Journal of Applied Mathematics and Stochastic Analysis, **12**:4 (1999), 357-370.

TWO SERVICE UNITS WITH INTERFERENCE IN THE ACCESS TO SERVERS

ROSA E. LILLO
Universidad Carlos III de Madrid
Dpto. de Estadistica y Econometria
C/ Madrid 126, Getafe 28903 Madrid, Spain
E-mail: lillo@est-econ.uc3m.es

MARCEL F. NEUTS
The University of Arizona
Dept. of Systems and Industrial Eng.
Tucson, AZ 85721 USA
E-mail: marcel@sie.arizona.edu

(Received: December, 1998; Revised July, 1999)

We examine the service mechanism of two queueing models with two units in tandem. In the first model, customers who complete service in Unit 1 must wait in an intermediate buffer until the ongoing service in Unit II ends. In the second model, jobs can be pre-positioned in an intermediate buffer to await service in Unit II. Under the assumption of phase-type service times, the steady-state regime of the service system is studied in detail.

The models are inspired by the gas pump model of A.B. Clarke and by phenomena observed in cafeteria lines and certain manufacturing systems. However, their primary interest may lie in the methodology of their exceptionally tractable analysis. We derive formulas for the throughput and other quantities by using the familiar *PH*-formalism. These formulas turn out to be unusually transparent and have probabilistic interpretations that do not depend on the *PH* assumptions. These interpretations therefore also hold for general service time distributions. The methodology is general and can be applied to other systems with interactions between servers. The models also present interesting algorithmic problems of didactic interest.

Key words: Queueing Systems, Phase-Type Distributions, Throughput Analysis.

AMS subject classifications: 60K25, 60K99.

1. Introduction

We discuss two systems of queues with two service stations. Access to the second server or departures from the first station may be blocked by customers present in the other unit. In 1977 and 1978, in a series of preprints, A.B. Clarke introduced the first model. Informally described, it represents the interference one might see at two gasoline pumps with a narrow lane so that a customer who has completed service at the second pump may be blocked by a vehicle that is still using the first pump.

It was recognized that, under the assumption of exponential service times, the model called *A.B. Clarke's Tandem Queue* can be studied as a continuous-time Markov chain of $GI/M/1$ type. A fairly detailed analysis is found in Section 5.3 of Neuts [5].

Our model I is just Clarke's tandem queue with service time distributions of phase-type that are discussed in Chapter 2 of the cited book by Neuts. It is a fairly routine matter to extend the analysis of Markovian queueing systems with exponential assumptions to the corresponding models with PH service time distributions. One usually obtains Markov chains having the classical block structures but with blocks of large, often huge, dimensions. The resulting algorithmic problems are challenging but their discussion is not the purpose of this paper.

We shall show that the analysis with phase-type distributions can provide significant new physical insight that is lost in the multitude of special and interrelated simplifications due to the exponential assumptions. Important quantities can be given probabilistic interpretations that are independent of the PH-formalism. These interpretations are therefore also valid for general service time distributions. That points the way to the computation of these quantities by other analytic means or to their measurement through efficient simulations.

We concentrate on the *throughput analysis* of the two-server systems rather than on a full analysis of the queues. That is, we assume that an unlimited number of jobs are waiting for service and we analyze the departure rate of customers and various other descriptors of the service mechanism.

The analysis of the models suggests interesting algorithmic problems of clear didactic interest. A numerical exploration will yield insight into queueing models with behavior not seen in the classical models.

2. Description of the Models

In both models, there are two servers I and II with I placed to the left of II. The job flow is from left to right. Each customer is served by only one of the servers. All service times are independent random variable but the duration of those dispensed by Server I have the probability distribution $F_1(\cdot)$; those of Server II obey the probability law $F_2(\cdot)$. These probability distributions are of phase type. For $j = 1, 2$, $F_j(\cdot)$ has m_j phases and the irreducible representation $(\beta(j), S(j))$. The column vectors $\mathbf{S}^0(j) = -S(j)\mathbf{e}$ play their usual role in the formalism of PH-distributions.

The interest of Clarke's tandem model to queueing analysis lies in the quantification of the interference between the two servers. When there is a job in course in Unit II, a customer whose service is completed in I cannot leave the system immediately. Server I remains blocked as long as his customer cannot move forward. To alleviate that blocking there is a finite waiting space between the two servers. The distinction between our two models lies in the way that this waiting space is used.

In distinguishing between the models, we see what happens at service completions in Unit I. In Model I, if upon such a completion Server II is free, the completed job leaves the system and two new jobs are initiated, one by each server. If Server II is not free, the completed job moves into a buffer of size n and Server I initiates a new job. Should that buffer be full, the completed job blocks Unit I. At the end of a job in II, that customer and all those that have already been served by I leave the system. Server II then remains idle until the job in I is finished.

In Model II, the buffer is used to pre-position customers for Server II. If upon a job completion in Unit I Server II is free, the completed job leaves the system and $m+1$ new jobs are brought into the service system; one for each server to start processing and $m-1$ to wait in the buffer for service by II. These customers leave one at a time as they finish service in II. Jobs completed by I can leave the system only when there are not active jobs downstream.

With phase-type service times, the service systems of each of the models can be described as finite Markov chains, typically with many states, but with highly structured infinitesimal generators. To make that structure clear, we display these generators for small but representative values of n or m. By exploiting their special structure, we obtain explicit matrix formulas for various steady-state descriptors of the two models. These formulas can be used in numerical explorations from which substantial physical insight can be gained. Our approach is an example of the general modeling methodology advocated in [8].

3. Discussion of Model I

The state space of the Markov chain for Model I consists of $n+3$ macro-states denoted by **0,1,...,n+2**. The macro-state **0** corresponds to the case where Unit II is empty and it contains the m_1 labels of the phases for service time distribution $F_1(\cdot)$. The set of states **n+2** contains the m_2 phases to service time distribution $F_2(\cdot)$. It corresponds to the case where Server I is blocked. The macro-states **1,...,n+1** contain the $m_1 m_2$ phases of the two service times in course. Within each macro-state **i** with $1 \leq i \leq n$ the phases are listed in lexicographic order. The macro-state **i** signifies that Server II is occupied and $i-1$ customers who have already been served by I are waiting in the buffer. For $i = n+1$, the buffer is full but the service in I is going on. If that service terminates before that in Unit II, blocking occurs. The Markov chain then moves into the macro-state **n+2**.

We make extensive use of the Kronecker produce $\otimes$ and sum $\oplus$ of matrices. The properties of these operations are discussed in many books on matrix algebra; a summary of these for use with PH-distributions is given in Chapter 2 of [5].

The infinitesimal generator Q_1, displayed here for $n = 3$ is given by

$$Q_1 = \begin{array}{c} \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \end{array} \begin{pmatrix} S(1) & \mathbf{S}^0(1)\boldsymbol{\beta}(1) \otimes \boldsymbol{\beta}(2) & 0 & 0 & 0 & 0 \\ I \otimes \mathbf{S}^0(2) & S(1) \oplus S(2) & \mathbf{S}^0(1)\boldsymbol{\beta}(1) \otimes I & 0 & 0 & 0 \\ I \otimes \mathbf{S}^0(2) & 0 & S(1) \oplus S(2) & \mathbf{S}^0(1)\boldsymbol{\beta}(1) \otimes I & 0 & 0 \\ I \otimes \mathbf{S}^0(2) & 0 & 0 & S(1) \oplus S(2) & \mathbf{S}^0(1)\boldsymbol{\beta}(1) \otimes I & 0 \\ I \otimes \mathbf{S}^0(2) & 0 & 0 & 0 & S(1) \oplus S(2) & \mathbf{S}^0(1) \otimes I \\ 0 & \boldsymbol{\beta}(1) \otimes \mathbf{S}^0(2)\boldsymbol{\beta}(2) & 0 & 0 & 0 & S(2) \end{pmatrix}$$

We partition the stationary probability vector $\boldsymbol{\pi}$ of Q_1 according to the macro-states as

$$\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \ldots, \boldsymbol{\pi}(n+1), \boldsymbol{\pi}(n+2).$$

The explicit matrix formulas for the steady-state probabilities involve two matrices M and N, each having a probabilistic significance to be explained later.

The matrix M is $m_2 \times m_2$ and is give by:

$$M = [\beta(1) \otimes I][-S(1) \oplus S(2)]^{-1}[\mathbf{S}^0(1) \otimes I]. \tag{1}$$

The matrix N is $m_2 \times m_1$ and is given by:

$$N = [\beta(1) \otimes I][-S(1) \oplus S(2)]^{-1}[I \otimes \mathbf{S}^0(2)]. \tag{2}$$

With irreducible representations of the two-phase-type distributions, the matrix M is irreducible and sub-stochastic. The matrix N is nonnegative and, as is easily verified, $M\mathbf{e} + N\mathbf{e} = \mathbf{e}$.

Theorem 3.1: *The vectors in the partitioned form of $\boldsymbol{\pi}$ are given by*

$$\boldsymbol{\pi}(0) = k^*\beta(2)(I - M^{n+1})(I - M)^{-1}N[-S(1)]^{-1}, \tag{3}$$

$$\boldsymbol{\pi}(i) = k^*\beta(2)M^{i-1}[\beta(1) \otimes I][-S(1) \oplus S(2)]^{-1} \tag{4}$$

for $1 \leq k \leq n+1$, and

$$\boldsymbol{\pi}(n+2) = k^*\beta(2)M^{n+1}[-S(2)]^{-1}. \tag{5}$$

The normalizing constant k^* is given by

$$[k^*]^{-1} = \mu_1'(2) + \beta(2)(I - M^{n+1})(I - M)^{-1}N[-S(1)]^{-1}\mathbf{e}, \tag{6}$$

where $\mu_1'(2)$ is the mean service time in Unit II. The quantity k^* is the rate at which simultaneously services in both units are initiated. It is also the steady-state rate of departure of customers processed by Server II.

Proof: The steady-state equations are:

$$\boldsymbol{\pi}(0)S(1) + \sum_{r=1}^{n+1} \boldsymbol{\pi}(r)[I \otimes \mathbf{S}^0(2)] = \mathbf{0}, \tag{7}$$

$$\boldsymbol{\pi}(0)[\mathbf{S}^0(1)\beta(1) \otimes \beta(2)] + \boldsymbol{\pi}(1)[S(1) \oplus S(2)] + \boldsymbol{\pi}(n+2)[\beta(1) \otimes \mathbf{S}^0(2)\beta(2)] = \mathbf{0}, \tag{8}$$

$$\boldsymbol{\pi}(i-1)[\mathbf{S}^0(1)\beta(1) \otimes I] + \boldsymbol{\pi}(i)[S(1) \oplus S(2)] = \mathbf{0} \tag{9}$$

for $2 \leq i \leq n+1$, and

$$\boldsymbol{\pi}(n+1)[\mathbf{S}^0(1) \otimes I] + \boldsymbol{\pi}(n+2)S(2) = \mathbf{0}. \tag{10}$$

We notice that

$$\boldsymbol{\pi}(0)[\mathbf{S}^0(1)\beta(1) \otimes \beta(2)] = [\boldsymbol{\pi}(0)\mathbf{S}^0(1)][\beta(1) \otimes \beta(2)], \tag{11}$$

$$\boldsymbol{\pi}(n+2)[\beta(1)\otimes \mathbf{S}^0(2)\beta(2)] = [\boldsymbol{\pi}(n+2)\mathbf{S}^0(2)][\beta(1)\otimes\beta(2)], \tag{12}$$

and

$$\beta(2)[\beta(1)\otimes I] = \beta(1)\otimes\beta(2). \tag{13}$$

By using (11) and (12) we may write (8) as

$$\boldsymbol{\pi}(1) = k^*[\beta(1)\otimes\beta(2)][-S(1)\oplus S(2)]^{-1}, \tag{14}$$

where

$$k^* = \boldsymbol{\pi}(0)\mathbf{S}^0(1) + \boldsymbol{\pi}(n+2)\mathbf{S}^0(2). \tag{15}$$

From (9) it follows, for $2 \leq i \leq n+1$,

$$\boldsymbol{\pi}(i) = \boldsymbol{\pi}(i-1)[\mathbf{S}^0(1)\otimes I][\beta(1)\otimes I][-S(1)\oplus S(2)]^{-1}.$$

Equation (4) now readily follows. Upon substituting these expressions into (7) and (10), the formulas (3) and (5) follow by routing calculations.

The quantity k^* has several interesting interpretations. From its definition in (15), it is clearly the steady-state rate at which service in both units are simultaneously started. The rate of terminations of jobs in Unit II is given by

$$\sum_{i=1}^{n+1}\boldsymbol{\pi}(i)[\mathbf{e}\otimes\mathbf{S}^0(2)] + \boldsymbol{\pi}(n+2)\mathbf{S}^0(2),$$

which after substitutions and routine simplification reduces to k^*.

From the normalizing equation starting that $\boldsymbol{\pi}$ is a probability vector, we obtain $[k^*]^{-1}$ as the sum of three expressions. The first of these, $\boldsymbol{\pi}(0)\mathbf{e}$, is just the second term in the equation (6).

To show that the sum

$$\beta(2)(I-M^{n+1})(I-M)^{-1}[\beta(1)\otimes I][-S(1)\oplus S(2)]^{-1}(\mathbf{e}\otimes\mathbf{e}) \\ + \beta(2)M^{n+1}[-S(2)]^{-1}\mathbf{e},$$

of the other two reduces to the mean service time $\mu_1'(2)$ in Unit II requires slightly intricate manipulations with Kronecker products. We show the essential steps.

In the first expression, we write

$$[-S(1)\oplus S(2)]^{-1}(\mathbf{e}\otimes\mathbf{e}) \\ = -[-S(1)\oplus S(2)]^{-1}[I\otimes S(2)][\mathbf{e}\otimes[-S(2)]^{-1}\mathbf{e}], \tag{16}$$

and we notice that

$$[-S(1)\oplus S(2)]^{-1}[I\otimes S(2)] = -I\otimes I - [-S(1)\oplus S(2)]^{-1}[S(1)\otimes I]. \tag{17}$$

Recalling the definition of M, routine matrix calculations now show that the sum of the two involved expressions reduces to $\beta(2)[-S(2)]^{-1}\mathbf{e} = \mu_1'(2)$. □

The Exponential Case

Upon replacing $S(1)$ and $S(2)$ by the scalars $-\mu_1$ and $-\mu_2$ and $\beta(1)$ and $\beta(2)$ by one, we obtain the case of exponential service times discussed in [5]. The infinitesimal generator Q_1 agrees with the matrix A of formula (5.3.3) except that the latter corresponds to $n = 4$.

The matrix M reduces to the scalar $p = \mu_1(\mu_1 + \mu_2)^{-1}$, and k^* is given by

$$k^* = [\mu_2^{-1} + (1 - p^{n+1})\mu_1^{-1}]^{-1}.$$

The steady-state probabilities $\pi(i)$ are as stated in formula (5.3.4) of [5].

The Departure Rate

Next we calculate the steady-state rate $\mu^*(n)$ at which jobs leave the service system or the *throughput* of the system. The differential $\mu^*(n)dt$ may be interpreted as the expected number of completed jobs departing in a time interval of length dt. We observe that, in steady-state, there are single departures at a rate $\boldsymbol{\pi}(0)\mathbf{S}^0(1)$. Customers can also leave in groups of sizes i, $2 \leq i \leq n+2$, consisting of one customer processed by II and $i-1$ by Server I.

Theorem 3.2: *The departure rule $\mu^*(n)$ is given by*

$$\mu^*(n) = k^*\beta(2)[2I - M - M^{n+1}](I - M)^{-1}\mathbf{e}. \tag{18}$$

The rate $\mu^*(n)$ is equal to $\lambda^*(n)$, the steady-state rate at which services are initiated.

Proof: We have that

$$\mu^*(n) = \boldsymbol{\pi}(0)\mathbf{S}^0(1) + \sum_{i=1}^{n+1} i\boldsymbol{\pi}(i)[\mathbf{e} \otimes \mathbf{S}^0(2)] + (n+2)\boldsymbol{\pi}(n+2)\mathbf{S}^0(2).$$

Using the expressions in Theorem 3.1, recalling that $N\mathbf{e} = (I - M)\mathbf{e}$, and

$$\sum_{i=1}^{n+1} iM^{i-1}(I - M) = (I - M^{n+1})(I - M)^{-1} - (n+1)M^{n+1},$$

the stated formula follows by routine calculations.

Services are either initiated singly or in pairs. Two services are stated at a transition from the macro-state **0** or after a blocked system is cleared. The rate $\lambda^*(n)$ is therefore given by

$$\lambda^*(n) = 2\boldsymbol{\pi}(0)\mathbf{S}^0(1) + \sum_{i=1}^{n} \boldsymbol{\pi}(i)[\mathbf{S}^0(1) \otimes \mathbf{e}] + 2\boldsymbol{\pi}(n+2)\mathbf{S}^0(2).$$

Routine calculations show that $\lambda^*(n) = \mu^*(n)$. □

For the exponential case,

$$\mu^*(n) = (\mu_1 + \mu_2)(2p - p^2 - p^{n+2})(1 - p^{n+1} + p^{n+2})^{-1},$$

which agrees with the right-hand side in the equilibrium condition (5.3.5) in [5].

Other Steady-State Descriptors

As for most other models analyzed by matrix-analytic methods, there are many useful descriptors of the steady-state regime of the service mechanism that are readily expressed once we have the stationary probability vector $\boldsymbol{\pi}$. The following is a selection of such descriptors.

The *fraction $p_2^*(n)$ of jobs served by Unit II*, is given by the ratio

$$p_2^*(n) = k^*[\mu^*(n)]^{-1}. \tag{19}$$

A typical *idle time of Server II* has the *PH*-distribution with representations $[\boldsymbol{\gamma}, S(1)]$, where the vector $\boldsymbol{\gamma}$ is given by

$$\boldsymbol{\gamma} = [\boldsymbol{\pi}(0)\mathbf{e}]^{-1}\boldsymbol{\pi}(0),$$

and its mean equals $\mu_1'(2_{idle}, n) = [\boldsymbol{\pi}(0)\mathbf{e}]^{-1}\boldsymbol{\pi}(0)[-S(1)]^{-1}\mathbf{e}$. In addition, the mean time between an arbitrary visit to the macro-state $\mathbf{0}$ and the next is just $[\boldsymbol{\pi}(0)\mathbf{e}]^{-1}$. The *fraction of time that Server II is idle* is given by

$$p_2(2_{idle}, n) = \boldsymbol{\pi}(0)[-S(1)]^{-1}\mathbf{e}, \tag{20}$$

a descriptor of interest in quantifying the effect of the buffer size on the performance of the system.

The point processes of the departure epochs and of the epochs of initiations of services are both *BMAP*s. The *BMAP* is a Markovian point process for which many descriptors can be expressed in explicit matrix formulas. From its extensive literature, we refer to Lucantoni [2, 3], Neuts [7], and Narayana and Neuts [4] for discussions of the basic definitions and properties of *BMAP*s.

We note that the order of the coefficient matrices of these *BMAP*s is the same as that of Q_1. The matrix formulas for most descriptors involve matrices of that order. The principal challenge therefore lies in the efficient organization of numerical computations with large but highly structured matrices.

From the present results, we immediately obtain expressions for various rates. The rate of *single initiations of service* is $\sum_{i=1}^{n}\boldsymbol{\pi}(i)[\mathbf{S}^0(1)\otimes\mathbf{e}]$ while *pairs of services* start at rate k^*.

Jobs leave the system in groups of sizes ranging from 1 to $n+2$. The steady-state departure rate $\mu^*(i;n)$ of a group of size i is given by

$$\mu^*(1;n) = \boldsymbol{\pi}(0)\mathbf{S}^0(1) + \boldsymbol{\pi}(1)[\mathbf{e}\otimes\mathbf{S}^0(2)],$$

$$\mu^*(i;n) = \boldsymbol{\pi}(i)[\mathbf{e}\otimes\mathbf{S}^0(2)]$$

for $2 \le i \le n+1$, and

$$\mu^*(n+2;n) = \boldsymbol{\pi}(n+2)\mathbf{S}^0(2).$$

The sum $\mu^{**}(n)$ of these quantities is the rate at which departures occur regardless of the group size. By using formulas (15) and (4), we readily obtain that

$$\mu^{**}(n) = k^*[2 - \beta(2)M^{n+1}\mathbf{e}].$$

The probability that an arbitrary departure consists of i customers is therefore $\mu^*(i;n)[\mu^{**}(n)]^{-1}$, for $1 \le i \le n+2$. The mean group size is $\mu^*(n)[\mu^{**}(n)]^{-1}$.

The Matrices *M* and *N*

These matrices arise naturally in studying the super-position of two independent *PH*-

renewal processes; they were introduced in Latouche and Neuts [1]. Consider the product Markov chain of the familiar Markov chains with generators $S(1)+\mathbf{S}^0(1)\beta(1)$ and $S(2)+\mathbf{S}^0(2)\beta(2)$, used in discussing the PH-renewal process.

The element $M_{j,j'}$ of M is the conditional probability that, given that we start at a renewal in the first process and with the second in phase j, the next renewal is also in the first process and when it occurs, the phase in the second process is j'. Similar ly, the element $N_{j,h'}$ of N is the conditional probability that, given that we start at a renewal in the first process and with the second in phase j, the next renewal is in the second process and when it occurs, the phase in the first process is h'.

Most constants arising as descriptors of the present models can be interpreted as expected values of random variables associated with the super-position of two independent PH-renewal processes. We given an illustration in the sequel.

3.1 Model I via Markov Renewal Theory

When, as in the case here, models based on phase-type or MAP analysis are unusually tractable, it is often possible to obtain slightly more general results by studying an embedded Markov renewal process. The succinct discussion that follows serves to illustrate that point.

We consider Model I at the ends of successive services in Unit II. We allow the service times in that unit to have a general probability distribution $F_2(\cdot)$. The service times in Unit I have the PH-distribution with representation $[\beta(1), S(1)]$. When a service in II ends while Server I is occupied, we include the phase of that service in the state description. The m_1 corresponding states form the macro-state $\mathbf{0}$. The single state 1 signifies a transition epoch in which new services simultaneously start in both units. It is readily seen that we obtain a Markov renewal process with m_1+1 states. We write its transition probability matrix $K(\cdot)$ in the partitioned form:

$$K(x) = \begin{array}{c} \mathbf{0} \\ 1 \end{array} \begin{pmatrix} K(\mathbf{0},\mathbf{0},x) & K(\mathbf{0},1,x) \\ K(,\mathbf{0},x) & K(1,1,x) \end{pmatrix}.$$

The elements of $K(x)$ are obtained by routine conditioning arguments. To write them concisely, we need the standard matrices $P(\nu,t)$ for the counting process of the PH-renewal process with the underlying distribution $F_1(\cdot)$. To remind us that these correspond to the service time distribution in Unit I, we add the subscript 1.

The matrix $K(\mathbf{0},\mathbf{0},x)$ is given by

$$K(\mathbf{0},\mathbf{0},x) = \int_0^x \int_0^y \exp[S(1)u]\mathbf{S}^0(1)du\beta(1)\sum_{r=0}^{n} P_1(r;y-u)dF_2(y-u),$$

for $x \geq 0$. To obtain that expression, we condition on the end of the service in course in Unit I and on the number of additional service time terminations in I during the service by II. There can be at most n such job completions; otherwise, blocking occurs.

There are similar expressions for the other elements of $K(x)$, but for the sake of brevity, we only give the expressions for their Laplace-Stieltjes transforms which facilitate the subsequent calculations. We write

$$A_1(r;s) = \int_0^\infty e^{-sy} P_1(r;y)dF_2(y), \tag{21}$$

for $r \geq 0$. These are the same matrices that arise in the analysis of the $PH/G/1$ queue.

The Laplace-Stieltjes transforms of the elements of $K(\cdot)$ are given by:

$$K^*(\mathbf{0},\mathbf{0},s) = [sI - S(1)]^{-1}\mathbf{S}^0(1)\beta(1)\sum_{r=0}^{n} A_1(r;s), \tag{22}$$

$$K^*(\mathbf{0},1,s) = [sI - S(1)]^{-1}\mathbf{S}^0(1)\beta(1)[f_2(s) - \sum_{r=0}^{n} A_1(r;s)\mathbf{e}], \tag{23}$$

$$K^*(1,\mathbf{0},s) = \beta(1)\sum_{r=0}^{n} A_1(r;s), \tag{24}$$

and

$$K^*(1,1,s) = f_2(s) - \beta(1)\sum_{r=0}^{n} A_1(r;s)\mathbf{e}. \tag{25}$$

We find that the row sums of the transform matrix are given by

$$K^*(\mathbf{0},\mathbf{0},s)\mathbf{e} + K^*(\mathbf{0},1,s) = f_2(s)[sI - S(1)]^{-1}\mathbf{S}^0(1), \tag{26}$$

and

$$K^*(1,\mathbf{0},s)\mathbf{e} + K^*(1,1,s) = f_2(s). \tag{27}$$

We shall use a symbol with the variable s suppressed to denote which its value for $s = 0$.

We do not present a detailed analysis of this embedded Markov renewal process as all explicitly tractable results agree with those derived by the approach using PH-distributions. We give a brief derivation of its *fundamental mean* E whose inverse is the steady-state rate of transitions, that is, ends of services in Unit II.

Theorem 3.3: *The fundamental mean E of the embedded Markov renewal process is given by*

$$E = \mu_1'(2) + \sum_{r=0}^{n} \beta(1)A_1(r)[-S(1)]^{-1}\mathbf{e}. \tag{28}$$

Proof: Let $(\boldsymbol{\phi}, \phi')$ be the (partitioned) invariant probability vector of the stochastic matrix $K = K(\infty)$; then an elementary calculation shows that

$$\boldsymbol{\phi} = \sum_{r=0}^{n} \beta(1)A_1(r), \qquad \phi' = 1 - \sum_{r=0}^{n} \beta(1)A_1(r)\mathbf{e}.$$

The vector $(\mathbf{b}, b')$ of the row sum means of the transition probability matrix is obtained by differentiating with respect to s in the equations (26) and (27). We find that

$$\mathbf{b} = \mu_1'(2)\mathbf{e} + [-S(1)]^{-1}\mathbf{e}, \qquad b' = \mu_1'(2).$$

By evaluating the inner product $\phi\mathbf{b}+\phi'b'$ we obtain formula (28). □

When the probability distribution $F_2(\cdot)$ is of phase type, the matrices $A_1(r)$ are given by $\beta(1)A_1(r)=\beta(2)M^rN$, for $r\geq 0$, so that the expressions obtained here agree with those obtained earlier.

The pervasiveness of lower order moments of the service and interarrival times in the equilibrium conditions and mean values of the descriptors of the classical queues can be deceptive. The quantities that arise in the corresponding descriptors of more complex queueing models are rarely insensitive. To enhance our understanding of the physics of such models, it is useful to have interpretations of various constants.

We shall not do this for all the constants we have encountered so far, but for the sake of an illustration we interpret the quantity

$$\kappa(n)=\sum_{r=0}^{n}\beta(1)A_1(r)[-S(1)]^{-1}\mathbf{e}, \tag{29}$$

in the formula for the fundamental mean E.

Consider two independent renewal processes with underlying probability distributions $F_1(\cdot)$ and $F_2(\cdot)$ that are continuous at 0. The sequences of successive renewal epochs (called, respectively, 1-*renewals* and 2-*renewals*) are denoted by $\{S_{1,\nu}\}$ and $\{S_{2,\nu}\}$ respectively. We agree that $S_{1,0}=S_{2,0}=0$, so that time 0 corresponds to an event in both renewal processes.

We define the event $\{S_{1,i}\leq S_{2,1}<S_{1,i+1}\}=C_i$ that there are exactly i events in the first renewal process prior to the first event in the second one. The summands in formula (29) are the expectations

$$E[(S_{1,i+1}-S_{2,1})I(C_i)],$$

when $F_1(\cdot)$ is of phase type. $I(C_i)$ is the indicator function of C_i. Informally, therefore, κ is the expected time to the first 1-renewal following the first 2-renewal, evaluated over the event where there are at most n 1-renewals prior to that first 2-renewal.

When there is positive probability of coincident renewals, κ may not be a continuous functional of $F_1(\cdot)$ and $F_2(\cdot)$. Application of the continuity argument in Section 5.1, p. 243 of Neuts [6] requires some caution. However, under mild restrictions - in the present case, that $F_2(\cdot)$ can be continuous - the continuity argument can be invoked. The interpretation of the fundamental mean E is then also valid for general (continuous) distributions.

Several other constants in this paper may be given similar interpretations. By the same argument and with the same proviso, these interpretations hold for general distributions.

4. Discussion of Model II

For Model II, we keep track of the number of jobs in the service system what are currently allocated to Server II. If that number is positive, we need to distinguish the cases where Server I is active or is blocked. The state space of the Markov chain for Model II therefore consists of $2m+1$ macro-states denoted by $\mathbf{0}$, $\mathbf{1},\ldots,\mathbf{m}$ and $\mathbf{1}^*,\ldots,\mathbf{m}^*$, where the asterisk indicates that Server I is blocked. The macro-state $\mathbf{0}$ corresponds to the case where Unit II is empty. It contains the m_1 labels of the phases for the service time distribution $F_1(\cdot)$. The macro-states $\mathbf{1},\ldots,\mathbf{m}$ contain the m_1m_2 phases of the two service times in course. The corresponding macro-states with an asterisk contain the phases of the service in course in Unit II. Server I has then completed a job but is blocked so we need not keep track of a service phase.

The infinitesimal generator Q_2 for Model II is displayed for $m = 3$. The structure for a general value of m is similar.

$$Q_2 =$$

$$\begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 1^* \\ 2^* \\ 3^* \end{array} \begin{pmatrix} S(1) & 0 & 0 & \mathbf{S}^0(1)\beta(1)\otimes\beta(2) & 0 & 0 & 0 \\ I\otimes\mathbf{S}^0(2) & S(1)\oplus S(2) & 0 & 0 & \mathbf{S}^0(1)\otimes I & 0 & 0 \\ 0 & I\otimes\mathbf{S}^0(2)\beta(2) & S(1)\oplus S(2) & 0 & 0 & \mathbf{S}^0(1)\otimes I & 0 \\ 0 & 0 & I\otimes\mathbf{S}^0(2)\beta(2) & S(1)\oplus S(2) & 0 & 0 & \mathbf{S}^0(1)\otimes I \\ 0 & 0 & 0 & \beta(1)\otimes\mathbf{S}^0(2)\beta(2) & S(2) & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{S}^0(2)\beta(2) & S(2) & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{S}^0(2)\beta(2) & S(2) \end{pmatrix}$$

As for Model I, two matrices M_1 and N_1 appear in the explicit matrix formulas for the steady-state probabilities. These matrices are defined as follows: The matrix M_1 is $m_1 \times m_1$ and

$$M_1 = [I\otimes\beta(2)][-S(1)\oplus S(2)]^{-1}[I\otimes\mathbf{S}^0(2)]. \tag{30}$$

The matrix N_1 is $m_1 \times m_2$ and

$$N_1 = [I\otimes\beta(2)][-S(1)\oplus S(2)]^{-1}[\mathbf{S}^0(1)\otimes I]. \tag{31}$$

The matrix M_1 is irreducible and sub-stochastic. The matrix N_1 is nonnegative and $M_1\mathbf{e} + N_1\mathbf{e} = \mathbf{e}$.

The stationary probability vector $\boldsymbol{\pi}$ of Q_2 is partitioned according to the macro-states as

$$\boldsymbol{\pi}(0), \boldsymbol{\pi}(1),\ldots, \boldsymbol{\pi}(m), \boldsymbol{\pi}(1^*),\ldots, \boldsymbol{\pi}(m^*).$$

Theorem 4.1: *For Model II, the vectors in the partitioned form of* $\boldsymbol{\pi}$ *are given by*

$$\boldsymbol{\pi}(0) = h^*\beta(1)M_1^m[-S(1)]^{-1}, \tag{32}$$

and

$$\boldsymbol{\pi}(i) = h^*\beta(1)M_1^{m-i}[I\otimes\beta(2)][-S(1)\oplus S(2)]^{-1} \tag{33}$$

and

$$\boldsymbol{\pi}(i^*) = h^*\beta(1)M_1^{m-i}N_1[-S(2)]^{-1} + h^*[1-\beta(1)M_1^{m-i}\mathbf{e}]\cdot\beta(2)[-S(2)]^{-1} \tag{34}$$

for $1 \le i \le m$. *The normalizing constant* h^* *is given by*

$$[h^*]^{-1} = m\mu_1'(2) + \beta(1)M_1^m[-S(1)]^{-1}\mathbf{e}, \tag{35}$$

where $\mu_1'(2)$ *is the mean service time in Unit II.*

Proof: The required calculations are similar to those for Model I. We present only the essential steps. The steady-state equations are as follows:

$$\boldsymbol{\pi}(0)S(1)+\boldsymbol{\pi}(1)[I\otimes \mathbf{S}^0(2)]=\mathbf{0}, \tag{36}$$

and for $1\leq i\leq m-1$,

$$\boldsymbol{\pi}(i)[S(1)\oplus S(2)]+\boldsymbol{\pi}(i+1)[I\otimes \mathbf{S}^0(2)\beta(2)]=\mathbf{0} \tag{37}$$

$$\boldsymbol{\pi}(0)[\mathbf{S}^0(1)\beta(1)\otimes \beta(2)]+\boldsymbol{\pi}(m)[S(1)\oplus S(2)] + \boldsymbol{\pi}(1^*)[\beta(1)\otimes \mathbf{S}^0(2)\beta(2)]=\mathbf{0}, \tag{38}$$

and for $1\leq i\leq m-1$

$$\boldsymbol{\pi}(i)[\mathbf{S}^0(1)\otimes I]+\boldsymbol{\pi}(i^*)S(2)+\boldsymbol{\pi}(i+1^*)\mathbf{S}^0(2)\beta(2)=\mathbf{0} \tag{39}$$

$$\boldsymbol{\pi}(m)[\mathbf{S}^0(1)\otimes I]+\boldsymbol{\pi}(m^*)S(2)=\mathbf{0}. \tag{40}$$

Defining h^* by

$$h^*=\boldsymbol{\pi}(0)\mathbf{S}^0(1)+\boldsymbol{\pi}(1^*)\mathbf{S}^0(2), \tag{41}$$

and proceeding as in Theorem 3.1, we get that $\boldsymbol{\pi}(i)$ for $1\leq i\leq m$ is as given by formula (33). That immediately gives the expressions for $\boldsymbol{\pi}(0)$ and $\boldsymbol{\pi}(m^*)$. Working backwards from that last formula, we obtain (34).

The constant h^* is obtained by using the normalizing equation. However, to get the simple expression in (35) we need the following equality that results in major cancellations:

$$[I\otimes \beta(2)][-S(1)\oplus S(2)]^{-1}(\mathbf{e}\otimes\mathbf{e})=\mathbf{e}\mu_1'(2)-N_1[-S(2)]^{-1}\mathbf{e}.$$

That equality is established by the same manipulations involving the formulas (16) and (17). □

By a direct calculation, we can verify that h^* is the steady-state departure rate of customers processed by Server I. The second term in formula (35) is the mean time that an arbitrary customer served by I is delayed by jobs in progress in Unit II.

The Exponential Case

Replacing the general PH-representations for the service time distributions by -μ_1 and -μ_2, we obtain formulas for the stationary probability vector for the exponential case. These are written most concisely in terms of the ratio $\xi=\mu_1/\mu_2$ and read as follows:

$$\pi(0)=[1+m\xi(1+\xi)^m]^{-1},$$

and

$$\pi(i)=\xi(1+\xi)^i[1+m\xi(1+\xi)^m]^{-1},$$

$$\pi(i^*)=\xi[(1+\xi)^m-(1+\xi)^i][1+m\xi(1+\xi)^m]^{-1}$$

for $1\leq i\leq m$. The constant h^* is given by $h^*=\xi(1+\xi)^m[1+m\xi(1+\xi)^m]^{-1}\mu_2$.

The Departure Rate

Customers leave the system singly or in pairs. A pair leaves when the end of a service in II also releases a customer blocked in Server I.

Theorem 4.2: *The steady-state rates $\mu^*(m)$ of departures and $\lambda^*(m)$ of service initiations are given by*

$$\mu^*(m) = \lambda^*(m) = (m+1)h^*. \tag{42}$$

Proof: By substitution and simplification in the expressions

$$\mu^*(m) = \boldsymbol{\pi}(0)\mathbf{S}^0(1) + \sum_{i=1}^{m} \boldsymbol{\pi}(i)[\mathbf{e} \otimes \mathbf{S}^0(2)] + 2\boldsymbol{\pi}(1^*)\mathbf{S}^0(2) + \sum_{i=2}^{m} \boldsymbol{\pi}(i^*)\mathbf{S}^0(2)$$

and

$$\lambda^*(m) = (m+1)\boldsymbol{\pi}(0)\mathbf{S}^0(1) + (m+1)\boldsymbol{\pi}(1^*)\mathbf{S}^0(2) = (m+1)h^*. \tag{43}$$

□

In Neuts [5] it is shown that, with exponential servers of unequal rates, the throughput in Model I is largest with the faster server is assigned to Unit I. For Model II and with the same m, the opposite conclusion holds.

To see this, we let ξ be less than one, so that the second server is the faster of the two. The departure rate $\mu^*(m)$ is then given by

$$\mu^*(m) = (m+1)\xi(1+\xi)^m[1 + m\xi(1+\xi)^m]^{-1}\mu_2.$$

If we interchange the role of the two servers, the throughput $\mu^{**}(m)$ is obtained by replacing ξ by ξ^{-1} and the final μ_2 by μ_1.

Now,

$$[\mu^{**}(m)/\mu^*(m) - 1][\xi^{m+1} + m(1+\xi)^m] = 1 - \xi^{m+1} - m(1-\xi)(1+\xi)^m, \tag{44}$$

and for $0 < \xi < 1$, the right-hand side is negative.

Given the complex dependence on the service time distributions it is by no means obvious that the same conclusion holds for non-exponential distributions. For *PH*-distributions, a numerical exploration of that question is easy.

A further interesting question is to find the value of m for which the throughput is largest. For the exponential case, that is elementary; for fixed ξ, we find the m for which the function in (43) attains its maximum. For *PH*-distributions, the numerical exploration is more substantial.

5. Some General Methodological Comments

While the two service systems in this paper may be of limited practical applicability, their analysis is similar to that required for other tandem queues and for the complex systems arising in manufacturing and telecommunications.

We have demonstrated how the formulation with *PH*-distributions can provide structural insights that are lost under exponential assumptions. That formulation also makes clear how much - or more commonly, how little - analytic tractability of the model is to be expected. The models in this paper are highly tractable. For others that are less, the computational effort required to study large, structured Markov chains becomes clear. Such an effort is worthwhile; it can be methodologically challenging and its results provide benchmarks for the more common simulation studies.

The throughput analysis is an important first step in studying complex queueing systems where the input process is independent of the service mechanism. Both models in this paper, even with a *BMAP* arrival process, lead to Markov chains of $GI/M/1$ type. Provided that the fundamental rate of arrivals is less than the departure rate of the saturated system, these Markov chains have a (modified) matrix-geometric stationary probability vector. The definition of the boundary matrices is a belabored task which we have avoided here. With realistic choices of the service time distributions and even for Poisson arrivals. The order of the rate matrix R of the matrix-geometric solution is so high that its evaluation requires a massive, if not infeasible numerical computation.

When an approach under *PH*-assumptions is inadequate, there remains the alternative of an empirical study by computer experimentation. Such a study is much more challenging than running a few simulations may suggest. We propose to investigate that approach, in combination with a *PH*-analysis, for other models of greater applied interest.

Acknowledgements

The research of R.E. Lillo was partially supported by a scholarship from the Carlos III Foundation; that of M.F. Neuts by NSF Grant Nr. DMI-9306828.

References

[1] Latouche, G. and Neuts, M.F., The super-position of two *PH*-renewal processes In: *Semi-Markov Models: Theory and Applications* (ed. by J. Janssen), Plenum Press, New York (1986), 131-177.

[2] Lucantoni, D.M., New results on the single server queue with a batch Markovian arrival process, *Commun. in Statistics: Stochastic Models* **7**:1 (1991), 1-46.

[3] Lucantoni, D.M., The $BMAP/G/1$ queue: A tutorial, In: *Performance Evaluation of Comp. and Commun. Systems: Joint Tutorial papers of Perf. '93 and Sigmetrics '93* (ed. by L. Donatiello and R. Nelson), Springer-Verlag, Berlin (1993), 330-358.

[4] Narayan, S. and Neuts, M.F., The first two moments matrices of the counts for the Markovian arrival process, *Comm. Statist. Stoch. Models* **8**:3 (1992), 459-477.

[5] Neuts, M.F., *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press 1981. Corrected reprint: Dover Publications, Inc., New York 1994.

[6] Neuts, M.F., *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, Marcel Dekker, New York 1989.

[7] Neuts, M.F., Models based on the Markovian arrival processes, *IEICE Trans. on Commun.* **E75-B**:12 (1992), 1255-1265.

[8] Neuts, M.F., Some promising directions in algorithmic probability, In: *Adv. in Matrix Analytic Methods for Stoch. Models* (ed. by A.S. Alfa and S.R. Chakravarthy), New Jersey (1998), 429-443.