S²-RAID: A New RAID Architecture for Fast Data Recovery

Jiguang Wan*, Jibin Wang*, Qing Yang+, and Changsheng Xie*

*Huazhong University of Science and Technology, China
+University of Rhode Island, USA
Overview

- A reconstruction solution - $S^2$-RAID
  - Using parallel data layout to boost data construction

- Online reconstruction performance
  - Average user response time
  - Shorten reconstruction time by a factor of 3~6
    - Comparing with the traditional RAID
Outline

- Reconstruction background
- Data layout strategy
- $S^2$-RAID prototype
- Evaluation results
- Performance analyse
Background

- High-capacity disk keep increasing.
- Offline reconstruction is result in service down time.
- Existing reconstruction solutions
  - Long reconstruction time and Average user response time
S²-RAID Idea

- **Our goals**
  - Reducing construction time sharply
  - Maximizing Parallel reconstruction
  - Minimizing the impact on front end performance.

- **S²-RAID data layout**
  - Parallel reconstruction model
  - Using “subRAID” concept
  - Each subRAID uses standard RAID
Traditional RAID 5 reconstruction

- Disk 0
- Disk 1
- Failed Disk 2
- Spare disk

Single reconstruction stream long reconstruction time
$S^2$-RAID data layout

subRAID0

Standard RAID 5

partition
S$^2$-RAID data layout structure

Group 0

\[
\begin{array}{ccc}
0 & 1 & 2 \\
3 & 4 & 5 \\
6 & 7 & 8 \\
\end{array}
\]

Init. matrix $M_0$

Group 1

\[
\begin{array}{ccc}
0 & 1 & 2 \\
5 & 3 & 4 \\
7 & 8 & 6 \\
\end{array}
\]

$M_1$

Group 2

\[
\begin{array}{ccc}
0 & 1 & 2 \\
4 & 5 & 3 \\
8 & 6 & 7 \\
\end{array}
\]

$M_2$
**S²-RAID data layout structure**

\[ P_{ij} \text{: subRAID numbers of the } (j+1)\text{th partition on disks of } (i+1)\text{th group in the RAID} \]

\[ K \text{: the partition number of the disk} \]

\[
\begin{align*}
m_0 &= \begin{pmatrix} P_{0.0} \\ P_{0.1} \\ \vdots \\ P_{0.K-1} \end{pmatrix} & m_1 &= \begin{pmatrix} P_{1.0} \\ P_{1.1} \\ \vdots \\ P_{1.K-1} \end{pmatrix} = \begin{pmatrix} SH_r^0(P_{0.0}) \\ SH_r^1(P_{0.1}) \\ \vdots \\ SH_r^{K-1}(P_{0.K-1}) \end{pmatrix} \\
m_i &= \begin{pmatrix} P_{i.0} \\ P_{i.1} \\ \vdots \\ P_{i.K-1} \end{pmatrix} = \begin{pmatrix} SH_r^0(P_{i-1.0}) \\ SH_r^1(P_{i-1.1}) \\ \vdots \\ SH_r^{K-1}(P_{i-1.K-1}) \end{pmatrix}
\end{align*}
\]

**Note:** the size of the group must be a prime number
S²-RAID data layout

- S²-RAID structure
  - 9 disks
  - 9 subRAID0s
  - RAID type
  - RAID 5, RAID 10, RAID 6 etc.
## $S^2$-RAID 5 reconstruction

<table>
<thead>
<tr>
<th>D0</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>D6</th>
<th>D7</th>
<th>D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>2.0</td>
<td>0.1</td>
<td>1.1</td>
<td>2.1</td>
<td>0.2</td>
<td>1.2</td>
<td>2.2</td>
</tr>
<tr>
<td>3.0</td>
<td>4.0</td>
<td>5.0</td>
<td>5.1</td>
<td>3.1</td>
<td>4.1</td>
<td>4.2</td>
<td>5.2</td>
<td>3.2</td>
</tr>
<tr>
<td>6.0</td>
<td>7.0</td>
<td>8.0</td>
<td>7.1</td>
<td>8.1</td>
<td>6.1</td>
<td>8.2</td>
<td>6.2</td>
<td>7.2</td>
</tr>
</tbody>
</table>

D4 was divided into 3 partitions
Reconstruction speed!
No bottleneck in reconstruction
No operation conflict (write or read)
$S^2$-RAID 5 reconstruction

Free disk can serve front end and **Spare disk**
S²-RAID prototype structure

- S²-RAID prototype based on MD, are using the open source
- The *iSCSI target* module modifies the IET SCSI command handling and disk IO parts.
- The *Config* module provides RAID setup and configuration functions using mdadm commands to realize different S²-RAID subRAID functions.
- The *S²-RAID* module realizes the basic functions of RAID10 and RAID5 including RAID rebuilder based on MD.
## Experimental Setup

- **Hardware of server and client**
- **Evaluation tools of the storage server and client**

<table>
<thead>
<tr>
<th>OS</th>
<th>Fedora Core 8.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>blktrace</td>
<td>blktrace 1.6</td>
</tr>
<tr>
<td>postmark</td>
<td>postmark 1.5</td>
</tr>
<tr>
<td>Disks</td>
<td>1 Seagate ST3160023AS, 160GB, 7200RPM.</td>
</tr>
<tr>
<td></td>
<td>12 Seagate ST3500320AS, 500GB, 7200RPM.</td>
</tr>
<tr>
<td>postgresql</td>
<td>postgresql 8.1.19</td>
</tr>
<tr>
<td>TPC-C</td>
<td>TPC-Cbeta 1.9.11</td>
</tr>
<tr>
<td>Mainboard</td>
<td>SUPER X7DVLJ</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel(R) Xeon(R) CPU 5110 @ 1.60GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>1GB DDR2</td>
</tr>
<tr>
<td>HBA</td>
<td>Highpoint 2240 RAID,</td>
</tr>
<tr>
<td>client</td>
<td></td>
</tr>
<tr>
<td>blktrace</td>
<td>blktrace 1.0</td>
</tr>
<tr>
<td>postmark</td>
<td>postmark 1.5</td>
</tr>
<tr>
<td>Disks</td>
<td>1 Seagate ST3160023AS, 160GB, 7200RPM.</td>
</tr>
<tr>
<td></td>
<td>12 Seagate ST3500320AS, 500GB, 7200RPM.</td>
</tr>
<tr>
<td>postgresql</td>
<td>postgresql 8.1.19</td>
</tr>
<tr>
<td>TPC-W</td>
<td>TPC-W 1.5</td>
</tr>
<tr>
<td>Mainboard</td>
<td>GA-945GCMX-S2</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel(R) Celeron(R) CPU 2.80GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>512MB DDR2</td>
</tr>
<tr>
<td>HBA</td>
<td></td>
</tr>
<tr>
<td>TPC-W</td>
<td>TPC-W 1.5</td>
</tr>
<tr>
<td>Mainboard</td>
<td>GA-945GCMX-S2</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel(R) Xeon(R) CPU 5110 @ 1.60GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>1GB DDR2</td>
</tr>
<tr>
<td>HBA</td>
<td>Highpoint 2240 RAID,</td>
</tr>
</tbody>
</table>
S²-RAID 5 reconstruction performance

- Two evaluation parameters
  - Average User Response Time
  - Reconstruction Time

<table>
<thead>
<tr>
<th>Trace File</th>
<th>Write Ratio</th>
<th>Ave Req Size: KB</th>
<th>Total Req</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial-1</td>
<td>76.84%</td>
<td>3.38</td>
<td>5,334,987</td>
</tr>
<tr>
<td>Financial-2</td>
<td>17.65%</td>
<td>2.39</td>
<td>3,699,195</td>
</tr>
<tr>
<td>Websearch</td>
<td>0%</td>
<td>15.07</td>
<td>4,579,809</td>
</tr>
</tbody>
</table>

Average User Response Time (ms)
$S^2$-RAID 5 reconstruction performance

Reconstruction Time (seconds)
S²-RAID 5 Normal Performance

Average User Response Time (ms)

- Fin1: 32%
- Fin2
- Web

MD
S²-RAID


S²-RAID 5 Degraded Performance

Average User Response Time (ms)

- Fin1
- Fin2
- Web

MD
S²-RAID
Other Benchmark Performance (MD vs $S^2$-RAID)

TPCC: 20 warehouses with 10 terminals per warehouse interval of 120 minutes

TPCW: 150 emulated browsers

Postmark: 20,000 files of size 4KB to 500KB and to perform 100,000 transactions
Sensitivity Parameters for Reconstruction

- Some sensitivity parameters
  - Reconstruction speed bandwidth
  - I/O request block size
  - Number of spare disk (additional disk not system disk)
Reconstruction speed bandwidth

The result is based on Financial-1 traces

Speed threshold: 1MB/s ~ 200MB/s

Average User Response Time
Reconstruction Time

The result is based on Financial-1 traces
Number of spare disk

![Graph showing speedup and reconstruction time vs. number of spare disks]

Speed bandwidth and No. of spare disk is insensitive to $s^2$-RAID
I/O request block size

- RAID5-random
- S2-RAID5-random
- RAID5-sequence
- S2-RAID5-sequence

S²-RAID has smaller stripe span
Conclusion

- A parallel reconstruction data layout
- Implement the $s^2$-RAID prototype and evaluation of this structure

- $S^2$-Raid reduces the reconstruction time greatly.
- User response time of $S^2$-Raid is comparable to that of MD.

- Optimization?
  - Embedding existing rebuilding process (distributed sparing)-----Reduce the number of disks
  - Tolerate the mulit-disk failures.
Thank you for your attention!

Questions?