Ten Fallacies of Availability and Reliability Analysis

Michael Grottke[1], Hairong Sun[2], Ricardo M. Fricks[3], and Kishor S. Trivedi[4]

[1] University of Erlangen-Nuremberg, Department of Statistics and Econometrics
Lange Gasse 20, D-90403 Nürnberg, Germany
Michael.Grottke@wiso.uni-erlangen.de

[2] Sun Microsystems, 500 Eldorado Blvd, Broomfield, CO 80021, USA
Hairong.Sun@sun.com

[3] Motorola Inc., 1501 West Shure Drive, Arlington Heights, IL 60004, USA
Ricardo.Fricks@motorola.com

[4] Duke University, Department of Electrical and Computer Engineering
Box 90291, Durham, NC 27708, USA
kst@ee.duke.edu

Abstract. As modern society becomes more and more dependent on computers and computer networks, vulnerability and downtime of these systems will significantly impact daily life from both social and economic point of view. Words like reliability and downtime are frequently heard on radio and television and read in newspapers and magazines. Thus reliability and availability have become popular terms. However, even professionals are in the danger of misunderstanding these basic concepts. Such misunderstandings can hinder advances in designing and deploying high-availability and high-reliability systems.

This paper delves into ten fallacious yet popular notions in availability and reliability. While the discussions on the first five fallacies clarify some misconceptions among reliability engineers working on modeling and analysis, the remaining five fallacies provide important insights to system engineers and companies focusing on system level integration.

1 Prologue

It is hard to discuss the reliability and availability concepts without first considering the lifetime of components and systems. We will mainly refer to systems in the explanation to follow but the same concepts will equally apply to components or units. In this section we review basic definitions that baseline our presentation to follow.

1.1 Basic Probability Theory Definitions

The lifetime or time to failure of a system can usually be represented by a random variable due to the intrinsic probabilistic nature of events that lead to system malfunction. Let the random variable X represent the lifetime or time to failure

T. Nanya et al. (Eds.): ISAS 2008, LNCS 5017, pp. 187–206, 2008.
© Springer-Verlag Berlin Heidelberg 2008

of a system. The continuous random variable X can be characterized by the (cumulative) distribution function (CDF) $F(t)$, the (probability) density function (PDF) $f(t)$, or the hazard (rate) function $h(t)$, also known as the instantaneous failure rate. The CDF represents the probability that the system will fail before a given time t, i.e.,

$$F(t) = \Pr(X \leq t). \tag{1}$$

The PDF describes the rate of change of the CDF, i.e.,

$$f(t) = \frac{dF(t)}{dt} = \lim_{\Delta t \to 0} \frac{\Pr(t < X \leq t + \Delta t)}{\Delta t}. \tag{2}$$

Hence, $f(t)\Delta t$ is the limiting (unconditional) probability that a system will fail in the interval $(t, t + \Delta t]$. However, if we have observed the system functioning up to some time t, we expect the conditional probability in the interval to be different from $f(t)\Delta t$. This leads us to the notion of the instantaneous failure rate, or the hazard rate function,

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < X \leq t + \Delta t \mid X > t)}{\Delta t} = \frac{f(t)}{1 - F(t)}. \tag{3}$$

Thus, $h(t)\Delta t$ represents the conditional probability that a system surviving to age t will fail in the interval $(t, t + \Delta t]$. Applied to a large population of systems, this conditional probability is the proportion of the survivors at time t that die during the immediately following small interval of time Δt.

The three functions $F(t)$, $f(t)$ and $h(t)$ are interrelated as shown in Table 1.

Table 1. Interrelationships between functions related to the lifetime distribution

$\boldsymbol{f(t)}$	$\frac{dF(t)}{dt}$	$h(t)e^{-\int_0^t h(\tau)d\tau}$
$\int_0^t f(\tau)d\tau$	$\boldsymbol{F(t)}$	$1 - e^{-\int_0^t h(\tau)d\tau}$
$\frac{f(t)}{\int_t^\infty f(\tau)d\tau}$	$\frac{dF(t)/dt}{1-F(t)}$	$\boldsymbol{h(t)}$

Any of these three functions can uniquely describe the lifetime distribution. For instance, if the time to failure of a system follows an exponential distribution with parameter λ then

$$F(t) = 1 - e^{-\lambda t}, \tag{4}$$

$$f(t) = \frac{d}{dt}\left(1 - e^{-\lambda t}\right) = \lambda e^{-\lambda t}, \tag{5}$$

$$h(t) = \frac{\lambda e^{-\lambda t}}{1 - [1 - e^{-\lambda t}]} = \lambda \tag{6}$$

for $t \geq 0$. Observe that the hazard rate function $h(t)$ shows that the exponential lifetime distribution is characterized by the age-independent failure rate λ. As a matter of fact, the exponential distribution is the only continuous probability distribution having a hazard function that does not change over time.

Therefore, whenever people refer to a lifetime distribution with constant failure rate, they are implicitly establishing the exponential distribution for the system lifetime.

1.2 Reliability Definitions

Recommendation E.800 of the International Telecommunications Union (ITU-T) defines reliability as the "ability of an item to perform a required function under given conditions for a given time interval." Therefore, for any time interval $(z, z+t]$ reliability $R(t \mid z)$ is the probability that the system does not fail in this interval, assuming that it is working at time z. Of specific interest are the intervals starting at time $z = 0$; reliability $R(t) := R(t \mid 0)$ denotes the probability that the system continues to function until time t. If the random variable X represents the time to system failure as before, then

$$R(t) = \Pr(X > t) = 1 - F(t), \tag{7}$$

where $F(t)$ is the system lifetime CDF.

Closely related to the reliability $R(t)$ is the definition of mean time to failure (MTTF). System MTTF is the expected time that a system will operate before the first failure occurs; i.e., on the average, a system will operate for MTTF hours and then encounter its first failure. The average of the system's lifetime distribution $E[X]$ is

$$E[X] = \int_0^\infty t f(t) dt = \int_0^\infty R(t) dt, \tag{8}$$

provided this integral is finite. If the right-hand side is not absolutely convergent, then $E[X]$ does not exist. Therefore, system MTTF can be computed by first determining its corresponding reliability function $R(t)$ and then applying (8). For example, if the system lifetime is exponentially distributed with failure rate λ then

$$R(t) = 1 - \left(1 - e^{-\lambda t}\right) = e^{-\lambda t} \tag{9}$$

and

$$\text{MTTF} = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}. \tag{10}$$

1.3 Availability Definitions

Availability is closely related to reliability, and is defined in ITU-T Recommendation E.800 as the "ability of an item to be in a state to perform a required function at a given instant of time or at any instant of time within a given time interval, assuming that the external resources, if required, are provided."

An important difference between reliability and availability is that reliability refers to failure-free operation of the system during an interval, while availability refers to failure-free operation of the system at a given instant of time.

Like in the case of reliability, we can restate the availability definition with the assistance of random variables. Let $Y(t) = 1$ if the system is operating at time t, and 0 otherwise. The most straightforward measure of system availability is the instantaneous availability $A(t)$, which is the probability that the system is operating correctly and is available to perform its functions at a specified time t, i.e.,

$$A(t) = \Pr(Y(t) = 1) = E[Y(t)]. \tag{11}$$

The instantaneous availability is always greater than or equal to the reliability; and in the absence of repairs or replacements, the instantaneous availability $A(t)$ is simply equal to the reliability $R(t)$ of the system.

Given $A(t)$ we can define the (*steady-state*) *availability* A of the system as

$$A = \lim_{t \to \infty} A(t). \tag{12}$$

The steady-state availability, or simply availability, represents the long-term probability that the system is available. It can be shown that the steady-state availability is given by

$$A = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}, \tag{13}$$

where the system mean time to repair (MTTR) is the average time required to repair system failures, including any time required to detect that there is a failure, to repair it, and to place the system back into an operational state; i.e., once the failure has occurred, the system will then require MTTR hours on the average to restore operation. It is known that the limiting availability depends only on the mean time to failure and the mean time to repair, and not on the nature of the distributions of failure times and repair times. There is an implied assumption in this model that repairs can always be performed which will restore the system to its best condition ("as good as new").

If the system lifetime is exponential with failure rate λ, and the time-to-repair distribution of the system is exponential with (repair) rate μ, then (13) can be rewritten as

$$A = \frac{\mu}{\lambda + \mu}. \tag{14}$$

Another concept of interest is the *interval* (or *average*) *availability* $A_I(t)$ of the system given by

$$A_I(t) = \frac{1}{t} \int_0^t A(\tau) d\tau. \tag{15}$$

The interval availability $A_I(t)$ is the expected proportion of time the system is operational during the period $(0, t]$. A property that can easily be verified if we represent the total amount of system uptime during $(0, t]$ by the random variable $U(t)$ is the following one:

$$A_I(t) = \frac{1}{t} \int_0^t E[Y(\tau)] d\tau = \frac{1}{t} E[U(t)]. \tag{16}$$

The *limiting average availability* A_I is the expected fraction of time that the system is operating:

$$A_I = \lim_{t \to \infty} A_I(t). \tag{17}$$

If the limit exists, then the steady-state and the limiting average availabilities are the same [1,2]; i.e.,

$$A_I = \lim_{t \to \infty} \frac{1}{t} \int_0^t A(\tau) d\tau = A. \tag{18}$$

2 Fallacies

2.1 "Fault Tolerance Is an Availability Feature and Not a Reliability Feature"

This fallacy comes from the misunderstanding of the reliability definition. The statement "the system continues to function throughout the interval $(0, t]$" does not imply the absence of internal system faults or error conditions during the interval $(0, t]$. Failure and recovery at component level is allowed as long as the system continues to function throughout the interval $(0, t]$. A simple example is Redundant Array of Independent (or Inexpensive) Disks (RAID) [3]. For RAID 1-5, it stores redundant data in different places on multiple hard disks. By placing data on multiple disks, I/O operations can overlap in a balanced way, improving performance and also increasing fault-tolerance.

Figure 1 is the state-transition diagram of a continuous-time Markov chain (CTMC) modeling the failure/repair behavior of a RAID 5 system. State 0 represents the state that all the N disks in the parity group are working, state 1 represents the failure of one disk. The parity group fails (data is lost) when there are double disk failures. The failure rate and repair rate for a disk are λ and μ, respectively.

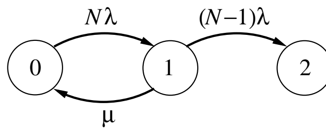


Fig. 1. CTMC for RAID 5 with N disks

Solving the CTMC model, it can be shown that the system reliability in the interval $(0, t]$ is given by [4]

$$R(t) = \frac{N(N-1)\lambda^2}{\alpha_1 - \alpha_2} \left(\frac{e^{-\alpha_2 t}}{\alpha_2} - \frac{e^{-\alpha_1 t}}{\alpha_1} \right), \tag{19}$$

where

$$\alpha_1, \alpha_2 = \frac{(2N-1)\lambda + \mu \pm \sqrt{\lambda^2 + 2(2N-1)\lambda\mu + \mu^2}}{2}. \tag{20}$$

From these expressions, the mean time to reach the absorbing state 2 (i.e., the system MTTF) is derived as

$$\text{MTTF} = \frac{2N-1}{N(N-1)\lambda} + \frac{\mu}{N(N-1)\lambda^2}. \tag{21}$$

If the MTTF of a disk is $\lambda^{-1} = 20$ years [5], and the MTTR is $\mu^{-1} = 20$ hours, using RAID 5 to add parity in a rotating way, the MTTF of the parity group with $N = 6$ disks will be $5,847.333$ years.

In a system without repair (i.e., $\mu = 0$) the time to failure follows a two-stage hypoexponential distribution with transition rates $N\lambda$ and $(N-1)\lambda$. The reliability function is then

$$\begin{aligned} R(t) &= \frac{N(N-1)\lambda^2}{N\lambda - (N-1)\lambda}\left(\frac{e^{-(N-1)\lambda t}}{(N-1)\lambda} - \frac{e^{-N\lambda t}}{N\lambda}\right) \\ &= Ne^{-(N-1)\lambda t} - (N-1)e^{-N\lambda t}, \end{aligned} \tag{22}$$

and the MTTF amounts to

$$\text{MTTF} = \frac{1}{N\lambda} - \frac{1}{(N-1)\lambda} = \frac{2N-1}{N(N-1)\lambda}. \tag{23}$$

For our RAID 5 example with parameter values given above follows a system MTTF of about 7.333 years, which is considerably less than the system MTTF in the presence of repair, and it is even less than the MTTF of a single disk; this stresses the importance of combining redundancy with effective and efficient repair.

However, we have seen that in the presence of adequate repair fault tolerance can improve system reliability. It is thus not only an availability feature, but also a reliability feature.

2.2 "Availability Is a Fraction While Reliability Is Statistical"

This statement seems to imply that availability is a deterministic concept, while reliability is a random quantity. In fact, as can be seen from (7), for a given time t reliability $R(t)$ is a fixed probability that depends on the distribution of the time to failure. Of course, the parameters of this distribution - or even the type of distribution - may be unknown; then the reliability needs to be estimated from measured data. For example, assume that we observe m new (or "as good as new") copies of the same system throughout the time interval $(0, t]$. If x_t of these copies do not fail during the observation period, then we can give a point estimate of $R(t)$ as the fraction x_t/m. Note that the number of non-failing copies X_t is random; therefore, the estimator $\hat{R}(t) = X_t/m$ is also a random variable. As a consequence, the point estimate x_t/m can be far from the true reliability $R(t)$; instead of merely calculating such a point estimate, it is therefore advisable to derive a confidence interval. Based on the fact that X_t

follows a binomial distribution with size m and success probability $R(t)$, it can be shown [6] that

$$\left[\left(1+\frac{m-x_t+1}{x_t f_{2x_t,2(m-x_t+1);\alpha}}\right)^{-1};1\right] \tag{24}$$

is the realized upper one-sided $100(1-\alpha)\%$ confidence interval for $R(t)$, where the expression $f_{2x_t,2(m-x_t+1);\alpha}$ denotes the (lower) $100\alpha\%$-quantile of the F-distribution with $2x_t$ numerator degrees of freedom and $2(m-x_t+1)$ denominator degrees of freedom. This means that if we repeat the experiment (of observing the number of non-failing systems among a set of m until time t) very often, then about $100(1-\alpha)\%$ of the confidence intervals constructed based on the respective measured values of x_t will contain the true but unknown reliability $R(t)$. Note that the estimator and the confidence interval given above are valid regardless of the distribution of the time to failure. As an example, assume that we observe 100 new copies of a system for 10 hours each. If one of them fails, then we estimate the reliability in the time interval $(0;10\text{ hr}]$ to be 0.99, while the realized upper one-sided 95% confidence interval is given by $[0.9534;1]$.

Similarly, the steady-state availability A can be estimated as follows. We can for example measure n consecutive times to failure $(Y_1,\dots,Y_n)$ and times to repair $(Z_1,\dots,Z_n)$ of a system in steady-state. All times to failure and times to repair, as well as the total up-time $U_n=\sum_{i=1}^{n} Y_i$ and the total downtime $D_n=\sum_{i=1}^{n} Z_i$ are random variables. Based on the values u_n and d_n actually observed, an obvious choice for a point estimate of steady-state availability is $\hat{A}=u_n/(u_n+d_n)$. Again, it is possible to derive a confidence interval. If all Y_i and Z_i are exponentially distributed with rate λ and μ, respectively, then $2\lambda U_n/(2\mu D_n)$ follows an F-distribution with $2n$ numerator degrees of freedom and $2n$ denominator degrees of freedom. Therefore, the realized upper one-sided $100(1-\alpha)\%$ confidence interval for steady-state availability A is given by [4]

$$\left[\left(1+\frac{d_n}{u_n f_{2n,2n;\alpha}}\right)^{-1};1\right]. \tag{25}$$

For example, if we have 10 samples of failures and repairs, and the total up time and total down time are 9990 hours and 10 hours, respectively, then the point estimate of availability is 0.999. Assuming that both the time to failure and the time to repair follow exponential distributions, the realized upper one-sided 95% confidence interval is $[0.9979;1]$.

This availability inference process is not always feasible since the system MTTF of commercial systems such as the ones supporting most computing and communications systems is of the order of months to years. So, a more practical approach aims at estimating the interval availability $A_I(t)$ for the interval $(0,t]$ with fixed length t (e.g., a week or a month) instead. One possible approach is to observe m statistically identical new (or "as good as new") copies of the system during the time interval $(0,t]$. For each copy $i=1,\dots,m$, the total downtime $d_i(t)$ (realization of the random variable $D_i(t)$) in the observation interval is recorded. Alternatively, we could observe the same system for m periods of

fixed duration t, provided that it is as good as new at the beginning of each of these periods. The random variables $D_i(t)$ would then represent the downtime of the system in the ith observation period, while $d_i(t)$ is the actual downtime experienced in this period.

Since the individual downtimes $D_i(t)$ are identically distributed random variables, the sample mean

$$\bar{D}(t) = \frac{1}{m} \sum_{i=1}^{m} D_i(t) \tag{26}$$

has an expected value that is equal to the true but unknown expected downtime in the interval $(0, t]$. We can therefore use the average of the observed downtimes to compute a point estimate of the system interval availability $A_I(t)$:

$$\hat{A}_I(t) = \frac{t - \frac{1}{m} \sum_{i=1}^{m} d_i(t)}{t} = 1 - \frac{\sum_{i=1}^{m} d_i(t)}{m \cdot t}. \tag{27}$$

If the individual downtimes $D_i(t)$ are independent, then the Central Limit Theorem [4] guarantees that for large sample sizes m the sample mean $\bar{D}(t)$, (26), approximately follows a normal distribution. This fact can be used for deriving an approximate confidence interval to the interval availability estimate.

As the observation period t increases, the interval availability estimated using (27) will eventually converge to the true steady-state availability A of the system, i.e.,

$$\lim_{t \to \infty} \hat{A}_I(t) = A. \tag{28}$$

Simulation experiments in [7] show that it is possible to produce a steady-state availability estimate with an efficient confidence interval based on a temporal sequence of interval availability estimates. This technique does not depend on the nature of the failure and repair time distributions.

All this shows the similarities between reliability and availability from a stochastic point of view: Both reliability and availability are fixed but unknown values; they can be estimated by fractions; the estimators are random variables; and based on the distributions of these random variables, we can come up with expressions for constructing confidence intervals.

2.3 "The Term 'Software Reliability Growth' Is Unfortunate: Reliability Is Always a Non-increasing Function of Time"

This fallacy is caused by the fact that reliability $R(t \mid z)$, the probability of no failure in the time interval $(z, z+t]$, can be considered as a function of interval length t, or as a function of interval start time z.

The latter aspect is often forgotten due to the importance of the reliability function $R(t) := R(t \mid 0)$ referring to the reliability in the interval $(0, t]$. By integrating both sides of (3) we get

$$\int_0^t h(\tau) d\tau = \int_0^t \frac{f(\tau)}{1 - F(\tau)} d\tau \; = \; \int_0^t \frac{-\partial R(\tau)/\partial \tau}{R(\tau)} d\tau; \tag{29}$$

using the boundary condition $R(0) = 1$, this yields [4]

$$R(t) = e^{-\int_0^t h(\tau)d\tau}. \qquad (30)$$

Since the hazard function $h(t)$ is larger than or equal to zero for all $t > 0$, $R(t)$ is always a non-increasing function of t. This result is reasonable: If the time interval examined is extended, then the probability of experiencing a failure may be higher, but it cannot be lower.

However, if the interval length t considered is set to a fixed value, say t_0, while the interval start time z is allowed to vary, then reliability $R(t_0 \mid z)$ can be a decreasing, constant, or increasing function of z. Software reliability growth models describe how the failure generating process evolves as testing and debugging proceed. Almost all of them assume that $R(t_0 \mid z)$ will eventually be a non-decreasing function of z; hence the term "software reliability growth."

For example, in the important class of non-homogeneous Poisson process models, the instantaneous failure rate of the software is a mere function of time and does not depend on the number of faults discovered so far, etc. It can be shown that according to these models the reliability $R(t \mid z)$ is given by [8]

$$R(t \mid z) = e^{-\int_z^{z+t} h(\tau)d\tau}; \qquad (31)$$

in this context, the function $h(t)$ is often called "program hazard rate." Obviously, (31) includes (30) as the special case $z = 0$. Regardless of the value of z, $R(t \mid z)$ is always a non-increasing function of t, starting out at $R(0 \mid z) = 1$. However, there will eventually be software reliability growth in the sense described above if and only if the instantaneous failure rate is eventually a non-increasing function of time. Figure 2 illustrates these aspects based on a so-called S-shaped software reliability growth model, featuring an instantaneous failure rate that is first increasing (e.g., due to learning effects on part of the testers) and then decreasing (because the software quality improves). The strictly decreasing function $R(t \mid 0)$ is depicted in the left diagram. For the arbitrarily chosen interval length t_0, the right diagram shows that $R(t_0 \mid z)$ as a function of z first decreases and then increases.

Thus, the term "software reliability growth" is correct, as it refers to the fact that the probability of no failure occurrence in a time interval of fixed length tends to be higher if the interval start time is increased.

2.4 "MTTF Is the Whole Story about Reliability"

This misconception is probably due to the fact that simple system reliability and availability (R&A) models often assume that the time to failure of individual components follows an exponential distribution. As we have seen in Section 1.1, the only parameter of this distribution is the constant failure rate λ, which according to (10) is the reciprocal value of the MTTF. Therefore, if we know that the time to failure is exponentially distributed, then this piece of information plus the MTTF indeed completely describe the distribution of the time to failure

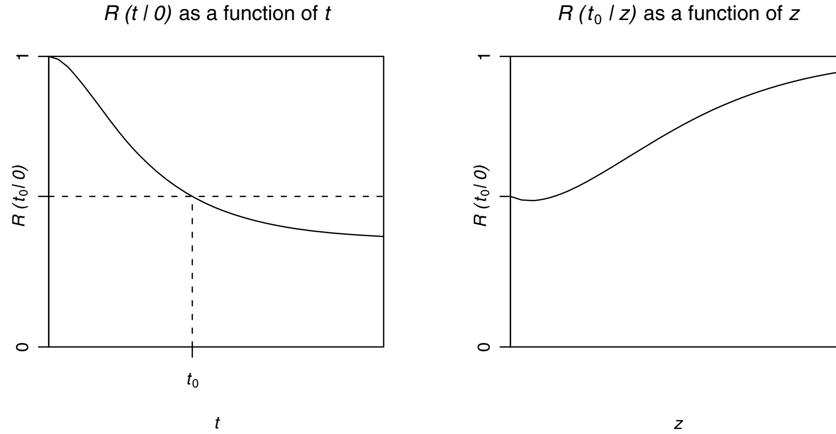


Fig. 2. Reliability $R(t \mid z)$ as functions of t and z

and hence the reliability function $R(t)$. However, it needs to be stressed that the distribution of the time to failure is never fully specified by its expected value, the MTTF, alone; we always need at least additional information about the type of distribution. Unfortunately, the exponential distribution assumption is sometimes not stated explicitly.

Even if the time to failure is assumed to follow a more complex distribution, the MTTF can suffice as an additional piece of information. For example, a two-stage Erlang distribution [4] features an increasing failure rate, but like the exponential distribution it only has one parameter, λ. From a given MTTF, the value of this parameter can be derived as $\lambda = \frac{2}{\mathrm{MTTF}}$.

Note that for distributions with more than one parameter, information on the type of distribution and the MTTF will not be enough for completely specifying the distribution - information about additional moments of the distribution (or about its additional parameters) will be needed. For example, if the time to failure is known to follow a k-stage Erlang distribution (where k is unknown), then in addition to the MTTF we would require further information, like the variance of the time-to-failure distribution Var.TTF, in order to derive the two model parameters $\lambda = \frac{\mathrm{MTTF}}{\mathrm{Var.TTF}}$ and $k = \mathrm{MTTF} \cdot \lambda = \frac{\mathrm{MTTF}^2}{\mathrm{Var.TTF}}$.

The fact that the MTTF by itself does not completely specify the time-to-failure distribution also means that decisions based on the MTTF alone can be wrong. As an illustration, consider the analysis of triple modular redundant (TMR) systems. The TMR technique is widely adopted in the design of high-reliability systems. Since two of the three components present in a TMR system need to function properly for the system to work, the reliability of such a system is [4]

$$R(t) = 3R_u^2(t) - 2R_u^3(t), \tag{32}$$

where $R_u(t)$ represents the reliability of any of the three statistically identical components. If the time to failure of each component follows an exponential

distribution with reliability function given by (9), then we get

$$R(t) = 3e^{-\lambda t} - 2e^{-\lambda t}. \tag{33}$$

It can be shown that $R(t) > R_u(t)$ for $t < t_0 \equiv \ln(2)/\lambda$. Therefore, the TMR type of redundancy clearly improves reliability for a mission time that is shorter than t_0. However, the MTTF of the TMR system,

$$\text{MTTF} = \int_0^\infty 3e^{-\lambda t}dt - \int_0^\infty 2e^{-\lambda t}dt = \frac{3}{2\lambda} - \frac{2}{3\lambda} \;=\; \frac{5}{6\lambda}, \tag{34}$$

is smaller than $1/\lambda$, the component MTTF. Based on the MTTF alone, a system designer would always favor the single component over the TMR system; as we have seen, this decision is wrong if the mission time is shorter than t_0.

Therefore, MTTF is not the whole story about reliability. It does not suffice to fully specify the time-to-failure distribution; decisions based on the MTTF alone can thus be wrong.

2.5 "The Presence of Non-exponential Lifetime or Time-to-Repair Distributions Precludes Analytical Solution of State-Space Based R&A Models"

One common misconception is that analytic solutions of state-space based R&A models are only feasible if all modeled distributions are exponential or geometric in nature; if that is not the case, simulation modeling is the only viable alternative. This assertion could not be further from the truth given the rich theory of non-Markovian modeling.

Markov models have often been used for software and hardware performance and dependability assessment. Reasons for the popularity of Markov models include the ability to capture various dependencies, the equal ease with which steady-state, transient, and cumulative transient measures can be computed, and the extension to Markov reward models useful in performability analysis [9]. For example, Markov modeling is quite useful when modeling systems with dependent failure and repair modes, as well as when components behave in a statistically independent manner. Furthermore, it can handle the modeling of multi-state devices and common-cause failures without any conceptual difficulty.

Markov modeling allows the solution of stochastic problems enjoying the property: the probability of any particular future behavior of the process, when its current state is known exactly, is not altered by additional knowledge concerning its past behavior. For a homogeneous Markov process, the past history of the process is completely summarized in the current state. Otherwise, the exact characterization of the present state needs the associated time information, and the process is said to be non-homogeneous. Non-homogeneity extends the applicability of Markov chains by allowing time-dependent rates or probabilities to be associated to the models. For instance, in case of a non-homogeneous CTMC, the infinitesimal generator matrix $Q(t) = [q_{ij}(t)]$ is a function of time. This implies that the transition rates $q_{ij}(t)$ and $q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t)$ are also functions of t.

A wide range of real dependability and performance modeling problems fall in the class of Markov models (both homogeneous and non-homogeneous). However, some important aspects of system behavior in stochastic models cannot be easily captured through a Markov model. The common characteristic these problems share is that the Markov property is not valid (if valid at all) at all time instants. This category of problems is jointly referred to as non-Markovian models and include, for instance, modeling using phase-type expansions, supplementary variables, semi-Markov processes (SMPs), and Markov regenerative processes (MRGPs). For a recent survey, see [10].

Thus, state-space based R&A models can be solved analytically, even if lifetime or time-to-repair distributions are non-exponential.

2.6 "Availability Will Always Be Increased with More Redundancy"

In a perfect world, availability increases with the degree of redundancy. However, if coverage ratio and reconfiguration delay are considered, availability does not necessarily increase with redundancy [11].

Assume there are n processors in a system and that at least one of them is needed for the system being up. Each processor fails at rate λ and is repaired at rate μ. The coverage probability (i.e., the probability that the failure of one processor can be detected and the system can be reconfigured successfully) is c. The average reconfiguration delay after a covered failure is $1/\delta$, and the average reboot time after an uncovered failure is $1/\beta$. In the CTMC model in Fig. 3 state i means there are i processors working, state D_i stands for the case that there are i processors working, the failure of a processor has been detected and the system is under reconfiguration, while state B_i means there are i processors working, the failure of a processor is undetected and the system is undergoing a reboot. The system is only available in states $1, 2, \ldots, n$.

According to the numerical results in Fig. 4, system availability is maximized when there are 2 processors. Therefore, availability will not always increase with more redundancy, and the coverage probability and reconfiguration delay play important roles. To realize redundancy benefits, coverage must be near perfect and reconfiguration delay must be very small.

2.7 "Using Low-Cost Components Can Always Build Highly Available Systems"

In Section 2.6, we argued that the coverage probability plays an important role for availability. From industry experience, low-cost components are usually designed with relatively poor fault management because component vendors are reluctant to increase expense to improve the quality of products. Thus, there are many cases in which low-cost components are accompanied with lower coverage probability, lower fault-detection probability, longer fault-detection time and larger no-trouble-found ratio (i.e., one cannot find where the problem is when the system fails). From Fig. 4, we can conjecture that we might not be able to build a highly-available system if the coverage probability is low and/or

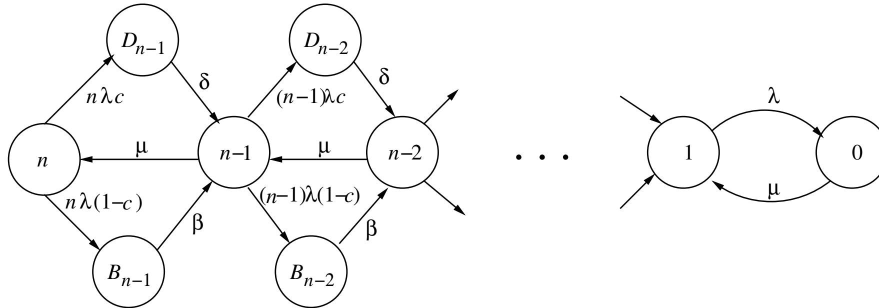
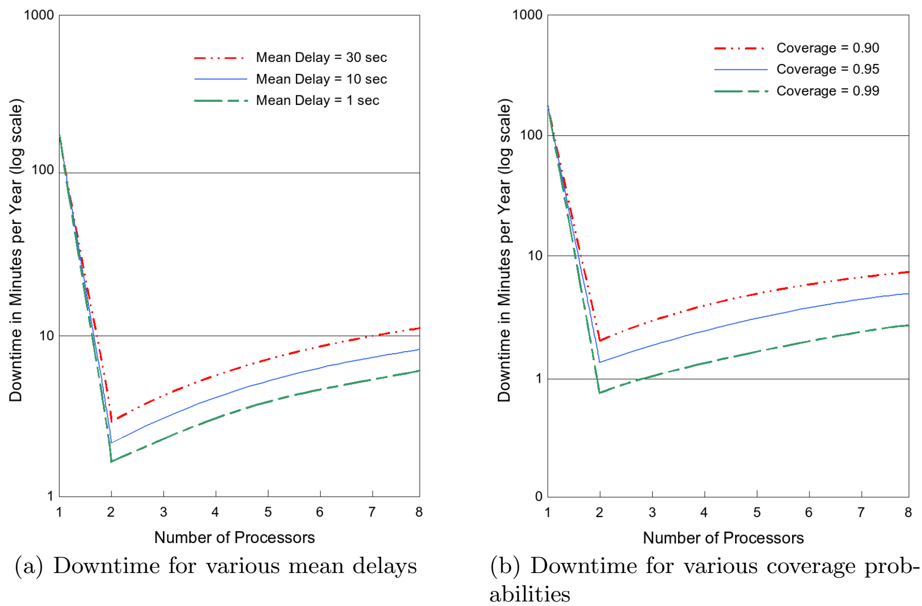


Fig. 3. CTMC model for a multi-processor system



(a) Downtime for various mean delays

(b) Downtime for various coverage probabilities

Fig. 4. System downtime as a function of the number of processors used

the reconfiguration delay is long, which are the attributes that usually come with low-cost components. So before choosing a low-cost component, make sure to assess its reliability and fault coverage probability and ensure that they meet the availability goal at the system level.

2.8 "A Ten-Times Decrease in MTTR Is Just as Valuable as a Ten-Times Increase in MTTF"

Equation (13) suggests that a ten-times decrease in MTTR is just as valuable as a ten-times increase in MTTF. That is correct from system availability point of

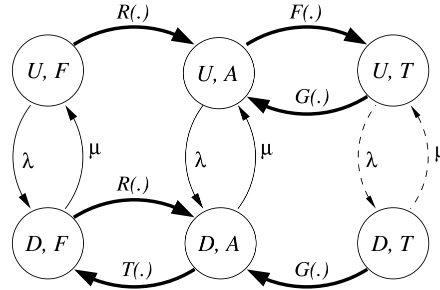


Fig. 5. MRGP Model for single-user-single-host Web browsing

view. However, for most of the applications running on the Internet, a decrease in MTTR is sometimes more valuable than the corresponding increase in MTTF, due to the automatic retry mechanism implemented at various layers of the Internet protocols which masks some short outages and makes them imperceptible [12].

Figure 5 is an MRGP model for single-user-single-host Web browsing. The circles in this figure represent the states of our model, and the arcs represent state transitions. Each state is denoted by a 2-tuple (s, u), where s is the state of the platform and u is the user status. $s = \{U, D\}$ includes the situations that the underlying system is up and down, respectively, and $u = \{T, A, F\}$ contains the user status of thinking, active, and seeing a failure, respectively. Our model's state space is the Cartesian product of s and u, $\{(U, T), (D, T), (U, A), (D, A), (U, F), (D, F)\}$.

The system fails at rate λ (from (U, u) to (D, u)), and is repaired at rate μ (from (D, u) to (U, u)). After the user has been active for a certain amount of time, which has a CDF of $F(.)$, she enters thinking state (from (s, A) to (s, T)), and comes back to active (from (s, T) to (s, A)) after some time (with CDF $G(.)$). If she is active and the network is down (state (D, A)), the browser retries after some time that follows a distribution with CDF $T(.)$. The repair of the system in state (D, A) will be detected immediately by the automatic HTTP recovery mechanism. If the retry fails, the user sees a failure (state (s, F)). The user reattempts to connect to the Web host, which is represented by transition with distribution $R(.)$. Note that transitions $F(.)$, $G(.)$, $T(.)$, and $R(.)$ have general distributions (solid thick arcs in Fig. 5); hence the model described above is not a CTMC, nor is it an SMP because of the existence of local behaviors, which are known as state changes between two consecutive regenerative points. For example, if the failure transition from (U, A) to (D, A) occurs, the user active transition $F(.)$ is not present in state (D, A). This exponential transition is known as competitive exponential transition (represented by solid thin arcs), and its firing marks a regenerative point. On the other hand, the transitions of the server going up and down in states (U, T) and (D, T) do not affect (add, remove or reset the general transitions) the user thinking process which is generally

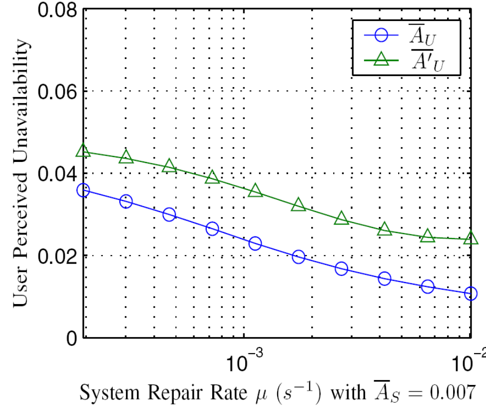


Fig. 6. User-perceived unavailability $\bar{A}_U$, $\overline{A'}_U$

distributed. They are called concurrent exponential transitions (represented by dashed thin arcs), and their occurrences are just local behaviors.

Following the methodology in [12], and using the assumptions on parameters and distributions [13], we can get the numerical results depicted in Fig. 6. For comparison purpose, we also constructed and solved the corresponding CTMC model, i.e., we replaced all the general distributions with exponential distributions with the same means.

We denoted the user-perceived service availability of the CTMC model by $\overline{A'}_U$. System unavailability was set to a constant 0.007, while the failure rate λ and repair rate μ varied accordingly. If we incorporate both the failure recovery behaviors of the service-supporting infrastructure and the online user behaviors and evaluate the dependency of the user-perceived unavailability on parameters including the service platform failure rate/repair rate, user retry rate, and user switching rate, we will find the user-perceived unavailability very different from the system unavailability.

For Web applications, the user-perceived availability is more sensitive to the platform repair rate; i.e., for two systems with same availability, the one with faster recovery is better than the one with higher reliability from an end user's perspective. We also found that the CTMC model overestimates the user-perceived unavailability by a significant percentage.

2.9 "Improving Component MTTR Is the Key to Improve System Availability"

This fallacy results from a common misunderstanding of the steady-state availability formula (13): If we maintain the MTTF invariant (e.g., by not investing in more reliable components) then we can still improve system availability by reducing component MTTR, right? Not necessarily, because the MTTF and MTTR parameters in the system availability formulas are related to system properties,

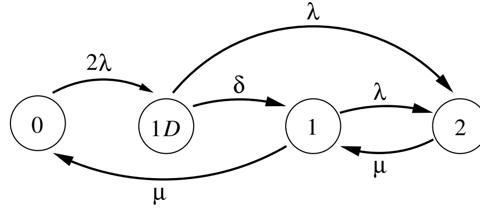


Fig. 7. CTMC model of a two-component parallel redundant system with fault recovery delay

not component ones. The system MTTR parameter in a fault-tolerant system, for instance, will also be a function of the quality of its fault management mechanism.

Consider for instance the fact that any given CTMC model can be reduced to an equivalent two-state model for the sole purpose of conducting steady-state analysis using the aggregation technique introduced in [14]. With this procedure we are collapsing all operational states into a single state, and all failed states into a single failure state. Failures in the equivalent model happen with rate λ_{eq} and are repaired with rate μ_{eq}. Therefore, we can define system MTTF or MTTF_{eq} as $1/\lambda_{eq}$ and system MTTR or MTTR_{eq} as $1/\mu_{eq}$ in reference to the trivial solution of a two-state availability model provided by (13). The problem is that these equivalent rates are numerical artifacts with complex formulae that most of the time cannot be physically interpreted (e.g., there is no simple relation mapping a system MTTR to component MTTR).

For example, consider a two-component parallel redundant system with a single shared-repair facility. The availability model is shown in Fig. 7. In the state transition diagram, state $1D$ represents the recovery behavior after the first fault in the system (i.e., the first component failure). All other states are labeled by the number of failed components. States $1D$ and 2 are assumed to be the system failure states in this example. The component failure and repair rates are λ and μ, respectively. Once the first system fault is triggered, the system will recover with rate δ. The time the CTMC stays in state $1D$ represents the combined time the system's fault manager needs to react to the first system fault. A second fault during this sojourn time in state $1D$ leads the system directly to the system failure represented by state 2. This event happens with rate λ.

The steady-state solution of the CTMC model in Fig. 7 results in the following state probabilities:

$$\pi_0 = \frac{\mu^2(\lambda+\delta)}{2\lambda^2(\lambda+\mu+\delta)E}, \qquad \pi_{1D} = \frac{\mu^2}{\lambda(\lambda+\mu+\delta)E}, \tag{35}$$

$$\pi_1 = \frac{\mu(\lambda+\delta)}{\lambda(\lambda+\mu+\delta)E}, \qquad \pi_2 = \frac{1}{E}, \tag{36}$$

with

$$E = 1 + \frac{\mu(\lambda+\delta)}{\lambda(\lambda+\mu+\delta)} + \frac{\mu^2}{\lambda(\lambda+\mu+\delta)} + \frac{\mu^2(\lambda+\delta)}{2\lambda^2(\lambda+\mu+\delta)}. \tag{37}$$

Equivalent system failure and repair rates can be determined applying the aggregation techniques introduced in [14]. For the system in Fig. 7 we obtain

$$\lambda_{eq} = \frac{2\lambda\pi_0 + \lambda\pi_1}{\pi_0 + \pi_1}, \tag{38}$$

$$\mu_{eq} = \frac{\delta\pi_{1D} + \mu\pi_2}{\pi_{1D} + \pi_2}, \tag{39}$$

with system availability given by (14). To better understand the impact of the equivalent rates on system availability look at the composition of λ_{eq} and μ_{eq} in (38) and (39). Take μ_{eq} as an example. A typical setting for the parameter values is: $\lambda \approx 10^{-5}$, $\mu \approx 10^{-1}$, and $\delta \approx 10^2$. Then $\delta\pi_{1D} \gg \mu\pi_2$ in the numerator of (39). This shows that the recovery rate δ, not μ, is the key to improving μ_{eq}; thus improving system availability.

This example has illustrated a case where a higher system availability can be reached much more effectively by increasing the system recovery rate rather than decreasing the component MTTR.

2.10 "High-Availability Systems Should Have No Single Point-of-Failure"

Single point-of-failure (SPOF) analysis is one of the traditional practices in reliability engineering. Naive interpretation of the topology of reliability block diagrams or other architectural diagrams may lead to the erroneous perception that the optimal improvement opportunity (without considering costs) in any high-availability architecture is always the removal of SPOFs. What the analyst may fail to realize is that the structure of the system is just one of many factors that determine the importance of a component in a high-availability system. Other determining factors are for instance the reliability/unreliability (or availability/unavailability) of the system components, the mission time, and target availability. Besides, the adoption of hierarchical modeling approaches may also lead to confusion. For instance, subsystems that appear as SPOFs in high-level diagrams may in fact correspond to highly redundant component structures.

Importance theory, a concept introduced by Birnbaum [15] in 1969, provides superior criteria than SPOFs alone for objective placement of redundancy. The reasoning behind the theory is that during the design of a system, the choice of components and their arrangement may render some components to be more critical with respect to the functioning of the system than others. The first quantitative ranking metrics proposed were the *structural importance* and *Birnbaum component importance*.

The structural importance of a component establishes the probability that the system shall fail when the component fails, i.e., the component is critical for system operation. Similar to an SPOF analysis, structural importance allows us to consider the relative importance of various components when only the structure of the system is known, but no other information is available.

When we do have additional information, improved measures such as the Birnbaum component importance provide a better framework to identify improvement

opportunities to the system. For instance, when we additionally know the individual reliability of system components, we can compute the Birnbaum component importance. Semantically, this new metric represents the rate at which the system reliability improves as the reliability of a particular component improves. Another way of interpreting the Birnbaum importance metric is the probability that at a given time the system is in a state in which the component is critical for system operation. The larger the Birnbaum importance measure is, the more important the component is, in agreement with the intuition that a component that is frequently critical should be considered important.

To exemplify the distinction of both importance measures, consider the series-parallel system represented by the reliability block diagram in Fig. 8.

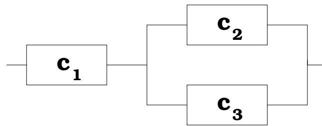


Fig. 8. Series-parallel reliability block diagram

The system is operational as long as component c_1 is functioning together with at least one of the two parallel components. Just based on the block diagram, we can determine, using the methods in [16] for instance, that the structural importance of c_1 is three times larger than those of either c_2 or c_3. This is an outcome that agrees with the intuition that series components are more fundamental to system reliability than parallel components, matching the SPOF reasoning. Now let us assume that we also know the intrinsic reliability of the system components. Suppose that the intrinsic reliability of the series component is 90% for a given mission time T, while the reliability of the parallel components are just 30% for the same mission time. Then, one can determine, using also the methods in [16] for instance, the Birnbaum importance of the components to be 0.51 for c_1, and 0.63 for the other two components. Therefore, the analysis indicates that components c_2 and c_3 should be the target of improvements (contrary to the results of a SPOF analysis) because at time T there is a 63% probability of the system being in a state that the functioning of these components is critical. For a comprehensive survey of other importance measures see [17].

3 Conclusions

Modern society and economy have been posing an increasingly imperative demand on the availability and reliability of computer systems and computer networks. The so called “24x7” (24-hours-a-day-and-7-days-a-week) requirement for these systems presents an unprecedented technical challenge. However, seemingly well-known concepts like availability and reliability are sometimes misunderstood

even by professionals; such misunderstandings may eventually hinder advances in designing and deploying high-availability and high-reliability systems. This paper introduced ten fallacies existing in availability and reliability analysis and traced them back to their theoretical flaws. The first five fallacies address misconceptions related to R&A modeling and analysis, while the remaining ones provide insights for system engineers and companies focusing on system level integration.

References

1. Barlow, R.E., Proschan, F.: Statistical Theory of Reliability and Life Testing - Probability Models. Holt, Rinehart and Winston, New York (1975)
2. Leemis, L.M.: Reliability: Probability Models and Statistical Methods. Prentice-Hall, Englewood Cliffs (1995)
3. Patterson, D.A., Gibson, G.A., Katz, R.H.: A case for redundant arrays of inexpensive disks (RAID). In: Proc. SIGMOD Conference, pp. 109–116 (1988)
4. Trivedi, K.S.: Probability & Statistics with Reliability, Queueing, and Computer Science Applications, 2nd edn. John Wiley and Sons, New York (2001)
5. Schroeder, B., Gibson, G.A.: Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? In: Proc. 5th USENIX Conference on File and Storage Technologies (2007)
6. Hald, A.: Statistical Theory with Engineering Applications. John Wiley and Sons, New York (1952)
7. Fricks, R.M., Ketcham, M.: Steady-state availability estimation using field failure data. In: Proc.Annual Reliability and Maintainability Symposium 2004, pp. 81–85 (2004)
8. Grottke, M., Trivedi, K.S.: On a method for mending time to failure distributions. In: Proc. International Conference on Dependable Systems and Networks 2005, pp. 560–569 (2005)
9. Trivedi, K.S., Muppala, J.K., Woolet, S.P., Haverkort, B.R.: Composite performance and dependability analysis. Performance Evaluation 14(3 & 4), 197–216 (1992)
10. Wang, D., Fricks, R., Trivedi, K.S.: Dealing with non-exponential distributions in dependability models. In: Kotsis, G. (ed.) Performance Evaluation - Stories and Perspectives, pp. 273–302. Österreichische Computer Gesellschaft, Wien (2003)
11. Trivedi, K.S., Sathaye, A., Ibe, O., Howe, R.: Should I add a processor? In: Proc. Twenty-third Hawaii International Conference on System Sciences, pp. 214–221 (1990)
12. Choi, H., Kulkarni, V.G., Trivedi, K.S.: Markov regenerative stochastic Petri nets. Performance Evaluation 20, 335–357 (1994)
13. Xie, W., Sun, H., Cao, Y., Trivedi, K.S.: Modeling of user perceived webserver availability. In: Proc. IEEE International Conference on Communications, vol. 3, pp. 1796–1800 (2003)
14. Lanus, M., Lin, Y., Trivedi, K.S.: Hierarchical composition and aggregation of state-based availability and performability models. IEEE Trans. Reliability 52(1), 44–52 (2003)

15. Birnbaum, Z.W.: On the importance of different components in a multicomponent system. In: Krishnaiah, P.R. (ed.) Multivariate Analysis - II, pp. 581–592. Academic Press, New York (1969)
16. Henley, E.J., Kumamoto, H.: Reliability Engineering and Risk Assessment. Prentice-Hall, Englewood Cliffs (1981)
17. Wang, D., Fricks, R.M., Trivedi, K.S.: Importance analysis with Markov chains. In: Proc. Annual Reliability and Maintainability Symposium, pp. 89–95 (2003)