Tri-layered QBD Processes with Boundary Assistance for Service Resources

D Stanford[†*], W Horn[†] & G Latouche[*]

July 5, 2005

Presentation Overview:

- Tri-layered Quasi-Birth and Death Processes
- Computational Philosophy
- Introduction of U and G matrices
- Relationship to Censoring
- Asymptotic Forms of R and U
- Numerical Examples

Tri-layered QBDs with Boundary Assistance

- Classical QBD structure: infinite level & finite phase

- Insufficient for systems with two unbounded queues without sacrificing structure.

- Many queues have service resources for one type of customer that can serve the the other type when free. For example, bilingual servers intended for minority language group in a Bilingual Call Centre (see Stanford & Grassmann (1993, 2000)) can serve majority language customers. We call this capability as "Boundary Assistance".

Selected Literature on Tri-layered QBDs:

Miller (1981): two-class priority M/M/1 queue

Alfa (1998): discrete-time priority queue.

Sapna-Isoptupa & Stanford (2002): non-preemptive priority queue.

Alfa, Liu and He (2003): multi-class preemptive priority queue.

- Simplifying aspects: rate matrix R has a block-upper-triangular form
- Full $R \longrightarrow$ substantially more complicated.

Computational Philosophy:

For QBDs with matrices of infinite dimension at the top layer, and the infinite number of finite-sized blocks they contain, we determine as many elements in that infinite collection as are needed until desired thresholds are met, rather than truncating the process for the sub-level at an arbitrary point, as in the traditional approach.

Model Structure

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & \cdots \end{array} \\ \left(\begin{array}{ccccc} B' & B'_{01} & 0 & 0 & \cdots \\ B'_{10} & A'_1 & A'_0 & 0 & \cdots \\ 0 & A'_2 & A'_1 & A'_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{array} \right) \end{array}.$$

where $A'_0 = \text{diag}\{A'_{01}, \ldots\}$, $A'_2 =$diag$\{A^{*\prime}_{21}, A'_{21}, \ldots\}$. A'_{01}, $A^{*\prime}_{21}$ and A'_{21} are $m \times m$ matrices. Boundary Assistance: $A^{*\prime}_{21} e_m \geq A'_{21} e_m$.

$$A'_1 = \begin{array}{c c} & \begin{array}{ccccc} 0 & 1 & 2 & 3 & \cdots \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} & \left(\begin{array}{ccccc} A^{*\prime}_{11} & A'_{10} & 0 & 0 & \cdots \\ A'_{12} & A'_{11} & A'_{10} & 0 & \cdots \\ 0 & A'_{12} & A'_{11} & A'_{10} & \ddots \\ 0 & 0 & A'_{12} & A'_{11} & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{array} \right) \end{array}.$$

Equivalent Uniformized MC:

Set $\tau = \max\{|q_{ii}|\}$, we obtain $P = \tau^{-1}Q + I$.

$$P = \begin{array}{c c} & \begin{array}{ccccc} 0 & 1 & 2 & 3 & \cdots \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \end{array} & \left(\begin{array}{ccccc} B & B_{01} & 0 & 0 & \cdots \\ B_{10} & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{array} \right) \end{array}.$$

Complete Specification of Boundary:

In a wide range of applications,

$$B'_{01} = \begin{bmatrix} A^{*\prime}_0 \\ A'_0 \end{bmatrix};$$

$$B'_{10} = \begin{bmatrix} A^{*\prime}_2 & A'_2 \end{bmatrix}$$

where the block $A^{*\prime}_2$ is typically null. B', like A'_1, possesses a tri-diagonal form:

$$B' = \begin{array}{c} \\ \phi \\ 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \phi & 0 & 1 & 2 & \cdots \end{array} \\ \begin{pmatrix} B^{*\prime}_1 & B^{*\prime}_0 & 0 & 0 & \cdots \\ B^{*\prime}_2 & B'_1 & A'_{10} & 0 & \cdots \\ 0 & B'_2 & B'_1 & A'_{10} & \ddots \\ 0 & 0 & B'_2 & B'_1 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \end{array}$$

where we set $B'_2 = A'_{12} + \Delta_2$ and $B'_1 = A'_{11} + \Delta_1$. Typically, $\Delta_2 \geq 0$.

Example: Spatial Queue, 2 areas. Level and sub-level equal number of waiting customers in areas 1 and 2, respectively. The phase, is finite and represents the server allocation. (Horn, 2004, Ph.D. Thesis). Clearly, $A'_{01} = \{\lambda_1 I_3\}$ and $A'_2 =$diag$\{A^{*'}_{21}, A'_{21}, A'_{21}, \ldots\}$.

$$A^{*'}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 2\mu_1 & 0 & 0 \\ \mu_2 & \mu_1 & 0 \\ 0 & 2\mu_2 & 0 \end{pmatrix} \end{array} A'_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mu_1 & 0 \\ 0 & 2\mu_2 & 0 \end{pmatrix} \end{array}.$$

A'_1 is a block tri-diagonal matrix where, for example $A'_{10} = \lambda_2 I_3$, and

$$A'_{12} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 0 & 2\mu_1 & 0 \\ 0 & \mu_2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array}.$$

$$P = \begin{array}{c} \\ \phi \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} \phi & 0 & 1 & 2 & 3 & \cdots \end{array} \\ \begin{pmatrix} B_1^* & \left[B_0^*, 0, \ldots\right] & 0 & 0 & 0 & \cdots \\ \begin{bmatrix} B_2^* \\ 0 \\ \vdots \end{bmatrix} & A_1 + \Delta & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \ddots \\ 0 & 0 & A_2 & A_1 & A_0 & \ddots \\ 0 & 0 & 0 & A_2 & A_1 & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \end{array}.$$

Matrix Geometric form of the Stationary Distribution:

$$\pi(k) = \pi(0)R^k, k \geq 0.$$

Relating U and G matrices to a Censored Process

Define matrix $U = [U_{jk}]; U_{jk} =$ prob. QBD will return to its initial level n before decreasing below it, and is in phase k when it returns, given it starts from phase j. Latouche & Ramaswami (1999) show

$$U = A_1 + A_0(I - U)^{-1}A_2 = A_1 + RA_2;$$

$$R = A_0(I - U)^{-1}.$$

For infinite-sized matrices, there is more than one inverse for $(I - U)$; we require $(I - U)^{-1} = \sum_{i=0}^{\infty} U^i$. Equivalently,

$$R = A_0 + RA_1 + R^2A_2.$$

The Level-0 Censored Process

The level-0 stationary vector is obtained by censoring, (see Grassmann & Stanford (2000), 173-180, based on Kemeny, Snell, and Knapp (1966)). A denumerable Markov chain partitioned into two sets E and E', possessing a transition matrix of the form

$$P = \begin{pmatrix} T & H \\ L & S \end{pmatrix} \tag{0.1}$$

yields a censored transition matrix P^E for E

$$P^E = T + HNL \tag{0.2}$$

where $N = \sum_{m=0}^{\infty} S^m$ represents the matrix containing the expected number of visits to states in E' starting from states in E'.

Let P(1) (respectively, P(0)) represent the resulting transition matrix when all states above level 1 (resp. 0) are censored. Direct application of the method for the general model yields

$$P(1) = \begin{pmatrix} B & B_{01} \\ B_{10} & U \end{pmatrix}; \qquad (0.3)$$

$$P(0) = \begin{array}{c} \\ \phi \\ \geq 0 \end{array}\!\!\begin{array}{c} \begin{array}{cc} \quad\phi\quad & \quad\geq 0\quad \end{array} \\ \begin{pmatrix} B_1^* & \left[B_0^*, 0, \ldots\right] \\ \begin{bmatrix} B_2^* \\ 0 \\ \vdots \end{bmatrix} & \Delta + U \end{pmatrix} \end{array}. \qquad (0.4)$$

Solution for the Level-0 Stationary Vector

Theorem 1 *The level process can be positive recurrent only if*

$$\omega_0\,(A^{*\prime}_{21} - A'_{21})\,e_m + \omega\,A'_{21}\,e_m > \omega\,A'_{01}\,e_m \quad (0.5)$$

Proof: Standard QBD drift requirement $\Omega\,A'_2\,e_\infty > \Omega\,A'_0\,e_\infty$ (see Neuts (1981), p. 32) still holds for infinite-sized blocks. Expanding and collecting terms, one obtains (0.5).

Remark: The term $\omega_0\,(A^{*\prime}_{21} - A'_{21})\,e_m$ is the average long-term extra service effort available to the level process. If it is not needed to keep the level process stable, the desired asymptotic forms will result.

Theorem 2 *Assume that the level process is stable without resorting to Boundary Assistance. Then the R matrix possesses the following asymptotically block Toeplitz form, when expanded at the middle layer:*

$$R = \begin{pmatrix} R_{00} & R_{01} & R_{02} & \cdots & R_{0k} & \ddots \\ R_{10} & R_{11} & R_{12} & \cdots & R_{1k} & \ddots \\ R_{20} & R_{21} & R_{22} & \cdots & R_{00} & \ddots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \ddots \\ R_{k0} & R_{k1} & R_{k2} & \cdots & R_{kk} & \ddots \\ \ddots & \ddots & \ddots & \dots & \ddots & \ddots \end{pmatrix} \quad (0.6)$$

where $\lim_{i\to\infty} R_{i,i+k} = R_k, \ \forall i$. *The U matrix possesses a corresponding asymptotic form as well:* $\exists\ U_k = \lim_{i\to\infty} U_{i,i+k}$.

Corollary 3: *The matrix P(0) as given by (0.4) is asymptotically of block Toeplitz form.*

A one-step transition matrix P is of *GI/G/1 type* if one can write

$$P = \begin{pmatrix} B_0 & C_1 & C_2 & C_3 & \dots \\ B_1 & Q_0 & Q_1 & Q_2 & \ddots \\ B_2 & Q_{-1} & Q_0 & Q_1 & \ddots \\ B_3 & Q_{-2} & Q_{-1} & Q_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \qquad (0.7)$$

where the matrices $Q_i, i = -\infty, \dots, \infty$ are all square of common finite size, the matrix B_0 is finite but may be a different size, and the matrices C_i and $B_i,\ i = 1, 2, \dots$ are finite matrices of appropriate dimension. We choose B_0 to be sufficiently large to contain those blocks not within a suitable tolerance of their asymptotic forms.

$Q_i = U_i$ (except for $Q_{-1} = U_{-1} + \Delta_2$). The blocks C_i and $B_i,\ i = 1, 2, \ldots$ found accordingly.

The solution proceeds by repetitive block elimination. Starting at some arbitrarily large "level" K, set all blocks to be null. The finite, censored Markov chain retaining matrices up to "level" (n+1) would look like

$$P^E = \begin{pmatrix} \ddots & \ddots & \ddots & \vdots \\ \ddots & Q_0 & Q_1 & Q_2^{(n+1)} \\ \ddots & Q_{-1} & Q_0 & Q_1^{(n+1)} \\ \cdots & Q_{-2}^{(n+1)} & Q_{-1}^{(n+1)} & Q_0^{(n+1)} \end{pmatrix}. \tag{0.8}$$

However, if one retains in a single matrix $P^{(n+1)*}$ of infinite dimension all of the matrices obtained at each stage of the elimination process down to the $(n+1)st$, one would find that

$$P^{(n+1)*} = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & Q_0 & Q_1 & Q_2^{(n+1)} & \ddots \\ \ddots & Q_{-1} & Q_0 & Q_1^{(n+1)} & \ddots \\ \ddots & Q_{-2}^{(n+1)} & Q_{-1}^{(n+1)} & Q_0^{(n+1)} & \ddots \\ \ddots & Q_{-3}^{(n+2)} & Q_{-2}^{(n+2)} & Q_{-1}^{(n+2)} & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{0.9}$$

We obtain the following equations at the nth iteration (see Grassmann & Stanford (2000) equations (87) and (88)):

$$Q_i^{(n)} = Q_i + \sum_{j=1}^{\infty} Q_{i+j}^{(n+j)} (I - Q_0^{(n+j)})^{-1} Q_{-j}^{(n+j)}, \ i \geq 0; \tag{0.10}$$

$$Q_i^{(n)} = Q_i + \sum_{j=1}^{\infty} Q_j^{(n+j)} (I - Q_0^{(n+j)})^{-1} Q_{i-j}^{(n+j)}, \ i \leq 0. \tag{0.11}$$

For a sufficiently large starting point K, as n decreases, these sequences of matrices converge to their respective limits Q_i^*, given by

$$Q_i^* = Q_i + \sum_{j=1}^{\infty} Q_{i+j}^* (I - Q_0^*)^{-1} Q_{-j}^*, \ \ i \geq 0; \tag{0.12}$$

$$Q_i^* = Q_i + \sum_{j=1}^{\infty} Q_j^* (I - Q_0^*)^{-1} Q_{i-j}^*, \ i \leq 0. \tag{0.13}$$

Similarly

$$C_i^* = C_i + \sum_{j=1}^{\infty} C_{i+j}^* (I - Q_0^*)^{-1} Q_{-j}^*, \ \ i > 0; \tag{0.14}$$

$$B_i^* = B_i + \sum_{j=1}^{\infty} Q_j^* (I - Q_0^*)^{-1} B_{i+j}^*, \ i > 0. \tag{0.15}$$

Lastly, one directly evaluates

$$B_0^* = B_0 + \sum_{i=1}^{\infty} C_i^* (I - Q_0^*)^{-1} B_i^*. \tag{0.16}$$

Define non-normalized probabilities $\alpha_j = t x_j(0)$ where t is an appropriate normalizing constant, to be determined later. Setting any component (say the first) of α_0 to one, we solve

$$\alpha_0 = \alpha_0 B_0^*. \tag{0.17}$$

The succession of non-normalized probabilities is given by

$$\alpha_n = \sum_{i=1}^{n-1} \alpha_{(n-i)} Q_i^* (I - Q_0^*)^{-1} + \alpha_0 C_n^* (I - Q_0^*)^{-1}$$
$$= \sum_{i=1}^{n-1} \alpha_{(n-i)} V_i^* + \alpha_0 C_n^* (I - Q_0^*)^{-1} \qquad (0.18)$$

for V_i^* defined as $Q_i^*(I - Q_0^*)^{-1}$.

The α_n's are determined successively, and the corresponding $x_n(0) = \alpha_n/t$ is found until an index J is reached such that $x_J(0)e < \epsilon$ for a suitably tight threshold ϵ. The remaining vectors $x_j(0), j > J$ are set to zero.

Lastly, we address the question of the independent determination of t. Since in the censored Markov chain we require that $\pi_\phi(0)e + \sum_{j=0}^{\infty} x_j(0)e = 1$, it follows that t is given by

$$t = \sum_{k=1}^{m} \alpha_{0k} + \sum_{n=1}^{\infty} \alpha_n e. \tag{0.19}$$

Define the matrix-based generating functions

$$V(z) = \sum_{n=1}^{\infty} V_n^* z^n; \quad C(z) = \sum_{n=1}^{\infty} C_n^* z^n (I - Q_0^*)^{-1}; \tag{0.20}$$

$$\alpha(z) = \sum_{n=1}^{\infty} \alpha_n z^n. \tag{0.21}$$

$\alpha(1)$ is determined from α_0 via

$$\alpha(1)(I - V(1)) = \alpha_0 C(1). \tag{0.22}$$

t depends on sufficient precision in the determination of $\sum_{k=0}^{\infty} V_k^*$ and $\sum_{k=0}^{\infty} C_k^* (I - Q_0^*)^{-1}$.

Numerical Examples

Table 1: Various Six Server Configurations with 3 Majority Language Servers

N_1	N_2	M	K	$E\left(W_{Maj}\right)$	$E\left(W_{Min}\right)$
3	0	3	0	5.13	5.72
3	0	3	1	4.70	7.57
3	0	3	2	3.50	13.25
3	1	2	0	7.82	4.57
3	1	2	1	6.99	6.42
3	2	1	0	21.31	3.03
3	2	1	1	17.77	5.16

Table 2: Joint Queue Length Distribution,

$\lambda_1 = 0.5, \lambda_2 = 0.75$

Areas 1 & 2:	0	1	2	3	4	5
0	.375	.146	.064	.031	.015	.008
1	.088	.057	.032	.017	.009	.005
2	.025	.021	.014	.009	.005	.003
3	.008	.008	.006	.004	.003	.002

References Alfa, A.S. (1998) Matrix geometric solution of discrete time MAP/PH/1 priority queue. *Naval Research Logistics* **45** 23–50.

Alfa, A.S., Liu, B., and Q. HE (2003) Analysis of MAP/PH/1 multiclass general preemptive priority queue. *Naval Research Logistics* **50** 662-682.

Grassmann, W.K. and Stanford, D.A. (2000) Matrix Analytical Methods. In W.K. Grassmann (Ed.), *Computational Probability*, pp. 153-203, Kluwer Academic Publishers.

Horn, W. and Stanford, D.A. (2005) A Matrix-Geometric Model of a Spatially Distributed Queueing System. *In preparation*.

Horn, W., Stanford, D.A. and Latouche, G. (2005) Bilingual Call Centres with Two Unilingual Server Pools. Submitted for publication, January 2005.

Latouche, G. and Ramaswami, V. (1999) *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA SIAM, Philadelphia.

Miller, D.R. (1981) Computation of Steady State Probabilities for M/M/1 Priority Queues. *Operations Research* **29** 945-958.

Neuts, M.F. (1981) *Matrix-Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore.

Sapna Isotupa, K.P. and Stanford, D.A. (2002) An infinite-phase quasi-birth-and-death model for the non-preemptive priority M/PH/1 queue. Stochastic Models **18** 387–424.

Stanford, D.A. and Grassmann, W.K. (1993) The bilingual server system: a queueing model featuring fully and partially qualified servers. *INFOR* **31** 261-277.

Stanford, D.A. and Grassmann, W.K. (2000) Bilingual server call centres. In D.R. McDonald and S.R.E. Turner, editors, *Analysis of Communication Networks: Call Centres, Traffic, and Performance. Fields Institute Communications* **28** 31–47.