Firebird and RAID

Choosing the right RAID configuration for Firebird.

Paul Reeves
IBPhoenix

mail: preeves@ibphoenix.com
Introduction

Disc drives have become so cheap that implementing RAID for a firebird server is now very affordable.
Intended Audience

- It is hard enough to sell clients your application, never mind get them to invest in a suitable server.
- Your clients don't have much of an in-house IT dept.
- Your Sys Admins don't see why you need a dedicated RAID array for your departmental server.

Unintended Audience

- If you are working for a company that can afford a million dollar SAN this talk may not be much use to you.
The focus of this talk

- Primarily looking at the day to day performance issues that underly RAID
- Data recovery is not really considered. However data recovery is heavily impacted by choice of RAID.
What is RAID?

**Redundant Array of Inexpensive Discs**

- **Redundant** – if one disc fails the system continues.
- **Array** – Discs grouped together bring better performance.
- **Inexpensive** – Many cheap discs combined can work better than more expensive discs.
**RAID is NOT**

- An alternative backup strategy
- RAID exists to overcome disc failure, not data corruption.
- Data corruption is copied to all discs so always make sure you make backups.
Redundancy has a price

- All RAID implementations require writing to more than one disc.
- Database updates can actually become slower (in extreme cases).
Forget about the fancy RAID names

There are just two basic types of RAID configuration
• Mirrored
• Parity

And there is two types of No RAID at all
• JBOD
• Disc striping (concatenating) – RAID 0
JBOD

- **Just a Bunch Of Discs**
- Where would we be without acronyms?
RAID 0

- Good for creating very large discs
- A disaster waiting to happen.
- Not much use for database storage.
- Becomes very useful when combined with other RAID levels.
Mirrored RAID

- Maintain identical data on two or more discs
- Each disc in the array requires a write.
- Usually implemented as RAID 1 or RAID 10
**Parity RAID**

- Writes data blocks on every N discs -1 plus parity block(s).

- Distribution of data and parity blocks is evenly distributed across all discs.

- All discs in array must be written to.

- Calculating parity costs read I/O.

- Usually implemented as RAID 5, RAID 50, RAID 6 or RAID 60.
Combining RAID levels.

- Two or more arrays are concatenated to make a larger array.
Choosing the correct RAID level
Calculating Hard Disc performance

IOPS – Input / Output Operations Per Second

For hard drives we first need to calculate average latency:

\[
\text{Avg Latency} = (60 / \text{RPMs} / 2) \times 1000
\]

We then take the average seek time for the drive and derive the IOPS:

\[
\text{IOPS} = 1000 / (\text{avg latency} + \text{avg seek})
\]
A rough guide to IOPS for different disc speeds

Manufacturers don't always provide full specifications but we can make a good guess.

<table>
<thead>
<tr>
<th>RPM</th>
<th>Avg Latency</th>
<th>Avg Read Seek</th>
<th>RIOPS</th>
<th>Avg Write Seek</th>
<th>WIOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,400</td>
<td>5.56</td>
<td>9.40</td>
<td>66</td>
<td>10.50</td>
<td>62</td>
</tr>
<tr>
<td>7,200</td>
<td>4.17</td>
<td>8.50</td>
<td>79</td>
<td>9.50</td>
<td>73</td>
</tr>
<tr>
<td>10,000</td>
<td>3.00</td>
<td>3.80</td>
<td>147</td>
<td>4.40</td>
<td>135</td>
</tr>
<tr>
<td>15,000</td>
<td>2.00</td>
<td>3.50</td>
<td>182</td>
<td>4.00</td>
<td>167</td>
</tr>
</tbody>
</table>
What does IOPS really mean?

- IOPS is a theoretical value.
- As such it has no relation to actual data throughput.
- IOPS indicates the maximum number of times per second that a drive could randomly read or write to a disc.
Random and Sequential Access - not what they seem

- Sequential access is almost non-existent on a server if more than one process is accessing the disc.

- Random is rarely random – usually several blocks can be written in a single I/O.
## The Write Penalty

<table>
<thead>
<tr>
<th>RAID Level</th>
<th>Min. No. Disks</th>
<th>Write Penalty</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>JBOD / RAID 0</td>
<td>1</td>
<td>1</td>
<td>One disc. One write.</td>
</tr>
<tr>
<td>RAID 1</td>
<td>2</td>
<td>2</td>
<td>Write penalty is directly related to the number of disks in the mirror. For three disks the penalty is 3.</td>
</tr>
<tr>
<td>RAID 5</td>
<td>3</td>
<td>4</td>
<td>Every write requires a read of the data block in the stripe, a read of the existing parity block then the actual write of the updated data block and parity block.</td>
</tr>
<tr>
<td>RAID 6</td>
<td>4</td>
<td>6</td>
<td>As for RAID 5 except extra parity block requires an additional read and write</td>
</tr>
</tbody>
</table>
Calculating RAID performance

There is a simple formula:

\[
\frac{(\text{DISK\_WIOPS} \times \text{NO\_DISCS} \times \%\text{WRITES})}{\text{WRITE\_PENALTY}} + \frac{(\text{DISK\_RIOPS} \times \text{NO\_DISCS} \times \%\text{READS})}{\text{WRITE\_PENALTY}}
\]

= Theoretical maximum random IOPS for a given array.
Here we compare a two disk RAID 1 array to a three disk RAID 5 array. RAID 5 manages to maintain a performance advantage.
Here we compare a four disk RAID 10 array to a five disk RAID 5 array. RAID 5 works better, as long as the reads are light.
Six Disc RAID configurations

Comparison of IOPS with Six Disks

A six disk array of RAID 10 consistently outperforms a six disk array of RAID 50
Eight Disc RAID configurations

Comparison of IOPS with Eight Disk RAID set

8 disk RAID 10 outperforms all other 8 disk RAID configurations
Can slower discs be better value than faster discs?

Comparison of slower disks using RAID 1 with faster discs using RAID 5

Here we compare a four and six disk RAID 10 arrays using cheaper 7,200 rpm discs with a three disk RAID 5 array of 10,000 rpm.
Summary of Theory

- Adding discs increases available IOPS.
- The Write Penalty is real.
- Write intensive applications always benefit from Mirrored RAID.
- For a given number of discs Mirrored RAID will always outperform Parity RAID in Random I/O unless the database is read only.
So much for theory.

- What about reality?
First, you need a RAID Controller

- Firmware based RAID controllers
- Software based RAID controllers
- Hardware based RAID controllers
**Firmware RAID**

- AKA Fake RAID or Bios RAID.
- Usually built into Motherboard or on cheap disc controller cards.
- Not easily portable in the event of failure.
- Requires CPU and RAM from host.
- It brings the benefit of configuration in the bios.
  
  By extension, this allows the O/S to boot from the RAID.
  
  But it also renders remote recovery a problem.
**Software RAID**

- Part of the O/S
- No special hardware required
- Requires CPU and RAM from Host.
- No Vendor Lock-in.
- Portable – if Host fails just pop the discs into another computer.
- Linux implementation rich in functionality – aims to provide top class RAID support.
- No BBU
Hardware RAID

- CPU independent
- Built-in cache
- Battery backup of cache
- Ease of configuration
- Multi-platform client GUI (HP, IBM?)
- Disc monitoring
- Hot spares
- Hot swapping
- Vendor lock-in.
A word about Stripe/Strip size

- What is it?
- Does it affect RAID choice?
- Does it affect FB page size?
RAID, Strip and Page Size

DB Restore based on RAID Type, Strip Size and Firebird Page Size. Smaller is better.
RAID Performance in the real world

- Case study – HP Smart Array 410
- Case study – s/w RAID on Linux
INSERTS comparison HW RAID

INSERTS - 5000 rows. Average of Five test runs

RAID 5 is at least 35% slower

- Raid 10 Strip 256: 115
- Raid 10 Strip 8: 131
- Raid 10 Strip 16: 135
- Raid 5 Strip 256: 155
- Raid 5 Strip 16: 172
- Raid 5 Strip 8: 183
INSERTS comparison SW RAID

INSERTS - 5000 rows. Average of Five test runs (S/W RAID)

RAID 5 is not even in the game.

- Raid 10 16K Pages: 122
- Raid 10 8K Pages: 130
- Raid 5 4K Pages: 268
- Raid 5 8K Pages: 274
- Raid 5 16K Pages: 286

IBPhoenix
THE POWER WITHIN
UPDATE - 50,000 rows. Average of Five test runs

RAID 5 is at least 20% slower
UPDATES comparison SW RAID

UPDATES - 50,000 rows. Average of Five test runs ( S/W RAID )

RAID 5 is 50% slower.
SELECTS comparison HW RAID

SELECT ~40,000 rows. Average of Five test runs

RAID type is not relevant

<table>
<thead>
<tr>
<th>RAID Configuration</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raid 5 Strip 8</td>
<td>6</td>
</tr>
<tr>
<td>Raid 5 Strip 16</td>
<td>6</td>
</tr>
<tr>
<td>Raid 10 Strip 8</td>
<td>6</td>
</tr>
<tr>
<td>Raid 10 Strip 16</td>
<td>5</td>
</tr>
<tr>
<td>Raid 10 Strip 256</td>
<td>6</td>
</tr>
<tr>
<td>Raid 5 Strip 256</td>
<td>7</td>
</tr>
</tbody>
</table>
SELECTS comparison SW RAID

Results similar to HW RAID
And what about SSD in all this?

- SSD raises the bar – IOPS do increase massively.
- Wear Levelling, TRIM, Garbage Collection and Write Amplification pose real problems for database use especially for MLC based flash drives.
- RAID and TRIM don't (yet) work together.
- Not all SSDs are created equal – check benchmarks from a reliable h/w test site.
- Don't believe the manufacturers specs. Do your own real world tests.
- Smaller drives seem to have poorer performance!
- Price / Capacity ratio is still a hindrance to uptake.
- SS Drives can still fail and when they do, failure is total.
Conclusion

We've compared parity and mirrored RAID at the fundamental, theoretical level.

We've also looked at some real world examples of RAID.

Although parity RAID can be tweaked it cannot out perform a mirrored RAID implementation of the same spec when deploying a database server.