COMPUTER & COMPUTATIONAL SCIENCES

Los Alamos National Laboratory RADIANT
www.lanl.gov/radiant

Tour de HPCycles

Wu Feng
feng@lanl.gov

Los Alamos National Laboratory

Allan Snavely
allans@sdsc.edu

San Diego Supercomputing Center

NNSA

Los Alamos NATIONAL LABORATORY

Abstract

- In honor of Lance Armstrong's seven consecutive Tour de France cycling victories, we present Tour de HPCycles. While the Tour de France may be known only for the yellow jersey, it also awards a number of other jerseys for cycling excellence.

  The goal of this panel is to delineate the "winners" of the corresponding jerseys in HPC. Specifically, each panelist will be asked to award each jersey to a specific supercomputer or vendor, and then, to justify their choices.

The Jerseys

- Green Jersey (a.k.a Sprinters Jersey): Fastest consistently in miles/hour.
- Polka Dot Jersey (a.k.a Climbers Jersey): Ability to tackle difficult terrain while sustaining as much of peak performance as possible.
- White Jersey (a.k.a Young Rider Jersey): Best "under 25 year-old" rider with the lowest total cycling time.
- Red Number (Most Combative): Most aggressive and attacking rider.
- Team Jersey: Best overall team.
- Yellow Jersey (a.k.a Overall Jersey): Best overall supercomputer.

Panelists

- David Bailey, LBNL
  - Chief Technologist. IEEE Sidney Fernbach Award.
- John (Jay) Boisseau, TACC @ UT-Austin
  - Director. 2003 HPCwire Top People to Watch List.
- Bob Ciotti, NASA Ames
  - Lead for Terascale Systems Group. Columbia.
- Candace Culhane, NSA
  - Program Manager for HPC Research. HECURA Chair.
- Douglass Post, DoD HPCMO & CMU SEI
  - Chief Scientist. Fellow of APS.

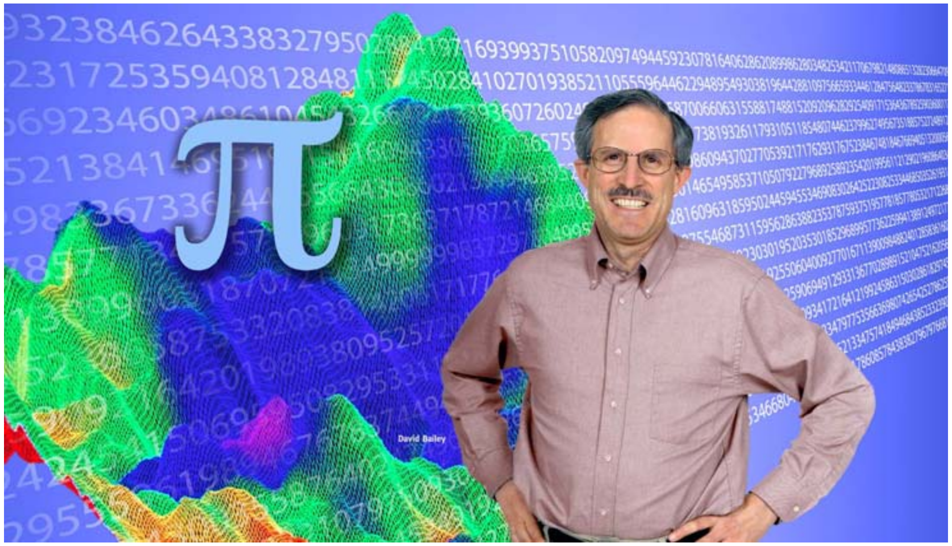
Ground Rules for Panelists

- Each panelist gets SEVEN minutes to present his position (or solution).
- Panel moderator will provide “one-minute-left” signal.
- During transitions between panelists, one question from the audience will be fielded.
- The panel concludes with 30-40 minutes of open discussion and questions amongst the panelists as well as from the audience.

Tour de HPCycles

David H Bailey

Lawrence Berkeley National Laboratory



What We’ve Seen at SC2006

- Remarkable performance:
  - 280.6 Tflop/s on Linpack.
- Remarkable application results:
  - At least six papers citing performance results over 10 Tflop/s.
  - Numerous outstanding papers and presentations.
- Remarkable system diversity:
  - Well-integrated “constellation” systems (e.g., IBM Power).
  - Several vector-based systems (e.g., Cray X1E, NEC).
  - Numerous commodity cluster offerings (e.g., Dell, HP, California Digital).
  - Impressive add-on components (e.g., Clearspeed).
  - FPGA-based systems (e.g., SRC, Starbridge).

Green Jersey (Sprinter's Jersey)

Fastest consistently in miles/hour:

- IBM BlueGene/L
  - 280.6 Tflop/s Linpack performance.
  - 101.7 Tflop/s on a molecular dynamics material science code.

No contest!

Polka Dot Jersey (Climber’s Jersey)

Ability to tackle difficult terrain while sustaining as much of peak performance as possible:

- The Japanese Earth Simulator (ES) system (by NEC): 67.6% of peak on 2048 processors, on a Lattice-Boltzmann MHD code.

Honorable mention:

- Cray’s X1E system: 41.1% of peak on 256 MSPs, on the Lattice-Boltzmann MHD code.
- IBM Power3: 39.8% on 1024 CPUs, on the PARATEC material science code.

These results are from Oliker et al (SC2005 paper 293).

White Jersey
(Young Rider Jersey)

Best under-25-year-old rider with the lowest total cycling time:

- IBM BlueGene/L: 101.7 Tflop/s on molecular dynamics material science code.

Red Number (Most Combative)

Most aggressive and attacking rider:

- Vendors of commodity clusters, including:
  - Dell – Sandia system #5 on Top500.
  - IBM – Barcelona system, #8 on Top500.
  - California Digital – LLNL system, #11 on Top500.
  - Hewlett-Packard – LANL system, #18 on Top500.
  - Apple Computer – Virginia Tech system, #20 on Top500.
  - Linux Networks – ARL system, #25 on Top500.
  - 360 commodity cluster systems in the latest Top500.

Warning to established HPC vendors: Beware the killer micros – fight them or join them.

Team Jersey

Best overall team:

- IBM
  - Strongest presence on Top500 list, with 219 systems and 52.8% of installed performance.
  - Variety of system designs: BlueGene/L, Power, clusters.

Honorable mention:

- HP
  - Second strongest presence on Top500 list, with 169 systems and 18.8% of installed performance.
- Cray
  - A rising star with impressive, well-balanced systems, designed specifically for real-world scientific computing.

Yellow Jersey

Best overall supercomputer:

- IBM BlueGene/L

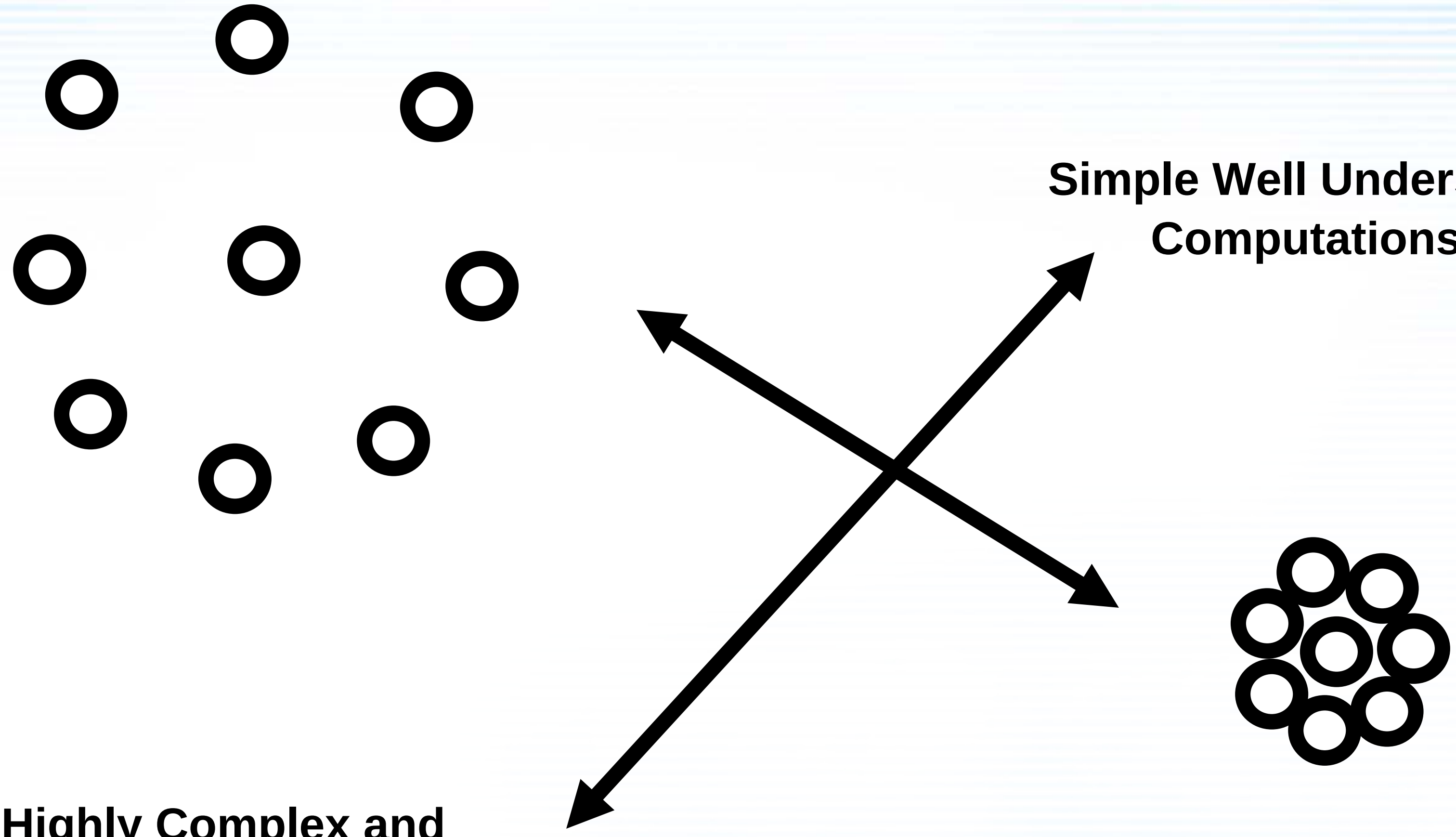
Tour de HPCycles

Bob Ciotti
Terascale Systems Lead
NASA Advanced Supercomputing Division (NAS)

Whats a Supercomputer gotta do?

NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING
AERONAUTICS RESEARCH MISSION DIRECTORATE
SCIENCE MISSION DIRECTORATE
SPACE OPERATIONS MISSION DIRECTORATE
EXPLORATION SYSTEMS MISSION DIRECTORATE
intel.
NASA
sgi
NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING

Computational Challenges



Embarrassingly Parallel

Simple Well Understood Computations

Highly Complex and Evolving Computations

Tightly Coupled

Classes of Computation

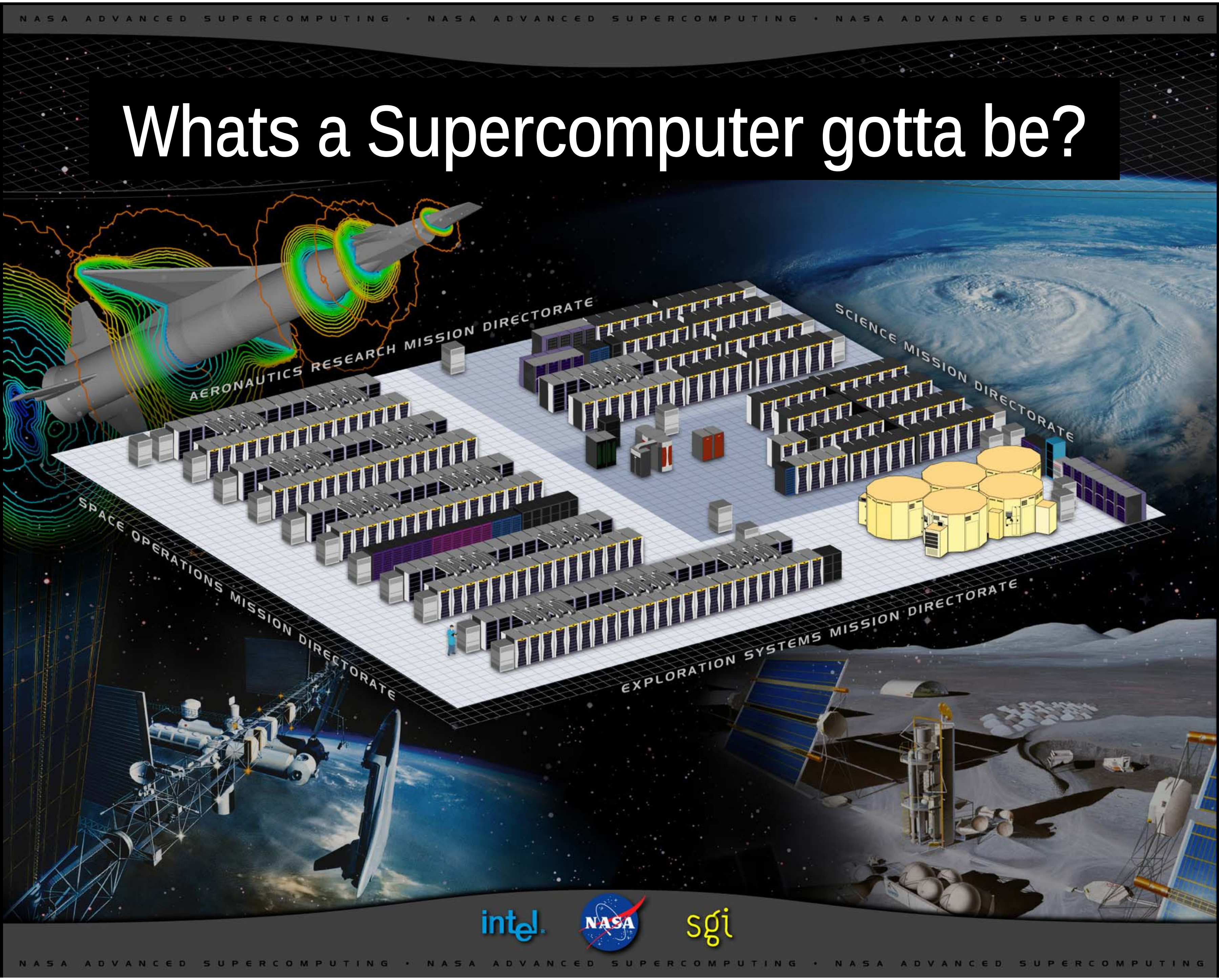
- Large Scale Breakthrough Investigations
  - Hurricane Forcast, Ocean Modeling, Shuttle Design
- Baseline Computational Workload – Daily Pedestrian Work
  - Existing Engineering/Science Workloads
- Emergency Response
  - Unexpected Highest Priority Work
    - Drop every thing else and solve this problem
  - Periodic requirement for mission critical analysis work
    - Shuttle Flight Support, STS fuel line, X37 heating

Productivity

HPC Development FACTORS

- Full Cost of Implementation
  - Design/Develop/Debug/Maintenance
- Time Sensitive Value
- Opportunity Cost
  - What aren't you doing because you are too busy developing parallel code?
- Flexibility in approach
  - OpenMP - MPI – pthreads – shmem – etc…
- Scalability/Performance
- Efficient access to data
  - High performance file systems
  - High sustained performance on entire problem
- Deployment
  - Quick and Straight Forward

Whats a Supercomputer gotta be?


NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING
AERONAUTICS RESEARCH MISSION DIRECTORATE
SCIENCE MISSION DIRECTORATE
SPACE OPERATIONS MISSION DIRECTORATE
EXPLORATION SYSTEMS MISSION DIRECTORATE
intel.
NASA
sgi
NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING • NASA ADVANCED SUPERCOMPUTING

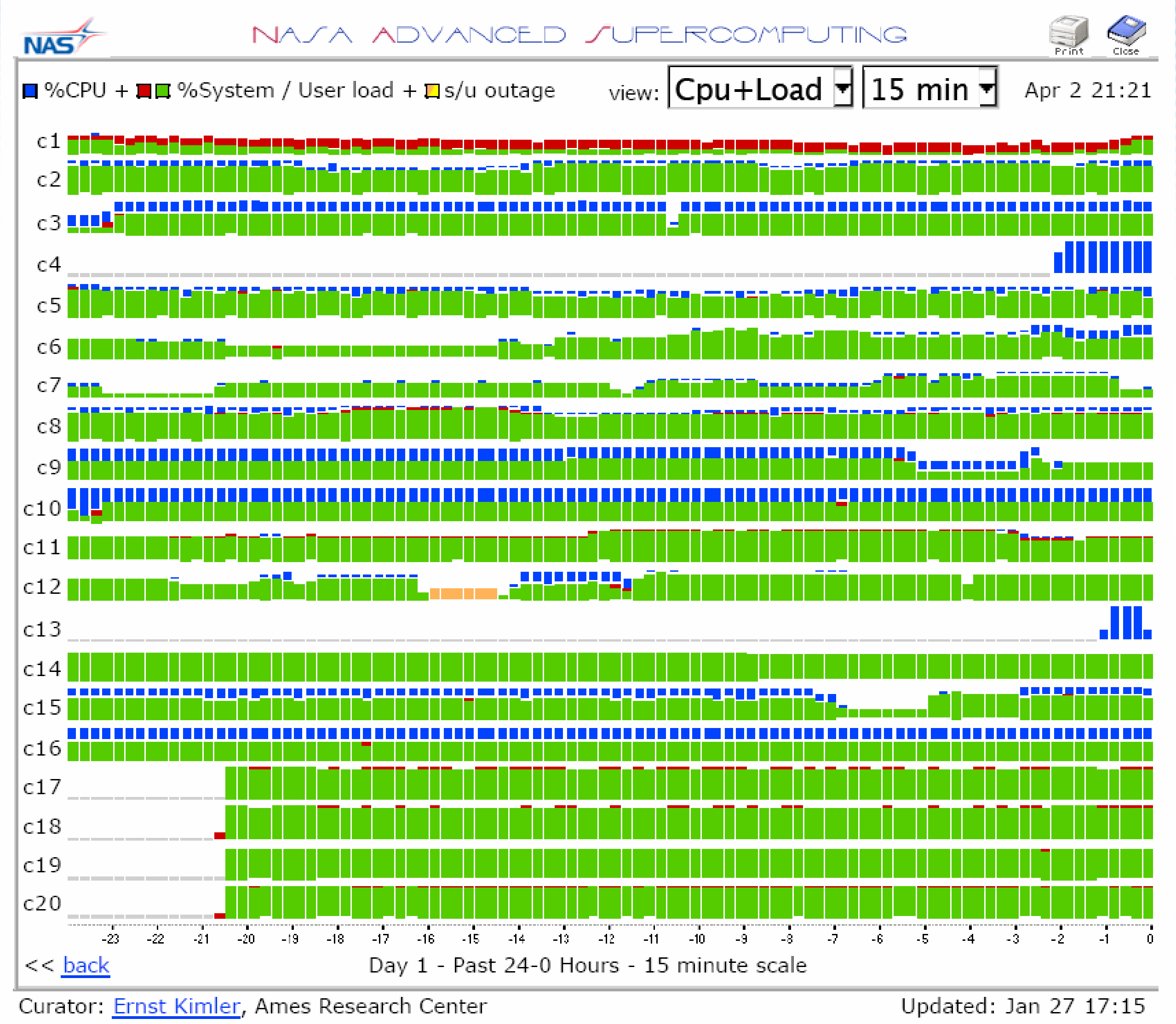
Operational Load

(solves all your problems)

20

N
o
d
e
s

2048



NASA ADVANCED SUPERCOMPUTING

🟦 %CPU + 🟥🟩 %System / User load + 🟨 s/u outage

view: Cpu+Load | 15 min | Apr 2 21:21

<< back

Day 1 - Past 24-0 Hours - 15 minute scale

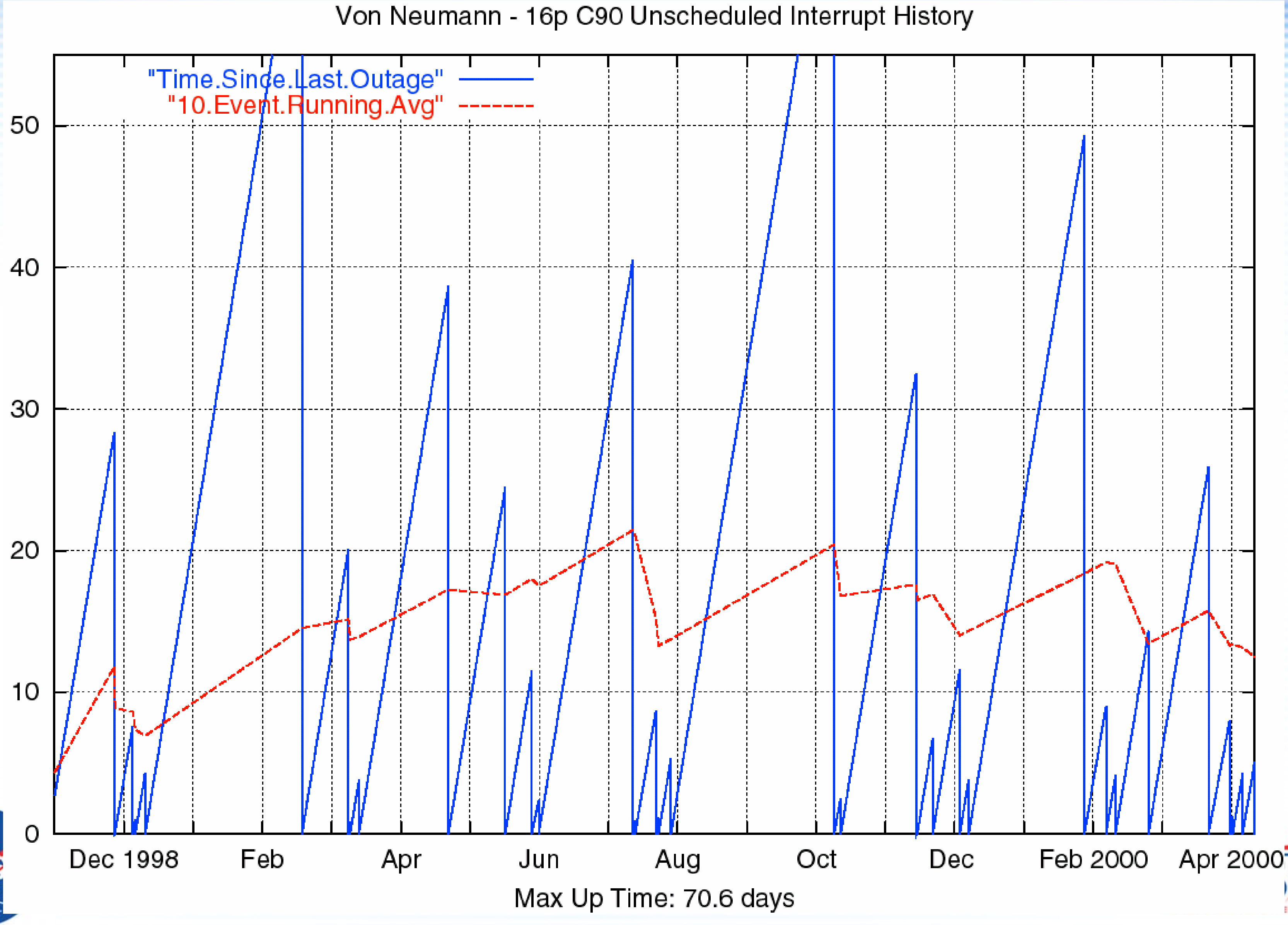
Curator: Ernst Kimler, Ames Research Center

Updated: Jan 27 17:15

System
Load
past
24 Hours

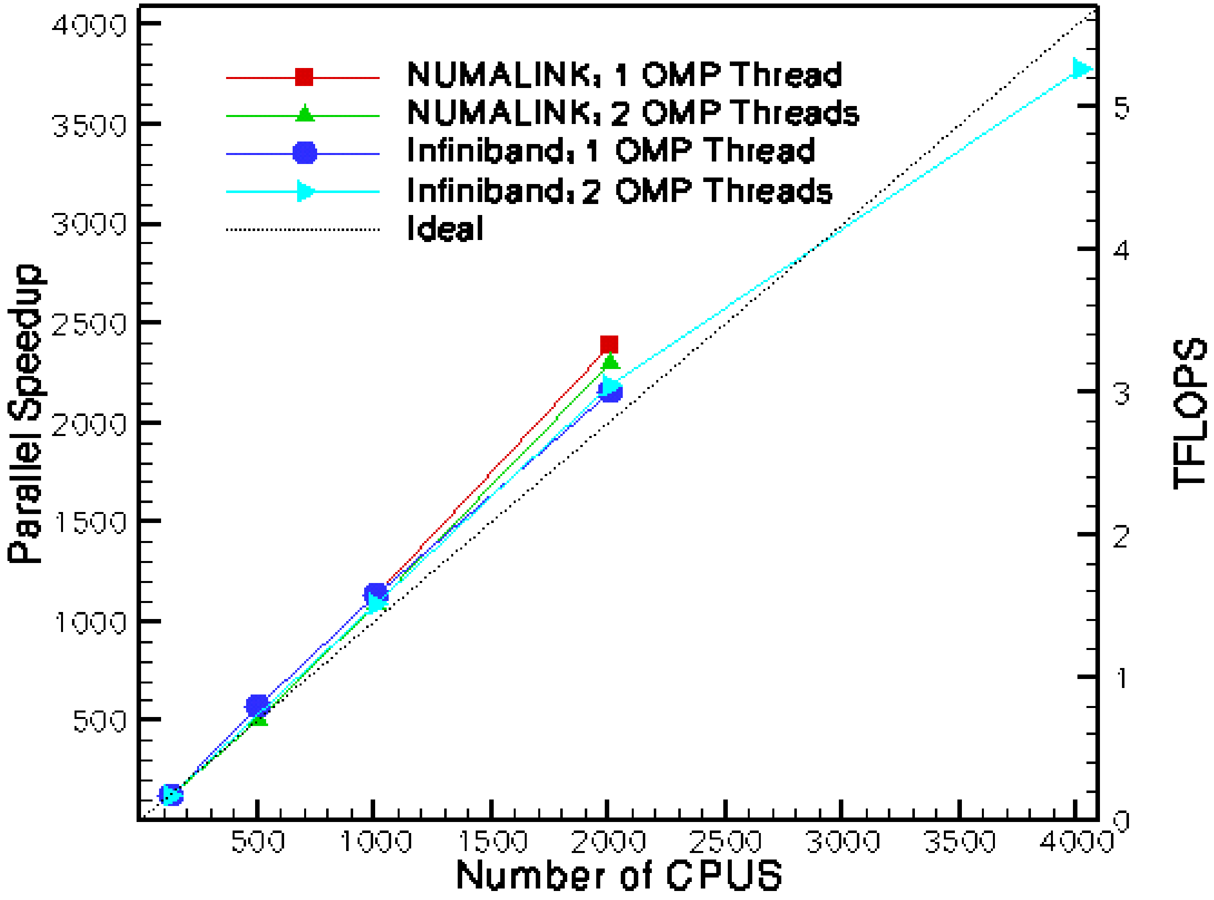
SC’05

Reliability - The Gold Standard: Cray C90



Performance

- 5.2 Tflops at 4016



Awards

Notable Retirements

- Single Level Programming
  - Multi-level implementations will draft behind Multi-core and fatter node system.
- Benchmarks that require single level programming

DNF – Did not Finalize

- MPI
  - Still not getting along with the Domain Scientists

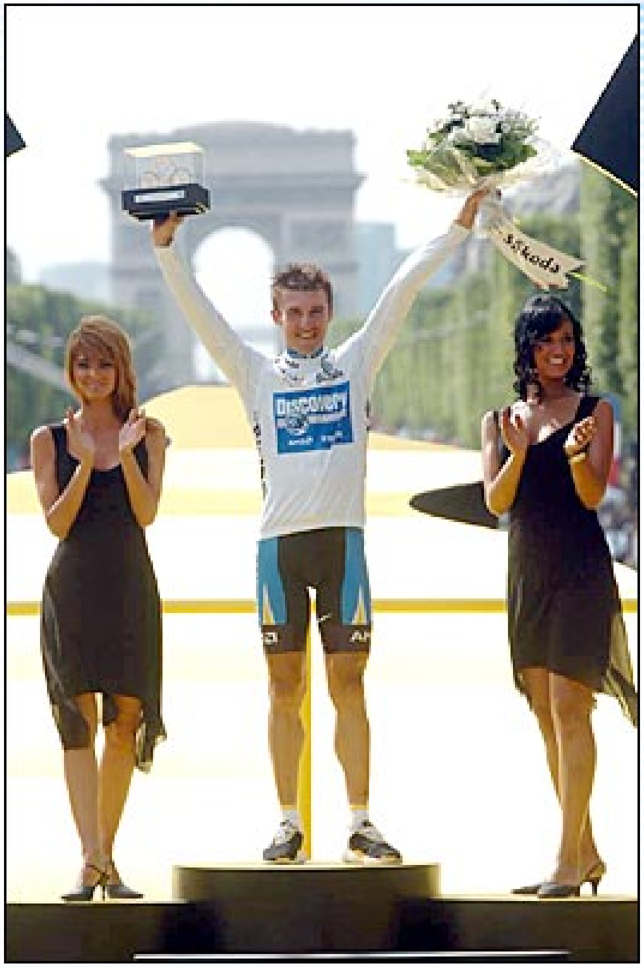
- BlueGeneL
  - Unable to establish a reliable track record

Red Jersey - Disruptive

- Luxtera

White Jersey

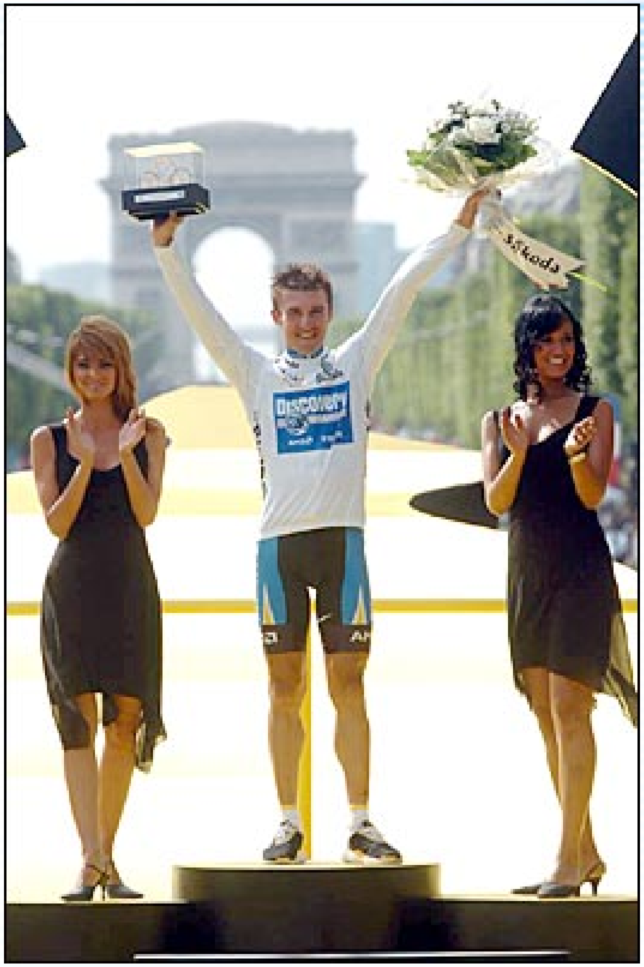
- Most Innovative
- Most likely to be a future repeat winner



White Jersey

- Most Innovative
- Most likely to be a future repeat winner

- Sun Microsystems
  - HERO System



The Contenders

DOE/NNSA/LLNL	**Bg/L**	**IBM**	**280**	**367**	**76%**
IBM TJ Watson	**BG/L**	**IBM**	**91**	**115**	**79%**
DOE/NNSA/LLNL	**ASC Purple**	**IBM**	**63**	**78**	**81%**
NASA/Ames	**Columbia**	**SGI**	**52**	**61**	**85%**
Sandia	**Thunderbird**	**Dell**	**38**	**65**	**58%**
Sandia	**Red Storm**	**Cray**	**36**	**44**	**82%**
Japan	Earth Simulator	**NEC**	**36**	**41**	**88%**

Polka Dots
Lance says climbing is “Hard Work”

Has to be:

- Widely accessible
- Reliable
- Ballanced
  - (I/O – Compute)
- Loaded up

Polka Dots
Lance says climbing is “Hard Work”

Has to be:

- Widely accessible
- Fairly Reliable
- Ballanced
  - (I/O – Compute)
- Loaded up

- Columbia

Yellow Jersey

- Still Fastest at the finish
- Unlimited team budget
- Didn't win every stage

Yellow Jersey

- Still Fastest at the finish
- Unlimited team budget
- Didn’t win every stage

- Earth Simulator

Tour de HPCycles

Tommy Minyard

November 18, 2005

Green Jersey (Sprinter)

- IBM BlueGene/L

Polka Dot Jersey (Climber)

- Cray X1E

White Jersey (Young Rider)

- Infiniband

Red Number (Aggressive)

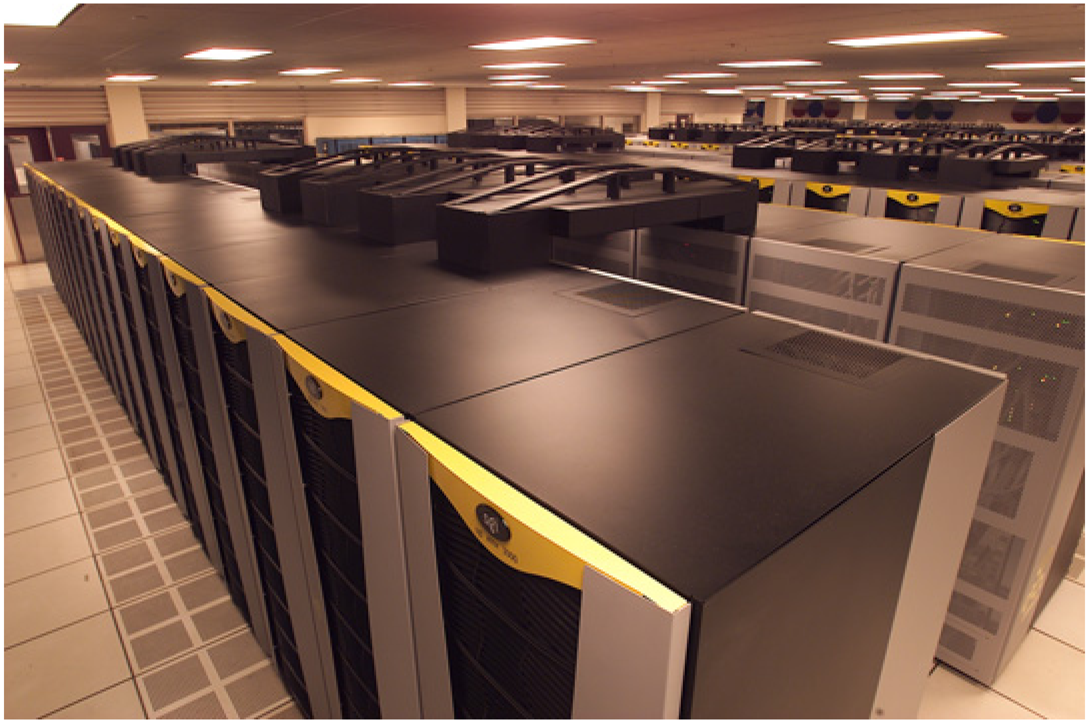
- Dell HPCC

Best Team

- IBM

Yellow Jersey (Best Overall)

- SGI Altix



Texas Advanced Computing Center

www.tacc.utexas.edu
(512) 475-9411

Department of Defense

High Performance Computing Modernization Program

Computer Performance: Computers and Codes

Douglass Post

Chief Scientist—HPCMP

Acknowledgements: Roy Campbell, Larry Davis, William Ward

Tour de HPCyles

18 November 2005

And the winners could be:

- Green (fastest sprinter): SGI Altix on Gamess, followed by IBM P4+ on Gamess, but depends on application
- Polka Dot (most capable): SGI Altix(2.41), Cray X1 (2.01), IBM P4+ (1.54), IBM Opteron (1.51): based on the weighted performance for the DoD benchmark suite
- White (best youngest): Linux Networx
- Red (most aggressive): no data
- Team Jersey (best team): HPCMP suite of computers
- Yellow Jersey (best overall computer): depends on application but the HPCMP suite comes closest

DoD High Performance Computing Modernization Program goal is to provide the best mix of computers for our mix of customers.

- HPCMP measures performance on prospective platforms using application benchmarks that represent our workload as part of the basis of our procurement decisions.
- 8 benchmark codes in 2005[1]
- 4920 users from approximately 178 DoD labs, contractors and universities
- 12 platforms from 5 vendors (Cray, IBM, HP/Compaq, Linux Networks, and SGI) at our four computer centers.
- Performance for a single code varies among platforms
  - Maximum performance/minimum performance ranges from 3.26 to 180.
- Performance for a single platform varies among codes
  - Maximum performance/minimum performance ranges from 1.42 to 47.
- No single benchmark measures useful performance over the range of applications

[1] **R. Campbell and W. Ward, HPCMP Guide to the Best Program Architectures Based on Application Results for TI-05, Proceedings of the 2005 DoD HPCMP Users' Group Conference, June 2005, Nashville, TN, IEEE Computer Society, Los Alamitos, CA.**

TI-05 Application Benchmark Codes

- Aero – Aeroelasticity CFD code
  (Fortran, serial vector, 15,000 lines of code)
- AVUS (Cobalt-60) – Turbulent flow CFD code
  (Fortran, MPI, 19,000 lines of code)
- GAMESS – Quantum chemistry code
  (Fortran, MPI, 330,000 lines of code)
- HYCOM – Ocean circulation modeling code
  (Fortran, MPI, 31,000 lines of code)
- OOCore – Out-of-core solver
  (Fortran, MPI, 39,000 lines of code)
- CTH – Shock physics code
  (~43% Fortran/~57% C, MPI, 436,000 lines of code)
- WRF – Multi-Agency mesoscale atmospheric modeling code
  (Fortran and C, MPI, 100,000 lines of code)
- Overflow-2 – CFD code originally developed by NASA
  (Fortran 90, MPI, 83,000 lines of code)

4 Major Computer Centers: HPCMP Systems (MSRCs)

HPC Center	System	Processors
Army Research Laboratory (ARL)	IBM P3	1,024 PEs
	SGI Origin 3800	256 PEs
		512 PEs
	IBM P4	768 PEs
		128 PEs
	Linux Networx Cluster	256 PEs
	LNX1 Xeon Cluster	2,100 PEs
	IBM Opteron Cluster	2,372 PEs
	SGI Altix Cluster	256 PEs
Aeronautical Systems Center (ASC)	Compaq SC-45	836 PEs
	IBM P3	528 PEs
	COMPAQ SC-40	64 PEs
	SGI Origin 3900	2,048 PEs
	SGI Origin 3900	128 PEs
	IBM P4	32 PEs
	SGI Altix Cluster	2,048 PEs
	HP Opteron	2,048 PEs
Engineer Research and Development Center (ERDC)	Compaq SC-40	512 PEs
	Compaq SC-45	512 PEs
	SGI Origin 3000	512 PEs
	Cray T3E	1,888 PEs
	SGI Origin 3900	1,024 PEs
	Cray X1	64 PEs
	Cray XT3	4,176 PEs
Naval Oceanographic Office (NAVO)	IBM P3	1,024 PEs
	IBM P4	1,408 PEs
	SV1	64 PEs
	IBM P4	3,456 PEs

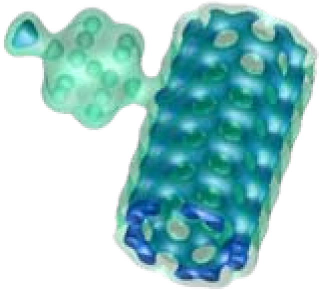
FY 01 and earlier
FY 02
FY 03
FY 04
FY 05
Retired in FY 05

As of: April 05

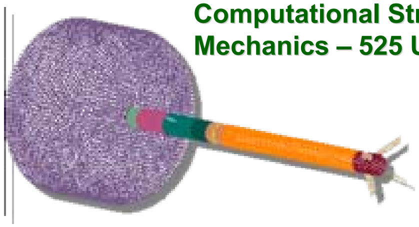
Current User Base and Requirements

- 613 projects and 4,920 users at approximately 178 sites
- Requirements categorized in 10 Computational Technology Areas (CTA)
- FY 2006 non-real-time requirements of 282 Habu-equivalents

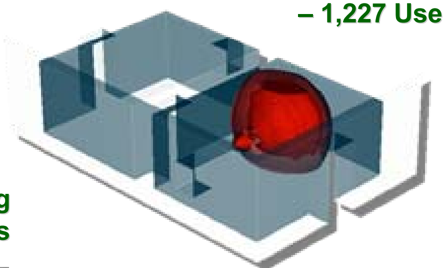
Electronics, Networking, and Systems/C4I – 34 Users



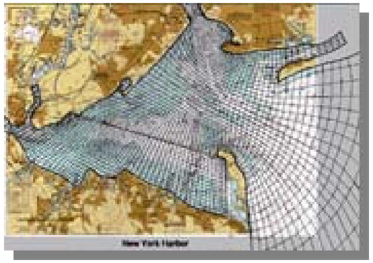
Computational Structural Mechanics – 525 Users



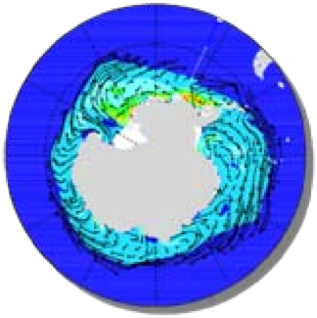
Computational Fluid Dynamics – 1,227 Users



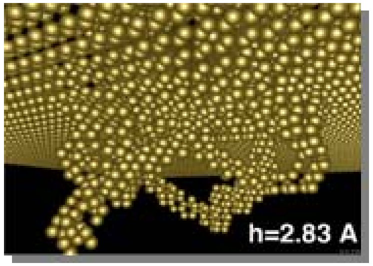
Environmental Quality Modeling & Simulation – 183 Users



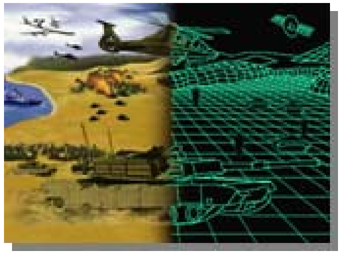
Climate/Weather/Ocean Modeling & Simulation – 233 Users



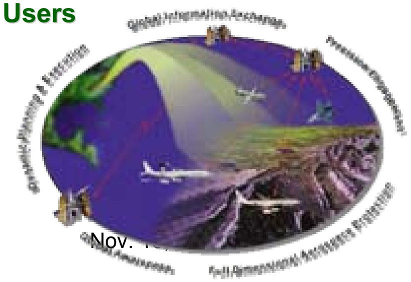
Computational Chemistry, Biology & Materials Science – 332 Users



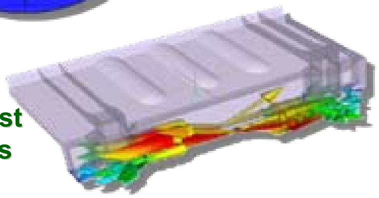
Forces Modeling & Simulation – 916 Users



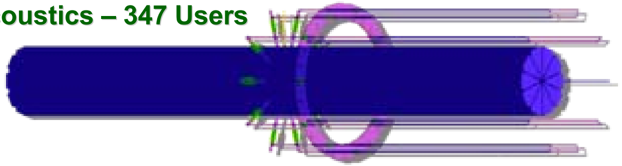
Signal/Image Processing – 439 Users


Global Information Exchange
Precision Engagement
Full Dimensional Aerospace Protection
Global Awareness
Dynamic Planning & Execution

Integrated Modeling & Test Environments – 617 Users



Computational Electromagnetics & Acoustics – 347 Users



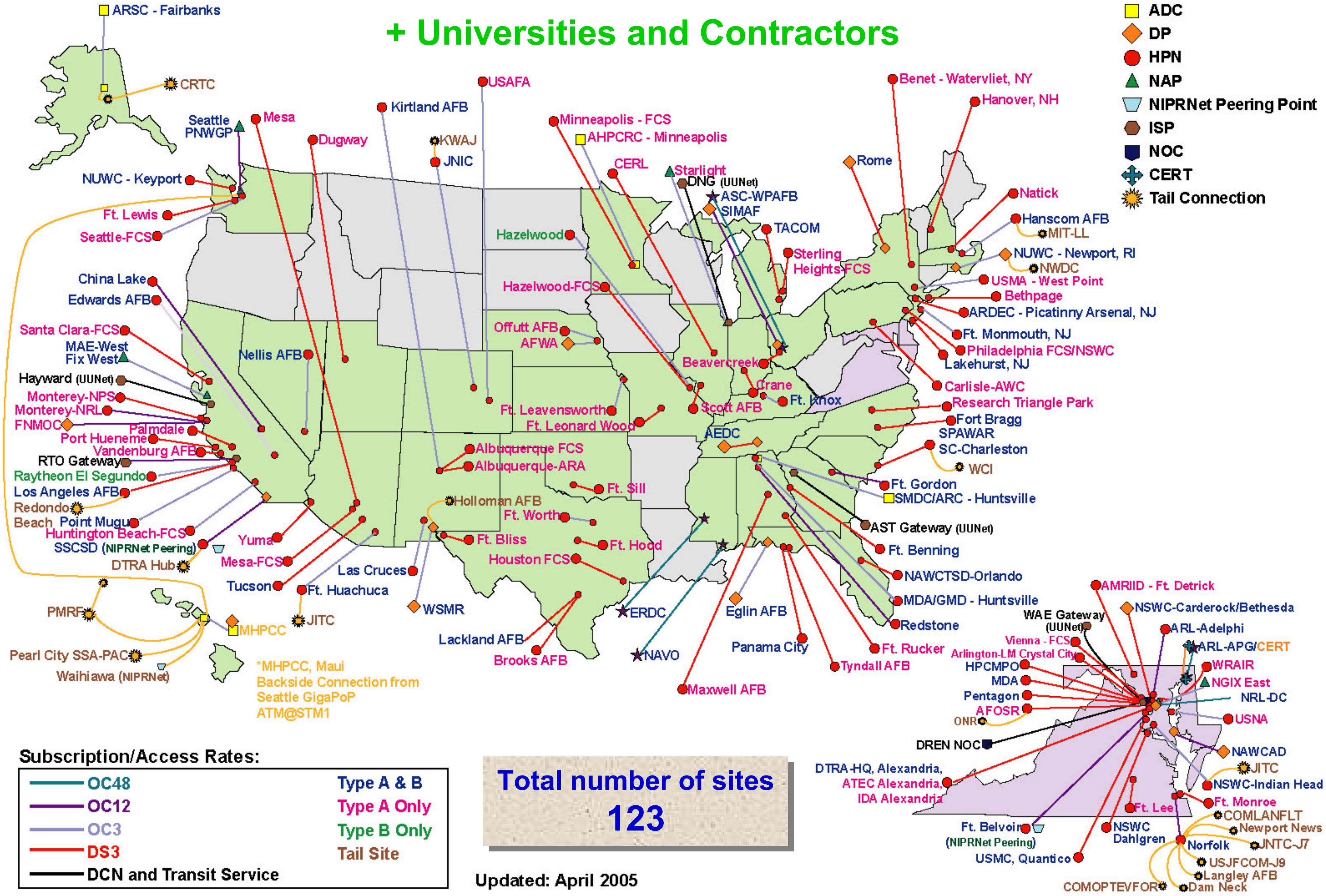
67 users are self characterized as "other"

Defense Research & Engineering Network (DREN) Current Sites

+ Universities and Contractors

Legend:

- ★ MSRC
- □ ADC
- ◆ DP
- ● HPN
- ▲ NAP
- ▽ NIPRNet Peering Point
- ⬢ ISP
- ■ NOC
- ✛ CERT
- ✹ Tail Connection



Subscription/Access Rates:

▬	OC48	Type A & B
▬	OC12	Type A Only
▬	OC3	Type B Only
▬	DS3	Tail Site
▬	DCN and Transit Service	

Total number of sites 123

Updated: April 2005

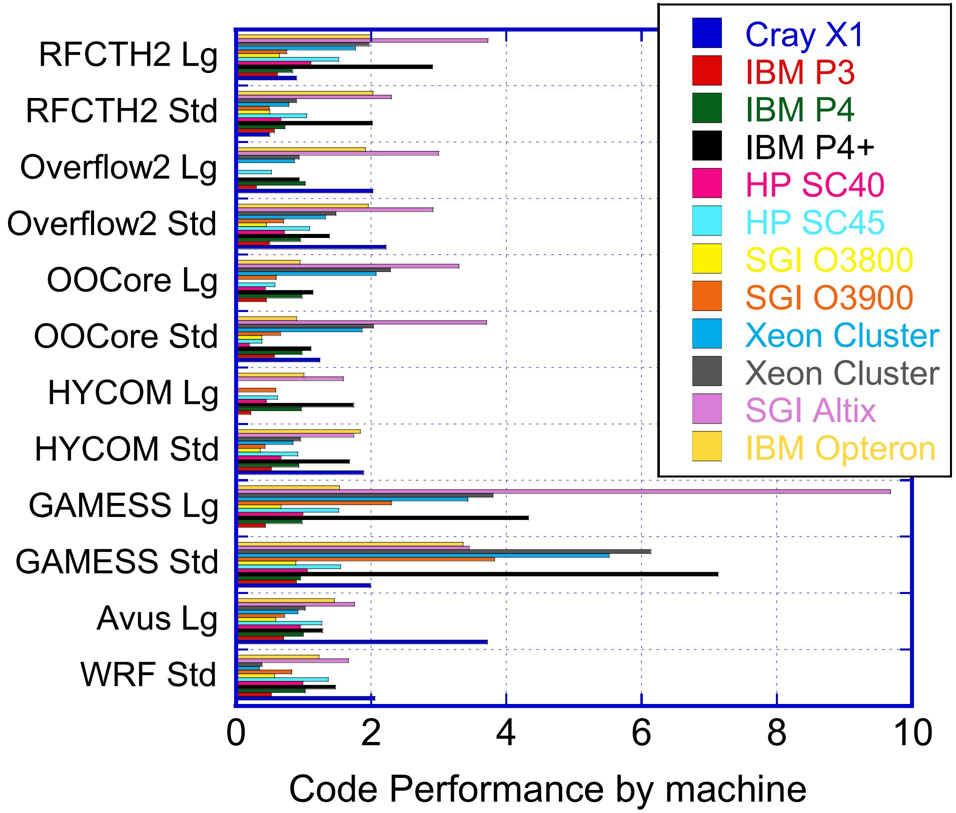
Performance depends on the computer and on the code.

- Normalized Performance = 1 on the NAVO IBM SP3 (HABU) platform with 1024 processors (375 MHz Power3 CPUs) assuming that each system has 1024 processors.

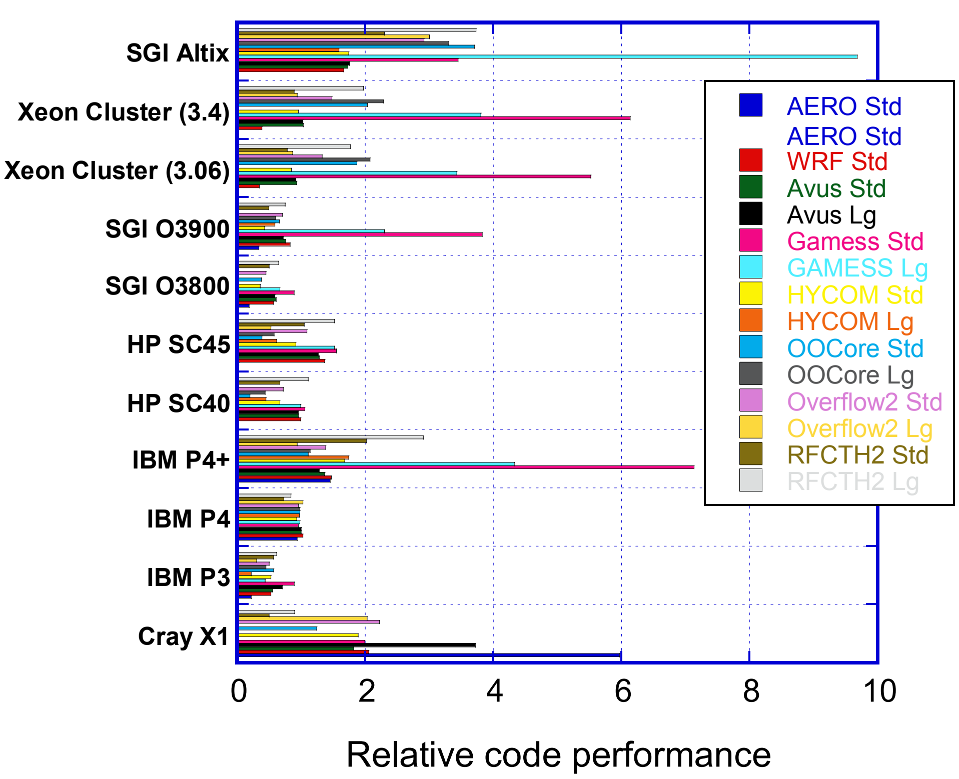
- GAMESS had the most variation among platforms.

Substantial variation of codes for a single computer.

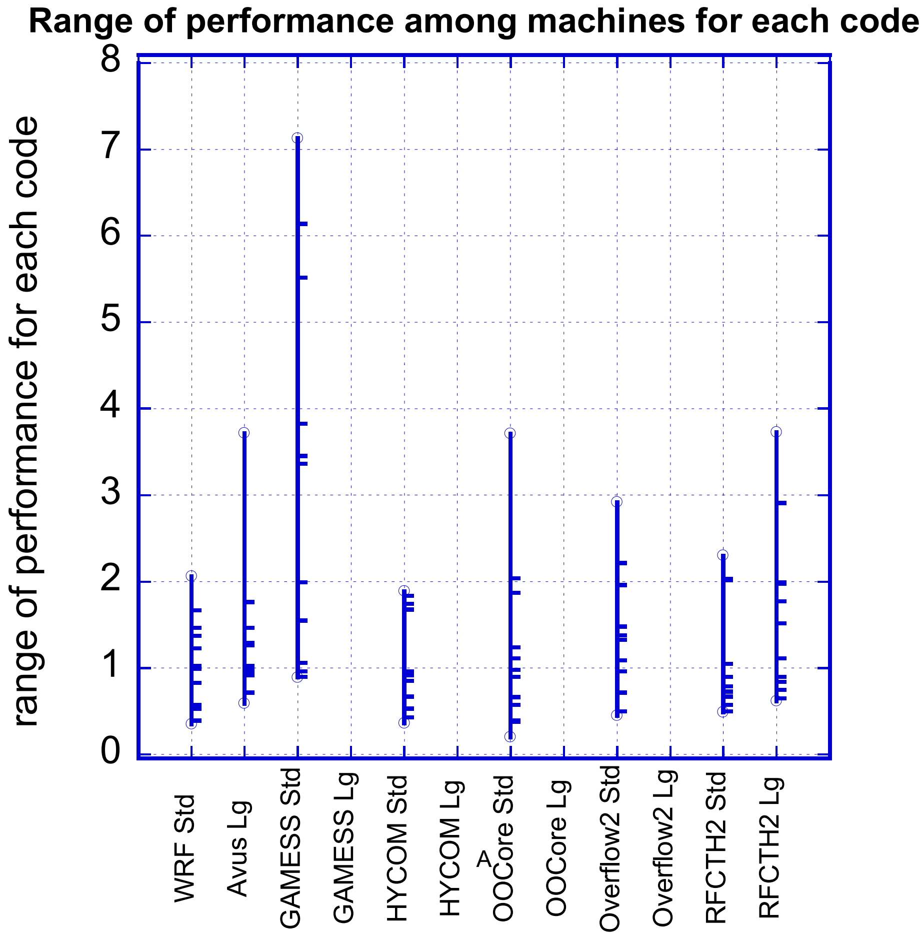
Code Performance (by machine)



Code performance (grouped by machine)



Performance range of codes is large.



General conclusions

- Performance depends on application and on the computer
- Tuning for a platform can pay off in a big way
- Shared memory is really good for some codes

And the winners could be:

- Green (fastest sprinter): SGI Altix on Gamess, followed by IBM P4+ on Gamess, but depends on application
- Polka Dot (most capable): SGI Altix(2.41), Cray X1 (2.01), IBM P4+ (1.54), IBM Opteron (1.51): based on the weighted performance for the DoD benchmark suite
- White (best youngest): Linux Networx
- Red (most aggressive): no data
- Team Jersey (best team): HPCMP suite of computers
- Yellow Jersey (best overall computer): depends on application but the HPCMP suite comes closest