



White Paper

metagroup.com



800-945-META [6382]

April 2004

A Comprehensive View of High-Availability Data Center Networking

*A META Group White Paper
Created for Cisco*

“A robust availability plan will include a combination of component level availability, networkwide recovery intelligence, system-level redundancy, interaction with endpoints (e.g., servers, storage), and operational best practices.”



METAGROUP

Contents

- Introduction 2**
- Understanding the Business Requirements 4**
 - Current-State Analysis..... 4*
 - Risk Assessment 4*
 - Process Assessment 6*
 - Conducting a Business Impact Analysis 6*
 - Securing Business Buy-In for Availability Planning 8*
- The High-Availability Framework..... 9**
 - Value-Based Metrics 9*
 - A Technology View of Business Continuance Levels 12*
- Defining Network-Centric Availability Metrics 14**
 - Data Center Network Recovery Time..... 14*
 - Storage Network Recovery Time..... 15*
 - Multiple Data Center Interconnect Network Recovery Time 16*
 - Access Network Recovery Time..... 17*
 - Recovery Access Objective..... 17*
- Building a Highly Available Data Center Network 18**
 - Network Availability Services..... 18*
 - Recovering Network Services 20*
 - High-Availability Implications for Network Operations..... 21*
- Closing the Loop 22**

Introduction

Furthering competitive advantage in the 2004 business environment requires the agility to adapt and evolve with shifting market dynamics. The extent to which information delivery is fluid yet controlled throughout the enterprise is directly related to the organization's ability to identify and react to market changes and thereby maintain and then improve its competitive positioning. At the heart of enterprise information systems is the data center. As the reliance on access to information has increased over time, so too has the importance of the data center. The consolidation of data center facilities has been a key industry trend, enabling enterprises to centralize resources and improve service levels to the business while decreasing ownership costs. Furthermore, the advent of server and storage consolidation has led to improved efficiencies through more effective use of processing power.

In addition to these major shifts in data center strategies, emerging trends such as application integration, Web services, and adaptive or utility computing promise to effect further change within the enterprise data center and increase the reliance that the business places on this critical infrastructure. Given the increased reliance on a centralized data center, the stakes associated with disruptions are higher than ever. As a result, it has become paramount that the data center be architected to ensure users have reliable and secure access to corporate services housed within the data center. Furthermore, the internal data center network must provide a highly resilient fabric to enable scalable, high-performance server interconnection, as well as the virtualization of computing, storage, and application resources.

Fundamental to a realistic and effective business continuity and high-availability data center strategy is the clear statement of its importance to the business. Specifically, the IT organization must seek and obtain business buy-in on the real-world business impact of system outages and application downtime. Indeed, many Global 2000 users fail to adequately consider or document the critical link between the IT organization's recovery plans and the business units' requirements. Taking the valuation of continuity to the next level requires the accurate measurement of the real costs of disruption. The business-case justification for data center availability must be tied directly to a credible estimate of business performance impact and should also be analyzed within the context of increasing regulatory requirements. Although gross industry-sector averages exist to guide users (see Figure 1), a thoroughly researched business impact analysis (BIA) should be a regularly updated part of the enterprise business continuity and high-availability data center strategy.

A Comprehensive View of High-Availability Data Center Networking

Figure 1 — Potential Loss of Revenue by Industry Sector

Industry Sector	Revenue/Hour	Revenue/Employee Hour
Energy	\$2,817,846.00	\$569.20
Telecommunications	\$2,066,245.00	\$168.98
Manufacturing	\$1,610,645.00	\$134.20
Financial Institutions	\$1,495,134.00	\$1,079.89
Information Technology	\$1,344,461.00	\$184.03
Insurance	\$1,202,444.00	\$370.92
Retail	\$1,107,274.00	\$244.37
Pharmaceuticals	\$1,082,252.00	\$167.53
Banking	\$996,802.00	\$130.52
Food/Beverage Processing	\$804,192.00	\$153.10
Consumer Products	\$785,719.00	\$127.98
Chemicals	\$704,101.00	\$194.53
Transportation	\$668,586.00	\$107.78
Utilities	\$643,250.00	\$380.94
Healthcare	\$636,030.00	\$142.58
Metals/Natural Resources	\$580,588.00	\$153.11
Professional Services	\$532,510.00	\$99.59
Electronics	\$477,366.00	\$74.48
Construction and Engineering	\$389,601.00	\$216.18
Media	\$340,432.00	\$119.74
Hospitality	\$330,654.00	\$38.62
Average	\$1,010,536.00	\$205.55

Organizations that have the most revenue and are most heavily dependent on online systems have the highest potential loss of revenue from (global) application and network outages. The above table shows average revenues by industry sector per hour (i.e., 24x365) and revenue by employee hour (i.e., assuming 2,000 hours per year). The percentage of the revenue actually lost depends on the criticality of the system that experiences the outage (e.g., degree of customer interaction, existing workarounds, peak periods) and the number of users affected by the outage or slowdown. Also, significant immediate losses can result in bad publicity and loss of customer trust that affects future revenues.

Source: META Group

To mitigate such potential revenue losses occurring from unplanned downtime, IT organizations and, more specifically, data center teams have dedicated substantial resources to improving the availability guarantee of corporate resources served by the enterprise data center. Although much attention has been given to the internal workings of the data center, server and storage components and facilities, IT organizations must not neglect the role of the data center network. This network must be capable of non-blocking, wire-speed connectivity while delivering a full suite of management, security, availability, application optimization, and replication services. The impact of a suboptimal network design can have grave consequences and negate all other investments the business may have made in the design and implementation of a high-availability data center.

This paper will explore the role the network plays in guaranteeing specific service levels and how it fits into the accepted methodology of data center availability metrics. This analysis seeks to expand the traditional parameters defining optimal data center design to include the network as a critical component in the implementation of a robust, highly available, yet manageable data center environment.

Understanding the Business Requirements

As part of any infrastructure project, the IT organization must ensure that it has a clear understanding of the needs of the business. Such alignment of business and IT priorities is critical to ensuring the infrastructure build-out meets its stated purpose and serves as a business enabler as opposed to a business hindrance. When planning a data center network, much thought must be given to the level of availability that is required and the way in which that network will meet the various availability metrics applied to world-class data centers. It is important to perform a current-state analysis of the business requirements, taking time to conduct a risk and process assessment to ensure optimal alignment between business and technology.

Current-State Analysis

High-availability solutions must reflect the operational and business exposures of the enterprise — not only the loss of physical assets, but also the risks associated with core organizational systems, including people, knowledge, reputation, and other intangibles.

Risk Assessment

Risk is not only about adverse events, but also about missed opportunities. Consideration of all operational areas is required to ensure that agility is not compromised by agreed-to risk mitigation solutions.

Many methods exist for grouping risks, starting with either categorizing risk (e.g., physical, reputation) or analyzing it from a functional approach (e.g., finance, distribution). Common business-level categorization groupings include business environment (competition, regulations), physical issues (fire, flood), compliance (governmental and industry regulations), data, application, system and network protection (security), health and safety (factories, environmental issues), extended enterprise failure (supplier or customer failure), and asset protection (secured assets). In parallel, the IT organization must assess the risk of disruptions due to system, storage, or network failures and data corruption caused by human error, technical failures, or directed/indiscriminate attacks.

In addition, the business must consider the fact that applications are increasingly interrelated to the degree that an outage or lower degree of availability of one application may cause another application to experience problems. As a result, it is important that the business identify these interrelations and take measures to ensure a lowest common denominator of availability across all applications. Identifiable risks need to be prioritized into manageable order before action can be focused on required solutions. For each significant risk, an assessment of both the likelihood of occurrence and impact on the organization is required. A potential scale for the likelihood of occurrence could be the following:

1. Rare
2. Moderately likely
3. Highly likely
4. Frequent occurrence

An impact assessment should be applied against the likelihood. A potential scale could be the following:

1. Minor impact
2. Minor to short term, significant impact to long term
3. Significant impact to medium term
4. Fundamental to continued operations

Once risks have been identified and prioritized (higher numbers first), the executive team has to evaluate the degree of the solution. This is where the BIA comes in, and the cost of the availability and business continuance solution is assessed against the financial impact.

Process Assessment

A growing number of IT organizations (ITOs) are analyzing operational process relationships and their links to business process requirements to determine service value/risk profiles and improve operational performance. These efforts often leverage the ongoing cooperation between lines of business (LOBs) and IT process owners aiming to close the gap between service quality expectations and perceptions as business needs and requirements evolve.

Identifying the most relevant operational processes to improve performance and evaluating the process innovation impact continue to challenge many organizations. Some users view process improvement and integration efforts as quick fixes, focusing on a few processes that are easy to analyze and monitor but that fail to produce the desired results. In these efforts, users should address four questions:

1. How should processes be performed to address future business requirements (where do we wish to be)?
2. How do we perform processes to deliver services (where are we now)?
3. Which process changes will be necessary (how do we get there)?
4. How will process performance be measured (how do we know we have arrived)?

Conducting a Business Impact Analysis

A key step in architecting a highly available data center infrastructure is assessing the current state of IT operations to identify optimum solutions for resuming key business processes and services in case of a disruption. The level of investment in data center infrastructure to mitigate risk of business service loss will depend on the results of the business impact analysis. The primary function of a business impact analysis is to identify and prioritize the minimum enterprise business continuity requirements to stay in business at certain levels of disruption. Such analysis is a necessity for creating an effective business continuity plan.

A business impact analysis of all business units and applications that are part of the business environment enables the project team to do the following:

- Identify critical systems, processes, and functions



A Comprehensive View of High-Availability Data Center Networking

- Assess the economic impact of incidents and disruptions that result from denial of access to systems, services, and other facilities
- Assess the “pain threshold” (i.e., business tolerance for loss), including the length of time business units, customers, and partners can survive without access to critical business applications and facilities — considering financial, legal, regulatory, customer attrition, and damage to public image
- Assess the loss associated with permanent loss or corruption of data to identify how much information, if any, an organization can risk losing

The business impact analysis report and findings should identify critical service functions and time frames that must be recovered after interruption. The BIA report presents the business rationale for identifying business recovery needs; the IT systems, networks, and resources needed to support the objectives; and other IT and business infrastructure components required to support the critical services provided by the business.

By correctly identifying key business functions and efficiently managing the change process, organizations will be in a better position to exploit the cost-versus-risk equation. Indeed, although it is possible to protect most functions via various methods (e.g., SANs, third-party services, outsourcing,

synchronous/asynchronous data transfers), not all functions are worth the cost. By identifying the necessary recovery time objectives (RTOs) and recovery point objectives (RPOs) for each function, organizations can efficiently allocate resources. An example of this could be a billing application. Although critical to the business (input of revenues), the billing process in many organizations has a fairly lengthy RTO (>72 hours). After all, if a client receives a bill on Thursday of one week versus Monday, does it really make any difference to the annual revenue stream? For this particular example, the cost to have an RTO measured in hours does not make any business sense, unless interdependencies present within the system cause another crucial business functions to fail. Ultimately, the owner of the business function must decide on the level and the type of recoverability that is appropriate for that function.

Recovery Time Objective (RTO)

RTO represents the maximum elapsed time to complete an application (and associated business process) and ensure that technology components recover and are functional to the extent that transactions, business functions, etc. can be resumed. RTO does not mean “100% recovered,” however; it usually indicates a degraded processing mode (e.g., less capacity, less performance).

Recovery Point Objective (RPO)

RPO is the point in time to which application data must be recovered to resume business transactions. It defines the point in a data stream to which you need to recover information — or, more simply put, how much data you can afford to lose.

When conducting such a business impact analysis to determine the amount of recoverability required for specific applications and business functions, the data center team must consider factual, user-specific performance history and include assumptions such as peak periods of performance, time required for users to regain productivity after an interruption, fully loaded employee burden rates (i.e., full-time equivalent cost), customer responses to incidents, stakeholder resistance to change, breakeven date estimates for investments, and IT asset life spans. Accurately and consistently calculating the value of potential business disruption involves staffing of new roles for performance analysis. In addition, permanent preventive strategies must be planned, and user impact and consumer/customer response to performance changes must be tracked and reported.

Securing Business Buy-In for Availability Planning

Management buy-in is a key factor in the success or failure of any business continuity and high-availability data center planning. Without management buy-in, these plans are doomed from the outset and eventually end up as untested, inefficient processes that are not updated. The importance of high availability needs to be acknowledged and communicated across all business functions and made an enterprisewide priority.

Without adequate management buy-in, data center architects are not sufficiently empowered to successfully implement a recovery strategy. Lack of support translates into poorly conceived strategies, sloppy project efforts, and lack of employee buy-in. It can also result in a lack of recovery priorities — a dangerous situation in terms of execution in the event of a disaster. At a tactical level, lack of authority will result in poor attendance at workshops, constant rescheduling of meetings, and other project delays. Until the IT organization uses its executive team to establish and communicate continuity and information availability priorities, it remains vulnerable.

Traditionally, one of the primary responsibilities of a corporate executive was to minimize cost, which meant business continuity and data center availability initiatives were considered “luxuries.” With a change in corporate governance and responsibility (due to various accounting

The Availability Numbers Game

With availability moving to a larger 24x365 base and planned downtime eliminated from most operational availability equations, traditional percentages-based availability numbers provide substantial availability “wiggle room.” For example, a 99.5% available end-user application and network scenario translates into about 50 minutes a week of service unavailability. Even pushing this up to 99.8% (an exceptionally strong end-to-end application percentage), it still provides operations groups with an average 20 minutes per week for outages. In an extreme but possible case, a once-a-month 80-minute outage could be acceptable operational service. In many cases, an outage of this duration could prove to be disastrous, but based on averages, operations groups have met their goals.

scandals, etc.), combined with the threat of terrorism, executives now have a responsibility to protect their business. This means high-availability initiatives have a much higher priority and are able to obtain the necessary budget to become successful. As always, organizations need to weigh the cost of availability against business risk and not spend excessively without consideration of risks (perceived and real).

The High-Availability Framework

Value-Based Metrics

Historically, the industry has been quick to focus on availability numbers as a means of benchmarking uptime. IT vendors and enterprises have evaluated availability based on the concept of five nines (99.999%) as a meaningful way of defining customer availability. Although more availability is a good thing, many users continue to be unimpressed by “nines” rhetoric and how it affects their business. Although the five-nines goal generally relates to platform hardware with meager amounts of system software, users at the end of complicated network and multilayer application suites are lucky to see three-nines (99.9%) availability. Although it will remain important for operations groups to monitor data center availability (keeping a set of books for internal metrics such as average availability), they must begin to think more like users and provide more meaningful operational service data such as mean-time statistics.

Availability averages such as average airplane stress have very little real value to customers. For example, operations groups can “save up” downtime, and the downtime reserve could be as high as almost nine hours for a three-nines offering over a 12-month period. Simply put, although operations could meet availability service-level agreements (SLAs) with just one nine-hour data center outage, would users or the business tolerate such an outage for their most business-critical applications? Although this is an extreme example, it highlights the fact that traditional availability thinking has limited value — especially in an e-everything IT world.

More meaningful parameters that have been used in many other industries to describe availability performance are mean time to failure (MTTF) and mean time to repair (MTTR). Both MTTF and MTTR move away from a pure “nines” discussion of availability and align more with accepted RPO and RTO metrics for determining availability requirements. We believe successful operations groups will move to these approaches during the next three to five years because smarter customers are painfully aware of the “availability average” hoax and are already asking for more meaningful availability criteria (a value solution versus an operational internal metric). With product-focused mean-time availability



A Comprehensive View of High-Availability Data Center Networking

scenarios, customers can more closely pay only for the level of availability that the market demands, not what operations on average delivers. More important, moving to mean-time availability will force operations to rethink recovery and production transition procedures, driving operations to a more valuable service model and a thorough housecleaning of 20+ years of availability misconceptions. Moreover, operations can, by introducing user-focused availability metrics, play a larger part in the delivery and crafting of value-driven business solutions.

In addition to the value-based metrics of MTTF and MTTR, enterprises should strive to take an application-centric view of availability. From the initial business impact assessment, the business should have a relatively good understanding of the importance that individual applications play in optimizing business success. By defining the hierarchy of application criticality, business and IT can better align business requirements with technology initiatives. Within the context of the data center environment, it is important to align the application performance requirements with the underlying service levels delivered by the infrastructure. The same requirement is true of the data center network. Traditional application networkability principles can be used to determine the application performance requirements and how the network must be architected to optimize the delivery of different types of applications.

The business should analyze and rank applications based on their hourly cost impact in the case of a data center outage. Hourly cost should include the business risk associated with an application loss. Furthermore, to the extent possible, this cost should include soft costs such as negative impact to the corporate reputation, image, stock valuations, and increases in regulatory costs. While it may not be possible to identify exact hourly revenues per application, by ranking applications the business and IT can best align themselves to ensure that the data center and the network are architected to the highest level of availability and data protection required by the application or service with the highest outage or data loss cost. It should be noted that the reason to focus on the application-specific cost impact of outages is that cost is the best indication of the importance of the application from an availability perspective. In some cases, it may be tempting to identify the most important application as the one responsible for the greatest revenue generation. However, this will not always be the application that has the most impact on the short-term cost to the business.

For example, a pharmaceutical company relies on a multitude of business applications, including ERP, CRM, e-commerce, and data warehousing, along with many research-focused applications like core drug analysis, protein modeling, and product life-cycle management. In addition to these, other communications applications such as e-mail, IP telephony, videoconferencing, and collaboration

A Comprehensive View of High-Availability Data Center Networking

tools may also rely on the network and data center for availability. There are also basic IT infrastructure services such as directory services, DNS, DHCP, and others that provide basic services, without which users cannot gain access to the services or the applications. It is tempting to identify the research applications as those being the most critical to the business given their potential for generating the greatest long-term revenues. However, when building a high-availability data center, it is more important to focus on the applications that have the greatest short-term impact on cost. A drug analysis application may be offline for several hours, or even days, with minimal impact to the long-term revenue-generating potential of the company. In the event the application is unavailable, researchers could simply shift focus to other aspects of the drug development process. However, in the case of a pharmaceutical company's ERP application, an outage could have severe short-term cost impacts based on the inability of the business to handle core functions such as order processing and supply chain issues.

Figure 2 profiles the real hourly dollar value to the business of the outages for nine specific business applications, expressed not in IT metrics, but in terms that drive the business (i.e., not transactions lost or calls missed, but dollars lost per hour). Such metrics are critical when building and operating a highly available data center environment and must be created, measured, and tailored to the specific service provided.

Figure 2 — Industry Outage Impact

Type of Business	Average Hourly Impact
Retail Brokerage	\$6,450,000
Credit Card Sales Authorization	\$2,600,000
Home Shopping Channel	\$113,750
Catalog Sales Centers	\$90,000
Airline Reservation Centers	\$89,500
Cellular Service Activation	\$41,000
Package Shipping Services	\$28,250
Online Network Connect Fees	\$22,250
ATM Service Fees	\$14,500

Source: META Group

A Technology View of Business Continuity Levels

Once the business has identified the most critical applications, it is up to the data center architects to match the level of availability required by the application to the technology. IT should provide a categorization framework for the evaluation by establishing base-level availability services that match most business requirements. Complexity should be minimized, and a few key applications should be chosen and scrutinized under the categorization framework. These applications should be looked at in relation to current disaster recovery (DR) capabilities to give organizations a baseline. It is important for IT operations to set accurate cost expectations for each category, and not forget to ensure any legal/governmental requirements are met in relation to the LOB applications.

META Group has defined the following four categories (see Figure 3) within a business continuity (BC)/DR framework:

- ***Platinum service:*** This level of service gives organizations continuous availability, with an RTO and RPO of zero. It will require synchronous replication to ensure zero data loss in case of disaster. Typically, to meet this service level, organizations would need two hot data centers fewer than 50 miles apart (to enable synchronous data replication without seriously impacting application performance). Both data centers should be active and configured to handle approximately 70% of the total service-level agreement (SLA). Due to availability requirements, the data centers will require proprietary, fault-tolerant hardware as well as a duplexed processing environment, and will include such items as transaction routers. A major expense is that a well-trained staff is required in both data centers. Platinum service should be limited to high-revenue-impact applications (typically large financial transactions) and is difficult to justify from a business/cost perspective. One argument in favor of platinum service is that, once the infrastructure is in place, other applications can “piggyback” for free and gain continuous availability for minimal additional cost. Although this may be true for some applications, most legacy data center applications have their own TP monitors, procedures for synchronization with the underlying database, etc. Thus, even with a full duplicate data center, a restoration may not happen instantaneously and will require highly trained personnel who are familiar with the unique requirements of each application. Typical costs of a platinum service offering are 6x-8x that of the bronze level.
- ***Gold service:*** This level of service will be based on storage controller-based data replication, with an RTO of eight hours and RPO of less than 60 minutes. Costs for gold service infrastructure are 4x-5x that of the bronze



A Comprehensive View of High-Availability Data Center Networking

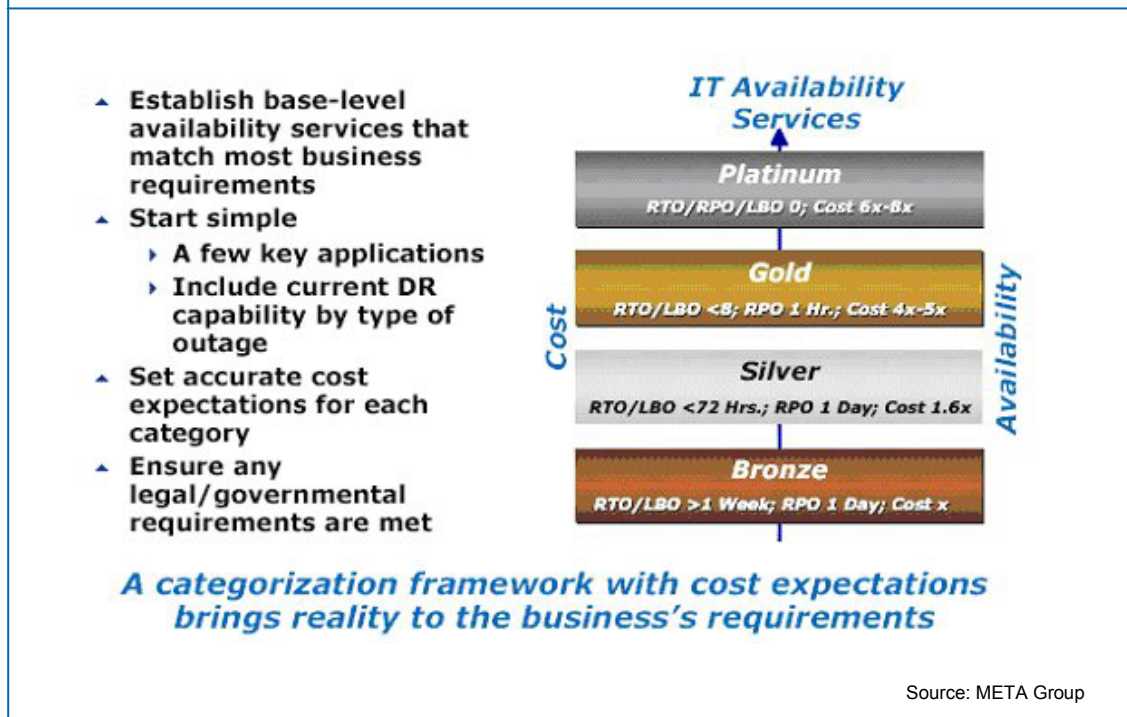
level. A second data center is still required, and it is necessary to have a complete copy of all data and applications. The data centers should be configured to handle less than 50% of the total SLA. Gold-level services are adequate for most high-availability requirements.

- **Silver service:** This category utilizes tape-based backup, with full backups on a weekly basis and daily incrementals. The RTO is less than 72 hours, and the RPO is one day (the last incremental backup), with costs 1.6x that of the bronze level. The data available in a DR scenario is only as good as the last available incremental. Data and server environments are rebuilt from tape, which can require days for complete environments. IT operations should have robust operational processes in place to expedite DR. The silver service model reduces complexity, and as a result, 70% of users with DR plans utilize silver-based solutions, most often with a third-party DR service provider.
- **Bronze service:** This category defines a best-effort disaster recovery with no third-party DR service. Typically, weekly backups are kept off-site, with incremental backups kept locally and off-site. Organizations utilizing a bronze service model are betting on a low probability of data center loss. The RTO is more than one week, with an RPO of one day (the last incremental backup).

Although the data center is critical, it is the business that keeps an organization afloat. The appropriate resources — employees, budget, and management support — must be allocated to cover all aspects of the strategy. Ideally, the BC should be developed for both data center and business units concurrently, or in less ideal circumstances, the data center plan should be followed with its business-focused counterpart.

A Comprehensive View of High-Availability Data Center Networking

Figure 3 — Establishing a Categorization Framework



Defining Network-Centric Availability Metrics

In the building of a state-of-the-art data center, it is clear that data center and network managers must work closely to ensure optimal integration of storage, servers, and the network. Traditional high-availability recovery metrics have largely focused at the macro level of data center availability and generally do not provide the network manager with a clear set of metrics by which to architect the network. Network managers should attempt to follow the same metrics used by data center managers, but should identify specific network-centric methodologies to ensure the correct mapping between overall availability goals and network availability is achieved. There are five basic metrics on which data center and network managers should agree, based on the different parts of the data center network.

Recovery Access Objective (RAO)

RAO is a subcomponent of RTO that identifies the point in time at which the users that were connected to applications and services running on one data center have access to the same applications and services running at an alternate data center. The RAO goal will typically be lower than the RTO for any specific application, so that a recovered application is not waiting on network access in order to resume providing services to users.

Data Center Network Recovery Time

The data center network is the backbone of the data center and is responsible for the interconnectivity of server resources to ensure optimal application delivery and availability. The main characteristics of this network include service virtualization, security, connectivity, performance, and resiliency. The data center network

recovery time (normally measured in seconds) is the time it takes for the network within the data center to recover in the case of a single failure in a device subcomponent, complete device, link, or server NIC. With the trend being to consolidate applications onto a consolidated data center network, the desired recovery time would be driven by the applications with the most stringent requirements. This typically results in the need for a fully redundant network infrastructure with no single point of failure, a homogeneous and hierarchical network design, and implementation of fast failover technologies.

For the ultimate level of availability, a data center could deploy completely parallel internal networks with dual homing into each server resource. While this may be a possibility for some very extreme cases, most data center managers will not find this cost-effective or advantageous from a management and control perspective. More realistic for most data centers would be redundant network components (e.g., routing switches, firewalls, load balancers), each with redundant subcomponents, that operate over the single network. Transactions should be failed over automatically and transparently, without losing connection, using stateful failover. It is important to note the recovery time required relates not only to highly scalable L2-L3 functionality, but also to the higher-level intelligent network services that are prevalent in the data center, such as L4-L7 and firewalling. One step lower in availability would be a single network and single components, but with each component having internal redundancy such as dual power supplies, dual management modules, and dual fans. In this case, there would be no alternative components in hot standby, and the recovery time would be determined purely by how long it took the individual component to fail over.

Storage Network Recovery Time

The growing adoption of wide-scale storage pooling is driving the creation of highly scalable storage-area networks (SANs). The criteria have been traditionally more stringent than the requirements for the IP network due to the sensitivity associated with losing any business-critical data, whereas packets that are lost rarely re-emerge. ITOs are increasingly realizing the importance of storage, particularly among midsize companies that have historically not had requirements for enterprise storage capabilities. Through 2005/06, most organizations will have organized around and created a storage infrastructure and operations team (the data and media center of excellence). Consequently, it is imperative that ITOs begin to view storage with regard to the strategic importance it will have — even in difficult economic environments that often (and, many times, incorrectly) result in solely tactical decisions. ITOs should undertake a strategic evaluation, examining how daily functions can be leveraged across (and automated by) multivendor capabilities and how tasks currently consuming significant resources can be dramatically reduced (e.g., recovery, provisioning, procurement).

Application/DBMS recoverability is one of the most compelling reasons for SAN participation. Based on the mean time to recovery (the time it takes the application to begin accepting transactions again — not simply restoring information from tape to server) discussed with the business, numerous technology options can assist in meeting those requirements. Snapshot or volume-based replication (either local or remote) can dramatically reduce the recovery times to minutes, preventing the need for a tape restore (but not completely eliminating the need for tape backup/recovery).

Although restoration is a significant and necessary capability, the “state” of the information/data can be much more important than the existence of the information. Recovering data in an inconsistent state — or data from one database that reflects a point in time that is significantly different from another reliant database — is often useless. Consistently managing replication and recovery options (possibly from multiple vendors) will prove to be operationally beneficial, as will rationalizing backup/recovery software and hardware across as many platforms as possible (often excluding the mainframe). Consistently managing various storage vendors’ snapshot/volume replication can provide further operational savings, striving for policy-based automation. Moreover, all applications are not created equal, and storage replication software is no different. Integration with operating systems, clustering software, applications, and databases, as well as quiescent time (time required for a consistent view of the application), will often differentiate offerings, given that replication software has become much more prominent during the past 36 months.

Moreover, as storage technologies continue to mature and more servers participate in networked storage (SAN and NAS), ITOs will be forced into measuring delivered capabilities as storage becomes a service that is delivered to the business in “business speak” (e.g., performance, availability, flexibility to change), and as associated cost tradeoffs based on their selection of tiered services are clearly understood.

Multiple Data Center Interconnect Network Recovery Time

META Group estimates that, by 2006, 40% of Global 2000 enterprises will support a dual data center approach for continuous availability requirements. The redundancy of data centers allows for a “live site” approach, with the alternate site having some backup processing, facilities for storage (usually with some mirroring or data replication) and communication, and some personnel locally available. The interconnect between multiple data centers must also have a recovery time

objective associated with it. The impact of the loss of an interconnected network depends very much on the types of business continuance applications that need to be supported. For example, if synchronous remote mirroring is deployed, then a loss of connectivity will directly and immediately impact the user experience.

Access Network Recovery Time

This relates to how quickly a user can regain application access in the case of a disruption to a network external to the data center. While not a part of the core data center network, the connectivity methods for ensuring user access to corporate information can be just as critical. When assessing the availability of the information contained within the data center, IT organizations must understand who is ultimately accessing that information and the means with which those individuals gain access. Two types of users can be identified, internal company employees and external third parties. In the case of the former, these users may access corporate resources over a private, enterprise-owned network infrastructure such as a local-area network or corporate-owned/managed remote access dial service. Alternatively, these internal employees may also access the information via a public network such as fixed broadband to the home, public Wi-Fi hot-spot service, or aggregated Internet dial services. From the business impact analysis, the IT organization should have some idea of the requirements for availability to certain classes of users and should then be able to provide specific network access options to best meet those requirements. In cases where the business requires a higher level of availability, a reliance on private, enterprise-owned/managed networks would be the preference.

The business must also assess the impact of loss of information delivery or service to third-party users such as business partners, suppliers, and consumers. For example, a news and entertainment organization that distributes content to a network of independent regional publishers may place a higher degree of importance on the availability of the partner network and its data center interface than on the internal network to connect its remote employees. Thus, defining the priority with which certain user types require access to the data center will allow the business to assign specific RAO goals and the IT organization to optimize network connectivity to meet those goals.

Recovery Access Objective

This metric measures the time it takes for the network to re-establish connectivity of users, customers, and partners with the applications at the alternate site once the primary site has been disrupted. While traditional RTO metrics have focused on the time it takes for the applications to return to full availability, it has not specifically addressed the network availability. Measuring the recovery access

objective (RAO) would allow network managers to align themselves with the overall data center RTO metric, while maintaining a specific focus on how the underlying network and communications infrastructure would handle the outage and connecting to a system reconstituted elsewhere. It is clear that the RAO should be viewed as a subcomponent of RTO and must therefore meet or beat the time specifications that have been established by the business for RTO. For example, for zero RTO applications, users should be dynamically and automatically redirected to the alternate site without even feeling it.

Building a Highly Available Data Center Network

This section focuses on the components that go into designing a highly available network for the data center environment. In the interest of being concise, we will focus primarily on the inter-server network, as this is generally accepted as being the most critical component of the overall data center networking environment. However, network managers should leverage the principles outlined in this section and apply them across the breadth of their data center network. It is important to recognize when planning for high availability that the highest levels of availability are achieved through layering. It may be tempting to equate network availability with redundancy; however, redundancy is only one component of the overall availability architecture. In fact, too much redundancy will have an adverse effect on availability due to the increase in complexity and management required to support the network.

A robust availability plan will include a combination of component-level availability, networkwide recovery intelligence, system-level redundancy, interaction with endpoints (e.g., servers, storage), and operational best practices. Furthermore, organizational issues such as training and staffing will also affect the degree to which high availability for the network can be maintained over the long term. Finally, sourcing and vendor partnership is also a key component, and IT organizations should seek partners who have the expertise and long-term commitment to the market to ensure the continued development and support of hardware and software.

Network Availability Services

The network has evolved into a utility providing connectivity into a multiservice infrastructure critical for delivering value-added services to the business. Many have argued that the network has become a commodity, yet when one considers the increasing role of the network in providing services greater than pure throughput and connectivity, it becomes clear that this is far from the case. Within the context of availability, network services can be categorized as follows:



A Comprehensive View of High-Availability Data Center Networking

- Security
- Resiliency
- Performance and optimization
- Connectivity

Each of these classifications has a role to play in the broader delivery of a highly available data center network. Furthermore, network architects and data center managers must validate existing and planned data center network build-outs against each of these parameters to ensure peak performance.

A key trend affecting the enterprise and data center network is the movement toward the virtualization of services. As the intelligence of the network continues to increase, so too does the ability to incorporate additional services directly into the thread of the network. This virtualization of services is playing an increasingly important role in the delivery of high availability. Historically, network and data center architects would rely on a multitude of devices and components to achieve a specified level of performance. Often, this heterogeneous environment of purpose-built, feature-specific products was more complex to manage and maintain due to the disparate, vendor-specific interfaces and controls.

Virtualization allows for much of the functionality previously delivered through point products to be delivered within the context of a single-network framework. As an example, data center and network architects historically have used discrete load-balancing switches as a means to more intelligently route packets across a large server farm. Through the virtualization of that capability, architects are now able to leverage a single platform that has load balancing capability integrated into the same platform used for aggregation, throughput, and connectivity.

A similar example can be made with network security. Traditionally, data traffic isolation or segmentation was the role of the standalone firewall. As networking technology advances, the ability to segment traffic through the use of virtual LANs (VLANs), access control lists (ACLs), and now embedded virtual firewall capabilities becomes reliable to a point where architects may forgo the use of point products. The advent of virtualized services also serves to enhance the management and control of application flows across the network. More specifically, as network infrastructure consolidates, network and data center managers will still require visibility into the performance of specific applications. Virtualized services such as VLANs and VSANs (virtual SANs) allow for this isolation and segmentation of traffic flows across a common underlying infrastructure in a way that improves control and management — and thus, ultimately, availability.

Connectivity services refer to features that optimize the internetworking of endpoints and servers. The most evident connectivity service is the bandwidth that is provisioned between endpoints. Most data centers currently rely on a mix of Fast Ethernet and Gigabit Ethernet. As the amount of information being passed across the data center network increases, network managers will migrate to higher bandwidth levels such as 10 Gigabit Ethernet. While bandwidth may seem like a basic service, underprovisioning can negatively impact data center performance to the detriment of availability. Other connectivity services include port aggregation (e.g., 802.3ad), DHCP, DNS, multicast protocols such as IGMP and DVMRP, and redundancy protocols such as Fast Spanning Tree and VRRP (Virtual Router Redundancy Protocol). All these services can have a beneficial impact on the level of availability and recovery time of the overall network when used correctly.

Performance services center around the optimization of traffic flows and network configuration. Within the context of the data center environment, these services would include quality-of-service protocols and policies, load balancing, caching, Web acceleration, encryption offload, and route optimization. These services have long been handled through point products (e.g., load balancing, caching) and are now being increasingly incorporated into the thread of the network. The use of these tools, whether in a virtualized or point product fashion, is essential in driving higher levels of network availability.

Finally, resiliency refers not only to the availability of the network, but also to the quality and performance of its individual components. When building a data center network, network architects should evaluate components based on their ability to operate with minimal interruption or failure. Metrics such as MTBF (mean time between failure) can provide insight into the durability of a product. Furthermore, architects should evaluate the reputation and ability of the manufacturer to engineer and produce high-quality products. IT organizations must gain greater insight into vendor testing and certification methods for ensuring hardware and software stability. Many leading vendors offer forums for customers to exchange information on product specifics. Those vendors with a clear feedback and response channel will be in a better position to identify and resolve issues in a timely fashion before they have the ability to seriously impact the customer network.

Recovering Network Services

Achieving the predetermined network recovery objectives requires a plan for network recovery and thus network services recovery. As we have outlined in the previous section, the intelligence of the network has increased to a point where services that were once separate from the network are now being delivered virtually within the network. However, one potential downside of this approach is

the interdependency between network and service. In other words, if the network were to fail, so too would the services. Conversely, the optimal performance of the network is now directly tied to the availability of the service. For this reason, it is critical that architects evaluate services based not only on the need to guarantee optimal application performance and availability, but also on their own inherent ability to recover after an attack or outage.

As a design principle, network architects should evaluate the ability of specific services to recover after an outage. Many network components have the ability to statefully fail over to components in hot standby — thus guaranteeing the continuation of the service. The virtualization of services into the network requires architects to plan for redundancy and failover not only at the component and network layers, but also at the services layer.

High-Availability Implications for Network Operations

Just as important in guaranteeing high availability as any technology or design principle is how the network will be managed and maintained. As the expectations placed on the network increase, so too does the care with which network and data center operations must monitor and manage the day-to-day performance of the network. High-availability requirements will dictate the level of operational support a network will require with mission-critical data centers demanding substantially more human capital resources than other less critical networks. Unless the business is prepared to allocate the human capital resources required to optimize the maintenance of the data center, it risks experiencing a decrease in the level of performance and availability of corporate services and information.

In addition to investment in human capital, the business must ensure that adequate operational processes have been established. Change management is of the utmost importance in the data center, because the slightest change in software or network configuration could have disastrous effects on performance and availability. Many leading hardware manufacturers help facilitate change management through the development of features such as automatic software rollback or reconfiguration. However, having a clear process for managing configuration and change is a must-have for any data center manager. To mitigate the risk of performance degradation and service impact, the business must also establish performance management and monitoring procedures. Through such planning, data center managers can make certain that potential problems are identified and rectified before snowballing into service degradation or, worse, loss of service.

Closing the Loop

Throughout this document, we have encouraged IT organizations to set specific objectives, including RTO, RPO, and SLAs. Indeed, setting specific, measurable goals is critical to creating a successful high-availability plan. Too often, however, ITOs limit success criteria to these objective metrics. In addition to numeric goals, IT organizations must continue to engage with the business units to define success around subjective metrics as well. For example, as the business and organization change, the high-availability infrastructure must respond to these changes as close to real time as possible. Thus, we recommend that organizations implement a business continuity group composed of business and IT stakeholders to periodically review the high-availability plan as compared to business changes.

The bottom line for ongoing system monitoring and management is year-over-year improvement. With improving technology and the increasing speed of business both being environmental “givens,” one can be certain that today’s best practices will be insufficient in three to five years. Thus, IT organizations must approach planning, implementing, measuring, and improving their high-availability systems as a continuous process.

Carl Greiner is a senior vice president and Chris Kozup is a program director with Infrastructure Strategies, a META Group advisory service. For additional information on this topic or other META Group offerings, contact info@metagroup.com.



About META Group

Return On IntelligenceSM

META Group is a leading provider of information technology research, advisory services, and strategic consulting. Delivering objective and actionable guidance, META Group's experienced analysts and consultants are trusted advisors to IT and business executives around the world. Our unique collaborative models and dedicated customer service help clients be more efficient, effective, and timely in their use of IT to achieve their business goals. Visit metagroup.com for more details on our high-value approach.

