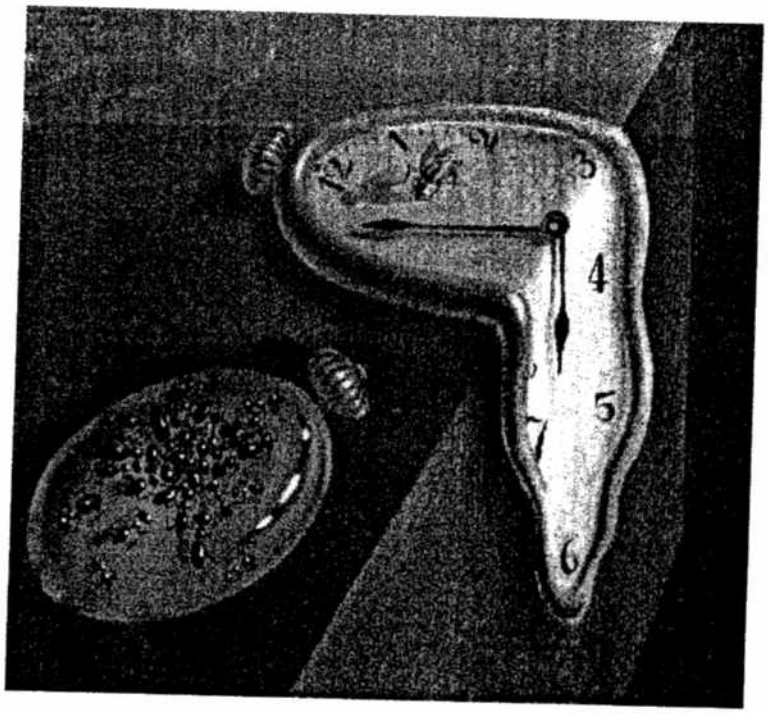
Institut für Automation
Abt. für Automatisierungssysteme

Technische
Universität
Wien

Projektbericht Nr. 183/1–17
März 1991

# Exponential Limiting Distributions in Queueing Systems with Deadlines

*M. Drmota, U. Schmid*



Ausschnitt aus: Salvador Dali, "Die Beständigkeit der Erinnerung"

# Exponential Limiting Distributions in Queueing Systems with Deadlines

M. DRMOTA[1] AND U. SCHMID[2]

Abstract. This paper contains some general theorems on the limiting distribution of a certain random variable $S_T$ arising in the context of recurrent events. $S_T$ is especially meaningful to the investigation of discrete time queueing systems subjected to service time deadlines. It may be viewed as a sum of mutually independent conditional random variables $\mathcal{B}_T$ having the same distribution. We assume that $\mathcal{B}_T$ depends on a parameter $T$ and tends to an unconditional random variable $\mathcal{B}$ for $T \to \infty$. Under some weak conditions concerning the conditional probability generating function $B_T(z)$ it follows that $S_T$ is approximately exponentially distributed with parameter $\lambda_T = 1/\mu_T$, $\mu_T = B_T'(1)/(1 - B_T(1))$, which tends to infinity for $T \to \infty$. We provide uniform asymptotic expansions for the appropriate probabilities and for all moments, too.

*Keywords:* applied probability theory, asymptotic probabilities, renewal processes, queueing systems with deadlines.

*AMS subject classification:* 60K, 68M20

## 1. INTRODUCTION

There is a relatively young and flourishing branch of computer industries, which has been somewhat neglected by traditional computer science: the development of *real-time systems* controlling spacecrafts, power plants, or automated factories, for example. Generally speaking, a real-time system is concerned with tasks, which have to be performed not only correctly, but also in a timely fashion; usually, they are are forced to finish within a predefined deadline. Otherwise, there might be severe consequences.

A well-known design problem for real-time systems concerns methods for a suitable *task scheduling*. Scheduling goals for real time systems are much different from those fitting the needs of ordinary computer systems, since timeliness is by no means equivalent to throughput or similar performance measures. The whole problem is sufficiently well-understood in the case of *deterministic* task arrivals ("total knowledge"), mainly *periodic* tasks. Requirements of this type may be scheduled in advance, i.e., *offline*; systems relying on this idea are usually called *static*. On the other hand, sufficient theoretical foundations for *indeterministic* task arrivals (without "total knowledge") are lacking. Scheduling in such *dynamic* systems has to be performed during normal operation, i.e., *online*. A brief survey of scheduling algorithms for real-time systems may be found in [**CSR**], for example.

Some of the recent research of one of the authors has been devoted to the problem of qualifying scheduling for indeterministic task arrivals in real-time systems. As a result, a mathematically tractable quality measure for such scheduling algorithms was found by means of a certain queueing system approach, which is based on the following idea: Consider a queueing system consisting of a task scheduler, a task list of (potential) infinite capacity, and a single server. Newly arriving tasks are inserted into the task list by the

[1] Department for Algebra and Discrete Mathematics (118/4) at the Technical University of Vienna, Wiedner Hauptstraße 8–10, A-1040 Wien.

[2] Department for Automation (183/1) at the Technical University of Vienna, Treitlstraße 3, A-1040 Wien.

scheduler, according to the scheduling discipline. The server always executes the task at the head of the list. A dummy task will be generated by the scheduler if the list becomes empty. If the server executes a dummy task, the system is called *idle*, otherwise *busy*.

Rearranging of the task list is assumed to occur at discrete points on the time axis only, without any scheduling overhead. The (constant) time interval between two such points is called a *cycle*. Due to this assumption we are able to model tasks formed by indivisible *actions* with duration of 1 cycle. The *task execution time* of a task is the number of cycles necessary for processing the task to completion if it might occupy the server exclusively. An ordinary task may have an arbitrary task execution time, a dummy task as mentioned above is supposed to consist of a single no-operation action (1 cycle). The *service time* of a task is the time (measured in cycles) from the beginning of the cycle in which the task arrives at the system to the end of the cycle which completes the execution of that task.

Obviously, the time axis is covered by a sequence of *busy periods*, which are supposed to include the initial idle cycle, too. We call a busy period *feasible*, if all tasks processed during the busy period meet a fixed *service time deadline* of $T$ cycles. A sequence of feasible busy periods followed by a non-feasible busy period (containing at least one deadline violation) is called a *run*, the sequence without the last (violating) busy period is refered to by a *successful run*.

Now, it turned out that the random variable *successful run duration* $\mathcal{S}_T$ provides a suitable point of application for gaining a quality criterion for a scheduling algorithm. $\mathcal{S}_T$ is obviously the time interval from the beginning of an initial idle cycle to the beginning of the (idle) cycle initiating the busy period containing the very first violation of a task's deadline $T$. Different scheduling algorithms may be compared via the distribution of $\mathcal{S}_T$, even if the arrival process is modeled very simple (as we did): We assume an arrival process, which provides an arbitrary distributed number of task arrivals within a cycle, independent from the arrivals in the preceding cycles, and independent from the arbitrary distributed task execution times, too.

Based on this model, the first few moments of $\mathcal{S}_T$ for a number of different scheduling algorithms were analyzed in some former papers (preemptive LCFS: [**BS1**], FCFS: [**SB1**], nonpreemptive LCFS: [**SB2**]). Note that the derivations relied on the combinatorial and asymptotic analysis of certain random trees representing feasible busy periods (and not on queueing theory!). The appropriate results, however, gave us the idea that $\mathcal{S}_T$ might (always) be approximately exponentially distributed, with a parameter $\lambda_T = 1/\mu_T$ depending on the scheduling discipline considered, of course.

This paper provides a generalization of the problem based on recurrent events and, most important, rigorously justifies the assumption mentioned above. It is outlined as follows: Section 2 contains the generalization of the problem and an overview regarding the derivation of our major theorems, actually provided in Section 3. The closing Section 4 is devoted to the application of our theorems to the scheduling algorithms investigated in the papers cited above.

As a consequence, it is possible to reduce the analysis of the distribution of $\mathcal{S}_T$ for an arbitrary scheduling algorithm to the computation of $\mu_T$, a quantity concerning a *single* feasible busy period only. A single feasible busy period, however, is usually tractable by a number of powerful (combinatorial and analytical) devices from the *analysis of algorithms*

*and data structures*, as mentioned above. Therefore, our method as a whole provides a noticeable extension to the limited power of queueing theory within this contex. Note that we recently solved the old problem of analyzing the duration of the successful operation of the well-known *slotted ALOHA collision resolution algorithm* (which we found exponentially distributed, too) by a very similar approach; cf. [**DS**] and [**D**] for details.

## 2 Generalization and Overview

Consider a certain repetitive pattern $\mathcal{E}$ connected with repeated trials. $\mathcal{E}$ is called a *recurrent event* if at each occurrence of $\mathcal{E}$ the trials start from the scratch again. Thus, the waiting times $\{\mathcal{X}^{(i)};\ i = 1, 2, \dots\}$ between the $i$-th and $i+1$-th occurrence of $\mathcal{E}$ are mutually independent nonnegative random variables having the same distribution. Common examples of recurrent events are the arrival of a job in a computer system or a telephone-call arrival to a switchboard. An introduction to the appropriate field of renewal theory may be found in [**FE**], for instance. We should mention that it is sometimes convenient to view a sequence $\{\mathcal{X}^{(i)};\ i = 1, 2, \dots\}$ of mutually independent nonnegative random variables having a common distribution without regarding the associated recurrent event. Such a sequence is therefore occasionally called a *renewal process*, cf. [**TR**], for example.

A well-known recurrent event is the end of a *busy period* in queueing systems. Consider a single server system employed with servicing arriving *tasks*. Starting from an *idle* server, a task arrival causes the transition to the *busy* state. Additional tasks arriving during the service become queued (according to a specific queuing discipline) until the server has finished the whole amount of earlier work. A transition back to the idle state, i.e., the end of the busy period, occurs when the server succeeded in emptying the queue (that is, if it is not kept busy by new arrivals). Note, that we restrict ourselves to discrete time queueing systems, where all time values are an integral multiple of some unit time interval.

Providing certain independent task arrival and service time distributions, it is not difficult to imagine that the times $\{\mathcal{B}^{(i)};\ i = 1, 2, \dots\}$ between the ends of successive busy periods are indeed mutually independent and identically distributed discrete random variables. Of course, $\mathcal{B}^{(i)}$ corresponds to the length of the $i$-th busy period. Let $(\Omega, \mathcal{F}, P)$ denote a suitable common probability space for $\mathcal{B}$ (that is, for any of the $\mathcal{B}^{(i)}$); $\Omega$ is a countable set and each $\omega \in \Omega$ represents a certain busy period. The random variable $\mathcal{B}$ is a mapping $\mathcal{B} : \Omega \to \{0, 1, 2, \dots\}$, by definition.

It is conceiveable to extend the model developed so far in the following way: We define another random variable $\mathcal{C}_T : \Omega \to \{0, 1\}$ on the probability space $(\Omega, \mathcal{F}, P)$, which depends on an arbitrary $T \in \{1, 2, \dots\}$. It provides a certain *interpretation* of a busy period $\omega \in \Omega$: if $\omega$ satisfies a certain condition (which depends on $T$) we define $\mathcal{C}_T(\omega) = 0$. otherwise $\mathcal{C}_T(\omega) = 1$. Obviously, the interpretation may be defined by the corresponding partition of $\Omega$ into sets $\Omega_{\mathcal{C}_T=0} = \Omega_T$ and the complement $\Omega_{\mathcal{C}_T=1} = \Omega_T^c$ as well. An example is the *feasibility* of a busy period with regard to *service time*† *deadlines*: $\mathcal{C}_T^{(i)} = 0 \iff$ all tasks serviced during the $i$-th busy period have a service time less than $T$ time units.

†In classical queueing theory, service time denotes the time a task occupies the server. We prefer the term *task execution time* for this quantity; the *service time* of a task denotes the time interval between the arrival of the task and its completion, i.e., the time spent in the queue plus the execution time.

That is, we are faced with a sequence of independent and identically distributed two-dimensional random variables $\{(\mathcal{B}^{(i)}, \mathcal{C}_T^{(i)});\ i = 1, 2, \dots\}$, where $\mathcal{B}^{(i)}$ represents the length of the $i$-th busy period and $\mathcal{C}_T^{(i)}$ its feasibility. Within this context, the following question is of importance: How long does it last until the first non-feasible busy period is encountered?

The rest of this section is devoted to some preliminaries and an informal overview to our derivations, which establish the following general result: The distribution of the random variable mentioned above is approximately exponential with a certain parameter $\lambda_T$, getting small for large $T$.

We start from an arbitrary discrete nonnegative random variable $\mathcal{B}$ defined on an probability space $(\Omega, \mathcal{F}, P)$. The appropriate probability distribution $P[\mathcal{B}]$ is defined via the probabilities

$$b_n = \text{prob}\{\mathcal{B} = n\}. \tag{2.1}$$

Regarding our example, we have $b_n = \text{prob}\{\text{a busy period has length } n\}$.

Let $\Omega_T \subseteq \Omega$, $\Omega_T \neq \Omega$ for $T \in \{1, 2, \dots\}$ denote an arbitrary sequence of subsets of $\Omega$ with the property that $\lim_{T\to\infty} \Omega_T = \Omega$‡. For any fixed $T$, the conditional probability distribution $P[\mathcal{B}|\Omega_T]$ is given by

$$b_{n,T} = \text{prob}\{\omega : \mathcal{B}(\omega) = n \,|\, \omega \in \Omega_T\}. \tag{2.2}$$

In our example, we have $b_{n,T} = \text{prob}\{\text{a feasible busy period has length } n\}$.

It is obvious that the "probabilistic condition" $\lim_{T\to\infty} \Omega_T = \Omega$ implies the convergence $b_{n,T} \to b_n$ for all $n$. Moreover, $b_{n,T} \leq b_n$ for all $n$ and $T$, and $\sum_{n\geq 0} b_{n,T} < \sum_{n\geq 0} b_n = 1$ for all finite $T$. Since our subsequent treatment is solely based on analytical methods, we will rely on the convergence condition

$$\lim_{T\to\infty} b_{n,T} = b_n \qquad \text{for all } n \tag{2.3}$$

in conjunction with some additional assumptions. However, all our further results remain valid under the probabilistic preconditions mentioned above.

The corresponding probability generating functions (PGFs) read

$$B(z) = \sum_{n\geq 0} b_n z^n \tag{2.4}$$

and

$$B_T(z) = \sum_{n\geq 0} b_{n,T} z^n. \tag{2.5}$$

First, we show the equivalence of condition (2.3) and a limiting condition on the corresponding PGFs†.

---

‡A sequence of sets $\{A_n\}_{n\geq 1}$ has limit $\lim_n A_n = A$, written $A_n \to A$, if $\liminf_n A_n = \limsup_n A_n = A$ where $\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$ and $\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$, as usual.

†This is a well-known fact, cf. [**FE**, p.280] for a similar *continuity theorem*.

LEMMA 2.1 (CONTINUITY LEMMA). *Under the basic assumption $b_{n,T} \leq b_n$ for all $n$ and $T$, the following holds:* $\limsup_{n\to\infty} \sqrt[n]{|b_n|} = 1/R$, $0 < R \leq \infty$, *and* $\lim_{T\to\infty} b_{n,T} = b_n$ *for all $n$ if and only if $B(z)$ has radius of convergence $R$ and $\lim_{T\to\infty} B_T(z) = B(z)$ uniformly for all $|z| \leq r < R$.*

**Proof:** Since $\limsup_{n\to\infty} \sqrt[n]{|b_{n,T}|} \leq \limsup_{n\to\infty} \sqrt[n]{|b_n|} = 1/R$ for all $T$, $B(z)$ has radius of convergence $R$ and $B_T(z)$ has radius of convergence which is at least $R$ for all $T$.

Now, for any $\varepsilon > 0$ it is possible to find some $N = N(\varepsilon, r)$ such that $\sum_{n>N} b_n r^n < \varepsilon$. Thus, if $|z| \leq r < R$ then

$$\begin{aligned} |B(z) - B_T(z)| &\leq \sum_{n=0}^{N} (b_n - b_{n,T}) r^n + \sum_{n>N} b_n r^n + \sum_{n>N} b_{n,T} r^n \\ &< \sum_{n=0}^{N} (b_n - b_{n,T}) r^n + 2\varepsilon \\ &< 3\varepsilon, \end{aligned}$$

provided that $T$ is chosen large enough to guarantee the $\varepsilon$-bound for the finite sum above, too. Therefore, $\lim_{T\to\infty} B_T(z) = B(z)$ uniformly for $|z| \leq r < R$.

On the other hand, the convergence $\lim_{T\to\infty} b_{n,T} = b_n$ for all $n$ is most easily seen by considering the trivial inequality

$$0 \leq (b_n - b_{n,T}) x^n \leq B(x) - B_T(x),$$

for an arbitrary $0 < x \leq r$, since the right hand side tends to 0 for $T \to \infty$, too.

Note, that the convergence $B_T(x) \to B(x)$ for an arbitrary argument value $x > 0$ suffices to conclude $b_{n,T} \to b_n$ for all $n$. However, the assumption $b_{n,T} \leq b_n$ for all $n$ and $T$ is crucial here.

This completes the proof of Lemma 2.1. ∎

A simple corollary of the lemma above (in conjunction with the well-known Weierstraßian double series theorem, cf. [**HE**, pp.133–136]) is the fact that $B_T^{(m)}(z) \to B^{(m)}(z)$ uniformly for $|z| \leq r < R$ and $T \to \infty$.

We require the following (technical) conditions for our derivations:

(1) $\lim_{T\to\infty} b_{n,T} = b_n$ for all $n \geq 0$.
(2) $b_{n,T} \leq b_n$ for all $n$ and $T$.
(3) $B_T(1) < B(1) = 1$ for all finite $T$.
(4) The radius of convergence $R$ of $B(z)$ should be larger than 1. Condition (3) only implies $R \geq 1$; however, this is no serious restriction for our practical purposes. Note, that by virtue of condition (2) the radius of convergence of $B_T(z)$ is $R_T \geq R$.
(5) $B'(1) > 0$, which is equivalent to $B(z) \not\equiv 1$, that is, $b_0 < 1$. This condition has a probabilistic meaning: the expectation of $\mathcal{B}$ should be greater than zero.

Note, that this condition and $B'_T(1) \to B'(1)$ for $T \to \infty$ according to the corollary above reveals $B'_T(1) \geq \epsilon > 0$ for $T$ sufficiently large.

(6) $d = \gcd\{n : b_n > 0\} = 1$, which ensures that $|B(z)| < B(|z|)$ for any $z$ not real and positive. The probabilistic interpretation is a certain non-periodicity property of the corresponding renewal process. Note however, that this is no real restriction for our derivations since $d > 1$ implies $B(z) = \overline{B}(z^d) = \sum_{k\geq 0} \overline{b}_k z^{dk}$ with $\overline{b}_k = b_{dk}$ and $\gcd\{k : \overline{b}_k > 0\} = 1$. Thus, we could use $\overline{B}(z)$ instead of $B(z)$.

Due to the convergence $B_T(z) \to B(z)$ it is obvious that the condition on $B(z)$ carries over to a condition on $B_T(z)$, namely $\gcd\{n : b_{n,T} > 0\} = 1$ for $T$ sufficiently large.

The random variable $\mathcal{S}_T$ in question is defined on a probability space $(\Omega', \mathcal{F}', P')$, where $\Omega' = \{\omega_1\omega_2 \ldots \omega_n\omega_{n+1} : \omega_i \in \Omega_T \text{ for } 1 \leq i \leq n,\ n \geq 0 \text{ and } \omega_{n+1} \in \Omega_T^c\}$ consists of arbitrary sequences of $\omega \in \Omega_T$ terminated† by a single $\omega^c \in \Omega_T^c$. The field $\mathcal{F}'$ and the probability measure $P'$ are the obvious extensions of $\mathcal{F}$ and $P$.

To be more specific, we define $\mathcal{S}_T : \Omega' \to \{0, 1, \ldots\}$ for $\omega' = \omega_1 \ldots \omega_n\omega_{n+1} \in \Omega'$ by

$$\mathcal{S}_T(\omega_1 \ldots \omega_n\omega_{n+1}) = \sum_{i=1}^{n} \mathcal{B}^{(i)}(\omega_i). \tag{2.6}$$

Note, that we omit the terminating $\omega_{n+1} \in \Omega_T^c$ in the summation above.

Regarding our introductional example, $\mathcal{S}_T$ denotes the sum of the lengths of an arbitrary number of feasible busy periods followed by a single non-feasible busy period. The terminating non-feasible busy period is not taken into account.

The probablility distribution $P[\mathcal{S}_T]$, i.e., the probabilities

$$s_{n,T} = \text{prob}\{\mathcal{S}_T = n\}, \tag{2.7}$$

are uniquely determined by the appropriate PGF, which reads

$$S_T(z) = \sum_{n\geq 0} s_{n,T} z^n = \frac{1 - B_T(1)}{1 - B_T(z)}. \tag{2.8}$$

This follows from the fact that

(1) the PGF of a sum of an arbitrary number of mutually exclusive random variables with PGF $B_T(z)$ is $\sum_{i\geq 0} B_T(z)^i$

(2) the probability of the terminating $\omega^c \in \Omega_T^c$ equals $1 - B_T(1)$.

After these preliminary discussions we give a short and informal overview to the following treatment, which establishes that the distribution of $\mathcal{S}_T$ for $T \to \infty$ is approximately exponential with parameter $\lambda_T = 1/\mu_T$. The latter is evaluated to

$$\mu_T = \frac{B_T'(1)}{1 - B_T(1)}, \tag{2.9}$$

†Note the resemblance with geometric trials.

which tends to infinity for $T \to \infty$ due to our conditions on $B(z)$. We provide two different ways for establishing the result:

(1) Using a well-known continuity theorem for characteristic functions we show the weak convergence (convergence in distribution) of the distribution of the (normalized) random variable $\mathcal{Y}_T = \mathcal{S}_T/\mu_T$ to the exponential distribution with parameter $\lambda = 1$. The related mathematical treatment is simple, but unfortunately it provides no informations concerning the quality of convergence.

(2) Using singularity analysis on $S_T(z)$ yields uniform asymptotic expressions for $s_{n,T}$ and $\sum_{k=0}^{n} s_{k,T}$ when $n \to \infty$ and $T \to \infty$. In addition to the general statement (i.e., the exponential type limiting distribution) we obtain remainder terms which provide the required informations concerning the quality of convergence. Moreover, using Mellin transform techniques, we derive uniform asymptotic expansions for the $m$-th moment $E[\mathcal{S}_T^m]$ of $\mathcal{S}_T$, too.

## 3. Major Results

First, we should mention that we rely on the notations and conditions stated in Section 2 (unless otherwise noticed). That is, we avoid their reformulation in the following theorems for the sake of readability.

The following theorem states the weak convergence (that is, convergence in distribution, cf. [BI, p.338]) of the probability distribution of the normalized discrete random variable $\mathcal{Y}_T = \mathcal{S}_T/\mu_T$ to the exponential distribution with parameter $\lambda = 1$. Note, that if a random variable $c\mathcal{X}$, $c > 0$, is exponentially distributed with parameter $\lambda$, then $\mathcal{X}$ is exponentially distributed with parameter $\lambda/c$.

Theorem 3.1 (Weak Convergence Property). *The distribution of the normalized random variable $\mathcal{Y}_T = \mathcal{S}_T/\mu_T$ converges weakly to the exponential distribution with parameter $\lambda = 1$.*

**Proof:** Using (2.8), the characteristic function $Y_T^*(t)$ of $\mathcal{Y}_T$ evaluates to

$$Y_T^*(t) = S_T(e^{it/\mu_T}) = \frac{1 - B_T(1)}{1 - B_T(e^{it/\mu_T})}.$$

The Taylor expansions of $B_T(z)$ around $z = 1$ reads

$$B_T(z) = B_T(1) + B_T'(1)(z-1) + O((z-1)^2) \qquad \text{for } z \to 1, \tag{3.1}$$

and it is easy to show that the remainder term is uniform for $T \to \infty$. Essentially, it is only necessary to use the fact that the remainder is bounded by a function involving $B_T''(z)$, which converges uniformly to $B''(z)$ for $|z| \le r < R$ according to the corollary of Lemma 2.1.

Providing the straightforward expansion

$$e^{it/\mu_T} = 1 + it/\mu_T + O(t^2/\mu_T^2) \qquad \text{for } t = o(\mu_T),$$

we obtain

$$\begin{aligned} Y_T^*(t) &= \frac{1 - B_T(1)}{1 - B_T(1) - B_T'(1)(e^{it/\mu_T} - 1) + O\big((e^{it/\mu_T} - 1)^2\big)} \\ &= \frac{1}{1 - \frac{B_T'(1)}{1-B_T(1)}\big(\frac{it}{\mu_T} + O(\frac{t^2}{\mu_T^2})\big) + O\big(\frac{1}{1-B_T(1)}\frac{t^2}{\mu_T^2}\big)} \\ &= \frac{1}{1 - it + O(t^2/\mu_T)} \qquad \text{for } t = o(\mu_T) \text{ and } T \to \infty \end{aligned}$$

by using (2.9) and condition (5). Thus, for any $t$ fixed, the expression above converges to $1/(1-it)$. This function is continuous at $t = 0$ and represents the characteristic function of the exponential distribution with parameter $\lambda = 1$. The application of the continuity theorem of characteristic functions (cf. [**BI**, p.359]) completes the proof. ∎

Next, we will perform singularity analysis on (2.8) in order to obtain asymptotic results for $s_{n,T}$. We start with

LEMMA 3.2 (SINGULARITIES OF $S_T(z)$). *The function $S_T(z)$ has radius of convergence*

$$\phi_T = 1 + 1/\mu_T + O(1/\mu_T^2) = e^{\mu_T^{-1}(1+O(1/\mu_T))} \qquad \text{for } T \to \infty,$$

*which is the solution of $B_T(\phi_T) = 1$ for $T$ sufficiently large. In addition, $S_T(z)$ has only a simple pole $z = \phi_T$ on its circle of convergence. The residue is*

$$\operatorname{res}\big(S_T(z); z = \phi_T\big) = -\frac{1 - B_T(1)}{B_T'(\phi_T)} = -\mu_T^{-1}\big(1 + O(1/\mu_T)\big) \qquad \text{for } T \to \infty.$$

**Proof:** Since $B_T(x)$ is a monotonically increasing function for positive $x$, it is clear from condition (3) that $S_T(z)$ has a pole at $z = \phi_T > 1$ resulting from the vanishing denominator $1 - B_T(z)$. Note, that $R > \phi_T$ for $T$ sufficiently large avoids difficulties with additional singularities originated in $B_T(z)$ itself. That is, there are no other singularities within the disk $|z| < \phi_T$. Moreover, condition (6) ensures that there are no other zeros of $1 - B_T(z)$ for $|z| = \phi_T$ but $z \neq \phi_T$, too.

An asymptotic expression for $\phi_T$ is easily determined via bootstrapping. Using simple geometric arguments, we find $1 < \phi_T \leq 1 + \big(1 - B_T(1)\big)/B_T'(1) = 1 + 1/\mu_T$. Hence, we provide the (coarse) first estimation

$$\phi_T = 1 + \psi_T \quad \text{where} \quad \psi_T = O(1/\mu_T) \text{ for } T \to \infty$$

and use the Taylor expansion of $B_T(z)$ at $z = 1$ to improve it. Substituting $z = \phi_T$ in

$$1 - B_T(z) = 1 - B_T(1) - B_T'(1)(z-1) + O\big((z-1)^2\big) \qquad \text{for } z \to 1$$

yields

$$\psi_T = \frac{1 - B_T(1)}{B_T'(1)} + O\big(\psi_T^2/B_T'(1)\big) = 1/\mu_T + O(1/\mu_T^2) \qquad \text{for } T \to \infty.$$

Note, that the remainder term is uniformly in $T$, cf. condition (5) and the remark on equation (3.1). Using $e^x = 1 + x + O(x^2)$ for small $x$, we find

$$\phi_T = 1 + 1/\mu_T + O(1/\mu_T^2) = e^{\mu_T^{-1}(1+O(1/\mu_T))} \qquad \text{for } T \to \infty$$

as asserted.

The asymptotic expansion for $S_T(z)$ near $z = \phi_T$ is easily derived by substituting the Taylor expansion

$$B_T(z) = 1 + B_T'(\phi_T)(z - \phi_T) + O\big((z - \phi_T)^2\big) \qquad \text{for } z \to \phi_T$$

in (2.8). By the same arguments as in the proof of Theorem 3.1 it follows that the remainder term is uniformly in $T$, cf. the remark on equation (3.1). We obtain

$$\begin{aligned} S_T(z) &= \frac{1 - B_T(1)}{-B_T'(\phi_T)(z - \phi_T)\big(1 + O(z - \phi_T)\big)} \\ &= -\frac{1 - B_T(1)}{B_T'(\phi_T)}\left(\frac{1}{z - \phi_T} + O(1)\right) \end{aligned}$$

for $z \to \phi_T$ and $T \to \infty$. Obviously, the uniform remainder $O(1)$ represents a function analytic at $z = \phi_T$. Using the Taylor expansion of $B_T'(z)$ at $z = 1$, namely $B_T'(z) = B_T'(1) + O(z-1)$ for $z \to 1$, we find $B_T'(\phi_T) = B_T'(1) + O(1/\mu_T)$ for $T \to \infty$. Substituting this in the expression above yields

$$S_T(z) = -\frac{1 - B_T(1)}{B_T'(1)}\big(1 + O(1/\mu_T)\big)\left(\frac{1}{z - \phi_T} + O(1)\right)$$

for $z \to \phi_T$ and $T \to \infty$, which completes proof of the lemma. ∎

**Remark:** It is easy to provide a more accurate asymptotic expansion for $\phi_T$ by additional bootstrapping steps. For example, a second step yields

$$\phi_T = 1 + \mu_T^{-1} - \frac{B_T''(1)}{2B_T'(1)}\mu_T^{-2} + O(\mu_T^{-3})$$

for $T \to \infty$. Thus, we could improve all our subsequent asymptotic results which involve this quantity.

Lemma 3.2 reveals that the Taylor coefficients $s_{n,T}$ of $S_T(z)$ are mainly determined by the simple polar singularity at $z = \phi_T$. Their asymptotics are easily evaluated via Cauchy's formula. However, some additional investigations are necessary in order to obtain error terms which are uniform in $T$.

We will point out that additional polar singularities of $S_T(z)$, i.e., points $|z_0| > \phi_T$ providing $1 - B_T(z_0) = 0$, lie completely outside the closed disk $\overline{D}(0, 1+\delta)$ for some $\delta > 0$. In addition, for $T$ sufficiently large, we will establish a uniform bound $|1 - B_T(z)| \geq \Delta > 0$ for $|z| = 1 + \delta$.

LEMMA 3.3 (UNIFORM BOUND). *There exists some $\delta > 0$ such that $1 - B_T(z) \neq 0$ for $|z| \leq \delta$ (with the only exception of the point $z = \phi_T$) for $T$ sufficiently large. Moreover, we have the uniform bound*

$$\min_{|z|=1+\delta} |1 - B_T(z)| \geq \Delta$$

*for some $\Delta > 0$ and $T$ sufficiently large.*

**Proof:** Since $B(1) = 1$ and $B'(1) \neq 0$ there exists a neighborhood of $z = 1$, i.e., an open disk $D(1, \eta)$, $\eta > 0$ where

$$B(z) \neq 1 \qquad \text{for } z \in D(1,\eta)\backslash\{1\}, \tag{3.2}$$

by virtue of the implicit function theorem.

Due to condition (6) we have $|B(z)| < 1$ for $|z| = 1$, $z \neq 1$ and hence $\max|B(z)| < 1$ when $|z| = 1$ and $|z - 1| \geq \eta/2$. But, continuity of $B(z)$ implies that there exists some $\delta > 0$ (restricted to $1 + \delta < R$, of course) such that the inequality above remains valid for $|z| = 1 + \delta$. We obtain

$$\max_{\substack{|z|=1+\delta \\ |z-1|\geq\eta/2}} |B(z)| \leq \Delta_0$$

for some $\Delta_0 < 1$.

Since $B_T(z) \to B(z)$ uniformly, it is clear that this property of $B(z)$ carries over to a property of $B_T(z)$ for $T$ sufficiently large, i.e., that there exists some $T_0$ such that

$$\max_{\substack{|z|=1+\delta \\ |z-1|\geq\eta/2}} |B_T(z)| \leq \Delta_0/2 < 1 \qquad \text{for all } T > T_0. \tag{3.3}$$

Provided that our constant $\delta$ is chosen not too large, that is, $\delta < \eta/2$, we obviously have $\{z : |z| = 1 + \delta\} \cap \overline{D}(1, \eta/2) \neq \emptyset$. Remembering (3.2) it is clear that

$$\min_{\substack{|z|=1+\delta \\ z\in\overline{D}(1,\eta/2)}} |1 - B(z)| \geq \Delta_1 > 0.$$

By the same argument as before it follows that there exists some $T_1$ such that

$$\min_{\substack{|z|=1+\delta \\ z\in\overline{D}(1,\eta/2)}} |1 - B_T(z)| \geq \Delta_1/2 > 0. \tag{3.4}$$

Putting inequality (3.3) into the minimum form like (3.4) it is obvious that

$$\min_{|z|=1+\delta} |1 - B_T(z)| \geq \Delta > 0$$

for $T > T' = \max\{T_0, T_1\}$, where $\Delta = \min\{\Delta_0/2, \Delta_1/2\}$. This completes the proof of the lemma. ∎

With this preparations we are able to state

THEOREM 3.4 (ASYMPTOTICS OF $s_{n,T}$). *There exists some $\delta > 0$ such that the $n$-th Taylor coefficient $s_{n,T}$ of $S_T(z)$ has the uniform asymptotic expansion*

$$\begin{aligned} s_{n,T} &= \frac{1 - B_T(1)}{\phi_T B_T'(\phi_T)} \phi_T^{-n} + O\big(\mu_T^{-1}(1+\delta)^{-n}\big) \\ &= \mu_T^{-1}\big(1 + O(1/\mu_T)\big) e^{-\mu_T^{-1}(1+O(1/\mu_T))n} + O\big(\mu_T^{-1}(1+\delta)^{-n}\big) \end{aligned}$$

*for $n \to \infty$ and $T \to \infty$.*

**Proof:** Using Cauchy's formula, we have

$$s_{n,T} = \frac{1}{2\pi i} \int_{\mathcal{C}:|z|=\gamma} \frac{S_T(z)}{z^{n+1}}\, dz$$

for some $\gamma < 1$, for example. Extending $\mathcal{C}$ to the circle $|z| = 1 + \delta$, we have to add the residue of the integrand at the newly enclosed† singularity $z = \phi_T$. Thus, we obtain

$$\frac{1}{2\pi i} \int_{\mathcal{C}':|z|=1+\delta} \frac{S_T(z)}{z^{n+1}}\, dz = s_{n,T} + \operatorname{res}\big(S_T(z); z = \phi_T\big) \phi_T^{-n-1}.$$

But, the integral on the left hand side is trivially bounded by $(1+\delta)^{-n} \max_{|z|=1+\delta} S_T(z)$. Substituting the uniform bound of Lemma 3.3 in (2.8) we find

$$\max_{|z|=1+\delta} S_T(z) \le \frac{1 - B_T(1)}{\Delta}$$

for $T$ sufficiently large. Remembering the result of Lemma 3.2 and (2.9), we finally obtain

$$\begin{aligned} s_{n,T} &= -\operatorname{res}\big(S_T(z); z = \phi_T\big) \phi_T^{-1} \phi_T^{-n} + O\left(\frac{1 - B_T(1)}{(1+\delta)^n}\right) \\ &= \mu_T^{-1}\big(1 + O(1/\mu_T)\big) e^{-\mu_T^{-1}(1+O(1/\mu_T))n} + O\big(\mu_T^{-1}(1+\delta)^{-n}\big) \qquad \text{for } n \to \infty, \end{aligned}$$

where the remainder is uniformly in $T$, of course. ∎

By the same devices it is possible to prove a similar theorem concerning the distribution function of $\mathcal{S}_T$, that is, $v_{n,T} = \operatorname{prob}\{\mathcal{S}_T \le n\} = \sum_{k=0}^{n} s_{k,T}$.

THEOREM 3.5 (ASYMPTOTICS OF $\sum s_{k,T}$). *There exists some $\delta > 0$ such that the distribution function $v_{n,T} = \sum_{k=0}^{n} s_{k,T}$ of $\mathcal{S}_T$ has a uniform asymptotic expansion*

$$\begin{aligned} v_{n,T} &= 1 - \frac{1 - B_T(1)}{(1-\phi_T)\phi_T B_T'(\phi_T)} \phi_T^{-n} + O\big(\mu_T^{-1}(1+\delta)^{n}\big) \\ &= 1 - \big(1 + O(1/\mu_T)\big) e^{-\mu_T^{-1}(1+O(1/\mu_T))n} + O\big(\mu_T^{-1}(1+\delta)^{-n}\big) \end{aligned}$$

†For $T$ sufficiently large, we obviously have $\phi_T < 1 + \delta$.

*for $n \to \infty$ and $T \to \infty$.*

**Proof:** The GF $V_T(z)$ of $v_{n,T}$ is obviously

$$V_T(z) = \frac{1}{1-z} S_T(z) = \frac{1}{1-z} \cdot \frac{1 - B_T(1)}{1 - B_T(z)}, \tag{3.5}$$

according to (2.8). Thus, $V_T(z)$ has an (additional) simple pole at $z = 1$, the residue is

$$\operatorname{res}\big(V_T(z); z = 1\big) = -1.$$

The next singularity $z = \phi_T$ comes from $S_T(z)$, its residue reads

$$\operatorname{res}(V_T(z); z = \phi_T) = \frac{\operatorname{res}\big(S_T(z); z = \phi_T\big)}{1 - \phi_T} = 1 + O(1/\mu_T) \qquad \text{for } T \to \infty$$

due to Lemma 3.2.

The rest of the proof runs along the proof of Theorem 3.3: Using Cauchy's formula, we have

$$v_{n,T} = \frac{1}{2\pi i} \int_{\mathcal{C}:|z|=\gamma} \frac{V_T(z)}{z^{n+1}}\, dz$$

for some $\gamma < 1$, for example. Extending $\mathcal{C}$ to the circle $|z| = 1 + \delta$, we have to add the residues of the integrand at the newly enclosed singularities $z = 1$ and $z = \phi_T$. Thus, we obtain

$$\frac{1}{2\pi i} \int_{\mathcal{C}':|z|=1+\delta} \frac{V_T(z)}{z^{n+1}}\, dz = v_{n,T} + \operatorname{res}(V_T(z); z = 1) + \operatorname{res}(V_T(z); z = \phi_T)\phi_T^{-n-1}.$$

Again, the integral on the left hand side is trivially bounded by $(1+\delta)^{-n} \max_{|z|=1+\delta} V_T(z)$. Substituting the uniform bound of Lemma 3.3 in (3.5) we find

$$\max_{|z|=1+\delta} V_T(z) \le \frac{1 - B_T(1)}{\delta \Delta}$$

for $T$ sufficiently large. We finally obtain

$$\begin{aligned} v_{n,T} &= -\operatorname{res}\big(V_T(z); z = 1\big) - \operatorname{res}\big(V_T(z); z = \phi_T\big)\phi_T^{-1}\phi_T^{-n} + O\left(\frac{1 - B_T(1)}{(1+\delta)^n}\right) \\ &= 1 - \big(1 + O(1/\mu_T)\big)e^{-\mu_T^{-1}(1+O(1/\mu_T))n} + O\big(\mu_T^{-1}(1+\delta)^{-n}\big) \qquad \text{for } n \to \infty, \end{aligned}$$

where the remainder is uniformly in $T$, of course. ∎

Theorem 3.4 provides all informations required for the computation of the $m$-th moment $E[\mathcal{S}_T^m]$ of $\mathcal{S}_T$. Using Mellin transform techniques, it is easy to obtain asymptotic expansions which are uniformly in $m$. We start with the following

LEMMA 3.6 (A BIVARIATE ASYMPTOTIC EXPANSION).

$$f_m(\alpha) = \sum_{n\geq 1} n^m e^{-\alpha n} = \frac{m!}{\alpha^{m+1}} - \frac{B_{m+1}}{m+1} + \frac{B_{m+2}}{m+2}\alpha + O\Big(\frac{m!}{(2\pi)^m}(m\alpha)^{3/2}\Big)$$

*uniformly for* $\alpha \to 0+$ *and* $m \to \infty$, $m \geq 1$. $B_m$ *denotes the* $m$*-th Bernoulli number, that is, the coefficient of* $[z^n/n!]$ *in* $z/(e^z-1)$. *Note, that* $B_n = 0$ *for* $n = 2k+1$, $k \geq 1$ *and* $|B_n| < 4n!/(2\pi)^n$.

**Proof:** The function $f_m(\alpha)$ is a so-called *harmonic sum*, which is tractable by Mellin transform techniques, cf. [**VF**] for an introduction. The Mellin transform

$$\begin{aligned} f_m^*(s) &= \int_0^\infty f_m(\alpha)\alpha^{s-1}\,d\alpha \\ &= \sum_{n\geq 1} n^m \int_0^\infty e^{-\alpha n}\alpha^{s-1}\,d\alpha = \sum_{n\geq 1} n^m n^{-s}\Gamma(s) \\ &= \Gamma(s)\zeta(s-m) \end{aligned}$$

is analytic within the *fundamental strip* $\Re(s) > m+1$ and has a simple pole resulting from Riemann's zeta function $\zeta(z)$ at $s = m+1$. Using the well-known inversion formula

$$f_m(\alpha) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} f_m^*(s)\alpha^{-s}\,ds,$$

where $c > m+1$ denotes an arbitrary real constant within the fundamental strip, an asymptotic expansion of $f_m(\alpha)$ for $\alpha \to 0+$ is obtained by extending the contour to a closed rectangle (beyond the left of the fundamental strip) and taking into account the residues of the enclosed singularities. We therefore have

$$\begin{aligned} f_m(\alpha) = {} & \operatorname{res}\big(f_m^*(s)\alpha^{-s}; s = m+1\big) + \operatorname{res}\big(f_m^*(s)\alpha^{-s}; s = 0\big) + \operatorname{res}\big(f_m^*(s)\alpha^{-s}; s = -1\big) \\ & - \frac{1}{2\pi i}\int_{c+i\infty}^{-3/2+i\infty} f_m^*(s)\alpha^{-s}\,ds \\ & - \frac{1}{2\pi i}\int_{-3/2+i\infty}^{-3/2-i\infty} f_m^*(s)\alpha^{-s}\,ds \\ & - \frac{1}{2\pi i}\int_{-3/2-i\infty}^{c-i\infty} f_m^*(s)\alpha^{-s}\,ds. \end{aligned} \tag{3.6}$$

The residues are easily evaluated. Since $\zeta(z)$ has a simple pole with residue 1 at $z = 1$, we obtain

$$\operatorname{res}\big(f_m^*(s)\alpha^{-s}; s = m+1\big) = \frac{m!}{\alpha^{m+1}}. \tag{3.7}$$

The singularities at $s = 0$ and $s = -1$ are simple poles resulting from $\Gamma(s)$. However, if $m = 2k$, $k \geq 1$ is even, the singularity at $s = 0$ is removeable since $\zeta(z) = 0$ for $z = -2k$. Similarily, the singularity at $s = -1$ is removeable if $m = 2k-1$, $k \geq 1$.

Using the well-known values (cf. [**AS**])

$$\zeta(1-2k) = -\frac{B_{2k}}{2k} \qquad \text{for } k \geq 1,$$

where $B_n$ denotes the $n$-th Bernoulli number, and the fact that

$$\operatorname{res}\big(\Gamma(z); z = -n\big) = \frac{(-1)^n}{n!} \qquad \text{for } n \geq 0,$$

we obtain

$$\operatorname{res}\big(f_m^*(s)\alpha^{-s}; s = 0\big) = -\frac{B_{2k}}{2k} = -\frac{B_{m+1}}{m+1} \tag{3.8}$$

for $m = 2k-1$, $k \geq 1$. However, formula (3.8) is valid for all $m \geq 1$ since $B_{m+1} = 0$ for $m = 2k$, $k \geq 1$. Similarily,

$$\operatorname{res}\big(f_m^*(s)\alpha^{-s}; s = -1\big) = \frac{B_{2(k+1)}}{2(k+1)}\alpha = \frac{B_{m+2}}{m+2}\alpha \tag{3.9}$$

for $m = 2k$, $k \geq 1$ at first, remaining valid for all $m \geq 1$ as before.

The last task is the estimation of the contour integrals in (3.6). We will use the intuitively meaningful abbreviations $I_{\leftarrow}$, $I_{\downarrow}$, and $I_{\rightarrow}$, respectively, and proceed with treating $I_{\downarrow}$. Using the functional equation for $\zeta(z)$, namely

$$\zeta(z) = 2^z \pi^{z-1} \sin(\pi z/2)\Gamma(1-z)\zeta(1-z)$$

we find

$$\begin{aligned} f_m^*(s) &= \Gamma(s)\zeta(s-m) \\ &= \frac{2^s \pi^{s-1}}{(2\pi)^m}\Gamma(s)\sin\big(\pi(s-m)/2\big)\Gamma(m+1-s)\zeta(m+1-s). \end{aligned} \tag{3.10}$$

Due to the estimate

$$\big|\sin\big(\pi(x-m+iy)/2\big)\big| \leq \cosh(\pi y/2) \leq e^{\pi|y|/2} \tag{3.11}$$

for real $x, m, y$, the functional relation $\Gamma(z+1) = z\Gamma(z)$, and $|\Gamma(1/2+iy)|^2 = \pi/\cosh(\pi y)$ we obtain

$$\begin{aligned} |\Gamma(-3/2+iy)\sin\big(\pi(x-m+iy)/2\big)| &\leq \frac{\sqrt{\pi}}{|-1/2+iy||-3/2+iy|} \cdot \frac{\cosh(\pi y/2)}{\cosh(\pi y)^{1/2}} \\ &= \frac{2\sqrt{\pi}}{(1+4y^2)^{1/2}(9+4y^2)^{1/2}} \left(\frac{\cosh(\pi y)+1}{2\cosh(\pi z)}\right)^{1/2} \\ &\leq \frac{2\sqrt{\pi}}{1+4y^2}. \end{aligned} \tag{3.12}$$

From $|\Gamma(x+iy)| \le |\Gamma(x)|$ and the well-known limes relation $\Gamma(z+a)/\Gamma(z+b) \sim z^{a-b}$ for $|z| \to \infty$ (which is valid for all values of $z$ bounded away from singular points of the functions involved) follows

$$|\Gamma(m+1+3/2-iy)| \le C_1\Gamma(m+2)\sqrt{m+5/2} \le C_2(m+1)!\sqrt{m+1} \tag{3.13}$$

for some constant $C_2$. At last, it is easy to see that

$$|\zeta(m+1+3/2-iy)| \le \sum_{n\ge 1} n^{-(m+5/2)} \le 2. \tag{3.14}$$

Substituting (3.12), (3.13) and (3.14) in (3.10), we obtain

$$\begin{aligned} |I_\downarrow| &\le \int_{-\infty}^{+\infty} |f_m^*(-3/2+it)\alpha^{3/2-it}|\,|dt| \\ &\le C_3\frac{(m+1)!\sqrt{m+1}}{(2\pi)^{m+1}}\alpha^{3/2}\cdot 2\int_0^\infty \frac{dt}{1+4y^2} \\ &= O\left(\frac{(m+1)!\sqrt{m+1}}{(2\pi)^{m+1}}\alpha^{3/2}\right) = O\left(\frac{m!}{(2\pi)^m}(m\alpha)^{3/2}\right) \end{aligned} \tag{3.15}$$

since the (arcus tangens) integral in the second line above obviously converges.

Finally, we have to deal with the contributions of $I_{\leftarrow}$ and $I_{\rightarrow}$, which will be shown to be neglible. We start with an additional estimation on $\Gamma(z)$ for large imaginary parts. By virtue of a weak form of Stirlings formula, namely

$$\Gamma(z) = O(e^{-z}z^{z-1/2}) \qquad \text{for } |z| \to \infty,$$

which is valid for $|\arg(z)| \le \pi - \delta$ and $|z| > 0$, we obtain

$$\begin{aligned} |\Gamma(x+iy)| &\le C_4 e^{-x} e^{(x-1/2)\log\sqrt{x^2+y^2} - y\arctan(y/x)} \\ &= C_4 e^{-x+(x-1/2)\log|y| + (x-1/2)\log\sqrt{1+x^2/y^2} - |y|(\pi/2 - x/|y| + O(x^2/y^2))} \\ &\le C_4|y|^{x-1/2}e^{-|y|\pi/2}e^{O(x^2/y^2)+O(x^2/y)} \\ &\le C_5|y|^{x-1/2}e^{-|y|\pi/2} \end{aligned}$$

provided that $x = o(\sqrt{y})$ as $|y| \to \infty$. In addition, by weakening well-known estimations concerning $\zeta(z)$ for large imaginary parts (cf. [**WW**, p.276]) we have

$$\zeta(x-m+iy) \le C_6|y|^{3/2-x+m} \qquad \text{for } |y| \to \infty.$$

Thus, choosing $c = m+3/2 > m+1$ we find

$$\begin{aligned} |I_{\leftarrow}| &= \lim_{t\to\infty}\left|\int_c^{-3/2} f_m^*(x+it)\alpha^{-x-it}\,dx\right| \\ &\le \lim_{t\to\infty} C_7|t|^{m+1}e^{-|t|\pi/2}\alpha^{-m-3/2}\cdot(m+3) \\ &= 0, \end{aligned}$$

and by the same devices $|I_-| = 0$, too. This completes the proof of Lemma 3.6; collecting (3.7), (3.8), (3.9), and (3.15) according to (3.6) establishes the result. ∎

**Remark:** It is not difficult to extend the derivations above in order to develop a more accurate expansion for $f_m(\alpha)$. In fact, we have

$$f_m(\alpha) = \frac{m!}{\alpha^{m+1}} + \sum_{k=0}^{K} \frac{B_{k+m+1}}{k+m+1} \cdot \frac{(-\alpha)^k}{k!} + O\left(\frac{m!}{(2\pi)^m}(m\alpha)^{K+1/2}\right),$$

where the remainder depends on $K$ but is still uniformly for $m \to \infty$ and $\alpha \to 0+$.

THEOREM 3.7 (UNIFORM EXPANSIONS FOR THE MOMENTS OF $\mathcal{S}_T$). *There exists some $\delta > 0$ such that the moments $E[\mathcal{S}_T^m]$ of $\mathcal{S}_T$ have the uniform asymptotic expansion*

$$\begin{aligned} E[\mathcal{S}_T^m] &= \sum_{n\geq 1} n^m s_{n,T} \\ &= m! \frac{1 - B_T(1)}{\phi_T B_T'(\phi_T)} (\log \phi_T)^{-m-1} + O\left(\mu_T^{-1} \frac{m!\sqrt{m}}{(2\pi e\delta)^m}\right) \\ &= m! \big[\mu_T(1 + O(1/\mu_T))\big]^m + O\left(\mu_T^{-1} \frac{m!\sqrt{m}}{(2\pi e\delta)^m}\right) \end{aligned}$$

*for $T \to \infty$ and $m \geq 1$.*

**Proof:** According to Theorem 3.4, we have for some $\delta > 0$

$$s_{n,T} = \frac{1 - B_T(1)}{\phi_T B_T'(\phi_T)} \phi_T^{-n} + O\big(\mu_T^{-1}(1+\delta)^{-n}\big) = \mu_T^{-1}\big(1 + O(1/\mu_T)\big)\phi_T^{-n} + O\big(\mu_T^{-1} e^{-\delta n}\big),$$

where $\phi_T = 1 + \mu_T^{-1} + O(\mu_T^{-2})$. Using the result of Lemma 3.6 (that is, the remark on this lemma) we obtain

$$\begin{aligned} \sum_{n\geq 1} n^m \phi_T^{-n} &= \frac{m!}{(\log \phi_T)^{m+1}} + O\left(\frac{m!\sqrt{m}}{(2\pi e)^m}\right) \\ &= \frac{m!}{\big(\mu_T^{-1} + O(\mu_T^{-2})\big)^{m+1}} + O\left(\frac{m!\sqrt{m}}{(2\pi e)^m}\right), \end{aligned}$$

where we used $\log(1+z) = z + O(z^2)$ for $z \to 0$. Similarily, we find

$$\sum_{n\geq 1} n^m e^{-\delta n} = O\left(\frac{m!\sqrt{m}}{(2\pi e\delta)^m}\right)$$

Putting this together yields the statement of Theorem 3.7. ∎

**Remark:** Using the more accurate asymptotic expansion of $\phi_T$ from the remark on Lemma 3.2 yields

$$\log \phi_T = \mu_T^{-1} - \frac{B_T'(1) + B_T''(1)}{2B_T'(1)} \mu_T^{-2} + O(\mu_T^{-3})$$

for $T \to \infty$. Hence we could improve the asymptotic expression for $E[\mathcal{S}_T^m]$, too.

## 4. Applications

Now, we shall return to the problem of qualifying scheduling techniques for indeterministic task arrivals in hard real-time systems presented in Section 1, i.e., to apply the theorems developed in Section 3 to the scheduling algorithms investigated in our earlier papers.

In order to do that, we need some additional details concering our discrete time queueing system model: With the notations from Section 1, the PGF of the number of task arrivals during a cycle is denoted by

$$A(z) = \sum_{k \geq 0} a_k z^k, \qquad \text{where } a_k = \text{prob}\{k \text{ tasks arrive during a cycle}\} \tag{4.1}$$

and should meet the constraint $a_0 = A(0) > 0$, i.e., the probability of no arrivals during a cycle should be greater than zero. This assures the existence of idle cycles. Note, that the definition implies the independence of arrivals within two arbitrary different cycles.

The PGF of the task execution times (measured in cycles) is denoted by

$$L(z) = \sum_{k \geq 0} l_k z^k, \qquad \text{where } l_k = \text{prob}\{\text{task execution time is } k \text{ cycles}\} \tag{4.2}$$

with the additional assumption $L(0) = 0$, i.e., all task execution times should be greater than or equal to one cycle. Again, this definition implies that task execution times are independent from each other and from the arrival process.

We should mention that the number of probability distributions meeting our constraints is considerably limited due to the required independency. An example for a suitable model is based on an interarrival distribution with the so-called memoryless property, i.e., an exponential or geometric distribution, leading to (well-thumbed) Poisson- or Bernoulli-type arrivals within a cycle, respectively.

It turns out that the overall execution time, i.e., the number of cycles necessary for processing all actions induced by task arrivals within a cycle, plays a central role. We obviously have $P(z) = A\big(L(z)\big)$. Note, that the function $B(z)$, which provides the connection to this paper, is (essentially) the solution of the functional equation $B = zP(B)$.

For technical reasons we need some additional conditions on $A(z)$, $L(z)$ and $P(z)$, respectively. We omit a detailed discussion for the sake of simplicity; most of them are analyticity conditions, which are usually easy to establish. Note however, that we explicitely exclude the trivial case $P(z) = p_0 + (1 - p_0)z$.

It turns out that it is necessary to distinguish three different situations, namely

(1) *Normal Case*

This (most important) case is characterized by an average offered load of less than 100%, which may be expressed by $P'(1) < 1$ (since $P'(1)$ equals the average number of actions caused by task arrivals within a cycle). That is, our system has to deal with task arrivals keeping it not totally busy on the average. This situation is reflected by the fact that $B(z)$ has a radius of convergence $R > 1$, which makes the theorems of Section 3 applicable.

(2) *Balanced Case*
Here, our system is kept 100% busy on the average, i.e., $P'(1) = 1$. Since $B(z)$ has radius of convergence $R = 1$ here our theorems are no longer applicable. However, some tedious computations showed that the expectation of $S_T$ fulfills $E[S_T] \sim T$ for $T \to \infty$ for all scheduling techniques investigated so far.

(3) *Overloaded Case*
This case may be characterized by an average offered load which is higher than the maximum load the system is able to cope with, that is, $P'(1) > 1$. Again, our theorems are not suitable here since the radius of convergence of $B(z)$ is less than 1. However, more or less simple computations showed a constant expectation $E[S_T] \sim C$ for $T \to \infty$.

In what follows we restrict ourselves to the normal case for FCFS and both nonpreemptive and preemptive LCFS scheduling. The very complete derivation of the basic results, which rely on some well-established (combinatorial and asymptotic) methods from the *analysis of algorithms and data structures*, is contained in a number of other papers, cf. **[SB1]**, **[SB2]**, **[BS1]**. They provide asymptotic results concerning the crucial quantity $\mu_T$, cf. equation (2.9). Using those results and our Theorem 3.4 of Section 3, we obtain

THEOREM 4.1. *(FCFS scheduling in the normal case, cf. [**SB1**, Theorem 1]). The successful run duration $S_T$ for FCFS scheduling in the normal case is approximately exponentially distributed with parameter $1/\mu_T^{FCFS}$ where*

$$\mu_T^{FCFS} = \frac{P'(\kappa) - 1}{(\kappa - 1)(1 - P'(1))^2} \kappa^T (1 + O(1/T)) \qquad \text{for } T \to \infty,$$

*$\kappa > 1$ is the solution of $x = P(x)$, $x > 1$.* ∎

THEOREM 4.2. *(nonpreemptive LCFS scheduling in the normal case, cf. [**SB2**, Theorem 5.1]). The successful run duration $S_T$ for nonpreemptive LCFS scheduling in the normal case is approximately exponentially distributed with parameter $1/\mu_T^{npLCFS}$ where*

$$\mu_T^{npLCFS} = C\, T^{3/2} \rho^T (1 + O(1/T)) \qquad \text{for } T \to \infty,$$

*and*

$$C = \frac{2\sqrt{\pi}(\rho - 1)(\tau - a_0)L(\tau)}{bL(\rho)(1 - P'(1))} \left( \frac{(1 - a_0)\,(L(\tau) - L(a_0))\, a_0(\rho - 1)}{L(a_0)(\tau - a_0)} - \frac{(\tau - 1)(\tau - a_0\rho)L'(\tau)}{L(\tau)} + \tau - a_0 \right)^{-1},$$

*$\tau > 1$ is the solution of $P(x) = xP'(x)$, $\rho = \tau/P(\tau) > 1$ and $b = \sqrt{2P(\tau)/P''(\tau)}$.* ∎

THEOREM 4.3. *(preemptive LCFS scheduling in the normal case, cf. [**BS1**, Theorem 2]). The successful run duration $S_T$ for preemptive LCFS scheduling in the normal case is approximately exponentially distributed with parameter $1/\mu_T^{pLCFS}$ where*

$$\mu_T^{pLCFS} = \left( \frac{2\pi P''(\tau)}{P(\tau)} \right)^{1/2} \frac{\tau^2(\rho - 1)}{\rho^2(1 - P'(1))} T^{3/2} \rho^T (1 + O(1/T)) \qquad \text{for } T \to \infty,$$

$\tau > 1$ *is the solution of* $P(x) = xP'(x)$ *and* $\rho = \tau/P(\tau) > 1$. ∎

Due to the exponential growth results for $\mu_T$ for any scheduling discipline mentioned one might expect that our system will operate properly a very long time, even for a high average offered load and a short deadline $T$. Numerical results concerning a particular example (assuming a constant task execution time of 1 cycle) showed indeed very impressive results: For example, in the case of FCFS scheduling with Poisson arrivals with rate $\lambda = 0.5$ tasks/cycle, a service time deadline of $T = 10$ cycles causes $\mathrm{E}[S_T] = \mu_T \approx 10^6$ cycles; $T = 20$ cycles yields $\mu_T \approx 10^{12}$ cycles. Note however, that FCFS scheduling shows always the best performance since $\kappa > \rho$. This is not too unexpected, since FCFS scheduling is equivalent to the so-called *earliest deadline first* scheduling† here (due to our fixed deadline assumption).

Moreover, it is obvious that Theorem 3.4 provides an answer to the following practical question: Given the probability distributions (4.1) and (4.2) for certain stress situations, and a (tolerable) probability $p$ for deadline missing (say, $p = 10^{-9}$), what is the maximum duration of such a stress period to guarantee a deadline missing probability of at most $p$?

Finally, we should repeat that the results of this paper reduce the investigation of different scheduling techniques (in the normal case) to the task of establishing an asymptotic expression for $B_T(1)$ and $B'_T(1)$. However, those derivations are sometimes complicated enough, as may be seen in our cited papers.

## References


AS. M. Abramowitz, I. A. Stegun, "Handbook of Mathematical Functions," Dover Publications, Inc., New York, 1972.

BI. P. Billingsley, "Probability and Measure," 2nd ed., John Wiley & Sons, Inc., New York, 1986.

BS1. J. Blieberger, U. Schmid, *Some Investigations on Preemptive LCFS Scheduling in Hard Real Time Applications.* (to appear in Performance Evaluation).

CSR. S. Cheng, J. Stankovic, K. Ramamritham, *Scheduling Algorithms for Hard Real-Time Systems—A Brief Survey*, in "Tutorial: Hard Real-Time Systems," Computer Society Press IEEE, Washington, 1988.

D. M. Drmota, *The Instability Time Distribution Behaviour of Slotted ALOHA.* (submitted).

DS. M. Drmota, U. Schmid, *Ultimate Characterization of the Successful Operation of Slotted ALOHA.* (submitted).

FE. W. Feller, "An Introduction to Probability Theory and Its Applications, Vol. 1," John Wiley & Sons, Inc., New York, 1968.

HE. P. Henrici, "Applied and Computational Complelx Analysis, Vol. 1," John Wiley & Sons, New York, 1974.

SB1. U. Schmid, J. Blieberger, *Some Investigations on FCFS Scheduling in Hard Real Time Applications.* (to appear in Journal of Computers and System Sciences).

SB2. U. Schmid, J. Blieberger, *On Nonpreemptive LCFS Scheduling with Deadlines.* (submitted).

TR. K. S. Trivedi, "Probability and statistics with reliability, queueing, and computer science applications," Prentice Hall, Englewood Cliffs, N.J., 1982.

WW. E. T. Whittaker, G. N. Watson, "A Course of Modern Analysis," Cambridge University Press, Cambridge, 1927.

VF. J. S. Vitter, Ph. Flajolet, *Average Case Analysis of Algorithms and Data Structures.* Handbook of Theoretical Computer Science (J. van Leeuwen, ed.) (1990), North Holland.


†Which is known to be optimal.