# Protecting Your Vital Applications

**Mark Pollans**

**HPE Sr. Worldwide Product Manager**

This article is based on the presentation, "Delivering Business Continuity for Vital Applications" that was presented at the HPE Integrity NonStop Technical Boot Camp (TBC) 2017.

## What is Business Continuity?

When first presenting this topic it became apparent that if you hadn't spent time studying or working on business continuity projects, then the term might be a bit elusive. For those less familiar with the term Business Continuity, here is a way to look at it. Take your mobile phone for example; is it mission-critical? Business-critical? Well, maybe not for most. For some people though, their mobile is essential for them to do their job, pay a bill, or get an urgent message from a family member. We all have different levels of what's critical to us when it comes to our mobiles.

Now imagine a disaster strikes. You lost your mobile or dropped it in a puddle and damaged it beyond use. What would you do? What is your business continuity strategy for your "critical" mobile functions? (Side thought – does anyone even remember a phone number anymore)?

One such business continuity strategy is to run to your local electronics store and buy a new mobile. That's a strategy, and it could work for some. For others it would take too long to recover (see RTO below). And even then, what about your contact data, was it backed up somewhere? How would you do a restore of it? And how far back in time was your last "save," meaning all the changes you made on your mobile since then would be lost (see RPO below). To mitigate these concerns, some people carry two mobiles. Still, with two "active" mobiles, you will need a strategy to keep them synchronized. Depending on how critical your mobile applications and data are to you, there are different methods that you could employ to do so, each with its own characteristics.

Do you have a disaster plan for your mobile? Thinking about this plan before the disaster hits, is the first step of business continuity.

## Business Continuity Planning Begins with a Business Impact Analysis

As with your mobile, proper business continuity planning is often overlooked. Continuity planning begins with a business impact analysis to determine the potential effects of an outage of each vital application. A vital application is any mission-critical application that if unavailable, could result in a loss of revenue, a tarnished business reputation, lost productivity, or regulatory violations.

The criticality of applications can change over time as business processes and needs change. An application that was not mission-critical last year may be mission-critical today. (After all, how important was that first mobile phone as compared to today?) Therefore, the business impact analysis should be reassessed at least every one to two years.

## The Costs of Outages

Disasters happen more often than you may think; 95% of enterprises have experienced at least one unplanned data center outage in the last two years. During that time, the typical financial services business experienced 1.8 complete data center outages; and those in the healthcare industry experienced three such outages.[1]

The real costs of an outage can be significant. IDC estimates an average downtime cost of about $1.7 million per hour. IDC also notes that some outages can reach up to $10 million per hour! The average outage duration is 90 minutes, with some outages lasting 24 hours or more! Adding to this cost is the reputational damage to the company as news of the outage can spread very rapidly through today's social media.[2]

## RTO and RPO

The key elements of a business impact analysis are the Recovery Time Objective (RTO) and the Recovery Point Objective (RPO). These two terms provide a common framework to express requirements and objectives of a business continuity strategy and plan (Figure 1). RTO is the maximum acceptable time for recovery from an outage. It is the time it takes for an application or business process to be restored after a disruption. If the outage goes beyond the RTO, unacceptable consequences associated with a break in business continuity should be expected. On the other hand, as RTO approaches zero, the effects of an outage become less visible to the end users.

RPO is the maximum amount of data loss due to an outage. It is the data that has been generated between the last backup of data and the outage. It is often based on the average value of a transaction. However, in certain cases, no data loss is acceptable. For example, if a loss of data could result in the loss of life or limb, then that data absolutely must be protected.

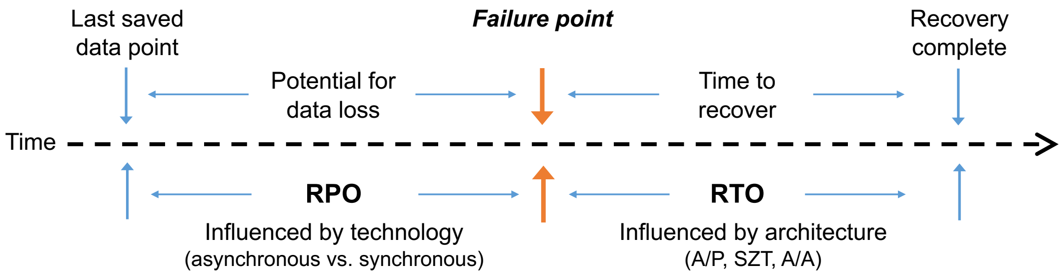

Figure 1: The RTO and RPO relationship

## Business Continuity Planning is a Strategic Imperative

"Critical applications" range from business-critical applications to mission-critical applications (Figure 2). Business-critical applications and data are necessary to run the business. Mission-critical applications and data are so valuable that any outage affecting them would be catastrophic. Critical applications can be at either end or in between.

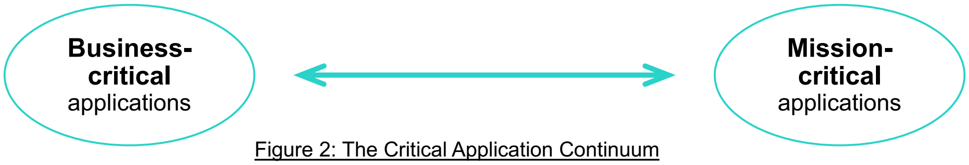

Figure 2: The Critical Application Continuum

Business continuity planning begins with the realization that it is not if, but when, a disaster will happen. Not all applications are critical. However, each critical application must be evaluated to determine what RTO and RPO are required in order to keep the business functioning. Here are some questions to ask when trying to place a particular application on the Critical Application Continuum: What is the potential lost revenue? What is the potential lost productivity? What are the potential lost business functions? What is the impact on customers? What are the legal and compliance issues? Since the needs of businesses change over time and applications evolve, this evaluation must be reassessed periodically.

Once an application's criticality has been assessed, a business continuity strategy then can be tailored to meet the required RTO and RPO objectives.Once an application's criticality has been assessed, a business continuity strategy can then be tailored for it to meet the required RTO and RPO objectives.

[1] Fingers Crossed? Or What is Your Business Continuity Plan for the Inevitable, Gravic, Inc., 2015 (original source Ponemon Institute).

[2] High-Value Business Applications on x86: The Need for True Fault-Tolerant Systems, IDC, May, 2015.

## Approaches to Business Continuity

Whatever approach is taken with the system architecture to provide the required availability for an application, at the very least, it should be geographically dispersed to survive local and regional disastrous events such as floods, fires, and earthquakes.

Here are three fundamental business continuity architectures described that provide a range of RTOs and RPOs:

- **Active/Passive systems using uni-directional data replication to maintain the passive (backup) system in synchronization with the active system.**
- **Active/almost Active systems, also known as Sizzling-Hot-Takeover systems (SZT), which are active/active systems with the application running on both of the systems, but only processing transactions on one.**
- **Active/Active systems, in which both (or all) systems are actively processing transactions. Their databases are kept synchronized via bi-directional replication.**

## Active/Passive Systems

An Active/Passive system is the classic disaster recovery solution. It is the minimally acceptable business continuity architecture for either business-critical or mission-critical applications.[3]

All transactions are executed on the active system, and changes to its database are replicated to the backup (or passive) system (Figure 3). The capacity of the backup system is mostly unused (unless other applications or read-only query/reporting functions are running on it).When a failover occurs, applications must be started on the backup system and the system tested before users can be connected to it. During this time, which could range from minutes to hours, the application is unavailable to the users.

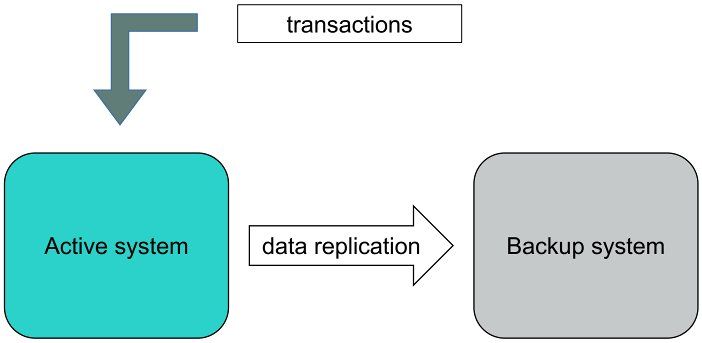

Figure 3: An Active/Passive Architecture

Failover testing generally requires an outage of the application on the active system. Therefore, failover testing is often not performed, or not performed to the fullest extent. Thus, there is a high risk that the backup system cannot be brought online when needed – this is known as a failover fault. Consequently, it is common practice to attempt to revive the original active system first, often resulting in lengthening the overall outage.

## Sizzling-Hot Takeover Systems

If an application cannot run in a distributed Active/Active environment, many of the advantages of an Active/Active architecture can still be achieved by running it in an SZT environment. In this architecture, the servers are configured as an Active/Active solution, but the application performs data changes on the active server only. On the standby server, the application is active, but is not performing data changes (Figure 4). The standby server's database is kept synchronized with the active server via data replication. If the active server fails, all that is required is to reroute transactions to the standby server and transaction processing continues uninterrupted.

Because the standby server has all the applications up and running and the database open for read/write access, it is easy to send test transactions at any time to verify that it is fully operational, which is a notable advantage over the Active/Passive architecture. Thus, when the need for a takeover arises, it is a known-working system and will not experience failover faults.

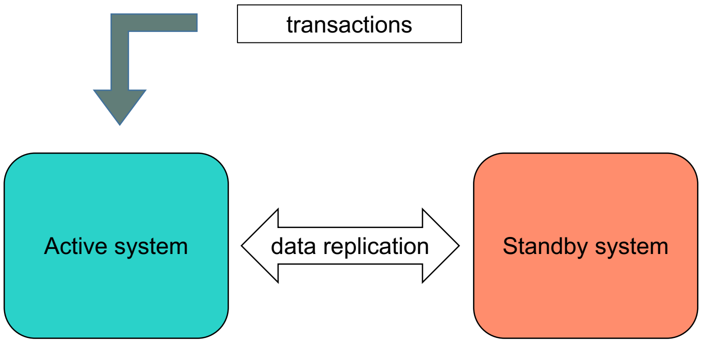

Figure 4: A Sizzling-Hot-Takeover Architecture

## Active/Active Systems

In an Active/Active environment, applications are active on all systems, and transactions can be sent to any system in the application network. Changes to the database in one system are replicated to all other systems via bi-directional replication to keep the databases synchronized (Figure 5).

Thus, the application has continuous availability even in the event of a system outage. If a system fails, all succeeding transactions are simply routed to surviving servers. Users connected to the surviving system(s) are not even aware that any outage has occurred.

Active/Active systems are more difficult to implement than Active/Passive systems. Some applications cannot run in a distributed environment (i.e., if the application assigns incremental invoice numbers from a memory-based counter).

[3] Although this is a bold statement to make in light of other recovery technologies such as virtual tape backup and restore, the recovery profiles of such approaches are so long and the data loss potential is so great that the applications being protected do not fit into the typical business-critical or mission-critical RTO and RPO categories.

Furthermore, data collisions are possible. A data collision occurs if the same data object in two different systems are modified at the same time. Both changes will be replicated to the other system. The data object values will be different in the two systems, and both will be wrong. Data collisions must be detected and corrected (i.e., the latest data change is accepted by both systems). Some data replication products offer a technology known as synchronous replication, which can prevent data collisions from occurring.

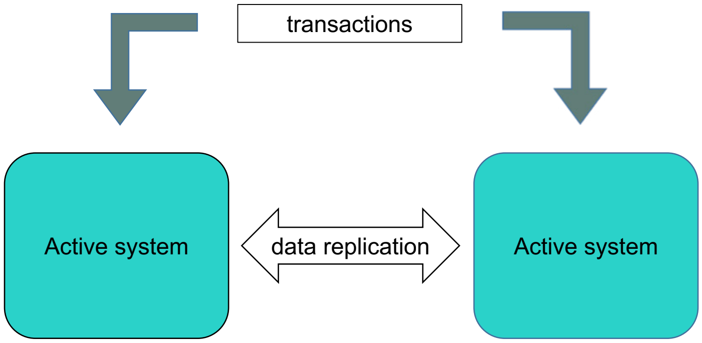

Figure 5: An Active/Active Architecture

## The Duration of an Outage

The architecture of the business continuity solution can significantly affect the duration of an outage. Average recovery times for different recovery architectures are[4]:

| | |
|---|---|
| Magnetic tape backup | 24 hours |
| Virtual tape | 12 hours |
| Active/Passive with failover faults | 3 hours (if at all) |
| Active/Passive without failover faults | 10 minutes |
| Sizzling-hot takeover | 30 seconds |
| Active/Active | 30 seconds |

Thus, the business continuity solution architecture should be chosen to match the criticality of the application to be protected and its RTO requirements.

## Summary

Business continuity is about minimizing or completely mitigating the impact of a catastrophe on your business applications. It is not a question of if a disaster will occur; it is a matter of when. A well thought out, planned, and tested business continuity strategy can provide a business with the assurance that an application can survive a disaster. The business impact analysis is a key part of that strategy. Understanding and creating appropriate RTO and RPO objectives for each application is a good next step towards meeting your business requirements.

Remember to review your impact analysis and RTO/RPO objectives at least every two years as business needs change and applications evolve. What was not mission-critical yesterday could be mission-critical today.

[4] This number is the time to switch affected users and transactions to surviving nodes; users/transactions already routed to surviving nodes see no outage at all.

Mark Pollans is HPE's Worldwide Senior Product Manager responsible for the HPE NonStop systems portfolio, including the NonStop X (Xeon® based) and the NonStop i (Itanium® based) systems, disk storage and new solid state drive technologies.

Most recently, Mark introduced the NonStop NS7 system. Previously, he orchestrated the release of the second and third generations of the HPE Integrity NonStop BladeSystem and four generations of HPE Integrity NonStop entry-class systems. Earlier, he introduced the NS16200 along with various NonStop platforms and storage solutions.

Mark has several years of experience at HPE, largely in enterprise computing and networking. During his tenure with HPE, he held various management and engineering positions in R&D and marketing for hardware and software projects.