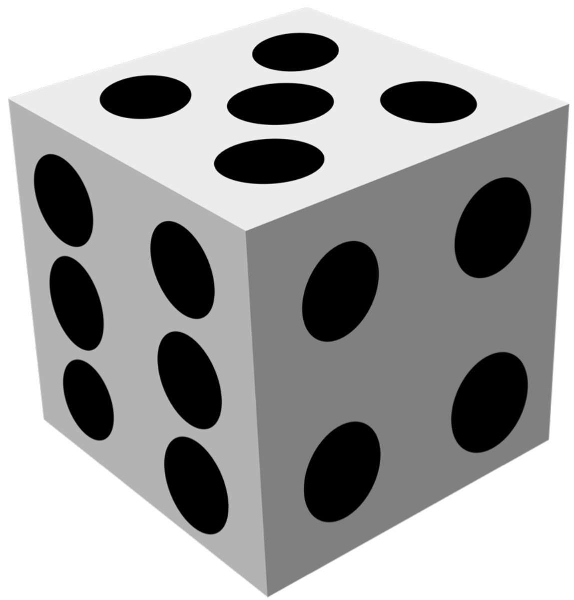
# COMP9334
# Capacity Planning for Computer Systems and Networks

## Week 3A: Queues with Poisson arrivals (2)

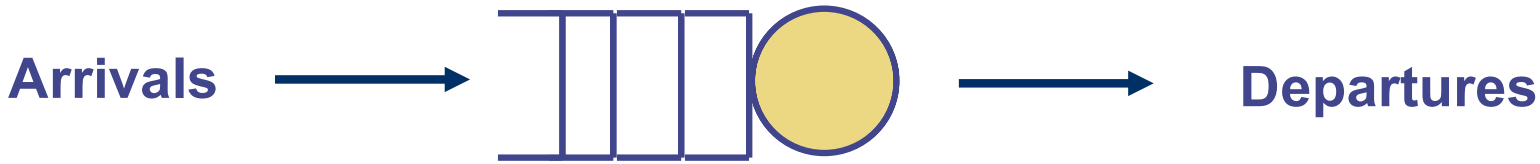
# Pre-lecture exercise

- You have a loaded die with 6 faces with values 1, 2, 3, 4, 5 and 6
- The probability that you can get each face is given in the table below
- What is the mean value that you can get?

| Value | Probability |
|---|---|
| 1 | 0.1 |
| 2 | 0.1 |
| 3 | 0.2 |
| 4 | 0.1 |
| 5 | 0.3 |
| 6 | 0.2 |

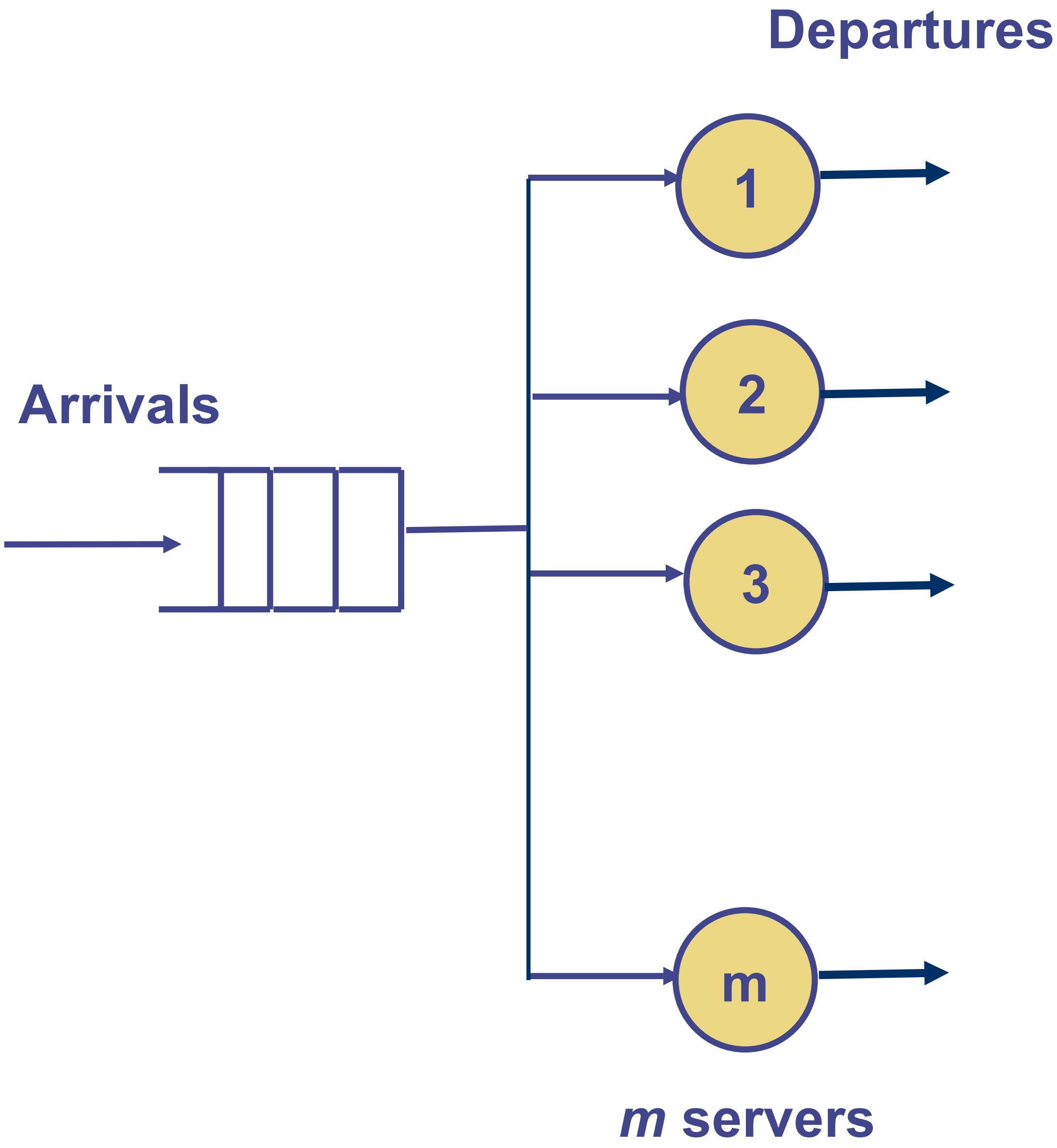


# Single-server queue




- Open, single server queues
- How to find:
  - Waiting time
  - Response time
  - Mean queue length etc.
- The technique to find waiting time etc. is called *Queueing Theory*

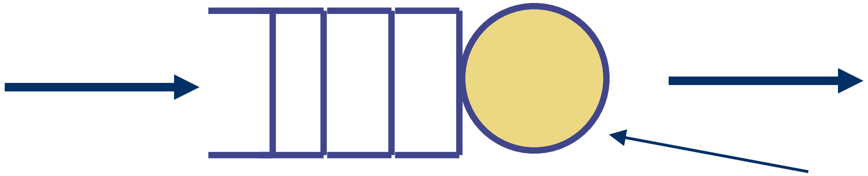
# Multiple server queue




- Open, multi-server queue
- How to find:
  - Waiting time
  - Response time
  - Mean queue length etc.

# What will you be able to do with the results?

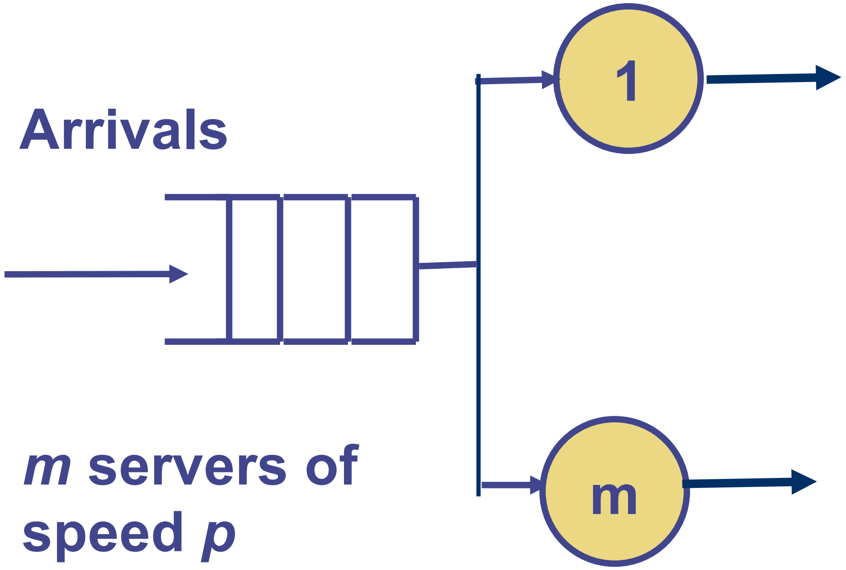
**Configuration 1:**



**processing speed =** ***m p***

**Configuration 2:**

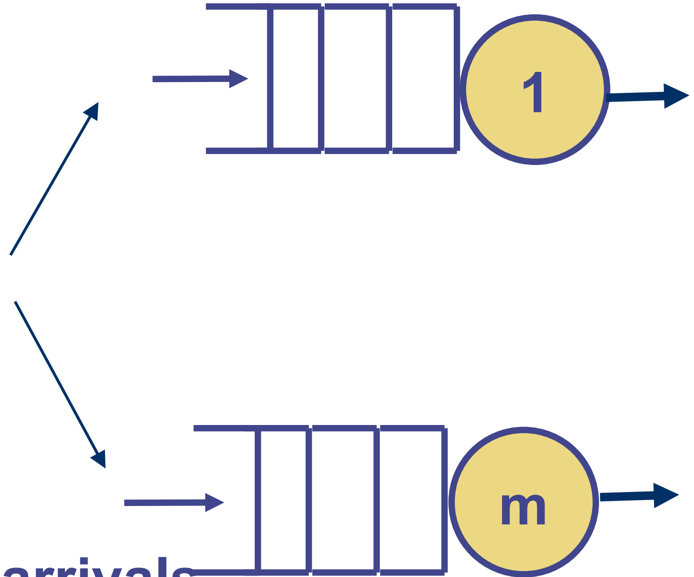
**Arrivals**




***m*** **servers of speed** ***p***

**Configuration 3:**

**Arrivals**

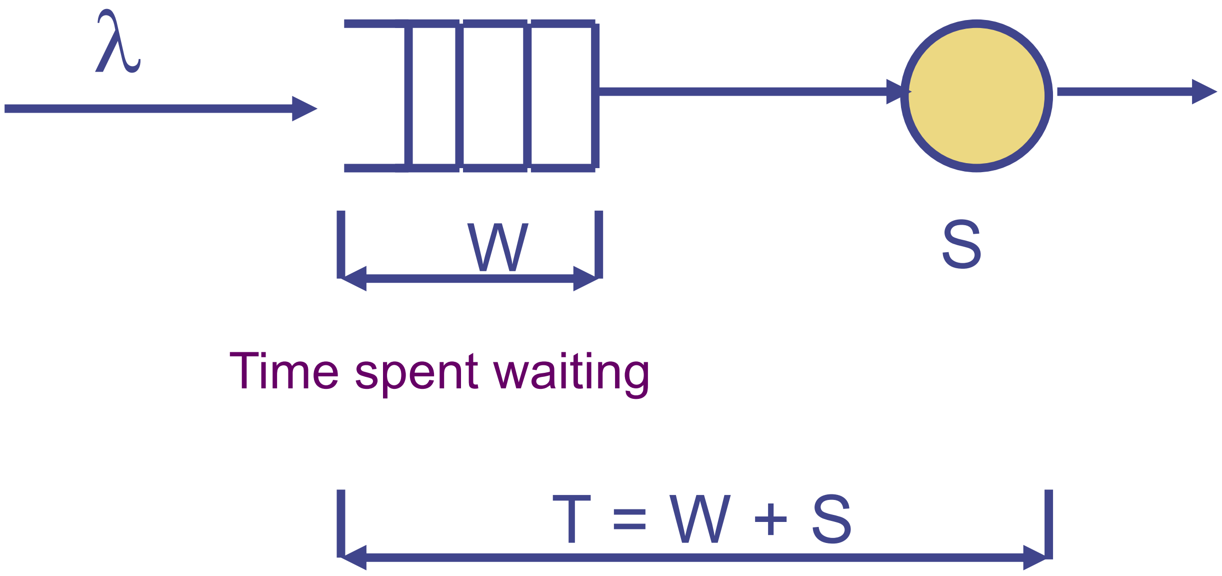



**Randomly assign the arrivals into** ***m*** **queues**

***m*** **servers of speed** ***p***

**Which configuration has the best response time?**

# Single Server Queue: Terminology




Response Time T
= Waiting time W + Service time S

Note: We use T for response time because this is the notation in many queueing theory books. For a similar reason, we will use $\rho$ for utilisation rather than U.

# Call centre analogy from Week 2B

- Consider a call centre
  - Calls are arriving according to Poisson distribution with rate $\lambda$
  - The length of each call is exponentially distributed with parameter $\mu$
    - Mean length of a call is $1/\mu$

**Call centre:**

**Arrivals**

**$m$ operators**
**If all operators are busy, the centre can put at most $n$ additional calls on hold.**
**If a call arrives when all operators and holding slots are used, the call is rejected.**

- We solved the problems for
  - ($m = 1$ and $n = 0$), and ($m = 1$ and $n = 1$)
- How about other values of $m$ and $n$? What about response time?

# Kendall's notation

- To represent different types of queues, queueing theorists use the Kendall's notation
- The call centre example on the previous page can be represented as:

$$M \,/\, M \,/\, s \;(/\, B)$$

Inter-arrival distribution is Markovian i.e. Exponential

Service time distribution is Markovian i.e. Exponential

Number of servers

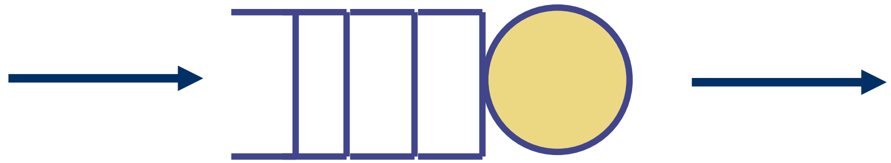
Buffer Positions (wait room)

The call centre example on the last page is a M/M/m/(m+n) queue

If n = ∞, we simply write M/M/m

# M/M/1 queue

Exponential
Inter-arrivals ($\lambda$)



Exponential
Service time ($\mu$)

Infinite buffer      One server

- Consider a call centre analogy
  - Calls are arriving according to Poisson distribution with rate $\lambda$
  - The length of each call is exponentially distributed with parameter $\mu$
    - Mean length of a call is $1/\mu$

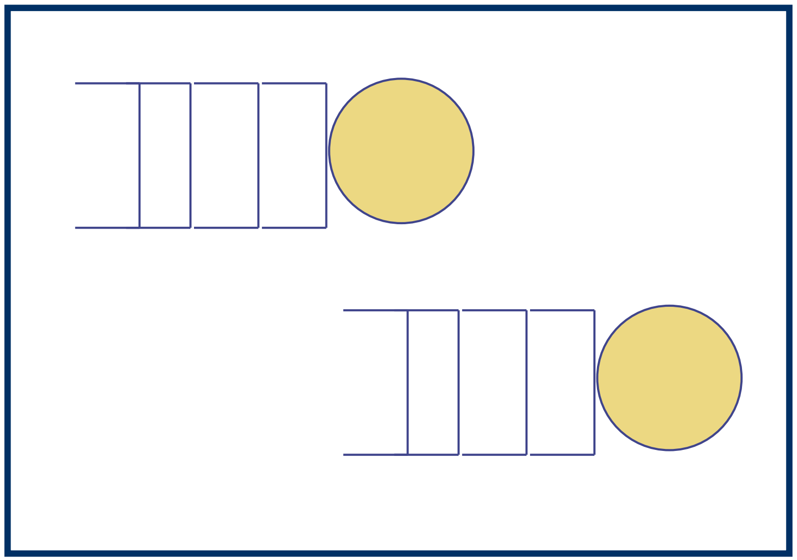
**Arrivals** →

**Call centre with *1* operator**
**If the operator is busy, the centre will put the call on hold.**
**A customer will wait until his call is answered.**

- Queueing theory will be able to answer these questions:
  - What are the mean waiting time, mean response time for a call?

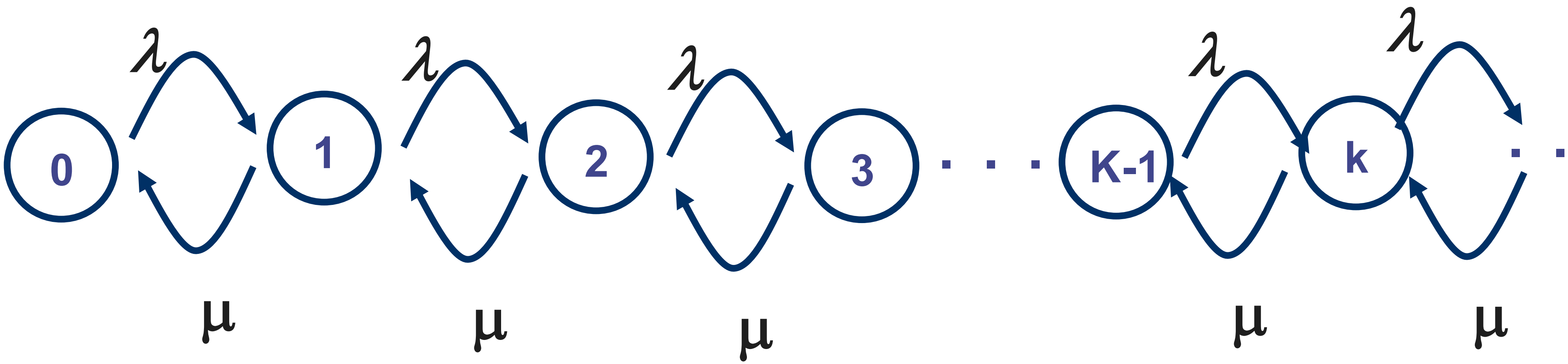
# Little's Law

- Applicable to any "box" that contains some queues or servers
- Mean number of jobs in the "box" =
  Mean response time x Throughput
- We will use Little's Law in this lecture to derive the mean response time
  - We first compute the mean number of jobs in the "box" and throughput



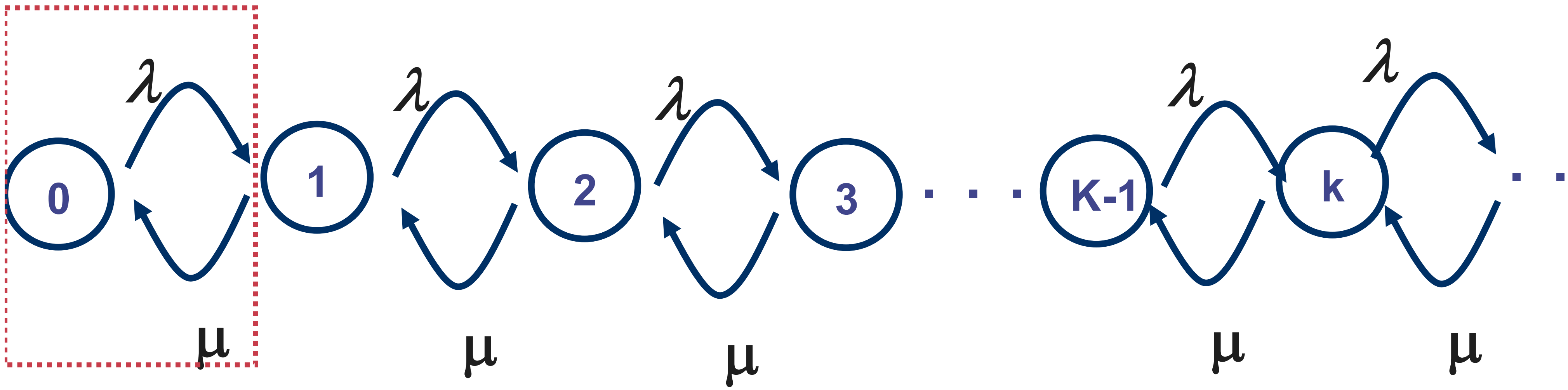
# M/M/1: State and transition diagram

- We will solve for the steady state response
- Define the states of the queue
  - State 0 = There is zero job in the system (= The server is idle)
  - State 1 = There is 1 job in the system (= 1 job at the server, no job queueing)
  - State 2 = There are 2 jobs in the system (= 1 job at the server, 1 job queueing)
  - State *k* = There are *k* jobs in the system (= 1 job at the server, *k-1* job queueing)
- The state transition diagram



# M/M/1 state balance:

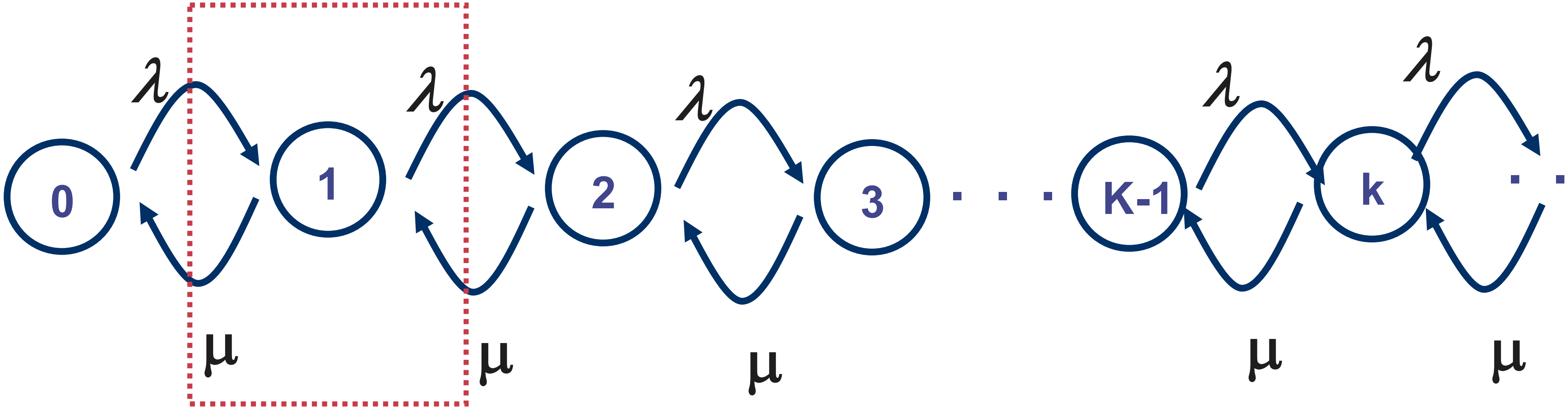
$$P_k = \text{Prob. } k \text{ jobs in system}$$




$$\lambda P_0 = \mu P_1$$

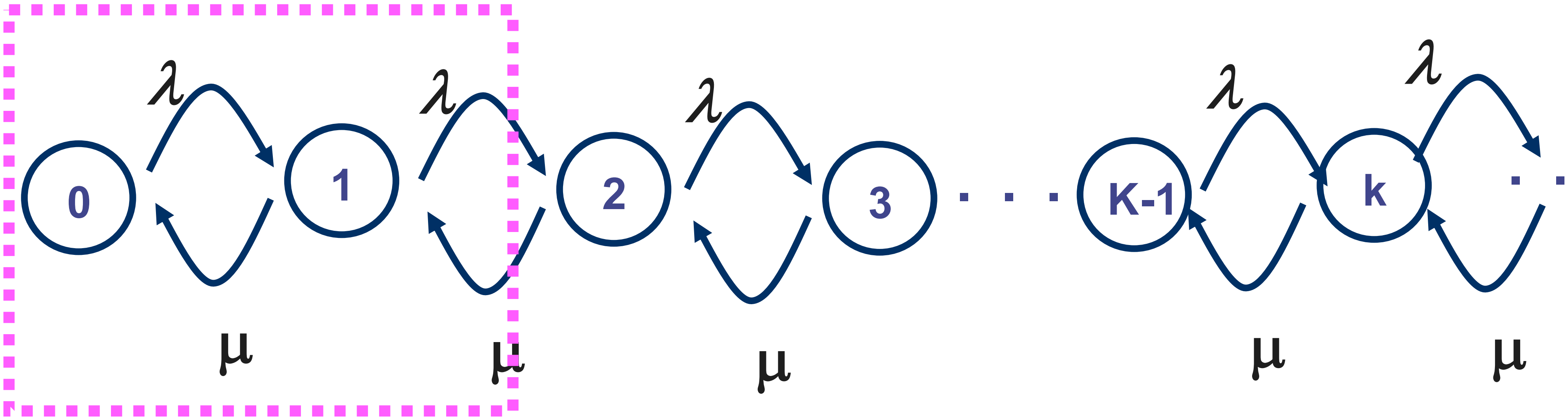
$$\Rightarrow P_1 = \frac{\lambda}{\mu} P_0$$

# M/M/1 state balance: Exercise 1




- Exercise: Write the state balance equation for State 1

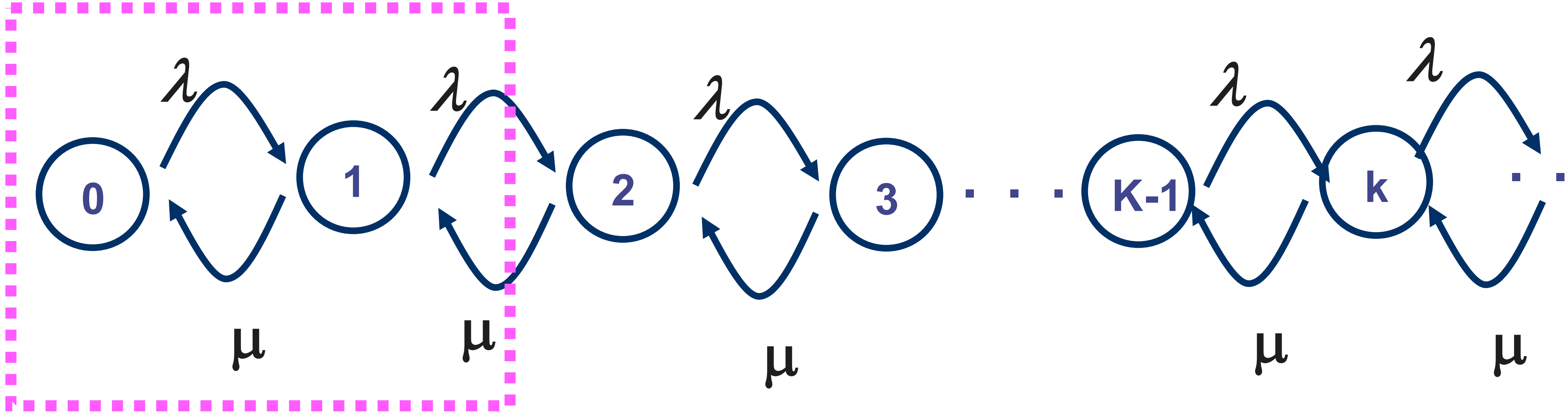
# M/M/1 state balance: Exercise 2



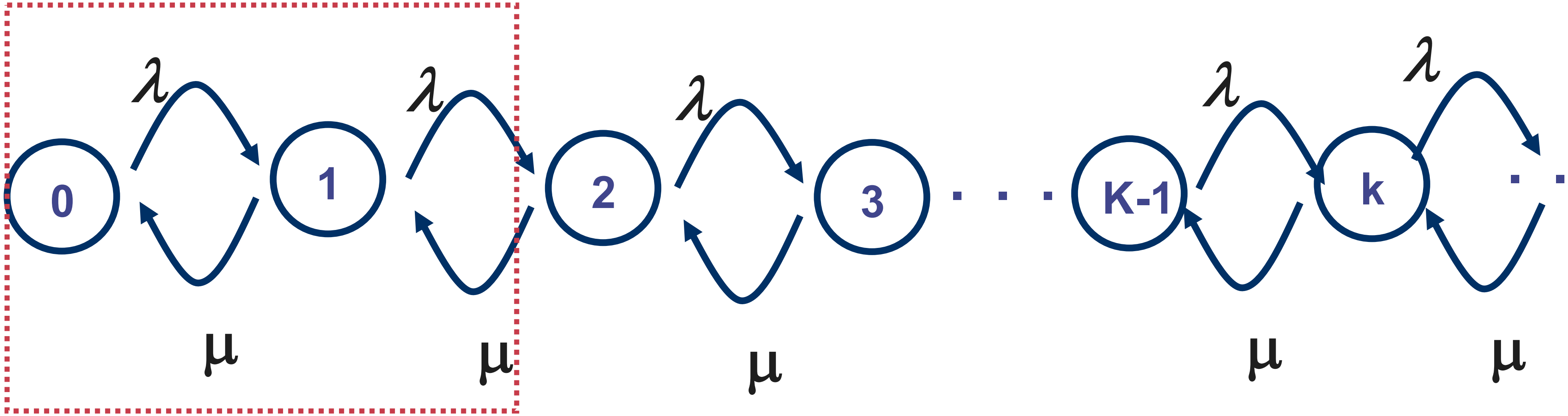

- Exercise: Write the state balance equation for magenta box, i.e.

Rate of transiting out of the magenta box
= Rate of transiting into the magenta box

# Which state balance is easier to work with?



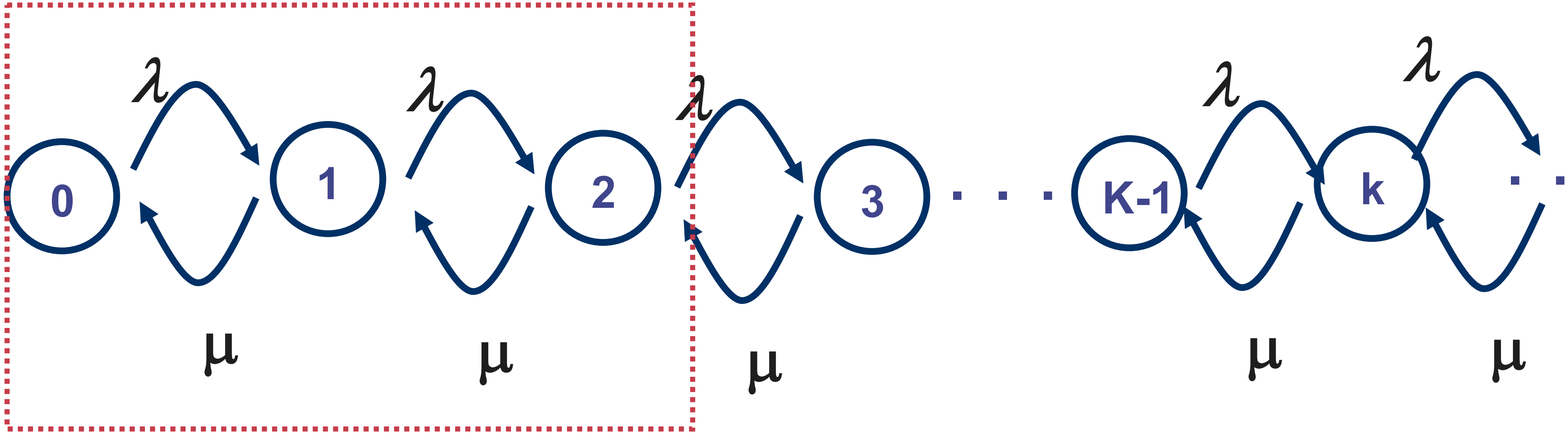
# M/M/1 state balance: Relating $P_2$ and $P_0$




$$\lambda P_0 = \mu P_1 \qquad \lambda P_1 = \mu P_2$$

$$\Rightarrow P_2 = \frac{\lambda}{\mu} P_1 \quad \Rightarrow P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_0$$

# M/M/1 state balance: Relating $P_3$ and $P_0$




$$\lambda P_2 = \mu P_3$$

$$\Rightarrow P_3 = \frac{\lambda}{\mu} P_2 \quad \Rightarrow P_3 = \left(\frac{\lambda}{\mu}\right)^3 P_0$$

# M/M/1 state balance: Relating $P_k$ and $P_0$

**In general** $$P_k = \left(\frac{\lambda}{\mu}\right)^k P_0$$

$$\text{Let } \rho = \frac{\lambda}{\mu}$$

**We have** $$P_k = \rho^k P_0$$

# Solving for $P_k$

**With** $P_k = \rho^k P_0$ **and**

$$P_0 + P_1 + P_2 + P_3 + \ldots = 1$$

$$\Rightarrow (1 + \rho + \rho^2 + \ldots)P_0 = 1$$

$$\Rightarrow P_0 = 1 - \rho \text{ if } \rho < 1$$

$$\Rightarrow P_k = (1 - \rho)\rho^k$$

**$\rho$ = utilisation**
**= Prob server is busy**
**= 1 - $P_0$**
**= 1- Prob server is idle**

**Since** $\rho = \dfrac{\lambda}{\mu}$ $, \rho < 1 \Rightarrow \lambda < \mu$

Arrival rate < service rate

# Exercise: Mean number of jobs

Recall that $P_k = \text{Prob. } k \text{ jobs in system}$

and we have calculated that $P_k = (1-\rho)\rho^k$

Determine the mean number of jobs in the system.

Hint 1: Look at pre-lecture exercise 1.

You can use the following formula to help you.

For $0 \le x < 1$,

$$p + x(p+q) + x^2(p+2q) + x^3(p+3q) + \ldots = \frac{p}{1-x} + \frac{xq}{(1-x)^2}$$

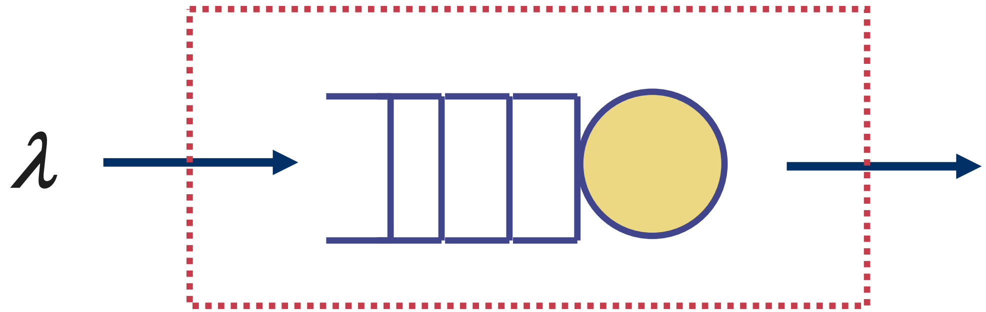
# Mean number of jobs

$$P_k = \text{Prob. } k \text{ jobs in system}$$

$$P_k = (1-\rho)\rho^k$$

The mean number of jobs in the system =

# M/M/1: mean response time

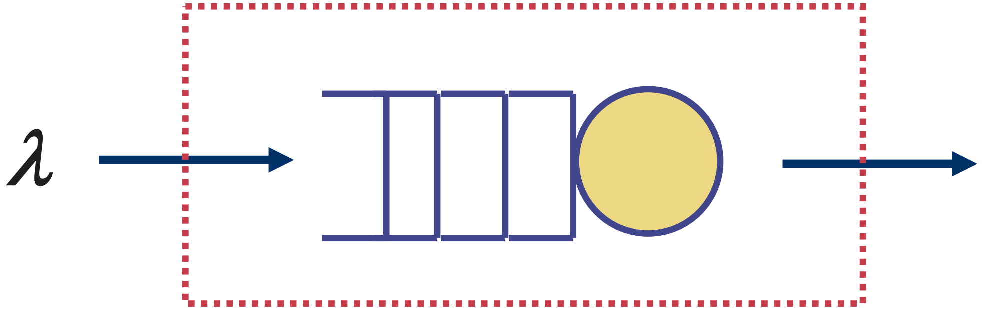



**Little's law:**
**mean number of customers = throughput x response time**

**Throughput is** $\lambda$ ***(why?)***

$$\text{Response time } T = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}$$

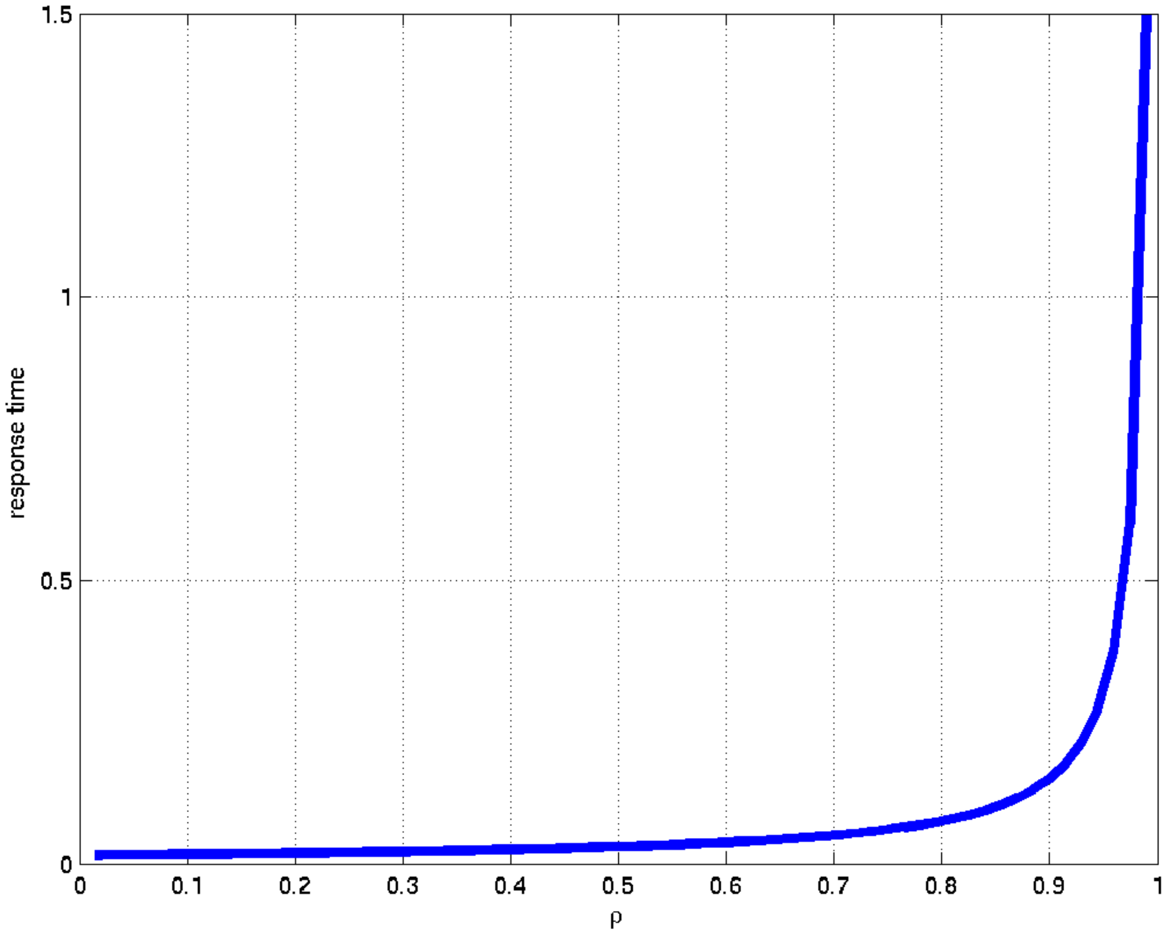
# Exercise: M/M/1 mean waiting time




**What is the mean waiting time at the queue?**

Using the service time parameter ($1/\mu$ = 15ms) in the example, let us see how response time T varies with $\lambda$

$$T = \frac{1}{\mu(1-\rho)}$$




Observation: Response time increases sharply when $\rho$ gets close to 1

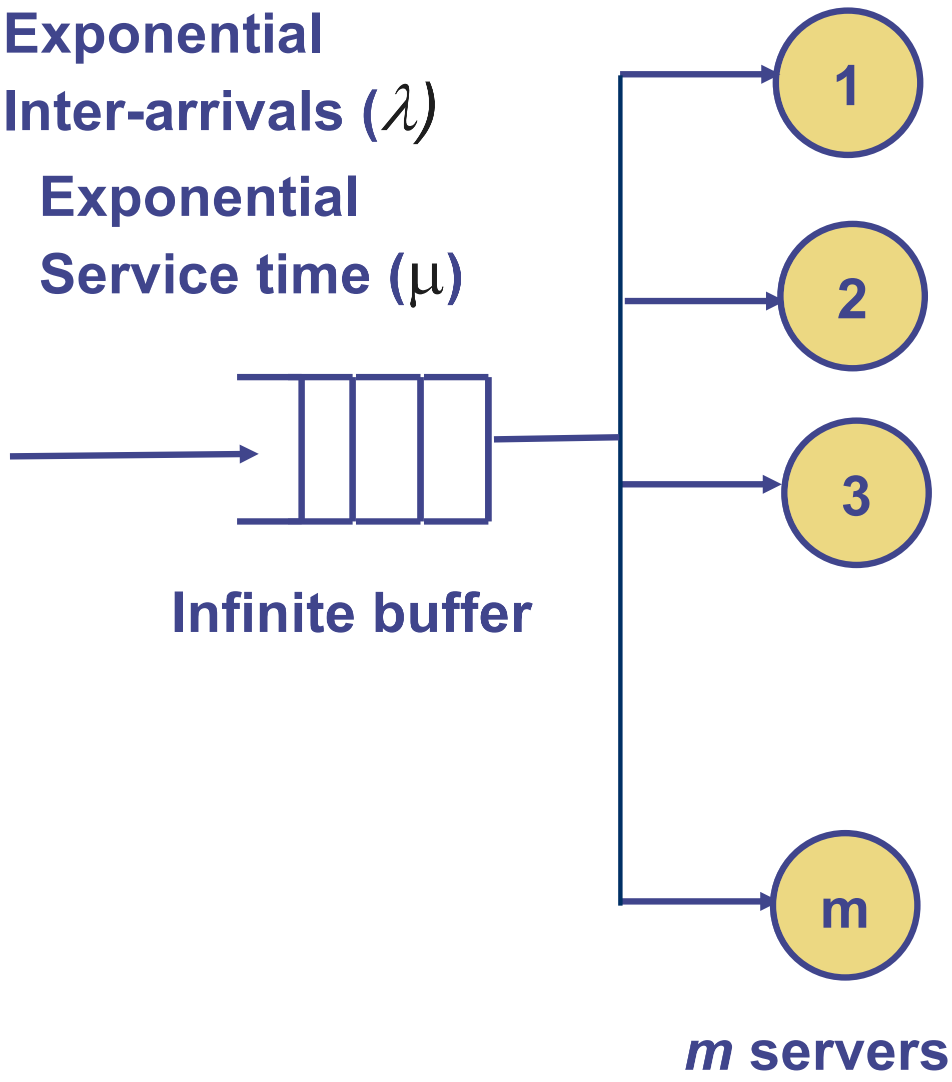
Infinite queue assumption means $\rho \to 1$, $T \to \infty$

# Non-linear effect on response time

- The response time of an M/M/1 queue $= \dfrac{1}{\mu - \lambda}$
- Assuming the mean arrival rate is 10 requests/s
- We will calculate the effect of service rate on response time
- Complete the following table and see what you can conclude

| Service rate | Utilisation $\lambda/\mu$ | Response time |
|---|---|---|
| 11 | 10/11 = 0.909 | 1 |
| 22 | 10/22 = 0.454 | 0.08 |

# Multi-server queues M/M/m




All arrivals go into one queue.

Customers can be served by any one of the *m* servers.

When a customer arrives

- If all servers are busy, it will join the queue
- Otherwise, it will be served by one of the available servers

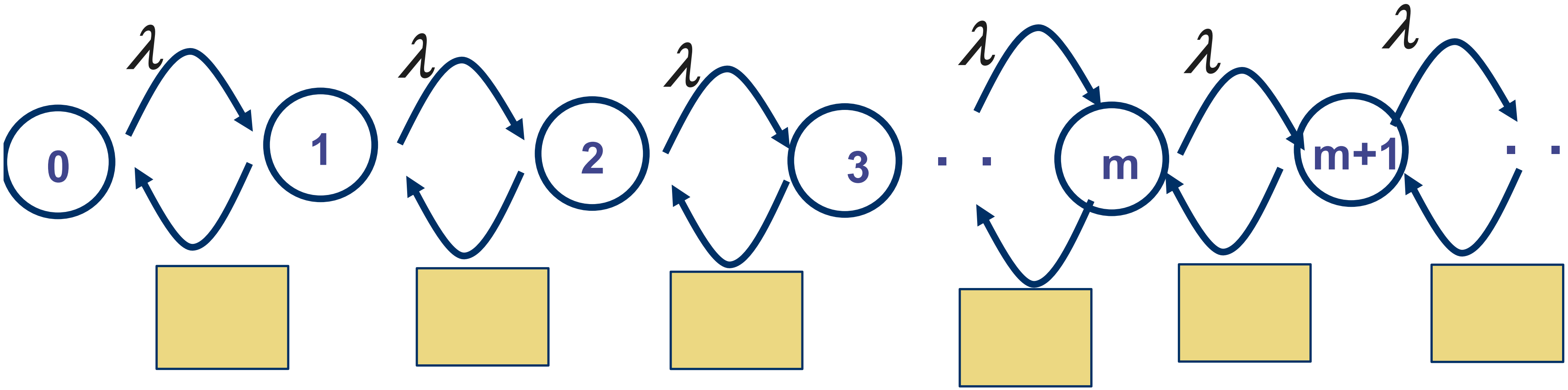
# A call centre analogy of M/M/m queue

- Consider a call centre analogy
  - Calls are arriving according to Poisson distribution with rate $\lambda$
  - The length of each call is exponentially distributed with parameter $\mu$
    - Mean length of a call is $1/\mu$

**Arrivals** →

**Call centre with *m* operators**
**If all *m* operators are busy, the centre will put the call on hold.**
**A customer will wait until his call is answered.**

# State transition for M/M/m



# M/M/m

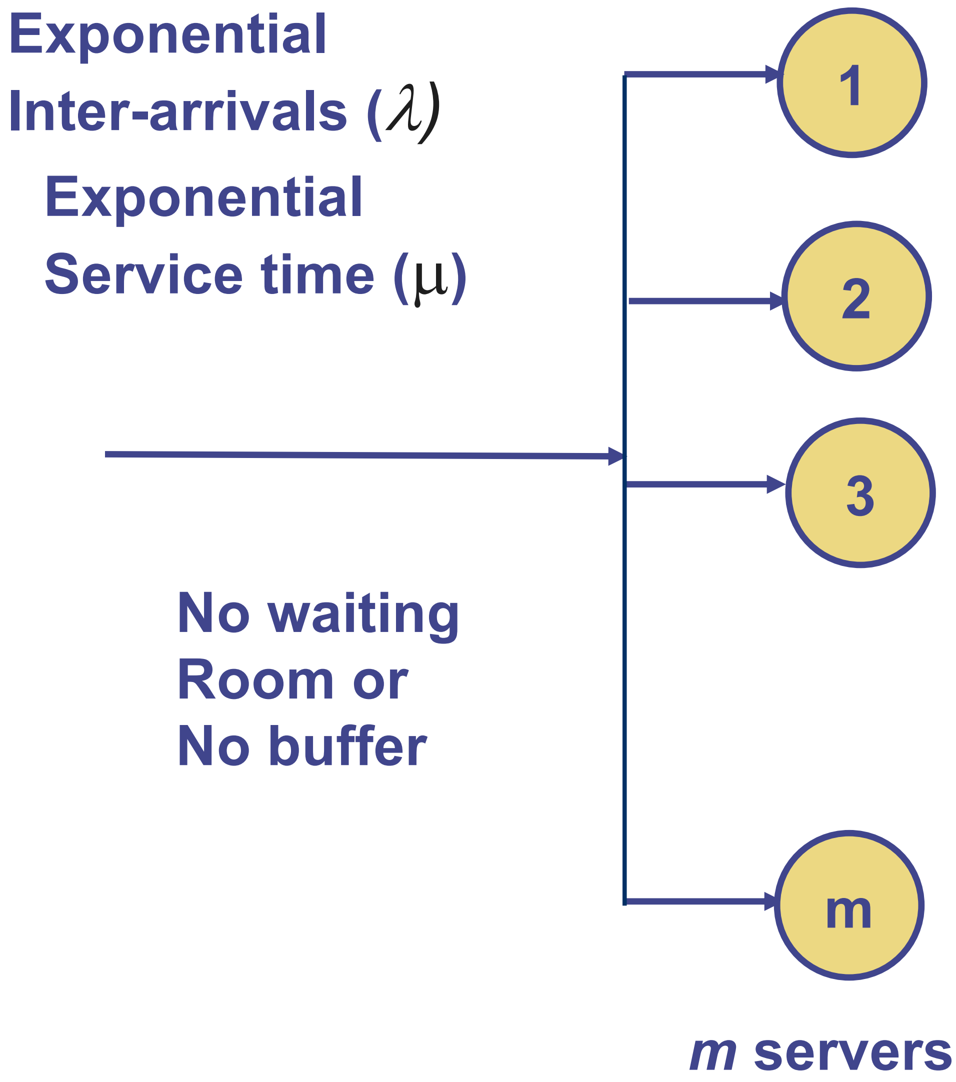
- Following the same method, we have mean response time *T* is

$$T = \frac{C(\rho, m)}{m\mu(1-\rho)} + \frac{1}{\mu}$$

where $$\rho = \frac{\lambda}{m\mu}$$

$$C(\rho, m) = \frac{\frac{(m\rho)^m}{m!}}{(1-\rho)\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!}}$$

# Multi-server queues M/M/m/m with no waiting room




**An arrival can be served by any one of the *m* servers.**

**When a customer arrives**

- **If all servers are busy, it will *depart* from the system**

- **Otherwise, it will be served by one of the available servers**

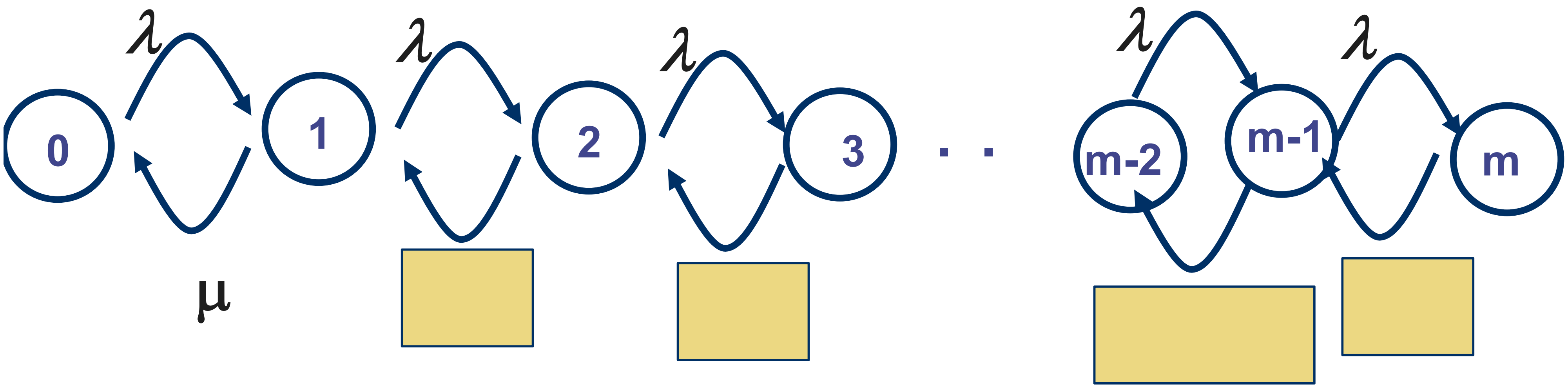
# A call centre analogy of M/M/m/m queue

- Consider a call centre analogy
  - Calls are arriving according to Poisson distribution with rate $\lambda$
  - The length of each call is exponentially distributed with parameter $\mu$
    - Mean length of a call is $1/\mu$

**Arrivals** →

**Call centre with *m* operators**
**If all *m* operators are busy, the call is dropped.**

# State transition for M/M/m/m

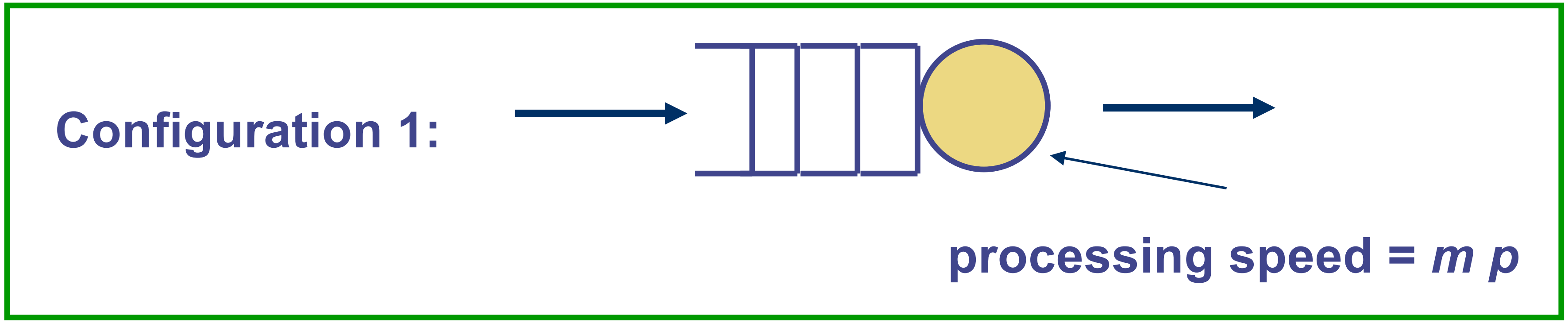


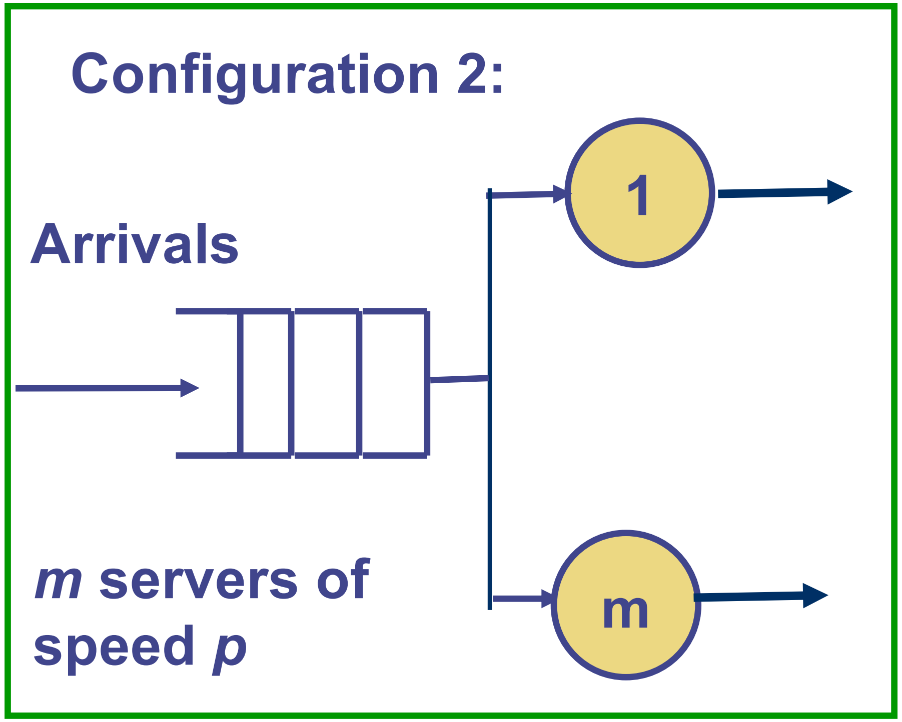

**Probability that an arrival is blocked**
**= Probability that there are m customers in the system**

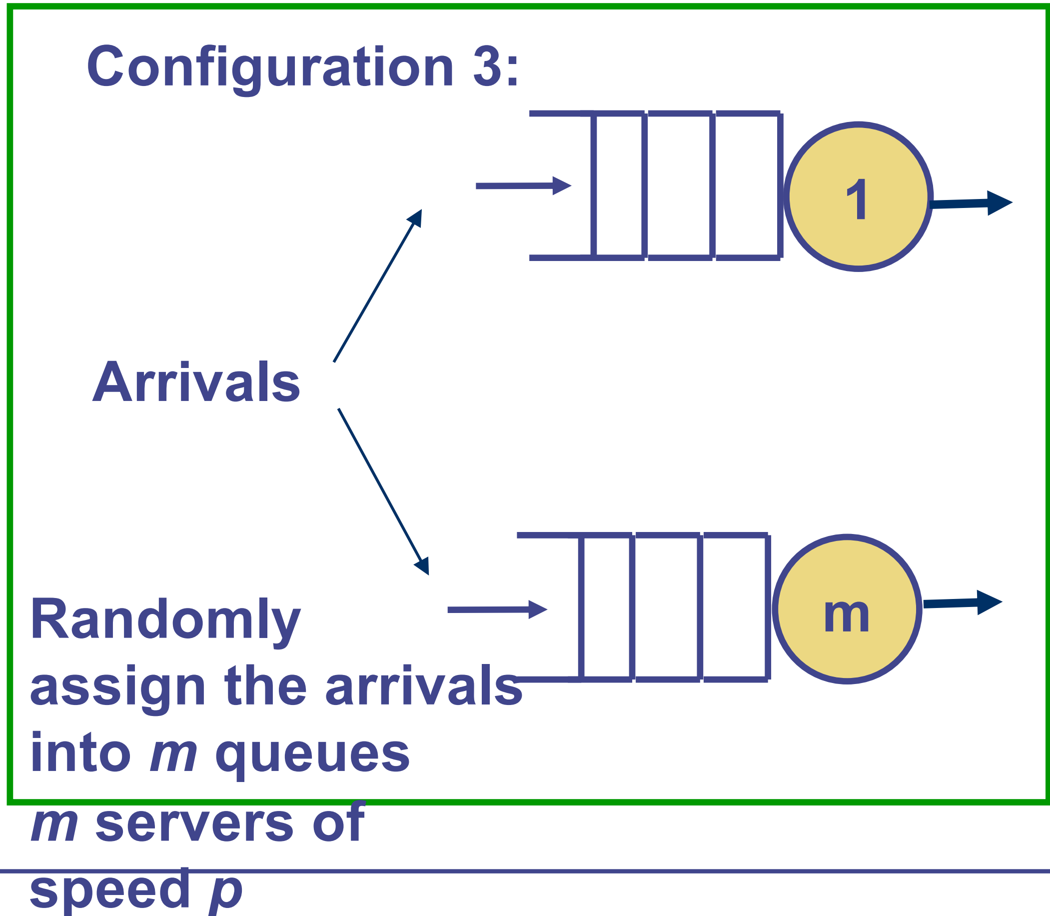
$$P_m = \frac{\frac{\rho^m}{m!}}{\sum_{k=0}^{m} \frac{\rho^k}{k!}} \quad \textbf{where} \quad \rho = \frac{\lambda}{\mu}$$

**"Erlang B formula"**

# What configuration has the best response time?




**Try out the tutorial question!**

# References

- Recommended reading
  - Queues with Poisson arrival are discussed in
  - Bertsekas and Gallager, *Data Networks*, Sections 3.3 to 3.4.3
  - Note: I derived the formulas here using continuous Markov chain but Bertsekas and Gallager used discrete Markov chain
  - Mor Harchal-Balter. Chapters 13 and 14