1. Consider a web server with processing capacity C = 10^5 op/sec. The server receives requests with a mean rate 2 req/sec.
The requests have different demand *Z*. Consider the following intervals:

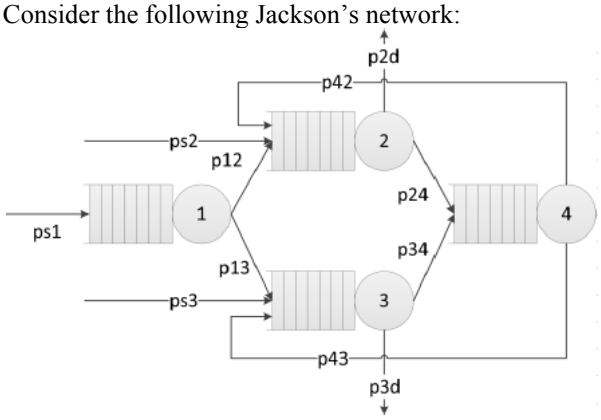
   - $Z < 20.000$ op
   - $20.000 \text{ op} \leq Z < 40.000$ op
   - $Z \geq 40.000$ op

   By assuming that:
   i. the mean size is 40.000 op, characterized by an exponential distribution;
   j. the arrival rate is characterized by a Poisson process;
   Define a management mechanism of the server to satisfy the following QoS requirements:
   req1. Mean response time ≤ 1.5 s for all requests
   req2. Mean waiting time ≤ 0.5 s, for $Z < 40.000$ op.ni.

   Evaluate
   a. The mean *throughput* for the server with the chosen management mechanism;
   b. The mean *conditional slowdown* for jobs with size x=0.1 s, 0.3 s
   c. Compare the mean slowdown obtained in b. with the corresponding mean slowdown for FIFO and PS scheduling.
   Please comment all the obtained results.

2. Assume the web server has a minimum memory capacity so that just two requests can be kept in the system, according to a PS discipline.
   Evaluate:
   a. The loss probability
   b. The mean number of requests in the system and its variance
   c. The probability the server is idle
   d. The system throughput.

3. Consider the following Jackson's network:



   the parameters values are:
   $p_{s1} = \frac{3}{5}$, $p_{s2} = \frac{3}{10}$, $p_{s3} = \frac{1}{10}$, $p_{42} = \frac{1}{5}$, $p_{43} = \frac{4}{5}$, $p_{2d} = p_{3d} = \frac{2}{3}$
   γ=10 j/s, μ_1=18 j/s, μ_2=18 j/s, μ_3=30 j/s, μ_4=15 j/s.

   1. By assuming p_{12}=2/3 and p_{13}=1/3, compute the system response time
   2. Compute the values of p_{12} and p_{13} such that the servers 2 and 3 have the same utilization
   3. By using the computed values of p_{12} and p_{13} compute the system response time.

---

1. Consider a web server with processing capacity C=10^3 op/sec. The server receives requests according to the following flow characteristics:

   80% with mean demand Z=1875 op/job
   20% with mean demand Z=7500 op/job

The arrival flow is uniformly partitioned in two classes and served according to abstract priority.
By assuming that the arrival process is exponential and the service process is Hyperexponential[1], determine:

a. Which is the maximum arrival flow admittable to guarantee a SLO (Service Level Object) of less than 10 sec for the mean global response time

b. By assuming an arrival rate of 0.2 req/sec, the abstract scheduling to guarantee a SLO of about 2 sec for the mean waiting time of the highest priority class.

c. Evaluate the mean waiting and response time for each priority class. Explain the obtained results by considering the Hyperexponential distribution characteristics.

d. By assuming an exponential service process, prove that the mean global response time is independent of the classes partition (two classes).

2. Consider a closed queueing network with the following characteristics:

- service demand $D_1 = 10$
- service demand $D_2 = 5$
- think time $Z = 10$
- number of users $N = 3$

Which is the response time of the system? Justify the applied relations.

1. Consider a system that offers two kinds of services: *service a* and *service b*. The arrival flow of both service requests is Poisson-like. The service requests are first elaborated by a dedicated server (server A and B for service a and b respectively) and then they are processed by a second common service stage (server C). All the service requests are exponential distributed. The server characteristics are the following:

server A:
$C_{Ai}=10^4$ op/msec single server capacity
$Z_A=8x10^3$ op/job mean demand
$\lambda_A=1.5$ job/msec mean arrival rate
$m_A=3$ parallel servers[2]
FIFO scheduling

server B:
$C_B=10^5$ op/msec single server capacity
$Z_B=4x10^4$ op/job mean demand
$\lambda_B=2$ job/msec mean arrival rate
SB-non-preemptive scheduling, with priority for jobs of size $S \leq E(S)$

server C:
$C_{Ci}=10^3$ op/msec single server capacity
$Z_C=2x10^3$ op/job mean demand
$m_C=10$ parallel servers
FIFO scheduling

Determine:

a. The queueing network model
b. The system response time for service request of type a

[1] For a Hyperexponential distribution, $\sigma^2(X) = g(p)E(X)^2$ where $g(p) = \frac{1}{2p(1-p)} - 1$ and p is the phase probability

[2] Note that for a multi-server center the probability to have all the servers busy is $P_Q = \frac{(m\rho)^m}{m!(1-\rho)} p(0)$, with $p(0) = \left[\sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)} \right]^{-1}$

c. The system response time for any service request
d. The mean conditional slowdown - for a *service a* request of size x= 0.4 msec at server A; - for a *service b* request of size x= 0.2 msec at server B; - for any request of size x= 0.2 msec at server C.

2. Consider a single-core server hosting a web service. Requests arrive to the server according to a Poisson, with an average inter-arrival time of 200 ms. [3]
   a. Knowing that the maximum buffer size is N = 4 (including the jobs in service) and that each request requires on average 200 ms of processing time, compute the throughput of the system.
   b. Consider a CPU upgrade to a faster single-core processor which can process a request in 150 ms. Compute the throughput of the upgraded system.
   c. Consider a CPU upgrade to a slower quad-core processor, which can process a request in 300 ms using one of its processor cores. Compute the throughput of the upgraded system.

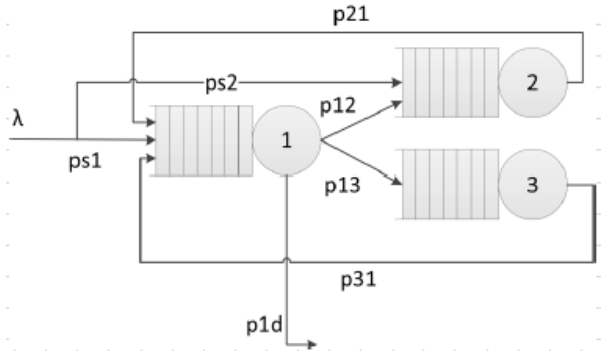
---

1. Consider a telephone company call center that receives requests for different services. The requests are sent from two types of users: *subscribers* and *new*. Moreover, maximum priority should be given to some *critical* requests. From historical data of the last six months, the following can be stated: 15% of critical requests; 60% of standard requests; 25% of information requests from potential new users. Consider the following requirements:
   req1. The new users abandon if the waiting time is over 3 min;
   req2. The critical requests must be completed within 5 min from their arrival
   req3. The subscribers change telephone service provider if the response time is over 90 min.

   By assuming a single server model and the following processes characteristics:
   - Poisson arrivals with a mean rate of 0.23 req/min;
   - Hyperexponential[4] services of mean 4 min and p=0.35.

   Determine:
   a. The mean service time for each stage of the Hyperexpo with relative percentage
   b. the system management mechanisms to satisfy the following:
      mechanism a – satisfies req1 and req2
      mechanism b – satisfies req2 and req3.
   c. the mean conditional slowdown for requests of size 1 min. Compare with the PS case.
   d. For one of the chosen mechanisms and in the general case, prove the relation between priority classes i and i+1 for the waiting time and for the response time.

2. Consider the following Jackson's network:



the parameters values are:

$p_{s1} = \frac{2}{5}$, $p_{1d} = \frac{1}{2}$, $p_{12} = \frac{1}{10}$,

[3] The geometric series have the following property: $\sum_{i=0}^{n} q^i = \frac{1-q^{n+1}}{1-q}$

[4] For a Hyperexponential distribution, $\sigma^2(X) = g(p)E(X)^2$ where $g(p) = \frac{1}{2p(1-p)} - 1$ and p is the phase probability

λ=3 j/s, μ_1=10 j/s, μ_2=15 j/s, μ_3=5 j/s.
Determine:

a. The expected number of visits at each station
b. The expected response time at each station
c. The system response time
d. The choice of p_{12} and p_{13} (p_{1d} is fixed) that minimizes the system response time.

---

1. A service provider apply two different rates for two kinds of users: users paying the highest fee obtain preemptive priority.

By considering the mean response time as the user satisfaction measure, evaluate the following statements:

1. The highest priority class experiments the minimum mean response time.
2. The lowest priority class experiments the maximum mean response time.
3. Globally, the mean response time is improved by the preemption.

Prove **if** the statements above are correct and for which arrival and service time distributions.
Let us assume the following system characteristics:

- Single processor with capacity 10^5 op./sec
- For both classes, exponential mean service demand $5x10^4$ op./job
- System utilization 75%.

Determine:

a. the mean waiting time and the mean response time for the highest priority class if this includes the 30% of the arrival requests.

The service provider wants to investigate the performance with a size-based scheduling, in particular with a SRPT[5]. Determine

b. The mean waiting time E(T_Q(x)) for job size x=1.
c. Which percentage of jobs would experiment a waiting time ≤ E(T_Q(1)).
d. Compare the results with the above abstract case.

2. Consider the following measurement data for an interactive system:
   measurement interval: 10 min
   number of users: 50
   number of servers: 10
   average response time per transaction:
   20 sec
   Dmax: 1 sec/transaction
   Dtot: 2 sec/transaction
   Number of completed transactions:
   90
   On average, how many users are thinking?

---

1. Consider a web server with the following system characteristics:

- Single processor with capacity 10^5 op./sec
- Exponential mean service demand $4x10^4$ op./job
- System utilization 60%.

By knowing the job size, the service provider adopts a simple Size Based - priority scheduling without preemption: jobs with size less (or equal) than the average will have the highest priority (class 1); jobs with size greater than the average have the lowest priority (class 2). Determine:

a. the mean response time for both classes and the global mean response time.

The service provider wants to investigate if a dual core server would improve the service performance.

---

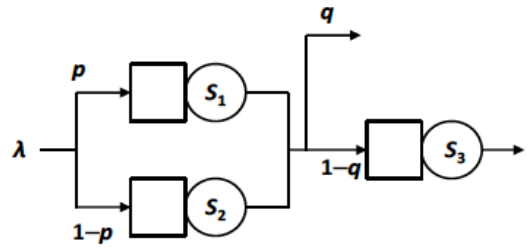
[5] $E(T_Q(x))=\dfrac{\frac{\lambda}{2}\int_0^x t^2 dF(t)+\frac{\lambda}{2}x^2(1-F(x))}{\left(1-\lambda\int_0^x t dF(t)\right)^2}$

b. Conjecture the behaviour of the performance measures for both classes, by writing the mean waiting and response time definition for the dual core case.

2. Consider the following measurement data for an interactive system:
   measurement interval: 5 min
   number of users: 50
   number of servers: 10
   average response time per transaction: 20 sec
   Dmax: 1 sec/transaction
   Dtot: 2 sec/transaction
   Number of completed transactions: 75

On average, how many users are thinking?

3. Consider the following queueing network model:



$S_1 = S_2 = S_3 = S$

Determine the following characteristics and performance measures:
- The visits at the three stations
- The response time of the three stations
- The response time of the system
- The maximum throughput of the system.