Delay Bounds for Multiclass FIFO

Yuming Jiang
Department of Telematics
Norwegian University of Science and Technology
jiang@item.ntnu.no

Vishal Misra
Department of Computer Science
Columbia University
misra@cs.columbia.edu

ABSTRACT
FIFO is perhaps the simplest scheduling discipline. For single-class FIFO, its delay guarantee performance has been extensively studied: The well-known results include a stochastic delay bound for $GI/GI/1$ by Kingman [2] and a deterministic delay bound for $D/D/1$ by Cruz [3]. However, for multiclass FIFO, few such results are available. To fill the gap, we prove delay bounds for multiclass FIFO in this work, considering both deterministic and stochastic cases. Specifically, delay bounds are presented for $D/D/1$, $G/D/1$, $GI/D/1$, and $GI/GI/1$, all under multiclass FIFO.

1. INTRODUCTION
Multiclass FIFO refers to the scheduling discipline where customers are served in the first-in-first-out (FIFO) manner and the services required by different classes may differ. Compared to single-class FIFO, multiclass FIFO provides a more natural way to model the system in many scenarios. One example is input-queueing on a switch, where packets are FIFO-queued at the input port before being forwarded to the output ports that may pick packets from the FIFO queue and serve at different rates. Another example is downlink-sharing in wireless networks, where a wireless base station, shared by multiple users, sends packets to them in the FIFO manner. Since the characteristics of the wireless channel seen by these users may differ, the data rates to them may also be different. For the concern of providing delay guarantees in such networks, it is crucial to have analytical results, particularly delay bounds for multi-class FIFO.

Surprisingly, while there are a lot of results for single-class FIFO, few such results exist for multiclass FIFO. The existing results for multiclass FIFO are mostly under the classic queueing theory (e.g. [4]). However, those available results are rather limited and their focus has been on the queue stability condition. In the context of input-queueing in packet-switched systems, multiclass FIFO has also been studied (e.g. [5]). However, in these studies, the focus has been on the throughput of the switch, assuming saturated traffic on each input port. None of these studies has focused on the delay bound performance of multiclass FIFO.

In terms of delay bounds for FIFO, available results are almost all for single-class FIFO. Among them are the two well-known delay bound results: one by Kingman [2] for a stochastic case: $GI/GI/1$, and one by Cruz [3] for a deterministic case: $D/D/1$. For other or more general arrival and service processes, various delay bounds have also been derived, mainly in the context of network calculus [6]. With those results, one might expect that they could be readily or easily extended to multi-class FIFO. Unfortunately, such an extension is surprisingly difficult and a direct extension may result in rather limited applicability of the obtained results.

This work is devoted to filling the gap. In the following, the difficulties are first discussed. Then, we prove delay bounds for multiclass FIFO, considering both deterministic and stochastic cases. Specifically, delay analysis is performed and delay bounds are derived for $D/D/1$, $G/D/1$, $GI/D/1$ and $GI/GI/1$, all under multiclass FIFO. The bound for $D/D/1$ is tight, and the bounds for $GI/D/1$ and $GI/GI/1$ are similar to the Kingman’s bound in form and accuracy.

2. DIFFICULTIES
For a $GI/GI/1$ FIFO queue, the following delay bound was proved by Kingman five decades ago [2]:

$$P\{W \geq \tau \} \leq e^{-\vartheta \tau}$$

(1)

where $W$ denotes the waiting time in queue of a customer, $\vartheta = \sup\{\theta > 0 : M_1(\theta) < 1\}$ with $X(1)$ being the interarrival time and $Y(1)$ the service time of the first customer. In this paper, for a random variable $Z$, $M_Z(\theta)$ denotes its moment generating function (MGF), i.e., $M_Z(\theta) \equiv E[e^{\theta Z}]$, $F_Z$ its CDF or a lower bound on the CDF, and $\bar{F}_Z$ its CCDF or an upper bound on the CCDF.

The i.i.d. condition on service times implies that the bound (1) is for single-class FIFO. Extending it to multi-class FIFO is formidable difficult. For one reason, due to the implicit dependence between $X(1)$ and $Y(1)$ if we were to treat the inputs of all classes as one aggregate input, finding $M_Y(1)X(1)(\theta)$ for the aggregated sequence would be extremely hard. As a result, if the Kingman’s bound were adopted for $GI/GI/1$ multiclass FIFO, it would be difficult to find the exponent parameter in the bound.

For a $D/D/1$ FIFO queue in communication networks, the following delay bound was initially shown by Cruz [4]

$$D \leq \frac{\sigma}{C}$$

(2)

where $D$ denotes the system delay of any packet, $C$ (in bps) the service rate of the queue, and $\sigma$ (in bits) the traffic burstiness parameter. The conditions of the Cruz’ delay bound are that the input traffic during any time period of length $t$ is upper-constrained by $r \cdot t + \sigma$ and that $r \leq C$.
Unlike (1), the Cruz’ bound can be readily used for multiclass FIFO. To illustrate this, consider a FIFO queue with $N$ classes, where the traffic of each class is upper-constrained by $r_n \cdot t + \sigma_n$ and the service rate of the class is $C_n$. Without difficulty, (2) can be extended to this multiclass queue and it can be shown that the delay of any packet is upper-bounded by $D \leq \frac{\sum_n \sigma_n}{\min_n \{C_n\}} + \epsilon t^0$, if $\sum_n r_n \leq \min_n \{C_n\}$. Unfortunately, the condition $\sum_n C_n \leq \min_n \{C_n\}$ can be too restrictive, particularly when the service rates differ much.

3. MAIN RESULTS

Consider a work-conserving multiclass FIFO queue. The queue is initially empty at time 0. There are $N$ classes. For each class $n$, let $\lambda_n$ denote the average customer arrival rate, $\mu_n$ the average customer service rate, both in number of customers per unit time, and $\rho_n = \frac{\lambda_n}{\mu_n}$. To gain insights, before studying the usual continuous time multiclass FIFO queue, we consider discrete time multiclass FIFO in communication networks, where the packets are the customers. For such a queue, let $C_n$ denote the service rate (in bps) of class $n$, which is constant. Also let $\tau_n$ (in bits) denote the expected packet length of class $n$. In addition, for each class $n$, let $A_n(0, t) \equiv A(t)$ denote the amount of traffic (in bits) which arrives within the time period $[0, t]$, and $A_n(s, t) \equiv A_n(t) - A_n(s)$ the traffic in $[s, t)$. For the aggregate traffic of all classes, $A(s, t) \equiv \sum_n A_n(s, t)$ and $A(t)$ are similarly defined. It is easily calculated: $\lambda_n = \lim_{t \to \infty} \frac{A_n(t)}{t}$ and $\mu_n = \frac{C_n}{t}$. For any packet/customer $p$, let $a^j$ and $d^j$ respectively denote its arrival time and departure time. The delay of the packet/customer is then $D^j = d^j - a^j$. In the rest, for time $t$, we define $t_+ \equiv t + \epsilon$ and $t_- \equiv t - \epsilon$ with $\epsilon \to 0$.

3.1 Delay Bound for Multiclass $D/D/1$

For the communication network queue, suppose the traffic of each class $n$ is constrained by $A_n(s, t) \leq r_n(t-s) + \sigma_n$ for all $0 \leq s \leq t$. For this multiclass $D/D/1$ queue, we have:

**Theorem 1.** If $\sum_n \frac{\sigma_n}{C_n} \leq 1$, the delay of any packet is bounded by:

$$D \leq \sum_n \frac{\sigma_n}{C_n}$$  \hspace{1cm} (3)

**Proof.** For any packet $p^j$, there exists some time $t_0$ that starts the busy period where it is in. Note that such a busy period always exists, since in the extreme case, the period is only the service time period of $p^j$ and in this case, $t^0 = a^j$.

Since the system is work-conserving, there holds: $d^j = t^0 + \sum_{n=1}^{N} S_n(t^0, d^j)$, where $S_n(t^0, d^j)$ denotes the total service time of class $n$ packets that are served in $[t^0, d^j]$. Because of FIFO and that the system is empty at $t^0$, $S_n(t^0, d^j)$ is hence limited by the amount of traffic that arrives in $[t^0, a^j_n)$, which is $A(t^0, a^j_n)$. Specifically, $S_n(t^0, d^j) \leq \frac{A_n(t^0, a^j_n)}{C_n}$, so we then have $d^j \leq t^0 + \sum_{n=1}^{N} \frac{A_n(t^0, a^j_n)}{C_n}$.

Under the condition $\sum_n \frac{\sigma_n}{C_n} \leq 1$, we then obtain:

$$D^j \leq \sum_{n=1}^{N} \frac{A_n(t^0, a^j_n)}{C_n} + t^0 - a^j$$ \hspace{1cm} (4)$$

where the second last step is due to the traffic constraint and the last step is from $\sum_n \frac{\sigma_n}{C_n} \leq 1$ and $a^j > t_0$.

Note that, for the system, consider that immediately after time 0, every traffic class generates a burst with size $\sigma_n$. In this case, the packet in the bursts, which receives service last, will experience delay $\sum_{n=1}^{N} \frac{\sigma_n}{C_n}$ that equals the delay bound. So, the bound is tight. \[\Box\]

It is worth highlighting that the condition $\sum_n \frac{\sigma_n}{C_n} \leq 1$ also leads to $\rho \leq 1$, because $\rho_n \leq \frac{\lambda_n}{\mu_n}$. This is consistent with the multiclass FIFO stability condition, $\rho < 1$, proved under the classic queueing theory analysis (see, e.g., [4]).

3.2 Delay Bounds for Multiclass $G/D/1$; $GI/D/1$

Still consider the communication network queue, but the traffic of each class is constrained to have generalized Stochastically Bounded Burstiness (gSBB) [8]: $P\{\sup_{0 \leq t \leq \tau} [A_n(s, t) - r_n(t-s)] > \sigma \} \leq \bar{F}_n(\sigma)$, for all $t > 0$ and $\sigma \geq 0$. A wide range of traffic processes have been proved to have gSBB [8]. Then, we have the following delay bounds.

**Theorem 2.** If $\sum_n \frac{\sigma_n}{C_n} \leq 1$, the delay of any packet is bounded by (a.s.):

$$P\{D > \tau\} \leq \inf_{p_{1,\ldots,N}=1} \sum_{n=1}^{N} \bar{F}_n(p_n C_n \tau)$$

and if the arrival processes of the $N$ classes are independent of each other, then the delay is bounded by (a.s.):

$$P\{D > \tau\} \leq 1 - F_1 * \cdots * F_N(\tau)$$

where $F_N(\tau) \equiv 1 - \bar{F}_n(C_n \tau)$, and $* \text{ denotes the convolution operation.}$

**Proof.** Following the proof of Th. [4], we have obtained [4]. By applying $\sum_{n=1}^{N} \frac{\sigma_n}{C_n} \leq 1$ to (4), we further obtain

$$D^j \leq \sum_{n=1}^{N} \frac{A_n(t^0, a^j_n) - r_n(a^j - t^0)}{C_n}$$ \hspace{1cm} (5)

Note that in (5), $t^0$ is a random variable. Taking all sample paths into consideration, with $A_n(t^0, a^j_n) - r_n(a^j - t^0) \leq \sup_{0 \leq s \leq a^j_n} [A_n(s, a^j_n) - r_n(a^j - s)]$, we get:

$$D^j \leq \sum_{n=1}^{N} \frac{\sup_{0 \leq s \leq a^j_n} [A_n(s, a^j_n) - r_n(a^j - s)]}{C_n}$$ \hspace{1cm} (6)

Since the traffic of each class has gSBB, with simple manipulation on the definition, we have,

$$P\{\sup_{0 \leq s \leq a^j_n} [A_n(s, a^j_n) - r_n(a^j - s)] + r_n \epsilon > \tau\} \leq \bar{F}_n(C_n \tau)$$

The theorem then follows from probability theory results on sum of random variables, together with applying $\epsilon \to 0$. \[\Box\]

In Th. [8] no intra-independence assumption on each $A_n(t)$ has been made. If the arrival processes satisfy some independence assumptions, the queue becomes $GI/D/1$ and improved delay bounds can be shown in Th. [4] (its proof is similar to that of Th. [4]).

In the remainder, in addition to waiting time and (system) delay, the concept of virtual waiting time is also used. The virtual waiting time at time $t$ is defined to be the time that a virtual packet/customer, which arrives immediately before time $t$, would experience: All arrivals at $t$ are excluded in the calculation of the virtual waiting time at $t$. 
Theorem 3. Suppose $A_1(t), \ldots, A_N(t)$ are independent and each has independent stationary increments. If there exists some small $\theta > 0$ such that $E[e^{\theta \sum_{n=1}^N A_n(t+1) - A_n(t)}] \leq 1$, then, for any packet $p^j$ and for all such $\theta$, the virtual waiting time at $a^j$ and the delay of the packet are respectively bounded by,

$$P\{W^j \geq \tau\} \leq e^{-\theta(\tau - 2\Delta)}$$

and its delay is bounded by (a.s.), \forall \tau \geq 2\Delta,

$$P\{D^j \geq \tau\} \leq 1 - e^{-\theta(\tau - 2\Delta)}$$

with $F_{W_j}(\tau) \equiv 1 - e^{-\theta(\tau - 2\Delta)}$ and $F_{W_j}(\tau)$ the CDF of the customer's service time $Y^j$.

Proof. Let $t^0$ denote the start of the busy period where the customer $p^j$ is in, and $e^j$ the time at which the queue can start serving the customer who arrives at $a^j$. Then:

$$e^j = t^0 + \sum_{n=1}^{N} S_n(t^0, a^n)$$

$$d^j = t^0 + \sum_{n=1}^{N} S_n(t^0, a^n_1) = e^j + Y^j$$

where $a^n_1$ is used such that $S_n(t^0, a^n_1)$ represents the system work that arrives in $[t^0, a^n]$.

We now map $a^n$ and $t^0$ to the discrete time system. Specifically, suppose $a^n$ falls between $[k \Delta, (k+1)\Delta)$ for some $K \geq 1$, and $t^0$ between $[k_0 \Delta, (k_0+1)\Delta)$ for some $k_0 \geq 0$. By these, we must have $k_0 < K$. In addition, $S_n(t^0, a^n) = S_n(k_0 \Delta, a^n)$, applying which to (11), the following inequality is further established for the waiting time of the customer, i.e. $W^j = e^j - a^j$:

$$W^j \leq \sum_{n=1}^{N} S_n(k_0 \Delta, K \Delta) - (K-k_0)\Delta + 2\Delta$$

$$\leq \sup_{0 \leq k < K} \sum_{n=1}^{N} S_n(k_\Delta, K \Delta) - (K-k)\Delta + 2\Delta$$

$$= V^j + 2\Delta$$

where the 2nd last step is to establish the inequality that holds for all sample paths, and in the last step,

$$V^j \equiv \sup_{0 \leq k < K} \sum_{n=1}^{N} S_n(k_\Delta, K \Delta) - (K-k)\Delta$$

Next, define $Z(k \Delta) = e^{\theta \sum_{n=1}^N S_n((k+1)\Delta, K \Delta) - k_0 \Delta}$. It can be proved that $\{Z(k \Delta) \}_{k = 1, 2, \ldots}$, forms a supermartingale if $M_{\sum_{n=1}^N S_n(\theta - \Delta)}(\theta) \leq 1$ for some small $\theta > 0$. Consequently, the following delay inequality is established for (15):

$$P\{V^j \geq \tau\} \leq E[Z(\Delta)]e^{-\theta \tau}$$

where $E[Z(\Delta)] = M_{\sum_{n=1}^N S_n(\Delta)}(\Delta)$.

Finally, the first part of the theorem, i.e. (10), follows immediately from (16) and (13). For the second part, the service time of the customer $p^j$ needs to be added to the waiting time. With the independence assumption, the service time of the customer is independent of the waiting time, and hence (10) is easily verified.

Remarks: (i) In obtaining (13), $W^j = e^j - a^j$ has been used. It is worth highlighting that, in general, $e^j - a^j$ is only the virtual waiting time at $a^j$ as used in Th. 3. For Th. 3 it equals the waiting time of the customer because of the Lévy process (continuous time) assumption that implies that, even under the multiclass FIFO setting, at any time, there is at most one customer arrival. If this assumption is not met and more than one customer may arrive at one time, $W^j$ and $D^j$ must take into consideration the service times of the concurrent arrivals at $a^j$ as shown in Th. 3. (ii) The bounds in Th. 4 and Th. 4 are similar to the Kingman’s bound in form and accuracy, e.g., for single-class $M/M/1$, by letting the unit time or $\Delta$ to 0. (ii) and (iii) give $e^{-(\mu - \lambda)\tau}$ that is the same as by the Kingman’s bound.

4. REFERENCES

[1] Y. Jiang. Performance bounds for multiclass FIFO in communication networks: A deterministic case. CoRR, abs/1306.4773, 2013.

[2] J.F.C. Kingman. A martingale inequality in the theory of queues. Proc. Camb. Phil. Soc., 59:359–361, 1964.

[3] R. L. Cruz. A calculus for network delay, part I: network elements in isolation. IEEE Trans. Information Theory, 37(1):114–131, Jan. 1991.

[4] H. Chen and H. Zhang. Stability of multiclass queueing networks under FIFO service discipline. Mathematics of Operations Research, 22(3):691–725, 1997.

[5] M. J. Karol, M. G. Hluchyj, and S. P. Morgan. Input versus output queueing on a space-division packet switch. IEEE Trans. Comm., 35(12):1347–1356, 1987.

[6] J.-Y. Le Boudec and P. Thiran. Network Calculus: A Theory of Deterministic Queueing Systems for the Internet. Springer-Verlag, 2001.

[7] Y. Jiang and Y. Liu. Stochastic Network Calculus. Springer-Verlag, 2008.

[8] Y. Jiang, Q. Yin, Y. Liu, and S. Jiang. Fundamental calculus on generalized stochastically bounded bursty traffic for communication networks. Computer Networks, 53(12):2011–2021, 2009.