Queue management

Outline

- Characteristics of a Waiting-Line System.
  - Arrival characteristics.
  - Waiting-Line characteristics.
  - Service facility characteristics.
- Waiting Line (Queuing) Models.
  - M/M/1: One server.
  - M/M/2: Two servers.
  - M/M/S: S servers.
  - Cost comparisons.

Waiting Lines

- First studied by A. K. Erlang in 1913.
  - Analyzed telephone facilities.
- Body of knowledge called queuing theory.
  - Queue is another name for waiting line.
- Decision problem:
  - Balance cost of providing good service with cost of customers waiting.

You've Been There Before!

The average person spends 5 years waiting in line!!

'The other line always moves faster.'

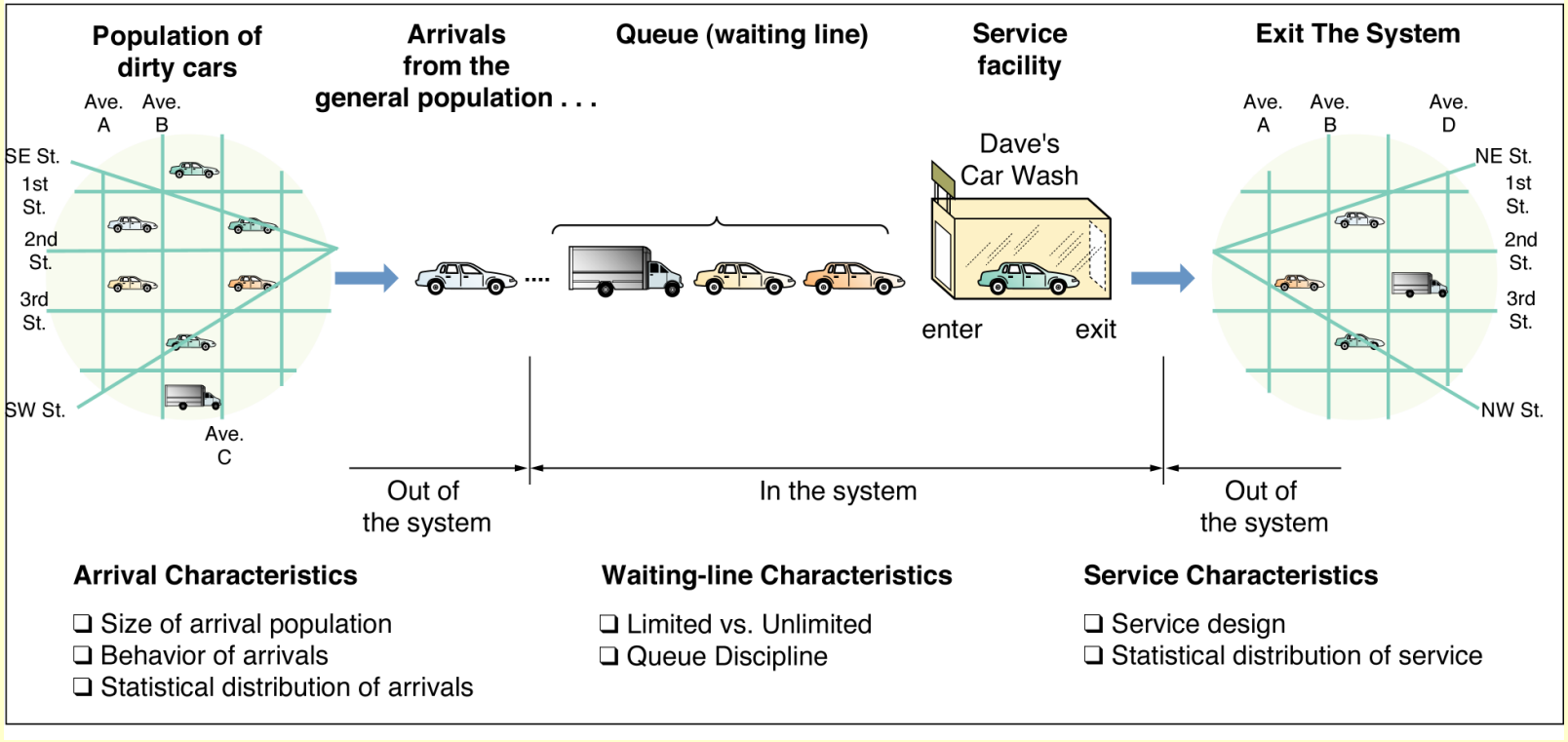
Waiting Line Examples

Situation	Arrivals	Servers	Service Process
Bank	Customers	Teller	Deposit etc.
Doctor's office	Patient	Doctor	Treatment
Traffic intersection	Cars	Traffic Signal	Controlled passage
Assembly line	Parts	Workers	Assembly

Waiting Line Components

- **Arrivals**: Customers (people, machines, calls, etc.) that demand service.
- **Waiting Line (Queue)**: Arrivals waiting for a free server.
- **Servers**: People or machines that provide service to the arrivals.
- **Service System**: Includes waiting line and servers.

Car Wash Example



Key Tradeoff

- Higher service level (more servers, faster servers)

⇒ Higher costs to provide service.

⇒ Lower cost for customers waiting in line (less waiting time).

Waiting Line Terminology

- **Queue**: Waiting line.
- **Arrival**: 1 person, machine, part, etc. that arrives and demands service.
- **Queue discipline**: Rules for determining the order that arrivals receive service.
- **Channels**: Parallel servers.
- **Phases**: Sequential stages in service.

Input Characteristics

- Input source (population) size.
  - Infinite: Number in service does not affect probability of a new arrival.
    - A very large population can be treated as infinite.
  - Finite: Number in service affects probability of a new arrival.
    - Example: Population = 10 aircraft that may need repair.
- Arrival pattern.
  - Random: Use Poisson probability distribution.
  - Non-random: Appointments.

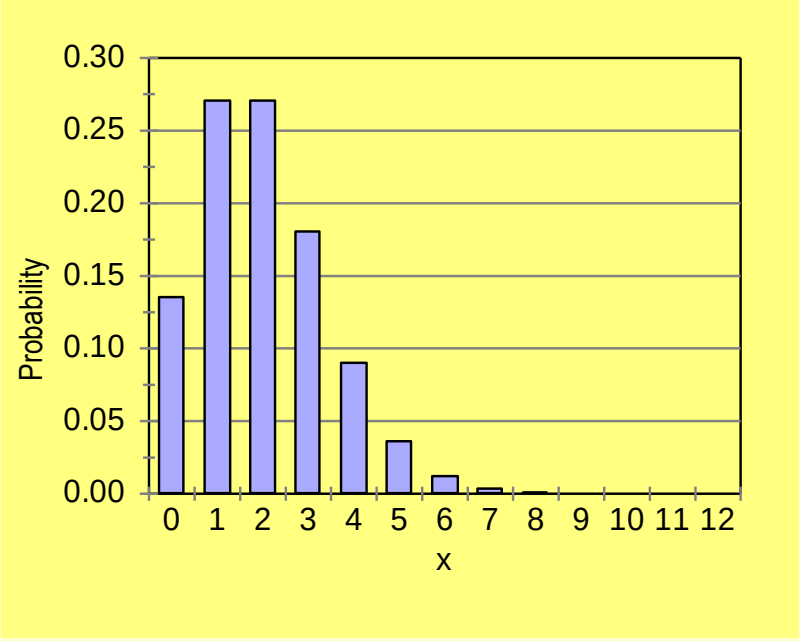
Poisson Distribution

- Number of events that occur in an interval of time.
  - Example: Number of customers that arrive each half-hour.
- Discrete distribution with mean = λ
  - Example: Mean arrival rate = 5/hour .
  - Probability:

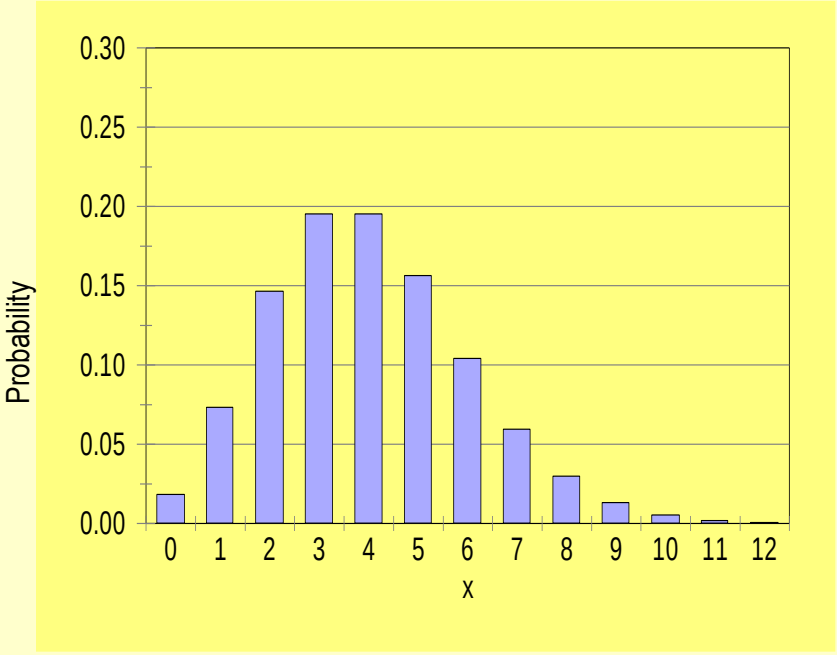
$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x}$$

- Time between arrivals has a negative exponential distribution.

Poisson Probability Distribution



$\lambda=2$



$\lambda=4$

Behavior of Arrivals

- Patient.
  - Arrivals will wait in line for service.
- Impatient.
  - Balk: Arrival leaves before entering line.
    - Arrival sees long line and decides to leave.
  - Renege: Arrival leaves after waiting in line a while.

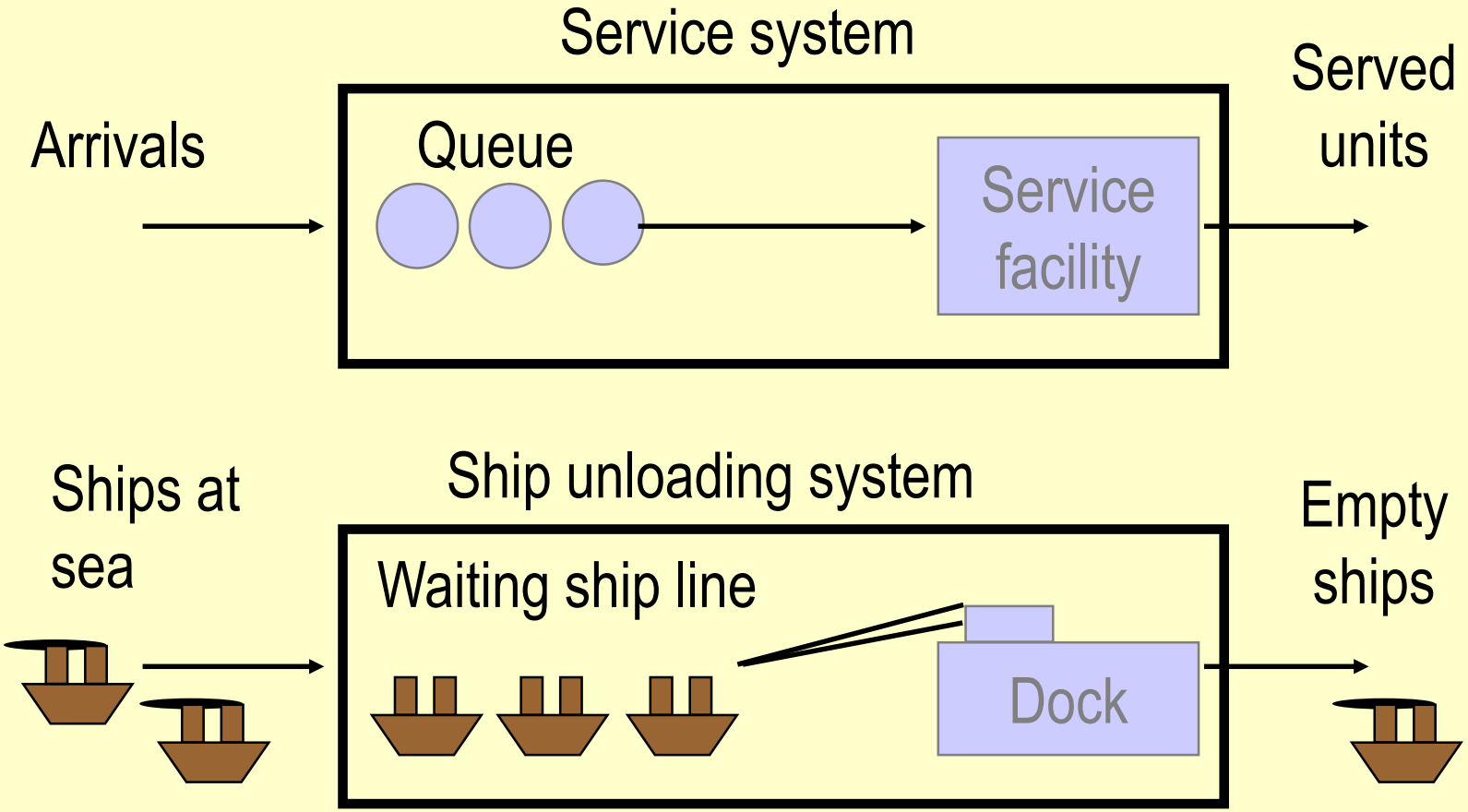
Waiting Line Characteristics

- **Line length:**
  - **Limited: Maximum number waiting is limited.**
    - **Example: Limited space for waiting.**
  - **Unlimited: No limit on number waiting.**
- **Queue discipline:**
  - **FIFO (FCFS): First in, First out. (First come, first served).**
  - **Random: Select arrival to serve at random from those waiting.**
  - **Priority: Give some arrivals priority for service.**

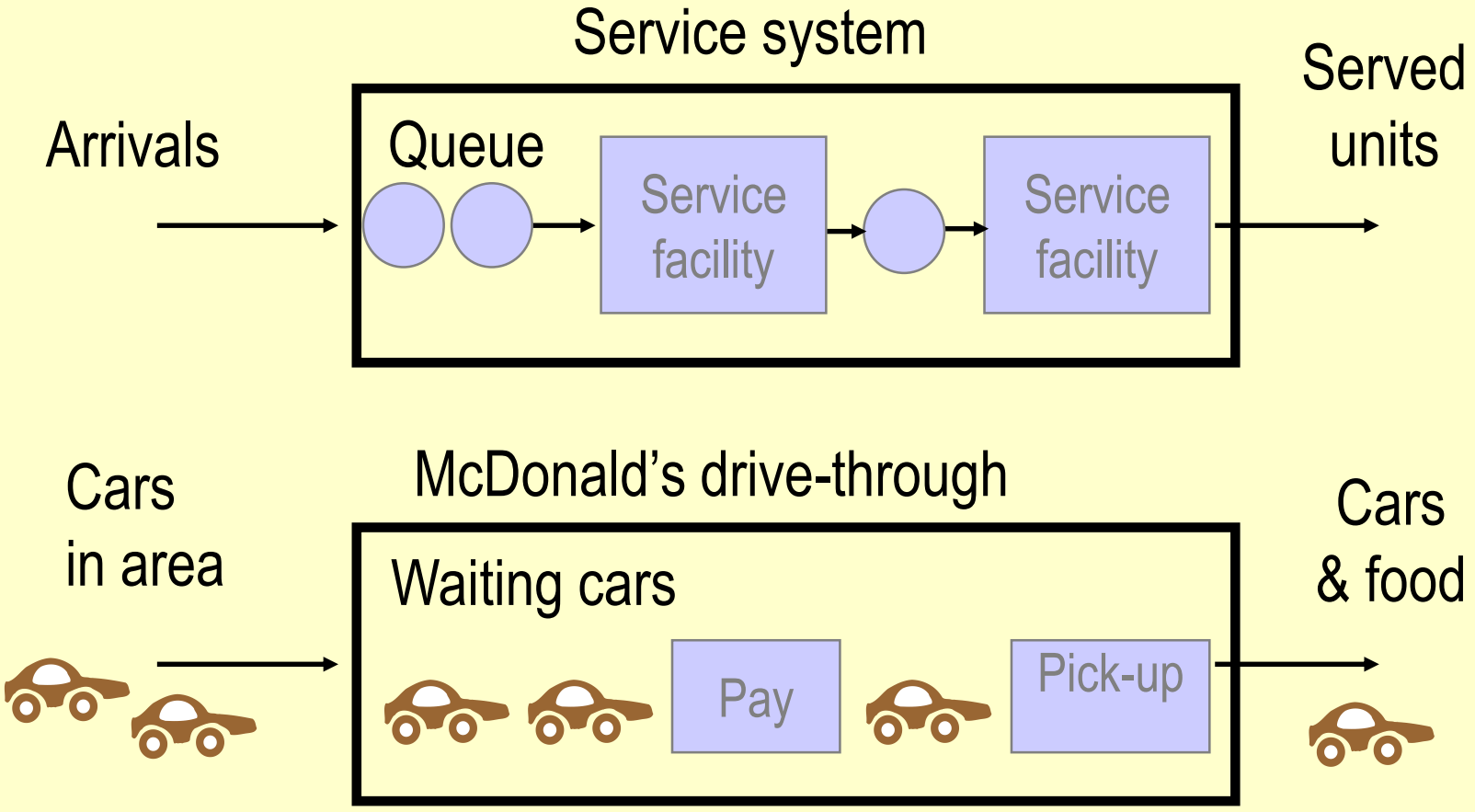
Service Configuration

- **Single channel, single phase.**
  - One server, one phase of service.
- **Single channel, multi-phase.**
  - One server, multiple phases in service.
- **Multi-channel, single phase.**
  - Multiple servers, one phase of service.
- **Multi-channel, multi-phase.**
  - Multiple servers, multiple phases of service.

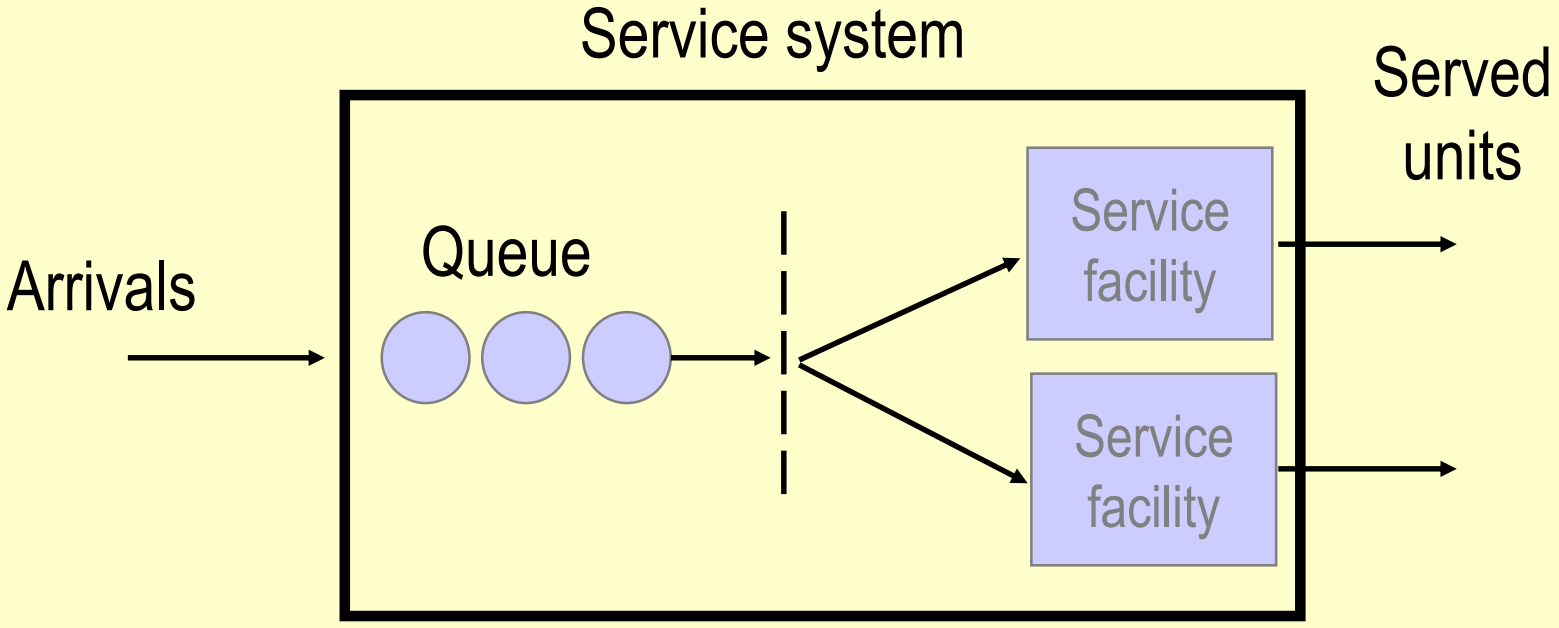
Single Channel, Single Phase



Single Channel, Multi-Phase

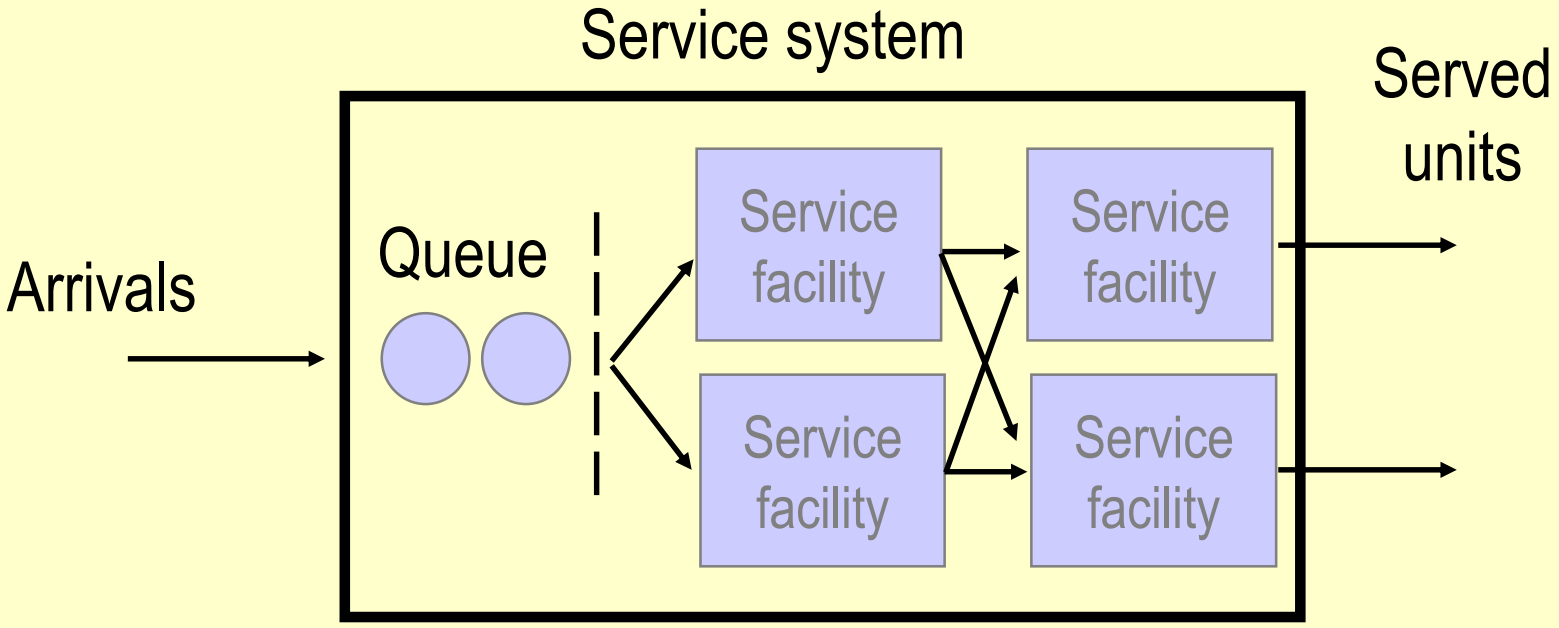


Multi-Channel, Single Phase



Example: Bank customers wait in ***single line*** for one of several tellers.

Multi-Channel, Multi-Phase



Example: At a laundromat, customers use one of several washers, then one of several dryers.

Service Times

- **Random**: Use Negative exponential probability distribution.
  - Mean service *rate* = μ
    - 6 customers/hr.
  - Mean service *time* = $1/\mu$
    - 1/6 hour = 10 minutes.

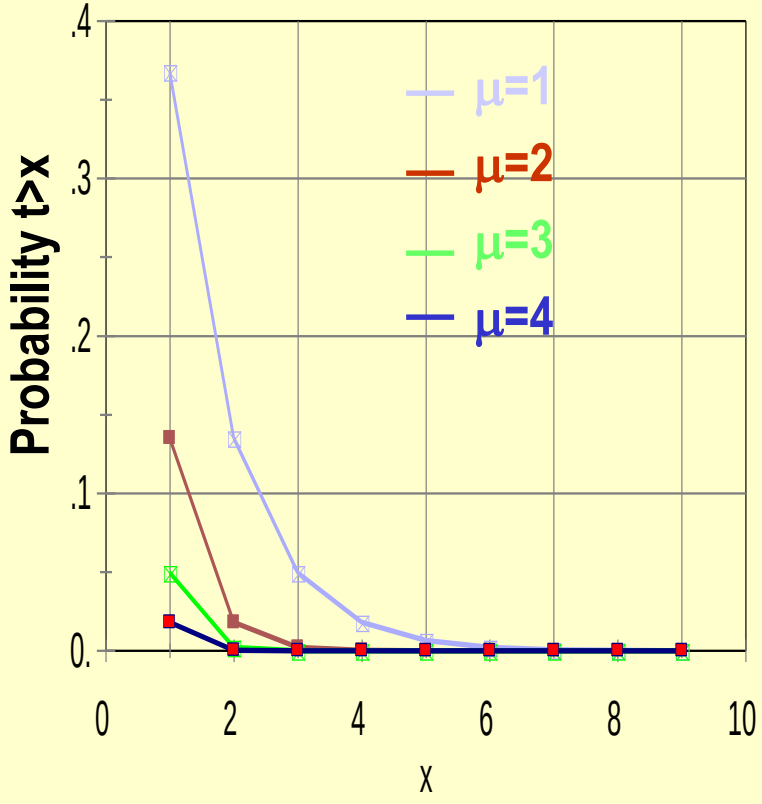
- **Non-random**: May be constant.
  - Example: Automated car wash.

Negative Exponential Distribution

- Continuous distribution.
  - Probability:

$$f(t > x) = e^{-\mu x}$$

- Example: Time between arrivals.
  - Mean service *rate* = μ
    - 6 customers/hr.
  - Mean service *time* = $1/\mu$
    - 1/6 hour = 10 minutes



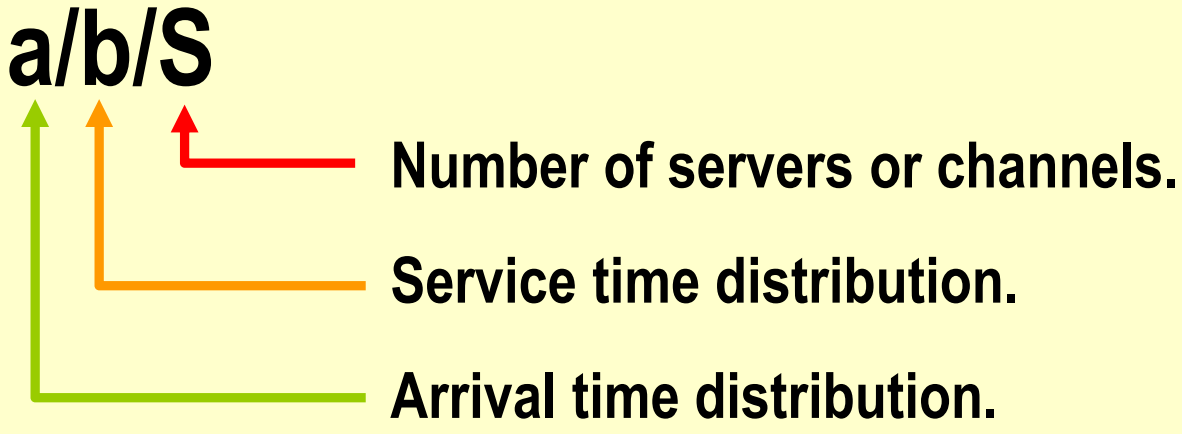
Assumptions in the Basic Model

- Customer population is *homogeneous* and *infinite*.
- Queue capacity is *infinite*.
- Customers are *well behaved* (no balking or reneging).
- Arrivals are served *FCFS (FIFO).*
- *Poisson arrivals*.
  - The time between arrivals follows a negative exponential distribution
- *Exponential service times*: Services are described by the negative exponential distribution.

Steady State Assumptions

- Mean arrival rate λ, mean service rate μ, and the number of servers are constant.
- The service rate is greater than the arrival rate.
- These conditions have existed for a long time.

Queuing Model Notation



- M = Negative exponential distribution (Poisson arrivals).
- G = General distribution.
- D = Deterministic (scheduled).

Types of Queuing Models

- Simple (M/M/1).
  - Example: Information booth at mall.
- Multi-channel (M/M/S).
  - Example: Airline ticket counter.
- Constant Service (M/D/1).
  - Example: Automated car wash.
- Limited Population.
  - Example: Department with only 7 drills that may break down and require service.

Common Questions

- Given λ, μ and S, how large is the queue (waiting line)?
- Given λ and μ, how many servers (channels) are needed to keep the average wait within certain limits?
- What is the total cost for servers and customer waiting time?
- Given λ and μ, how many servers (channels) are needed to minimize the total cost?

Performance Measures

- Average queue time = W_q
- Average queue length = L_q
- Average time in system = W_s
- Average number in system = L_s
- Probability of idle service facility = P_0
- System utilization = ρ
- Probability of more than k units in system = $P_{n > k}$
  - Also, fraction of time there are more than k units in the system.

General Queuing Equations

$$\rho = \frac{\lambda}{s\mu}$$

$$W_s = W_q + \frac{1}{\mu}$$

$$L_s = L_q + \frac{\lambda}{\mu}$$

$$L_q = \lambda W_q$$

$$L_s = \lambda W_s$$

Given one of W_s, W_q, L_s, or L_q you can use these equations to find all the others.

M/M/1 Model

- **Type:** Single server, single phase system.
- **Input source**: Infinite; no balks, no reneging.
- **Queue**: Unlimited; single line; FIFO (FCFS).
- **Arrival distribution**: Poisson.
- **Service distribution**: Negative exponential.

M/M/1 Model Equations

Average # of customers in system: $L_s = \frac{\lambda}{\mu - \lambda}$

Average time in system: $W_s = \frac{1}{\mu - \lambda}$

Average # of customers in queue: $L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$

Average time in queue: $W_q = \frac{\lambda}{\mu(\mu - \lambda)}$

System utilization $\rho = \frac{\lambda}{\mu}$

M/M/1 Probability Equations

Probability of 0 units in system, i.e., system idle:

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu}$$

Probability of more than k units in system:

$$P_{n>k} = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

This is also probability of k+1 or more units in system.

M/M/1 Example 1

Average arrival rate is 10 per hour. Average service time is 5 minutes.

λ = 10/hr and μ = 12/hr

($1/\mu$ = 5 minutes = 1/12 hour)

Q1: What is the average time between departures?

5 minutes? 6 minutes?

Q2: What is the average wait in the system?

$$W_s = \frac{1}{12/hr - 10/hr} = 0.5 \text{ hour or } 30 \text{ minutes}$$

M/M/1 Example 1

λ = 10/hr and μ = 12/hr

Q3: What is the average wait in line?

$$W_q = \frac{10}{12\,(12-10)} = \text{O.41667 hours} = 25 \text{ minutes}$$

Also note: $W_s = W_q + \frac{1}{\mu}$

so $W_q = W_s - \frac{1}{\mu} = \frac{1}{2} - \frac{1}{12} =$ O.41667 hours

M/M/1 Example 1

λ = 10/hr and μ = 12/hr

Q4: What is the average number of customers in line and in the system?

$$L_q = \frac{10^2}{12\,(12-10)} = 4.1667 \text{ customers}$$

$$L_s = \frac{10}{12-10} = 5 \text{ customers}$$

Also note: $L_q = \lambda W_q = 10 \times 0.41667 = 4.1667$

$L_s = \lambda W_s = 10 \times 0.5 = 5$

M/M/1 Example 1

λ = 10/hr and μ = 12/hr

Q5: What is the fraction of time the system is empty (server is idle)?

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu} = 1 - \frac{10}{12} = 16.67\% \text{ of the time}$$

Q6: What is the fraction of time there are more than 5 customers in the system?

$$P_{n>5} = \left(\frac{10}{12}\right)^6 = 33.5\% \text{ of the time}$$

More than 5 in the system...

Note that "more than 5 in the system" is the same as:

- "more than 4 in line"
- "5 or more in line"
- "6 or more in the system".

All are $P_{n>5}$

M/M/1 Example 1

λ = 10/hr and μ = 12/hr

Q7: How much time per day (8 hours) are there 5 or more customers in line?

$P_{n>5}$ = *0.335* so *33.5%* of time there are 6 or more in line.

0.335 x 480 min./day = *160.8* min. = ~2 hr 40 min.

Q8: What fraction of time are there 3 or fewer customers in line?

$$1 - P_{n>4} = 1 - \left(\frac{10}{12}\right)^5 = 1 - 0.402 = 0.598 \text{ or } 59.8\%$$

M/M/1 Example 2

Five copy machines break down at UM St. Louis per eight hour day on average. The average service time for repair is one hour and 15 minutes.

λ = 5/day (λ = 0.625/hour)

$1/\mu$ = 1.25 hours = 0.15625 days

μ = 1 every 1.25 hours = 6.4/day

Q1: What is the average number of "customers" in the system?

$$L_s = \frac{5/day}{6.4/day - 5/day} = 3.57 \text{ broken copiers}$$

M/M/1 Example 2

λ = 5/day (or λ = *0.625*/hour)

μ = 6.4/day (or μ = 0.8/hour)

Q2: How long is the average wait in line?

$$W_q = \frac{5}{6.4(6.4 - 5)} = 0.558 \text{ days (or } 4.46 \text{ hours)}$$

$$W_q = \frac{0.625}{0.8(0.8 - 0.625)} = 4.46 \text{ hours}$$

M/M/1 Example 2

λ = 5/day (or λ = *0.625*/hour)

μ = 6.4/day (or μ = 0.8/hour)

Q3: How much time per day (on average) are there 2 or more broken copiers waiting for the repair person?

2 or more "in line" = more than 2 in the system

$$P_{n>2} = \left(\frac{5}{6.4}\right)^3 = 0.477$$ (*47.7%* of the time)

0.477x 480 min./day = *229* min. = 3 hr 49 min.

M/M/1 Example 3

A coffee shop sees on average one arrival every two minutes in the morning. The average service time (for preparing the drink, paying, etc.) is 90 seconds. If you leave your house at 9:00 am, and it is a 10 minutes drive to the coffee shop, and then a 15 minute drive to school, what time would you expect to arrive at school (on average)?

λ = 30/hr μ = 40/hr = 1 every 90 seconds

$$W_q = \frac{30}{40(40 - 30)} = 0.075 \text{ hrs} = 4.5 \text{ minutes}$$

Arrival at school on average is at 9:31 am
(9:00 am + 10 min + 4.5 min + 1.5 min + 15 min = 9:31 am)

M/M/S Model

- **Type:** Multiple servers; single-phase.
- **Input source**: Infinite; no balks, no reneging.
- **Queue**: Unlimited; multiple lines; FIFO (FCFS).
- **Arrival distribution**: Poisson.
- **Service distribution**: Negative exponential.

M/M/S Equations

Probability of zero people or units in the system:

$$P_0 = \frac{1}{\left[\sum_{n=0}^{M-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{M!}\left(\frac{\lambda}{\mu}\right)^M \frac{M\mu}{M\mu - \lambda}}$$

Average number of people or units in the system:

$$L_s = \frac{\lambda\mu\left(\lambda/\mu\right)^M}{(M-1)!(M\mu - \lambda)} P_0 \frac{\lambda}{\mu}$$

Average time a unit spends in the system:

$$W_s = \frac{\mu\left(\lambda/\mu\right)^M}{(M-1)!(M\mu - \lambda)} P_0 + \frac{1}{\mu}$$

Note: M = number of servers in these equations

M/M/S Equations

Average number of people or units waiting for service:

$$L_q = L_s - \frac{\lambda}{\mu}$$

Average time a person or unit spends in the queue:

$$W_q = W_s - \frac{1}{\mu}$$

M/M/2 Model Equations

Average time in system: $W_s = \frac{4\mu}{4\mu^2 - \lambda^2}$

Average time in queue: $W_q = \frac{\lambda^2}{\mu(2\mu + \lambda)(2\mu - \lambda)}$

Average # of customers in queue: $L_q = \lambda W_q$

Average # of customers in system: $L_s = \lambda W_s$

Probability the system is empty: $P_0 = \frac{2\mu - \lambda}{2\mu + \lambda}$

M/M/2 Example

Average arrival rate is 10 per hour.

Average service time is 5 minutes for each of 2 servers.

λ = 10/hr, μ = 12/hr, and S=2

Q1: What is the average wait in the system?

$$W_s = \frac{4 \times 12}{4(12)^2 - (10)^2} = 0.1008 \text{ hours} = 6.05 \text{ minutes}$$

M/M/2 Example

λ = 10/hr, μ = 12/hr, and S=2

Q2: What is the average wait in line?

$$W_q = \frac{(10)^2}{12\,(2\times12 + 10)(2\times12 - 10)} = 0.0175 \text{ hrs} = 1.05 \text{ minutes}$$

Also note: $W_s = W_q + \frac{1}{\mu}$

so $W_q = W_s - \frac{1}{\mu} = 0.1008 \text{ hrs} - 0.0833 \text{ hrs} = 0.0175 \text{ hrs}$

M/M/2 Example

λ = 10/hr, μ = 12/hr, and S=2

Q3: What is the average number of customers in line and in the system?

$L_q = \lambda W_q$ = 10/hr $\times$ 0.0175 hr = 0.175 customers

$L_s = \lambda W_s$ = 10/hr $\times$ 0.1008 hr = 1.008 customers

M/M/2 Example

λ = 10/hr and μ = 12/hr

Q4: What is the fraction of time the system is empty (server are idle)?

$$P_0 = \frac{2 \times 12 - 10}{2 \times 12 + 10} = 41.2\% \text{ of the time}$$

M/M/1, M/M/2 and M/M/3

	1 server	2 servers	3 servers
W_q	25 min.	1.05 min.	0.1333 min. (8 sec.)
	= 0.417 hr	= 0.0175 hr	= 0.00222 hr
W_S	30 min.	6.05 min.	5.1333 min.
L_q	4.167 cust.	0.175 cust.	0.0222 cust.
L_S	5 cust.	1.01 cust.	0.855 cust.
P_0	16.7%	41.2%	43.2%

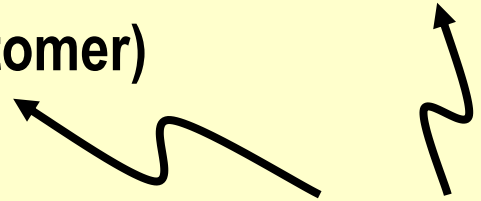
Waiting Line Costs

Service cost per day

= (# of servers) x (cost per day of each server) +
(# customers per day) x (marginal cost per customer)

Customer waiting cost per day

= (# of customers per day) x (average wait per customer)
x (time value for customer)



Time units must agree

Service Cost per Day

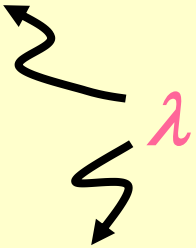
λ = 10/hr and μ = 12/hr

Suppose servers are paid $7/hr and work 8 hours/day. Also, suppose the marginal cost to serve each customer is $0.50.

M/M/1 Service cost per day

= *$7*/hr x *8* hr/day + *$0.5*/cust *x 10* cust/hr *x 8* hr/day

= *$56 + $40 = $96*/day



M/M/2 Service cost per day

= 2 *x* $7/hr x 8 hr/day + *$0.5*/cust x *10* cust/hr *x 8* hr/day

= *$112 + $40 = $152*/day

Customer Waiting Cost per Day

λ = 10/hr and μ = 12/hr

Suppose customer waiting cost is $10/hr.

M/M/1 Waiting cost per day

= *0.417* hr/cust *x 10* cust/hr x *8* hr/day x $10/hr = *$333.33*/day

M/M/1 total cost per day= $96 + $333.33 = $429.33/day

M/M/2 Waiting cost per day

= *0.0175* hr/cust *x 10* cust/hr x *8* hr/day x $10/hr = *$14*/day

M/M/2 total cost per day= $152 + $14 = $166/day

Unknown Waiting Cost

Suppose customer waiting cost is not known = *C*.

M/M/1 Waiting cost per day

= *0.417* hr/cust *x 10* cust/hr *x 8* hr/day *x C* = *33.33C* $/day

M/M/1 total cost per day = 96 + 33.33*C*

M/M/2 Waiting cost per day

= *0.0175* hr/cust *x 10* cust/hr *x 8* hr/day *x C* =*1.4C* $/day

M/M/2 total cost per day= 152 + 1.4*C*

M/M/2 is preferred when *152 + 1.4C* < *96 + 33.33C* or *C > $1.754/hr*

M/M/2 and M/M/3

Q: How large must customer waiting cost be for M/M/3 to be preferred over M/M/2?

M/M/2 total cost = 152 + 1.4*C*

M/M/3 Waiting cost per day

= *Cx 0.00222* hr/cust *x 10* cust/hr *x 8* hr/day = *0.1776C* $/day

M/M/3 total cost = 208 + 0.1776*C*

M/M/3 is preferred over M/M/2 when

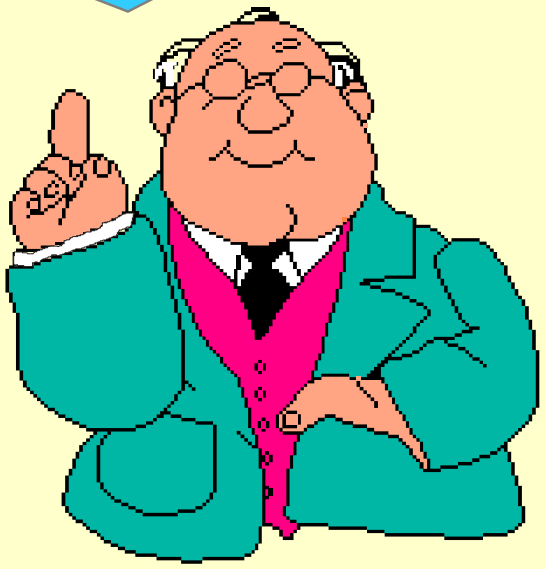
$$208 + 0.1776C < 152 + 1.4C$$

$$C > \$45.81/hr$$

Remember: λ & μ Are Rates

- λ = Mean number of arrivals per time period.
  - Example: 3 units/hour.
- μ = Mean number of arrivals served per time period.
  - Example: 4 units/hour.
    - $1/\mu$ = 15 minutes/unit.

If average service time is 15 minutes, then μ is 4 customers/hour



Other Queuing Models

- M/D/S
  - Constant service time; Every service time is the same.
  - Random (Poisson) arrivals.
- Limited population.
  - Probability of arrival depends on number in service.
- Limited queue length.
  - Limited space for waiting.
- Many others...

Other Considerations

- Wait time & queue length increase rapidly for $\rho > 0.7$
  - Queue is small until system is about 70% busy; then queue grows very quickly.
- Pooling servers is usually advantageous.
  - Airport check-in vs. Grocery stores.
- Variance in arrivals & service times causes long waits.
  - Long service times cause big waits.
- Cost of waiting is nonlinear.
  - Twice as long wait may be more than twice as bad.

More Considerations

- Reduce effect of waiting.
  - Distract customers with something to do, look at or listen to.
    - Music, art, mirrors, etc.
  - Provide feedback on expected length of wait.
    - "Your call will be answered in 6 minutes…"
- Use self service can reduce load on servers.
  - Self-service at grocery stores.