

A New Class of Non-Iterative Bounds for Closed Queueing Networks

Giuliano Casale
Politecnico di Milano
DEI
Via Ponzio 34/5
Milan, I-20133, Italy
giuliano.casale@polimi.it

Richard R. Muntz
UCLA
Computer Science Dep.
3277A Boelter Hall
Los Angeles, CA 90095-1596, US
muntz@cs.ucla.edu

Giuseppe Serazzi
Politecnico di Milano
DEI
Via Ponzio 34/5
Milan, I-20133, Italy
giuseppe.serazzi@polimi.it

Abstract

A new emerging class of problems related to the online configuration and optimization of computer systems and networks requires the solution in a very short amount of time of a large number of analytical performance models, often based on queueing networks. In this paper we propose the Geometric Bounds (GB), a new family of fast non-iterative bounds on performance metrics of closed product-form queueing networks. In spite of their simplicity, the proposed bounds are more accurate than the popular Balanced Job Bounds (BJB), even in the difficult case of networks with multiple bottlenecks or including large delays.

1. Introduction

The optimization of enterprise networks requires adaptive techniques in order to cope with the highly variable load characteristics and the resulting dynamic bottleneck shifting. This new class of problems can be approached by solving online a very large number of performance models. As an example, consider self-optimizing and self-configuring systems [1] where the best configuration is selected by maximizing a weighted sum of performance metrics with respect to a number of cost and QoS constraints using nonlinear programming or heuristic methods. Depending on the complexity of the system, hundreds of thousands or even millions of models may be analyzed before a global (or a nearly global) optimum can be found, and the speed at which the optimization algorithm evaluates the performance trade-off of each alternative is a critical issue. Furthermore, bounds on solution accuracy is required for applications where violations to service level agreements are associated to penalties. Similar difficulties emerge in capacity planning and tuning of large host-

ing centers (see, e.g., [5]), where adaptive resource and energy allocation decisions must be evaluated across a wide range of different service demands, and fast algorithms for the solution of the performance models are therefore required. Initial applications of these techniques can also be found in the configuration and in the management of storage area networks, disk arrays, parallel and grid applications [16].

As suggested by the above examples, there is an increasing request for fast bounding techniques. Analytic queueing network models, both open and closed, are often used in this context [3, 21] due to their robustness and the availability of simple solution algorithms and formulas. However, the computational complexity of exact solution techniques, even for basic single class models, make them unfeasible for problems with a very large number of instances to be solved. Since estimates of the performance indices, rather than exact values, are often sufficient to satisfy the requirements of the majority of performance analyses, efficient approximate techniques are typically adopted.

Approximate techniques for the analysis of queueing networks can be divided into two main categories, namely iterative local approximations [2, 4, 19] and single-step bounding techniques [6, 7, 10, 11, 15, 22]. The computational requirements of iterative local approximations are lower than those of exact solution algorithms, but are usually much higher than single-step bounding techniques. On the other hand, the accuracy of iterative techniques is usually higher. Thus, there is a trade-off between computational costs and result accuracy. In what follows we will show that the bounds accuracy can be significantly improved with a small increase in the computational costs. Therefore, using the proposed bounding techniques we do not have to sacrifice accuracy in order to obtain low computational complexity.

This paper describes a single-step bounding tech-

nique for closed single class queueing networks that provides inexpensive and accurate results regardless of the dimensions of the system and of the workload. We introduce the *Geometric Bounds* (GB), a new family of performance bounds that are more accurate than previously proposed bounding techniques. GB bounds are derived by describing the queue-lengths with a geometric sequence of terms related to the resource utilizations. The validation of GB bounds has been performed using known stress cases proposed in the literature for the evaluation of other bounding techniques. In particular, we show in the paper that the GB bounds provide very good results also in the critical case of strongly unbalanced networks, where the BJB are very loose.

The paper is organized as follows. In Section 2 we review popular bounding techniques presented in previous work. Section 3 describes the proposed bounding technique, and Section 4 presents the numerical examples and discusses the computational requirements. Conclusions are drawn in Section 5. Theorem proofs are reported at the end of the paper. Results essential to bound proofs are developed in Appendix A.

2. Related Work

We consider closed networks with M fixed-rate queues, N customers and D delays. The average loading at queue i is $L_i = V_i S_i$, where V_i and S_i are the average visit ratio and the average service time of the station. Queue indices are sorted according to the relative loading of the device, so that queue 1 has the maximum loading L_1 , queue M the minimum L_M . Without loss of generality, when $D > 0$ we consider an equivalent network where the delays are aggregated into a single delay with average loading Z equal to the sum of delays' loadings.

For queue i , we denote the utilization by $U_i(N)$ and the queue-length by $Q_i(N)$. System throughput is denoted by $X(N)$ and can be computed from the queue-lengths of models with $N - 1$ customers using the MVA relation [18]

$$X(N) = \frac{N}{Z + \sum_{j=1}^M L_j [1 + Q_j(N - 1)]} \quad (2.1)$$

the queue-length recurrence relation

$$Q_i(N) = L_i X(N) [1 + Q_i(N - 1)] \quad (2.2)$$

and some general formulas derived from Little Law (e.g., see [12]), i.e., the Utilization Law

$$U_i(N) = L_i X(N) \quad (2.3)$$

the Response time Law

$$R(N) = \frac{N}{X(N)} - Z \quad (2.4)$$

and the population constraint [11]

$$\sum_{i=1}^M Q_i(N) = N - ZX(N). \quad (2.5)$$

Let us review the most popular non-iterative bounds proposed in the literature [10, 11, 15, 22]. Iterative bounding techniques [9] are not considered in this paper because they require much higher computational costs than the methods presented in the next sections. We use the notation α^+ and α^- to indicate upper and lower bounds on a performance metric α (e.g., X^+ and X^- indicate respectively an upper and a lower bound on system throughput).

Let $L = \sum_{j=1}^M L_j$ be the sum of network loadings, and let $R_0 = Z + L$ be the minimum cycle time of the network. The Asymptotic Bounds (ABA) [15] of system throughput can be obtained by assuming that the queue-lengths seen by an arriving customer at all queues are either empty or set to the maximum value $N - 1$ and are given by

$$\frac{N}{R_0 + L(N - 1)} \leq X(N) \leq \frac{N}{R_0} \quad (2.6)$$

The ABA bounds also include the upper bound

$$X_{max}^+ = \frac{1}{L_1} \quad (2.7)$$

where $1/L_1$ is the throughput at which the primary bottleneck queue reaches maximum utilization (i.e., $U_1 = 1$). Note that the system reaches maximum throughput $X(N) = 1/L_1$ for finite populations only when $M = 1$ and $D = 0$. We do not address this case, and we assume that $X(N) < 1/L_1$ for all $N < +\infty$.

It is known that the ABA bounds provide accurate results only when the network is lightly loaded or heavily congested. Let us denote by $L_{ave} = L/M$ the average queue loading. The Balanced Job Bounds (BJB) [22]

$$\frac{N}{R_0 + L_1(N - 1)} \leq X(N) \leq \frac{N}{R_0 + L_{ave}(N - 1)} \quad (2.8)$$

offer greater accuracy than the ABA bounds with a small increase in computational cost. The BJB bounds may be interpreted as throughputs of balanced networks where all loadings have been set equal to L_1 or to L_{ave} . It has been proven in [22] that the throughputs of these systems are upper and lower bounds for the throughput of the original network.

The main limitation of the BJBS is that they cannot distinguish between networks with identical L_1 , L_{ave} and R_0 . Hence, for instance, the lower bound is the lowest of the throughputs of all possible networks with identical L_1 and R_0 . Conversely, the Proportional Bounds (PB) [10] have the important characteristics of considering each individual value of the L_i . For a network without delays, the PB throughput bounds are given by

$$\frac{N}{R_0 + \sigma_N(N-1)} \leq X(N) \leq \frac{N}{R_0 + \delta(N-1)} \quad (2.9)$$

where $\delta = \sum_{i=1}^M L_i^2/L$, $\sigma_N = \sum_{i=1}^M L_i^N (\sum_{j=1}^M L_j^{1-N})$. Note that if $X(N)$ has to be bounded on several populations for the same model, then the δ coefficient can be computed a single time, being independent of N . It has also been proven in [10] that the PB are always tighter than the BJB. In particular, the best accuracy is achieved for lightly-loaded networks, where the average fraction of jobs at queue i is approximately proportional to L_i .

Extended bounds that accurately approximate networks with delays are discussed in [10, 11]. These typically use an estimate of the number of customers waiting at the queues $N_q = \sum_i Q_i(N) = N - ZX(N)$ to improve accuracy. An accurate evaluation of N_q is critical for a tight bound on the performance of a model with delay queues, because a bad estimate of the numerator may greatly affect bound accuracy. Previous work has addressed this point by expressing (2.5) for population $N-1$, and incorporating the resulting equation in the denominators of (2.8). The resulting formulas are nonlinear recurrence equations in the form, e.g.,

$$X^-(N) = \frac{N}{\dots + ZX^-(N-1)}$$

which can be solved in an iterative fashion. Iterative bounding algorithms based on this approach are discussed in [10, 11]. We also point out that non-iterative approximations are also available, and provide an accuracy level that is competitive with that achievable using iterative techniques. For instance, the Square Root Bounds (SQ) [11] employ the relations (see e.g. [8, 11])

$$\frac{N-1}{N} X(N) \leq X(N-1) \leq X(N) \quad (2.10)$$

to remove the dependency from $X(N-1)$ at the denominator of the nonlinear recurrence.

3. Geometric Bounds (GB)

3.1. Queue-length Bounds

We first introduce the Geometric Bounds (GB) for queue-lengths.

Theorem 1 (Pessimistic Queue-length) *In closed product-form networks, the queue-length $Q_i(N)$ is lower bounded by*

$$Q_{i,gb}^-(N) = \frac{y_i(N)}{1-y_i(N)} - \frac{y_i(N)^{N+1}}{1-y_i(N)} \quad (3.1)$$

where $y_i(N) = NL_i/(R_0 + L_1N)$.

Theorem 2 (Optimistic Queue-length) *In closed product-form networks, the queue-length $Q_i(N)$ is upper bounded by*

$$Q_{i,gb}^+(N) = \frac{Y_i(N)}{1-Y_i(N)} - \frac{Y_i(N)^{N+1}}{1-Y_i(N)} \quad (3.2)$$

where $Y_i(N) = U_i^+(N) < 1$, and $U_i^+(N)$ is an upper bound on $U(N)$.

The expressions for $Q_{i,gb}^-(N)$ and $Q_{i,gb}^+(N)$ derive from partial sums of a geometric sequence with common ratio $y_i(N)$ or $Y_i(N)$. Therefore, our bounds are similar to the well-known queue-length formula for open networks with input workload rate λ

$$Q_i(\lambda) = \frac{U_i(\lambda)}{1-U_i(\lambda)} \quad (3.3)$$

that is derived from geometric series.

From (2.3) we see that (3.2) is parametric in the choice of the bound X^+ used to define Y_i . Thus, several possible alternatives may be considered. We can find some useful indications on the best ones by recalling the monotonicity of the partial sum of a geometric sequence $\sum_i r^i$ with respect to the common ratio r . As a consequence, if $X_1^+(N) < X_2^+(N)$, then $Y_i(N) = X_1^+(N)L_i$ gives a tighter bound $Q_{i,gb}^+(N)$ than $Y_i(N) = X_2^+(N)L_i$. Nevertheless, when a large number of bounds are to be computed (e.g., in optimization studies) it may be convenient to use bounds with sub-optimal accuracy, but with reduced computational complexity.

Parametrization of the upper bound. The upper bound X^+ must satisfy the condition

$$Y_i(N) = U_i^+(N) = X^+(N)L_i < 1$$

Let us denote by $n^*(X^+)$ the population corresponding to the intersection point between X^+ and $X_{max} = 1/L_1$. Assuming the saturation condition

$$X^+(N) = X_{max}, \text{ for all } N \geq n^*(X^+)$$

then we have $U_i^+(N) < 1$ for all $L_i < L_1$ because

$$U_i^+(N) \leq X_{max} L_i < X_{max} L_1 = 1$$

When $L_i = L_1$ it is instead $U_1(N) = 1$, for all $N \geq n^*(X^+)$, and the upper GB cannot be employed. Note that the straightforward extension $Q_{1,gb}^+(N) = N$ may not be accurate when the network has $m_1 > 1$ bottleneck queues with loading L_1 . To overcome this limitation, it is sufficient to observe that from (2.5)

$$m_1 Q_1(N) = N - ZX(N) - \sum_{k=m_1+1}^M Q_k(N) \quad (3.4)$$

and defining

$$Q_{1,gb}^+(N) = \frac{1}{m_1} \left[N - ZX_Z^-(N) - \sum_{k=m_1+1}^M Q_{k,gb}^-(N) \right]$$

provided that $0 \leq X_Z^-(N) \leq X(N)$, we finally get

$$Q_{1,gb}^+(N) \geq Q_1(N)$$

The above derivation shows how to account in bounds both for the presence of multiple bottlenecks and for part of the population at the non-bottleneck stations. This concept is fundamental to the throughput bounds presented in the next section.

3.2. Throughput Bounds

From $Q_{i,gb}^-$ and $Q_{i,gb}^+$ we obtain throughput and response time bounds. In the following, in order to simplify the notation we assume $m_1 = 1$. The generalization to networks with $m_1 > 1$ can be done with slight modifications to formulas.

Theorem 3 (Pessimistic Response Time) *In closed product-form queueing networks, the response time $R(N)$ is upper bounded by*

$$R_{gb}^+(N) = L + L_1(N - 1 - ZX_Z^-(N - 1)) + \sum_{i=2}^M d_i Q_{i,gb}^-(N - 1) \quad (3.5)$$

where $d_i = L_i - L_1 \leq 0$, and $X_Z^-(N - 1) \leq X(N - 1)$.

Corollary 1 (Pessimistic Throughput) *In closed product-form networks, the throughput $X(N)$ is lower bounded by*

$$X_{gb}^-(N) = \frac{N}{Z + R_{gb}^+(N)} \quad (3.6)$$

Theorem 4 (Optimistic Response Time) *In closed product-form networks, the response time $R(N)$ is lower bounded by*

$$R_{gb}^-(N) = L + L_1(N - 1 - ZX_Z^+(N - 1)) + \sum_{i=2}^M d_i Q_{i,gb}^+(N - 1) \quad (3.7)$$

where $d_i = L_i - L_1 \leq 0$, and $X_Z^+(N - 1) \geq X(N - 1)$.

Corollary 2 (Optimistic Throughput) *In closed product-form queueing networks, the throughput $X(N)$ is upper bounded by*

$$X_{gb}^+(N) = \frac{N}{Z + R_{gb}^-(N)} \quad (3.8)$$

As observed before, an important innovation of our throughput and response time bounds is that we embed (2.5) directly into the equations by replacing $Q_1(N - 1)$ in all summations by the equivalent expression (3.4). This let us explicit at the denominator of both formulas the term NL_1 which grants the asymptotic correctness of the two bounds. In fact, observing that the GB queue-lengths bounds for the non-bottleneck queues have finite value (being partial sums of geometric sequence with common ratio $y_i(N)$ and $Y_i(N)$ less than one), and observing that

$$\lim_{N \rightarrow \infty} L_1 ZX(N - 1) = L_1 Z \frac{1}{L_1} = Z$$

it is easy to see that (3.5)-(3.8) converge asymptotically to the exact values

$$\lim_{N \rightarrow \infty} X(N) = \frac{1}{L_1}$$

and

$$\lim_{N \rightarrow \infty} R(N) = \lim_{N \rightarrow \infty} R(N) = NL_1 - Z$$

In general, as shown later in the paper, the convergence is remarkably quicker than for BJB and PB bounds.

3.2.1. Geometric Square-root Bounds (GSB)

We discuss how GB throughput bounds could be extended for improved accuracy for queueing networks with delays. The extension to the response times bounds follows straightforwardly by (2.4). Our approach is similar to Kriz's square root bounds [11], and depends on the substitutions

$$\begin{aligned} X_Z^+(N - 1) &= X(N) \\ X_Z^-(N - 1) &= \frac{N - 1}{N} X(N) \end{aligned} \quad (3.9)$$

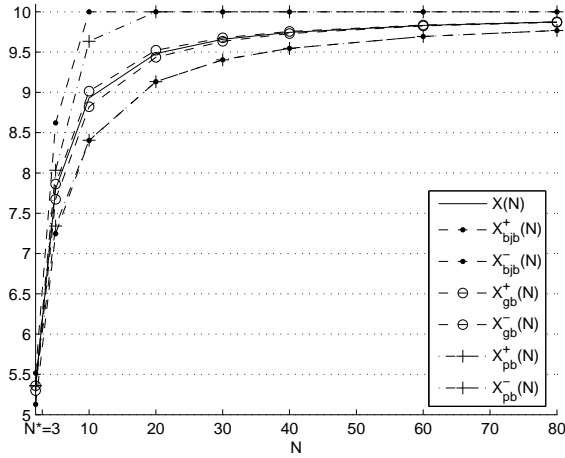


Figure 1. Comparison of the BJB and PB bounds with the GB bounds for Stress Case 3

in the denominator of (3.6) and (3.8). These remove the dependency on X_Z^- and X_Z^+ . (We point to [8, 11] for proofs that (3.9) are bounds on $X(N-1)$ for all populations).

We can derive the following Geometric Square-root Bounds (GSB):

Theorem 5 (Pessimistic Throughput) *In closed product-form networks with $Z > 0$, the throughput $X(N)$ is lower bounded by*

$$X_{gsb}^-(N) = 2N \left[b_1 + \sqrt{b_1^2 - 4ZL_1(N-1)} \right]^{-1} \quad (3.10)$$

where $b_1 = R_0 + L_1(N-1) + \sum_{i=2}^M d_i Q_{i,gb}^-(N)$, $d_i = L_i - L_1$.

Theorem 6 (Optimistic Throughput) *In closed product-form networks with $Z > 0$, the throughput $X(N)$ is upper bounded by*

$$X_{gsb}^-(N) = 2N \left[b_2 + \sqrt{b_2^2 - 4ZL_1N} \right]^{-1} \quad (3.11)$$

where $b_2 = R_0 + L_1(N-1) + \sum_{i=2}^M d_i Q_{i,gb}^+(N)$, $d_i = L_i - L_1$.

We show in the next section that the non-iterative GSB bounds are tighter than the iterative extension of the BJB and PB bounds for networks with delays presented in [10, 11].

4. Numerical Examples and Computational Requirements

We compared the accuracy of the GB and GSB bounds with that of BJB and PB on several hundreds models with different loadings and number of queues. The conclusion that emerges from this kind of analysis is that the relative accuracy of the methods is mainly determined by the number and strength of network bottlenecks, and by the level of balance of network loadings. For this reason, it is more interesting to consider critical cases that are known to be representative of an entire family of models, instead of solving randomly generated networks.

We compare throughput bounds on four known stress cases. The characteristics of the four considered models are known to be critical for bounding techniques, and have been used in previous work [10, 11] to evaluate bound quality.

Stress Case 1 *The network is almost balanced, with $M = 4$ queues and loadings $L_1 = 0.1$, $L_2 = 0.1$, $L_3 = 0.09$, $L_4 = 0.08$. The ABA saturation point is $n^*(X_{aba}^+) = R_0/L_1 = 3.7$ and $X_{max}^+ = 10.000$.*

Stress Case 2 *The network has the same queues of Stress Case 1 and additionally a delay with $Z = 1$. The ABA saturation point is now $n^*(X_{aba}^+) = 13.7$.*

Stress Case 3 *The network is strongly unbalanced, with $M = 4$ queues with $L_1 = 0.1$, $L_2 = 0.1$, $L_3 = 0.05$, and $L_4 = 0.04$. For this model $n^*(X_{aba}^+) = 2.9$ and $X_{max}^+ = 10.000$.*

Stress Case 4 *The network has the same queues of Stress Case 3 and an additional delay $Z = 1$. The new value of the ABA saturation point is $n^*(X_{aba}^+) = 12.9$.*

There are at least two conditions that make the above models stress cases:

1. none of the models is balanced, thus we do not obtain good approximations using the BJB. This let us understand to which degree bounds are able to account for variabilities in loadings
2. a second critical aspect is that $X(N)$ converges very slowly to the asymptotic value $1/L_1$ due to the presence of multiple bottleneck queues [13]. Hence, the range of populations for which the throughput grows before showing saturation effects is larger than in the single bottleneck case. This complicates the approximation, and in particular for the optimistic bounds, whose saturation point is usually independent of the number of bottleneck queues.

N	X(N)	Lower Bounds					
	exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	4.317	4.255	4.317	4.304	-0.062	-0.000	-0.013
5	6.715	6.494	6.660	6.686	-0.221	-0.055	-0.029
10	8.206	7.874	8.022	8.152	-0.332	-0.184	-0.054
15	8.835	8.475	8.569	8.766	-0.360	-0.266	-0.068
20	9.168	8.811	8.867	9.095	-0.358	-0.301	-0.074
30	9.499	9.174	9.194	9.429	-0.325	-0.305	-0.070
40	9.654	9.368	9.375	9.593	-0.286	-0.279	-0.061
60	9.792	9.569	9.570	9.750	-0.223	-0.222	-0.042
80	9.853	9.674	9.674	9.823	-0.179	-0.179	-0.030

Table 1. Results for the lower bounds on throughput for Stress Case 1 of [10, 11]

Tables 1-8 show the lower and upper bound results for the four considered scenarios, respectively. Let us remark that the BJBS are probably the most popular among the existing bounds, while the PB currently provide the best results. The upper GB is computed using the utilization bound $U_i^+ = L_i X_{pb}^+$. The PB and BJB bounds in Stress Case 2 and 4 are computed using their iterative extensions defined in [10,11] with $i = 2$ recursion steps. Thus, the values given by the GB and GSB bounds can be easily compared with the ones given by the BJB and PB bounds. Figure 1 illustrates results for the strongly unbalanced model of Stress Case 3. As can be seen from the results provided by GB and GSB, the proposed bounds are very accurate and much closer to the exact values for the great majority of the models. They are less precise than the PB and occasionally of the upper BJB only for small values of N , that is, $N = 2, 5, 10$. The increase of accuracy is particularly evident in the rapid convergence to the exact value when the network becomes congested (i.e., usually after crossing the $n^*(X_{aba}^+)$ saturation point). Note also that the maximum absolute error of GB is often very close to the minimum absolute error of BJB and PB for medium and heavy load conditions. In conclusion, the results strongly indicate the effectiveness and the robustness of the proposed bounds on both nearly balanced and strongly unbalanced models. We also point out that the observations of this section and of Example 1 extend to larger networks, even those composed by hundreds of servers, with large delays and populated by thousands of customers. In particular, the best improvements with respect to previous work are obtained in the most difficult case of networks with strongly unbalanced queues, and with multiple primary and secondary bottlenecks.

Computational Complexity Let us now consider the computational cost of the GB bounds, and let us begin with the queue-length bounds. In general, the number of operations required to compute (3.1) and

N	X(N)	Lower Bounds					
	exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	1.434	1.432	1.434	1.432	-0.002	-0.000	-0.001
5	3.364	3.267	3.288	3.347	-0.097	-0.075	-0.017
10	5.872	5.390	5.440	5.783	-0.482	-0.432	-0.088
15	7.468	6.680	6.728	7.294	-0.788	-0.740	-0.174
20	8.375	7.489	7.525	8.162	-0.886	-0.850	-0.213
30	9.186	8.393	8.408	8.990	-0.793	-0.777	-0.195
40	9.502	8.858	8.864	9.348	-0.644	-0.638	-0.154
60	9.740	9.306	9.307	9.646	-0.433	-0.432	-0.094
80	9.828	9.514	9.514	9.768	-0.314	-0.314	-0.060

Table 2. Results for the lower bounds on throughput for Stress Case 2 of [10, 11]

N	X(N)	Lower Bounds					
	exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	5.360	5.128	5.360	5.299	-0.232	-0.000	-0.062
5	7.803	7.246	7.341	7.673	-0.556	-0.461	-0.130
10	8.930	8.403	8.407	8.822	-0.527	-0.523	-0.109
15	9.302	8.876	8.876	9.231	-0.427	-0.427	-0.071
20	9.483	9.132	9.132	9.435	-0.350	-0.350	-0.048
30	9.659	9.404	9.404	9.634	-0.255	-0.255	-0.025
40	9.746	9.547	9.547	9.730	-0.199	-0.199	-0.015
60	9.831	9.693	9.693	9.824	-0.138	-0.138	-0.007
80	9.874	9.768	9.768	9.870	-0.106	-0.106	-0.004

Table 3. Results for the lower bounds on throughput for Stress Case 3 of [10, 11]

(3.2) depend on the time required to compute the terms $y(N)$ and $Y(N)$, and on the additional operations for determining the final expressions of $Q_{i,gb}^-$ and $Q_{i,gb}^+$. Let us denote the computational cost of the most expensive arithmetic operations as follows

- K is the time for performing a multiplication
- D is the time for performing a division
- E is the time for performing an exponentiation

Let us denote by T^- and T^+ the time required to compute X^- and X^+ in $y_i(N)$ and $Y_i(N)$, respectively. Then $Q_{i,gb}^-$ requires

$$T^- + K + 2D + 1E$$

time. Similarly $Q_{i,gb}^+$ has a time requirement of

$$T^+ + 1D + 1E$$

In all cases where even a slight decrease of the above requirements may produce considerable speedups, it is possible to avoid the exponentiation by introducing an additional approximation. This is in general very tight for the queue-lengths of the non-bottleneck queues, which we recall are the only ones required to compute

N	Lower Bounds						
	X(N) exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	1.528	1.524	1.528	1.527	-0.004	-0.000	-0.001
5	3.628	3.476	3.489	3.608	-0.152	-0.139	-0.020
10	6.422	5.684	5.685	6.281	-0.738	-0.737	-0.141
15	8.133	6.978	6.979	7.842	-1.155	-1.155	-0.291
20	8.945	7.767	7.767	8.640	-1.179	-1.179	-0.306
30	9.483	8.618	8.618	9.302	-0.864	-0.864	-0.181
40	9.659	9.042	9.042	9.555	-0.617	-0.617	-0.104
60	9.797	9.437	9.437	9.752	-0.360	-0.360	-0.045
80	9.856	9.614	9.614	9.831	-0.241	-0.241	-0.024

Table 4. Results for the lower bounds on throughput for Stress Case 4 of [10, 11]

N	Upper Bounds						
	X(N) exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	1.434	1.434	1.434	1.522	0.000	0.000	0.088
5	3.364	3.402	3.400	3.568	0.038	0.036	0.204
10	5.872	6.270	6.263	6.174	0.398	0.391	0.302
15	7.468	8.621	8.606	7.872	1.153	1.138	0.405
20	8.375	9.081	9.053	8.666	0.706	0.678	0.291
30	9.186	9.592	9.549	9.344	0.406	0.363	0.159
40	9.502	9.870	9.818	9.616	0.368	0.316	0.113
60	9.740	10.000	10.000	9.804	0.260	0.260	0.065
80	9.828	10.000	10.000	9.859	0.172	0.172	0.032

Table 6. Results for the upper bounds on throughput for Stress Case 2 of [10, 11]

N	Upper Bounds						
	X(N) exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	4.317	4.324	4.317	4.355	0.007	0.000	0.038
5	6.715	6.757	6.730	6.800	0.042	0.015	0.085
10	8.206	8.316	8.270	8.300	0.110	0.064	0.094
15	8.835	9.009	8.953	8.926	0.174	0.118	0.091
20	9.168	9.401	9.339	9.258	0.233	0.171	0.089
30	9.499	9.828	9.759	9.590	0.329	0.260	0.090
40	9.654	10	9.984	9.748	0.346	0.330	0.094
60	9.792	10.000	10.000	9.836	0.208	0.208	0.043
80	9.853	10.000	10.000	9.877	0.147	0.147	0.024

Table 5. Results for the upper bounds on throughput for Stress Case 1 of [10, 11]

N	Upper Bounds						
	X(N) exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	5.360	5.517	5.360	5.499	0.157	0.000	0.138
5	7.803	8.621	8.033	7.947	0.818	0.231	0.144
10	8.930	10.000	9.635	9.042	1.070	0.704	0.112
15	9.302	10.000	10.000	9.375	0.698	0.698	0.073
20	9.483	10.000	10.000	9.524	0.517	0.517	0.041
30	9.659	10.000	10.000	9.677	0.341	0.341	0.018
40	9.746	10.000	10.000	9.756	0.254	0.254	0.010
60	9.831	10.000	10.000	9.836	0.169	0.169	0.005
80	9.874	10.000	10.000	9.877	0.126	0.126	0.003

Table 7. Results for the upper bounds on throughput for Stress Case 3 of [10, 11]

the GB bounds for throughput and response time. Observing that $y_i^n = e^{n \log y_i}$, we have immediately the following bound

$$Q_{i,gb}^-(N) \geq \frac{y_i(N) - e^{(N+1)\lceil \log y_i(N) \rceil}}{1 - y_i(N)}$$

where $\lceil x \rceil$ rounds x towards $+\infty$. The above lower bound is denoted by $Q_{i,gblog}^-(N)$. Since $0 \leq \lceil \log y_i(N) \rceil \leq 4$ for $y_i(N) \in [0.01, 1]$, the derivation shows that it is sufficient to precompute a few values of the integer function e^n , e.g., for $n \in \{0, 1, \dots, 4\}$, to save the computational cost of the exponentiation in $Q_{i,gb}^-$ on the great majority of the queue-length bounds. With an analogous derivation we obtain the following upper bound $Q_{i,gblog}^+$ on the queue-length

$$Q_{i,gb}^+(N) \leq \frac{Y_i(N) - e^{(N+1)\lfloor \log Y_i(N) \rfloor}}{1 - Y_i(N)}$$

where $\lfloor x \rfloor$ rounds x towards $-\infty$ and similar considerations apply. The reduced time complexities of the queue-length bounds approximations $Q_{i,gblog}^-$ and $Q_{i,gblog}^+$ are

$$T^- + 3K + 2D$$

and

$$T^+ + 2K + 1D$$

respectively. The modest accuracy loss due to the introduced approximation is estimated in Table 9 for different subsets of queues, all with utilization U_i below the indicated threshold. The estimates were obtained on a sample of 1000 random models with randomly chosen $2 \leq M \leq 100$, and $2 \leq N \leq 1000$. The estimates refer to the following error function Let us now consider the GB throughput bounds (3.6) and (3.8). We do not explicitly consider the complexity of (3.5) and (3.7), because it can be immediately obtained from the requirements of (3.6) and (3.8) with one less division operation. We make the assumption that all queue-length bounds in the denominator of (3.6) and (3.8) are obtained from a single pair of bounds X^- or X^+ which can be computed in T^- and T^+ time. Denoting by Z^- and Z^+ the time required to compute X_Z^- and X_Z^+ , the overall time requirements are

$$Z^- + T^- + (4M + 2)K + (M + 1)D + (M + 1)E$$

for X_{gsb}^- , and

$$Z^+ + T^+ + (4M + 1)K + (M + 1)D + (M + 1)E$$

N	X(N)	Upper Bounds					
	exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	1.528	1.531	1.528	1.642	0.003	0.000	0.113
5	3.628	3.690	3.664	3.916	0.062	0.036	0.288
10	6.422	6.960	6.858	6.882	0.538	0.436	0.460
15	8.133	9.494	9.245	8.527	1.361	1.112	0.394
20	8.945	10.000	9.814	9.134	1.055	0.868	0.188
30	9.483	10.000	10.000	9.534	0.517	0.517	0.051
40	9.659	10.000	10.000	9.681	0.341	0.341	0.022
60	9.797	10.000	10.000	9.805	0.203	0.203	0.007
80	9.856	10.000	10.000	9.859	0.144	0.144	0.004

Table 8. Results for the upper bounds on throughput for Stress Case 4 of [10, 11]

	U_i	Δ_{gblog}			
		median	mean	std	max
$Q_{i,gblog}^-$	< 1	0.02164	0.02633	0.01974	0.20920
$Q_{i,gblog}^-$	< 0.9	0.01317	0.01671	0.01389	0.13526
$Q_{i,gblog}^-$	< 0.75	0.00479	0.00855	0.01084	0.12050
$Q_{i,gblog}^-$	< 0.5	0.00000	0.00000	0.00001	0.00019
$Q_{i,gblog}^+$	< 1	0.09095	0.12781	0.12937	0.76343
$Q_{i,gblog}^+$	< 0.9	0.01413	0.02283	0.02665	0.25232
$Q_{i,gblog}^+$	< 0.75	0.00461	0.00913	0.01660	0.24800
$Q_{i,gblog}^+$	< 0.5	0.00000	0.00001	0.000s20	0.00632

Table 9. Error estimate based for employing the GBLOG queue-length bound instead of the GB bounds. Only the subset of stations with utilization U_i below the indicated threshold is considered in each row.

for X_{gsb}^+ . Hence, the computational complexity of the proposed throughput and response time bounds is $O(M)$, that grows linearly with the number of queues M of the model and is independent of N .

5. Conclusions

New problems related, among others, to self-optimizing and self-configurable systems require the online solution of a large number of queueing models. In this paper we proposed the Geometric Bounds (GB), a fast and accurate bounding technique for the computation of performance measures which are frequently used by QoS control algorithms, as queue-lengths, throughputs and response times. We have shown that the proposed GB bounds, in spite of their simple formulas which are related to partial sums of geometric sequences, are more accurate than known bounding techniques even on unbalanced models with delays and multiple bottlenecks, which are the hardest to approximate.

6. Acknowledgments

The authors wish to thank Emilia Rosti for her helpful suggestions that significantly improved this paper.

References

- [1] *Proceedings of the 2nd IEEE International Conf. on Autonomic Computing (ICAC-05)*. IEEE Press, 2005.
- [2] Y. Bard. Some extensions to multiclass queueing network analysis. In M.Arato, A.Butrimenko, and E.Gelembe, editors, *Proc. 3rd Int'l Symp. on Model. and Perf. Eval. of Comp. Sys.*, pages 51–62, 1979.
- [3] M.N. Bennani and D. A. Menascè. Resource allocation for autonomic data centers using analytic performance. In *2nd IEEE Int'l Conf. on Autonomic Comp.*, Jun 2005.
- [4] K.M. Chandy and D. Neuse. Linearizer: A heuristic algorithm for queueing network models of computing systems. *Comm. ACM*, 25(2):126–134, 1982.
- [5] J.S. Chase, D.C. Anderson, P.N. Thakar, A.M. Vahdat, and R. P. Doyley. Managing energy and server resources in hosting centers. In *18th Symposium on Operating Systems Principles (SOSP)*, Oct 2001.
- [6] W.C. Cheng and R.R. Muntz. Bounding errors introduced by clustering of customers in closed product-form queueing networks. *J.ACM*, 43(4):641–669, 1996.
- [7] P.J. Denning and J.P. Buzen. The operational analysis of queueing network models. *ACM Comp. Surv.*, 10(3):225–261, 1978.
- [8] L.W. Dowdy, D.L. Eager, K.D. Gordon, and L. V. Saxton. Throughput concavity and response time convexity. *Inform. Proc. Letters*, 19(4):209–212, 1984.
- [9] D.L. Eager and K.C. Sevcik. Bound hierarchies for multiple-class queueing networks. *J.ACM*, 33(1):179–206, 1986.
- [10] C.H. Hsieh and S. Lam. Two classes of performance bounds for closed queueing networks. *Perf. Eval.*, 7(1):3–30, 1987.
- [11] J. Kriz. Throughput bounds for closed queueing networks. *Perf. Eval.*, 4(1):1–10, 1984.
- [12] E.D. Lazowska, J. Zahorjan, G.S. Graham, and K.C. Sevcik. *Quantitative System Performance*. Prentice-Hall, 1984.
- [13] L. Lipsky, C.H. Lieu, A. Tehranipour, and A. van de Liefvoort. On the asymptotic behavior of time-sharing systems. *Comm. ACM*, 25(10):707–714, 1982.
- [14] G.S. Lueker. Some techniques for solving recurrences. *ACM Comp. Surv.*, 12(4):419–436, 1980.
- [15] R.R. Muntz and J.W. Wong. Asymptotic properties of closed queueing network models. In *Proc. of the 8th Annual Princeton Conf. on Inf. Science and Sys.*, pages 348–352, 1974.
- [16] T. Nowicki, M. S. Squillante, and C. W. Wu. Fundamentals of dynamic decentralized optimization in autonomic computing systems. In *Self-star Properties in Complex Information Systems – LCNS 3460*, 2005.

- [17] F.W.J. Olver. *Asymptotics and Special Functions*. Academic Press, 1974.
- [18] M. Reiser and S.S. Lavenberg. Mean-value analysis of closed multichain queueing networks. *J.ACM*, 27(2):312–322, 1980.
- [19] P.J. Schweitzer. Approximate analysis of multiclass closed networks of queues. In *Int'l Conf. on Stochastic Control and Optim., Amsterdam*, pages 25–29, 1979.
- [20] K.C. Sevcik and I. Mitrani. The distribution of queueing network states at input and output instants. *J.ACM*, 28(2):358–371, 1981.
- [21] L. Wynter, C.H. Xia, and F. Zhang. Parameter inference of queueing models for it systems using end-to-end measurements. In *Proc. of the 2004 ACM Sigmetrics/Performance Intl. Conf.*, pages 408–409, 2004.
- [22] J. Zahorjan, K.C. Sevcik, D.L. Eager, and B. Galler. Balanced job bound analysis of queueing networks. *Comm. ACM*, 25(2):134–141, 1982.

Appendix A

This appendix discusses exact and approximate solutions of the recurrence relation

$$f(N) = C(N) [1 + f(N-1)] = \sum_{n=1}^N \prod_{k=1}^n C(N-k+1) \quad (6.1)$$

where $C(N)$ is an arbitrary function of N , and with termination condition $f(0) = 0$. Clearly, (2.2) belongs to this class of recurrence relation. An overview of solution techniques can be found in [14]. In order to obtain simple non-iterative approximations of (6.1), we seek for tight upper and lower bounds on $f(N)$. Denote by

$$Geom(r, n) = \sum_{i=1}^n r^i = \frac{r - r^{n+1}}{1 - r} \quad (6.2)$$

a geometric sequence with common ratio $r \neq 1$. For $r = 1$ we set $Geom(1, n) = n$.

Theorem 7 *The solution of (6.1) is bounded by*

$$Geom(C^-, N) \leq f(N) \leq Geom(C^+, N) \quad (6.3)$$

for all $N \geq 1$, where

$$C^- = \min_{n:1 \leq n \leq N} C(n), \quad C^+ = \max_{n:1 \leq n \leq N} C(n) \quad (6.4)$$

In general, the quality of the approximation greatly depends on the structure of the $C(N)$ terms. Henceforth, we focus on coefficients $C(N)$ that may be written in the form

$$C(n) = \frac{an}{b+n} \quad \text{for } 1 \leq n \leq N \quad (6.5)$$

where $a > 0$ and $b \geq 0$ are two real constants. Furthermore, let us note that for $b = 0$ the recurrence (6.1) becomes a simple geometric sequence which may be solved non-iteratively by (6.2). Hence, we focus in the rest of the section on the recurrences where $b > 0$. In this case we see that $C^- = C(1)$ and $C^+ = C(N)$. We introduce an exact solution formula for (6.1). Denote by $(x)_k = x(x+1) \cdots (x+k-1)$ and $x^{(k)} = x(x-1) \cdots (x-k+1) = (-1)^k (-x)_k$ the *rising factorial* and the *falling factorial*, respectively, and introduce the following special function

$${}_2F_1(\alpha, \beta; \gamma; x) = \sum_{k=0}^{\infty} \frac{(\alpha)_k (\beta)_k}{(\gamma)_k} \frac{x^k}{k!}. \quad (6.6)$$

For properties of ${}_2F_1$ we point to [17].

Theorem 8 *The solution of (6.1) is given by*

$$f(N) = {}_2F_1(1, -N; -b - N; a) - 1 \quad (6.7)$$

for all $N \geq 1$.

We remark that the above expression relates the solution of a recurrence with a finite number of terms to a series ${}_2F_1$. Currently, non-iterative expressions of (6.7) are available only for special values of the coefficients of (6.6) [17]. Let us then evaluate the accuracy of the bounds in Theorem 7. A numerical inspection leads to the conclusion that $Geom(C(N), N)$ is a very tight upper bound, while the lower bound $Geom(C(1), N)$ is rather loose, as suggested by the following example.

Example 1 Let us consider a recurrence (6.1) with coefficients (6.5) where $a = 0.53$, $b = 0.62$ and $N = 45$. A recursive evaluation yields $f(N) = 1.0947$. Equivalently, we obtain from (6.7) the value $f(N) = {}_2F_1(1, -45; -45.62; 0.53) - 1 = 1.0947$. The bounds (6.3) are $Geom(C(N), N) = 1.0956$, with a relative error $\Delta_{rel,ub} = 5\%$, and $Geom(C(1), N) = 0.4863$, with a relative error $\Delta_{rel,lb} = 56\%$.

The tightness of $Geom(C(N), N)$ can be explained by unfolding $f(N)$ as

$$f(N) = C(N) + C(N)C(N-1) + C(N)C(N-1)C(N-2) + \cdots \quad (6.8)$$

and noting that sequence of terms is monotonically decreasing, that is,

$$C(N) > C(N)C(N-1) > C(N)C(N-1)C(N-2) > \dots$$

it is easy to conclude that, the approximation of the largest terms provided by the upper bound $Geom(C(N), N) = C(N) + C(N)^2 + \dots$ is more accurate the one given by $Geom(C(1), N) = C(1) + C(1)^2 + \dots$, because the leading terms in

$Geom(C(N), N)$ are very close to the related terms in $f(N)$. We introduce an improved lower bound by an additional assumption on the coefficients (6.5).

Theorem 9 *If $a \leq 1$ in (6.5), then the related $f(N)$ is lower bounded by*

$$f(N) \geq Geom\left(\frac{N}{N+1}C(N+1), N\right) \quad (6.9)$$

for all $1 \leq N < +\infty$.

Note that the interpolation of $C(N+1)$ may be bigger or smaller of the terms $C(n)$, $1 \leq n \leq N$, and this means that the $NC(N+1)/(N+1)$ in general is not an upper bound on $C(n)$ as C^+ . We conclude by showing the accuracy of (6.9) with the following example.

Example 2 Let us consider the case of Example 1. Denoting by $C' = NC(N+1)/(N+1) = 0.5116$, the lower bound on $f(N)$ becomes $Geom(C', N) = 1.0474$, with a new error of $\Delta'_{rel,lb} = 4\% \ll \Delta_{rel,lb} = 56\%$.

Appendix B

PROOF OF THEOREM 1. Consider the relation

$$Q_i^-(N) = X_{bjb}^-(N)L_i[1 + Q_i^-(N-1)] \quad (6.10)$$

with $Q_i^-(0) = 0$, where $X_{bjb}^- = N/(R_0 + L_1(N-1))$ is the lower BJB (2.8). By comparing the above expression with (2.2) we see that it is indeed a lower bound on $Q_i(N)$. Further, if we set $C(N) = X_{bjb}^-(N)L_i$, we see that $0 < a \equiv L_i/L_1 \leq 1$ and $b \equiv (R_0/L_1 - 1) \geq 0$. Then, the theorem follows easily from Theorem 9, and from (6.2) in the case $b = 0$.

PROOF OF THEOREM 2. It is sufficient to apply Theorem 7 of Appendix A and (6.2) to the recurrence

$$Q_i^+(N) = U_i^+(N)[1 + Q_i^+(N-1)] \quad (6.11)$$

with $Q_i^+(0) = 0$ that defines an upper bound on $Q_i(N)$.

PROOF OF THEOREM 3. From the Arrival Theorem [18, 20] we know that $R(N) = L + \sum_i L_i Q_i(N-1)$. Expressing (2.5) for $N-1$, and solving for $Q_1(N-1)$ we get $Q_1(N-1) = N-1 - ZX(N-1) - \sum_{i \neq 1} Q_i(N-1)$. Substituting in $R(N)$ we finally obtain

$$R(N) \geq L + L_1(N-1 - ZX(N-1)) + \sum_{i=2}^M d_i Q_i^-(N-1)$$

where $d_i = L_i - L_1 \leq 0$.

PROOFS OF COROLLARIES 1 AND 2. The corollaries are proved immediately by (2.4).

PROOF OF THEOREM 4. Analogous to the proof of Theorem 3.

PROOF OF THEOREMS 5-6. The proofs are analogous to the original Square-root bounds proofs [11] applied to equations (3.6) and (3.8).

PROOF OF THEOREM 7. The theorem follows easily from (6.1) and (6.2) observing that

$$\sum_{n=1}^N (C^-)^n \leq \sum_{n=1}^N \prod_{k=1}^n C(N-k+1) \leq \sum_{n=1}^N (C^+)^n$$

PROOF OF THEOREM 8. From (6.1) and (6.5) we have

$$\begin{aligned} f(N) &= \sum_{k=1}^N \frac{a^k N \cdots (N-k+1)}{(b+N) \cdots (b+N-k+1)} \\ &= \sum_{k=1}^N \frac{a^k (-N)_k}{(-b-N)_k} \quad (6.12) \end{aligned}$$

Now using the relation $(1)_k = k!$, by (6.6) we get

$$f(N) = \sum_{k=1}^N \frac{(-N)_k (1)_k a^k}{(-b-N)_k k!} = {}_2F_1^N(1, -N; -b-N; a) - 1$$

where ${}_2F_1^N$ denotes the N -th partial sum of ${}_2F_1$. The theorem follows noting that $(-N)_k = 0$ for $k \in \{N+1, \dots, +\infty\}$, we have that the terms of the series ${}_2F_1(1, -n; -b-n; a)$ are all equal to zero after the N -th term and hence ${}_2F_1^N = {}_2F_1$

PROOF OF THEOREM 9. Proof by induction on N .

Case $N = 1$: we need to prove that $f(1) \geq Geom(C(2)/2, 1)$ that simplifies to $C(1) \geq C(2)/2$. Inserting (6.5) this can be simplified as $b \geq 0$ that is always true by definition of $C(N)$.

Induction step: let us define $D(N) = (N-1)C(N)/N = a(N-1)/(b+N)$. Note that it is $0 \leq D(N) < 1$ by the assumptions $0 < a \leq 1$, $b \geq 0$, and for $N \geq 1$. The induction hypothesis is

$$f(N-1) \geq Geom(D(N), N-1)$$

Inserting the above expression in (6.1) and using (6.2), the induction hypothesis implies

$$f(N) = C(N)[1 + f(N-1)] \geq C(N) \frac{1 - D(N)^N}{1 - D(N)}$$

We wish to prove that

$$f(N) \geq Geom(D(N+1), N) = D(N+1) \frac{1 - D(N+1)^N}{1 - D(N+1)}$$

Note that the theorem is proved if we can show that

$$\Delta = C(N) \frac{1 - D(N)^N}{1 - D(N)} - D(N+1) \frac{1 - D(N+1)^N}{1 - D(N+1)} \geq 0$$

but this can be easily verified, using the conditions on a and b , by substituting the definitions of $D(N)$ and $C(N)$ into the inequality.