04/03/22

Performance Modeling of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

Analytical models

(single resource)

Università degli studi di Roma Tor Vergata
Department of Civil Engineering and Computer Science Engineering

1

Performance evaluation techniques

Computational and mathematical techniques to *model*, *simulate* and *analyze* the performance of *stochastic systems*

Modeling: conceptual framework describing a system

Simulate: perform experiments using computer implementation of the model

Analyze: draw conclusions from output

Simulation models

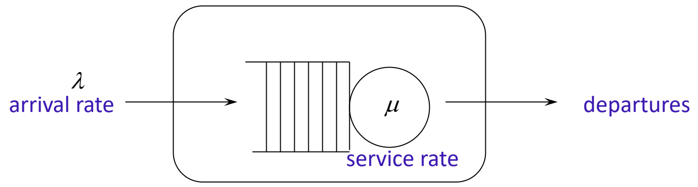
Analytical models

2

p 22 discrete

conceptual model

Single server center



Terminology

size ≥ service time S (es: Job A chiede 5s) $S=1/\mu$

time in the queue T_Q waiting time

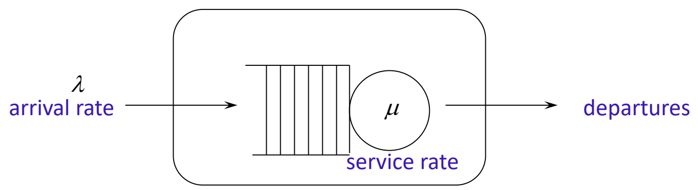
time in the system T_S residence/response time (da quando arrivo a quando esco)

number in the system N_S

number in the queue N_Q

number in the service U, ρ

Single server center

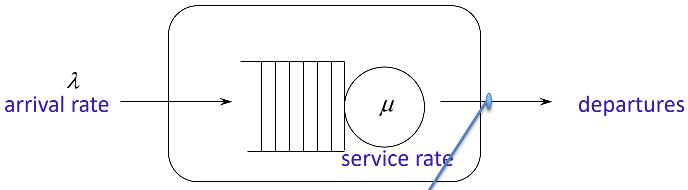


E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

↳ utile nei requisiti di qualità

1. As λ, the mean arrival rate, increases, all the performance metrics mentioned above increase. (cresce tutto)
2. As μ, the mean service rate, increases, all the performance metrics mentioned above decrease.

Single server center



E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

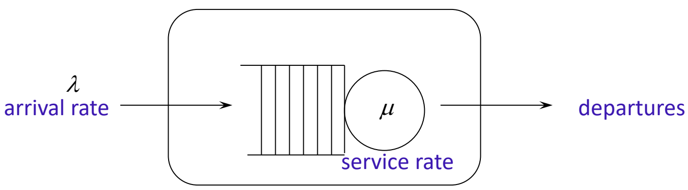
Def. throughput (produttività)

t=1, E(n)$_1$ n° of completions (departures) in the time unit

Non tempo totale !

5

Single server center



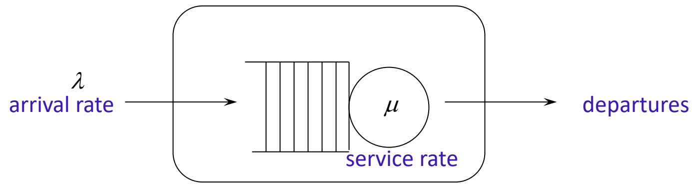
Def. utilization

How can we “mathematically” define the utilization?

$$\rho = \lambda / \mu$$

6

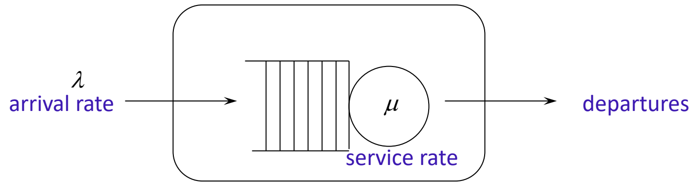
Single server center



$E(T_Q)$, $E(T_S)$, $E(N_S)$, $E(N_Q)$, Prob{ $T_S > t$ }, $E(n)_t$

$$\underbrace{E(T_S)}_{T_{TOT}} = \underbrace{E(T_Q)}_{\text{in coda}} + \underbrace{E(S)}_{\text{servizio}}$$

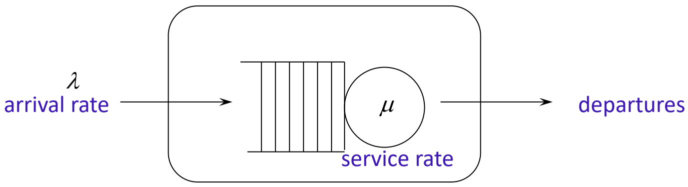
Single server center



$E(T_Q)$, $E(T_S)$, $E(N_S)$, $E(N_Q)$, Prob{ $T_S > t$ }, $E(n)_t$

$$\underbrace{E(N_S)}_{\text{Pop. TOTALE servente singolo}} = \underbrace{E(N_Q)}_{\text{in coda}} + E\big(\underbrace{\text{number in service}}_{\rho \div \text{media in corso di servizio}}\big)$$

Single server center

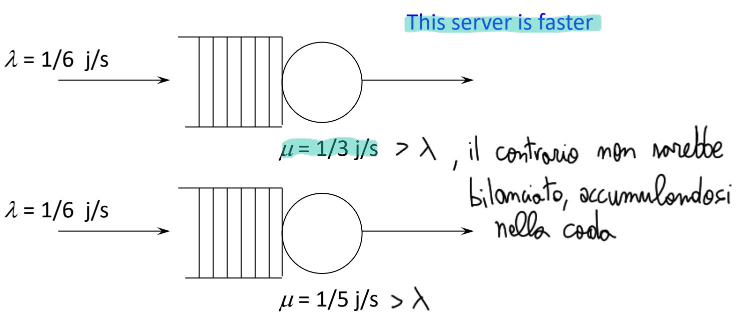


E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

$$E(N_S) = E(N_Q) + \rho$$

Analytical models

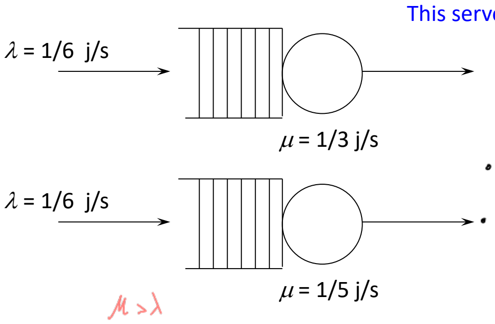
Single server center



Which system has greater throughput?

Single server center

This server is faster



- throughput $X = \frac{\text{Job completati } C}{\text{tempo } \tau}$
- $\rho = \frac{\text{tempo Busy } B}{\text{tempo } \tau \text{ osservazione}}$

$$\frac{C}{\tau} = \underbrace{\frac{C}{B}}_{\frac{\text{completati}}{\text{tempo occupato}} \left[\frac{job}{s}\right] = \mu} \cdot \frac{B}{\tau}$$

$\mu > \lambda$

By assuming *job flow balance*, the throughput is the same !!

For both systems $X = \lambda$ = 1/6 j/s

$\rightarrow X = \mu \cdot \rho = \mu \cdot \frac{\lambda}{\mu} = \lambda$

BUT the faster server shows the shorter queue and so shorter mean response time

In other words, improving the mean response time does not necessarily improve the throughput

Single server center

If the center is in $\underbrace{\text{stochastic}}_{\text{random}}$ equilibrium (stationary condition),

$$\lambda < \mu, \quad \rho = \lambda / \mu \; < 1$$

E$(n)_1 = X = \lambda$

$$\text{throughput} = \text{MIN}(\lambda, \mu)$$

Throughput is independent of the service rate μ

If the center is NOT in stochastic equilibrium, (NO bilanciamento flusso)

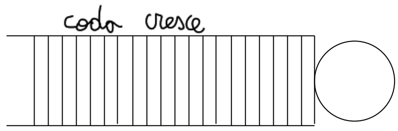
$\lambda > \mu$, (nella coda c'è sempre qualcuno)

E$(n)_1 = X = \mu$

the center cannot work off the arrival rate, the queue grows unlimited

Single server center

What's up if $\lambda > \mu$?

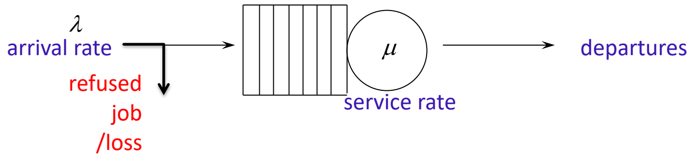


the center cannot work off the arrival rate, the queue grows unlimited

media in coda
nel tempo T

$$\mathrm{E}(N_Q \text{ in } T) \geq \underbrace{\lambda T}_{\text{entrati in } T} - \underbrace{\mu T}_{\text{usciti in } T} = T\overbrace{(\lambda - \mu)}^{>0} \rightarrow \infty \text{ as } T \rightarrow \infty$$

Single server center with finite buffer (coda finita)

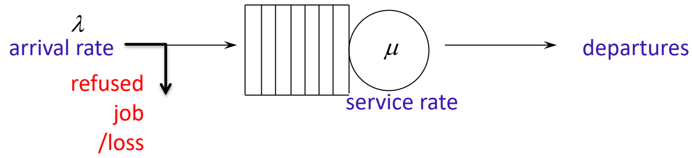


Each arrival when the queue is *full* will be lost

Which is the throughput?

Single server center with finite buffer

coda finita SEMPRE STAZIONARIA, la coda non è infinita



Each arrival when the queue is *full* will be lost

Which is the throughput? (riceve MENO di λ)

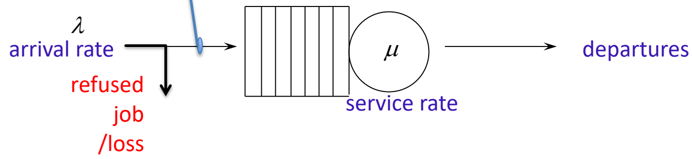
$X \neq \lambda$ $\quad \cancel{\rho = \lambda / \mu}$

No!
On the contrary

$X < \lambda$

Single server center with finite buffer

$\lambda' = X$ (quelli effettivi)



Each arrival when the queue is *full* will be lost

Which is the throughput?

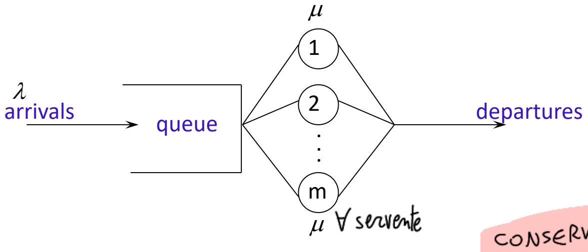
$X \neq \lambda$ $\quad \cancel{\rho = \lambda / \mu}$

No!
On the contrary

$X < \lambda$

P.6 discrete

Multi Server Queue (ho m serventi in // identici)



CONSERVATIVO

$E(S_i)$ $E(S)$

$N_S = \begin{cases} \{0, \dots, m\} & \text{se } N_Q = \emptyset \\ N_Q + m & \text{se } N_Q > \emptyset \end{cases}$ (coda vuota)

(quanti ce ne sono?)

"Come è la media?"

Nel servente singolo $N_S = N_Q + \{0,1\}$

↳ se servo o meno qualcuno

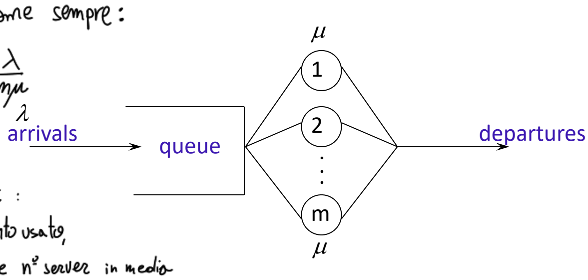
media: $E(N_S) = E(N_Q) + \rho$

definisco

- $\rho_i = \left(\frac{\lambda}{m}\right) \cdot \frac{1}{\mu}$
- $\rho_{GLOBALE} \doteq$ def. come sempre:

tasso medio IN → λ / tasso MAX OUT → $m\mu$ $= \frac{\lambda}{m\mu}$

Multi Server Queue



anche se uguali:

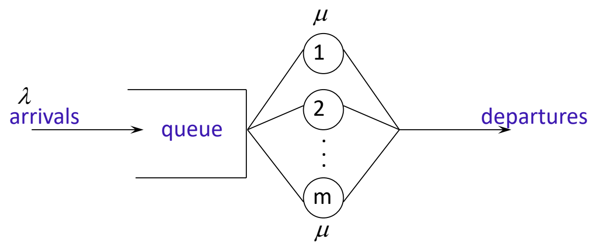
- ρ_i dice server "i" quanto usato,
- con >1 server mi dice n° server in media su m totali

$$E(S_i) = 1/\mu \qquad E(S) = 1/m\mu = E(S_i)/m$$

se $m = 10$, $\rho = 0{,}5$ allora:
- server singolo: usato al 50%
- server multipli: di 10 server, mediamente occupati $0{,}5 \cdot 10 = 5$

NB: se popolazione $N_S \to \infty$, sempre 'm' saranno occupati, e calcolo ρ

Multi Server Queue

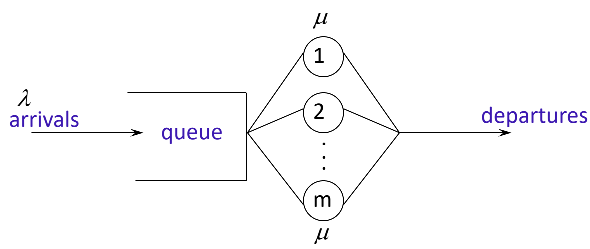


$$E(N_S) = \begin{cases} E(N_Q) + \rho & if\ m = 1 \text{ (singolo)} \\ E(N_Q) + m\rho & f\ m > 1 \text{ (multiplo)} \end{cases}$$

But how is the utilization defined for the multiserver case?

19

Multi Server Queue

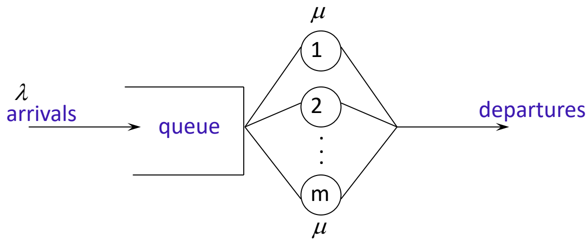


$$\rho_i = \frac{\lambda_i}{\mu} = \frac{\lambda}{m\mu} \qquad \rho_{glob} = \frac{\lambda}{\mu_{glob}} = \frac{\lambda}{m\mu}$$

$\mu_1 = \mu_2 = \dots$

20

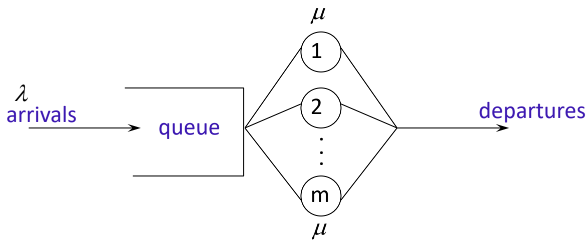
Multi Server Queue



$$\rho = \begin{cases} \frac{\lambda}{\mu} = \lambda E(S_i) & \text{if } m = 1 \\ \frac{\lambda}{m\mu} = \frac{\lambda E(S_i)}{m} & \text{if } m > 1 \end{cases}$$

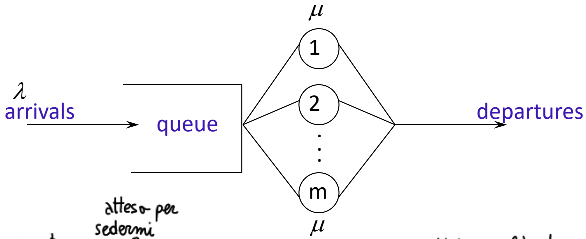
$\underbrace{\frac{1}{\mu}}_{[\frac{1}{s}]} = E(S_i) \quad [s]$

Multi Server Queue



$$\rho_i = \rho_{glob} = \frac{\lambda}{m\mu}$$

Multi Server Queue



tempo risposta:

atteso per sedermi

$$E(T_S) = E(T_Q) + E(S_i)$$

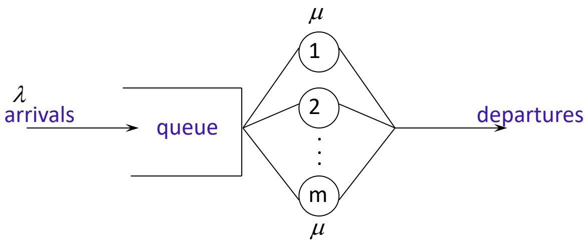
$E(S_i) = \frac{1}{\mu}, \forall i$: qualsiasi servente (tutti uguali), da quando uno si siede, fino ad uscire, calcolo media, è $\frac{1}{\mu}$

quando arriva a quando va via

- introduco $E(S) = \frac{1}{m\mu}$, da quando uno entra e uno qualsiasi (diverso da chi è entrato) esce, "dopo quanto tempo se ne libera uno di quegli m, in media"
- $E(T_q)$ = tempo attesa medio (arrivo - siedo sul server)

23

Multi Server Queue

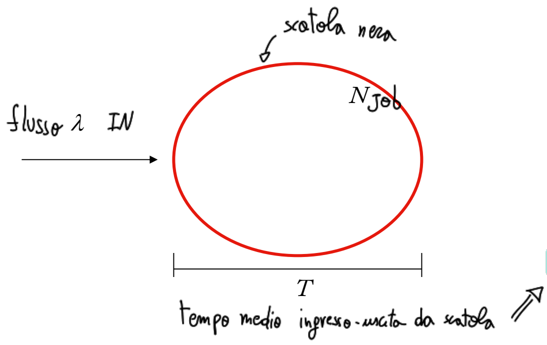


$$E(T_S) = E(T_Q) + E(S_i)$$

$$E(T_Q) = f(\lambda, \rho, E(S))$$

24

Little's law is very important for its broad applicability.
In general, we can see Little's law as applied at a black box:
it states relations between mean values



Little's Law (1961)

(a) queue discipline is FIFO,

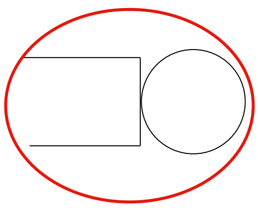
(b) service node capacity is infinite,

(c) flow balance

$N = \lambda T$ (sono medie)

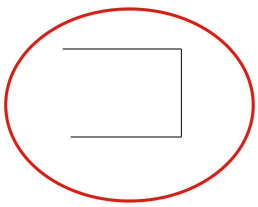
If λ is the mean arrival rate, T is the mean residence time in the black box, N is the mean population in the black box, the theorem asserts that, if the system is "stable", the mean population is given by the "mean arrival flow" multiplied the mean time the jobs spend in the black box

25



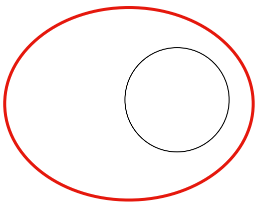
If the black box is the whole center, the theorem is applied to the center mean population:

$$E(N_S) = \lambda E(T_S)$$



If the black box is just the queue, the theorem is applied to the queue mean population:

$$E(N_Q) = \lambda E(T_Q)$$



If the black box is just the server, the theorem is applied to the server "mean population", in other words to the utilization:

$\rho = \lambda E(S)$ (legge utilizzazione)

tempo che ci sto dentro

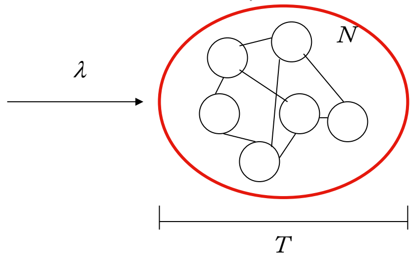
Prof. Vittoria de Nitto Personè 26

26

But if the black box is a network of centers, anyway interconnected,

↳ sia singoli che multi!

(Non mi interessa dentro come sia fatto)



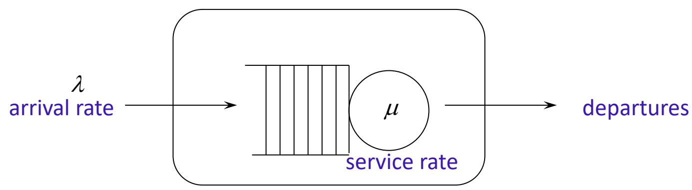
$N = \lambda T$

The theorem is applied to the entire network!!

27

Analytical models

Single server center



$$E(T_S) = E(T_Q) + E(S)$$

$$E(N_S) = E(N_Q) + \rho$$

Little's law aggiunge

$$E(N_S) = \lambda E(T_S)$$

$$E(N_Q) = \lambda E(T_Q)$$

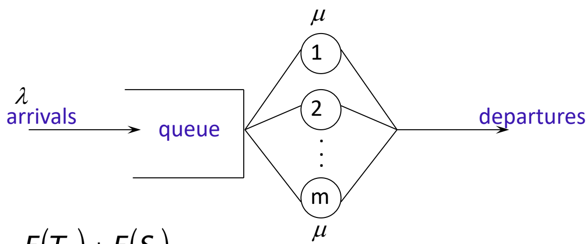
$$\Longleftrightarrow$$

$$E(T_S) = \frac{E(N_S)}{\lambda}$$

$$E(T_Q) = \frac{E(N_Q)}{\lambda}$$

28

Multi Server Queue



$$E(T_S) = E(T_Q) + E(S_i)$$ ↳ attesa del "singolo"

Little's law

$$E(N_S) = \lambda E(T_S)$$

$$E(N_Q) = \lambda E(T_Q)$$

tempi ↗ Pop. media

$$E(T_S) = \frac{E(N_S)}{\lambda}$$

$$E(T_Q) = \frac{E(N_Q)}{\lambda}$$

#1)

Consider a web server with a mean processing rate of 1.2 job/s.
If the server receives requests with a rate of 0.45 job/s and it has 0.225 enqueued jobs on average, determine: (serv. singolo)
(in media, nella coda)

a) the average utilization
b) the average response time.

During rush hours (picco) the arrival rate grows of 20% and the average number of enqueued jobs becomes 0.3681818.
Determine:
c) the performance metrics a) and b) (calcolo di nuovo stessi indici)
d) which further increasing in arrival rate makes the server collapsing (coda → ∞)
e) the performance metrics a) and b) for the limiting case d).

#1)

SVOLGIMENTO

- "mean processing rate" $\doteq$ service rate $\mu = 1.2$ J/s
- "server receivs requests" $\doteq$ arrival rate $\lambda = 0.45$ J/s
- "enqueued Job average" $\doteq$ popolazione media in coda $E[N_q] = 0.225$

Richieste

a) average utilization $\rho = \frac{\lambda}{\mu} = 0.375$

b) average response time (tempo TOT nel sistema) $E[T_s] = \frac{E[N_s]}{\lambda}$

dove $E[N_s] = \underbrace{E[N_q]}_{\text{persone in coda}} + \underbrace{\rho}_{\text{persone in server}} = 0,225 + 0,375 = 0,6$, allora $E[T_s] = \frac{0,6}{0,45} = 1.\bar{3}$

λ incrementa del 20% $\rightarrow \lambda' = \lambda \cdot 1,2 = 0,54$ J/s ; $\mu = 1.2$ J/s e $E[N_q'] = 0,3681818$

c) ricalcola le metriche

c.a) average utilization $\rho' = \frac{\lambda'}{\mu} = 0.45$

c.b) average response time (tempo TOT nel sistema) $E[T_s'] = \frac{E[N_s']}{\lambda'}$

dove $E[N_s'] = \underbrace{E[N_q']}_{\text{persone in coda}} + \underbrace{\rho'}_{\text{persone in server}} = 0,3681818 + 0,45 = 0,8181818$, allora $E[T_s'] = \frac{0,81818}{0.54} = 1.\overline{51}$

Analogamente $E[T_q] = \frac{E[N_q]}{\lambda'} = 0,6818188\ s$ e $E[T_s] = E[T_q] + \underbrace{E[S]}_{= \frac{1}{\mu}} = 0,681818\ s + \frac{1}{1.2} = 1,\overline{51}$

d) aumento in % tale che server collassa e coda $\rightarrow \infty$

Vorrei $\rho \rightarrow 1$, cioè $\lambda' \rightarrow \mu$, quindi $\lambda' \cdot \underbrace{x}_{\text{aumento \%}} \overset{?}{=} \mu$, trovo $x = \frac{1.2}{0.54} = 2,\bar{2}$ (aumento 120%)

e) in tale condizione, $\rho = \frac{\lambda}{\mu} = 1$ e $E[T_s] = \frac{E[N_s]}{\lambda'} = \infty$ (arrivi coda > partenze server)

#2)

Let us consider a server that processes jobs with rate 0.8 jobs/s.
By assuming that the server receives jobs with a rate depending on the time slot as follows:

8.00 a.m. – 12.00 a.m. average arrival rate 1.5 jobs/s
12.00 a.m. – 2.00 p.m. average arrival rate 0.5 jobs/s
2.00 p.m. – 7.00 p.m. average arrival rate 1.5 jobs/s
7.00 p.m. – 9.00 p.m. average arrival rate 0.5 jobs/s
9.00 p.m. – 8.00 a.m. average arrival rate 0.05 jobs/s

Determine:

a) average arrival rate per day (24 hours)
b) average utilization per day (" ")
c) average throughput per day
d) average throughput for each time slot ($\forall$ fascia oraria)

Please, justify and comment the results by indicating the used laws.

Soluzione in lect5Dex1introAM.pdf