

Business Continuity and Disaster Recovery - Strategic Imperative for the Enterprise Management

Jolanta H. Lapkiewicz

Board of Education NJ State, New Jersey, USA

e-mail: j0036823@netzero.net

Kazimierz Frączkowski

Wydziałowy Zakład Informatyki, Wydział Informatyki i Zarządzania Politechniki Wrocławskiej

e-mail: fraczkowski@ci.pwr.wroc.pl

Abstract

This paper will discuss the differences between disaster recovery and business continuity planning (BCP). We will look at the principal strategies, technologies and best practices that support an effective business continuity plan. We will further consider the business continuity service provider market.

Few businesses can make an easy calculation of how they would perform, and how much money they would lose, without their computers. The best evidence comes from both ends of the scale - those who have survived a disaster and those for whom the loss of the computer is the death of the company. With an increased overall requirement for 24x7 access to applications and data (e.g., due to Internet-based applications, customer service and globalization), many enterprises are planning for both high availability and business continuity.

By clearly defining the key business processes that must be re-established in the event of a disaster, enterprises can prioritize the most-critical resources to be made available quickly. The overall goals of a business continuity plan are to maintain confidence in the business by all external constituents including customers, trading partners and regulatory agencies.

Building Continuous Availability Into E-Applications Enterprises incur significant costs when critical Web-based OLTP applications are down. We provide advice for systematically building Web application infrastructures that boost application availability. Achieving continuous 24x7 application availability requires a multipronged strategy that addresses and mitigates risks of unplanned failures and planned maintenance/upgrade. Why is it so hard to achieve? Because continuous availability is highly systematic - it cannot be bought off-the-shelf, and it requires substantial levels of cross-organizational people/process planning, discipline and control.

Most IT projects, however, are not systematic, but opportunistic in nature, where timeliness takes precedence over planning, discipline and control. Moreover, building continuous availability is expensive, costing approximately 3.5 times as much as a standard application. Consequently, despite the Internet driving significantly increased desire for continuous availability, through 2005, fewer than 20 percent of mission-critical Web-based applications will achieve it (0.8 probability). Around 40 percent will achieve high availability (99 percent to 99.5 percent or 43 to 87 hours per year of unplanned downtime and four to 16 hours per month of planned downtime) at lower cost (around 2.5 times that of a standard application) (0.8 probability). The balance will not meet quality-of-service objectives (0.7 probability).

1. Definitions

A *highly available* application provides end-user access to applications and data during a high percentage (e.g., greater than 99 percent) of scheduled uptime, despite unscheduled incidents. High availability implies the ability to minimize unscheduled (also called unplanned) outages. Implied in high availability is application and data integrity as well as acceptable (as defined by the user) application performance.

A *continuously operable* application provides end-user access to applications and data during expanded times — typically 24 hours a day, seven days a week (24x7). It implies the ability to minimize scheduled (also called planned) outages (e.g., outages scheduled for maintenance or upgrade of hardware, networks, facilities or software).

A *continuously available* application combines high availability and continuous operations to avoid or minimize both unplanned and planned downtime. It implies that the application is scheduled to be available 24x7 with a minimum of 99 percent end-to-end application availability (or less than 88 hours of unplanned downtime per year).

A key differentiator in enterprises achieving continuous or high availability is an investment in internal architecture standards for design, development, testing and operations. Once standards are implemented, it costs little more money *or time* to deliver a continuously or highly available application. Disregarding availability requirements during design frequently results in re-architecting the application at a later date to meet growing availability requirements. As most enterprises do not adequately “design in” application availability, by 2005, more than 50 percent of new applications built and deployed before 2002 will be re-architected or rewritten to meet availability and performance requirements (0.7 probability). Designing for continuous or high availability minimizes unplanned and planned outages. Furthermore, it minimizes user impact for any outage that does occur. While unexpected component outages cannot be completely eliminated, masking outages from users creates the *perception* of uninterrupted availability. This approach is the underpinning of any availability strategy. Further, inevitable planned downtime can be mitigated in a number of ways. One method sets user uptime expectations (e.g., four hours per week of scheduled downtime). Another method enables 24x7 site accessibility, but not to all functions, lessening the impact of planned outages. Yet another method programmatically hides the effects of planned downtime (e.g., transactions may be queued for later processing or processing may be temporarily re-directed to a replica of the primary system). Although there is no single infrastructure on which mission-critical transactional Web applications rely, many of the business-to-business (B2B) and business-to-consumer (B2C) application architectures and underlying technology infrastructures have similar underpinnings. Summarizing the experience of many projects where continuous or high availability was systematically designed, we offer the architecture as a reflection of the current industry best practices.

2. Application Architecture

Advanced transactional Web sites are designed to have no single point of failure (SPOF), with a fully redundant, *n*-tier application architecture across two or more geographical locations. They implement multiple Web servers on the front end, multiple application servers in the middle tier and highly available databases on the back end. In addition, most have links to legacy, enterprise and/or external systems.

By partitioning Web site functions into components that can reside separately on different systems, enterprises can achieve greater availability (through no-SPOF architectures), scalability (by scaling components horizontally through load balancing and vertically through system upgrades) and flexibility (since underutilized systems can be redeployed where they are needed most). At each level of the n -tier architecture, the issues of availability are solved differently, as we describe below. While enterprises would like to rely on the underlying technology infrastructure exclusively for the quality of service of their applications, the truth is that application availability is also dependent on application design. For example, component architectures make it easier to upgrade an application “in flight,” thereby reducing planned downtime. Stateless application components allow for a more flexible and powerful use of application server load balancing, providing for greater levels of availability as well as user access persistency, despite component outages. Moreover, asynchronous connections between components and applications are much more tolerant of failures of hardware and software than are synchronous connections. These choices, however, are made during application design and cannot be reversed once the application is developed and deployed. Thus, consideration of application availability must be present from the start of the project and not delayed until application deployment.

2.1. Tier 1: The Web Infrastructure

Web load balancers are used to ensure that Web servers are not a SPOF. The load balancer is configured with policy information regarding the Web servers to which it hands off requests, such as maximum number of connections or relative server size. The more sophisticated load balancers can make loading decisions based on differentiation of users. Additional servers can be easily added for scalability. If the balancer finds a server is unavailable, it is removed from the loading table, and subsequent requests are directed toward the remaining Web servers. Moreover, planned downtime can more easily be carried out nondisruptively simply by taking the server out of use, which automatically disables client loading. To ensure the load balancer is not itself a SPOF, enterprises should implement redundant load balancers.

2.2. Tier 2: The Application Servers

The application server platform is used to host and manage the software that implements the core business of the transaction. In some applications, this is limited to simply the role of an intermediary between the Web and the database. Typically, more-advanced applications use application servers to host such additional functions as business process management, transaction management, personalization, e-commerce services (such as shopping cart), data analysis, transformation and interpretation. The application server is used to enable immediate and uninterrupted access to application functions for each of the typically large number of concurrent and competing user transactions generated from the Web. These transactions also require security, monitoring and administration, and integrity. To meet these requirements, application servers deploy their own replication, context management, transaction management and load balancing. These are software solutions, complementary to the balancing and clustering that occurs at the network, hardware and database management system (DBMS) levels. The application servers, having control over individual component interfaces, replicate at the component level as well as applying access control and performance optimization at that level. In addition, some high-end application servers (iPlanet, PowerTier/J) allow for replication of the in-flight transaction context. This feature makes it possible, when a failure occurs in an individual application server, to have transactions automatically transferred to another application server for processing, transparent to the user. (Failures in application servers not using transaction context replication would result in transaction failure, requiring re-submittal by the user.) Some application servers are much better than others at these tasks. Many enterprises also separate their user-facing and data-facing logic and deploy multiple application server platforms. This approach results in a multiplatform software architecture at the application

server level, offering greater flexibility (e.g., supporting user devices), but adding complexity, and requiring additional availability considerations.

2.3. Tier 3: The Database Servers

It is important to ensure that the hardware on which the database resides and the database itself are designed for no SPOFs. To protect against hardware and operating-system failures, enterprises implementing databases — e.g., Oracle, SQL Server and DB2 Universal Database (UDB) — on Unix and Windows NT implement database clusters, such as Hewlett-Packard's MC/ServiceGuard, IBM's High-Availability Clustered Multiprocessing (HACMP), Microsoft's Cluster Server, Sun Microsystems' SunCluster and third-party alternatives from Legato Systems and Veritas Software. Database clusters enable fast failover to an alternative system when the primary system fails. Recovery time averages 10 minutes (which is not transparent to users) and is dependent on the amount of in-flight data recovery. Unix and NT database clusters also reduce planned downtime by enabling a manual switchover from primary to secondary for planned maintenance/upgrades. Enterprises' running databases on S/390, or fault-tolerant systems from Stratus Computer and Compaq Computer (Tandem), gain stronger hardware and operating-system fault avoidance mechanisms and the benefit of application transparency when a failure does occur. In addition, enterprises implementing S/390 Parallel Sysplex, or Stratus or Compaq (Tandem) fault-tolerant system clusters, reduce planned downtime by transparently enabling workload movement to other cluster nodes. Because Unix and NT database server failures have significant end-user impact, many enterprises implementing e-commerce applications based on Oracle implement Oracle Parallel Server (OPS). OPS enables concurrent database access from multiple cluster nodes, resulting in failures appearing as a transaction failure (of about 30 seconds) but not necessarily downtime, since the user can re-submit the transaction, which is automatically applied to another system in the cluster. Successful use of OPS, however, requires proper initial design to ensure database scalability. Enterprises running databases on S/390, or Compaq (Tandem) or Stratus fault-tolerant systems, do not incur long failover times (failover is transparent to users), but they often implement clustering (e.g., Parallel Sysplex on S/390) to reduce planned downtime; here, applications can move to another system complex (transparently to users) so that the primary system complex can be maintained or upgraded. Should the database itself become corrupt, then database clusters will not help (because there is only one database). Best practices replicate the database so that it can be recovered to a point in time with integrity (and rolled forward to the last transaction prior to the corruption if possible). There are many ways of handling the database replication, and no consistent method is used across e-commerce sites. Some implement database replication technologies (typically log-based approaches that are set to be two to four hours behind the primary), whereas others utilize hardware- or software-based snapshot technology and take many point-in-time replicas of the database throughout the day. Both methods enable reasonably fast recovery to a consistent database state in the event that a logical corruption error cannot quickly be corrected by some other means. Most transactional Web sites integrate with internal and/or external applications. Sometimes this link is direct and synchronous; at other times, the connection is asynchronous, via a messaging interface.

Asynchronous connections generally lower the overall risk of downtime because of their loosely coupled nature, but only if the application is effectively designed (so that the application continues despite component downtime). From an availability perspective, synchronous connections are often the critical bottlenecks in the entire topology and also may be the most difficult and expensive to cluster or replicate.

To facilitate the integration of multiple participating applications, enterprises often deploy another separate infrastructure platform — an integration broker or suite. Integration brokers are not application servers. Their role is to be placed between applications and facilitate the transformation and flow management required between external and internal applications and the new business process. At present, most integration brokers operate asynchronously and can tolerate some downtime. Modern integration brokers (e.g., STC's eGate or Crossworlds) can be clustered, thus elimi-

nating the integration broker as a SPOF. The newer and fast-growing application integration patterns of composite applications, however, are synchronous and rely generally on the features of application servers, including the application server replication and fault tolerance. The tightly coupled synchronous links to internal and external applications may put the availability of the overall application at substantial risk. As a result, we recommend that applications requiring continuous or high availability rely on asynchronous integration middleware in addition to well-designed application code.

2.4. Multisite Architecture

Many e-business sites have disaster recovery built into the application and technology infrastructure by operating over multiple physical sites. Some enterprises operate on both sites actively, meaning that client traffic is routed (load balanced) to any of the physical sites for complete processing. Complete transactional redundancy across physical sites requires significant forethought on transaction and data replication as well as synchronization. Depending on the application and its complexity, the design may call for load balancing of Web server connections across the physical sites, for load balancing and replicating of the business logic (and transactions) across independent locations, and for replicating the DBMS to a backup location for disaster recovery, or for a combination of some or all of the above. With replication of the entire application environment, in the event of a catastrophic failure at the primary location, all traffic is routed to the alternative location. The cost of such a replicated topology, however, can be very high.

Bottom Line: Replication of databases, hardware servers, Web servers, application servers and integration brokers/suites helps increase availability of the application services. The best results, however, are achieved when, in addition to the reliance on the system's infrastructure, the design of the application itself incorporates considerations for continuous availability. Users looking to achieve continuous availability for their Web applications should not rely on any one tool but should include the availability considerations systematically at every step of their application projects.

3. Key Trends

- * How will enterprises best prepare for impending disasters?
- * What tools, technologies and products will enterprises employ to protect critical applications and business processes?
- * How will the market for business continuity services evolve?

Few business can make an easy calculation of how they would perform, and how much money they would lose, without their computers. The best evidence comes from both ends of the scale — those who have survived a disaster and those for whom the loss of the computer is the death of the company. With an increased overall requirement for 24x7 access to applications and data (e.g., due to Internet-based applications, customer service and globalization), many enterprises are planning for both high availability and business continuity. By clearly defining the key business processes that must be re-established in the event of a disaster, enterprises can prioritize the most-critical resources to be made available quickly. The overall goals of a business continuity plan are to maintain confidence in the business by all external constituents including customers, trading partners and regulatory agencies, and to resume business as usual for as many employees as possible as soon as possible. BCP must be a top-level concern for enterprises, considering the potentially devastating financial and organizational impact of a disaster. The main reason behind the shift from disaster recovery to business continuity is the realization that most processes in the enterprise are now totally dependent on IT services for some stages, if not all, of their life cycles, and that the

business is the owner of the system. Along with dispersed ownership, IT becomes harder to manage and control as it is distributed across the enterprise.

Effective BCP is vital to maintaining confidence in the business, making a quick recovery and minimizing losses. It is important to understand that complete BCP includes all four components listed above. In addition, the potential scope of a failure must be determined from an organizational perspective and all interdependencies appropriately planned for. At the time of the business interruption, these plans overlap, e.g., if the data center and the business areas are co-located, then a failure at the one location will result in the invocation of the disaster recovery and business continuity plans; also, business resumption plans go into effect for the time frame between the event occurring and full recovery at the alternate site.

BC Components

	Disaster Recovery	Business Recovery	Business Resumption	Contingency Planning
Objective	Mission-critical applications	Mission-critical business processing	Business process workarounds	External event
Focus	Site or component outage (external)	Site outage (external)	Application outage (internal)	External behavior forcing change to internal
Deliverable	Disaster recovery plan	Business recovery plan	Alternate processing plan	Business contingency plan
Sample Event(s)	Fire at the data center; critical server failure	Electrical outage in the building	Credit authorization system down	Main supplier cannot ship due to its own problem
Sample Solution	Recovery site in a different location	Recovery site in a different power grid	Manual procedure	25% backup of vital products; Backup supplier

Through 2003, only 25 percent of large enterprises will leverage their year 2000 continuity planning efforts and improve the overall quality of business continuity programs and plans (0.8 probability). By 2003, there will be a polarization in recovery windows, with critical business processes and application systems requiring recovery in under 24 hours and noncritical ones requiring recovery in four days or more (0.8 probability).

3.1. Key Issue: How will enterprises best prepare for impending disasters?

The foundation of BCP success is senior management sponsorship and participation. The business impact analysis (BIA) is the most critical step, as it identifies what and how much the enterprise has at risk, as well as which business processes are most critical, thereby prioritizing risk management and recovery investments.

Direct financial impact will arise via lost sales, increased costs of working, material losses or other loss exposure. Indirect financial impacts, e.g. reduction in future earnings, may arise in the longer term via loss of customer confidence or competitive advantage or damage to the brand value. In a 1997 study, Knight and Pretty observed an immediate decline in share value equivalent to the material loss. They argued that share price recovery is, to a large part, dependent on investor confidence in the management to effect a recovery. Skipping the BIA because it involves senior management participation often results at the end of the planning phase in only the costs of risk

mitigation being presented, with no convincing indication of what is being protected. Risk analysis identifies the vulnerability of the enterprise to different categories of risk and its probability. The recovery strategy outlines in broad terms the approaches to risk mitigation, incident management and recovery from the incident. Detailed plans and procedures are then created by those responsible for the daily operation of the processes. The recovery process must be tested. Last, and NOT least, a process is established to keep the plan up to date. *Action Item: Use a formal BIA and risk analysis as the basis for building business continuity plans.*

Often, disaster recovery (DR)/BCP planners are "lone rangers" within their enterprises, espousing a need, but having little authority and budget control for implementation. With the distributed nature of today's business processes, responsibility for DR/BCP must reside within business units (with policy set centrally). Senior management must understand how important its level of commitment is to the success of the program. For example, in the event of a disaster, funding must be available in a timely manner. If the person who would normally sign off is unavailable (e.g., the CFO), he or she must identify a successor or some other method to ensure the timely release of money. As another example, declaring a disaster may mean notifying the DR hot-site service provider and the start of expending large amounts of money. The timeliness of this decision is extremely important, as a regional disaster may mean that firms are competing for time at the local DR site. Declaring a disaster will cost money, even if the DR site is never used due to the disaster not being as severe as originally thought. Senior management needs to understand the risk to the enterprise and the need for a corporate sponsor or oversight committee. Examples such as these can be used to enlighten the executive team and shore up necessary commitment for DR/BCP. However, avoid making BCP sound overly complicated — it is NOT rocket science. Often, disaster recovery (DR)/BCP planners are "lone rangers" within their enterprises, espousing a need, but having little authority and budget control for implementation. With the distributed nature of today's business processes, responsibility for DR/BCP must reside within business units (with policy set centrally). Senior management must understand how important its level of commitment is to the success of the program. For example, in the event of a disaster, funding must be available in a timely manner. If the person who would normally sign off is unavailable (e.g., the CFO), he or she must identify a successor or some other method to ensure the timely release of money. As another example, declaring a disaster may mean notifying the DR hot-site service provider and the start of expending large amounts of money. The timeliness of this decision is extremely important, as a regional disaster may mean that firms are competing for time at the local DR site. Declaring a disaster will cost money, even if the DR site is never used due to the disaster not being as severe as originally thought. Senior management needs to understand the risk to the enterprise and the need for a corporate sponsor or oversight committee. Examples such as these can be used to enlighten the executive team and shore up necessary commitment for DR/BCP. However, avoid making BCP sound overly complicated — it is NOT rocket science. Security events (malicious hacking, disgruntled employees and viruses) and denial-of-service attacks can bring down systems, applications and expose sensitive corporate and customer data to public disclosure. The first step of prevention is thorough planning — ensuring security policies and practices are adhered to in systems, networks and applications, as well as knowing what the appropriate response will be if an incident does occur. Despite solid planning, however, most enterprises will be attacked, most likely through the exploitation of recently discovered software bugs. Should attacks occur, enterprises must be able to detect them as quickly as possible and will need to implement intrusion detection technology to assist with this process. The response to an incident should be coordinated by a specialized team of people skilled in investigating technical anomalies and capable of establishing a virtual crime scene. From an availability perspective, the key is in planning to avoid potential incidents as much as possible, and in detection, containment and incidence response to recover as quickly as possible after an event occurs.

Action Item: Proactively implement preventive security breach measures, including establishing a computer incident response team (CIRT).

With the migration of mission-critical systems to distributed and Internet-based computing systems comes the need to decentralize responsibility for DR/BCP planning. Enterprises must ensure that business units accept responsibility for the availability (or lack thereof) of mission-critical systems or information in the event of disaster.

To assist the business units and ensure that company interests are consistently protected, many enterprises appoint a business continuity manager (BCM) to coordinate the BCP activities of the enterprise. The BCM typically reports to the chief financial or operations officer. The BCM has oversight responsibility for setting BCP policy, reviewing individual business-unit progress for BCP development and reporting BCP status to senior management or Board of Directors. Critical success factors for IT include: the ability to restore systems under different recovery situations (e.g., different scenarios, hot sites and team members), the ability to identify cost-effective alternatives and the ability to meet agreed-to deliverables. Enterprises must ensure that business continuity plans cover systems managed by third parties. Critical success factors for business units include: communications plan to customers, external service providers and regulators; documented business recovery and resumption procedures; and restoration procedures.

To ensure a credible degree of recovery preparedness, every plan must be regularly tested. Testing familiarizes all BC team members with the experience of a sudden and unexpected interruption in business processing and exposes potential problems and unforeseen situations. Further, the variables that have been captured and rehearsed as part of a successful plan — equipment, tasks, applications and personnel — are constantly changing. Unless the test is flawless, there will be lessons and modifications to be incorporated into the plan and reflected in the next test, at which point other modifications will come to light. This cycle is the key to recovery preparedness and maximizing the chances of successfully surviving a disaster.

Reporting BCP status and progress is a key element of getting the plans created, but more important, of keeping them up to date. By using the normal line management to ensure that functional and process management take responsibility for their own domains, each manager should sign the business continuity plan and be held accountable for its accuracy.

Action Item: When contracting for test time with service providers, enterprises should ensure sufficient test time (usually two tests annually, with the first one lasting 48 to 72 hours).

Each enterprise must establish its own level of investment in business continuity services based on the nature of its business, the risks it faces and management's attitude toward risk (risk-averse, risk-takers, risk ostriches). This explains why investments by enterprises in the same industry and the same region may vary significantly. An average figure should not be used as an indicator of what the enterprise should spend, but as a sanity check to assess its own investment. The criticality of e-business applications increases the risk of financial impact due to a poorly planned, executed or nonexistent BC plan. E-business applications are as critical, if not more critical, to the success of some businesses, and the recovery strategies that need to be implemented are in essence duplicating the business environment (e.g., load balancing or 24x7 availability). As a result, IT costs of recovery increase. We believe that e-business will drive increased spending on disaster recovery to an average of 7 percent of the data center budget in 2004 from 3 percent today (0.7 probability). Of equal importance is the perception in the marketplace. The world knows when critical Web applications are out, and the BC plan must address customer and partner confidence.

Action Item: A BIA identifies the enterprise's risk profile and a direction as to the appropriate spending levels. An average spending number is meaningless; each enterprise must determine spending to meet its recovery requirements given risk appetite.

To determine appropriate availability investments, enterprises should first understand the consequences of downtime. Understanding these effects will aid in justifying investments for day-to-day operational availability and for business continuity. One of the first steps in developing a business continuity plan is performing a BIA, where critical business processes are identified and pri-

oritized, and costs of downtime are evaluated over various time periods (e.g., one day, two days, one week or two weeks). The BIA is performed by a project team consisting of business unit, security and IS personnel. Key goals of the BIA are to: 1) agree on the cost of business downtime over varying time periods; 2) identify business process recovery time objectives; and 3) identify business process recovery point objectives. The results of the BIA feed into the business recovery planning process. In order to meet recovery time and point objectives, enterprises will evaluate many different service and technology offerings, some of which are discussed in the following presentation pages. BIA is the critical first step in the conceptual transition from recovery to continuity. BIA is designed to establish a common understanding, explicitly endorsed by executive management, of what the enterprise sees as its key processes, the priorities it assigns to the processes and the quantified impact on the enterprise of disruption to those processes. Enterprises embarking on a BIA may develop their own procedures from first principles or use the services of external consultants. Although each BIA is unique to its enterprise, those that use a tool can avoid some degree of “reinventing the wheel,” and gain efficiencies in analysis and reporting. The results of the BIA feed the detailed business continuity plan, which includes all contingency, business recovery, business resumption and disaster recovery plans for each business process. Here, tools can help build the plan and provide a central repository. The increasing amount of data continues to outpace the scalability of backup tools, making backup a “never-ending project.” In addition, the increase in application architectural complexity (components across many platforms) is driving new requirements for the ability to “snapshot” or create point-in-time copies of data that can be used for quick recovery in the event of a disaster.

Action Item: Focus on continuous improvement in backup and BCP processes. Tools can help the efficiency of these processes.

3.2. Key Issue: What tools, technologies and products will enterprises employ to protect critical applications and business processes?

Traditional business continuity plans provide 24- to 72-hour application and business process recoverability. With technology more intrinsic to business processes and costs of outages escalating, many enterprises are seeking shorter recovery times for critical applications. Use of high-availability techniques is escalating, especially for ERP and e-business applications, enabling enterprises to achieve reduced recovery time objectives (RTOs) and recovery point objectives (RPOs). With e-business, hot standby (an idle standby application environment that waits to be turned on in the event of disaster affecting the primary physical site) often is not good enough, and most design application architectures for two or more physical sites. This way, even if one physical data center experiences an outage, the others continue processing the requests. Although load balancing across two or more physical sites is becoming more common, often the transactional database is located in a single physical site, with hot standby to another site. This configuration reduces the complexity and chances for conflict resolution. For both hot standby and load-balanced sites, data is replicated between physical sites via either mirroring or shadowing. Shadowing builds a replica of databases or file systems by continuously capturing changes and applying them at the recovery site. Mirroring builds a replica of databases or file systems, by applying changes at the DR facility in lockstep with the primary site.

Action Item: Evaluate application RTOs and RPOs; they may not meet business unit requirements.

Data replication solutions are specific to a database, file system, OS or disk subsystem; thus, enterprises often must use multiple solutions to protect their critical data. The most common approach to mirroring for DR is disk-to-disk remote copy. Remote copy is relatively easy to set up, well proven and has low day-to-day administration. The cons are the costs and the potential impact on production applications (when in synchronous mode). Further, databases are not accessible at the secondary site for inquiry purposes. Database replication enables replication between sites (which

can be bi-directional). For some applications this works well, but for others, there is too great a chance for conflict resolution (updates made on both sites on the same record that must be resolved). In addition, for DR purposes, database replication is typically far too complex. Log-based DBMS replication products have less management overhead, work well for DR and enable a copy of the database in an alternative location for use for inquiry and reporting (horizontal scalability). Server-based block replication tools are just emerging on the market (Legato Replication, Veritas SRVM) and work similarly to disk-to-disk remote copy, but operate on the host. They require more management overhead than remote copy, but cost significantly less. File-based replication products such as those from NSI and Legato are popular on NT and Windows 2000 platforms.

Action Item: For RTOs under 24 hours, evaluate data replication through mirroring and shadowing.

There are many emerging technologies/service offerings designed to increase business continuity flexibility, reduce costs and reduce recovery times. IBM, Hewlett-Packard and Sun have introduced capacity-on-demand/emergency backup programs to enable an increase in CPU capacity under emergency conditions (e.g., disaster or peak volumes). Typically, increases are permanent and paid for as used; however, there is a desire to enable throttling of capacity to meet e-business workload variability. This will require significant changes in software licensing practices. Wide-area clusters are being deployed by Type A enterprises with very short RTO/RPO. These include IBM's Geographic Parallel Sysplex service offering (40 km distance limitations) and HP's Continental Clusters (unlimited distances). In addition, the combination of available fiber-optic networks and remote copy mirroring technologies has fueled new cascading service offerings. Cascading service offerings enable synchronous mirroring over a shorter distance, with asynchronous mirroring to the recovery site. Service providers unwilling or late to invest in their regional facilities will be left behind. Further, service providers will perform a greater amount of operational tasks such as tape backup/archival at regional facilities. These services will have lower margins than traditional DR hot-site subscriptions but will be required to enable "one-stop shopping" and high-availability services.

Action Item: Select a service provider with local presence and resources.

As business process dependence on technology increases, so too does the business cost of outages. E-business and ERP systems are examples of where significant outages directly affect an enterprise's operations and revenue, and quite possibly, survival. As a result, requirements for recovery time and point objectives are shortening and driving changes in the BC services industry. A greater number of applications will require recovery in less than 24 hours, and many, especially e-business applications, will be designed for processing simultaneously across multiple data centers (e.g., building disaster recovery into the architecture). Consequently, the proportion of DR services using high-availability methods (mirroring, shadowing, hot standby, load-balanced physical sites) will greatly increase in overall proportion. In fact, the primary DR/BC service providers (Comdisco, IBM, SunGard) are already gearing up to enable production operations within their facilities to capture e-business market share. At the other end of the spectrum, we believe there will be a polarization in DR windows, and most noncritical applications will not require DR windows of less than four days; quick ship programs will be the most common alternative for the DR requirements of these applications. Although some applications will continue to require a two- to three-day recovery window, these will be much reduced. As such, the profile of recovery services used will change.

3.3. Key Issue: How will the market for business continuity services evolve?

While a substantial part of the business continuity process cannot be outsourced, much of the equipment and labor required to plan for and support the recovery can be. Outsourcing the resourcing of the recovery plan can be cost-effective, as the cost of the equipment is syndicated in the United States among 100 (6 to 35 in Europe) other enterprises. It is sometimes impractical to out-

source because there are no local resources to support the recovery or the recovery window is too small (and risk too high) to consider sharing equipment. In these cases, enterprises may use dedicated equipment at their own or a service provider's location, or, for Web applications, operate in a load-balanced mode across multiple facilities to enable continuous availability (at their own or a service provider's location). In large installations, the BC vendor may not have the necessary capacity or the workload may not be able to be partitioned across multiple machines. Other enterprises choose to outsource because they do not want to capitalize the necessary equipment or they are not confident of their ability to maintain the critical capacity in the "recovery" location. Service providers can offer the resources necessary to support testing without disrupting an organization's normal operations. Its staff also has the experience of supporting live invocations, which are always times of stress for operational staff. It may also be that the organization's own staff is unavailable to run the recovery.

Action Item: When considering e-business service providers, evaluate their BC/DR experience.

The business continuity market is extremely fragmented, with many players in each market niche. Therefore, comparison of the different players is almost impossible; each one has a different area of emphasis and expertise. The range of products and services and the number of individual suppliers of varying degrees of specialization mean that the business continuity market in aggregate does not lend itself to Gartner's "Magic Quadrant" assessment. However, the players fall broadly into two categories: full-service providers and product- or service-specific suppliers. Full-service suppliers offer, either directly from their own resources or indirectly through subcontracted integration capabilities with other vendors, the spectrum of services listed in the chart above. In terms of revenue and expressed client interest, the three full-service providers (i.e., Comdisco Recovery Services, IBM Business Continuity Recovery Services and SunGard Recovery Services) are the most substantial participants in the North American market, and enterprises report that competition among them is increasingly close on grounds of cost. In addition, due to intense competition from Web-hosting service providers, Comdisco, IBM and SunGard have broadened their offerings to include full-service Web hosting, with availability and continuity services included.

Action Item: For high-end Web sites, evaluate Comdisco, IBM and SunGard for hosting and continuity services.

Business continuity service providers traditionally make their money from investing in equipment and people and then syndicating them to as many clients as possible without compromising risk or service. Enterprises must understand the subscription levels, and who and where their fellow subscribers are, to ensure they are mitigating their own risks. Beware of the cheapest supplier, which may be oversyndicating or managing risk badly. Changing service providers can be a painful process requiring modifications to the business continuity plan, more testing and familiarization with the new facilities. However, there are substantial savings to be made if competitive tendering is introduced. Make sure the components of the proposal from the vendors are understood. Costs will continue to drop over the next five years, so keep the contract to three years. The contract should include termination clauses related to changes in circumstances such as an acquisition or outsourcing of the data center, increases in costs beyond an agreed limit and test failures due to recovery facilities proving unsuitable. There should be a clear indication of the incremental costs of adding additional capacity. Some vendors insist on a declaration fee, but the value can be negotiated. Always check the right of access and what happens if the vendor has multiple invocations. For dedicated equipment that the service provider will provide uniquely for the contract, check the cost of acquiring the equipment. Unbundle contract costs so that they can be compared to other service providers' proposals.

Conclusions

Key Issues

1. E-commerce is driving a need for shorter recovery times and points.
2. E-commerce is driving greater interest in dedicated (nonshared) recovery solutions.
3. Enterprises are more organized for business continuity planning due to greater risk (i.e., e-commerce-based business processes).
 - Obtain senior management support for BCP
 - Establish organization structure, including budget management
 - Perform BIA/risk analysis
 - Understand critical business processes and risks
 - Determine business process RTO/RPO
 - Develop BC plan
 - Business resumption
 - Business recovery
 - Disaster recovery
 - Contingency planning
 - Crisis management
 - Evaluate technologies to reduce RTO/RPO (chances are you need them)
 - Evaluate BC service providers to help plan, implement and host
 - Test, test, test