

Overview of I/O Performance and RAID in an RDBMS Environment

Edward Whalen
Performance Tuning Corporation

Updated May 2005

Abstract

This paper covers the fundamentals of I/O topics and an overview of RAID levels commonly found in an RDBMS environment. This paper will cover both the performance and fault tolerant properties of those RAID levels. By understanding the base concepts of I/O performance and technology used in your system you will be better able to size and configure your system. By understanding the various RAID levels you will be better able to design and configure your own system using the most appropriate RAID level for the type of activity being done.

This paper first starts with an overview of disk drive performance followed by a survey of RAID levels including RAID 0, RAID 1, RAID 0+1 and RAID 5. Once you have read this paper you should have a better understanding of the I/O subsystem and how it relates to the performance of your system and your database.

Following the overview of RAID and disk performance is a survey of different I/O storage systems including SAN, NAS and direct attached storage. An explanation of these types of storage systems and their performance characteristics will be provided.

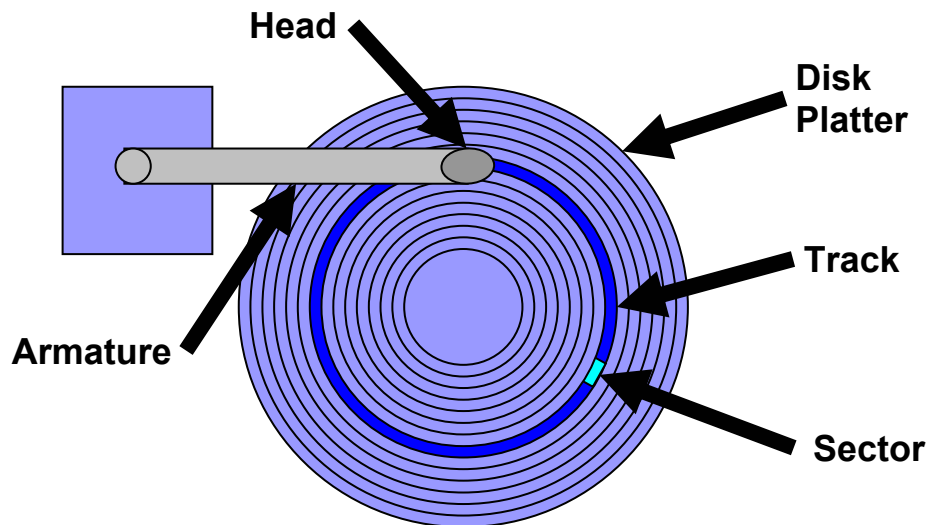
Disk Drive Performance

Disk drive performance is a combination of two components of the disk drive itself, the rotational speed and the seek time. As you will learn, one of these components is typically much more important than the other. This section of the paper will cover both of those topics, but first an overview of how a disk drive works will be covered.

How Does a Disk Drive Work?

A disk drive is made up of a number of disk platters that spin at a very high rotational speed. These disk platters are made up of a number of tracks that store the data, just like tracks on a CD-ROM. The disk drive is made up of multiple platters that ride on top of each other. The tracks that ride on top of each other make up a cylinder. Within each of these tracks are sectors. A sector is typically 512 bytes and it is the sector that holds the data.

Disk Drive Components



The disk drive has an armature that moves in and out of the disk drive in order to move the magnetic heads over the cylinders where the requested data is stored. The act of the armature and heads moving in and out of the disk drive to where it is needed is known as seeking. The time it takes for the heads to move from where they currently were, to where they need to be is known as the *seek time*.

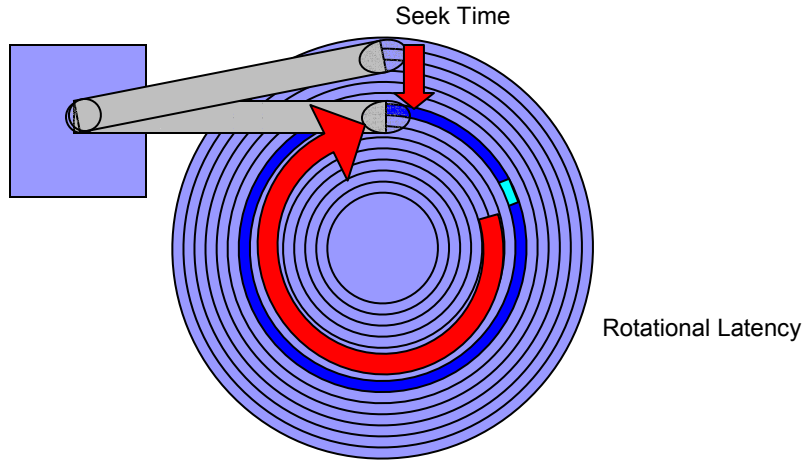
Disk Drive Performance

Once the heads move over the cylinders where the desired data is held it must wait for the desired sector or sectors to rotate under the head. This wait time is known as *rotational latency*.

It is the sum of these two times that make up the time it takes for a single I/O operation to complete. Of course there are other components that take time as well, such as electrical transfer time, etc. but they are minor compared with the seek time and rotational latency.

So, the important components of disk drive performance that make up the time it takes to perform an I/O operation are the seek time and the rotational latency. You might be wondering by now that type of times are we talking about.

Disk Drive Seek Time and Rotational Latency



Let's look at the top of the line 32GB SCSI disk drive. According to the specification sheet here are some of the statistics.

Statistics	Value
Rotational Speed	15,000 RPM
Average Seek Time	4.2 ms
Full Disk Seek	8.4 ms
Track to Track Seek	0.6 ms
Average Rotational Latency	2.0 ms

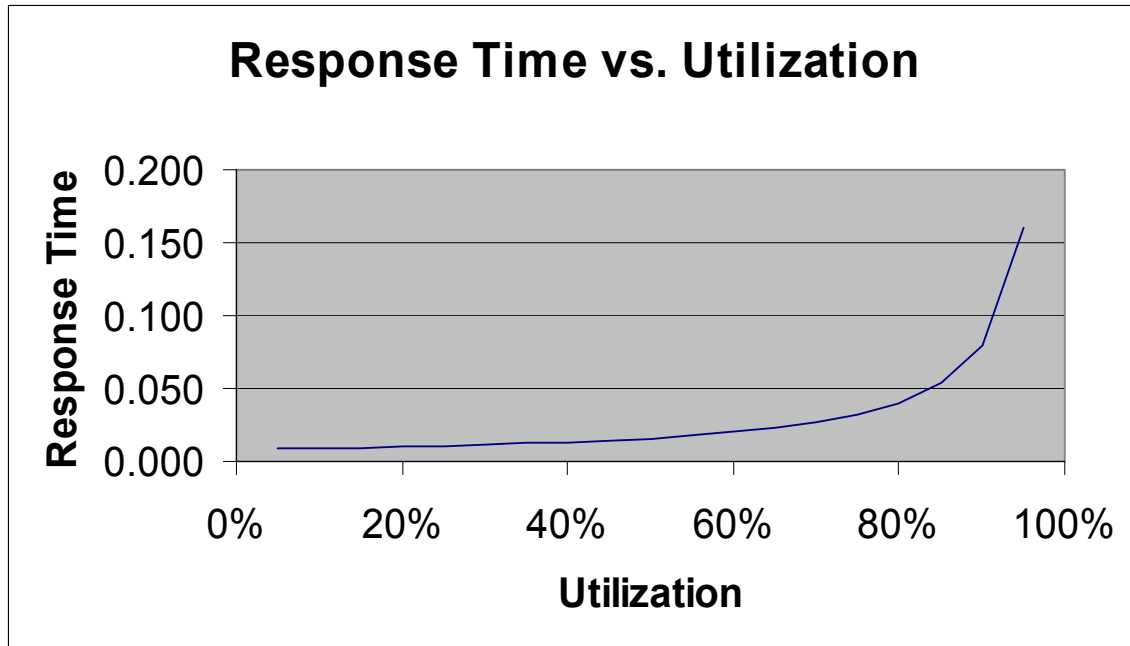
From these statistics we can determine the number of I/Os per second that a single disk drive can perform. The number of I/Os per second is determined by the following formula:

$$\text{I/Os per second (IOPS)} = 1 / \text{seconds per I/O}$$

Where the seconds per I/O = seek time + rotational latency

$$\begin{aligned} \text{IOPS} &= 1 / (4.2\text{ms} + 2.09\text{ms}) \\ &= 1 / (6.2 \text{ ms}) = 1 / (0.0062 \text{ ms}) \\ &= 161 \end{aligned}$$

What happens if you exceed this 161 I/Os per second. If you exceed the rate of 161 I/Os per second, the I/O latency, or time it takes on average for an I/O to complete will increase. The following graph illustrates the relationship between percent utilization vs. increased latency.



A disk drive that is overdriven, or is pushed to hard will cause increased response times which will drastically affect the performance of your RDBMS. So it is important to make sure that you have enough disk drives in your system to support the required I/O rates without exceeding this performance.

As you can see from the above chart, you really should not exceed 80% of the maximum capacity if you want I/O latencies to be within reasonable levels. So, for this 15,000 RPM disk drive, optimal I/O rates are 128 IOPS.

With sequential I/Os the disk armature and heads don't have to move much and much higher I/O rates can be achieved. Instead of 4.2ms seek times for random I/Os, the track-to-track seek time is on the order of 0.6ms. In addition, entire tracks are read without incurring any seeks at all. With sequential I/Os it is not uncommon to do 250 IOPS without experiencing any increased latencies.

Overcoming Disk Drive Performance Issues

So how do you overcome this limitation of 128 IOPS. The answer is very simple, since most I/Os in a database environment are random, by adding disk drives you should be able to achieve higher I/O rates. For example, if you don't want to run your disk drives more than 100 IOPS, 10 disk drives can run at 1,000 IOPS (assuming random I/O).

But how do you configure your system to use multiple disk drives without having to spend a great deal of time balancing I/Os across the various disk drives. The answer is RAID.

RAID Systems

Managing a large number of individual disk drives can be very difficult, since you must balance the database files across all of these disk drives in order to spread out the I/O load. In order to simplify this task, provide for optimal performance and to provide a fault tolerant system, RAID was developed. RAID stands for Redundant Array of Inexpensive Disks.

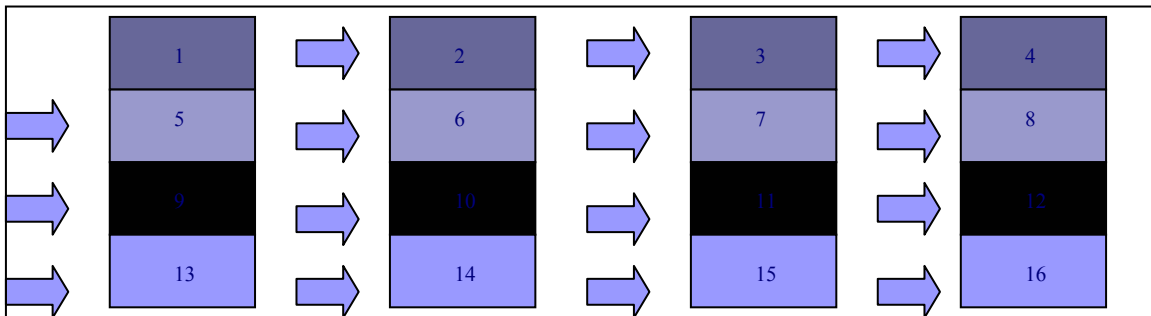
RAID systems are very configurable and can be configured in different ways, depending on what your needs are. These different configurations have different performance and fault tolerant properties and are known as RAID levels. The different RAID levels work differently but essentially serve the same purpose, to create a *logical disk drive* out of two or more physical disks.

A logical disk drive or *logical volume* looks to the operating system and RDBMS like a disk drive, but in reality might be the combination of many disk drives. RAID volumes are combinations of multiple disk drives configured in a RAID array to provide the desired performance and fault tolerant properties.

RAID 0

RAID 0 is considered a RAID level even though there are no redundant properties associated with this RAID level. A RAID 0 takes a number of disk drives and stripes them into a larger logical volume. By using RAID 0 you can combine or *stripe* multiple disk drives into what appears to the operating system as a single large disk drive.

RAID 0 works by taking the data in the logical volume and striping that data across the array. The data in the logical volume is broken down into what are known as *chunks* or *stripes* (depending on the vendor). These chunks are typically 64K, 32K or configurable in size. The chunks are then allocated to the physical disk drives in a round-robin fashion as shown here.



RAID 0 Volume

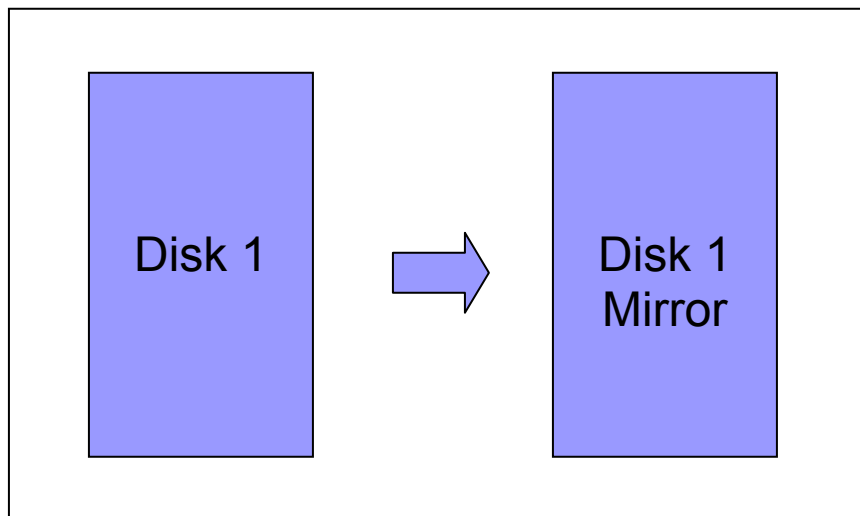
There are advantages and disadvantages to RAID 0.

RAID 0 Advantages	RAID 0 Disadvantages
No overhead from RAID processing. So maximum performance is reached.	No fault tolerance. If a single disk drive were to fail all data would be lost.
All disk space is used.	

In an RDBMS environment we never recommend using RAID 0. In the event of a disk failure (and disk failures are probably the most likely type of failure to occur) all of the data or programs would be lost and you must recover from backup.

RAID 1 and RAID 0+1

RAID 1 is known as mirroring. With RAID 1 the entire contents of your disk drive has an exact copy on another disk drive, known as the mirror. With RAID 1 a disk drive failure is transparent to the user. If a disk drive were to fail, the mirrored disk drive immediately takes over. If a spare disk drive has been configured into the system it will immediately begin to copy data from the surviving disk drive in order to assume fault tolerance. The term *fault tolerance* refers to the fact that the system can tolerate a fault, such as the loss of a disk drive and continue processing seamlessly.

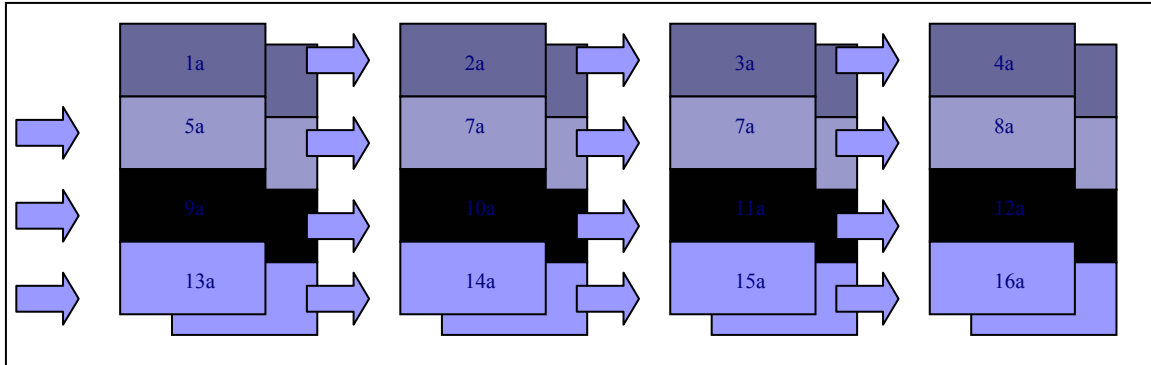


RAID 1 Volume

As with RAID 0 there are advantages and disadvantages.

RAID 1 Advantages	RAID 1 Disadvantages
Excellent fault tolerance. RAID 1 can tolerate the loss of a disk drive.	RAID overhead. When writing to the RAID 1 volume two physical I/Os are required, one to each disk.
Read performance is increased since reads occur on both disk drives.	RAID 1 is expensive since you must double the number of disk drives that you purchase.

RAID 0+1 or RAID 10 are combinations of RAID 0 and RAID 1. With a RAID 0+1 configuration disk drives are mirrored and then striped. Thus you can take advantage of the RAID 0 disk volume where you can increase space and performance as well as taking advantage of the mirroring properties of RAID 1. A RAID 0+1 volume is shown here.



RAID 0+1 Volume

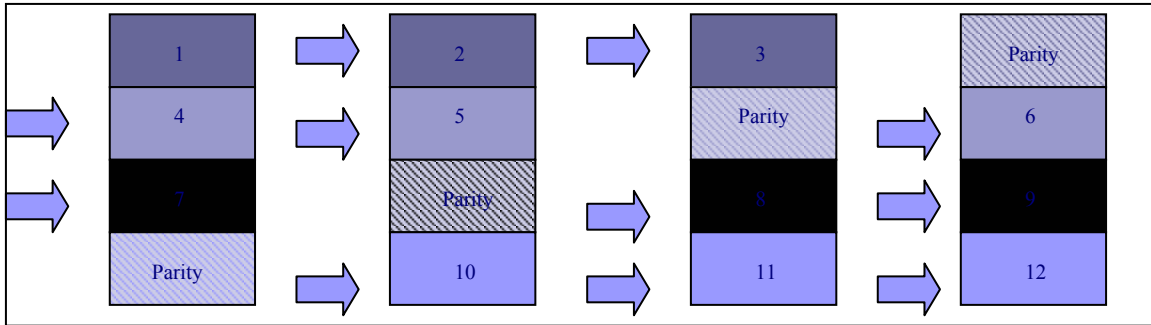
As with the other RAID levels there are advantages and disadvantages to RAID 0+1 or RAID 10.

RAID 0+1 Advantages	RAID 0+1 Disadvantages
Excellent fault tolerance. RAID 0+1 can tolerate the loss of a disk drive or even the loss of all of the mirrors..	RAID overhead. When writing to the RAID 0+1 volume two physical I/Os are required, one to each disk.
Read performance is increased since reads occur on both disk drives.	RAID 0+1 is expensive since you must double the number of disk drives that you purchase.
Striping provides for greater performance since there are multiple disk drives in the RAID volume.	
In the event of a failure performance is not severely degraded. All I/Os are routed to the surviving mirror.	

RAID 1 and RAID 0+1 are the most recommended RAID levels for RDBMS use because of the fault tolerance and performance characteristics.

RAID 5

RAID 5 uses parity for fault tolerance. The advantage of using parity is that instead of having to double the number of disk drives in the system, you only have to add one disk drive to store the parity. RAID 5 uses parity, but distributes the parity among all of the disk drives in the RAID volume as shown here.



RAID 5 Volume

RAID 5 is very popular because it provides a fault tolerant solution at a relatively low cost. For the cost of one additional disk drive fault tolerance is achieved, but this is at a relatively high performance cost.

In order to maintain the parity, when a logical write (a write to the logical volume) occurs a number of steps are required:

1. The parity and data disks must be read.
2. The new data is compared to the data already on the disk drive and changes are noted.
3. A new parity is calculated based on step 2.
4. Both the party and data disks are written to.

So, for a single logical write, four physical I/Os must take place. So when calculating the number of disk drives that are needed in your system you must take into account the additional overhead due to RAID 5.

As with the other RAID levels there are both advantages and disadvantages to RAID 5.

RAID 5 Advantages	RAID 5 Disadvantages
Fault tolerance. RAID 5 can tolerate the loss of one disk drive in the RAID volume.	RAID overhead. When writing to the RAID 5 volume four physical I/Os are required.
Read performance is increased since reads occur on both disk drives.	RAID 5 fault tolerance can only tolerate the loss of one disk drive in the RAID volume.
Striping provides for greater performance since there are multiple disk drives in the RAID volume.	In the event of a failure, performance is severely affected since all remaining drives must be read for each I/O in order to recalculate the missing disk drives data.

RAID Comparison

Each RAID level has its own attributes and performance characteristics as described above. Let's compare those attributes and characteristics.

RAID Level	Read Performance	Write Performance	Fault Tolerance	Cost
RAID 0	Good	Good	None	Low
RAID 1 and RAID 0+1	Good	OK 1 logical write = 2 physical I/Os	Excellent Can potentially tolerate the loss of multiple disks	Highest Requires that you purchase 2x disk drives
RAID 5	Good	Poor 1 logical write = 4 physical I/Os (2 then 2)	OK Can survive the loss of 1 disk at severely degraded performance level	Best for fault tolerance

As you can see, there are vast differences among the most popular RAID levels.

Storage Recommendations

There are a number of recommendations that make sense in an RDBMS environment. These recommendations revolve around the primary goal of database stability and uptime with a secondary goal of cost. Since the RDBMS's are different, recommendations are provided for both Oracle and MS SQL Server.

Oracle Recommendations

Oracle is sensitive to read performance and sensitive to write performance on the redo log files and on the archive log files. Thus, the following recommendations are given.

OS Volume	The OS should be installed on a RAID 1 disk volume. It is important that you do not need to restore/rebuild the OS in the event of a disk failure. This can be very time consuming and expensive. The OS will certainly fit on one disk drive, and RAID 5 makes no sense in a 2 drive configuration, since both RAID 1 and RAID 5 would require the same number of disk drives. In addition, the Oracle binary files can be placed on this volume.
Redo Log Files	The Redo Log files should be placed on a RAID 1 or RAID 0+1 volume. The I/Os to the Redo Log files are 100% sequential and 100% writes, thus RAID 5 is inappropriate.
Data Files	The Data files should be RAID 0+1 if the I/Os are 90% reads or less. If the I/O pattern is 90% or greater reads, then RAID 5 is OK. Again, your budget may help determine this.
Archive Log Files	The Archive Log files can either be RAID 0+1 or RAID 5 depending on your budget. Archiving might take longer if it is RAID 5.

By using RAID fault tolerant volumes, much pain and expense can be avoided in the event of a disk failure.

MS SQL Server Recommendations

SQL Server is sensitive to read performance and sensitive to write performance on the transaction log files and on the transaction log backup volume. Thus, the following recommendations are given.

OS Volume	The OS should be installed on a RAID 1 disk volume. It is important that you do not need to restore/rebuild the OS in the event of a disk failure. This can be very time consuming and expensive. The OS will certainly fit on one disk drive, and RAID 5 makes no sense in a 2 drive configuration, since both RAID 1 and RAID 5 would require the same number of disk drives. In addition, the SQL Server executable files can be placed on this volume.
Transaction Log	The Transaction Log should be placed on a RAID 1 or RAID 0+1 volume. The I/Os to the Transaction Log files are mostly sequential and mostly writes, thus RAID 5 is inappropriate.
Data Files	The Data files should be RAID 0+1 if the I/Os are 90% reads or less. If the I/O pattern is 90% or greater reads, then RAID 5 is OK. Again, your budget may help determine this.
Trans Log Backup	The Transaction Log backup files can either be RAID 0+1 or RAID 5 depending on your budget. Backing up the transaction log is a critical task where performance is important. Keep this in mind when deciding.

By using RAID fault tolerant volumes, much pain and expense can be avoided in the event of a disk failure.

Storage Systems Overview

There are a number of different storage systems available today from many different vendors. Among these storage systems are several primary types including SAN, NAS, DAS, iSCSI and JBOD. After a brief definition of these storage systems a more detailed description is provided.

JBOD	The term JBOD has been around for some time and is an acronym for Just a Bunch Of Disks. A single disk drive or a number of disk drives that use no RAID striping are considered JBOD.
DAS	Direct Attached Storage (DAS) is considered to be a RAID system that utilizes any storage controller that is not part of a network (see SAN and NAS). This includes Fibre Channel Arrays and SCSI Arrays.
NAS	Network Attached Storage (NAS) is a virtual storage device that is available via your Local Area Network (LAN). A windows share or a Novell file share is considered NAS storage, however, this paper will focus more on NAS appliances.

SAN Storage Area Network (SAN) is a storage system that allows a network of systems to access storage over a dedicated storage network.

iSCSI iSCSI is a variant on NAS storage. Although it is storage that is available over the network, it uses the host system's SCSI subsystem to manage the storage.

Let's look at these storage systems in a little more detail.

JBOD (Just a Bunch Of Disks)

As described previously, JBOD disks have no fault tolerance associated with them and are limited to the performance of a single disk drive. Because of both of these limitations JBOD is really not very interesting in a database environment.

DAS (Direct Attached Storage)

Direct attached storage can provide a high performance and fault tolerant environment and thus is often used for database storage. SCSI RAID controllers are an excellent type of direct attached storage, but Fibre Channel storage systems can be configured as direct attached storage. In fact, often the only difference between a direct attached fibre channel storage system and a SAN is the addition of fibre channel switches.

A limitation of Direct Attached Storage (DAS) is that it can only support one or two hosts, because each host must be connected directly to the storage device. Fibre channel storage systems often have 4 or more connections which will accommodate a few hosts, but in order to connect more systems you must use a switch, which turns DAS into SAN.

NAS (Network Attached Storage)

NAS (Network Attached Storage) systems are simply put, storage that is available over the network. Although windows file shares, Novell file shares and NFS shares are technically considered to be NAS storage, I usually think of NAS as an appliance. There are excellent NAS appliances from companies such as Network Appliance, IBM and EMC.

Traditionally NAS systems have been discouraged as database storage, but with performance enhancements both within the storage systems and with NFS and networking NAS systems are an excellent choice for database storage. If you are planning on using NAS storage there are several things that you can do to improve its performance.

Isolate the NAS Network	NAS storage should be accessed over a private network dedicated to storage. It is not difficult or expensive to add network cards to your system.
Use a high speed network	NAS access should only be done over Gigabit or faster networks.
Tune network packet size	If possible; tune the network packet size to be at least the size of a database block (variable for Oracle, 8K for SQL Server).

Minimize network protocols If possible; reduce the number and type of network protocols running on your storage network.

Avoid routers if possible; configure storage near the systems that use that storage and avoid intelligent routers, firewalls, etc. in between the storage and the systems that use them.

If you properly configure and tune your NAS system, you can achieve high performance and stability. There are several advantages and some limitations to NAS storage.

- Oracle RAC on NAS. Oracle RAC is supported and runs very well on NAS. Because of the nature of network attached storage it is unnecessary to use a clustered filesystem such as OCFS (Oracle Cluster FileSystem). Because NAS is filesystem based storage Oracle ASM (Automatic Storage Management) is not supported on NAS.
- Microsoft SQL Server Clusters. Failover clusters are not supported on NAS storage at this time. This is due to the inability of Windows to lock the storage by inserting a cluster layer in the I/O stack.

NAS storage is quite popular and Performance Tuning Corporation has installed multiple RAC clusters using NAS storage.

iSCSI (internet SCSI)

The iSCSI storage system has most of the features and benefits of the NAS system with the exception that it looks like a disk drive rather than a network drive. This can offer an advantage. Since iSCSI looks like a disk drive, Microsoft Clusters and Oracle RAC with ASM can be configured on it. In addition, since iSCSI looks like a disk, you can monitor the performance of the iSCSI disk along with other disk drivers in addition.

iSCSI storage should be treated as NAS in terms of optimization and configuration. Use an isolated fast network and avoid additional overhead. As with all storage, the underlying RAID and disk configuration will also affect performance.

SAN (Storage Area Network)

A SAN is a network that is dedicated to serving storage. Storage Area Networks are characterized by fibre channel controllers attached to fibre channel switches attached to a SAN storage controller. SAN storage controllers are available from a variety of vendors including EMC, HP, IBM and others.

The SAN network is dedicated to serving storage and utilizes a low overhead protocol and a high speed network. This combination provides a high level of performance and fault tolerance. Every component in a SAN system can be made redundant, including the controllers and switches.

As with the NAS system, the SAN storage system supports Oracle RAC, however, with SAN you have the choice of a clustered filesystem such as OCFS, raw devices and Oracle

ASM. SAN storage also supports Microsoft clustering as well as Linux clusters and Veritas clusters.

Storage Consolidation

Both NAS and SAN storage offer benefit of storage consolidation. Storage consolidation allows one storage system to provide storage to a number of client systems. This allows for storage to be administered in a central location and to be allocated on an as-needed basis.

Storage consolidation can be both a benefit and a curse. Because individual disk drives can be shared within a NAS or SAN system, your I/O performance can be degraded by other systems accessing the same disk drives as you. In addition, low-end SAN systems often don't have enough CPU power to service all of the incoming requests, so beware.

Recommendations

I recommend either NAS or SAN storage for your database environment. Whether you choose NAS or SAN should be determined by your individual needs, your budget and the relationship that you have with your storage vendors. Just a few years ago I would never have recommended NAS storage for databases, however, with the introduction of Gigabit networks, improved NFS and TCP/IP performance and high performance NAS systems I will now recommend NAS.

Conclusions

This paper has provided information on the basic concepts of disk drive technology and what affects I/O performance. In addition, the most popular RAID levels have been explained. As you have learned, there are a number of tradeoffs involved in configuring your system. It really comes down to a tradeoff between performance and price. By choosing the best solution for your needs, you can avoid costly downtime and avoid the possibility of losing data.

Remember that the use of a RAID system is not an excuse for not doing backups. The RAID volume will save you from the loss of a disk drive. It will not, however, save you from a mistake by a user or a software error. It is important to perform regular backups and periodically testing those backups.

In addition, several different storage system types were introduced and some of the characteristics of these systems was presented. Regardless of whether you choose direct attached, NAS or SAN storage, the fundamental concepts of disk performance and RAID applies. Careful planning and sizing can save you a lot of performance problems later.

About the Author

Edward Whalen is CEO and founder of Performance Tuning Corporation (www.perftuning.com). Performance Tuning Corporation provides database performance

tuning, load testing and troubleshooting services on MS SQL Server. Edward Whalen was a co-author on for SQL Server books from Microsoft Press;

- SQL Server 7 Administrator's Companion,
- SQL Server 7 Performance Tuning Technical Reference,
- SQL Server 2000 Administrator's Companion
- SQL Server 2000 Performance Tuning Technical Reference

Edward Whalen has also authored four Oracle books;

- Oracle Performance Tuning and Optimization (Oracle7)
- Teach Yourself Oracle8 in 21 Days
- Oracle Performance Tuning (Oracle8/9i)
- Oracle10g Linux Administration (2005)

Edward Whalen is considered a leader in database performance tuning.