Downtime-Frequency Curves for Availability Characterization

Ilkka Norros, Urho Pulkkinen and Jorma Kilpi
VTT Technical Research Centre of Finland
P.O. Box 1000
FIN-02044 VTT, Finland
firstname.lastname@vtt.fi

Abstract

Availability data typically consist of start and end times of a system's down periods. We propose a natural way to plot their statistics so that Service Level Agreements concerning availability can be formulated as the condition that an empirical curve lies below a given curve.

1. Introduction

In this paper we have in mind data communications systems, whose dependability was recently considered in VTT's IPLU project [2], but most of the discussion applies more generally. The traditional way to set a quantitative requirement for the availability of a system is to give a single number like 0.99999. Such numbers are typically used in Service Level Agreements (SLA) concerning, for example, transmission links. A single-number characteristic is, however, quite uninformative: it tells nothing about the lengths of the individual downtimes, which may have great significance.

2. Definition and properties

We propose instead the use of *downtime-frequency curves* that characterize the frequency of each down-period length separately in an appropriate form. They are defined as follows.

Consider first the characterization of the reliability of a system or, similarly, availability of a resource, with binary nature: at each timepoint t, it can be unequivocally stated whether the system is up or down. Thus, its performance is described by a $\{0,1\}$-valued stochastic process:

$$I_t = 1_{\{\text{system down at time } t\}}.$$

The probability of failure, $\mathbb{P}(\text{system down at time } t) = \mathbb{E}\, I_t$ is already a characteristic of the reliability of the system. Assuming stationarity and ergodicity, this number is independent of t and obtained almost surely as the limit of the observed relative frequency:

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T I_t \,\mathrm{d}t = \mathbb{E}\, I_0 \quad \text{a.s.}$$

Let us define the ongoing down-period length at time t as

$$W_t = \inf\{s \geq t:\ I_s = 0\} - \sup\{s \leq t:\ I_s = 0\}.$$

When the system is up, we have $W_t = 0$. The relative share of time spent in down-periods lasting longer than τ during an observation period of length T is then given by the random variable

$$\varphi_T(\tau) = \frac{1}{T}\int_0^T 1_{\{W_t > \tau\}}\,\mathrm{d}t.$$

Considered as a random function of τ, $\varphi_T(\tau)$ is non-increasing. Its initial value $\varphi_T(0)$ equals the relative overall downtime of the system in the observation period (for example one year).

If the system is stationary, W_t is a stationary stochastic process, and we find that the expectation of the random function $\varphi_T(\tau)$ equals the tail distribution function of the random variable $W = W_0$:

$$F_T(\tau) = \mathbb{E}\,\varphi_T(\tau) = \frac{1}{T}\int_0^T \mathbb{P}(W_t > \tau)\,\mathrm{d}t = P(W > \tau).$$

Using this framework, we can now formulate reliability criteria that take into account the down-period lengths also: let us consider the performance of the system acceptable if

$$\varphi_T(\tau) \leq \psi(\tau)$$

for some selected function ψ. The function ψ can be specified in a SLA. The network operator has to build the system in such a way that the expected curve $F_T(\tau)$ lies sufficiently much below $\psi(\tau)$. Since the relevant values of both the down-periods and the probabilities extend over many orders of magnitude, the curves should be drawn in a log-plot or even log-log plot. When the axes are selected appropriately, the curve $\psi(\tau)$ can often be given as a straight line.

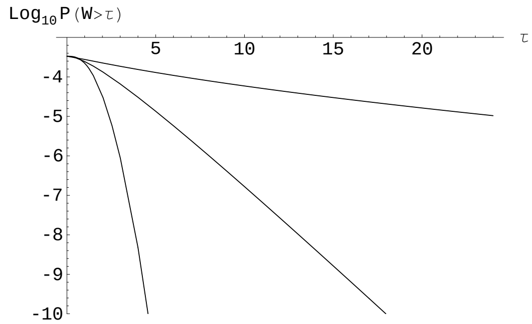

Figure 1. The expected downtime-frequency curves when the downtimes are Weibull distributed with exponents $\alpha = 0.5$ (highest curve), 1 (middle curve), 2 (lowest curve).

3 On/off-availability and quality-availability

The binary notion of availability is often insufficient for communication systems. The packet transfer may work 'reliably' in both directions but proceed with much lower rate and/or higher delays than in normal conditions. From a mathematical point of view, however, this problem can be reduced to the binary case simply by considering the *set* of binary processes

$$1_{\{q(S_t)\le r\}}, \quad r \in R,$$

where S_t is the system state at time t, q is some characteristic of it (rate, delay,...), and R is the set of possible or relevant values of that characteristic.

For example, an SLA may require that the bandwidth of an MPLS path be higher than 50 Mbit/s with an availability of at least 0.999. Then, the set R may contain the value 50 Mbit/s alone. However, since IP-based services are usually quite flexible with respect to bandwidth requirements, it would make sense to require additionally that the availability of 5 Mbit/s be at least 0.99999.

One can also, at least in principle, let R be a whole interval and replace the binary-case criterion that the empirical values should lie below a curve to the two-dimensional criterion that they should lie below a surface. If higher $q(\cdot)$ means better quality, the monotonicities behave similarly in both dimensions if r is replaced by some inverse parameter β by writing, for example, $r = 1/\beta$.

4. Examples

A standard mathematical model of this kind of process is the alternating renewal process, where the up- and down-periods are independent random variables with distributions G_{up} and G_{down} and means μ_{up} and μ_{down}, respectively. (The resulting formulae can in fact be generalized using the Palm theory of stationary processes, see [1].) When I_t is a stationary version on an alternating renewal process, the distribution of W is

$$\mathbb{P}(W > \tau) = \frac{1}{\mu_{\mathrm{up}} + \mu_{\mathrm{down}}} \int_{\tau}^{\infty} y\, G_{\mathrm{down}}(\mathrm{d}y).$$

Note that the distribution G_{up} has an effect only through the expectation μ_{up}.

Here is a formal example of such plots. Assume that the down-periods and up-periods are independent, time unit is one hour, the up- and down-periods have means 3000 and 1, respectively, and the down-period length has a Weibull distribution

$$1 - G_{\mathrm{down}}(y) = \exp(-\beta y^{\alpha}),$$

where α and β are parameters. The choice $\beta = \Gamma(1 + 1/\alpha)^{\alpha}$ yields the desired mean 1. We can now compute and plot the functions $F_T(\tau)$ for three qualitatively different parameter values $\alpha = 0.5$, 1 and 2. This example also illustrates the usability of linear, log-linear and log-log plots for various purposes.

As an example how empirical data might look like in this framework, assume that the downtimes of a system within a year consist of intervals with lengths 2, 2, 2, 3, 3, 5, 5, 6, 7, 9, 25, 35, 240 minutes (in ascending order). The empirical tail distribution function of W is then determined by the points marked as triangles in Figure 2. Note that the few long down-periods have the effect that the whole point set looks almost horizontal. The other three point sets show the corresponding plot when 1, 2 and 3 largest values are removed from the data set, respectively.

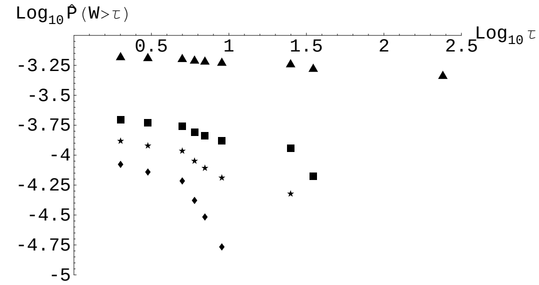

Figure 2. Data plot example (see text).

References

[1] F. Baccelli and P. Bremaud. *Elements of Queueing Theory*. Springer Verlag, Berlin, 2003.

[2] IPLU project homepage. http://iplu.vtt.fi.