CSE 597 – Winter 1998
Final Exam

Name: Answer Key

NOTE: The following answers are written to be concise. I did not expect that you would produce exactly these arguments during the exam.

1. [10 points]
A database system is measured for 100 seconds. The measurements show a total of 1000 transactions completed, a CPU utilization of 50%, and utilizations of 40%, 25%, 20%, 10%, and 5% for each of 5 disk devices. If there were an average of 3 transactions in the system, what fraction of an average transaction’s response time is spent waiting for service?

[This question was intended to be about fundamental laws.]

The arrival rate \( X \) is 1000 transactions / 100 seconds = 10 trans./sec. The total utilization, 150%, represents the average total number of customers in service. Since there is an average of 3 customers in the system 3-1.5=1.5 customers are waiting on average. By Little’s theorem, \( R_{\text{System}} = \frac{N_{\text{System}}}{X} \) and \( R_{\text{Waiting}} = \frac{N_{\text{Waiting}}}{X} \). So, \( \frac{R_{\text{Waiting}}}{R_{\text{System}}} = \frac{1.5}{3.0} = 50\% \).

2. [10 points]
Give a technique that could be used in a simulation to generate samples from the distribution with density function \( f(x)=x \), \( 0 \leq x \leq 1 \), \( f(x)=2-x \), \( 1 \leq x \leq 2 \), and \( f(x)=0 \) elsewhere:

[This is a question about random number generation.]

This is a continuous distribution, so we use the inverse transformation method. The CDF, \( F(x) \), is \( x^2/2 \), \( 0 \leq x \leq 1 \). So we have \( U=F(x)= x^2/2 \), or \( x = \sqrt{2U} \), when the random sample \( U \) is between 0 and 0.5.
When $0.5 < U < 1.0$, we must generate samples between 1 and 2. By symmetry, we can do this as $x = 2 - \sqrt{2(1-U)}$.

3. [10 points]
A novice simulator creates what is really a Monte Carlo simulation. Instead of analyzing all $N \times M$ samples generated by his simulation as a single group, though, he batches them into groups of $M$ consecutive samples and uses the batch mean method. Show that this technique is no worse than what he would have done had he realized that the simulation was Monte Carlo (and had analyzed it that way).

[This question was intended to be about confidence interval analysis.]

The estimate of mean is the same computed both ways: $(\text{sum of all samples})/(NM)$. The question is whether the confidence interval is different or not.

The width of the confidence interval is determined by the sample variance. Let $\sigma^2$ be the actual variance of the underlying distribution (that we are sampling). Then the expected value of the variance of the estimator of the mean using straight Monte Carlo is $\sigma^2/\sqrt{NM}$. Using the batch mean method, the variance of each of the $N$ batches of $M$ samples is $\sigma^2/\sqrt{M}$. Thus, the expected value of the final estimator of the mean (obtained by averaging the $N$ samples corresponding to the batch means) is $(\sigma^2/\sqrt{M})/\sqrt{N} = \sigma^2/\sqrt{NM}$.

4. [10 points]
A processor contains a four stage instruction pipeline. The first stage, instruction fetch, can supply instructions at the rate of 1 every 5 ns. (200 million per second). If the instruction fetched is a branch instruction, the instruction fetch state stalls until the branch exists the pipeline and the next instruction address is known. (No other stalls occur.) All other stages take time less than or equal to 5 ns.

Assuming that fraction $p$ of an instruction stream consists of branches:

(A) What is the effective instruction execution rate?

[This was just a question about using measurement data, and differentiating between rates and times.]

Looking at the output stage of the pipe, a non-branch is followed immediately by another instruction, while a branch is followed by three “bubbles.” Thus the effective execution rate is $(200 \times 10^6)/(1+3p)$.

(B) What is the mean number of stages performing a stall?

[This is (another) question about Little’s theorem.]
By Little's theorem, the average number of stall stages is sum over all pipeline stages of the arrival rate of branches times the amount of stall time a single branch causes.

The branch arrival rate is \((200 \times 10^6) p/(1+3p)\). A branch causes each stage to stall 3 cycles = \(15 \times 10^{-9}\) seconds. So, there is an average of \(3p/(1+3p)\) stalls per stage, or \(12p/(1+3p)\) total.

[Had you had time, you could have checked the answer you gave to make sure it made sense at the extremes: at \(p=0\) your answer should be 0, and at \(p=1\) it should be 3 (or possibly 4, depending on how you interpreted the question – I was indifferent in grading to the assumption of 3 or 4 stall cycles).]

5. [20 points]
Consider the following closed queueing network model in which customers are allowed to change classes as they move from one center to another:

Here the notation \(p_{A \rightarrow B}\) indicates the probability that a class A customer leaving the source center takes this arc and turns into a class B customer on arrival at the destination center (and similarly for \(p_{B \rightarrow A}\)). There are a total of two customers in the model (each switching classes randomly according to the transition probabilities). Service times are exponential and the scheduling discipline is processor sharing (the limiting case of round-robin). Both classes have the same service time (per visit) at each center.

(A) Draw the state transition diagram for this system.

[Well, this was intended as a question about Markov modelling, with a side effect of illustrating that product-form networks with class changes allowed always have equivalent, non-class-changing networks.]

I’ll ignore the transient states (those that have any class B customers at center 0). A semicolon separates the center 0 state from the center 1 state in my labels.
(B) Solve for the equilibrium state probabilities. (You should be able to do this by inspection; i.e., no calculator, laptop, or significant pencil-and-paper calculation should be needed.)

Well, we can guess that everything will balance if we balance each pair of arcs between each pair of states. We start with “unnormalized probabilities,” which we denote W. Let W(2A;0) be 4. Then W(A;A) = W(A;B) = 2, and W(0;2A) = W(0;2B) = 1, and W(0;AB) = 2. Thus the sum of the W’s is 12. The actual probability of any state S, then, is W(S)/12. (This solution procedure worked only because this is a special network. In general, you’d have to solve the global balance equations.)

(C) Show that the equilibrium mean queue lengths (of all customers) in the model above are equal to those in the following single class model.

\[ N=2 \]

\[ D_0=1 \]

\[ D_1=2 \]

[This was a question about MVA, and would also have been about using state probabilities to compute performance measures if I hadn’t told you to forget about part (B) during the exam.]

We start with N=0, for which all queue lengths are 0. For N=1, the residence times are 1 and 2, and the queue lengths must be in this proportion and sum to 1, i.e., \( Q_0[1]=1/3 \) and \( Q_1[1]=2/3 \).

For N=2, we have \( R_0 = 1(1+1/3) = 4/3 \) and \( R_1=2(1+2/3) = 10/3 \). The queue lengths must sum to 2 and be in this proportion, so they are \( 4/7 \) and \( 10/7 \).

For the original network, \( Q_0 = 2*4/12 + 1*4/12 = 1 \), and \( Q_1 = 1*4/12 + 2*4/12 = 1 \), which isn’t equal to the network in this question. Why? Well, unfortunately I made a little last second edit to the class change network to try to make solving it easier, and forgot to change this part of the question. (\( S_1 \) in part (A) should be 2.0, or else part (C) should be changed (to non-integers).)

6. [10 points]

A token ring has \( N \) stations. There is a single token that moves from station to station. When a station has the token, it can server one customer from the head of its queue in time exactly \( S \). (Arriving customers queue FCFS.) If there is no customer at the station, the token still requires time exactly \( S \) to detect this before moving on. In either case, it takes time \( d \) to propagate the token to the next station.

Under the assumption that the interarrival times to each station are exponentially distributed with mean \( 1/\lambda \):
(A) Give a response time equation that describes the response time each customer experiences.

[This was intended to be a question about approximate MVA equations, much like one of your homework questions.]

In the system as described, an arriving customer waits time exactly $N(S+d)$ for every customer ahead of it, plus time $S$ for itself, plus time $N(S+d)/2$ on average for the token to make its way around to that customer’s station for the first time after the customer’s arrival. Thus we have

$$R = N(S+d)(1.5 + Q_{\text{arrival}})$$

(B) Explain how you would use your equation to compute the average response time of customers in this system.

We need to find a value for $Q_{\text{arrival}}$. If the model is open, we just substitute $\lambda R$ and solve for $R$. If the model is closed, we can use the approximate MVA iteration by setting $Q_{\text{arrival}}$ to $\lambda R (N-1)/N$ using our most recent estimation of $R$, then computing a new $R$ using the equation above, and continuing until it converges.

7. [10 points]

I’m thinking of building a new kind of 1xN crossbar. There is a single input port on which all packets arrive, and N symmetric output ports to which they can be routed. An incoming packet is routed (in zero time) to a randomly chosen output port (i.e., there are equal, independent probabilities of choosing output ports).

In one version of my design there is no “excess buffering” – if a packet is sent to a port that is busy, the new packet is dropped. Assuming that there are 100 output ports, that a packet at an output port takes unit time to be transmitted (i.e., occupies that port for unit time), and that packets arrive as a Poisson stream with rate 50, what fraction of packets are lost?

[This is another Markov model question, plus splitting of Poisson processes.]

The system has only two states, 0 and 1 customers:

![Diagram of states](image)

By inspection, the state probabilities are 2/3 and 1/3, respectively. Since arrivals to system are Poisson and they’re split in a Bernoulli fashion, the arrivals to each center are Poisson. Thus, 1/3 of the arrivals will be lost.
8. [10 points]
Briefly explain the tradeoff between factors affecting the choice of the number of
slots and the width of the slots in a calendar queue. (That is, what makes you want to
have many slots? few slots? narrow slots? wide slots?)

[A simulation question.]

We want to have only a few (a constant) number of events in each slot. Having slots
that are too wide wastes time sorting the events in that slot. Having them too narrow
means we have many empty slots, and so waste time looking for the next event.
Similarly, too few slots results in many events per slot, and too many results in empty
slots [at the “tail” of the time line, so we’re basically wasting memory].

9. [20 points]
The department is currently re-architecting its in-building network.

(A) What information would we need to construct an asymptotic bounds model of a
proposed new network?

The service demands of the components of the network, and the “think time” (inter-
packet generation times).

(B) What could the ABA model tell us? What could it not tell us?

It would give us upper and lower bounds on throughput and response time. It would
not estimate average response time or queue lengths.

(C) Suppose I’m thinking of building a simulation model of the proposed system.
What factors would keep me from getting perfectly accurate predictions from this
model?

Primarily uncertainty about what the workload will really look like in the future. In
addition, any model must leave out aspects of the system; if I’m not very careful,
some of these could be important. Finally, simulations are basically random trials,
so if I’m unlucky (or negligent) I could get random results that are outliers.

(D) Suppose I’m thinking of building a queueing model (to be analyzed, not
simulated). What factors would keep me from getting perfectly accurate predictions
from this model?

All the drawbacks of simulation EXCEPT for the outlier problem. In addition, the
analysis techniques impose limitations on the arrival and service processes.