A Survey of Bottleneck Analysis in Closed Networks of Queues *

P. J. Schweitzer
W. E. Simon Graduate School of Business Administration,
University of Rochester, Rochester, NY 14627, USA

G. Serazzi, M. Broglia
Politecnico di Milano, Dip. Elettronica e Informazione,
P.za L. da Vinci 32, 20133 Milano, Italy
serazzi@ipmel2.elet.polimi.it

Abstract

Several of the principal results in bottleneck analysis for closed queueing networks are surveyed. Both product-form closed queueing networks, where exact bottleneck analysis is possible, and non-product-form closed queueing networks, where approximations are given for asymptotic bottleneck behavior, are considered. Algorithms for the asymptotic bottleneck analysis are presented and the switching surfaces of bottlenecks are described.

1 Introduction and Scope

1.1 Importance

Identification of bottlenecks (BNs) in queueing networks is an important step in systems performance evaluation and upgrades: investing resources at a BN will have dramatic effect on systems performance; investing at a non-BN will have negligible benefit. However, over-investing at a BN is unwise because, beyond a certain point, the secondary (or tertiary) BN becomes the primary BN. Beyond this point, the investment should be split among the several nearties for BN. This complication, along with the diminishing marginal benefits associated with investment, make BN analysis — especially of multiple BNs — somewhat complex.

*This work was partially supported by CNR "Progetto Finalizzato Sistemi Informatici e Calcolo Parallelo" by grant N. 92.01615.PF69.115.23757 and by M.U.R.S.T. 40% Project

In addition, BN analysis unavoidably leads to deep technical difficulties, because the creation of BNs is inherently a *non-linear* phenomenon: *small* changes in relative loads or capacities can lead to *large* shifts in BN locations. On the other hand, BN analysis can be simpler than a full performance analysis, because much less is being demanded, at the minimum merely requesting the location of the most-congested system resources. Furthermore, additional simplification is possible if one performs only *asymptotic* BN analysis, where the load on the system (e.g., customer population) approaches 100% saturation.

1.2 Purpose and Scope

The purpose of this paper is to survey several of the principal results in BN analysis for *closed* queueing networks (CQNs) (open queueing networks are much less challenging to analyze because knowledge of external arrival rates and visit ratios permits immediate prediction of resource utilizations).

We include both product-form closed queueing networks (PF-CQNs), where exact BN analysis are surveyed and non-product-form networks, where approximations are given for asymptotic BN behavior. However, there are several assumptions made to limit the scope, and thereby make the survey manageable in size:

- all servers are constant-rate, and either FCFS (no parallel servers) or ample server (AS);
- queue space is unlimited (no blocking) system;
- a customer is at only one resource at any given time, and makes *instantaneous* transfers from one resource to another (e.g., ignore bus transmission times);
- all customer classes are *closed*, and customers do not change class;
- only an equilibrium steady-state analysis is presented. This assumes, among other things, that loads (e.g., populations) remain constant over time, so that one does not have to forecast time-varying BNs;
- we do not distinguish between user-workload and system overhead.

1.3 Definition of Bottlenecks

At least two concepts of BNs are in common use:

- *physical bottleneck*:
  - resource with highest utilization;
  - resource with highest mean sojourn time;

- resource with highest mean queueing (delay) time;
- resource with highest mean queue length (queue length for us means number present either in service or in queue; this definition makes meaningful mean queue length at an AS (= mean number of customers present));
- device with highest 90% percentile (or other percentile) of sojourn time (or queue length);

- *economic bottleneck*:
  - resource with largest value of the derivative: rate of improvement in the system performance measure per dollar invested at the resource (the systems performance measure could be any scalar such as weighted average throughput, weighted average response time, etc.).

It is evident that many possible definitions of BNs exist, and that they are *not equivalent*: the device with highest utilization need not have highest response time, nor be the economic BN. However, in the case where one — and only one — device is running at a high level of congestion, all definitions will agree.

The device with highest value of the performance measure (e.g., highest utilization) is called the primary *bottleneck* (if more than one, we speak of the *BN set*). Next highest is called the secondary BN, etc.

It is noteworthy that physical BNs differ from economic BNs in two ways

- physical BN uses a *physical* measure of performance and does not involve money;
- physical BN looks at *average* level of performance while economic BN looks at *marginal return on performance.*

In this survey, we use *device with highest utilization* as our definition of BN. This is the most common approach, is the easiest to measure, and in any case acts as a reasonable surrogate for the congestion at the device.

1.4 Notation

To keep the presentation simple, in the sequel of the paper the index r will always be implicitly assumed to range from 1 to R and the indexes i, j to range from 1 to M.

1.4.1 Model Inputs

- M servers, labelled $\{1, 2, \ldots, M\} = SFCFS + SAS$, where $SFCFS$ = set of FCFS servers and SAS = set of AS;

- customer classes labelled $\{1, 2, \ldots, R\}$;
- K_r = (fixed) population of class r;
- S_{ri} = mean service time of a class-r customer for each visit to server i (for product-form networks, S_{ri} is independent of r if i is FCFS);
- V_{ri} = mean number of visits a class-r customer makes to server i;
- $L_{ri} = V_{ri}S_{ri}$ = mean load a class-r customer makes to server i in all its visits.

We assume

- $\sum_r L_{ri} > 0 \qquad 1 \le i \le M \qquad$ (otherwise delete server i);
- $\sum_i L_{ri} > 0 \qquad 1 \le r \le R \qquad$ (otherwise delete class r);
- $K_r \ge 1 \qquad 1 \le r \le R \qquad$ (otherwise delete class r);
- $R \ge 1$;
- $S_{ri}, V_{ri}, L_{ri} \ge 0 \qquad 1 \le r \le R \quad 1 \le i \le M$;
- $K_{\text{sum}} \equiv \sum_r K_r$ = total customer population;
- $\beta_r = K_r/K_{\text{sum}}$ = fraction of population being of class $r \qquad 1 \le r \le R$;
- $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_R)$ = mix vector;
- $\underline{K} = (K_1, K_2, \ldots, K_R) = \underline{\beta}K_{\text{sum}}$ = population vector.

1.4.2 Model Outputs

- W_{ri} = mean sojourn time;
- W_r = response time of class r;
- X_r = throughput of class r;
- X_{ri} = throughput of class-r customers at server i;
- Q_{ri} = mean queue length;
- Q_i = mean queue at server i;
- U_{ri} = utilization at server i due to class r;
- U_i = utilization of server i.

Only $\{W_{ri}\}$ are independent, the rest being obtained from

$$W_r = \sum_i V_{ri} W_{ri}$$

$$X_r = \frac{K_r}{W_r}, \qquad X_{ri} = X_r V_{ri} \quad \text{(forced-flow law)}$$

$$Q_{ri} = X_{ri} W_{ri} \quad \text{(Little's law)}, \qquad Q_i = \sum_r Q_{ri}$$

$$U_{ri} = X_{ri} S_{ri}, \qquad U_i = \sum_r U_{ri}$$

There are all nonnegative and also satisfy $W_{ri} \geq S_{ri}$ (so $W_r \geq \sum_i L_{ri} > 0$) and $\sum_i Q_{ri} = K_r$.

2 Asymptotic Bottleneck Analysis

Here $K_{\text{sum}} \to \infty$, i.e., at least one customer class becomes very large. We assume that each class r with $K_r \to \infty$ also satisfies

$$\sum_{i \in SFCFS} L_{ri} > 0$$

i.e., class r visits at least one FCFS server. This means that at least one FCFS server will be saturated.

For product-form networks, where exact solutions are available (by looking, for example, at the integral representation of the generating function), one is interested in the asymptotic behavior of U_i, X_r, W_{ri}, Q_{ri} and Q_i as $K_{\text{sum}} \to \infty$.

These are typically *asymptotic expansions* of the form

$$U_i = U_i^* + \frac{U_i^{**}}{K_{\text{sum}}} + \mathrm{O}\left(\frac{1}{K_{\text{sum}}^2}\right) \qquad (U_i^* \leq 1)$$

$$X_r = X_r^* + \frac{X_r^{**}}{K_{\text{sum}}} + \mathrm{O}\left(\frac{1}{K_{\text{sum}}^2}\right)$$

$$W_{ri} = K_{\text{sum}} W_{ri}^* + W_{ri}^{**} + \mathrm{O}\left(\frac{1}{K_{\text{sum}}}\right)$$

$$Q_{ri} = K_{\text{sum}} Q_{ri}^* + Q_{ri}^{**} + \mathrm{O}\left(\frac{1}{K_{\text{sum}}}\right)$$

$$Q_i = K_{\text{sum}} Q_i^* + Q_i^{**} + \mathrm{O}\left(\frac{1}{K_{\text{sum}}}\right) \qquad \left(Q_i^* = \sum_r Q_{ri}^*\right)$$

Our notation uses $*$ for the leading term, $**$ for the next term, etc. Note U_{ri} and X_{ri} approach finite limits (U_{ri}^* and X_{ri}^*) while Q_{ri} and W_{ri} diverge *linearly* with K_{sum} for at least one $i \in SFCFS$. Also note $0 \leq U_i^* \leq 1$, $Q_i^* = 0$ if i is an AS, $\sum_{i \in SFCFS} Q_i^* = 1$.

The *asymptotic bottleneck set* $BN(\underline{\beta})$ is defined as the set of servers whose utilization approaches 100%:

$$BN(\underline{\beta}) \equiv \{i : i \in SFCFS \text{ and } U_i^* = 1\}$$

Note the set $\{i : i \in SFCFS \text{ and } Q_i^* > 0\}$ of servers whose queue length approaches infinity is a *subset* of $BN(\underline{\beta})$.

The primary goals of asymptotic BN analysis are to find

$$BN(\underline{\beta}) = \text{asymptotic bottleneck set}$$

and

$$\gamma_i(\underline{\beta}) = \frac{Q_i^*}{\sum_{j \in SFCFS} Q_j^*} = Q_i^* = \text{asymptotic fraction of population at server } i$$

Note

$$\gamma_i(\underline{\beta}) = 0 \quad i \in SAS, \qquad \gamma_i(\underline{\beta}) \geq 0, \qquad \sum_{i \in SFCFS} \gamma_i(\underline{\beta}) = 1$$

The remaining performance measures are then given by

$$W_{ri}^* = S_i \gamma_i \quad \text{(independent of } r\text{)}, \qquad W_r^* = \sum_i L_{ri} \gamma_i \tag{1a}$$

$$X_r^* = \frac{\beta_r}{\sum_i L_{ri} \gamma_i}, \qquad X_{ri}^* = \frac{\beta_r V_{ri}}{\sum_j L_{rj} \gamma_j} \tag{1b}$$

$$Q_{ri}^* = \frac{\beta_r L_{ri} \gamma_i}{\sum_j L_{rj} \gamma_j}, \qquad Q_i^* = \gamma_i \sum_r \frac{\beta_r L_{ri}}{\sum_j L_{rj} \gamma_j} \tag{1c}$$

$$U_{ri}^* = \frac{\beta_r L_{ri}}{\sum_j L_{rj} \gamma_j}, \qquad U_i^* = \sum_r \frac{\beta_r L_{ri}}{\sum_j L_{rj} \gamma_j} \tag{1d}$$

At least 4 different cases of asymptotic analysis must be distinguished

1. $K_{\text{sum}} \to \infty$, M fixed, at least one AS exists [RM82], [KB91];

2. $K_{\text{sum}} \to \infty$, $M \to \infty$, K_{sum}/M fixed [KT90];

3. $K_{\text{sum}} \to \infty$, $M \to \infty$, K_{sum}/M fixed (or has a limit), service in random order [BGPS87];

4. $K_{\text{sum}} \to \infty$, M fixed, no AS.

Case 1. is the best known due to the PANACEA code. A scheme is given to find as many terms as desired in the asymptotic expansion, along with a bound on the truncation error. However, due to the assumption of "normal usage", where all FCFS servers have utilizations bounded away from unity, the model is restricted to the case where an infinite number of customers accumulates at, and only at, the ample servers. These always act as "bottlenecks".

Case 2. models a computer network with a very large number of terminals. The asymptotic analysis is quite delicate.

Case 3. shows asymptotic normality of the joint queue lengths, and provides algorithms for the means and covariances.

Case 4. differs from Case 1 because customers demands must accumulate at, and only at, FCFS servers [BS93a], [BS93b], [SSB92] etc. It differs from Case 2 and 3 because M is fixed. Its extensions to ample servers is straightforward. This case is the least known, and therefore merits presentations of some of the technical details. It also leads to convenient approximations for the non-product-form case.

3 Algorithms for Asymptotic Bottleneck Analysis

3.1 Introduction

This section shows how to carry out the asymptotic BN analysis for CQN as $K_{\text{sum}} \to \infty$. In particular, we describe how to compute the asymptotic fractions

$$\gamma_i(\underline{\beta}) \equiv \lim_{K_{\text{sum}} \to \infty} \frac{Q_i(K_{\text{sum}}\underline{\beta})}{K_{\text{sum}}}$$

and the asymptotic bottleneck set $BN(\underline{\beta})$. Note that $\sum_i \gamma_i = 1$, $\gamma_i > 0$ implies $U_i^* = 1$.

Values of $\underline{\beta}$, $i \in BN(\underline{\beta})$ and $U_i^* = 1$ implies the converse, namely that $\gamma_i > 0$.

Note that the technical difficulty lies in computing $\gamma_i(\underline{\beta})$, since $BN(\underline{\beta})$ can then be easily computed from

$$BN(\underline{\beta}) = \{i \in SFCFS : U_i^* = 1\} = \left\{ i \in SFCFS : \sum_r \frac{\beta_r L_{ri}}{\sum_j L_{rj}\gamma_j(\underline{\beta})} = 1 \right\} \tag{2}$$

The algorithm include

1. fixed point methodology;
2. optimization methodology;
3. simultaneous non-linear equation methodology;
4. simultaneous linear equation methodology.

The remaining subsections discuss the higher order terms and the extensions to non-product form CQNs.

3.2 Fixed Point Methodology

This approach exploits the properties

$$\gamma_i(\underline{\beta}) \geq 0 \tag{3a}$$

$$U_i^*(\underline{\beta}) = \sum_r \frac{\beta_r L_{ri}}{\sum_j L_{rj}\gamma_j(\underline{\beta})} \leq 1 \tag{3b}$$

$$\gamma_i(\underline{\beta}) > 0 \implies U_i^*(\underline{\beta}) = 1 \tag{3c}$$

If we define the function

$$f_i(\underline{z}) \equiv \sum_r \frac{\beta_r L_{ri}}{\sum_j L_{rj} z_j} \qquad 1 \leq i \leq M, \quad \underline{z} \in [0,1]^M \tag{4}$$

where $\underline{z} = (z_1, z_2, \ldots, z_M)$ is constrained to satisfy $z_i \geq 0$ for all i and $\sum_j L_{rj} z_j > 0$ for all r, then (4) may be rewritten as the *non-linear complementarity problem*

$$\gamma_i \geq 0 \qquad 1 \leq i \leq M \tag{5a}$$

$$1 - f_i(\underline{\gamma}) \geq 0 \qquad 1 \leq i \leq M \tag{5b}$$

$$\gamma_i[1 - f_i(\underline{\gamma})] = 0 \qquad 1 \leq i \leq M \tag{5c}$$

(i.e., $\underline{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_M)$ and $U_i^* = f_i(\underline{\gamma}) \quad 1 \leq i \leq M$).

Summation of the last of these shows that, as expected,

$$\sum_i \gamma_i = 1$$

In addition, the last may be understood as the *fixed point equation* for the γ's

$$\gamma_i = \gamma_i f_i(\underline{\gamma}) \qquad 1 \leq i \leq M \tag{6a}$$

subject to the side constraints

$$\gamma_i \geq 0 \qquad 1 \leq i \leq M \tag{6b}$$

$$\sum_i L_{ri}\gamma_i > 0 \qquad 1 \leq r \leq R \tag{6c}$$

$$f_i(\underline{\gamma}) \leq 1 \qquad 1 \leq i \leq M \tag{6d}$$

An alternate derivation of (5c) and (6a) is based directly upon the MVA recursion

$$Q_i(\underline{K}) = \sum_r \frac{K_r L_{ri}[1 + Q_i(\underline{K} - \underline{e}^r)]}{\sum_j L_{rj}[1 + Q_j(\underline{K} - \underline{e}^r)]} \qquad 1 \leq i \leq M \tag{7}$$

namely, divide both sides by K_{sum} and let $K_{\text{sum}} \to \infty$.

The fixed point equation (6a) has been obtained by [Schw79], [Chow83]. In most cases it has a *unique* solution $\underline{\gamma}$. However there are some exceptional cases where (6a) does *not* have a unique solution, and one must go to higher order in the expansion

$$\frac{Q_i(K_{\text{sum}}\underline{\beta})}{K_{\text{sum}}} = \gamma_i(\underline{\beta}) + \mathrm{O}\left(\frac{1}{K_{\text{sum}}}\right)$$

in order to resolve ambiguities. Even in the case where the fixed point equation does not satisfy $\underline{\gamma}$ uniquely, it is possible to show that $f_i(\gamma)$ is unique for each i, hence $BN(\underline{\beta})$ is well-defined.

A convenient way to solve (6) is by *successive substitutions*

$$\gamma_i^{(n+1)} = \gamma_i^{(n)} f_i(\underline{\gamma}^{(n)}) \qquad 1 \leq i \leq M \tag{8}$$

starting from an initial guess which satisfies (6 c,d) but with (6b) replaced by $\gamma_i^{(0)} > 0$ (else $\gamma_i^{(n)}$ remain zero for all n). The scheme (8) is easy to program and usually has geometric convergence to $\underline{\gamma}$. However convergence is not guaranteed and sometimes the scheme fails. Ways of ensuring convergence are discussed in subsection 3.3 below, hence computation of $\gamma_i(\underline{\beta})$ and $BN(\underline{\beta})$ (and higher order terms) may be considered to be *routine*, except at (or near) the exceptional values of $\underline{\beta}$ where the fixed point is not unique. This algorithm appears to be *simpler* than other asymptotic expansions, such as those based upon integral representations.

3.3 Optimization Methodology

The scheme (8) may be shown [Schw79] to be a projected gradient approach for solving the concave optimization problem

$$\max\left\{h(\underline{z}) : \underline{z} \in [0,1]^M, \quad z_i \geq 0 \text{ and } \sum_i L_{ri} z_i > 0\right\} \tag{9}$$

where

$$h(\underline{z}) \equiv \sum_r \beta_r \log\left[\sum_i L_{ri} z_i\right] - \sum_i \gamma_i$$

and indeed, the complementary slackness conditions (6) are the Kuhn-Tucker conditions for (9), so (8) and (9) are equivalent characterizations of $\underline{\gamma}$.

The step-direction $\underline{\gamma}^{(n+1)} - \underline{\gamma}^{(n)}$ in successive substitutions is an *uphill* direction for maximizing h, i.e.,

$$\sum_i [\underline{\gamma}^{(n+1)} - \underline{\gamma}^{(n)}]_i \frac{\partial}{\partial z_i} h(\underline{\gamma}^{(n)}) = \sum_i \gamma_i^{(n)} [f_i(\underline{\gamma}^{(n)}) - 1]^2 > 0$$

However the step length could be too long, causing overshooting and lack of convergence of successive substitutions. To enforce convergence, one must merely check that $h(\underline{\gamma}^{(n+1)})$ is *strictly longer* than $h(\underline{\gamma}^{(n)})$, and if this is not so, one reduces the step length sufficiently that this criterion is met.

We found it simplest to repeatedly halve the step length (i.e., $\underline{\gamma}^{(n+1)} = \frac{1}{2}(\underline{\gamma}^{(n)} + \underline{\gamma}^{(n+1)})$) until $h(\underline{\gamma}^{(n+1)}) > h(\underline{\gamma}^{(n)})$ is met. This procedure works flawlessly, producing (at least) 6 digit accuracy without difficulty.

For other reduction of the leading asymptotic term to an optimization problem, see [Pitt79], [BGPS87], [PKT90].

3.4 Simultaneous Non-Linear Equation Methodology

Here one applies any root-finding technique for the M simultaneous equations (6a) and then checking for satisfaction of (6 b,c,d). For non-exceptional $\underline{\beta}$, where $BN(\underline{\beta}) \equiv \{i : U_i^*(\underline{\beta}) = 1\} = \{i : \gamma_i(\underline{\beta}) > 0\}$, this is especially easy in the usual case where there is only a *small* set of bottlenecks $BN(\underline{\beta})$, because (6a) reduces to only $|BN(\underline{\beta})|$ simultaneous (non-linear) equations

$$1 = \sum_r \frac{\beta_r L_{ri}}{\sum_{j \in BN(\underline{\beta})} L_{rj} \gamma_j(\underline{\beta})} \qquad i \in BN(\underline{\beta}) \tag{10}$$

for $|BN(\underline{\beta})|$ unknowns $\{\gamma_i(\underline{\beta}) : i \in BN(\underline{\beta})\}$.

If $BN(\underline{\beta})$ can be guessed correctly, then (say) Newton's method applied to (10) gives $\{\gamma_i(\underline{\beta}) : i \in BN(\underline{\beta})\}$ while $\gamma_i(\underline{\beta}) = 0$ if $i \notin BN(\underline{\beta})$. One then checks if (6 b,c,d) hold in order to confirm the guess for $BN(\underline{\beta})$.

3.5 Simultaneous Linear Equation Methodology

This approach uses

$$X_r^* = \frac{\beta_r}{\sum_{i \in BN(\underline{\beta})} L_{ri} \gamma_i(\underline{\beta})} > 0 \qquad 1 \le r \le R \tag{11}$$

as primary unknowns, thereby transforming (10) into a set of *linear* equations

$$1 = \sum_r X_r^* L_{ri} \qquad i \in BN(\underline{\beta}) \qquad (= U_i^*) \tag{12}$$

for the X^*'s (if the X^*'s are under-determined, one adds additional equations involving the reciprocals of the X^*'s, to reflect the fact that the rows of $L(\underline{\beta}) \equiv [L_{ri}]$ $1 \le r \le R$, $i \in BN(\underline{\beta})$ may be linearly dependent). Once the X^*'s are known, the γ's can be chosen anywhere in the polytope

$$\left\{ \underline{\gamma} \in [0,1]^M \;:\; \gamma_i = 0 \text{ for } i \notin BN(\underline{\beta}), \quad \gamma_i \ge 0 \text{ for } i \in BN(\underline{\beta}), \right.$$
$$\left. \sum_i \gamma_i = 1, \quad \sum_i L_{ri}\gamma_i = \beta_r / X_r^* \text{ for all } r \right\} \tag{13}$$

preferably the *maximal* solution $\gamma_i > 0$ (strict) for all $i \in BN(\underline{\beta})$.

As the simplest illustration this, consider the most common case where there is just one BN, say

$$BN(\underline{\beta}) = \{b\}$$

Then

$$\gamma_i = \delta_{ib} \tag{14}$$

and (10) reduces to the identity

$$1 = \sum_r \frac{\beta_r L_{rb}}{L_{rb}}$$

Evidently

$$X_r^* = \frac{\beta_r}{L_{rb}} \qquad 1 \le r \le R$$

is easily determined from (11) and (14). Note $L(\underline{\beta})$ is a $R \times 1$ matrix, hence only one row of it is independent. Also, one must check that $U_i^* = \sum_r \frac{\beta_r L_{ri}}{L_{rb}} < 1 = U_b^*$ for $i \ne b$.

The next complicated case is of two BNs,

$$BN(\underline{\beta}) = \{b_1, b_2\}$$

This is most tractable if there are 2 types of customers, since (12) then consists of 2 simultaneous linear equations for the two y's. Finally, γ_{b_1} and γ_{b_2} are obtained from the latter two linear equations from the trio in (13)

$$\begin{cases} \gamma_{b_1} + \gamma_{b_2} = 1 \\ L_{rb_1}\gamma_{b_1} + L_{rb_2}\gamma_{b_2} = \beta_r / X_r^* \qquad r = 1, 2 \end{cases}$$

Assuming that the 2×2 matrix

$$L(\underline{\beta}) = \begin{bmatrix} L_{1b_1} & L_{1b_2} \\ L_{2b_1} & L_{2b_2} \end{bmatrix}$$

is non-singular, the result is

$$\begin{cases} \gamma_{b_1} = \dfrac{\beta_1 L_{2b_2}}{L_{2b_2} - L_{2b_1}} - \dfrac{\beta_2 L_{1b_2}}{L_{1b_1} - L_{1b_2}} \\ \gamma_{b_2} = \dfrac{\beta_2 L_{1b_1}}{L_{1b_1} - L_{1b_2}} - \dfrac{\beta_1 L_{2b_1}}{L_{2b_2} - L_{2b_1}} \end{cases}$$

More generally, this approach works if the number of bottlenecks $|BN(\underline{\beta})|$ agrees with the number of classes R, provided the $R \times R$ matrix

$$L(\underline{\beta}) \equiv \{L_{ri} : 1 \leq r \leq R, \quad i \in BN(\underline{\beta})\}$$

is non-singular.

3.6 Higher Order Terms

By inserting the asymptotic expansion

$$Q_i(K_{\text{sum}}\underline{\beta}) = K_{\text{sum}} Q_i^*(\underline{\beta}) + Q_i^{**}(\underline{\beta}) + \mathrm{O}\left(\frac{1}{K_{\text{sum}}}\right) \tag{15}$$

into (7), one gets a *sequence* of fixed point problems for Q^*, Q^{**}, etc. We have already investigated the first of these, for Q^*. As long as $\underline{\beta}$ is not exceptional, the higher order terms can be evaluated recursively. The authors have obtained explicit expressions for Q^{**} and Q^{***}. We found that the first three terms in the series (15) are sufficient provided

- K_{sum} is sufficiently large (typically 1000's); the relatively slow convergence as $K_{\text{sum}} \to \infty$ was also noted in [Lave82] if ample servers occur;
- $\underline{\beta}$ is not too close to an exceptional value, where the asymptotic series becomes singular. The bad cases can usually be detected because of symptoms like $|Q_i^{**}|$ is very large for some $i \in BN(\underline{\beta})$, $\gamma_i(\underline{\beta})$ is very close to zero for some $i \in BN(\underline{\beta})$, some $\left|\dfrac{\partial}{\partial \beta_r}\gamma_i(\underline{\beta})\right|$ gets very large, etc. We call the set of β's where the expansions break down *switching surfaces* (i.e., singularities), because the set $BN(\underline{\beta})$ of BNs is discontinuous there.

3.7 Examples

Consider the case $M = 4$, $R = 2$ with

$$L_{ri} = \begin{bmatrix} 100 & 90 & 50 & 40 \\ 50 & 70 & 90 & 40 \end{bmatrix}$$

Server 4 can never be a bottleneck since it is masked-off by the other servers. Then, from (2), we have

$$BN(\underline{\beta}) = \begin{cases} \{3\} & 0 \le \beta_1 < \frac{5}{23} \\ \{2,3\} & \frac{5}{23} \le \beta_1 \le \frac{9}{23} \\ \{2\} & \frac{9}{23} < \beta_1 < \frac{18}{25} \\ \{1,2\} & \frac{18}{25} \le \beta_1 \le \frac{4}{5} \\ \{1\} & \frac{4}{5} < \beta_1 \le 1 \end{cases}$$

and $\beta_2 = 1 - \beta_1$.

The exceptional points (switching surfaces) are $\underline{\beta} = \left(\frac{5}{23}, \frac{18}{23}\right) \doteq (0.217, 0.783)$, $\underline{\beta} = \left(\frac{9}{23}, \frac{14}{23}\right) \doteq (0.391, 0.609)$, $\underline{\beta} = \left(\frac{18}{25}, \frac{7}{25}\right) = (0.72, 0.28)$ and $\underline{\beta} = \left(\frac{4}{5}, \frac{1}{5}\right) = (0.8, 0.2)$ where $BN(\underline{\beta})$ changes. The corresponding first terms in the asymptotic expansion are

$$\underline{\gamma}(\underline{\beta}) = \underline{Q}^*(\underline{\beta}) = \begin{cases} (0,0,1,0) & 0 \le \beta_1 < \frac{5}{23} \\ \left(0, -\frac{5}{4} + \frac{23}{4}\beta_1, \frac{9}{4} - \frac{23}{4}\beta_1, 0\right) & \frac{5}{23} \le \beta_1 \le \frac{9}{23} \\ (0,1,0,0) & \frac{9}{23} < \beta_1 < \frac{18}{25} \\ \left(-9 + \frac{25}{2}\beta_1, 10 - \frac{25}{2}\beta_1, 0, 0\right) & \frac{18}{25} \le \beta_1 \le \frac{4}{5} \\ (1,0,0,0) & \frac{4}{5} < \beta_1 \le 1 \end{cases}$$

The switching surfaces can be detected from the expressions for $Q_i^*(\underline{\beta})$, from (1c), that violate the requirements $0 \le Q_i^* \le 1$. In addition, at the switching surface we have the unusual situation where some server i has $U_i^* = 1$ but $\gamma_i = 0$.

As can be seen from Figure 1, the migration of the bottleneck from one server to another yields a bottleneck set in which both these servers saturate.

With $R = 3$ and K_{sum} constant all the possible mix vectors $\underline{\beta}$ belong to the triangle $\{\underline{\beta} : \beta_1 + \beta_2 + \beta_3 = 1 \text{ and } 0 \le \beta_1, \beta_2, \beta_3 \le 1\}$. The switching surfaces become straight lines identifying polyhedral regions in which one, two or (at most) three servers saturate. Each value of $BN(\underline{\beta})$ occurs only once, i.e., the regions are all connected sets. Bottleneck set switches from regions with one to regions with two and with three components.

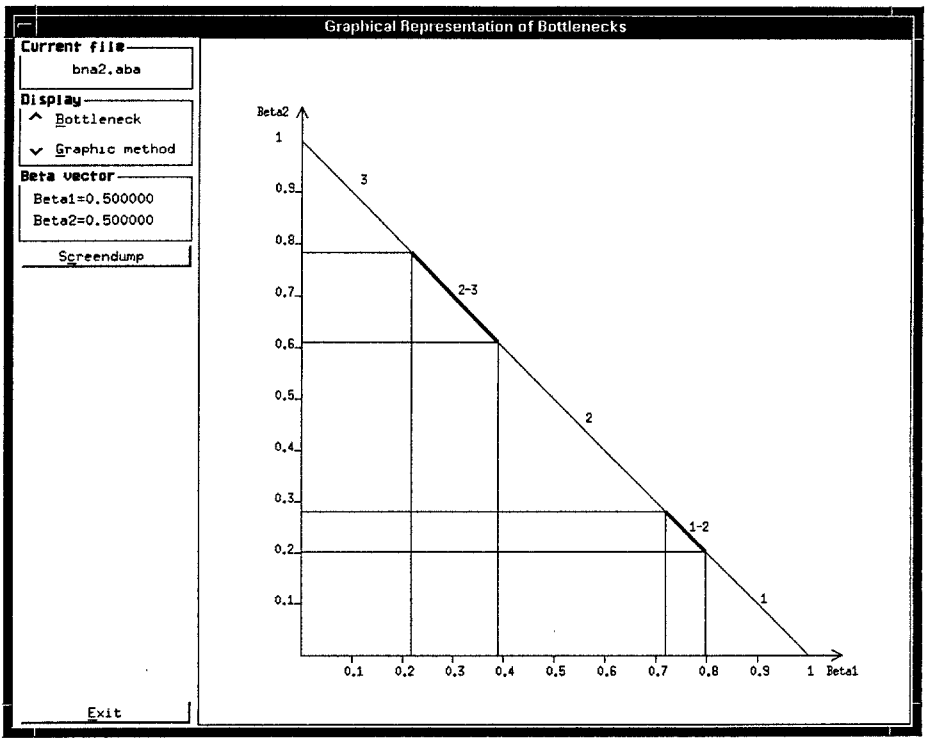

Figure 1: Bottleneck set $BN(\underline{\beta})$ vs. population mix $\underline{\beta}$ for $R = 2$.

Let us consider now the case $M = 5$, $R = 3$ with

$$L_{ri} = \begin{bmatrix} 90 & 50 & 60 & 80 & 70 \\ 40 & 80 & 30 & 40 & 30 \\ 50 & 80 & 90 & 50 & 70 \end{bmatrix}$$

Figure 2 represents the various bottleneck sets of this case. As can be seen, with the loading matrix considered we have only one bottleneck set in which three servers saturate together, i.e., the internal triangle. However, depending on the relative values of L_{ri} and on the number of servers none or several of such regions may exist. The regions with all the other bottleneck sets are also represented. A complete description of the switching surfaces can be found in [BS93b].

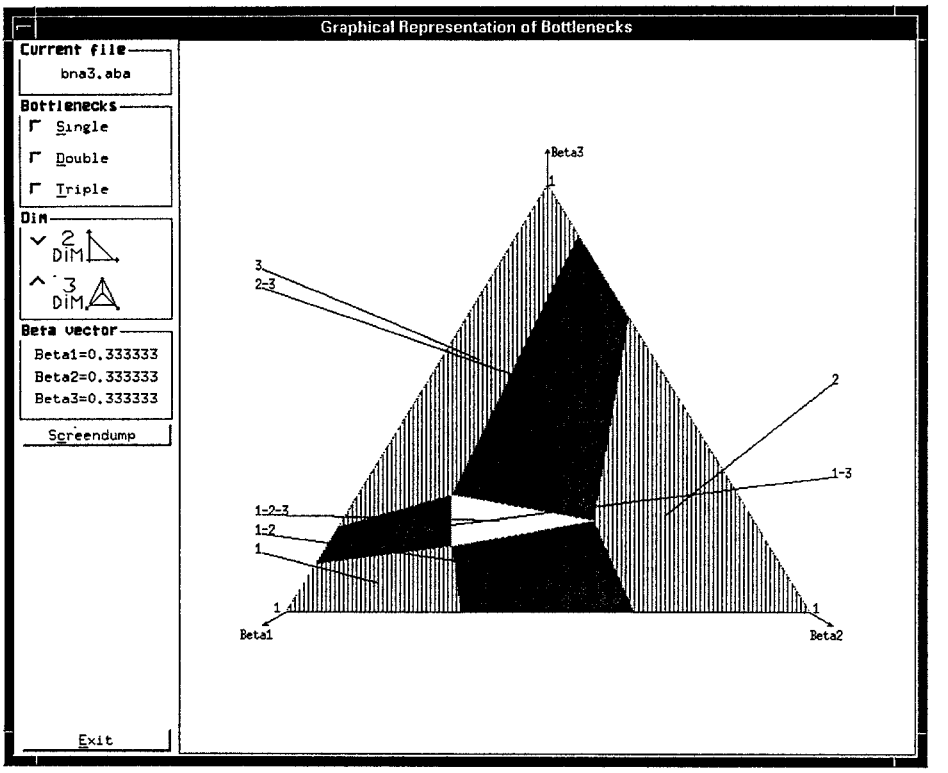

Figure 2: Bottleneck set $BN(\underline{\beta})$ vs. population mix $\underline{\beta}$ for $R = 3$.

3.8 Non-Product Form Networks

If S_{ri} depends upon r at one or more FCFS server i, the CQN lacks product-form. If we assume

$$W_{ri} \sum_t Q_{ti} S_{ti} + \mathrm{O}(1) \text{ for large } K_{\mathrm{sum}} \tag{16}$$

$$\frac{Q_{ri}}{K_{\mathrm{sum}}} \to Q^*_{ri} \text{ as } K_{\mathrm{sum}} \to \infty$$

then

$$\frac{Q_i}{K_{\mathrm{sum}}} \to Q^*_i = \sum_r Q^*_{ri}$$

$$\frac{W_{ri}}{K_{\text{sum}}} \rightarrow W_i^* = \sum_t Q_{ti}^* S_{ti} \qquad \text{(independent of } r \text{ due to (16))}$$

$$X_{ri} = \frac{K_r V_{ri}}{\sum_j V_{rj} W_{rj}} \rightarrow \frac{\beta_r V_{ri}}{\sum_j V_{rj} W_j^*}$$

$$U_i = \sum_r X_{ri} S_{ri} \rightarrow U_i^* = \sum_r \frac{\beta_r L_{ri}}{\sum_j V_{rj} W_j^*}$$

We expect that $W_i^* > 0$ if and only if $Q_i^* > 0$ (since all throughputs X_{ri} are bounded, this follows from Little's law), so we still expect the relationships

$$W_i^* \geq 0 \qquad 1 \leq i \leq M$$

$$U_i^* \leq 1 \qquad 1 \leq i \leq M$$

$$W_i^* > 0 \implies U_i^* = 1 \qquad 1 \leq i \leq M$$

$$\sum_i V_{ri} W_i^* > 0 \qquad 1 \leq r \leq R$$

which is a non-linear complementarity problem for $\underline{W}^*$ similar to (3) and (5). It may still be solved by successive substitutions as in (8) or by a root-finder as in (10). However, we *lose* the interpretation of successive approximation being an optimization problem, and therefore lose the ability to force convergence by reducing the step length.

The one exception is if $S_{ri} = a_r b_i$ for all r and i, in which the optimization interpretation still survives [Schw79]. This case arises in models of telecommunications where r = message type, i = transmission link, and where the transmission time on a link depends on both on the (constant) message length S_r and the link speed b_i.

The empirical result is that successive substitution with stepsize reduction converges well, for arbitrary S_{ri}, despite the absence of an explanation. So asymptotic BN analysis is possible (for the dominant term) for general CQNs, although the accuracy of the higher-order terms (Q^{**}, Q^{***}, etc.) is doubtful.

Note that approximate MVA [Schw79], [Bard79] and both the linearizer [CN82] and Chow [Chow83] approximations will reduce to the fixed point problem when populations get very large. This help explains why all are accurate when the populations are very large.

4 Conclusions

One of the major problem that arise in modelling actual computer systems and networks is that the computational complexity of the exact solution techniques

becomes prohibitively expensive as the number of classes, customers, and stations grows.

As a consequence, different methods are becoming fundamental for the future of systems modelling. Among them, approximation techniques and asymptotic bottleneck analysis techniques will play an important role in the near future.

In this paper several of the principal results in bottleneck analysis for closed queueing networks, either product-form and non product-form, are described. Algorithms for the asymptotic bottleneck set identification have been presented and their applicability has been shown through examples.

References

[Bard79] Y. Bard. Some extensions to multiclass queueing network analysis. In A. Butrimenko M. Arato and E. Gelenbe, editors, *Proceedings of the 4th International Symposium on Modelling and Performance Evaluation of Computer Systems*, Performance of Computer Systems, pages 51–62, Amsterdam, Netherlands, 1979. North Holland Publishing Company.

[BGPS87] O. Bronshtein, I. Gertsbakh, B. Pittel, and S. Shahaf. One-node closed multichannel service system: Several types of customers and service rates, and random pick-up from the waiting line. *Advanced Applied Probability*, 19:487–504, 1987.

[BS93a] G. Balbo and G. Serazzi. Asymptotic analysis of multiclass closed queueing networks: Common bottleneck. Submitted for publication, 1993.

[BS93b] G. Balbo and G. Serazzi. Asymptotic analysis of multiclass closed queueing networks: Multiple bottlenecks. Submitted for publication, 1993.

[Chow83] W. M. Chow. Approximations for large scale closed queueing networks. *Performance Evaluation*, 3:1–12, 1983.

[CN82] K. M. Chandy and D. Neuse. Linearizer: A heuristic algorithm for queueing network models of computing systems. *Communications of the ACM*, 25(2):126–134, February 1982.

[KB91] Y. Kogan and A. Birman. Asymptotic analysis of closed queueing networks with bottlenecks. In *Proceedings of International Conference Performance Distributed Systems and Integr. Comm. Networks*, Kyoto, Japan, September 1991.

[KT90] C. Knessl and C. Tier. Asymptotic expansions for large closed queueing networks. *Journal of the ACM*, 37(1):144–174, January 1990.

[Lave82] S. S. Lavenberg. Closed multichain product form queueing networks with large population sizes. *Applied Probability*, pages 219–249, 1982.

[Pitt79] B. Pittel. Closed exponential networks of queues with saturation: the jackson-type stationary distribution and its asymptotic analysis. *Mathematics of Operations Research*, 4(4):357–378, November 1979.

[PKT90] K. R. Pattipati, M. M. Kostreva, and J. L. Teele. Approximate mean value analysis algorithms for queuing networks: Existence, uniqueness, and convergence results. *Journal of the ACM*, 37(3):643–673, July 1990.

[RM82] K. G. Ramakrishnan and D. Mitra. An overview of PANACEA, a software package for analyzing markovian queueing networks. *Bell Systems Technical Journal*, 61(10):2849–2872, 1982.

[Schw79] P. J. Schweitzer. Approximate analysis of multiclass closed networks of queues. In *Proceedings of International Conference on Stochastic Control and Optimization*, Free University, Amsterdam, Netherlands, April 1979.

[SSB92] P. J. Schweitzer, G. Serazzi, and M. Broglia. A fixed-point approximation to product-form networks with large population. Presented at Second ORSA Telecommunications Conference, Boca Raton, Florida, March 1992.