Realtime publishers

The Essentials Series:
Making High Availability Pay For Itself

High Availability's Spectrum of Cost and Capabilities

sponsored by

MARATHON
Run to Infinity

by Ron Barrett

Copyright Statement

© 2009 Realtime Publishers. All rights reserved. This site contains materials that have been created, developed, or commissioned by, and published with the permission of, Realtime Publishers (the "Materials") and this site and any such Materials are protected by international copyright and trademark laws.

THE MATERIALS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. The Materials are subject to change without notice and do not represent a commitment on the part of Realtime Publishers or its web site sponsors. In no event shall Realtime Publishers or its web site sponsors be held liable for technical or editorial errors or omissions contained in the Materials, including without limitation, for any direct, indirect, incidental, special, exemplary or consequential damages whatsoever resulting from the use of any information contained in the Materials.

The Materials (including but not limited to the text, images, audio, and/or video) may not be copied, reproduced, republished, uploaded, posted, transmitted, or distributed in any way, in whole or in part, except that one copy may be downloaded for your personal, non-commercial use on a single computer. In connection with such use, you may not modify or obscure any copyright or other proprietary notice.

The Materials may contain trademarks, services marks and logos that are the property of third parties. You are not permitted to use these trademarks, services marks or logos without prior written consent of such third parties.

Realtime Publishers and the Realtime Publishers logo are registered in the US Patent & Trademark Office. All other product or service names are the property of their respective owners.

If you have any questions about these terms, or if you would like information about licensing materials from Realtime Publishers, please contact us via e-mail at info@realtimepublishers.com.

High Availability's Spectrum of Cost and Capabilities

One of the most difficult achievements in technology is gaining the resolve to spend on the possibilities of what if. This is understandable especially in lean economic times such as we are now experiencing. However, the idea of losing data, services, or the ability to access an application that is usually readily available highlights the need to keep these IT services up and running at all costs (sort of). But what is high availability (HA) and how do you determine whether you have a need for it in your organization?

HA is sometimes mistakenly defined as a means of *recovering* from an unexpected loss or outage within your organization. Although recovery plays a part in HA technologies, HA is more about making sure applications, services, and data are always available to end users at all times regardless of what is taking place behind the scenes. Advances in technology have improved the way we work and made us an 'always on' society, which naturally means that the technology running this society must itself be always on.

It was not long ago that when we spoke about HA, it was in the context of wants. Businesses wanted to be able to have mission-critical applications and data available with almost no downtime. And there are some industries that have been working in this realm of HA for a long time: The need for consumers to access banking, surf the Internet, shop, or execute a stock trade any time of the day or night means that for these companies HA has become mission critical in and of itself. Lately, HA has moved beyond these specific industries and has become more widely adopted in what might seem to be non-traditional businesses. What trends have moved HA out of the nearly unattainable and into the mainstream?

Trends that Are Transforming HA

The need to be available all the time has transformed over the years; it is no longer specific industries that need to think in terms of HA. We have gone from the 9×5 local or regional businesses to a 24×7 global marketplace. Sectors that could not touch a large marketplace are suddenly able to market, produce, and sell anywhere in the world. And that paradigm shift has changed how we now think, work, and purchase goods and services. The perfect storm of communications, mobility, and affordability has transformed HA from a want to a need.

As these trends have changed, so has the consumers' view of what is an acceptable level of service. It has also changed the business' view of meeting the expectations of their clients. This naturally has driven demand for technology to meet the needs of both. These trends have not only changed the view of HA but also the composition of it. In the past, we thought of HA as an all-or-nothing technology. The complexity and expense of implementing an HA solution meant that this was only a solution for companies with deep pockets. And even then, it meant they could apply HA only to 'mission-critical' applications.

The technology to meet the needs of business has become both less complex and much more affordable. This has given HA a broader appeal and now has all sorts of organizations looking at how they can be both more competitive and more responsive to the needs of their clients and even their employees.

As business has become a 24×7 operation, the workforce behind that business has also changed. The need to access email, documents, data, and so on means that back-end systems need to be just as available as the front-end–facing systems for an organization. For some businesses, even the normally-accepted downtime for backup and maintenance is no longer an option, and this is where HA solutions shine.

So here we are at the crossroads of want and need again but with a twist: organizations that **want** to meet their business objectives **need** to have some kind of viable HA solution. Thankfully, meeting that need is no longer about available or not available but rather about levels of availability. What needs to be available? What is acceptable downtime (if any) for a particular application, service, or data set? And what solutions can help you meet your business objectives for HA? To find the answers to these questions, we need a way to measure the cost of the solution versus the cost of downtime—or what we will refer to as the *availability scale.*

Measures to Determine HA Solution Needs

When considering which services, data, and applications should be part of an HA solution, you need to determine a few key factors. Which services, data, and applications do I need to have available? And what is the cost to me if they are not available? Although they may seem over simplified, these two questions are at the core of all the others that you will ask yourself. So if you are going to find out what needs to be available and what the possible loss is for any downtime to that service, data, or application, you need to have a means of measuring the two factors. You also need a baseline to measure against. The baseline in this case should be a predetermined service level agreement (SLA) based on Recovery Time Objectives (RTOs) and Recovery Point Objectives (RPOs):

- SLA—The common understanding about services, priorities, responsibilities, and guarantees of services. These include both a target level of service as well as a minimum acceptable level of service. SLAs often include a promise of percentage of uptime (that is, 99.9999%). This is the standard you look to achieve or exceed in creating an HA solution.
- RTO—The duration and service level that needs to be restored after a disruption to avoid negative consequences on the business.
- RPO—The point in time in which services and data must be restored; this is also defined as the acceptable amount of lost data.

Once you have established your baseline SLAs, you then use them to compare against the actual performance of your systems. A set of metrics that work well for measuring against the SLA are the ITILv.3 Availability Management key performance indicators (KPIs).

The **ITIL v.3 Availability Management KPIs** provide metrics for the following:

- Service availability—Measures the availability of services compared with the established SLAs
- Number of IT service disruptions—Measures the total number of service, data, or application interruptions in an organization; these can be further segmented into root causes (that is, hardware failure, customer action, security issue, and so on)
- Duration of IT service disruptions—Measures the accumulative duration of service, data, or application disruptions in an organization

Note

Opinions vary on whether to include both planned and unplanned disruptions into the calculations of availability. Some argue that excluding planned downtime provides higher-than-actual uptime values. When planning for an HA solution, the right answer probably depends on what you are discussing. An application that is being used 24×7 should have planned and unplanned downtime calculated. However, there may be services that do not require 24×7 access; for these, the exclusion of planned downtime would be acceptable when determining an HA solution.

- Availability monitoring—Determines a percentage of all services and infrastructure components that have or will be monitored
- Availability measures—Implements measures with the objective of increased availability

To illustrate the process, let's consider Application A and Service B, which both have an SLA of 100% uptime. Using the current in-place solution (even if that solution is no solution), apply these metrics to gain a baseline of availability such as the one shown in Figure 1.

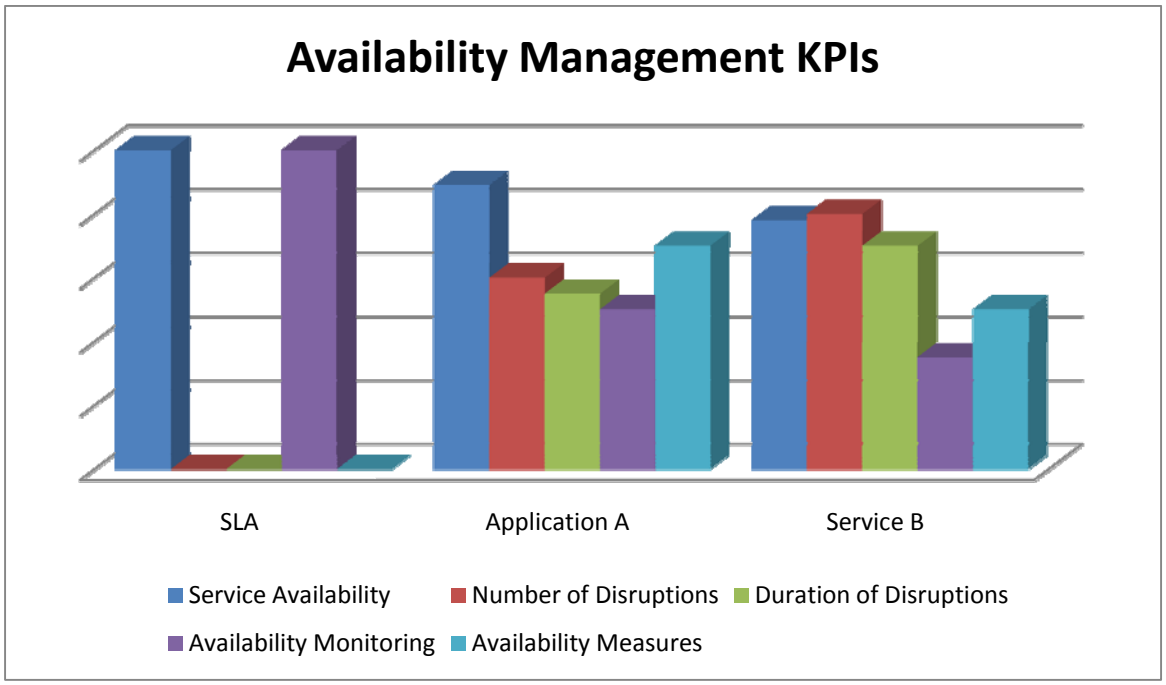

Figure 1: ITIL Availability Management KPIs as measured against existing SLAs.

As you can see from Figure 1, the baseline SLA measurements of 100% uptime (left) are significantly different from the reality of the actual measurements of Application A and Service B. Once you have metrics to determine the actual availability of your systems, you can measure them against the RTO and RPO in your SLAs to get a better idea of where you need HA protection.

However, there is more to consider than the purely technical. HA is a business process as much as it is a set of technologies. So, naturally, the cost impact of these metrics weighs heavily on the decision of what solution to implement; this is where ROI comes into play.

Calculating ROI

To calculate the ROI, you need to first determine the fiscal impact of disruptions. To do so, you need to consider direct/indirect costs such as:

- Loss of business operations—Revenues from client transactions, orders, production, shipping, and so on
- Lost employee productivity—Employee salaries multiplied by lost hours
- Business reputation—Value of lost business due to the disruption
- Operation costs—Loss of revenue due to penalties, compensations, lost interest, discounts, stock price losses, and so on

Once you understand the fiscal impact, you can take the metrics for current service level, look at your RTO and RPO, then consider the HA solutions that will help you achieve the levels of availability for that IT service. Taking the cost of the solution and subtracting it from the fiscal impact report will provide a ROI for the HA solution.

What you need to remember is that all applications, services, and data that are considered for HA need to be put through these measurements to determine a baseline. That baseline along with several other factors will help you begin to create an availability scale.

The Availability Scale

Rather than thinking of HA in linear terms, you need to forego the all-or-nothing approach and weigh potential solutions in terms of need, technology, and cost. HA is not about trying to fit all your business needs into your solution; it is about your solution fitting your business needs. After you weigh the impact of an application, service, or data disruption, you can apply that weight to your availability scale. The availability scale uses a weighted number (1 to 10) for each IT service, desired level of availability, and cost.

We usually think of availability as having five levels:

- Unprotected—IT services that are not part of any HA solution; these may use some kind of redundancy, but by and large they are unprotected
- Reliable—Applications and services or hot-swappable components that can eventually be recovered; they are not business impacting
- Recoverable—Redundant infrastructure components that have some automatic recoverability built in; some downtime is acceptable
- Highly Available—Mission-critical IT services that drive the business; demand for availability is high, so these need high redundancy at the hardware, software, and communications levels
- Always (continuously) Available—Few if any of these IT services exist in most organizations, as they require absolute zero acceptable downtime; these are triple or even quadruple redundant

The idea behind this scale is to simplify the decision-making process. IT services with a weight of 10, desired availability of 8, and a cost of 8 make sense to implement at the desired level. IT services with a weight of 5, a desired availability of 9, and a cost of 8 might need a revaluation as to the desired availability level. Does the cost and weight of the service justify that level of availability? The scale helps to take a balanced approach to planning for HA solutions. Table 1 provides a sample of the availability scale. A simple way to make sure the three elements are in sync is to use the following formula:

*IT Svc. Weight * Desired Availability / Cost = Availability Factor*

	IT Service Weight	Desired Availability	Cost	Availability Factor
Email	10	8	8	10
Web Site	5	9	8	6

Table 1: A sample availability scale.

In the first instance, the email application has a service weight of 10 and a desired availability of 8. Multiplying the two factors and dividing them by the cost gives you an availability factor of 10. In contrast, a Web site with a service weight of 5 and a desired availability of 9, once divided by the cost of 8, gives you an availability factor of 6 (rounded up). You can then take this information and match an availability factor with an HA solution within the availability scale.

Availability Factor	Availability Level	HA Solutions
Factor 0 to 1	Unprotected	No Availability Solution
Factor 2 to 3	Reliable	Hot Swappable Components Snapshots
Factor 4 to 6	Recoverable	Replication & Failover
Factor 7 to 9	Highly Available	Virtualization, HA Clustering, Synchronized Systems
Factor 10	Continuously Available	System Fault Tolerance

Table 2: Comparing availability factors to HA solutions.

Coming Up Next…

The second article of this series will look at how to match availability wants to business needs and will use the availability scale to determine the best HA solution to meet the common ground of these two factors.