Operations Research Letters 45 (2017) 40–45

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Lowest priority waiting time distribution in an accumulating priority Lévy queue

Offer Kella [a], Liron Ravner [b,*]

[a] Department of Statistics, The Hebrew University of Jerusalem, Israel

[b] Department of Statistics and Operations Research, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Received 3 July 2016
Received in revised form
6 November 2016
Accepted 23 November 2016
Available online 1 December 2016

Keywords:
Priority queues
Accumulating priority
Lévy driven queues
Lévy processes
First passage time

ABSTRACT

We derive the waiting time distribution of the lowest class in an accumulating priority (AP) queue with positive Lévy input. The priority of an infinitesimal customer (particle) is a function of their class and waiting time in the system, and the particles with the highest AP are the next to be processed. To this end we introduce a new method that relies on the construction of a workload overtaking process and solving a first-passage problem using an appropriate stopping time.

1. Introduction

Suppose that a single server provides service to non-atomic (infinitesimal) customers, that we refer to as particles, of different types according to a dynamic non-preemptive accumulating priority (AP) service regime. That is, the priority of every particle in the queue is a function of their class and their accumulated waiting time in the queue. Upon a service completion the server admits the particle with the highest accumulated priority. Such regimes are common in health-care applications where the condition of a patient can deteriorate while waiting (e.g. [19]). The analysis of the accumulating priority M/G/1 queue goes back to [11]. Note that it was then called the delay-dependent priority regime, but this has come to mean different things over the years and so we opt to use the accumulating-priority terminology of [20]. This paper analyses the waiting time distribution of the particle class with the lowest AP rate in the general setting of a Lévy driven queue with positive input (see [2]) by means of a novel method. In this setting particles may arrive as a continuous flow or in batches with a whole mass of other particles. Thus, this formulation can also be useful for inventory or insurance models that include heterogeneous input and priorities. The method introduced here relies on a martingale representation of the composite workload accumulation process brought on by customer overtaking, and the subsequent solution of a first-passage time problem.

The M/G/1 queue with a linear AP regime has been extensively studied. The expected waiting times are known to satisfy a recursive formula, (3.47) on p. 131 of [12], which we will refer to as the Kleinrock formula. The initial condition for the recursion is the expected waiting time of the lowest priority classes, which can be computed on its own. The mean-value analysis was extended to other AP functions (without closed form solutions such as the Kleinrock formula): e.g. power law [13], affine [5], and concave [17]. In [20] the distributions of the waiting times for all customer classes were derived as a system of recursive equations for the LST of the different classes. This was done by the construction of an auxiliary process, the maximum-priority process, and the derivation of its LST. The distributional analysis was extended to a multi-server model in [19,16] (the latter included heterogeneous servers), with the additional assumption that service times are exponential. The method of [20] was also shown to be useful for certain non-linear accumulating functions in [15], and for a preemptive priority regime in [4]. In [18] the lowest priority waiting time distribution was derived for a dynamic priority model in which customers jump to a higher priority class after waiting for a certain time threshold. Economic analysis of this model where customers can purchase their AP rates appeared in [6].

Our analysis is the first to address the AP queue in the general case of a Lévy driven queue, i.e. the workload arrival process is

* Corresponding author.
E-mail addresses: offer.kella@huji.ac.il (O. Kella), lravner@post.tau.ac.il (L. Ravner).

http://dx.doi.org/10.1016/j.orl.2016.11.007
0167-6377/

a positive Lévy process and not necessarily a compound Poisson process as in the regular M/G/1. This means that the incoming workload can be any subordinator, that is, a non-decreasing Lévy process. To achieve this we suggest an alternative approach to derive the distribution of the waiting time of the lowest priority customer class. An up to date review of the research of Lévy driven queues can be found in [2] and in greater detail in [3]. Our method involves constructing an accumulative-overtaking process and a stopping time with respect to it that is distributed as the total waiting time. A Wald-type martingale is then used to derive the LST of the waiting time.

In the sequel we establish a decomposition result for the stationary distribution of the workload that will be served before an arriving particle. Such a particle may arrive as part of a batch of particles and thus it will face more workload then just the amount that was present a moment before the batch arrived. This distinction will be important when analysing the overtaking process brought on by the accumulating priority regime in the following sections.

2. Preliminary technicalities

This somewhat more technical section is needed to justify some of the analysis that appears later. Let $J_0 = \{J_0(t)|t \geq 0\}$ be a subordinator, i.e. non-decreasing right continuous Lévy process (see p. 71 of [1]), with respect to some filtration $\{\mathcal{F}_t|t \geq 0\}$ satisfying the usual conditions (right continuous, augmented). This means that $J_0(t) \in \mathcal{F}_t$ for all $t \geq 0$ and that $J_0(t+s) - J_0(t)$ is independent of $\mathcal{F}_t$ for all $s, t \geq 0$. In addition, we assume that $\rho_0 \equiv EJ_0(1) < \infty$.

As in [10], we recall that there are a constant $c_0 \geq 0$, a Lévy measure ν_0 satisfying $\int_{(0,\infty)} x\nu_0(dx) < \infty$, and a Poisson random measure N_0 on $(0,\infty)\times[0,\infty)$ with mean measure $\nu_0 \otimes \ell$, where ℓ is Lebesgue measure, such that

$$J_0(t) = c_0 t + \int_{[0,t]\times(0,\infty)} xN_0(dx, ds).$$

Also, as in (23), (25) and Lemma 1 of [10] it follows that if W is some non-negative càdlàg adapted process and F is a continuous function for which $\int_{(0,\infty)}(F(w+x)-F(w))^2\nu_0(dx)$ is bounded (in w) on $[0,\infty)$, then

$$\begin{aligned} M(t) = &\int_0^t \int_{(0,\infty)} F(W(s-)+x) - F(W(s-))N_0(dx, ds) \\ &- \int_0^t \int_{(0,\infty)} F(W(s)+x) - F(W(s))\nu_0(dx)ds \end{aligned}$$

is a zero mean L^2 martingale satisfying $M(t)/t \to 0$ both a.s. and in L^2. In particular, if F is differentiable with bounded derivative f and we denote by Y_e an independent (of all other processes) random variable with

$$P(Y_e \leq t) = \frac{c_0 + \int_0^t \nu_0(y,\infty)dy}{\rho_0},$$

(see (4.6) of [7]) then it follows with $g(w) = Ef(w+Y_e)$ that

$$\begin{aligned} &c_0 f(W(s)) + \int_{(0,\infty)} F(W(s)+x) - F(W(s))\nu_0(dx) \\ &= c_0 f(W(s)) + \int_{(0,\infty)} \int_0^x f(W(s)+y)dy\nu_0(dx) \\ &= c_0 f(W(s)) + \int_0^\infty f(W(s)+y)\nu_0(y,\infty)dy \\ &= \rho_0 g(W(s)). \end{aligned}$$

With all of the above we have, since $J_0(t)/t \to \rho_0$ a.s., that

$$\begin{aligned} &\frac{1}{J_0(t)}\int_0^t \frac{1}{\Delta J_0(s)}\int_0^{\Delta J_0(s)} f(W(s-)+x)dxdJ_0(s) \\ &- \frac{1}{t}\int_0^t g(W(s))ds \end{aligned} \qquad (1)$$

converges almost surely and in L^2 to zero, where for the case $\Delta J_0(s) = 0$ we define by convention

$$\frac{1}{\Delta J_0(s)}\int_0^{\Delta J_0(s)} f(W(s-)+x)dx \equiv f(W(s-)).$$

Now, note that

$$\begin{aligned} &\int_0^t \frac{1}{\Delta J_0(s)}\int_0^{\Delta J_0(s)} f(W(s-)+x)dxdJ_0(s) \\ &= c_0\int_0^t f(W(s))ds + \sum_{0<s\leq t}\int_0^{\Delta J_0(s)} f(W(s-)+x)dx, \end{aligned}$$

and observe that if $W(\cdot)$ is some content that is found at time t, then the right hand side aggregates the function values of the content in front of all the $J_0(\cdot)$ particles that arrived by time t. That is, if a particle arrives on its own, then the content in front of it is just $W(s-) = W(s)$. If it is in some $x \geq 0$ location in a batch (=jump) then the amount is $W(s-) + x$. If we divide the right hand side by $J_0(t)$, then we have the average function value of the content in front of an arriving particle until time t.

From the fact that (1) vanishes almost surely, it follows that if $W(s)$ has an ergodic distribution of some random variable W, that is

$$\frac{1}{t}\int_0^t g(W(s))ds \to Eg(W) = Ef(W+Y_e),$$

then the long run average distribution of the content in front of particles will be distributed like $W + Y_e$ where W, Y_e are independent. This is some generalized form of the well known PASTA (Poisson Arrivals See Time Averages) property for Poisson processes. See also Remark 2 of [9].

We note that with

$$\eta_0(\alpha) \equiv -\log \mathrm{E}e^{-\alpha J_0(1)} = c_0\alpha + \int_{(0,\infty)}(1-e^{-\alpha x})\nu_0(dx)$$

we have, as in (4.8) of [7], that

$$Ee^{-\alpha Y_e} = \frac{\eta_0(\alpha)}{\rho_0\alpha}. \qquad (2)$$

We conclude that if the process we are dealing with is regenerative with finite mean, nonarithmetic regeneration epochs, then the ergodic, stationary and limiting distributions all coincide. This will be the situation in the sequel.

All one needs to remember from this section is that under the assumptions that will soon appear, the limiting=stationary=ergodic distribution of the content in front of an arriving particle is that of $W + Y_e$ where W, Y_e are independent, W has the steady state distribution of the total content in the system while Y_e has the excess distribution associated with the particular stream of particles (lowest priority) we are interested in.

3. Lévy driven AP queue

A single server processes the workload of N types of particles at a constant rate of r per unit of time. The AP rates of the particle classes are ordered: $b_1 < b_2 < \cdots < b_N$. Thus, the AP at time t of a type i particle that arrived at time s, $s \leq t$, is $b_i(t-s)$. The priority dynamics are illustrated in Fig. 1 for a two-class example.