# Pittsburgh SC Terascale

# NSF Terascale Computing Initiative


Ralph Roskies

Scientific Director

Pittsburgh Supercomputing Center


September 21,2000


*Roskies@psc.edu*

# Solicitation Synopsis

- Single, new, terascale computing system to enable U.S. researchers in all science and engineering disciplines to gain access to leading edge computing capabilities.
  - System balanced in processor speed, memory, communication and storage systems.
  - System software comparable to that on other high-performance systems

# Motivation from solicitation

- High-end computing is essential to science and engineering research
- Both for the sake of fundamental scientific research and to enable applications to benefit from the research, the research community needs access to systems at the leading edge of capability.

# System requirements from discussion with users

- Excellent single processor performance
- Adequate memory per processor for large data structures and codes, including those from ISV's
- High bandwidth, low latency inter-processor communication
- Ability to take periodic snapshots with minimal effect on computational speed (major I/O demands are for snapshots and checkpointing)
- Large scratch disk space
- Fast networks for real-time use

# Software requirements from discussions with users

- Scheduling to dedicate a large fraction of the machine to single users
- Good Fortran, C and C++ compilers,
- Effective debuggers and performance tools.
- Widely available instruction set allowing development on remote, low-cost, commodity systems
- Very little need for global shared memory.

# Proposed System

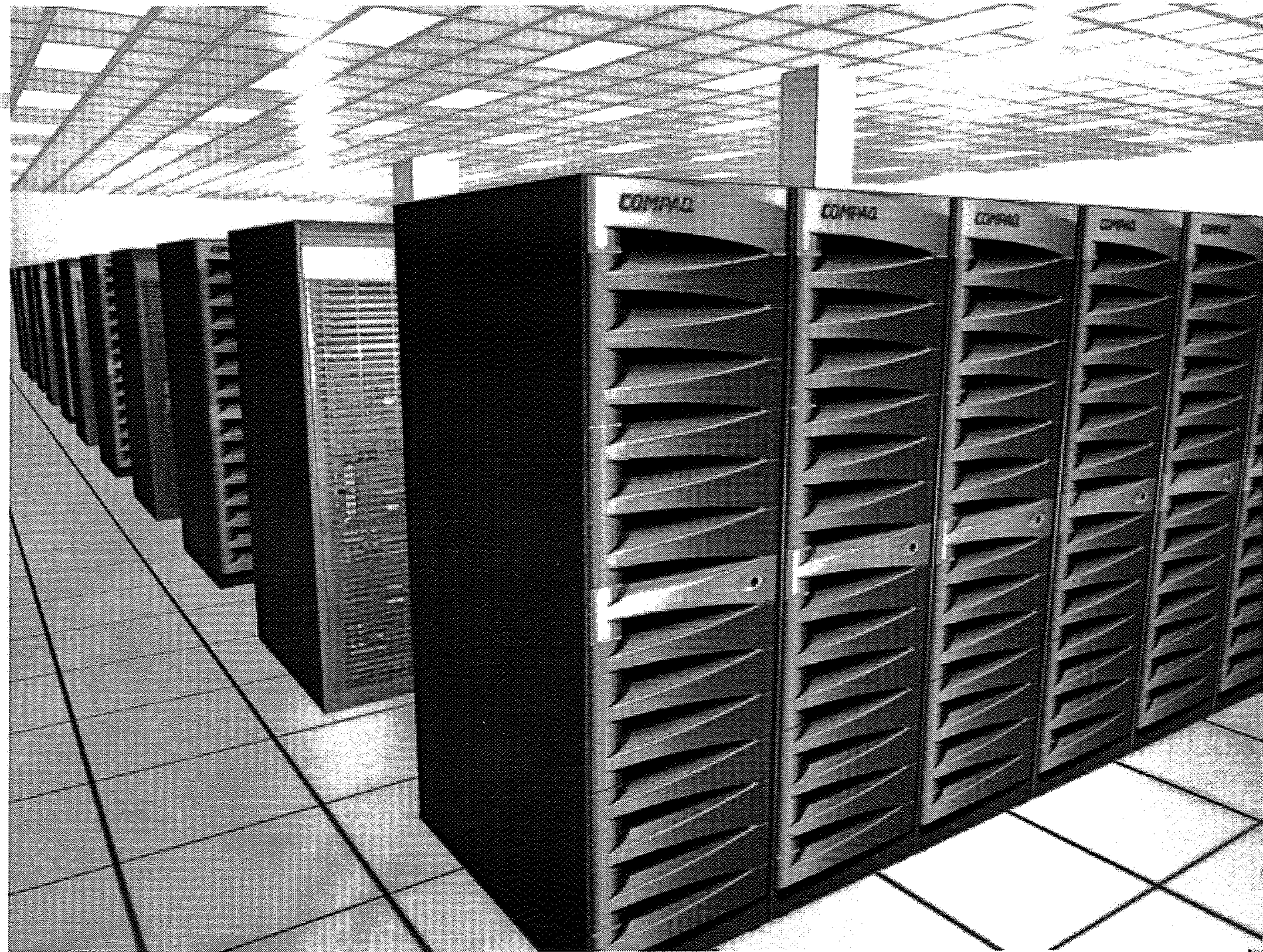
- 682 Compaq nodes, each with 4 GB memory and 4 next-generation Compaq Alpha processors. In aggregate, 6 teraflops peak, 2.7 terabytes memory
- 25 TB of disk local to the individual nodes for booting, local system functions, and local scratch space, with an aggregate bandwidth of 20 GB/sec
- 30 of the nodes also serve as I/O nodes. They have attached RAID disks, with a total of 27 TB of storage and an aggregate bandwidth of over 18 GB/s.

# Proposed System

- Interprocessor network from Quadrics Supercomputing World- nodes can receive and send at a bandwidth of 400MB/sec each, with application code latencies of ~5 μs.
- Visualization subsystem with hardware support for parallel, high-speed, on-the-fly rendering
- High speed links to a file server initially having ~300TB capacity.

# Summary

- Extraordinary computational capability
- Excellent interprocessor communication
- Ability to snapshot memory to disk in less than 3 minutes
- Ability to write to tape at 1TB/hour.
- Considerable attention to redundancy for robustness.

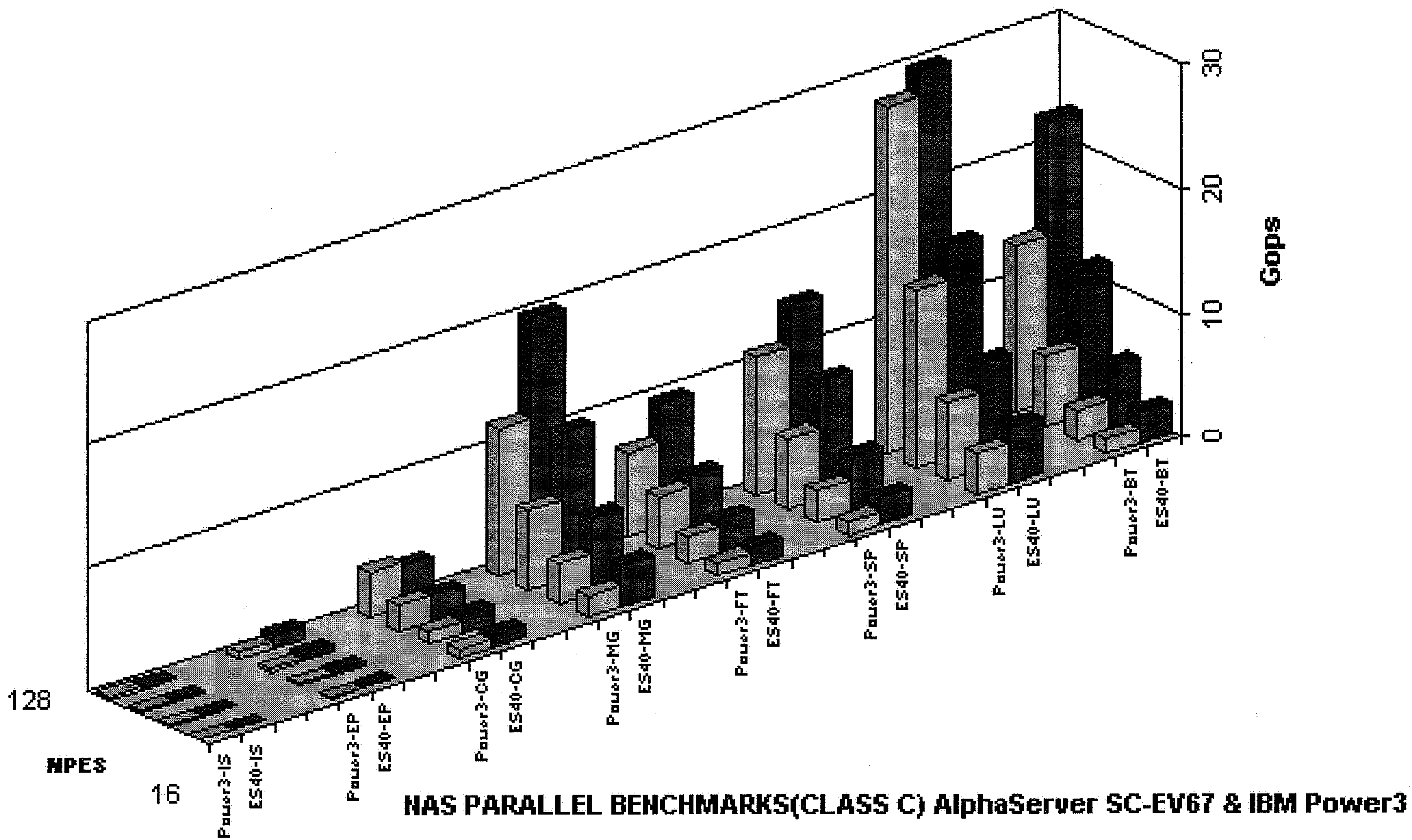

COMPAQ

# Why Compaq?

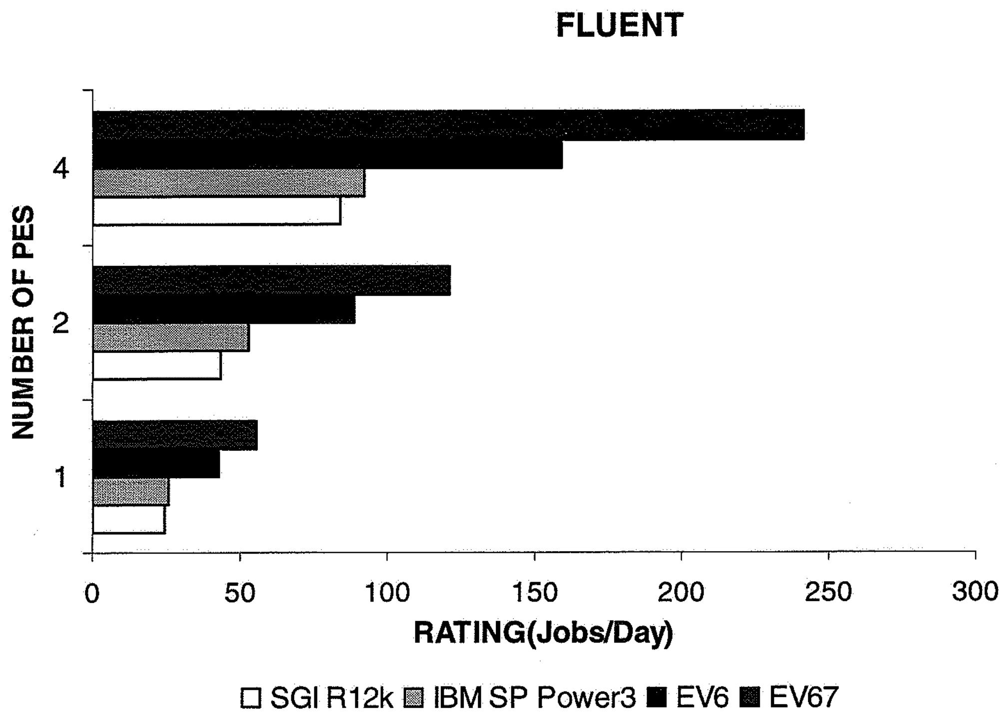
- Superior technical performance
- Excellent credible upgrade path
- Diversity for the PACI program, which has large machines from IBM and SGI

# Compaq processors

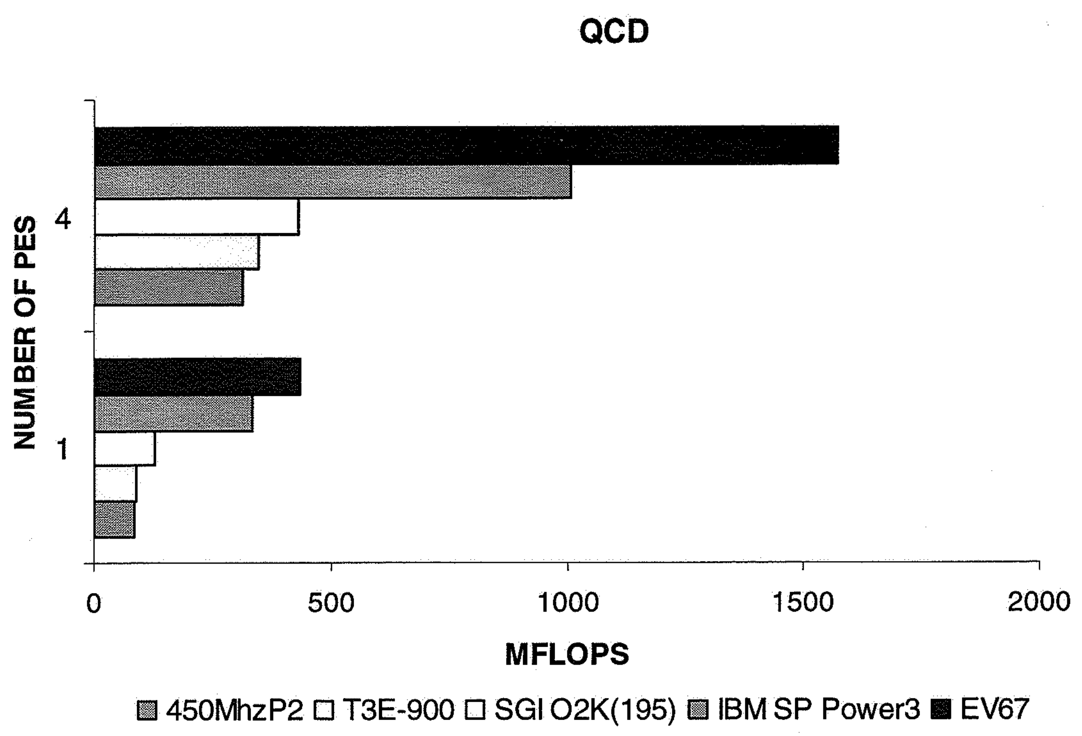
| CPU | System | Clock (MHz) | FP Per Clock | FP Peak (Gf) | SPECint95 | SPECfp95 |
|---|---|---|---|---|---|---|
| EV4 | T3D | 300 | 1 | 0.30 | 4.5 | 6.5 |
| EV5 | T3E-900 | 450 | 2 | 0.90 | 14.1 | 27.0 |
| EV5 | T3E-1200 | 600 | 2 | 1.20 | 18.8 | 29.2 |
| EV67 | ES40 | 667 | 2 | 1.34 | 40.0 | 82.7 |
| Next | Generation | >1000 | 2 | >2.00 | Est 66 | Est 132 |

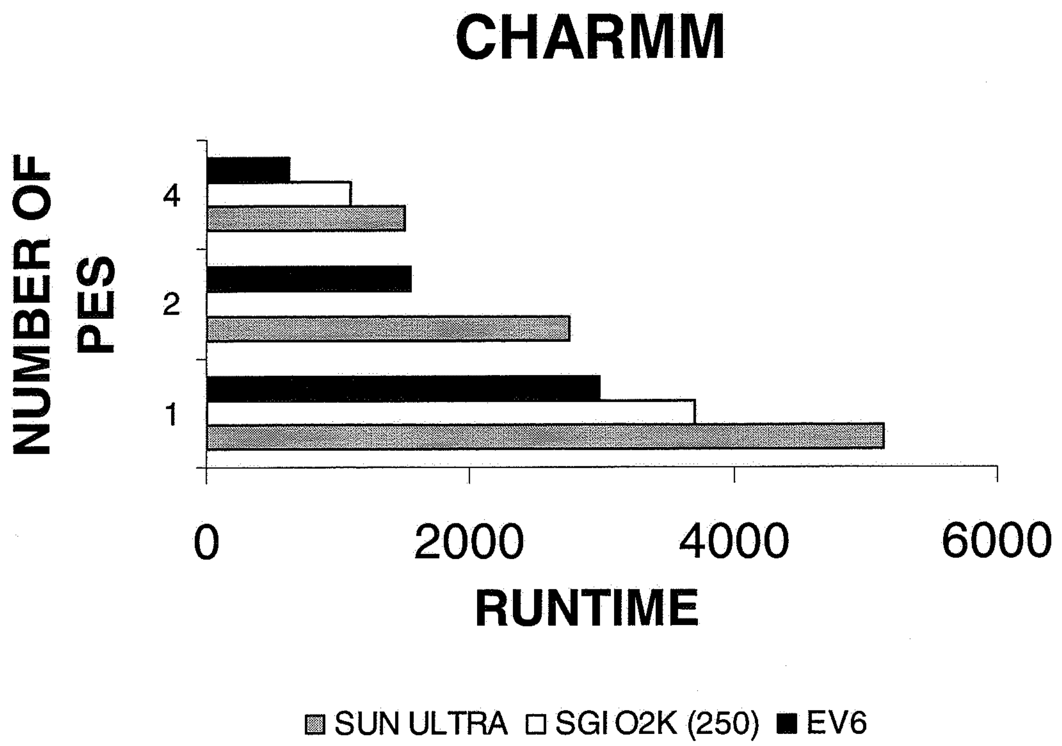


30
20
10
0
Gops
128
NPES
16
Power3-IS
ES40-IS
Power3-EP
ES40-EP
Power3-CG
ES40-CG
Power3-MG
ES40-MG
Power3-FT
ES40-FT
Power3-SP
ES40-SP
Power3-LU
ES40-LU
Power3-BT
ES40-BT
NAS PARALLEL BENCHMARKS(CLASS C) AlphaServer SC-EV67 & IBM Power3


FLUENT
NUMBER OF PES
4
2
1
0
50
100
150
200
250
300
RATING(Jobs/Day)
SGI R12k
IBM SP Power3
EV6
EV67

QCD


QCD
NUMBER OF PES
4
1
0
500
1000
1500
2000
MFLOPS
450MhzP2
T3E-900
SGI O2K(195)
IBM SP Power3
EV67


CHARMM
NUMBER OF PES
4
2
1
0
2000
4000
6000
RUNTIME
SUN ULTRA
SGI O2K (250)
EV6

# Performance Evaluation of the IBM SP and the Compaq AlphaServer SC

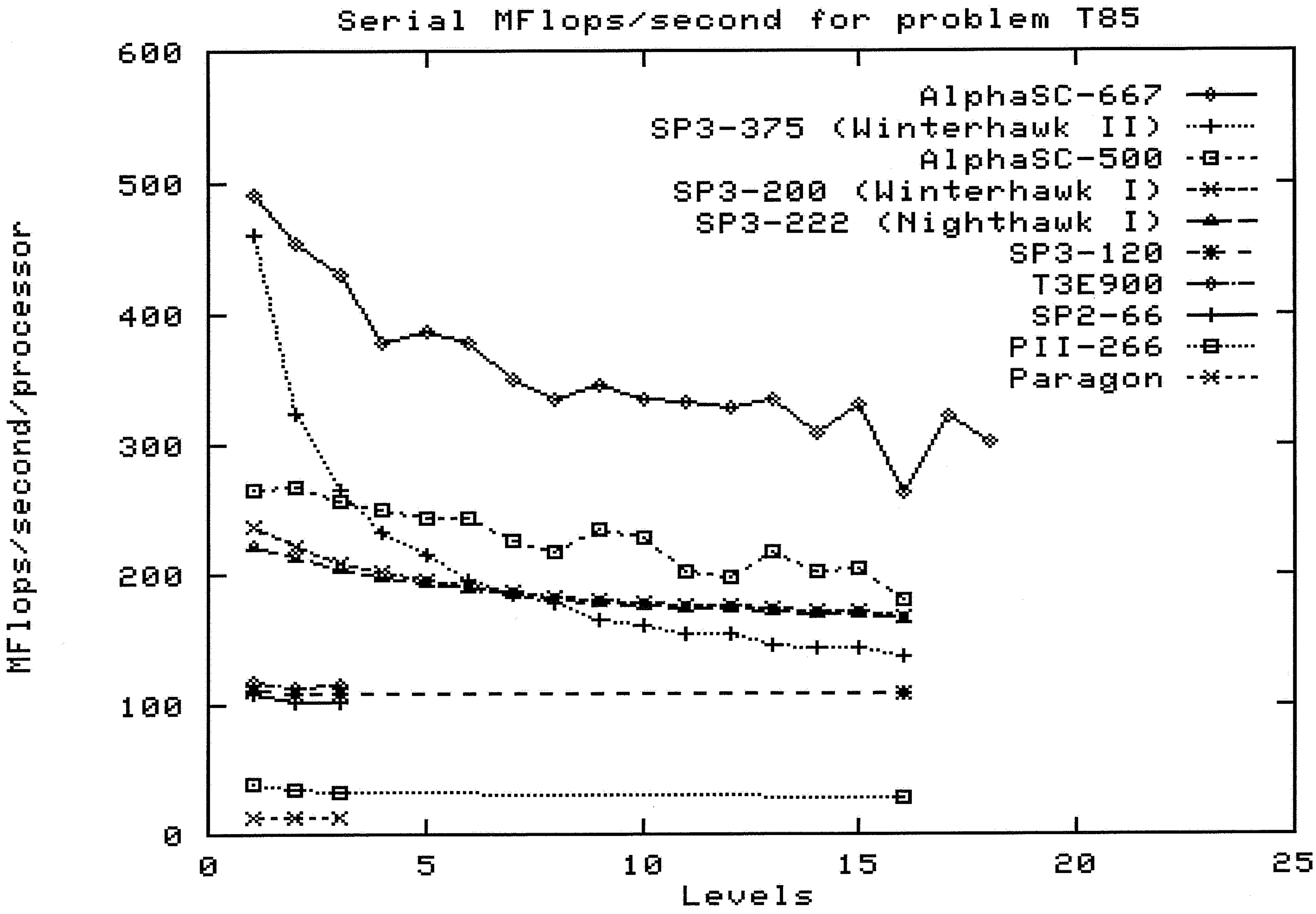
Patrick H. Worley
Computer Science and Mathematics Division
Oak Ridge National Laboratory

ACM International Conference on Supercomputing 2000
May 10, 2000

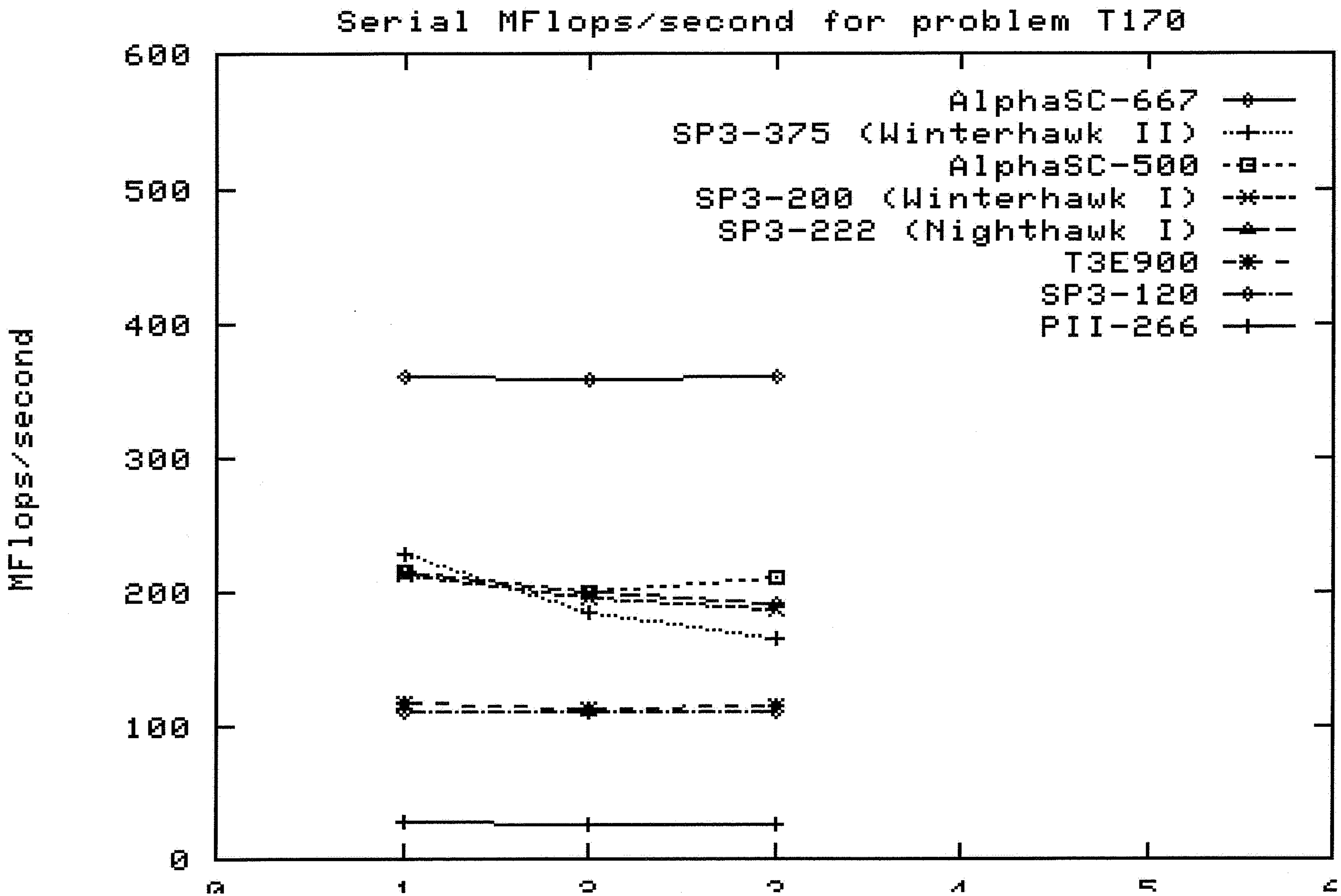
# Spectral Dynamics

- PSTSWM
  - solves the nonlinear shallow water equations on a sphere using the spectral transform method
  - accessing memory linearly, but not much reuse
  - (longitude, vertical, latitude) array index ordering
    - computation independent between horizontal layers (fixed vertical index)
    - as vertical dimension size increases, demands on memory increase

# PSTSWM – From PHW Web Site

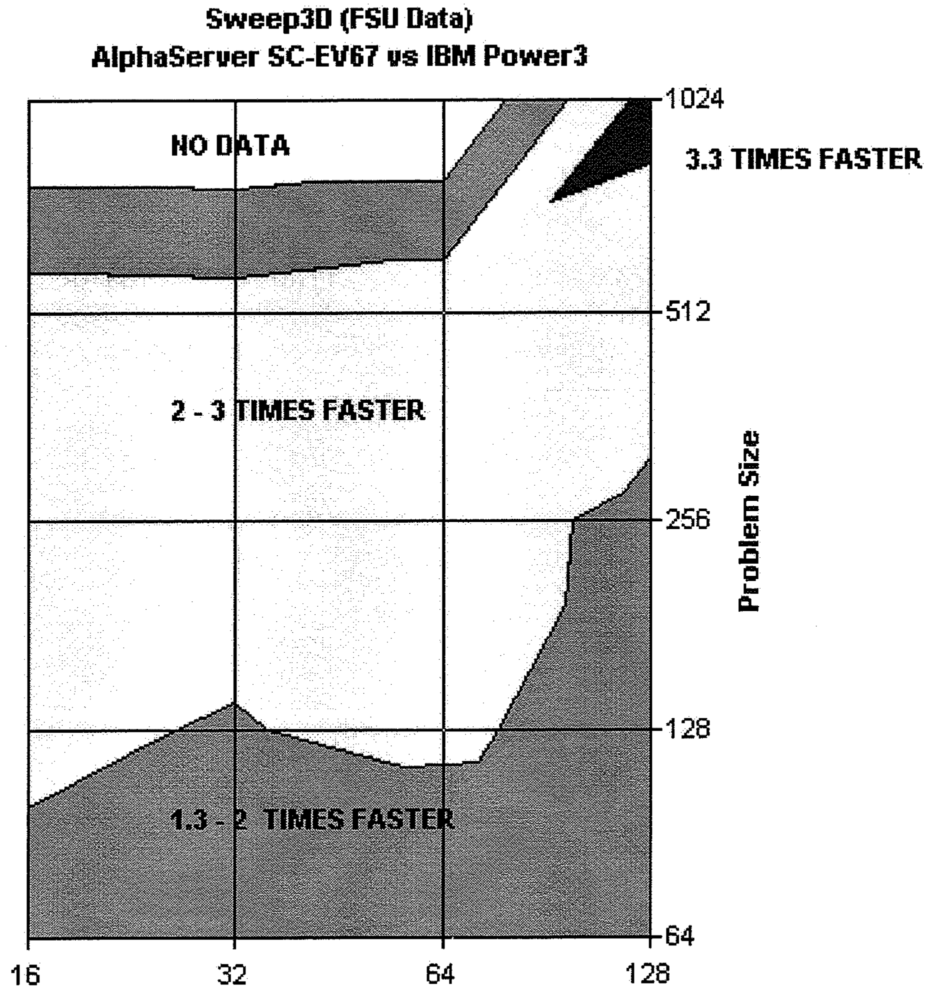


# PSTSWM - From PHW Web Site



# Spectral Dynamics

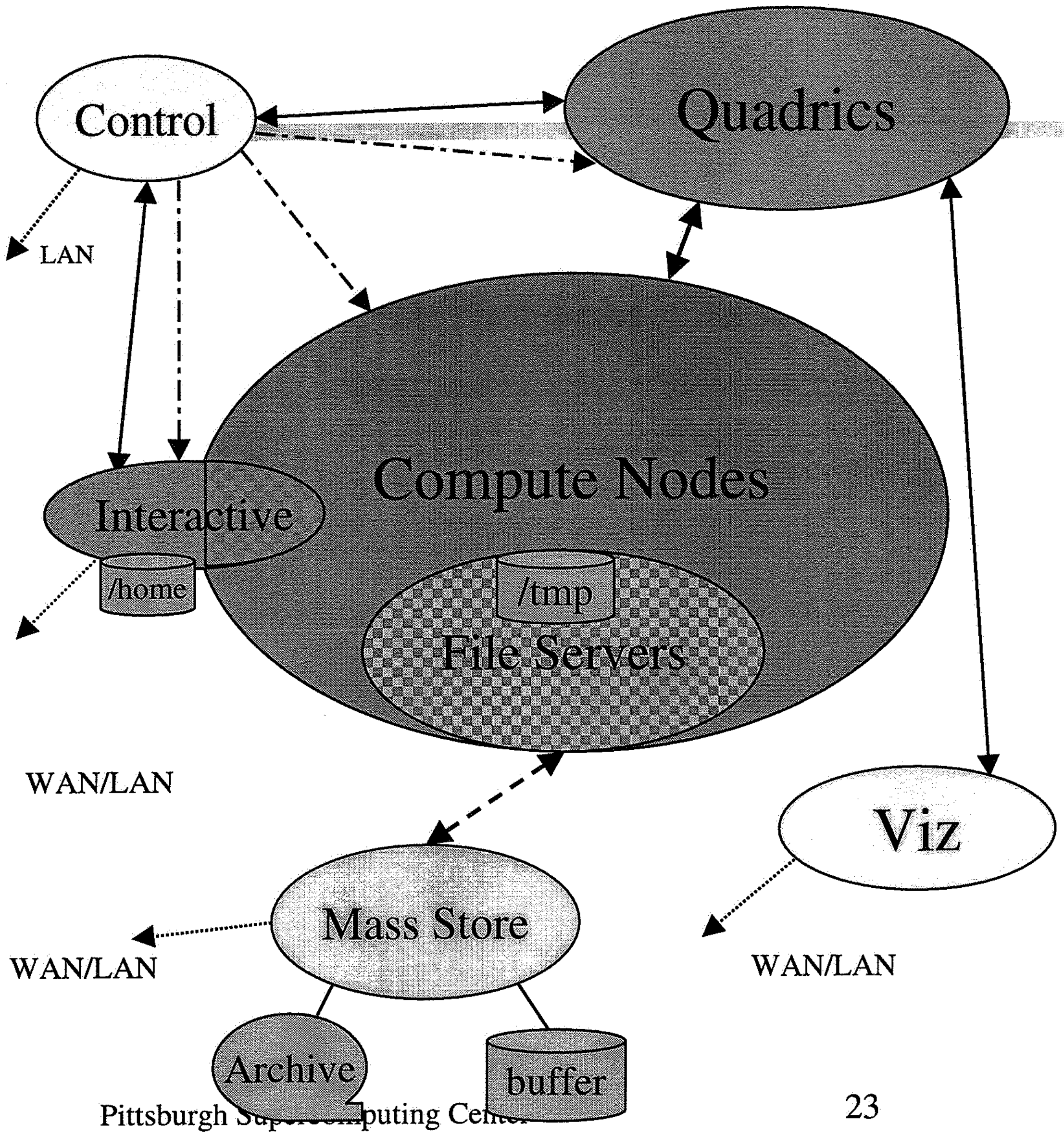
- Summary
  - ➤ Performance of both the IBM and Compaq systems is significantly improved over that of previous generations of the same architectures.
  - ➤ Node memory bandwidth is important for this kernel code.
  - ➤ The Compaq system performance is better than that of the IBM system.


Sweep3D (FSU Data)
AlphaServer SC-EV67 vs IBM Power3
NO DATA
3.3 TIMES FASTER
2 - 3 TIMES FASTER
1.3 - 2 TIMES FASTER
1024
512
256
128
64
Problem Size
16
32
64
128

# For chemistry benchmarks

- See Martyn Guests' recent work http://www.dl.ac.uk/CFS/benchmarks/compchem.html

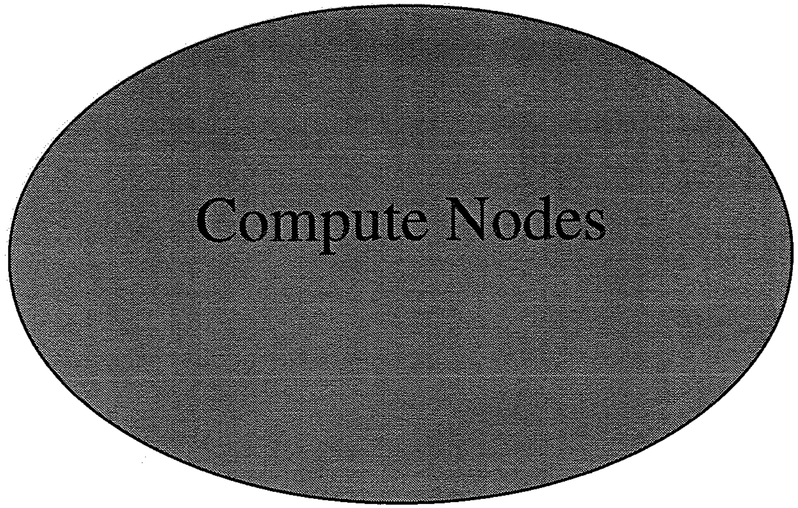
# Terascale Computing System




Terascale Computing System

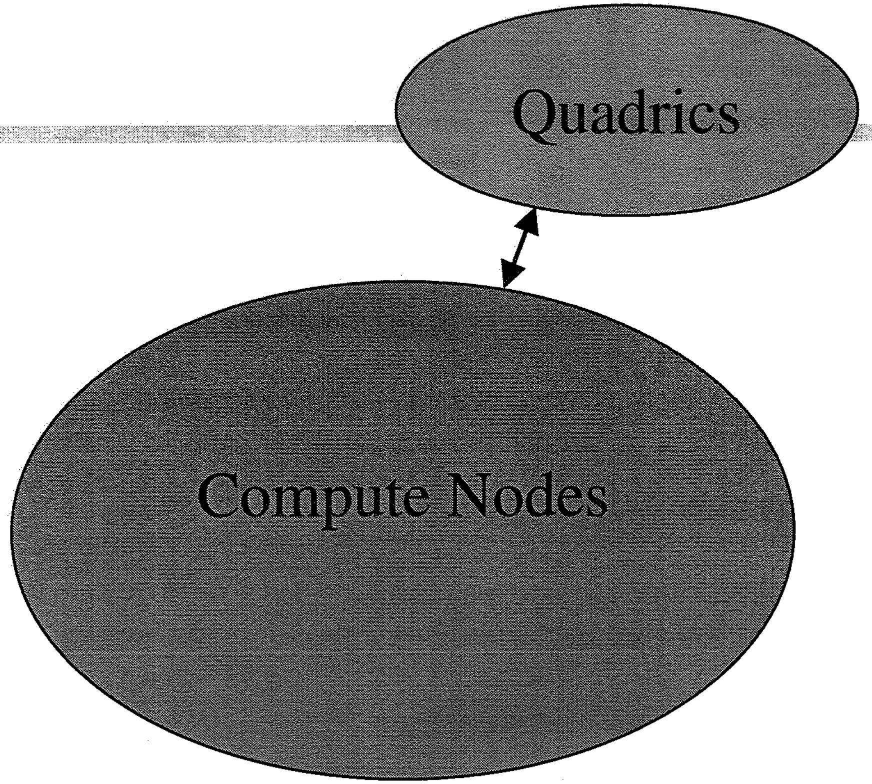
# Terascale Computing System

## Compute Nodes




- Next generation alpha
>2 Gf/processor peak
- 4 processors/node
for bandwidth reasons,
(also price)
- 4 GB memory [2.7 TB]
- 36 GB local disk [25 TB]
- Tru64 Unix

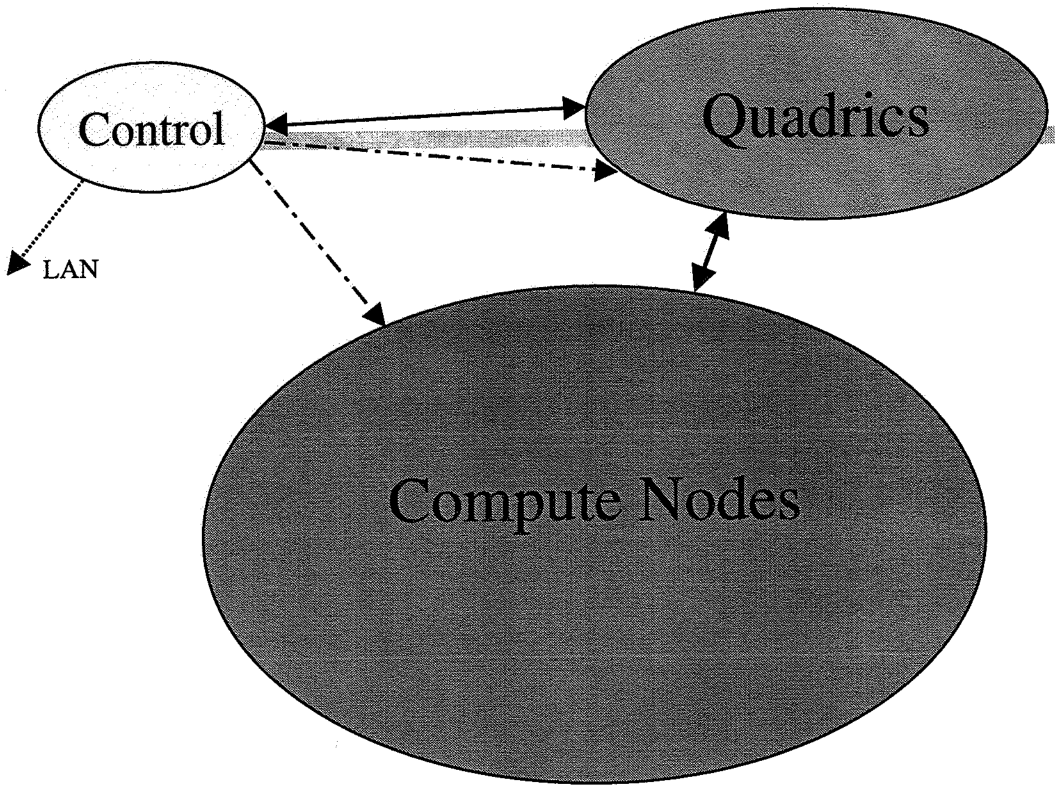
# Terascale Computing System




## **Quadrics Network**

- *Full* "fat-tree"
- Multiple "rails", each sustaining 200MB/sec each direction
- MPI latency ~5 μs
- Fault tolerant- multiple routes multiple rails

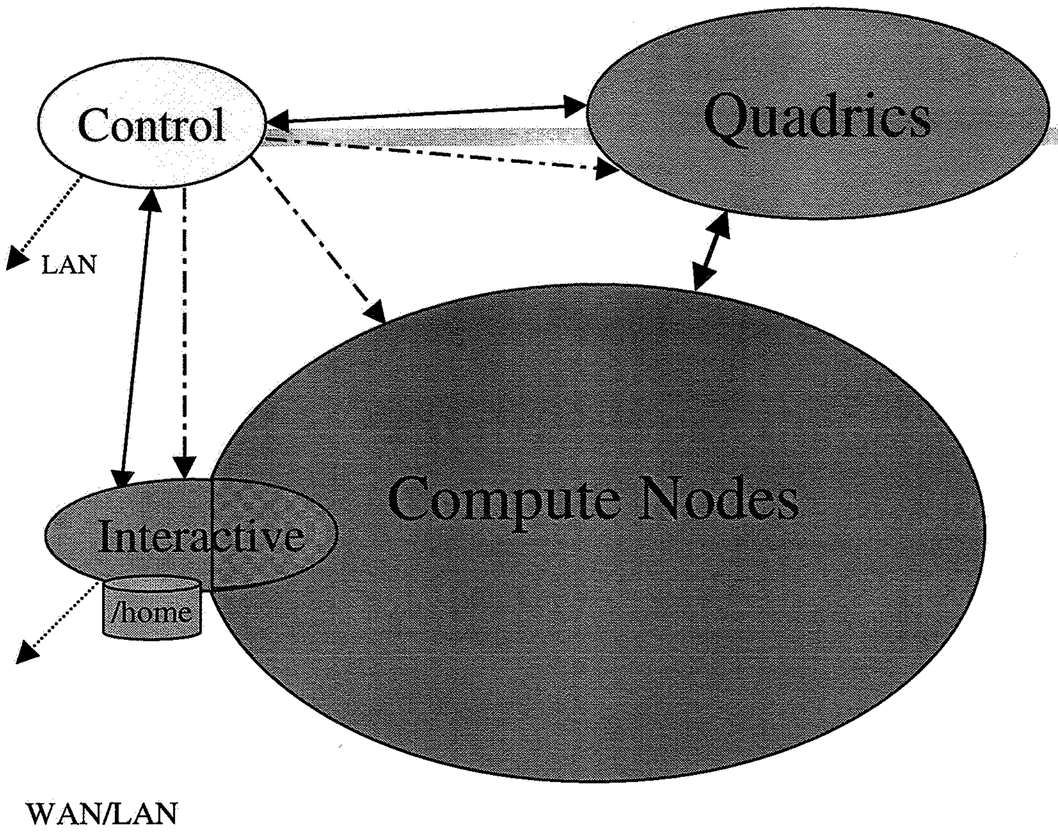
# Terascale Computing System




## Control Nodes

- Node monitoring & control
- 2 for redundancy, with own network
- also Quadrics connected
- RMS database

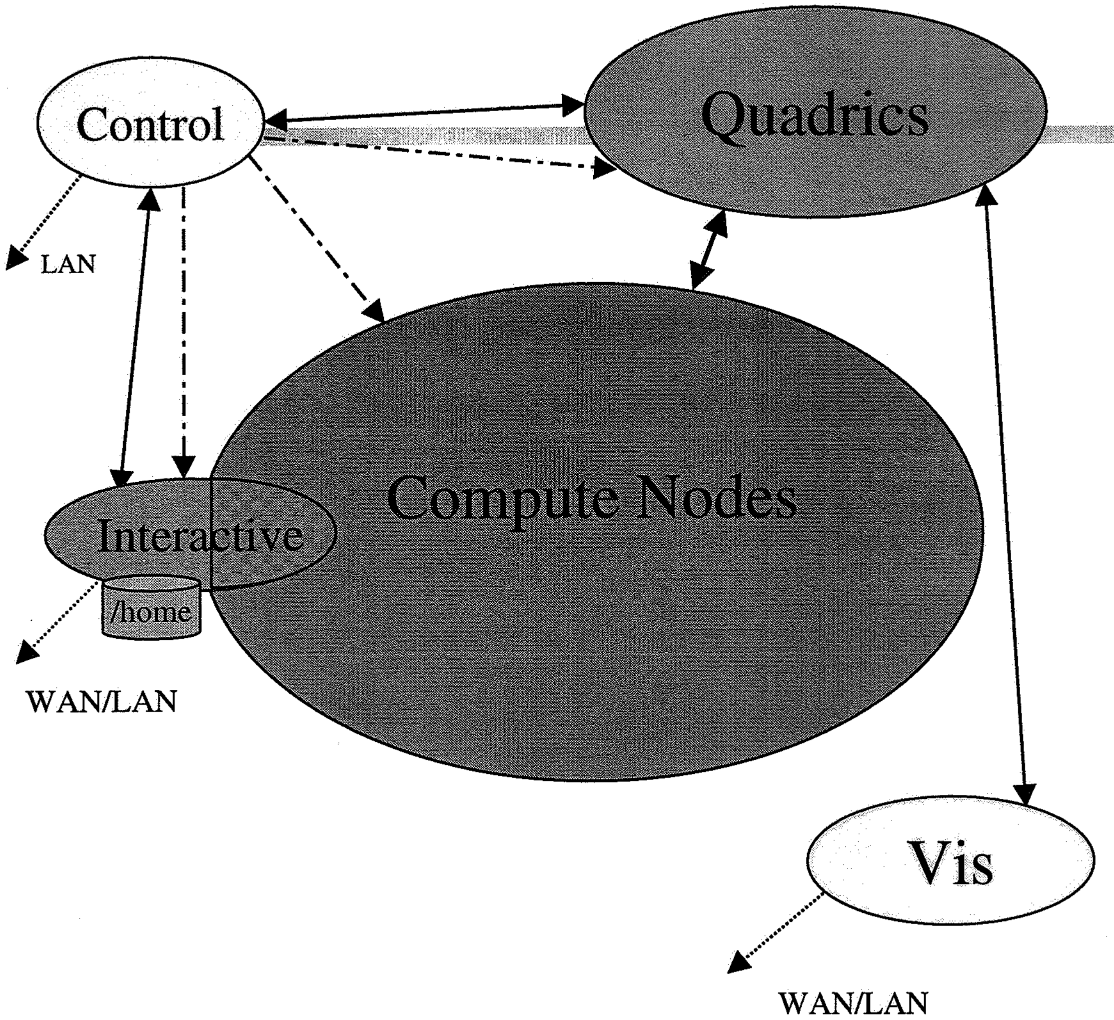
# Terascale Computing System




## Interactive Nodes

- 2 Dedicated single processor nodes, and up to 8 on the compute nodes
- User access
- Gigabit Ethernet
- /home

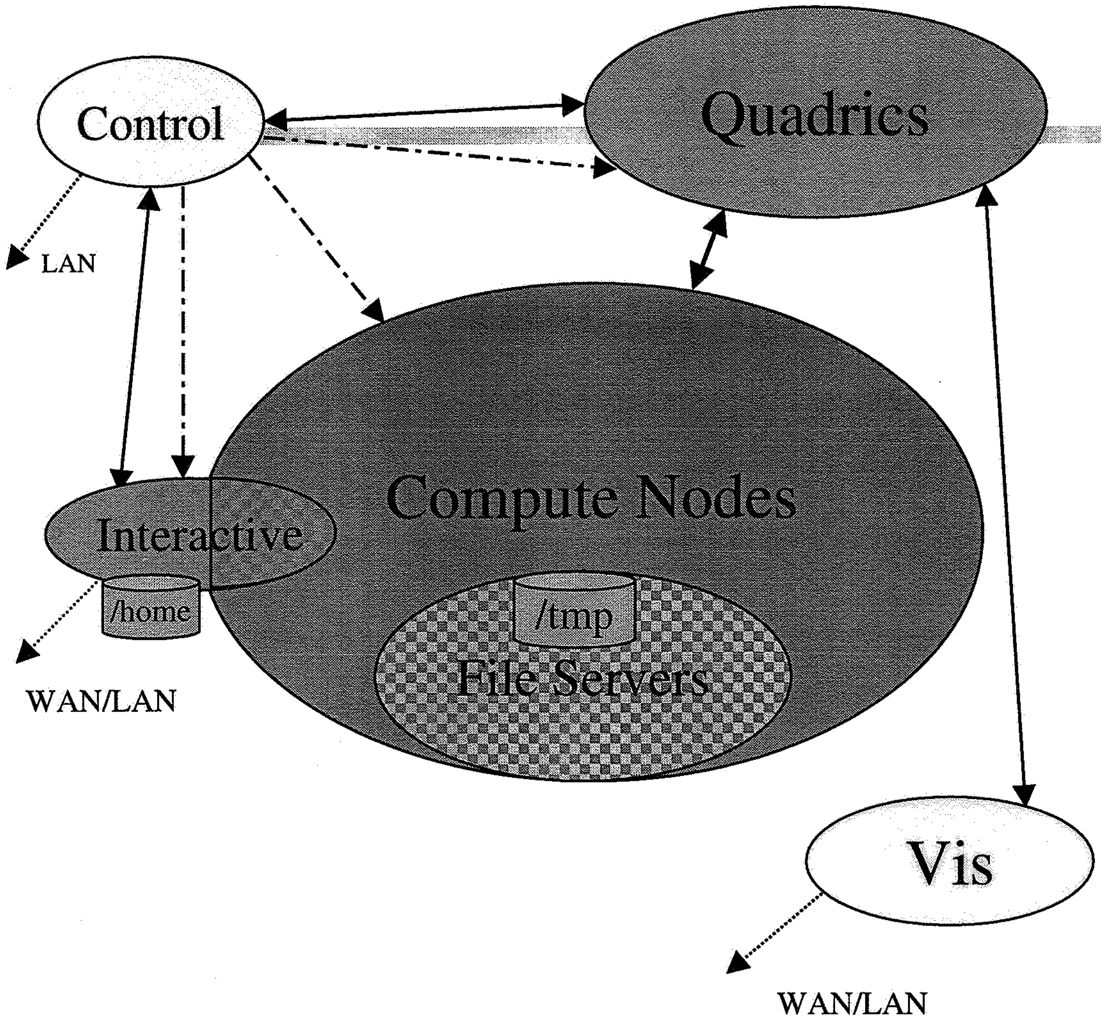
# Terascale Computing System




## Visualization

- Intel/Linux
- ~8 nodes (initially)
- Parallel rendering
- HW/SW compositing
- Quadrics connected
- Image output

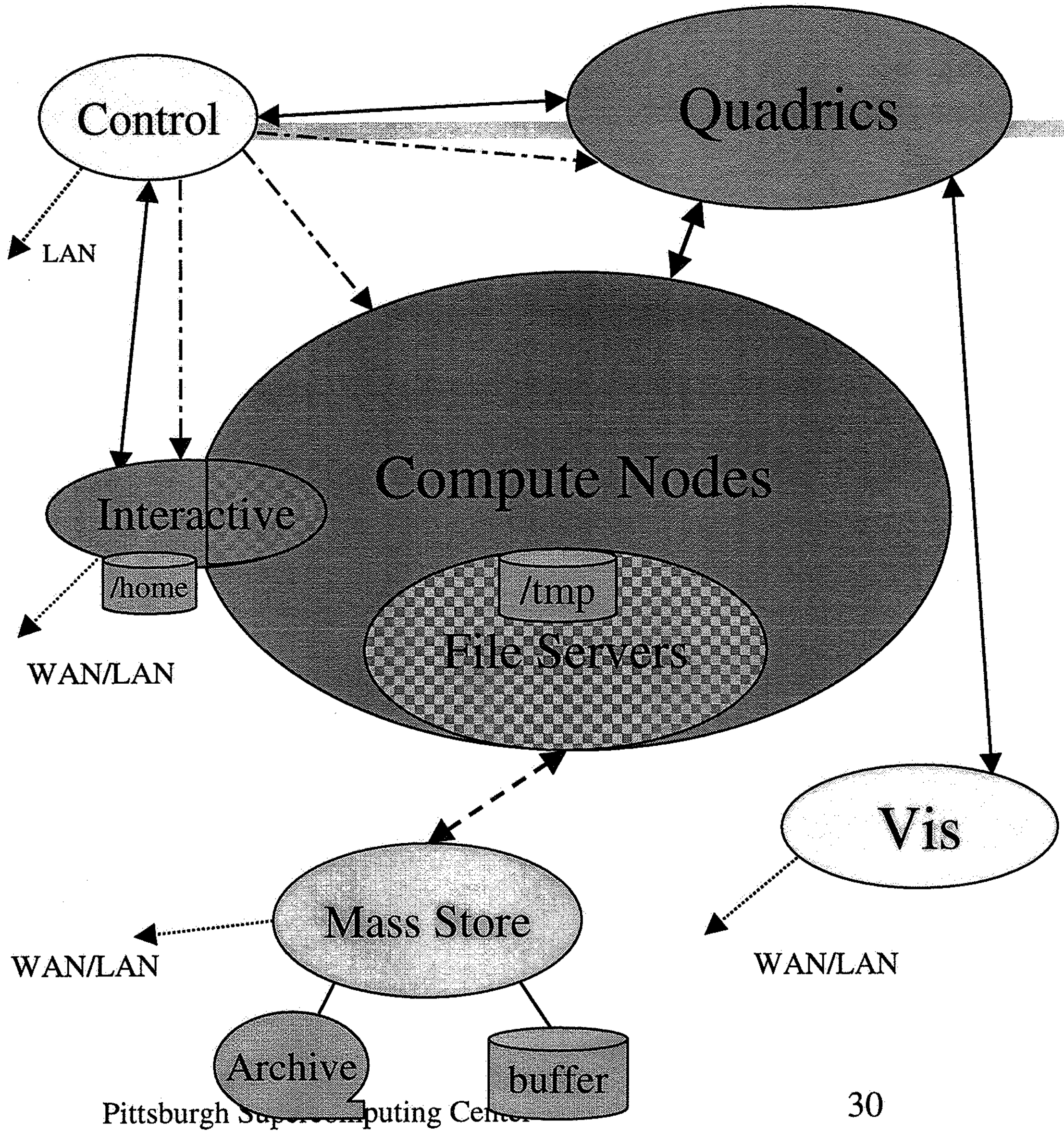
# Terascale Computing System




## File Servers

- 30, on compute nodes
- Network allows memory dump in ~ 3 minutes
- 0.9 TB/server [27 TB]
- RAID
- ~600 MB/s [18 GB/s]
- /tmp

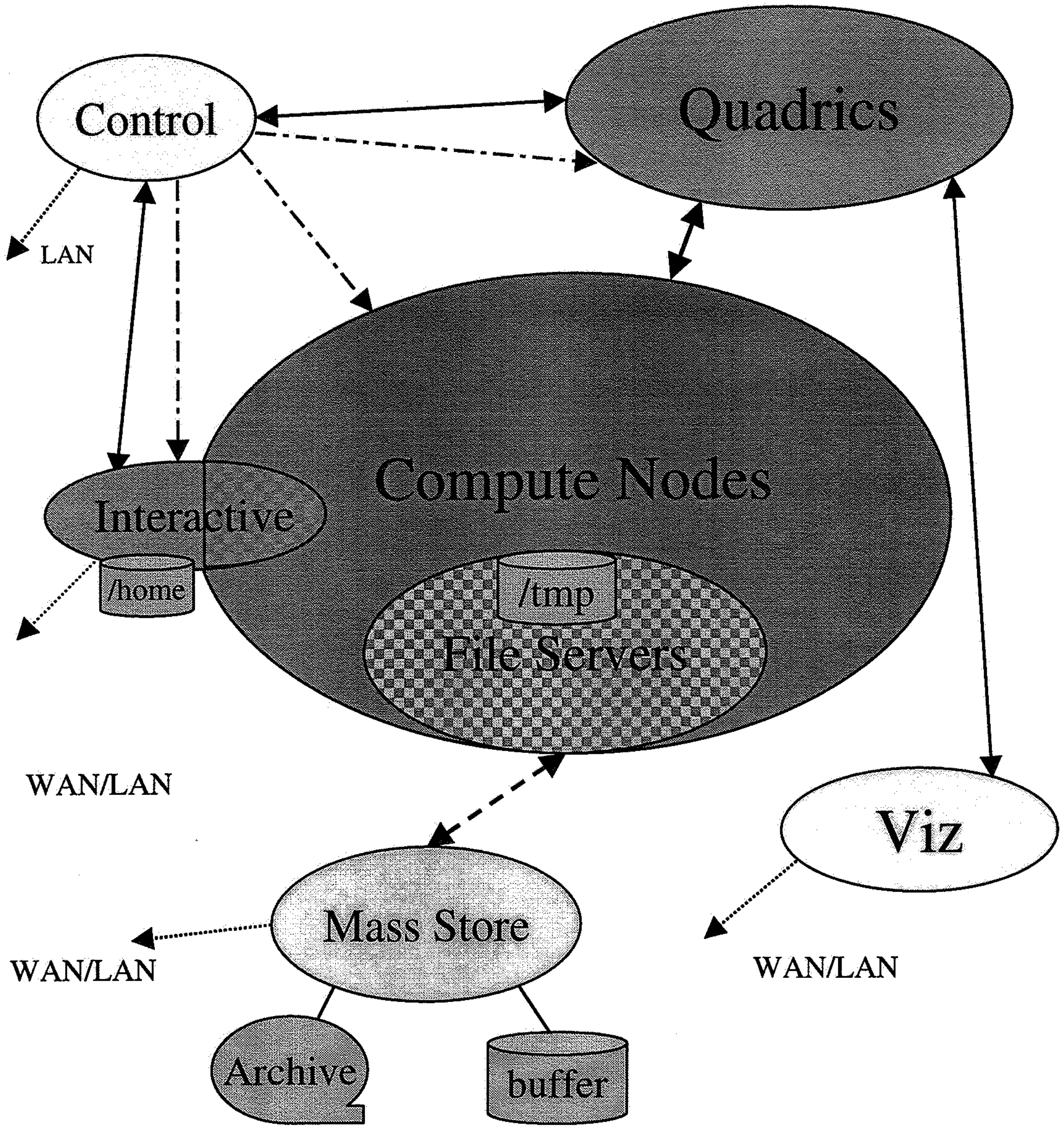
# Terascale Computing System




## Mass Storage

- > 300 TB *Nearline*
- Hippi coupled
- > 1 TB buffer
- ~ 1 TB/hr to tape
- WAN/LAN accessible

## Terascale Computing System




## Summary

- 682 Compute Nodes
- 2728 Alpha processors
- 6 Tf peak,
- 2.7 TB memory
- 25 TB local disk
- Multi-rail fat-tree network
- Redundant monitor/ctrl
- WAN/LAN accessible
- Parallel visualization
- File servers:
  27TB, 18 GB/s
- Mass store, ~1 TB/hr

# Unprecedented scale, not technology

Versions of all components already working

- Alpha EV6x processors
- Quad-processor servers
- A network interconnect
- Standard Compaq software
- AlphaCluster SC cluster software.

# Connections with grid?

- Will be a node on the Grid in US
- Testbed for grid technology, without some of the complications, because of large number of components

# Redundancy to mitigate component failures

- Redundant nodes at critical points: Login, Control
- Redundant power supplies (& hot-swap) in nodes and switches.
- If node goes down, only jobs using that node are affected, can patch in a hot spare, do not have to reboot system,
- Dynamically reconfigurable.
- Network is fully redundant (multi-rail)
- Strong snap-shot capability (IO) to do production even with low MTTI, even for very large jobs.

# Emphasis on checkpointing

- If MTTI for node is 1 year, MMTI for system is 13 hours
- If can checkpoint in 3 minutes, and do it once an hour, can get lots of work done, at <10% impact.

# How frequently to checkpoint

If M=MTTI

S= time for single checkpoint

N= number of checkpoints in time M

Time lost= $M/N/2 + NS$

Minimize in N, find

$N=\sqrt{(M/2S)}$, Time lost = $\sqrt{(2MS)}$

If M=10 hours, S=3 minutes,

then N=10, Time lost=1 hour.

# Some needed software developments

- Scale applications to thousands of processors
- Scheduling- enhancements to PBS
- Scale Quadrics switch to >256 nodes
- High performance I/O (to be based on MPI-IO syntax)

# Applications

- Molecular biology
  - ➤ protein folding (largest simulation to date done at PSC, 2 months, 1/2 of T3D)
  - ➤ how mechanical proteins unfold when stretched
- Fluids and combustion (design of next generation turbine)
- Cosmology
- QCD
- Materials Science

# Applications

- Astrophysical turbulence (Toomre) wants to dump 56 TB in 3 day run
- Storm modeling- quasi real-time
- Truly real-time
  - Lanier- teleimmersion (users at geographically distributed sites collaborate in real time in a shared, simulated environment as if in same physical room).
  - Kanade- sports from arbitrary perspectives

# Focus on capability computing

- Preference for projects which will exploit unique capabilities, rather than capacity
  - e.g. exploit large memory, or I/O capability
  - dedicate the processors to single job
  - real-time applications

# Schedule

- 256 processor system by November (built on EV67, ES40's)
- Final system to be built up over summer of 2001 with next generation chips and boxes

# Team effort of PSC, Compaq, computer and computational science community