Derivation of Product-Form Approximation for Tandem Queues via Matrix Geometric Method

Giuliano Casale, Peter G. Harrison, Maria Grazia Vigliotti
Department of Computing
Imperial College London
g.casale@imperial.ac.uk, \{pgh, mgv98\}@doc.ic.ac.uk

Abstract—We introduce a product-form approximation for tandem networks with Poisson arrivals and non-exponential service times. The proposed technique perturbs the model state space to match the sufficient conditions for product-form solution provided by the Compound Reversed Agent Theorem (RCAT). After characterizing the relation between RCAT product-forms and matrix geometric solutions, we develop an algorithm based on nonlinear programming that automatically searches for an approximating product-form model.

I. INTRODUCTION

We consider a tandem model composed by two queues processing requests arriving at the first queue according to a Poisson process with rate $\lambda$. Both queues have infinite capacity and schedule requests according to a first-come first-served policy. Service times are phase-type distributed, thus they can fit a wide range of distributions including exponential, hyper-exponential, hypo-exponential, and heavy-tail distributions. Phase-type distributions are defined using the method of phases, thus time is modeled using passage times in continuous-time Markov chains; upon activation of a tagged transition, a service time sample is generated as the cumulative time elapsed from the last activation of a tagged transition.

Throughout this paper, we represent phase-type distributed service times using the Markovian arrival process (MAP) notation. Denote with $\mu_{k,k'}$ the rate of a tagged transition between states $k$ and $k'$, a MAP is described by a pair of matrices $(D_0, D_1)$ where $\mu_{k,k'}$ is the entry in row $k$ and column $k'$ of $D_1$, while $D_0 = Q - D_1$, being $Q$ the infinitesimal generator of the underlying continuous-time Markov chain that does not distinguish between tagged and untagged transitions. Let us call queue $a$ the first queue to receive jobs, and let queue $b$ be the second queue. We denote with $(D^a_0, D^a_1)$ and $(D^b_0, D^b_1)$ the service processes respectively of the first and second queue and assume that the number of phases are $K$ and $H$.

II. RCAT CONDITIONS FOR TANDEM QUEUES

Product-form approximations of state probabilities are obtained using the RCAT theorem [4]. RCAT provides sufficient conditions for communicating Markov processes to enjoy a product-form solution in their joint state space. RCAT has also been shown to unify existing product-form results for queueing networks, stochastic Petri nets, and to drive the definition of new product-forms [1]. In RCAT, a feed-forward tandem network is represented by two communicating Markov processes with infinitesimal generators $Q^a : (n, k) \rightarrow (n', k')$ for queue $a$ and $Q^b : (m, h) \rightarrow (m', h')$ for queue $b$, where $n$ is the number of jobs in queue $a$ and $k$ is the active phase in its service process, the population $m$ and phase $h$ are similarly defined for queue $b$. The equilibrium probabilities of $Q^a$ and $Q^b$ are denoted respectively by the vectors $\alpha_n$, $n \geq 0$, and $\beta_{m,h}$, $m \geq 0$, where $\alpha_n$ is the $k$th element of $\alpha_n$ describing the equilibrium probability of state $(n, k)$; similarly, the element $\beta_{m,h}$ is the probability of state $(m, h)$ in queue $b$. The tandem network state space is then obtained by composing $Q^a$ and $Q^b$, see [4] for related background.

RCAT’s conditions for a product form solution in the joint process given by the composition of $Q^a$ and $Q^b$ are as follows [4]:

\begin{itemize}
  \item [C1)] queue $b$ can accept an incoming job in any of its states $(m, h);
  \item [C2)] each state $(n, k)$ of queue $a$ has an incoming transition for each tagged event corresponding to a job arrival at queue $b$ – such transitions are called active transitions;
  \item [C3)] the sum of reversed rates of all active transitions incoming to any state $(n, k)$ is constant (within each class of transition in a multi-class model).
\end{itemize}

where we here consider the formulation of C3 proposed in [8]. If all the above conditions are simultaneously satisfied, then the equilibrium distribution of the model becomes

$$\pi((n, k), (m, h)) = \alpha_n \cdot k \cdot \beta_{m,h} \tag{1}$$

where $\beta_{m,h}$ is determined after setting to an appropriate value the arrival rate at the second queue $b$ according to RCAT’s rate equations [4]. We refer to (1) as RCAT product-form since other product-form conditions exist, e.g., ERCAT product-forms [1].

In the tandem models we consider, condition C1 is always satisfied since queue $b$ has infinite buffer size. We instead propose to modify the state space of $Q^a$ to match RCAT conditions C2 and C3 and define a new approximating model that has product-form solution. Due to the complexity of the problem, we here focus on models with a single-class of transitions synchronizing between $Q^a$ and $Q^b$, i.e., all active

The work of G. Casale was supported by the Imperial College Junior Research Fellowship. The work of P. G. Harrison and M. G. Vigliotti was supported by the Engineering and Physical Sciences Research Council of the United Kingdom, research grant number EP/D047587/1.
transitions are marked with an identical label [4]. While the results provided here are general under this assumption, the single-class case restricts in practice the number of models where we can find an approximating model that admits a RCAT product-form. We plan to characterize in future work the class of models that admit such single-class RCAT product-form as well the improvements resulting from using multiple classes. In particular, it would be interesting to assess the applicability to nonrenewal service times which are hard to approximate using a single-class of transitions. Under the single-class assumption, condition C3 is equivalent to verifying the existence of a constant \( \bar{q} \) such that

\[
\sum_{(n',k')} \alpha_{n,k} q((n',k') \rightarrow (n,k)) / \alpha_{n',k'} = \bar{q}, \quad \forall(n,k),
\]

where the summation considers only active transition rates \( q((n',k') \rightarrow (n,k)) \) between states \((n',k')\) and \((n,k)\) in \(Q^a\) that result in an arrival to queue \( b \).

The basic idea of the product-form approximation we consider in this paper is to introduce perturbations on the structure and rates of the \( Q^a \) process to make the tandem network a product-form model. Throughout the next sections, we take the same qualitative approach proposed in [5] for \( M/E_2/1 \rightarrow -/M/1 \) networks, where \( E_2 \) is an Erlang-2 service process, but greatly extend its scope to queues with phase-type distributed service times. This leads to substantial differences in the approach since, without the special structure of an \( E_2 \) process, it is no longer possible to symbolically solve for \( \bar{q} \) in (2). Determining the \( \bar{q} \) value is fundamental to verify RCAT’s conditions and to determine the equilibrium probabilities of \( Q^b \) using RCAT’s rate equations.

### A. Matrix Geometric Solutions and RCAT Product-Forms

We begin with observing that since the behavior of the first queue is independent of the second queue, we can solve for the equilibrium state probabilities of \( Q^a \) using the quasi-birth-death (QBD) process

\[
Q^a = \begin{bmatrix}
L_0 & F_0 & 0 & 0 & 0 & \ldots \\
B_0 & L & F & 0 & 0 & \ldots \\
0 & B & L & F & 0 & \ldots \\
0 & 0 & B & L & F & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\]

where the arrival and service rates define the forward transition matrix \( F = \lambda I \), the local transition matrix \( L = -\lambda I + D_0 \), and the backward transition matrix \( B = D_1 \). The transition matrices \( L_0, B_0, \) and \( F_0 \) describe boundary conditions for states where the first queue is empty. The aim of this section is to develop a characterization of the relations between RCAT condition C3 and matrix geometric solutions of the above QBD. We find in particular that a RCAT product-form impose special properties on the equilibrium distribution of the QBD.

We begin by writing the global balance equations for states where queue \( a \) has population of two or more jobs, i.e.,

\[
\alpha_{n-1} F + \alpha_n L + \alpha_{n+1} B = 0, \quad n \geq 2.
\]

We now make the fundamental observation that, for a product-form model that satisfies C3, inserting (2) in the global balance equations of \( Q^b \) for state \((n',k')\) always cancels out the probability flux of the incoming transitions from the states \((n,k)\) contributing to the summation in (2). This can be equivalently stated in matrix notation by first noting that C3 implies

\[
\alpha_n \hat{L} = \alpha_{n+1} B, \quad n \geq 1,
\]

in which we define

\[
\hat{L} = \begin{bmatrix}
\bar{q}\delta_1 & 0 & 0 & \cdots & 0 \\
0 & \bar{q}\delta_2 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
0 & 0 & 0 & \cdots & \bar{q}\delta_K
\end{bmatrix}
\]

where \( \delta_k \) is 1 if the \( k \)th column of \( D_k^T \) has at least a nonzero element, 0 otherwise. Inserting (4) into (3), C3 implies that

\[
\alpha_{n-1} F + \alpha_n (L + \hat{L}) = 0, \quad n \geq 2.
\]

Since \( L \) is invertible also \( L + \hat{L} \) is invertible, thus

\[
\alpha_n = \alpha_{n-1} H, \quad n \geq 2,
\]

where \( H = -F(L + \hat{L})^{-1} \). Note that this is also the solution of the matrix equation \( F + HL + HL = 0 \). However, the matrix geometric solution requires that there exist a rate matrix \( R \) such that

\[
\alpha_n = \alpha_{n-1} R, \quad n \geq 2
\]

where \( R \) is the non-negative minimum norm solution of \( F + RL + R^2 B = 0 \). Thus, we find that if the QBD process satisfies C3 there exist a rate matrix

\[
H = -F(L + \hat{L})^{-1} = R(L + RB)(L + \hat{L})^{-1},
\]

that provides an additional geometric relation (6) for the computation of the equilibrium solution. Note that \( H \) is different from \( R \) whenever \( RB \neq L \), yet (6) and (7) clearly imply dependencies between the spectral properties of \( H \) and \( R \). In fact, we note that in the limit for \( n \rightarrow \infty \) it is \( \alpha_n \propto \eta^n \alpha_{n-j} \), where \( \eta \in [0,1] \) is the largest eigenvalue of \( R \) that describes the asymptotic decay rate of queue-length probabilities. However, by (6), \( \eta \) must also be the largest eigenvalue of \( H \). Thus, \( R \) and \( H \) must have the same spectral radius

\[
\rho(R) = \rho(H) = \eta
\]

and impose identical stability conditions on the QBD, i.e., \( \rho(R) < 1 \). The importance of the last equation is that one can find \( \bar{q} \) by searching for the \( \hat{L} \) matrix that imposes \( \rho(R) = \rho(H) \) and then extract directly \( \bar{q} \) from its main diagonal.

To complete the characterization of the \( Q^a \) process when C3 holds true, we are left with verifying the implications of C3 on the global balance conditions relating the vectors \( \alpha_0 \) and \( \alpha_1 \); such global balance equations are not considered in (3). Following the same argument used before, we see that C3
implies that the following conditions hold true simultaneously
\[
\begin{align*}
\alpha_0 F_0 + \alpha_1 L + \alpha_1 RB &= 0 \\
\alpha_0 L_0 + \alpha_1 B_0 &= 0 \\
\alpha_0 1 + \alpha_1 (1 - R)^{-1} 1 &= 1 \\
\alpha_0 \tilde{L}_0 - \alpha_1 B_0 &= 0 \\
\alpha_1 \tilde{L} - \alpha_1 RB &= 0
\end{align*}
\]
where we have used that \(\alpha_2 = \alpha_1 R, 1\) is a vector of ones of proper size, (10) normalizes the probabilities, and \(L_0\) is defined similarly to \(\tilde{L}\) but relatively to the states where the QBD is empty. Note that in (8)-(12) only the last two relations follow by C3, since (8)-(10) are valid for all QBDs.

III. PRODUCT-FORM APPROXIMATION ALGORITHM

In order to obtain a product-form model that behaves as closely as possible to the original tandem network, we propose to introduce two perturbations in the state space. The first perturbation assumes that the rates in \(F_0\) can be assigned to different values in order to impose condition C3 on the entire state space. The second perturbation involves adding a self-looping active transition to all states \((n, k)\) that do not satisfy the RCAT condition C2; such transitions are referred to as invisible transitions and to match condition C3 must have rate \(\tilde{q}\). Note that the activation of an invisible transition does not change the steady state distribution of queue \(a\), and its only effect is to have a new job arriving to queue \(b\) but not due to a departure from queue \(a\). Following an argument similar to [5], it is easy to see that invisible transitions describe the action of an exogenous Interrupted Poisson Process (IPP) that is superposed to the departure flow of the QBD and that is active if and only if queue \(a\) is in a state with a self-looping invisible transition. Thus, the modified QBD process now communicates to \(Q^b\) also job arrivals from the IPP arrival process\(^3\).

Since C2 can be easily imposed, the main challenge of the proposed approximation is to search for appropriate rates in \(F_0\) ensuring an equilibrium probability distribution satisfying (4) and (8)-(12) which are equivalent to C3. In order to impose the infinite conditions (4) we first observe that any set of \(K\) powers of \(R\) must be linearly dependent being \(R\) a square matrix of order \(K\) [7]. Thus, C3 is satisfied whenever (8)-(12) and (4), for \(n = 1, \ldots, K - 1\), hold true. To obtain this result, we propose to search over the space of the rates in \(F_0\) using a nonlinear optimization program. At each iteration we determine the values of \(\alpha_0\) and \(\alpha_1\) for the current assignment of the rates in \(F_0\); this may be done using tools such as [2] which automatically determine \(\alpha_0\) and \(\alpha_1\) from conditions (8)-(10). The values of \(\alpha_0\), \(\alpha_1\), and \(F_0\) can then be used to drive the minimization of the objective function
\[
f_{\text{obj}} = ||\alpha_0 \tilde{L}_0 - \alpha_1 B_0||_2 + \sum_{i=1}^{K-1} ||\alpha_1 R^{i-1} \tilde{L} - \alpha_1 R^i B||_2
\]
which equals zero if C3 holds on all states of the QBD. We have observed in several occasions that the above formulation avoids stagnation problems that arise when searching directly for a feasible solution to the nonlinear system of equations (4) and (8)-(12) in the unknowns \(F_0, \alpha_0\), and \(\alpha_1\). A summary of the proposed approximation is given in the following pseudocode.

Single-Class RCAT Product-Form Approximation:
1) Determine \(R\) for given \(F, L, B\)
2) Compute numerically the largest eigenvalue \(\eta\) of \(R\)
3) Determine \(\tilde{q}\) such that
\[
\rho(\bar{F}(\bar{L} - \tilde{L})^{-1}) = \eta
\]
for unknown \(\tilde{q}\) in \(\tilde{L}\).
4) Using the rate \(\tilde{q}\) found in the previous step, search for values of the entries in \(F_0\) satisfying (4) and (8)-(12).
   At each iteration, \(\alpha_0\), \(\alpha_1\) are immediately determined by the linear conditions (8)-(12) for assigned \(F_0\).

IV. CASE STUDY

We illustrate our product-form approximation on a case study. We consider a tandem network \(M/H_{\text{ypo3}}/1 \rightarrow /E_2/1\), where \(E_2\) denotes an Erlang-2 distribution and \(H_{\text{ypo3}}\) indicates a hypo-exponential process composed by three sequential stages with rates \(\mu_{1,1} = 1, \mu_{2,2} = 2, \mu_{3,3} = 3\), where \(\mu_{1,1}, \mu_{2,2} \in D_0\) and \(\mu_{3,3} \in D_1\). The arrival process to the first queue is Poisson with rate \(\lambda = 0.25\). The QBD for \(Q^a\) is then defined by
\[
B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} -1.25 & 1 & 0 \\ 0 & -2.25 & 2 \\ 0 & 0 & -3.25 \end{bmatrix},
\]
\[
F = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}.
\]
The rate matrix \(R\) is computed using the tool presented in [2] and its eigenvalues are found to be \(\eta(R) = 0.3522, 0.0856 + 0.008i, 0.0856 - 0.008i\), thus \(\eta = \rho(R) = 0.3522\). Using this value, we determine numerically the minimum of \(\rho(\bar{F}(\bar{L} - \tilde{L})^{-1}) - \eta\); this function is plotted in Figure 1 as a function of the \(\tilde{q}\) element in \(\bar{L}\). The minimization finds that a product form requires a reversed rate \(\tilde{q} = 0.54\), thus
\[
\tilde{L} = \begin{bmatrix} 0.54 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}
\]
where the zero diagonal elements correspond to the empty columns in \(B\).

We now define \(F_0 = [\lambda + x_0, x_1, \ldots, x_K]\), where \(x_j\) are the perturbations we introduce in \(F_0\) to obtain a product-form. We set up a non-linear optimization program searching for the values of \(x_0, x_1, \ldots, x_K\) generating a distribution that satisfies (4) and (8)-(12). We use MATLAB’s fmincon function to solve the nonlinear optimization program. This provides the following result: \(x_0 = 0, x_1 = 0.1623, x_2 = 0.1278\). Thus
\[
F_0 = \begin{bmatrix} 0.2500 & 0.1623 & 0.1278 \end{bmatrix}, \quad L_0 = \begin{bmatrix} -0.5401 \end{bmatrix}.
\]

\(^3\)Since the above approximation only injects new load into the system, the resulting product-form model may be used as a pessimistic estimate on the performance of the tandem network, see [3], [6] and references therein for related techniques.
Fig. 1. Plot of the absolute spectral radius difference between \( H \) and \( R \). The minimum is for \( q = 0.54 \).

Fig. 2. The single-class RCAT product-form approximation is based on the (invisible) self-looping transitions in phases 2 and 3 and on the new transitions \( x_1 \) and \( x_2 \) (highlighted in blue) introduced in \( F_0 \). All active transitions (highlighted in red) generate arrivals to queue \( b \).

where \( L_0 \) is defined by the element \(-\lambda - x_0 - x_1 - x_2 = -0.5401\) that appears on the main diagonal of the QBD; note that this is equal to \(-q\) similarly to what found in [5] for the Erlang-2 case.

We now compute numerically the entire probability distribution \( \alpha_n, n \geq 0 \), which gives the reversed rate \( q \) in (2) that need to be equal to satisfy the C3 condition of RCAT. Table I shows that this is the case with good precision, hence condition C3 holds. Based on this result, the state space of the first queue is corrected as shown in Figure 2, where the red transitions represent marked events that produce an arrival to the second queue. Note in particular that red self-loops are used to denote arrivals to the second queue due to the extra IPP flow with rate \( \bar{q} \).

We conclude the case study illustrating the marginal probabilities of queue \( b \) in the above case study. Probability values are obtained by simulation initializing both queues in the empty state. Table II reports simulation results; similar differences are observed if the Erlang-2 distribution is replaced by an exponential or hyper-exponential distribution. In this experiment, queue \( a \) utilization is \( \rho_a = 0.458 \) in the original model and \( \rho'_a = 0.462 \) after the modification of \( F_0 \). For queue \( b \) it is \( \rho_b = 0.25 \) in the original model and becomes \( \rho'_b = 0.60 \) in the product-form approximation. Thus, we see that the additional jobs injected in the system by the IPP flow and by the \( x_3 \) perturbations result in an increased load at queue \( b \). In fact, all product-form values for states where the queue is busy are greater than the corresponding values for the original model. Indeed, depending on the service characteristics at the second queue, there exist perturbations that result in saturation at the second queue, in this case the approximation does not return a valid product-form model. We plan to characterize conditions under which these cases arise in future work.

V. CONCLUSION

We have proposed a new approximation method for tandem network of queues with Markovian service. We have found that product-form conditions applied to QBDs determine a novel matrix geometric relation between state probabilities. This allows us to formulate the RCAT conditions compactly and integrate them into an optimization program that generates an approximate product-form.

Possible generalization of the above results include considering generalized arrival processes and multiclass service rates; the latter are handled immediately within RCAT’s theory [4].

REFERENCES