A Non Monotone Waiting Time Queueing Model*

M. Hlynka, S. Drekic, G. McPhail
Department of Mathematics and Statistics
University of Windsor
Windsor, ON N9B 3P4

ABSTRACT. We create an example of a FCFS queueing model with a particular service distribution for which the expected waiting time of an arriving customer is, surprisingly, NOT monotone increasing as a function of the customers already waiting in line.

1 Introduction

Hlynka, Stanford, Poon, and Wang [2] showed the difficulties of deciding which queue to join when selecting from two parallel queues. In Whitt [4], an example of a system with two parallel servers, each with identical service distributions, was presented to show that join-the-shortest-queue strategies were not always optimal.

In our present article, we consider a single server queueing model. We assume exponentially distributed interarrival times. We will construct a Coxian (Cox [1]) service distribution for which the expected time to completion of an arriving customer is not monotonically increasing in the number of customers present in the line when a customer arrives. This is an unexpected characteristic, since a shorter line generally means a shorter expected waiting time.

In our (counter)example, as in Whitt's, the service times are not exponentially distributed. Whitt mentioned the difficulty of constructing counterexamples for his system. Fortunately, our service time distribution can be considered to be the service time within a network with several exponential subservers. This allows us to use a conventional continuous time

*This research is supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Markov chain analysis. The nature of Whitt's example required a different analysis.

A relationship exists between our example of a non monotone expected waiting time and the choice of which queue to join in a two identical server parallel queueing system. Certainly if a longer queue leads to shorter waiting times, then a customer would want to join a longer queue. However, in the parallel queue model, one must consider the queue joining strategy of ALL the customers. The state space in that case thus becomes more cumbersome. The one server model that we present allows an simpler analysis and gives an novel result.

Assume that interarrival times are exponentially distributed at rate λ. We have the following

$$- - - - \rightarrow - - - [\text{Server}] - - - - - \rightarrow - -$$

The single server and its service time distribution, can be replaced by the following network of servers.

$$\begin{array}{c} \uparrow - - - - \rightarrow - - - - - \downarrow \\ - - \rightarrow [\mu_1] - - \rightarrow [\mu_2] - - \rightarrow [\mu_3] - - \rightarrow \end{array}$$

Assuming that the network is empty, an arriving customer enters and is served by the first exponential server at rate μ_1. With probability p, the customer exits the system completely. With probability $1-p$, the server is served by a second exponential server at rate μ_2, then by a third exponential server at rate μ_3. One could think of a customs station with secondary and tertiary inspections by the same inspector.

If there is already at least one customer in the system, an arriving customer must wait in front of server 1. Service on a new customer does not begin until the customer in the network has exited.

An arriving customer does not actually see the structure of the network. The arriving customer merely sees a single server, (at most) one customer being served and (perhaps) other customers waiting for service. The unobserved network is a special case of a Coxian distribution. There are n customers in the single server queue iff there are n customers in the network.

2 Analysis

For the network, define the states to be of the form (i, j, k), where the first component of the ordered triple represents the number of customers waiting or being served by the first server. The second and third components represent the customers receiving service from the second and third servers, respectively. We impose the restriction that $j + k \leq 1$. We write the states

in the following order: $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(2,0,0)$, $(1,1,0)$, $(1,0,1)$, $(3,0,0)$, ...

Then the rate matrix can be written in the form

$$\Lambda = \begin{bmatrix} A_{00} & A_{01} & 0 & 0 & 0 & \dots \\ A_{10} & A & B & 0 & 0 & \dots \\ 0 & C & A & B & 0 & \dots \\ 0 & 0 & C & A & B & \dots \\ 0 & 0 & 0 & C & A & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix},$$

where

$$A_{00} = [-\lambda], \quad A_{01} = [\lambda\ 0\ 0],$$

$$A_{10} = \begin{bmatrix} p\mu_1 \\ 0 \\ \mu_3 \end{bmatrix}, \quad A = \begin{bmatrix} -(\lambda+\mu_1) & (1-p)\mu_1 & 0 \\ 0 & -(\lambda+\mu_2) & \mu_2 \\ 0 & 0 & -(\lambda+\mu_3) \end{bmatrix},$$

$$B = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \quad C = \begin{bmatrix} p\mu_1 & 0 & 0 \\ 0 & 0 & 0 \\ \mu_3 & 0 & 0 \end{bmatrix},$$

Thus the first few rows of Λ are look like

$$\Lambda = \begin{matrix} (0,0,0) \\ (1,0,0) \\ (0,1,0) \\ (0,0,1) \end{matrix} \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & 0 & \dots \\ p\mu_1 & -(\lambda+\mu_1) & (1-p)\mu_1 & 0 & \lambda & 0 & 0 & \dots \\ 0 & 0 & -(\lambda+\mu_2) & \mu_2 & 0 & \lambda & 0 & \dots \\ \mu_3 & 0 & 0 & -(\lambda+\mu_3) & 0 & 0 & \lambda & \dots \end{bmatrix}$$

A limiting probability vector with all non zero limiting probabilities will exist iff $\lambda < \min\{\mu_1, \frac{\mu_2}{1-p}, \frac{\mu_3}{1-p}\}$, $0 < \lambda$, $\mu_1, \mu_2, \mu_3 < \infty$, $0 < p < 1$.

Let $v = (v_{000}, v_{100}, v_{010}, v_{001}, v_{200}, v_{110}, v_{101}, \dots)$ be the limiting probability vector. To find the components of v, we need to solve the equation $v\Lambda = 0$. We obtain the following equations.

$$-\lambda v_{000} + p\mu_1 v_{100} + \mu_3 v_{001} = 0 \tag{1}$$

$$\lambda v_{000} - (\lambda+\mu_1)v_{100} + p\mu_1 v_{200} + \mu_3 v_{101} = 0 \tag{2}$$

$$(1-p)\mu_1 v_{100} - (\lambda+\mu_2)v_{010} = 0 \tag{3}$$

$$\mu_2 v_{010} - (\lambda+\mu_3)v_{001} = 0 \tag{4}$$

For $n \geq 1$,

$$\lambda v_{n00} - (\lambda+\mu_1)v_{n+1,0,0} + p\mu_1 v_{n+2,0,0} + \mu_3 v_{n+1,0,1} = 0 \tag{5}$$

$$\lambda v_{n-1,0,0} + (1-p)\mu_1 v_{n+1,0,0} - (\lambda+\mu_2)v_{n10} = 0 \tag{6}$$

$$\lambda v_{n-1,0,0} + \mu_2 v_{n10} - (\lambda+\mu_3)v_{n01} = 0 \tag{7}$$

Property 2.1: *If* $p = 1$, $\lambda < \mu_1$, *and* $n \geq 1$, *then* $v_{000} = 1 - \frac{\lambda}{\mu_1}$, $v_{n00} = (1 - \frac{\lambda}{\mu_1})(\frac{\lambda}{\mu_1})^n$, $v_{n-1,1,0} = 0$, *and* $v_{n-1,0,1} = 0$.

Proof: For $p = 1$, we have an $M/M/1$ queueing system and the result follows.

Property 2.2: *Let* $\alpha = \lambda + \mu_2$ *and* $\beta = \lambda + \mu_3$. *For* $p < 1$, *the following hold.*

$$\begin{aligned} v_{100} &= \frac{\lambda\alpha\beta}{\mu_1[\lambda p\beta + \mu_2(\lambda p + \mu_3)]} v_{000} \\ v_{010} &= \frac{\lambda(1-p)\beta}{\lambda p\beta + \mu_2(\lambda p + \mu_3)} v_{000} \\ v_{001} &= \frac{\lambda(1-p)\mu_2}{\lambda p\beta + \mu_2(\lambda p + \mu_3)} v_{000}. \end{aligned} \tag{8}$$

Letting $A = \lambda(\mu_1 + \mu_2 + \mu_3)$, $B = \lambda^2(1-p)$, *we also have*

$$\begin{aligned} v_{200} &= \frac{\lambda^2\alpha\beta[A + \lambda(\lambda - p\mu_1) + (1-p)\mu_1(\mu_2 + \mu_3) + \mu_2\mu_3] - B\mu_1\mu_2\mu_3(\alpha + \beta)}{\mu_1^2[\lambda p\beta + \mu_2(\lambda p + \mu_3)]^2} v_{000} \\ v_{110} &= \frac{B\beta[\lambda^2 + A + \mu_3(\mu_1 + \mu_2)] + \lambda^3(1-p)^2\mu_1\mu_2}{\mu_1[\lambda p\beta + \mu_2(\lambda p + \mu_3)]^2} v_{000} \\ v_{101} &= \frac{B\mu_2[\lambda^2 + A + \mu_1(\lambda p + \mu_2 + \mu_3) + \mu_2\mu_3]}{\mu_1[\lambda p\beta + \mu_2(\lambda p + \mu_3)]^2} v_{000}. \end{aligned} \tag{9}$$

Proof: From (3) and (4), we obtain $v_{100} = \frac{\alpha}{(1-p)\mu_1}$ and $v_{001} = \frac{\mu_2}{\beta} v_{010}$. By substituting these expressions into (1), we obtain our expressions for v_{100}, v_{010}, v_{001}. From (7) and (8), with $n = 1$, we can eliminate v_{110}. Using that expression together with (2), we obtain an expression for v_{200}. Using (6) and (7) again, we obtain our expressions for v_{110} and v_{010}. □

Our limiting probabilities already found all involve v_{000} which is unknown. However, it turns out that we can get useful information without knowing this value.

Property 2.3: *Let* $E(T|n)$ *be the expected time for a randomly arriving customer to enter service given that there are already* n *customers in the system. Let*

$$A_1 = v_{100} + v_{010} + v_{001}, \quad A_2 = v_{200} + v_{110} + v_{101}, \tag{10}$$

and $\gamma = \frac{1}{\mu_1} + \frac{1}{\mu_2}$. *Then*

$$E(T|1) = \frac{1}{A_1}[v_{100}[\frac{1}{\mu_1} + \gamma(1-p)] + \gamma v_{010} + v_{001}\left(\frac{1}{\mu_3}\right)]$$
$$E(T|2) = \frac{1}{A_2}[v_{200}[\frac{2}{\mu_1} + 2\gamma(1-p)] + v_{110}[\frac{1}{\mu_1} + 2\gamma(2-p)]$$
$$+ v_{101}[\frac{1}{\mu_1} + \frac{1}{\mu_3} + \gamma(1-p)]].$$

Proof: Let ET_{ijk} denote the expected waiting time for an arriving customer to enter service (first server), given that the arriving customer encounters the state (i,j,k) on arrival. By the memoryless property for the exponential distribution, the expected time will not depend on the time since the last movement, but merely on the configuration (i,j,k). We find

$$E(T|1) = \frac{1}{A_1}[v_{100}ET_{100} + v_{010}ET_{010} + v_{001}ET_{001}]$$
$$E(T|2) = \frac{1}{A_2}[v_{200}ET_{200} + v_{110}ET_{110} + v_{101}ET_{101}] \qquad (11)$$

The result follows. □

Intuitively, one would think that under all circumstances, the expected waiting time for an arriving customer to enter service would be longer if there were more customers ahead of it upon arrival. After all, the total waiting time would be the sum of the service times of the other waiting customers plus a residual service time of the customer currently in service. The expected sum of the service times of the waiting customers is clearly larger if the line is longer, since the service times are identically distributed.

We will show that our intuition is faulty, due to the residual service time of the customer in service. Our goal is to show that the function $E(T|n)$ is not always monotone increasing in n for all parameters μ_1, μ_2, μ_3, p.

Theorem 2.4. *There exist values of values of λ, μ_1, μ_2, μ_3, p, for which $E(T|n)$ is not an increasing function of n.*

Proof: For the model already presented, choose $\lambda = 1$, $\mu_2 = \mu_3$ and let $\mu_1 \to \infty$. We thus have only two variables to work with and we rename them by letting $\mu_2 = \mu_3 = x$ and $p = y$. With this notation, our equations

(8)-(10) become

$$v_{100} = 0 \qquad v_{010} = \frac{(1+x)(1-y)}{y(1+x)+x(x+y)}v_{000}$$

$$v_{001} = \frac{x(1-y)}{y(1+x)+x(x+y)} \qquad A = \frac{(1+2x)(1-y)}{y(1+x)+x(x+y)}v_{000}$$

$$v_{200} = 0 \qquad v_{110} = \frac{[(1+x)^2+x(1-y)](1-y)}{[y(1+x)+x(x+y)]^2}v_{000}$$

$$v_{101} = \frac{x(1+2x+y)(1-y)}{[y(1+x)+x(x+y)]^2}v_{000} \qquad B = \frac{(1+4x+3x^2)(1-y)}{[y(1+x)+x(x+y)]^2}v_{000}$$

Since $v_{000} \neq 0$, we obtain from (11)

$$\begin{aligned} E(T|1) &= \frac{y(1+x)+x(x+y)}{(1+2x)(1-y)}\left[\frac{(1+x)(1-y)}{h(1+x)+x(x+y)}\left(\frac{2}{x}\right)\right. \\ &\quad \left. + \frac{x(1-y)}{y(1+x)+x(x+y)}\left(\frac{1}{x}\right)\right] \\ &= \frac{2+3x}{x(1+2x)} \end{aligned}$$

$$\begin{aligned} E(T|2) &= \frac{[y(1+x)+x(x+y)]^2}{(1+4x+3x^2)(1-y)}\left[\frac{[(1+x)^2+x(1-y)](1-y)}{[y(1+x)+x(x+y)]^2}\left(\frac{1(2-y)}{x}\right)\right. \\ &\quad \left. + \frac{x(1+2x+y)(1-y)}{[y(1+x)+x(x+y)]^2}\left(\frac{3-2y}{x}\right)\right] \\ &= \frac{2(1+x)^2(2-y)+2x(1-y)(2-y)+x(1+2x+y))(3-2y)}{x(1+4x+3x^2)} \end{aligned}$$

Let $F(x,y) = E(T|2) - E(T|1)$. If we fix $x = .7$, for example, and plotted $F(.7,y)$ versus $y(0 \leq y \leq 1)$, then we would observe that $F(.7,y)$ takes on negative values for y near 1. This show that $E(T|2) < E(T|1)$ and completes the proof. □

Comment The actual service distribution c.d.f. given in the previous property is

$$F(t) = \begin{cases} 0 & \text{for } t < 0 \\ p + \int_0^t \mu_1 w^2 e^{-\mu_1 w} dw & \text{for } t \geq 0. \end{cases}$$

3 Conclusions

We have constructed a service distribution such that an arriving customer would not always prefer to encounter a short line (one customer) as compared to a longer line (two customers). We have used standard Markov process methods on a particular Coxian service distribution.

Our service distribution was equivalent to one with three exponential subservers, the first of which had infinite service rate. Could we have used only two servers, assuming that the first server had infinite service rate? The answer is no. If we had only two servers, and we observed $n > 0$ customers in the system, then we would know that one customer is being served by the second server. Since the second server has exponentially distributed (memoryless) service times, there is no advantage (and in fact there is a disadvantage) to seeing more customers upon arrival.

Our service distribution is in fact not uncommon. At a customs booth, for example, most customers (travellers) pass through with a very short service time. Somtimes, however, a customer is given a much longer detailed inspection, and even more rarely, is thorough searched.

What we have shown is that we are not always happy to encounter a queue with a small number of customers waiting. Some papers (e.g. Kulkarni and Sethi [3]) deal with situations where an arriving customer can return to a queueing system later if the number of customers already waiting is a number that the arriving customer considers undesirable. In the example that we have presented, we have the unusual situation that an arriving customer might encounter one customer already in the system and choose to leave and return later with the hope of finding two customers in the system!

References

[1] D.R. Cox. The Analysis of Non-Markovian Stochastic Processes by the Inclusion of Supplementary Variables, *Proc. Camb. Philos. Soc.* **51** (1955), 433–441.

[2] M. Hlynka, D.A. Stanford, W.H. Poon, W.H., and T. Wang. Observing Queues Before Joining, *Operations Research* **42** (1994), 365–371.

[3] V.G. Kulkarni and S.P. Sethi. Deterministic Retrial Times are Optimal in Queues with Forbidden States, *INFOR* **27** (1989), 374–386.

[4] W. Whitt. Deciding Which Queue to Join: Some Counterexamples, *Operations Research* **34** (1986), 55–62.