Lévy-driven polling systems and continuous-state branching processes

Boxma, O.; Ivanovs, J.; Kosiński, K.M.; Mandjes, M.

DOI
10.1214/10-SSY008

Publication date
2011

Document Version
Final published version

Published in
Stochastic Systems

Stochastic Systems
2011, Vol. 1, No. 2, 411–436
DOI: 10.1214/10-SSY008

LÉVY-DRIVEN POLLING SYSTEMS AND CONTINUOUS-STATE BRANCHING PROCESSES

By Onno Boxma[†,‡], Jevgenijs Ivanovs[†,§], Kamil Kosiński[*,†,§] and Michel Mandjes[†,§,¶]

EURANDOM[†] and Eindhoven University of Technology[‡] and University of Amsterdam[§] and CWI[¶]

In this paper we consider a ring of $N \geq 1$ queues served by a single server in a cyclic order. After having served a queue (according to a service discipline that may vary from queue to queue), there is a switch-over period and then the server serves the next queue and so forth. This model is known in the literature as a *polling model*.

Each of the queues is fed by a non-decreasing Lévy process, which can be different during each of the consecutive periods within the server's cycle. The N-dimensional Lévy processes obtained in this fashion are described by their (joint) Laplace exponent, thus allowing for non-independent input streams. For such a system we derive the steady-state distribution of the joint workload at embedded epochs, i.e. polling and switching instants. Using the Kella-Whitt martingale, we also derive the steady-state distribution at an arbitrary epoch.

Our analysis heavily relies on establishing a link between fluid (Lévy input) polling systems and multi-type Jiřina processes (continuous-state discrete-time branching processes). This is done by properly defining the notion of the *branching property* for a discipline, which can be traced back to Fuhrmann and Resing. This definition is broad enough to contain the most important service disciplines, like exhaustive and gated.

1. Introduction. Consider a queueing model consisting of multiple queues attended by a single server, visiting the queues one at a time in a cyclic order. Moving from one queue to another, the server incurs a non-negligible switch-over time. Such single-server multiple-queue models are commonly referred to as *polling models*. Stimulated by a wide variety of applications, polling models have been extensively studied in the literature, see [28, 30, 31] for a series of comprehensive surveys and [20, 29] for extensive overviews of the applicability of polling models.

Received July 2010.

[*]The author was supported by NWO grant 613.000.701.

AMS 2000 subject classifications: Primary 60K25; secondary 90B22.

Keywords and phrases: Polling system, Lévy processes, branching processes.

Throughout the vast polling literature, it is almost always assumed that customers arrive at the queues according to independent Poisson processes, where, in addition, the service requirements brought along by these customers are i.i.d. sequences; the resulting input processes in the queues thus constitute independent compound Poisson processes (CPPs). Correlated arrivals in polling models have received little attention; see Levy and Sidi [19] for a treatment of polling models with correlated CPP input. Classical analysis of polling systems heavily focuses on keeping track of customers in the system at embedded epochs, i.e., instants of specific changes in the system, like *polling instants* or *switching instants*.

A key feature of polling models is the *service discipline*. A service discipline specifies the rule that determines how long a server will visit a queue (and process any workload found there). The most important and well known disciplines include the *exhaustive* discipline, *gated* discipline and *1-limited* discipline. Under the exhaustive discipline, the server will stay at the queue until this queue has become empty. Under the gated discipline, the server serves exactly the customers (or: the amount of work) present upon the beginning of the visit. Under the 1-limited discipline, the server serves only one customer – if there is one. The 'system's service discipline' can be any 'mixture' of the individual disciplines; for instance: some of the queues are served according to a gated discipline, whereas others are served exhaustively.

In Resing [23] (see also Fuhrmann [14]) it is shown that for a large class of classical polling models, including those with exhaustive and gated service at all queues (but not 1-limited), the evolution of the system at successive polling instants at a fixed queue can be described as a multi-type *branching process* (MTBP) with immigration. Models that satisfy this MTBP-structure allow for an exact analysis, whereas models that violate the MTBP-structure are considerably more intricate, and therefore usually intractable. It turns out that it is exactly the nature of the service disciplines which determines the MTBP structure of the system. The structure is preserved if each service discipline satisfies a special property called the *branching property*, see e.g. Property 1 of [23]. The exhaustive and gated disciplines do satisfy this property whereas the 1-limited does not. A key result for polling models with the MTBP structure is the joint steady-state distribution of the queue length (i.e., in terms of the number of customers) at polling or switching instants of a particular queue.

In this paper we generalize the classical assumptions in several ways. We consider polling models with Lévy-driven, possibly correlated, input streams. More specifically, we assume that the input process $\boldsymbol{W}$ is an N-dimensional Lévy subordinator, where $N \geq 1$ corresponds to the number of queues, and where 'subordinator' means that the corresponding sample paths are non-decreasing (in all N coordinates). We refer to this model as a *Lévy-driven polling model* with input process $\boldsymbol{W}$. If the queue under consideration, say queue i, is not in service, its workload evolves according to the subordinator $W_i(t)$, whereas during its service time it is described by $W_i(t)$ minus drift t. It is important to note that, in fact, this paper considers a slightly richer class of models, in which the workload level while in service behaves as a spectrally positive Lévy process A_i with negative drift (that decreases on average); here 'spectrally positive' means that the underlying process has positive jumps only. We remark that the class of spectrally positive Lévy processes with negative drift is used frequently in the theory of *storage processes* to model the storage level (workload) of queues, dams or fluid models, see e.g. Kyprianou [18], and Prabhu [22] for an early reference.

We recall that Lévy processes are processes with stationary, independent increments; it is stressed, however, that the components of the N-dimensional Lévy process need not necessarily be independent. The class of Lévy processes is rich and covers Brownian motion, linear increment processes and CPPs as special cases. The generalization from CPPs to Lévy input implies that we can no longer speak of notions such as customers and queue lengths; this explains why we focus on the (joint) *workload* process. While quite a few studies have been devoted to a single server single queue model with Lévy input (see, e.g., Prabhu [22, Chapter 4] and Asmussen [5, Chapter 14]), there is hardly any literature on Lévy-driven polling systems. An exception is Eliazar [13], who considers such systems only for the gated discipline, independent input processes and does not allow for spectrally positive Lévy processes. His analysis follows a dynamical-systems approach: a stochastic Poincaré map, governing the one-cycle dynamics of the polling system is introduced, and its statistical characteristics are studied. This approach differs from ours; we identify a branching structure in Lévy-driven polling models as will be explained later. By considering the input as an N-dimensional Lévy process $\boldsymbol{W}$ instead of N one-dimensional processes W_i, we accomplish an easy incorporation of *correlation* between the inputs to different queues. This is due to the fact that every Lévy process is uniquely characterized by its characteristic exponent, which in the multidimensional case also captures the correlation structure between the individual components.

Considering polling models with Lévy input opens several new perspectives. Firstly, the theory of Lévy processes was strongly developed in recent years, and its application appears to lead to more simplified derivations of many results which, for the case of compound Poisson input, are only obtained after detailed calculations. Secondly, having Lévy input leads to significant generalizations of known results. Such generalizations are theoretically interesting, but also, owing to the inherent flexibility of Lévy processes, offer various new possibilities from the viewpoint of applications. Polling models have found applications in many different areas, like (i) Maintenance (a patrolling repairman); (ii) Stochastic Economic Lotsizing (a machine producing products of various types upon demand); (iii) road traffic (traffic lights at signalized intersections); and (iv) protocols in computer and communication networks (Bluetooth; token ring protocols; protocols for web servers and routers). Almost invariably, it has been assumed in the polling literature that the input process is composed of a number of independent compound Poisson processes. We allow Lévy input, and correlation between the various input streams, and different input processes during different visit and switch-over periods. This gives much additional modelling capability. E.g., in stochastic economic lotsizing it is quite natural to have correlations between the arrival processes of demands of different product types. And in road traffic as well as in communications, it is sometimes better to model traffic as a fluid than as separate customers; indeed, a special case of our model is the situation in which there is a constant fluid input in one queue of the polling model, and a compound Poisson input in an other queue. As another example, while a Brownian motion component may not be natural in representing work inflow, it may represent realistic fluctuations in the speed of the server. Recall that a served queue is modelled by a spectrally positive Lévy process with negative drift allowing for incorporation of a Brownian component.

The transition from CPPs to Lévy subordinators deprives us from the possibility of using the branching property from Resing [23], which is stated in terms of customers (which are of a discrete nature) in the system, and therefore has no simple translation to our continuous state-space setting. In our paper we identify the analogous property, also referred to as the branching property, for the disciplines in the Lévy framework, that enables to identify a branching structure in our system. This allows us to mimic Resing's approach, and to describe the multidimensional workload in the system at successive polling instants at a fixed queue as a multi-type continuous state-space (discrete time) branching process. This branching process is referred to in the sequel as multi-type Jiřina branching process (MTJBP) due to Jiřina [16], who introduced the notion of continuous state-space branching processes and paid special attention to discrete-time processes (called *Jiřina processes* in the literature). The relation between Lévy-driven polling models and continuous state-space branching processes has been observed before by

Altman and Fiems [4] in a special case strongly relying on the assumption imposed that all the queues are fed by *identical* Lévy subordinators. This relation was only used to derive the first two waiting-time moments and did not focus on the underlying structure of the branching process. The observation that Lévy-driven polling models can be completely characterized (in terms of distribution) by a continuous-state space branching process of the MTJBP form is therefore novel.

Another performance measure which is analyzed in the paper, again for disciplines satisfying our new branching property, is the Laplace-Stieltjes Transform (LST) of the steady-state distribution of the joint workload in the queues at an *arbitrary* epoch. The classical polling literature focuses strongly on joint queue lengths at polling epochs, and contains results for *marginal* queue lengths and workloads at arbitrary epochs, but we are not aware of any general results for *joint* queue-length or workload distributions at arbitrary epochs – with the exception of the recent paper by Czerniak and Yechiali [12] that considers constant fluid input at all queues for a very special case of our model. We employ the Kella-Whitt martingale [17] to obtain this result. A similar approach has been used before; e.g., Boxma et al. [10] give the steady-state storage level transform for a Lévy-driven queueing model with service vacations.

Contribution. This paper casts a broad class of queueing models into a single general framework. More specifically the contributions are the following. First, we consider general, Lévy-driven polling models instead of the classical models with CPP inflow. Second, we let the input $\boldsymbol{W}$ change at polling and switching instants, whereas in classical polling models the input processes are typically fixed once and for all. Third, we allow for correlation between the individual input processes (correlated arrivals received little attention so far, see [19]). Fourth, we introduce a new class of service disciplines satisfying a novel branching property, and we relate Lévy-driven polling models to MTJBP. This class is broad and contains the well known exhaustive and gated disciplines. Fifth, we provide the LST of the joint steady-state workload distribution at an arbitrary epoch, which is a new result even for classical polling models. Finally, we show that the stability of our system does not depend on the disciplines used at different queues, and can be formulated in terms of rates of input (which leads to an intuitively appealing criterion).

Organization of the paper. The remainder of this paper is organized as follows. In Section 2 we describe the model and the service disciplines that are considered in this paper. Section 3 presents a brief introduction on MTJBPs, and some additional intuition behind these processes. We also state a limit theorem for MTJBPs with immigration. Section 4 contains one of the two main theorems in this paper. It is shown that in our model the workload level at different queues at successive epochs that the server reaches a fixed queue is an MTJBP with immigration. This leads to an expression for the LST of the stationary joint workload distribution at different queues at these epochs. Section 5 contains the second main theorem of this paper. We derive the LST of the stationary joint workload distribution at an arbitrary epoch. In Section 6 we show that our results carry over in a straightforward way to the situation in which $\boldsymbol{W}$ changes between polling and switching instants. In Section 7 we present a discussion of the ergodicity of the most general model, i.e., the model addressed in Section 6. We conclude in Section 8 by suggesting possible further generalizations.

Notation. In the sequel, for any random variable X we denote its LST by $\tilde{X}$, i.e., $\tilde{X}(u) = \mathbb{E}e^{-uX}$ for $u \in \mathbb{R}$ such that $\tilde{X}$ exists. Throughout this paper, we will only use *spectrally positive* Lévy processes, that is, processes which are allowed to have positive jumps only (and therefore containing the class of all aforementioned Lévy subordinators). Recall from Section 1 that the class of spectrally positive Lévy processes is used frequently in the theory of storage processes, see, e.g., [18, 22]. For a background on Lévy processes see e.g. [18, 24]. Such a process $\{L(t), t \geq 0\}$ can be uniquely characterized by its Laplace exponent: a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ say, such that $\mathbb{E}e^{-uL(t)} = e^{-t\phi(u)}$.

To explicitly distinguish the multidimensional case from the single-dimensional case, we will use bold symbols to denote vectors and plain symbols to denote their coordinates, so that $\boldsymbol{u} \equiv (u_1, \ldots, u_d)$ for some $d > 1$. The inner product of two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ will be denoted by $\boldsymbol{u} \cdot \boldsymbol{v}$. Finally, for a random vector $\boldsymbol{X}$ and multidimensional spectrally positive Lévy process $\{\boldsymbol{L}(t), t \geq 0\}$ we define its LST and Laplace exponent analogously to the corresponding one-dimensional objects. All inequalities for vectors should be understood coordinate-wise. For sequences of one-dimensional objects we will use subscripts and write them as (a_n), whereas in the multidimensional case we will sometimes use superscripts and write them as $(\boldsymbol{a}^n)$ to avoid double (or sometimes even triple) subscripts.

As mentioned above, we focus on the system's workload process, and would like to show that it possesses a specific branching structure. To this end, we talk about *children* of a workload portion x. Because we can treat a portion of workload x as any number n of smaller portions x/n, we need to be able to talk about infinitesimally small portions of workload. That is why we shall adopt the language of fluid queues, and an infinitesimally small portion of workload can be regarded as a *drop*. Hence, from now on we shall use the terms workload and amount of fluid interchangeably.

2. Model description. We consider a system of N infinite-buffer fluid queues, $Q_1, \dots, Q_N$, and a single server. The server moves along the queues in a cyclic order. When leaving Q_j and before moving to Q_{j+1} (where, by convention, Q_{N+1} should be understood as Q_1), the server incurs a switch-over period whose duration is a positive random variable S_j independent of anything else. Queues are fed by an N-dimensional Lévy subordinator $\boldsymbol{W} = \{\boldsymbol{W}(t), t \geq 0\}$ with Laplace exponent ϕ. The server's work at Q_i is modelled by a spectrally positive Lévy process A_i with negative drift (i.e., $\mathbb{E}A_i(1) < 0$), so that the work in the system evolves (while the server is at Q_i) according to a Lévy process

$$\boldsymbol{A}_i(t) := \Big(W_1(t), \dots, W_{i-1}(t), A_i(t), W_{i+1}(t), \dots, W_N(t)\Big).$$

The Laplace exponent of $\boldsymbol{A}_i(t)$ is denoted through $\phi_i^{\boldsymbol{A}}(\boldsymbol{u})$.

Remark 1. Taking a spectrally positive Lévy process A_i rather than just a Lévy subordinator W_i allows for usage of a slightly bigger class of input processes. For instance one can have a Brownian component as a component in an input process. The use of reflected Brownian motion is quite common in queueing theory, as it is the limiting model for a wide class of queueing models under the functional central limit theorem, see, e.g., Whitt [32].

Remark 2. In Section 2, Section 4 and Section 5 the input process $\boldsymbol{W}$ remains fixed. In Section 6 we show that one can still analyze the joint workload process if $\boldsymbol{W}$ is allowed to change at polling and switching instants. In classical polling models this could correspond to, e.g., having arrival rate λ_{ij} at Q_i when the server is at Q_j.

The service disciplines that we consider in this paper satisfy the following property.

Property 1. *If the server arrives at Q_i to find the workload level x there, then during the course of the server's visit, this workload is replaced by $\boldsymbol{H}_i(x)$, where $\{\boldsymbol{H}_i(x), x \geq 0\}$ is an N-dimensional Lévy subordinator with Laplace exponent η_i, which can be any Laplace exponent corresponding to an N-dimensional subordinator.*

In other words, if the server finds workload vector $\boldsymbol{x}$ at the time of arrival at Q_i then the workload vector at the end of the service of this queue becomes $\boldsymbol{x} - x_i\boldsymbol{e}_i + \boldsymbol{H}_i(x_i)$. Note that any replacement process should stay positive so that work does not become negative in Q_i and does not decrease in the other queues. It is also obvious that such a process should be increasing in x. Therefore the assumption that $\boldsymbol{H}_i(x)$ is a subordinator is intuitively clear and natural. Moreover, due to the independent stationary increments property of any Lévy process, we have that $\boldsymbol{H}_i(x+y) =_{\mathrm{d}} \bar{\boldsymbol{H}}_i(x) + \bar{\boldsymbol{H}}_i(y)$, where $\bar{\boldsymbol{H}}_i(x)$ and $\bar{\boldsymbol{H}}_i(y)$ are independent with the same distribution as $\boldsymbol{H}_i(x)$ and $\boldsymbol{H}_i(y)$, respectively (and '$=_{\mathrm{d}}$' denotes equality in distribution). Note that this properties essentially says that each *drop* of the fluid in the served queue is treated in an i.i.d. manner. It is further observed that Property 1 is a continuous analogue of the branching property from Fuhrmann [15]. Note that we allow different service disciplines at different queues, as long as they obey Property 1.

Examples. It is readily verified that the important exhaustive and gated disciplines both satisfy Property 1.

The gated discipline. Under the gated discipline, the server only serves the workload that was present at the start of the visit. Fluid flowing into the queue during the course of the visit is served in the next visit. Assuming (c.f., Remark 3 below) that the server works with rate 1, i.e. $A_i(t) = W_i(t) - t$, and finds an amount of work x upon arrival, the time $\tau_i(x)$ spent in Q_i is simply $\tau_i(x) = x$, so that

$$\boldsymbol{H}_i(x) = \boldsymbol{W}(\tau_i(x)) = \boldsymbol{W}(x)$$

and

$$\eta_i(\boldsymbol{u}) = \phi(\boldsymbol{u}). \tag{1}$$

REMARK 3. Observe that in the case of the gated discipline we assumed that $A_i(t) = W_i(t) - t$, so the workload level in Q_i, during the server's visit, behaves as $x + W_i(t) - t$, where x is the starting level. This is a standard assumption in the theory of *storage processes*, although, in principle, A_i could be any spectrally positive Lévy process with negative drift. Such processes are used frequently to model the workload level in fluid queues. The gated discipline, however, becomes ill-defined in such a general setting, thus every time we speak of the gated discipline we tacitly assume that the server works with rate 1.

The exhaustive discipline. Verification of the validity of Property 1 for exhaustive service is somewhat more involved. To this end, first recall that the server continues to work until the queue becomes empty. Fluid arriving during the course of the visit is served in the current visit. Let $T_i(x) := \inf\{t \geq 0 : A_i(t) = -x\}$ be the time needed to empty the queue with initial workload level x. It is known that for $A_i(t)$ a spectrally positive Lévy process with negative drift, $T_i(x)$ is an a.s. finite stopping time. Using the property of stationary and independent increments generalized to stopping times [18, Theorem 3.1] one can see that $\boldsymbol{H}_i(x)$ is a Lévy process. We need to compute

$$\mathbb{E}\exp(-\boldsymbol{u}\cdot\boldsymbol{H}_i(x)) = \mathbb{E}\exp\left(-\sum_{j\neq i} u_j W_j(T_i(x))\right),$$

which does not depend on u_i. Now consider, for a fixed vector $\boldsymbol{u} \geq \boldsymbol{0}$, the function $\phi_{i,\boldsymbol{u}}(\theta) := \phi_i^{\boldsymbol{A}}(u_1, \dots, u_{i-1}, \theta, u_{i+1}, \dots, u_N)$, where $\theta \geq 0$. It is easy to see that $\phi_{i,\boldsymbol{u}}(0) \geq 0$, because $W_j(t), j \neq i$ are subordinators. Moreover, $\lim_{\theta\to\infty} \phi_{i,\boldsymbol{u}}(\theta) = -\infty$, because $A_i(t)$ is not a subordinator. It follows from the continuity of $\phi_{i,\boldsymbol{u}}(\theta)$ that there exists a number $\psi_i(\boldsymbol{u}) \geq 0$ such that $\phi_{i,\boldsymbol{u}}(\psi_i(\boldsymbol{u})) = 0$. Then consider Wald's martingale $\exp(-\boldsymbol{u}\cdot\boldsymbol{A}_i(t) + \phi_i^{\boldsymbol{A}}(\boldsymbol{u})t)$ and pick $u_i = \psi_i(\boldsymbol{u})$. Application of the optional sampling theorem to the a.s. finite stopping time $T_i(x)$ gives $\mathbb{E}\exp(-\sum_{j\neq i} u_j W_j(T_i(x)) + \psi_i(\boldsymbol{u})x) = 1$. Hence $\mathbb{E}e^{-\boldsymbol{u}\boldsymbol{H}_i(x)} = e^{-\psi_i(\boldsymbol{u})x}$ and

$$\eta_i(\boldsymbol{u}) = \psi_i(\boldsymbol{u}). \tag{2}$$

Mixed disciplines. For a fixed $p \in [0,1]$ one can require that the fraction p of the present workload (upon arrival) is handled according to some branching discipline $\boldsymbol{H}^{(1)}$ and the rest according to a different branching discipline $\boldsymbol{H}^{(2)}$. Then,

$$\boldsymbol{H}_i(x) =_{\mathrm{d}} \boldsymbol{H}^{(1)}(px) + \boldsymbol{H}^{(2)}((1-p)x)$$

is a branching discipline. For instance, one can consider a *p-exhaustive* discipline, where it is required that the initial amount of workload x be reduced to the level px. Then,

$$\boldsymbol{H}_i(x) =_{\mathrm{d}} px + \boldsymbol{H}^{\mathrm{exhaustive}}((1-p)x)$$

is a branching discipline with Laplace exponent

$$\eta_i(\boldsymbol{u}) = pu_i + \psi_i(\boldsymbol{u})(1-p).$$

Composition of disciplines. Because the composition of two Lévy processes is again a Lévy process, composition of two branching disciplines is again a branching discipline. In other words, one can consider a discipline, when upon finishing its service with initial workload level x according to a branching discipline $\boldsymbol{H}^{(1)}$ (with Laplace exponent $\eta^{(1)}$), the server immediately starts to work again with initial workload level $\boldsymbol{H}_i^{(1)}(x)$ according to a branching discipline $\boldsymbol{H}^{(2)}$ (with $\eta^{(2)}$). That is,

$$\boldsymbol{H}_i(x) = \boldsymbol{H}^{(2)}\left(\boldsymbol{H}^{(1)}(x)\right)$$

is a branching discipline with Laplace exponent

$$\eta_i(\boldsymbol{u}) = \eta^{(2)}\left(\eta^{(1)}(\boldsymbol{u})\right).$$

Naturally, one can think of the composition of more than two branching disciplines.

General method. Recall that $T_i(x) := \inf\{t \geq 0 : A_i(t) = -x\}$. In the above examples we saw that as soon as we know the time $\tau_i(x) \leq T_i(x)$ that the server spends serving Q_i if it finds the workload level x upon its arrival, the replacement process $\boldsymbol{H}_i(x)$ can be written as $x\boldsymbol{e}_i + \boldsymbol{A}_i(\tau_i(x))$. This relation will be further exploited in Section 5.

3. Multi-type Jiřina branching processes. The observation that the theory of branching processes of Bienaymé-Galton-Watson type can be extended to random variables taking their values in a continuous state-space appears to be due to Jiřina [16], who points out that the key to such an extension is to make the *offspring distribution* infinitely divisible. The effect of an initial quantity of *parent*, may then be described by raising the Laplace exponent of the number of offspring per unit parent to the appropriate non-negative power; note the similarity with the discrete case, where the generating function of the distribution of offspring per individual is raised to a non-negative integral power.

More formally, let $\{R_{i,j}, i, j \geq 0\}$ be a sequence of integer-valued independent random variables, all distributed as R, and let $\{G, G_j, j \geq 1\}$ be a sequence of i.i.d. integer-valued random variables. Then the Bienaymé-Galton-Watson process with immigration is defined as

$$X_{n+1} = \sum_{j=1}^{X_n} R_{n+1,j} + G_{n+1}, \quad n = 0, 1, \ldots,$$

where X_0 is a given (integer-valued) starting random variable. In the terms of LST's, we have

$$\mathbb{E}\left(z^{X_{n+1}}|\mathcal{F}_n^X\right) = \left(\mathbb{E}z^R\right)^{X_n}\mathbb{E}z^G, \quad |z| \le 1.$$

Now consider an i.i.d. sequence of Lévy subordinators $\{R_n(x), x \ge 0\,, n = 1, 2, \ldots\}$ characterized by a common Laplace exponent κ and independent from $\{G, G_j, j \ge 1\}$. Then the sequence X_n, where

$$X_{n+1} = R_{n+1}(X_n) + G_{n+1}, \quad n = 0, 1, \ldots, \tag{3}$$

is called *Jiřina branching process* with immigration. Note that every part of X_n reproduces in an i.i.d. way, because R_{n+1} is a Lévy subordinator. The distribution corresponding to κ may be interpreted as describing the quantity of offspring per unit quantity of parent. Models of this type have been analyzed in various papers, see e.g. [21, 26, 27]. Finally, observe that from (3) it follows that

$$\mathbb{E}\left(e^{-uX_{n+1}}|\mathcal{F}_n^X\right) = e^{-X_n\kappa(u)}\tilde{G}(u), \quad u \ge 0. \tag{4}$$

This relation expresses what has been said in the first paragraph of this section.

In a natural way this concept extends to the *multi-type* case, where each type reproduces independently from others and gives rise to a multi-type population. For $i = 1, \ldots, N$, let $\{\boldsymbol{R}_i, \boldsymbol{R}_i^n(x), x \ge 0, n \ge 0\}$ be mutually independent sequences of i.i.d. N-dimensional Lévy subordinators with Laplace exponent κ_i, and let $\{\boldsymbol{G}, \boldsymbol{G}^n, n \ge 1\}$ be a sequence of non-negative N-dimensional i.i.d. random vectors. We define the one-step evolution of the process $\boldsymbol{X}$ through

$$\boldsymbol{X}^{n+1} = \sum_{i=1}^{N} \boldsymbol{R}_i^{n+1}(X_i^n) + \boldsymbol{G}^{n+1}, \tag{5}$$

where $\boldsymbol{R}^{i,n+1}$ and $\boldsymbol{G}^{n+1}$ are assumed to be independent of $\mathcal{F}_n^{\boldsymbol{X}}$ and $\boldsymbol{X}^0$ is a given starting random vector. Equation (4) then becomes

$$\mathbb{E}\left(e^{-\boldsymbol{u}\cdot\boldsymbol{X}^{n+1}}|\mathcal{F}_n^{\boldsymbol{X}}\right) = e^{-\boldsymbol{X}^n\cdot\boldsymbol{\kappa}(\boldsymbol{u})}\tilde{\boldsymbol{G}}(\boldsymbol{u}). \tag{6}$$

Such a sequence will be called a *multi-type Jiřina branching process* (MTJBP) with *branching mechanism* $\boldsymbol{\kappa}$ and immigration $\boldsymbol{G}$.

Let $m_{i,j}$ be the expected quantity of type j offspring per unit quantity of parent population of type i, i.e.

$$m_{i,j} = \mathbb{E}R_{i,j}(1) = \frac{\partial \kappa_i}{\partial u_j}(0).$$

An essential role is played by what we will call the *mean matrix* $M \equiv (m_{i,j})_{i,j=1,\ldots,N}$. Note that M is a non-negative matrix, so by the Perron-Frobenius theory the spectral radius ρ_M of M is an eigenvalue such that any other eigenvalue is strictly smaller in absolute value.

Define $\boldsymbol{\kappa}^{(i)}(\boldsymbol{u})$ inductively by $\boldsymbol{\kappa}^{(0)}(\boldsymbol{u}) = \boldsymbol{u}$ and $\boldsymbol{\kappa}^{(i)}(\boldsymbol{u}) = \boldsymbol{\kappa}^{(i-1)}(\boldsymbol{\kappa}(\boldsymbol{u}))$, for $i = 1, 2, \ldots$. Finally, let $\|\cdot\|$ denote any norm on $\mathbb{R}^N$.

THEOREM 1. *Let $\|\mathbf{G}\|$ be integrable then the following holds*

- *if $\rho_M < 1$ (subcritical case) then $\boldsymbol{X}^n$ converges in distribution to a random vector $\boldsymbol{X}^\infty \in \mathbb{R}_+^N$ satisfying*

$$\mathbb{E}e^{-\boldsymbol{u}\cdot\boldsymbol{X}^\infty} = \prod_{k=0}^{\infty} \tilde{\boldsymbol{G}}(\boldsymbol{\kappa}^{(k)}(\boldsymbol{u})), \quad \boldsymbol{u} \geq \boldsymbol{0}; \tag{7}$$

- *if $\rho_M > 1$ (supercritical case) and $G_i > 0$ with positive probability for all i then $\|\boldsymbol{X}^n\| \to \infty$ a.s. as $n \to \infty$.*

We remark that the case of $\rho_M = 1$ is substantially more involved and its treatment is beyond the scope of this paper.

PROOF. The first claim follows from [3]. We prove the second claim. Let $\boldsymbol{v}$ be a non-negative eigenvector associated to $\rho_M > 1$; such a vector exists according to the Perron-Frobenius theory. Note that $\kappa_i(\theta\boldsymbol{v})$ as a function of θ, is the Laplace exponent of the Lévy process $\boldsymbol{v}\cdot\boldsymbol{R}_i(t)$, and hence it is concave with derivative at 0 equal to $(m_{i,1}, \ldots, m_{i,N})\cdot\boldsymbol{v} = \rho_M v_i > v_i$. Hence we can choose $\theta_i^* > 0$ such that $\kappa_i(\theta\boldsymbol{v}) \geq \theta v_i$ for all $\theta \in [0, \theta_i^*]$. Let $\theta^* > 0$ be the minimum over θ_i^*, then

$$\boldsymbol{\kappa}(\theta^*\boldsymbol{v}) \geq \theta^*\boldsymbol{v}. \tag{8}$$

Combining (6) and (8) we get

$$0 \leq \mathbb{E}e^{-\theta^*\boldsymbol{v}\cdot\boldsymbol{X}^{n+1}} \leq \mathbb{E}e^{-\theta^*\boldsymbol{v}\cdot\boldsymbol{X}^n}\tilde{\boldsymbol{G}}(\theta^*\boldsymbol{v}) \leq (\tilde{\boldsymbol{G}}(\theta^*\boldsymbol{v}))^{n+1} \to 0$$

as $n \to \infty$, because $\mathbb{E}e^{-\theta^*\boldsymbol{v}\cdot\boldsymbol{G}} < 1$. Hence $\theta^*\boldsymbol{v}\cdot\boldsymbol{X}^n \to \infty$ and so also $\|\boldsymbol{X}^n\| \to \infty$ with probability 1. $\square$

It is easy to see from the final part of the proof that the additional condition in the supercritical case, namely $G_i > 0$ with positive probability for all i, can be substituted with the following one. It is enough to assume that $\|\boldsymbol{G}\|$ is not identically 0 and there exists a positive vector $\boldsymbol{v}$ such that $M\boldsymbol{v} > \boldsymbol{v}$. It is known that such a vector exists if M is irreducible. We do not assume irreducibility of M, because our polling system with exhaustive discipline at the N-th queue will correspond to a matrix M which is not irreducible. We elaborate more on the stability issue in Section 7.

4. Polling systems and multi-type continuous-state branching processes. In this section we shall prove that for our model the amounts of fluid in the N queues on successive epochs that the server reaches Q_1 form an MTJBP with immigration, which is one of the main results of this paper. Define t_n as the (random) time point that the server reaches Q_1 for the nth time. Let t_0 correspond to time 0.

THEOREM 2. *Consider a polling system from Section 2 with switch-over times S_i. Assume that the service discipline at Q_i satisfies Property 1 with Laplace exponent η_i, $i = 1, \ldots, N$. Then the amount of fluid $\boldsymbol{B}^n$ in the different queues at time points t_n constitutes a multi-type Jiřina branching process with immigration, where the branching mechanism $\boldsymbol{\kappa}$ is given by the recursive equations*

$$\kappa_i(\boldsymbol{u}) = \eta_i\left(u_1, \ldots, u_i, \kappa_{i+1}(\boldsymbol{u}), \ldots, \kappa_N(\boldsymbol{u})\right), \quad i = 1, \ldots, N, \tag{9}$$

and the immigration LST $\tilde{\boldsymbol{G}}(\boldsymbol{u})$ is given by

$$\tilde{\boldsymbol{G}}(\boldsymbol{u}) = \prod_{i=1}^{N} \tilde{S}_i\left(\phi(u_1, \ldots, u_i, \kappa_{i+1}(\boldsymbol{u}), \ldots, \kappa_N(\boldsymbol{u}))\right). \tag{10}$$

We tacitly assume that κ_N is read as η_N, the argument of $\tilde{S}_N$ in (10) is $\phi(\boldsymbol{u})$ and $\boldsymbol{B}^0$ is a given starting distribution. Importantly, an immediate consequence of Theorem 2 is that we can use Theorem 1 to obtain the limiting (*steady-state*) distribution of the joint workload $\boldsymbol{B}^\infty$ at polling epochs for our polling model, cf. (7).

PROOF. Consider the polling system at time t_n and assign the color c_i to the fluid in Q_i and denote its amount through x_i, for all $i = 1, \ldots, N$. Now suppose that fluid arriving during switch-overs has color c_0, and fluid arriving during the service of c_i-colored fluid has the same color c_i for all $i = 0, \ldots, N$. We stress that our coloring depends on the color of the fluid in service, i.e., *not* on the queue. Given x_i – the amount of fluid of color c_i at time t_n, denote the amount of fluid of color c_i at time t_{n+1} through $\boldsymbol{R}_i^{n+1}(x_i)$ (offspring of type i) if $i \neq 0$, and $\boldsymbol{G}^{n+1}$ (immigration) if $i = 0$, so that

$$\boldsymbol{B}^{n+1} = \sum_{i=1}^{N} \boldsymbol{R}_i^{n+1}(B_i^n) + \boldsymbol{G}^{n+1}, \quad n \geq 0. \tag{11}$$

Obviously the sequences $\{\boldsymbol{R}_i^n(x), x \geq 0, n \geq 1\}$ and $\{\boldsymbol{G}^n, n \geq 1\}$ constitute sequences of i.i.d. increasing stochastic processes and random vectors, respectively. Moreover, $\boldsymbol{G}^n$ is independent from $\boldsymbol{R}_i^n$ for $i = 1, \ldots, N$, because the amount of fluid arriving during switch-over periods depends only on the lengths S_i (independent from anything else) of those periods and the input process $\boldsymbol{W}$, but not on the amount of fluid already in the system. Denote the common distribution of $\{\boldsymbol{R}_i^n(x), x \geq 0, n \geq 1\}$ and $\{\boldsymbol{G}^n, n \geq 1\}$ by $\boldsymbol{R}_i$ and $\boldsymbol{G}$, respectively.

Note that at the time instant when the server starts polling Q_{i+1}, the amount of fluid of color c_i is given by $\boldsymbol{H}_i(x_i)$, so

$$\boldsymbol{R}_i(x_i) =_{\mathrm{d}} \sum_{j=1}^{i} \boldsymbol{e}_j H_{i,j}(x_i) + \sum_{j=i+1}^{N} \boldsymbol{R}_j(H_{i,j}(x_i)), \tag{12}$$

where $H_{i,j}(x)$ denotes the jth element of $\boldsymbol{H}_i(x)$. Note that the color c_i fluid can appear in the system only as a replacement of the fluid already present in Q_i at the beginning of the polling cycle (which corresponds to the part $\sum_{j=1}^{i} \boldsymbol{e}_j H_{i,j}(x_i)$), or as a replacement of the fluid that arrived to the, yet to be served, queues $Q_{i+1}, \ldots, Q_N$, during the service of Q_i (which corresponds to the part $\sum_{j=i+1}^{N} \boldsymbol{R}_j(H_{i,j}(x_i))$).

Backward induction (from $i = N$ to $i = 1$) and stationarity and independence of increments of $\boldsymbol{H}_i$ imply the same properties for $\boldsymbol{R}_i$. Hence, $\boldsymbol{R}_i$ are Lévy subordinators, for $i = 1, \ldots, N$. Finally, the mutual independence of $\boldsymbol{R}_i^n$ for $i = 1, \ldots, N$ follows from the Property 1. Hence $\{\boldsymbol{B}^n, n \geq 0\}$ is a MTJBP.

Next, we compute the Laplace exponent κ_i of $\boldsymbol{R}_i$. Using (12) and conditioning on $\boldsymbol{H}_i(x)$ we obtain

$$\begin{aligned} \mathbb{E} \exp\left(-\boldsymbol{u} \cdot \boldsymbol{R}_i(x)\right) &= \mathbb{E} \exp\left(-\sum_{j=1}^{i} u_j H_{i,j}(x) - \sum_{j=i+1}^{N} H_{i,j}(x) \kappa_j(\boldsymbol{u})\right) \\ &= \exp\left(-x \eta_i(u_1, \ldots, u_i, \kappa_{i+1}(\boldsymbol{u}), \ldots, \kappa_N(\boldsymbol{u}))\right), \end{aligned}$$

so that (9) holds.

It is left to compute the LST of $\boldsymbol{G}$. First note that we can write $\boldsymbol{G} = \sum_i \boldsymbol{G}_i$, where $\boldsymbol{G}_i$ are mutually independent and represent fluid at the end of the polling cycle generated by the ith switch-over. That is,

$$\boldsymbol{G}_i =_{\mathrm{d}} \sum_{j=1}^{i} \boldsymbol{e}_j W_j(S_i) + \sum_{j=i+1}^{N} \boldsymbol{R}_j(W_j(S_i)).$$

Similarly as above we obtain

$$\begin{aligned}\mathbb{E}\exp(-\boldsymbol{u}\cdot\boldsymbol{G}_i) &= \mathbb{E}\exp\left(-\sum_{j=1}^{i} u_j W_j(S_i) - \sum_{j=i+1}^{N} W_j(S_i)\kappa_j(\boldsymbol{u})\right)\\ &= \mathbb{E}\exp\left(-S_i\phi(u_1,\dots,u_i,\kappa_{i+1}(\boldsymbol{u}),\dots,\kappa_N(\boldsymbol{u}))\right),\end{aligned}$$

which proves (10). □

Let $\boldsymbol{B}_i$ and $\boldsymbol{E}_i$ denote the random variable having the steady-state distribution of the joint amount of fluid in each queue at the beginning of a visit (polling instant) to Q_i and at the end of a visit (switching instant) to Q_i, respectively.

COROLLARY 1. *$\boldsymbol{B}_i$ and $\boldsymbol{E}_i$ can be related to each other by*

$$\tilde{\boldsymbol{B}}_{i+1}(\boldsymbol{u}) = \tilde{\boldsymbol{E}}_i(\boldsymbol{u})\tilde{S}_i(\phi(\boldsymbol{u})) \tag{13}$$

and

$$\tilde{\boldsymbol{E}}_i(\boldsymbol{u}) = \tilde{\boldsymbol{B}}_i(u_1,\dots,u_{i-1},\eta_i(\boldsymbol{u}),u_{i+1},\dots,u_N),$$

where

$$\tilde{\boldsymbol{B}}_1(\boldsymbol{u}) = \prod_{k=0}^{\infty} \tilde{\boldsymbol{G}}(\boldsymbol{\kappa}^{(k)}(\boldsymbol{u})),$$

with $\tilde{\boldsymbol{G}}$ and $\boldsymbol{\kappa}$ given by Theorem 2.

REMARK 4. By the same token, we can, for arbitrary $i = 1,\dots,N$, find $\tilde{\boldsymbol{B}}_i(\boldsymbol{u})$ as well (renumber such that Q_i becomes Q_1). The infinite product formula for the Laplace transform of the distribution of $\boldsymbol{B}_1$ is typical in the area of (classical) polling models. This kind of formula can be numerically inverted to obtain various performance metrics, see Abate and Whitt [1, 2] and Choudhury and Whitt [11]. Such an infinite product already arises in the case of a M/G/1 queue ($N = 1$) with gated vacations. For example [28, Section 2.5, Formula (5.19)] gives the following relation for $Q(z)$, the generating function of the queue length distribution at the end of a vacation:

$$Q(z) = Q(B^*(\lambda(1-z))V^*(\lambda(1-z)),$$

where λ denotes arrival rate and $B^*(\cdot)$, $V^*(\cdot)$ are the LST of service time and vacation length, respectively. This immediately results in

$$Q(z) = \prod_{i=0}^{\infty} V^*(\lambda(1-\delta^{(i)}(z))),$$

where $\delta^{(i)}(z) = B^*(\lambda(1-\delta^{(i-1)}(z)))$, $i = 1, 2, \ldots$, $\delta^{(0)}(z) = z$. In the special case of a single queue ($N = 1$) and exhaustive service, the infinite product degenerates to $V^*(\lambda(1-z))$ because at the end of each visit, the system has become empty.

For arbitrary N and classical Poisson input processes, Resing [23] presents the joint queue length generating function at epochs the server begins a visit to Q_1. This generating function is also given in the form of an infinite product, cf. [23, Theorem 1 and Theorem 3].

REMARK 5. The interpretation of the above infinite-product expression for $\tilde{\boldsymbol{B}}_1(\boldsymbol{u})$ is the following. The terms of the infinite product correspond to independent contributions to the workload vector at a polling instant of Q_1. The 0th term represents work still present, that has arrived during the switch-over periods in the cycle that has just ended. The 1st term represents work that has arrived during the service of work that had arrived one cycle earlier during switch-over periods. And so on: the kth term, $k = 1, 2, \ldots$, represents work that was initiated k cycles before the last cycle by ancestral work arriving during a switch-over period.

REMARK 6. A special case of our model is the fluid polling model studied in Czerniak and Yechiali [12]: there the Lévy input reduces to N linear deterministic processes. Another special case of our model is the classical polling model in which the Lévy input corresponds to N independent CPPs.

REMARK 7. From the proof of Theorem 2 we can clearly see the importance of the branching property. Most importantly, it implied the distributional equality (12) for the distribution of $\boldsymbol{R}_i$. Then the distribution of $\boldsymbol{G}$ follows in terms of the Laplace exponents of the $\boldsymbol{R}_i$'s. However, to establish the MTJBP structure of the Lévy-driven polling system, all we need to show is that (11) (or equivalently (5)) holds with independent components.

One can think of disciplines that do not satisfy Property 1, but for which still the workload evolution can be described by (11). An example is the *globally-gated* discipline (see [8]), which works as follows. At the beginning of each cycle, all fluid in $Q_1, \dots, Q_N$ is marked. During the next cycle, the server serves all the marked fluid. The newly arrived fluid, however, has to wait until being marked at the next cycle-beginning, and will be served during the next cycle. This discipline does not satisfy Property 1, but assuming that the server works with rate 1, an equation of the form (11) can be derived with $\boldsymbol{R}_i^n =_{\mathrm{d}} \boldsymbol{W}$ and $\boldsymbol{G} = \sum \boldsymbol{G}_i$, where $\boldsymbol{G}_i =_{\mathrm{d}} \boldsymbol{W}(S_i)$. Therefore such a model also has the MTJBP structure with branching mechanism $\boldsymbol{\kappa}(\boldsymbol{u}) = (\phi(\boldsymbol{u}), \dots, \phi(\boldsymbol{u}))$ and immigration LST $\tilde{\boldsymbol{G}}(\boldsymbol{u}) = \prod \tilde{S}_i(\phi(\boldsymbol{u}))$.

As in the introduction, it should be noted that a relation between Lévy-driven polling systems and continuous-state space branching processes was considered before by Altman and Fiems [4]. In that work, the assumption is imposed that all queues are fed by *identical* Lévy subordinators, and it is precisely this assumption that enables the construction of a continuous state-space branching process. Moreover, the relation in [4] is used only to derive the first two waiting-time moments and not to determine the structure of the branching process itself.

5. Steady-state distribution at an arbitrary epoch. Having determined the LST of the joint steady-state workload at polling and switching instants in the previous section, we now concentrate on its counterpart *at an arbitrary instant in time*. It should be noted that this distribution was not even found before for the classical polling models with independent CPP inputs, except for the marginal distributions. In order to do so, we need to make the notion of service disciplines more precise.

Firstly, recall that work in the system (while the server is at Q_i) evolves according to a Lévy process

$$\boldsymbol{A}_i(t) := \Big(W_1(t), \dots, W_{i-1}(t), A_i(t), W_{i+1}(t), \dots, W_N(t)\Big),$$

where $A_i(t)$ can be any spectrally positive Lévy process with negative drift. The Laplace exponent of $\boldsymbol{A}_i(t)$ is denoted through $\phi_i^{\boldsymbol{A}}(\boldsymbol{u})$. Let $\mathcal{F}_i$ be an augmented right-continuous filtration, such that $\boldsymbol{A}_i(t)$ is a Lévy process with respect to $\mathcal{F}_i$ (one can take an augmented natural filtration of $\boldsymbol{A}_i(t), t \geq 0$). Let $\tau_i(x)$ denote the time the server spends at Q_i when it finds x units of work in this queue upon arrival. Loosely speaking, $\tau_i(x)$ is a stopping rule for the server, which observes the process $\boldsymbol{A}_i(t)$. This motivates the following assumption.

Assumption 1. *$\tau_i(x)$ is an $\mathcal{F}_i$-stopping time for every x.*

Assumption 1 ensures that the disciplines are 'non-anticipating'. For example, we exclude the cases, when the server decides to stop the service if it 'sees' that, for instance, the cumulative input in the next T units of time is less than some ε. The above assumption expresses the fact, that a service strategy or discipline should be based only on the knowledge of the evolution of the system up to the current time-point. Observe that for the gated discipline $\tau_i(x) = x$ and for exhaustive $\tau_i(x) = T_i(x) = \inf\{t \geq 0 : A_i(t) = -x\}$ (recalling the definition of $T_i(x)$ from Section 2). It is easily verified that both are $\mathcal{F}_i$-stopping times, as desired.

We also require that a discipline is 'work conserving', that is, the server never stays at a queue after it became empty. This is made precise in the following assumption.

ASSUMPTION 2. *For every x it holds that $\tau_i(x) \leq T_i(x)$ a.s.*

Note that the gated discipline ($A_i(t) = W_i(t) - t$) and the exhaustive discipline are always work conserving. Because of Assumption 2 one does not need to consider reflection of the workload process at 0, hence the workload replacement $\boldsymbol{H}_i(x)$ is given by

$$\boldsymbol{H}_i(x) = x\boldsymbol{e}_i + \boldsymbol{A}_i(\tau_i(x)).$$

Moreover, $\mathbb{E}\tau_i(x) \leq \mathbb{E}T_i(x) < \infty$, because $A_i(t)$ has negative drift. Hence using Wald's identity twice to Lévy processes $\boldsymbol{H}_i$ and $\boldsymbol{A}_i$ we get

$$\mathbb{E}\tau_i(x) = x\mathbb{E}\tau_i(1). \tag{14}$$

The following result presents a Kella-Whitt type martingale, which is a key tool in deriving the workload LST at an arbitrary time.

PROPOSITION 1.

$$M_i(t) := e^{-\boldsymbol{u}\cdot\boldsymbol{A}_i(t)} - 1 + \phi_i^{\boldsymbol{A}}(\boldsymbol{u}) \int_0^t e^{-\boldsymbol{u}\cdot\boldsymbol{A}_i(s)}\, ds, \quad t \geq 0,$$

is a zero mean martingale with respect to filtration $\mathcal{F}_i$.

PROOF. Apply Kella and Whitt [17, Theorem 1] to the one-dimensional Lévy process $\boldsymbol{u} \cdot \boldsymbol{A}_i$ (with respect to filtration $\mathcal{F}_i$). □

Applying Doob's Optional Sampling theorem to the martingale M_i from Proposition 1 and stopping time $\tau_i(x) \wedge n$ yields

$$\mathbb{E} \int_0^{\tau_i(x)\wedge n} e^{-\boldsymbol{u}\cdot\boldsymbol{A}_i(s)}\, ds = \frac{1}{\phi_i^{\boldsymbol{A}}(\boldsymbol{u})} \mathbb{E}\left(1 - e^{-\boldsymbol{u}\cdot\boldsymbol{A}_i(\tau_i(x)\wedge n)}\right).$$

Taking $n \to \infty$ and applying the monotone convergence theorem on the left and the dominated convergence theorem on the right ($e^{-\boldsymbol{u}\cdot\boldsymbol{A}_i(\tau_i(x)\wedge n)} \leq e^{u_i x}$) we obtain

$$\mathbb{E}\int_0^{\tau_i(x)} e^{-\boldsymbol{u}\cdot\boldsymbol{A}_i(s)}\,ds = \frac{1}{\phi_i^{\boldsymbol{A}}(\boldsymbol{u})}\mathbb{E}\left(1 - e^{-\boldsymbol{u}\cdot\boldsymbol{A}_i(\tau_i(x))}\right). \tag{15}$$

We are now ready to state the second main result of this paper.

THEOREM 3. *Consider the polling system described in Section 2 and suppose that the disciplines at every queue satisfy Assumption 1, Assumption 2 and Property 1. Then the LST of the steady-state distribution of the joint amount of fluid* $\boldsymbol{F}$ *at an arbitrary epoch is given by*

$$\mathbb{E}e^{-\boldsymbol{u}\cdot\boldsymbol{F}} = \frac{N(\boldsymbol{u})}{\sum_{i=1}^N \mathbb{E}S_i + \sum_{i=1}^N \mathbb{E}\tau_i(1)\mathbb{E}B_{i,i}}, \tag{16}$$

where

$$N(\boldsymbol{u}) = \sum_{i=1}^N \left(\frac{\tilde{\boldsymbol{B}}_i(\boldsymbol{u}) - \tilde{\boldsymbol{E}}_i(\boldsymbol{u})}{\phi_i^{\boldsymbol{A}}(\boldsymbol{u})} + \frac{\tilde{\boldsymbol{E}}_i(\boldsymbol{u}) - \tilde{\boldsymbol{B}}_{i+1}(\boldsymbol{u})}{\phi(\boldsymbol{u})}\right) \tag{17}$$

and $\tilde{\boldsymbol{B}}_i$, $\tilde{\boldsymbol{E}}_i$ *are as given in Corollary 1.*

PROOF. Let $\boldsymbol{F}(t)$ be the amount of fluid in each queue at time t within a cycle C, assuming that we start in stationarity. The LST of $\boldsymbol{F}$ is calculated by dividing the expected area of the function $e^{-\boldsymbol{u}\cdot\boldsymbol{F}(t)}$ over the cycle C, by the expected cycle time $\mathbb{E}C$:

$$\tilde{\boldsymbol{F}}(u) = \frac{\mathbb{E}\int_0^C e^{-\boldsymbol{u}\cdot\boldsymbol{F}(t)}\,dt}{\mathbb{E}C}.$$

Note that

$$C =_{\mathrm{d}} \sum_{i=1}^N (S_i + \tau_i(B_{i,i})),$$

from which, in combination with (14), the denominator in (16) follows.

An arbitrary cycle of length C can be divided into intervals corresponding to visit periods V_i to Q_i and switching (idle) periods I_i between Q_i and Q_{i+1}.

Conditioning on S_i and using Fubini's theorem, we find

$$\begin{aligned}\mathcal{S}_i(\boldsymbol{u}) &:= \mathbb{E}\int_{I_i} e^{-\boldsymbol{u}\cdot\boldsymbol{F}(t)}\,dt = \mathbb{E}\int_0^{S_i} e^{-\boldsymbol{u}\cdot(\boldsymbol{E}_i+\boldsymbol{W}(t))}\,dt = \tilde{\boldsymbol{E}}_i(\boldsymbol{u})\frac{1-\tilde{S}_i(\phi(\boldsymbol{u}))}{\phi(\boldsymbol{u})}\\ &= \frac{\tilde{\boldsymbol{E}}_i(\boldsymbol{u})-\tilde{\boldsymbol{B}}_{i+1}(\boldsymbol{u})}{\phi(\boldsymbol{u})}.\end{aligned}$$

On the intervals V_i we have

$$\mathcal{L}_i(\boldsymbol{u}) := \mathbb{E}\int_{V_i} e^{-\boldsymbol{u}\cdot\boldsymbol{F}(t)}\,dt = \mathbb{E}\int_0^{\tau_i(B_{i,i})} e^{-\boldsymbol{u}\cdot(\boldsymbol{B}_i+\boldsymbol{A}_i(t))}\,dt.$$

Conditioning on $\boldsymbol{B}_i$ and applying (15) yields

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{u}) &= \frac{1}{\phi_i^{\boldsymbol{A}}(\boldsymbol{u})}\left(\mathbb{E}e^{-\boldsymbol{u}\cdot\boldsymbol{B}_i} - \mathbb{E}e^{-\boldsymbol{u}\cdot(\boldsymbol{B}_i+\boldsymbol{A}_i(\tau_i(B_{i,i})))}\right)\\ &= \frac{1}{\phi_i^{\boldsymbol{A}}(\boldsymbol{u})}\left(\mathbb{E}e^{-\boldsymbol{u}\cdot\boldsymbol{B}_i} - \mathbb{E}e^{-\boldsymbol{u}\cdot\boldsymbol{E}_i}\right) = \frac{\tilde{\boldsymbol{B}}_i(\boldsymbol{u})-\tilde{\boldsymbol{E}}_i(\boldsymbol{u})}{\phi_i^{\boldsymbol{A}}(\boldsymbol{u})}.\end{aligned}$$

Finally,

$$\mathbb{E}\int_0^C e^{-\boldsymbol{u}\cdot\boldsymbol{F}(t)}\,dt = \sum_{i=1}^N\left(\mathcal{L}_i(\boldsymbol{u})+\mathcal{S}_i(\boldsymbol{u})\right).$$

□

Note that all the quantities appearing in the statement of Theorem 3 are computable, in the sense that they follow from the results presented in Section 4: the transforms $\tilde{\boldsymbol{B}}_i(\boldsymbol{u})$ and $\tilde{\boldsymbol{E}}_i(u)$ are given in Corollary 1 and $\mathbb{E}B_{i,i} = -\partial\tilde{\boldsymbol{B}}_i/\partial u_i(\boldsymbol{0})$.

6. Varying input processes. In Section 2 and Section 4 we considered a polling model, with fixed input $\boldsymbol{W}$ characterized by its Laplace exponent ϕ. However, to derive our results, all we needed was the knowledge of ϕ and $\phi_i^{\boldsymbol{A}}$ for $i = 1,\dots,N$ as remarked at the end of Section 2 and not the fact that $\phi_i^{\boldsymbol{A}}$ are related to each other or that ϕ is fixed. That is why we can allow our input to change between embedded epochs.

More precisely, let $\boldsymbol{W}_i$ and $\hat{\boldsymbol{W}}_i$ be sequences of N-dimensional subordinators for $i = 1,\dots,N$. When the server arrives at Q_i then the input process changes to $\boldsymbol{W}_i$, and when the server leaves Q_i the input process switches to $\hat{\boldsymbol{W}}_i$. Let us denote the Laplace exponents of these processes by ϕ_i and $\hat{\phi}_i$, respectively. The process $\boldsymbol{A}_i(t)$ becomes

$$\boldsymbol{A}_i(t) := \Big(W_{i,1}(t),\dots,W_{i,i-1}(t),A_i(t),W_{i,i+1}(t),\dots,W_{i,N}(t)\Big),$$

where $A_i(t)$ is an arbitrary spectrally positive Lévy process with negative drift modelling the server's work at Q_i. Again, denote the Laplace exponent of $\boldsymbol{A}_i$ by $\phi_i^{\boldsymbol{A}}$.

We still consider disciplines satisfying Property 1, so that for the gated discipline (1) changes to

$$\eta_i(\boldsymbol{u}) = \phi_i(\boldsymbol{u}),$$

and for the exhaustive discipline (2) changes to

$$\eta_i(\boldsymbol{u}) = \psi_i(\boldsymbol{u}),$$

with ψ_i such that $\phi_{i,\boldsymbol{u}}(\psi_i(\boldsymbol{u})) = 0$ for $\phi_{i,u}(\theta) = \phi_i^{\boldsymbol{A}}(u_1, \ldots, u_{i-1}, \theta, u_{i+1}, \ldots, u_N)$, where $\theta \geq 0$. The resulting process $\{\boldsymbol{B}^n, n \geq 1\}$ of the joint amount of the fluid in the different queues at time points t_n also constitutes an MTJBP with branching mechanism $\boldsymbol{\kappa}$ given by (9) and immigration LST given by

$$\tilde{\boldsymbol{G}}(\boldsymbol{u}) = \prod_{i=1}^{N} \tilde{S}_i(\hat{\phi}_i(u_1, \ldots, u_i, \kappa_{i+1}(\boldsymbol{u}), \ldots, \kappa_N(\boldsymbol{u}))) \tag{18}$$

instead of (10). The argument given in the proof of Theorem 2 stays valid. Then (13) in Corollary 1 changes into

$$\tilde{\boldsymbol{B}}_{i+1}(\boldsymbol{u}) = \tilde{\boldsymbol{E}}_i(\boldsymbol{u}) \tilde{S}_i(\hat{\phi}_i(\boldsymbol{u})). \tag{19}$$

The statement of Theorem 3 also still holds with

$$N(\boldsymbol{u}) = \sum_{i=1}^{N} \left(\frac{\tilde{\boldsymbol{B}}_i(\boldsymbol{u}) - \tilde{\boldsymbol{E}}_i(\boldsymbol{u})}{\phi_i^{\boldsymbol{A}}(\boldsymbol{u})} + \frac{\tilde{\boldsymbol{E}}_i(\boldsymbol{u}) - \tilde{\boldsymbol{B}}_{i+1}(\boldsymbol{u})}{\hat{\phi}_i(\boldsymbol{u})} \right).$$

REMARK 8. For independent compound Poisson input processes, the case of varying input has been studied in [7]. There it is assumed that the arrival process at Q_i, when the server is at Q_j, is a Poisson process with rate λ_{ij}. Under the assumption of branching-type service disciplines, the joint queue-length distribution at polling instants is derived in [7].

7. Ergodicity. Consider the polling model with varying input, as was presented in Section 6. The stability condition of such a system is given in terms of the Perron-Frobenius eigenvalue ρ_M of the mean matrix M associated to a certain branching process, see Theorem 1. This branching process in turn is specified in Theorem 2 (or, more precisely, in its generalization to varying input, as presented in Section 6) in terms of η_i among other quantities. This leads to a rather non-transparent stability criterion, as opposed to the '$\rho < 1$' type of criteria one usually encounters in queueing theory. In addition, it is not clear how the criterion depends on the disciplines used at different queues. The goal of this section is to make the stability condition more explicit, and to show that it is, under quite general circumstances, *not* affected by the service discipline.

In this section we assume that the disciplines at all queues satisfy Assumption 1 and Assumption 2. Moreover, we exclude the degenerate case when $\tau_i(x) = 0$ (never serving Q_i). In this setting the stability condition can be greatly simplified. Importantly, we show that it can be expressed in terms of properties of the *rate matrix* $A = (a_{ij})$ rather than properties of the *mean matrix* M; here $a_{ij} = \mathbb{E}A_{i,j}(1)$, that is, a_{ij} is the rate of the work evolution at Q_j while the server is at Q_i. Note that A has non-negative off-diagonal elements, hence by Perron-Frobenius theory it has a real eigenvalue ρ_A which is larger than the real part of any other eigenvalue of A.

LEMMA 1. *Let A be irreducible. Then it holds that*

- *if $\rho_A < 0$ then $\rho_M < 1$ (subcritical);*
- *if $\rho_A = 0$ then $\rho_M = 1$ (critical);*
- *if $\rho_A > 0$ then $\rho_M > 1$ (supercritical).*

In the supercritical case there exists a positive vector $\boldsymbol{w}$ such that $M\boldsymbol{w} > \boldsymbol{w}$.

PROOF. Let us consider the polling model from Section 6 and denote the MTJBP associated to it by $(\hat{\boldsymbol{B}}^n)$. This MTJBP has a corresponding branching mechanism $\boldsymbol{\kappa}$ and immigration $\boldsymbol{G}$. Let us construct a new MTJBP $(\boldsymbol{B}^n)$ with the same branching mechanism $\boldsymbol{\kappa}$ but without immigration, i.e., with $\boldsymbol{G}$ set to 0. Such an MTJBP obviously corresponds to a polling model with the same characteristics as the starting model, but with switch-over times set to 0. Moreover the mean matrix $\hat{M}$ is the same as the mean matrix M.

From the definition of M its ith row is given by $\boldsymbol{m}_i = \mathbb{E}(\boldsymbol{B}^1|\boldsymbol{B}^0 = \boldsymbol{e}_i)$. In the following we assume without loss of generality, that $\boldsymbol{B}^0 = \boldsymbol{e}_i$. We can write

$$\boldsymbol{B}^1 = \boldsymbol{e}_i + \sum_{k=i}^{N} \boldsymbol{A}_k\left(\tau_k(F_k)\right),$$

where F_k denotes the fluid in Q_k upon the server's arrival to this queue (in the first cycle). Using Wald's identity and the linearity of $\mathbb{E}\tau_k(x)$, as given in (14), we obtain

$$\boldsymbol{m}_i = \mathbb{E}\boldsymbol{B}^1 = \boldsymbol{e}_i + \sum_{k=i}^{N} \mathbb{E}\tau_k(1)\mathbb{E}F_k\boldsymbol{a}_k,$$

where $\boldsymbol{a}_k$ is the kth row of A. Let $\boldsymbol{w} > \boldsymbol{0}$ be an eigenvector of A with positive elements associated to ρ_A, which exists by the Perron-Frobenius theory. Then

$$\boldsymbol{m}_i \cdot \boldsymbol{w} = w_i + \rho_A \sum_{k=i}^{N} \mathbb{E}\tau_k(1)\mathbb{E}F_k w_k.$$

Note that $\mathbb{E}\tau_k(1) > 0$ and $\mathbb{E}F_k \geq 0$ for all k, and $\mathbb{E}F_i = 1$, because $\boldsymbol{X}^0 = \boldsymbol{e}_i$ and $S_k = 0$ for all k. Therefore, according to $\rho_A < 0, \rho_A = 0$, and $\rho_A > 0$ we obtain $M\boldsymbol{w} < \boldsymbol{w}, M\boldsymbol{w} = \boldsymbol{w}$, and $M\boldsymbol{w} > \boldsymbol{w}$, respectively. Now the claim follows from Lemma 2 in the Appendix. □

In the following we assume that the total work arriving during the switch-overs in one polling cycle is not identically 0, that is we can not erase the switch-over periods without changing the system. The next corollary is an immediate consequence of Lemma 1 and Theorem 1, in conjunction with the comments following the proof of Theorem 1.

COROLLARY 2. *Let A be irreducible. Then it holds that*

- *if $\rho_A < 0$, then the polling system is stable;*
- *if $\rho_A > 0$, then the polling system is unstable.*

Note that the stability of our polling system depends only on the input and does not depend on the disciplines used at different queues. Clearly, this result strongly relies on the fact that the disciplines are work conserving, see Assumption 2, and satisfy Property 1.

Finally we make a comment on a simplified model from Section 2, where the input does not depend on the location of the server, and in which the server works at unit speed. That is $A_i = W_i(t) - t$ for a fixed $\boldsymbol{W}$. Denote the mean rate of the input into Q_i by $\rho_i > 0$ and the mean total rate by $\rho = \sum_{i=1}^N \rho_i$. This means that $a_{ij} = \rho_j - \delta_{ij}$, so that A is irreducible and $A\mathbf{1} = (\rho - 1)\mathbf{1}$. Apply Lemma 2 in the Appendix and Corollary 2 to see that we obtain the expected stability condition: the system is stable if $\rho < 1$ and is unstable if $\rho > 1$.

8. Discussion and concluding remarks. In this paper we analyzed a general class of Lévy driven polling models. Exploiting the relation with multi-type Jiřina processes, we determined the LST of the joint stationary workload distribution. The collection of results presented in this paper is rich, as they cover various known results as special cases, but also constitute a broad range of new results, even for the class of classical polling models with compound Poisson input.

Various extensions and ramifications can be thought of. For instance, where the present paper considers cyclic polling (i.e., in every cycle all N queues are served in a cyclic manner), any system with a fixed polling table can be dealt with similarly. The class of models in which the polling order is *random*, however, does *not* fit in the class of MTJBPs, as was already observed for MTBP in [23], and can therefore not be analyzed by our methods. Another direction for future research relates to the use of the LST of the workload distribution to obtain insight into the corresponding tail probabilities, as was done for a specific polling model in [9].

It would also be interesting to investigate whether a work decomposition property for the polling model with/without switch-over times exists, cf. Boxma [6]. Yet another direction for future research is to derive the joint steady-state queue length distribution *at an arbitrary epoch* for the classical polling model with CPP input, using a martingale technique as we have employed for the workload in Section 5.

We envisage that a variant of our approach can deal with an even larger class of service disciplines than the class of branching disciplines. As was noted in Remark 7 of Section 4, the so-called *globally-gated* service discipline does not qualify as branching type, but we showed that it is nevertheless possible to find a corresponding MTJBP. The ideas presented in Remark 7 indicate under what circumstances still MTJBPs can be constructed.

APPENDIX

The following Lemma is a variation of [25, Theorem 1.6].

LEMMA 2. *Let M be a square matrix with non-negative off-diagonal elements. Let ρ_M be its Perron-Frobenius eigenvalue (the eigenvalue with maximal real part) and $\boldsymbol{w}$ be a positive vector. Then*

$$M\boldsymbol{w} < \boldsymbol{w} \textit{ implies } \rho_M < 1,$$
$$M\boldsymbol{w} = \boldsymbol{w} \textit{ implies } \rho_M = 1,$$
$$M\boldsymbol{w} > \boldsymbol{w} \textit{ implies } \rho_M > 1.$$

PROOF. Let $\boldsymbol{v} \geq \boldsymbol{0}$ be a left eigenvector (i.e., a row vector) of M corresponding to ρ_M. If $M\boldsymbol{w} < \boldsymbol{w}$, then $\rho_M(\boldsymbol{v} \cdot \boldsymbol{w}) = \boldsymbol{v}M\boldsymbol{w} < \boldsymbol{v} \cdot \boldsymbol{w}$. Hence $\rho_M < 1$. The other two statements follow similarly. ◻

Acknowledgments. The authors thank E. Altman (INRIA, France), S. Meyn (University of Illinois at Urbana-Champaign, United States), and J.A.C. Resing (Eindhoven University of Technology, the Netherlands) for useful discussions.

REFERENCES

[1] J. Abate and W. Whitt. Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comp.*, 7:36–43, 1995.

[2] J. Abate and W. Whitt. A unified framework for numerically inverting Laplace transforms. *INFORMS J. Comput.*, 18:408–421, 2006. MR2274708

[3] E. Altman. Stochastic recursive equations with applications to queues with dependent vacations. *Annals of Operations Research*, 112:43–61, 2002. MR1960670

[4] E. Altman and D. Fiems. Expected waiting time in symmetric polling systems with correlated walking times. *Queueing Syst.*, 56:241–253, 2007. MR2336110

[5] S. Asmussen. *Applied Probability and Queues*. Springer, 2nd edition, 2002. MR1978607

[6] O.J. Boxma. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Syst.*, 5:185–214, 1989. MR1032554

[7] O.J. Boxma. Polling systems. In K. Apt, A. Schrijver, and N. Temme, editors, *Liber Amicorum for P.C. Baayen.*, pages 215–230. CWI, Amsterdam, 1994. MR1490592

[8] O.J. Boxma, H. Levy, and U. Yechiali. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Ann. Oper. Res.*, 35:187–208, 1992.

[9] O.J. Boxma, Q. Deng, and J.A.C. Resing. Polling systems with regularly varying service and/or switchover times. *Adv. Perf. Anal.*, 3:71–107, 2000.

[10] O.J. Boxma, M. Mandjes, and O. Kella. On a queueing model with service interruptions. *Probab. Engrg. Infor. Sci.*, 22:537–555, 2008. MR2452336

[11] G.L. Choudhury and W. Whitt. Computing distributions and moments in polling models by numerical transform inversion. *Performance Evaluation*, 25:267–292, 1996.

[12] O. Czerniak and U. Yechiali. Fluid polling systems. *Queueing Syst.*, 63:401–435, 2009. MR2576020

[13] I. Eliazar. Gated polling systems with Lévy inflow and inter-dependent switchover times: a dynamical systems approach. *Queueing Syst.*, 49:49–72, 2005. MR2119657

[14] S.W. Fuhrmann. Performance analysis of a class of cyclic schedules. Technical Report 81-59531-1., Bell Laboratories, 1981.

[15] S.W. Fuhrmann. A decomposition result for a class of polling models. *Queueing Syst.*, 11:109–120, 1992. MR1178645

[16] M. Jiřina. Stochastic branching processes with continuous state space. *Czechosl. Math. J.*, 8:292–313, 1958. MR0101554

[17] O. Kella and W. Whitt. Useful martingales for stochastic storage processes with Lévy input. *J. Appl. Probab.*, 29:396–403, 1992. MR1165224

[18] A.E. Kyprianou. *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer-Verlag, Berlin Heidelberg, 2006. MR2250061

[19] H. Levy and M. Sidi. Polling systems with correlated arrivals. In *Technology: Emerging or Converging, IEEE*, volume 3 of *Proceedings of the Eighth Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 907–913. INFOCOM '89, 1989.

[20] H. Levy and M. Sidi. Polling models: applications, modeling and optimization. *IEEE Trans. Commun.*, 38:1750–1760, 1990.

[21] A.G. Pakes. Some limit theorems for Jiřina processes. *Period. Math. Hungar.*, 1:55–66, 1979. MR0494544

[22] N.U. Prabhu. *Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication*. Springer-Verlag, New York, 2nd edition, 1998. MR1492990

[23] J.A.C. Resing. Polling systems and multitype branching processes. *Queueing Syst.*, 13:409–426, 1993. MR1231052

[24] K. Sato. *Lévy Processes and Infinitely Divisible Distributions*, volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1999. MR1739520

[25] E. Seneta. *Non-negative Matrices and Markov Chains*. Springer-Verlag, New York, 2nd edition, 1981. MR2209438

[26] E. Seneta and D. Vere-Jones. On the asymptotic behaviour of subcritical branching processes with continuous state space. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 10: 212–225, 1968. MR0239667

[27] E. Seneta and D. Vere-Jones. On a problem of M. Jiřina concerning continuous state branching processes. *Czechosl. Math. J.*, 19:277–283, 1969. MR0246379

[28] H. Takagi. *Queueing Analysis*, volume 1. North-Holland, Amsterdam, 1991. MR1149382

[29] H. Takagi. Application of polling models to computer networks. *Comput. Netw. ISDN Syst.*, 22:193–211, 1991.

[30] H. Takagi. Analysis and application of polling models. In G. Haring, C. Lindemann, and M. Reiser, editors, *Performance Evaluation: Origins and Directions*, volume 1769 of *Lecture Notes in Computer Science*, pages 423–442. Springer, Berlin, 2000.

[31] V.M. Vishnevskii and O.V. Semenove. Mathematical methods to study the polling systems. *Autom. Remote Control*, 67:173–220, 2006. MR2210457

[32] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, 2002. MR1876437

Eindhoven University of Technology
Department of Mathematics
and Computer Science
PO Box 513
5600 MB Eindhoven
The Netherlands
E-mail: boxma@win.tue.nl

Korteweg-de Vries Institute
for Mathematics
University of Amsterdam
PO Box 94248
1090 GE Amsterdam
The Netherlands
E-mail: ivanovs@eurandom.tue.nl

EURANDOM
PO Box 513
5600 MB Eindhoven
The Netherlands
E-mail: kosinski@eurandom.tue.nl

CWI
PO Box 94079
1090 GB Amsterdam
The Netherlands
E-mail: m.r.h.mandjes@uva.nl