Improving Systems Availability

Background

With the emerging global economy, organizations of every kind and in all parts of the world are becoming increasingly dependent on their IT systems. E-commerce is making it possible to conduct business 24 hours a day, 7 days a week. New, powerful applications allow businesses and institutions to introduce unprecedented levels of computerization into their daily operations. The trend is clear. We all depend on reliable access to computer systems at all times.

This need has dramatically increased the importance of systems availability. Since companies and organizations rely heavily on computer systems to conduct their business, any downtime can seriously cripple their business. More than just lost productivity, downtime has come to mean lost revenues and even weakened market position.

At the very least, IT downtime can severely impact a business' operations and increase cost enormously. According to Computer Economics & Infocorp Consulting, systems downtime in 1996 cost American businesses $4.54 billion due to lost productivity and revenues. This cost is estimated to rise to $6.6 billion in 1999.

While the financial impact of IT outages is significant, regardless of industry, some are at far more risk than others according to a 1998 report by the Gartner Group.

Industry	Business Operation	Average Cost per Hour of Downtime
Financial	Brokerage operations	$6.5 million
Financial	Credit card/sales authorization	$2.6 million
Media	Pay-per-view television	$1.1 million
Retail	Home shopping (TV)	$113.0 thousand
Retail	Home catalog sales	$90.0 thousand
Transportation	Airline reservations	$89.5 thousand

As a result of these pressures, there is a strong demand for IT systems that are available almost continuously. Studies by Gartner Group and Dataquest indicate that there is substantial growth in the placements of highly available systems and technologies such as clustering, data shadowing and mirroring. The research demonstrates that this growth should continue.

Further, according to the same sources, some 58% of US businesses believe they require some form of computer processing, 24 hours a day, 7 days a week. Nearly 31% of US businesses require this 24x7 processing with uptime greater than 99.5%. Approximately 27% of these businesses require 24x7 availability with 99.9% uptime.

But what exactly does this mean? The expense of building systems that are continuously available can be daunting. Do all users really need this degree of availability?

What is high availability?

Availability means that a system is on-line and ready for access. A variety of factors can take a system off-line, ranging from planned downtime for maintenance to catastrophic failure. The goals of high availability solutions are to minimize this downtime and/or to minimize the time needed to recover from an outage. Exactly how much downtime can be tolerated will dictate the comprehensiveness, complexity and cost of the solution.

High availability is a convenient label, but its meaning is often misunderstood. High availability is not a specific technology nor a quantifiable attribute. Rather, it is a goal to be reached — one that has different meanings according to need. A variety of strategies, technologies and services are used to accomplish that goal.

At one end of the spectrum, high availability might simply mean a disaster recovery plan that will put an organization back on its feet within 24 hours. For small systems, this could mean something as simple as an uninterruptible power supply and a rigorous backup methodology. At the other end of the spectrum is the pinnacle of continuous availability, exemplified by robust workload-sharing solutions spread across multiple locations. Between these two extremes, are varying degrees of availability.

More than just technology

There is no such thing as a simple, easy and inexpensive high availability solution. Any given approach must strike a balance between true need and economics — and there are many ways to approach the problem.

Our approach to sizing up this as yet unknown entity is called the “bounded system,” which we define most simplistically as the environment — including hardware, software, processes and more — to be encompassed by our support agreement or guarantee. More specifically, it’s a group of interacting, interrelated, and interdependent system elements forming a complex whole and performing complex procedures, work flows, and tasks associated with accomplishing business requirements.

A bounded system may be configured and measured as either a *server centric* or a *network centric* configuration. A *server centric* configuration is typically a centralized server or server cluster configuration with availability measured as the percentage of time the online services are functioning at the server level and are potentially available to a minimum group of clients anywhere within the system’s domain.

A *network centric* configuration is typically based on a distributed architecture. Availability is measured as the percentage of time the online services are functioning at an end-user’s level and that a critical mass of end users are functioning during the customer’s specified online window.

Under either scenario, the bounded system configuration forms the basis for managing system availability and allows us to include or exclude certain system components.

At its simplest, a bounded system might include nothing but a CPU or data on a direct access storage device (DASD). These historically have been among the aspects of a system first addressed in providing enhanced availability. In today’s environment, however, it’s recognized that applications and/or communications paths will generally be included in a bounded system. So a more complex bounded system might include redundant CPUs, hot-swappable RAID drives with multiple access

paths, controlling software for the high availability cluster, multiple LANs and WANs with all the redundant hardware and lines this implies, end user devices with multiple network adapters and so forth. A bounded system might even extend to an application and its performance, with a requirement that the application be available at all times to *x* percent of users and that the average response to any user request be completed in not more than *y* seconds. This approach to availability differs significantly from the other companies that are attempting to sell availability on the basis of CPU reliability. IBM's methodologies allow us to encompass the entirety of the business system; from the physical processor through and including the end-user application.

Understanding the bounded system is a crucial and complex task that forms the basis for further assessments of the computing environment. Only through the creation of a bounded system can the customers current availability level (CAL) be accurately determined. While determining the parameters of the bounded system, we are almost certain to uncover numerous deficiencies in the system and associated system management practices. These can inhibit or prevent the system from achieving the customer's target availability level (TAL). By determining the gap between the CAL and the TAL, and working with the customer, we can develop a unique solution to address the customer's availability goals.

Here are descriptions of the technology and process areas we evaluate in the Bounded System creation effort:

Application management
Availability management
Capacity management
Change management
Metric management
Network management
Performance management
Problem management
Service level management
System recovery management

A high availability solution should focus on both preventing and avoiding problems in all the listed disciplines, that might lead to an interruption of service. Also it should focus on recovering quickly and minimizing the impact from any outages that do occur. Taking a proactive approach requires not only the proper hardware, but also the right mix of software and services to arrive at a total solution.

The need for software and services, in addition to infrastructure, is an important point. High availability is not achieved through technology alone; although technology is a key component of any solution. Purchasing a high-cost, fault-tolerant, state-of-the-art system does not necessarily mean that your business will achieve its desired level of availability. The technology requires proper service, administration, preventative support, recovery management and planning.

Key to any high availability solution is knowledge and planning. Systems availability must be assessed, preventive and remedial measures taken, and outage control plans put in place. A high degree of expertise is required if the solution is to be effective. Supporting such a solution may require a wide range of services that reach far beyond the initial stages. For instance, the system and software must be kept updated to meet changing demands. Often outside vendors may provide key components of the solution, such as off-site disk mirroring, monitoring services, or fully equipped recovery facilities.

All of these aspects of high availability require serious commitment. As the requirements for systems availability increase, so does the magnitude of this commitment. Few organizations are able to muster the resources, experience and expertise needed to properly deliver on the promise of high availability. Consequently, they must turn to others for help.

IBM Global Services is uniquely qualified to provide, support and be an active part of high availability solutions. It is one of the few vendors able to provide complete, end-to-end solutions for businesses and organizations in all industries. In addition to exclusive, world-class technologies, IBM offers an unparalleled breadth of high availability skills and services. These include:

- IT consulting
- systems and network management
- business continuity services
- hardware and software support services
- operations and administrative services
- site management and technology enablement.

Depending on the need, IBM can provide everything from planning and assessments to fully operational disaster recovery centers that take over an organization's entire data center operations in the event of catastrophe. IBM can also bring to bear its unique, proven methodologies that increase the effectiveness of its services while reducing cost.

Summary

In closing, consider the following points:

- High availability is an often misunderstood concept. It is a goal to be reached and one that has various meanings.
- Technology is a key part of high availability, but it does not stop there. Planning and ongoing support of the solution are just as critical.
- High availability solutions can be complex, including a mix of technology and services such as disaster recovery, consulting, assessment, management and more. The Bounded System is the key to creating a solution for the complex customer environment.
- It is important to team with a vendor qualified to provide the high level of support required to deliver an effective, meaningful, high availability solution.

IBM Global Services
Information Development Center
3200 Windy Hill Road
Atlanta, GA 30339
U.S.A.

Printed in the United States of America
8/99

IBM's Product Support Services organization in the United States, part of IBM Global Services, has successfully achieved registration to the ISO