SciDAC PDSI Update (part 1)
CS/VIS PI Meeting, October 23, Germantown, MD

Garth Gibson
Carnegie Mellon University and Panasas Inc.

SciDAC Petascale Data Storage Institute (PDSI)
www.pdsi-scidac.org
w/ LANL (G. Grider), LBNL (W. Kramer), SNL (L. Ward),
ORNL (P. Roth), PNNL (E. Felix),
UCSC (D. Long), U.Mich (P. Honeyman)
Clearing Path thru Petascale to Exascale

- Scaling 100%/yr given disk realities is hard
  - Disk BW @ 20%/yr, IO/s @ 5%/yr
  - Storage problem renews itself each year
• PETASCALE DATA STORAGE INSTITUTE 06-11
  • 3 universities, 5 labs, G. Gibson, CMU, PI
  • Enabling HEC storage to meet SciDAC needs
• SciDAC @ Petascale storage issues
  • Community building: ie. PDSW @ SCxy
  • APIs & standards: ie., Parallel NFS, POSIX
  • Failure data collection, analysis: ie., cfdr.usenix.org
  • Performance trace collection & benchmark publication
  • IT automation applied to HEC systems & problems
  • Novel mechanisms for core (esp. metadata, wide area)
Annual PDSI Sponsored Workshops

HEC FSIO '07

HEC FSIO R&D Workshop/HECERA FSIO PI Meeting '07 AGENDA

Welcome Session
Monday, June 5, 2007
Welcome and Overview of HEC FSIO
Welcome from NSF

Session 1: HECERA

1. Quality of Service Guarantee for Scalable File-System Parallel Storage Systems
2. Performance Modeling for Large-Scale Distributed Storage
3. Open-loop of paper progressed

LANL ISSRM and INFIT
LANL Data Management
Research Session 2:
Interconnection, Understanding, Cache, and Architectures
File System Testing, Replication, Profiling, and Analysis on HEC

4. Memory caching and prefetching
5. Open-loop of paper progressed
6. Research Session 3: HECERA

7. Petascale I/O for High-End Computing

WORKSHOP ABSTRACT
Petascale computing infrastructure makes petascale demands on storage and management. In the last decade it has been shown that parallel file systems can meet these demands, but the underlying challenge is how to design and implement systems that can do so. The key to success lies in the ability to store and access large amounts of data on a system that is not tightly coupled to the compute nodes. The goal of this workshop is to bring together experts in the field to discuss the latest developments in parallel file systems and their applications.

PetaScale Data Storage Workshop Introduction
Garth Gibson

SESSION 1: Scalable Systems


SESSION 2: Scalable Services

Jonah Koenig (presenter), Y. Zhang, Univ. of California, Santa Cruz

SESSION 3: Scalable System Design

Swarni V. Patt (presenter), Garth A. Gibson, Sam Lang, Mike Polte, Carnegie Mellon University

POSTER SESSION 1 - see info below

POSTER SESSION 2 - see info below

POSTER SESSION 3 - see info below

POSTER SESSION 4 - see info below

OPEN MIC AVAILABLE

www.pdsi-scidac.org/SC07/
Failure Data Collection

- Los Alamos root cause logs
  - 22 clusters & 5,000 nodes
  - covers 9 years & continues
  - cfdr.usenix.org publishes this and many other failure datasets

<table>
<thead>
<tr>
<th>HW</th>
<th>ID</th>
<th>Nodes</th>
<th>Procs</th>
<th>Procs/node</th>
<th>Production Time</th>
<th>Mem (GB)</th>
<th>NICs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>N/A</td>
<td>12/99</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>32</td>
<td>32</td>
<td>N/A - 12/03</td>
<td></td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>N/A - 04/03</td>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>164</td>
<td>328</td>
<td>04/01 - now</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>256</td>
<td>512</td>
<td>12/01 - now</td>
<td></td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>128</td>
<td>1024</td>
<td>09/01 - 01/02</td>
<td></td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>7</td>
<td>1024</td>
<td>4096</td>
<td>05/02 - now</td>
<td></td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1024</td>
<td>4096</td>
<td>05/02 - now</td>
<td></td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>128</td>
<td>512</td>
<td>05/02 - now</td>
<td></td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>128</td>
<td>512</td>
<td>05/02 - now</td>
<td></td>
<td>352</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>128</td>
<td>512</td>
<td>10/02 - now</td>
<td></td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>32</td>
<td>128</td>
<td>10/02 - now</td>
<td></td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>128</td>
<td>256</td>
<td>10/02 - now</td>
<td></td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>256</td>
<td>512</td>
<td>10/02 - now</td>
<td></td>
<td>352</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>256</td>
<td>512</td>
<td>10/02 - now</td>
<td></td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>256</td>
<td>512</td>
<td>10/02 - now</td>
<td></td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>256</td>
<td>512</td>
<td>10/02 - now</td>
<td></td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>512</td>
<td>1024</td>
<td>10/02 - now</td>
<td></td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>19</td>
<td>16</td>
<td>2048</td>
<td>03/05 - 06/05</td>
<td></td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>49</td>
<td>6152</td>
<td>12/06 - 09/02</td>
<td></td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>5</td>
<td>544</td>
<td>12/06 - 09/02</td>
<td></td>
<td>64</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>1</td>
<td>256</td>
<td>11/04 - now</td>
<td></td>
<td>1024</td>
<td>0</td>
</tr>
</tbody>
</table>
Projections: More Failures

- Con’t top500.org 2X annually
  - 1 PF Roadrunner in 2008 (May 26)
- Cycle time flat, but more of them
  - Moore’s law: 2X cores/chip in 18 mos
- # sockets, 1/MTTI = failure rate up 25%-50% per year
  - Optimistic 0.1 failures per year per socket (vs. historic 0.25)
Checkpointing Failure Tolerance in Trouble

- Periodic \( (p) \) pause to checkpoint \( (t) \)
  - On failure, roll back & restart
- Balanced systems
  - disk speed tracks FLOPS & mem size, so checkpoint capture \( (t) \) is constant time
  - \( 1 - \text{AppUtilization} = \frac{t}{p} + \frac{p}{2 \times \text{MTTI}} \)
  - \( p^2 = 2 \times t \times \text{MTTI} \)

- but dropping MTTI kills app utilization!
Bolster HEC Fault Tolerance

- More storage bandwidth
  - disk speed 1.2X/yr
    - # disks +67%/y
      just for balance!
  - to also counter MTTI
    - # disks +130%/yr!
  - poor appetite for the cost

- Compress checkpoints
  - plenty of cycles available
  - smaller fraction of memory each year
    - 25-50% smaller / yr
Different Approaches

- Dedicated checkpoint device (ie., PSC Zest)
  - Stage checkpoint through fast memory
  - Cost of dedicated memory large fraction of total
  - Cheaper memory (flash?) now bandwidth limited

- Classic enterprise process pairs duplication
  - Flat 50% efficiency cost, plus message duplication
Tools for Understanding IO in Apps

**NEWEST TRACE DATA, REDSTORM, SANDIA NAT’L LAB**
- A physics simulation problem for a common Sandia application, Alegra
  - Runs were performed alongside regular user runs
  - Each run generated 4 restart dumps, and ran for 20 simulation cycles
- Both single core per node, and 2 core (virtual node mode) per node
  - Repeated with and without tracing enabled
- The single core per node jobs ran at a client size of 2744 processes
  - Non-tracing elapsed run time 10:42 minutes
  - Tracing elapsed run time 11:07 minutes
- The 2 core per node jobs ran at 2916 nodes, 5832 processes.
  - Non-tracing elapsed run time 15:52 minutes
  - Tracing elapsed run time 16:37 minutes
- Raw trace file sizes 30K-50K per MPI rank, except rank zero (600KB-700KB)
  - Rank 0 I/O to terminal records progress in the job.

sourceforge.net/projects/libsysio
**pNFS: Scalable NFS Standard & Code Soon**

- Open source & competitive offerings!
- NetApp, Sun, IBM, EMC, Panasas ....

---

**From: Tigiran Mkrtchyan <tigiran.mkrtchyan@desy.de>**
Date: July 16, 2008 4:18:13 AM PDT
To: pnfs@linux–nfs.org
Subject: [pnfs] pnfs becomes real!

today we ran the first real physics analysis job using dCache-pnfs server and linux pnfs client:

tigiran@nairi:~/work/linux-pnfs> git show | head -5
commit 6ae52464ba2c77f1bf2365e415305fd9b51dd20
Author: Benny Haley <bhaleyv@panasas.com>
Date: Tue Jul 15 20:22:51 2008 +0300

Anyway, fist time we can show that NFSv4.1 is something real (and not my hobby only).

---

**From: Spencer Shepler <Spencer.Shepler@Sun.COM>**
Date: August 1, 2008 4:34:46 PM GMT–04:00

2. IETF status

All of the current working group internet drafts are moving forward for publication. This means that they have submitted to the area director and will start their way through the process (IETF last call and IESG review).
Spyglass design

- Partition file system hierarchy by subtree
  - Each subtree is an independent subindex
- Summarize contents of each subindex
  - Quickly rule out entire subindexes that can’t satisfy the query
- Log incremental changes
  - Rebuild index when there are “enough” changes
- Integrity is much easier
  - Rebuild subindex, not entire index
Metadata: Huge Directories (PVFS, FUSE)

- Eliminate serialization
- All servers grow directory independently, in parallel, without any co-ordinator

Local representation of huge directory in Giga
- No synchronization & consistency bottlenecks
  - Servers only keep local “view”, no shared state
SciDAC PDSI Update (part 1)
CS/VIS PI Meeting, October 23, Germantown, MD

Garth Gibson
Carnegie Mellon University and Panasas Inc.

SciDAC Petascale Data Storage Institute (PDSI)
www.pdsi-scidac.org
w/ LANL (G. Grider), LBNL (W. Kramer), SNL (L. Ward), ORNL (P. Roth), PNNL (E. Felix), UCSC (D. Long), U.Mich (P. Honeyman)