The University of Manchester
Alliance Manchester Business School

Midlands Decision Support Network
Midlands Analyst Network - 21 March 2024

The 85% bed occupancy fallacy

Dr Nathan Proudlove
nathan.proudlove@manchester.ac.uk

MANCHESTER 1824
The University of Manchester
Alliance Manchester Business School

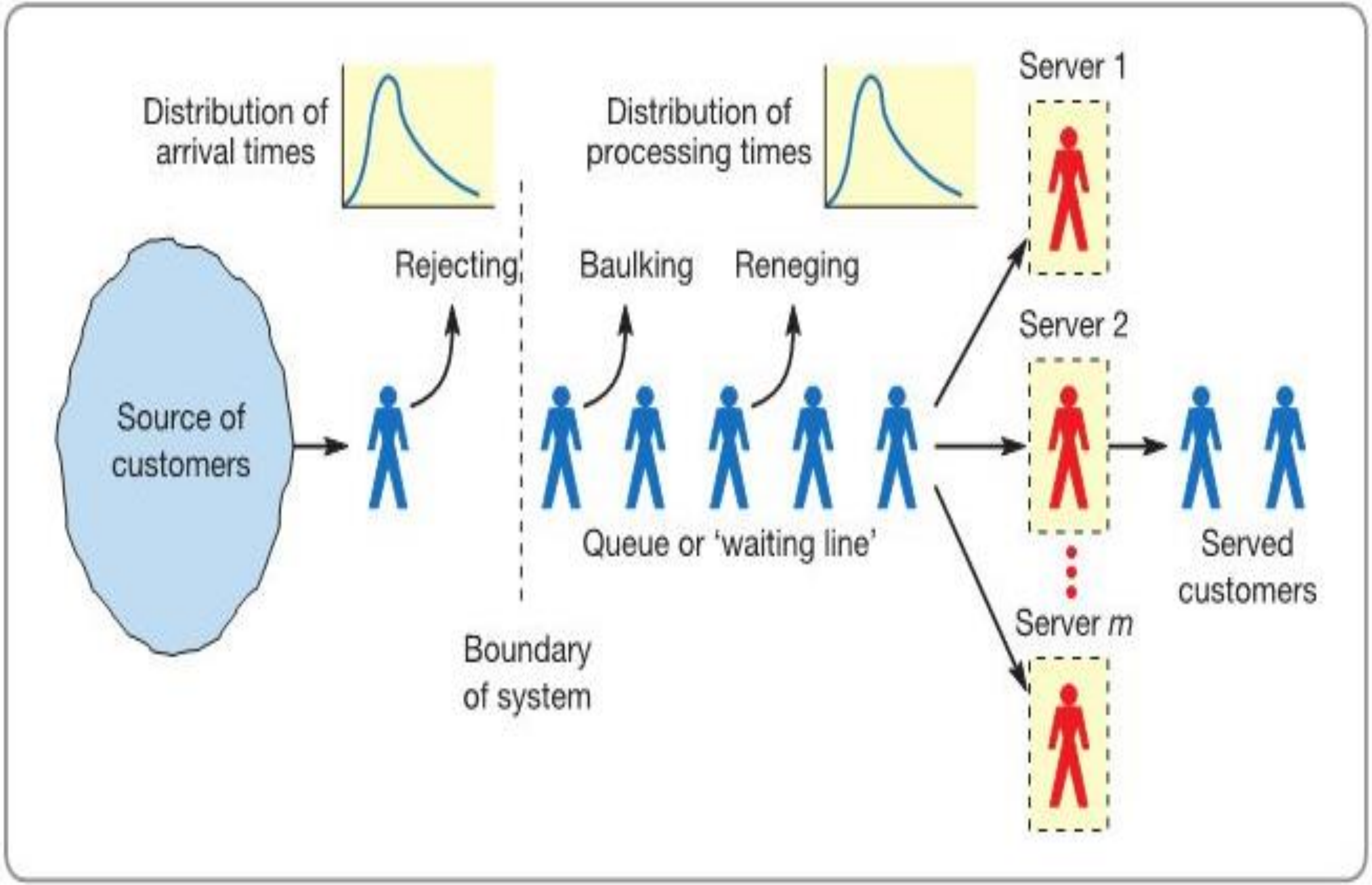
Insights from Queuing Systems

Prof Neil Walton (Fong *et al.*, 2022) [Huddle 2 Nov 2023]

- www.midlandsdecisionsupport.nhs.uk/communities-of-practice/midlands-analyst-network/
- With queuing models, we usually consider the mean e.g. waiting time as a performance measure
- In healthcare we usually want some % of patients to be withing a target time
- The shape of the distribution of individual patients waiting times is (negative) exponential

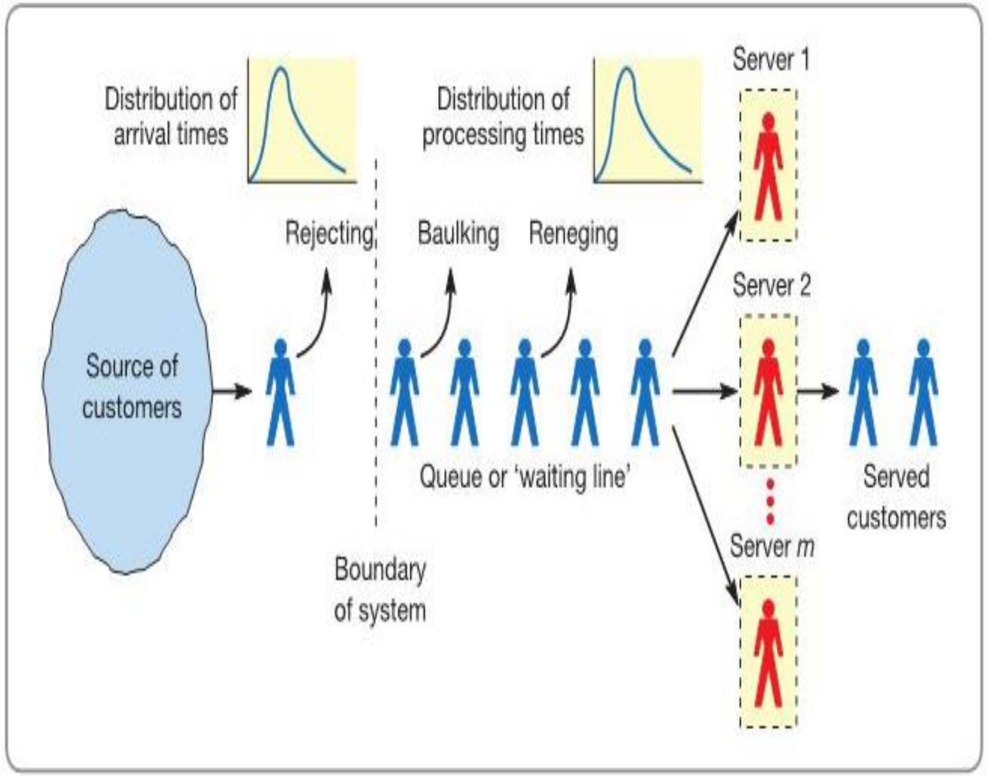
MANCHESTER 1824
The University of Manchester
Alliance Manchester Business School

Single Echelon Queuing Models



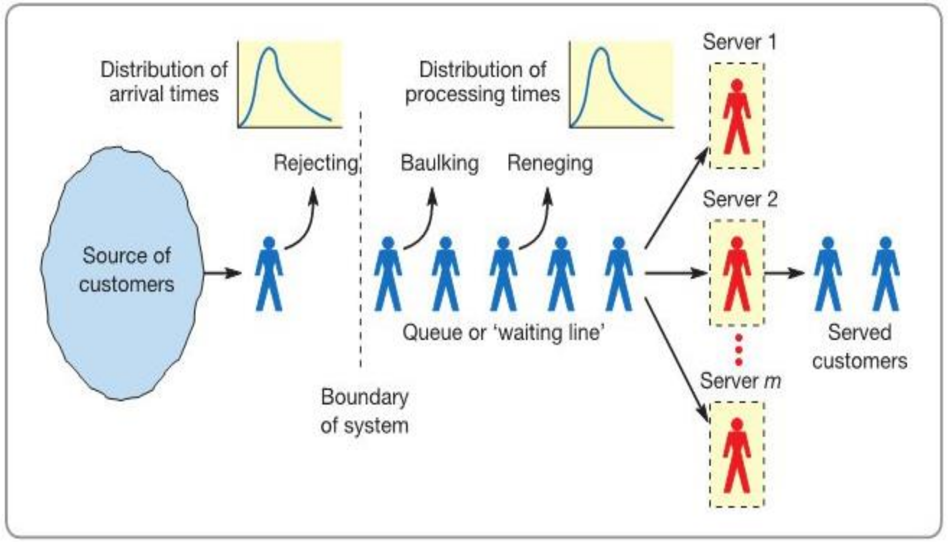
(Slack *et al*, 2016) p.349

Generally interested in Performance metrics such as …



- Expected waiting time in queue
  - or probability (risk) that waiting time exceeds some target time
- Expected queue length
- Probability (risk) all servers busy (so customer has wait)
  - e.g., patient waits in A&E for an inpatient bed)
- Probability (risk) a restricted queue is full (so customer is rejected)
  - e.g., patient becomes an outlier or transferred to another hospital
- Utilisation of servers (e.g., staff, cubicles, beds)

Modelling queuing systems



Analytical

- Lots of assumptions…
- Steady state (long-run averages)

Simulation

- Freeform

Markov → Exact equations

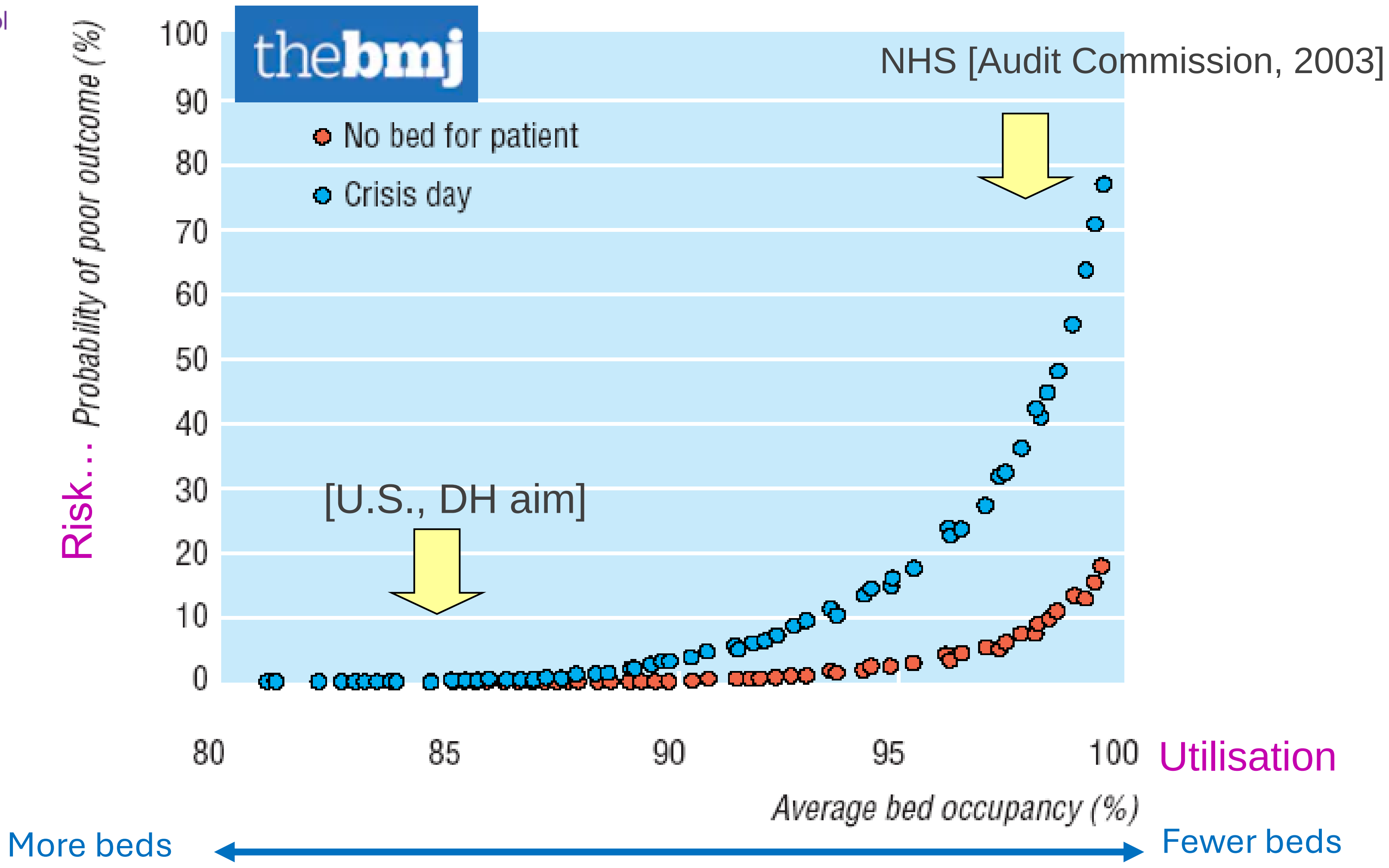
Generalised → Approximate equations

Monte Carlo → Spreadsheets / coding

Discrete Event → Specialist software / lots of coding

MANCHESTER 1824
The University of Manchester
Alliance Manchester Business School

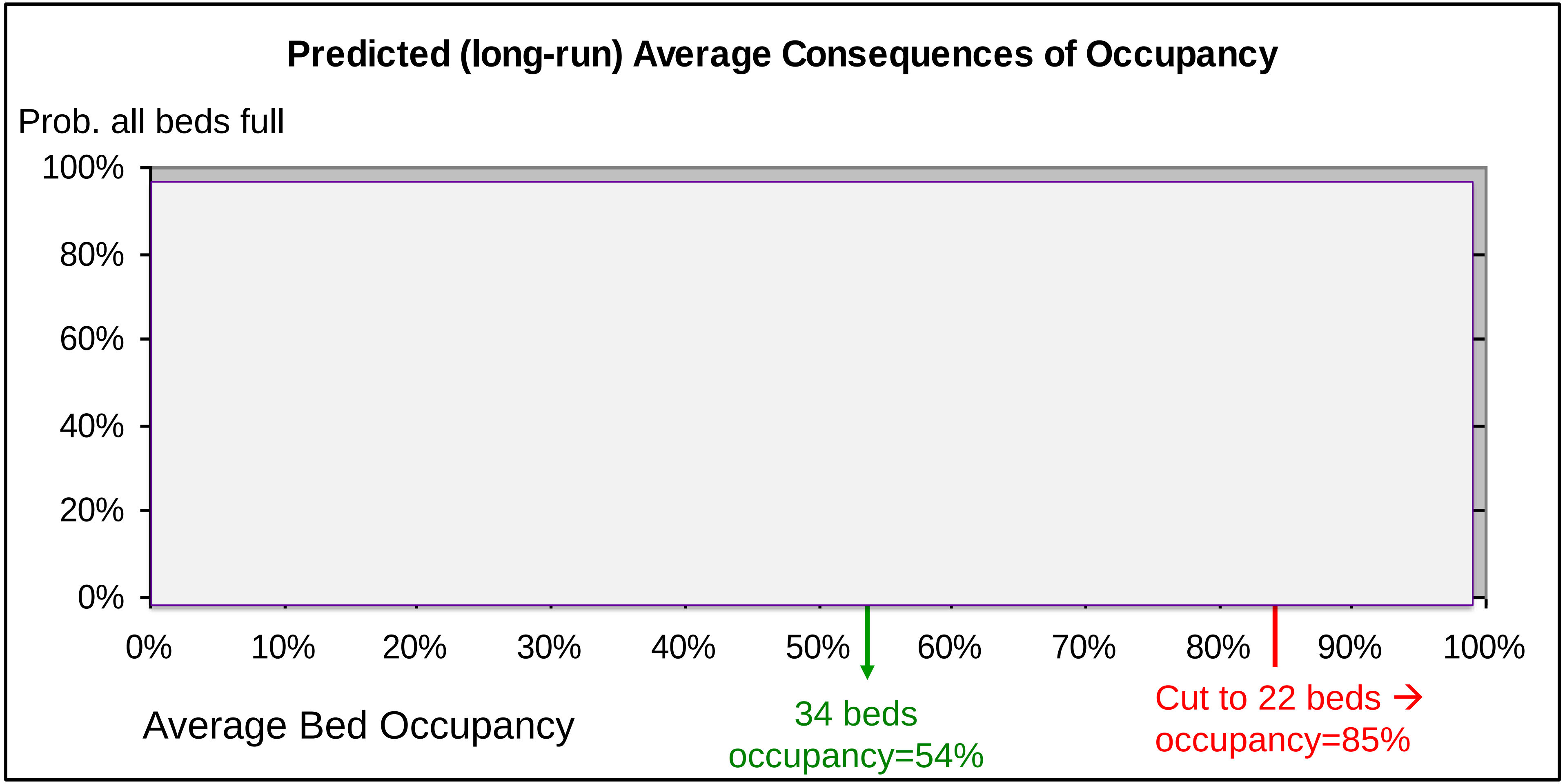
"bed occupancy should be 85% (or 82 or 84…)"



results from computer simulations of large, medical inpatient bed pool (Bagust *et al.*, 1999, p.156)

The University of Manchester
Alliance Manchester Business School

A Greater Manchester hospital's Paediatric Bed Pool (simple model!)



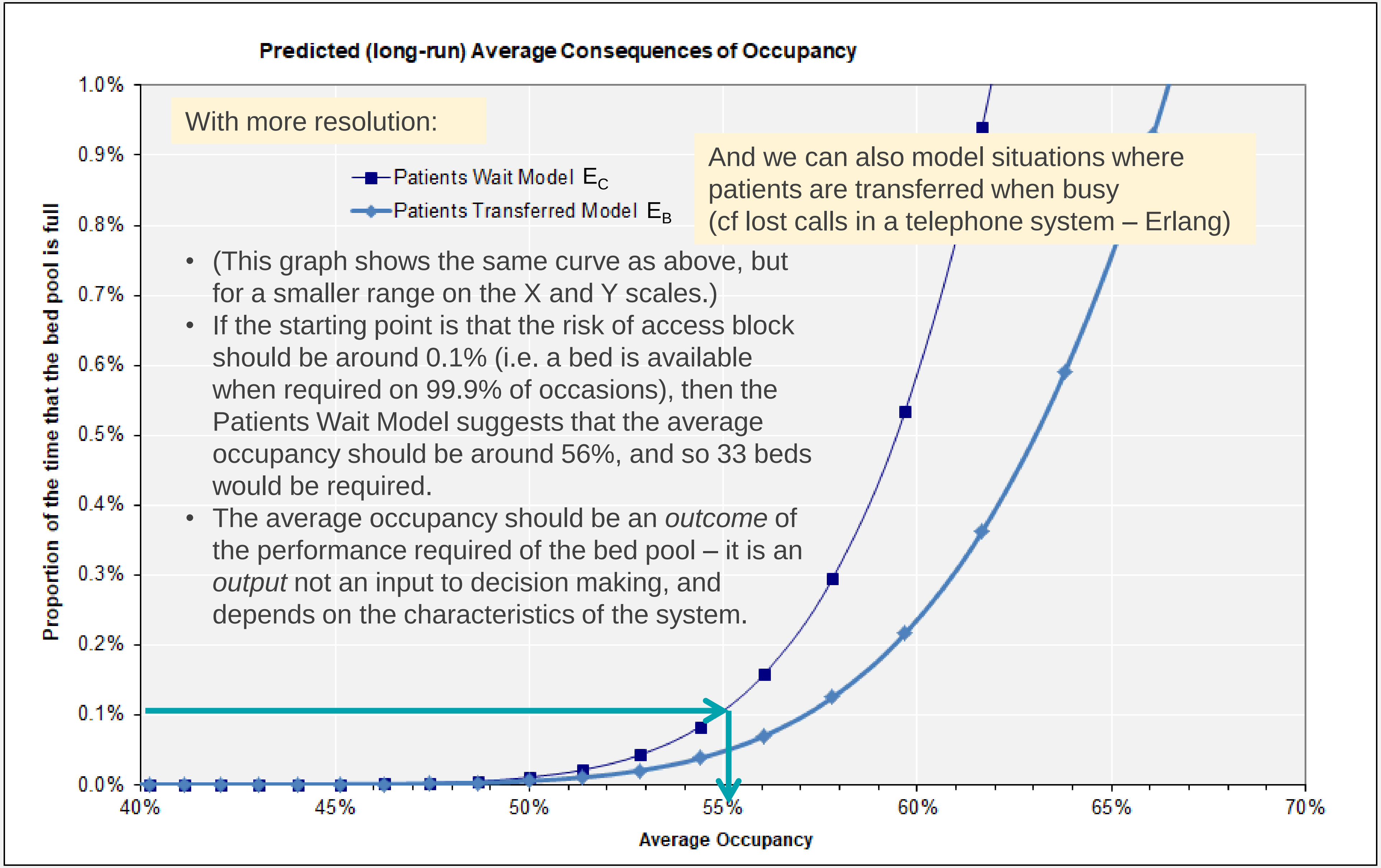
Occupancy is *not a target*!
the appropriate level is a consequence of need to *absorb variation*

"The 85% bed occupancy fallacy: The use, misuse and insights of queuing theory" (Proudlove, 2020)

The curve was drawn using results from queuing theory (Erlang equations)
- makes simplistic assumptions, but gives you a quick idea for simple situations; for more complex situations you need simulation (more laborious!)

With more resolution:

And we can also model situations where patients are transferred when busy (cf lost calls in a telephone system – Erlang)



- (This graph shows the same curve as above, but for a smaller range on the X and Y scales.)
- If the starting point is that the risk of access block should be around 0.1% (i.e. a bed is available when required on 99.9% of occasions), then the Patients Wait Model suggests that the average occupancy should be around 56%, and so 33 beds would be required.
- The average occupancy should be an *outcome* of the performance required of the bed pool – it is an *output* not an input to decision making, and depends on the characteristics of the system.

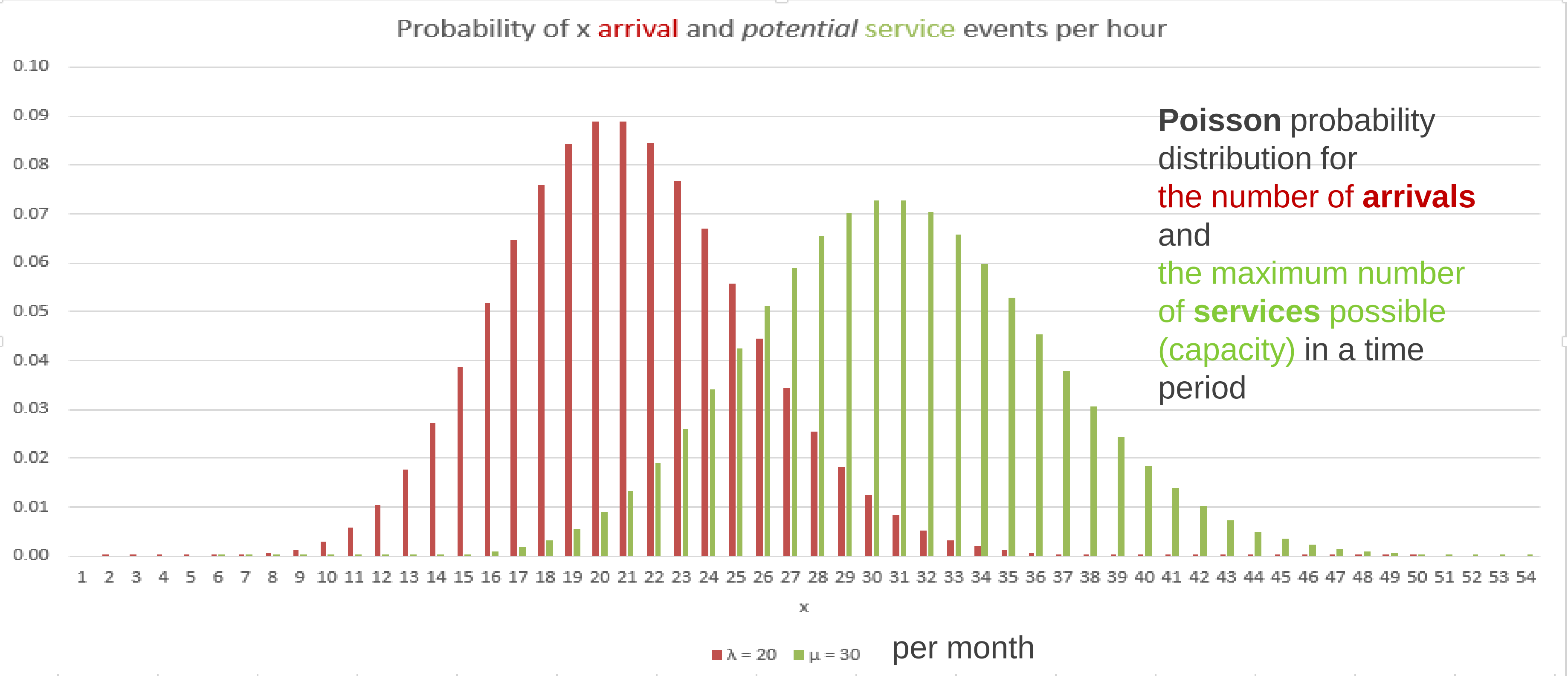
Occupancy						46%		47%	49%	50%	51%		53%	54%		56%		58%		60%		62%		64%		66%			69%	
Beds						40		39	38	37	36		35	34		33		32		31		30		29		28			27	

See Proudlove (2020)

The University of Manchester
Alliance Manchester Business School

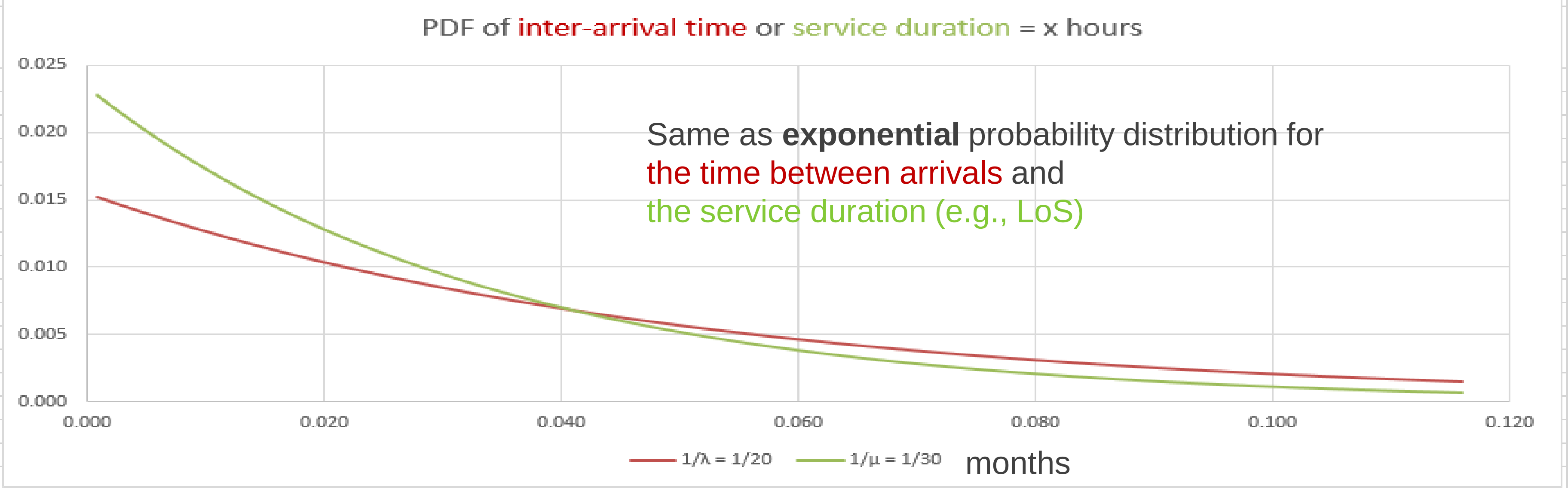
The simplest models make lots of assumptions including:

- steady state
- 'Markovian' probability distributions



Poisson probability distribution for the number of **arrivals** and the maximum number of **services** possible (capacity) in a time period

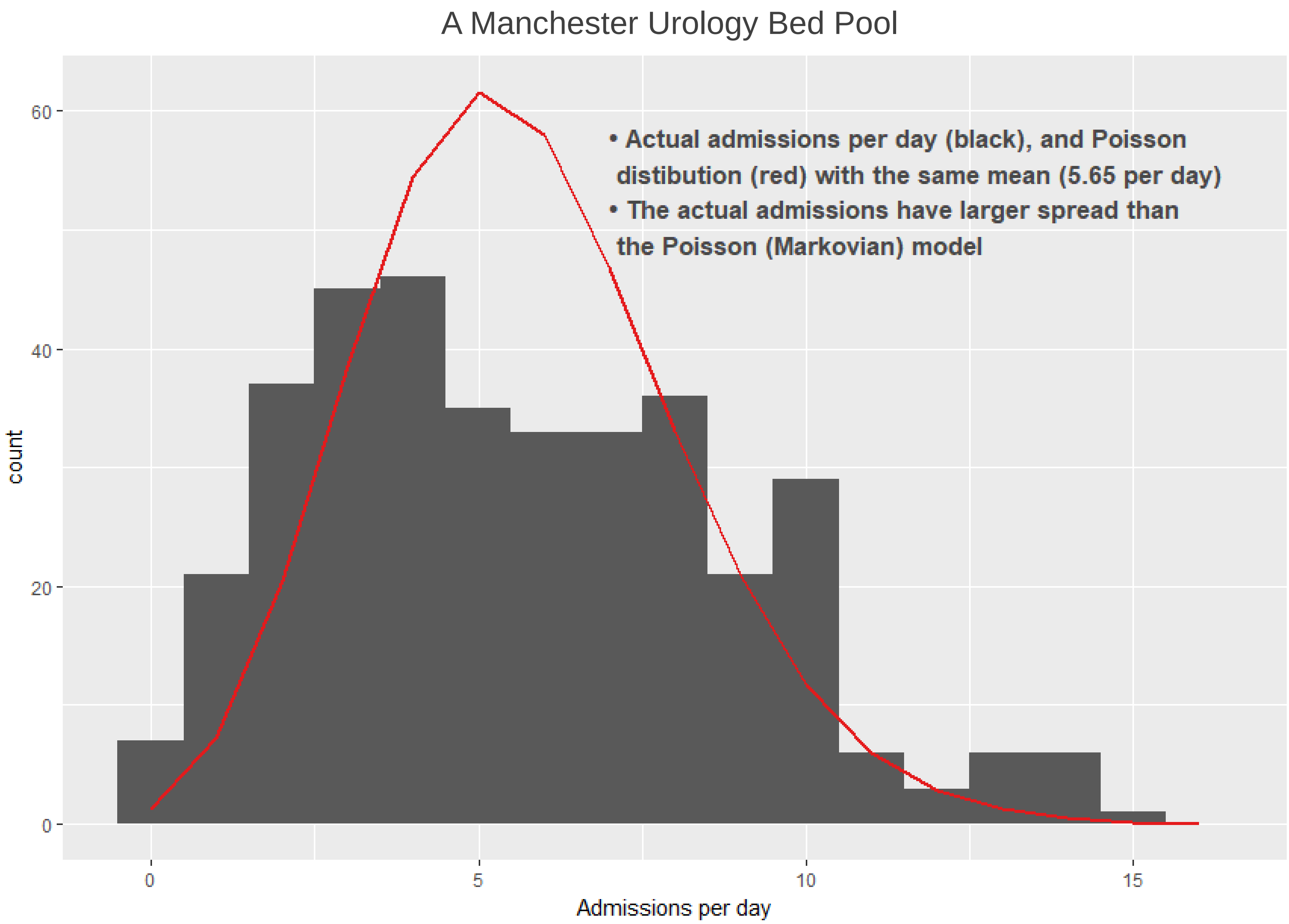
per month



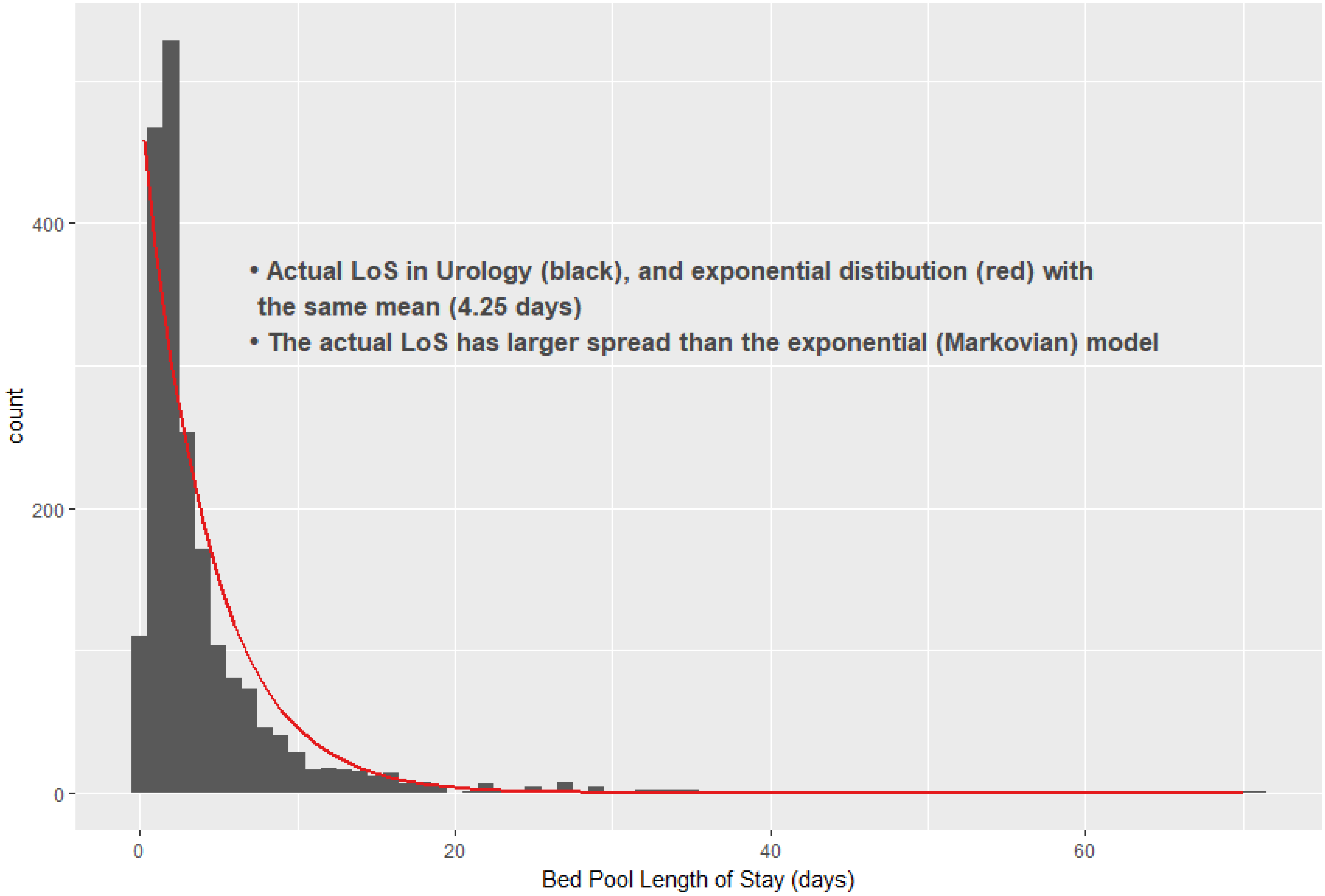
Same as **exponential** probability distribution for the time between arrivals and the service duration (e.g., LoS)

months

Do the assumptions fit?



See Proudlove (2020)



See Proudlove (2020)

MANCHESTER 1824

The University of Manchester
Alliance Manchester Business School

W_q = expected waiting time in queue
λ = mean arrival rate
μ = mean service rate (potential, if customers)
$1/\mu$ = mean service duration [e.g., ALoS]
s = number of servers [e.g., beds]
ρ = utilisation = $\lambda/(s\mu)$
c_a = coefficient of variation of arrivals
[std dev of time between arrivals / its mean]
c_e = coefficient of variation of service
[std dev of service duration / its mean (t_e)]

Simplest model (Markovian, 1 server, customers wait)
(**M/M/1**):(GD/∞/∞)

$$W_q = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu}$$

Relaxing Markovian assumptions
(to any 'General' distribution)**:**
(**G/G/1**):(GD/∞/∞)

$$W_q \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu}$$

The Kingman Formula

Delay (time in queue) | Variation term | Utilisation term | Processing time

(Markovian *c*'s are 1, so *V* term = 1)

Delay ≈ V × U × T

Modelling multiple servers (from the same queue)**:**
(**G/G/s**):(GD/∞/∞)

$$W_q \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho^{\sqrt{2(s+1)}-1}}{s(1-\rho)}\right)\left(\frac{1}{\mu}\right)$$

The VUT Relationship

Delay (time in queue) | Variation term | Utilisation term | Processing time

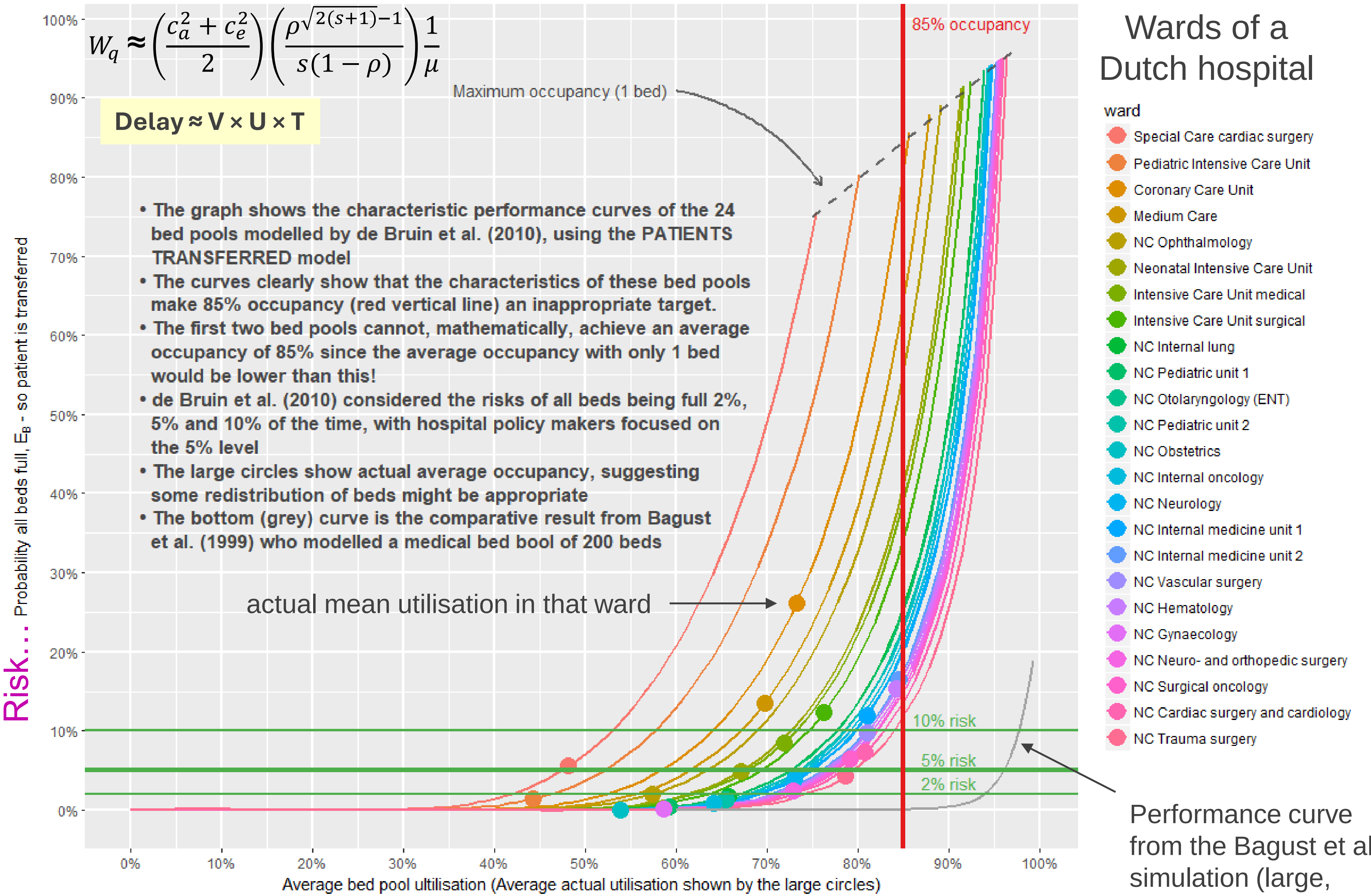
MANCHESTER 1824
The University of Manchester
Alliance Manchester Business School

Insight from the behaviour of systems:

Different bed pools have *different* queuing system performance curves depending on

- Variation
- Utilisation
- (and number of servers)
- Service Duration
- ➢ *VUT curves*

○ So, the same risk of all beds being full when needed would require different average utilisations (so numbers of beds)



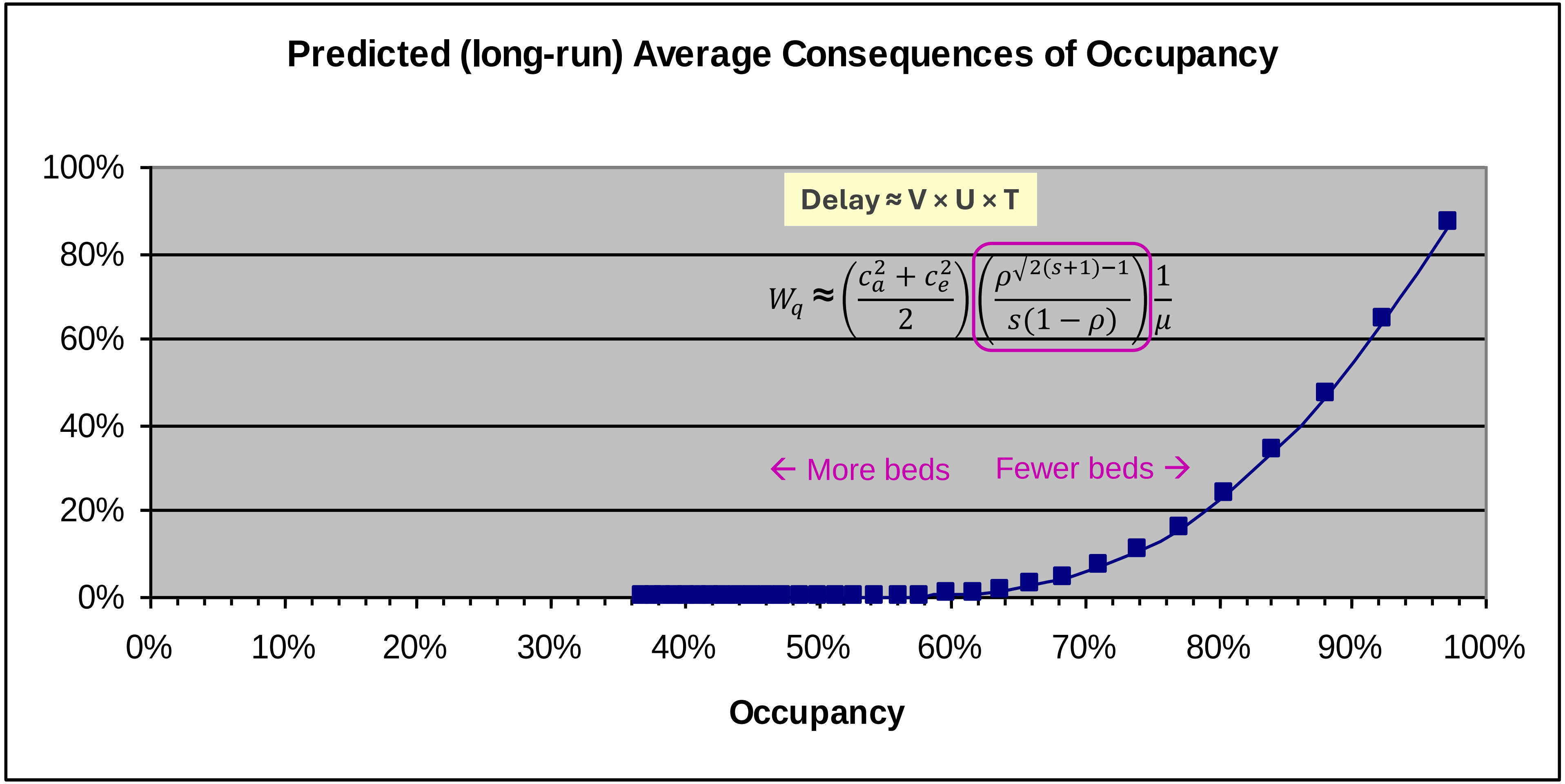
Performance curve from the Bagust et al simulation (large, medical bed pool)

Occupancy

See Proudlove (2020)

The University of Manchester
Alliance Manchester Business School

Accepting the systems' characteristics (variation and service times) you can slide up or down the performance curve



14

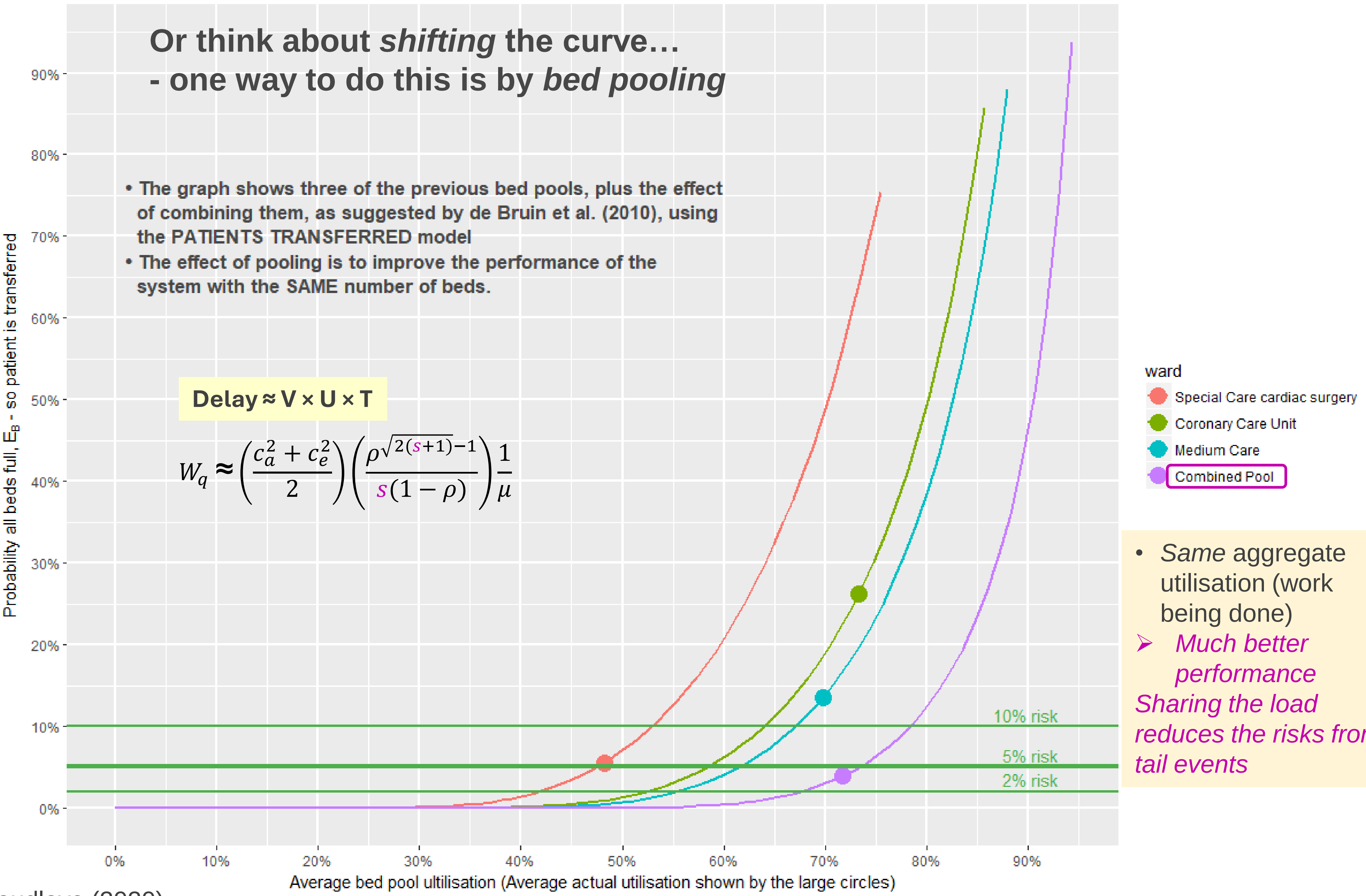
MANCHESTER 1824
The University of Manchester
Alliance Manchester Business Sch

Or think about *shifting* the curve…
- one way to do this is by *bed pooling*

- The graph shows three of the previous bed pools, plus the effect of combining them, as suggested by de Bruin et al. (2010), using the PATIENTS TRANSFERRED model
- The effect of pooling is to improve the performance of the system with the SAME number of beds.

Delay ≈ V × U × T

$$W_q \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho^{\sqrt{2(s+1)}-1}}{s(1-\rho)}\right)\frac{1}{\mu}$$



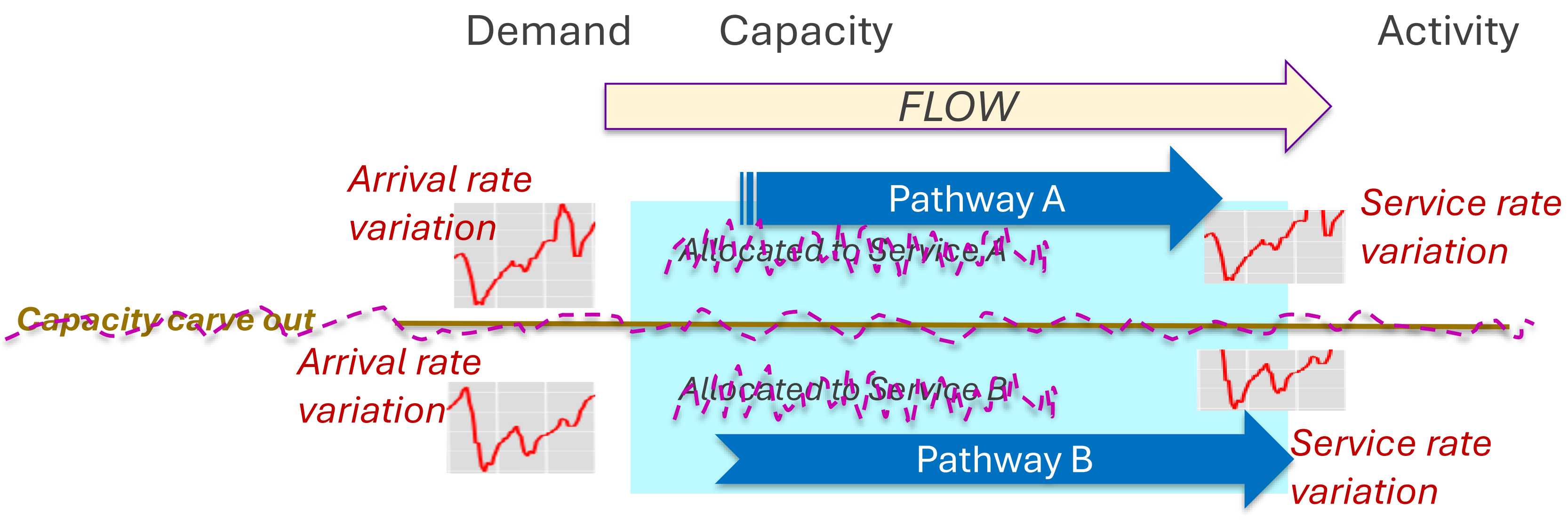
- *Same* aggregate utilisation (work being done)
  - ➢ *Much better performance*

Sharing the load reduces the risks from tail events

See Proudlove (2020)

The University of Manchester
Alliance Manchester Business School

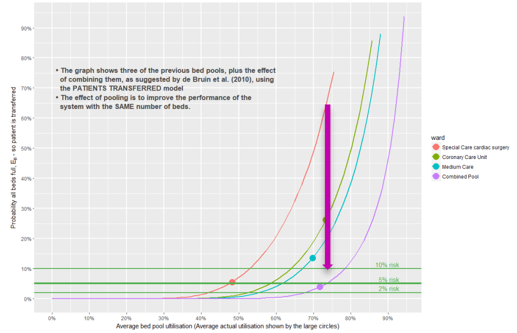
Capacity *Carve-out* vs. *Pooling* or *Segmentation*

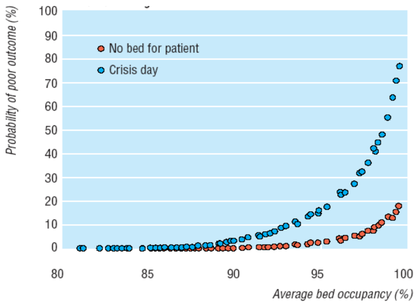


Segmentation

- Tailoring to customer segment: faster service rates and/or lower variety of job types
- More efficient pathways outweigh carve-out

Capacity Pooling





Relation of performance risks to average bed occupancy

What happened in NHS trusts?!

- There may be good reasons for 'carve-out'? (depends on system and objectives)
  - But can you increase flexibility? [e.g., short-notice call-in to unused appointment slots carved-out for expected urgent demand?]

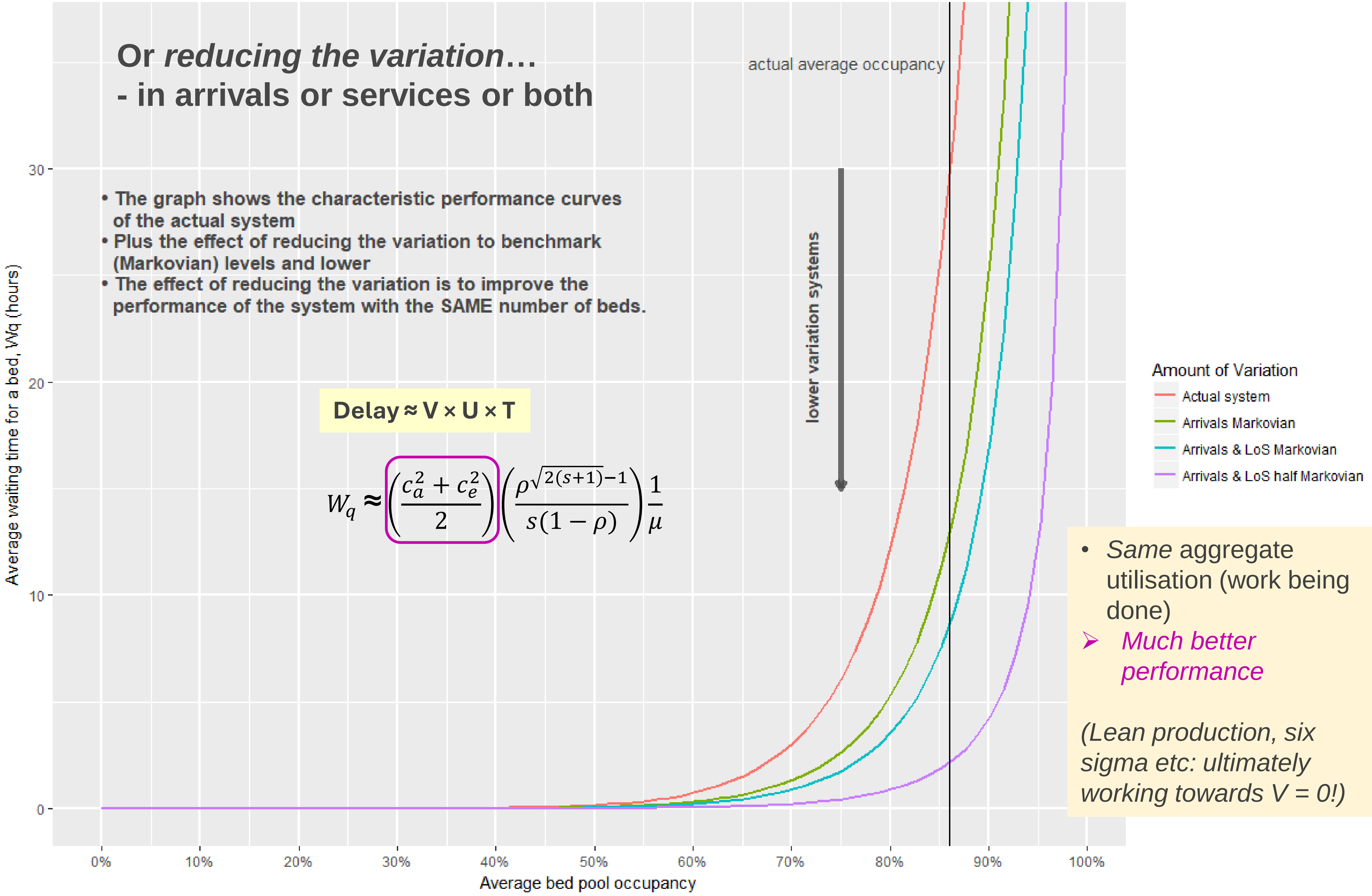
The University of Manchester
Alliance Manchester Business School

Or *reducing the variation*…
- in arrivals or services or both

- The graph shows the characteristic performance curves of the actual system
- Plus the effect of reducing the variation to benchmark (Markovian) levels and lower
- The effect of reducing the variation is to improve the performance of the system with the SAME number of beds.

Delay ≈ V × U × T

$$W_q \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho^{\sqrt{2(s+1)}-1}}{s(1-\rho)}\right)\frac{1}{\mu}$$



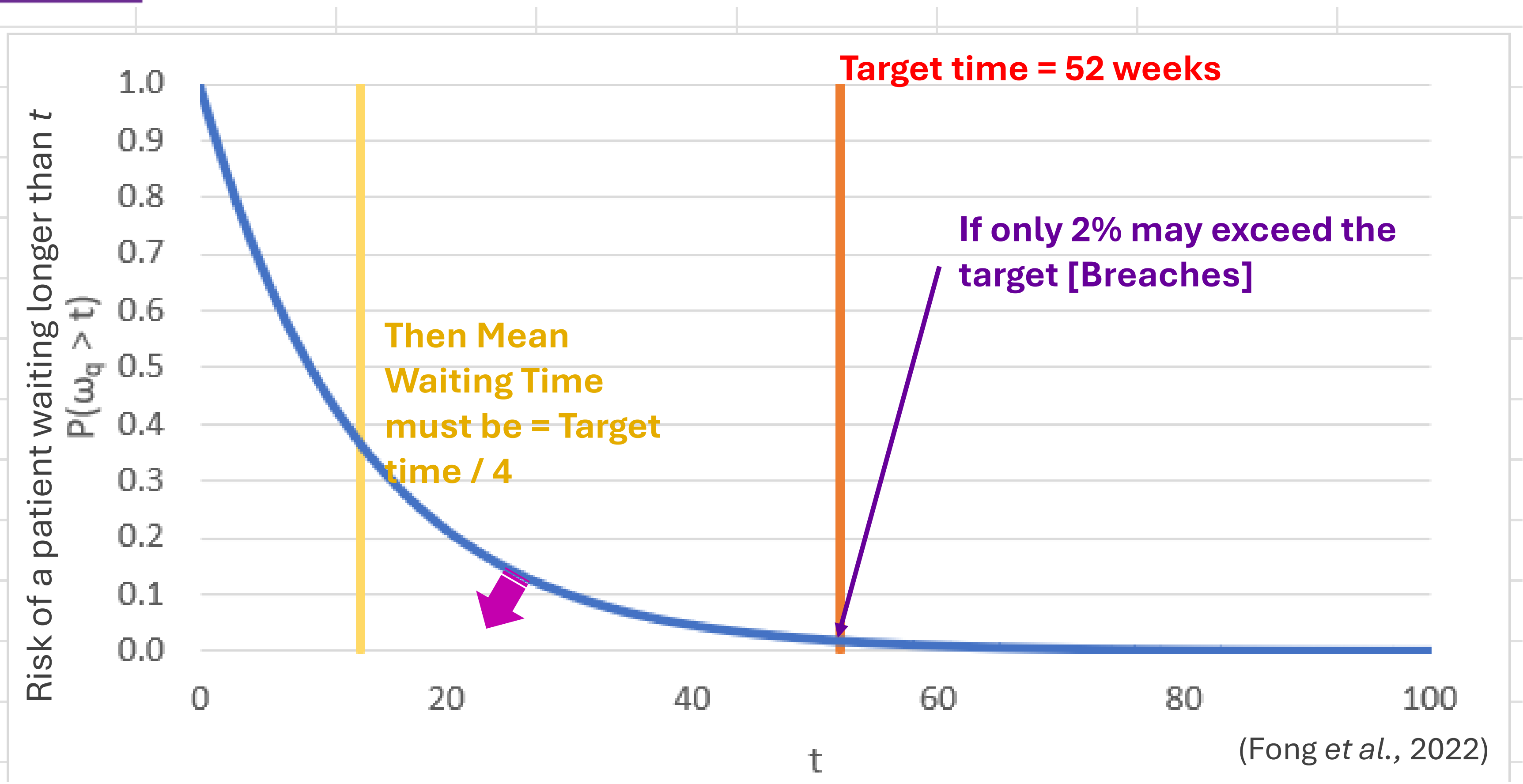
- *Same* aggregate utilisation (work being done)
- ➢ *Much better performance*

(Lean production, six sigma etc: ultimately working towards V = 0!)

See Proudlove (2020)

The University of Manchester Alliance Manchester

More insights from Queuing Systems



- The **design of the system** and the **variation** make system performance highly non-linear
  - In particular, the long and fat tails
- Meaning low risk of poor performance [low breaches, trolley waits etc] requires very much better **average** performance
  - So lower utilisation → more resource
  - (and) or **improve the design of the system** and/or **reduce the variation**!

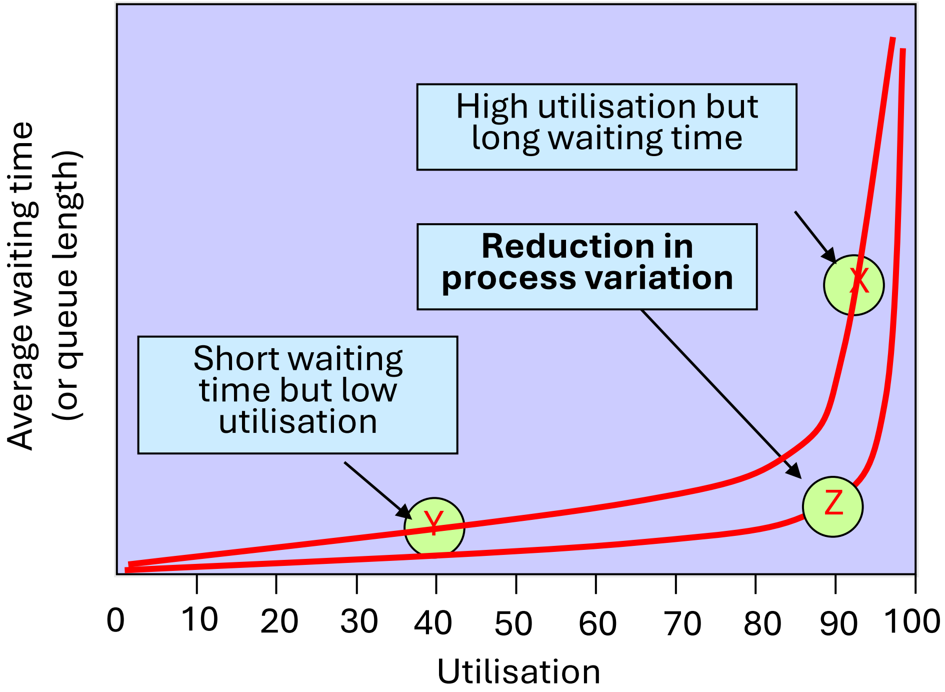
MANCHESTER 1824
The University of Manchester
Alliance Manchester Business School

Take aways

- Characteristics of system drive performance
  - "85% occupancy" does not fit all environments
  - Mean occupancy levels should be a consequence of the demand characteristics you need to absorb
    - 'empty' capacity protects the system
  - e.g., sensible utilisation levels for knee surgery vs. ICU

- Queuing theory models give quick, first-cut results
  - Make *a lot* of assumptions…
  - But give good **insights** – e.g. the VUT relationship
  - Beyond that is simulation (laborious, data-hungry, requires specialist knowledge and software)

- Can *shift* the trade-off
  - **Bed pooling**, but
    - Pooling *vs* carve-out *or* segmentation?
    - Behavioural impacts?!
  - **Reducing variation**
    - How?!



(Slack et al., 2011)

The University of Manchester
Alliance Manchester Business School

References

Bagust A, Place M and Posnett J (1999). "Dynamics of bed use in accommodating emergency admissions: stochastic simulation model". *British Medical Journal* 319, 155-158. doi: 10.1136/bmj.319.7203.155

Fong K, Mushtaq Y, House T, Gordon D, Chen Y, Griffths D, Ahmad S and Walton N (2022). "Understanding waiting lists pressures". *medRxiv*, 2022.08.23.22279117. doi: 10.1101/2022.08.23.22279117

Proudlove NC (2020). "The 85% bed occupancy fallacy: The use, misuse and insights of queuing theory". *Health Services Management Research* 33:3, 110-121. doi: 10.1177/0951484819870936

Proudlove NC (2023). "Chapter 11: Use and misuse of Queueing Theory for hospital capacity decisions" in Vissers J, Elkhuizen S and Proudlove N (Eds.) *Operations Management for Healthcare* 2nd ed. Routledge: Abingdon, UK. doi: 10.4324/9781003020011

Slack N, Brandon-Jones A and Johnston R (2016). *Operations management* 8th ed. Pearson Education: Harlow.

20