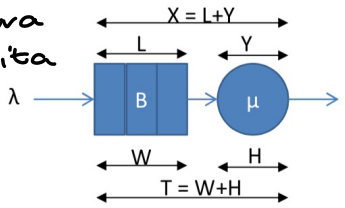
Sistemi di servizio

→ un'astrazione utile per studiare le prestazioni di molti sistemi reali.

→ è caratterizzato da un processo di arrivo, da una coda di attesa e da (uno o più) server.

B=0 → sistema a pura perdita

B= ∞ → sistema a pura attesa



λ = frequenza di arrivo	Y = # in servizio
1/μ = tempo medio di servizio	X = # nel sistema
B = dimensione del buffer	W = tempo speso in coda
N = numero di server	H = tempo di servizio
L = # in coda	T = tempo speso nel sistema

Formula di Little

→ vale per sotto condizioni molto generali riguardanti i sistemi di servizi a pura attesa:

sistema stabile, sistema work-conserving

↳ il lavoro fatto dal server non si perde.

- $\overline{X} = \lambda \overline{T}$
- Vale anche per i sottosistemi:

$\overline{L} = \lambda \overline{W}$ e $\overline{Y} = \lambda H = \frac{\lambda}{\mu} = \rho$

carico del sistema

Distribuzioni degli arrivi dei servizi

Supponiamo di osservare il flusso di auto su un'autostrada e la coda ad un casello. Se il flusso è di 10 auto al minuto e la coda è stabile, e lunga in media 100 auto.

Formula di Little?

- Il tempo medio $\overline{T} = 100/10 = 10$ min (100 → n° auto; 10 → λ = frequenza di arrivo)
- $\mu > 10$ auto/min

Modelli Markoviani

- la distribuzione dei tempi di interarrivo e di servizio è esponenziale

$$P\{t_i \leq t\} = 1 - e^{-\lambda t}$$

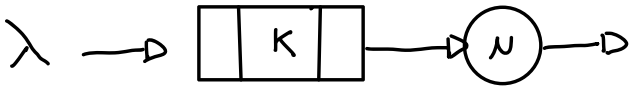
$$\overline{A(t)} = \lambda t$$

Sistemi senza memoria

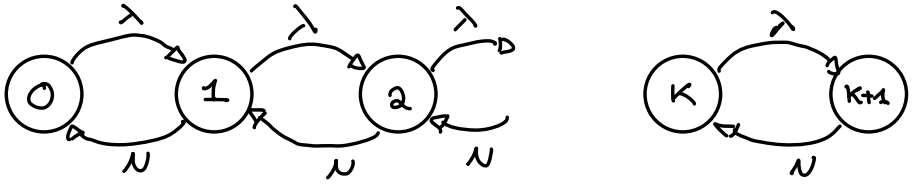
- il comportamento del sistema a partire da un momento in cui viene osservato, non dipende da quanto successo prima.

Coda M/M/1/K

- Una coda con un servente, interarrivo e servizio esponenziali e buffer con K posizioni si chiama coda M/M/1/K



- Il sistema si risolve con una catena di Markov in cui ogni stato rappresenta il numero di elementi nel sistema:



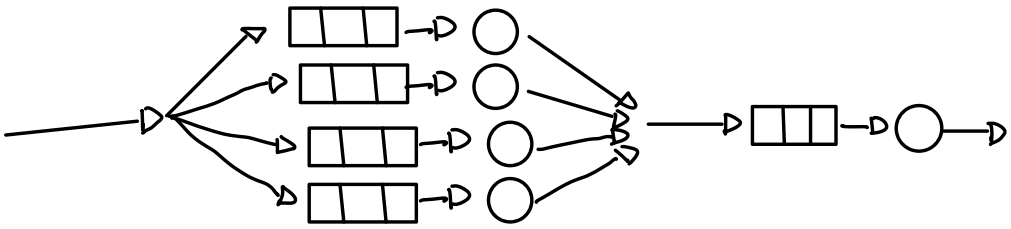
- Se P_i è la probabilità dello stato i-esimo

$$\begin{cases} \lambda P_0 = \nu P_1 \\ \lambda P_1 = \nu P_2 \\ \lambda P_K = \nu P_{K+1} \\ \sum_{i=0}^{K+1} P_i = 1 \end{cases}$$

$$P_0 = \frac{1-\rho}{1-\rho^{K+2}} \qquad P_j = \frac{1-\rho}{1-\rho^{K+2}} \rho^j$$

Esempio

- Un sistema informativo Web-based sia costituita da quattro server di front-end operanti a divisione di carico e un server di backoffice che gestisce tutte le richieste.
  Si supponga che ciascun server di front-end elabori una richiesta in media 20ms e il server 10 ms. Se al sistema arrivano 80 richieste al secondo e il tempo di risposta è di 95 ms. Quanti server di front-end stanno probabilmente funzionando.



- Abbiamo due sottosistemi in cascata e supponiamo che tutto il traffico offerto al primo venga offerto al primo venga offerto anche al secondo.

$$T = \frac{1}{(\mu - \lambda)} = \frac{1}{(100 - 80)} = 50 \text{ msec}$$

Otteniamo:

- 1 server $T = 1/(\mu - \lambda)_{n \text{ server}} = 1/(50 - 80) =$ congestione
- 2 server $T = 1/(50 - 40) = 100 \text{ ms}$
- 3 server $T = 1/(50 - 26{,}6..) = 48 \text{ ms}$
- 4 server $T = 1/(50 - 20) = 33{,}\bar{3} \text{ ms}$

La somma dei ritardi con 3 server è di 92,85 ($\sim$ 95 ms)

Sistemi a perdita e principio PASTA

PASTA

(Poisson Arrivals see Time Averages)

Gli arrivi «campionano» lo stato del sistema e la distribuzione di probabilità dello stato

Se un sistema a coda ha un buffer finito e il processo degli arrivi permette che giungano al sistema più job di quanti ne può contenere.

$$p_{K+1} = \frac{1-\rho}{1-\rho^{K+2}} \rho^{K+1}$$

Coda M/D/1

Una coda con un servente, tempo di interarrivo esponenziale e tempo di servizio deterministico

L'analisi del sistema è matematicamente meno semplice di quella relativa ad una coda M/D/1

$$\bar{X} = \left(1 - \frac{\rho}{2}\right)\frac{\rho}{1-\rho} \qquad \bar{Y} = \rho$$

$$\bar{T} = \left(1 - \frac{\rho}{2}\right)\frac{1}{\nu - \lambda} \qquad \bar{W} = \frac{1}{\nu}\,\frac{\rho}{2(1-\rho)}$$

$$\bar{L} = \frac{\rho^2}{2(1-\rho)} \qquad \bar{H} = \frac{1}{\nu}$$