ANALYSIS OF THE QUEUE LENGTHS IN A PRIORITY RETRIAL QUEUE WITH CONSTANT RETRIAL POLICY

ARNAUD DEVOS*
Department of Telecommunications and Information Processing
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

JORIS WALRAEVEN
Department of Telecommunications and Information Processing
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

TUAN PHUNG-DUC
Division of Policy and Planning Sciences
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

HERWIG BRUNEEL
Department of Telecommunications and Information Processing
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

(Communicated by Wuyi Yue)

Abstract. In this paper, we analyze a priority queueing system with a regular queue and an orbit. Customers in the regular queue have priority over the customers in the orbit. Only the first customer in the orbit (if any) retries to get access to the server, if the queue and server are empty (constant retrial policy). In contrast with existing literature, we assume different service time distributions for the high- and low-priority customers. Closed-form expressions are obtained for the probability generating functions of the number of customers in the queue and orbit, in steady-state. Another contribution is the extensive singularity analysis of these probability generating functions to obtain the stationary (asymptotic) probability mass functions of the queue and orbit lengths. Influences of the service times and the retrial policy are illustrated by means of some numerical examples.

1. Introduction. In queueing systems, it is not uncommon that some customers who find the server busy upon arrival leave the system, but return after some random time to request for service again. This class of customers can be visualised as being located in a virtual waiting room, hereafter called the orbit [11]. In queueing theory, such queueing systems are called retrial queues. There exists a vast literature on the theory of retrial queues [3]. Perhaps the most well known books are those of [4, 10], which can serve as a solid introduction to retrial queues. In the literature, two types of retrial policies are typically identified. The first one, called the classical retrial policy, assumes that all the customers in the orbit retry to get access to the

2020 Mathematics Subject Classification. Primary: 60K25, 90B22; Secondary: 68M20.
Key words and phrases. Queueing theory, priority, retrials, constant retrial policy, probability generating functions, singularity analysis.

* Corresponding author: Arnaud Devos.
server. For analytical feasibility reasons, these retrials are typically modelled as independent Poisson processes with a common arrival rate. An excellent survey of these models can be found in [8]. The second policy, referred to as the constant retrial policy, assumes that only one customer of the orbit can retry for service. Therefore, in the constant retrial policy, the orbit is commonly presented as a FCFS queue. In contrast to the classical retrial policy, the retrial times in the constant retrial policy are independent of the orbit length. This property may be important when modeling practical applications, as illustrated in [20, 21] where the principle of call blending in call centres is modelled as queues with a constant retrial policy.

The constant retrial policy was first introduced in [12], where a telephone exchange model with only one line is examined. In [12], blocked customers (incoming calls) form a feedback queue (this is the orbit) and only the customer at the head of the feedback queue can retry for service after an exponentially distributed retrial time. The model presented in [12], but with a general retrial time distribution is studied in [7]. Finally, in [16] the constant retrial policy with general retrial times and general service times is extensively studied. The author has obtained the probability generating function of the orbit length and the Laplace-Stieltjes transform of the waiting time, in steady-state. The most important feature of a retrial queue with constant retrial policy and general retrial times is that, after each service completion,\footnote{Remark that for exponential retrial times, it is not necessary to know exactly when the retrial has started.} a competition between an exponential law (the arrivals) and a general retrial time determines the next customer who receives service. Retrial queues with constant retrial policy and general retrial times are therefore related to many queueing systems where the server spends some random amount of time to find the next customer to serve after each service completion.

If retrial queues also possess a regular queue, priority occurs naturally. Customers in the regular queue have priority over the customers in the orbit. This means that the customer at the head of the orbit can only start retrying when both the regular queue and the server becomes empty. If another customer arrives during a retrial time, this customer is served and the retrial has to start over when the regular queue becomes empty again. The priority can in principal be non-preemptive or preemptive, the non-preemptive again being the most natural one. In [15], a preemptive constant retrial policy is studied. The author has obtained the probability generating functions of the stationary queue and orbit length, and the Laplace-Stieltjes transforms of the waiting time of both types of customers. In [5], a non-preemptive constant retrial policy is studied. The probability generating functions of the stationary queue and orbit lengths are obtained. An interesting result from this paper, is that the distribution of the priority queue length is independent of the retrial time distribution. While in [15] it is assumed that the service time distributions of the customers in the priority queue and customers in the orbit are different, this is not the case in [5]. However, there might be situations where the service time distributions are not necessarily the same. It might even be the reason why a customer joins either the queue or the orbit. For instance, customers with a small (non-urgent) service time who face a queue upon arrival, are (in certain situations) more likely to join the orbit. In [9], we generalized the model of [5] in two ways. We assumed different service time distributions between both classes of customers. Further, in contrast to [5], we analyzed the waiting times of both classes of customers.
Our present paper is an extended version of [9]. In [9], we obtained explicit expressions for the probability generating functions of the stationary queue and orbit length. Theoretically, these contain all information of the queue and orbit lengths. For example, we can compute relatively easy the mean and the variance of the queue and orbit lengths from their probability generating functions. However, obtaining the distribution from the probability generating function is a much more delicate task. If we want to know the probability that the orbit length equals \( n \) customers, we have to compute the \( n \)-th derivative of the probability generating function and evaluate at zero. This becomes quickly problematic for large \( n \). An elegant way to approximate these probabilities is to asymptotically invert the probability generating function, resulting into explicit expressions for the probability mass function [14], which are accurate for large \( n \). Asymptotics for the M/G/1 retrial queue, for the classical retrial policy, is studied for light-tailed service time distributions in [17] and for a subclass of heavy-tailed service time distributions in [22]. The asymptotics of the marginal and joint queue lengths in the (non-preemptive) M/M/1 priority queue is studied in [18]. We extend our analysis of [9] to include asymptotics of the queue and orbit lengths, for several different types of service time distributions (e.g. light-tailed, fat-tailed, \ldots). Since we assume different service time distributions between customer classes, we have to examine several possible combinations of types of distributions. By obtaining these asymptotics, we now demonstrate a complete characterisation of the queue and orbit lengths of this queueing model. Our methodology is systematic in the sense that we use the same singularity analysis of the probability generating functions to obtain the asymptotics for all combinations of the service time distributions. This is the novelty and difference in comparison with previous studies in which adhoc methods were developed for specific types of service time distributions [17, 22].

The paper is organized as follows. In Section 2, we provide a detailed description of the queueing model under consideration. Next, in Section 3 we discuss the stability condition and analyze the steady-state distributions using the supplementary variable technique. In Section 4 we give an overview of singularity analysis, which is the method we apply in Section 5 and Section 6 to analyze the asymptotic distributions of the queue and orbit lengths respectively. Section 7 discusses the most important performance measures of the queueing system. Finally, we discuss some numerical examples in Section 8.

2. Mathematical model. We assume a queueing system with an infinite-sized queue, an infinite-sized orbit, and one server. Class-1 and class-2 customers arrive to the system according to two independent Poisson processes with rate \( \lambda_1 \) and \( \lambda_2 \) respectively. If a new customer arrives to the system and the server is free, he gets service immediately. If the server is busy upon arrival, a class-1 customer waits in the queue, while a class-2 customer joins the orbit. We assume that only the first customer in the orbit retries to get access to the server when the queue and server are idle. Retrial times are characterized by means of i.i.d. continuous random variables with common cumulative distribution (cdf) \( R(x) \) and common Laplace-Stieltjes transform (LST) \( R^*(s) \). Service times of class-\( j \) customers, \( j = 1, 2 \), are specified by i.i.d. continuous random variables with cdf \( B_j(x) \), LST \( B^*_j(s) \) and mean \( \beta_j \). We define the arrival load as \( \rho \triangleq \lambda_1 \beta_1 + \lambda_2 \beta_2 \). Further, we assume that service times, retrial times and inter-arrival times are all mutually independent.
In the rest of this section we fix notations for the rest of the paper. We use the notation $\mathbb{P}$ for the probability measure, $\mathbb{E}$ for the expectation operator and $\mathbb{I}$ for the indicator function.

Let $N_1(t)$ be the number of class-1 customers in the priority queue at time $t$, not counting the one in the server. Similarly, let $N_2(t)$ the number of class-2 customers in the orbit at time $t$. Further, we add four auxiliary processes. First, $C(t)$ summarizes the state of the system, more precisely $C(t) = 0$ when the server is idle at time $t$, $C(t) = 1$ when the server is serving a class-1 customer at time $t$ and $C(t) = 2$ when the server is serving a class-2 customer at time $t$. When $C(t) = 0$ and $N_2(t) > 0$, $\xi_0(t)$ represents the elapsed retrial time at time $t$; when $C(t) = 1$, $\xi_1(t)$ denotes the elapsed class-1 service time at time $t$; when $C(t) = 2$, $\xi_2(t)$ denotes the elapsed class-2 service time at time $t$. The stochastic vector $\{C(t), N_1(t), N_2(t), \xi_0(t), \xi_1(t), \xi_2(t)\}$ is then a Markov Process. We assume that the stationary distribution of this stochastic vector exists. Therefore we define the following limiting probabilities:

\[
p_0 = \lim_{t \to \infty} \mathbb{P}[C(t) = N_1(t) = N_2(t) = 0],
\]
\[
p_{i,n}^{(0)}(x)dx = \lim_{t \to \infty} \mathbb{P}[C(t) = 0, N_1(t) = 0, N_2(t) = n, x < \xi_0(t) \leq x + dx], n \geq 1
\]
\[
p_{i,n}^{(1)}(x)dx = \lim_{t \to \infty} \mathbb{P}[C(t) = 1, N_1(t) = i, N_2(t) = n, x < \xi_1(t) \leq x + dx], i, n \geq 0
\]
\[
p_{i,n}^{(2)}(x)dx = \lim_{t \to \infty} \mathbb{P}[C(t) = 2, N_1(t) = i, N_2(t) = n, x < \xi_2(t) \leq x + dx], i, n \geq 0.
\]

We also define the following probability generating functions (pgf)

\[
P_0(x, z_2) = \sum_{n=1}^{\infty} p_{0,n}^{(0)}(x) z_2^n,
\]
\[
P_1(x, z_1, z_2) = \sum_{i=0}^{\infty} \sum_{n=0}^{\infty} p_{i,n}^{(1)}(x) z_1^i z_2^n,
\]
\[
P_2(x, z_1, z_2) = \sum_{i=0}^{\infty} \sum_{n=0}^{\infty} p_{i,n}^{(2)}(x) z_1^i z_2^n.
\]

Next, we define $\bar{r}(x)$ as the probability density function (pdf) of the retrial time $R$, given that the retrial time is at least $x$. We define $b_j(x)$ as the pdf of the service time of a class-$j$ customer, given that the service time is at least $x$. We thus have

\[
\bar{r}(x) = \frac{R'(x)}{1 - R(x)} \quad \text{and} \quad \bar{b}_j(x) = \frac{B_j'(x)}{1 - B_j(x)}.
\]

Let $\lambda \triangleq \lambda_1 + \lambda_2$ the total arrival rate to the system. Let $p_1$ and $p_2$ the probability that a customer is of class-1 and of class-2 respectively. These probabilities are given by $p_j = \lambda_j/\lambda$, $j = 1, 2$.

For further use, we define $A_j(z_1, z_2)$ as the joint pgf of the class-1 and class-2 arrivals during a class-$j$ service time and $A(z_1, z_2)$ as the joint pgf of the class-1 and class-2 arrivals during a randomly chosen service time. They are calculated as

\[
A_j(z_1, z_2) = B_j^r(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)),
\]
\[
A(z_1, z_2) = p_1 A_1(z_1, z_2) + p_2 A_2(z_1, z_2).
\]
In particular, we are interested in the steady-state distributions of the orbit and queue lengths. We define
\[ Q(z) = p_0 + \int_0^\infty P_0(x,z)dx, \]
as the partial pgf of the number of customers in the orbit when the server is idle, and
\[ U_j(z_1,z_2) = \int_0^\infty P_j(x,z_1,z_2)dx, \quad j = 1, 2, \]
as the partial joint pgf of the number of customers in queue and orbit when the server is serving a type-\( j \) customer.

3. Queuing analysis. In this section, we use the supplementary variable technique to derive expressions for the steady-state (partials) pgfs of the number of customers in the orbit when the server is idle and of the number of customers in queue and orbit when the server is serving a type-\( j \) customer, \( j = 1, 2 \). First, we indicate the stability condition of the queueing system.

3.1. Stability condition. A stability condition could be derived by studying the embedded Markov Chain at departure epochs, similarly as in [5]. It can be shown that this yields the same condition as in [5], namely
\[ \rho < \frac{R^*(\lambda)}{p_1R^*(\lambda) + p_2}. \]
Due to the competition of two exponentials with rate \( \lambda_j, \ j = 1, 2 \) and a general retrial time, \( R^*(\lambda) \) is the probability that there are no new arrivals during a retrial time and is therefore also the probability that a single retrial is successful. We can also rewrite (9) as
\[ p_2\rho < R^*(\lambda)(1 - p_1\rho), \]
to which we will now give an intuitive explanation. From (6) it follows that \( p_2\rho \) is the mean number of class-2 arrivals during a randomly chosen service time. In stability, the probability that the regular queue is empty after a departure instant is equal to \( 1 - p_1\rho \). Notice now that between two departure instants, a customer from the orbit (if any) can only have entered the server if at the first departure instant the regular queue was left empty and his retrial was successful. Since only the first customer from the orbit can try to get access to the server, \( R^*(\lambda)(1 - p_1\rho) \) is the expected number of customers from the orbit entering the server between two departure instants. Condition (10) can therefore be interpret as follows: the average number of customers entering the orbit during a randomly chosen service time must be strictly less then the average number of customers from the orbit entering the server between two departure instants.

3.2. Steady-state distributions. If the regular queue is non-empty, a customer from that queue is served next. Otherwise, a customer from the orbit (if any) retries to access the server. If his retrial time finishes before a new arrival occurs, he is served. Otherwise, the newly arriving customer starts service, irrespective of his type, and the customer at the head of the orbit stops retrying until the server and the priority queue are empty again. This leads to the following equilibrium equations
In terms of the pgfs, the boundary conditions can be written as

\[ \lambda p_0 = \int_0^\infty p_{0,0}^{(1)}(x)\tilde{b}_1(x)dx + \int_0^\infty p_{0,0}^{(2)}(x)\tilde{b}_2(x)dx , \]  
\[ \frac{\partial}{\partial x} p_{0,n}^{(0)}(x) = -\left(\lambda + \bar{r}(x)\right)p_{0,n}^{(0)}(x) , \quad n \geq 1 , \]  
\[ \frac{\partial}{\partial x} p_{1,n}^{(1)}(x) = -\left(\lambda + \bar{b}_1(x)\right)p_{1,n}^{(1)}(x) + \lambda_1 p_{1-1,n}^{(1)}(x)\mathbb{1}(i \geq 1) \]  
\[ + \lambda_2 p_{1,n-1}^{(1)}(x)\mathbb{1}(n \geq 1) , \quad i \geq 0 , n \geq 0 , \]  
\[ \frac{\partial}{\partial x} p_{1,n}^{(2)}(x) = -\left(\lambda + \bar{b}_2(x)\right)p_{1,n}^{(2)}(x) + \lambda_1 p_{1-1,n}^{(2)}(x)\mathbb{1}(i \geq 1) \]  
\[ + \lambda_2 p_{1,n-1}^{(2)}(x)\mathbb{1}(n \geq 1) , \quad i \geq 0 , n \geq 0 . \]

The boundary conditions are

\[ p_{0,n}^{(0)}(0) = \int_0^\infty p_{0,0}^{(1)}(x)\tilde{b}_1(x)dx + \int_0^\infty p_{0,0}^{(2)}(x)\tilde{b}_2(x)dx , \quad n \geq 1 , \]  
\[ p_{1,n}^{(1)}(0) = \mathbb{1}(i = 0, n = 0)\lambda_1 p_0 + \mathbb{1}(i = 0, n > 0)\lambda_1 \int_0^\infty p_{0,n}^{(0)}(x)dx \]  
\[ + \int_0^\infty p_{1+1,n}^{(1)}(x)\tilde{b}_1(x)dx + \int_0^\infty p_{1+1,n}^{(2)}(x)\tilde{b}_2(x)dx , \quad i \geq 0 , n \geq 0 , \]  
\[ p_{1,n}^{(2)}(0) = \mathbb{1}(i = 0, n = 0)\lambda_2 p_0 + \mathbb{1}(i = 0, n > 0)\lambda_2 \int_0^\infty p_{0,n}^{(0)}(x)dx \]  
\[ + \mathbb{1}(i = 0)\int_0^\infty p_{0,n+1}^{(0)}(x)\bar{r}(x)dx , \quad i \geq 0 , n \geq 0 . \]

Assuming stationarity, the equilibrium Equations (12), (13) and (14) are transformed to the following partial differential equations for the pgfs \( P_0(x,z) \), \( P_1(x,z_1, z_2) \) and \( P_2(x,z_1, z_2) \):

\[ \frac{\partial}{\partial x} P_0(x,z_2) = -\left(\lambda + \bar{r}(x)\right)P_0(x,z_2) , \]  
\[ \frac{\partial}{\partial x} P_1(x,z_1, z_2) = -\left(\lambda + \bar{b}_1(x) - \lambda_1 z_1 - \lambda_2 z_2\right)P_1(x,z_1, z_2) , \]  
\[ \frac{\partial}{\partial x} P_2(x,z_1, z_2) = -\left(\lambda + \bar{b}_2(x) - \lambda_1 z_1 - \lambda_2 z_2\right)P_2(x,z_1, z_2) , \]

whose solutions are given by

\[ P_0(x,z_2) = P_0(0,z_2)[1 - R(x)]e^{-\lambda x} , \]  
\[ P_1(x,z_1, z_2) = P_1(0,z_1, z_2)[1 - B_1(x)]e^{-\left(\lambda_1(1-z_1) + \lambda_2(1-z_2)\right)x} , \]  
\[ P_2(x,z_1, z_2) = P_2(0,z_1, z_2)[1 - B_2(x)]e^{-\left(\lambda_1(1-z_1) + \lambda_2(1-z_2)\right)x} . \]

In terms of the pgfs, the boundary conditions can be written as
where we also used (11) in the first equation.

If we substitute (18), (19) and (20) in the previous equations and by using (5), we get

\[
P_0(0, z_2) = \int_0^\infty P_1(x, 0, z_2) b_1(x) dx + \int_0^\infty P_2(x, 0, z_2) b_2(x) dx - \lambda p_0 , \tag{21}
\]

\[
P_1(0, z_1, z_2) = \lambda_1 p_0 + \lambda_1 \int_0^\infty P_0(x, z_2) dx
+ \frac{1}{z_1} \int_0^\infty [P_1(x, z_1, z_2) - P_1(x, 0, z_2)] b_1(x) dx
+ \frac{1}{z_1} \int_0^\infty [P_2(x, z_1, z_2) - P_2(x, 0, z_2)] b_2(x) dx , \tag{22}
\]

\[
P_2(0, z_1, z_2) = \lambda_2 p_0 + \lambda_2 \int_0^\infty P_0(x, z_2) dx + \frac{1}{z_2} \int_0^\infty P_0(x, z_2) f(x) dx , \tag{23}
\]

where we also used (11) in the first equation.

If we substitute (18), (19) and (20) in the previous equations and by using (5), we get

\[
P_0(0, z_2) = P_1(0, 0, z_2) A_1(0, z_2) + P_2(0, 0, z_2) A_2(0, z_2) - \lambda p_0 , \tag{24}
\]

\[
P_1(0, z_1, z_2) = \lambda_1 p_0 + p_1 (1 - R^*(\lambda)) P_0(0, z_2) + \frac{1}{z_1} P_1(0, z_1, z_2) A_1(z_1, z_2)
- \frac{1}{z_1} P_1(0, 0, z_2) A_1(0, z_2) + \frac{1}{z_1} P_2(0, z_1, z_2) A_2(z_1, z_2)
- \frac{1}{z_1} P_2(0, 0, z_2) A_2(0, z_2) , \tag{25}
\]

\[
P_2(0, z_1, z_2) = \lambda_2 p_0 + \frac{1}{z_2} [p_2 z_2 (1 - R^*(\lambda)) + R^*(\lambda)] P_0(0, z_2) . \tag{26}
\]

We are left with five unknowns: \( p_0, P_0(0, z_2), P_1(0, z_1, z_2), P_1(0, 0, z_2) \) and \( P_2(0, 0, z_2) \).\(^2\) First, we substitute (26) in (24) and solve for \( P_0(0, z_2) \),

\[
P_0(0, z_2) = \frac{z_2 A_1(0, z_2) P_1(0, 0, z_2) + (p_2 A_2(0, z_2) - 1) z_2 \lambda p_0}{z_2 - [p_2 z_2 + R^*(\lambda) (1 - p_2 z_2)] A_2(0, z_2)} . \tag{27}
\]

Substituting (26) in (25) and solving to \( P_1(0, z_1, z_2) \), we can write

\[
z_2 [z_1 - A_1(z_1, z_2)] P_1(0, z_1, z_2) = \{z_1 p_1 + p_2 A_2(z_1, z_2) - p_2 A_2(0, z_2)\}z_2 \lambda p_0
+ \{z_2 (1 - R^*(\lambda)) [p_1 z_1 + p_2 A_2(z_1, z_2) - p_2 A_2(0, z_2)]
+ R^*(\lambda) [A_2(z_1, z_2) - A_2(0, z_2)]\} P_0(0, z_2) - z_2 A_1(0, z_2) P_1(0, 0, z_2) . \tag{28}
\]

It is well-known that \( f(z_1) := z_1 - A_1(z_1, z_2) \) has one zero inside the complex unit disk for each \( z_2 \) with \( |z_2| < 1 \), see for instance [25]. Denote this zero by \( Y(z_2) \). It can be proven that this is an analytic function inside the complex unit disk and that \( Y(1) = 1 \), at least when the queueing system is stable. For more details, we refer to Appendix A in [23]. By substituting \( z_1 \) by \( Y(z_2) \) in (28), we find

\[
z_2 A_1(0, z_2) P_1(0, 0, z_2) = \{A(Y(z_2), z_2) - p_2 A_2(0, z_2)\}z_2 \lambda p_0
+ \{z_2 (1 - R^*(\lambda)) [A(Y(z_2), z_2) - p_2 A_2(0, z_2)]
+ R^*(\lambda) [A_2(Y(z_2), z_2) - A_2(0, z_2)]\} P_0(0, z_2) . \tag{29}
\]

\(^2\)Note from Equation (26) that \( P_2(0, z_1, z_2) \) is independent of \( z_1 \).
Substituting the above equation into (27) yields

\[ P_0(0, z_2) = \frac{[A(Y(z_2), z_2) - 1]z_2\lambda p_0}{z_2^2 - z_2(1 - R^*(\lambda))A(Y(z_2), z_2) - R^*(\lambda)A_2(Y(z_2), z_2)} . \] (30)

Substituting Equation (30) back into (27) and solving for \( P_1(0, 0, z_2) \) yields

\[ P_1(0, 0, z_2) = \frac{R^*(\lambda)\lambda p_0}{z_2 - z_2(1 - R^*(\lambda))A(Y(z_2), z_2) - R^*(\lambda)A_2(Y(z_2), z_2)} \times \frac{[1 - A(Y(z_2), z_2)][A_2(0, z_2) - z_2 + [p_2A_2(0, z_2) - 1][A_2(Y(z_2), z_2) - z_2]}{A_1(0, z_2)} . \] (31)

Inserting (30) and (31) into (28) gives

\[ P_1(0, z_1, z_2) = \frac{R^*(\lambda)\lambda p_0}{z_2 - z_2(1 - R^*(\lambda))A(Y(z_2), z_2) - R^*(\lambda)A_2(Y(z_2), z_2)} \times \left\{ \frac{[1 - p_1z_1 - p_2A_2(0, z_2)]|A_2(Y(z_2), z_2) - z_2|}{z_1 - A_1(z_1, z_2)} + \frac{[A_2(z_1, z_2) - z_2][A(Y(z_2), z_2) - 1]}{z_1 - A_1(z_1, z_2)} \right\} . \] (32)

The forth unknown \( P_2(0, z_1, z_2) \) is given as

\[ P_2(0, z_1, z_2) = \frac{[p_2z_2 + p_1Y(z_2) - 1]R^*(\lambda)\lambda p_0}{z_2 - z_2(1 - R^*(\lambda))A(Y(z_2), z_2) - R^*(\lambda)A_2(Y(z_2), z_2)} . \] (33)

Finally, the last unknown \( p_0 \) can be found from the normalisation condition \( Q(1) + U_1(1, 1) + U_2(1, 1) = 1 \). Using the Definitions (7), (8) and the Expressions (18), (19) and (20), one easily finds

\[ Q(z_2) = p_0 + \frac{1 - R^*(\lambda)}{\lambda}P_0(0, z_2) , \]

\[ U_j(z_1, z_2) = \frac{1 - A_j(z_1, z_2)}{\lambda_1(1 - z_1) + \lambda_2(1 - z_2)}P_j(0, z_1, z_2), \quad j = 1, 2 . \]

It now follows that

\[ p_0 = 1 - \frac{R^*(\lambda)p_1 + p_2}{R^*(\lambda)}\rho . \] (34)

Remark that the stability condition implies that \( p_0 > 0 \), i.e. the total system will occasionally be idle.

Summarized, we have obtained partial pgfs of the number of customers in the queue and orbit when the server is either busy or idle. The complete expressions
are given by:

\[ Q(z) = \frac{z - A_2(Y(z), z)}{z - z(1 - R^*(\lambda))A(Y(z), z) - R^*(\lambda)A_2(Y(z), z)} \]
\[ \times \left(1 - \frac{R^*(\lambda)p_1 + p_2}{R^*(\lambda)}\right) R^*(\lambda), \tag{35} \]

\[ U_1(z_1, z_2) = \left\{ \begin{array}{ll}
\frac{[1 - p_1 z_1 - p_2 A_2(z_1, z_2)] [A_2(Y(z_2), z_2) - z_2]}{z_2 - z_2(1 - R^*(\lambda))A(Y(z_2), z_2) - R^*(\lambda)A_2(Y(z_2), z_2)} \\
+ \frac{[A_2(z_1, z_2) - z_2][A(Y(z_2), z_2) - 1]}{z_2 - z_2(1 - R^*(\lambda))A(Y(z_2), z_2) - R^*(\lambda)A_2(Y(z_2), z_2)}
\end{array} \right. \]
\[ \times \left(1 - \frac{1}{z_1 - A_1(z_1, z_2)}\right) \frac{1}{1 - p_1 z_1 - p_2 z_2} \left(1 - \frac{R^*(\lambda)p_1 + p_2}{R^*(\lambda)}\right) R^*(\lambda), \tag{36} \]

\[ U_2(z_1, z_2) = \frac{p_2 z_2 + p_1 Y(z_2) - 1}{z_2 - z_2(1 - R^*(\lambda))A(Y(z_2), z_2) - R^*(\lambda)A_2(Y(z_2), z_2)} \]
\[ \times \frac{1 - A_2(z_1, z_2)}{1 - p_1 z_1 - p_2 z_2} \left(1 - \frac{R^*(\lambda)p_1 + p_2}{R^*(\lambda)}\right) R^*(\lambda). \tag{37} \]

From these three important pgfs, the marginal pgfs of the numbers of customers in
the queue and orbit are easily deduced. The pgf \( N_1(z) \) of \( N_1 \) is given by

\[ N_1(z) = Q(1) + U_1(z, 1) + U_2(z, 1) \]
\[ = \frac{1 - p_1 \rho - p_1 (1 - \rho)z - p_2 A_2(z, 1)}{p_1 [A_1(z, 1) - z]} . \tag{38} \]

The pgf of \( N_1 \) is thus independent of the distribution of the retrial times. As
mentioned before, this was also observed in [5]. The pgf \( N_2(z) \) of \( N_2 \) is given by

\[ N_2(z) = Q(z) + U_1(1, z) + U_2(1, z) \]
\[ = \frac{p_1 Y(z) + p_2 z - 1}{z - z(1 - R^*(\lambda))A(Y(z), z) - R^*(\lambda)A_2(Y(z), z)} \]
\[ \times \frac{1}{p_2} \left(1 - \frac{R^*(\lambda)p_1 + p_2}{R^*(\lambda)}\right) R^*(\lambda). \tag{39} \]

4. **Singularity analysis of a probability generating function.** Now we have
obtained the stationary pgf of the queue and orbit length, we are interested in deriving
asymptotic expressions for the stationary queue and orbit length distribution. This section gives a brief overview of the method we apply to obtain such asymptotics in Section 5 of the queue length distribution and in Section 6 of the orbit
length distribution.

Denote by \( X \) a generic discrete random variable with pgf \( X(z) \). By definition, the
Taylor expansion of \( X(z) \) at the origin is equal to

\[ X(z) = \sum_{n=0}^{\infty} \mathbb{P}[X = n] z^n . \tag{40} \]

Let \( \zeta \) be the radius of convergence of \( X(z) \). Because \( X(z) \) is a pgf, we have that \( \zeta \geq 1 \). According to Pringsheim’s theorem [14, Th. IV.6], \( \zeta \) is a singularity of \( X(z) \). Throughout the rest of this paper, we will assume that the encountered pgfs
have a single dominant singularity\(^3\). If the dominant singularity of a pgf \(X(z)\) is a simple pole, it then relatively easily follows, from the Laurent series of \(X(z)\) at \(\zeta\), that \(\Pr[X = n]\) has an asymptotic geometric behaviour. In priority queueing systems, other type of singularities can occur (c.f. \([18, 19]\)), even if the service time distributions are well-behaved functions. Therefore, we need a more general method. The method we apply, is a singularity analysis of the pgf \(X(z)\) of interest.

For a detailed explanation of this method we refer the reader to \([14, \text{Chap. } 6]\) or \([24, \text{Sect. } 3]\). We provide Corollary VI.1 in \([14]\) as a theorem, which we will make use of frequently.

Theorem 4.1. (Corrolary VI.1 in \([14]\)).
Assume the following two conditions on an analytic function at the origin \(f(z)\), with Taylor coefficients \(f_n\):

1. There exists numbers \(r, \phi, r > 1\) and \(0 < \phi < \frac{\pi}{2}\), such that the function \(f(z)\) is analytic in \(\Delta(\phi, r)\),
\[
\Delta(\phi, r) := \{z \parallel |z| < r, z \neq 1, |\arg(z - 1)| > \phi\}.
\]
(41)

2. Let \(\alpha \in \mathbb{R} \setminus \mathbb{Z} \leq 0\). We have that
\[
f(z) \sim (1 - z)^{-\alpha}, \quad \text{as } z \to 1, z \in \Delta,
\]
where \(f(z) \sim g(z)\) means that \(f(z)/g(z)\) tends to 1 as \(z \to 1\) in \(\Delta\).

Then, the coefficients of \(f(z)\) satisfy,
\[
f_n \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)},
\]
where we write \(f_n \sim g_n\) for \(n \to \infty\) if \(\lim_{n \to \infty} g_n/f_n = 1\) and \(\Gamma(\cdot)\) stands for the Gamma function.

Remark 1. 1. The first assumption is rather technical. It means that the function can be analytic continued in a domain, larger than its disk of convergence. This domain has a specific form, defined by (41), and is called a \(\Delta\)-domain.\(^4\).

A function that is analytic in a \(\Delta\)-domain is then called a \(\Delta\)-analytic function.

2. By the rescaling rule, for \(z \to \zeta\) in some \(\Delta\)-domain, we have that
\[
f(z) \sim (\zeta - z)^{-\alpha} \quad \rightarrow \quad f_n \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)} \zeta^{-n-\alpha}.
\]

Our strategy to derive tail asymptotics for a discrete random variable \(X\) consists of determining \(\zeta, c\) and \(\alpha\), such that we can write
\[
X(z) \sim \sum_{j=0}^{[-\alpha]} d_j (\zeta - z)^j + c(\zeta - z)^{-\alpha} \quad \text{as } z \to \zeta, z \in \Delta.
\]
(43)

Applying Theorem 4.1 then yields
\[
\Pr[X = n] \sim c n^{\alpha-1} \frac{1}{\Gamma(\alpha)} \zeta^{-n-\alpha}.
\]
(44)

Notice that the analytic terms \(d_j (\zeta - z)^j, j = 0, \ldots, [-\alpha]\) do not contribute to the asymptotic coefficients. Therefore, unless needed, we never explicitly compute the constants \(d_j\). Usually, we will write such unspecified constants as \(x_i, \hat{x}_i\) or \(\hat{x}_i\). We

\(^3\)Dominant singularities are singularities that lie on the boundary of the disk of convergence. In other words, dominant singularities are singularities with smallest norm.

\(^4\)We refer the reader to \([14, \text{Fig. VI.6}]\) for a picture of this region.
will assume that $B_1^*(s)$ and $B_2^*(s)$ are $\Delta$-analytic to ensure $N_1(z)$ and $N_2(z)$ are $\Delta$-analytic. For notational purposes, we will not write explicitly that the limit is taken in a $\Delta$-domain in the remainder.

The singularities of $N_1(z)$ and $N_2(z)$ depend on the pgfs $A_1(z_1, z_2)$, $A_2(z_1, z_2)$ and $Y(z)$. These pgfs are defined through the LST of the service time distributions $B_1^*(s)$ and $B_2^*(s)$. Therefore, it is clear that the asymptotics strongly depend on the service time distributions. Let us define $-\tau_j$ as the dominant singularity of $B_j^*(s)$ for $j = 1, 2$. Recall that for an LST, the dominant singularity is the rightmost singularity on the negative real axis; hence, $\tau_j \geq 0$. To proceed carefully, we classify the service time distributions according to the behaviour of their LST $B_j^*(s)$ in their rightmost singularity $-\tau_j$:

1. Type-1: $\tau_j > 0$ and $B_j^*(-\tau_j) = \infty$, 
2. Type-2: $\tau_j > 0$ and $1 < B_j^*(-\tau_j) < \infty$, 
3. Type-3: $\tau_j = 0$ and $B_j^*(-\tau_j) = 1$.

In case the distribution has no singularities, we say it is of type-1 ($\tau_j = \infty$). This classification for probability distributions on the positive halfline was first introduced in [2]. Type-1 distributions are well-behaved functions and are also referred to as light-tailed distributions. It is worth noting that the dominant singularity of these distributions is not necessary a pole, for instance the Gamma distribution with $LST (1 + \theta s)^{-k}$, with parameters $\theta$ and $k$ real and strictly positive. Type-2 distributions are called semi-exponential distributions, because the tail is dominated by an exponential but is not a pure exponential one, like type-1 distributions. Lastly, type-3 distributions are the heavy-tailed distributions, because their tail is dominated by an exponential one. Often encountered type-3 distributions in practice are the Pareto distribution and the lognormal distribution. For more discussion, we refer the reader to [1].

In our analysis, we will consider (almost) every possible distributions types for $B_1^*(s)$ and $B_2^*(s)$. In case of type-3 distributions, we restrict ourself to the important subclass of distributions of the form

$$1 - B_j(t) \sim t^\alpha_j, \quad \text{as } t \to \infty,$$  

where $-\alpha_j > 1$. We refer to these distributions as fat-tailed distributions. Such distributions include the Pareto-like distributions. It can be shown that for distributions satisfying (45), it holds that [6, Th. 8.1.6]

$$B_j^*(s) \sim 1 + \sum_{l=1}^{[-\alpha_j]} \frac{B_j^{(l)}(0)}{l!} s^l - \Gamma(1 + \alpha_j)s^{-\alpha_j}, \quad \text{as } s \to 0.$$  

From this expression, we see that these are distributions where the first $m$ moments exist, $m = [-\alpha_j]$, and that have an infinite $(m + 1)^{st}$ moment. The assumption $-\alpha_j > 1$ is to ensure that the first moment $\beta_j$ of $B_j$ exists. We want to emphasize that the singularity analysis of regularly varying distributions\(^5\) and fat-tailed distributions does not differ significantly, since the slowly varying function $L(t)$ can in fact be treated as a constant. The analysis can be easily extended to regularly varying tails, using Theorem 5 in [13]. However we prefer not to, since this only leads to more complicated notations, without making the analysis different.

\(^5\)A regularly varying distributions is of the form (45), but multiplied with a slowly varying function.
For type-2 distributions, we restrict ourself to type-2 distributions with an LST similar as (46), namely as follows

\[ B^*_2(s) \sim \sum_{l=0}^{|-\alpha_j|} \frac{B_j^{(l)}(-\tau_j)}{l!}(s + \tau_j)^l + k_j(s + \tau_j)^{-\alpha_j}, \quad \text{as } s \to -\tau_j, \quad (47) \]

where \( \alpha_j < 0 \) and \( k_j \) is a real constant such that \( k_2 \Gamma(\alpha_j) > 0 \).

5. **Asymptotics of the queue length.** In order to obtain asymptotics for the queue length, we perform a singularity analysis of the pgf \( N_1(z) \), given by (38), as described in Section 4. The singularities of \( N_1(z) \) are those of \( A_1(z, 1) \), \( A_2(z, 1) \) and the possible zeros of the denominator. Let us define \( \xi_1 \) and \( \xi_2 \) as the dominant singularity of \( A_1(z, 1) \) and \( A_2(z, 1) \) respectively. Using Definition (5) it follows that

\[
\begin{align*}
\xi_1 &= 1 + \frac{\tau_1}{\lambda_1}, \\
\xi_2 &= 1 + \frac{\tau_2}{\lambda_1}.
\end{align*}
\]

Lastly, we define \( \xi \) as the dominant singularity of \( N_1(z) \).

5.1. **Type-1 service time distributions for class-1 customers.** Because in this case, \( A_1(\xi_1, 1) = B_1^*(\tau_1) = \infty \), \( A_1(1, 1) = 1 \), \( A_1^{(1)}(1, 1) < 1 \) and \( A_1(z, 1) \) is a strictly increasing, convex function on the positive real axis, the denominator of (38) has a simple zero \( z_0 \), such that \( 1 < z_0 < \xi_1 \). The dominant singularity \( \xi \) of \( N_1(z) \) is then either \( z_0 \) or \( \xi_2 \).

5.1.1. **Type-1 service time distributions for class-2 customers.** We have two possible cases depending on the minimum of \( z_0 \) and \( \xi_2 \).

**Case 1:** \( z_0 < \xi_2 \). The dominant singularity of \( N_1(z) \) is \( z_0 \). The singular expansion of \( N_1(z) \) in the neighbourhood of \( z_0 \) singularity can be obtained from its Laurent series and is given by

\[
N_1(z) \sim \frac{1 - p_1 \rho - p_1(1 - \rho)z_0 - p_2 A_2(z_0, 1)}{p_1(1 - A_1^{(1)}(z_0, 1))}(z_0 - z)^{-1}, \quad \text{as } z \to z_0. \quad (50)
\]

Transferring to coefficients yields

\[
P[N_1 = n] \sim \frac{1 - p_1 \rho - p_1(1 - \rho)z_0 - p_2 A_2(z_0, 1)}{p_1(1 - A_1^{(1)}(z_0, 1))} z_0^{-n-1}. \quad (51)
\]

Remark that the asymptotic distribution of \( N_1 \) consists exclusively of an exponential term \( z_0^{-n} \), besides a constant factor. We say that \( N_1 \) has, in this case, a geometric behaviour.

**Case 2:** \( \xi_2 < z_0 \). The dominant singularity of \( N_1(z) \) is now \( \xi_2 \). The singularity depends on the LST \( B_2^*(s) \). Without loss of generality, we write that \( B_2^*(s) \) has a singular expansion at \( s = \tau_2 \) of the form

\[
B_2^*(s) \sim k_2(s + \tau_2)^{-\alpha_2} \quad \text{as } s \to -\tau_2, \quad (52)
\]

where \( \alpha_2 > 0 \) and \( k_2 \) is a real constant wherefore \( k_2 \Gamma(\alpha_2) > 0 \). Because \( A_2(z, 1) = B_2^*(\lambda_1(1 - z)) \), we find that

\[
A_2(z, 1) \sim k_2\lambda_1^{-\alpha_2}(\xi_2 - z)^{-\alpha_2}, \quad \text{as } z \to \xi_2.
\]
Notice that \( A_1(z, 1) \) is analytic at \( z = \xi_2 \). It is sufficient to approximate \( 1/(A_1(z, 1) - z) \) as
\[
\frac{1}{A_1(\xi_2, 1) - \xi_2} + O(\xi_2 - z), \quad \text{as } z \to \xi_2,
\]
where we write \( f(s) = O(g(s)) \) as \( s \to s_0 \) if \( \lim_{s \to s_0} \sup |f(s)/g(s)| < \infty \).

Using the expansion of \( A_1(z, 1) \) and \( A_2(z, 1) \) at \( z = \xi_2 \), we obtain that
\[
N_1(z) \sim \frac{\lambda_2k_2\lambda_1^{\alpha_2-1}}{\xi_2 - A_1(\xi_2, 1)}(\xi_2 - z)^{-\alpha_2}, \quad \text{as } z \to \xi_2. \tag{53}
\]

Using Theorem 4.1, we conclude that
\[
\mathbb{P}[N_1 = n] \sim \frac{\lambda_2k_2\lambda_1^{\alpha_2-1}}{\xi_2 - A_1(\xi_2, 1)} \frac{n^{\alpha_2-1}}{\Gamma(\alpha_2)} \xi_2^{-n\alpha_2}. \tag{54}
\]

If \( \alpha_2 \neq 1 \), the distribution of \( N_1 \) is, in this case, is the product of a power law \( n^{\alpha_2-1} \) and an exponential \( \xi_2^{-n} \). We say that \( N_1 \) has, in this case, a semi-geometric behaviour. Note that \( \alpha_2 = 1 \) means that \( -\tau_2 \) is a simple pole of \( B_2^*(s) \). If this is the case, \( N_1 \) has a geometric behaviour.

5.1.2. Type-2 service time distributions for class-2 customers. The singularity analysis yields the same result as the previous section. If \( z_0 < \xi_2 \), the analysis is completely the same. If \( \xi_2 < z_0 \), we make use of \( (47) \), it follows that
\[
A_2(z, 1) \sim \sum_{j=0}^{\lceil -\alpha_2 \rceil} \frac{B_2^{(j)}(-\tau_2)}{j!} \lambda_1^j(\xi_2 - z)^j + k_2\lambda_1^{-\alpha_2}(\xi_2 - z)^{-\alpha_2} \quad \text{as } z \to \xi_2, \tag{55}
\]
where now \( -\alpha_2 > 0 \). This yields a singular expansion for \( N_1(z) \) of the form
\[
N_1(z) \sim \sum_{j=0}^{\lceil -\alpha_2 \rceil} x_j(\xi_2 - z)^j + \frac{\lambda_2k_2\lambda_1^{\alpha_2-1}}{\xi_2 - A_1(\xi_2, 1)}(\xi_2 - z)^{-\alpha_2}, \quad \text{as } z \to \xi_2. \tag{56}
\]

Where \( x_j \) are some constants. Applying Theorem 4.1 results again into \( (54) \).

5.1.3. Type-3 service time distributions for class-2 customers. Since \( B_2 \) is a type-3 distribution, we have that \( \xi_2 = 1 \), hence \( \xi = 1 \). Further, by making use of \( (46) \) we find that
\[
A_2(z, 1) \sim 1 + \sum_{j=1}^{\lceil -\alpha_2 \rceil} \frac{B_2^{(j)}(0)}{j!} \lambda_1^j(1 - z)^j - \Gamma(1 + \alpha_2)\lambda_1^{-\alpha_2}(1 - z)^{-\alpha_2}, \quad \text{as } z \to 1. \tag{57}
\]

The numerator of \( (38) \) can therefore be estimated for \( z \to 1 \) as
\[
p_1(1 - \rho)(1 - z) - \sum_{j=1}^{\lceil -\alpha_2 \rceil} \frac{B_2^{(j)}(0)}{j!} p_2\lambda_1^j(1 - z)^j + p_2\Gamma(1 + \alpha_2)\lambda_1^{-\alpha_2}(1 - z)^{-\alpha_2}. \tag{58}
\]

Since the denominator of \( (38) \) vanishes for \( z = 1 \), we now estimate the denominator of \( (38) \) as \( p_1(1 - \lambda_1\beta_1)(1 - z) \) for \( z \to 1 \). If we now cancel the common factor \( (1 - z) \) in the estimated numerator and estimated denominator of \( (38) \) near \( z = 1 \), we find the following singular expansion for \( N_1(z) \) at \( z = 1 \)
\[
N_1(z) \sim 1 + \sum_{j=1}^{\lceil -\alpha_2 \rceil - 1} x_j(1 - z)^j + \frac{\lambda_2\Gamma(1 + \alpha_2)\lambda_1^{-\alpha_2-1}}{(1 - \lambda_1\beta_1)}(1 - z)^{-\alpha_2-1}. \tag{59}
\]
Transferring to coefficients yields
\[ P[N_1 = n] \sim \frac{\lambda_2}{\Gamma(1 + \alpha_2)} \frac{\lambda_1^{-\alpha_2} n^{\alpha_2}}{1 - \lambda_1 \beta_1}. \] (60)

The distribution of \( N_1 \) consists in this case solely of a power term \( n^{\alpha_2} \).

5.2. **Type-2 service time distributions for class-1 customers.** We assume the LST of \( B_1^*(s) \) is of the form (47) in this section. Since now \( A_1(\xi_1, 1) = B_1(-\tau_1) < \infty \), it is no longer guaranteed that there exists a (simple) zero \( z_0 \) of \( A_1(z, 1) - z \) in the interval \([1, \xi_1]\). We now have three candidates for the dominant singularity \( \xi \) of \( N_1(z) \), namely \( z_0 \) (if any), \( \xi_1 \) and \( \xi_2 \).

5.2.1. **Type-1 and type-2 service time distributions for class-2 customers.**

**Case 1:** \( z_0 \) does not exist, and \( \xi_1 < \xi_2 \). We deduce that
\[ A_1(z, 1) \sim \sum_{j=0}^{[-\alpha_1]} \frac{B_1^{(j)}(-\tau_1)}{j!} \lambda_1^j (\xi_1 - z)^j + k_1 \lambda_1^{-\alpha_1} (\xi_1 - z)^{-\alpha_1} \quad \text{as } z \to \xi_1. \]

For \( z \to \xi_1 \), we estimate the factor \( 1/(A_1(z, 1) - z) \) as follows
\[
\frac{1}{A_1(z, 1) - z} \sim \sum_{j=0}^{[-\alpha_1]} \frac{B_1^{(j)}(-\tau_1)}{j!} \lambda_1^j (\xi_1 - z)^j + k_1 \lambda_1^{-\alpha_1} (\xi_1 - z)^{-\alpha_1} - \xi_1 + (\xi_1 - z)
\]
\[
\sim \sum_{j=0}^{[-\alpha_1]} \frac{k_1 \lambda_1^{-\alpha_1}}{(B_1^{(j)}(-\tau_2) - \xi_1)^j} (\xi_1 - z)^{-\alpha_1},
\]
where we in the second step expanded the denominator and kept only the terms of order \( -\alpha_1 \) or lower (remark that \( -\alpha_1 > 0 \)). Plugging this expansion into (38) yields
\[ N_1(z) \sim \sum_{j=0}^{[-\alpha_1]} \tilde{x}_j (\xi_1 - z)^j + \frac{p_2 A_2(\xi_1, 1) + p_1 (1 - \rho) \xi_1 + p_1 \rho - 1}{(A_1(\xi_1, 1) - \xi_1)^j} k_1 \lambda_1^{-\alpha_1} (\xi_1 - z)^{-\alpha_1}. \]

Transferring to coefficients then yields
\[ P[N_1 = n] \sim k_1 \lambda_1^{-\alpha_1} \frac{p_2 A_2(\xi_1, 1) + p_1 (1 - \rho) \xi_1 + p_1 \rho - 1}{(A_1(\xi_1, 1) - \xi_1)^2} n^{\alpha_1} \frac{1}{\Gamma(\alpha_1)} \xi_1^{-\alpha_1}. \] (61)

**Case 2:** \( z_0 \) does not exist, and \( \xi_2 < \xi_1 \). This case yields the same singularity analysis as in Case 2 of Section 5.1.1.

**Case 3:** \( z_0 \) exists. This case yields the same singularity analysis as in Section 5.1.1 for both \( z_0 < \xi_2 \) and \( \xi_2 < z_0 \).

5.2.2. **Type-3 service time distributions for class-2 customers.** The singularity analysis and asymptotics are the same as in Section 5.1.3.

5.3. **Type-3 service time distributions for class-1 customers.** If the class-1 service time distributions are of type-3, the denominator no longer possesses a simple zero \( z_0 \), wherefore \( z_0 > 1 \). Because \( \xi_1 = 1 \), the dominant singularity of \( N_1(z) \) is that of \( A_1(z, 1) \), i.e. \( \xi = \xi_1 \). If we assume that the LST of \( B_1 \) is of the form (46), it follows that
\[ A_1(z, 1) \sim \sum_{j=0}^{[-\alpha_1]} \frac{B_1^{(j)}(0)}{j!} \lambda_1^j (1 - z)^j - \Gamma(1 + \alpha_1) \lambda_1^{-\alpha_1} (1 - z)^{-\alpha_1}, \quad \text{as } z \to 1. \] (62)
Consequently,
\[ p_1(A_1(z), 1 - z) \sim p_1 \sum_{j=1}^{(-\alpha_1)} \frac{B_1^{(j)}(0)}{j!} (-1)^j (1 - z)^j - p_1 \Gamma(1 + \alpha_1) \lambda_1^{-\alpha_1} (1 - z)^{-\alpha_1 - 1} + p_1(1 - z), \]  
for \( z \to 1. \)

5.3.1. Type-1 and type-2 service time distributions for class-2 customers. Note that the numerator of (38) vanishes for \( z = 1 \). In case of type-1 and type-2 distributions for \( B_2 \), we have that \( \xi_2 > 1 \). Hence we can estimate \( A_2(z, 1) \sim 1 - \lambda_1 \beta_2 (1 - z) \) for \( z \to 1 \). Therefore the numerator of (38) can be written as
\[ p_1(1 - \lambda_1 \beta_2)(1 - z), \quad \text{for} \; z \to 1. \]
Dividing this expansion by Expansion (63), cancelling the common factor \((1 - z)\), and expanding thereafter the denominator yields the following singular expansion for \( N_1(z) \) for \( z \to 1 \),
\[ N_1(z) \sim 1 + \sum_{j=1}^{(-\alpha_1)} x_j (1 - z)^j + \frac{\Gamma(1 + \alpha_1)}{1 - \lambda_1 \beta_1} \lambda_1^{-\alpha_1} (1 - z)^{-\alpha_1 - 1}. \]  
Hence, applying Theorem 4.1, we obtain that
\[ \mathbb{P}[N_1 = n] \sim \frac{\lambda_1}{1 - \lambda_1 \beta_1} \lambda_1^{-\alpha_1 - 1} n^{\alpha_1}. \]  
The asymptotic distribution of \( N_1 \) is again a pure power-law.

5.3.2. Type-3 service time distributions for class-2 customers. We have that \( \xi_2 = 1 \) and the singular expansion of \( A_2(z, 1) \) at \( z = 1 \) is given by (57). The expansion for the numerator is the same as (58), while the expansion for the denominator is given by (63). We divide (58) by (63) and cancel the common factor \((1 - z)\). If we now expand the denominator and keep only the dominant terms, the result depends on the minimum between \(-\alpha_1\) and \(-\alpha_2\). Namely, if \(-\alpha_1 < -\alpha_2\), we obtain the singular expansion (64). If \(-\alpha_2 < -\alpha_1\), we obtain the singular Expansion (59). Suppose now the boundary case \( \alpha_1 = \alpha_2 \). In this case, the singular expansion at \( z = 1 \) yields
\[ N_1(z) \sim 1 + \sum_{j=1}^{(-\alpha_1)} x_j (1 - z)^j + \lambda \frac{\Gamma(1 + \alpha_1)}{1 - \lambda_1 \beta_1} \lambda_1^{-\alpha_1 - 1} (1 - z)^{-\alpha_1 - 1}, \]  
which leads to
\[ \mathbb{P}[N_1 = n] \sim \frac{\lambda}{1 - \lambda_1 \beta_1} \lambda_1^{-\alpha_1 - 1} n^{\alpha_1}. \]  
Notice that constant factor in this expression is the sum of the constants in Expressions (60) and (65).

6. Asymptotics of the orbit length. In this section we derive the asymptotic distribution of \( N_2 \). Expression (39) contains the implicit defined function \( Y(z) \). Therefore, the singularity analysis of \( N_2(z) \) is more involved than the singularity analysis of \( N_1(z) \). It will be convenient to define the following function
\[ D(z) \triangleq (1 - R^*(\lambda))zA(Y(z), z) - R^*(\lambda)A_2(Y(z), z) \]
\[ = (1 - R^*(\lambda))p_1zY(z) + [p_2z + (1 - p_2z)R^*(\lambda)]A_2(Y(z), z), \]  
(68)
so that $N_2(z)$ can be rewritten as
\[ N_2(z) = \frac{p_0 R^*(\lambda)}{p_2} \frac{p_1 Y(z) + p_2 z - 1}{z - D(z)}. \]  
(69)
The possible singularities for $N_2(z)$ arise from the functions $Y(z)$ and $A_2(Y(z), z)$ and the zeros of $z - D(z)$. In order to perform a singularity analysis on $N_2(z)$, we first have to determine the dominant singularities of $Y(z)$ and $A_2(Y(z), z)$. We emphasize that both $Y(z)$ and $A_2(Y(z), z)$ are in fact pgfs (c.f. [23]). To make the analysis more clear, we introduce some more new variables,
\[
\begin{align*}
\zeta_1 &:= \text{Dominant singularity of } Y(z), \tag{70} \\
\zeta_2 &:= \text{Dominant singularity of } A_2(Y(z), z), \tag{71} \\
\zeta_3 &:= \text{Dominant singularity of } D(z), \tag{72} \\
\zeta &:= \text{Dominant singularity of } N_2(z). \tag{73}
\end{align*}
\]
We have to determine successively $\zeta_1$, $\zeta_2$, $\zeta_3$ and $\zeta$. First, we will show that the dominant singularity of the implicit defined function $Y(z)$ is of square-root type, if $B_1$ is a light-tailed distribution. If $B_1$ is a fat-tailed distribution, the dominant singularity of $Y(z)$ will be that of $B_1^*(s)$. In case of type-2 distributions, both situations can occur. Secondly, to determine $\zeta_2$, we introduce the auxiliary function $h(z)$:
\[ h(z) := \lambda_1 (1 - Y(z)) + \lambda_2 (1 - z) \]  
(74)
so that $A_2(Y(z), z) = B_2^*(h(z))$. Notice that the function $h(z)$ is a decreasing function for $z \in [0, \zeta_1]$ with $h(1) = 0$. There are three possible cases for the singularity of $A_2(Y(z), z)$, depending on the location of $-\tau_2$ compared to $h(\zeta_1)$ [14, Ch. VI.9]:

1. Supercritical case: $h(\zeta_1) < -\tau_2$. 
   By the intermediate value theorem, there exists a $\theta$, $1 < \theta < \zeta_1$, such that $h(\theta) = -\tau_2$. The dominant singularity of $A_2(Y(z), z)$ is $\theta$, thus $\zeta_2 = \theta$. Around this point, $Y(z)$ is analytic because $\theta < \zeta_1$. A singular expansion of $A_2(Y(z), z)$ around $\theta$ is obtained by combining the singular expansion of $B_2^*(s)$ and the Taylor expansion of $Y(z)$ at $\theta$.

2. Subcritical case: $h(\zeta_1) > -\tau_2$.
   When $z$ increases from 0, the singularity of $h(z)$ is achieved before the function $h(z)$ achieves the critical value $-\tau_2$. Therefore the singularity of $A_2(Y(z), z)$ is now driven by the singularity of $h(z)$. Hence, $\zeta_2 = \zeta_1$ and the singular expansion of $A_2(Y(z), z)$ at $\zeta_1$ is obtained by combining the Taylor expansion of $B_2^*(s)$ and the singular expansion of $Y(z)$ at $\zeta_1$.

3. Critical case: $h(\zeta_1) = -\tau_2$.
   We have that $\zeta_2 = \zeta_1$ and $A_2(Y(\zeta_2), \zeta_2) = B_2^*(-\tau_2)$. The singular expansion is obtained by composition rules of the singular expansions.

Remark that not all cases are possible if $\zeta_1 = 1$ or $\tau_2 = 0$. Indeed, if we have that $\zeta_1 = 1$, the supercritical case can never occur, because $h(1) = 0$ and $\tau_2 \geq 0$. If we have that $\tau_2 = 0$, the subcritical case never occurs, because $h(\zeta_1) \leq 0$. Except when both service time distributions are of type-3, we will never treat the critical case because this case only occurs for a very specific choice of parameters and service time distributions.

Thirdly, from the definition of $D(z)$, see (68), it is clear that $\zeta_3 = \min(\zeta_1, \zeta_2)$. But from the above considerations we already know that $\zeta_2 \leq \zeta_1$. Finally, if there
exists a $z_0 \in [1, \zeta_3]$ such that $z_0 - D(z_0) = 0$, it necessary holds that $\zeta = z_0$. If this is not the case, we have that $\zeta = \zeta_3$ because of the previous reasonings.

6.1. **Type-1 service time distributions for class-1 customers.** In this section we assume that the service time distribution of class-1 customers are type-1 distributions. We start with the singularity analysis of $Y(z)$. This is done by applying the Smooth implicit-function scheme [14, Th. VII.3, pg 468].

**Lemma 6.1.** We have that $\zeta_1 = \omega$, where $\omega$ is a square-root singularity. The tuple $(\omega, Y(\omega))$ is the positive solution $(z, w)$ of the following characteristic system of equations,

$$A_1(z, w) = z, \quad A_1^{(1)}(z, w) = 1.$$  \hfill (75)

The expansion of $Y(z)$ in the neighbourhood of $\omega$ is equal to

$$Y(z) \sim Y(\omega) - c_Y \sqrt{\omega - z} + O(\omega - z), \quad \text{for } z \to \omega,$$  \hfill (76)

the expansion being valid in a $\Delta$-domain. The constant $c_Y$ is given by

$$c_Y = \sqrt{2} \lambda_2 B_1^* \left( \lambda_1 (1 - Y(\omega)) + \lambda_2 (1 - \omega) \right).$$  \hfill (77)

**Proof.** See [14]. \hfill $\square$

**Remark 2.**

1. If a solution of the characteristic system exists within the positive quadrant, where $A_1(z, w)$ is analytic, then it will be necessary unique, because the Taylor coefficients of $Y(z)$ cannot admit two asymptotic expressions with different parameters [14].

2. We have that $A_1^{(1)}(Y(1), 1) = \lambda_1 \beta_1 < 1$. Hence, by the implicit function theorem, $Y(z)$ is analytic in $z = 1$. Therefore, it necessary holds that $\omega > 1$.

In the remainder of this section, we use notation $\zeta_1$ instead of $\omega$ as the dominant singularity of $Y(z)$.

6.1.1. **Type-1 service time distributions for class-2 customers.**

**Supercritical case.** In the supercritical case and since $B_2$ is a type-1 distribution, we have that $A_2(Y(\zeta_2), \zeta_2) = B_2(-\tau_2) = \infty$. Therefore, $D(\zeta_3) = \infty$. This property, together with the fact that $D(z)$ is an increasing, convex function on the positive real-axis, $D(1) = 1$ and

$$D'(1) = 1 - \frac{p_0 R^*(\lambda)}{1 - \lambda_1 \beta_1} < 1,$$  \hfill (78)

implies that $z - D(z)$ has a simple zero $z_0$ wherefore $1 < z_0 < \zeta_3$. The singular expansion of $N_2(z)$ in the neighbourhood of $\zeta = z_0$ is given by

$$N_2(z) \sim \left( \frac{p_0 R^*(\lambda)}{p_2} \right) \left( \frac{1 - p_1 Y(z_0) - p_2 z_0}{1 - D'(z_0)} \right) (z_0 - z)^{-1} \quad \text{as } z \to z_0.$$  \hfill (79)

Transferring to coefficients yields

$$P[N_2 = n] \sim \left( \frac{p_0 R^*(\lambda)}{p_2} \right) \left( \frac{1 - p_1 Y(z_0) - p_2 z_0}{1 - D'(z_0)} \right) z_0^{-n-1}.$$  \hfill (80)
Subcritical case. In this case we have that the dominant singularity of \( Y(z) \) is transferred to \( A_2(Y(z), z) \) and \( D(z) \). It follows that \( D(\zeta_3) < \infty \). Hence, unlike in the supercritical case, the simple zero \( z_0 \) of \( z - D(z) \) does not necessarily exist. Therefore, there are two possibilities for \( \zeta \). Either \( \zeta = \zeta_0 \), if \( z_0 \) exists, or \( \zeta = \zeta_3 \). The first case will lead to the same coefficient asymptotics as (80). In order to obtain the subexponential factor in the latter case, we consecutively write down singular expansions of the pgfs \( A_2(Y(z), z) \) and \( D(z) \) at \( z = \zeta_1 \). Because we are in the subcritical case, \( B_2^\ast(s) \) is analytic in \( h(\zeta_1) \). Hence, we can write

\[
A_2(Y(z), z) \sim B_2^\ast(h(\zeta_1)) + B_2^\ast(\lambda_1c_Y \sqrt{\zeta_1 - z}), \quad \text{as } z \to \zeta_1. \tag{81}
\]

A singular expansion of \( D(z) \) is simply obtained by using those of (81) and (76). This yields

\[
D(z) \sim D(\zeta_1) - c_Yc_D \sqrt{\zeta_1 - z}, \quad \text{as } z \to \zeta_1. \tag{82}
\]

The constant \( c_D \) is given by

\[
c_D = (1 - R^\ast(\lambda)p_1\zeta_1[1 - \lambda_2B_2^\ast(\lambda_1)] - R^\ast(\lambda)\lambda_1B_2^\ast(\lambda_1)). \tag{83}
\]

We can now make use of (76) and (82) to obtain a singular expansion of \( N_2(z) \). For \( z \to \zeta_2 \), we have that

\[
N_2(z) \sim \left( \frac{p_0R^\ast(\lambda)}{p_2} \right) p_1 Y(\zeta_1) + p_2 \zeta_1 - 1 - p_1c_Y \sqrt{\zeta_1 - z} + O(\zeta_1 - z) + c_Yc_D \sqrt{\zeta_1 - z} + O(\zeta_1 - z)
\]

\[
= \left( \frac{p_0R^\ast(\lambda)}{p_2} \right) \left( p_1 Y(\zeta_1) + p_2 \zeta_1 - 1 - p_1c_Y \sqrt{\zeta_1 - z} + O(\zeta_1 - z) \right)
\]

\[
\times \left( \frac{1}{\zeta_1 - D(\zeta_1)} - c_Yc_D (\zeta_1 - D(\zeta_1))^2 \sqrt{\zeta_1 - z} + O(\zeta_1 - z) \right)
\]

\[
= N_2(\zeta_1) - c_Y \sqrt{\zeta_1 - z} + O(\zeta_1 - z). \tag{84}
\]

The constant \( c_1 \) is given by

\[
c_1 = \left( \frac{p_0R^\ast(\lambda)}{p_2} \right) \frac{p_1(\zeta_1 - D(\zeta_1)) + (p_1 Y(\zeta_1) + p_2 \zeta_1 - 1)c_D}{(\zeta_1 - D(\zeta_1))^2} c_Y. \tag{85}
\]

Using Theorem 4.1, we obtain

\[
P[N_2 = n] \sim c_1 \frac{1}{2\sqrt{\pi n}} \zeta_1^{-n+1/2}. \tag{86}
\]

In case that both service time distributions are of type-1, we can conclude that the asymptotic distribution of \( N_2 \) either has a geometric behaviour \( (\zeta_0^{-n}) \) or a semi-geometric behaviour \( (n^{-3/2}\zeta_1^{-n}) \), depending on the existence of \( z_0 \).

6.1.2. Type-2 service time distributions for class-2 customers.

Supercritical case. In the supercritical case, we have that \( \zeta_3 = \zeta_2 < \zeta_1 \) and \( h(\zeta_2) = -\tau_2 \). Hence, \( \zeta \) is either the simple zero in \( [1, \zeta_2] \) of \( z - D(z) \), if any, or \( \zeta = \zeta_2 \). The first case yields the same result as (80). In the rest of this paragraph, we will assume that \( \zeta = \zeta_2 \). We assume that \( B_2^\ast(s) \) has a singular expansion near \( -\tau_2 \) like (47). Using this expansion, we can write for \( h(z) \to -\tau_2 \)

\[
A_2(Y(z), z) \sim \sum_{j=0}^{\lfloor -\alpha_2 \rfloor} \frac{B_2^\ast(-\tau_2)}{j!}(h(z) + \tau_2)^j + k_2(h(z) + \tau_2)^{-\alpha_2}. \tag{87}
\]
Notice now that \( h(z) \) is analytic in \( \zeta_2 \) and admits therefore a Taylor expansion around \( \zeta_2 \). Inserting the Taylor expansion of \( h(z) \) around \( z = \zeta_2 \) into (87) yields

\[
A_2(Y(z), z) \sim \sum_{j=0}^{\lfloor -\alpha_2 \rfloor} x_j (\zeta_2 - z)^j + (\lambda_1 Y'(\zeta_2) + \lambda_2)^{-\alpha_2} k_2 \zeta_2^{-\alpha_2} (\zeta_2 - z)^{-\alpha_2},
\]

where the \( x_i \) are some constants and where we used that \( h'(z) = -(\lambda_1 Y'(z) + \lambda_2) \). Further, we can write

\[
D(z) \sim \sum_{j=0}^{\lfloor -\alpha_2 \rfloor} \hat{x}_j (\zeta_2 - z)^j + k_2 (\lambda_1 Y'(\zeta_2) + \lambda_2)^{-\alpha_2} (1 - \zeta_2 p_2) R^*(\lambda) + \zeta_2 p_2 \zeta_2^{-\alpha_2} (\zeta_2 - z)^{-\alpha_2},
\]

Using a similar procedure as in the previous subsection, it follows that

\[
N_2(z) \sim \sum_{j=0}^{\lfloor -\alpha_2 \rfloor} \hat{x}_j (\zeta_2 - z)^j + c_2 \zeta_2^{-\alpha_2} (\zeta_2 - z)^{-\alpha_2}.
\]

The constant \( c_2 \) is given as

\[
c_2 = \left( \frac{p_0 R^*(\lambda)}{p_2} \right) \left( 1 - \frac{p_1 Y(\zeta_2) - p_2}{(\zeta_2 - D(\zeta_2))^2} \right) k_2 (\lambda_1 Y'(\zeta_2) + \lambda_2)^{-\alpha_2} \left( 1 - \zeta_2 p_2 R^*(\lambda) + \zeta_2 p_2 \right).
\]

Finally we obtain the asymptotic distribution of \( N_2 \) using the transform theorem,

\[
P[N_2 = n] \sim c_3 \frac{n^{\alpha_2 - 1}}{\Gamma(\alpha_2)} \zeta_2^{-n - \alpha_2}.
\]

Subcritical case. The analysis for the subcritical case is completely the same as for the case of a type-1 distribution for the class-2 service time. This is due to the fact that in the subcritical case, the singularity is driven by that of \( Y(z) \).

6.1.3. Type-3 service time distributions for class-2 customers. As already pointed out, the subcritical case can never occur if \( \tau_2 = 0 \). Moreover, because \( \zeta_1 > 1 \), the scheme is always supercritical. We assume that \( B_2 \) is a fat-tailed distribution. Then we can write for \( z \to 1 \)

\[
A_2(Y(z), z) \sim 1 + \sum_{j=1}^{\lfloor -\alpha_2 \rfloor} x_j (1 - z)^j - \Gamma(1 + \alpha_2) \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_2} (1 - z)^{-\alpha_2},
\]

and

\[
D(z) \sim 1 + \sum_{j=1}^{\lfloor -\alpha_2 \rfloor} \hat{x}_j (1 - z)^j - (p_1 R^*(\lambda) + p_2) \Gamma(1 + \alpha_2) \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_2} (1 - z)^{-\alpha_2}.
\]

From the singular expansion of \( D(z) \) near \( z = 1 \), we find that

\[
N_2(z) \sim 1 + \sum_{j=1}^{\lfloor -\alpha_2 \rfloor - 1} \hat{x}_j (1 - z)^j + c_3 \Gamma(1 + \alpha_2) (1 - z)^{-\alpha_2 - 1}, \quad \text{as } z \to 1,
\]

with \( c_3 \) defined as

\[
c_3 = \frac{\lambda_2}{\rho p_0} \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_2 - 1}.
\]
To obtain this singular expansion of $N_2(z)$, we cancelled first a common factor $(1 - z)$ in the expansions of the numerator and denominator. Then we expanded the denominator again. We used that $\bar{x}_1 = -D'(1)$. We have now obtained the following asymptotic:

$$P[N_2 = n] \sim c_3 n^{\alpha_2}.$$  \hfill (95)

Remark that $R^*(\lambda)$ is incorporated in the constant $p_0$, so the retrial time distribution does have an influence on the constant $c_3$. Similar as for the priority queue, we observe that in this case the distribution of $N_2$ follows a power-law $n^{\alpha_2}$.

6.2. Type-2 service time distributions for class-1 customers. Because $B_1^*(s)$ is no longer infinite in $-\tau_1$, a solution of the characteristic system (75) in the positive quadrant no longer necessarily exists. See also Appendix A in [24] for a detailed discussion. If a solution exists, the previous results apply. Therefore, in this section we will only discuss the case that Lemma 6.1 does not hold, i.e. there exists a $\theta$, wherefore $h(\theta) = -\tau_1$. Moreover, because $Y(\theta) = B_1^*(-\tau_1)$, we obtain an explicit expression for $\theta$:

$$\theta = 1 + \frac{\tau_1 + \lambda_1(1 - B_1^*(-\tau_1))}{\lambda_2}.$$  \hfill (96)

As before, we will write the dominant singularity of $Y(z)$ as $\zeta_1$, in fact $\zeta_1 = \theta$. Note that $A_1(w, z)$ is still bivariate analytic in $(1, 1)$, because of $\tau_1 > 0$, and $A_1^{(1)}(Y(1), 1) \neq 1$. Hence, the implicit function theorem ensures us that $\zeta_1 > 1$. We assume the expansion of $B_1^*(s)$ in $-\tau_1$ is like (47). The singular expansion of $Y(z)$ near $\zeta_1$ can then be written as

$$Y(z) \sim \sum_{j=0}^{[-\alpha_1]} y_j (\zeta_1 - z)^j + k_1 (\lambda_1 Y'(\zeta_1) + \lambda_2)^{-\alpha_1}(\zeta_1 - z)^{-\alpha_1}.$$  \hfill (97)

6.2.1. Type-1 and type-2 service time distributions for class-2 customers.

Supercritical case. The supercritical scheme is the same as in Section 6.1.1 and Section 6.1.2.

Subcritical case. We have two possible instances for the dominant singularity $\zeta$ of $N_2(z)$. Either $\zeta = \zeta_1$, or $\zeta = z_0$, where $z_0$ is a simple zero of $z - D(z)$ contained in the interval $[1, \zeta_1]$, if any. The latter case leads to the same asymptotics as (80). In the following analysis, we assume that such a $z_0$ does not exist. A singular expansion for $A_2(Y(z), z)$ at $z = \zeta_1$ is obtained by combining the Taylor expansion of $B_2^*(s)$ at $s = -\tau_1$ and the singular expansion of $Y(z)$, given by (97). This yields

$$A_2(Y(z), z) \sim \sum_{j=0}^{[-\alpha_1]} x_j (\zeta_1 - z)^j - \lambda_1 B_2^{*'}(-\tau_1)k_1 (\lambda_1 Y'(\zeta_1) + \lambda_2)^{-\alpha_1}(\zeta_1 - z)^{-\alpha_1},$$  \hfill (98)

for $z \to \zeta_1$. Consequently,

$$D(z) \sim \sum_{j=0}^{[-\alpha_1]} \bar{x}_j (\zeta_1 - z)^j - c_D k_1 (\lambda_1 Y'(\zeta_1) + \lambda_2)^{-\alpha_1}(\zeta_1 - z)^{-\alpha_1}, \text{ as } z \to \zeta_1,$$  \hfill (99)

where the constant $c_D$ is given by

$$c_D = R^*(\lambda)\lambda_1 B_2^{*}(-\tau_1) + (1 - R^*(\lambda)p_1\zeta_1 (\lambda_2 B_2^{*}(-\tau_1) - 1).$$  \hfill (100)
With similar techniques as before, we obtain a singular expansion of \( N_2(z) \),
\[
N_2(z) \sim \sum_{j=0}^{[\alpha_1]} x_j (\zeta_1 - z)^j + c_4 (\zeta_1 - z)^{-\alpha_1}, \quad \text{as } z \to \zeta_1. \tag{101}
\]

The constant \( c_4 \) is given as
\[
c_4 = \frac{-c_D \tau_1 / \lambda + p_1 (\zeta_1 - D(\zeta_1))}{(\zeta_1 - D(\zeta_1))^2} k_1 (\lambda_1 Y'(\zeta_1) + \lambda_2)^{-\alpha_1}. \tag{102}
\]

We obtain the following asymptotic
\[
P[N_2 = n] \sim c_4 n^{\alpha_1 - 1} \frac{\zeta_1^{-n-\alpha_1}}{\Gamma(\alpha_1)}. \tag{103}
\]

Remark that \( k_1 \) and \( \Gamma(\alpha_1) \) have the same sign by assumption. Further, based on the properties of an LST, it can easily be seen that \( c_D < 0 \). Finally, \( \zeta_1 - D(\zeta_1) > 0 \), exactly because we assumed that \( z_0 \) does not exist. From these three remarks one can verify, as a minor check, that the obtained tail asymptotic is positive.

6.2.2. Type-3 service time distributions for class-2 customers. The subcritical case cannot occur if \( B_2 \) is a type-3 distribution. The supercritical case is the same as in Section 6.1.3.

6.3. Type-3 service time distributions for class-1 customers. As explained in Appendix A [24], Lemma 6.1 never holds, hence \( \zeta_1 = 1 \). As a consequence, the scheme is always subcritical if the class-2 distribution is type-1 or type-2. The scheme is critical in case that \( \tau_2 = 0 \). Using the Expansion (46), we find the following singular expansion of \( Y(z) \) for \( z \to 1 \)
\[
Y(z) \sim 1 + \sum_{j=1}^{[\alpha_1]} y_j (1 - z)^j - \Gamma(\alpha_1 + 1) \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_1} (1 - z)^{-\alpha_1}, \tag{104}
\]

6.3.1. Type-1 and type-2 service time distributions for class-2 customers. In these two cases, \( B_2^2(s) \) is analytic in \( s = 0 \). Hence, for the singular expansion of \( A_2(Y(z), z) \) for \( z \to 1 \) we get
\[
A_2(Y(z), z) \sim 1 + \sum_{j=1}^{[\alpha_1]} x_j (1 - z)^j - \lambda_1 \beta_2 \Gamma(1 + \alpha_1) \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_1} (1 - z)^{-\alpha_1}. \tag{105}
\]

We also obtain the singular expansion of \( D(z) \) for \( z \to 1 \),
\[
D(z) = 1 + \sum_{j=1}^{[\alpha_1]} \tilde{x}_j (1 - z)^j - c_D \Gamma(\alpha_1 + 1) \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_1} (1 - z)^{-\alpha_1}, \tag{106}
\]
with \( c_D \) defined as
\[
c_D = (1 - R^*(\lambda)) p_1 (1 + \lambda_2 \beta_2) + R^*(\lambda) \lambda_1 \beta_2 . \tag{107}
\]

Using the expansion of \( D(z) \) and \( Y(z) \) for \( z \to 1 \), yields
\[
N_2(z) \sim 1 + \sum_{j=1}^{[\alpha_1] - 1} \tilde{x}_j (z - 1)^j + c_5 \Gamma(\alpha_1 + 1)(1 - z)^{-\alpha_1 - 1}, \quad \text{as } z \to 1. \tag{108}
\]
The constant $c_5$ is equal to
\[ c_5 = \frac{\lambda_1(1 - \lambda_1\beta_1)}{\rho} \frac{1 - p_0}{p_0} \left( \frac{\lambda_2}{1 - \lambda_1\beta_1} \right)^{-\alpha_1 - 1}, \tag{109} \]
where we used that $\tilde{x}_1 = -D'(1)$. Finally, using the transfer theorem, we find that
\[ \mathbb{P}[N_2 = n] \sim c_5 n^{\alpha_1}. \tag{110} \]
Again, similar as for the priority queue, we conclude that the asymptotic distribution of $N_2$ follows a power-law $n^{\alpha_1}$ in this case.

6.3.2. Type-3 service time distributions for class-2 customers. In this particular case, both service time distributions are of type-3, i.e. their dominant singularities are zero. The scheme is now critical and the expansion of $A_2(Y(z), z)$ for $z \to 1$ will depend on the minimum of $-\alpha_1$ and $-\alpha_2$, namely
\[ A_2(Y(z), z) \sim \begin{cases} 
1 + \sum_{j=1}^{[-\alpha_2]} \tilde{x}_j (1 - z)^j + \lambda_1 \beta_2 k_1 \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_1} (1 - z)^{-\alpha_1} & \text{if } \alpha_1 > \alpha_2, \\
1 + \sum_{j=1}^{[-\alpha_1]} \tilde{x}_j (1 - z)^j + k_2 \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_2} (1 - z)^{-\alpha_2} & \text{if } \alpha_2 > \alpha_1, \\
1 + \sum_{j=1}^{[-\alpha_1]} \tilde{x}_j (1 - z)^j + (\lambda_1 \beta_2 k_1 + k_2) \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_1} (1 - z)^{-\alpha_1} & \text{if } \alpha_1 = \alpha_2.
\end{cases} \tag{111} \]
The first case $-\alpha_1 < -\alpha_2$ corresponds to (105). The asymptotics will yield the same result as (110). The second case $-\alpha_2 < -\alpha_1$ corresponds to (92). The corresponding asymptotics will yield the same result as (95). The last case $-\alpha_1 = -\alpha_2$ is a boundary case. However, since it is an important boundary case (both service time distributions are equally “fat”), we treat it explicitly. It can be shown that, if $-\alpha_1 = -\alpha_2$, the asymptotic expansion of $D(z)$ in $z = 1$ is given by
\[ D(z) \sim 1 + \sum_{j=1}^{[-\alpha_1]} \tilde{x}_j (1 - z)^j + c_D \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_1} (1 - z)^{-\alpha_1}, \quad \text{as } z \to 1, \tag{112} \]
where the auxiliary constant $c_D$ is given by
\[ c_D = k_1 p_1 (1 - R^*(\lambda)) + (k_2 + \lambda_1 \beta_2 k_1) (p_1 R^*(\lambda) + p_2). \tag{113} \]
Using similar mathematically manipulations as before, we find that
\[ N_2(z) \sim 1 + \sum_{j=1}^{[-\alpha_1]-1} \tilde{x}_j (1 - z)^j + c_6 \Gamma(1 + \alpha_1) (1 - z)^{-\alpha_1 - 1}, \tag{114} \]
where the constant $c_6$ is given by
\[ c_6 = \frac{1 - p_0}{p_0 \rho} (\lambda_1 (1 - \lambda_1 \beta_1) + \lambda_2) \left( \frac{\lambda_2}{1 - \lambda_1 \beta_1} \right)^{-\alpha_1 - 1}. \tag{115} \]
Notice that $c_6 = c_3 + c_5$. Finally, we obtain that
\[ \mathbb{P}[N_2 = n] \sim c_6 n^{\alpha_1}. \tag{116} \]
7. Performance measures. In this section we list some important performance measures of the queueing system. The probability that the server is empty is given by \( Q(1) \),

\[ Q(1) = 1 - \rho. \] (117)

It is worth noting that this is not equal to the probability that the system is empty, which is given by \( p_0 \), see (34). Also, this performance measure indicates that the probability that a new arrival will be served immediately is independent of the retrial time distribution.

The mean length of the priority queue is found by taking the first derivative of (38), yielding

\[ E[N_1] = \frac{\lambda_1 (\lambda_1 B_1^{\infty}(0) + \lambda_2 B_2^{\infty}(0))}{2(1 - \beta_1 \lambda_1)}. \] (118)

The mean orbit length is found by taking the first derivative of (39), which leads to

\[ E[N_2] = \frac{\lambda_2 [p_2 + p_1 R^*(\lambda)] [\lambda_1 B_1^{\infty}(0) + \lambda_2 B_2^{\infty}(0)]}{2(1 - \lambda_1 \beta_1)(R^*(\lambda) - [p_2 + R^*(\lambda)p_1]\rho)} + \frac{\lambda_2 [1 - R^*(\lambda)](p_1 \beta_1 + p_2 \beta_2)}{R^*(\lambda) - [p_2 + R^*(\lambda)p_1]\rho}. \] (119)

From the derived asymptotics in Sections 5 and 6, we can derive tail asymptotics, i.e. \( P[N_j > n] \) as \( n \to \infty \). It holds that

\[ \sum_{n=0}^{\infty} P[N_j > n]z^n \sim \frac{N_j(z) - 1}{z - 1}, \quad j = 1, 2. \] (120)

If the dominant singularity is strictly greater than one, we have that

\[ P[N_j > n] \sim \frac{P[N_j = n]}{\zeta - 1}, \quad \text{if} \quad \zeta > 1. \] (121)

In case the dominant singularity is equal to one, the singular expansion of \( N_j(z) \) is of the form \( 1 + \sum_{j=1}^{[-\alpha]-1} x_j (1 - z)^j + c \Gamma(1 + \alpha)(1 - z)^{-\alpha-1} \). Therefore, after cancelling \( (1 - z) \), the singular expansion of (120) is of the form

\[ \sum_{n=0}^{\infty} P[N_j > n]z^n \sim \sum_{j=1}^{[-\alpha]^{-1}} (-1)x_j (1 - z)^j - c \Gamma(1 + \alpha)(1 - z)^{-\alpha-2}, \quad z \to 1. \]

Using Theorem 4.1, we obtain that

\[ P[N_j > n] \sim \frac{c}{-\alpha - 1} n^{\alpha+1}, \quad \text{if} \quad \zeta = 1, \] (122)

where we used that \( \Gamma(\alpha + 2) = (\alpha + 1)\Gamma(\alpha + 1) \). Notice that \( -\alpha > 1 \) by assumption. Hence (122) is always positive.

8. Numerical examples. To conclude this paper, we show and discuss some numerical examples. We assume the service times of type-1 customers to be exponentially distributed with mean \( \beta_1 \), i.e.

\[ B_1^*(s) = (1 + \beta_1 s)^{-1}. \] (123)

In order to study the influence of different service times, we assume the service times of type-2 customers to be of the gamma type with shape parameter \( 1/2 \) and mean \( \beta_2 \), i.e.

\[ B_2^*(s) = (1 + 2\beta_2 s)^{-1/2}. \] (124)
The reason for this distribution is twofold. The first reason is to have two different service time distributions. The second reason is to have a coefficient of variation\(^6\) that is bigger than one; it is equal to \(\sqrt{2}\). The coefficient of variation for \(B_1^*(s)\) obviously equals one. Next, we suppose the retrial times follow an Erlang distribution of order two and rate \(\nu > 0\), i.e.

\[
R^*(s) = \left(\frac{\nu}{\nu + s}\right)^2.
\]

(125)

Finally, we define \(\alpha\) as the fraction of class-1 customers load in the overall traffic mix, i.e.,

\[
\alpha = \frac{\lambda_1 \beta_1}{\rho},
\]

(126)

with \(\rho = \lambda_1 \beta_1 + \lambda_2 \beta_2\) as before.

![Figure 1](image_url)

**Figure 1.** Mean values of the queue and orbit lengths versus the mean class-2 service times (\(\rho = 0.7, \beta_1 = 1\) and \(\nu = 3\)).

Figure 1 shows the mean queue length and mean orbit length versus the mean service time of class-2 customers for \(\rho = 0.7, \beta_1 = 1, \nu = 3\) and \(\alpha = 0.25, 0.5\) and 0.75. From Figure 1, it can be seen that the mean queue length increases linearly with \(\beta_2\), while the mean orbit length decreases with increasing \(\beta_2\). The decrease of the mean orbit length is caused by the decrease of \(\lambda_2\) when \(\beta_2\) increases (\(\rho\) is kept constant), hence less class-2 customers arrive to the system. The increase in the mean queue length can be explained as follows. When \(\beta_2\) increases, more class-1 customers will arrive during the time periods that a class-2 customer is being served. Due to the non-preemptive priority policy, arriving class-1 customers cannot interrupt the service of class-2 customers. Hence, they have to wait until this class-2 customer leaves the system, which results in a larger mean queue length.

\[^{6}\]The coefficient of variation of a distribution is defined as the ratio of the standard deviation to the mean.
In Figure 2 the mean queue and orbit lengths are shown as a function of the (mean) class-1 service time with \( \rho = 0.7, \beta_1 = 1.5, \nu = 3 \) and \( \alpha = 0.25, 0.5 \) and 0.75. On the one hand, we observe that the mean queue length decreases with increasing \( \beta_1 \). This is because less class-1 customers arrive when \( \beta_1 \) increases (since \( \lambda_1 \beta_1 \) is kept constant). The mean orbit length on the other hand first decreases rapidly and then increases almost linearly, for increasing \( \beta_1 \). The latter effect is due to the priority policy, the former to the retrial policy. To motivate this, we present in Figure 3 the same model as in Figure 2, but where the retrial times are zero, i.e. \( R^*(s) = 1 \). In this way, we obtained the ordinary non-preemptive priority queueing model. The mean queue lengths are the same in both figures, because the mean queue length is independent of the retrial time distribution. The decrease for small \( \beta_1 \) observed in Figure 2 can be explained as follows. The probability of a successful retrial \( R^*(\lambda_1 + \lambda_2) \) increases, for decreasing \( \lambda_1 \) (for fixed \( \lambda_2 \)). Intuitively, this is because less class-1 customers arrive and therefore there is less competition between the retrial and the arriving customers. Mathematically, this can be seen by taking the first derivative of (125). Consequently, if \( \beta_1 \) increases (hence \( \lambda_1 \) decreases), the effective time until a successful retrial decreases. This initially results in a decrease of the orbit length because the retrials gets more successful. This positive effect continues until \( \beta_1 \) achieves a critical value. When the (mean) class-1 service time gets longer, the variance in the lengths of the periods that both queue and server are empty grows, which implies higher (mean) orbit lengths [23].

The variances of the queue and orbit lengths can be derived from the pgfs (38) and (39). In Figure 4, the variances of the queue and orbit lengths are shown as functions of the load when \( \alpha = 0.25, 0.5 \) and 0.75 respectively.
Figure 3. Mean values of the queue and orbit lengths versus the mean class-1 service times ($\rho = 0.7$, $\beta_2 = 1.5$ and $R^*(s) = 1$).

Figure 4. Variances of the queue and orbit length versus the load ($\beta_1 = \beta_2 = 1.1$ and $\nu = 4$).
We now take a look at the tail probabilities of the orbit length. We assume that both service time distributions are of type-1, with $\tau_1 = 1/\beta_1$ and $\tau_2 = 1/(2\beta_2)$. Figure 5 shows which scheme applies, according to the analysis in Section 6.1.1, for $\lambda_1 = \lambda_2 = 0.3$ and $\nu = 3$. The line $\beta_1 + \beta_2 = c$, for a certain constant $c$, indicates the stability border for the queuing system. The parameterspace $(\beta_1, \beta_2)$ is first divided into two parts, the supercritical and the subcritical scheme. In the supercritical scheme, the tail of the orbit has a geometric behaviour. In the subcritical scheme, there are two cases, geometric or the product of a power law with a geometric (semi-geometric). This yields the second partition of the parameterspace $(\beta_1, \beta_2)$. For a large set of the parameters $\beta_1$ and $\beta_2$, the tail behaviour is thus geometric. Which scheme applies, i.e. the supercritical or the subcritical one, is independent of the retrial time distribution. However, which case applies within the subcritical scheme does depend on the retrial time distribution, because this is determined by the existence of a positive real zero greater than one of the denominator of (39). To illustrate this, we show in Figure 6 which scheme applies if the retrial times are zero. Notice, that in this case the stability condition is different, which explains the move upwards of the stability border in Figure 6. We indeed see that the area that indicates semi-geometric behaviour is changed compared to Figure 5.

Finally, Figure 7 shows the tail probabilities of the orbit length for $\lambda_1 = \lambda_2 = 0.3$, $\beta_1 = 1.2$, $\beta_2 = 1$ and $\nu = 3$, when one of the service time distributions is changed to a fat-tailed distribution, for $\alpha_1 = \alpha_2 = -2.5$ (see also (46)). The straight line in Figure 7 is the case when both service time distributions are of type-1. From Figure 5 we know that the tail behaviour is then geometric. As expected, the tail of the orbit length is much heavier if (one of) the service time distributions is fat-tailed.
Figure 6. Regions for the tail behaviour as a function of the mean service times of both classes ($\lambda_1 = \lambda_2 = 0.3$ and $R^*(s) = 1$).

Figure 7. Tail behaviour of the orbit length for different service time distributions ($\lambda_1 = \lambda_2 = 0.3$, $\beta_1 = 1.2$, $\beta_2 = 1$, $\nu = 3$ and $\alpha_1 = \alpha_2 = -2.5$).
9. Conclusion. In this paper, We have investigated a priority retrial queue with constant retrial policy and different service time distributions between class-1 and class-2 customers. We were able to derive expressions for the pgfs of the stationary queue and orbit lengths. The main contribution of this paper is the singularity analysis of these pgfs, for several different types of service time distributions (including a large subclass of heavy-tailed distributions). The result of this analysis is that, it characterizes the effects of the service time distributions and the retrial time distributions on the queue and orbit lengths distributions. It turns out that the sort of asymptotic behaviour for the orbit length is strongly determined by the service time distributions. However, if both class-1 and class-2 service time distributions are not heavy-tailed, then also the retrial time distribution has an influence on the asymptotic behaviour of the orbit length, since the pgf of the orbit length may then possess a simple pole as dominant singularity. The existence of this simple pole is determined by the service time distributions, but also of the retrial time distribution if the scheme is subcritical. In case one of the service time distributions is a fat-tailed distribution, both queue and orbit length follow a power-law distribution.

REFERENCES

[1] J. Abate and W. Whitt, Asymptotics for M/G/1 low-priority waiting-time tail probabilities, Queueing Systems, 25 (1997), 173–233.
[2] J. Abate, G. L. Choudhury and W. Whitt, Waiting-time tail probabilities in queues with long-tail service-time distributions, Queueing Systems, 16 (1994), 311–338.
[3] J. R. Artalejo, Accessible bibliography on retrial queues: Progress in 2000–2009, Mathematical and Computer Modelling, 51 (2010), 1071–1081.
[4] J. R. Artalejo and A. Gómez-Corral, Retrial Queueing Systems, Springer-Verlag Berlin Heidelberg, 2008.
[5] I. Atencia and P. Moreno, A single-server retrial queue with general retrial times and Bernoulli schedule, Applied Mathematics and Computation, 162 (2005), 855–880.
[6] N. H. Bingham, C. M. Goldie and J. L. Teugels, Regular Variation, Cambridge University Press, 1986.
[7] B. D. Choi, K. K. Park and C. E. M. Pearce, An M/M/1 retrial queue with control policy and general retrial times, Queueing Systems, 14 (1993), 275–292.
[8] B. D. Choi and Y. Chang, Single server retrial queues with priority calls, Mathematical and Computer Modelling, 30 (1999), 7–32.
[9] A. Devos, J. Walraevens and H. Bruneel, A priority retrial queue with constant retrial policy, International Conference on Queueing Theory and Network Applications, (2018), 3–21.
[10] G. Falin and J. G. Templeton, Retrial Queues, CRC Press, 1997.
[11] K. Farahmand, Single line queue with repeated demands, Queueing Systems, 6 (1990), 223–228.
[12] G. Fayolle, A simple telephone exchange with delayed feedbacks, Proc. of the International Seminar on Teletraffic Analysis and Computer Performance Evaluation, (1986), 245–253.
[13] P. Flajolet and A. Odlyzko, Singularity analysis of generating functions, SIAM Journal on Discrete Mathematics, 3 (1990), 216–240.
[14] P. Flajolet and R. Sedgewick, Analytic Combinatorics, Cambridge University Press, 2009.
[15] S. Gao, A preemptive priority retrial queue with two classes of customers and general retrial times, Operational Research, 15 (2015), 233–251.
[16] A. Gómez-Corral, Stochastic analysis of a single server retrial queue with general retrial times, Naval Research Logistics (NRL), 46 (1999), 561–581.
[17] J. Kim, B. Kim and S. S. Ko, Tail asymptotics for the queue size distribution in an M/G/1 retrial queue, Journal of Applied Probability, 44 (2007), 1111–1118.
[18] H. Li and Y. Q. Zhao, Exact tail asymptotics in a priority queue characterizations of the non-preemptive model, Queueing Systems, 68 (2011), 165–192.
[19] T. Maerents and J. Walraevens and H. Bruneel, Priority queueing systems: from probability generating functions to tail probabilities, Queueing Systems, 55 (2007), 27–39.
[20] T. Phung-Duc, W. Rogiest, Y. Takahashi and H. Bruneel, Retrial queues with balanced call blending: Analysis of single-server and multiserver model, *Annals of Operations Research*, 239 (2016), 429–449.

[21] T. Phung-Duc and W. Rogiest, Two way communication retrial queues with balanced call blending, *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, (2012), 16–31.

[22] W. Shang and L. Liu and Q. Li, Tail asymptotics for the queue length in an M/G/1 retrial queue, *Queueing Systems*, 52 (2006), 193–198.

[23] J. Walraevens, *Discrete-Time Queueing Models With Priorities*, Ph.D thesis, Ghent University, 2004.

[24] J. Walraevens, D. Claeys and T. Phung-Duc, Asymptotics of queue length distributions in priority retrial queues, *arXiv:1801.06993*, (2018).

[25] J. Walraevens and B. Steyaert and H. Bruneel, Performance analysis of a single-server ATM queue with a priority scheduling, *Computers & Operations Research*, 30 (2003), 2003.