Tips learned while implementing Oracle Solutions on Linux for IBM System z

Speaker Names: David Simpson & Tom Kennelly
IBM – simpson.dave@us.ibm.com , kennelly@us.ibm.com

Date of Presentation: Aug 4, 2010
Session Number: 7164

Agenda

- zEnterprise – Some Features for Oracle Databases
- Fit for Purpose / System z196 – Cache is King
- SLES 10 SP3 , SLES 11 SP1 and Red Hat 5.5 Updates
- Reports From the Field…
- Oracle Support Features for System z
- Questions entertained anytime.

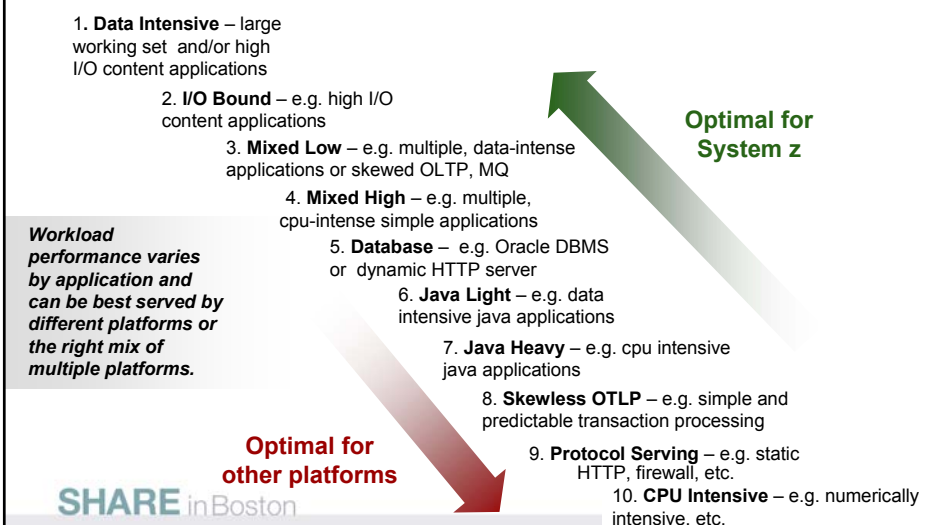
What Does zEnterprise mean for Oracle Workloads?

- More than 100,000 virtualized servers can be managed as a single system and there's 60% more capacity compared to z10 while using same amount of energy.
- Offers 40% lower acquisition costs and reduces cost of ownership by 55%.
- Complex database queries can experience up to a 10-x performance improvement in a hybrid environment.
- IBM relied on internal experts and researchers, as well as customers, in designing zEnterprise. In fact the zEnterprise System architecture was developed with **direct involvement from a team of IBM's 30 top customers**, which provided direct input at every stage of the development process.
- The core server, called zEnterprise 196, can execute more than 50 billion instructions per second and optimizes performance of data-heavy workloads, including up to a **60% improvement in data intensive and Java workloads**.

z196 Oracle Testing

- The IBM Oracle team has had an opportunity to do test on a Linux guest on a z196. The Oracle Database was installed and a small functional test was completed. More extensive tests are planned for later in the year.

Application performance characteristics – What fits on which platform?

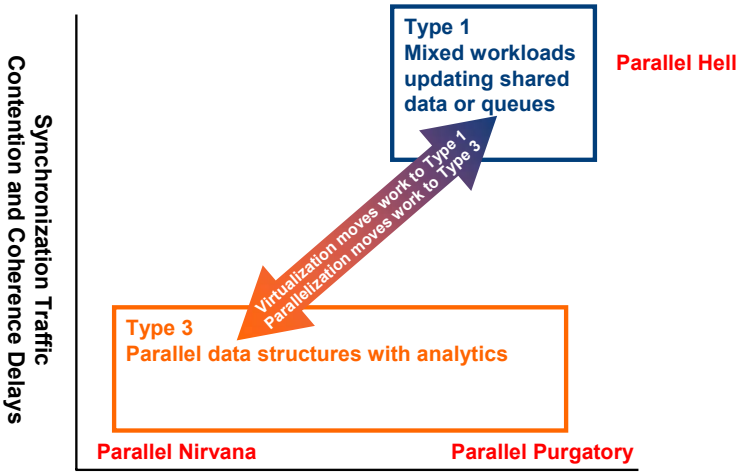


Workload performance varies by application and can be best served by different platforms or the right mix of multiple platforms.

Source: **Ray Kilgore (IBM)**

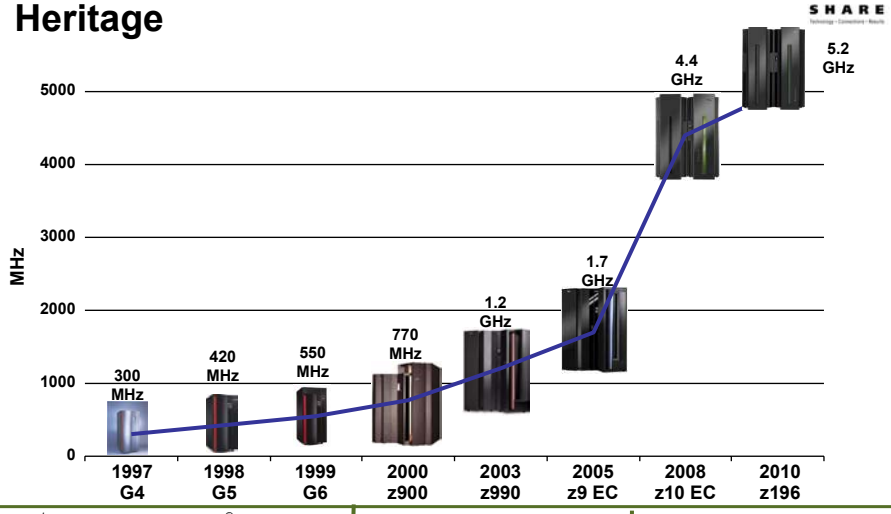
Courtesy- Joe Temple (IBM)

Parallelization and Virtualization are at Odds



Courtesy- Joe Temple (IBM)

z196 Continues the CMOS Mainframe Heritage



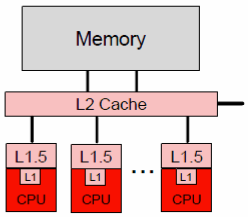
- **G4** – 1st full-custom CMOS S/390®
- **G5** – IEEE-standard BFP; branch target prediction
- **G6** – Copper Technology (Cu BEOL)

- **z900** – Full 64-bit z/Architecture
- **z990** – Superscalar CISC pipeline
- **z9 EC** – System level scaling

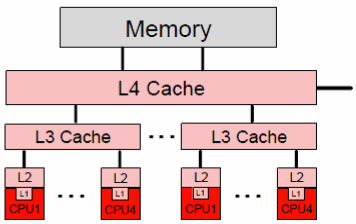
- **z10 EC** – Architectural extensions
- **z196** – Additional Architectural extensions and new cache structure

zEnterprise 196 CPU Cache: Cache is King

- z10 EC
  - CPU
    - 4.4 Ghz
  - Caches
    - L1 private 64k i, 128k d
    - L1.5 private 3 MBs
    - L2 shared 48 MBs / book
    - book interconnect: star

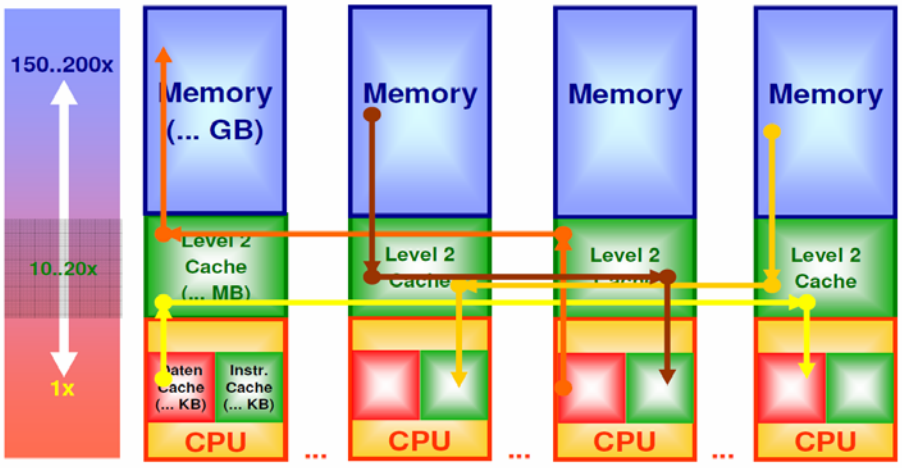


- z196
  - CPU
    - 5.2 Ghz
    - Out-Of-Order execution
  - Caches
    - L1 private 64k i, 128k d
    - L2 private 1.5 MBs
    - L3 shared 24 MBs / chip
    - L4 shared 192 MBs / book
    - book interconnect: star



Scalability: System-Structures optimized for data

The key problem of current microprocessor-systems:
Memory access does not scale with CPU-cycletime !

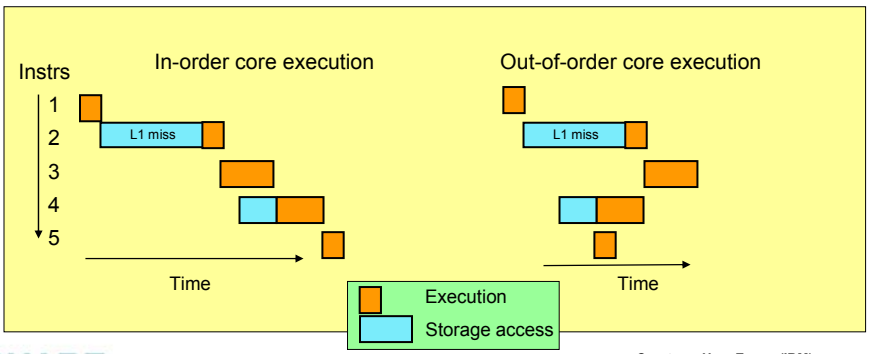


SHARE in Boston

Source: Klaus-Dieter Mueller (IBM)

z196 Out-of-Order (OOO) Value

- **z196 has the first System z CMOS out-of-order core**
- **z196 has the first System z out-of-order core since 1991**
- **OOO yields significant performance benefit for applications through**
  - Re-ordering instruction execution
    - Later (younger) instructions can execute ahead of an older stalled instruction
  - Re-ordering storage accesses and parallel storage accesses
- **OOO maintains good performance growth for traditional apps**



SHARE in Boston

Courtesy- Harv Emery (IBM)

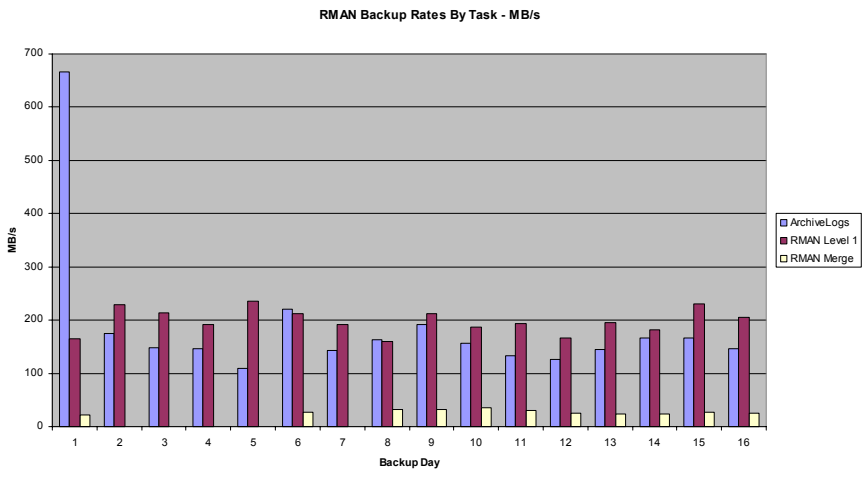
SLES 10 – SP3 Update

- rpm -qa --queryformat="%{n}-%{v}-%{r}.%{arch}.rpm\n" | **grep compat-libstdc++**
  Oracle Support Note: - **556906.1 (SLES 10 Oracle Release Notes)**
- With Fresh Installs of SLES 10 SP3.... The **libstdc++33** package replaces the compat-libstdc++ package
- **# rpm -q --provides libstdc++33-3.3.3-7.8.1**

```
compat-libstdc++
libstdc++5 = 3.3.3-7.8.1
libstdc++.so.5()(64bit)
libstdc++.so.5(CXXABI_1.2)(64bit)
libstdc++.so.5(CXXABI_1.2.1)(64bit)
libstdc++.so.5(CXXABI_1.2.2)(64bit)
libstdc++.so.5(GLIBCPP_3.2)(64bit)
libstdc++.so.5(GLIBCPP_3.2.1)(64bit)
libstdc++.so.5(GLIBCPP_3.2.2)(64bit)
libstdc++.so.5(GLIBCPP_3.2.3)(64bit)
libstdc++.so.5(GLIBCPP_3.2.4)(64bit)
libstdc++.so.5(GLIBCPP_3.2.5)(64bit)
libstdc++.so.5(libstdc++.so.5)(64bit)
libstdc++33 = 3.3.3-7.8.1
```

Conclusion -> It's safe to ignore any errors in regards to the missing **compat-libstdc++** rpm for Oracle Installations

Backup rates (MB/s) – Large System z DB – 35TB



Backup and Recovery Deployment

- **RMAN Level 0, Fast Incremental Level 1, Incrementally Updated Merge**
  - Initial level 0 image copy to disk, followed by nightly incremental backups
  - Roll forward image copy with incremental, to produce new on-disk full backup, on regular basis (e.g. daily)
  - Archived logs are backed up and retained on-disk
  - Great for:
    - VLDBs - no more than several hours for restore & recovery
      - *Fast recovery from disk*
      - *Environments where disk can be allocated for full database image copy or image copy of most critical tablespaces*

  - Support Note: **Merged Incremental Strategy creates backups larger than expected [ID 413265.1]**

    backup incremental level 1 tag 'DAVE' **for recover of copy** database skip offline skip inaccessible ;

Backup Considerations

- Tivoli storage manager is certified with Oracle RMAN backup utility to go directly to tape.

  http://www-1.ibm.com/support/docview.wss?&uid=swg21297510

- Veritas Netbackup, is currently certified for file system backups with System z, so you will need to backup to disk, then use file system utilities to backup to tape, which will require more disk.

http://ftp.support.veritas.com/pub/support/products/NetBackup_Enterprise_Server/279048.pdf
http://ftp.support.veritas.com/pub/support/products/NetBackup_Server/278064.pdf

- FDR Upstream

Reasons For Oracle Node Evictions:

- **Hang Check Timer** - replaced by **oprocd** process in 10.2.0.4 and tuned with CRS diagwait parameter
  - Still needs to be set in 10g & 11g though:
    /sbin/insmod hangcheck-timer hangcheck_tick=1 hangcheck_margin=10 hangcheck_reboot=1
  - Oracle oprocd process sets a timer, then sleeps. When oprocd wakes up again and gets scheduled onto the cpu if it sees that a longer time has passed than the acceptable margin, oprocd will reboot the node.

Richard Lewis Script - /var/log/messages
Apr 17 20:36:23 eagmnmnbp040 logger: WARNING Fri Apr 17 20:36:23 CDT 2009 The date loop took longer than 2 seconds from 1240018575 to 1240018583
Apr 17 20:42:51 eagmnmnbp040 logger: WARNING Fri Apr 17 20:42:51 CDT 2009 The date loop took longer than 2 seconds from 1240018963 to 1240018971

/etc/oracle/oprocd/<nodename>.oprocd.lgl
Apr 24 05:20:03.665 | INF | TrackHistoricalTrends: added first sample 3242554777 in 10 to 50 percentile
Apr 26 05:26:23.593 | INF | TrackHistoricalTrends: added first sample 2642327278 in 10 to 50 percentile
2 entries where the delay in scheduling was approximately 3.24 and 2.64 seconds respectively

- **I/O** - Inability for a Node to communicate with the Voting disk – after 200 seconds will evict.
- **Network** Interconnect cannot ping other node for more than 60 seconds
  - z/VM uses OSA Cards with redundancy connected with 10GB Switch
- **Lack of Memory** - causing Linux to do emergency page scans and kill processes
  - Increase swap space to provide a "**safety net**" for unexpected workloads
- **VM Page Reordering: For every 8GB ~1 second of delay.**
  Turned off for Really Large Oracle Guests. http://www.vm.ibm.com/perf/tips/reorder.html

Monitoring CPU Run Levels / Oracle Parallel Query

Watch the run queue!

 # vmstat 3 (on 2 Virtual CPU Machine)

```
procs -----------memory---------- ---swap-- -----io---- -system-- -----cpu------
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 4  0 276900 286468   1164 468472    0    0     5    26    7    8  0  0 100 0  0  ← Typically Ignore 1st
 1  0 276896 284772   1256 468900    0    0   267    76  257  760 43  7 49  1  0
 2  0 276888 272052   1392 470320    0    0   475   107  218  439 47  4 47  1  2
 3  0 275672   8988   1228 464564  277 42971 1224 47888 1332  350 67 11  0 15  6
 2  0 273636   8884    652 489576  524    3   889 20575  397  321 59  4 37  0  1
 1  0 271560   8580    788 536964  599    5   984 29069  470  255 61  3 34  1  1
 1  0 267576   8732   1068 591056 1412    0  3772 31208  796  696 50 11 22 16  1
 6  5 283124   6168    240 586176  299 5451  2148 17865 1220  528 15 24  6 53  1
 0  8 307192   5840    432 614808  437 8451 12868 26735 1249  575 14 21  2 59  4
16 12 307192   6668    136 572948    3   17 46792   701 1744  963  0 87  0 13  1
15 15 307192   7796    120 570384    0    0 13271     0  393  188  0 99  0  0  1
```

- r –run queue –how many processes currently waiting for or running on the CPU try to keep < # of Virtual IFLs for Oracle Parallel Query
- b –how many processes waiting – want to keep low
- Steal time is the percentage of time a virtual CPU waits for a real CPU while the hypervisor is servicing another virtual processor.

Oracle RAC Concurrency

Partitioning / Global Temporary Tables – Network Interconnect

- **Enqueue Waits with TEMP observed**
  - Possible PGA memory was constrained
- **Truncating tables concurrently from different RAC instances does not scale well**, (especially if direct read operations such as parallel queries are taking place at same time)
  - While the first TRUNCATE command is processing, the second has to wait until the first one completes. There are different types of cross-instance calls. However, all use the same serialization mechanism. The cache flush for a partitioned table with many partitions may add latency to a corresponding parallel query.
  - Each cross-instance call is serialized at the cluster level, and one cross-instance call is needed for each partition at the start of the parallel query for direct read purposes.

Source: Joel Goodman, Oracle University - Oracle 10g RAC Tuning Tips

Data Block Contention: Optimal Design

- Small tables with high row density and frequent updates and reads can become “globally hot” with serialization e.g.
  - Queue tables
  - session/job status tables
  - last trade lookup tables
- Higher PCTFREE for table reduces # of rows per block
- Adding a row with the RAC Instance Number was used at large Telco, to reduce Block Contention between RAC Instances.

Oracle Memory Sizing – Peak Memory

Memory element:	Memory Allocated	Max Memory (GB)	Linux Guest Size	%Utilized	Swap
SGA – DB1	176 GB	176.00			
PGA – DB1	16 GB	25.50			
ASM - (Global)	1 GB	1.00			
SGA – DB2	10 GB	10.00			
PGA – DB2	4GB	5.50			
Linux Kernal	512 K	0.50			
Linux Page Tables (from meminfo)	17214928 kB	16.40			
Patrol? Other Linux Jobs		0.50			
Swap Allocated					20 GB
Oracle OEM Agent	71 mb	0.07			
CRS Processes	200mb	0.20			
473 User Processes (4.5 MB each)	4.5mb x 473	2.08			
BI DB User Processes (UGA)	40 mb	0.04			
5% Linux Overhead Safety Net		12.60			
	Totals:	250.39	**256 + 20**	0.9936	20

Monitoring Memory

- It's important to Note all Linux platforms use File System Cache which will use all available free memory. This Memory is available though, Linux (feature) does not un-cache the file in case it may be needed.

```
procs -----------memory---------- ---swap-- -----io---- -system-- -----cpu------
 r b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0  0 177252  14856 118800 1736556    0    0     5    26    7    8  0  0 100  0  0   <- First Row should ignore (Since boot time)
 0  0 177252  14856 118808 1736548    0    0     7    21   12   18  0  0 100  0  0   <- Watch si (swap), some so OK
 0  0 177252  14856 118808 1736548    0    0     5    16    9   12  0  0 100  0  0
 0  0 177252  14856 118812 1736544    0    0     5    20    8   13  0  0 100  0  0
```

You can manually free file system cache (not for Prod) with => echo 3 > /proc/sys/vm/drop_caches;
(you lose the benefit of cached files though!)

```
procs -----------memory---------- ---swap-- -----io---- -system-- -----cpu------
 r b   swpd   free      buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0  0 177252 301040       224 1588884    0    0     5    26    7    8  0  0 100  0  0
 0  0 177252 301164       240 1589384    0    0     8    24   16   20  0  0 100  0  0
 0  0 177252 301660       240 1589384    0    0     5    16    8   13  0  0 100  0  0
```


PL/SQL Memory Leak – Patch - 5866410

v$process_memory (for PL/SQL Insert session only)

BEFORE:

CATEGORY	ALLOCATED	USED	MAX_ALLOCATED
SQL	60,080	23,072	438,560
PL/SQL	31,224	26,400	31,224
Freeable	131,072	0	
Other	980,813		980,813

AFTER (Running PL/SQL FOR ALL insert)

CATEGORY	ALLOCATED	USED	MAX_ALLOCATED
SQL	52,096	27,752	691,096
PL/SQL	**977,385,024**	**4,434,584**	**977,389,224**
Freeable	1,048,576	0	
Other	**121,035,317**		822,479,285

********** **AFTER APPLYING PGA PATCH 5866410**
****

BEFORE:

CATEGORY	ALLOCATED	USED	MAX_ALLOCATED
SQL	60,080	23,072	323,080
PL/SQL	31,224	26,400	31,224
Freeable	196,608	0	
Other	915,277		915,277

AFTER (Running PL/SQL FOR ALL insert):

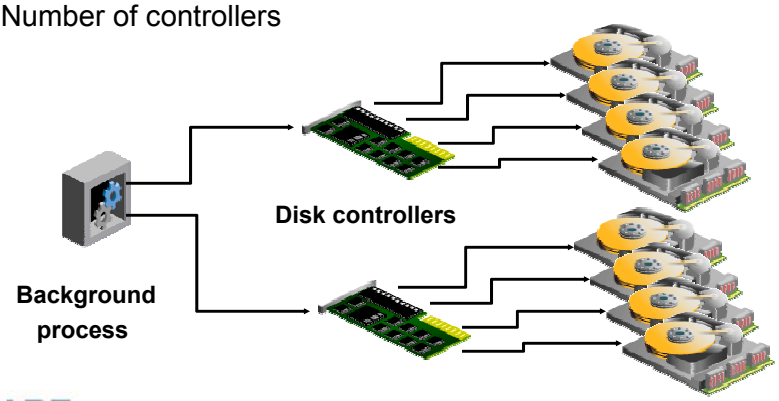
CATEGORY	ALLOCATED	USED	MAX_ALLOCATED
SQL	37,968	15,184	329,568
PL/SQL	**6,181,248**	**4,548,968**	**6,185,448**
Freeable	1,048,576	0	
Other	**816,677**		702,739,117

Meeting the I/O Challenge

- I/O avoidance
  - Larger server memory cache
  - Larger I/O subsystem cache
  - Application redesign
- More Efficient I/O
  - Data Layout Strategies
  - Data Compression
  - Smarter cache management algorithms
  - Application redesign
- Getting beyond HDD
  - In Memory Database
  - Selective use of Solid State Disk (SSD)

Bandwidth Versus Size

- I/O performance depends on bandwidth.
  - Number of disks, not size
  - Number of controllers



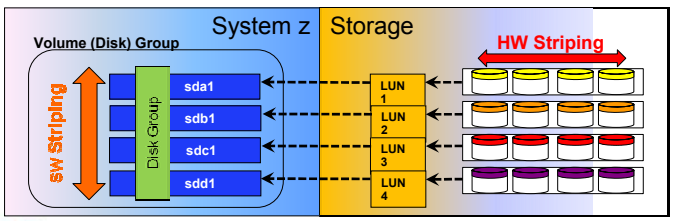
Improving Oracle Database I/O Performance

- Striping Data to avoid I/O hotspots (not so much for XiV)
- Tuning the Database Buffer Cache
- Implementing Data Compression
- Using Solid State Disk

- Using faster HDD based storage subsystems
  - 15k RPM SAS drives
  - Ample read/write cache
  - Enterprise Class cache management algorithms

Data Striping to Avoid I/O Hotspots

- Old Wisdom: **Isolate files based on function and/or usage**
  - Manually intensive effort
  - Leads to I/O hotspots that impact throughput capacity and performance
- New Wisdom: **Stripe objects across as many physical disks as possible**
  - Minimal manual intervention
  - Evenly balanced I/O across all available physical components
  - Good average I/O response time object throughput capacity with no hotspots
- Implementation Options:
  - ASM can do this automatically within a given disk group or filesystem
  - Can be implemented with conventional Volume Managers and filesystems: http://www-1.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP100319



Tuning Oracle DB Buffer Cache

- Buffer Cache is the primary database I/O avoidance option
- Old Wisdom: **If the buffer hit% is > 90% it's good enough**
- New Wisdom: **Depending on workload, a higher hit% may be optimal**
  - For a given workload with a buffer hit% of 98%, a 1% increase (to 99%) will reduce physical I/O requests by 50%
  - Reducing IOPS will also improve response time for remaining I/Os
  - In many cases, adding server memory may be cheaper than adding I/O subsystem cache memory or short-stroking disks
- Evaluate impact of increasing db_cache_size on physical I/O
  - Oracle AWR "Buffer Cache Advisory" report is a good place to start
- Monitor for and address potential impacts:
  - Increased peak CPU demand
  - Increased logical read rates
  - System paging (add physical memory as necessary)

Oracle AWR Buffer Cache Statistics (Real Example)

- **Instance Efficiency Percentages**

Buffer Nowait %:	99.97	Redo NoWait %:	100.00
Buffer Hit %:	98.89	In-memory Sort %:	100.00
Library Hit %:	70.53	Soft Parse %:	26.01
Execute to Parse %:	28.44	Latch Hit %:	99.96
Parse CPU to Parse Elapsd %:	30.81	% Non-Parse CPU:	89.14

Buffer Hit% = 98.89

- **Buffer Pool Advisory Report**

P	Size for Est (M)	Size Factor	Buffers for Estimate	Est Phys Read Factor	Estimated Physical Reads
D	256	0.64	16,080	1.11	97,368,882
D	288	0.72	18,090	1.11	96,868,286
D	320	0.80	20,100	1.08	94,323,210
D	352	0.88	22,110	1.05	91,776,695
D	384	0.96	24,120	1.02	89,228,794
D	400	1.00	25,125	1.00	87,480,193
D	416	1.04	26,130	0.98	85,731,549
D	448	1.12	28,140	0.94	82,232,582
D	480	1.20	30,150	0.90	78,731,330
D	512	1.28	32,160	0.86	75,225,110
D	544	1.36	34,170	0.82	71,715,825
D	576	1.44	36,180	0.78	68,209,778
D	608	1.52	38,190	0.72	63,357,042
D	640	1.60	40,200	0.67	58,494,659

- **Predicts 29 (of 87) million block reads could be eliminated over 30 minute period by adding 240 MB of buffer pool cache:**
  - **2,000 read IOs /second**
  - **16,000 blocks /second**
  - **125 MB/second**
  - **A 33% savings**

28

AWR Wait Events

- Top 5 Timed Events

	Event	Waits	Time(s)	Avg Wait(ms)	% Total Call Time	Wait Class
1	db file sequential read	188,158	2,757	15	61.8	User I/O
	DB CPU		1,259		28.2	
2	log file sync	20,320	325	16	7.3	Commit
	db file scattered read	6,135	54	9	1.2	User I/O
	read by other session	1,203	13	11	.3	User I/O

3

1. **If "User I/O" events (especially "db file sequential read") account for a high percentage of total Oracle wait time and Average Wait time is high, selective use of SSD may improve performance**
2. **If "log file sync" event appears in top 5 list and Average Wait time is high, the online redo logs are SSD candidates**
3. **The higher the "Average Wait" time for these events, the greater the potential SSD benefit**

SQL Statement Tuning

- AWR "SQL ordered by Reads" report is a good place to look for SQL statements that may benefit from tuning
  - **SQL ordered by Reads**
- If desired, semi-automatic tools are available to evaluate SQL statement performance and recommend possible remedial actions

Physical Reads	Executions	Reads per Exec	%Total	CPU Time (s)	Elapsed Time (s)	SQL Idu	SQL Module	SQL Text
51,014,572	27	1,889,428.59	10.62	4512.75	20648.44	cpb5vktsrxxnp		BEGIN APPS.IKN_FS_TRAC_SR_PKG....
49,905,589	1,913	26,087.61	10.39	6777.33	13408.17	9kfwnug5cbszd		SELECT CONTACT_ID FROM IKN_O...
22,956,796	26	882,953.69	4.78	2002.24	9528.41	52rzwx2917kd2		SELECT SUMMARY FROM CS_INCIDEN...
20,329,085	23	883,873.26	4.23	1817.61	8203.42	am77sx96jujqk		UPDATE CS_INCIDENTS_ALL_B SET ...
16,650,797	5	3,330,159.40	3.47	1564.46	9668.13	f9zpsu61dq7cy	SQL*Plus	SELECT wdd.delivery_name ||...

I/O Recommendations:

- Continue to Tune TOP queries in AWR reports for I/O reduction
- Continue to Monitor Average I/O read response times from AWR Reports when > 15ms investigate adding more disk spindles to spread the I/Os out further and or move high read Tablespaces to the fastest storage devices.
- Potentially Add more HBA's channels to allow for more concurrent I/Os to occur and increase queue depth (hardware dependent)
- Investigate Oracle Buffer Cache areas to ensure memory that is not being used (streams_pool_size) is utilized and areas that require more cache and keep buffer pools are adequitly sized.

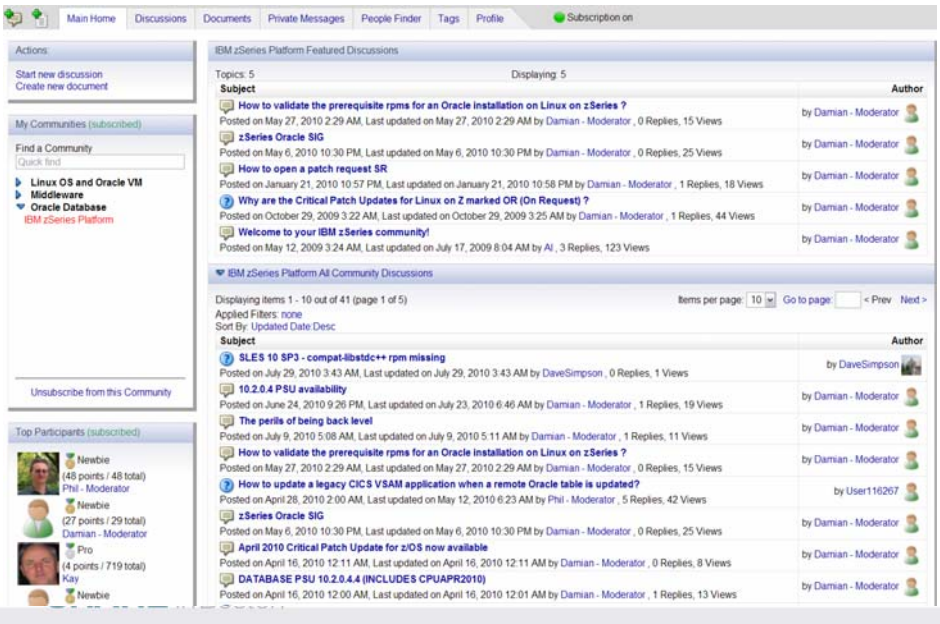
Oracle Data Compression

- Sequential scans typically limited by sequential I/O bandwidth of backend disk storage
- Data Compression increases the number of rows stored per physical data block
- If sufficient CPU capacity is available, the potential increase in sequential data scan rate should correlate to the Compression Ratio (CR)
- Oracle Compression Options:
  - **Standard table compression** CR: up to 4x (2x typical)
    - Only supported for bulk loads/inserts
  - **Advanced Compression (11g)** CR: up to 4x (2x typical)
    - Supports OLTP as well as DW workloads
    - Separately licensed feature

My Oracle Support Community

- Encourages user posts
- Spotlight and highlight posts
- Specific community for System z customers
- Accessed via My Oracle Support (Metalink)
  - support.oracle.com
- Announcements will be made in Community Featured Section !
  - “Webinars” One hour “brownbag” type presentations given by Oracle Support
  - First one planned in December 2010
  - Topic will be ASM
- Join the Oracle zLinux community and help to shape the future:
  - Provide feedback
  - Exchange ideas
  - Get answers
  - Expand networks
  - Share successes

Choose IBM zSeries Platform (Sign on From Oracle Required)



Oracle Recommended Patches as of Aug 4, 2010

- Follow Oracle Support Note: **756671.1** for the latest recommended patches
- Apply 10.2.0.4 (Patch **6810189**) and then latest PSU Patch **9654991 (10.2.0.4.5)**
  - http://updates.oracle.com/ARULink/PatchDetails/process_form?patch_num=6810189 (10.2.0.4)
  - http://updates.oracle.com/ARULink/PatchDetails/process_form?patch_num=6880880 (OPatch 10.2 Update)
  - http://updates.oracle.com/ARULink/PatchDetails/process_form?patch_num=9119284 (10.2.0.4.4)
  - http://updates.oracle.com/ARULink/PatchDetails/process_form?patch_num=9654991 (10.2.0.4.5)
- **CRS/RAC**, patch to 10.2.0.4 and apply the latest **CRS Rollup Patch 9294403**
  - http://updates.oracle.com/ARULink/PatchDetails/process_form?patch_num=9294403
- 10.2.0.5.2 Grid Control Patch Set Update (PSU) 986539.1
  - http://updates.oracle.com/ARULink/PatchDetails/process_form?patch_num=9162498
- Patch Operating Systems to Latest Levels SLES 10 SP3 or RHEL 5.4 or 5.5
  - e.g. Oracle **Bug# 8666385** - CRSD.BIN CORE DUMP FREQUENTLY
  - at least glibc-2.5-34.el5_3.1.s390x.rpm (Red Hat)
  - at least glibc-2.9-13.3.1.s390x.rpm (SLES)