ORAC: Demand, Capacity and **Queues**

Teow Kiok Liang
Kiok_liang_teow@nhg.com.sg
1-2 Aug 2022

HSOR@NHG
The evidence behind your decisions

Overview

Module objective: From understanding formation of a queue to resource planning

- Understand a Queue
  - Why a queue forms and essential concepts
- Model a Queue
  - What-if, trade-off analysis
- Improve a Queue
- Case studies

Resources

Reading materials

Palvannan RK, Teow KL. Queueing for Healthcare. Journal of Medical Systems. 04/2012; 36(2):541-547.

Green, L. 2011. "Queueing Theory and Modeling". In Handbook of Healthcare Delivery Systems
https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/5474/queueing%20theory%20and%20modeling.pdf

And many others

Free Excel add-in for Queueing

ftp://ftp.wiley.com/public/sci_tech_med/queueing_theory/

1. Understand a Queue

Intuition

Science of Q

Queueing examples

- Non-healthcare
  - Call centre, Lift, HDB housing, Taxi, Ponding, Restaurant games
- Healthcare examples?
  - All the counter waits
  - Appointment lead time (e.g. SOC, Elective surgery)
  - Inpatient beds
  - Transplant
  - Screening, treatment

MCQ: What is the main cause of Q?

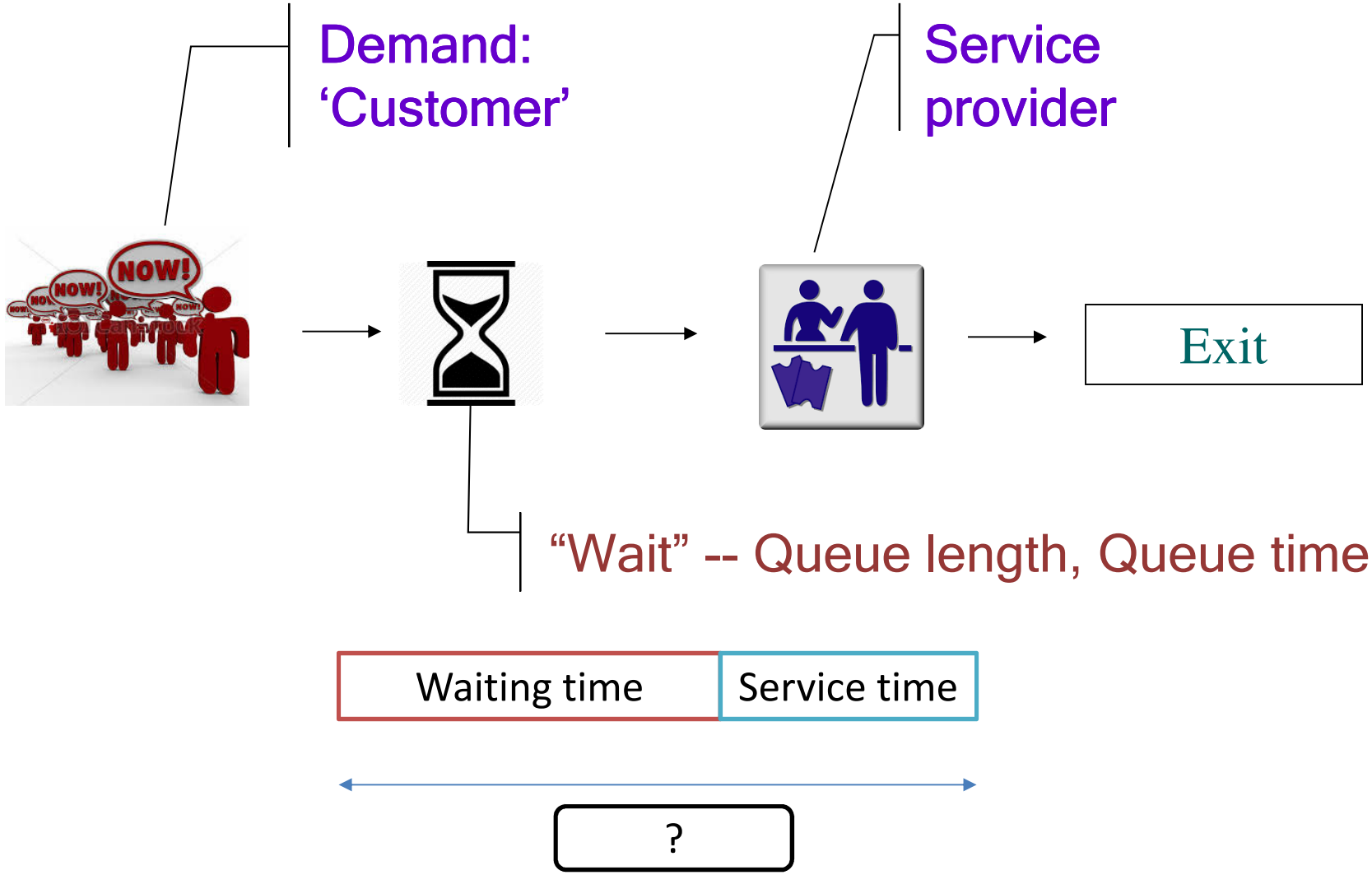
a) Demand more than Supply (utilisation > 100%)
b) Heavy workload
c) Not enough staff
d) Some other thing

Singapore Rainfall Example

- Wiki
  1. Average annual rainfall 2,340 mm
  2. Highest 24-hour rainfall 512 mm (2/12/1978) at Paya Lebar
- CNA 8 Jan 18
  3. The heaviest rainfall of 118.8mm was recorded at the Kim Chuan Road rain gauge between 6.20am to 10.25am - that's about half of Singapore's average monthly rainfall in January, PUB said.
  4. Rainfall in the area peaked between 8am and 8.30am, with an intensity of 56mm per hour

Reference	Hourly rate (mm)
1	?
2	21
3	30
4	?

Terms used in Queueing



Queueing language

	Demand ('Customer')	Service Provider
Rate	**Arrival rate**, λ	Service rate, μ
	How fast is the demand coming in?	How fast can you serve?
Time (=1/rate)	Inter-arrival time: Mean and Variation	**Service time**: Mean and Variation

Utilisation Ratio, $\rho = \lambda / \mu$

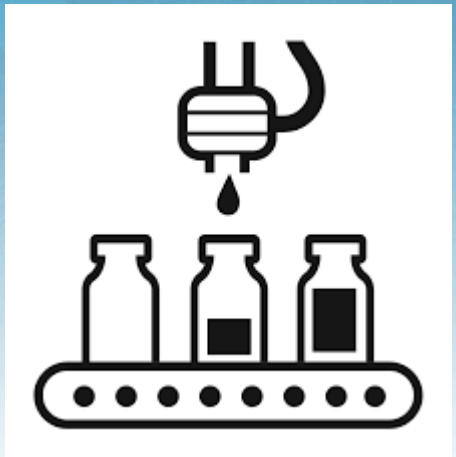
Utilisation

- Demand / Capacity (same unit of measure)
- Examples
  - Bed utilisation (hospitals, hotels)
  - Call centre utilisation
  - Staff utilisation
- Food for thoughts
  - What is a good utilisation rate?
  - How can we use “utilisation” for resource planning?

How do you describe the Q?	Arrival (Demand)	Service (Service)	Utilisation
Call Centre	Call arriving	Resolution time	Staff util
Escalator			
Lift / Bus			
Bike rental			
COVID vaccination			
Medical appt			
Beds			

Bottle filling

- Empty bottle arrives at 60 per min
- Filling up each bottle takes 1 min
- What are the arrival rate, service rate, utilisation ratio and waiting time?



A GP clinic

- Clinic opens from 10 am to 6 pm (no lunch break!)
- Average 40 patients per day
- Each consult took 10 min
- What are the arrival rate, service rate, utilisation ratio and waiting time?

Key elements

- Arrival & Service processes: Deterministic or Random?
  - D: Deterministic
  - M: Markovian – i.e. randomly following a Markov (exponential) distribution

M/M/c/k

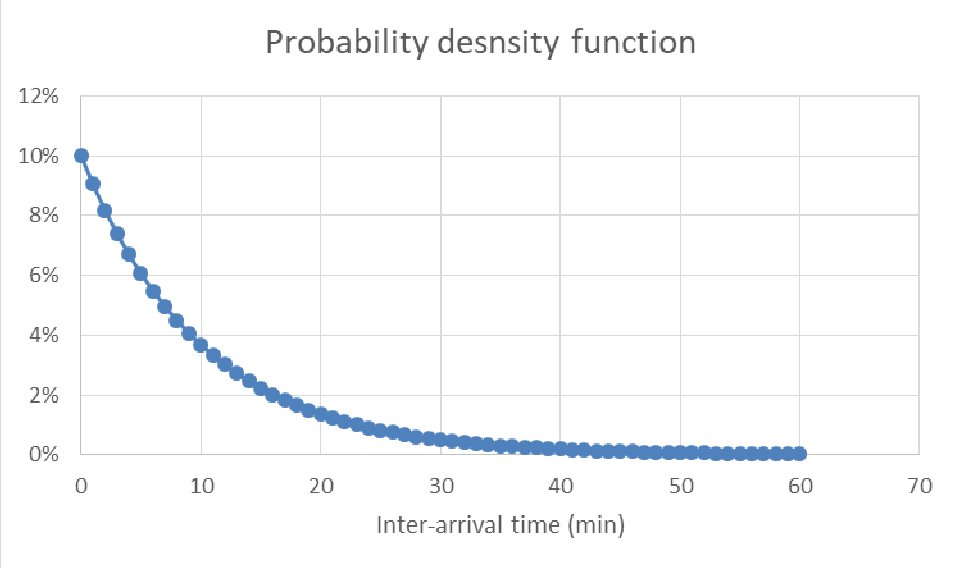
demand

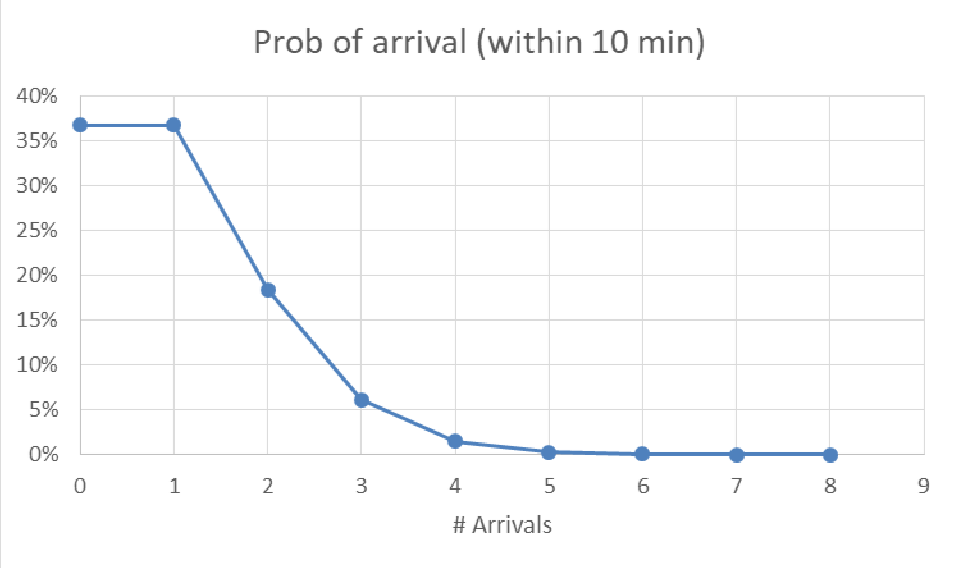
service

No of service providers

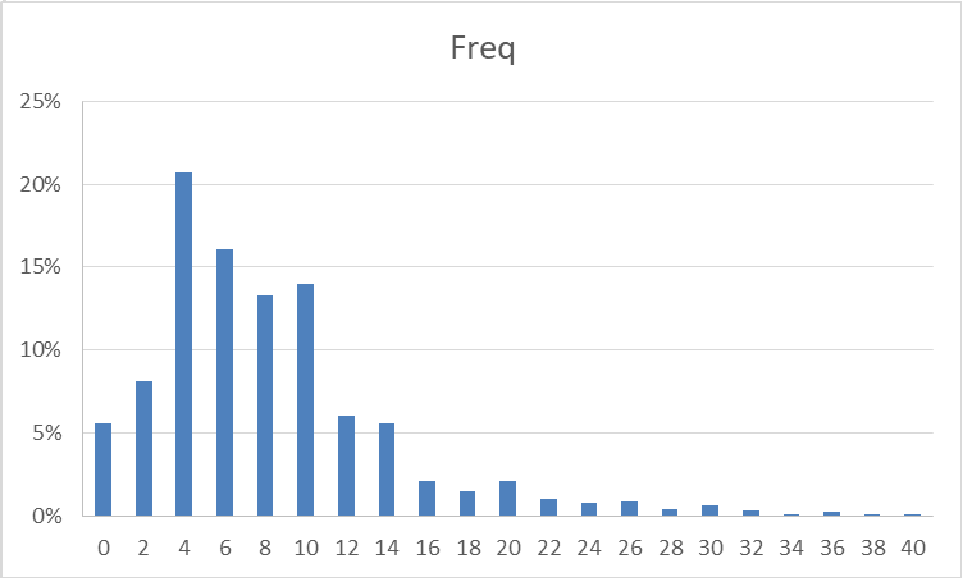
Service providers+ Waiting room

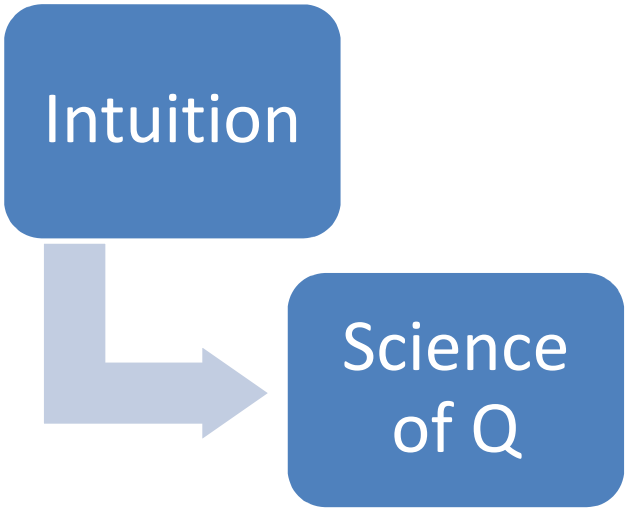
(Negative) Exponential distribution & Poisson Distribution with mean 10 min





A possible consult (service) duration distribution





Recap: basic framework of Queueing

2. Model a Queue

Probabilistic thinking

3 approaches to test/model queue

- Real trial
- Computer simulation
- Mathematical model

Basic Queueing model M/M/1

- What is M/M/1? Hint: Slide 14
- Why M/M/1?
- So what if M/M/1?

FYI only

- State space diagram and then solving equations to derive queueing formula

So what if it is M/M/1

Only need to know 2 parameters:

At steady-state

- **λ: arrival rate**
- **μ: service rate**

And we can derive the following:

- Utilisation $\rho = \lambda / \mu$
- Average # customers in System, L $= \lambda / (\mu - \lambda)$
- Average # customers in Queue, L_q $= L - \rho =$ **ρ^2 / (1 - ρ)**
- Average Waiting Time in system, W $= L / \lambda$
- Average Waiting Time in queue, W_q $= L_q / \lambda$
- System idle time (probability) $= 1 - \rho$

Apply the formula to GP example

QtsPlus

Microsoft Excel - QtsPlus.xls

File Edit View Insert Format Tools Data Window Help

Queueing Theory Software Plus Toolbox 3.0

(c) 2000-2008. All rights reserved.

Select model category, then model and press "Run Model" button.

Model Category: Multi-Server

Single-Server
Multi-Server
Bulk
Priority
Networks
Simulation
Tutor
Index

Category Description: Collection of models for multiple-server

Model Description: Poisson/exponential multi-server queu ... apacity, FIFO: calculates the major measures of effec

Run Model

Help

M/M/c: Poisson Arrivals to Multiple Exponential Servers
M/M/c/K: Multiple-Servers/Limited Capacity
M/G/c/c: Pure Overflow Model
Optimal Number of Servers for an M/G/c/c
M/M/infinity: Poisson Input and Unlimited Exponential Service
M/G/infinity: Poisson Input, Unlimited Number General Servers
Markov Multi-Server, Finite-Source Queue without Spares
Markov Multi-Server, Finite-Source Queue with Spares
M/E(k)/c: Multi-Server Poisson-Erlang(k) Queue

Start-up | GlobalParameters

Microsoft Excel - M_M_c.xls

File Edit View Insert Format Tools Data Window Help

H9

M/M/c: POISSON ARRIVALS TO MULTIPLE EXPONENTIAL SERVERS

Input Parameters:

Arrival rate (λ)	5.
Mean service time ($1/\mu$)	0.4
Number of servers in the system (c)	3

Plot Parameters:

Maximum size for probability chart	15
Total time horizon for probability plotting	2.

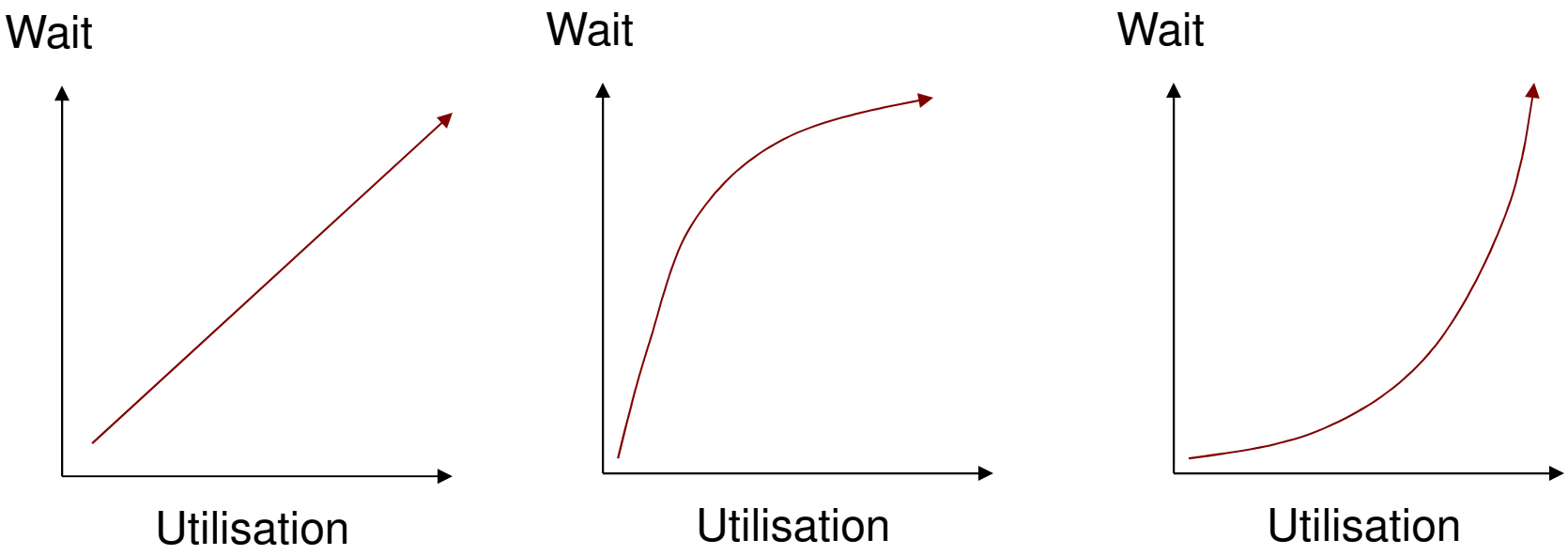
Results:

Mean interarrival time ($1/\lambda$)	0.2
Service rate (μ)	2.5
Average # arrivals in mean service time (r)	2.
Server utilization (ρ)	66.67%
Fraction of time all servers are idle (p_0)	0.111111
Mean number of customers in the system (L)	2.888889
Mean number of customers in the queue (Lq)	0.888889
Mean wait time (W)	0.577778
Mean wait time in the queue (Wq)	0.177778
Probability arriving customer is delayed in queue (1-Wq(0))	0.444444

M_M_c | SizeDistributionChart | TimeDistributionChart | PlotData

ftp://ftp.wiley.com/public/sci_tech_med/queueing_theory/

Q: Waiting time vs Utilisation



Which graph is correct?

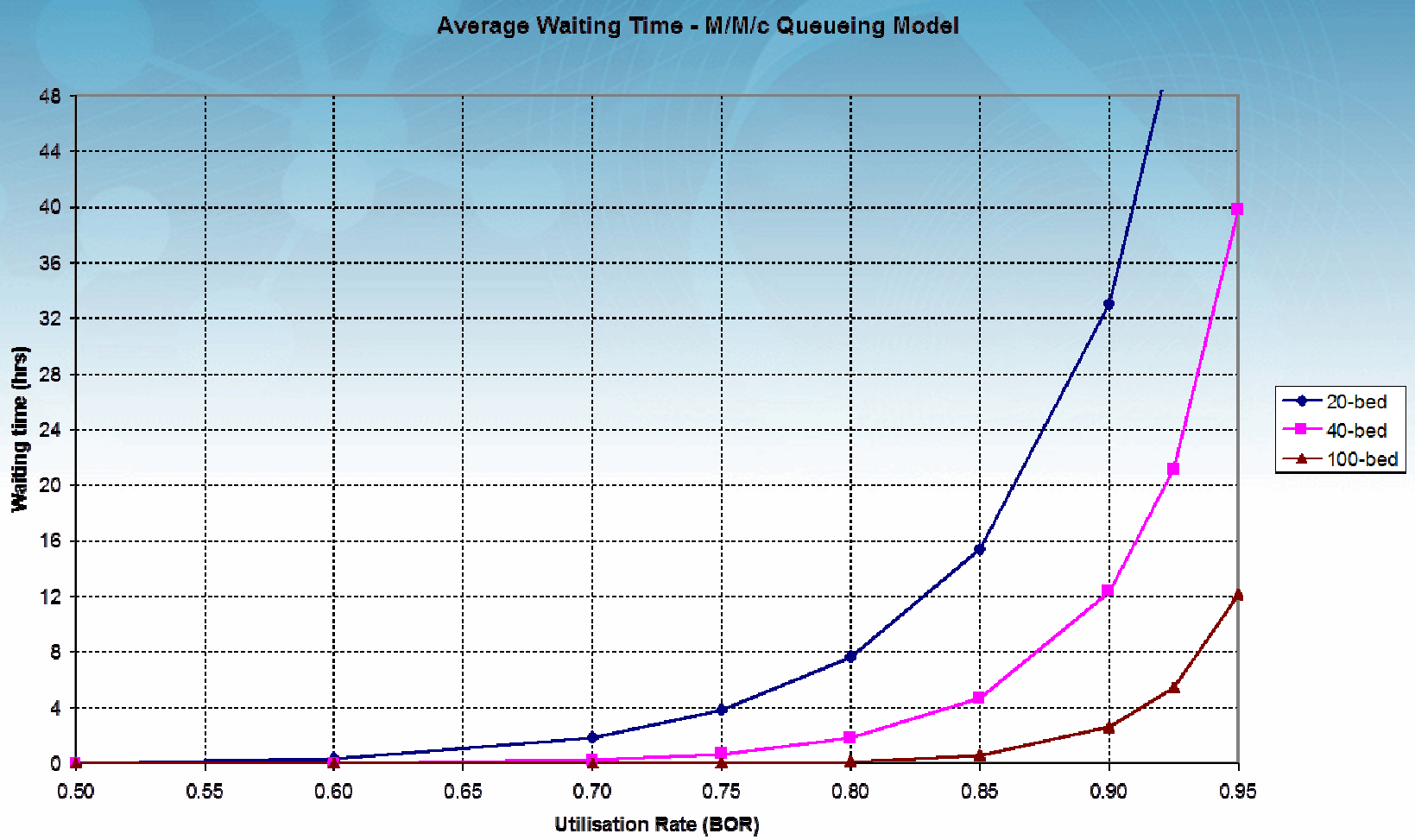
Insights / Observations

- From M/M/1
- From GP and Cola
- And using M/M/1 for planning

What is beyond M/M/1?

- First extension: from Solo to Group practice
  - Notation: M/M/c
  - Does it matter? E.g. is it similar to a few M/M/1? GP example with QtsPlus
  - Formula?
  - Examples

Pooling and Partitioning



2 Model a Queue

Examples of 'Group practice'?

- Non-healthcare
- Healthcare

Another Q type: 'Cannot wait'

- Err, then there is no waiting time by definition – customers are 'rejected'
  - measure the performance of the system by Probability of 'Failure'
- Examples?
- QTSPLUS: M/G/C/C or M/M/c/k model
  - What is 'G'?

Probability of a patient finding no bed at post-recovery

Prob of not finding a bed after Ops		# of theatres * % req beds * Recovery time / Ops duration (Min beds req'd (λ/μ))									
		5	6	7	8	9	10	11	12	13	14
# of Recovery beds	22	0.0%	0.0%	0.0%	**0.0%**	0.0%	0.0%	**0.1%**	0.3%	0.7%	1.2%
	21	0.0%	0.0%	0.0%	**0.0%**	0.0%	0.1%	**0.2%**	0.6%	1.1%	2.0%
	20	0.0%	0.0%	0.0%	**0.0%**	0.1%	0.2%	**0.5%**	1.0%	1.8%	3.0%
	19	0.0%	0.0%	0.0%	**0.0%**	0.1%	0.4%	**0.8%**	1.6%	2.8%	4.4%
	18	0.0%	0.0%	0.0%	**0.1%**	0.3%	0.7%	**1.5%**	2.7%	4.3%	6.3%
	17	0.0%	0.0%	0.1%	**0.2%**	0.6%	1.3%	**2.5%**	4.1%	6.2%	8.6%
	16	0.0%	0.0%	0.1%	**0.5%**	1.1%	2.2%	**3.9%**	6.0%	8.6%	11.5%
	15	0.0%	0.1%	0.3%	**0.9%**	2.0%	3.6%	**5.9%**	8.6%	11.6%	14.8%
	14	0.0%	0.2%	0.7%	**1.7%**	3.4%	5.7%	**8.5%**	11.7%	15.1%	18.6%
	13	0.1%	0.5%	1.4%	**3.1%**	5.4%	8.4%	**11.9%**	15.5%	19.2%	22.8%
	12	0.3%	1.1%	2.7%	**5.1%**	8.3%	12.0%	**15.9%**	19.9%	23.7%	27.5%
	11	0.8%	2.3%	4.8%	**8.1%**	12.1%	16.3%	**20.6%**	24.8%	28.7%	32.4%
	10	1.8%	4.3%	7.9%	**12.2%**	16.8%	21.5%	**26.0%**	30.2%	34.1%	37.7%

E.g. 6 OT, 80% req beds, 75min recovery, 30min ops duration (column "12"); and 20 beds (row) will result in 1% of patient not finding a bed

Know the trade-offs!!
Aka, No-free-lunch

% demand met

Ave Demand	# providers 1	2	3	4	8
1	86%	100%	100%	100%	100%
2	67%	96%	100%	100%	100%
4	43%	78%	95%	99%	100%
8	24%	48%	69%	86%	100%
16	12%	25%	37%	50%	92%

% Utilisation unused

Ave Demand	# providers 1	2	3	4	8
1	57%	75%	83%	88%	94%
2	33%	52%	67%	75%	88%
4	14%	22%	37%	50%	75%
8	5%	5%	8%	14%	50%
16	1%	1%	1%	1%	8%

More details:

- Average service time = 0.5 hour
- Hourly demand given
- M/M/c/2c model, that is, patient will wait for at most 1 patient in front of him

Is there an universal Q model?

General approximation, Ave Waiting time ~

$$\frac{\left(\sigma_x^2 + \left(\frac{\sigma_c}{c}\right)^2\right)\lambda}{2(1-\rho)}$$

Service time variance

Arrival time variance

Arrival rate

# of server

Utilisation rate

Other Queueing models with QtsPlus

- M/D/1, D/M/1 etc
- Faster or 2 machines?
- Network
- Time varying?

More realistic Q,
harder to analyse

A useful formula: Little's law

Ave no of people in system = Ave arrival rate x Ave time in system

$$L = \lambda \times W$$

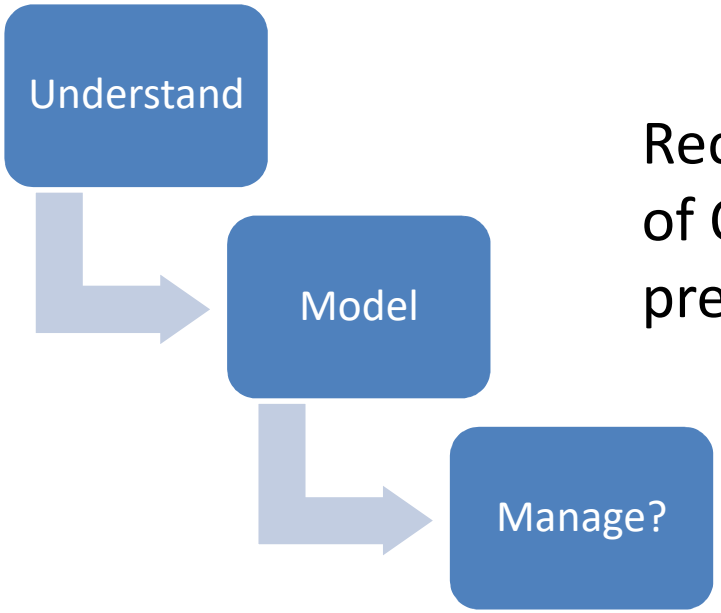
- You receive 50 emails a day and your inbox has 100 emails. How long do you take to clear (delete) your mails
- 10 patients enter into observation ward per hour. Average time spent in ward is 2.5 hr. What is the average number of patients in the ward?
- Chronic disease prevalence
- # of COVID beds occupied

Similar to Stock and Flow!

Recap

We learn a few models. These models (and queueing concepts) help to plan the <u>capacity</u> needed for a given demand and service level target (such as waiting time, loss rate)

So, what are the Queueing models covered?



Recap: understand the causes of Q, and learn some models to predict Q

3. Improve a Queue

How to reduce Q?

We see why Q forms, how to quantify variation, some basic models to predict Q

1. Reduce demand (possible?)
2. Increase supply (capacity planning)
   - Identify and work on the bottleneck

What else?

Some suggestions to improve Q

- **Reduce variation** (??)
  - Demand levelling (macro and operational level)
  - Supply: work standardization (?)
- Pooling, or 'flexible partition'
  - E.g. cross-training
- Demand – supply matching
- Put 'uncertain' cases last
- Shortest job first

Case study

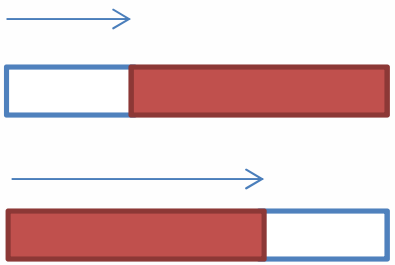
Tactical planning

- Waiting for pharmacy dispensing
- Cannot change demand nor capacity
- What else can be done?

Queue 'discipline'

- "First-come-first-serve" (or FIFO) is fairest and gives lowest variance, but not necessarily gives lowest <u>average</u> waiting time
- "Shortest-job-first" has lower waiting time, but potentially starve the longer ones

First jobs have 0 waiting time, waiting time for the second jobs differ depending on which job is scheduled first



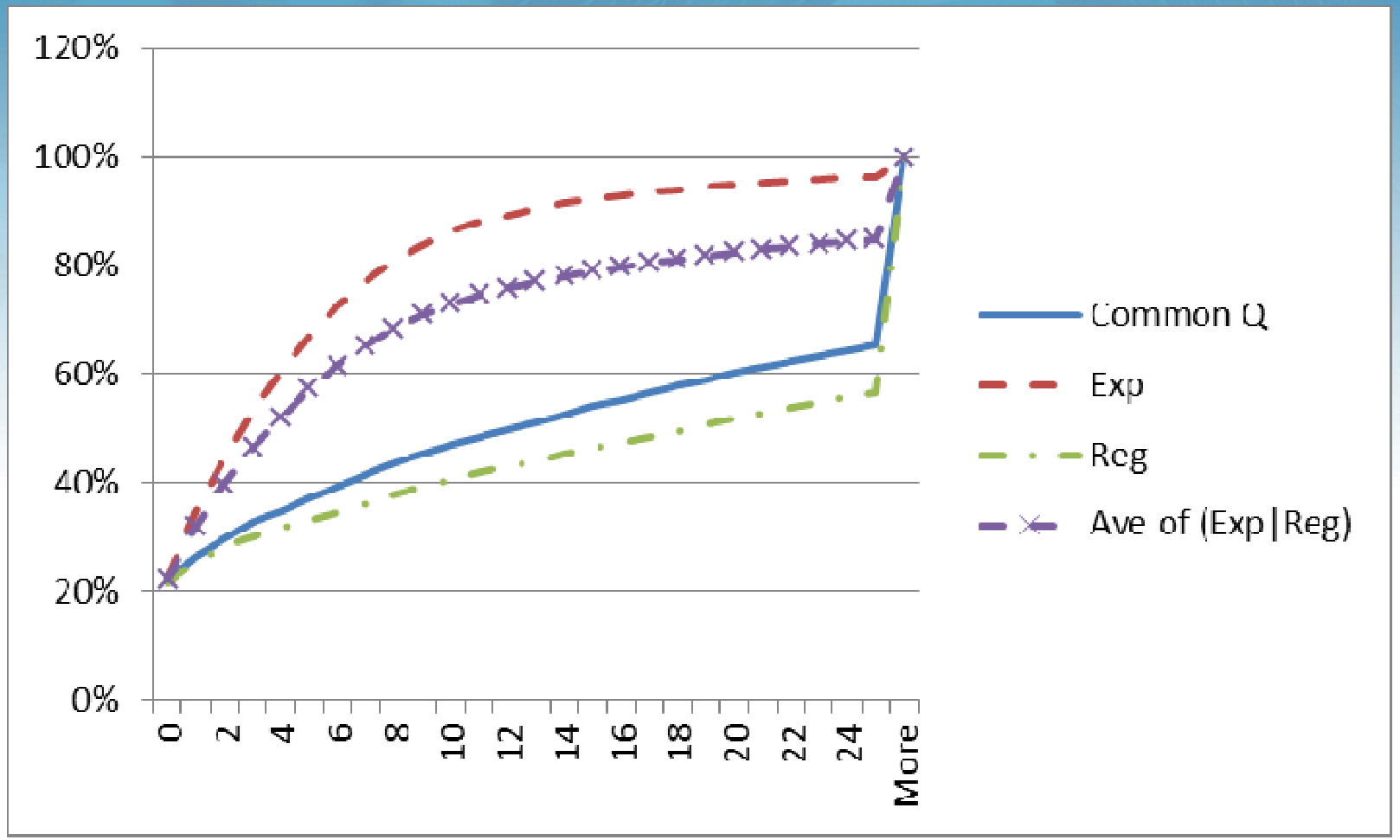
Support for 'Express Q'

- Scenarios

1. Common queue (i.e. no differentiation) between 'Express' and ' Regular'
2. (Soft) Split of 'Regular' and 'Express' counters

Q: 'how' to implement 'express' queue?

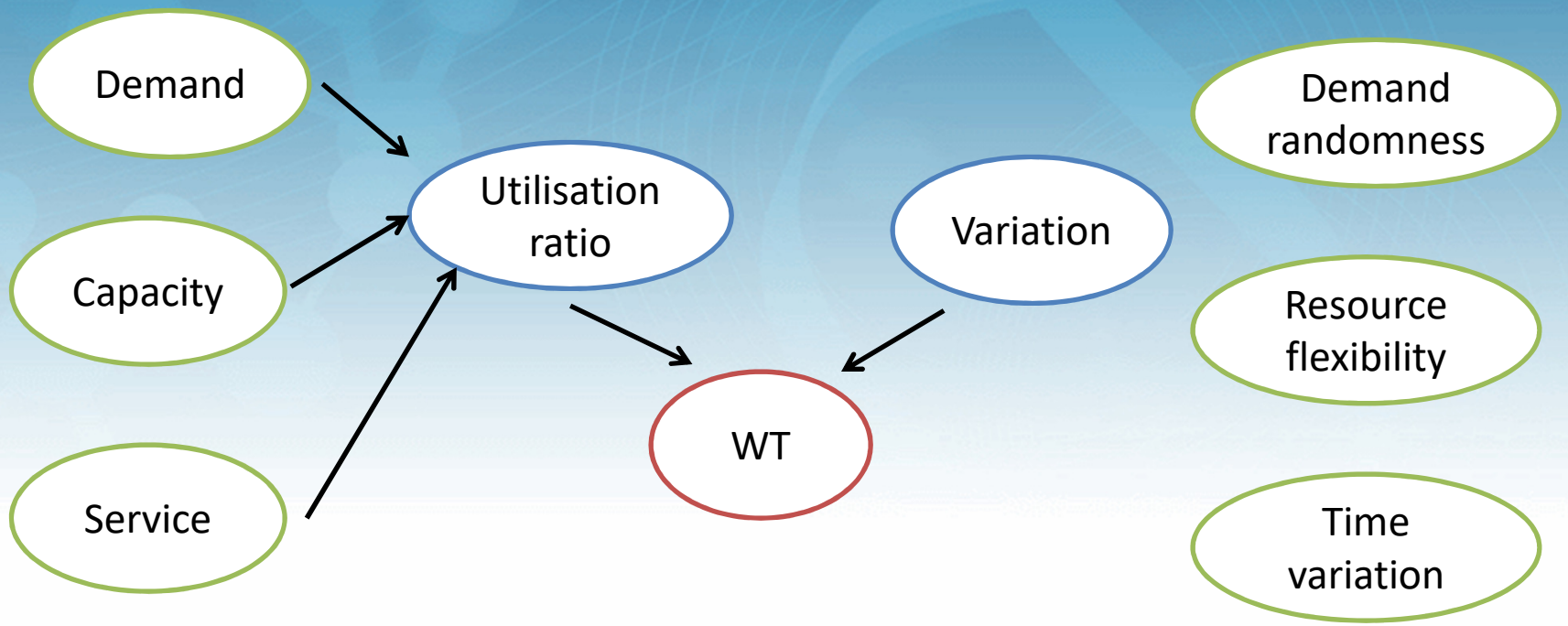
Results: % of cases waiting within 'x' min



Using Queueing's perspective, how to read the following?

- Process improvement
- Batching
- Service differentiation
- Prioritisation of patients
- Create flexibility
- Psychology of queue

Summary


Demand
Capacity
Service
Utilisation ratio
Variation
WT
Demand randomness
Resource flexibility
Time variation