Modeling of User Perceived Webserver Availability

Wei Xie*, Hairong Sun†, Yonghuan Cao* and Kishor S. Trivedi*
{wxie, hairong, ycao, kst}@ee.duke.edu
* Center for Advanced Computing and Communications
Department of Electrical and Computer Engineering
Duke University, Durham, NC 27708
† High Reliability and Availability Technology Center
Motorola, Elk Grove Village, IL 60007, USA

***Abstract*— We propose to use Markov regenerative process (MRGP) models to study the availability of Internet-based services perceived by a Web user, which capture the interactions between the service facility and the user. The necessity of the sophisticated MRGP modeling is evidenced by the comparisons with the corresponding continuous time Markov chain (CTMC) models, which show that the popular convenient CTMC models tend to overestimate user-perceived service unavailabilities by 26% to 125%.**

We study two different online service scenarios: (1) single-user-single-host and (2) single-user-multiple-host. It is found that user-perceived service unavailability depends not only on the infrastructure's failure-recovery characteristics but also, more importantly, on the user's behavior. Also, for a service provider, to improve users' satisfaction, inventing a fast recovery mechanism is more effective than striving for a more reliable platform given the platform availability is the same.

***Index Terms*— User-perceived online service availability, Web user behavior, Markov regenerative process (MRGP)**

I. Introduction

The trend of e-commerce poses an increasingly imperative demand on the availability and reliability of the Internet-based services. The so-called "24×7" (24-hours-a-day-and-7-days-a-week) requirement for online services presents an unprecedented technical challenge given the fact that the exponentially growing Internet is of such a large-scaled, vastly distributed and heterogeneous nature. To design high-availability (HA) service systems, it is critical to deepen our understanding of not only the causes of the failure-and-recovery behaviors of the service infrastructure, but also the users behaviors and their subjective perceptions and reactions to the provided services. There have been separated research efforts on *either* behaviors of the underlying infrastructures *or* the online user behaviors. However, the lack of effort connecting the two is obvious. This paper intends to fill this gap by providing a more complete modeling for online service availability that is a result of the interactions between service platforms and users.

This research was supported in part by the Air Force Office of Scientific Research under MURI Grant No. F49620-00-1-0327, and in part by DARPA and US Army Research Office under Award No. C-DAAD19 01-1-0646. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and does not necessarily reflect the view of the sponsoring agencies.

This work was done while K. Trivedi was a visiting Professor in the Department of Computer Science and Engineering holding the Poonam and Prabhu Goel Chair at the Indian Institute of Technology, Kanpur.

The unavailability of the Internet-based services stems from various type of failures, malfunctions, and planned outages from a broad range of network components, service provider equipments, and user accessing facilities. Govindan *et al.* revealed that both the route availability and the mean reachability duration have degraded with the Internet growth [1]. Li *et al.* studied Webserver aging phenomenon and proactive software rejuvenation techniques [2]. Long *et al.* evaluated mean time to failure (MTTF), mean time to repair (MTTR), and availability and reliability of a sample of hosts by repeatedly polling the hosts and discovered that daily and weekly shutdowns appeared very commonly in the Internet [3]. By periodically collecting data on a set of nearly 100 popular Web sites, Kalyanakrishnan *et al.* in [4] found that the mean availability of Internet hosts is two-nines, *i.e.*, about 0.99, which is far below that of telephone systems. The aforementioned research efforts all focused on the study of platform outage-recovery of Internet-based services. However, for a particular Web user, a more important performance index is service availability perceived by himself, the probability that the users service request is fulfilled. Studying the platform availability alone is apparently inadequate for this purpose. We have yet to characterize the behavior of online users and reveal the interplay between the service platform and users.

It is widely accepted that the behavior of Web browsers is fairly complicated. Deng of then-GTE lab proposed a tractable empirical model, which was able to capture the behavior of World-wide-web (WWW) browsers [5]. The activity of a Web browser is modeled as an ON-OFF process, with the ON period having a Weibull distribution and the OFF time following a long-tailed Pareto distribution. ON periods are initiated by the users clicking on the hypertext links on a Web service page while OFF periods are those in which the user is reading and/or thinking and hence no requests is generated. In this study, we adopt this model as the starting point of user behavior modeling.

The purpose of this paper is to evaluate the service availability for the Web users. We assume that the time to failure (TTF) and the time to repair/recovery (TTR) of the Internet

hosts are exponentially distributed, and leave the more general case to future work. The behavior of Web user is captured by Deng's ON-OFF model. Due to the non-exponential Pareto and Weibull distributions in users' behavior, Markov regenerative process (MRGP) is needed to model the interactions between a Web user and the supporting platform. We then solve the MRGP model for steady-state probabilities, from which the service availability expression is derived.

This paper is organized as follows: In Section II we discuss the failure and recovery in the Internet, the retry mechanism of HTTP 1.1, and the user behaviors. Section III describes the user-perceived MRGP availability models of two different scenarios, and a brief introduction to solution techniques and availability indices. Numerical results and detailed discussion of the two MRGP models are included in Section IV. Comparisons with corresponding CTMC models are included as well. Section V is the conclusion.

II. Background and Related Work

To proceed, we would like to use the terms *platform*, *infrastructure*, and *system* interchangeably to refer to the underlying infrastructure that supports the online service.

Following [6], we classify platform failures into three types: "near-user", "in-middle", and "near-host". A failure in "near-user" portion, which typically is the user's subnet, disallows the user to access the rest of the Internet. Analogously, a "near-host" failure makes the Web host unreachable from the outside world. The "in-middle" failure usually refers to the Internet backbone connection malfunctions that separate the user and the specific service host, but the user still may visit a non-trivial part of the Internet. In this study, we use a unified failure-recovery model that assumes time to failure (TTF) and time to recover (TTR) are exponentially distributed for all of the three cases discussed above. The model thus can be parameterized with MTTF and MTTR, *i.e.*, λ^{-1} and μ^{-1}, respectively. To study failures from different portions in the platform, we may simply vary parameters to reflect the difference in failure behaviors in different parts of the supporting platform. The steady-state unavailability of the platform can be simply given (*e.g.*, see [7]) as $\overline{A_S} = \lambda/(\lambda+\mu)$.

HTTP 1.1 (Hypertext Transfer Protocol) [8], the major protocol for WWW services, has an automatic retry and recovery mechanism to achieve reliable data transfers. When the destined Web site is unreachable, the user is unaware of the problem although his request is not fulfilled during the HTTP retries. This is a temporary interruption of service because the remote Web host may be reconnected during this time. If the HTTP retry fails, an error message is explicitly returned to the user, and the user knows the existence of the problem and may switch to another Web site or try again later.

An empirical and tractable ON-OFF model of Web user behavior was proposed by Deng [5]. The ON period follows a Weibull distribution with the cumulative distribution function (cdf) $F(t) = 1 - e^{-(t/\theta)^k}$. Typically, the shape parameter k = 0.77 to 0.91 and the scale parameter θ = $e^{4.4}$ to $e^{4.6}$ as in [5]. The duration of OFF period follows a general Pareto distribution with the cdf $G(t) = C(1 - (\frac{m}{t})^\alpha), m \le t \le n$, and 0, otherwise, where α, m, n are constants, with typical values are α = 0.5 to 0.9 (α is the "shape parameter" of the Pareto distribution), $m = 60$, and $n = 6000$. The constant m is the so-called "ON-OFF threshold" (m is also referred as the *scale parameter* of the general Pareto distribution) which means a series of requests with inter-arrival times within m is considered constituting an ON period, and a request occurs more than time m after the previous request marks an OFF period. The constant n is the "session threshold", which indicates that requests with inter-arrival times greater than n are considered as belonging to separate sessions. $C = 1/(1 - (\frac{n}{m})^\alpha)$ is the normalization factor.

III. Service Availability MRGP Models

In this paper, we build two different service availability models for two different scenarios.

- Model-1: Single-User-Single-Host
  The first scenario includes one online user and one service host. The user is *dedicated* to the host, meaning that the user will not switch to other sites even in the presence of service outage.
- Model-2: Single-User-Multiple-Host
  The second scenario includes one online user and a large number of hosts, each of which provides a certain service that the user is interested in. The user in this case is not dedicated, *i.e.*, in the event of service failure, the user is able to switch to another service site. The failure and recovery characteristics of all the hosts in the pool are assumed to be homogeneous. In fact, the MTTF and MTTR parameters used in the next section are the average of a sizeable group of real world Internet data.

A. Model-1: Single-User-Single-Host Model

As mentioned above, only one Web user and one Web host are involved in this model. In real life, this corresponds to the case that the failure occurs in user's subnet or in the *only* Web site of interest, and the user cannot or does not access any web sites until the failure is recovered. The failure in user subnet, or near-user failure, may be caused by link disconnection, or DNS failure, or congestion in the subnet gateway, and the Web host failure, or near-host failure, may be caused by overloaded server, expected/unexpected server outages, *etc.*

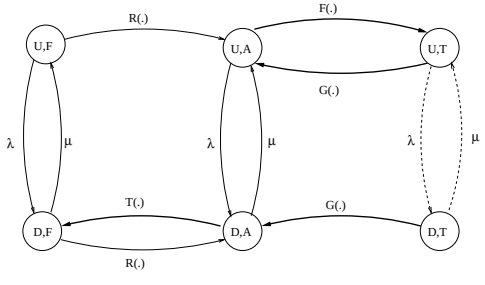

Fig. 1. Service Availability Model-1: Single-User-Single-Host

Fig. 1 depicts the MRGP service availability model for the user's subnet failure case. The circles in Fig. 1 represent the

states of our model, and the arcs represent state transitions. Each state is denoted by a 2-tuple (s, u), where $s \in \Omega_S$ (the state space of the platform) and $u \in \Omega_U$ (the state space of the user status). $\Omega_S = \{U, D\}$ includes the situations that underlying system is up and down, respectively, and $\Omega_U = \{T, A, F\}$ contains the user status of thinking, active, and seeing a failure, respectively. Our model's state space $\Omega = \{(U,T), (D,T), (U,A), (D,A), (U,F), (D,F)\}$ is the combination of Ω_S and Ω_U. 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 is active for certain amount of time, which has a distribution of $F(.)$, s/he enters thinking state (from (s, A) to (s, T)), and comes back to active (from (s, T) to (s, A)) after some time (with distribution $G(.)$). If s/he is active and the network is down (state (D, A)), the browser retries after some time with distribution $T(.)$. 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 re-attempts 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. 1), hence the model described above is not a continuous-time Markov chain (CTMC) nor is it a semi-Markov process (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 distributed. They are called "concurrent exponential transitions" (represented by dashed thin arcs), and their occurrences are just local behaviors. Refer to [9], [10] for definitions of these MRGP concepts. In this paper, we assume the user retry distribution $R(.)$ is exponential with rate r. [1]

B. Model-2: Single-User-Multiple-Host Model

In this model we consider the failures of individual Web hosts (near-host failures) which are located on the other side of the Internet, and the in-middle failures. After such a failure occurs, the user may switch to another Web hosts after some time which is assumed to be exponentially distributed with a mean of $1/\gamma$.[2]

The MRGP model is shown in Fig. 2, which resembles the previous one in Fig. 1. One of the differences is that we do not have state (U, F), *i.e.*, the user may visit another Web site after his browser's HTTP retry fails (transition from (D, F) to (U, A)). Since the unavailability of all Web hosts is $\overline{A}_S$, the chance the user sees another bad Web host is $\overline{A}_S$ (transition from (D, F) to (D, A)), and that for a good one is $(1 - \overline{A}_S)$ (transition from (D, F) to (U, A)).

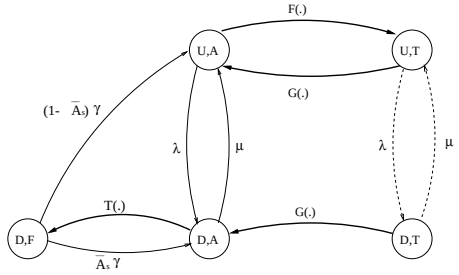

Fig. 2. Service Availability Model-2: Single-User-Multiple-Host

C. Model Solution Techniques and Unavailability Indices

We are interested in the steady-state unavailability perceived by the online user. The following procedure [9], [11], [10] is used to obtain the steady-state probabilities for the MRGP models.

1) An MRGP model is governed by its local and global kernels, $\mathbf{E}(t)$ and $\mathbf{K}(t)$. So the first step is to construct $\mathbf{E}(t)$ and $\mathbf{K}(t)$ of the MRGP.
2) To calculate $\boldsymbol{\alpha} = \int_0^\infty \mathbf{E}(t)dt$, and $\boldsymbol{\mu} = \boldsymbol{\alpha}\mathbf{e}^T$, where $\mathbf{e}$ is a row vector with all elements of 1.
3) To obtain the one-step transition probability matrix $\mathbf{K}(\infty)$, and solve $\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{K}(\infty)$ for $\boldsymbol{\nu}$, where $\boldsymbol{\nu}\mathbf{e}^T = 1$.
4) Steady-state probability vector is given by $\mathbf{P} = \frac{\boldsymbol{\nu}\boldsymbol{\alpha}}{\boldsymbol{\nu}\boldsymbol{\mu}}$.

We are interested in the unavailability the user experiences, *i.e.*, the fraction of time during which the user cannot send requests (state (s, F)) or his requests are not fulfilled (state (D, A)) divided by the fraction of time that the user is seeking service (state (s, A) and (s, F)). This user-perceived unavailability is denoted by

$$\overline{A}_U \equiv \begin{cases} \dfrac{P_{U,F} + P_{D,F} + P_{D,A}}{P_{U,F} + P_{D,F} + P_{D,A} + P_{U,A}}, & \text{for Model-1,} \\ \dfrac{P_{D,F} + P_{D,A}}{P_{D,F} + P_{D,A} + P_{U,A}}, & \text{for Model-2.} \end{cases}$$

D. Comparison with CTMC Models

The traditional and convenient CTMC models are often used in stochastic analysis mostly for simplicity reasons. Cao *et al.* found that under certain conditions, an all-exponential model is a good approximation of a stochastic model with general distributions [12]. However, CTMC is not a good substitution of MRGP model in our case as we will see in the next Section.

For comparison purpose, we also construct and solve the corresponding CTMC models of Model-1 and Model-2, *i.e.*, replacing all the general distributions with exponential distributions with the same means (in other words, our original MRGP models are simplified to CTMC models). We denote the user-perceived service unavailability of the CTMC models by $\overline{A'}_U$. As we will see in the numerical analysis, the convenient all-exponential assumption may predict user-perceived

[1] If the user retry time is generally distributed, our model is still an MRGP because there is still only one general transition enabled in any single state.

[2] As in Model-1, even if the user switch time is generally distributed, our model is still an MRGP.

TABLE I

SUMMARY OF MODEL PARAMETERS

Parameter	Default Value	Comment
k	0.88	Shape parameter of Weibull distribution
θ	$e^{4.5}$	Scale parameter of Weibull distribution
α	0.5	Shape parameter of Pareto distribution
m	60 seconds	Scale parameter of Pareto distribution (ON-OFF threshold)
n	6000 seconds	Truncation point of Pareto distribution (session threshold)
$1/r$	100 seconds	Mean time between user retries upon failure
$1/\gamma$	100 seconds	Mean time to switch site upon failure
T	10 seconds	HTTP retry time
$1/\lambda$	10^5 seconds	Platform MTTF
$1/\mu$	2581 seconds	Platform MTTR

unavailability inaccurately. The over- or under-estimation is indicated by $\vartheta = (\overline{A'}_U - \overline{A}_U)/\overline{A}_U$.

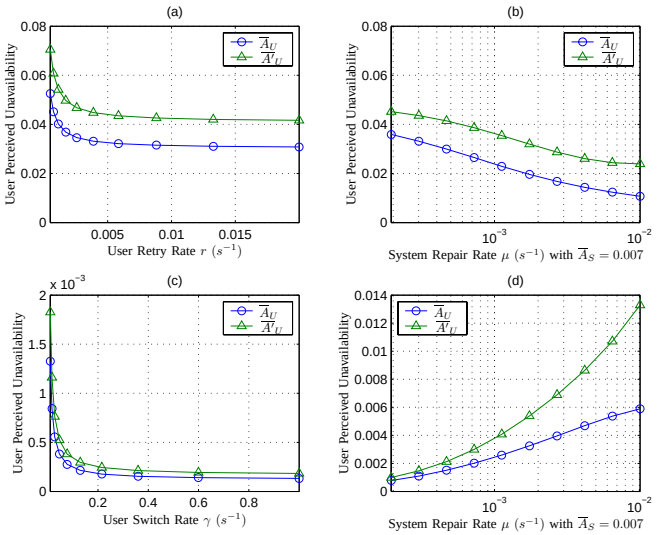

Fig. 3. User-Perceived Unavailability $\overline{A}_U$, $\overline{A'}_U$ with Platform Unavailability $\overline{A}_S = 0.007$. (a)(b) for Model-1, (c)(d) for Model-2.

IV. NUMERICAL ANALYSIS

A. Parameters Used

We tabulate the parameters with default values used in numerical solution in Table I partly based on [5] and [13]. The mean time between user retries r^{-1} and mean time to switch Web site γ^{-1} upon failure are assumed to have default value of 100 seconds.

The platform unavailabilities and intervals between failures or outages and the durations of the outages (repair or recovery) of the supporting platform from different studies appear to vary in fairly large ranges. It is found that the average platform unavailability $\overline{A}_S$ is from 0.0036 to 0.07, the typical MTTF λ^{-1} is 346,982 seconds (about 4 days), and the MTTR μ^{-1} is from 200 seconds to 2581 seconds (about 4 minutes)[6], [4].

In the following evaluation, we do not explicitly differentiate the failure and repair rates of various parts in the supporting platform, such as user's subnet, routers between subnets, average Web hosts, or commercial Web hosts. Instead, broad spectrums of λ, μ, and $\overline{A}_S$ are used to accommodate the wide ranges of failure and repair rates.

B. Model-1 Results

One might think that the user-perceived service unavailability is as simple as the probability that the platform being unavailable $\overline{A}_S$. This is not true as shown in Fig. 3, in which $\overline{A}_S = 0.007$, because the user's behavior is coupled with the platform failure-recovery process. As seen in Fig. 3, user-perceived unavailability could be several times larger than system unavailability, due to the fact that the user's behavior is not independent of the platform behavior. Whenever the user sees a failure, s/he is no longer able to go back to "thinking" states ((s,T)). Instead, s/he enters (s,F) states, assuming the service is not accessible. This actually extends the failure duration that user experiences, and makes the user-perceived unavailability higher than the system unavailability.

The relationship between the user-perceived unavailability and user retry rate is given by Fig. 3(a). If the user retries more frequently (r becomes higher), the user wastes less time idling after the system is recovered, resulting in a lower $\overline{A}_U$. The shape of $\overline{A}_U$ curve in Fig. 3(a) indicates that when the user retry rate r is small, a small increase of r can bring a significant improvement of the availability seen by the user. However, indefinitely increasing the user retry rate after a certain value may no longer decrease $\overline{A}_U$ significantly, since retries before the completion of repair or recovery are useless in terms of improving the user-perceived availability. More retries means more unsuccessful HTTP attempts, and more time wasted. The total unavailability $\overline{A}_U$ is nearly independent of r when r is sufficiently large. Fig. 3(a) shows that a retry rate of 0.003 to 0.005, or mean time to retry of about 200 seconds to 300 seconds, is a good candidate.

Fig. 3(b) deserves more attention. In this scenario the system unavailability is set to a constant 0.007 while the failure rate λ and repair rate μ vary accordingly. Note that when $\overline{A}_S$ is small, $\lambda \approx \mu\overline{A}_S$, which implies a higher failure rate requires a higher repair rate to maintain the total unavailability level. Fig. 3(b) illustrates that given a $\overline{A}_S$, increasing μ will lower $\overline{A}_U$. In other words, if there are two systems with the same availability, the one with shorter MTTF and MTTR looks better from a user's perspective than the other. This conclusion is not evident. Our analysis indicates that the originality of this characteristic is complex, including the asymmetry of the underlying process. This finding suggests a valuable strategy of maximizing the customer satisfaction with limited resources.

Fig. 3(a)(b) showed that $\overline{A'}_U$ (user-perceived unavailability predicted by corresponding CTMC model) is about 26% to 124% larger than $\overline{A}_U$, *i.e.*, the corresponding CTMC model overestimates the user-perceived unavailability by a significant percentage. We note that when μ is higher, the CTMC model gives a more pessimistic result, and when μ is unchanged, the overestimation appears unchanged. This μ-dependency is the

result of underlying process asymmetry, the ON-OFF threshold m and the special shape of the Pareto pdf. When μ becomes larger, the overestimation of the competence of $G(.)$ worsens. This result justifies the necessity and importance of MRGP modeling.

C. Model-2 Results

As shown in Fig. 3(c)(d), the user-perceived unavailability $\overline{A}_U$ of Model-2 ($\overline{A}_U$ is from 2% to 84% of $\overline{A}_S = 0.007$) is much smaller than that of Model-1. In other words, in most cases only a small fraction of the system unavailability is seen by the user, while in Model-1, the user-perceived unavailability is several times larger than the platform unavailability. This is because in the single-user-multiple-host case, the user may switch to other Web sites without waiting for the completion of repair or recovery.

This correlation of $\overline{A}_U$ and γ is given in Fig. 3(c). Apparently, when the user waits for less time to switch Web site, the probability that s/he receives service from another Web host is relatively high ($A_S = 1 - \overline{A}_S$). As a result when γ rises, $\overline{A}_U$ drops considerably. If we keep $\overline{A}_S$ unchanged and increase μ, λ is increased almost proportionally, and Fig. 3(d) demonstrates that $\overline{A}_U$ grows too. Although this appears radically different from Fig. 3(b) at the first glance, it stands on its own reason. Again, in Model-2, since the user can switch to another Web site quickly when s/he encounters a failure, the repair rate μ does not play a role as important as the failure rate λ, given $\overline{A}_S$ is fixed. Usually the user does not know (or care) the status of the failed Web host as long as s/he has the choice of going to other Web sites. However, a higher failure rate λ for all Web hosts does impact the user's visiting directly. Thus in Fig. 3(d), higher μ means higher λ, which leads to higher $\overline{A}_U$.

Similar to Model-1, Fig. 3(c) and 3(d) show that the all-exponential counterpart of the MRGP model overestimates the user-perceived unavailability by about 28% to 125%.

V. Conclusion

Although many efforts have been dedicated to identifying causes of the Internet unavailability [6], [4], [13], [3] and statistically quantifying activities of online users [5], there have been very few studies that connect these two and analyze the Web service availability from an end user's perspective. In this paper, we have developed Markov Regenerative Process (MRGP)-based models that incorporate both the failure-recovery behaviors of the service-supporting infrastructure and the online user behaviors, and evaluated 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 have found that the corresponding all exponential models (*i.e.*, replacing all general distributions by exponential distributions with same means) overestimated the user-perceived unavailability by about 26% to 125%. This substantial difference shows that the oversimplified but popular all-exponential assumption, which under certain conditions serves as a good approximation, is inadequate in our case, and MRGP modeling is necessary and important.

The MRGP models are solved numerically for steady-state probabilities, and the user-perceived unavailability was found to be very different from the system unavailability. For Model-1, the ratio of user-perceived unavailability and the platform unavailability is much larger than that of Model-2. This disparity of the two models is because in Model-2, the user may switch to another Web host without having to wait for the repair of a failed Web host.

From the models, we have also discovered some interesting properties of user-perceived unavailability of online services that may be useful for online service providers/designers as well as end-users.

- For Web end users, they should try to switch to other Web site first when seeing a failure. If for some reason s/he has to stay with the failed site, do not wait too long to retry. On the other hand, unreasonably frequent retries do not help in improving the user-perceived availability. Our analysis showed that a mean time to retry from 200 seconds to 300 seconds seems appropriate with given parameters.
- For Web site/subnet owners, fast recovery is more effective than high reliability. Our analysis indicated that 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.

Our future work includes comparisons of the aforementioned analytical results with those from testbed experiments and trace-driven simulations.

References

[1] R. Govindan and A. Reddy, "An analysis of Internet inter-domain topology and route stability," in *INFOCOM'97*, 1997, pp. 850–857.

[2] L. Li, K. Vaidyanathan, and K. Trivedi, "An approach to estimation of software aging in a web server," in *ISESE 2002*, Nara, Japan, Oct. 2002.

[3] D. Long, A. Muir, and R. Golding, "A longitudinal survey of Internet host reliability," in *Proc. of the 14th Symposium on Reliable Distributed Systems*, Bad Neuenahr, Germany, September 1995, pp. 2–9.

[4] M. Kalyanakrishnan, R. K. Iyer, , and J. U. Patel, "Reliability of Internet hosts: A case study from the end user's perspective," *Computer Networks*, vol. 31, pp. 45–57, 1999.

[5] S. Deng, "Empirical model of WWW document arrivals at access link," in *ICC'96*, 1996, pp. 1797–1802.

[6] B. Chandra, M. Dahlin, L. Gao, and A. Nayate, "End-to-end WAN service availability," in *USITS01*, Jan. 2001.

[7] K. S. Trivedi, *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, John Wiley & Sons, 2nd Ed., 2001.

[8] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, *Hypertext Transfer Protocol - HTTP/1.1*, June 1999.

[9] V. G. Kulkarni, *Modeling and analysis of stochastic systems*, Chapman Hall, 1995.

[10] H. Choi, V. G. Kulkarni, and K. S. Trivedi, "Markov Regenerative Stochastic Petri Nets," *Perf. Eval.*, vol. 20, pp. 335–357, 1994.

[11] V. G. Kulkarni, *Lecture Notes on Stochastic Models in Operations Research*, University of North Carolina, Chapel Hill, U.S.A., 1990.

[12] Y. Cao, H. Sun, and K. S. Trivedi, "System availability with non-exponentially distributed outages," *IEEE Transactions on Reliability*, vol. 51, no. 2, pp. 193–198, June 2002.

[13] B. Liu, G. Abdulla, T. Johnson, and E. A. Fox, "Web response time and proxy caching," in *WebNet98*, Orlando, Nov. 1998.