

Queueing Models

Queueing Theory

Where is there waiting?

Service facility

- Fast-food restaurants
- post office
- grocery store
- bank

Manufacturing

Equipment awaiting repair

Phone or computer network

Product orders

Why is there waiting?

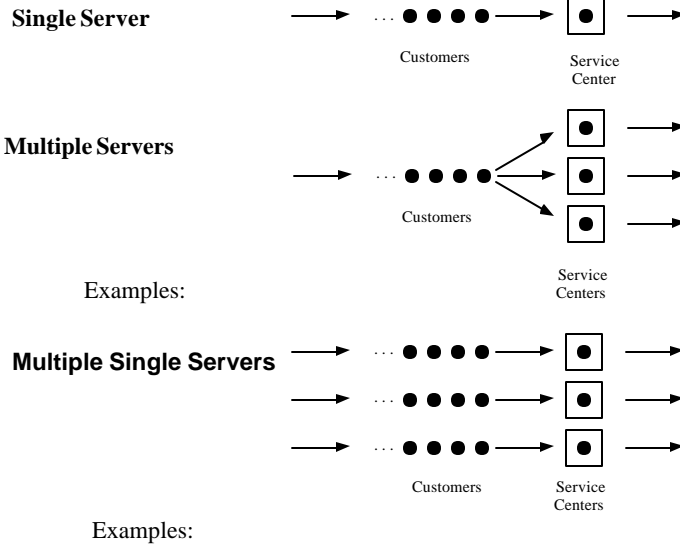
System Characteristics

- Number of servers
- Arrival and service pattern
- Queue discipline

Measures of System Performance

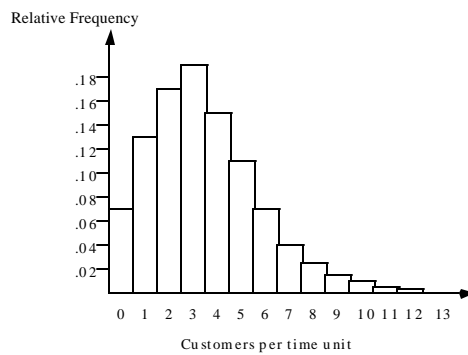
- Average number of customers waiting
- Average time customers wait
- System utilization

Number of Servers



Arrival Pattern

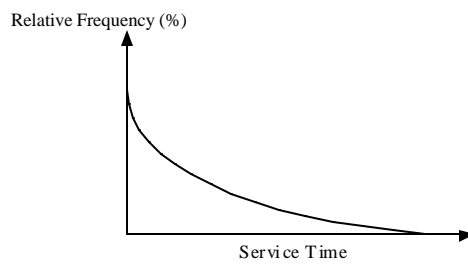
A Poisson distribution is usually assumed:



This also is referred to as having random arrivals.

Service Time

Either an exponential distribution is assumed



Examples:

OR any distribution (only single-server model is easily solved)

Examples:

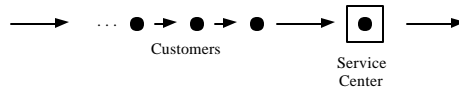
Irwin/McGraw-Hill

14-7

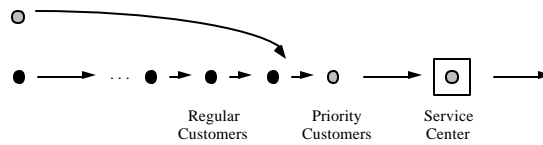
© The McGraw-Hill Companies, Inc., 2000

Queue Discipline

First come -- first served (FCFS):



Multiple Priorities:



Examples:

Irwin/McGraw-Hill

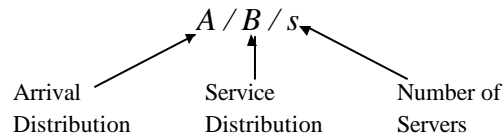
14-8

© The McGraw-Hill Companies, Inc., 2000

Some Models

1. Single server, exponential service time ($MM1$)
2. Single server, general service time ($MG/1$)
3. Multiple servers, exponential service time (MMs)

A Taxonomy



where

M = exponential distribution ("Markovian")
 D = deterministic (constant)
 G = general distribution

Given

I = customer arrival rate
 m = service rate ($1/\mu$ = average service time)
 s = number of servers

Calculate

L_q = average number of customers in the queue
 L = average number of customers in the system
 W_q = average waiting time in the queue
 W = average waiting time (including service)
 P_n = probability of having n customers in the system
 r = system utilization

Basic Relationships

The following relationships hold for *all* of the models.

- The system utilization is $r = \frac{I}{sm}$
- The average number of customers in the system is $L = L_q + \frac{I}{m}$
- The average waiting time in the queue is $W_q = \frac{L_q}{I}$
- The average total waiting time (including service) is $W = W_q + \frac{1}{m}$

Model 1 (M/M/1)

Formulas

Probability that system is empty: $P_0 = 1 - \frac{I}{m}$

Probability of n customers in system: $P_n = P_0 \left(\frac{I}{m}\right)^n$

Average number in queue: $L_q = \frac{I^2}{m(m-I)}$

Example

The reference desk at a library receives request for assistance at an average rate of 10 per hour (Poisson distribution). There is only one librarian at the reference desk, and he can serve customers in an average of 5 minutes (exponential distribution). What are the measures of performance for this system?

Template for M/M/s Queueing Model

Data		
$\lambda =$	10	(mean arrival rate)
$\mu =$	12	(mean service rate)
$s =$	1	(# servers)

Pr($w > t$) =	0.1353353
when t =	1

Prob($w_q > t$) =	0.1127794
when t =	1

Results	
L =	5
$L_q =$	4.16666667
W =	0.5
$W_q =$	0.41666667
$\rho =$	0.83333333
$P_0 =$	0.16666667
$P_1 =$	0.13888889
$P_2 =$	0.11574074
$P_3 =$	0.096450617
$P_4 =$	0.080375514
$P_5 =$	0.066979595

Model 2 (M/G/1)

Formulas

Average number in line:	$L_q = \frac{l^2 s^2 + r^2}{2(1-r)}$
Probability that system is empty:	$P_0 = 1 - \frac{l}{m}$
(Special case: M/D/1	$L_q = \frac{l^2}{2m(m-1)})$

Example

ABC Car Wash is an automated car wash. Each customer deposits four quarters in a coin slot, drives the car into the auto-washer, and waits while the car is automatically washed. Cars arrive randomly at an average rate of 20 cars per hour. The service time is exactly 2 minutes. What are the measures of performance?

Template for M/D/1 Queuing Model

Data		Results	
$\lambda =$	20	(mean arrival rate)	$L =$
$\mu =$	30	(mean service rate)	$L_q =$
$s =$	1	(# servers)	$W =$
			$W_q =$
			$\rho =$
			$P_0 =$

Irwin/McGraw-Hill
14-15
© The McGraw-Hill Companies, Inc., 2000

Model 3 (M/M/s)

Formulas

Probability that system is empty:
$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(1/m)^n}{n!} + \frac{(1/m)^s}{s!(1-r)}}$$

Probability of n customers in system:
$$P_n = \begin{cases} \frac{(1/m)^n}{n!} P_0 & \text{for } n = 1, \dots, s \\ \frac{(1/m)^n}{s!s^{n-s}} P_0 & \text{for } n \geq s \end{cases}$$

Probability a new arrival will have to wait:
$$P_w = \left(\frac{1}{m}\right)^s \frac{P_0}{s!(1-r)}$$

Average number in queue:
$$L_q = P_0 \left[\frac{(1/m)^{s+1}}{(s-1)!(s-1/m)^2} \right]$$

Irwin/McGraw-Hill
14-16
© The McGraw-Hill Companies, Inc., 2000

Example

The Federal Bank of Washington has three tellers at their Seattle branch. Customers arrive randomly at the branch at an average rate of 1 per minute. The service time averages 2 minutes, and follows the exponential distribution. What are the measures of performance?

Template for M/M/s Queueing Model		
Data		
$\lambda =$	60	(mean arrival rate)
$\mu =$	30	(mean service rate)
$s =$	3	(# servers)
$\text{Pr}(w > t) =$	1.341E-12	
when $t =$	1	
$\text{Prob}(w_q > t) =$	4.159E-14	
when $t =$	1	
Results		
$L =$	2.888888889	
$L_q =$	0.888888889	
$W =$	0.048148148	
$W_q =$	0.014814815	
$\rho =$	0.666666667	
$P_0 =$	0.111111111	
$P_1 =$	0.222222222	
$P_2 =$	0.222222222	
$P_3 =$	0.148148148	
$P_4 =$	0.098765432	
$P_5 =$	0.065843621	

Irwin/McGraw-Hill

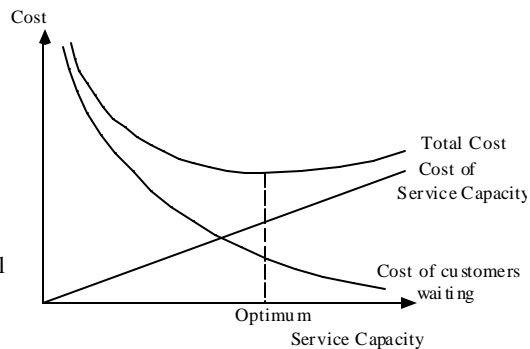
14-17

© The McGraw-Hill Companies, Inc., 2000

Application of Queueing Theory

We can use the results from queueing theory to make the following types of decisions:

- How many servers to employ
- Whether to use a single fast server or a number of slower servers
- Whether to have general purpose or faster specific servers



Goal:

Minimize total cost = cost of servers + cost of waiting

Irwin/McGraw-Hill

14-18

© The McGraw-Hill Companies, Inc., 2000

Example #1: How Many Servers?

In the service department of an auto repair shop, mechanics requiring parts for auto repair present their request forms at the parts department counter. A parts clerk fills a request while the mechanics wait. Mechanics arrive at an average rate of 40 per hour (Poisson). A clerk can fill requests in 3 minutes (exponential). If the parts clerks are paid \$6 per hour and the mechanics are paid \$18 per hour, what is the optimal number of clerks to staff the counter.

Template for Economic Analysis of M/M/s Queueing Model		
Data		
$\lambda =$	40	(mean arrival rate)
$\mu =$	20	(mean service rate)
$s =$	3	(# servers)
Pr($\omega > t$) = 0.0012726		
when t =	1	
Prob($a_k > t$) = 0.0008484		
when t =	1	
$C_s =$	\$6	(cost/server/unit time)
$C_w =$	\$18	(waiting cost/unit time)
Cost of Service =	\$18.00	
Cost of Waiting =	\$52.00	
Total Cost =	\$70.00	
Results		
L =	2.88889	
$L_q =$	0.88889	
W =	0.07222	
$W_q =$	0.02222	
$\rho =$	0.6666667	
$P_0 =$	0.11111	
$P_1 =$	0.14815	
$P_2 =$	0.09877	
$P_3 =$	0.06584	
$P_4 =$	0.04390	
$P_5 =$	0.02926	
$P_6 =$	0.01951	

Irwin/McGraw-Hill

14-19

© The McGraw-Hill Companies, Inc., 2000

Template for Economic Analysis of M/M/s Queueing Model		
Data		
$\lambda =$	40	(mean arrival rate)
$\mu =$	20	(mean service rate)
$s =$	4	(# servers)
Pr($\omega > t$) = 4.54E-05		
when t =	1	
Prob($a_k > t$) = 2.27E-05		
when t =	1	
$C_s =$	\$6	(cost/server/unit time)
$C_w =$	\$18	(waiting cost/unit time)
Cost of Service =	\$24.00	
Cost of Waiting =	\$39.13	
Total Cost =	\$63.13	
Results		
L =	2.17391	
$L_q =$	0.17391	
W =	0.05435	
$W_q =$	0.00435	
$\rho =$	0.5	
$P_0 =$	0.13043	
$P_1 =$	0.17391	
$P_2 =$	0.08696	
$P_3 =$	0.04348	
$P_4 =$	0.02174	
$P_5 =$	0.01087	
$P_6 =$	0.00543	

Irwin/McGraw-Hill

14-20

© The McGraw-Hill Companies, Inc., 2000

Template for Economic Analysis of M/M/s Queuing Model

Data			Results	
$\lambda =$	40	(mean arrival rate)	$L =$	2.03980
$\mu =$	20	(mean service rate)	$L_q =$	0.03980
$s =$	5	(# servers)	$W =$	0.05100
$\Pr(\omega > t) =$	6.144E-06		$W_q =$	0.00100
when $t =$	1		$\rho =$	0.4
$\text{Prob}(a_i > t) =$	2.458E-06		$P_0 =$	0.13433
when $t =$	1		$P_1 =$	0.17910
$C_s =$	\$6	(cost/server/unit time)	$P_2 =$	0.08955
$C_w =$	\$18	(waiting cost/unit time)	$P_3 =$	0.03582
Cost of Service =	\$30.00		$P_4 =$	0.01433
Cost of Waiting =	\$36.72		$P_5 =$	0.00573
Total Cost =	\$66.72		$P_6 =$	0.00229

So $s = 4$ has the smallest total cost.

Example #2: How Many Servers?

Beefy Burgers is trying to decide how many registers to have open during their busiest time, the lunch hour. Customers arrive during the lunch hour at a rate of 98 customers per hour (Poisson distribution). Each service takes an average of 3 minutes (exponential distribution). Management would not like the average customer to wait longer than five minutes in line. How many registers should they open?

Template for M/M/s Queueing Model

Data		Results	
$\lambda =$	98 (mean arrival rate)	$L =$	51.4655
$\mu =$	20 (mean service rate)	$L_q =$	46.5655
$s =$	5 (# servers)	$W =$	0.5252
$\Pr(\omega > t) =$	0.1429016	$W_q =$	0.4752
when $t =$	1	$\rho =$	0.9800
$\text{Prob}(\omega_q > t) =$	0.1286114	$P_0 =$	0.0008
when $t =$	1		

Template for M/M/s Queueing Model

Data		Results	
$\lambda =$	98 (mean arrival rate)	$L =$	7.3593
$\mu =$	20 (mean service rate)	$L_q =$	2.4593
$s =$	6 (# servers)	$W =$	0.0751
$\Pr(\omega > t) =$	1.19E-08	$W_q =$	0.0251
when $t =$	1	$\rho =$	0.8167
$\text{Prob}(\omega_q > t) =$	1.54E-10	$P_0 =$	0.0053
when $t =$	1		

Choose $s = 6$ since $W = 0.0751$ hour is less than 5 minutes.

Irwin/McGraw-Hill 14-23 © The McGraw-Hill Companies, Inc., 2000

Example #3: One Fast Server or Many Slow Servers?

Beefy Burgers is considering changing the way that they serve customers. For most of the day (all but their lunch hour), they have three registers open. Customers arrive at an average rate of 50 per hour. Each cashier takes the customer's order, collects the money, and then gets the burgers and pours the drinks. This takes an average of 3 minutes per customer (exponential distribution). They are considering having just one cash register. While one person takes the order and collects the money, another will pour the drinks and another will get the burgers. The three together think they can serve a customer in an average of 1 minute. Should they switch to one register?

3 Slow Servers

Template for M/M/s Queueing Model	
Data	
$\lambda =$	50 (mean arrival rate)
$\mu =$	20 (mean service rate)
$s =$	3 (# servers)
$\Pr(\omega > t) =$ 6.376E-05 when $t =$ 1	
$\text{Prob}(\omega_q > t) =$ 3.188E-05 when $t =$ 1	
Results	
$L =$	6.0112
$L_q =$	3.5112
$W =$	0.1202
$W_q =$	0.0702
$\rho =$	0.8333
$P_0 =$	0.0449

1 Fast Server

Template for M/M/s Queueing Model	
Data	
$\lambda =$	50 (mean arrival rate)
$\mu =$	60 (mean service rate)
$s =$	1 (# servers)
$\Pr(\omega > t) =$ 4.54E-05 when $t =$ 1	
$\text{Prob}(\omega_q > t) =$ 3.783E-05 when $t =$ 1	
Results	
$L =$	5.0000
$L_q =$	4.1667
$W =$	0.1000
$W_q =$	0.0833
$\rho =$	0.8333
$P_0 =$	0.1667

W is less for one fast server, so choose this option.

Example #4: General or Specific Servers?

A small bank in a mall has two tellers. One handles only merchant customers and one handles regular customers. Merchant customers and regular customers each arrive at an average rate of 20 per hour (for a total arrival rate of 40 customers per hour). The service time for both tellers averages 2 minutes (exponential). The bank manager is considering changing the setup to allow each teller to handle both merchant customers and regular customers. Since the tellers would have to handle both types of jobs, their efficiency would decrease to a mean service time of 2.2 minutes. Should they switch to the new setup?

Specific Tellers (1 for Merchants, 1 for Regular)

Template for M/M/s Queueing Model

Data		Results	
$\lambda =$	20 (mean arrival rate)	L =	2.0000
$\mu =$	30 (mean service rate)	$L_q =$	1.3333
s =	1 (# servers)	W =	0.1000
Pr($\omega > t$) = 4.54E-05		$W_q =$	0.0667
when t =	1	$\rho =$	0.6667
Prob($\omega_q > t$) = 3.027E-05		$P_0 =$	0.3333
when t =	1		

General Tellers (2 Tellers Handle Both Jobs)

Template for M/M/s Queueing Model

Data		Results	
$\lambda =$	40 (mean arrival rate)	L =	3.1731
$\mu =$	27.2727 (mean service rate)	$L_q =$	1.7064
s =	2 (# servers)	W =	0.0793
Pr($\omega > t$) = 6.408E-07		$W_q =$	0.0427
when t =	1	$\rho =$	0.7333
Prob($\omega_q > t$) = 2.991E-07		$P_0 =$	0.1538
when t =	1		

Waiting times are less when both tellers handle both jobs, so choose this option.