Some Asymptotic Results for the $M/M/\infty$ Queue with Ranked Servers

CHARLES KNESSL*  knessl@uic.edu

Department of Mathematics, Statistics and Computer Science (M/C 249), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045, USA

Received 6 November 2002; Revised 26 January 2004

Abstract. We consider an $M/M/\infty$ model with $m$ primary servers and infinitely many secondary ones. An arriving customer takes a primary server, if one is available. We derive integral representations for the joint steady state distribution of the number of occupied primary and secondary servers. Letting $\rho = \lambda/\mu$ be the ratio of arrival and service rates (all servers work at rate $\mu$), we study the joint distribution asymptotically for $\rho \to \infty$. We consider both $m = O(1)$ and $m$ scaled to be of the same order as $\rho$. We also give results for the marginal distribution of the number of secondary servers that are occupied.

Keywords: $M/M/\infty$ queue, asymptotic expansions, memory fragmentation, saddle point method

AMS subject classification: 41A60, 41A63, 30E20, 60K25, 60K30, 90B22

1. Introduction

We consider an infinite server queue with a Poisson arrival rate $\lambda$. The servers are divided into two classes. There are $m$ primary servers, and we can think of these as being ranked $1, 2, \ldots, m$. Then there are infinitely many secondary servers, ranked $m + 1, m + 2, \ldots$. Each server (primary or secondary) works at rate $\mu$ and we let $\rho = \lambda/\mu$ be the traffic intensity. An arriving customer is served by one of the $m$ primary servers, if one is available. Otherwise the customer is served by a secondary server.

This model has a variety of applications and has an extensive literature. For example, the servers can represent parking spaces near a restaurant, with the primary ones being near the restaurant and the secondary ones further away. If the spaces are ranked according to distance from the restaurant then an arriving customer will naturally take the lowest ranked space (if possible, the lowest ranked primary space).

Another application is in the fragmentation of computer memory. Here a server represents, say, a page of memory. Of particular interest here is to compute the amount of wasted space, which is defined as the difference between the highest ranked occupied page and the total number of occupied pages. Thus if, say, pages 1, 2, 4, 6 and 9 are used and those numbered 3, 5, 7, 8 and $\geq 10$ are not used, the wasted memory is $9 - 5 = 4$.

* Supported in part by NSF grants DMS-99-71656 and DMS-02-02815, and NSA grant MDA 904-03-1-0036.
In this application the “servers” are all identical, but breaking them up into two types proves useful for the mathematical analysis.

This model has been analyzed by many authors, including Kosten [9], Coffman et al. [4], Newell [10] and Aldous [2]. Of primary interest is the joint steady state distribution of the numbers of occupied primary and secondary servers. Denoting by $N_1$ and $N_2$ the number of primary and secondary servers occupied, we let $\pi(k, r) = \text{Prob}[N_1 = k, N_2 = r]$ be the steady state distribution, for $0 \leq k \leq m, r \geq 0$. We note that $N_1 + N_2$ corresponds to the total number of customers in an $M/M/\infty$ queue, and thus follows a Poisson distribution. The marginal distribution of $N_1$ follows a truncated Poisson distribution, as the number of occupied primary servers behaves as the classic Erlang loss model ($M/M/m/m$ queue). However, the marginal distribution of $N_2$ and the joint distribution are of a more complicated structure; the purpose of this paper is to better understand these.

Exact expressions for $\pi(k, r)$ appear in [4,9], but these are of a complicated form and do not yield much immediate insight. Here we obtain an alternate (integral) representation for $\pi(k, r)$. Then we study the joint distribution asymptotically for $\rho \to \infty$, which corresponds to “heavy traffic”. Our starting point for the asymptotic analysis is the two-dimensional generating function derived in [4], from which we obtain various integral forms for $\pi(k, r)$. It proves necessary to consider four cases of $m$:

1. $m = O(1)$,
2. $m \to \infty$ with $m/\rho \in (0, 1)$,
3. $m \to \infty$ with $m/\rho > 1$,
4. $m = \rho + O(\sqrt{\rho})$.

Within each case of $m$ it is furthermore necessary to consider different ranges of $k$ and $r$, where the asymptotic form of $\pi(k, r)$ is different. We write $\pi(k, r) = \pi(k, r; m, \rho)$ to emphasize the dependence on $m$ and $\rho$.

Previous work on the asymptotic analysis of this model for $\rho \to \infty$ includes [2,10]. In [10] Newell uses classical probabilistic arguments to obtain a variety of results, depending on the relative size of $m$ to $\rho$. In [2] Aldous uses more modern probabilistic techniques, approximating the various processes by simpler ones. In particular, the distribution of $N_1$ is approximated by an $M/M/1$ queue ($m/\rho < 1$), an exponential process ($m/\rho < 1, m - \rho \gg \sqrt{\rho}$), an Ornstein–Uhlenbeck process ($m = \rho + O(\sqrt{\rho})$), or an extremal process ($m \gg \rho + O(\sqrt{\rho})$).

The results we obtain are “global” in nature in that we treat the entirety of the lattice strip $\{0 \leq k \leq m, r \geq 0\}$, even ranges where the distribution is asymptotically small as $\rho \to \infty$. Also, we obtain explicit expressions for the critical scaling $m = \rho + O(\sqrt{\rho})$, which leads to a nontrivial approximation to $\pi(k, r)$ by a two-dimensional density function. This case is discussed in [10] using diffusion approximations and Mellin transforms, but the limiting density was not explicitly obtained.

Previously [6] we studied the functional $\sum_{k=0}^m \pi(k, 0; m, \rho)$ for $\rho \to \infty$ and the four ranges of $m$ indicated above. This quantity is important in estimating the amount of wasted memory in the computer storage model.

The paper is organized as follows. In section 2 we summarize the main results (cf. theorems 1–5). In theorem 1 we give integral representations for $\pi(k, r)$. The asymptotic expansions are summarized in theorems 2–5, for the four cases of $m$. In the corollaries that follow these theorems, we give the asymptotic expansions for the mar-