

The Value of Availability

June 2011

In an earlier issue, we reviewed *Blueprints for High Availability: Designing Resilient Distributed Systems*, a classic book on high availability.¹ The authors, Evan Marcus and Hal Stern, have since published a second edition.² Their new book focuses less on cluster technology (their core experience) and extends itself to the broader issues of high availability.

We will not review the entire book in this article. Rather, we pass on the authors' insights into the financial justification for highly available systems as presented in Chapter 3, The Value of Availability.

What is High Availability?

Unfortunately, there is no commonly accepted definition for the term "high availability." System vendors have molded the term to suit their own purposes. Almost every system in the marketplace today touts its high-availability virtues.

The problem lies with both terms – "high" and "availability." Just how is *availability* defined? Does it mean that the system hardware and software are up and running? Does it mean that the users are able to use the system for its intended purpose? Is the system available if only a subset of users is down?

The term *high* is equally ambiguous. Does it mean that the system will never fail? Does it mean that the system is down for only seconds or minutes following a system fault? Does it mean that the system is down for only seconds or minutes a year? Does it mean that the system is operational only when it is needed?

The authors suggest that a system can be considered highly available if it is available enough of the time to generate the return for which it was implemented in the first place. High availability requires that a system be protected against all possible events. But the accurate prediction of these events is impossible. Therefore, *high availability* is considered a design goal. It should be clear to the system designers what the availability requirements for the system are. If the system is designed to meet those requirements, it can be considered to be highly available.

The authors propose a definition that seems to meet with reality:

"High availability is a level of system availability implied by a design that is expected to meet or exceed the business requirements for which the system was implemented."

¹ *Blueprints for High Availability*, *Availability Digest*, May 2007.

http://www.availabilitydigest.com/private/0205/blueprints_for_ha.pdf

² *Blueprints for High Availability: Second Edition*, Wiley Publishing, 2003.

In other words, it is a business decision; and it is a matter of *expectation* rather than of future experience.

The Costs of Downtime

There are two categories of downtime cost – direct and indirect.

Direct Costs

Direct costs are those that can be measured by the bean counters in the company or its auditors. They include:

- *Lost user productivity.* This includes not only the time lost (and paid for) because the system users are sitting around with nothing to do but also the cost of overtime as users recover the lost work.
- *Opportunity losses.* An opportunity loss is the loss of a chance to do something regardless of the outcome. Such losses can be lost sales experienced by a retailer, lost commissions on trades suffered by a brokerage firm, or fraudulent credit-card activity if transactions cannot be verified.
- *Regulatory and litigation costs.* These costs include regulatory fines due to violating governmental regulations as well as the results of legal actions by third parties seeking compensation for losses due to a system outage.³

Various industry analysts have made passes at quantifying direct costs of downtime for various industries. The authors report the following findings:

Average Downtime Cost per Hour (U.S. dollars)

Brokerage	\$6.5 million	Financial	\$1.5 million	Chemicals	\$704,000
Energy	\$2.8 million	Manufacturing	\$1.6 million	Health care	\$636,000
Credit card	\$2.6 million	Retail	\$1.1 million	Media	\$340,000
Telecom	\$2.0 million	Pharmaceutical	\$1.0 million	Airlines	\$90,000

Sources: Network Computing, the Meta Group, Contingency Planning Research.

Though these numbers are average across the industry and do not reflect the costs of any one company, they clearly show that the financial cost of downtime can be significant.

Indirect Costs

Indirect costs are harder to quantify but can have a greater impact and can extend over a longer period of time. Examples of indirect costs are:

- *Customer satisfaction.* If a customer cannot immediately place an order for an item, he is likely to go elsewhere. He may or may not return as a repeat customer.
- *Bad publicity.* The press loves to report bad news. Furthermore, it focuses on sensationalism and ignores the technical details.

³ We added this one.

- *Stock prices.* A common consequence of bad press is a run on a company's stock, devaluing it and reducing the company's market capitalization. This can reverberate throughout the company, reaching the company's C-level executives and the board of directors.
- *Legal liability.* Those financially impacted by an outage can take legal action for remuneration. In some cases, top executives can be held liable for gross negligence.
- *Employee morale.* Bad press, a stock drop, and legal action can cause key employees to leave the company.
- *External reputation.* If the word gets out that key employees are leaving, the reputation of the company can be damaged. This can lead to a vicious cycle of worsening fortunes for the company.

Is High Availability Worth It?

Downtime costs money. Protecting against downtime costs money. How much downtime protection is warranted?

It is difficult to evaluate the tradeoff for the indirect costs of downtime. This must be a management evaluation. However, determining the balance between direct costs of downtime and the cost of availability can be quantified.

Simply put, given a risk, there is a downtime cost associated with that risk over the life of the system. The cost savings to the company is the difference between the risk cost without availability and the risk cost with availability:

$$\begin{aligned} \text{risk} &= \text{downtime cost over the life of the system} \\ \text{cost savings} &= (\text{risk without availability}) - (\text{risk with availability}) \end{aligned}$$

The return of an investment in high availability, its ROI, is

$$\text{ROI} = (\text{cost savings}) / (\text{cost of availability})$$

For instance, if an availability solution costing \$100,000 results in a savings of \$300,000, the ROI is $\$300,000 / \$100,000 = 3.0$, or 200%.⁴ If the cost of downtime is \$50,000 per year, the investment will be returned in two years ($\$100,000 / \$50,000$).

The calculation of risk involves four factors:

- Likelihood (L)* – The number of failure events that will happen over the life of the system.
- Duration (D)* – The length of time that the outage event will last.
- Impact (I)* – The percentage of the user community that the outage will affect.
- Cost (C)* – The cost of downtime per unit time.

Risk is the product of these factors:

$$\text{Risk} = \text{Likelihood} \times \text{Duration} \times \text{Impact} \times \text{Cost}.$$

For instance, if an outage will occur ten times over the life of the system, if it will last two hours, if it will affect 100% of the users, and if downtime costs \$10,000 per hour, then the risk is

⁴ The \$300,000 savings covers the investment of \$100,000, leaving an additional \$200,000. Thus, the investment is paid off, and a benefit of \$200,000 is realized. This is 200% of the original investment of \$100,000.

$$\text{Risk} = 10 \times 2 \times 1 \times \$10,000 = \$200,000$$

A complexity not considered above is that the cost of downtime may not be constant. A brief outage of a minute or so may not have any cost impact. An outage of an hour may have a measureable impact. An outage of a day may have a significant impact. An outage of a week may put the company out of business. An outage at 1 AM may have a lesser cost impact than an outage at noontime. An outage in August may have a lesser cost impact than an outage just before the Christmas holidays. The cost of downtime has to be some justifiable (and hopefully conservative) average over all of the downtime possibilities.

An Example

The authors give an illuminating example, which is summarized here. A company that is considering expanding its single system into a two-node clustered system with a lifetime of five years. The company's cost of downtime is \$75 per minute, or \$4,500 per hour. The company's cost of moving to high availability is estimated as follows:

Second server	\$42,000
Licenses for clustering software	12,100
Extra networking gear	4,500
Five years of vendor support for the clustering software	15,660
Cluster training for two staff members	6,000
Cluster software implementation services	12,000
File-system software for the second node	6,000
Five years file-system support for the second node	<u>6,900</u>
	\$105,160

The company considered several downtime scenarios as detailed in the following table:

Failure	Risk Without Availability (minutes)				Risk With Availability (minutes)			
	L	D (min)	I	Risk (min)	L	D (min)	I	Risk (min)
Crash	10	60	1.00	600	10	5	1.00	50
Crash (off hours)	10	120	0.75	900	10	5	0.75	37.5
Scheduled reboot	60	30	0.50	900	60	5	0.50	150
Hardware	2	1,440	1.00	2,880	2	5	1.00	10
Network	2	240	1.00	480	2	4.5	1.00	9
Application	20	60	1.00	1,200	20	3	1.00	60
Sched. maintenance	20	240	0.50	2,400	20	5	0.50	50
Failover testing	0	0	0.00	0	20	5	0.50	50
Total effect (minutes)				9,360				416.5

Thus, the risk without availability is 9,360 minutes. The risk with availability is 416.5 minutes. At \$75 per minute, the risk cost without availability over the lifetime of the system is \$702,000. The risk cost with availability is \$31,238. The savings in downtime costs provided by the high availability cluster solution is \$702,000 – \$31,238 = \$670,762.

At a cost of \$105,160, The ROI of the availability solution is \$670,762 / \$105,160 = 6.38, or 538%. The time to pay back this five-year investment is \$105,160 / \$670,762 = .157 of five years, or 286 days. Not a bad investment!

What If You Don't Know the Cost of Downtime?

The cost of downtime is not always possible to reliably estimate. In this case, a different approach can be taken if the downtime minutes with and without availability can be estimated, as in the previous table. The trick is to use the above equations to calculate the amount of downtime that will result in zero cost savings, given a particular availability solution. A business decision can then be made as to whether this is a reasonable solution.

From the above analysis, the total cost savings given a particular application is

$$\text{total cost savings} = (\text{downtime savings}) \times (\text{downtime cost}) - (\text{availability cost})$$

where

$$\text{downtime savings} = (\text{downtime without availability}) - (\text{downtime with availability})$$

Note that *downtime savings* is expressed in time and that *total cost savings* is expressed in currency.

We want to find the downtime cost that will yield zero *total cost savings*:

$$0 = (\text{downtime savings}) \times (\text{downtime cost}) - (\text{availability cost})$$

$$\text{downtime cost} = (\text{availability cost}) / (\text{downtime savings})$$

This equation gives us the downtime cost at which we will realize no savings. If we think that our downtime cost is greater than this, the availability solution should be considered further. If we think that our downtime cost is less than this, we should proceed no further.

In our previous example, the downtime cost that produces zero savings is

$$\text{downtime cost} = \$105,160 / [(9,360 - 416.5) \text{ minutes}] = \$11.76/\text{minute}$$

or \$706 per hour. In all likelihood, our example company would have considered this a good investment without knowing exactly what its downtime cost was.

Summary

High availability is a business decision. Whether to adopt a particular high-availability solution depends upon many factors. From a cost viewpoint, it is a comparison between the cost of the solution and the cost of the downtime that it will save. Factored into this must be other considerations with respect to indirect costs, such as customer satisfaction and company reputation.

In this article, we have summarized the techniques put forth by Marcus and Stern to evaluate the financial savings and return on investment that an availability solution might bring to a company.