Probabilistic estimation of response times through large deviations

Nicolas Navet
LORIA-INRIA
Vandoeuvre, France
nicolas.navet@loria.fr

Liliana Cucu
LORIA-INPL
Vandoeuvre, France
liliana.cucu@loria.fr

René Schott
IECN-LORIA
Vandoeuvre, France
schott@loria.fr

Abstract

We apply large deviation theory to assess the probability that the average, or the sum, of the response times of a sequence of consecutive aperiodic jobs is below a given threshold. This coarse-grained performance metric is for instance adapted to evaluate the responsiveness of a soft-real system or the freshness of input data consumed by an algorithm. The technique proposed works with distribution of response times as input but does not require that the distribution obeys a closed-form equation. Indeed, it can accept empirical distributions given under the form of frequency histograms obtained, for instance, by monitoring the system. Future work should be devoted to further assess the applicability of the proposal and relax some technical assumptions.

1 Introduction

Context of the study In the field of real-time systems, the real-time performances of periodic activities (tasks, messages) have been extensively studied. Response times, be they worst-case or average, and jitters can be evaluated by simulation or analysis but it requires that the activation model of the tasks and its parameters are known. The problem is quite different for aperiodic activities since, in many practical cases, it is not possible to have a precise knowledge of the activation pattern before the implementation of the system. For example, this is generally the case for aperiodic frames, usually subject to soft real-time constraints, exchanged among the Electronic Control Units (ECUs) in the body network of a vehicle. Such frames are usually assigned a low priority, and thus they do not delay the hard real-time periodic or sporadic traffic. However their own response times are difficult to estimate.

Problem definition In this paper we discuss the problem of evaluating the real-time performances of aperiodic activities. The metric of interest here is the average response times of successive activations of an activity. Activities are termed tasks in the following, but what is said equally holds for frames. We want to estimate the probability that the average response times of the aperiodic tasks remain below a given threshold. We are generally here in the realm of soft real-time constraints, but their satisfaction is important since large response times may jeopardize the execution of a function, and may even raise safety concerns in some cases (e.g. headlights flashes in a vehicle). In addition, low responsiveness is negatively perceived by the user. It is worth mentioning that activities that are periodic per essence are sometimes implemented in an aperiodic manner in order to save resources. For instance, in some control systems, a frame is transmitted only if the value it contains belongs to a certain interval.

Overview of our approach We do not assume any knowledge of the aperiodic tasks activation pattern, however we assume that it is possible to monitor the system, or a detailed simulation model of it, and gather data about the response times. Precisely, from the measurements, we build a frequency histogram of the response times that will be used later on as an empirical response time probability distribution for a single aperiodic task. Now, the problem is to assess the average response times over a set of successive activations of the same task. This can be done by Monte-Carlo simulation or analytically. The latter option is developed in the present paper by applying results from Large Deviation (LD) theory. The interest of LD with regard to Monte-Carlo simulations is threefold. First, simulation is not well suited to estimate rare events (e.g., less frequent than 10^{-4}) because of the size of the sample that is needed to achieve reasonable error bounds[1]. Second an analytical approach does not suffer the uncertainties of simulation (e.g.,

[1] Central Limit Theorem tells us that the convergence rate is of order $n^{1/2}$ where n is the number of random draws, which means that adding one significant digit requires increasing n by a factor 100.

quality of the random number generators). Finally, results presented here can be integrated into a broader probabilistic temporal analysis.

Existing work Our approach belongs to the class of stochastic analyses for real-time systems. Probabilistic approaches are promising because they answer questions that cannot be addressed in a deterministic manner (e.g., distribution of response times) and consider models that are more realistic, for instance, regarding the task activation patterns or the way to express soft real-time constraints.

These approaches can be classified in two main classes. One class consists in extracting quantitative information for one or more parameters (e.g., distribution of the execution time) from samples of observations collected by monitoring the system [5]. Such approaches actually belong more to the realm of statistical methods. The other class of stochastic approaches concerns the temporal analysis of systems that have at least one parameter being a random variable. Among the studies in this area, one can for instance mention [3, 1, 9, 6, 4].

To the best of our knowledge, there is no study on aperiodic task response times, which is the focus of this study. In addition, another novelty with respect to previous work comes from the use of LD, which to our knowledge has never been applied to real-time systems. LD is a theory of rare events that is focused on the analysis of tails of probability distribution, and is classically used to study how random processes deviate from their expected value. If upper bounds on this quantity can be obtained through Chernov, Markov and Tchebychev inequalities, LD provides the exact rate of convergence, and not an upper bound that is often not tight enough for real-time applications. LD has been a very active field of investigation over the last 10 years with numerous practical applications, for instance for evaluating performance of algorithms or telecommunication infrastructures [7] or assessing the risks in finance [8, 10].

Contribution of the paper We apply large deviation theory to assess probability bounds on the average response times, or sum of the response times, of successive instances of aperiodic tasks. We provide the analysis enabling to deal with empirical distributions given under the form of histograms, which is nice in practice since actual response times might not always obey a closed-form equation.

Another advantage is that the technique is independent of the scheduling and can be used whatever the policy (preemptive, non-preemptive, fixed-priority, dynamic-priority, etc) and whatever the task model. However, as it will be discussed in Section 2, the fulfillment of some rather strong assumptions is necessary.

2 System model and assumptions

We consider a system made of a set of tasks comprising aperiodic and possibly periodic tasks. The kth job of task τ_i is denoted by $\tau_{i,k}$. Let $R_{i,k}$ be the response time of $\tau_{i,k}$, that is the time between the release time of the job and its completion. Let $(R_{i,n})_{n\in\mathbb{N}}$ be the sequence of the response time values for task τ_i. We assume that the system can be monitored during a sufficiently long time period and that an empirical distribution of the response times can be obtained.

Since several tasks compete for the processor with execution times that may vary over time, the response time of a task τ_i may change from one instance to another job of this task. We assume that the response times of successive instances of an aperiodic task form a sequence of mutually independent and identically distributed (i.i.d) values. One denotes by $(\mathcal{R}_{i,n})$ the sequence of i.i.d. values of the response times of $\tau_{i,k}$ over successive activations. Departure from the i.i.d. property, caused by non-stationarity, linear and non-linear dependences, can be estimated, for instance using the BDS test (Brock, Dechert and Scheinkman - [2]), but a simple autocorrelation analysis alone will give us a good deal of information by detecting linear temporal dependencies.

This study is specifically targeted at aperiodic tasks, first, because periodic tasks can generally be well handled using existing results from scheduling theory and, second, because response times of periodic tasks (under WCET assumption) forms a sequence that is periodic after some time instant and thus do not verify the i.i.d. assumption.

3 Probabilistic estimation with large deviation: a recap

In this section, we recap some basic results from the field of large deviation, and some recent results presented by the authors in [8] that enable to handle distributions given as histograms.

For a given sequence of mutually independent, identically distributed random variables $(\mathcal{R}_{i,n})$, $n \in \mathbb{N}$, let $\mathcal{M}_n = \frac{1}{n}\sum_{k=1}^{n} \mathcal{R}_{i,k}$ be the mean of this sequence $(\mathcal{R}_{i,n})$. We obtain using Cramer's theorem[2], that $P(\mathcal{M}_n \in G)$ *satisfies a rate deviation principle with rate-function* I:

$$P(\mathcal{M}_n \in G) \asymp e^{-n \inf_{x\in G} I(x)}$$

[2] see Appendix 6 for details and notations

where G is any subset of $\mathbb{R}$. In our case, G is the subset we want to assess the probability that the response time belongs to. For instance, if we are interested in $P(\mathcal{M}_n \geq k)$ with $k \geq E[\mathcal{R}_{i,n}]$ then $G = [k, +\infty)$ and we estimate the decay rate of the right-hand tail of the distribution. From LD theory, we know that

$$\begin{aligned} I(x) &= \sup_{\tau>0}[\tau x - \log E(e^{\tau x})] \\ &= \sup_{\tau>0}[\tau x - \log \textstyle\sum_{k=-\infty}^{+\infty} p_k e^{k\tau}] \end{aligned} \tag{1}$$

If there is a closed form for the law of the $\mathcal{R}_{i,n}$, or if the $\mathcal{R}_{i,n}$ form a finite Markov chain, it is possible to obtain an explicit expression for the rate function. In our case, where the law of $\mathcal{R}_{i,n}$ is given by a density histogram, this is not possible and a numerical method has to be used to obtain an estimate of $I(x)$.

As $I(x)$ is a supremum of affine functions, it is a convex function and it is enough to compute the point x^* where $I(x)$ reaches its minimum to obtain the asymptotic behavior

$$P(\mathcal{M}_n \in G) \asymp e^{-nI(x^*)} \tag{2}$$

x^* is the point where the first derivative of $I(x)$ with respect to t is equal to 0 (see equation 1). This point is reached for τ_0 s.t. $x \sum_{k=-\infty}^{+\infty} p_k e^{k\tau_0} = \sum_{k=-\infty}^{+\infty} k p_k e^{k\tau_0}$ which can be rewritten as $\sum_{k=-\infty}^{+\infty} (k-x) p_k e^{k\tau_0} = 0$. Let $u = e^t$ and

$$F(u) = \sum_{k=-\infty}^{+\infty} (k-x) p_k u^k \tag{3}$$

The problem consists in finding numerically $u_0 > 0$ s.t. $F(u_0) = 0$. This problem can be solved with Newton-like methods, which are available in any numerical or symbolical computation software.

4 An example

Let us consider here a numerical example for an aperiodic task having the empirical distribution of the response times given in Figure 1.

Imagine we want to evaluate the probability that the average response time over a certain number of instances n is greater than a value x (or that the sum of the response times is larger than nx). We replace x by its value in Equation 3 and we obtain numerically u_0 such that $F(u_0) = 0$. The value of t for which the right side of Equation 1 is maximized is $t_0 = \ln(u_0)$. Then, we compute $I(x)$ over the interval of interest G (here $[x, +\infty)$) and the infimum is the decay rate $I(x^*)$ we are looking for (see Equation 2).

The upper bound on the probability that $P[\mathcal{M}_n \geq x]$ for $x \in \{45, 50, 55\}$ is shown in Figure 2. For instance, over 10 instances, the probability to get an average response time greater than 45 is less than 0.25, while the probability to get more than 50 and 55 are respectively lower than 0.021 and 0.0004.

RT interval	Probability	k
$[0, 10)$	$1/25$	5
$[10, 20)$	$2/25$	15
$[20, 30)$	$3/25$	25
$[30, 40)$	$10/25$	35
$[40, 50)$	$4/25$	45
$[50, 60)$	$3/25$	55
$[60, 70)$	$2/25$	65

Figure 1: Empirical distribution of the response time (in ms), with an expectation equal to 37.4. The value of k is the mean of the interval.

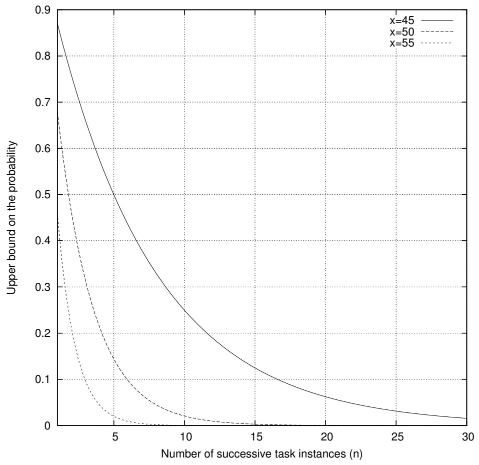

Figure 2: Upper bound on the probability that $P[\mathcal{M}_n \geq x]$ with $x \in \{45, 50, 55\}$ and $n \in [1, 30]$ obtained for the empirical distribution of the response times given in Figure 1.

It is worth mentioning that it is possible to study the sum of the response times of a set of aperiodic tasks in the same way as done previously for a single aperiodic task, provided that the individual distribution of the response times are independent. The probability distribution of the sum of two independent discrete random variables X and Y with probability distribution f and g is given by their convolution $f \star g$. In practice, the most efficient way to compute a convolution is the use of the Fast Fourier Transform (FFT) (see [8] for more details).

5 Conclusion

In this paper we propose a new approach for estimating a probability on the average and the sum of the response times of a sequence of consecutive aperiodic tasks. The approach is based on large deviation theory and can be applied under any scheduling policies (fixed-priorities, EDF, preemptive, non-preemptive) as long as the system can be monitored.

The results hold under the assumption that the response times are i.i.d.. In practice, this assumption can be easily tested using statistical tests such as the BDS statistics but it is clear that it will not hold for all kinds of systems and workloads. Future work should be devoted to experimental studies aimed at determining the practical conditions ensuring the i.i.d. property. It would be also interesting to study, for instance by simulation, how departure from the i.i.d. property impacts the accuracy of the results.

To some extent, it is possible in theory to relax the i.i.d. assumption and consider some correlation among the response times. It might thus enable us to handle periodic tasks as well. The extent to which it can be applied with the type of correlations one may expect in real-time systems, for both periodic and aperiodic tasks, remains to be investigated.

Finally, it is worth mentioning that the same approach can be used to study other quantities of interest, such as task execution times or inter-arrival times.

6 Appendix : Large deviation theory - notations and reminder

Let us recall that $(\mathcal{R}_{i,n})$ is the sequence of i.i.d. random variables modelling the response times of $\tau_{i,k}$ over successive activations and $\mathcal{M}_n = \frac{1}{n}\sum_{k=1}^{n} \mathcal{R}_{i,k}$. The Cramer theorem states that:

$$-\inf_{x \in G^\circ} I(x) \le \liminf_{n\to\infty} \frac{1}{n} \ln P(\mathcal{M}_n \in G) \le \limsup_{n\to\infty} \frac{1}{n} \ln P(\mathcal{M}_n \in G) \le -\inf_{x \in \bar{G}} I(x)$$

where G is any subset of $\mathbb{R}$, with G° the open subset and $x \in \bar{G}$ the closed subset. From the previous inequalities, one derives

$$-\inf_{x \in G^\circ} I(x) \le \frac{1}{n} \ln P(\mathcal{M}_n \in G) \le -\inf_{x \in \bar{G}} I(x)$$

which gives us the behavior of the logarithm of the quantity of interest. Taking the exponential, we obtain

$$e^{-n \inf_{x \in G^\circ} I(x)} \le P(\mathcal{M}_n \in G) \le e^{-n \inf_{x \in \bar{G}} I(x)}$$

which, since in our case G is a subset of $\mathbf{R}$, can be simplified into

$$P(\mathcal{M}_n \in G) \asymp e^{-n \inf_{x \in G} I(x)}.$$

References

[1] L. Abeni and G. Buttazzo. QoS guarantee using probabilistic deadlines. In *IEEE Euromicro Conference on Real-Time Systems (ECRTS99)*, 1999.

[2] W.A. Brock, W.D. Dechert, B. LeBaron, and J.A. Scheinkman. A test for independence based on the correlation dimension. Working papers 9520, Wisconsin Madison - Social Systems, 1995.

[3] A. Burns, G. Bernat, and I. Broster. A probabilistic framework for schedulability analysis. In *Third International Embedded Software Conference (EMSOFT03)*, pages 1–15, 2003.

[4] L. Cucu and E. Tovar. A framework for response time analysis of fixed-priority tasks with stochastic inter-arrival times. *ACM SIGBED Review*, 3(1), 2006.

[5] S. Edgar and A. Burns. Statistical analysis of WCET for scheduling. In *22nd of the IEEE Real-Time Systems Symposium (RTSS01)*, 2001.

[6] J.P. Lehoczky. Real-time queueing theory. In *10th of the IEEE Real-Time Systems Symposium (RTSS96)*, pages 186–195, 1996.

[7] J. T. Lewis and R. Russel. An introduction to large deviations for teletraffic engineers, 1997. Available at http://www.stp.dias.ie/APG/dias_apg_pub.html.

[8] N. Navet and R. Schott. Assessing the risk and return of financial trading systems - a large deviation approach. In *6th International Conference on Computational Intelligence in Economics and Finance (CIEF2007)*, 2007.

[9] N. Navet, Y.-Q. Song, and F. Simonot. Worst-case deadline failure probability in real-time applications distributed over CAN (Controller Area Network). *Journal of Systems Architecture*, 46(7):607–618, 2000.

[10] H. Pham. Some applications and methods of large deviations in finance, February 2007. available at http://arxiv.org/abs/math.PR/0702473.