the Availability Digest

Calculating Availability – Failover Faults

March, 2007

In our last article, we analyzed the effect of failover time on system availability. We found that failover times measured in minutes or more could have a profound effect on the availability of a system that was otherwise designed to be highly available. Only if failover times can be kept to a few seconds are extremely available systems such as active/active configurations relatively immune to failover time.

However, the effect of failover time on availability is only half the story. The problem is that failover doesn't always work. When it doesn't, the system often goes down and has to be recovered. Instead of a few seconds or minutes of failover time, there now may be hours of system recovery time.

In this article, we look at the effect of these failover faults on system availability.

Where We Left Off

In our analysis of failover faults, we evaluated redundant systems in which every subsystem is backed up by one or more subsystems. The system is provisioned with one or more spare subsystems. If all spare subsystems fail, the system is left open to failure should one more subsystem fail.

When a subsystem fails, a failover procedure is executed. The failover takes a time of MTFO (mean time for failover). During this interval, some or all of the system's services are unavailable to some or all of the users.

We assumed that a failed node has to be *repaired* and *recovered* before it can be returned to service. Following a total system failure, the system can be returned to service as soon as the first repaired node is available. Returning the system to service requires some additional *restoration* tasks.

Because of failover time, there are two components to the availability equation:

- The probability that the system will be down due to a multiple subsystem failure.
- The probability that the system will be down due to failover processing.

The resulting failure probability, *F*, is (for the case of a singly-spared system with parallel repair)

$$F \approx \frac{r/2 + R}{r/2} \frac{n(n-1)}{2}(1-a)^2 + \frac{MTFO}{r} n(1-a) \qquad (1)$$

where

F	is the probability of failure of the system.
a	is the availability of a subsystem.
n	is the number of subsystems in the system.
r	is the repair and recovery time for a subsystem. That is, it is the time required to return the subsystem to service. This term is often referred to as the mean time to repair (mtr) for the subsystem.
R	is the restore time of the system. It is the time that it takes to perform system-wide functions required to return the system to service once it has a full complement of subsystems. For instance, restoration functions may include database resynchronization and the reentry of transactions submitted during the system's downtime.
MTFO	is the time required to failover (the mean time for failover).

The first term in Equation (1) is the probability that the system will be down due to a multiple system failure. The second term is the probability that the system will be down during a failover process.

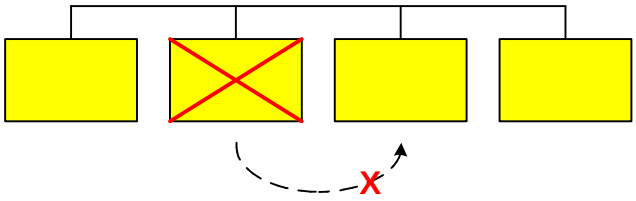
The system availability, *A*, is, of course, 1-*F*.

We showed by example that a typical cluster with a three-minute failover time might have its probability of failure increased by an order of magnitude but that a typical active/active system retains its extreme availability if failover times can be kept to seconds.

On the other end of the scale, failover time for a cold standby with a failover time of two hours (which is reasonably fast for a cold standby) completely dominates the availability relationship. The probability of a dual failure of both the active and standby systems is not an issue, though such a failure could lead to a long recovery time.

What is a Failover Fault?

The purpose of a failover is to direct an operating system component to take over the functions of a component that has just failed. This can be a very complex task and can take some time (seconds to hours).



Failover Fault

Not only are failover processes complex, but they must cover a myriad of failure scenarios, some of which cannot even be visualized during the design of the failover process. Furthermore, the failover process can be very difficult to test. A means must be provided to inject faults into the system, and even the best fault-injection facilities cannot reasonably inject all known possible faults, much less those that are unknown. Consequently, because of their complexity and the difficulty in testing them, failovers are subject to failure themselves.

In practice, failover procedures are substantially debugged in the field. This is a neverending process, leading to continual improvement but never perfection in the failover process.

When a failover attempt fails and takes down the system, this is known as a *failover fault*. It has been estimated that about 1% of failovers are unsuccessful.[1] In a three 9s system, this may not

[1] Estimates from a vendor of highly available systems.

be much of a factor. But if we are looking for availabilities measured in centuries, failover faults can indeed be a factor in system availability.

Failover faults are a consideration for fault-tolerant systems, for clusters, and for active/active systems. However, they do not generally apply to active/backup configurations because the failover process is really the bringing up of the backup system. If this fails, it is because of a dual system failure, not a failover fault.

Failure Probability with Failover Faults

By introducing the possibility of failover faults, the availability relationship given by Equation (1) must be extended. There are now three classes of system downtime that we must consider:

- The system may be down due to a multiple subsystem failure.
- The system may be down due to failover processing.
- The system may be down due to a failover fault.

The first two classes are covered by Equation (1). Let us determine the third term that will define the impact of failover faults on system availability.[2]

Let *p* be the probability that a failover attempt will end in failure. That is, given a single-subsystem failure, *p* is the probability of a failover fault.

The probability that a failover attempt will be made is the probability that there is the failure of a single subsystem. If a subsystem has an availability of *a*, the probability of a single-subsystem failure is (1-*a*). If there are *n* subsystems in the system, the probability that there will be a single-system failure and thus a failover attempt is *n*(1-*a*).

If *p* of these failover attempts fail, the probability of a failover fault is

$$\text{failover fault probability} = pn(1-a) \qquad (2)$$

The measured availability of the subsystem, *a*, is based on it being returned to service after it is repaired and recovered. This is our term *r* defined above. However, recovery from a failover fault does not require a subsystem repair and recovery. Rather, it requires a system restoration of *R*, as defined above. Therefore, to convert the above failover fault probability to the probability that the system is down,[3] it must be adjusted by the factor *R*/*r*:

$$\text{probability that system will be down due to a failover fault} = \frac{R}{r}pn(1-a) \qquad (3)$$

If *p* of all system faults are caused by failover faults, then (1-*p*) of all faults are caused by multiple system faults or failover times. Applying these observations, we have an expanded availability relation that covers all of the three cases in which we are interested:

$$F \approx (1-p)\left[\frac{r/2+R}{r/2}\frac{n(n-1)}{2}(1-a)^2 + \frac{MTFO}{r}n(1-a)\right] + p\frac{R}{r}n(1-a) \qquad (4)$$

---

[2] Failover faults are discussed in some detail in Chapter 5, The Facts of Life, *Breaking the Availability Barrier: Survivable Systems for Enterprise Computing*, by Dr. Bill Highleyman, Paul J. Holenstein, and Dr. Bruce Holenstein.

[3] This relationship is derived formally with Markov models in Appendix 3, Failover Fault Models, *Breaking the Availability Barrier: Survivable Systems for Enterprise Computing*, referenced above.

Since a node failure in an active/active system affects only 1/*n* of the users, Equation (4) becomes, for active/active systems,

$$F \approx (1-p)\left[\frac{r/2+R}{r/2}\frac{n(n-1)}{2}(1-a)^2+\frac{MTFO}{r}(1-a)\right]+p\frac{R}{r}(1-a) \text{ (active/active)} \qquad (5)$$

The first term of Equations (4) and (5) is the probability that the system will be down due to a dual-system failure. The second term is the probability that it will be down while it is undergoing a failover procedure. The third term is the probability that it will be down because of a failover fault. The sum of these probabilities is the probability that the system will be down.

Examples

To get a feel for the effect of failover faults, let us repeat the clustering example and the active/active example from our previous article, which evaluated the effects of failover time, and add a failover fault of 1% to these examples.

A Clustered System

We consider the case of a single-spared, four-node cluster, which is made up of nodes with availabilities of .999. The recovery time for a node is two hours, as is the system restore time. The failover time is three minutes, and failovers have a 1% chance of failing. Thus,

r	= 2 hours
R	= 2 hours
n	= 4
a	= .999
MTFO	= .05 hours
p	= .01

From Equation (4), the probabilities of failure for this case are:

Probability that the system is down due to a multiple node failure = 1.8×10^{-5}
Probability that the system is down during failover = 9.9×10^{-5}
Probability that the system is down due to a failover fault = 4×10^{-5}
Probability that the system is down = 1.6×10^{-4}

A 1% chance of a failover fault has added 34% to the system availability that would be achieved if there had been no failover faults.

An Active/Active System

In this example, we consider a two-node active/active system. The node repair and recovery time, *r*, and the system restore time, *R*, are the same as the clustered example above. The nodal availability is four 9s, and the failover time is 1 second. The failover fault probability is again 1%.

r	= 2 hours
R	= 2 hours
n	= 2
a	= .9999
MTFO	= 1 seconds = .00028 hours
p	= .01

From Equation (5),the results for this case are

Probability that the system is down due to a multiple node failure = 3×10^{-8}
Probability that the system is down during failover = 1.4×10^{-8}
Probability that the system is down due to a failover fault = 1×10^{-6}
Probability that the system is down = 1×10^{-6}

In this case, failover faults dominate the system availability, reducing it by almost two 9s. Clearly, as the availability of the nodes increases and as the failover time decreases, failover faults play an ever more dominant role in system availability.

A Limiting Case

We can get an insight into this phenomenon by considering a limiting case. Let us consider a two-node system with a restore time, *R*, that is half of the nodal repair and recovery time, *r*. That is, *R* = *r*/2. Assume that MTFO is small enough so that it can be ignored and that the probability of failover is small (*p* << 1). Then equation (4) reduces to

$$F \approx 2(1-a)^2 + p(1-a) = [2(1-a)][(1-a) + p/2] \quad (6)$$

The limiting case that we want to consider is the case in which the probability of a failover fault is very much greater than the probability of a node failure as it is in the active/active example above (.01 versus .0001). In this case, *p*/2 >> (1 – *a*); and Equation (6) further reduces to

$$F \approx 2(1-a)\frac{p}{2} = (1-a)p \quad (7)$$

This equation says that if a failover fault is much more likely than a node fault, then following a node failure the surviving system acts as if it has a probability of failure of *p* rather than of (1-*a*).

Thus, in high-availability systems, failover faults may cause a disproportionate decrease in system availability.

Summary

Failover faults can have a serious impact on system availability. For systems with modest availability, failover faults are not terribly serious so far as overall system availability is concerned. However, as the inherent reliability of a system improves, that is, as the nodes become more reliable and as failover time decreases, the impact of failover faults can increase dramatically.

In the limit, once one node in a single-spared system has failed, the system availability is determined by the probability of a failover fault rather than by the probability that a second node will fail.