white paper

Systems Availability

On the IBM System i

CCSS
Monitoring your system

What does Systems Availability mean to you?

In today's global economy, poor Availability Management can cost a manufacturing company as much as 10k in lost revenue per day if they are unable to ship their products. For financial organisations, the scenario is worse still, with conservative lost trade estimates reaching 10k per second, should they be unable to process credit card orders or provide real-time trading to their customers. Many companies who have been unfortunate enough to experience these financially devastating circumstances have been able to trace the root cause directly back to unexpected outages and system downtime.

By understanding your organisation's level of Systems Availability, much of the fear and financial horror of 'What if' scenarios can be eliminated.

There are many phrases currently in use within our industry, such as Systems Management, Service Management, Performance Management, Asset Management, Change Management, Problem Management, Operations Management and High Availability, that allow simple Systems Availability to be eclipsed. For the purposes of this document we will define Availability Management as the bridge between simple Systems Management and the High Availability product offerings.

Availability Management, in the current plethora of offerings, can encompass various types of hardware (computer systems, cabling systems, telephone systems and mobile devices), physical and software security, systems software, internal business applications, external Internet applications, multiple sites, staff and various combinations of these items, across date, time and geographical boundaries.

IT budgets have been continually driven down over the last decade, so it is more likely that less experienced staff have replaced technical personnel that may have been employed earlier. Also, due to budgetary concerns or the sheer volume of work involved, it is not always possible to implement an enterprise wide solution, so more often than not the focus turns to certain key areas to which a solution is sought.

As IT managers have come under continuing pressure to assure the high quality of IT services, governed by internal service level agreements and by the requirement to deliver application responsiveness to web users coming in from outside the firewall, the systems management software used to oversee the network, has had to evolve from simple device management to support applications and provide evidence for service levels and IT budget spending. It is no longer acceptable for an IT infrastructure to simply be managed via discrete tasks, such as server availability or network bandwidth although these are still important. There is now a need to manage the level of service provided by all of the individual pieces, including your business applications, viewed as a single entity across the enterprise and allow any problems that may arise to be dealt with quickly and efficiently.

In the last few years, with the further development of LPAR functionality and extensive support for applications written for other operating systems, such as AIX, Windows or LINUX, the System i has changed almost beyond recognition, when compared to the early AS400s.

At first glance, server consolidation is a mechanism that appears to reduce costs and simplify the systems management tasks. However, maybe this tactic simply reinforces the need for better Availability Management in the first place.

Why does my organisation need Availability Management?

What if you were in the driver's seat of an automobile built for optimum performance and then were told that you had to drive at high speeds without the assurance of a speedometer, fuel gauge or any of the other controls? Would you feel comfortable?

What if you had a ticket to fly on the safest plane in the world, but the pilot was flying without any flight instrumentation? Would you feel secure? More to the point, would you risk it?

Eighty percent of people asked, said that they would avoid these risky situations at all costs. Ironically, these same people are running their multi-million businesses in this same way, managing business critical systems and applications blindly.

Effectively managed companies use a combination of processes, tools and metrics to identify and visualise potential pitfalls in business and to measure long-term success.

By reporting today's performance metrics and analysing past trends, Availability Management bridges the proactive steps of performance management to ensure optimal high availability.

What benefits can I expect?

Organisations, regardless of size, can experience significant benefits from deploying a successful Availability Management strategy:

- Significant decrease in downtime
- Reduction in related business & revenue losses
- Improved service levels & overall customer service
- Decreased support related costs
- Increased employee productivity from proactive management
- Ability to capitalise on future business opportunities
- Provides a competitive advantage
- Key to optimal output from business applications & processes

System and message monitoring tools may not be glamorous or sexy, but without them it's possible for you to lose large amounts of revenue very quickly. If any part of your system is slow to respond, information is unavailable or, in the worst case, access is completely down (even if just for a few minutes), it costs you money and potentially profits. With your 24/7 business comes a greater commitment of uptime, availability, and functionality. All

organisations need the ability to identify any potential system problems and quickly notify IT staff of these issues, before the situation reaches a crisis point.

Implementing Availability Management

Organisations considering implementing a successful Availability Management strategy need to consider the following points of standardisation:

- ***Data replication software and business survival without its core applications***

What is the implication to your business if the backup system falls too far behind the central system and you have a central system failure?

If you were not aware that there was a lag and a failure should occur, you would be in danger of losing a portion of your business data and most probably future revenue as well. Whether you have a complete Disaster Recovery plan or simply a Backup and Recovery procedure you need to be aware of any potential gaps to plug. The nearer you can get to a complete recovery of data at the same point in time, the easier it will be for your business to commence trading again.

Equally a loss of revenue cannot be afforded through a simple controller or router failure cutting off a sales office for a period of time through a normal working day. Communications hardware and software is more resilient today, but contractors can still cut through cables and cleaning staff can still unplug equipment to plug in their vacuum cleaners. These events need to be monitored for and contingency plans created.

Application monitoring is also required now, to ensure that critical subsystems or jobs are running as required, for example as part of your Customer Order application. Modern applications tend to have a level of monitoring built into them, but if the application malfunctions you may not be able to access that information and by then it will be too late.

- ***Service Level Agreements, departmental budgets or hosting services***

How long does it take to generate your daily, weekly or monthly reports and how much revenue could potentially be lost if you are unable to produce these reports on time?

Should a Director, Manager or Customer query a response problem, can you quickly and easily explain and perhaps even provide evidence of the cause of the problem?

If you need to account for systems usage due to billing or service agreements, you need to ensure that you can provide reliable, easily understood reports and keep a history of them, as well as the base data that was used to create them, in case it is required to settle any queries that may arise. Being able to provide this information efficiently in a graphical, as well as a detailed format, is essential.

- ***Poor response times or resource hungry users***

How often do your users ring up complaining of poor response times, only for you to discover after the event that somebody has been running an interactive query, or using a poorly written SQL statement, or that a developer ran an interactive restore operation?

Those sorts of problems can be dealt with and a response automated. However, one of the most pernicious problems is the job that loops. Jobs can misbehave because of users doing things that they shouldn't or because of programming errors. A job can go into an unending loop, consuming more and more resources, at the expense of other users.

If a program goes into an endless loop, and that loop is actually doing something, then it will not always show up as the job using the highest amount of CPU. It may be performing a lot of output to the database, and the only thing the user will see is that something is filling up all the disk space.

Once a rogue job has been found, the operator has to make a note of the job name, user and number, and then use a different interface to actually do something about it. If this problem occurs out-of-hours do you have procedures in place to ensure that the business does not suffer an outage?

Do a particular group of users (e.g. POS tills during a seasonal Sales period) appear to be having more performance problems than others? Do they appear to need more resources made available periodically?

It may simply be that you have discovered a requirement for “on demand” computing resources to be budgeted for. If so, you need to be able to identify what additional resources will be required and the period during which this requirement exists.

- ***LPAR systems or capped processors***

Partitioning is a method that allows the user to divide a single system into a number of logical systems, each of which appears to be a separate machine, with its own name, clock, network address, disk space, data and operating system.

When a system is divided into partitions, each partition is given a fraction of the total processing power and memory available.

Partitions do such a good job of pretending to be different systems that unless you already know the physical layout, it's very difficult to identify what's a partition and what's a system. It's vital to know the physical layout in a number of situations:

- You can only move processor power and memory between partitions on the same machine.
- If you need to shut down systems for maintenance, you need to know which partitions will be affected.

Therefore you need to easily confirm the affected systems before making critical decisions that could otherwise result in unscheduled downtime. Before amending the resource allocations of each partition on the box, you need to know whether there are any spare resources that have not been allocated to any partitions, and are currently being under utilised.

The monitoring of processors, especially in an LPAR environment where small percentages of a shared processor can be used to run a system, is perhaps even more critical than the disk units. What would be the implication to the business if a processor failed or is only partly functional and eventually brings the system down?

Database CPU and Secondary CPU are types of processor usage that are limited on some models. When the usage is at the maximum allowed, you have no way of knowing why this is so, or which job is using it. On many System i models, the machine limits the amount of CPU processing power that can be used for certain purposes. On some systems, for instance, the amount of processing power used for interactive work is limited to a certain fraction of the total available to the machine.

On other models, the way in which this limit is implemented by the machine has a serious effect on the entire partition. As the amount of processor being used for interactive work approaches the limit, the system starts reducing the amount of power available to everyone - not just the offending jobs. By the time the system has reached the limit, there is no power left for any other work at all.

So a system can be using 8% of the processor for interactive work, and have none at all left for the web server and the payroll cheque printing run. It would help to monitor the absolute limit, and the point at which the system will start throttling the processor. This would allow you to determine thresholds for limited resources, and take action before this sort of activity becomes a problem.

- ***Sudden disk growth***

Have you ever had situations where disk usage suddenly increases at an alarming rate for no obvious reason?

In the past, the System i had a single way of storing data, by putting files into libraries. Now, we can store data in a complex structure of directories that are known as the IFS. A large number of modern applications, whether they are ERP, CRM, Domino or WebSphere can store vast amounts of information within a deep, complex structure.

It's very difficult to see what's in there - there are few tools that let you look into it at all and getting to grips with the complex structure is challenging. This becomes a particular problem with some applications that are constantly writing log files in unexpected places, consuming large amounts of disk space.

Can you easily generate lists of objects that are of particular concern?

- Large files created during the last week.
- All Log files in an ERP application’s directory.
- Files that haven't been used for more than a year.
- Files that are more than 10 gigabytes in size.

Originally, the type of programs that utilise the IFS were usually written for UNIX type systems, and later converted to the System i. A programmer on UNIX assumes that there are experts around who know exactly what the programs and the operating system are

doing. Such a program expects that it can make files as large as it pleases, and that someone else will come along and clean up after it.

Part of the problem is in knowing whether the change in disk usage is a sudden jump, related to special processing (such as year end), or whether it's something gradual, in which case the growth is likely to continue.

How will you determine the pattern of growth? Can you use this pattern confidently to decide when additional disk space is required? Will you know which specific files have grown, or quickly identify when a particular file has been deleted?

- ***Early warning of potential disk drive problems***

What would be the implication to the business if a disk drive, RAID set or a processor failed, consequently bringing the system down? The System i isn't a small machine any more and a large machine needs significantly more disk storage and major processing power. Early warning of potential problems that can be investigated before they become critical is vital.

Disks are grouped into sets, called an Auxiliary Storage Pool (ASP). Some ASPs are always available to a partition but others can be switched between partitions and are called *Independent ASPs*. A system can have up to 255 ASPs, of which 223 can be independent.

ASPs tend to be created, for example, to be used by journals to improve application performance. If there's a lot of work being processed for a single ASP, the disks can become overloaded. This will affect the performance of all work that uses data from the ASP, so you need to monitor the ASPs to see how busy they are. This can highlight performance bottlenecks that could affect your applications.

If you keep putting data into an ASP, it will fill up. Typically, when an ASP fills up, the system will start putting the extra data into another ASP, unless it is an independent ASP. This has a serious effect on performance; it can also pose a serious risk to the data in some situations. At the very least, the system needs to be taken away from the users to recover from this problem.

Monitoring the percentage of space used in each ASP is important. Do you know what action will be taken by the system when the ASP becomes full, and whether data compression is being used to squeeze more information onto the disks?

In Summary

This document is not meant to be a definitive answer to the myriad problems we meet day to day in business life, but it should highlight some of the potential pitfalls that may catch the unwary among us and provide guidance on creating a successful Availability Management strategy.

Analyse where your business is today and identify “what if” scenarios and their potential impacts on revenue and business. The points raised should be used as a basis to review your current operating procedures and ensure that any gaps can be plugged or at least budgets created to cover the eventuality. Ensure that your availability goals are synchronised with your business goals.

There are plenty of tools out there to review. Choose a tool or solution that will enable your Availability Management strategy. The best method is to first highlight any known problem areas, then those areas that are essential to be monitored, then those that are important and finally those that would be ideal once the first three categories have been dealt with.

Then try and pick a suite of products that give the best overall fit, rather than separate products for each area. Ideally the nearer you can get to a “plug and go” solution the easier it will be for your staff to quickly see the benefits of their new tools.

Integrated tools available on the same desktop will provide the best route for your technical staff to start using the products and then suggest their own enhancements as their knowledge of the products increase.

Achieving optimal Systems Availability with CCSS

QSystem Monitor is designed specifically for the System i and produces a detailed overview of the status of systems across the network. Critical functions that dictate system performance are measured and recorded. User defined events can also be configured and monitored. Historical data targets problems and allows staff to anticipate future requirements. System Managers can resolve recurring and existing problems throughout the network and support staff can detect and react to events on single or networked systems from a central console.

Thresholds can be devised and linked to **QMessage Monitor**, to allow support staff to be notified via various methods, of potential system problems before they develop into fully-fledged disasters.

QMessage Monitor provides the total solution to System i Message and Event Management, eliminating the need for manual monitoring. Capable of filtering, answering and escalating messages and log entries or user-defined events, QMessage Monitor makes easy work of Event Management.

Automatic notification via audio, visual, email, pager or mobile phone allows staff to keep in touch with their systems, whilst performing other tasks or outside of normal working hours. By restricting what an Operator sees to what the business regards as essential, you can reduce the chance of errors and manage all of your systems from one console.

QRemote Control brings benefits to those companies who value the ability to manage their systems, access application data, run commands and programs directly from their mobile phone, as well as simply receiving and replying to messages sent from the System i. Users benefit from the ability to access information and control multiple System i without the restrictions that come with onsite management.

Organisations can make use of this very flexible application to help them manage not only their System i systems management, but also business applications can be accessed via a mobile phone wherever their staff may happen to be, whenever they feel the need.

If you cannot monitor it, you cannot manage it.

Pro-active or re-active, let your Business choose!

For further information or customer stories on any of these products please visit the CCSS (Europe) Ltd website http://www.ccssltd.com or contact our sales or support staff via sales@ccssltd.com or support@ccssltd.com.