Technical Case Study:

RAID & SAN for Storage - Infrastructure Management Services

This is subsequent to Summary of Projects –

VMware - ESX Server to Facilitate:
IMS, Server Consolidation, Storage & Testing with Production Server

PRIMA - Panel for Remote Infrastructure Management Applications

VAssure | Virtualization Labs | trRIMS | Offshore-QA | BI | Portals

http://www.vassure.com

26 Nagarjuna Hills Panjagutta Hyderabad - 500 082 AP India

INTRODUCTION

VMware ESX Server was designed for high performance,server consolidation, Disaster recovery and its architecture is streamlined to provide high-speed I/O. In this paper, we focus on one component of ESX Server's I/O architecture, its storage subsystemsuch as RAID array.

The use of RAID, a Redundant Array of Inexpensive Disks, up until a few years ago was pretty much limited to servers and high end workstations; this was primarily due to the cost of the controller and the accompanying hard drives. Today that's not at all the case! Most of the newer motherboards provide one or more onboard RAID controllers capable of delivering configurations up to and including RAID 5. With the cost of disk storage at an all time low the two primary barriers to using this once esoteric form of data storage have been lifted.

In 2004 onboard RAID controllers were just beginning to appear; today they're commonplace and considerably better. In 2004 SATA drives were not the standard as they are today.

My goal is to give our readers from beginner to expert a guide that will benefit them. The beginning users should come away with an understanding of all the intricacies of RAID enabling them to make an informed decision whether RAID is an appropriate modality for their use. The more advanced users should garner additional information helping them decide if moving to a different level of RAID in hopes of reaching some predetermined level of performance or redundant protection is feasible. Most important of all, I want this to be a positive learning experience for both you and me!

Rather than trying to fix something that's not broken, I plan through research to take what has already been written, coupled with my own thoughts and experiences and update it to a very thorough, yet easy to read guide to RAID that's applicable to 2006. Initially I plan for the guide to have parts.

Part 1, will be function and procedure oriented covering the basics of RAID including the history, terminology, types of RAID, functionality, pros and cons, and descriptions.

Part 2, will be dealing with the implementation and configuration of Software and Hardware RAID controller

Part 3, Which RAID controller is suitable for our project and how to implement

Part 1

HISTORY OF RAID

Most credit the beginning of RAID research to Norman Ken Ouchi at IBM. He was issued U.S. Patent 4,092,732 titled "System for recovering data stored in failed memory unit" in 1978 and the claims for this patent describe what would later be termed RAID 5 with full stripe writes. This 1978 patent also mentions that disk mirroring or duplexing (what would later be termed RAID 1) and protection with dedicated parity (what would later be termed RAID 4) were prior art at that time.

RAID levels 1 through 5 were formally defined by David A. Patterson, Garth A. Gibson and Randy H. Katz in the paper, "A Case for Redundant Arrays of Inexpensive Disks (RAID)". This paper also listed a mathematical calculation that many believe is still accurate today which is used to determine Meant Time To Failure (MTTF), a key factor in a systems's fault tolerance. While there have been many ground breaking advances in RAID development since this article, one must fully credit these researchers form the University of California at Berkeley as having the paternity rights to what is considered modern day RAID.

While this history is extremely meaningful to the development of modern day RAID, it did not stop there. Annually the RAID Symposia, an international conference which began in 1998 on RAID development is held. Here researchers, developers, and other interested parties from all over the world meet to present and discuss key research oriented information that is instrumental to the continuing development of even more efficient RAID hardware and software for the present and future.

Technical Terms

RAID (Redundant Array of Inexpensive or Independent Drives) - is a technology that uses multiple hard drives to increase the speed of data transfer to and from hard disk storage, and also to provide instant data backup and fault tolerance for any information you might store on a hard drive.

RAID Array - A group of hard drives linked together as a single logical drive, connected to one or more hardware RAID controllers, or be attached normally to a computer using a RAID capable operating system, such as Windows XP Professional.

Redundancy - In a redundant system, if you lose part of the system, you can continue to operate. For example, if you have two power supplies and one takes over if the other one dies, that is a form of redundancy. You can take redundancy to extreme levels, but you spend more money.

JBOD (Just a Bunch of Disks) - One or more disk drives that form a single volume. However, the information on these disks is not striped in any way or protected--a JBOD is not a RAID. The term JBOD can also be used to refer to a volume on a single drive, where anything that's not a RAID is a JBOD.

Bit - The smallest unit of measure in a computer. It is represented by a 0 (off) or 1 (on). You can think of a bit as a switch. If it's in the on position it's a 1, and if the switch is off it's a 0. All parts of your computer communicate in bits at the lowest level.

Byte - A contiguous sequence of binary bits within a binary computer, that comprises the smallest addressable sub-field of the computer's natural word-size. That is, the smallest unit of binary data on which meaningful computation, or natural data boundaries, could be applied.

Fault-tolerance - Simply put this is the ability of a RAID array to continue to function after the degradation or loss of one or more of its constituent components.

Hard disk drive (HDD) - is a non-volatile data storage device that stores data on a magnetic surface layered onto hard disk platters.

Volume - A fixed amount of storage on a disk or tape. The term volume is often used as a synonym for the storage medium itself, but it is possible for a single disk to contain more than one volume or for a volume to span more than one disk.

Disk Mirroring - A procedure in which data sent to a RAID array is duplicated and written onto two or more drives at once.

Disk Duplexing - A procedure much like disk mirroring, but each drive is on a separate controller. This speeds up the normally slow write operations and also adds an additional level of redundancy, in case one of your controller cards dies.

Striping - A procedure in which data sent to a RAID array is broken down and portions of it written to each drive in the array. This can dramatically speed up hard drive access when the data is read back, since each drive can transfer part of the data simultaneously

Parity - In its simplest form, parity is an addition of all the drives used in an array. Recovery from a drive failure is achieved by reading the remaining good data and checking it against parity data stored by the array. Parity is used by RAID levels 2, 3, 4, and 5. RAID 1 does not use parity because all data is completely duplicated (mirrored). RAID 0, used only to increase performance, offers no data redundancy at all.

Mean Time to Data Loss (MTDL) - The average time before the failure of an array component causes data to be lost or corrupted.

Mean Time between Data Access / Availability (MTDA) - The average time before non-redundant components fail, causing data inaccessibility without loss or corruption.

Mean Time To Repair (MTTR) - The average time required to bring an array storage subsystem back to full fault tolerance.

Mean Time Between Failures (MTBF) - Used to measure computer component average reliability/life expectancy. MTBF is not as well-suited for measuring the reliability of array storage systems as MTDL, MTTR or MTDA because it does not account for an array's ability to recover from a drive failure. In addition, enhanced enclosure environments used with arrays to increase uptime can further limit the applicability of MTBF ratings for array solutions.

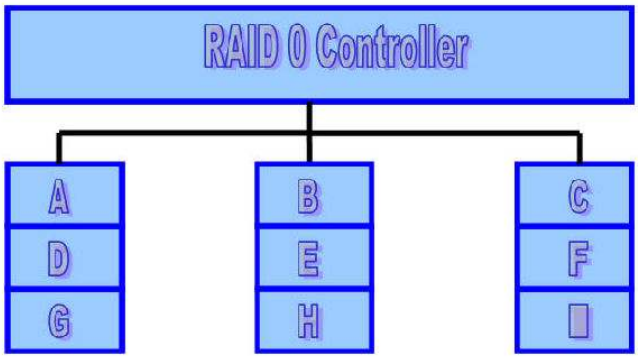
TYPES OF RAID LEVELS

There are several different RAID levels or redundancy schemes, each with an inherent cost, performance, and availability (fault-tolerance) characteristics designed to meet different storage needs. No individual RAID level is inherently superior to any other. Each of the five array architectures is well-suited for certain types of applications and computing environments. For client/server applications, storage systems based on RAID levels 1, 0/1, and 5 have been the most widely used.

RAID Level 0 (Non-Redundant)

A non-redundant disk array, or RAID level 0, has the lowest cost of any RAID organization because it does not employ redundancy at all. This scheme offers the best write performance since it never needs to update redundant information but it does not have the best read performance. Redundancy schemes that duplicate data, such as mirroring, can perform better on reads by selectively scheduling requests on the disk with the shortest expected seek and rotational delays. Without, redundancy, any single disk failure will result in data-loss. Non-redundant disk arrays are widely used in super-computing environments where performance and capacity, rather than reliability, are the primary concerns.

Sequential blocks of data are written across multiple disks in stripes. The size of a data block, which is known as the stripe width, varies with the implementation, but is always at least as large as a disk's sector size. When it comes time to read back this sequential data, all disks can be read in parallel. In a multi-tasking operating system, there is a high probability that even non-sequential disk accesses will keep all of the disks working in parallel.



Advantages:

- RAID 0 implements a striped disk array, the data is broken down into blocks and each block is written to a separate disk drive
- I/O performance is greatly improved by spreading the I/O load across many channels and drives
- Best performance is achieved when data is striped across multiple controllers with only one drive per controller
- No parity calculation overhead is involved
- Very simple design
- Easy to implement

Disadvantages:

- Not a "True" RAID because it is NOT fault-tolerant
- The failure of just one drive will result in all data in an array being lost
- Should never be used in mission critical environments

Recommended Use:

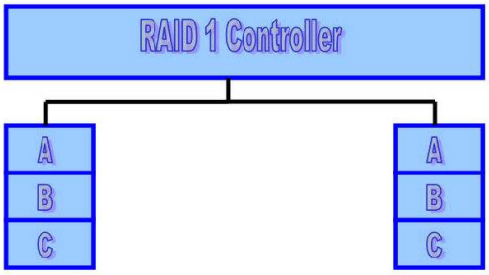
- Video Production and Editing
- Image Editing
- Pre-Press Applications
- Any application requiring high bandwidth

Conclusion:

Minimum number of drives: 2
Strengths: Highest performance
Weaknesses: No data protection; One drive fails, all data is lost

RAID Level 1 (Mirrored)

The traditional solution, called mirroring or shadowing, uses twice as many disks as a non-redundant disk array. Whenever data is written to a disk the same data is also written to a redundant disk, so that there are always two copies of the information. When data is read, it can be retrieved from the disk with the shorter queuing, seek and rotational delays. If a disk fails, the other copy is used to service requests. Mirroring is frequently used in database applications where availability and transaction time are more important than storage efficiency



Advantages:

- One Write or two Reads possible per mirrored pair
- Twice the Read transaction rate of single disks, same Write transaction rate as single disks
- 100% redundancy of data means no rebuild is necessary in case of a disk failure, just a copy to the replacement disk
- Transfer rate per block is equal to that of a single disk
- Under certain circumstances, RAID 1 can sustain multiple simultaneous drive failures
- Simplest RAID storage subsystem design

Disadvantages:

- Highest disk overhead of all RAID types (100%) – inefficient
- Lower write performance

Recommended Use:

- Accounting , Payroll , Financial
- Any application requiring very high availability
- Dual data storage from important files

Conclusion:
Minimum number of drives: 2
Strengths: Very high performance; Very high data protection; Very minimal penalty on write performance.
Weaknesses: High redundancy cost overhead; Because all data is duplicated, twice the storage capacity is required.

RAID Level 2 (Memory Style)

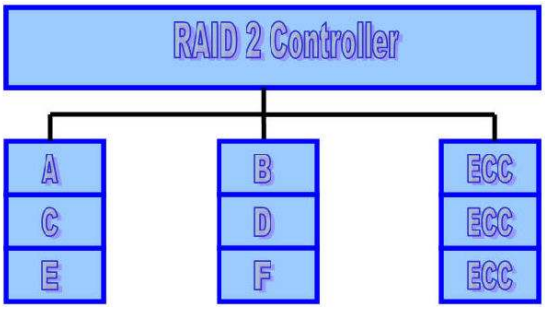
Memory systems have provided recovery from failed components with much less cost than mirroring by using Hamming codes. Hamming codes contain parity for distinct overlapping subsets of components. In one version of this scheme, four disks require three redundant disks, one less than mirroring. Since the number of redundant disks is proportional to the log of the total number of the disks on the system, storage efficiency increases as the number of data disks increases.

If a single component fails, several of the parity components will have inconsistent values, and the failed component is the one held in common by each incorrect subset. The lost information is recovered by reading the other components in a subset, including the parity component, and setting the missing bit to 0 or 1 to create proper parity value for that subset. Thus, multiple redundant disks are needed to identify the failed disk, but only one is needed to recover the lost information.

In you are unaware of parity, you can think of the redundant disk as having the sum of all data in the other disks. When a disk fails, you can subtract all the data on the good disks form the parity disk; the remaining information must be the missing information. Parity is simply this sum modulo 2.

A RAID 2 system would normally have as many data disks as the word size of the computer, typically 32. In addition, RAID 2 requires the use of extra disks to store an error-correcting code for redundancy. With 32 data disks, a RAID 2 system would require 7 additional disks for a Hamming-code ECC.

For a number of reasons, including the fact that modern disk drives contain their own internal ECC, RAID 2 is not a practical disk array scheme.



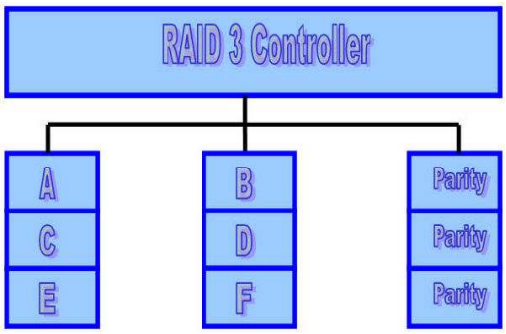
Conclusion :
Minimum number of drives: Not used in LAN
Strengths: Previously used for RAM error environments correction (known as Hamming Code) and in disk drives before he use of embedded error correction.
Weaknesses: No practical use; same performance can be achieved by RAID 3 at lower cost.

RAID Level 3 (Bit-Interleaved Parity)

One can improve upon memory-style ECC disk arrays by noting that, unlike memory component failures, disk controllers can easily identify which disk has failed. Thus, one can use a single parity rather than a set of parity disks to recover lost information.

In a bit-interleaved, parity disk array, data is conceptually interleaved bit-wise over the data disks, and a single parity disk is added to tolerate any single disk failure. Each read request accesses all data disks and each write request accesses all data disks and the parity disk. Thus, only one request can be serviced at a time. Because the parity disk contains only parity and no data, the parity disk cannot participate on reads, resulting in slightly lower read performance than for redundancy schemes that distribute the parity and data over all disks. Bit-interleaved, parity disk arrays are frequently used in applications that require high bandwidth but not high I/O rates. They are also simpler to implement than RAID levels 4, 5, and 6.

Here, the parity disk is written in the same way as the parity bit in normal Random Access Memory (RAM), where it is the Exclusive Or of the 8, 16 or 32 data bits. In RAM, parity is used to detect single-bit data errors, but it cannot correct them because there is no information available to determine which bit is incorrect. With disk drives, how ever, we rely on the disk controller to report a data read error.Knowing which disk's data is missing; we can reconstruct it as the Exclusive Or (XOR) of all remaining data disks plus the parity disk.



Conclusion:

Minimum number of drives: 3
Strengths: Excellent performance for large, sequential data requests.
Weaknesses: Not well-suited for transaction-oriented network applications; Single parity drive does not support multiple, simultaneous read and write requests.

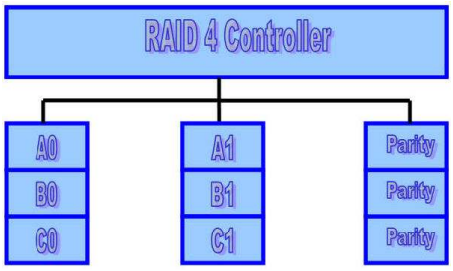
RAID Level 4 (Block-Interleaved Parity)

The block-interleaved, parity disk array is similar to the bit-interleaved, parity disk array except that data is interleaved across disks of arbitrary size rather than in bits. The size of these blocks is called the striping unit. Read requests smaller than the striping unit access only a single data disk. Write requests must update the requested data blocks and must also compute and update the parity block. For large writes that touch blocks on all disks, parity is easily computed by exclusive-or'ing the new data for each disk. For small write requests that update only one data disk, parity is computed by noting how the new data differs from the old data and applying those differences to the parity block. Small write requests thus require four disk I/Os: one to write the new data, two to read the old data and old parity for computing the new parity, and one to write the new parity. This is referred to as a read-modify-write procedure. Because a block-interleaved, parity disk array has only one disk, which must be updated on all write operations, the parity disk can easily become a bottleneck. Because of this limitation, the block-interleaved distributed parity disk array is universally preferred over the block-interleaved, parity disk array.

A write request for one block is issued by a program:

- RAID software determines which disks contain data, parity, and which block they are in.
- The disk controller reads the data block from disk.
- The disk controller reads the corresponding parity block from disk.
- The data block just read is XORed with the parity block just read.
- The data block to be written is XORed with the parity block.
- The data block and the updated parity block are both written to disk.

It can be seen from the above example that a one block write will result in two blocks being read from disk and two blocks being written to disk. If the data blocks to be read happen to be in a buffer in the RAID controller, the amount of data read from disk could drop to one, or even zero blocks, thus improving the write performance.

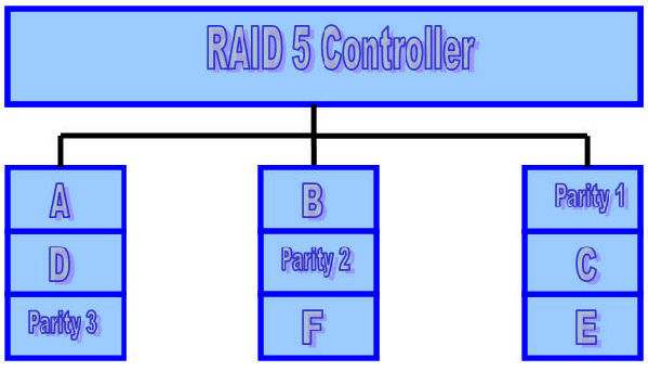


Conclusion:
Minimum number of drives:3 (Not widely used)
Strengths: Data striping supports multiple simultaneous read requests.
Weaknesses: Write requests suffer from same single parity-drive bottleneck as RAID 3; RAID 5 offers equal data protection and better performance at same cost. For small writes, the performance will decrease considerably. To understand the cause for this, a one-block write will be used as an example.

RAID Level 5 (Block-Interleaved Distributed Parity)

The block-interleaved distributed-parity disk array eliminates the parity disk bottleneck present in the block-interleaved parity disk array by distributing the parity uniformly over all of the disks. An additional, frequently overlooked advantage to distributing the parity is that it also distributes data over all of the disks rather than over all but one. This allows all disks to participate in servicing read operations in contrast to redundancy schemes with dedicated parity disks in which the parity disk cannot participate in servicing read requests. Block-interleaved distributed-parity disk array have the best small read, large write performance of any redundancy disk array. Small write requests are somewhat inefficient compared with redundancy schemes such as mirroring however, due to the need to perform read-modify-write operations to update parity. This is the major performance weakness of RAID level 5 disk arrays.

The exact method used to distribute parity in block-interleaved distributed-parity disk arrays can affect performance. The following figure illustrates left-symmetric parity distribution.



Each square corresponds to a stripe unit. Each column of squares corresponds to a disk. Parity 1 computes the parity over stripe units A and B; Parity 2 computes parity over stripe units C and D; Parity 3 computes the parity over stripe units E and F.

A useful property of the left-symmetric parity distribution is that whenever you traverse the striping units sequentially, you will access each disk once before accessing any disk device. This property reduces disk conflicts when servicing large requests.

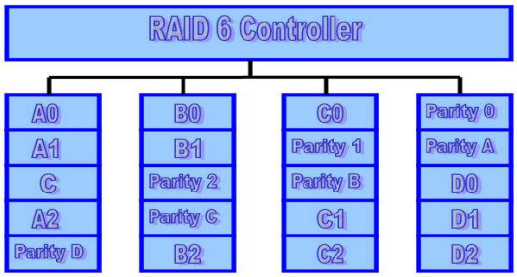
Conclusion :

Minimum number of drives: 3
Strengths: Best cost/performance for transaction-oriented networks; Very high performance, very high data protection; Supports multiple simultaneous reads and writes; Can also be optimized for large, sequential requests.
Weaknesses: Write performance is slower than RAID 0 or RAID 1.

RAID Level 6 (P+Q Redundancy)

Parity is a redundancy code capable of correcting any single, self-identifying failure.As large disk arrays are considered, multiple failures are possible and stronger codes are needed. Moreover, when a disk fails in parity-protected disk array, recovering the contents of the failed disk requires successfully reading the contents of all non-failed disks. The probability of encountering an uncorrectable read error during recovery can be significant. Thus, applications with more stringent reliability requirements require stronger error correcting codes.

Once such scheme, called P+Q redundancy, uses Reed-Solomon codes to protect against up to two disk failures using the bare minimum of two redundant disk arrays The P+Q redundant disk arrays are structurally very similar to the block interleaved distributes-parity disk arrays and operate in much the same manner. In particular, P+Q redundant disk arrays also perform small write operations using a read-modify-write procedure, except that instead of four disk accesses per write requests, P+Q redundant disk arrays require six disk accesses due to the need to update both the `P' and `Q' information.



RAID Level 10 (Striped Mirrors)

RAID 10 is now used to mean the combination of RAID 0 (striping) and RAID 1 (mirroring). Disks are mirrored in pairs for redundancy and improved performance, and then data is striped across multiple disks for maximum performance.

RAID 10 uses more disk space to provide redundant data than RAID 5. However, it also provides a performance advantage by reading from all disks in parallel while eliminating the write penalty of RAID 5. In addition, RAID 10 gives better performance than RAID 5 while a failed drive remains un-replaced. Under RAID 5, each attempted read of the failed drive can be performed only by reading all of the other disks. On RAID 10, a failed disk can be recovered by a single read of its mirrored pair.

Conclusion:
Minimum number of drives: 4
Strengths: High performance, highest data protection (tolerates multiple drive failures)
Weaknesses: High redundancy cost overhead; Because all data is duplicated, twice the storage capacity is required; Requires minimum of four drives.

Part 2:

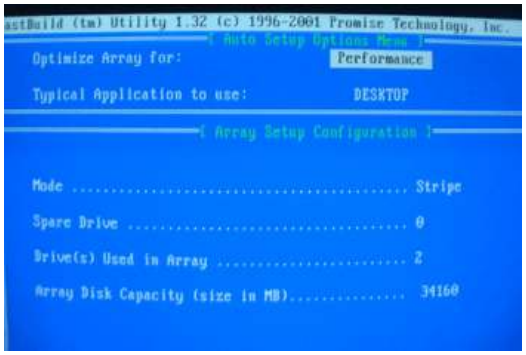
Configuration of RAID

Note that for the purposes of hardware RAID 0 (striping) it is strongly recommended that you use two disks of the exact same model. For mirror (RAID 1) setups, this is not so essential, but the two drives should be of the same capacity.

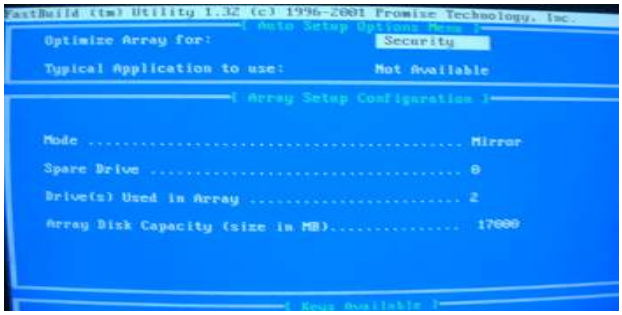
Attach the drives to the RAID controller, one drive per channel, set as master for the best performance, and boot the computer. Note that while you can attach both SATA drives to a single IDE port on your RAID controller, you will tend to get better performance with a pair of drives if you plug one into each port during startup, the RAID controller drive detection and setup screen will appear.

Press <CTRL-H> or F10 other key combination as instructed to enter RAID setup

From the main menu, press '1' to enter Auto Setup. From here, you can choose either a RAID 0 or 1 configuration, referred to in this case as either 'performance' or 'security.' Note the separate drive configurations in the screen shots



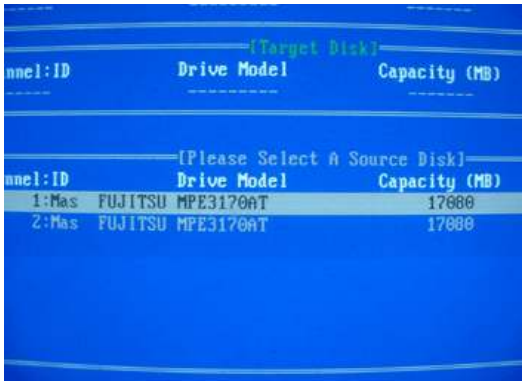
Steps to configure the Hardware RAID



Choose and accept the desired RAID type. If you select a stripe (RAID 0) array, no further configuration is necessary. Accept the change and reboot.

If you elect to setup a RAID 1 (mirror) configuration, you must then choose whether you wish to simply create a mirror array (if you have two blank disks and want them to be exact copies when adding data in the future), or create the array and then copy the contents of one disk to the other (if you have a data drive and you wish to create a mirror copy of it for redundancy).

If you elect to mirror and copy data, you will be asked to choose a source drive for the data.



BE CAREFUL. Choosing the wrong drive here can be disastrous, so ensure that you know which drive is which. Paying attention to which port you plugged each drive into should help here, as they will be labeled on the motherboard or card. Once you have created the array, reboot.

How to set up Software RAID in windows XP Professional

Like most other hard drive and storage options, RAID is managed through Windows XP's disk management window, found by right clicking on 'my computer,' then selecting 'manage' followed by 'disk management.' Windows XP Professional is only capable of creating RAID 0 striped arrays, while the various Windows Server operating systems can also create software RAID 1 mirror arrays.

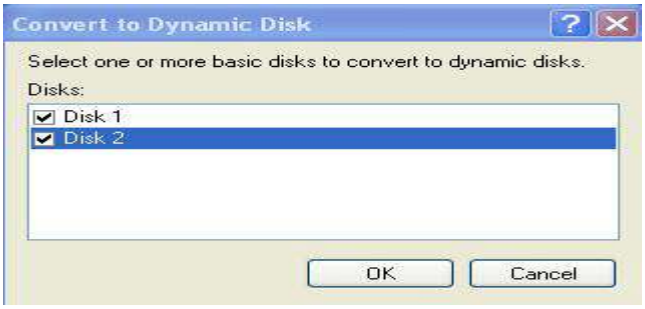
Creating a striped RAID array in XP:

For the purpose of this section of the guide we installed two blank 17GB hard drives on a test system. To create a striped array you must first have at least two drives with a portion of 'unpartitioned space' free. The largest stripe you can create will be twice the size of the smallest unused space on either of the disks. If you have two disks, one with 4GB of unpartitioned space and one with 3GB, the largest striped array you could create would be 6GB, as the area of space used by the stripe on each disk must be the same.

The first step is to convert both disks from basic to dynamic disks within Windows. A dynamic disk is a disk that contains an additional database of other dynamic disks on the system. Dynamic disks can only be read by Windows 2000, XP Professional and the various Windows Server operating systems, and are required to create software RAID arrays within Windows.

To convert the disks from basic to dynamic, right click the grey box on the left that contains the disk names (disk 1, disk 2, etc.) and select 'convert to dynamic disk…'

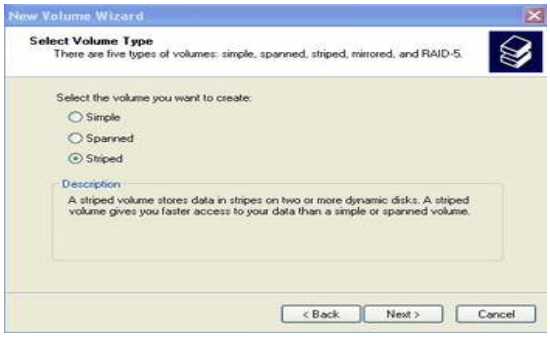
From the next Window you can check both blank drives and click 'ok' to convert them.



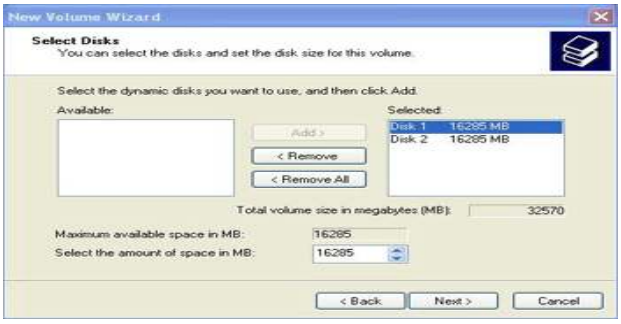
Once both disks are listed as dynamic, right click the 'unpartitioned space' of either drive and select 'new volume.' On the next page we'll set these drives to be striped, and configure the software RAID options.

Setting up a hardware RAID array

In the 'select volume type' Window, select 'striped.'



Add all disks you wish to use, then decide on the amount of space on both disks you wish to use for the striped volume you are about to create. If you wish, you use only part of each disk for the stripe, leaving the rest free for other uses.



Choose a drive letter or folder to use, and the method of formatting, and you are done. The striped array will format and be ready for use.

How to set up hardware RAID:

For this section, we used a Highpoint HPT 372 ATA/133 RAID controller built into an Epox EP-8K5A2+ motherboard. The drives we used to test our RAID configuration were a pair of SATA Disks 149GB hard disks. Similarly we can also set up a second hardware RAID configuration on a Promise 20276 ATA/133 RAID controller built into an MSI KT3 Ultra2 motherboard, attached to the same pair of 17GB drives used in the software RAID setup above. These two controllers are typical of hardware RAID solutions found on modern motherboards and add-in PCI cards.

We wanted to include instructions for both Highpoint and Promise controllers, as these two companies dominate the home desktop and enthusiast market for RAID controllers. Most RAID setup functions are standard, so if you do not have the same exact controller, these instructions should still translate well.

The following instructions assume two identical blank hard disks. It also assumes that you have correctly installed the Windows drivers for your RAID controller. We used the most recent BIOS versions for both controllers, and we recommend that you obtain these from the manufacturer's website if you have not done so already.

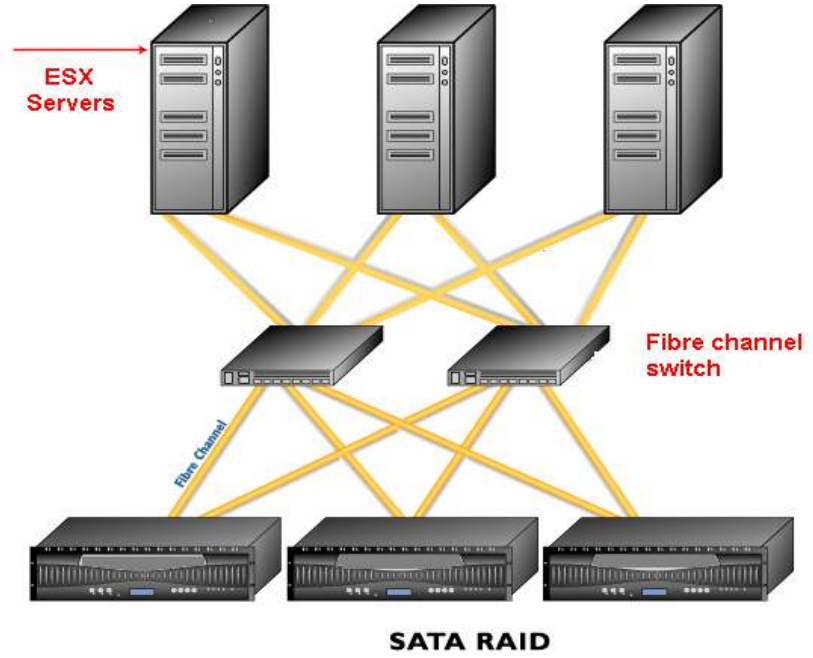
Part 3:

Infrastructure Management service:

This Project concentration on the management services, storage availability and management of the clients. This server consolidation is done by virtualization Technology. We use VMware ESX server for multiple environments and server consolidation. SAN is used as storage, this helps us in easy storage management as it is more reliable and can be assigned to any server that needs more storage, this avoids in purchasing new storages. SAN storage helps in remote connectivity to the server, helping us in centralizing data and also for data backup, and disaster recovery

Here we use the Two Xeon Processor. One processor is configured with ESX server. Second Xeon Processor configured with VMware ESX server connected to SAN which is configured for data recovery (RAID array).If any disaster occurring in primary data centre, then applications running on VMware can be recovered from the secondary data centre. Data Backup we using the RAID array.

VMware ESX Server is a software platform that efficiently multiplexes the hardware resources of a server among virtual machines. VMware ESX server Support for additional storage and server hardware



Production Environment

Configuring and Working:

Storage Area Network:

A storage area network (SAN) is a specialized high-speed network of storage devices and computer systems (also referred to as servers, hosts, or host servers). Currently, most SANs use the Fibre Channel protocol.

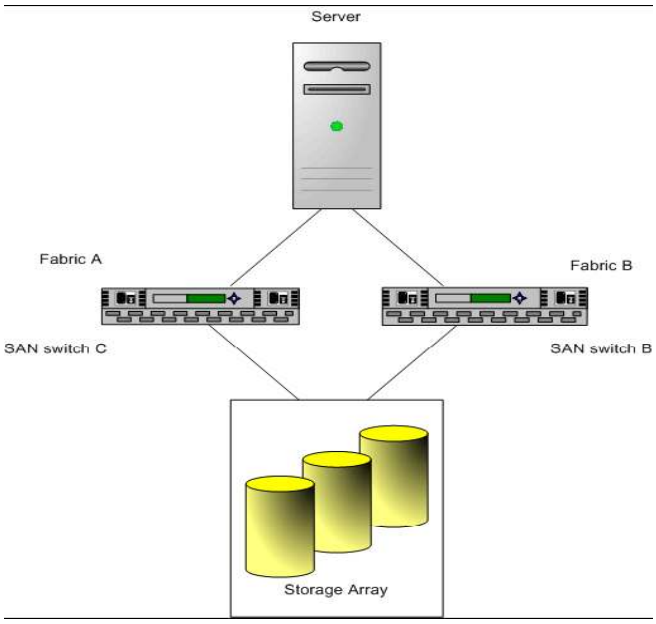
A storage area network presents shared pools of storage devices to multiple servers.Each server can access the storage as if it were directly attached to that server. The SAN makes it possible to move data between various storage devices, share databetween multiple servers, and back up and restore data rapidly and efficiently. In addition, a properly configured SAN provides robust security, which facilitates both disaster recovery and business continuance. Components of a SAN can be grouped closely together in a single room or connected over long distances. This makes SAN a feasible solution for businesses of any size: the SAN can grow easily with the business it supports.

SAN Components
In its simplest form, a SAN is a number of servers attached to a storage array using a switch. The following components are involved:

• SAN Switches — Specialized switches called SAN switches are at the heart of the typical SAN. Switches provide capabilities to match the number of host SAN connections to the number of connections provided by the storage array.Switches also provide path redundancy in the event of a path failure from host server to switch or from storage array to switch.

• Fabric — When one or more SAN switches are connected, a fabric is created.The fabric is the actual network portion of the SAN. A special communications protocol called Fibre Channel (FC) is used to communicate over the entire network. Multiple fabrics may be interconnected in a single SAN, and even for a simple SAN it is not unusual to be composed of two fabrics for redundancy.

• Connections: HBA and Controllers — Host servers and storage systems are connected to the SAN fabric through ports in the fabric. A host connects to a fabric port through a Host Bus Adapter (HBA), and the storage devices connect to fabric ports through their controllers. Each server may host numerous applications that require dedicated storage for applications processing. Servers need not be homogeneous within the SAN environment.



How a SAN Works

The SAN components interact as follows:
1. When a host wishes to access a storage device on the SAN, it sends out a block based access request for the storage device.
2. The request is accepted by the HBA for that host and is converted from its binary data form to the optical form required for transmission on the fiber optic cable.
3. At the same time, request is “packaged” according to the rules of the Fiber Channel protocol.
4. The HBA transmits the request to the SAN.
5. Depending on which port is used by the HBA to connect to the fabric, one of the SAN switches receives the request and checks which storage device the host wants to access. From the host perspective, this appears to be a specific disk, but it is actually just a logical device that corresponds to some physical device on the SAN. It is up to the switch to determine which physical device has been made available to the host for its targeted logical device.
6. Once the switch has determined the appropriate physical device, it passes the request to the appropriate storage device. The remaining sections of this chapter provide additional details and information about the components of the SAN and how they interoperate. These sections also present general information on the different ways in which a SAN can be configured, and the considerations to be made when designing a SAN configuration.

SAN Components

Host Components
Fabric Components
Storage Controllers

Host Components

The host components of a SAN consist of the servers themselves and the components that enable the servers to be physically connected to the SAN:

• Host bus adapters (HBAs) are located in the servers, along with a component that performs digital-to-optical signal conversion. Each host connects to the fabric ports from its HBA.
• Cables connect the HBAs in the servers to the ports of the SAN fabric.
• HBA drivers run on servers to enable server's operating system to communicate with HBA.

Fabric Components

All hosts connect to the storage devices on the SAN through the fabric of the SAN.
The actual network portion of the SAN is formed by the fabric components.
The fabric components of the SAN can include any or all of the following:

• Data Routers
• SAN Hubs
• SAN Switches
• Cables

Data Routers

Data routers provide intelligent bridges between the Fibre Channel devices in the SAN and the SCSI devices. Specifically, servers in the SAN can access SCSI disk or tape devices in the SAN through the data routers in the fabric layer.

SAN Hubs

SAN hubs were used in early SANs and were the precursors to today's SAN switches. A SAN hub connects Fibre Channel devices in a loop (called a Fibre Channel Arbitrated Loop, or FC-AL). Although some current SANs may still be based on fabrics formed by hubs, the most common use today for SAN hubs is for sharing tape devices, with SAN switches taking over the job of sharing disk arrays.

SAN Switches

SAN switches are at the heart of most SANs. SAN Switches can connect both servers and storage devices, and thus provide the connection points for the fabric of the SAN.

• For smaller SANs, the standard SAN switches are called modular switches and can typically support 8 or 16 ports (though some 32-port modular switches are beginning to emerge). Sometimes modular switches are interconnected to create a fault-tolerant fabric.
• For larger SAN fabrics, director-class switches provide a larger port capacity (64 to 128 ports per switch) and built-in fault tolerance. The type of SAN switch, its design features, and its port capacity all contribute its overall capacity, performance, and fault tolerance. The number of switches, types of switches, and manner in which the switches are interconnected define the topology of the fabric.

Cables

SAN cables are special fiber optic cables that are used to connect all of the fabric components. The type of SAN cable and the fiber optic signal determine the maximum distances between SAN components, and contribute to the total bandwidth rating of the SAN.

Storage Components

The storage components of the SAN are the disk storage arrays and the tape storage devices. Storage arrays (groups of multiple disk devices) are the typical SAN disk storage device. They can vary greatly in design, capacity, performance, and other features. Tape storage devices form the backbone of the SAN backup capabilities and processes.

• Smaller SANs may just use high-capacity tape drives. These tape drives vary in their transfer rates and storage capacities. A high-capacity tape drive may exist as a stand-alone drive, or it may be part of a tape library.

• A tape library consolidates one or more tape drives into a single enclosure. Tapes can be inserted and removed from the tape drives in the library automatically with a robotic arm. Many tape libraries offer very large storage capacities—sometimes into the petabyte (PB) range. Typically, large SANs, or SANs with critical backup requirements, configure one or more tape libraries into their SAN.

WORKING:

One Xeon processor is configured with VMware ESX server to SAN which is configured for the storage system such as RAID. Which is RAID -1 (Mirror) technology or mirroring or shadowing, uses twice as many disks as a non-redundant disk array Whenever data is written to a disk the same data is also written to a redundant disk, so that there are always two copies of the information. When data is read, it can be retrieved from the disk with the shorter queuing, seek and rotational delays. If a disk fails, the other copy is used to service requests. Mirroring is frequently used in database applications where availability and transaction time are more important than storage efficiency.

It requires Minimum 2 drives, Very high performance, Very high data protection, Very minimal penalty on write performance.
But High redundancy cost overhead; because all data is duplicated, twice the storage capacity is required

Each LUN should have the right RAID level and storage characteristic for applications in virtual machines that will use it.One LUN should contain only one single VMFS volume. If multiple virtual machines accessing same LUN, use disk shares to prioritize virtual machines.

Overall Operation:

The project deals with the infrastructure service, storage management, and testing and server consolidation of the client, using virtualization technologies and latest storage technologies. The production server is built on VMware ESX server for server consolidation. ESX server consolidates different applications and infrastructure services running on different operating systems onto fewer highly scalable, reliable enterprise class servers. Production server is connected to SAN storage setup.

SAN helps us to assign the storage remotely to the server. No downtime and perhaps not even a reboot is required if the OS can handle it. All the storage can be managed globally from a single console.

Using the snapshot of VMware ESX server and the data replication the availability of the server is managed during any disasters or downtime of the storage. Using SAN for centralizing, data backup can backup can improve recovery time dramatically while reducing overall costs by sharing tape resources and eliminating backup windows.

The instances which are taken as the snapshot are used for testing the application. But we can't move the snapshot. So we are taking the clone of the snapshot which is copy of snapshot it contains whole virtual machine. The applications are tested and the bugs are reported in the Bugzilla, defect tracking system used. Then the patch is being developed using the bug report as the source. This developed patch is deployed on to the Testing server maintained at Virtualization lab.

Compiled by: Narendar K
narendar.k@vassure.com

This paper is not intended to be a definitive implementation guide, it is compilation of data. Many factors are not addressed in this document. Expertise may be required to solve logistical problems when the system is designed and built. VAssure team has not tested this procedure with all the combinations of hardware and software options available on all ESX or guest OS variants. There may be significant differences in your configuration that will alter the procedures necessary to accomplish the objectives outlined in this paper.