Midterm Exam for Capacity Planning (CIS 4930/6930)

Spring 2001

>>> SOLUTIONS <<<

Welcome to the midterm exam in *Capacity Planning* (CIS 4930/6930). You have 75 minutes. Read each problem carefully. There are six required problems (each worth 16 points - you get 4 points for "free") and one extra credit problem worth 5 points. You may have with you a calculator, pencils, erasers, blank paper, lucky rabbit's foot, and one 8.5 x 11 inch "formula sheet". On this formula sheet you may have anything you want (definitions, formulas, etc.) *handwritten by you*. You may use both sides. Computer generated text, photocopies, and scans are not allowed on this sheet. Please submit your formula sheet with your exam. Please start each numbered problem on a new sheet of paper and do not write on the back of the sheets (I really do not care about saving paper!). Submit everything in problem order. No sharing of calculators. Good luck and be sure to show your work!

Problem #1 (10 minutes)

a) What is capacity planning (you *knew* this problem would be on the test!).

Capacity planning is a process for determining the saturation point for computer systems. Capacity planning is also used to determine what can be done (i.e., alternatives) when saturation is reached. The inputs to capacity planning are workload evolution (the inputs to the system), system parameters (the system itself), and the desired service level (e.g., 99% of all transactions must complete in 2 seconds or less).

b) Give three examples (one each for software development, hardware development, and systems) of applications of capacity planning and/or performance evaluation. Be precise in your answers (e.g., what are the constraints and trade-offs of interest?).

Software development: Capacity planning can held determine what programming methods should be used given a constraint of development time. For example, OOP language (fast development, possibly poor performance) vs. procedural language (slower development, better performance).

Hardware development: Capacity planning can be used for determining design trade-offs given a constraint of development time or chip space. For example, which is better... a bigger cache or a faster floating point unit.

Systems: Capacity planning can be used to determine when a server should be upgraded given a contstaint of unacceptable service , but also not too early so that technology cost drops are missed (this then is the trade-off - price/performance).

Problem #2 (15 minutes)

Here are some measurements (e.g., of web server response time in milliseconds)...

110, 120, 140, 120, 120, 100, 110, 100, 140, 120

a) Compute the mean and variance of this population of measurements

$$E[X]=\frac{1}{10}(110+120+140+120+120+100+110+100+140+120)=118$$

$$\sigma^2=\frac{1}{10}\left(110^2+120^2+140^2+120^2+120^2+100^2+110^2+100^2+140^2+120^2\right)-118^2=176$$

b) Plot a histogram for the measurements.

```
 Pr[X]  |
        |
        |                40%
        |  20%    20%    |            20%
        |   |      |     |             |
        +----------------------------------
           100    110    120    130    140
```

c) Describe (give the formulas) how the mean for time measurements and rate measurements is computed. Explain why the means are computed in a different way for time and rate measurements. This has nothing to do with parts (a) and (b) above.

An arithmetic mean $\overline{X}_A$ is used for time measurements and a harmonic mean $\overline{X}_H$ for rate measurements. This is so that the mean rate is inversely proportional to an increase in time values.

$$\overline{X}_A = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \overline{X}_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

Problem #3 (5 minutes)

Answer the following questions regarding time series.

a) Sketch an example time series and label key items of interest.

```
 Job size  |
           |         Job size
           |       |
           |       |     |      |
           |    |  |     |      |    |  |  |  |
           +-----------------------------------
                   <----->                Time
              Interarrival time
```

b) What is "stationarity" and why do we care about it?

Stationary means that the mean of the underlying probablity distribution is constant. A time series with a daily cycle or trend is not stationary. We need a time series to stationary in order to do a "meaningful" analysis on the type of underlying distribution.

c) What does autocorrelation measure?

Autocorrelation measures the independence of subsequent events (whether job sizes or interarrival times) in a time series. An autocorrelation of zero indicates independence, one indicates complete correlation.

Problem #4 (15 minutes)

Answer the following questions about workload

a) What is "workload"?

Workload is the input to a system. Workload parameters include job size and job interarrival time. A workload can be described as a time series.

b) Why do we want to be able to generate synthetic workload (give at least two reasons)?

Here are four reasons:

1. To drive real systems (e.g., in benchmarks)
2. To drive simulation models
3. To drive analytical models
4. For insight into the users and the system

c) What are some of the attributes of a good workload generator (give at least three attributes)?

Here are five good attributes:

1. Parsimonious (few parameters)
2. Easily generated
3. Easily tunable
4. Tractible (if for an analytical model)
5. Accurate (same characteristics as a real, measured workload)

d) Write a C function that will return a value from an empirical distribution based on the measurements given in problem #2. About a half-dozen lines of code should do the trick.

```
#include <stdio.h>     // Needed for printf()
#include <stdlib.h>    // Needed for ran() and RAND_MAX

int emp(void);

void main(void)
{
  printf("%d \n", emp());
}
```

```
int emp(void)
{
  double z;         // unif(0,1)

  z = (double) rand() / RAND_MAX;

  if (z <= 0.20) return(100);
  if (z <= 0.40) return(110);
  if (z <= 0.80) return(120);
  return(140);
}
```

Problem #5 (15 minutes)

Answer the following questions regarding queueing.

a) What is an M/M/1 queue? Describe it carefully and completely.

An M/M/1 queue is a single-server queue with infinite buffer size and infinite customer population. Arrivals and service times are exponentially distributed.

b) The formula for mean number of customers in the system (*L*) for an M/1/1 is

$$L = \frac{\rho}{1-\rho} \text{ where } \rho = \frac{\lambda}{\mu} = \frac{\text{arrival rate}}{\text{service rate}}$$

Derive *W* (mean wait in system) and W_q (mean wait in queue). This is *not* a difficult problem, do not let the word "derive" scare you. Show all of your steps. Your answers must be in a minimized form.

We use $L = \lambda W$, and $W = W_q + (1/\mu)$. Then...

$$W = \frac{\frac{\rho}{1-\rho}}{\lambda} = \frac{\frac{1}{u}}{1-\rho} = \frac{1}{\mu-\lambda} \text{ and } W_q = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda}$$

c) The P-K formula for L for an M/G/1 queue is

$$L = \rho + \rho^2\left[\frac{\left(1+C_s^2\right)}{2(1-\rho)}\right] \text{ where } C_s \text{ is the coefficient of variation of the service time}$$

Show that the P-K formula reduces to *L* (in (b) above) for the M/M/1 case (i.e., where the G for service time is an M). This too is simple.

For M/M/1, $C_s^2 = 1$, so...

$$L = \rho + \rho^2\left[\frac{(1+1)}{2(1-\rho)}\right] = \rho + \frac{\rho^2}{1-\rho} = \frac{\rho(1-\rho)+\rho^2}{1-\rho} = \frac{\rho}{1-\rho}$$

Problem #6 (15 minutes)

The following table contains test measurements for the response time of two systems. A table of T-scores is at the end of this exam.

Test #	System #1	System #2
1	110 ms	110 ms
2	105	90
3	110	100
4	108	80
5	102	85
6	112	85
7	98	105

Can you with 95% confidence say that system #1 or system #2 is better (remember, "better" would be smaller response time)? Show your work.

D = 0, 15, 10, 28, 17, 27, -7

$$E[D]=\frac{1}{7}(0+15+10+28+17+27-7)=12.857$$

$$s=\sqrt{\frac{1}{7-1}\left(0^2+15^2+10^2+28^2+17^2+27^2-7^2\right)-12.857^2}=13.013$$

$$H=2.45\cdot\left(\frac{13.013}{\sqrt{7}}\right)=12.050$$ (we use 2.45 for the T score for N - 1 = 6 and $\alpha/2=0.025$)

With 95% confidence the population mean for the difference lies between 0.807 and 24.907. Since this range is entirely above zero, we can say (with 95% confidence) that system #2 is better (lower response time) than system #1.

Extra Credit problem:

Fill in the blanks for the following. Each blank is one word. The answer is not on your formula sheet. Think.

The "fastest" possible web server would be able to _predict_ user requests and serve these requests _before_ a user _wants_ to view a given item (e.g., view an HTML web page).

T-scores. Selected values of $t_{\alpha/2;N-1}$

	$\alpha/2=0.05$	$\alpha/2=0.025$
N - 1	**t**	**t**
4	2.13	2.78
5	2.02	2.57
6	1.94	2.45
7	1.90	2.37
8	1.86	2.31
9	1.83	2.26
10	1.81	2.23
11	1.80	2.20
12	1.78	2.18