Xinetica White Paper

Title:	**Availability Architecture White Paper**
Content:	**Explaining fundamental 'Availability' concepts and terminology.**
Developed by:	**Lee Hezzlewood. (lee.hezzlewood@xinetica.com)**

Xinetica Ltd.,
Asher House,
Barsbank Lane,
Lymm,
Cheshire.
WA13 0ED.
Tel: +44 (0)1925 759838
Fax: +44 (0)1925 759839
www.xinetica.com

For related topics and information on Xinetica - see www.xinetica.com

---

Document Contents:

---

1. Introduction

This document is intended to give an overview of the technologies involved in creating a highly available environment, along with the business and cost implications. While it will give a feel of the steps involved it is not designed to be an implementation guide for building an HA environment, merely a discussion of things that need to be considered.

2. Definition of ‘Availability’

What makes a system or environment 'Available'? Or, rather, what makes a system or environment more 'Available' than another system or environment?

An available system is one that minimises disruption to the service provided by that system. This disruption can be caused by both planned and unplanned outages. This minimising of disruption can take many forms and can be applied to various scenarios, from users being able to access a screen at certain times to a business critical, 24 by 7 operation, where downtime can cost vast sums of money.

Making a system highly available can require considerable thought, effort and investment. In simple terms an ‘Available’ system could just be something running on a platform that is known to be reliable, and therefore reduces downtime. In other forms it consists of clustering, disaster recovery and business continuity planning, all of which are used to minimise the effect

of an outage of varying severity.

Backups play an important part in a system's availability. Good backups make recovering from a problem easier and therefore quicker. Many organisations believe that backups are something that should be considered once an environment has been established. However, it should be considered at the architecture design stage in order to maximise its effectiveness.

Availability also incorporates many areas of system architecture. These areas include disk storage, network infrastructure, server architecture, application design and data management.

3. Network Infrastructure

In a day when more and more business is being performed electronically and more organisations are using technology to improve their businesses, the network is critical.

If the network cannot handle the amount of data required then systems become unavailable for the purpose they have been designed. This is essentially an outage, and a great deal of thought needs to be put in to designing a network infrastructure to handle both current business needs, and be scaleable enough to handle any predicted growth.

The network also needs to be secure. A security breach can disrupt service. There are a number of ways that security breaches can be prevented and attempts detected. For further discussion of security issues and our related services visit http://www.xinetica.com/what_we_can_do/

4. Storage Technology

Having the correct storage technology plays a vital part in ensuring system availability. There are a number of different technologies available ranging from stand alone disks through to large storage arrays that can be used directly attached or via a network. Some of the more common technologies are discussed below.

4.1. Storage Arrays

Storage arrays are directly attached units which house multiple disks. They are usually attached to the server(s) via one or more SCSI or Fibre Channel connections. The disks in the storage arrays can either be accessed individually or as an array using RAID technology (Redundant Array of Inexpensive Disks).

RAID technology allows a number of disks to be configured as a single logical unit in one of a number of configurations. This single logical disk unit is generally referred to as a LUN (Logical Unit). The RAID configurations are :

- Level 0 - A simple stripe where the LUN is placed across a number of physical disks. This configuration does not on its own offer any resilience. Should a physical disk fail the entire LUN would be unavailable.
- Level 1 - A mirror where the LUN consists of 2 disks that mirror each other. This offers good resilience as the loss of a single physical disk does not disable the LUN. However the number of disks required is doubled and the LUN can only be as large as the physical device.
- Level 0+1 (or 10 or 01) - This is a combination of Level 0 and Level 1. In essence a mirrored stripe. This provides excellent resilience as no single failure will disable the LUN and it provides for a LUN size greater than available on a single physical disk. The disadvatage is that double the disks are required.
- Level 4 - This is similar to Level 0 but with one of the disks in the stripe holding parity data for the other disks. This means that a single disk failure will not disable the LUN and a replacement disk can be added and have its data 'rebuilt' from the parity on the

parity disk. The problem with this layout is that the parity disk is accessed every time data is written to the LUN, causing a write bottleneck in the performance.

- Level 5 - This is similar to Level 4 but with the parity being spread between all of the disks in the stripe. This provides a performance improvement over Level 4 in most implementations as all of the disks are written to evenly.

RAID can be implemented on storage arrays either by using hardware, or using software such as a Volume Manager. The implementation can also be a combination of both by, for example, configuring a Level 0 stripe on the array using hardware and then implementing mirroring of 2 (or more) stripes using a Volume Manager. The LUN defined by the Volume Manager software would then be accessed by the operating system as though it were a single device.

Disk storage arrays tend to be used where the data is accessed by a single system (or cluster) in the same location. There are limits to the physical distance a storage array can be from the hardware it is connected to. Fibre Channel offers greater flexibility than SCSI in this regard.

Array technology (such as RAID) is also widely used in both Network Attached Storage and Storage Area Networks which are discussed later in this document.

4.2. Volume Managers

Volume Managers are software packages or operating system additions that allow the management of disks from the servers themselves. They allow mirroring and RAID striping to be configured on any type of supported disk. Volume Managers offer great flexibility with disks and LUN's being simple to add or reconfigure.

Volume managers, in conjunction with file system software can manage file system resizing without loss of service to applications using the file system.

Volume Managers, similarly to Storage Arrays, create LUNs. The LUNs, rather than disks, are then accessed by the operating system. The Volume Management software then handles physical disk access in the way required.

Another advantage of Volume Managers is the ability to add and split mirrors. This can be useful if there is a requirement for off-line backups, where a copy of the data can be created from a split off mirror. The mirror is recreated once the backup is complete.

4.3. Storage Area Network (SAN)

Storage Area Networks (SANs) are dedicated networks, devoted to data storage and backups. A SAN has one or more storage arrays and/or high capacity tape units that are accessed by a number of servers. Each server must have disks allocated from the disk arrays. Servers cannot share data with other systems apart from using the normal shared file systems approach as implemented on various operating systems.

Usually, SAN disk arrays are connected to the host systems using Fibre Channel network technology.

SAN storage devices usually offer RAID, mirroring and hot-swap technology as standard. They offer good levels of redundancy and resilience.

4.4. Network Attached Storage (NAS)

Network Attached Storage (NAS) systems are large disk arrays connected directly to the network. They have little by way of operating system, just enough to be able to handle requests from client systems and manage the disks in the array. NAS arrays usually provide access via both IP and NetBIOS protocols.

NAS systems can be used to store data specific to one host system or data that is shared between a number of different systems, even those using different protocols for access. They offer the option of RAID technology, as in the Storage Arrays, but generally have a specific implementation (i.e. RAID 4 or RAID 5). They can also offer mirroring, within the array and to a separate array, hot-swap disks and the ability to take snapshots of data to use for backup purposes.

Due to the ever increasing performance of network technology NAS systems can potentially offer the same or even better performance than directly attached storage. NAS systems can also be directly attached to individual systems using high speed network connectivity in order to maximise security and performance. Note that in this context direct connection means that the system and the NAS are the only systems on a private network. Gigabit ethernet technology is now available for these networks with even more advanced technology currently in the pipeline.

NAS systems also offer clustering which is discussed later.

5. Clustering Technology

5.1. Load Balancing

Load Balancing can be broken down into two areas, Network and Component (or Application) Load Balancing.

Network Load Balancing is generally used for web hosting and e-business. Network Load Balancers usually sit in front of application servers and receive incoming requests. They then distribute the requests between a number of application servers, leveling out network activity and preventing a single bottleneck.

Component Load Balancing is used in applications, distributing workload between a number of servers that make up the application layer. This again prevents a single server causing a bottleneck while other systems are lightly loaded.

Both of these technologies are very useful in Internet applications, where single systems may produce poor levels of response and so potentially lose business.

5.2. Clustering

Clustering is where two or more servers, usually sharing resources such as storage arrays, are configured to provide the same service. A minimal cluster is implemented with one server providing a service with another being available should a failure occur in the primary system. If a failure does occur the second system takes over the resources of the first with minimal interruption in service.

In more complex configurations, multiple servers share access to resources while running their own applications. In this case, should a failure occur, the resources of the failed server are taken over by one of the remaining servers. This may result in a slightly reduced service capability, but offers a high level of resilience. The cost of availability per application is also reduced with no servers just waiting (for the inevitable) but all servers being used at all times.

Clustering technology can be used to provide a very high level of availability for business critical applications. Clustering offers a high level of resilience and minimises service interruption should a failure occur. Most cluster failovers will take only minutes before the application becomes available on the secondary server.

Most clustering technologies are very similar. Hardware requirements may differ slightly between implementations but are generally:

- two identically configured servers - usually with multiple network adapters, multiple

SCSI or Fibre Channels

- shared storage, again usually with multiple channels (at least 1 channel to each server) or network based storage (see SAN and NAS)
- heartbeat connections, these are usually dedicated network connections that are used by the clustering software.

The heartbeat connections are critical to the management of the cluster. These connections maintain communication between the servers in the cluster and ensure that any failures are noted and acted upon. Some implementations of clustering also use these heartbeat connections to maintain and distribute the cluster configuration data, i.e. data about the cluster itself, on all the nodes in the cluster.

5.3. NAS Clusters

NAS technology offers the ability to cluster filers (disk controller units) for a given set of storage. This provides a level of resilience above that offered by a single stand alone NAS system.

Because NAS systems use IP for connectivity to client hosts, the clients use an IP address to access their data. This means that all that is required in a NAS cluster is the ability for the filer to manage the common disk units and the ability to 'fail-over' the IP address(es) used to access those units. This makes NAS cluster fail over simple and easy with fewer complex fail over management issues to worry about.

5.4. Dynamic Multi Pathing (DMP) and Dynamic Reconfiguration

Dynamic Multi-Pathing and Dynamic Reconfiguration are technologies used to improve availability. DMP is where there are multiple channels to the same device. This allows the system to automatically reconfigure should one of the channels fail. It is therefore useful in situations when the loss of access to a device will result in an outage. DMP is usually software or operating system controlled and will be activated if the primary route to a device should be interrupted.

Dynamic reconfiguration is the ability to reconfigure the available hardware, while in use, with no outage and perhaps a small degradation in service. Dynamic reconfiguration is used with Sun E10000 Starfire systems and allows the addition and removal of hardware components without the need to shutdown the servers. System boards containing CPU's, memory and peripheral devices can be added and removed almost at will thus helping availability.

5.5. Failover

Failover occurs in a clustered environment when the service on the primary server is required to switch to a secondary server. This may be due to a failure at the primary or could be scheduled by an administrator for maintenance purposes.

When a failover is initiated, all running applications (if there are any still running) are shutdown, the filesystems are unmounted and disk resources are stopped and taken offline. The process is then begun to start these resources on the secondary server, with the disk resources being started, filesystems mounted and finally any applications being restarted.

The time taken for the failover process to occur will depend on the applications running on the servers and the number of resources configured in the environment. For example, a small Oracle database running on a cluster, with a single instance will take only a few minutes to failover.

6. Backup Technology

Backups form a major part of system and data availability. Without good backups, critical data can be lost through hardware faults, data corruption, accident and malicious damage. The loss of business critical data can be immeasurable and could result in great loss to a business.

Backup requirements should be fully considered as part of the initial architecture design. If it is done later, then the infrastructure may not be sufficient to handle the backup requirements. This is very important and should not be overlooked.

All backups should also be verified and even restored on a regular basis to ensure that no system or data is at risk, and that the backups themselves are working properly. Offsite storage of backup media should be considered to ensure that recovery from a more serious disaster is possible.

6.1. Media Technologies

There is a number of options for backup media. The most common technologies involve the use of tape devices. These device vary quite considerably in format and capacity, even those that on first appearance seem to be the same.

Some of the main tape technologies are:

- DAT - Digital Audio Tape. This is a magnetic medium that comes in a number of different formats. The current most popular being DDS 3 or DDS 2. The tapes are 4mm cassettes. The drives are available as either stand alone units, auto-loaders, with multiple tapes allocated to a single device, or tape banks, with multiple tapes allocated to multiple devices. Current maximum, uncompressed, capacity per tape is approximately 24 Gb.
- DLT - Digital Linear Tape. These tapes are much larger than DAT tapes, both physically and in capacity, with a storage capacity in the region 35Gb/70Gb (uncompressed/compressed). DLT drives are again available in single unit, jukebox, and multi-device configurations. DLT's are also much faster than DAT as data can be sent to them in a multi-threaded manner, improving throughput.
- 8mm Tape - Largely obsolete technology using 8mm videotapes. These devices are fairly slow and do not have a large capacity.
- QIC - Quarter Inch Cartridge. Old technology only able to store about 150Mb. No longer very common.

6.2. Backup Software

Backup software is available from a number of vendors. Most of the technologies have similarities in their use and implementation. The most popular are currently network backup technologies.

These use one or more servers with attached backup storage (DLT or DAT) connected to the network. Other servers then use the backup server across the network to save their files to the tape media.

The available network backup software packages all tend to have similar implementation considerations. They use configuration files or a bespoke database to define the files or filesystems to be backed up and the schedules to use.

6.3. Warm/Hot/Cold Backup Options

When performing backups there are a number of different options as to how they should be done. The method chosen will depend on a number of things - time allowed, impact on usage, media, schedules etc.

One option worth considering is whether it is possible to back up your data without preventing access to your services.

Hot backups are done online, with the system up and running. The problem here is that data files that are open could potentially be under modification when they are saved to the backup media. This would result in an incomplete file being saved and an invalid backup.

To avoid this problem most database management software vendors allow hot or online backups. By setting a flag within their software they can allow access to data files even though these files are being updated. The most common implementation of this is the use of ‘redo’ files, where a data file is frozen at a point in time. All modifications are then written to a temporary file to be applied to the data when the backup has finished.

Warm backups are similar to hot backups but are usually taken from a copy of the actual data, such as from a mirror. This means that the actual data is a snapshot but there is minimal impact to the service being provided. This reduces down time from what a cold backup would impose.

A cold backup involves the application being completely shutdown while the backup takes place. This may take many hours for large databases.

Some other backup options that could be considered are implemented on NAS or SAN technology. These are known as snap-shots, where a copy of the data is taken at a point in time and replicated on another set of disks, perhaps even on a separate SAN or NAS storage array. Some of these technologies will even allow a limited number of ‘archives’ of your data to remain on the same set of disks that your live data is on.

Using these technologies it is possible to have multiple warm or even hot copies of your data, available to your production systems should a problem occur with your live data.

7. DR/BCP

Disaster Recovery (DR) and Business Contingency Planning (BCP) play an important role in today's business critical IT infrastructure. Without these being implemented, a major incident affecting the physical IT environment can result in major loss of business and even the complete collapse of the business.

DR is the technical process and technologies involved in recovering from a disaster, while BCP is the overall procedure for the business, covering risk assessment, contingency planning and disaster recovery.

A number of issues need to be addressed in considering your DR and BCP requirements. These issues are outlined below.

7.1. Costs

An important consideration is the costs involved in implementing a DR environment. A complete duplicate of your existing production systems will be expensive, doubling the costs of your production implementation.

A possible way of reducing costs is to share resources with other companies. Hosting DR/BCP equipment at a specialist company, who also host these facilities for other organisations, could reduce the costs of implementing your DR/BCP environment.

Other options would depend on the overall requirements. Sharing server resources in your DR/BCP environment would provide a reduced service should a disaster occur. But it would provide a level of continuity in order to keep your business running, hence minimising the impact of the disaster.

7.2. Requirements

Careful consideration needs to be paid to the requirements of a potential DR/BCP environment. Is there a need to provide an exact level of service comparable with the existing production systems? Would a reduced level of service be acceptable and allow the consolidation of servers in the DR/BCP environment? What are the legal, regulatory or business requirements that need to be considered? Are there any Service Level Agreements (SLAs) that need to be met, internal or external, to your organisation?

7.3. Infrastructure

If your DR/BCP environment needs to be available quickly in the event of a disaster, consideration needs to be paid to the technologies involved. The use of SAN or NAS storage can assist as they have the ability to snapshot data between storage arrays, making DR/BCP implementations as simple as copying data across the network. This obviously depends on the network infrastructure available although the bandwidth necessary to implement this may not be as great as it may appear due to the methods used to copy the data.

Good backups are essential in implementing a DR/BCP environment. They provide the ability to restore data quickly at the remote site, and can be used for testing the integrity of the backup infrastructure at the production site.

7.4. Standby Options

Again there are a number of options available for the way the DR/BCP environment is made available in the event of a disaster. Some of the options available are :

- having to implement the hardware infrastructure and restore backups
- having a standby system available with data that is regularly restored but which may not be fully up to date at the time of a DR situation arising
- having a hot standby system which replicates data from the production system, making it essentially a mirror.

The options available all have varying costs and will depend on the requirements and available budget.

7.5. Scenario Testing

Whatever option is chosen, the process of switching to the DR/BCP systems needs to be tested thoroughly and on a regular basis, with all changes to the production environment being filtered through to the standby environment to ensure integrity. The switching processes and procedures should be tested on a regularly to ensure that they still work, that the processes do not need modifying and that all staff members are familiar with them. They should also be re-tested when any significant changes are made to the existing production infrastructure.

If these tests are not carried out, to verify the implementation, it may well end up being pointless having a DR/BCP environment. Switching a production system to an untested DR solution at the point of failure may well involve just as much work as implementing a completely new environment.

A permanent DR site also offers the opportunity to test that backups are satisfactory, by making the DR environment available for testing the restore procedures. This way any faults with backup processes can be identified and rectified before they become a serious issue.

8. Conclusions

Implementing a high availability solution involves more than just having some shared storage between two servers. It involves a whole range of issues, from system and application

architecture, physical environment, hardware resilience, data availability and integrity and business continuity.

It is not enough to implement a solution and then never touch it again. Testing of the installed architecture is important in ensuring ongoing resilience. Changes in the production environment, whether infrastructure or application, needs to be tested in the DR environment to ensure integrity.

More often than not, any solution will be a trade off between cost, performance, security and availability. Many of these areas go hand-in-hand but they can also impact one another and priorities may need to be decided in order to reach a compromise.

If corners are cut in the design phase or at implementation time then the results could turn out to be disastrous. The best solution may not be the most expensive but likewise, it is unlikely to be the cheapest. It is worth spending the time in advance to ensure that all of the components fit together, that all of the issues have been considered and that the solution offers the best value for money.