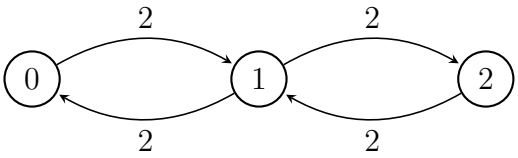
**SOLUTIONS**
**Final exam Stochastic Modelling**
December 22, 2021

**Question 1.** In an auto repair shop, there is a single repairman and room for (at most) two cars at a time (one undergoing repair and one waiting for repair). Cars drop by this repair shop according to a Poisson process of rate 2. A car that drops by and finds the shop already occupied by two other cars can not be taken in and leaves immediately. The repair times of the different cars are independent and distributed exponentially with rate 2. The order of repair is the order of arrival.

**(a) [2pt]** Formulate a CTMC based on which you can answer the subsequent parts of the question.

**Solution** $L(t)$ = number of cars in the shop at time $t$ is a CTMC with the transition diagram




**(b) [2pt]** At a certain moment the repair shop is empty. What is the expected time until it is full?

**Solution** Let $T_2 := \min\{t\colon L(t) = 2\}$ and $m_i := E(T_2 \mid L(0) = 0)$. The question is what is $m_0$.
By conditioning on the 1st jump we get the system

$$\begin{cases} m_0 = 1/2 + 1 * m_1, \\ m_1 = 1/4 + 1/2 * m_0 + 1/2 * 0. \end{cases}$$

As we plug the 2nd equation into the 1st, we get

$$m_0 = 1/2 + 1/4 + 1/2m_0 \quad \Leftrightarrow \quad 1/2m_0 = 3/4 \quad \Leftrightarrow \quad m_0 = 3/2.$$

**(c) [3pt]** What is the fraction of time that each of the following three situations occur: (i) the shop is empty, (ii) there is exactly one car in the shop,

(iii) the shop is full?

**Solution** The question is what is (i) $p_0^{occ}$, (ii) $p_1^{occ}$, (iii) $p_2^{occ}$. Since $L(\cdot)$ is an irreducible CTMC on a finite state space, the occupancy distribution exists and solves the balance and normalization equations:

$$\begin{cases} p_0 * 2 = p_1 * 2, & \text{balance for state 0} \\ p_1 * 2 = p_2 * 2, & \text{global balance for set } \{0,1\} \\ p_0 + p_1 + p_2 = 1. \end{cases}$$

From the two balance equations it follows that $p_0 = p_1 = p_2$ and then the normalization equation gives $p_0 = p_1 = p_2 = 1/3$.

I.e. the answers are (i) 1/3, (ii) 1/3, (iii) 1/3.

**(d) [2pt]** What fraction of cars drop by a full shop and have to leave without repair?

**Solution** By PASTA, it is $p_2^{occ} = 1/3$.

**(e) [2pt]** What is the time-average number of cars in the shop?

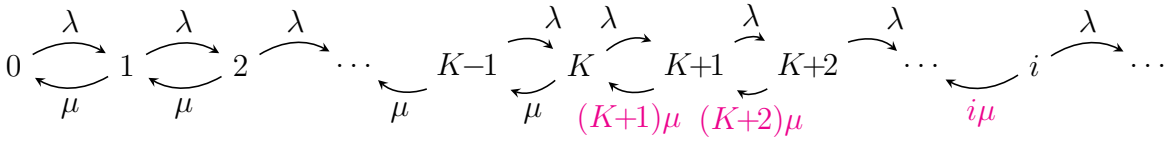
**Solution** It is $EL = p_0^{occ} * 0 + p_1^{occ} * 1 + p_2^{occ} * 2 = 1$.

**Question 2.** Customers arrive at a service facility according to a Poisson process of rate $\lambda$. Their service times are independent and distributed exponentially with mean $1/\mu$. The number of servers depends on how crowded it is in the system relative to a certain threshold $K$. There is one *permanent* server and a virtually unlimited amount of *support servers*. At all times when there are $K$ or less customers in the system, the permanent server is handling them on its own, one at a time in the order of arrival. At all times when there are more than $K$ customers in the system, the support servers are involved so that each customer is served at its dedicated server. (Mind that as soon as the number of customers drops down to $K$, the support is discontinued and it is the permanent server on its own again.)

**(a) [4pt]** Argue that the number of customers in the system is a CTMC. Argue *intuitively* whether it is a stable CTMC.

**Solution** $L(t)$ = number of cars in the shop at time $t$ is a CTMC with

the transition diagram




Intuitively, this CTMC is stable because, for large states, the arrival rate is below the service rate.

**(b) [5pt]** Find the limit and occupancy distribution. In particular, derive that

$$p_0^{lim} = p_0^{occ} = \left[ \frac{1-\rho^{K+1}}{1-\rho} + K! \left( e^{\rho} - \sum_{i=0}^{K} \frac{\rho^i}{i!} \right) \right]^{-1},$$

where $\rho := \lambda/\mu$.

**Solution** Below we find a solution to balance and normalization equations. This solution is both $p^{lim}$ and $p^{occ}$ since $L(t)$ is an irreducible CTMC.

The system for $p^{lim}$ and $p^{occ}$ is,

$$\begin{cases} \text{global balance for sets } \{0,\dots,i-1\}: \\ p_{i-1} * \lambda = p_i * \mu, \quad i = 1,\dots,K, \\ p_{i-1} * \lambda = p_i * i\mu, \quad i = K+1, K+2, \dots \\ \text{normalization: } \sum_{i=0}^{\infty} p_i = 1. \end{cases}$$

Hence, for $i = 1,\dots,K$,

$$p_i = \rho p_{i-1} = \rho^2 p_{i-2} = \dots = \rho^K p_0 \quad \text{(also true for } i = 0).$$

As for $i \geq K+1$, it follows that

$$p_{K+1} = \frac{\rho}{K+1} p_K,$$

$$p_{K+2} = \frac{\rho}{K+2} p_{K+1} = \frac{\rho^2}{(K+2)(K+1)} p_K,$$

$$\dots$$

$$p_i = \frac{\rho}{i} p_{i-1} = \frac{\rho^2}{i(i-1)} p_{i-2} = \dots = \frac{\rho^{i-K}}{i(i-1)\dots(K+1)} p_K = \frac{\rho^{i-K}}{i!/K!} \rho^K p_0 = K! \frac{\rho^i}{i!} p_0.$$

Now we plug the black and red relations into the normalization equation and get

$$1 = \sum_{i=0}^{\infty} p_i = \sum_{i=0}^{K} \rho^i p_0 + \sum_{i=K+1}^{\infty} K! \frac{\rho^i}{i!} p_0 = p_0 \left( \sum_{i=0}^{K} \rho^i + K! \sum_{i=K+1}^{\infty} \frac{\rho^i}{i!} \right) =$$
$$= p_0 \left( \frac{1-\rho^{K+1}}{1-\rho} + K!(e^\rho - \sum_{i=0}^{K} \frac{\rho^i}{i!}) \right).$$

To summarize, the last derivation implies that $p_0^{occ} = p_0^{lim}$ are as given in the question, and

$$\text{for } i = 0, 1, \ldots, K, \quad p_i^{occ} = p_i^{lim} = \rho^i p_0^{occ},$$
$$\text{for } i \geq K+1, \quad p_i^{occ} = p_i^{lim} = K! \frac{\rho^i}{i!} p_0^{occ}.$$

**(c) [4pt]** The higher the threshold $K$, the higher the fraction of customers that experience waiting times; *you can use this monotonicity fact without proof*. The table below provides the occupancy and limit probabilities for $\lambda = 5.5$, $\mu = 1$ and a few different values of $K$. These probabilities are rounded. Based on this table, if $\lambda = 5.5$ and $\mu = 1$, what is the biggest $K$ under which the fraction of customers that experience waiting does not exceed 5%?

| | $p_0$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ |
|---|---|---|---|---|---|---|---|---|
| $K=2$ | 0.002 | 0.011 | 0.063 | 0.115 | 0.158 | 0.174 | 0.159 | 0.125 |
| $K=3$ | 0.001 | 0.004 | 0.022 | 0.121 | 0.166 | 0.183 | 0.168 | 0.132 |
| $K=4$ | 0.0002 | 0.001 | 0.006 | 0.034 | 0.187 | 0.206 | 0.189 | 0.148 |
| $K=5$ | 0.00005 | 0.0002 | 0.002 | 0.008 | 0.046 | 0.252 | 0.231 | 0.181 |
| $K=6$ | 0.00001 | 0.00006 | 0.0003 | 0.002 | 0.01 | 0.056 | 0.311 | 0.244 |

**Solution** Customers that, upon arrival, see 1 to $K-1$ other customers in the system will experience waiting. By PASTA, the fraction $\Pi_W$ of such customers is

$$\Pi_W = p_1^{occ} + p_2^{occ} + \ldots + p_{K-1}^{occ}.$$

From the table we find that,

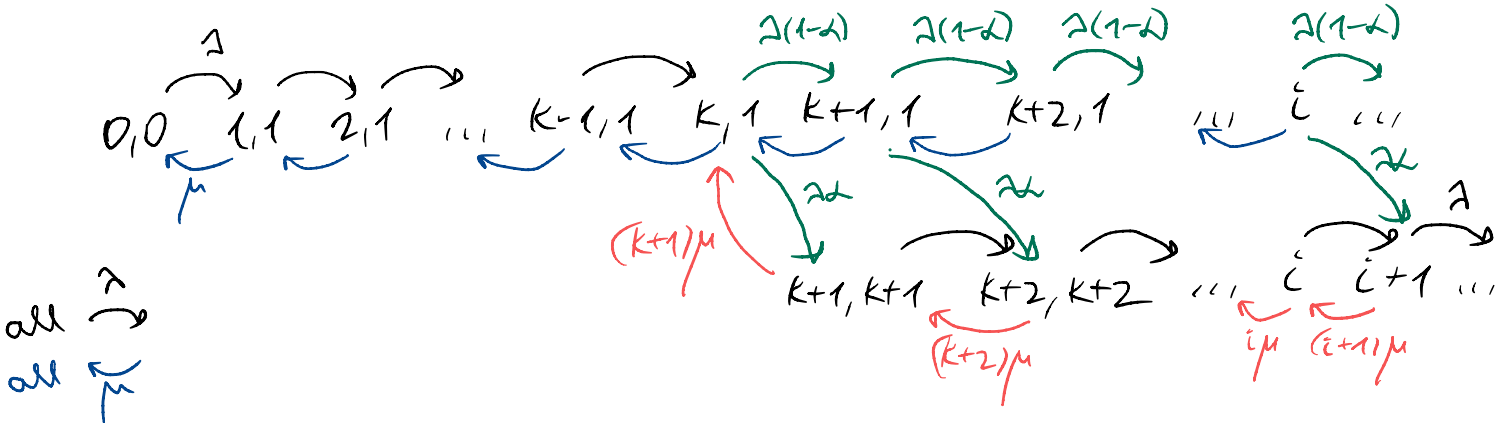
$$\text{for } K = 4, \quad \Pi_W = p_1^{occ} + p_2^{occ} + p_3^{occ} \approx 0.041 < 5\%,$$
$$\text{for } K = 5, \quad \Pi_W = p_1^{occ} + p_2^{occ} + p_3^{occ} + p_4^{occ} \approx 0.0562 > 5\%.$$

It is given that $\Pi_W$ increases in $K$, hence $K = 4$ is the biggest under which $\Pi_W$ does not exceed 5%.

**(d)** [**5pt**] Consider a new situation where there is a different procedure to involve the support servers. It is not the case anymore that the support servers necessarily get involved as soon as the number of customers exceeds $K$. With each new arrival that leads to more than $K$ customers in the system, the permanent server makes a request for support. The request is satisfied with probability $\alpha$ independently of all previous such requests. In case the request is satisfied, the support servers get involved immediately (so that each customer is served at its dedicated server) and remain involved until the number of customers drops down to $K$ (then it is the permanent server on its own again). In case the request is not satisfied, the permanent server remains on its own until the next opportunity to request support (i.e. until the next arrival that leads to more than $K$ customers in the system).

Formulate a CTMC that models this new situation and, in particular, keeps track of the number of customers in the system. Do you have to include any additional information in the state?

**Solution** $X(t)$ = (number of customers in the system, number of working servers) at time $t$ is a CTMC with the transition diagram




To explain: in states $i \geq K$, the arrival $\text{Exp}(\lambda)$ transition, once it happens, turns out to be to $(i+1, i+1)$ wp $\alpha$ (in case the request for support is satisfied) and to $(i+1, 1)$ wp $1-\alpha$ (in case the request for support is not satisfied). This is equivalent to two separate transitions $\text{Exp}(\lambda\alpha)$ to $(i+1, i+1)$ and $\text{Exp}(\lambda(1-\alpha))$ to $(i+1, 1)$ since $\text{Exp}(\lambda) \sim \min(\text{Exp}(\lambda\alpha), \text{Exp}(\lambda(1-\alpha)))$.

**Question 3.** There are two communication channels. Channel $A$ handles type $A$ messages, which are generated at rate $1/4$ and are all of fixed size 3. Channel $B$ handles type $B$ messages, which are generated at rate $1/30$ and are varied in size, the size distribution is (approximately) Normal with mean 18 and variance 9. The rates are per second and the sizes are transmission times in seconds at a unit transmission speed. Both channels transmit messages at a unit speed. There is an option of merging the two channels with the benefit of doubling the transmission speed and cutting transmission times by half.

**(a) [5pt]** Both channels transmit messages in the order in which the messages are generated. In particular, transmissions are delayed on average by 4.5 seconds at channel $A$ and by 13.875 seconds at channel $B$ *(these two values you can use without derivation)*. *Show* that, after merging, the average transmission delay would be approximately 5.13 seconds. Is there an improvement to the average transmission delay of type $A$ messages, the average transmission delay of type $B$ messages, or the weighted average delay?

**Solution** At the merger channel, the number of messages changing over time is an $M/G/1$ queue with arrival rate and service (transmission) time

$$\lambda = 1/4 + 1/30, \quad B = \begin{cases} 3/2, & \text{wp } \dfrac{1/4}{1/4+1/30}, \\ N(18,9)/2 = N(9,9/4), & \text{wp } \dfrac{1/30}{1/4+1/30}. \end{cases}$$

Since the order of service (transmission) is FIFO, the Pollaczek-Khinchine formula applies and the average waiting time (transmission delay) is given by

$$EW = \frac{\rho}{1-\rho}\frac{E(B^2)}{2E(B)}.$$

We have

$$E(B) = \frac{1/4}{1/4+1/30} * 3/2 + \frac{1/30}{1/4+1/30} * \overbrace{E(N(9,9/4))}^{=9},$$

$$E(B^2) = \frac{1/4}{1/4+1/30} * (3/2)^2 + \frac{1/30}{1/4+1/30} * \underbrace{E\left(N(9,9/4)\right)^2}_{=V+(E)^2=9/4+9^2}$$

$$\rho = \lambda E(B) = 1/4 * 3/2 + 1/30 * 9.$$

As we plug these values into the PK formula, we indeed obtain $EW \approx 5.13$.

*Comparison:* before merging, the delay at channel A is $EW_A = 4.5$ (as given), the delay at channel B is $EW_B = 13.875$ (as given), and the weighted average delay is

$$\overline{EW} = \frac{1/4}{1/4 + 1/30} * EW_A + \frac{1/30}{1/4 + 1/30} * EW_B \approx 5.61.$$

Comparing these to $EW \approx 5.13$ after merging, we conclude that the delay of type $A$ messages will become worse after merging, but the delay of type $B$ messages will improve by a lot and the weighted average delay will improve somewhat after merging.

**(b) [3pt]** For messages of type $A$, would their average sojourn time (from the moment a message is generated till its transmission is finished) improve after merging? And for messages of type $B$?

**Solution** Before merging, the average sojourn times of type $A$ and type $B$ messages are, respectively,

$$ES_A = EW_A + 3 = 4.5 + 3 = 7.5,$$
$$ES_B = EW_B + E(N(18, 9)) = 13.875 + 18 = 31.875.$$

After merging, the average sojourn times of type $A$ and type $B$ messages will be, respectively,

$$ES_A^m = EW + 3/2 \approx 5.13 + 3/2 = 6.63,$$
$$ES_B^m = EW + E[N(18, 9))/2] \approx 5.13 + 9 = 14.13.$$

We conclude that, after merging, the average sojourn time improves for both types of messages.

**(c) [3pt]** Will your answers to (a) and (b) change in case the order of transmission is non-preemptive LIFO presently and will remain such after merging?
*Hint:* no new calculation is required.

**Solution** No because the individual channels as well as the merger channel are $M/G/1$ models and PK applies under all non-preemptive non-size based service disciplines, including FIFO and non-preemptive LIFO. The calculations above done for FIFO will be the exactly the same for non-preemptive LIFO.

**(d)** [**5pt**] Will your answers to (a) and (b) change in case (the order of transmission is FIFO presently and will remain such after merging but) after merging there are *start-up delays* that are distributed exponentially with mean $1/\theta = 8$ seconds? In order to answer this question, do Mean Value Analysis for the average transmission delay and average number of delayed transmissions. *You can use without proof the fact that the system is stable and the fraction of time the channel is transmitting equals the load.*

*To clarify:* when, after a period of time with no messages to transmit, a new message is generated, the channel does not start transmitting immediately but remains idle for an additional period of time. This additional idling time is a start-up delay.

**Solution** The MVA equations for the delayed messages at the merger channel with Exp($\theta$) start-up delays are:

$$\begin{cases} \text{Little's law} & EL^q = \lambda EW, \\ \text{arrival relation} & EW = \rho * ER + (1-\rho) * E\text{Exp}(\theta) + EL^q * EB \\ & \quad = \rho * \dfrac{E(B^2)}{2E(B)} + (1-\rho) * \dfrac{1}{\theta} + EL^q * EB. \end{cases}$$

The arrival relation above comes up as follows,
- proportion $\rho$ of messages are generated while the channel is transmitting (by PASTA) and they have to wait for the remaining transmission time $R$;
- proportion $1-\rho$ of messages are generated while the channel is idling or starting up (by PASTA) and they have to wait for the full or remaining start-up time, Exp($\theta$) in either case (by the memorylessness of Exponential distribution);
- finally, each message has to wait for full transmission times of messages they find in the queue in front of them upon arrival (on average $EL^q$ of them by PASTA).

Solving the MVA equations (plug the Little's law into the arrival relation), we get

$$EW = \rho * \frac{E(B^2)}{2E(B)} + (1-\rho) * \frac{1}{\theta} + \underbrace{\lambda EW * EB}_{=\rho EW} \quad \Leftrightarrow \quad EW = \underbrace{\frac{\rho}{1-\rho}\frac{E(B^2)}{2EB}}_{\text{PK as in (a)}} + \frac{1}{\theta}.$$

I.e. to the average delay $EW$ in (a) and the average sojourn times $ES_A^m$, $ES_B^m$ in (b), we have to add the average start-up delay $1/\theta = 8$.

*Comparison:* after merging with start-up delays, the average delay and sojourn times for type $A$, type $B$ are

$$EW \approx \underbrace{5.13}_{(a)} + 8 = 13.13, \quad ES_A^m \approx \underbrace{6.63}_{(b)} + 8 = 14.63, \quad ES_B^m \approx \underbrace{14.13}_{(b)} + 8 = 22.23.$$

Comparing $EW$ to $EW_A = 4.5$, $EW_B = 13.875$, $\overline{EW} \approx 5.61$, we conclude that, after merging with start-up delays, there is a slight improvement in delay for type $B$ messages, but the delay of type $A$ messages and the weighted average delay become much worse. (The answer to (a) changes when start-up delays are added.)

Comparing $ES_A^m$, $ES_B^m$ to $ES_A = 7.5$, $ES_B = 31.875$, we conclude that, after merging with start-up delays, the average sojourn time of type $A$ messages becomes worse and the the average sojourn time of type $B$ messages improves. (The answer to (b) changes when start-up delays are added.)

**FORMULA SHEET**

**Erlang distribution.** If $S_n$ has an Erlang$(n, \mu)$ distribution, then

$$P(S_n > t) = \sum_{k=0}^{n-1} e^{-\mu t} \frac{(\mu t)^k}{k!} \qquad \text{and} \qquad f_{S_n}(t) = \mu\, e^{-\mu t} \frac{(\mu t)^{n-1}}{(n-1)!}.$$

**Residual time till next event.** Let $X$ be a generic inter-event time and $R$ the residual time till next event. Then

$$P(R \le x) = \frac{1}{E(X)} \int_0^x P(X > u)\, \mathrm{d}u \qquad \text{and} \qquad E(R) = \frac{E(X^2)}{2E(X)}.$$

**M/G/1 queue.** The waiting time $W$ under FIFO and the busy period $BP$ under work-conserving disciplines satisfy

$$E(W) = \frac{\rho}{1-\rho} \frac{E(B^2)}{2E(B)} = \frac{1}{2} \frac{\rho}{1-\rho}(1 + c_B^2)E(B), \quad \text{where } c_B^2 = \frac{VB}{(EB)^2}$$
$$E(BP) = \frac{E(B)}{1-\rho}.$$

**M/M/$c$ queue.** The probability of waiting $\Pi_W$, waiting time $W$ and sojourn time $S$ satisfy

$$\Pi_W = \frac{(c\rho)^c/c!}{(1-\rho)\sum_{i=0}^{c-1}(c\rho)^i/i! + (c\rho)^c/c!},$$
$$E(W) = \Pi_W \frac{1}{c\mu(1-\rho)} \qquad \text{and} \qquad P(W > t) = \Pi_W\, e^{-c\mu(1-\rho)t},$$
$$P(S > t) = \frac{\Pi_W}{1-c(1-\rho)}\, e^{-c\mu(1-\rho)t} + \left(1 - \frac{\Pi_W}{1-c(1-\rho)}\right) e^{-\mu t}.$$

**M/G/$c$/$c$ queue.** The blocking probability is

$$B(c, a) = \frac{a^c/c!}{\sum_{i=0}^{c} a^i/i!} \qquad \text{with } a = \lambda\, E(B) = c\rho.$$