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

$$\begin{aligned} \lambda_1 &= \lambda_1^0 + \lambda_1 p_1 \\ \lambda_2 &= \lambda_2^0 + \lambda_2 p_2 + \lambda_1(1-p_1) \\ \lambda_3 &= \lambda_3^0 + \lambda_3 p_3 + \lambda_2(1-p_2) \end{aligned} \quad \Rightarrow \quad \begin{aligned} \lambda_1 &= \frac{\lambda_1^0}{1-p_1} = \frac{\lambda}{3(1-p_1)} \\ \lambda_2 &= \frac{\lambda_1^0 + \lambda_2^0}{1-p_2} = \frac{2\lambda}{3(1-p_2)} \\ \lambda_3 &= \frac{\lambda_1^0 + \lambda_2^0 + \lambda_3^0}{1-p_3} = \frac{\lambda}{1-p_3} \,. \end{aligned}$$

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

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

   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

$$\begin{array}{llll} \hat\lambda_{A,1} = \lambda p_A, & \hat\lambda_{A,2} = 2\lambda p_B, & \hat\lambda_{A,3} = 2\lambda p_C, \;\Rightarrow & \hat\lambda_A = \lambda(1 + p_B + p_C) \\ \hat\lambda_{B,1} = 0, & \hat\lambda_{B,2} = \lambda p_B, & \hat\lambda_{B,3} = 0, \;\Rightarrow & \hat\lambda_B = \lambda p_B \\ \hat\lambda_{C,1} = 0, & \hat\lambda_{C,2} = 0, & \hat\lambda_{C,3} = \lambda p_C, \;\Rightarrow & \hat\lambda_C = \lambda p_C \\ \hat\lambda_{D,1} = \lambda p_A, & \hat\lambda_{D,2} = \lambda p_B, & \hat\lambda_{D,3} = \lambda p_C, \;\Rightarrow & \hat\lambda_D = \lambda(p_A + p_B + p_C) = \lambda. \end{array}$$

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}).