How Disaster-Tolerant Is Your Company?

By Bob Laliberte

Since 9/11 and the 2005 hurricanes, business continuity and disaster recovery have come off the back burner.

The events of September 11, 2001 didn't change everything— but they did bring into stark relief a collection of trends in business continuity/disaster recovery (BC/DR) that had been building since the 1990s: Businesses had become increasingly reliant on IT and communication links, while the Internet and, more recently, Web 2.0 were changing the way they interact, collaborate and conduct business. At the same time, datacenter and communications technologies that enable continuous operations had improved, and they had become more affordable.

Among the painful lessons of 9/11 was the realization in many IT shops that they needed a more robust BC/DR plan. Boardroom-level directors and officers also began to pay increased attention to BC/DR, and many made it a CEO-level mandate to improve the survivability of the IT and communications systems.

This article explores these trends and their impact on today's approaches to BC/DR. The current focus, as we shall see, is on keeping continuity of information access and communications flows. This is why the most prominent new term in the field is "disaster tolerant" environments. We no longer want to "recover" and "continue"—we want to avoid outages and ensure that our systems and services continue running no matter what.

Bob Laliberte is an analyst with the Enterprise Strategy Group, www.enterprisestrategy-group.com, specializing in strategic guidance and service to technology vendors, IT professionals, venture capitalists, and institutional investors. He can be reached at bob.laliberte@enterprisestrategy-group.com

What it Used to Be Like

The bottom line in today's business environment is simple: Companies cannot afford to have downtime. But it wasn't always like this.

Fifteen years ago, most companies had only one datacenter. Typically, this was a hardened facility with no single points of failure, including redundant power and communications connections, backup power, etc. Being disaster tolerant in the early 1990s meant making arrangements for disaster recovery, by performing a nightly tape back-up and then shipping that off to a secure vault somewhere. True business continuity was rare, even for organizations that could fail over to redundant components within the datacenter.

If and when a disaster struck, tapes and personnel were shipped to a designated recovery center, usually a shared facility operated by IBM, Sungard or another regional provider. The tapes were loaded and the people went to work to resume operations as quickly as possible.

The problem with this shared model is evident when there is a wide-scale disaster and everyone and their tapes converge on the out-of-region recovery center. Then, not only are personnel away from their families, but recovery point objectives (RPOs) are usually around 24 hours and recovery time objectives (RTOs) are usually 3–5 days (see "Decoding BC/DR Acronyms").

A few large financial organizations and government institutions had found the funding and motivation to create multiple datacenters, but the replication technology they used was very expensive and distance-limited. Communication links were costly and complicated to set up, so most of these secondary sites were located within 10 kilometers of the main datacenter.

If the local telco or some other provider couldn't offer dedicated fiber between the two locations, then expensive specialized equipment was necessary to move data over legacy telcom links. On the upside, these nearby sites were easily accessible to company employees, and often these sites were leveraged for application testing and backup functions. Using a nearby secondary site improved the RPOs of mission-critical applications to nearly zero, although RTOs were still measured in hours.

Few of these dual-site designs included voice communications—the focus was almost always on the back-end data stores and applications. While 9/11 brought increased executive attention to BC/DR, it was the terrible hurricane season of 2005 that awakened the understanding that data is of little use when there are no voice or Web connections with which to talk, email and post updates to customers and co-workers.

Decoding BC/DR Acronyms

The following terms often are used interchangeably, but they have different meanings:

Business Continuity Planning: This is the process of developing and preparing procedures and infrastructure that enable an organization to continue mission-critical business processes in the event of a disaster with little or no interruption to service.

Basically, this enables all work flows and transactions to take place without any noticeable disruption to the end user.

Disaster Recovery Planning: This is the process of developing and preparing procedures and infrastructure that enable an organization to resume mission critical business processes after a certain amount of time. In this case, service disruption is inevitable and acceptable. The concept here is to minimize the amount of time required to restore all business processes.

In the most common disaster recovery scenario, backup tapes are recovered from an off-site vault, new equipment is brought in and data is reloaded. Typically recovery times can be measured in hours or days.

So how does a company decide between implementing a business continuity plan or a disaster recovery plan? The answer is determined by asking the following questions:

- How much data can I afford to lose?
- How long can I afford to be without the applications that run my business?

These questions are commonly expressed as Recovery Point Objectives and Recovery Time Objectives and are defined as follows:

Recovery Point Objective (RPO): The acceptable amount of data loss when recovering from a disaster. Many enterprises accept that in the event of a disaster, they will recover from the last tape backup, which would equate to an RPO of 24 hours. Enterprises with more stringent recovery point objectives, such as in the financial markets, require zero data loss and have an RPO of 0.

Recovery Time Objective (RTO): The amount of time required to recover business processes; typically this includes recovery of applications, and data, as well as restoring end-user access to those applications.

Ideally, every company would choose to have a business continuity solution for every business process, unfortunately it can be cost-prohibitive to do this for all applications. To accurately determine which business processes (applications, communications—phone, email, messaging) require business continuity solutions and which ones only need disaster recovery, most enterprise customers perform a Business Impact Analysis (BIA).

Business Impact Analysis (BIA): The BIA is a comprehensive program where critical business processes are identified and prioritized and costs of downtime are evaluated over various time periods. Typically a BIA is performed by business units and IT departments. The goal of a BIA is to agree on a cost of downtime, establish Recovery Time Objectives (RTOs) and identify Recovery Point Objectives (RPOs).

Once the BIA is complete, the organization can make informed decisions about which aspects of their business need to be protected and to what degree or cost. The challenge is to match your needs and budget to the appropriate technology. With technology and services rapidly advancing, and as the costs are reduced, best practices would dictate a regular review of the BIA against existing technology to evaluate which processes could be upgraded□

Many companies simply cannot tolerate any downtime whatsoever

No More Downtime, Period

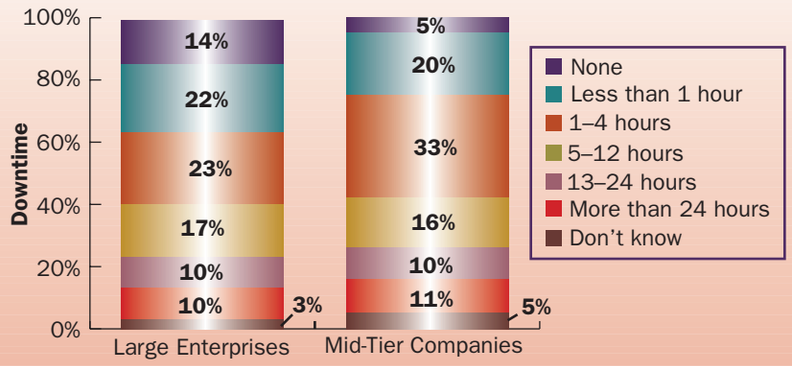
We have found that many companies and organizations simply cannot tolerate any downtime. In an ESG survey of mid-tier and large enterprise companies, 36 percent of large enterprise respondents said they will incur significant revenue loss or other adverse business impact if they have even an hour of downtime on their mission-critical applications. And 14 percent of large enterprises said they cannot tolerate any downtime at all.

Virtually any amount of downtime can mean lost productivity, lost revenue, lost customers and lost opportunities—not to mention damage to brand. Figure 1 (p. 46) illustrates the tolerance for downtime in large enterprises and mid-tier companies. All told, more than 80 percent of respondents reported that they cannot tolerate more than 24 hours of application unavailability. Survey respondents were from the financial sector, government, manufacturing, retail and health care (including pharmaceutical).

Regardless of the industry or company size, the trend is clear: More businesses require highly available, disaster tolerant solutions. Overall, today's approaches aim, not just to recover data, or to maintain financial record-keeping, but to continue virtually all operations.

Today, the goal is to have all end users able to access critical voice and data applications, regardless of their platform—Web application, call center or back office app. Large communications providers also are adapting their infrastructure to suit the needs of their enterprise customers.

FIGURE 1 How Much Application Downtime Can Your Organization Tolerate Before You Experience Significant Revenue Loss Or Other Adverse Business Impact?



Summarizing The Changes

In short, in less than a decade, the following major changes have occurred:

- Companies that had only one datacenter now have two. This also coincides with a trend to in-source business continuity functions instead of relying on a third party.
- Those datacenters are now much farther apart. Whereas synchronous data replication distances used to be 10 km, they now span distances of 100 km or more. Telco infrastructures are commonly dedicated, redundant optical networks to ensure high availability and adequate throughput.
- Datacenter technologies have advanced to enable immediate failover not only of stored data, but also of applications and servers. Clustering technologies enable any operation on a server to be monitored and when a signal is lost (i.e., no heartbeat), the secondary site automatically continues operations. RTOs are now measured in seconds and minutes.
- The large organizations that had two datacenters now have three. Typically the first two are in the same metro area (100 km) and the third is located out of region, thousands of km away (Figure 2). This configuration enables synchronous data replication to the first site and asynchronous data replication (data is not consistent due to latency) to the third site. This ensures zero data loss, and operations can resume as soon as the asynchronous data catches up. RTOs here are typically in minutes or hours, depending on configuration and amount of changed data.

Today, almost every business understands the need for some way to recover data, critical applications and communications in the event of a disaster. Depending on the size of the company and the industry, however, different levels of protection are being utilized. These solutions range from simply ensuring that backup tapes are stored off-site and having an emergency phone tree, to complex environments with double and triple sites that are capable of resuming operations within seconds (Figure 3).

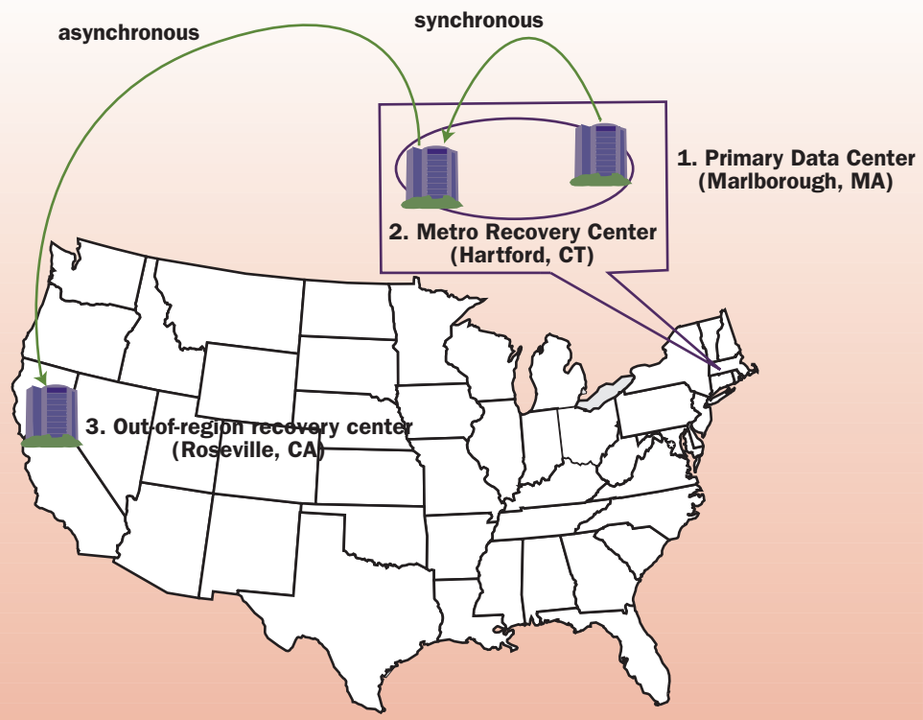
Different Levels Of Protection

Establishing a disaster tolerant infrastructure can be a very complex and costly undertaking. Having the ability to recover all mission critical applications, retain all data without any loss and ensure that end users have access to these applications is extremely difficult. The technologies involve mirroring of data and applications, remotely clustering servers, and ensuring redundant communication links to a duplicate workspace (phones, PCs etc.) for employees to resume their business functions and processes as quickly as possible. The need for ongoing voice communications and website services is also critical.

Fortunately, technology has advanced and costs have come down. However, it is still true that the more complex the environment and the shorter the RPOs and RTOs, the higher will be the cost of a solution.

For example, in a recent

FIGURE 2 Multisite Disaster Tolerant Environments



failover test, a Microsoft SQL database and a Unix system on a high-end server running a financial program were exposed to an actual disaster. Both systems were clustered and data was mirrored using array- and host-based replication respectively. In both cases, a quorum server—that is, a server that monitors the heartbeat of both servers at a third site—negotiated the failover for the servers. The high-end server and Unix had zero data loss and was accessible in about 14 seconds at the recovery site. The Microsoft SQL database was back on line in less than two minutes, also without any data loss.

Both solutions offer impressive recovery times, but the Unix system is typically more than twice the cost (depending on configuration) of the SQL solution, and both solutions are far more costly than backup to tape. Figure 4 (p. 48) provides a rough comparison of the costs of implementing a disaster tolerant solution (for more about this test, see www.hp.com/go/disasterproof)

In the upper left hand corner of Figure 4, solutions can easily cost in the millions of dollars; if whole new datacenters are required, the cost can run into the hundreds of millions of dollars. But when the cost of downtime is millions of dollars of revenue lost for every minute of downtime, these solutions don't look so costly.

ESG estimates that without investing in some type of backup solution, about one-third to one-half of companies that experience a disaster will be out of business in less than a year. In fact many companies that invest in business continuity solutions highlight their extensive BC capabilities as a selling feature that enables them to guarantee higher levels of service to customers and partners.

Depending on your company philosophy and size, you can choose to outsource your business continuity and disaster recovery solutions, or to build them yourself—usually with the help of selected strategic vendors.

Vendors like IBM, Sungard and HP, as well as a host of other regional/specialty providers, offer services to enable business continuity and disaster recovery via remote locations and shared or dedicated infrastructures. These companies leverage the cost of their servers and storage systems across all their customers, based on a shared model, which that assumes that not all their customers will declare disasters at the same time. In some cases there will be some data lost, and RTOs can vary.

Therefore, while these sites provide additional protection from a regional disaster, they cannot always ensure zero data loss and continuous operations. These vendors also offer dedicated equipment and data replications, but it is very expensive and, in order to have zero data loss, the company must be in close proximity (~100 km—due to latency) to the recovery center.

For companies that need this very high level of protection, building out their own secondary site makes the most sense. A second site also allows for continuous access for testing and for new application development. Typically, equipment vendors will provide blueprints and in some cases even "solution guarantees" that the infrastructure will work as advertised.

Almost every business needs to recover not just data, but also critical applications and communications capabilities

Looking Ahead

ESG expects that technology will continue to advance and virtualization services will help to enable greater efficiencies and higher availability. These solutions will have more intelligent and higher levels of automation, thus enabling a seamless operational experience.

For example, imagine an environment where software has abstracted the underlying infrastructure.

FIGURE 3 Complex Multisite Environment

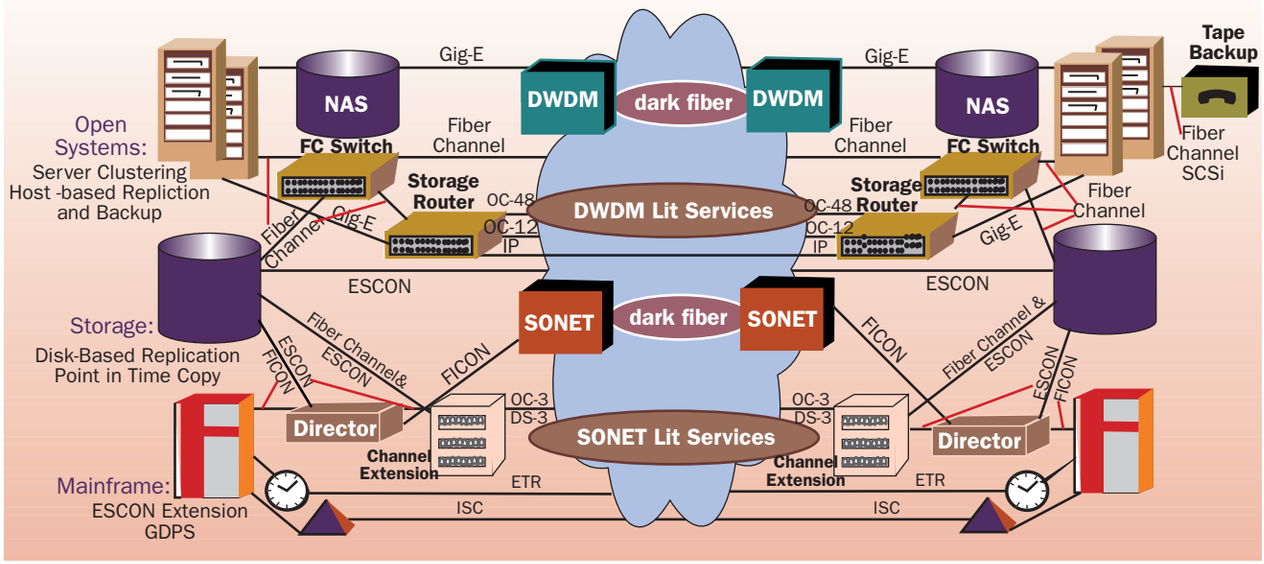
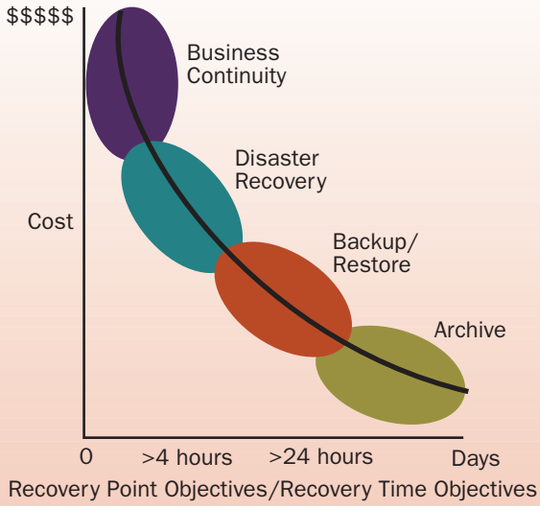


FIGURE 4 Cost Of Solutions vs. Recovery Point Objectives And Recovery Time Objectives



In this scenario there are computing resources, connectivity resources, data and communication resources. These resources are distributed among multiple sites, and protection levels vary based the value of the information, communication, etc. As new services are turned up, they will consume these virtualized resources based on desired RPO and RTO levels. Since all the infrastructure is virtualized and distributed, *n*-level redundancy is possible, and disruption to the business could be virtually eliminated, or measured in the milliseconds.

Certainly, we have a long way to go before this is achieved. Standards are still in development and not all vendors have embraced them. Many virtualization techniques for servers and storage are still in their infancy, but are developing rapidly.

The larger vendors continue to expand their offerings in order to deliver end-to-end solutions, while innovative startups are producing advanced features for specific areas. In fact, there are several small companies today working on the ability to verify data consistency across multiple sites, basically promising that, in the event of a disaster, you will be able to recover all data. These software packages would continuously monitor the replicated data and warn users as soon as something gets out of synch. This approach would certainly be better than the typical model of finding problems when a DR test is conducted on an annual or semi-annual basis.

Probably the biggest challenge to continuous uptime in the future will come from the newer types of disaster. While today most planners take into consideration natural disasters like hurricanes, floods, earthquakes and tornados and man-made issues like sabotage, equipment failure and terrorist attacks, they have just begun in the past few years to consider necessary procedures in the event of a pandemic.

In a pandemic, not only must the planners ensure a disaster tolerant infrastructure, but they must also plan for one that can be sustained in a "lights out" environment. Almost every company is set up to allow workers remote access to applications, but how well will that solution scale up when every employee has to have remote access? How ready is the datacenter set up for remote operation? In addition to just the sheer volume of voice and data traffic, there will be security implications. If you have a call center, how will all the calls be routed if all the employees are working from home? How will information be disseminated to all employees?

The answers to these kinds of questions must be weighed against the costs of solutions, and against the possibility of this occurring. Unlike many natural disasters, which are limited to certain regions, (earthquakes, hurricanes tornados etc.), a pandemic could be a global disaster.

Conclusion

The bottom line is that, in today's business environment, the ability to recover your critical applications, including communications, in real time is essential for a growing number of businesses of all types and sizes. The technology exists today to accomplish this, and can be consumed either directly or through a service.

The key to balancing a highly available, disaster tolerant environment and a desire to cut costs is an accurate Business Impact Analysis, then aligning the appropriate resources or services to ensure RPOs and RTOs are achievable.

Once a solution is implemented, be prepared to conduct regular testing to ensure operational readiness. If the everyday moves, adds and changes that take place in the primary datacenter are not mirrored to the secondary site, or updated to the recovery site, the sites can drift out of sync and thus be useless if disaster strikes. Regular testing guarantees that these faults are found and corrected before an actual disaster occurs.

In addition to regular testing, the company's business continuity plan (BCP) should be reviewed regularly against what potential threats may exist. A solution that was implemented a year ago may not handle future risks. BCPs should always be preparing for and anticipating the next possible threat and working to educate their executives about the risk so they can make informed decisions about the expense of implementing technology to minimize that risk. Companies should be prepared to continuously adapt and evolve their environments□

Companies Mentioned In This Article

HP (www.hp.com)
IBM (www.ibm.com)
Sungard (www.sungard.com)