



“The Availability Corner” Advice and Solutions for Enterprise Computing

**As Seen in *The Connection*, An ITUG Publication
November 2004 – September 2006**

About the Authors:

Dr. Bill Highleyman, Paul J. Holenstein, and Dr. Bruce Holenstein, have a combined experience of over 90 years in the implementation of fault-tolerant, highly available computing systems. This experience ranges from the early days of custom redundant systems to today's fault-tolerant offerings from HP (NonStop) and Stratus.

Series Topics:

Testing Your System Recovery Plan (09/06)
Is IBM's Parallel Sysplex a NonStop Competitor? (06/06)
Grid Computing (03/06)
The Net Present Value of Active/Active Systems (01/06)
TCO for Active/Active Systems (11/05)
Fault Tolerance vs. High Availability (09/05)
The Great Tape Backup Paradigm Shift (07/05)
The Language of Availability (05/05)
What Reliability Do We Really Need? (01/05)
Let's Measure System Reliability in Centuries (11/04)

Gravic, Inc.

Shadowbase Products Group
17 General Warren Blvd.
Malvern, PA 19355
610-647-6250
www.ShadowbaseSoftware.com

The Availability Corner

The Language of Availability

May/June, 2005

Dr. Bill Highleyman
Dr. Bruce Holenstein
Paul J. Holenstein

Every technology has its own buzz words, and the study of reliability is no exception. Though many of the terms that we will use in this series have obvious everyday interpretations, their precise meaning can subtly vary from person to person.

To get us all on the same wavelength, we present below several terms that we will use in this series and define what we really mean by them. They are arranged in an order that let's us build definitions upon prior definitions. First, some definitions that define reliability:

user	A <i>user</i> (or end user) is a person, an application, or a system that is requesting a service.
service	A <i>service</i> is a function provided to a user.
SLA	A <i>service level agreement (SLA)</i> is an agreement between the service provider and the users of that service defining what constitutes acceptable service (such as response time, capacity, availability, data loss following a failure, etc.).
available	A service is <i>available</i> if it is functioning correctly and with acceptable performance, often as defined by an SLA.
unavailable	A service is <i>unavailable</i> if it is not available.
up	A service is <i>up</i> if it is available.
down	A service is <i>down</i> if it is unavailable.
failure	A <i>failure</i> is any event that causes a service to be down.
MTBF	<i>MTBF (mean time before failure)</i> is the average amount of time that a service will be up. MTBF is also sometimes used as mean time <i>between</i> failures which is the average amount of time from one failure to the next.
MTR	<i>MTR</i> can be a confusing term. It can mean <i>mean time to repair</i> which is the average amount of time required to repair a service. It can mean

mean time to restore which is the average amount of time to repair a service and return it to service. It can mean *mean time to recover* which is the average time that it takes to restore a service and to recover its state (typically its database). Its usage must be defined in the context of its use.

availability	<i>Availability</i> is the probability, or proportion of time, that a service is up.
failure probability	<i>Failure probability</i> is the probability, or proportion of time, that a system is down.
reliability	<i>Reliability</i> is an inverse function of failure probability. If system A is down 1% of the time and system B is down 10% of the time, then system A is ten times more <i>reliable</i> than system B.
nines	<i>Nines</i> is a logarithmic measure of availability expressed as a number of nines. For instance, a service with three nines availability will be up 99.9% of the time.
uptime	<i>Uptime</i> is a synonym for <i>availability</i> .
downtime	<i>Downtime</i> is a synonym for <i>failure probability</i> .
RTO	<i>RTO</i> (the <i>recovery time objective</i>) is the goal for the maximum allowable amount of time that it will take for a service to recover from a failure.
RPO	<i>RPO</i> (the <i>recovery point objective</i>) is the goal for the maximum allowable amount of data that may be lost due to a failure.

Then there also are a series of terms that we will use to describe reliable systems:

fault tolerant	A <i>fault tolerant</i> system will survive any single hardware or software component failure.
highly available	A <i>highly available</i> system is built from highly reliable components but may not survive a single component failure.
redundant	A <i>redundant</i> system comprises one or more spares for each component. This is a requirement for fault tolerance.
distributed	A <i>distributed</i> system has redundant components that are geographically dispersed.
disaster	A <i>disaster tolerant</i> system has <i>redundant</i> components <i>distributed</i>

- tolerant over a sufficient distance that any anticipated disaster will not cause a failure of all components of the same type.
- active/standby Also known as a *primary/backup* system, an active/standby system comprises two *redundant* systems, one providing services to end users and the other prepared to take over should the active system fail.
- active/active An *active/active* system is one in which several *redundant* subsystems cooperate by independently processing user requests to share the load.

With this as background, join us in our next column as we discuss various measurements of availability and reliability that are important to understand where we are and where we want to go.