Performance Engineering Tutorial
Revision

Exercise 1. Consider the following probability transition matrix

$$P = [p_{ij}] = \begin{array}{c} \\ E \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ X \end{array} \begin{array}{c} \begin{array}{cccccc} E & S_1 & S_2 & S_3 & S_4 & X \end{array} \\ \begin{bmatrix} 0 & 1.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0.5 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \end{array}$$

describing user interactions with an IT system hosting services S_1, S_2, S_3, S_4, and where E and X respectively denote entry and exit states.

Question 1.1 Determine the mean session length.
Solution:
The visits are obtained from the system of linear equations

$$\begin{aligned} V_E &= 1 \\ V_1 &= V_E + 0.5V_3 \\ V_2 &= 0.4V_1 \\ V_3 &= 0.2V_1 + 0.5V_4 \\ V_4 &= 0.4V_1 \\ V_X &= 1 \end{aligned}$$

which has solutions $V_E = V_X = 1$, $V_1 = \frac{5}{4}$, $V_2 = V_3 = V_4 = \frac{1}{2}$. Thus the mean session length is $L = V_1 + V_2 + V_3 + V_4 = \frac{11}{4} = 2.75$ requests.

Question 1.2 Give a formula to compute the probability of ending the session after exactly three requests.
Solution:

$$\begin{aligned} \pi^0 &= [\pi^0_E, \pi^0_1, \pi^0_2, \pi^0_3, \pi^0_4, \pi^0_X] = [1, 0, 0, 0, 0, 0] \\ \pi^2 &= [\pi^2_E, \pi^2_1, \pi^2_2, \pi^2_3, \pi^2_4, \pi^2_X] = \pi^0 P^2 \\ \pi^3 &= [\pi^3_E, \pi^3_1, \pi^3_2, \pi^3_3, \pi^3_4, \pi^3_X] = \pi^0 P^3 \end{aligned}$$

The requested value is $\pi^3_X - \pi^2_X$.

Question 1.3 Assume that the application has $n = 10$ users, each starting a new session at rate $\lambda = 0.11$ sessions/min. The front server is hosted on two $M = 2$ virtual machines (VMs) whereas the database server is hosted on a single machine. The requests require the following service times:

Time $[min]$	FS VMs	DB
S_1	0.25	0.1
S_2	0.10	0.15
S_3	0.33	0.15
S_4	0.20	0.15

For a load balancer using a round-robin policy, what would be the expected CPU utilization at the front servers and at the database?
Solution:
The arrival rate is $\lambda' = n\lambda = 1.1$ sessions/min. The arrival rate of requests for each service type is thus

$$\begin{aligned}\lambda_1 =& V_1\lambda' = \frac{5}{4}\frac{11}{10} = \frac{55}{40} = 1.375\\ \lambda_2 =& V_4\lambda' = \frac{1}{2}\frac{11}{10} = \frac{11}{20} = 0.550\\ \lambda_2 =& \lambda_3 = \lambda_4\end{aligned}$$

Due to round-robin, a VM hosting a front server instance receives requests at a rate reduced by a factor $1/M$, since only one request is M is forwarded to that particular VM.
Let $T_{i,j}$ be the processing time of service i at VM j, then by the utilization law

$$U_1 = \sum_{i=1}^{4} \lambda_i T_{i,1} = 1.375{\cdot}0.25/M{+}0.55{\cdot}0.10/M{+}0.55{\cdot}0.33/M{+}0.55{\cdot}0.20/M = 0.692/M = 0.346 \ (34.6\%)$$

$$U_2 = \sum_{i=1}^{4} \lambda_i T_{i,2} = 1.375 \cdot 0.10 + 0.55 \cdot 0.15 + 0.55 \cdot 0.15 + 0.55 \cdot 0.15 = 0.385 \ (38.5\%)$$

We may observe that the database is the bottleneck for utilization.

Exercise 2. Suppose we investigate the throughput X of a database using a 2^k factorial design without replication and with $k = 2$ factors. The first factor is *cache size* (C) with levels 512MB and 1GB. The second factor is *threading level* (T) with levels 256 and 512. The following throughput measurements are obtained:

X $[ms]$	512MB	1GB
256	4	5
512	8	6

Question 2.1 Give the sign table for the design and quantify the effects q_0, q_C, q_T, q_{CT}.
Solution:

I	C	T	CT
+1	-1	-1	+1
+1	+1	-1	-1
+1	-1	+1	-1
+1	+1	+1	+1

Therefore,

$$\begin{aligned}q_0 =& (4 + 5 + 8 + 6)/4 = 5.75,\\ q_C =& (-4 + 5 - 8 + 6)/4 = -.25,\\ q_T =& (-4 - 5 + 8 + 6)/4 = 1.25,\\ q_{CT} =& (4 - 5 - 8 + 6)/4 = -.75,\end{aligned}$$

Question 2.2 Quantify the percentages of variation explained by the factors and by their interaction. Discuss your findings.

Solution:

$$SST = 4(q_C^2 + q_T^2 + q_{CT}^2) = 8.75$$
$$SS - C = 4(q_C^2) = 0.25$$
$$SS - T = 4(q_T^2) = 6.25$$
$$SS - TC = 4(q_{CT}^2) = -2.25$$

Therefore, the threading level explains $6.25/8.75 = 71\%$ of the total variation and the rest is mainly due to the interaction between threading levels and cache.

Question 2.3 Assume a third factor H (hyper-threading) is also included in the experiments, with levels ON and OFF. Give the sign table for a 2^{3-1} fractional factorial design. Indicate all the confoundings.

Solution:
This is obtained from the previous design by the substitution $H = CT$.

I	C	T	H
+1	-1	-1	+1
+1	+1	-1	-1
+1	-1	+1	-1
+1	+1	+1	+1

The confoundings are generated by the generating relation $H = CT$. We have $H^2 = I = CTH$, $HT = CT^2 = C, CH = C^2T = T$.

Exercise 3. A server I-O is described in terms of the transfer function

$$H(z) = \frac{Y(z)}{U(z)} = \frac{4}{z}$$

where $Y(z) = \mathcal{Z}[y_t]$ and $U(z) = \mathcal{Z}[u_t]$ are the z-transforms of the input and output signals.

We wish to describe the system response in the time domain as a function h_t that produces the output according to a convolution of the inputs, i.e.,

$$y_t = \sum_{k=-\infty}^{+\infty} h_{t-k} u_k \tag{1}$$

You are asked to determine the discrete time series h_t, assuming that $h_t = 0$ for $t < 0$. ***Hint***: note that $Y(z) = H(z)U(z)$.
Solution:
By the convolution property of the z-transform, h_t is the time series such that $H(z) = \mathcal{Z}[h_t]$, since

$$y_t = \sum_{k=-\infty}^{+\infty} h_{t-k} u_k \iff Y(z) = H(z)U(z) \tag{2}$$

By the definition of z transform, we also know that

$$H(z) = h_0 + h_1 z^{-1} + h_2 z^{-2} + \dots$$

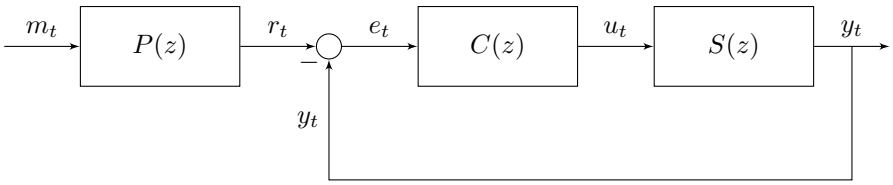
By matching the last expression to the particular expression $H(z) = 4z^{-1}$ given in this exercise, we see that h_n is the time series

$$h_0 = 0, h_1 = 4, h_2 = 0, h_3 = 0, \dots$$

Exercise 4. A software probe monitors the memory usage m_t of a software system ($m_t = 1$ implies 1 Gb). Based on the current value, the probe determines a reference queue-length threshold r_t that is passed to an admission controller that controls the parallelism level u_t in the server. The server and the admission controller have the following transfer functions

$$P(z) = \frac{z}{z^2 + \theta} \qquad S(z) = \frac{1}{z+1} \qquad C(z) = \frac{z}{z-1}$$

and a block diagram



where $e_t = r_t - y_t$ is an error signal.

Question 4.1 Determine if the control system is stable for some choices of θ.

Solution:
The system features a mix of open-loop and closed-loop controllers. We can first determine the transfer function of the closed-loop subsystem as

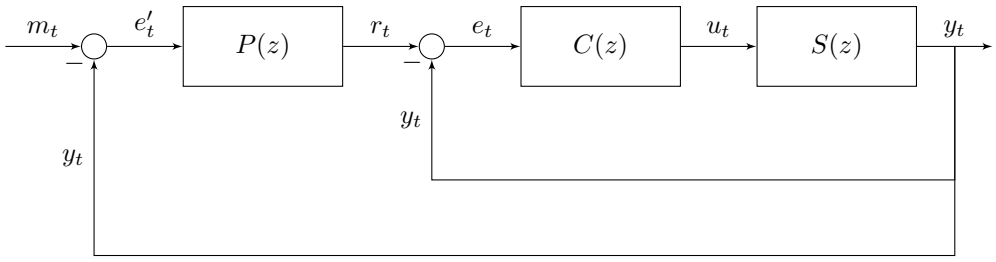
$$H_c(z) = \frac{S(z)C(Z)}{1 + S(z)C(Z)} = \frac{z}{z^2 + z - 1}$$

For the system as a whole we then have the transfer function

$$H(z) = P(z)H_c(z) = P(z)\frac{S(z)C(Z)}{1 + S(z)C(Z)} = \frac{z^2}{(z^2 + z - 1)(z^2 + \theta)}$$

Since the denominator is a product of two second-order polynomials, we can solve each of them and find the poles $\lambda_1 = -\sqrt{-\theta}$, $\lambda_2 = +\sqrt{-\theta}$, $\lambda_3 = -\sqrt{5}/2 - 1/2 = -1.6180$, $\lambda_4 = \sqrt{5}/2 - 1/2 = 0.6180$. It is clear that no choice of θ ensures stability, where we need to require that $\lambda = \max_k |\lambda_k| \leq 1$, since the system pole at -1.6180 cannot be changed if we alter θ.

Question 4.2 Assume now that $\theta = 0$ and that the memory consumption of a running job is exactly 1 Gb, so that y_t is also the instantaneous memory usage in gigabytes. How would the previous answer change if we were to modify the system topology as follows?



Solution:
In general, the transfer function is now given by

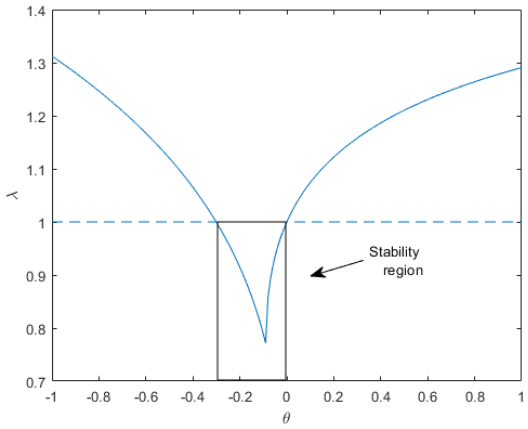
$$H(z) = \frac{P(z)H_c(z)}{1 + P(z)H_c(z)} = \frac{z^2}{z^4 + z^3 + \theta z^2 + \theta z - \theta}$$

Thus, for $\theta = 0$ this reduces to

$$H(z) = \frac{P(z)H_c(z)}{1 + P(z)H_c(z)} = \frac{1}{z^2 + z}$$

that has poles $\lambda_1 = 0$ and $\lambda_2 = -1$. Therefore, the control system is now stable.

It should be noted that since $\lambda = \max\{|\lambda_1|, |\lambda_2|\} = 1$, the settling time of this control is infinite. The following plot shows how alternative parameterizations of θ produce in some cases controllers with faster settling times ($\lambda < 1$):



For example, for $\theta = -0.08$, we get $\lambda = 0.8577$ and the settling time is $T \approx 26$ time steps.