Accelerating Research Applications:
Enterprise Storage for HPC Environments

04 May 2007
NAS is best?

- NAS is most applicable architecture, invented for simultaneous access from many clients, easy to use
- *Performance and scalability bottlenecks of “legacy NAS” are not trivial*
- Clustered DAS attractive, *but offers neither scalability of SAN nor ease of NAS*
- Is Clustered DAS cheap? *Not as inexpensive as it sounds when total costs are considered*
- SAN scales well, performs well, *but largely proprietary, very expensive, and requires specific skill set*
- SAN access physically controlled, *not good at sharing data with many clients or across large distances*
Titan 2000 Chassis

Fully redundant power supplies
Titan 2000 Chassis

6 x Gigabit Ethernet ports

Fully redundant power supplies
Titan 2000 Chassis

- Fully redundant power supplies
- Dedicated 10/100 management LAN, DB9 console port
- 6 x Gigabit Ethernet ports

© 2007 BlueArc, Corp. Proprietary and Confidential
Titan 2000 Chassis

- Dedicated 10/100 management LAN, DB9 console port
- 6 x Gigabit Ethernet ports
- Fully redundant power supplies
Titan 2000 Chassis

- Fully redundant power supplies
- 6 x Gigabit Ethernet ports
- Dedicated 10/100 management LAN, DB9 console port
- 4 x 4Gb FC ports
Titan 2000 Chassis

- Dedicated 10/100 management LAN, DB9 console port
- 6 x Gigabit Ethernet ports
- Fully redundant power supplies
- 4 x 4Gb FC ports
- 2 x 10GbE XFP ports, dedicated HCI

© 2007 BlueArc, Corp. Proprietary and Confidential
Titan 2500 Chassis

2 x 10GbE XFP ports
Titan Storage Solution

Hybrid NAS-on-SAN architecture
- Fault tolerant solution with redundant components
- Upgradeable, “future-proof” blade design
- Full 4Gb FC switched SAN fabric back-end
- Hardware-based RAID controllers
- Parallel RAID striping for increased performance
- Global hot spares for high availability
- Three levels of media error detection
- On-the-fly LUN expansion (optionally automated)
- Simultaneous FC and SATA disk options for true multi-tiered storage architectures, lower TCO
- Disk-to-Disk-to-Tape archiving
Titan Storage Solution

Clustered solutions too!

- Active-Active clustering with CNS™
- Automated failover
- Advanced replication options
- Extreme scalability
Documented Performance
The Tolly Group: Independently Verified

Tolly Certified Results

- >717 MB/s random mixed read/writes
- Single Titan performance
- 48 clients, GigE network, no jumbo frames
- 146GB FC disk, 35TB usable
- Throughput invariant with record size

Source: The Tolly Group, January 2006

Figure 1
SC|06 Bandwidth Challenge

Titan Cluster IOPS

Date / Time

TITAN-1 (Node 1)  TITAN-2 (Node 2)  TITAN-CLUSTER (Aggregate)
SC|06 Bandwidth Challenge

Titan Cluster Throughput

MB/Sec

Date / Time


TITAN-1 (Node 1) TITAN-2 (Node 2) TITAN-CLUSTER (Aggregate)
SC|06 Bandwidth Challenge

- 32kB block size
- 44,221 IOPS peak concurrent value
- Max concurrent R/W bandwidth delivered: 1382MB/s

- 2-node Titan 2200 cluster, 12 GbE links
- 3 filesystems per Titan node
- Single namespace (common mount-point)
- 140 146GB 15k FC disks

- 36 compute nodes (144 cores), 8 processes each
- 50% R/W mix; NFSv3, no particular optimizations
**SPECsfs Results: Titan Cluster**

SPECsfs97_R1.v3 results, AA clustered Titan 2200s, single namespace showing Titan cluster response time as a function of delivered I/O performance

<table>
<thead>
<tr>
<th>Throughput (ops/sec)</th>
<th>Response (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20208</td>
<td>0.8</td>
</tr>
<tr>
<td>40214</td>
<td>1.1</td>
</tr>
<tr>
<td>60618</td>
<td>1.4</td>
</tr>
<tr>
<td>80937</td>
<td>1.8</td>
</tr>
<tr>
<td>101499</td>
<td>2.2</td>
</tr>
<tr>
<td>121719</td>
<td>2.7</td>
</tr>
<tr>
<td>142104</td>
<td>3.1</td>
</tr>
<tr>
<td>162759</td>
<td>3.7</td>
</tr>
<tr>
<td>181640</td>
<td>5.2</td>
</tr>
<tr>
<td>195502</td>
<td>5.6</td>
</tr>
</tbody>
</table>

**About 764MB/s**
SpecSFS to BWC comparison

- 4kB block size
- 195,502 IOPS peak concurrent value
- Max concurrent R/W bandwidth delivered: 764MB/s

- 2-node Titan 2200 cluster, 12 GbE links
- 8 filesystems per Titan node
- Single namespace (common mount-point)
- 416 73GB 15k FC disks

- 28 compute nodes (56 cores), 32 processes each
- Spec R/W mix; NFSv3, no particular optimizations
Real World Performance

- 48 hour cyclical workload, single Titan
- Large computational Linux cluster
- High performance
  - Peak throughput over 700MB/sec Reads, 250MB/sec Writes
  - Over 222k aggregate peak IOPs
Multi-Tiered Storage

- Match disk to application and price requirements
- Any combination of FC and SATA
- Immediately benefit from new disk technology and prices
- Disk-to-disk-to-tape eliminates the backup window
- LAN-free data migration

Tape Library
Multi-tiered storage design

- 146GB FC, 15k – max spindles for max IOPS
- 300GB FC, 10k – good for IOPS, bandwidth, many tasks or unknown storage I/O profiles
- 500GB → 750GB → 1TB SATA – cost effective, high-density, decent bandwidth

- 4kB blocks to maximize IOPS
- 32kB blocks for flexible storage I/O needs
- Larger blocks to maximize bandwidth
Cluster Name Space

- Features:
  - Single name space
  - NFS and CIFS Support
  - Dual 10GbE Cluster Interconnect
  - Request Redirection in HW
  - Multi-node Read Caching

- Benefits:
  - Single mount point and file system for simplified user administration
    - Universal CNS Access
    - Unified Directory Structure
  - Load balancing
Shared GNS Cluster Storage Pool

Balanced & efficient storage use

BlueArc Shared Storage Pool

GNS

Root

FS1

FS2

GNS

SAN

Used

55%

Available

45%

Separate unbalanced storage

DAS-like inefficiency

Used

80%

Available

20%

Used

70%

Available

30%
Balanced, efficient, scalable storage

- Titan not particularly optimized for small-file, high-IOPS nor large-file, high-bandwidth profiles, just happens to be good at both
- Design tiers to meet just about any I/O profile
- Add more Titans for higher performance without needing to add capacity
- Add more disk for higher capacity if needed
- True separation-of-function in storage design
Any Questions?