33rd European Modeling & Simulation Symposium
18th International Multidisciplinary Modeling & Simulation Multiconference

ISSN 2724-0029 ISBN 978-88-85741-57-7 © 2021 The Authors.
doi: 10.46354/i3m.2021.emss.020

Time-to-idle Control Variate Performance in the Single Queue Case

Andrés Suárez-González[1,*], Cándido López-García[1], José C. López-Ardao[1], Raúl Rodríguez Rubio[1] and Miguel Rodríguez Pérez[1]

[1]atlanTTic research center, *Universidade de Vigo*, *Escola de Enxeñaría de Telecomunicación*, 36310 Vigo, Spain

[*]Corresponding author. Email address: asuarez@det.uvigo.es

Abstract

Control Variates (CV) is a Variance Reduction technique used in order to shorten simulation experiments. In a previous work we presented *Time-to-idle* as a stochastic process strongly correlated with the queue waiting time processes in the different queues of a polling service discipline network. *Time-to-idle* sample values are asynchronous with respect to those of queuing times, that is, they are generated at unpredictable times in an unpredictable order with respect to each other. This inherent characteristic allows it to be used in a network of queues (through batch means methods and taking care of synchronization between batches of both processes) but can hinge its performance in the single queue case. In this paper we evaluate its performance through simulation of the single queue case, comparing it with the service time and/or interarrival time synchronous random variables in the D/M/1, M/D/1 and M/M/1 queues where actual mean queue waiting times are known. We observe a slightly lower efficiency of *Time-to-idle* CV as was expected and we conclude that new techniques for synchronization of batches should be explored in order to minimize it.

Keywords: Analysis Methodology, Output Analysis, Variance Reduction Techniques, Control Variates, Batch Means

1. Introduction

Polling service discipline networks have been present in the telecommunication arena since the inception of computer local area networks like Token Bus, Token Ring or FDDI. Nowadays polling service discipline is still pervasive in networks such as Wireless Sensor Networks (Siddiqui et al., 2018), Wireless Metropolitan Area Networks (Yang et al., 2017) or IoT (Guan et al., 2019), for example.

A performance measure of interest when evaluating these systems is the steady state queue waiting time of packets in each node. This quantity is random in nature and can be represented by an stochastic process $W = \left\{W_i; i = 1, \ldots, \infty\right\}$ for each traffic source, and we assume it is a covariance-stationary process. As a basic measure of performance, we estimate its mean—$\mathrm{E}(W) = \overline{W}$—from the observations of a single simulation run generating a sequence of size n and computing its average

$$\overline{W}\,[n] \equiv \frac{1}{n} \cdot \sum_{i=1}^{n} W_i \tag{1}$$

and its confidence interval for the mean value through the batch means method. This method estimates the variance of (1) through m batches of size l ($n = m \cdot l$)

$$\overline{W_i}\,[l] \equiv \frac{1}{l} \cdot \sum_{j=(i-1)\cdot l+1}^{i\cdot l} W_j \tag{2}$$

under the assumption that $\left\{\overline{W_i}\left[l\right];i=1,\ldots,\infty\right\}$ behaves asymptotically as $l\to\infty$ like a Gaussian renewal process. We use a simple algorithm due to Law and Carson (1979) for the experiments commented in this paper. This way, after gathering a group of $m = 400$ batches of size l of W, $\left\{\overline{W_i}\left[l\right];i=1,\ldots,400\right\}$, we check the amount of autocorrelation; when it is under certain level we may compute the confidence interval over the 40 batches of size $10 \cdot l$, $\left\{\overline{W_i}\left[10\cdot l\right];i=1,\ldots,40\right\}$. As a matter of fact the sample size needed to comply with a given confidence interval requirement will be approximately proportional to the variance of the batches once they are approximately uncorrelated.

In order to reduce the simulation time needed for complying with a given confidence interval requirement when studying a polling service queue, Suárez-González et al. (2000) proposed a new control variate process and measured the performance of a simple implementation in a polling service network of queues.

With the goal of gaining a better understanding of its potential and shortcomings, in the present paper we compare its performance with respect to straightforward Control Variates (CV) variables available for the simplest scenario, the single queue model—service time S and interarrival time Υ (S. S. Lavengerg and Sauer, 1979)—using the TiTI simulation tool—https://icarus.det.uvigo.es/TiTI/.

In Section 2 we summarize the state of the art of the CV method applied to mean value estimation. In Section 3 we explain the use of the *Time to idle* process as a CV process. In Section 4 we present the results of the comparison with the two simple CV variables. In Section 5 we summarize the conclusions derived from our study and propose future lines of research.

2. State of the art

Although Control Variates (CV) is not restricted to mean value estimation—e.g. Portier and Segers (2019), Ortiz-Gracia (2020)—, we will focus on its use with this classical purpose. CV method—see for example Adewunmi and Aickelin (2012)—takes advantage of the knowledge about a stochastic process C—with known mean $\mathrm{E}\left(C\right) = \overline{C}$—strongly correlated with W to estimate its mean, $\mathrm{E}\left(W\right) = \overline{W}$, defining the controlled stochastic process $Y = \left\{Y_i; i=1,\ldots,\infty\right\}$ as

$$Y_i \equiv W_i - \beta \cdot \left(C_i - \overline{C}\right) \tag{3}$$

so we hope its average $\overline{Y}\left[n\right]$ will have less variance than (1). The controlled stochastic process Y with smallest variance is obtained with $\beta^* = \mathrm{Cov}\left(W, C\right)/\mathrm{Var}\left(C\right)$. As this value is usually unknown, it is estimated through $\widehat{\beta}\left[n\right]$ from the same samples of W and C used to compute $\overline{W}\left[n\right]$ and $\overline{C}\left[n\right]$ in

$$\overline{Y}\left[n\right] = \overline{W}\left[n\right] - \widehat{\beta}\left[n\right] \cdot \left(\overline{C}\left[n\right] - \overline{C}\right) \tag{4}$$

As a consequence, although W and C would be renewal processes, $\overline{Y}\left[n\right]$ will be in general a biased estimator of $\overline{W}$.

Nevertheless, if $\left(W, C\right) \equiv \left\{\left(W_i, C_i\right); i=1,\ldots,\infty\right\}$ are i.i.d. and distributed as a multivariate normal, S. S. Lavenberg and Welch (1982) show that (4) is an unbiased estimator of $\overline{W}$. They also develop an unbiased estimator $\widehat{\sigma^2}_{\overline{Y}[n]}$ of $\mathrm{Var}\left(\overline{Y}\left[n\right]\right)$, and show that $\left(\overline{Y}\left[n\right] - \overline{W}\right)/\widehat{\sigma}_{\overline{Y}[n]}$ has a Student's t distribution with $n-2$ degrees of freedom. S. S. Lavenberg and Welch (1982) also show that the loss in potential variance reduction when the optimum coefficient β^* is estimated by $\widehat{\beta}\left[n\right]$ is $(n-2)/(n-3)$.

If W and C are both correlated stochastic processes under a joint functional central limit theorem assumption, Loh (1997) shows, applying the previous result to m batches of size l $(n = l \cdot m)$ of W and C, that $\left(\overline{Y}\left[m,l\right] - \overline{W}\right)/\widehat{\sigma}_{\overline{Y}[m,l]}$ behaves asymptotically as $l \to \infty$ like a Student's t random variable with $m-2$ degrees of freedom. Hence, it is possible to use both control variates and batch means methods simultaneously.

Although control variates proposed in the literature are specific of the system being simulated, usually they are synchronous with the process whose mean is being estimated. In the case of the single server queue the seminal work of S. S. Lavengerg and Sauer (1979) proposed both the service time S and interarrival time Υ.

3. Time-to-idle process

Given a task aliquot share of idle time

$$\mathsf{L} \equiv \left(1-\rho\right) \cdot \overline{\Upsilon} \tag{5}$$

Suárez-González et al. (2000) define T_i—*Time-to-idle*—as the amount of time needed to arrive to the idle time share of the i-th task from that of the $(i-1)$-th one, composing the stochastic process $T = \{T_i; i=1,\ldots,\infty\}$ of known mean value

$$\mathrm{E}\left(T\right) = \overline{T} = \frac{\mathsf{L}}{1-\rho} = \overline{\Upsilon}. \tag{6}$$

Although it shares the same mean value as the interarrival time process, it captures the global variation of load in the polling service discipline due to the mixed variations in both service time and interarrival time processes for every traffic in the network.

An advantage of the *Time-to-idle* process as CV is its wide usability. Although derived in the case of a polling service discipline, it just needs a common server shared sequentially by one or several queues.

A disadvantage of *Time-to-idle* is its asynchronous nature itself that has to be dealt with, though.

3.1. Synchronization with W

The ratio of the amount of values (sample size) of T and that of W, tend to one as the simulated time increases. Due to the stochastic nature of the queuing systems themselves, it will happen that we will have different amount of values of T and W, and hence we will have a different number of batches $\overline{T_i}\left[l\right]$ than that of $\overline{W_i}\left[l\right]$. Nevertheless, Suárez-González et al. (2000) take advantage of the batch means method to help in the synchronization, with a simple (not the only one possible) strategy of five points:

1. Using the auxiliary stochastic process $K = \left\{K_i; i = 1, \ldots, \infty\right\}$ where K_i is the instant (simulation clock) when the system has been idle exactly $i \cdot \mathrm{L}$ units of time. Batches of size l of K are represented by the first value of the batch, that is, $K_i\left[l\right] \equiv K_{(i-1)\cdot l+1}$. The batches of T are obtained from those of K by differentiation: $\overline{T_i}\left[l\right] = (K_{i+1}\left[l\right] - K_i\left[l\right])/l$.
2. Each batch $\overline{W_j}\left[l\right]$ is marked with the arrival time to the queue of its first frame, $A_j\left[l\right] \equiv A_{(j-1)\cdot l+1}$, where $A = \left\{A_i; i = 1, \ldots, \infty\right\}$ is the arrival process. Some initial values of W can be deleted to limit the impact of the transient period of the simulation, and we still are able to synchronize both stochastic processes in an easy way. Moreover, there is no need to begin to construct pairs at the beginning or at the end of both sequences.
3. K is stored with smaller batch size ($l/2$) than W (l), that is, more values of K are stored than batches of W, and some amount of extra stored values of K are allowed to deal with the non-perfect synchronization of both processes.
4. Matching of both set of batches in pairs begins from the middle batch of W, $\overline{W_{m/2}}\left[l\right]$, pairing it with the value $K_{i'}\left[l\right]$ nearest to $A_{m/2}\left[l\right]$, and continuing toward both sides from there.
5. The coefficient $\widehat{\beta}\left[m', 10 \cdot l\right]$, the variance of the average $\widehat{\sigma^2}_{\overline{Y}[m', 10 \cdot l]}$ and the confidence interval, are estimated with a higher batch size than that used to compute the average $\overline{Y}\left[m, l\right]$ itself (that uses $\widehat{\beta}\left[m', 10 \cdot l\right]$ estimation).

4. Performance on the single queue

In order to obtain a better understanding of the capabilities and limitations of the *Time-to-idle* as a CV process, the single queue with *first in first out* discipline allows a comparison with more straightforward variables in a scenario of known mean queue waiting time $\overline{W}$: the service time (S) and interarrival time random variables (Υ). It is important to notice that these two random variables would loose their straightforward advantage in the single queue (inherent synchronization with the waiting time and unbiased controlled estimator) when applied to the multiple queue polling service discipline.

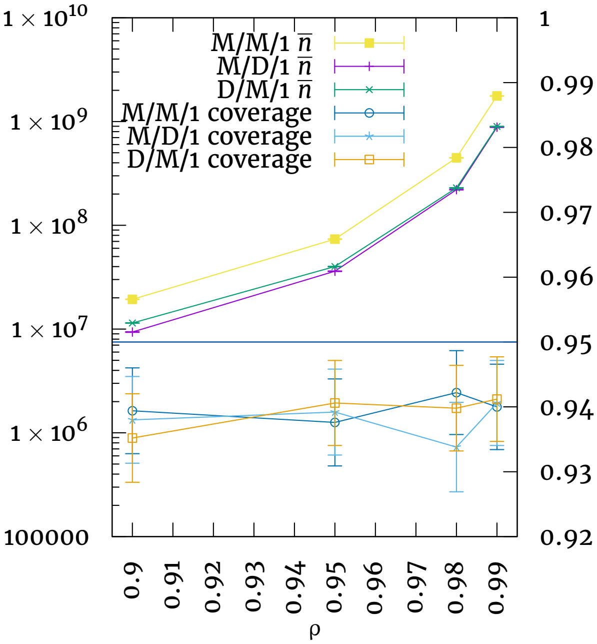

Figure 1. Simulation results without CV

We will compare the performance of *Time-to-idle* against service time and/or interarrival time in G/G/1 queues with known mean waiting time.

We will focus on only the deterministic distribution apart the exponential one since it permits us to isolate the efficiency of S and Υ in a best case scenario for them. Hence we will study the M/D/1 and D/M/1 queues.

As a final comparison for the case where both S and Υ can be CV candidates, we will study the M/M/1 queue.

For each model we do 5000 simulation runs for each load value $\rho \in \{0.9, 0.95, 0.98, 0.99\}$ with a requirement of 95% confidence intervals narrower than ±1% of the average value. We simulate both using a given CV and no CV at all and then compute:

- the average number of sample size of W in the no CV case and its average reduction when applying CV, and
- the actual coverage of the 95% confidence intervals from the simulations.

During the first 20 seconds sample values of W are discarded to limit any transient period effect.

4.1. Case study

All of the models are simulated with $\overline{\Upsilon} = 1$ and $\overline{S}$ varying for each ρ.

Figure 1 shows for the batch means method without CV both 95%-CI for the mean sample size $\overline{n}$ needed to comply with the stopping requirement and the 95%-CI for the actual coverage attained in each model. Sample size increases as ρ does, as expected. Coverage of the

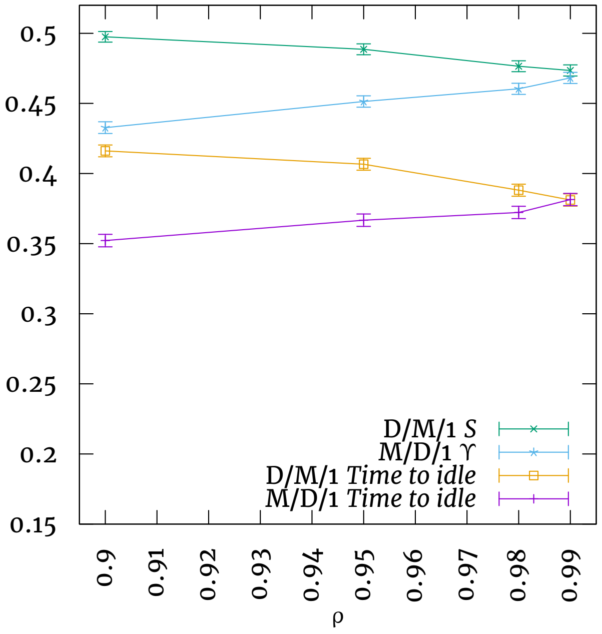

Figure 2. Sample size reduction in M/D/1 and D/M/1 with CV

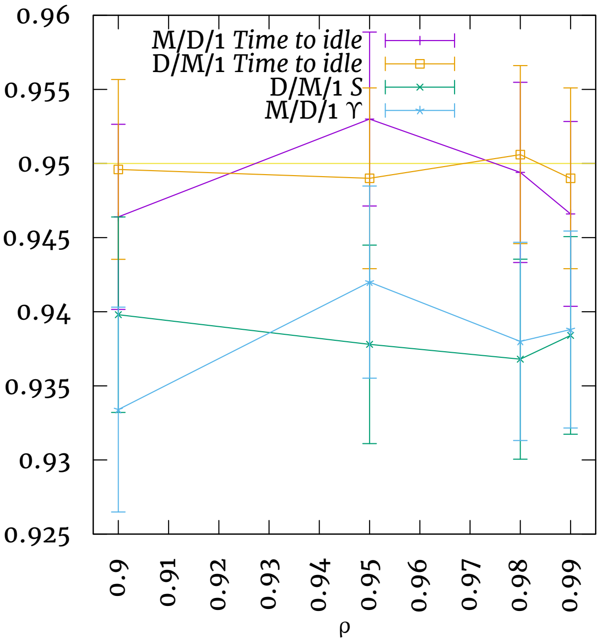

Figure 3. Actual coverage in M/D/1 and D/M/1 with CV

batch means method approaches the 95% requirement but it is clearly lower. We should point out that it will produce asymptotically correct ones as batch sizes increases.

4.1.1. CV on M/D/1 and D/M/1

Figure 2 shows the sample size reduction attained by each CV process. In the D/M/1 case the service time S reduces the needed sample size by approximately an additional 9 percent points with respect to *Time-to-idle*. A similar result appears for the interarrival time Υ in the M/D/1 case.

Figure 3 shows the coverage attained by each CV process. Only the *Time-to-idle* 95%-CIs include the asked requirement of 0.95, while both S and Υ shares their behavior with the batch means method without CV. One possible cause of *Time-to-idle* arriving to a higher coverage is due to its use of more sample values when computing the mean value estimator than when estimating its CI coverage, as commented on item 5 in Section 3.1.

4.1.2. CV on M/M/1

Figure 4 shows the sample size reduction attained by each CV process in the M/M/1 model. In this case *Time-to-idle* achieves a much better reduction than any of the other two by itself. Nevertheless, the sum of the reductions achieved by S and Υ, expected to be the reduction achieved when applied both together, would mean approximately an additional 4 percent points with respect to *Time-to-idle* alone. Being nearer to the attainable reduction by the straightforward pair when applied together shows the potential of *Time-to-idle* as a CV process.

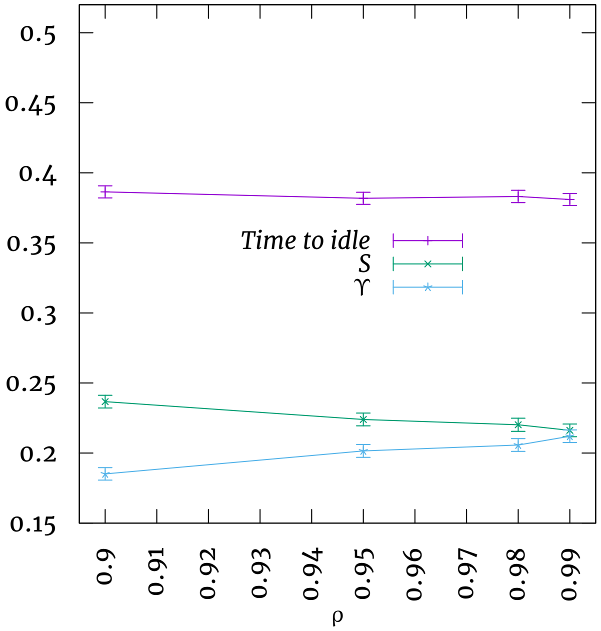

Figure 4. Sample size reduction in M/M/1 with CV

Figure 5 shows the same behavior in the M/M/1 model already observed in Figure 3, with *Time to idle* again the only one attaining the requested coverage.

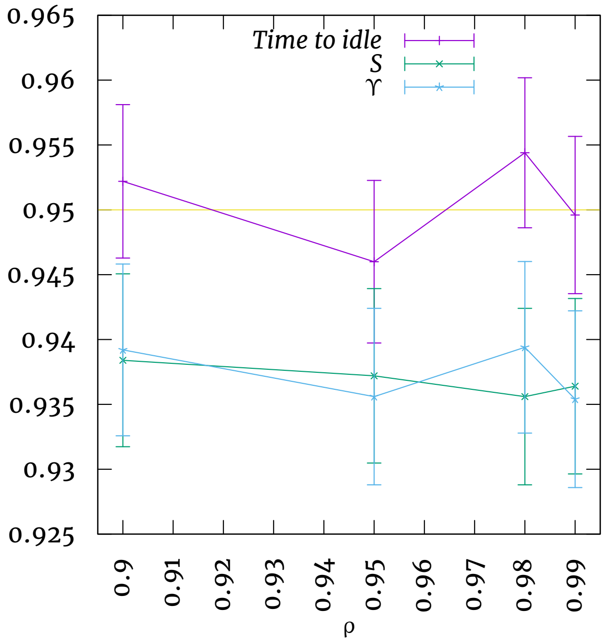

Figure 5. Actual coverage in M/M/1 with CV

5. Conclusions

Time-to-idle is a process following the load variations in a polling service network or any model with queues sharing sequentially a unique server. It is a useful control variate in order to shorten simulation time. We have compared its efficiency with respect to the straightforward variables service time and interarrival time in a single queue scenario, where both of them make sense and are perfectly synchronous.

Studying the M/D/1 and D/M/1 queues allows us to isolate the efficiency of both competing variables in their most favorable configuration. In these queues the asynchronous nature of *Time to idle* shows itself, with a slightly lower efficiency with respect to the competing ones. Nevertheless, *Time-to-idle* is the one computing the more accurate confidence interval for the mean waiting time though.

Studying the M/M/1 queue shows that *Time-to-idle* performs better than any of the two competing ones alone. Nevertheless, it is expected than using both of the competing ones at once would attain a slightly higher sample size reduction with respect to *Time to idle*, although shorter than in the other two queues. Since this is not a clear straightforward result, it could mean a higher potential of *Time-to-idle* if a better synchronization strategy can be found. We deem a deeper study of the synchronization technique of *Time-to-idle* appropriate for a future work.

6. Funding

This work has received financial support from the *Xunta de Galicia* (*Centro singular de investigación de Galicia* accreditation 2019–2022) and the European Union (European Regional Development Fund—ERDF), and by the *Ministerio de Ciencia e Innovación* through grant PID2020-113240RB-I00 *Proyectos I+D+i 2020* (partly financed with ERDF funds), and RED2018-102585-T (GO2EDGE).

References

Adewunmi, A. and Aickelin, U. (2012). Investigating the effectiveness of variance reduction techniques in manufacturing, call center and cross-docking discrete event simulation models. In *Use Cases of Discrete Event Simulation*, pages 1–26. Springer.

Guan, Z., Jia, Y., and He, M. (2019). A bidirectional polling MAC mechanism for IoT. *Electronics*, 8(6):715.

Law, A. M. and Carson, J. S. (1979). A sequential procedure for determining the length of a steady-state simulation. *Operations Research*, 27(5):1011–1025.

Loh, W. W. (1997). *On the Method of Control Variates*. PhD thesis, Stanford University.

Ortiz-Gracia, L. (2020). Expected shortfall computation with multiple control variates. *Applied Mathematics and Computation*, 373:125018.

Portier, F. and Segers, J. (2019). Monte carlo integration with a growing number of control variates. *Journal of Applied Probability*, 56(4):1168–1186.

S. S. Lavenberg, T. L. M. and Welch, P. D. (1982). Statistical results on control variables with application to queuing network simulation. *Operations Research*, 30:182–202.

S. S. Lavengerg, T. L. M. and Sauer, C. H. (1979). Concomitant control variables applied to the regenerative simulation of queueing systems. *Operations Research*, 27:134–160.

Siddiqui, S., Ghani, S., and Khan, A. A. (2018). ADP-MAC: An adaptive and dynamic polling-based MAC protocol for wireless sensor networks. *IEEE Sensors Journal*, 18(2):860–874.

Suárez-González, A., López-García, C., López-Ardao, J. C., and Fernández-Veiga, M. (2000). On the use of control variates in the simulation of medium access control protocols. In *2000 Winter Simulation Conference Proceedings (Cat. No. 00CH37165)*, volume 1, pages 782–787. IEEE.

Yang, Z.-J., Su, Y., Ding, H.-W., and Ding, Y.-Y. (2017). Strategies for improving the quality of polling service in wireless metropolitan area network. In *MATEC Web of Conferences*, volume 125, page 04019. EDP Sciences.