Common Exadata Mistakes

Andy Colvin

Practice Director, Enkitec

IOUG Collaborate 2014

ORACLE Platinum Partner

enkitec

About Enkitec

- Global systems integrator focused on the Oracle platform
- Established in August, 2004
- Headquartered in Dallas, Texas
- Consultants average 15+ years of Oracle experience
- Worldwide leader in Exadata implementations

ORACLE Platinum Partner

enkitec

About Me

- 8 years at Enkitec
- Supporting Oracle since 1999
- Working with Exadata since 2010
- Oracle ACE
- Author of Expert Oracle Exadata second edition
- Blog – blog.oracle-ninja.com
- Twitter - @acolvin

ORACLE Platinum Partner

enkitec

Enkitec E4 2014

- 2 day Exadata conference in Dallas
  - June 1 – 3, 2014
  - http://www.enkitec.com/e4

- Tom Kyte
- Maria Colgan
- Tanel Poder
- Kerry Osborne
- Sue Lee
- Doug Burns
- Martin Bach

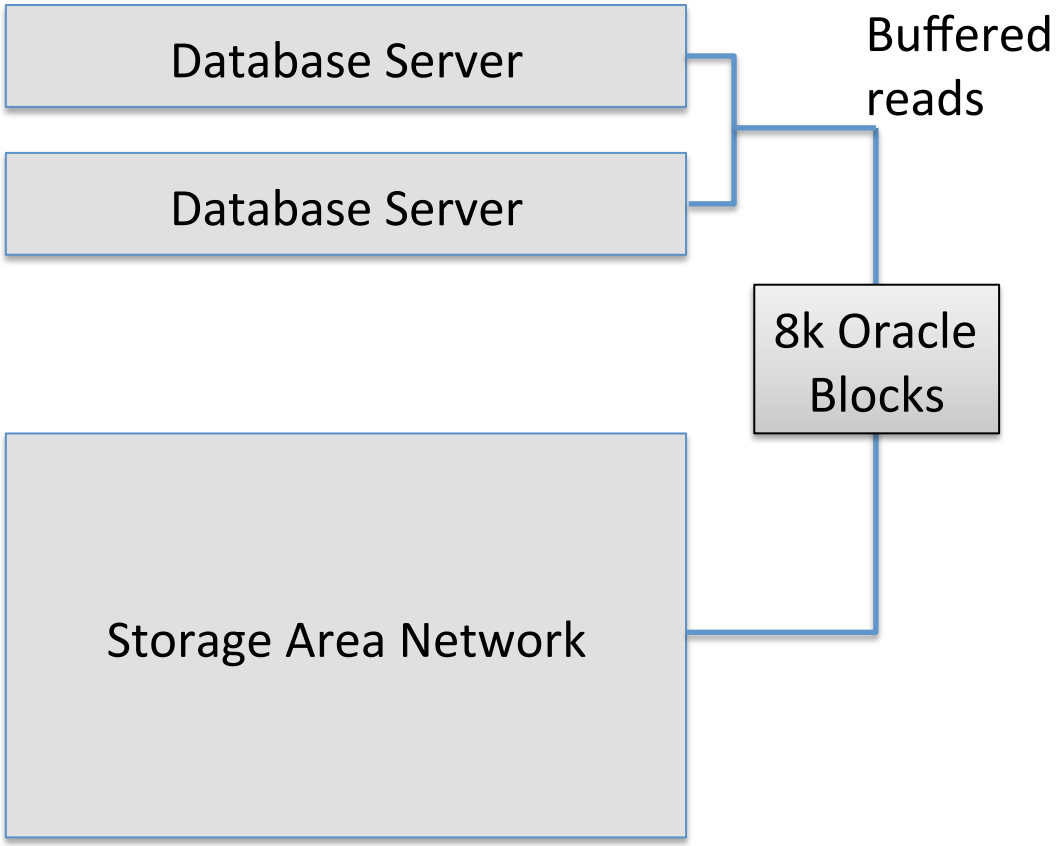
Why Exadata?

- Many databases are I/O bound
  - Slower fibrechannel
  - Older hardware
  - Shared infrastructure
- Exadata solves these problems
  - 40gbps InfiniBand
  - Dedicated Storage

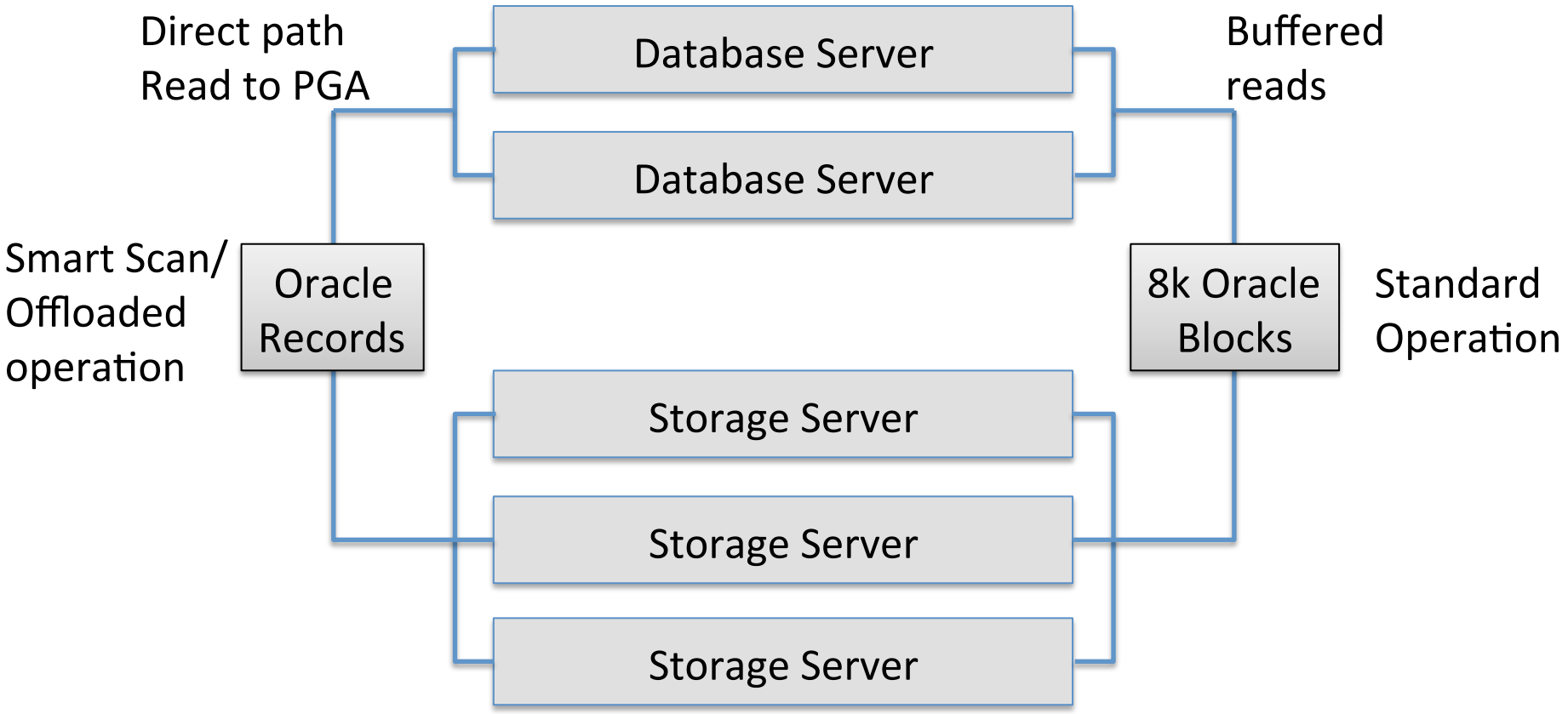
Quick Exadata Primer

- Smart scans are the secret sauce
- Smart scans kick in with full object scans
- Without smart scans, you're probably in "The 3X Club"

Conventional RAC System



Quick Exadata Primer



Common Exadata Mistakes

- Classified into 2 types
  - Performance degrading issues
  - Issues that make Exadata management tougher
- Rule #1 – You aren't special

SGA Sizing

Improperly Sized SGAs

- Large SGAs can be counterproductive (in a warehouse)
- A determining factor for full scans is the small table threshold (`_small_table_threshold` parameter)

Small Table Threshold

- Default setting based on size of buffer cache
- Relative size of a segment to the buffer cache
- Calculation has changed over various versions
- Direct path reads required for smart scans

What Size SGA?

- For pure DW, smaller is better – think 8GB
  - You'll get more smart scans, which use PGA memory
- OLTP is trickier
  - OLTP queries typically need larger SGA
- Start small - you'll never get to shrink your SGA later

Hugepages

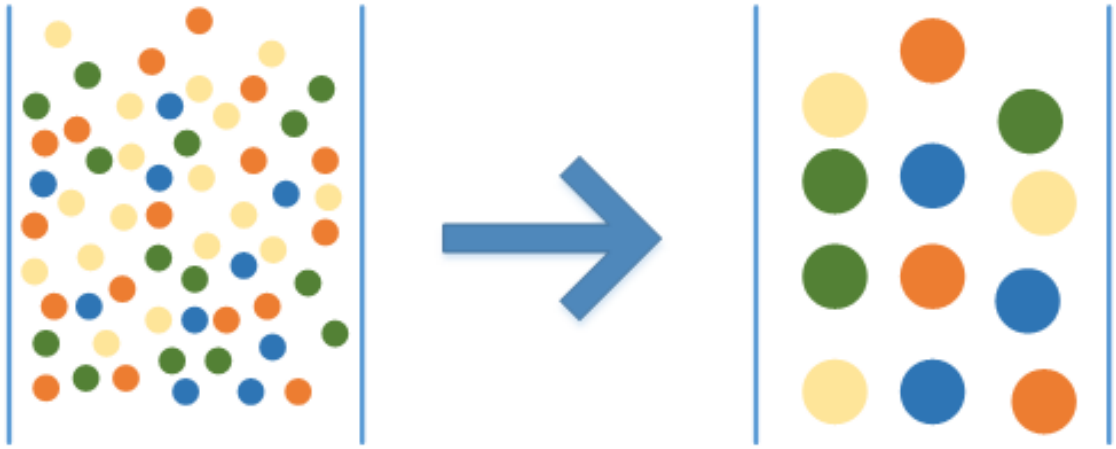
Hugepages

- Allocates memory in bigger “chunks”
  - Default page size – 4KB
  - Huge page size – 2MB
- Controlled via `vm.nr_hugepages` kernel parameter

```
[root@enkx3db01 ~]# grep hugepages /etc/sysctl.conf
vm.nr_hugepages=26014
```

Benefits of using Hugepages

- Less pages to manage = less resource consumption
- Huge pages cannot be swapped to disk



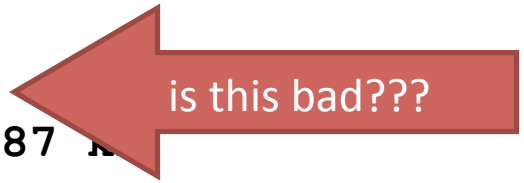
Before Hugepages

After Hugepages

Life Without Hugepages

- X2-2 Half Rack – 96GB RAM per node

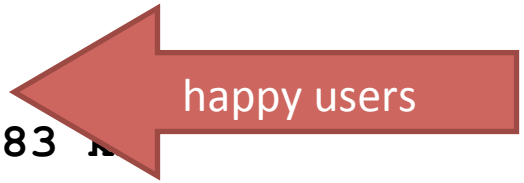
```
[root@xxxxx ~]# cat /proc/meminfo
MemTotal: 98848968 kB
MemFree: 3093112 kB
Buffers: 487884 kB
. . . .
PageTables: 28487608 kB
VmallocChunk: 34359438787 k
HugePages_Total: 0
HugePages_Free: 0
HugePages_Rsvd: 0
Hugepagesize: 2048 kB
```



After Enabling Hugepages

- X2-2 Half Rack – 96GB RAM per node

```
[root@xxxxx ~]# cat /proc/meminfo
MemTotal: 98848968 kB
MemFree: 25423728 kB
Buffers: 360304 kB
. . . .
PageTables: 408268 kB
VmallocChunk: 34359442083 k
HugePages_Total: 30726
HugePages_Free: 10070
HugePages_Rsvd: 8018
Hugepagesize: 2048 kB
```



Hugepages in the Database

- Controlled with `use_large_pages` parameter
  - **`TRUE`** – start up with huge pages if available
    - 11.2.0.3 can use a mix of huge/small pages
  - **`FALSE`** – ignore huge pages that are available
  - **`ONLY`** – instance starts if enough hugepages are available

Hugepages in the Database

- Excerpt from 11.2.0.4 alert log

```
Starting ORACLE instance (normal)
********** Large Pages Information **********
Per process system memlock (soft) limit = 53 GB

Total Shared Global Region in Large Pages = 24 GB (100%)

Large Pages used by this instance: 12289 (24 GB)
Large Pages unused system wide = 1425 (2850 MB)
Large Pages configured system wide = 13714 (27 GB)
Large Page size = 2048 KB
*****************************************
```

Caveats to Hugepages

- Only for SGA – PGA cannot utilize hugepages
- Incompatible with AMM (`memory_target`)
- Compatible with ASMM (`sga_target`)

Allocating Hugepages

- Allocate enough pages to account for all SGA memory
  - See MOS note #401749.1 for script
- Don’t allocate everything to huge pages
  - Over allocation leads to memory starvation, node evictions

Over Indexing

(and under indexing)

Index Use on Exadata

- Index use is #1 smart scan killer
  - Direct path reads required for smart scans
- Mark indexes invisible in a warehouse
  - Global change at the database level

Index Use on Exadata (non-OLTP)

- Look at long running queries with SQL monitoring report
- If it's using indexes, try running with FULL hint to force full table scans
- Try marking indexes invisible (in test)

Index Use on Exadata (OLTP)

- You'll still need many indexes in OLTP environment
- It can be a delicate balance
- Test, test, test before dropping indexes
- Start without them, add as necessary

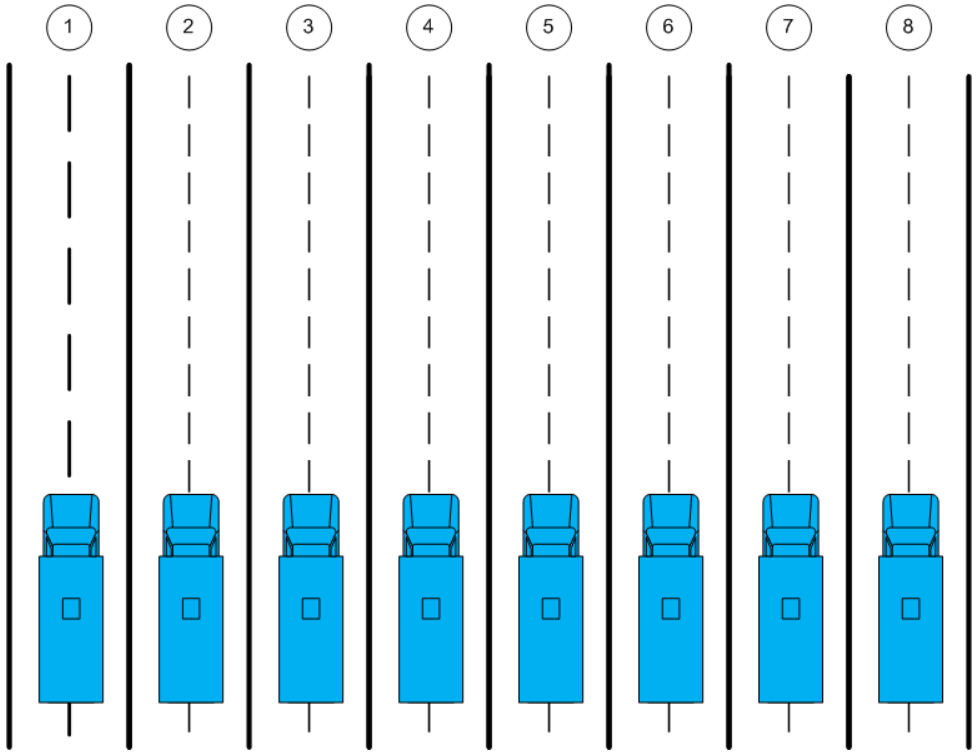
Parallelization

Parallelization

- Parallel will (typically) force direct path reads
- Direct path reads are required for smart scans
- Beware allocating too many slaves
- Remember that smart scans already parallelize on I/O

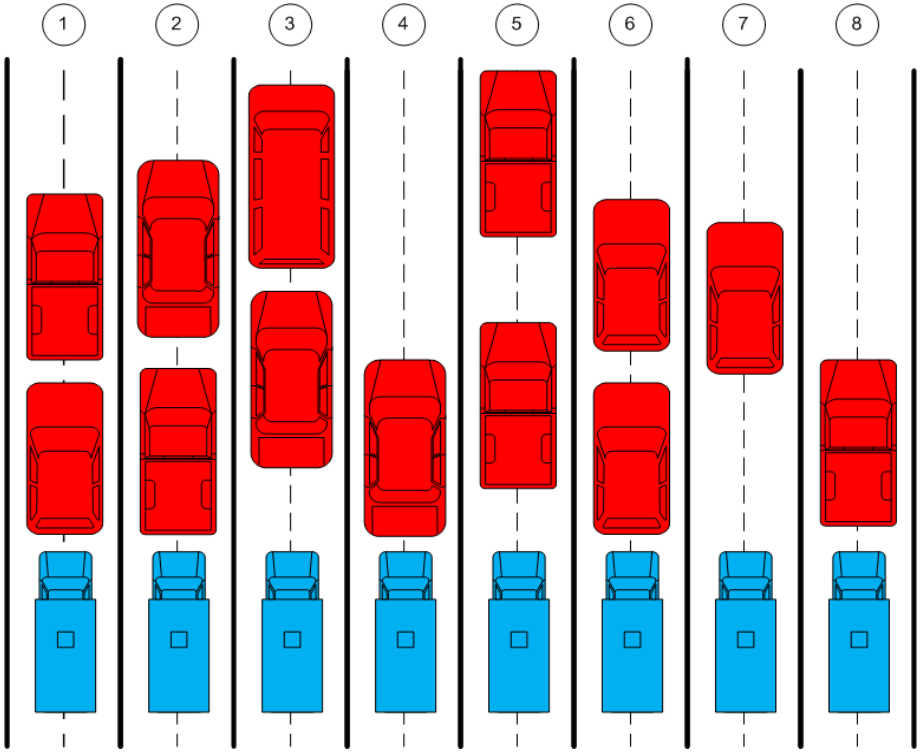
Parallel Slave Allocation

- With no other activity, the more PX is better



Parallel Slave Allocation

- More PX slaves = a busier system



Auto Degree of Parallelism

- Oracle calculates DOP on the fly
- Beware of Auto DOP
  - Test extensively before implementing
  - It's getting better (big changes in 11.2.0.4)
- Requires `DBMS.RESOURCE_MANAGER.CALIBRATE_IO`
- Set `PARALLEL_DEGREE_POLICY = AUTO`

Controlling Parallelism

- Use Database Resource Manager (DBRM)
- Don’t set “DEFAULT” for segment PX degree
- Try lower PX degree to start – storage is already parallel
- Slowly ratchet up DOP operations – powers of 2

Disk Selection

Exadata Storage Concepts

- Exadata is all about smart storage
  - Offload your workload via full scans
  - Single block reads come from flash

Exadata Disk Types

- 2 varieties of disks
  - High performance
    - V2 – X3: 600GB 15,000RPM
    - X4: 1.2TB 10,000RPM
  - High capacity
    - 2TB – 4TB 7,200RPM

Disk Read Performance Comparison

X3-2 High Capacity

```
drive [20:4] random read throughput: 127.24 MBPS, and 200 IOPS
drive [20:5] random read throughput: 126.07 MBPS, and 200 IOPS
```

X3-2 High Performance

```
drive [20:4] random read throughput: 162.67 MBPS, and 407 IOPS
drive [20:5] random read throughput: 162.66 MBPS, and 422 IOPS
```

X4-2 High Capacity

```
drive [20:4] random read throughput: 143.54 MBPS, and 203 IOPS
drive [20:5] random read throughput: 143.76 MBPS, and 203 IOPS
```

X4-2 High Performance

```
drive [20:4] random read throughput: 155.27 MBPS, and 323 IOPS
drive [20:5] random read throughput: 155.99 MBPS, and 329 IOPS
```

results pulled from “cellcli –e calibrate force”

Flash Read Performance Comparison

X3-2 Flash Read Test

```
[FLASH_2_2] random read throughput: 543.70 MBPS, and 37332 IOPS
[FLASH_2_3] random read throughput: 543.33 MBPS, and 37303 IOPS
```

X4-2 Flash Read Test

```
[FLASH_2_2] random read throughput: 539.26 MBPS, and 45683 IOPS
[FLASH_2_3] random read throughput: 539.42 MBPS, and 45184 IOPS
```

X4-2 High Capacity

```
drive [20:4] random read throughput: 143.54 MBPS, and 203 IOPS
drive [20:5] random read throughput: 143.76 MBPS, and 203 IOPS
```

X4-2 High Performance

```
drive [20:4] random read throughput: 155.27 MBPS, and 323 IOPS
drive [20:5] random read throughput: 155.99 MBPS, and 329 IOPS
```

results pulled from “cellcli –e calibrate force”

ORACLE Platinum Partner

enkitec

Which Disk Type?

- High performance for latency-sensitive applications
  - Heavy OLTP workloads
  - Write back flash cache changes this significantly
- For everything else, high capacity
  - Flash cache masks slower disks
  - 3.2TB of flash per cell on X4 (before compression)

Patching

Exadata Patching

- If you don’t patch, you WILL hit bugs
- Patching isn’t something to be scared of any more
  - GI/RDBMS patch (bundle patch)
  - OS/Firmware patch (Storage server patch)
- See MOS note #888828.1 for details

Exadata Patching

- If you can apply a RAC PSU, you can apply an Exadata bundle patch
- Storage server patches are infrequent
- Patches include bug fixes, new features

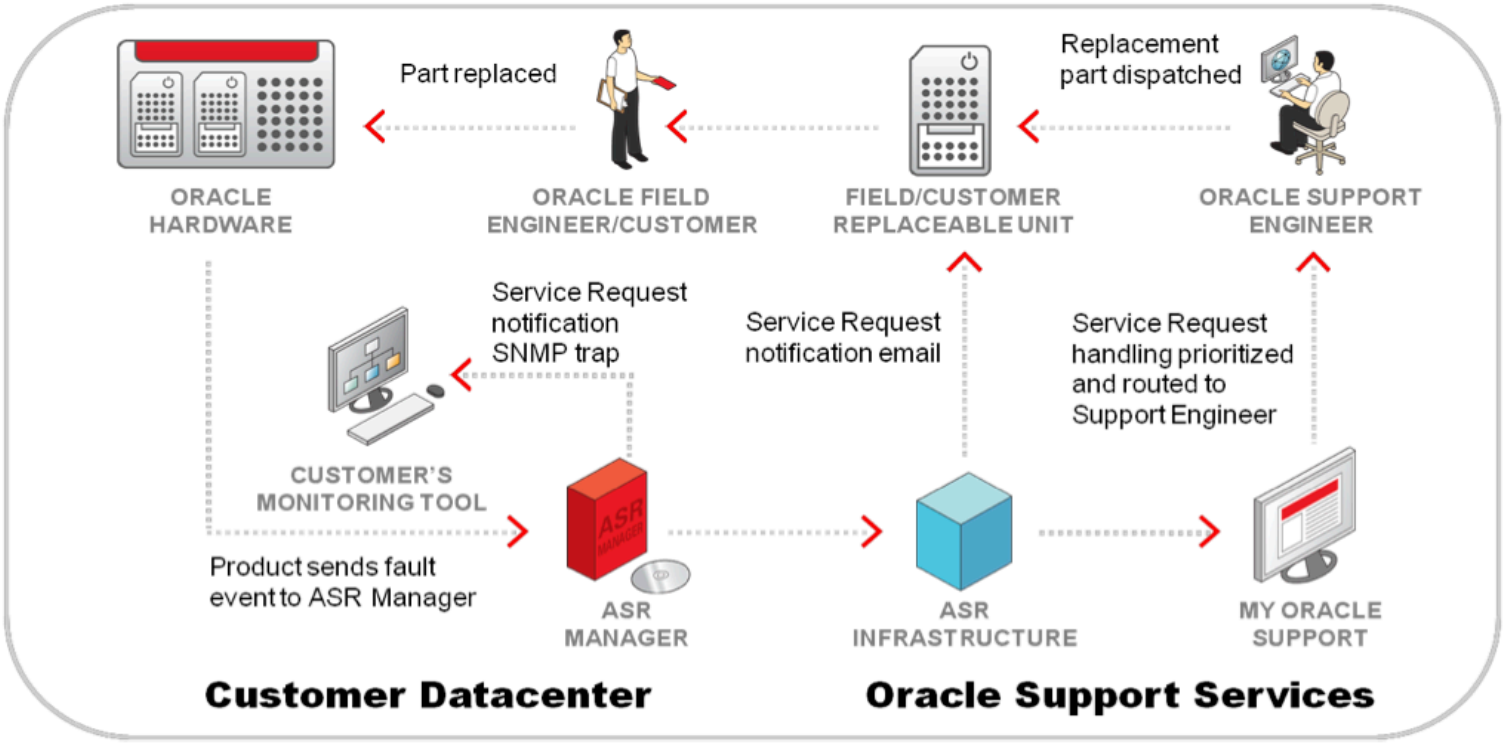
ASR

Automatic Service Request

- Free with your support contract
- No privacy issues – one way communication
- Requires external VM or host running Solaris or OEL
- Speeds up service request resolution time

Automatic Service Request

- ASR creates service requests when hardware breaks



exachk

exachk

- MOS note #1070954.1
- Checks cluster and database against Oracle “best practices” at the time of release
- Requires root passwords for initial configuration
- Generates HTML reports with scorecard

exachk Recommendations

- Configure exachk to run in daemon mode
  - Prevents requirement for root passwords
- Focus on the failures, not the score

Other Common Issues

- Not using OEM 12c
- Role separated installations
- Larger RECO vs DATA diskgroup split
- Trying to make Exadata like your old platform

Questions

andy.colvin@enkitec.com
@acolvin
blog.oracle-ninja.com

ORACLE Platinum Partner

enkitec