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).

a. $$\frac{\frac{\lambda}{2} E(S^2)}{1-\rho} + E(S) \le 10 s$$

$$\sigma^2(S) = E(S^2) - E(S)^2$$

$$E(S^2) = \sigma^2(S) + E(S)^2$$

$$\rho = 0,8$$

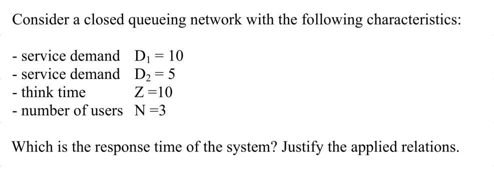
$$28.125 \cdot \lambda \le 14 - 42 \lambda$$

$$\lambda \le 14 / 70.125 = 0,1996432$$

b. Prova con PREEMPTIVE

c. // //

d. // //



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.

Unico modo per risolvere è MVA

$$E(t_i(N)) = E(S_i)(1+E(n_i(N-1))$$

$$R_i(N) = D_i(N)(1+Q_i(N-1))$$

$N=0$

$Q_1(0) = Q_2(0) = 0$

$R_1(1) = D_1(1)(1+Q_1(0)) =$

$R_2(1) = D_2(1)(1+Q_2(0)) =$

$$X(1) = \frac{1}{Z+R_1(1)+R_2(1)}$$ da LITTLE

$Q_1(1) = X(1)\, R_1(1)$

$Q_2(1) = X(1)\, R_2(1)$

$N=2$

$R_1(2) = D_1(2)\left(1+Q_1(1)\right) =$

$R_2(2) = D_2(2)\left(1+Q_2(1)\right) =$

$$X(2) = \frac{2}{Z + R_1(2) + R_2(2)} =$$

$Q_1(2) = X(2)\,R_1(2)$

$Q_2(2) = X(2)\,R_2(2)$

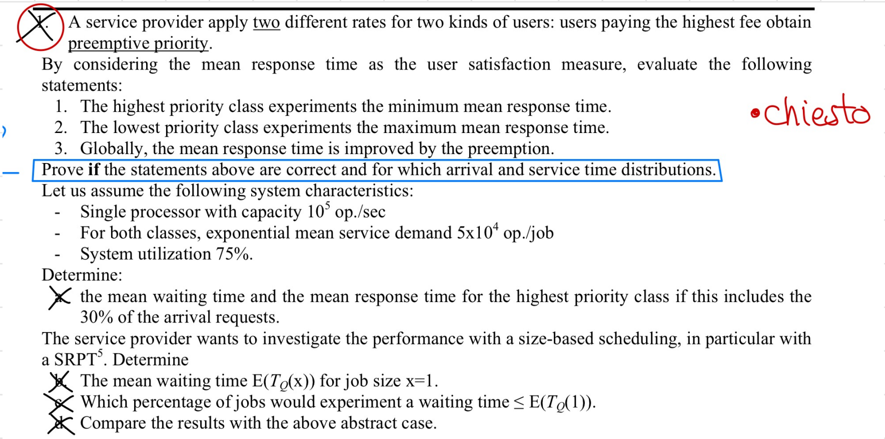
$N=3$

$R_1(3) = D_1(3)\left(1+Q_1(2)\right) =$

$R_2(3) = D_2(3)\left(1+Q_2(2)\right) =$

$$R_{TOT} = R_1(3) + R_2(3) = \frac{79}{3} = 26{,}33\ s$$

Può essere utile usare le frazioni e approx alla fine



~~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.

•chiesto

1. Vero sempre, ma a patto che gli arrivi siano random (con una distribuzione di Poisson)

2. Vero sempre, ma a patto che gli arrivi siano random (con una distribuzione di Poisson)

3. Nel caso di una distribuzione ESPONENZIALE il tempo di RISPOSTA del caso preemptive non cambia, per qualsiasi distribuzione invece il caso PREEMPTIVE migliora le prestazioni.

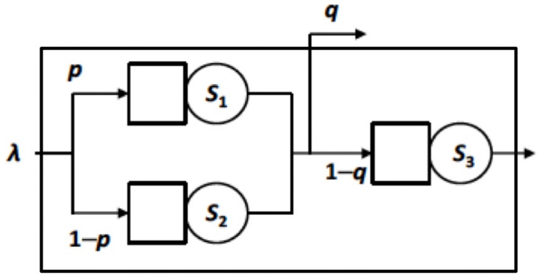
d. Confrontiamo il TEMPO DI ATTESA di tutti quelli con size x =1 con scheduling:

1 - SRJF

2 - ASTRATTO CON PRELAZIONE

Dato che le 2 non so in quale coda può finire un JOB di size 1 (minore di 1) allora devo confrontare le PRESTAZIONI GLOB (spoiler: SJRF va meglio)

3. Consider the following queueing network model:



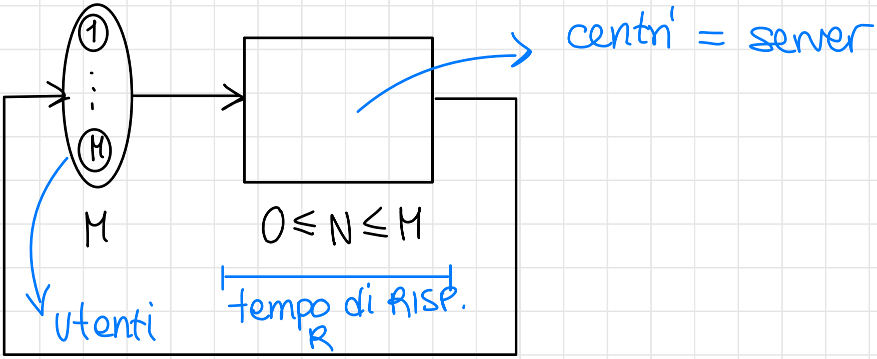
$S_1 = S_2 = S_3 = S$

Determine the following characteristics and performance measures:

~~a~~ - The visits at the three stations
b - The response time of the three stations
c - The response time of the system
d - The maximum throughput of the system.

Assumere ESPONENZIALE

~~2.~~ Consider the following measurement data for an interactive system:
measurement interval: 10 min
number of users: 50 la differenza?
number of servers: 10 → 10 centri
so server paralleli
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? • chiesto


1
M
M
utenti
$0 \leq N \leq M$
tempo di RISP.
R
centri = server