Using Preferred Read Groups in Oracle ASM
Introduction

Important Oracle databases require:

- **High Performance**
  - Queries, reports, and screens must return quickly
  - Scale to high user loads

- **Reliability**
  - 100% uptime
  - Single system fault can not be fatal
  - Loss of processing impacts bottom line

- **Cost Effectiveness**
  - Effective use of resources
  - Leverage tech to achieve outsized performance gains for the cost
  - Reliability can not be compromised
Oracle Performance
Performance

- Issues that most affect the performance are related to the IO subsystem

![Relative Improvement Over Time](chart.png)
IO is the Biggest Issue

- The IO subsystem is the weakest link
- Many complex techniques are used to squeeze the last bit of performance from disk drives
- Disk drives
  - Limited to 15K RPM
  - Latency ~5 milliseconds
    - The main component of a disk drive’s latency is its rotational latency
How Do Disk Systems Compensate?

- Increase the number of active disks
- Each disk has a max of 200 random IOPS
- To achieve 10,000 IOPS
  - ~ 50 disk drives for 2-5 ms response time

- Does NOT reduce latency below that of what a single disk can achieve
Solid State Disks

- Solid State Disks have 10 to 50x better latency than HDDs
  - 80 microsecond write performance
  - 250 microsecond read performance
Oracle and Queries
Where does latency matter?

READS

- Reads - Cache miss
- Foreground Waits:
  - DB file sequential read
  - DB file scattered read
  - 5 ms

Storage latency

User’s Query

Oracle Processes

Memory
SGA & PGA

Tables & Indexes

Logs
Oracle and Insert/update/delete
Where does latency matter?

LOG WRITES

Memory
SGA & PGA

Oracle Processes

Users Insert
Commit

DBWR
(Background)

LGWR
(Foreground)

Tables & Indexes

Logs

The World’s Fastest Storage®
What is the Solution?

- Adding disks may help, up to a point!
- SSDs help but can be expensive
- Mirroring to both disk and SSD can cause convoy effect
- ASM Preferred Read Groups offer the solution
Optimized Architecture for ASM PRG
Optimized Architecture for ASM PRG

Lots of RAM and CPU resources

2, 10, or 20 TB of SSD

6, 30, or 60 TB of Enterprise SAS HDD
Architectue

- Integrates Oracle 11g ASM as the Preferred Read Mirror (Group) option
  - Mirror created between SSD and HDD
  - Writes to both
  - Reads are only serviced by the SSD
- Redo Undo and Temp
  - Write performance matters
  - Stored on mirrored SSD
- Other disk managers also offer this!
Optimized PRG - Reads

Reads - Cache miss
Foreground Waits:
DB file sequential read
DB file scattered read
0.25 ms
Optimized PRG – Writes Insert, Update, Delete

Memory
SGA & PGA

Oracle Processes

LGWR
SSD (foreground)

DBWR (background)
Disk speed

Users Insert
Commit
Demonstration Setup

- Three Tables
  - Part (600m rows)
  - Supplier (30m rows)
  - PartSupp (2.4b rows)

- Three Indexes
  - (partkey, suppkey, partkey+suppkey)
A Simple query

Finds the total amount owed to all suppliers for a particular part:

```sql
select sum(s_acctbal) into sum_s_acctbal
from supplier
where s_suppkey
in (
    select ps_suppkey
    from partsupp
    where ps_partkey = (x)
);
```
Run many times

- From each server (4 total), 50 simulated users run a stored procedure 10 times that submits this query 1000 times

- $4 \times 50 \times 10 \times 1000 = 2,000,000$ Queries

- *Demo with disks or SSD set to preferred*

```
- SQL> alter system set
  ASM_PREFERRED_READ_FAILURE_GROUPS = 'HYBRID.RAMSAAN';
  System altered.
- SQL> alter system set
  ASM_PREFERRED_READ_FAILURE_GROUPS = 'HYBRID.DISK';
  System altered.
```
With the Disks Alone (PRG=DISK)

- ~4000 IOPS per RAC node
  - 16,000 IOPS total
- 12.25 minutes to complete with 4 nodes running (2m queries).

```
[oracle@opera1 ~]$ time ./spawn_50.sh
real    12m15.434s
user    0m5.464s
sys     0m4.031s
```
With the SSD (PRG=SSD)

- 40,000 IOPS per RAC node
  - 160,000 total in this test
- 1.3 minutes to complete with 4 nodes running (2m queries).

```
[oracle@opera1 ~]$ time ./spawn_50.sh

real    1m19.838s
user    0m4.439s
sys     0m3.215s
```
Comparison - AWR

- **Disk (13 ms per read):**

  Top 5 Timed Foreground Events

<table>
<thead>
<tr>
<th>Event</th>
<th>Waits</th>
<th>Time(s)</th>
<th>Avg wait (ms)</th>
<th>% DB time</th>
<th>Wait Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>db file sequential read</td>
<td>257,293</td>
<td>3,293</td>
<td>13</td>
<td>82.54</td>
<td>User I/O</td>
</tr>
<tr>
<td>db file parallel read</td>
<td>30,915</td>
<td>567</td>
<td>18</td>
<td>14.22</td>
<td>User I/O</td>
</tr>
<tr>
<td>DB CPU</td>
<td></td>
<td>75</td>
<td></td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>gc cr grant 2-way</td>
<td>199,215</td>
<td>36</td>
<td>0</td>
<td>0.91</td>
<td>Cluster</td>
</tr>
<tr>
<td>reliable message</td>
<td>346</td>
<td>10</td>
<td>28</td>
<td>0.24</td>
<td>Other</td>
</tr>
</tbody>
</table>

- **SSD (<1 ms per read):**

  Top 5 Timed Foreground Events

<table>
<thead>
<tr>
<th>Event</th>
<th>Waits</th>
<th>Time(s)</th>
<th>Avg wait (ms)</th>
<th>% DB time</th>
<th>Wait Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>gc cr grant 2-way</td>
<td>1,703,359</td>
<td>1,344</td>
<td>1</td>
<td>35.93</td>
<td>Cluster</td>
</tr>
<tr>
<td>db file sequential read</td>
<td>2,250,261</td>
<td>1,253</td>
<td>1</td>
<td>33.51</td>
<td>User I/O</td>
</tr>
<tr>
<td>DB CPU</td>
<td></td>
<td>637</td>
<td></td>
<td>17.02</td>
<td></td>
</tr>
<tr>
<td>gc cr multi block request</td>
<td>367,691</td>
<td>356</td>
<td>1</td>
<td>9.52</td>
<td>Cluster</td>
</tr>
<tr>
<td>db file parallel read</td>
<td>276,130</td>
<td>111</td>
<td>0</td>
<td>2.96</td>
<td>User I/O</td>
</tr>
</tbody>
</table>
# Tablespace IO Stats

**Disk:**

<table>
<thead>
<tr>
<th>Tablespace</th>
<th>Reads</th>
<th>Av Reads/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS_S</td>
<td>131,487</td>
<td>1,677</td>
</tr>
<tr>
<td>TS_I_LORDERKEY</td>
<td>124,720</td>
<td>1,590</td>
</tr>
<tr>
<td>TS_PS</td>
<td>58,061</td>
<td>740</td>
</tr>
<tr>
<td>SYSAUX</td>
<td>3,761</td>
<td>48</td>
</tr>
<tr>
<td>UNDOTBS3</td>
<td>178</td>
<td>2</td>
</tr>
<tr>
<td>DISKS_TEMP</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>68</td>
<td>1</td>
</tr>
</tbody>
</table>

**SSD:**

<table>
<thead>
<tr>
<th>Tablespace</th>
<th>Reads</th>
<th>Av Reads/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS_S</td>
<td>1,181,958</td>
<td>15,562</td>
</tr>
<tr>
<td>TS_I_LORDERKEY</td>
<td>1,117,768</td>
<td>14,970</td>
</tr>
<tr>
<td>TS_PS</td>
<td>520,385</td>
<td>6,969</td>
</tr>
<tr>
<td>SYSAUX</td>
<td>2,448</td>
<td>33</td>
</tr>
<tr>
<td>UNDOTBS3</td>
<td>713</td>
<td>10</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>296</td>
<td>4</td>
</tr>
<tr>
<td>DISKS_TEMP</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>UNDOTBS1</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
PRG in Oracle ASM

- ALL blocking IO is handled by the SSD
  - >10 times faster performance than HDDs!
- Disks provide redundancy in order to keep costs reasonable.
- No sacrificing redundancy
- Allows reuse of legacy hardware
The World’s Fastest Storage®