Fixed points for multi-class queues

James B. Martin and Balaji Prabhakar
University of Oxford and Stanford University

Abstract

Burke’s theorem can be seen as a fixed-point result for an exponential single-server queue; when the arrival process is Poisson, the departure process has the same distribution as the arrival process. We consider extensions of this result to multi-type queues, in which different types of customer have different levels of priority. We work with a model of a queueing server which includes discrete-time and continuous-time M/M/1 queues as well as queues with exponential or geometric service batches occurring in discrete time or at points of a Poisson process. The fixed-point results are proved using interchangeability properties for queues in tandem, which have previously been established for one-type M/M/1 systems. Some of the fixed-point results have previously been derived as a consequence of the construction of stationary distributions for multi-type interacting particle systems, and we explain the links between the two frameworks. The fixed points have interesting “clustering” properties for lower-priority customers. An extreme case is an example of a Brownian queue, in which lower-priority work only occurs at a set of times of measure 0 (and corresponds to a local time process for the queue-length process of higher priority work).

1 Introduction

One of the most famous results in queueing theory is Burke’s theorem [7]. Consider a queue in which available services occur as a Poisson process of rate $\mu$ (a so-called $\text{M}/\text{M}/1$ queueing server). If the arrival process is a Poisson process of rate $\lambda < \mu$ (independent of the service process), then the departure process is also a Poisson process of rate $\lambda$. We may say that the arrival process is a fixed point for the server.

In this paper we consider the question of fixed points for queues with two or more classes of customer (with different levels of priority). When a service occurs in such a queue, it is used by a customer whose priority is highest out of those currently present in the queue. We will see that a two-type fixed point can be constructed using the output processes (consisting of departures and unused services) from a one-type queue. Then in a recursive way, a fixed point with $m \geq 3$ classes of customer can be constructed using the output of a queue whose arrival process is itself a fixed point with $m - 1$ classes.

Except in the familiar one-type case, the fixed points are not Markovian. In particular, one observes clustering of the lower-priority customers.

In the paper we work with a queueing model which is somewhat more general than the $\text{M}/\text{M}/1$ queue described above. Our basic model is of a discrete-time queue with batch arrivals...
and services. Let $S_n$ be the amount of service offered at time $n$. We obtain fixed-point results for the case where $S_n$ are i.i.d. and each $S_n$ has so-called “Bernoulli-geometric” distribution, i.e. is equal to the product of a geometric random variable and an independent Bernoulli random variable. By taking appropriate limits where necessary, this model covers a variety of previously considered queueing servers, for example discrete-time $/M/1$ queues [17], discrete-time queues with geometric or exponential service batches [6, 9, 27], continuous-time $/M/1$ queues as described above, continuous-time queues with geometric or exponential service batches occurring at times of a Poisson process, and Brownian queues [15, 29]. Versions of Burke’s theorem and related reversibility results were proved for this Bernoulli-geometric model in [21].

Some such fixed-point processes were already constructed in certain cases ($M/M/1$ queues in continuous or discrete time) in [12] in the context of stationary distributions for certain multi-type interacting particle systems. In this paper we give a more direct proof of the fixed-point property, which relies on properties of interchangeability for queueing servers. Weber [30] showed that for a tandem queueing system consisting of two independent $/M/1$ servers with service rates $\mu_1$ and $\mu_2$, and an arbitrary arrival process, the distribution of the departure process is unchanged if $\mu_1$ and $\mu_2$ are exchanged.

This interchangeability result was subsequently proved in a number of different ways, for example in [3], [19], and [29]. The coupling proof given by Tsoucas and Walrand in [29] is important for our purposes, since we can use their approach to extend the interchangeability result to multi-type queues.

Before developing the general batch queueing model, we begin in Section 2 by giving a guide to the main results and methods of proof in the particular case of the continuous-time $/M/1$ queue. Since this model is already rather familiar, we give an informal account without introducing too much notation. (Everything is developed rigorously in later sections). In addition, certain aspects are simpler in the $/M/1$ case; for example, the service process has only one parameter, so all such service processes are interchangeable, and given any vector of arrival rates $\lambda_1, \ldots, \lambda_m$ (corresponding to customers of classes 1, \ldots, $m$ respectively), there is a unique fixed-point arrival process which is common to all $/M(\mu)/1$ queues for $\mu > \lambda_1 + \cdots + \lambda_m$.

Our general model is introduced in Section 3, which describes the set-up of a discrete-time batch queue. Multi-class systems are introduced in Section 4. The Bernoulli-geometric distribution, and corresponding queueing servers, are described in Section 5.

Interchangeability results are given in Section 6. These extend the results for one-type $/M/1$ queues described above, to cover multi-type systems and to the more general queueing server model. In Section 7 we give the construction of multi-type fixed points, and prove the fixed-point property using the interchangeability results. The main result is given in Theorem 7.1 (the corresponding results in the $/M/1$ case are Theorem 2.1 and Theorem 2.2). The proof of the interchangeability result itself is given in Section 8.

In Section 9 we give examples of the application of the results to several of the particular queueing systems described above. The final example is that of the Brownian queue. Here the lower-priority work in the fixed-point process corresponds to the local-time process of a reflecting Brownian motion; this process is non-decreasing and continuous but is constant.
except on a set of measure 0. This is an extreme case of the "clustering of lower-priority customers" referred to above.

The connections with interacting particle systems are discussed in Section 10. The fixed points for \( M/M/1 \) servers in discrete time and in continuous time correspond to stationary distributions for multi-type versions of the TASEP \([13]\) and of Hammersley’s process \([14]\), respectively. Time in the queueing systems corresponds to space in the particle systems; with this identification, questions of fixed points for queues and stationary distributions for particle systems are closely analogous.

Finally in Section 11 we mention a limit as the number of classes goes to infinity, with the density of each class going to 0. In this limit, the class-label of each customer becomes, for example, a real number in \([0,1]\). This gives another illustration of the clustering phenomenon; although a priori, any given label has probability 0 of occurring, nonetheless for any realisation of the process, each label that does occur will occur infinitely often with probability 1.

2 Continuous-time \( ./M/1 \) case

2.1 \( ./M/1 \) queueing servers and Burke’s theorem

Let \( A \) and \( S \) be independent Poisson processes, of rate \( \lambda \) and \( \mu \) respectively, with \( \lambda < \mu \). We can use these processes to define an \( M/M/1 \) queue with arrival rate \( \lambda \) and service rate \( \mu \).

Arrivals occur at points of the process \( A \); at these points the queue-length increases by 1. At a point of the process \( S \), a departure occurs and the queue-length decreases by 1, unless the queue-length is already 0, in which case it stays the same and we say that an unused service has occurred.

We write \( D \) for the process of departures and \( U \) for the process of unused services.

Perhaps the most famous result in queueing theory is Burke’s theorem, which states that the departure process \( D \) is itself a Poisson process of rate \( \lambda \). We may regard the service process \( S \) as an operator (a “\( ./M/1 \) queue”, or a “\( ./M(\mu)/1 \) queue” if we want to emphasise the service rate) which maps the distribution of an arrival process to the distribution of a service process. In this sense, Burke’s theorem \([7]\) says that a Poisson process of rate \( \lambda \) is a fixed point for a \( ./M(\mu)/1 \) queue whenever \( \mu > \lambda \).

In \([4]\), Anantharam showed that in fact Poisson processes are the only such ergodic fixed points. Mountford and Prabhakar \([24]\) showed further that these fixed points are attractive: if one starts with any ergodic process of rate \( \lambda \) and applies the “\( ./M(\mu)/1 \) queueing operator” repeatedly, the sequence of distributions obtained converges weakly to a Poisson process of rate \( \lambda \).

2.2 Multi-class queues

In this paper we consider fixed-points in the context of multi-class queues.

Consider again a queue whose service process is a Poisson process of rate \( \mu \). The queue may now contain several types of customer, say types \( 1, 2, \ldots, m \). Each arrival to the system is of one of these types. Customers of type 1 (“first-class customers”) have the highest priority, followed by those of type 2 (“second-class customers”) and so on. When a service occurs in
the queue, if there are any customers present, then a customer with the highest priority out
of those present departs from the system. Hence each departure from the system also has
a type. Again one may have “unused services”, when the queue is completely empty at the
time of an event in the service process.

An \(m\)-type queue may be seen as a coupling of \(m\) one-type queues, which share the same
service process. Given \(r\) with \(1 \leq r \leq m\), consider only customers of classes 1 up to \(r\), ignoring
the differences between these customers. Since any such customer has priority any customer
whose class is higher than \(r\), the process obtained behaves exactly as a one-type queue.

2.3 Multi-class fixed points

The distribution of a multi-type arrival process is said to be a fixed point if the process of
departures from the system has the same distribution as the process of arrivals. (We assume
that arrivals and services are independent).

We will consider arrival processes which are stationary and ergodic, in which case for each
\(m\) there is a deterministic long-run intensity of arrivals of customers of type \(m\).

A coupling argument analogous to that used by Mountford and Prabakhar in [24] can be
used to show that for any \(\lambda_1, \ldots, \lambda_m\) with \(\lambda_1 + \cdots + \lambda_m < \mu\), there is a unique stationary
and ergodic \(m\)-type arrival process with intensity \(\lambda_r\) of customers of type \(r\) which is a fixed
point for the \(\cdot/M/\mu\) queue.

As observed above, the system comprising only customers of types 1 up to \(r\) can be seen
as a single one-type queue (for each \(1 \leq r \leq m\)). Since the only one-type fixed points are
Poisson processes, this shows that in any multi-type fixed point, the combined process of
customers of types 1, \ldots, \(r\) must be a Poisson process of rate \(\lambda_1 + \cdots + \lambda_r\).

2.4 Construction of a 2-type fixed point

First we describe how to construct such a two-type fixed point, with intensities \(\lambda_1\) and \(\lambda_2\).

Consider the output process \((D, U)\) of an \(M/M/1\) queue as described above, with arrival
rate \(\lambda_1\) and service rate \(\lambda_1 + \lambda_2\). We regard \((D, U)\) as a two-type process, in which first-class
customers occur at the points of \(D\) and second-class customers occur at the points of \(U\).
Write \(F_2 = F_{2, \lambda_1, \lambda_2}\) for the distribution of this process.

**Theorem 2.1** The distribution \(F_{2, \lambda_1, \lambda_2}\) is a fixed point for a \(\cdot/M/1\) queue with service rate
\(\mu\), for any \(\mu > \lambda_1 + \lambda_2\).

Note that indeed the sub-process of first-class customers is a Poisson process of rate \(\lambda_1\)
(by Burke’s theorem) and that the combined process of first-class and second-class customers
is a Poisson process of rate \(\lambda_1 + \lambda_2\) (since \(D + U = S\)).

However, note also that the process \(U\) of second-class customers is not a Poisson process.
Rather, the second-class customers tend to cluster. For example, fix \(\lambda_1\) and suppose that \(\lambda_2\)
 is very small. The process \((D, U)\) is obtained as the output of a queue whose arrival rate is \(\lambda_1\)
and whose service rate is \(\lambda_1 + \lambda_2\), so that it is operating very close to capacity. The long-run
rate of unused services is only \(\lambda_2\). However, if we observe an unused service, we know that
the queue is currently empty; hence just after an unused service, the instantaneous rate at which another unused service occurs is $\lambda_1 + \lambda_2$ which may be much larger.

### 2.5 Recursive construction of multi-type fixed points

Now we show how to construct fixed points with a larger number of classes recursively. Fix $\lambda_1, \ldots, \lambda_m$. Suppose we have already constructed a distribution of an $(m-1)$-type process $F_{m-1}$, with intensity $\lambda_i$ of $i$th-class customers for $1 \leq i \leq m-1$, which is a fixed point for a $./M/1$ queueing server (whenever $\mu > \lambda_1 + \cdots + \lambda_{m-1}$). Now consider a $./M(\lambda_1 + \cdots + \lambda_m)/1$ queue whose arrival process has distribution $F_{m-1}$. Write $(D_1, D_2, \ldots, D_{m-1}, U)$ for the output process of this queue, comprising departures of types $1, 2, \ldots, m-1$ and unused services. Now identify points of $U$ as $m$th-type customers, and write $F_m = F_{m, \lambda_1, \ldots, \lambda_m}$ for the distribution of the $m$-type process obtained.

**Theorem 2.2** The distribution $F_{m, \lambda_1, \ldots, \lambda_m}$ is a fixed point for a $./M/1$ queue with service rate $\mu$, whenever $\mu > \lambda_1 + \cdots + \lambda_m$.

The construction of $F_2$ and $F_3$ is illustrated in Figure 2.1.

Note that if we take the $a$ process with distribution $F_m$ and ignore the $m$th-class customers, we obtain a process with distribution $F_{m-1}$. (This had to be the case, because of the uniqueness of the fixed-points with given intensities, and because the ignoring the $m$th class of an $m$-class system simply gives an $m-1$-class system).

Theorem 2.2 was already proved in [14]; it emerged as a corollary of a multi-type tandem queue construction which was used to give the stationary distribution of a multi-type version of an interacting particle system called “Hammersley’s process”. In this paper we give a much more direct proof, using properties of interchangeability of queueing servers.
2.6 Tandems and interchangeability

In 1979 Richard Weber [30] proved an interchangeability property for \(/M/1\) queueing servers. Consider two independent \(/M/1\) queueing servers in tandem, with service rates \(\mu_1\) and \(\mu_2\). The first queue has some arrival process \(A\), with an arbitrary distribution (for example, it could even be deterministic), which is independent of the service processes. By “tandem” we mean that a customer leaving the first queue immediately joins the second queue; the departure process from the first server is the arrival process of the second.

The result of [30] is that the law of the departure process from the system (that is, the departure process from the second queue), is unchanged if \(\mu_1\) and \(\mu_2\) are exchanged. Given a queue with arrival process \(A\) and service process \(S\), we write \(D(A, S)\) for the departure process from the queue. So in a system with two queues in series, with arrival process \(A\) and service process \(S_1\) and \(S_2\) at the first and second queues respectively, the process of departures from the system is \(D(D(A, S_1), S_2)\). Then the result of [30] can be written as follows:

**Theorem 2.3** Let \(S_1\) and \(S_2\) be independent Poisson processes of rates \(\mu_1\) and \(\mu_2\) respectively. Then for any arrival process \(A\),

\[
D(D(A, S_1), S_2) \overset{d}{=} D(D(A, S_2), S_1).
\]

Thus, for any arrival process, the order of the queues does not affect the law of the output of the system. By induction, the result extends easily to tandems containing more than two queues.

Alternative proofs of this interchangeability result were subsequently given by Anantharam [3], by Lehtonen [19], and by Tsoucas and Walrand [29]. The proof by Tsoucas and Walrand in [29] is particularly important for our purposes since it allows an extension of the result which will also apply to multi-class systems. Their methods can be used to give the following result.

**Theorem 2.4** Let \(S_1\) and \(S_2\) be independent Poisson processes of rates \(\mu_1\) and \(\mu_2\). There is a coupling of \(S_1\) and \(S_2\) with two further processes \(\tilde{S}_1\) and \(\tilde{S}_2\) such that:

(i) \((S_1, S_2) \overset{d}{=} (\tilde{S}_2, \tilde{S}_1)\); and

(ii) If \(A\) is any arrival process, then \(D(D(A, S_1), S_2) = D(D(A, \tilde{S}_1), \tilde{S}_2)\).

The key point is that we can couple a system with service rates \(\mu_1, \mu_2\) and another system with service rates \(\mu_2, \mu_1\) in such a way that for any arrival process, the departure processes from the two systems are the same. (Weber’s result already implies the existence of such a coupling for any fixed arrival process, but not necessarily that the same coupling works for all arrival processes simultaneously).

Above we explained how a queue with \(m\) types of customer can be regarded as a coupling of \(m\) one-type systems, with the same service process but with different arrival processes. (In the \(r\)th system, we only look at customers whose class in the original system is between 1 and \(r\), and neglect all the customers whose class is greater than \(r\)). Since Theorem 2.4 gives a coupling which yields interchangeability simultaneously for all arrival processes, it immediately implies interchangeability for any multi-type arrival process.
Corollary 2.5  Let $S_1$ and $S_2$ be independent Poisson processes of rates $\mu_1$ and $\mu_2$. Then for any multi-type arrival process $A$,

$$D(D(A, S_1), S_2) \overset{d}{=} D(D(A, S_2), S_1).$$

The coupling described in Theorem 2.4 is quite easy to construct. (The construction in [29] is essentially the same although not expressed in quite the same way). Assume without loss of generality that $\mu_1 < \mu_2$. Now consider a queue whose arrival process is $S_1$ and whose service process is $S_2$. Let $D = D(S_1, S_2)$ be the departure process of this queue, and let $U$ be the process of unused services (so that $S_2 = D + U$).

Now let $\tilde{S}_2 = D$, and let $\tilde{S}_1 = S_1 + U$. Loosely speaking, we have transferred the unused service events from one process to the other.

The fact that $(\tilde{S}_2, \tilde{S}_1)$ has the same distribution as $(S_1, S_2)$ is an extension of Burke’s Theorem, and can be proved using simple reversibility arguments (see for example Theorem 3 of [25]). To prove the interchangeability statement in part (ii) of Theorem 2.4 one can verify pathwise that, whatever the arrival process is, transferring “unused services” in this way cannot affect the overall departure process.

A full proof of this result in the more general setting is given below (Theorem 6.1). The construction in the general setting is slightly more complicated; again we remove the process of “unused services” from $S_2$ to obtain $\tilde{S}_2$, but now we do not add precisely the same process to $S_1$ to obtain $\tilde{S}_1$, but instead something more like an independent copy of the process. However, the same ideas of reversibility are again the key part of the proof.

2.7 Proving multi-class fixed points using interchangeability

Still in the context of the $\mathcal{M}/1$ queue, we finally indicate how the interchangeability result of Corollary 2.5 can be used to prove that the processes $F_m$ defined above are indeed fixed points. The arguments in this section are informal but can be made fully rigorous, as they are below in the proof of Theorem 7.1. We will show the argument for $m = 2$; the case $m > 2$ follows by induction in a natural way, and can be found in the general setting below.

Recall that $F_2 = F_{2, \lambda_1, \lambda_2}$ is the distribution of a two-type process which can be obtained by taking the departure and unused service processes $(D, U)$ from an $M/M/1$ queue whose arrival rate is $\lambda_1$ and whose service rate is $\lambda_2$. So the first-class customers occur as a Poisson process of rate $\lambda_1$ and whose service rate is $\lambda_2$. The first-class and second-class customers together occur as a Poisson process of rate $\lambda_1 + \lambda_2$. Let $\mu > \lambda_1 + \lambda_2$, and consider feeding the arrival process $F_2$ into a system of two queues in series, with service rates $\lambda_1 + \lambda_2$ and $\mu$. The two possible orderings of the two queues can be visualized as follows:

$$F_2 \rightarrow S_{\lambda_1 + \lambda_2} \rightarrow G \rightarrow S_{\mu} \rightarrow H$$

$$F_2 \rightarrow S_{\mu} \rightarrow J \rightarrow S_{\lambda_1 + \lambda_2} \rightarrow K$$

For example, $G$ is the distribution of the departure process of the first queue with service rate $\lambda_1 + \lambda_2$, and is then itself used as the arrival process for the second queue with service rate $\mu$. 
Note that the total arrival rate in the process $F_2$ is $\lambda_1 + \lambda_2$. But the first queue in the first system only has service rate $\lambda_1 + \lambda_2$, so it is saturated; every service is used by some customer. Since the arrival process of first-class customers is a Poisson process of rate $\lambda_1$, the first-class departures in $G$ occur as the departures from an $M(\lambda_1)/M(\lambda_1 + \lambda_2)/1$ queue; the second-class customers fill up all the other services which are unused by first-class departures. Hence this two-type process has the distribution of the departure and unused service process $(D,U)$ from an $M(\lambda_1)/M(\lambda_1 + \lambda_2)/1$ queue, and we have $G = F_2$.

The same argument shows also that $K = F_2$, since the process $J$ has total intensity $\lambda_1 + \lambda_2$, and (by Burke’s theorem) its first-class customers occur as a Poisson process of rate $\lambda_1$.

Finally, the interchangeability result in 2.5 tells us that $H = K$, so that also $H = F_2$. Since we have $G = H = F_2$, we can concentrate only on the second queue in the first line, to obtain

$$F_2 \rightarrow S_{\mu} \rightarrow F_2$$

so that indeed $F_2$ is a fixed point for the $/M(\mu)/1$ queue as desired.

3 Batch queueing model

We now move to the more general queueing model. We begin by defining a model of a batch queue in discrete-time. By taking particular values, or appropriate limits, this model will cover various interesting cases, including discrete-time and continuous-time $M/M/1$ queues.

The batch queue is driven by an arrival process $(A(n), n \in \mathbb{Z})$ and a service process $(S(n), n \in \mathbb{Z})$.

At time-slot $n \in \mathbb{Z}$, $A(n)$ customers arrive at the queue. Then service is available for $S(n)$ customers; if the queue-length is at least $S(n)$, then $S(n)$ customers are served, while if the queue length is less than $S(n)$ then all the customers are served.

We define various processes as functions of the basic data $A$ and $S$ of the queue.

Let $X(n)$ be the queue length after the service $S(n-1)$, before the arrival $A(n)$. Formally, we define the process $X$ by

$$X(n) = \sup_{m \leq n} \sum_{r=m}^{n-1} (A(r) - S(r)) \quad (3.1)$$

(where a sum from $n$ to $n-1$ is understood to be 0). In accordance with the description of the queue above, we have the basic recurrences

$$X(n+1) = \lceil X(n) + A(n) - S(n) \rceil_+ \quad (3.2)$$

(where $\lceil x \rceil_+$ denotes $\max\{x, 0\}$).

Let $D(n)$ be the number departing from the queue at the time of the service $S(n)$. So

$$D(n) = \min(X(n) + A(n), S(n)) = X(n) + A(n) - X(n+1).$$

Finally let $U(n) = S(n) - D(n)$ be the unused service at the time of the service $S(n)$. 

See Figure 3.1 for a representation of the evolution of the queue along with its inputs and outputs.

Since $X, D, U$ are all functions of the data $A$ and $S$, we sometimes write $D = D(A, S)$, and so on.

Note that we allow the possibility that $X(n) = \infty$. Indeed, we don’t impose any stability condition on the queue; so, for example, if the average rate of arrivals exceeds the average rate of service, then the queue will become saturated. In fact, the following simple observation will be useful later:

**Lemma 3.1** Suppose the arrival and service processes are independent, and $A_n$ are i.i.d. with mean $\lambda$ and $S_n$ are i.i.d. with mean $\mu \leq \lambda$. Then with probability 1,

- $X_n = \infty$ for all $n$,
- $D_n = S_n$ for all $n$,
- $U_n = 0$ for all $n$.

The lemma follows immediately from the definition (3.1), since the random walk whose increment at step $r$ is $A(-r) - S(-r)$ is either recurrent (if $\mu = \lambda$) or escapes to $+\infty$ with probability 1 (if $\mu < \lambda$); in either case it attains arbitrarily high values.

If the $A(n)$ and $S(n)$ take integer values, then it is natural to talk in terms of “number of customers” arriving, or departing, or in the queue, and so on. However we will also consider cases where the values are more general, in which case one could talk of “amount of work” rather than “number of customers”.

## 4 Multi-class queues

We will now define a *multi-class* batch queue. The system can now contain different types of customers (or work) with different priorities. When service occurs at the queue, it is first available to first-class customers. If there is more service available than there are first-class customers present in the queue, the remaining service (unused by the first-class customers) is then offered to customers of lower class, starting with second-class customers, then third-class and so on.

For example, suppose that at the start of time-slot $n$, there are 7 customers in the queue, of whom 3 are first-class, 1 is third-class and 3 are fourth-class. Suppose that 5 units of
service are available at time $n$. Then the departures at time $n$ will be 3 first-class customers, 1 third-class customer and 1 fourth-class customer, leaving 2 fourth-class customers remaining in the queue.

Let $m$ be the total number of classes. We will have a collection of arrival processes $A^{=1}, A^{=2}, \ldots, A^{=m}$, where $A^{=r}(n)$ is the number of $r$-th-class customers arriving at time $n$.

We will also denote $A^{\leq r} = \sum_{i=1}^{r} A^{=i}$, for $r = 0, 1, \ldots, m$.

Let $S$ be the service process of the queue.

Similarly we will write $D^{=r}(n)$ for the number of $r$-th-class customers departing at time $n$, and $X^{=r}(n)$ for the number of $r$-th-class customers present in the queue at the beginning of time-slot $n$. Write also $D^{\leq r} = \sum_{i=1}^{r} D^{=i}$ and $X^{\leq r} = \sum_{i=1}^{r} X^{=i}$.

There are two natural ways to construct the multi-class queue (which are equivalent). One way would be to look at the queueing process of $r$-th-class customers for each $r$. This is a queue with arrival process $A^{=r}$ and service process $S - D^{\leq r}$; the services available to $r$-th-class customers are those that have not been used by any higher-priority customer.

Alternatively, we will consider, for each $r$, the queueing process of customers of classes $1, 2, \ldots, r$ combined. This is a queue with arrival process $A^{\leq r}$ and service process $S$. So in particular we define

$$D^{\leq r} = D(A^{\leq r}, S),$$
$$X^{\leq r} = X(A^{\leq r}, S).$$

This second description turns out to be more useful since it describes the $m$-class queue as a coupling of $m$ single-class queues, each with the same service process. Once we come to consider interchangeability of queues, the fact that we are working with a common service process becomes crucial.

### 5 Bernoulli-geometric distribution

We define a Bernoulli-geometric distribution, with parameters $p$ and $\alpha$. A random variable with this distribution has the distribution of the product of two independent random variables, one with $\text{Ber}(p)$ distribution and the other with $\text{Geom}(\alpha)$ distribution. That is, $A \sim \text{Ber}(p)\text{Geom}(\alpha)$ if

$$
\mathbb{P}(A = k) = \begin{cases} 
1 - p, & k = 0 \\
p\alpha(1 - \alpha)^{k-1}, & k \geq 1.
\end{cases}
$$

We have $\mathbb{E}A = p/\alpha$.

We will consider a queue where $A(n)$ is an i.i.d. sequence with $A(n) \sim \text{Ber}(p)\text{Geom}(\alpha)$, and $S(n)$ is an i.i.d. sequence (independent of $A(n)$) with $S(n) \sim \text{Ber}(q)\text{Geom}(\beta)$. (We will say that $A$ and $S$ are Bernoulli-geometric processes). Queues of this type are investigated in [21], in particular regarding their reversibility properties.

For stability we will assume that $\mathbb{E}A(n) < \mathbb{E}S(n)$, i.e. that $p/\alpha < q/\beta$, and further we assume that

$$c(p, \alpha) = c(q, \beta),$$

(5.1)
where for \( p, \alpha \in [0, 1] \) we define

\[
    c(p, \alpha) = \frac{p}{1 - p} \frac{\alpha}{1 - \alpha}
\]

(5.2)

When the arrival and service processes share a value of the parameter \( c \) in this way, a version of Burke’s theorem holds. We use an asterisk to denote the time reversal of a process, so that \( A^*(n) = A(-n) \). The following result is included in Theorem 4.1 of [21]:

**Theorem 5.1** Under the assumptions above, the departure process \((D(n), n \in \mathbb{Z})\) has the same law as the arrival process \((A(n), n \in \mathbb{Z})\). Moreover, the queue is reversible in the sense that

\[
    (A(n), D(n), n \in \mathbb{Z}) \overset{d}{=} (D^*(n), A^*(n), n \in \mathbb{Z}).
\]

Note that for given \( q, \beta \) and \( \lambda \), there is precisely one choice of the parameters \( p \) and \( \alpha \) such that \( p/\alpha = \lambda \) and \( c(p, \alpha) = c(q, \beta) \). So for a Bernoulli-Geometric service process with given parameters \( q \) and \( \beta \), Theorem 5.1 gives a one-parameter family of fixed-point arrival processes, indexed by the arrival intensity \( p/\alpha \).

In the next section we see further that two service processes are interchangeable precisely when they fall within the same one-parameter family (that is, when they share a value of the parameter \( c \)).

### 6 Interchangeability

We consider two queues in tandem. Suppose we are given an arrival process \( A_1 \) and service processes \( S_1, S_2 \). From this we construct a system of two queues. The first queue has arrival process \( A_1 \) and service process \( S_1 \). The second queue has service process \( S_2 \) and arrival process \( A_2 \) defined by \( A_2 = D(A_1, S_1) \). (The departure process from the first queue becomes the arrival process from the second queue).

Note that a departure from queue 1 at time \( n \) becomes an arrival at queue 2 at the same time-slot \( n \) (and may indeed depart from queue 2 at time \( n \) also).

The following result says that two service processes, which fall within the same one-parameter family mentioned above, are interchangeable:

**Theorem 6.1** Suppose that the processes \( S_1 \) and \( S_2 \) are independent of each other and each are i.i.d., with \( S_1(n) \sim \text{Ber}(p) \text{Geom}(\alpha) \) and \( S_2(n) \sim \text{Ber}(q) \text{Geom}(\beta) \). Suppose further that condition [5.1] holds.

Then there is a coupling of \((S_1, S_2)\) with another pair of processes \((\tilde{S}_1, \tilde{S}_2)\) such that:

(i) \( (S_1, S_2) \overset{d}{=} (\tilde{S}_2, \tilde{S}_1) \);

(ii) If \( A \) is any arrival process, then \( D_2 = \tilde{D}_2 \) where \( D_2 = D(D(A, S_1), S_2) \) and \( \tilde{D}_2 = D(D(A, \tilde{S}_1), \tilde{S}_2) \).

Note that in particular it follows that for any \( A \), the distributions of \( D(D(A, S_1), S_2) \) and \( D(D(A, S_2), S_1) \) are the same. The result gives something rather stronger; namely a coupling
of the two pairs of service processes such that, simultaneously for all arrival processes, the output of the system is the same for both pairs.

Recall that a multitype queue can be seen as a coupling of several queues, with different arrival processes but the same service process. Because Theorem 6.1 gives a coupling which works for all arrival processes simultaneously, we immediately have as a corollary an interchangeability result for multitype queues.

**Theorem 6.2** Suppose the conditions of Theorem 6.1 hold, and let \( A \) be a multitype arrival process. Then \( D(D(A, S_1), S_2) \overset{d}{=} D(D(A, S_2), S_1) \).

### 7 Multitype fixed points

In this section we will define a series of distributions of multi-class processes, denoted by \( F_{m,c,\lambda_1,\ldots,\lambda_m} \).

\( F_{m,c,\lambda_1,\ldots,\lambda_m} \) will be the distribution of an \( m \)-type process, with intensity of \( j \)-th-class customers equal to \( \lambda_j \). We will show that it is a fixed point arrival process for a Ber(\( q, \beta \)) service process, for all \( q \) and \( \beta \) such that \( q/\beta \geq \lambda_1 + \cdots + \lambda_m \) and \( c(q, \beta) = c \) (where the function \( c \) is defined at (5.2)). When \( c \) and \( \{\lambda_j\} \) are fixed, we may abbreviate and write simply \( F_m \). The distributions are defined recursively, using \( F_{m-1} \) to construct \( F_m \).

\( F_1 \) is the distribution of a 1-type process; simply a Ber(\( p \))Geom(\( \alpha \)) process where \( p \) and \( \alpha \) are chosen so that \( p/\alpha = \lambda_1 \) and \( c(p, \alpha) = c \).

Suppose we have constructed \( F_{m-1} \). Now \( F_m \) is constructed as follows. We consider a queue whose service process is Ber(\( q_m \))Geom(\( \beta_m \)), where \( q_m \) and \( \beta_m \) are chosen so that the total service intensity \( q_m/\beta_m \) is equal to \( \lambda_1 + \cdots + \lambda_m \), and so that \( c(q_m, \beta_m) = c \). (As observed in Section 5 there is a unique such choice).

The arrival process to the queue has distribution \( F_{m-1} \) (and is independent of the service process). This leads to an \((m - 1)\)-type departure process. We will extend the departure process to an \( m \)-type process by replacing the unused service in the queue by customers of type \( m \). Let \( F_m \) be the distribution of the \( m \)-type process obtained in this way.

Note that a consequence of the construction and of the fixed point result is that the restriction of a process with distribution \( F_m \) to its first \( m - 1 \) types gives a process with distribution \( F_{m-1} \).

**Theorem 7.1** Let \( S \) be a Ber(\( q \))Geom(\( \beta \)) service process, with \( q/\beta = \mu \) and \( c(q, \beta) = c \). Then for all \( \{\lambda_j\} \) with \( \lambda_1 + \cdots + \lambda_m \leq \mu \), the distribution \( F_{m,c,\lambda_1,\ldots,\lambda_m} \) is a fixed point for the process \( S \).

That is, suppose \( A \) is an \( m \)-type arrival process with distribution \( F_{m,c,\lambda_1,\ldots,\lambda_m} \) and is independent of \( S \). If \( D = D(A, S) \) is the \( m \)-type departure process, then \( D \overset{d}{=} A \).

**Proof:** The case \( m = 1 \) is the version of Burke’s theorem given in Theorem 5.1 above. The arrival process is a one-type Ber(\( p \))Geom(\( \alpha \)) process and (5.1) holds, so indeed the departure process has the same law as the arrival process.

Now let \( m \geq 2 \) and suppose that the result holds as claimed for \( F_{m-1} \). We wish to establish the result for \( F_m \). We will use the interchangeability result in Theorem 6.2.
We consider two independent service processes $S_{\mu}$ and $S_{\lambda_1 + \cdots + \lambda_m}$, which are Bernoulli-geometric processes with parameters $\text{Ber}(r, \gamma)$ and $\text{Ber}(q_m, \beta_m)$ respectively. We choose the parameters such that:

- $r/\gamma = \mu$ and $c(r, \gamma) = c$;
- $q_m/\beta_m = \lambda_1 + \cdots + \lambda_m$ and $c(q_m, \beta_m) = c$.

We will consider an $m$-type arrival process $A$ distributed according to $F_m$. We consider the effect of feeding this arrival process into two tandem systems, comprising the service processes $S_{\mu}$ and $S_{\lambda_1 + \cdots + \lambda_m}$ in their two possible orders. Theorem 6.2 tells us that $D(D(A, S_{\mu}), S_{\lambda_1 + \cdots + \lambda_m})$ and $D(D(A, S_{\lambda_1 + \cdots + \lambda_m}), S_{\mu})$ have the same distribution. To make use of this result we first need two simple properties. The first concerns projections of multi-type processes onto processes with fewer types, and the second concerns the behaviour of a queue which is saturated (i.e. whose service rate is only just sufficient to serve all the arriving customers).

**Claim 1:** Suppose $A$ has distribution $F_m$, and let $G_m$ be the distribution of $D(A, S_{\mu})$. Then the restrictions of $F_m$ and $G_m$ to the first $m-1$ coordinates have the same distribution, and the combined process of all $m$ types under $G_m$ has the same distribution as $S_{\lambda_1 + \cdots + \lambda_m}$.

**Proof of Claim 1:** Restricting $F_m$ to its first $m-1$ coordinates gives the distribution $F_{m-1}$. Since $F_{m-1}$ is an $(m-1)$-type fixed point for the service process $S_{\mu}$, the first $m-1$ types under $G_m$ do indeed have distribution $F_{m-1}$ also.

Finally, from the definition of $F_m$, the combined process of all $m$ types under $F_m$ is a $\text{Ber}(q_m)\text{Geom}(\beta_m)$ process. The service process $S_{\mu}$ is a $\text{Ber}(r)\text{Geom}(\gamma)$ process. By assumption $c(q_m, \beta_m) = c(r, \gamma) = c$. So condition (5.1) is satisfied for the queue, and we can apply the 1-type fixed point result in Theorem 5.1 to show that the distribution of the combined process of all $m$ types in the departure process has the same distribution as the combined process of all $m$ types in the arrival process, namely $S_{\lambda_1 + \cdots + \lambda_m}$. This completes the argument for Claim 1.

**Claim 2:** Suppose $A$, restricted to its first $m-1$ coordinates, has distribution $F_{m-1}$. Suppose also that the combined process of all $m$ types of arrival in $A$ is an i.i.d. process with intensity $\lambda_1 + \cdots + \lambda_m$. Then $D(A, S_{\lambda_1 + \cdots + \lambda_m})$ has distribution $F_m$.

**Proof of Claim 2:** Recall that $F_m$ is the distribution obtained by passing an $(m-1)$-type arrival process with distribution $F_{m-1}$ through a queue with service process $S_{\lambda_1 + \cdots + \lambda_m}$, and putting customers of type $m$ in place of all unused service.

The first $m-1$ components of $A$ indeed have distribution $F_{m-1}$. Certainly the $m$th component of the departure process is a subset of the unused service from the $m-1$ first types. To show that the departure process has distribution $F_m$, it remains to show that all the service unused by the first $m-1$ types is used by customers of type $m$; that is, that if we look at all $m$ types combined, then there is no unused service.

The arrival process of all customers combined is an i.i.d. process with rate $\lambda_1 + \cdots + \lambda_m$; the same is true of the service process, and the arrival and service processes are independent. Hence by Lemma 3.1 there is no unused service in the queue as desired. So indeed the output process has distribution $F_m$, and we have established Claim 2 as required.
Now we put together Claims 1 and 2 to complete the proof. We consider the two possible orderings of the service processes $S_\mu$ and $S_{\lambda_1 + \ldots + \lambda_m}$. The two systems that arise can be illustrated as follows:

$$
\begin{align*}
F_m &\rightarrow S_\mu \rightarrow G_m \rightarrow S_{\lambda_1 + \ldots + \lambda_m} \rightarrow H_m \\
F_m &\rightarrow S_{\lambda_1 + \ldots + \lambda_m} \rightarrow J_m \rightarrow S_\mu \rightarrow K_m
\end{align*}
$$

By Claim 1, $G_m$ satisfies the conclusion of Claim 1, which is also the assumption on the arrival process of Claim 2. Then by Claim 2, $H_m = F_m$.

Certainly $F_m$ itself also satisfies the assumption of Claim 2, so by Claim 2, $J_m = F_m$.

Theorem 6.1 says that $H_m = K_m$.

Then the second box of the second line down shows us that if we pass $J_m$ through $S_\mu$ we get $K_m$.

But $J_m = K_m = F_m$. So indeed $F_m$ is a fixed point for $S_\mu$ as desired. $\square$

## 8 Proof of interchangeability result

In this section we will prove Theorem 6.1. We first need two simple lemmas about queues in tandem.

**Lemma 8.1** Consider a system of two queues in tandem with arrival process $A$ and service processes $S_1$ and $S_2$, all taking integer values. For $s \in \mathbb{Z}$, define $A|_{(-s, \infty)}$ to be the arrival process truncated before time $-s$. That is,

$$
A|_{(-s, \infty)}(n) = \begin{cases} 
0 & \text{if } n < -s, \\
A(n) & \text{if } n \geq -s.
\end{cases}
$$

Define $D_2 = D(D(A, S_1), S_2)$ and $D_2^{(s)} = D(D(A|_{(-s, \infty)}, S_1), S_2)$. For any $n$, if $s$ is large enough then $D_2(n) = D_2^{(s)}(n)$.

**Proof:** We compare the original system and the system with the arrival process before time $-s$. Write $D_1$ and $D_2$ for the departure processes from the first and second queues in the original system, and similarly $D_1^{(s)}$, $D_2^{(s)}$ for the processes in the system with the truncated arrival process.

Take any $n \in \mathbb{Z}$. From the definitions in Section 3 we have

$$
D_1(n) = \min \left( S_1(n), A(n) + \sup_{u \leq n} \sum_{r = u}^{n-1} [A(r) - S_1(r)] \right),
$$

$$
D_1^{(s)}(n) = \min \left( S_1(n), A(n) + \sup_{-s \leq u \leq n} \sum_{r = u}^{n-1} [A(r) - S_1(r)] \right).
$$

Certainly $D_1^{(s)}(n)$ is increasing in $s$, and $D_1^{(s)}(n) \leq D_1(n)$ for all $s$.

We first observe that $D_1^{(s)}(n) = D_1(n)$ for all large enough $s$. Suppose the sup in the first line is finite. Then (since the variables all take integer values) it is attained for some $u = u^*$, and we have $D_1(n) = D_1^{(s)}(n)$ for all $s > -u^*$. If instead the sup in the first line is infinite,
then by taking \( s \) large we may make the sup in the second line as large as desired; if \( s \) is large enough that the second term in the min exceeds \( S_1(n) \), then we have \( D_1(n) = D_1(s)(n) \), as required.

Now we have the corresponding expressions for the departure process from the second queue (whose arrival process is \( D_1 \) and whose service process is \( S_2 \)):

\[
D_2(n) = \min \left( S_2(n), D_1(n) + \sup_{u \leq n} \sum_{r = u}^{n-1} [D_1(r) - S_2(r)] \right),
\]

\[
D_2(s)(n) = \min \left( S_2(n), D_1(s)(n) + \sup_{u \leq n} \sum_{r = u}^{n-1} [D_1(s)(r) - S_2(r)] \right).
\]

Since \( D_1(s)(n) \) is increasing in \( s \) and bounded above by \( D_1(n) \), we also have that \( D_2(s)(n) \) is increasing in \( s \) and bounded above by \( D_2(n) \).

Suppose that the sup in the first line is finite. Then as before it is attained at some \( u^* \). Now if we take \( s \) large enough that \( D_2(s)(r) = D_1(r) \) for all \( r \) with \( u^* \leq r \leq n \), then we see from the second line that indeed \( D_2(s)(n) \geq D_2(n) \), and hence in fact \( D_2(s)(n) = D_2(n) \) as required. A similar argument applies if the sup in the first line is infinite; by taking \( s \) large enough, we can make the sup in the second line as large as desired.

**Lemma 8.2** Consider a tandem of two queues, with arrival process \( A \) and service processes \( S_1 \) and \( S_2 \). Suppose that \( A(n) = 0 \) for all \( n < 0 \). Let \( D_2 \) be the departure process from the second queue. Then for all \( t \geq 1 \),

\[
\sum_{r=0}^{t-1} D_2(r) = \inf_{0 \leq u_1 \leq u_2 \leq t} \left\{ \sum_{r=0}^{u_1-1} A(r) + \sum_{r=u_1}^{u_2-1} S_1(r) + \sum_{r=u_2}^{t-1} S_2(r) \right\}.
\] (8.1)

**Proof:** An equivalent result for a related queueing model was given in [28], and similar properties appear in many places in the literature. Formula (8.1) is essentially a special case of equation (10) of [27].

Let \( D_1 \) and \( D_2 \) be the departure processes from the first and second queues, and \( X_1 \) and \( X_2 \) be the queue-length processes at the first and second queues.

Since \( A_n = 0 \) for all negative \( n \), the queues start empty at time 0: \( X_1(0) = X_2(0) = 0 \). So the total number of departures from the system (that is, from the second queue) before time \( t \) is given by the total number of arrivals during that time, minus the queue-lengths at time \( t \):

\[
\sum_{r=0}^{t-1} D_2(r) = \sum_{r=0}^{t-1} A(r) - X_1(t) - X_2(t).
\] (8.2)

Similarly, considering the evolution of the first queue alone between times \( u \) and \( t \), we have that for any \( u < t \),

\[
\sum_{r=u}^{t-1} D_1(r) = \sum_{r=u}^{t-1} A(r) - X_1(t) + X_1(u).
\]

Now \( X_2 \) is the queue-length of a queue with arrivals \( D_1 \) and services \( S_2 \). So from the definition (8.1), we have

\[
X_1(t) + X_2(t) = X_1(t) + \sup_{u_2 \leq t} \sum_{r=u_2}^{t-1} [D_1(r) - S_2(r)]
\]

15
\[ X_1(t) + \sup_{u_1 \leq t} \left\{ \sum_{r=0}^{t-1} A(r) - \sum_{r=0}^{t-1} S_2(r) \right\} \]

\[ = \sup_{u_2 \leq t} \left\{ X_1(u) + \sum_{r=0}^{u_2-1} [A(r) - S_2(r)] \right\} \]

\[ = \sup_{u_2 \leq t} \left\{ \sum_{u_1 \leq u_2} \sum_{r=0}^{u_2-1} [A(r) - S_1(r)] + \sum_{r=0}^{t-1} [A(r) - S_2(r)] \right\} \]

\[ = \sup_{u_1 \leq u_2 \leq t} \left\{ \sum_{r=0}^{u_2-1} A(r) - \sum_{r=0}^{u_1-1} S_1(r) - \sum_{r=0}^{t-1} S_2(r) \right\} . \]

Since \( A_n = 0 \) for negative \( n \), the sup will be attained for some \( u_1 \geq 0 \). Then using (8.2) gives (8.1) as desired.

Now we proceed to prove Theorem 6.1. To do this, we need to construct \( S_1, S_2, \tilde{S}_1 \) and \( \tilde{S}_2 \) on the same probability space, in such a way that:

(i) \( (S_1, S_2) \cong (\tilde{S}_2, \tilde{S}_1) \)

(ii) For all \( A, D(D(A, S_1), S_2) = D(D(A, \tilde{S}_1), \tilde{S}_2) \).

We assume that \( S_2 \) has higher intensity than \( S_1 \). (If not, then swap them around. If the intensity is the same, then the assumptions ensure that they have the same distribution.)

The construction is as follows. We think of a queue with arrival process \( S_1 \) and service process \( S_2 \). We can decompose \( S_2 \) into \( D(S_1, S_2) \) and \( U(S_1, S_2) \) — that is, \( S_2 = D(S_1, S_2) + U(S_1, S_2) \).

We let \( \tilde{S}_2 = D(S_1, S_2) \). So we have obtained \( \tilde{S}_2 \) from \( S_2 \) by removing the unused service process. By Theorem 5.1, we have that \( (S_1, \tilde{S}_2) \cong (\tilde{S}_2, S_1^*) \). (Recall that the asterisk denotes the reverse of a process). Knowing this, we can extend the probability space to include a random variable \( \tilde{S}_1 \), with its time-reversal \( \tilde{S}_1^* \), such that \( (S_1, \tilde{S}_2, S_2) \cong (\tilde{S}_2, S_1^*, \tilde{S}_1^*) \).

In particular this implies that \( (S_1, S_2) \cong (\tilde{S}_2^*, \tilde{S}_1^*) \). But the distribution of \( (S_1, S_2) \) is invariant under time reversal, so this definition gives us (i) above as desired.

(Note that since \( (S_1, \tilde{S}_2, S_2) \cong (\tilde{S}_2^*, S_1^*, \tilde{S}_1^*) \), we have in particular that \( S_2^* = D(\tilde{S}_2^*, \tilde{S}_1^*) \), or equivalently that \( S_1^* + U(\tilde{S}_2^*, \tilde{S}_1^*) = \tilde{S}_1^* \). So we can think of \( \tilde{S}_1 \) as being obtained from \( S_1 \) by adding an unused service process, for a queue operating in reverse time).

Now we need to verify (ii) above. Using Lemma 8.1, it will be sufficient to verify it for every arrival process which is 0 up to some finite time. Without loss of generality, suppose that \( A(n) = 0 \) for all \( n < 0 \). Let \( D = D(D(A, S_1), S_2) \). From Lemma 8.2 we have the following representation of the total number of departures from the tandem of two queues from time 0 up to time \( t - 1 \):

\[ \sum_{r=0}^{t-1} D(r) = \inf_{0 \leq u_1 \leq u_2 \leq t} \left\{ \sum_{r=0}^{u_1-1} A(r) + \sum_{r=0}^{u_2-1} S_1(r) + \sum_{r=0}^{t-1} S_2(r) \right\} . \]

It will be enough to show that for any \( A \), this quantity is unchanged if we replace \((S_1, S_2)\) by \((\tilde{S}_1, \tilde{S}_2)\). For this, in turn it’s enough that for all \( s \) and \( t \), the quantity

\[ \inf_{s \leq u \leq t} \left\{ \sum_{r=s}^{u-1} S_1(r) + \sum_{r=u}^{t-1} S_2(r) \right\} \]
is unchanged if we replace \((S_1, S_2)\) by \((\tilde{S}_1, \tilde{S}_2)\).

We consider the queue with arrival process \(S_1\) and service process \(S_2\), along with its queue-length, unused service and departure processes \(X, U\) and \(D\) which are functions of \(S_1\) and \(S_2\). Note that \(U(n) > 0\) precisely if \(\tilde{S}_2(n) < S_2(n)\).

First we show that the quantity we are interested in is unchanged if we replace \((S_1, S_2)\) by \((S_1, \tilde{S}_2)\); that is:

\[
\inf_{s \leq u \leq t} \left\{ \sum_{r=s}^{u-1} S_1(r) + \sum_{r=u}^{t-1} S_2(r) \right\} = \inf_{s \leq u \leq t} \left\{ \sum_{r=s}^{u-1} S_1(r) + \sum_{r=u}^{t-1} \tilde{S}_2(r) \right\} .
\] (8.3)

If there is no unused service between times 0 and \(t - 1\), then \(S_2\) and \(\tilde{S}_2\) agree on the whole interval and the equality is obvious.

Otherwise, let \(n\) be the latest time in \(\{s, \ldots, t - 1\}\) for which there is unused service, i.e. for which \(S_2(n) > \tilde{S}_2(n)\), or equivalently that \(U(n) > 0\) where \(U = U(S_1, S_2)\). In particular there are no more arrivals than departures in the interval \(\{s, \ldots, n\}\) (otherwise there would still be customers in the queue after time \(n\), and hence there could not have been unused service at time \(n\)). That is, \(\sum_{i=s}^{n} S_1(i) \leq \sum_{i=s}^{n} \tilde{S}_2(i)\). This implies that we can take \(u > n\) in minimising the RHS of (8.3). Then consider using the same \(u\) on the LHS as on the RHS. Since \(u > n\) and \(n\) was the last moment of unused service, the processes \(S_2\) and \(\tilde{S}_2\) agree after time \(n\). So in fact the LHS is at least as small as the RHS. But since \(S_2 \geq \tilde{S}_2\) it is also clear that the LHS is no smaller than the RHS. So in fact the two are the same.

Now we want to show that we can further replace \(S_1\) by \(\tilde{S}_1\); that is,

\[
\inf_{s \leq u \leq t} \left\{ \sum_{r=s}^{u-1} S_1(r) + \sum_{r=u}^{t-1} \tilde{S}_2(r) \right\} = \inf_{s \leq u \leq t} \left\{ \sum_{r=s}^{u-1} \tilde{S}_1(r) + \sum_{r=u}^{t-1} \tilde{S}_2(r) \right\} .
\] (8.4)

This is in fact precisely the same relation as (8.3), but now for the queue with arrival process \(\tilde{S}_2\) and service process \(\tilde{S}_1\). So an exactly analogous argument can be applied.

This completes the proof of (ii) and hence of Theorem 6.1.

9 Summary and examples

In this section we point out various interesting cases that can be obtained by taking particular values or appropriate limits in the general model considered up to now.

Our general model is of a service process in discrete time, whose batches are i.i.d. with \(\text{Ber}(q)\text{Geom}(\beta)\) distribution. Such a process has service intensity \(q/\beta\). This model has two parameters. For each choice \((q, \beta)\) there is a one-dimensional family of service process which are interchangeable with it; namely an \((r, \gamma)\)-process is interchangeable with a \((q, \beta)\)-process whenever \(c(r, \gamma) = c(q, \beta)\). This relation partitions the parameter space into equivalence classes (and each class contains precisely one set of parameters for each service intensity).

Suppose we are given service parameters \((q, \beta)\) with service intensity \(\mu = q/\beta\), and also \(\lambda_1, \ldots, \lambda_m\) such that \(\lambda_1 + \cdots + \lambda_m < \mu\). Then there is precisely one \(m\)-type fixed point arrival process with intensities \(\lambda_r\) of \(r\)-th-class customers for \(1 \leq r \leq m\). For example, the two-type fixed point with intensities \(\lambda_1\) and \(\lambda_2\) is obtained by taking the departure and unused service processes from a queue with arrival process \(A\) and service process \(S\) where \(A\) and \(S\) are chosen
so that $A$ has intensity $\lambda_1$, $S$ has intensity $\lambda_1 + \lambda_2$, and both $A$ and $S$ are interchangeable with the $(q, \beta)$-process. (As explained in the last paragraph, this uniquely determines $A$ and $S$).

This framework extends to various particular cases which we describe below. In some cases, the model has only one parameter, and any pair of processes in the class considered are interchangeable.

### 9.1 $/M/1$ queue in discrete time

If we take $\beta = 1$, the service process is a “Bernoulli process”. The batches are i.i.d., each equal to 0 with probability $1 - q$ and 1 with probability $q$, so that $\mu = q$. Hence we have a one-parameter model.

This is the discrete-time equivalent of the $/M/1$ server described in Section 2. A version of Burke’s theorem for such systems was proved by Hsu and Burke in [17], any Bernoulli process with lower intensity $p < q$ is a one-type fixed point arrival process.

Any two such Bernoulli processes are interchangeable. For any $m$ and $\lambda_1 + \cdots + \lambda_m < 1$, there is an arrival process $F_{m,\lambda_1,\ldots,\lambda_m}$ with intensities $\lambda_i$ and which is a fixed for any such $/M(\mu)/1$ server whenever $\mu > \lambda_1 + \cdots + \lambda_m$. The multi-type fixed points were constructed in [13], and they can also be seen as stationary distributions for multi-type versions of the TASEP (totally asymmetric simple exclusion process).

### 9.2 Geometric and exponential batches in discrete time

If we take $q = 1$, then the service batches are i.i.d. geometric distributions. (A similar case arises for $q = 1 - \beta$, except that the geometric distribution obtained starts from 0 rather than from 1). Again we have a one-parameter model, and any two such geometric processes are interchangeable. A related result (for a particular form of the arrival process) was given by Draief, O’Connell and Mairesse in [9]. A version of Burke’s theorem had been given by Bedekar and Azizoğlu in [6].

The multi-type fixed points in this case are related to stationary distributions for a multi-type version of the totally asymmetric zero-range process [22].

Now take the limit $\delta \to 0$, putting $\beta = \delta/u$ and rescaling work by a factor $\delta$. In this way we can obtain the case where the batches are i.i.d. exponential with mean $u$. Again, any two such exponential processes are interchangeable.

### 9.3 Batch queues in continuous time

Suppose we let $\epsilon \to 0$, take $q = \nu \epsilon$ and rescale time by a factor $\epsilon$.

Then we have batches which are geometric with parameter $\beta$, which occur at times of a Poisson process of rate $\nu$.

Now we have a two-parameter distribution, and for each pair of parameters $(\nu, \beta)$ there is a one-parameter family of processes, one for each service intensity, which are interchangeable with it. Namely, processes of this kind with parameters $(\nu, \beta)$ and $(\rho, \gamma)$ are interchangeable.
If
\[ \frac{\nu\beta}{1-\beta} = \frac{\rho\gamma}{1-\gamma}. \]

If we further take \( \delta \to 0 \), putting \( \beta = \delta/u \) and rescaling work by a factor \( \delta \), then we obtain the case where the batches are exponential with mean \( u \) (again occurring at the times of a Poisson process of rate \( \nu \)). Two such “Poisson-exponential” processes with parameters \((\nu, u)\) and \((\rho, v)\) are interchangeable if \( \nu/u = \rho/v \).

If on the other hand we take \( \beta \to 1 \), the batches all have size 1 and we obtain the familiar \( /M/1 \) queueing server which we described in detail in Section 2.

9.4 Brownian queues

Queues in continuous time (as in the previous example) can be represented using a notation analogous to that introduced for discrete-time queues in Section 3.

Let \( S_t, t \in \mathbb{R} \) be a service process and \( A_t, t \in \mathbb{R} \) be an arrival process. We interpret \( S_t - S_s \) as the amount of service offered in the interval \((s, t]\), and similarly \( A_t - A_s \) as the amount of work arriving in \((s, t]\).

Adding a constant to either of the processes \( A \) and \( S \) makes no difference, so we may take for example \( A_0 = S_0 = 0 \).

We define processes \( Q_t, D_t \) and \( U_t \), as functions of the processes \( A_t \) and \( S_t \), by
\[
Q_t = \sup_{s < t} \left\{ (A_t - A_s) - (S_t - S_s) \right\},
\]
\[
D_t = A_t + Q_0 - Q_t,
\]
\[
U_t = S_t - D_t.
\]

Now \( D_t - D_s \) is the amount of work departing in \((s, t]\), \( U_t - U_s \) is the amount of unused service in \((s, t]\), and \( Q_t \) is the queue-length at time \( t \).

Note that under these definitions we have \( U_0 = D_0 = 0 \) (again, this normalization is not important).

A frequently studied example is that of the Brownian queue. Let \( A_t \) be a two-sided Brownian motion with drift \( \lambda \) and variance 1, and let \( S_t \) be a two-sided Brownian motion with drift \( \mu \) and variance 1, where \( \lambda, \mu \in \mathbb{R} \) and \( \lambda < \mu \).

This model arises naturally as the scaling limit of queues in the so-called “heavy traffic” regime [15], [31]. For example, for \( n \in \mathbb{N} \) consider a discrete time queue whose arrival process \( A(r), r \in \mathbb{Z} \) is a Bernoulli process with rate \( 1/2 + \lambda/\sqrt{n} \) and whose service process \( S(r), r \in \mathbb{Z} \) is a Bernoulli process with rate \( 1/2 + \mu/\sqrt{n} \). We recentre and rescale by defining, for \( t \in \mathbb{R} \),
\[
\tilde{A}^{(n)}(t) = \frac{\sum_{r=0}^{nt} A(r) - \frac{nt}{2}}{\sqrt{n/2}} , \tilde{S}^{(n)}(t) = \frac{\sum_{r=0}^{nt} S(r) - \frac{nt}{2}}{\sqrt{n/2}} .
\]

Then as \( n \to \infty \), the distribution of the processes \( \tilde{A}^{(n)}(t) \) and \( \tilde{S}^{(n)}(t) \) converges to that of \( A_t \) and \( S_t \) above.

A version of Burke’s theorem holds for such a Brownian queue: namely that the processes \( D_t \) and \( A_t \) have the same distribution. This was shown by Harrison and Williams in [16] (see also [20] in particular, where the result is generalized in a number of ways). The process \( Q_t \)
has the law of a stationary reflected Brownian motion with drift \(-(\mu - \lambda)\), and its stationary distribution is exponential with rate \(\mu - \lambda\). The process \(U_t\) is the local-time process of \(Q_t\) at 0. Note that \(U\) is non-decreasing, and grows at average rate \(\mu - \lambda\), but is constant except on a set of measure 0 (namely, the set of times when \(Q = 0\)).

Thus we may say that a Brownian queueing server (with drift \(\mu\)) has Brownian motions of drift \(\lambda < \mu\) as fixed point arrival processes. Again this can be extended to multi-type arrival processes.

Given a service process \(S_t\) and a two-type arrival process \(A^{(1)}_t, A^{(2)}_t\), the two-type departure process \(D^{(1)}_t, D^{(2)}_t\) is defined in a natural way by

\[
D^{(1)}_t = D\left(A^{(1)}_t, S_t\right),
\]

\[
D^{(1)}_t + D^{(2)}_t = D\left(A^{(1)}_t + A^{(2)}_t, S_t\right).
\]

Suppose we wish to construct a two-type fixed point for a Brownian service process of drift \(\mu\), which has drifts \(\lambda_1\) and \(\lambda_2\) where \(\lambda_1 < \lambda_1 + \lambda_2 < \mu\). To do this, consider a Brownian queue whose arrival process is a Brownian motion with drift \(\lambda_1\), and whose service process is a Brownian motion with drift \(\lambda_2\). The process \((D, U)\) of departures and unused service from this queue is a two-type fixed point of the required type.

In this example we can see an extreme form of the “clustering of lower-class customers” already observed in the discrete case. Here the second-class work, represented by the local-time process \(U\), is concentrated on a set of times of measure 0. Note also the difference in nature between the process \(D\) of first-class work (which is non-monotonic, and in fact has unbounded variation) and the process \(U\) of second-class work (which is non-decreasing).

Fixed points with larger numbers of classes can be constructed recursively in the same way as in the discrete case. The subprocesses of \(m\)-th class work in these fixed points are non-decreasing and singular, for each \(m \geq 2\).

One can also obtain an interchangeability result; two Brownian service processes, with different drifts but with the same variance, are interchangeable.

Arrival processes obtained as the local time of a reflecting diffusion have been considered before in various contexts, for example in [20] and [18]. These models were also motivated in part by the modelling of a second-class departure process as a local-time process. However, the models analysed are somewhat different; the service processes are deterministic rather than Brownian \((S_t = ct)\) and the local-time arrival process has access to all the service capacity (rather than sharing the queue with a higher-priority stream).

10 Interacting particle systems

There is a close analogy between fixed-point arrival processes for queues and stationary distributions for interacting particle systems.

Consider for example the TASEP (totally asymmetric simple exclusion process). In the one-type version of the process, each site of \(Z\) contains either a particle or a hole. The state-space of the process can be written as \(\{1, \infty\}^Z\), where 1 denotes a particle and \(\infty\) denotes a hole. The dynamics of the process are as follows: each particle tries to jump to its left as a
Poisson process of rate 1, and a jump succeeds if the site to the left of the particle is empty (in which case the values 1 and \(\infty\) are exchanged between the two sites).

One can also consider a multi-type TASEP. The process with \(m\) types has state space \((\{1,2,\ldots,m\} \cup \{\infty\})^2\). As before, each particle tries to jump to its left as a Poisson process of rate 1, and a jump by a particle of type \(r\) succeeds if the site to its left is occupied by a particle with lower priority (higher-numbered class) or is empty (in which case the values at the two sites are interchanged as before). As in the case of multi-type queues, this process can be seen in a natural way as a coupling of \(m\) single-type TASEPs.

Stationary distributions for TASEPs with two types of particle were constructed by Derrida, Janowsky, Lebowitz and Speer [8], and related construction were also given in [11], [5] and [10]. In [13], this construction was extended to multi-type processes, and made particularly explicit using a construction using queues in tandem which corresponds to the construction of multi-type fixed points for queues above.

Given \(\lambda_1,\ldots,\lambda_m\) with \(\lambda_1 + \cdots + \lambda_m < 1\), consider the fixed-point arrival process distribution \(F_{m,\lambda_1,\ldots,\lambda_m}\) for the discrete-time /M/1 queue as described in Section 9.1. Now regard this distribution as a distribution over configurations of the \(m\)-type TASEP; an arrival of type \(r\) at time \(n\) corresponds to a particle of type \(r\) at site \(n\), and an empty arrival slot at time \(n\) corresponds to a hole at site \(n\).

The result of [13] is that \(F_{m,\lambda_1,\ldots,\lambda_m}\) is a stationary distribution for the \(m\)-type TASEP (indeed, it is the unique ergodic stationary distribution which has intensities \(\lambda_r\)).

Note that time in the context of the queue corresponds to space in the context of the particle system.

The analogy between queues and particle systems can be made even clearer by considering particle systems evolving in discrete time rather than continuous time. The same distributions \(F_m\) are also stationary for certain versions of the multi-type TASEP in discrete time [23]. (Here many particles may try to jump at the same time-step, and there are various possible natural conventions for the order in which the jump attempts are to be processed). Now a queueing server, which uses a service process to transform an arrival process into a departure process, can be seen as analogous to a set of jump attempts at particular sites at a given time-step of the particle system, which transforms the particle configuration before that time-step into the new particle configuration after that time-step. Successive updates in the particle system correspond to consecutive queueing servers in a system of queues in tandem.

The fixed-point property for the distributions \(F_m\) can be derived from the result for the particle system (as was done in [13]). To see this, note that if we take an \(m\)-class TASEP and treat particles of class \(m\) as holes, then the resulting restriction of the process to the first \((m-1)\) types is an \((m-1)\)-class TASEP. Hence restricting the stationary distribution \(F_m\) of the \(m\)-class system to its first \((m-1)\) types must give a stationary distribution of the \((m-1)\)-class system, namely \(F_{m-1}\). That is, \(F_{m-1}\) can be obtained from \(F_m\) by restricting to the first \((m-1)\) types. But also, the first \((m-1)\) types of \(F_m\) are defined as the departure process from a queue whose arrival process is \(F_{m-1}\) (the remaining \(m\)th class is the unused service process). So indeed \(F_{m-1}\) is also the distribution of the departure process from the queue; this is the fixed-point result we obtained earlier.

However, such a proof via the particle system is rather indirect and perhaps less natural...
than the approach using interchangeability above.

In fact, in some cases, results previously obtained for stationary distributions of particle systems can themselves be derived from the fixed-point results for queues. For example, the stationary distributions for “Hammersley’s process” \[1\] correspond to the fixed points for the \( /M \cdot M \) queue in continuous time, described in Section \[2\] \[14\] \[12\]. The dynamics of Hammersley’s process can be obtained by taking appropriate limits starting from a \( M/M/1 \) queueing server in discrete time (first one lets the rates of arrival and service approach 1 together in an appropriate way; after rescaling space, the particles in a discrete-time version Hammersley’s process correspond to the gaps in the customer processes in the queue. Then the more familiar continuous-time version of Hammersley’s process can be obtained by letting the rate of jumps in the discrete version go to 0 and rescaling time). In this way the result on stationary distributions for Hammersley’s process can be seen as a consequence of the fixed-point results given here. (However, we don’t know how to use similar methods to derive the results for the TASEP, for example).

11 Continuous class-labels

By considering the limit as the number of classes goes to infinity, and the intensity of each class goes to zero, one can arrive at models where the classes are indexed by real numbers.

This could be done in the general case above, but for simplicity we discuss it in the case of a \( /M/1 \) server (where the service process is a Bernoulli process). We will consider an arrival process in which at most one customer arrives at each time slot, and a customer arriving at time \( n \) has a real-valued label. As before, lower-labelled customers have priority over higher-labelled customers.

By taking appropriate limits in the results above, one can obtain the following result:

**Proposition 11.1** There exists a unique distribution over sequences \((L(n), n \in \mathbb{Z}) \in [0, 1]^\mathbb{Z}\) with the following properties:

(i) The distribution is stationary and ergodic.

(ii) \(L(0) \sim U[0, 1]\).

(iii) Let \(0 < \lambda < 1\). Define an arrival process as follows: if \(L(n) \leq \lambda\), then a customer with label \(L(n)\) arrives at time \(n\), while if \(L(n) > \lambda\), then no customer arrives at time \(n\). This arrival process is a fixed point for the \( /M(\mu)/1 \) queue for all \(\mu > \lambda\).

Property (ii) is included just as a normalization (since the mechanics of the queue are unchanged if the labels of all the customers are transformed by an increasing function). Then in property (iii), the condition \(\mu > \lambda\) ensures that the queue is not saturated.

The distribution of \((L(n), n \in \mathbb{Z})\) is also a stationary distribution for the TASEP. See \[2\] for an investigation of this process, including in particular an interpretation as the “speed process” for a multi-type TASEP started out of equilibrium.

One interesting property of the process is a manifestation of the “clustering” effect described above. Although \(L(n)\) has a continuous distribution for each \(n\), nonetheless one has
that for any $n$, $P(L(0) = L(n)) > 0$. (For example, $P(L(0) = L(1)) = 1/6$). In fact, with probability one there exist infinitely many $n$ such that $L(0) = L(n)$. Hence clustering occurs in the following sense: although any class-label has probability 0 of being seen a priori, if one sees the label at any particular time the same label has high probability of being seen nearby, and will be seen infinitely often in the process.

Acknowledgments

JBM thanks Pablo Ferrari for many valuable conversations related to this work, and Mike Harrison and Ilkka Norros for discussions about the results on Brownian queues.

References

[1] Aldous, D. and Diaconis, P., (1995) Hammersley’s interacting particle process and longest increasing subsequences. *Probab. Theory Related Fields* **103**, 199–213.

[2] Amir, G., Angel, O. and Valko, B., (2008). The TASEP speed process. Preprint arXiv:0811.3706.

[3] Anantharam, V., (1987) Probabilistic proof of the interchangeability of $\cdot / M / 1$ queues in series. *Queueing Systems Theory Appl.* **2**, 387–392.

[4] Anantharam, V., (1993) Uniqueness of stationary ergodic fixed point for a $\cdot / M / K$ node. *Ann. Appl. Probab.* **3**, 154–172.

[5] Angel, O., (2006) The stationary measure of a 2-type totally asymmetric exclusion process. *J. Combin. Theory Ser. A* **113**, 625–635.

[6] Bedekar, A. S. and Azizoğlu, M., (1998) The information-theoretic capacity of discrete-time queues. *IEEE Trans. Inform. Theory* **44**, 446–461.

[7] Burke, P. J., (1956) The output of a queuing system. *Operations Res.* **4**, 699–704.

[8] Derrida, B., Janowsky, S. A., Lebowitz, J. L. and Speer, E. R., (1993) Exact solution of the totally asymmetric simple exclusion process: shock profiles. *J. Statist. Phys.* **73**, 813–842.

[9] Draief, M., Mairesse, J. and O’Connell, N., (2005) Queues, stores, and tableaux. *J. Appl. Probab.* **42**, 1145–1167.

[10] Duchi, E. and Schaeffer, G., (2005) A combinatorial approach to jumping particles. *J. Combin. Theory Ser. A* **110**, 1–29.

[11] Ferrari, P. A., Fontes, L. R. G. and Kohayakawa, Y., (1994) Invariant measures for a two-species asymmetric process. *J. Statist. Phys.* **76**, 1153–1177.

[12] Ferrari, P. A. and Martin, J. B., (2006) Multiclass processes, dual points and $M/M/1$ queues. *Markov Proc. Related Fields* **12**, 175–201.
[13] Ferrari, P. A. AND Martin, J. B., (2007) Stationary distributions of multi-type totally asymmetric exclusion processes. *Ann. Probab.* **35**, 807–832.

[14] Ferrari, P. A. AND Martin, J. B., (2009) Multiclass Hammersley-Aldous-Diaconis process and multiclass-customer queues. *Ann. Inst. Henri Poincaré Probab. Stat.* **45**, 250–265.

[15] Harrison, J. M., (1985) *Brownian motion and stochastic flow systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

[16] Harrison, J. M. AND Williams, R. J., (1990) On the quasireversibility of a multiclass Brownian service station. *Ann. Probab.* **18**, 1249–1268.

[17] Hsu, J. AND Burke, P. J., (1976) Behavior of tandem buffers with geometric input and Markovian output. *IEEE Trans. Comm.* **COM-24**, 358–361.

[18] Konstantopoulos, T., Kyprianou, A. E., Salminen, P. AND Sirvio, M., (2008) Analysis of stochastic fluid queues driven by local-time processes. *Adv. in Appl. Probab.* **40**, 1072–1103.

[19] Lehtonen, T., (1986) On the ordering of tandem queues with exponential servers. *J. Appl. Probab.* **23**, 115–129.

[20] Mannersalo, P., Norros, I. AND Salminen, P., (2004) A storage process with local time input. *Queueing Syst.* **46**, 557–577.

[21] Martin, J. B., (2009) Batch queues, reversibility and first-passage percolation. *Queueing Syst.* **62**, 411–427.

[22] Martin, J. B., (2010) Stationary distributions for multi-class zero-range processes. In preparation.

[23] Martin, J. B. AND Schmidt, P., (2010) Multi-type TASEP in discrete time. Preprint arXiv:1002.3539.

[24] Mountford, T. AND Prabhakar, B., (1995) On the weak convergence of departures from an infinite series of -/M/1 queues. *Ann. Appl. Probab.* **5**, 121–127.

[25] O’Connell, N., (2003) Random matrices, non-colliding particle systems and queues. In *Séminaire de Probabilités XXXVI*, no. 1801 in Lecture Notes in Mathematics, pages 165–182. Springer-Verlag.

[26] O’Connell, N. AND Yor, M., (2001) Brownian analogues of Burke’s theorem. *Stochastic Process. Appl.* **96**, 285–304.

[27] O’Connell, N. M., (2000) Directed percolation and tandem queues. HP Labs technical report, HPL-BRIMS-2000-28, http://www.hpl.hp.com/techreports/2000/.

[28] Szczotka, W. AND Kelly, F. P., (1990) Asymptotic stationarity of queues in series and the heavy traffic approximation. *Ann. Probab.* **18**, 1232–1248.
[29] Tsoucas, P. and Walrand, J., (1987) On the interchangeability and stochastic ordering of $\cdot/M/1$ queues in tandem. *Adv. in Appl. Probab.* **19**, 515–520.

[30] Weber, R. R., (1979) The interchangeability of $\cdot/M/1$ queues in series. *J. Appl. Probab.* **16**, 690–695.

[31] Williams, R. J., (1996) On the approximation of queueing networks in heavy traffic. In F. P. Kelly, S. Zachary and I. Ziedins, eds., *Stochastic Networks: Theory and Applications*, pages 35–56. Clarendon Press, Oxford.

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK

martin@stats.ox.ac.uk

Department of Electrical Engineering and Computer Science, Packard Building, Stanford University, Stanford, CA 94305-9510, USA

balaji@stanford.edu