Internet Reliability and Availability Analysis Using Markov Method

Lela Mirtskhulava
Department of Computer Sciences
Iv. Javakhishvili Tbilisi State University
Tbilisi, Georgia
e-mail: lela.mirtskhulava@tsu.ge

Natela Ananiashvili
Department of Computer Sciences
Iv. Javakhishvili Tbilisi State University
Tbilisi, Georgia
e-mail: ia.ananiashvili@gmail.com

Revaz Kakubava
Department of Computer Engineering
Georgian Technical University
Tbilisi, Georgia
e-mail: r.kakubava@gtu.edu.ge

Giorgi Gugunashvili
Department of Computer Sciences
Iv. Javakhishvili Tbilisi State University
Tbilisi, Georgia
e-mail: g.gugunashvili@gmail.com

Abstract— Modeling based approach is described for analyzing and evaluating Internet server system reliability and availability in this paper. In the given model the states are defined by the different kinds of failures of the server system. The Markov model evaluates the probability of jumping from one known state into the next logical state. The probabilities between transitions of the states are a function of the failure rates of the transitional probabilities from one to another state. The number of first-order differential equations is equal the number of the states of the servers. First-order differential equations is developed by describing the probability of being in each state in terms of states of the model.

Keywords—availability; markov model; reliability; Internet; server system; MTTF; state transition diagram

I. INTRODUCTION

Today, billions of dollars are being spent annually worldwide to develop reliable and good quality products and services. Global competition and other factors are forcing manufacturers and others to produce highly reliable and good-quality products and services. Needless to say, nowadays reliability and quality principles are being applied across many diverse sectors of economy and each of these sectors has tailored reliability and quality principles, methods, and procedures to satisfy its specific need. There is a definite need for reliability and quality professionals working in diverse areas to know about each other’s work activities because this may help them, directly or indirectly, to perform their tasks more effectively [1].

The history of the Internet goes back to 1969 with the development of Advanced Research Projects Agency Network (ARPANET). It has grown from 4 hosts in 1969 to over 147 million hosts and 38 million sites in 2002. In 2000, the Internet economy generated around $830 billion in revenues in the United States alone. In 2001, there were 52,658 Internet-related incidents and failures. Needless to say, today a reliable and stable Internet is extremely important to the global economy and other areas, because Internet failures can easily generate millions of dollars in losses and interrupt the daily routine of hundreds of thousands of end users [3].

Most communication systems are sensitive to random errors and synchronization failures. The Physical Layer analysis of any Computer Communication network is very important. This is because many different problems concerning network execution and utilization are caused by errors and failures on the Physical Layer. [4]

An accurate estimation of network performance is vital for the success of a network of any kind. Networks, whether voice or data, are designed around many different variables. Two of the most important factors that you need to consider in network design are service and cost. Service is essential for maintaining customer satisfaction. Cost is always a factor in maintaining profitability. [5]

The demand for Internet reliability continues to escalate as the Internet evolves to support various applications including telephony and banking [34]. However, various studies conducted over the past decade indicate that the reliability of Internet paths falls far short of the 99.999% availability expected in the public switched telephone network (PSTN) [35]. Furthermore, small-scale studies conducted in 1994 and 2000 revealed that the probability of encountering a major routing pathology along a path is approximately 1.5% to 3.3% [34, 36]. Over the years various means have been used to improve Internet reliability, including server replication, multi-homing, and overlay networks [10].

Markov chains are a relatively simple but very interesting and useful class of random processes. A Markov chain describes a system whose state changes over time. The changes are not completely predictable, but rather are governed by probability distributions. These probability distributions incorporate a simple sort of dependence structure, where the conditional distribution of future states of the system, given some information about past states, depends only on the most recent piece of information. That is, what matters in predicting the future of the system is its present state, and not the path by which the system got to its present state. Markov chains illustrate many of the
important ideas of stochastic processes in an elementary setting. This classical subject is still very much alive, with important developments in both theory and applications coming at an accelerating pace in recent decades. [9]

II. MODEL DESCRIPTION

There are many mathematical models that can be used to perform reliability-related analysis in various areas of Internet [5, 6].

We describe Internet model for analyzing the reliability of the Internet server system. The model assumes that it represents the process of Internet server system functioning the alternation between the following five states:

- S₁ – system is in active (operating) state at the moment of time \( t \);
- S₂ – systems is operable and is periodically controlled at the moment of time \( t \);
- S₃ – system is inoperable due to occurrence of failure, detected by continuous hardware-based control;
- S₄ – system is inoperable and in process of restoration;
- S₅ – system is inoperable due to occurrence of failure and is under process of periodic control, after that it will be transferred for restoration with intensity: \( \gamma = \alpha + \beta \).

Figure 1 shows Internet System state transition diagram.

![System state transition diagram](image)

We define the following symbols:

- \( \alpha \) is system failure rate (intensity of occurrence of failures) detected only by continuous hardware (test) control.
- \( \beta \) is failure rate or intensity of failures, detected only by periodic hardware (test) control.

Time of hardware testing is exponentially distributed, where distribution function is described by the following equation:

\[
A(t) = 1 - e^{-\alpha t} 
\]

distribution function of periodic testing (control) is equal to, respectively:

\[
B(t) = 1 - e^{-\beta t} 
\]

Testing periodicity is also exponentially distributed and corresponding distribution function is described by the following equation:

\[
\Gamma(t) = 1 - e^{-\gamma t} 
\]

where \( \gamma \) is intensity of periodic testing distribution. \( Z(t) \) is distribution function of queuing system being in periodic testing, distributed exponentially and described by the following equation:

\[
Z(t) = 1 - e^{-\tau t} 
\]

During continuous hardware based control, when the system fails, a repair process is initiated. The repair process will bring the system back to functioning state and may recover its initial reliability. System is served by one repair team. During the process of periodic control, in the system may occur the same kinds of failures like in the idle state. During the recovery process, failures do not occur in the system. Additional periodic control is held during the system transitions between the idle and busy states, followed by further recovery if necessary.

Time recovery of the system is also exponentially distributed and described by the distribution function:

\[
G(t) = 1 - e^{-\mu t} 
\]

where \( \mu \) is intensity of recovery time.

Creating the equations for different states of the server system:

Let us create an equation for the state \( S₁ \) of queuing system, for these purposes we’ll consider three incompatible events. In order to keep a system in active state (\( S₁ \)) at the moment of time \( t \):

- event \( A \) – failures ought not to occur and a time of periodic control \( \tau \) ought not to come. The probability of this event is equal to:

\[
P_A = P₁(t)[1 - (\alpha + \beta + \gamma)\Delta t] 
\]

- event \( B \) – transition of the system from the state \( S₂ \) to the state \( S₁ \) takes place, if periodic control is completed at time interval \( t+\Delta t \) and the system will remain operable. The probability of the event \( B \) is described by the equation:

\[
P_B = \tau\Delta tP₂(t) 
\]

- event \( C \) – the system moves from the state \( S₃ \) to the state \( S₁ \) at time interval \( t+\Delta t \), after repair with the probability:

\[
P(C) = \mu\Delta tP₄(t) 
\]

\[
P₃(t + \Delta t) = P₁(t)[1 - (\alpha + \beta + \gamma)\Delta t] + P₂(t)\tau\Delta t + P₄(t)\mu\Delta t 
\]

So the equation of the system being in the state \( S₁ \) has the following form:
Let’s create an equation for the state $S_2$. Let’s consider the sum of the two incompatible events: A – failures ought not to take place neither with $\alpha$-intensity and nor with $\beta$ – intensity and control will not be completed.

$$P(A) = P_2(t) \left[ 1 - (\alpha_1 + \beta_1 + \tau) \right]$$ (11)

B – The system jumps from the state $S_1$ to the state $S_2$ if control will start with probability:

$$P(B) = P_2(t) \gamma \Delta t$$ (12)

$$P_2(t + \Delta t) = P(A) + P(B)$$ (13)

So the state $S_3$ is described by the following equation:

$$P_2(t + \Delta t) = P_2(t) \left[ 1 - (\alpha_1 + \beta_1 + \tau) \Delta t \right] + P_1(t) \gamma \Delta t$$ (14)

The system is in the state $S_3$ if the sum of two events take place: A – if failure with intensity $\alpha$ will not take place and the time of periodic control will not come, i.e.:

$$P(t) = P_3(t) \left[ 1 - (\gamma + \alpha) \Delta t \right]$$ (15)

B – the system moves from the state $S_1$ to the state $S_3$ if failure with intensity $\beta$ will occur, i.e.:

$$P(B) = P_3(t) \beta \Delta t$$ (16)

$$P_3(t + \Delta t) = P(A) + P(B)$$ (17)

So the equation for the state $S_3$ is described by the following expression:

$$P_3(t + \Delta t) = P_3(t) \left[ 1 - (\alpha_1 + \gamma + \alpha_2) \Delta t \right] + P_1(t) \beta \Delta t$$ (18)

The system being in the state $S_4$ could be considered as the sum of four the following events:

A – if the system is repairing and the process of repairing is not completed yet:

$$P(A) = P_4(t) \left[ 1 - \mu \Delta t \right]$$ (19)

B – the system moves from the state $S_1$ to the state $S_4$, if failure with intensity $\alpha$ will occur:

$$P(B) = P_4(t) \alpha \Delta t$$ (20)

C – the system moves from the state $S_3$ to the state $S_4$, if failure with intensity $\gamma$ will occur:

$$P(C) = P_4(t) \gamma \Delta t$$ (21)

D – the systems will move from the state $S_5$ to the states $S_4$, if the periodic control is not completed:

$$P(D) = P_4(t) \tau \Delta t$$ (22)

So

$$P_4(t + \Delta t) = P(A) + P(B) + P(C) + P(D)$$ (23)

We finally get the equation of the state $S_4$:

$$P_4(t + \Delta t) = P_4(t) \left[ 1 - \mu \Delta t \right] + \gamma P_4(t) \Delta t + P_4(t) \alpha \Delta t$$ (24)

Now let’s consider the being of the system in the state $S_5$ which can be considered by the sum of the following three events:

A – the system will be in the state $S_5$ if the periodic control will not complete:

$$P(A) = P_5(t) \left[ 1 - \tau \Delta t \right]$$ (25)

B – the system will jump from the state $S_2$ to the state $S_5$, if failure with the intensity $\alpha_1$ and $\beta_1$, i.e. the probability of the event B is described by the following equation:

$$P(B) = P_2(t) \left[ \alpha_1 + \beta_1 \right]$$ (26)

C – the system will move from the state $S_3$ to the state $S_5$, if the periodic control will take place, i.e.:

$$P(C) = P_3(t) \gamma \Delta t$$ (27)

As a result, we have the equation of the state $S_5$:

$$P_5(t + \Delta t) = P_5(t) \left[ 1 - \mu \Delta t \right] + P_1(t) [\alpha_1 + \beta_1] \Delta t + P_4(t) \gamma \Delta t$$ (28)

We finally have the following system of equations:

$$P_1(t + \Delta t) = P_1(t) \left[ 1 - (\alpha + \beta + \gamma) \Delta t \right] + P_2(t) \tau \Delta t + P_3(t) \mu \Delta t$$ (29)

$$P_2(t + \Delta t) = P_2(t) \left[ 1 - (\alpha_1 + \beta_1 + \tau) \Delta t \right] + P_1(t) \gamma \Delta t$$ (30)

$$P_3(t + \Delta t) = P_3(t) \left[ 1 - (\alpha_1 + \beta_1 + \gamma) \Delta t \right] + P_1(t) \beta \Delta t$$ (31)

$$P_4(t + \Delta t) = P_4(t) \left[ 1 - (\alpha_1 + \gamma + \alpha_2) \Delta t \right] + P_1(t) \alpha \Delta t + P_3(t) \gamma \Delta t$$ (32)

$$P_5(t + \Delta t) = P_5(t) \left[ 1 - \tau \Delta t \right] + P_1(t) [\alpha_1 + \beta_1] \Delta t + P_4(t) \gamma \Delta t$$ (33)

Taking the limit as $\Delta t \to 0$:

$$\frac{dP_1(t)}{dt} = - (\alpha + \beta + \gamma) P_1(t) + \tau P_2(t) + \mu P_4(t)$$ (34)

$$\frac{dP_2(t)}{dt} = - (\alpha_1 + \beta_1 + \tau) P_2(t) + \gamma P_1(t)$$ (35)

$$\frac{dP_3(t)}{dt} = - (\alpha_1 + \beta_1 + \gamma) P_3(t) + \beta P_1(t)$$ (36)

$$\frac{dP_4(t)}{dt} = - (\alpha + \gamma) P_2(t) + \alpha P_3(t) + \beta P_1(t)$$ (37)

$$\frac{dP_5(t)}{dt} = - (\alpha_1 + \beta_1 + \gamma) P_1(t) + \gamma P_3(t)$$ (38)

Moving to Laplace transform:

$$\mathcal{L} \left[ P_1(s) - P_1(0) \right] = - (\alpha + \beta + \gamma) \mathcal{L} \left[ P_1(s) \right] + \tau \mathcal{L} \left[ P_2(s) \right] + \mu \mathcal{L} \left[ P_4(s) \right]$$ (39)

$$\mathcal{L} \left[ P_2(s) - P_2(0) \right] = \gamma \mathcal{L} \left[ P_1(s) \right] - (\alpha_1 + \beta_1 + \gamma) \mathcal{L} \left[ P_2(s) \right]$$ (40)

$$\mathcal{L} \left[ P_3(s) - P_3(0) \right] = \beta \mathcal{L} \left[ P_1(s) \right] - (\alpha + \gamma) \mathcal{L} \left[ P_3(s) \right]$$ (41)

$$\mathcal{L} \left[ P_4(s) - P_4(0) \right] = \alpha \mathcal{L} \left[ P_3(s) \right] + \gamma \mathcal{L} \left[ P_1(s) \right] - \mu \mathcal{L} \left[ P_2(s) \right] + \tau \mathcal{L} \left[ P_5(s) \right]$$ (42)

$$\mathcal{L} \left[ P_5(s) - P_5(0) \right] = (\alpha_1 + \beta_1) \mathcal{L} \left[ P_2(s) \right] + \gamma \mathcal{L} \left[ P_3(s) \right] - (\alpha_1 + \beta_1) \mathcal{L} \left[ P_1(s) \right]$$ (43)

We have the following Boundary conditions:

$$P_1(0) = 1; P_2(0) = 0; P_3(0) = \frac{2}{5}$$ (44)

Let us consider the following case:

$$\beta = 0; \beta_1 = 0; \gamma = 0$$ (45)

We get:

$$\left\{ \begin{array}{l}
S_1 = 1 - \frac{1}{\alpha + \mu + \alpha_1 + \beta_1 + \gamma} \\
S_2 = \frac{1}{\alpha + \mu + \alpha_1 + \beta_1 + \gamma} \\
S_3 = \frac{1}{\alpha_1 + \beta_1 + \gamma} \\
S_4 = \frac{1}{\alpha + \mu + \alpha_1 + \beta_1 + \gamma} \\
S_5 = \frac{1}{\alpha_1 + \beta_1 + \gamma}
\end{array} \right.$$ (46)

Solving the latter system of equations, we obtain:

$$P_0 = \frac{1}{\alpha + \mu + s + \alpha_1 + \beta_1 + \gamma}$$ (47)

$$P_1 = \frac{1}{\alpha + \mu + s + \alpha_1 + \beta_1 + \gamma}$$ (48)

$$P_2 = \frac{1}{\alpha + \mu + \alpha_1 + \beta_1 + \gamma}$$ (49)

Moving to original in equations (47) and (48), we get:

$$P_0 = \frac{1}{\alpha + \mu + s + \alpha_1 + \beta_1 + \gamma}$$ (50)

$$P_1 = \frac{1}{\alpha + \mu + s + \alpha_1 + \beta_1 + \gamma}$$ (51)
\[ P_a = \frac{\alpha}{\alpha + \mu} (1 - e^{-(\alpha + \mu)t}) \] (39)

Availability ratio:
\[ AV = P_1(\infty) = \frac{1}{1 + \alpha \tau_r} \] (40)

Where:
\[ \tau_r = \frac{1}{\mu} \]

For \( \mu=0 \), equation (35) reduces to
\[ R_1(t) = e^{-at} \] (41)

III. CONCLUSION

Markov analysis is a powerful tool in the analyses of dynamic systems such as server systems. Markov techniques decrease the analyst's task by reducing the problem from one of mathematical computation to that of state modeling. Given Markov model breaks the system configuration into five states. Each of these states is connected to all other states by transition rates as shown in the state transition diagram. Using model reduction techniques allows to simplify model with insignificant impact on model accuracy. First-order differential equations is developed by describing the probability of being in each state in terms of states of the model. The expressions for reliability and availability were obtained. MTTF has been defined for the server system.

![Reliability graph for different meanings of failure rates](image)

Figure 2. Reliability graph for different meanings of failure rates

Where \( R_1(t) \) is the server system reliability at time \( t \). Thus, mean time to failure (MTTF) is given by:
\[ MTTF = \int_0^\infty R_1(t) dt = \int_0^\infty e^{-at} dt = \frac{1}{a} \] (42)

IV. BACKGROUND AND RELATED WORK

This section provides a brief overview of relevant background and related studies in reliability analysis and Markovian chains.

Dhillon in his book “Applied Reliability and Quality: Fundamentals, Methods and procedures”, Chapter 7 presents various aspects of computer and Internet reliability including computer system failure causes and measures, fault masking, software reliability evaluation models, Internet failure examples and outage categories, and Internet reliability models. This book presents mathematical model, which is concerned with evaluating the reliability and availability of a server system using the Markov method. The model assumes that the Internet server system can either be in an operating or a failed state and its failure/outage and restoration/repair rates are constant [1].

J Chang studied Markov chains. Markov chains are a relatively simple but very interesting and useful class of random processes. A Markov chain describes a system whose state changes over time. The changes are not completely predictable, but rather are governed by probability distributions. These probability distributions incorporate a simple sort of dependence structure, where the conditional distribution of future states of the system, given some information about past states, depends only on the most recent piece of information. That is, what matters in predicting the future of the system is its present state, and not the path by which the system got to its present state.

Markov chains illustrate many of the important ideas of stochastic processes in an elementary setting. This classical subject is still very much alive, with important developments in both theory and applications coming at an accelerating pace in recent decades [9].

The above-mentioned and the other related works analysis the reliability and availability for server system by considering only two main: operating and main states. Therefore, this paper first presents an innovative reliability model for Internet server system, which considers various types of failures that have significant influences on the success/failure of Internet Server system. They are: \( S_1 \) – system is in active (operating) state at the moment of time \( t \); \( S_2 \) – systems is operable and is periodically controlled at the moment of time \( t \); \( S_3 \) – system is inoperable due to occurrence of failure, detected by continuous hardware-based control; \( S_4 \) – system is inoperable and in process of restoration; \( S_5 \) – system is inoperable due to occurrence of failure and is under process of periodic control, after that it will be transferred for restoration with intensity: \( \gamma = \alpha + \beta \).

REFERENCES


