# Estimating Characteristics of Queueing Networks Using Transactional Data

Avi Mandelbaum
Faculty of Industrial Engineering and Management
Technion
Haifa 32000, Israel

Sergey Zeltyn
Faculty of Industrial Engineering and Management
Technion
Haifa 32000, Israel

January 21, 1997

Revised December 14, 1997

**Abstract.** We are motivated by queueing networks in which queues are difficult to observe but services are easy to record. Our goal is to estimate the queues from service data. More specifically, we consider an open queueing network with Poisson external arrivals, multi-server stations, general service times and Markovian switches of customers between stations. Customers' transitions between stations may be either immediate or of exponentially distributed durations. Each customer is supplied with an Identification Number (ID) upon entering the network. Operational data is collected which includes transaction times (starts and terminations of services) and ID's of served customers. Our objective is to estimate the evolution of the queues in the network, given the collected data. We cover estimation at both end of busy periods and in real time. The applicability of the theory is demonstrated by analyzing a service operation.

**Key words.** Queues, Queueing Networks, Non-parametric Inference, Data Analysis, Hidden Markov Models, Performance Evaluation, Queueing Inference Engine.

**AMS 1991 Subject Classification.** Primary 60G17, 60J20, 60K25, 60K30, 62G05, 62M05; secondary 68M20, 90B22.

**Running Head.** Estimating Characteristics of Queueing Networks.

**Acknowledgments.**
A.M. was supported by the Fund for Promotion of Research at the Technion.
S.Z. acknowledges the support of the Technion, Miriam Gutwirth and Aaron Gutwirth.

# Contents

Frequently used notation.

**1. General.**

$\mathcal{D} = \{1, \ldots, d\}$ — set of network stations.

$\mathcal{D}_0 = \{0, 1, \ldots, d\}$ — extended set of stations including station 0 (network entrance).

$\alpha(t)$ — non-homogeneous rate of external arrivals.

$\lambda$ — homogeneous rate of external arrivals.

$P = [p_{jk}],\ j \in \mathcal{D}_0,\ k \in \mathcal{D}$ — matrix of routing probabilities.

$Q_j(t)$ — queue length at station $j$ at time $t$.

$E_t$ — observed event at time $t$ (described and defined in §1.5, 3.1).

$\mathcal{G}_i$ — set of admissible states of a Markov chain (defined in §2.2).

$A \Rightarrow B$ — event $A^c \bigcup B$.

**2. Single busy period at a given station.**

$t_0, t_1, \ldots, t_n, t_{n+1}$ — transaction times during a busy period.

$A_0, A_1, \ldots, A_n$ — arrival times of customers during a busy period.

$A(t)$ — cumulative number of arrivals until $t$; $\hat{A}(t) = \mathrm{E}[A(t) \mid \bullet]$ (§2.2).

**3. Queueing network. Times and Transitions.**

Capital and Greek letters denote random variables.

Small letters are used for their realizations.

$T^j_{ki} = t^j_{ki}$ — time of $i$-th transaction during busy period number $k$ at station $j$.

$A^j_{ki}$ — arrival time to station $j$ of a customer that started service at $t^j_{ki}$.

$\bar{A}^n_{ml}$ — arrival time to his next station of a customer that terminated service at $t^n_{ml}$.

$A^0_l$ — time of $l$-th external arrival to station 0.

$A^{j0}_l$ — time of $l$-th external arrival to station $j$.

$\xi^j_{ki}$ — duration of $i$-th service during busy period number $k$ at station $j$.

$R^n_{ml}$ — station which the customer that terminated service at $t^n_{ml}$ switched to.

**4. Queueing network. Counting and Routing.**

$\mathcal{D}_b(t) = d_b(t)$ — set of busy stations at time $t$ (§3.1).

$B_j(t) = b_j(t)$ — number of busy periods which started at station $j$ up to time $t$.

$M_j(t) = m_j(t)$ — number of service starts before $t$ during the current busy period at station $j$.

$\mathcal{R}(n, m, l)$ — Routing Operator (§3.1).

$\mathcal{R}^{-1}(j, k, i)$ — Backward Routing Operator (§3.1).

$V_R(t)$ — Visible-Routing times (§3.2).

$I_R(t)$ — Invisible-Routing times (§3.2).

$C_j(t)$ — arrival time to station $j$ of the customer served there at $t$ (§4.2) or arrival time to station $j$ of the last internal customer (§4.3);

$\mathcal{M}_t(t^n_{ml})$ — available stations process (§4.2).

# 1 Introduction

## 1.1 Motivation

We are interested in queueing systems with the following two features:

- *invisible queues*, namely queues of which direct measurements are either impossible or too expensive to collect;
- *transactional data* (times of service starts and service terminations) is relatively easy to access.

The goal is to estimate the invisible queues from transactional data. Examples of such systems are Automatic Teller Machines (ATM) in a bank (transactions are insertions and ejections of customers' ATM cards), telephone service centers (transactions are starts and terminations of calls), transportation queues at traffic intersections (transactions correspond to times when vehicles cross a measuring cable) and face-to-face services such as banks, government offices, health clinics, and so on.

Note that some applications constitute single-station queues (ATM, for example) but others are more naturally described within a queueing network context (traffic intersections, some face-to-face services in a bank). Previous research covered the single-station case exclusively, and the contribution here is an extension to a network setting.

## 1.2 Survey of Literature

The history of the subject is short, hence an attempt is made to survey all existing literature.

Larson opened up the subject in his pioneering paper [13], where he introduced the terminology Queueing Inference Engine (QIE) to describe the algorithm that infers queue length from transactional data.

Larson considered a service station with several servers working in parallel, Poisson arrivals and general distribution of service times. The service mechanism is truly general. In particular, it generalizes the classical $M/G/s$ in that servers need not be statistically identical and service times are not required to be independent or identically distributed. (However, an assumption of the independence between services and arrivals must be imposed.)

A simple key observation in [13] was that busy periods (time periods when all servers are working) coincide with the time intervals over which new services begin instantaneously after service termination times (in practice — almost immediately). Larson's objective was to estimate the evolution of the queue length over a specified busy period, given the data (transactions) accumulated over that period. The estimation is carried out at the end of the period. Notably, the algorithm does not require knowledge of the arrival rate $\lambda$. The original algorithm [13] required $O(n^5)$ calculations, where $n$ is the number of transactions during the busy period. In fact, a simple modification [14] requires $O(n^3)$ calculations.

S.A.Hall [10] incorporates additional information. Specifically, one is informed each time that the queue size hits 0 or a predetermined upper limit $M$. Hall also calculated additional statistics, such as the density function of the arrival time for the $k$-th customer.

Jones and Larson [12] derive results concerning order statistics and apply them to calculate conditional distributions of queue characteristics (for example, the conditional probability that a random customer, served during the busy period, waited in queue less then $\tau$ minutes).

An alternative approach was developed by Daley and Servi. In [3] they used the technique of taboo probabilities for Markov chains to solve Larson's problem and additional problems of this type. For example, algorithms for $E_k/G/1$ (interarrival times are independent with Erlang-$k$ distribution) and $M/G/s/m$ (finite buffer) systems were analyzed. (In every case, except $M/G/s$, the arrival rate $\lambda$ must be known.) An approximate $O(n^2 \ln n)$ algorithm was also developed.

In [4] Daley and Servi found that many models can be accommodated within the general framework of Markov Chain Boundary Value problems which, in turn, can be solved using taboo probabilities. A single algorithm was developed and applied to models such as queues with reneging (a customer leaves if the wait is too long), and balking (any arriving customer does not join with probability $p$ whenever the queue size is at or above a threshold $M$).

Markov Chain Boundary Value problems are part of an extensive class of models which are called Hidden Markov Models (HMM). The problem addressed in these models is the inference of different characteristics of a Markov chain, given partial information about its state. See [8] or [16] for an account on HMM methods.

In [5] asymptotic conditional mean queue length was studied for busy periods with a large number of customers. Brownian excursions were used for the approximations.

Several important ideas and results were introduced in Bertsimas and Servi [2]. First, an algorithm for queue estimation in real-time (within a busy period) was derived. (A real-time algorithm was also presented in [3].) Other results from [2] included an original $O(n^3)$ algorithm for the $M/G/s$ problem (the first one submitted for publication), an extension to $M_t/G/s$ (after a simple time-change, the initial $M/G/s$ algorithm pertains; the arrival rate $\alpha(t)$ must be a priori known), and a solution for the $GI/G/s$ problem (formulae include multidimensional integrals and convolutions that are, in general, non-computable).

An application of Larson's QIE algorithm [13] to real data, originating from telecommunications, was made in Gawlick [9]. Satisfactory results were reported, despite the fact that arrivals were not exactly Poisson.

## 1.3 Network Description

Consider an open queueing network with $d$ stations, denoted as a set by $\mathcal{D} = \{1, 2, \ldots, d\}$. Prevailing assumptions about the network are as follows:

- External arrivals to the network constitute a Poisson process with a non-homogeneous arrival rate $\alpha = \{\alpha(t), t \geq 0\}$.
- Switches of customers between stations are governed by a $d$-dimensional routing matrix $P = [p_{jk},\ j, k \in \mathcal{D}]$.
- Service times $\xi_{ji},\ j \in \mathcal{D},\ 1 \leq i < \infty$ are completely general random variables (not necessarily independent or identically distributed).
- The stochastic components of the network, namely external arrivals, services and switches between stations, are independent of each other.
- Stations may be multi-server stations.
- Queue-size capacity at all stations is infinite.
- First Come First Served (FCFS) queue discipline is assumed at all stations.
- Work-conserving principle applies: a server cannot be idle if there is a customer waiting for service.
- An *Identification Number* (ID) is attached to each customer upon entering the network, and it accompanies him until exit.
- No simultaneous arrivals to stations take place.

In specific problems, some of these assumptions will be weakened or even omitted. The last technical assumption is used because, in the case of simultaneous arrivals, FCFS does not determine the service order.

Available data includes transaction times and ID's of corresponding customers. It will become clear later that the availability of ID's and the FCFS queue discipline play a key role in our approach. (Both assumption are not relevant in the single-station case.) Moreover, this information improves considerably the queue estimates, as demonstrated in the applications of Section 6.

Two different models are treated:

- Transitions between stations are immediate.
- Transition times between stations are exponentially distributed with parameters $\eta_{jk},\ j, k \in \mathcal{D}$, independently of the other stochastic components of the network.

We now elaborate on several features of the network. Most often we assume that the network has a single entrance, which is conveniently referred to as "station 0". (The extended set of stations $\{0, 1, \ldots, d\}$ is denoted by $\mathcal{D}_0$). Thus, at time $t$, new customers arrive to station 0 at a rate $\alpha(t)$, after which they switch to other stations according to the routing probabilities $p_{0j}$, $j \in \mathcal{D}$ (and transition rates $\eta_{0j}$, if transition times are exponential).

It is reasonable to assume sometimes that a measuring device is installed at the entrance of the network (see Section 6). Therefore, two cases are considered.

- External arrivals are *known* (registered).
- External arrivals are *unknown* (not registered).

As mentioned above, external arrivals constitute a non-homogeneous Poisson process. However, it is worthwhile to consider separately external arrivals with a *constant* rate $\lambda$, mostly because in several special cases the rate value can be assumed unknown. The solution for the non-homogeneous case is derived using a simple time-change (see discussion after Algorithm 2.1 in §2.2).

It turns out that if external arrivals are known (registered), we do need to impose on them any assumptions, except their independence of the other stochastic components.

Non-external arrivals to specific station from other stations of the network (except station 0) will be called *internal arrivals.* Sometimes we also use the terminology of "internal (external) customer" for customers that correspond to internal (external) arrivals.

## 1.4 Objectives

We define the *observed information* $\mathcal{E} = \{\mathcal{E}_t, t \geq 0\}$ as the filtration generated by observed events: $\mathcal{E}_t$ is the $\sigma$-field that models the information accumulated during the interval $[0, t]$. (Rigorous description of an observed event will be given in §3.1.) Introduce the right-continuous queue length process

$$Q = \{Q_t,\, t \geq 0\}; \quad Q_t = (Q_1(t), \ldots, Q_d(t)),$$

where $Q_j(t)$ is the number of customers waiting for service at station $j$ at time $t$. Our objective is to estimate the unobserved queue process $Q$, given the observed information $\mathcal{E}$. Three different estimation problems are posed:

**Busy-Period Interpolation** (Section 2) An interval $[T_1, T_2]$ is a *busy period* for station $j$ if *all* the servers at station $j$ are busy during that interval but just before $T_1$ and just after $T_2$ some servers are idle.

The busy-period interpolation problem is to estimate the sample path $Q_j(t)$, $t \in [T_1, T_2]$, based on the observed information $\mathcal{E}_T$ at a fixed time $T \geq T_2$,

**Real-Time Estimation** (Section 4) Estimate $Q_t$, based on the observed information $\mathcal{E}_t,\ t \geq 0$.

**General Interpolation** (Section 5) Estimate the sample path $Q_s,\ 0 \leq s \leq t$, based on the observed information $\mathcal{E}_t,\ t \geq 0$.

To summarize, we deal with 12 cases, distinguished by three estimation objectives, transition times (immediate or exponential) and external arrivals (known or unknown): all in all $3 \times 2 \times 2 = 12$ models.

## 1.5 Summary of Results

In Table 1 we introduce briefly our results for the twelve special cases mentioned above. These are covered in Sections 2, 4 and 5. Section 3 provides formal prerequisites for the complicated models of Sections 4 and 5. In Section 6 we apply some of our algorithms to real data, drawn from a branch of a commercial bank. We conclude, in Section 7, with some directions worthy of future research.

| | Immediate Transitions | | Exponential Transitions | |
|---|---|---|---|---|
| | External Arrivals | | | |
| | Known | Unknown | Known | Unknown |
| Busy-Period Interpolation | Queue length computable exactly (§2.1) | Adaptation of Daley and Servi [4] HMM algorithm (§2.2) | HMM; Properties of truncated exponential random variables (§2.3, §2.4) | HMM; truncated exponentials; properties of $M_t/M/1$ queue §2.5, [19] |
| Assumptions | None. | FCFS; Poisson external arrivals. | FCFS; transition rates $\eta_{ij}$. | FCFS; Poisson external arrivals; $\eta_{ij}$. |
| Real-time Estimation. General Interpolation | Simple combinatorics (§4.2, §5.2) | New algorithm; uses Bertsimas and Servi [2] real-time algorithm (§4.3, §5.3) | HMM with cumbersome state space; truncated exponentials §4.4, [19] | HMM in principle; infinite state-space; no practical solution provided (§4.4) |
| Assumptions | FCFS; routing probabilities $p_{ij}$ | FCFS; $p_{ij}$; external arrival rate $\lambda$ | FCFS; $p_{ij}$; transition rates $\eta_{ij}$. | FCFS; $p_{ij}$ $\lambda$; $\eta_{ij}$ |

**Table 1. Summary of Models, Methods and Results.**

Entries in Table 1 correspond to methods used and assumptions made. They will be elaborated on in the suitable places subsequently. (Explicit mention of the parameters $\lambda$, $p_{ij}$, $\eta_{ij}$ indicates that their values are required for the algorithm.) For every model we provide a reference to the corresponding paragraph of our paper.

Two models will be discussed only briefly: the first model is busy-period interpolation with exponential transitions and unknown external arrivals; the second model is real-time estimation with exponential transitions and known external arrivals. See [19] for complete solutions.

**Remarks.**

— The complexity of the problems and the number of assumptions increase rapidly when one moves from the upper left corner of the table to the lower right one.

— Homogeneous Poisson arrivals were assumed throughout the table. If arrivals are non-homogeneous Poisson then the arrival rate $\alpha(t)$ must be known for all real-time cells of the table. From a practical point of view, this is reasonable: estimating $\alpha(t)$, say a periodic function, requires data over many periods; such estimation should be carried out prior to queue-inference.

— Real-time estimation and general interpolation turn out to require similar analysis (see §5.2, 5.3), hence they are combined in Table 1.

## 1.6 Observed and relevant information

The following question is fundamental: which part of the observed information is relevant for our estimation of the queue length? The question must be addressed separately for each of the 12 special cases. (See, for example, §4.2.)

For technical reasons, we prefer to work with *observed events* rather then $\sigma$-fields. Such observed events often contain information that is irrelevant for estimating the queues. This gives rise to the notion of relevant information. More formally, let $E_t \in \mathcal{E}_t$ be an observed event. (See Section 3 for a description, especially 1.1-1.8 in §3.1.) Any event $F_t^s \in \mathcal{E}_t,\ s \leq t$, will be called *relevant information* for estimating $Q_s$, if $E_t \subseteq F_t^s$ and the two conditional distributions $\mathrm{P}\{Q_s \in \bullet \mid F_t^s\}$ and $\mathrm{P}\{Q_s \in \bullet \mid E_t\}$ coincide. In specific applications, the event $F_t^s$ will be chosen so that $\mathrm{P}\{Q_s \in \bullet \mid F_t^s\}$ is convenient to calculate while $F_t^s$ excludes only information in $E_t$ that is irrelevant for estimating $Q_s,\ s \leq t$.

The events $E_t$ and $F_t^s$ can have probability zero due to specifying a value of a continuous random variable, namely a service time. In this case, conditioning on $E_t$ or $F_t^s$ takes the meaning of regular conditioning (see [7], pages 198-200), for example $\mathrm{P}\{X \leq x \mid Y = y\}$, where $Y$ is a continuous random variables.

For Busy-Period Interpolation, treated in Section 2, the structure of the relevant information is simple, hence only intuition-based statements will be presented. This conforms to the level of treatment in all previous publications on the subject. Intuition does not suffice for Real-Time Estimation (Section 4) and General Interpolation (Section 5), where the structure of relevant information is less intuitive, hence it must be treated rigorously. The mathematical framework to support such a treatment is developed in Section 3.

# 2 Busy-Period Interpolation

Busy-Period Interpolation problems cover a specific busy period, which terminated before our estimation time $T$. Here we focus on such a busy period, hence notation can and will be simplified (throughout this section only) by omitting the indices of the busy period and its associated station. Transactions during the busy period will be denoted by $t_0,\ t_1,\dots,t_n,t_{n+1}$, where
$t_0$ is the start of the busy period,
$t_1,\dots,t_n$ are service starts, which are also service terminations and
$t_{n+1}$ is the end of the busy period.
Without loss of generality, we assume that the estimation time is $T = t_{n+1}$.

Note that in the multi-server case, a customer that starts service at $t_i$ need not leave at $t_{i+1}$.

Let $A_0 = t_0, A_1,\dots,A_n$ stand for the arrival times of the customers that started service at $t_0,\ t_1,\dots,t_n$, respectively. The cumulative number of arrivals during $(t_0,t]$ will be denoted by $A(t),\ \ t_0 \le t \le t_n\ \ (A(t_0)=0,\ A(t_n)=A(t_{n+1})=n)$.

It is worth mentioning that the ID's of all the customers, that were served during the busy period, are available. The problem, then, reduces to making inferences about arrival times of a completely specified customer set. Knowledge of this set also renders the Markovian assumptions on customers routing unnecessary.

## 2.1 Immediate Transitions, Known External Arrivals

This problem is deterministic in the sense that one can calculate the queues exactly. To clarify this point, represent the queue process on the $j$-th station as

$$Q_j(t) = Q_j^\alpha(t) + Q_j^\beta(t),$$

where $Q_j^\alpha(t)$ is the part of the queue due to external customers (customers that arrived from station 0) and $Q_j^\beta(t)$ is due to internal ones.

The quantity $Q_j^\alpha(t)$ is known exactly, according to problem definition. Consider the second term. If a customer arrived from some other station, then his arrival time to the $j$-th station coincides with the termination time of his most recent service. But we know the time of the most recent termination because all the network is observed and all the

ID's are registered. Information about arrivals to the station provides us with the queue size. Note that Poisson arrivals are not necessary here. This is also the only case where the FCFS queue discipline plays no role.

## 2.2 Immediate Transitions, Unknown External Arrivals

**Model Definition and Relevant Information** We distinguish again between external and internal customers. The arrivals of internal ones are known, following the analysis in §2.1. We need, therefore, to estimate only external arrivals. Suppose that the external arrivals to a station constitute a homogeneous Poisson process with a known arrival rate $\lambda_j = p_{0j}\lambda$, where $\lambda$ is an external arrival rate to the entrance of the network. (At the end of this paragraph, we modify our results to accommodate the non-homogeneous case.) By the memoryless property of the exponential distribution, we can assume that this process starts at $t_0$.

Suppose that $t_i$ is a service start of some internal customer; his most recent service-termination time (which is known, according to §2.1) will be denoted by $s_i$. The relevant information can now be represented as the intersection of the following four events:
$E_1$. $A_0 = t_0$ (the first arrival time coincides with the start of our busy period);
$E_2$. $A_i = s_i$, if $A_i$ is an internal arrival time (arrival time of an internal customer coincides with his most recent termination-time);
$E_3$. $A_i \le t_i$, if $A_i$ is an external arrival time (a customer must arrive before his service starts);
$E_4$. If $A_k < A_l$ are any adjacent internal arrival times, then

$$s_k = A_k < A_{k+1} < A_{k+2} < \ldots < A_{l-1} < A_l = s_l.$$

If the first or last arrival of the busy period is not internal then

$$t_0 < A_1 < A_2 < \ldots < A_{l-1} < A_l = s_l,$$

in case $A_l$ is the first internal arrival and

$$s_k = A_k < A_{k+1} < A_{k+2} < \ldots < A_n \le t_n,$$

in case $A_k$ is the last internal arrival.

Event $E_4$ amounts to $E_2$ combined with $A_{i+1} > A_i$, $k \le i \le l-1$, the latter being a consequence of FCFS.

Consider two adjacent internal arrival times $A_k = s_k$ and $A_l = s_l$. The interval $[0, t_n]$ is divided into intervals of the type $[s_k, s_l)$ and then cumulative arrivals $A(t)$ and queue length are estimated separately for every such interval (see Figure 2.1).

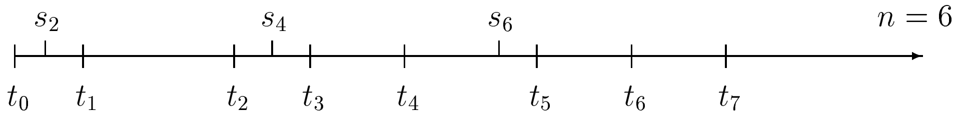


**Figure 2.1. Example of Busy Period. Immediate Transitions, Unknown External Arrivals.** *Customers 2, 4 and 6 are internal. Queue estimation is carried out separately on* $[t_0, s_2)$, $[s_2, s_4)$, $[s_4, s_6)$ *and* $[s_6, t_6)$.

We consider two cases, according to the location of $s_l$ relative to $t_{k+1}$.

**First case:** $t_{k+1} > s_l$ (see Figure 2.2).

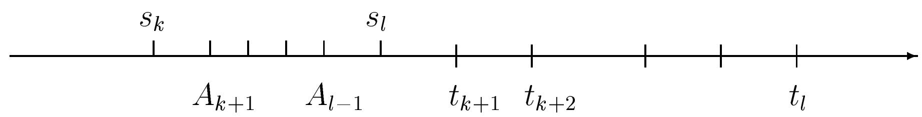


**Figure 2.2. Case 1. Immediate Transitions, Unknown External Arrivals.**

Then event $E_3$ from the relevant-information list follows from event $E_4$, for $i \in \{k+1, \ldots, l-1\}$, and it can be omitted. From event $E_4$ and the order-statistics property of the Poisson process, we conclude that $A_{k+1}, \ldots, A_{l-1}$ are distributed as uniform order statistics on $[s_k, s_l]$. This means that for $t \in [s_k, s_l)$, $0 \le i \le l-k-1$,

$$\mathrm{P}\{A(t) = k+i \mid F_T\} = \binom{i}{l-k-1} \frac{(t-s_k)^i (s_l - t)^{l-i-1}}{(s_l - s_k)^{l-k-1}}, \tag{2.1}$$

$$\hat{A}(t) = \mathrm{E}[A(t) \mid F_T] = k + (l-k-1)\frac{t-s_k}{s_l-s_k}, \tag{2.2}$$

$$\mathrm{Var}[A(t) \mid F_T] = (l-k-1)\frac{(t-s_k)(s_l-t)}{(s_l-s_k)^2}, \tag{2.3}$$

where $F_T$ is the relevant information. As for the queue length, if $t \in [t_i, t_{i+1})$, $0 \le i \le n-1$, then

$$\hat{Q}(t) = \hat{A}(t) - i \tag{2.4}$$

**Second case:** $t_{k+1} < \ldots < t_m < s_l < t_{m+1} < \ldots < t_l$ (see Figure 2.3).

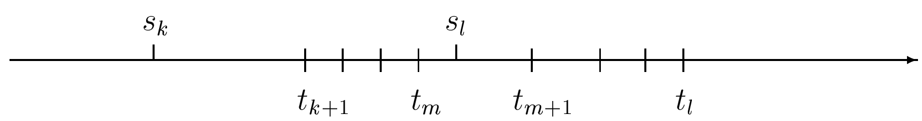


**Figure 2.3. Case 2. Immediate Transitions, Unknown External Arrivals.**

Now $A_{k+1}, \ldots, A_{l-1}$ are distributed as uniform order-statistics on $[s_k, s_l]$, conditioned on $\{A_{k+1} \le t_{k+1},\ A_{k+2} \le t_{k+2}, \ldots, A_m \le t_m\}$.

**Formulation as HMM** We formalize the second-case problem in HMM terms, following [4]. Our HMM applies to a discrete-time Markov chain, evolving on a finite time set and a countable state space. (This Markov chain may be time-inhomogeneous.) The HMM ingredients are: Markov chain (time set, state space, transition probabilities); boundary conditions (constraints on the way the chain starts and ends its evolution); taboo conditions (for each time, a set of states that cannot be visited at this time).

Conditioning on the boundary and taboo conditions, the HMM algorithm calculates the distribution of the Markov chain at a given time.

The ingredients in our case are as follows.

$$\textit{Markov chain:}\ \{A(s_k), A(t_{k+1}), \dots, A(t_m), A(s_l-)\}, \tag{2.5}$$

where

$$A(s_l-) = \lim_{\epsilon \downarrow 0} A(s_l - \epsilon).$$

*Boundary conditions:* $A(s_k) = k$ and $A(s_l-) = l-1$.

*Admissible (non-taboo) states*: $A(t_i) \in \mathcal{G}_i$, where

$$\mathcal{G}_i = \{i, i+1, \dots, l-1\}, \quad k+1 \le i \le m,$$

since at least $i$ arrivals must occur before the start of the $i$-th service.

**Remark.** Our Markov chain has been constructed in such a way that the events that constitute the boundary and taboo conditions are identical to the relevant information.

*Transition probabilities.* To simplify the presentation, we maintain index continuity by assigning

$$z_k = s_k,\ z_{k+1} = t_{k+1}, \dots, z_m = t_m,\ z_{m+1} = s_l.$$

The transition probabilities between the admissible states of the Markov chain in (2.5) are given by

$$\begin{aligned} \tilde{p}_i(u,v) &= \mathrm{P}\{A(z_i) = v \mid A(z_{i-1}) = u\} = \\ &= e^{-\lambda_j(z_i - z_{i-1})} \frac{(\lambda_j(z_i - z_{i-1}))^{v-u}}{(v-u)!}, \quad k+1 \le i \le m,\ i \le u \le v \le l-1, \end{aligned} \tag{2.6}$$

$$\begin{aligned} \tilde{p}_{m+1}(u, l-1) &= \mathrm{P}\{A(z_{m+1}-) = l-1 \mid A(z_m) = u\} = \\ &= e^{-\lambda_j(z_{m+1} - z_m)} \frac{(\lambda_j(z_{m+1} - z_m))^{l-1-u}}{(l-1-u)!}, \quad m \le u \le l-1. \end{aligned} \tag{2.7}$$

All other transition probabilities between admissible states are equal to zero.

Further introduce the events

$$B^{r_1,r_2} = \{A(z_i) \in \mathcal{G}_i,\ r_1 \le i \le r_2\}, \quad k+1 \le r_1 < r_2 \le m.$$

The taboo probability matrices are defined as $W^{r_1,r_2} = [w^{r_1,r_2}_{u,v}]$, where

$$w^{r_1,r_2}_{u,v} = \mathrm{P}\{A(z_{r_2}) = v; B^{r_1+1,r_2-1} \mid A(z_{r_1}) = u\}, \quad k \le r_1 \le r_2 \le m+1,$$

in which $u$ and $v$ are admissible states.

We now adapt the algorithm presented in Theorem 1 from Daley and Servi [4] to calculate queue distribution at transaction times. The first three steps of our Algorithm 2.1 coincide, in essence, with the first three steps of the algorithm from Theorem 1 of Daley and Servi [4]. The only distinction is that we use the model with partial state observations (Hidden Markov Model), which generalizes the Two-Point Markov Chain Boundary Value Problem treated in Theorem 1 [4]. (Daley and Servi [4] provided the alterations of definitions which are necessary for such a generalization.)

**Algorithm 2.1**

1°. Using (2.6) and (2.7), define transition matrices between admissible states:

$$\tilde{P}_i = \{\tilde{p}_i(u,v),\ u \in \mathcal{G}_{i-1},\ v \in \mathcal{G}_i\}, \quad k+1 \le i \le m+1,$$

where $\mathcal{G}_k = k,\ \mathcal{G}_{m+1} = l$. The dimensions of these matrices are $1 \times (l-k-1)$ for $i = k+1$, $(l-i+1) \times (l-i)$ for $k+2 \le i \le m$, and $(l-m) \times 1$ for $i = m+1$.

2°. Calculating the matrices of taboo probabilities:

$$\begin{aligned} W^{k,k+1} &= \tilde{P}_{k+1}, \\ W^{k,i} &= W^{k,i-1}\tilde{P}_i, \quad k+2 \le i \le m+1, \\ W^{m,m+1} &= \tilde{P}_{m+1}, \\ W^{i,m+1} &= \tilde{P}_{i+1}W^{i+1,m+1}, \quad k+1 \le i \le m-1. \end{aligned}$$

3°. Calculating the distribution for cumulative number of arrivals at transaction times:

$$\begin{aligned} &\mathrm{P}\{A(z_i) = u \mid E_r\} = \\ &= \mathrm{P}\{A(z_i) = u \mid A(z_k) = k; A(z_{m+1}) = l; A(z_r) \in \mathcal{G}_r,\ k+1 \le r \le m-1\} \\ &= \frac{w^{k,i}_{k,u} w^{i,m+1}_{u,l}}{W^{k,i}W^{i,m+1}}. \end{aligned}$$

(The denominator of the last expression is the scalar product of two vectors.)

4°. Calculating the queue distribution at transaction times, using

$$\hat{Q}(z_i) = \hat{A}(z_i) - i, \quad k+1 \le i \le m.$$

**Discussion and Extensions**

• Algorithm 2.1 is used for queue estimation also in §2.4 and in other special cases treated in [19]. In general, our technical contribution is to define a proper Hidden Markov Model (state space, transitional probabilities, taboo conditions) for every special case and, then, use the algorithm. The algorithm produces estimates of cumulative arrivals at transaction times. These, combined with cumulative departures, yield queue estimates.

• Implementation of Algorithm 2.1 to real data is presented in Section 6.

• Suppose that external arrivals to the network constitute a non-homogeneous Poisson process with a known time-dependent arrival rate $\alpha(t)$. To accommodate such a generalization, define $\alpha_j(s) = p_{0j}\alpha(s)$ and let $\Lambda_j(t) = \int_0^t \alpha_j(s)ds$ stand for the cumulative arrival rate. The transition probabilities from (2.6) are now:

$$\tilde{p}_i(u,v) = e^{-(\Lambda_j(z_i)-\Lambda_j(z_{i-1}))}\frac{(\Lambda_j(z_i)-\Lambda_j(z_{i-1}))^{v-u}}{(v-u)!}, \quad k+1 \le i \le m,\ i \le u \le v \le l-1. \tag{2.8}$$

Equation (2.7) is modified similarly.

Note that the same solution may be derived by performing a time-change $t \to \int_0^t \alpha_j(s)ds$, assuming that the arrival rate $\lambda = 1$ and, finally, performing the reverse time-change.

• Suppose that $t$ is not a transaction point and one needs to estimate the queue at time $t$. If only conditional expectation of the queue is of interest, then one can use its linearity between the transaction points (see [4],[13]). Otherwise, $t$ can be added to the domain of the Markov chain (2.5). If $t \in (t_i, t_{i+1})$, then the set of the admissible states for $t$ is $\{i, i+1, \ldots, l-1\}$. The transition probabilities for $A(t_i) \to A(t)$ and $A(t) \to A(t_{i+1})$ are derived using the Poisson distribution.

• Given the conditional distribution of queue-length, it is easy to calculate conditional expectation, variance and other distributional characteristics.

• As in the single-station case, the queue-length distribution does not depend on $\lambda$. Indeed, this follows from formulae (2.1)-(2.4) for the first case. For the second case, the distribution of external arrivals is that of uniform order statistics, conditioned on an event whose probability does not depend on $\lambda$.

The independence on $\lambda$ is useful for debugging of any software implementation of the algorithm: simply substitute two distinct values of $\lambda$ and check that probabilities do not change.

• The jump magnitude of our hidden Markov chain is bounded. It follows from Theorem 2 in [4] that Algorithm 2.1 involves $O(m^3)$ calculations, where $m$ is the number of estimation points in an interval between two successive internal arrivals. It is important, therefore, from a computational point of view, that our algorithm is executed separately on each interval between successive internal arrivals, since such intervals typically include less points than a whole busy period. Note also that for many queueing networks it is reasonable to expect that the number of internal arrivals in a busy period with $n$ customers is of order $n$. For example, in the steady-state of a Jackson network, the long-run proportion of internal arrivals to station $j$ is given by $(\lambda_j - \alpha_j)/\lambda_j$, where $\alpha_j$ is the external arrival rate and $\lambda_j$ is the solution of flow conservation equations (see [18], for example).

If every $k$-th arrival is internal, the number of calculations needed to estimate the queue at transaction points during a busy period will be reduced to $O(n)$, where $n$ is the number of transactions in the period. (Recall that in the single-station case, $O(n^3)$ calculations are needed.)

If internal arrivals are "approximately uniformly" dispersed between external ones, the same result can be expected. The knowledge of external arrival times enables, therefore, a decrease in the number of calculations even in comparison with the case of a single service station.

## 2.3 Truncated Exponential Random Variables

In the case of exponential transition times between the stations we need several facts about truncated exponential random variables.

**Notation.** Define a truncated exponential random variable with rate $\lambda$ and domain $[a, b]$ as a random variable $\zeta$ with density:

$$f(y) = \frac{\lambda e^{-\lambda y}}{e^{-\lambda a} - e^{-\lambda b}}, \quad a \le y \le b.$$

It is denoted by $\zeta \sim \text{truncexp}(\lambda; a, b)$ and its interpretation is that of an exponential random variable $\zeta_0 \sim \exp(\lambda)$, first shifted by $a$ and then conditioned to have values in $[a, b]$. Specifically, suppose that $\zeta_1, \zeta_2, \ldots, \zeta_n$ are independent truncated exponential random variables. In [19] we calculate the probability $\mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n\}$, expressions for which are presented in the following two propositions.

**Proposition 2.1**
Let $\zeta_i \sim \text{truncexp}(\lambda_i; x_i, t)$, $1 \le i \le n$ and $x_1 < x_2 < \ldots < x_n < t$.
Denote $P_\Lambda^X(t) = \mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n\}$, where $\Lambda = (\lambda_1, \ldots, \lambda_n)$ and $X = (x_1, \ldots, x_n)$.
Then

$$\begin{aligned} P_\Lambda^X(t) &= \frac{1}{e^{-\lambda_n x_n} - e^{-\lambda_n t}} \left[ e^{-\lambda_n x_n} \frac{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i x_n})}{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i t})} P_{\Lambda_1}^{X_1}(x_n) + \right. \\ &+ \left. \frac{\lambda_{n-1}}{\lambda_{n-1} + \lambda_n} \frac{e^{-(\lambda_{n-1}+\lambda_n)x_n} - e^{-(\lambda_{n-1}+\lambda_n)t}}{e^{-\lambda_{n-1}x_{n-1}} - e^{-\lambda_{n-1}t}} P_{\Lambda_2}^{X_2}(t) - e^{-\lambda_n t} P_{\Lambda_1}^{X_1}(t) \right], \end{aligned} \quad (2.9)$$

where

$$\Lambda_1 = (\lambda_1, \ldots, \lambda_{n-1}),$$

$$
\begin{aligned}
X_1 &= (x_1, \dots, x_{n-1}), \\
\Lambda_2 &= (\lambda_1, \dots, \lambda_{n-1} + \lambda_n), \\
X_2 &= (x_1, \dots, x_{n-2}, x_n).
\end{aligned}
$$

**Proposition 2.2**
Let $\zeta_i \sim \text{truncexp}(\lambda_i; x_i, t)$, $1 \le i \le n$, where $0 \le x_i \le t$. Let

$$r_i = \max_{j \le i} x_j \quad \text{(see Figure 2.4) and} \quad R = (r_1, r_2, \dots, r_n).$$

Then

$$P_\Lambda^X(t) = P_\Lambda^R(t) \prod_{i=1}^{n} \frac{e^{-\lambda_i r_i} - e^{-\lambda_i t}}{e^{-\lambda_i x_i} - e^{-\lambda_i t}}, \tag{2.10}$$

where $P_\Lambda^R(t)$ can be calculated using Proposition 2.1.

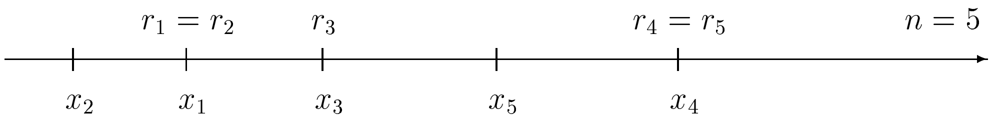


**Figure 2.4. Illustration to Proposition 2.2.**

## 2.4 Exponential Transitions, Known External Arrivals

**Model Definition and Relevant Information** In the present model we eliminate the assumption of immediate transitions between stations. This assumption is indeed too restrictive for many applications, such as transportation models and human service systems. (It is appropriate, however, for many communication networks.)

Assume that the transition time of a customer, routed from station $j$ to station $k$, is exponentially distributed with a known parameter $\eta_{jk}$, $j \in \mathcal{D}_0$, $k \in \mathcal{D}$. New customers are registered upon arrival to station 0 and then proceed to station $j$ with probability $p_{0j}$, $j \in \mathcal{D}$. Finally, every transition time is assumed independent of all other stochastic components of the model. Note that our model allows overtaking between customers.

We now determine the part of the available information $\mathcal{E}_t$, which is relevant for queue estimation.

Consider a busy period at some station of the network, with transactions $t_0, t_1, \dots,$ $t_n, t_{n+1}$, as before. For the customer whose service-start time is $t_i$, $0 \le i \le n$, let $s_i$ denote the most recent transaction time prior to $t_i$. These times will be referred to as *transition-start times* (see Figure 2.5). They may be either service termination or arrival times to station 0. According to our model specification, $s_0, s_1, \dots, s_n$ can be deduced from our data.

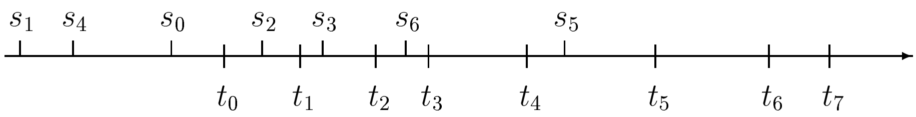


**Figure 2.5. Example of Busy Period.**
**Exponential Transitions, Known External Arrivals.**

Enumerate by $1, \ldots, n$ the customers that started service at $t_1, \ldots, t_n$. Letting $A_1, \ldots, A_n$ denote their unknown arrival times, we have

$$A_i = s_i + \tau_i, \quad 1 \le i \le n, \tag{2.11}$$

where $\tau_i \sim \exp(\gamma_i)$, $\gamma_i$ are the corresponding known transition rates and $\tau_i$ are independent of each other.

Two important facts about arrival times constitute the relevant information:

$E_1$. Every customer arrived before his service start, namely

$$A_1 \le t_1,\ A_2 \le t_2, \ldots, A_n \le t_n;$$

$E_2$. From the FCFS queueing discipline,

$$t_0 = A_0 < A_1 < A_2 < \ldots < A_n \le t_n.$$

The arrivals to the station are distributed according to (2.11), conditioned on the event $E_1 \bigcap E_2$.

**Formulation as HMM** We introduce the ingredients of HMM.

*Markov chain:* $\tilde{A} = \{\tilde{A}(t_0), \ldots, \tilde{A}(t_n)\}$,
whose state at $t_i$ is the set of customers that arrived during $(t_0, t_i]$. The set is ordered according to arrival times. In other words, the state space $\mathcal{A}$ is

$$\{\emptyset\} \bigcup \{m_1, \ldots, m_l\}, \quad m_j \in \{1, \ldots, n\},\ \ m_i \ne m_j.$$

(The state space is more complicated than in the previous subsection since event $E_2$ must be accommodated in the taboo conditions.)

*Boundary conditions:*

$$\tilde{A}(t_0) = \{\emptyset\};\ \tilde{A}(t_n) = \{1, \ldots, n\}.$$

*Admissible states* at $t_i$ are given by

$$\mathcal{G}_i = \{\ \{1, 2, \ldots, i\},\ \{1, 2, \ldots, i, i+1\},\ \{1, 2, \ldots, n\}\ \}.$$

*Transition probabilities* between admissible states. We shall use capital letters as a convenient notation for these states. For example

$$I \stackrel{\text{def}}{=} \{1,2,\ldots,i\}, \quad N \stackrel{\text{def}}{=} \{1,2,\ldots,n\}, \quad K+L \stackrel{\text{def}}{=} \{1,2,\ldots,k+l\}.$$

Then

$$\mathrm{P}\{\tilde{A}(t_{i+1}) = N \mid \tilde{A}(t_i) = N\} = 1, \quad 0 \le i \le n-1,$$

and for $0 \le k \le n-1,\ k+l \le n$,

$$\begin{aligned}
\mathrm{P}\{\tilde{A}(t_{i+1}) = K \mid \tilde{A}(t_i) = K\} &= \prod_{j=k+1}^{n} \exp\{-\gamma_j(t_{i+1} - \max(t_i, s_j))_+\}, \\
\mathrm{P}\{\tilde{A}(t_{i+1}) = K+L \mid \tilde{A}(t_i) = K\} &= 0, \quad \text{if } \exists s_m \in \{s_{k+1},\ldots,s_{k+l}\} \ni\ s_m > t_{i+1}.
\end{aligned}$$

Otherwise,

$$\begin{aligned}
&\mathrm{P}\{\tilde{A}(t_{i+1}) = K+L \mid \tilde{A}(t_i) = K\} = \\
&= \mathrm{P}\{t_{i+1} < A_{k+l+1},\ldots,t_{i+1} < A_n;\ t_i < A_{k+1} \le t_{i+1},\ldots,t_i < A_{k+l} \le t_{i+1}; \\
&\quad A_{k+1} < A_{k+2} < \ldots < A_{k+l} \mid t_i < A_{k+1},\ldots,t_i < A_n\} \\
&= \prod_{j=k+l+1}^{n} \exp\{-\gamma_j(t_{i+1} - \max(t_i,s_j))_+\} \prod_{j=k+1}^{k+l} (1 - \exp\{-\gamma_j(t_{i+1} - \max(t_i,s_j))\}) \times \\
&\times \mathrm{P}\{A_{k+1} < A_{k+2} < \ldots < A_{k+l} \mid t_i < A_{k+1} \le t_{i+1},\ldots,t_i < A_{k+l} \le t_{i+1}\}. \qquad (2.12)
\end{aligned}$$

Consider the last term of (2.12). Given $\{t_i < A_j \le t_{i+1}\}$, we have

$$A_j \sim \text{ truncexp}(\gamma_j;\ \max(s_j,t_i), t_{i+1}).$$

A recursive algorithm for the calculation of the probabilities associated with this term is presented in §2.3. The key formulas are (2.9) and (2.10).

Now we can use Algorithm 2.1 for calculating the queue distribution.

**Discussion**

— Unlike the case in §2.2, it is no longer true that the expectation of the queue is linear between transaction points.

— If queue distribution at some non-transaction time $t$ is of interest, then $t$ must be added to the domain of the Markov chain. Specifically, if $t \in (t_i, t_{i+1})$, then the set of the admissible states for $t$ is $\{I,\ldots,N\}$.

## 2.5 Exponential Transitions, Unknown External Arrivals

The problem was posed and solved in [19]. The treatment there resembles that of the previous case, in which external arrivals were known. The state-space and taboo-states of the relevant HMM are identical. The difference is the calculation of transition probabilities.

Moreover, external arrivals to station $j$ behave as departures from a $M_t/M/1$ queue, which is used in the calculation of arrival rates and transition probabilities. These probabilities include multidimensional integrals and, unfortunately, we have not been able to derive recursive algorithm for their calculations. (Specifically, the expression that corresponds to the last term of (2.12) contains $A_i$, which are not truncated exponential). These integrals, however, can always be computed approximately by numerical methods.

# 3 Representations of an Observed Event

As already mentioned in §1.6, our results and algorithms for the Interpolation case were not treated as carefully as they could have been. For example, information that is relevant for queue inference was identified intuitively.

We have used, in essence, the terminology and approaches of the previous papers by Larson and Daley & Servi on single-station models [13], [3], [4]. It seemed inappropriate to complicate considerably our models and methods in order to derive rigorous proofs for facts that had been intuitively clear. However, the cases of Real-Time Estimation and General Interpolation are more difficult, in particular the structure of the relevant information is less intuitive, so formal methods will now be developed for these problems.

Representations of the observed event play an important role in our approach. They serve as a mathematical framework within which Interpolation and Real-Time problems are conveniently formulated and solved.

Three types of equivalent representations are introduced. The first, via *transaction times* (representations 1 and 1' in §3.1), describes the real data that is actually observed. The third, via *stochastic components* (namely external arrivals, service times, switches between stations and transition times), is helpful for our analysis: indeed, conditioning on the observed event, represented in this third form, identifies the part of the event that is relevant for queue estimation. (See representations 3 and 4 in §4.2 and §4.3, respectively.) The second representation, via *arrival times* to stations (representations 2 and 2' in §3.2), helps to ascertain the equivalence between the two former representations.

**Remark.** In §4.3 we shall consider the model with unknown external arrivals and immediate transitions between stations. In this case, our original assumption that all arrivals join station zero is cumbersome to model. It turns out convenient to assume, instead, that external arrivals join each station directly. In the case of immediate transitions, the new model coincides with the basic one if external arrival rate is constant and generalizes it if the arrival rate is time-dependent. This is our reason to introduce representations 1' and 2' for the case of unknown external arrivals.

Summarizing, we establish equivalence between representations 1, 2 and 3 if external arrivals are known, and between 1', 2' and 4, otherwise.

## 3.1 Representations via Transaction Times

For simplicity of notation (which is somewhat cumbersome anyway) we shall assume, everywhere throughout Sections 3, 4 and 5, that only a single server is present at every station. However, all our results remain valid for the multi-server case if one adopts our definition of a busy period (see §1.4).

Recall that we always register start and termination times of services with ID's of customers and sometimes also arrival times and ID's of external customers.

At first, a notation for transaction times will be introduced. Let $N_k^j$ denote the number of customers that were served during the $k$-th busy period at station $j$ (without counting the customer that starts the busy period, since customers at the $k$-th busy period are numbered by $0, 1, 2, \ldots, N_k^j$). Introduce $T_{ki}^j$, $0 \leq i \leq N_k^j + 1$, as the time of the $i$-th transaction during busy period number $k$ at station $j$. In particular, $T_{k0}^j$ and $T_{k,N_k^j+1}^j$ stand for the service start of customer 0 and for the service termination of customer $N_k^j$ respectively; time $T_{ki}^j$, $1 \leq i \leq N_k^j$, denotes the time of both the $i$-th service completion and the $(i+1)$-th service start.

Suppose that we observe the network during the time interval $[0, t]$. Let $B_j(t)$, $j \in \mathcal{D}$, denote the number of busy periods, which started at station $j$ up to time $t$. Let $\mathcal{D}_b(t)$ stand for the *set of busy stations* at time $t$ (i.e. stations whose servers are busy). For $j \in \mathcal{D}_b(t)$ introduce $M_j(t)$ to be the number of customers that started service before $t$ during the current busy period. (Here, again, we do not count customer 0, which started this busy period.)

Finally, introduce a notation for external arrivals: let $A^0(t)$ stand for the number of external arrivals during $[0, t]$ and $A_i^0$ denote the time of the $i$-th external arrival to the network.

We now develop a theoretical framework for describing the information embodied in the ID's of customers. Using ID's, it is possible to obtain the trace of a customer (the sequence of stations visited up to time $t$) and the times of transactions up to $t$. To this end, introduce two routing operators. The *Routing operator* $\mathcal{R}$ locates the next service start of a customer, the *Backward Routing operator* $\mathcal{R}^{-1}$ locates his previous service termination.

**Definition of the Routing Operator $\mathcal{R}$.** The domain of $\mathcal{R}$ consists of triplets of natural numbers $(n, m, l)$ and pairs of the type $(0, l)$. The value of the routing operator is given by

$$\mathcal{R}(n, m, l) = (j, k, i) \quad (\text{ or } \mathcal{R}(0, l) = (j, k, i)\ ) \tag{3.1}$$

if the customer, which terminated service at time $T_{ml}^n$ at station $n$ (or entered the network at $A_l^0$), starts his next service at time $T_{ki}^j$ at station $j$. If that customer leaves the network, let

$$\mathcal{R}(n, m, l) = 0 \quad (\text{or } \mathcal{R}(0, l) = 0\,).$$

We introduce also a special notation for the station to where the customer switches: if formula (3.1) prevails, then

$$R^n_{ml} = j \quad (\text{or } R^0_l = j).$$

**Definition of the Backward Routing operator $\mathcal{R}^{-1}$** The domain of $\mathcal{R}^{-1}$ is formed by triplets of natural numbers. It takes the value

$$\mathcal{R}^{-1}(j,k,i) = (n,m,l)$$

if the customer, which starts service at time $T^j_{ki}$, terminated his previous service at $T^n_{ml}$. If the service that starts at time $T^j_{ki}$ was the first one for that customer let

$$\mathcal{R}^{-1}(j,k,i) = (0,l),$$

where $l$ is the index of the customer's transaction at station 0.

Roughly speaking, the information embodied in ID's is identical to the information about values of the Backward Routing operator. Indeed, an ID of a customer enables the derivation of all his previous transactions. But it is possible to get that same information recursively from the Backward Routing operator.

**Representation 1 of $\mathbf{E_t}$ via Observations.**
**Known External Arrivals.**

The observed event $E_t$ can be represented as the intersection of the eight events 1.1-1.8 below. Here and later we adopt a uniform style of event representations: small letters will be used for realizations of random variables; a verbal description of an event is followed by formulae; the main statements are printed in bold.

**1.1** The number of busy periods at every station is known.
$\mathbf{B_j(t) = b_j(t)}, \quad j \in \mathcal{D}.$

**1.2** The set of busy stations is known.
$\mathcal{D}_\mathbf{b}\mathbf{(t) = d_b(t)}.$

**1.3** The number of transactions is known for all complete busy periods.
$\mathbf{N^j_k = n^j_k}, \quad j \in \mathcal{D},$
$1 \le k \le b_j(t)$ for $j \notin d_b(t)$,
$1 \le k \le b_j(t) - 1$ for $j \in d_b(t)$.

**1.4** The number of transactions until $t$ for incomplete busy periods is known.
$\mathbf{M_j(t) = m_j(t)}, \quad j \in d_b(t).$

**1.5** The transaction times are known.

$\mathbf{T^j_{ki} = t^j_{ki}}, \quad j \in \mathcal{D},\ 1 \le k \le b_j(t),$
$0 \le i \le m_j(t) \quad$ for $\ j \in d_b(t),\ k = b_j(t),$
$0 \le i \le n^j_k + 1, \quad$ otherwise.

**1.6** Complete information about the backward routing.

For $(j,k,i)\ \ni\ t^j_{ki} \le t,$
$\mathcal{R}^{-1}(\mathbf{j},\mathbf{k},\mathbf{i}) = (\mathbf{n},\mathbf{m},\mathbf{l})$ **or** $\mathcal{R}^{-1}(\mathbf{j},\mathbf{k},\mathbf{i}) = (\mathbf{0},\mathbf{l}).$

**1.7** The number of external arrivals is known.

$\mathbf{A^0(t) = a^0(t)}.$

**1.8** The transactions at station 0 are known.

$\mathbf{A^0_i = t^0_i},\ 1 \le i \le a^0(t).$

**Representation 1' of $\mathbf{E_t}$ via Observations.**
**Unknown External Arrivals.**

Here we assume a separate arrival process to each station of the network. Let $A^{0j}_l$ denote the time of the $l$-th external arrival to station $j$. If the corresponding customer starts his first service at station $j$ at time $T^j_{ki}$, we modify the definition of the Routing and Backward Routing operators to:

$$\mathcal{R}(0,j,l) = (j,k,i), \quad \mathcal{R}^{-1}(j,k,i) = (0,j,l).$$

Now we define the observed event $E_t$ as the intersection of the six events 1.1'-1.6'. The events 1.1'-1.5' coincide with their counterparts 1.1-1.5 from Representation 1. Event 1.6' is defined by:

**1.6'** Partial information about the backward routing.

For $(j,k,i)\ \ni\ t^j_{ki} \le t,$
$\mathcal{R}^{-1}(\mathbf{j},\mathbf{k},\mathbf{i}) = (\mathbf{n},\mathbf{m},\mathbf{l}),\ \mathbf{n > 0}\ $ or $\ \mathcal{R}^{-1}(\mathbf{j},\mathbf{k},\mathbf{i}) = (\mathbf{0},\mathbf{j},\mathbf{l}).$

**Remark.** The difference between 1.6 and 1.6' is due to the fact that, in the latter (with external arrivals unknown), information about a customer's history is incomplete in the following sense: if the ID of a customer, starting service at $T^j_{ki}$, has never been registered before, it must be that he arrived to station $j$ from the outside; in addition, his arrival number $l$ is known due to the FCFS discipline; however, his arrival time to station $j$ is unknown.

## 3.2 Representations via Arrival Times

Recall that $t^j_{ki}$ denotes the realization of the $i$-th transaction time from the $k$-th busy period at station $j$. Let $\xi^j_{ki}$ denote the duration of service which started at $t^j_{ki}$. Let also

$A^j_{ki}$ stand for the arrival time to station $j$ of the customer, which started service at $t^j_{ki}$. Finally, $\bar{A}^n_{ml}$ stands for the arrival time to his next station of the customer that terminated service at $t^n_{ml}$. A connection between the two last definitions is given by:

$$\mathcal{R}(n,m,l) = (j,k,i) \quad \Rightarrow \quad \bar{A}^n_{ml} = A^j_{ki}.$$

(Note that $\bar{A}^n_{ml} = t^n_{ml}$ in the case of immediate transitions between stations. The following representations via arrival times are valid for arbitrary transition times.)

Consider a service-termination time $t^n_{ml} \leq t$. It will be called a *Visible-Routing* service termination time (or VR-time in short) if

$$\exists\, (j,k,i) \ \ni \ t^j_{ki} \leq t,\ \mathcal{R}(n,m,l) = (j,k,i).$$

In words, this customer registered at least once during the time interval $[t^n_{ml}, t]$. Otherwise, time $t^n_{ml}$ will be called *Invisible-Routing* (IR-time). A time of service termination is IR-time if either the corresponding customer leaves the network or if his next service start takes place after $t$. We use the notation $(n,m,l) \in V_R(t)$ and $(n,m,l) \in I_R(t)$ for VR- and IR-times respectively. The corresponding customers are called *visible customers* and *invisible customers.* In words, a customer is visible at time $t$ if the identity of the station he switched to, after his last service-termination at time $t^n_{ml}$, is known.

The definition of VR- and IR-times can be easily extended to the external arrival times: $A^0_l$ (external customers arrive to station 0) or $A^{0j}_l$ (external customers arrive directly to station j).

**Representation 2 of $\mathbf{E_t}$ via Arrival Times.**
**Known External Arrivals.**
Here we identify conditions that must be imposed on service times and arrivals (internal and external) in order to get the observed event from Representation 1.

We use the notation $A \Rightarrow B$ to denote the event $A^c \bigcup B$.

**2.1** Services that terminated up to $t$ are known.
$\boldsymbol{\xi^j_{ki} = t^j_{k,i+1} - t^j_{ki}}, \quad j \in \mathcal{D},\ 1 \leq k \leq b_j(t),$
$0 \leq i \leq m_j(t) - 1, \qquad \text{for } j \in d_b(t),\ k = b_j(t),$
$0 \leq i \leq n^j_k, \qquad \text{otherwise}.$

**2.2** Last service at a busy station have not terminated until $t$.
For $j \in d_b(t) \quad \boldsymbol{\xi^j_{b_j(t),m_j(t)} > t - t^j_{b_j(t),m_j(t)}}.$

**2.3** The start of a busy period coincides with the first arrival.
For $(n,m,l) \in V_R(t) \ \ni \ \mathcal{R}(n,m,l) = (j,k,0)$
$\boldsymbol{\bar{A}^n_{ml} = A^j_{k0} = t^j_{k0}}.$

**2.4** Truncation rule. Arrivals take place before corresponding service starts.
For $(n,m,l) \in V_R(t) \ni \mathcal{R}(n,m,l) = (j,k,i), \ \ i \geq 1$
$\mathbf{\bar{A}^n_{ml} = A^j_{ki} \leq t^j_{ki}}$.

**2.5** FCFS discipline.
For $(n_i,m_i,l_i) \in V_R(t) \ni \mathcal{R}(n_i,m_i,l_i) = (j,k,i)$,
$\mathbf{t^j_{k0} \leq \bar{A}^{n_1}_{m_1,l_1} = A^j_{k1} \leq \bar{A}^{n_2}_{m_2,l_2} = A^j_{k2} \leq \ldots \leq \bar{A}^{n_v}_{m_v,l_v} = A^j_{kv}}$,
$v = m_j(t)$ for $j \in d_b(t), \ \ k = b_j(t)$,
$v = n^j_k$, otherwise.

**2.6** Complete information about the switches of the visible customers.
For $(n,m,l) \in V_R(t)$ $\mathbf{R^n_{ml} = j}$.

**2.7** Partial information about the invisible routing.
For $(n,m,l) \in I_R(t)$
$\mathbf{if\ j \in d_b(t), \quad R^n_{ml} = j \ \Rightarrow \ \bar{A}^n_{ml} \geq A^j_{b_j(t),m_j(t)}}$,
$\mathbf{if\ j \notin d_b(t), \quad R^n_{ml} = j \ \Rightarrow \ \bar{A}^n_{ml} > t}$.

**2.8** The number of external arrivals is known.
$\mathbf{A^0(t) = a^0(t)}$.

**2.9** The times of external arrivals are known.
$\mathbf{A^0_i = t^0_i}, \quad 1 \leq i \leq a^0(t)$.

**Remark.** Events 2.3-2.6 and event 2.7 include also external arrivals $(0,l) \in V_R(t)$ and $(0,l) \in I_R(t)$ respectively. They are treated in the same way as $(n,m,l)$ (for example, $R^l_0 = j$ in event 2.6).

**Representation 2' of $\mathbf{E_t}$ via Arrival Times.**
**Unknown External Arrivals.**

Event $E_t$ is the intersection of the seven events 2.1'-2.7'. Events 2.1', 2.2' and 2.6' coincide with their counterparts 2.1, 2.2 and 2.6 from Representation 2. Events 2.3'-2.5' are equivalent to events 2.3-2.5 supplemented with the events:

**2.3$^\mathbf{0}$** For $(0,j,l) \in V_R(t) \ni \mathcal{R}(0,j,l) = (j,k,0)$
$\mathbf{A^{0j}_l = A^j_{k0} = t^j_{k0}}$.

**2.4$^\mathbf{0}$** For $(0,j,l) \in V_R(t) \ni \mathcal{R}(0,j,l) = (j,k,i), \ i \geq 1$
$\mathbf{A^{0j}_l = A^j_{ki} \leq t^j_{ki}}$.

**2.5$^\mathbf{0}$** For $(0,j,l) \in V_R(t) \ni \mathcal{R}(0,j,l) = (j,k,i), \ i \geq 1$
$\mathbf{A^{0j}_l \geq A^j_{kv}, \ v < i; \ \ A^{0j}_l \leq A^j_{kv}, \ v > i}$.
(In words, $A^{0j}_l$ inserted at its proper place in the main inequality of event 2.5.)

Finally, Event 2.7' is equal to 2.7 supplemented with the event

**$2.7^0$** Partial information about invisible routing.
For $(0, j, l) \in I_R(t)$
$\mathbf{j} \in \mathbf{d_b(t)}, \quad \mathbf{A_l^{0j}} \geq \mathbf{A^j_{b_j(t),m_j(t)}}.$

**Equivalence of Representations 1 and 2.** First we show that the event that constitutes Representation 1 implies Representation 2.

Event 2.1 follows from the definition of $T^j_{ki}$ (times of service starts and service terminations). Events 1.2, 1.4, 1.5 determine event 2.2 (the last service on a busy station have not terminated before $t$).

Then note that complete information about the values of the Backward Routing operator (event 1.6) is identical to complete information about the values of the Routing operator for the VR-times. Events 2.3-2.5 are implied by the knowledge of $\mathcal{R}(n, m, l)$ for $(n, m, l) \in V_R(t)$ and by the apparent facts written at the headings of these events. Event 2.6 follows from 1.6.

The first statement of event 2.7 means that if an invisible customer switched to a busy station, then his arrival time took place later then the arrival time of the last serviced customer. (This fact follows from the FCFS queue discipline.) The second statement means that if he switched to an idle station, then his arrival there took place after $t$. (This is the second part of the event).

Finally, events 1.7 and 1.8 coincide with events 2.8 and 2.9.

Now, for the converse, suppose all events from Representation 2 are taking place. Then arrival times and service times uniquely define service starts and terminations from Representation 1. Specifically, 2.1-2.4 and 2.6 allow service starts at $t^j_{ki}$, 2.7 guarantees that no arrivals, which can prevent terminations of busy periods at the proper times, take place during $[0, t]$ and 2.5 implies 1.6 (the proper customers start services at the proper times).

The equivalence of Representations 1' and 2' can be derived similarly.

# 4 Real-Time Estimation.

## 4.1 Stochastic Components of the Network.

As already indicated, there are four stochastic components:

- external Poisson arrivals to the network;
- service durations;
- switches of customers between the stations (Markovian routing);

- transition durations between the stations (in the case of exponential transitions).

| | Immediate Transitions | | Exponential Transitions | |
|---|---|---|---|---|
| | External Arrivals | | | |
| | Known | Unknown | Known | Unknown |
| Busy-Period Interpolation. | All stochastic components are known. | External arrivals. | Transition times. | External arrivals; transition times. |
| Real-time Estimation. General Interpolation. | Switches. | Switches; external arrivals. | Switches; transition times. | Switches; external arrivals; transition times. |

**Table 2. Unknown stochastic components.**

In Table 2 we summarize the components that are known exactly and those which are necessary to infer, for various special cases. We actually cover here all problems: those that were already analyzed (Busy-Period Interpolation), the new ones that will be analyzed in this section and those that will be only briefly discussed.

To facilitate understanding, consider, as an example, the cell that corresponds to the case of Real-Time Estimation, Immediate Transitions and Unknown External Arrivals. In this case, service times and transition times between stations (all equal to 0) are known. To estimate the queue-length, we must infer switches between stations and external arrivals.

Note that, in all cases, we have complete information about the service times up to the estimation time $t$. We see also that the second row of Table 2 differs from the first one only in the knowledge of customers' switches between the stations.

Recall that in Busy-Period Interpolation, queue estimation was carried out separately for every station; customers served over a busy period were all known, and the objective was to infer, if necessary, their arrival times (internal or external). The critical parameter, from a computational point of view, was the length of a busy period (equivalently, the number of customers served during a busy period), while the number of stations in the network had a minor effect.

In contrast, problems of real-time estimation and general interpolation require also inference about the set of customers that stand in queue at a specific station. Indeed, the current location of a specific customer, having been served, is often unknown. A basic concept here turns out to be the set of *invisible customers*, namely those customers that possibly wait in some queue at the time of estimation. The "double inference" nature of real-time problems, i.e. the need to infer both the set of customers standing in the queue, and their arrival times increases considerably the computational complexity.

## 4.2 Immediate Transitions, Known External Arrivals

Our first goal is to introduce the third representation of the observed event $E_t$, namely via stochastic components of the network. (The two other representations were introduced in Section 3.) Then $Q_j(t)$, $j \in \mathcal{D}$, will be estimated at a given observation time $t \geq 0$.

**Notation and Definitions.** Recall the notation for the *stochastic components* of the network: $A_i^0$ stand for external arrivals, $\xi_{ki}^j$ denote service durations and $R_{ml}^n \in \mathcal{D}_0$ (or $R_l^0 \in \mathcal{D}_0$) stand for switches of customers after their service-termination times $t_{ml}^n$.

Then introduce a $d$-dimensional stochastic process, $C = \{C(t),\, t \geq 0\}$, where the $j$-th coordinate of $C(t)$, $C_j(t)$ denotes the arrival time to station $j$, of the customer who is served there at time $t$. If station $j$ is idle, put $C_j(t) = t$. Using the notation of §3.2, we can write $C_j(t) = A^j_{b_j(t),m_j(t)}$ for $j \in d_b(t)$. Note that $C_j(t)$ is determined by the observed event $E_t$. Indeed, information about the history of this customer provides us with his most recent service-termination time at the predecessor station to $j$ (or with his arrival time to station 0), which is precisely $C_j(t)$ because transitions in the network are immediate.

Recall that the customers that terminated service at IR-times (see §3.2 for the definition) are called invisible customers. For every IR-time $t_{ml}^n$ we introduce the *available stations process*

$$\mathcal{M}_t(t_{ml}^n) = \{\{j : C_j(t) < t_{ml}^n\} \bigcup \{0\}\}.$$

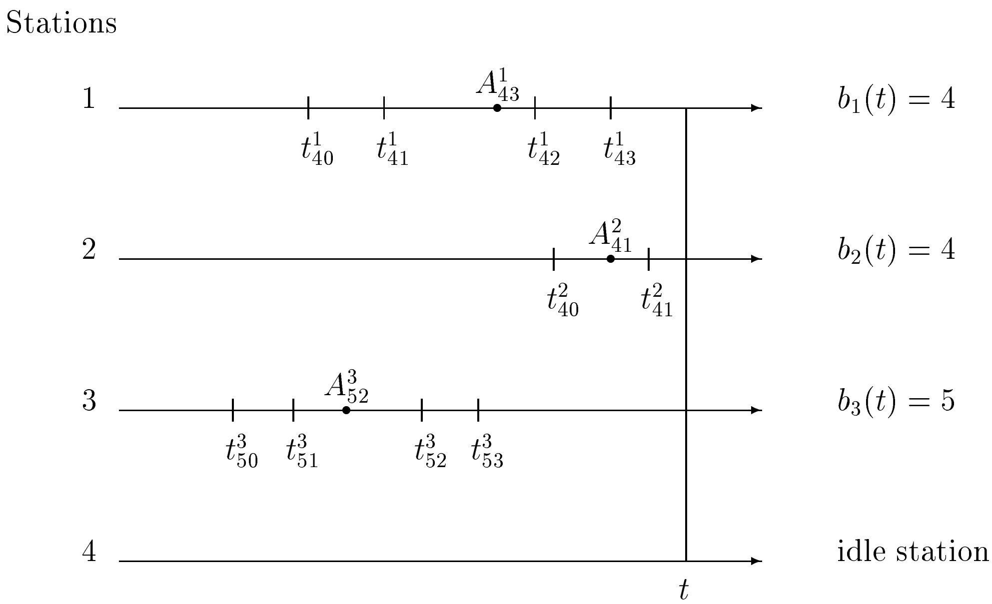


**Figure 4.1. Illustration to Definitions of $\mathbf{C_j(t)}$ and $\mathcal{M}_\mathbf{t}(\mathbf{t^n_{ml}})$**
*Transactions of the last incomplete busy periods at several stations are given. Here* $C_1(t) = A^1_{43}$, $C_2(t) = A^2_{41}$, $C_3(t) = A^3_{52}$, $C_4(t) = t$ *and, for example,* $\mathcal{M}_t(t^2_{41}) = \{0,1,2,3\}$, $\mathcal{M}_t(t^3_{52}) = \{0,3\}$, $\mathcal{M}_t(t^3_{51}) = \{0\}$.

Figure 4.1 illustrates the definition.

If a customer stands in some queue at $t$, then his most recent transaction has been either arrival to the network or service termination (it could not have been service start). If this transaction occurred before $C_j(t)$, he can not queue at station $j$ according to the FCFS queue discipline and the assumption of immediate transitions. This argument provides us with two important statements.

- Only invisible customers occupy the queues of the network.
- The invisible customer that terminated service at $t^n_{ml}$ must reside at one of the stations of $\mathcal{M}_t(t^n_{ml})$.

**Remark.** If the IR-time $t^n_{ml} < \min_{j\in\mathcal{D}} C_j(t)$ we know definitely that $\mathcal{M}(t^n_{ml}) = \{0\}$. Therefore, the corresponding invisible customer has left the network.

**Representation 3 of $\mathbf{E_t}$ via Stochastic Components.**
**Immediate Transitions, Known External Arrivals.**
To each index of an event we add parenthesized letters that correspond to the stochastic components involved in the event's description. Specifically:

**A** — external arrivals;

**S** — service times;

**R** — routing between stations;

**3.1(S)** Services that terminated up to $t$ are known.
$\xi^{\mathbf{j}}_{\mathbf{ki}} = \mathbf{t}^{\mathbf{j}}_{\mathbf{k,i+1}} - \mathbf{t}^{\mathbf{j}}_{\mathbf{ki}}, \quad j \in \mathcal{D},\ 1 \le k \le b_j(t),$
$0 \le i \le m_j(t) - 1, \quad$ for $j \in d_b(t),\ k = b_j(t),$
$0 \le i \le n^j_k, \quad$ otherwise.

**3.2(S)** Last service at a busy station have not terminated until $t$.
For $j \in d_b(t) \quad \xi^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}} > \mathbf{t} - \mathbf{t}^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}}$.

**3.3(R)** Complete information about the switches of the visible customers.
For $(n,m,l) \in V_R(t)$ and $(0,l) \in V_R(t)$
$\mathbf{R}^{\mathbf{n}}_{\mathbf{ml}} = \mathbf{j}\ \ (\mathbf{R}^{\mathbf{0}}_{\mathbf{l}} = \mathbf{j})$.

**3.4(R)** Partial information about the invisible routing.
For $(n,m,l) \in V_R(t)$ and $(0,l) \in I_R(t)$
$\mathbf{R}^{\mathbf{n}}_{\mathbf{ml}} \in \mathcal{M}_{\mathbf{t}}(\mathbf{t}^{\mathbf{n}}_{\mathbf{ml}})\ \ (\mathbf{R}^{\mathbf{0}}_{\mathbf{l}} \in \mathcal{M}_{\mathbf{t}}(\mathbf{t}^{\mathbf{0}}_{\mathbf{l}}))$.

**3.5(A)** The number of external arrivals is known.
$\mathbf{A}^{\mathbf{0}}(\mathbf{t}) = \mathbf{a}^{\mathbf{0}}(\mathbf{t})$.

**3.6(A)** The times of external arrivals are known.
$\mathbf{A_i^0} = \mathbf{t_i^0}$, $1 \le i \le a^0(t)$.

**Equivalence of Representations 2 and 3.** First note that events 3.1, 3.2, 3.3, 3.5 and 3.6 coincide with events 2.1, 2.2, 2.6, 2.8 and 2.9 respectively. The assumption of immediate transitions between the stations can be expressed as

$$R(n,m,l) = (j,k,i) \;\Rightarrow\; A^j_{ki} = \bar{A}^n_{ml} = t^n_{ml}. \tag{4.1}$$

Now recall event 2.7: for $(n,m,l) \in I_R(t)$
if $j \in d_b(t)$, $\quad R^n_{ml} = j \;\Rightarrow\; \bar{A}^n_{ml} \ge A^j_{b_j(t),m_j(t)}$,
if $j \notin d_b(t)$, $\quad R^n_{ml} = j \;\Rightarrow\; \bar{A}^n_{ml} > t$.

Since $\bar{A}^n_{ml} = t^n_{ml} \le t$, the second part of event 2.7 means that the event $\{R^n_{ml} = j\}$ cannot prevail if $j \notin d_b(t)$. The first part of 2.7 changes to:
if $j \in d_b(t)$, $\quad R^n_{ml} = j \;\Rightarrow\; t^n_{ml} \ge A^j_{b_j(t),m_j(t)} = C_j(t)$.

Using the definition of $\mathcal{M}_t$ we get the equivalence between events 2.7 and 3.4. Summarizing, we derived that 3.1-3.6 are equivalent to 2.1-2.2, 2.6-2.9. Therefore, Representation 3 follows from 2.

To prove the second direction, we note that each arrival $\bar{A}^n_{ml}$ is known and equals $t^n_{ml}$. Hence, events 2.3-2.5 trivialize. (They are either the whole sample space or the empty set.)

**Queue Estimation at Time t.** We proceed to compute the conditional expectation and the variance of the queue at station $j$.

**Proposition 4.1**

$$\mathrm{E}[Q_j(t) \mid E_t] = \sum_{t^n_{ml} \in I_R(t)} 1_{\{j \in \mathcal{M}_t(t^n_{ml})\}} \frac{p_{nj}}{\sum_{k \in \mathcal{M}_t(t^n_{ml})} p_{nk}}, \tag{4.2}$$

$$\mathrm{Var}[Q_j(t) \mid E_t] = \sum_{t^n_{ml} \in I_R(t)} 1_{\{j \in \mathcal{M}_t(t^n_{ml})\}} \frac{p_{nj}(1-p_{nj})}{(\sum_{k \in \mathcal{M}_t(t^n_{ml})} p_{nk})^2}. \tag{4.3}$$

*Proof.*

$$\begin{aligned}
\mathrm{E}[Q_j(t) \mid E_t] &= \mathrm{E}[\sum_{t^n_{ml} \in I_R(t)} 1_{\{R^n_{ml} = j\}} \mid E_t] && (4.4)\\
&= \sum_{t^n_{ml} \in I_R(t)} \mathrm{P}\{R^n_{ml} = j \mid E_t\} \\
&= \sum_{t^n_{ml} \in I_R(t)} \mathrm{P}\{R^n_{ml} = j \mid R^n_{ml} \in \mathcal{M}_t(t^n_{ml})\} && (4.5)\\
&= \sum_{t^n_{ml} \in I_R(t)} 1_{\{j \in \mathcal{M}(t^n_{ml})\}} \frac{p_{nj}}{\sum_{k \in \mathcal{M}(t^n_{ml})} p_{nk}} && (4.6)
\end{aligned}$$

Formula (4.4) is true since only invisible customers occupy the queues of the network. Routing of the invisible customers is involved in Representation 3 only in event 3.4. This fact and the independence of the stochastic components of the network imply (4.5). Finally, (4.6) follows from a straightforward calculation.

Formula (4.3) prevails since transitions between the stations are conditionally independent given $E_t$. □

**Remark.** Proposition 4.1 provides the estimate of queue-length at some fixed time $t$ (static estimate). In general, it is not desirable to calculate the estimate separately for each estimation time. A dynamic algorithm for queue re-estimation at successive estimation times is needed (see real-time algorithm in [2], for example).

A dynamic algorithm for the current problem was presented in [19]. We omit it here since the static formulae (4.2) and (4.3) are not complicated and the derivation of the dynamic algorithm is straightforward.

**Multi-Server Case.** All our results remain valid for the multi-server case. The only alteration concerns the definition of process $C$. With multi-servers, $C_j(t)$ denotes the arrival time to station $j$, of the customer who has started service last among all customers currently being served there.

## 4.3 Immediate Transitions, Unknown External Arrivals

**Notation and Definitions.** In this paragraph, it will be convenient to introduce a separate Poisson arrival process to every station and not to consider station 0. The corresponding external arrival rates are denoted by $\lambda_j, j \in \mathcal{D}$. In the case of immediate transitions and constant arrival rates. the two models are equivalent. (See the end of the paragraph for a brief discussion on the non-homogeneous case.)

Let $d_b^0(t) \subseteq d_b(t)$ denote the following subset of busy stations: station $j \in d_b^0(t)$ if the last service that starts at station $j$ before $t$ corresponds to an external customer.

As in §4.2, we introduce a $d$-dimensional stochastic process $C = \{C(t),\ t \geq 0\}$ but our current definition is somewhat different. For an idle station $j$ put $C_j(t) = t$. If a station is busy but no internal customer started service during the current busy period, $C_j(t) = t$ again. Otherwise, $C_j(t)$ stands for arrival time of the last internal customer, which started service at station $j$.

The available stations process for the internal IR-times $t_{ml}^n$

$$\mathcal{M}_t(t_{ml}^n) = \{\{j : C_j(t) < t_{ml}^n\} \bigcup \{0\}\}.$$

is defined identically to §4.2.

Naturally, all internal customers, which occupy the queues at time $t$, are invisible. Again, we can represent the queue at station $j$ as

$$Q_j(t) = Q_j^{\beta}(t) + Q_j^{\alpha}(t), \tag{4.7}$$

where $Q_j^{\alpha}(t)$ consists of external customers and $Q_j^{\beta}(t)$ of internal ones.

**Representation 4 of $\mathbf{E_t}$ via Stochastic Components.**
**Immediate Transitions, Unknown External Arrivals.**

**4.1(S)** Services that terminated up to $t$ are known.

$\xi^{\mathbf{j}}_{\mathbf{ki}} = \mathbf{t}^{\mathbf{j}}_{\mathbf{k,i+1}} - \mathbf{t}^{\mathbf{j}}_{\mathbf{ki}}, \quad j \in \mathcal{D},\ 1 \le k \le b_j(t),$
$0 \le i \le m_j(t) - 1, \quad \text{for } j \in d_b(t),\ k = b_j(t),$
$0 \le i \le n_k^j, \quad \text{otherwise.}$

**4.2(S)** Last service at a busy station have not terminated until $t$.

For $j \in d_b(t) \quad \xi^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}} > \mathbf{t} - \mathbf{t}^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}}.$

**4.3(A)** The start of a busy period coincides with the first arrival.

For $(0,j,l) \in V_R(t) \ \ni\ \mathcal{R}(0,j,l) = (j,k,0),$
$\mathbf{A}^{\mathbf{0j}}_{\mathbf{l}} = \mathbf{A}^{\mathbf{j}}_{\mathbf{k0}} = \mathbf{t}^{\mathbf{j}}_{\mathbf{k0}}.$

**4.4a(A)** Truncation rule for complete busy periods.

For $(0.j,l) \in V_R(t) \ \ni\ \mathcal{R}(0,j,l) = (j,k,i),\ i \ge 1,$
$\mathbf{A}^{\mathbf{0j}}_{\mathbf{l}} = \mathbf{A}^{\mathbf{j}}_{\mathbf{ki}} \le \mathbf{t}^{\mathbf{j}}_{\mathbf{ki}},$
$1 \le k \le b_j(t)$ for $j \notin d_b(t)$, and
$1 \le k \le b_j(t) - 1$ for $j \in d_b(t)$.

**4.4b(A)** Truncation rule for incomplete busy periods.

For $(0,j,l) \in V_R(t) \ \ni\ j \in d_b(t),\ \mathcal{R}(0,j,l) = (j, b_j(t), i),\ i \ge 1,$
$\mathbf{A}^{\mathbf{0j}}_{\mathbf{l}} = \mathbf{A}^{\mathbf{j}}_{\mathbf{b_j(t),i}} \le \mathbf{t}^{\mathbf{j}}_{\mathbf{b_j(t),i}}.$

**4.5a(A)** FCFS discipline for complete busy periods.

$\mathbf{t}^{\mathbf{j}}_{\mathbf{k0}} \le \mathbf{A}^{\mathbf{j}}_{\mathbf{k1}} \le \ldots \le \mathbf{A}^{\mathbf{j}}_{\mathbf{k,n^j_k}},$
$1 \le k \le b_j(t)$ for $j \notin d_b(t)$, and
$1 \le k \le b_j(t) - 1$ for $j \in d_b(t)$,
where some $A^j_{ki}$ are unknown external arrivals ($\mathcal{R}(0,j,l) = (j,k,i),\ A^{0j}_l = A^j_{ki}$) and the others are known internal ones ($\mathcal{R}(n,m,l) = (j,k,i),\ A^j_{ki} = t^n_{ml}$).

**4.5b(A)** FCFS discipline for incomplete busy periods.

$\mathbf{t}^{\mathbf{j}}_{\mathbf{b_j(t),0}} \le \mathbf{A}^{\mathbf{j}}_{\mathbf{b_j(t),1}} \le \ldots \le \mathbf{A}^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}}, \quad \mathbf{j} \in \mathbf{d_b(t)}.$

**4.6(R)** Complete information about the switches of the visible customers.

For $(n,m,l) \in V_R(t) \quad \mathbf{R}^{\mathbf{n}}_{\mathbf{ml}} = \mathbf{j}.$

**4.7(RA)** Partial information about the invisible routing.

**4.7a** For $(n,m,l) \in I_R(t)$
$\mathbf{R^n_{ml} = j \ \Rightarrow \ j \in \mathcal{M}_t(t^n_{ml})}$.

**4.7b** If $j \in d^0_b(t)$, $(n,m,l) \in I_R(t)$
$\mathbf{R^n_{ml} = j \ \Rightarrow \ t^n_{ml} \geq A^j_{b_j(t),m_j(t)}}$.

**4.7c** Invisible external arrivals.
Let $N^0_j(t)$ denote the number of visible external customers that arrived at station $j$ during $[0,t]$. Then
$\mathbf{j \in d^0_b(t) \ \Rightarrow \ A^{0j}_{N^0_j(t)+1} > A^j_{b_j(t),m_j(t)}}$.
$\mathbf{j \in \{d_b(t) \setminus d^0_b(t)\} \ \Rightarrow \ A^{0j}_{N^0_j(t)+1} > C_j(t)}$.
$\mathbf{j \notin d_b(t) \ \Rightarrow \ A^{0j}_{N^0_j(t)+1} > t}$.

**Equivalence of Representations 2' and 4.**
Events 2.1', 2.2' and 2.6' coincide with events 4.1, 4.2 and 4.6 respectively. Similar to Representation 3, events 2.3-2.5 trivialize. Their supplements $2.3^0$-$2.5^0$ coincide with 4.3-4.5.

Now we prove identity of event 2.7 and the intersection of events 4.7a and 4.7b. First, consider $j \notin d_b(t)$. In the case of immediate transitions between stations, the event, described in the second line of 2.7, turns out to be $\{R^n_{ml} \neq j\}$, since $\bar{A}^n_{ml} = t^n_{ml}$ and $t^n_{ml} \leq t$. Event 4.7a turns out to be the same event $\{R^n_{ml} \neq j\}$, due to the definition of $\mathcal{M}_t$ and the equality $C_j(t) = t$.

If $j \in d_b(t)$, the equivalence between the event, described in the first line of 2.7, and the intersection of 4.7a and 4.7b follows from the inequality $C_j(t) \leq A^j_{b_j(t),m_j(t)}$.

To derive equivalence of the events $2.7^0$ and 4.7c, note that the triple $(0,j,l)$ corresponds to the invisible-routing time $A^{0j}_l$ if and only if $l > N^0_j(t)$ and $A^{0j}_l \leq t$. □

**Queue estimation at time t. Relevant information.** According to formula (4.7),

$$\mathrm{E}[Q_j(t) \mid E_t] = \mathrm{E}[Q^\beta_j(t) \mid E_t] + \mathrm{E}[Q^\alpha_j(t) \mid E_t]. \tag{4.8}$$

Consider the two terms of (4.8) separately. Define events

$$\begin{aligned} H^1_t &= \{E^1_t, E^2_t, E^3_t, E^{4a}_t, E^{5a}_t, E^6_t\} \\ H^2_t &= \{E^{4b}_t, E^{5b}_t, E^{7a}_t, E^{7b}_t, E^{7c}_t\} \end{aligned}$$

where $E^X_t$ correspond to event X from Representation 4.

Then

$$\begin{aligned}
\mathrm{E}[Q_j^{\beta}(t) \mid E_t] &= \sum_{t_{ml}^n \in I_R(t)} \mathrm{P}\{R_{ml}^n = j \mid E_t\} \\
&= \sum_{t_{ml}^n \in I_R(t)} \frac{\mathrm{P}\{R_{ml}^n = j, H_t^2 | H_t^1\}}{\mathrm{P}\{H_t^2 | H_t^1\}} \\
&= \sum_{t_{ml}^n \in I_R(t)} \frac{\mathrm{P}\{R_{ml}^n = j, H_t^2\}}{\mathrm{P}\{H_t^2\}} \qquad (4.9) \\
&= \sum_{t_{ml}^n \in I_R(t)} \mathrm{P}\{R_{ml}^n = j / E_t^{4b}, E_t^{5b}, E_t^{7a}, E_t^{7b}, E_t^{7c}\}
\end{aligned}$$

Formula (4.9) follows from independence of the stochastic components and conditional independence of external arrivals from different busy periods given starts and terminations of these busy periods.

We treat the second term of (4.8) similarly.

$$\mathrm{E}[Q_j^{\alpha}(t) \mid E_t] = \mathrm{E}[Q_j^{\alpha}(t) \mid E_t^{4b}, E_t^{5b}, E_t^{7a}, E_t^{7b}, E_t^{7c}]$$

$$= \begin{cases} \mathrm{E}[\#\text{ (Poisson arrivals in } [A^j_{b_j(t),m_j(t)}, t]) \mid E_t^{4b}, E_t^{5b}, E_t^{7a}, E_t^{7b}], & j \in d_b^0 \\ \mathrm{E}[\#\text{ (Poisson arrivals in } [C_j(t), t])\ ], & j \in d_b \setminus d_b^0 \\ 0, & j \notin d_b \end{cases} \qquad (4.10)$$

(Due the memoryless property of the exponential distribution, the increments of the Poisson processes, considered in (4.10), are independent of other stochastic components.)

Now we shall simplify notation and get a convenient representation of the relevant information.

Note that the external arrival times which took place before $C_j(t)$, are conditionally independent of those which happened after $C_j(t)$, given $C_j(t)$ is known (and it is known). Therefore, the parts of events 4.4b and 4.5b that correspond to arrivals before $C_j(t)$ are irrelevant.

Suppose that there are $K$ invisible customers at time $t$, all in all. Let $s_k$, $1 \le k \le K$, denote their most recent service-termination times. Let $R_k$, $1 \le k \le K$, stand for the station where they switch to next.

Consider $j \in d_b^0(t)$. Let $u_j$ denote the service-start time of the internal customer, which arrived at $C_j(t)$. (Till the end of the current section, we omit the time index whenever convenient.) The transactions at station $j$ between $u_j$ and $t$ are denoted by $u_j < t_1^j < \ldots < t_{m_j}^j$. We denote $t_0^j = C_j(t)$ (as though the busy period starts at $C_j(t)$).

Let $A_1^j, \ldots, A_{m_j}^j$ denote the successive times of external arrivals to station $j$, which correspond to the service starts $t_1^j, \ldots, t_{m_j}^j$. Due to the memoryless property of exponential random variables, we can assume that they are the jump points of a Poisson process starting at $C_j(t) = t_0^j$.

Then the relevant information is given by the intersection of the 3 events:

**4.4b** $A_1^j \le t_1^j, \dots, A_{m_j}^j \le t_{m_j}^j, \quad j \in d_b^0;$

**4.7a** $R_k = j \ \Rightarrow \ j \in \mathcal{M}_k, \quad 1 \le k \le K.$

**4.7b** For $j \in d_b^0, \quad R_k = j \ \Rightarrow \ A_{m_j}^j \le s_k.$

Note that in the event 4.4b, we consider only those external arrivals which took place after $C_j(t)$ (they are independent of the previous external arrival times given $C_j(t)$). The event 4.5b prevails since $A_1^j \le A_2^j \le \dots \le A_{m_j}^j$ by their definition. Event 4.7c prevails since invisible external arrivals constitute a Poisson process starting at $A_{b_j(t),m_j(t)}^j$.

**Queue estimation at time $t$.** We start with some additional definitions, proceed with the description of the algorithm and, finally, prove its correctness. Here we treat $t$ fixed, and there seems to be no way to recalculate the estimate for time $t+h$ with significant reduction in the computational effort.

Let $B$ denote the event 4.4b. Then

$$B = \bigcap_{j \in d_b^0} B_j,$$

where $B_j = \{A_1^j \le t_1^j, \dots, A_{m_j}^j \le t_{m_j}^j\}, \ \ j \in d_b^0$. We also use the notation

$$B^j \stackrel{\text{def}}{=} B \setminus B_j.$$

Consider the event 4.7b. It can be represented as

$$D = \bigcap_{k=1}^{K} \bigcap_{j \in d_b} D_{kj}, \tag{4.11}$$

where

$$D_{kj} = \begin{cases} \{R_k = j \ \Rightarrow \ C_j \le s_k\} = \{R_k \ne j\} \bigcup \{j \in \mathcal{M}_k\}, & j \in d_b \setminus d_b^0 \\ \{R_k = j \ \Rightarrow \ A_{m_j}^j \le s_k\} = \{R_k \ne j\} \bigcup \{A_{m_j}^j \le s_k\}, & j \in d_b^0 \end{cases} \tag{4.12}$$

Note that $D$ is an intersection of events that are organized in matrix form (the rows correspond to invisible customers and the columns to stations). Observe that events, which do not belong to the same row or to the same column, are independent. We also define

$$D^{lm} \stackrel{\text{def}}{=} \bigcap_{k \ne l} \bigcap_{j \ne m} D_{kj}; \quad D^{l*} \stackrel{\text{def}}{=} \bigcap_{k \ne l} \bigcap_{j \in d_b^0} D_{kj}; \quad D^{*m} \stackrel{\text{def}}{=} \bigcap_{k=1}^{K} \bigcap_{j \ne m} D_{kj}.$$

Here $D^{lm}$ equals the event $D$ after removal of the $l$-th row and the $m$-th column. If only the $l$-th row was removed we get $D^{l*}$ and, finally, $D^{*m}$ corresponds to $D$ after removal of the $m$-th column.

To simplify conditioning on the event 4.7a, define the routing matrix "invisible customers $\to$ stations". It is denoted by $\tilde{R} = (\tilde{r}_{mu},\ 1 \le m \le K,\ u \in \mathcal{D}_0)$, where

$$\tilde{r}_{mu} = 1_{\{u \in \mathcal{M}_m\}} \frac{p_{N_m,u}}{\sum_{l \in \mathcal{M}_m} p_{N_m,l}} \tag{4.13}$$

(Here $N_m$ denotes the station at which customer $m$ have terminated his most recent service). The conditioning on the event 4.7a is equivalent to the assumption that the routing of invisible customers is governed by the matrix $\tilde{R}$ (instead of the routing matrix $P$). It is convenient to denote conditioning on 4.7a by $\mathrm{P}_{\tilde{R}}$ or $\mathrm{E}_{\tilde{R}}$.

Now define

$$\omega_{kj} = \mathrm{P}\{A^j_{m_j} \le s_k \mid B_j\}, \tag{4.14}$$

and, finally, introduce the events $F_{kj},\quad 1 \le k \le K,\ j \in \mathcal{D}$, meaning "the invisible customer number $k$ was the first invisible customer that switched to station $j$". Event $F_{0j},\ j \in \mathcal{D}$, corresponds to "no invisible customer switched to station $j$".

**Remark.** In the following algorithm, we use results from the two alternative sources. The first is Bertsimas and Servi [2]. The second is Algorithm 5.1 in §5.1 later. (This algorithm is verified independently of the present analysis.)

**Algorithm 4.1. Real-Time Estimation.**
**Immediate Transitions, Unknown External Arrivals.**

1°. Basic queue decomposition.

$$\mathrm{E}[Q_j(t) \mid E_t] = \mathrm{E}_{\tilde{R}}[Q^\alpha_j(t) \mid B; D] + \mathrm{E}_{\tilde{R}}[Q^\beta_j(t) \mid B; D]. \tag{4.15}$$

2°. Calculating $\omega_{kj}$.

$$\begin{aligned} \omega_{kj} &= \frac{\mathrm{P}\{A^j_{m_j} \le s_k; B_j\}}{\mathrm{P}\{B_j\}} = \\ &= \frac{\mathrm{P}\{A^j_1 \le \min(s_k, t^j_1); A^j_2 \le \min(s_k, t^j_2); \ldots; A^j_{m_j} \le \min(s_k, t^j_{m_j})\}}{\mathrm{P}\{B_j\}}. \end{aligned} \tag{4.16}$$

The numerator of (4.16) and $\mathrm{P}\{B_j\}$ can be calculated using definition (5.1) and formulae (5.3–5.6). (Formula (46) from Bertsimas and Servi [2] provides an alternative calculation).

3°. Calculating $\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\}$.

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{1j}\} = \tilde{r}_{1j};\ \ \mathrm{P}_{\tilde{\mathrm{R}}}\{F_{0j}\} = \prod_{l=1}^{K}(1-\tilde{r}_{lj});\ \ \mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\} = \tilde{r}_{kj}\prod_{l=1}^{k-1}(1-\tilde{r}_{lj}),\ \ 2 \le k \le K. \tag{4.17}$$

4°. Calculating the internal queue.

$$\mathrm{E}_{\tilde{\mathrm{R}}}[Q^\beta_j(t) \mid D; B] = \frac{\sum_{k=1}^K \mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; D \mid B\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B\}}. \tag{4.18}$$

The denominator of (4.18) is computed recursively by:

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B\} = \sum_{l \in d_b^0} \tilde{r}_{1l}\omega_{1l}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l} \mid B^l\} + \sum_{l \notin d_b^0} \tilde{r}_{1l}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1*} \mid B\}, \tag{4.19}$$

starting with $k = 1$:

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B\} = \sum_{l \in d_b^0} \tilde{r}_{1l}\omega_{1l} + \sum_{l \notin d_b^0} \tilde{r}_{1l}. \tag{4.20}$$

Finally, the numerator of (4.18) is calculated by

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; D \mid B\} = \tilde{r}_{kj}\omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*} \mid B^j; B_j \bigcap \{A^j_{m_j} \le s_k\}\}, \tag{4.21}$$

where $\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*} \mid B^j; B_j \bigcap \{A^j_{m_j} \le s_k\}\}$ is computed recursively using (4.19).

5°. Calculating the external queue. For $j \in d_b \setminus d_b^0$,

$$\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B; D] = \lambda_j(t - C_j(t)), \tag{4.22}$$

and the conditional distribution of the external queue $Q_j^\alpha(t)$ is Poisson with parameter $\lambda_j(t - C_j(t))$.

For $j \in d_b^0$,

$$\begin{aligned}\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B; D] &= \frac{1}{\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B\}}[\sum_{k=1}^{K} \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B_j; A^j_{m_j} \le s_k]\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\}\omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}_k}\{D^{kj} \mid B^j\} + \\ &+ \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B_j]\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{0j}\}\mathrm{P}_{\tilde{\mathrm{R}}_0}\{D^{*j} \mid B^j\}].\end{aligned} \tag{4.23}$$

Here $\tilde{R}_k = \tilde{R} \mid F_{kj}$, i.e. if $k \ne 0$, $m < k$, we must replace in (4.15) $\mathcal{M}_m$ by $\mathcal{M}_m \setminus \{j\}$. If $k = 0$, $\mathcal{M}_m$ must be replaced by $\mathcal{M}_m \setminus \{j\}$, for $1 \le m \le K$. The expression $\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B_j; A^j_{m_j} \le s_k]$ can be computed using formulae (5.2)–(5.7) from Algorithm 5.1 or the algorithm in §4.1 of Bertsimas and Servi [2].

*Proof of correctness and comments.*

1°. Formula (4.17) is equivalent to (4.8).

2°. Formula (4.16) follows from the definition of $\omega_{kj}$.

3°. The following straightforward representation of events $F_{kj}$ implies formula (4.17):

$$F_{1j} = \{R_1 = j\}; \quad F_{0j} = \bigcap_{l=1}^{K}\{R_l \ne j\}; \quad F_{kj} = \{\bigcap_{l=1}^{k-1}\{R_l \ne j\}\}\bigcap\{R_k = j\}, \quad 2 \le k \le K.$$

4°. Formula (4.18) follows from

$$Q_j^\beta(t) = \sum_{k=1}^{K} 1_{\{R_k = j\}}.$$

In order to derive formula (4.19), we condition on the switch of the first invisible customer (recall that $s_1 < s_2 < \ldots < s_K$) :

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B\} = \sum_{l \in \mathcal{D}_0} \mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B; R_1 = l\} \tilde{r}_{1l}.$$

The following calculation is typical for this section. It is carried out in detail only once, and further similar calculations will be presented briefly.

If $l \in d_b^0$

$$\begin{aligned}
\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B; R_1 = l\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\left\{\bigcap_{k=1}^{K} \bigcap_{j \in d_b^0} \{\{R_k \neq j\} \bigcup \{A^j_{m_j} \leq s_k\}\} \mid B; R_1 = l\right\} \\
&= \frac{\mathrm{P}_{\tilde{\mathrm{R}}}\left\{A^l_{m_l} \leq s_1; \bigcap\bigcap_{(k,j)\neq(1,l)} \{\{R_k \neq j\} \bigcup \{A^j_{m_j} \leq s_k\}\}; R_1 = l; B\right\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l; B\}} \\
&= \frac{\mathrm{P}_{\tilde{\mathrm{R}}}\{A^l_{m_l} \leq s_1; \bigcap_{k=2}^{K} \bigcap_{j \neq l} D_{kj}; R_1 = l; B^l; B_l\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l; B^l; B_l\}} \qquad (4.24) \\
&= \frac{\mathrm{P}_{\tilde{\mathrm{R}}}\{A^l_{m_l} \leq s_1; B_l\} \mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l\} \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l}; B^l\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l\} \mathrm{P}_{\tilde{\mathrm{R}}}\{B_l\} \mathrm{P}_{\tilde{\mathrm{R}}}\{B^l\}} \\
&= \mathrm{P}_{\tilde{\mathrm{R}}}\{A^l_{m_l} \leq s_1 \mid B_l\} \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l} \mid B^l\} \\
&= \omega_{1l} \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l} \mid B^l\}.
\end{aligned}$$

Formula (4.24) prevails since the event $\{R_1 = l\}$ implies $\{R_1 \neq j, \;\; j \neq l\}$ and the event $\{A^l_{m_l} \leq s_1\}$ implies $\{A^l_{m_l} \leq s_k, \;\; 2 \leq k \leq K\}$.

Then, in the same way, we prove that for $j \in d_b \setminus d_b^0$,

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B; R_1 = l\} \;=\; \mathrm{P}_{\tilde{\mathrm{R}}}\{\bigcap_{k=2}^{K} \bigcap_{j \in d_b^0} D_{kj} \mid B\} \;=\; \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1*} \mid B\}.$$

As for formula (4.20), note that if $k = 1$

$$\begin{aligned}
\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{\bigcap_{j \in d_b^0} D_{1j} \mid B\} \\
&= \sum_{l \in d_b} \mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l; \bigcap_{j \in d_b^0} D_{1j} \mid B\} \;=\; \sum_{l \in d_b^0} \tilde{r}_{1l}\omega_{1l} + \sum_{l \neq d_b^0} \tilde{r}_{1l}.
\end{aligned}$$

Note, that all these computations are in some way similar to the computation of a matrix determinant.

Finally, (4.21) is derived from the following calculation:

$$\begin{aligned}
\mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; D \mid B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; \bigcap_{l \neq k} \bigcap_{j \in d_b^0} D_{kj}; A^j_{m_j} \leq s_k \mid B\} \\
&= \tilde{r}_{kj} \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*}; A^j_{m_j} \leq s_k \mid B\} \\
&= \tilde{r}_{kj}\omega_{kj} \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*} \mid A^j_{m_j} \leq s_k; B\} \\
&= \tilde{r}_{kj}\omega_{kj} \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*} \mid B^j; B_j \bigcap \{A^j_{m_j} \leq s_k\}\}.
\end{aligned}$$

5°. Formula (4.22) follows from (4.10).

For $j \in d_b^0$, we condition on the first internal customer which arrived to station $j$:

$$\mathrm{E}_{\tilde{R}}[Q_j^\alpha(t) \mid B; D] = \sum_{k=0}^{K} \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid F_{kj}; B; D] \mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj} \mid B; D\}. \tag{4.25}$$

If $k \neq 0$,

$$\begin{aligned} \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid F_{kj}; B; D] &= \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid R_k = j; \bigcap_{l<k} R_l \neq j; B; D] \\ &= \mathrm{E}_{\tilde{\mathrm{R}}}\left[Q_j^\alpha(t) \mid R_k = j; \bigcap_{l<k} \{R_l \neq j\}; B_j \bigcap B^j; A_{m_j}^j < s_k; D^{kj}\right] \qquad (4.26) \\ &= \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B_j; A_{m_j}^j < s_k]. \qquad (4.27) \end{aligned}$$

Formula (4.26) prevails since $\{R_l \neq j\}$ implies $D_{lj}$, $l < k$, and $\{A_{m_j}^j < s_k\}$ implies $D_{lj}$, $l > k$. Formula (4.27) is true since $Q_j^\alpha(t)$ is the number of Poisson arrivals in $[A_{m_j}^j, t]$, which depends only on the events that are left in the condition of (4.26). If $k = 0$,

$$\begin{aligned} \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid F_{0j}; B; D] &= \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid \bigcap_{l \in d_b^0} \{R_l \neq j\}; B; D] \\ &= \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid \bigcap_{l \in d_b^0} \{R_l \neq j\}; B; D^{*j}] = \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B_j]. \qquad (4.28) \end{aligned}$$

Now consider the second term of (4.25):

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}; D \mid B\} = \mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid F_{kj}; B\}\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj} \mid B\} = \mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid F_{kj}; B\}\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\}, \tag{4.29}$$

since events $F_{kj}$ and $B$ are independent. Then if $k \neq 0$:

$$\begin{aligned} \mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid F_{kj}; B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid \bigcap_{l<k} \{R_l \neq j\}; R_k = j; B\} = \\ &= \mathrm{P}_{\tilde{\mathrm{R}}}\left\{A_{m_j}^j \leq s_k; D^{kj} \mid \bigcap_{l<k} \{R_l \neq j\}; R_k = j; B_j \bigcap B^j\right\} = \\ &= \omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{kj} \mid \bigcap_{l<k} \{R_l \neq j\}; B^j\} = \omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}_k}\{D^{kj} \mid B^j\}. \qquad (4.30) \end{aligned}$$

Similarly, for $k = 0$:

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D \mid F_{0j}; B\} = \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{*j} \mid F_{0j}; B\} = \mathrm{P}_{\tilde{\mathrm{R}}_0}\{D^{*j} \mid B^j\}. \tag{4.31}$$

Summarizing (4.25)–(4.31) we deduce (4.23).

**Non-Homogeneous External Arrival Rate.**

Assume time-dependent external arrival rate $\alpha_j(t)$ to station $j$. Then formula (4.22) of Algorithm 4.1 must be modified to:

$$\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^\alpha(t) \mid B; D] = \int_{C_j(t)}^{t} \alpha_j(t)dt.$$

The formulae from Bertsimas and Servi [2], that were used for the calculation of (4.16) and (4.23), are valid for the non-homogeneous case. The alternative approach involves formulae (5.2)–(5.7) from §5.1. In this case, the time-change $t \to \Lambda_j(t) = \int_0^t \alpha_j(s)\mathrm{d}s$ must be performed.

All other formulae from Algorithm 4.1 remain intact.

### 4.4 Exponential Transitions between Stations. Known and Unknown External Arrivals.

The case of exponential transitions and known external arrivals was analyzed in [19] within the HMM framework. However, the Markov chain in this case is considerably more complicated than that corresponding the case of Busy-Period Interpolation. (Here we cannot focus on a specific station and all stations must be analyzed simultaneously.)

The complexity of the algorithm increases very rapidly when the number of invisible customers increases. Therefore, it is impractical to apply it to "large" systems, and the need arises for approximations. We leave this topic for future research.

As for the case of real-time estimation with exponential transitions and unknown external arrivals, a Hidden Markov Model can be also constructed. The state space of HMM and, consequently, notation will be very cumbersome. Indeed, here the state space will be infinite, due to an unrestricted number of external invisible customers. Again any practical implementation of the algorithm must involve approximations and truncation of the state space.

## 5 General Interpolation

### 5.1 Single-Station Case

Bertsimas and Servi [2] carried out the analysis of the Real-Time problem for a single-station. They have not considered the General Interpolation problem (queue path reconstruction during a busy period that has not terminated yet), which we now do.

Assume that the busy period under consideration begins at time 0. The times $t_0 = 0, t_1, \ldots, t_n, \ldots$ denote the realizations of service starts. As before, $A_0 = 0, A_1, \ldots, A_n$, stand for arrivals to the station and $A(t)$ is the cumulative number of arrivals in $(0, t]$. Arrivals to the station constitute a homogeneous Poisson process with rate $\lambda$. (The case of non-homogeneous Poisson input may be treated, as before, performing the time change $t \to \Lambda(t) = \int_0^t \alpha(s)\mathrm{d}s$.) Introduce the event

$$B_n \;=\; \{A(t_1) \geq 1, A(t_2) \geq 2, \ldots, A(t_n) \geq n\} \;=\; \{A_1 \leq t_1, A_2 \leq t_2, \ldots, A_n \leq t_n\}, \quad (5.1)$$

which means that our busy period has not terminated up to $t_n$.

The objective is the recursive calculation of $\mathrm{E}\,[A(t) \mid B_{n+1}]$, for a fixed $t > 0$. Specifically, we update our estimate at successive service-termination times.

Bertsimas and Servi [2] assumed that $t = t_n$ (real-time estimation) or $t > t_n$ (extrapolation), while we present an algorithm for general $t$ and clarify the difference between extrapolation and interpolation ($t \leq t_n$ for the latter).

Define the events

$$C_n = \{A(t_1) \geq 1, A(t_2) \geq 2, \ldots, A(t_{n-1}) \geq n-1, A(t_n) = n\}$$

and

$$D_n = C_n \bigcap \{A(t_{n+1}) = n\}.$$

**Algorithm 5.1. General Interpolation. Single-station case.**

— Main formula:

$$\mathrm{E}\,[A(t) \mid B_{n+1}] \;=\; \frac{\mathrm{E}\,[A(t) \mid B_n]\mathrm{P}\{B_n\} - \mathrm{E}\,[A(t) \mid D_n]\mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}}{\mathrm{P}\{B_n\} - \mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}}. \quad (5.2)$$

$$\mathrm{E}\,[A(t) \mid B_0] \;=\; \lambda t.$$

— Recursive calculation of $\mathrm{P}\{B_{n+1}\}$:

$$\mathrm{P}\{B_{n+1}\} = \mathrm{P}\{B_n\} - \mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}. \quad (5.3)$$

$$\mathrm{P}\{B_0\} = 0.$$

— Recursive calculation of $\mathrm{P}\{C_{n+1}\}$:

$$\mathrm{P}\{C_{n+1}\} = e^{-\lambda t_{n+1}}\lambda^{n+1} g(t_1, t_2, \ldots, t_n, t_{n+1}), \quad (5.4)$$

where

$$g(t_1, t_2, \ldots, t_n, t_{n+1}) \stackrel{\mathrm{def}}{=} \int_{x_1=0}^{t_1}\int_{x_2=x_1}^{t_2}\cdots\int_{x_n=x_{n-1}}^{t_n}\int_{x_{n+1}=x_n}^{t_{n+1}} dx_{n+1}dx_n \ldots dx_2 dx_1 \quad (5.5)$$

and

$$g(t_1, t_2, \ldots, t_n, t_{n+1}) = \sum_{k=0}^{n}(-1)^{n-k}\frac{t_{k+1}^{n+1-k}}{(n+1-k)!}g(t_1, t_2, \ldots, t_k). \quad (5.6)$$

$$\mathrm{P}\{C_0\} = g_0 = 1.$$

— Calculation of $\mathrm{E}\,[A(t) \mid D_n]$ If $t \geq t_n$ (extrapolation)

$$\mathrm{E}\,[A(t) \mid D_n] = n + \lambda(t - t_{n+1})_+. \tag{5.7}$$

If $t < t_n$ (interpolation), specifically if $t_i \leq t < t_{i+1},\ i \leq n-1$ :

$$\begin{aligned}\mathrm{E}\,[A(t) \mid D_n] = \frac{1}{\mathrm{P}\{C_n\}} e^{-\lambda t_n} \lambda^n [\sum_{k=i}^{n} \frac{t^k}{(k-1)!}\, g(t_{k+1}-t, t_{k+2}-t, \ldots, t_n - t) \ - \\ - \ \sum_{k=i}^{n} k\, g(t_{k+1}-t, t_{k+2}-t, \ldots, t_n - t)] \sum_{m=0}^{i-1} \frac{(t-t_{m+1})^{k-m}}{(k-m)!} g(t_1, \ldots, t_m).\end{aligned} \tag{5.8}$$

**Remark.** Formulae (5.2)–(5.6) demonstrate that $\mathrm{P}(B_{n+1})$ and $\mathrm{P}(C_{n+1})$ can be computed recursively using $O(n)$ operations for every recalculation of each event.

*Verification of Algorithm 5.1.*
Formula (5.2) prevails since

$$\begin{aligned}\mathrm{E}\,[A(t) \mid B_{n+1}] &= \mathrm{E}\,[A(t) \mid \{B_n \setminus D_n\}] = \\ &= \frac{\mathrm{E}\,[A(t) \mid B_n]\mathrm{P}\{B_n\} - \mathrm{E}\,[A(t) \mid D_n]\mathrm{P}\{D_n\}}{\mathrm{P}\{B_n\} - \mathrm{P}\{D_n\}} = \\ &= \frac{\mathrm{E}\,[A(t) \mid B_n]\mathrm{P}\{B_n\} - \mathrm{E}\,[A(t) \mid D_n]\mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}}{\mathrm{P}\{B_n\} - \mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}}.\end{aligned}$$

Formula (5.3) follows from the definitions of $B_n$, $C_n$ and $D_n$, the equalities

$$B_{n+1} = B_n \setminus \{C_n \bigcap \{A(t_{n+1}) - A(t_n) = 0\}\} = B_n \setminus D_n,$$

and the independence between $C_n$ and $A(t_{n+1}) - A(t_n)$.
Formula (5.4) prevails since

$$\begin{aligned}\mathrm{P}\{C_{n+1}\} &= \mathrm{P}\{A_1 \leq t_1, \ldots, A_n \leq t_n, A_{n+1} \leq t_{n+1}, A(t_{n+1}) = n+1\} \\ &= \mathrm{P}\{A_1 \leq t_1, \ldots, A_n \leq t_n, A_{n+1} \leq t_{n+1} \mid A(t_{n+1}) = n+1\} e^{-\lambda t_{n+1}} \frac{(\lambda t_{n+1})^{n+1})}{(n+1)!} \\ &\qquad \text{(conditional joint density of uniform order statistics is } \frac{(n+1)!}{t_{n+1}^{n+1}}) \\ &= e^{-\lambda t_{n+1}} \lambda^{n+1} g(t_1, t_2, \ldots, t_n, t_{n+1}).\end{aligned}$$

Formula (5.6) is almost identical to formula 42 (page 225) of Bertsimas and Servi [2]. We present a simpler proof, by computing multidimensional integrals according to the definition (5.5).

$$
\begin{aligned}
&g(t_1, t_2, \ldots, t_n, t_{n+1}) \\
&= \int_{x_1=0}^{t_1} \int_{x_2=x_1}^{t_2} \cdots \int_{x_n=x_{n-1}}^{t_n} (t_{n+1} - x_n) dx_n \ldots dx_1 \\
&= t_{n+1} g(t_1, \ldots, t_n) - \int_{x_1=0}^{t_1} \int_{x_2=x_1}^{t_2} \cdots \int_{x_{n-1}=x_{n-2}}^{t_{n-1}} \frac{t_n^2 - x_{n-1}^2}{2} dx_{n-1} \ldots dx_1 \\
&= t_{n+1} g(t_1, \ldots, t_n) - \frac{t_n^2}{2} g(t_1, \ldots, t_{n-1}) + \int_{x_1=0}^{t_1} \cdots \int_{x_{n-2}=x_{n-3}}^{t_{n-2}} \frac{t_{n-1}^3 - x_{n-2}^3}{3!} dx_{n-2} \ldots dx_1 \\
&= \ldots = \sum_{k=0}^{n} (-1)^{n-k} \frac{t_{k+1}^{n+1-k}}{(n+1-k)!} g(t_1, t_2, \ldots, t_k).
\end{aligned}
$$

Formula (5.7) holds since, for $t \geq t_n$:

$$
\begin{aligned}
\mathrm{E}\,[A(t) \mid D_n] &= \mathrm{E}\,[A(t_n) \mid D_n] + \mathrm{E}\,[(A(t) - A(t_n)) \mid D_n] = \\
&= n + \lambda(t - t_{n+1})_+.
\end{aligned}
$$

We still need to verify (5.8). For $t_i \leq t < t_{i+1} \leq t_n$,

$$\mathrm{E}\,[A(t) \mid D_n] = \mathrm{E}\,[A(t) \mid C_n].$$

Then

$$
\begin{aligned}
\mathrm{E}\,[A(t) \mid C_n] &= \sum_{k=i}^{n} k \mathrm{P}\{A(t) = k \mid C_n\} = \\
&= \frac{1}{\mathrm{P}\{C_n\}} \sum_{k=i}^{n} k \mathrm{P}\{A(t) = k\} \mathrm{P}\{C_n \mid A(t) = k\} = \\
&= \frac{1}{\mathrm{P}\{C_n\}} \sum_{k=i}^{n} e^{-\lambda t} \frac{(\lambda t)^k}{(k-1)!} \mathrm{P}\{B_i \mid A(t) = k\} \mathrm{P}\{C_n \setminus B_k \mid A(t) = k\}. \qquad (5.9)
\end{aligned}
$$

Formula (5.9) follows from the conditional independence of $B_i$ and $C_n \setminus B_i$, given $A(t)$, combined with

$$\mathrm{P}\{C_n \setminus B_i \mid A(t) = k\} = \mathrm{P}\{C_n \setminus B_k \mid A(t) = k\}.$$

We now calculate separately the two conditional probabilities in (5.9).

$$
\begin{aligned}
\mathrm{P}\{B_i \mid A(t) = k\} &= \mathrm{P}\{A_1 \leq t_1, \ldots, A_i \leq t_i, \ldots, A_{i+1} \leq t, A_k \leq t\} \\
&= \frac{k!}{t^k} \int_{x_1=0}^{t_1} \int_{x_2=x_1}^{t_2} \cdots \int_{x_i=x_{i-1}}^{t_i} \int_{x_{i+1}=x_i}^{t} \cdots \int_{x_k=x_{k-1}}^{t} dx_k \ldots dx_1 \qquad (5.10) \\
&= \frac{k!}{t^k} \int_{x_1=0}^{t_1} \int_{x_2=x_1}^{t_2} \cdots \int_{x_i=x_{i-1}}^{t_i} \frac{(t - x_i)^{k-i}}{(k-i)!} dx_i \ldots dx_1 \qquad (5.11) \\
&= \frac{k!}{t^k} \Big(\frac{t^k}{k!} - \sum_{m=0}^{i-1} \frac{(t - t_{m+1})^{k-m}}{(k-m)!} g(t_1, \ldots, t_m)\Big) \qquad (5.12) \\
&= 1 - \frac{k!}{t^k} \sum_{m=0}^{i-1} \frac{(t - t_{m+1})^{k-m}}{(k-m)!} g(t_1, \ldots, t_m).
\end{aligned}
$$

Here (5.10) follows from the order statistics property, Lemma 1 of Bertsimas and Servi [2] implies (5.11), and (5.12) is derived by integrating (5.11).

As for the second probability in (5.9),

$$\mathrm{P}\{C_n \setminus B_k \mid A(t) = k\} = e^{-\lambda(t_n - t)}\lambda^{n-k} g(t_{k+1} - t, t_{k+2} - t, \ldots, t_n - t), \tag{5.13}$$

and $g(t_{k+1} - t, t_{k+2} - t, \ldots, t_n - t)$ can be computed recursively using (5.6).

Formulas (5.9), (5.12) and (5.13) imply (5.8).

**Remark.** Note that extrapolation requires $O(n)$ computations for each updating of the estimate. However, for the interpolation problem the number increases to $O(n^2)$, which is the complexity of calculating $\mathrm{E}\,[A(t) \mid D_n]$.

## 5.2 Immediate Transitions, Known External Arrivals

The queue is estimated at station $j$ at time $s$, before the observation time $t$.

**Proposition 5.1.**
If $s \le C_j(t)$, $Q_j(s)$ is known exactly given $E_t$ and is given by

$$Q_j(s) = A_j(s) - A_j(C_j(s)). \tag{5.14}$$

If $s > C_j(t)$,

$$Q_j(s) = [A_j(C_j(t)) - A_j(C_j(s))] + [A_j(s) - A_j(C_j(t))],$$

where the first term is known exactly given $E_t$ and

$$\mathrm{E}[A_j(s) - A_j(C_j(t)) \mid E_t] = \sum_{t^n_{ml} \in I_R(t) \text{ s.t. } t^n_{ml} < s} 1_{\{j \in \mathcal{M}_t(t^n_{ml})\}} \frac{p_{nj}}{\sum_{k \in \mathcal{M}_t(t^n_{ml})} p_{nk}}.$$

*Proof.*
Let $s \le C_j(t)$. Formula (5.14) prevails since, due to FCFS, customers that wait in queue at $s$ arrived during $[C_j(s), s]$. All arrivals to station $j$ which took place before $C_j(t)$ are known because the corresponding customers started service before $t$ and we can trace their previous transactions using their ID's. Therefore, $A_j(C_j(t)) - A_j(C_j(s))$ is known exactly.

If $s > C_j(t)$, arrivals that took place during $[C_j(t), s]$ are estimated using the same argument as for formula (4.2) in Proposition 4.1. Figure 5.1 might be helpful.

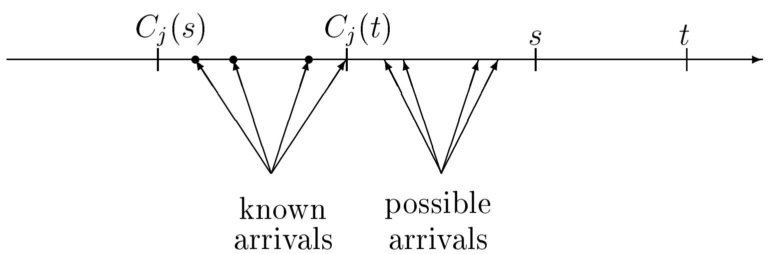


**Figure 5.1. Illustration to Proposition 5.1.**

### 5.3 Immediate Transitions, Unknown External Arrivals

We provide only an outline of the solution, since it is similar to the one in §4.3. In order to calculate $\mathrm{E}[Q_j(s) \mid E_t]$, $s < t$, two special cases must be considered.

If $s \leq C_j(t)$, all internal arrivals to station $j$ until $s$ are known. The external arrivals can be estimated using the approach from §2.2.

Otherwise ($s > C_j(t)$), one must estimate the arrivals in $[C_j(s), C_j(t)]$ and in $[C_j(t), s]$ separately. Arrivals that took place during the first interval can be estimated using HMM approach. As for the second interval, the technique of §4.3 must be used. External arrivals are estimated using formulae (4.22) and (4.23), where we replace $t$ by $s$. Internal queue is estimated using (4.18)-(4.21). In (4.23) and (4.18), only those internal customers, whose most recent service-termination times $s_k \leq s$, must be included in the sum.

## 6 Applications.

Here we apply our algorithm for Busy-Period Interpolation and immediate transitions to real data. Real-Time applications are left for future research.

The service system that we analyze is a typical branch of a commercial bank. Measurements from 14 days of work were collected. The data constitutes of records, each corresponding to a specific service that is characterized as follows: service date, customer ID, station ID, server ID, time of customer arrival to the station, time of service start, time of service termination, waiting time and service time.

Table 3 contains part of such a data file for Day 8. Note that customers with ID's 3006, 3031, 3036, 3040 visited more than one station (for example, customer 3006 visited stations 5 and 3). Several stations have multi-servers (for example, station 5 is occupied by servers with ID's 12, 13, 14, 15, 16).

Times of service starts, service terminations and external arrivals to the bank entrance (station 0) were recorded directly by the measurement system but internal arrival times were not. The measurement system assumes that an arrival time of a customer to the first internal station on his route is equal to his external arrival time. Each successive arrival time to a station is assumed to be equal to the preceding service termination time. In essence, the assumption of immediate transitions between stations is implicitly used by the measurement system. This assumption seems to be typically a good approximation for the bank, where walking distances are short. (However, we shall see below that sometimes this assumption is violated in a way that affects calculations of the queue.)

This data set is convenient for our application, since it adheres to our simplest model: transitions between stations are immediate and external arrivals are known. In this case, the exact queue is computable (see §2.1).

| DATE | CUST ID | STAT ID | SERVER ID | ARRIV | SERV START | SERV TERM | WAIT TIME | SERV TIME |
|---|---|---|---|---|---|---|---|---|
| 300594 | 3003 | 5 | 14 | 8:26:17 | 8:29:49 | 8:33:47 | 3.533 | 3.967 |
| 300594 | 3005 | 5 | 12 | 8:26:55 | 8:34:15 | 8:37:31 | 7.333 | 3.267 |
| 300594 | 3006 | 5 | 14 | 8:26:58 | 8:33:47 | 8:39:06 | 6.817 | 5.317 |
| 300594 | 3006 | 3 | 10 | 8:39:06 | 8:39:06 | 8:43:49 | 0.000 | 4.717 |
| 300594 | 3007 | 5 | 16 | 8:27:01 | 8:29:10 | 8:35:40 | 2.150 | 6.500 |
| 300594 | 3008 | 5 | 12 | 8:27:04 | 8:30:42 | 8:34:15 | 3.633 | 3.550 |
| 300594 | 3009 | 5 | 15 | 8:27:07 | 8:28:58 | 8:32:54 | 1.850 | 3.933 |
| 300594 | 3010 | 5 | 15 | 8:27:54 | 8:32:54 | 8:34:01 | 5.000 | 1.117 |
| 300594 | 3011 | 5 | 12 | 8:28:01 | 8:28:25 | 8:30:42 | 0.400 | 2.283 |
| 300594 | 3013 | 5 | 15 | 8:28:08 | 8:34:01 | 8:35:31 | 5.883 | 1.500 |
| 300594 | 3014 | 8 | 22 | 8:28:11 | 8:29:15 | 8:42:32 | 1.067 | 13.283 |
| 300594 | 3015 | 8 | 19 | 8:28:14 | 8:30:34 | 8:42:36 | 2.333 | 12.033 |
| 300594 | 3016 | 5 | 16 | 8:28:17 | 8:35:40 | 8:40:43 | 7.383 | 5.050 |
| 300594 | 3017 | 1 | 5 | 8:28:21 | 8:36:02 | 9:06:05 | 7.683 | 30.050 |
| 300594 | 3018 | 5 | 12 | 8:28:24 | 8:37:31 | 8:38:25 | 9.117 | 0.900 |
| 300594 | 3019 | 5 | 15 | 8:28:27 | 8:37:00 | 8:44:21 | 8.550 | 7.350 |
| 300594 | 3020 | 5 | 12 | 8:28:30 | 8:41:54 | 8:43:11 | 13.400 | 1.283 |
| 300594 | 3023 | 5 | 12 | 8:28:48 | 8:38:25 | 8:40:08 | 9.617 | 1.717 |
| 300594 | 3024 | 5 | 15 | 8:28:55 | 8:35:31 | 8:37:00 | 6.600 | 1.483 |
| 300594 | 3026 | 4 | 11 | 8:30:09 | 8:31:38 | 8:34:59 | 1.483 | 3.350 |
| 300594 | 3027 | 5 | 16 | 8:30:23 | 8:40:43 | 8:45:28 | 10.333 | 4.750 |
| 300594 | 3029 | 5 | 12 | 8:31:05 | 8:40:08 | 8:41:54 | 9.050 | 1.767 |
| 300594 | 3030 | 5 | 16 | 8:31:49 | 8:45:28 | 8:53:34 | 13.650 | 8.100 |
| 300594 | 3031 | 3 | 10 | 8:31:54 | 8:32:56 | 8:36:41 | 1.033 | 3.750 |
| 300594 | 3031 | 4 | 11 | 8:36:41 | 8:50:37 | 8:52:47 | 13.933 | 2.167 |
| 300594 | 3032 | 5 | 14 | 8:31:57 | 8:40:30 | 8:43:53 | 8.550 | 3.383 |
| 300594 | 3035 | 5 | 12 | 8:33:12 | 8:43:11 | 8:48:09 | 9.983 | 4.967 |
| 300594 | 3036 | 1 | 2 | 8:33:24 | 8:34:38 | 8:41:28 | 1.233 | 6.833 |
| 300594 | 3036 | 4 | 11 | 8:41:28 | 8:52:47 | 8:54:13 | 11.316 | 1.433 |
| 300594 | 3037 | 5 | 14 | 8:33:28 | 8:43:53 | 8:46:01 | 10.416 | 2.133 |
| 300594 | 3038 | 5 | 15 | 8:33:32 | 8:44:21 | 8:47:18 | 10.816 | 2.950 |
| 300594 | 3040 | 5 | 14 | 8:33:47 | 8:48:51 | 8:50:16 | 15.066 | 1.417 |
| 300594 | 3040 | 5 | 16 | 8:50:16 | 8:53:34 | 8:54:28 | 3.300 | 0.900 |
| 300594 | 3040 | 6 | 17 | 8:54:28 | 8:54:39 | 8:59:47 | 0.183 | 5.133 |

**Table 3. Example of Data. Day 8.**

*The entries of columns are date, ID's of customers, ID's of stations, ID's of servers, arrival times, times of service starts, times of service terminations, waiting times and service times, respectively.*

Examples of the queue length at station 5 during several working days are presented in Figure 6.1. (Tellers 12-16 work there simultaneously.) We observe very different queue patterns for different days.

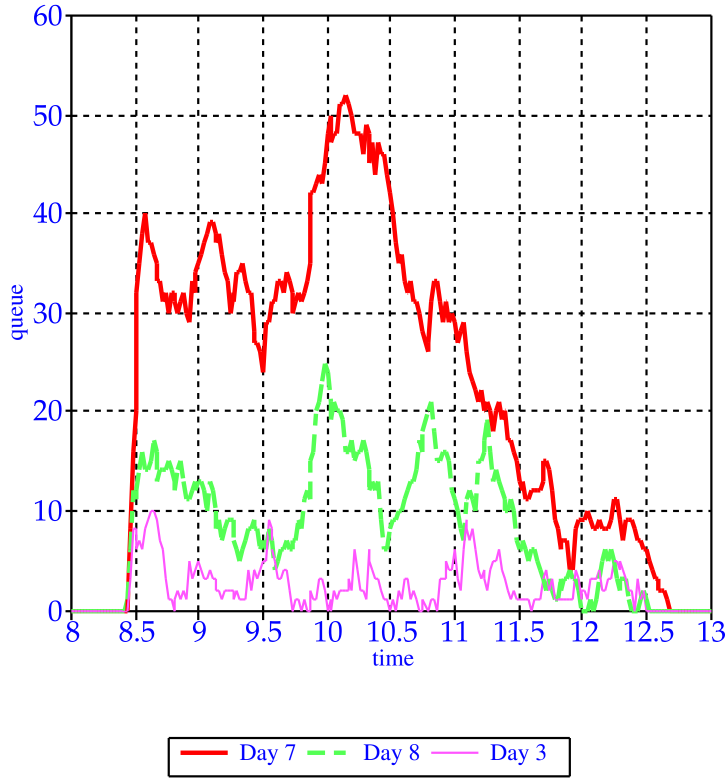


**Figure 6.1. Examples of queues at a multi-server station.**

Now we may pretend that external arrivals are unknown, run Algorithm 2.1 with incomplete data and, finally, compare the queue estimate with the exact queue. We assume a constant arrival rate during busy periods in the following applications. At the station considered, the queue empties several times during the day and the constant-rate assumption seems like a good approximation.

This approach is applied to the data of station 4 which is occupied by the single teller 11. Estimates of the queue length were calculated for several days. Some of the results are presented in the figures below. (For clarity, the figures are restricted to the time periods that are most relevant for the points that they illustrate.) Every figure contains three graphs: exact queue, its estimate (conditional expectation of the queue length given transactional data), and the 90% quantile of the conditional distribution of the queue length. The exact queue is computed using

$$Q_t = A_t - D_t,$$

where $Q_t$ is the cumulative number of arrivals to the station until $t$ and $D_t$ is the cumulative number of departures. Figure 6.2 exhibits good correspondence between the estimated and the exact queue.

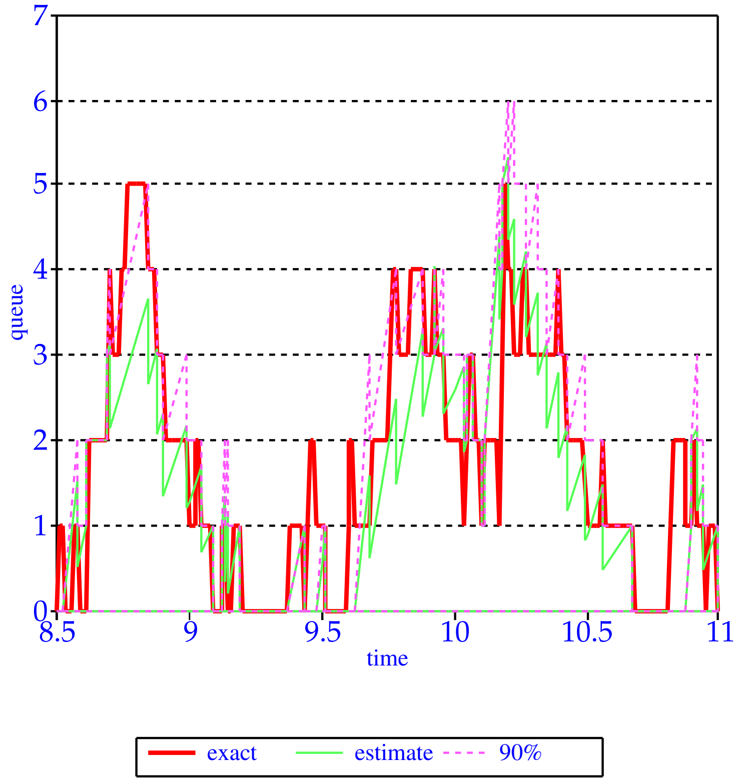


**Figure 6.2. Exact queue length and its estimate on day 8.**

When the estimate on day 8 is computed under the assumption that all arrivals, external and internal, are unknown (using, in essence, the one-station algorithm of Larson), we get Figure 6.3.

We observe that, in comparison with Figure 6.2, the correspondence is worse during two time intervals: after 9:00 and after 10:00. (Note also the difference between the 90% quantiles of two estimates.) The reason for the discrepancy is that, given FCFS, the queue length is known *exactly* at times of internal arrivals (see §2.2). Indeed, during the longest busy period (approximately between 9.6 and 10.7) only four arrivals out of 20 were internal. However, their knowledge reveals the exact queue and helps to improve the estimate significantly.

When working with real-data, some assumptions of our model, such as immediate transitions and FCFS, do not always prevail. For example, consider the queue estimate on day 11 (Figure 6.4).

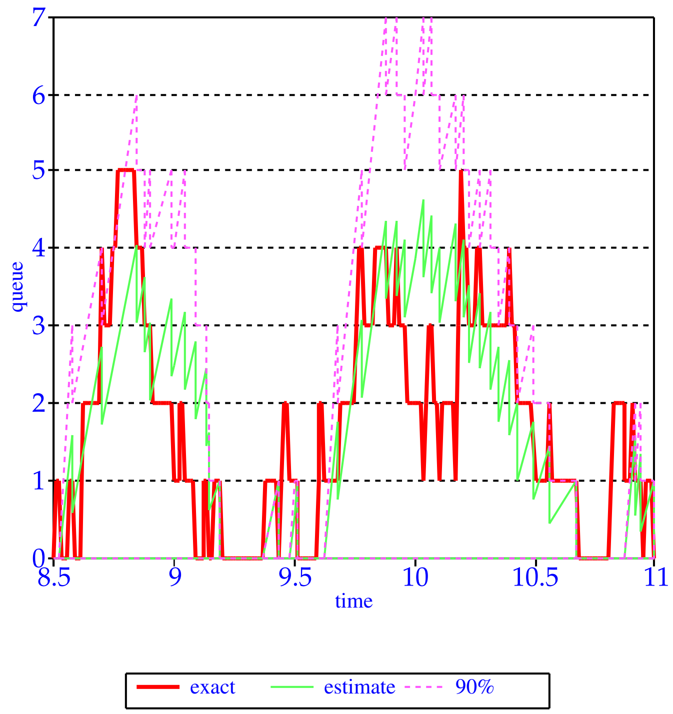


**Figure 6.3. Queue length and its estimate on day 8. All arrivals assumed unknown.**

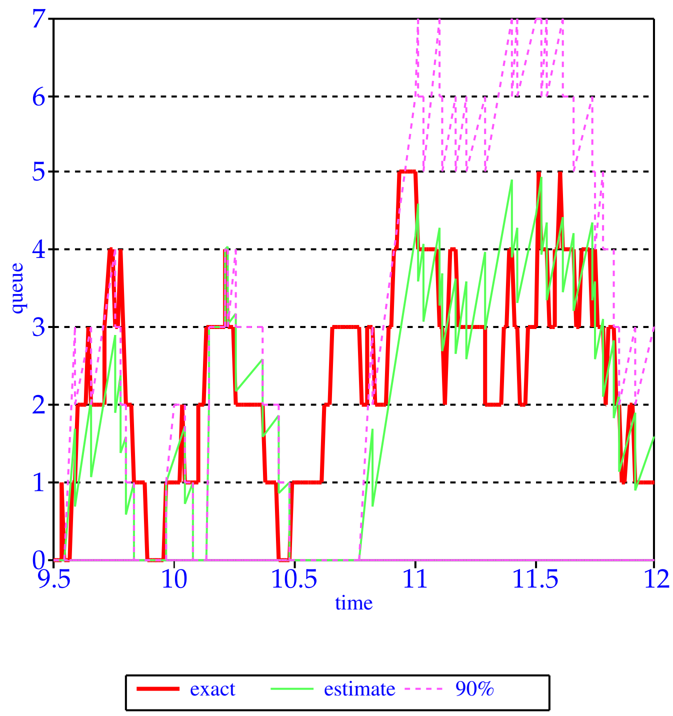


**Figure 6.4. Queue length and its estimate on day 11.**

The fit is satisfactory everywhere except during the time interval between 10.5 and 10.8 where the estimated queue vanishes but the exact queue increases from 0 to 3. To understand this, consider the corresponding part of the data file in Table 4.

| Station | ID | Arrival | Service start | Service termination |
|---|---|---|---|---|
| 4 | 3281 | 10.2150 | 10.4344 | 10.4778 |
| 4 | 3354 | 10.4722 | 10.4778 | 10.7283 |
| 4 | 3380 | 10.6556 | 10.7744 | 10.8311 |
| 4 | 3376 | 10.6258 | 10.8311 | 11.0192 |
| 4 | 3400 | 10.8011 | 11.0192 | 11.0433 |
| 4 | 3356 | 10.4806 | 11.0433 | 11.1022 |

**Table 4. Day 11. Part of data file.**

We observe the violation of the FCFS assumption for customers 3380, 3376 and 3356. This may be due to non-immediate transitions between stations. For example, customer 3356 terminated his previous service at 10.4806. Our algorithm that calculates the exact queue, given known external arrivals, assumes immediate transitions and registers him in the queue since 10.4806. However, customer 3556 possibly arrived to station 4 later. Therefore, the actual queue may in fact be closer to our estimate than the measured (exact) queue.

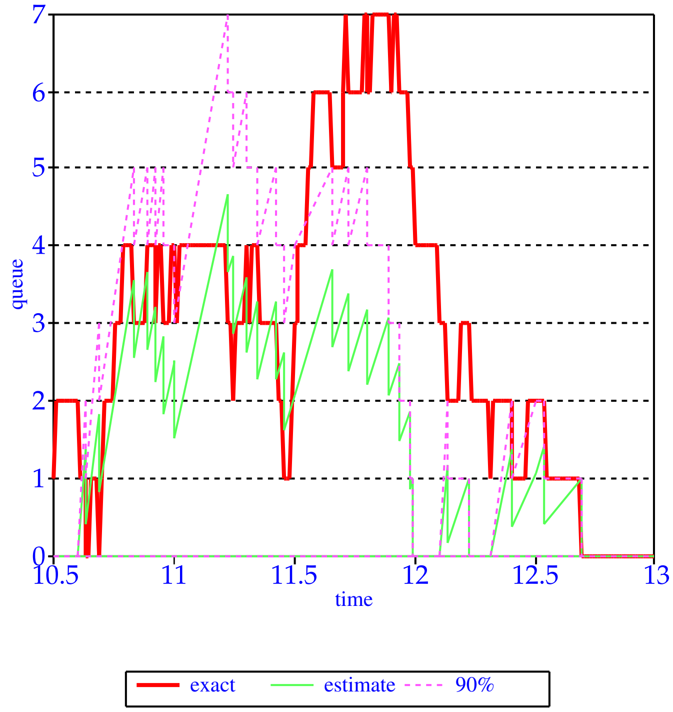


**Figure 6.5. Queue length and its estimate on day 10.**

It turns out that most problems with our real-data arise due to violations of the work-conserving principle (a human server often stops working for some periods of time even when there are customers waiting for service). Consider, for example, Figure 6.5. During the time interval [10.5,11.5] there is a reasonable fit, but the fit deteriorates between 11.5 and 12.5.

To understand this, consider Table 5.

| Station | ID | Arrival | Service start | Service termination |
|---|---|---|---|---|
| 4 | 3609 | 11.5517 | 11.9347 | 11.9783 |
| 4 | 3616 | 11.5781 | 11.9783 | 11.9953 |
| 4 | 3638 | 11.7042 | 11.9953 | 12.0700 |
| 4 | 3646 | 11.7189 | 12.1014 | 12.1353 |
| 4 | 3670 | 11.8286 | 12.1353 | 12.2317 |
| 4 | 3659 | 11.7814 | 12.2317 | 12.2931 |
| 4 | 3688 | 11.9175 | 12.3125 | 12.4094 |
| 4 | 3740 | 12.1894 | 12.4094 | 12.5467 |

**Table 5. Day 10. Part of data file.**

We observe that at 12.07 and at 12.2931 there were customers standing in queue but, regardless, the service stopped for a short period. Our algorithm interprets this fact as the end of busy period which causes the bad fit. To overcome such a problem, we prolong the service times of these customers. Then a very different graph emerges in Figure 6.6: the estimate is still below the exact queue but the 90% quantile covers the exact queue almost everywhere; this is a testimony to the quality of our algorithm,

This example suggests that a server should register the intervals during which his work is interrupted but there are still customers in queue. Then, in the algorithm, one can assume that the service before the interruption is prolonged until the start of the next service. However, if the customer switches to another station and one needs to estimate the queue there, then his actual service-termination time must be used as the arrival time to that station. (This in order to account for his presence there.)

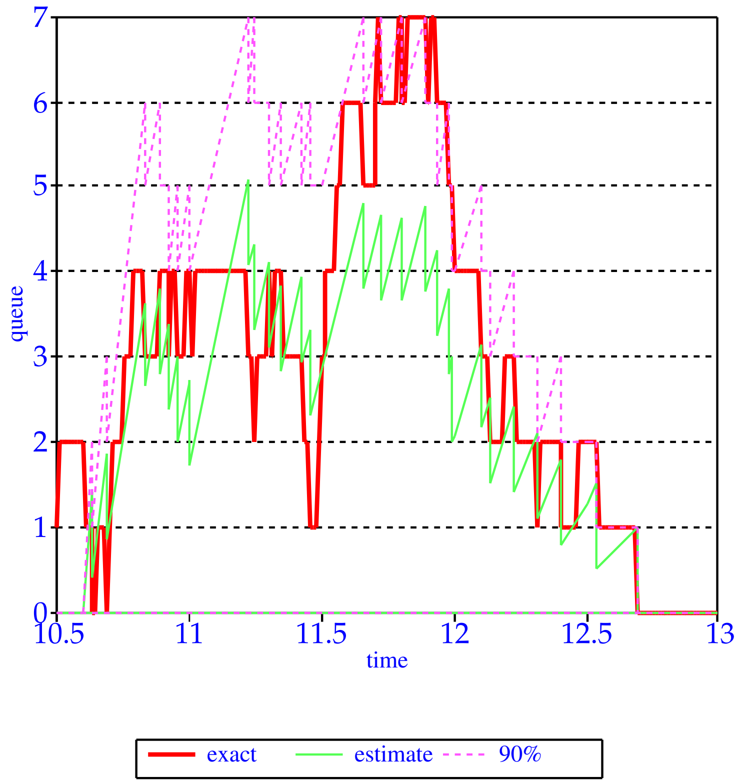


**Figure 6.6. Queue length and its estimate on day 10 after data correction.**

**Remarks.** Working with real data, we encountered also the following data error: a server forgets to register the end of a busy period. (The last service of the busy period is erroneously prolonged until the start of the next service.) As a result, two separate busy periods are interpreted by the algorithm as one large busy period and, consequently, the QIE-estimate deteriorates. Here, we believe that inaccuracy can be reduced by filtering unusually long service times.

In our applications, we used the assumption of Markovian routing between stations. However, one could use more general Kelly/BCMP Network (multi-type, deterministic routing per type). We completely agree that many service applications can be better modeled within the BCMP framework and leave this topic for future research.

On the other hand, a Jackson-like representation enjoys simplicity. Moreover, had we used a Kelly/BCMP like model, the information embodied in classes would have rarely mattered since in our applications, most customers actually left after a single or two visits. As it happens, the first visit does not determine the route. Hence, after the first visit, one must estimate transition probabilities and the model becomes similar to that of a Jackson model.

# 7 Possible Future Research.

We conclude with some possible research directions that could build upon the work presented here.

- Approximations. These can be applied to problems that are otherwise non-computable (general arrival pattern, no information about ID's, real-time estimation problems). An appropriate mathematical framework is the theory of conditional weak convergence ([11],[17]), excursion theory (for example, [1]) and large deviations theory ([6],[15]). For example, queue length during a busy period, after proper normalization, would converge to a Brownian excursion when the number of services is large. (Daley and Servi began working in this direction in [5].)
- Adding new modeling features, such as buffers, abandonment, reneging, multi-type customers and alternatives to FCFS (for example priorities).
- Analyzing new performance measures, such as waiting-time distributions during a busy period or fraction of customers who wait more then $\tau$ minutes. The topic of total time in system (sojourn times of customers), and breaking it down into components, namely service time, travel time and waiting time, is of special interest.
- Applications of real-time algorithms.
- Developing robust algorithms that are less sensitive to measurement errors and violations of model assumptions.
- Assessing the numerical complexity of our algorithms.

# References

[1] R.M.Blumenthal "Excursions of Markov Processes", Birkhauser, Boston, (1992).

[2] D.J.Bertsimas, L.D.Servi "Deducing queueing from transactional data: the Queue Inference Engine, revisited", *Operations Research*, 40 (1992) 217-228.

[3] D.J.Daley, L.D.Servi "Exploiting Markov chains to infer queue length from transactional data", *Journal of Applied Probability*, 29 (1992) 713-732.

[4] D.J.Daley, L.D.Servi "A two-point Markov chain boundary value problem", *Advances of Applied Probability*, 25 (1993) 607-630.

[5] D.J.Daley, L.D.Servi "Approximating last exit probabilities of a random walk, with application to conditional queue length moments within busy periods of $M/GI/1$ queues", *Journal of Applied Probability*, 31A (1994) 251-267.

[6] A.Dembo, O.Zeitouni "Large Deviations Techniques and Applications", Jones and Bartlett, Boston, 1992.

[7] R.Durrett "Probability: Theory and Examples", Wadsworth & Brooks, (1991).

[8] R.G.Elliott, L.Argoun, J.B.Moore "Hidden Markov Models. Estimation and Control", Springer Verlag, (1995).

[9] R.Gawlick "Estimating disperse queueing networks: the Queue Inference Engine", *Computer Communication Review*, 20 (1990) 111-118.

[10] S.A.Hall "New directions in queue inference for management implementations", Massachusetts Institute of Technology, Ph.D. Thesis, (1992).

[11] Iglehart D.L. "Conditioned limit theorems for random walks", in M.Puri (Ed.) "Stochastic Processes and Related Topics", vol.1, Academic Press, New York.

[12] L.K.Jones, R.C.Larson "Efficient computation of probabilities of events described by order statistics and applications to queue inference", preprint (1992).

[13] R.C.Larson "The Queue Inference Engine: deducing queue statistics from transactional data", *Management Science*, 36 (1990) 586-601.

[14] R.C.Larson "The Queue Inference Engine: deducing queue statistics from transactional data, addendum", *Management Science*, 36 (1990) 1062.

[15] A.Shwartz, A.Weiss "Large Deviations for Performance Analysis: Queues, Communications and Computing", Chapman & Hall (1995).

[16] L.R.Rabiner "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proceedings of the IEEE*, 77 (1989) 257-286.

[17] Takacs, L "Queueing methods in the theory of random graphs", in J.H.Dshalow (Ed.) "Advances in Queueing: Models, Methods and Problems", CRC Press 1995, 45-78.

[18] J.Walrand "Introduction to Queueing Networks", Prentice Hall, (1988).

[19] S.Zeltyn "Estimating characteristics of queueing networks using transactional data", Technion, M.Sc.Thesis (1996).