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

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

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

2

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

$E(T_Q) = 0.26$ s $E(T_Q)^{\text{Abstract-P}} = 0.2243$ s

- Dual-core of C/2 each one

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

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

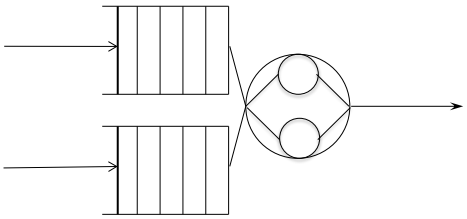
3

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

Analytical models
priority scheduling

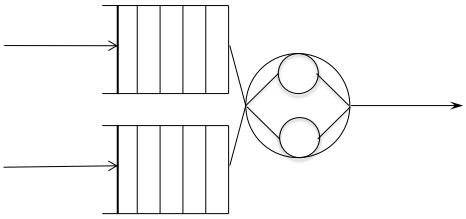
$\lambda =$ 1 j/s, $E(S) = 0.4$ s ⟹ $\rho =$ 0.4

- Dual-core of C/2 each one



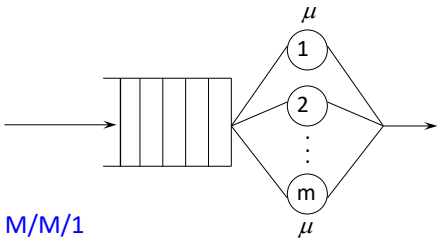
4

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

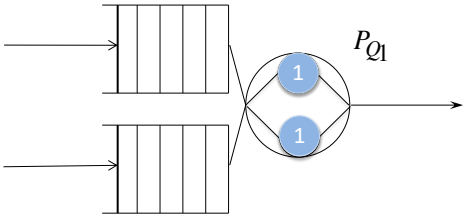
Multiserver with priority classes

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

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

7

Multiserver with priority classes

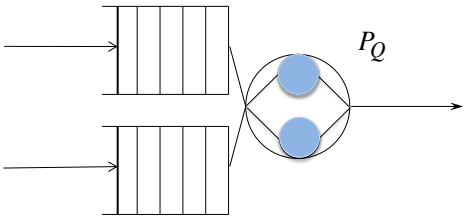


m=2
p_1=0.35
p_2=0.65
ρ = 0.4
ρ_1=0.14
ρ_2=0.26

$$P_{Q1} = Erlang(\rho_1) = 0.03438$$

8

Multiserver with priority classes



$$P_{Q1} = Erlang(\rho_1) = 0.03438 \qquad P_Q = 0.22857$$

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

QoS requirements:

- Average waiting $T_Q < 0.15$ s **!!**

9

QoS requirements:

- 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

- Dual-core of C/2 each one

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

10

QoS requirements:

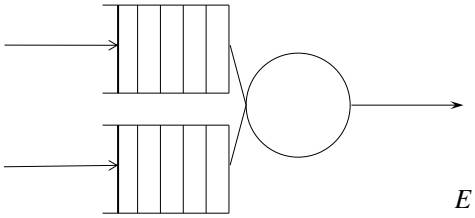
- For at least 35% of arrivals average response time $T_S < 0.5$ s

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

Abstract-P



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

$$E\left(T_{S1}\right) = 0.4651162$$