Take the Guesswork Out of Database Layout and I/O Tuning with Automatic Storage Management

An Oracle Technical White Paper
December 2005

ORACLE

Take the Guesswork Out of Database Layout And I/O Tuning with Automatic Storage Management

Take the Guesswork Out of Database Layout and I/O Tuning with Automatic Storage Management

Storage arrays or disk drives fundamentally perform two basic functions: seek and transfer data. The challenge is to find the right formula and compromise to minimize seek time and maximize data transfer that benefits all workloads.

EXECTUTIVE SUMMARY

Database layout and I/O performance tuning is considered by some to be an art vs. science and for others, still a mystery. Oracle is in the leading edge of database technologies and continues to innovate and bring more sophistication to meet the growing demand of enterprise customers. Simplification and automation of multiple technology layers is critical to success without compromising functionality nor service layer agreements expected by our customers.

The Automatic Storage Management (ASM) feature in Oracle Database 10g takes the guesswork out of database tuning and contributes it's fare share to simplify the storage management stack as well as provide a cluster file system and volume management functionality that delivers the optimal database I/O performance automatically.

Often the database, system and storage administrators are challenged to make very difficult decisions when designing database. Database layout strategies, maximizing performance while minimizing overhead, maintaining flexibility and minimizing cost and designing different workload environments are typical questions that often need to be addressed. Getting to the right answer usually requires not only DBA knowledge but also server and storage administrative expertise as well. The necessity to involve multiple personnel and expertise further impedes and complicates database deployment.

The goal of this paper is to discuss why database layout and I/O tuning does not have to be a rocket science anymore. Automatic Storage Management is Oracle's solution that incorporates the knowledge gained by many years of research at Oracle and it's partners as well as customers. Simply deploy ASM and realize the optimal configuration and performance irrespective of different workload types and storage variations that you may have in your computing environments. Follow the simple principles discussed in this paper and never worry about making the wrong decision when faced with many tough configuration questions.

STORAGE TECHNOLOGY ANALYSIS

Oracle Database I/O Sub System Fundamentals

The Oracle I/O sub system services and its clients are designed to perform 1MB I/O when possible or I/O equal to the block size specified by DBA to optimize performance.

Kernel Service Direct File (ksfd) I/O is the central module providing disk I/O services to all its clients. The main interfaces provided allow a client to create database files and perform I/O operations on them. Every read/write operation is done in multiples of the block size that is specified at file creation time.

- Operating System Dependent (OSD) interface is used to access raw and file system files (used with ASM when ASMLIB is not configured) by default
- ASMLIB is a disk access interface defined by Oracle and implemented by storage/system vendors. Oracle currently has an ASMLIB implementation for Linux only. ASMLIB provides I/O enhancements for persistent binding, I/O optimization and ease of management
- Oracle Disk Manager (ODM) – Interfaces defined by Oracle and implemented by storage/system vendors – not discussed in this paper

Both synchronous and asynchronous interfaces are provided to do I/Os to files. Asynchronous I/O is enabled for the server by default.

Oracle I/O Clients Perform 1MB I/O when possible

Various I/O clients and their usage model is discussed in this section to provide insight to I/O block size and parallelization at the Oracle level

The log writer writes the very important redo buffers into log files. These writes are sequential and synchronous by default. The max size of any I/O request is set at 1MB on most platforms.

Redo log reads are sequential and issued either during recovery or by logminor or log dumps. The size of each I/O buffer is limited to 1MB in most platforms. There are two asynchronous I/Os pending at anytime for parallelization.

DBWR is the main server process that submits asynchronous I/O's in a big batch. Most of the I/O request size are equal to the database block size. DBWR also tries to coalesce the adjacent buffers in the disk up to a max size of 1MB whenever it can and submits them as one large I/O.

ASM is one of the clients of ksfd. This module calls ksfd to issue I/O when ASM feature is enabled and there is no ASMLIB library present. A database file is identified as an ASM managed file from the name pattern.

The Kernel Sequential File I/O (ksfq) provides support for sequential disk/tape access and buffer management. Ksfq allocates 4 sequential buffers by default. The size of the buffers is determined by *db*file_direct_io_count parameter set to 1MB by default. Some of the ksfq clients are Datafile. Redolog file, RMAN, Archive log file, Datapump, Data Guard and the File Transfer Package.

ASM Encapsulates The SAME Methodology

ASM extends the SAME principles to further add value to the Oracle GRID architecture and vision:
- Automates and simplifies tedious tasks
- Increases storage utilization
- Improves IT service level agreement for performance and availability

Oracle's Stripe And Mirror Everything (SAME) methodology was first presented at Oracle Open World in 1999. The SAME technical white paper can be found at http://www.oracle.com/technology/deploy/availability/pdf/oow2000_same.pdf. The SAME concepts were introduced to simplify storage configurations for Oracle databases and achieve optimal performance with minimal overhead. These principles largely hold true even today.

ASM is essentially the implementation of the SAME methodologies that was introduced to Oracle customers as a feature of Oracle Database 10g in 2004. It is a storage manager with volume management, file system and cluster capabilities that provides a unique way of evenly distributing data across all available storage for even distribution and best I/O performance maintained over time.

ASM (consistent with the SAME methodology) optimizes for sequential throughput and random access without making tradeoffs. This section reviews the technical reasons supporting this hypothesis.

Disk drive capacity, rotational speed, seek time and transfer rate are important properties that control and limit their performance. Disk speed improvements have been slower relative to network bandwidth and CPU speed creating further potential bottlenecks at the disk I/O level.

A simple way to increase disk bandwidth is to maximize disk drives (sometime referred to as spindles) and spread the data evenly across all. Very efficient disk striping is possible at the ASM level to allocate space evenly across disks. The striping strategies will be discussed in more detail later.

Maximizing sequential performance

ASM chooses a 1MB strip size (coarse striping) to achieve the delicate balance between high though put and efficiency. ASM also provides fine striping of 128KB best suited for redolog data types.

Future releases of ASM will have performance regions or zones defined to best take advantage of the disk drive physics and simplify optimization.

A key observation was made in the SAME research. To maximize sequential performance, we only need to make sure that Transfer_Time > 5*Positioning_Time. The SAME research demonstrates that a sequential transfer that is 1 MB or larger achieves a good balance for high throughput and efficiency. Small sequential access is much less efficient and larger than 1MB access improve performance by a small factor.

The SAME research also shows that the transfer rate of a disk drive varies by a factor of two from the inner edge to the outer edge of the drive (i.e. 11-22 MB/sec transfer rates). Placing most frequently accessed data towards the outer edge, and less frequently data towards the inside half achieves close to optimal sequential performance.

Optimizing random access

The key to optimizing random access is to limit the length that the disk head moves (seeks) between data and access. The time to position the disk head is dominated by the rotational delay (1ms to seek, 3ms to rotate) for small seek distances. The SAME

research concludes that it is enough to position data that is frequently accessed roughly in the same half or quarter of disk drive.

Therefore, placing frequently accessed data towards the outside half of the disk achieves most of the performance benefits. If the minimum length seek time is 4ms, the average read time for a random access read within the outer half of disk is approximately 5.3ms which seems to be a great compromise.

Optimizing disk bandwidth recommendations

ASM implements all features consistent with SAME methodology to be aligned with the following recommendations:

- ASM stripes data to equalize the workload across disks and eliminate hot spots
- ASM implements 1MB striping to achieve high sequential bandwidth
- ASM will take advantage of LUNs created from the outer half of disks to provide fast transfer rates and to minimize seek overhead

WORKLOAD DISCUSSION

ASM is the best file system and volume management of choice for all workload types.

ASM eliminates the need to get into the details and complexities of designing and customizing for different workloads since the underlying disk technology is independent of theses factors. Different workload types such as OLTP, DSS, batch and on-line users depend on Oracle DB operation types (scan, lookup, load, insert, etc) and file types (data, log, temp, archive, undo, backup, etc).

ASM is an ideal environment for all workloads because it spreads the I/O evenly across all disks. The following are some key attributes that will help explain how ASM satisfies these requirements.

Sequential I/O is treated specially by Oracle.

Oracle Database issues large I/O operations that span block boundaries for sequential I/O types. The best way to achieve large I/O operation is to use ASM which provides the 1MB extent striping. Set db_file_multiblock_read_count parameter to 1MB and the OS I/O size limits should be set to at least 1MB.

ASM ensures best possible execution for asynchronous read ahead operations.

Use ASM to achieve maximum *direct IO* throughput allowing high parallelism for the scan and load operations to use the bandwidth of multiple disks with 1MB I/O extents. *Direct IO* bypasses the Oracle buffer cache and do large scan and load operation directly to disk.

ASM optimizes Oracle parallel execution.

ASM should be deployed to ensure that all subsets of database data are spread evenly across all available disks. Oracle automatically parallelizes many operations including scans, join, sort, hash, load, create index, etc placing a high demand on the I/O sub system. The conventional database design typically configured to have subset of data spread over just one partition limiting scalability.

ASM allows fine grained (128kb) striping for log files to parallelize at the disk level

The on line redo log file writes are serial and can not be parallelized at the Oracle level. Therefore, it is a best practice to spread the log file writes on as many disk as possible to get disk write parallelization. ASM provides 128kb striping within the

same disk group as other files are stored and get the benefit of maximizing the number of spindles and optimal performance. Archive log file writes however can be in parallel and does not have the same limitation.

STORAGE OPTIONS

Mirroring (RAID1) vs. Parity Protection (RAID 5)

The tradeoffs between RAID 1 and 5 are well known in the industry. To put it simply, RAID 1 has the best performance but require twice the storage capacity. RAID 5 is a much more economical solution but with a performance penalty. With RAID 5, data is protected thru the 'parity' information and both data and parity are striped across all disks. If a dive fails, the original data can be reconstructed using the surviving drives and the parity.

Calculating and writing the parity is the problem since it requires at least four I/O writes for each I/O. Therefore, the write performance in RAID 5 suffers the most. Although, the RAID 5 implementations have come a long way, however, the performance is vastly different from one storage array product to another and caution should be exercised when choosing a storage platform. The ORION tools (http://www.oracle.com/technology/deploy/availability/htdocs/lowcoststorage.html) from Oracle can be used to help determined the pro/cons of arrays for your application.

The general rule of thumb is to deploy RAID 5 where cost of storage is critical, performance is not the primary goal, and applications with primary read operations such as data warehouse applications. The Oracle Flash Recovery Area Diskgroup can be another good use of RAID 5 where the storage capacity requirement is the highest and predominantly sequential I/O.

To Stripe or Not to Stripe

DBA and storage administrators are typically given too many choices and are often confused about making the right choice for the right application. The usual answer they get from experts is "it depends". Although the SAME methodology from Oracle streamlined the choices and made it easier, however, the customer was still faced with several difficult decisions to make in terms of using storage array hardware RAID 0 (striping) or a server based logical volume manager and unclear about what to do about stripe configurations within each layer.

ASM striping alone or hardware RAID 0 with ASM are equally good alternatives.

For all the reasons stated in this paper, using ASM striping is the right answer since it automatically performs two of the key recommendations mainly 1MB striping and constant distribution of data for optimal performance.

ASM striping with hardware RAID 0 is equally good based on internal testing and customer benchmarks if configured correctly. Configuration best practices is discussed in the next chapter. We strongly recommend against using Logical Volume Manager (LVM) with ASM since it would a complete duplication of functionality, overhead and complexity.

The storage arrays today typically are designed to perform striping in small stripe widths ranging from 64k to 512kb and some will do 1MB or higher. The smaller than 1MB stripe size is inefficient since it breaks down the 1MB I/O operation into multiple smaller I/Os and less efficient because of physical disk constraints such as head positioning, latency and rotation speeds. A stripe size that is not in multiples of power of 2 is also inefficient it breaks up the 1MB IO operation and alignment into multiple reads/writes causing excessive unnecessary IO operations.

Storage Sub System Cache – Does it Matter

Maximize your CPU memory vs. bigger RAID cache for better results

Storage vendors typically offer large amounts of cache on the RAID controller (4GB to 64GB or more). RAID cache can have significant impact to increase performance. Writes to cache is immediately acknowledged and reads I/O can benefit specially when combined with read-ahead algorithms anticipating where the next read is going to happen can speedup performance. The dual redundant controllers with cache consistency typically experience a heavy performance overhead.

RECOMMENDED STORAGE CONFIGURATIONS

Oracle has conducted experiments with different storage configurations and database layouts and compared results. Stripe sizes of 256KB to 1MB were used with several layout configurations under different workloads. The full disclosure of these reports are available on http://www.oracle.com/technology/deploy/availability/pdf/ora_cbook1.pdf (Oracle EMC and Veritas Joint White Paper, and http://www.oracle.com/technology/deploy/availability/pdf/SAME_HP_WP_112002.pdf (Oracle and HP White Paper).

The SAME research varied the stripe size depth between 256KB and 1MB and found that the 1MB offered the best overall performance.

CONFIGURATION ALTERNATIVES

This section discusses best practices that may be deployed based on your storage configurations. These are recommendations that provide the best compromise that applies to all environments regardless of storage or workload variances.

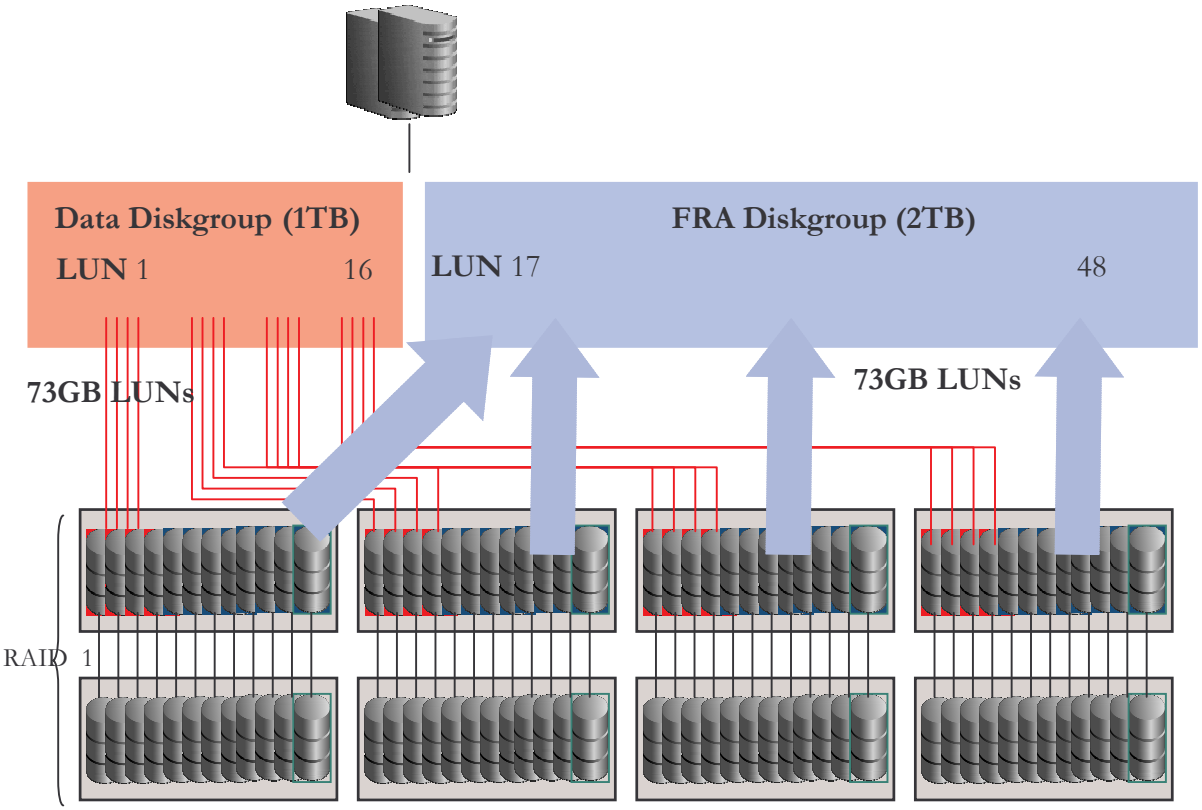
The examples are based on the following resource assumptions:

- Oracle DB size: 1TB database with 2TB Flash Recovery Area
- Storage resource: 8 storage trays with 12x73GB disk per tray

ASM Striping Only

This is an example that provisions entire disk drives and allows ASM to stripe data for optimal performance utilizing a simple storage configuration. Please note that you need to create a partition that defines the whole disk skipping the 1st block, the VTOC disk label.

Four RAID 1 mirrored LUN sets are created from each tray and provisioned to ASM Data disk groups to ensure we are spreading the load evenly across the available storage arrays.



ASM striping only (no hardware RAID 0)

The advantages of this configuration are simple storage configuration, even data load across all storage arrays (4 disks from each array for Data and 8 disks for the FRA), small 73GB incremental growth for database, high IO bandwidth, and no drive contention with other applications or with the Flashback Recovery Area.

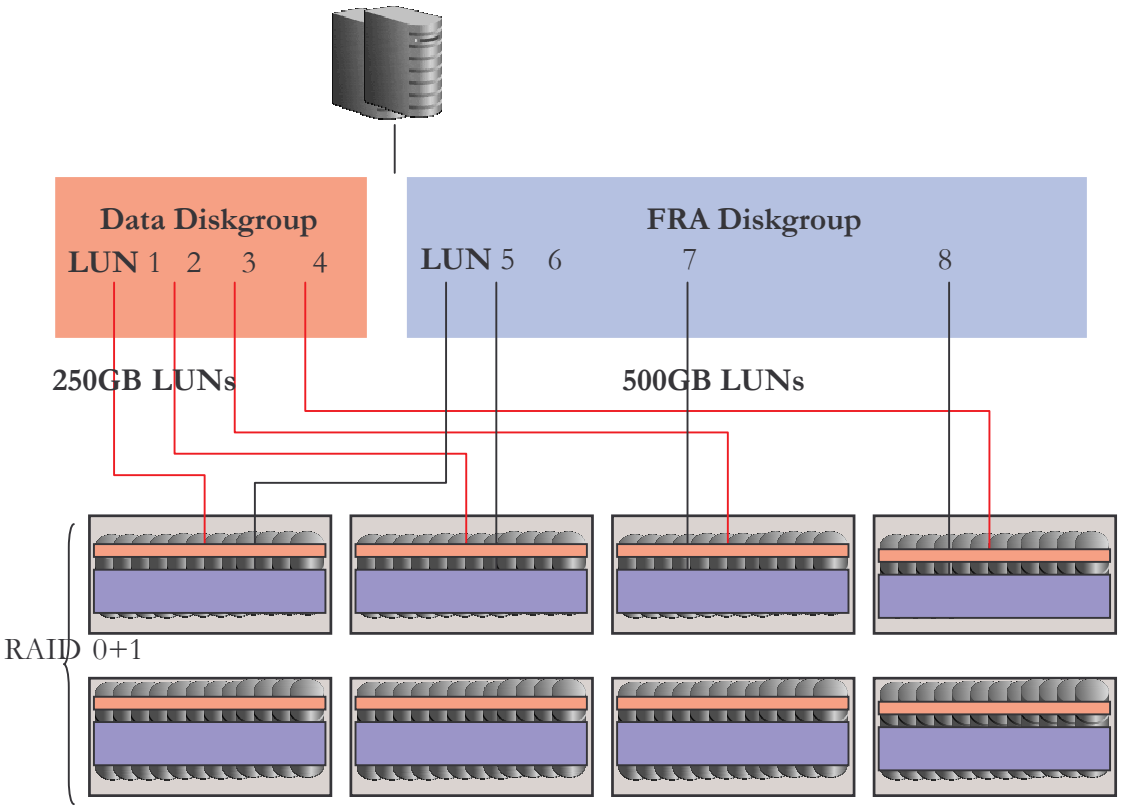
The disadvantages are that each diskgroup is not leveraging all the drives in all arrays, LUN size is limited to the drive size and failure of an array port will cause

both Data and FRA to be unavailable. An alternate higher availability (but lower bandwidth) solution would be to dedicate separate arrays to both the Data as well as the FRA diskgroups to eliminate the dependency.

ASM Striping and Hardware RAID 0

An alternate method is to create hardware RAID 0 striped LUNs and provision those to an ASM diskgroup. The diagram below is based on the same 1TB database and storage configurations stated earlier.

Two striped and mirrored LUNs are created for each array which includes a 250GB LUN for the Data diskgroup and 500GB LUN for the Flash Recovery Area diskgroup. 4x250GB LUNs for Data and 4x500GB LUNs for FRA diskgroups. It is a good idea to create the 250GB LUNs for the Data diskgroup from the outer edge of the disk platter higher performance if necessary.



ASM Striping with hardware RAID 0

The advantages of this configuration are fast regions for the Data diskgroup, well balanced distribution of data utilizing all the disks for higher bandwidth, fewer larger LUNs to create and manage and efficient use of storage capacity available.

The trade offs of this configuration are large incremental growth for a diskgroup (250GB and 500GB incremental growth for Data and FRA diskgroup respectively),

and friendly (as opposed to having contention with other applications with unknown IO patterns) IO contention between the Data and FRA diskgroups. A more highly available alternative would be to create the Data and FRA LUNs using separate arrays or even controllers to reduce the risk of downtime in case one array fails.

Customer Benchmark

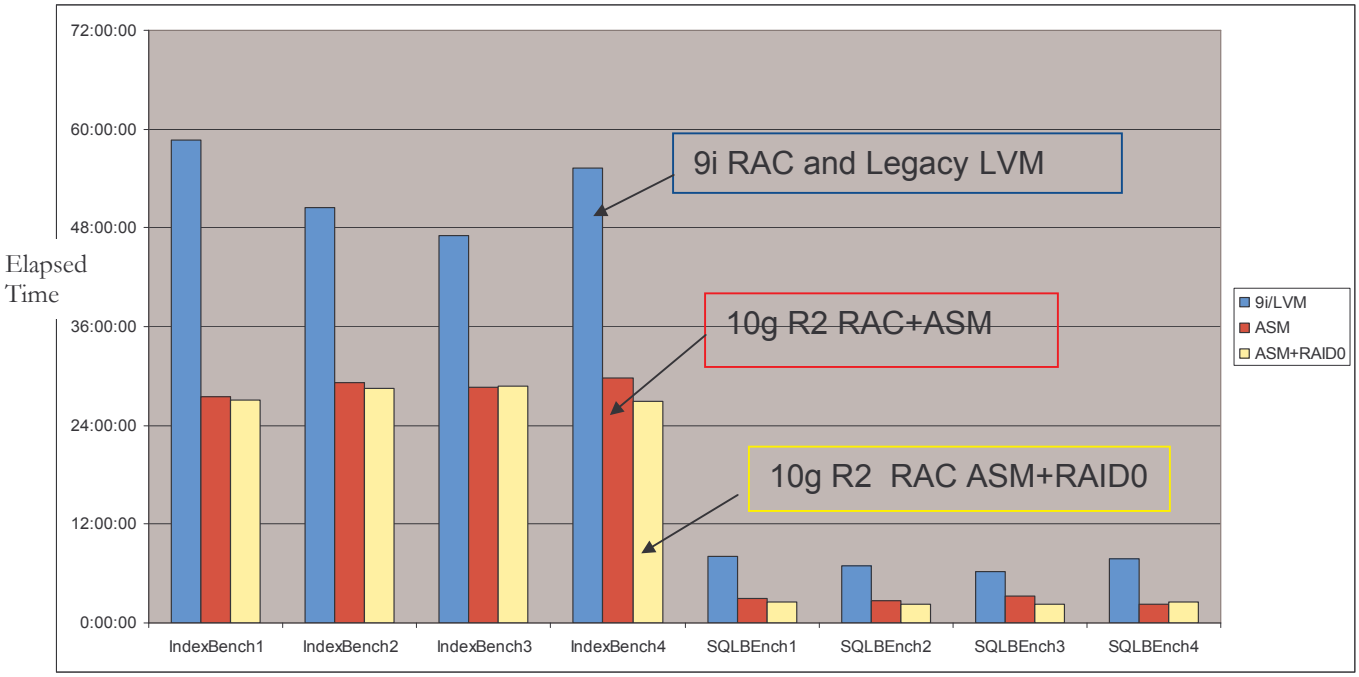
Oracle Database 10g Release 2 with ASM performed up to 200% better than Oracle Database 9i with LVM/RAW

University of Vanderbilt , in the process of evaluation to migrate from 9i to Oracle Database 10g, engaged in a series of performance benchmarks that resulted in measuring 100-200% improvement in performance with Oracle Database 10g and ASM compared to their 9i and legacy LVM environment.

Identical server and storage configurations were used. Two different SM/storage configurations were deployed using EMC Symetrix arrays as follows:

- 16 drives, 4 LUNs and each 4 way striped thru hardware RAID 0 (4 LUNs in an ASM diskgroup)
- 16 drives, each drive=one LUN (16 LUNs in an ASM diskgroup)

The benchmark environment is a very demanding home grown sql-bench exercising most intensive sql queries, and index-bench which included massive drops and re-builds in parallel simulating their production environment.

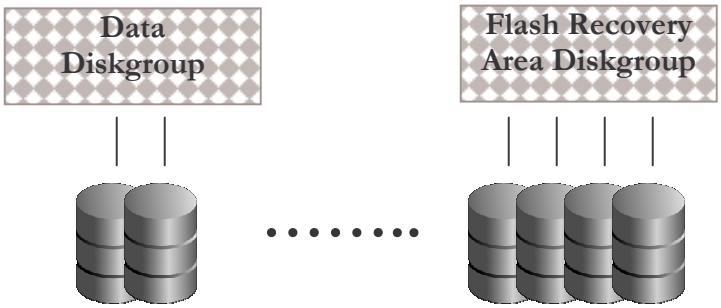


9i RAC/LVM/RAW vs. 10g Release 2 RAC/ASM

The Oracle Database 10g was significantly faster (up to 200%) across the board. The ASM configurations were equivalent with a few % better performance when configured with hardware RADI 0.

RECOMMENDED ASM DISKGROUP CONFIGURATIONS

Oracle recommends two disk groups for a given database. The Data diskgroup contains the data file, log file and control file data types. The Flash Recovery Area (backup) diskgroup stores the backup, temp, archive log, dumpsets and other data types.



A typical ASM diskgroup configuration

It is a best practice to maximize the number of spindles in a given diskgroup while creating the simplest configuration possible. It is a common misperception that separating the redo log files in a separate diskgroup will give you better performance. ASM provides a feature to define 'fine grained' (128KB) striping in ASM templates to accommodate for the special case of small sequential I/O. This method maximizes the spindles and allows for small I/O efficiency.

The notable exception to the 2 diskgroup rule is when you are creating a tiered database structures and need to introduce different capacity of characteristic storage such as low cost ATA storage into your configuration. Mixing different characteristics storage into one ASM diskgroup is not a good idea since it will create an unbalanced data distribution and revert to the lowest common denominator.

Another research study (http://www.oracle.com/technology/deploy/availability/pdf/SAME_HP_WP_112002.pdf) comparing the SAME configuration (similar to ASM) and modified SAME separating the log files into a separate volume shows no advantages when measuring transaction throughput and average I/O response time. The conclusion of this research was:

- The SAME configuration (ASM) dramatically reduces the complexity of initially configuring your Oracle database to optimize I/O
- SAME configuration (ASM) consistently out performs the traditional methods for database layout or nay variations of the SAME configuration

- Isolating the redo log files with their more sequential workload from the remainder of the Oracle datafiles did not produce any significant benefits
- Varying the strip size did not produce any significant changes in the test results for OLTP workloads

ASM BEST PRACTICES AND PRINCIPALS

The following are simple guidelines and best practices when configuring ASM diskgroups:

- Configure 2 diskgroups, one for the Datafile and the other for the Flash Recovery Area (one is a backup for the other for availability purposes)
- LUNs (disk drives of partitions) provisioned to ASM diskgroups should have same storage performance and availability characteristics. Configuring mixed speed drives (i.e. 10k and 15k rpm) will default to the lowest common denominator
- ASM data distribution policy is capacity based. Therefore, LUNs provided to ASM should be the same capacity for each diskgroup to avoid an imbalance and hot spots
- Leverage the storage array hardware RAID 1 mirroring protection when possible to offload the mirroring overhead from the server. Use ASM mirroring (redundancy) in the absence of a hardware RAID or where you need hosted based volume management functionality i.e. mirroring across storage systems. ASM mirroring may be used in geo-cluster configurations where mirroring between remote sites over dark fibre or DWDM. In addition, we have observed that hardware RAID 1 in some ATA storage products are inefficient and degrades the performance of the array even more. Using ASM redundancy has proven to deliver much better performance in ATA arrays http://www.oracle.com/technology/deploy/availability/pdf/lcs_OW.doc.pdf
- Maximize the number of spindles (disks) in your diskgroup for maximum data distribution and higher IO bandwidth
- Create LUNs using outside half of disk drives for higher performance if you need the extra performance boost. Use small disks with highest RMP if possible
- Create large LUNs to reduce LUN management overhead
- Minimize IO contention between ASM disks (LUNs) and other applications when feasible (dedicate disks or arrays to ASM diskgroups that are not shared with other applications)
- Choose a hardware RAID stripe size to be ideally 1TB. If the storage array does not support a 1TB stripe size, then choose the max size up to 1MB which is a multiple of power of 2 (128/256/512 etc)

- Do not use a Logical Volume Manager (LVM) if possible since it would be redundant. There are situations where certain multhipathing or 3rd party cluster solutions would require the use of an LVM. In this case, use the LVM to represent a single LUN with not striping or mirroring to minimize the performance impact (you would still have an undesired LVM layer which complicates the management stack)
- Use the Oracle ASMLIB feature that is available for Linux to address the device naming and permission consistency in between re-boots. This is a great management feature that reduces admin overhead considerably specially in larger cluster configurations

CONCLUSION

The right answer is always ASM! when choosing a volume manager, a file system or raw partitions to deploy an Oracle Database 10g database regardless of the underlying storage configuration and application workload characteristics. Simplicity and automation are even more desirable in RAC and GRID computing environments.

Automatic Storage Management takes the guesswork out of database I/O layout and tuning:

- ASM provides the best compromised single solution for all scenarios and dramatically reducing the management overhead. You no longer have to worry about making the wrong choices and decisions for customizing your storage to your database application
- ASM minimizes disk contention by dynamically distributing I/O across all disks
- ASM implements 1MB striping which has demonstrated optimal performance for all workloads
- ASM consolidates all your Oracle data into two diskgroups improving storage utilization while eliminating wasted space significantly
- ASM saves you money specially in cluster and GRID environments eliminating the need for costly 3rd party cluster volume management and file system solutions

ORACLE

Take the Guesswork Out Of Database IO Tuning
With Automatic Storage Management
December 2005
Author(s): Ara Shakian
Contributing Authors:
Optimal Storage Configuration Made Easy
Configuring Oracle Databases with Veritas and EMC
SAME and HP XP512
Low Cost Storage Initiative White Paper
Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com