Computer Engineering 4DK4
Lab 5
Scheduling for Mobile Cloud Computation Offloading

This lab considers a system that uses mobile *computation offloading*, where an infrastructure-based cloud server executes jobs on behalf of one or more mobile devices. This may be advantageous for the mobile device since energy that would otherwise be expended to execute jobs locally, can be expended on a cloud server to which the mobile device communicates. As shown in Figure **??**, M mobile devices have wireless cellular connections to the Internet where the cloud server can be reached.

An issue of concern is that the channel quality of the cellular connections can be vastly different. A mobile device which is close to the base station, for example, may have a very fast cellular channel connection. Conversely, a device farther away may be forced to communicate at a much lower bit rate. When round-trip latency between the mobile device and the cloud server is important, this can create unfairness where devices farther from the base station may be prevented from using computation offloading. This lab considers what can be done at the cloud server to compensate for this unfairness. A computer simulation is used to assess the performance of the system when $M = 2$.

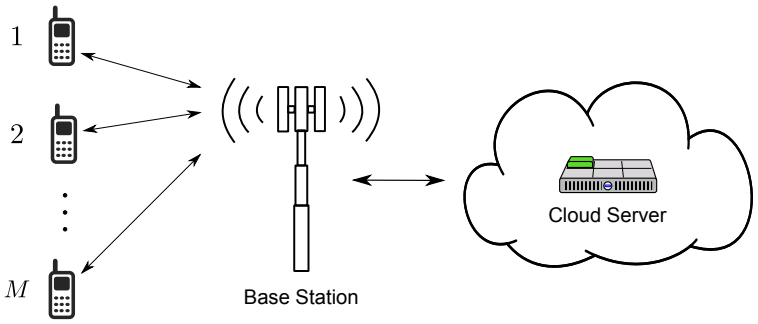

Figure 1: Mobile Computational Offloading. M mobile wireless devices access an infrastructure-based cloud server and use computation offloading for energy reduction.

1 System Model

Assume that there are $M = 2$ mobile devices, one with a good cellular channel connection and one with a bad one. Remote execution job arrivals occur to the devices via a Poisson arrival process with a total arrival rate of λ. Each job has an equal probability to arrive at one of the two devices. The job execution times needed are fixed and given by J_g and J_b for the good and bad connection devices, respectively. Each job has a fixed cellular upload time for it to reach the cloud server, given by U_g and U_b for the two devices.

When a job arrives to a station, it is placed into a FIFO queue and transmitted to the cellular base station (with a service time of U_g or U_b). The cellular base station forwards it to the cloud server with a latency which can be ignored compared to the upload times. Once the job arrives at the cloud server it is placed in a FIFO queue and executed (using a service time of J_g or J_b).

You need to write a simulation of this system and collect the mean delay performance of the each of the devices.

2 Experiments

1. Using the simulation that you have written, generate performance results that show the mean delay of each of the devices, plotted on the same graphs. Assume that $U_b = 10\,U_g$ and $J_g = J_b$.

   If the mean delay for *each device* must not exceed some threshold value, d_{max}, determine the maximum arrival rate, λ^*, that can be supported. This can be done by generating a graph of mean delay vs. λ (i.e., showing the mean delay of both devices), and finding the value of λ (i.e., λ^*) where the mean delay reaches d_{max} for one of the devices. Assume that the arrival rate is the same for both devices.

2. (Bonus) What can you do at the cloud server that would improve the fairness between the experienced delay of the two devices, i.e., keep the two delay curves as close as possible? Simulate this and show results which confirm your idea.

3. (Bonus) What can you do at the cloud server that would increase the value of λ^*? Simulate this and show results which confirm your idea.

3 Writeup

Submit a writeup for the lab. Each group is responsible for their own experiments and writeup. Include in your writeup a description of everything that you did including all data (and random number generator seeds) that were used to obtain the graphs. Include the plots and a listing of the modified program with your writeup. Explain the results that you obtain in all the experiments.