Title:

A Queuing Network Model with Blocking: Analysis of Congested Patient Flows in Mental Health Systems

AUTHOR INFORMATION

Naoru Koizumi (Corresponding author)
Department of Electrical and Systems Engineering, University of Pennsylvania
278 Towne Building, 200 South 33rd Street, Philadelphia, PA 19104
Phone: +1-215-880-6095 Fax: +1-215-898-5020
E-mail: koizumi@seas.upenn.edu

Eri Kuno
Center for Mental Health Policy Services Research
Department of Psychiatry, School of Medicine, University of Pennsylvania

Tony E. Smith
Department of Electrical and Systems Engineering,
University of Pennsylvania

ACKNOWLEDGEMENTS

We would like to express our appreciation to Dr. Michael Fu at the Department of Business & Information Technology of the University of Maryland and Dr. Aileen Rothbard at the Center for Mental Health Policy Services Research of the University of Pennsylvania for their tremendous input to this paper.

A Queuing Network Model with Blocking: Analysis of Congested Patient Flows in Mental Health Systems

ABSTRACT

The downsizing and closing of state mental health institutions in Philadelphia in the 1990's led to the development of a continuum care network of residential-based services. Although the diversity of care settings increased, congestion in facilities caused many patients to unnecessarily spend extra days in intensive facilities. This study applies a queuing network system *with blocking* to analyze such congestion processes. "Blocking" denotes situations where patients are turned away from accommodations to which they are referred, and are thus forced to remain in their present facilities until space becomes available. Both mathematical and simulation results are presented and compared. Although queuing models have been used in numerous healthcare studies, the inclusion of blocking is still rare. We found that, in Philadelphia, the shortage of a particular type of facilities may have created "upstream blocking". Thus removal of such facility-specific bottlenecks may be the most efficient way to reduce congestion in the system as a whole.

KEYWORDS

Open queuing network model, Blocking, Single node decomposition method, Simulation, Mental health system

1. INTRODUCTION

During the past several decades, many U.S. states, including Pennsylvania, have downsized and/or closed state mental hospitals [1]. The intention behind this downsizing and/or closure was to provide mentally ill patients with more adequate care by replacing the state hospital function with a continuum care network which consists of the facilities with various levels of structure and support. In response, several types of less structured residential alternatives have been developed [2]. In Philadelphia, the state mental hospital was closed in 1990. Several structural changes have been observed since the closure. The three relevant changes are listed below.

- *Increase in the number of residential facilities*: The budget allocation for the state hospital was redirected to increase existing residential facilities. By 1997, the number of residential beds had increased from around 450 to approximately 1,400.
- *Emergence of Supported Housing*: A new type of community setting accommodation, "supported housing", was created to provide housing to clients with basic self-care and medication management skills. Among the various community setting options, supported housing provides the least care. It is commonly an individual home visited by a mental healthcare specialist on a regular basis. About 400 supported housing units were created by 1997.
- *Development of EAC Hospitals*: To accommodate the acute hospital patients whose condition would not stabilize during the maximum of 30-day acute care period, extended acute care (EAC) beds were created in general hospitals to provide relatively intensive care for up to 90 additional days.

The Philadelphia mental health system has been the forefront of the movement [3]. Inflow and outflow of clients into the continuum care network is controlled by the centralized gate-keeping agency, Access to Alternative Services (AAS). However, planning frameworks are lacking to provide mental health system planners with information on how many, and which type of residential services are required to optimize the placement match. Despite the initial intention of providing the patients with adequate level of care, the data collected by AAS suggest that a certain proportion of the clients find the facilities that they were referred to are full, and they are forced to wait at their current, often unnecessarily intensive, care facilities. Table 1 shows the observed empirical waiting time at three main types of facilities obtained from the 1997-1999 data.

[INSERT TABLE 1]

These clients unnecessarily "block" the current beds while waiting, preventing utilization by potential clients who require care at these facilities. This "blocking" phenomenon often generates some financial loss because the blocking clients are, in many cases, ready to move to a less intensive and hence less expensive facility. Hence, controlling the congestion in the existing mental health system is important not only from the clinical and human rights perspectives but also from the budgetary perspective for the mental health policy makers.

Our model analyzes the congestion in the system by applying a queuing network model with blocking. Two performance indicators, the number of patients waiting to enter each type of facility in the system and the associated waiting time at the steady state, are derived in the steady-state analysis. Queuing with blocking is a relatively new topic and most applications have been done in the engineering fields. To the author's knowledge, an applied study of a blocking model in the mental health field has not been published. Moreover, we believe that this is the first study that introduced a queuing model as a tool to plan and manage resource allocation in a mental

health system. Traditionally, needs assessment strategies have been the major approach to mental health system planning[1]. In this approach, predicted needs level for defined geographical areas or gaps between needed and currently provided services are used to guide mental health resource allocation decisions. A drawback of the needs assessment approach is its cross sectional nature that limits the ability of decision makers to examine interdependency among various components of the system concurrently and longitudinally (Peterson, 1987). The needs assessment approach can identify the proportion of clients who are in higher levels of care than needed among the resident population at the moment, and estimate the required number of beds at each level of care to place the currently served clients in their appropriate need level. However, the system in reality is not static; new clients are constantly arriving and consumer need level changes over time. What is desirable is a method that takes into account these dynamic factors in predicting required bed capacity. In addition, because each setting is dependent on the behavior of the other settings, it is important to consider how changes in the capacity of one setting affect needs and utilization in another. The queuing approach introduced in our study addresses both drawbacks: First, queuing theory is fundamentally based on a Markov process which is inherently dynamic. Second, by introducing the "blocking" concept into a traditional infinite buffer network model (i.e., Jackson model), the degree of interdependencies between settings can be examined. Specifically, our model can project the relationship between capacities and unnecessary stays at various treatment facilities. Furthermore, the model can show service planners various relationships among the model parameters, for example, how unnecessary stays can be minimized by changing residential capacity as well as by other strategies to reduce clients' length of stay (e.g., efficient discharge planning, increasing intensive case management support, etc.). .

Although the work presented here applies the basic blocking model and a simplified structure of the real-world mental health system in Philadelphia, future extensions of this work could lead to the development of service configurations that provide a better match between level of care and needs of consumers than those based on the current decision making practice.

2. METHODOLIGICAL BACKGROUND: THEORY AND APPLICATION

Theoretically, "blocking" occurs when waiting spaces between stations are finite (i.e., finite buffers). The pioneering work of Hunt (1956) [4] examined a finite buffer between two stations. The model is based on Markov chains and has Poisson arrival rates and exponential service times. The methodology used in his model provides the "exact solutions", and thus the probabilities of all possible states at the two stations are obtained. But as the number of stations and/or the number of servers becomes larger, the set of possible states increases dramatically and computations become far more complex. A methodology that approximates the solution was thus proposed to address this computational challenge. One of the most well-known approximation method decomposes a network into smaller subsystems, analyzes each subsystem in isolation, and uses the subsystem results to analyze the overall network. Hillier and Boling (1967) [6] is probably the first publication in this area. Since then, the method was further researched and extended by many publications [7, 8, 9, 10]. Our study fundamentally applies the decomposition algorithm introduced by Takahashi, Miyahara, and Hasegawa [7], which dealt with blocking in an open queuing network system with feed-forward flows and finite buffers between stations. Our

[1] The methods of need assessment approach include epidemiological surveys, social indicator analysis, treatment utilization review, key informant interviews, and community opinion surveys (Rabkin, 1986). More recently, an individually based assessment approach has been explored, in which a panel of stakeholders prescribed service needs for groups of clients with similar levels of functioning (e.g., Leff, Mulkern, Lieberman, & Raab, 1994; Shern, Wilson, Ellis, Bartsch, & Coen, 1986; Srebnik, Uehara, & Smukler, 1998).

work, however, extends their model by modifying their single-server model to a multiple-server model. To author's knowledge, the direct application of Takahashi, Miyahara, and Hasegawa [7] in the context of multiple servers is only found in Korporal, Ridder, Kloprogge, and Dekker (2000) [11], which analyzed the congestion level in a penitentiary system in the Netherlands.

While there are numerous published healthcare studies that apply queuing theories, those analyzing a blocking mechanism are still rare. Four healthcare publications were found that involve analyses of blocking. El-Darzi, et al. (1998) [12] and Cohen, Hershey, and Weiss (1980) [13] are both *simulation* studies. El-Darzi, et al. [12] analyzed the congestion in geriatric patient flows in a U.K. hospital system. The model framework used in this study is similar to our mental facility system, with the exception that their system has a "tandem" structure as opposed to our "arbitrarily-linked" system (i.e., patients can skip stations). The other two articles, Hershey, Weiss, and Cohen (1981) [14] and Weiss and McClain (1987) [15] employ mathematical approaches in analyzing a blocking problem. However, neither of these methods could be directly applied to analyze congestion in the mental health system in Philadelphia. Hershey, Weiss, and Cohen [14] dealt with blocking in the same context as this paper, but blocking occurs only when entities enter a specific station (Unit 1), while other stations were assumed to have infinite waiting space in front of the stations. The methodology introduced by Weiss and McClain [15] may approximate the congestion in the Philadelphia mental health system with reasonable accuracy. However, their methodology does not utilize either blocking or the single-node decomposition approach that has advanced significantly since Hillier and Boling [6].

3. MODEL FRAMEWORK

Our system consists of three types of psychiatric institutions: *extended acute hospitals* (*E*), *residential facilities* (*R*), and *supported housing* (*S*). *E* is the most structured institution in the system with the patients who require follow-up care after being discharged from acute hospitals. *R* accommodates those clients who require basic daily living support with full-time monitoring, while *S* is the least structured institution in the system and provides clients with a minimum daily living support on a part-time basis. The accommodations outside the system are categorized into two groups: *acute hospitals* (*A*) and *all other accommodations* denoted as *X*. Station *X* is composed of various accommodations for psychiatric patients ranging from housing with family or friend houses to homeless shelters, jails, or even living on the streets. Figure 1 illustrates the three internal stations, two external stations, and the flows between these stations. As seen in the figure, patients have an overall tendency to flow from the most structured institution *E*, to the least structured institution, *S* in the system.

[INSERT FIGURE 1]

In the figure, there are two dotted backflows, (i) $R \rightarrow A$ and (ii) $S \rightarrow A$. These flows reflect clients at *R* and *S* who experience relapse and hence flow backwards to acute hospitals. Under the current policy, most residential (*R*) and supported housing (*S*) clients keep their beds while temporarily receiving acute care in *A*, and thus effectively occupy two types of beds in the system. Clients at residential facilities and supported housing must give up their beds only if they are away for longer than a specified length of time (i.e., 30 days for residential facilities and 60 days for supported housing). The actual data shows that, among those who relapse and move from *S* or *R* to *A*, only a few patients per year are forced to give up their beds at *S* or *R*. In fact, these patients constitute less than 0.01% of total patients who leave *S* or *R*. Thus, these two backflows were omitted in the present study. As to the patients who occupied two beds temporarily at (*R*,*A*) or (*S*,*A*) and came back to their primary bed within the specified period, the model treated these patients as if they remained at *R* or *S* throughout the hospitalization. Since *A* exists outside the

system and is also considered to have an infinite number of beds (i.e., no resource constraints), capturing the temporal backflows and associated occupancies of acute beds adds nothing of significance to our analysis[2].

Blocking at station i occurs when the patient out-flow from station i is hampered due to full occupancy at the immediate downstream station. There are two key characteristics of this process: (i) patients remain at station i even after completing their treatment, and (ii) these patients potentially block incoming patients to station i. In this paper, we define "blocking" as a situation where both of these characteristics are observed. Given that stations E, R and S have only a *finite* number of beds (and no other waiting space), blocking can occur at the flows $E \rightarrow R$ and $R \rightarrow S$. Blocking will *not* occur if the immediate downstream station has an *infinite* capacity (either beds or other forms of waiting space). In our model, the only station assumed to have an unlimited number of beds is A, and the only station assumed to have unlimited waiting space is X. Accordingly, the paths $E \rightarrow X$, $R \rightarrow X$, and $S \rightarrow X$ are "blocking-free" flows in the model. The flows $X \rightarrow A$, $A \rightarrow X$ are also "blocking-free" flows, but both X and A are external stations, and thus these flows are not modeled. There are four remaining flows in the model: $A \rightarrow E$, $A \rightarrow R$, $X \rightarrow R$ and $X \rightarrow S$. When patients at A or X find R or S full, they must wait at the current location [characteristic (i) above is observed]. But doing so would not block anyone exiting the system, since A and X have infinite capacities [characteristic (ii) is *not* observed]. This type of congestion is analyzed in traditional queuing models, and thus is here designated as "classic congestion". The following table summarizes the types of congestion associated with each flow in the model.

[INSERT TABLE 2]

In principle, patients arriving to any station in the system are treated equally regardless of the station of origin. This means that there is basically a single First-Come-First-Served (FCFS) queue at each station although the waiting location of the patients (blocked or unblocked) may vary.

4. STEADY-STATE ANALYSIS

In this section, we first provide a quick review of the standard queuing theory in our model framework but without considering blockings between stations. We will then explain how the standard algorithm needs to be modified in order to incorporate a blocking phenomenon. The derivation of the equations shown in the first subsection can be found in any basic queuing textbooks such as Gross and Harris (1998) [5], while the description of the blocking model is fundamentally based on Takahashi, Miyahara, and Hasegawa [7].

I. Steady-state Analysis *without* Blocking

As with most Markovian models, the present network queuing model exhibits long-run steady-state behavior (as long as inflows are compatible with system capacities). The analysis of such steady states thus provides a useful way of establishing long-run "performance indicators"

[2] There few publications that analyze blocking with feedback flows. This is partly due to inappropriateness of simple of standard Poisson-arrival assumptions when feedback flows exist, as shown by Disney (1981) [16]. Another reason could be that the model potentially faces a "deadlock flow" problem. Deadlock refers to the situation in which entities at two or more stations block each other, and occurs only when feedback flows are allowed in the system. To the author's knowledge, all existing articles make the simplifying assumption that deadlock is detected and resolved automatically by exchanging the entities between the stations. It should be noted, however, that deadlock could potentially violate the common assumption of a service discipline in a queuing model, "First Come First Served" [17].

for each station in the system. The two most commonly used performance indicators are mean length of queue and mean waiting time, while other indicators include the mean percentage of time that a station is full and the mean number of patients at a station. Essentially, these steady-state performance indicators are obtained mathematically by equating the arrival and exit rates at each station. The three key input parameters used in the analysis are: (i) *arrival rates* (i.e., expected number of patients arriving at each station per unit of time, (ii) *mean service time* (i.e., expected length of time that a patient spends at each station), and (iii) and the *number of servers* (beds at each station). The latter two parameters determine the *maximum service rate* at the station (i.e., the service rate achieved when the station is full). If we denote daily *service rate per bed* at station i by μ_i, then the *mean service time per bed* is given by $1/\mu_i$. If we further denote the number of *beds* at station i by c_i, the *service rate* for that station is given by $c_i\mu_i$. When only $n_i\ (< c_i)$ beds are occupied, the service rate must be $n_i\mu_i$.

In a queuing network model, while external arrival rates (for the arrivals from outside the system) are obtained from data, internal arrival rates (for the arrivals between stations) are obtained endogenously from external arrival rates and routing probabilities between (internal) stations (i.e., the fraction, r_{ij} of the patients departing from station i who are directed to station j in any given unit of time).

In general, the *routing matrix* is expressed as:

$$\boldsymbol{R} = \left[\begin{array}{ccc:c} r_{EE} & r_{ER} & r_{ES} & r_{EX} \\ r_{RE} & r_{RR} & r_{RS} & r_{RX} \\ r_{SE} & r_{SR} & r_{SS} & r_{SX} \end{array}\right],$$

where the vertical line separates flows with internal and external destinations. Referring to Figure 1, the specific *routing matrix* for this model is seen to be of the form:

$$\boldsymbol{R} = \left[\begin{array}{ccc:c} 0 & r_{ER} & 0 & r_{EX} \\ 0 & 0 & r_{RS} & r_{RX} \\ 0 & 0 & 0 & 1 \end{array}\right].$$

Given the *routing probabilities* $(r_{ER}, r_{RS}, r_{EX}, r_{RX})$ and *external arrivals* (λ_{Aj} and λ_{Xj}, $j = E, R, S$), the internal and total arrival rates $\left(\lambda_j : j = E, R, S\right)$ are generally obtained by solving the following system of general *traffic equations*:

$$\lambda_j = \lambda_{Aj} + \lambda_{Xj} + \sum_{i=1}^{m} r_{ij}\lambda_i$$

where i indicates any internal station from which patients flow directly to station j. Referring to Figure 1, it can be seen that the *traffic equations* for our system are:

$$\begin{aligned} \lambda_E &= \lambda_{AE} \\ \lambda_R &= \lambda_{AR} + \lambda_{XR} + r_{ER}\lambda_E \\ \lambda_S &= \lambda_{XS} + r_{RS}\lambda_R \end{aligned} \qquad (1)$$

In a traditional multi-server queuing network model *without* blocking (like an open Jackson network model), patients at a station arrive, receive services, and depart to the next station (if it is not full) or to infinite buffer space (if it is full). In such a system, the steady state of each station is analyzed in isolation without considering the impact of congestion at any particular station on the flows at other stations in the system [18, 19]. For mathematical convenience, total arrival rates are often assumed to follow *Poisson distribution* with *mean* λ (so that inter-arrival times follow an *exponential distribution* with mean $1/\lambda$)[3]. Similarly, mean service times are often assumed to follow an *exponential distribution* with *mean* $1/\mu$. In this study, we also assumed that the total arrival rates follow Poisson distribution, while definition of "service time" and its distribution need to be modified in order to incorporate blocking into the model. These will be discussed in detail in the following subsection.

Given the values for all input parameters λ_i, μ_i, and numbers of beds, c_i, at each station i, the *expected steady-state queue length* at station *i* in the case of no blocking is obtained by the following equation:

$$L_i = \left[\sum_{n_i=0}^{c_i-1} \frac{\omega_i^{n_i}}{n_i!} + \frac{\omega_i^{c_i}}{(1-\rho_i)c_i!} \right]^{-1} \frac{\rho_i \omega_i^{c_i}}{(1-\rho_i)^2 c_i!} \tag{2}$$

where $\omega_i = \lambda_i / \mu_i$ and $\rho_i = \omega_i / c_i < 1$.

In the equation, $\rho_i = \lambda_i /(c_i \mu_i)$ is known as the "traffic intensity" at station *i*. The case $\rho_i > 1$ represents a situation where the mean arrival rate, λ_i, is larger than the maximum service rate, $c_i \mu_i$, at station i. Under this circumstance, the queue increases without bound over time, and a steady-state cannot be achieved. Thus, $\rho_i = \omega_i / c_i < 1$, is essentially the stability condition for equation (2) to hold.

The *expected waiting time* to enter station *i* at steady-state is obtained by applying Little's formula [21], $W_i = L_i / \lambda_i$, as;

$$W_i = \left[\sum_{n_i=0}^{c_i-1} \frac{\omega_i^{n_i}}{n_i!} + \frac{\omega_i^{c_i}}{(1-\rho_i)c_i!} \right]^{-1} \frac{\rho_i \omega_i^{c_i}}{(1-\rho_i)^2 \lambda_i c_i!} \tag{3}$$

II. Steady-state Analysis *with* Blocking

When there is no buffer between stations (so that potential blocking exists), congestion at any particular station could potentially affect congestion levels at all upstream stations. Thus the Jackson approach needs to be modified to capture such interactions between stations. This can be accomplished by introducing the concept of "*effective* service time" [7]. To be more specific, suppose there are only two stations, i and $i+1$, linked in tandem. Then patients at station $i+1$ always depart system (to an infinite space) and never face blocking. Hence for any inflow to $i+1$, the steady state for this station can be solved using the traditional Jackson type approach. Since blocking at station *i* is caused by congestion at station $i+1$, the impact of this blocking can be analyzed by using "*effective* service time" rather than "service time". The *effective* service time is comprised of two types of service times, namely, "treatment time" and "blocked (or maintenance) time". *Treatment time* is the time between admission to station *i* and referral to

[3] Burke's theorem [20, 21] assures that internal and total arrival rates follow Poisson distribution as long as external arrival are assumed to follow Poisson distribution and there is no feedback flows in the system.

station $i+1$. Note that treatment time is equal to service time if no blocking exists. *Blocked time* represents the time between referral to station $i+1$ and physical exit from station i and is thus equivalent to the waiting time to enter station $i+1$. As stated earlier, waiting time at station $i+1$ is one of the performance indicators obtained from the steady-state analysis of station $i+1$. Therefore, the congestion impact of station $i+1$ is captured in the analysis of station i through *effective* service time. Figure 2 shows the relationship between the two stations. Here the waiting (or blocked) time to enter station $i+1$ is seen to be part of the *effective* service time from station i's perspective.

For mathematical convenience, we assume here that e*ffective* service times follow an *exponential distribution*[4]. The mean *effective* service time at station i (i = E, R) is henceforth denoted by $1/\tilde{\mu}_i$, where by definition $1/\tilde{\mu}_i$ consists of the mean treatment time, $1/\mu_i$, plus the mean waiting time, W_{i+1} to enter station $i+1$. The steady state for station i is then analyzed by applying equations (2) and (3), where the *effective* service time, $1/\tilde{\mu}_i \left[=(1/\mu_i)+W_{i+1}\right]$, replaces $1/\mu_i$. Accordingly, any tandem network system is solved sequentially from station $i+1$ to station i. Figure 1 however shows that in our system only fraction of the patients (r_{iX}) at stations $i=E,R$ leave the system after treatment without facing any wait. Hence the more general *effective* service time at stations $i=E,R$ are thus given by the convex combination of effective waiting times as:

$$1/\tilde{\mu}_i = r_{iX}(1/\mu_i) + \sum_j r_{ij}(1/\mu_i + W_j)$$

where j represents those stations downstream from i with limited capacities.

[INSERT FIGURE 2]

For E and R, the above equation respectively becomes:

$$1/\tilde{\mu}_E = r_{EX}(1/\mu_E) + r_{ER}(1/\mu_E + W_R), \text{ and} \quad (4)$$

$$1/\tilde{\mu}_R = r_{RX}(1/\mu_R) + r_{RS}(1/\mu_R + W_S) \quad (5)$$

Thus, equations (2) and (3) are applied to obtain steady-state mean queue lengths and waiting times in terms of *effective* service times. The number of patients waiting at station i to enter station j at steady state (L_{ij}) is then given by:

$$\begin{cases} L_{ij} = L_j(\lambda_{ij}/\lambda_j) & (i = A, X) \\ L_{ij} = r_{ij}L_j & (i = E, R, S) \end{cases} \quad (6)$$

The expected waiting time for arriving patients at immediate upstream stations should be the same regardless of the patient's station origin. This is because all new patients are treated FCFS as if they formed a single queue upon arrival.

[4] More generally *effective* service time can be modeled using an Erlang distribution. However, as observed by Perros [23], it is common practice to approximate effective service time distributions by the simpler exponential model.

Consequently, the steady state of every station in the system can be solved individually. The method is thus known as the "single node decomposition" approximation [6].

It should be noted that this analysis treats the waiting space to enter a particular station as infinite when analyzing the steady state of that station. In the two-station example above, this means that the number of patients waiting to enter station $i+1$ can be bigger than the number of beds at station *i*. When analyzing station $i+1$, the beds at station *i* play the role of "waiting space". The most congested scenario for station $i+1$ is where all at station *i* are occupied by patients waiting to be admitted to station $i+1$. In this scenario, no additional patient can enter station i, so that the maximum number of waiting patients is equal to the number of beds at station *i*. Therefore, it should be clear that our assumption only holds when the total number of beds at adjacent upstream stations is large enough to accommodate the steady-state number of waiting patients. The results of the numerical simulation below will indicate that our assumption does hold for the case studied here. However, an important direction for extending the present model would be to relax this assumption.

In summary, steady state of each station is analyzed using this single node decomposition method, and the overall model can be described in succinct terms by the standard notational shorthand, M/M/C/∞/FCFS. The notations respectively represent Markov arrival and departure (thus Poisson arrival rates and exponential service times) processes, a fixed number of servers (beds), an infinite capacity (waiting space), and the FCFS queuing discipline. The algorithm to solve the steady state of the system is outlined in Figure 3.

[INSERT FIGURE 3]

5. DATA SOURCES, DATA, PARAMETER VALUES AND ESTIMATES

Several administrative data files from the State and the County Office of Mental Health (OMH) were integrated to construct individually based histories of extended acute care (EAC) hospital and residential service use. The files contain data on dates and sources of referrals, dates of admission and discharge, and placement decisions. The integrated data consists of two data files maintained by the Access to Alternative Services (AAS), a centralized gate-keeping system that manages client entries to the institutional and publicly funded residential system in Philadelphia. In addition, the Center for Mental Health Policy and Services Research (CMHPSR), Department of Psychiatry, School of Medicine, University of Pennsylvania has access to Medicaid claims records that were used to gather data on general hospital admissions and discharges. The analysis involved both residential service utilization and placement referral data for public seriously mentally ill clients in Philadelphia from 1997 to 1999. The numbers of beds at each station were obtained from the OMH residential directory, and were validated by interviews with AAS staff.

The mean service times were obtained from the length-of-stay (LOS) data. The service times at community setting facilities (*R* and *S*), however, required estimation. The data for these two types of facilities revealed that a large share of patients had not exited from these facilities, thereby creating truncated LOS data. The reasons for this appear to be that many facilities are relatively new, and that the most community-setting facilities (especially supported housing) have been developed with an intention to provide the patients with a long-term care. The method of survival analysis was employed to estimate the service times at each of the community setting facilities. In particular, the commercial statistical software package, STATA (StataCorp, 2001, ver.7), was used for the analyses, and exponential survival distributions were employed (in a manner consistent with the exponential distributions in both the mathematical and simulation analyses).

A total of 1,415 observations were used for the survival analysis of station *R*, of which 806 were censored at the end of 1999. The estimate of the service rate, μ, and the summary of the survival regression results are presented in Table 3. The hypothesis of zero time effect was rejected (with P-value < 0.0001). The average daily service rate at station *R* was estimated to be 0.001, which in turn indicates that the average service time per patient is about 893 (=1/0.00112) days.

[INSERT TABLE 3]

There were 368 observations for the survival analysis of station *S*, of which 258 were censored at the end of 1999. The estimate of μ and the summary of the survival information are shown in Table 4. As seen from the table, the hypothesis of zero time effect was again rejected (with P-value < 0.0001). The estimated service rate is much smaller (0.0004) compared to that for the residential facilities. The mean service time per patient was thus estimated to be 2500 (=1/0.0004) days.

[INSERT TABLE 4]

The external arrival rates were estimated using both referral and actual admission data (the data indicated that a substantial number of patients referred to *R* or to *S* from *X* reneged from the queues). Henceforth we let λ_{ij}^* denote the *admission rate* from *i* to *j*, and denote the associated *total arrival rate* at station *i* by λ_i^*. The corresponding *referral rates* are denoted without star, i.e., λ_{ij} and λ_i. As shown in Table 5 below, about 48% of patients referred to *R* from *X* were actually admitted to *R*, while the corresponding percentage of patients referred to *S* was even lower (26%). Although the reasons for reneging are not discernible from the data, one can speculate from the results of the following steady-state analyses that the expected long wait may discourage patients from staying in the queue. In addition, the conditions of some patients may have improved while they were waiting, thus removing the need for further treatment.

The *routing matrix* estimated from the above data is the following:

$$R = \left[\begin{array}{ccc:c} 0 & 0.252 & 0 & 0.748 \\ 0 & 0 & 0.057 & 0.943 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

By solving equations (1) using both referral and admission rates from outside the system, we obtained the internal and total arrival rates at each station. The total referral rates to the stations were $\lambda_E = 0.674$, $\lambda_R = 2.175$, and $\lambda_S = 0.446$, while total admission rates were $\lambda_E^* = 0.674$ (unchanged), $\lambda_R^* = 1.313$, and $\lambda_S^* = 0.165$. The internal referral and admission rates are presented in Table 5, together with other input parameters used in the steady-state analyses.

[INSERT TABLE 5]

6. RESULTS OF MATHEMATICAL ANALYSIS

The case with no reneging was investigated first. Thus, all the patients at *X* who were referred to *R* or to *S* were assumed to enter the stations without reneging from the queue. In this

scenario, the traffic intensity at S turned out to be $\rho_S = \lambda_S/(c_S\mu_S) = 2.800 > 1$. This implies that, if all referred patients stay in the queue until they are admitted by S, there will be more patients arriving than exiting. In this situation, the queue becomes unbounded with time and the mean length of queue and the mean waiting time cannot be calculated. The result supports the argument that the number of supported housing beds needs to be increased for the system to be stabilized.

For station R, the traffic intensity was also above 1 [$\rho_R=\lambda_R/(c_R\mu_R) = 1.611$]. If all referred patients stay in the queue until they are admitted by R, the queue will become longer without limit. As for the *effective* service time, $1/\tilde{\mu}_R$, which would be larger than $1/\mu_R$ for any positive queue at station S, the traffic intensity is expected to be even higher. Hence the number of residential facilities must be increased if policy makers intend to accommodate all patients referred to residential facilities.

The case with reneging patients was analyzed next using the actual admission rates from X to R or to S. The results are summarized in Table 6. The steady-state analysis of station S with the admission rate ($\lambda^*_{RS} = 0.075$) revealed that the traffic intensity was below 1, but as high as 0.988. The mean queue length and the mean waiting time to enter S were 60 patients ($L_S = 60$) and about a year (W_S =366), respectively, and the station would be full about 72% of the time ($\Pr[n_S \geq c_S] = 0.722$). Reflecting this year-long wait, the *effective* service time at station R was calculated to be 914 days [$1/\tilde{\mu}_R = r_{RX}(1/\mu_R) + r_{RS}(1/\mu_R + W_S) = 0.943 \cdot 893 + 0.057 \cdot (893+366) = 914$]. The steady-state analysis of station R showed that the impact brought about by this increase in the mean time of bed occupancy (from 893 days to 914 days) would be drastic. Without blocking, the total number of patients waiting to enter R would be only 11 patients, as compared to 179 waiting patients in the case of blocking. The waiting times also showed a considerable difference. Without blocking, the mean waiting time is 9 days, as opposed to 136 days in the case of blocking. The station would be full more than 80% of the time instead of 32%. The impact of blocking on station E turned out to be even more significant. The mean time of bed occupancy at E increased from 60 days ($1/\mu_E$) to 94 days [$1/\tilde{\mu}_E = r_{EX}(1/\mu_E) + r_{ER}(1/\mu_E + W_R) = 0.748 \cdot 60 + 0.252 \cdot (60+136) = 94$] in the case of blocking. The results showed that, without blocking, an average of less than patient would be waiting to enter E for more than a day, while about 140 patients would be waiting more than 200 days in the case of blocking. The station would be full more than 94% of the time. The above results lead to our primary finding, that is, in Philadelphia, the shortage of supported housing has created "upstream blocking" of patients at both extended acute hospitals and residential facilities. Thus, allocating more resources to increase supported housing beds would be the most cost-efficient way to reduce congestion in the system as a whole.

[INSERT TABLE 6]

A comparison between the empirical observations (Table 1) and the analytical results shows that the empirical congestion is much less serious at all stations. There are two possible explanations for this discrepancy. First, the empirical system may be in the process of reaching the steady state. In this case, if capacities and arrival/service rates remain the same, the system congestion would intensify in the future, resulting in serious congestion at all stations in the steady states. AAS reports that the congestion at the two bottom stations, R and S have been intensified in recent years, which triggered the increase in the number of beds at these facilities. If this is the case,

Secondly, our analysis used the external arrival rates obtained from 1997-1999 data. While this data was assumed to be representative, it was later discovered from AAS that the arrival rates to community type institutions (i.e., residential facilities and supported housing) have been increasing since 1995 in parallel to the expansion of the number of the beds at these facilities. This is primarily because fewer patients renege from the queue when they learn that the

waiting time or the queue is shorter. This suggests that the arrival rates to these facilities prior to 1997 were smaller. Thus, the empirically observed waiting times are more likely to be comparable to analytical results based on parameter values with smaller arrival rates and capacities.

7. SCENARIO/SENSITIVITY ANALYSES AND PARAMETER ESTIMATION ISSUES

The results of the steady-state analysis indicated that, while congestion at supported housing is serious, the number of patients waiting to enter *R* and *S* and the associated waiting times are marginal if the bottleneck at supported housing does not exist. Therefore, we investigated how effectively the increase in the number of supported housing bed could reduce the steady-state congestion level in the system. Note that increasing other types of beds would intensify the congestion at supported housing, which, in turn, would aggravate the blockages at all upstream stations.

Figures 4-6 show the effect of increasing the number of supported housing beds on the congestion at each station. All the figures show a significant and sharp drop in mean queue lengths at the steady state. Similar dramatic decreases were observed for waiting times at each station.

[INSERT FIGURES 4, 5, and 6]

A sensitivity analysis was conducted to examine the magnitude of impacts resulting from different mean service (treatment) times at *S*. Sensitivity analysis was considered to be particularly relevant here because of possible biases in the estimated parameter values. As explained earlier, the mean service times at *R* and *S* were estimated using survival analysis under the assumption of exponential distributions. However, it is possible that this distribution does not accurately reflect the actual service time distribution. In particular, our survival analysis indicated that the services times for a substantial percentage (about 70%) of supported housing residents were censored, thus introducing considerable uncertainty into the estimates of services times at *S*. The following figures show how queue lengths at *R* and *S* vary with different values of mean service time at *S*. Not surprisingly, these figures indicate that queue lengths are quite sensitive to mean service times. This indicates that further empirical investigation of actual service time distributions is essential for establishing more reliable indicators of performance.

[INSERT FIGURES 7 and 8]

It should also be noted that the arrival rate estimates used in this study are somewhat biased. Our steady-state results in the previous section are based on the *admission rates* to *R* and *S* and do not include those patients who were referred to the facilities but reneged from the waiting lists[5]. Hence the arrival rates corresponding to the "true demand" are likely to be somewhat higher. Whether the *referral rates* reflect true demand or not is another issue. It is likely that *referral rates* depend on the congestion levels of those facilities to which patients are being referred. Referring physicians (or gatekeepers) may be more inclined to refer patients when the relevant facilities have vacancies. Thus one may conjecture that the arrival rates reflecting "true demand" lie somewhere between these referral and admission rates.

[5] The system did not support the steady state when we used the *referral rates* to *R* and *S* as the arrival rates.

8. SIMULATION ANALYSIS

While simplified mathematical models are very useful for long-run (steady-state) analyses of performance indicators, simulation models are essential for the analysis of short-term transient system performance. In conjunction with the theoretical model above, we have conducted a simulation study for two purposes; (i) to investigate the transient behavior of the system, and (ii) to test the robustness of our mathematical model. The input parameters for the study are those used in the mathematical analysis where the admission rates are used as arrival rates. The parameters are summarized in the following table.

[INSERT TABLE 7]

The commercial simulation software, ARENA, was used for the study. The length of the simulation was taken to be 150,000 days (i.e., 410 years). Here it was observed that a "warm-up" period of 10,000 days (27 years) was sufficient to negate any effects of initial conditions used for the study. ARENA computes batch means[6], creates 95% confidence intervals, and reports the "half width" of these intervals[7]. It then conducts a statistical test of independence between the adjacent batch means. In particular, it tests whether the correlation (Corr) of two adjacent batch means is zero. If this test fails at the 0.05 level, ARENA signifies this by writing "Corr" in the "half width" column[8].

Table 8 summarizes the results of the simulation analysis. The steady states resulting from this analysis produced results that are almost identical to those of the corresponding steady states for the mathematical model above – with the exception of the performance indicators for station *E*. The half-width columns and the Figure A1 in Appendix together indicate that the queue at *E* exhibits the strongest degree of autocorrelation. The reason for this appears to be due mainly to the fact that station *E* is furthest upstream, and hence is most affected by downstream congestion. This may indicate that the steady-state performance indicators obtained from the mathematical model for station *E* are less useful than for stations further down stream. Although Table 8 also indicates the correlation in the queue at *R* to enter *S*, such trend could not be observed in the plot (Figure A2 in Appendix).

[INSERT TABLE 8]

9. CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

In this study, congestion levels in the Philadelphia mental health system were analyzed using a queuing network model with blocking. The model focused on blocking phenomena

[6] For mean waiting-time statistics, observations are initially grouped into batches of size 16, and the corresponding batch means are computed. When the number of batches reaches 40, adjacent pairs are combined to create 20 larger batches and the means are recomputed. From this point on, the process continues with batches of size 32. For mean queue-length statistics the procedure is the same, where observations of queue length are made at every quarter of a time unit (such as four times a year).

[7] The confidence interval is calculated from the *t*-distribution as $\bar{x} \pm t_{n-1,.025} \frac{s}{\sqrt{n}}$, where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the number of batches, and the second term is the "half width" of the interval where $t_{n-1,.025}$ is the critical point from Student's t distribution with $n-1$ degree of freedom.

[8] The data collected during the warm-up period are not used to calculate the confidence intervals [24].

between three types of mental facilities, namely, extended acute hospitals, residential facilities and supported housing. In addition, the flows entering the system from both acute hospitals (i.e., general hospitals) and the community at large were examined. Our study is new in two aspects: (i) This is the first application of queueing with blocking to analyze the congestion in a mental health system; (ii) We propose this queueing approach for the first time as an alternative to the conventional needs assessment used by planners for purposes of mental health resources allocation. Our queuing approach addresses the drawbacks that are inherent to the needs assessment approach by introducing a dynamic model to replace a static model, and also capturing interdependencies between the various treatment settings.

Our model was analyzed both in terms of derived steady states and numerical simulations. Both of these analyses showed that for the parameters estimates used, all types of facilities in the system experience serious congestion in the resulting steady states. Perhaps the single most important finding is that, in contrast to popular perception, system congestion is not always a simple cumulative effect of shortages across all facility types. In the Philadelphia case, the model suggests that system-wide congestion is due primarily to shortages in one specific facility type, namely, supported housing. Here the shortage of supported housing in Philadelphia has created "upstream blocking" of patients at both extended acute hospitals and residential facilities. From a policy viewpoint, this analysis suggests that removal of such facility-specific bottlenecks may often be the most cost-efficient way to reduce congestion in the system as a whole. A second important finding is that current congestion levels in the Philadelphia system appear to be significantly lower than those predicted by the model at steady state. Hence the model suggests that congestion levels in this system may continue to increase for some time. In fact, it has been reported that congestion levels in the Philadelphia metal health system have increased across the board in recent years, leading to increases in the number of beds at various community type facilities.

However it should be emphasized that the analysis presented here is only a first attempt at modeling this complex system. Thus the conclusions above should be regarded as tentative at best. For example, the gap between observed congestion and the results of our study may in part be due to our assumption of constant arrival rates based on 1997-1999 data. In fact, AAS findings indicate that both arrival rates at residential facilities and supported housing have been increasing since 1995, implying that the arrival rates to these facilities prior to 1997 were smaller. Hence the steady-state congestion levels predicted by our model may in fact be too high.

This suggests that one important direction for extending the present model would be to allow parameters such as arrival rates to change over time. For example, reneging behavior can in fact be viewed as changes in arrival rates as a response to queue lengths. As mentioned in section 7 above, there may also be a change in referral rates by physicians (or gatekeepers) in response to changing vacancy rates. While such interaction effects are often difficult to model formally, they are easily introduced into simulation models. In addition, simulation study is much less restrictive in terms of the probability distributions that can be employed to characterize both arrival and service patterns. Hence to achieve more realistic models of system behavior, our future work will focus more heavily on simulation.

Finally, it should be pointed out that the results presented here suggest that transient behavior may in fact be more important than steady-state behavior for mental health systems. In particular, the severe autocorrelation observed at upstream stations suggest that steady states may not even be meaningful within a practical time frame. Hence such findings serve to reinforce the need for more detailed simulation studies of mental health systems.

TABLES AND FIGURES

LIST OF TABLES

LIST OF FIGURES

Table 1. Summary of Empirically Observed Congestion

Destination	Average Waiting Days
Extended Acute Hospitals	23
Residential Facilities	33
Supported Housing	105

Source: OMH community residential rehabilitation reporting system. Housing Corporation Annual Report.

Table 2. Flows in the System and Congestion Types

Flow	Cause of Congestion	Station Facing Congestion	Characteristics	Congestion Type
A to *E* *A* to *R*	*E* is full *R* is full	*A*	(i) only	Classic congestion
E to *R* *R* to *S*	*R* is full *S* is full	*E* *R*	(i) and (ii)	Blocking
E to *X* *R* to *X* *S* to *X*	Not Applicable	Not Applicable	None	No congestion
X to *R* *X* to *S*	*R* is full *S* is full	*X*	(i) only	Classic congestion

Table 3. Service Time Estimate for Station *R*

The Model Fit				
Prob > chi2	< 0.0001			
Parameter	**Estimate**	**Std. Error**	**Confidence Interval**	
μ (service rate)	.00112	.00003	.0011 to .0012	
Median	617.6466	18.2283	581.9190 to 653.3742	
Percentiles of Survival Distribution				
Survival	.25	.50	.75	.95
Time	1235.29	617.65	256.35	45.71

Table 4. Service Time Estimate for Station *S*

The Model Fit				
Prob > chi2	< 0.0001			
Parameter	**Estimate**	**Std. Error**	**Confidence Interval**	
μ (service rate)	.00040	.00004	.0003 to .0005	
Median	1719.95021	173.16541	1380.5460 to 2069.3544	
Percentiles of Survival Distribution				
Survival	.25	.50	.75	.95
Time	3439.90	1719.95	713.84	127.28

Table 5. Input Parameters for Steady-State Analysis

Station	Parameter	Values	Description
E	$\lambda_E = \lambda_{AE}$	0.674	External arrival rate from *A* to *E*
	c_E	64	Number of beds at *E*
	$1/\mu_R$	60	Mean treatment time (days) at *E*
R	λ_{AR}	0.337	External arrival (= referral) rate from *A* to *R*
	λ_{XR}	1.668	Arrival (= referral) rate from *X* to *R*
	λ^*_{XR}	0.806	Admission rate from *X* to *R*
	λ_{ER}	0.170	Arrival (= referral) rate from *E* to *R*
	c_R	1206	Number of beds at *R*
	$1/\mu_R$	893	Mean treatment time at *R* Estimated using survival analysis
S	λ_{XS}	0.342	Arrival (= referral) rate from *X* to *S*
	λ^*_{XS}	0.090	Admission rate from *X* to *S*
	λ_{RS}	0.124	Arrival (= referral) rate from *R* to *S*
	λ^*_{RS}	0.075	Admission rate from *R* to *S*
	$1/\mu_S$	2500	Mean treatment time at *S* Estimated using survival analysis
	c_S	416	Number of beds at *S*

Table 6. Outputs of Steady-State Analyses

Station		Performance Indicator	Mean	Traffic Intensity (ρ_i)	$\Pr(n_i \geq c_i)$
E	L_E	The queue length to enter E	137 (< 1)	0.993 (0.630)	0.936 (0.0001)
	W_E	The waiting time to enter E (days)	203 (< 1)		
R	L_{AR}	The queue length for the flow $A \rightarrow R$	46 (3)	0.995 (0.972)	0.811 (0.315)
	L_{ER}	The queue length for the flow $E \rightarrow R$	23 (1)		
	L_{XR}	The queue length for the flow $X \rightarrow R$	110 (7)		
	L_R	The queue length to enter R	179 (11)		
	W_R	The waiting time to enter R (days)	136 (9)		
S	L_{RS}	The queue length for the flow $A \rightarrow R$	27	0.988	0.722
	L_{XS}	The queue length for the flow $A \rightarrow R$	33		
	L_S	The queue length to enter S	60		
	W_S	The waiting time to enter S (days)	366		

Note: The numbers in parentheses present the results of "no blocking" scenarios.

Table 7. Input Parameters for Simulation Analysis

Station	Parameter	Values	Description
E	$\lambda_E = \lambda_{AE}$	0.674	External arrival rate from A to E
	c_E	64	Number of beds at E
	$1/\mu_R$	60	Mean treatment time (days) at E
R	λ_{AR}	0.337	External arrival (= referral) rate from A to R
	λ^*_{XR}	0.806	Admission rate from X to R
	c_R	1206	Number of beds at R
	$1/\mu_R$	893	Mean treatment time at R Estimated using survival analysis
	λ^*_{XS}	0.090	Admission rate from X to S
	$1/\mu_S$	2500	Mean treatment time at S Estimated using survival analysis
	c_S	416	Number of beds at S

Table 8. Summary of Simulation Results

Station	Statistic	Average	Half Width	Minimum	Maximum
E	L_{AE}	237.68	Corr	0.00	598.00
	W_{AE}	352.07	Corr	0.00	865.99
R	L_{AR}	46.33	0.78	19.00	81.00
	L_{XR}	110.04	0.24	97.00	477.00
	L_{ER}	23.25	0.33	7.00	39.00
	W_{AR}	136.66	1.11	91.76	84.54
	W_{XR}	135.43	1.00	89.59	186.63
	W_{ER}	136.13	1.08	93.24	184.22
S	L_{XS}	32.63	0.02	26.00	208.00
	L_{RS}	28.28	Corr	7.00	56.00
	W_{XS}	359.72	12.42	169.42	570.36
	W_{RS}	369.59	9.74	165.43	582.17

Figure 1. In-flows and Out-flows between Stations

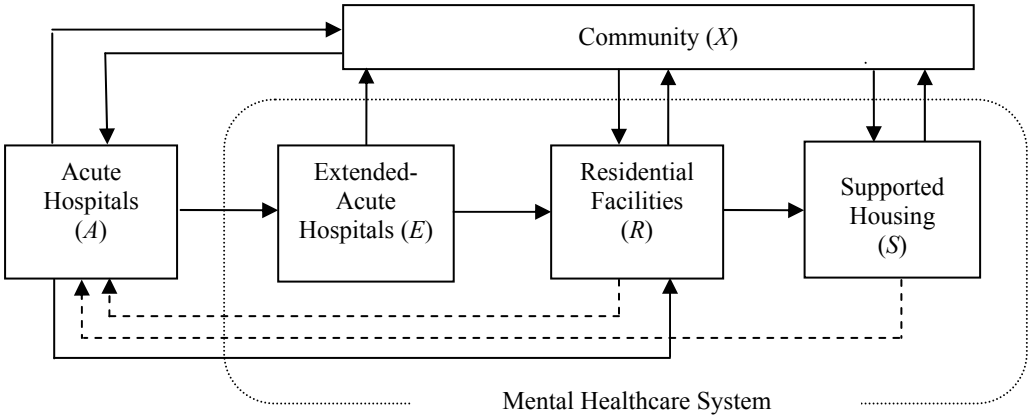


Figure 2. Tandem Two-Station System with No Buffer

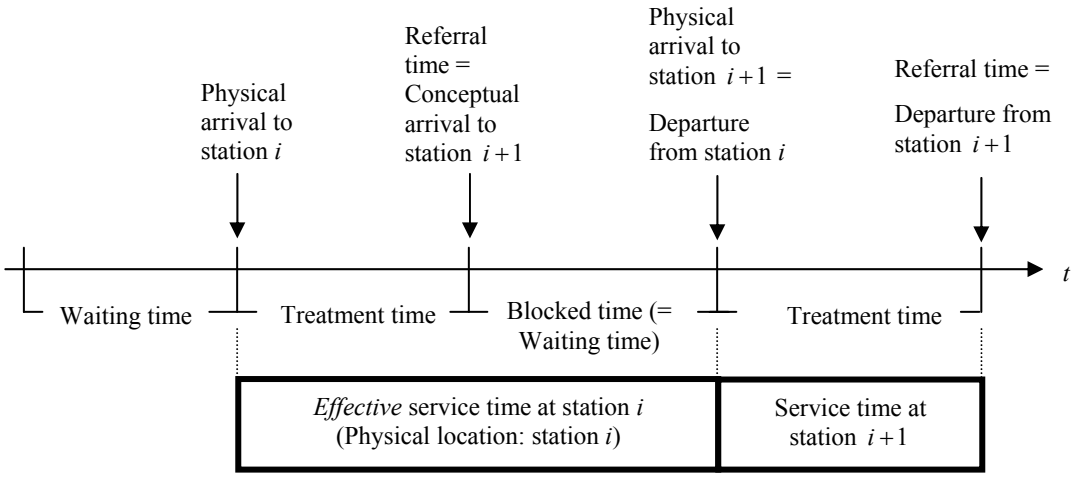


Figure 3. Algorithm to obtain Steady-States with Blocking

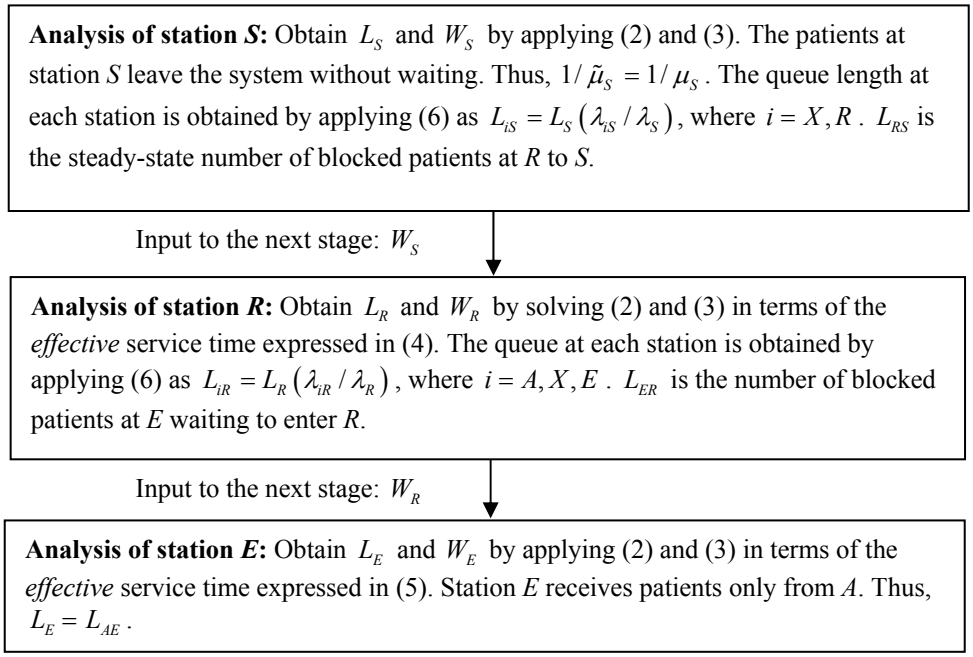


Figure 4. Congestion at Station *E*

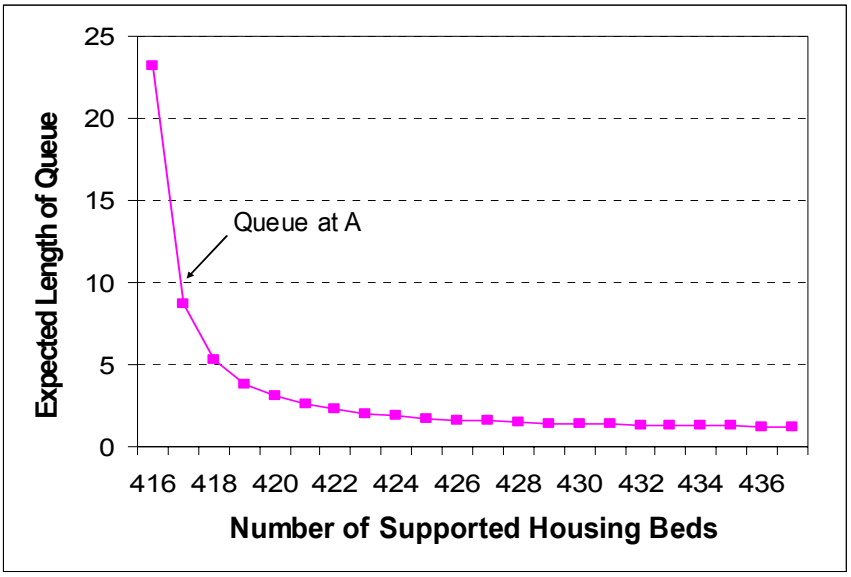


Figure 5. Congestion at Station *R*

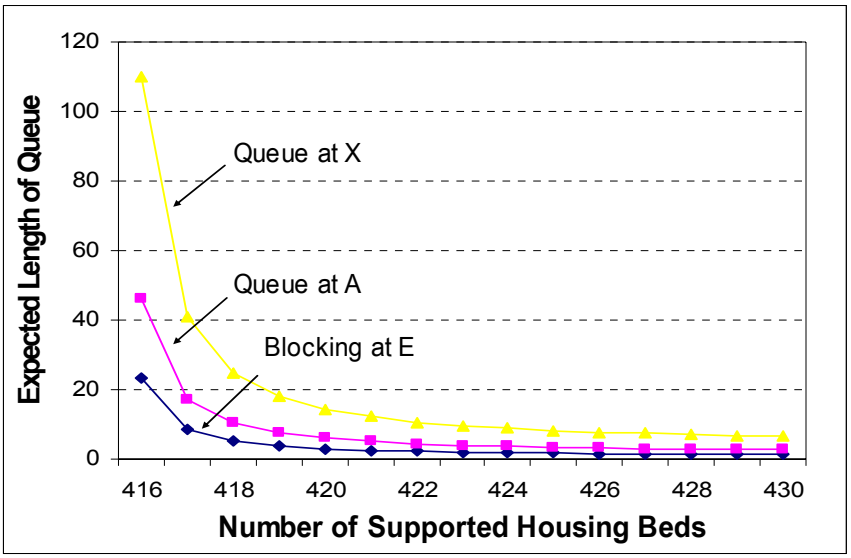
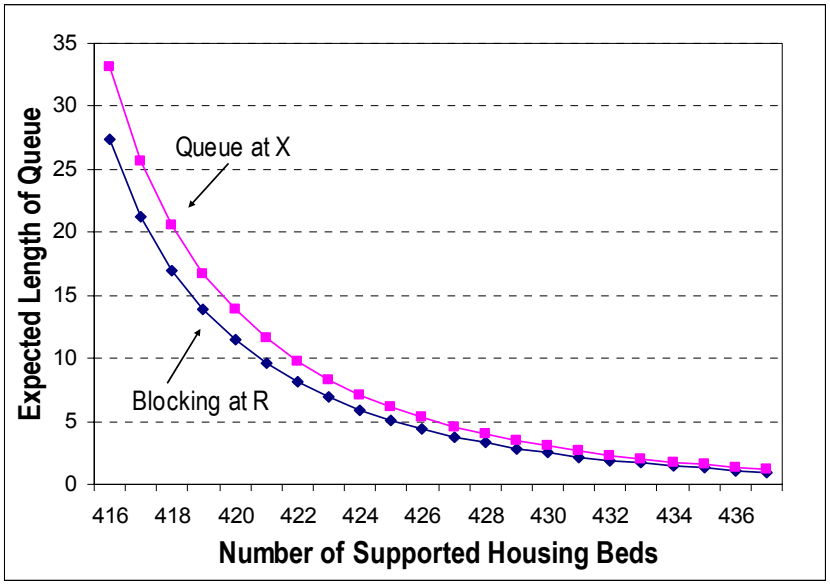


Figure 6. Congestion at Station *S*



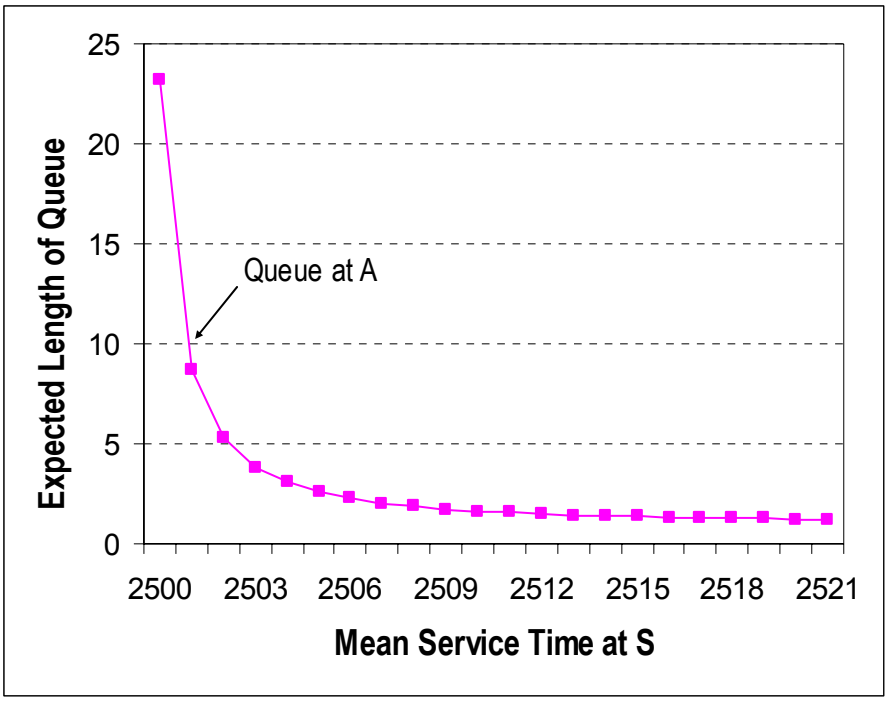

25
20
15
10
5
0
Expected Length of Queue
Queue at A
2500
2503
2506
2509
2512
2515
2518
2521
Mean Service Time at S

Figure 7. Congestion at Station *R*

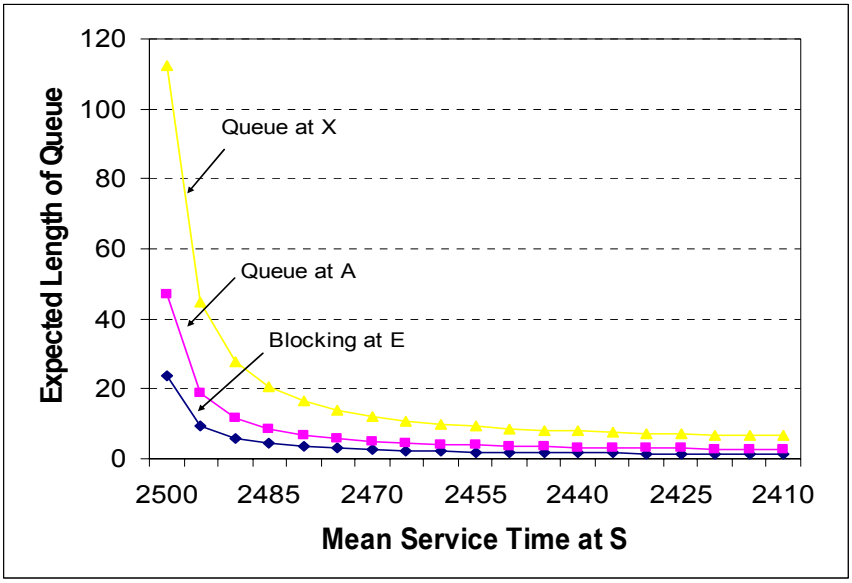
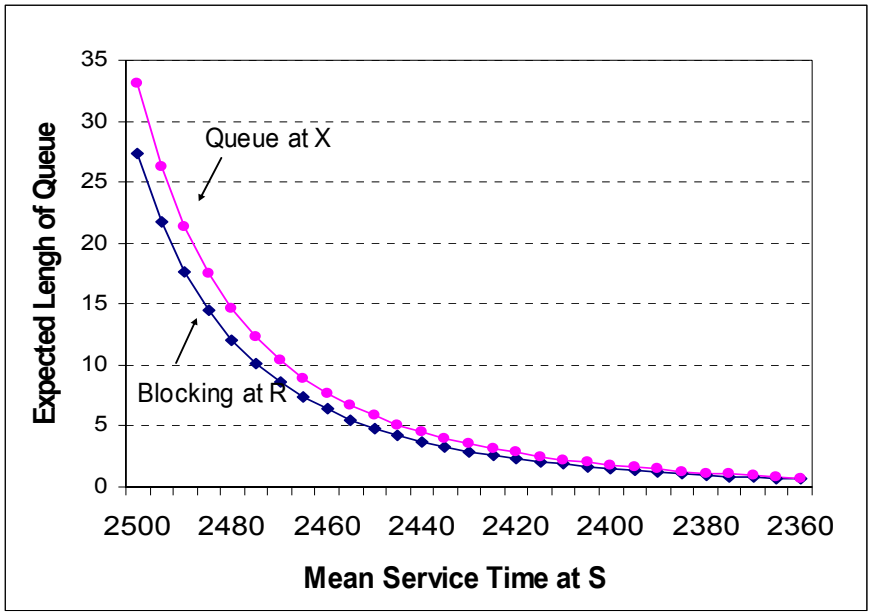


Figure 8. Congestion at Station *S*



REFERENCES

[1] NASMHPD Research Institute (State Profile Highlight, No. 1, Aug 2000)

[2] Randolph FL. Ridgway P. Carling PJ., Residential programs for persons with severe mental illness: a nationwide survey of state-affiliated agencies, (Hospital & Community Psychiatry, 42(11) 1991) 1111-1115.

[3] Rothbard, Aileen B; Kuno, Eri; Schinnar, Arie P; Hadley, Trevor R; Turk, Roland, Service utilization and cost of community care for discharged state hospital patients: A 3-year follow-up study, (American Journal of Psychiatry, 156(6), 1999) 920-927.

[4] Hunt, G.C., Sequential Arrays of Waiting Lines, (Operations Research, 4, 1956) 674-683.

[5] Hillier, F.S. and Boling, R.W, Finite Queues in Series with Exponential or Erlang Service Times – A numerical Approach, (Operations Research, 15, 1967) 286-303.

[6] Takahashi, Y., Miyahara, H., Hasegawa, T., An Approximation Method for Open Restricted Queuing Networks, (Operations Research, 28, 1980) 594-602.

[7] Perros, H.G. and Altiok, T., Approximate Analysis of Open Networks of Queues with Blocking: Tandem Configurations, (IEEE transactions on software engineering, 12, 1986) 450-461.

[8] Brandwajn, A. and Jow Y.L., An Approximation Method for Tandem Queuing Systems with Blocking, (Operations Research, 1, 1988) 73-83.

[9] Brandwajn, A. and Jow Y.L., Tandem Exponential Queues with Finite Buffers, (Proceedings of International Seminar on Computer Networking and Performance Evaluation, Tokyo, Sept 18-20, 1985).

[10] Korporaal, R., Ridder, A., Kloprogge P. and Dekker, R., An Analytic Model for Capacity Planning of Prisons in the Netherlands, (Journal of the Operations Research Society, 51, 2000) 1228-1237.

[11] El-Darzi, et al., A Simulation Modelling Approach to Evaluating Length of Stay, Occupancy, Emptiness and Bed Blocking in a Hospital Geriatric Department, (Health Care Management Science, 1, 1998) 143-149.

[12] Cohen, M.A., Hershey, J.C., Weiss, E.N., Analysis of Capacity Decisions for Progressive Patient Care Hospital Facilities, (Health Services Research, 15, 1980) 145-160.

[13] Gross, D. and Harris, C., Fundamentals of Queuing Theory – 3rd edition, (John Wiley & Sons, 1998)

[14] Hershey, J.C., Weiss, E.N. and Cohen, M.A., A Stochastic Service Network Model with Application to Hospital Facilities, (Operations Research, 29, 1981) 1-22

[15] Weiss, E. N. and McClain, J.O., Administrative Days in Acute Care Facilities: A Queuing-Analytic Approach, (Operations Research, 35, No. 1, 1987) 35-44.

[16] Disney, R. L., Queuing Networks, (Proceedings of Symposia in Applied Mathematics, American Mathematics Society, 25, 1981) 53 – 83.

[17] Jun, K-P and Perros, H.G., (Proceedings of First International Workshop on Queuing Network with Blocking, North-Holland, Amsterdam 1989) 259-280.

[18] Jackson, J. R., Networks of Waiting Lines, (Operations Research, 5, 1957) 518-521.

[19] Jackson, J.R., Jobshop-like Queuing Systems, (Management Science, 10, 1963) 131-142.

[20] Burke, P.J., The Output of a Queuing System, (Operations Research, 4, 1956) 699- 714.

[21] Melamed, B., Characterization of Poisson Traffic System in Jackson Queuing Networks, (Advances in Applied Probability, 11, 1979) 422-438.

[22] Little, J. D. C., A Proof for the Queuing Formula L=λW, (Operations Research, 9, 1961) 383-387.

[23] Perros, H. G., Queuing Networks with Blocking, (Oxford University Press, New York, 1994).

[24] Kelton, W.D., Sadowski, R.P., & Sadowski, D.A., Simulation with Arena – 2nd Edition, (McGraw-Hill, Washington, DC, 2001).

APPENDIX

Figure A1. Queue Length *A*→*E*

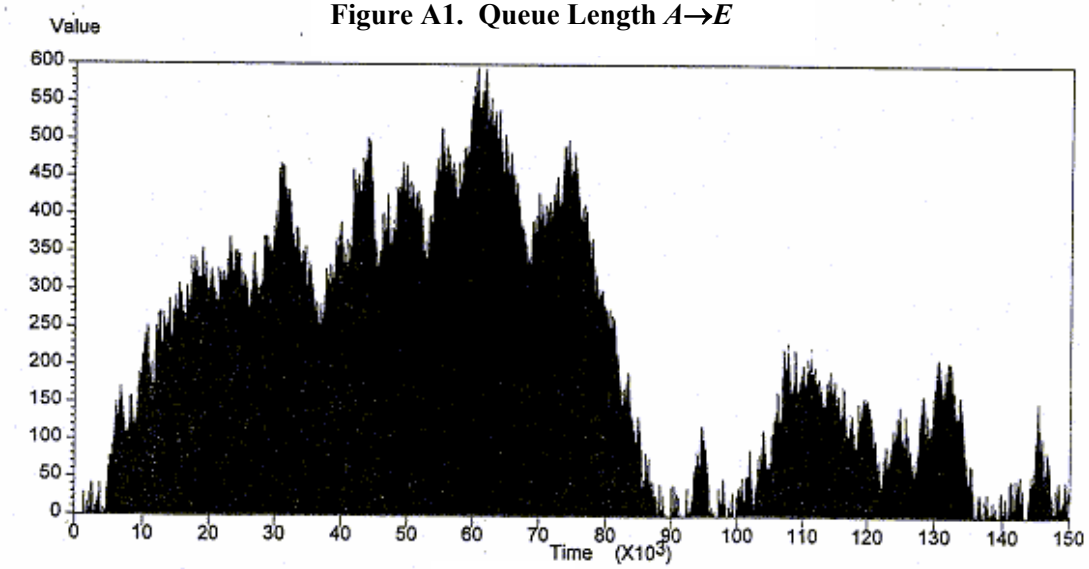


Figure A2. Queue Length *R*→*S*