CSE 538 – Fall 2014 Midterm Exam
140 points total - 100 points max

Your Name:

Problem No. 1 [50 points]

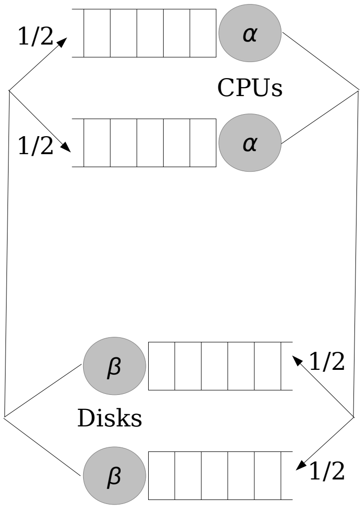

Figure 1: Closed network model of disks+CPUs system.

Consider the system depicted in Fig. 1 that consists of two parallel CPUs, both with an average processing rate of α jobs/sec, connected to two parallel disks that both complete jobs' R/W operations at an average rate of β jobs/sec, where $\beta < \alpha$. When jobs complete work on a CPU, they are randomly assigned to one of the two disks with equal probabilities. Similarly, after a job completes its R/W operations on a disk, it is immediately reassigned to a CPU. The selection of a CPU is again random, with each CPU selected with equal probability.

1. **[5 points]** Assuming initially that there is only a single job in the system, *i.e.,* $N = 1$, obtain expressions for the throughput X and the average response time $E[T]$ of the system.

   Little's Law for closed (batch) systems states that $N = X \cdot E[T]$, where X is the system throughput, $E[T]$ the average response time, and N the multi-programming level or number of jobs in the system. When $N = 1$, this implies $X = \frac{1}{E[T]}$. Furthermore, when there is a single job, the job never waits at either the processors or the disks, so that its average response time is simply the sum of its average service times, *i.e.,*

   $$E[T] = \frac{1}{\alpha} + \frac{1}{\beta} = \frac{\alpha+\beta}{\alpha\beta} \qquad \Rightarrow \qquad X = \frac{\alpha\beta}{\alpha+\beta}$$

2. **[5 points]** Can you reduce the response time by a factor 2 simply by speeding-up the processors, and if yes by how much do you need to speed them up?
   Conversely, can you do the same by speeding the disks up, and if yes, again by how much?

   If we make the processor faster by a factor $x > 1$, we have

   $$E[T_{Px}] = \frac{1}{x\alpha} + \frac{1}{\beta} = \frac{x\alpha+\beta}{x\alpha\beta}$$

   We want $E[T_{Px}] \leq \frac{E[T]}{2}$, which requires

   $$\frac{x\alpha+\beta}{x\alpha\beta} \leq \frac{\alpha+\beta}{2\alpha\beta} \qquad \Rightarrow \qquad x(\alpha-\beta) \leq -2\beta$$

which is infeasible.

A similar derivation for the case of speeding up the disks by a factor $y > 1$ yields

$$\frac{\alpha + y\beta}{y\alpha\beta} \le \frac{\alpha+\beta}{2\alpha\beta} \quad \Rightarrow \quad y(\alpha-\beta) \ge 2\alpha \quad \Rightarrow \quad y \ge \frac{2\alpha}{\alpha-\beta}$$

3. **[10 points]** Considering the original system of Fig. 1, assume now that there is an arbitrary number N of jobs in the system. Derive as functions of N, α and β, an upper bound for the system's throughput X_N and a lower bound for the system's response time $E[T_N]$.

   We have four components in the system; two processors and two disks, with expected service times of the form $E[S_1] = E[S_2] = \frac{1}{\alpha}$ for the two processors, and $E[S_3] = E[S_4] = \frac{1}{\beta} > \frac{1}{\alpha}$ for the two disks respectively. Any job visits a processor or disk with probability $\frac{1}{2}$ in each cycle, so that the total expected service times $E[D_i], i = 1, 2, 3, 4$ for each of the four components during any job cycle are given by

$$E[D_1] = E[D_2] = \frac{1}{2\alpha} \quad \text{and} \quad E[D_3] = E[D_4] = \frac{1}{2\beta}$$

   This gives

$$D = \sum_{i=1}^{4} E[D_i] = \frac{\alpha+\beta}{\alpha\beta} \quad \text{and} \quad D_{\max} = \frac{1}{2\beta}$$

   This gives the following bounds for the system throughput and response time respectively

$$X_N \le \min\left\{\frac{N\alpha\beta}{\alpha+\beta}, 2\beta\right\}, \qquad E[T_N] \ge \max\left\{\frac{\alpha+\beta}{\alpha\beta}, \frac{N}{2\beta}\right\}$$

4. **[5 points]** Assume that $N = 5$. Give tight upper and lower bounds for the system throughput X_5 and response time $E[T_5]$, respectively; again as functions of α and β?
   Hint: Compute N^* and use it to determine what side of the bounds of question 3 applies.

   The crossover N value, N^*, for the above bounds verifies

$$N^* = \frac{D}{D_{\max}} = \frac{2(\alpha+\beta)}{\alpha} = 2 + \frac{2\beta}{\alpha} \le 4 \quad \text{since} \quad \beta < \alpha\,.$$

   Hence $5 > N^*$, so that tighter bounds satisfy

$$X_5 \le 2\beta \quad \text{and} \quad E[T_5] \ge \frac{2.5}{\beta}\,.$$

5. **[10 points]** The administrator of the system considers next replacing one of the two disks by one that is twice as fast as the CPUs instead of slower, *i.e.,* $\beta' = 2\alpha$ rather than $\beta < \alpha$. The other disk remains unchanged. How does this change the results from the previous question? Rigorously justify your answer.

   Note that although replacing one of the disks by a faster one does not change $D_{\max}$ that remains equal to $\frac{1}{2\beta}$, it reduces, say, $E[D_3]$, and therefore D, and consequently N^*. We therefore need to recompute N^* to determine if upgrading one disk helps.

Unfortunately, one readily sees that decreasing D also decreases N^*, so that the bounds from the previous question remain unchanged. This is because, when $N > N^*$ the upper bound for X_N and lower bound for $E[T_N]$ are of the form

$$X_N \leq \frac{1}{D_{\max}} \quad \text{and} \quad E[T_N] \geq N \cdot D_{\max}\,.$$

Since we still have $N > N^*$ and $D_{\max}$ is unchanged, so are the bounds.

6. **[15 points]** Can you suggest a change in the scheduling policy used for either the CPUs or the disks, which would help extract further performance improvements from the use of the faster disk of the previous question, *i.e.,* improve the system's performance bound? What would this improvement be for $N = 5$?

   In order to leverage the faster disk, we need to give it more work, *i.e.,* load balance between the fast and slow disks. Let p denote the fraction of jobs assigned to the fast disk. Increasing p always improves D, but in order to "minimize" (more on this below) $D_{\max}$ we need to select p such that $E[D_3] = E[D_4]$. In other words

$$\frac{p}{2\alpha} = \frac{1-p}{\beta} \quad \Rightarrow \quad p = \frac{2\alpha}{2\alpha+\beta}\,,$$

   which gives

$$D = \frac{1}{\alpha} + \frac{2}{2\alpha+\beta} = \frac{4\alpha+\beta}{(2\alpha+\beta)\alpha} \quad \text{and} \quad D_{\max} = \frac{1}{2\alpha}$$

   where we note from the new value of $D_{\max}$ that the processors have become the bottleneck. This points to the fact that we could actually have chosen a p value that further lowered D by increasing the load of the faster disk until either itself or the slower disk reached an expected service demand equal to that of the processors. This would not have increased $D_{\max}$ and would have lowered D. We will omit this added refinement. The above choice for p then yields the following improved bounds for the system

$$\tilde{X}_5 \leq \min\left\{\frac{5\alpha(2\alpha+\beta)}{4\alpha+\beta}, 2\alpha\right\} = 2\alpha \quad \text{and} \quad E[\tilde{T}_5] \geq \max\left\{\frac{4\alpha+\beta}{(2\alpha+\beta)\alpha}, \frac{2.5}{\alpha}\right\} = \frac{2.5}{\alpha}\,,$$

   where we have used the fact that $\beta < \alpha$.

Problem No. 2 [70 points]

Consider the model of Fig. 2 that describes the operation of a single processor system to which a type of iterative jobs are submitted. Jobs arrive at the *end* of a slot with a fixed probability p. A job that is in service completes with probability q at the *end* of the slot (just before arrivals), and is then either immediately resubmitted with probability r or departs with probability $(1-r)$. If the job is resubmitted, it is enqueued at the end of the queue or immediately moves into service if there are no other jobs waiting. We initially assume that the waiting room in front of the processor can hold an arbitrary number of jobs.

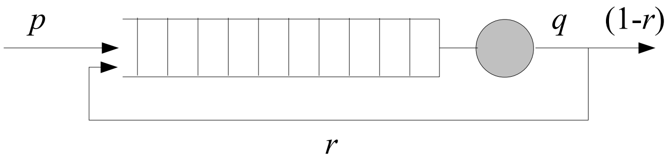

Figure 2: Single processor system operation.

1. **[5 points]** What is the *total* average processing time $E[S]$ of a job?
   Hint: The total service time of a job is essentially a random sum of i.i.d. random variables (one for each iteration in the processor).

The total service time S of a job is of the form $S = \sum_{i=0}^{N} X_i$, where the X_i's are i.i.d. and geometrically distributed with mean $E[X] = \frac{1}{q}$, and N is geometrically distributed with mean $E[N] = \frac{1}{1-r}$ (*e.g.*, when $r = 0$ a job goes through the processor only once). Hence, the total average processing time of a job is $E[S] = E[N] \cdot E[X] = \frac{1}{q(1-r)}$.

2. **[10 points]** Identify a Markov chain representation for the model, and justify why it verifies the Markov property. In specifying the Markov chain, identify its transition probability matrix P.

Let i denote the number of jobs in the system, and consider a Markov chain whose state is i. This state definition satisfies the Markov property since both inter-arrival times and service times are geometrically distributed, and hence memoryless. Note that the fact that a job that completes its current processing can be resubmitted does not affect this statement.

The transition probabilities for the chain are shown in Fig. 3 and verify

$$\begin{aligned}
P_{0,0} &= 1 - p \\
P_{0,1} &= p \\
P_{i,i-1} &= q(1-r)(1-p), i > 0 \\
P_{i,i} &= (1-p)(1-q) + (1-p)qr + pq(1-r), i > 0 \\
P_{i,i+1} &= p(1-q) + pqr, i > 0
\end{aligned}$$

$\alpha = p(1\text{-}q)\text{+}pqr, \quad \beta = q(1\text{-}r)(1\text{-}p)$

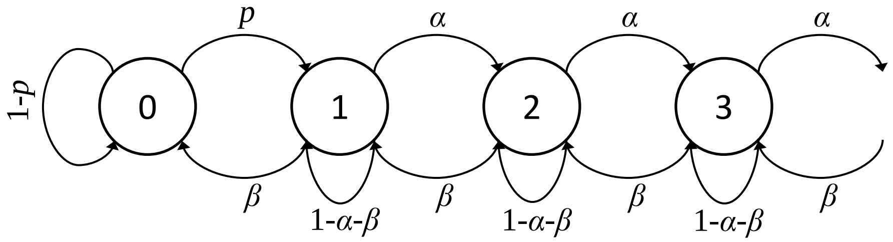

Figure 3: Markov chain representation of iterative processing system.

3. **[10 points]** Assuming that the Markov chain you just defined has a stationary distribution, obtain without deriving this distribution an expression function of p, q, and r for the probability π_0 that the server is idle. Document your derivation steps.
Hint : Obtain an expression for the *total* arrival rate to the server, and apply Little's Law to the server.

The arrival rate to the server consists of both *new* arrivals that arrive with a rate of p in every time slot, and jobs that return after completing their current processing requirements. The rate of arrivals of those jobs is $(1 - \pi_0)qr$, where $1 - \pi_0$ denotes the probability that the processor is busy (there is a job that can depart –with probability q– and return –with probability r). Hence the total arrival rate to the server is $\lambda = p + qr(1 - \pi_0)$.

A job spends on average $\frac{1}{q}$ when in service, and the average number of jobs in service is $E[N] = 1 \times (1 - \pi_0) + 0 \times \pi_0 = 1 - \pi_0$. Hence, applying Little's Law gives the following expression for π_0:

$$\begin{aligned}
1 - \pi_0 &= \frac{p + qr(1 - \pi_0)}{q} \\
\Rightarrow \pi_0 &= 1 - \frac{p}{q(1-r)}
\end{aligned}$$

4. **[5 points]** Use the expression for π_0 from the previous question to obtain a condition for the chain to have a limiting distribution.

   The chain is easily seen to be aperiodic and irreducible, so that we only need to ensure that the stationary distribution exists to ensure that the limiting distribution exists as well (and is equal to it). From the previous expression for π_0, this requires that $\pi_0 > 0$ or in other words $p < q(1-r)$. Note that this is intuitive in that it states that the external arrival rate p must be lower than the actual departure rate $q(1-r)$.

5. **[15 points]** Assuming again that the Markov chain has a stationary distribution, obtain an expression function of p, q, and r for the probability $\pi_i, i = 0, 1, 2, \dots$ that there are i jobs in the system.

   Under the assumption that the stationary distribution exists, we can compute the π_i's by writing the balance equations $\pi = \pi P$, where P is the transition probability matrix that we derived in question 2. Using the notation $\alpha = p(1-q) + pqr$ and $\beta = q(1-r)(1-p)$, this readily yields

$$\begin{aligned} p\pi_0 &= \beta\pi_1 \\ \alpha\pi_i &= \beta\pi_{i+1}, i > 0\,, \end{aligned}$$

   which implies

$$\begin{aligned} \pi_1 &= \frac{p}{\beta}\pi_0 \\ \pi_i &= \left(\frac{\alpha}{\beta}\right)^{i-1}\pi_1, i > 0\,, \end{aligned}$$

   which together with the normalization condition $\sum_{i=0}^{\infty}\pi_i = 1$ implies

$$\begin{aligned} 1 &= \pi_0 + \sum_{i=1}^{\infty}\left(\frac{\alpha}{\beta}\right)^{i-1}\pi_1 = \pi_0\left[1 + \frac{p}{\beta}\sum_{i=0}^{\infty}\left(\frac{\alpha}{\beta}\right)^{i}\right] = \pi_0\left[1 + \frac{p}{\beta\left(1-\frac{\alpha}{\beta}\right)}\right] \\ \Rightarrow \quad \pi_0 &= \frac{\beta-\alpha}{\beta-\alpha+p} = 1 - \frac{p}{q(1-r)} \quad \text{and} \quad \pi_i = \frac{p}{\beta}\left(\frac{\alpha}{\beta}\right)^{i-1}\pi_0\,, i > 0 \end{aligned}$$

6. **[15 points]** Derive an expression for the *total* average time a job spends waiting in the *queue*.
   Hint: Use the answer to the previous question to compute the average number of jobs in the system, and then apply Little's Law and your answer to question 1, where you computed a job's total processing time (the total time in the system is the sum of the queueing and service times).

   From the result of the previous question, we can compute the expected number of jobs in the system, $E[N]$.

$$\begin{aligned} E[N] &= \sum_{i=1}^{\infty} i\pi_i = \sum_{i=1}^{\infty} i\left(\frac{\alpha}{\beta}\right)^{i-1}\pi_1 \\ &= \pi_1\frac{1}{\left(1-\frac{\alpha}{\beta}\right)^2} = \frac{p}{\beta}\frac{(\beta-\alpha)}{(\beta-\alpha+p)}\frac{\beta^2}{(\beta-\alpha)^2} = \frac{\beta p}{(\beta-\alpha)(\beta-\alpha+p)} \\ &= \frac{pq(1-r)(1-p)}{(q(1-r)-p)\,q(1-r)} = \frac{p(1-p)}{q(1-r)-p}\,. \end{aligned}$$

Since the external arrival rate to the whole system is p, we have from Little's Law that the expected total time in the system $E[T]$ is given by

$$E[T] = \frac{E[N]}{p} = \frac{(1-p)}{q(1-r)-p}$$

Since the total time in the system is the sum of the total waiting time W and the total processing S, we have that$E[W] = E[T] - E[S]$, where $E[S]$ was derived in question 1. This gives

$$E[W] = \frac{p(1-p)}{q(1-r)-p} - \frac{1}{q(1-r)} = \frac{p\left(1-q(1-r)\right)}{q(1-r)\left(q(1-r)-p\right)}$$

7. **[5 points]** We consider now a system where only *one* (1) job can wait if the processor is busy, *i.e.,* there can be at most two jobs in the system. Under which conditions does a stationary distribution exists for this system? Rigorously justify your answer.

   This is just a truncated version of the original chain, and so it remains irreducible and aperiodic. Since the chain is now finite, this is all we need to ensure that a limiting distribution exists that is equal to the stationary distribution, *i.e.,* the stationary distribution always exists.

8. **[5 points]** Assuming that the stationary distribution exists, give an expression for the probability $\widehat{\pi}_0$ that the processor is idle?
   Hint: This is just a truncated version of the earlier chain.

   Because this is simply a truncation of the earlier chain, the balance equations still hold and the only thing that changes is the normalization condition. This yields

$$1 = \widehat{\pi}_0\left(1 + \frac{p}{\beta} + \frac{\alpha p}{\beta^2}\right) \qquad \Rightarrow \qquad \widehat{\pi}_0 = \frac{\beta^2}{\beta^2 + \beta p + \alpha p}$$

   where recall that $\alpha = p(1-q) + pqr$ and $\beta = (1-p)q(1-r)$.

Problem No. 3 [20 points]

Consider a processing system that operates according to a discrete clock cycle, where in each clock cycle a new job arrives with a fixed probability p. All jobs require exactly 2 cycles from the processor. Jobs that arrive to find the processor busy with another job wait in a queue that has infinite capacity.

1. **[10 points]** Formulate a Markov chain representation. Note that a chain whose state is simply the number of jobs in the system does not satisfy the Markov property.
   Hint: Having a state that includes not just the number of jobs in the system, but also the current stage of the service time of the job currently in service, if any, might help.

   Define the state space as (i,t), where $i \geq$ is the number of jobs in the system and $t = 1, 2$ is the stage of the job in service if any. The chain's transition probability matrix is as given below and illustrated in Fig. 4.

$$\begin{aligned} P_{0,0} &= 1-p, P_{0,11} = p \\ P_{12,0} &= 1-p \\ P_{i1,i2} &= 1-p, P_{i2,i1} = p, P_{i1,(i+1)2} = p, P_{(i+1)2,i1} = 1-p, i \geq 1 \end{aligned}$$

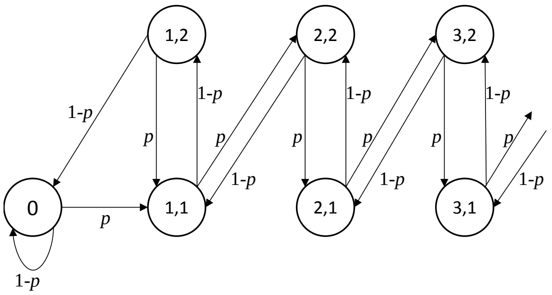

Figure 4: Markov chain representation.

2. **[10 points]** Does the chain admit a limiting distribution, and if yes under which condition? Rigorously justify your answer.
   Hint: Little's Law might again help.

   The chain is easily seen to be aperiodic and irreducible. In order to admit a stationary probability, it must also be positive recurrent. Applying Little's law to the server and using the fact that the arrival rate is p and the service time equal to 2, we get $(1 - \pi_0) = 2p$, so that if we want $\pi_0 > 0$, we need $p < \frac{1}{2}$. Note that this is the usual condition that the average service time must be less than the average inter-arrival time.