Realtime publishers

The Essentials Series: Configuring High Availability for Windows Server 2008 Environments

The Art of High Availability

by Richard Siddaway

Copyright Statement

The Art of High Availability

All organizations are becoming increasingly reliant upon their computer systems. The availability of those systems can be the difference between the organization succeeding and failing. A commercial organization that fails is out of business with the consequences rippling out to suppliers, customers, and the community.

This series will examine how we can configure our Windows Server 2008 environments to provide the level of availability our organizations need. The topics we cover will comprise:

- The Art of High Availability—What do we mean by high availability? Why do we need it, and how do we achieve it?
- Windows Server 2008 Native Technologies—What does Windows Server 2008 bring to the high-availability game, and how can we best use it?
- Non-Native Options for High Availability—Are there other ways of achieving high availability, and how can we integrate these solutions into our environments?

The first question we need to consider is why we need highly available systems.

Why Do We Need It?

This question can be turned on its head by asking "Do all of our systems need to be highly available?" The answer for many, if not most, organizations is no. The art of high availability comes in deciding which systems need to be made highly available and how this is going to be achieved. When thinking about these systems, we need to consider the effects of the systems not being available.

Downtime Hurts

Downtime is when the system is unavailable to the user or customer and the business process cannot be completed. If the server is up and the database is online but a network problem prevents access, the system is suffering downtime. Availability is an end-to-end activity. Downtime hurts in two ways: If a system is unavailable, the business process it supports cannot be completed and there is an immediate loss of revenue. This could be due to:

- Customer orders not being placed or being lost
- Staff not working
- Orders not being processed

The second way that downtime hurts is loss of reputation. This loss can be even more damaging in the long term if customers decide that your organization cannot be trusted to deliver and they turn to a competitor. The ability to gain business increases with ease of communication and access. The converse is that the ability to lose business increases just as fast if not faster.

Critical Systems on Windows

Critical business systems are hosted on the Windows platform. These can be customer facing or internal, but without them, the business grinds to a halt. Email may not seem to be a critical system, but it is essential to the modern business. More than 60% of person-to-person communication is via email in most businesses. This includes internal and external communications. If a company is non-responsive to communications, it is judged, perhaps harshly, as being out of business. This can become reality if it progresses too long.

24 × 7 Business Culture

The "Global Village" concept has been accelerated by the adoption of the Internet for business purposes. Globalization in this case means that business can come from anywhere in the world—not necessarily your own time zone. If your business competes at this level, high availability isn't an option, it's a necessity.

Legislation

Industries such as the financial services and health sector have a requirement to protect the data they store. This requirement can involve the availability of the data. In other cases, the systems must be highly available to meet safety requirements.

Once you know why you need it, you need to define what is meant by high availability.

What Is High Availability?

High availability is usually expressed in terms of a number of "9"s. Four nines is 99.99% availability. The ultimate goal is often expressed as 5 "9"s availability (99.999%), which equates to five and a quarter **minutes** of downtime per **year**. The more nines we need, the greater the cost to achieve that level of protection.

One common argument is scheduled downtime. If downtime is scheduled, for example, for application of a service pack, does that mean the system is unavailable? If the system is counted as unavailable, any Service Level Agreements (SLAs) on downtime will probably be broken. In hosting or outsourcing scenarios, this could lead to financial penalties. However, if scheduled downtime doesn't mean the system is counted as unavailable, impressive availability figures can be achieved—but are they a true reflection of availability to the users? There is no simple answer to these questions, but all systems require preventative maintenance or they will fail. The disruption to service can be minimized (for example, the patching nodes of a cluster in sequence) but cannot be completely eliminated. Probably the best that can be achieved is to ensure that maintenance windows are negotiated into the SLA.

These measurements are normally taken against the servers hosting the system. As we have seen, the server being available doesn't necessarily mean the system is available. We have to extend our definition of highly available from protecting the server to also include protecting the data.

The Clustering Service built-in to Windows is often our first thought for protecting the server. In the event of failure, the service automatically fails over to a standby server, and the business system remains available. However, this doesn't protect the data in that a failure in the disk system, or even network failures, can make the system unavailable.

> **Cross-Reference**
> We will return to these ideas in Articles 2 and 3.

Do We Still Need to Back Up?

One common question is "Do I still need to take a backup?" The only possible answer is YES! High availability is not, and never can be, a substitute for a well-planned backup regimen. Backup is your ultimate "get out of jail card." When all else fails, you can always restore from backup. However, this pre-supposes a few points:

- Test restores have been performed against the backup media. The last place you want to be is explaining why a business-critical system cannot be restored because the tapes cannot be read.
- A plan exists to perform the restore that has been tested and practiced. Again, you don't want to be performing recoveries where the systems and steps necessary for recovery are not understood.

Backup also forms an essential part of your disaster recovery planning.

Disaster Recovery vs. High Availability

These two topics, high availability and disaster recovery, are often thought of as being the same thing. They are related but separate topics. High availability can be best summed up as "keeping the lights on." It is involved with keeping our business processes working and dealing with day-to-day issues. Disaster recovery is the process and procedures required to recover the critical infrastructure after a natural or man-made disaster. The important point of disaster recovery planning is restoring the systems that are critical to the business in the shortest possible time.

Traditionally, these are two separate subjects, but the technologies are converging. One common disaster recovery technique is replicating the data to a standby data center. In the event of a disaster, this center is brought online and business continues. There are some applications, such as relational database systems and email systems, that can manage the data replication to another location. At one end of the scale, we have a simple data replication technique with a manual procedure required to bring the standby data online in place of the primary data source. This can range up to full database mirroring where transactions are committed to both the primary and mirror databases and failover to the mirror can be automatically triggered in the event of applications losing access to the primary. In a geographically dispersed organization where systems are accessed over the WAN, these techniques can supply both high availability and disaster recovery.

We have seen why we need high availability and what it is. We will now consider how we are going to achieve the required level of high availability.

Achieving High Availability

When high availability is discussed, the usual assumption is that we are talking about clustering Windows systems. In fact, technology is one of three areas that need to be in place before high availability works properly:

- People
- Processes
- Technology

People and Processes

These are the two points that are commonly overlooked. I have often heard people say that clustering is hard or that they had a cluster for the application but still had a failure. More often than not, these issues come down to a failure of the people and processes rather than the technology.

The first question that should be asked is "Who owns the system?" The simple answer is that IT owns the system. This is incorrect. There should be an established business owner for all critical systems. They are the people who make decisions regarding the system from a business perspective—especially decisions concerning potential downtime. A technical owner may also be established. If there is no technical owner, multiple people try to make decisions that are often conflicting. This can have a serious impact on availability. Ownership implies responsibility and accountability. With these in place, it becomes someone's job to ensure the system remains available.

A second major issue is the skills and knowledge of the people administering highly available systems. Do they really understand the technologies they are administering? Unfortunately, the answer is often that they don't. We wouldn't make an untrained or unskilled administrator responsible for a mainframe or a large UNIX system. We should ensure the same standards are applied to our highly available Windows systems. I once worked on a large Exchange 5.5 to Exchange 2003 migration. This involved a number of multi-node clusters, each running several instances of Exchange. One of the Exchange administrators asked me "Why do I need to know anything about Active Directory?" Given the tight integration between Exchange and Active Directory (AD), I found this an incredible question. This was definitely a case of untrained and unskilled.

Last, but very definitely not least, we need to consider the processes around our high-availability systems. In particular, two questions need to be answered:

- Do we have a change control system?
- Do we follow it?

If the answer to either of these is no, our system won't be highly available for very long. In addition, all procedures we perform on our systems should be documented and tested. They should always be performed as documented.

Technology

Technology will be the major focus of the next two articles, but for now, we need to consider the wider implications of high availability. We normally concentrate on the servers and ensure that the hardware has the maximum levels of resiliency. On top of this, we need to consider other factors:

- Network—Do we have redundant paths from client to server? Does this include LAN, WAN, and Internet access?
- Does the storage introduce a single point of failure?
- Has the operating system (OS) been hardened to the correct levels? Is there a procedure to ensure it remains hardened?
- Does our infrastructure in terms of AD, DNS, and DHCP support high availability?
- Does the application function in a high-availability environment?

Costs

Highly-available systems explicitly mean higher costs due to the technology and people we need to utilize. The more availability we want, the higher the costs will rise. A business decision must be made regarding the cost of implementing the highly-available system when compared against the risk to the business of the system not being available.

This calculation should include the cost of downtime internally together with potential loss of business and reputation. When a system is unavailable and people can't work, the final costs can be huge leading to the question "We lost how much?"

Summary

We need high availability to ensure our business processes keep functioning. This ensures our revenue streams and business reputation are protected. We achieve high availability through the correct mixture of people, processes, and technology.