

## Waiting Line Management

- Process Flow Concepts
- Waiting Line Models
- Examples
- Exercise
- More examples!

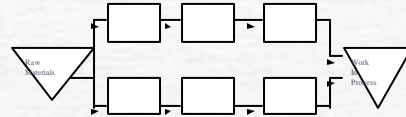


Developed by Jim Grayson, Ph.D.

1

## Example

The bakery operates two parallel lines, each equipped with a mixer, a proofing, and an oven. In addition, the bakery operates a single packaging line which is shared by the two bread making lines. A process flow diagram is shown below for the bakery.



Note: Mix CT =  $\frac{3}{4}$  hour / 100 loaves  
 Proof CT =  $\frac{3}{4}$  hour / 100 loaves  
 Bake CT = 1 hour / 100 loaves

Developed by Jim Grayson, Ph.D.

3

## Process Flow Terminology

**Throughput time** = amount of time each unit spends in the manufacturing process or the sum of the times for each of the production steps.

**Process flow diagram** = diagram depicting the activities in a process and the flows between them.

**Capacity** = the maximum rate of output of a process, measured in units of output per unit of time.

**Bottleneck** = the production resources that limits the capacity of the overall process (this is the step with the lowest capacity or longest cycle time).

**Cycle time** = average time between completion of successive units. It is directly related to output rate. (An output rate of 4 units per hour has a cycle time of 15 minutes.)

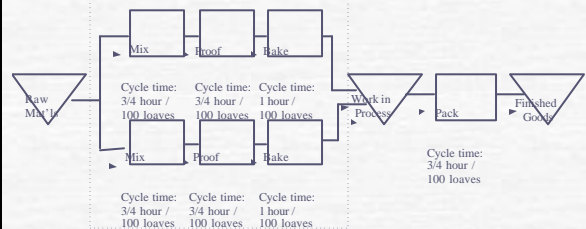
**Batch** = number of units of a particular product type that is produced before beginning production of another product type.

Developed by Jim Grayson, Ph.D.

2

Source: Capacity Analysis: Sample Problems. HBS case 9-696, 058.

## Bread making with two parallel lines

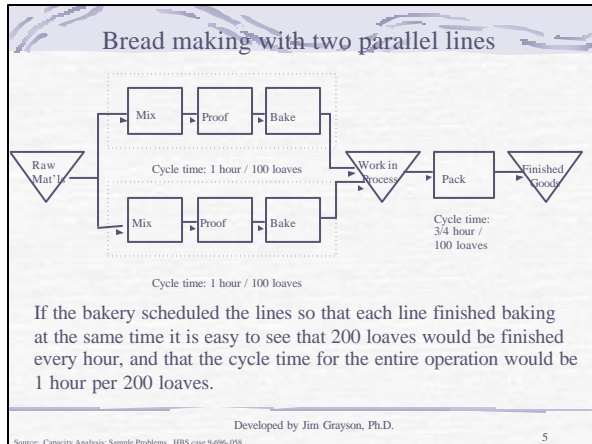


The ovens are the bottleneck in bread-making and constrain the cycle time of each bread-making line to 1 hour per 100 loaves.

Developed by Jim Grayson, Ph.D.

4

Source: Capacity Analysis: Sample Problems. HBS case 9-696, 058.



### Observations:

The entire process has a cycle time of 3/4 hour per 100 loaves. In order to minimize work in process the bakery operates the bread making process at 3/4 hour per 100 loaves by alternating beginning batches between the two mixers every 3/4 hour.

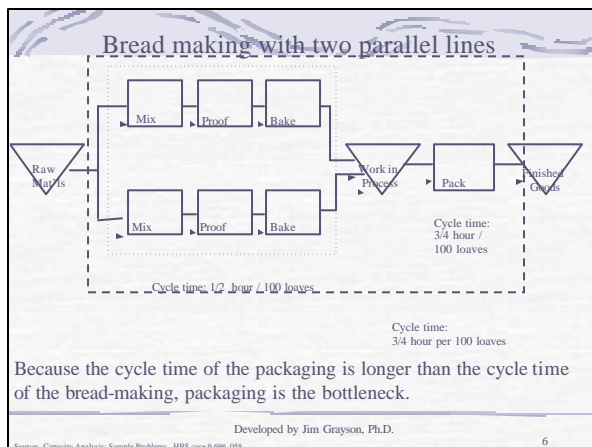
- If the bakery didn't slow down its bread making operation to match the packaging cycle time, loaves of bread would build up over time in the WIP inventory and become stale.
- Packaging, unless it ran overtime, would never have the capacity to keep up.

The bakery's overall daily capacity, assuming it operates the packaging line for 8 hours a day is:

$$(100 \text{ loaves} / 3/4 \text{ hour}) * 8 \text{ hours} = 1,066 \text{ loaves}$$

Developed by Jim Grayson, Ph.D. 7

Source: Capacity Analysis: Sample Problems. HBS case 9-696.058



### Waiting Line Model Notation:

1. Arrival pattern: M=exponential
2. Service times: M=exponential, D=deterministic
3. Number of parallel servers: 1=one server, etc.
4. Queue discipline: GD = general queue discipline
5. Maximum number of customers: 4 = infinite
6. Population size: 4 = infinite

Model	Parallel Servers	Service Pattern	Notation
1	1	Exponential	M/M/1/GD/4/4
2	1	Constant	M/D/1/GD/4/4
3	n	Exponential	M/M/n/GD/4/4

Developed by Jim Grayson, Ph.D. 8

### Example: Model I (M/M/1)

Drive up window at a fast food restaurant. Customers arrive at the rate of 25 per hour. Employee can serve one customer every two minutes. Assume Poisson arrival and exponential service times.

- A. Average utilization of employee?
- B. Average number of customers in line?
- C. Average number of customers in the system?
- D. Average waiting time in line?
- E. Average waiting time in the system?
- F. Probability that exactly two cars will be waiting in line?

Developed by Jim Grayson, Ph.D.

9

- B. Average number of customers in line?

$$L_q = \bar{n}_l = \frac{\lambda^2}{(m)(m-1)} = \frac{25^2}{30(30-25)} = 4.167$$

- C. Average number of customers in the system?

$$L_s = \bar{n}_s = \frac{\lambda}{m-1} = \frac{25}{30-25} = 5$$

Developed by Jim Grayson, Ph.D.

11

$$\begin{aligned} \lambda &= 25 \text{ customers / hour} \\ \mu &= 1/2\text{min} = 30 \text{ customers / hour} \end{aligned}$$

- A. Average utilization of employee?

$$\begin{aligned} UOF = r &= \lambda / m \\ &= \frac{25}{30} = 0.833 \end{aligned}$$

Developed by Jim Grayson, Ph.D.

10

- D. Average waiting time in line?

$$\begin{aligned} W_q = \bar{t}_l &= \frac{\lambda}{(m)(m-1)} = \frac{25}{30(30-25)} \\ &= .1667 \text{ hrs or 10 min} \end{aligned}$$

- E. Average waiting time in system?

$$\begin{aligned} W_s = \bar{t}_s &= \frac{1}{(m-1)} = \frac{1}{30-25} \\ &= .2 \text{ hrs or 12 min} \end{aligned}$$

Developed by Jim Grayson, Ph.D.

12

F. Probability exactly 2 cars waiting in line?

$$P(N) = \left[1 - \frac{1}{m}\right] \left[\frac{1}{m}\right]^N =$$

$$\left[1 - \frac{25}{30}\right] \left[\frac{25}{30}\right]^2 = 0.1157$$

Developed by Jim Grayson, Ph.D.

13

$\lambda = 1$  per 6 minutes = 10 per hour  
 $\mu =$  every 4 minutes = 15 per hour

A. Average number of customers in line.

$$L_q = \bar{n}_l = \frac{1^2}{(2m)(m-1)} =$$

$$\frac{10^2}{2(15)(15-10)} = 0.6667$$

Developed by Jim Grayson, Ph.D.

15

### Model 2 Example

An automated pizza vending machine heats and dispenses a slice of pizza in 4 minutes.

Customers arrive at a rate of one every 6 minutes with the arrival rate exhibiting a Poisson distribution.

Determine:

- A. Average number of customers in line.
- B. Average total waiting time in the system.

Developed by Jim Grayson, Ph.D.

14

B. Average total waiting time in the system?

$$W_q = \bar{t}_l = \frac{1}{(2m)(m-1)} = \frac{\bar{n}_l}{1} = \frac{L_q}{1}$$

$$= \frac{0.6667}{10} = 0.0667 \text{ hrs}$$

$$W_s = \bar{t}_s = W_q + \frac{1}{m} = 0.0667 + \frac{1}{15} =$$

0.1333 hrs or 8min

Developed by Jim Grayson, Ph.D.

16

**Model 3 Example:**

Same as Model 1 example, but with another server.

What is the effect of an additional server on the average number of cars in the system and total time customers wait before being served?

$$\begin{aligned} \lambda &= 25 \text{ customers / hour} \\ \mu &= 1/2\text{min} = 30 \text{ customers / hour} \\ S &= 2 \end{aligned}$$

$$P(0) = \frac{1}{\sum_{n=0}^{S-1} \frac{1}{n!} \left(\frac{I}{m}\right)^n + \frac{1}{S!} * \frac{S}{[S - (I/m)]} \left(\frac{I}{m}\right)^S}$$

Developed by Jim Grayson, Ph.D.

17

$$\begin{aligned} L_q = \bar{n}_l &= \frac{(I)(m)}{(S-1)!} * \frac{1}{(Sm-I)^2} \left(\frac{I}{m}\right)^S P(0) \\ &= \frac{(25)(30)}{1} * \frac{1}{[(2)(30)-25]^2} \left(\frac{25}{30}\right)^2 (0.413) \\ &= \frac{(25)(30)}{1225} (0.83)^2 (0.413) = 0.1742 \\ L_s = \bar{n}_s &= L_q + \frac{I}{m} = 0.1742 + \frac{25}{30} = 1.007 \end{aligned}$$

Developed by Jim Grayson, Ph.D.

19

$$P(0) =$$

$$\begin{aligned} &= \frac{1}{\left[\left(\frac{1}{0!}\right)\left(\frac{25}{30}\right)^0 + \left(\frac{1}{1!}\right)\left(\frac{25}{30}\right)^1\right] + \frac{1}{2!} * \frac{2}{\left[2 - \left(\frac{25}{30}\right)\right]} \left(\frac{25}{30}\right)^2} \\ &= \frac{1}{1.83 + \frac{1}{1.167} (.83)^2} = 0.413 \end{aligned}$$

Developed by Jim Grayson, Ph.D.

18

Total time customers wait before being served:

$$\begin{aligned} W_q = \bar{t}_l &= \frac{m}{(S-1)!} * \frac{1}{(Sm-I)^2} \left(\frac{I}{m}\right)^S P(0) \\ &= \frac{P(\text{wait})}{Sm-I} = \frac{L_q}{I} = \frac{0.1742}{25} = 0.0069 \text{ min} \end{aligned}$$

Developed by Jim Grayson, Ph.D.

20

**In Class Exercise:**

Customers arrive at the coffee urn at a local company at an average rate of 5/minute according to a Poisson distribution. The coffee urn has an adjustable spigot to vary the pouring rate. Most customers are regulars who know how to adjust the spigot and are able to obtain their coffee quickly. However a few are real klutzes and seem to take forever. As a result, the average time it takes a customer to pour a cup of coffee from the urn is 10 seconds and the pouring times follow an exponential distribution:

- A. On average, how many customers would be waiting in line?
- B. On the average, how long would you have to wait in line before getting to the coffee urn to pour yourself a cup of coffee?
- C. What percentage of the time is the coffee urn in use?
- D. What's the probability of finding 3 or more people around the urn?

[p. 512, class]

Developed by Jim Grayson, Ph.D.

21

B. Average times waiting in line and in the system:

$$W_q = \bar{t}_l = \frac{1}{(m)(m-1)} = \frac{5}{(6)(6-5)}$$

$$= 0.833 \text{ min or } 50 \text{ sec}$$

$$W_s = \bar{t}_s = \frac{1}{(m-1)} = \frac{1}{6-5} = 1 \text{ min}$$

Developed by Jim Grayson, Ph.D.

23

$$\lambda = 5 \text{ and } \mu = (1/10 \text{ seconds})(60 \text{ seconds/minute}) = 6/\text{minute}$$

A. Average number in line and in the system:

$$L_q = \bar{n}_l = \frac{1^2}{(m)(m-1)} = \frac{5^2}{(6)(6-5)} = 4.167$$

$$L_s = \bar{n}_s = \frac{1}{m-1} = \frac{5}{6-5} = 5.0$$

Developed by Jim Grayson, Ph.D.

22

C. Percentage of the time the urn is used:

$$UOF = r = 1 / m = \frac{5}{6} = 0.833 = 83.3\%$$

D. Probability of finding 3 or more people:

$$P(N \geq 3) = 1 - P(N < 3) = 1 - P(N \neq 2) = 1 - \{ P(N=0) + P(N=1) + P(N=2) \}$$

In this case, we can use a shortcut:

$$P(N > K) = (UOF)^{K+1} = P(N > 2)$$

$$= \left(\frac{5}{6}\right)^{2+1} = 0.5787 = 57.87\%$$

Developed by Jim Grayson, Ph.D.

24

If the manager installs an automatic device on the coffee urn to dispense a cup of coffee in a constant time of 10 seconds, how does that affect the answers previously computed?

The average numbers in the waiting line and in the system;

$$L_q = \bar{n}_l = \frac{I^2}{(2m)(m-1)} = \frac{5^2}{(2)(6)(6-5)} = 2.083$$

$$L_s = \bar{n}_s = L_q + r = 2.083 + \frac{5}{6} = 2.917$$

Developed by Jim Grayson, Ph.D.

25

Suppose that the manager installs a second coffee urn, each providing service on an exponential distribution with a mean service time of 10 seconds. How does this affect things?

The number of service channels, S, is 2.  
 $\lambda = 5/\text{minute}$  and  $\mu = 6/\text{minute}$ .

$$P(0) = \frac{1}{\left[1 + \frac{5}{6}\right] + \frac{1}{2} * \frac{2}{\left[2 - \left(\frac{5}{6}\right)\right]} \left(\frac{5}{6}\right)^2}$$

$$= \frac{42}{102} = 0.411765$$

Developed by Jim Grayson, Ph.D.

27

The average times waiting in line and in the system;

$$W_q = \bar{t}_l = \frac{I}{(2m)(m-1)} = \frac{5}{(2)(6)(6-5)} =$$

0.4167 min or 25 sec

$$W_s = \bar{t}_s = W_q + \frac{1}{m} = 25 + 10 \text{ seconds} = 35 \text{ sec}$$

Developed by Jim Grayson, Ph.D.

26

$$L_q = \bar{n}_l = \frac{(5)(6)}{(2-1)} * \frac{1}{((2)(6)-5)^2} \left(\frac{5}{6}\right)^2 (0.411765)$$

$$= \frac{125}{714} = 0.175070$$

$$L_s = \bar{n}_s = 0.175070 + 0.83333 = 1.008403$$

$$W_q = \bar{t}_l = \frac{L_q}{I} = \frac{0.175070}{5} = 0.035014 \text{ min or } 2.1 \text{ sec}$$

$$W_s = \bar{t}_s = 0.035014 + 0.166667 = 0.201681 \text{ min or } 12.1 \text{ sec}$$

Developed by Jim Grayson, Ph.D.

28

## Centralized Phone Scheduling System at Lourdes Hospital

[source: Chapter 12, *Applied Management Science* by Claus.]

Lourdes Hospital in Binghamton, New York, uses a centralized telephone system to schedule appointments for outpatients, inpatients, and other ambulatory services requested by patients, physicians, and hospital personnel (Agnihotri and Taylor, 1991). The central system handles calls for 13 departments, such as X-ray, the laboratory, the maternity clinic, and physical therapy. Complaints of poor service because of long delays in answering incoming calls plagued the system from the start. Management's initial response was to install an answering machine to queue calls on a FCFS basis. This solution was unacceptable because customers then complained about being placed on hold for several minutes.

Developed by Jim Grayson, Ph.D.

29

A sample of almost 1000 calls showed that service times could be closely approximated by an exponential distribution with a mean of 3.11 minutes. This gave an average service rate of 19.3 calls/hour ( $60/3.11 = 19.3$ ).

The number of arriving calls during the day varied with time, from lows of about 4 calls/hour at 7:00 a.m. and 5:00 p.m. to peaks of slightly more than 40 calls/hour from 9:00 to 11:30 a.m. and about 38/hour from 2:00 to 3:45 p.m. Because the intervals with similar arrival rates were reasonably long, those with similar rates were grouped together. The problem then was to determine the appropriate staffing level for each interval.

source: Chapter 12, *Applied Management Science* by Claus. Developed by Jim Grayson, Ph.D.

31

Patricia Taylor was a manager at the hospital at the time and was given the assignment of determining appropriate staffing levels for providing a satisfactory level of service. Taylor was then attending Agnihotri's class on service operations, and she chose the problem for her student project. A service level of 95% was selected—that is, the probability that an incoming telephone call would be put on hold and wait in the queue was to be no greater than 5%.

Developed by Jim Grayson, Ph.D.

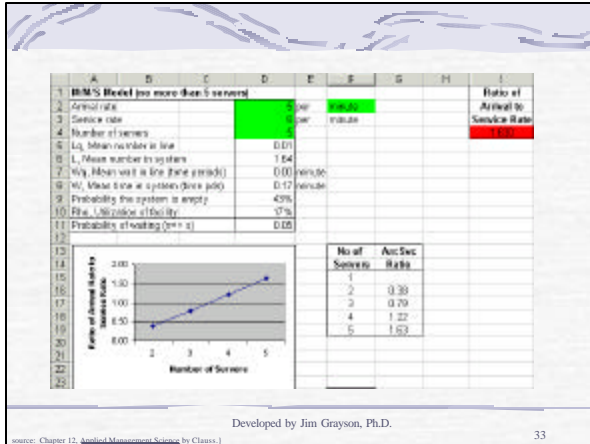
30

To determine the number of servers (operators) for a 95% service level we use the EXCEL Solver and this expression for the probability of waiting.

$$P(\text{wait}) = \frac{(I/m)^S}{(S-1)! \left[ S - \frac{I}{m} \right] \left[ \sum_{k=0}^{S-1} \frac{1}{k!} \left( \frac{I}{m} \right)^k \right] + \frac{(I/m)^S}{(S-1)! [S - (I/m)]}}$$

source: Chapter 12, *Applied Management Science* by Claus. Developed by Jim Grayson, Ph.D.

32



Developed by Jim Grayson, Ph.D.

33

Source: Chapter 17, Applied Management Science by Chase, J.

## St. Luke's Hematology Lab

(source: Chapter 15, Decision Modeling by Moore and Weatherford.)

St. Luke's treats a large number of patients on an outpatient basis: that is, there are many patients who come to the hospital to see the staff doctors for diagnosis and treatment but who are not admitted to the hospital. Outpatients plus those admitted to the 600-bed hospital produce a large flow of new patients each day. Most new patients must visit the hematology laboratory as part of the diagnostic process. Each such patient has to be seen by a technician. The system works like this: After seeing a doctor, the patient arrives at the laboratory and checks in with a clerk. Patients are assigned on a first-come, first-served basis to test rooms as they become available. The technician assigned to that room performs the tests ordered by the doctor. When the testing is complete, the patient goes on to the next step in the process (perhaps X-ray), and the technician sees a new patient.

Developed by Jim Grayson, Ph.D.

35

Developed by Jim Grayson, Ph.D.

34

Monte must decide how many technicians to hire. Superficially, at least, the trade-off is obvious. More technicians mean more expense for the hospital, but quicker service for the patients.

In the blood-testing model each patient joins a common queue and, on arriving at the head of the line, enters the first examining room that becomes available. This type of system must not be confused with a system in which a queue forms in front of each server, as in the typical grocery store.

Assume that the inter-arrival time is given by an exponential distribution with parameter  $\lambda = 0.20$  per minute. This implies that a new patient arrives every 5 minutes on the average, since

$$\text{mean inter-arrival} = 1/\lambda = 1/0.20 = 5$$

Developed by Jim Grayson, Ph.D.

36

Source: Chapter 15, Decision Modeling by Moore and Weatherford.

Also, assume that each server is identical and that each service time is given by an exponential distribution with parameter  $\mu = 0.125$  per minute. This implies that the mean service time is 8 minutes, since

$$\text{mean service time for an individual server} = 1/\mu = 1/0.125 = 8$$

Note that if there were only one server, the queue would grow without limit, since  $\lambda > \mu$  ( $0.20 > 0.125$ ). For a multi-server queue, however, a steady state will exist as long as  $\lambda < s\mu$ , where  $s$  is the number of servers. For example, if we have two servers, we will achieve a steady state because  $0.20 < 0.25$  ( $= 2 * 0.125$ ).

Monte selected the number of lab technicians to hire by looking at the operating characteristics and using his judgment. This is not an unusual approach in queuing models and is especially common in the not-for-profit sector. Monte realizes that he is balancing the cost of hiring more technicians against the costs he incurs by forcing the patients to wait. The cost of hiring additional technicians is fairly clear. The waiting cost is not.

Developed by Jim Grayson, Ph.D.

37

(source: Chapter 15, Decision Modeling by Moore and Weatherford.)

If you are willing and able to estimate certain costs, you can build expected cost models of queuing systems. Consider, for example, the hematology lab model (in general terms any multi-server queue with exponential inter-arrival and service times), and suppose the manager is willing to specify two costs:

$C_s$  = cost per hour of having a server available

$C_w$  = per hour of having a person wait in the system (a very "fuzzy" or qualitative cost)

Developed by Jim Grayson, Ph.D.

39

(source: Chapter 15, Decision Modeling by Moore and Weatherford.)

Monte first notes that the cost to the patient is irrelevant to his decision, except as it affects the patient's willingness to use the hospital. It really does not matter who is waiting—a consultant who charges \$250 per hour for his services or an unemployed person with no opportunity cost—unless the waiting time persuades the patient to use another health facility. This observation explains why certain monopolies like government agencies and utilities can be so casual about your waiting time. There is no place else to go!

Besides the possible effect on demand, the hematology lab could cost the hospital money if it reduced the output of the hospital. Suppose, for example, that the outpatient clinics could process 50 new patients each day, but that the hematology lab could handle only 10 patients. (This is clearly and extreme example to establish a point.) In this case, the hospital would be wasting a valuable resource, the doctors and other staff in the clinics, because of a bottleneck in the hematology lab. However, having stated this, it still is not easy to assess an explicit cost of a patient waiting.

Developed by Jim Grayson, Ph.D.

38

(source: Chapter 15, Decision Modeling by Moore and Weatherford.)

With these it is possible to calculate the total costs associated with the decision to use any particular number of servers. Let us start by calculating the total cost of hiring 2 servers for an 8-hour day. There are two components:

$$\text{server cost} = (C_s)(2)(8)$$

where  $C_s$  is the cost per hour for one server, 2 is the number of servers, and 8 is the number of hours each server works, and

$$\text{waiting cost} = (C_w)(L_2)(8)$$

where  $L_2$  is the number of people in the queue when there are 2 servers. This second calculation may not be as obvious, but the rationale is the same as for the server cost. If there are, on the average,  $L_2$  people waiting when the system has 2 servers, then  $L_2$  times 8 is the average number of waiting "hours" per day. Hence,  $(C_w)(L_2)(8)$  is the average waiting cost for the 8-hour day.

Developed by Jim Grayson, Ph.D.

40

(source: Chapter 15, Decision Modeling by Moore and Weatherford.)

If we wanted to calculate the total cost of using 4 servers for a 6-hour day, we would take

$$(C_s)(4)(6) + (C_w)(L_4)(6)$$

or

$$[(C_s)(4) + (C_w)(L_4)]6$$

The term in square brackets,  $[(C_s)(4) + (C_w)(L_4)]$ , then, is the total cost per hour of using 4 servers.

We now define

$$TC(s) = \text{total cost per hour of using } s \text{ servers}$$

and we see that

$$TC(s) = (C_s)(S) + (C_w)(L_s).$$

Developed by Jim Grayson, Ph.D.

41

Source: Chapter 15, Decision Modeling by Moore and Weatherford.

Next Monte creates a data table to determine the sensitivity of this decision to the “fuzzy” cost,  $C_w$ . He decides he wants to explore values for  $C_w$  from 0 to \$180. The steps for Monte to do this in his spreadsheet are:

- Set up the spreadsheet as shown.
- Enter the formulas for the quantities we want to track (total cost with 2 servers, total cost with 3 servers, total cost with 4 servers) in cells B10:D10. These formulas are =E6, =E7, and =E8, respectively.
- Highlight the range A12:D23, and click on Data, then “Table.”
- Enter the Column Input Cell as B2. Click on OK.
- Excel automatically fills in the table.
- Graph the results and reach conclusions.

Developed by Jim Grayson, Ph.D.

43

Source: Chapter 15, Decision Modeling by Moore and Weatherford.

Our goal is to choose  $s$ , the number of servers, to minimize this function. We can see that as  $S$  increases, the waiting cost will decrease and the server cost will increase. The idea is to find that value of  $s$  that minimizes the sum of these two costs.

In this example, let's put a relatively large cost on waiting and see if the decision changes from Monte's original decision of choosing two servers. We establish  $C_s = \$50/\text{server}/\text{hour}$  and  $C_w = \$100/\text{customer}/\text{hour}$ , and then we can calculate the server cost and waiting cost for 2, 3, and 4 servers. We'll assume we want to compare the cost over an 8-hour shift, and we must enter in the values for  $L$  for each value of  $s$  we want to explore. We can see that 3 servers minimize the Total Cost at \$2,730.

Developed by Jim Grayson, Ph.D.

42

Source: Chapter 15, Decision Modeling by Moore and Weatherford.

Number of Servers	Avg. No. in Queue	Server Cost	Waiting Cost	Total Cost
2	2			\$
3	3			\$
4	4			\$
		2 Servers	3 Servers	4 Servers
Cost of Waiting/hour				
0				
20				
40				
60				
80				
100				
120				
140				
160				
180				
200				

Developed by Jim Grayson, Ph.D.

44

Source: Chapter 15, Decision Modeling by Moore and Weatherford.