Queueing Theory

M/M/1 and M/M/m Queues

Why queuing theory?

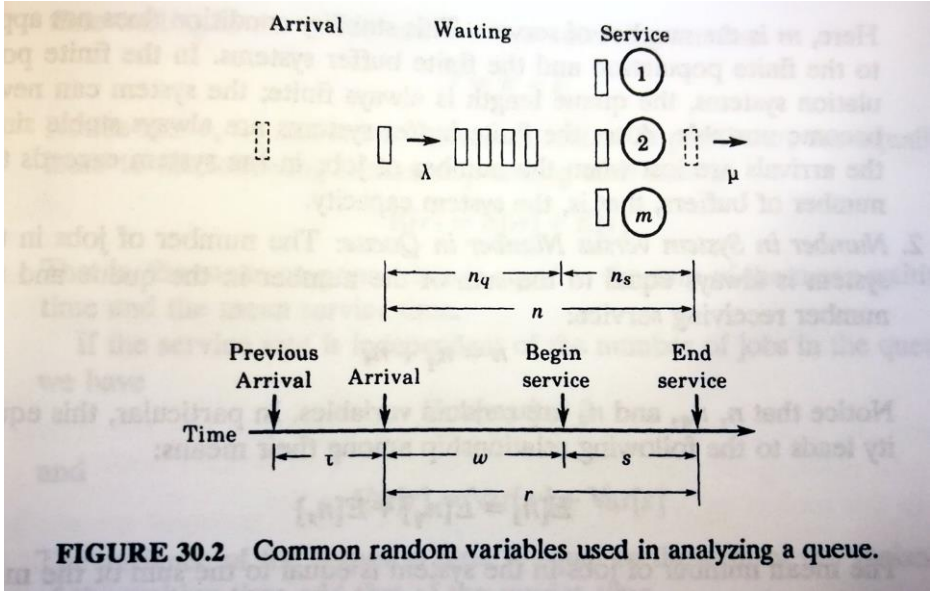


FIGURE 30.2 Common random variables used in analyzing a queue.

Art of Computer Systems Performance Analysis - p. 511

- Reason about the system
- Analyze different components
- Determine bottlenecks
- Queues are a core aspect of complex systems

Little's Law

Number of requests in the system = arrival rate * mean response time

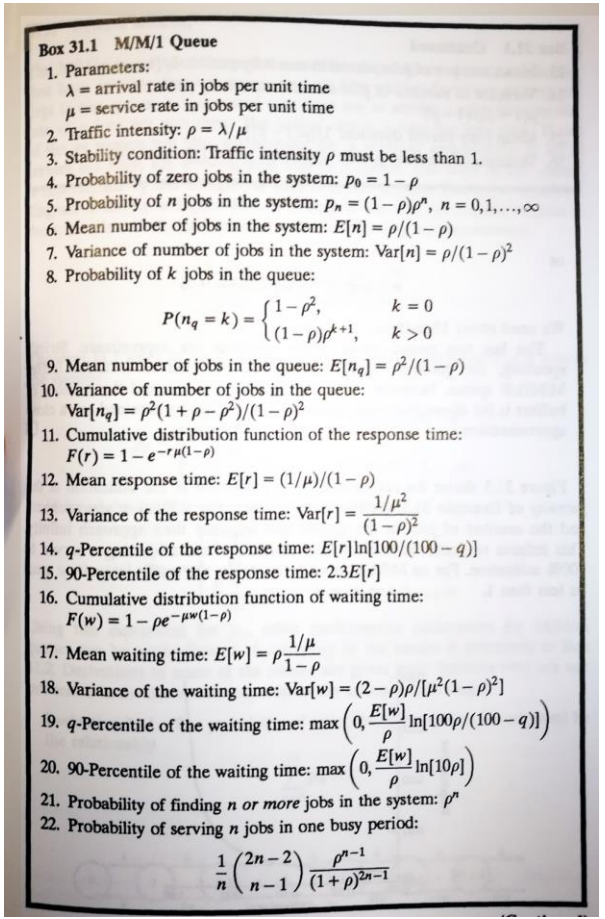
- Relates the number of requests to the response time
- Valid for any type of queueing system
- Valid for systems in its entirety or for parts of the system

Number of requests in the queue = arrival rate * mean waiting time in the queue

What you are trying to predict?

- Be clear what the input to the model is. This depends on what you have measured!
- Be clear what elements you want to predict
- Explain which aspects of your system correspond to the model and which ones do not
- Consider all the data you have gathered
  There are many options to proceed!

Input	Anything that has been measured
Goal	Arrival rate, service rate
Output	Length of the queues, number of request in the system, graphs, behaviour, ...



Box 31.1 M/M/1 Queue

1. Parameters:
   λ = arrival rate in jobs per unit time
   μ = service rate in jobs per unit time
2. Traffic intensity: $\rho = \lambda/\mu$
3. Stability condition: Traffic intensity ρ must be less than 1.
4. Probability of zero jobs in the system: $p_0 = 1 - \rho$
5. Probability of n jobs in the system: $p_n = (1-\rho)\rho^n,\ n = 0,1,\ldots,\infty$
6. Mean number of jobs in the system: $E[n] = \rho/(1-\rho)$
7. Variance of number of jobs in the system: $\mathrm{Var}[n] = \rho/(1-\rho)^2$
8. Probability of k jobs in the queue:

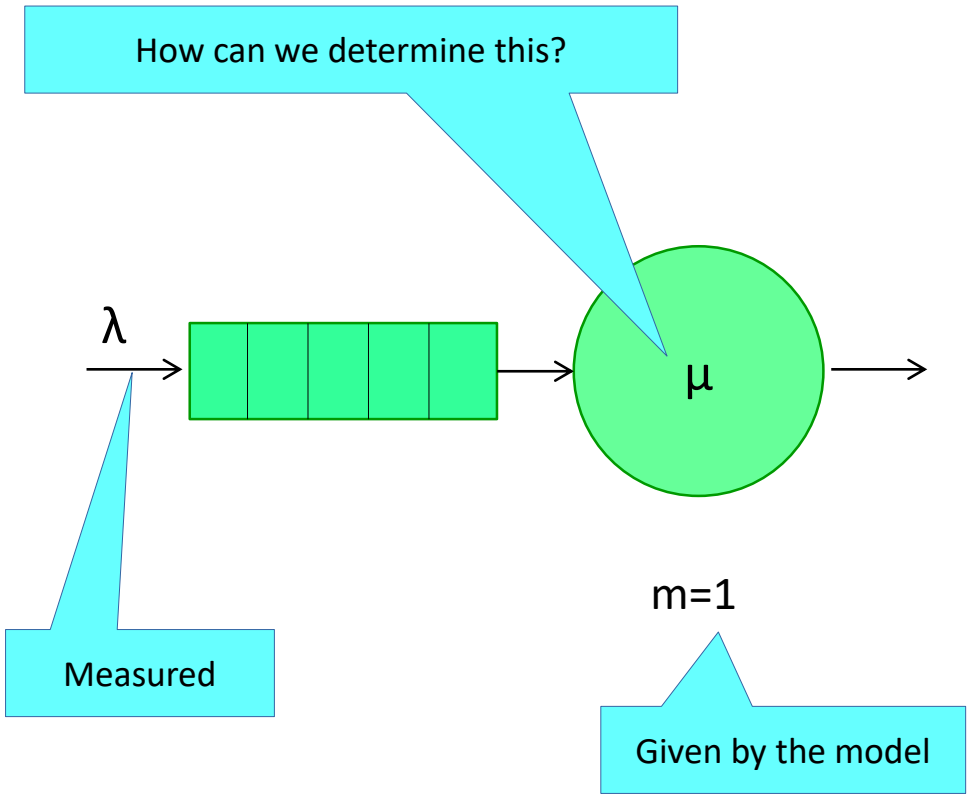
$$P(n_q = k) = \begin{cases} 1-\rho^2, & k = 0 \\ (1-\rho)\rho^{k+1}, & k > 0 \end{cases}$$

9. Mean number of jobs in the queue: $E[n_q] = \rho^2/(1-\rho)$
10. Variance of number of jobs in the queue: $\mathrm{Var}[n_q] = \rho^2(1+\rho-\rho^2)/(1-\rho)^2$
11. Cumulative distribution function of the response time: $F(r) = 1 - e^{-r\mu(1-\rho)}$
12. Mean response time: $E[r] = (1/\mu)/(1-\rho)$
13. Variance of the response time: $\mathrm{Var}[r] = \dfrac{1/\mu^2}{(1-\rho)^2}$
14. q-Percentile of the response time: $E[r]\ln[100/(100-q)]$
15. 90-Percentile of the response time: $2.3E[r]$
16. Cumulative distribution function of waiting time: $F(w) = 1 - \rho e^{-\mu w(1-\rho)}$
17. Mean waiting time: $E[w] = \rho\dfrac{1/\mu}{1-\rho}$
18. Variance of the waiting time: $\mathrm{Var}[w] = (2-\rho)\rho/[\mu^2(1-\rho)^2]$
19. q-Percentile of the waiting time: $\max\left(0, \dfrac{E[w]}{\rho}\ln[100\rho/(100-q)]\right)$
20. 90-Percentile of the waiting time: $\max\left(0, \dfrac{E[w]}{\rho}\ln[10\rho]\right)$
21. Probability of finding n *or more* jobs in the system: ρ^n
22. Probability of serving n jobs in one busy period:

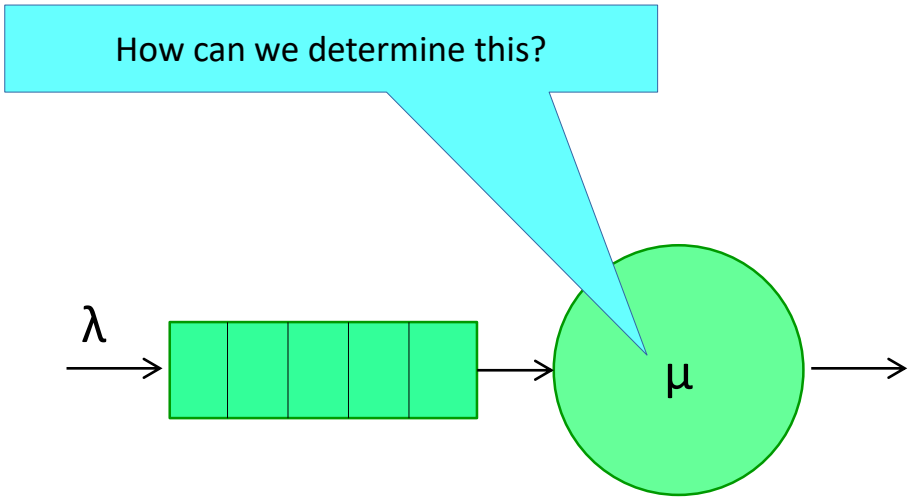
$$\frac{1}{n}\binom{2n-2}{n-1}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}$$

Art of Computer Systems Performance Analysis - p. 525

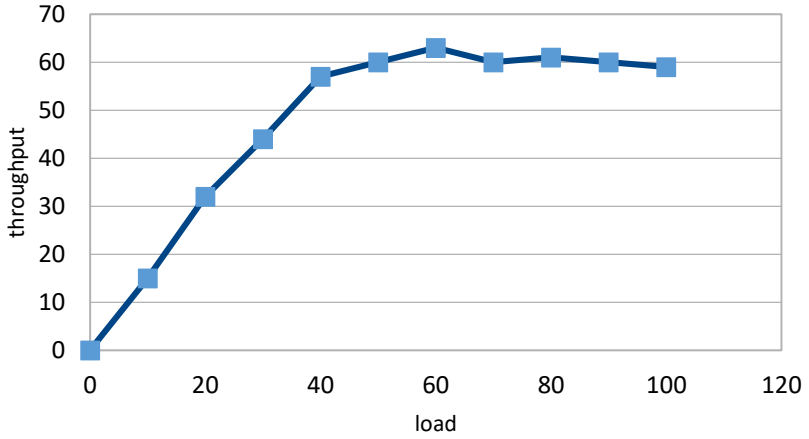
M/M/1 Queue



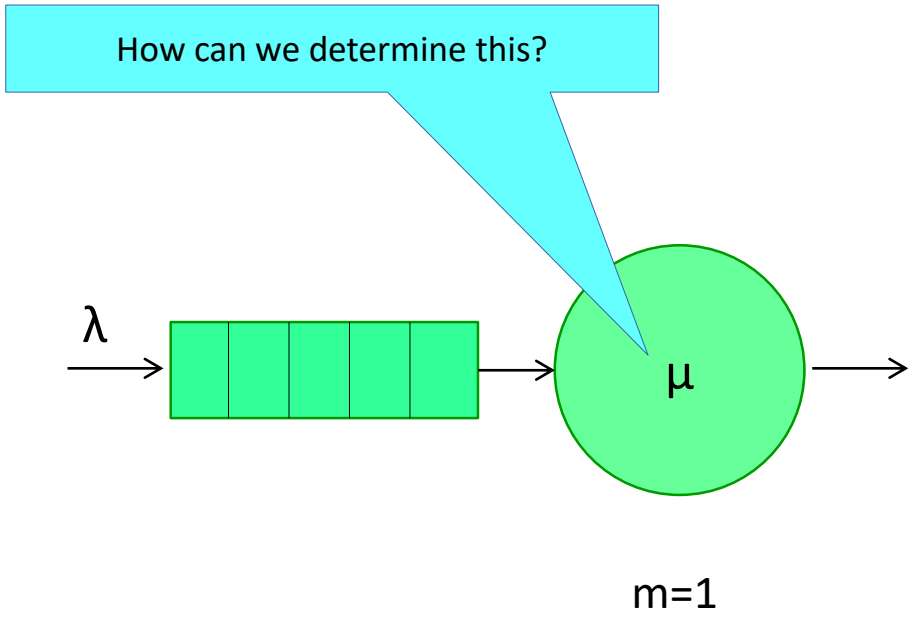
M/M/1 Queue

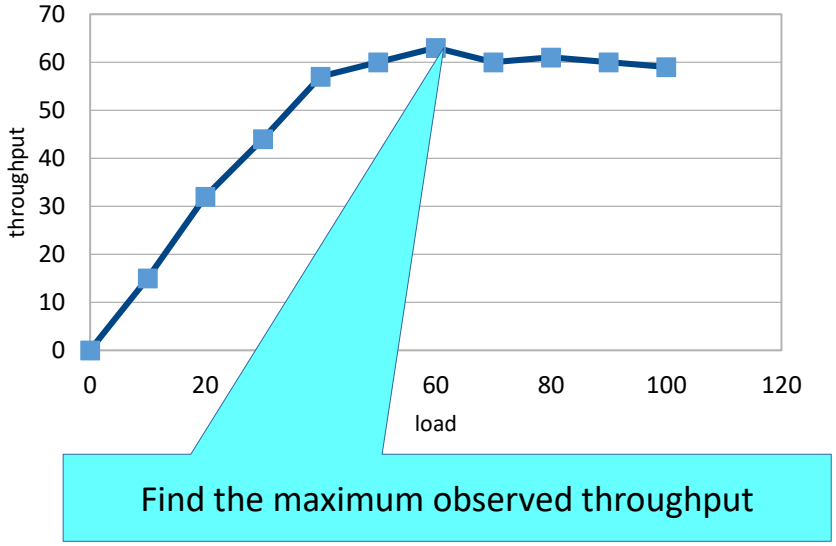

How can we determine this?
λ
μ

m=1


70
60
50
40
30
20
10
0
throughput
0
20
40
60
80
100
120
load

M/M/1 Queue



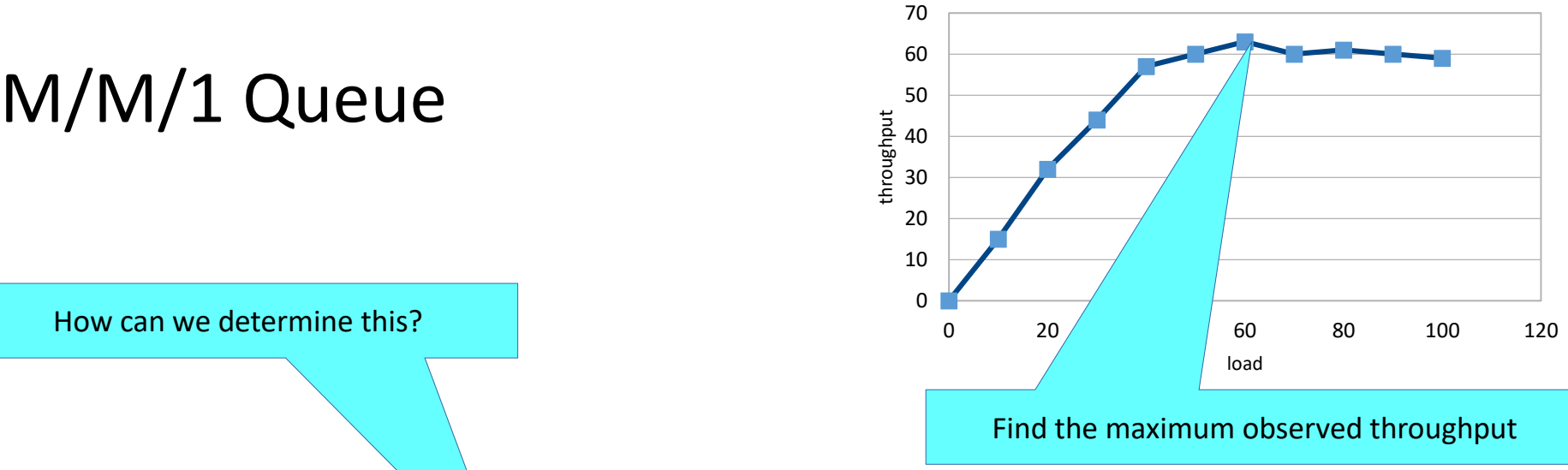


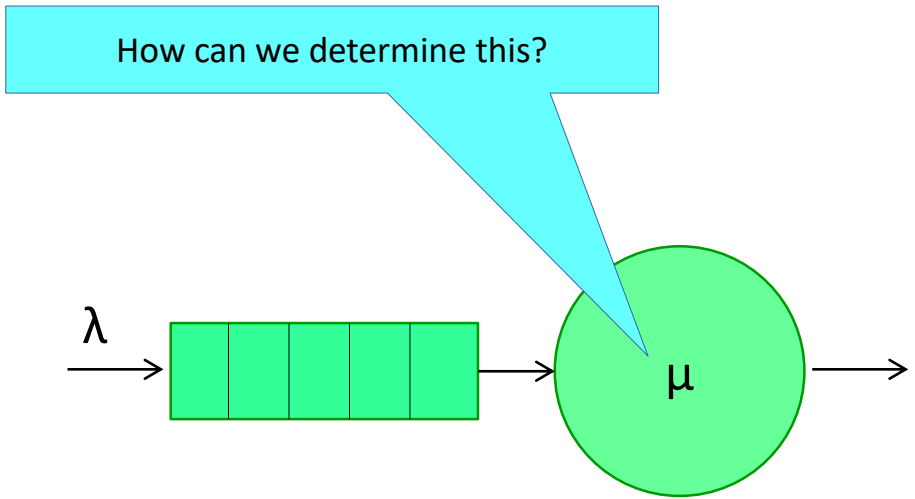
- Input to the model: λ, μ
- How to measure the service rate μ?

There are many approaches, depending what aspect of your system you want to model!

E.g., use configuration that gives maximum throughput

M/M/1 Queue





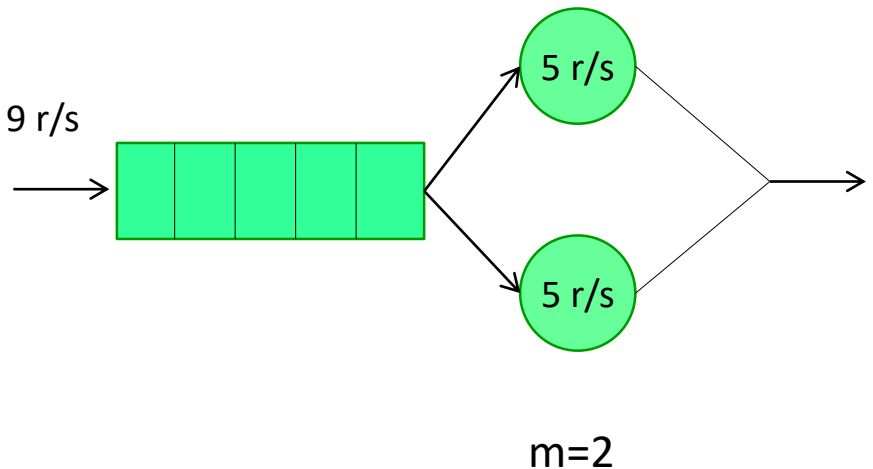
m=1

- Input to the model: λ, μ
- How to measure the service rate μ?
  - There are many approaches, depending what aspect of your system you want to model!
  - E.g., use configuration that gives maximum throughput

Service rate > arrival rate. ρ < 1.

Example of computing the output of M/M/m models

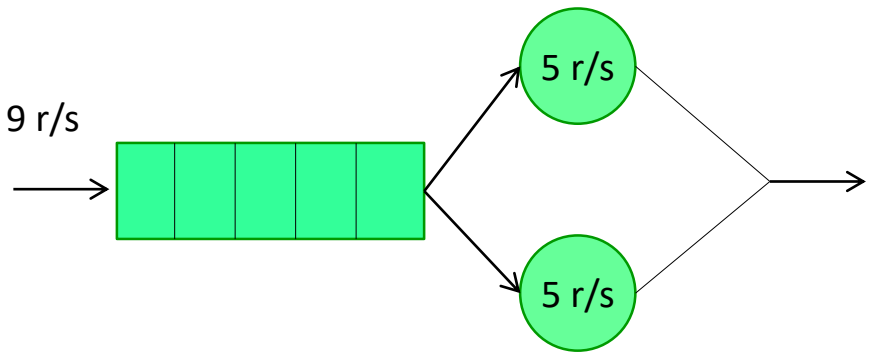
Example 1: M/M/2



Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization:
- Average service time per worker:
- Average number of requests in queue:
- Average waiting time in queue:
- Average response time:

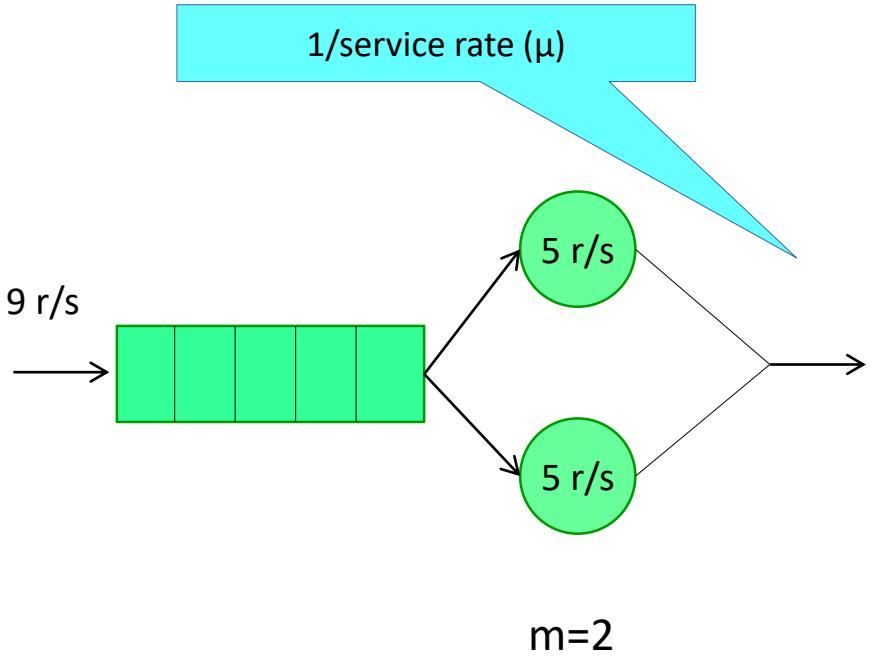
Example 1: M/M/2



Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization: ρ = λ/(m*μ) = 90%
- Average service time per worker:
- Average number of requests in queue:
- Average waiting time in queue:
- Average response time:

Example 1: M/M/2



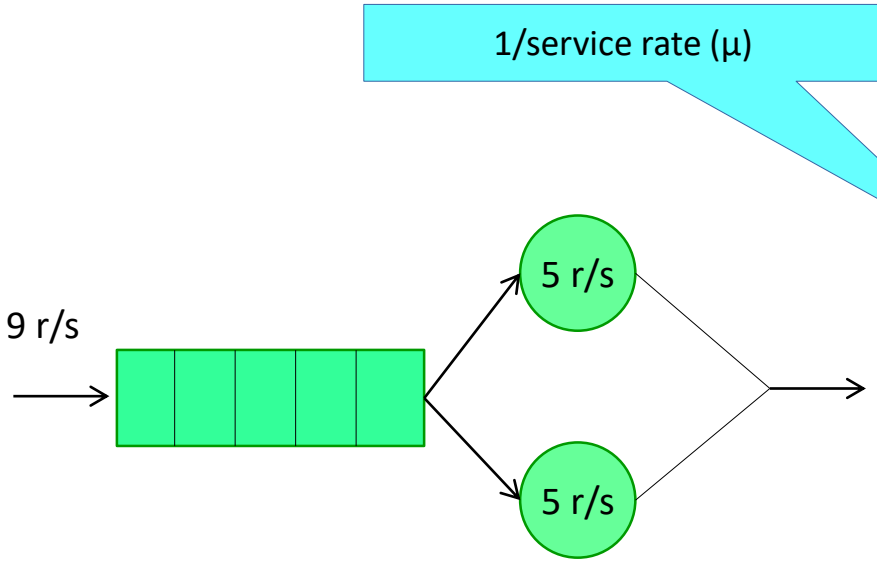
Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization: ρ = λ/(m*μ) = 90%
- Average service time per worker:

  E[s] = 0.20 s
- Average number of requests in queue:

  E[nq] = 7.674
- Average waiting time in queue:
- Average response time:

Example 1: M/M/2



Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization: ρ = λ/(m*μ) = 90%
- Average service time per worker:

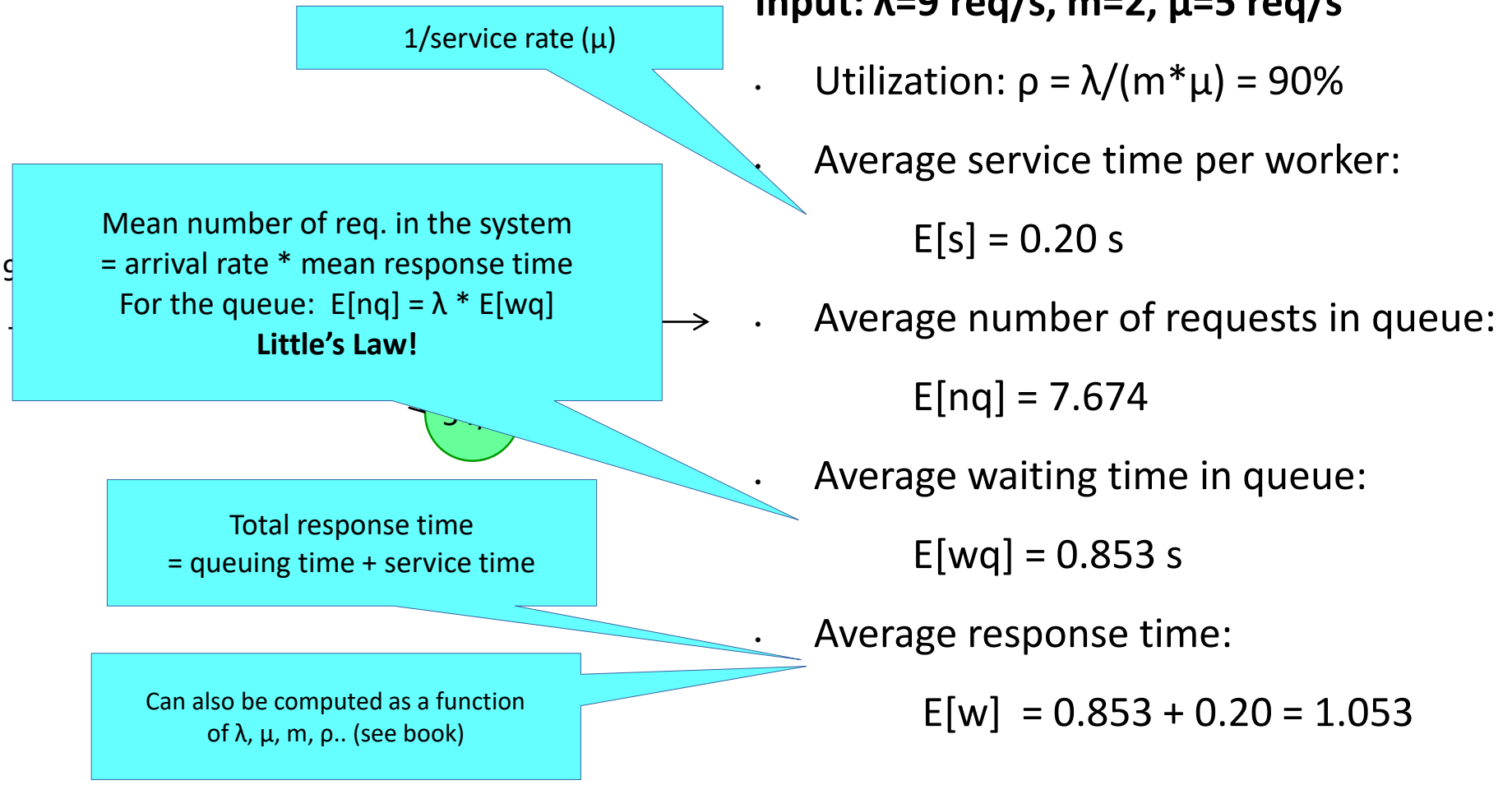
  E[s] = 0.20 s
- Average number of requests in queue:

  E[nq] = 7.674
- Average waiting time in queue:

  E[wq] = 0.853 s
- Average response time:

Mean number of req. in the system
= arrival rate * mean response time
For the queue: E[nq] = λ * E[wq]
Little's Law!

Example 1: M/M/2



Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization: $\rho = \lambda/(m*\mu) = 90\%$
- Average service time per worker:

  $E[s] = 0.20$ s
- Average number of requests in queue:

  $E[nq] = 7.674$
- Average waiting time in queue:

  $E[wq] = 0.853$ s
- Average response time:

  $E[w] = 0.853 + 0.20 = 1.053$

Example 1: M/M/2

Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization: ρ = λ/(m*μ) = 90%
- Average service time per worker:

  E[s] = 0.20 s
- Average number of requests in queue:

  E[nq] = 7.67
- Average waiting t n queue:

  E[wq
- Average re

  E[w]

1/service rate (μ)

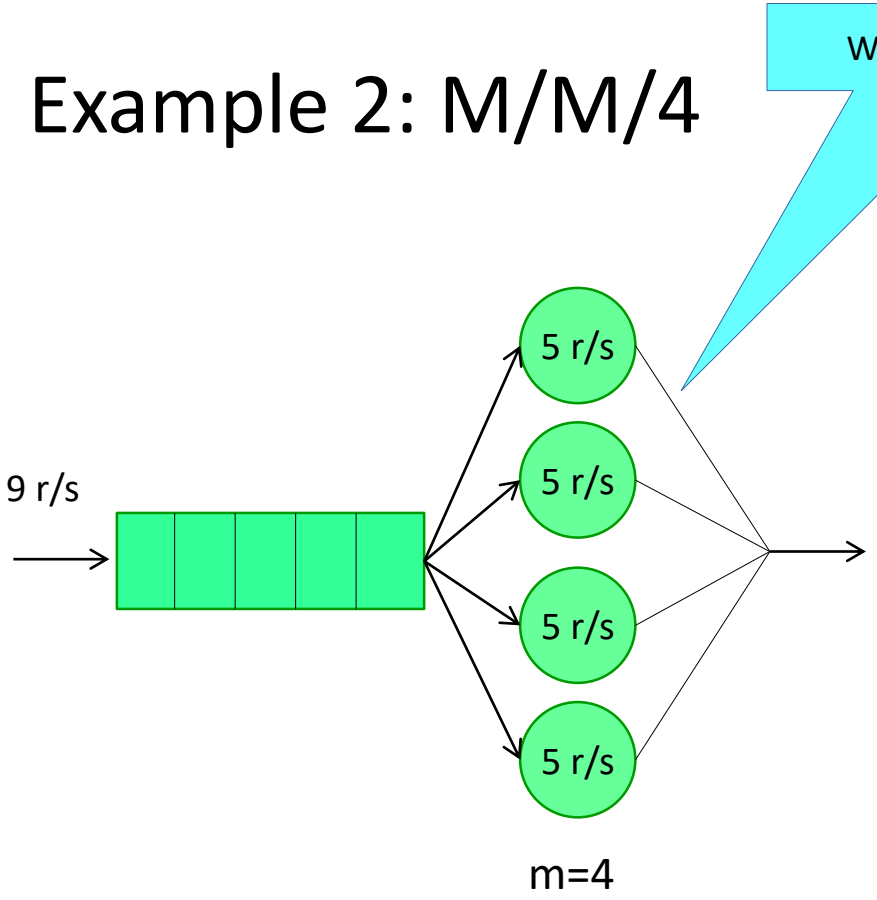
Mean number of req. in the system
= arrival rate * mean response time
For the queue: E[nq] = λ * E[wq]
Little's Law!

Total response time
= queuing time + service time

Can also be computed as a function
of λ, μ, m, ρ.. (see book)

In your report, you need to compare the output of the model to your measurements. Explain the difference!

Example 2: M/M/4

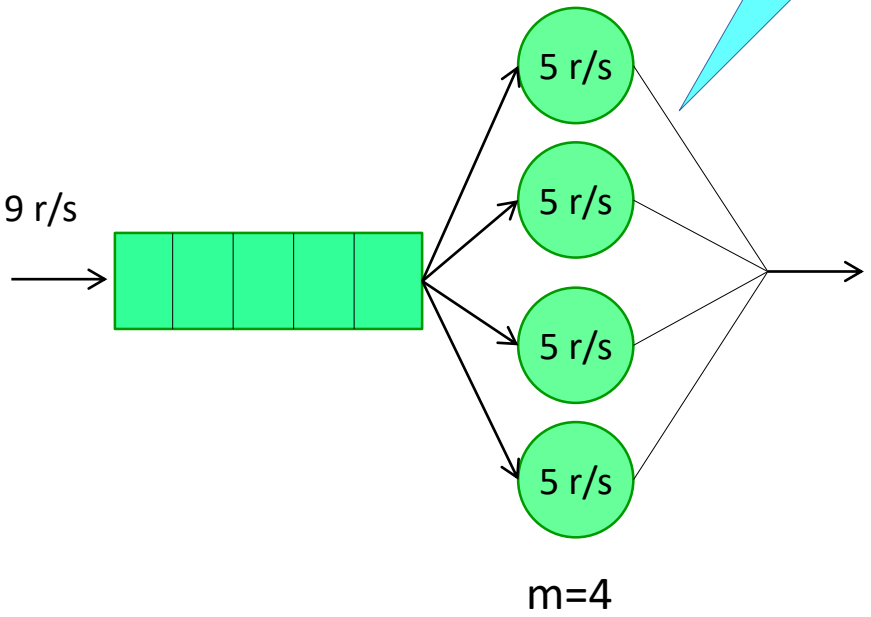


Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization:
- Average service time per worker:
- Average number of requests in queue:
- Average waiting time in queue:
- Average response time:

Example 2: M/M/4

What changes for m=4?



Input: λ=9 req/s, m=2, μ=5 req/s

- Utilization: $\rho = \lambda/(m*\mu) = 45\%$
- Average service time per worker:

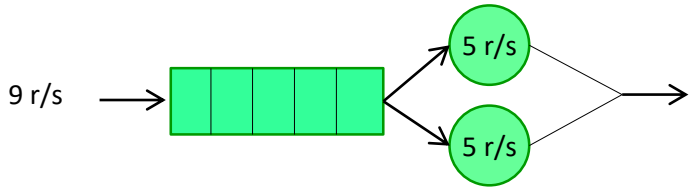
  $E[s] = 0.20$ s
- Average number of requests in queue:

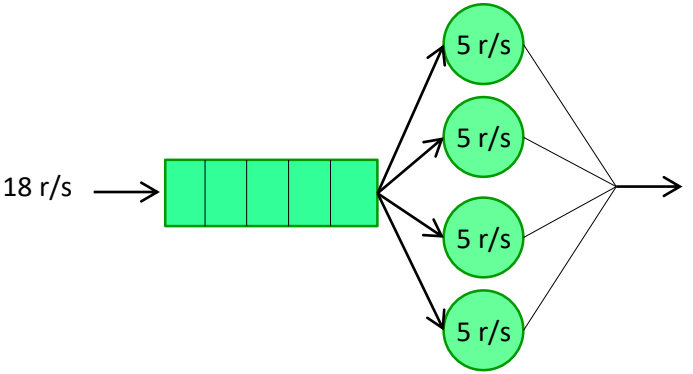
  $E[nq] = 0.105$
- Average waiting time in queue:

  $E[wq] = 0.012$ s Lower
- Average response time:

  $E[w] = 0.012 + 0.20 = 0.212$ s

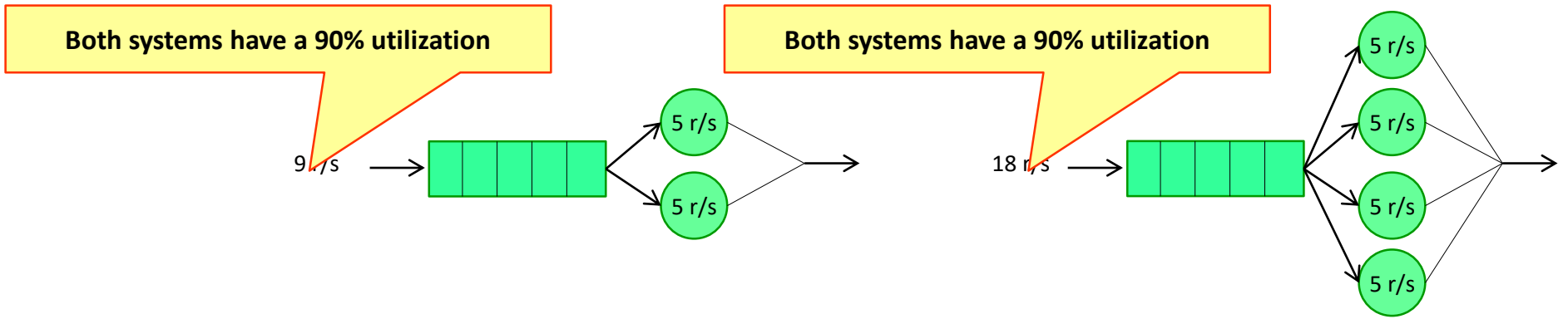
Example 3: How do these systems differ?





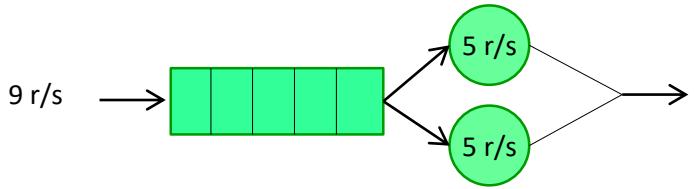
	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization		
Service time		
Queue Length		
Queuing Time		
Response Time		

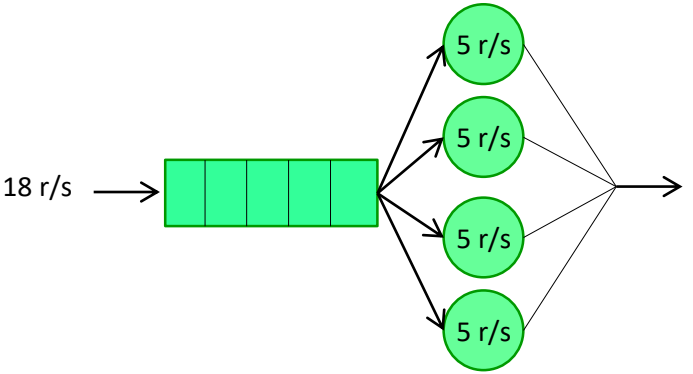
Example 3: How do these systems differ?



	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization	90%	90%
Service time		
Queue Length		
Queuing Time		
Response Time		

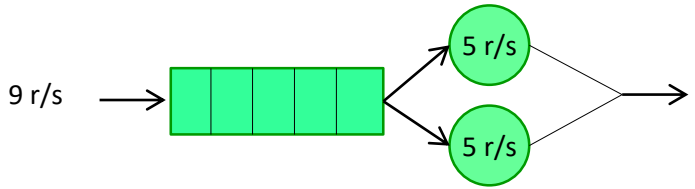
Example 3: How do these systems differ?

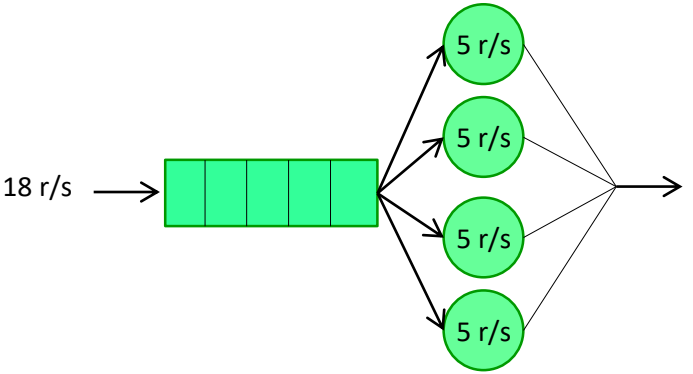




	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization	90%	90%
Service time	0.200 s	0.200 s
Queue Length		
Queuing Time		
Response Time		

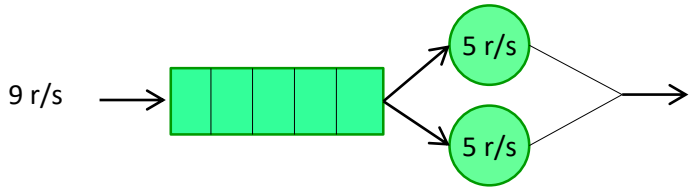
Example 3: How do these systems differ?

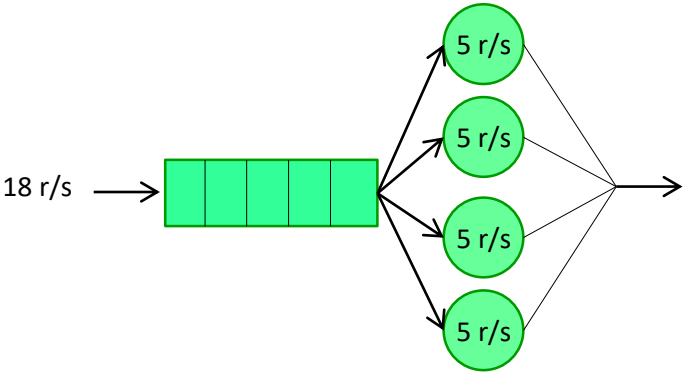




	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization	90%	90%
Service time	0.200 s	0.200 s
Queue Length	7.67 requests	7.10 requests
Queuing Time		
Response Time		

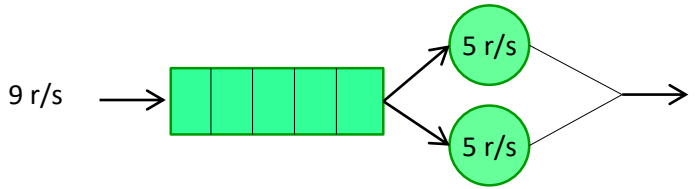
Example 3: How do these systems differ?

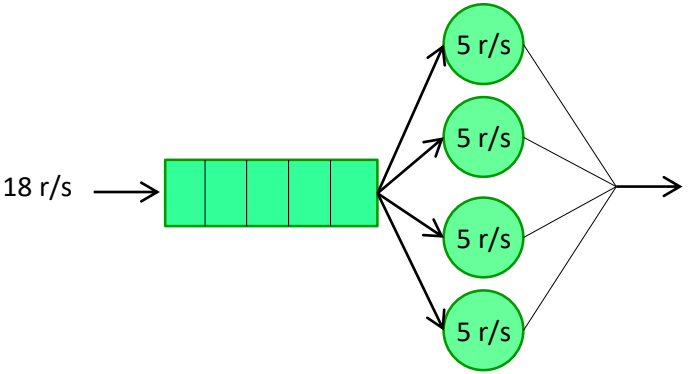




	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization	90%	90%
Service time	0.200 s	0.200 s
Queue Length	7.67 requests	7.10 requests
Queuing Time	0.853 s	0.394 s
Response Time		

Example 3: How do these systems differ?

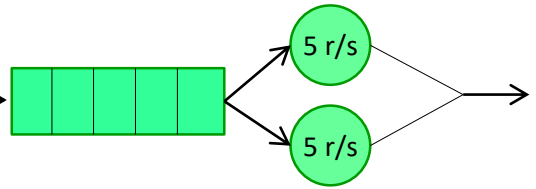


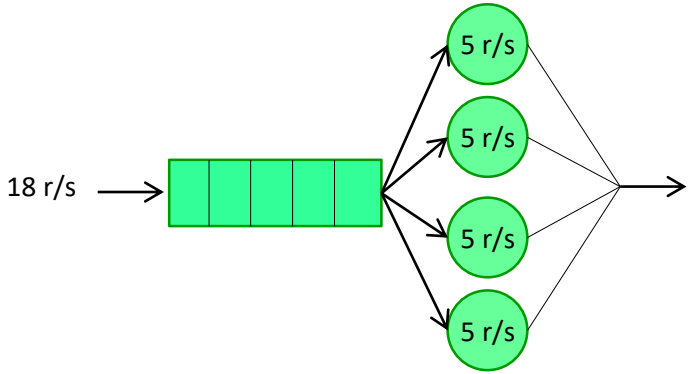


	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization	90%	90%
Service time	0.200 s	0.200 s
Queue Length	7.67 requests	7.10 requests
Queuing Time	0.853 s	0.394 s
Response Time	1.053 s	0.594 s

Example 3: How do these systems differ?

What would be better?
2x M/M/2 or 1x M/M/4

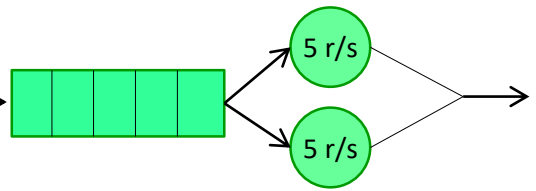


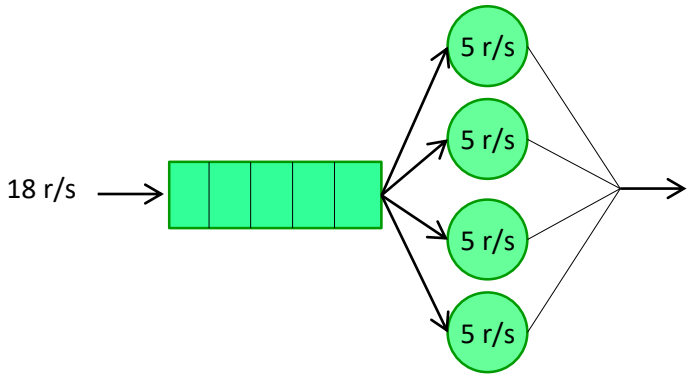


	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization	90%	90%
Service time	0.200 s	0.200 s
Queue Length	7.67 requests	7.10 requests
Queuing Time	0.853 s	0.394 s
Response Time	1.053 s	0.594 s

Example 3: How do these systems differ?

What would be better?
2x M/M/2 or 1x M/M/4
M/M/4 has less queueing!





	λ=9 req/s, m=2, μ=5 req/s	λ=18 req/s, m=4, μ=5 req/s
Utilization	90%	90%
Service time	0.200 s	0.200 s
Queue Length	7.67 requests	7.10 requests
Queuing Time	0.853 s	0.394 s
Response Time	1.053 s	0.594 s