

CMPT 815

Computer Systems and Performance Evaluation

Assignment Two

Due Date: Tuesday, October 22nd

Total Marks: 36

1. In a Web server benchmark run with 36 clients, the average system residence time for a web page request was 1.5 seconds. The throughput at a particular disk on the web server was 16 accesses per second, and a Web request required, on average, 2 accesses of this disk.
 - (a) (2 marks) What was the average client think time, in seconds?
 - (b) (2 marks) How many clients were “thinking” on average?

2. Consider an Internet service implemented with two co-located server machines, one relatively fast (at which requests have an average service time of S_{fast}), and one relatively slow (at which requests have an average service time of S_{slow} , $S_{\text{slow}} > S_{\text{fast}}$). A fraction f of the incoming requests are directed to the fast server, and $1 - f$ to the slow server.
 - (a) (2 marks) What value of f maximizes the throughput capacity of this service (and, therefore, approximately minimizes the average request delay when the load is very heavy)?
 - (b) (2 marks) What value of f minimizes the average request delay when the load is so light that there is no queueing?

3. Measurements of a Web server with two identical disks, under a benchmark load generated by a number of client workstations, yielded the following data:

Observation interval: 1000 seconds
Average think time: 2 seconds
Completed requests: 5000
Processor busy time: 400 seconds
Disk 1 busy time: 500 seconds
Disk 2 busy time: 600 seconds

 - (a) (2 marks) Give the service demands for the processor and disks.
 - (b) (4 marks) Suppose that the processor is upgraded to make it 10x faster. Using the lower bounds on response time from asymptotic bounds analysis, graph an estimate of the percentage decrease in response time as a function of the number of clients, for numbers of clients from 1 to 50.
 - (c) (4 marks) Suppose now that instead of upgrading the processor, the disk subsystem is upgraded by adding another disk, identical in speed to the current disks, and perfectly balancing the load among all three disks. Using the lower

bounds on response time from asymptotic bounds analysis, graph an estimate of the percentage decrease in response time as a function of the number of clients. Consider numbers of clients from 1 to 50.

4. Consider a network link for which the packet arrival process can be accurately modeled as Poisson, with no correlations between the sizes of successive packets or between packet sizes and interarrival times. The packet arrival rate is 8000 packets/second, and the mean packet service (transmission) time is 0.1 milliseconds. Unless stated otherwise, assume that service is First-Come-First-Served (FCFS).
 - (a) (2 marks) What is the link utilization?
 - (b) (4 marks) Give the mean packet residence time and mean queue length, assuming that packet transmission times are exponentially distributed.
 - (c) (4 marks) Give the mean packet residence time and mean queue length, assuming that all packets have exactly the same transmission time (i.e., have the same size).
5. Consider the request routing problem described in problem 2, with $S_{\text{slow}} = 3$ and $S_{\text{fast}} = 1.5$, FCFS scheduling at each server, and Poisson request arrivals at each server.
 - (a) (4 marks) Assume that the request service time at each server is a constant (i.e., exactly 3 at the slow server, and 1.5 at the fast server). Give the value of f that minimizes the average request delay (i.e., $fR_{\text{fast}} + (1-f)R_{\text{slow}}$), to two decimal places, for each of the following values of the request rate: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9.
 - (b) (4 marks) Repeat part (a), but now assuming that the request service times at each server are independent and identically distributed, following a Pareto distribution, with shape parameter $\alpha=2.25$, and minimum value $k = 5/3$ at the slow server and $5/6$ at the fast server. Give some intuition for why the value of f that minimizes the average request delay is typically smaller than in part (a).