BUAN/OPRE 6398
Prescriptive Analytics

Queuing Theory

Rasoul Ramezani

School of Management
The University of Texas at Dallas

Lecture Outline

- Objective of Queuing Model
- Characteristics of Queuing Systems:
  - Arrival Rate
  - Interarrivals Time
  - Service Rate
  - Service Time
- Kendall Notation
  - M/M/s model
  - M/M/s model with finite queue length
  - M/M/s model with finite arrival population
  - M/G/1 model
  - M/D/1 model

Purpose of Queuing Models

- The focus of a queuing model is on determining various characteristics of the system of interest, such as:
  - the average waiting time in line,
  - the average length of the waiting line,
  - the average service time,
  - the average number of units in the system (including waited in line and being served),
  - the probability of waiting for more than a certain amount of time,
  - and so on.
- These are then used for subsequent cost/profit analysis.

Objective of Queuing Model

- Achieving a balance between two conflicting costs so that the combined total cost (TC) is minimized:
  1. $C_1(Q)$ = Cost of providing service (increases with the service level)
  2. $C_2(Q)$ = Cost of customers waiting time -- i.e., cost of customers dissatisfaction (decreases with service level)

- **Total Cost:**

$$TC = C_1(Q) + C_2(Q)$$

- **Optimal Service Level:**

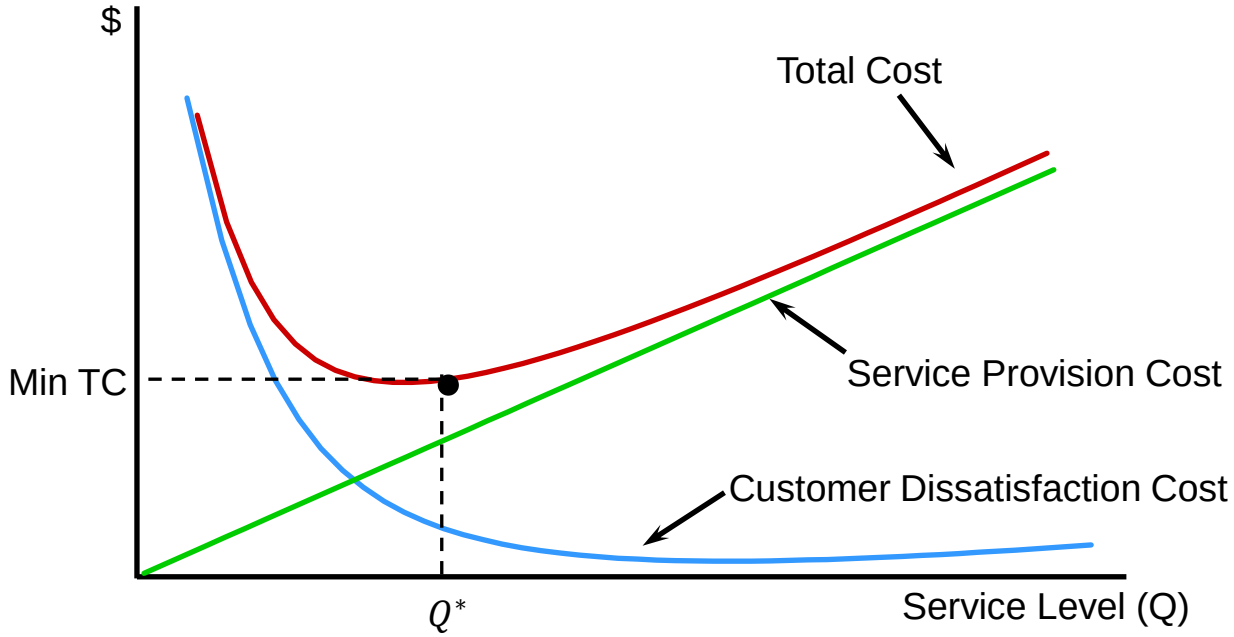
FOC:

$$\frac{dTC}{dQ} = \frac{dC_1}{dQ} + \frac{dC_2}{dQ} = 0$$

$$\frac{dC_1}{dQ} = -\frac{dC_2}{dQ} \Rightarrow Q^*$$

SOC:

$$\frac{d^2TC}{dQ^2}\Big|_{Q^*} > 0$$



Characteristics of Queuing Systems: Arrival Rate

- **Arrival rate (λ):** The average number of arrivals per time period. E.g.,
  - In a carwash, cars arrive at the rate of 12 cars per hour.
  - In a technical support department, calls arrive at a rate of 5 calls per hour.
- Arrivals are often described by a *Poisson* random variable.
  - Let X = the number of arrivals per time period
  - The probability of having x arrivals per time period

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \text{ for } x = 0, 1, 2, \ldots$$

    - $e = 2.7183$ is the base of natural logarithm
    - Note: $E(X) = V(X) = \lambda$
- In Excel: $P(X = x) = POISSON.DIST(x, \lambda, FALSE)$
- Cumulative Probability: $P(X \leq x) = POISSON.DIST(x, \lambda, TRUE)$

Interarrivals Time

- The average time period between two consecutive arrivals.
- If arrivals (X) follow a Poisson distribution with mean λ per time period,
  - Interarrivals time follows an *exponential* distribution with the mean of $1/\lambda$ time periods.
- E.g., If calls arrive according to a Poisson distribution with a mean of 5 calls per hour.
  - Then, interarrivals time follows an exponential distribution with a mean of 0.2 hours.
  - I.e., on average, calls arrive every 0.2 hours (or every 12 minutes).

Characteristics of Queuing Systems: Service Rate and Service Time

- **Service rate (μ):** The average number of customers served per time period.
  - E.g., on average, ten cars are washed at a carwash per hour ($\mu = 10$).
- **Service time:** The time spent until a customer is fully serviced (not including time in the queue).
- **Note:** The *average service time* per customer $= 1/\mu$ time period.
  - E.g., If the service rate = 10 per hour. Then, on average, each customer is served for 0.10 hours (or 6 minutes).

Probability Distribution of Service Time

- Service times are often described by an *exponential* random variable.
- Let $\mu =$ service rate per time period
- The probability density function for service time (T) is:

$$f(t) = \mu e^{-\mu t} \text{ for } t > 0$$

  - Note: $E(T) = SD = 1/\mu$
- With the service rate of μ, the probability that the service time (T) is between t_1 and t_2 is:

$$P(t_1 < T < t_2) = \int_{t_1}^{t_2} \mu e^{-\mu t} dt = e^{-\mu t_1} - e^{-\mu t_2}$$

- In Excel:

$$P(t_1 < T < t_2) = EXPON.DIST(t_2, \mu, 1) - EXPON.DIST(t_1, \mu, 1)$$

Kendall Notation

- Queuing models are described by three characteristics in the following general format:

1/2/3

  - **Characteristic 1**
    - M = Markovian (random) interarrival times (following an exponential distribution)
    - D = Deterministic interarrival times (not random)
  - **Characteristic 2**
    - M = Markovian (random) service times (following an exponential distribution)
    - G = General service times (following a nonexponential distribution)
    - D = Deterministic service times (not random)
  - **Characteristic 3**
    - A number indicating the number of servers in the system

- Examples: M/M/2, M/G/3, M/D/1

Operating Characteristics

- Typical operating characteristics of interest:
  - U: Utilization factor, percentage of time that all servers are busy
  - P_0: Prob. that there are no units in the system
  - P_n: Prob. of n units in the system
  - P_w: Prob. that an arriving unit has to wait for service
  - L_q: Avg. number of units in the line (queue length)
  - L: Avg. number of units in the system (in line & being served)
  - W_q: Avg. time a unit spends in the line
  - W: Avg. time a unit spends in the system

Models Analyzed in this Session

- We analyze four types of queuing models:
  - M/M/s model
  - M/M/s model with finite queue length
  - M/M/s model with finite arrival population
  - M/G/1 model
  - M/D/1 model
- Formulas for the operating characteristics of these queuing models have been derived analytically.
- An Excel template called *Q.xls* implements the formulas for several common types of models.
  - Q.xlsx was created by David Ashley of the University of Missouri at Kansas City.

The M/M/s Model

Assumptions:

- There are s (a positive integer) servers.
- Arrivals follow a Poisson distribution and occur at an average rate of λ per time period.
  - Hence, interarrival times follow an exponential distribution with a mean of $1/\lambda$.
- Each server provides service at an average rate of μ per time period, and actual service times follow an exponential distribution.
- Arrivals wait in a single FIFO queue and are serviced by the first available server.
- Total service capacity must be strictly greater than the arrival rate ($\lambda < s\mu$).
  - Otherwise, the queue becomes infinitely long (even if $\lambda = s\mu$ (why?)), and some customers would not get the service at the end.

Formulas describing the operating characteristics of an M/M/s queue

$$U = \lambda/(s\mu)$$

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

$$L_q = \frac{P_0(\lambda/\mu)^{s+1}}{(s-1)!(s-\lambda/\mu)^2}$$

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

$$W_q = L_q/\lambda$$

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

$$P_w = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s\left(\frac{s\mu}{s\mu-\lambda}\right)P_0$$

$$P_n = \begin{cases}\dfrac{(\lambda/\mu)^n}{n!}P_0, \text{ for } n \le s \\ \dfrac{(\lambda/\mu)^n}{s!s^{(n-s)}}P_0, \text{ for } n > s\end{cases}$$

An M/M/s Example

- The customer support hotline for Bitway Computers is currently staffed by a single technician.
- Calls arrive randomly at a rate of 5 per hour and follow a Poisson distribution.
  - So, the interarrival time follows an exponential distribution with a mean of 0.2 hours.
- The technician services the calls at a mean rate of 7 per hour.
  - So, the service time follows an exponential distribution with a mean of 1/7 hours.
- Bitway's president has received numerous complaints from customers about the length of time they must wait "on hold" for service when calling the hotline.
- He wants to determine:
  1. The average waiting time (W_q)
  2. If $W_q > 5$ minutes, the number of technicians required to have $W_q \leq 2$.

See file Q.xlsx (MMS)

Summary of Results
Bitway Computers

Arrival rate (per hour)	5	5
Service rate (per hour)	7	7
Number of servers	1	2
Utilization (% time all servers are busy)	71.43%	35.71%
P_0, probability that the system is empty	0.2857	0.4737
L_q, expected queue length	1.7857	0.1044
L, expected number in system	2.5	0.8187
W_q, expected time in queue (hours)	0.3571	0.0209
W, expected total time in system (hours)	0.5	0.1637
P_w, probability that a customer waits	0.7143	0.188

M/M/s Model with Finite Queue Length

- In the above example,
  - It is assumed that ALL arrivals can join the queue and wait for the service (i.e., the capacity of the waiting area is infinite).
- Sometimes, the capacity of the waiting area is limited.
  - I.e., the queue length is finite.
  - E.g., no more than 10 calls could be put on hold.
- Formulas describing the operating characteristics of an M/M/s model with a finite queue length of $K \Rightarrow$

$$U = (L - L_q)/s$$

$$P_0 = \left(1 + \sum_{n=1}^{s} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s+1}^{K} \left(\frac{\lambda}{s\mu}\right)^{n-s}\right)^{-1}$$

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0, \text{ for } n = 1,2,\ldots,s$$

$$P_n = \frac{(\lambda/\mu)^n}{s!s^{n-s}} P_0, \text{ for } n = s+1, s+2, \ldots, K+s$$

$$P_n = 0, \text{ for } n > K+s$$

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}(1 - \rho^{K-s} - (K-s)\rho^{K-s}(1-\rho)), \text{ where } \rho = \lambda/(s\mu)$$

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right)$$

$$W_q = \frac{L_q}{\lambda(1 - P_K)}$$

$$W = \frac{L}{\lambda(1 - P_K)}$$

M/M/s Model with Finite Queue Length Example

- Suppose Bitway's telephone system can keep a maximum of 5 calls on hold at any time.
  - Thus, if a new call is made to the hotline when five calls are already in the queue, the new call receives a busy signal.
- One way to reduce the number of calls facing busy signals is …
  - increasing the length of the queue (i.e., increasing the number of calls that can be put on hold).
- Bitway's president wants to investigate the effect of adding a second technician on…
  1. The number of calls receiving busy signals
  2. The average waiting time before receiving service

See file Q.xlsx (finite Q length)

Summary of Results: Bitway Computers with Finite Queue

Arrival rate (per hour)	5	5
Service rate (per hour)	7	7
Number of servers	1	2
Maximum queue length	5	5
Utilization (% time all servers are busy)	68.43%	35.69%
P_0, probability that the system is empty	0.3157	0.4739
L_q, expected queue length	1.082	0.1019
L, expected number in system	1.7664	0.8157
W_q, expected time in queue (hours)	0.2259	0.0204
W, expected total time in system (hours)	0.3687	0.1633
P_w, probability that a customer waits	0.6843	0.1877
Probability that a customer is rejected	0.0419	0.0007

M/M/s Model with Finite Population

- In the above examples, it is assumed that the population of potential customers is extremely large/ infinite.
- Sometimes, the possible number of arriving customers is finite.
  - So, the mean arrival rate (λ) depends on the average queue length (L_q).

- **Assumptions:**
  - There are s servers.
  - There are N potential customers in the arrival population.
  - The arrival pattern of *each customer* follows a Poisson distribution with a mean arrival rate of λ per time period.
  - Each server provides service at an average rate of μ per time period, and service times follow an exponential distribution.
  - Arrivals wait in a single FIFO queue and are serviced by the first available server.

M/M/s model with a finite arrival population of size N

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

$$P_n = \frac{N!}{(N-n)!n!}\left(\frac{\lambda}{\mu}\right)^n P_0, \text{ if } 0 \le n \le s$$

$$P_n = \frac{N!}{(N-n)!s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n P_0, \text{ if } s < n \le N$$

$$P_n = 0, \text{ if } n > N$$

$$L_q = \sum_{n=s}^{N} (n-s)P_n$$

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right)$$

$$W_q = \frac{L_q}{\lambda(N-L)}$$

$$W = \frac{L}{\lambda(N-L)}$$

M/M/s with Finite Population
Example

- Miller Manufacturing owns 10 identical machines that produce colored nylon thread.
- Machine breakdowns follow a Poisson distribution with an average of 0.01 breakdowns per machine per hour.
- The company loses $100 each hour a machine is down.
- One technician is employed (at a rate of $20/hour) to fix broken machines.
- Service times to repair the machines are exponentially distributed, with an average of 8 hours per repair (i.e., the service rate = 1/8 machines per hour).
- Management wants to …
  1. Analyze the impact of adding another service technician on …
     i. The average time needed to fix a machine.
     ii. The total cost associated with machine breakdowns.
  2. Determine the optimal number of servers.

See file Q.xlsx (finite population)

Summary of Results: Miller Manufacturing

Optimal

Arrival rate	0.01	0.01	0.01
Service rate	0.125	0.125	0.125
Number of servers	1	2	3
Population size	10	10	10
Utilization (% time all servers are busy)	67.80%	36.76%	24.67%
P_0, probability that the system is empty	0.322	0.4517	0.4623
L_q, expected queue length	0.8463	0.0761	0.0074
L, expected number in system	1.5244	0.8112	0.7476
W_q, expected time in queue (hours)	9.9856	0.8282	0.0799
W, expected total time in system (hours)	17.986	8.8282	8.0799
P_w, probability that a customer waits	0.678	0.1869	0.0347
Hourly cost of service technicians	$20.00	$40.00	$60.00
Hourly cost of inoperable machines	$152.44	$81.12	$74.76
Total hourly costs	$172.44	**$121.12**	$134.76

M/G/1 Model

- Not all service times can be modeled accurately using an exponential distribution (where the service time can be any number > 0).
- E.g.,
  - Changing the oil in a car may require at least 15 minutes.
  - Getting a haircut may require at least 20 minutes.

- **M/G/1 Model Assumptions:**
  - Arrivals follow a Poisson distribution with a mean of λ.
  - Service times follow *any* random distribution with a mean of μ and a standard deviation of σ.
  - There is a single server.

M/G/1 Queue

$$P_0 = 1 - \lambda/\mu$$

$$L_q = \frac{\lambda^2\sigma^2 + (\lambda/\mu)^2}{2(1 - \lambda/\mu)}$$

$$L = L_q + \lambda/\mu$$

$$W_q = L_q/\lambda$$

$$W = W_q + 1/\mu$$

$$P_w = \lambda/\mu$$

M/G/1 Model
Example

- Zippy-Lube is a drive-through automotive oil change business, which operates 10 hours a day, 6 days a week, with a marginal profit of $15 per oil change.
- Cars arrive at the oil change center following a Poisson distribution with an average of 3.5 cars per hour.
- The average service time per car is 15 minutes with a standard deviation of 2 minutes.
  - Note: No assumption about the service time probability distribution is made.
- The manufacturer's representative claims that a new automated oil dispensing device (which costs $5000) will reduce the average service time by 3 minutes per car.
- The owner wants to analyze the impact of the new automated device on his business and determine the payback period for this device.

See file Q.xlsx (MG1)

Summary of Results
Zippy Lube

	With Automated Machine?		
	NO	YES	YES
Arrival rate	**3.5**	**3.5**	**4.3702**
Average service TIME	**0.25**	**0.20**	**0.20**
Standard dev. of service time	0.0333		0.333
Utilization (% time all servers are busy)	87.50%	70.00%	87.40%
P_0, probability that the system is empty	0.1250	0.3000	0.1260
L_q, expected queue length	**3.1169**	0.8394	**3.1168**
L, expected number in system	3.9919	1.5394	3.9908
W_q, expected time in queue (hours)	0.8906	0.2398	0.7132
W, expected total time in system (hours)	1.1406	0.4398	0.9132

- The arrival rate could increase by 0.87 per hour with the dispensing machine before the expected queue length exceeds its level without the dispensing machine.

Financial Impact

Change in …	
Arrival per hour	0.87
Profit per hour	$13.05
Profit per day	$130.53
Profit per week	$783.19
Cost of Machine	$5,000
Payback period (weeks)	6.38

M/D/1 Model

- Service times may not be random in some queuing systems.
- Examples
  - In manufacturing, the time to machine an item might be exactly 10 seconds per piece.
  - An automatic car wash might spend the same amount of time on each car it services.
- The M/D/1 model can be used when service times are deterministic (not random).
- The results for an M/D/1 model can be obtained using the M/G/1 model by setting the standard deviation of the service time to 0 ($\sigma = 0$).

End of Lecture 11