UBINET: Performance Evaluation of Networks

Homework 1
To be returned on 17 September 2024 at 9 am

Homeworks are a personal effort. Copied solutions will get 0 for a grade.

1.1 A dysfunctional laptop

A laptop can be in one of four states: running CPU-intensive jobs, running memory-intensive jobs, swapping heavily, rebooting. A monitoring of the laptop's state on a minute basis reveals the following.

- If the laptop is running CPU-intensive jobs at a given minute, it will do the same in the following minute with a probability 19/20, and with the complementary probability it will be running memory-intensive jobs.
- If the laptop is running memory-intensive jobs at a given minute, it will do the same in the following minute with a probability 1/2, and with the complementary probability it will be swapping heavily.
- If the laptop is swapping heavily at a given minute, it will continue swapping with probability 2/3 or it will reboot with probability 1/3.
- If the laptop is rebooting, then it will continue rebooting with probability 3/4 or it will be running CPU-intensive jobs with probability 1/4.

For convenience, the four states of the laptop will be denoted A, B, C, and D in the order they were presented.

1. Say why the laptop's state can be described by a discrete-time Markov chain.
   Write the state-space $\mathcal{E}$.
   Draw the probability transition diagram.
   Write the transition matrix.
2. The laptop is running CPU-intensive jobs at minute 1. What is the probability that it is doing the same at minute 3?
   What is the probability that it is doing the same at minute 5?
3. Does the limiting distribution exist? If yes, say why.
   If yes, compute it.
4. The laptop's power consumption is 15 W when running CPU-intensive jobs, 20 W when running memory-intensive jobs, 25 W when swapping heavily, and 10 W when rebooting. What is the expected power consumption of the laptop in the stationary regime?

1.2 Exponential variables and Poisson processes

Let X and Y be two independent random variables having cumulative distribution function $F_X(\cdot)$ and $F_Y(\cdot)$. Let $Z = \min\{X, Y\}$ and $V = \max\{X, Y\}$. We define the following events:
A: "$X \leq t$"
B: "$Y \leq t$"
C: "$Z \leq t$"
D: "$V \leq t$"

1. Express events C and D in terms of events A and B or their complementary events.

2. Express the cumulative distribution function $F_Z(\cdot)$ of Z and $F_V(\cdot)$ of V in terms of $F_X(\cdot)$ and $F_Y(\cdot)$.

We assume now that the density function of X is $f_X(t) = \lambda e^{-\lambda t}$ for $t > 0$ and 0 otherwise, and that of Y is $f_Y(t) = \mu e^{-\mu t}$ for $t > 0$ and 0 otherwise, with $\lambda > 0$ and $\mu > 0$.

3. Compute the cumulative distribution functions $F_X(t)$ and $F_Y(t)$.

4. Compute the expectations $\mathrm{E}[X]$ et $\mathrm{E}[Y]$.

5. Compute $F_Z(t)$ and $\mathrm{E}[Z]$.
   What can you say about the distribution of the random variable Z?

Application: A data center is composed of two clusters and each time one of the clusters violates the server-level agreement (SLA), the data center itself violates the SLA which raises penalties to be paid by the provider. One of the clusters is more reliable than the other being more recent and better provisioned. The administrator of the data center estimates that the older cluster runs smoothly for a period of time that is exponentially distributed with a mean equal to 1 month. As for the newer cluster, the administrator estimates the period without SLA violation to be exponentially distributed with a mean equal to 2 months.

6. What is the nature of the *stochastic process* describing the number of occurrences of SLA violation over time of each of the clusters?

7. What is the nature of the process describing the number of occurrences of SLA violation over time of the data center?
   What is the rate of SLA violations of the data center?

8. The last SLA violation of the data center was in December 2020. What is the probability that a SLA violation will occur in March 2021?

9. We are in January 2021 and there has been no SLA violation since August 2020. What is the probability that no SLA violation will occur within 3 months?

UBINET: Performance Evaluation of Networks

Correction of homework 1

1.1 A dysfunctional laptop

1. The laptop's state is A when running CPU-intensive jobs, B when running memory-intensive jobs, C when swapping heavily and D when rebooting. The state-space is $\mathcal{E} = \{A, B, C, D\}$. It is enough to know the previous state of the laptop to know the transition probabilities to the future state. The Markov property is verified and we have a homogeneous discrete-time Markov chain. The transition matrix is

$$\mathbf{P} = \begin{bmatrix} \frac{19}{20} & \frac{1}{20} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} \end{bmatrix} .$$

   Sanity check: the elements in each row of $\mathbf{P}$ sum to 1.

2. The state at minute 1 is A then $\pi(1) = (1, 0, 0, 0)$. We want to know $\pi_A(3)$ and $\pi_A(5)$. We know that

$$\pi(3) = \pi(1)\mathbf{P}^2 , \quad \text{also} \quad \pi(5) = \pi(3)\mathbf{P}^2 .$$

   Let us compute $\mathbf{P}^2$, we get

$$\mathbf{P}^2 = \begin{bmatrix} \frac{361}{400} & \frac{29}{400} & \frac{10}{400} & 0 \\ 0 & \frac{1}{4} & \frac{7}{12} & \frac{1}{6} \\ \frac{1}{12} & 0 & \frac{4}{9} & \frac{17}{36} \\ \frac{17}{40} & \frac{1}{80} & 0 & \frac{9}{16} \end{bmatrix} .$$

   Sanity check: the elements in each row of $\mathbf{P}^2$ sum to 1.

   Therefore

$$\pi(3) = \left(\frac{361}{400}, \frac{29}{400}, \frac{10}{400}, 0\right) \quad \Rightarrow \quad \pi_A(3) = \frac{361}{400} = 0.9025$$

$$\pi_A(5) = \left(\frac{361}{400}, \frac{29}{400}, \frac{10}{400}, 0\right) \cdot \begin{bmatrix} \frac{361}{400} \\ 0 \\ \frac{1}{12} \\ \frac{17}{40} \end{bmatrix} \quad \Rightarrow \quad \pi_A(5) = \frac{391963}{480000} \approx 0.81659 .$$

3. The diagonal terms of $\mathbf{P}$ are all strictly positive, which means that each state in $\mathcal{E}$ has a loop. Therefore each state is aperiodic and so is the DTMC. There is a path that cycles through all states, namely $A \to B \to C \to D \to A$, which means that all states communicate with each other and so the DTMC is irreducible. As the state-space is finite, we can conclude that the DTMC is positive recurrent. The limiting distribution exists and is equal to the stationary distribution, which we compute next.

The stationary equations are

$$\begin{cases} \pi_A = \frac{19}{20}\pi_A + \frac{1}{4}\pi_D & \Rightarrow \quad \pi_D = \frac{1}{5}\pi_A \\ \pi_B = \frac{1}{20}\pi_A + \frac{1}{2}\pi_B & \Rightarrow \quad \pi_B = \frac{1}{10}\pi_A \\ \pi_C = \frac{1}{2}\pi_B + \frac{2}{3}\pi_C & \Rightarrow \quad \pi_C = \frac{3}{2}\pi_B = \frac{3}{20}\pi_A \\ \pi_D = \frac{1}{3}\pi_C + \frac{3}{4}\pi_D \ . \end{cases}$$

The normalizing equation allows to compute π_A. We can write

$$1 = \pi_A + \pi_B + \pi_C + \pi_D = \pi_A \left(1 + \frac{1}{10} + \frac{3}{20} + \frac{1}{5}\right) \quad \Rightarrow \quad \pi_A = \frac{20}{29} \ .$$

The stationary solution is

$$\pi = \left(\frac{20}{29}, \frac{2}{29}, \frac{3}{29}, \frac{4}{29}\right) \ .$$

Sanity check: each probability is in the interval $(0, 1)$ and the sum of the probabilities is equal to 1.

4. The expected power consumption in steady-state is

$$P = 15 \cdot \pi_A + 20 \cdot \pi_B + 25 \cdot \pi_C + 10 \cdot \pi_D = \frac{455}{29} \approx 15.69 \text{ W}.$$

1.2 Exponential variables and Poisson processes

1. We have

$$Z \leq t \quad \Leftrightarrow \quad \min\{X, Y\} \leq t \quad \Leftrightarrow \quad X \leq t \text{ or } Y \leq t \ .$$

Therefore $C = A \cup B$. We observe that $\overline{C} = \overline{A} \cap \overline{B}$. Also

$$V \leq t \quad \Leftrightarrow \quad \max\{X, Y\} \leq t \quad \Leftrightarrow \quad X \leq t \text{ and } Y \leq t \ .$$

Consequently $D = A \cap B$.

2. We observe first that as X and Y are independent rvs the events A and B are also independent. We also observe that $P(A) = F_X(t)$ and $P(B) = F_Y(t)$. The cumulative distribution functions of Z and V are

$$\begin{aligned} F_Z(t) &= P(Z \leq t) = P(C) = 1 - P(\overline{C}) = 1 - P(\overline{A} \cap \overline{B}) \overset{\text{indep.}}{=} 1 - P(\overline{A})P(\overline{B}) \\ &= 1 - (1 - F_X(t))(1 - F_Y(t)) \\ F_V(t) &= P(V \leq t) = P(D) = P(A \cap B) \overset{\text{indep.}}{=} P(A)P(B) = F_X(t)F_Y(t) \end{aligned}$$

3. We can compute

$$F_X(t) = \int_0^t f_X(x)dx = \int_0^t \lambda \mathrm{e}^{-\lambda x} dx = \lambda \left[\frac{\mathrm{e}^{-\lambda x}}{-\lambda}\right]_0^t = \lambda \left[\frac{\mathrm{e}^{-\lambda t} - 1}{-\lambda}\right] = 1 - \mathrm{e}^{-\lambda t}.$$

Similarly $F_Y(t) = 1 - \mathrm{e}^{-\mu t}$.

4. The computation can be done as in Example 23 page 64 of the lecture notes. A faster computation is possible as X and Y are positive rvs taking values in the interval $[0, +\infty)$. We can compute the expectations $\mathrm{E}[X]$ and $\mathrm{E}[Y]$ as follows.

$$\mathrm{E}[X] = \int_0^\infty P(X > t)dt = \int_0^\infty \mathrm{e}^{-\lambda t} = \left[\frac{\mathrm{e}^{-\lambda t}}{-\lambda}\right]_0^\infty = \frac{0-1}{-\lambda} = \frac{1}{\lambda}.$$

Similarly $\mathrm{E}[Y] = \frac{1}{\mu}$.

5. We use the expression found for $F_Z(t)$ in 2 and the expressions for $F_X(t)$ and $F_Y(t)$ found in 3. We obtain

$$F_Z(t) = 1 - \mathrm{e}^{-\lambda t}\mathrm{e}^{-\mu t} = 1 - \mathrm{e}^{-(\lambda+\mu)t}.$$

Similarly to what was done in 4, we can derive $\mathrm{E}[Z] = \frac{1}{\lambda+\mu}$. According to the cumulative distribution function found for Z, we can say that Z follows an exponential distribution with parameters $\lambda + \mu$. We can simply write $Z \sim \mathrm{Exp}(\lambda + \mu)$.

We have shown that the minimum among exponentially distributed rvs is an exponentially distributed rv whose parameter is the sum of the parameters.

6. Let X be the rv representing the time without SLA violations in the less reliable cluster, and Y be the same for the more reliable cluster. According to the statement, $\mathrm{E}[X] = 1$ month and $\mathrm{E}[Y] = 2$ months. Also $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\mu)$. We then have $\mathrm{E}[X] = 1/\lambda$ and $\mathrm{E}[Y] = 1/\mu$, therefore $\lambda = 1$ violation/month and $\mu = 0.5$ violation/month.

The number of occurrences of SLA violations in a given cluster is a counting stochastic process. The time between two increments of this process is the time between two consecutive violations. This time is exponentially distributed as observed by the administrator. By assuming that all inter-violations durations are independent, then the counting process is a Poisson process with rate equal to the parameter of the exponential distribution of the inter-event time.

Therefore, the SLA violations in the less reliable cluster form a Poisson process with rate $\lambda = 1$ violation/month, and the SLA violations in the more reliable cluster form a Poisson process with rate $\mu = 0.5$ violation/month.

7. The number of occurrences of SLA violations of the data center is the aggregation of the violation processes in the two clusters. Assuming the SLA violations in the two cluster are independent, then we can use Proposition 32 page 73 in the lecture notes to conclude that the SLA violations in the datacenter form a Poisson process with rate $\lambda + \mu = 1.5$ violations/month.

Alternatively, one can consider the time without SLA violation in the datacenter. It is nothing but $\min\{X, Y\} = Z$. We know that $Z \sim \mathrm{Exp}(\lambda + \mu)$, therefore the SLA violations in the datacenter form a Poisson process with rate $\lambda + \mu = 1.5$ violations/month.

8. As there are 3 months between December 2020 and March 2021, we are looking for the probability that Z is less than 3. We have

$$P(Z \leq 3) = F_Z(3) = 1 - \mathrm{e}^{-(\lambda+\mu)3} = 1 - \mathrm{e}^{-4.5} \approx 0.988891 \ .$$

9. We know that Z is already larger than 5 months and we are looking for the (conditional) probability that Z is larger than 8 $(= 5 + 3)$ months. We use the memoryless property of the exponential distribution (Example 21 in page 63) to write

$$P(Z \geq 8 | Z \geq 5) = P(Z \geq 3) = 1 - F_Z(3) = \mathrm{e}^{-(\lambda+\mu)3} = \mathrm{e}^{-4.5} \approx 0.0111 \ .$$

UBINET/SI5: Performance Evaluation of Networks

Homework 2

To be returned on 24 September 2024 at 9 am

Homeworks are a personal effort. Copied solutions will get 0 for a grade.

2.1 A functional database

A database can be in one of five states: idle (I), read operation (R), add operation (A), update operation (U), and delete operation (D) such that $\mathcal{E} = \{I, R, A, U, D\}$. The IT department has observed that the database remains idle for a time that is exponentially distributed with parameter μ_I. Read, add, update and delete operations all require an exponentially distributed time to complete, with respective parameters μ_R, μ_A, μ_U, and μ_D. It has been observed that a read operation is followed by either an update or a delete, with equal chances. After being idle, the database handles a read request or add request with equal probabilities. After an update, a delete or an add operation, the database becomes idle.

1. Explain why the database's state can be described by a continuous-time Markov chain.
2. Write the infinitesimal generator.
   Draw the transition rate diagram.
3. Is this CTMC ergodic? Explain why.
4. Compute the stationary distribution. (It may be helpful to use the notation $\frac{1}{C} = \frac{4}{\mu_I} + \frac{2}{\mu_R} + \frac{2}{\mu_A} + \frac{1}{\mu_U} + \frac{1}{\mu_D}$.)
5. What is the utilization rate of this database?
6. The power consumption of the server storing the database is 10 W when the database is idle, 60 W when read operations are ongoing, and 70 W with all write operations (add, update or delete). What is the expected power consumption in the stationary regime?
7. According to you, if one looks to minimize the power consumption, which operation better be optimized?
   Explain your reasoning.

UBINET/SI5: Performance Evaluation of Networks

Correction of homework 2

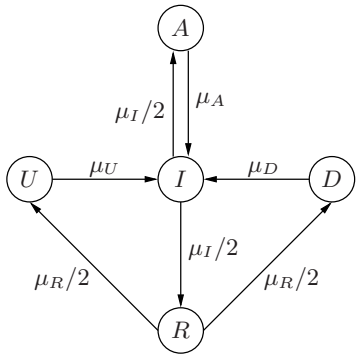
2.1 A functional database

1. According to the description of the process, we can see that it corresponds to construction rule number 1. The time spent in each state is exponentially distributed and once this time expires, the process changes state with some probability. These probabilities are 1 to go from states A, U or D to state I, and 1/2 to go from state I to R or A, and from state R to U or D. Therefore, according to construction rule 1 the stochastic process is a continuous-time Markov chain.

2. We have the following transition rates and infinitesimal generator (states are ordered according to the state-space $\mathcal{E} = \{I, R, A, U, D\}$):

$$\begin{array}{lcll} I & \to & R & \text{with rate } \mu_I/2 \\ I & \to & A & \text{with rate } \mu_I/2 \\ R & \to & U & \text{with rate } \mu_R/2 \\ R & \to & D & \text{with rate } \mu_R/2 \\ A & \to & I & \text{with rate } \mu_A \\ U & \to & I & \text{with rate } \mu_U \\ D & \to & I & \text{with rate } \mu_D \end{array} \qquad Q = \begin{bmatrix} -\mu_I & \mu_I/2 & \mu_I/2 & 0 & 0 \\ 0 & -\mu_R & 0 & \mu_R/2 & \mu_R/2 \\ \mu_A & 0 & -\mu_A & 0 & 0 \\ \mu_U & 0 & 0 & -\mu_U & 0 \\ \mu_D & 0 & 0 & 0 & -\mu_D \end{bmatrix}.$$

   The transition diagram of the CTMC is



3. The chain is irreducible as we can find a path that visits all states: $I \to R \to U \to I \to R \to D \to I \to A \to I$. As the state-space is finite, the irreducibility implies the ergodicity. The CTMC is ergodic.

4. The vector $\pi = (\pi_I, \pi_R, \pi_A, \pi_U, \pi_D)$ is the solution of the steady-state equations (including the normalization equation)

$$\begin{cases} \mu_I\,\pi_I = \mu_A\,\pi_A + \mu_U\,\pi_U + \mu_D\,\pi_D \\ \mu_R\,\pi_R = \frac{1}{2}\mu_I\,\pi_I \\ \mu_A\,\pi_A = \frac{1}{2}\mu_I\,\pi_I \\ \mu_U\,\pi_U = \frac{1}{2}\mu_R\,\pi_R \\ \mu_D\,\pi_D = \frac{1}{2}\mu_R\,\pi_R \\ \pi_I + \pi_R + \pi_A + \pi_U + \pi_D = 1\ . \end{cases}$$

Expressing all probabilities in terms of π_I yields

$$\pi_R = \frac{1}{2}\frac{\mu_I}{\mu_R}\pi_I \ , \quad \pi_A = \frac{1}{2}\frac{\mu_I}{\mu_A}\pi_I \ , \quad \pi_U = \frac{1}{4}\frac{\mu_I}{\mu_U}\pi_I \ , \quad \pi_D = \frac{1}{4}\frac{\mu_I}{\mu_D}\pi_I \ .$$

Using the notation suggested in the exercize statement, namely $\frac{1}{C} = \frac{4}{\mu_I} + \frac{2}{\mu_R} + \frac{2}{\mu_A} + \frac{1}{\mu_U} + \frac{1}{\mu_D}$, the normalization equation becomes

$$\frac{\mu_I}{4}\pi_I \frac{1}{C} = 1 \quad \Rightarrow \quad \pi_I = C \ \frac{4}{\mu_I} \ .$$

The steady-state distribution is then

$$\pi \ = \ C\left(\frac{4}{\mu_I}, \frac{2}{\mu_R}, \frac{2}{\mu_A}, \frac{1}{\mu_U}, \frac{1}{\mu_D}\right) \ .$$

Sanity check: each probability is in the interval $(0, 1)$ and the sum of the probabilities is equal to 1.

5. The database is utilized every time the process leaves state I. This occurs with rate $\mu_I \pi_I = 4C$. This is then the utilization rate.

6. The expected power consumption in steady-state is

$$P = 10\pi_I + 60\pi_R + 70(\pi_A + \pi_U + \pi_D) = C\left(\frac{40}{\mu_I} + \frac{120}{\mu_R} + \frac{140}{\mu_A} + \frac{70}{\mu_U} + \frac{70}{\mu_D}\right) \text{ W.}$$

7. The two operations that mostly impact the power consumption are the add operation followed by the read operation. To minimize the power consumption, it is important to reduce the expected duration of these operations, namely $1/\mu_A$ and $1/\mu_R$. In other words, one should try to increase μ_A and μ_R as much as possible. We observe that μ_I will be affected by changes in the duration of any operation time. Indeed, for the same number of requests by users (same utilization rate, C is constant), the idle time $1/\mu_I$ increases when the read/add/update/delete operations take less time, and the length of the change is the same. However, given the coefficients in the expression of the power consumption, the increase in $1/\mu_I$ is largely compensated by the decrease in the operation time, especially in $1/\mu_A$ or $1/\mu_R$. Therefore, the power consumption will decrease.

UBINET/SI5: Performance Evaluation of Networks

Homework 3

To be returned on 1 October 2024 at 9 am

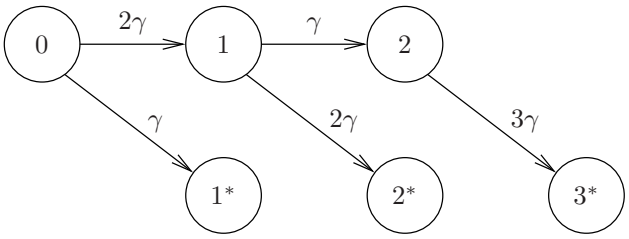
Homeworks are a personal effort. Copied solutions will get 0 for a grade.

3.1 A routing problem

Cars are getting equipped with wireless cards enabling direct wireless communications between them. The networks that are formed thanks to vehicle-to-vehicle communications are in the category of delay-tolerant networks. Routing messages in such networks can be achieved by using the so-called two-hop routing. In this protocol, the source of a message transmits a copy of its message to any node it meets in the network; such a node becomes a relay. Relays can only transmit messages to their respective destinations. As such a message can reach the destination after doing either one hop (case of source to destination transmission) or two hops (case of source to relay, then relay to destination transmissions). But there could be a potentially very large number of relays having a copy of the message when the destination receives the message.

Measurements collected from real-life encounters reveal that the inter-meeting times between *any two* nodes is roughly an exponential distribution. By assuming that inter-meeting times are independent of each other, the meeting process between any two nodes is Poisson with rate γ. We assume that when any two nodes meet they will exchange instantly all messages that they are allowed to transmit to each other, according to the two-hop routing protocol. Let $X(t)$ be the number of times a given message has been transmitted in the interval $[0, t[$ ($t = 0$ is seen as the message generation instant at the source). We have $X(0) = 0$.

We consider a network with only 4 nodes. Consequently, the state-space of the process $\{X(t), t > 0\}$ is $\mathcal{E} = \{0, 1, 2, 1^*, 2^*, 3^*\}$, where the asterisk denotes the fact that the message has reached its destination. For instance if the message has been transmitted a total of i times and has reached its destination by time t then $X(t) = i^*$. The transition diagram is



The performance of the two-hop routing is assessed through two metrics: (i) $T(0)$, the expected time to deliver a message to its destination given that $X(0) = 0$, and (ii) $\mathrm{E}[X]$, the expected number of transmissions until a message is delivered. (X is the stationary version of $X(t)$.) $T(0)$ is a quality metric while $\mathrm{E}[X]$ is a cost metric. Since b_{0,j^*} is the probability that j transmissions are needed for the message to be delivered given that $X(0) = 0$, we have $\mathrm{E}[X] := \sum_{j=1}^{3} j b_{0,j^*}$.

1. Say why $\{X(t), t > 0\}$ is an absorbing CTMC over $\mathcal{E}$. Mention specifically the transient/absorbing states.

2. Write the infinitesimal generator $\mathbf{Q}$.

3. Use appropriately $\mathbf{Q}$ to compute $T(0)$.

4. Derive the transition matrix $\mathbf{P}$ of the *embedded* Markov chain at *jump* times.

5. Use appropriately $\mathbf{P}$ to compute $\mathrm{E}[X]$.

We consider now a network with 5 nodes.

6. Write the new state-space $\mathcal{E}'$.

7. Explain the new transition rates between the possible states.

8. Draw the new transition diagram.
   Write the new infinitesimal generator $\mathbf{Q}'$.

9. Compute the expected time to deliver a message to its destination given that $X(0) = 0$.
   How does it compare with the same when there are four nodes in the network?
   Why could this be expected without performing the calculation?

10. Without making the computation, how does the expected number of transmissions until a message is delivered vary as the number of nodes in the network increases?
    Explain your reasoning.

UBINET/SI5: Performance Evaluation of Networks

Correction of homework 3

3.1 A routing problem

1. There is only one event that affects $X(t)$, the number of transmissions of a given packet at time t: it is a meeting between two nodes in the network. The inter-meeting time is an exponentially distributed rv. There could be multiple inter-meeting times that compete to change the state of the system (that is $X(t)$), so by using construction rule 2 we can say that $\{X(t), t > 0\}$ is a CTMC. It is an absorbing CTMC as states $\{1^*, 2^*, 3^*\}$ are absorbing. There are 3 transient states $\{0, 1, 2\}$. If the chain is being absorbed in state j^*, then this means that j transmissions were necessary to deliver the message to destination and the expected delivery time is the expected time until absorption in state j^* given that the chain was initially in state 0.

2. The infinitesimal generator is

$$\mathbf{Q} = \begin{bmatrix} -3\gamma & 2\gamma & 0 & \gamma & 0 & 0 \\ 0 & -3\gamma & \gamma & 0 & 2\gamma & 0 \\ 0 & 0 & -3\gamma & 0 & 0 & 3\gamma \\ & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

3. The infinitesimal generator can be seen as

$$\mathbf{Q} = \begin{bmatrix} \tilde{\mathbf{Q}} & \tilde{\mathbf{R}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

According to Proposition 11 in the lecture notes, the matrix giving the expected sojourn times in transient states is $\tilde{\mathbf{T}} = [t_{i,j}]_{0 \le i,j \le 2} = -\tilde{\mathbf{Q}}^{-1}$ and $T(0) = t_{0,0} + t_{0,1} + t_{0,2}$. We have

$$|\tilde{\mathbf{Q}}| = \begin{vmatrix} -3\gamma & 2\gamma & 0 \\ 0 & -3\gamma & \gamma \\ 0 & 0 & -3\gamma \end{vmatrix} = -27\gamma^3 > 0 \;\Rightarrow\; \tilde{\mathbf{T}} = \frac{1}{27\gamma} \begin{bmatrix} 9 & 6 & 2 \\ 0 & 9 & 3 \\ 0 & 0 & 9 \end{bmatrix} \;\Rightarrow\; T(0) = \frac{17}{27\gamma}.$$

Sanity check: all terms in $\tilde{\mathbf{T}}$ are positive or null.

Alternative solution: According to Corollary 2 in the lecture notes, the column vector giving the expected absorption times when initially in a transient state is the solution of $\tilde{\mathbf{Q}}\mathbf{T} = -\mathbf{1}$. We can write

$$\begin{bmatrix} -3\gamma & 2\gamma & 0 \\ 0 & -3\gamma & \gamma \\ 0 & 0 & -3\gamma \end{bmatrix} \cdot \begin{bmatrix} T(0) \\ T(1) \\ T(2) \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} \;\Rightarrow\; \begin{cases} -3\gamma T(0) + 2\gamma T(1) = -1 \\ -3\gamma T(1) + \gamma T(2) = -1 \\ -3\gamma T(2) = -1 \end{cases}$$

$$\Rightarrow \quad T(2) = \frac{1}{3\gamma}, \quad T(1) = \frac{4}{9\gamma}, \quad T(0) = \frac{17}{27\gamma}.$$

Sanity check: no absorbing time is negative.

4. We obtain the embedded Markov chain by observing $\{X(t),\, t \geq 0\}$ at *jump times*. The embedded Markov chain is a homogeneous, absorbing, DTMC. The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{2}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ & & & & & \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

5. The transition matrix can be seen as

$$\mathbf{P} = \begin{bmatrix} \mathbf{A} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

According to Proposition 9 in the lecture notes, the matrix giving the absorption probabilities is $\mathbf{B} = [b_{i,j^*}]_{\substack{0 \leq i \leq 2 \\ 1 \leq j \leq 3}} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{R}$. We have

$$|\mathbf{I} - \mathbf{A}| = \begin{vmatrix} 1 & -\frac{2}{3} & 0 \\ 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 1 \end{vmatrix} = 1 > 0 \quad \Rightarrow \quad \mathbf{B} = \begin{bmatrix} 1 & \frac{2}{3} & \frac{2}{9} \\ 0 & 1 & \frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{4}{9} & \frac{2}{9} \\ 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix}.$$

Sanity check: all terms in $\mathbf{B}$ are between 0 and 1 (they are probabilities). The sum of terms in each row is 1.

The expected number of transmissions per message delivery is

$$\mathrm{E}[X] = 1 \cdot b_{0,1^*} + 2 \cdot b_{0,2^*} + 3 \cdot b_{0,3^*} = 1 \cdot \frac{1}{3} + 2 \cdot \frac{4}{9} + 3 \cdot \frac{2}{9} = \frac{17}{9}.$$

6. The new state-space is $\mathcal{E}' = \{0, 1, 2, 3, 1^*, 2^*, 3^*, 4^*\}$.

7. The network is composed of 5 nodes so for each message there can be at most 3 relays. When the Markov chain is in state $i \in \{0, 1, 2\}$, this means that i relays have a copy of the message and $3 - i$ relays do not have a copy of it. The source can meet each of the $3-i$ relays after a time that is $\mathrm{Exp}(\gamma)$, so the first meeting between the source and any of the $3-i$ relays occurs after a time that is $\mathrm{Exp}((3-i)\gamma)$ and once this meeting occurs the Markov chain jumps to transient state $i+1$ as an additional transmission has just been made. On the other hand, when the Markov chain is in state $i \in \{0, 1, 2, 3\}$, this means that there are $i + 1$ nodes (i relays and one source) that can meet the destination and deliver it the message. Each of these $i + 1$ nodes can meet the destination after a time that is $\mathrm{Exp}(\gamma)$, so the first meeting between any of the $i + 1$ nodes and the destination occurs after a time that is $\mathrm{Exp}((i+1)\gamma)$ and once this meeting occurs the Markov chain jumps to absorbing state $(i+1)^*$ as an additional transmission has just been made and the message has been delivered.

8. The new infinitesimal generator is

$$\mathbf{Q}' = \begin{bmatrix} -4\gamma & 3\gamma & 0 & 0 & \gamma & 0 & 0 & 0 \\ 0 & -4\gamma & 2\gamma & 0 & 0 & 2\gamma & 0 & 0 \\ 0 & 0 & -4\gamma & \gamma & 0 & 0 & 3\gamma & 0 \\ 0 & 0 & 0 & -4\gamma & 0 & 0 & 0 & 4\gamma \\ & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

9. To distinguish it from the same quantity for the case when there are 4 nodes in the network we add a prime to the notation. So we are looking for $T'(0)$. We use Corollary 2. We need to find the solution of $\tilde{\mathbf{Q}}'\mathbf{T}' = -\mathbf{1}$. We have

$$\begin{bmatrix} -4\gamma & 3\gamma & 0 & 0 \\ 0 & -4\gamma & 2\gamma & 0 \\ 0 & 0 & -4\gamma & \gamma \\ 0 & 0 & 0 & -4\gamma \end{bmatrix} \cdot \begin{bmatrix} T'(0) \\ T'(1) \\ T'(2) \\ T'(3) \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \Rightarrow \begin{cases} -4\gamma T'(0) + 3\gamma T'(1) = -1 \\ -4\gamma T'(1) + 2\gamma T'(2) = -1 \\ -4\gamma T'(2) + \gamma T'(3) = -1 \\ -4\gamma T'(3) = -1 \end{cases}$$

$$\Rightarrow \quad T'(3) = \frac{1}{4\gamma}, \quad T'(2) = \frac{5}{16\gamma}, \quad T'(1) = \frac{13}{32\gamma}, \quad T'(0) = \frac{71}{128\gamma}.$$

Sanity check: no absorbing time is negative.

With 4 nodes, the expected time to deliver the message is $T(0) = \frac{17}{27\gamma} \approx \frac{0.63}{\gamma}$. With 5 nodes, $T'(0) = \frac{71}{128\gamma} \approx \frac{0.55}{\gamma}$. We have $T'(0) < T(0)$. As the number of nodes increases, the expected delivery time decreases. This result could be expected for the following reason. By allowing the use of relays, we enable the possibility to have more nodes carrying the message meet the destination and deliver the message sooner. By increasing the number of nodes, what we are increasing is the number of relays, so the time until one of them meets the destination decreases on average. As the inter-meeting time is exponentially distributed, by having more possible meetings, the time that one of them occurs is the minimum among all of the inter-meeting times, with a rate that increases with the number of possible meetings. Therefore we expect $T(0)$ to decrease as the number of relays (i.e. the number of nodes in the network) increases.

10. As the number of nodes in the network increases, the number of relays increases and so does the number of transmissions. There will be more copies of each message in the network. Therefore, the cost $\mathrm{E}[X]$ should increase with the number of nodes.

UBINET/SI5: Performance Evaluation of Networks

Homework 4
To be returned on 8 October 2024 at 9 am

Homeworks are a personal effort. Copied solutions will get 0 for a grade.

4.1 Fake news

A malicious person is hired to spread fake news over a social network. Each person on this social network that is exposed to the fake news will spread it in turn. We assume that each person—beside the source of the fake news—will eventually realize that the news is fake and will stop spreading it. However, the fake news will continue to spread over the social network as its membership is infinite.

It is assumed that when $i \geq 1$ persons are spreading the fake news

- it takes a random time for *each* spreader to convince a new person to start spreading the fake news; this convincing time is exponentially distributed with rate $\alpha > 0$.
- it takes a random time for *each* person (other than the source) spreading the fake news to awaken and realize that it is fake and stop then diffusing it; this awakening time is exponentially distributed with rate $\beta > 0$.

All convincing and awakening durations introduced above are assumed to be mutually independent. Denote $\rho := \alpha/\beta$. Let $X(t)$ be the number of persons spreading the fake news at time t and π_i be the stationary probability that i persons are spreading the fake news.

1. What is the nature of the stochastic process $\{X(t), t \geq 0\}$?
   What is the state-space (denoted by $\mathcal{E}$) on which this process is irreducible?
   Draw the transition rate diagram of this process.
2. Compute π_i for all $i \in \mathcal{E}$.
   What is the stability condition?
3. Compute $\overline{N}$, the expected number of fake news spreaders in steady-state.
4. Compute $\overline{\lambda}$, the average convincing rate (the rate of turning persons to become spreaders of fake news).
5. Compute $\overline{T}$, the average duration a person acts as a spreader of fake news.
   Any comment?
6. Assume now that the awakening time is exponentially distributed with rate $\beta/(i-1)$ when there are i spreaders of fake news. (This captures the idea that when more persons are spreading a fake news, it takes more time for a person to realize that it is fake.)
   Under what condition would this system be stable?

4.2 Dimensioning a server

A company's server consists of a single processor computer that serves a queue of jobs in a FIFO fashion. Jobs arrive according to a Poisson process with rate λ and require a service time that is exponentially distributed with rate μ. This server is stable because $\lambda < \mu$. Let $\overline{T}$ denote the mean response time of the server.

After a merger, the company starts offering a much wider range of services. The manager expects then the jobs arrival rate to double in the days following the merger. It is crucial that customers will not notice any degradation in their services. In a hurry, the manager orders a new server with twice the speed of the actual company's server.

1. What is the utilization ρ_{new} of the new server with the new arrival rate?
   Comment on the comparison with the old server before the merger.

2. What is the response time $\overline{T}_{\text{new}}$ of the new server with the new arrival rate?
   Comment on the comparison with the old server before the merger.

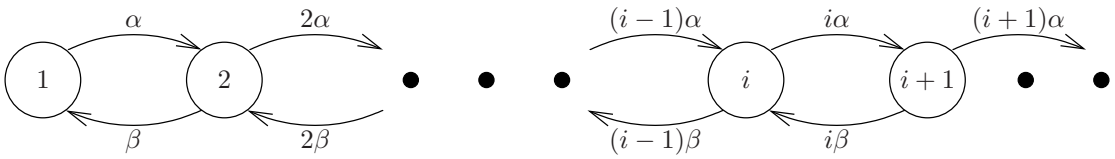
3. If the main objective is to maintain the same quality of service for customers, what would you have suggested to buy as new server for the company?

UBINET/SI5: Performance Evaluation of Networks

Correction of homework 4

4.1 Fake news

1. The number of spreaders of fake news at time t, $X(t)$, takes values in $\mathcal{E} = \mathbb{N}^* = \{1, 2, 3, \ldots\}$ (there is always at least one spreader that is the source). The convincing and awakening processes compete to change the state of the system (the number of spreaders), and each duration is exponentially distributed. By construction rule 2, $\{X(t), t > 0\}$ is a CTMC. Furthermore, the convincing and awakening processes make the number of spreaders of fake news increase/decrease by one at each transition. Clearly, $\{X(t), t > 0\}$ is a birth and death process with birth rate $\lambda_i = i\alpha$ (i processes are competing) and death rate $\mu_i = (i-1)\beta$ (if i persons are spreading the fake news, only $i-1$ of them will eventually stop doing so, so there will be $i-1$ awakening processes initiated in parallel). The transition diagram is



2. We can write the global balance equations for $S = \{1, \ldots, i-1\}$, and for $i \geq 2$,

$$(i-1)\alpha\pi_{i-1} = (i-1)\beta\pi_i \quad \Rightarrow \quad \pi_i = \frac{\alpha}{\beta}\pi_{i-1} = \rho\pi_{i-1} = \rho^2\pi_{i-2} = \ldots = \rho^{i-1}\pi_1$$

where we have used $\rho := \alpha/\beta$.

Instead, we can start directly with the stationary distribution of a birth and death process that verifies the following equation, for $i = 2, 3, \ldots$

$$\pi_i = \frac{\lambda_{i-1}\lambda_{i-2}\cdots\lambda_1}{\mu_i\mu_{i-1}\cdots\mu_2}\pi_1 = \frac{(i-1)!\alpha^{i-1}}{(i-1)!\beta^{i-1}}\pi_1 = \rho^{i-1}\pi_1.$$

This equation is also true for $i = 1$. The normalizing condition gives

$$\sum_{i=1}^{\infty}\pi_i = 1 \quad \Rightarrow \quad \pi_1 = \frac{1}{1 + \sum_{i=2}^{\infty}\rho^{i-1}} = \frac{1}{\sum_{i=0}^{\infty}\rho^i}.$$

This probability is positive if the sum in the denominator of π_1 converges. The denominator is the sum of the terms of a geometric progression, it converges for $\rho < 1$. In this case

$$\pi_1 = 1 - \rho\ , \qquad \Rightarrow \quad \pi_i = (1-\rho)\rho^{i-1}, \text{ for } i = 1, 2, \ldots$$

The stability condition is $\rho < 1$ or equivalently that $\alpha < \beta$.

3. The expected number of fake news spreaders in steady-state is

$$\overline{N} = \sum_{i=1}^{\infty} i\pi_i = (1-\rho)\sum_{i=1}^{\infty} i\rho^{i-1} = (1-\rho)\cdot\frac{1}{(1-\rho)^2} = \frac{1}{1-\rho}\ .$$

4. The average convincing rate is

$$\overline{\lambda} = \sum_{i=1}^{\infty} \lambda_i\pi_i = \sum_{i=1}^{\infty} i\alpha\pi_i = \alpha\overline{N} = \frac{\alpha}{1-\rho}\ .$$

5. We will use Little's law on the set of spreaders of fake news (this is the black box), there are on average $\overline{N}$ items in the box (spreaders of fake news), and the arrival rate into the box is $\overline{\lambda}$. The average time a person acts as a spreader of fake news (expected time inside the box) is

$$\overline{T} = \frac{\overline{N}}{\overline{\lambda}} = \frac{1}{\alpha}.$$

   A spreader of fake news will spread it for an average duration of $1/\alpha$. This is also the average time a person takes to convince another one to start spreading the fake news.

   Remark: Since $\alpha < \beta$ then $1/\alpha > 1/\beta$, that is the average time a random person acts as a spreader of fake news is larger than the average time needed by a spreader other than the source (the malicious person) to awaken and stop spreading the news. This is a consequence of the fact that the source never awakens and this increases the average time spent by a random person as a fake news spreader.

6. The new death rate is $\mu_i = (i-1)\beta/(i-1) = \beta$. The stationary distribution, for $i = 2, 3, \ldots$, becomes

$$\pi_i = \frac{(i-1)!\alpha^{i-1}}{\beta^{i-1}}\pi_1 = (i-1)!\rho^{i-1}\pi_1,$$

   where the last equality holds also for $i = 1$. The normalizing condition gives

$$\sum_{i=1}^{\infty}\pi_i = 1 \quad \Rightarrow \quad \pi_1 = \frac{1}{\sum_{i=1}^{\infty}(i-1)!\rho^{i-1}}$$

   By the ratio test, the series in the denominator diverges since $\lim_{i\to\infty}\frac{i!\rho^i}{(i-1)!\rho^{i-1}} = \lim_{i\to\infty} i\rho \to \infty$. The system is always unstable regardless of the values of ρ (or α and β).

4.2 Dimensioning a server

1. The new server has twice the speed of the old server. Therefore the customers service rate is doubled, we have $\mu_{\text{new}} = 2\mu$. The utilization of the new server is $\rho_{\text{new}} = \lambda_{\text{new}}/\mu_{\text{new}} = 2\lambda/(2\mu) = \rho$. The new faster server has the same utilization as the older one.

2. The original server can be modeled as an M/M/1 queueing system with arrival rate λ and service rate μ. By Little's formula we know that the mean response time (or the expected sojourn time) is $\overline{T} = 1/(\mu - \lambda)$. After the merger, the arrival rate has doubled so $\lambda_{\text{new}} = 2\lambda$. The new server serves client with a rate $\mu_{\text{new}} = 2\mu$. The new server can be modeled as an M/M/1 queueing system with arrival rate 2λ and service rate 2μ. The expected sojourn time of the new server is $\overline{T}_{\text{new}} = 1/(\mu_{\text{new}} - \lambda_{\text{new}}) = \overline{T}/2$. The expected sojourn time is halved in the new server.

3. To maintain the same quality, we need to have $\overline{T}_{\text{new}} = \overline{T}$. Let $\mu_{\text{new}} = \alpha\mu$, with $\alpha > 1$ (new server is faster), be the new server's speed. The resulting mean response time is

$$\overline{T}_{\text{new}} = \frac{1}{\mu_{\text{new}} - \lambda_{\text{new}}} = \frac{1}{\alpha\mu - 2\lambda} = \overline{T} = \frac{1}{\mu - \lambda} \quad \Rightarrow \quad \alpha = 1 + \frac{\lambda}{\mu}.$$

   The CPU should be increased by a ratio $\rho := \lambda/\mu$ (original speed is μ, new speed is $\mu + \lambda$). Since $\lambda < \mu$, it is clear that $\alpha < 2$. We observe that the new server must be stable. We must have $2\lambda < \alpha\mu$, i.e. $\alpha > 2\rho$. Therefore $2\rho < \alpha < 2$.

   Our recommendation is to buy a server with a speed $\alpha\mu$ with $\alpha = 1 + \frac{\lambda}{\mu}$. It would be cheaper than a server with speed 2μ ($\alpha < 2$) and will achieve the same response time for customers. We note that the utilization of this suggested new server would be $\rho_{\text{new}} = \lambda_{\text{new}}/\mu_{\text{new}} = 2\lambda/(\alpha\mu) = (2/\alpha)\rho > \rho$. Our suggested faster server will be more utilized than the old one.

UBINET/SI5: Performance Evaluation of Networks

Homework 5

To be returned on 15 October 2024 at 9 pm

Homeworks are a personal effort. Copied solutions will get 0 for a grade.

5.1 $M/G/1$ queue with priorities

Consider an $M/G/1$ queueing system with two customers classes where all service times and all inter-arrival times are independent. Customers in class k, or type k customers, for $k = 1, 2$, arrive according to a Poisson process with rate λ_k and their generic service time σ_k is exponentially distributed with mean $\overline{\sigma}_k = 1/\mu_k$. Denote $\rho_k = \lambda_k/\mu_k$, for $k = 1, 2$. Customers within each class are served according to the first-come-first-served discipline, however type 1 customers have non-preemptive priority over type 2 customers. In other words,

- if there are both a type 1 customer and a type 2 customer in the waiting room, the server serves the type 1 customer even if the type 2 customer entered the queue first;
- if a type 2 customer is being served and a type 1 customer arrives, the service of the type 2 customer will not be interrupted.

We assume the system is stable and at steady-state.

1. Draw over time the residual service time $R(t)$.
2. Compute $\overline{R}$, the time-average residual service time, using the curve $R(t)$.
3. Prove that the expected waiting time of customers in top priority class 1 is
$$\overline{W}_1 = \frac{\overline{R}}{1 - \rho_1} \ .$$
4. A type 2 customer entering the queue needs to wait for the service of all type 1 and type 2 clients in the queue, and all type 1 customers that arrive while waiting, before getting served. Prove that the expected waiting time of customers in low priority class 2 satisfies the relation
$$\overline{W}_2 = \overline{R} + \rho_1 \overline{W}_1 + (\rho_1 + \rho_2)\overline{W}_2 \ .$$
Hint: use the fact that the expected number of arrivals of a Poisson process with rate λ in an interval of duration d is equal to λd.
5. Find an expression for $\overline{W}_2$ as a function of $\overline{R}$, ρ_1 and ρ_2.
6. What is the stability condition of the system?

5.2 M/G/1 with different job types

Consider an M/G/1 queue that serves two types of jobs: red and blue. Red jobs arrive according to a Poisson process with rate $\lambda_R = \frac{1}{4}$ jobs/second, and blue jobs arrive according to a Poisson process with rate $\lambda_B = \frac{1}{2}$ jobs/second. Red job sizes have mean 1 and variance 1, whereas blue job sizes have mean 0.5 and variance 1. All jobs arrive to the same FCFS queue, so that, at any time, the server might be serving a red job or a blue one, and there might be jobs of one or both types in the queue.

1. What is the nature of the arrival process to this queue?
   Write its arrival rate λ.

2. What is the expected service time $\mathrm{E}[\sigma]$ of a job picked at random?

3. Compute the second moment $\mathrm{E}\left[\sigma^2\right]$ of a job picked at random?

4. Compute the load ρ of this M/G/1 queue.

5. What is the expected waiting time of red jobs? Of blue jobs?

6. What is the mean response time of red jobs $\overline{T}_R$?
   What is the mean response time of blue jobs $\overline{T}_B$?

UBINET/SI5: Performance Evaluation of Networks

Correction of homework 5

5.1 $M/G/1$ queue with priorities

1. The curve $R(t)$ is similar to the one in the $M/G/1$ FIFO queue, composed of half-square triangles separated by idle periods. The heights of the triangles come from two different distributions, according to the type of the customer being served.

2. The computation is similar to the one done for the $M/G/1$ FIFO queue without priority and without vacations. We need to account for the number of customers of each class that are served in an interval $(0, C)$, given that the system is empty at instants 0 and C.

$$\begin{aligned}
\overline{R} &= \lim_{C\to\infty} \frac{1}{C} \left(\sum_{i=1}^{Y_1(C)} \frac{\sigma_{1,i}^2}{2} + \sum_{j=1}^{Y_2(C)} \frac{\sigma_{2,j}^2}{2} \right) \\
&= \lim_{C\to\infty} \left(\frac{Y_1(C)}{C} \right) \lim_{C\to\infty} \left(\frac{1}{Y_1(C)} \sum_{i=1}^{Y_1(C)} \frac{\sigma_{1,i}^2}{2} \right) + \lim_{C\to\infty} \left(\frac{Y_2(C)}{C} \right) \lim_{C\to\infty} \left(\frac{1}{Y_2(C)} \sum_{j=1}^{Y_2(C)} \frac{\sigma_{2,j}^2}{2} \right) \\
&= \lambda_1 \frac{E[\sigma_1^2]}{2} + \lambda_2 \frac{E[\sigma_2^2]}{2} \ .
\end{aligned}$$

   As the system is at steady-state and stable, the output rate of each class is equal to its input rate.

3. The computation is similar to the one done for the $M/G/1$ FIFO queue without priority.

$$\overline{W}_1 = \overline{R} + \frac{\overline{X}_1}{\mu_1}$$

   and $\overline{X}_1 = \lambda_1 \overline{W}_1$ by applying Little's formula to type 1 customers in the waiting room. By combining the two equalities we find the result.

4. Let $\overline{X}_2$ denote the expected number of type 2 customers waiting in the queue, then $\overline{X}_1 = \lambda_1 \overline{W}_1$ by applying Little's formula to type 1 customers in the waiting room. Let $\overline{Z}_1$ denote the expected number of type 1 customers that arrive during $\overline{W}_2$. Since type 1 customers arrive according to a Poisson process with rate λ_1, by conditioning on the waiting room W_2 we get $\overline{Z}_1 = \lambda_1 \overline{W}_2$. We have

$$\overline{W}_2 = \overline{R} + \frac{\overline{X}_1}{\mu_1} + \frac{\overline{X}_2}{\mu_2} + \frac{\overline{Z}_1}{\mu_1} = \overline{R} + \rho_1 \overline{W}_1 + (\rho_1 + \rho_2) \overline{W}_2 \ .$$

5. By combining the results found in the previous two questions we derive

$$\overline{W}_2 = \frac{1}{1-\rho_1-\rho_2} \left(\overline{R} + \frac{\rho_1 \overline{R}}{1-\rho_1} \right) = \frac{1}{1-\rho_1-\rho_2} \frac{\overline{R}}{1-\rho_1} \ .$$

6. The stability condition is $\rho_1 + \rho_2 < 1$, otherwise the expression found for $\overline{W}_2$ does not make sense.

5.2 M/G/1 with different job types

1. The arrival process to the M/G/1 queue is the aggregation of two independent Poisson processes, it is a Poisson process whose rate is $\lambda = \lambda_B + \lambda_B = \frac{3}{4}$.

2. We first observe that a job picked at random is red with probability $\frac{\lambda_R}{\lambda} = \frac{1}{3}$ and is blue with probability $\frac{\lambda_B}{\lambda} = \frac{2}{3}$. The expected service time of a job picked at random is

$$\mathrm{E}[\sigma] = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot \frac{1}{2} = \frac{2}{3}.$$

3. The second moment of a variable X is $\mathrm{E}\left[X^2\right] = \mathrm{Var}(X) + (\mathrm{E}[X])^2$. Therefore, the second moment of red job sizes is $1 + 1^2 = 2$ and that of blue job sizes is $1 + 0.5^2 = \frac{5}{4}$. The second moment of a job picked at random is

$$\mathrm{E}\left[\sigma^2\right] = \frac{1}{3} \cdot 2 + \frac{2}{3} \cdot \frac{5}{4} = \frac{3}{2}.$$

4. The load of this M/G/1 queue is

$$\rho = \lambda \mathrm{E}[\sigma] = \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2}.$$

5. All jobs, red or blue, experience the same waiting time. We use the Pollaczek-Khinchin formula to write the expected waiting time

$$\overline{W} = \frac{\lambda \mathrm{E}\left[\sigma^2\right]}{2(1-\rho)} = \frac{\frac{3}{4} \cdot \frac{3}{2}}{2\left(1 - \frac{1}{2}\right)} = \frac{9}{8}.$$

6. To get the mean response time of a particular type of jobs we add its expected service time (job size) to its expected waiting time. We obtain

$$\overline{T}_R = 1 + \frac{9}{8} = \frac{17}{8}, \qquad \overline{T}_B = \frac{1}{2} + \frac{9}{8} = \frac{13}{8}.$$

UBINET/SI5: Performance Evaluation of Networks

Homework 6

To be returned on 5 November 2024 at 9 am

Homeworks are a personal effort. Copied solutions will get 0 for a grade.

6.1 A labyrinth

A labyrinth is divided in 3 sectors having each an entry point. There is only one exit point that is in sector 3. Anyone entering a sector spends a time that is exponentially distributed to find a door. In sector $i = 1, 2$, with some probability, the door opens to the entry point of the sector (the person will have to cross the sector all over again) and with the complementary probability it opens to the next sector, sector $i + 1$. In sector 3, the door opens either to the entry point of the sector or to the exit point. Persons wishing to cross this labyrinth arrive according to a Poisson process with rate λ and choose one of the three entry points at random, with equal probability. Persons do not know that only sector 3 has an exit point. Only one person can be inside a sector at any given time. If a sector is not empty, then one must wait at the entry point until the sector empties.

1. Say why this labyrinth can be modeled as the Jackson network of Figure 1.
   Write explicitly the parameters of the Jackson network:
   – exogenous arrival rates at all nodes,
   – routing matrix $\mathbf{P}$,
   – probabilities to leave the network.

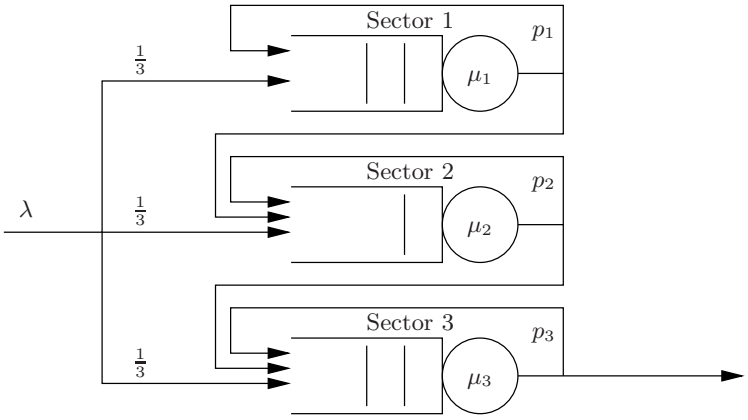


Figure 1: A Jackson network with 3 nodes.

2. Write and solve the traffic equations.
3. What are the irreducibility conditions?
4. Find an upper bound on λ such that the network is stable.
5. What is the stationary probability that only sector 3 is not empty?
6. What time a person needs on average to exit the labyrinth?

6.2 Caching in a Web server

The aim of this problem is to evaluate the impact of a caching mechanism on the performance of a web server. The web server is composed of four components: the processing unit (denoted node A) that processes all requests and manages the cache of the web server, two disks (denoted nodes B and C) and one communication unit (node D) whose role is to send the requested documents to the users. A table storing mappings between documents and their location (disk B or C) is available at node A.

Any document requested by users must be in one and only one of the disks at nodes B or C and possibly in the cache at node A. Let p_A be the probability of this last case. The event of having a document in disk B but not in cache A occurs with probability p_B. p_C is defined similarly, so that $p_A + p_B + p_C = 1$. The web server handles the requests according to the following algorithm:

1. check whether the requested document is in the local cache of the processing unit or not; if yes, send a copy of the document to node D and go to step 4; if not, retrieve its location from a table at node A;

2. access the disk containing the document, according to the information retrieved from the table; make a copy of the requested document and go to node A;

3. update the local cache at the processing unit by inserting in it a copy of the document; send the copy of the document to node D;

4. at the communication unit (node D), transmit the copy of the document to the user that has requested it.

Steps 1 and 3 are handled by node A. The time needed to perform any of these steps is modeled by an exponential random variable with rate μ_A. Step 2 is handled by either node B or node C. The time needed to perform this step is assumed to be exponentially distributed with rate μ_B, when it is node B performing it or rate μ_C, when it is node C performing it. Node D handles step 4 in a time that is exponentially distributed with rate μ_D. Observe that steps 1 and 4 are always performed for any request, whereas steps 2 and 3 are additionally performed only for those requests asking for documents that are only on disks.

We will assume that requests arrive to the web server according to a Poisson process with rate $\lambda > 0$.

1. Explain why this web server can be modeled as a Kelly network with $K = 4$ nodes (A, B, C, D) and $R = 3$ classes $(1, 2, 3)$ with routes r_1, r_2, and r_3 respectively.
   What are the three routes?
   What is the traffic intensity λ_k along route r_k for $k = 1, 2, 3$?

2. Compute $\hat{\lambda}_{i,k}$ the global arrival rate of customers of class $k = 1, 2, 3$ in node $i \in \{A, B, C, D\}$ and $\hat{\lambda}_i$ the global arrival rate in node $i \in \{A, B, C, D\}$.

3. Under which condition is the network stable?

4. Compute the expected response time of the web server (denoted as $\overline{T}$).

5. Numerical applications: assume that $1/\mu_B = 12.5$ ms, $1/\mu_C = 20$ ms, $1/\mu_D = 10$ ms. We will consider two different scenarios:

   - the cache is small: $1/\mu_A = 5$ ms, $p_B = p_C = 0.4$.
   - the cache is large: $1/\mu_A = 25$ ms, $p_B = p_C = 0.1$.

   For each case:
   (i) compute TH, the maximum throughput of the web server (in requests per second),
   (ii) find the bottleneck node, and
   (iii) compute the expected response time when $\lambda = 20$ requests/sec.
   Which configuration is the best?

6. Is it possible to model this web server as a Jackson network? Justify your answer.

UBINET/SI5: Performance Evaluation of Networks

Correction of homework 6

6.1 A labyrinth

1. The splitting of a Poisson process is a Poisson process. Therefore the exogenous arrival process to each node is a Poisson process with rate $\lambda/3$. The "service time" of a sector is the time needed to find a door which is exponentially distributed. There is an infinite waiting room at each sector and there is only one server at each queue since at most one person can be inside a sector. There are routing probabilities between the queues. All the conditions are met to have a Jackson network. The exogenous arrival rates are $\lambda_1^0 = \lambda_2^0 = \lambda_3^0 = \lambda/3$. The routing matrix is

$$\mathbf{P} = \begin{bmatrix} p_1 & 1-p_1 & 0 \\ 0 & p_2 & 1-p_2 \\ 0 & 0 & p_3 \end{bmatrix}$$

   The probabilities to leave the network are $p_{10} = p_{20} = 0$ and $p_{30} = 1 - p_3$. This is an open Jackson network.

2. The traffic equations are

$$\begin{array}{lcl} \lambda_1 = \lambda_1^0 + \lambda_1 p_1 & & \lambda_1 = \dfrac{\lambda_1^0}{1-p_1} = \dfrac{\lambda}{3(1-p_1)} \\ \lambda_2 = \lambda_2^0 + \lambda_2 p_2 + \lambda_1(1-p_1) & \Rightarrow & \lambda_2 = \dfrac{\lambda_1^0 + \lambda_2^0}{1-p_2} = \dfrac{2\lambda}{3(1-p_2)} \\ \lambda_3 = \lambda_3^0 + \lambda_3 p_3 + \lambda_2(1-p_2) & & \lambda_3 = \dfrac{\lambda_1^0 + \lambda_2^0 + \lambda_3^0}{1-p_3} = \dfrac{\lambda}{1-p_3} \ . \end{array}$$

3. For the associated Markov chain to be irreducible, a customer in any queue must have a possibility to leave the network. Therefore, the irreducibility conditions are $p_1 < 1$, $p_2 < 1$, $p_3 < 1$. We can easily see that if any of these probabilities is 1 then the traffic equations do not have a solution.

4. The network is stable if each queue is stable, that is if $\lambda_i < \mu_i$ for $i = 1, 2, 3$. We must have

$$\begin{array}{l} \lambda_1 < \mu_1 \\ \lambda_2 < \mu_2 \\ \lambda_3 < \mu_3 \end{array} \quad \Leftrightarrow \quad \begin{array}{l} \lambda < 3(1-p_1)\mu_1 \\ \lambda < 3(1-p_2)\mu_2/2 \\ \lambda < (1-p_3)\mu_3 \ . \end{array} \quad \Leftrightarrow \quad \lambda < \min\left\{ 3(1-p_1)\mu_1, \frac{3(1-p_2)\mu_2}{2}, (1-p_3)\mu_3 \right\} .$$

   We have found an upper bound on λ.

5. We need to apply Jackson theorem and use the traffic rates in each queue. We have

$$\begin{aligned} \pi(n_1, n_2, n_3) \ = \ & \left(1 - \frac{\lambda}{3(1-p_1)\mu_1}\right)\left(\frac{\lambda}{3(1-p_1)\mu_1}\right)^{n_1} \\ & \times \left(1 - \frac{2\lambda}{3(1-p_2)\mu_2}\right)\left(\frac{2\lambda}{3(1-p_2)\mu_2}\right)^{n_2} \times \left(1 - \frac{\lambda}{(1-p_3)\mu_3}\right)\left(\frac{\lambda}{(1-p_3)\mu_3}\right)^{n_3} . \end{aligned}$$

It is a product-form solution where $(1-\rho_i)\,(\rho_i)^{n_i}$ is the stationary probability to have n_i persons in sector i. Therefore, the probability that only sector three is not empty is

$$\begin{aligned} P(\text{only sector 3 non-empty}) &= P(\text{sector 1 empty})P(\text{sector 2 empty})\big(1-P(\text{sector 3 empty})\big) \\ &= \left(1-\frac{\lambda}{3(1-p_1)\mu_1}\right)\left(1-\frac{2\lambda}{3(1-p_2)\mu_2}\right)\frac{\lambda}{(1-p_3)\mu_3}\ . \end{aligned}$$

6. To apply Little's law, we compute first the expected number of customers in the network. The expected number of customers in queue i is

$$\overline{N}_i = \frac{\lambda_i}{\mu_i - \lambda_i}\ .$$

The expected number of customers in the system (the labyrinth) is

$$\overline{N} = \sum_{i=1}^{3}\overline{N}_i = \frac{\lambda}{3(1-p_1)\mu_1-\lambda} + \frac{2\lambda}{3(1-p_2)\mu_2-2\lambda} + \frac{\lambda}{(1-p_3)\mu_3-\lambda}\ .$$

The expected sojourn time is:

$$\overline{T} = \frac{\overline{N}}{\lambda} = \frac{1}{3(1-p_1)\mu_1-\lambda} + \frac{2}{3(1-p_2)\mu_2-2\lambda} + \frac{1}{(1-p_3)\mu_3-\lambda}\ .$$

6.2 Caching in a Web server

1. We can distinguish three classes of requests according to where the requested document is retrieved. Class 1 traffic follows route $r_1 = (A, D)$ and corresponds to all requests that hit on the cache at node A. Class 2 traffic follows route $r_2 = (A, B, A, D)$ and corresponds to all requests for documents found only in disk B. Class 3 traffic follows route $r_3 = (A, C, A, D)$ and corresponds to all requests for documents found only in disk C. Since requests arrive to the web server according to a Poisson process with rate λ, and given that a request follows route r_1 with probability p_A, route r_2 with probability p_B and route r_3 with probability p_C, we have then an independent thinning of a Poisson process, and consequently
• the requests arrival process of class 1 traffic is Poisson with rate $\lambda_1 = \lambda p_A$,
• the requests arrival process of class 2 traffic is Poisson with rate $\lambda_2 = \lambda p_B$, and
• the requests arrival process of class 3 traffic is Poisson with rate $\lambda_1 = \lambda p_C$,
and they are all independent. The processing time at any of nodes A, B, C or D is exponentially distributed and independent of the arrival processes. The queues associated to nodes A, B, C and D are infinite. All the conditions are met to model the web server as a Kelly network.

2. The global arrival rates are

$$\begin{array}{llll} \hat{\lambda}_{A,1} = \lambda p_A, & \hat{\lambda}_{A,2} = 2\lambda p_B, & \hat{\lambda}_{A,3} = 2\lambda p_C, & \Rightarrow \quad \hat{\lambda}_A = \lambda(1+p_B+p_C) \\ \hat{\lambda}_{B,1} = 0, & \hat{\lambda}_{B,2} = \lambda p_B, & \hat{\lambda}_{B,3} = 0, & \Rightarrow \quad \hat{\lambda}_B = \lambda p_B \\ \hat{\lambda}_{C,1} = 0, & \hat{\lambda}_{C,2} = 0, & \hat{\lambda}_{C,3} = \lambda p_C, & \Rightarrow \quad \hat{\lambda}_C = \lambda p_C \\ \hat{\lambda}_{D,1} = \lambda p_A, & \hat{\lambda}_{D,2} = \lambda p_B, & \hat{\lambda}_{D,3} = \lambda p_C, & \Rightarrow \quad \hat{\lambda}_D = \lambda(p_A+p_B+p_C) = \lambda. \end{array}$$

3. The stability conditions of the network are $\hat{\lambda}_i < \mu_i$ for $i = A, B, C, D$, namely,

$$\left.\begin{array}{l} \lambda(1 + p_B + p_C) < \mu_A \\ \lambda p_B < \mu_B \\ \lambda p_C < \mu_C \\ \lambda < \mu_D \end{array}\right\} \quad \Rightarrow \quad \lambda < \min\left\{\frac{\mu_A}{1 + p_B + p_C}, \frac{\mu_B}{p_B}, \frac{\mu_C}{p_C}, \mu_D\right\}.$$

4. From a result seen in class, the expected response time in the network is

$$\overline{T} \;=\; \frac{1}{\lambda} \sum_{i \in \{A,B,C,D\}} \frac{\hat{\lambda}_i}{\mu_i - \hat{\lambda}_i} \;=\; \frac{1 + p_B + p_C}{\mu_A - \lambda(1 + p_B + p_C)} + \frac{p_B}{\mu_B - \lambda p_B} + \frac{p_C}{\mu_C - \lambda p_C} + \frac{1}{\mu_D - \lambda}.$$

5. In steady-state, the throughput of the system is the same as the arrival rate (expressed in requests per second). The *maximum* throughput TH is the maximum value of λ according to the stability condition. The bottleneck node is the one yielding the tighter constraint in the stability condition.

   • The cache is small: $\mu_A = 200$ requests per second, $\mu_B = 80$ requests per second, $\mu_C = 50$ requests per second, and $\mu_D = 100$ requests per second. The stability condition is

   $$\lambda < \min\left\{\frac{200}{1 + 0.4 + 0.4}, \frac{80}{0.4}, \frac{50}{0.4}, 100\right\} = \min\left\{111.\overline{1}, 200, 125, 100\right\} \;\Rightarrow TH = 100 \text{ req. per second.}$$

   Node D is the bottleneck node.
   When $\lambda = 20$, the system is stable and its expected response time is

   $$\overline{T} = \frac{1.8}{200 - 20 \cdot 1.8} + \frac{0.4}{80 - 20 \cdot 0.4} + \frac{0.4}{50 - 20 \cdot 0.4} + \frac{1}{100 - 20} = \frac{9}{820} + \frac{1}{180} + \frac{1}{105} + \frac{1}{80} = \frac{7967}{206640} \approx 0.03855$$

   • The cache is large: $\mu_A = 40$ requests per second. The stability condition is

   $$\lambda < \min\left\{\frac{40}{1 + 0.1 + 0.1}, \frac{80}{0.1}, \frac{50}{0.1}, 100\right\} = \min\left\{33.\overline{3}, 800, 500, 100\right\} \;\Rightarrow TH = 33.\overline{3} \text{ req. per second.}$$

   Node A is the bottleneck node.
   When $\lambda = 20$, the system is stable and its expected response time is

   $$\overline{T} = \frac{1.2}{40 - 20 \cdot 1.2} + \frac{0.1}{80 - 20 \cdot 0.1} + \frac{0.1}{50 - 20 \cdot 0.1} + \frac{1}{100 - 20} = \frac{3}{40} + \frac{1}{780} + \frac{1}{480} + \frac{1}{80} = \frac{189}{2080} \approx 0.09087 \text{ s.}$$

   The best configuration is the first one since it yields a smaller expected response time.

6. The network cannot be modeled by a single class Jackson network (the one seen in class) since, in the latter, routing at any node is identical for all customers leaving a node: any customer leaving node i would be routed to node j (resp. outside the network) with the probability p_{ij} (resp. p_{i0}).