Performance Modeling of Computer Systems and Networks

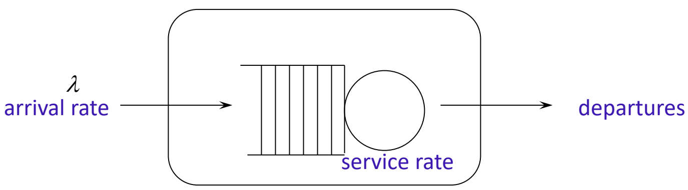
Prof. Vittoria de Nitto Personè

The model for a service center: analytical results

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

1

Server center



Little's law

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

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

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

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

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

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

OK se conosco popolazione coda ... ma senza ?

2

The Kendall notation

A/S/m/B/N/D

- Interarrival distribution
- Service distribution
- n° of servers
- Queue capacity (Finito)
- Population size (Popolazione fissa)
- Queue discipline

Distributions: D, M, E_k, H_2, G → generale (qualsiasi)

H_2 → esp. 2 stati

D: deterministica

M: markov espon.

E_k: erlang, con param. K

$\underbrace{\text{Non-preemptive}}_{\text{non interrompibili}}$ abstract scheduling (Non conosco/uso dimensioni) → M/G/1 (nostra notazione)

FIFO, LIFO-non-preemp, Random] cambia variabilita'

It seems like

- FIFO should have the best mean response time because jobs are serviced most closely to the time they arrive
- LIFO may make a job wait a very long time (difficile uno venga scelto)

all the above policies have exactly the *same* mean response time.

Stesso tempo di risposta medio.

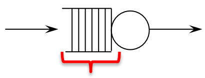
S.403 'Performance'

1930: The Khinchin Pollaczek equation (KP)

Markov espon. ← generale ↑ → servente singolo

M/G/1 abstract scheduling

$$E(N_Q) = \underbrace{\frac{\rho^2}{2(1-\rho)}}_{\text{dipende da utilizzazione}}\left[1 + \frac{\sigma^2(S)}{E(S)^2}\right]$$



$= C^2$
Squared coefficient of variation
Service time dispersion

1. Any service time distribution
2. Poisson arrivals
3. Abstract discipline (FIFO, LIFO, RAND…)

5

Phase-type distributions

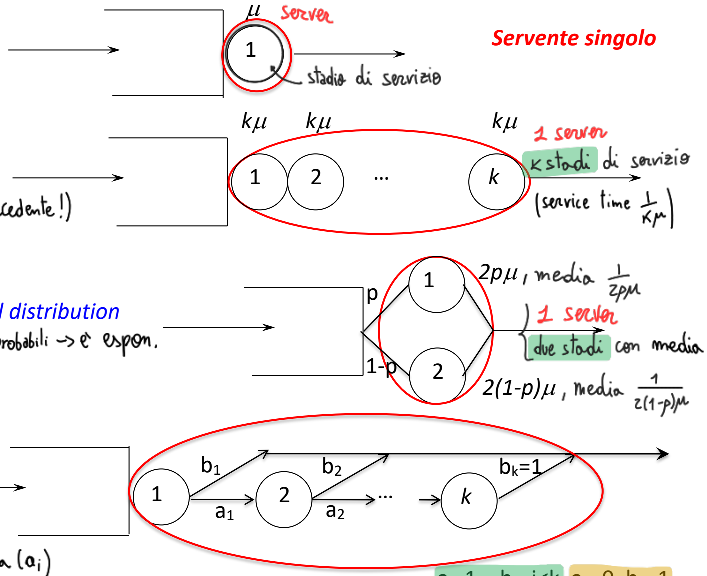
Servente singolo

Exponential

k-Erlang
(se k=1 è caso precedente!)

hyperexponential distribution
se p=0,5 → stadi equiprobabili → è espon.

Cox distribution
può modellare ogni distribuzione, fasi arbitrarie ed esponenziali, o esco (b_i) o fase successiva (a_i)



successione di k tempi, ogni fase è esponenziale, uguale tra loro, di tasso μk. Il Job esce dopo aver completato tutte le fasi.

esponenziali a 2 stadi (≈ 2 freq.)

$a_i = 1 - b_i$, $i<k$, $a_k=0$, $b_k=1$
Finale

6

The Khinchin Pollaczek equation (KP)

$$E(N_Q) = \frac{\rho^2}{2(1-\rho)}\left[1 + C^2\right]$$

The mean queue population grows as C^2 (coeff. quadratico, dispersione tempi servizio attorno)

var. crescente

D $\longrightarrow$ $C^2=0$

E_k $\longrightarrow$ $C^2 = \frac{1}{k},\ k \geq 1$

(espon.) M $\longrightarrow$ $C^2=1$

H_2 $\longrightarrow$ $C^2 = g(p) = \frac{1}{2p(1-p)} - 1$

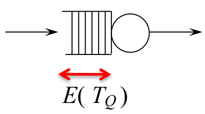
60% esperimenti media $\frac{1}{2p\mu}$, 40% $\frac{1}{2(1-p)\cdot\mu}$

$p = 0.6 \quad C^2 = 1.08\overline{3}$
$p = 0.7 \quad C^2 = 1.38095$
$p = 0.8 \quad C^2 = 2.125$
$p = 0.9 \quad C^2 = 4.\overline{5}$

Congestione coda = pop. media nella coda

The Khinchin Pollaczek equation (KP)

M/G/1 abstract scheduling



$$\underbrace{E(T_Q) = \frac{E(N_Q)}{\lambda}}_{\text{Little}} = \frac{\rho^2}{\lambda 2(1-\rho)}\left[1 + C^2\right] = \frac{\rho}{1-\rho}\,\frac{C^2+1}{2}\,E(S)$$

$C^2 = \frac{var}{media^2}$

The Khinchin Pollaczek equation (KP)

generali

$g(p) = \frac{1}{2p(1-p)} - 1$

$E(N_Q) = \frac{\rho^2}{2(1-\rho)}[1+C^2]$, $E(T_Q) = \frac{\rho}{1-\rho}\frac{C^2+1}{2}E(S)$

Service time	$E(N_Q)$	$E(T_Q)$
Deterministic, M/D/1	$\frac{\rho^2}{2(1-\rho)}$	$\frac{\rho E(S)}{2(1-\rho)}$
Markovian, M/M/1	metà $\frac{\rho^2}{1-\rho}$	metà $\frac{\rho E(S)}{1-\rho}$
K-Erlang, M/E$_k$/1 $\sigma^2(S) = \frac{E(S)^2}{k}$	$\frac{\rho^2}{2(1-\rho)}\left(1+\frac{1}{k}\right)$	$\frac{\rho E(S)}{2(1-\rho)}\left(1+\frac{1}{k}\right)$
Hyperexpo, M/H$_2$/1 $\sigma^2(S) = E(S)^2 g(p)$	$\frac{\rho^2}{2(1-\rho)}(1+g(p))$	$\frac{\rho E(S)}{2(1-\rho)}(1+g(p))$

tutti i valori sono indipendenti da C^2.

9

• esempio provider che deve fronteggiare gli arrivi raddoppiati:

$\lambda' = 2\lambda$
$\mu' = 2\mu$
$\rho' = \rho$

$$\begin{cases} E(S') = \frac{E(S)}{2} \\ E(T_S') = E(T_Q') + E(S') = \frac{E(T_Q)}{2} + \frac{E(S)}{2} = \frac{E(T_S)}{2} \end{cases}$$

NB: $E(T_Q)$ dip da $(E(S), \rho)$

Service time Sensitivity

$$E(N_Q)_D \le E(N_Q)_{E_k} \le E(N_Q)_M \le E(N_Q)_{H_2}$$

$$\sigma^2(N_Q)_D \le \sigma^2(N_Q)_{E_k} \le \sigma^2(N_Q)_M \le \sigma^2(N_Q)_{H_2}$$

By considering $E(N_S)= E(N_Q)+\rho$, the same order holds for the variable N_S

By considering the Little's equation, the same order can be derived for the mean times $E(T_S)$ and $E(T_Q)$, but just for the 1° order moment, not for the variance

10

Discipline Sensitivity (scheduling)

By definition, KP holds for any abstract service discipline, so

$$E(N_Q)_{FIFO} = E(N_Q)_{LIFO} = E(N_Q)_{RAND} = E(N_Q)_{abstract}$$

$$\sigma^2(N_Q)_{FIFO} = \sigma^2(N_Q)_{LIFO} = \sigma^2(N_Q)_{RAND} = \sigma^2(N_Q)_{abstract}$$

By considering $E(N_S)= E(N_Q)+\rho$, the same equalities hold for the variable N_S

By considering the Little's equation, the same holds for $E(T_S)$ and $E(T_Q)$,

$$E(T_Q)_{FIFO} = E(T_Q)_{LIFO} = E(T_Q)_{RAND} = E(T_Q)_{abstract}$$

?? Is $\sigma^2(T_Q)$ the same for all these policies? ??

lo scheduling conta!

Discipline Sensitivity

No!

LIFO can generate some extremely high response times because we have to wait for system to become empty to take care of that first arrival

"b" vede arrivare dei job, e il suo tempo di esecuzione ne risente.

$$\sigma^2(T_Q)_{FIFO} \le \sigma^2(T_Q)_{RAND} \le \sigma^2(T_Q)_{LIFO}$$

Qui parlo di varianza (non media), e vale se ρ è grande, cioè se "b" ha alta probabilità che coda sia piena.