Queueing models with service speed adaptations at arrival instants of an external observer

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Queueing models with service speed adaptations at arrival instants of an external observer

Rudesindo Núñez-Queija[1] · Balakrishna Prabhu[2] · Jacques Resing[3]

1 Introduction

Motivated by dynamic speed scaling, which enables a balance between performance and energy consumption, we got interested in queues in which the server can work at different service speeds and in which the speed of the server can only be changed at arrival instants of an external observer. In the past, several works dealt with models in which these types of service speed adaptations occurred. Bekker, Boxma and Resing [1] studied an $M/M/1$ queue with a two-stage service rule. Whenever the amount of work in the system is below a threshold, the server wants to work at low service speed, otherwise the server wants to work at high service speed. However, speed adaptations can only occur at Poisson instants. The amount of work in the system could either be measured by the number of customers or by the total workload in the system. Later on, the work was extended to the $M/G/1$ workload process, and to Lévy processes in Bekker, Boxma and Resing [2]. Here, we look at the model in which the server can not only work at two different service speeds but at infinitely many server speeds. When there are j customers in the system, the server wants to work at speed j.

2 Model description

Consider a single-server queue to which arrivals occur according to a Poisson process of rate λ. Each arrival requires an exponentially distributed service time with

✉ Jacques Resing
j.a.c.resing@tue.nl

Rudesindo Núñez-Queija
nunezqueija@uva.nl

Balakrishna Prabhu
Balakrishna.Prabhu@laas.fr

[1] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands

[2] LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France

[3] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

parameter μ. The speed of the server can be dynamically assigned values in the set $\{0, 1, 2, \ldots\}$. Adjustments to the speed can be made only at control instants which are assumed to occur according to an independent Poisson process of rate ν. At a control instant, the speed of the server is set equal to the number of customers observed at that instant. Between any two consecutive control instants, the speed of the server cannot be changed. Let $Q(t)$ denote the number of customers in the system at time t and $S(t) = Q(\tau_t)$, with τ_t the last control instant at or before time t. $S(t)$ can be interpreted as the speed of the server at time t. The process $(Q(t), S(t))_{t \geq 0}$ is a Markov process with transition rates

$$(Q(t), S(t)) \rightarrow \begin{cases} (Q(t) + 1, S(t)) & \text{with rate } \lambda; \\ (Q(t) - 1, S(t)) & \text{with rate } \mu S(t); \\ (Q(t), Q(t)) & \text{with rate } \nu. \end{cases} \tag{1}$$

Customer arrivals and service completions lead to jumps to neighbouring states, arrivals of the observer lead to jumps to states on the diagonal of the positive quadrant.

3 Problem statement

For arbitrary $\rho = \lambda/\mu$, the system will be stable. Denote with $\pi(i, j) = \lim_{t \to \infty} P(Q(t) = i, S(t) = j)$ the steady-state probabilities of the two-dimensional Markov process and let

$$P(x, y) = \sum_{i \geq 0, j \geq 0} \pi(i, j) x^i y^j \tag{2}$$

be the corresponding joint probability generating function. From the balance equations, one can obtain the following proposition.

Proposition 1 $P(x, y)$ *is the solution of the functional equation*

$$(\nu + \lambda(1 - x)) P(x, y) + \mu y \left(1 - \frac{1}{x}\right) \frac{\partial}{\partial y} [P(x, y) - P(0, y)] = \nu P(xy, 1). \tag{3}$$

Main problem: What is the solution of functional equation (3)? In particular, our focus is on characterizing the steady-state distribution for some asymptotic regimes.

4 Discussion

It is easily seen for this model that the marginal steady-state distributions of the processes $Q(t)$ and $S(t)$ are the same (i.e. $P(y, 1) = P(1, y)$). Furthermore, given that the server works at speed j, the number of customers in the system behaves as the number of customers in an $M/M/1$ queue with arrival rate λ, service rate $j\mu$ and in which, after exponentially distributed times with parameter ν, the number of customers in the system is set equal to j again. Let $p_j(\ell)$ be the stationary conditional probability of having ℓ customers in the system when the service rate is $j\mu$ and let $f_j(z)$ be the generating function of this stationary conditional distribution. Furthermore, define

$$\beta_j(\nu) = \frac{\lambda + j\mu + \nu - \sqrt{(\lambda + j\mu + \nu)^2 - 4\lambda(j\mu)}}{2\lambda}. \tag{4}$$

Proposition 2 *For the generating function $f_j(z)$, we have*

$$f_0(z) = \frac{\nu}{\nu + \lambda(1-z)} \tag{5}$$

$$f_j(z) = \frac{\nu\tilde{\beta}_j(\nu)\sum_{k=0}^{\infty} c_{k,j}z^k}{\lambda(1-\tilde{\beta}_j(\nu)z)}, \quad j = 1, 2, \ldots, \tag{6}$$

where

$$\tilde{\beta}_j(\nu) = \frac{\lambda\beta_j(\nu)}{j\mu}, \tag{7}$$

$$c_{k,j} = \begin{cases} (1-\beta_j(\nu))^{-1}\beta_j(\nu)^j & k = 0; \\ \beta_j(\nu)^{j-k} & k = 1, \ldots, j; \\ 0 & k > j. \end{cases} \tag{8}$$

Proposition 2 might be helpful in the proof of the conjecture below.

4.1 Asymptotics for $\nu \to \infty$ and $\nu \to 0$

If $\nu \to \infty$, then $P(x, y) \to e^{\rho(xy-1)}$. In this case $\pi(i, j) \to 0$ for $i \neq j$ and the processes $Q(t)$ and $S(t)$ will, simultaneously, behave as an ordinary $M/M/\infty$ system.

For $\nu \to 0$, due to the increasing lengths of periods in which the service rate is smaller than the arrival rate, we have to consider the scaled pair of random variables $(\nu Q(t), \nu S(t))$. Define $\hat{P}_\nu(x, y) = P(x^\nu, y^\nu)$ to be the joint probability generating function of the steady-state distribution of this scaled pair of random variables and let $\hat{P}(x, y) := \lim_{\nu \to 0} \hat{P}_\nu(x, y)$. Then, we conjecture that

$$\hat{P}(x, y) = \frac{1}{2}\frac{1}{1-\lambda\log(x)} + \frac{1}{2}\frac{1}{1-\lambda\log(y)}. \tag{9}$$

This means that as $\nu \to 0$ the probability mass of the joint process concentrates around the two axes. On each of these two axes, the scaled stationary process behaves like an exponentially distributed random variable of rate λ^{-1}. The coefficients $1/2$ in (9) correspond to the proportion of time spent by the process on each of the two axes.

For future work, it is interesting to allow the rate ν of the control process to depend upon the state of the system. For example, a higher backlog may require a higher frequency of observations.

References

1. Bekker, R., Boxma, O., Resing, J.: Queues with service speed adaptations. Statistica Neerlandica **62**, 441–457 (2008)
2. Bekker, R., Boxma, O., Resing, J.: Lévy processes with adaptable exponent. Adv. Appl. Probab. **41**, 177–205 (2009)