Modelling large timescale and small timescale service variability*

Marco Gribaudo[1], Illés Horváth[2], Daniele Manini[3], Miklós Telek[4]

[1] Dip. di Elettronica, Informazione e Bioingengeria, Politecnico di Milano, Italy,
marco.gribaudo@polimi.it
[2] MTA-BME Information Systems Reseach Group, Hungary,
horvath.illes.antal@gmail.com
[3] Dip. di Informatica, Università di Torino, Italy
manini@di.unito.it
[4] Dept. of Networked systems, Technical University of Budapest, Hungary
telek@hit.bme.hu

Abstract. The performance of service units might depend on various randomly changing environmental effects. It is quite often the case that these effects varies on different time scales. In this paper we consider short and long scale service variability, where the short scale variability affects the instantaneous service speed of the service unit and the large scale effect is defined by a modulating background Markov chain. The main modelling challenge in this work is that the considered short and long range variation results randomness along different axes, the short scale variability along the time axis and the long scale variability along the work axis.

Keywords: short and long term service variability, Brownian motion, Markov modulation, performance analysis

1 Introduction

Service speed variability is a problem that has been measured in many practical application scenario. For example in [3], it has been observed for vehicular traffic. More recently this problem has been recognized in data-center [2]. The effect of variability was also studied in [1] with application to video-streaming. Most of the previous literature however, focused only on large-time scale variability, where Markov-modulating models represent the random effect of the environment. All of those models can be handled with matrix analytic methods, summarized e.g., by Latouche and Ramaswami in [4].

The variation in the service speed can be modelled by dividing the amount of job to be executed into "infinitesimal quantities of work to be done" and consider the "speed at which this infinitesimal work is performed", i.e., the random amount of time needed to execute the infinitesimal amount of work. Then, if a

* This work is partially supported by the OTKA K-123914 grant.

model that defines how speed changes over time, the complete system can be modelled in a straight-forward way where the amount of work increases gradually along the analysis and the time required to execute the given amount of work is a random process.

If the service process depends on a time dependent random process, e.g., on a modulating background CTMC representing the environmental state, whose "clock" evolves according to the time, then the natural performance analysis is based on the gradually increasing time and randomly varying time dependent environment state.

However, in many real applications, variability is not easily predictable and works at different time-scales. Modulating CTMCs (whose "clock" evolves according to the time) works very well to model variability where the parameters of the job execution remains constant for a longer random period of time, and there are few jumps during the execution of one job. Apart of this large scale variability, in this work, we focus also on variability that occurs at much smaller time scales, where the execution speeds changes thousands, if not millions, of times during the execution of the main job, and combine it with the more classical modulation that works on a larger time scale.

The remainder of this paper is structured as follows. In Section 2 we start with considering only the small time-scale variability. In Section 3 we additionally introduce also the large time-scale variability. The effects of the considered variability is studied in Section 4 through numerical examples, and Section 5 concludes the paper.

2 Small time-scale variability

In this section, we omit the large time-scale variability and instead focus only on small time-scale variability. So assume that the environmental state is unchanged for now.

We introduce a second order fluid model for the short time-scale variability: assuming that a job is composed of quantums of size Δx, each such quantum is served in a random amount of time with distribution $N(\mu\Delta x, \sigma^2\Delta x)$ (with $\mu > 0$). Assuming that the service times of the different quantums are independent, the progress of service is modeled by a Brownian motion $B(x)$ with parameters μ and σ^2. We emphasize that in this model, the Brownian motion corresponds to *the time required to service a job as a function of the size of the job* (see Figure 1). A job of size x thus requires a random time T with distribution $N(\mu x, \sigma^2 x)$, whose probability density function is $\frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x\sigma^2}}$. Note that a Brownian motion may take negative values as well, which does not make sense physically, but, since $\mu > 0$, for macroscopic values of w, the probability that T is negative is negligible.

We focus on the service of a job in a queue whose work requirement, W, is generally distributed according to probability density function $f_W(x)$.

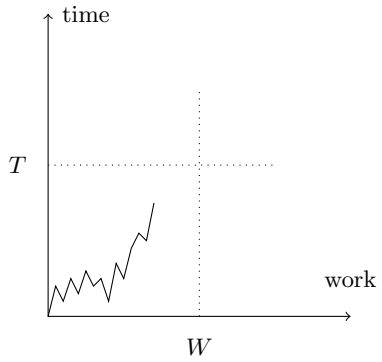


Fig. 1: The time T required to service a job as a function of the job size W

Using the second order fluid model assumption, the probability density function of the service time of a job, denoted by $f_T(t)$, can be computed as:

$$f_T(t) = \int_0^\infty f_W(x) \cdot \frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x\sigma^2}} dx \tag{1}$$

2.1 Moments of the scaled distribution

There are also some interesting relations between the moments of W, the moments of T and the parameters μ and σ^2. In particular, the K-th moment of T can be expressed as:

$$\begin{aligned} E[T^K] &= \int_0^\infty t^k f_T(t) dt = \int_0^\infty t^k \int_0^\infty f_W(x) \cdot \frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x\sigma^2}} dx \cdot dt = \\ &= \int_0^\infty f_W(x) \int_0^\infty t^k \cdot \frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x\sigma^2}} dt \cdot dx = \\ &= \int_0^\infty f_W(x) E[N(x\mu, x\sigma^2)^k] dx \end{aligned}$$

Now, since $E[N(x\mu, x\sigma^2)^k]$ can be expressed as a polynomial in $x\mu$ and $x\sigma^2$, where σ appears only for even exponents:

$$E[N(x\mu, x\sigma^2)^k] = \sum_{j=0}^{k} u_{k,j} (x\mu)^j (\sqrt{x}\sigma)^{k-j} \tag{2}$$

we can compute the moments of T as:

$$\begin{aligned} E[T^K] &= \int_0^\infty f_W(x) \sum_{j=0}^{k} u_{k,j} (x\mu)^j (\sqrt{x}\sigma)^{k-j} dx = \\ &= \sum_{j=0}^{k} u_{k,j} \mu^j \sigma^{k-j} \int_0^\infty f_W(x) x^j (\sqrt{x})^{k-j} dx = \\ &= \sum_{j=0}^{k} u_{k,j} \mu^j \sigma^{k-j} E[W^{\frac{k+j}{2}}] \end{aligned} \tag{3}$$

Since σ^2 appears only for even exponents, $k+i$ is always even, so $E[T^{\frac{k+i}{2}}]$ is always an integer moment of T. For example, for the first and second moment, since $E[N(x\mu, x\sigma^2)^2] = x\mu$ and $E[N(x\mu, x\sigma^2)^2] = x^2\mu^2 + x\sigma^2$, we have:

$$\begin{aligned} E[T] &= \mu E[W], \\ E[T^2] &= \mu^2 E[W^2] + \sigma^2 E[W]. \end{aligned}$$

3 Combining large and small time-scale variability

Large scale variability can be considered using a discrete state Markov modulating process (MMP) of K states, denoted by $M(t)$. The MMP is a CTMC with infinitesimal generator matrix denoted by Q. In state i, the service is characterised by rate μ_i and variance σ_i.

Only considering large scale variability (that is, assuming $\sigma_k \equiv 0$) would lead to a standard first order Markov-modulated fluid model. However, including small-scale variability makes for an interesting and complex model.

Assume that a job of size $W = x$ starts service at time $t = 0$, with the background modulating process in state i. Then the evolution of the service time $B(x)$ as a function of the job size is the following:

- Let a_1 denote the time of the first transition of $M(t)$. As long as $B(x)$ is smaller than a_1, $B(x)$ evolves according to a BM(μ_i, σ_i).
- At time a_1, $M(t)$ changes to some state j. Accordingly, assuming that the first passage of $B(x)$ to a_1 occurs at work amount w_1, for $x \geq w_1$, $B(x)$ evolves according to a BM(μ_j, σ_j) (starting from the point w_1 and from level a_1).
- This is repeated for further transitions of $M(t)$ at times $a_2, a_3, \ldots$, up to the point $x = W$.

Note that in visualization, the x axis denotes the job size, and the y axis denotes time, see Figure 2. Thus for $B(x)$, the behaviour can be described as a type of *level-dependent Brownian motion*: the parameters μ and σ of the Brownian motion change upon first passage to levels $a_1, a_2, \ldots$. This is different from usual second order Markov-modulated fluid models, where parameter changes

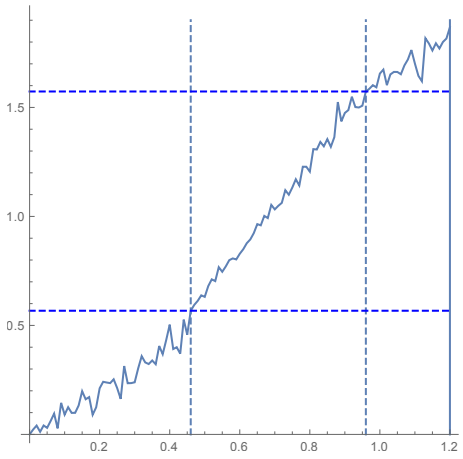


Fig. 2: A possible realization of $B(x)$ for job size $W = 1.2$

occur upon the variable of the Brownian motion (x in our case) reaching some transition points instead of the level reaching transition points.

Keeping in mind that $M(t)$ is a CTMC, the entire distribution of $B(x)$ is determined by the initial points $t = 0$ and $x = 0$ and the initial state of the modulating process $M(0) = i$. The process $B(x)$ can be simulated as follows:

- $B(x)$ starts from $t = 0$, $x = 0$, with $M(0) = i$ and job size W.
- Generate the first transition time a_1 of $M(t)$.
- $B(x)$ runs as a $\mathrm{BM}(\mu_i, \sigma_i^2)$ until either the value of $B(x)$ reaches a_1 or x reaches W, whichever occurs first.
- If $x = W$ occurred first, then the simulation is finished.
- If $B(w_1) = a_1$ for some $w_1 < W$, then we generate the next state j and also the next transition time a_2 according to the CTMC $M(t)$, then continue $B(x)$ as a Brownian motion with parameters (μ_j, σ_j^2) starting from the point (w_1, a_1) until either the value of $B(x)$ reaches a_2 or x reaches W, whichever occurs first.
- We keep generating new transitions and new Brownian motion sections until we reach W. The service time of the job is $T = B(W)$.

The main question, similar to Section 2, is the distribution of T and performance measures derived from T. In this case, an analytical answer is non-trivial even for a given job size $W = x$. One possible analytic formulation is to first introduce the cumulative distribution type functions (for fixed x)

$$G_{ij}(x,t) = \Pr\left(B(x) \le t, M(B(x)) = j | M(0) = i, W = x\right) \tag{4}$$

which include information about the initial and final background state of $M(t)$ along with the distribution of the service time. An analytic formula for $G_{ij}(W,t)$ is subject to ongoing research.

4 Simulation results

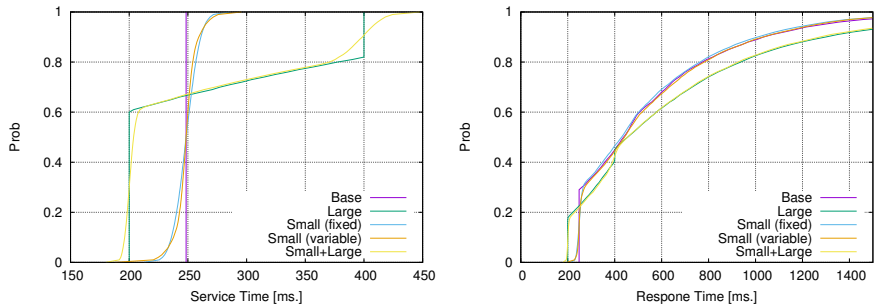
To study the effects of variability, we have applied the procedure outlined in Section 3 to simulate the behaviour of the queue with short and long scale variability. In particular, to find the intersection between the Brownian motion and the level determined by the time at which the modulating process changes state, we have discretised the work with a quantum Δx, and during the period when the MMP stays in state i, for each quantum we have set the evolution of the time according to a normal distribution $N(\mu_i \Delta x, \sigma_i^2 \Delta x)$ (following the procedure outlined at the beginning of Section 2). The MMP leaves state i at the first time instant in which the discretised BM crosses the level T_n, where T_n is the time of the nth state transition of the MMP. When the nth state transition occurs in state i, then $T_n = T_{n-1} + \tau_i$, where T_{n-1} is the time of the previous state transition and τ_i is exponentially distributed with parameter $-Q_{ii}$ (the $i$the diagonal element of the generator matrix of the modulating CTMC). This simulation approach is indeed an approximation, but it can be made arbitrarily precise by choosing appropriately small values of Δx.

In our numerical experiment, we have considered a two-state modulating process with jump rates γ_{12} and γ_{21}, and studied the effects of different service speed and variability parameters μ_i and σ_i. To show a possible application, we have used the proposed process to describe the variable service rate in an M/G/1 queue, where jobs arrive according to a Poisson process of rate λ and are served by a single server subject to short and long range variability according to a first-come-first-served discipline. To compare the results for different service time distributions we assumed that the mean service time $E[W]$ is identical in each cases. The arrival rate, λ, is selected such that the queue is stable. Unless otherwise stated, the used parameters have been the following:

$$\begin{gathered} \lambda = \tfrac{1000}{350}\text{job/s}, \quad E[W] = 100\text{ms}, \quad \Delta X = 0.05\text{ms}, \\ \mu_1 = 2, \quad \mu_2 = 4, \quad \sigma_1 = 0.4, \quad \sigma_2 = 1.5, \\ \tfrac{1}{\gamma_{12}} = 1.25\text{s}, \quad \tfrac{1}{\gamma_{21}} = 0.8\text{s}. \end{gathered} \tag{5}$$

In this framework, the discretisation interval has been chosen so that on average, the BM for each job requires 2000 samples, and in the average sojourn time in the two modulating states, the BM is samples respectively 25000 and 16000 times. Each simulation considers the execution of $N = 10000$ jobs.

We start focusing on jobs requiring a fixed amount of work (i.e. $W = E[W]$ is deterministic. Figure 3a shows the service time distribution for different server variability configuration. The *Base* case, considers the case in which no variability is used: in particular to $\mu_1 = \mu_2 = 2.4848$ and $\sigma_1 = \sigma_2 = 0$. As it is expected, all the probability mass is centred along $\mu E[W] = 248.48$. The *Small* variability cases, differs from the *Base* one by adding a small variability. In the *Small (fixed)* case $\sigma_1 = \sigma_2 = 0.98773$ and in the *Small (variable)* case we have the state dependent variability $\sigma_1 = 0.4$ and $\sigma_2 = 1.5$. As it can be seen, they both destroy the deterministic behaviour, in a slightly different way: the fixed σ case has a more uniform effect, while the variable one presents larger tails. The

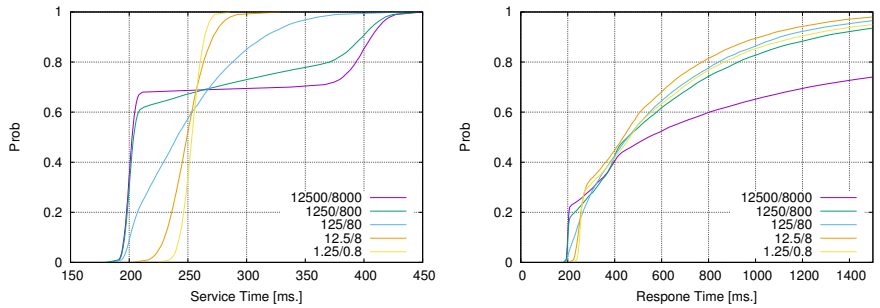


a) b)

Fig. 3: Considering different small scale and large scale variability configurations for a fixed job length: a) service time distribution, b) response time distribution

case called *Large* considers only large scale variability only, i.e., $\sigma_1 = \sigma_2 = 0$. During a sojourn in a state of the MMP the service time of a job is deterministic. In state 1, with $\mu_1 = 2$, the service time is exactly 200ms, and in state 2, with $\mu_2 = 4$, it is exactly 400ms. The jumps in Figure 3a at 200ms, and 400ms are associated with the cases when the MMP stays in state 1 (2, respectively) for the whole period of the service. The cases when the MMP experiences state transition during the service are represented by the continuously increasing part of the *Large* curve. The case that combines both small and large scale variability (*Small+Large*, $\mu_1 = 2, \mu_2 = 4, \sigma_1 = 0.4, \sigma_2 = 1.5$) further smooths the curves, and the effect is more evident near the two probability masses at 200ms, and 400ms. Figure 3b shows the response time distribution of the corresponding queuing models. In this case it is interesting to see that in the cases where small variability is considered there are no jumps due to its perturbation effect.

We then study the effect of the modulating process, by changing the average sojourn time in its two states, while maintaining the state probabilities. Figure 4 considers different combinations of sojourn times ranging from 12.5s and 8s down to 1.25ms and 0.8ms for the deterministic job length distribution W, and the other parameters defined as in (5). When the sojourn time is very large, service times are correlated, and the service time distribution tend to concentrate the probability mass near the times required in both modulating states. On the other hand, when the switching process changes very fast, the distribution tend to concentrate in the average case, producing results very similar to the one seen in Figure 3 for the cases with small variability only: in this case, there is almost no difference between large scale and small scale variability, because the quick alternation of the modulating process eliminates the large scale effect. As a final remark, in order to consider a switching process with 1.25ms and 0.8ms, we had to reduce the sampling time $\Delta x = 0.01$ to allow a sufficient number of samples during the sojourn in a modulating state. For what concerns response time (Figure 4c), when the modulating process present deep correlation by spending



a) b)

Fig. 4: Considering different durations in the modulating process for a fixed job length: a) service time distribution, b) response time distribution

longer times in a single state, bursts are created, decreasing considerably the performances of the system.

We finally consider the effect of variability on different job length distributions. In particular, Figure 5a shows the service time distribution when the job length follows, beside the deterministic distribution already discussed, an exponential distribution, an Erlang distribution with 4 stages, and the following Hyper-Exponential ($w_H(x)$) and Pareto ($w_P(x)$) distributions characterised by the following probability density functions:

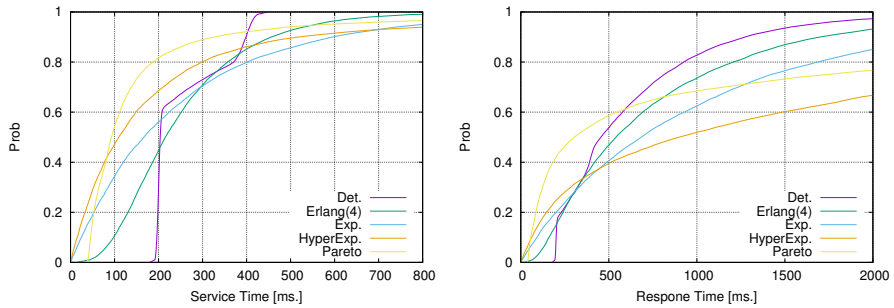
$$w_H(x) = \frac{1}{2}\lambda_1 e^{-\lambda_1 x} + \frac{1}{2}\lambda_2 e^{-\lambda_2 x},$$

$$w_P(x) = \begin{cases} \frac{20^{\frac{5}{4}}\frac{5}{4}}{x^{\frac{9}{4}}} & x > 20, \\ 0 & x < 20, \end{cases}$$

where $\lambda_1 = \frac{1}{100(1+\sqrt{\frac{3}{5}})}$ and $\lambda_2 = \frac{1}{100(1-\sqrt{\frac{3}{5}})}$. As it can be noted, the effect of service variability is more evident on job length distributions with a lower coefficient of variation. Figure 5b shows the effect on response time: indeed, combining the effect of service variability with heavy tailed distribution, as for the Pareto case, can create very long queues which can lead to extremely long response times.

5 Conclusions

In this work, we have introduced a queue with a service model where the large timescale variability is modelled by a modulating background Markov process, and small timescale variability is modelled by a second-order fluid process for the service time of a job. The resulting service model can be interpreted as a certain type of level-dependent Brownian motion.



a) b)

Fig. 5: Considering small scale and large scale variability for different job length distributions: a) service time distribution, b) response time distribution

We have presented simulation results for the service time and response time of a job for various job size distributions. In future work, we hope to give a full analytic description of the system, most notably by giving an analytic solution for (4).

References

1. Anjum, B., Perros, H.: Bandwidth estimation for video streaming under percentile delay, jitter, and packet loss rate constraints using traces. Computer Communications 57, 73 – 84 (2015), http://www.sciencedirect.com/science/article/pii/S0140366414003089
2. Guo, J., Liu, F., Huang, X., Lui, J.C., Hu, M., Gao, Q., Jin, H.: On efficient bandwidth allocation for traffic variability in datacenters (04 2014)
3. Kimber, R., Daly, P.: Time-dependent queueing at road junctions: Observation and prediction. Transportation Research Part B: Methodological 20(3), 187 – 203 (1986), http://www.sciencedirect.com/science/article/pii/0191261586900160
4. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. Society for Industrial and Applied Mathematics (1999), http://epubs.siam.org/doi/abs/10.1137/1.9780898719734