# Sink or Swim Together: Necessary and Sufficient Conditions for Finite Moments of Workload Components in FIFO Multiserver Queues

Alan Scheller-Wolf
Tepper School of Business
Carnegie Mellon University, Pittsburgh, USA
Ph: (412) 268-5066; Fax: (412) 268-7064
awolf@andrew.cmu.edu

Rein Vesilo
Department of Electronics
Macquarie University, Sydney, Australia
Ph: +61 2 9850 9133; Fax: +61 2 9850 9128
rein@ics.mq.edu.au

March 12, 2008

**Abstract**

Previously established necessary and sufficient conditions for finite stationary moments in stable FIFO $GI/GI/s$ queues exist only for the first component of the workload vector, the delay. In this paper, we derive moment results for *all* the components of the stationary workload vector in stable FIFO $GI/GI/s$ queues. As in the case of stationary delay, the moment conditions for workload components incorporate the interaction between service time distribution, traffic intensity and the number of servers in the queue. If we denote a generic service time random variable by $S$, a generic interarrival time by $T$, and define the traffic intensity as $\rho = \mathrm{E}S/\mathrm{E}T$, then sufficient conditions so that $\mathrm{E}W_i < \infty$, where $W_i$ is the $i$'th smallest component of the ordered workload vector, depend crucially on the traffic intensity relative to $i$ — specifically, on whether $i \leq \lceil \rho \rceil$ or $i > \lceil \rho \rceil$, where for any real $x$, $\lceil x \rceil$ denotes the smallest interger greater than or equal to $x$. Explicitly, for $i \leq \lceil \rho \rceil$, $\mathrm{E}W_i^\alpha < \infty$, provided that $\mathrm{E}S^{\beta_1(i)} < \infty$, where $\beta_1(i) = (s - \lfloor \rho \rfloor + \alpha)/(s - \lfloor \rho \rfloor)$, for $\alpha \geq 1$. Furthermore, components with indices lower than $\lceil \rho \rceil$ all share the same finite moment conditions. This is not true for $i > \lceil \rho \rceil$; these components have individual finite moment conditions: $\mathrm{E}W_i^\alpha < \infty$ provided that $\mathrm{E}S^{\beta_2(i)} < \infty$, where $\beta_2(i) = (s - i + \alpha)/(s - i)$, for $\alpha \geq 1$. Finally, for $S$ in the class $\mathcal{L}_1^{\beta_j(i)}$, $j = 1, 2$, defined in [9], these conditions are also necessary.

**KeyWords:** Multiserver Queues, Finite Moment Conditions, Workload Components, Necessary and Sufficient Conditions.

# 1 Introduction

We consider necessary and sufficient conditions for finite moments of the workload vector as seen by an arriving customer in stable FIFO $GI/GI/s$ queues: Queues operating under the first-in-first-out service discipline, in which the time between customer arrivals and the customer service times are mutually independent sequences of i.i.d. random variables. throughout, we denote a generic service time as $S$ and a generic interarrival time is T and the traffic intensity, or as is sometimes used, the load, as $\rho \stackrel{\text{def}}{=} \mathrm{E}(S)/\mathrm{E}(T)$. When the traffic intensity is less than the number of servers $s$, the system is stable. In [8] it was established that moment conditions for delay – the smallest component of the workload vector – incorporate the interaction between the service time distribution, the number of servers in the queue, and the traffic intensity. In particular, [8] showed that for *non-integral* $\rho$ and any $\alpha \geq 1$

$$\mathrm{E}\left(S^{1+\frac{\alpha}{s-\lfloor\rho\rfloor}}\right) < \infty \Rightarrow \mathrm{E}(D^\alpha) < \infty\,. \tag{1}$$

The corresponding necessary condition, still for *non-integral* $\rho$ obtained by [8] and then extended in [9], is that if $S$ is in the class $\mathcal{L}_1^\beta$, where $\beta = 1 + \alpha/(s - \lfloor\rho\rfloor)$, then

$$\mathrm{E}\left(S^{1+\frac{\alpha}{s-\lfloor\rho\rfloor}}\right) < \infty \Leftarrow \mathrm{E}(D^\alpha) < \infty\,. \tag{2}$$

The class $\mathcal{L}_1^\beta$ is defined to be such that $S \in \mathcal{L}_1^\beta$, $1 < \beta < \infty$, implies:

1. $\mathrm{E}(S) < \infty$; and
2. If $S_1, \cdots, S_m$ are $m$ i.i.d random variables distributed as $S$ then

$$\mathrm{E}(S^\beta) = \infty \quad \text{implies} \quad \mathrm{E}((\min(S_1, \cdots, S_m))^{m\beta}) = \infty\,.$$

For more information on the class $\mathcal{L}_1^\beta$, please see [9].

In this paper we extend these delay moment results, deriving necessary and sufficient conditions for the finiteness of the moments of *all* the components of the stationary ordered workload vector, $\mathbf{W} = (W_1, \cdots, W_s)$, (where $W_1 \leq W_2 \leq \cdots \leq W_s$ and thus delay $D = W_1$) in a stable FIFO $GI/GI/s$ queue. Similar to [8] we will only consider the case of *non-integral* $\rho$. When $\rho$ is integral the analysis is much more delicate; we leave this case for further research.

A primary goal of this work is to investigate in what way the delay moment dynamics extend to the workload components, answering questions such as: Does traffic intensity still play a role in the workload component moment conditions (as it did in the delay case)? If it does, what is the nature of that role? Is this role uniform across components or does it vary with component? And if it varies, how?

We show, as in the case of delay, that there is a load dependence on the moment conditions for the workload components; and that this load dependence incorporates the interaction between service time distribution, traffic intensity, $\rho$, the number of servers in the queue, and the specific workload component under consideration. In fact, we show that moment conditions for $\mathrm{E}(W_i) < \infty$ depend crucially on the traffic intensity *relative to* $i$: There is dichotomy in behavior and the moment conditions depend on whether $i \leq \lceil\rho\rceil$ or $i > \lceil\rho\rceil$, where for any rwal $x$, $\lceil x\rceil$ denotes the smallest interger greater than or equal to $x$. Similaryly $\lfloor x\rfloor$ denotes the greatest interger smaller than or equal to $x$.

We now state the main theorem, and principal contribution, of the paper, that gives necessary and sufficient conditions for the $\alpha \geq 1$ moment of a particular workload component to be finite:

**Theorem 1.1** *Given a FIFO $GI/GI/s$ queue with non-integral traffic intensity $\rho$, for $\alpha \geq 1$:*

*(i) The condition*

$$(3) \qquad ES^{(s-\lfloor\rho\rfloor+\alpha)/(s-\lfloor\rho\rfloor)} < \infty \text{ is sufficient for } EW_i^\alpha < \infty \qquad i = 1, \cdots, \lceil\rho\rceil\,.$$

*If $S \in \mathcal{L}_1^\beta$ where $\beta = (s - \lfloor\rho\rfloor + \alpha)/(s - \lfloor\rho\rfloor)$ then Condition (3) is a necessary condition.*

*(ii) If $\lceil\rho\rceil \leq s - 1$, the condition*

$$(4) \quad ES^{(s-i+1+\alpha)/(s-i+1)} < \infty \text{ is sufficient for } EW_i^\alpha < \infty, \qquad i = \lceil\rho\rceil + 1, \cdots, s\,.$$

*If $S \in \mathcal{L}_1^\beta$ where $\beta = (s-i+1+\alpha)/(s-i+1)$ then Condition (4) is a necessary condition.*

Note that $(i)$ establishes *identical* conditions for each of the $\lceil\rho\rceil$ smallest components of the workload vector – they "sink or swim" together. In contrast, $(ii)$ establishes that the $s - \lceil\rho\rceil$ largest components, i.e. components $\lceil\rho\rceil + 1$ to $s$, have *individual* moment conditions.

To help illustrate the nature of the moment conditions in Theorem 1.1, we give an example of a 5-server system in Table 1: Table 1 shows which moment indices of the service time must be finite in order for the first moment of each workload component to be finite (i.e. $\alpha = 1$), for a range of different traffic intensities. In Table 1 the underlined bold numbers indicate cases in which the traffic intensities are such that a server may be removed and the system remains stable (i.e. $i > \lceil\rho\rceil$). These are the components with *individual* moment conditions. Once no more servers may be removed while preserving stability – the non-boldface elements of Table1 – all of the remaining components share the same moment conditions.

An easy method of determining the appropriate moment conditions on the service time distribution is to use the following algorithm:

1. Start with the largest workload component, $W_s$, and set the workload component index $i$ to $s$. The sufficient condition for $\mathrm{E}W_s < \infty$ is the same as that of a $GI/GI/1$ FIFO queue. That is, $\mathrm{E}W_s < \infty$, iff $\mathrm{E}S^2$ is finite. Express this condition in the form $\mathrm{E}S^{(j+1)/j} < \infty$, where $j = 1$.

2. If $i = \lceil\rho\rceil$, then all workload components with index smaller than $i$ have the same moment condition as $W_i$.

3. Otherwise (when $i > \lceil\rho\rceil$) if we remove one server from the system the system remains stable. Then the condition $\mathrm{E}S^{(j+2)/(j+1)} < \infty$ becomes sufficient for $\mathrm{E}W_{i-1} < \infty$. In this case set $i$ to $i - 1$ and $j$ to $j + 1$ and return to 2.

| $\rho$ | 0.5 | 1.5 | 2.5 | 3.5 | 4.5 |
|---|---|---|---|---|---|
| $W_1$ | 6/5 | 5/4 | 4/3 | 3/2 | 2 |
| $W_2$ | **5/4** | 5/4 | 4/3 | 3/2 | 2 |
| $W_3$ | **4/3** | **4/3** | 4/3 | 3/2 | 2 |
| $W_4$ | **3/2** | **3/2** | **3/2** | 3/2 | 2 |
| $W_5$ | **2** | **2** | **2** | **2** | 2 |

Table 1: 5 server queue ($s = 5$)

We proceed to prove our result in the rest of the paper. We begin by presenting some background on related work in Section 2. After describing the notation used in the paper, in Section 3, we give the formal proof of Theorem 1.1 in Section 4. We discuss future directions for work in Section 5.

## 2 Background and Related Work

Research on properties of stationary delay in multiserver queues gained prominence with the work of Kiefer and Wolfowitz [5, 6]. This latter work provided necessary and sufficient conditions for finite stationary delay moments in stable $GI/GI/1$ queues, and showed that these were sufficient conditions for stable $GI/GI/s$ queues.

Analysis of multiserver queues has long been recognized as a difficult problem, and it was not until the work of Scheller-Wolf and Sigman [7] that less stringent sufficient moment conditions than the Kiefer-Wolfowitz conditions were developed for the $GI/GI/s$ multi-server queue. These sufficient conditions for finite delay moments reached their most refined form in [8], which also provides necessary conditions for the class of service time distributions $\mathcal{L}^\beta$; see [8] for a definition of the class $\mathcal{L}^\beta$.

The work of [9] provided an alternative, more intuitive explanation for the moment conditions of [8], and extended the class of service time distributions for which the necessary conditions were valid from $\mathcal{L}^\beta$ to $\mathcal{L}_1^\beta$. The proof of the upper bounds in [9] is simpler than in [8], generalizing techniques used by Foss and Korshunov [2]. In deriving the lower bounds, [9] utilized techniques developed by Whitt [10].

For a more comprehensive literature review on delay moments in multiserver FIFO queues we refer the reader to [8]. Other relevant references on the topic are [1, 2, 7, 10].

One of the implications of the results in this paper and of the results in [7, 8, 9] is that that when faced with service times with finite mean but infinite $\beta^{\text{th}}$ moment, for some finite $\beta$, replacing one fast server (operating at rate 1) with $s$ servers (each operating at rate $1/s$) will significantly reduce the tail of the stationary delay distributions provided $s > 1/(1-\rho)$. In fact, each time the number of servers passes the threshold $s > i/(1-\rho)$, where $i$ is an integer, there is quantum reduction in delay. In the construction and operation of telecommunication networks, such as the Internet, where service times with infinite moments are often used for

modeling, this result shows that to reduce delays it would better to use a larger number of inexpensive slower servers than one expensive fast server. For a more detailed discussion of the relevance of heavy tailed job size distributions in designing computer and telecommunications systems see [3, 4].

## 3 Notation

We consider a $GI/GI/s$ queue, with $s \geq 1$ servers. We introduce the following notation:

- $\tau_n$ denotes the arrival time of customer $n$.
- $\nu_n$ denotes the time customer $n$ enters service.
- $S_n$ denotes the service time of customer $n$, which we assume is brought by the customer. $\{S_n, n = 0, 1, \cdots\}$ is an i.i.d sequence with mean $0 < 1/\mu \stackrel{\text{def}}{=} \mathrm{E}(S_n) < \infty$.
- $T_n \stackrel{\text{def}}{=} \tau_{n+1} - \tau_n$ denotes the time between the arrival of arrival $n$ and arrival $n+1$. $\{T_n, n = 0, 1, \cdots\}$ is an i.i.d sequence with mean $0 < 1/\lambda \stackrel{\text{def}}{=} \mathrm{E}(T_n) < \infty$. $\{T_n\}$ and $\{S_n\}$ are assumed to be mutually independent sequences.
- $W_{n,i}$ denotes the $i^{th}$ component of the Kiefer and Wolfowitz [5] workload vector $\mathbf{W}_n = (W_{n,1}, \cdots, W_{n,s})$; $W_{n,i}$ corresponds to the $i^{th}$ largest workload at a server observed by customer $n$ upon arrival. This vector is defined by the recursion

$$W_{n+1,i} = \mathbf{R}(W_{n,1} + S_n - T_n, W_{n,2} - T_n, \cdots, W_{n,s} - T_n)^+ ,$$

  where $\mathbf{R}$ places the components in ascending order.
- $D_n \stackrel{\text{def}}{=} \nu_n - \tau_n$ is the delay which customer $n$ experiences while waiting for service. Under FIFO, $D_n = W_{n,1}$.
- $\rho \stackrel{\text{def}}{=} \lambda/\mu$. this is the *traffic intensity* experienced by the system. According to this definition, the basic stability condition is $\rho < s$, not $\rho < 1$.

Any time a random variable that is defined above as having a subscript $n$ appears in an unsubscripted form, it will refer to the stationary version of the subscripted random variable.

## 4 Proof of Main Theorem

We prove Theorem 1.1 in this section. Before embarking on the proof, we give the reader an intuitive understanding of the conditions given in Theorem 1.1, again for the special case of first moments ($\alpha = 1$).

Consider the workload components, starting with the largest component, $W_s$. To obtain a large value in $W_s$, it takes the arrival of only a single large service time. This is analogous to

a single server queue, for which the condition $\mathrm{E}S^2 < \infty$ is sufficient for $\mathrm{E}W < \infty$. Hence, this condition is sufficient for $\mathrm{E}W_s < \infty$ in the multiserver system.

Now consider the second largest workload vector, $W_{s-1}$. Suppose that $s-1 < \rho < s$ and a large service time has recently arrived to make $W_s$ large. This causes $W_s$ to be blocked. The arriving load now has to be served by the remaining servers, $W_1, \cdots, W_{s-1}$, but since $s-1 < \rho$, the remaining system will be unstable. By the SLLN, arriving work will be allocated roughly equally between these remaining servers, causing the work at each of these remaining servers to grow in proportion to the work at $W_s$, and thus be of the same order as $W_s$. Hence, $W_{s-1}$ will have a similar service time moment condition to that of $W_s$, i.e. $\mathrm{E}W_{s-1} < \infty$ if $\mathrm{E}S^2 < \infty$. The remaining servers, $W_{s-2}, \cdots, W_1$ will likewise have this service time moment conditions: These servers all "sink" together.

Suppose now that $s-2 < \rho < s-1$. In this case, if one large service time arrives, it is allocated to $W_s$ and blocks that server. However, the arriving load is such that the remaining servers still constitute a stable system; hence, there is not likely to be a significant build up of work in the remaining servers. They all "swim" together. If $W_{s-1}$ is to become large, this will most likely happen due to a *second* large service time arriving soon after the first large service time has arrived. This will only happen, with a large enough probability to affect the finiteness of the stationary workload moments, if the service times have a heavier tail than the $\mathrm{E}S^2 < \infty$ condition entails. The precise condition for $W_{s-1}$ to become large is given in Theorem 1.1. As $s-2 < \rho$, when two large service times have arrived, the remaining $s-2$ server system becomes unstable and the work on servers $s-2, \cdots, 1$ will be over the same order as $W_{s-1}$ and, hence, have the same moment condition as $W_{s-1}$. Now these servers all "sink" together.

Generalizing this to smaller traffic intensities, one large service time is enough need to make $W_s$ large, two large service times are needed to make $W_{s-1}$ large, $\cdots$, $s - \lfloor \rho \rfloor$ large service times are need to make $W_{\lceil \rho \rceil}$ large. Thus the service time distributions needed to make $W_{s-j}$ large ($j = 0, \cdots, s - \lfloor \rho \rfloor$) become progressively heavier as $j$ increases. However, if $s - \lfloor \rho \rfloor$ large service times have arrived, the arrival of another large job will cause the remaining system to becoms unstable; the work on $W_1, \cdots, W_{\lfloor \rho \rfloor}$, because of the SLLN, will become of the same order as $W_{\lceil \rho \rceil}$. Hence, the moment conditions for $W_1, \cdots, W_{\lfloor \rho \rfloor}$ are the same as for $W_{\lceil \rho \rceil}$; these servers all "sink" or "swim" together.

The theorem is proved separately for the cases $i = 1, \cdots, \lceil \rho \rceil$ and $i = \lceil \rho \rceil + 1, \cdots, s$.

## 4.1 Proof of sufficiency for $i = 1, \cdots, \lceil \rho \rceil$

Sufficiency is proved by using the SLLN to show that the workload components $W_1, \cdots, W_{\lceil \rho \rceil}$ are all roughly of the same order, in a probabilitistic sense.

**Proposition 4.1** *For any fixed $x_0 \geq 0$, for all $x > x_0$, there exist constants $C > 0$ and $K > 0$, independent of $x$, such that*

$$\mathbf{P}(W_i > x) \geq C\mathbf{P}(W_{i+1} > Kx) \qquad i \leq \lfloor \rho \rfloor .$$

*Proof :* The proof is given in Appendix A.1 ■

*Proof :*(Sufficiency for $i = 1, \cdots, \lceil \rho \rceil$)

To prove sufficiency for given $1 \leq i \leq \lceil \rho \rceil$, use Proposition 4.1 iteratively to give $\mathbf{P}(W_1 > x) \geq C^{i-1}\mathbf{P}(W_i > K^{i-1}x)$, for $i = 2, \cdots, \lceil \rho \rceil$. Hence,

$$\int_{x_0}^{\infty} \mathbf{P}(W_1 > x)\, dx \geq \int_{x_0}^{\infty} C^{i-1}\mathbf{P}(W_i > K^{i-1}x)\, dx\,.$$

By Theorem 2.1 of [9], if $\mathrm{E}S^{1+\alpha/(s-\lfloor \rho \rfloor)} < \infty$ then $\mathrm{E}W_1^{\alpha} < \infty$. This then implies $\mathrm{E}W_i^{\alpha} < \infty$ by integrating the tail of the distribution to get the moment of order $\alpha$ (see for example page 37 of [11]).

■

### 4.2 Proof of necessity for $i = 1, \cdots, \lceil \rho \rceil$

Suppose $S \in \mathcal{L}_1^{\beta}$. By Theorem 3.2 of [9], $\mathrm{E}S^{1+\alpha/(s-\lfloor \rho \rfloor)} = \infty$ implies $\mathrm{E}W_1^{\alpha} = \infty$. Since $W_i \geq W_1$, this implies $\mathrm{E}W_i^{\alpha} = \infty$, which proves necessity.

### 4.3 Proof of sufficiency for $i = \lceil \rho \rceil + 1, \cdots, s$

Denote the $i^{th}$ component of the workload seen by customer $n$ in an $m$-server system by $W_{n,i}^m$. Sufficiency for $i = \lceil \rho \rceil + 1, \cdots, s$ is obtained by showing that the workload for component $i$ seen by customer $n$ in an $s$-server system is less than the workload seen by customer $n$ for component $i-1$ in an $(s-1)$-server system, facing the same $\{S_n\}$ and$\{T_n\}$ sequences, provided both systems are empty at index 0. This is as stated in the following Lemma.

**Lemma 4.1** *Suppose that $W_{0,i}^s = 0$ for $i = 1, \cdots, s$ and $W_{0,i}^{s-1} = 0$ for $i = 1, \cdots, s-1$, then if the $s-1$ and $s$ server queues face identical $\{S_n\}$ and$\{T_n\}$ sequences,*

$$W_{n,i}^s \leq W_{n,i-1}^{s-1} \qquad i = 2, \cdots, s \quad a.s.$$

*Proof :* See Appendix B ■

*Proof :* (Proof of sufficiency for $i = \lceil \rho \rceil + 1, \cdots, s$)

Apply Lemma 4.1 $j$ times to give

$$W_{n,i}^s \leq W_{n,i-j}^{s-j}\,.$$

Now, set $i - j = \lceil \rho \rceil$ to give

$$W_{n,i}^s \leq W_{n,\lceil \rho \rceil}^{s-i+\lceil \rho \rceil}\,. \tag{5}$$

By using the sufficiency results for $1 \leq i \leq \lceil \rho \rceil$ in equation (3), with $s$ replaced by $s - i + \lceil \rho \rceil$, it follows from $\lceil \rho \rceil = \lfloor \rho \rfloor + 1$ that

$$ES^{(s-i+1+\alpha)/(s-i+1)} < \infty$$

is sufficient for $E(W^{s-i+\lceil\rho\rceil}_{\lceil\rho\rceil})^\alpha < \infty$. Hence, using the inequality (5), this also implies $E(W^s_i)\alpha < \infty$.

■

### 4.4 Proof of necessity for $i = \lceil\rho\rceil + 1, \cdots, s$

We prove necessity by obtaining a lower bound for $W^s_i$ for $i = \lceil\rho\rceil + 1, \cdots, s$ by using Lemma 4.3 of [8].

**Lemma 4.2** *(Lemma 4.3 [8]) Suppose two initially empty FIFO queues having $s-1$ and $s$ servers, respectively, are fed by two identical interarrival and service time sequences. In addition, suppose the $s-1$ server queue operates at twice the rate of the $s$ queue system. Then, for $i = 2, \cdots, s$ and all $n$, $W^{s-1}_{n,i-1} \le W^s_{n,1} + W^s_{n,i}$.*

Suppose we have a set of systems, labelled $m = 1, \cdots, s$, so that in system $m$ there are $m$ servers in the system and the service rate of a server in the $m$-server system is $2^{s-m}$ times the rate of a server in the $s$-server system.

From Lemma 4.3 [8], we have for an $(s-1)$-server system, since $W^s_{n,1} \le W^s_{n,i}$, that

$$W^{s-1}_{n,i-1} \le 2W^s_{n,i}\,.$$

Applying this inequality $j$ times gives, for an $(s-j)$-server system,

$$W^{s-j}_{n,i-j} \le 2^j W^s_{n,i}\,.$$

In the particular case that $j = i-1$, we have

$$W^{s-i+1}_{n,1} \le 2^{i-1} W^s_{n,i}\,. \tag{6}$$

Let $\rho_m$ $(1 \le m \le s)$ denote the traffic intensity for an $m$-server system. For our set of systems, $\rho_m$ is equal to $\rho/2^{s-m}$. The traffic intensity on the $s-i+1$-server system given on the left hand side of equation (6) is thus

$$\rho_{s-i+1} \;=\; \frac{\rho}{2^{i-1}}\,. \tag{7}$$

By the conditions of Theorem 1.1, $i - 1 \ge \lceil\rho\rceil$, we get from equation (7) that

$$0 < \rho_{s-i+1} \;<\; \frac{\lceil\rho\rceil}{2^{\lceil\rho\rceil}} < 1\,,$$

where the last inequality is true for all $\lceil\rho\rceil$.

If $S \in \mathcal{L}^\beta_1$, a sufficient condition for $E(W^{s-i+1}_1)^\alpha = \infty$ is obtained by using equation (2) for an $(s-i+1)$-server system with traffic intensity $\rho_{s-i+1} < 1$. To determine this condition, replace $s$ with $s-i+1$ and replace $\lfloor\rho\rfloor$ with 0, to give

$$\mathrm{E}S^{1+\alpha/(s-i+1)} = \mathrm{E}S^{(s-i+1+\alpha)/(s-i+1)} = \infty\,. \tag{8}$$

Equation (6) then implies that the condition in equation (8), is sufficient for $\mathrm{E}(W^s_{n,i})\alpha = \infty$, which proves the theorem.

# 5 Conclusion

In this paper we have extended the previously know finite delay moment conditions for stationary FIFO multiserver queues to all of the components of the Kielfer and Wolfowitz workload vector. We have shown that while the conditions for the higher components share the delay moment's dependence on the service time distribution, traffic intensity, and number of servers in the queue, they do so in a unique way: While the $s - \lceil \rho \rceil$ largest workload components have individual finite moment conditions, that grow increasingly weaker as the component index grows smaller, the $\lceil \rho \rceil$ smallest components all share the same conditions as those for finite delay moments – these sink or swim together.

# A Appendix

## A.1 Proof of Proposition 4.1

*Proof :*

The proof of the proposition uses the SLLN to obtain bounds on sample path probabilities. We assume that the arriving workload process is defined by two doubly infinite stationary sequences of service times and interarrival times $\{S_j, T_j \ : \ j = -\infty, \cdots \infty\}$, where $S_j$ is the service time of customer $j$ and $T_j$ is the time between the arrival of customer $j$ and the arrival of customer $j+1$. The sequences are $\{S_j\}$ and $\{T_j\}$ are mutually independent. The proof examines the workload vector seen by arrival 0: $(W_{0,1}, \cdots, W_{0,s}) \equiv (W_1, \cdots, W_s)$.

For ease of notation, define, for $-m \geq -n$,

$$S[-n,-m] \equiv \sum_{i=-n}^{-m} S_i \quad \text{and} \quad T[-n,-m] \equiv \sum_{i=-n}^{-m} T_i \, ;$$

$S[-n,-m]$ is the total work bought by customers $-n$ to $-m$ inclusively, and $T[-n,-m]$ is the time between the arrival of customer $-n$ and the arrival of customer $-m+1$. For $-m < -n$, define $S[-n,-m] = T[-n,-m] = 0$.

We now define important events needed in our proof. Given $0 < \delta < 1$, since $\mathrm{E}S < \infty$, by the SLLN there exists $n_1$ such that, for all $n \geq n_1$,

$$\mathbf{P}\{(\mathrm{E}S - \epsilon)n < S[-n,-1] < (\mathrm{E}S + \epsilon)n\} > 1 - \delta/2 \, . \tag{9}$$

Similarly, since $\mathrm{E}T < \infty$, there exists $n_2$ such that, for all $n \geq n_2$,

$$\mathbf{P}\{(\mathrm{E}T - \epsilon)n < T[-n,-1] < (\mathrm{E}T + \epsilon)n\} > 1 - \delta/2 \, . \tag{10}$$

Define the event,

$$\begin{aligned} SLLN_{\tilde{n}} \ &= \ \{(\mathrm{E}S - \epsilon)n < S[-n,-1] < (\mathrm{E}S + \epsilon)n, \\ &\qquad (\mathrm{E}T - \epsilon)n < T[-n,-1] < (\mathrm{E}T + \epsilon)n \, , n \geq \tilde{n}\} \end{aligned} \tag{11}$$

and define the event $SLLN$ to be

$$SLLN \equiv SLLN_{n_0} \, , \tag{12}$$

where $\tilde{n} = n_0 \equiv \max(n_1, n_2)$. It follows from the independence of $\{S_n\}$ and $\{T_n\}$ and equations (9) and (10) that $\mathbf{P}(SLLN) > (1-\delta/2)^2 > 1-\delta$.

Given $x > 0$, define the events:

$$E_x = \{W_{-N_x,i+1} > K_x x\} \quad \text{and} \tag{13}$$

$$H_x = \{W_{-N_x,1} = \cdots = W_{-N_x,i} = 0\} \tag{14}$$

where $N_x$ and $K_x$ are chosen such that

$$N_x > \max\left(\frac{ix}{(\mathrm{E}S-\epsilon) - i(\mathrm{E}T+\epsilon)}, n_0\right) \quad \text{and} \tag{15}$$

$$K_x > \frac{(\mathrm{E}S + \mathrm{E}T + 2\epsilon)N_x}{x}\,. \tag{16}$$

Define also $x_0 \equiv n_0((\mathrm{E}S-\epsilon)/i - (\mathrm{E}T+\epsilon))$.

$E_x$ is the event that the workload at component $i+1$ a "long" time ($-N_x$ arrivals) in the past was "large" (as defined by $K_x$), while $H_x$ lower bounds the work at components $1, \ldots i$. Note that by definition, $-N_x < -n_0$.

To prove the proposition, we begin with a lower bound on $\mathbf{P}(W_{0,i} > x)$ given by

$$\mathbf{P}(W_{0,i} > x) \geq \mathbf{P}(W_{0,i} > x, E_x) = \mathbf{P}(W_{0,i} > x | E_x)\mathbf{P}(E_x)\,. \tag{17}$$

We examine the right-most terms of this inequality.

By stationarity,

$$\mathbf{P}(E_x) = \mathbf{P}(W_{-N_x,i+1} > K_x x) = \mathbf{P}(W_{0,i+1} > K_x x)\,. \tag{18}$$

In addition, $\mathbf{P}(W_{0,i} > x | E_x)$ is lower bounded as follows:

$$\begin{aligned} \mathbf{P}(W_{0,i} > x | E_x) &\geq \mathbf{P}(W_{0,i} > x, SLLN | E_x) \\ &= \mathbf{P}(W_{0,i} > x | E_x, SLLN)\mathbf{P}(SLLN | E_x)\,. \end{aligned} \tag{19}$$

As the events $SLLN$ and $E_x$ are independent (as $-n_0 > -N_x$), $\mathbf{P}(SLLN|E_x) = \mathbf{P}(SLLN)$. Also, defining $H^c$ as the complement of event $H$:

$$\begin{aligned} \mathbf{P}(W_{0,i} > x | E_x, SLLN) &= \mathbf{P}(W_{0,i} > x | E_x, SLLN, H_x)\mathbf{P}(H_x | E_x, SLLN) \\ &\quad + \mathbf{P}(W_{0,i} > x | E_x, SLLN, H_x^c)\mathbf{P}(H_x^c | E_x, SLLN) \\ &\geq \mathbf{P}(W_{0,i} > x | E_x, SLLN, H_x)\mathbf{P}(H_x | E_x, SLLN) \\ &\quad + \mathbf{P}(W_{0,i} > x | E_x, SLLN, H_x)\mathbf{P}(H_x^c | E_x, SLLN) \\ &= \mathbf{P}(W_{0,i} > x | E_x, SLLN, H_x) = 1\,, \end{aligned} \tag{20}$$

where the inequality in the second line follows from the monotonicity of the workload and the last line follows from Lemma A.1. Combining equations (17), (18), (19) and (20) gives,

$$\begin{aligned} \mathbf{P}(W_{0,i} > x) &\geq \mathbf{P}(SLLN)\mathbf{P}(W_{0,i+1} > K_x x) \\ &\geq (1-\delta)\mathbf{P}(W_{0,i+1} > K_x x) \\ &\geq C\mathbf{P}(W_{0,i+1} > K_x x) \end{aligned}$$

where $C \equiv 1 - \delta > 0$, which is independent of $x$.

To complete the proof, note that from (15) and (16) $K_x$ is a non-increasing function of $x$, so we may define $K \stackrel{\text{def}}{=} K_{x_0} \geq K_x$, for $x > x_0$ yielding:

$$\mathbf{P}(W_{0,i} > x) \quad \geq \quad C\mathbf{P}(W_{0,i+1} > K_x x) \geq C\mathbf{P}(W_{0,i+1} > Kx), \qquad x \geq x_0\,.$$

We now show that $K$ is strictly positive under the assumptions in the statement of the proposition. From equations (15) and (16), and dividing the numerator and denomonator by $i\mathrm{E}T$:

$$K \quad > \quad \frac{i(\mathrm{E}S + \mathrm{E}T + 2\epsilon)}{(\mathrm{E}S - \epsilon) - i(\mathrm{E}T + \epsilon)} = \frac{(\rho + 1 + 2\epsilon')}{\rho/i - 1 - \epsilon'(1 + 1/i)}\,, \tag{21}$$

where $\epsilon' \equiv \epsilon/\mathrm{E}T$. Since $i \leq \lfloor \rho \rfloor$ $\epsilon$ can be chosen small enough so that the right hand side of equation (21) is $> 0$.

■

### A.1.1 Lemma A.1

**Lemma A.1** *For $N_x$ and $K_x$ defined by equations (15) and (16), respectively, events $SLLN$, $E_x$ and $H_x$ defined by equations (12) (13) and (14), respectively, and $i \leq \lfloor \rho \rfloor$,*

$$\mathbf{P}(W_{0,i} > x | E_x, SLLN, H_x) = 1\,.$$

*Proof :*

The first part of the proof of the Lemma involves showing that under events $SLLN$, $E_x$ and $H_x$ all the work arriving with customers $-N_x$ to $-1$ is allocated to the servers associated with workload components $1, \cdots, i$ seen by arrival $-N_x$.

First, for $-N_x \leq -n \leq 0$, we have the following lower bound for $W_{-n,i+1}$:

$$W_{-n,i+1} \quad \geq \quad W_{-N_x,i+1} - T[-N_x, -n-1]\,. \tag{22}$$

This follows because, after the arrival of each customer $-m \in \{-N_x, \cdots, -n-1\}$, the most each workload component can reduce is $-T_{-m}$, with the rearrangement of workload components only possibly increasing $W_{-n,i+1}$, due to the addition of a service time.

Since $T[-N_x, -n-1] \leq T[-N_x, -1]$ we get from inequality (22):

$$W_{-n,i+1} \quad \geq \quad W_{-N_x,i+1} - T[-N_x, -1]\,. \tag{23}$$

Since $W_{-N_x,i+1} > K_x x$ on event $E_x$ and $T[-N_x, -1] < (\mathrm{E}T + \epsilon)N_x$ on event $SLLN$ (and since $N_x > n_0$) we get from (23) that

$$W_{-n,i+1} > K_x x - (\mathrm{E}T + \epsilon)N_x\,. \tag{24}$$

In addition, from the definition of $K_x$ in equation (16),

$$K_x x - (\mathrm{E}T + \epsilon)N_x \quad > \quad (\mathrm{E}S + \epsilon)N_x\,.$$

Inserting this into inequality (24) gives

$$W_{-n,i+1} > (\mathrm{E}S + \epsilon)N_x \,. \tag{25}$$

We use this lower bound on $W_{-n,i+1}$ to show that under the conditions of the Lemma the workloads seen on the servers associated with $W_{-N_x,j}$, $j = 1, \cdots, i$, never exceed the workload seen on the server associated with $W_{-N_x,i+1}$, after any arrival with index between $-N_x$ and 0.

To show this, we first specify servers more precisely. Suppose that when customer $-N_x$ arrives, workload components $1, \cdots, s$ correspond to servers $r_1, \cdots, r_s$ and when customer 0 arrives, workload components $1, \cdots, s$ correspond to servers $r_1^0, \cdots, r_s^0$.

On the event $SLLN \cap H_x$, the most the workload seen by customer $-n$ on any of the servers $r_1, \cdots, r_i$ is given by:

$$W_{-n,j} \le S[-N_x, -n] \le S[-N_x, -1] \,, \qquad j = 1, \cdots, i \,,$$

Applying the the upper bound $S[-N_x, -1] \le (\mathrm{E}S + \epsilon)N_x$ on event $SLLN$ to this inequality gives

$$W_{-n,j} \;\; \le \;\; (\mathrm{E}S + \epsilon)N_x \,, \qquad j = 1, \cdots, i \,. \tag{26}$$

As the upper bound on workloads for $W_{-n,j}$, $j = 1, \cdots, i$, given in inequality (26) equals the the lower bound for $W_{-n,i+1}$ given in inequality (25), we get $W_{-n,j} \le W_{-n,i+1}$, for $j = 1, \cdots, i$ and $-N_x \le -n \le -1$, on the event $E_x \cap SLLN \cap H_x$. Hence, no work is assigned to any of the servers associated with workload components $W_{-n,j}$, $j = i+1, \cdots, s$, for $-N_x \le -n \le -1$, i.e. no work is assigned to servers $r_{i+1}, \cdots, r_s$ and $r_1^0, \cdots, r_i^0$ is a permutation of $r_1, \cdots, r_i$.

We use this result to show that when arrival 0 occurs, workload component $i$ seen by arrival 0 is at least $x$.

We have $W_{0,i}$ bounded below by

$$W_{0,i} \;\; \ge \;\; \frac{1}{i}\sum_{j=1}^{i} W_{0,i} = \frac{W_{0,1} + \cdots + W_{0,i}}{i} \,. \tag{27}$$

Let $\overline{W}_{-n,r_j}$ be the work on server $r_j$ seen by arrival $-n$ ($-N_x \le -n \le 0$); note that this may not be $W_{-n,j}$ as this is server $r_j$ defined at $-N_x$. Let $I_{r_j} \ge 0$ be the total idle time for server $r_j$ in the time interval between the arrival of customer $-N_x$ and the arrival of customer 0. Then, by the conservation of work, as all work is allocated to servers $r_1, \cdots, r_i$, on the event $E_x \cap SLLN \cap H_x$,

$$\begin{aligned} \sum_{j=1}^{i} W_{0,j} \;\; &= \;\; S[-N_x, -1] + \sum_{j=1}^{i} \overline{W}_{-N_x,r_j} - \left( iT[-N_x, -1] - \sum_{j=1}^{i} I_{r_j} \right) \\ &\ge \;\; S[-N_x, -1] - iT[-N_x, -1] \,, \end{aligned}$$

where we have used the fact that on event $H_x$, $\overline{W}_{-N_x,r_j} = 0$, $j = 1, \cdots, i$.

Since $S[-N_x, -1] \ge (\mathrm{E}T - \epsilon)N_x$ on the event $SLLN$, this gives

$$\sum_{j=1}^{i} W_{0,j} \;\; \ge \;\; (\mathrm{E}S - \epsilon)N_x - iT[-N_x, -1] \,. \tag{28}$$

Hence, combining inequalities (27) and (28), under the conditions of the Lemma,

$$W_{0,i} \geq \frac{(\mathrm{E}S - \epsilon)N_x}{i} - T[-N_x, -1]\,.$$

Using the upper bound $T[-N_x, -1] \leq (\mathrm{E}T + \epsilon)N_x$, we get on the event $SLLN$,

$$W_{0,i} \geq \frac{(\mathrm{E}S - \epsilon)N_x}{i} - (\mathrm{E}T + \epsilon)N_x\,.$$

Applying the definition of $N_x$ in equation (15) to this inequality, we get

$$W_{0,i} \geq x\,.$$

Thus,

$$\mathbf{P}(W_{0,i} > x | H_x, E_x, SLLN) = 1\,.$$

■

# B Proof of Lemma 4.1

We first begin with a preliminary Lemma.

**Lemma B.1** *Given two real vectors* $\mathbf{X} = \{X_1, X_2, X_3\}$ *and* $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$*, define* $X_{(i)}$ *and* $Y_{(i)}$ *as the* $i^{tth}$ *largest elements of* $\mathbf{X}$ *and* $\mathbf{Y}$*, respectively.*

*If* $X_i \leq Y_i$ *for* $i = 1, 2, 3$*, then* $X_{(2)} \leq Y_{(2)}$.

*Proof :* Without loss of generality, assume $Y_1 \leq Y_2 \leq Y_3$, and assume in addition that $X_{(2)} > Y_{(2)}$. This implies that $Y_1 \leq Y_2 < X_{(2)} \leq X_{(3)}$. Thus either $Y_1 < X_1$ or $Y_2 < X_2$, contradicting the statement of the Lemma. ■

*Proof :*[Proof of Lemma 4.1]

Induction is used to prove the result. The result is clearly true for $n = 0$. Assume it is true for $n$, that is, $W^s_{n,i} \leq W^{s-1}_{n,i-1}$, $i = 2, \cdots, s$. We will use the induction hypothesis to prove that $W^s_{n+1,i} \leq W^{s-1}_{n+1,i-1}$, $i = 2, \cdots, s$.

Within the framework of Lemma B.1, define the vector $\mathbf{X}$ to be $\{(W^s_{n,i} - T_n)^+, (W^s_{n,i+1} - T_n)^+, (W^s_{n,1} + S_n - T_n)^+\}$ and the vector $\mathbf{Y}$ to be $\{(W^{s-1}_{n,i-1} - T_n)^+, (W^{s-1}_{n,i} - T_n)^+, (W^{s-1}_{n,1} + S_n - T_n)^+\}$. Then it holds that $W^s_{n+1,i} = X_{(2)}$ and $W^{s-1}_{n+1,i-1} = Y_{(2)}$.

Given the induction hypotesis, the result follows immediately from Lemma B.1. ■

## References

[1] Asmussen, S., *Applied Probability and Queues*, Springer Verlag, New York, 2003.

[2] Foss, S. and Korshunov, D., Heavy Tails in Multi-Server Queues. *Submitted to Queueing Systems*, 2004.

[3] Harchol-Balter, M., The Effect of Heavy-Tailed Job Size. Distributions on Computer System Design *Proceedings of ASA-IMS Conference on Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics*, Washington, DC, June 1999.

[4] Harchol-Balter, M., Task Assignment with Unknown Duration *Journal of the ACM* , Vol. 49, No. 2, March 2002, pp. 260-288.

[5] Kiefer, J. and Wolfowitz, J., On the Theory of Queues with Many Servers. *Transactions of the American Mathematical Society*, 78, pp. 1 – 18, 1955.

[6] Kiefer, J. and Wolfowitz, J., On the Characteristics of the General Queueing Process with Applications to Random Walk. *Annals of Mathematical Statistics*, 27, pp. 147 - 161, 1956.

[7] Scheller-Wolf, A. and Sigman, K., Delay Moments for FIFO $GI/GI/c$ Queues. *Queueing Systems*, 25, pp. 77 – 95, 1997.

[8] Scheller-Wolf, A., Necessary and Sufficient Conditions for Delay Moments in FIFO Multiserver Queues: Why s Slow Servers are Better than One Fast Server for Heavy-Tailed Systems. *Operations Research*, 51, pp. 748 – 758, 2003.

[9] Scheller-Wolf, A. and Vesilo, R., Structural Interpretation and Derivation of Necessary and Sufficient Conditions for Delay Moments in FIFO Multiserver Queues, *Queueing Systems*, 54, 221 – 232, 2006.

[10] Whitt, W., The impact of a heavy-tailed service-time distribution upon the $M/GI/s$ waiting-time distribution. *Queueing Systems*, 36, pp. 71 – 87, 2000.

[11] Wolff, R., *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, New Jersey, 1989.