



BUAD 467/667
Service Management
Spring 2008

Professor Patrick T. Harker

Class 4a
Capacity Design I

Copyright P.T. Harker 2008 Page 1



Outline for the Class

- Strategies for Managing Capacity
- What is Capacity?
- Queues: The Psychology of Waiting
- Is A Line Always Bad?
- What is a Queuing System?

Copyright P.T. Harker 2008 Page 2

Strategies for Managing Capacity

Supply-Side

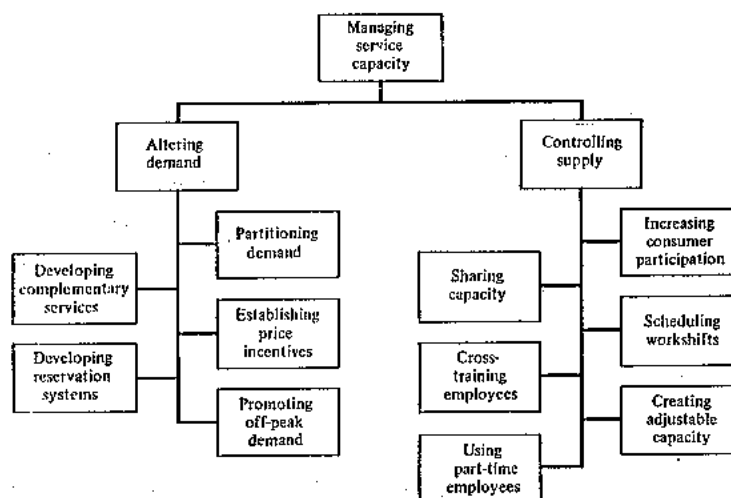
- add staff or resources
- reorganize the work process
- co-produce!

Demand-Side

- take 'em as they come
- reservation systems
- yield management systems

Copyright P.T. Harker 2008 Page 3

Capacity Management Strategies



Copyright P.T. Harker 2008 Page 4

What is Capacity?

- is there a finite limit to the number of customers that can be handled?
- if not, how do you define the limit?
 - ◆ number of customers?
 - ◆ length of the line?
 - ◆ waiting time?
- they're all related!! The need for *queuing theory*.

Queues

Queues arise in a variety of service situations:

- physical lines
- on-hold on the telephone
- multitasking on computer networks
- transportation: **BULK** queues
- Police and EMV's: service comes to you!

We wait 5 years of our lives in line!!!!

The Psychology of Waiting

- waiting as psychological punishment
 - ◆ put mirrors in elevators
 - ◆ spread the wait between various servers
 - ◆ make the customer do some work while waiting
 - ◆ numb the wait: bars in restaurants
 - ◆ distract the folks in the queue

Carmon: *it's the end that matters!*

Copyright P.T. Harker 2008 Page 7

The Psychology of Waiting

- waiting as ritual insult:
 - ◆ sensitivity training
 - ◆ make **some** initial contact
 - ◆ keep the customer informed: Disney
- waiting as social interaction

These things matter more than you think!!!

Copyright P.T. Harker 2008 Page 8

Is a Line Always Bad?

- queues as quality signals
 - ◆ restaurants
 - ◆ doctors
 - ◆ UD?
- quality as a function of others in the “game”
 - ◆ the lonely restaurant patron
 - ◆ the barn-size nightclub

Copyright P.T. Harker 2008 Page 9

If You Have a Persistent Line, Why Not Add Capacity?

Why doesn't the NFL raise the price for Superbowl tickets?

The Risks of Being “Out”

- if quality depends on others being there, can be “out” fast in capacity is increased too much.
- risk-hating owners
- stay away from societal pressure

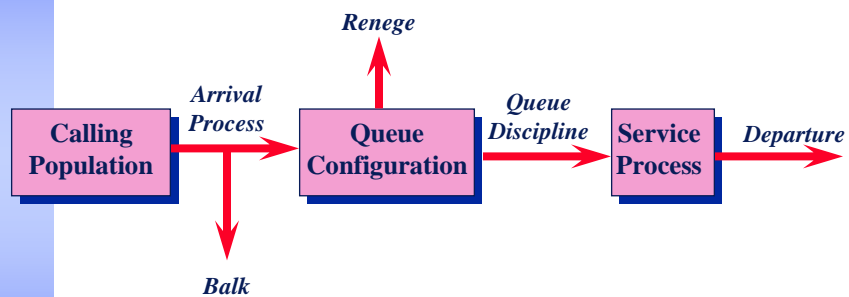
Copyright P.T. Harker 2008 Page 10

What is a Queuing System?

Components of a queuing system include:

- calling population
- arrival process
- queue configuration
- queue discipline
- service process
- departure process

Queueing System



Calling Population

Homogeneous

vs.

Heterogeneous

Arrival Process

- time of day/ week/ year
- choose the time period of the analysis carefully (e.g., investing on plows based on the average annual snowfall!)
- common assumption: *Poisson distribution*

Example of an Arrival Process with Poisson/ Negative Exponential Characteristics

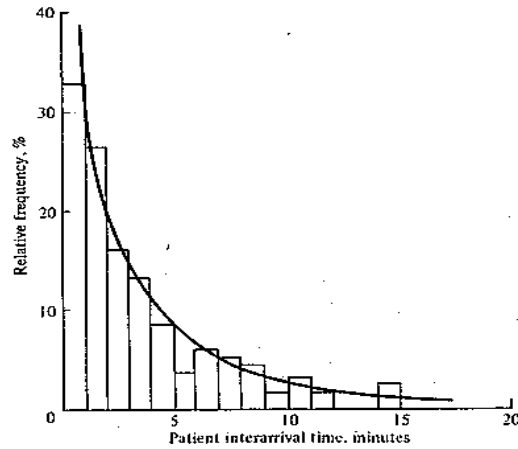
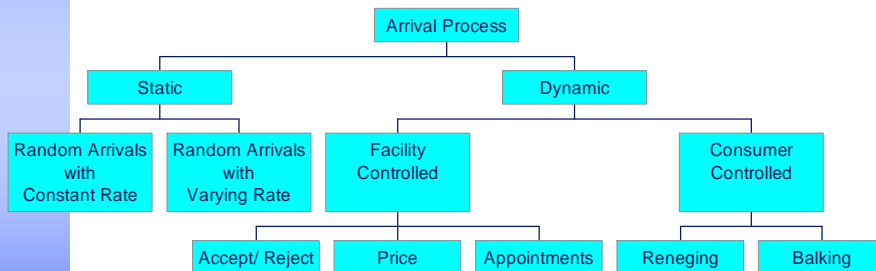


Figure 11.3 Distribution of patient interarrival times for a university health clinic. [Adapted with permission from E. J. Rising, R. Baron, and B. Averill, "A Systems Analysis of a University Health-Service Out-patient Clinic, *Operations Research*, September 1972, p. 1038.]

Copyright P.T. Harker 2008 Page 15

Classification of Arrival Processes



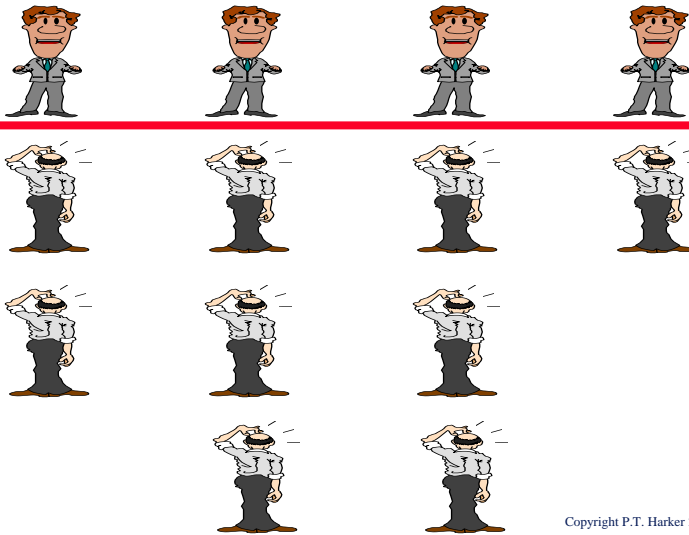
Copyright P.T. Harker 2008 Page 16

Queue Configurations

- multiple queue
- single queue
- take a number
- random

Copyright P.T. Harker 2008 Page 17

Queue Configurations: *Multiple Queues*



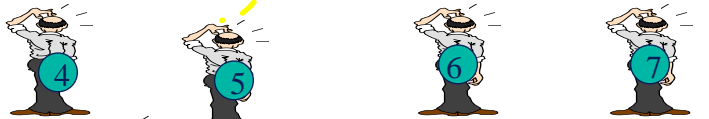
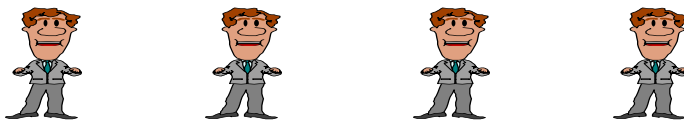
Copyright P.T. Harker 2008 Page 18

Queue Configurations: *Single or Disney Queues*



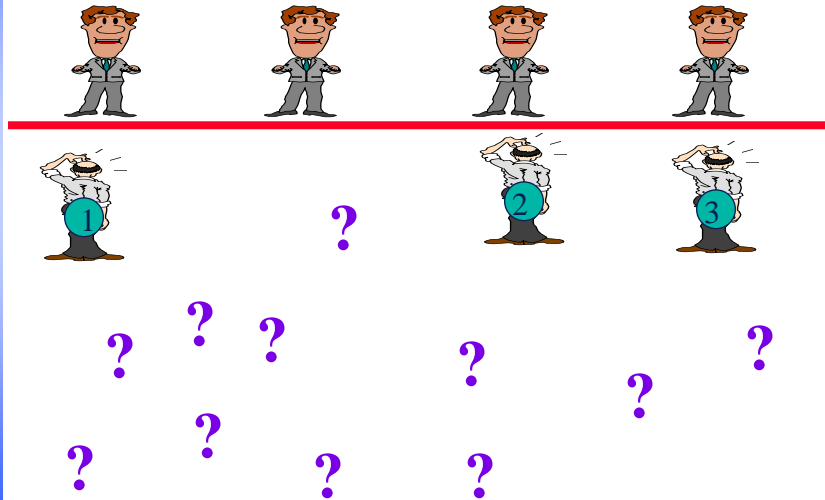
Copyright P.T. Harker 2008 Page 19

Queue Configurations: *Take a Number*



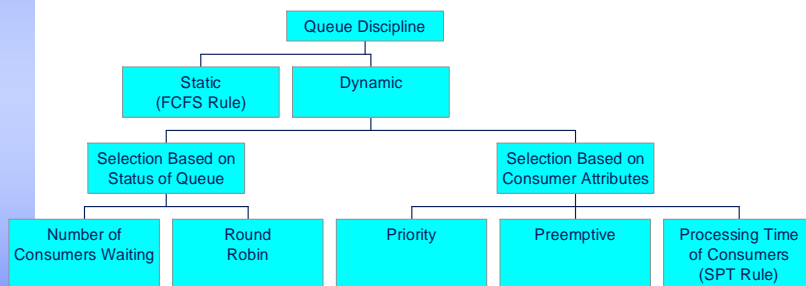
Copyright P.T. Harker 2008 Page 20

Queue Configurations: *Random Choice*



Copyright P.T. Harker 2008 Page 21

Queue Disciplines



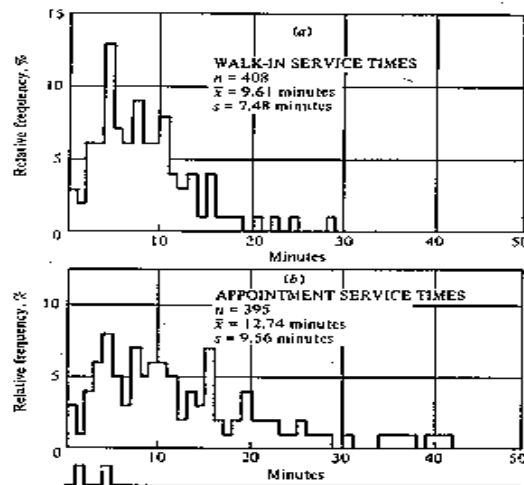
Copyright P.T. Harker 2008 Page 22

Service Process

- Poisson rises again!
- Organization can include
 - ◆ self-serve (parking lot)
 - ◆ servers in parallel (toll booth)
 - ◆ servers in series (factory)
 - ◆ self-serve first, parallel later (supermarket)
 - ◆ many service centers in parallel and series, not all used by a consumer (hospital)

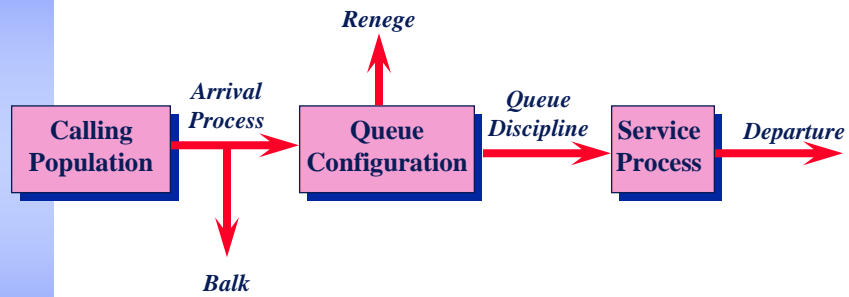
Copyright P.T. Harker 2008 Page 23

Service Process Examples



Copyright P.T. Harker 2008 Page 24

Queueing System: *The Challenge*



Matching the server capacity with demand arrivals:

- change the number of servers
- change the configuration of the servers and/or queue
- change the demand arrival process

Copyright P.T. Harker 2008 Page 25

BUAD 467/667 Service Management Spring 2008

Professor Patrick T. Harker

Class 4b
Capacity Design II

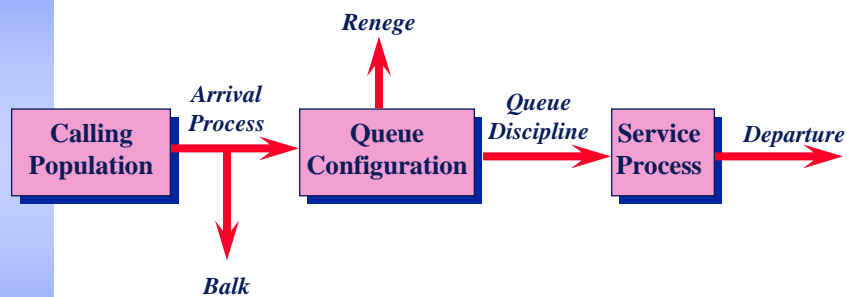
Copyright P.T. Harker 2008 Page 26

Outline for the Class

- Simple Analytical Models for Queues
- Excel to the Rescue!
- The Pooling Effect in Queues
- Capacity Planning with Queuing Models
- Next Class

Copyright P.T. Harker 2008 Page 27

Queueing System



Copyright P.T. Harker 2008 Page 28

Classification of Queues

A/B/C classification scheme

A = distribution of the interarrival times

B = distribution of the service times

C = number of parallel servers

M = exponential

D = deterministic

G = general

Note the difference between steady-state and transient behavior - queuing versus simulation

Copyright P.T. Harker 2008 Page 29

Basic Notation

λ = mean arrival rate (units per time period)

μ = mean service rate (units per time period)

$\rho = \lambda / \mu < 1$ (traffic intensity)

c = number of servers

P_0 = probability that there are zero in the system

P_n = probability that there are n in the system

L_s = mean number of units in the system

L_q = mean number of units in the queue

W_s = mean time in the system

W_q = mean time in the queue

Copyright P.T. Harker 2008 Page 30

Little's Formula

$$L_q = \lambda W_q$$

Queue length = arrival rate * time in the queue

Copyright P.T. Harker 2008 Page 31

Negative Exponential Distribution

- related to a *Poisson process*

- ◆ no memory in the system
- ◆ equal likelihood of arrivals & independent sources
- ◆ quite common in practice

Coefficient of Variation = Std dev. of Arrival Time

Expected Value

= 0 *deterministic*

= 1 *Poisson*

Same for processing times

Copyright P.T. Harker 2008 Page 32

Example of an Arrival Process with Poisson/ Negative Exponential Characteristics

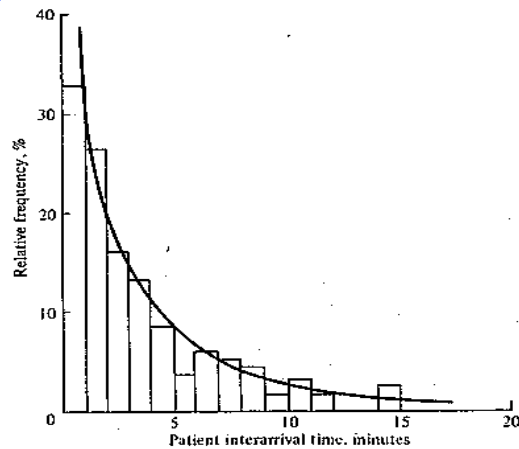


Figure 11.3 Distribution of patient interarrival times for a university health clinic. [Adapted with permission from E. J. Rising, R. Baron, and B. Averill, "A Systems Analysis of a University Health-Service Out-patient Clinic, *Operations Research*, September 1972, p. 1038.]

Copyright P.T. Harker 2008 Page 33

M/M/1 Model: The Workhorse

1. calling population - infinite with independent arrivals
2. arrival process - Poisson/ exponential
3. queue configuration - simple line with no queue capacity, balking or reneging
4. queue discipline - FIFO
5. service process - exponential distribution

Copyright P.T. Harker 2008 Page 34

M/M/1 Facts

$$P_0 = 1 - \rho$$

$$P_n = P_0 \rho^n$$

$$L_s = \lambda / (\mu - \lambda)$$

$$W_s = 1 / (\mu - \lambda)$$

$$L_q = \rho \lambda / (\mu - \lambda)$$

$$W_q = \rho / (\mu - \lambda)$$

$$\rho < 1$$

Copyright P.T. Harker 2008 Page 35

Example: *The Boat Ramp*

arrivals = 6 boats per hours

can launch 10 boats per hour or 1 boat every 6 minutes

$\rho = 6/10 = 0.6$ ramp is busy 60% of the time

$W_q = (0.6)/(10-6) = 0.15$ hours or 9 minutes wait

$W_s = 1/(10-6) = 0.25$ hours or 15 minutes total

$L_s = 6/(10-6) = 1.5$ boats in the system

$L_q = 0.6*6/(10-6) = 0.9$ boats waiting

$$\rho < 1$$

Copyright P.T. Harker 2008 Page 36

The Boat Ramp: continued

Problem: not enough parking spots.

Want enough spots for 90% of the time

n	P_n	$P(\text{no. of boats} < n)$
0	0.4	0.4
1	0.24	0.64
2	0.144	0.784
3	0.0864	0.8704
4	0.05184	0.92224

Copyright P.T. Harker 2008 Page 37

Excel to the Rescue!

ggs.xls G/G/s Queuing Formula Spreadsheet

Inputs:

lambda	6
mu	10
Ca ²	1
Cb ²	1

Definitions of terms:

lambda	= arrival rate
mu	= service rate
s	= number of servers
Ca ²	= squared coeff. of variation of arrivals
Cb ²	= squared coeff. of variation of service times
Nq	= average length of the queue
Ns	= average number in the system
Wq	= average wait in the queue
Ws	= average wait in the system (lambda/mu)
P(0)	= probability of zero customers in the system
P(delay)	= probability that an arriving customer has to wait

0.6

Outputs:

s	Nq	Ns	Wq	Ws	P(delay)	Utilization
1	0.900000	1.500000	0.150000	0.250000	0.600000	0.600000
2	0.059341	0.659341	0.009890	0.109890	0.138462	0.300000
3	0.006164	0.606164	0.001027	0.101027	0.024658	0.200000
4	0.000615	0.600615	0.000103	0.100103	0.003486	0.150000
5	0.000055	0.600055	0.000009	0.100009	0.000404	0.120000
6	0.000004	0.600004	0.000001	0.100001	0.000040	0.100000
7	0.000000	0.600000	0.000000	0.100000	0.000003	0.085714

Copyright P.T. Harker 2008 Page 38

The Pooling Effect in Queues

One secretary is assigned to each of four departments: accounting, finance, marketing and OM.

The Dean's getting complaints and is **not** a happy camper!

Two requests per hour arrive in all departments except OM, where 3 requests per hour arrive

All secretaries can service requests in 15 minutes (on average) or 4 requests per hour.

Copyright P.T. Harker 2008 Page 39

When Run as Separate Departments:

4 M/M/1 queues

$W_s = 1/(4-2) = 0.5$ hours or 30 minutes for three departments

$W_s = 1/(4-3) = 1$ hour in OM!

Copyright P.T. Harker 2008 Page 40

When We Have a Secretary Pool:

One M/M/4 queue

$\lambda = 2 + 2 + 2 + 3 = 9$ requests per hour

$\mu = 4$ requests per hour

$c = 4$ servers

s	Nq	Ns	Wq	Ws	P(delay)	Utilization
1	infinity	infinity	infinity	infinity	1.000000	1.000000
2	infinity	infinity	infinity	infinity	1.000000	1.000000
3	1.703271	3.953271	0.189252	0.439252	0.567757	0.750000
4	0.310086	2.560086	0.034454	0.284454	0.241178	0.562500

$W_s = 0.28$ hours or 17 minutes!

Copyright P.T. Harker 2008 Page 41

Capacity Planning With Queuing Models

One can plan the needed capacity based on:

- waiting times (phone centers)
- probability of excess waiting
- probability of lost sales
- maximize revenue (LL Bean methods)

Copyright P.T. Harker 2008 Page 42

Example: *Gas Pumps*

48 cars arrive per hour; 50% for self-serve and 50% for full-serve

On average, takes 5 minutes to service (12 cars/hour)

Two independent M/M/c systems with 24 cars per hour each.

Want to pick the number of pumps so that a customer only needs to wait 5% of the time or less

Copyright P.T. Harker 2008 Page 43

Gas Pumps: (continued)

s	Nq	Ns	Wq	Ws	P(delay)	Utilization
0						
1	infinity	infinity	infinity	infinity	1.000000	1.000000
2	infinity	infinity	infinity	infinity	1.000000	1.000000
3	0.888889	2.888889	0.037037	0.120370	0.444444	0.666667
4	0.173913	2.173913	0.007246	0.090580	0.173913	0.500000
5	0.039801	2.039801	0.001658	0.084992	0.059701	0.400000
6	0.009009	2.009009	0.000375	0.083709	0.018018	0.333333
7	0.001924	2.001924	0.000080	0.083414	0.004811	0.285714
8	0.000382	2.000382	0.000016	0.083349	0.001146	0.250000

At 6 pumps, probability of delay < 5%

Copyright P.T. Harker 2008 Page 44

Basic Lessons in Queuing

- excess/ idle capacity is not necessarily bad!
- pooling will create *economies of scale*; may need to fight against this strong tendency
- be careful of expansion risks!

Copyright P.T. Harker 2008 Page 45

Next Class:

- Read the materials on call centers to be prepared to discuss with our guest speaker
- Understand *Pronto Pizza*
- Project proposals are due

Copyright P.T. Harker 2008 Page 46