Performance Modeling
of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

Abstract Priority

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

1

QoS management

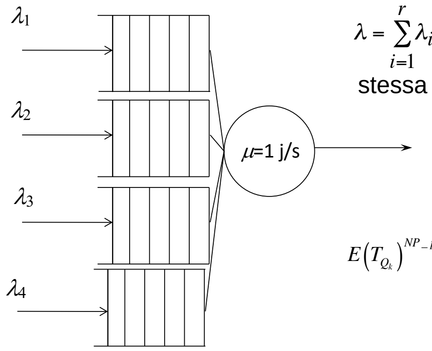
- Service provider
- Traffic flows with different QoS
- QoS: mean response time

2

Abbiamo tempo medio uguali per tutti. (Abstract)

M/M/1 – NP_priority

Prob di essere di classe "k" = 1/4



$$\lambda = \sum_{i=1}^{r} \lambda_i \quad \text{uniform partition:} \quad \lambda_i = \frac{\lambda}{4}, \; p_i = \frac{1}{4}$$

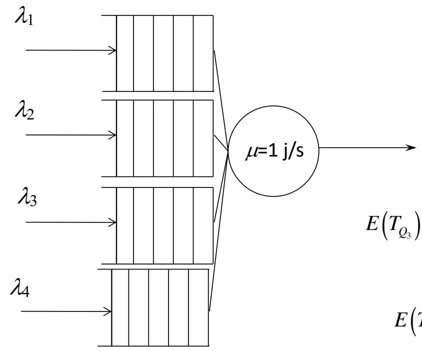
stessa occupazione per ogni classe

$$\rho = 0.1, 0.2, 0.4, 0.6$$

$$E(T_{Q_k})^{NP\_priority} = \frac{\frac{\lambda}{2} E(S^2)}{\left(1 - \sum_{i=1}^{k} \rho_i\right)\left(1 - \sum_{i=1}^{k-1} \rho_i\right)}$$

3

NP priority



$$E(T_{Q_1}) = \frac{\rho E(S)}{(1-\rho_1)}$$

prima classe va sempre meglio, ovvio.

$$E(T_{Q_2}) = \frac{\rho E(S)}{(1-(\rho_1+\rho_2))(1-\rho_1)}$$

$$E(T_{Q_3}) = \frac{\rho E(S)}{(1-(\rho_1+\rho_2+\rho_3))(1-(\rho_1+\rho_2))}$$

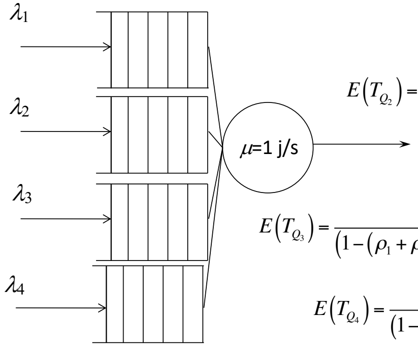
$$E(T_{Q_4}) = \frac{\rho E(S)}{(1-\rho)(1-(\rho_1+\rho_2+\rho_3))}$$

l'ultima va sicuramente peggio.

4 Il numeratore vede sempre l'occupazione di tutte le classi.

NP priority

$$\rho = 0.1$$



$$E\left(T_{Q_1}\right) = \frac{\rho E(S)}{(1-\rho_1)} \approx 0.1025641s$$

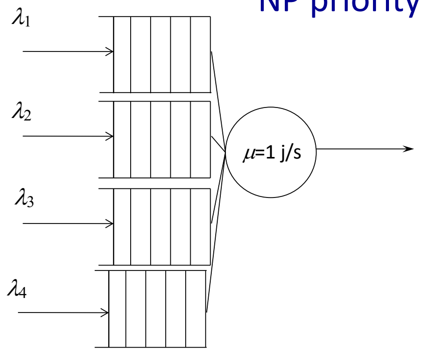
$$E\left(T_{Q_2}\right) = \frac{\rho E(S)}{\left(1-(\rho_1+\rho_2)\right)(1-\rho_1)} \approx 0.1079622s$$

$$E\left(T_{Q_3}\right) = \frac{\rho E(S)}{\left(1-(\rho_1+\rho_2+\rho_3)\right)\left(1-(\rho_1+\rho_2)\right)} \approx 0.113798s$$

$$E\left(T_{Q_4}\right) = \frac{\rho E(S)}{(1-\rho)\left(1-(\rho_1+\rho_2+\rho_3)\right)} \approx 0.1201201s$$

5

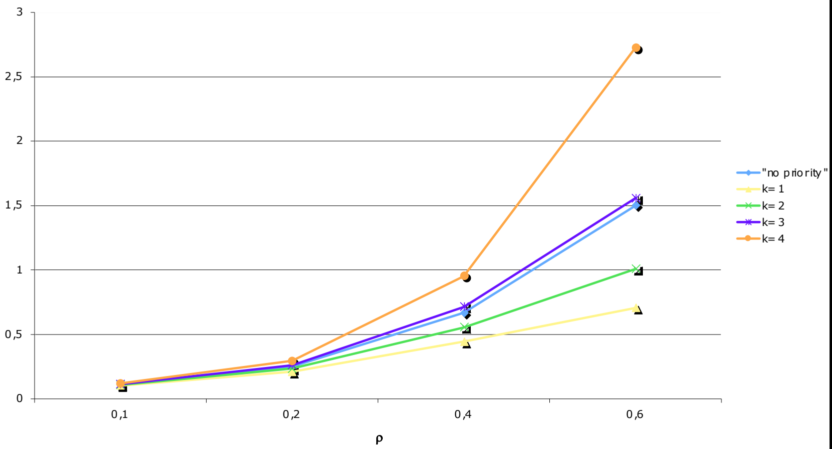
NP priority



ρ	$E(T_{Q1})$	$E(T_{Q2})$	$E(T_{Q3})$	$E(T_{Q4})$	$E(T_Q)$ (noprior)
0,1	0,1025641	0,1079622	0,113798	0,1201201	0,11111111
0,2	0,2105263	0,2339181	0,2614379	0,2941176	0,25
0,4	0,44444444	0,55555555	0,7142857	0,9523809	0,66666666
0,6	0,7058823	1,0084033	1,5584415	2,7272727	1,5

6

$E(T_{Q_i})$



7 "No priority" divide in due gruppi i vari grafici di "k".

Goals:

- given a QoS requirement, decide if adopt priority classes
- note that if the policy is non-size-based, we can reason just in terms of waiting time

Mean service demand: 0.4 s

QoS requirement
the waiting time should not exceed the service demand, in particular:
the service provider will not incur in penalties if $T_Q \leq 0.45$ s;
the service provider will gain revenue if $T_Q < 0.4$ s

By simple "costless" analysis we can offer good insights

8

Analytical models
priority scheduling

$E(S) = 0.4$ s exponential

	Low load	medium load	high load
$\rho =$	0.4	0.6	0.8
$\lambda =$	1	1.5	2 job/s
$E(T_Q) =$	0.26	0.6	1.6 job/s

without priority classes

2 priority class

medium load

ρ_1=0.3	ρ_2=0.3	ρ_1=0.18	ρ_2=0.42	ρ_1=0.42	ρ_2=0.18
$E(T_{Q1})$ 50%	**$E(T_{Q2})$ 50%**	**$E(T_{Q1})$ 30%**	**$E(T_{Q2})$ 70%**	**$E(T_{Q1})$ 70%**	**$E(T_{Q2})$ 30%**
0.342857	0.85714	0.2926829	0.731707317	0.413793	1.03448

high load

ρ_1=0.4	ρ_2=0.4	ρ_1=0.24	ρ_2=0.56	ρ_1=0.56	ρ_2=0.24
$E(T_{Q1})$ 50%	**$E(T_{Q2})$ 50%**	**$E(T_{Q1})$ 30%**	**$E(T_{Q2})$ 70%**	**$E(T_{Q1})$ 70%**	**$E(T_{Q2})$ 30%**
0.5333333	2.6666666	0.4210526	2.105263	0.727272	3.636363

prof. Vittoria de Nitto Personè 9

9

In "medium load", con caso 30% e 70%, le prestazioni vanno meglio per entrambi. Basta alternare le percentuali, e le prestazioni vanno entrambe peggio. Perchè 30% e 70% meglio?
Alla classe 1 do un privilegio maggiore, ma a "pochi". La seconda classe è superata dal 30% di quelli di classe 1, quindi superata da "pochi".
Nel caso 70% e 30% va peggio perchè do la priorità a troppi job.

high load ρ=0.8

not penalties if $T_Q \le 0.45$ s	gain revenue if $T_Q < 0.4$ s
$p_1 = 0.36$, $p_2 = 0.64$	$p_1 = 0.22$, $p_2 = 0.78$
$E(T_{Q1}) = 0.4494$ s	$E(T_{Q1}) = 0.3883$ s
$E(T_{Q2}) = 2.2472$ s	$E(T_{Q2}) = 1.9417$ s

$$E(T_Q)_{glob} = E(T_Q)_{KP} = 1.6 \text{ s}$$

prof. Vittoria de Nitto Personè 10

10

Come ho trovato le percentuali?
Quando cerco la divisione delle classi, devo imporre per la prima classe il soddisfacimento del QoS. Risolta l'equazione sotto, troviamo che p1<= 0.36, cioè 36%

$$\frac{\rho \cdot E[S]}{1 - \rho \cdot p_1} < Q_{0.5} = 0.45$$

Osservazione: Nel secondo caso, sia E[Tq1] che E[Tq2] sono migliori del primo caso. Tuttavia devo sempre rapportarli alle loro percentuali: Nel secondo caso, la classe 1 ha un miglioramento rispetto alla classe 1 del primo caso, ma questo miglioramento non è di TUTTA la classe 1, perchè nel secondo caso il miglioramento tocca il 22% dei job, nel primo il 36%.

Euristica per la ripartizione in classi di priorità astratta

coda singola

$\rho = 0.92,\ E(S) = 1$ j/s $\quad E(T_Q) = 11.5$ s , $E(T_S) = 12.5$ s

60%, 25 %, 15 %,

$E(T_{Q1}) = 2.05357$ s
$E(T_{Q2}) = 9.42005$ s
$E(T_{Q3}) = 52.75229$ s

$E(T_{S1}) = 3.05357$ s
$E(T_{S2}) = 10.42005$ s
$E(T_{S3}) = 53.75229$ s

15 %, 25 %, 60%,

$E(T_{Q1}) = 1.067285$ s
$E(T_{Q2}) = 1.688743$ s
$E(T_{Q3}) = 18.196203$ s

$E(T_{S1}) = 2.067285$ s
$E(T_{S2}) = 2.688743$ s
$E(T_{S3}) = 19.196203$ s

A sinistra, miglioro classe 1 e classe 2 (miglioro nell'85% = 60% + 25%) rispetto coda unica. Il problema è che Q3 cresce molto rispetto a coda singola.

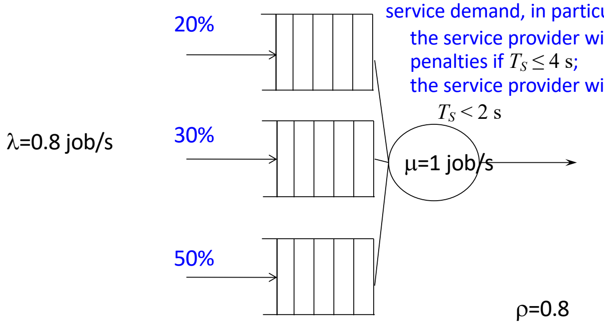
11

Analytical models
abstract priority

exp

arrivi random/ exp (quite variable) / 1 servente

M/M/1 Arrival flow: random, rate 0.8 job/s
Service process: quite variable rate 1 job/s
QoS requirements
the response time should not exceed twice the service demand, in particular:
the service provider will not incur in penalties if $T_S \leq 4$ s;
the service provider will gain revenue if $T_S < 2$ s



ρ=0.8

con coda singola $E(T_Q) = \frac{\rho E(S)}{1-\rho} = \frac{0.8}{0.2} = 4 \quad E(T_S)=5$ non rispetto i requisiti.

12

M/M/1

Not penalties if $T_S \leq 4$ s;
gain revenue if $T_S < 2$ s

	NP	P	NP	P
class	$E(T_Q)$		$E(T_S)$	
1 - 20%	0.9523809523809524	0.19047619047619052	$\boxed{1.9523809523809526}$	1.1904761904761905
2 - 30%	1.5873015873015874	0.7936507936507937	$\boxed{2.5873015873015874}$	1.9841269841269842
3 - 50%	6.666666666666669	6.666666666666669	7.666666666666669	8.333333333333336
global	4.000000000000001	3.6095238095238105	5.000000000000001	5.000000000000001

La classe 3, ha tempo E[Tq] NP e P uguali, essendo ultima classe non fa prelazione.
La classe 1 e classe 2, facendo prelazione, hanno guadagno rispetto alla variante senza prelazione.
Globalmente ho guadagno rispetto a KP, in virtù della prelazione. Con KP ho tempo globale 4.
Passiamo ai tempi di risposta: perdo il guadagno perchè devo considerare il tempo di servizio che aumenta per i job che vengono buttati fuori (colpa della MEMORYLESS).
Vedendo E[Ts] NP, in due casi riesco a stare sotto 4s, con prelazione ho anche guadagno.

M/M/1

Not penalties if $T_S \leq 4$ s;
gain revenue if $T_S < 2$ s

	NP	P	NP	P
class	$E(T_Q)$		$E(T_S)$	
1 - 20%	0.9523809523809524	0.19047619047619052	$\boxed{1.9523809523809526}$	1.1904761904761905
2 - 30%	1.5873015873015874	0.7936507936507937	2.5873015873015874	1.9841269841269842
3 - 50%	6.666666666666669	6.666666666666669	7.666666666666669	8.333333333333336
global	4.000000000000001	3.6095238095238105	5.000000000000001	5.000000000000001

Analytical models
abstract priority

M/M/1

Not penalties if $T_S \leq 4$ s;
gain revenue if $T_S < 2$ s

	NP	P	NP	P
class	$E(T_Q)$		$E(T_S)$	
1 - 20%	0.9523809523809524	0.19047619047619052	1.9523809523809526	1.1904761904761905
2 - 30%	1.5873015873015874	0.7936507936507937	2.5873015873015874	1.9841269841269842
3 - 50%	6.666666666666669	6.666666666666669	7.666666666666669	8.333333333333336
global	4.000000000000001	3.6095238095238105	5.000000000000001	5.000000000000001

prof. Vittoria de Nitto Personè 15

15

Aggiungiamo una variabilità maggiore

Analytical models
bstract priority

No priority classes
g(0.1)=4.5556,
$E(T_Q)$=11.1111 s, $E(T_S)$=12.1111 s

Not penalties if $T_S \leq 4$ s;
gain revenue if $T_S < 2$ s

Hypergeometrica $M/H_2/1$

	NP	P	NP	P
class	$E(T_Q)$		$E(T_S)$	
1 - 20%	2.645502645502645	0.5291005291005291	3.645502645502645	1.529100529100529
2 - 30%	4.409171075837742	2.204585537918871	5.409171075837742	3.3950617283950617
3 - 50%	18.51851851851852	18.51851851851852	19.51851851851852	20.185185185185187
global	11.11111111111111	10.026455026455027	12.11111111111111 >	11.416931216931218

piccolo guadagno
con la prelazione, nella memoryless perdevo tale vantaggio.

p=0.1 → 10% E(S)=5 s

90% E(S)=0.55555 s

prof. Vittoria de Nitto Personè 16

16

E[S] circa 1 anche nell'hypergeometrica, coincide con quella di partenza. Stessa media globale del servizio, sennò non ha senso con un carico diverso.
Nel grafico, nel caso E[Ts] Preemptive riesco a non avere penalità (tempi sotto i 4 secondi), ma non riesco ad avere ricompense, poichè solo la prima classe scende sotto i due secondi.

Mean service demand (expo): 0.4 s

QoS requirement

the waiting time (average) should not exceed 0.1 s, in particular:

the service provider will gain c_1 for each service within QoS

the service provider will pay c_2 for each service violates QoS

quanti soddisfano*guadagno - quanti non soddisfano*perdita

$R = p_1 c_1 - p_2 c_2$

Abstract-P → max R

17

esempio

$E(S) = 0.4$ s, $\lambda = 0.8$ j/s, $\rho = 0.32$

trovato con la formula che abbiamo visto qualche slide fa.

$p_1 = 0.6$, $p_2 = 0.4$, $c_1=5$, $c_2=3$ → $R=2.2$

$E(T_{Q1}) = 0.095$ s , $E(T_{S1}) = 0.495$ s

$E(T_{Q2}) = 0.233$ s , $E(T_{S2}) = 0.728$ s

tale vincolo è imposto sulla prima coda, che "vede solo se stessa"

Il vincolo è : $E(T_{q1}) = \frac{\rho_1 E[S]}{1-\rho_1} \leq 0.1$ se e solo se $\rho_1 \leq 0.2$

$\rho_1 = \rho \cdot p_1 \leq 0.2 \iff p_1 \leq 0.625 \approx 60\%$ (p1 è la probabilità)

sfruttando tale risultato trovo E[Tq1] = 0.095 (con 60%, non 62%)

18