Delivering business continuity for vital applications

Best practices for surviving data center disasters

Hewlett Packard Enterprise

Imagine that you are driving down the road on the way to the market. Out of the corner of your eye, something distracts you for just a moment. In that moment, the car in front of you stops and you end up crumpling the rear of that car and the front end of yours. That is traumatic enough if it happens, but what if you don't have insurance or the cash to cover the damages? What happens next?

As individuals, we seldom give a second thought to the need to carry insurance to secure us from accident, theft, or other loss. Businesses take a similar approach, sometimes spending millions of dollars annually to assess and mitigate risk. In many industries, IT organizations have taken a "money-is-no-object" approach to securing data from unauthorized access or exposure. When it comes to enabling the continuous availability of the applications that process data, or preventing loss from system and data center failures, some organizations may be operating with an antiquated idea of what really needs to be "always on." As a result, these organizations are often reluctant to expend the resources necessary to assure an acceptable level of business continuity.

IDC estimates the mean cost of downtime is roughly $1.7 million USD per hour across industries, with some approaching $10 million USD per hour. The average downtime incident is 90 minutes in length, with some lasting over 24 hours.[3]

The risk is real

Ask yourself: What does it cost in lost revenue if a customer-facing system is unavailable? What is the impact, to the business or your customers, of losing customer transactions in-flight? For that matter, what does it cost your business in terms of lost productivity if your email or unified communication and collaboration systems fail? What if there is a fire in the data center, power is lost, or an earthquake causes damage, effectively taking the center offline? What happens next?

These things happen more often than you might think. Ninety-five percent of enterprises have experienced at least one unplanned data center outage in the past 24 months—not just an individual system but an entire data center. The average financial services business experienced 1.8 complete data center outages over the last 24 months. In healthcare, the average is three outages in the past 24 months.[1]

IDC estimates that the mean cost of downtime is roughly $1.7 million USD per hour across industries, with some outages approaching $10 million USD per hour.[2] The average incident is 90 minutes in length, with some lasting over 24 hours. Along with lost revenue, the real cost of downtime can include damaged reputation, lost customer confidence and loyalty, damaged competitive position, and even regulatory compliance exposure. In today's social-media–driven world, it doesn't take long for news of an outage to go viral, and it can take years to unwind the damage.

Recovery objectives

If an application service fails, what is the acceptable timeframe for recovery? This recovery-time objective (RTO) varies from application to application: it could be five seconds, five minutes, or five days. The cost of downtime is usually the driving factor in setting an application's RTO. Some apps have an RTO of zero, meaning that the customer or end user should never be aware of a failure.

If an application service fails, what is the maximum acceptable amount of data loss? This recovery-point objective (RPO) can be set based on the nature of the data and how important it is to the enterprise, the value of the data to the enterprise (i.e., the costs in terms of lost revenue, including potential legal and compliance liability), or a combination of both.

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

[2, 3] "**High-Value Business Applications on x86: The Need for True Fault-Tolerant Systems**," Peter Rutten, IDC, May 2015

RTO: the maximum acceptable time for recovery from an outage

RPO: the maximum acceptable amount of data loss from a system outage

RPOs are often based on the average value of a lost transaction. While that might sound reasonable on the surface, when you dig a little deeper that approach doesn't always make sense. For example, for a financial institution, the average electronic funds transfer (EFT) might be $1,000 USD, but the largest might be millions of dollars.

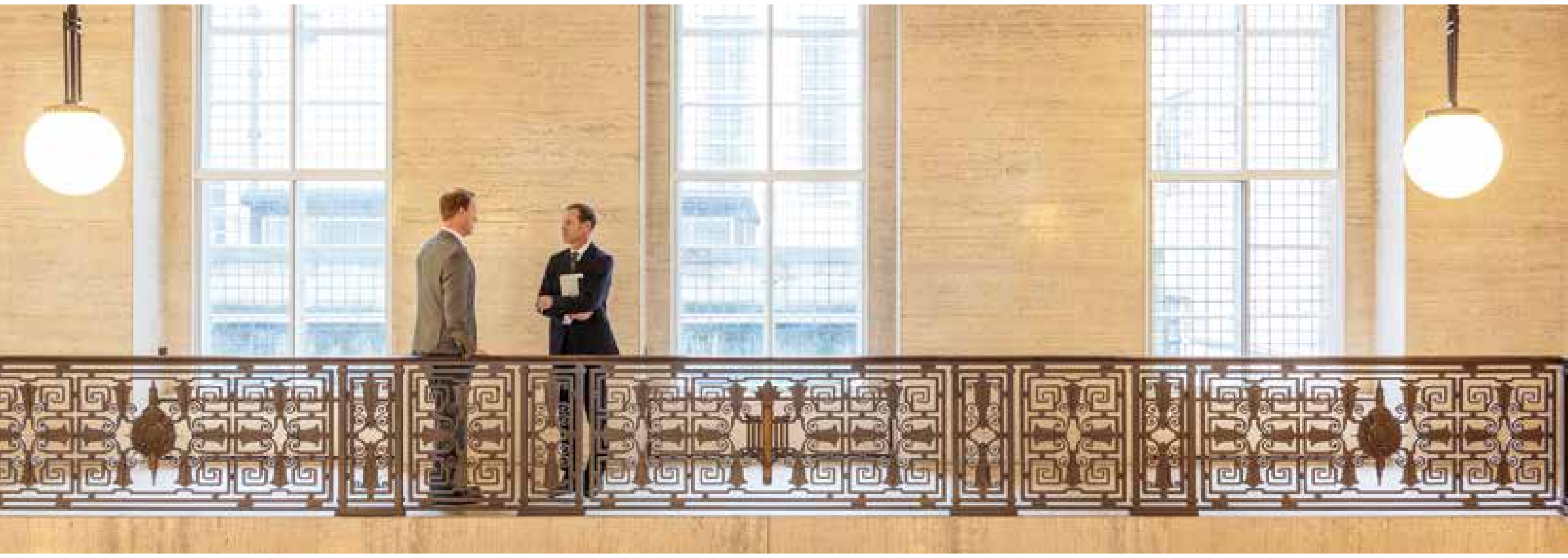
Since you can't predict which transactions might fail, the real potential cost of an outage is the cost of the most valuable data that could be lost (for example, the largest EFT transaction, the biggest sales order, the largest volume trading transaction, the greatest legal liability from lost data, or the largest account). This is the value that should determine RPO goals.

Business continuity

It usually starts with a conversation about lost transactions, lost revenue, or lost productivity. These business problems prompt organizations to start thinking about business continuity as a strategic imperative. It's not a question of if systems will fail or if a catastrophic event will occur; it's a question of when. Then the question becomes: What happens next? Business continuity is the practice of ensuring that your business is able to continue functioning regardless of what happens.

From a practical perspective, business continuity requires examination (and re-examination, with best practices suggesting reassessment every one to two years) of the tolerable RTOs and RPOs for every system necessary to store or serve vital applications and their data. In some cases, determining this might be fairly simple. For example, a medical-records system always must be available and there can never be any data loss. Patient lives depend on it. If a failure occurs, the system must be immediately recoverable to the point that users never know there was a problem.

In other cases, tolerable RTOs and RPOs might be a little more difficult to determine. For example, how valuable is your online store, VoIP phone, or customer relationship management (CRM) system? What happens if one of these goes down? Is it mission critical or business critical? Could much of your business still survive while these systems are down? How long can you afford to spend recovering lost data? What is the impact if customers can't reach you? This type of examination often reveals that many systems are more vital to your business than you might think. These systems require stronger business continuity support with more stringent RTOs and RPOs than are currently in place. On further review, these applications or services might even turn out to be mission critical, meaning they can never go down without significant business impact. Keep in mind that what is not mission critical today may be so tomorrow.



Defining what is "critical"

We've established that "business-critical" applications and data are necessary to effectively run the business and "mission-critical" apps and data are so valuable that any outage would be catastrophic. Now ask yourself: What is the impact if a mission-critical system is unavailable or if transaction data is lost? What's the impact on your customers? On your business revenue? Does loss of data create potential legal or compliance issues?

Support for business continuity must embrace a continuum from mission-critical to business-basic apps and data. Evaluation of your applications and systems along this continuum should be based on the needs of your customers and your revenue goals. An application with very limited tolerance for its availability (zero or near-zero RTO) or for the amount of data you could lose (zero or near-zero RPO) should be considered mission critical.

Applications that either can't go down or must be back up and running so quickly that no one notices should be considered mission critical.

Some examples of vital applications and services include:

- **Financial services:** payment processing, fraud prevention, high-frequency trading
- **Telco:** mobile network management; machine-to-machine, real-time customer service
- **Retail:** point of sale, e-commerce, online transactions, and order processing
- **Manufacturing:** continuous production control processes and multi-channel distribution
- **Healthcare:** real-time patient and lab data, provider information retrieval
- **Transportation:** reservations, ticketing, scheduling

In an always-on world, downtime for complex, interconnected, customer-facing workloads is usually not an option.

Business continuity approaches

True business continuity requires a level of geographic dispersion (or distance) to survive both localized events (such as a fire in the data center) as well as regional failures (such as regional power grid collapse). There are three basic approaches to creating geographically distributed business continuity infrastructure, each with different RTO and RPO profiles:

- **Asynchronous Active/Passive:** This is a classic disaster recovery (DR) scenario where all transactions execute on an active system and data asynchronously replicates to a passive backup node. In the event of failure, applications must be started up on the backup node, which can result in delay and a longer RTO. Failover procedures to restart these applications are often difficult to test or execute and the risk of failure is high.
- **Asynchronous Active/Almost-Active:** Also known as Sizzling-Hot-Takeover (SZT) or Sizzling-Hot-Standby, this approach is similar to an Active/Passive architecture using replication, except that the backup node is immediately ready to start processing transactions with the local copy of the application database already open for read or write access. It is basically an Active/Active architecture, except that all user transactions are directed to the primary node. This greatly improves the chances of success when failover occurs and gives a much better and repeatable RTO than classic DR.
- **Asynchronous Active/Active:** In a disaster-tolerant architecture, production processing is split across multiple nodes and each node has a copy of the database synchronized using bi-directional data replication. If one node fails, its traffic can be automatically routed to other active nodes. Users connected to the surviving nodes are unaware that an outage has occurred. Failover becomes a simple process that can be easily tested and practiced, because all nodes are known to be working at all times.

For mission-critical applications, a disaster tolerant, multi-node architecture is the best alternative and provides the best available RTO and RPO. Of course, there are multiple technologies that assist with creating business continuity solutions, from software based transactional data replication to hardware-based clustering and RAID technologies. Each will have limits on the RTO and RPO capabilities they can provide.

One other consideration is atomicity, one of the four atomicity, consistency, isolation, and durability (ACID) concepts for database design and application architecture. Atomicity defines an "all-or-nothing" rule for transaction processing: a transaction begins at a certain point in time, and if anything goes wrong before the transaction is fully complete, the transaction is unwound all the way back to the beginning as if it never occurred. Business continuity approaches should be designed to observe this principle, especially for mission-critical applications.

Total cost of ownership

If you have one failure, one event, or one disaster that causes an application outage that's longer than acceptable to your customers, you should have had a disaster tolerant architecture in place to avoid the loss of business and loss of reputation. In the age of social media, customers will vent their frustration in seconds and the issue can go viral in minutes. For businesses with strict compliance restrictions around downtime, a disaster-tolerant architecture will mitigate governmental penalties and reporting. For example, if a system goes down and takes three hours to fully recover, using the average downtime costs from the IDC study cited previously, the cost could be over $5 million USD. The same failure might last only a few seconds or even be invisible to an application that resides in a well-designed business continuity environment, resulting in a much smaller loss.

When looked at in terms of total cost of ownership (TCO), classic Active/Passive DR architectures present high TCO because of the cost of downtime.
In general:

- The better the availability, the greater the complexity and implementation costs
- The better the availability, the lower the outage costs

The net result is that as implementation cost increases, overall TCO decreases, but at a much faster rate.[4] To put it another way, you can buy a lot of fault tolerance for the cost of a single outage.

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

Business Continuity offerings

HPE Services
To help customers with today's business-critical data center, Hewlett Packard Enterprise offers skilled and experienced professionals to deliver a comprehensive range of advisory, design, deployment, and management services for disaster-tolerant architecture.

HPE Integrity NonStop X
HPE Integrity NonStop X systems are uniquely designed for industries that never stop and have the highest level of availability, system-wide security, massive scalability, and lowest TCO in class. According to IDC's highest Availability Level 4 (AL4) definition, AL4 is the combination of multiple hardware and software components that allows a near-instantaneous failover to alternate resources so that business processing continues as before without interruption.[5] HPE Integrity NonStop X, paired with HPE NonStop Shadowbase software, delivers AL4 fault tolerance across geographic locations, providing unmatched continuous availability for zero planned or unplanned downtime, nearly eliminating the possibility of an application outage. Reimagine mission critical compute fine-tuned for maximum availability, scalability, and data integrity with HPE NonStop X.

HPE XP Storage
HPE XP7 Storage is designed for hybrid Flash storage and is ideal for mission-critical applications that require continuous data availability, scalability, and performance. Array-based virtualization technology enables multi-site and multi-array virtualization, replication, and management to increase availability, avoid disasters, and improve resource utilization by helping eliminate storage silos. HPE XP7 Storage is the most highly available SAN array that HPE offers with a number of software solutions that help achieve the highest recovery objectives, remote replication, and DR capabilities.

[5] "Worldwide and U.S. High-Availability Server 2014–2018 Forecast and Analysis," IDC, Doc #250565

Conclusion

It's not a question of if some catastrophic event will impact a mission-critical system, it's a question of when. Once the event occurs, what happens next? A business continuity solution architected for continuous disaster tolerance can help minimize the damage, providing the best available recovery-time and recovery-point capabilities at a TCO that makes sense for your business.[6]

Learn more at
hpe.com/info/nonstop
hpe.com/storage/xp
hpe.com/info/dcm

[6] "**High-Value Business Applications on x86: The Need for True Fault-Tolerant Systems**," Peter Rutten, IDC, May 2015

Sign up for updates

★ Rate this document

Hewlett Packard Enterprise

4AA6-5326ENN, May 2016