UvA-DARE (Digital Academic Repository)

The Foreground-Background Queue
Nuijens, M.F.M.

Link to publication

Citation for published version (APA):
Nuijens, M. F. M. (2004). The Foreground-Background Queue

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 1

The Foreground-Background discipline

1.1 Queueing theory

Imagine a person who solves (mathematical) problems as a profession. Every month he receives a number of problems. Many of them can be solved in a few minutes, most of them within an hour, but some problems may take weeks or even years to solve. Since eventually every problem is solved in an instant of insight, it is impossible beforehand to say how much time he will spend on a problem. A reasonable strategy to solve as many problems as possible is to dedicate a few minutes to a problem, and if it cannot be solved, to put it aside and start with a new problem. When all problems present on his desk have received a few minutes of attention, he returns to a problem that was put aside and spends half an hour on that problem. Then he returns to another problem that was put aside, and so on.

The situation in the example above may be modelled as a queue. The strategy is called the service discipline, queue(ing) discipline or simply discipline. A large number of real world phenomena can be thought of in a queueing context. One should think of communication networks like internet and call centres, queues for a printer,
in the post office and supermarket, but also inventory systems may be modelled as queues. Hence it is no wonder that a whole branch of (applied) probability theory is devoted to queueing.\(^1\)

It is clear that by means of the service discipline one may influence the behaviour of a queueing system. Consider for example what happens to the queue described above when the problem solver uses the FIFO (First in first out) service discipline, which gives priority to the first customer in line, as is tradition in for example post offices. This strategy may work for some time, until the mathematician hits upon a problem that takes years to solve. Then the size of the pile of problems waiting to be solved is likely to grow very large. The question which service discipline to use, is therefore a crucial one.

Queueing theory has been around since the beginning of the twentieth century. The Danish mathematician Erlang may be considered as the founder of queueing theory, since his studies for the Copenhagen Telephone Company between 1909 and 1920 form pioneering work in the theory of queues. Since then queueing theory has been the object of growing interest. From different angles and with different levels of mathematical rigour people have investigated queueing models. This has resulted in a rich variety of models and techniques in the literature. Nowadays hundreds of articles on the subject appear yearly and several journals are devoted to queueing theory. For an account of the (early) history of queueing theory up to 1960 we refer to Saaty (1961, \[56\]). The developments in the second half of the twentieth century are amply discussed in Stidham (2001, \[62\]).

1.2 The Foreground-Background (FB) model

In this thesis we analyse queues that use the so-called FB service discipline. This not very well-known discipline has some appealing features. These will be described in Section 1.4 below. The FB discipline works according to the following rule.

**FB rule:** the customer in the queue that has received the least amount of service is served. If there are \(n\) such customers, for some \(n \in \mathbb{N}\), then they are served simultaneously, i.e., each of them is served at rate \(1/n\).

Let the *age* of a customer be the amount of work he has received. Then a server using the FB discipline always serves the youngest customer(s). As a consequence

\(^1\)It is claimed that ‘queueing’ is the only non-artificial English word containing five subsequent vowels.
customers tend to cluster together in cohorts, groups of customers having the same age.

When a new customer arrives in an FB queue, he is (strictly) the youngest customer in the queue. Hence he is served immediately, see Figure 1 below. If the server stops serving another customer(s) to serve the new customer, then the server is said to have preempted the other customer(s). After that time, basically three things may happen.

1. If the customer needs at least as much service as the age of the customer(s) that was (were) preempted at his arrival, then at a certain time their ages will be equal and they are served together. This happens to the second customer in Figure 1 below. We say that the customer joins a cohort. Later this cohort may join another cohort.

2. The customer may need less service than the age of the customer that he preempted, and hence he leaves the queue before joining the older cohort (see the third customer in Figure 1 below) and the server returns to the cohort that was preempted.

3. Before leaving the queue or joining another cohort the customer may be preempted himself by the arrival of a new customer. This happened to the first customer in Figure 1 below.

![Figure 1](image-url)  
*Figure 1 A realisation of the age process in the FB queue. The (large) full circles denote the departure of a customer.*

By the FB rule, a customer with service time $\tau$ has, throughout his stay in the queue, priority over customers older than $\tau$, since he is always younger than $\tau$. The time
this customer spends in the system\(^2\) is the same as in the queue with service times truncated at level \(\tau\). We call such a queue a \(\tau\)-queue. The concept of the \(\tau\)-queue is used frequently in this thesis.

To further illustrate how the FB model works, let us consider a busy period in the queue with constant service times. Customers arriving in the same busy period cluster together as described above. By the FB priority rule however none of them is allowed to leave the queue before another customer leaves: since service times are constant, the customers with the smallest residual service time have the lowest priority. Hence at the end of the busy period a large cohort, containing all customers who arrived during that busy period, leaves the queue. Kleinrock (1976, [32]) uses this example to emphasise the disastrous effect that using the FB discipline may have.

### 1.3 History of the FB model

In its initial stages in the second half of the sixties, the term FB, or rather FB\(_n\), was used as an abbreviation for both *Foreground-Background* and *Feedback* queueing systems. These different names referred to the same models. See Schrage (1967, [58]), Coffman and Kleinrock (1968, [13]), and the survey article by McKinney (1969, [38]). The FB\(_n\) queue with so-called *quantum size* \(q\) at first was a one-server queue with \(n\) states, or priority classes. This queue operated as follows. A job upon arrival enters the first - or highest priority - state. Within each priority class, the priority of customers depends on their arrival time to that class, in a FIFO manner. Customers are served alone and uninterruptedly for a time period of length \(q\). After the server has completed a customer’s service request in a certain state, one of the customers with the highest priority is selected for service, according to the FIFO rule. If a customer does not leave the queue during his time in the \(k\)th state, he moves to the \(k + 1\)st state - this state has lower priority - and waits until he is served in that state. In the \(n\)th and last state, customers are served uninterruptedly until they leave the system.

The interest in the FB\(_n\) model with \(n\) states and positive quantum size \(q\) faded after a few years. The only model to survive was the limiting case where (first) \(n \to \infty\) and (then) \(q \to 0\). After Kleinrock devoted a section of his *Queueing Systems* (1976, [32]) to this model, the term Foreground-Background (FB) is generally used for the FB\(_{\infty}\) model with quantum size \(q \to 0\). In this thesis the term FB refers to this model.

To distinguish FB from the FB\(_n\) model, some authors prefer to use the term

\(^2\)The words ‘system’ and ‘queue’ are used as synonyms, as are ‘customer’ and ‘job’.
1.4 Why the FB queue?

By means of the service discipline the behaviour of the queueing process may be influenced. There are many ways to choose the service discipline. A queue with a certain discipline may behave well for certain service-time distributions, while for other distributions the queue exhibits less favourable properties. Consider for example the classical FIFO service discipline. This discipline performs very well in case the service times are constant (deterministic) or are bounded random variables. For light-tailed unbounded service times the FIFO discipline still behaves fine. However in case of heavy-tailed service times, the number of customers in the queue is likely to grow very large now and then: imagine what happens to the queue length when a customer with a very large service time is served. In fact the Pollaczek-Khinchin mean value formula, see for instance Kleinrock (1975, [31]), states that the mean of the stationary queue length does not even exist if the second moment of the service time does not exist.

Heavy tails arise in nature, for example, in job lengths in internet traffic, length of telephone calls, etc., see for example Crovella and Bestavros (1996, [18]) and Taqqu et al. (1996, [64]). For such heavy-tailed systems a different scheduling discipline than FIFO is called for, to make sure that in the presence of a large job other jobs do not suffer too much delay.

One candidate is the FB discipline: by the FB priority rule, a customer \( m \), with service time \( x \) say, has priority over customers older than \( x \) that are present in the queue. The service requirements of jobs that arrive during the stay of customer \( m \) are only relevant for him up to the value \( x \): customer \( m \) leaves the queue the moment these new jobs become older than \( x \), if at all. Hence small jobs do not notice the presence of large jobs in their midst and are therefore insensitive to the shape of tail of the service-time distribution. For a number of these heavy-tailed service-time distributions the FB service discipline is optimal in the following sense: it minimises the queue length in certain ways over the class of service disciplines that do not use knowledge of residual service times, see for instance Righter's optimality theorem in
Section 2.2.

Another discipline that limits the influence of large jobs present in the system is the PS (Processor Sharing) discipline. Under the PS discipline all jobs receive an equal share of the server's capacity. Hence if there are \( n \geq 1 \) jobs present, then every job is served at rate \( l/n \). We shall see that for a number of (asymptotic) results the behaviour of FB and PS is comparable.

The overloaded system

Let the random variable \( B \) be the generic service time and denote the rate at which customers arrive, the arrival rate, by \( \lambda \). Then \( \rho = \lambda EB \leq \infty \) is the load of the queue. If \( \rho > 1 \) then on average more work enters the system than the server can manage and the system is called unstable or overloaded. It may be seen that the queue length in unstable queuing systems asymptotically grows linearly in time.

Under the FB discipline a certain part of the customers, namely those with service times that are small enough, are still served in an efficient way. Indeed, by the insensitivity property above, customers with service times up to a certain critical value \( c^* \) do not notice that the load is larger than one. Hence for those customers the busy period is finite a.s. (almost surely) and they are not affected by overloading, see Balkema and Verwijmeren (2000, [6]).

Numerical calculations show that for heavy-tailed distributions with much mass to the left of \( \lambda^{-1} \), the asymptotic growth rate of FB is smaller than that of PS. See Section A.1 for further discussion on this issue and Section 8.5 for a result on the output process in overloaded queues.

1.5 Goal of the thesis

This thesis studies several aspects of the Foreground-Background (FB) queue. We mention the stationary queue length, the maximum queue length in a busy period and over a time interval, the departure process, the cohort process, and the sojourn time. The sojourn time, also known as response time or total waiting time, is the total time a customer spends in the queue, i.e. the time between his arrival and his departure.

Many of the results illustrate the concept that for heavy-tailed service-time distributions the FB discipline performs efficiently. For light-tailed service times, the behaviour is exactly the opposite and the queue under the FB discipline may behave badly, see Kleinrock (1976, [32]).
In this thesis we provide evidence that for certain models the FB discipline is markedly superior to FIFO. Especially for queues with heavy-tailed service times, the FB discipline behaves very efficiently and outperforms the FIFO discipline. If the overhead, caused by monitoring the ages of the customers, is not a problem, then the FB discipline is an efficient scheduling mechanism in case the value of the service times is unknown and the service-time distribution is heavy-tailed.

It is surprising that the FB discipline has received so little attention in the literature\(^3\), since in a certain sense it is the natural counterpart to the FIFO discipline: FB serves the youngest customers, while FIFO gives priority to the oldest customer. This may have to do with the fact that the interest in queues with heavy-tailed characteristics is relatively recent, as well as with the difficulties that arise in the analysis of the FB queue. The literature review in Chapter 2 gives an overview of different types of results obtained for the FB queue. An overview of this type does not exist in the literature. The review article by Yashkov (1992, [69]) on processor-sharing, including FB queues, dates from 1992 and does not mention the optimality results in Section 2.2, or the recent results on the slowdown in Subsection 2.3.3.

1.6 The FB queue: properties and basic notions

In this section we describe some basic notions and properties of the FB queue that are used in this thesis.

**Distributions: heavy tails and log-convex densities**

There is no consensus on when a random variable is heavy-tailed. Some authors require that the variance is infinite, while for others a distribution is heavy-tailed if no exponential moments exist. We adopt this latter notion. For heavy-tailed service times the FB queue may behave better than for light-tailed service times (with the same mean). This counter-intuitive phenomenon appears on a number of occasions in this thesis, see Chapters 4, 6 and 9. In this thesis a special place is reserved for the class of distributions with log-convex densities. For this class the FB discipline exhibits certain optimality properties, as will be discussed in Section 2.2. This class contains heavy-tailed distributions, like the Pareto and (certain) Weibull distributions, as well as distributions that are light-tailed, like the exponential distribution and (certain) gamma distributions.

\(^3\)We estimate that this thesis roughly doubles the number of words on queues with the FB discipline in the literature.
FB as a stochastic version of SRPT and as the opposite of FIFO

If one knows the exact value of the service time of the customer when he enters the system, then one can use this knowledge to minimise the queue length. The service discipline SRPT (Shortest remaining processing time) indeed does so, see Baccelli and Brémaud (2003, [5]). However, precise knowledge about the customer’s (remaining) service time is in general not available to the server. Dealing with this uncertainty, the server may choose, if the probability distribution of the job sizes is known, to serve the customer that has the shortest expected remaining processing time (or residual life). For so-called IMRL (Increasing mean residual life, see Section 2.2) service-time distributions, this is the youngest customer and hence the discipline reduces to the FB discipline.

The server may also favour the customer with the highest failure rate, i.e. the instantaneous departure ‘probability’, cf. Section 2.2. For DFR (Decreasing failure rate) service-time distributions, this policy is again the FB service discipline. Finally, choosing the customer with the smallest likelihood ratio (see again Section 2.2) results in the FB discipline if the service times are from the class of DLR distributions. Hence in a certain sense the FB discipline is the stochastic equivalent of SRPT.

For the opposite classes of DMRL, IFR or ILR distributions, the natural discipline turns out to be FIFO, or any other non-preemptive discipline. A discipline is called non-preemptive if customers are not preempted, i.e. the service of a customer is not interrupted. This again shows that the FB discipline and FIFO are, in a certain sense, opposite disciplines.

Age discrimination

The FB discriminates customers on their age. Small jobs have the server almost to themselves. It may be shown that the ratio of the time spent by a customer in the system and his service time converges to 1 as the size of the customer’s service time converges to zero, see Theorem 2.11. The consequence of this quick service to younger customers is felt by customers with large demands. These customers are mostly served in the otherwise idle time of the system, i.e. when no other customers are present, as is indicated by Theorem 7.2 and the proof of Theorem 7.14.

It is an interesting question how high the price is that large jobs have to pay for the priority given by FB to short jobs. One performance measure is the (mean) slowdown. The slowdown of a job of size \( x \) in the stationary queue is its sojourn time divided by \( x \), see Definition 2.16 below. By simulations Rai et al. (2003, [48])
compare the slowdown under FB, PS, SRPT and FIFO for service times with the so-called high-variability property. For such service-time distributions less than 1% of the jobs accounts for more than half the load. According to recent studies internet traffic has this property, see Crovella and Bestavros (1996, [18]). The simulations show that a very large percentage of the jobs has a significantly smaller slowdown under the FB discipline than under PS or FIFO, and only a negligible part of the jobs has a larger slowdown.

Criteria to measure performance

There are many ways to measure the performance of a queuing system. Usually, the behaviour of only one characteristic of the queuing process is considered, for example the queue length, the maximum queue length, the vector of ordered residual service times, the sojourn time or the slowdown. See for instance Chapter 4 of Baccelli and Brémaud (2003, [5]), Harchol-Balter and Wierman (2003, [25]), Chapter 5 of Stoyan (1983, [63]), and Chapters 2 and 7 of this thesis.

Measuring the performance of queues by analysing just one characteristic is subjective, since good behaviour of one characteristic does not necessarily imply good behaviour of other characteristics. Furthermore, there may be a large difference in the performance of the queue experienced by small jobs and by customers with a large demand. Finally, there exists an infinite number of more or less exotic ways to judge the performance of a queue. For instance in the example at the beginning of this chapter, one could give more credit for solving difficult problems quickly. Another criterion is ‘the closer the queue length is to 7 at time 23, the better’. This poses the question: what is a service discipline? Although the answer is intuitively clear, no answer to this question has been published so far.

In this thesis we often use the queue length measured in the number of customers present in the system as a measure to compare the performance of queues (see Sections 3 to 9 and Section 10). The queue with the smaller queue length is then said to be better. The queue length is an important characteristic when designing buffer sizes for finite buffer systems. Little’s law states that in a stationary queue the mean queue length $Q$ is related to the mean of the sojourn time $V$ by $EQ = \lambda EV$, where $\lambda$ is the arrival intensity. Hence minimising $EQ$ and minimising $EV$ are equivalent.

Random quantities like the length of the first busy period, or the workload, at time zero in a stationary queue do not depend on the discipline and will therefore not play an important role in this thesis.
Analysis of the FB queue

The FB queue with exponentially distributed interarrival times is a Markov process whose state is described by the sizes and ages of the different cohorts. In general all customers have a positive age. Analysis often turns out to be more difficult than in non-preemptive queues (like FIFO) where at most one customer has positive age, or than in queues where the service rate only depends on the number of customers in the system, like the PS queue. Usually, we have to be satisfied with lower and upper bounds of the probabilities under consideration. An exception is the sojourn time, which, by the notion of the $\tau$-queue described above, is relatively easy to handle, see for example Chapter 7.

Let us conclude this section with a very short description of the type of results that the reader will encounter in this thesis. Part of this thesis consists of calculating characteristics in a quantitative way, either in an exact form (Chapters 4 and 7), through upper or lower bounds (Chapters 3 and 8) or by showing finiteness or infiniteness (Chapter 9). Another part consists of a qualitative analysis of the FB queue. We compare the behaviour of characteristics of the queue under different service-time distributions in Chapters 4 and 6.

1.7 Short summary of used notation and vocabulary

To describe the type of queue under consideration we use the following shorthand version of Kendall's notation. A queue is referred to by an expression of the form A/B/1 C, where A and B stand for the distributions of the interarrival times and service times, and C denotes the service discipline, such as FB, PS, FIFO and LIFO, which is written in sans-serif capitals. In this thesis A and B take the shape of M (exponential distribution), D (deterministic distribution) and G (general, meaning unspecified distribution). Unless stated otherwise we assume that the interarrival times are i.i.d. and independent of the service times, which are also i.i.d. Hence the M/G/1 queue is determined by $\lambda$ and $F$, where $\lambda > 0$ is the arrival rate and $F$ is the distribution of the service times.

A short inspection of standard books and papers on queueing theory shows that the capitals $L, S, T, V$ and $W$ are used with very different meanings. To avoid confusion, the next table shows the meaning of the random variables used most frequently in this thesis. For a complete list of notation we refer to the List of
notation on page 155.

$B$ the generic service time.
$B_n$ the service time of the $n$th customer: an independent copy of $B$
$F$ the distribution function of $B$ (and of $B_n$)
$Q$ the stationary queue length (in number of customers present)
$Q(t)$ the queue length at time $t$
$V$ the sojourn time in the stationary queue
$V(x)$ the sojourn time of a customer with service time $x$ in the stationary queue
$\lambda$ the rate of the Poisson arrival process
$\rho$ the load or the traffic intensity of the system, given by $\rho = \lambda E B$
$\rho(x)$ the load of the $x$-queue (with generic service time $B \wedge x$): $\rho(x) = \lambda E (B \wedge x)$
$\mathbb{N}$ the set $\{1, 2, \ldots\}$
$a \wedge b$ $\min\{a, b\}$

By $\mathcal{D}$ we denote the class of service disciplines that do not use knowledge of the residual service times. Unless specified otherwise, we consider only work-conserving (or non-idling) disciplines, i.e. the server is not idle when there are customers in the system. Unless specified otherwise, we assume that the queue is stable, $0 < \rho < 1$.

1.8 Outline of the thesis

In the second chapter we give an overview of the results on the FB model obtained in the literature so far. The results are of two types. First we describe the optimality of FB: for certain service-time distributions the FB discipline minimises the queue length in a certain way. Secondly we describe the results on characteristics of the M/G/1 FB queue that are known in the literature, like the generating function of the stationary queue length, the sojourn time conditioned on the job size, and recent results on the slowdown.

In the third chapter we show that for log-convex service times the tail of the maximum queue length in the busy period in the M/G/1 FB queue is bounded by an exponential tail. This result yields bounds for the maximum queue length over a time interval. Using these bounds, the probability of the overflow of a finite buffer is studied.

In chapters four, five and six we examine the impact of an increase in the variability of the service-time distribution on several characteristics of the queue. To measure the variability of a random variable we use the convex order, defined in
Definition 4.1. It turns out that a higher variability in the service-time distribution often, but not always, leads to better behaviour of the FB queue.

A toy model is studied in the fourth chapter. We compare the maximum queue length in a busy period in two queues. In one queue the service times are deterministic, in the other queue the service times have the same expectation, but the customers either leave immediately after their service has started, or have service time $1 + c$ for some $c > 0$. This is the simplest example of two service times that are *convexly ordered*. It is shown that in this case the more variable distribution yields a stochastically smaller maximum queue length. We conclude with a sketch how to simulate the maximum queue length in the FB queue.

In chapter five the convex ordering of service times is studied for other service disciplines. We study the impact of a more variable service time on the distribution of the maximum queue length in busy periods in M/G/1 queues, but now with disciplines other than FB. The considered disciplines are the LIFO discipline (both preemptive resume and repeat) and non-preemptive disciplines. For these disciplines we show that a more variable service time yields a smaller expected maximum queue length.

In chapter six we return to the FB queue and describe the impact of a more variable service-time distribution on the mean queue length. For two queues with service-time distributions from two specific classes we show the following. The service times from the first class are more variable than the service times from the second class, but yield a smaller mean queue length. In the second part of the chapter we show this does not hold in general, adding variability to an arbitrary service-time distribution $F$ does not necessarily result in a smaller value of $EQ$.

In chapter seven we study the sojourn time of a customer in the stationary M/G/1 FB queue. For light-tailed service times we show that the (asymptotic) decay rate of the tail of the sojourn time is as small as possible, namely equal to the decay rate of the busy period. A smaller decay rate corresponds to a larger tail probability. On the other hand, for a class of heavy-tailed service times Núñez Queija (2000, [41]) has shown that the sojourn time and the service-time distributions in the M/G/1 FB have the same tail behaviour. In the second part of this chapter we relax the conditions of this theorem.

The departure process of the FB queue is the subject of chapter eight. We give an exponential upper and lower bound for the tail of the departure-time distribution in the stationary queue and generalise the method to the unstable queue. For M/D/1 queues the bounds are sharp.

Chapter nine features a number of related results on the stationary queue length.
Q. asymptotics of the rate $Q(t)/t$ in the unstable queue and the stationary cohort process. We prove the existence of all moments of $Q$ under a (weak) condition on the service-time distribution and derive limiting behaviour of $EQ$ as the load increases to 1 (also known as heavy traffic). It turns out that the behaviour of the mean queue length is different for service-time distributions with a finite end-point, than for those with infinite end-points. Furthermore we study the cohort process in the stationary queue. The chapter concludes with some results for the asymptotic growth of the queue length in unstable queues under several disciplines.

In chapter ten we give a theoretical description of the single server queue, including a precise definition of the service discipline. By means of a game we show that the deterministic queue is not interesting from a scheduling point of view. In the third part of the chapter it is shown that a continuous time queue may be approximated by discrete time queues. This last result is combined with a limiting argument to validate optimality properties, for example those in chapter two, so far only obtained for discrete time queues.

**Final remarks** The lion’s share of Chapter 3 is formed by the paper *The maximum queue length for heavy-tailed service times in the M/G/1 FB queue* that will appear in *Queueing Systems*. The paper *Sojourn times in the M/G/1 FB queue with light-tailed service times* [36] appears in this thesis as the first part of Chapter 7. Chapter 5 is based on the paper *The effect of service-time variability on maximum queue lengths in M/G/1 queues* [43].

Apart from this introduction and to a lesser extent Chapter 2, the chapters in this thesis do not depend heavily on each other and may be read in an arbitrary sequence. Cross-references indicate relationships between results in different chapters. In this thesis we assume that the reader is familiar with the basic concepts of queueing theory that are listed in Section 1.7. New concepts are explained when considered.