Realtime
publishers

The Essentials Series:
Making High Availability Pay For Itself

Matching High Availability Technology with Business Needs

sponsored by

MARATHON
Run to Infinity

by Ron Barrett

Copyright Statement

Matching High-Availability Technology with Business Needs

As we discussed in Article 1, there are a few ways to tackle the issue of high availability (HA). Finding a solution—or more appropriately a *set* of solutions—is about matching the available technologies to the needs of the business. The first step was to look at the applications, services, and data that need to be available. Then we used the metrics to determine the current reliability of your environment and the fiscal impact of a disruption.

After using the availability scale to provide a weighted metric for the IT service, desired availability, and cost, you need to next look at availability versus affordability. This is where you can discern whether the HA solution fits both the business needs and budget.

Availability vs. Affordability

Ideally, we would like to see all our IT services run at the level of continuous availability. After all, that is what end users and clients expect. The reality is that we are not likely to fit our needs into a single HA solution. Rather a set of solutions aligns better with how we actually do business. In any typical organization, there exists a hierarchy of job duties; not all are equally important, although all are necessary to get the job done. The same is true of IT services—all of them are necessary to keep the whole organization running but not all carry the same importance in reaching business objectives.

Keeping the balance between the two is a matter of picking the best availability to match affordability. This is not just about having the dollars to throw at a solution. It makes no business sense to make an internal Web server that has no direct business impact highly available. The nature of the service would in most cases allow for a recoverable solution with some downtime. Likewise, it would make no business sense to take a core business application and classify it as recoverable when in fact it should be continuously available. Continuous availability is an expense, so in some cases, the business just cannot support that expense. The next logical step is to opt for HA; you should not accept a recoverable system to save a few dollars. If in fact this core application should fail, the ROI will be greater than the cost of the HA solution. Figure 1 provides a matrix of HA solutions, availability levels, and relative affordability. It also classifies examples of IT services that fall into those levels of availability.

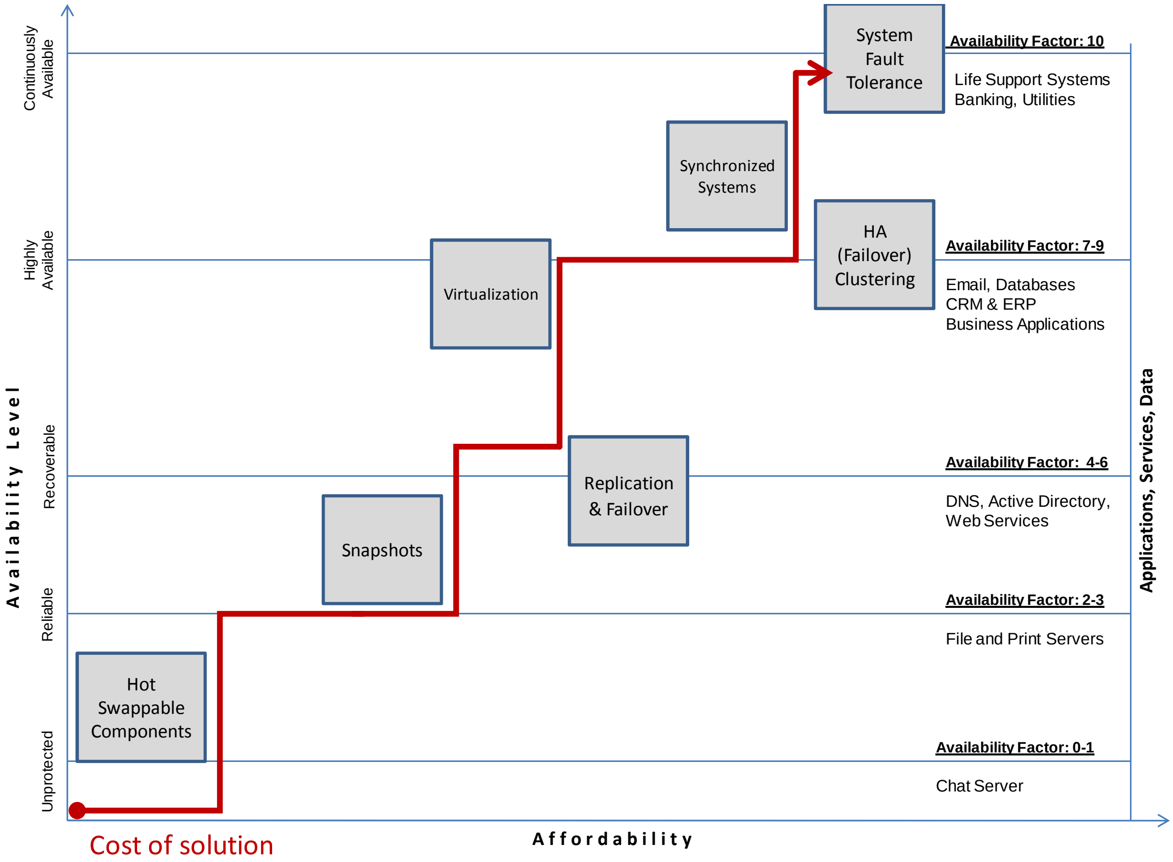

Figure 1: Example matrixes highlighting availability vs. affordability.

Deciding which solution works best is important in making sure that the HA solution pays for itself. Remember that in many organizations, an HA solution is not sold to decision makers based on the technology itself. It is typically sold based on meeting business objectives at a particular price.

Matching Technology to Availability Needs

In line with choosing the right availability and affordability is understanding what each HA technology brings to the table. Let's review the levels of availability and define how they meet (or fail to do so) the business objectives:

- Unprotected—These applications, services, and data do not have an impact on business objectives. A disruption means a complete loss of usability, and the Recovery Time Objective (RTO) is typically days or unrecoverable.
- Reliable—These applications, services, and data do not have an impact on business objectives. A disruption means a complete loss of usability, and the RTO is typically days.

- Recoverable—These applications, services, and data typically are infrastructure components. Recovery is sometimes automatic. A disruption means some impact at the user level while the systems replicate and failover to redundant systems. The RTO is typically 1 day.
- Highly Available—These applications, services, and data are core or mission-critical applications, services, and data. Disruptions are minimal and typical brief. Due to the nature of these IT services, the RTO is typically 1 to 4 hours.
- Always (continuously) Available—These applications, services, and data are highly critical and include utilities, trading systems, banking, telecom systems, and so on. Disruptions are not an option, and due to the nature of these IT services, the RTO is 0.

Understanding the levels of availability and the potential impact on business objectives helps you to better relate how HA pays for itself. It is important to 'weigh' the service against the desired availability level, then measure the potential cost of the solution against both. Using the availability scale, you then create a mix and match approach. This will keep critical systems up while reducing overall downtime. Thus, you can meet or exceed the SLA you have put into place. It also means you can control costs.

To make a determination of which HA solution matches your needs, you need to consider the various technologies behind HA. Once you determine the availability need of an application, service, or data set, you must understand what technologies will help you reach those objectives.

HA Technologies

We will wrap up this article by defining the various technologies that can be used to help you meet your availability goals. These definitions can be useful in presenting an HA solution.

Hot-Swappable Components

Hot-swappable components provide the ability to change disk drives, controllers, power supplies, and so on without turning off the system. This allows the system to be available without much downtime (certain RAID configurations may require a rebuild, which can degrade service levels although not necessarily service itself). This solution offers no protection from sudden interruptions or disruptions due to software factors. This option is relatively inexpensive to implement because the cost of hardware is relatively low.

Snapshots

Snapshots provide a point-in-time copy of data and files that can then be backed up to various media. Using snapshots to take a copy of the data at timed intervals and moving snapshots to external storage such as a NAS/SAN will make the data and files recoverable if there is a service disruption. This solution again may provide a reliable or recoverable system but not a highly-available one because this method requires restoring from a snapshot in the case of a disruption or failure. This option can be a fairly inexpensive solution and requires usually only a licensing fee from most SAN/NAS vendors.

Replication and Failover

Replication and failover solutions provide a complete replica of the primary system and replicate both the hardware and software elements of the systems. This solution usually can detect imminent disruptions and move the workload over to a failover system to ensure availability. These systems are usually failed over automatically with minimal impact on data and users. These solutions are a bit more costly compared with the previously mentioned options because they require hardware to replicate and failover to should a disruption take place.

HA (Failover) Clusters

HA clusters not only replicate systems and provide automatic failover but also go an extra step in attempting to eliminate any single point of failure (SPOF). These solutions usually will detect a hardware or software fault and start the application on another node immediately. Clusters require at least two nodes (computers) but can consist of dozens of nodes. This setup provides a very high level of availability but comes with a very high cost because redundancy is required not only at the system levels but also within the entire infrastructure. However, in situations where the impact of downtime on the business is significant, the higher costs can be easily justified.

Synchronized Systems

Synchronized systems combine two Windows servers and present them to the world as a single server. The servers are monitored and tested by the synchronization software, and if one server fails, the other is available as an exact duplicate to continue providing the applications to support the business. The two servers are completely synchronized at the OS, application, and data levels by ensuring that changes happen to both servers simultaneously. The advantage of this approach is that it is a single solution that can be implemented without a high degree of skill with the individual components that make up systems.

Virtualization

This newest technology has the ability to create a highly-available solution. Virtual machines can be used to create a failover cluster without the need of a physical machine to place that node onto. However, it is important not to underestimate the hardware needs for using virtualization. You must also account for the SPOF: While an OS failure will have no impact, placing the nodes on the same hardware once again presents a SPOF for the cluster. Costs can be lowered by using virtualization, but the solution needs to be planned so as not to inadvertently create a less-available solution.

Coming Up Next…

We'll next look at how you can use the metrics you have gathered to gain approval for your HA solution. We will look at how to explain the business impact of those metrics as well as, how to translate the technology into business reports. In addition, we will explore where HA has real business value overall.