Due: March 27, 2017

Task 1 (M/M/1/K System) Let $\mu=1$ as before, the size of the queue $K=40$, and the number of customers served before a simulation run terminates $C=100,000$. Plot the average customer waiting time against the value of ρ, for $\rho=0.05$ to $\rho=0.95$, in increments of 0.10 (remember to include the confidence intervals). Compile five plots, one for each of the service disciplines. In particular, for the Priority service disciplines, plot the average waiting time of each of the four classes of customers, as well as the overall average. Compare and discuss the results, and submit these graphs along with your code.
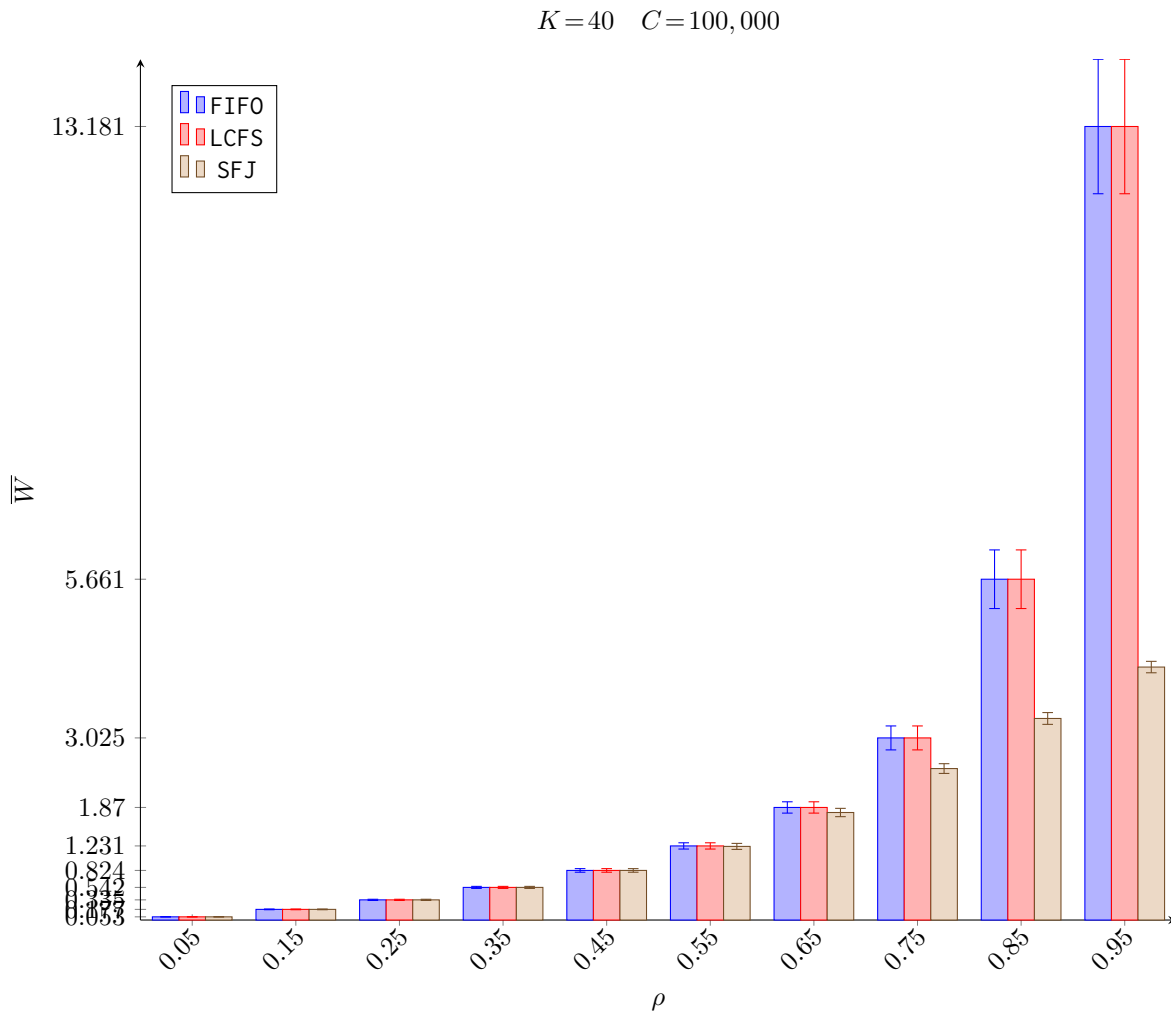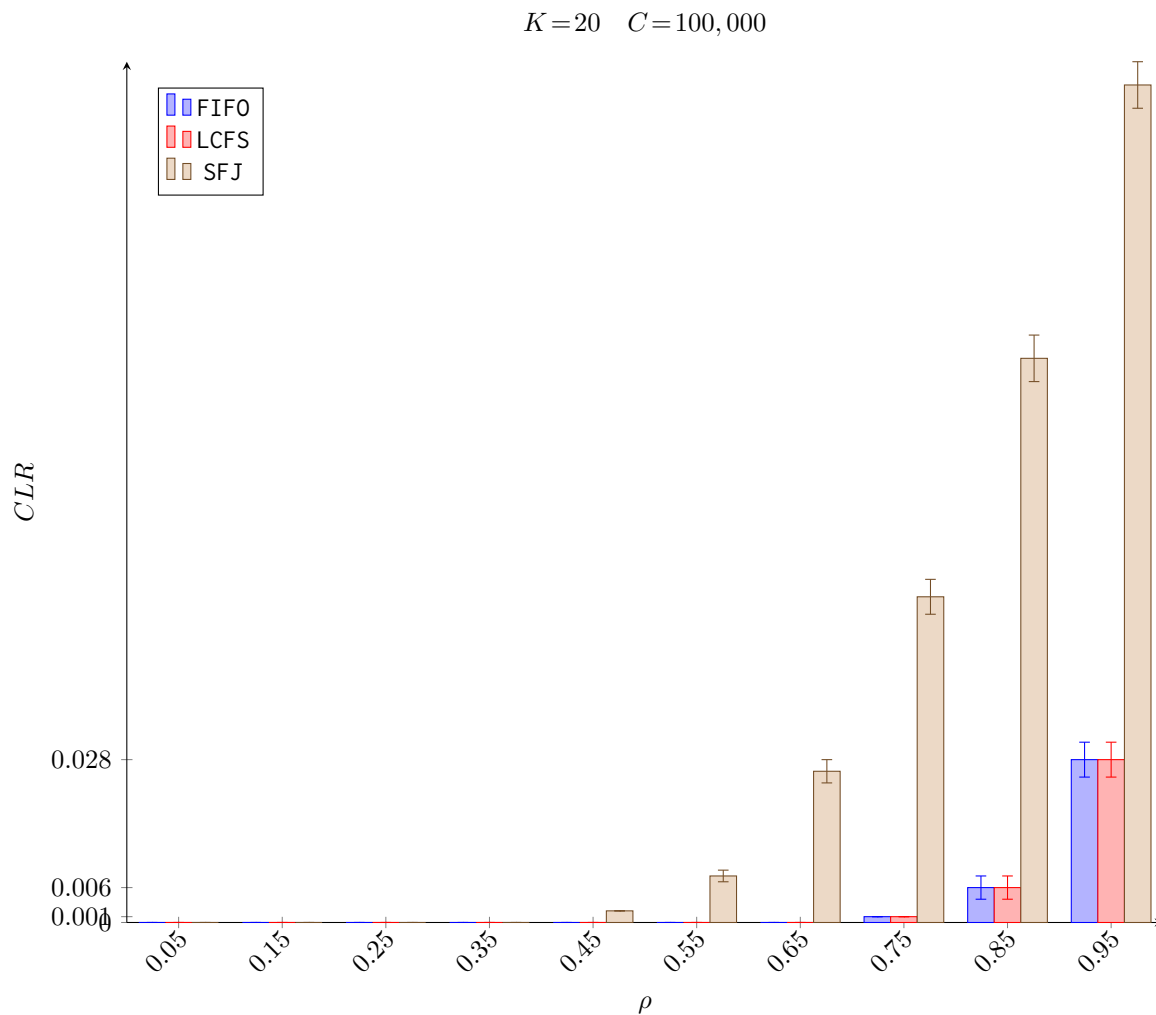
$$K = 40 \quad C = 100,000$$
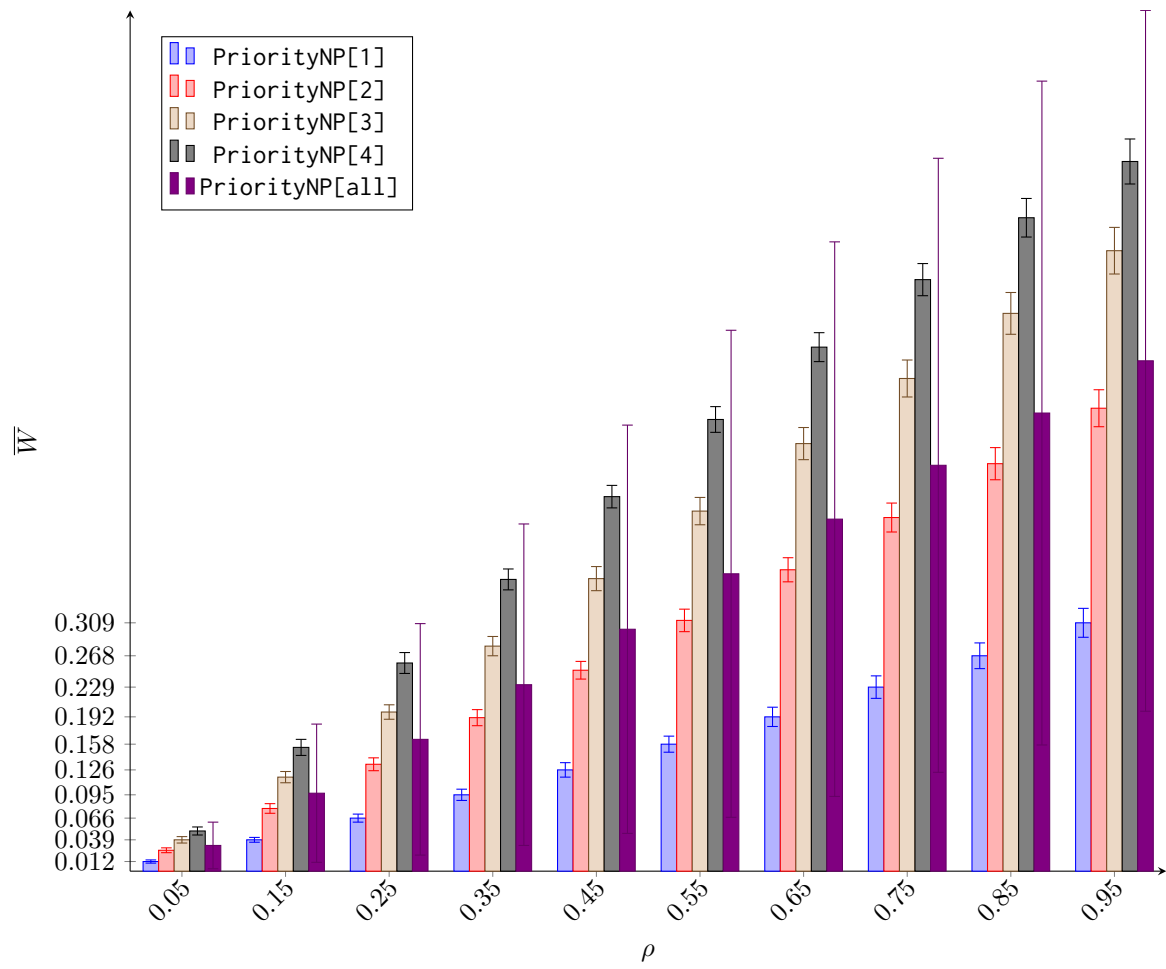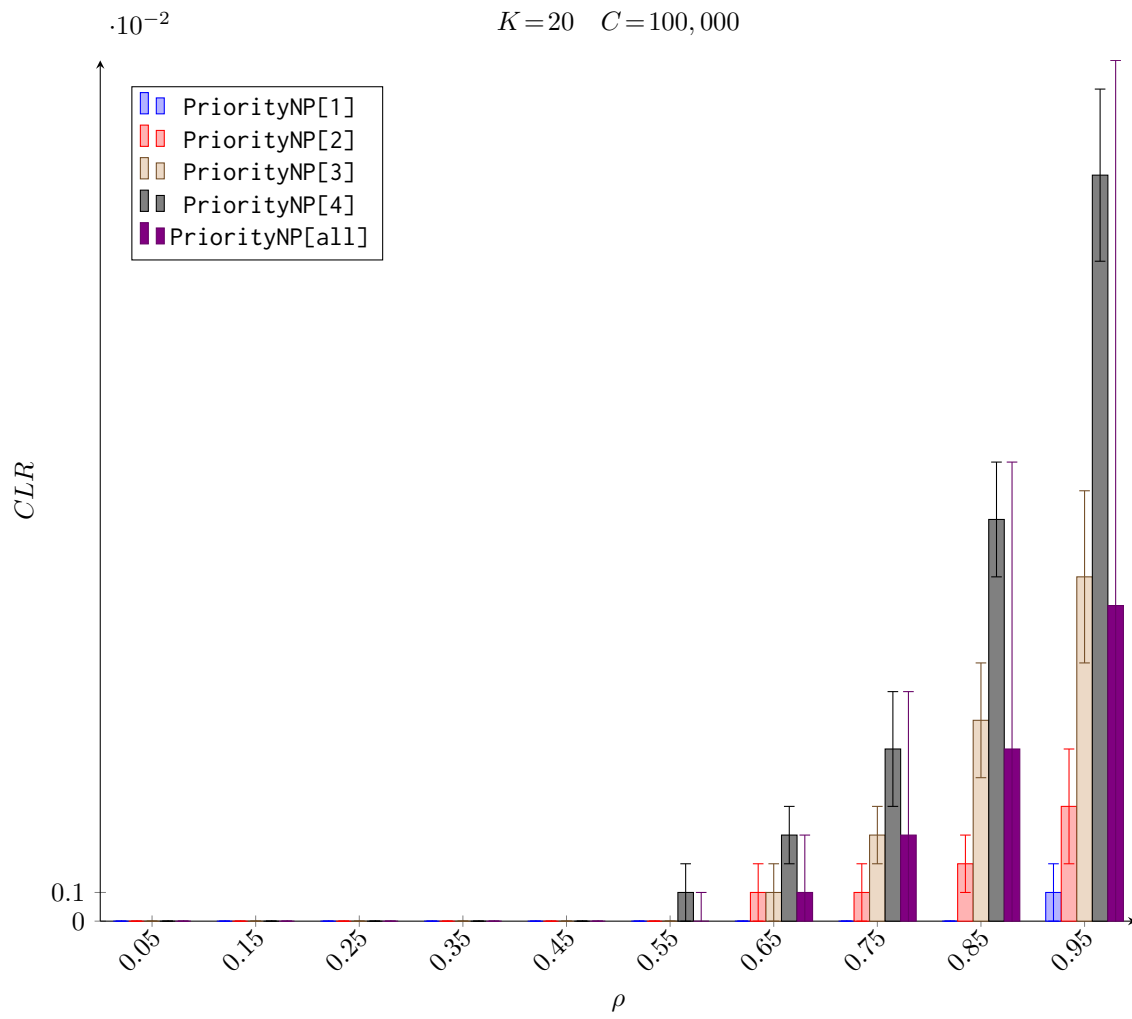
$$K = 40 \quad C = 100,000$$
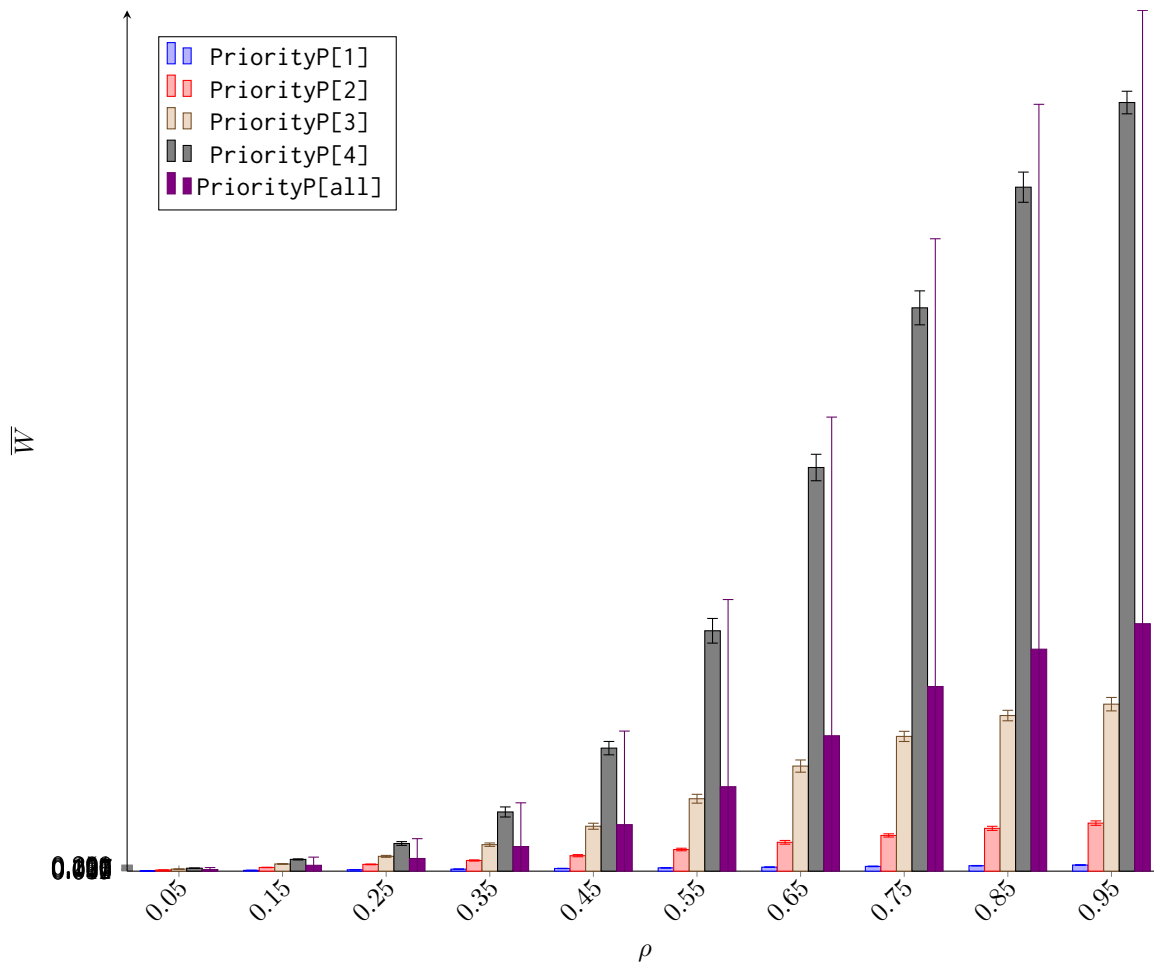
For a constant queue capacity and completion size, where the queue capacity is large relative to the average number of customers in the system, the $\overline{W}$ increases quickly at higher values of ρ. The FIFO and LCFS scheduling policies yield exactly the same average wait times, since there are only two components to calculating average wait time: the residual customers currently in service (because non-preemptive), and the waiting time due to jobs that arrive after the a specific job arrives, but before it enters service. These do not change between FIFO and LCFS.
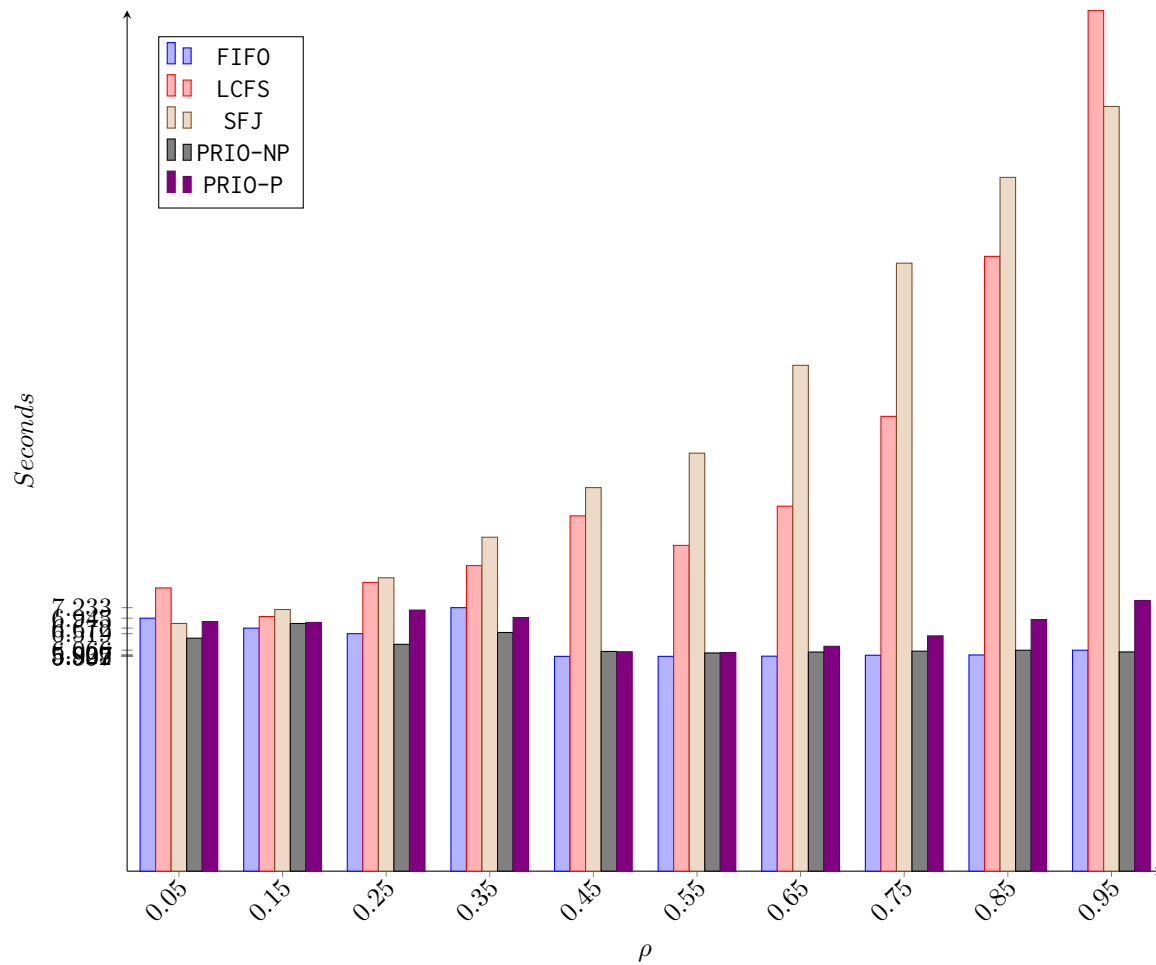SJF shows a lower overal wait time.
The preemptive priority queue has a slightly reduced $\overline{W}$ for queue1 compared to the non-preemptive implementation at higher values of ρ, but at the high cost of much longer wait times for the other queues.

Task 2 (M/M/1/K System) Let $K=20$ and $C=100,000$. Compute and plot the CLR for $\rho=0.05,,0.95$, in increments of 0.10. Also determine and plot the running time of your simulation for the same values of ρ (use $C=1,000,000$ customers). Produce five plots as in Task 1 for each performance metric (CLR and running time); also compute confidence intervals. How does the behavior of the plots change as a function of (i) the service discipline, and (ii) preemption/non-preemption, for the Priority disciplines? Can you explain this behavior?

$K=20$ $C=100,000$

$K\!=\!20$ $C\!=\!100,000$

$\cdot 10^{-2}$
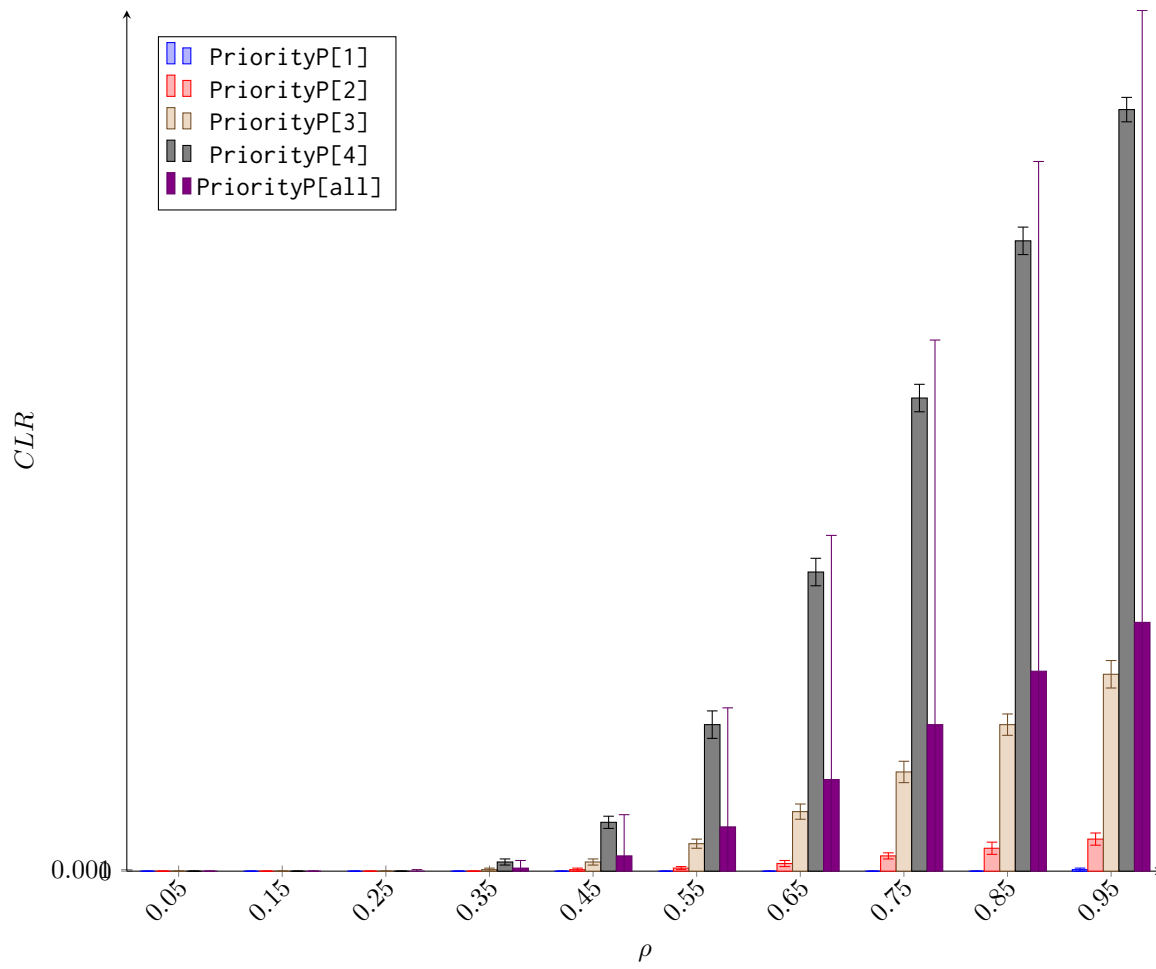
$K = 20 \quad C = 100,000$

$$K = 20 \quad C = 1,000,000$$

For a constant queue capacity and completion size, where the queue capacity is large relative to the average number of customers in the system, the CLR increases quickly at higher values of ρ. The FIFO and LCFS scheduling policies yield exactly the same CLR, since residual customers currently in service (because non-preemptive) does not change with the two policies. SJF shows a higher overal CLR, since this policy trades shorter average wait times for longer queues. The priority queues show higher CLR for the lower priority queues, since these will be serviced less often. The preemptive priority queue has a higher CLR for all queues compared to the non-preemptive implementation. This is because there is a higher queue length due to pre-empted customers.

For running time, the SJF and LCFS policies show increasing times, due to the linear computation needed to manage nonpreemption with these implementations. The other implementations show stable (or even decreasing) times. The decreasing times has no explanation, but the stable times are expected for this type of event-based simulation.

Task 3 (Web Server with I/O) Let $K_{CPU}=50, K_{I/O}=30, C=100,000$ (this is the number of customers leaving the system, possibly after each has made several trips to the CPU). Plot the average customer waiting time (and corresponding confidence intervals) against the value of $\rho=\lambda/\mu_{CPU}$, for $\rho=0.05,,0.95$, in increments of 0.10. Plot five curves: the average waiting time over all four queues, the average for the CPU queue only, and the average for each of the I/O disk queues. Also, define the CLR at any of the queues as the ratio of the number of customers dropped as they try to enter the queue over the total number of customers who tried to enter the same queue. Note that the denominator in this ratio may be different for the different queues; also assume that if a customer is dropped at any queue, it is lost. Plot the CLR for each of the four queues and the overall CLR for the system (for the overall CLR, any customer dropped at any queue is counted as lost).

$$K_{cpu}=50, K_{i/o}=30 \quad C=100,000$$
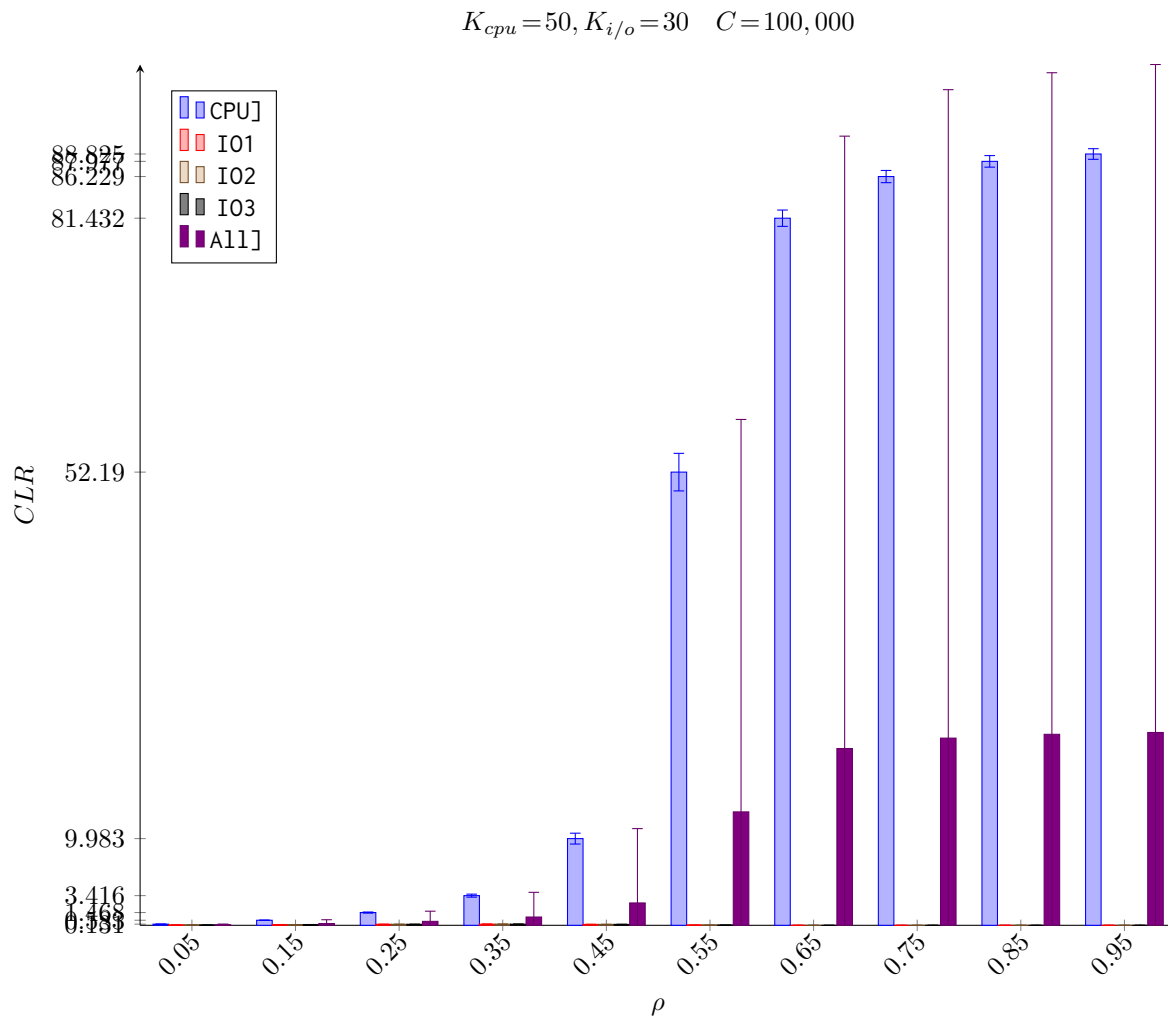
At low values of ρ, the CLR for the CPU and IO queues remains low but >0. As ρ increases, the CPU queue becomes saturated and its CLR rises, resulting in a drop in CLR to the IO queues, since fewer (and then zero) jobs will be passed on these queues.

Task 4 (Web Server with I/O) Let $K_{CPU} = 40, C = 100,000$, and vary ρ from 0.10 to 0.60 in increments of 0.10. For each value of ρ, determine the minimum value $K_{I/O}$ of the I/O queues so that the CLR for any of these queues is less than 0.01 percent.

$$K_{cpu} = 40; CLR < 0.0001 \quad C = 100,000$$
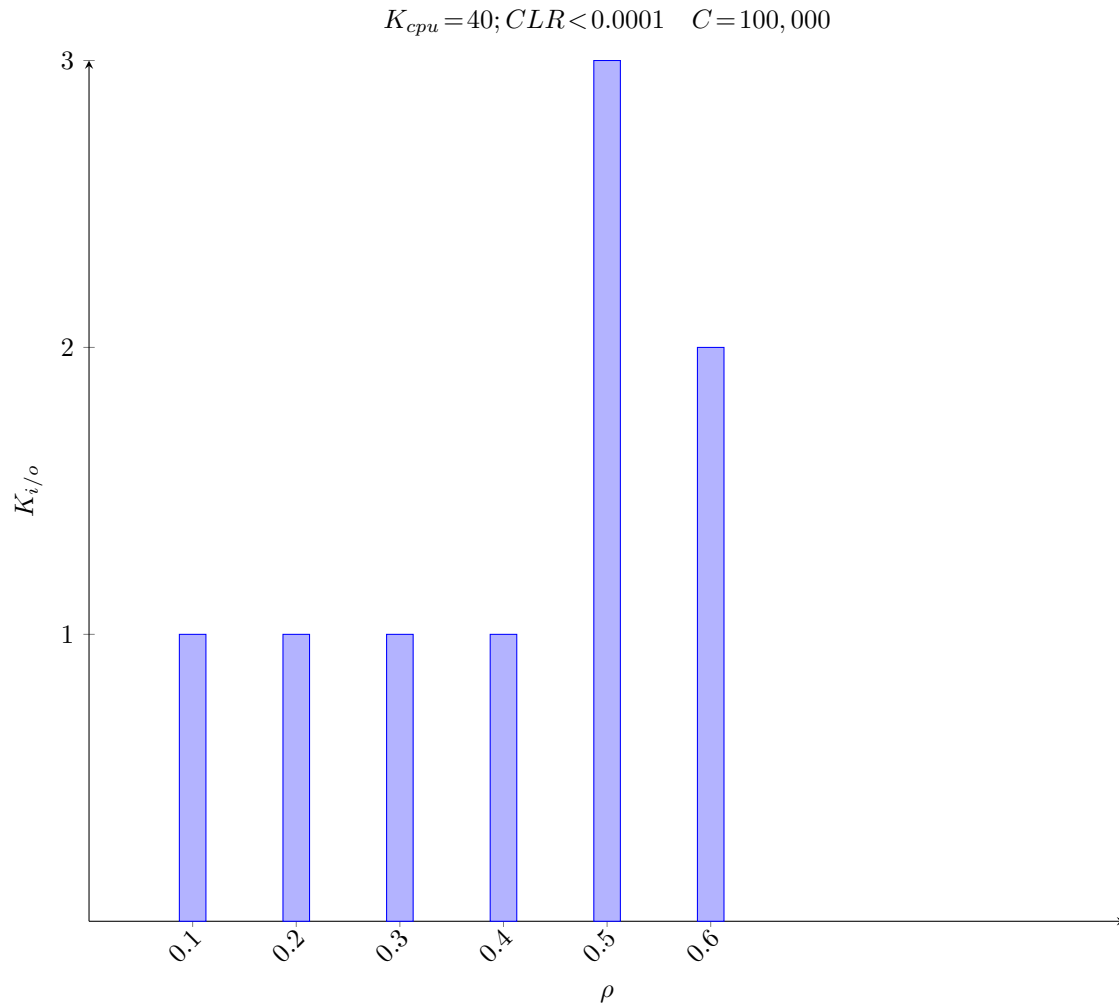
At higher values of ρ for a given K_{CPU}, a higher K_{IO} is needed to maintain a low CLR, but at higher values of rho, the size of the IO queues is irrelevant.

Task 5 (Web Server with I/O) Estimate the average number of times that a customer visits the CPU queue and each of the I/O disk queues. Plot these values against the value of ρ. Do they change as ρ increases? Why or why not?

$$K = 20 \quad C = 100,000$$
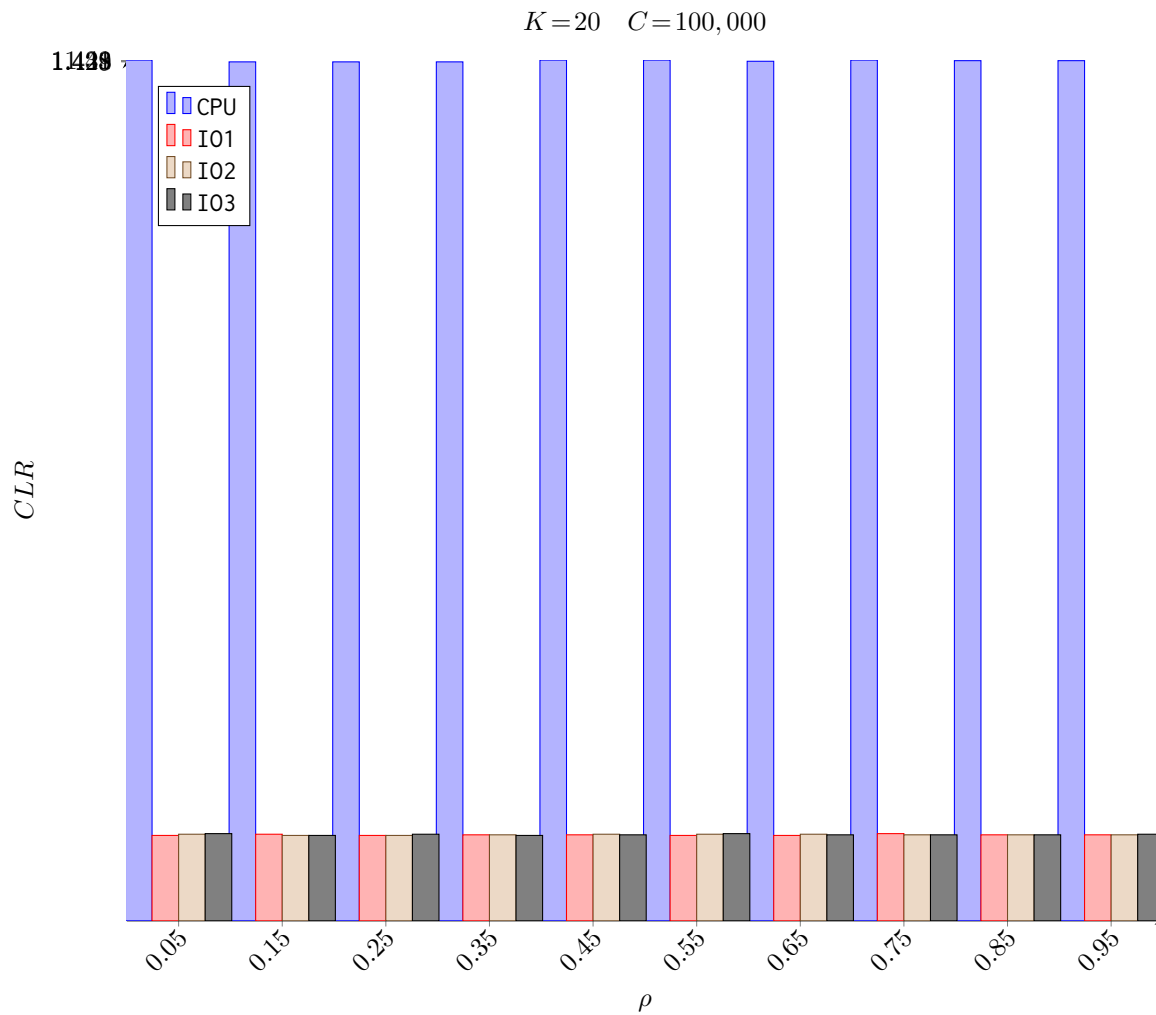
A customer that exits the system will have visited the CPU and 0 or more IO queues based only the transition probabilities, not on the rate of customers entering the system.