Bounds for simulation

van Dijk, N.M.; van der Sluis, H.J.

INSTITUTE OF
ACTUARIAL SCIENCE
&
ECONOMETRICS

REPORT AE 4/2006

Bounds for Simulation

N.M. van Dijk

E. van der Sluis

University of Amsterdam

BOUNDS FOR SIMULATION

N.M. van Dijk and E. van der Sluis

University of Amsterdam, Fac. of Economics and Business
Roetersstraat 11, NL-1018 WB, Amsterdam, The Netherlands

Abstract. *To support the simulation of production line structures a modification approach is promoted and illustrated which provides analytic performance bounds. For a number of reasons these bounds can be useful for simulation.*

1 Introduction

A variety of simulation models, such as for manufacturing, telecommunications and service logistics includes serial structures of successive service stations. These service stations quite generally have finite capacity limitations on the number of jobs that they can accommodate. Unfortunately, analytic expressions, most notably so-called product-forms, are usually not available under finite capacity constraints.

Nevertheless, by modifying the system, for the purpose of its evaluation, an analytic solution might be regained which may lead to rough but secure performance bounds, such as for a loss probability, throughput or mean delay. This approach has been described in [1], Chapter 4, and extended in [2] to queuing networks with finite clusters of stations.

So far, however, this approach has not been advocated for the purpose of simulation. More precisely, despite their inaccuracy, these analytic performance bounds can be useful for simulation such as for:

- determining secure orders of magnitude
- verification purposes
- optimization

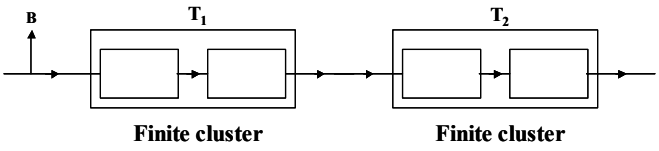
In addition, under specific service disciplines, such as a multi-server or processor sharing discipline, these product form bounds are also insensitive, i.e., they only depend on the service distributions by their means. Sensitivity analysis on the effect of service variability is thus not required.

This note therefore simply aims to promote and illustrate this modification approach as of possible interest to support the simulation for systems of a sequential or production line structure.

2 An instructive example

This section provides an instructive and generic example to illustrate the modification approach and the nature of its results. In the next section four more examples are briefly presented. A more extended outline and formal support of the approach can be found in [1] and [2].

Consider a simple assembly line structure with 4 service stations, numbered 1,…,4 and finite capacity constraints T_1 for the total number of jobs at stations 1 and 2 (cluster1) and T_2 at stations 3 and 4 (cluster 2).



The system has an arrival rate of λ jobs per unit of time and assume that station i has (an exponential) service rate $\mu_i(k)$ when k jobs are present. Let n_i denote the number of jobs at station i, $i = 1,\ldots,4$ and t_j the total number of jobs at cluster j, $j = 1, 2$. ($t_1 = n_1 + n_2$ and $t_2 = n_3 + n_4$). When the first cluster is saturated ($t_1 = T_1$) an arriving job is lost. When the second cluster is saturated ($t_2 = T_2$) the service at cluster 1 (that is at both stations) is stopped.

As simple as the system may look to analyze, there is no simple expression for the loss probability $\boldsymbol{B}$ or the throughput $\boldsymbol{H} = \lambda(1-\boldsymbol{B})$.

As outlined in [2], in order for a system to exhibit a closed product-form expression, a notion of both balance per station (as also presented in [1]) and of balance per cluster (that is, as if a cluster is regarded as one aggregated station) is to be satisfied. But clearly, both notions are violated in the present example, since when $t_1 < T_1$ but $t_2 = T_2$,

- the out-rate of stations 1 and 2 and the out-rate of cluster 1 are necessarily equal to 0 while the in-rate for station 1 (and possibly also for 2) and for cluster 1 are positive.

The following artificial modification to enforce these notions can therefore be suggested.

- When cluster 2 is saturated ($t_2 = T_2$): stop the input.
- When cluster 1 is saturated ($t_1 = T_1$): stop cluster 2 (that is, *both stations* at cluster 2).

Indeed, with $\boldsymbol{n} = (n_1, n_2, n_3, n_4)$ the state description, e_i the unit vector for the i^{th} component and $\mathbf{1}_{\{A\}}$ the indicator of event A, under the above modification one easily verifies the station balance equations at S_U the set of admissible states:

$$S_U = \{\boldsymbol{n} \mid t_1 = n_1+n_2 \le T_1;\ t_2 = n_3+n_4 \le T_2;\ t_1+t_2 \ne T_1+T_2\}$$

as

$$\left\{\begin{array}{ll}
\pi(\mathbf{n})\mu_1(n_1)\mathbf{1}_{(t_2<T_2)} & = \pi(\mathbf{n}-e_1)\lambda\mathbf{1}_{(t_2<T_2)} \\
\pi(\mathbf{n})\mu_2(n_2)\mathbf{1}_{(t_2<T_2)} & = \pi(\mathbf{n}-e_2+e_1)\mu_1(n_1+1)\mathbf{1}_{(t_2<T_2)} \\
\pi(\mathbf{n})\mu_3(n_3)\mathbf{1}_{(t_1<T_1)} & = \pi(\mathbf{n}-e_3+e_2)\mu_2(n_2+1)\mathbf{1}_{(t_1<T_1)} \\
\pi(\mathbf{n})\mu_4(n_4)\mathbf{1}_{(t_1<T_1)} & = \pi(\mathbf{n}-e_4+e_3)\mu_3(n_3+1)\mathbf{1}_{(t_1<T_1)} \\
\pi(\mathbf{n})\lambda\mathbf{1}_{(t_1<T_1)}\mathbf{1}_{(t_2<T_2)} & = \pi(\mathbf{n}+e_4)\mu_4(n_4+1)\mathbf{1}_{(t_1<T_1)}\mathbf{1}_{(t_2<T_2)}
\end{array}\right\} \qquad (1)$$

by the product-form:

$$\pi(\mathbf{n})=\mathrm{c}\lambda^{n_1+n_2+n_3+n_4}\prod_{i=1}^{4}\left[\prod_{k=1}^{n_i}\mu_i(k)\right]^{-1}, \qquad \mathbf{n}\in S_U \qquad (2)$$

with c a normalizing constant. Clearly, the modification leads to an upper bound $\boldsymbol{B}_U \geq \boldsymbol{B}$ for the loss probability. Conversely, also a lower bound product-form modification $\boldsymbol{B}_L$ can be suggested. Hence,

$$\boldsymbol{B}_U = \sum_{\{\mathbf{n}|t_1=T_1 \text{ or } t_2=T_2\}} \pi_U(\mathbf{n}) \tag{3}$$

Below some numerical results are given for the case of single server stations. Here μ_i represents the service speed of station i, $\boldsymbol{B}_L$ and $\boldsymbol{B}_U$ are the easily obtained lower and upper bound for the blocking probability, $\boldsymbol{B}_{av} = (\boldsymbol{B}_L + \boldsymbol{B}_U) / 2$ and $\boldsymbol{B}$ is obtained by numerical computation.

$\mu 1$	$\mu 2$	$\mu 3$	$\mu 4$	$\boldsymbol{T}_1$	$\boldsymbol{T}_2$	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$	$\boldsymbol{B}_{av}$	$\boldsymbol{B}$
1	1	1	1	3	5	.33	.52	.43	.42
1	1	1	1	6	6	.25	.40	.33	.30
1	1	1	1	8	8	.20	.33	.27	.24
2	2	1	1	10	10	.10	.17	.14	.12
1	2	3	2	10	10	.054	.101	.078	.084
1.1	2	3	2	10	10	.021	.065	.048	.049

Table 1. Lower and upper bounds of the loss probability $\boldsymbol{B}$
(and throughput $\boldsymbol{H}$ by $\boldsymbol{H} = \lambda(1\text{-}\boldsymbol{B})$) for finite two-cluster tandem example.

Remark 2.1 (Insensitive bounds)

Product form results are also known to be related to the so-called insensitivity property, provided specific service disciplines are in order such as a multi-server or processor sharing discipline. This property states that the steady state distribution only depends on the service distributions by their means.

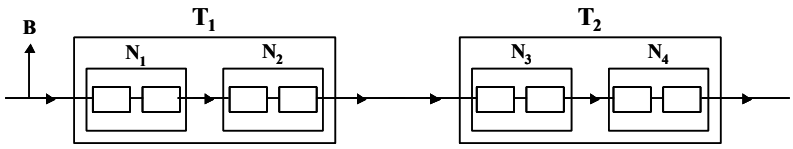
For example, pure multi-sever or for processor sharing disciplines, the product form expression (2) remains valid for arbitrary service distributions with means $1/\mu_i$. Simulation can thus be restricted to one or at most a few service distributions to get a 'good' order of accuracy.

3 Further Illustrative Examples

In this section, we provide some more examples, in line with the instructive example in section 2. For each of these there is no analytic expression known while the modifications guarantee closed product form expressions similar to (2). These in turn will lead to easily computable bounds similar to (3). Some numerical results will be provided which indicate the possible usefulness for simulation.

3.1 A Nested Blocking Structure

As a nested analogue of the example from section 2, in addition to the total cluster constraints $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, we can also allow capacity constraints $\boldsymbol{N}_i$ for each individual station i, $i = 1, ..., 4$.



The following modification now secures (2):

- When $t_1 = \boldsymbol{T}_1$ stop both stations 3 and 4.
- When $t_2 = \boldsymbol{T}_2$ stop both stations 1 and 2 and arrivals.
- When $n_i = \boldsymbol{N}_i$, stop arrivals and all other stations $j \neq i$.

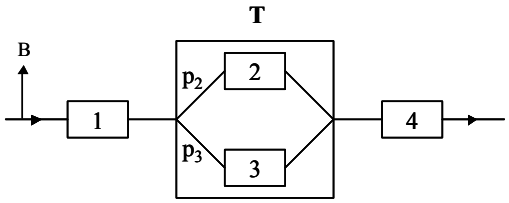
Clearly, this modification leads to an upper bound $\boldsymbol{B}_U$ for the loss probability $\boldsymbol{B}$. Conversely, a lower bound $\boldsymbol{B}_L$ can be suggested. Some numerical results are presented below with μ_i the service parameter of single server station i.

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	$\boldsymbol{N}_1$	$\boldsymbol{N}_2$	$\boldsymbol{N}_3$	$\boldsymbol{N}_4$	$\boldsymbol{T}_1$	$\boldsymbol{T}_2$	$\mathbf{B}_L$	$\mathbf{B}_U$	$\mathbf{B}_{av}$	$\mathbf{B}$
1	1	1	1	1	1	1	1	3	2	4	2	4	5	.471	.724	.598	.572
2	3	4	5	1	2	3	4	3	2	4	2	4	5	.158	.398	.278	.204

Table 2. Results for the nested blocking structure ($\lambda = 1$).

3.2 A Cluster With Parallel Stations

This example contains a random routing to either one of two stations in parallel within one cluster with a capacity constraint $\boldsymbol{T}$ for the total number of jobs at stations 2 and 3, next to capacity constraints $\boldsymbol{N}_i$ at each station i, $i = 1, ..., 4$.



By regarding the cluster as one aggregated station as in section 2, the following modifications lead to product-form expressions:

- Stop arrivals and all stations either when one of stations ($n_i = \boldsymbol{N}_i$) or the cluster ($n_2 + n_3 = \boldsymbol{T}$) is saturated, or
- Stop arrivals when the total number of jobs is equal to $\boldsymbol{N}_1 + \boldsymbol{T} + \boldsymbol{N}_4 = S$, while each station may contain up to S jobs.

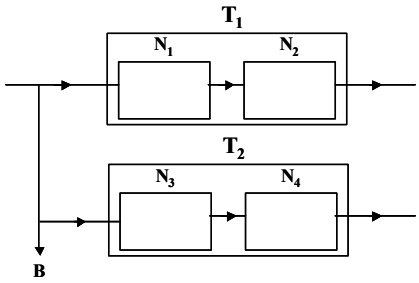
Clearly, the first modification leads to an upper bound $\boldsymbol{B}_U$ and the second to a lower bound $\boldsymbol{B}_L$ for the loss probability $\boldsymbol{B}$ of the original system. Some numerical results are shown below.

μ_1	μ_2	μ_3	μ_4	$\boldsymbol{N}_1$	$\boldsymbol{N}_2$	$\boldsymbol{N}_3$	$\boldsymbol{N}_4$	$\boldsymbol{T}$	p_2	p_3	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$	$\boldsymbol{B}_{av}$	$\boldsymbol{B}$
2	2	2	2	2	2	2	2	3	0.5	0.5	0.03	0.30	0.16	0.16
10	10	10	10	2	2	2	2	4	0.5	0.5	0.00	0.02	0.01	0.01
1	1	1	1	5	5	5	5	10	0.5	0.5	0.10	0.30	0.18	0.20
1	1	1	1	10	5	5	10	10	0.75	0.75	0.06	0.17	0.12	0.10

Table 3. Results for the finite cluster with parallel stations ($\lambda = 1$).

3.3 An Overflow Example

Consider two finite clusters in parallel with arrivals at cluster 1. If a job cannot enter cluster 1 it is rerouted to cluster 2. Each cluster consists of two finite stations in tandem. In addition to the total cluster constraints $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, we also allow capacity constraints $\boldsymbol{N}_i$ for each individual station i, $i = 1, ..., 4$. We assume that $\mu_1 \leq \mu_3$ and $\mu_2 \leq \mu_4$.



For this example, the so-called notion of cluster balance is violated when cluster 2 is busy while cluster 1 is not saturated. In that case the outflow at cluster 2 is positive, but the inrate is 0. The following two modifications are therefore suggested:

- Stop both stations in cluster 2 when cluster 1 is not saturated ($t_1 < \boldsymbol{T}_1$), or
- Assign arriving jobs randomly to either one of the clusters proportional to the free buffer capacity at the two clusters.

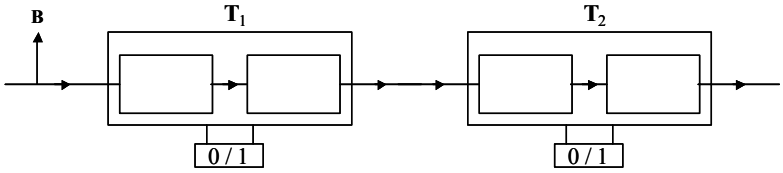
By the first modification cluster 2 is slowed down and kept more congested. The arrival loss probability will thus be enlarged which leads to an upper bound $\boldsymbol{B}_U$ for the loss probability $\boldsymbol{B}$ of the original system. With the second modification, the faster overflow cluster is used more frequently than in the original system, which leads to a lower bound $\boldsymbol{B}_L$.

λ	μ_1	μ_2	μ_3	μ_4	$\boldsymbol{N}_1$	$\boldsymbol{N}_2$	$\boldsymbol{N}_3$	$\boldsymbol{N}_4$	$\boldsymbol{T}_1$	$\boldsymbol{T}_2$	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$	$\boldsymbol{B}_{av}$	$\boldsymbol{B}$
1	1	1	1	1	1	1	1	1	2	2	0.095	0.444	0.270	0.300
2	1	1	4	4	3	3	1	1	6	2	0.005	0.174	0.090	0.075
3	1	1	4	4	3	3	2	2	6	4	0.023	0.126	0.075	0.063

Table 4. Results for parallel finite clusters with overflow.

3.4 Breakdown Model

Reconsider two finite clusters in tandem, which are both subject to breakdowns. In addition to the cluster constraints $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, we assume repair and breakdown rates γ_{10} and γ_{11} for cluster 1, and similarly, γ_{20} and γ_{21} for cluster 2.



Clearly, cluster balance is violated when either cluster is down. The following two modifications are therefore suggested:

- Stop both stations in cluster i when cluster j is down ($j \neq i$), or
- The breakdown rate for both clusters is 0 (breakdowns do not take place).

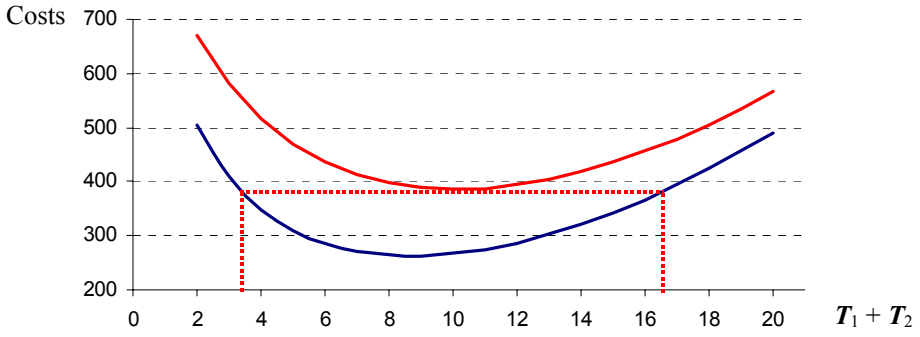
Again, the first modification leads to an upper bound $\boldsymbol{B}_U$ and the second to a lower bound $\boldsymbol{B}_L$ for the loss probability $\boldsymbol{B}$ of the original system. Some numerical results are shown below.

μ_1	μ_2	μ_3	μ_4	N_1	N_2	N_3	N_4	T_1	T_2	γ_{10}	γ_{11}	γ_{20}	γ_{21}	B_L	B_U	B_{av}	B
2	2	2	2	2	2	1	1	4	4	50	1	50	1	0.04	0.42	0.23	0.20
2	1	2	1	2	4	2	4	6	6	50	1	50	1	0.16	0.48	0.32	0.28

Table 5. Results for finite clusters with breakdowns ($\lambda = 1$).

3.5 An Optimal Design Example

Reconsider the finite cluster tandem example from section 2 in which the numbers $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are still be determined by trading off capacity costs $(\boldsymbol{T}_1+\boldsymbol{T}_2)^2$ and opportunity losses $1000\boldsymbol{B}$ due to rejections. Based on the lower and upper bounds for the loss probability, lower and upper bound curves for the costs are easily computed. Despite the large discrepancy between the lower and upper bound values, the qualitative curving behavior seems to almost pinpoint the same optimal number (9 or 10). A simulation can thus at first instance be restricted to these numbers. If one wishes to simulate more, in any case one can be 100% sure that the optimal number is within the region 4-16.



Evaluation

- This technical note advocated the use of a simple analytic approach to support the simulation of production type systems.
- The approach leads to secure analytic (lower and upper) bounds for system performance.
- These bounds can also be insensitive for service distributional forms.
- The approach is based on simple balance insights (to provide analytic modifications for systems that are analytically unsolvable).
- The approach is illustrated and numerically supported by a number of simple but generic structures, which indicate a possible practical usefulness for simulation.

References

[1] Van Dijk, N. M. (1993) *Product forms for Queueing Networks: A system's Approach*, Wiley.

[2] Van Dijk, N. M., Van der Sluis, E. (2002). Simple Product-Form Bounds for Queueing Networks with Finite Clusters. *Annals of Operations Research* 113, 175-195.