

Class Problem Set

Instructor: Dr. Neil Gunther

© 2003 Performance Dynamics Educational Services. All Rights Reserved.

Last revision : Friday, August 29, 2003

Mathematica Configuration

Crib Sheet

T	Measurement period
A	Number of arrivals
C	Number of completions
B	Busy time (cumulative)
$\lambda = \frac{A}{T}$	Average arrival rate
$X = \frac{C}{T}$	Average throughput
$S = \frac{B}{C}$	Average service time
$\rho = \frac{B}{T}$	Per server utilization ($\leq 100\%$)
$\rho = XS = \lambda S$	Little's microscopic law (no wait ti
$U = \sum_{k=1}^m \rho_k$	Total utilization for m servers
$Q = \frac{\rho}{1-\rho}$	M/M/1 queue length
$W = QS$	Average wait time
$R = S + W$	Average residence time
$R = \frac{S}{1-\rho}$	M/M/1 residence/response tim
$R = \frac{S}{1-\rho^m}$	M/M/m residence/response tim
$R = S + \frac{\rho S}{1-\rho} \left(\frac{C^2+1}{2} \right)$	M/G/1 residence/response tim
$R = \frac{N}{X} - Z$	Finite request (closed queue) response
$Q = XR = \lambda R$	Little's macroscopic law (with wait

Question 1

A grocery store checkout is monitored for 120 mins (T). In that period, 60 customers (C) have their groceries rung up, and depart from the store. The checker was observed to be idle 25% of the time.

- a) What is the number of arrivals (A) at that checkout?

The measured number of completions, $C=60$ customers. By assumption $C = A$ (i.e., no customers are lost or created) when measured over a sufficiently long measurement period. Therefore:

$$\text{Completions} = 60; T_{\text{groc}} := 120; \text{Arrivals} = \text{Completions}$$

$$60$$

- b) What is the average arrival rate (λ) at that checkout?

$$\lambda = \text{Arrivals} / T_{\text{groc}}$$

$$\frac{1}{2}$$

What does half a customer mean?

- c) What is the average throughput (X) at that checkout?

By the same assumptions we used to determine the number of Arrivals (A), we can say:

$$X_{\text{groc}} = \lambda$$

$$\frac{1}{2}$$

- d) What is the checker utilization (U)?

$$\text{Idle} := 0.25; U_{\text{groc}} = 1 - \text{Idle}$$

$$0.75$$

- e) What is the aggregate busy time (B) at that checkout?

By definition, $U = B/T$. Therefore:

$$\text{BusyTime} = U_{\text{groc}} * T_{\text{groc}}$$

90.

- f) What is the average service time (S) at that checkout?

By definition, $S=B/C$. Therefore:

$$S_{\text{groc}} = \text{BusyTime} / \text{Completions}$$

1.5

What are the units for this result?

Question 2

- a) What is the average residence time (R) for a customer in Question 1?

Treating the checkout stand as a single-server queueing center, we can apply the formula:

$$R_{\text{groc}} = \frac{S_{\text{groc}}}{1 - U_{\text{groc}}}$$

6.

- b) What is the average queue length (Q) at that checkout?

The average queue length can be calculated from **Little's Law**:

$$\text{Queue} = X_{\text{groc}} * R_{\text{groc}}$$

3.

Question 3

Suppose we were asked by the grocery store management, to predict what would happen to **response time** if he added another register-server stand at each checkout?

- a) First, assume the total utilization $U = 0.75$ and $m = 2$ (i.e., added capacity)

Treating the upgraded checkout stand as a dual-server queueing center, we can apply the formula:

$$\rho = U_{\text{groc}} / 2$$

$$S_{\text{groc}}$$

$$\rho^2$$

$$1 - \rho^2$$

$$R_{\text{dualA}} = \frac{S_{\text{groc}}}{1 - \rho^2}$$

$$0.375$$

$$1.5$$

$$0.140625$$

$$0.859375$$

$$1.74545$$

- b) Second, what if the load $\rho = 0.75$ on each server (i.e., scaled demand)?

Treating the checkout stand as a single-server queueing center, we can apply the formula:

$$\rho = U_{\text{groc}}$$

$$\rho^2$$

$$S_{\text{groc}}$$

$$R_{\text{dualB}} = \frac{S_{\text{groc}}}{1 - \rho^2}$$

$$0.75$$

$$0.5625$$

$$1.5$$

$$3.42857$$

This is **still** a 50% improvement over R for the single server case! A truly impressive result.

Question 4

An image file takes 1 minute to scan (on average). A computer system needs to be able to scan 75,000 files per month. Your manager wants you to size an appropriate server for this application.

- a) What is the minimum system capacity to ensure this throughput is maintained? Assume this is a 7 x 24 operation with 4 weeks in a month. (*Hint: Try Little's law: $U = \lambda S$*)

Identify the performance metrics correctly and express the time-base in **minutes**.

$$X_{\text{scan}} = 75000 / (4 * 7 * 24 * 60) \text{ (* files per minute *)}$$

$$S_{\text{scan}} = 1.0 \text{ (* minute per file *)}$$

$$U_{\text{scan}} = X_{\text{scan}} * S_{\text{scan}}$$

$$\frac{625}{336}$$

$$1.86012$$

$$1.$$

$$1.86012$$

$$N[U_{\text{scan}} * 100, 10] \quad (* \text{ percent total utilization } *)$$

186.012

Which is also equivalent to **2** servers running at:

This is a good definition of **"maxed out!"**

- b) What is the average response time to get a file scanned with this minimum server capacity?
Assume there is a single queue.

$$\rho_{\text{scan}} = N[U_{\text{scan}}, 10] / 2 \quad (* \text{ per server } *)$$

0.93006

$$\rho_{\text{percent}} = \rho_{\text{scan}} * 100 \quad (* \text{ as a percentage } *)$$

93.006

Use the response time formula for a dual server:

$$R_{\text{dualC}} = \frac{S_{\text{scan}}}{1 - \rho_{\text{scan}}^2} \quad (* \text{ minutes per file } *)$$

7.408

- c) Let's call the server load in the original system, ρ . Suppose the scan demand grows in such a way that the load (ρ) remains fixed as more servers (m) are added the system. How many servers would be required to meet a **service level objective** of better than 1.3 minutes per image file? (*Hint: Keep 4-5 decimal places for accuracy*)

$$m = \{2, 3, 4, 5, 6, 7, 8\}; \quad \rho_{\text{multi}} = U_{\text{scan}} / m;$$

$$R_{\text{multi}} = \frac{S_{\text{scan}}}{1 - \rho_{\text{multi}}^m} \quad (* \text{ minutes per file } *) // N$$

$$\{7.408, 1.31298, 1.04906, 1.00718, 1.00089, 1.00009, 1.00001\}$$

Question 5

Measurements of a UNIX database server that supports 100 active users, show that the average response time is 1.5 seconds per transaction. The average CPU time per transaction is found to be 0.30 seconds, at 25% CPU time spent in the kernel and 50% in user space. What is the average think time per transaction?

Hint: CPU busy is the sum of the system and user times

Treat as a **closed** queueing circuit and use the **Finite User Response Time** formula: $R = \frac{N}{X} - Z$

$$N_{db} = 100;$$

$$R_{db} = 1.5;$$

$$S_{cpu} = 0.30;$$

$$U_{cpu} = 0.75;$$

$$X_{db} = U_{cpu} / S_{cpu} \text{ (* Little's Law *)};$$

$$X_{db} // N$$

$$2.5$$

$$Z_{db} = \frac{N_{db}}{X_{db}} - R_{db} \text{ (* seconds *)}$$

$$38.5$$

Question 6

A server supports 70 active clients. You use a stopwatch to estimate the average time between the completion of one client transaction and the submission of another. It is found to be around 30 seconds. The paging disk has a measured service demand of 250 mSecs and is 50% busy. What is the average response time of the server?

Hint: The system throughput (X) is the same, no matter which queueing center is used

Treat as a **closed** queueing circuit and use the **Finite User Response Time** formula:

$$R = \frac{N}{X} - Z$$

$$N_{\text{sys}} = 70;$$

$$Z_{\text{sys}} = 30;$$

$$S_{\text{dsk}} = 0.250; (* \text{ seconds } *)$$

$$U_{\text{dsk}} = 0.5;$$

$$X_{\text{sys}} = U_{\text{dsk}} / S_{\text{dsk}} (* \text{ Little's Law } *);$$

$$X_{\text{sys}}$$

2.

$$R_{\text{sys}} = \frac{N_{\text{sys}}}{X_{\text{sys}}} - Z_{\text{sys}} (* \text{ seconds } *)$$

5.

Question 7

A computer system receives transactions at the rate of 8 per second. If each transaction is in the computer for an average of 0.7 seconds, how many transactions (on average) are simultaneously in the computer system?

$$\lambda_{\text{sys}} = X_{\text{sys}} = 8;$$

$$R_{\text{sys}} = 0.7; (* \text{ residence time } *)$$

$$Q_{\text{sys}} = X_{\text{sys}} * R_{\text{sys}} (* \text{ number of transactions in the system } *)$$

5.6

Question 8

A hotel bartender knows that, on average, 18 customers per hour arrive at his bar. There are typically 6 customers to be seen at the bar. What is the average length of time each customer spends at the bar?

This is an application of Little's Law: $Q_{\text{bar}} = X_{\text{bar}} R_{\text{bar}}$

$$\lambda_{\text{bar}} = X_{\text{bar}} = 18;$$

$$Q_{\text{bar}} = 6.0;$$

$$R_{\text{bar}} = Q_{\text{bar}} / X_{\text{bar}}$$

$$0.333333$$

Question 9

In a data comm network, messages arrive to be transmitted over a particular link. The average time required to transmit a message is 600 milliseconds, and messages arrive at an average rate of 1 message every second. How long does each message take to traverse the link?

$$\lambda_{\text{msg}} = 1;$$

$$S_{\text{msg}} = 0.6;$$

$$U_{\text{link}} = \lambda_{\text{msg}} * S_{\text{msg}}$$

$$0.6$$

$$R_{\text{sys}} = \frac{S_{\text{msg}}}{1 - U_{\text{link}}} (* \text{ seconds } *)$$

$$1.5$$

Question 10

From time to time, on a crowded New York street, a passer-by decides to make a call on a lone public telephone. The average length of the call is 2 minutes, and on average people arrive to make calls once every 5 minutes.

- How long on average does each person have to wait to place a call?

$$\lambda_{\text{caller}} = 1/5; (* \text{ people per minute} *)$$

$$S_{\text{caller}} = 2.0; (* \text{ minutes} *)$$

$$U_{\text{phone}} = \lambda_{\text{caller}} * S_{\text{caller}}$$

$$0.4$$

$$R_{\text{caller}} = \frac{S_{\text{caller}}}{1 - U_{\text{phone}}} \quad (* \text{ minutes} *)$$

$$3.33333$$

From the fundamental **Residence Time** definition: $R_{\text{caller}} = W_{\text{caller}} + S_{\text{caller}}$

$$W_{\text{caller}} = R_{\text{caller}} - S_{\text{caller}} \quad (* \text{ minutes} *)$$

$$1.33333$$

- How many people are waiting on average?

$$Q_{\text{caller}} = \lambda_{\text{caller}} R_{\text{caller}}$$

$$0.666667$$

The average number of callers **in service** is given by

U_{phone} The average number of callers **waiting** is the difference between the queue length and the number already in service.

$$Q_{\text{caller}} - U_{\text{phone}}$$

$$0.266667$$

Question 11

A country post office only has a single clerk to help customers. Seventy percent of the customers require 1 minute to have their mail serviced, 20 percent take 3 minutes and 10 percent take 10 minutes. Calculate the average time a person can expect to spend in the post office when people arrive at a rate of 1 every 3 minutes?

Hint: Calculate the variance, the standard deviation, the squared coefficient of variation, and use M/G/1 formula.

$$\lambda_{\text{cust}} = 1/3; (* \text{ people per minute} *)$$

Weighted mean service time:

$$S_{\text{cust}} = (0.7 * 1) + (0.20 * 3) + (0.10 * 10)$$

2.3

Variance of the service time:

$$\text{VarS} = 0.7 * (1 - S_{\text{cust}})^2 + 0.20 * (3 - S_{\text{cust}})^2 + 0.10 * (10 - S_{\text{cust}})^2$$

7.21

Standard deviation of the service time:

$$SD_{\text{clerk}} = \sqrt{\text{VarS}}$$

2.68514

$$\rho_{\text{clerk}} = \lambda_{\text{cust}} S_{\text{cust}}$$

0.766667

Now, calculate the Squared COV:

$$C_{\text{clerk}}^{\text{sq}} = \left(\frac{SD_{\text{clerk}}}{S_{\text{cust}}} \right)^2$$

1.36295

Using the **Pollacek –**

Kinchine formula (Eqn. 2 – 62 in [The Practical Performance Analyst](#)),

we can calculate the average number of **minutes** a person spends in the post office :

$$R_{\text{cust}} = \left(1 + \frac{\rho_{\text{clerk}} \left(1 + C_{\text{clerk}}^{\text{sq}} \right)}{2(1 - \rho_{\text{clerk}})} \right) S_{\text{cust}}$$

11.2286

Question 12

A message switch receives random traffic at an average rate of 240 messages per minute. The bandwidth of the connection is 800 Kbps. The measured distribution of message length (including control characters) is close to exponential with an average length of 17.6 KBytes.

- a) Calculate the important performance metrics: S, λ , ρ , Q, R.

$$\text{msglength} = 17.6 * 1024 * 8; \text{ (**** Bytes ****)}$$

$$\text{bandwidth} = 800 * 1024; \text{ (**** bps ****)}$$

$$S_{\text{msg}} = \frac{\text{msglength}}{\text{bandwidth}} \text{ // N (**** seconds ****)}$$

0.176

$$\text{mpm} = 240; \text{ (**** msgs per minute ****)}$$

$$\lambda_{\text{msg}} = \text{mpm} / 60 \text{ (* msgs per second *)}$$

4

$$\rho_{\text{msg}} = \lambda_{\text{msg}} S_{\text{msg}}$$

0.704

$$Q = \frac{\rho_{\text{msg}}}{1 - \rho_{\text{msg}}} \text{ (* average number of messages in the switch *)}$$

2.37838

$$R_{\text{msg}} = \frac{S_{\text{msg}}}{1 - \rho_{\text{msg}}} \text{ (* average residence time *)}$$

0.594595

- b) What is the 90th percentile time in the switch?

$$P90_{\text{msg}} = \frac{7}{3} R_{\text{msg}} \text{ (* approximate result *)}$$

1.38739

$$P90_{\text{exact}} = R_{\text{msg}} \text{Log}[10] \text{ (* exact result *)}$$

1.3691

- c) What is the increase in the response time if the traffic rate is increased by 10% into the switch?

$$\rho_{10} = (1.0 + 0.1) * \rho_{\text{msg}}$$

0.7744

$$R10_{\text{msg}} = \frac{S_{\text{msg}}}{1 - \rho_{10}} \text{ (* new residence time *)}$$

0.780142

$$\frac{R_{\text{msg}}}{R10_{\text{msg}}} 100 \text{ (* percent increase in residence time *)}$$

76.2162