Using Preferred Read Groups in Oracle ASM

Introduction

Texas Memory Systems

Important Oracle databases require:

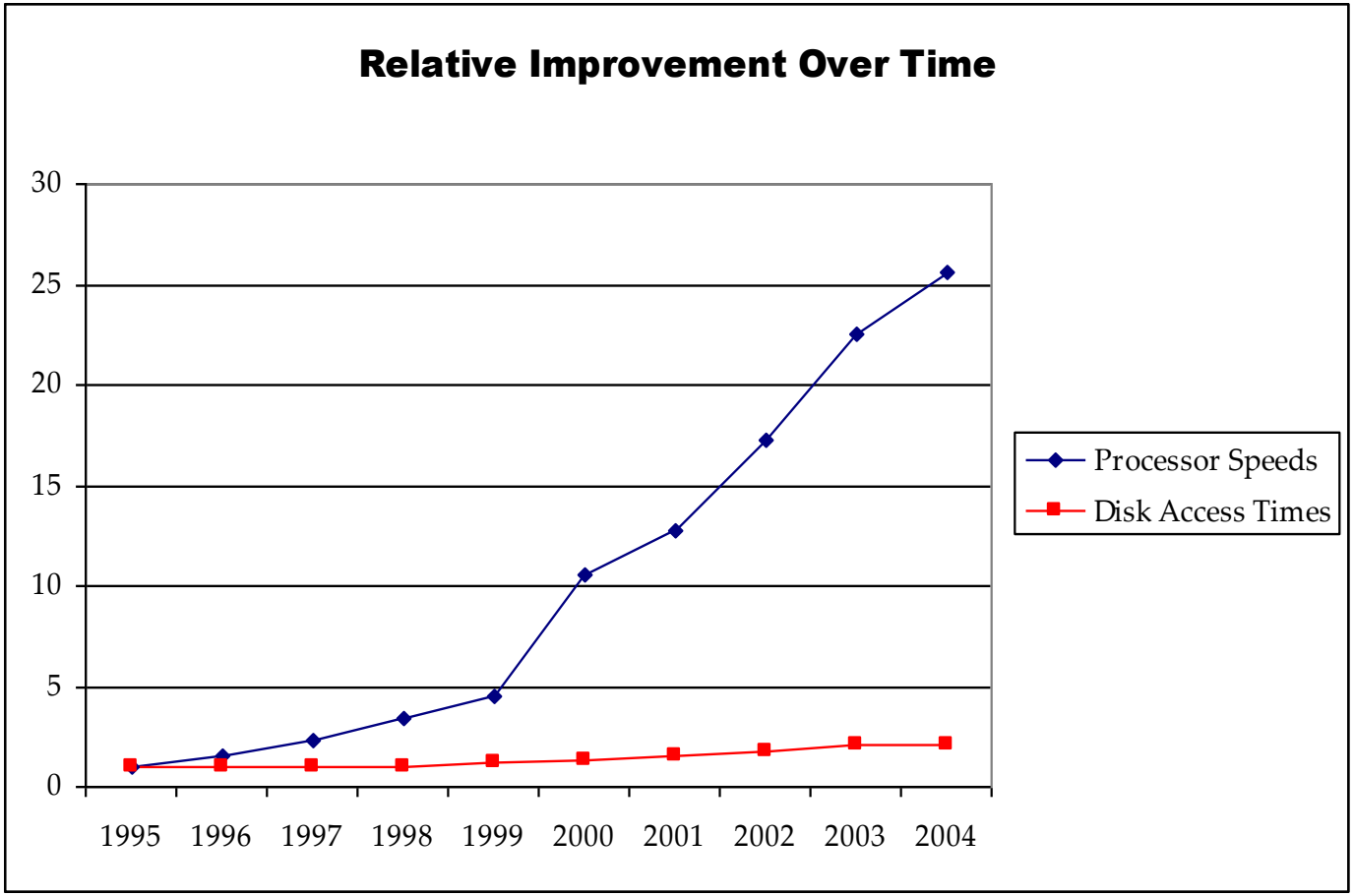
- High Performance
  - Queries, reports, and screens must return quickly
  - Scale to high user loads
- Reliability
  - 100% uptime
  - Single system fault can not be fatal
  - Loss of processing impacts bottom line
- Cost Effectiveness
  - Effective use of resources
  - Leverage tech to achieve outsized performance gains for the cost
  - Reliability can not be compromised

Oracle Performance

Texas Memory Systems

Performance

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



Texas Memory Systems

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

Texas Memory Systems

Solid State Disks

- Solid State Disks have 10 to 50x better latency than HDDs
  - 80 *micro*second write performance
  - 250 *micro*second read performance

Oracle and Queries
Where does latency matter?

READS

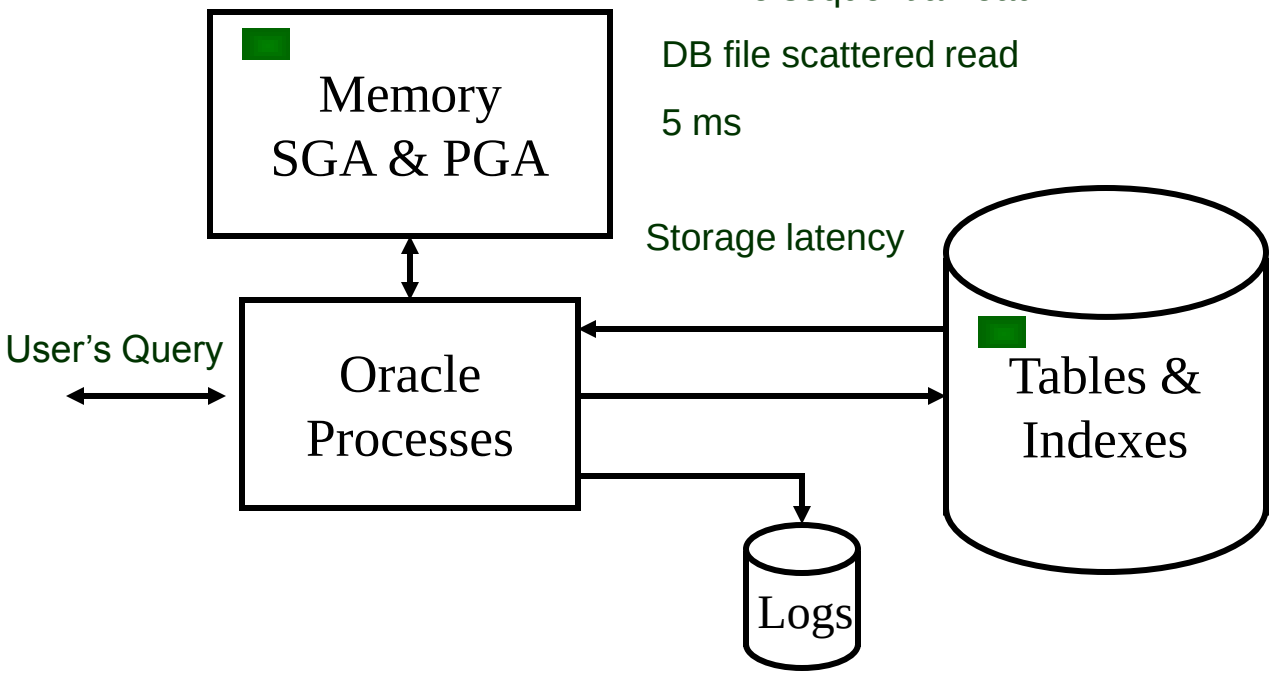
Reads - Cache miss

Foreground Waits:

DB file sequential read

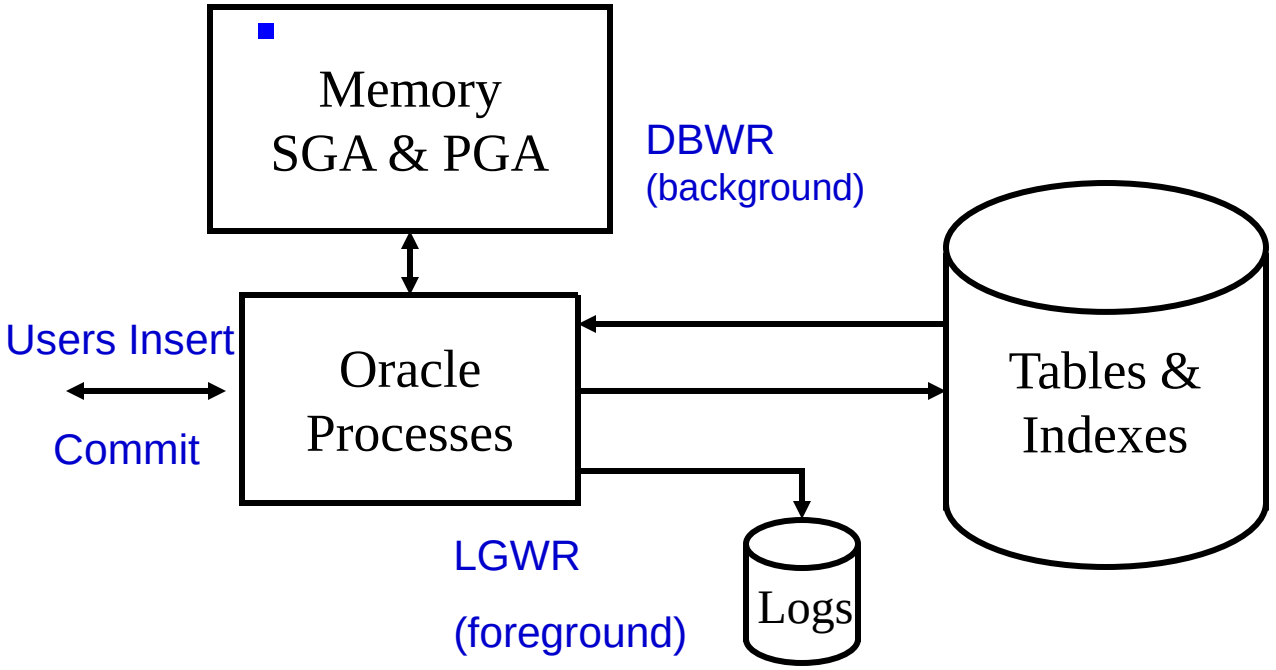
DB file scattered read

5 ms



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

LOG WRITES



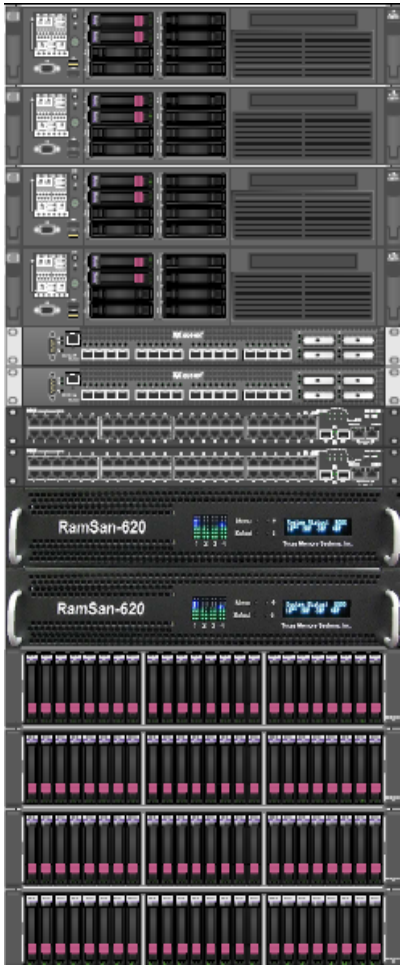
Texas Memory Systems

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

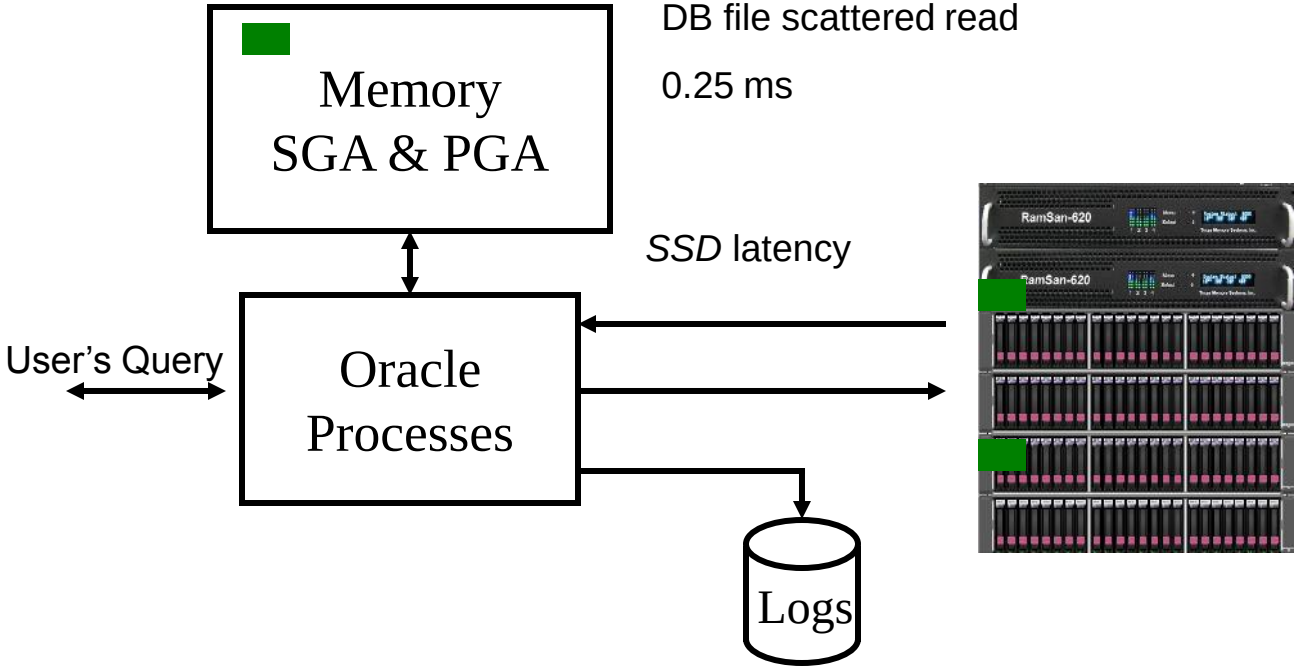
Reads - Cache miss

Foreground Waits:

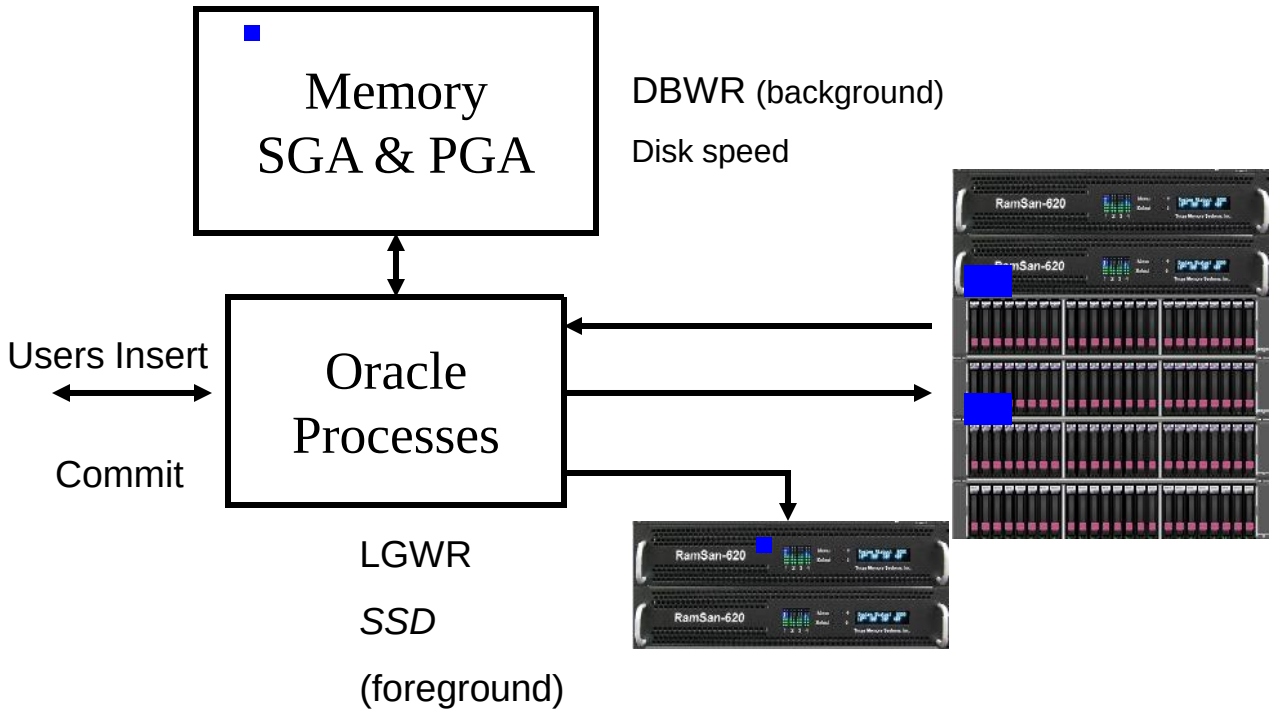
DB file sequential read

DB file scattered read

0.25 ms

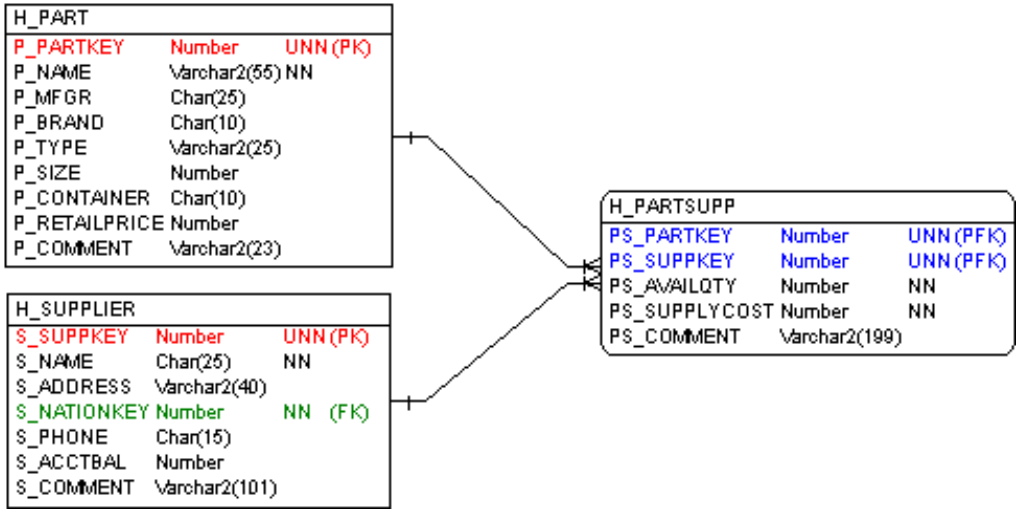


Optimized PRG – Writes Insert, Update, Delete



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:

```
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*50*10*1000 = 2,000,000 Queries
- *Demo with disks or SSD set to preferred*

```
SQL> alter system set
ASM_PREFERRED_READ_FAILURE_GROUPS =
'HYBRID.RAMSAN';

System altered.

SQL> alter system set
ASM_PREFERRED_READ_FAILURE_GROUPS = 'HYBRID.DISK';

System altered.
```

Texas Memory Systems

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

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

- SSD(<1 ms per read):

Top 5 Timed Foreground Events

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

Tablespace IOStats

- Disk:

Tablespace IO Stats

- ordered by IOs (Reads + Writes) des

Tablespace	Reads	Av Reads/s
TS_S	131,487	1,677
TS_I_LORDERKEY	124,720	1,590
TS_PS	58,061	740
SYSAUX	3,761	48
UNDOTBS3	178	2
DISKS_TEMP	38	0
SYSTEM	68	1

- SSD:

Tablespace IO Stats

- ordered by IOs (Reads + Writes) desc

Tablespace	Reads	Av Reads/s
TS_S	1,161,958	15,562
TS_I_LORDERKEY	1,117,768	14,970
TS_PS	520,385	6,969
SYSAUX	2,448	33
UNDOTBS3	713	10
SYSTEM	296	4
DISKS_TEMP	41	1
UNDOTBS1	3	0

Texas Memory Systems

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®