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 - CS 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 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 are required for applications where violations to service level agreements are associated to penalties.

Hence, there is an increasing demand for fast bounding techniques. Analytic queueing network models, both open and closed, are often used in this context (e.g., [2]) 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 evaluation studies, efficient approximate techniques may be adopted.

The computational requirements of iterative local approximations (e.g., [3]) are lower than those of exact solution algorithms, but are usually much higher than single-step bounding techniques [4, 5, 8, 9, 11, 15]. 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 to obtain low computational complexity.

In this paper 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 the GB bounds has been performed using well-known stress cases proposed in the literature for the evaluation of other bounding techniques. In particular, we will show that the GB bounds provide very good results also in the critical case of unbalanced networks, where existing bounds 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 numerical examples and discusses the computational requirements. Conclusions are drawn in Section 5. Theorem proofs are reported in the appendices.

*The work of G. Casale and G. Serazzi was partially supported by the Italian FIRB-Perf project.

2 Related Work

We consider closed networks with M fixed-rate queues, N customers and D delay stations. The average loading of queue i is $L_i = V_i S_i$, where V_i and S_i are respectively 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, we consider an equivalent network where the delay stations are aggregated into a single delay station with delay Z. For a model with N customers, we denote the utilization of queue i by $U_i(N)$, and its queue-length by $Q_i(N)$. The throughput is denoted by $X(N)$ and can be computed from the queue-lengths of models with $N-1$ customers using the MVA relation [13]

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

and the queue-length recurrence relation

$$Q_i(N) = L_i\, X(N) \left[1 + Q_i(N-1)\right]. \tag{2.2}$$

Throughout the paper, we will use some formulas derived from Little's Law, i.e., the Utilization Law

$$U_i(N) = L_i\, X(N), \tag{2.3}$$

for all queues i, $1 \le i \le M$; the Response time Law

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

which allows us to compute the average time spent at the queues, and the population constraint

$$\sum_{i=1}^{M} Q_i(N) + Z X(N) = N, \tag{2.5}$$

which captures the fact that the number of customers in the network is always equal to N.

Let us now review the most popular non-iterative bounds proposed in the literature [8, 9, 11, 15]. Iterative bounding techniques [7–9] are not considered in this paper because they require much higher computational costs than the non-iterative ones. 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 throughput bound).

Let $L = \sum_{j=1}^{M} L_j$ be the sum of loadings, and let $R_0 = Z + L$ be the minimum customer cycle time of the network. The Asymptotic Bounds (ABA) [11] 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

$$\frac{N}{R_0 + L(N-1)} \le X(N) \le \frac{N}{R_0}. \tag{2.6}$$

The ABA bounds also include the upper bound

$$X^+_{max} = \frac{1}{L_1}, \tag{2.7}$$

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 will assume that $X(N) < 1/L_1$ for all $N < +\infty$. Hence, the utilization terms will be $U_i = L_i\, X(N) < 1$, for all $1 \le i \le M$.

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) [15]

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

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 shown [15] that the throughputs of these systems are upper and lower bounds on $X(N)$.

The main limitation of the BJB 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) [8] have the important characteristics of considering each individual value of the L_i. For a network without delay stations, the PB throughput bounds are given by

$$\frac{N}{R_0 + \sigma_N(N-1)} \le X(N) \le \frac{N}{R_0 + \delta(N-1)}, \tag{2.9}$$

where

$$\sigma_N = \frac{\sum_{i=1}^{M} L_i^N}{\sum_{j=1}^{M} L_j^{N-1}}, \qquad \delta = \frac{\sum_{i=1}^{M} L_i^2}{\sum_{j=1}^{M} L_j} = \sigma_2.$$

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 shown in [8] 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.

3 Geometric Bounds (GB)

3.1 Queue-length Bounds

In this section we derive the upper and lower Geometric Bounds (GB) for queue-lengths.

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

$$Q^-_{i,gb}(N) = \frac{y_i(N)}{1-y_i(N)} - \frac{y_i(N)^{N+1}}{1-y_i(N)} \tag{3.1}$$

where $y_i(N) = NL_i/(R_0 + L_1 N)$.

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

$$Q^+_{i,gb}(N) = \frac{Y_i(N)}{1-Y_i(N)} - \frac{Y_i(N)^{N+1}}{1-Y_i(N)} \tag{3.2}$$

where $Y_i(N) = U^+_i(N) = L_i\, X^+(N)$, and $U^+_i(N)$ is any upper bound on $U_i(N)$ obtained from any X^+ such that $X(N) \le X^+ \le X_{max}$, so that $U^+_i(N) < 1$ for all queues with $L_i < L_1$.

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 $Y_i(N)$. This explains why our bounds are similar to the well-known queue-length formula for open networks with input workload rate λ

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

which is derived from geometric series.

Parametrization of the upper bound.

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 the most accurate by recalling the monotonicity of the partial sum of a geometric sequence $\sum_i r^i$ with respect to the common ratio $r \ge 0$. As a consequence, if $X^+_1(N) < X^+_2(N)$, then $Y_i(N) = X^+_1(N)L_i$ gives a tighter $Q^+_{i,gb}(N)$ than $Y_i(N) = X^+_2(N)L_i$. Nevertheless, when a large number of bounds are to be computed, as in optimization studies, it may be convenient to use bounds with sub-optimal accuracy, but with reduced computational complexity. Note that, in any case, 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 the ABA bound $X_{max} = 1/L_1$. Assuming the saturation condition

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

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

$$U^+_i(N) \le X_{max}L_i < X_{max}L_1 = 1.$$

When $L_i = L_1$ it is instead $U^+_1(N) = 1$, for all $N \ge n^*(X^+)$, 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 all with the same loadings $L_1 = L_2 = \ldots = L_{m_1} = \max_{1\le j\le M} L_j$. To overcome these limitations, it is sufficient to observe that from (2.5) it is

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

where the summation considers the non-bottleneck stations only, and define

$$Q^+_{1,gb}(N) = \frac{1}{m_1}\left(N - ZX^-_Z(N) - \sum_{k=m_1+1}^{M} Q^-_{k,gb}(N)\right),$$

provided that $0 \le X^-_Z(N) \le X(N)$. By construction, it is clear that the last quantity is an upper bound on $Q_1(N)$, i.e.

$$Q^+_{1,gb}(N) \ge Q_1(N).$$

The above derivation shows how to account in bounds both for the presence of multiple bottlenecks and for the fraction of the population at the non-bottleneck stations. This derivation is exploited also in 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 easily done.

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

$$R^+_{gb}(N) = L + L_1(N - 1 - ZX^-_Z(N-1)) + \sum_{i=2}^{M} d_i Q^-_{i,gb}(N-1), \tag{3.5}$$

where $d_i = L_i - L_1 \le 0$, and X^-_Z is a lower bound on X, i.e., $X^-_Z(N-1) \le X(N-1)$.

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

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

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

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

where $d_i = L_i - L_1 \leq 0$, and X_Z^{+} is an upper bound on X, i.e., $X_Z^{+}(N-1) \geq X(N-1)$.

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

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

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 lets us make explicit at the denominator of both formulas the term NL_1 that 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

$$\lim_{N\to\infty} L_1 Z X(N-1) = L_1 Z \frac{1}{L_1} = Z,$$

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

$$\lim_{N\to\infty} X_{gb}^{-}(N) = \lim_{N\to\infty} X_{gb}^{+}(N) = \lim_{N\to\infty} X(N) = \frac{1}{L_1},$$

and

$$\lim_{N\to\infty} \left[R_{gb}^{+}(N) - (NL_1 - Z)\right] = \lim_{N\to\infty} \left[R_{gb}^{-}(N) - (NL_1 - Z)\right] = 0.$$

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

3.2.1 Geometric Square-root Bounds (GSB)

We discuss how the GB throughput bounds could be extended for improved accuracy in queueing networks with delay stations. The extension of the response times bounds follows straightforwardly by (2.4). Our approach is similar to the Square Root bounds [9], and depends on the substitutions

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

in the denominators of (3.6) and (3.8). These remove the dependency on X_Z^{-} and X_Z^{+}, so that we can solve the resulting equation for $X(N)$. (We point to [6, 9] for proofs that (3.9) are bounds on $X(N-1)$ for all populations).

We then propose the following Geometric Square-root Bounds (GSB):

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

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

where $b_1 = R_0 + L_1(N-1) + \sum_{i=2}^{M} d_i Q_{i,gb}^{-}(N-1) \geq 0$, $d_i = L_i - L_1 \leq 0$.

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

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

where $b_2 = R_0 + L_1(N-1) + \sum_{i=2}^{M} d_i Q_{i,gb}^{+}(N-1) \geq 0$, $d_i = L_i - L_1 \leq 0$.

4 Numerical Examples and Computational Requirements

We compared the accuracy of GB and GSB with that of BJB and PB on several hundreds models with different loadings and number of queues. The conclusion that emerges from this statistical 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 important to consider network models that are known to be critical for bounding techniques.

We compare throughput bounds on well-known stress cases. The characteristics of the four considered models make them stress cases for bounding techniques, and hence have been used in previous works [8, 9] for comparison.

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 an additional delay station with $Z = 1$. The ABA saturation point is now $n^*(X^+_{aba}) = 13.7$.*

Stress Case 3 *The network is* 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 with $Z = 1$. 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 perfectly balanced, thus we do not obtain good approximations using the BJB. This lets 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 [10]. 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 typically independent of the number of bottleneck queues.

Tables 1-8 report the lower and upper bounds on X obtained with the BJB, PB, GB and GSB, for the four considered cases. Let us remark that the BJB are probably the most popular among the existing bounds, while the PB typically provide the best results among non-iterative techniques. The upper GB is computed using the utilization bound $U_i^+ = \min\{L_i X^+_{pb}, L_i/L_1\}$. The PB and BJB bounds in the stress cases 2 and 4 are computed using their iterative extensions defined in [8, 9] with $i = 2$ recursion steps. 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 than the upper BJB only for very small values of N. 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

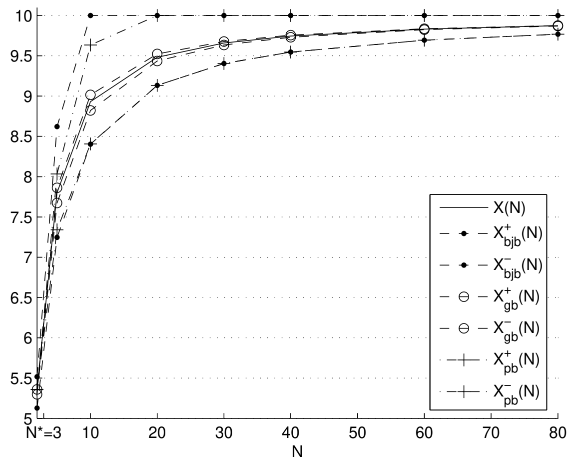

Figure 1. Comparison of the BJB and PB bounds with the GB bounds for Stress Case 3

absolute error of BJB and PB for medium and heavy load conditions.

These 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 Stress Case 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.

Finally, we point out that the computational cost of the throughput 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 performance measures that are frequently used by QoS control algorithms. We have shown that the proposed GB bounds, in spite of their simple formulation, which is 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.

	X(N)	Lower Bounds					
N	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 [8, 9]

	X(N)	Lower Bounds					
N	exact	bjb	pb	gsb	Δ_{bjb}	Δ_{pb}	Δ_{gsb}
2	1.434	1.432	1.434	1.432	–0.002	–0.000	–0.002
5	3.364	3.267	3.288	3.343	–0.097	–0.075	–0.020
10	5.872	5.390	5.440	5.757	–0.482	–0.432	–0.115
15	7.468	6.680	6.728	7.237	–0.788	–0.740	–0.230
20	8.375	7.489	7.525	8.086	–0.886	–0.850	–0.289
30	9.186	8.393	8.408	8.902	–0.793	–0.777	–0.284
40	9.502	8.858	8.864	9.264	–0.644	–0.638	–0.238
60	9.740	9.306	9.307	9.579	–0.433	–0.432	–0.161
80	9.828	9.514	9.514	9.715	–0.314	–0.314	–0.113

Table 2. Results for the lower bounds on throughput for Stress Case 2 of [8, 9]

	X(N)	Lower Bounds					
N	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 [8, 9]

	X(N)	Lower Bounds					
N	exact	bjb	pb	gsb	Δ_{bjb}	Δ_{pb}	Δ_{gsb}
2	1.528	1.524	1.528	1.527	–0.004	-0.000	–0.002
5	3.628	3.476	3.489	3.600	–0.152	–0.139	–0.028
10	6.422	5.684	5.685	6.233	–0.738	–0.737	–0.189
15	8.133	6.978	6.979	7.764	–1.155	–1.155	–0.369
20	8.945	7.767	7.767	8.558	–1.179	–1.179	–0.388
30	9.483	8.618	8.618	9.238	–0.864	–0.864	–0.245
40	9.659	9.042	9.042	9.508	–0.617	–0.617	–0.151
60	9.797	9.437	9.437	9.725	–0.360	–0.360	–0.072
80	9.856	9.614	9.614	9.814	–0.241	–0.241	–0.042

Table 4. Results for the lower bounds on throughput for Stress Case 4 of [8, 9]

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] 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.

[3] K.M. Chandy and D. Neuse. Linearizer: A heuristic algorithm for queuing network models of computing systems. *Comm. ACM*, 25(2):126–134, 1982.

[4] W.C. Cheng and R.R. Muntz. Bounding errors introduced by clustering of customers in closed product-form queuing networks. *J.ACM*, 43(4):641–669, 1996.

[5] P.J. Denning and J.P. Buzen. The operational analysis of queueing network models. *ACM Comp. Surv.*, 10(3):225–261, 1978.

[6] 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.

[7] D.L. Eager and K.C. Sevcik. Bound hierarchies for multiple-class queueing networks. *J.ACM*, 33(1):179–206, 1986.

[8] C.H. Hsieh and S. Lam. Two classes of performance bounds for closed queueing networks. *Perf. Eval.*, 7(1):3–30, 1987.

[9] J. Kriz. Throughput bounds for closed queueing networks. *Perf. Eval.*, 4(1):1–10, 1984.

[10] 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.

[11] R.R. Muntz and J.W. Wong. Asymptotic properties of closed queueing network models. In *Proc. of the*

	X(N)	Upper Bounds					
N	exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	4.317	4.324	4.317	4.317	0.007	0.000	0.000
5	6.715	6.757	6.730	6.752	0.042	0.015	0.038
10	8.206	8.316	8.270	8.269	0.110	0.064	0.063
15	8.835	9.009	8.953	8.905	0.174	0.118	0.071
20	9.168	9.401	9.339	9.243	0.233	0.171	0.075
30	9.499	9.828	9.759	9.582	0.329	0.260	0.083
40	9.654	10.000	9.984	9.743	0.346	0.330	0.089
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 [8, 9]

	X(N)	Upper Bounds					
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 [8, 9]

	X(N)	Upper Bounds					
N	exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	5.360	5.517	5.360	5.360	0.157	0.000	0.000
5	7.803	8.621	8.033	7.862	0.818	0.231	0.059
10	8.930	10.000	9.635	9.016	1.070	0.704	0.086
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 [8, 9]

	X(N)	Upper Bounds					
N	exact	bjb	pb	gb	Δ_{bjb}	Δ_{pb}	Δ_{gb}
2	1.528	1.531	1.528	1.636	0.003	0.000	0.108
5	3.628	3.690	3.664	3.885	0.062	0.036	0.257
10	6.422	6.960	6.858	6.708	0.538	0.436	0.286
15	8.133	9.494	9.245	8.190	1.361	1.112	0.057
20	8.945	10.000	9.814	8.871	1.055	0.868	0.074
30	9.483	10.000	10.000	9.414	0.517	0.517	0.069
40	9.659	10.000	10.000	9.624	0.341	0.341	0.035
60	9.797	10.000	10.000	9.794	0.203	0.203	0.003
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 [8, 9]

8th Annual Princeton Conf. on Inf. Science and Sys., pages 348–352, 1974.

[12] F.W.J. Olver. *Asymptotics and Special Functions*. Academic Press, 1974.

[13] M. Reiser and S.S. Lavenberg. Mean-value analysis of closed multichain queueing networks. *J.ACM*, 27(2):312–322, 1980.

[14] K.C. Sevcik and I. Mitrani. The distribution of queuing network states at input and output instants. *J.ACM*, 28(2):358–371, 1981.

[15] 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

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

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 relations. In order to obtain simple non-iterative approximations of (6.1), we seek for tight upper and lower bounds on $f(N)$. Let us denote a geometric sequence with common ratio $r \neq 1$ by

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

For $r = 1$, we set $Geom(1, n) = \sum_{k=1}^{n} 1 = n$.

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

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

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. Many bounds on utilization obtained by the BJB and the PB bounds can be rewritten in the form (6.5). Note that since the ratio of two consecutive $C(n)$ coefficients is a rational function of n, the coefficients (6.5) form a hypergeometric sequence [12]. Furthermore, let us note that for $b = 0$ the recurrence (6.1) becomes a simple geometric sequence that may be solved non-iteratively by (6.2). Hence, we focus in the rest of this appendix on recurrences where $b > 0$. In this case we see that $C^- = C(1)$ and $C^+ = C(N)$. Further, we can introduce an exact solution formula for (6.1). Let us denote $(x)_k = x(x+1)\cdots(x+k-1)$ and $x^{(k)} = x(x-1)\cdots(x-k+1) = (-1)^k(-x)_k$. Introducing the Gaussian hypergeometric function [12]

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

we have the following result:

Theorem 8 *The solution of (6.1), for all $N \geq 1$, is given by*

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

Currently, non-iterative expressions of (6.7) are available only for special values of the coefficients of ${}_2F_1$ [12]. Concerning the accuracy of the bounds in Theorem 7, a numerical inspection quickly leads us to conclude that *Geom*$(C(N), N)$ is a very tight upper bound, while the lower bound *Geom*$(C(1), N)$ is quite loose. Therefore, we introduce an improved lower bound by taking 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, for all $1 \leq N < +\infty$, by*

$$ f(N) \geq \textit{Geom}\left(\frac{N}{N+1}C(N+1), N\right) \qquad (6.8)$$

Note that $N\,C(N+1)/(N+1)$ may be bigger or smaller of the terms $C(n)$, $1 \leq n \leq N$, and this means that it is not an upper bound on $C(n)$ as C^+. This result is fundamental for the derivation of the pessimistic GB queue-length bound.

Appendix B

Proof of Theorem 1.

Consider the relation $Q_i^-(N) = X_{bjb}^-(N)L_i[1+Q_i^-(N-1)]$, with $Q_i^-(0) = 0$, and 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).

Proof of Theorem 2.

It is sufficient to apply Theorem 7 of Appendix A and (6.2) to the recurrence $Q_i^+(N) = U_i^+(N)[1 + Q_i^+(N-1)]$, with initial condition $Q_i^+(0) = 0$, that defines an upper bound on $Q_i(N)$.

Proof of Theorem 3.

From the Arrival Theorem [13, 14], we know that $R(N) = L + \sum_i L_iQ_i(N-1)$. Expressing (2.5) for $N-1$, and solving for $Q_1(N-1)$ we get $Q_1(N-1) = N-1-ZX(N-1) - \sum_{i\neq 1} Q_i(N-1)$. Inserting this relation into $R(N)$, it is

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

$$ \geq L + L_1(N-1-ZX_Z^-(N-1)) + \sum_{i=2}^{M} d_iQ_i^-(N-1),$$

where $d_i = L_i - L_1 \leq 0$. This proves the theorem.

Proofs of Corollaries 1 and 2.

The two corollaries are proved immediately by (2.4).

Proof of Theorem 4.

The proof is analogous to that of Theorem 3.

Proof of Theorems 5-6.

The proofs are analogous to the original Square-root bounds proofs [9] applied to equations (3.6) and (3.8).

Proof of Theorem 7.

The theorem follows easily from (6.1) and (6.2) observing that

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

Proof of Theorem 8.

From (6.1) and (6.5) we have

$$ f(N) = \sum_{k=1}^{N} \frac{a^k N\cdots(N-k+1)}{(b+N)\cdots(b+N-k+1)} = \sum_{k=1}^{N}\frac{a^k(-N)_k}{(-b-N)_k}$$

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

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

where ${}_2F_1^N$ denotes the N-th partial sum of ${}_2F_1$. The theorem follows noting that $(-N)_k = 0$ for $k \in \{N+1,\ldots,+\infty\}$, 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.

The proof is obtained by induction on N.

Case $N = 1$: we need to prove that $f(1) \geq \textit{Geom}(C(2)/2, 1)$ that simplifies to $C(1) \geq C(2)/2$. Inserting (6.5), this simplifies to $a/(b+1) \geq a/(b+2)$, which is always true by the assumptions $a > 0$ and $b \geq 0$ in (6.5).

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

$$f(N-1) \geq \textit{Geom}(H(N), N-1).$$

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

$$f(N) = C(N)\,[1 + f(N-1)] \geq C(N)\frac{1 - H(N)^N}{1 - H(N)}. \quad (6.9)$$

We wish to prove that

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

Note that this follows by (6.9) if we can show that

$$C(N)\frac{1 - H(N)^N}{1 - H(N)} \geq H(N+1)\frac{1 - H(N+1)^N}{1 - H(N+1)},$$

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