A Processor-Sharing model for the Performance of Virtualized Network Functions

Fabrice Guillemin, Veronica Quintuna Rodriguez, Alain Simonian
Orange Labs, France
email: firstname.lastname@orange.com

***Abstract*—The parallel execution of requests in a Cloud Computing platform, as for Virtualized Network Functions, is modeled by an $M^{[X]}/M/1$ Processor-Sharing (PS) system, where each request is seen as a batch of unit jobs. The performance of such paralleled system can then be measured by the quantiles of the batch sojourn time distribution. In this paper, we address the evaluation of this distribution for the $M^{[X]}/M/1$-PS queue with batch arrivals and geometrically distributed batch size. General results on the residual busy period (after a tagged batch arrival time) and the number of unit jobs served during this residual busy period are first derived. This enables us to provide an approximation for the distribution tail of the batch sojourn time whose accuracy is confirmed by simulation.**

***Index Terms*—Processor-Sharing discipline; Queues with Batch Arrivals; Busy Period; Distribution Tail; Cloud Computing; Virtualized Networks; Performance Analysis.**

I. Introduction

Network Function Virtualization (NFV) [1] is deeply modifying the architecture and the operation of telecommunication networks. As a matter of fact, network functions, which were so far hosted on dedicated hardware, are now implemented owing to virtualization technologies on common hardware. A Virtualized Network Function (for short, VNF) is actually most of the time composed of sub-functions which can be executed in parallel or in series. A VNF thus appears as a set of tasks to be executed on a computing system, such as for instance a multi-core platform.

Some sub-functions of a global VNF can be executed in parallel. This is notably the case of the Radio Access Network (RAN) functions such as the channel coding see [2]. In this case, it is fundamental to investigate which resource allocation strategy is the most adapted to execute virtualized (sub-)functions on a multi-core platform and, moreover, how cores must be allocated to the tasks that must be executed in parallel.

From a modelling point of view, a global VNF or a set of tasks, which can be processed in parallel, appears as a batch of unit jobs to be executed on a multi-server system. In this view, batch arrivals correspond to instants when VNF-jobs have to be processed. This leads us to consider multi-server queuing systems with batch arrivals. Several core allocation procedures have been considered by simulation in [3], notably the $M^{[X]}/M/C$ and $M^{[X]}/M/1$ Processor Sharing (PS). The analytical study of the $M^{[X]}/M/C$ queue has been performed in [2], thus extending earlier results obtained in [4]. In particular, the asymptotic behavior of the waiting time distribution of an entire batch has been derived.

In this paper, we consider the PS discipline which is a popular method of sharing a common resource between competing tasks. In the context of a multi-core platform, PS consists of sharing the whole computing capacity among all tasks present in the system. The allocation of cores is achieved by the scheduler of the operating system managing the multi-core system. When the PS discipline is performed, all batches are treated in parallel and receive equal portion of the computing capacity in a fair basis. It is worth noting that more sophisticated methods could also be envisaged. For instance, the Early Deadline First (EDF) discipline which is often implemented in Linux OS for dealing with real-time applications. However, it is still much more difficult to analyze [5].

In this context, the PS queue with batch arrivals has been already envisaged [6] to calculate the distribution of the sojourn time W of a single job, extending the results obtained by Kleinrock *et al.* for the mean value [7]. In this paper, we consider the evaluation of the sojourn time Ω of an entire batch. Although the distribution of W could be given as an explicit integral representation, the exact calculation of the distribution of Ω proves much more challenging. To overcome this difficulty, we propose an approximation for the distribution tail of Ω which compares reasonably well to simulation and can be easily handled to quantify the system performance. This approximation is derived from (i) general results for the *residual* busy period starting after the arrival time of a tagged batch and (ii) an equi-probability assumption for the departing order of jobs within the residual busy period.

The analysis performed in this paper allows us to explicitly compute the exponential decay rate of the sojourn time of an entire batch. This decay rate globally gives a means of estimating the performance of a resource sharing discipline. In particular, when real time constraints have to be met while executing VNFs (notably in Cloud RAN systems), the decay rate is an indication of the reneging rate of VNFs. In fact, if some VNFs are not executed within prescribed delay bounds, they fail (or renege from a modeling point of view). The results obtained in this paper allow us to compare the PS discipline against the FIFO discipline considered in [2] for scheduling channel coding tasks in a virtual RAN context.

The paper is organized as follows. Sections II and III address the exact characterization of the full (resp. residual)

busy period of the PS queue with batch arrivals and the number of jobs served during this period. On the basis of the latter results, Section IV then addresses an approximation for the distribution of the batch sojourn time Ω. This approximation is then compared to simulation experiments in Section V. Some concluding remarks are finally presented in Section VI.

II. Characteristics of the busy period

Consider a general $M^{[X]}/G/1$ single server queue with a work-conserving service discipline. This queue is fed by a Poisson process of batches with mean arrival rate ϱ; the size (in number of jobs) of any batch is denoted by B. The mean service time of a unit job is set equal to 1 and we assume that

$$\varrho \stackrel{def}{=} \varrho\,\mathbb{E}(B) < 1, \tag{1}$$

to ensure the existence of a stationary regime for this queue. Let T be the duration of a busy period; during such a busy period, the number of jobs served is denoted by M. After [8, Chap.2, Sect.3], let $\widetilde{F}_m(x) = \mathsf{P}(T \quad x, M = m)$ for $m > 1$ and $x > 0$, and define the double transform ν by

$$\nu(r,s) = \sum_{m \ 1} r^m \int_0^{+1} e^{\ s\,x}\,\mathrm{d}\widetilde{F}_m(x), \qquad jrj < 1,\ s > 0.$$

It is known that, given $jrj < 1$ and $s > 0$, $\nu(r,s)$ is equal to the smallest root (in modulus) to the equation [ibid., Eq. (2.15)]

$$\nu = B\ (r \quad D\ (s + \varrho \quad \varrho\nu)) \tag{2}$$

where D (resp. B) denotes the Laplace transform of the distribution of the job service time (resp. the generating function of the number of jobs contained in a batch).

In the rest of this paper, it is assumed that

- the identically and independently distributed (i.i.d.) job service times are exponentially distributed with parameter 1 so that $D\ (s) = 1/(1+s)$ for $s > 0$;
- the size (in number of jobs) of a given batch is geometrically distributed with parameter q 2 $]0,1[$, so that

$$B\ (z) = \frac{(1 \quad q)\,z}{1 \quad q\,z}, \qquad jzj < 1. \tag{3}$$

After Eq. (3), in particular, the general stationarity condition (1) now specifies into

$$\varrho < 1 \quad q. \tag{4}$$

Under these assumptions, we can assert the following.

Lemma 1: **For the $M^{[X]}/M/1$ queue with geometrically distributed batch size, the Laplace transform T of the busy period duration T is given by**

$$T\ (s) = \frac{s+1 \quad q+\varrho \quad \sqrt{\Delta_q(s)}}{2\varrho} \tag{5}$$

for s 2 $\mathbb{C}\,n\,[\sigma_q\ ,\sigma_q^+]$**, where**

$$\Delta_q(s) = (s+1+\varrho \quad q)^2 \quad 4\varrho(1 \quad q) = (s \quad \sigma_q^+)(s \quad \sigma_q\ ) \tag{6}$$

and

$$\sigma_q \ = \ (\sqrt{1 \quad q} \quad \mathsf{p}\overline{\varrho})^2. \tag{7}$$

The distribution tail of the busy period T decays exponentially fast with rate $j\sigma_q^+j = (\mathsf{p}\overline{1 \quad q} \quad \mathsf{p}\overline{\varrho})^2$, specifically

$$\mathsf{P}(T > x) \quad \frac{(1 \quad q)^{\frac{1}{4}}}{2\mathsf{p}\overline{\pi}\,\varrho^{\frac{3}{4}}\,j\sigma_q^+j}\ \frac{e^{\sigma_q^+ x}}{x^{\frac{3}{2}}} \tag{8}$$

for large positive x.

Proof: Let $T\ (s) = \nu(1,s) = \mathsf{E}(e^{\ s\,T})$, $s > 0$. Applying Eq. (2) for $r = 1$ with $D\ (s) = 1/(1+s)$ and the definition (3) of B entails that the Laplace transform T verifies

$$T\ (s) = \frac{1 \quad q}{1+s+\varrho \quad q \quad \varrho\,T\ (s)},$$

that is, $\varrho\,T\ (s)^2 \quad (1+s+\varrho \quad q)\,T\ (s) + 1 \quad q = 0$, $s > 0$, which solves for $T\ (s)$ (equal to the smallest root) into

$$T\ (s) = \frac{1+s+\varrho \quad q \quad \sqrt{\Delta_q(s)}}{2\,\varrho}, \qquad s > 0,$$

with $\Delta_q(s)$ defined by (6). It is readily verified that T then defines an analytic function in the cut plane $\mathbb{C}\,n\,[\sigma\ ,\sigma_+]$.

Besides, it is known [9, Sect. 3.46] that the Laplace inverse of transform $s > 0$ 7! $s \quad \mathsf{p}\overline{s^2 \quad a^2}$ is $t > 0$ 7! $a\,I_1(at)/t$ for any constant $a > 0$, where I_1 is the modified Bessel function with order 1; applying the latter inverse with $a = 2\sqrt{\varrho(1 \quad q)}$, the Laplace inversion of (5) entails that the busy period T has the probability density

$$\mathsf{P}(T = t) = \sqrt{\frac{1 \quad q}{\varrho}}\ \frac{e^{\ (1+\varrho \quad q)t}}{t}\ I_1(2\sqrt{\varrho(1 \quad q)}\ \ t) \tag{9}$$

for $t > 0$. Using the fact that $I_1(X) \quad e^X/\mathsf{p}\overline{2\pi X}$ for large positive X [10, Chap.5, Eq. (5.11.8)], the tail of this density is therefore asymptotic to

$$\mathsf{P}(T = t) \qquad \sqrt{\frac{1 \quad q}{\varrho}}\ \frac{e^{\ (1+\varrho \quad q)t}}{t}\ \frac{e^{2\mathsf{p}\overline{\varrho(1 \quad q)t}}}{\sqrt{2\pi \quad 2\sqrt{\varrho(1 \quad q)}t}}$$
$$= \quad \frac{(1 \quad q)^{\frac{1}{4}}}{\varrho^{\frac{3}{4}}}\ \frac{e^{\sigma_q^+ t}}{2\mathsf{p}\overline{\pi}\,t^{\frac{3}{2}}}$$

for large positive t. The estimate (8) of $\mathsf{P}(T > x)$ for large positive x follows. ∎

Note that the random variable T is the length of the busy period seen by an external observer. It is not easy to relate the distribution of T to the sojourn time of a batch. In fact, upon arrival, an arbitrary batch sees the system in equilibrium owing to the PASTA property, and not the empty state; we can thus claim that the sojourn time of an arbitrary batch is less than the residual busy period following the batch arrival instant. This sojourn time is further studied in the next section.

Lemma 2: **The generating function M of the number M of customers served in a busy period of the $M^{[X]}/M/1$ queue with geometric batch arrivals is given by**

$$M\ (z) = \frac{1+\varrho \quad qz \quad \sqrt{\delta_q(z)}}{2\varrho} \tag{10}$$

for z 2 $\mathbb{C}\,n\,[\zeta_q\ ,\zeta_q^+]$**, where**

$$\delta_q(z) = (1+\varrho \quad qz)^2 \quad 4\varrho z(1 \quad q) = q^2(z \quad \zeta_q\ )(z \quad \zeta_q^+) \tag{11}$$

and

$$\zeta_q = \left(\frac{\sqrt{\varrho + q} - \sqrt{\varrho(1-q)}}{q} \right)^2 . \quad (12)$$

The distribution tail of M decays exponentially fast with rate ζ_q , specifically

$$\mathsf{P}(M = m) \sim \frac{q\sqrt{(\zeta_q^+ - \zeta_q)\zeta_q}}{4\varrho\sqrt{\pi}} \frac{1}{m^{\frac{3}{2}}} \left(\frac{1}{\zeta_q}\right)^m \quad (13)$$

for large integer m.

Proof: Setting $s = 0$ in Eq. (2), we deduce that $M(z)$ satisfies

$$M(z) = \frac{(1-q)z}{1 + \varrho - \varrho M(z) - q},$$

which quadratic equation has the smallest solution given by expression (10). This equation defines a analytic function in the cut plane $\mathbb{C} \setminus [\zeta_q, \zeta_q^+]$, where ζ_q are defined by (12). A direct application of Darboux's method [11, Theorem VI.14] further yields asymptotics (13), as claimed. ■

III. The residual busy period

As argued in the following, the distribution of the batch sojourn time Ω for the Processor-Sharing $M^{[X]}/M/1$ queue can be upper bounded to that of the residual busy period after the arrival of a tagged batch. To analyze this residual busy period, we here follow the treatment of [8, p.249, Section II.4.4] for the analysis of the (full) busy period for the $M^{[X]}/G/1$ queue with any work-conserving service discipline. This obtained results will apply, in particular, to the Processor-Sharing discipline subsequently considered.

A. Joint Laplace transform of $\widetilde{T}$ and $\widetilde{M}$

Let an $M^{[X]}/M/1$ queue with a work-conserving discipline. Consider a tagged batch with size $B = b$ (in terms of number of jobs) arriving during a busy period; this batch sees a number $N_0 = n > 0$ of jobs already present in the queue. Let $\widetilde{T}$ (resp. $\widetilde{M}$) further denote the residual duration of the busy period after the arrival time of the test batch (resp. the number of jobs served during this residual duration $\widetilde{T}$). For an illustration, Figure 1 displays a sample busy period with duration T and the residual busy period $\widetilde{T}$ associated with the arrival of this tagged batch (recall that for any work-conserving service discipline, the busy period is determined by the smallest interval where the unfinished workload does not reach 0).

Let us introduce the random variable $\widetilde{M}$ equal to the number of jobs served in the residual busy period of length $\widetilde{T}$. We can then state the following result.

Proposition 1: **Given $N_0 = n > 0$ and $B = b > 1$, the conditional distribution of the pair $(\widetilde{T}, \widetilde{M})$ is given by**

$$\mathsf{E}_{n,b}(r^{\widetilde{M}} e^{-s\widetilde{T}}) = \left(\frac{r}{1 + s + \varrho - \varrho\nu(r,s)} \right)^{n+b} \quad (14)$$

for $|r| < 1$ and $s > 0$, where ν is the solution to the functional equation (2).

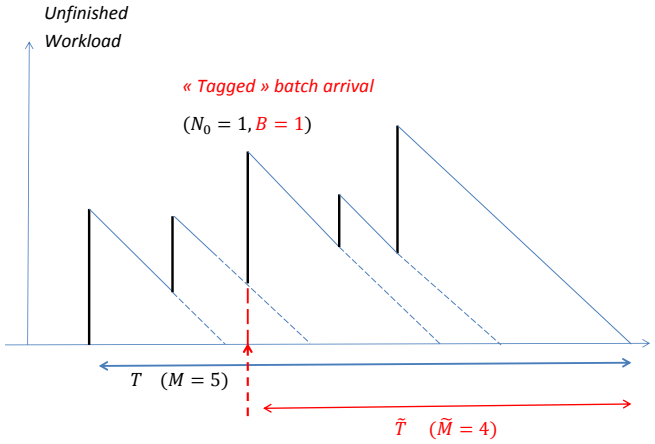

Fig. 1. *Busy period T and residual Busy period $\widetilde{T}$.*

Proof: Fix $N_0 = n > 0$, $B = b > 1$ and define τ as the cumulative service time of all the $n + b$ jobs present in queue just after the arrival of the tagged batch; A_τ as the number of batch arrivals in the time interval with duration τ.

By definition of $\widetilde{T}$ and $\widetilde{M}$,

(a) if $A_\tau = 0$, then $\widetilde{T} = \tau$ and $\widetilde{M} = n + b$;

(b) if $A_\tau > 1$, with regard to the duration of the residual busy period, it is indifferent to postpone the service of the remaining $A_\tau - 1$ batches to the end of the busy period generated by the first batch $\mathfrak{B}_1$ arrived among this number A_τ. For $1 \le k \le A_\tau$, setting then

- T_k equal to the duration of the busy period generated by the k-th batch $\mathfrak{B}_k$ arrived during the interval τ (with the convention $T_0 = 0$),
- M_k the number of jobs served during the busy period generated by this k-th batch $\mathfrak{B}_k$ (with the convention $M_0 = 0$),

we have the defining equalities $\widetilde{T} = \tau + T_1 + ... + T_{A_\tau}$ and $\widetilde{M} = n + b + M_1 + ... + M_{A_\tau}$.

By using the above observations for both variables $\widetilde{T}$ and $\widetilde{M}$, the double transform defining the joint distribution of the pair $(\widetilde{T}, \widetilde{M})$ then satisfies

$$\mathsf{E}_{n,b}(r^{\widetilde{M}} e^{-s\widetilde{T}}) = \mathsf{E}_{n,b}(r^{n+b+M_1+...+M_{A_\tau}} e^{-s(\tau+T_1+...+T_{A_\tau})})$$

so that

$$\mathsf{E}_{n,b}(r^{\widetilde{M}} e^{-s\widetilde{T}}) = r^{n+b} \sum_{k=0} \mathsf{E}_{n,b}(e^{-s\tau} \mathbf{1}_{A_\tau = k} r^{M_1+...+M_{A_\tau}} e^{-s(T_1+...+T_{A_\tau})}) =$$

$$r^{n+b} \sum_{k>0} \int_0^{+1} \mathrm{dP}_\tau(t) e^{-st} e^{-\varrho t} \frac{(\varrho t)^k}{k!} \left[\mathsf{E}(r^M e^{-sT})\right]^k \quad (15)$$

after conditioning with respect to the variable τ, by noting that the distribution of A_t is Poisson with parameter ϱt and

by using the essential fact that all pairs (T_k, M_k), $k > 1$, are independent and identically distributed. Performing the summation with respect to index k in (15), we are therefore left with

$$\mathsf{E}_{n,b}(r^{\widetilde{M}} e^{-s\widetilde{T}}) = r^{n+b} \int_0^{+1} \mathrm{d}\mathsf{P}_\tau(t) e^{-st} e^{-\varrho t} \exp\left[\varrho t\, \mathsf{E}(r^M e^{-sT})\right] = r^{n+b}\, \tau(s+\varrho-\varrho\,\mathsf{E}(r^M e^{-sT})) \quad (16)$$

where τ denotes the Laplace transform of variable τ. By the memory-less property of the exponential distribution applied to the remaining service duration of the n jobs present at the arrival instant of the tagged batch, τ is the sum of $(n+b)$ i.i.d. variables with exponential distribution with parameter 1; hence $\tau(s) = 1/(1+s)^{n+b}$, $s > 0$. By equality (16) and the latter expression of $\tau(s)$, formula (14) follows. ∎

B. Marginal distributions of $\widetilde{T}$ and $\widetilde{M}$

By using the joint Laplace transform determined in Proposition 1, we now derive the Laplace transform of the duration of the residual busy period and the generating function of the number of jobs served during such a residual busy period.

First note that the number N_0 of jobs present in the queue at the tagged batch arrival instant and the size B of this batch are independent variables. From [6], Eq. (3.2), we know that the generating function of the number N_0 is given by

$$\eta(z) = \mathsf{E}(z^{N_0}) = (1-\varrho)\frac{1-qz}{1-(\varrho+q)z}, \qquad |z| < 1, \quad (17)$$

where $\varrho = \varrho/(1-q)$.

From the definition (3) of $\mathsf{E}(z^B) = B(z)$, the generating function φ for the sum $N_0 + B$ is consequently given by $\varphi(z) = \mathsf{E}(z^{N_0+B}) = \eta(z)\, B(z)$ which reduces by (17) to

$$\varphi(z) = \frac{(1-\varrho-q)z}{1-(\varrho+q)z}, \qquad |z| < 1. \quad (18)$$

Proposition 2: **The Laplace transform $\widetilde{T}$ of the residual busy period $\widetilde{T}$ is given by**

$$\widetilde{T}(s) = \frac{(1-q-\varrho)\left[(s+1-\varrho-q) + \sqrt{\Delta_q(s)}\right]}{2\varrho s} \quad (19)$$

for $s \in \mathbb{C} \setminus [\sigma_-, \sigma_+]$, with $\Delta_q(s)$ defined by (6).

The distribution tail of $\widetilde{T}$ is asymptotic to

$$\mathsf{P}(\widetilde{T} > x) \sim \frac{(1-q-\varrho)\sqrt{\sigma_q^+ - \sigma_q}}{4\sqrt{\pi}\varrho(\sigma_q^+)^2} \frac{e^{\sigma_q^+ x}}{x^{\frac{3}{2}}} \quad (20)$$

for large x, with σ_q defined by (7).

Proof: By setting $r = 1$ in Equation (14) and deconditioning on $N_0 + B$, we deduce that $\widetilde{T}(s)$ is given by

$$\widetilde{T}(s) = \varphi\left(\frac{1}{1+s+\varrho-\varrho T(s)}\right)$$

with transform T given in (5) and function φ defined in (18); simple algebra then provides expression (19), which defines an analytic function in the cut plane $\mathbb{C} \setminus [\sigma_-, \sigma_q^+]$. Besides, (19) entails that $\widetilde{T}$ has an algebraic singularity at point σ_q^+ with the expansion

$$\widetilde{T}(s) = T_q + S_q(s-\sigma_q^+)^{1/2} + o\left((s-\sigma_q^+)^{1/2}\right) \quad (21)$$

when $s \to \sigma_q^+$, where constants $T_q = \widetilde{T}(\sigma_q^+)$ and S_q are easily calculated as

$$T_q = 1 + \sqrt{\frac{1-q}{\varrho}}, \quad S_q = \frac{(1-q-\varrho)\sqrt{\sigma_q^+ - \sigma_q}}{2\varrho\sigma_q^+}. \quad (22)$$

A direct application of a classical Tauberian theorem [12, Theorem 25.2] then yields asymptotics (20). ∎

Proposition 2 has an immediate consequence for the distribution of the batch sojourn time in the $M^{[X]}/M/1$-PS queue. In the sequel, we denote by Ω the sojourn of an entire batch in this PS queue, that is, the time elapsed between the batch arrival time in queue and the time when all its component jobs have completed their service.

Corollary 1: **In the *Processor-Sharing* $M^{[X]}/M/1$ queue, the distribution tail of the batch sojourn time Ω decreases exponentially fast with rate $|\sigma_q^+|$ introduced in Eq. (7).**

Proof: As derived in [6, Cor. 5.2.1] for the $M^{[X]}/M/1$-PS queue, the exponential decay rate $|\sigma_q^+|$ of the distribution of $\widetilde{T}$ (and T) equals that of the distribution of the sojourn time W of a *single job*.

The inequalities

$$W \le \Omega \le \widetilde{T}, \qquad \text{a.s.}, \quad (23)$$

then entail that $\mathsf{P}(W > x) \le \mathsf{P}(\Omega > x) \le \mathsf{P}(\widetilde{T} > x)$ for all $x > 0$, which enables us to conclude that the distribution tail of the batch sojourn time Ω also decreases exponentially fast with rate $|\sigma_q^+|$. ∎

It is worth noting that $\widetilde{T}$ is asymptotically greater than T. Indeed, for large x, we have

$$\frac{\mathbb{P}(\widetilde{T} > x)}{\mathbb{P}(T > x)} \sim \frac{1-\varrho-q}{|\sigma_q^+|} > 1.$$

Proposition 3: **The generating function $\widetilde{M}$ of the number $\widetilde{M}$ of jobs served during the residual busy period is given by**

$$\widetilde{M}(z) = \frac{(1-q-\varrho)\left[1+\varrho-(q+2\varrho)z - \sqrt{\delta_q(z)}\right]}{2\varrho(\varrho+q)(z-1)} \quad (24)$$

withe $\delta_q(z)$ defined by (11).

For large m, we further have

$$\mathbb{P}(\widetilde{M} = m) \sim \frac{(1-q-\varrho)q\sqrt{(\zeta_q^+ - \zeta_q)\zeta_q}}{4\sqrt{\pi}\varrho(\varrho+q)(\zeta_q - 1)} \frac{1}{m^{\frac{3}{2}}} \left(\frac{1}{\zeta_q}\right)^m \quad (25)$$

with ζ_q defined by (12).

Proof: By setting $s = 0$ in Equation (14) and deconditioning on $N_0 + B$, we deduce that $\widetilde{M}(z)$ is given by

$$\widetilde{M}(z) = \varphi\left(\frac{z}{1+\varrho-\varrho M(z)}\right)$$

with generating function M given in (10) and function φ defined in (18); simple algebra then yields expression (24) for $\widetilde{M}(z)$, which defines an analytic function in the cut plane $\mathbb{C} \setminus [\zeta_q, \zeta_q^+]$. When z tends to ζ_q, we then derive

$$\widetilde{M}(z) = \frac{(1-q-\varrho)(1+\varrho)(q+2\varrho)\zeta_q}{2\varrho(\varrho+q)(\zeta_q-1)} \frac{(1-q-\varrho)q\sqrt{(\zeta_q-z)(\zeta_q^+-\zeta_q)}}{2\varrho(\varrho+q)(\zeta_q-1)} + o\left(\sqrt{\zeta_q-z}\right)$$

and a direct application of Darboux's method provides estimate (25). ∎

Define the sequence $(a_k)_{k>0}$ by

$$a_k = \frac{1}{(2k-1)2^{2k}}\binom{2k}{k}$$

so that $\sqrt{1-x} = \sum_{k>0} a_k x^k$ for $|x| < 1$ [13, Eq.(3.6.11)].

Corollary 2: **The distribution of variable $\widetilde{M}$ is given by**

$$\mathsf{P}(\widetilde{M}=m) = \frac{(1-q-\varrho)(1+\varrho)}{2\varrho(\varrho+q)} \sum_{\ell=m+1}^{+\infty} b_\ell \qquad (26)$$

for $m > 1$, where we define

$$b_k = \frac{1}{(\zeta_q)^k} \sum_{\ell=0}^{k} a_\ell a_{k-\ell} \left(\frac{q\zeta_q}{1+\varrho}\right)^{2\ell}, \quad k > 1. \qquad (27)$$

Proof: Using the fact that $\zeta_q^+\zeta_q = (1+\varrho)^2/q^2$, we have

$$\sqrt{\delta_q(z)} = (1+\varrho)\sqrt{1-\frac{z}{\zeta_q^+}}\sqrt{1-\frac{z}{\zeta_q}}$$

so that $\sqrt{\delta_q(z)} = (1+\varrho)\sum_{k>0} b_k z^k$ where b_k is defined by (27). If then follows from (24) that

$$\widetilde{M}(z) = \frac{(1-q-\varrho)}{2\varrho(\varrho+q)} \frac{1+\varrho-(q+2\varrho)z-(1+\varrho)\sum_{k>0} b_k z^k}{z-1};$$

by analyticity of $\widetilde{M}$, the numerator of the latter fraction vanishes for $z = 1$ so that $1-\varrho-q = (1+\varrho)\sum_{k>0} b_k$ and thus

$$\widetilde{M}(z) = \frac{(1-q-\varrho)}{2\varrho(\varrho+q)}\left[q-2\varrho-(1+\varrho)\sum_{k=1}^{+\infty} b_k \frac{z^k-1}{z-1}\right].$$

By definition of the residual busy period, we have $\widetilde{M} > 1$ a.s., hence $\widetilde{M}(0) = 0$ and the latter power series expansion consequently reduces to

$$\widetilde{M}(z) = \frac{(1-q-\varrho)(1+\varrho)}{2\varrho(\varrho+q)} \sum_{m=1}^{+\infty} z^m \sum_{\ell=m+1}^{+\infty} b_\ell,$$

whence (26). ∎

IV. Estimating the distribution of Ω

While Corollary 1 has provided us with the exponential decay rate for the distribution of the batch sojourn time Ω in the Processor-Sharing $M^{[X]}/M/1$ queue, the exact computation of the distribution of Ω remains, however, extremely challenging. In the present section, we use results of Sections II and III to propose an approximation for the distribution of sojourn time Ω.

Let us first introduce a few preliminary definitions. Given the numbers $N_0 = n > 0$ and $B = b > 1$, we denote by $I_1 < I_2 < \ldots < I_b$ the respective departure rank from the queue for each of the b jobs building up the tagged batch; by the above definition of the residual number of jobs served $\widetilde{M}$ after the tagged batch arrival, we certainly have

$$\forall k \in \{1,\ldots,b\}, \quad 1 \leq I_k \leq \widetilde{M}. \qquad (28)$$

All departure ranks I_k, $1 \leq k \leq b$, being distinct integers by construction, the maximal departure rank I_b also satisfies $b \leq I_b \leq \widetilde{M}$ (see illustration in Fig.2).

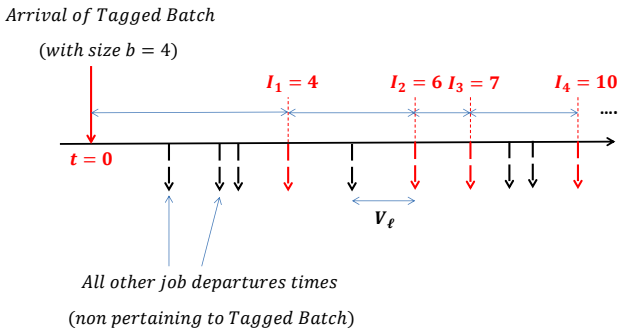

Fig. 2. *Consecutive departure times of jobs pertaining to the tagged batch (red arrows) and of jobs non pertaining to this batch (black arrows).*

Let V_ℓ denote the inter-departure duration between the consecutive completion time of job $\ell-1$ and job ℓ, $\ell > 1$, within the residual busy period (jobs $\ell-1$ and ℓ may belong to the tagged batch or not). The sojourn time Ω can then be written as

$$\Omega = V_1 + V_2 + \ldots + V_{I_b}. \qquad (29)$$

Given the residual number of jobs served $\widetilde{M}$, durations V_ℓ, $1 \leq \ell \leq I_b$, are dependent random variables: in fact, we have $V_1 + \ldots + V_{I_b} \leq \widetilde{T}$ so that, given $\widetilde{M}$ hence $\widetilde{T}$, the distribution of the V_ℓ's depends on the residual busy period and we cannot simply assert that it is exponential with parameter 1.

A. Equi-probability assumption

At this stage, we make the following assumption.

Hypothesis 1: **Given $N_0 = n$, $B = b$ and $\widetilde{M} = m$, all possible departure configurations $(I_1,\ldots,I_b)$ with constraints (28) for the set of departure ranks of the test tagged are equally probable.**

Setting then ξ_k as the indicator of the event "one of the jobs of the tagged batch has departure rank k", $k \in \{1,...,m\}$, Hypothesis 1 entails that

$$\mathsf{P}(\xi_1 = \varepsilon_1, ..., \xi_m = \varepsilon_m) = 1 \Big/ \binom{m}{b} \tag{30}$$

for any tuple $(\varepsilon_1, ..., \varepsilon_m) \in \{0,1\}^m$ such that $\varepsilon_1 + ... + \varepsilon_m = b$; $\binom{m}{b}$ is indeed the number of ways that b balls can be placed into $m > b$ boxes, each box containing at most one ball. Defining then the variable

$$J = \max \left\{ j \in \{1,...,m\}, \quad \sum_{k>j} \varepsilon_k = 0 \right\} \tag{31}$$

as the largest box index above which no other boxes contain any ball, **Hypothesis 1** thus consists in approximating the distribution of largest index I_b by that of J. In the following, we study the distribution of random variable J.

Lemma 3: **Given $N_0 = n > 0$, $B = b > 1$, $\widetilde{M} = m > 1$ and within the equi-probability Hypothesis 1, the conditional distribution of the maximum index J is given by**

$$\mathsf{P}_{n,b,m}(J = j) = \binom{j}{b} \Big/ \binom{m}{b} \tag{32}$$

for $b \le j \le m$.

Proof: By (30) and for $b \le j \le m$, we derive

$$\mathsf{P}_{n,b,m}(J \le j) = \mathsf{P}(\xi_1 + ... + \xi_j = b) = \binom{j}{b} \Big/ \binom{m}{b}$$

so that $\mathsf{P}_{n,b,m}(J = j) = \mathsf{P}_{n,b,m}(J \le j) - \mathsf{P}_{n,b,m}(J \le j-1)$ readily reduces to (32). ∎

The unconditional distribution of the random variable J turns out to be very difficult to compute. We can, nevertheless, estimate its asymptotic behavior at infinity as follows.

Corollary 3: **The unconditional distribution tail of J is given by**

$$\mathbb{P}(J = j) \sim K_q \frac{1}{j^{\frac{5}{2}}} \left(\frac{1}{\zeta_q} \right)^j \tag{33}$$

for large j, with constant

$$K_q = \frac{\kappa_q \zeta_q}{\zeta_q - 1} \frac{(1 - \varrho - q) r_q}{(1 - q r_q)^2} \frac{1 - q(\varrho + q) r_q^2}{(1 - (\varrho + q) r_q)^2} \tag{34}$$

where we set

$$r_q = \frac{2\zeta_q}{1 + \varrho + q\zeta_q}, \quad \kappa_q = \frac{q\sqrt{(\zeta_q^+ - \zeta_q)\zeta_q}}{2\sqrt{\pi}(1 + \varrho + q\zeta_q)}.$$

The proof of Corollary 3 is detailed in Appendix A.

B. Estimation of the sojourn time of a batch

We now formulate another assumption in order to approximate the distribution tail of sojourn time Ω of an entire batch. The customers pertaining to a given residual busy period leave the queue after service completion; as mentioned in the introduction of Section IV, we do not actually know the distribution of the inter-departure duration V_ℓ, $\ell > 1$; as the queue is work conserving, however, we may reasonably assume that they are independent and identically distributed (recall that this independence assumption can only be an approximation since theses inter-departures are considered conditionally to the fact that they are included in a given residual busy period). This motivates the following assumption.

Hypothesis 2: **The job inter-departure times in a residual busy period are i.i.d.**

Let then U denote an arbitrary job inter-departure time and $U(s)$ its Laplace transform. Given the event $\widetilde{M} = m$, **Hypothesis 2** then entails $U(s)^m = E(e^{-s\widetilde{T}} \mid \widetilde{M} = m)$ which, by deconditioning on $\widetilde{M}$ gives $\widetilde{M}(U(s)) = \widetilde{T}(s)$; using the expression (24) for $\widetilde{M}(z)$, the latter equation readily solves for $U(s)$ into

$$U(s) = \frac{\left[1 - q - \varrho(q+\varrho)(1 - \widetilde{T}(s))\right] \widetilde{T}(s)}{R(\widetilde{T}(s))}, \ s > 0, \tag{35}$$

where $R(t) = (\varrho t + 1 - \varrho - q)((q+\varrho)t + 1 - \varrho - q)$.

With the above evaluation of the inter-departure time U, we now approximate the sojourn time Ω of a tagged batch of size b as the departure time of the last customer among b customers picked up at random among those customers of the residual busy period. Let $\widetilde{\Omega}$ denote this approximate departure time.

Hypothesis 3: **Given $J = j > b$ and following (29), the distribution of the sojourn time Ω is approximated by the sum $\widetilde{\Omega} = U_1 + U_2 + ... + U_j$ where the U_ℓ's are i.i.d. random variables with the distribution of U defined by Eq.** (35)**.**

Invoking **Hypothesis 2** and **Hypothesis 3** now enable us to obtain the following evaluation for the distribution tail of Ω.

Proposition 4: **The distribution tail of the sojourn time Ω of of an entire batch can be approximated by**

$$\mathsf{P}(\widetilde{\Omega} > x) \sim \frac{H_q L_q}{2\sqrt{\sigma_q^+ \pi}} \frac{e^{\sigma_q^+ x}}{x^{\frac{3}{2}}} \tag{36}$$

for large x, with multiplying factor

$$H_q = \frac{\mathrm{d}J}{\mathrm{d}z}(U(\sigma_q^+))$$

where J denotes the generating function of variable J and with argument

$$U(\sigma_q^+) = \frac{1 + \varrho - \sqrt{\varrho(1-q)}}{q + \sqrt{\varrho(1-q)}}, \tag{37}$$

along with

$$L_q = \frac{\sigma_q^+ \sqrt{\sigma_q^+ - \sigma_q}}{2(q + \sqrt{\varrho(1-q)})^2}. \tag{38}$$

Proof: Following **Hypothesis 3**, the Laplace transform of $\widetilde{\Omega}$ is given by

$$\mathbb{E}(e^{-s\widetilde{\Omega}}) = J(U(s)), \qquad s > 0. \tag{39}$$

We claim that the smallest singularity of transform (39) in the complex plane is algebraic and located at $s = \sigma_q^+$. In fact, we make the following points:

After (22), the value $\widetilde{T}\ (\sigma_q^+) = T_q$ is finite and positive. Besides, the function $s \mapsto \widetilde{T}\ (s)$ decreases on the real interval $[\sigma_q^+, +\infty[$ from $T_q > 0$ to 0. In fact, we calculate

$$\frac{\mathrm{d}\widetilde{T}}{\mathrm{d}s}(s) = \frac{1 \quad q \quad \varrho}{2\varrho s^2\sqrt{\Delta_q(s)}} \left[\quad (1 \quad q \quad \varrho)^2 \quad s(1 \quad q+\varrho) + (1 \quad q \quad \varrho)\sqrt{\Delta_q(s)} \right].$$

For $s > \sigma_q^+$, we have (*) $(1 \quad q \quad \varrho)^2 + s(1 \quad q+\varrho) > 0$ (if this quantity were negative, we would have $s < \quad (1 \quad q \quad \varrho)^2/(1 \quad q+\varrho)$; but the inequality $(\sqrt{1 \quad q} + \sqrt{\varrho})^2 > 1 \quad q+\varrho$ implies in turn $\quad (1 \quad q \quad \varrho)^2/(1 \quad q+\varrho) < \sigma_q^+$ and then $s < \sigma_q^+$, a contradiction). It follows that for $s > \sigma_q^+$, inequality (*) and the identity

$$(1 \quad q \quad \varrho)^2\Delta_q(s) \quad \left((1 \quad q \quad \varrho)^2 + s(1 \quad q+\varrho)\right)^2 = \quad 4(1 \quad q)\varrho s^2$$

imply that $\mathrm{d}\widetilde{T}(s)/\mathrm{d}s > 0$ for $s > \sigma_q^+$ and the function $\widetilde{T}$ is monotonic decreasing on $[\sigma_q^+, +\infty[$, as claimed.

From definition (35), polynomial $R(t)$ has negative roots which cannot therefore be attained by $T\ (s) > 0$, $s > \sigma_q^+$. We conclude that $R(T\ (s))$ cannot vanish on this interval. As being well-defined on interval $[\sigma_q^+, +\infty[$, the Laplace transform U introduced in Eq. (35) is thus well-defined over the whole half-plane $\{s \in \mathbb{C},\ \Re(s) > \sigma_q^+\}$.

By Proposition 3, the generating series $J\ (z)$ is convergent for $|z| < \zeta_q$. We further verify that the value $U\ (\sigma_q^+)$ of the argument of J in (39) for $s = \sigma_q^+$ is less than this convergence radius ζ_q . In fact, expression (35) and simple algebra easily provide formula (37) given in the Proposition for $U\ (\sigma_q^+)$. It is then first easily checked that $U\ (\sigma_q^+) > 1$; in addition, the difference

$$\zeta_q \quad U\ (\sigma_q^+) = \frac{\sqrt{\varrho(1 \quad q)}}{q^2(q+\sqrt{(1 \quad q)\varrho})} \left[q + \sqrt{(1 \quad q)\varrho} \quad \sqrt{\varrho+q}\right]^2$$

is non negative and vanishes for $\varrho = 1 \quad q$ only, which is excluded by the stability condition (4); this consequently shows that $1 < U\ (\sigma_q^+) < \zeta_q$, as claimed.

Setting $U\ (s) = U(T\ (s))$ for short and using expansion (21) for $\widetilde{T}\ (s)$, we then have

$$U\ (s) = U(T_q) + L_q(s \quad \sigma_q^+)^{1/2} + o\left((s \quad \sigma_q^+)^{1/2}\right) \quad (40)$$

in the neighborhood of the singularity $s = \sigma_q^+$, where we set $L_q = U^0(T_q)S_q$ with constants T_q and S_q given in (22). We calculate $U^0(t) = (1 \quad \varrho \quad q)^2(1 \quad q \quad \varrho(q+\varrho)(1 \quad t)^2)/R(t)^2$ so that

$$U^0(T_q) = \frac{\varrho(\sigma_q^+)^2}{(1 \quad q \quad \varrho)(q+\sqrt{\varrho(1 \quad q)})^2}$$

hence the explicit expression (38) given in the Proposition for $L_q = U^0(T_q)S_q$. By expansion (40), transform (39) consequently expands at first order in $(s \quad \sigma_q^+)^{1/2}$ as

$$\mathbb{E}(e^{\ s\widetilde{\Omega}}) = J\ \left(U(T_q) + L_q(s \quad \sigma_q^+)^{1/2} + ...\right) = J\ (U\ (\sigma_q^+)) + H_q L_q(s \quad \sigma_q^+)^{1/2} + ...$$

where $H_q = \mathrm{d}_z J\ (U\ (\sigma_q^+))$ denotes the first derivative of J at point $U\ (\sigma_q^+)$. Applying the Tauberian theorem [12, Theorem 25.2] then provides estimate (36), as claimed. ■

V. Numerical results

To validate the accuracy of the propositions asserted in the previous sections, we simulate a Processor-Sharing system where jobs have exponentially distributed service times with unit mean and arrive in batches with geometrically distributed size with parameter q, according to a Poisson process with rate ϱ such that $\rho = \frac{\varrho}{1 \quad q} < 1$. We simulate batches arriving to the system in equilibrium. We have simulated more than 10^7 batches to compute distributions of random variables Ω and I_b as well as the associated random variable J.

In a first step, we examine the equi-probability Hypothesis 1. We compare the index of the last job of the tagged batch leaving the system (denoted by I_b) to the index J computed by randomly picking up a number of jobs equal to the size of the tagged batch. In Figures 3 and 4, we plot the probability density distribution of these two random variables as well as the approximation given by Equation (33).

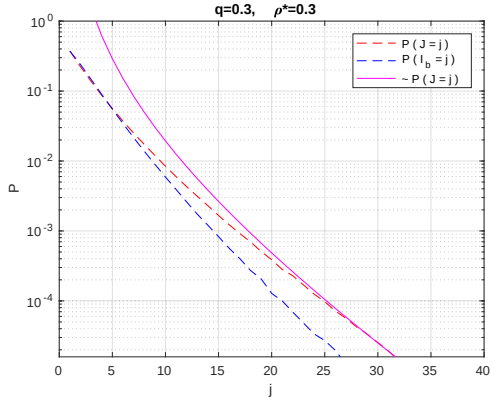

Fig. 3. *P(J=j) for* $q = 0.3$, $\varrho = 0.3$

In Figure 3, the load of the system and the mean batch size are rather small and the proposed approximation is quite accurate. In Figure 4, we increase the load and the batch size; the proposed approximation is still relevant for small indexes but becomes loose for larger ones. Nevertheless, we empirically observe that the proposed approximation yields an upper bound for the index of the last job of the tagged batch leaving the system.

We now consider the sojourn time Ω. Because of Hypotheses 2 and 3, the random variable $\widetilde{\Omega}$ cannot be easily estimated because the probability distribution of inter-departure times of

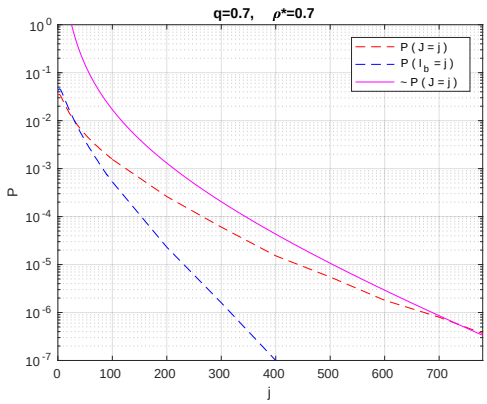

Fig. 4. *P(J=j) for* $q = 0.7$, $\varrho = 0.7$

jobs within a busy period is not known. Instead, we introduce another random variable $\hat{\Omega}$ equal to the departure time of the last batch, when picking up at random a number of jobs equal to the batch size and when setting the time origin equal to the tagged job arrival time.

In Figure 5, we plot the complementary cumulative distribution function of random variables Ω and $\hat{\Omega}$ for a light load and for both small and moderate mean batch size. We observe that the approximation is reasonably accurate. We have also represented approximation (36) for $\widetilde{\Omega}$. For computing the multiplying factor

$$H_q = \sum_{j=1}^{+1} j\mathbb{P}(J=j)\left(\frac{U\ (\sigma_q^+)}{\zeta_q}\right)^{j\ \ 1}$$

introduced in (36), we use the values of $\mathbb{P}(J=j)$, $j > 1$, obtained by simulation. It turns out that this approximation is much better than $\hat{\Omega}$ for large values of the mean batch size. The random variable $\hat{\Omega}$ is easy to simulate but difficult to study analytically while it is exactly the contrary for $\widetilde{\Omega}$.

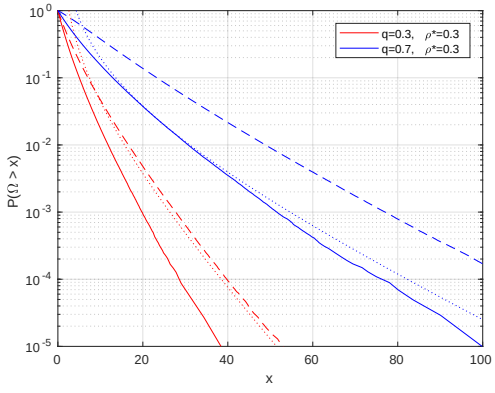

Fig. 5. *solid line,* ^ *dashed line,* ~ *dotted line*

As previously observed for the evaluation of variable J, the approximation is reasonably accurate for small values of the system load but becomes less accurate for larger values. As observed earlier, approximation (36) yields better results.

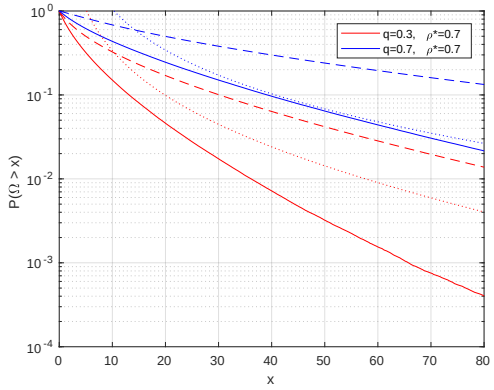

Fig. 6. *solid line,* ^ *dashed line,* ~ *dotted line*

VI. Conclusion

In this paper, we have considered the sojourn time of an entire batch in the $M^{[X]}/M/1$-PS system. Since this quantity is difficult to study analytically, we have introduced two approximations, one for the index of the last job of a tagged batch leaving the system (the index is obtained by labeling jobs according to their departure instants after the batch arrival) and another for the sojourn time of the entire batch. Simulations show that the proposed approximations give reasonable results.

From a practical point of view, we conclude from the computations carried out in this paper that the decay rate of the sojourn time of batch in the $M^{[X]}/M/1$-PS system is $j\sigma_q^+ j$ defined by Equation (7). By using results from [2], we can easily see that this decay rate is less than the one associated with the $M^{[X]}/M/C$ queue. Hence, if we introduce deadlines in the execution of VNFs, the rate of overrun will be higher in the $M^{[X]}/M/1$-PS than in the $M^{[X]}/M/C$ system. This confirms the earlier results obtained in [3] by simulation.

References

[1] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Network function virtualization in 5G," *IEEE Communications Magazine*, vol. 54, no. 4, pp. 84–91, 2016.

[2] V. Q. Rodriguez and F. Guillemin, "Cloud-ran modeling based on parallel processing," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 457–468, 2018.

[3] V. K. Quintuna Rodriguez and F. Guillemin, "Performance analysis of resource pooling for network function virtualization," in *Networking Conference*, Nov. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01621281

[4] M. Cromie, M. Chaudhry, and W. Grassman, "Further results for the queueing systems $M^{[X]}/M/C$," *J. Opl Res. Soc.*, vol. 30, no. 8, pp. 755–763, 1979.

[5] M. Andrews, "Probabilistic end-to-end delay bounds for earliest deadline first scheduling," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, vol. 2, March 2000, pp. 603–612 vol.2.

[6] F. Guillemin, V. K. Q. Rodriguez, and A. Simonian, "Sojourn time in a processor sharing queue with batch arrivals," *Stochastic Models*, vol. 34, no. 3, pp. 322–361, 2018.

[7] L. Kleinrock, R. Muntz, and E. Rodemich, "The processor sharing queueing model for time shared systems with bulk arrivals," *Networks*, 1971.
[8] J. Cohen, *The Single Server Queue*. North Holland Company, 1982.
[9] F. Oberhettinger and L. Badii, *Table of Laplace Transforms*. Springer Verlag, 1973.
[10] N. Lebedev, *Special functions and their applications*. Prentice Hall, 1965.
[11] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*. Cambridge University Press, 2009.
[12] G. Doetsch, *Einfuhrung in Theorie und Anwendung der Laplace Transformation*. Birkhauser, 1958.
[13] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. Dover Publications, 1965.

APPENDIX

A. Proof of Corollary 3

After the identity [13, Equ. (6.2.2)]

$$\int_0^1 t^{\alpha\ \ 1}(1\quad t)^{\beta\ \ 1}\mathrm{d}t = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

write

$$1\Big/\binom{m}{b} = b\int_0^1 t^{b\ \ 1}(1\quad t)^{m\ \ b}\,\mathrm{d}t.$$

Consequently, expression (32) equivalently reads

$$\mathbb{P}_{n,b,m}(J=j) = \frac{(j\quad 1)!}{(b\quad 1)!(j\quad b)!}\int_0^1 \frac{b\,t^{b\ \ 1}}{(1\quad t)^b}(1\quad t)^m\,\mathrm{d}t$$

and deconditioning with respect to variable $\widetilde{M}$ gives

$$\mathbb{P}_{n,b}(J=j) = \frac{(j\quad 1)!}{(b\quad 1)!(j\quad b)!} \int_0^1 \frac{b\,t^{b\ \ 1}}{(1\quad t)^b}\sum_{m=j}^{+\infty}(1\quad t)^m\mathbb{P}_{n,b}(\widetilde{M}=m)\,\mathrm{d}t. \qquad (41)$$

We now evaluate $\mathbb{P}_{b,n}(\widetilde{M}=m)$ for large $m > j$. Using the expression (10) of $M\ (z)$ with smallest singularity located at $z=\zeta_q$, we first have

$$M\ (z) = \frac{1+\varrho\quad q\zeta_q\quad q\sqrt{(\zeta_q^+\quad \zeta_q\ )(\zeta_q\quad z)}}{2\varrho} + ...$$

where dots denote $o\big(\sqrt{\zeta_q\quad z}\big)$ terms when $z\ !\ \zeta_q$; applying relation (14) to $s=0$ and with $\nu(z,0) = M\ (z)$, we deduce that

$$\mathbb{E}_{n,b}(z^{\widetilde{M}}) = r_q^{n+b} \left[1\quad (n+b)\frac{q\sqrt{(\zeta_q^+\quad \zeta_q\ )\zeta_q}}{(1+\varrho+q\zeta_q\ )}\sqrt{1\quad \frac{z}{\zeta_q}} + ...\right]$$

when $z\ !\ \zeta_q$, where we set $r_q = 2\zeta_q\ /(1+\varrho+q\,\zeta_q\ )$ for short. A direct application of Darboux's method [11, Theorem VI.14] then yields the asymptotics

$$\mathbb{P}_{n,b}(\widetilde{M}=m)\quad \kappa_q(n+b)r_q^{n+b}\frac{1}{m^{\frac{3}{2}}}\left(\frac{1}{\zeta_q}\right)^m$$

for large m, with constant κ_q set as in (34). Using the latter estimate of $\mathsf{P}_{n,b}(\widetilde{M}=m)$, we consequently deduce that

$$\sum_{m=j}^{+\infty}(1\quad t)^m\mathbb{P}_{b,n}(\widetilde{M}=m) \quad \frac{\kappa_q\zeta_q\ (n+b)r_q^{n+b}}{\zeta_q\quad 1+t}\frac{1}{j^{\frac{3}{2}}}\left(\frac{1\quad t}{\zeta_q}\right)^j$$

for large j so that expression (41) yields in turn

$$\mathbb{P}_{n,b}(J=j)\quad (n+b)r_q^{n+b}\frac{1}{j^{\frac{3}{2}}}\left(\frac{1}{\zeta_q}\right)^j \quad \frac{b}{(b\quad 1)!}\int_0^1 (jt)^{b\ \ 1}(1\quad t)^{j\ \ b}\frac{\kappa_q\zeta_q}{\zeta_q\quad 1+t}\,\mathrm{d}t \qquad (42)$$

where we have used the fact that $(j\quad 1)!/(j\quad b)!\quad j^{b\ \ 1}$ for large j and fixed b. To finally evaluate the integral appearing in (42) for large j, the variable change $u=jt$ simply provides

$$\frac{b}{(b\quad 1)!}\int_0^1 (jt)^{b\ \ 1}(1\quad t)^{j\ \ b}\frac{\kappa_q\zeta_q}{\zeta_q\quad 1+t}\,\mathrm{d}t = \frac{b}{j(b\quad 1)!}\int_0^j u^{b\ \ 1}\left(1\quad \frac{u}{j}\right)^{j\ \ b}\frac{\kappa_q\zeta_q}{\zeta_q\quad 1+u/j}\,\mathrm{d}u$$

$$\frac{b\kappa_q\zeta_q}{j(\zeta_q\quad 1)}\frac{\Gamma(b)}{(b\quad 1)!} = \frac{b\kappa_q\zeta_q}{j(\zeta_q\quad 1)}$$

by definition of the Euler Γ function; using the latter and estimate (42), we deduce

$$\mathbb{P}_{n,b}(J=j)\quad \frac{\kappa_q\zeta_q}{\zeta_q\quad 1}\frac{1}{j^{\frac{5}{2}}}\left(\frac{1}{\zeta_q}\right)^j\quad b(n+b)r_q^{n+b} \qquad (43)$$

for large j. We finally note that

$$r_q = \frac{2\zeta_q}{1+\varrho+q\zeta_q} < \frac{1}{\varrho+q}; \qquad (44)$$

in fact, calculating

$$\delta_q\left(\frac{1+\varrho}{2\varrho+q}\right) = \quad \frac{4\varrho(1+\varrho)(1\quad q\quad \varrho)(q+\varrho)}{(q+2\varrho)^2} < 0$$

together with condition (4) show that

$$\frac{1+\varrho}{2\varrho+q} > \zeta_q$$

hence inequality (44); this consequently ensures that $\mathbb{E}(r_q^{N_0+B}) < +\infty$ after (18). Deconditioning each side of (43) on variables N_0 and B then provides asymptotics (33), with associated constant

$$K_q = \frac{\kappa_q\zeta_q}{\zeta_q\quad 1}\ \ \mathsf{E}\left[B(N_0+B)r_q^{N_0+B}\right]. \qquad (45)$$

Using the respective definitions (3) and (17) of generating function B and η , it is easily verified that the expectation in (45) equals

$$\mathsf{E}\left[B(N_0+B)r_q^{N_0+B}\right] = r_q^2\frac{\mathrm{d}B}{\mathrm{d}z}(r_q)\,\frac{\mathrm{d}\eta}{\mathrm{d}z}(r_q) + r_q\left(\frac{\mathrm{d}B}{\mathrm{d}z}(r_q) + r_q\frac{\mathrm{d}^2B}{\mathrm{d}z^2}(r_q)\right)\eta(r_q);$$

the latter together with (45) yield the final expression (34) of constant K_q after simple algebra.