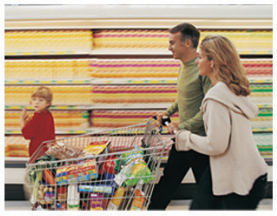
THE COST OF CRITICAL APPLICATION FAILURE







Stratus® Technologies

For an **Always-On** World

www.stratus.com

The Cost of Critical Application Failure

Make Better Technology Decisions By Understanding Your Cost of Downtime

Executive Summary

Server downtime is an issue that plagues the vast majority of organizations. When servers go down, many—if not all—of an organization's most critical applications become unavailable, and the cost of being unable to do business mounts minute by minute.

But how much does server downtime actually cost you? The reality is that the majority of organizations do not know the answer to this question. Few measure their cost of application downtime, and if they do, they measure it incorrectly. Without knowing this exact cost, an organization's ability to make sound investments in data center technology and availability protection is impaired.

Ultimately, high availability is a business decision that is based on the value a computer system has to an enterprise. When thinking about availability solutions, it is important to consider that the applications that levy the highest downtime cost on your organization are likely to be the ones you want up and running first after a server outage. What is the value of these applications, and what does it cost you when they are inaccessible? Without knowing the true cost of downtime, your organization can't properly and cost-effectively protect itself.

In this paper, you will learn:

- Three different levels of availability, what they mean, and the impact of downtime on your organization
- Six costs of downtime—some of which may be surprising
- Four options for protecting your critical applications against downtime
- How relying on a virtualized environment impacts your choice of availability solution

It is essential to know exactly how much downtime is costing your organization so you can properly protect your critical applications with the best solution for your needs.

What Exactly Is "Availability"?

The term "availability" is a characterization of how reliably a computing system can function and run the tasks it was designed to run. When talking about servers, there are different levels of availability:

1. **Backups and restores:** Basic backup, data-replication, and failover procedures are in place via conventional servers. Recoverability translates into 99% to 99.9% availability.

2. **High availability:** Applications are accessible a very high percentage of the time. Users perceive little or no interruption if there is a failure. High availability translates into 99.95% to 99.99% availability.

3. **Continuous availability:** Even if there is a failure, server operations are not interrupted. Downtime is eliminated and data is not lost in the event of a server failure. Continuous availability translates into 99.999% availability.

At first glance, there doesn't seem to be a huge difference among all these percentages of 9s. However, when considering that Aberdeen Research recently found that downtime now costs an organization an average of $138,000 per hour,[1] the cost differences become readily apparent:

Figure 1: Understanding the Nines

Availability Level	99%	99.9%	99.95%	99.99%	99.999%
Downtime Per Year	87.6 hours	8.76 hours	4.38 hours	57 minutes	5 1/2 minutes
Total cost of downtime per year	$12,088,800	$1,208,880	$604,440	$131,100	$12,650

Note the large difference in the cost of downtime between 99.95% and 99.99%. Just a 0.04% difference in downtime means a cost difference of greater than 4.5x. The value of greater availability is evident.

In June 2010, Aberdeen found that downtime cost, on average, were $98,000 per hour. In 2012, that number rose to $138,000—a 19% year-over-year increase.[2]

The Cost of Downtime Involves Much More than Just Lost Wages

It may be a surprise that the cost of downtime involves much more than just lost wages and actually affects the entire company. But when calculated correctly, the cost of downtime impacts the whole organization, with no one group seeing the entire impact.[3] Sales are lost. Employee productivity goes down. Customers become frustrated. And competitors can benefit.

The costs of downtime include both direct and indirect costs. Direct costs are expenses that can be completely attributed to the production of specific goods or to a particular function or service, whereas indirect costs are more difficult to quantify—but can be even more damaging to an organization.

The costs of downtime can be broken down into the following categories:

Business costs. These are the first costs that come to mind for most people. Lost wages, overtime, and remedial labor costs all add up during an outage.

Sales can be lost and so can future repeat business. For example, imagine the consequences if a retailer experiences downtime during the holiday season. Potential customers won't have the time or patience to put up with an outage and will take their business elsewhere. During the 2011 holiday season, major retailers such as GameFly, Brookstone, Sears, Sephora, and Abercrombie & Fitch all experienced major outages.[4] The cost to these retailers was, no doubt, extensive.

Other business costs include lost inventory and the scrap of work in progress, potential legal penalties for not delivering on service-level agreements, and litigation costs due to third parties seeking compensation for losses incurred during a system outage.

Productivity costs. During an outage, employees can't perform their regular duties. The impact of this idle time varies by industry. For example, in an office environment, an employee may not be able to access the Internet but can work on a desktop spreadsheet program, so perhaps his or her productivity would be cut in half. But in a manufacturing environment, if the line stops, employees may be 100% unproductive.

A common way to calculate productivity loss is:

(average employee wage x number of hours production stops) + overtime costs for employees to make up lost work time

If employees are salaried and not paid overtime, they may be forced to put in extra work hours, which can take a toll on both employee productivity and morale—tougher to measure but certainly still costly.

Recovery costs. These costs include the price paid to repair the system, IT staff overtime, and third-party consultants or technicians needed to restore services.[5] Another consideration: the opportunity cost sacrificed when IT needs to focus on system recovery instead of working on other critical projects for the organization.

Customer loss. The effects of indirect costs can be felt long after an outage is resolved. Previously loyal customers can lose faith and take their business to competitors. Once a company is seen by its customers as unreliable, it can be very difficult to undo the perception.

Reputation damage. Bad publicity can cause major damage to an organization—and not just large ones. It's true that the traditional press loves a good headline about bad news at a big company. But what can a complaint on Twitter or a negative post on Facebook cost you? Convergys found that one bad tweet can cost a company 30 customers.[6] And while industry websites and bloggers don't always have a large audience base, they do have the rapt attention of your target market. This means that one negative blog post about your company can make a huge impression on your customers and prospects.

Shareholder value impact. Bad press can also devalue a company's stock and reduce its market capitalization. Especially in shaky economic times, the stock market reacts to negative press about a company, even more so if the news is about a significant sales loss—an event that is entirely possible when servers go down.

For the highly demanding, most critical workloads accessed by large numbers of end users, lengthy periods of downtime—especially unplanned downtime—are costly to the business and should not be tolerated.[7]

Downtime Effects Vary by Industry

Downtime can affect different types of organizations in different ways, and it's important to take these additional factors into consideration when thinking about the cost of downtime.

In **manufacturing**, unexpected downtime can mean lost inventory, a lower-quality product, and/or unsellable products. In some cases, a momentary disruption in production can cause an entire run to be scrapped due to regulatory guidelines—a potentially devastating scenario and a harsh reality for food and pharmaceutical manufacturers. When production deadlines are missed, the business can be impacted both financially and in terms of reputation.

The **retail sector** is hit hard by IT downtime, losing $18.18 billion per year due to outages.[8] A single downtime event for a retailer can be a huge blow to its financials, especially when such an event happens during a holiday shopping season.

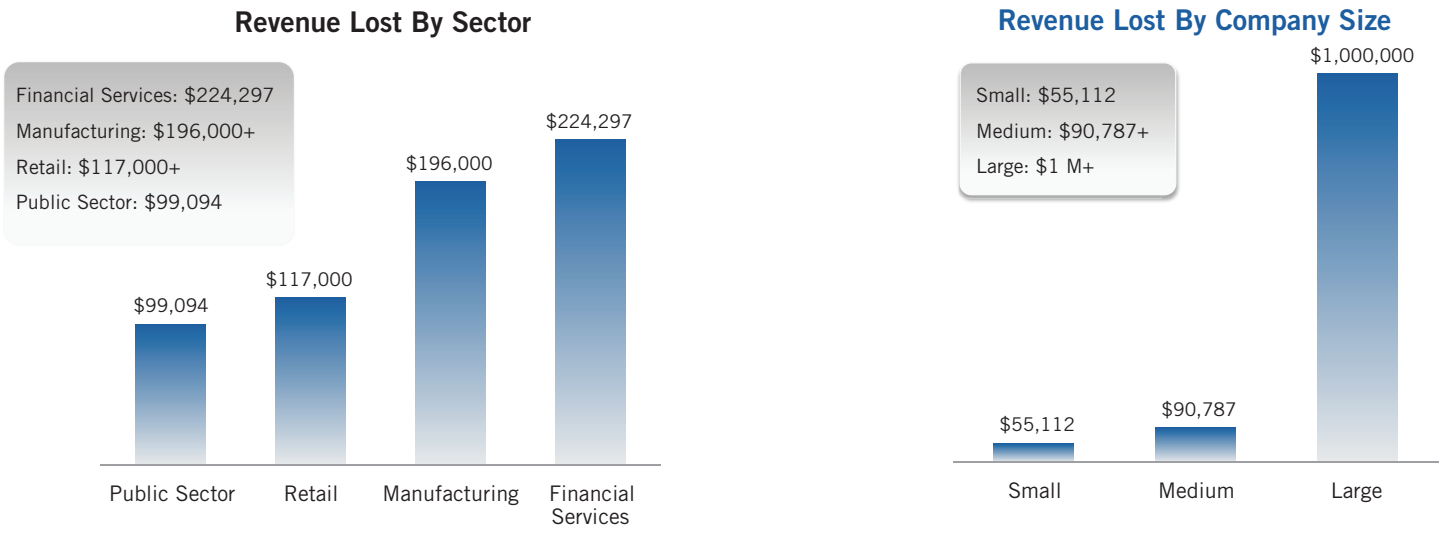
When a server goes down for an online retailer, website performance is compromised, which frustrates customers and may cause them to abandon their online shopping carts and shop elsewhere. In a store setting, point-of-sale (POS) systems need to be up and running to process sales and maintain the flow of customers throughout the store. Server downtime can mean poor customer service when employees can't check on product availability and long, slow checkout lines for customers trying to make their purchases.

Public safety organizations have their own unique set of concerns. Public safety 911 call takers, dispatchers, and first responders all depend on the applications and information managed by computer systems to protect lives and property. Downtime of public safety answering point (PSAP) applications causes slower emergency response times and can tragically result in the worst type of loss: loss of life.

If computer systems are offline, they cannot give location history and fire inspection data to first responders. And sudden system downtime for police departments means, for example, an officer would not be able to tell if an individual he or she just pulled over is a wanted criminal, which puts both the officer and the public in danger. Downtime also harms the reputation and public perception of a department and even opens the department up to potential lawsuits.

For **financial services** organizations, downtime affects transactions. Customers want to quickly and securely complete their transactions, whether over the Internet, by telephone, at a local branch office, through an ATM, or via debit/credit card. When downtime occurs, financial institutions are hit hard on a company level: A CA Technologies report says that revenue loss due to IT downtime is $224,297 per company each year.[9]

Figure 2: Annual Revenue Losses



Source: CA Technologies, *The Avoidable Cost of Downtime* (CA Technologies, 2010), 4.

Four Options for Protecting Your Critical Applications

After your cost of downtime is thoroughly understood, it's important to consider the level of availability your most important applications need. For business-critical applications such as CRM, ERP, back-office databases that run the business, financial software, and email servers, service interruption and data loss are very expensive.[10] Experiencing hours of downtime for these applications is likely an unacceptable scenario.

Depending upon your business or organization, you may also have your most critical applications for which no downtime is acceptable. Some examples include:

- Manufacturing execution systems (MESs)
- Security systems
- Trading and banking systems
- Electronic medical record (EMR) systems
- Applications that support emergency response operations
- Applications that control life-sustaining processes
- Military and civilian security applications

Keeping in mind your desired availability for your critical applications, there are four options to consider for guarding against downtime.

Standard Servers: Always-On Level of 99%

A standard x86-based server typically stores data on RAID (redundant arrays of independent disks) storage devices. The capabilities of x86 servers range from vendor to vendor and support a variety of operating systems and processors.

However, a standard x86 server may have only basic backup, data-replication, and failover procedures in place, which means it would be susceptible to catastrophic server failures. A standard server is not designed to prevent downtime or data loss. In the event of a crash, the server stops all processing and users lose access to their applications and information, so data loss is likely. Standard servers also do not provide protection for data in transit, which means if the server goes down, this data is also lost.

Though a standard x86 server does not come from its vendor as highly available, there is always the option to add availability software following initial deployment and installation.

Traditional High-Availability Solutions: Always-On Level of 99.9% to 99.95%

Traditional high-availability solutions that can bring a system back up quickly are based on server clustering: two or more servers that are running with the same configuration and are connected with cluster software to keep the application data updated on both/all servers.

Servers (nodes) in a high-availability cluster communicate with each other by continually checking for a heartbeat that confirms other servers in the cluster are up and running. If a server fails, another server in the cluster (designated as the failover server) will automatically take over, ideally with minimal disruption to users.

Computers in a cluster are connected by a local area network (LAN) or a wide area network (WAN) and are managed by cluster software. Failover clusters require a storage area network (SAN) to provide the shared access to data required to enable failover capabilities. This means that dedicated shared storage or redundant connections to the corporate SAN are necessary.

While high-availability clusters improve availability, their effectiveness is highly dependent on the skills of specialized IT personnel. Clusters can be complex and time-consuming to deploy, and they require programming, testing, and continuous administrative oversight. As a result, the total cost of ownership (TCO) is often high.

It is also important to note that downtime is not eliminated with high-availability clusters. In the event of a server failure, all users who are currently connected to that server lose their connections. Therefore, data not yet written to the database is lost.

Advanced High-Availability Solutions: Availability of 99.99%

The most advanced high-availability solutions are software designed to prevent downtime, data loss, and business interruption, with a fraction of the complexity and at a fraction of the cost of high-availability clusters. These solutions are equipped with predictive features that automatically identify, report, and handle faults before they become problems and cause downtime.

Two important features of advanced high-availability software are that it works with standard x86 servers and doesn't require the skills of highly advanced IT staff to install or maintain it. SANs are also not required, which makes the system easier to manage and lowers an organization's TCO. Advanced high-availability software is designed to configure and manage its own operation, making the setup of application environments easier and more economical.

There is a key difference between high-availability clusters and advanced high-availability software: The software continuously monitors for issues to prevent downtime from occurring, whereas cluster solutions are designed to recover after a failure has already occurred. The prevention of downtime is the goal of high-availability software, and the most effective solutions provide more than 99.99% availability, which translates to less than one hour of unscheduled downtime per year.

Fault-Tolerant Solutions: Availability of 99.999%

Fault-tolerant solutions are also referred to as continuous availability solutions. A fault-tolerant server provides the highest availability because it has system component redundancy with no single point of failure. This means that end users never experience an interruption in server availability because downtime is preempted.

67% of best-in-class organizations use fault-tolerant servers to provide high availability to at least some of their most critical applications.[11]

Fault tolerance is achieved in a server by having a second set of completely redundant hardware components in the system. The server's software automatically synchronizes the replicated components, executing all processing in lockstep so that "in flight" data is always protected. The two sets of CPUs, RAM, motherboards, and power supplies are all processing the same information at the same time. So if one component fails, its companion component is already there, and the system keeps functioning.

Fault-tolerant servers also have built-in, failsafe software technology that detects, isolates, and corrects system problems before they cause downtime. This means that the operating system, middleware, and application software are protected from errors. In-memory data is also constantly protected and maintained.

A fault-tolerant server is managed exactly like a standard server, making the system easy to install, use, and maintain. No software modifications or special configurations are necessary, and the sophisticated back-end technology runs in the background, invisible to anyone administering the system.

"Organizations often overlook fault-tolerant hardware-based solutions because they believe that they're simply too expensive. In a full analysis of costs of hardware, software, staff-related costs, and the cost of downtime, they just might find that this is one of the lower-cost solutions."[12]

Considering the Specialized Needs of a Virtualized Environment

"High availability is an issue that most consolidated virtual environments face. After all, if a company puts all of its eggs (in this case, virtual servers) into one basket, extreme care must be taken to not to drop that basket or allow it to fall on the floor."[13]

Virtualization is the practice of using a software layer to let one physical computing server run multiple virtual machines. Server virtualization allows an organization to save money by consolidating a number of applications on the same physical server and provides an environment that is easier for IT staff to manage. In fact, the capital costs and operating costs for a data center can be significantly reduced with virtualization, often by a factor of four or greater.[14]

If you already have a high percentage of your applications in a virtualized environment, you are in good company. An Aberdeen survey found that in a 10-month period, there was an 18% increase in the percentage of corporate applications virtualized. As of March 2011, the surveyed organizations reported that 49% of their applications were currently virtualized, and at the end of all current projects, 66% of all applications would be virtualized.[15]

Relying on virtualization, however, means that a greater number of applications are exposed to downtime. The practice of virtualization does increase redundancy and availability, but it does not guarantee continuous operation. The virtualization server still has the potential to be a single point of failure for all the applications it supports. Therefore, when you consolidate dozens of virtual machines to a single physical machine, that system requires continuous availability - even if all the applications it's running do not.

If you are depending upon a virtualized environment, it is crucial to employ a continuous availability strategy in order to guarantee that virtual machines can continue to function. Many industry experts recommend using a fault-tolerant solution in a virtualized environment. Says managing editor of *The Availability Digest* Dr. Bill Highleyman, "Especially when the cost of downtime is considered, virtualized fault tolerance can bring continuous availability to a data center at a competitive cost and with no special administrative skills."[16]

Conclusion

The need for availability has become critical to most organizations. "The server is down" is not an acceptable excuse for systems not working. Downtime affects the whole organization, and its costs are both direct (lost wages, lost sales, lost customers) and indirect (loss of employee productivity, reputation damage, opportunity costs).

Understanding the cost of downtime to your organization is the first step in creating a plan to manage the risk. The next step is to think about the level of availability that your most critical applications need and whether your goal is to recover as quickly as possible after a failure has occurred or to prevent failure altogether. The answers to these questions will help you determine the appropriate course of action for your organization.

About Stratus

Stratus Technologies is the leading provider of infrastructure-based solutions that keep your applications running continuously in today's always-on world.

Stratus always-on solutions can be rapidly deployed without changes to your applications. Our platform solutions provide end-to-end operational support with integrated hardware, software and services. Our software solutions are designed to provide always-on capabilities to applications running in your chosen environment – physical, virtualized or cloud. Our approach and our people enable us to identify problems that others miss and prevent application downtime before it occurs. Multiple layers of proactive diagnostic, monitoring and self-correcting services are backed by a global team of engineers who provide immediate support no matter where in the world your system is located.

If **always-on** is an application requirement, Stratus Technologies has a solution that fits.

[1] *Dick Csaplar, Four Steps to Setting the Right Budget for Downtime Prevention (Aberdeen Group, 2012), 1.*
[2] *Csaplar, Four Steps, 1.*
[3] *Csaplar, Four Steps, 4.*
[4] *"82 Sites with Recent Outages," Panopta Availability Index, accessed June 11, 2012, http://holiday.panopta.com/#less-than-100.*
[5] *Deepak Mane, "Calculating the Costs of Downtime," Disaster Recovery in Cloud Computing (blog), July 16, 2010, http://deepak4278.blogspot.com/2010/07/calculating-costs-of-downtime.html.*
[6] *"One Bad Twitter 'Tweet' Can Cost 30 Customers, Survey Shows," Bloomberg.com, accessed June 11, 2012, http://www.bloomberg.com/apps/news?pid=newsarchive&sid=afod9i5PqoMQ.*
[7] *Jean S. Bozman and Lloyd Cohen, Worldwide and U.S. High-Availability Server 2011 - 2015 Forecast and Analysis (IDC 2011), 44.*
[8] *CA Technologies, The Avoidable Cost of Downtime (CA Technologies, 2010), 4.*
[9] *CA Technologies, The Avoidable Cost of Downtime, 5.*
[10] *Citrix Systems Inc., The Three Levels of High Availability, 4.*
[11] *Dick Csaplar, Best Practices for Protecting Virtualized Applications (Aberdeen Group, 2011), 2.*
[12] *Dan Kuznetzky, "Hardware Fault Tolerance in a Virtual Environment," Virtually Speaking (blog), March 10, 20*
[13] *Dan Kuznetzky, "Stratus Uptime Appliance for VMware vCenter Server," Virtually Speaking (blog), February 22, 2012, http://www.zdnet.com/blog/virtualization/stratus-uptime-appliance-for-vmware-vcenter-server/4644.*
[14] *W.H. Highleyman, Fault Tolerance for Virtual Environments: Part Three (Sombers Associates, 2008), 2.*
[15] *Dick Csaplar, Best Practices for Protecting Virtualized Applications, 1.*
[16] *W.H. Highleyman, Fault Tolerance for Virtual Environments: Part Three, 8*

"If you are depending upon a virtualized environment, it is crucial to employ a continuous availability strategy in order to ensure uptime and guarantee that virtual machines can continue to function. Many industry experts recommend using a fault-tolerant solution in a virtualized environment. Especially when the cost of downtime is considered, virtualized fault tolerance can bring continuous availability to a data center at a competitive cost and with no special administrative skills."

Dr. Bill Highleyman
Managing editor of *The Availability Digest*

Stratus Technologies
www.stratus.com