Head of the line processor sharing for many symmetric queues with finite capacity

J.A. Morrison

AT&T Bell Laboratories, Room 2C-378, 600 Mountain Avenue, Murray Hill, NJ 07974, USA

Received 8 June 1992; revised 9 November 1992

In this paper the steady-state behavior of many symmetric queues, under the head of the line processor-sharing discipline, is investigated. The arrival process to each of $n$ queues is Poisson, with rate $\lambda$, and each queue has $r$ waiting spaces. A job arriving at a full queue is lost. The queues are served by a single exponential server, which has a mean rate $\mu n$, and splits its capacity equally amongst the jobs at the head of each nonempty queue. The normal traffic case $\rho = \lambda / \mu < 1$ is considered, and it is assumed that $n \gg 1$ and $r = O(1)$. A 2-term asymptotic approximation to the loss probability $L$ is derived, and it is found that $L = O(n^{-r})$, for fixed $\rho$. If $\delta = (1 - \rho) / \rho < 1$, then the approximation is valid if $n\delta^2 \gg 1$ and $(r + 1)^2 \ll n\delta$, and in this case $L \sim r! / (nr)^r$. Numerical values of $L$ are obtained for $r = 1, 2, 3, 4$ and $5, n = 1000, 500$ and $200$, and various values of $\rho < 1$. Very small loss probabilities may be obtained with appropriate values of these parameters.

Keywords: Asymptotics; head of the line processor sharing; loss probability; many queues.

1. Introduction

In this paper we investigate the steady-state behaviour of many symmetric queues with Poisson arrivals, a finite number of waiting spaces, and a single exponential server with head of the line processor sharing. This provides a model for the buffering of channels (virtual circuits) in a wide-area data network. The head of the line processor-sharing discipline is a close approximation to the head of the line round-robin discipline, which is used for reasons of fairness. Hundreds or thousands of channels may be served by a single server.

Head of the line processor sharing for two queues with infinitely many waiting spaces has been investigated by several authors. Fayolle and Iasnogorodski [2] derived an explicit, albeit complicated, expression for the generating function of the steady-state joint distribution of the queue lengths, in the general case of two asymmetric queues. In a later paper, Konheim, Meilijson and Melkman [11]...
derived this generating function, in an elegant manner, in the case of two symmetric queues.

Knessl [10] investigated the tail behavior of the (heavy-traffic) diffusion approximation to the joint distribution of the queue lengths, when the arrival rates are equal, but the service rates for the queues are different. He used the ray method [8] and the method of matched asymptotic expansions [1] to solve the equations for the diffusion approximation asymptotically, but his results are not complete. However, in the case in which the service rates are also equal, he obtained an exact solution for the diffusion approximation. Recently, Morrison [13] derived an explicit expression for the diffusion approximation to the joint distribution of the queue lengths, in the general case of two asymmetric queues. He also investigated the tail behavior of the joint and marginal distributions.

Hooghiemstra, Keane and van de Ree [7] derived a power-series expansion in the traffic intensity which yields a representation of the joint distribution of the queue lengths that applies to multiple queues with head of the line processor sharing. However, the established radius of convergence decreases rapidly as the number of queues increases.

Fendick and Rodrigues [3] recently investigated a system of symmetric queues with head of the line processor sharing, when the number of buffers (waiting spaces) is finite. The arrival processes are general, although independent and identical, and the service time of each job has an independent and identical distribution. They modeled the system as a Brownian flow [6], under the assumption of heavy-traffic conditions, and obtained approximate loss probabilities for shared and segregated buffers.

In this paper we assume that the arrival process to each of n queues is Poisson, with rate $\lambda$, and that the arrival processes are mutually independent. (This is equivalent to a single Poisson arrival process with rate $\lambda n$, after which arrivals are assigned to one of the queues independently with equal probability.) Each queue has $r$ waiting spaces, and jobs arriving at a full queue are lost. The queues are served by a single exponential server, which has a mean rate $\mu$, and splits its capacity equally amongst the jobs at the head of each nonempty queue. We consider the normal traffic case in which $\rho = \lambda / \mu < 1$, and assume that $n \gg 1$ and $r = O(1)$. If $r = 1$ the system is equivalent to a finite population $M/M/1/n/n$ queue, so we suppose that $r \geq 2$ in the analysis.

Greenberg and Whitt [5] considered the same system as we do, but in the very heavy traffic case $\rho > 1$. They investigated the transient problem, and used a fluid approximation to analyze the proportion of the $n$ queues that have $j$ jobs at time $t$, $j = 0, \ldots, r$. They determined the equilibrium points of the corresponding system of differential equations, but the loss probability is too large for our purposes when $\rho > 1$.

In section 2 we formulate the problem, and introduce a generating function, which satisfies a second order linear partial differential equation with $r$ independent variables. The loss probability is given in terms of the solution of this equa-