

COMP9334

Capacity Planning for Computer Systems and Networks

Assignment Project Exam Help

Week 3: <https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

Pre-lecture exercise 1:

- Let X and Y be two events
- Let $\text{Prob}[X]$ = Probability that event X occurs
- Let $\text{Prob}[Y]$ = Probability that event Y occurs
- Question: Under what condition will the following equality hold?

Assignment Project Exam Help

- $\text{Prob}[X \text{ or } Y] =$ <https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

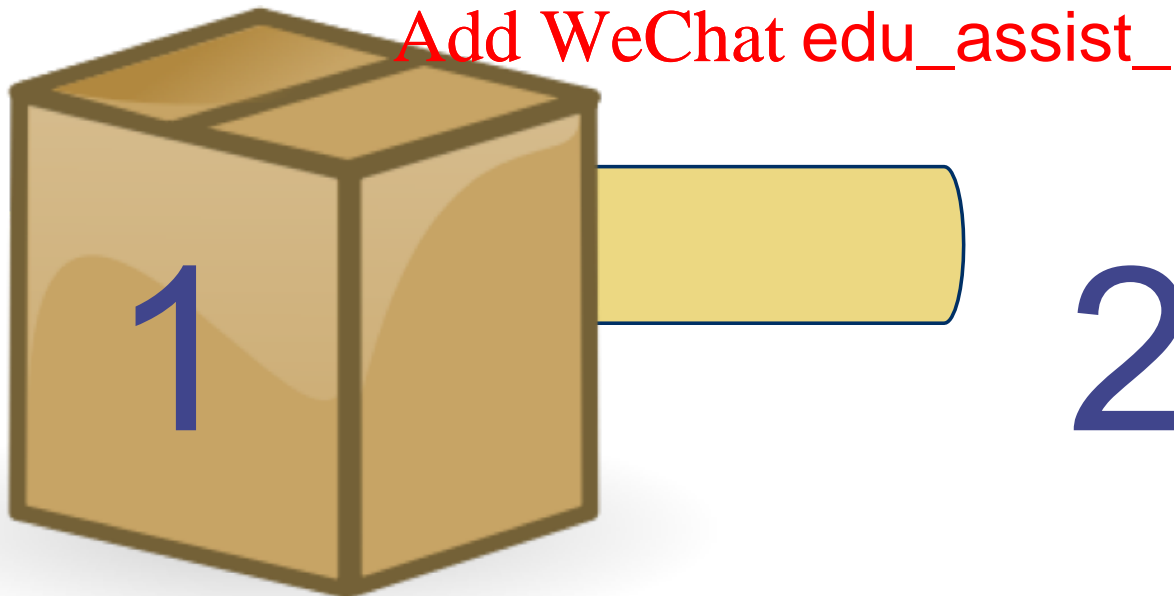
Pre-lecture exercise 2: Where is Felix? (Page 1)

- You have two boxes: Box 1 and Box 2, as well as a cat called Felix
- The two boxes are connected by a tunnel
- Felix likes to hide inside these boxes and travels between them using the tunnel.
- Felix is a very fast cat so the probability of finding him in the tunnel is zero
- You know Felix is in one of the boxes but you don't know which one

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro



Pre-lecture exercise 2: Where is Felix? (Page 2)

- Notation:

- $\text{Prob}[A]$ = probability that event A occurs
- $\text{Prob}[A \mid B]$ = probability that event A occurs given event B

- You do know

- Felix is in one of the boxes at times 0 and 1
- $\text{Prob}[\text{Felix is in Box 1 at time 0}] = 0.4$
- $\text{Prob}[\text{Felix will be in Box 1 at time 1} \mid \text{Felix is in Box 1 at time 0}] = 0.4$
- $\text{Prob}[\text{Felix will be in Box 1 at time 1} \mid \text{Felix is in Box 2 at time 0}] = 0.2$

- Calculate

- $\text{Prob}[\text{Felix is in Box 1 at time 1}]$
- $\text{Prob}[\text{Felix is in Box 2 at time 1}]$



Pre-lecture exercise 3

- 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?

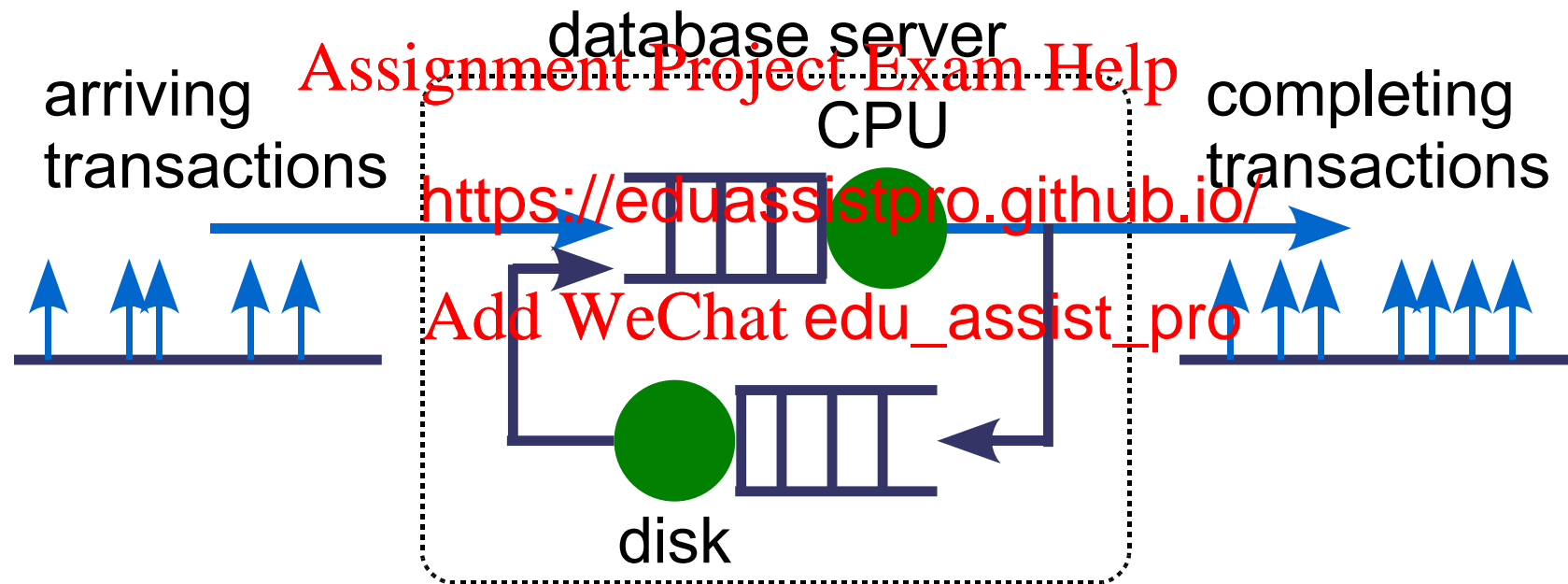
Value	Probabilit
1	0.1
2	0.1
3	0.2
4	0.1
5	0.3
6	0.2

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

Week 1:

- Modelling a computer system as a network of queues
- Example: Open queueing network consisting of two queues



Week 2:

- Operational analysis
 - Measure #completed jobs, busy time etc
 - Operational quantities: utilisation, response time, throughput etc.
 - Operational laws relate the operational quantities

Assignment Project Exam Help

- Bottleneck anal

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

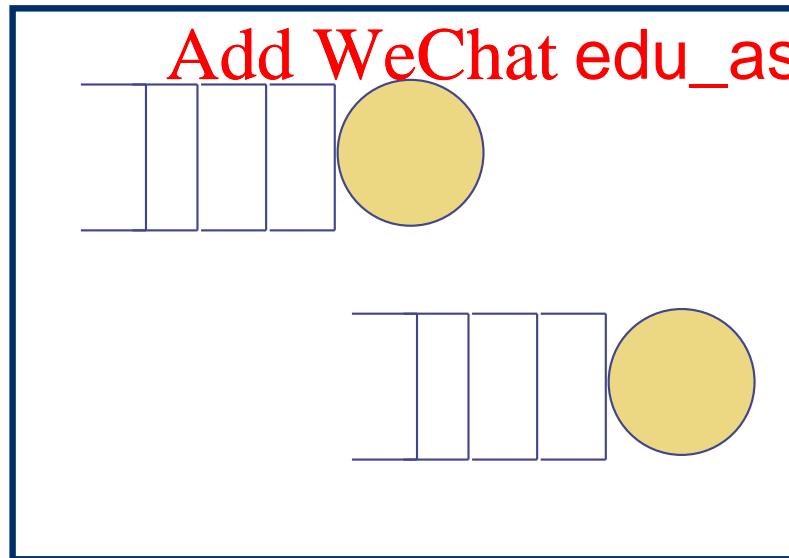
Little's Law

- 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

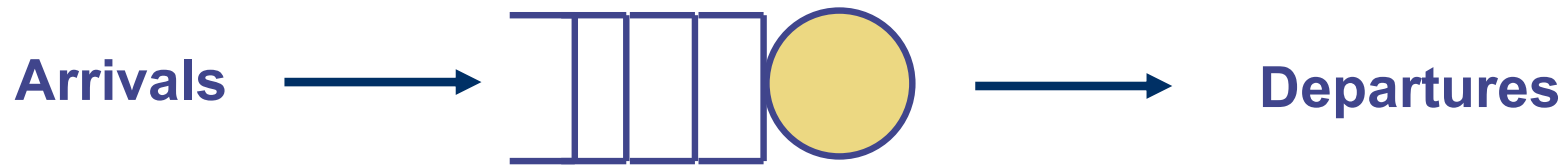
Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

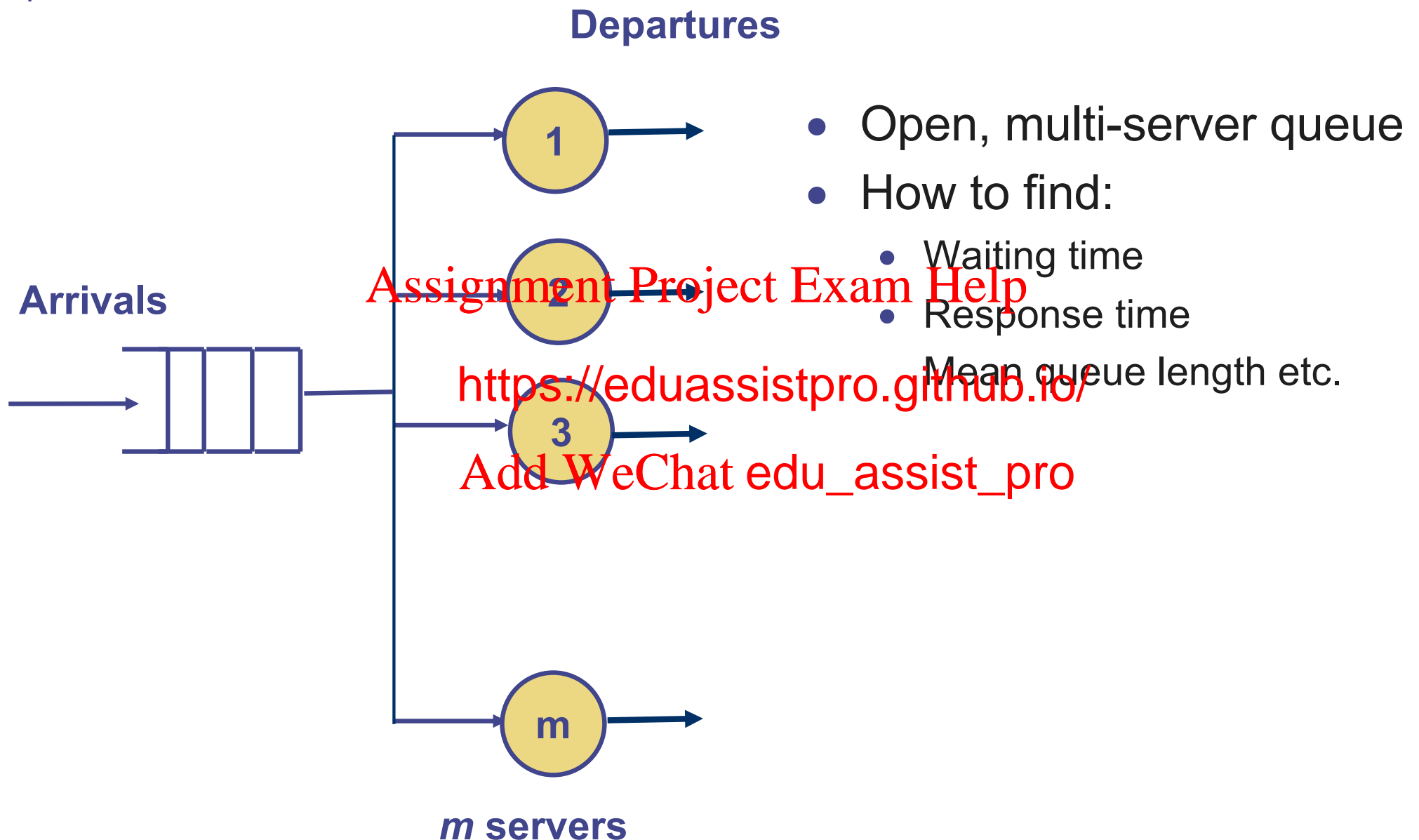


This week (1)



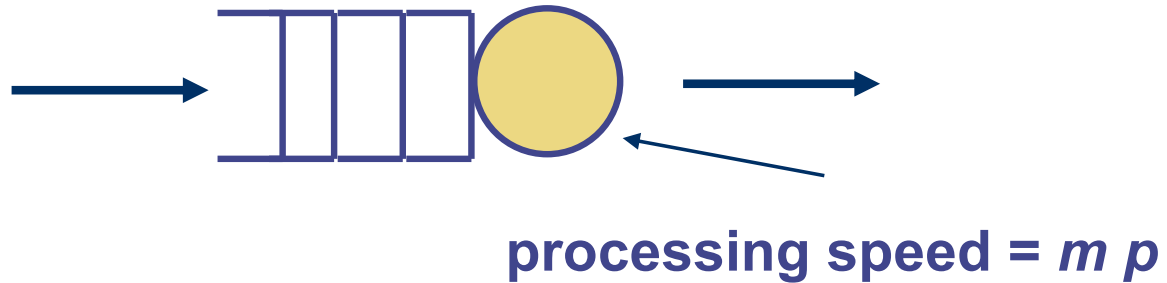
- Open, single server queues and
- How to find:
 - Waiting time <https://eduassistpro.github.io/>
 - Response time
 - Mean queue length etc.
- The technique to find waiting time etc. is called *Queueing Theory*

This week (2)

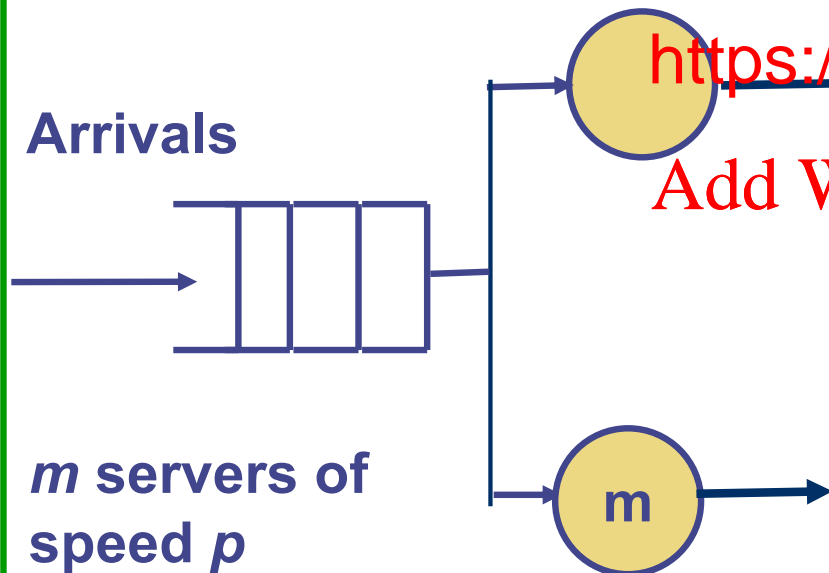


What will you be able to do with the results?

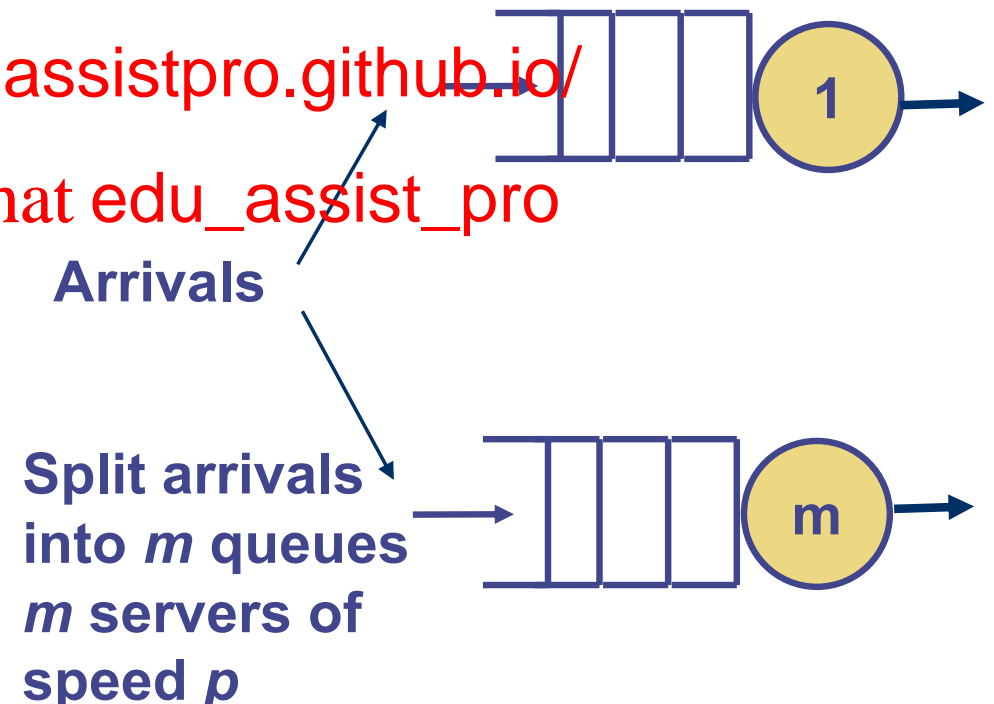
Configuration 1:



Configuration 2:



Configuration 3:



Which configuration has the best response time?

Be patient

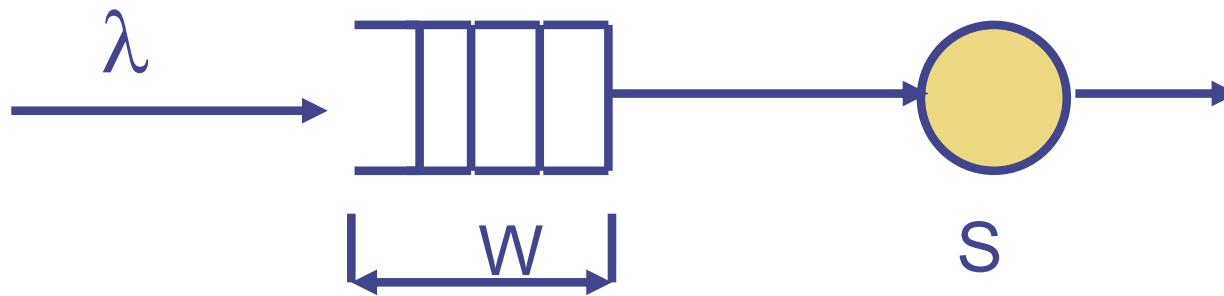
- We will show how we can obtain the response time
 - It takes a number of steps to obtain the answer
- It takes time to stand in a queue, it also takes time to derive results in queuing theory!

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

Single Server Queue: Terminology



Time spent waiting
Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

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 ρ for utilisation rather than U .

Single server system

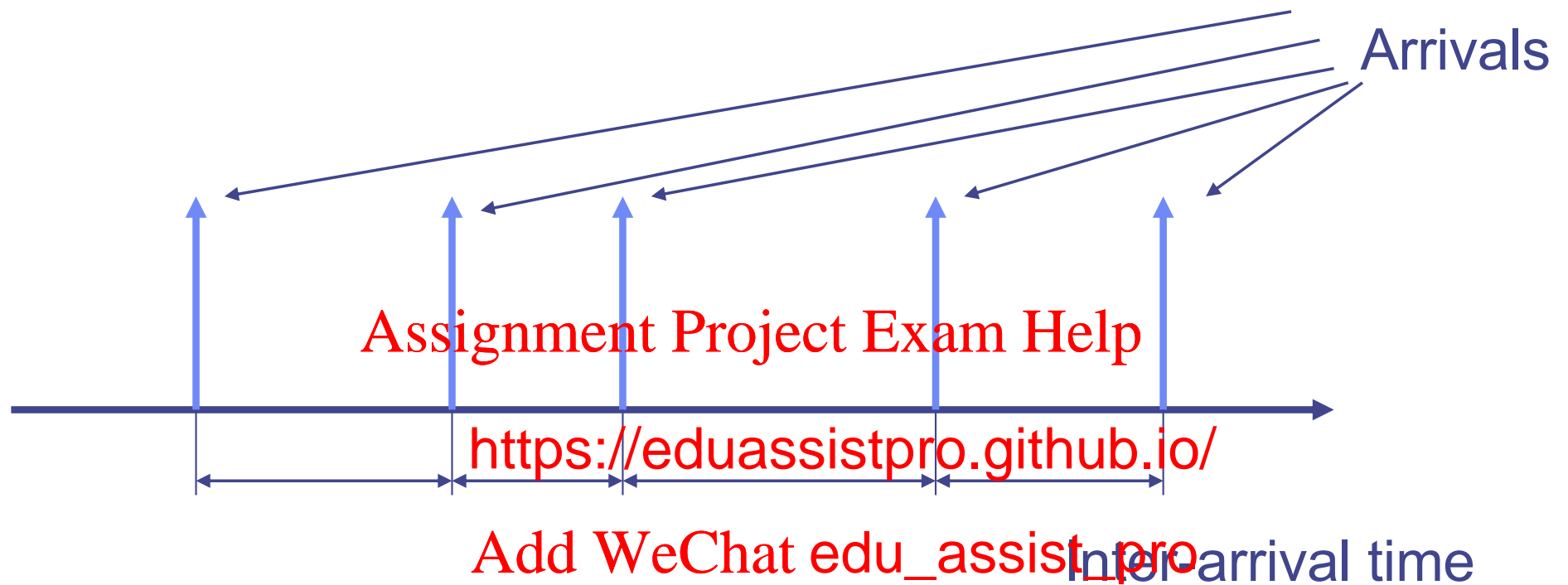
- In order to determine the response time, you need to know
 - The inter-arrival time probability distribution
 - The service time probability distribution
- Possible distributions
 - Deterministic
 - Constant inter-arrival time
 - Constant service time
 - Exponential distribution
- We will focus on exponential

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

Exponential inter-arrival with rate λ



We assume that successive arrivals are independent

Probability that inter-arrival time is between x and $x + \delta x$
 $= \lambda \exp(-\lambda x) \delta x$

Poisson distribution (1)

- The following are equivalent
 - The inter-arrival time is independent and exponentially distributed with parameter λ
 - The number of arrivals in an interval T is a Poisson distribution with parameter λ

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

- Mean inter-arrival time = $1 / \lambda$
- Mean number of arrivals in time interval $T = \lambda T$
- Mean arrival rate = λ

Poisson distribution (2)


- Poisson distribution arises from a large number of independent sources
 - An example from Week 2:
 - N customers, each with a probability of p per unit time to make a request.
 - This creates a Poisson arrival with $\lambda = Np$
- Another interpretation of Poisson arrival:
 - Consider a small time interval δ
 - This means $\delta^n \rightarrow 0$
 - Probability [no arrival in δ] = 1
 - Probability [1 arrival in δ] = $\lambda \delta$
 - Probability [2 or more arrivals in δ] ≈ 0
- This interpretation can be derived from:

$$Pr[k \text{ arrivals in a time interval } T] = \frac{(\lambda T)^k \exp(-\lambda T)}{k!}$$

Service time distribution

- Service time = the amount of processing time a job requires from the server
- We assume that the service time distribution is exponential with parameter μ
 - The probability that the service time is between t and $t + \delta t$ is:

<https://eduassistpro.github.io/>

- Here: μ = service rate = $1 / \text{mean}$
- Another interpretation of exponential time:
 - Consider a small time interval δ
 - Probability [a job will finish its service in next δ seconds] = $\mu \delta$
 - Probability [a job will **not** finish its service in next δ seconds] = 

Sample queueing problems

- Consider a call centre
 - Calls are arriving according to Poisson distribution with rate λ
 - The length of each call is exponentially distributed with parameter μ
 - Mean length of a call is $1/\mu$ (in, e.g. seconds)

Arrivals



Assignment Project Exam Help

Call centre:

m op

If all

at most n additional

If a call arrives when

slots are used, the call is rejected.

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

entre can put

ld.

rs and holding

- Queueing theory will be able to answer these questions:
 - What is the mean waiting time for a call?
 - What is the probability that a call is rejected?

Road map

- We will start by looking at a call centre with one operator and no holding slot
 - This may sound unrealistic but we want to show how we can solve a typical queueing network problem
 - After that we go into queues that are more complicated

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

Call centre with 1 operator and no holding slots

- Let us see how we can solve the queuing problem for a very simple call centre with 1 operator and no holding slots
- What happens to a call that arrives when the operator is busy?
 - [Assignment Project Exam Help](https://eduassistpro.github.io/)
- What happens then the operator is idle?
 - [Add WeChat edu_assist_pro](https://eduassistpro.github.io/)
- We are interested to find the probability that an arriving call is rejected.

Arrivals



Call centre:

1 operator. No holding slot.

Solution (1)

- There are two possibilities for the operator:
 - Busy or
 - Idle
- Let

- State 0 = Operator is idle (no #calls in the call centre = ?)

- State 1 = Operator is busy (1 call in the call centre = ?)

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

$P_1(t)$ = Prob. 1 call in the call centre at time t

Solution (2)

We try to express $P_0(t + \Delta t)$ in terms of $P_0(t)$ and $P_1(t)$

- No call at call centre at $t + \Delta t$ can be caused by
 - No call at time t and no call arrives in $[t, t + \Delta t]$, or

-

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

**Question: Why do we NOT have to consider the following possibility:
No customer at time t & 1 customer arrives in $[t, t + \Delta t]$ &
the call finishes within $[t, t + \Delta t]$.**

Solution (3)

- Similarly, we can show that

$$P_1(t + \Delta t) = P_0(t)\lambda\Delta t + P_1(t)(1 - \mu\Delta t)$$

- If we let $\Delta t \rightarrow 0$, we have

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

$$\frac{dP_1(t)}{dt} = P_0(t)\lambda - P_1(t)\mu$$

Solution (4)

- We can solve these equations to get

$$P_0(t) = \frac{\mu}{\lambda + \mu} - \frac{\mu}{\lambda + \mu} e^{-(\mu + \lambda)t}$$

Assignment Project Exam Help

<https://eduassistpro.github.io/>

- This is too complicated, let us look for a steady state solution

$$P_0 = P_0(\infty) = \frac{\mu}{\lambda + \mu}$$

$$P_1 = P_1(\infty) = \frac{\lambda}{\lambda + \mu}$$

Solution (5)

- From the steady state solution, we have
 - The probability that an arriving call is rejected
 - = The probability that the operator is busy
 - =

Assignment Project Exam Help

<https://eduassistpro.github.io/>

- Let us check whether it makes
 - For a constant μ , if the arrival rate λ eases, will the probability that the operator is busy go up or down?
 - Does the formula give the same prediction?

An alternative interpretation

- We have derived the following equation:

$$P_0(t + \Delta t) = P_0(t)(1 - \lambda\Delta t) + P_1(t)\mu\Delta t$$

- Which can be rewritten as:

Assignment Project Exam Help

<https://eduassistpro.github.io/>

- At steady state: Add WeChat edu_assist_pro

Change in Prob in State 0 = 0

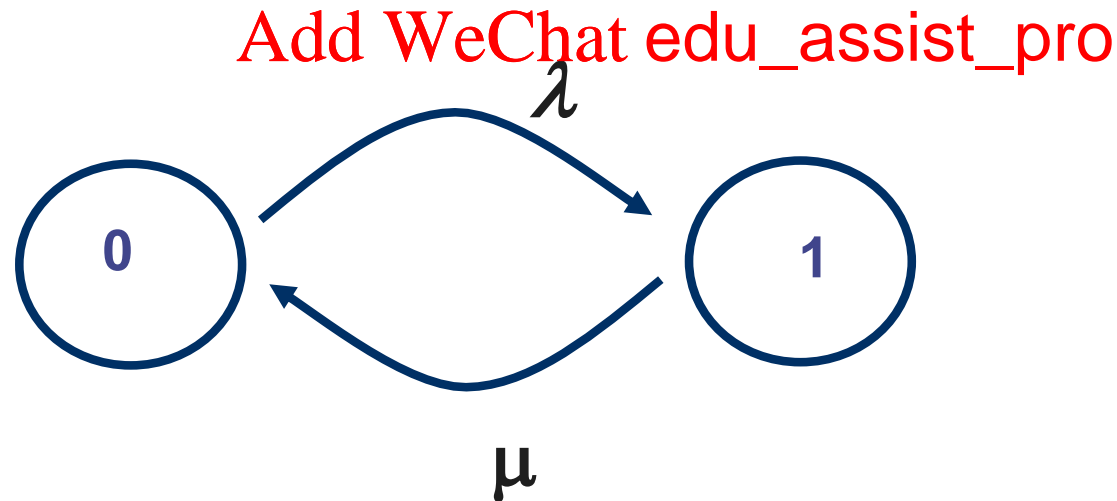
$$\Rightarrow 0 = -\boxed{P_0\lambda}\Delta t + \boxed{P_1\mu}\Delta t$$

Rate of leaving state 0

Rate of entering state 0

Faster way to obtain steady state solution (1)

- Transition from State 0 to State 1
 - Caused by an arrival, the rate is λ
- Transition from State 1 to State 0
 - Caused by a completed service, the rate is μ
- State diagram representation
 - Each circle is a
 - Label the arc b



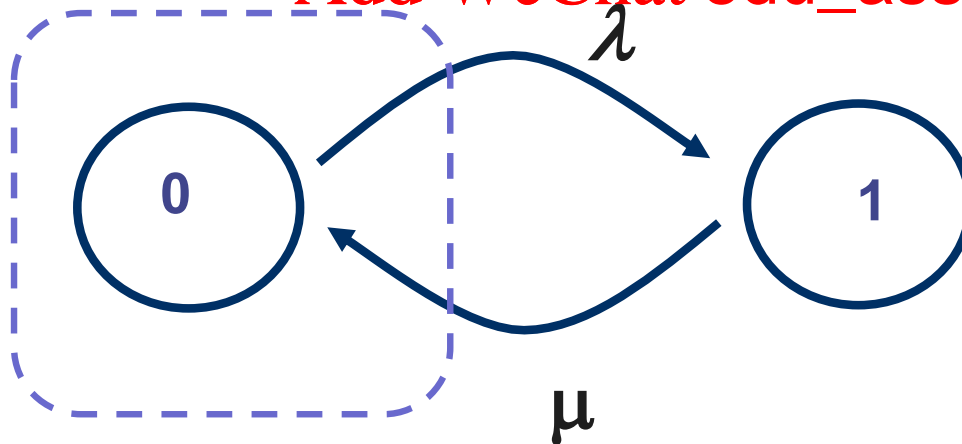
Faster way to obtain steady state solution (2)

- Steady state means
 - rate of transition out of a state = Rate of transition into a state**
- We have for state 0:

$$\lambda P_0 = \mu P_1$$

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro



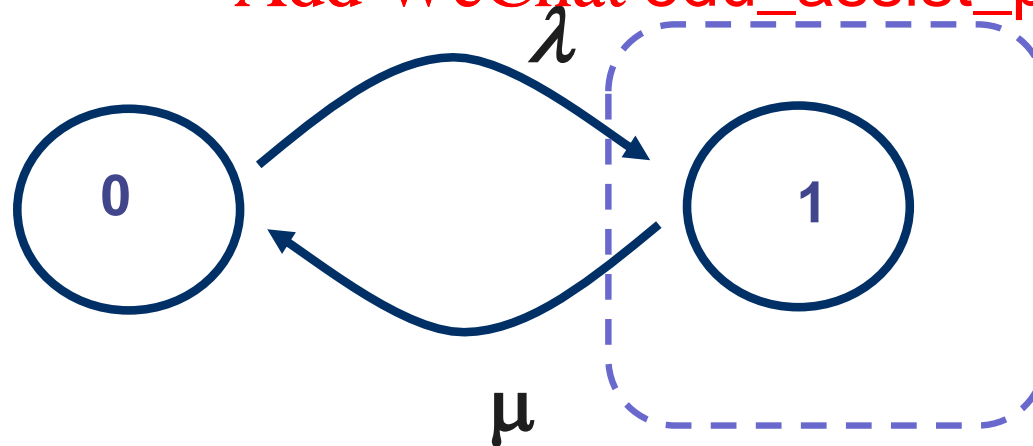
Faster way to obtain steady state solution (3)

- We can do the same for State 1:
- Steady state means
 - **Rate of transition into a state** = **rate of transition out of a state**
- We have for state 1:

$$\lambda P_0 = \mu P_1$$

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro



Faster way to obtain steady state solution (4)

- We have one equation $\lambda P_0 = \mu P_1$
- We have 2 unknowns and we need one more equation.
- Since we must be either one of the two states:

Assignment Project Exam Help

<https://eduassistpro.github.io/>

- Solving these two equations, we get steady state solution as before

Add WeChat edu_assist_pro

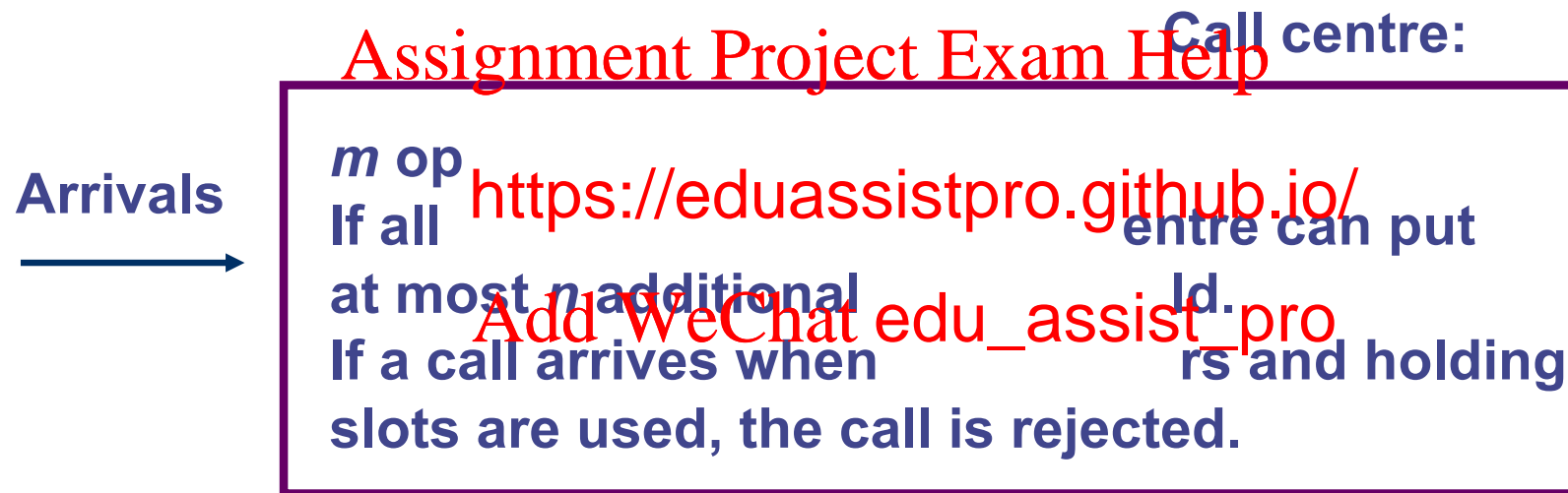
$$P_0 = \frac{\mu}{\lambda + \mu} \quad P_1 = \frac{\lambda}{\lambda + \mu}$$

Summary

- Solving a queueing problem is not simple
- It is harder to find how a queue evolves with time
- It is simpler to find how a queue behaves at steady state
 - Procedure:
 - Draw a diagram with the states
 - Add arcs between states
 - Derive flow balance equations, i.e.
 - Rate of entering a state = Rate of leaving a state
 - Solve the equation for steady state probabilities

Let us have a look at our call centre problem again

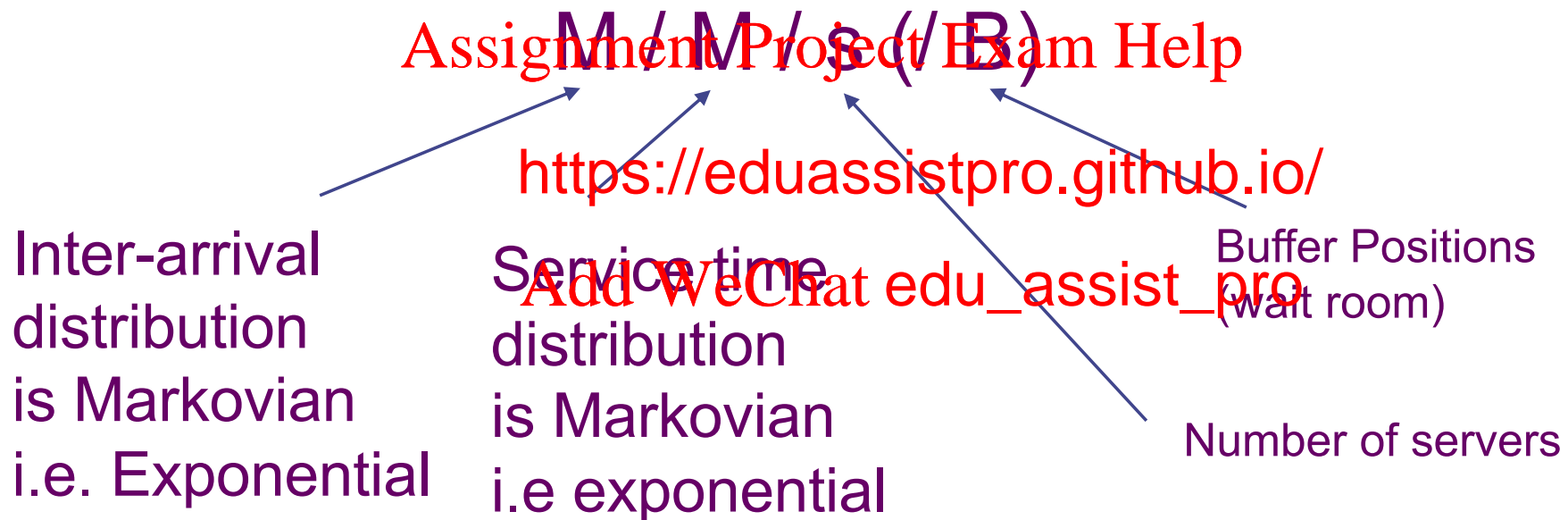
- Consider a call centre
 - Calls are arriving according to Poisson distribution with rate λ
 - The length of each call is exponentially distributed with parameter μ
 - Mean length of a call is $1/\mu$



- We solve the problem for $m = 1$ and $n = 0$
 - We call this a M/M/1/1 queue (explanation on the next page)
- How about other values of m and n

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:

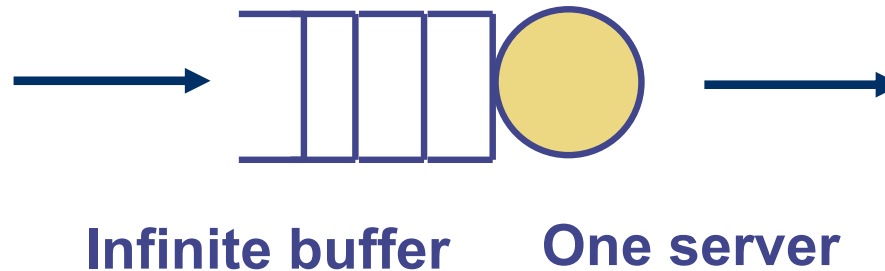


The call centre example on the last page is a $M/M/m/(m+n)$ queue
If $n = \infty$, we simply write $M/M/m$

M/M/1 queue

Exponential
Inter-arrivals (λ)

Exponential
Service time (μ)



- Consider a call centre analogy

- Calls are arriving with an exponential distribution with rate λ
- The length of each call is exponentially distributed with parameter μ
 - Mean length of a call is $1/\mu$

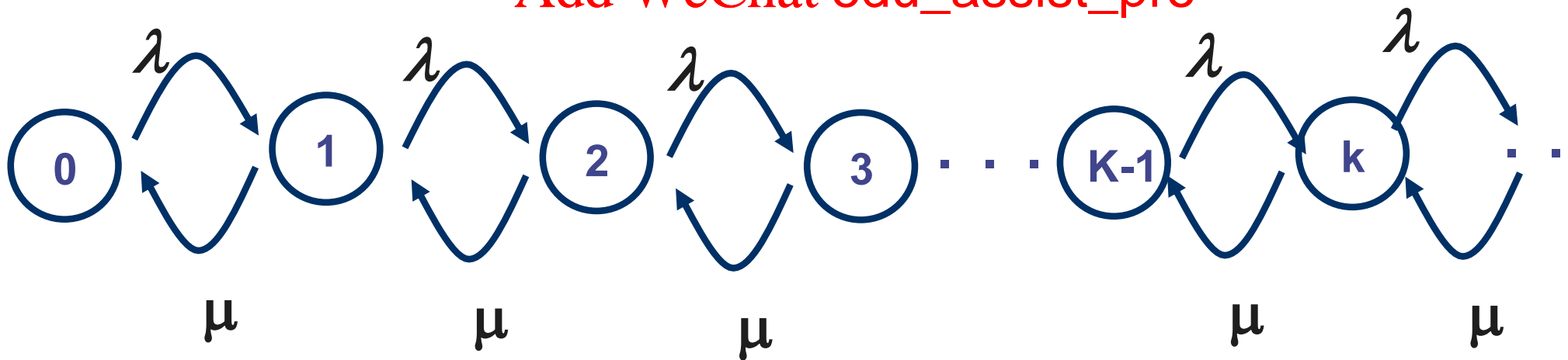
Arrivals
→

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 is the mean waiting time for a call?

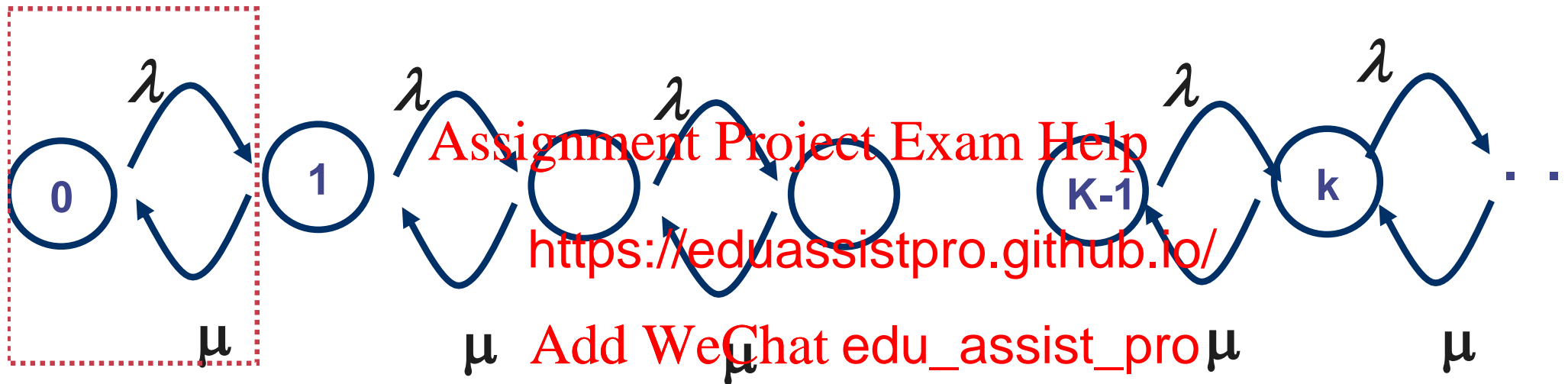
Solving M/M/1 queue (1)

- 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



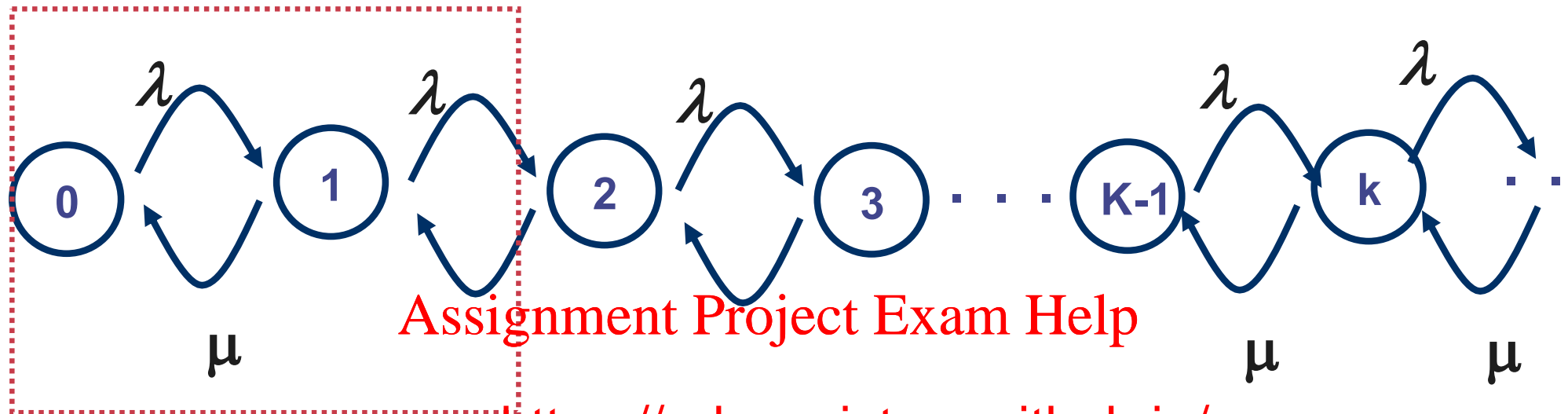
Solving M/M/1 queue (2)

P_k = Prob. k jobs in system



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

Solving M/M/1 queue (3)

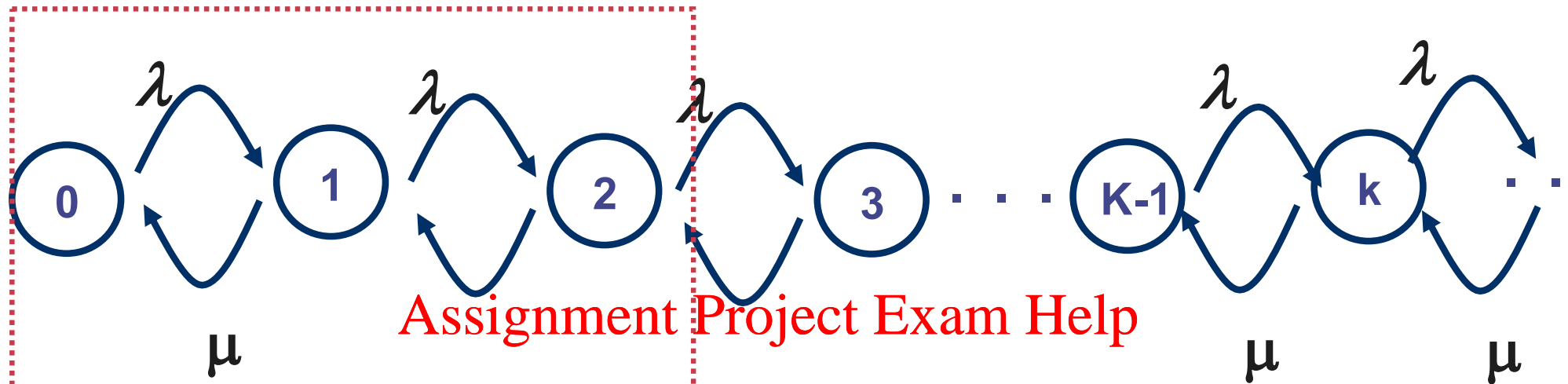


<https://eduassistpro.github.io/>

$$\lambda P_1 = \mu P_2 \quad \text{Add WeChat edu_assist_pro}$$

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

Solving M/M/1 queue (4)



<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

$$\lambda P_2 = \mu P_3$$

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

Solving M/M/1 queue (5)

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

Assignment Project Exam Help

Let $\rho = \frac{\lambda}{\mu}$ <https://eduassistpro.github.io/>
Add WeChat edu_assist_pro

We have $P_k = \rho^k P_0$

Solving M/M/1 queue (6)

With $P_k = \rho^k P_0$ and

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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

ρ = utilisation
= Prob server is busy
= $1 - P_0$
= 1 - Prob server is idle

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

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

Arrival rate < service rate

Solving M/M/1 queue (7)

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

This is the probability that there are k jobs in the system.
To find the response time, we will make use of Little's law.
First we need to find the mean number of customers in the system.

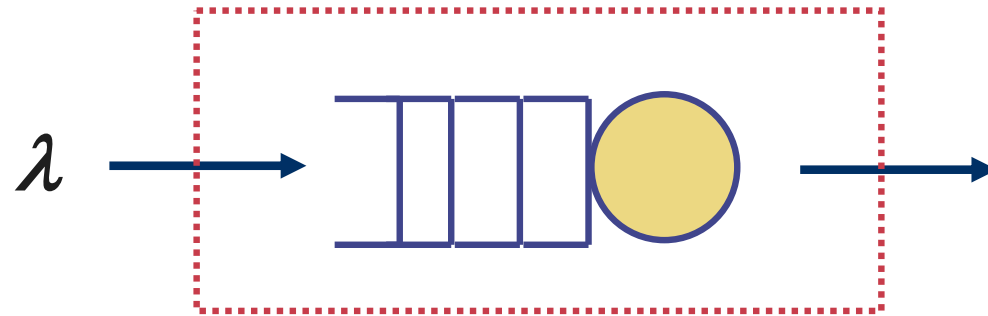
$$\sum_{k=0}^{\infty} k P_k$$

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

$$= \frac{\rho}{1 - \rho}$$

Solving M/M/1 queue (8)



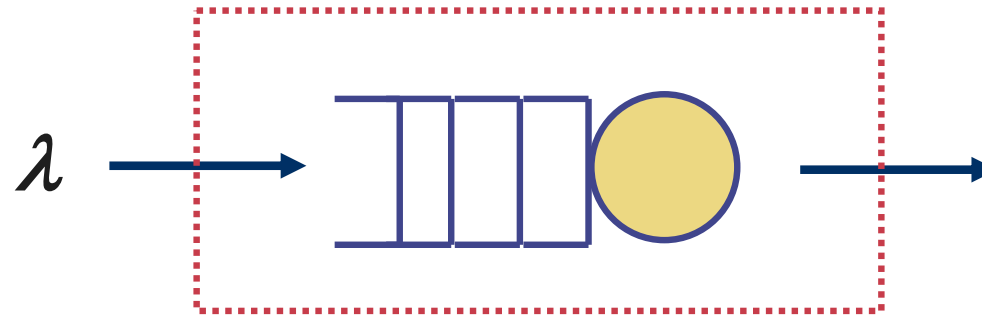
Little's law: **Assignment Project Exam Help**
mean number of cust **ponse time**

<https://eduassistpro.github.io/>

Throughput is λ (**why?**)
Add WeChat edu_assist_pro

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

Solving M/M/1 queue (9)



What is the mean waiting time at the queue?

Mean waiting time = m <https://eduassistpro.github.io/> service time

We know mean response time (from I [Add WeChat edu_assist_pro](#)

Mean service time is $= 1 / \mu$

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

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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

Observation:
Response time
increases

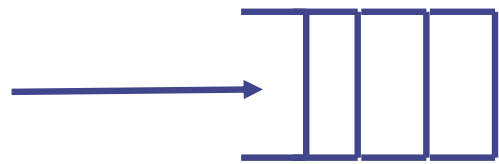
sharply when
 ρ gets close
to 1

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

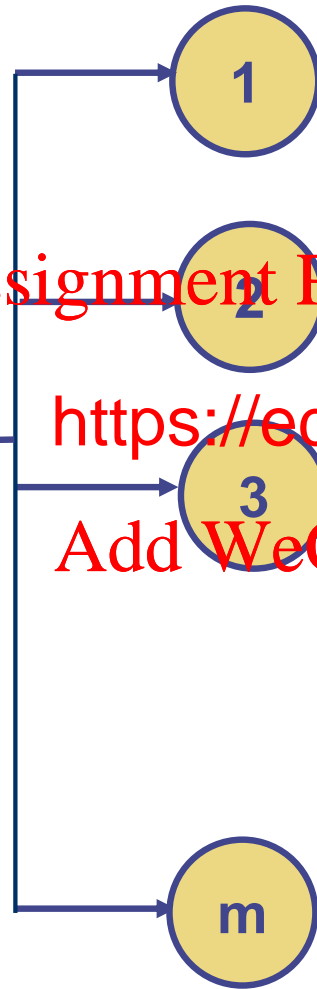
Multi-server queues M/M/m

Exponential
Inter-arrivals (λ)

Exponential
Service time (μ)



Infinite buffer



m servers

All arrivals go into one queue.

Customers can be served by any one of the m s.

customer arrives
servers are busy, it
will join the queue

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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

A call centre analogy of M/M/m queue

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

Arrivals



Call centre with m operators

If all m op
the call on

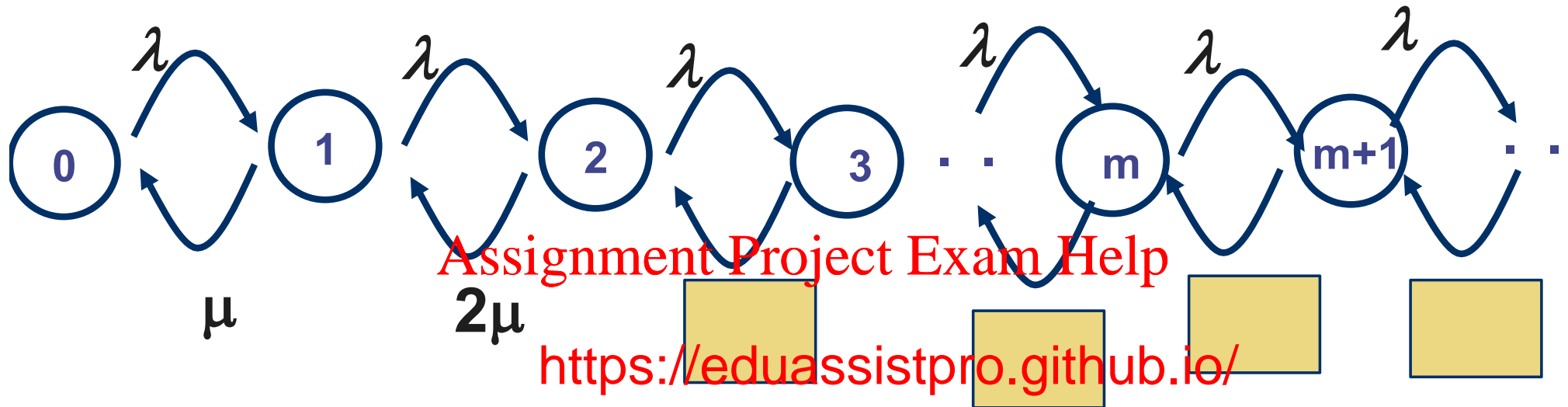
A customer will wait until

the will put

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

State transition for M/M/m



Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

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

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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

where

Add WeChat edu_assist_pro

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

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

Exponential
Inter-arrivals (λ)

Exponential
Service time (μ)

No waiting
Room or
No buffer



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

When a customer arrives
servers are busy, it
part from the

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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

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

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

Arrivals

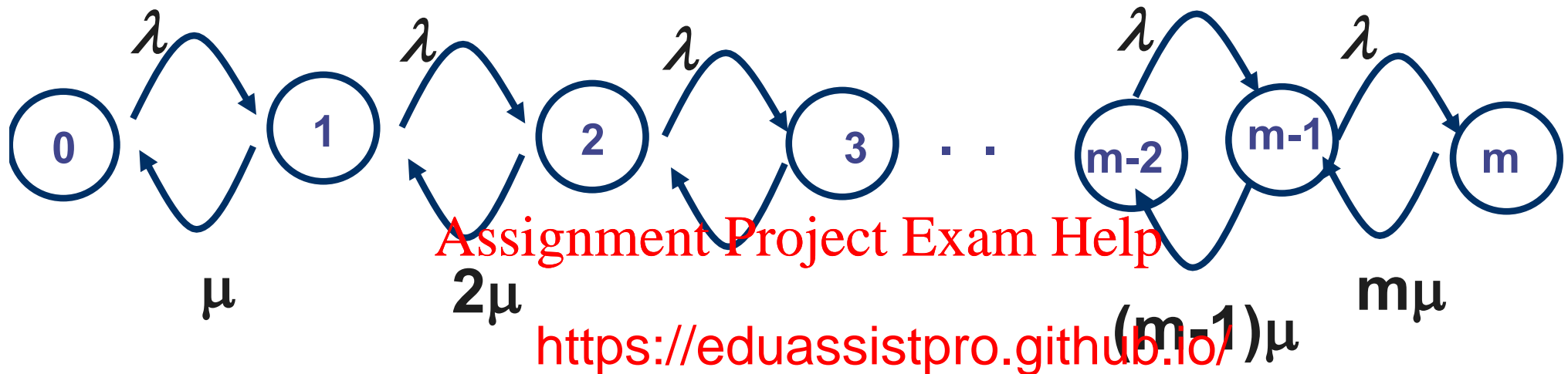


Call centre with m operators

If all m op is dropped.

Add WeChat edu_assist_pro

State transition for M/M/m/m



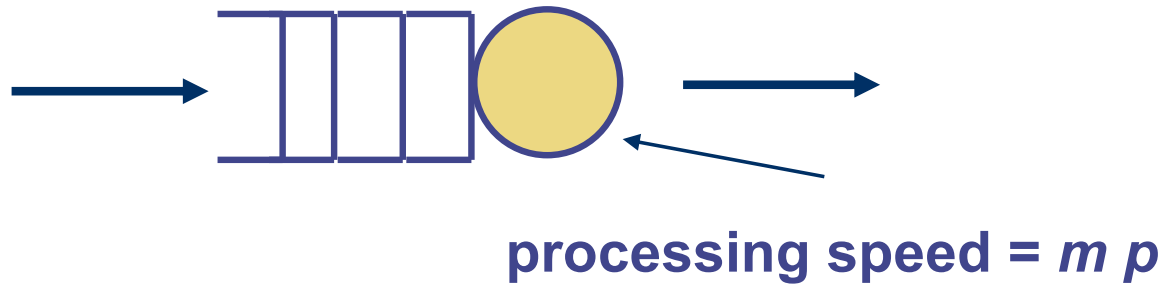
Add WeChat edu_assist_pro
**Probability that an arrival is
= Probability that there are m customers in the system**

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

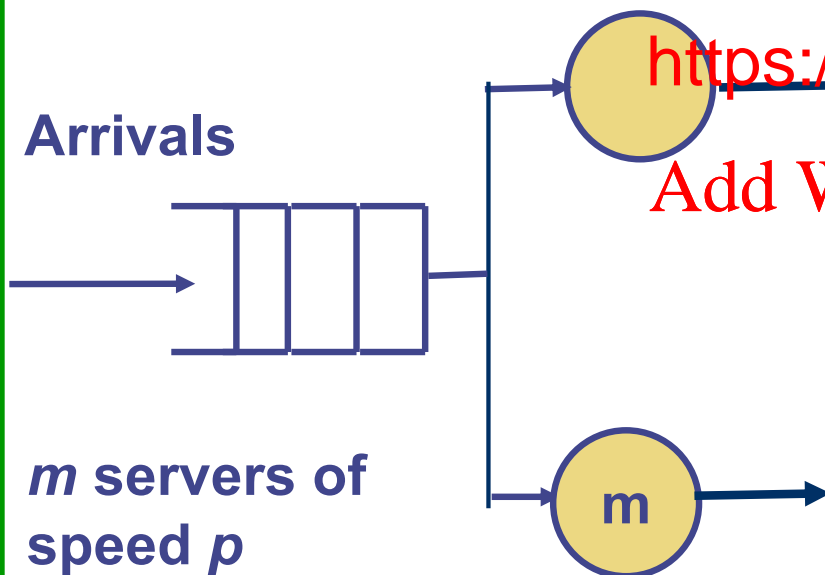
“Erlang B formula”

What configuration has the best response time?

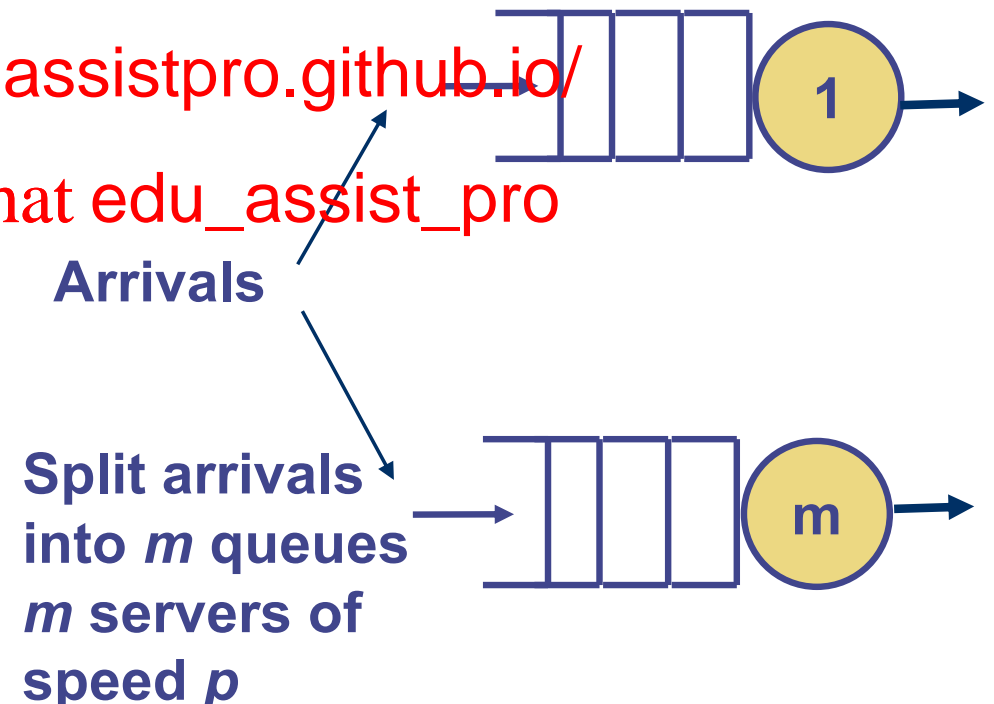
Configuration 1:



Configuration 2:



Configuration 3:



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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro