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

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

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.,

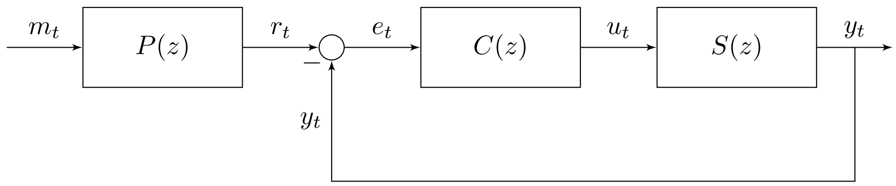
$$y_t = \sum_{k=-\infty}^{+\infty} h_{t-k} u_k \tag{1}$$

You are asked to determine the discrete time series h_t, assuming that $h_t = 0$ for $t < 0$. *Hint*: note that $Y(z) = H(z)U(z)$.

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 probe, 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.
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?