Best Practice Guide

Topic:
5 Steps to Achieving High Availability with Systems Management

Summary:

How to meet the demands of a 24/7 availability environment and reduce the risk of downtime events

CCSS
Monitoring your systems

CCSS develops, markets, and supports performance monitoring, message management and automation solutions for IBM® i servers, including Power Systems™ and System i™. An Advanced IBM Business Partner, CCSS develops powerful solutions to support some of the world's most demanding IBM i environments across many industries including pharmaceutical, insurance, banking, and logistics.

5 Steps to Achieving High Availability with Systems Management

Step 1: Create a Real-World Definition of a High Availability Environment

A hard working System i environment is never more visible (or expensive) to an organization than when it becomes unavailable. The expectation is for optimal performance on a 24/7 basis and beyond that, any system issues should be the remit of the IT Manager – not seen, not heard and most definitely not felt by the user community. Sadly, the realities of system availability can fall short of these exacting standards – whether the systems are truly available or not. This is because the definition of high availability means different things to different groups within an organization.

Some may say that unless a system fails, it is, technically speaking, still available. In the real-world however, these technical definitions will be of little persuasion to the user who is unable to access their application because a TCP port has stopped listening, or others who are experiencing poor response times, due to the 'lack of availability'. Similarly, when a system performs below agreed SLA standards but does not necessarily fail altogether, then the financial penalties will resonate louder than any justification. Creating a high availability environment must account for all definitions and perceptions. In short, when it comes to high availability, the system must be all things to all people at all times.

Available for Comment

"We've been accruing an unacceptable number of financial penalties because our SLAs are breached when HA software that relies on Journal Receivers, like the Audit Journal, is compromised by rogue jobs. The financial impact is extensive in each instance as we're obligated to make DASD available until the issue is resolved. Extra investigation time (sometimes overtime) is usually required to find the cause of the problem. How can we guard against this unnecessary expense?"

System Administrator, Manufacturing Industry, Asia

"Our organization invested heavily in a HA solution to ensure we remained operational following an unforeseen downtime event. We experienced a rapidly escalating CPU situation and the machine failed and switched over without incident, just as it should. The problem that caused the fail continued to consume CPU on the back up system. We had to work extremely fast to resolve the issue and at the same time feed the system with enough CPU to keep it from failing again. We don't want to be in that position again – the expense overall was huge - how can we guard against these types of problems in the future?"

IT Manager, Pharmaceutical Industry, USA USA

Step 2: Identify Threats to Availability

Identifying the threats to availability will require both a historical knowledge of past events, for example those leading to a specific downtime incidence, and also an examination of the lines of dependence that stretch from the user group back to the system. In some cases, these can also include dedicated High Availability (HA) solutions or other resources used for disaster recovery. Analysis of all these links between the system and the people who rely on them can help to determine potential areas of vulnerability where adverse conditions could not only impact availability, but also create serious secondary issues on the system.

Applications, jobs, sub systems and communication elements can all give rise to potential availability issues and as such, these key areas should be identified and monitored in real-time for abnormal status or performance conditions. The chart below looks at how some of these system elements could transform into availability issues and what their impact could mean for the organization experiencing them.

Area	Problem	Result
Objects		
e.g. **Journal Status**	Journal receiver is inactive	• **Performance degrades and entries are not processed in good time** • **The system falls over, HA switch occurs and large amounts of transactions are lost to users**
Communications		
e.g. **TCP Port Status**	TCP Port is inactive	• **Application generated messages sit in queues unanswered** • **Critical levels of unanswered messages are breached and the application is in danger of falling over**
Jobs		
e.g. **Subsystem Status**	Scheduler subsystem is inactive	• **Important jobs are not being scheduled to run e.g. Payroll** • **Employees are not paid on time**

Step 3: Automate Response to Issues

Once key areas have been identified, IT Managers can take steps to implement a pro-active approach to monitoring these in real-time. This will help to reduce any investigative time required, should these elements become inactive or perform in such a way as to potentially impact availability. Immediate notification can help to eradicate additional resource expenditure – as seen in circumstances where vast amounts of CPU or DASD are consumed and at such a rate as to possibly cause the system to fail or incur significant additional cost. As availability issues can occur at any time, the most effective real-time monitoring is used in conjunction with escalation procedures that synchronize to a defined chain of response within the organization and take account of staff working shifts and include a variety of communication channels (e.g. email, cell phone and pagers.) Managers can also attach pre-defined commands for certain conditions where this course of action is appropriate and sufficient, fully automating the process from real-time identification to response.

Step 4: Add Value to Reactive Solutions with Pro-Active Systems Management

For large organizations, the solution to the availability issue might be seen to be solved with a suitable HA solution or global mirroring environment. Whilst these are incredibly valuable in a downtime event, their purpose is to create an acceptable and immediate alternative system's environment following a downtime event – they do not prevent the event from occurring in the first place. That is the job of systems management. Reducing the risk of threats to availability - be it the risk of a downtime event at one extreme, or poor performance at the other, can be achieved with a pro-active approach to systems management issues.

By employing systems management as the first line of defense against threats to availability, IT Managers can achieve tremendous cost savings and ensure the system meets the 24/7 availability demands placed upon it.

Step 5: Select What to Monitor for Optimal Availability

Each System i environment may differ from the next when it comes to the number of systems, amount of resources and types of applications that run, however, there are common areas which should be considered when choosing what to monitor for optimal availability. The list below identifies these areas as issues detected at an early stage can help managers make a speedy resolution so they don't escalate to a stage where they pose a serious threat to availability.

Objects

- Journal Status
- Output Queue Status

Communications Status

- Controller Status
- Device Status
- Distribution Queue Status
- Line Status
- Network Interface Status
- Network Server Status
- TCP Port Status

Job Status

- Job Queue Status
- Job Status
- Subsystem Status