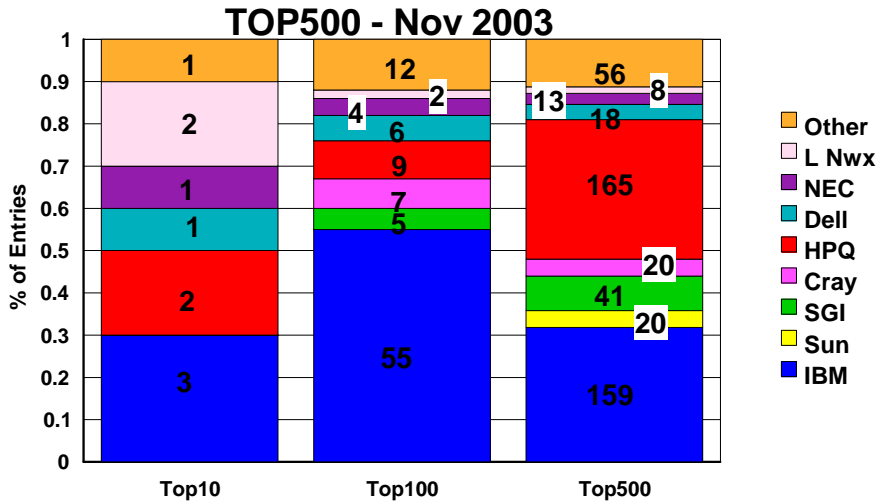


IBM Weather Market Update



- **David Blaskovich**
IBM Global Sales Executive
for HPC Weather & Environmental Markets
IBM Deep Computing

Top500 Leadership

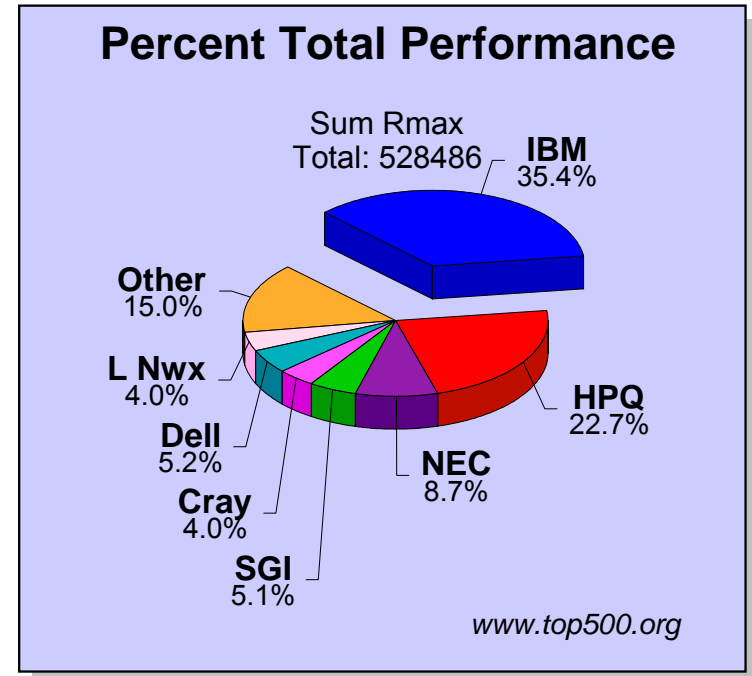


Source: www.top500.org List of Supercomputers, Nov 2003

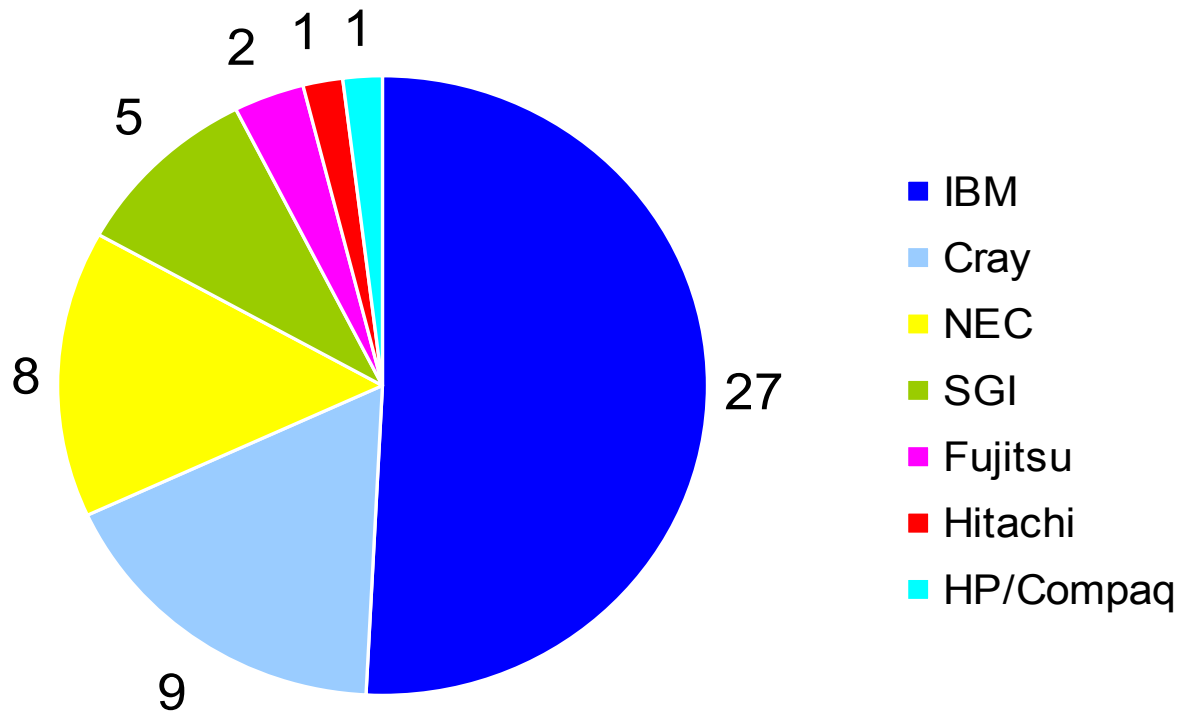
Semiannual independent ranking of top 500 supercomputers in the world

IBM is dominant vendor:

- ✓ 30% of Top 10
- ✓ 55% of Top 100
- ✓ 52.7% of systems over 1 Teraflop
- ✓ Virtual tie for lead of most systems (159 IBM vs. 165 HPQ)
- ✓ Increased lead of installed aggregate performance: 186.9 Teraflops, 55.9% higher than HPQ
- ✓ 43.6% of Linux Clusters
- ✓ New architecture of IBM BlueGene/L ranked #73 for small prototype
- ✓ #3 system at Virginia Tech based on IBM PowerPC 970 chip technology



Weather/Environmental Market HPC Suppliers



IBM Worldwide Weather Customers - I

- U.S. NOAA National Center for Environmental Prediction (NCEP) 2x22xp690
- U.S. NOAA National Climatic Data Center (NCDC) p690 HPSS
- U.S. National Center for Atmospheric Research (NCAR) WH-2/31, 3NH2
47xp690
- U.S. Naval Oceanographic Office (NavO) WH-2/334, 44xp690/32
- U.S. Navy Fleet Numeric Meteor & Oceanography Center (FNMOC) X-series
- U.S. Air Force Weather Agency (AFWA) 14xp655
- U.S. Environmental Protection Agency (EPA) NH-2/2, p690
- Raytheon NPOESS 2xp690/32
- Duke Power p690/8
- Williams Energy & University of Oklahoma 4xp690/32
- Environment Canada (EC) Meteor. Services of Canada (MSC) 27xp690
- European Centre for Medium-Range Weather Forecasts (ECMWF) 2x30xp690
- Finnish Meteorological Institute (FMI@CSC) WH-2/32
- German Deutscher Wetterdienst (DWD) NH-2/80
- German Potsdam Institute fur Klimate (PIK) 30xp655/8-way

IBM Worldwide Weather Customers - II

- Ireland Met Eireann WH-2/10
- Norway Meteorological (met.no) Linux Cluster 1350
(2 GHz Opteron) 2x40
- UK Proudman Oceanographic Lab. (POL) X-series
- Hungarian Meteorological Service (HMS) p690/32
- Morocco Direction de la Meteorologie Nationale (DMN) NH-2/3
- Slovakia Hydrometeorology p690/32
- Turkey General Dir. of State Meteorological Works (DMI) p690/32
- Tunisia l'Institut National de la Météorologie (INM) p690/8
- China National Climate Centre (NCC) NH-2/10
- China National Meteorological Centre (NMC) SP2/32
- China Hong Kong Observatory (HKO) WH-2/6, p690/20
- Thailand Meteorological Department (TMD) SP2

ECMWF - European Centre for Medium-Range Weather Forecasts



**eServer Cluster 1600 of POWER4
processors building up to a 20 TF
supercomputer**

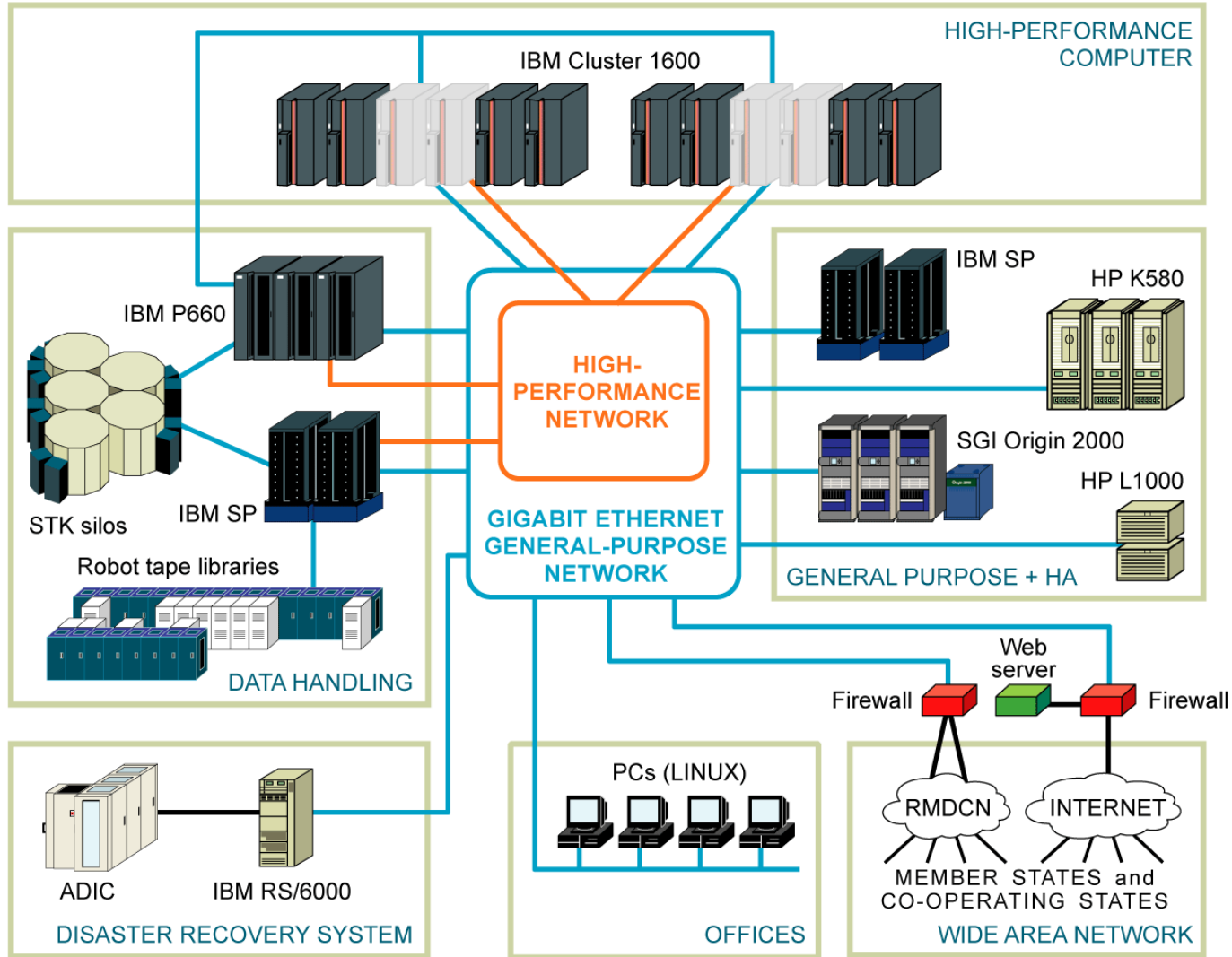
**Data Handling system based on IBM SAN
and HPSS**

Linux Intellistations for all desktop systems

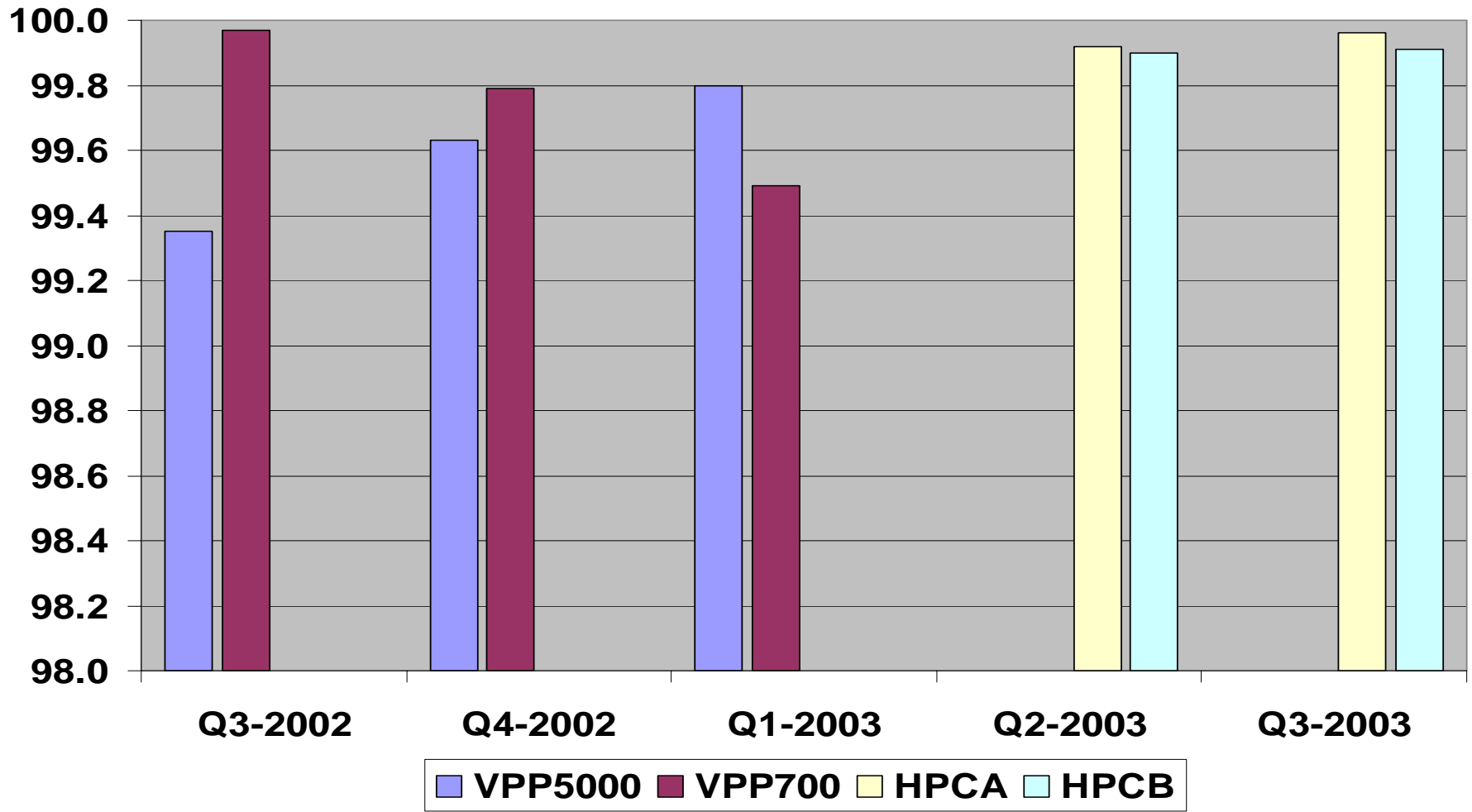


ECMWF Computer configuration

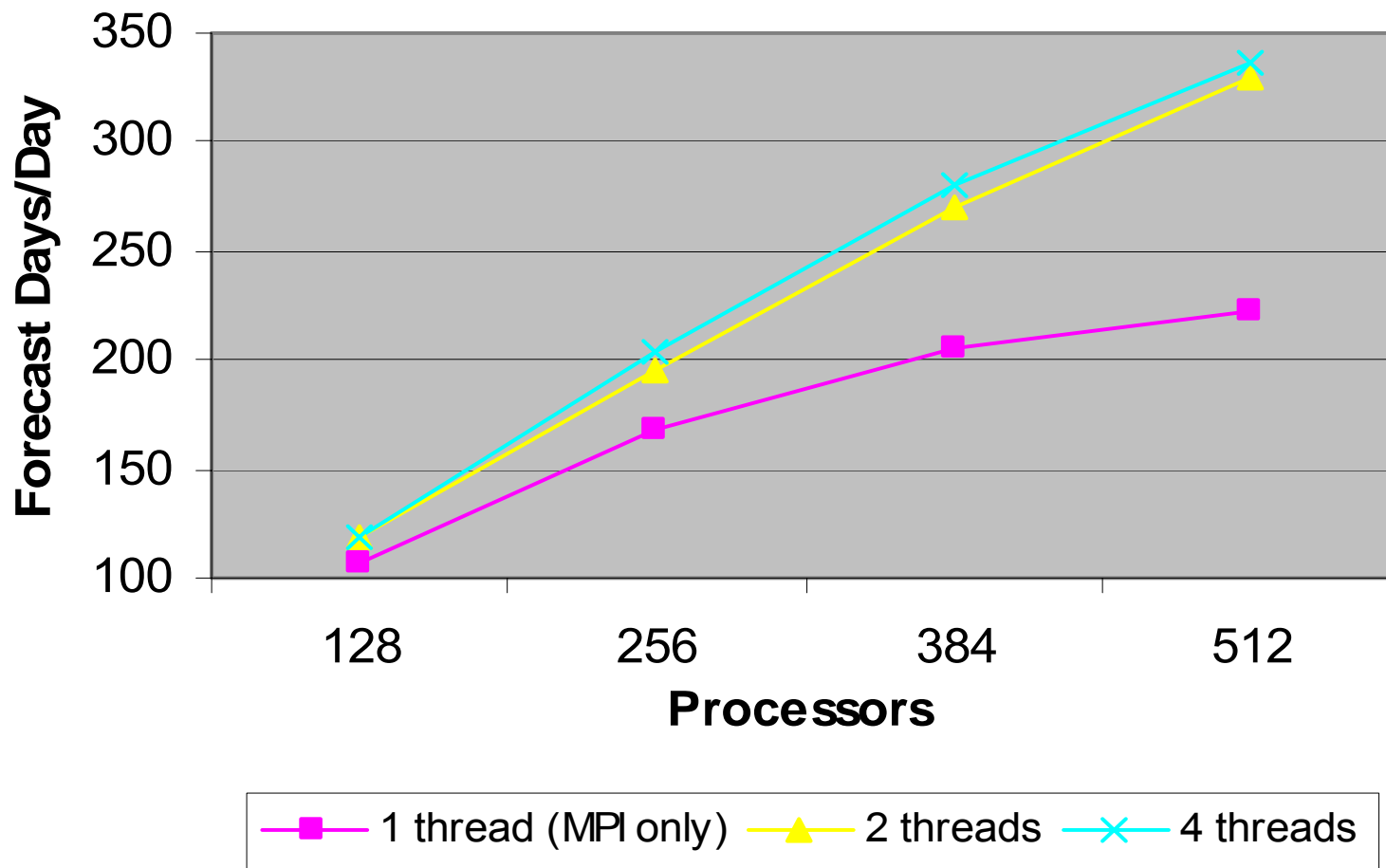
System Solution: HPC, DHS, & desktop, for ECMWF, NCEP, DWD



2003 ECMWF User availability of the HPC systems



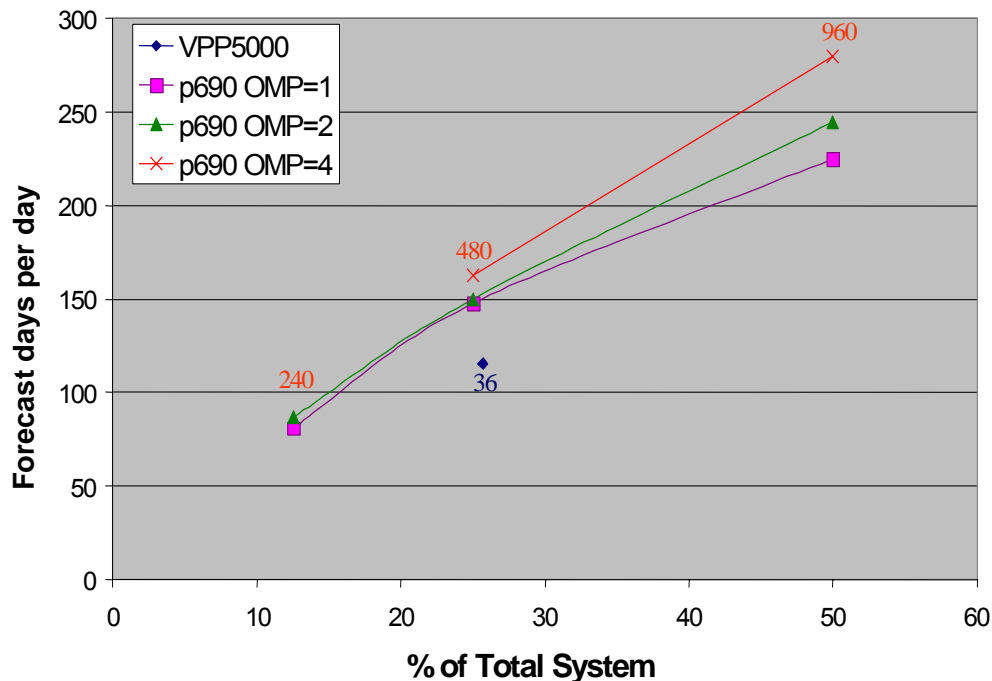
TL511L60 forecast model (26R3)



Scaling from 128 to 512 processors: x3.12 or 78%

ECMWF Performance

High Resolution forecast : T799 L90



- Timestep = 15 minutes
- Two clusters of
 - 30 p690 frames
 - with 32 CPUs/frame.
- Results given as percentage of total system.
- Operational 10 day forecast runs on 256 processors.
- VPP system is 140 VPP500 processors = 400 GFlops sust

OMP = number of OpenMP (Shared Memory) threads.

The number of MPI tasks = total number of processors divided by OMP.

The scaling with number of processors was greatly improved (by up to a factor of two) by organising the communication so that MPI tasks send and receive data at the same time, but do not overlap with application processing.

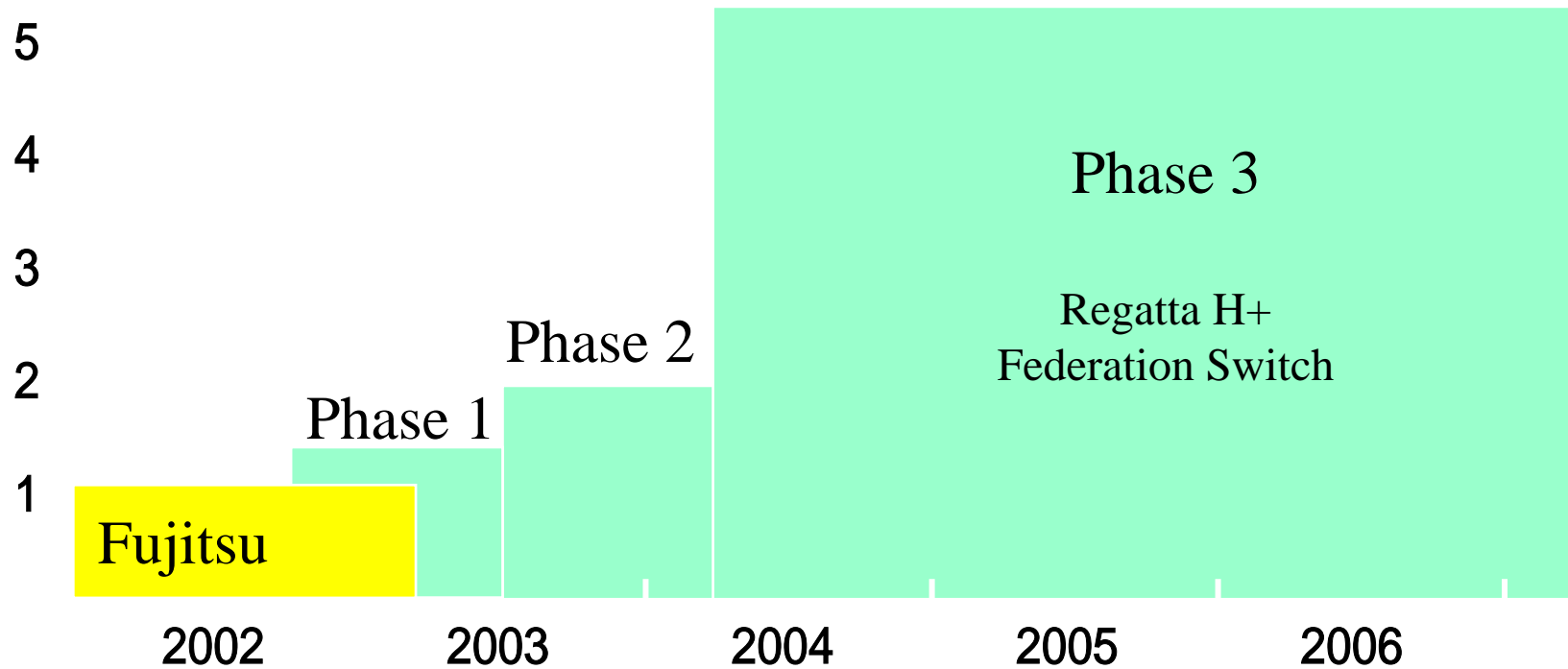
This is best because communication uses a significant amount of CPU time, and do not want application processing to interfere with communication.

Secondly, by using OpenMP, we reduce the number of MPI tasks involved in communication, and because only the master thread participates,

there are spare CPUs available to handle spurious AIX related interrupts.

Source: John Hague & Deborah Salmond

Performance profile of the contracted IBM solution relative to VPP systems



Performance on ECMWF codes (Fujitsu = 400 GF sustained)

ECMWF Data handling system

- ECMWF has used Tivoli Storage Manager (TSM) for several years but encountered scalability issues
- In 2001, HPSS was chosen as a replacement
- Projected data storage volumes (including backups)
 - mid-2003: 1.4 petabytes
 - mid-2004: 2.1 petabytes
 - mid-2005: 3.2 petabytes
- Experience with HPSS is very positive:
 - Complex system but very scalable
 - Excellent support when problems are encountered
- Timely migration of the data was a major concern at the project planning stage

*40+ instances of HPSS installed at
20+ organizations throughout the world*

The World of HPSS



U.S. NOAA National Centers for Environmental Prediction

<http://wwwt.ncep.noaa.gov/>



2002

\$224.4 M contract over 9 years

Largest Supercomputer Contract for IBM

Supercomputer, services, and hosting facility (Gaithersburg, MD)

Committed performance levels throughout contract to improve forecast accuracy

7.3 TFLOPS in 2002

2x Regatta/22 Systems + 42 TB storage

2004

36 additional Regattas at 1.8 GHZ

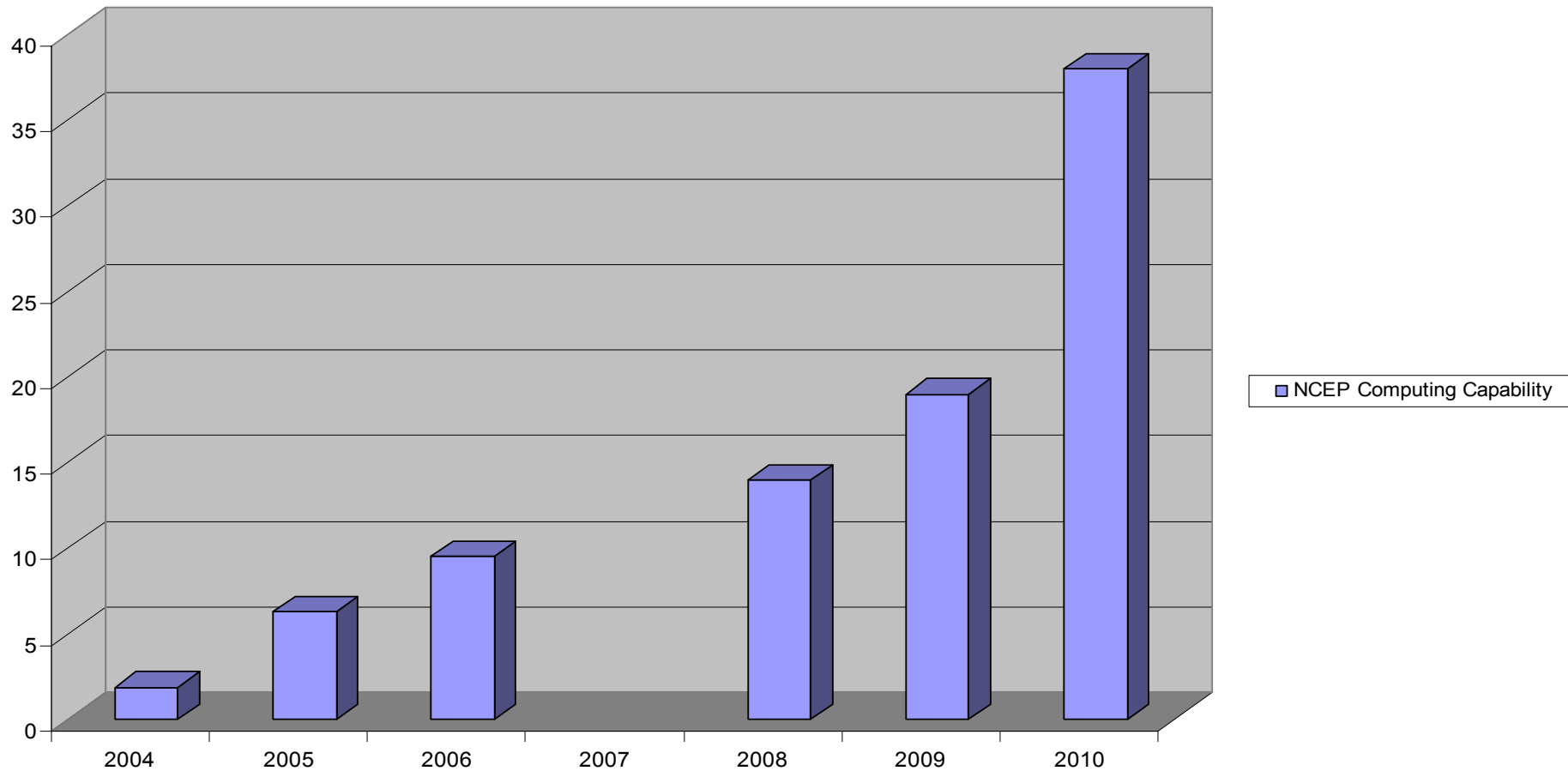
2005 and beyond - Power5

2009

100 TFLOPS

Forecast model from 80km to 40km - 2006

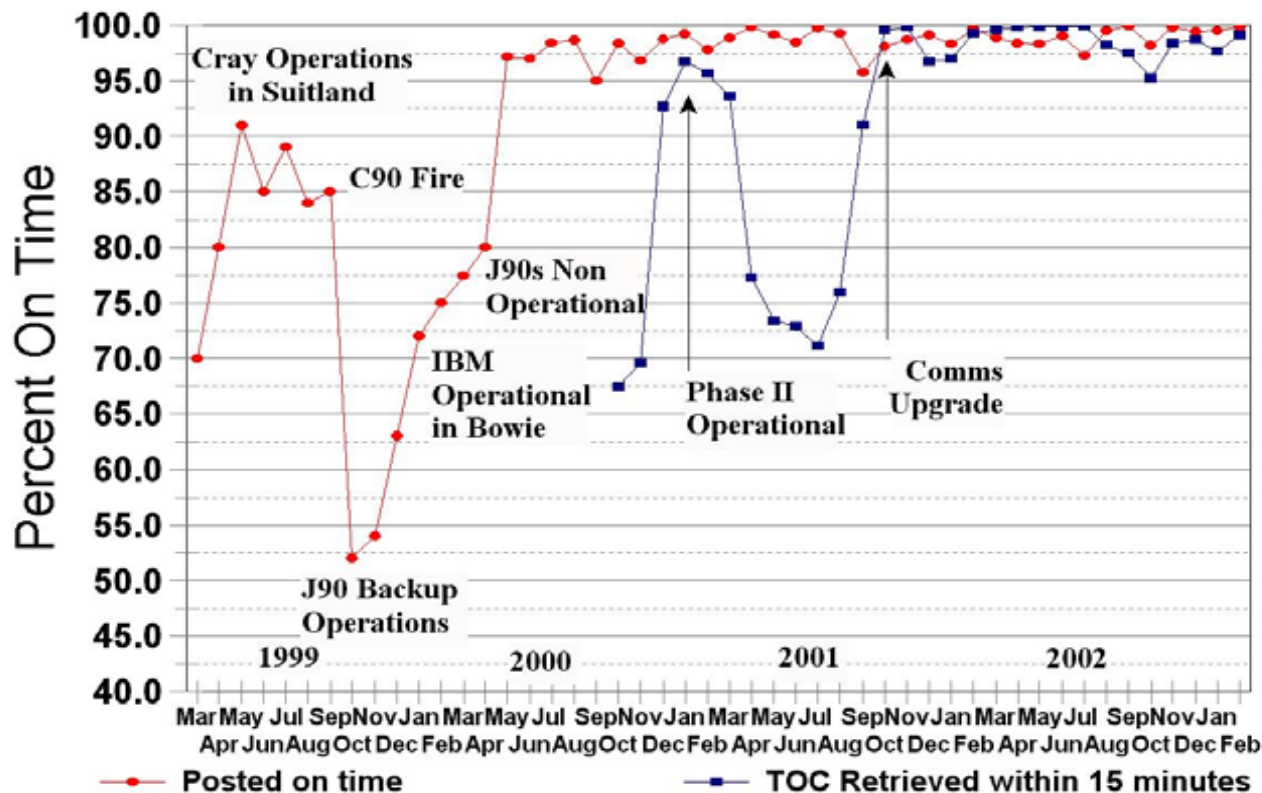
NCEP Weather Computing Capability Relative to December 2002



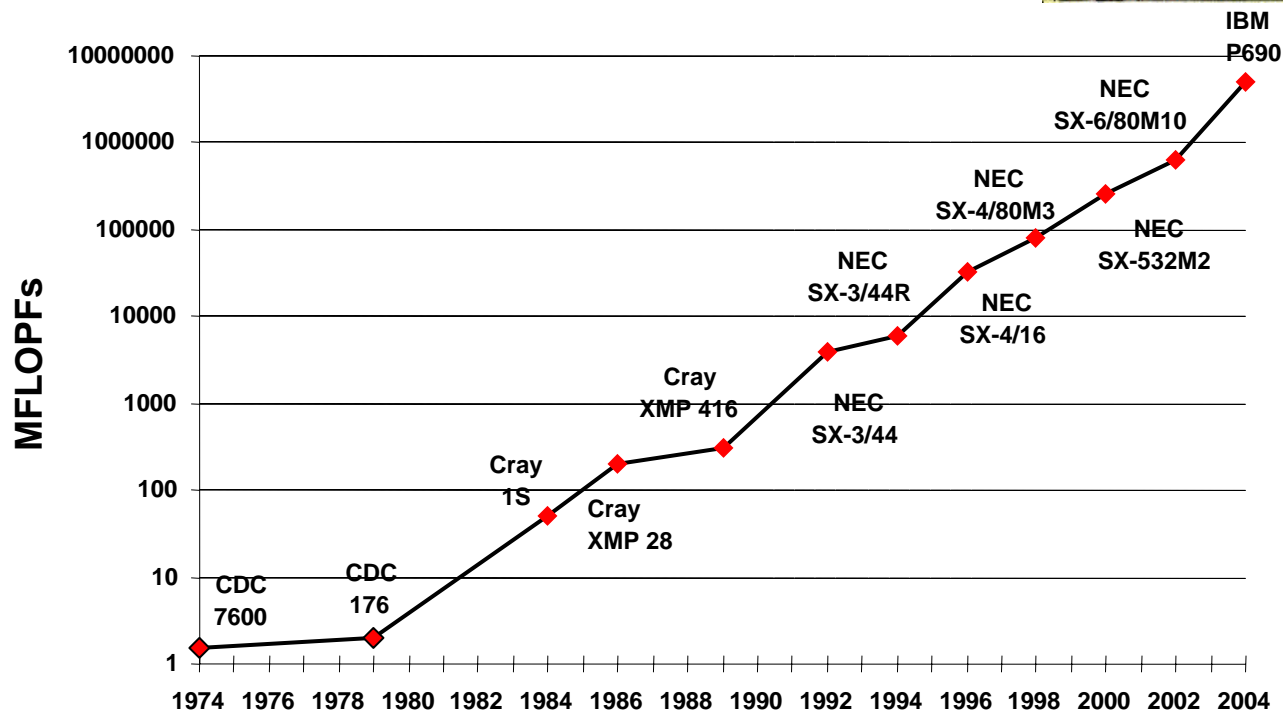
IBM System Reliability

Source: Dr. Lou Uccellini, Director NOAA NCEP 173,260 forecast products delivered daily

Product Generation Summary for NCEP Computer Systems



**Environment Canada (EC)
Canadian Meteorological Centre (CMC)
Meteorological Service of Canada (MSC)
Supercomputing History**



MSC Supercomputers: Then and Now

NEC

- 10 node
- 80 PE
- 640 GB Memory
- 2 TB disks
- 640 Gflops (peak)



IBM

- 128 nodes
- 936 PE
- 2.12 TB Memory
- 15 TB Disks
- 4.825 Tflops (peak)



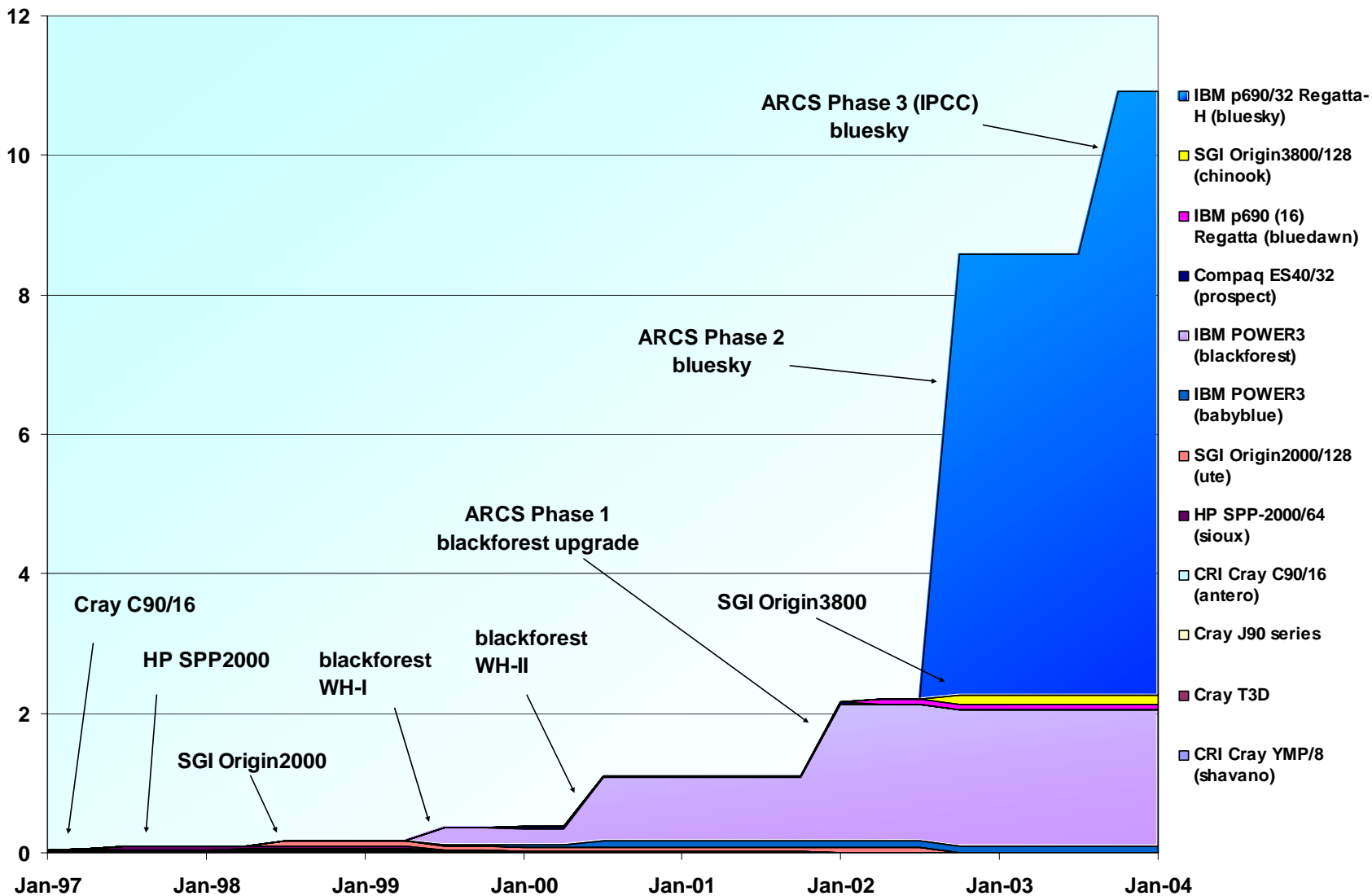
IBM Clusters at NCAR

- ***Bluesky***: 1024 IBM 1.3 GHz Power-4 cluster
32 P690/32 compute servers
736 in 92, 8 way “nodes” (bluesky8)
288 in 9, 32 way “nodes” (bluesky32)
Peak: 5.234 TFlops
Dual “Colony” interconnect
- ***Blackforest***: IBM 375 MHz Power-3 cluster
283 “winterhawk” 4-way SMP’s
Peak: 1.698 TFlops
TBMX interconnect

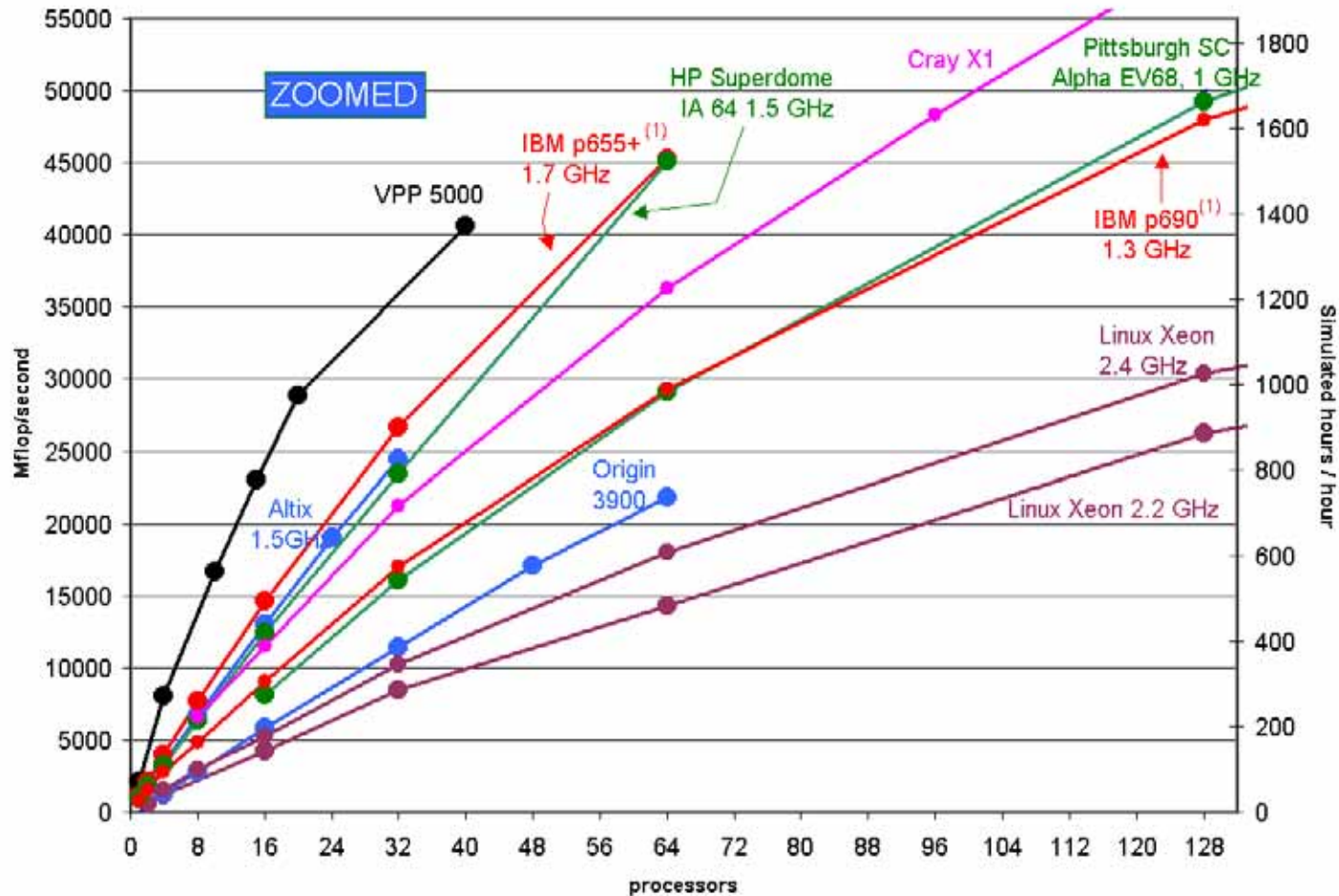


<http://www.ncar.ucar.edu/>

Peak TFLOPs at NCAR



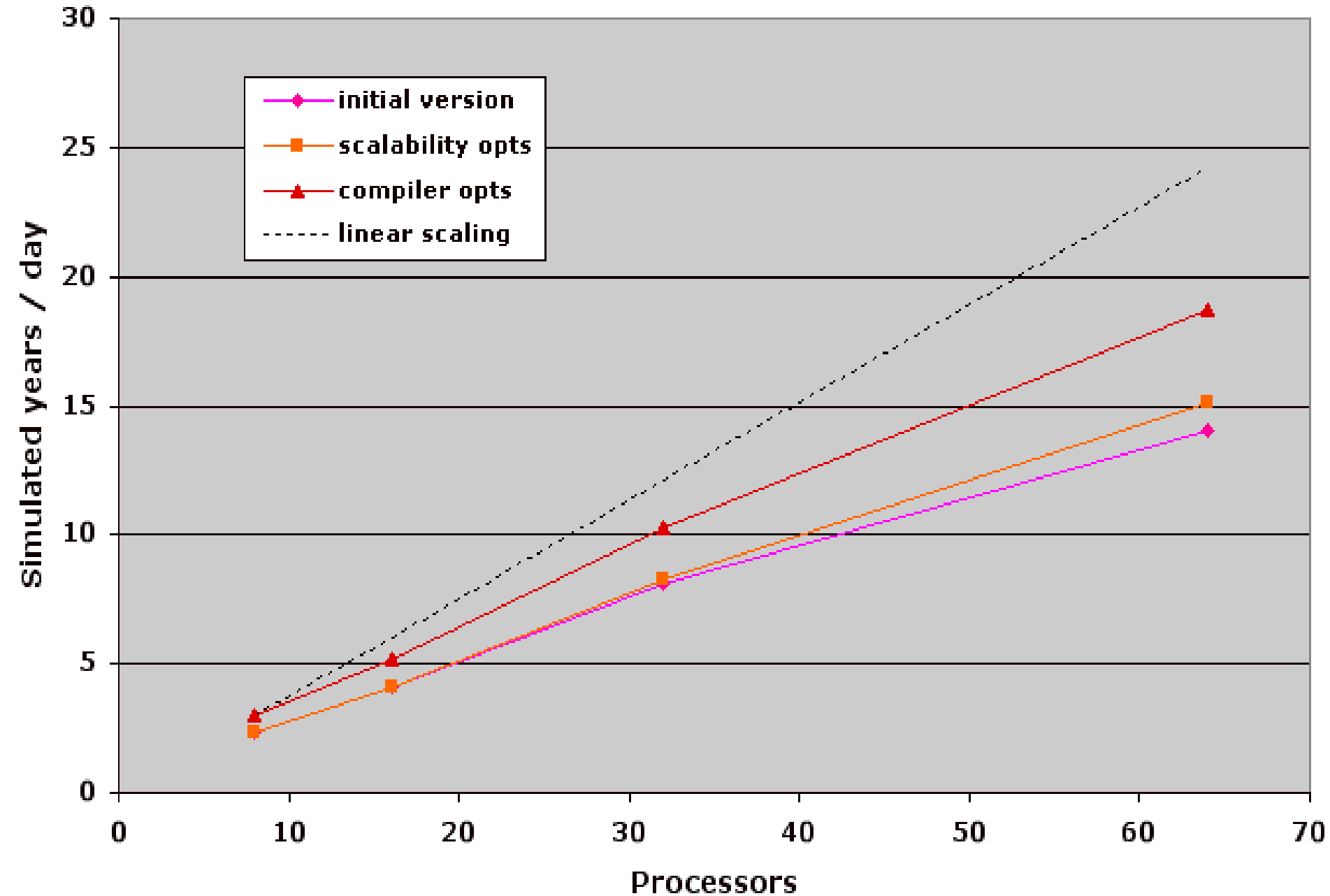
Comparative Sustained Performance



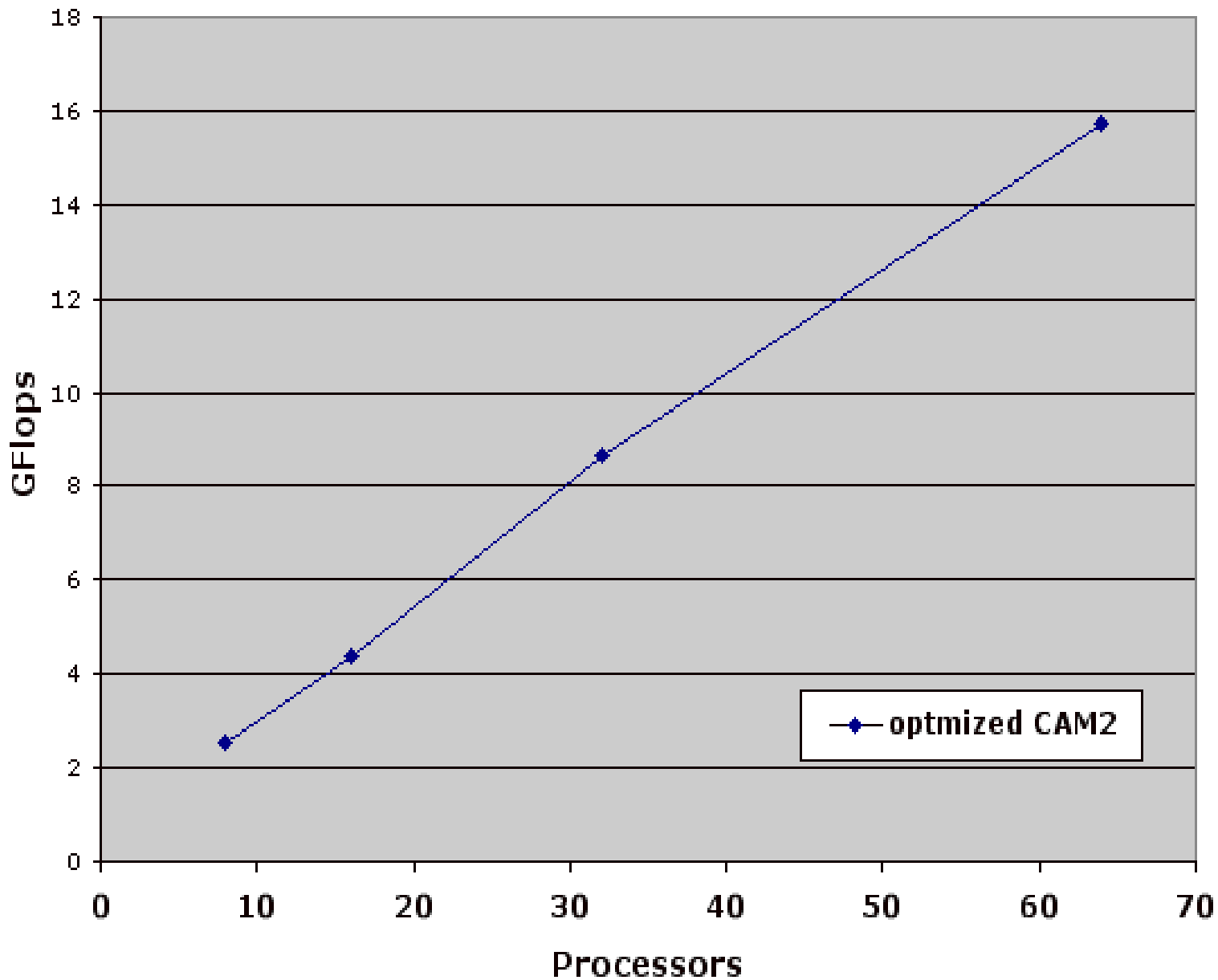
MM5 Floating Point Performance on Various Platforms - October 2003 - Jim Abeles

<http://box.mmm.ucar.edu/mm5/mpp/helpdesk/20030923.html>

T42L26 CAM2.0.2 Performance Improvements

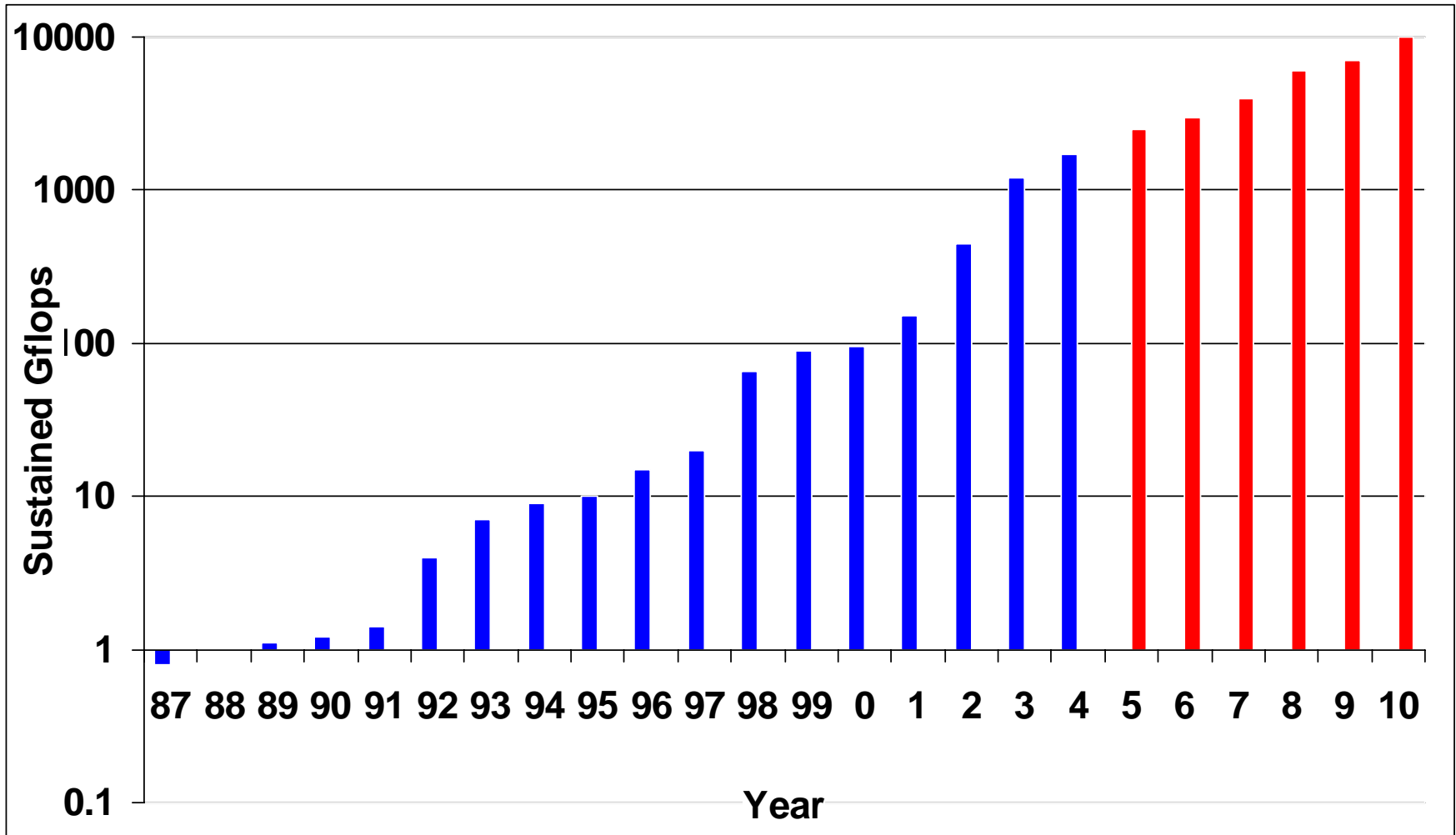


T42L26 CAM2.0.2 Floating Point Performance



Sustained Performance Trend

(Sustained performance for atmospheric and ocean models)
Collated from several weather HPC customers.
(NCEP, ECMWF, NCAR, UK Met, DWD, etc.)



IBM's Deep Computing Strategy

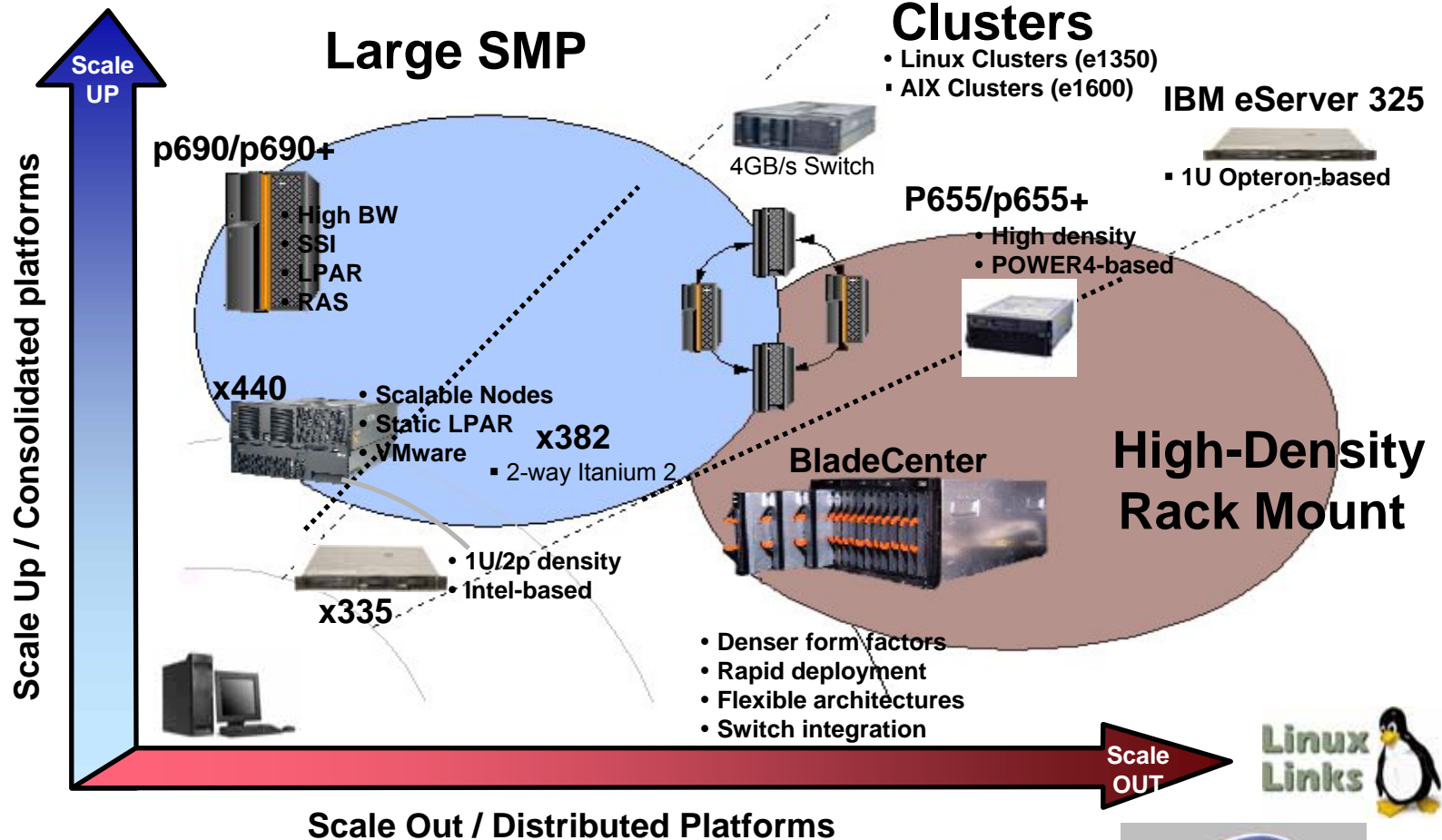
Solving Problems More Quickly at Lower Cost

- Aggressively evolve the POWER-based Deep Computing product line
- Develop advanced systems based on loosely coupled clusters
- Deliver supercomputing capability with new access models and financial flexibility
- Research and overcome obstacles to parallelism and other revolutionary approaches to supercomputing



IBM @server

Positioning by Scaling UP or Out

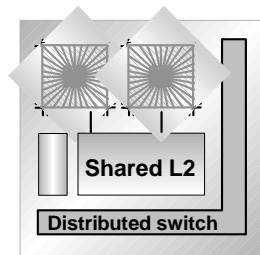
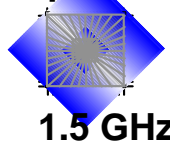


Power4+ Building Block

p655 Chip / Module Structure Build-up

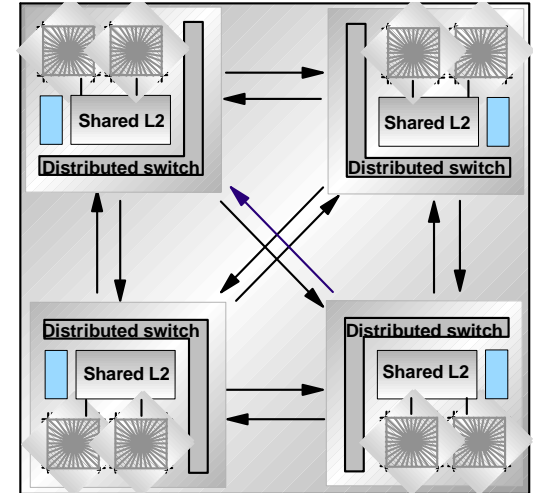
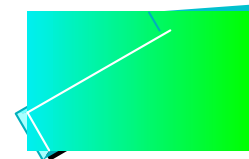
2-way POWER4+ SMP with
shared L2 cache

POWER4+
Microprocessor



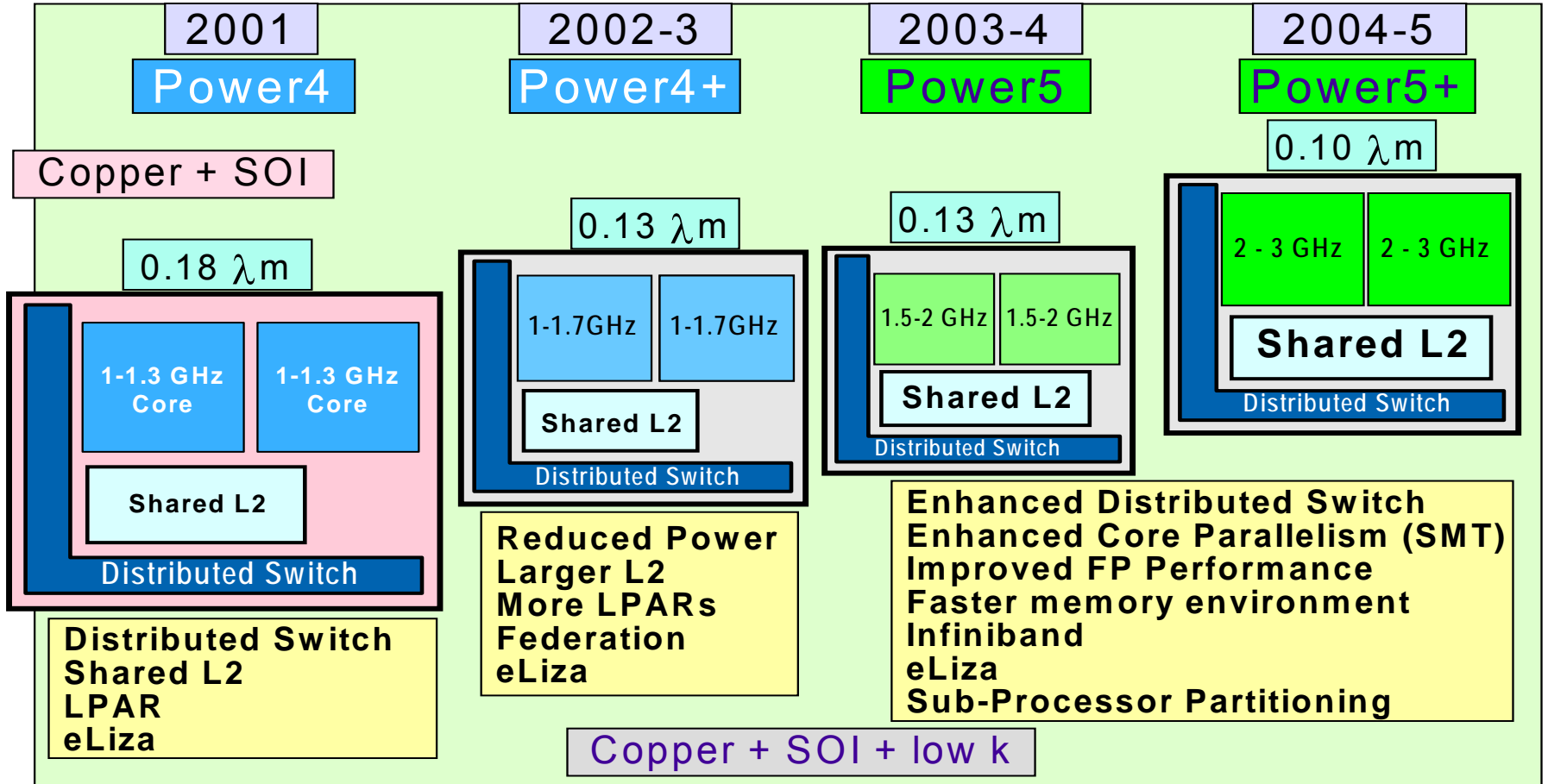
2-way POWER4+ SMP system
on a single chip!

(184 million transistors)



8-way (4 chips) POWER4+
SMP system on a
Multi-Chip Module (MCM)

POWER Processor Roadmap



Power5 Goals

- Improve performance per cycle
- Dynamic power management
- Lower power mode
- Larger systems: 64 cores (128-way SMT)
- Improve Cache and Memory performance
- Fast hardware lock, page mover, barrier synchronization register
- Add Simultaneous Multi-threading (SMT) with dynamic switching
ST to/from SMT, dynamic thread priority
- Improve scalability
- Sustained Bandwidth per MCM: 4X over P4
- Sustained Floating Point Performance per MCM: 1.5-2.5X P4

IBM's contribution to Linux

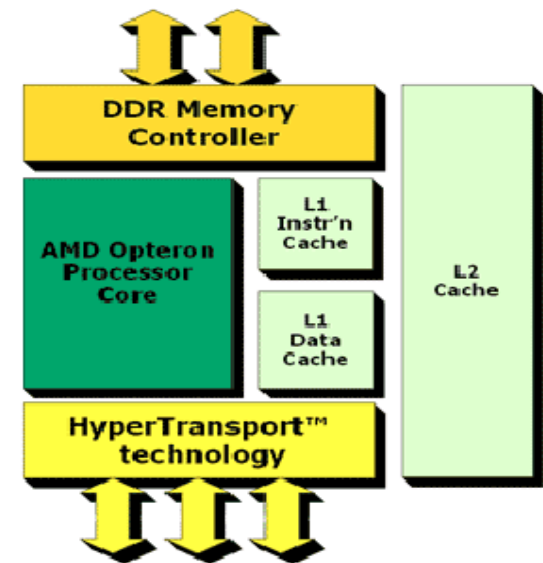
Over \$2 Billion invested

- **250+ developers worldwide**
- **70+ active projects**
- **Critical contributions to current and in-progress versions of Linux:**
 - Scalability
 - System Performance
 - Security
 - Serviceability
 - Internationalization
 - Availability
 - Cluster Management
 - System Management
 - Stress Testing
 - Networking
 - Standards & Documentation
 - Reliability
- **Members of the Linux Community**
 - Linux Standards Base
 - Free Standards Group
 - Linux Internationalization
 - Kernel Summit
 - Linux Weekly News
 - Open Source Development Lab
 - USAGI (IPV6) Project
 - GNOME Foundation
 - KDE League
 - OSDL Carrier Grade
 - OSDL Data Center



IBM eServer 325 Opteron-based Solution

- **Unparalleled price/performance for 32- and 64-bit high performance computing workloads**
 - ▶ Exceptional scalability
 - ▶ #1 TPC-H benchmark
 - 2X overall performance of all other reported results
- **Investment Protection**
 - ▶ Allows customers to deploy 32- and 64-bit applications on the same platform
 - ▶ Enables organizations to deploy new 64-bit applications while continuing to leverage existing investments in legacy programs
- **IBM worldwide capability**
 - ▶ IBM's seamless end-to-end offerings help enable maximum uptime and optimal productivity for global companies





Norway's Met Office Selects IBM's Linux Cluster for Atmosphere and Ocean Prediction

Solution

Linux Cluster 1350 with e325 nodes

40, 2-CPU 2 GHz Opteron connected with Myrinet

Norway Met 's use of the e325 Cluster:

- HIRLAM (High-resolution limited area model) consortium
 - Norway, Denmark, Finland, Iceland, Sweden, Netherlands, Ireland, Spain
- Daily production of air pollution forecast for some of the municipalities in Norway
- Development of new methods and models for operational weather forecasting.
- Research in the areas of atmospheric science, oceanography, climate change and transport of air pollution.

Why IBM

- the high sustained performance of met.no's models on the system,
- an integrated solution in the form of the IBM e1350 cluster populated with the e325 nodes, which are based on Opteron processors, and
- the flexibility of Opteron processors operating in mixed 32-bit and 64-bit mode.



- 1100 e325 Opteron Nodes, 140 IA-64 4-way nodes
- Applications: Basic research, life sciences

AIST (National Institute of Advanced Industrial Science & Technology)

Challenge

AIST, Japan's largest national research organization needed to provide an on-demand computing infrastructure which dynamically adapts to support various research requirements of its collaborators focusing in areas of Grids, life science, and nanotechnology.

Story

IBM, HP, Hitachi, and Fujitsu all competed for this opportunity. IBM provided the most powerful single Grid system planned to be installed in the spring of 2004.

Solution

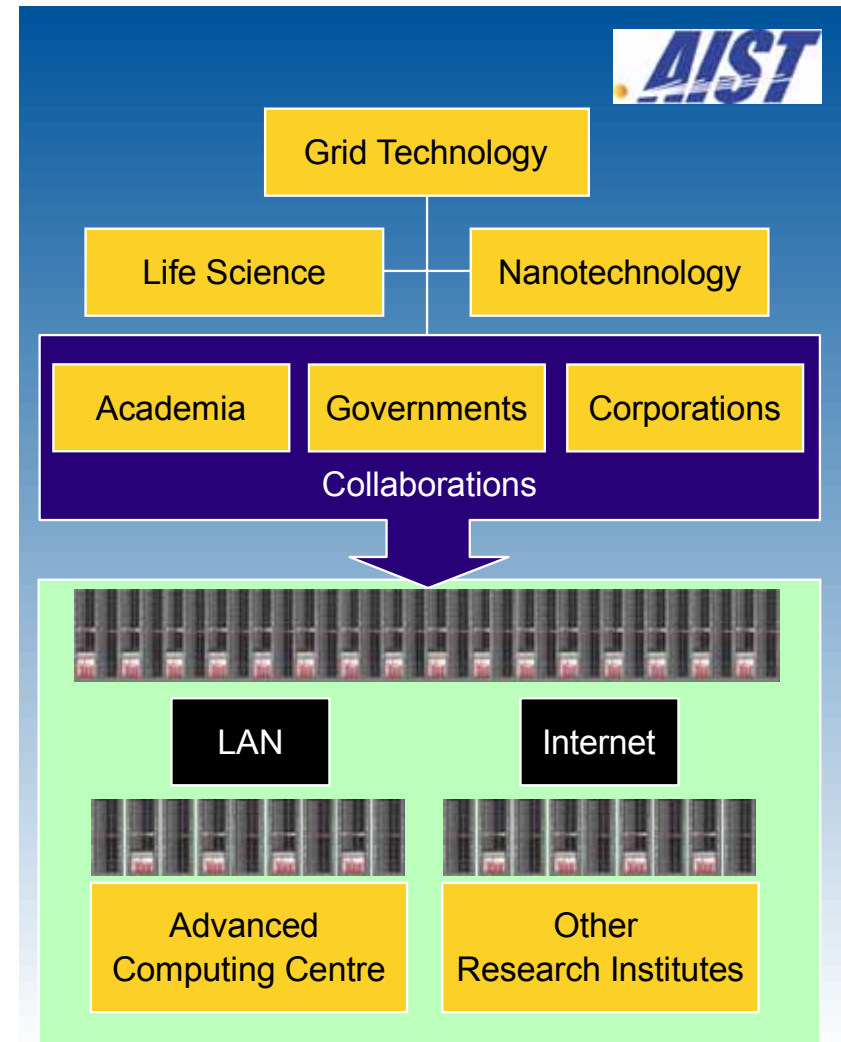
Linux Cluster

- 2116 CPU AMD Opteron Cluster
 - 520 CPU Intel Madison Cluster
- Globus Toolkit 3.0 (OGSA)

World's most powerful Linux-based supercomputer

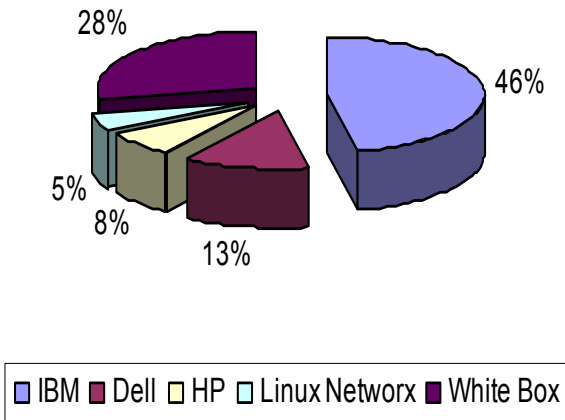
- More than 11 TFLOPS Rpeak
- More powerful than the current third most powerful supercomputer in the world

Research & Development

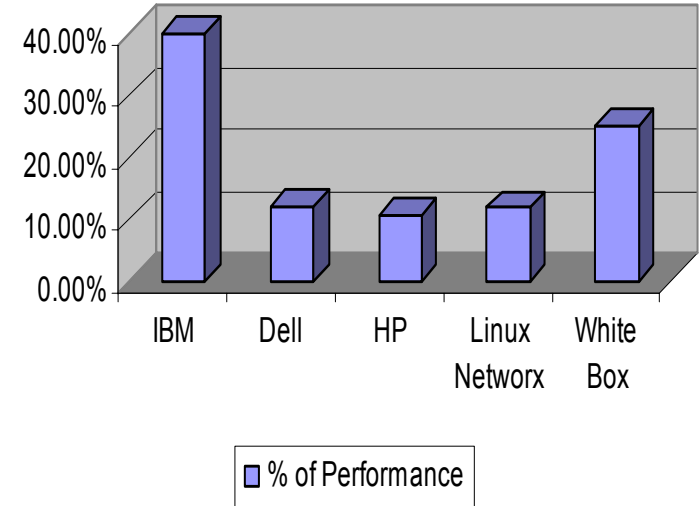


TOP500 List of Supercomputers – Linux Clusters

of Clusters - June 2003 Top500



Aggregate R-max - June 2003 Top500



IBM Pathways to Deep Computing

- **Single Integrated Systems**

 - Single systems image SMP's

 - Clusters of dense systems

- **Grids**

 - Access to remote resources available on demand

 - Computational and data-intensive

 - Self-managing and self-healing systems

 - React to conditions and redistribute workload

- **On Demand Computing**

 - Delivery of standardized processes, applications and infrastru network

 - Utility-like model -- quick access to incremental capacity

