

EE549: Problem Set #2

Due Monday Feb. 4

Here we provide some preliminary definitions and formulas that will be useful for the problem set. Consider a single-server queue. Service is First-In-First-Out (FIFO) if packets are placed in the server in the order of arrival. We say that a packet does not depart the system until its last bit departs the system. Let $D(t)$ represent the total number of packet departures during the interval $[0, t]$. Let $L(t)$ represent the total number of packets in the system at time t (so that $L(t) = N(t) - D(t)$ for all $t \geq 0$, assuming the system is initially empty). We say that the queue has a *time average number of packets* \bar{L} if:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L(\tau) d\tau = \bar{L} \quad \text{with prob. 1}$$

The $M/D/1$ queue: We say that a packet arrival process $N(t)$ is a *Poisson process with rate* λ if inter-arrival times are i.i.d. and exponentially distributed with rate parameter λ (so that the inter-arrival density is $p_\tau(t) = \lambda e^{-\lambda t}$ for $t \geq 0$). Consider a Poisson packet arrival process with arrival rate λ , where all packets have the same size of B kb. Suppose the packets are served in FIFO order according to a constant rate server with service rate μ kb/sec. This is called a $M/D/1$ queue, where M stands for *memoryless (Poisson) arrivals*, D stands for *deterministic (constant) service times*, and 1 stands for *single server*. If $\lambda B < \mu$, it can be shown that the time average number of packets \bar{L} in a $M/D/1$ queue is given by the following formula:¹

$$\bar{L}_{M/D/1} = \frac{\rho^2}{2(1-\rho)} + \rho$$

where ρ is defined $\rho \triangleq \lambda B / \mu$ and satisfies $0 \leq \rho < 1$. The parameter ρ is called the *utilization* or the *loading parameter*.

I. DEPARTURES

Packets arrive to a single-server queue according to an arrival process $N(t)$, and are served in FIFO order. Consider a *service time queueing model* and let S_1, S_2, \dots represent the packet service times, where S_i is the service time required for packet i . Suppose the system is initially empty, and let $D(t)$ represent the number of packet departures from the system up to and including time t (where a packet can only depart after it completes its service time in the FIFO server).

a) Draw a picture and write a paragraph (with at least two or three sentences) describing why the following inequality is true for all time $t \geq 0$:

$$\sum_{i=1}^{D(t)} S_i \leq t$$

b) Suppose the input process $N(t)$ has rate λ . Suppose the service times $\{S_i\}_{i=1}^{\infty}$ are i.i.d. with mean $\mathbb{E}\{S\}$. Prove that the maximum possible departure rate of packets is $\min[1/\mathbb{E}\{S\}, \lambda]$.

c) Suppose that $\lambda > 1/\mathbb{E}\{S\}$. Describe what will happen to the number of packets in the system over time (you might define $L(t)$ as the current number of packets in the system at time t).

II. FIXED LENGTH PACKETS AND THE MULTIPLEXING INEQUALITY

Let $X(t)$ be an arrival process where all packets have a fixed size of B kb (so that $X(t) = N(t)B$). Let $U_{single}(t)$ represent the unfinished work in a single server system with server rate $\mu(t)$ and input process $X(t)$. Let $U_{multi}(t)$ represent the unfinished work in a K -server system with input process $X(t)$ and with server rate processes $\mu_1(t), \dots, \mu_K(t)$, where $\mu_1(t) + \dots + \mu_K(t) = \mu(t)$ for all t . Assume both queues are initially empty, and assume the single-server queue uses FIFO service. We say that a packet does not depart until its *last bit departs*. Let $L_{single}(t)$ and $L_{multi}(t)$ respectively represent the total number of packets in the single and multi-server systems.

¹We will talk more about Poisson queueing models and prove such formulas in the second part of this course.

a) Explain why $L_{single}(t) = \left\lceil \frac{U_{single}(t)}{B} \right\rceil$ for all t . Explain why $L_{multi}(t) \geq \left\lceil \frac{U_{multi}(t)}{B} \right\rceil$ for all t (where $\lceil x \rceil$ is the *ceiling function*, defined as the smallest integer greater than or equal to x).

b) Prove that $L_{single}(t) \leq L_{multi}(t)$ for all t .

c) Suppose that we have a single-server queue with a constant transmission rate of $\mu = 6$ kb/sec. There are two arrival processes $X_1(t)$ and $X_2(t)$ that enter this queue (see Fig. 1). Both processes have fixed size packets of size $B = 2$ kb. The first process $X_1(t)$ is Poisson with arrival rate $\lambda = 1.5$ packet/sec. The second process $X_2(t)$ is periodic with one packet arrival every second. Assume service is FIFO. Let $L(t)$ represent the total number of packets in the system (including packets from both streams). Compute the tightest upper bound you can for \bar{L} . That is, prove that $\bar{L} \leq C$, where C is some finite constant that you try to make as small as possible, using only methods we have learned in the course.

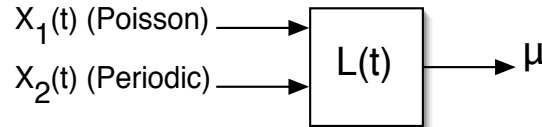


Fig. 1. A single server queue with two input streams $X_1(t)$ and $X_2(t)$.

III. BROADCASTING WITH THE LAW OF LARGE NUMBERS

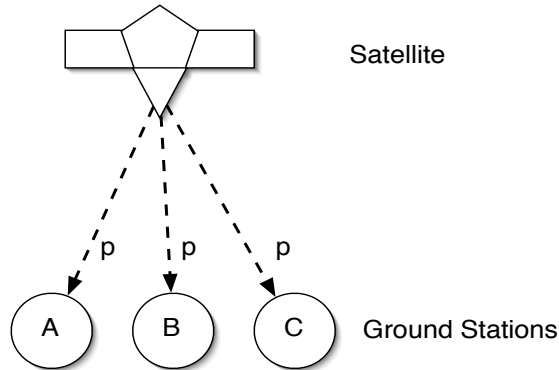


Fig. 2. A satellite that broadcasts a k packet file to 3 ground stations.

A satellite system that operates over fixed timeslots broadcasts data to three different ground stations (Fig. 2). The satellite has an infinite buffer of packets to send, and each packet must be delivered to each of the 3 ground stations (for example, the packets may represent data from a news broadcast that is desired by each ground station). All 3 stations are within range of the satellite and can overhear each single packet transmission. However, due to independent noise at each ground station, every packet transmission is successfully received at each station independently with probability p (where $p < 1$), and is unsuccessfully received with probability $1 - p$. Packets are transmitted on slot boundaries $t \in \{0, T, 2T, 3T, \dots\}$ (where T is the slot size). The total time required until all three ground stations have a particular packet can be viewed as the *service time* of the packet. Thus, all service times are an integer multiple of T seconds, and the minimum possible service time is T seconds exactly (which happens with probability p^3).

a) Suppose that each ground station gives ACK/NACK feedback to the satellite after each packet transmission (so that the satellite knows which stations successfully received the packet after each transmission, and before the next timeslot boundary occurs). Consider an algorithm for packet transmission as follows: Transmit a packet every slot until all stations have the packet, then transmit the next packet, etc. Let \bar{S}_0 , \bar{S}_1 , and \bar{S}_2 represent the expected number of remaining transmissions required until all stations have the current packet, given that 0 stations have the current packet, 1 station has the current packet, and 2 stations have the current packet, respectively. In 2 or 3

sentences, intuitively explain the following equations (see also the corresponding “Markov chain” shown in Fig. 3):²

$$\begin{aligned}\bar{S}_0 &= 1 + (1-p)^3\bar{S}_0 + 3p(1-p)^2\bar{S}_1 + 3p^2(1-p)\bar{S}_2 \\ \bar{S}_1 &= 1 + (1-p)^2\bar{S}_1 + 2p(1-p)\bar{S}_2 \\ \bar{S}_2 &= 1 + (1-p)\bar{S}_2\end{aligned}$$

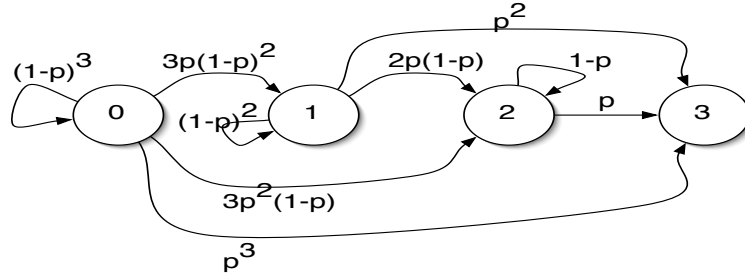


Fig. 3. The Markov chain for problem 1(a).

b) Assume $p = 0.9$. Solve the equations in part (a) to compute \bar{S}_0 , the expected number of times required to transmit a new packet before all 3 users successfully receive it.

c) Again assume $p = 0.9$. Assuming an infinite buffer of packets to send, let $D(t)$ represent the number of original packets that have been successfully delivered to all three ground stations (so that $D(0) = 0$, and $D(t_1) = 1$, where t_1 is the time when all three ground stations get the first packet.) Compute the departure rate from the satellite, i.e., $\lim_{t \rightarrow \infty} D(t)/t$. Hint: Look for *renewals* where the law of large numbers can be applied.

IV. MULTIPLEXING INEQUALITY, SLOW TRUCKS, AND DELAY

Suppose we have 100 packets in a single-server queueing system at time $t = 0$. Suppose the server rate is constant and given by $\mu = 2$ kb/sec. The packet lengths are given by $\{B_1, B_2, \dots, B_{100}\}$. There are no arrivals, and service is FIFO with packet B_1 served first. Let $L_{single}(t)$ represent the total number of packets in the system at time t (so that $L_{single}(0) = 100$). Now consider a 2-server system with constant and equal server rates $\mu_1 = \mu_2 = 1$ kb/sec. Suppose the *same* 100 packets (with the same lengths $\{B_1, B_2, \dots, B_{100}\}$) are initially in the system. Suppose service is work conserving and FIFO, in the sense that packets are immediately placed to a server when it becomes available, and packets B_1 and B_2 are placed first, B_3 is placed next, etc. Let $L_{multi}(t)$ be the total number of packets in the multi-server system (so that $L_{multi}(0) = 100$). There are never any new arrivals in either system.

a) Prove that the single-server system empties first.

b) Let W_i^{single} represent the delay of packet i in the single-server system, being the time when packet i departs. For example, W_{100}^{single} is the time when the single-server system empties. Similarly let W_i^{multi} be the delay of packet i in the multi-server system. The empirical average delays \bar{W}_{single} and \bar{W}_{multi} are defined:

$$\begin{aligned}\bar{W}_{single} &= \frac{1}{100} \sum_{i=1}^{100} W_i^{single} \\ \bar{W}_{multi} &= \frac{1}{100} \sum_{i=1}^{100} W_i^{multi}\end{aligned}$$

Give an example where $\bar{W}_{multi} \leq \frac{1}{30} \bar{W}_{single}$. Hint: You must design the packet sizes $\{B_1, \dots, B_{100}\}$, and compute the exact empirical average delays \bar{W}_{single} and \bar{W}_{multi} . This shows that, despite the multiplexing theorem, there can be cases where delay in the multi-server system can be much better.

²Markov chains and steady state theory will be formally introduced later in the course. This problem is so intuitive that it can be done without knowledge of Markov chain definitions.

V. DYNAMIC ROUTING TO PARALLEL SERVERS

Consider a set of K parallel queues with separate buffers and with constant server rates μ_1, \dots, μ_K . Assume that $\mu_1 \leq \mu_2 \leq \dots \leq \mu_K$. Packets arrive to the system and must be routed to one of the K queues *immediately upon arrival*. The GREEDY strategy routes every arriving packet to the queue that will allow it to exit the system first. The Join-the-Shortest-Queue (JSQ) strategy routes the packet to the queue that will allow it to begin its service first. Assume all packet sizes are less than B_{max} bits.

a) Suppose a packet arrives at time t and has length B (where $B \leq B_{max}$). Give a mathematical definition of the two strategies in terms of B and the unfinished works $U_1(t^-), \dots, U_K(t^-)$ seen by the arriving packet. Show that the two strategies are equivalent if all service rates are equal.

In class we saw that the JSQ strategy ensures the total unfinished work is no more than $(K-1)B_{max}$ in excess of any other routing strategy (using the tracking theorem). Here we analyze the GREEDY strategy.

c) Let $U_i(t)$ represent the unfinished work in queue i at time t , assuming the GREEDY strategy is used. Show that if $U_i(t) = 0$ at any time t , then $U_j(t) \leq B_{max} \frac{\mu_j}{\mu_i}$.

d) Show that the GREEDY strategy ensures the total unfinished work $\sum_{i=1}^K U_i(t)$ is no more than $B_{max} \sum_{i=1}^K \mu_i / \mu_1$ in excess of any other strategy. (Hint: Define a *fully loaded period* to be an interval of time in which all servers of the GREEDY system are busy).

e) Which is “better,” the JSQ strategy or the GREEDY strategy? (write a few sentences).

VI. RATE STABILITY FOR MULTI-SERVER, SHARED BUFFER SYSTEMS

Consider a K -server, shared buffer system with server rates $\mu_1(t), \dots, \mu_K(t)$. Let $X(t)$ be an arrival process with rate r kb/sec. Assume $X(t)$ is composed of packets with a maximum packet size $B_{max} < \infty$. Suppose that all server processes have well defined time average rates $\bar{\mu}_i$ for all $i \in \{1, \dots, K\}$. We define the multi-server, shared buffer system to be *rate stable* if:

$$\lim_{t \rightarrow \infty} \frac{U_{multi}(t)}{t} = 0 \quad \text{with prob. 1}$$

Suppose that the service policy in the multi-server, shared buffer system is *work conserving*. We want to prove that the system is rate stable if and only if $r \leq \sum_{i=1}^K \bar{\mu}_i$.

a) Show that if $r \leq \sum_{i=1}^K \bar{\mu}_i$ then the system is rate stable.

b) Show that if $r > \sum_{i=1}^K \bar{\mu}_i$ then the system is *not* rate stable.

(Note that we only have rate-stability results for single-server systems, and so you will need to compare to a single-server system.)

VII. RATE STABILITY FOR A SERVER SCHEDULING PROBLEM

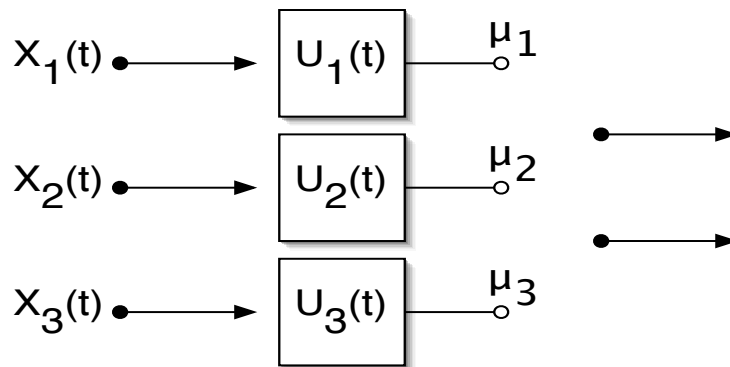


Fig. 4. A 3-queue, 2-server problem

Consider the 3-queue wireless downlink as shown in Fig. 4. Data arrives to the three different queues according to arrival processes $X_1(t)$, $X_2(t)$, and $X_3(t)$, with rates r_1, r_2, r_3 kb/sec, respectively. The system is time-slotted with slots equal to T seconds. There are two orthogonal frequency bands, so that exactly two queues can be served

on each timeslot. The problem is to decide which two queues to serve every slot, or equivalently, how to allocate the two servers amongst three queues. Once a server is allocated to a queue on a particular slot, it is kept there for the duration of the slot and can only be changed on slot boundaries. The server rates are constant and given by μ_1, μ_2, μ_3 , so that if a server is allocated to queue i on a particular slot, it can shift out as much as $\mu_i T$ kb of data. Thus, each server rate process can be viewed as a time varying process $\mu_i(t)$ given by:

$$\mu_i(t) = \begin{cases} \mu_i & \text{if a server is allocated to } i \text{ during the slot containing time } t \\ 0 & \text{otherwise} \end{cases}$$

We say that the system is rate stable if all three queues are rate stable.

a) Let $r_1 = .9$ kb/sec, $r_2 = .4$ kb/sec, and $r_3 = .5$ kb/sec. Suppose that $\mu_1 = \mu_2 = \mu_3 = 1$ kb/sec. Design a server scheduling policy that ensures the system is rate stable. (Note: You should use the rate stability theorem).

b) Let $r_1 = .9$ kb/sec, $r_2 = .45$ kb/sec, and $r_3 = .6$ kb/sec. Suppose that we have the same system as part (a), so that $\mu_1 = \mu_2 = \mu_3 = 1$ kb/sec. Design a server scheduling policy that ensures the system is rate stable.

c) For a general r_1, r_2, r_3 , and for $\mu_1 = \mu_2 = \mu_3 = 1$ kb/sec, prove that it is *impossible* to design a server scheduling policy that is rate stable if any one of the following inequalities is violated:

$$\begin{aligned} r_1 &\leq 1 \\ r_2 &\leq 1 \\ r_3 &\leq 1 \\ r_1 + r_2 + r_3 &\leq 2 \end{aligned}$$

d) Consider now a system with *four queues* and two servers. The service rates are given by $\mu_1 = \mu_2 = \mu_3 = 1$ kb/sec, $\mu_4 = 2$ kb/sec. Design a server scheduling policy to support bit arrival rates $(r_1, r_2, r_3, r_4) = (.4, .5, .5, .6)$ kb/sec.

e) (Extra credit +7) Suppose we have the three queue system again with $\mu_1 = \mu_2 = \mu_3 = 1$ kb/sec. Suppose we have a general set of bit arrival rates (r_1, r_2, r_3) that satisfy all four of the inequalities in part (c) (also assume that $r_i \geq 0$ for all i). Design a general scheduling algorithm that ensures all three queues are rate stable. This shows that the set of four inequality constraints in part (c) define the *capacity region* Λ of the 3-queue system, being the set of all possible rate vectors that the system can support, considering all possible scheduling algorithms.