15-829 **HW 3 – Due start of class 10/7**

You have 2 weeks to do this one. Homework is due at the start of class on 10/7. Remember that we don't have class on 9/23. Please turn in a hardcopy. It can be handwritten. Please show all computations. Include a copy of the code your wrote.

Reading from your textbook: Chapter 3 (again) and Chpt 23. We're also covering bits and pieces of Chpts 13, 14. Remember that when reading you should cover all the "Answers" and derivations and try them yourself first.

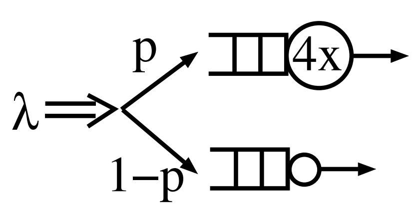
Exercises:

1. Exercise 13.1 from your textbook.

2. Exercise 14.5 from the PnC book (this is the Chpt 14 handout that I gave out in class). After you complete the exercise, please add a part (f), where you check all your simulation results from part (d) by computing the answer based on the formula you learned for the M/G/1 queue.

3. Exercise 23.3 from your textbook. This should be done analytically. If you're unsure about your result, please run a simulation.

4. **[Does load balancing make sense?]** The figure below shows a system with two servers, where the top server is 4x faster than the bottom server. Jobs arrive to the system according to a Poisson process with average rate λ jobs/sec. Assume that job sizes are denoted by r.v. $S \sim \text{Exp}(1)$. Note that when a job runs on a server of speed 4x, its service requirement becomes $S/4$. Each job is immediately dispatched to the fast server with probability p and to the slow server with probability $1 - p$.



   (i) In a load balanced system, what would p be? This exercise will explore whether that value of p yields minimum mean response time, $\mathbf{E}[T]$.

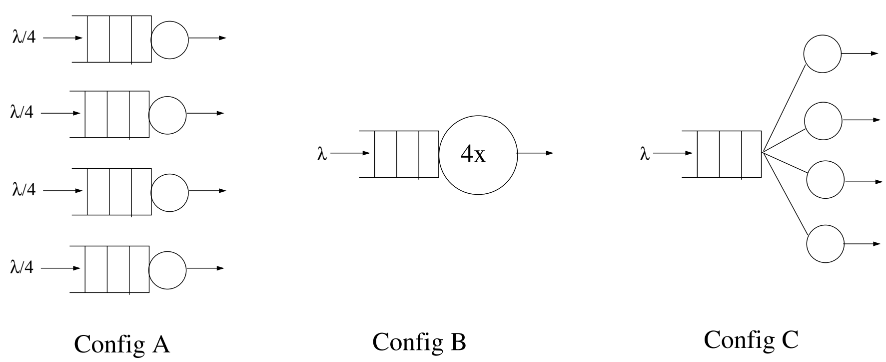
   (ii) Write an expression for $\mathbf{E}[T]$ as a function of p and λ. Simplify your expression as much as you can.

(iii) Now determine the value of p (as a function of λ) that minimizes $\mathbf{E}[T]$. You will do this by differentiating the expression in (ii) with respect to p.

(iv) Using your formula from (iii) determine the optimal value of p for each of these arrival rates: $\lambda = 5, 4, 3, 2$. Note that the only legal arrival rates for the system are $0 \leq \lambda < 5$, so when you evaluate $\lambda = 5$, you're really looking at the limit of λ approaching 5. If you're not sure about your formula in (iii), please run a simulation and determine the optimal value of p from your simulation.

(v) When is load balancing the optimal thing to do? When is it suboptimal? Can you explain intuitively what is going on?

5. **[Comparison of 3 server organizations]** The figure below shows three server organizations. Outside arrivals occur according to a Poisson process with rate λ. Job sizes are denoted by r.v. S. When a job runs on a server of speed 4, its service time is $S/4$.



(i) The load ρ is the same under all three server organizations. What is ρ as a function of λ and S?

(ii) Assume $S \sim \text{Exp}(\frac{1}{3000})$. What is $\mathbf{E}[S]$? What is C_S^2? Now use simulation to determine the mean response time, $\mathbf{E}[T]$ for all three server organizations, for the case of $\rho = 0.3, 0.5, 0.7, 0.9$. Which server organization is best? Which is worst? Why do you think this is?

(iii) Now assume $S \sim \text{BoundedPareto}(k = 1333.33, p = 10^{10}, \alpha = 1.8)$. This distribution is defined by the following p.d.f.:

$$f_S(x) = \alpha \cdot x^{-\alpha-1} \cdot \frac{k^\alpha}{1 - \left(\frac{k}{p}\right)^\alpha}, \qquad \text{where } k \leq x \leq p,$$

where the parameters k, p, and α are defined above. What is $\mathbf{E}[S]$? What is C_S^2? Now use simulation to determine the mean response time, $\mathbf{E}[T]$, for all three server organizations, for the case of $\rho = 0.3, 0.5, 0.7, 0.9$. Which server organization is best? Which is worst? Why do you think this is?

(iv) The ranking of the server organizations is different in parts (ii) and (iii). Why do you think this is, intuitively?