Performance Modeling of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

Multiserver and Priority scheduling

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

1

esempio:

Assumptions:
- Arrival rate 1 j/s random
- Average demand Z=4x10^5 oxat, expo, do not know size

(Non è che tutti chiedono uguale)

Possible configurations:
- 1 server of capacity C=10^6 oxat/s
- Dual-core of C/2 each one

QoS requirements:
- Average waiting $T_Q < 0.15$ s
- For at least 35% of arrivals average response time $T_S < 0.5$ s

Def. $E(S) = Z/C = 0.4$ s (Z indipendente da C, C è caratteristica fisica, Z no!)
↳ costante ↳ variabile

Capacità processamento macchina != quanto chiede job. alcuni elementi sono caratteristiche fisiche della macchina, non modellabile.

2

"il job" = domanda media, quanto chiede?

QoS requirements:
- Average waiting $T_Q < 0.15$ s

Analytical models
priority scheduling

$\lambda = 1$ j/s, $E(S) = 0.4$ s ⟹ $\rho = 0.4$

- 1 server of capacity C=10^6 oxerat/s

(M/M/1 KP) $E(T_Q) = 0.26$ s = $\frac{\rho E(S)}{1-\rho}$ $E(T_Q)^{\text{Abstract-P}} = 0.2243$ s

- Dual-core of C/2 each one

ogni coda è dimezzato!

$$E(S_i) = \frac{Z}{\frac{C}{2}} = 2\frac{Z}{C} = 2E(S) = 0.8\ s$$

$$E(T_Q)_{Erlang} = \frac{P_Q E(S)}{1-\rho} = 0.15238 \text{ s} > 0.15$$

$$\begin{cases} p_1 = 0.35 \cdot \rho = 0.14 \\ E(T_{a_1}) = \frac{p_1 E(S)}{1-0.14} = 0.065 \\ E(T_{a_2}) = \frac{\rho E(S)}{(1-p_1)(1-\rho)} = 0.31 \end{cases}$$

NON SI INTERROMPE

$E(T_a) = 0.065 \cdot 35\% + 0.31 \cdot 65\% = 0.2243$ s

NB: $\frac{\lambda}{2} E(S^2) = \rho E(S)$ se EXP

- P_a = tutti e due pieni ; $m = 2$, $p(0) = \left[\sum_{i=0}^{1} \frac{(2\rho)^i}{i!} + \frac{(2\rho)^2}{2!(1-\rho)}\right]^{-1} = \left[1 + 2\rho + \frac{(2\rho)^2}{2!(1-\rho)}\right]^{-1} =$ Prob. centro vuoto

e calcolo $P_a = \frac{(m\rho)^m}{m!(1-\rho)} p(0)$, utile per E(Ta along)

3

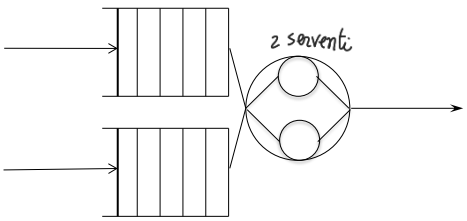
in E(Tq) erlang calcolo Pq e P0 con le formule citate, usando rho = 0.4 ed E(s) = 0.4

QoS requirements:
- Average waiting $T_Q < 0.15$ s

Analytical models
priority scheduling

$\lambda = 1$ j/s, $E(S) = 0.4$ s ⟹ $\rho = 0.4$

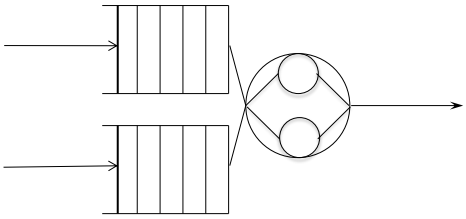
- Dual-core of C/2 each one, classi priorità con multi server



Con due code verrebbe molto più complesso, sconsigliato.

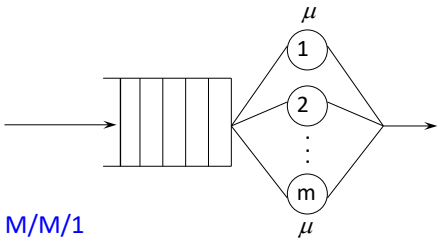
4

Multiserver with priority classes



5

The *Erlang* formula



M/M/m

$$E(T_Q)_{Erlang} = \frac{P_Q E(S)}{1-\rho}$$

M/M/1

$$E(T_Q)_{KP} = \frac{\rho E(S)}{1-\rho} = \frac{E(S_{rem})}{1-\rho}$$

$$E(S) = \frac{E(S_i)}{m}$$

6

Multiserver with priority classes

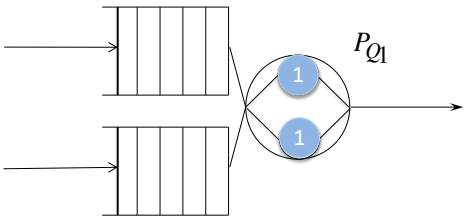
$$E(T_Q) = p_1 \frac{\rho_1 E(S)}{(1-\rho_1)} + p_2 \frac{\rho E(S)}{(1-\rho)(1-\rho_1)}$$

$$E(T_Q) = p_1 \frac{P_{Q1} E(S)}{(1-\rho_1)} + p_2 \frac{P_Q E(S)}{(1-\rho)(1-\rho_1)}$$

C'è prelazione, poichè il primo rapporto vede Pq1, ed è in funzione di rho1, non rho generico.

7

Multiserver with priority classes



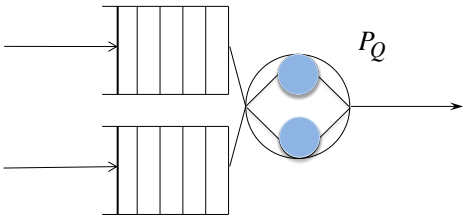
m=2
p_1=0.35
p_2=0.65
ρ = 0.4
ρ_1=0.14
ρ_2=0.26

2 classi, 2 server.
indici '1' e '2' associati alle classi, non ai server!
Classe 1 occupa centro al 14%.

$P_{Q1} = Erlang(\rho_1) = 0.03438$ = tutti server occupati da job di classe 1, usando erlang.

8

Multiserver with priority classes



$P_{Q1} = Erlang(\rho_1) = 0.03438 \qquad P_Q = 0.22857$ (pieni indipendenti dalle classi: solo classe 1, solo classe 2, misti...)

$$E(T_Q) = p_1 \frac{P_{Q1} E(S)}{(1-\rho_1)} + p_2 \frac{P_Q E(S)}{(1-\rho)(1-\rho_1)} = 0.12077$$

(il peso usato per P_Q è sempre p_1)

← classe 1, con prelazione, vede solo il "suo P_1", gli altri NON li vede.

(dual core + prelazione) QoS requirements:

- Average waiting $T_Q < 0.15$ s !! "bound" rispettato

GLOBALMENTE sono sotto il requisito, non solo la classe 1.

9

QoS requirements: (altro requisito)

- For at least 35% of arrivals average response time $T_S < 0.5$ s

Analytical models
priority scheduling

$\lambda = 1$ j/s, $E(S) = 0.4$ s ⟹ $\rho = 0.4$

- 1 server of capacity C=10^6 oxerat/s

$E(T_Q) = 0.26$ s + 0,4 > 0,5, devo dividere le code!

- Dual-core of C/2 each one

$$E(S_i) = \frac{Z}{\frac{C}{2}} = 2\frac{Z}{C} = 2E(S) = 0.8\ s > 0{,}5,$$ senza contare E(Ta)

10

QoS requirements:

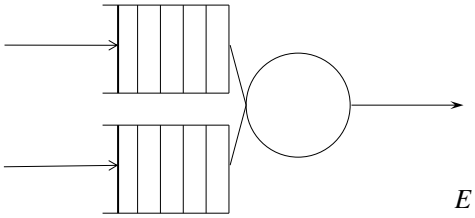
- For at least 35% of arrivals average response time $T_S < 0.5$ s

$\lambda =$ 1 j/s, $E(S) = 0.4$ s ⟹ $\rho =$ 0.4

- 1 server of capacity C=10^6 oxerat/s

p_1=0.35
p_2=0.65
$\rho = 0.4$
ρ_1=0.14
ρ_2=0.26

Abstract-P



$E(T_{S1}) = 0.4651162$

∀ obiettivo ∃ soluzione diversa.

servente singolo + abstract priority