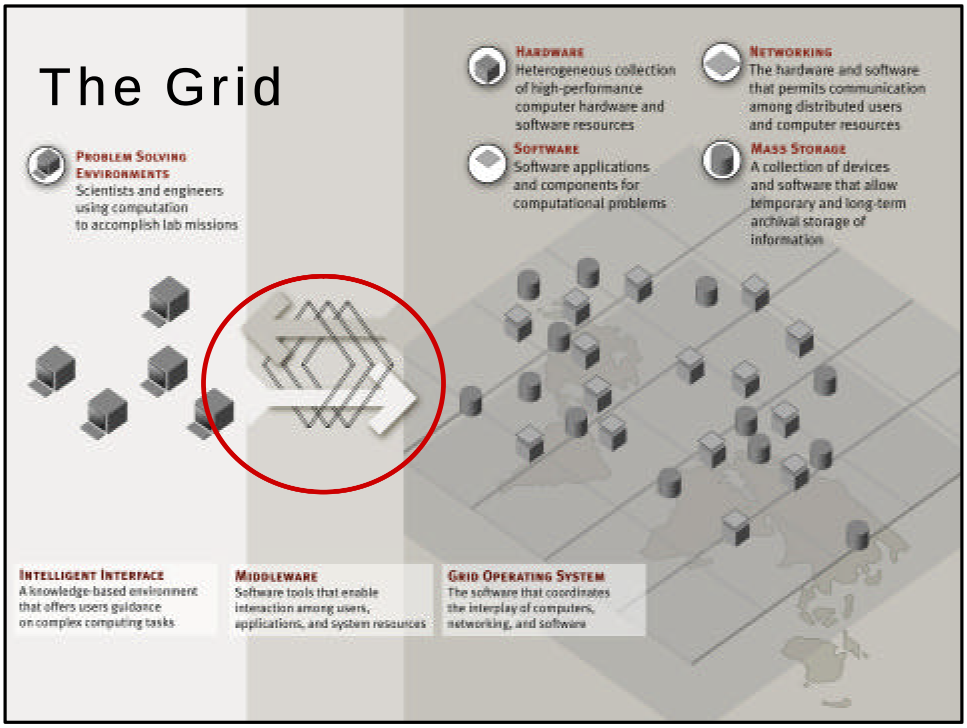


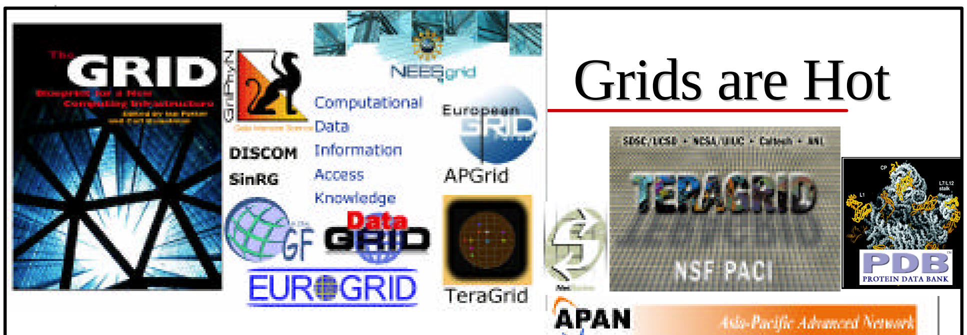
- Today there is a complex interplay and increasing interdependence among the sciences.
- Many science and engineering problems require widely dispersed resources be operated as systems.
- What we do as collaborative infrastructure developers will have profound influence on the future of science.
- Networking, distributed computing, and parallel computation research have matured to make it possible for distributed systems to support high-performance applications, but...
  - Resources are dispersed
  - Connectivity is variable
  - Dedicated access may not be possible

Today: Collaboration

24

The Grid



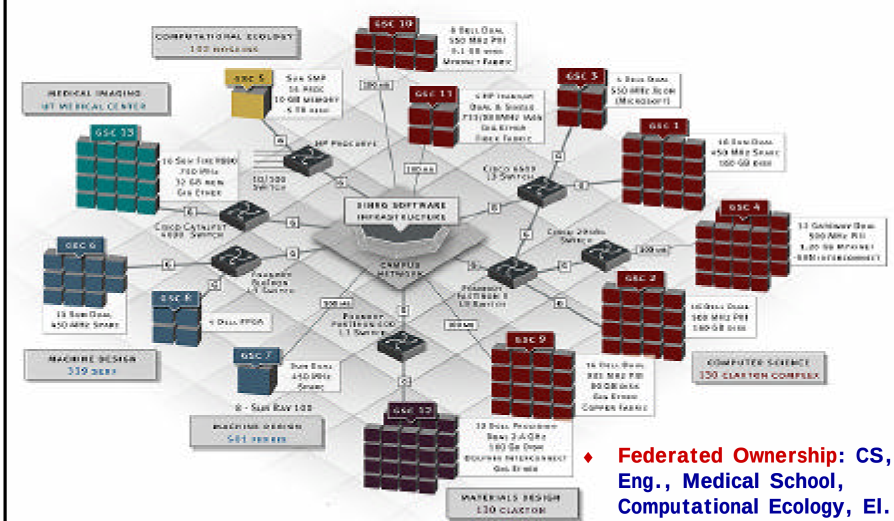


Grids are Hot

- **IPG NASA** http://nas.nasa.gov/~wej/home/IPG
- **Globus** http://www.globus.org/
- **Legion** http://www.cs.virgina.edu/~grimshaw/
- **AppLeS** http://www-cse.ucsd.edu/groups/hpcl/
- **NetSolve** http://www.cs.utk.edu/netsolve/
- **NINF** http://phase.etl.go.jp/ninf/
- **Condor** http://www.cs.wisc.edu/condor/
- **CUMULVS** http://www.epm.ornl.gov/cs/
- **WebFlow** http://www.npac.syr.edu/users/gcf/
- **NGC** http://www.nordicgrid.net

26

University of Tennessee Deployment: Scalable Intracampus Research Grid: SInRG



The Knoxville Campus has two DS-3 commodity Internet connections and one DS-3 Internet2/Abilene connection. An OC-3 ATM link routes IP traffic between the Knoxville campus, National Transportation Research Center, and Oak Ridge National Laboratory. UT participates in several national networking initiatives including Internet2 (I2), Abilene, the federal Next Generation Internet (NGI) initiative, Southern Universities Research Association (SURA) Regional Information Infrastructure (RII), and Southern Crossroads (SoX).

The UT campus consists of a meshed ATM OC-12 being migrated over to switched Gigabit by early 2002.

- **Federated Ownership: CS, Chem Eng., Medical School, Computational Ecology, El. Eng.**
- **Real applications, middleware development, logistical networking**

27

Grids vs. Capability Computing

- **Not an "either/or" question**
  - **each addresses different needs**
  - **both are part of an integrated solution**
- **Grid strengths**
  - **coupling necessarily distributed resources**
    - **instruments, archives, and people**
  - **eliminating time and space barriers**
    - **remote resource access and capacity computing**
  - **Grids are not a cheap substitute for capability HPC**
- **Capability computing strengths**
  - **supporting foundational computations**
    - **terascale and petascale "nation scale" problems**
  - **engaging tightly coupled teams and computations**

28

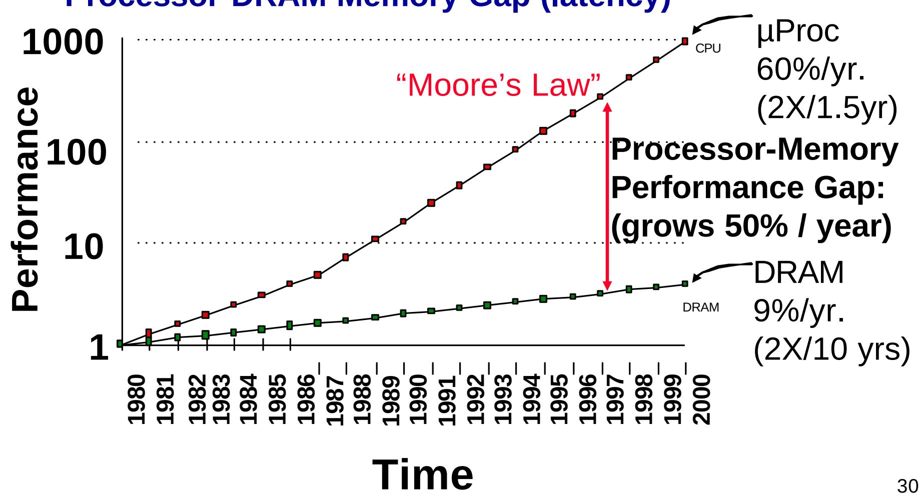
Software Technology & Performance

- **Tendency to focus on hardware**
- **Software required to bridge an ever widening gap**
- **Gaps between usable and deliverable performance is very steep**
  - **Performance only if the data and controls are setup just right**
    - **Otherwise, dramatic performance degradations, very unstable situation**
    - **Will become more unstable**
- **Challenge of Libraries, PSEs and Tools is formidable with Tflop/s level, even greater with Pflops, some might say insurmountable.**

29

Where Does the Performance Go? or Why Should I Care About the Memory Hierarchy?

Processor-DRAM Memory Gap (latency)



30

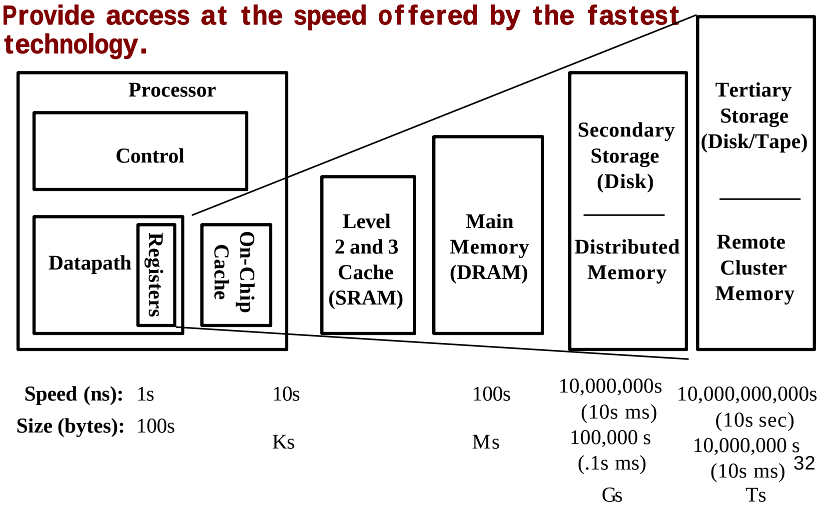
Optimizing Computation and Memory Use

- **Computational optimizations**
  - **Theoretical peak:(# fpus)*(flops/cycle) * Mhz**
    - Pentium 4: (1 fpu)*(2 flops/cycle)*(2.53 Ghz) = 5060 MFLOP/s
- **Operations like:**
  - $a = x^T y$: 2 operands (16 Bytes) needed for 2 flops; at 5060 Mflop/s will requires 5060 MW/s bandwidth
  - $y = a\,x + y$: 3 operands (24 Bytes) needed for 2 flops; at 5060 Mflop/s will requires 7590 MW/s bandwidth
- **Memory optimization**
  - **Theoretical peak: (bus width) * (bus speed)**
    - Pentium 4: (32 bits)*(533 Mhz) = 2132 MB/s = 266 MW/s

31

Memory Hierarchy

- **By taking advantage of the principle of locality:**
  - **Present the user with as much memory as is available in the cheapest technology.**
  - **Provide access at the speed offered by the fastest technology.**



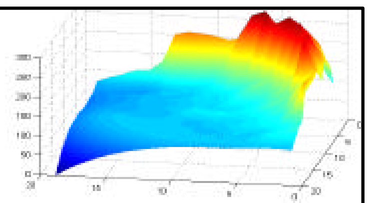
32

Self Adapting Software

- **Software system that …**
  - **Obtains information on the underlying system where they will run.**
  - **Adapts application to the presented data and the available resources perhaps provide automatic algorithm selection**
  - **During execution perform optimization and perhaps reconfigure based on newly available resources.**
  - **Allow the user to provide for faults and recover without additional users involvement**
- **The moral of the story**
  - **We know the concepts of how to improve things.**
  - **Capture insights/experience – do what humans do well**
  - **Automate the dull stuff**

33

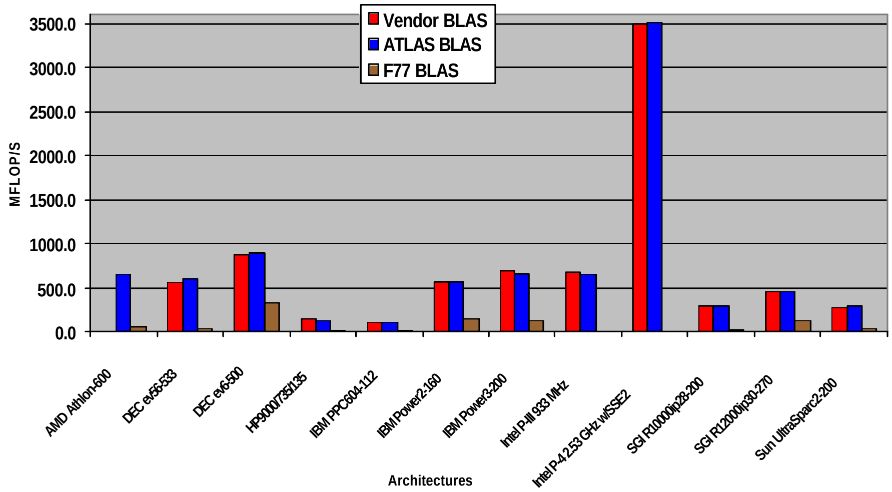
Software Generation Strategy - ATLAS BLAS



- **Parameter study of the hw**
- **Generate multiple versions of code, w/difference values of key performance parameters**
- **Run and measure the performance for various versions**
- **Pick best and generate library**
- **Level 1 cache multiply optimizes for:**
  - **TLB access**
  - **L1 cache reuse**
  - **FP unit usage**
  - **Memory fetch**
  - **Register reuse**
  - **Loop overhead minimization**
- **Takes ~ 20 minutes to run, generates Level 1,2, & 3 BLAS**
- **"New" model of high performance programming where critical code is machine generated using parameter optimization.**
- **Designed for modern architectures**
  - **Need reasonable C compiler**
- **Today ATLAS in used within various ASCI and SciDAC activities and by Matlab, Mathematica, Octave, Maple, Debian, Scyld Beowulf, SuSE,…**

34

ATLAS (DGEMM n = 500)



- ATLAS is faster than all other portable BLAS implementations and it is comparable with machine-specific libraries provided by the vendor.
- Looking at sparse operations

35

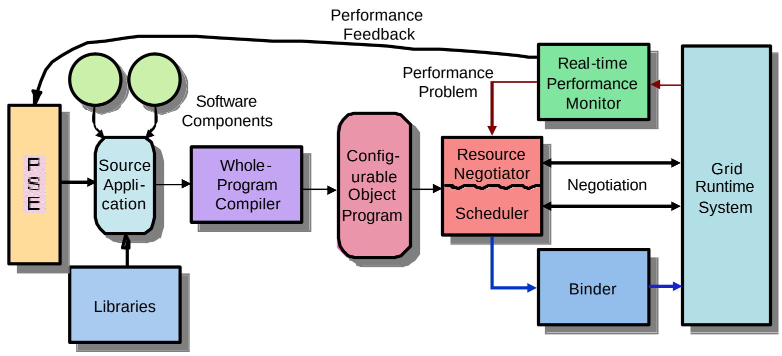
GrADS - Grid Application Development System

- Problem: Grid has distributed, heterogeneous, dynamic resources; how do we use them?
- Goal: reliable performance on dynamically changing resources
- Minimize work of preparing an application for Grid execution
  - Provide generic versions of key components (currently built in to applications or manually done)
    - E.g., scheduling, application launch, performance monitoring
- Provide high-level programming tools to help automate application preparation
  - Performance modeler, mapper, binder

36

NSF/NGS
GrADS - GrADSoft Architecture

- **Goal: reliable performance on dynamically changing resources**

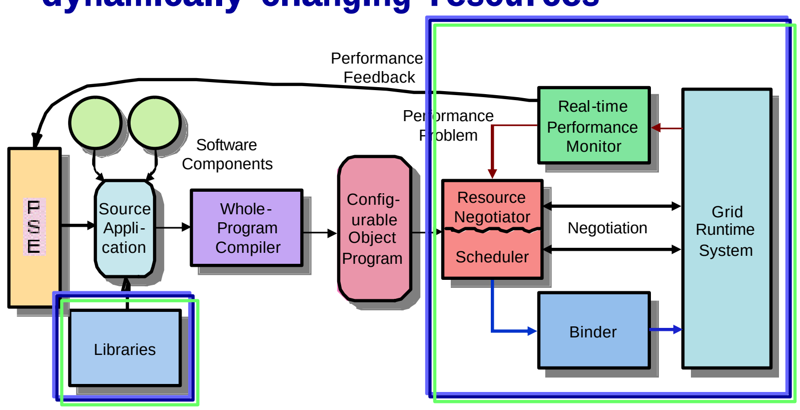


PIs: Ken Kennedy, Fran Berman, Andrew Chein, Keith Cooper, JD, Ian Foster, Lennart Johnsson, Dan Reed, Carl Kesselman, John Mellor-Crummey, Linda Torczon & Rich Wolski

37

NSF/NGS
GrADS - GrADSoft Architecture

- **Goal: reliable performance on dynamically changing resources**

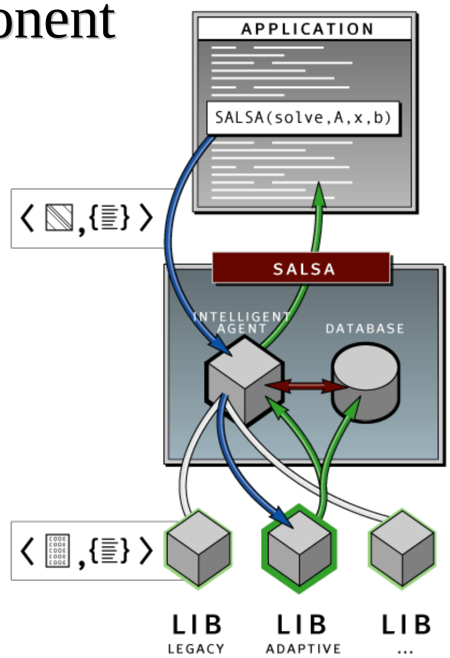


PIs: Ken Kennedy, Fran Berman, Andrew Chein, Keith Cooper, JD, Ian Foster, Lennart Johnsson, Dan Reed, Carl Kesselman, John Mellor-Crummey, Linda Torczon & Rich Wolski

38

Intelligent Component

- System to mediate between user application and multiple possible libraries
- Self-Adaptivity and Learning Behavior
  - Heuristics are tuned based on data
    - System gradually gets smarter (database)
  - The system can educate the user
- User Interaction
  - User can guide the system by providing further information
  - System teaches user about properties of the data



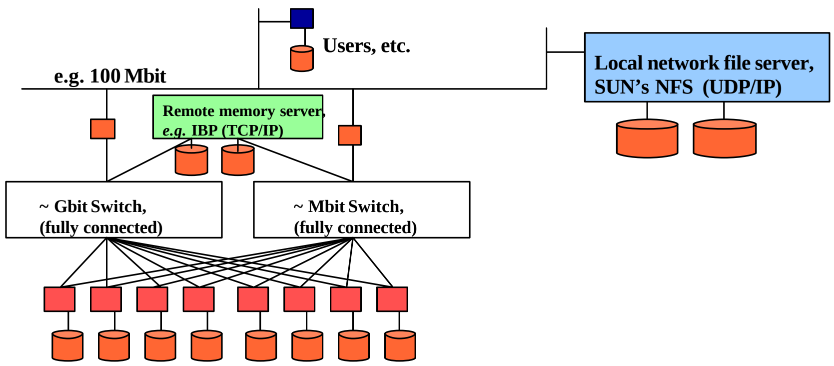
Research Areas

- Automatically generating performance models (e.g. for ScaLAPACK) on Grid resources
- Evaluating Performance "Contracts"
- Near Optimal Scheduling (execution) on the Grid
- Rescheduling for changing resources
- Checkpointing and fault tolerance
- High-latency tolerant algorithms (SANS ideas)
- Porting applications/libraries to GrADS framework
- Developing generic GrADSoft interfaces (API's)

40

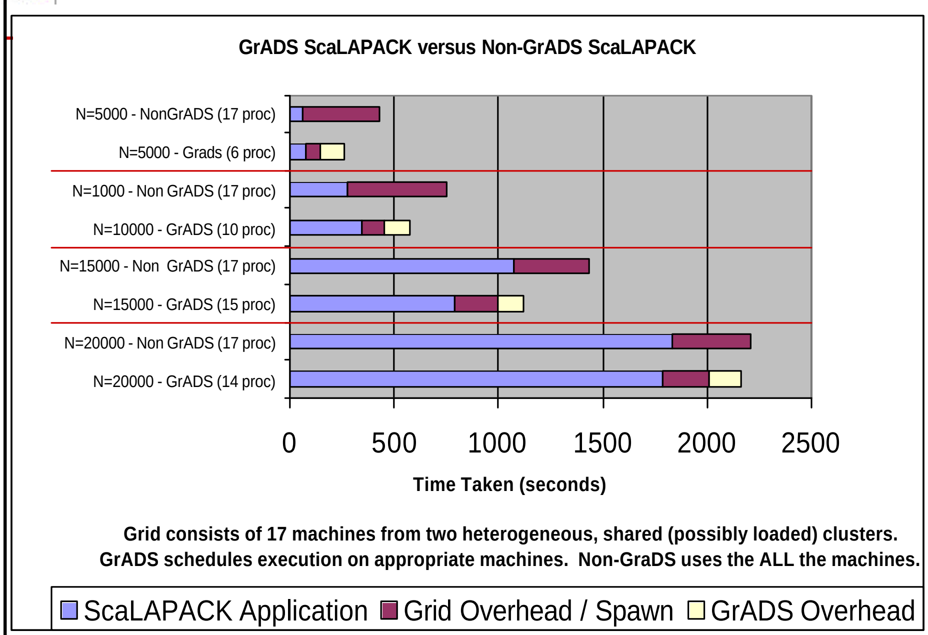
LAPACK For Clusters

- Developing middleware which couples cluster system information with the specifics of a user problem to launch cluster based applications on the "best" set of resource available.



- Using ScaLAPACK as the prototype software, but developing a framework

41



Grid consists of 17 machines from two heterogeneous, shared (possibly loaded) clusters. GrADS schedules execution on appropriate machines. Non-GraDS uses the ALL the machines.

42

Research Directions

- Parameterizable libraries
- Fault tolerant algorithms
- Annotated libraries
- Hierarchical algorithm libraries
- "Grid" (network) enabled strategies

A new division of labor between compiler writers, library writers, and algorithm developers and application developers will emerge.

43

Futures for Numerical Algorithms and Software

- Numerical software will be adaptive, exploratory, and intelligent
- Determinism in numerical computing will be gone.
  - After all, its not reasonable to ask for exactness in numerical computations.
  - Auditability of the computation, reproducibility at a cost
- Importance of floating point arithmetic will be undiminished.
  - 16, 32, 64, 128 bits and beyond.
- Reproducibility, fault tolerance, and auditability
- **Adaptivity** is a key so applications can effectively use the resources.

44

Collaborators / Support

- **TOP500**
  - **H. Mauer, Mannheim U**
  - **H. Simon, NERSC**
  - **E. Strohmaier, NERSC**
- **GrADS**
  - **Sathish Vadhiyar, UTK**
  - **Asim YarKhan, UTK**
  - Ken Kennedy, Fran Berman, Andrew Chein, Keith Cooper, Ian Foster, Carl Kesselman, Lennart Johnsson, Dan Reed, Linda Torczon, & Rich Wolski

- **Thanks**

NPACI

NSF
Next Generation Software (NGS)

GrADS
Grid Application Development Software Project

Scientific Discovery through
Advanced Computing (SciDAC)

TOPS
Terascale Optimal PDE Simulations

45