Question 1 (4 marks)

An interactive computer system consists of a CPU and a disk. The system was monitored for 90 minutes and the following measurements were taken:

Number of completed jobs 676

Number of CPU accesses 1,377

Number of disk accesses 1,515

CPU busy time 4,729 seconds

Disk busy time 2,565 seconds

(a) Determine the service demand for each device of the system.

According to service demand law:

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

we can determine the service demand by utilization of devices (U(j)) and the throughput of the system (X(0)).

Since we can get the utilization of devices by devices busy time (second), we should change the unit from minute to second, and we can get:

$$U(CPU) = \frac{B(CPU)}{T} = \frac{4729}{90 * 60} \approx 0.876$$

$$U(Disk) = \frac{B(Disk)}{T} = \frac{2565}{90 * 60} = 0.475$$

Since we can get the throughput by the number of requests completed, we can get:

$$X(0) = \frac{C(0)}{T} = \frac{676}{90 * 60} \approx 0.125$$

Thus, we can get the service demand respectively:

$$D(CPU) = \frac{U(CPU)}{X(0)} \approx 7(second)$$

$$D(Disk) = \frac{U(Disk)}{X(0)} = 3.8(second)$$

(b) In the lecture, we told you that you could identify the bottleneck of the system by using service demand. For the setting of this question, do you think it is possible to determine the bottleneck of the system without calculating the service demands? Justify your answer.

Yes, we can determine the bottleneck of the system without calculating the service demands. We can use the Utilisation law:

$$U = SX \leq 1, \; when \; S = \frac{B}{C}$$

Thus,

$$\frac{B}{C} * X \leq 1 \Rightarrow X \leq \frac{C}{B}$$

In this problem, we can get:

$$X(0) \leq \frac{C(0)}{max(B_i)} = \frac{676}{4729} \approx 0.143$$

Thus, we can get the bottleneck of the system without calculating the service demands:

$$X(0) \leq 0.143(job/s)$$

(c) Use bottleneck analysis to determine the asymptotic bound on the system throughput when there are 30 interactive users and the think time per job is 31 seconds.

Since the bottleneck analysis is:

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

and D(CPU) > D(Disk), we can get

$$\frac{1}{maxD_i} = \frac{1}{D(Disk)} = 0.143$$

Since we should consider the think time, we can get:

$$\frac{N}{\sum_{i=1}^{K} D_i} = \frac{N}{D(CPU) + D(Disk) + Think\,time} = \frac{30}{7 + 3.8 + 31} \approx 0.72$$

Thus, we can get the asymptotic bound:

$$X(0) \leq 0.143(job/s)$$

(d) Using your results in Part (c), compute the minimum possible response time of the computer system when the number of interactive users is 30.

From Part(c), we can get the throughput is 0.143 (job/s), from little's law, we can get:

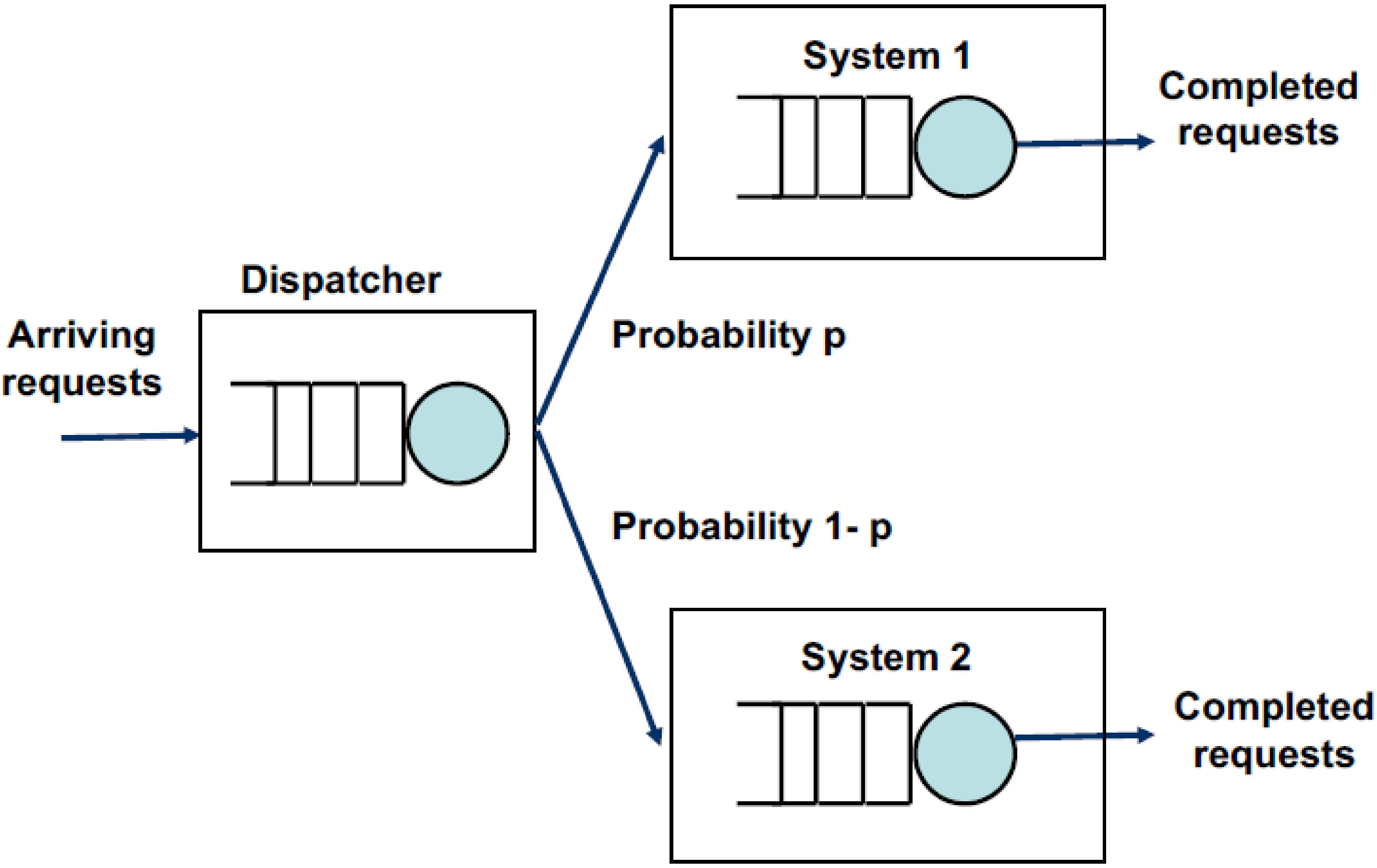
$$R = \frac{N}{X(0)}$$

And R = Response time + Think time, so we can get the minimum response time is:

$$R_{min} = \frac{N}{X(0)} - Think\,time = \frac{30}{0.143} - 31 \approx 178.79(second)$$

Question 2 (4 marks)

This question is based on the server farm in Figure 1. The server farm consists of a dispatcher and two computer systems, which are labelled as Systems 1 and 2. The purpose of the dispatcher is to route the incoming requests to either of the two systems. The policy of the dispatcher is to route an incoming request to System 1 with a probability of p, and to System 2 with a probability of 1 - p. You can make the following assumptions.



Answer the following questions:

(a) Find the probability p so that Systems 1 and 2 have the same utilisation.

Since that are two M/M/1, we can get:

$$U = \frac{\lambda}{\mu}$$

we should let U(System1) equal to U(System2), and the probability of request to System 1 is p and System 2 is 1 - p, Thus, we can get:

$$U(System1) = U(System2) \Rightarrow \frac{p\lambda}{\mu_1} = \frac{(1-p)\lambda}{\mu_2} \Rightarrow p\mu_2 = (1-p)\mu_1 \Rightarrow p = \frac{\mu_1}{\mu_1 + \mu_2} = \frac{10}{10+15} = \frac{2}{5}$$

Therefore, the probability p is equal to 0.4 so that Systems 1 and 2 have the same utilisation.

(b) Determine the mean response time of the server farm for the value of p that you have calculated in (a).

In this problem, we have two M/M/1 queue and we should calculate the mean response time respectively.

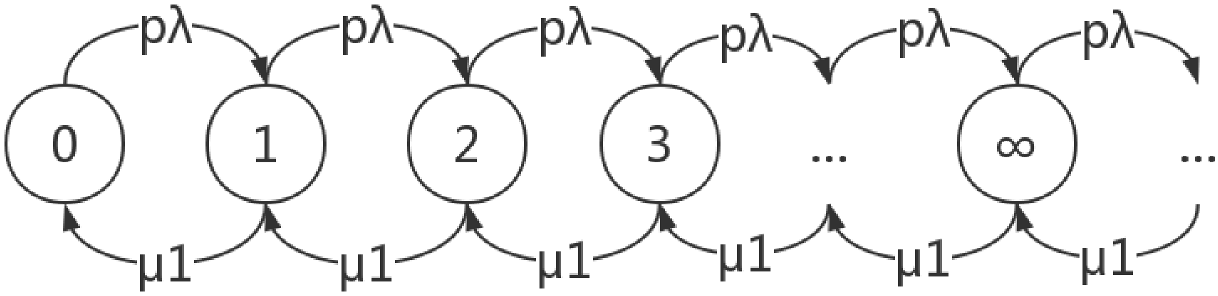
In System 1, we can define the states of the queue:

- 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). Since there are two M/M/1 queue and the

probability of requests to arrive at System 1 is p, the arrival rate to the System 1 is pλ. And there is only one server, so the processing rate is μ1. The below states are the same.

- State 2 = There is 1 job in the system (= 1 job at the server, 1 job queueing).
- State 3 = There are 3 jobs in the system (= 1 job at the server, 2 jobs queueing).
- State k = There are k jobs in the system (= 1 job at the server, k-1 jobs queueing).

The state transition diagram in System1:



M/M/1 state balance:

$$\begin{gathered} P_k = Prob.\, k\ jobs\ in\ system \\ p\lambda P_0 = \mu_1 P_1 \\ p\lambda P_1 = \mu_1 P_2 \\ p\lambda P_2 = \mu_1 P_3 \\ \cdots \end{gathered}$$

Therefore, in general, we can get:

$$P_k = (\frac{p\lambda}{\mu_1})^k P_0$$

The sum of probability is 1 and from above, p = 0.4, therefore:

$$\begin{gathered} Let\ \frac{p\lambda}{\mu_1} = \frac{0.4 * 20}{10} = 0.8 = \rho, \\ P_0 + P_1 + P_2 + \ldots + P_k + \ldots = 1 \Rightarrow (1 + \rho + \rho^2 + \ldots + \rho^k) P_0 = 1 \Rightarrow \\ P_0 = 1 - \rho\ Since\ \rho < 1 \Rightarrow P_k = (1 - \rho)\rho^k \end{gathered}$$

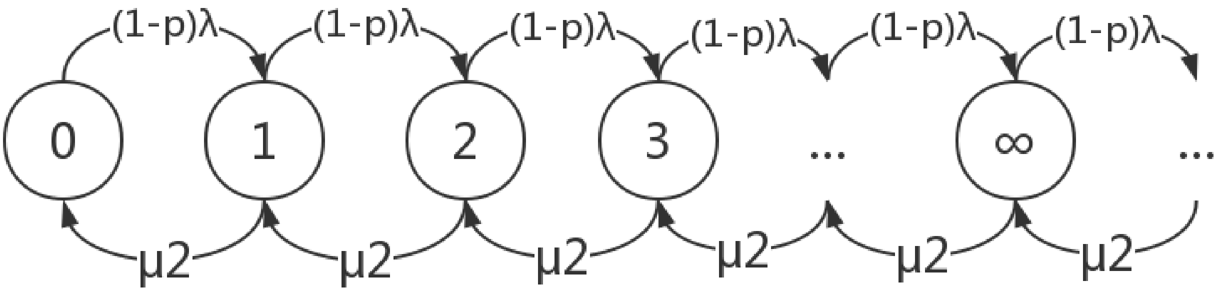
Thus, we can get the mean number of jobs in the system:

$$\begin{gathered} N_1 = \sum_{k=0}^{\infty} kP_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k = 0 + 1 * (1-\rho)\rho + 2 * (1-\rho)\rho^2 + \ldots + k(1-\rho)\rho^k + \ldots \\ Since\ p + x(p+q) + x^2(p+2q) + x^3(p+3q) + \ldots = \frac{p}{1-x} + \frac{xq}{(1-x)^2} \\ Let\ p = 0, q = 1 - \rho, x = \rho, \\ N_1 = \frac{0}{1-\rho} + \frac{\rho(1-\rho)}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{0.8}{0.2} = 4 \end{gathered}$$

For little's law: Mean number of jobs = Throughput * Response time, thus, we can get:

$$Mean\ Response\ time\ in\ System1 = \frac{N_1}{p\lambda} = \frac{4}{0.4*20} = 0.5$$

Same as above, in System2, we can get the diagram:



And we can get:

$$N_2 = \frac{0}{1-\rho} + \frac{\rho(1-\rho)}{(1-\rho)^2} = \frac{\rho}{1-\rho}\ when\ \rho = \frac{(1-p)\lambda}{\mu_2} = \frac{(1-0.4)*20}{15} = \frac{4}{5}$$

$$Thus, N_2 = \frac{\frac{4}{5}}{1-\frac{4}{5}} = 4$$

Therefore, we can get the mean respone time in System2:

$$Mean\ Response\ time\ in\ System2 = \frac{N_2}{(1-p)\lambda} = \frac{4}{0.6*20} = \frac{1}{3}$$

Thus, we can get the mean respon time in the whole system:

$$\begin{aligned} Mean\ Response\ time &= pMean\ Response\ time\ in\ System1 + (1-p)Mean\ Response\ time\ in\ System2 \\ &= 0.4*0.5 + 0.6*\frac{1}{3} \\ &= 0.4(second) \end{aligned}$$

(c) Determine the value of p so that the mean response time of the server farm is the smallest possible.

From above and little's law, we can get:

$$Mean\ Response\ time = pMean\ Response\ time\ in\ System1 + (1-p)Mean\ Response\ time\ in\ System2$$

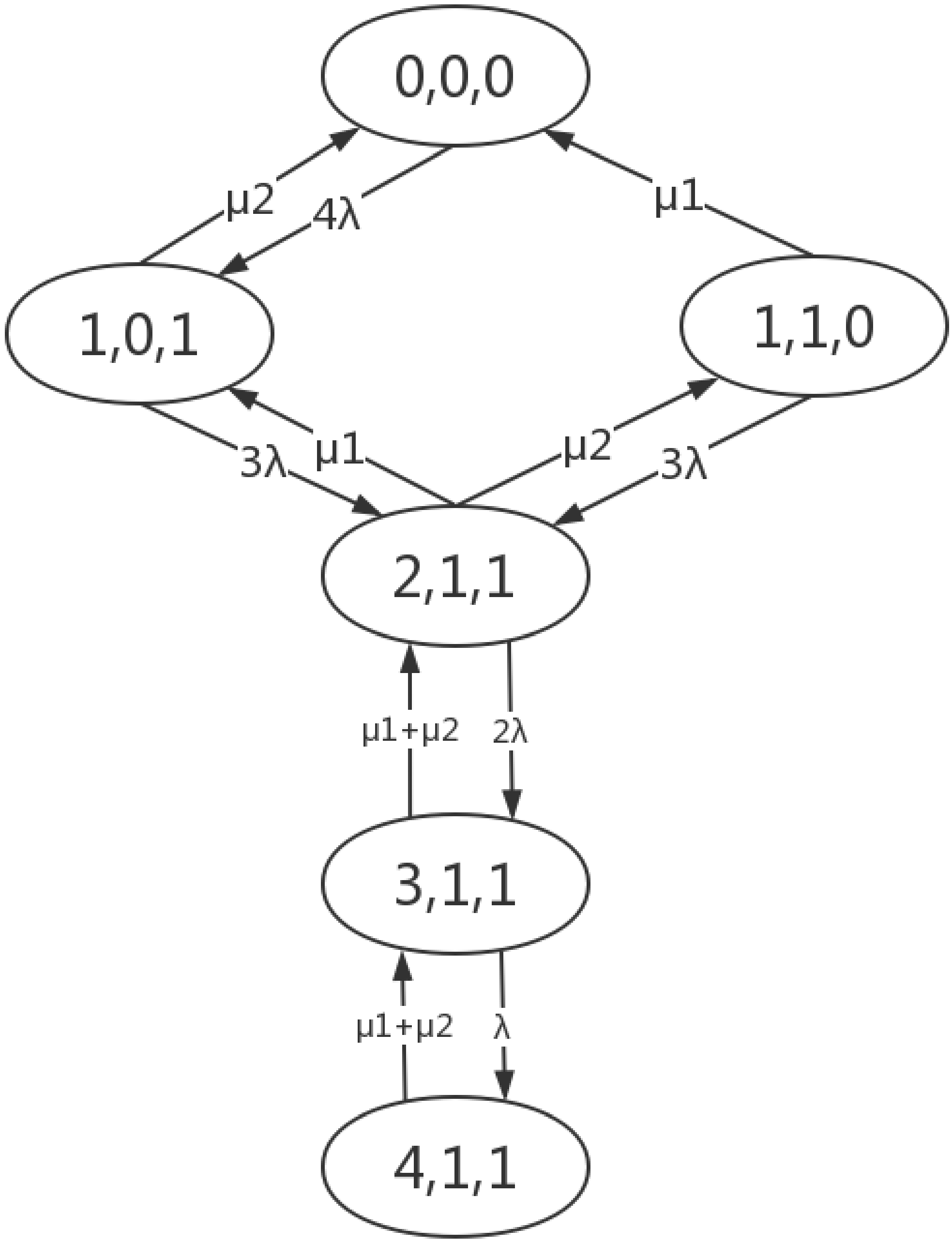
$$
\begin{aligned}
&= p\frac{N_1}{p\lambda} + (1-p)\frac{N_2}{(1-p)\lambda} \\
&= \frac{N_1 + N_2}{\lambda} \\
&= \frac{\frac{\frac{p\lambda}{\mu_1}}{1-\frac{p\lambda}{\mu_1}} + \frac{\frac{(1-p)\lambda}{\mu_2}}{1-\frac{(1-p)\lambda}{\mu_2}}}{\lambda} \\
&= \frac{p}{\mu_1 - p\lambda} + \frac{1-p}{\mu_2 - (1-p)\lambda} \\
&= \frac{p}{10 - 20p} + \frac{1-p}{15 - 20(1-p)}
\end{aligned}
$$

Thus, if we want to the smallest value of the mean response time, we should find the value of p to get the smallest value. We can use MATHLAB to solve this problem. The solution is in the attachment which is **a1_2c.m** in supp.zip
Finally, we can get the p which makes the mean response time of the server farm get the smallest value:

$$p = 0.3876$$

Question 3 (7 marks)
A data centre has 4 machines. For each machine, the time-to-next-failure is exponentially distributed with a mean of 600 minutes. The failure of a machine is independent of the others.
The data centre has a repair team consisting of a team leader and a trainee. The time required by a repair staff to repair a machine is exponentially distributed. The mean time taken by the team leader to repair a machine is 60 minutes. However, the trainee takes on average 90 minutes to repair a machine.
(a) Derive the state transition diagram for the Markov chain that describes the above problem. The diagram needs to include both the states and the transition rates. Explain how you arrive at your state transition diagram.
From above, we can get:

$$
\begin{aligned}
Let\ \lambda = 1/Mean - time - to - failure = \frac{1}{600} \\
\mu_1 = 1/Leader's\ mean\ service\ time\ to\ repair\ a\ machine = \frac{1}{60} \\
\mu_2 = 1/Trainee's\ mean\ service\ time\ to\ repair\ a\ machine = \frac{1}{90}
\end{aligned}
$$



State(0,0,0) represents 0 machine has failed

State(0,0,0) -> State(1,0,1) represents one of 4 machines has failed and the trainee repaired the machine, the rate of failure is 4λ

State(1,0,1) -> State(2,1,1)(leader repaired the incoming the failed machine) and State(1,1,0) -> State(2,1,1)(trainee repaired the incoming the failed machine) represent one of 3 remaining good machines has failed, the rate of failure is 3λ

State(1,1,0) -> State(0,0,0) and State(2,1,1) ->State(1,0,1) represent the 1 machine is repaired by the leader, the rate of repaired is µ1

State(1,0,1) -> State(0,0,0) and State(2,1,1) ->State(1,1,0) represent the 1 machine is repaired by the trainee, the rate of repaired is µ2

State(2,1,1) -> State(3,1,1) and State(3,1,1) -> State(4,1,1) represent one of remaining good machines has failed and it had to wait for repairing, the rate of failure is 2λ and λ respectively

State(3,1,1) -> State(2,1,1) and State(4,1,1) -> State(3,1,1) represent one of failed machines was repaired, since leader and trainee are repairing simultaneously, the both rate of repaired are µ1+µ2

(b) Derive the state balance equations for the Markov chain.

From above, we can get the value of λ, μ1 and μ2, so the state balance equations for each state:

$$\begin{gathered}
\frac{1}{150}P(0,0,0)-\frac{1}{90}P(1,0,1)-\frac{1}{60}P(1,1,0)+0*P(2,1,1)+0*P(3,1,1)+0*P(4,1,1)=0\\
-\frac{1}{150}P(0,0,0)+\frac{29}{1800}P(1,0,1)+0*P(1,1,0)-\frac{1}{60}P(2,1,1)+0*P(3,1,1)+0*P(4,1,1)=0\\
0*P(0,0,0)+0*P(1,0,1)+\frac{13}{600}P(1,1,0)-\frac{1}{90}P(2,1,1)+0*P(3,1,1)+0*P(4,1,1)=0\\
0*P(0,0,0)-\frac{1}{200}P(1,0,1)-\frac{1}{200}P(1,1,0)+\frac{7}{225}P(2,1,1)-\frac{1}{36}P(3,1,1)+0*P(4,1,1)=0\\
0*P(0,0,0)+0*P(1,0,1)+0*P(1,1,0)-\frac{1}{300}P(2,1,1)+\frac{53}{1800}P(3,1,1)-\frac{1}{36}P(4,1,1)=0\\
0*P(0,0,0)+0*P(1,0,1)+0*P(1,1,0)+0*P(2,1,1)-\frac{1}{600}P(3,1,1)+\frac{1}{36}P(4,1,1)=0\\
P(0,0,0)+P(1,0,1)+P(1,1,0)+P(2,1,1)+P(3,1,1)+P(4,1,1)=1
\end{gathered}$$

(c) Determine the steady state probability of all the states of the Markov chain.

To solve this problem, we can use MATHLAB, the program is in **a1_3c.m** in supp.zip.

Thus, we can get the steady state probability of all the states:

$$\begin{gathered}
P(0,0,0)=0.5918\\
P(1,0,1)=0.3081\\
P(1,1,0)=0.0313\\
P(2,1,1)=0.0611\\
P(3,1,1)=0.0073\\
P(4,1,1)=0.0004
\end{gathered}$$

(d) Compute the probability that at least three machines are available.

From above, we can get:

$$P(At\ least\ three\ machines\ are\ avilable)=P(0,0,0)+P(1,0,1)+P(1,1,0)=0.9312$$

(e) Compute the mean number failed machines.

For this problem, we can get this diagram:

State	Probability	Number of failure
0,0,0	P(0,0,0)	0
1,0,1	P(1,0,1)	1
1,1,0	P(1,1,0)	1
2,1,1	P(2,1,1)	2
3,1,1	P(3,1,1)	3

State	Probability	Number of failure
4,1,1	P(4,1,1)	4

Thus, the mean number failed machines is:

$$\begin{aligned}\overline{N_f} &= 0 * P(0,0,0) + 1 * P(1,0,1) + 1 * P(1,1,0) + 2 * P(2,1,1) + 3 * P(3,1,1) + 4 * \lambda P(4,1,1) \\ &= 0 * 0.5918 + 1 * 0.3081 + 1 * 0.0313 + 2 * 0.0611 + 3 * 0.0073 + 4 * 0.0004 \\ &= 0.3394 + 0.1222 + 0.0219 + 0.0016 \\ &= 0.4851\end{aligned}$$

(f) Compute the mean-time-to-repair (MTTR) for this data centre.
We should know the mean failure rate, so we can get this diagram:

State	Probability	Failure rate
0,0,0	P(0,0,0)	4λ
1,0,1	P(1,0,1)	3λ
1,1,0	P(1,1,0)	3λ
2,1,1	P(2,1,1)	2λ
3,1,1	P(3,1,1)	λ
4,1,1	P(4,1,1)	0

Thus, the mean failure rate is:

$$\begin{aligned}\overline{X_f} &= 4\lambda P(0,0,0) + 3\lambda P(1,0,1) + 3\lambda P(1,1,0) + 2\lambda P(2,1,1) + \lambda P(3,1,1) + 0 * P(4,1,1) \\ &= \frac{4}{600} * 0.5918 + \frac{3}{600} * 0.3081 + \frac{3}{600} * 0.0313 + \frac{2}{600} * 0.0611 + \frac{1}{600} * 0.0073 + 0 * 0.0004 \\ &\approx 0.00394533 + 0.0015405 + 0.0001565 + 0.00020367 + 0.00001217 \\ &= 0.00585817\end{aligned}$$

Since the mean-time-to-repair(MTTR) is the mean time of failure and for little’s law, we can get:

$$MTTR = \frac{Mean\ number\ \ failed\ machines}{Mean\ throughput} = \frac{\overline{N_f}}{\overline{X_f}} = \frac{0.4851}{0.00585817} \approx 82.81(minute)$$