A STUDY ON THE SHORTEST QUEUEING MODELS WITH PARALLEL QUEUES

K. ALAGESAN A. CHARLES SAGAYARAJ, M. RENI SAGAYARAJ, S. ANAND GNANA SELVAM

Abstract: In this thesis, we consider a stochastic model of two queues (queue I and queue II) with an interconnection between them is considered. When the Principal queue length exceeds a certain number N, secondary server open his own queue. Customer arrives according to a Poisson process with rate λ. Customers arrive at a constant rate λ and join the shortest of the two queues the service time for each server is exponentially distributed but the rates may be different. Jockeying between two queues is allowed. We assume that customers that customers always join the shortest queue. Parallel queueing system is obtained.

Keywords: Asymptotically bounds, exponentially distributed, jockeying, linear programming techniques, single- server queue

Introduction: We consider many other situations; customers are often confronted with two queues, each with its own server. Customers generally enter the shortest queue available. In some situations, customers move from one queue to another. This moment is called jockeying. Such queueing systems, where customers must choose from servers, each with its own queue are called parallel queues. This concept was first studied by [5] Height (1958). His work was followed by Keilson (2004) and others. Because each server has its own queue. Several authors have studied assignment of customers to system with Parallel queues. Kulkarni (1983) assumed that a customer would join the shortest queue available. If both queues were join the same size then a customer would join the queue. He also assumed that both servers had the same service rate. The problem to be analyzed is generally referred to as the "shortest queue problem". Wang (2004) formulae based on Flatto and Mckean which were superior for numerical purposes. In this thesis, we consider the following modification of the shortest queue model with jockeying. It is assumed that there is a Principal server who is always available when customer demand exists. When the Principal queue length exceeds a certain number N, secondary server open his own queue. Customer arrives according to a Poisson process with rate λ.

Model Description: A Poisson stream of customer with rate λ arrives at a service system which consists of two single-server queues in parallel[3].The service time of the customers are independent and exponentially distributed with rate μ.Both servers serve at an equal rate of 1.Each server has an associated queuing space of unlimited capacity. An arriving customer joins the shorter of the two queues, if their sizes are unequal, otherwise he joins any queue. Jockeying between the queues is not allowed.

This system, which is known as the 'shortest queue', has received a lot of attention in the literature, because it (with its generalizations to many servers and to general service times) models many real-life situations such as vehicles going through toll booths, jobs scheduled on a multiprocessor system, etc, .On the other hand, no simple method for analyzing this system is known. The problem was originally introduced by Haight (1958).Kingman (1961) proved that an equilibrium distribution exists whenever $\lambda/\mu < 2$.Both he and Flatto and McKean(1977)treated the problem by applying techniques of complex-function theory to obtain representation for the general functions of the state probability. The above mentioned papers derive some asymptotic approximations for the state probabilities for large number of customers in the system, and for heavy traffic. Conolly(1984) discussed the finite –waiting –room version of the problem. In the current paper we derive upper and lower bounds for the state probabilities, tail probabilities, and mean number of customers in the system, in equilibrium. The bound are derived by considering a subset of the balance equation, which gives rise to linear programs which in turn produce the bounds .We also derive lower bounds for the tail distribution by comparison with an M/M/2 system. These rather elementary techniques produce bound which are within 10% of the true values for $1 \le \lambda/\mu < 2$, and which are asymptotically tight in heavy traffic.

Transition state results with shortest queue: The state space consists of pairs (i,j) where i,j=0,1,2,3.....,and I ≥ j. We say that the system is in state (i,j) if number of customers in the longer queue is i and the number in the shorter queue is j. Note that under this description, servers are not associated with a particular Component of the state vector, and, there is no need to specify what happens when a customer finds that both queues are of equal length. Clearly the system behaves as a Markov chains on the state space, with transition, intensities

as described in following Figure. Let a= λ/μ .We assume that a <2, which implies that the system in stable. Let p_{ij} be the equilibrium probability of the state (i, j).

Let $\pi_n = \sum_{i+j=n} p_{ij}$, n=0, 1, 2.....Thus π_n is the probability that there are exactly n customers in the system in equilibrium.

Denote by q_k the sum of the probabilities in the k[th] diagonal.

We get

$$\lambda p_{i-1i-1} = \mu \sum_{j=0}^{i-1} p_{ij}, i = 1,2,..... \quad (3.1)$$

Summing (3.1) over i we get

$$\lambda q_0 = \mu q^* \quad (3.2)$$

Since $q_0 + q^* = 1$,we can solve them:

$$q_0 = \frac{1}{1+a}, q^* = \frac{a}{1+a} \quad (3.3)$$

Next, define the 'diagonals'

$$D_k = \bigcup_{i=0}^{\infty} (i+k, i) \quad k = 0,1,... \quad (3.4)$$

and let us separate the diagonals

$D_0, D_1, ...D_k$ from $D_{k+1}, D_{k+2}.....$

The resulting cut yields

$$(1+a)q_1 = (2+a)q_0 - 2p_{00}, \text{for } k = 0,... \quad (3.5)$$

$$(1+a)q_{k+1} = q_k - p_{k0}, \text{ for k= } 0,1,2,... \quad (3.6)$$

Summing (3.5) and (3.6) over k we get

$$(1+a)q^* = (2+a)q_0 + q^* - (2p_{00} + \sum_{k=1}^{\infty} p_{k0}) \quad (3.7)$$

and subtracting the values of qo and q*, we get the following result .

Lemma:

$$2p_{00} + \sum_{k=1}^{\infty} p_{k0} = 2 - a, \text{for } 0 \le a < 2 \quad (3.8)$$

It is introduce to note that $q_0, q_1, ...$ exist even when a ≥ 2,if we interpret q_k as the limit of the probability of being in D_k at time t ,when t→∞, independent of the initial conditional. To see that, when a ≥ 2 the probability that any queue is empty at time t converges to o for t → ∞. Thus the processes behaves asymptotically as birth and death process on the diagonals D_k, with the birth rate being μ and the death rate being $\lambda + \mu$ for k>o,and for k=o the birth rate being $\lambda + 2\mu$ and the death rate o.

Thus a limiting distribution exists, independent of the initial conditions, and it can be calculated explicitly as follows.

Theorem: 1

$q_k = \sum_{i=0}^{\infty} p_{i+ki,}$, k=o, 1, 2...

and let q* = $\sum_{k=1}^{\infty} q_k$ be the probability that the queues are unequal.

It is well known that for a Markov chain in equilibrium, if the state space is divided into two disjoint subsets, then the 'intensity flows' from one subset into the other are equal for both subsets.

First, for a fixed i, we divide the states into those with first component less than i, and those with first component greater than or equal to i.

$$q_0 = \begin{cases} \frac{1}{1+a}, for 0 \le a \prec 2 \\ \frac{a}{2(1+a)} for 2 < a \end{cases}$$

$$q_1 = \frac{a(a+2)}{2(1+a)^2} \quad \text{for} \quad 2 < a,$$

and $q_k = q_1 \frac{1}{(1+a)^{k-1}}$ for $2 < a.$

4. The total number in the system; Let us divide the state space into those states with n or more customers, and those with n-1 or less. The intensity flow equation, over these cuts yields:

$$\pi_n = a/2\pi_{n-1} + 1/2 p_{n0} \qquad n = 1,2... \qquad (4.1)$$

Iterating (4.1) by substituting $\pi_{n-1}, \pi_{n-2}, ...$ and noticing that $\pi_0 = p_{00}$

$\pi_n = (a/2)^n p_{00} + 1/2((a/2)^{n-1} p_{10} + (a/2)^{n-2} p_{20} + ... + p_{n0}),$ n = 0,1,2.. (4.2) We are now going to use (4.2) to derive bounds for the $\pi_n S$, by employing the relation found in section 3.From(3.3),(3.5)and (3.6)we get

$$p_{00} = \frac{2+a}{2(1+a)} - \frac{1+a}{2} q_1 \qquad (4.3)$$

$p_{i0} = q_i - (1+a)q_{i+1}$, i=1,2,.....n,

Substituting in (4.3) and rearranging we get

$$\pi_n = (a/2)^n \frac{2+a}{2(1+a)} + \frac{1}{4}(2-a-a^2)\sum_{i=1}^{n}(a/2)^{n-i} q_i - \frac{1+a}{2} q_{n+1} \; n = 0,1,... \qquad (4.4),$$

Consider the linear program with n+ 1 variable $X_1, X_2, ...X_{n+1},$ with the objective function

$$f(X) = \frac{1}{4}(2-a-a^2)\sum_{i=1}^{n}(a/2)^{n-i} X_i - \frac{1+a}{2} X_{n+1} \qquad (4.5)$$

and with the constraints:

$$\begin{cases} X_k \ge 0, k = 1,2,.....n+1 \\ C_0 : X_1 \le \frac{2+a}{(1+a)^2} \\ \overline{C} : X_1 + ...X_n + X_{n+1} \le \frac{a}{1+a} \\ \underline{C} : X_1 + ...X_n + \frac{1+a}{a} X_{n+1} \ge \frac{a}{1+a}. \end{cases}$$

Conclusion: The approach of the probability to its heavy traffic limit is very steep. So that this limit, which is easily computable, will not be a good approximation for most values of the load. Asymmetry in the service rates tends to extend to increase this probability in light traffic but to decrease it in moderate to heavy traffic.

Reference:

1. Aissani.A and J.R.Artalejo, On the single server retrial queue subject to breakdown, Queueing systems, 30(1998), pp. 309-321.
2. Artalejo,J. R.,(1998), Retrial queues with a finite number of sources, Journal of the Korean Mathematical Society, 35, pp. 503 - 525.

3. Chakravarthy,SR. Dudin,AN , An multi-server retrial queue with BMAP arrivals and group services, queueing systems 2002, 42, pp. 5-31.
4. Conolly,B.W.(1984), The autostrada queueing problem.J.Appl. Prob, 21, pp. 394- 403.
5. Falin,G.I.(1986)single-line repeated orders queueing systems Mathematics operations for schung and statistick, optimations, no 5, pp. 649-670.
6. Fayolly, G.,A Simple Telephone Exchange with Delayed Feedbacks, in: O. J. Boxma,
7. J.W. Cohen, H. C.Tijms(Eds.),Traffic Analysis and Computer Performance Evaluation, Elsevier, Amsterdam, pp. 245-253(1986).
8. Gaver, D.P., Jacobs, P.A. and Latouche, G., (1984), Finite birth-and-death models in randomly changing environments, Advances in Applied Probability, 16, pp. 715-731.
9. Haight, F.A.(1958) Two queues in parallel .Biometrika 45,401-410.
10. Keilson,J., Cozzolino,J. and Yang,H., A service system with unfilled requests repeated, Operations research, Vol 16, pp. 1126-1137, 2004.
11. V.G.Kulkarni,On queueing systems with retrials, Journal of Applied probability, 20(1983), pp. 380-389.
12. Mokaddis,G.S., Metwally,S.A and Zadi,B.M., 2007. A Feedback Retrial queueing system with starting Failures and single vacation, Tamkang Journal of Science and Engineering,Vol 10, No.3, pp. 183-192.
13. Netus.M.F, Two further closure properties of Ph-Distribution, Asia pacific Journal of Operational Research, 23, pp. 241-260.
14. Takagi,H 1991. A foundation of performance Evaluation.Vol.1, Vacation and priority system,Part1 ; Elsevier Science, Newyork.
15. Wang,J., An queue with second optional service and server breakdowns, Computers and Mathematics with applications, Vol.47, pp. 1713-1723, 2004.
16. Yang.T., Posner ,M.J..M.,and Templton J.G.C (1990), The M/G/1 retrial queue with non-persistant customers, Queueing systems 7(2); pp. 209-218.

K. Alagesan, A. Charles Sagayaraj
Department of Mathematics, Kandaswamy Kandar College, Velur.
M. Reni Sagayaraj, S. Anand Gnana Selvam
Department of Mathematics, Sacred Heart College, Tirupattur, India.