KEMP TECHNOLOGIES WHITE PAPER

How to obtain high availability for MS applications

High availability describes a concept that can also be called operational continuity over a particular period of time. This level of service is reflected by the amount of up time you are able to provide to your suppliers, clients and staff to have access to the system as a whole, the services you provide and the core of your business; the applications.

Availability is completely subjective and really depends on your business's needs and the success of your operational continuity is a combination of achieving business goals and how well you deal with your technical requirements. As you optimise the performance of your Microsoft applications your target needs to be to keep your applications and workloads up and running by combining the performance of hardware platforms and mission critical applications.

MICROSOFT: COMMITMENT TO HIGH AVAILABILITY

Microsoft is fully comitted to maintaining high availability and offer a series of suggestions about how to prepare your network to maintain operational continuity, here are some key recommendations:

- Always select the most robust hardware platform that your budget can afford, this will help you greatly
- Adopt the Windows Server ®2008 High Availability Program for Windows Server ® Enterprise
- Follow the Windows Server® for Itanium-Based System
- Embrace the Windows Server® 2008 Datacentre program

Having followed this advice you will have access to the programs, support and useful guidance from Microsoft that will ensure you make the right decisions and choose the right hardware platforms that will keep your business running.

Having given its advice on how to deploy the right hardware platforms, Microsoft turns its attention to what you need to do at the application level. Microsoft has built in high availability into their applications that could well be essential to your business. This is done by being able to replicate data across multiple instances of the applications.

IMPORTANT KEY POINTS

- Ensures service availability
- Guarantees business continuity
- Implemented in all versions since Microsoft® Exchange 2007
- Implemented in all versions of Microsoft® SQL Server 2005

With the launch of Windows Server 2008, both the Enterprise and Datacentre editions shipped with the ability to keep the operating systems and applications running in high availability, this is done by clustering the servers, how does this work?

WHAT IS A SERVER CLUSTER?

A cluster consists of a group of networked servers that are often called nodes. These nodes are able to run the server content of any of the other servers in their cluster and are ready to take up the load should one or more of the other servers in the network fail because of hardware or software failure. Clustered computers are aware of their peers, continually monitor peer level performance and can make decisions when to take over from a server in the cluster when it is considered, thanks to pre-programmed algorithms, that intervention is required.

WHAT ARE THE KEY APPLICATIONS THAT REQUIRE SERVER CLUSTERING?

Business critical applications are at the top of the list for key applications that require the benefits of server clustering. You could make a list of the applications that you run on your network servers and decide on the impact that loss of these applications will have, these can include:

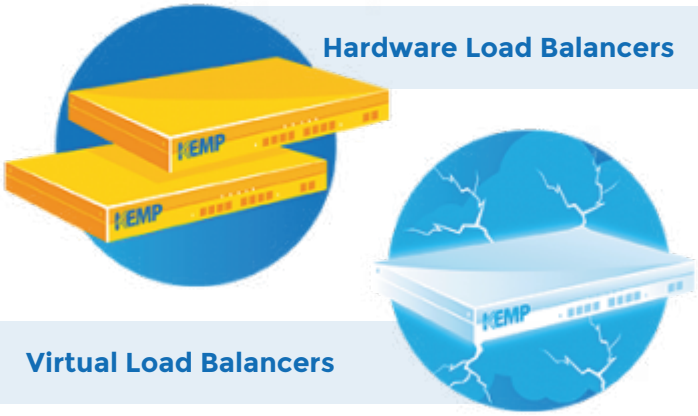
CRITICAL

- Online banking services
- e-commerce applications
- Company email applications
- All line-of-business (LOB) class applications

SERIOUS

- Unified communications and conferencing applications
- CRM systems
- Collaborative applications for example Microsoft SharePoint

LOAD BALANCING FOR ANY INFRASTRUCTURE



WHAT CAN CAUSE SERVICE OUTAGE AND APPLICATIONS TO STOP RESPONDING?

It is important to take into account two 'loss of service' scenarios. In order to do this analysis it is important to spend a moment talking about the OSI seven layer model, basically each layer defines a function of an element in the network, starting from the most basic and finishing with the top level or most complex elements. On this basis we find that the physical cabling of your network is categorised as Layer 1 networking and the applications are categorised at Layer 7, with 5 other layers in between. With this in mind, let's focus on the 2 Layers that are important for high availability, these are:

- Layer 4 the network layer
- Layer 7 the application layer

Layer 4 load balancers check the performance of the servers themselves, each server could be virtualised and running many applications. Layer 4 load balancers are able to monitor the health of the server and decide whether to take it out of the network or to continue to use it; these load balancers cannot monitor the health of the applications. The result is that the server could be performing perfectly but the application (for example Microsoft Lync) is not and the load balancer continues sending access requests to it. A good example of a popular Layer 4 load balancer is Windows Network Load Balancer or WNLB.

Layer 7 load balancers however health check the performance of the applications on the servers. These hardware and/or virtual load balancers are able to monitor the performance of each individual application and should performance fall below the KPIs that have been defined by the network management, switch the service to back-up servers. This form of load balancing is naturally more precise and effective than only using Layer 4 load balancing.

MICROSOFT WINDOW SERVER 2008 HYPER-V® DEFINES THE CHANGES

One of the inherent weaknesses of failover clusters is that the applications (as they run) need to monitor the performance of each other. This is a weakness because the applications should be optimised to perform for their users and not be obliged to act as "traffic cops" and monitor performance.

Windows Server 2008 Hyper-V® has introduced the ability for the supervisor layer to intervene without troubling the application server functions. Windows Server 2008 Hyper-V® offers new frontiers for high availability this functionality is known as quick migration. It combines failover clusters with server virtualization; quick migration is aware of virtualised servers and the physical hosts that run them. This combination is a fundamental step forward as it ensures that no single physical server becomes a vulnerable point of failure for a network.

DEALING WITH DOWNTIME

Downtime effectively occurs for two reasons that are either planned or unplanned. Serious downtime can be caused by:

- Network failure (either LAN or WAN)
- A server fault resulting in it becoming offline

Sometimes it is necessary to take servers offline for planned events including maintenance of the hardware or upgrades to applications or server operating systems. Unplanned downtime can take place at any moment and is beyond the control of the IT department administrators. Causes can be minor issues such as a hard disk or power supply that fails due to a catastrophic event such as a fire, a flood or an earthquake. One of the important points to take note of is that downtime, be it planned or unplanned, will eventually take place and it is not a case of *'if it happens'* but rather *'when will it happen'.*

Making sure your servers are located in a secure setting is of prime importance. For example, if your servers are located in a region susceptible to hurricanes, the premises should be constructed as hurricane proof. In addition comprehensive firefighting installations should be installed to protect your servers from that risk. However you can never be 100% sure your premises are invincible and so making provision for back-up facilities in a different location makes good sense. In-turn, the intelligent use of geographic load balancers also makes sense, allowing the diversion of traffic to your back-up site(s) should the primary site be taken off line.

Regular server maintenance allows a clean-up of the server, restoring it to its original performance levels. Moreover, having installed load balancers, back-up servers and increased the redundancy in your network means these server outages will have less effect on your users. If maintenance is not performed, minor problems in your servers will eventually grow more serious and the server will stop working. Therefore, as you plan your back-up facilities, consider the cost to the business of unplanned downtime both in terms of business lost as well as damage to the reputation of the organisation.

HARDWARE RELIABILITY NEEDS TO BE MEASURED

Importantly, the terms server up-time and server availability should not be confused. Servers could be operational but might not be available to users because a component in the network (such a router, a firewall or WAN equipment) could have failed; this counts against server availability. By selecting servers with dual power supplies and multiple network cards you can increase their reliability, however to really achieve a High Availability network it is important to install two or more load balancers configured in high availability mode.

DEFINING THE DOWNTIME RULES

If you ask an IT Manager about the permitted levels of downtime an organisation targets, the reply needs to be more than just an approximate percentage, but instead a specific figure. Actual downtime values, set on an annual basis, are as follows:

- 99% = 87 hours 35 minutes
- 99.9% = 8 hours 45 minutes
- 99.99% = 52 minutes 35 seconds
- 99,999% = 5 minutes 16 seconds

The cost of minimizing your permitted downtime varies server by server and is more complex because different server functions have different levels of criticality. A print server going off line is more likely to be annoying than critical, however it is a different matter if a mission critical database server fails as the damage to the business is immediate. These different levels of criticality should be considered when estimating the costs for raising the reliability of a system. With the above figures in mind, it could cost $95,000 to raise your reliability on a server from 99.99% to 99.999% but where a business would only loose $1,000 a minute due to downtime, the investment would not make a good return.

LOAD BALANCING FOR ANY INFRASTRUCTURE

Bare Metal Load Balancers

Cloud Load Balancers

Perhaps the most intelligent method of measuring server performance is not whether it can handle 80, 100 or 200 sessions simultaneously but instead, the effective time it takes users to complete their transactions. Using an e-commerce site as an example; if users are unable to complete their purchases successfully (despite servers still running), the organisation in question would lose revenue as disappointed potential customers abandon the site. Therefore, the primary point of concern should not be the number of users who can connect but instead the percentage of users who can complete their transactions at peak traffic periods.

MICROSOFT APPROACH TO HIGH AVAILABILITY

High availability solutions owe their success to how much redundancy is deployed in a network to minimize the risk of a single point of failure taking out mission critical servers. By employing a combination of high performance network servers from leading vendors together with load balancers deployed in high availability mode, the impact of a server failure can be reduced. Microsoft have taken the ability of its products and programs to aid the hardware platforms to maintain high availability and keep critical business IT systems up and running.

ADVANTAGES OF MICROSOFT HIGH AVAILABILITY

Microsoft has maintained a strong strategic objective to help its users maintain high availability with the Microsoft applications that they use. High availability is integrated into Microsoft products and programs including Windows Server® 2012 which comes complete with the High Availability Program for Windows Server 2012 Enterprise and Windows Server® 2012 Datacentre. It is recommended that, in order to deploy a sufficiently powerful and reliable system, the relevant high availability recommendations for the type of hardware platforms are followed.

Compared with Microsoft Exchange 2007 the design of Microsoft Exchange 2010 is completely different and high availability is actually built into the core of the application, allowing the support cluster service availability automatic recovery and data availability on an end-to end basis. The introduction of this new streamlined method of core architecture design, known as database availability group (DAG), means that the task of cluster implementation and maintenance has been greatly simplified.

The support overhead for IT departments maintaining Microsoft applications and programs on server clusters has been considerably simplified with the later generations of the core applications Lync, SharePoint and Exchange. This is a great benefit in terms of the complexity of the skills IT departments require just to maintain their clusters and the amount of time necessary for maintenance work.

In addition, Microsoft has expanded their recommendations for best practice (extending beyond the advice for network servers) to now include the deployment of hardware network load balancers. In-fact, since the launch of Microsoft Exchange 2010, network managers have been specifically told to use certified network load balancers instead of relying on Windows Network Load Balancer (WNLB) as they had done previously for Exchange 2003 or Exchange 2007.

In conclusion Microsoft has continued with its role of acting as a trusted adviser to its users regarding the set up implementation and clustering of servers for Microsoft applications and programs. Over time this role has extended to cover the complete cluster architecture and supporting infrastructure.

01285 868 500
sales_uk@zycko.com

Zycko Limited
Inda House
The Mallards
Broadway Lane
South Cerney
Cirencester GL7 5TQ