CMPT 815
Computer Systems and Performance Evaluation
Assignment Two

SOLUTIONS

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?

   $N = X(R+Z)$ (where $X = X_{disk}/V_{disk}$)
   $36 = (16/2 \text{ requests/second}) \times (1.5 \text{ seconds} + Z)$
   $Z = 3$ seconds

   (b) (*2 marks*) How many clients were "thinking" on average?

   From Little's Law, number of "thinking" clients is XZ, so 8 x 3 = 24 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}, $S_{slow} > S_{fast}$). 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)?

   The throughput capacity is maximized when the loads are balanced; i.e., when $f\,S_{fast} = (1 - f)\,S_{slow}$, yielding $f = S_{slow}/(S_{slow} + S_{fast})$.

   (b) (*2 marks*) What value of f minimizes the average request delay when the load is so light that there is no queueing?

   $f = 1$

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:

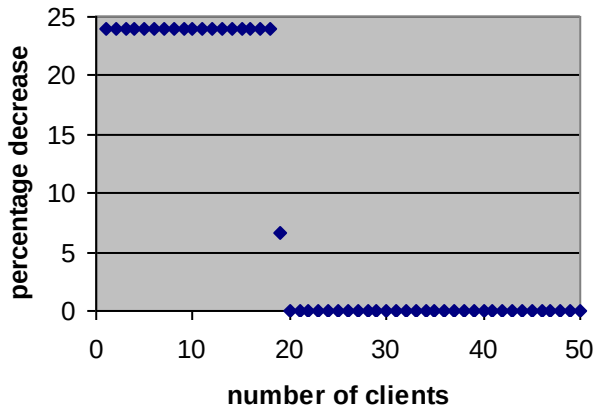
   Observation interval: 1000 seconds
   Average think time: 2 seconds
   Completed requests: 5000
   Processor busy time: 400 seconds
   Disk 1 busy time: 500 seconds
   Disk 2 busy time: 600 seconds

   (a) (*2 marks*) Give the service demands for the processor and disks.

$D_{processor}$ = (400/5000) = 0.08 seconds
D_{disk1} = (500/5000) = 0.1 seconds
D_{disk2} = (600/5000) = 0.12 seconds

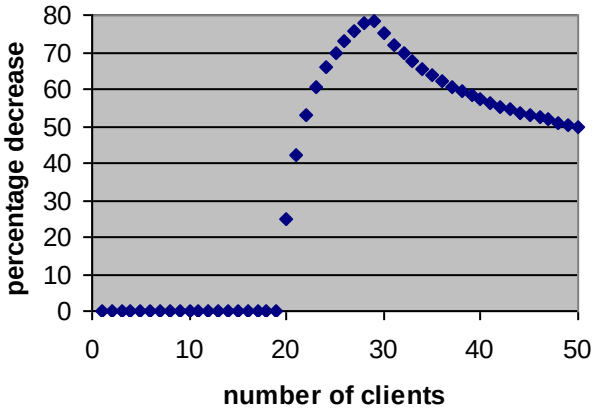
(b) (*4 marks*) Suppose that the processor is upgraded to make it 10x faster. Using the lower bounds on response time from asymptotic bounds analysis, graph an estimate of the percentage decrease in response time as a function of the number of clients, for numbers of clients from 1 to 50.

From asymptotic bounds analysis, $R \geq \max[D, ND_{max} - Z]$.
Have $D_{original}$ = 0.08 + 0.1 + 0.12 = 0.3 seconds, $D_{original\_max}$ = 0.12 seconds,
D_{new} = 0.008 + 0.1 + 0.12 = 0.228 seconds, and D_{new_max} = 0.12 seconds.
This gives an estimate of the percentage decrease in response time of:
(max[0.3, 0.12N - 2] - max[0.228, 0.12N - 2])/max[0.3, 0.12N - 2] x 100%.



(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.

From asymptotic bounds analysis, $R \geq \max[D, ND_{max} - Z]$.
Have $D_{original}$ = 0.08 + 0.1 + 0.12 = 0.3 seconds, $D_{original\_max}$ = 0.12 seconds,
D_{new} = 0.08 + 3(0.22/3) = 0.3 seconds, and D_{new_max} = 0.08 seconds.
This gives an estimate of the percentage decrease in response time of:
(max[0.3, 0.12N - 2] - max[0.3, 0.08N - 2])/max[0.3, 0.12N - 2] x 100%.



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?

   U = XS
   U = (8000 packets/second) x (1/1000 seconds/ms) x (0.1 ms/packet) = 0.8

   (b) (*4 marks*) Give the mean packet residence time and mean queue length, assuming that packet transmission times are exponentially distributed.

   $R = S(1 + Q) = S(1 + \lambda R)$ which yields $R = S/(1 - U)$
   $R = 0.1/(1 - 0.8) = 0.5$ milliseconds
   $Q = \lambda R = 8000 \times 0.5 / 1000 = 4$

   (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).

   $R = S(1 + Q - U) + (S/2)U = S(1 + \lambda R) - (S/2)U$ which yields $R = S(1 - U/2)/(1 - U)$
   $R = 0.1 \times (1 - 0.4)/(1 - 0.8) = 0.3$ milliseconds
   $Q = \lambda R = 8000 \times 0.3 / 1000 = 2.4$

5. Consider the request routing problem described in problem 2, with $S_{slow} = 3$ and $S_{fast} = 1.5$, 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., $fR_{fast} + (1-f)R_{slow}$), 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.

   $R = fR_{fast} + (1-f)R_{slow}$,
   where $R_{fast} = 1.5 + 0.75(1.5)f\lambda /(1-(1.5)f\lambda)$ and $R_{slow} = 3 + 1.5(3)(1-f)\lambda/(1-3(1-f)\lambda)$.

Numerically find that the following values for f minimize the average request delay:

λ	f
0.1	1.00
0.2	1.00
0.3	0.97
0.4	0.84
0.5	0.77
0.6	0.73
0.7	0.70
0.8	0.69
0.9	0.68

(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 α=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).

$R = fR_{fast} + (1-f)R_{slow}$,

where $R_{fast} = 1.5 + 0.5[(5/6)^2(2.25)/(0.25)]f\lambda/(1-(1.5)f\lambda)$ and

$R_{slow} = 3 + 0.5[(5/3)^2(2.25)/(0.25)](1-f)\lambda/(1-3(1-f)\lambda)$.

Numerically find that the following values for f minimize the average request delay:

λ	f
0.1	1.00
0.2	0.92
0.3	0.83
0.4	0.77
0.5	0.74
0.6	0.72
0.7	0.70
0.8	0.69
0.9	0.68

The high variability in service times results in a greater sensitivity to load (i.e., greater increases in queueing delays as load increases) than with deterministic service times, and thus the optimal division of load tends to be somewhat more balanced (i.e., the optimal f is a bit closer to 2/3).