MAXIMUM QUEUE LENGTH AND WAITING TIME REVISITED: MULTISERVER G|G|c QUEUE

August 24, 1996

John S. Sadowsky*
School of Electrical Engineering
Purdue University
W. Lafayette, IN 47907
U.S.A.

Wojciech Szpankowski†
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

We characterize the probabilistic nature of the maximum queue length and the maximum waiting time in a *multiserver* $G|G|c$ queue. We assume a general i.i.d. interarrival process and a general i.i.d. service time process for each server with the possibility of having different service time distributions for different servers. Under a weak additional condition we will prove that the maximum queue length and waiting time grow asymptotically *in probability* as $\log_\omega n^{-1}$ and $\log n^{1/\theta}$, respectively, where $\omega < 1$ and $\theta > 0$ are parameters of the queueing system. Furthermore, it is shown that the maximum waiting time – when appropriately normalized – converges *in distribution* to the extreme distribution $\Lambda(x) = \exp(-e^{-x})$. The maximum queue length exhibits similar behavior, except that some oscillation caused by discrete nature of the queue length must be taken into account. The first results of this type were obtained for the $G|M|1$ queue by Heyde, and for the $G|G|1$ queue by Iglehart. Our analysis is similar to that of Heyde and Iglehart. The generalization to $c > 1$ servers is made possible due to the recent characterization of the tail of the stationary queue length and waiting time in a $G|G|c$ queue (cf. Sadowsky and Szpankowski [17]).

*This research was supported by the NSF Grant ECS-9003007.

†This research was supported by the NSF Grant CCR-8900305, in part by AFOSR Grant 90-0107, and in part by Grant R01 LM05118 from the National Library of Medicine.

1. INTRODUCTION

The $G|G|c$ queue is a single queue with an i.i.d. interarrival time process and $1 \leq c < \infty$ servers each having an i.i.d. service time process. This model occurs in numerous applications including industrial process modeling, multiprocessor computer systems, telecommunications networks and service counters. In some of these applications it is required that different servers work with different speeds, or even more generally, that different servers have different service time distributions. For example, in a (heterogeneous) multiprocessor system there are efficient (task oriented) processors and slower (general-purpose oriented) processors. When the service time distributions differ, we say the $G|G|c$ queueing system is *heterogeneous*. It is known (cf. Kiefer and Wolfowitz [9, 10], Loynes [11]) that such a system is stable if and only if the rate of the arrival of new customers is smaller than the total service rate. This paper investigate the maximum queue length and the maximum waiting time of a stable $G|G|c$ queue in its stationary mode of operation. We also give some partial results on the maximum total workload.

Some important information about dynamics of a system can be obtained by investigating the small tail of probabilities of large queue length and waiting time, or simply the maximum size of the queue over a period of time. Such information, without any doubt, has obvious significance to issues of resource allocation (e.g., the design of a buffer size in a distributed system). Moreover, such an investigation can be used to assess space complexity of other dynamic data structures that share common features with queues. We mention here dictionaries, linear lists, stacks, priority queues, symbol tables, hashing and so forth (cf. Szpankowski [19] and Aldous *et al* [1]).

The maximum queue length and the maximum waiting time were extensively studied in the 1970's. Heyde [7] was the first who predicted the asymptotic growth of the maximum queue length in a $G|M|1$ system. Iglehart [8] continued this investigation by providing the rate of growth and the limiting law for the maximum waiting time in $G|G|1$. The maximum queue length – as shown by Anderson [2] – does not possess limiting distribution due to some oscillation caused by the discrete nature of the queue length. Nevertheless, this oscillation can be taken into account, and Anderson [2] derived the asymptotic behavior of the maximum queue length. These results are obtained as a consequence of the exponential (resp. geometric) tail distribution for the waiting time (resp. queue length) due to Feller [4], and Iglehart [8] who derived the tail distribution of the maximum waiting time in a busy period. Recently, we have obtained a tail characterization for the waiting time and queue length distributions in the *multiserver* $G|G|c$ queue. More importantly for the present application, we have characterized the distribution tails for the maximum waiting time and queue length over a stationary full busy period (to be defined below) [17]. These results will play the same role as Iglehart's result for the maximum waiting time in a $G|G|1$ busy period.

We note that Neuts and Takahashi [12] have also characterized the stationary queue length and waiting time distribution tails for the $G|PH|c$ queue. However, their analysis is not directly related to busy-idle cycles, and as a result, their results are not directly applicable to the analysis of Anderson [2] and Iglehart [8].

This paper is organized as follows. In the next section we present a summary of our results from [17] (see also [16]), as well as some important extensions of them that are directly applicable to the maximum size of $G|G|c$. In Section 3 we present our main results. In particular, after discussing one general result on the maximum order statistic, we show the growth *in probability* of the maximum queue length, the maximum waiting time and the maximum total workload. Finally, we extend these results to the convergence *in distribution.*

Throughout the paper we assume a *homogeneous* $G|G|c$ queue for simplicity of presentation, however – as discussed in Remarks 2.1 and 3.5 – extension to heterogeneous case is straightforward using the constructions of [17].

2. PRELIMINARIES

We consider a $G|G|c$ queue with $1 \leq c < \infty$ servers, and general interarrival times and service times distributions. The interarrival time process is denoted $\{A_k\}$, and the service time process for the i'th server is denoted $\{B_j^{(i)}\}$. The processes $\{A_k\}$ and $\{B_j^{(i)}\}$, $i = 1,..,c$, are independent and i.i.d. with distribution functions $A(t) = \mathcal{P}(A_k \leq t)$ and $B(t) = \mathcal{P}(B_j^{(i)} \leq t)$ (which does not depend on the server index i for a homogeneous queueing system). The Laplace-Stieltjes Transforms (LST) are $A^*(s) = E[\exp(-sA_k)]$ and $B^*(s) = E[\exp(-sB_j^{(i)})]$. To avoid trivial cases we also assume throughout that $A(0) < 1$ and $B(0) < 1$. For waiting time analysis, the service discipline is FIFO (first in – first out), and work-conserving (that is, a server cannot stays idle if there is a job in the queue). Of course, queue length does not depend on service disciplines.

We denote the queue length at the instant of arrival of the k'th customer as Q_k. The queue length Q_k does *not* include customers in service. The definition of the waiting time for multiserver queues is a little more involved. W_k will denote the waiting time of the k'th customer, not including service time. A FIFO queueing system can be thought of as c parallel queues, one for each server. Let $W_k^{(i)}$ denote the waiting time that would be experienced by the k'th customer if it were assigned to the i'th queue. Then the FIFO service priority is equivalent to assignment of the k'th job to the queue having the minimal waiting time, and hence, $W_k = \min\{W_k^{(1)}, ..., W_k^{(c)}\}$. (We assume some deterministic or random assignment rule for the case of ties.) It will be convenient to denote the c waiting times as a vector $\mathbf{W}_k = (W_k^{(1)}, ..., W_k^{(c)})$. Define $\rho = \lambda/(c\mu)$ where $\lambda = \mathrm{E}[A_k]^{-1}$ and $\mu = \mathrm{E}[B_j^{(i)}]^{-1}$ for homogeneous $G|G|c$ queue. It is well known that the system is *stable* if and only if $\rho < 1$ (cf. [9, 11]). If $\rho < 1$, then regardless of the initial state of the system $\mathbf{W}_k$ and Q_k have unique stationary (limiting) distributions. Q_∞ and W_∞ will denote random variables that are distributed according to the stationary distributions of Q_k and W_k. Likewise, $\mathbf{W}_\infty$ will denote a random vector which is distributed according to the stationary distribution of the waiting time vector $\mathbf{W}_k$. Throughout the paper we shall assume that $\rho < 1$ and the waiting time W_k as well as the queue length Q_k are stationary processes.

Our interest is in estimating a probabilistic behavior of the maximum queue length Q_n^{max} and the maximum waiting time W_n^{max} attained by the time the n'th customer has arrived, that is,

$$Q_n^{max} = \max_{1 \le k \le n} \{Q_k\} \qquad \text{and} \qquad W_n^{max} = \max_{1 \le k \le n} \{W_k\} .$$

Our analysis follows that of Heyde [7] and Iglehart [8] for the $c = 1$ server case which we briefly review here. As is very well known, the queueing process regenerates when the entire system empties out and successive busy periods are i.i.d. Let L_n denote the number of busy periods completed prior to the n-th arrival. Busy periods are independent, and the expected length (number of customers) of a busy period is finite [10]. Hence, from renewal theory $L_n/n \to \alpha$ (a.s.) for some $\alpha > 0$. Let $\overline{Q}_\ell$ and $\overline{W}_\ell$ denote the maximum queue length and the maximum waiting time in the ℓ'th busy period. Then, we have

$$\max_{1 \le \ell \le L_n} \{\overline{Q}_\ell\} \le Q_n^{max} \le \max_{1 \le \ell \le L_n + 1} \{\overline{Q}_\ell\} \tag{1}$$

and

$$\max_{1 \le \ell \le L_n} \{\overline{W}_\ell\} \le W_n^{max} \le \max_{1 \le \ell \le L_n + 1} \{\overline{W}_\ell\} . \tag{2}$$

The busy period maximums $\overline{Q}_\ell$ and $\overline{W}_\ell$, $\ell = 1, 2, ...$, are i.i.d. random variables. Therefore, knowing the tail distributions of $\overline{Q}_\ell$ and $\overline{W}_\ell$ we can apply standard approach of the extreme statistics for independent random variables (cf. Galambos [5], Gniedenko [6]), and obtain the limiting distribution of the maximum queue length and the maximum waiting time. The maximum queue length needs some additional care since some oscillations can occur due to discretization (cf. Anderson [2]).

In order to apply the ideas of the previous paragraph to a *multiserver* queue (which is our contribution here), we need two results. First, we will require a sufficiently detailed estimate of tail probabilities for $\overline{Q}_\ell$ and $\overline{W}_\ell$. We have recently obtained such an estimate in [17]. Second, we require a regeneration structure. As in [7] and [8], we will appeal to the regeneration that occurs due to busy/idle cycles, but to do this we will have to be careful about the definition of such cycles.

In a multiserver queue, a *full busy period* is a maximal contiguous time interval during which *all* servers are continuously busy. A *partial busy period* is a maximal contiguous time interval during which *at least* one server is busy. Full busy periods are separated by *partial idle periods* which are maximal contiguous time intervals during which there is always at least one idle server. Conversely, partial busy periods are separated by *full idle periods* which are maximal contiguous time intervals during which there is all servers are idle. Notice that in the $c = 1$ case partial and full busy periods are the same thing. A *busy cycle* is defined as a partial busy period followed by a full idle period. This conventional definition has the advantage that successive cycles are i.i.d. However, in the multiserver case, these cycles do not necessarily occur i.o. (infinitely often). *Regeneration by partial busy period / full idle period cycles must be assumed.* We will refer to the shorter cycles consisting of a full busy period followed by a partial idle period as *c-cycle*. These are not i.i.d. but they do form a Markov chain.

The following hypothesis is required to ensure the existence of both cycles and c-cycles.

(R) *Assume that* $\rho < 1$, $0 < \mathcal{P}(\mathbf{W}_\infty = \mathbf{0}) < 1$, *and* $\mathcal{P}($*exactly one* $W_\infty^{(i)} = 0) > 0$.

The inequality $\mathcal{P}(\mathbf{W}_\infty = \mathbf{0}) < 1$ rules out the trivial case that queue is always empty when new customers arrive. This occurs when there is a constant M such that $B_j^{(i)} \leq M$ and $A_k > M$ almost surely. The inequalities $0 < \mathcal{P}(\mathbf{W}_\infty = \mathbf{0}) < 1$ together are equivalent to $\mathcal{P}($*infinitely many distinct full idle periods*$) = 1$ by the ergodicity of the queue (cf. [9]). As noted above, infinitely many cycles is not automatic. For example, when $c > 1$, it is possible to have a constant $M > 0$ such that $A_k < M$ and $B_j^{(i)} > M$ almost surely, and still have $\rho < 1$. However, in this case there will always be at least one server busy at all times, and hence, full idle periods never occur. Whitt [20] gives some sufficient conditions that insures infinitely many full idle periods, in particular, $\mathcal{P}(A_k - B_j^{(i)} > 0) > 0$ is sufficient. The inequality $\mathcal{P}($*exactly one* $W_\infty^{(i)} = 0) > 0$ is equivalent to $\mathcal{P}($*infinitely many distinct full busy periods*$) = 1$. Again, this condition is also not automatic. For example, if $c > 2$, $B_j^{(i)} < M$ and $A_k > (c-1)M$ almost surely for some constant M, then there will always be at least $c - 1$ idle servers when a new customer arrives. However, $\mathcal{P}($*exactly one* $W_\infty^{(i)} = 0) > 0$ is a less significant hypothesis than the other inequalities in (R) because it simply rules out the trivial cases that $Q_k \equiv 0$.

Naturally, when a full idle period occurs the system is empty and the queueing process restarts with the arrival of the next customer. That is, full idle periods are regeneration events and successive busy cycles are i.i.d. Hence, we refer to assumption (R) as the *regeneration hypothesis.*

Assume (R) and define $\tilde{\mathbf{B}}_m = (\tilde{B}_m^{(1)}, ..., \tilde{B}_m^{(c)})$ as the c-dimensional vector representing the residual service times for the customers being processed at the beginning of the $(m+1)$'th c-cycle. Notice that $\tilde{\mathbf{B}}_m = \mathbf{W}_k$ when k is the index of the customer that initiates the $(m+1)$'th c-cycle, and hence, $\xi(\cdot) = \mathcal{P}(\mathbf{W}_\infty \in \cdot \mid \textit{exactly one } W_\infty^{(i)} = 0)$ is the stationary distribution of $\tilde{\mathbf{B}}_m$. (Notice that this conditional probability is well defined under (R).) It turns out that $\{\tilde{\mathbf{B}}_m\}$ is a Markov chain, hence $c-cylces$ form a Markov chain too [15, 17]. This property is strong enough to obtain a full characterization of the asymptotic behavior the maximum queue length in a $c-cycle$.

Define $\overline{Q}_m$ and $\overline{\overline{Q}}_\ell$ (resp. $\overline{W}_m$ and $\overline{\overline{W}}_\ell$) as the maximum queue length (resp. waiting time) in the m'th $c-cycle$ and ℓ'th busy cycle respectively. Furthermore, under assumption (R) we note that (1) and (2) hold if $\overline{Q}_m$ (resp. $\overline{W}_m$) is replaced by $\overline{\overline{Q}}_\ell$ (resp. $\overline{\overline{W}}_\ell$).

Now we are ready to summarize results of Sadowsky and Szpankowski [17]. In general, define

$$\theta \quad = \quad \sup\{s \le \bar{s} : A^*(s)B^*(-s/c) \le 1\}, \tag{3}$$

where $\bar{s} = \sup\{s : B^*(s) < \infty\}$. Furthermore, we define

$$\omega \quad = \quad A^*(\theta) \ . \tag{4}$$

Under some additional regularity, it turns out that θ is the unique positive solutions of the *characteristic equation*

$$A^*(\theta)B^*(-\theta/c) \quad = \quad 1 \tag{5}$$

The reader is referred to [17] (see also [16]) for a detailed presentation of the properties of the characteristic equation, or more generally (3), for the heterogeneous queue. For some results we require an additional hypothesis:

(E) *$\theta > 0$ satisfies (5), and* $\frac{d}{ds}B^*(s) \mid_{s=\theta} = \mathrm{E}[B_k \exp(-\theta B_k)] < \infty$.

Let $\overline{Q}$ and $\overline{W}$ denote the maximum queue length and the maximum waiting time respectively in a full busy period that starts with residual service time vector $\tilde{\mathbf{B}}_0$ having the stationary distribution $\xi(\cdot)$. In Sadowsky and Szpankowski [17] the following results are proved.

Theorem 1. (i) *Assume $\rho < 1$. Then*

$$\log\left(\mathcal{P}(Q_\infty \geq n)\right) \;\sim\; \log(\omega^n) \quad and \quad \log\left(\mathcal{P}_\xi\left(\overline{Q} \geq n\right)\right) \;\sim\; \log(\omega^n). \tag{6}$$

(ii) In addition assume (E) and the service times distribution $B(t)$ is spread-out.[1] *Then there exists a constants $K_{\overline{Q}}, K_Q$ such that*

$$\mathcal{P}\left(Q_\infty \geq n\right) \;\sim\; K_Q\,\omega^n. \quad and \quad \mathcal{P}_\xi\left(\overline{Q} \geq n\right) \;\sim\; K_{\overline{Q}}\,\omega^n \tag{7}$$

where $0 < K_{\overline{Q}}, K_Q < \infty$. ■

Theorem 2. (i) *Assume $\rho < 1$ and FIFO queueing discipline. Then*

$$\log\left(\mathcal{P}(W_\infty \geq w)\right) \;\sim\; -\theta w \quad and \quad \log\left(\mathcal{P}_\xi\left(\overline{W} \geq w\right)\right) \;\sim\; -\theta w. \tag{8}$$

(ii) In addition assume (E), the service times distribution $B(t)$ is spread-out, *and $A(t)$ is non-atomic. Then there exists a constants $K_{\overline{W}}, K_W$ such that*

$$\mathcal{P}\left(W_\infty \geq w\right) \;\sim\; K_W\,e^{-\theta w} \quad and \quad \mathcal{P}_\xi\left(\overline{W} \geq w\right) \;\sim\; K_{\overline{W}}\,e^{-\theta w}. \tag{9}$$

where $0 < K_{\overline{W}}, K_W < \infty$. ■

For the purpose of this paper we need an extension of Theorems 1 and 2, which deals with *partial* busy period maximum queue length an waiting time. Let $\overline{\overline{Q}}$ and $\overline{\overline{W}}$ denote the maximum queue length and waiting time over a partial busy period (i.e., busy cycle).

Corollary 3. *Let appropriate hypotheses of Theorem 1(i) and 2(i) hold, and in addition we adopt assumption (R). Then,*

$$\log\left(\mathcal{P}(\overline{\overline{Q}} \geq n)\right) \;\sim\; \log(\omega^n) \quad \text{and} \quad \log\left(\mathcal{P}(\overline{\overline{W}} \geq w)\right) \;\sim\; -\theta w \tag{10}$$

where $\overline{\overline{Q}}$ and $\overline{\overline{W}}$ represent the maximum queue length and the maximum waiting time in a busy cycle. Assume in addition the appropriate hypothesis of Theorem 1(ii) and Theorem 2(ii). Then,

$$\mathcal{P}\left(\overline{\overline{Q}} \geq n\right) \;\sim\; K_{\overline{\overline{Q}}}\,\omega^n \quad \text{and} \quad \mathcal{P}\left(\overline{\overline{W}} \geq w\right) \;\sim\; K_{\overline{\overline{W}}}\,e^{-\theta w} \tag{11}$$

where $0 < K_{\overline{\overline{Q}}}\ ,\ K_{\overline{\overline{W}}} < \infty$.

[1] A distribution is *spread-out* if some convolution power has a component that is absolutely continuous with respect to Lebesgue measure.

Proof: We prove only the result (11) for the queue length. As in Sadowsky and Szpankowski [17], let $\mathbf{C}_m = (\tilde{\mathbf{B}}_{m-1}, \mathbf{X}_m)$ denote the m'th c-cycle Markov chain where $\mathbf{X}_m$ is a random element that contains all of the service times and interarrival times for customers that arrive during the m'th c-cycle. Then $\tilde{\mathbf{B}}_m$ is determined by $\mathbf{C}_{m-1}$, and $\{\mathbf{C}_m\}$ is a regenerative positive recurrent Markov chain under hypothesis (R). Define $E_m = \{$ *no full idle periods before the m'th full busy period* $\}$. The stationary distribution for the c-cycle chain $\{\mathbf{C}_m\}$ is $\mathcal{P}_\xi(\mathbf{C}_1 \in \cdot)$. Let $\nu(\cdot) = \mathcal{P}_{\mathbf{b}}(\tilde{\mathbf{B}}_1 \in \cdot \,|$ *regeneration in* $\mathbf{C}_1)$ (which does not depend on the initial value $\tilde{\mathbf{B}}_0 = \mathbf{b}$). Then

$$\mathcal{P}_\xi(\mathbf{C}_1 \in \cdot) \quad = \quad \frac{\sum_{m=1}^\infty \mathcal{P}_\nu(\mathbf{C}_m \in \cdot\, ; E_m)}{\sum_{m=1}^\infty \mathcal{P}_\nu(E_m)}. \tag{12}$$

The above representation is easily verified to be the unique invariant, hence, the stationary measure. See also Theorem 5.2 in [13]. Define $F_{m,n} = \{$ $\overline{Q}_k < n$ *for all* $k < m$ $\}$. Then,

$$\begin{aligned}
\mathcal{P}_\nu\left(\overline{\overline{Q}} \geq n\right) &= \mathcal{P}_\nu\left(\overline{Q}_1 \geq n\right) + \sum_{m=2}^\infty \mathcal{P}_\nu\left(\overline{Q}_m \geq n; E_m \cap F_{m,n}\right) \\
&= \sum_{m=1}^\infty \mathcal{P}_\nu\left(\overline{Q}_m \geq n\, ;\, E_m\right) - \sum_{m=2}^\infty \mathcal{P}_\nu\left(\overline{Q}_m \geq n; E_m \cap F^c_{m,n}\right).
\end{aligned}$$

Applying (12) to the first term in the last line above, we conclude that

$$\mathcal{P}_\nu\left(\overline{\overline{Q}} \geq n\right) \quad = \quad \left[\sum_{m=1}^\infty \mathcal{P}_\nu(E_m)\right] \mathcal{P}_\xi\left(\overline{Q}_1 \geq n\right) - \sum_{m=2}^\infty \mathcal{P}_\nu\left(\overline{Q}_m \geq n; E_m \cap F^c_{m,n}\right).$$

The first term in the last line above is $\sim K_{\overline{\overline{Q}}}\omega^n$ where $K_{\overline{\overline{Q}}} = K_{\overline{Q}} \sum_{m=1}^\infty \mathcal{P}_\nu(E_m)$, by Theorem 1. We will now show that the second term in the last line above is $o(\omega^n)$ as $n \to \infty$. It follows from the proof of Theorem 2.1 in [17] that $\mathcal{P}(\,\overline{Q}_m \geq n\,|\,\tilde{\mathbf{B}}_{m-1} = \mathbf{b}\,) \leq \exp(\frac{\theta}{c}\sum_{i=1}^c b^{(i)})\,\omega^n$. Thus, for $m \geq 2$ we have

$$\begin{aligned}
&\mathcal{P}_\nu\left(\overline{Q}_m \geq n\, ;\, E_m \cap F^c_{m,n}\right) \\
&\qquad \leq \int \mathcal{P}(\,\overline{Q}_m \geq n\,|\,\tilde{\mathbf{B}}_{m-1} = \mathbf{b}\,)\,\mathcal{P}_\nu\left(E_m \cap F^c_{m,n}\, ;\, \tilde{\mathbf{B}}_{m-1} \in d\mathbf{b}\right) \\
&\qquad \leq \mathrm{E}_\nu\left[\exp\left(\frac{\theta}{c}\sum_{i=1}^c B^{(i)}_{m-1}\right)\, ;\, E_m \cap F^c_{m,n}\right]\omega^n.
\end{aligned}$$

Notice that $\mathcal{P}_\nu\left(E_m \cap F^c_{m,n}\right) \to 0$ for each m as $n \to \infty$, hence, for each m the expectation in the last line above vanishes $n \to \infty$. Moreover,

$$\sum_{m=2}^\infty \mathrm{E}_\nu\left[\exp\left(\frac{\theta}{c}\sum_{i=1}^c B^{(i)}_{m-1}\right)\, ;\, E_m \cap F^c_{m,n}\right]$$

$$\leq \sum_{m=1}^{\infty} \mathrm{E}_{\nu}\left[\exp\left(\frac{\theta}{c}\sum_{i=1}^{c} B_{m-1}^{(i)}\right) ; E_m\right]$$
$$= \left[\sum_{m=1}^{\infty} \mathcal{P}_{\nu}(E_m)\right] \mathrm{E}_{\xi}\left[\exp\left(\frac{\theta}{c}\sum_{i=1}^{c} B_{0}^{(i)}\right)\right] \quad < \quad \infty$$

where the convergence of this upper bound is proved in Lemma 4.8 in [17]. Thus, by the dominated convergence theorem we have

$$\sum_{m=2}^{\infty} \mathcal{P}_{\nu}\left(\overline{Q}_m \geq n; E_m \cap F_{m,n}^c\right)$$
$$\leq \left(\sum_{m=2}^{\infty} \mathrm{E}_{\nu}\left[\exp\left(\frac{\theta}{c}\sum_{i=1}^{c} B_{m-1}^{(i)}\right) ; E_m \cap F_{m,n}^c\right]\right) \omega^n \quad = \quad o(1)\,\omega^n$$

and this completes the proof. ■

Another variable of interest in some applications is the *total workload* $U_k = W_k^{(1)} + W_k^{(2)} + ... + W_k^{(c)}$. It is quite likely that result analogous to Theorem 1 and Theorem 2 can be proved using the methods of [17], but we shall present some more restricted results here.

Consider a slight generalization of the workload definition. Let $U_k = \sum_{j=1}^{Q_k} C_j^{(k)}$ where the $C_j^{(k)}$'s are i.i.d. random that are independent of Q_k. In particular, if $C_j^{(k)}$'s are the service times of the jobs in queue at the instant that customer k arrives, then U_k is precisely the total workload defined above. Another example occurs in computer system analysis. The $C_j^{(k)}$'s might represent the memory requirement for computer jobs in queue. We shall prove an asymptotic result for the stationary total workload U_∞.

Corollary 4. *Assume hypothesis of Theorem 1(ii) is satisfied and that the queue is operating under its stationary distribution. Let the $C_j^{(k)}$'s be i.i.d. random variables independent of Q_k. Let $C^*(s) = \mathrm{E}[\exp(-sC_j^{(k)})]$ denote the LST of the $C_j^{(k)}$'s, and we assume it is finite in some neighborhood of zero. Define s^* as a unique positive solution of the following equation $C^*(-s^*) = \omega^{-1}$. Then*

$$\mathcal{P}(U_\infty \geq u) \quad \sim \quad K_U\, e^{-s^* u} \tag{13}$$

as $u \to \infty$ for some constant $K_U \in (0, \infty)$.

Proof. Under stationary operation, let $Q^*(z) = \mathrm{E}[z^{Q_k}] = \mathrm{E}[z^{Q_\infty}]$ be the generating function of the stationary queue length distribution. Then clearly, $U^*(s) = \mathrm{E}[\exp(-sU_k)] = Q^*(C^*(s))$. An abelian theorem (cf. Postnikov [14]) together with Theorem 1(ii) imply that

$$Q^*(z) \quad \sim \quad 1 - \frac{(1-z)K_Q}{1-\omega z}$$

as $z \to \omega^{-1}$. Thus, as $s \downarrow -s^*$, we have

$$U^*(s) \;=\; Q(C^*(s)) \;\sim\; 1 - \frac{(1-C^*(s))K_Q}{1-\omega C^*(s)} \;\sim\; 1 - \frac{(1-C^*(s))K_Q}{C^{*\prime}(-s^*)(s+s^*)\omega} \; . \tag{14}$$

To obtain the tail of U from (14) we use a tauberian theorem. This needs some care. Fortunately, according to our basic assumptions the *average* value of the total total workload is finite, and this implies that $\mathcal{P}\{U > t\} = o(1/t)$. Hence we can apply Hardy and Littlewood's theorem (cf. Postnikov [14]) to (14), and this completes the proof. ■

Remark 2.1. In Corollary 4, if U_k is the total workload, that is, $C^*(s) = B^*(s)$, then by (5) it follows that $s^* = \theta/c$.

Remark 2.2. Theorems 1 and 2, as well as their extension Corollary 3, hold in fact under more general assumptions, namely for *heterogeneous* $G|G|c$ queues. In such a system there are c sequences of service times, each one associated with different server (e.g., servers might have different speeds). Let $\{B_k^{(i)}\}$ denote the service time required by the jth customer processed by server i, and $B_i^*(s_i) = \mathrm{E}[\exp(-s_i B_j^{(i)})]$ is the LST of $\{B_j^{(i)}\}$. To formulate our results in such a situation, we need to generalize the characteristic equation (5). This is done by Sadowsky and Szpankowski [17]. We briefly sketch this generalization here. For a fixed p define a vector $s_i(p)$, $i = 1, \ldots, c$ such that $\sum_{i=1}^{c} s_i(p) = p$ and $B_i^*(s_i(p)) = B_1^*(s_1(p))$. Then, under mild assumptions (for details see [17]) $s_i(p)$ is a function of $s_1(p)$ such that on the curve $s_1(p)$ the following holds $B_i^*(s_i(p)) = B_1^*(s_1(p))$. Then, the characteristic equation (5) becomes

$$A^*(\theta) B_1^*(s_1(\theta)) = 1 \; . \tag{15}$$

If all of the LSTs of $B_i^*(s)$ are defined on the same region, then Theorems 1 and 2, and Corollaries 3 and 4, hold with θ defined as in (15) provided assumption (E) is satisfied. For "logarithmic" results (Theorems 1(i) and 2(i)) the characteristic equation (15) should be replaced by a weaker form as in (4), that is,

$$\theta = \sup\{p : A^*(p) B_1^*(s_1(p)) \le 1\} \; .$$

Note that in the homogeneous case, $s_1(p) = p/c$ as needed to transform (15) into (5).

3. MAIN RESULTS

In this section we present our main results regarding the maximum queue length Q_n^{max}, the maximum waiting time W_n^{max}, and the maximum total workload U_n^{max}.

Many of the results stated here follow directly from well know results on the maximum of a set of i.i.d. random variables. For example, see Galambos [5]. We include some proofs here for completeness.

We discuss only the queue length problem. The reasoning for maximum waiting time and total workload are obviously analogous to our queue length arguments. Note that

$$\max_{1\le \ell \le L_n} \{\overline{\overline{Q}}_\ell\} \;\le\; Q_n^{max} \;=\; \max_{1\le k\le n}\{Q_k\} \;\le\; \max_{1\le \ell\le L_n+1}\{\overline{\overline{Q}}_\ell\}\ , \tag{16}$$

where (assuming (R)) L_n denotes the number of *busy cycles* completed prior to the nth arrival. By the ergodicity of the queueing process, $L_n/n \to \alpha$ (a.s) for some $\alpha \in (0,1]$.

Lemma 5. *Let $\{X_k\}$ be an i.i.d. sequence of random variables with common distribution function $F(\cdot)$. Assume that for some constant $\beta \in (0,\infty)$ we have $\log(1-F(x)) \sim -\beta x$ as $x\to\infty$. Let $\{L_n\}$ be a sequence of random variables such that $L_n/n \to \alpha \in (0,\infty)$ (pr.) and define $M_n = \max_{1\le k\le L_n} X_k$. Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers such that $a_n - \beta^{-1}\log(n\alpha) \to -\infty$ and $b_n - \beta^{-1}\log(n\alpha)\to +\infty$. Then $\mathcal{P}(a_n \le M_n \le b_n) \;\to\; 1$.*

Proof. For a fixed $\delta > 0$, define $\underline{M}_n = \max_{1\le k\le (1-\delta)\alpha n} X_k$ and $\overline{M}_n = \max_{1\le k\le(1+\delta)\alpha n} X_k$. We first have $\mathcal{P}(M_n > b_n) \le \mathcal{P}(\overline{M}_n > b_n) + \mathcal{P}(L_n > (1+\delta)\alpha n)$. Since $L_n/n \to \alpha$ (pr.), we only need to show that $\mathcal{P}(\overline{M}_n > b_n) \to 0$. By Boole's inequality, $\mathcal{P}(\overline{M}_n > b_n) \le (1+\delta)\alpha n(1-F(b_n))$. Thus,

$$\begin{aligned}\log(\mathcal{P}(\overline{M}_n > b_n)) &\le \log(1+\delta) + \log(1-F(b_n)) + \log(n\alpha)\\ &\sim -\beta b_n + \log(n\alpha) \quad\to\quad -\infty\end{aligned}$$

and this implies $\mathcal{P}(M_n > b_n) \to 0$ by the condition on the sequence $\{b_n\}$. Next we have $\mathcal{P}(M_n < a_n) \le \mathcal{P}(\underline{M}_n \le a_n) + \mathcal{P}(L_n < (1-\delta)\alpha n)$ and again it is clear that we only need to show that $\mathcal{P}(\underline{M}_n \le a_n) \to 0$. Using the independence of the X_k's we have $\mathcal{P}(\underline{M}_n \le a_n) \;=\; F(a_n)^{\lfloor(1-\delta)\alpha n\rfloor}$. Using $\log(1+x) \le x$ we have

$$\begin{aligned}-\log(\mathcal{P}(\underline{M}_n \le a_n)) &= -\lfloor(1-\delta)\alpha n\rfloor\,\log\left(1-(1-F(a_n))\right)\\ &\ge \lfloor(1-\delta)\alpha n\rfloor\,(1-F(a_n)),\end{aligned}$$

and hence,

$$\log(-\log(\mathcal{P}(\underline{M}_n \le a_n))) \;\ge\; \log(1-F(a_n)) \;+\; \log(\alpha n) \;+\; \log(\lfloor(1-\delta)\rfloor).$$

By the assumption on the sequence $\{a_n\}$, $\log(1-F(a_n)) + \log(n\alpha) \to +\infty$. This implies that $-\log(\mathcal{P}(\underline{M}_n < a_n)) \to +\infty$, and hence, $\mathcal{P}(\underline{M}_n < a_n) \to 0$. ■

As an immediate consequence of (16), Lemma 5 and part (i) of Corollary 3 we have the following result.

Corollary 6. *Assume for stationary queue ($\rho < 1$) that (R) holds, and there exists a positive solution, $\theta > 0$, of (3).*

(i) *For any sequences of numbers $\{a_n\}$ and $\{b_n\}$ such that $a_n - \log_\omega(\alpha n) \to -\infty$ and $b_n - \log_\omega(\alpha n) \to +\infty$ we have $\mathcal{P}(a_n \leq Q_n^{max} \leq b_n) \to 1$, and hence, $Q_n^{max}/\log_\omega(\alpha n) \to 1$ (pr.).*

(ii) *For any sequence of numbers $\{a_n\}$ and $\{b_n\}$ such that $a_n - \theta^{-1}\log(\alpha n) \to -\infty$ and $b_n - \theta^{-1}\log(\alpha n) \to +\infty$ we have $\mathcal{P}(a_n \leq W_n^{max} \leq b_n) \to 1$, and hence, $\theta W_n^{max}/\log(\alpha n) \to 1$ (pr.).* ■

Remark 3.1. The assumption $\theta > 0$ is important. It is easy to see that for heavy tail service time distribution (e.g., $1 - B(t) \sim 1/t^2$), one can construct a stable queueing system for which $\theta = 0$. Then, the tail of the queue length decays slower than geometric, and consequently the maximum queue length may grow faster than logarithmic.

Remark 3.2. Our results cannot be extended to $c = \infty$ as the $M|G|\infty$ example shows. Indeed, in this case the stationary distribution is subexponential, that is, more precisely $\mathcal{P}\{Q_\infty \geq n\} \sim e^{-\rho}\rho^n/n!$ (cf. Wolff [21]). In this case, we can prove that $Q_n^{max} \sim \log n/(\log\log n)$ (pr.) (cf. Aldous *et al* [1]).

Remark 3.3. How long one must wait until the asymptotics for the maximum queue length and waiting time become valid? Naturally this depends on ρ. For example, for $\rho = 1$ the growth of Q_n^{max} is almost linear (cf. Serfozo [18]). However, when $\rho \to 0$ the growth is much slower. Consider – as an example – the case when $n = \omega^{-1/\rho}$. Then, the rate of the convergence is exponential. In practice one requires the exponential rate of convergence, but then n must increase exponentially fast in $1/\rho$ for the asymptotics to be valid. Hence, one must wait "exponential time" before the maximum queue reaches its value $O(\log n)$ predicted by Corollary 6. For practical applications, it might be much sensible to consider (the time of observation) n being at most polynomially large in $1/\rho$.

Remark 3.4. If additionally we assume (E) in Corollary 6, then one can characterize the rate of convergence. For example, a simple modification of Lemma 5 leads to the following estimates

$$\begin{aligned}
\mathcal{P}((1-\varepsilon)\log_\omega(n\alpha)^{-1} \leq Q_n^{max} \leq (1+\varepsilon)\log_\omega(n\alpha)^{-1}) &= 1 - O(n^{-\varepsilon}) \\
\mathcal{P}((1-\varepsilon)\log(n\alpha)^{1/\theta} \leq W_n^{max} \leq (1+\varepsilon)\log(n\alpha)^{1/\theta}) &= 1 - O(n^{-\varepsilon}) .
\end{aligned}$$

A similar result to the one presented in Corollary 6, can be obtained for the generalized total workload U_n. However, since we need slightly different approach to prove it, we present it separately in the following theorem.

Theorem 7. *Assume hypotheses of Corollary 6 together with (E). Then,* $s^* U_n^{max}/\log n \to 1$ *(pr.).*

Proof. For an upper bound we use $U_n^{max} = \max_{1\le k\le n} U_k$ and Corollary 4. Then, by Boole's inequality we have

$$\mathcal{P}(U_n^{max} \le (1+\varepsilon)\frac{1}{s^*}\log n) \le n\mathcal{P}(U_k \le (1+\varepsilon)\frac{1}{s^*}\log n) \sim \frac{1}{n^\varepsilon} \ .$$

For the lower bound we note that $U_n^{max} \ge \max_{1\le k\le L_n} \overline{U}_k$ where $\overline{U}_k$ is the maximum generalized workload in a busy period. But, we can bound it from the below by the following

$$\overline{U}_k \ge \sum_{j=1}^{\overline{\overline{Q}}_k} C_j^{(k)} = \widetilde{U}_k \ .$$

Using the same approach as in the proof of Corollary 4 we can show that $\mathcal{P}\{\widetilde{U}_k \ge u\} \sim K_{\widetilde{U}} e^{-s^* u}$. Since $\widetilde{U}_k$ are i.i.d. with exponential tail, then by Lemma 5 $s^*\widetilde{U}_k/\log n \to 1$ (pr.), and this, together with the upper bound proved above, establishes the theorem. ■

Finally, we present our strongest results regarding convergence *in distribution* of the maximum waiting time and the maximum queue length.

Theorem 8. *Let* $\rho < 1$ *with* $c < \infty$*, and assumptions (R) and (E) hold together with hypotheses of Theorem 1(ii) and Theorem 2(ii). Then,*

$$\lim_{n\to\infty} \mathcal{P}(\theta W_n^{max} < x + \log(nK_{\overline{\overline{W}}})) = \exp(-\alpha e^{-x}) \tag{17}$$

for every nonnegative real x*. Furthermore, the maximum queue length behaves for large* n *as*

$$\lim_{n\to\infty} \max_x \mid \mathcal{P}(Q_n^{max} < x) - \exp(-nK_{\overline{\overline{Q}}}\, \alpha\omega^x) \mid = 0 \ , \tag{18}$$

or in another form

$$\begin{aligned} \exp(-\alpha\omega^{m-1}) \ &\le \ \lim_{n\to\infty} \inf \mathcal{P}(Q_n^{max} < m - \log_\omega(nK_{\overline{\overline{Q}}})) \\ &\le \ \lim_{n\to\infty} \sup \mathcal{P}(Q_n^{max} < m - \log_\omega(nK_{\overline{\overline{Q}}})) \le \exp(-\alpha\omega^m) \ , \end{aligned} \tag{19}$$

where m *is an integer.*

Proof. The proof is standard and along the lines of Iglehart's proof of $G|G|1$ results. For example, for the maximum waiting time we first consider fixed number, say N, of busy periods, and apply Corollary 3 to (16) in order to obtain

$$\mathcal{P}(W_N^{max} \leq (x + \log(NK_{\overline{\overline{W}}}))/\theta) = \mathcal{P}^N(\overline{\overline{W}} \leq (x + \log(NK_{\overline{\overline{W}}}))/\theta) \qquad =$$
$$= \Big(1 - K_{\overline{\overline{W}}} \exp(-x - \log(NK_{\overline{\overline{W}}}) + o(\exp(-x - \log(NK_{\overline{\overline{W}}}))\Big)^N \quad \to \exp(-e^{-x}) \ . \quad (20)$$

Now, to prove (17) it is enough to make N random such that $N/n \to \alpha$ (a.s), and apply Berman's lemma [3]. For the maximum queue length additional care is needed in order to consider some fluctuation due to discretization as in Anderson [2]. This completes the proof. ■

Remark 3.5. As discussed in Remark 3.3 this analysis cannot be expanded to the case of infinite number of servers. For example, for $M|G|\infty$ it is proved in Aldous *et al* [1] that for some $t_0 > 0$

$$|\mathcal{P}(\sup_{t \leq t_0} Q_t - 1 \leq a) - \exp(-t_0 \lambda e^{-\rho} \rho^{a+1}/(a+1)!)| \to 0 \quad \text{as} \qquad n, a \to \infty \ ,$$

and this is quite different than the limiting law in Theorem 7.

Remark 3.6. As discussed in Remark 2.1 our estimates on the tails for the maximum queue length and waiting time in a busy period work for a heterogeneous $G|G|c$ queue, if one computes θ as a positive solution of (15). Naturally, in such a case Theorem 6 and Theorem 7 are still valid with θ and ω appropriately evaluated.

References

[1] Aldous, D., Hofri, M. and Szpankowski, W. (1991). Maximum size of a dynamic data structure: hashing with lazy deletion revisited. *SIAM J. Computing*, to appear.

[2] Anderson, C.W. (1970), Extreme value theory for a class of discrete distributions with applications to some stochastic processes, *J. Appl. Prob.*, **7**, 99-113.

[3] Berman, S. (1962), Limiting distribution of the maximum term in sequences of dependent random variables, *Ann. Math. Statist.*, **33**, 894-908.

[4] Feller, W. (1971), *An Introduction to Probability and its Applications*, Vol. II, John Wiley & Sons, New York.

[5] Galambos, J. (1978), *The Asymptotic Theory of Extreme Order Statistics*, John Wiley & Sons, New York.

[6] Gniedenko, B.V. (1943), Sur la distribution limite du terme maximum d'une série abéatorie, *Ann. of Math.*, **44**, 423-453.

[7] Heyde, C.C. (1971), On the growth of the maximum queue length in a stable queue, *Oper. Res.*, **44**, 423-452.

[8] Iglehart, D. (1972), Extreme values in the $GI|GI|1$ queue, *The Ann. Math. Statist.*, **43**, 627-635.

[9] Kiefer, J., and Wolfowitz, J. (1955). On the theory of queues with many servers. *Trans. Amer. Math. Soc.*, **78**, 1-18.

[10] Kiefer, J., and Wolfowitz, J. (1956). On the characteristics of the general queueing process, with applications to random walk, *Ann. Math. Stat.*, **27**, 147-161.

[11] Loynes, R. (1962). The stability of a queue with non-independent inter-arrival and service times, *Proceedings of the Camb. Philos. Soc.,* **58** , 497–520.

[12] Neuts, M. and Takahashi, Y. (1981), Asymptotic behavior of the stationary distribution in the $GI|PM|c$ queue with heterogeneous servers, *Z. Wahrscheinlich*, **57**, 441-452.

[13] Nummelin, E. (1987), *General Irreducible Markov Chains and Non-negative Operators*, Cambridge University Press, Cambridge.

[14] Postnikov, A.G. (1980). *Tauberian Theory and Its Applications.* Proc. Steklov Institute of Mathematics, AMS Providence 1980.

[15] Sadowsky, J. (1991). Large deviation theory and efficient simulation of excessive backlogs in a $GI|GI|c$ queue. *Trans. on Automatic Control*, (in press).

[16] Sadowsky, J. and Szpankowski, W., (1990). On the analysis of the tail queue length and waiting time distributions of a $G|G|c$ queue, *Proc. PERFORMANCE'90*, pp. 93-107, Edinburgh 1990.

[17] Sadowsky, J. and Szpankowski, W., (1990). The probability of large queue lengths and waiting times in a heterogeneous $GI|GI|c$ queue, submitted to a journal.

[18] Serfozo, R. (1988). Extreme values of queue length in $M|G|1$ and $GI|M|1$ systems. *Mathematics Oper. Res.*, **13**, 349-357.

[19] Szpankowski, W. (1989), On the maximum queue length with applications to data structures analysis, *Proc. Allerton Conference*, Monticello, 263-272.

[20] Whitt, W. (1972). Embedded renewal processes in the $GI|G|s$ queue. *J. Appl. Prob.*, **9**, 650-658.

[21] Wolff, W. (1989). *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs .