Croatian Operational Research Review
CRORR **16:2**(2025), 191–203


# Aspects of blocking on time-varying tandem queueing network

**Anjale Ramesh**[1,*] **and M. Manoharan**[1]

[1] *Department of Statistics, University of Calicut, Kerala-673 635, India*
*E-mail: ⟨{anjaleramesh, manumavila}@gmail.com⟩*

**Abstract.** Most real-world service systems are susceptible to queue capacity constraints, leading to a blockage of entities. Several blocking mechanisms can be implemented when capacity constraints influence the flow of entities through the system. This study investigates the comparative performance of blocking-after-service (BAS) and blocking-before-service (BBS) mechanisms by modelling a hospital emergency department with a finite capacity queue. The model is based on a two-station tandem network with finite capacity on the intermediate queue, and we developed transient performance measures for the system in both mechanisms. Using a numerical approach, we highlight how these mechanisms influence the time-varying number of patients and the virtual workload in the system. Our results demonstrate that the BAS mechanism slightly outperforms the BBS mechanism in reducing unwanted congestion.

**Keywords**: blocking-after-service (BAS), blocking-before-service (BBS), tandem queues, transient analysis, virtual workload

Received: August 7, 2024; accepted: January 22, 2025; available online: April 2, 2025

DOI: 10.17535/crorr.2025.0016

**Original scientific paper.**

## 1. Introduction

In queueing theory, capacity restriction on waiting line is a crucial aspect to study and it is common for real-world service systems to have queues of finite capacity. In such systems, the flow of customers from the source node will be blocked if the waiting room at the destination node is full. There are mainly two blocking mechanisms, i.e. blocking-after-service (BAS) and blocking-before-service (BBS) that describe different scenarios when there are restrictions on waiting rooms. BAS occurs when an entity after service from a node finds that waiting room of the next station is full (saturated). So, they are being blocked before entering the waiting room of the next node. They must have to wait until the space becomes available. In BBS system, before starting service at current node, entities are blocked if there is no available space in the waiting room of next station. Once the space is available at the destination node, the blocked entity resumes service at the source node.

Tandem queueing networks with finite capacity are useful for modelling healthcare, communication, and manufacturing systems [6, 13, 14, 10]. A BAS mechanism is also known as manufacturing blocking. In manufacturing and production lines, items move through the workstations which can only process a limited number of items at a time. If the next workstation is full, it is not possible to move the items that have been processed from the current workstation, leading to a blockage. In [2], a steady-state analysis was performed under the BAS mechanism with two single-server queues connected in tandem. In [1], the same model is extended to a k-station tandem network with general arrival times, deterministic service times and finite

*Corresponding author.

waiting room between stations. Transient behaviour of a two station tandem network with no restriction on first station and no queue allowed for second station was investigated in [11]. Zychlinski et. al [18] developed time-varying fluid models for tandem network with a general time-varying arrival rate, a finite waiting room before first station and no intermediate waiting room. The BBS mechanism, which is also known as communication blocking, is commonly used in telecommunication networks [15, 8]. A detailed description of the different types of BBS mechanisms is presented in [5]. In communication networks, data packets move through a series of nodes/router, where each node processes the packets and forward it to the next node. If the buffer of the next node is full, the current node cannot forward the packet causing the packet to be blocked before the current node. Healthcare systems can also use the BBS mechanism in short medical procedures, such as cataract surgery, laparoscopic surgery and cardiac catheterization. These procedures can only begin when the room is available in the recovery area. Avi-Itzhak and Levy[4] introduced a k-stage blocking scheme as a generalisation of the results presented in [2, 1]. In [3], they analyzed the steady-state performance measures of a single-server network of k stations with no intermediate queue and an unlimited buffer prior to the first station under both BAS and BBS mechanisms. Fluid limits for tandem model of time-varying multi-server queues with finite buffers before the first station and between stations under BBS mechanism is considered in [19]. To facilitate comparison, they also developed steady-state closed-form expressions for system performance measures under the BAS and BBS mechanisms.

There has been extensive research conducted on tandem networks with finite capacity queues. However, there is limited research on time-varying tandem queues with blocking. In this study, we provide an analytical comparison between BBS and BAS in time-varying tandem queues, with special reference to a case of healthcare system. We develop a stochastic model for a two station finite capacity tandem network under different blocking mechanisms. In the second section, explicit expressions for transient performance measures such as number of patients and average virtual workload at time $t$ under BAS and BBS mechanisms are discussed. We also conducted a numerical study with the blocking mechanisms under different traffic intensity and queue capacity.

## 2. Two station Tandem Model with finite queue capacity

We consider a health care system, such as a hospital emergency department with a triage system where patients arrive according to a non-homogeneous Poisson process. The model considered here is a tandem network with two stations. Patients are first assessed by a triage nurse to determine their level of need for medical assistance. Subsequently, the patients are sent to the consultation room, where medical professionals provide the necessary treatment. There is an unlimited waiting room for triage and finite waiting room for treatment. Suppose that only a limited number of patients can wait in the treatment area due to space and resource constraints. In the above situation, there are two ways to manage heavy traffic of patients. In BAS, if the waiting space of treatment room is full (saturated), patients cannot join the queue for treatment after the triage process, causing a blockage. In BBS, even though the triage nurse is present, the treatment waiting room is full (saturated), preventing patients from being triaged and causing them to be blocked. In this section, we establish transient performance measures for the two-station tandem queueing network model, under the two blocking mechanisms. Ramesh and Manoharan [12] derived explicit expressions for time-varying measures, such as queue length and virtual workload, for a time-varying tandem queueing network of k stations. Building on that work, this paper extends these measures by incorporating the effects of blocking within the context of an outpatient clinic to facilitate meaningful comparisons.

## 2.1. Blocking After Service (BAS)

Initially, we model a healthcare system with a two-station tandem queueing network with finite queue capacity in second station. In the first-come-first-served (FCFS) model illustrated in Figure 1, patients arrive according to non-homogeneous Poisson process and the time-dependent service time following an exponential distribution, i.e. triage node is $M(t)/M(t)/1/\infty$ and treatment node is $M(t)/M(t)/1/K$. Following are the parameters that characterise the model.

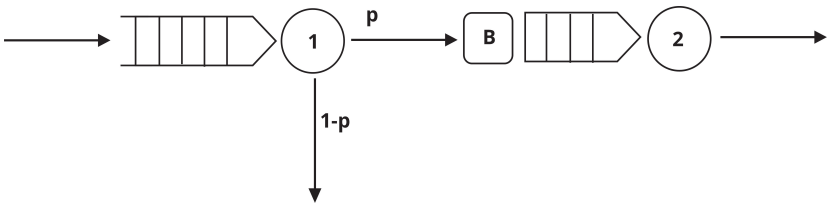



Figure 1: *A BAS two station tandem queueing network model*

1. $\{A_1(t), t \geq 0\}$ is the external arrival process to the triage node with arrival rate $\lambda_1(t)$.

2. $\{V(t), t \geq 0\}$ is the service requirement of the patient, i.e., the total amount of service (in terms of time) that a patient requires. $V_1(t)$ is the service requirement of a patient arriving at the triage at time $t$. Similarly, $V_2(t)$ is the service requirement of a patient arriving at the treatment area at time $t$. $\mu_i(t),\ i = 1, 2$ is the service processing rate at time $t$.

3. After triage process, patients move to the treatment area with probability $p$ and leave the system with probability $1 - p$ and at a rate of $(1 - p)\mu_1(t)$.

4. $\{A_2(t), t \geq 0\}$ denote the arrival of patients to the treatment area after triage process. Therefore, the arrival rate is $\lambda_2(t) = p\mu_1(t)$.

   The instantaneous traffic intensity at the triage node, $\rho_1(t)$ is defined as,

   $$\rho_1(t) = \lambda_1(t)/\mu_1(t).$$

   Thus, the two arrival rate functions are related as $\lambda_2(t) \ = p\ \lambda_1(t)/\rho_1(t)$.

5. In this study, the instantaneous traffic intensity, $\rho(t) = \lambda(t)/\mu(t)$, which measures the utilisation of a service system is chosen to be invariant of time. This adaptation is made to choose service rate function properly in order to adjust with the arrival rate and traffic intensity.

   We use the principle of rate-matching control, as discussed in [16], to determine the service rate function. In rate-matching control, the service rate is set to be proportional to the arrival rate for a fixed traffic intensity $\rho$. Thus, for a constant traffic intensity $\rho_i$, the time-dependent service rate function, $\mu_i(t)$ can be written as,

   $$\mu_i(t) \equiv \lambda_i(t)/\rho_i, \quad i = 1, 2 \quad t \geq 0. \tag{1}$$

6. There is an infinite waiting room at the triage node and a finite waiting room at the treatment node with a maximum capacity of $K$. If the finite buffer at treatment node is full, patients will be blocked. When the capacity of the waiting room of treatment node is $K - 1$, blocked patients will join the queue for consultation.

The following are the formulations of some transient performance measures associated with the model under consideration.

1. $\{W_1(t), t \geq 0\}$ is the waiting time of a patient who arrives at triage node at time t. An explicit expression for the probability distribution of waiting time $W(t)$ for $M(t)/M(t)/1/\infty$ is derived by Whitt[16]. If a patient who arrives at time $s$ is still waiting for service in the queue at time $t$, then we can express the probability that the waiting time of the patient who arrives at time $s$, is larger than $t-s$, for $0 \leq s \leq t$ as,

$$P(W_1(s) > t-s) = \rho_1 e^{-((1-\rho_2)\Lambda_{t,1}(s))/\rho_1}, \tag{2}$$

where $\Lambda_{t,1}(s) = \Lambda_1(t) - \Lambda_1(s)$, $\Lambda(.)$ is the cumulative arrival rate function defined as,

$$\Lambda_1(u) = \int_0^u \lambda_1(r)\ dr, \quad r \geq 0 \tag{3}$$

and $\Lambda_{t,1}(u)$ need to be strictly increasing and continuous, see [16].

2. $\{W_2(t), t \geq 0\}$ is the waiting time of a patient who joins the queue of treatment area at time $t$. Since the queue capacity is finite, i.e. $M(t)/M(t)/1/K$, here we derive a closed form expression for probability distribution of waiting time.

   Let $P_n$ be the probability that there are $n$ patients in the queue in front of treatment area, see [9].

$$P_n = \frac{(1-\rho_2)\ \rho_2^n}{1-\rho_2^{K+1}},$$

   where $\rho_2$ and $K$ are the traffic intensity and queue capacity of the treatment area. Let $Q_n$ be the probability of the arrival point, that is, the probability that there are $n$ patients in the queue at the time of arrival, for $n < K$. This is derived in Gross and Harris[9] using Baye's theorem.

$$Q_n = \frac{P_n}{1-P_{K+1}}. \tag{4}$$

   Then the probability distribution of waiting time for the stationary $M/M/1/K$ system can be obtained by reducing the expression for multi-server system in [9] to single server, i.e.,

$$\begin{aligned} P\{W > t\} &= \sum_{n=1}^{K-1} Q_n \sum_{i=0}^{n-1} \frac{(\mu t)^i\ e^{-\mu t}}{i!} \\ &= \sum_{n=1}^{K-1} Q_n \sum_{i=0}^{n-1} \frac{(\lambda t/\rho)^i\ e^{-(\lambda t/\rho)}}{i!}. \end{aligned} \tag{5}$$

   The parameter $\mu$ is replaced by $\lambda/\rho$. "From this, the waiting time distribution for a non-stationary system can be derived using Corollary 5.1 from Whitt [16]. Specifically, under the assumptions of a non-stationary system, $\lambda t$ in equation (5) becomes the cumulative arrival rate function $\Lambda(t)$.

   For the non-stationary system $M(t)/M(t)/1/K$, let $P(W_2(s) > t-s)$ denote the probability that the waiting time of a patient joining the queue of the treatment node at time $s$ exceeds $t-s$, for $0 \leq s \leq t$,

$$P(W_2(s) > t-s) = \sum_{n=1}^{K-1} Q_n \sum_{i=0}^{n-1} \frac{(\frac{\Lambda_{t,2}(s)}{\rho_2})^i\ e^{\left(\frac{\Lambda_{t,2}(s)}{\rho_2}\right)}}{i!}. \tag{6}$$

The cumulative rate function $\Lambda_{t,2}(s) = \Lambda_2(t) - \Lambda_2(s)$ with

$$\Lambda_2(u) = \int_0^u \lambda_2(r)\ dr, \quad r \geq 0. \tag{7}$$

This gives the waiting time distribution for an $M(t)/M(t)/1/K$ system under constant traffic intensity.

3. To account for the effects of blocking, we incorporate a blocking probability into the formulation of transient performance measures, i.e. the probability that a patient arriving at time $s$ is blocked after triage at time $t$. In other words, probability that the blocking time of a patient who arrive at time $s$ is greater than $t$-$s$, $P\{B(s) > t - s\}$.

4. $L_1(t)$ represents the number of patients present at the triage node. These patients arrived during the interval $[0, t]$ and have not yet completed their service. Mathematically, this corresponds to the arrivals $\{A_1(s), \quad 0 \leq s \leq t\}$, i.e.,

$$L_1(t) = \int_0^t (I_{\{W_1(s)>t-s\}})\ dA_1(s),$$

where $I_{\{W_1(s)>t-s)\}}$ denotes the number of patients who entered the queue in front of the triage node at time $s$ and are still waiting for service at time $t$, $0 \leq s \leq t$.

By using Campbell–Mecke formula in [7] for taking expectations of stochastic integrals, we get the average number of patients present at the triage node at time t, i.e.,

$$E(L_1(t)) = \int_0^t (P\{W_1(s) > t - s\})\ \lambda_1(s)\ ds, \tag{8}$$

where $E(I_{\{W_1(s)>t-s\}}) = P\{W_1(s) > t - s\}$ and $E(dA_1(s)) = \lambda_1(s)$ ds.

5. $L_2(t)$ denotes the number of patients at treatment area, including those in the blocking space. These patients completed their service at triage node and moved to treatment area during the interval $[0, t]$. They are either waiting to join the queue in front of the treatment area or already in the queue. This corresponds to the arrivals $\{A_2(s), \quad 0 \leq s \leq t\}$, i.e.,

$$L_2(t) = \int_0^t (I_{\{W_2(s)>t-s\}} + I_{\{B(s)>t-s\}})\ dA_2(s),$$

where $I_{\{W_2(s)>t-s)\}}$ represents the number of patients who entered the queue of the treatment area at time $s$ and are still waiting for service at time $t$, $0 \leq s \leq t$. Similarly, $I_{\{B(s)>t-s\}}$ denotes the number of patients who entered the blocking space in front of the treatment area at time $s$ and are still waiting to join the queue of treatment area at time $t$.

While applying expectations on both sides, Campbell–Mecke formula together with additive property of expectation we get $E(I_{\{W_2(s)>t-s\}} + I_{\{B(s)>t-s\}}) = P\{W_2(s) > t - s\} + P\{B(s) > t - s\}$. Therefore,

$$E(L_2(t)) = \int_0^t \left(P\{W_2(s) > t - s\} + P\{B(s) > t - s\}\right)\ p\ \mu_1(s)\ ds. \tag{9}$$

This represents the average number of patients present in both the blocking space and the queue of the treatment area at time $t$.

6. $Z_1(t)$ represents the time required to triage all patients who arrived at first node up to time $t$, i.e.

$$Z_1(t) = \int_0^t I_{\{W_1(s)>t-s\}}\ V_1(s)\ dA_1(s) + \int_0^t \frac{V_1(s)^2}{2}\ dA_1(s).$$

While applying expectations on both sides, Campbell–Mecke formula, we get,

$$E(Z_1(t)) = \int_0^t P\{W_1(s) > t-s\}\ E(V_1(s))\ E(A_1(s))\ ds + \int_0^t \frac{E(V_1(s)^2)}{2}\ E(A_1(s))ds.$$

Introducing another term, squared coefficient of variation $c^2 = Var(V(s))/E(V(s))^2$. Using the formula of variance, this can be rewritten as,

$$c^2 = E(V(s)^2)\ (E(V(s)))^2 - 1 = \frac{E(V(s)^2)}{\mu(s)^2} - 1. \tag{10}$$

Therefore the average workload at the triage node at time t can be represented as,

$$E(Z_1(t)) = \int_0^t P\{W_1(s) > t-s\}\ \frac{\lambda_1(s)}{\mu_1(s)}\ ds + \int_0^t \frac{c_1^2+1}{2}\ \frac{\lambda_1(s)}{\mu_1(s)^2}ds, \tag{11}$$

where $c_1^2$ is the squared coefficient of variation or relative variability in service times at triage node.

7. $Z_2(t)$ denotes the time required to complete the consultation of patients who arrived up to time t after triage, taking into account the queue and blocking space, i.e.

$$Z_2(t) = \int_0^t \left(I_{\{W_2(s)>t-s\}} + I_{\{B(s)>t-s\}}\right)\ V_2(s)\ dA_2(s) + \int_0^t \frac{V_2(s)^2}{2}\ dA_2(s).$$

By using campell-Mecke formula and additive property of expectation, $E(Z_2(t))$ can be written as,

$$E(Z_2(t)) = \int_0^t \big(P\{W_2(s) > t-s\} + P\{B(s) > t-s\}\big)E(V_2(s))E(A_2(s))ds + \int_0^t \frac{E(V_2(s)^2)}{2}\ E(A_2(s))ds.$$

Let $c_2^2$ be the squared coefficient of variation of service times in treatment node and applying equation (10),

$$E(Z_2(t)) = \int_0^t \left(P\{W_2(s) > t-s\} + P\{B(s) > t-s\}\right)\ \frac{\lambda_2(s)}{\mu_2(s)}\ ds + \int_0^t \frac{c_2^2+1}{2}\ \frac{\lambda_2(s)}{\mu_2(s)^2}ds \tag{12}$$

This represents the average workload at the treatment node, taking into account the patients in the blocking space and the queue at time t.

## 2.2. Blocking Before Service (BBS)

Here, we examine the application of the BBS mechanism in a hospital emergency department by modelling it with a two-station tandem queueing network. This model, illustrated in Figure 2, is built under the non-stationary Markovian assumption, and patients are served according to the FCFS discipline.

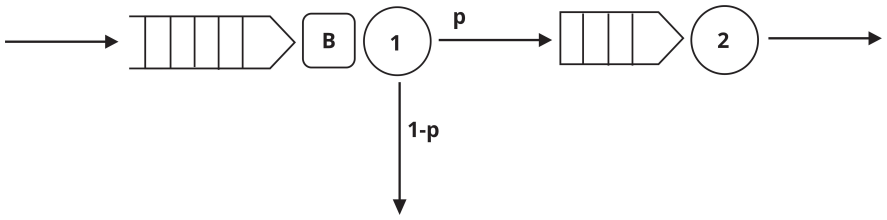



Figure 2: *A BBS two station tandem queueing network model*

Similar to the BAS tandem model, $\{A_i(t), t \geq 0\}, i = 1, 2$ is the arrival process with arrival rates $\lambda_1(t)$ and $\lambda_2(t) = p\ \mu_1(t)$. $p$ represents the probability of moving to treatment after triage, while patients leave the system with probability $1 - p$. $\{V_i(t), t \geq 0\}, i = 1, 2$ is the service requirement of a patient arriving at the triage node and treatment node at time $t$ with service processing rate at time $t$, $\mu_i(t),\ i = 1, 2$. An infinite waiting room is available for patients at the triage node, whereas the treatment node has a finite waiting room with a maximum capacity of $K$.

Before providing service from triage node, nurse checks whether the waiting room in front of treatment area is saturated or not. If it contains less than $K$ patients, triage continues. If the waiting room has attained maximum $K$ patients (saturated), stops service at triage node until the next waiting room can accommodate a new patient.

The following are the formulations of some transient performance measures related to the BBS tandem model considered here.

1. $\{W_1(t), t \geq 0\}$ is the waiting time of a patient who arrives at triage node at time t. Since the queue capacity is infinite, the probability distribution of waiting time is defined similar to BAS system, i.e.,

$$P(W_1(s) > t - s) = \rho_1 e^{-((1-\rho_1)\Lambda_{t,1}(s))/\rho_1},$$

where $\Lambda_{t,1}(s) = \Lambda_1(t) - \Lambda_1(s)$ and $\Lambda_1(.)$ is defined in (3).

2. $\{W_2(t), t \geq 0\}$ is the waiting time of a patient who arrives at the queue of treatment area at time $t$. Since the queue capacity is finite, i.e., $M(t)/M(t)/1/K$, the probability that the waiting time of the patient who arrives at time $s$, is larger than $t - s$, for $0 \leq s \leq t$ is,

$$P(W_2(s) > t - s) = \sum_{n=1}^{K-1} Q_n \sum_{1=0}^{n-1} \frac{(\frac{\Lambda_{t,2}(s)}{\rho_2})^i\ e^{\left(\frac{\Lambda_{t,2}(s)}{\rho_2}\right)}}{i!},$$

where $\Lambda_{t,2}(s) = \Lambda_2(t) - \Lambda_2(s)$ and $\Lambda_1(.)$ is defined in (7).

3. $L_1(t)$ represents the number of patients present at the triage node, including those blocked before service. These patients arrived during the interval $[0, t]$, i.e., $\{A_1(s),\ \ 0 \leq s \leq t\}$ and have not yet completed their service. They have not completed their service at the triage node, either because they are in the queue or are blocked due to capacity constraints at the treatment node. Therefore,

$$L_1(t) = \int_0^t (I_{\{W_1(s)>t-s\}} + I_{\{B(s)>t-s\}})\ dA_1(s),$$

where $I_{\{W_1(s)>t-s)\}}$ represents the number of patients who entered the queue of the triage node at time $s$ and are still waiting for service at time $t$, $0 \leq s \leq t$. Similarly, $I_{\{B(s)>t-s\}}$

denotes the number of patients who entered the blocking space in front of the triage node at time $s$ and are still waiting for triage at time $t$. When taking expectations,

$$E(L_1(t)) = \int_0^t \left(P\{W_1(s) > t-s\} + P\{B(s) > t-s\}\right)\ \lambda_1(s)\ ds. \tag{13}$$

This represents the average number of patients present at the triage node at time $t$, including those blocked before service.

4. $L_2(t)$ represents the number of patients present at treatment node at time $t$. These patients moved after triage to treatment node during the interval $[0,t]$ and have not yet completed their service. Mathematically, this corresponds to the arrivals $\{A_1(s),\ \ 0 \leq s \leq t\}$, i.e.,

$$L_2(t) = \int_0^t (I_{\{W_2(s)>t-s\}})\ dA_2(s),$$

where $I_{\{W_2(s)>t-s)\}}$ represents the number of patients who entered the queue of the treatment area at time $s$ and are still waiting for service at time $t$, $0 \leq s \leq t$. Then the average number of patients present in the queue of the treatment area at time $t$ is,

$$E(L_2(t)) = \int_0^t \left(P\{W_2(s) > t-s\}\right)\ p\ \mu_1(s)\ ds. \tag{14}$$

5. $Z_1(t)$ represents the time required to triage all patients who arrived at first node up to time t, including those blocked patients, i.e.,

$$Z_1(t) = \int_0^t \left(I_{\{W_1(s)>t-s\}} + I_{\{B(s)>t-s\}}\right)\ V_1(s)\ dA_1(s) + \int_0^t \frac{V_1(s)^2}{2}\ dA_1(s).$$

Then the average workload at the triage node, taking into account the patients in the blocking space and the queue at time t can be represented as,

$$E(Z_1(t)) = \int_0^t \left(P\{W_1(s) > t-s\} + P\{B(s) > t-s\}\right)\ \frac{\lambda_1(s)}{\mu_1(s)}\ ds + \int_0^t \frac{c_1^2+1}{2}\ \frac{\lambda_1(s)}{\mu_1(s)^2} ds. \tag{15}$$

6. $Z_2(t)$ denotes the time required to complete the consultation of patients who arrived up to time t after triage, i.e.,

$$Z_2(t) = \int_0^t (I_{\{W_2(s)>t-s\}})\ V_2(s)\ dA_2(s) + \int_0^t \frac{V_2(s)^2}{2}\ dA_2(s).$$

Then the average workload at the treatment node at time t can be represented as,

$$E(Z_2(t)) = \int_0^t P\{W_2(s) > t-s\}\ \frac{\lambda_2(s)}{\mu_2(s)}\ ds + \int_0^t \frac{c_2^2+1}{2}\ \frac{\lambda_2(s)}{\mu_2(s)^2} ds, \tag{16}$$

where $c_i^2$, $i = 1, 2$ is the coefficient of variation of service process in station i.

## 3. Numerical Study

We consider a two-station tandem network with non-stationary Markovian queues, in which first station has infinite queue capacity and second station has finite queue capacity. We compare BAS system and BBS system by computing the transient performance measures. First, we outline the prerequisites for conducting the numerical study.

1. A more realistic choice for the arrival rate function would be a sinusoidal or periodic function, which is useful for modelling daily or weekly fluctuations in patient arrivals. However, for simplicity, we choose the identity function as the external arrival rate.

   Let the time-dependent arrival rate of patients to the triage node (external arrival rate), $\lambda_1(t)$ be the identity function, $t$, $t \geq 0$.

2. Transition rate or arrival rate of patients from triage station to treatment station, $\lambda_2(t)$ is,

$$\lambda_2(t) = p\ \mu_1(t) = p\ \lambda_1(t)/\rho_1 \tag{17}$$

   where $p$ is the transition probability from station 1 to 2. Here we take $p = 0.75$.

3. The squared coefficient of variation of service time $(c_i^2),\ i = 1, 2$ appears in expressions of virtual workload. Since we are considering a Markovian queueing model, $c_i^2$ is assumed to be 1.

4. In the BAS system, blocking occurs only if a customer completes service at the triage node and attempts to move towards treatment node, but finds the queue is full. Therefore, the probability depends on the queue capacity and traffic intensity of the treatment area.

   For a BAS system, the blocking probability is defined by,

$$P(B_{BAS}) = \frac{(1-\rho_2)\rho_2^K}{1-\rho_2^{K+1}}, \tag{18}$$

   where $K$ is the queue capacity and $\rho_2$ is the traffic intensity of the treatment area. Gross and Haris [9] and Ziya [17] developed formulations for blocking probability.

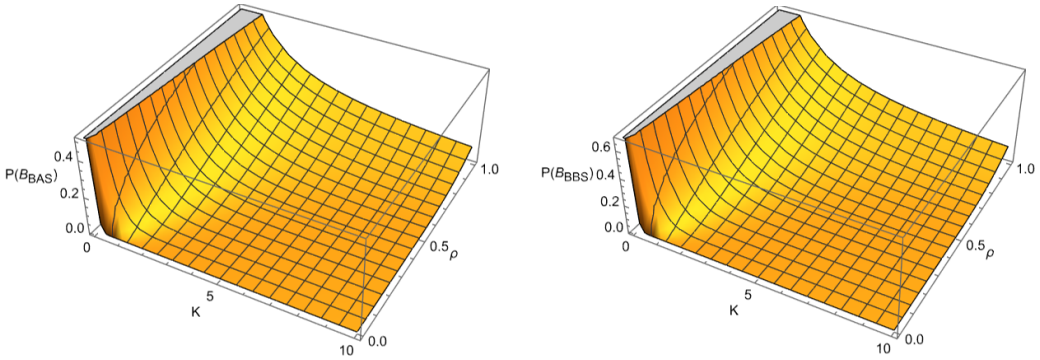
   Now we present an approximation for the blocking probability in the BBS system. In the BBS system, blocking occurs before the triage starts based on the availability of space in queue of treatment area. This means all patients might be blocked regardless of whether they would have moved for treatment or left the system after triage. As a result blocking probability is inflated.

   Since $P(B_{BAS})$ is applying only to transitioning customers, we can approximate $P(B_{BBS})$ by scaling $P(B_{BAS})$ with the proportion of transition, $p$. Here we assume that patient's decision to leave the system after triage or transition is independent of the congestion in the treatment area, so it only depends on the service at triage. $P(B_{BAS})$ is already calculated on transitioned patients, while formulating $P(B_{BBS})$, we need to undo this effect by dividing the proportion of transition, i.e.,

$$P(B_{BBS}) \approx \frac{P(B_{BAS})}{p} = \frac{(1-\rho_2)\rho_2^K}{(1-\rho_2^{K+1})\ p}. \tag{19}$$

   In this study, we have chosen a constant queue capacity and a constant traffic intensity using the rate-matching control principle. Consequently, the parameters in the above expressions are time-invariant, making them applicable to a non-stationary model.

   The Figure 3 illustrates the relationship between the queue capacity and traffic intensity and how their combined variation influences the blocking probabilities.




(a) BAS (b) BBS

Figure 3: *Blocking probability for queue capacity $K = 0$ to 10 and traffic intensity $\rho = 0$ to 1.*

5. In this study, we consider four cases, A, B, C, and D, by taking arbitrary values for traffic intensities and queue capacity of the second station, as shown in the Table 1. The blocking probabilities corresponding to each case are calculated and included in the table. The stations with traffic intensity close to 1 are considered bottleneck stations.

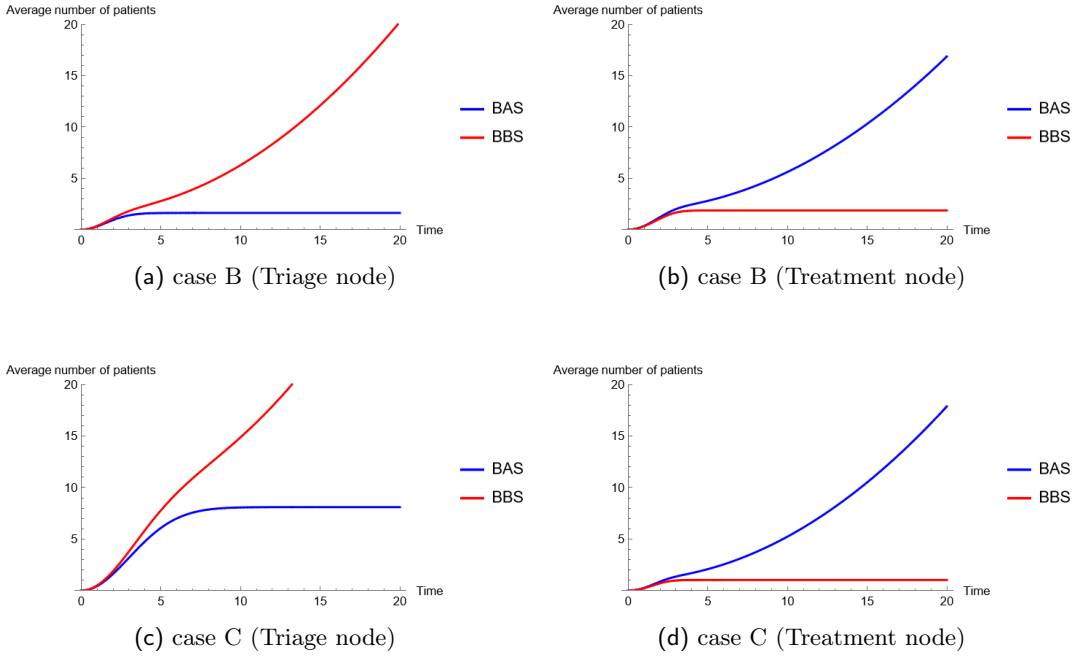
| cases | $\rho_1$ | $\rho_2$ | $K$ | $P(B_{BAS})$ | $P(B_{BBS})$ |
|---|---|---|---|---|---|
| A | 0.80 | 0.60 | 4 | 0.056 | 0.074 |
| B | 0.70 | 0.90 | 8 | 0.070 | 0.093 |
| C | 0.90 | 0.90 | 6 | 0.101 | 0.136 |
| D | 0.90 | 0.90 | 10 | 0.050 | 0.067 |

Table 1: *Four cases of traffic intensities, queue capacity and corresponding blocking probabilities.*

The probability of blocking is observed to be relatively high in cases where the traffic intensity at the second station increases significantly and the queue capacity decreases. Among the cases considered, cases A and D exhibit the lowest blocking probabilities. In contrast, Cases B and C show comparatively higher probabilities of blocking due to the combination of high traffic intensity and low queue capacity. Therefore, we have focused on presenting numerical illustrations for these two cases to obtain critical observations.

Figure 4 illustrates the number of patients at both nodes under the two mechanisms. Blocking after service occurs at the treatment node, while blocking before service occurs at the triage node. As a result, the number of patients increases at the corresponding nodes with time. When the traffic intensity at a node is high, it leads to a significant increase in the number of patients. A similar trend is observed in the average workload, as shown in Figure 5. Bottleneck nodes exhibit a relatively higher workload compared to others. In both cases, the treatment node experiences a higher workload under the BAS and BBS mechanisms, as it serves as a bottleneck.

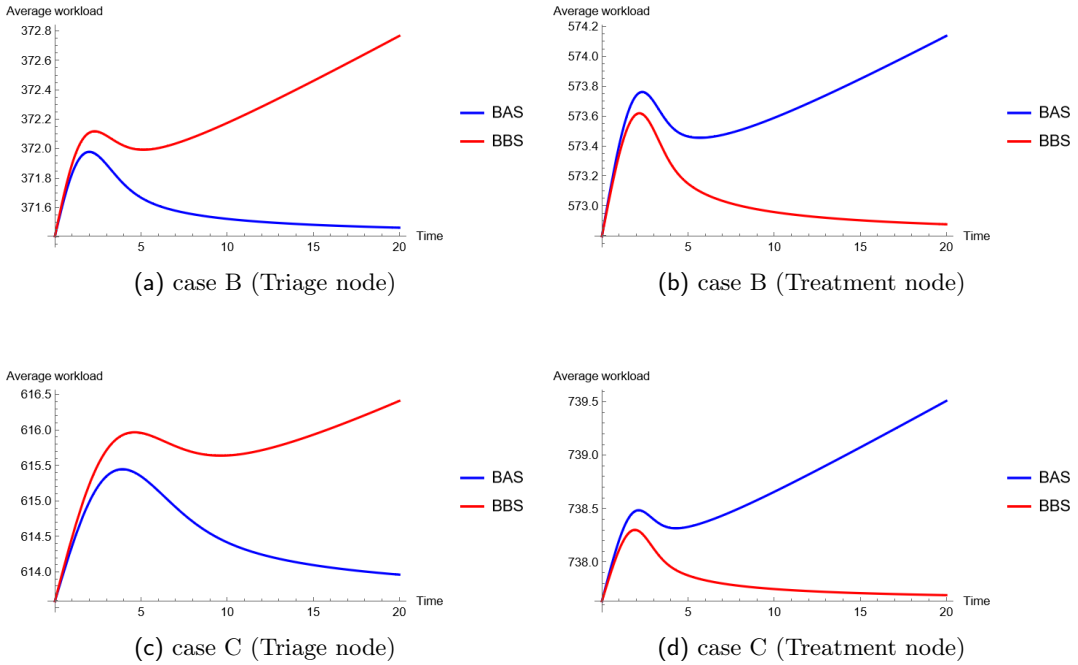
Figure 6 illustrates the average number of patients in the system over time. This analysis considers the four cases discussed in Table 1. In all cases, the total number of patients is consistently higher in the BBS system.




(a) case B (Triage node) (b) case B (Treatment node)

(c) case C (Triage node) (d) case C (Treatment node)

Figure 4: *Average number of patients under BAS and BBS mechanisms for cases B and C.*




(a) case B (Triage node) (b) case B (Treatment node)

(c) case C (Triage node) (d) case C (Treatment node)

Figure 5: *Average workload under BAS and BBS mechanisms for cases B and C.*

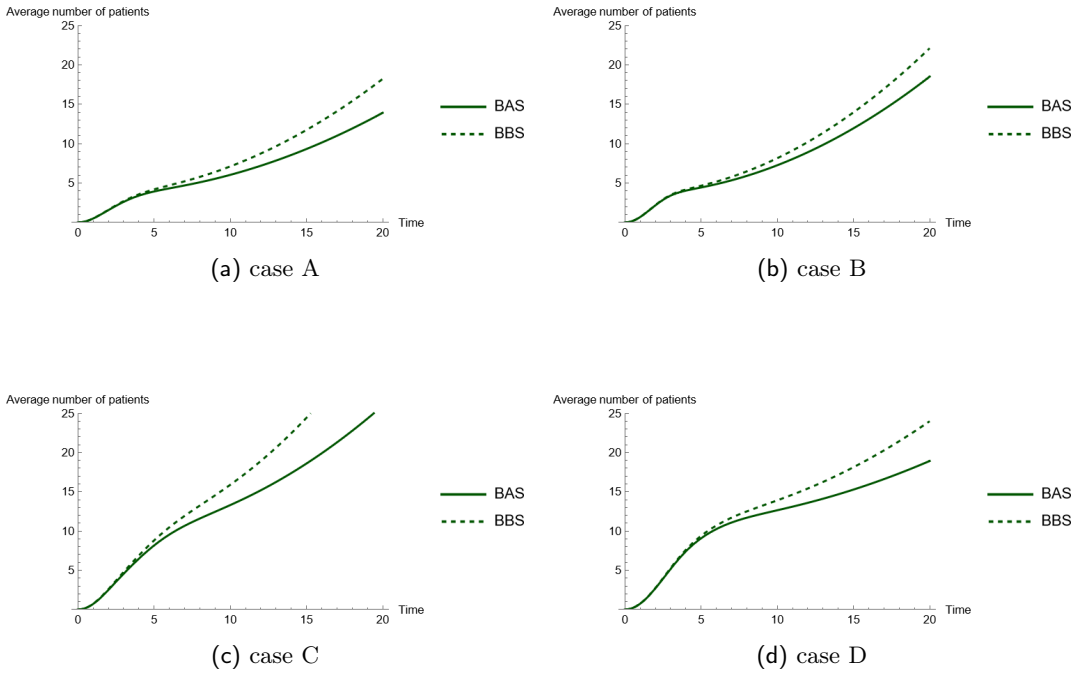



Figure 6: *Average number of patients in the system (including both stations) for the four cases considered in this study.*

## 4. Conclusion

Finite capacity queues are a realistic and common feature of many real-world service systems, such as hospitals, call centres, and manufacturing units, due to physical and budgetary constraints. However, in most of the theoretical studies of queues, infinite capacity is often assumed to simplify the analysis and results. In this study, we examined various blocking mechanisms that can be applied in queueing systems with capacity restrictions. Most commonly, BAS and BBS blocking mechanisms are used in such systems. Here, we have modelled a time-dependent hospital emergency department using a tandem network of two stations and derived transient performance measures based on these mechanisms. These measures enable an effective comparison of the BAS and BBS mechanisms through numerical study. The results highlight the impact of capacity restrictions on both mechanisms, demonstrating how they influence the time-varying number of patients and the virtual workload in the system. In our model, patients who do not need immediate emergency treatment can leave after triage. But in the BBS system, these patients also get blocked, causing unwanted congestion in the system. So, BAS is slightly better than BBS in the above scenario. Generally, BAS is preferred when intermediate congestion can be managed, whereas BBS is preferred when smooth flow of entities through the system is more important. This study examined a two-station tandem network of single-server queues with capacity restriction on the intermediate queue. This framework can be generalised to a k-station tandem network, and the single-server queues can be extended to multi-server systems. Deriving more rigorous mathematical expressions for performance measures under both blocking mechanisms would enable a more extensive study of these systems. These are some directions for future research.

## References


[1] Avi-Itzhak, B. (1965). A sequence of service stations with arbitrary input and regular service times. Management Science, 11(5), 565-571. doi: 10.1287/mnsc.11.5.565

[2] Avi-Itzhak, B. and Yadin, M. (1965). A sequence of two servers with no intermediate queue. Management Science, 11(5), 553-564. doi: 10.1287/mnsc.11.5.553

[3] Avi-Itzhak, B. and Halfin, S. (1993). Servers in tandem with communication and manufacturing blocking. Journal of Applied Probability, 30(2), 429-437. doi: 10.2307/3214851

[4] Avi-Itzhak, B. and Levy, H. (1995). A sequence of servers with arbitrary input and regular service times revisited: in memory of Micha Yadin, Management Science, 41(6), 1039-1047. doi: 10.1287/mnsc.41.6.1039

[5] Balsamo, S., de Nitto Personé, V., and Onvural, R. (2001). Analysis of queueing networks with blocking. Springer Science & Business Media. doi: 10.1007/978-1-4757-3345-7

[6] de Bruin, A. M., van Rossum, A. C. and Visser, M. C. et al. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. Health Care Management Science, 10(2), 125-137. doi: 10.1007/s10729-007-9009-8

[7] Fralix, B.H. and Riaño, G. (2010). A new look at transient versions of little's law, and m/g/1 preemptive last-come-first-served queues. Journal of Applied Probability, 47(2), 459-473. doi: 10.1239/jap/1276784903

[8] Frein, Y. and Dallery, Y. (1989). Analysis of cyclic queueing networks with finite buffers and blocking before service. Performance Evaluation, 10(3), 197–210. doi: 10.1016/0166-5316(89)90010-2

[9] Gross, D. and Harris, C. (1998) Fundamentals of Queuing Theory. John Wiley, Chichester. doi: 10.1002/9781119453765

[10] Meerkov, S. and Yan, C. B. (2016). Production lead time in serial lines: Evaluation, analysis, and control. IEEE Transactions on Automation Science and Engineering, 13(2), 663–675. doi: 10.1109/TASE.2014.2365108

[11] Prabhu, N. (1967). Transient behaviour of a tandem queue. Management Science, 13(9), 631–639. doi: 10.1287/mnsc.13.9.631

[12] Ramesh, A. and Manoharan, M. (2024). Transient Behaviour of Time-Varying Tandem Queueing Networks. OPSEARCH. doi: 10.1007/s12597-024-00790-0

[13] Seo, D. W., Lee, H. C. and Ko, S. S. (2008). Stationary waiting times in m-node tandem queues with communication blocking. Management Science and Financial Engineering, 14(1), 23-34.

[14] Seo, D. W. and Lee, H. C. (2011). Stationary waiting times in m-node tandem queues with production blocking. IEEE Transactions on Automatic Control, 56(4), 958-961. doi: 10.1109/TAC.2011.2105290

[15] Suri, R and Diehl, G. (1984). A new 'building block' for performance evaluation of queueing networks with finite buffers. ACM SIGMETRICS Performance Evaluation Review, 12, 134–142. doi: 10.1145/800264.809321

[16] Whitt, W. (2015). Stabilizing performance in a single-server queue with time-varying arrival rate. Queueing Systems, 81(4), 341–378. doi: 10.1007/s11134-015-9462-x

[17] Ziya, S. (2008). On the Relationships Among Traffic Load, Capacity, and Throughput for the M/M/1/m, M/G/1/m-PS, and M/G/c/c Queues. IEEE Transactions on Automatic Control, 53(11), 2696-2701. doi: 10.1109/TAC.2008.2007173.

[18] Zychlinski, N., Mandelbaum, A. and Momcilovic, P. (2018a). Time-varying tandem queues with blocking: modeling, analysis, and operational insights via fluid models with reflection. Queueing Systems, 89(1), 15–47. doi: 10.1007/s11134-018-9578-x

[19] Zychlinski, N., Momcilovic, P. and Mandelbaum, A. (2018b). Time-varying many-server finite-queues in tandem: Comparing blocking mechanisms via fluid models. Operations Research Letters 46(5), 492-499. doi: 10.1016/j.orl.2018.07.002