

COMP9334

Capacity Planning for Computer Systems and Networks

Assignment Project Exam Help

Week 2A: Operational Analysis (2).

Workload Characterisation

Last lecture

- Modelling a computer system as a queueing network
- Operational analysis on queueing networks
- We have derived these operational laws
 - Utilisation law $U(j) = X(j) S(j)$
 - Forced flow law $X(j) = V(j) X(0)$
 - Service demand law $D(j) = V(j) S(j) = U(j) / X(0)$
 - Little's law $N = X R$

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

This lecture

- Operational analysis (Continued)

- Using operational law for
 - Performance analysis
 - Bottleneck analysis

- Workload characterisation

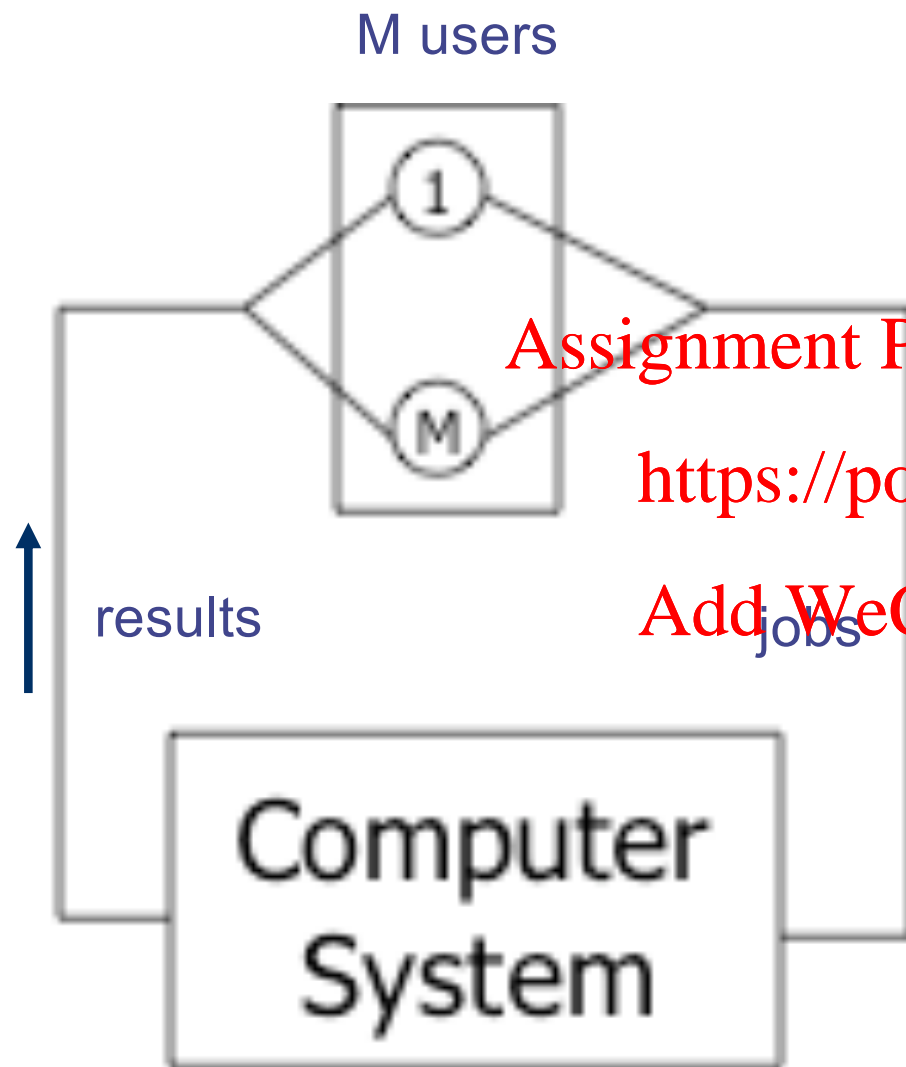
- Poisson processes and its properties

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Interactive systems



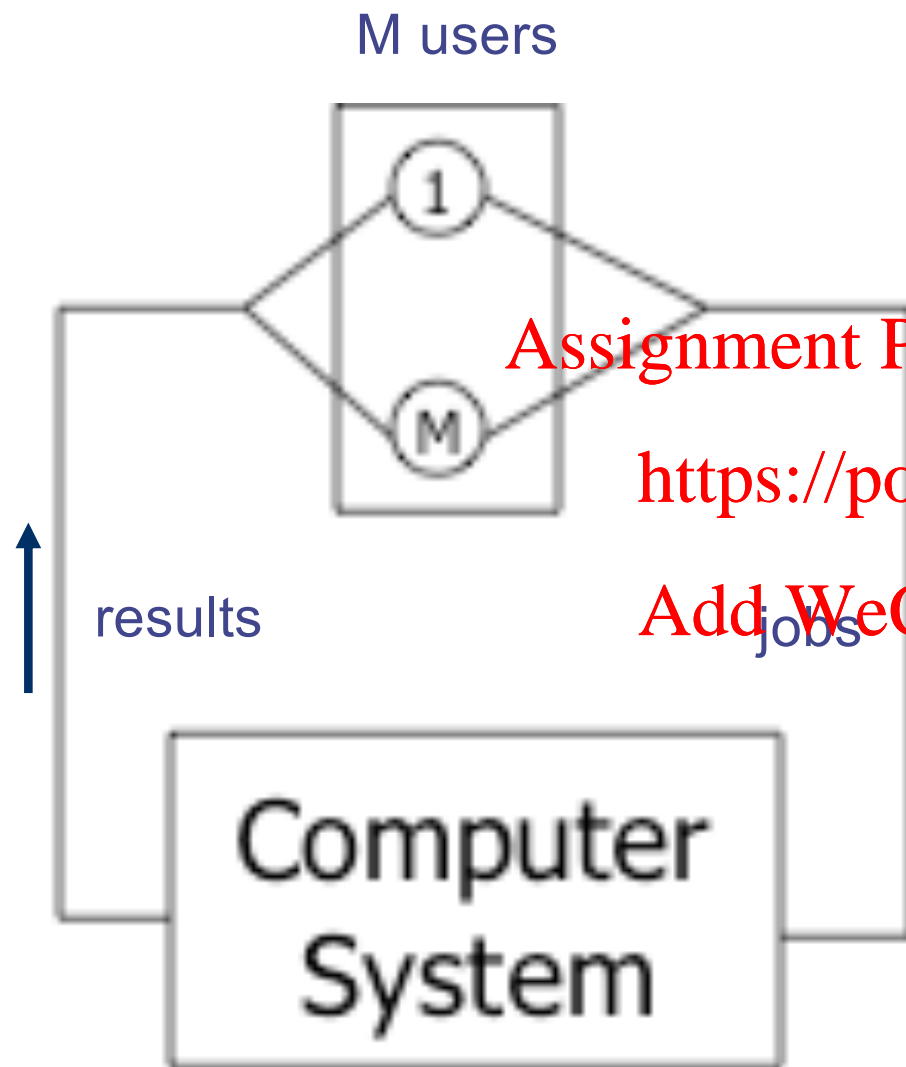
- An interactive system is used to model the interaction between humans (users) and computers

<https://powcoder.com>

Add WeChat powder

- The system consists of
 - A number of users
 - A computer system

Interactive systems (Cont'd)



- Interactions

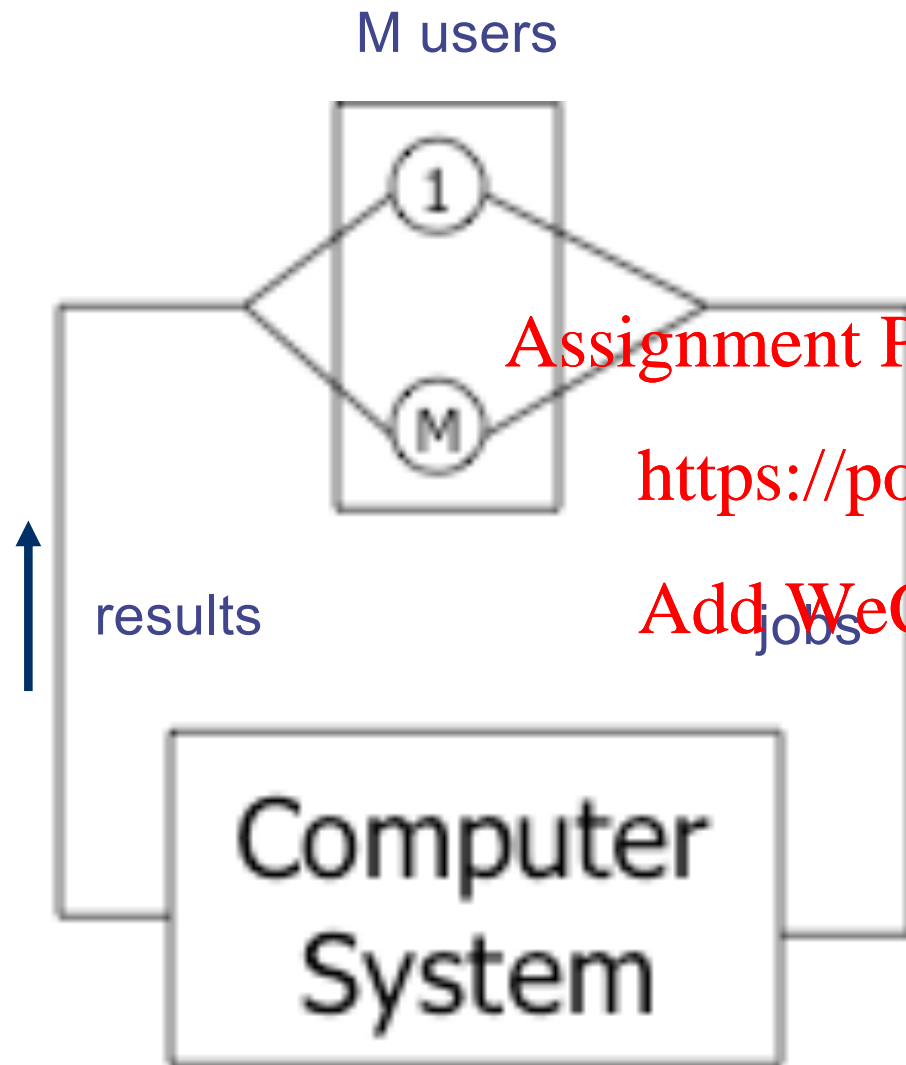
- Users send jobs to computer systems
- After finishing processing a job, the computer system returns the result to the user
- A user, after inspecting the results from the computer system, will send another job to the system

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Interactive systems: Modelling assumptions

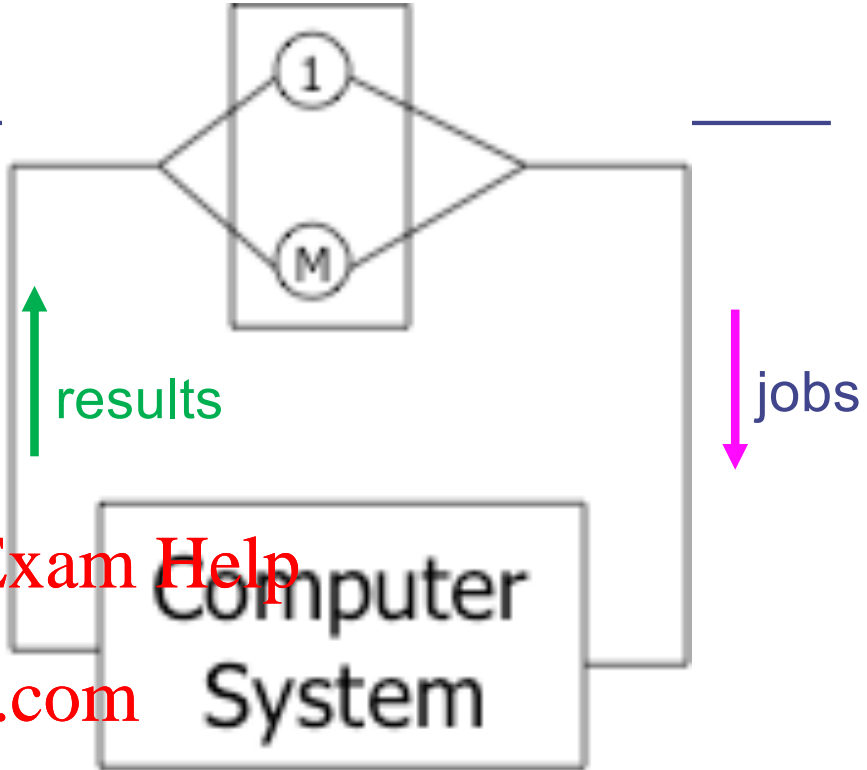


- Analyze interactive systems with specific assumptions
 - Fixed number of users denoted by M
 - Each user can have at most 1 job at the computer system
 - Each user goes through a cycle consisting of
 - Thinking time
 - Waiting for result time

Interactive cycle

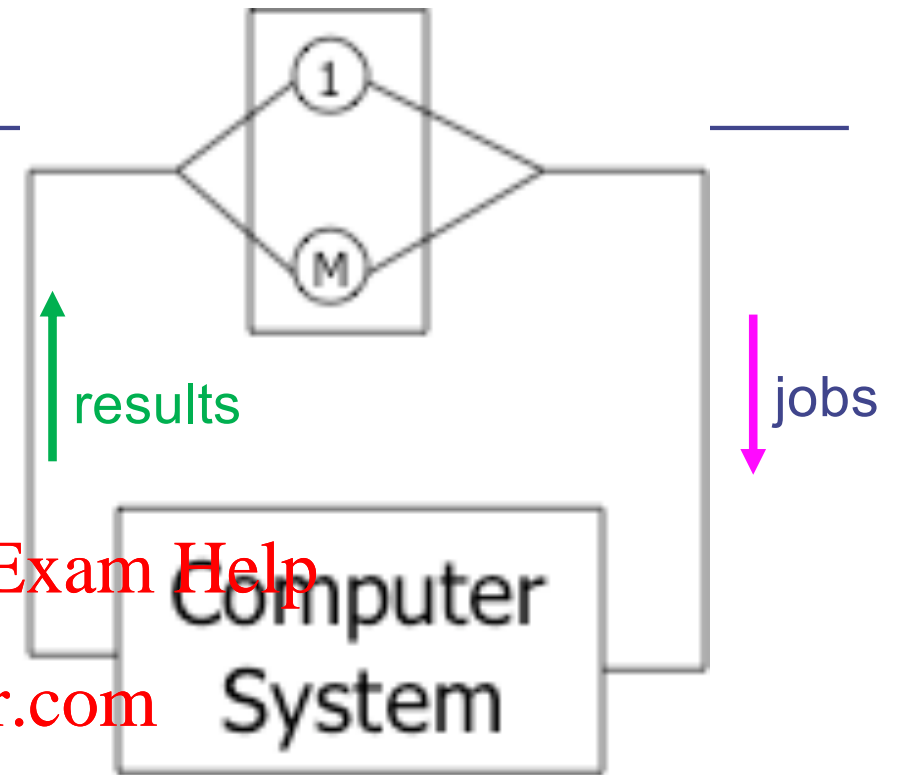
- # Assignment Project Exam Help

Add WeChat powcoder



Interactive cycle

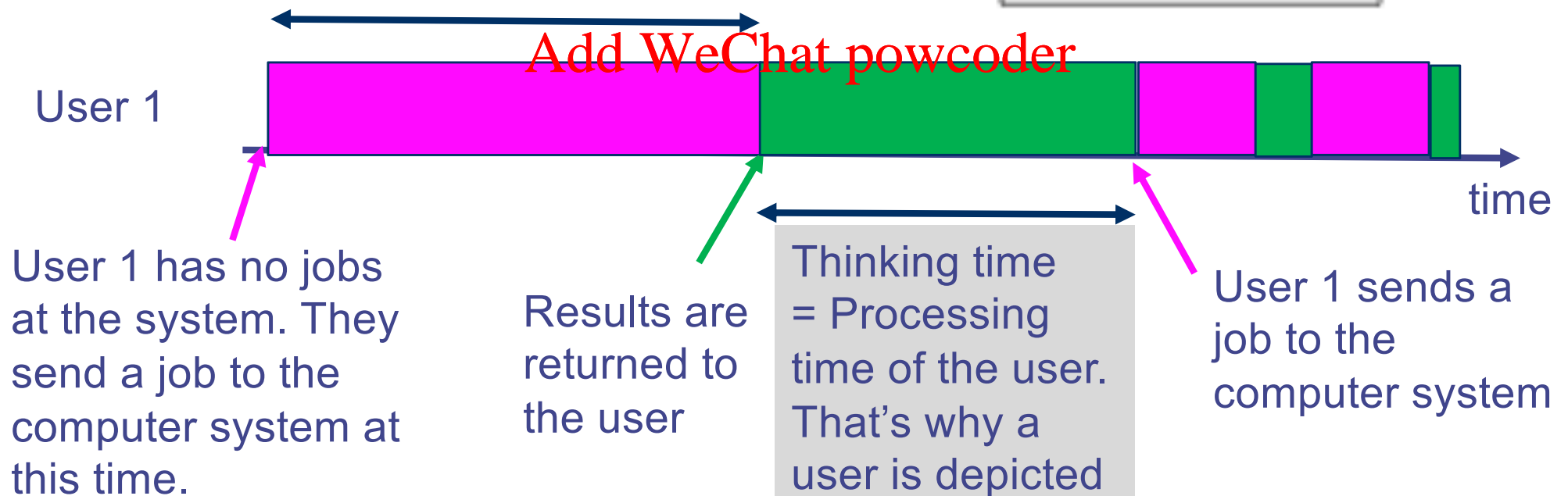
- User 1's perspective: waiting for the result of their job
- Computer system's perspective: Response time of the job from User 1



Assignment Project Exam Help

<https://powcoder.com>

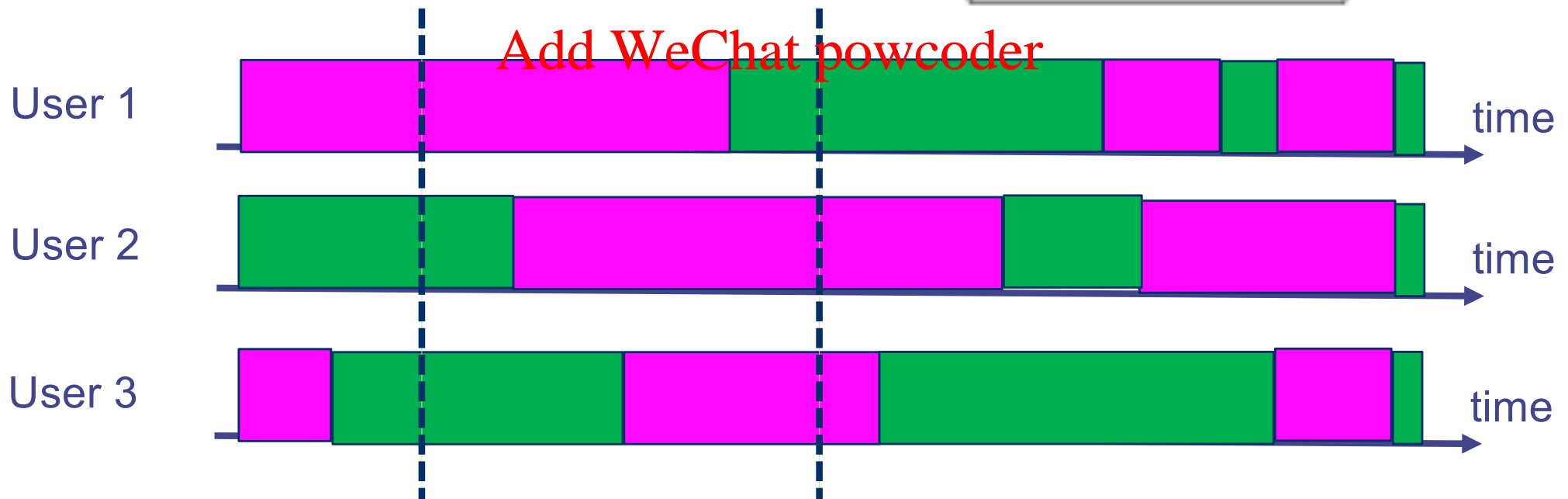
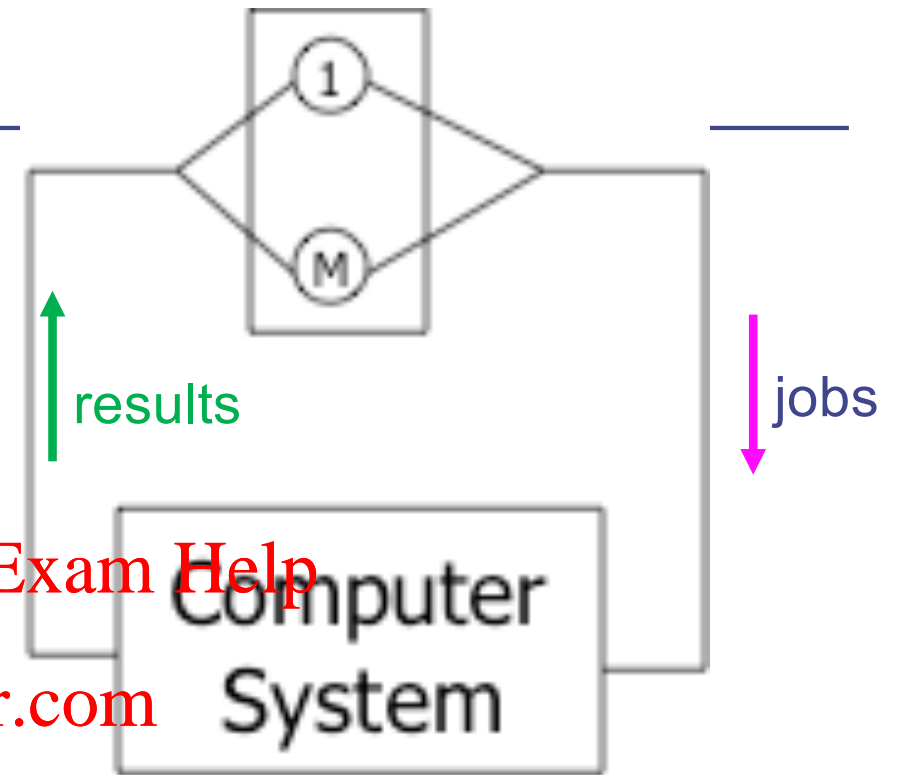
Add WeChat powcoder



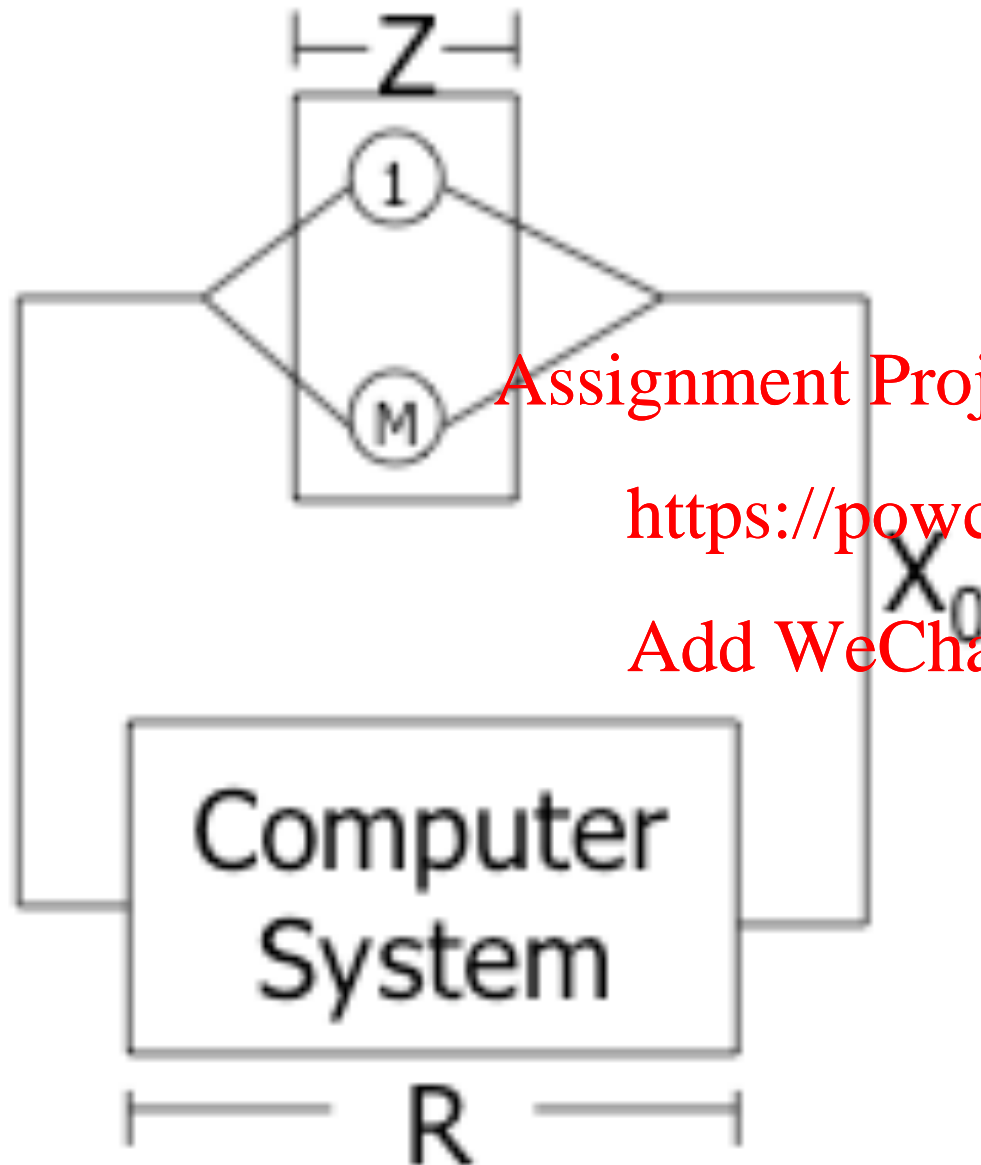
Thinking time
= Processing
time of the user.
That's why a
user is depicted
as a circle

Quiz

- Question: At any time, what is the sum of the number of busy users and the number of jobs at the computer system?



Interactive system: Parameters



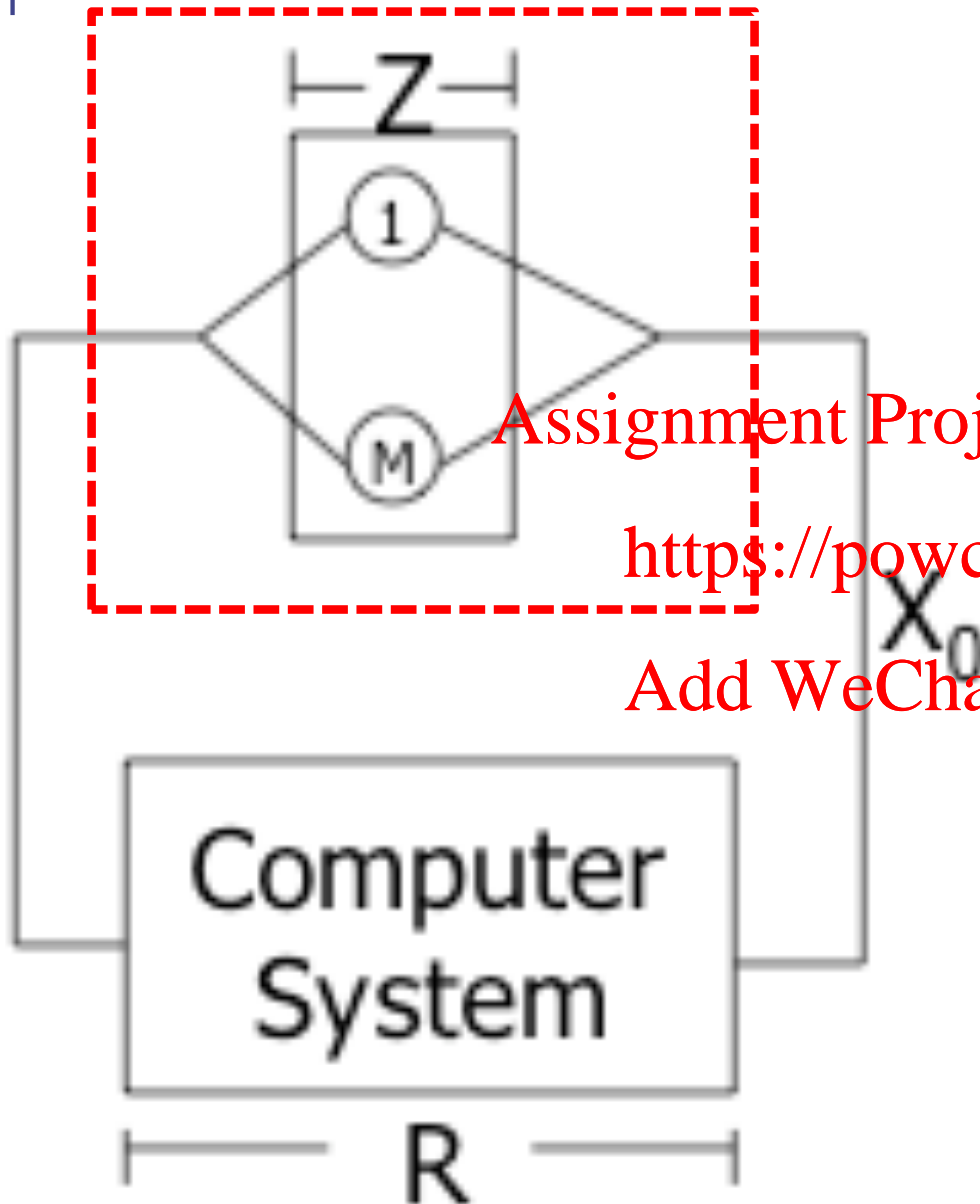
- M interactive users
- Z = mean thinking time
- R = mean response time of the computer system
- X_0 = throughput

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Analyzing interactive system: Quiz 1



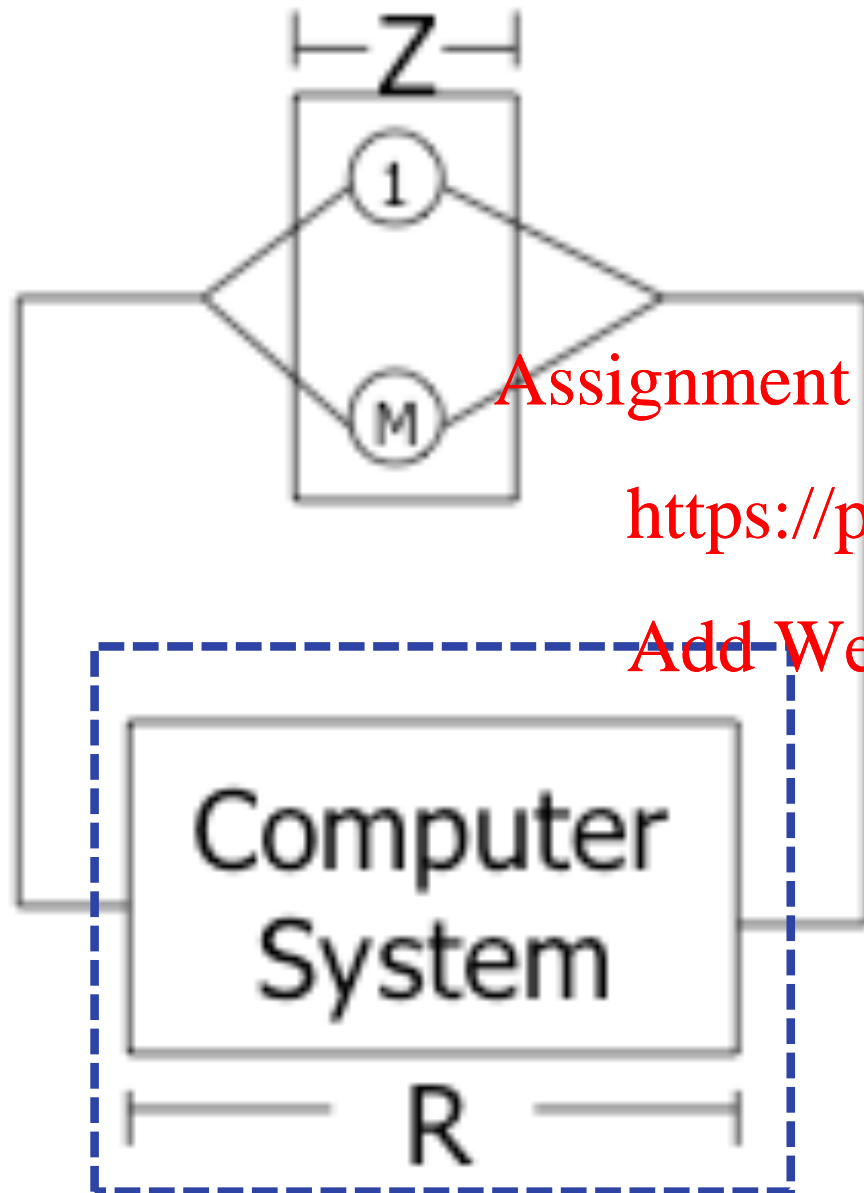
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- M_{avg} = mean # busy users
- Z = mean thinking time
- X_0 = throughput
- Apply Little's Law to the red box. What do you get?

Analyzing interactive system: Quiz 2



Assignment Project Exam Help

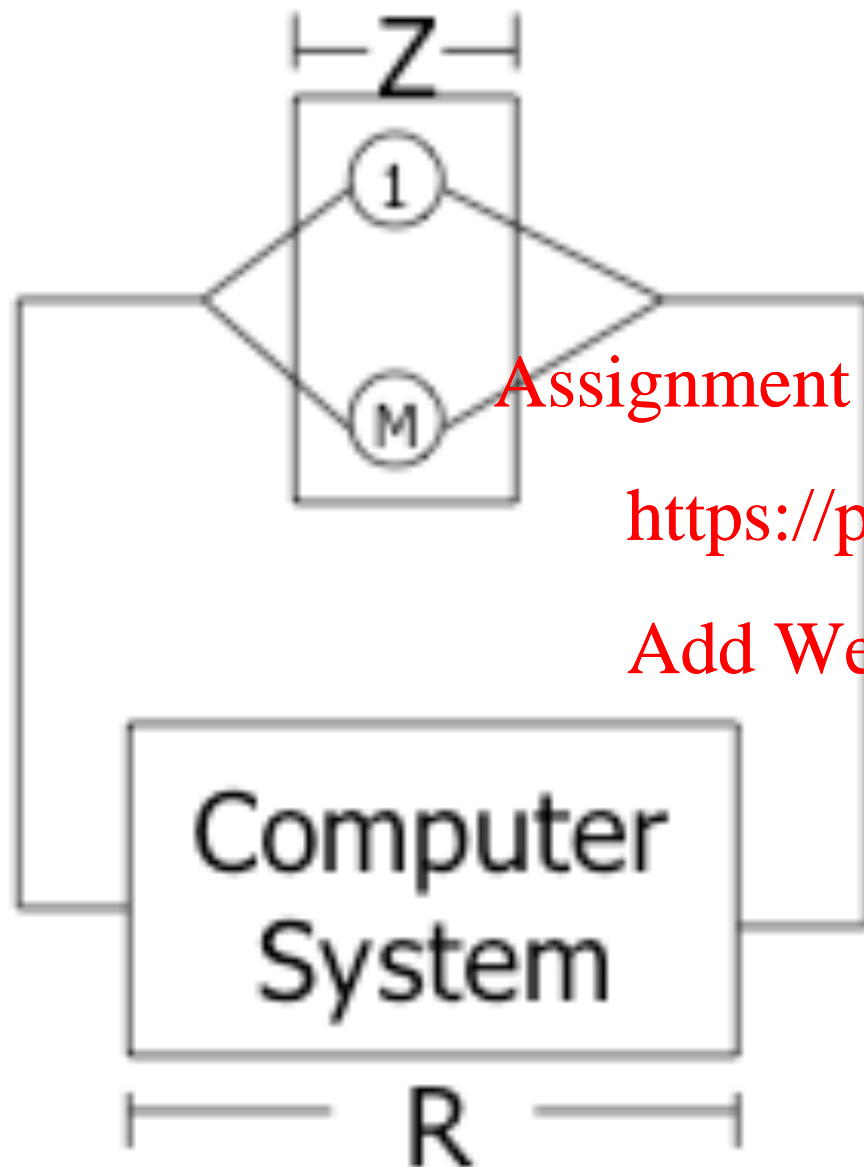
<https://powcoder.com>

Add WeChat powcoder

- N_{avg} = average # jobs in the computer system
- R = mean response time at the computer system
- X_0 = throughput
- Apply Little's Law to the computer system (i.e. the blue box), what do you get?

-

Analyzing interactive system: Quiz 3



- Quiz 1:

- Quiz 2:

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- What is $M_{avg} + N_{avg}$?

- Interactive response time law

The operational laws

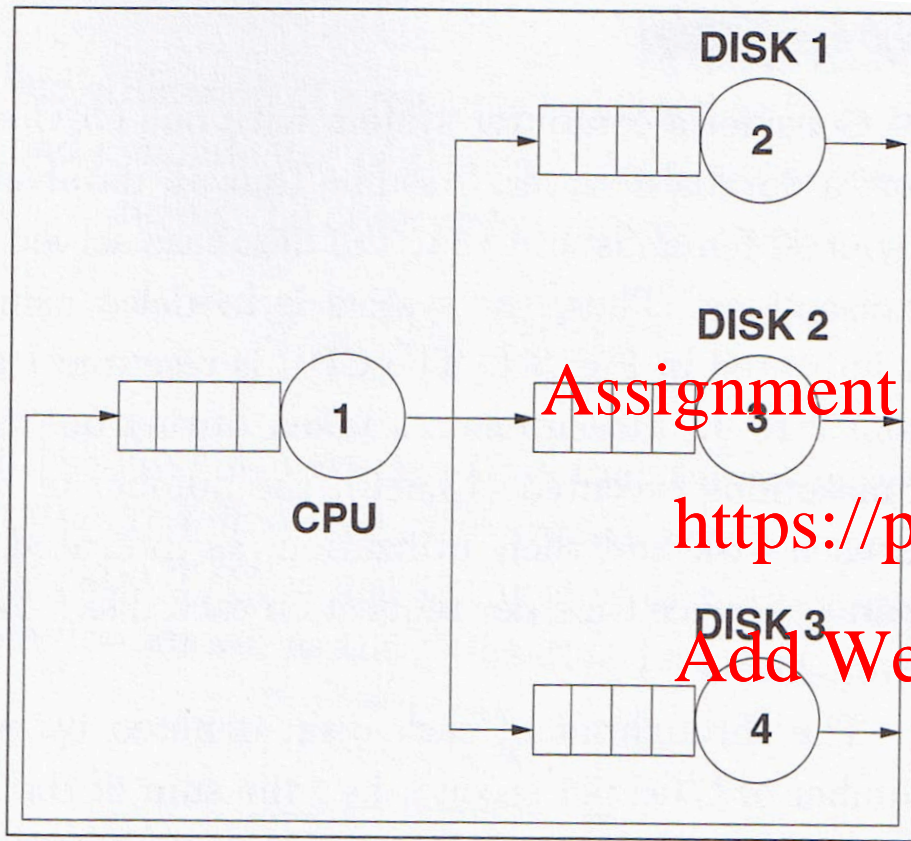
- These are the operational laws
 - Utilisation law $U(j) = X(j) S(j)$
 - Forced flow law $X(j) = V(j) X(0)$
 - Service demand law $D(j) = V(j) S(j) = U(j) / X(0)$
 - Little's law $N = X R$
 - Interactive response time $M = X(0) (R + Z)$
- Applications
 - Mean value analysis (later in the course)
 - Bottleneck analysis
 - Modification analysis

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Bottleneck analysis - motivation



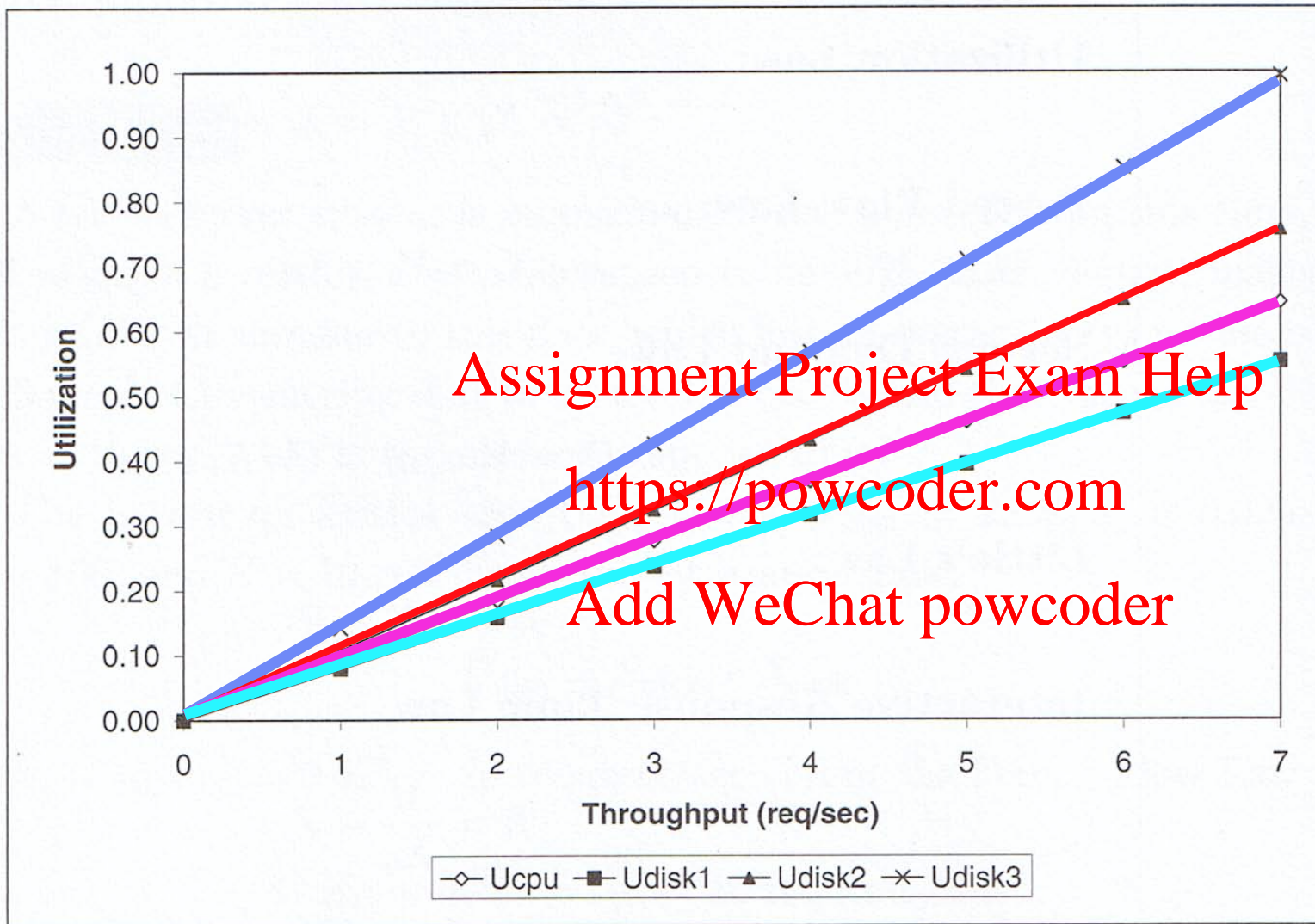
	D(j)	Utilisation
Disk 1	79ms	0.30
Disk 2	108ms	0.41
Disk 3	142ms	0.54
CPU	92ms	0.35

Service demand law: $D(j) = U(j) / X(0)$

$\Rightarrow U(j) = D(j) X(0)$

Utilisation increases with increasing throughput and service demand

Utilisation vs. throughput plot $U(j) = D(j) X(0)$



Disk 3

Disk 2

CPU

Disk 1

What determines this order?

Observation: For all system throughput:
Utilisation of Disk 3 > Utilisation of Disk 2 >
Utilisation of CPU > Utilisation of Disk 1

Bottleneck analysis

- Recall that utilisation is the busy time of a device divided by measurement time
 - What is the maximum value of utilisation?
- Based on the example on the previous slide, which device will reach the maximum utilisation first?

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Bottleneck (1)

- Disk 3 has the highest service demand
- It is the bottleneck of the whole system

Assignment Project Exam Help

Operational law:

$$X(0) = \frac{U(j)}{D(j)}$$

Utilisation limit:

$$U(j) \leq 1$$

$$X(0) \leq \frac{1}{D(j)}$$

Bottleneck (2)

$$X(0) \leq \frac{1}{D(j)}$$

Should hold for all K devices in the system

Assignment Project Exam Help

$$i.e. X(0) \leq \frac{1}{D(1)}, \dots, X(0) \leq \frac{1}{D(K)}$$

<https://powcoder.com>

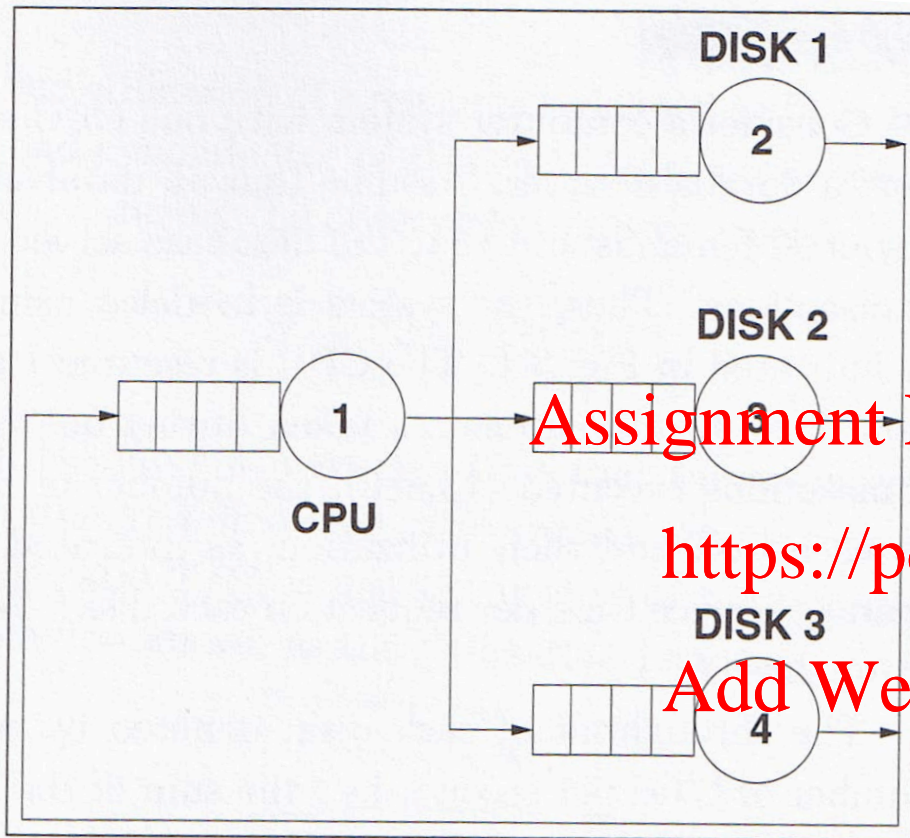
Add WeChat powcoder

$$\Rightarrow X(0) \leq \min \frac{1}{D(j)}$$

$$\Rightarrow X(0) \leq \frac{1}{\max D(j)}$$

Bottleneck throughput is limited by the maximum service demand

Bottleneck exercise



	D(j)	Utilisation
Disk 1	79ms	0.30
Disk 2	108ms	0.41
Disk 3	142ms	0.54
CPU	92ms	0.35

The system throughput is upper bounded by $\frac{1}{0.142} = 7.04$ jobs/s

If we upgrade Disk 3 by a new disk which is 2 times faster, which device will be the bottleneck after the upgrade? You can assume that service time is inversely proportional to disk speed.

Another throughput bound

- Little's law

$$N = R \times X(0) \geq \left(\sum_{i=1}^K D_i \right) \times X(0)$$

Assignment Project Exam Help

$$\Rightarrow X(0) \leq \frac{N}{\sum_{i=1}^K D_i}$$

Add WeChat powcoder

Previously, we have

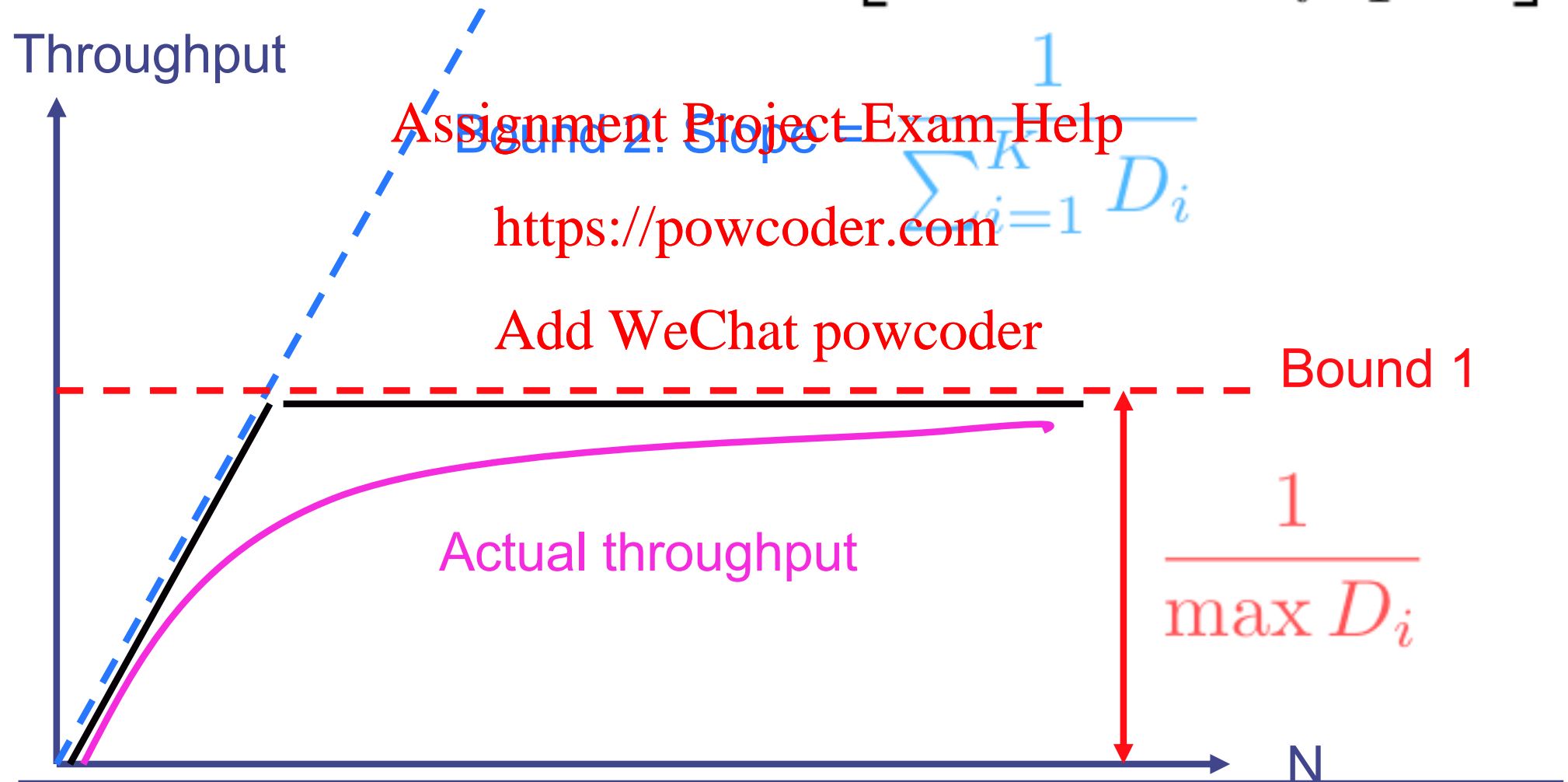
$$X(0) \leq \frac{1}{\max D(j)}$$

Therefore:

$$X(0) \leq \min \left[\frac{1}{\max D_i}, \frac{N}{\sum_{i=1}^K D_i} \right]$$

Throughput bounds

$$X(0) \leq \min \left[\frac{1}{\max D_i}, \frac{N}{\sum_{i=1}^K D_i} \right]$$



Bottleneck analysis

- Simple to use
 - Needs only utilisation of various components
- Assumes service demand is load independent

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Modification analysis (1)

- (Reference: Lazowska Section 5.3.1)
- A company currently has a system (3790) and is considering switching to a new system (8130). The service demands for these two systems are given below:

System	Service demand (seconds)	
	CPU	Disk
3790	4.6	4.0
8130	5.1	1.9

- The company uses the system for interactive application with a think time of 60s.
- Given the same workload, should the company switch to the new system?
- Exercise: Answer this question by using bottleneck analysis. For each system, plot the upper bound of throughput as a function of the number of interactive users.

Modification analysis (2)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Operational analysis

- Operational analysis allows you to bound the system performance but it does NOT allow you to find the throughput and response time of a system
- To order to find the throughput and response time, we need to use queueing analysis
- To order to use queueing analysis, we need to specify the workload

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Workload analysis

- Performance depends on workload
 - When we look at the performance bound earlier, the bounds depend on **number of users** and **service demand**
 - Queue response time depends on the **job arrival probability distribution** and **job service time distribution**
 - Recall from Lecture 1A:
 - Uniform arrival times and uniform processing times result in zero waiting time
 - But non-uniform distributions give non-zero waiting time
- Need to specify workload by using probability distribution.
- We will look at a well-known arrival process called Poisson process today.

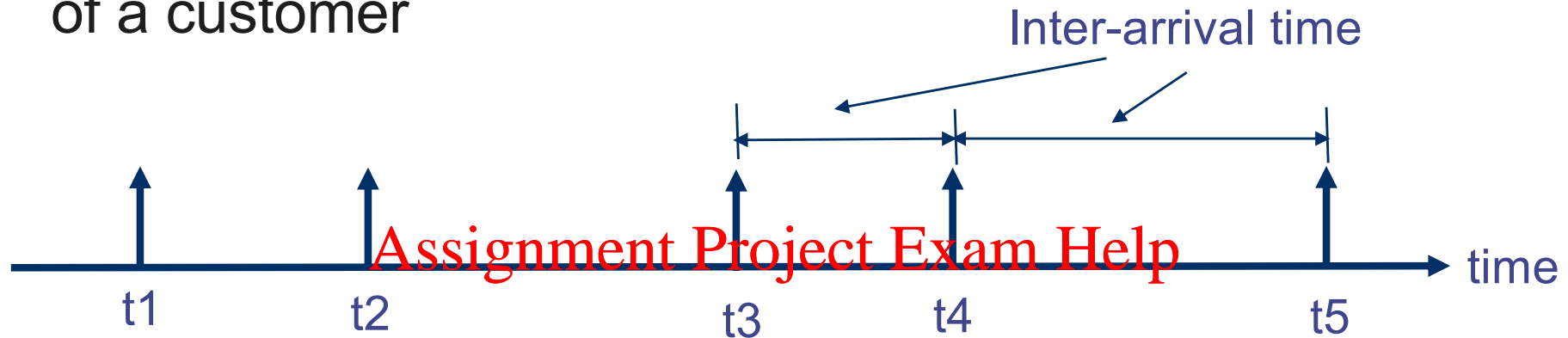
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Arrival process

- Each vertical arrow in the time line below depicts the arrival of a customer



Assignment Project Exam Help

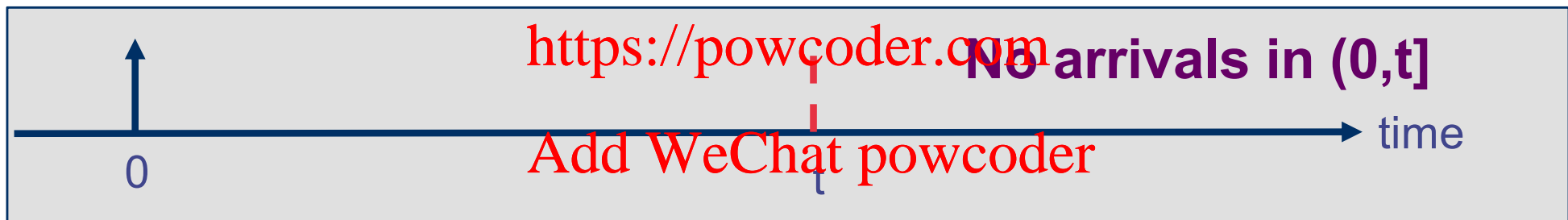
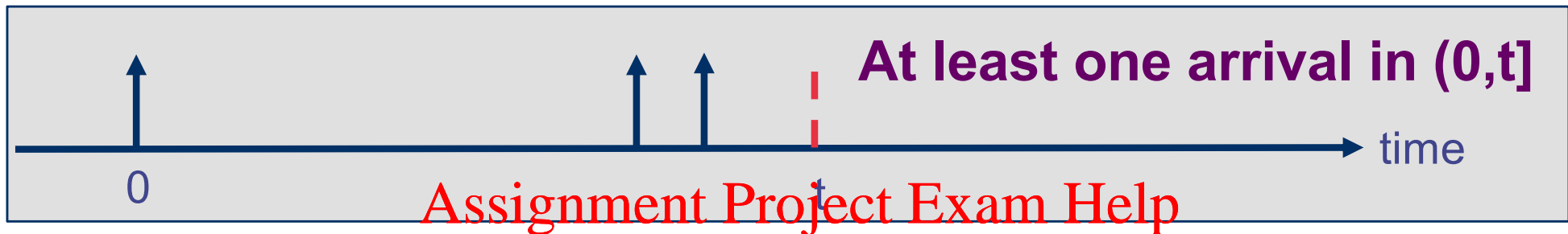
<https://powcoder.com>

Add WeChat powcoder

- An arrival can mean
 - A telephone call arriving at a call center
 - A transaction arriving at a computer system
 - A customer arriving at a checkout counter
 - An HTTP request arriving at a web server
- The **inter-arrival time** distribution will impact on the response time.
- We will study an inter-arrival distribution that results from a large number of **independent** customers.

Describing arrivals probabilistically

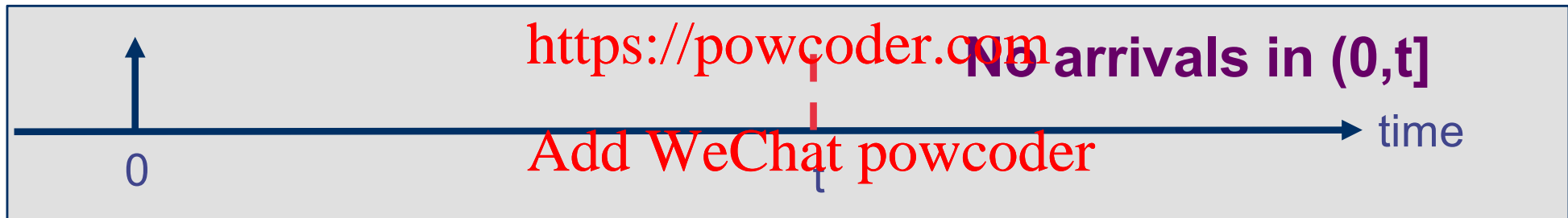
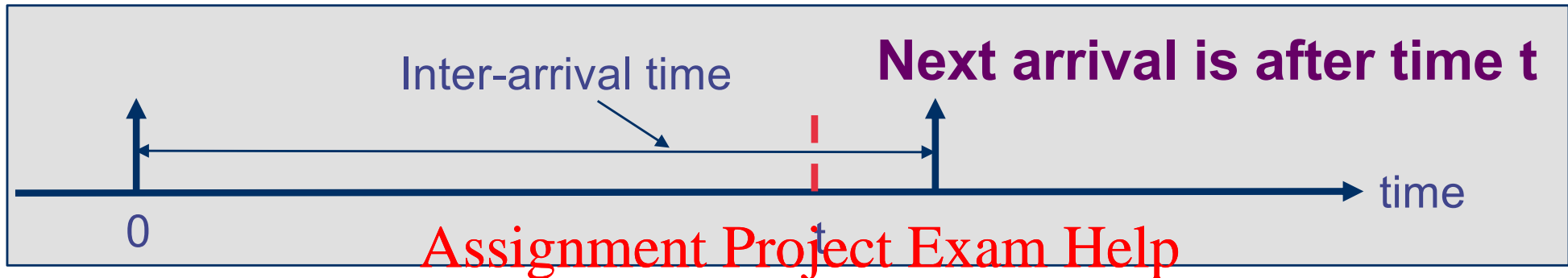
- Assume a customer arrives at time 0



- Quiz: What is the relation between the following two probabilities?
 - Prob[at least one arrival in $(0,t]$]
 - Prob[no arrivals in $(0,t]$]
- Answer:
- Moral: "No arrivals" is not boring, it tells you something

Inter-arrival probability

- Assume a customer arrives at time 0

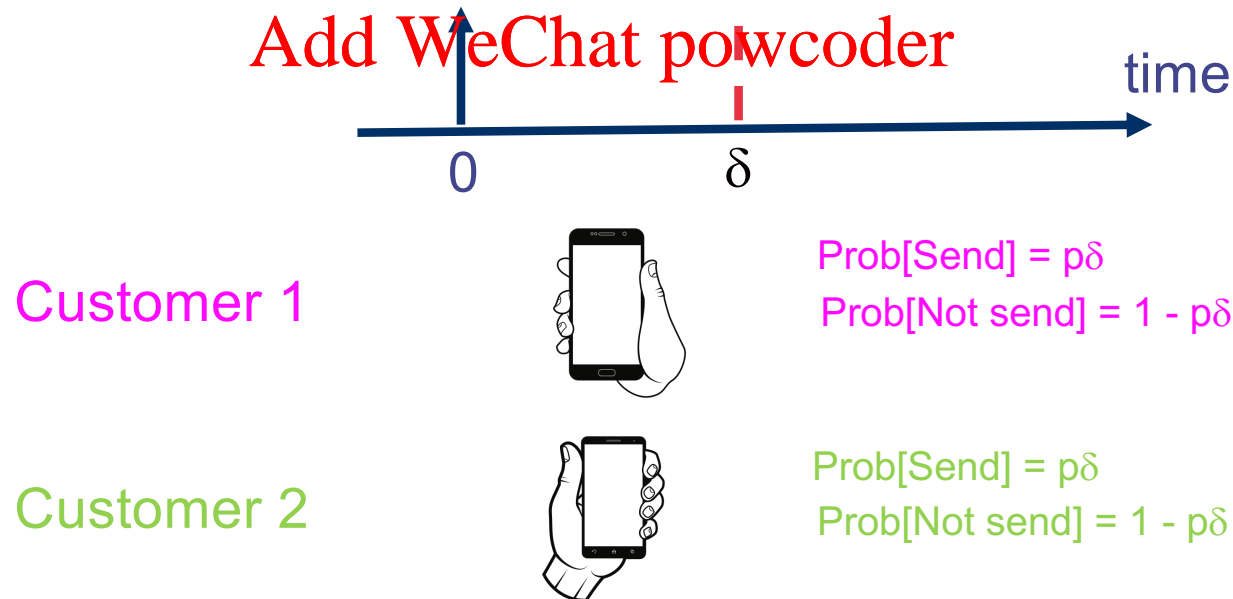


- Quiz: What is the relation between the following two probabilities?
 - Prob[Inter-arrival time is $\geq t$]
 - Prob[no arrivals in $(0, t]$]
- Answer:
- Next step: Find Prob[no arrivals in $(0, t]$] for independent customers

Many independent arrivals (1)

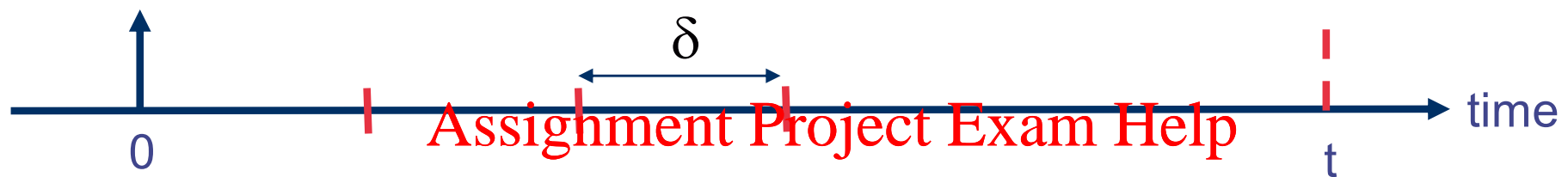
- Problem set up:
 - An arrival at time 0
 - A large pool of N independent customers
 - Behaviour of each customer: Within a small time interval of δ , a customer sends a request (or arrives) with a probability of $p\delta$
 - p is a constant


- Quiz: If there are 2 ($= N$) customers, what is the probability that both of them do not send any request in the time interval δ
 - Answer:



Many independent arrivals (2)

- Aim: Want to find the probability of no arrivals in $(0, t]$
- Divide the time t into intervals of width δ



- No arrival in $(0, t]$ = no arrival in each interval δ from N users
- Probability of no arrival in $\delta =$ 
- There are t / δ intervals
- Probability of no arrival in $(0, t]$ is

$$(1 - Np\delta)^{\frac{t}{\delta}} \rightarrow e^{-Npt} \text{ as } \delta \rightarrow 0$$

Exponential inter-arrival time

- We have showed

$$\text{Probability(no arrival in } (0,t]) = \exp(-Npt)$$

- Probability(inter-arrival time $> t$) = $\exp(-Npt)$

Assignment Project Exam Help

- This means

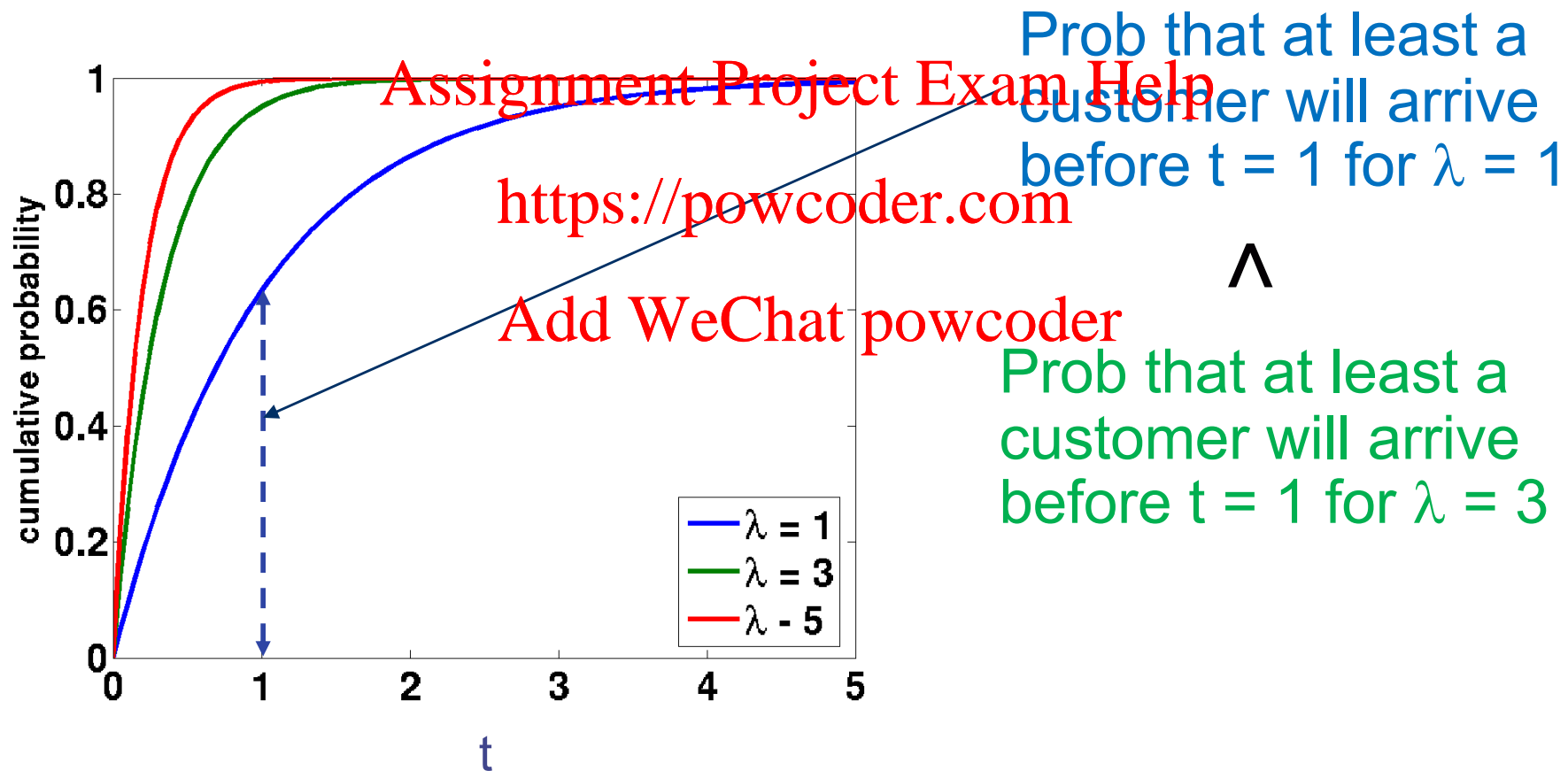
$$\text{Probability(inter-arrival time } \leq t) = 1 - \exp(-Npt)$$

Add WeChat powcoder

- What this shows is the inter-arrival time distribution for independent arrival is exponentially distributed
- Define: $\lambda = Np$
 - λ is the mean arrival rate of customers

Exponential distribution - cumulative distribution

- Cumulative distribution of inter-arrival time with customer arrival rate λ
 - $\text{Prob}(\text{inter-arrival time} \leq t) = 1 - \exp(-\lambda t)$



Exponential distribution

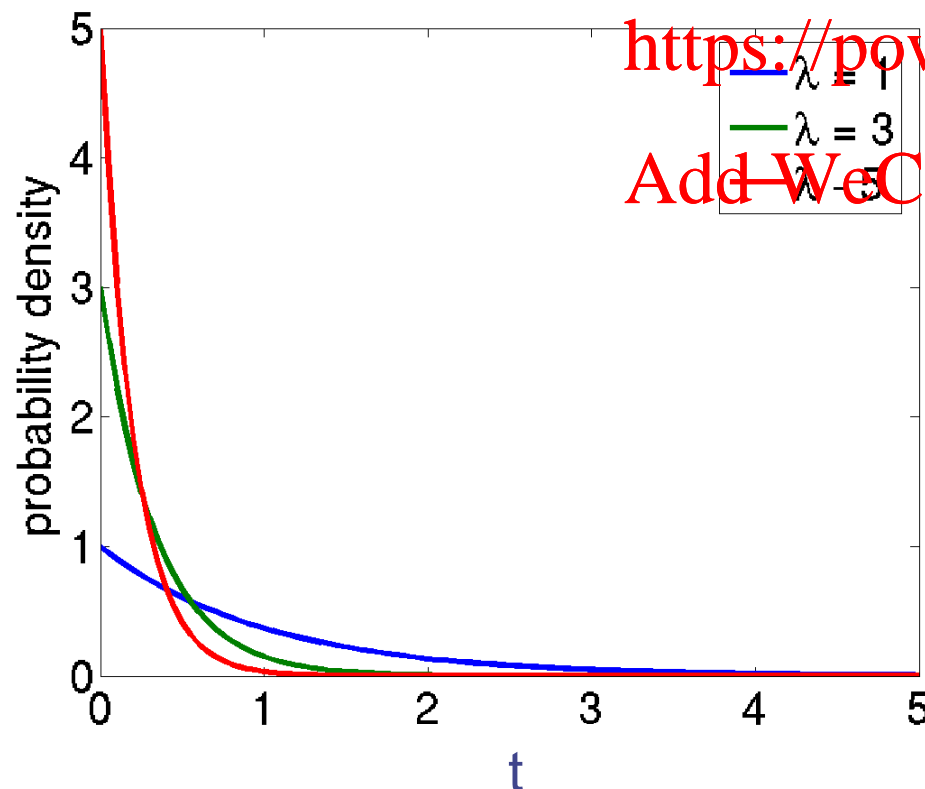
- A continuous random variable is exponentially distributed with rate λ if it has probability density function

$$f(t) = \begin{cases} \lambda \exp(-\lambda t) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Reminder:

Probability
density
function

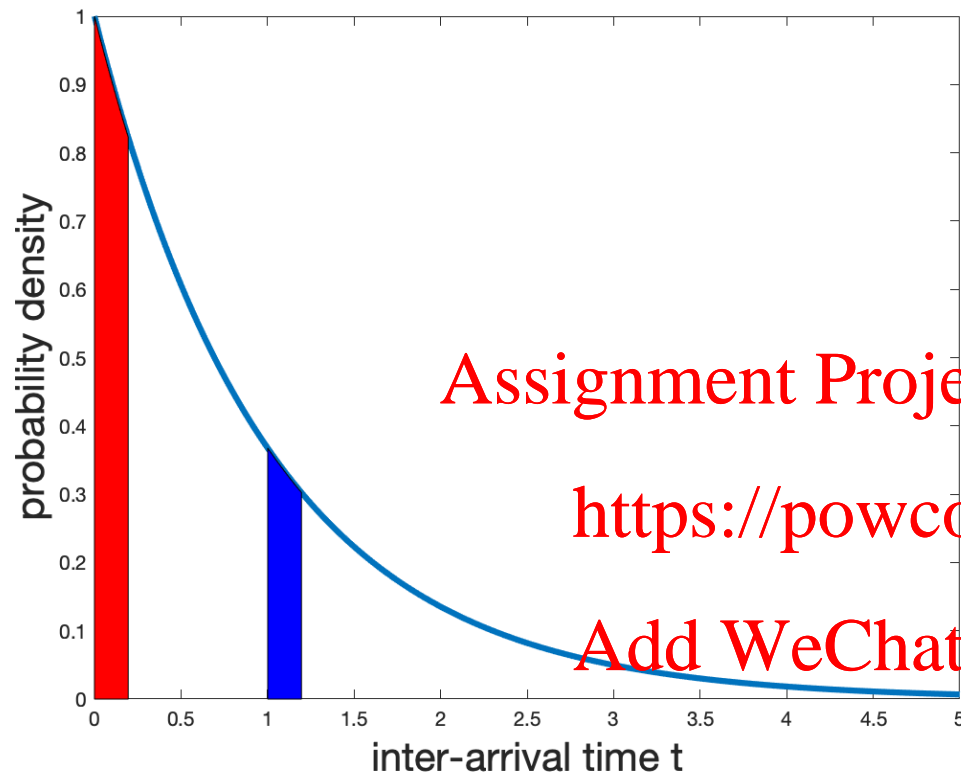
integration

differentiation

Cumulative
probability

$$1 - \exp(-\lambda t)$$

Probability density function (PDF)



Reminder: PDF $f(t)$

Probability($t \leq T \leq t + \delta t$)

$= f(t) \delta t$

Assignment Project Exam Help

<https://powcoder.com>

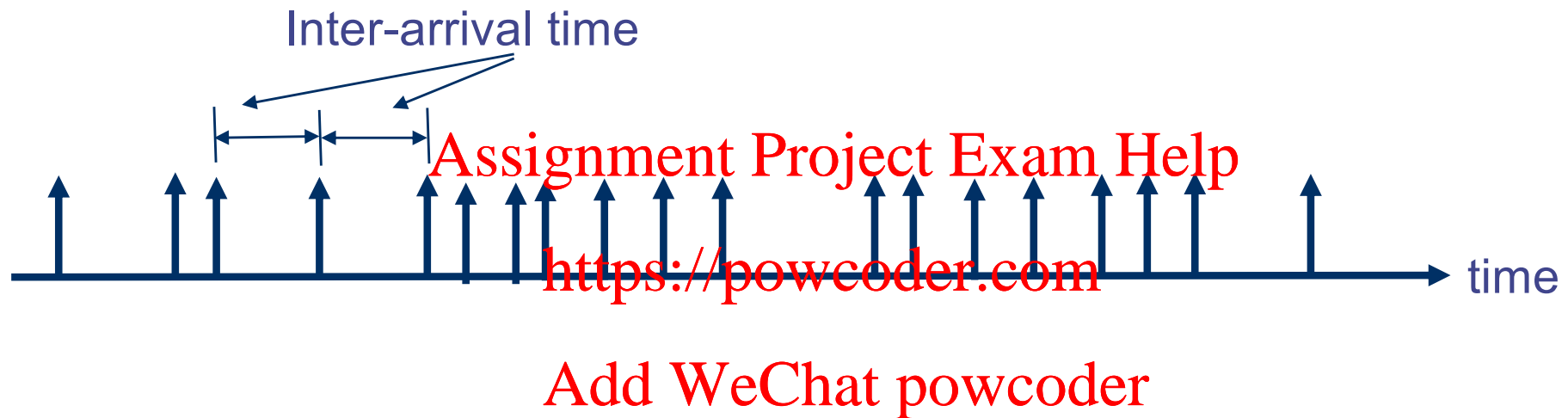
Add WeChat powcoder

Red area = probability
that inter-arrival time is in
the interval $[0, 0.2]$

Blue area = probability
that the inter-arrival is in
the interval $[1, 1.2]$

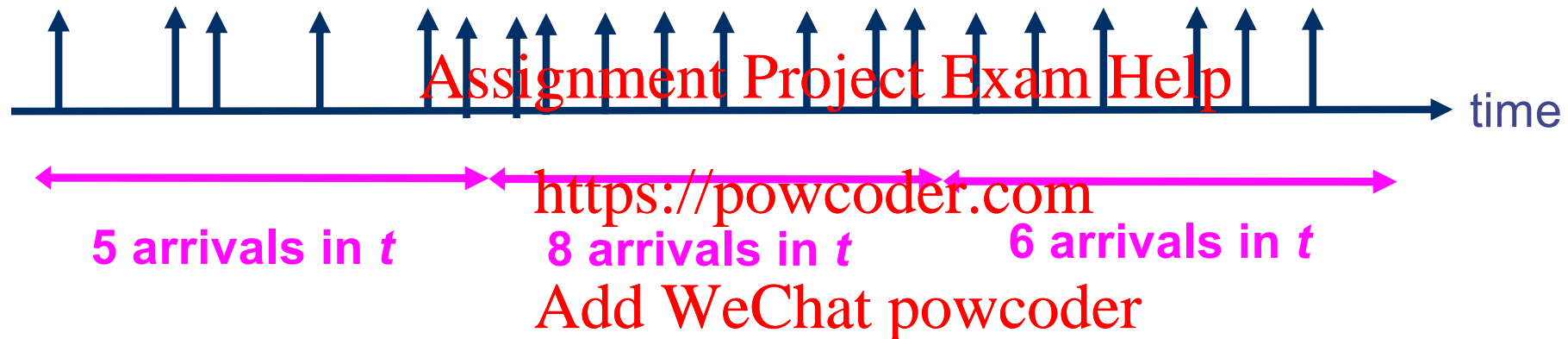
Two different methods to describe arrivals

Method 1: Continuous probability distribution of inter-arrival time



Two different methods to describe arrivals

Method 2: Use a fixed time interval (say t), and count the number of arrivals within t .



- The number of arrivals in t is random
- The number of arrivals must be a non-negative integer
- We need a discrete probability distribution:
 - $\text{Prob}[\text{\#arrivals in } t = 0]$
 - $\text{Prob}[\text{\#arrivals in } t = 1]$
 - etc.

Poisson process (1)

- Definition: An arrival process is Poisson with parameter λ if the probability that n customer arrive in any time interval t is

$$\frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

Assignment Project Exam Help

<https://powcoder.com>

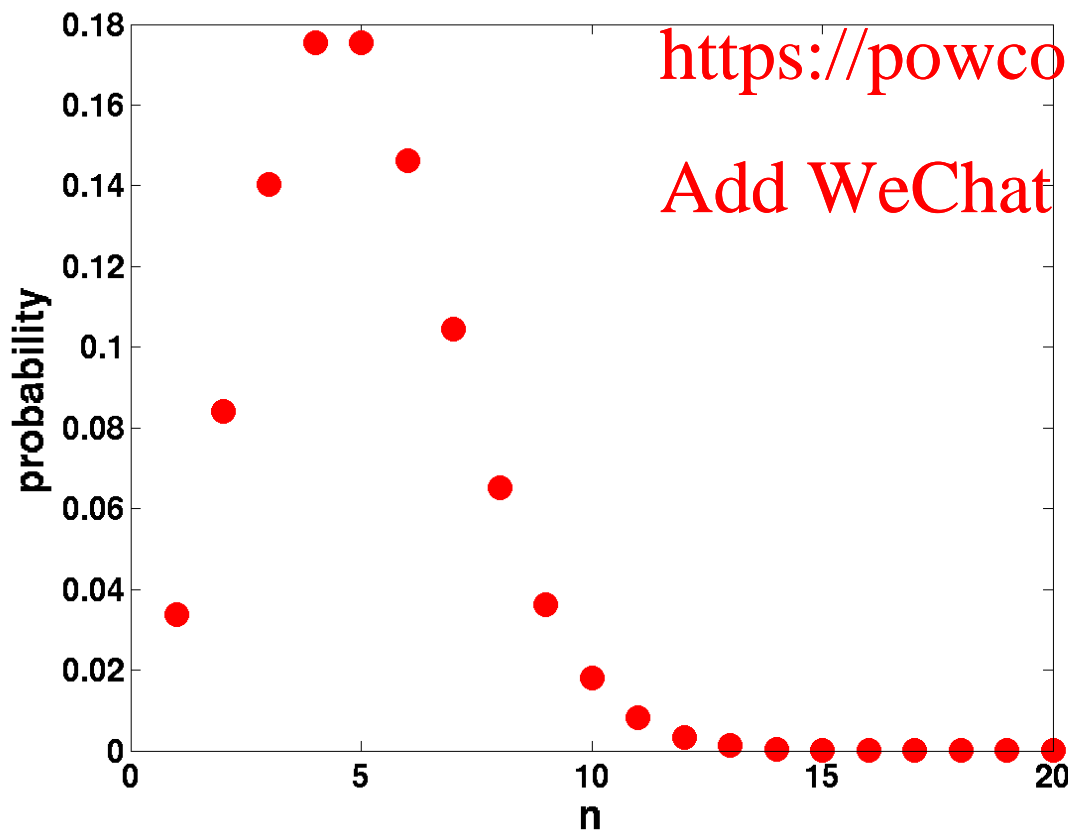
Add WeChat powcoder

Example:

Example:

$\lambda = 5$ and $t = 1$

Note: Poisson is a discrete probability distribution.



Poisson process (2)

- Theorem: An exponential inter-arrival time distribution with parameter λ gives rise to a Poisson arrival process with parameter λ
- How can you prove this theorem?
 - A possible method is to divide an interval t into small time intervals of width δ . A finite δ will give a binomial distribution and with $\delta \rightarrow 0$, we get a Poisson distribution.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Customer arriving rate

- Given a Poisson process with parameter λ , we know that the probability of n customers arriving in a time interval of t is given by:

$$\frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

Assignment Project Exam Help

- What is the mean number of customers arriving in a time interval of t ?

<https://powcoder.com>
Add WeChat powcoder

$$\sum_{n=0}^{\infty} n \frac{(\lambda t)^n e^{-\lambda t}}{n!} = \lambda t$$

- That's why λ is called the arrival rate.

Customer inter-arrival time

- You can also show that if the inter-arrival time distribution is exponential with parameter λ , then the mean inter-arrival time is $1/\lambda$
- Quite nicely, we have

Mean arrival rate = $1/\text{mean inter-arrival time}$

<https://powcoder.com>

Add WeChat powcoder

Application of Poisson process

- Poisson process has been used to model the arrival of telephone calls to a telephone exchange successfully
- Queueing networks with Poisson arrival is tractable
 - We will see that in the next few weeks.
- Beware that not all arrival processes are Poisson! Many arrival processes we see in the Internet today are not Poisson. We will see that later.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

References

- Operational analysis
 - Lazowska et al, Quantitative System Performance, Prentice Hall, 1984.
(Classic text on performance analysis. Now out of print but can be download from <http://www.cs.washington.edu/homes/lazowska/qsp/>)
 - Chapters 3 and 5 (For Chapter 5, up to Section 5.3 only)
 - Alternative 1: You can read Menasce et al, “Performance by design”, Chapter 3. Note that Menasce doesn't cover certain aspects of performance bounds. So, you will also need to read Sections 5.1-5.3 of Lazowska.
 - Alternative 2: You can read Harcol-Balter, Chapters 6 and 7. The treatment is more rigorous. You can gross over the discussion mentioning ergodicity.
- Poisson process: Harcol-Balter Chapter 11