Final Review

Lecturer:
萧逸寒

COMP9334
Capacity Planning

考试注意事项

开卷考试

- 开卷的唯一意义：避免遗忘或记错公式
- 不要现场理解或推导公式
- 不要把课件全部打印后带入考场
- 不会做的题不要空着

Review Outline

1. Operational analysis
2. Queue model:
    1. M/M/1 M/M/m M/M/m/m+k
    2. M/G/1 residual service time
    3. Priority queue
3. Closed queue: Markov-Chain analysis/MVA
4. Simulation
5. Integer Programming

Operational analysis

Forced Flow Law

$$V(j) = X(j)/X(0)$$

Utilization Law

$$U(j) = S(j) * X(j)$$

Service Demand Law

$$D(j) = U(j)/X(0)$$

Little’s Law

$$N = X * R$$

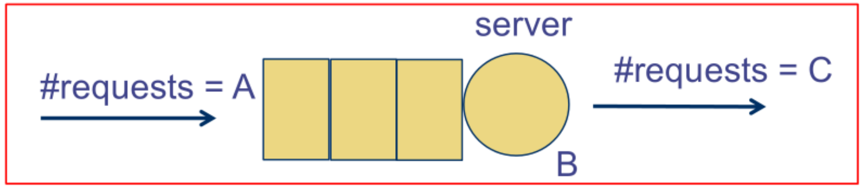
Bottleneck analysis

$$X(0) \leq \min\left\{\frac{1}{\max D_i}, \frac{N}{\sum D_i}\right\}$$

Operational analysis

描述系统性能所需的变量：

A = 到达任务的数量 #arrivals
B = 忙碌时间 busy time
C = 完成任务的数量 #complete
S = 服务时间 service time
T = 总时间 total observation time



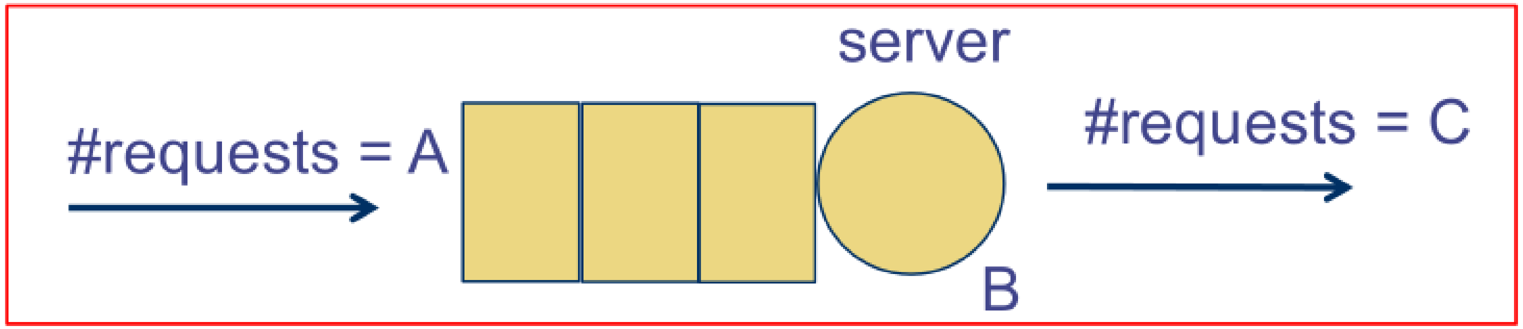
λ = 到达速率 arrival rate = A / T
U = 占用率 utilization = B / T
X = 输出速率 throughput = C / T

X(j) = 设备j的输出速率

X(0) = 整个系统的输出速率

标号 *j* 表示设备 device
标号 0 表示系统 system

Operational analysis



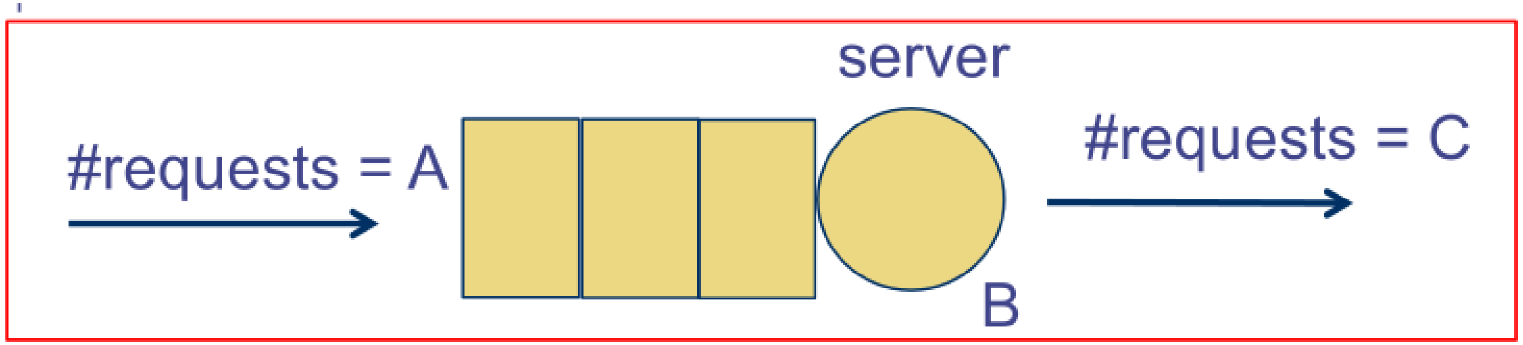
Forced Flow Law

$$V(j) = X(j)/X(0)$$

描述了单个设备的访问速率

* 一个任务有可能访问设备多次！所以通常 $X(j) \neq X(0)$

Operational analysis



Utilization Law

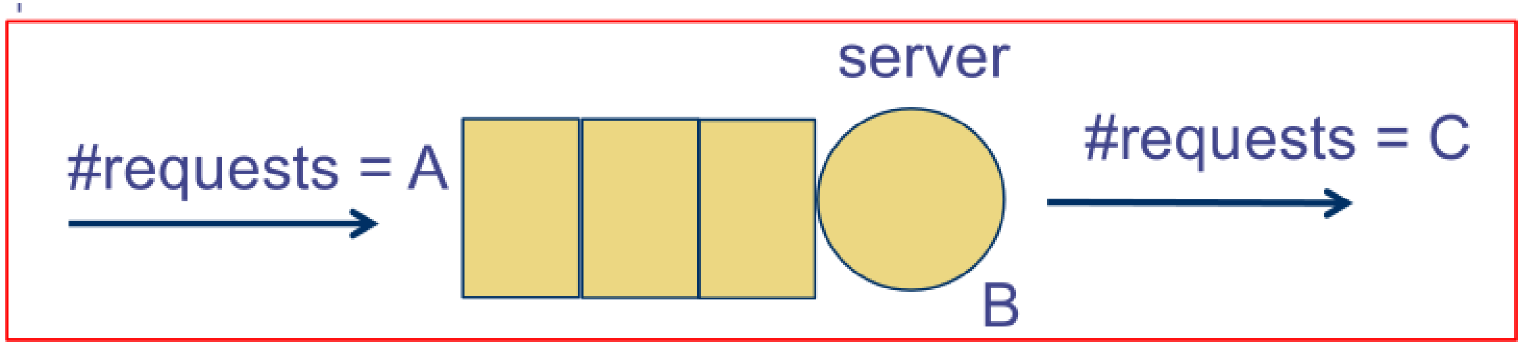
$$U(j) = S(j) * X(j)$$

描述了单个设备的占用率 记住U的基本定义： $U(j) = B(j)/T$

$$U(j) = \frac{B(j)}{T} = \frac{B(j)}{\mathrm{C}} * \frac{\mathrm{C}}{T} = S(j) * X(j)$$

* U也可以理解为设备有多少概率被占用（解题时十分有用）

Operational analysis



Service Demand Law

$$D(j) = U(j)/X(0)$$

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

描述了单个设备处理任务所需的时间（注意单位是时间）

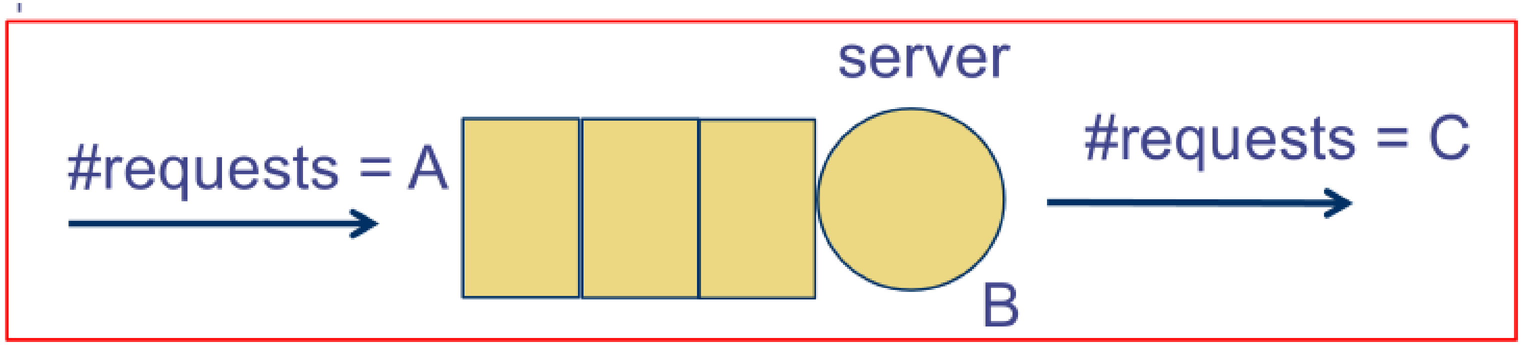
Tips：

S(j)为每次访问设备所需的时间
V(j)为每个任务需要访问设备的次数

所以D(j)等于每个任务访问设备所需的总时间

* D(j)和S(j)的实际物理意义有什么关系？

Operational analysis



Little’s Law

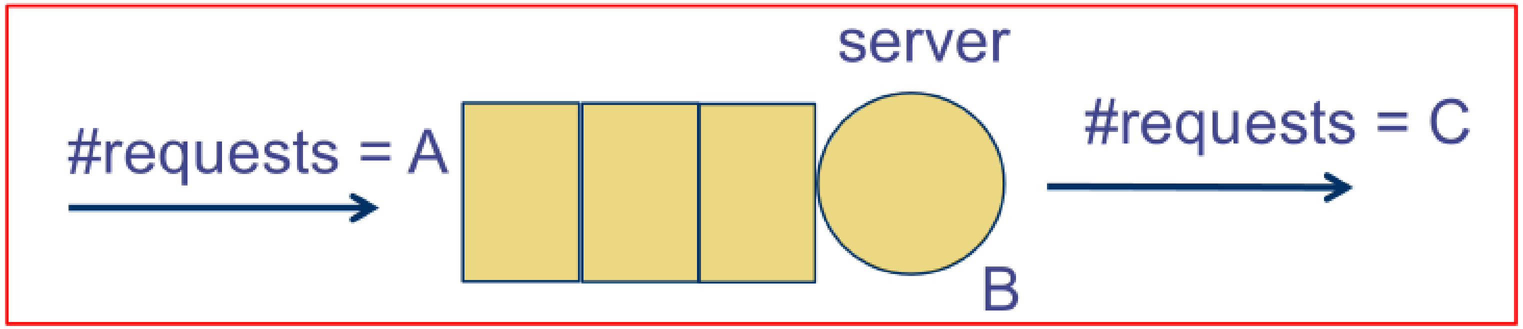
$$N = X * R$$

任意时刻系统内的任务数量N = 系统的输出速率X * 每个任务所需的响应时间R

* 响应时间response time和服务时间service time有什么关系?

Response time = service time + waiting time

Operational analysis



Little’s Law

$$N = X * R$$

从公式形式上来看，Little’s Law和哪个公式很相似?

对比一下 Utilization Law：

$$U(j) = X(j) * S(j)$$

Tips：

可以发现Utilization Law其实是Little’s Law的一种特殊形式，两者是等价的。

U可以视为针对server（除开队列）使用Little’s Law，所以U也相当于描述了server当中有多少任务。

Operational analysis

结合之前我们讨论过的情况，Utilization的实质是什么?

1. 在设备中的任务数量
2. 设备被占用的概率
3. 设备的忙碌时间占总时间的比例

为什么这些量居然可以等价?

假设 S = 2s, X = 0.2/s

U = SX = 0.4

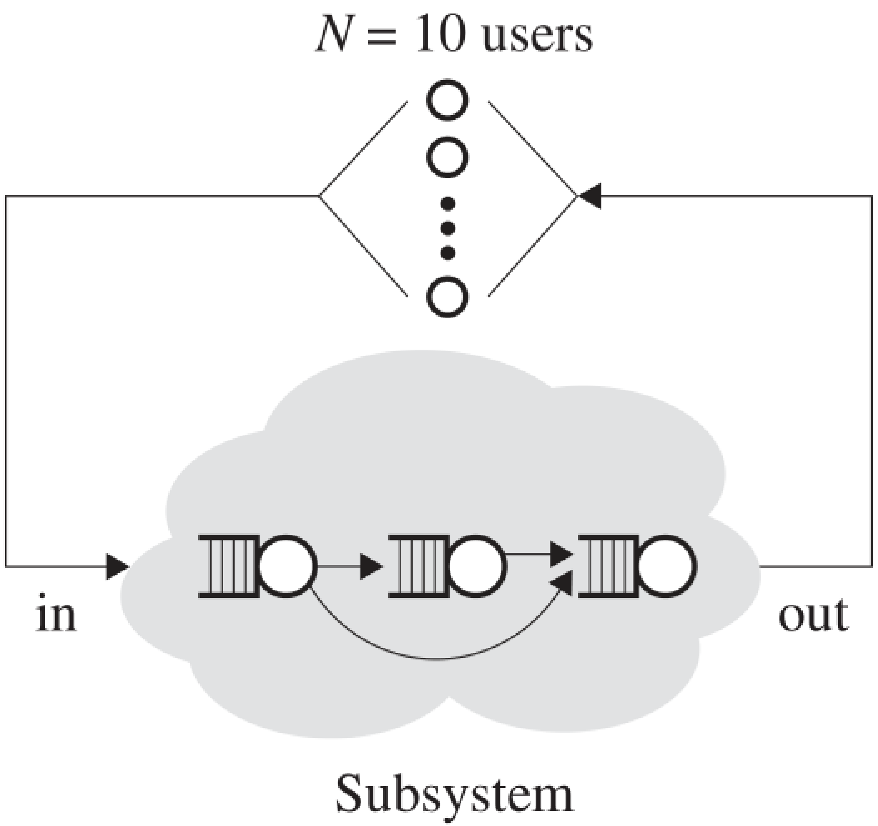
这个0.4可以理解为设备被占用的比例是0.4，也即有40%的可能被占用

$$\frac{1}{S}=\frac{1}{2}=0.5\ \ and\ \ \frac{0.2}{0.5}=0.4$$

同时也说明了设备中有0.4个任务

（假设设备满负荷运行，任意时刻其中最多也只能有1个任务）

Little’s Law practice: Interaction system - 1



We have an interactive system with N = 10 users, as shown in Figure. We are told that the expected think time is E[Z] = 5 seconds and that the expected response time is E[R] = 15 seconds.

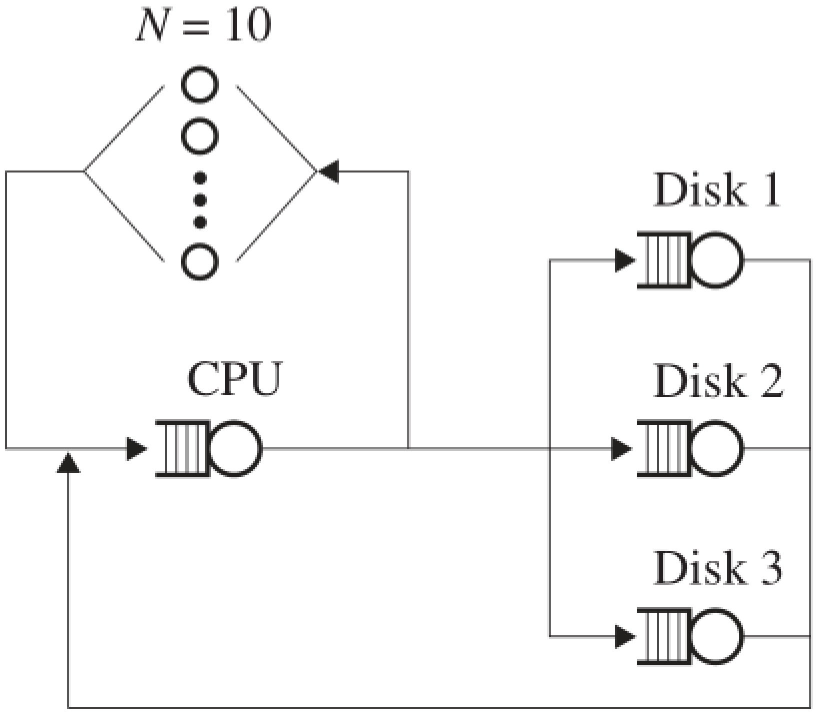
Question: What is the throughput, X , of the system?

$$N = X \cdot \mathbf{E}[T] = X(\mathbf{E}[Z] + \mathbf{E}[R])$$

$$\Rightarrow X = \frac{N}{\mathbf{E}[R] + \mathbf{E}[Z]} = \frac{10}{5 + 15} = 0.5 \text{ jobs/sec.}$$

Little's Law practice: Interaction system - 2

- The throughput of disk 3 is 40 requests/sec ($X_{\text{disk3}} = 40$).
- The service time of an average request at disk 3 is 0.0225 sec ($\mathbf{E}[S_{\text{disk3}}] = .0225$).
- The average number of jobs in the system consisting of disk 3 and its queue is 4 ($\mathbf{E}[N_{\text{disk3}}] = 4$).

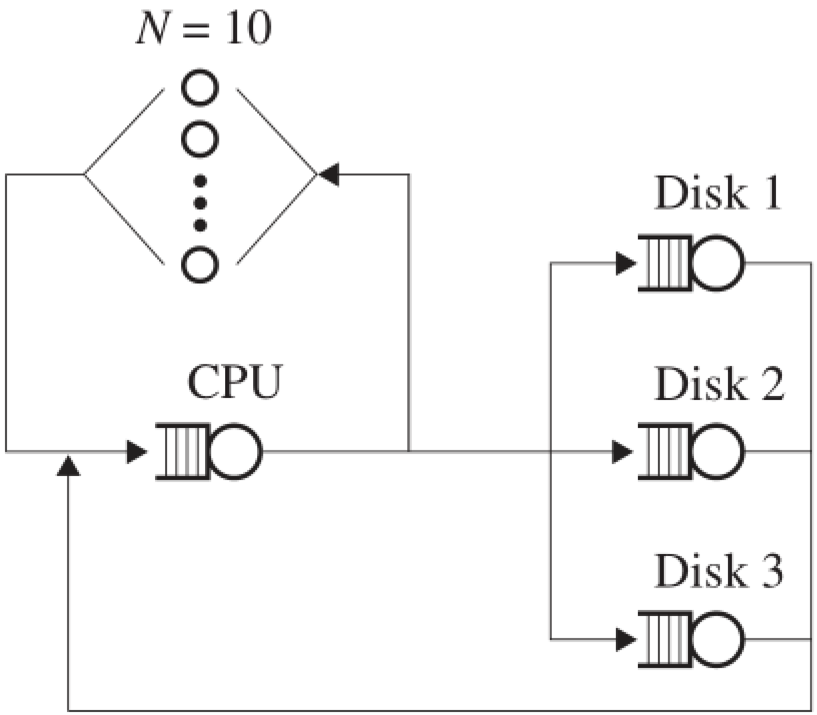


Question: What is the utilization of disk 3?

$$\rho_{\text{disk3}} = X_{\text{disk3}} \cdot \mathbf{E}[S_{\text{disk3}}] = 40 \cdot (0.0225) = 90\%.$$

Little's Law practice: Interaction system - 2

- The throughput of disk 3 is 40 requests/sec ($X_{\text{disk3}} = 40$).
- The service time of an average request at disk 3 is 0.0225 sec ($\mathbf{E}\left[S_{\text{disk3}}\right] = .0225$).
- The average number of jobs in the system consisting of disk 3 and its queue is 4 ($\mathbf{E}\left[N_{\text{disk3}}\right] = 4$).



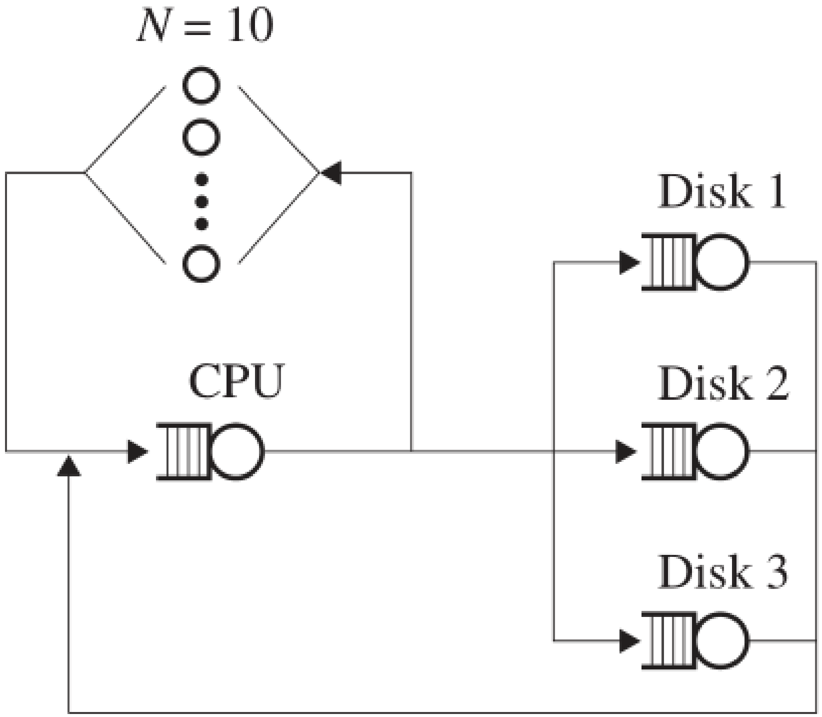
Question: What is the mean time spent queueing at disk 3?

$$\mathbf{E}\left[T_{\text{disk3}}\right] = \frac{\mathbf{E}\left[N_{\text{disk3}}\right]}{X_{\text{disk3}}} = \frac{4}{40} = .1 \text{ sec.}$$

$$\mathbf{E}\left[T_Q^{\text{disk3}}\right] = \mathbf{E}\left[T_{\text{disk3}}\right] - \mathbf{E}\left[S_{\text{disk3}}\right] = 0.1 \text{ sec} - 0.0225 \text{ sec} = 0.0775 \text{ sec.}$$

Little’s Law practice: Interaction system - 2

- The throughput of disk 3 is 40 requests/sec ($X_{\text{disk3}} = 40$).
- The service time of an average request at disk 3 is 0.0225 sec ($\mathbf{E}\left[S_{\text{disk3}}\right] = .0225$).
- The average number of jobs in the system consisting of disk 3 and its queue is 4 ($\mathbf{E}\left[N_{\text{disk3}}\right] = 4$).



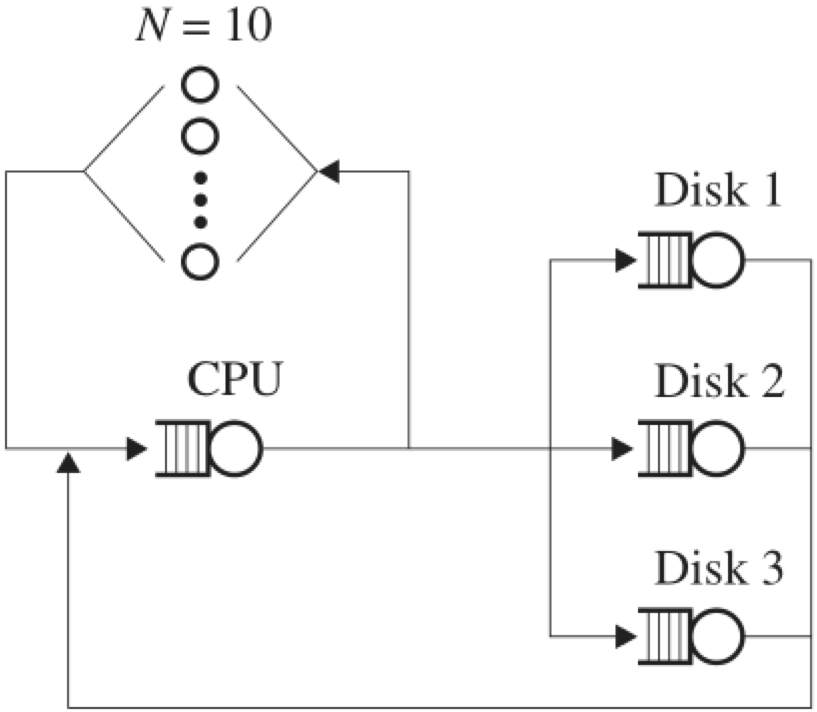
Question: Find $\mathbf{E}$ [Number of requests queued at disk 3].

排队的任务数 = Disk3内的总任务数 – 正在被处理的任务数

$$\begin{aligned}\mathbf{E}\left[N_Q^{\text{disk3}}\right] &= \mathbf{E}\left[N_{\text{disk3}}\right] - \mathbf{E}\left[\text{Number requests serving at disk 3}\right] \\ &= \mathbf{E}\left[N_{\text{disk3}}\right] - \rho_{\text{disk3}} \\ &= 4 - 0.9 \\ &= 3.1 \text{ requests.}\end{aligned}$$

Little's Law practice: Interaction system - 2

- The throughput of disk 3 is 40 requests/sec ($X_{\text{disk3}} = 40$).
- The service time of an average request at disk 3 is 0.0225 sec ($\mathbf{E}[S_{\text{disk3}}] = .0225$).
- The average number of jobs in the system consisting of disk 3 and its queue is 4 ($\mathbf{E}[N_{\text{disk3}}] = 4$).



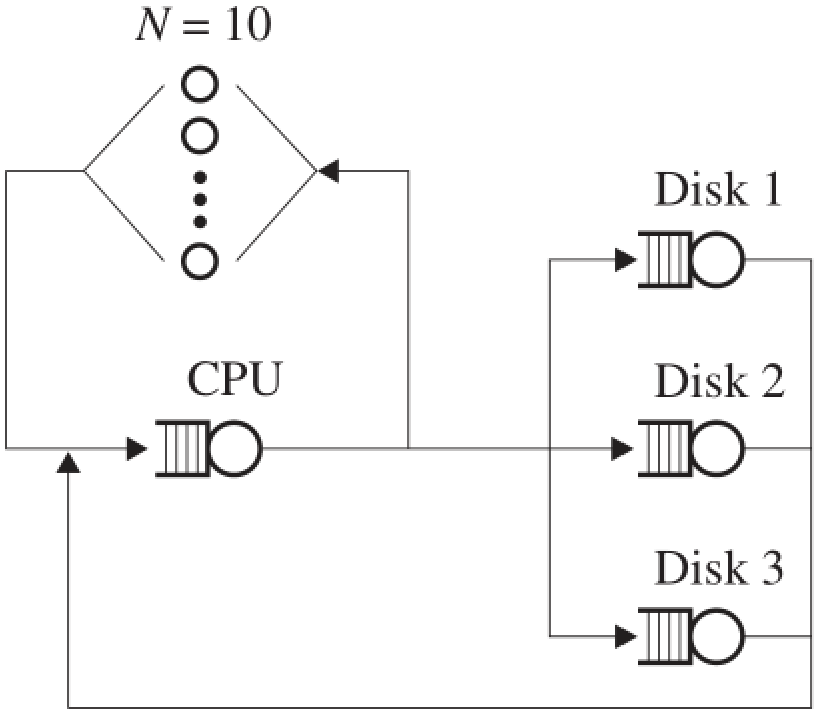
Question: Find $\mathbf{E}$ [Number of requests queued at disk 3].

换一种思路，第二种解法：
排队的任务数 = 对队列本身使用Little's Law

$$\mathbf{E}\left[N_Q^{\text{disk3}}\right] = \mathbf{E}\left[T_Q^{\text{disk3}}\right] \cdot X_{\text{disk3}} = 0.775 \cdot 40 = 3.1 \text{ requests}.$$

Little's Law practice: Interaction system - 2

- The throughput of disk 3 is 40 requests/sec ($X_{\text{disk3}} = 40$).
- The service time of an average request at disk 3 is 0.0225 sec ($\mathbf{E}[S_{\text{disk3}}] = .0225$).
- The average number of jobs in the system consisting of disk 3 and its queue is 4 ($\mathbf{E}[N_{\text{disk3}}] = 4$).

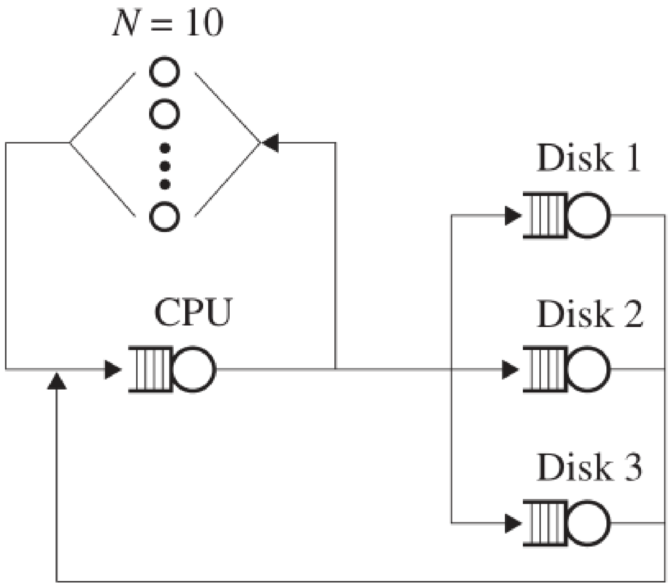


Next we are told that

- $\mathbf{E}$ [Number of ready users (not thinking)] $= 7.5$.
- Number of terminals N is 10.
- $\mathbf{E}$ [Think time] $= \mathbf{E}[Z] = 5$ sec.

Question: What is the system throughput?

Little’s Law practice: Interaction system - 2



Next we are told that

- $\mathbf{E}$ [Number of ready users (not thinking)] $= 7.5$.
- Number of terminals N is 10.
- $\mathbf{E}$ [Think time] $= \mathbf{E}[Z] = 5$ sec.

Question: What is the system throughput?

先对整个系统使用Little’s Law:

$$X = \frac{N}{\mathbf{E}[R] + \mathbf{E}[Z]} = \frac{10}{\mathbf{E}[R] + 5}$$

再对非用户部分使用Little’s Law:

$$\mathbf{E}[R] = \frac{\mathbf{E}[N_{\text{not-thinking}}]}{X} = \frac{7.5}{X}$$

最后可以联立解得: $\mathrm{E}[R] = 15, \mathrm{X} = 0.5$

Operational analysis

Bottleneck analysis

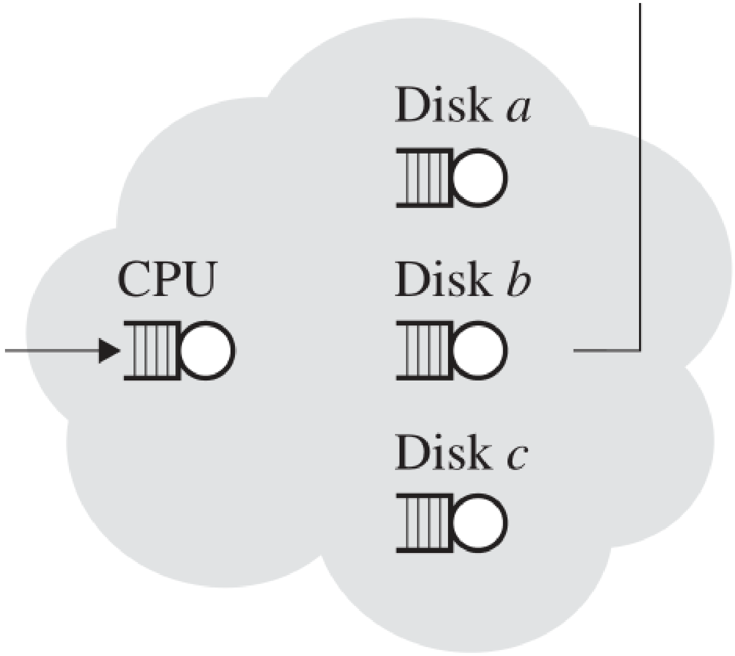
对于任何一个系统，我们可以很容易看到其瓶颈应该在于最慢的设备（木桶原理）

最慢的设备 throughput： $\dfrac{1}{\max D_i}$

同时对整个系统使用Little’s Law，可以得到：

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

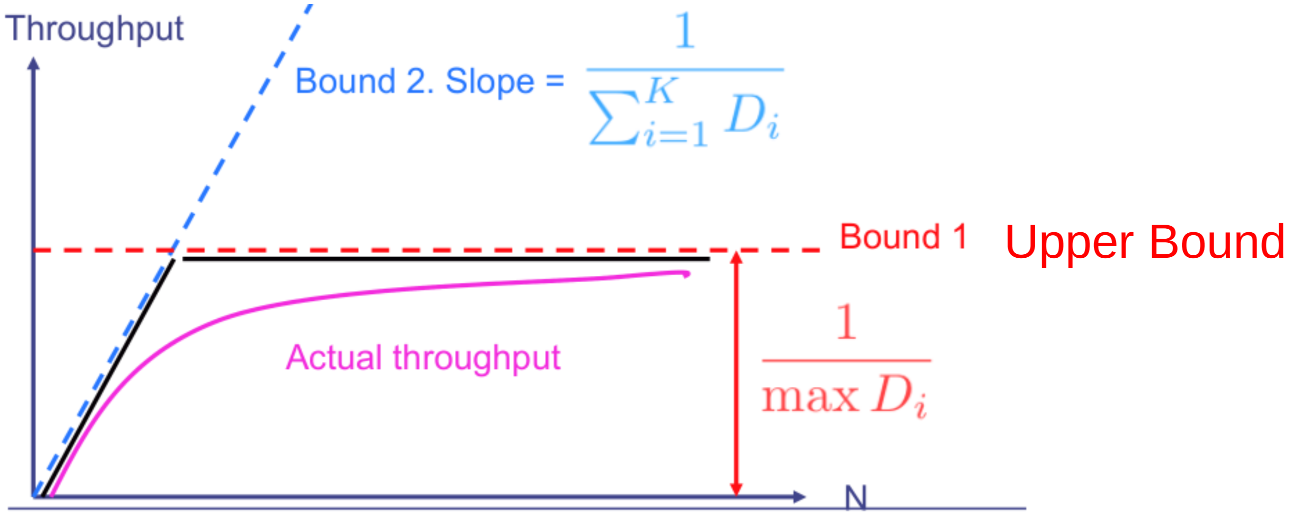
$$\Rightarrow X(0) \leq \frac{N}{\sum_{i=1}^{K} D_i}$$

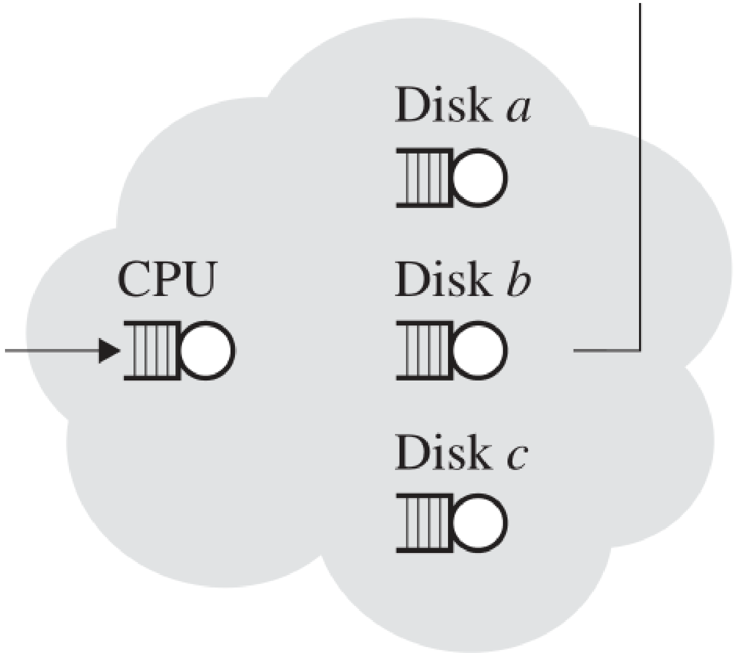


$$X(0) \leq \min\left\{\frac{1}{\max D_i}, \frac{N}{\sum D_i}\right\}$$

Operational analysis

Bottleneck analysis





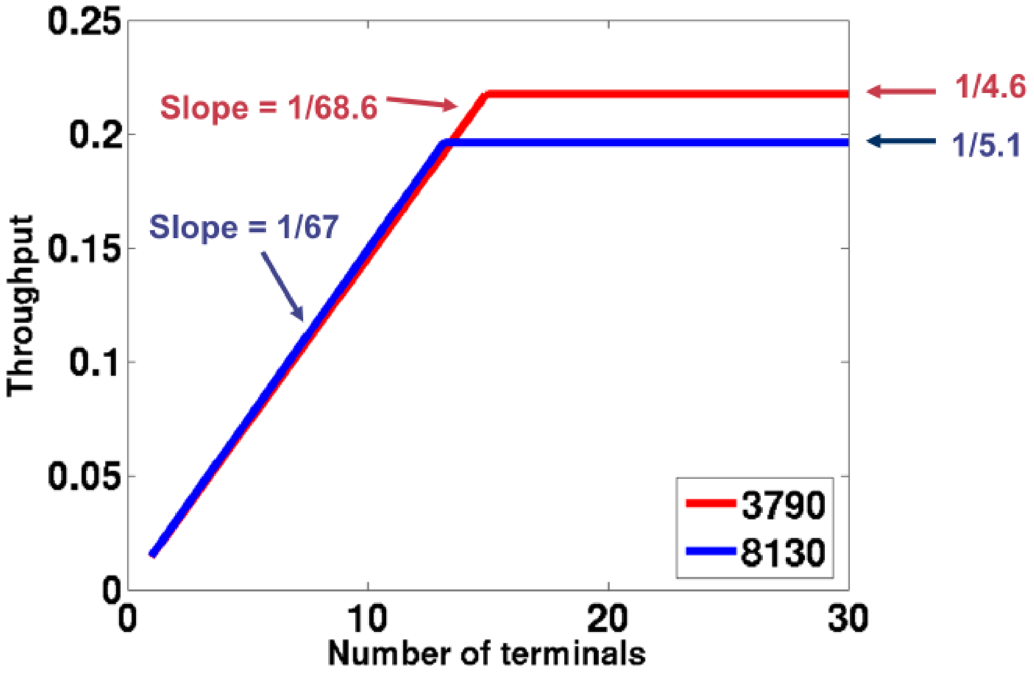
$$X(0) \leq \min\left\{\frac{1}{\max D_i}, \frac{N}{\sum D_i}\right\}$$

Bottleneck analysis

- 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

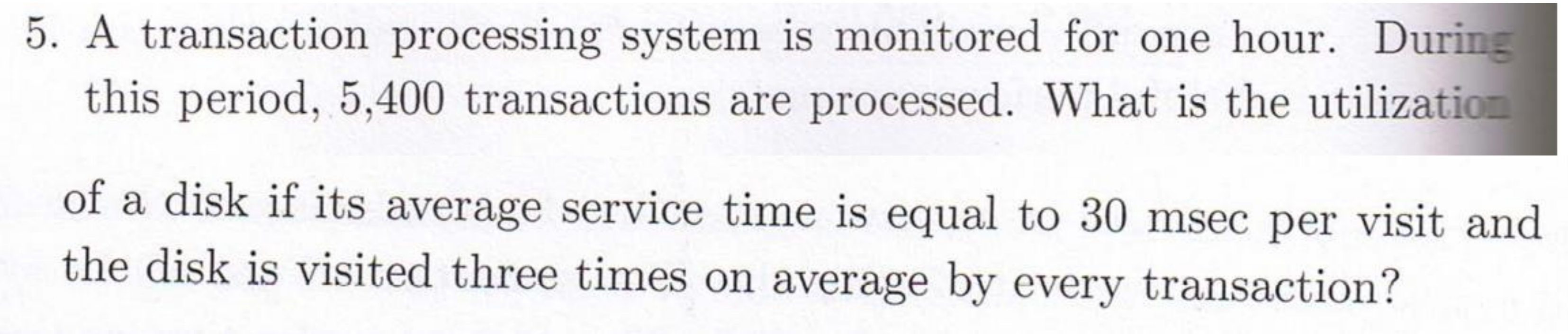
- The company uses the system for interactive application with a think time of 60s.
- Given the same workload, should the company switch to the new system?



升级后的系统Slope斜率几乎一样，但Upper Bound反而更低了，可以认为性能并没有提高

Operational analysis

Exercise



5. A transaction processing system is monitored for one hour. During this period, 5,400 transactions are processed. What is the utilization of a disk if its average service time is equal to 30 msec per visit and the disk is visited three times on average by every transaction?

Step1: 列出所有已知条件
T = 1 hour = 3600s
C = 5400
S(j) = 30ms = 0.03s   V(j) = 3

Step2: 从问题出发寻找关联
要求得U，哪些公式和U相关?
U(j) = S(j)*X(j)   D(j)=U(j)/X(0)

X(0) = C/T = 5400/3600 = 1.5
方法1：
Utilization Law + Forced Flow Law
U(j) = S(j)*X(0)*V(j) = 0.135

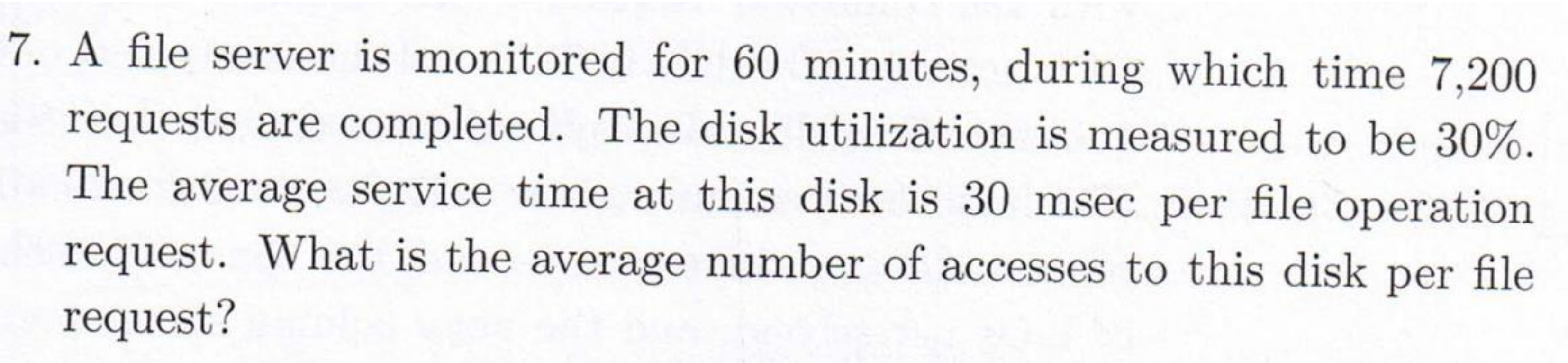
方法2：
Service Demand Law
U(j) = D(j)*X(0) = S(j)*V(j)*X(0) = 0.135

可以发现Utilization Law和Service Demand Law其实是可以相互转化的

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

Operational analysis

Exercise



7. A file server is monitored for 60 minutes, during which time 7,200 requests are completed. The disk utilization is measured to be 30%. The average service time at this disk is 30 msec per file operation request. What is the average number of accesses to this disk per file request?

Step1: 列出所有已知条件
T = 60min = 3600s
C = 7200
U(disk) = 0.3
S(disk) = 30ms = 0.03s

Step2: 从问题出发寻找关联
要求得V(disk)，哪些公式和V相关?
V(disk) = X(disk)/X(0)

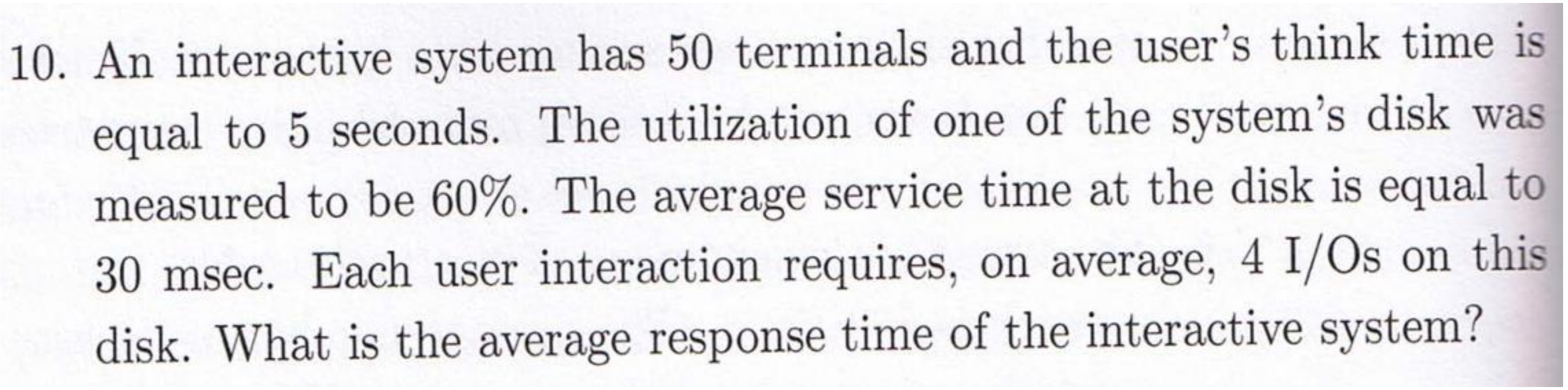
Utilization Law: U(disk) = S(disk)*X(disk)
可得
X(disk) = U(disk)/S(disk) = 0.3/0.03 = 10

X(0) = C/T = 7200/3600 = 2/s

最后
V(disk) = X(disk)/X(0) = 10/2 = 5

Operational analysis

Exercise



10. An interactive system has 50 terminals and the user's think time is equal to 5 seconds. The utilization of one of the system's disk was measured to be 60%. The average service time at the disk is equal to 30 msec. Each user interaction requires, on average, 4 I/Os on this disk. What is the average response time of the interactive system?

Step1: 列出所有已知条件

$N = 50$

$Z = 5s$

$U(disk) = 0.6$

$S(disk) = 30ms = 0.03s$

$V(disk) = 4$

Step2: 从问题出发寻找关联

要求得R，哪些公式和R相关?

$N = R*X$

$X(disk) = U(disk)/S(disk) = 0.6/0.03 = 20$

$X(0) = X(disk)/V(disk) = 20/4 = 5$

注意Interaction system

$R = (Z + T)$

$R = N/X = 50/5 = 10$

最后可得

$T = R – Z = 10 – 5 = 5s$

Queue Model

Arrival process

指数分布
Exponential distribution

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

泊松分布/过程
Poisson distribution/process

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

Queue Model

Arrival process

泊松分布/过程
Poisson distribution/process

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

日常生活中，大量事件是有固定频率的。

- 某医院平均每小时出生3个婴儿
- 某公司平均每10分钟接到1个电话
- 某超市平均每天销售4包xx牌奶粉
- 某网站平均每分钟有2次访问

它们的特点是：我们可以预估这些事件的总数，但是没法知道具体的发生时间。比如已知平均每小时出生3个婴儿，请问下一个小时，会出生几个？

等号的左边，P 表示概率，N表示某种函数关系，t 表示时间，n 表示数量，1小时内出生3个婴儿的概率，就表示为 P(N(1) = 3) 。等号的右边，λ 表示事件的频率。

Queue Model

Arrival process

泊松分布/过程
Poisson distribution/process

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

已知平均每小时出生3个婴儿：$\lambda = 3$

接下来两个小时，一个婴儿都不出生的概率是0.25%，基本不可能发生。

$$P(N(2)=0)=\frac{(3\times 2)^0 e^{-3\times 2}}{0!}\approx 0.0025$$

接下来一个小时，至少出生两个婴儿的概率是80%。

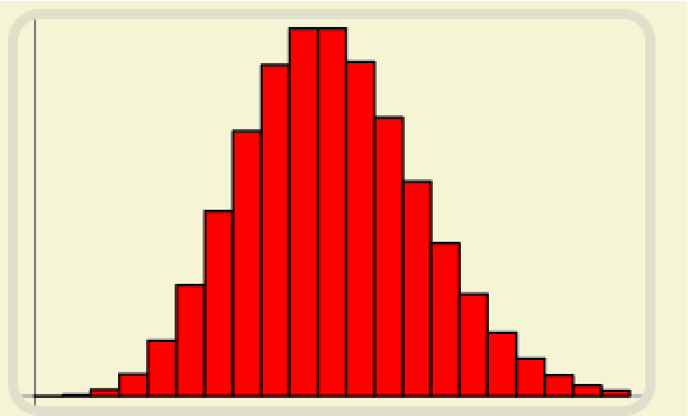
$$\begin{aligned}
P(N(1)\geq 2) &= 1-P(N(1)=1)-P(N(1)=0)\\
&= 1-\frac{(3\times 1)^1 e^{-3\times 1}}{1!}-\frac{(3\times 1)^0 e^{-3\times 1}}{0!}\\
&= 1-3e^{-3}-e^{-3}\\
&= 1-4e^{-3}\\
&\approx 0.8009
\end{aligned}$$

Queue Model

Arrival process

泊松分布/过程
Poisson distribution/process

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



可以看到，在频率附近，事件的发生概率最高，然后向两边对称下降，即变得越大和越小都不太可能。每小时出生3个婴儿，这是最可能的结果，出生得越多或越少，就越不可能。

Queue Model

Arrival process

指数分布
Exponential distribution

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- The cumulative distribution function F(x) = Prob(X ≤ x) is:

$$F(x) = \int_0^x \lambda e^{-\lambda z} dz = 1 - e^{-\lambda x} \text{ for } x \geq 0$$

指数分布是事件的时间间隔的概率。
下面这些都属于指数分布：
•婴儿出生的时间间隔
•来电的时间间隔
•奶粉销售的时间间隔
•网站访问的时间间隔

指数分布的公式可以从泊松分布推断出来。如果下一个婴儿要间隔时间 t，就等同于 t 之内没有任何婴儿出生

$$P(X > t) = P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!}$$

$$= e^{-\lambda t}$$

反过来，事件在时间 t 之内发生的概率，就是1减去上面的值。

$$P(X \leq t) = 1 - P(X > t) = 1 - e^{-\lambda t}$$

Queue Model

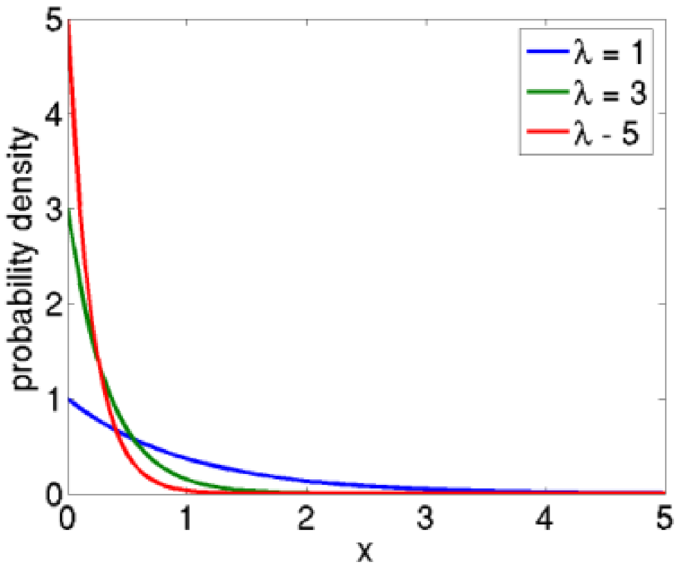
Arrival process

指数分布
Exponential distribution

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- The cumulative distribution function F(x) = Prob(X ≤ x) is:

$$F(x) = \int_0^x \lambda e^{-\lambda z} dz = 1 - e^{-\lambda x} \text{ for } x \geq 0$$



接下来15分钟，会有婴儿出生的概率是52.76%。

$$P(X \leq 0.25) = 1 - e^{-3\times 0.25}$$

$$\approx 0.5276$$

接下来的15分钟到30分钟，会有婴儿出生的概率是24.92%。

$$P(0.25 \leq X \leq 0.5) = P(X \leq 0.5) - P(X \leq 0.25)$$

$$= (1 - e^{-3\times 0.5}) - (1 - e^{-3\times 0.25})$$

$$= e^{-0.75} - e^{-1.5}$$

$$\approx 0.2492$$

Queue Model

Arrival process

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

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

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

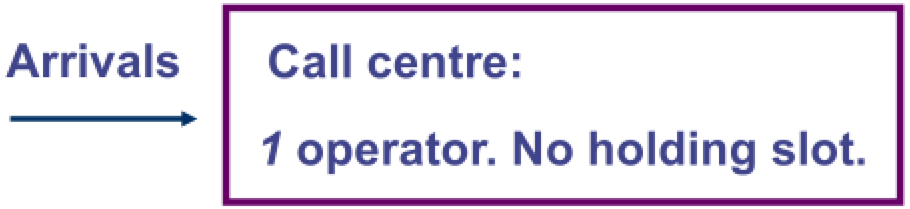
- You can also show that if the inter-arrival time distribution is exponential with parameter λ, then the mean inter-arrival time is $1/\lambda$
- Quite nicely, we have

  Mean arrival rate = 1 / mean inter-arrival time

由此可以知道，arrival的平均间隔时间就是 $\frac{1}{\lambda}$

Queue Model

Call center with 1 operator and no holding slots



Arrivals

Call centre:

1 operator. No holding slot.

- 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)

- What happens to a call that arrives when the operator is busy?
  - The call is rejected
- What happens to a call that arrives when the operator is idle?
  - The call is admitted without delay.
- We are interested to find the probability that an arriving call is rejected.

Queue Model

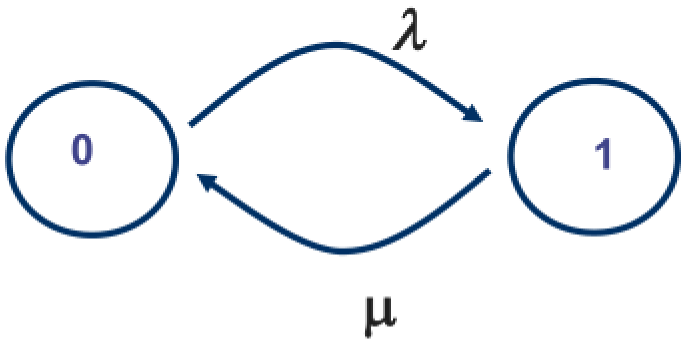
Call center with 1 operator and no holding slots

Arrivals

Call centre:

1 operator. No holding slot.

使用马尔科夫链(Markov chain)来描述模型



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

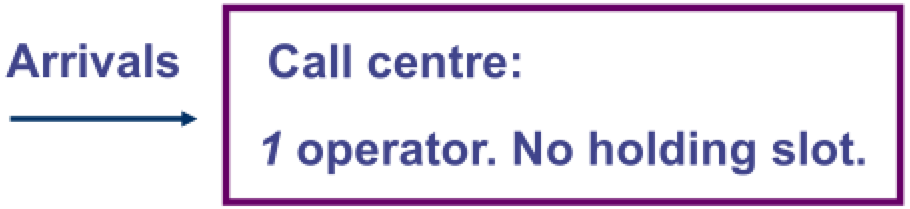
$$\underline{\lambda P_0} = \underline{\mu P_1}$$

- 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:

$$\underline{\lambda P_0} = \underline{\mu P_1}$$

Queue Model

Call center with 1 operator and no holding slots

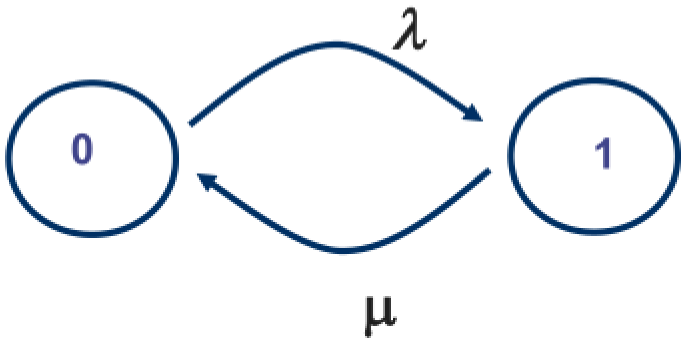


Arrivals

Call centre:

1 operator. No holding slot.

使用马尔科夫链(Markov chain)来描述模型



- 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:

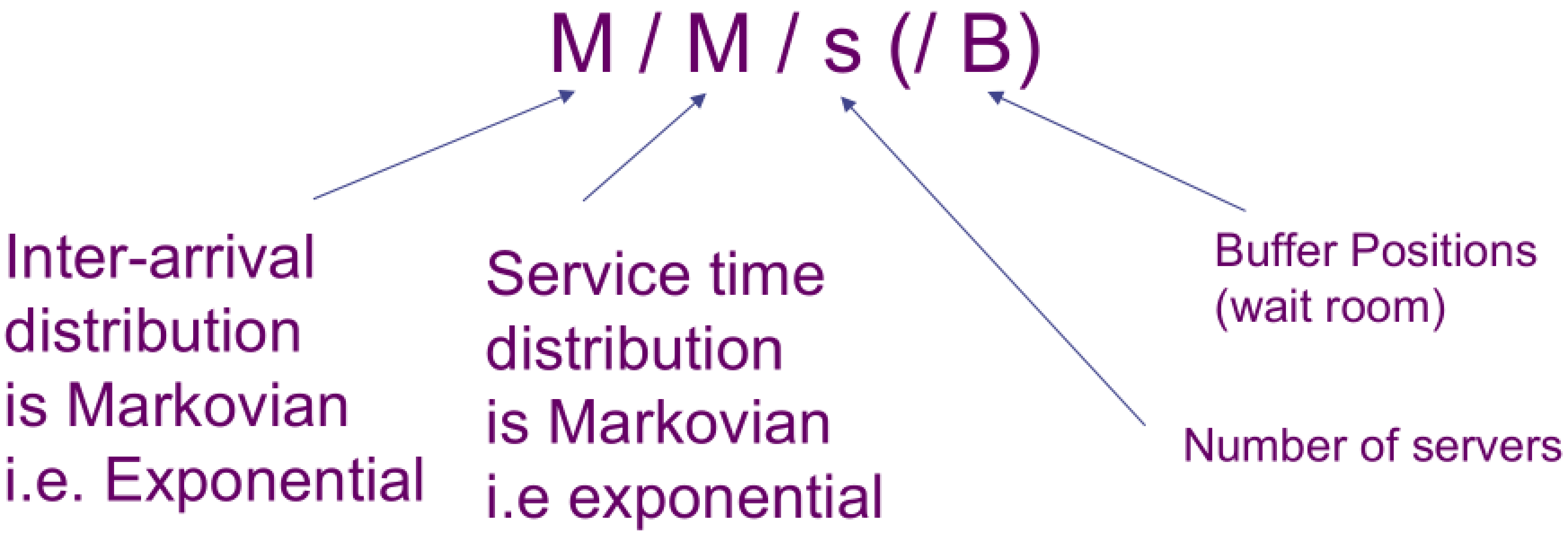
$$P_0 + P_1 = 1$$

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

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

Queue Model

Kendall's notation



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

如果队列长度无限则可以省略最后一项

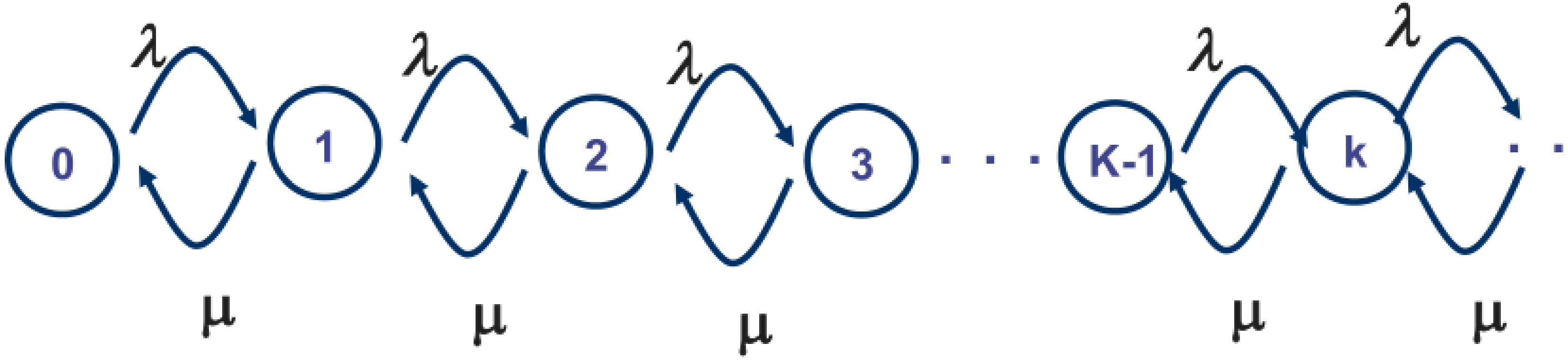
Queue Model

Call center with 1 operator and infinite holding slots – M/M/1

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.

和上一个例子一样，我们先把问题转化为马尔科夫链(Markov chain)



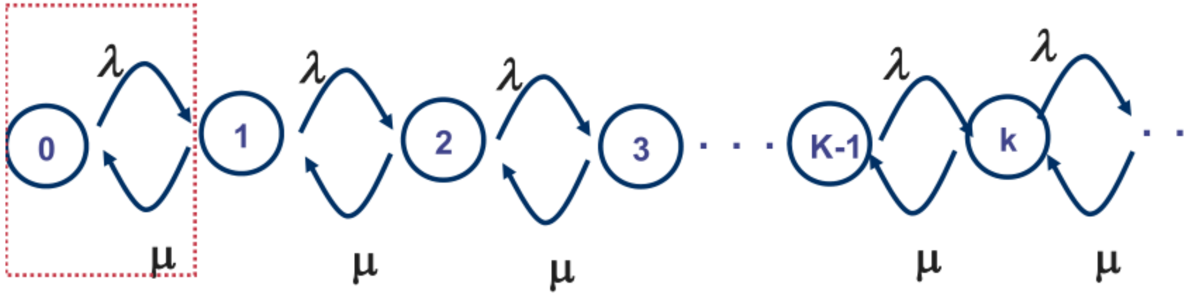
Queue Model

Call center with 1 operator and infinite holding slots – M/M/1

Arrivals →

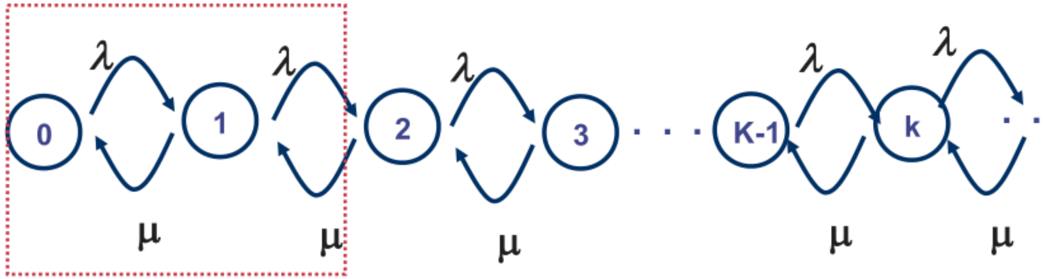
Call centre with *1* operator
If the operator is busy, the centre will put the call on hold.
A customer will wait until his call is answered.

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



$$\lambda P_0 = \mu P_1$$

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



$$\lambda P_1 = \mu P_2$$

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

Queue Model

Call center with 1 operator and infinite holding slots – M/M/1

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.

不断递推下去我们可以观察得到每个状态概率P的一般形式：

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

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

We have $P_k = \rho^k P_0$

With $P_k = \rho^k P_0$ and

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

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

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

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

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

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

Arrival rate < service rate

Queue Model

Call center with 1 operator and infinite holding slots – M/M/1

用期望求出系统内的平均客户数量

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 =

$$\sum_{k=0}^{\infty} kP_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k$$

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

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

Throughput is λ ***(why?)***

系统稳定时输出速率就等于输入速率

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

Queue Model

Call center with 1 operator and infinite holding slots – M/M/1

用期望求出系统内的平均客户数量

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 =

$$\sum_{k=0}^{\infty} kP_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k$$

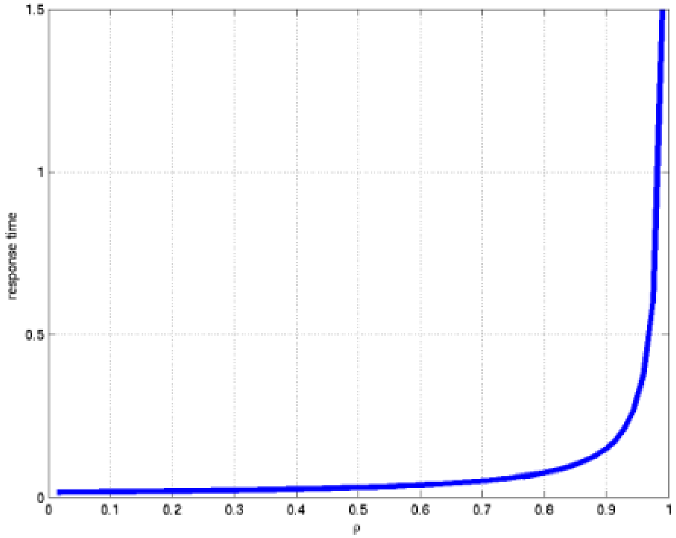
$$= \frac{\rho}{1-\rho}$$

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

Throughput is λ ***(why?)***

系统稳定时输出速率就等于输入速率

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

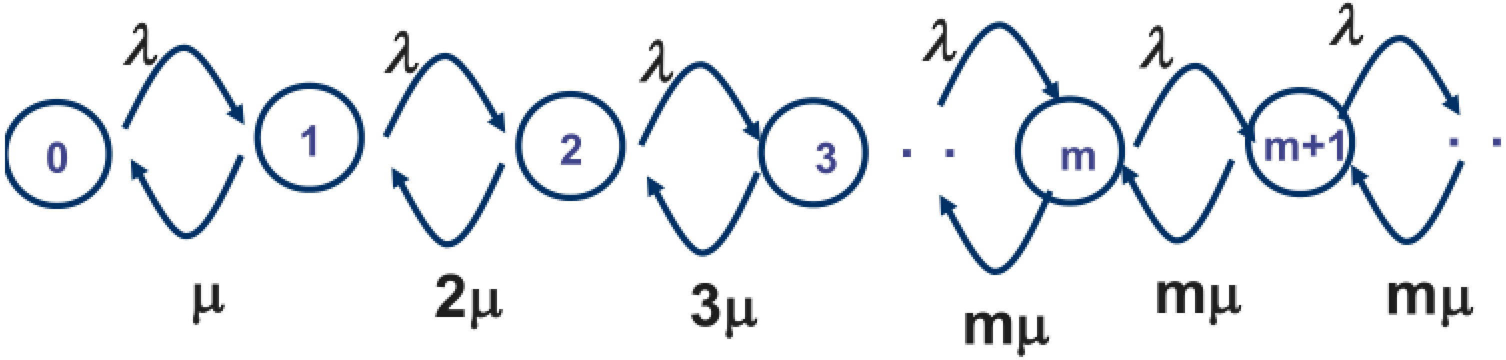


当ρ趋向1时，T趋向无穷大

可以理解为因为系统处理速度开始慢于任务到达的速度，任务于是无限堆积

Queue Model

Call center with m operator and infinite holding slots – M/M/m



和M/M/1系统相比，可以发现唯一的区别在于每个状态的处理速率u不同

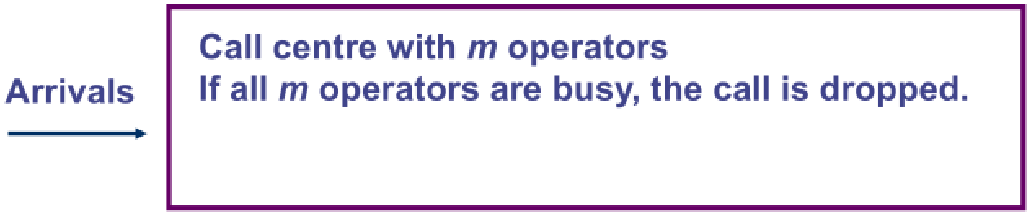
$$T = \frac{C(\rho, m)}{m\mu(1-\rho)} + \frac{1}{\mu}$$

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

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

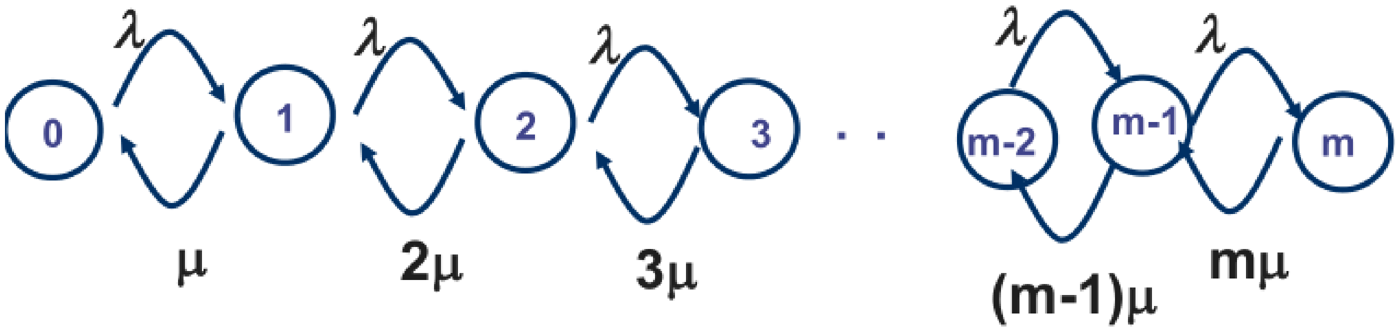
Queue Model

Call center with m operator and finite holding slots – M/M/m/m



Arrivals

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



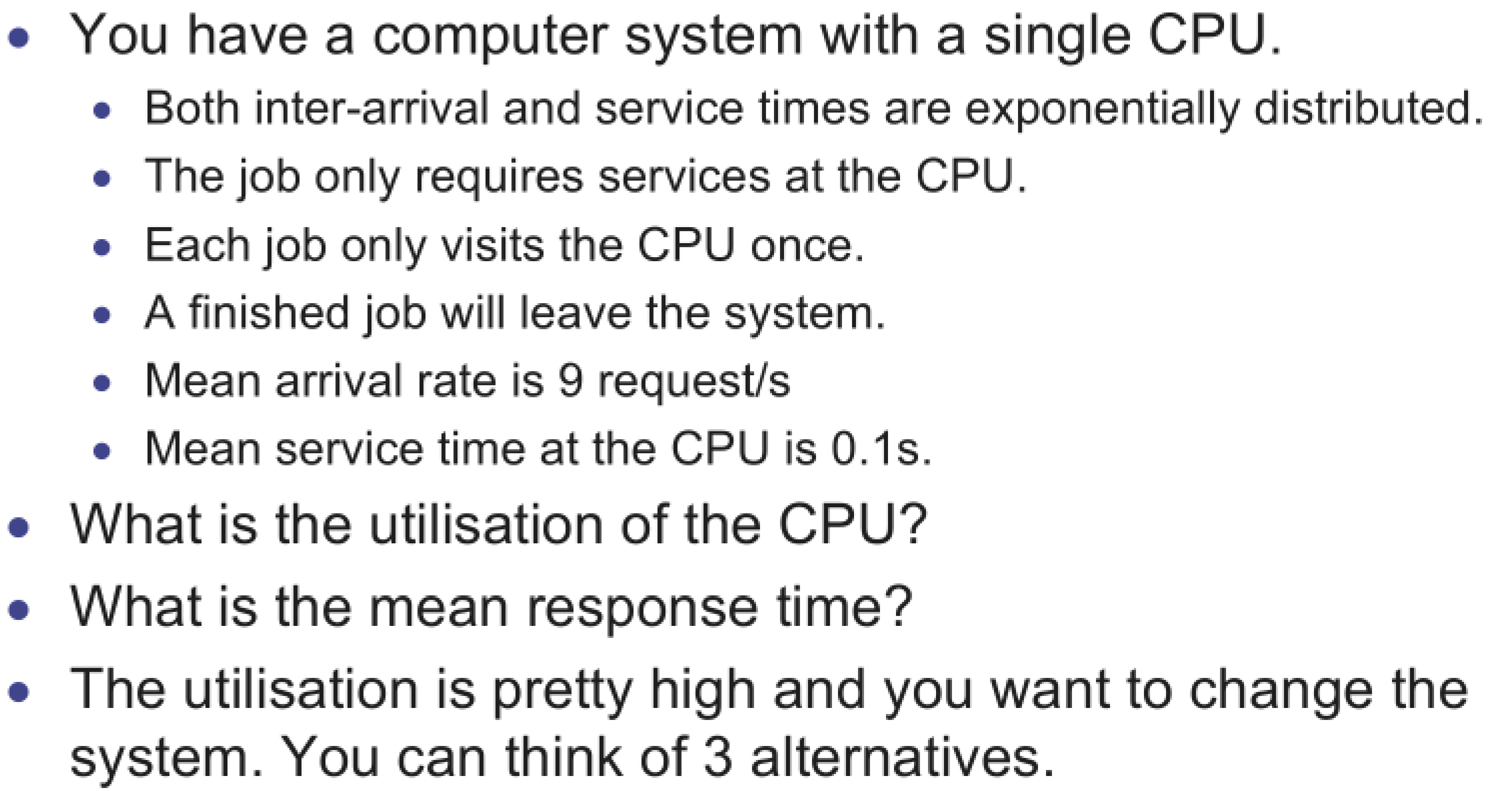
Probability that an arrival is blocked
= Probability that there are m customers in the system

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

"Erlang B formula"

Queue Model

Exercise



- You have a computer system with a single CPU.
  - Both inter-arrival and service times are exponentially distributed.
  - The job only requires services at the CPU.
  - Each job only visits the CPU once.
  - A finished job will leave the system.
  - Mean arrival rate is 9 request/s
  - Mean service time at the CPU is 0.1s.
- What is the utilisation of the CPU?
- What is the mean response time?
- The utilisation is pretty high and you want to change the system. You can think of 3 alternatives.

Queue Model

Exercise

从题目条件可以知道：

$$\lambda = 9, \qquad \mu = \frac{1}{service\ time} = \frac{1}{0.1} = 10$$

可得 Utilization (这里我们统一用ρ表示)

$$\rho = \frac{\lambda}{\mu} = \frac{9}{10} = 0.9$$

达到和服务都是指数分布，且只有一个处理器（可认为等待队列无限）可以用M/M/1描述

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

代入可得 T = 1/1 = 1s

Queue Model

Exercise

考虑三种升级策略

方案1： 升级CPU, 处理速率翻倍

$$\mu' = 2\mu$$

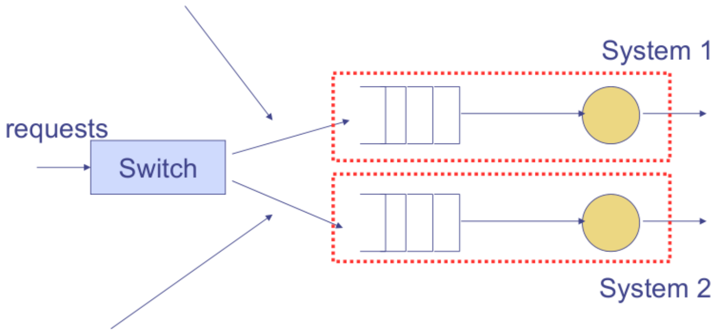
$$\mathrm{T}_1 = \frac{1}{\mu' - \lambda} = \frac{1}{20 - 9} = 0.0909$$

方案2： 用两个一样的处理器，彼此队列独立

可以发现由于两个系统分担任务，相当于各自的到达速率减半

$$\mathrm{T}_2 = \frac{1}{\mu - \lambda'} = \frac{1}{10 - 4.5} = 0.1818$$

1st,3rd,5th,… requests to system 1



2nd,4th,6th … requests to System 2

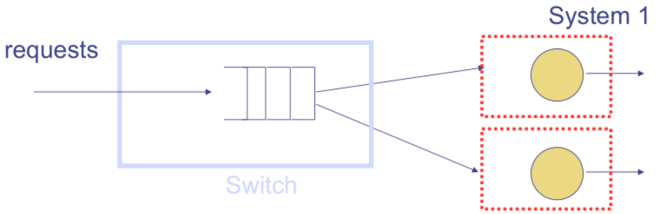
方案3： 用两个一样的处理器，队列共享

即变成M/M/2

$$\rho = \frac{\lambda}{m\mu} = \frac{0.9}{2} = 0.45\ ,\ \mathrm{T}_3 = \frac{C(\rho,m)}{m\mu(1-\rho)} + \frac{1}{\mu} = \frac{0.2793}{2*10*0.55} + \frac{1}{10} = 0.1254$$

$$C(\rho,m) = \frac{\frac{(m\rho)^m}{m!}}{(1-\rho)\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!}} = \frac{0.405}{0.55*(1+0.9)+0.405}$$

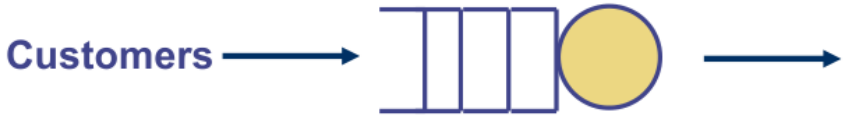
$$= 0.2793$$



The switch dispatches the requests to the first available server.

Queue Model

Exercise



- Consider a single server queue as shown above
- Part (a): Consider the situation
    - The inter-arrival time is a constant and is given by 1 second.
    - The service time required by each customer is always 0.5 second.
  - What is the mean waiting time per customer?
- Part (b): Consider the situation
    - The inter-arrival time is exponentially distributed with mean 1 second
    - The service time required by each customer is exponentially distributed with mean 0.5s
  - What is the mean waiting time per customer?
- Compare the answers of Parts (a) and (b). What conclusions can you draw?

(a) 这是一个M/M/1模型吗?

从题设可以发现，到达时间和服务时间都是常数，而不是指数分布，所以不是M/M/1模型。

这里并没有任何任务需要等待，因为每个任务的处理时间0.5s小于到达间隔1s

(b) 这是一个M/M/1模型吗?

是的，可以直接使用公式求解

$$T = \frac{1}{u - \lambda} = \frac{1}{1} = 1s$$

Queue Model

Exercise

- An Internet Service Provider has 4 dial-up ports. Connection requests obey Poisson distribution with a mean arrival rate of 3 requests per hour. The session duration of each connection request is exponentially distributed with a mean of 1.5 hours. What is the probability that a connection request will be rejected?

这是一个M/M/4/4模型

通话被拒绝的概率就等于系统里面有4个通话的概率（即系统所有dial-up port被占满的概率）

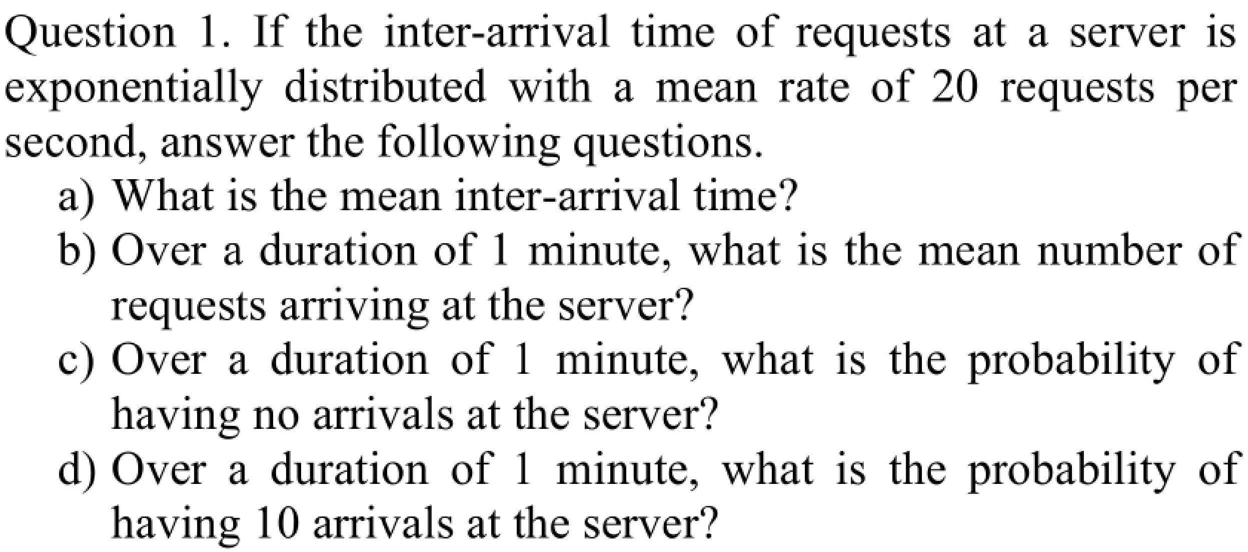
$$P_m = \frac{\frac{\rho^m}{m!}}{\sum_{k=0}^{m} \frac{\rho^k}{k!}}, \rho = \frac{\lambda}{\mu} = \frac{3}{1/1.5} = 4.5$$

$$\Rightarrow P_4 = \frac{\frac{4.5^4}{24}}{1 + 4.5 + \frac{4.5^2}{2} + \frac{4.5^3}{6} + \frac{4.5^4}{24}}$$

$$= 0.3567$$

Queue Model

Exercise: Arrival Process



Question 1. If the inter-arrival time of requests at a server is exponentially distributed with a mean rate of 20 requests per second, answer the following questions.

a) What is the mean inter-arrival time?

b) Over a duration of 1 minute, what is the mean number of requests arriving at the server?

c) Over a duration of 1 minute, what is the probability of having no arrivals at the server?

d) Over a duration of 1 minute, what is the probability of having 10 arrivals at the server?

a) Mean inter-arrival time = 1/20 = 0.05s

b) Mean number of requests = 20*60 = 1200

c) 没有任务到达，即n=0

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

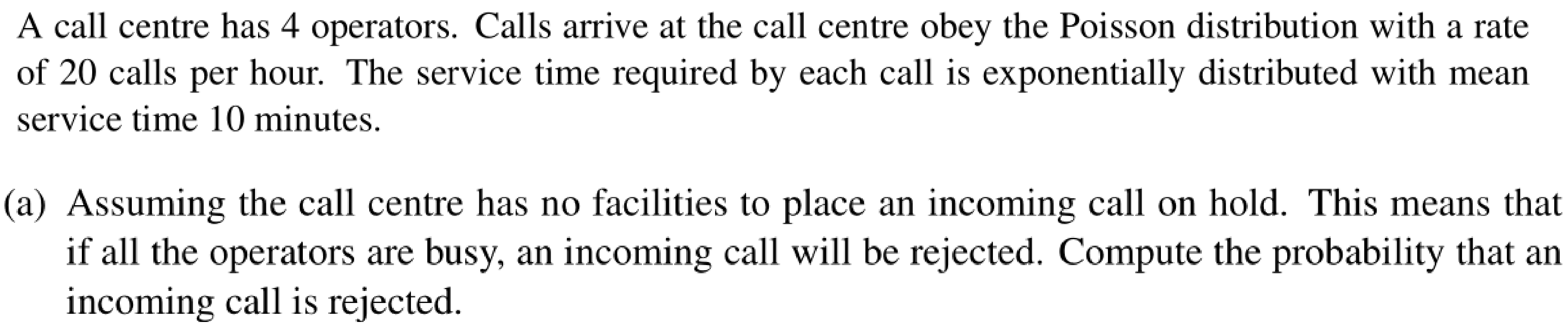
$$\exp(-\lambda t) = \exp(-1200)$$

d) n=10

$$\frac{(1200)^{10} \exp(-1200)}{10!}$$

Queue Model

Exercise: Queue Model - Inference



A call centre has 4 operators. Calls arrive at the call centre obey the Poisson distribution with a rate of 20 calls per hour. The service time required by each call is exponentially distributed with mean service time 10 minutes.

(a) Assuming the call centre has no facilities to place an incoming call on hold. This means that if all the operators are busy, an incoming call will be rejected. Compute the probability that an incoming call is rejected.

先判断题目属于什么模型

a) This is a M/M/4/4 model

Probability that an incoming call is rejected

= Probability that there are 4 calls in the system

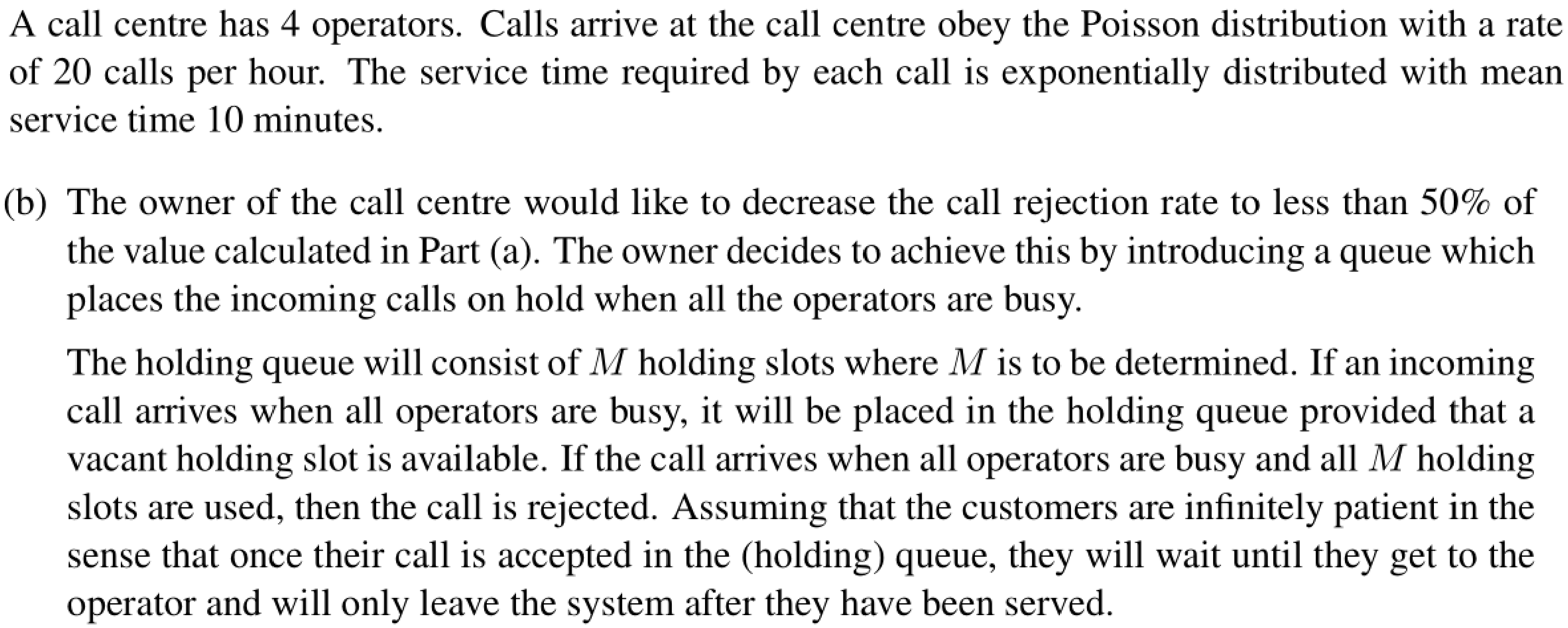
= P(4)

$$\mathrm{P}(4)=\frac{\frac{\rho^4}{4!}}{\sum_{k=0}^{4}\frac{\rho^k}{k!}}\quad where\ \rho=\frac{\lambda}{\mu}=\frac{20/h}{1/10min}=\frac{10}{3}=3.3333$$

$$\mathrm{P(rej)}=\mathrm{P}(4)=0.2425$$

Queue Model

Exercise: Queue Model - Inference



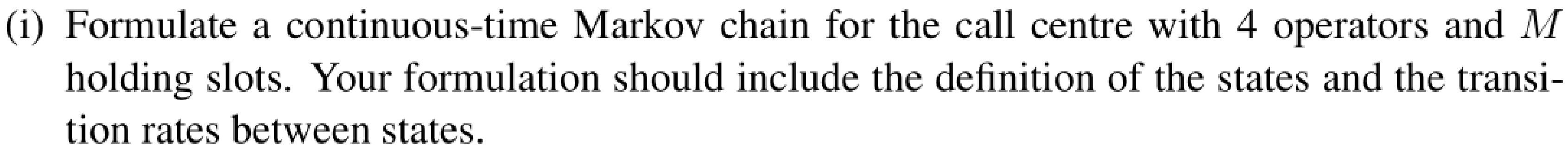
A call centre has 4 operators. Calls arrive at the call centre obey the Poisson distribution with a rate of 20 calls per hour. The service time required by each call is exponentially distributed with mean service time 10 minutes.

(b) The owner of the call centre would like to decrease the call rejection rate to less than 50% of the value calculated in Part (a). The owner decides to achieve this by introducing a queue which places the incoming calls on hold when all the operators are busy.

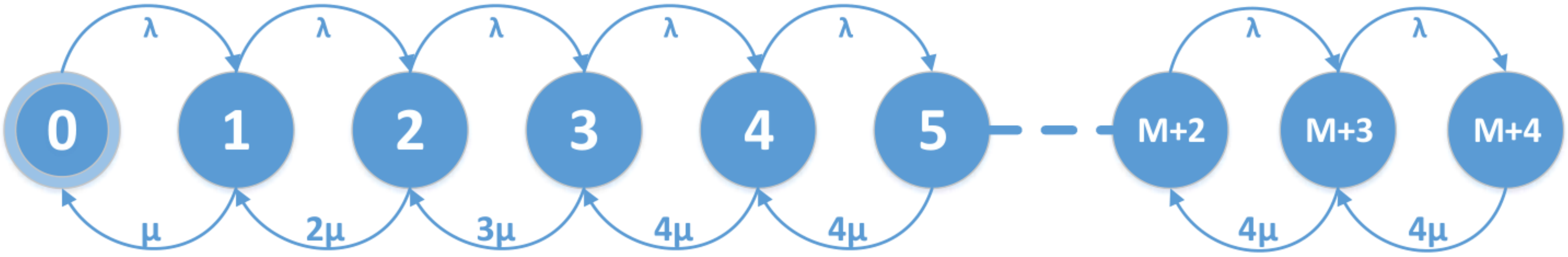
The holding queue will consist of M holding slots where M is to be determined. If an incoming call arrives when all operators are busy, it will be placed in the holding queue provided that a vacant holding slot is available. If the call arrives when all operators are busy and all M holding slots are used, then the call is rejected. Assuming that the customers are infinitely patient in the sense that once their call is accepted in the (holding) queue, they will wait until they get to the operator and will only leave the system after they have been served.

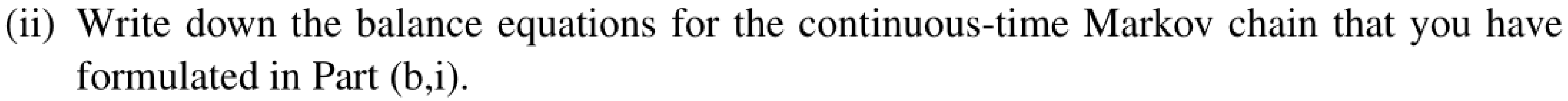
Queue Model

Exercise: Queue Model - Inference



(i) Formulate a continuous-time Markov chain for the call centre with 4 operators and M holding slots. Your formulation should include the definition of the states and the transition rates between states.





(ii) Write down the balance equations for the continuous-time Markov chain that you have formulated in Part (b,i).

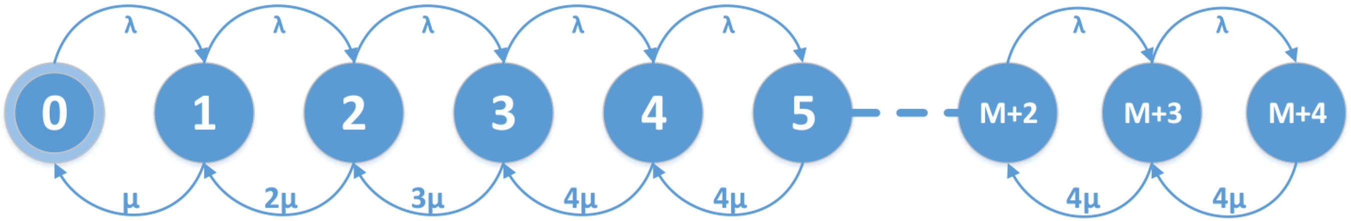
$\lambda P(0) = \mu P(1) \quad \lambda P(1) = 2\mu P(2) \quad \lambda P(2) = 3\mu P(3) \quad \lambda P(3) = 4\mu P(4)$

$\lambda P(4) = 4\mu P(5)$ …… $\lambda P(M+3) = 4\mu P(M+4)$

Queue Model

Exercise: Queue Model - Inference

(iii) Derive expressions for the steady state probabilities of the continuous-time Markov chain that you have formulated.



可以发现，前4项关系式为$P(k) = P(0) * \left(\frac{1}{k!}\right)\rho^k$ 从第5项之后，关系式为$P(k) = P(0) * \left(\frac{1}{24*4^{k-4}}\right)\rho^k$

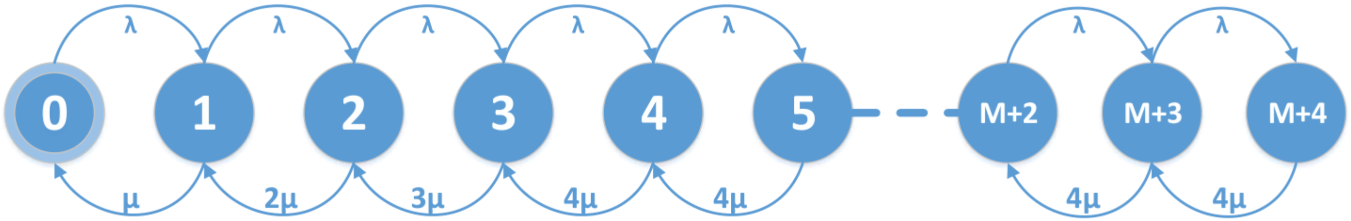
The expression for the steady state can be shown as:

$$\mathrm{P(k)} = \begin{cases} P(0)\dfrac{1}{k!}\rho^k & k \le 4 \\ P(0)\dfrac{1}{24 * 4^{k-4}}\rho^k & 4 < k \le \mathrm{M} + 4 \end{cases}$$

Queue Model

Exercise: Queue Model - Inference

(iv) Use your answer in Part (b,iii) to determine the smallest value of M required to reduce the call rejection rate to less than 50% of the value calculated in Part (a).



let ρ = λ/μ then we have

$$P(0)\left(1+\rho+\frac{1}{2}\rho^2+\frac{1}{6}\rho^3+\frac{1}{24}\rho^4+\cdots+\frac{1}{24*(4)^{M-1}}\rho^{M+3}+\frac{1}{24*(4)^M}\rho^{M+4}\right)=1$$

the probability of all states is 1

$$P(0)+P(1)+P(2)+\cdots+P(M+3)+P(M+4)=1$$

对不同的M求解P(0)即可

which can be written as:

$$P(0)\left[\sum_{k=0}^{4}\frac{1}{k!}\rho^k+\sum_{k=5}^{M+4}\frac{1}{24*4^{k-4}}\rho^k\right]=1$$

The smallest M we find here is 3, which means we need 3 waiting slots.

$$P_{M=3}(\text{rej})=P_{M=3}(7)=\ 0.0929\ <\ 0.5*P_{M=0}(\text{rej})=\ 0.1213$$

Queue Model

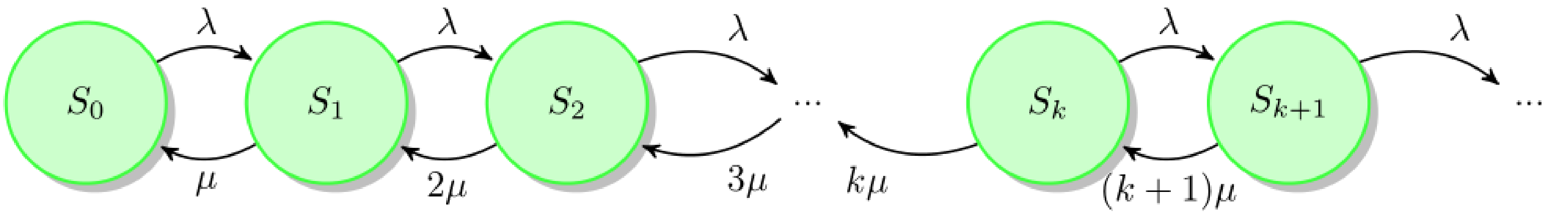
Exercise: Queue Model – Practice

A queuing system has one server and infinite queuing capacity. The number of customers in the system can be modeled as a birth-death process with $\lambda_k = \lambda$ *and* $\mu_k = k\mu$, $k = 0, 1, 2, \ldots$ *thus, the server increases the speed of the service with the number of customers in the queue. Calculate the average number of customers in the system as a function of* $\rho = \lambda/\mu$.

Hint: $$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Queue Model

Exercise: Queue Model – Practice



$$\lambda P_0 = \mu P_1 \to P_1 = \rho P_0$$

$$\lambda P_1 = 2\mu P_2 \to P_2 = \tfrac{1}{2}\rho P_1 = \tfrac{1}{2}\rho^2 P_0$$

$$\lambda P_2 = 3\mu P_3 \to P_3 = \tfrac{1}{3}\rho P_2 = \tfrac{1}{2\cdot 3}\rho^3 P_0$$

....

$$\lambda P_{k-1} = k\mu P_k \to P_k = \tfrac{1}{k}\rho P_{k-1} = \ldots = \tfrac{1}{k!}\rho^k P_0$$

....

$$\textstyle\sum_{k=0}^{\infty} P_k = 1 \quad \text{(normalization)}$$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

$$\sum_{k=0}^{\infty} \frac{\rho^k}{k!} P_0 = 1 \to P_0 e^{\rho} = 1 \to P_0 = e^{-\rho}.$$

$$P_k = \frac{\rho^k}{k!} e^{-\rho}$$

注意这就是泊松分布的表达式！

回忆一下泊松分布的均值求解过程

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

所以系统内的平均用户数量为

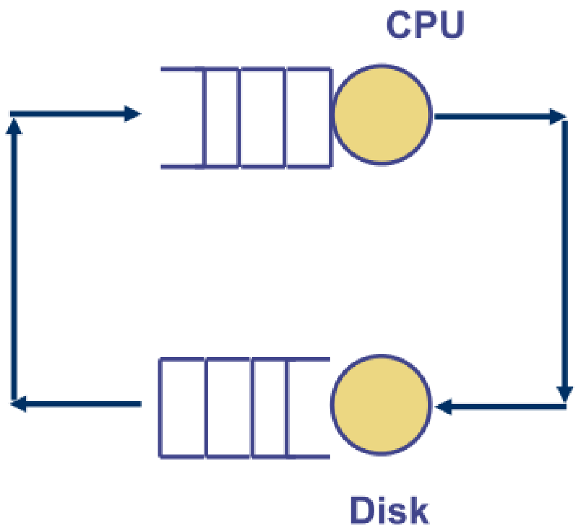
$$\overline{N} = \sum_{k=0}^{\infty} k P_k = \rho$$

如果结合Little’s Law，可以神奇地发现客户的response time就等于service time

$$E[T_{total}] = \frac{\overline{N}}{\lambda} = \frac{1}{\mu}.$$

Markov Chain

- You can use Markov Chain to analyse
  - Closed queueing network (see example below)
  - Reliability problem

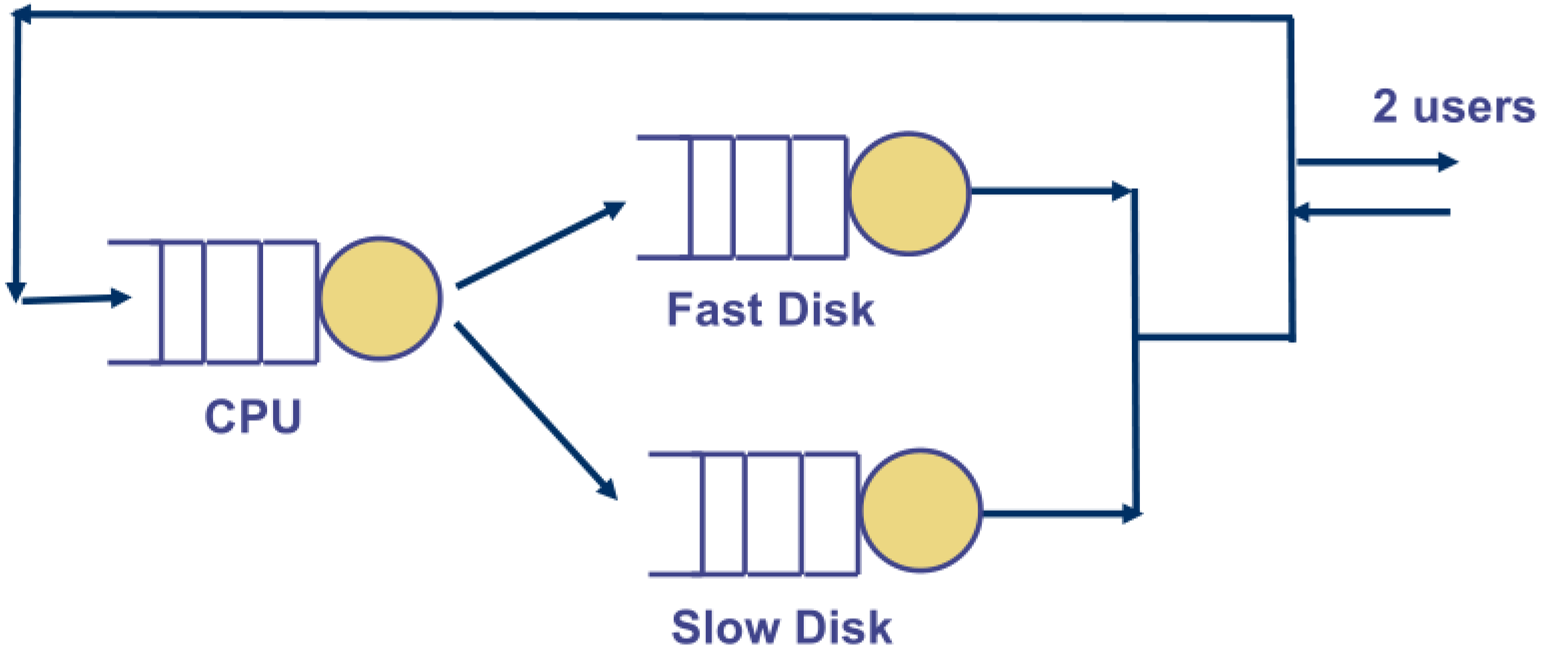


- There are *n* jobs in the closed system
- What is the response time of one job?
- What is the response time if we replace the CPU with one that is twice as fast?

- A Markov chain can be solved by
  - Identifying the states (may not be easy)
  - Find the transition rate between the states
  - Solve the steady state probabilities
- You can then use the steady state probabilities as a stepping stone to find the quantity of interest (e.g. response time etc.)

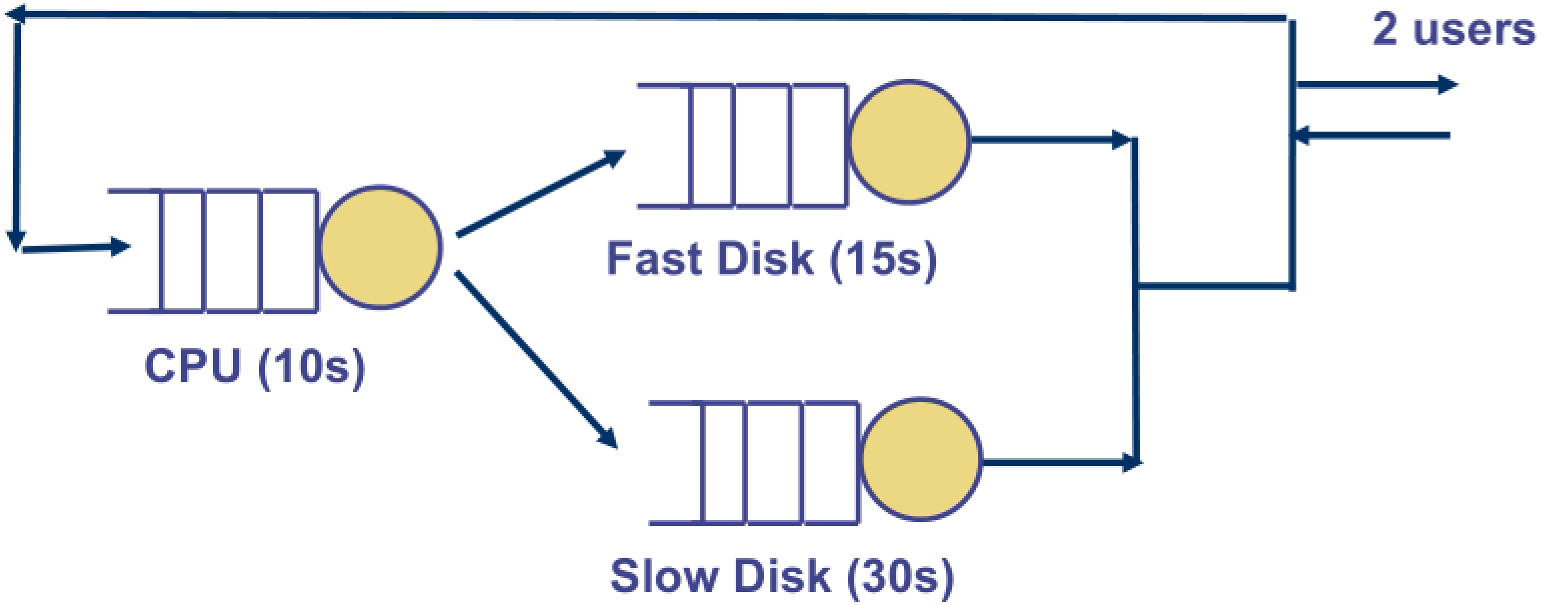
Markov Chain

- A database server with a CPU, a fast disk and a slow disk
- At peak demand, there are always two users in the system
- Transactions alternate between the CPU and the disks
- The transactions will equally likely find the file on either disk

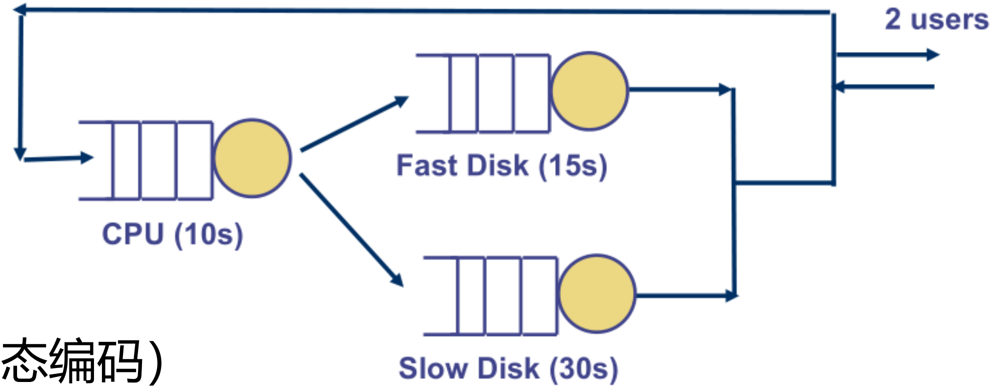


Markov Chain

- Fast disk is twice as fast as the slow disk
- Typical transactions take on average 10s CPU time
- Fast disk takes on average 15s to serve all files for a transactions
- Slow disk takes on average 30s to serve all files for a transactions
- The time that each transaction requires from the CPU and the disks is exponentially distributed



Markov Chain



第一步：
选择描述状态的方法（状态编码）

比如我们可以用2元变量(A,B)来表示
变量的值为该任务(或用户)所在的位置

Use a 2-tuple (A,B) where

- A is the location of the first user
- B is the location of the second user
- A, B are drawn from {CPU,FD,SD}
  - FD = fast disk, SD = slow disk
- Example states are:
  - (CPU,CPU): both users at CPU
  - (CPU, FD): 1st user at CPU, 2nd user at fast disk
- Total 9 states

也可以用3元变量(A,B,C)来表示
变量的值为该设备中的任务(或用户)数量

We use a 3-tuple (X,Y,Z)

- X is # users at CPU
- Y is # users at fast disk
- Z is # users at slow disk

第二种编码方式较为常见

Examples

- (2,0,0): both users at CPU
- (1,0,1): one user at CPU and one user at slow disk

Six possible states

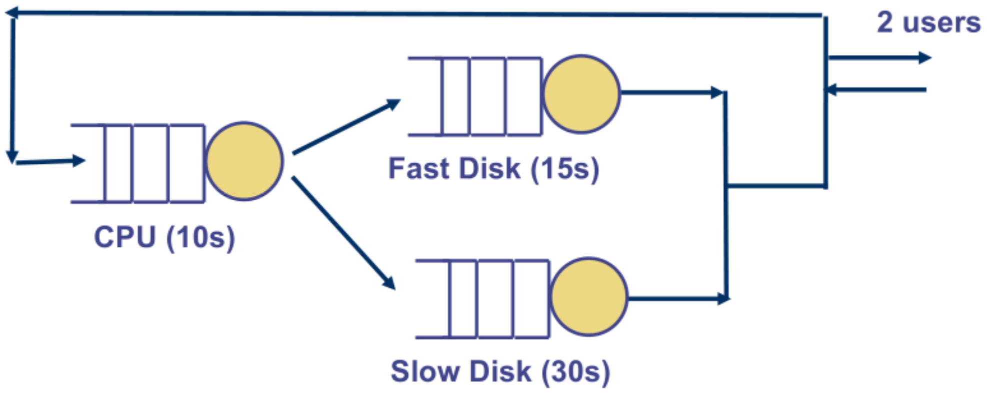
- (2,0,0) (1,1,0) (1,0,1) (0,2,0) (0,1,1) (0,0,2)

If there are n users, how many states do you need?

$$\frac{(n+1)(n+2)}{2}$$

Choice #2 requires less #states but loses certain information.

Markov Chain



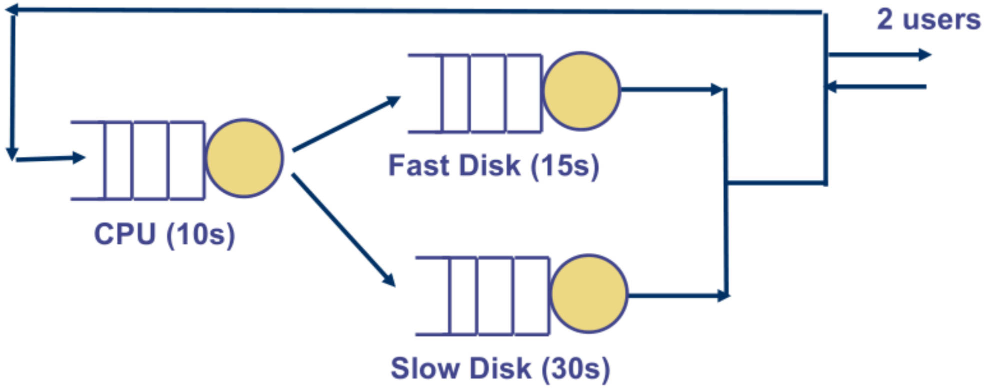
第二步：
描述状态转移

- A state is: (#users at CPU, #users at fast disk, #users at slow disk)
- What is the rate of moving from State (2,0,0) to State (1,1,0)?
  - This is caused by a job finishing at the CPU and move to fast disk
  - Jobs complete at CPU at a rate of 6 transactions/minute
  - Half of the jobs go to the fast disk
- Transition rate from (2,0,0) ➔ (1,1,0) = 3 transactions/minute
- Similarly, transition rate from (2,0,0) ➔ (1,0,1) = 3 transactions/minute

我们知道CPU的处理单个任务需要10s = 每分钟6个任务 6/min
同理
Fast_Disk: 4/min
Slow_Disk: 2/min

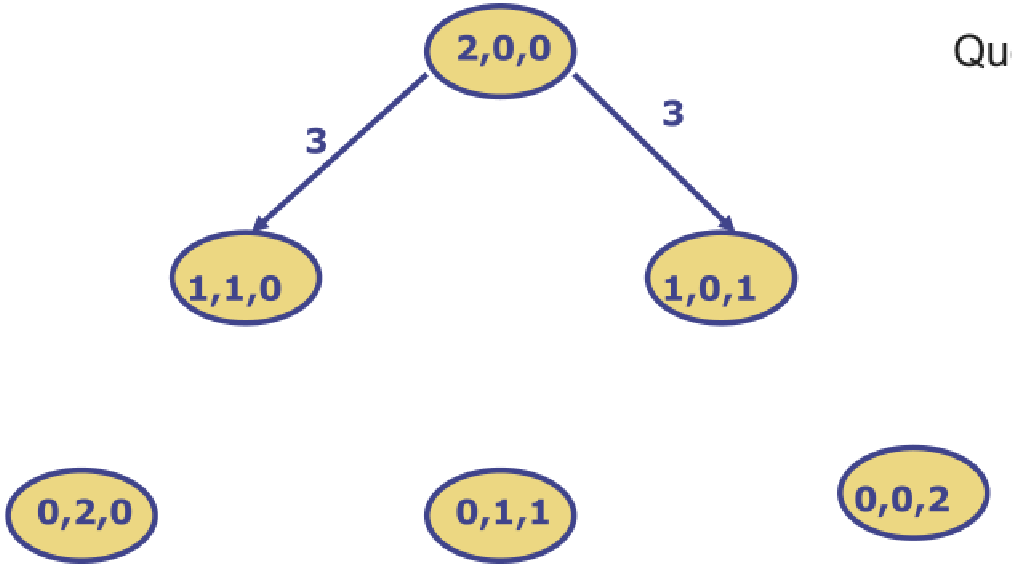
所以任务从CPU转移到两个硬盘的速率各自等于3/min

Markov Chain



第二步：
描述状态转移

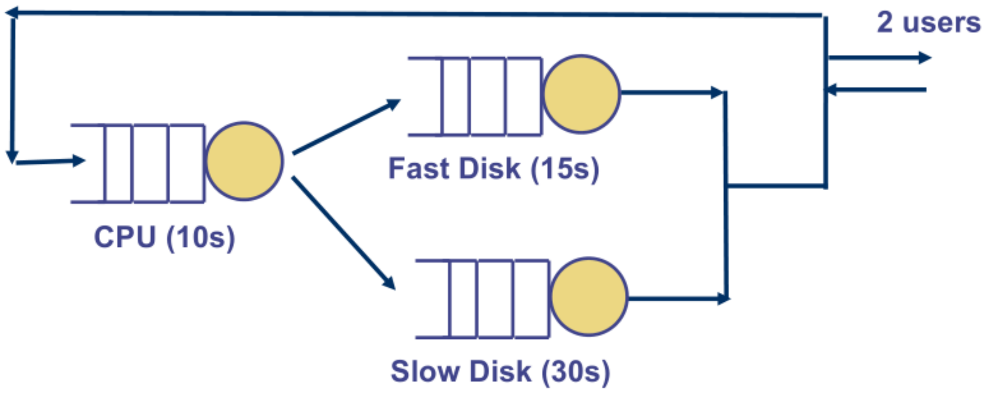
- Transition rate from (2,0,0) ➔ (1,1,0) = 3 transactions/minute
- Transition rate from (2,0,0) ➔ (1,0,1) = 3 transactions/minute



Question: What is the transition rate from (2,0,0) ➔ (0,1,1)?

从(2,0,0)不能直接转移到(0,1,1)
这两个状态并不相邻

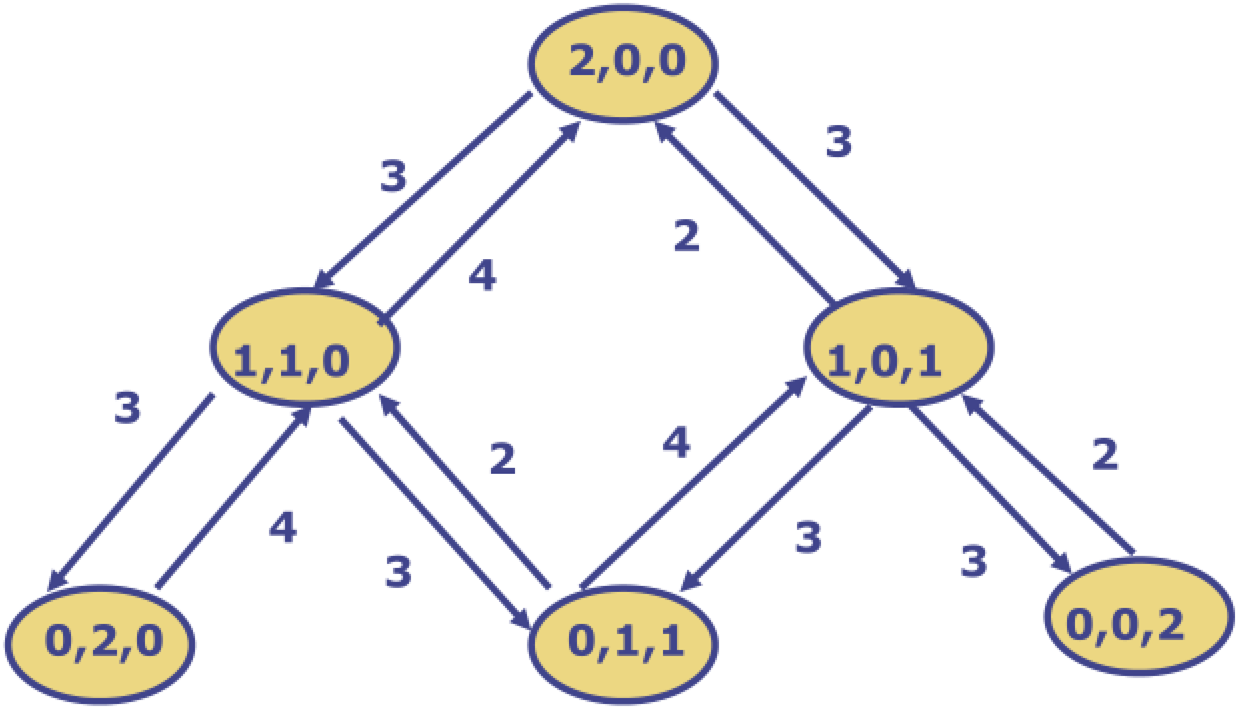
Markov Chain



第二步：
描述状态转移

From (1,1,0) there are 3 possible transitions

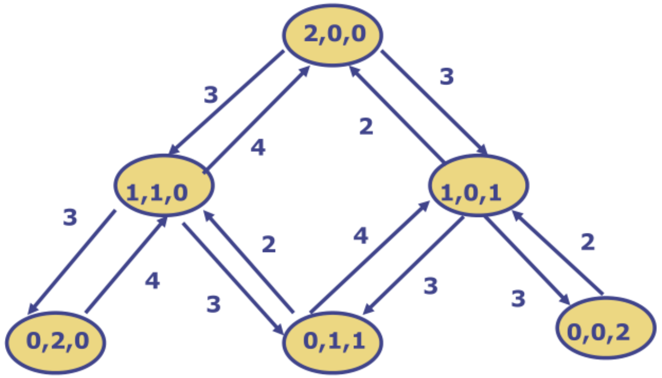
- Fast disk user goes back to CPU (2,0,0)
- CPU user goes to the fast disk (0,2,0), or
- CPU user goes to the slow disk (0,1,1)



Define
$P_{(2,0,0)} =$ Probability in state (2,0,0)
$P_{(1,1,0)} =$ Probability in state (1,1,0) etc.

Markov Chain

第三步：
建立状态平衡方程并求解



- You can write one flow balance equation for each state:

$$6\,P_{(2,0,0)} - 4\,P_{(1,1,0)} - 2\,P_{(1,0,1)} + 0\,P_{(0,2,0)} + 0\,P_{(0,1,1)} + 0\,P_{(0,0,2)} = 0$$

$$-3\,P_{(2,0,0)} + 10\,P_{(1,1,0)} + 0\,P_{(1,0,1)} - 4\,P_{(0,2,0)} - 2\,P_{(0,1,1)} + 0\,P_{(0,0,2)} = 0$$

$$-3\,P_{(2,0,0)} + 0\,P_{(1,1,0)} + 8\,P_{(1,0,1)} + 0\,P_{(0,2,0)} - 4\,P_{(0,1,1)} - 2\,P_{(0,0,2)} = 0$$

$$0\,P_{(2,0,0)} - 3\,P_{(1,1,0)} + 0\,P_{(1,0,1)} + 4\,P_{(0,2,0)} + 0\,P_{(0,1,1)} + 0\,P_{(0,0,2)} = 0$$

$$0\,P_{(2,0,0)} - 3\,P_{(1,1,0)} - 3\,P_{(1,0,1)} + 0\,P_{(0,2,0)} + 6\,P_{(0,1,1)} + 0\,P_{(0,0,2)} = 0$$

$$0\,P_{(2,0,0)} + 0\,P_{(1,1,0)} - 3\,P_{(1,0,1)} + 0\,P_{(0,2,0)} + 0\,P_{(0,1,1)} + 2\,P_{(0,0,2)} = 0$$

- However, there are only 5 linearly independent equations.
- Need one more equation:

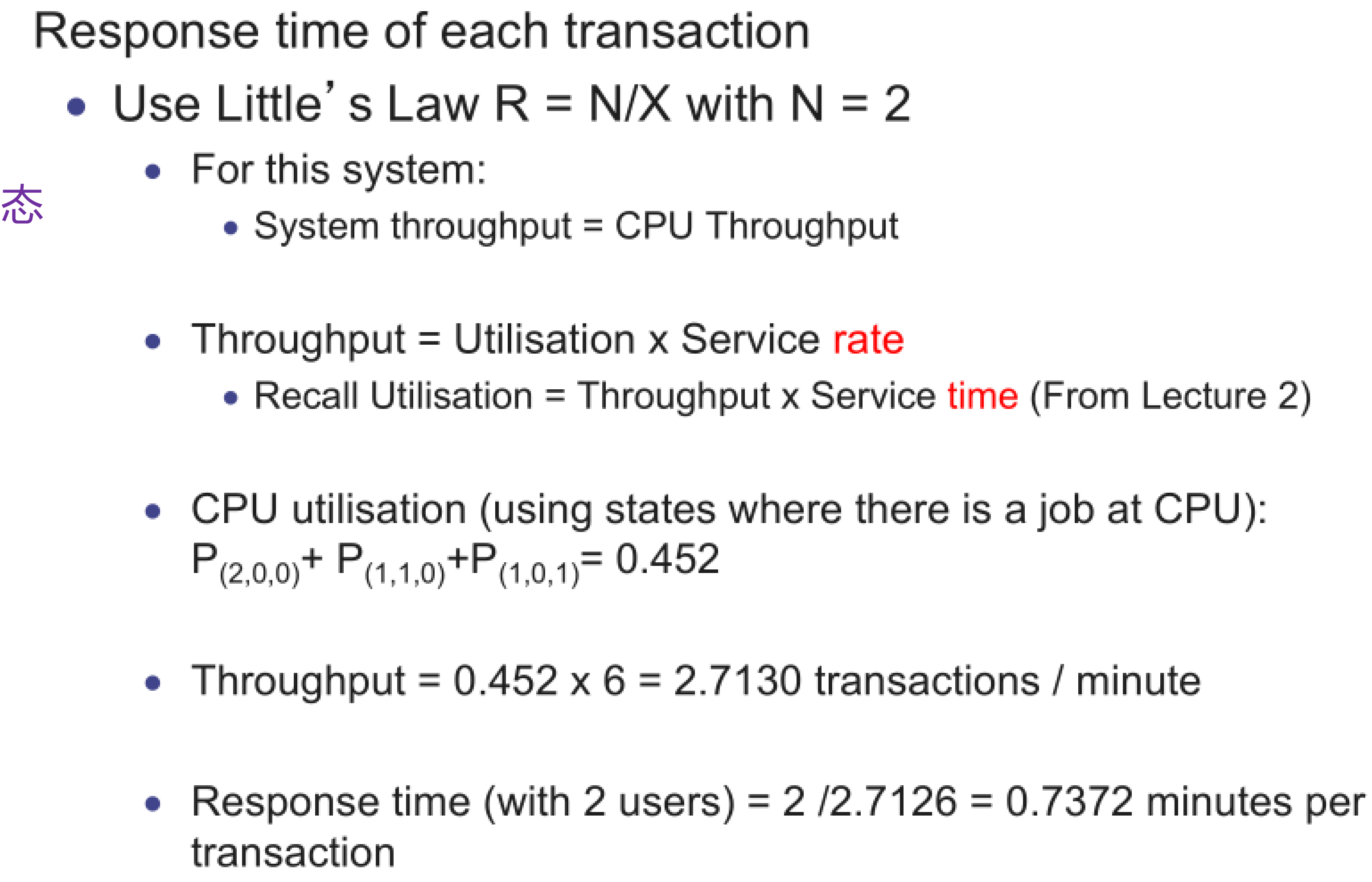
$$P_{(2,0,0)} + P_{(1,1,0)} + P_{(1,0,1)} + P_{(0,2,0)} + P_{(0,1,1)} + P_{(0,0,2)} = 1$$

Markov Chain

第四步：
使用结果进行分析

从状态的概率可以很容易得到
某个设备的Utilization（不为0的各个状态之和）

进而得到设备的Throughput
再利用Little’s Law得到
Response time



Response time of each transaction

- Use Little’s Law R = N/X with N = 2
  - For this system:
    - System throughput = CPU Throughput
  - Throughput = Utilisation x Service rate
    - Recall Utilisation = Throughput x Service time (From Lecture 2)
  - CPU utilisation (using states where there is a job at CPU): $P_{(2,0,0)}+ P_{(1,1,0)}+P_{(1,0,1)}= 0.452$
  - Throughput = 0.452 x 6 = 2.7130 transactions / minute
  - Response time (with 2 users) = 2 /2.7126 = 0.7372 minutes per transaction

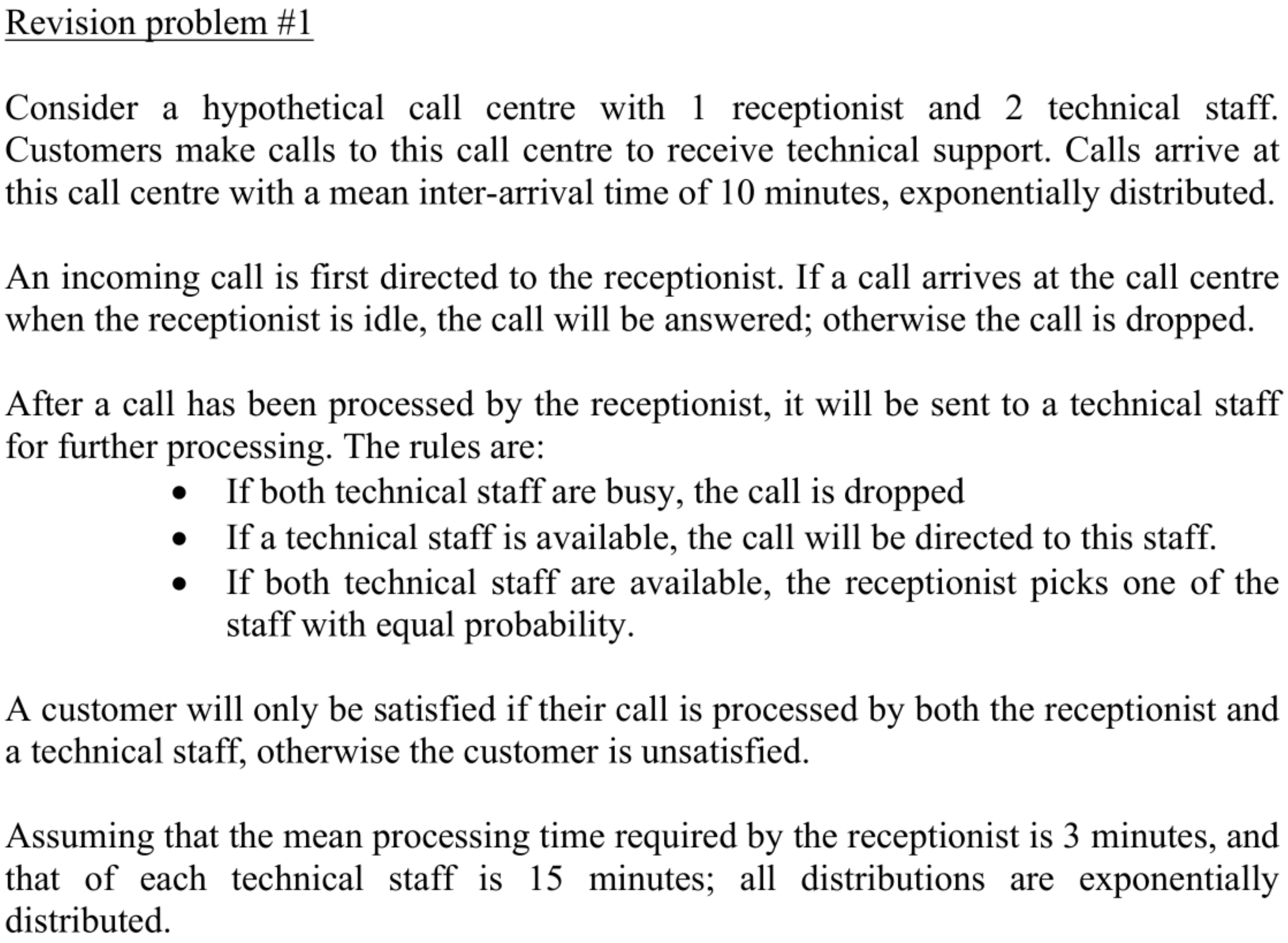
Markov Chain

What is the response time if the system have up to 4 users instead of 2 users only?

- You can’t use the previous Markov chain
- You need to develop a new Markov chain
  - The states are again (#users at CPU, #users at fast disk, #users at slow disk)
  - States are (4,0,0), (3,1,0), (1,2,1) etc.
  - There are 15 states
  - Determine the transition rates
  - Write down the balance equations and solve them.
  - Use the steady state probabilities and Little’s Law to determine the new response time

Markov Chain

Exercise:



Revision problem #1

Consider a hypothetical call centre with 1 receptionist and 2 technical staff. Customers make calls to this call centre to receive technical support. Calls arrive at this call centre with a mean inter-arrival time of 10 minutes, exponentially distributed.

An incoming call is first directed to the receptionist. If a call arrives at the call centre when the receptionist is idle, the call will be answered; otherwise the call is dropped.

After a call has been processed by the receptionist, it will be sent to a technical staff for further processing. The rules are:

- If both technical staff are busy, the call is dropped
- If a technical staff is available, the call will be directed to this staff.
- If both technical staff are available, the receptionist picks one of the staff with equal probability.

A customer will only be satisfied if their call is processed by both the receptionist and a technical staff, otherwise the customer is unsatisfied.

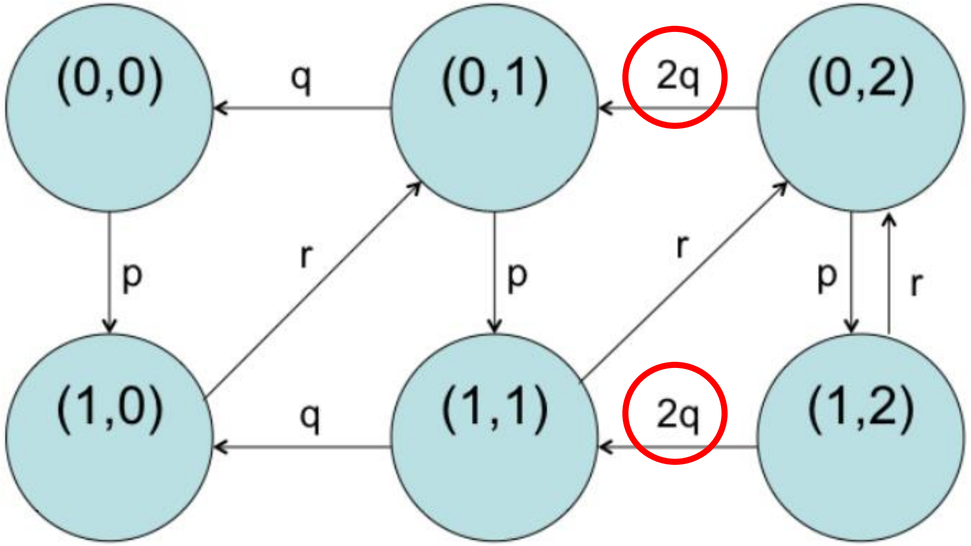
Assuming that the mean processing time required by the receptionist is 3 minutes, and that of each technical staff is 15 minutes; all distributions are exponentially distributed.

Markov Chain

Exercise:

Draw the appropriate Markov model. Label all arcs.

The states are 2-tuple *(x,y)* where *x* is the number of calls at the receptionist and *y* is the number of calls at the technical staff. We have *x* = 0 or 1; and, *y* = 0, 1, 2. There are altogether 6 states (0,0), (0,1), (0,2), (1,0), (1,1) and (1,2). The state space diagram is as follows:



Where p = 1/10, q = 1/15, r = 1/3

回忆一下之前的data server例子
我们是如何编码的

P(0,0) p = P(0,1) q
P(0,1) (p+q) = P(1,0) r + P(0,2) 2q
P(0,2) (p+2q) = P(1,1) r + P(1,2) r
P(1,0) r = P(0,0) p + P(1,1) q
P(1,1) (q+r) = P(0,1) p + P(1,2) 2q
P(1,2) (2q+r) = P(0,2) p

Where P(0,0) = Probability in State (0,0) etc.

P(0,0) + P(0,1) + P(0,2) + P(1,0) + P(1,1) + P(1,2) = 1

Markov Chain

Exercise:

Find the probability that a technical staff is busy

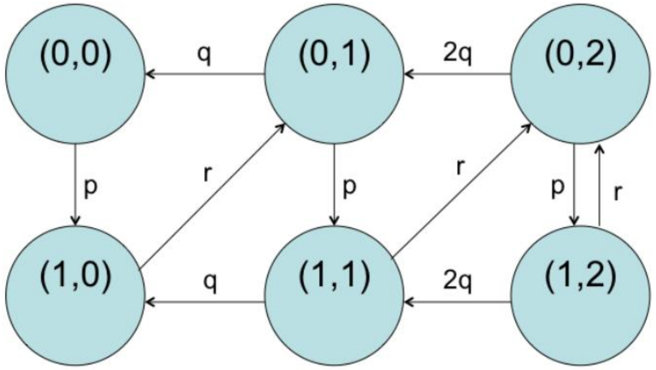
To find the probability that a technical staff is busy: This is the probability in the states (0,1), (0,2), (1,1) and (1,2) where there is at least one call being answered by a technical staff. The required probability is the sum P(0,1)+P(0,2)+P(1,1)+P(1,2)

What is the probability that an arriving call is dropped by the receptionist?

To find the probability that an arriving call is dropped by the receptionist: This is the probability that the receptionist is busy. The receptionist is busy in the states (1,0), (1,1) and (1,2). The required probability is the sum P(1,0) + P(1,1) + P(1,2)

What is the "good" throughput of the call centre (i.e. the rate of which satisfied customers leave the call centre) ?

P(0,1) q + P(0,2) 2q + P(1,1) q + P(1,2) 2q



Where p = 1/10, q = 1/15, r = 1/3