The M/M/c Queueing Model

CB2201 – Operations Management

Lecture 7

Lecture Overview

(1) How to deliver superior customer service

Poka-yoke – How to reduce mistakes via mistake-proofing

Ritz-Carlton – How to create a customer service culture

(2) How to manage customer wait time

Queueing model – How to decide the optimal service capacity

Psychology of waiting – How to make waiting more pleasant

Companies need to make staffing decisions

- Suppose that you are the manager of a Starbucks store
- How many employees to schedule?
  - Too few staff ⇒ long wait times for customers
  - Too many staff ⇒ high salary expenses for the company
- Need to strike the right balance between waiting time and salary costs

Let’s apply the scientific method to study queueing systems

- Queueing theory is the mathematical study of waiting lines or queues
- We will use a [M/M/c queue](#) to model a queueing system
- In case you were wondering…
  - The 1st M refers to memoryless interarrival times
  - The 2nd M refers to memoryless service times
  - The c refers to the number of servers

A queueing model of a service system

A customer **arrives**

If the server is busy, then
the **customer waits in the queue**
for his turn

The server **starts to serve**
the customer

Service is completed
and the customer departs

The entities in a queueing system

- Terminology:
    - A **customer** arrives to the system
    - Needs to receive service from the **server**
    - May need to wait for service in the **queue**
- Examples:

System	Customer	Server
Bank	Customer	ATM
CityU Health Clinic	Patient	General Practitioner

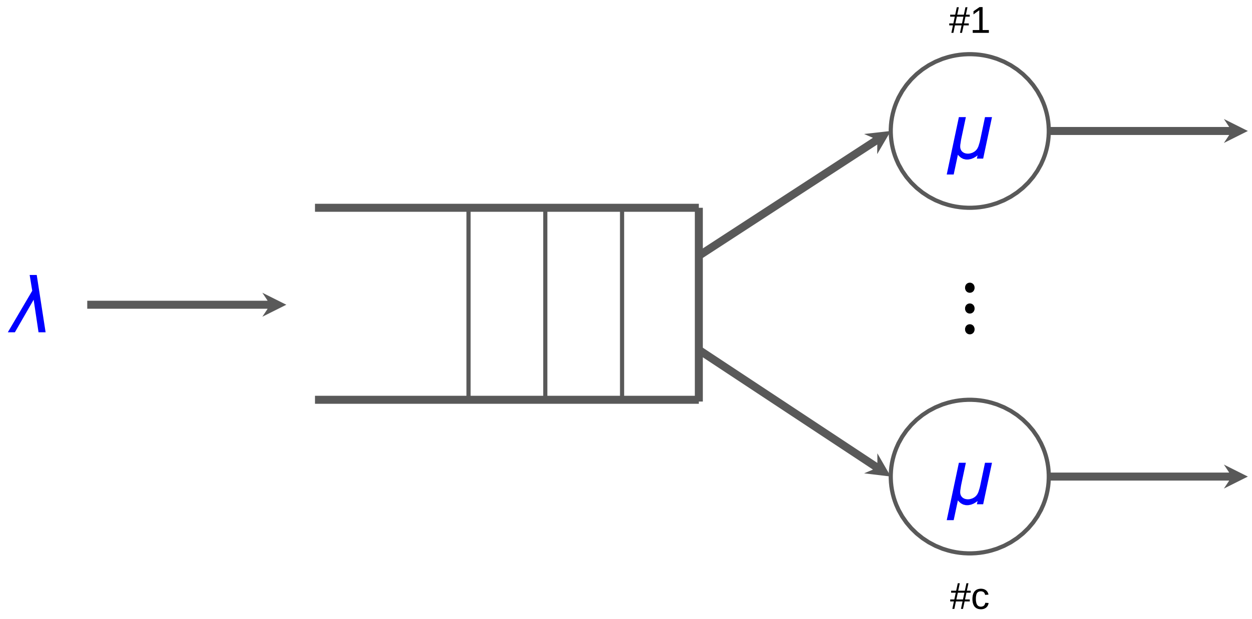
M/M/c queue – input parameters

- The arrival rate (denoted by **λ** which is pronounced “lambda”)
  - Defined as the number of customers arriving per unit time
  - Example: 4 customers per minute
- The service rate (denoted by **μ** which is pronounced “mew”)
  - Defined as the number of customers that a **single server** can serve per unit time
  - Example: 10 minutes per customer ⇒ 6 customers per hour
- The number of servers (denoted by **c**)
  - It is assumed that that one server can serve only one customer at a time

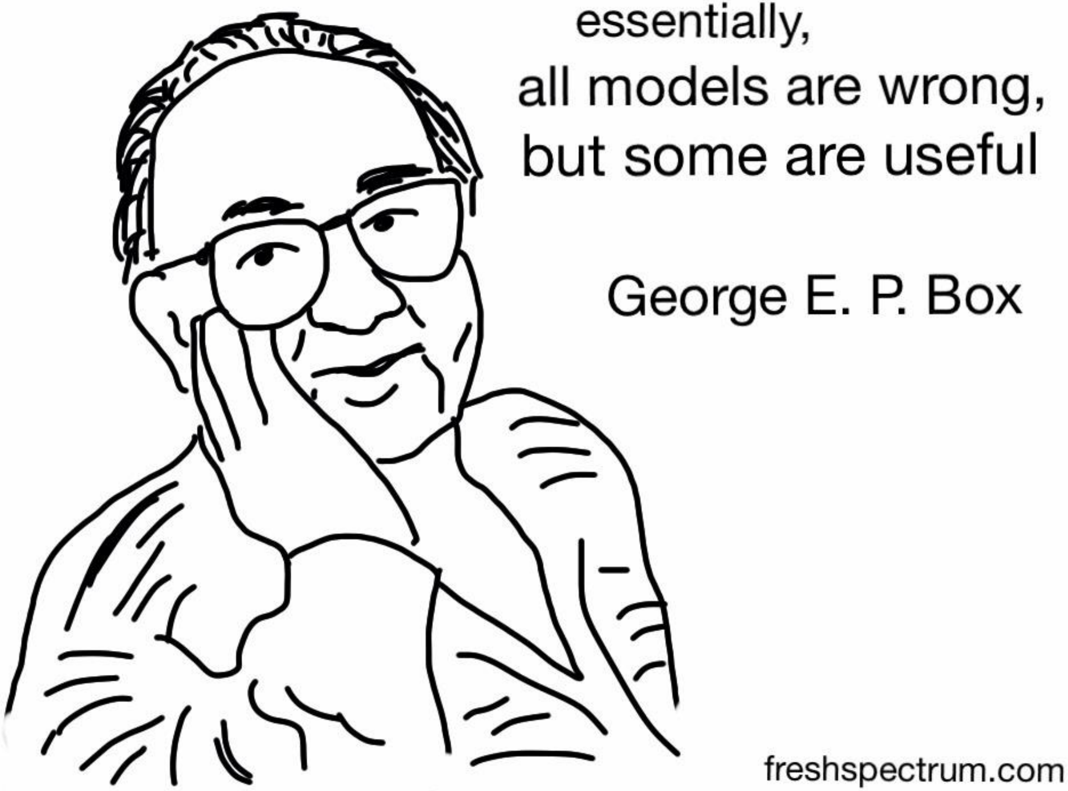
It’s all Greek to me

- The letters **λ** and **μ** may appear strange
- This is because they are letters from the Greek alphabet
- Greek letters are often used in mathematics

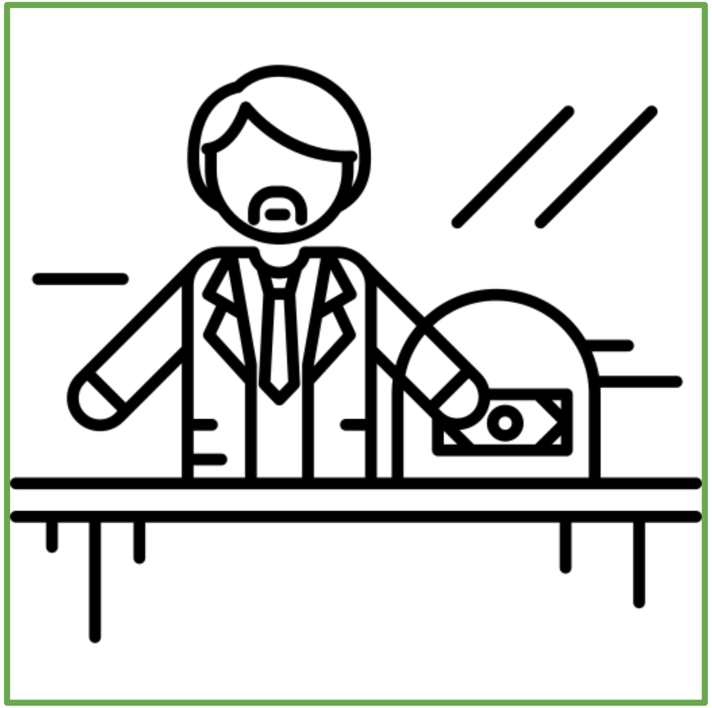
The M/M/c queueing model



The M/M/c model is not 100% correct, but it is still useful



The M/M/c model is not 100% correct, but it is still useful



Bank Teller

One queue for everyone

Supermarket Cashier

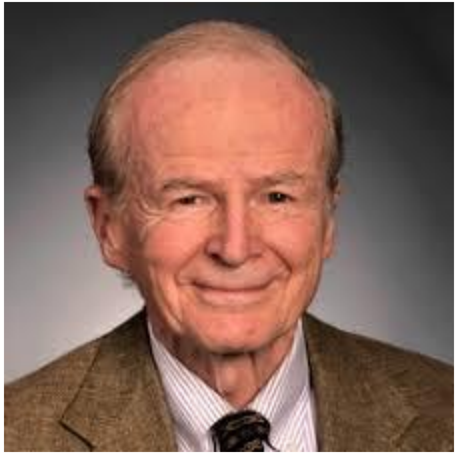
One queue per cashier

M/M/c queue – output metrics

- The average queue length (denoted by **L**)
- The average waiting time (denoted by **W**)
- The average total time in process

  = average wait time + average service time

A useful math result called “Little’s Law”



- Named after John Little (a prof from MIT)
- Little’s Law states that $L = \lambda W$
  - L = average queue length
  - λ = average arrival rate
  - W = average waiting time
- Example:
  - CB admits $\lambda = 800$ students per year
  - Average graduation time $W = 4$ years
  - Average number of students $L = 4 \times 800 = 3200$ students

Will we wait? Compare arrival rate with system service rate

- Case 1: $\lambda > c\mu$
  - Arrival rate is greater than the system service rate
  - The queue is unstable and will grow to infinity… and beyond

- Case 2: $\lambda < c\mu$
  - Depends on whether there is variability in the interarrival and service times

- Case 3: $\lambda = c\mu$
  - This situation is weird so just forget about it
  - (Very unlikely to encounter this situation in real life)

Will we wait? Is there variability?

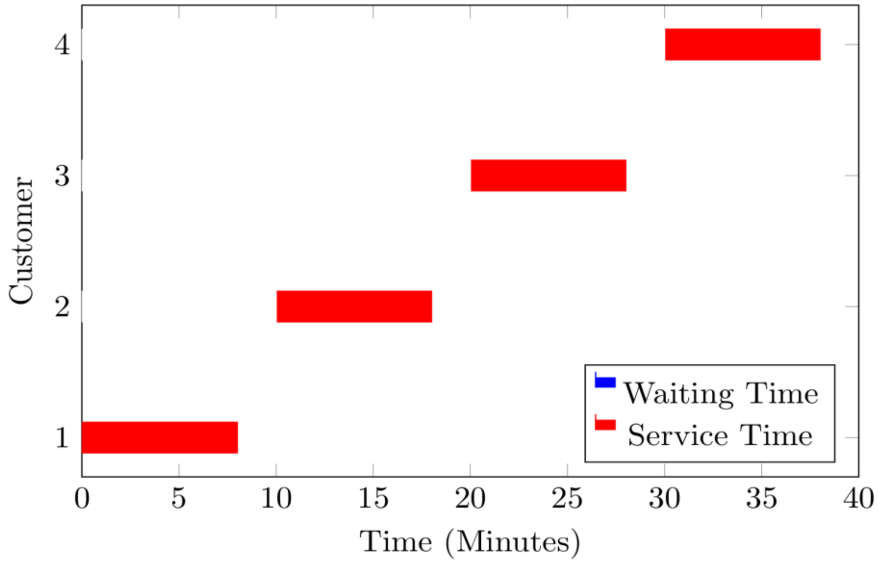
- Case 2A: $\lambda < c\mu$ **without variability** in interarrival and service times
  - Zero queue and zero waiting time

- Case 2B: $\lambda < c\mu$ **with variability** in interarrival and service times
  - Some customers may have to wait

- In general, $\uparrow$ variability leads to $\uparrow$ waiting

An illustration of Case 2A: without variability

- A customer arrives exactly every 10 minutes
- It takes exactly 8 minutes to serve every customer
- No variability ⇒ No waiting! :-)



An illustration of Case 2B: with variability

- A customer arrives, on average, every 10 minutes
- It takes, on average, 8 minutes to serve every customer
- With variability ⇒ Some customers may have to wait :-(

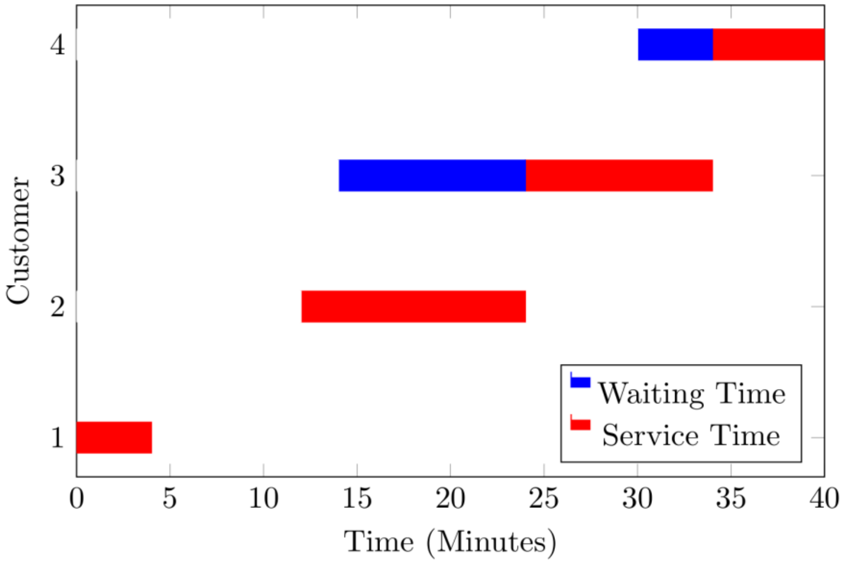


Table: Average Queue Length for M/M/c Queueing Model

- Note: We will give you a copy of the table during the quiz / exam
- [Link to Table on Google Drive]()

Average Queue Length for M/M/c Queuing Model

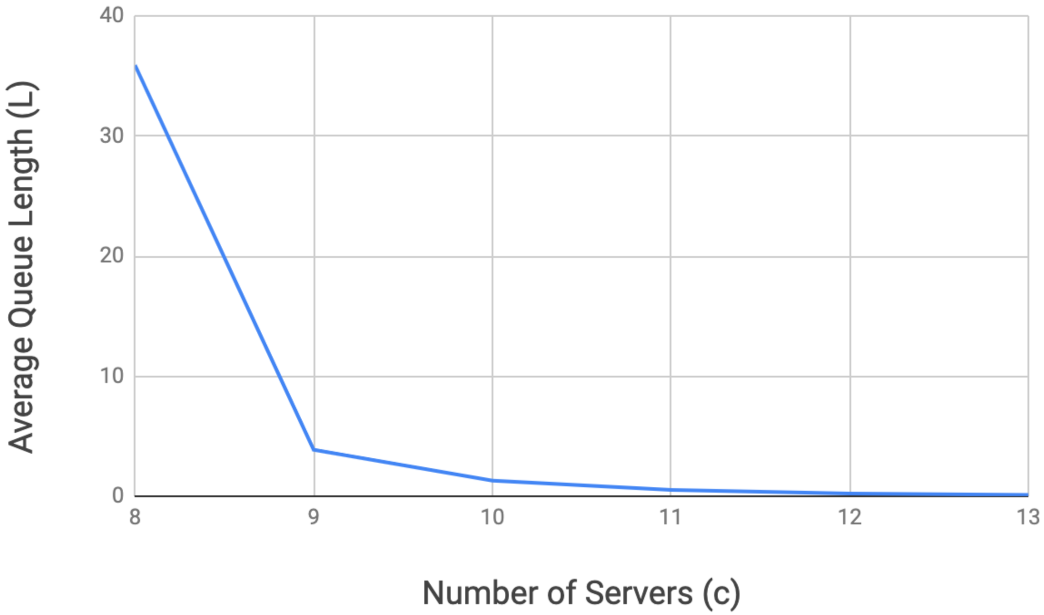
Arrival Rate / Service Rate	Number of Servers: 1	2	3	4	5	6	7	8
0.20	0.0500	0.0020						
0.40	0.2666	0.0166						
0.60	0.9090	0.0593	0.0061					
0.80	3.2000	0.1523	0.0189					
1.00		0.3333	0.0454	0.0067				
1.20		0.6748	0.0940	0.0158				
1.40		1.3449	0.1178	0.0324	0.0059			
1.60		2.8441	0.3128	0.0604	0.0121			
1.80		7.6731	0.5320	0.1051	0.0227	0.0047		
2.00			0.8888	0.1730	0.0390	0.0090		
2.20			1.4907	0.2770	0.0660	0.0158		
2.40			2.1261	0.4205	0.1047	0.0266	0.0065	
2.60			4.9322	0.6581	0.1609	0.0425	0.0110	
2.80			12.2724	1.0000	0.2411	0.0659	0.0180	
3.00				1.5282	0.3541	0.0991	0.0282	0.0077

Legend
Queue length is almost zero
Queue length is significant
Queue length is infinite

The effect of service capacity on queue length is nonlinear

- From 8 to 9 servers
  - Service capacity increases by 12.5%
  - Wait time decreased by 89.3%!

Small change in service capacity
→ Big change in wait time



Three ways to reduce queue length & wait time

More servers

	Number of Servers							
Arrival Rate / Service Rate	1	2	3	4	5	6	7	8
0.20	0.0500	0.0020						
0.40	0.2666	0.0166						
0.60	0.9090	0.0593	0.0061					
0.80	3.2000	0.1523	0.0189					
1.00		0.3333	0.0454	0.0067				
1.20		0.6748	0.0940	0.0158				
1.40		1.3449	0.1178	0.0324	0.0059			
1.60		2.8441	0.3128	0.0604	0.0121			
1.80		7.6731	0.5320	0.1051	0.0227			
2.00			0.8888	0.1730	0.0390			
2.20			1.4907	0.2770	0.0660			
2.40			2.1261	0.4205	0.1047			
2.60			4.9322	0.6581	0.1609			
2.80			12.2724	1.0000	0.2411			
3.00				1.5282	0.3541			

Decreasing variability in interarrival or service times

Reduce service time ⇔ Increase service rate

M/M/c queue – Steps to calculate the output metrics

- Step 1: Calculate λ / μ
- Step 2: Read average queue length L from the table
  - It is the entry in the **λ / μ row** and **c column** in the `Average Queue Length` table
- Step 3: Calculate average wait time W using Little's Law

$$W = L / \lambda$$

M/M/c example – Taste supermarket

- Use an M/M/c queueing model for the Taste supermarket
- Input parameters:
  - Arrival rate of customers is 780 customers/hour
  - A cashier requires 36 seconds on average to serve a customer
  - Taste employs 8 cashiers
- Goal – Calculate the key performance metrics of the system:
  - The average queue length
  - The average waiting time
  - The average process time

M/M/c example – Taste supermarket

Arrival rate	λ	780 customers/hour
Service rate of one server	μ	???
Arrival rate ÷ (service rate of one server)	λ / μ	???
Number of servers	c	8
Average queue length	L	???
Average waiting time in queue	W	???
Average total time in process		???

M/M/c example – Taste supermarket

Calculate the service rate from the service time

- Service time = 36 seconds / customer
- Note that 1 hour = 3600 seconds
- So # of customers that one server can serve in 1 hour
  = 3600 seconds ÷ (36 seconds / customer)
  = 100 customers
- Service rate = 100 customer / hour

Remember: Units are important!

M/M/c example – Taste supermarket

Arrival rate	λ	780 customers/hour
Service rate of one server	μ	100 customers/hour
Arrival rate ÷ (service rate of one server)	λ / μ	7.8
Number of servers	c	8
Average queue length	L	???
Average waiting time in queue	W	???
Average total time in process		???

M/M/c example – Taste supermarket

$c = 8$

	A	B	C	D	E	F	G	H	I	J	K
1											
2			**Average Queue Length for M/M/c Q**								
3											
4										**Number of Servers**	
5			1	2	3	4	5	6	7	8	9
45		5.20						4.3004	1.0804	0.3680	0.1345
46		5.40						6.6609	1.4441	0.5871	0.1779
47		5.60						11.5178	1.9436	0.6313	0.2330
48		5.80						26.3726	2.6481	0.8225	0.3032
49		6.00							3.6878	1.0707	0.3918
50		6.20							5.2979	1.3967	0.5037
51		6.40							8.0768	1.8040	0.6454
52		6.60							13.7992	2.4198	0.8247
53		6.80							31.1270	3.2441	1.0533
54		7.00								4.4471	1.3471
55		7.20								6.3133	1.7288
56		7.40								9.5102	2.2324
57		7.60								16.0379	2.9113
58		7.80								35.8956	3.8558
59		8.00									5.2264

$L = 35.9$

$\lambda / \mu = 7.8$

M/M/c example – Taste supermarket

Arrival rate	λ	780 customers/hour
Service rate of one server	μ	100 customers/hour
Arrival rate ÷ (service rate of one server)	λ / μ	7.8
Number of servers	c	8
Average queue length	L	35.9 customers
Average waiting time in queue	W	???
Average total time in process		???

M/M/c example – Taste supermarket

- Average waiting time in queue (formula $W = L / \lambda$)

  W = (35.9 customers) / (780 customers / hr)

  = 0.0460 hr

  = 2.76 minutes

Remember: Units are important!

- Average total time in process

  = average wait time + average service time

  = 2.76 minutes + 36 seconds

  = 3.36 minutes

M/M/c example – Taste supermarket

Arrival rate	λ	780 customers/hour
Service rate of one server	μ	100 customers/hour
Arrival rate ÷ (service rate of one server)	λ / μ	7.8
Number of servers	c	8
Average queue length	L	35.9 customers
Average waiting time in queue	W	2.76 minutes
Average total time in process		3.36 minutes

Key takeaways from M/M/c model

1. The M/M/c model is not 100% correct, but it gives useful predictions
2. Service capacity has a nonlinear effect on queue length and wait time
3. There are three ways to reduce waiting time