# Calculating Availability – Heterogeneous Systems Part 1

March 2008

In all of our availability analyses to date, we have assumed that the nodes in a system are identical. In particular, if there are *n* nodes in a redundant system, each with an availability of *a*, and if the system can withstand the loss of *s* nodes, then the availability *A* of the system is given by

$$A = 1 - \frac{n(n-1)}{2}(1-a)^{s+1} \tag{1}$$

But what if the nodal availabilities are not the same? What if one node is in a safe area, and the other is in Hurricane Alley in Florida? The Florida node will have an availability less than the other node because it stands to be destroyed by a hurricane at some time. What, then, is the availability of the redundant system?

In this article, we review some simple probability relationships necessary to analyzing this situation and those like it. In our next article, we apply these relationships to heterogeneous systems.

## Probability 101

When it comes to availability, we are often concerned about binary states. For instance, we are concerned about whether the state of a system is up (operational) or down (failed). This is a binary state – the statement that the system is up is either true or false.

The value (true or false) of a binary state can be specified as a Boolean function with operators AND, OR, and NOT. For instance, it may be that a certain state is true if *x* AND *y* are true OR if *z* is NOT true. Knowing the probabilities of *x*, *y*, and *z*, what is the probability of the system being in that state?

For instance, in a two-node redundant system with Nodes 1 and 2, the system is up if Node 1 OR Node 2 is up. Similarly, the system is up if Nodes 1 AND 2 are NOT down.

Let p(*k*) be the probability that *k* is true. These Boolean functions transform into the following probability equations.

***AND***

The AND operator implies multiplication. The probability that *x* AND *y* are true is

$$p(x \text{ AND } y) = p(x)p(y) \tag{2}$$



© 2008 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com

For instance, consider rolling a pair of dice. The probability of rolling a 2 is 1/6. The probability of rolling a four is 1/6. The probability of rolling a 2 on the first roll AND rolling a 4 on the second roll is

$$p(2\ \text{AND}\ 4)=p(2)p(4)=(1/6)(1/6)=1/36$$

The chance of rolling a 2 followed by a 4 is one time in 36 tries.

***OR***

The OR operator implies addition. The probability that *x* OR *y* is true is

$$p(x\ \text{OR}\ y) = p(x) + p(y) \qquad (3)$$

*x* and *y* are taken from a set of events, any one of which can be true. Therefore, p(x)+p(y) is necessarily less that one.

This can cause confusion. For instance, consider our pair of dice. The probability of rolling a number greater than 1 is 5/6. Therefore, the probability of rolling a number greater than 1 on the first throw OR rolling a number greater than 1 on the second throw is 5/6 + 5/6, or 10/6, right? Wrong. This is a probability that is greater than one, which makes no sense.

This demonstrates the need to set up probability equations properly. What we are really trying to solve is the probability that the first roll will be a number greater than 1AND the second roll will be a 1, OR the first roll will be a 1 AND the second roll will be greater than 1. Since the probability of rolling a 1 is 1/6, this results in a probability of

p(rolling a number greater than 1 on the first roll OR on the second roll)
= (5/6)(1/6) + (1/6)(5/6) = 10/36

or ten tries out of 36.

This begs another question. Did we mean rolling a number greater than 1 on the first roll OR on the second roll but NOT on both? Or did we mean rolling a number greater than 1 on either the first roll OR the second roll OR on both? The above equation satisfies the “NOT both” condition. If we are interested in the case that either or both of the rolls will be a number greater than 1, we are trying to calculate the probability that

- the first roll will be a number greater than 1 AND the second roll will be a 1, OR
- the first roll will be a 1 AND the second roll will be greater than 1, OR
- the first roll will be greater than 1 AND the second roll will be greater than 1.

In this case,

p(rolling a number greater than 1 only on the first roll OR only on the second roll OR on both)
= (5/6)(1/6)+(1/6)(5/6)+(5/6)(5/6) = 35/36.

Note that the only case we have left out is the probability that both rolls will be a 1. This probability is (1/6)(1/6) = 1/36. Thus, the probability of any occurrence is one, as we would expect.

***NOT***

The probability that event *x* is not true is


© 2008 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com

$$p(x\ \text{NOT true})=p(\text{NOT } x) = 1 - p(x) \quad (4)$$

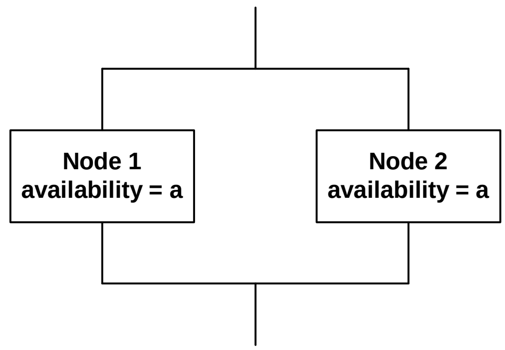
This is obvious since event A is either true or not true. Therefore, the probability that A is true OR the probability that A is NOT true is one. That is, p(A) + p(NOT A) = 1.

For instance, if the probability of rolling a 1 is 1/6, the probability of NOT rolling a 1 is (1–1/6) = 5/6.

We can use this to solve one of the previous problems in another way. The probability of rolling a number greater than 1 on either roll or on both rolls is the probability that both rolls will NOT result in a 1. The probability that both rolls will be a 1 is (1/6)(1/6) = 1/36. Therefore, the probability that both rolls will not result in a 1 is (1–1/36) = 35/36. This is the probability that a number greater than 1 will be rolled on either or both attempts, as we concluded above.

## Application to System Availability

Let us use these principles to review the simple case of a two-node active/active system, which we have analyzed exhaustively in previous articles. We take the case of an active/active system comprising two nodes, Node 1 and Node 2. Each node has an availability (the probability that it will be up) of *a*.




We can use either of two techniques to calculate the probability that the system will be up. This is the system's *availability*, *A*. These two techniques follow.

***The system will be up if both Nodes 1 AND 2 are NOT down.***

The probability that Node 1 will be down is the probability that it will NOT be up:

p(Node 1 is down) = p(Node 1 is NOT up) = (1-a)

Likewise,

p(Node 2 is down) = p(Node 2 is NOT up) = (1-a)

The probability that both Node 1 AND Node 2 are down is:

p(Node 1 AND Node 2 are down) = $(1\text{-}a)^2$

The availability, *A*, of the system is the probability that both Nodes 1 and 2 are NOT down:

A = p(both nodes are NOT down) = 1 – p(both nodes are down) = 1 - $(1\text{-}a)^2$ (5)

This is, of course, the availability equation for a two-node system with which we have been working [see Equation (1) for $n$ = 2 and $s$ = 1].

***The system will be up if either Node 1 or Node 2 is up.***

This can be expressed as the system will be up if both Nodes 1 AND 2 are up OR if Node 1 is up AND Node 2 is down OR if Node 1 is down AND Node 2 is up.

This is the same statement as above but in a different form. It leads to a different analysis.


© 2008 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com

The probability that Node 1 is up is *a*. The probability that Node 2 is up is also *a*. Thus,

$$p(\text{Node 1 is up AND Node 2 is up}) = a^2$$

The probability that Node 1 is down is the probability that it is NOT up and is (1-*a*). Likewise, the probability that Node 2 is down is (1-*a*). Thus,

$$p(\text{Node 1 is up AND Node 2 is down}) = a(1-a)$$

$$p(\text{Node 1 is down AND Node 2 is up} = (1-a)a$$

The availability of the system is the OR of these values, which implies summation. Thus,

$$A = a^2 + 2a(1-a) \qquad (6)$$

***Some Observations***

<u>The Value of Intuition</u>

Equation (6) looks strikingly different from Equation (5), yet they both represent system availability. How can this be? A little algebraic manipulation will show that they both reduce to *a*(2-*a*) and are therefore equivalent:

$$\text{Availability} = a(2-a) \qquad (7)$$

Why, then, have we been using the more complex expression of Equation (5) in our availability analyses rather than the simpler Equation (7)? The answer is that Equation (5) is more intuitive. It comes directly from the statement that the system will be up so long as both nodes are not down. The author, at least, cannot intuitively explain why Equation (7) is true.

It is from the simple reasoning behind Equation (5) that we have intuitively derived in our previous articles the effects of repair time, recovery time, restore time, failover faults, and many other factors that affect availability.

<u>Approximations</u>

In our analyses, we often eliminate terms to come up with reasonable approximations that are simpler and more intuitive. However, this can sometimes be a trap.

Consider Equation (6) above. In the cases which we typically analyze, subsystem availability, *a*, is often very close to one (.999 or higher). Therefore, (1–*a*) is very small. In Equation (6), we could then reasonably conclude that the second term, $2a(1-a)$ is very much smaller than the first term, $a^2$, and can be ignored. Thus, to a reasonable approximation, $A = a^2$.

However, since *a* is always less than one, we would then conclude from this approximation that adding a redundant node reduces system availability, which, of course, is not true. Sometimes, the minor terms have major importance; and we have to be careful how we simplify expressions into reasonable approximations. That is why we like to verify our intuitive approaches with formal proofs. The use of failure state diagrams is one way to do this (see our *Geek Corner* series on Failure State Diagrams).

## Summary

As shown above, we have multiple ways to look at the same problem and come up with the same answer in two different forms. This is generally true in probability analyses. There are many ways


© 2008 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com

to look at a problem. One way is often more intuitive than another, and that is the one we use. However, it is imperative that the probability equations be set up properly, and this is not always immediately obvious. Furthermore, we have to exercise caution when we attempt to simplify our results by using approximations.

In our next article, we will apply these rather elementary principles to the calculation of availability for systems with heterogeneous nodes. We will consider not only redundant systems that can stand the failure of one or more nodes, but also serial systems that will fail if any one component fails. An example of this sort of architecture is a tiered architecture in which the system will fail if the application layer fails or if the database layer fails.


© 2008 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com