Tutorial "Performance Evaluation Techniques" Fifth Problem Sheet

Dr.-Ing. Andreas Willig

Hasso-Plattner-Institute, University of Potsdam

email: willig@hpi.uni-potsdam.de

June 10, 2003

Date of assignment	June 11, 2003 (in the tutorial)
Date of submission	July 2, 2003 (in the tutorial)

Problem 1:

Assume that we have two M/M/1 systems in parallel, system one having a service rate of μ_1 customers per second, system two serves μ_2 customers per second. The overall arrival rate is λ. Each arriving customer is subjected to an independent Bernoulli random experiment: with probability $p \in (0,1)$ the customer enters system one, with probability $1 - p$ he enters system two.

- What is the stability condition for this system?

- Find p such that the expected system response time of a customer is minimized.

	2

Problem 2:

	10

Consider the following system: there is a number N of queueing stations (QS), each having a single server and an infinite waiting room. The service times at QS i are independent and identically distributed with arbitrary service time distributions. However, the mean service time of QS i is known and given by $\frac{1}{\mu_i}$. Different QS might have different service time distributions. Customers arrive to the overall system according to a Poisson process with rate $\lambda > 0$. An arriving customer is subjected to one of the following policies:

Random Selection (RS) : The customer enters a randomly chosen QS, all QS are equiprobable

1

Minimum Selection (MS) : The customer enters the QS which at the time of his arrival has the fewest number of customers in the system (if multiple queues have the same minimal number of customers in the system, one of them is chosen randomly).

Load-Balancer Selection (LBS) : The decision is made by a separate load-balancer entity, which works as follows:

- The load-balancer queries periodically with period $\Delta_t > 0$ the number of customers in all queues. Lets say the resulting values from sampling at time $n \cdot \Delta_t$ are $M_1(n), \ldots, M_N(n)$. For each QS i we compute the expected time to serve all the present customers, this time is given by:

$$S_i = \frac{M_i(n)}{\mu_i}$$

After sorting these times, we can determine indices $i_1, \ldots, i_N$ such that $S_{i_1} \leq S_{i_2} \leq \ldots \leq S_{i_N}$. A customer which arrives before the next sampling instant $(n+1) \cdot \Delta_t$ is subjected to one of the following policies:

LBS-I : the customer is sent to QS i_1 (i.e. to the fastest one)

LBS-II : assign QS i_1 the weight $w_{i_1} = 1$, and assign QS i_k ($k \geq 2$) the weight:

$$w_{i_k} = \begin{cases} \frac{S_{i_1}}{S_{i_k}} & : & S_{i_k} > 0, S_{i_1} > 0 \\ \frac{1}{1+S_{i_k}} & : & S_{i_k} > 0, S_{i_1} = 0 \\ 1 & : & S_{i_k} = 0 \end{cases}$$

An arriving customer is assigned to queue i_k with probability

$$p_{i_k} = \frac{w_{i_k}}{\sum_{j=1}^{N} w_{i_j}}$$

We want to compare the four policies RS, MS, LBS-I, and LBS-II with respect to the following steady-state performance measures:

- System throughput

- Mean delay of a single customer

- Variance of a single customers delay

- Distribution of a single customers delay

Develop a simulation model with OMNet++ which allows to vary the following parameters (e.g. by setting them in the `omnetpp.ini` file):

- The number of queueing stations N

- For each station i the distribution of the service times (give the mean service time $\frac{1}{\mu_i}$ as a separate parameter, which must be consistent with the given distribution)

- The interval Δ_t

- The arrival rate λ

Use the Replication/Deletion method (including proper transient removal!) to jointly estimate the steady-state system throughput and the mean delay of a single customer at a confidence level of $\alpha = 10\%$ and relative error (confidence interval half-width) of at most 5%. Run the following experiments:

- $\lambda = 100$ Customers per second, $N = 5$, all servers have exponential service times with a rate of 30 Customers per second, $\Delta_t = 1$s

- $\lambda = 100$ Customers per second, $N = 5$, all servers have exponential service times with a rate of 30 Customers per second, $\Delta_t = 0.01$s

- $\lambda = 100$ Customers per second, $N = 5$, all but one servers have exponential service times with a rate of 30 Customers per second, the remaining server has lognormal service times with mean 30 Customers per second and coefficient of variation of 10, $\Delta_t = 0.01$s

Submit:

- Your simulation program

- The `.ini` files for the different experiments

- For all simulation runs the following data both as tables and as figures, where appropriate: the length of the initial transient, the overall number of observations, the steady state throughput, the steady state mean system response time per customer, the variance of the steady state system response times per customer, minimum and maximum observed system response times.

- Your evaluation of the actually needed number of replications for each configuration and the achieved confidence interval width.