

## ORNL Experience with the Compaq AlphaServer SC

**Trey White**  
**Center for Computational Sciences**  
**whitejbiii@ornl.gov**

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

1

## AlphaServer SC at ORNL

- **Good news**
  - Hardware
  - Performance
- **Not-so-good news**
  - ***“It’s the Software, Stupid”*** SC99 Panel
  - SC system software
  - CCS infrastructure software
- **Summary and outlook**

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

2

## AlphaServer SC hardware

- **High-end cluster**
- **64 nodes (128 max)**
- **4 processors per node**
  - 5.2 GB/s aggregate memory bandwidth
- **Alpha EV67 processors**
  - 667 MHz
  - 8 MB L2 cache
- **Quadrics interconnect**

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

3

## Quadrics interconnect

- **64-bit 33 MHz PCI**
- **One-sided virtual-memory operations**
- **Fat tree**
- **Up to 128 ports**
- **Built-in hardware reductions**
  - Only for subtrees (consecutive nodes)

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

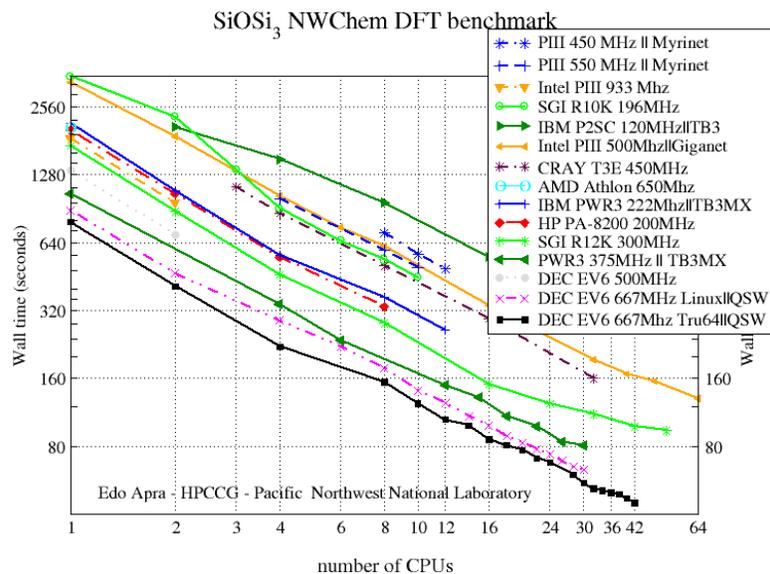
4

## AlphaServer SC performance

- **Great node performance**
  - Fast chip, big cache, fat pipe to memory
- **Great interconnect performance**
  - 200 MB/s MPI ping-pong
  - 6  $\mu$ s latency on MPI send
- **Easy performance**
  - “-o” pretty good
  - Not much tuning needed

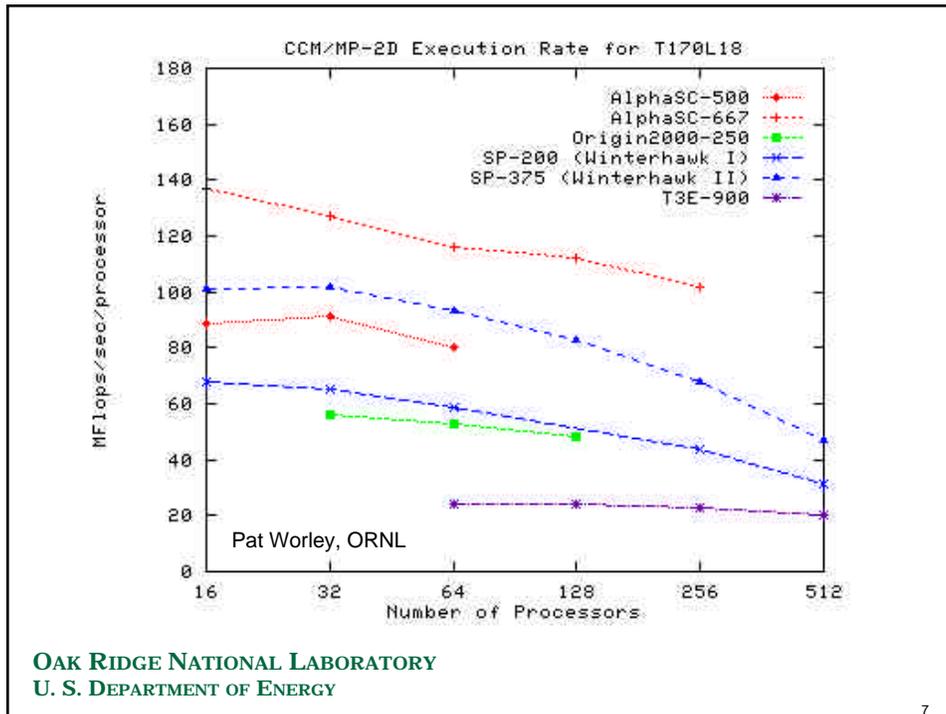
OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

5



OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

6



## AlphaServer SC system software

- **Operating system?**
  - Tru64 5.0
- **Parallel-job scheduler?**
  - Resource Management System (RMS)
- **System-wide file system?**
  - Cluster File System (CFS)
- **Fast, temporary file system?**
  - Parallel File System (PFS)

## Tru64 5.0

- **Marketing**
  - Stable, industrial, production-ready OS
- **Reality**
  - 5.0 pushed out to meet deadlines
  - “Real” release was later, 5.0A
  - SC based on 5.0 to meet deadlines
  - 5.0A so different that it isn’t supported
  - Nodes crash

## RMS

- **Marketing**
  - Job scheduler, resource manager, configuration and accounting database
- **Reality**
  - Not enough to replace a batch system
  - Too much to integrate easily with existing batch systems (like PBS)

## CFS marketing

- **Single cluster-wide file system**
- **Includes root file system**
- **Provides single-system view of cluster**
- **Single network alias represents whole cluster**

## CFS reality

- **Only scales to 32 nodes**
  - Based on old TruCluster software
  - Not originally designed for massive parallelism
  - 64 nodes = 2 CFS clusters
- **Worst of both worlds?**
  - No tools to keep clusters synchronized
  - Cluster alias messes up Kerberos, license managers, etc.
  - Must cross-mount user CFS files through NFS

## PFS

- **Marketing**
  - Fast, striped file system
  - Built on CFS
- **Reality**
  - Striped across CFS servers, all in same cluster
    - Spans one CFS cluster, not whole SC
  - No memory-mapping
    - No running executable files from PFS
  - Unstable

## Cluster? Clusters?

- **1 AlphaServer SC**
- **64 AlphaServer nodes**
- **1 RMS domain**
  - 1 parallel job can use all 64 nodes
- **2 CFS (and/or PFS) domains**
  - 2 cluster aliases, 1 per CFS domain
  - 32 nodes per CFS domain
  - 1 parallel job can span CFS domains
- **RMS and CFS are independent**

## Center for Computational Sciences (CCS)

- **Production system?**
  - IBM SP (184 Winterhawk-II nodes)
- **Account management and authentication?**
  - Distributed Computing Environment (DCE)
- **Center-wide home directories?**
  - Distributed File System (DFS)
- **Archival storage?**
  - High-Performance Storage System (HPSS)

## DCE marketing

- **Centralized user-account management**
  - Cell directory service
  - User info, including passwords
- **Centralized authentication/authorization**
  - Kerberos
  - Tickets/credentials
- **Integrated login**
  - Automatically acquire DCE credentials

## DCE + SC reality

- **Cluster alias is inconsistent with DCE**
  - Multiple interfaces per host OK
  - Multiple hosts per interface?
- **RMS doesn't distribute DCE credentials**
  - Parallel jobs don't have credentials
  - Important for DFS and HPSS

## DFS marketing

- **Center-wide file system**
- **More secure and scalable than NFS**
- **Access-control lists (ACLs)**
  - Like AFS, only more so
- **Uses DCE**
- **Features of NFSv4, available now**
- **CCS home directories**

## DFS + SC reality

- **DFS and CFS don't get along *at all***
- **NFS mount DFS to CFS through external DFS-NFS gateway server (whew!)**
  - Store DCE credentials on gateway server
  - Lose DFS security
  - Must “dfs\_login” every once in a while (messy)
  - Slow (around a MB/s)

## Summary and outlook

- **AlphaServer SC is a high-end cluster**
- **Great performance**
- **Not-so-great system software**
  - May get better in coming releases
- **Differs little with Quadrics+Linux**
- **Differences may increase**
  - Much bigger nodes, beyond Linux scalability
  - Improved proprietary system software
  - Improved integration