Performance Bounds for Multiclass FIFO in Communication Networks: A Deterministic Case

Yuming Jiang

May 7, 2014

Abstract

Multiclass FIFO is used in communication networks such as in input-queueing routers/switches and in wireless networks. For the concern of providing service guarantees in such networks, it is crucial to have analytical results, e.g. bounds, on the performance of multi-class FIFO. Surprisingly, there are few such results in the literature. This paper is devoted to filling the gap. Specifically, a single hop deterministic case is studied, for which, delay and backlog bounds are derived, in addition to guaranteed rate and service curve characterizations that may be exploited to extend the analysis to network cases.

1 Introduction

Multiclass FIFO refers to the scheduling discipline where packets (or customers in the queueing theory jargon) are served in the first-in-first-out (FIFO) manner and the service rates to packets of different classes may differ.

Multiclass FIFO is perhaps the most intuitive scheduling discipline for many scenarios in communication networks. One example is input-queueing on a router/switch, where packets are FIFO-queued at each input port before being forwarded to the corresponding output ports that may have different capacities. Another example is downlink-sharing in wireless networks, where a wireless base station, shared by multiple users, sends packets to these users in the FIFO manner while the characteristics of the wireless channel seen by packets of different users may be different.

For the concern of providing service guarantees in such networks, it is crucial to have analytical results, e.g. bounds, on the performance of multi-class FIFO. Surprisingly, there are few such results in the literature. In addition, as to be shown in the paper, the bounds that can be directly deduced from the literature results require conditions that may be too restrictive, limiting the applicability
of these bounds. This paper is devoted to filling this gap. The focus is on a deterministic case. Specifically, both delay and backlog bounds are derived. In addition, the guaranteed rate and service curve characterizations for each class are also obtained, which may be further exploited to extend the analysis to network cases.

The rest is organized as follows. In the next section, the system model and performance metrics are introduced. In Sec. 2 the direct bounds are presented and their limitations discussed. In Sec. 3 improved bounds with much relaxed conditions are derived and discussed. Concluding remarks are given in Sec. 5.

2 The System Model

2.1 The System Model

We consider a work-conserving multiclass system serving packets in a communication network. It is a discrete-time system with time indexed by \( t = 0, 1, 2, \ldots \), where the length of the time unit is not specified, but is supposed to be small enough to even count one bit service time. There are \( N \) classes of packets. Packets are served in the FIFO manner. If multiple packets arrive at the same time, the tie is broken arbitrarily. The service rate (in bps) of each class, denoted by \( C_n \), is assumed to be constant. Let \( C_{\min} = \min_n \{C_n\} \) and \( C_{\max} = \max_n \{C_n\} \).

By convention, a packet is said to have arrived to (respectively served by) the system when and only when its last bit has arrived to (respectively left) the system. When a packet arrives seeing the system busy, the packet will be queued and the buffer size for the queue is assumed to be large enough ensuring no packet loss. The queue is initially empty.

For each class \( n \), let \( f_n \) denote its traffic flow and \( p_n^{i,j} \) the \( i \)th packet \( (i = 1, 2, \ldots) \) of the flow. For each packet \( p_n^{i,j} \), we denote by \( a_n^{i,j} \) its arrival time to the system, \( d_n^{i,j} \) its departure time from the system, and \( l_n^{i,j} \) its length (in bits). For the aggregate flow of \( f_n \), \( (n = 1, \ldots, N) \), denoted by \( g \), we use \( p^{j} \), \( a^{j} \), \( d^{j} \) and \( l^{j} \) to respectively denote the \( j \)th packet of the aggregate flow, its arrival time, its departure time and its length. Let \( L_n \) be the maximum packet length of flow \( f_n \) and \( L = \max_n \{L_n\} \).

In addition, we use \( A_n(s,t) \) and \( A^*_n(s,t) \) to respectively denote the amount of traffic (in bits) that has arrived and departed from flow \( f_n \) within time period \([s,t] \), and let \( A_n(t) = A_n(0,t) \) and \( A^*_n(t) = A^*_n(0,t) \). For the aggregate flow, \( A(t) \) and \( A^*(t) \) are similarly defined.

The traffic of each class is assumed to be constrained by a leaky-bucket arrival curve \[ a_n(t) = \]
or in other words: for all $t \geq s \geq 0$,

$$A_n(s, t) \leq \alpha_n(t - s) = r_n(t - s) + \sigma_n.$$  

2.2 Performance Metrics of Interest

In this paper, we focus on finding bounds on packet delay, backlog, guaranteed rate and service curve, which are defined as follows.

(i) The delay of packet $p^{f_n,i}$, denoted by $D^{f_n,i}$, is naturally

$$D^{f_n,i} = d^{f_n,i} - a^{f_n,i}. \quad (1)$$

(ii) The backlog at time $t$, denoted by $B(t)$, is defined as:

$$B(t) = A(t) - A^*(t).$$

(iii) A system is said to be a Guaranteed Rate (GR) server with rate $R_n$ and error term $E_n$ to traffic class $n$ or flow $f_n$, if it guarantees that for any packet $p^{f_n,i}$ of the flow, its departure time satisfies $[3]$:

$$d^{f_n,i} \leq GR^{f_n,i}(R_n) + E_n \quad (2)$$

where the guaranteed rate clock (GRC) function is iteratively defined as:

$$GR^{f_n,i}(R_n) = \max\{a^{f_n,i}, GR^{f_n,i-1}\} + \frac{I^{f_n,i}}{R_n}, \quad (3)$$

with $GR^{f_n,0} = 0$.

(iv) A system is said to provide a service curve $\beta_n(t)$ to traffic class $n$ or flow $f_n$ if there holds $[6]$:

for all $t \geq 0$,

$$A^*_n \geq A_n \otimes \beta_n(t) \quad (4)$$

where $A_n \otimes \beta_n(t) \equiv \inf_{0 \leq s \leq t}\{A_n(s) + \beta_n(t - s)\}$.

3 Direct Results and the Limitations

In this section, we introduce performance bounds that can be derived by making easy application of the literature results.

Lemma 1. The multiclass FIFO system is a Guaranteed Rate server with rate $R = C_{\text{min}}$ and error term $E = 0$ to the aggregate input.
Proof. For the multiclass FIFO system, there holds for every packet \( p^{g,j} \) of the aggregate flow, \( j = 1, 2, \ldots \),
\[
d^{g,j} = \max\{a^{g,j}, d^{g,j-1}\} + \frac{l^{g,j}}{C_*}
\]
where \( C_* \) is the service rate of the class of packet \( p^{g,j} \).
Since \( C_* \geq C_{\text{min}} \), it can be easily verified:
\[
d^{g,j} \leq GRC^{g,j}(C_{\text{min}})
\]
with which, the lemma follows from the definition of GR.

With Lemma 1, the following result follows from the relationship between GR and service curve 4 [6].

Lemma 2. The multiclass FIFO system provides to the aggregate input a service curve \( \beta(t) = (C_{\text{min}} \cdot t - L)^+ \).

With Lemmas 1 and 2, the following corollaries follow immediately from the related literature results. Specifically, Cor. 1, Cor. 2 and Cor. 4 follow respectively from the delay bound, backlog bound and leftover service results of the deterministic network calculus [1, 6]. In addition, Cor. 3 further follows from the relationship between GR and service curve [4, 6].

Corollary 1. (i) Delay Bound: If \( \sum_n r_n \leq C_{\text{min}} \), then the delay of any packet \( p^{g,j} \) is bounded by:
\[
D^{g,j} \leq \sum_n \frac{\sigma_n}{C_{\text{min}}}
\]

Corollary 2. (ii) Backlog Bound: If \( \sum_n r_n \leq C_{\text{min}} \), then the backlog at any time \( t \) is bounded by:
\[
B(t) \leq \sum_n \sigma_n + (\sum_n r_n) \frac{L}{C_{\text{min}}}
\]

Corollary 3. (iii) Guaranteed Rate to Each Traffic Class: For each traffic class \( n \), if \( \sum_{m \neq n} r_m < C_{\text{min}} \), then the system is a guaranteed rate server to it with rate \( R_n = C_{\text{min}} - \sum_{m \neq n} r_m \) and error term \( E_n = \frac{\sum_{m \neq n} \sigma_m + L}{C_{\text{min}} - \sum_{m \neq n} r_m} \).

Corollary 4. (iv) Service Curve to Each Traffic Class: For each traffic class \( n \), if \( \sum_{m \neq n} r_m < C_{\text{min}} \), then the system provides to it a service curve \( \beta_n(t) = [(C_{\text{min}} - \sum_{m \neq n} r_m) \cdot t - L - \sum_{m \neq n} \sigma_m]^+ \).

3.1 The Limitations

The alert reader may have noticed that the conditions for the performance bounds presented above may be too restrictive. Specifically, in obtaining the delay and backlog bounds, it has been required
that \( \sum_n r_n \leq C_{\text{min}} \). This requirement may greatly limit the applicability of these bounds particularly when \( C_n, n = 1, \ldots, N \), differ much. For a simple example where there are two classes, i.e. \( N = 2 \), suppose \( C_1 = 1 \) Mbps and \( C_2 = 100 \) Mbps. This requirement basically limits the total traffic rate of both classes to below 1 Mbps in order to apply the delay and backlog bounds shown above.

A similar limitation is found for the guaranteed rate and service curve to each traffic class \( n \), where, it is required that \( \sum_{m \neq n} r_m < C_{\text{min}} \), i.e., the total traffic rate of all other classes must be smaller than the minimum capacity \( C_{\text{min}} \). Take the two-class example again. Cor. \( \text{b} \) and Cor. \( \text{d} \) essentially suggest that the system only guarantees the 2nd class a surprisingly low long-term service rate \( 1 - r_1 \) Mbs, even though \( C_2 = 100 \) Mbs!

Due to these limitations, the directly obtained bounds may have limited use or are even in contrary to intuitions. In the next section, we aim to relax the limitations and derive better bounds.

4 Main Results

We start with the main results and then present their proofs, followed by a discussion comparing them with the direct results.

4.1 Summary of Main Results

Theorem 1. (i) Delay Bound: For the multiclass FIFO system, if \( \rho \equiv \sum_n \frac{r_n}{C_n} \leq 1 \), the delay of any packet \( p^{g,j} \) is bounded by:

\[
D^{g,j} \leq \sum_n \frac{\sigma_n}{C_n}
\]

and the delay bound is tight in the sense that it may be reached.

Theorem 2. (ii) Backlog Bound: For the multiclass FIFO system, if \( \rho \equiv \sum_n \frac{r_n}{C_n} \leq 1 \), the backlog at any time \( t \) is bounded by:

\[
B(t) \leq \min \left\{ \left( \sum_n r_n \right) \sum_n \frac{\sigma_n}{C_n} + \sum_n \sigma_n, \right. 
\]

\[
C_{\text{max}} \cdot \sum_n \frac{\sigma_n}{C_n} + \rho C_{\text{max}} \cdot \max_n \frac{L_n}{C_n} \right\}.
\]

Theorem 3. (iii) Guaranteed Rate to Each Traffic Class: If \( \bar{\rho} \equiv \sum_{m \neq n} \frac{r_m}{C_m} < 1 \), the system is a guaranteed rate server to class \( n \) with rate \( R_n = (1 - \bar{\rho})C_n \) and error term

\[
E_n = \sum_{m \neq n} \frac{\sigma_m}{(1 - \bar{\rho})C_m}.
\]
Theorem 4. (iv) Service Curve to Each Traffic Class: If \( \bar{\rho} \equiv \sum_{m \neq n} \frac{\bar{r}_m}{C_m} < 1 \), the system provides to class \( n \) a service curve

\[
\beta_n(t) = [(1 - \bar{\rho})C_n \cdot t - L_n - C_n \cdot \sum_{m \neq n} \frac{\sigma_m}{C_m}]^+.
\]

4.2 Proofs

4.2.1 Proof of Theorem 1

For any packet \( p_{g,j} \), there exists time \( t_0 \) that starts the busy period of which the packet is in. Note that such a busy period always exists, since in the extreme case, the period is only the service time period of \( p_{g,j} \) and in this case, \( t^0 = a_{g,j} \).

Since the system is work-conserving and is busy with serving between \( t_0 \) and \( d_{g,j} \), there holds:

\[
d_{g,j} = t_0 + \sum_{m=1}^{N} T_m(t_0, d_{g,j}),
\]

where \( T_m(t_0, d_{g,j}) \) denotes the service time of packets from class \( m \), served in \([t_0, d_{g,j}]\). Because of FIFO and that the system is empty at \( t_{g,j} \), \( T_m(t_0, d_{g,j}) \) is hence limited by the amount of traffic that arrives in \([t_0, a_{f,n,i}]\):

\[
T_m(t_0, d_{g,j}) \leq \frac{A_m(t_0, a_{g,j})}{C_m}.
\]

We then have,

\[
d_{g,j} \leq t_0 + \sum_{m=1}^{N} \frac{A_m(t_0, a_{g,j})}{C_m},
\]

(6)

Under the condition that \( \sum_{m=1}^{N} \frac{r_m}{C_m} \leq 1 \), we obtain:

\[
D_{g,j} = d_{g,j} - a_{g,j} \leq \sum_{m=1}^{N} \frac{A_m(t_0, a_{g,j})}{C_m} + t_0 - a_{g,j}
\]

\[
\leq \sum_{m=1}^{N} \frac{r_m(a_{g,j} - t_0) + \sigma_m}{C_m} - (a_{g,j} - t_0)
\]

(7)

(8)

Note that, for the system, consider that immediately after time 0, every traffic class generates a burst with size \( \sigma_m \). In this case, the packet in the bursts, which receives service last, will experience delay \( \sum_{m=1}^{N} \frac{\sigma_m}{C_m} \) that equals the delay bound. So, the bound is tight.

4.2.2 Proof of Theorem 2

The backlog bound has two parts.

The first part follows from Th. 1 that the system provides to the aggregate a service curve \( \bar{\beta}(t) = \bar{\delta} \sum_m \frac{\bar{r}_m}{C_m}(t) \) [6], where \( \bar{\delta}(t) = 0 \) if \( 0 \leq t \leq D \) and \( \bar{\delta}(t) = \infty \) when \( t > D \). With this and that
the aggregate input is constrained by an arrival curve as: \( A(t) \leq (\sum_n r_n) \cdot t + \sum_n \sigma_n \equiv \alpha(t) \). Then, the first part follows from the delay bound result of the deterministic network calculus [6].

To obtain the second part, we propose the following mapping method. Recall that, the system behavior is essentially determined by (5), which can be re-written as

\[
d^{g,j} = \max\{a^{g,j}, d^{g,j-1}\} + \hat{l}^{g,j}
\]

where \( \hat{l}^{g,j} \equiv \frac{l^{g,j}}{C_n} \) and \( C_n \) is the service rate of the class of \( p^{g,j} \).

Equation (5) suggests that the multiclass FIFO system is equivalent to a reference FIFO system that has constant unit service rate for every packet whose length is equal to that in the original multiclass FIFO system normalized by the service rate of its corresponding class. Under this normalization, it can be verified that: \( \hat{A}_n(s, t) = \frac{A_n(s, t)}{C_n} \) and \( \hat{A}_n^*(s, t) = \frac{A_n^*(s, t)}{C_n} \) for all \( n, s, t \), and \( \hat{B}(t) = \sum_n \frac{B_n(t)}{C_n} \), where \( \hat{A}_n(s, t) \) and \( \hat{A}_n^*(s, t) \) respectively denote the (normalized) arrival and departure processes in the reference FIFO system and \( \hat{B}(t) \) the backlog at time \( t \) in the reference FIFO system.

Since the reference FIFO system has constant unit service rate with maximum packet length \( \max_n \frac{l_n}{C_n} \), it has a service curve as \( \hat{\beta}(t) = (t - \max_n \frac{l_n}{C_n})^+ \). Since its input is constrained by a leaky-bucket arrival curve as \( \hat{A}(s, t) = \sum_n \hat{A}_n(s, t) \leq \sum_n \frac{r_n}{C_n}(t - s) + \sum_n \frac{\sigma_n}{C_n} \), the literature deterministic network calculus results give \( \hat{B}(t) \leq \sum_n \frac{\sigma_n}{C_n} + \sum_n \frac{r_n}{C_n} \max_n \frac{l_n}{C_n} \). The second part is finally obtained from \( B(t) = C_{\max} \sum_n \frac{B_n(t)}{C_{\max}} \leq C_{\max} \sum_n \frac{B_n(t)}{C_n} = C_{\max} \cdot \hat{B}(t) \). \( \square \)

### 4.2.3 Proof of Theorem [3]

For any packet \( p^{f_n,i} \) of class \( n \), we suppose its departure time \( d^{f_n,i} \) is within the busy period that starts at \( t^0 \) and suppose \( p^{f_n,i_0} \) is the first class \( n \) packet served in this busy period. By these definitions, \( a^{f,j_0} \geq t^0 \). Note that this busy period always exists, since in the worst case, the period is only the service time period of \( p^{f_n,i} \) and in this case, \( t^0 = a^{f,n,i} \) and \( i_0 = i \).

Since the busy period starts at \( t^0 \), this implies that immediately before \( t^0 \), the system is idle. In other words, all packets, which arrived before \( t^0 \), have been served by \( t^0 \). Additionally since the node is work-conserving, FIFO and busy with serving between \( t^0 \) and \( d^{g,j} \), there holds:

\[
d^{f_n,i} \leq t^0 + \sum_{k=i_0}^{i} \frac{l^{f_n,k}}{C_n} + \sum_{m \neq n} \frac{A_m(t^0, a^{f_n,i})}{C_m},
\]

where, by definition, \( A_m(t^0, a^{f_n,i}) \) represents the total length (in bits) of packets from another flows.
m \neq n$, which can maximally been served in $[t^0, a^{f_{n,i}}]$ before packet $p^{f_{n,i}}$. We then have:

\[ d^{f_{n,i}} \leq t^0 + \sum_{k=0}^{i} \frac{t^{f_{n,k}}}{C_n} + \sum_{m \neq n} r_m (a^{f_{n,i}} - t^0) + \sigma_m \]

\[ = t^0 + \sum_{k=0}^{i} \frac{t^{f_{n,k}}}{C_n} + \sum_{m \neq n} \frac{C_m}{C_n} + \sum_{m \neq n} r_m (a^{f_{n,i}} - t^0) \]

\[ \leq t^0 + \sum_{k=0}^{i} \frac{t^{f_{n,k}}}{C_n} + \sum_{m \neq n} \frac{C_m}{C_n} + \sum_{m \neq n} \frac{r_m (d^{f_{n,i}} - t^0)}{C_m} \]

\[ \leq t^0 + \sum_{k=0}^{i} \frac{t^{f_{n,k}}}{(1 - \bar{\rho}) C_n} + \sum_{m \neq n} \frac{\sigma_m}{(1 - \bar{\rho}) C_m} \]

where \( \bar{\rho} = \sum_{m \neq n} \frac{C_m}{C_n} \) and the last step follows from some manipulation together with the left hand side.

Recall function (3), it is easily verified that:

\[ V^{f_{n,i}}((1 - \bar{\rho}) C_n) \geq a^{f_{n,i0}} + \sum_{k=0}^{i} \frac{t^{f_{n,k}}}{(1 - \bar{\rho}) C_n} \geq t_0 + \sum_{k=0}^{i} \frac{t^{f_{n,k}}}{(1 - \bar{\rho}) C_n} \]

Hence

\[ d^{f_{n,i}} \leq V^{f_{n,i}}((1 - \bar{\rho}) C_n) + \sum_{m \neq n} \frac{\sigma_m}{(1 - \bar{\rho}) C_m}. \]

This ends the proof.

\[ \square \]

4.2.4 Proof of Theorem 4

With Th. 3, Th. 4 follows immediately from the relationship between GR and service curve [4, 6].

4.3 Comparison

Comparing with Cor. 1 – Cor. 4, Th. 1 – Th. 4 provide performance bounds better in two aspects. One is that the conditions for the latter are much relaxed. The new condition for the delay and backlog bounds, i.e. \( \rho < 1 \), is indeed consistent with the multiclass FIFO stability condition found from the classic queueing theory (e.g. [2]).

Another is that, the obtained bounds are also generally better. Specifically, the delay bound in Th. 4 is tight in the sense that it may be reached by a packet as shown in the proof. In other words, there is no delay bound better than Th. 4 which may be found. In addition, it can be verified that the service curve obtained in Th. 4 is always tighter than that in Cor. 4. Also, the same conclusion can be made for the guaranteed rate characterization.

For backlog bound, the comparison is not straightforward. When \( C_1 = \cdots = C_N \), it is easily verified that the bound in Th. 2 may only be slightly better due to the difference between \( \max_{n} \frac{L_n}{C_n} \)
and \( \frac{C}{C_{\text{min}}} \). In general, when \( C_n \) differs little, the bound in Th. 2 is better. However, when \( C_n \) differs significantly, the bound in Cor. 2 may become better but under the condition that \( \sum_n r_n \leq C_{\text{min}} \). If this condition is not met, then the bound in Th. 2 should be used.

5 Concluding Remarks

In this paper, we have presented performance bounds for a multiclass FIFO system that has deterministically constrained traffic inputs. The analysis and results can be extended to obtain bounds for other performance metrics of interest, such as burstiness increase. Also, they can be extended to perform network case analysis by exploiting the GR and the service curve results. Moreover, the extension may be conducted for stochastic cases, which is our ongoing work.

A final remark is that, many results for multiclass FIFO can be found under the classic queueing theory (see e.g. [2] and references therein). However, the arrival and service processes assumed in those results do not match with the traffic constraints considered in this paper. In the context of input-queueing in packet switching systems, multiclass FIFO has also been studied (e.g. [5]). However, in those studies, the focus has been on the throughput of the switch, assuming saturated traffic on each input port.

References

[1] C.-S. Chang. Performance Guarantees in Communication Networks. Springer-Verlag, 2000.
[2] H. Chen and H. Zhang. Stability of multiclass queueing networks under fifo service discipline. Mathematics of Operations Research, 22(3):691–725, 1997.
[3] P. Goyal, S. S. Lam, and H. M. Vin. Determining end-to-end delay bounds in heterogeneous networks. Multimedia Systems, 5:157–163, 1997.
[4] Y. Jiang. Relationship between guaranteed rate server and latency rate server. Computer Networks, 43(3):307–315, 2003.
[5] M. J. Karol, M. G. Hluchyj, and S. P. Morgan. Input versus output queueing on a space-division packet switch. IEEE Transactions on Communications, 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.