Performance Modeling of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

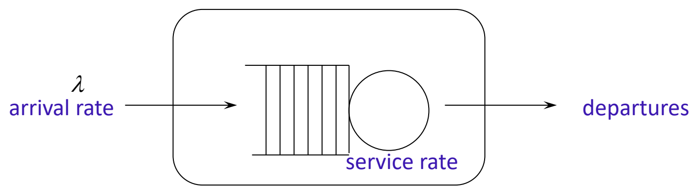
The model for a service center: analytical results

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

Copyright © Vittoria de Nitto Personè, 2021
https://creativecommons.org/licenses/by-nc-nd/4.0/
(CC BY-NC-ND 4.0)

1

Server center



Little's law

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

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

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

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

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

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

2

The Kendall notation

A/S/m/B/N/D

- A: Interarrival distribution
- S: Service distribution
- m: n° of servers
- B: Queue capacity
- N: Population size
- D: Queue discipline

Distributions: D, M, E_k, H_2, G

deterministica, esponenziali M, sequenziali di esponenziali Erlang con k numero di fasi con stessa media, parallele esponenziali H (due fasi //, cioè iperesponenziale) e distribuzione generale G (qualsiasi distribuzione)

Non-preemptive abstract scheduling

FIFO, LIFO-non-preemp, Random

It seems like

FIFO should have the best mean response time because jobs are serviced most closely to the time they arrive (rispetta ordine di arrivo)

LIFO may make a job wait a very long time

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

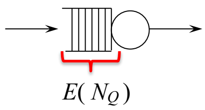
(hanno stessa media ma NON HANNO STESSA VARIANZA)

1930: The Khinchin Pollaczek equation (KP)

interrarivi esponenziali, distribuzione servente generale, 1 servente

M/G/1 abstract scheduling

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



$= C^2$
Squared coefficient of variation
Service time dispersion
rapporto tra varianza e quadrato delle media

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

Più il coefficiente cresce, più le prestazioni peggiorano!

Rho è la vera misura del carico, non lambda, il quale pesa pochissimo come carico, in quanto rho = lambda * E[S]

5

TIPO M/G/1, SI ANALIZZA UN SOLO JOB, QUINDI QUESTI SONO I PERCORSI CHE PUO' SEGUIRE UN JOB. IL SERVENTE E' SEMPRE SINGOLO

Phase-type distributions

unica fase di servizio
sto modellando tempo job con variabile esponenziale

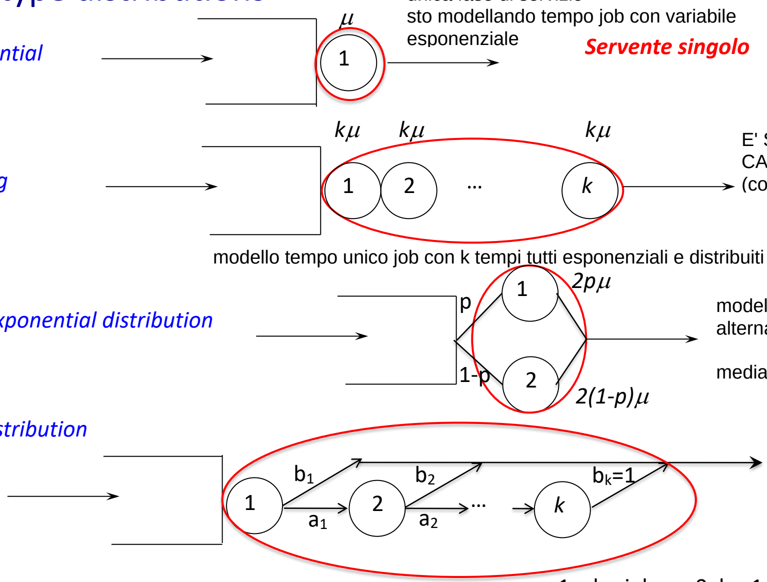
Servente singolo

Exponential

k-Erlang

hyperexponential distribution

Cox distribution



che senso ha modellare queste cose, se mediamente sono uguali?

modello diverse variabilità!!

E' SEMPRE UNICO SERVENTE, CAMBIA COME MODELLO TEMPO (comparo a parità di media)

modello tempo unico job con k tempi tutti esponenziali e distribuiti allo stesso modo.

modello 1 tempo di servizio con 2 fasi alternative esponenziali con due tassi diversi

media: 1/u (media del servizio uguale)

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

6

serie di fasi esponenziali, nella erlang passo per tutti gli stati, qui il job potrebbe fare un solo stadio ed uscire, fare due stadi e uscire,..., farli tutti! Ma sempre un job c'è.

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

- D $\longrightarrow$ $C^2=0$
- E_k $\longrightarrow$ $C^2 = \frac{1}{k}$, $k \geq 1$
- M $\longrightarrow$ $C^2=1$
- H_2 $\longrightarrow$ $C^2 = g(p) = \frac{1}{2p(1-p)} - 1$

variabilità cresce nel verso della freccia.

in funzione della probabilità,
dipende da p.
per p = 0.5, ottengo 1 come l'esponenziale.

$p = 0.6 \quad C^2 = 1.08\overline{3}$

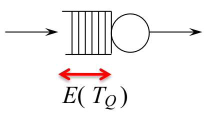
$p = 0.7 \quad C^2 = 1.38095$

$p = 0.8 \quad C^2 = 2.125$

$p = 0.9 \quad C^2 = 4.\overline{5}$

The Khinchin Pollaczek equation (KP)

M/G/1 abstract scheduling



$\rho = \lambda E[S]$

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

Se c^2, anche per utilizzazione grande, questo tempo nella coda più esplodere,
anche essere 30 volte il tempo di servizio è tanto!

fine

The Khinchin Pollaczek equation (KP)

$$g(p) = \frac{1}{2p(1-p)} - 1 \qquad E(N_Q) = \frac{\rho^2}{2(1-\rho)}\left[1 + C^2\right], \quad E(T_Q) = \frac{\rho}{1-\rho}\frac{C^2+1}{2}E(S)$$

Service time	$E(N_Q)$	$E(T_Q)$
Determinisctic, M/D/1	$\frac{\rho^2}{2(1-\rho)}$	$\frac{\rho E(S)}{2(1-\rho)}$
Markovian, M/M/1	$\frac{\rho^2}{1-\rho}$	$\frac{\rho E(S)}{1-\rho}$
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 $\sigma^2(S) = E(S)^2 g(p)$	$\frac{\rho^2}{2(1-\rho)}(1+g(p))$	$\frac{\rho E(S)}{2(1-\rho)}(1+g(p))$

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

Discipline Sensitivity

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?

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

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