Uptime and Downtime Analysis for Hierarchical Redundant Systems in Telecommunications

E. G. Coffman, Jr.
Columbia University
500 W. 120th St.
New York, NY 10027
coffman@cs.columbia.edu

Y. Kogan & W. Lai
AT&T, Operations & IT
Middletown, NJ 07748
yk1251, wlai @att.com

V. Ramaswami
AT&T, Research
Florham Park, NJ 07932
vramaswami@att.com

ABSTRACT

We consider non-degradable hierarchical redundant systems having multiple working and failure modes with restoration time depending on failure type. We evaluate these systems using two measures: generalized uptime and traditional downtime. We define the Impact Weighted System Uptime (IWSU) and illustrate its usefulness in practical terms, viz., an IP router. Next, we provide an analysis that fits the downtimes by a heavy-tailed log PH distribution. For these downtime distributions, we study whether it is more cost effective to reduce failure rates or to speed up the response to failures The first option is a vendor problem, but the second is a service provider problem. A numerical example is given to help appreciate the tradeoff.

1. INTRODUCTION

It is well recognized that classical reliability theory is not adequate for evaluating hierarchical telecommunication systems. In particular, an opening statement in a survey paper [7] by John F. Meyer says that "their evaluation calls for continuous innovation with regard to measure definition, model construction/solution, and tool development". Existing performability theory emphasizes degradable performance of such systems and combines their performance and reliability in one metric. Moreover, [8] has an explicit assumption that "each component, representing an independent failure or degradable mechanism, has a single working mode and arbitrary number of failure modes". In contrast, we consider non-degradable redundant systems at the upper hierarchical levels which have multiple working and failure modes with restoration time depending on failure type, and we evaluate these systems using two measures: generalized uptime and traditional downtime. The performability analysis of a general such hierarchical system leads only to numerical results under exponential assumptions on failure and repair times. In this paper, we identify several important network elements (e.g., IP routers, Radio Network Controller, etc.) as well the entire segment of a Mobility network, where the general performability approach can be significantly simplified, as follows. First, we use independent Markov models of each hierarchical level. Second, our model is an absorbing Markov chain rather than ergodic one and its absorption states correspond to failures which are not mitigated by redundancy. Third, we quantify reliability of each hierarchical level by the expected absorption time starting from the "perfect" state where all components are operational. That expected absorption time is referred to as the level uptime. Finally, we weight level uptimes by the respective impact of their failures and obtain impact weighted system uptime (IWSU). Applications of absorbing Markov chains to reliability started as early as 1974 [3]. Reference [4] applies an M-out-of-(M+1) redundancy model with absorption to an analysis of disk arrays and the expected absorption time is referred to as mean time to data loss (MTTDL). Our paper is different from the previous applications of absorbing Markov models in reliability in the following two key respects. First, we have a special redundancy structure of subsystems that include silent failures and unsuccessful failovers. This structure allows us to obtain explicit expressions for the IWSU. Second, we use these expressions to analyze typical architectures and maintenance strategies. Then, in the second part of the paper, we study the downtimes of a set of failures that are illustrative of current systems with silent failures. Our analysis fits their distribution to a heavy-tailed log-PH distribution. The set of failures is presumed to occur in redundant systems at upper hierarchical levels of traffic processing. The heavy tail is a natural consequence of silent failures. Long-duration failures at upper hierarchical levels are always a concern because they impact a large number of end customers. It is useful, therefore, to compute the cumulative distribution of the total downtime generated by a Poisson failure arrival process where the corresponding downtimes are i.i.d. with a log PH distribution. This result is used to compare two different ways of reducing the 95th and 90th percentiles of the total-downtime distribution. As will be seen in Section 4, failure rate reduction is much more cost effective for the service provider than techniques aimed at reducing the failure detection and service restoration time.

2. IMPACT WEIGHTED SYSTEM UPTIME

Consider a system that consists of three subsystems with increasing hierarchical levels $i = 1, 2, 3$ which we illustrate later by an IP router. In general, all subsystems are redundant. We count only the so-called *traffic impacting* failures which occur because of the limitations of the redundancy that has been adopted. Let N be the number of elements in subsystem 1. The failure impact of an element in subsystem 1 is 0 if the redundancy protects against that failure, and is 1 otherwise. However, the failure impact of the highest-level subsystem is N because it impacts all elements in subsystem 1. The impact of a failure of a subsystem at level 2 is the number of elements K in subsystem 1 that are connected to it. Let U_i be the expected uptime of subsystem $i = 1, 2, 3$. Then we define the impact-weighted system uptime as (cf. [5])

$$IWSU = \left(1/U_1 + K/U_2 + N/U_3\right)^{-1}.$$

It is not difficult to generalize this example to a hierarchy with a greater number of levels. For a 3-level example, consider a multi-chassis router consisting of L line-card chassis which are interconnected by a fabric card chassis. (Hereafter, we simplify these terms to *card chassis* and *fabric chassis*.) The total number of line cards is N and each card chassis has $K = N/L$ line cards and two route processors. A line card is a single point of failure, but the two route processors are in 1+1 configuration and the fabric chassis has M out of $M+1$ redundancy. Here, subsystems 1, 2 and 3 correspond to line cards, the route processors, and fabric chassis, respectively. Each hierarchical level is generally modeled as a system consisting of $M+1, M \geq 1$ independent identical units which may fail and be repaired upon the failure. We assume that the failure is detected with probability $0 < C \leq 1$. Upon failure detection of an active unit, the traffic is switched over to the hot standby unit or redistributed over the remaining M units without any service interruption. The system fails in case of a silent (undetected) failure of an active unit or when another unit fails before the previously failed unit is repaired. Under exponential assumptions, such a system can be modeled by an absorbing Markov chain whose absorption states correspond to failures of the system. Then the expected absorption time starting in the state where all $M+1$ units are operational is referred to as the level uptime (LU). Explicit expressions for the LU are derived in [6].

3. HEAVY TAILED DOWNTIME

Our motivation for fitting a distribution to downtime data arose from an attempt to answer an important question: Were silent failures outliers or were they in some way endemic to the system? Figure 1 provides the downtimes (the omitted units are not needed for present purposes), and Figure 2 provides a log log plot of the empirical complementary distribution function that exhibits a heavy tail tending to a linear asymptote. We chose the heavy tailed class of log-phase type distributions introduced by Ramaswami for fitting the data. By modeling a random variable Y as Y=e^X, where X is phase type, we get a log phase type random variable which has a power-law tail of the form $1/y^\eta$. For a mathematical treatment of log-phase type distributions and their applications, see Ahn et al [1]. As noted in [1], the log-PH class has many advantages over many of the classical models used in the context of heavy tails and can approximate any distribution in [1, ∞) arbitrarily closely, the restriction of the interval to the right of 1 not itself resulting in a loss of generality assuming an appropriate rescaling of the data if needed. We attempted to fit a phase type distribution to the logarithms, and after some experimentation with different parameter values, we obtained, using the EM algorithm of Asmussen et al. [2], a phase type distribution of order 6 characterized by the following parameters

$$\alpha = (0.5867, 0.0001, 0.3989, 0, 0, 0.0143)$$

and the matrix T given by

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	-2.675	2.675	0.000	0.000	0.000	0.000
[2,]	0.000	-2.675	2.675	0.000	0.000	0.000
[3,]	0.000	0.000	-2.949	2.949	0.000	0.000
[4,]	0.000	0.000	0.000	-2.949	2.949	0.000
[5,]	0.000	0.000	0.000	0.000	-2.949	2.949
[6,]	0.000	0.000	0.000	0.000	0.000	-3.719

We applied the Kolmogorov-Smirnov test to the empirical and fitted complementary distribution functions and the fit yielded a p-value of 0.9776, thus confirming the excellent agreement seen in Figure 3. Note the small number of distinct parameters identified. These we found to be quite stable upon testing with different values of the parameters of the iterative scheme. As noted earlier, this demonstrated that silent failures create unacceptable heavy tailed phenomena.

4. TOTAL ANNUAL DOWNTIME

In this section, we describe the computation of the distribution function of the random variable $\tau = \sum_{i=1}^{N} D_i$, where D_1, D_2, …, are i.i.d. random variables with the log PH distribution given in Section 3, and where N has the Poisson distribution with parameter λ. We first discretize the range of D_i into an equidistant set of points d_0, d_1, …, and compute $f(m)$ = Prob[$d_m - \frac{1}{2}h < D_i < d_m + \frac{1}{2}h$], where $h = d_m - d_{m-1}$, using the PH-distribution formulas governing the random variable e^{D_i}. This is converted to a distribution on 0, 1, 2,… by a linear transformation of the discretized points. A total of 2^M points are taken for some M that makes computations by Fast Fourier Transform (FFT) convenient. Given λ, the discretized version of τ then has the density $g(d_i) = \sum_{k=1}^{\infty} P(k,\lambda)\, f^{*k}(d_i)$, where $*k$ denotes k-fold convolution, and $P(k, \lambda) = (\lambda^k/k!)\, e^{-\lambda}$ are the Poisson probabilities. We compute each convolution of the density on the integers using FFT methods and then rescale that distribution back to the original scale accumulating the values for the sum yielding the vector of values $g(.)$. To speed up computations, truncations using the Chebychev inequality both to the left and the right are made so that we only compute values above a pre-assigned threshold ε. Finally, an accuracy check is made by comparing the computed discrete distribution's mean with that resulting from the theoretical formula derived using the known results for the mean of the log-PH distribution [1]. We omit the details as these are routine. For an appropriate value of λ, the random variable τ can be interpreted as the annual downtime. A particular question of interest is the following. In the presence of very long downtimes (due to long detection and/or restoration times), what is the tradeoff between the cost to a service provider of reducing failure rates and the cost of speeding up the detection of failures and restoration of service? These two methods are compared using the 95th and 90th percentiles of the annual downtime distribution. Table 1 shows the results of decreasing λ.

Table 1. Reduction by changing λ

Percentile	λ		
	120	90	60
95%	1	0.78	0.55
90%	0.91	0.70	0.49

Next, for given thresholds A we modeled the speed-up of detection and restoration by using the log-PH distribution function $\Pr\{\tau < x\}$ for $x \leq A$ and an exponential distribution with a given mean m for $\Pr\{\tau > A + x \mid \tau > A\}$. The respective results for different values of m and A are given in Table 2. (Tables 1 and 2 are illustrative of real systems but are not drawn from any existing system.)

Table 2. Reduction by changing A and m (λ=120)

m, hours	Percentile	A, hours		
		12	8	4
8	95%	0.74	0.74	0.82
	90%	0.71	0.71	0.78
4	95%	0.66	0.62	0.57
	90%	0.63	0.59	0.55

.

We see that the reduction of the 95th (90th) percentile of the distribution of the annual downtime by half can be achieved either by reducing the failure rate λ by half (see Table 1) or by a very aggressive reduction of detection and restoration time (see Table 2), the cost of which is much higher for the service provider.

5. FUTURE RESEARCH

Our work has raised a number of research questions aimed at a deeper understanding of reliability of hierarchical redundant systems in the presence of silent failures. The applicability of traditional availability analysis based on simple averages is greatly reduced in the presence of failures with varying impacts, and the presence of heavy-tailed distributions of restoration time after silent failures. Instead, we need to introduce appropriate weights varying with the hierarchical level of a failure, and we need to study in-depth the entire downtime distribution. We must also 1) validate the absorbing-state Markov model for the uptime derivation with silent failures of the hot/standby unit, and 2) validate the exponential approximation for the uptime distribution for model parameters of practical interest. There are also questions related to our log-PH distribution of downtime. Can we obtain any mathematical insights into the quality of the estimator of the tail decay parameter? What are good methods to estimate it from data as part of the fitting procedure? Having used one such method, can we force the log-PH fit to result in tail decay equal to the value obtained as the estimate? In-depth studies of the log PH class of distributions from the perspective of extreme value theory would become a worthwhile addition to the literature. The log PH approach provides a way to re-examine with improved models many networking issues which we have considered in the past only with ad-hoc procedures.

6. REFERENCES

[1] Ahn, S., Kim, J.H.T. & Ramaswami, V. 2012. A new class of models for heavy tailed distributions in finance and insurance risk, *Insurance: Mathematics and Economics,* 51, 43-52.

[2] Asmussen,S, Nerman, O. & Olsson, M. 1996. Fitting phase type distributions via the EM algorithm, *Scandinavian Actuarial J. of Stat.,* 23 (4), 419-441.

[3] Britney, R.R., The reliability of complex systems with dependent subsystems failures: An absorbing Markov Chain model, Technometrics, 16, NO. 2, pp.245-250, 1974.

[4] Kari, H.H., Latent sector faults and reliability of disk arrays, Doctor Thesis, Espoo, Finland, 1997.

[5] Kogan, Y. 2010. Reliability Metrics for Routers in IP Networks. In *Guide to Reliable Internet Services and Applications*, C.R. Kalmanek, S. Misra, Y.R. Yang, Eds. Springer, 97-112.

[6] Kogan, Y., and Zhang, X. 2012. System MTBF and its application to IP routers. Accepted by *18th ISSAT International Conference*, July 26-28, 2012, Boston. (Available from the author.)

[7] Meyer, J.F. 1992. Performabilty: a retrospective and some pointers to the future, *Performance Evaluation*, 14, 139-156.

[8] Oikonomou, K.N. 2010. Network performability evaluation, in *Guide to Reliable Internet Services and Applications*, pp. 113-135, eds. C.R. Kalmanek, S. Misra, Y.R. Yang, Springer.

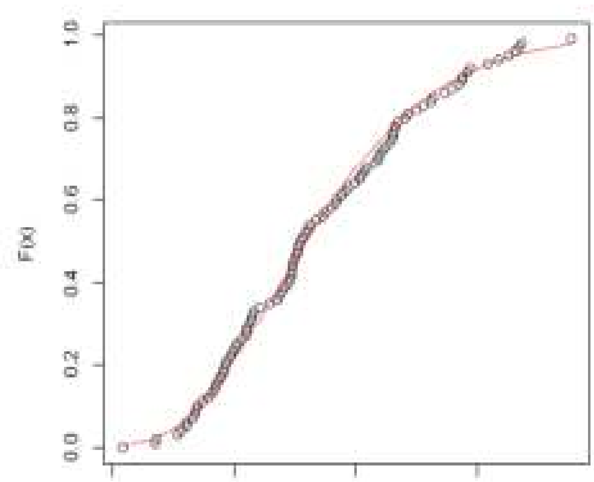

Figure 3: Empirical and fitted CDF

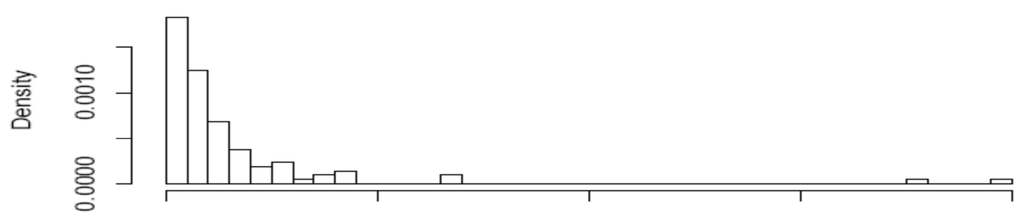

Figure 1: Empirical downtime distribution

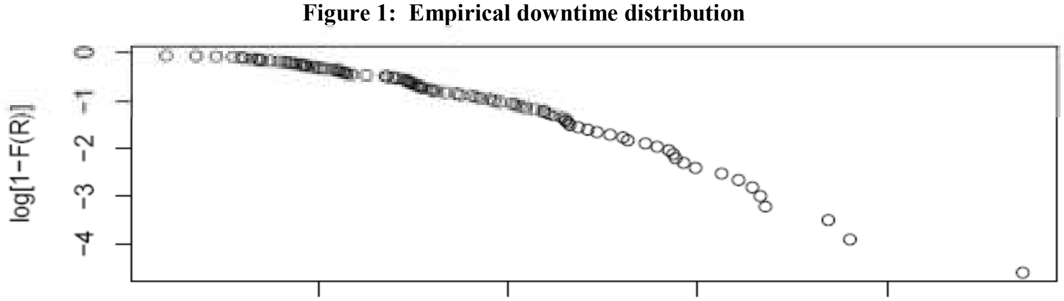

Figure 2: The log log plot of the downtime distribution