

Lecture Notes 5

EE 549 — Queueing Theory

Instructor: Michael Neely

I. BACKLOG AND DELAY BOUNDS FOR LEAKY BUCKET (r, σ) INPUTS

Recall that $X(t) \sim (r, \sigma)$ if $X[t_1, t_2] \leq r(t_2 - t_1) + \sigma$ for all intervals $[t_1, t_2]$. Consider a single server work conserving queue, and assume:

- $X(t) \sim (r, \sigma)$ input stream
- single server queue
- constant server rate μ
- initially empty system ($U(0) = 0$)

Fact 1: (Queue Bound) If $r \leq \mu$, then $U(t) \leq \sigma$ for all time t .

Proof: Fix a particular time t . We want to show that $U(t) \leq \sigma$. If $U(t) = 0$, then we are done. Else, $U(t) > 0$; and so we are in a busy period that started at some time t_b , where $t_b \leq t$. Thus:

$$\begin{aligned} U(t) &= X[t_b, t] - \mu \cdot (t - t_b) \\ &\leq r(t - t_b) + \sigma - \mu \cdot (t - t_b) \\ &\leq \sigma \end{aligned}$$

■

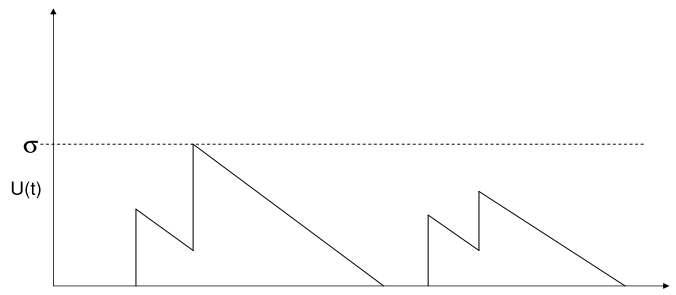


Fig. 1. An illustration of the unfinished work bound for leaky bucket inputs.

Fig. 1 illustrates the bound on unfinished work.

Corollary 1: (Delay Bound) If service is FIFO then delay $\leq \frac{\sigma}{\mu}$.

Proof: Take any packet i . Let t_i be the time of the packet arrival, and let $U(t_i)$ be the unfinished work at this time (which includes all bits that were in the queue just before the packet arrived, plus the total number of bits in packet i). Then under FIFO, the delay of packet i is exactly $U(t_i)/\mu$, which is less than or equal to σ/μ . ■

Fact 2: (Summing Leaky Bucket Inputs) If $X_1(t) \sim (r_1, \sigma_1)$ and $X_2(t) \sim (r_2, \sigma_2)$, then $X_1(t) + X_2(t) \sim (r_1 + r_2, \sigma_1 + \sigma_2)$.

Proof:

$$\begin{aligned} X[t, t+T] &= X_1[t, t+T] + X_2[t, t+T] \\ &\leq (r_1T + \sigma_1) + (r_2T + \sigma_2) \\ &= (r_1 + r_2)T + (\sigma_1 + \sigma_2) \end{aligned}$$

■

Fact 3: (Input-Output Parameter Invariance) Consider a leaky bucket input stream $X(t)$ with parameters (r, σ) entering a single server work conserving queue with constant processing rate μ bits/sec. Assume $r \leq \mu$, and let $Y(t)$ represent the departure process. Then $Y(t)$ is leaky bucket with parameters (r, σ) .

Proof: We want to show that for any time t and any interval duration $T \geq 0$, we have:

$$Y[t, t + T] \leq rT + \sigma$$

To show this, fix any time t and any $T \geq 0$, and consider the interval $[t, t + T]$. Note that the maximum amount of bits that depart over this interval is no more than $U(t)$ plus the total new bits that enter during $(t, t + T]$:

$$Y[t, t + T] \leq U(t) + X(t, t + T) \quad (1)$$

We now have two cases:

- Case 1: $U(t) = 0$. In this case, we have from (1):

$$Y[t, t + T] \leq 0 + X(t, t + T) \leq rT + \sigma$$

and we are done.

- Case 2: $U(t) > 0$. In this case, there is a time $t_b \leq t$ (at which the current busy period started) such that:

$$U(t) = X[t_b, t] - \mu \cdot (t - t_b) \quad (2)$$

Using (1) and (2)

$$\begin{aligned} Y[t, t + T] &\leq X[t_b, t] - \mu \cdot (t - t_b) + X(t, t + T) \\ &= X[t_b, t + T] - \mu \cdot (t - t_b) \\ &\leq r(t + T - t_b) + \sigma - \mu \cdot (t - t_b) \\ &= rT + \sigma - (\mu - r) \cdot (t - t_b) \\ &\leq rT + \sigma \end{aligned}$$

■

Note that Fact 3 does not say that the output process $Y(t)$ is the same as $X(t)$, it just says that it has the same leaky bucket parameters as $X(t)$.

A. Packet-Based Departures

The Fact 3 is useful for multi-stage networks if we have a *fluid departure model*, where the fluid bit stream $Y(t)$ is considered the output that enters the next stage (so that data enters the next stage “one bit at a time”). A more realistic departure model is the *packet based model*, where we say a packet does not depart until its last bit departs. Let $\tilde{Y}(t)$ represent the *packet based* bit departure process, being a non-decreasing staircase function that satisfies $\tilde{Y}(t) \leq Y(t)$ for all time t , and $\tilde{Y}(t_d) = Y(t_d)$ for any time t_d that is a packet departure time. We have the following modifications of Fact 3:

Fact 3b: If $X(t) \sim (r, \sigma)$ and enters a single-server work conserving queue with a constant service rate μ with *non-preemptive* service (such as FIFO, LIFO, non-preemptive priority, etc.), then if $r \leq \mu$ we have:

$$\tilde{Y}(t) \sim (r, \sigma + B_{max})$$

where B_{max} is an upper bound on the maximum packet size.

Proof: Because service is non-preemptive, there is at most one partially processed packet in the system at any time. Now fix a particular interval $[t_1, t_2]$, and note that:

$$\tilde{Y}[t_1, t_2] \leq Y[t_1, t_2] + r(t_1)$$

where $r(t_1)$ represents the residual amount of bits in the packet (if any) that was partially processed before time t_1 , but which departed during the interval $[t_1, t_2]$. It follows that $r(t_1) \leq B_{max}$, and hence: $\tilde{Y}[t_1, t_2] \leq Y[t_1, t_2] + B_{max}$. However, we know that $Y[t_1, t_2] \leq r \cdot (t_2 - t_1) + \sigma$, and hence:

$$\tilde{Y}[t_1, t_2] \leq r \cdot (t_2 - t_1) + \sigma + B_{max}$$

This is true for any interval $[t_1, t_2]$, and hence $\tilde{Y}(t) \sim (r, \sigma + B_{max})$. ■

Thus, the packet based departure model adds an extra B_{max} to the burstiness of the output. It turns out that if packets are *fixed length* with some constant size B bits, then the burstiness of the output process does not increase, as characterized below:

Fact 3c: Under the same assumptions of Fact 3b, if all packets have a fixed length of size B bits, and if $r \leq \mu$, then the packet-based departure process $\tilde{Y}(t)$ satisfies:

$$\tilde{Y}(t) \sim (r, \sigma)$$

Proof: See Appendix A ■

B. Multiple Output Streams

Fact 4: Consider two leaky bucket inputs $X_1(t)$ and $X_2(t)$ with parameters (r_1, σ_1) and (r_2, σ_2) that enter a single server work conserving queue with constant server rate μ . Let $Y_1(t)$ and $Y_2(t)$ represent the corresponding bit departure processes from the queue due to type 1 and type 2 data, respectively (where $Y(t) = Y_1(t) + Y_2(t)$ is the full bit departure process). If $r_1 + r_2 \leq \mu$, then:

$$Y_1(t) \sim (r_1, \sigma_1 + \sigma_2)$$

$$Y_2(t) \sim (r_2, \sigma_1 + \sigma_2)$$

Proof: Fix any two times t_1, t_2 such that $t_1 \leq t_2$. We want to show that $Y_1[t_1, t_2] \leq r_1(t_2 - t_1) + \sigma_1 + \sigma_2$. First note that:

$$Y_1[t_1, t_2] \leq U(t_1) + X_1(t_1, t_2) \tag{3}$$

- Case 1: $U(t_1) = 0$. In this case, we have:

$$Y_1[t_1, t_2] \leq 0 + X_1(t_1, t_2) \leq r_1(t_2 - t_1) + \sigma_1$$

and so we are done (note that $\sigma_1 \leq \sigma_1 + \sigma_2$).

- Case 2: $U(t_1) > 0$. In this case, there must be a time $t_b \leq t_1$ such that:

$$U(t_1) = X_1[t_b, t_1] + X_2[t_b, t_1] - \mu(t_1 - t_b)$$

Using this equality in (3) yields:

$$\begin{aligned} Y_1[t_1, t_2] &\leq X_1[t_b, t_1] + X_2[t_b, t_1] - \mu(t_1 - t_b) + X_1(t_1, t_2) \\ &= X_1[t_b, t_2] + X_2[t_b, t_1] - \mu(t_1 - t_b) \\ &\leq r_1(t_2 - t_b) + \sigma_1 + r_2(t_1 - t_b) + \sigma_2 - \mu(t_1 - t_b) \\ &= r_1(t_2 - t_1) + \sigma_1 + \sigma_2 + (r_1 + r_2 - \mu)(t_1 - t_b) \\ &\leq r_1(t_2 - t_1) + \sigma_1 + \sigma_2 \end{aligned}$$

■

C. Fluid Flow Networks: Example

Consider the system in Figure 2. Stability conditions:

$$r_1 \leq \mu_A, \quad r_1 + r_2 + r_3 \leq \mu_B$$

Assuming the above conditions hold, then under the fluid departure model:

$$U_A(t) \leq \sigma_1, U_B(t) \leq \sigma_1 + \sigma_2 + \sigma_3$$

$$\text{delay in A} \leq \frac{\sigma_1}{\mu_A}, \quad \text{delay in B} \leq \frac{\sigma_1 + \sigma_2 + \sigma_3}{\mu_B}$$

$$\text{End to end delay of data from stream 1} \leq \frac{\sigma_1}{\mu_A} + \frac{\sigma_1 + \sigma_2 + \sigma_3}{\mu_B}$$

$$\text{End to end delay of data from streams 2 or 3} \leq \frac{\sigma_1 + \sigma_2 + \sigma_3}{\mu_B}$$

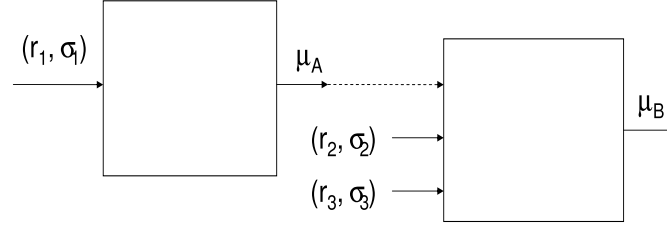


Fig. 2. An example tandem of queues.

Under a packet-based departure model, the backlog and delay in queue A does not change. The backlog and delay in queue B would satisfy (where B_{max} is the maximum packet size of stream 1 packets):

$$U_B(t) \leq \sigma_1 + B_{max} + \sigma_2 + \sigma_3$$

$$\text{Delay in B} \leq (\sigma_1 + B_{max} + \sigma_2 + \sigma_3)/\mu_B$$

and so the end-to-end delay bounds under the packet-based departure model would increase by B_{max}/μ_B .

II. TIME VARYING SERVER RATES

Definition 1: A time-varying server process $\mu(t)$ has a $(\bar{\mu}, \gamma)$ service envelope if over any interval $[t, t+T]$ we have:

$$\int_t^{t+T} \mu(\tau) d\tau \geq \bar{\mu}T - \gamma$$

γ is known as the “service lag”.

Consider a system with a (r, σ) arrival process and a server process $\mu(t)$ with a $(\bar{\mu}, \gamma)$ service envelope.

Fact 5: (Queue Bound for Varying Servers) If the system is initially empty and work conserving, and if $r \leq \bar{\mu}$, then:

$$U(t) \leq \sigma + \gamma$$

Proof: Fix any time t . If $U(t) = 0$ then we are done. Else, there exists a time $t_b \leq t$ such that:

$$U(t) = X[t_b, t] - \int_{t_b}^t \mu(\tau) d\tau$$

Thus:

$$\begin{aligned} U(t) &\leq r(t - t_b) + \sigma - \int_{t_b}^t \mu(\tau) d\tau \\ &\leq r(t - t_b) + \sigma - [\bar{\mu} \cdot (t - t_b) - \gamma] \\ &\leq \sigma + \gamma \end{aligned}$$

Fact 6: (Departures for Time Varying Servers) If the system is initially empty and work conserving, and if $r \leq \bar{\mu}$, then the departure process $Y(t)$ is leaky bucket constrained with parameters $(r, \sigma + \gamma)$. ■

Proof: The proof is almost exactly the same as the proof of Fact 3, and is omitted for brevity. ■

Fact 7: If an input process $X(t)$ with leaky bucket parameters (r, σ) enters a work conserving queue with service envelope $(\bar{\mu}, \gamma)$ and $r \leq \bar{\mu}$, and if service is FIFO, then:

$$\text{Delay} \leq \frac{\sigma + \gamma}{\bar{\mu}}$$

Proof: I leave this proof as an interesting exercise. Hint: Note that for any packet i that arrives at time t_i , we have under FIFO:

$$\int_{t_i}^{t_i + \text{Delay}_i} \mu(\tau) d\tau = U(t_i)$$

It is easy to prove that $\text{Delay}_i \leq (\sigma + 2\gamma)/\bar{\mu}$ from the above equality, but the tighter bound $\text{Delay}_i \leq (\sigma + \gamma)/\bar{\mu}$ can also be proven with just a little more work. ■

Fact 8: (I-O Relationship for Two Arrival Streams in a Time Varying Server) Let $X_1(t)$ and $X_2(t)$ be two arrival processes that enter a single-server, work conserving queue with transmission rate process $\mu(t)$. Suppose that: $X_1(t) \sim (r_1, \sigma_1)$, $X_2(t) \sim (r_2, \sigma_2)$, and $\mu(t) \sim (\bar{\mu}, \gamma)$. If $r_1 + r_2 \leq \bar{\mu}$ then:

$$Y_1(t) \sim (r_1, \sigma_1 + \sigma_2 + \gamma)$$

$$Y_2(t) \sim (r_2, \sigma_1 + \sigma_2 + \gamma)$$

Further, for packet-based departures, we have:

$$\tilde{Y}_1(t) \sim (r_1, \sigma_1 + \sigma_2 + \gamma + B_{max})$$

$$\tilde{Y}_2(t) \sim (r_2, \sigma_1 + \sigma_2 + \gamma + B_{max})$$

Proof: The proof is similar to previous proofs and is omitted for brevity. ■

Example: In Figure 3,

$$\mu_A(t) \sim (\bar{\mu}, \gamma) \text{ service envelope}$$

$$\mu_B \sim \text{constant service rate}$$

Conditions for stability:

$$r_1 + r_2 \leq \bar{\mu}_A$$

$$r_1 + r_2 \leq \mu_B$$

Suppose the above conditions hold, and that we have a fluid flow departure model. Because the combined input to queue A is leaky bucket with parameters $(r_1 + r_2, \sigma_1 + \sigma_2)$, we have:

$$U_A(t) \leq \sigma_1 + \sigma_2 + \gamma$$

Further, because the output $Y(t)$ at queue A is leaky bucket with parameters $(r_1 + r_2, \sigma_1 + \sigma_2 + \gamma)$, we have:

$$U_B(t) \leq \sigma_1 + \sigma_2 + \gamma$$

If service in all queues is FIFO, then from Fact 7 we have:

$$\text{Delay}_1 \leq (\sigma_1 + \sigma_2 + \gamma)/\bar{\mu}_A$$

$$\text{Delay}_2 \leq (\sigma_1 + \sigma_2 + \gamma)/\mu_B$$

Therefore, the worst case end-to-end delay under FIFO (and the fluid departure model) is less than or equal to:

$$\text{End-to-end Delay} \leq (\sigma_1 + \sigma_2 + \gamma) \left[\frac{1}{\bar{\mu}_A} + \frac{1}{\mu_B} \right]$$

In a packet-based departure model, we have:

$$U_B(t) \leq \sigma_1 + \sigma_2 + \gamma + B_{max}$$

and end-to-end delay under a FIFO model would be bounded as follows:

$$\text{End-to-end Delay} \leq \frac{(\sigma_1 + \sigma_2 + \gamma)}{\bar{\mu}_A} + \frac{\sigma_1 + \sigma_2 + \gamma + B_{max}}{\mu_B}$$

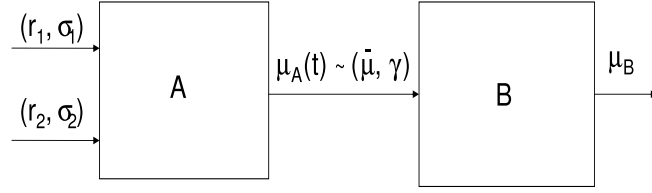


Fig. 3. An example tandem with a time varying input.

III. PERFORMANCE WITH PRIORITY

Consider two input streams entering a system with constant server rate μ :

$$\begin{aligned} X_1(t) &\sim (r_1, \sigma_1) \\ X_2(t) &\sim (r_2, \sigma_2) \end{aligned}$$

Definition 2: Packets from stream X_1 have *preemptive priority* over packets from stream X_2 if they are always placed before X_2 data in the buffer, and an arriving X_1 packet can interrupt service of any X_2 packet currently in the server. We assume that the interrupted X_2 packet can resume processing its residual service time when all X_1 data is gone. (This is sometimes called *preempt-resume* service).

Let X_1 packets have preemptive priority over X_2 packets. Effectively, stream X_1 does not “see” stream X_2 . In this case:

$$\begin{aligned} U_{X_1}(t) &\leq \sigma_1 \\ U_{X_2}(t) &\leq U_{X_1}(t) + U_{X_2}(t) \\ &\leq \sigma_1 + \sigma_2 \end{aligned}$$

where $U_1(t)$ and $U_2(t)$ respectively represent the unfinished work due to type 1 data and type 2 data. If service is FIFO:

$$\text{Delay}_1 \leq \frac{\sigma_1}{\mu}$$

What is the resulting delay for X_2 data?

It is tempting to say that $\text{Delay}_2 \leq \frac{\sigma_1 + \sigma_2}{\mu}$. However, this is not necessarily the case, as data from stream 2 may be preempted several times by the X_1 traffic. To understand how the low priority data views the queue, we define the notion of an *effective server process*.

Let $\mu_2(t)$ be the effective service rate seen by the low priority X_2 stream.

$$\mu_2(t) \triangleq \begin{cases} \mu & \text{if } X_1 \text{ data is not being served at time } t \\ 0 & \text{if } X_1 \text{ data is being served at time } t \end{cases}$$

Fact 9: If $r_1 + r_2 \leq \mu$, then $\mu_2(t)$ has service envelope $(\mu - r_1, \sigma_1)$.

Proof: Consider a time interval $[t, t + T]$. We need to show that:

$$\int_t^{t+T} \mu_2(t) d\tau \geq (\mu - r_1)T - \sigma_1$$

Let $Y_1(t)$ be the total number of bits of stream 1 processed in $[t, t + T]$. Then,

$$\int_t^{t+T} \mu_2(t) d\tau = \mu T - Y_1(t) \tag{4}$$

Also, we know

$$Y_1(t) \sim (r_1, \sigma_1) \quad (5)$$

Using (4) and (5),

$$\begin{aligned} \int_t^{t+T} \mu_2(\tau) d\tau &\geq \mu T - (r_1 T + \sigma_1) \\ &= (\mu - r_1)T + \sigma_1 \end{aligned}$$

■

It follows from Facts 9 and 7 that for the example of two leaky bucket streams with parameters $(r_1, \sigma_1), (r_2, \sigma_2)$ entering a single server queue where $r_1 - r_2 \leq \mu$ and where X_1 data has preemptive priority:

$$\text{Delay}_2 \leq \frac{\sigma_1 + \sigma_2}{\bar{\mu} - r_1}$$

A similar network calculus can be derived for non-preemptive priority (where priority data is placed first in the buffer, but cannot interrupt a packet currently in service) as well as for non-fluid models where departures take place as packets.

A. Multiple input streams with preemptive priority

Consider multiple input streams $X_i(t)$ (for $i \in \{1, \dots, n\}$), and assume each stream i is leaky bucket constrained with rate and burst parameters (r_i, σ_i) . All streams enter a single server work conserving queue with constant service rate μ . Assume that $\sum_{i=1}^n r_i \leq \mu$, so that the system is rate stable. Further assume that service follows a preemptive priority service rule, with priority as follows:

$$\text{Priority}(X_1) > \text{Priority}(X_2) > \dots > \text{Priority}(X_n)$$

so that X_1 data preempts all other data, X_2 data preempts all other data except for X_1 , etc. Assume that service within the same stream of data is FIFO (so that a packet from stream k that enters the queue first will be served before a packet from stream k that arrives later).

Note that for any $k \in \{1, \dots, n\}$, data from stream X_k is completely unaffected by streams $(k+1)$ or higher. Thus, from the above delay analysis, is it not difficult to show that for any $k \in \{1, \dots, n\}$:

$$\begin{aligned} U_k(t) &\leq \sum_{i=1}^k \sigma_i \\ \text{Delay}_k &\leq \frac{\sum_{i=1}^k \sigma_i}{\mu - \sum_{i=1}^{k-1} r_i} \end{aligned}$$

Exercise: Let $\mu = 1$ kilobit/sec. Let $\sigma_i = 1$ kilobit for $i \in \{1, \dots, 4\}$. Let $r_i = 0.2$ kilobits/sec for $i \in \{1, \dots, 4\}$. Thus, there are 4 leaky bucket input streams with identical rate and burst parameters. Assuming preemptive priority as above, make a bar plot of the worst case delay for streams $i \in \{1, \dots, 4\}$.

IV. SUMMARY OF NETWORK CALCULUS INEQUALITIES

Consider a queue with an input process $X(t)$ (leaky bucket constrained with parameters (r, σ)) entering a single server, work conserving queue with a time varying server rate $\mu(t)$ (with service envelope parameters $(\bar{\mu}, \gamma)$). If $r \leq \bar{\mu}$, then:

$$U(t) \leq \sigma + \gamma \quad (6)$$

$$\text{Delay}_{FIFO} \leq \frac{\sigma + \gamma}{\bar{\mu}} \quad (7)$$

$$\text{Output : } Y(t) \sim (r, \sigma + \gamma) \quad (8)$$

Note that the lag parameter γ satisfies $\gamma = 0$ iff the service rate is a constant μ for all time (except possibly for a set of times of measure 0). Hence, the above properties can be viewed as containing the special case of a constant rate server. Indeed, if $\mu(t) = \mu$ and $\gamma = 0$, the above bounds become: $U(t) \leq \sigma$, $\text{Delay}_{FIFO} \leq \sigma/\mu$ and $Y(t) \sim (r, \sigma)$.

A. Fluid Departures

Let two streams $X_1(t)$ and $X_2(t)$ enter a work conserving queue with time varying server rate $\mu(t)$. Suppose that: $X_1(t) \sim (r_1, \sigma_1)$, $X_2(t) \sim (r_2, \sigma_2)$, $\mu(t) \sim (\bar{\mu}, \gamma)$. If $r_1 + r_2 \leq \bar{\mu}$, then:

$$Y_1(t) \sim (r_1, \sigma_1 + \sigma_2 + \gamma) \quad (9)$$

$$Y_2(t) \sim (r_2, \sigma_1 + \sigma_2 + \gamma) \quad (10)$$

B. Packet Based Departures

Let B_{max} be the maximum packet size. Assume service is non-preemptive, so there is at most one partially processed packet in the system at any time. The equation (8) under packet-based departures becomes:

$$\tilde{Y}(t) \sim (r, \sigma + B_{max} + \gamma) \quad (11)$$

The equations (9) and (10) under the packet-based departure model become:

$$\tilde{Y}_1(t) \sim (r_1, \sigma_1 + \sigma_2 + B_{max} + \gamma)$$

$$\tilde{Y}_2(t) \sim (r_2, \sigma_1 + \sigma_2 + B_{max} + \gamma)$$

If the server rate is constant (so that $\mu(t) = \mu$ for all t , and so $\gamma = 0$), and if all packets have a fixed size of B , then inequality (11) can be improved to:

$$\tilde{Y}(t) \sim (r, \sigma)$$

APPENDIX A – PROOF OF FACT 3C

Here we prove Fact 3c. Suppose we have a single-server, work conserving queue with a constant transmission rate μ . Let $X(t)$ be a leaky bucket input with parameters (r, σ) where $r \leq \mu$. Suppose that all packets have fixed packet size B , and that service is non-preemptive. Note that we must have $B \leq \sigma$.

Let $T = B/\mu$ be the service time of a packet. Note that departures are spaced apart by at least T seconds. Consider any interval $[t_1, t_2]$ (where $t_1 \leq t_2$). We want to show that $\tilde{Y}[t_1, t_2] \leq r(t_2 - t_1) + \sigma$.

- Case 1: If $t_2 < t_1 + T$, then at most one packet could have departed over the interval $[t_1, t_2]$ (as the minimum spacing between departures is T seconds), and hence $\tilde{Y}[t_1, t_2] \leq B \leq r(t_2 - t_1) + \sigma$, so we are done.
- Case 2: If $U(t_1) = 0$, then:

$$\tilde{Y}[t_1, t_2] \leq X[t_1, t_2] \leq r(t_2 - t_1) + \sigma$$

and hence we are done.

- Case 3: If $t_2 \geq t_1 + T$, and if $U(t_1) > 0$, then there must be a time $t_b \leq t$ such that:

$$U(t_1) = X[t_b, t_1] - \mu \cdot (t_1 - t_b) \quad (12)$$

However, because $U(t_1) > 0$, the system is non-empty and there is a packet in the server. Let $\tau(t_1)$ represent the total time already spent processing this packet. We thus have:

$$0 \leq \tau(t_1) \leq \min[T, (t_1 - t_b)]$$

which follows because $\tau(t_1) < T$ (else, the packet would not be in the system), and because $\tau(t_1) \leq (t_1 - t_b)$ (as the time spent processing this packet up to time t_1 is no more than the duration of the busy period up to time t_1). The only bits that can depart the system during $[t_1, t_2]$ in the actual (packet-based) departure process are the $U(t_1) + \tau(t_1)\mu$ bits in the system plus the $X(t_1, t_2 - T]$ new bits that arrive during $(t_1, t_2 - T]$ (note that no arrivals less than T seconds before time t_2 can depart before time t_2 , as the service time of all packets is T). Thus:

$$\tilde{Y}[t_1, t_2] \leq U(t_1) + X(t_1, t_2 - T] + \tau(t_1)\mu \quad (13)$$

From (13) and (12) we have:

$$\begin{aligned} \tilde{Y}[t_1, t_2] &\leq X[t_b, t_1] - \mu \cdot (t_1 - t_b) + X(t_1, t_2 - T] + \tau(t_1)\mu \\ &= X[t_b, t_2 - T] - \mu \cdot (t_1 - t_b) + \tau(t_1)\mu \\ &\leq r(t_2 - T - t_b) + \sigma - \mu \cdot (t_1 - t_b) + \tau(t_1)\mu \\ &= r(t_2 - t_1) + \sigma - (\mu - r)(t_1 - t_b) - rT + \tau(t_1)\mu \end{aligned}$$

However, $\tau(t_1) \leq (t_1 - t_b)$ and hence (because $\mu \geq r$):

$$\begin{aligned}\tilde{Y}[t_1, t_2] &\leq r(t_2 - t_1) + \sigma - (\mu - r)\tau(t_1) - rT + \tau(t_1)\mu \\ &= r(t_2 - t_1) + \sigma - r(T - \tau(t_1)) \\ &\leq r(t_2 - t_1) + \sigma\end{aligned}$$

where the last inequality follows because $\tau(t_1) \leq T$. This completes the proof.