Processi di Markov e code markoviane

Gianluca Reali

sono solo esempi

Motivazioni

Supponiamo di dover affrontare le seguenti problematiche:

- Progettare una cloud ibrida facendo in modo che la probabilità di offloading sia inferiore a un valore predifinito
- Dato un sistema di calcolo distribuito, valutate l'effetto della presenza di un broker sulla latenza media e massima di accesso a un servizio
- Determinare quanti server includere in un tenant per far sì che le richieste di servizio possano essere servite immediatamente con una probabilotà predefinita.
- Valutare quante indirizzi IP configurare in una subnet di OpenStack affinché la probabilità di non trovarne disponibili da parte di una istanza sia inferiore a un valore predefinito.
- Calcolare il numero minimo e massimo di istanze da configurare in un hpa di kubernetes affinché la la latenza media di accesso al servizio sia inferiore a un valore predefinito
- Calcolare il numero minimo di connessioni da include a un pool predefinito per la connessione dal front-end al back-end di un servizio web affinché la probabilità di non trovarne disponibili sia inferiore a un valore predefinito.
- In un servizio faas in ambito edge, determinare la frequenza massima accettabile di accesso a una funzione affiché la latenza della risposta sui minore o uguale al valore massimo tollerabile.

Obiettivi della lezione

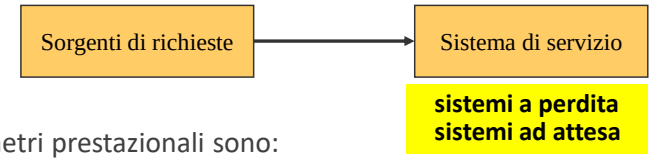
Comprendere i concetti di base relativi alle prestazioni di un sistema dinamico, sia concentrato sia distribuito.

- efficienza
- utilizzazione
- ritardo
- perdita
- Tempi di inattività

Essere in grado di valutare quantitativamente tali parametri in casi semplici ma significativi --> il goal

Modello del sistema

Per valutare quantitativamente le prestazioni di un sistema dinamico è necessario rappresentare in modo astratto le sue funzionalità:



sistemi a perdita
sistemi ad attesa

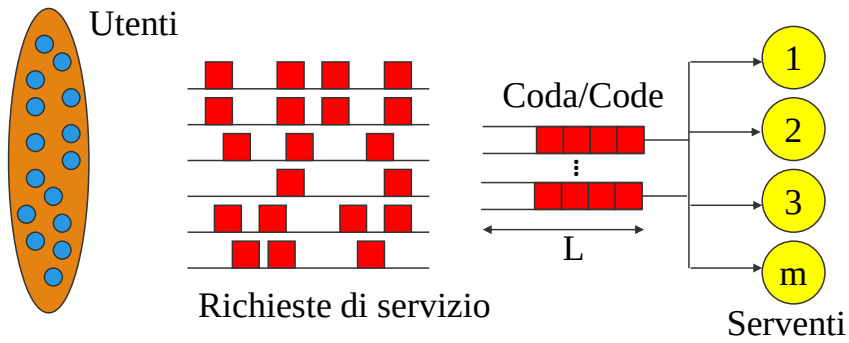
Tipici parametri prestazionali sono:

- sistemi a perdita:
  - frequenza con cui le richieste sono rifiutate
  - intensità del carico smaltito
  - durata dei periodi di congestione
- sistemi ad attesa:
  - tempo (…?) di attesa
  - tempo (…?) di permanenza
  - numero (…?) di richieste in attesa

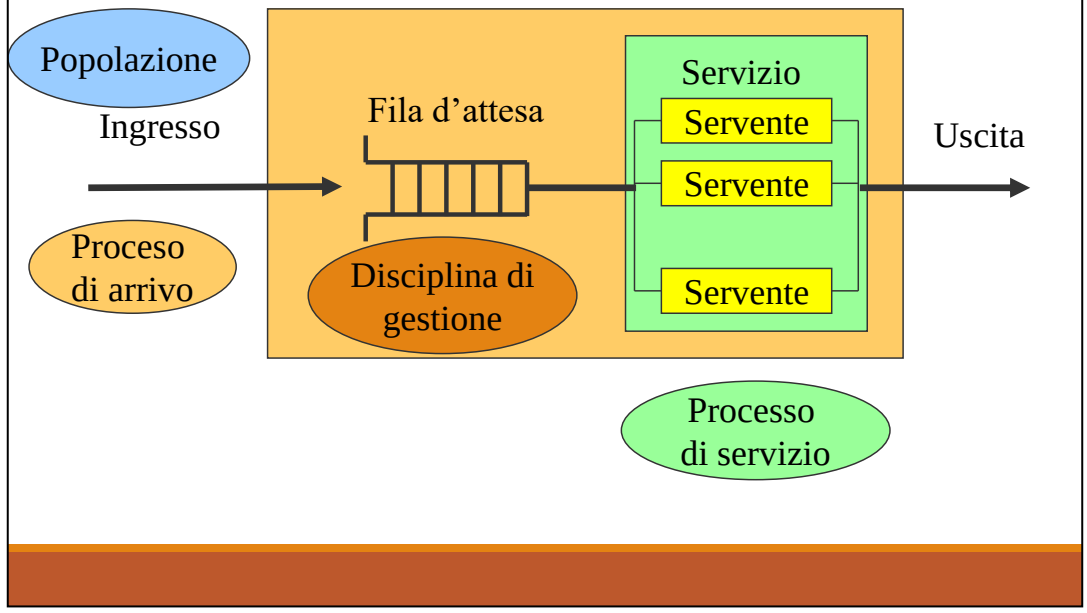
Sistema di servizio

Risoluzione delle contese di utilizzazione:

- sistema a perdita pura (L=0)
- sistema orientato alla perdita (L piccolo)
- sistema orientato al ritardo (L grande)
- sistema a ritardo senza perdita (L→∞ oppure L≥utenti)



Caratterizzazione dei Sistemi a coda



Elementi descrittivi del sistema a coda

Cardinalità della popolazione:
- finita, infinita

Processo di Arrivo :
- frequenza media, varianza …

Accodamento:
- dimensione della coda:
  - finita, infinita
- numero di code

Disciplina di gestione, o Selezione:
- discipline di coda
  - primo arrivato primo servito (FIFO)
  - shortest job first (SJF)
  - ...
- classi di priorità

Numero dei serventi

Fissiamo le grandezze principali:

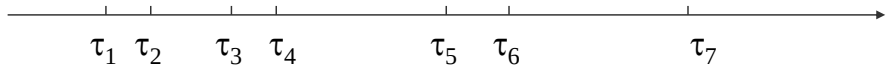
Caratterizzazione del servizio

- Siano
  - **n** = cardinalità della popolazione
  - **L**= dimensione della coda
  - **m**=numero di serventi
  - **C**=m+L capacità del sistema
- Classificazione:
  - se $n \leq C$ e $L>0$: sistema ad attesa (senza perdita)
  - se $n>C$ e $L>0$: sistema a perdita (con attesa)
  - se $n>C$ e $L=0$: sistema a perdita pura (senza attesa)
- Tipico aarametro prestazionale:
  - Tempo di sistema (s) = tempo di attesa (w) + tempo di servizio (x)

Caratterizzazione della domanda

La domanda è caratterizzata dalle richieste di servizio presentate dagli utenti del sistema

- consideriamo una sequenza di istanti di richiesta di servizio (τ_i)



- tali istanti sono distribuiti secondo una specifica descrizione statistica sull'asse dei tempi e costituiscono un insieme numerabile

Tempo di interarrivo:

- *i-esimo* tempo di interarrivo t_i è l'intervallo che intercorre tra l'istante di presentazione della richiesta *(i-1)-esima* (τ_{i-1}) e quello della richiesta *i-esima* (τ_i)

$$t_i = \tau_i - \tau_{i-1} \qquad i = 1,2,\ldots$$

Tempo di servizio

Sia L_i la 'quantità di lavoro', espresso in modo *quantitativo*, che i serventi del sistema devono erogare per soddisfare la richiesta *i-esima*

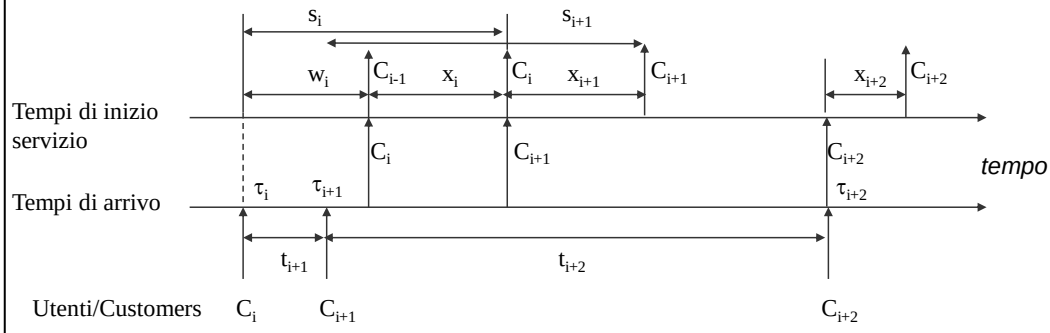
Definiamo l'*i-esimo* tempo di servizio x_i come l'intervallo di tempo che un servente impiega per soddisfare la richiesta *i-esima*

Se supponiamo che la capacità di erogare lavoro Γ (quantità di lavoro erogato nell'unità di tempo) di ogni servente del sistema sia identica, per servire la richiesta *i-esima* sarà necessario un tempo uguale a

$$x_i = L_i / \Gamma \qquad i = 1,2,...$$

Diagramma temporale del sistema

- C_i: *i-esima* richiesta di servizio (o *cliente*) ad entrare nel sistema
- Grandezze temporali:
  - τ_i: tempo di arrivo *i-esima* richiesta di servizio
  - t_i: tempo di interarrivo tra la richiesta *(i-1)-esima* e la *i-esima*
  - w_i: tempo di attesa in coda della *i-esima* richiesta di servizio
  - x_i: tempo di servizio dell'*i-esima* richiesta
  - s_i: tempo di sistema dell'*i-esima* richiesta



Processi di ingresso e di servizio

La sequenza dei tempi di interarrivo $\{t_i\}$ e quella dei tempi di servizio $\{x_i\}$ costituiscono delle realizzazioni di due processi stocastici:

- il processo di ingresso
- il processo di servizio

Ogni valore t_i e x_i è una realizzazione di una variabile aleatoria

N.B.

Normalmente si suppone i due processi siano stazionari, almeno WSS, e statisticamente indipendenti

Grandezze limite

Comportamento al limite delle variabili aleatorie:

- tempi di interarrivo

$$\tilde{t} = \lim_{n\to\infty} t_n$$

$$P[t_n \le t] = A_n(t) \xrightarrow[n\to\infty]{} P[\tilde{t} \le t] = A(t)$$

$$E[t_n] = \bar{t}_n \xrightarrow[n\to\infty]{} E[\tilde{t}] = \bar{t} = \frac{1}{\lambda}$$

Analogamente si possono definire le stesse grandezze per

- i tempi di attesa
- i tempi di servizio
- i tempi di sistema

$$\tilde{w} = \lim_{n\to\infty} w_n, \quad E[\tilde{w}] = \overline{w} = W$$

$$\tilde{x} = \lim_{n\to\infty} x_n, \quad E[\tilde{x}] = \bar{x} = \frac{1}{\mu}$$

$$\tilde{s} = \lim_{n\to\infty} s_n, \quad E[\tilde{s}] = \bar{s} = T$$

Caratterizzazione di un sistema a coda

Una coda è definita da:

- processo degli arrivi (D.d.p.) $A(t)$
- tempi di servizio (D.d.p.) $B(t)$
- numero di serventi m
- dimensione del sistema L
- cardinalità della popolazione n
- disciplina di servizio

Notazione sintetica di **Kendall** (*A/B/m/L/n*)

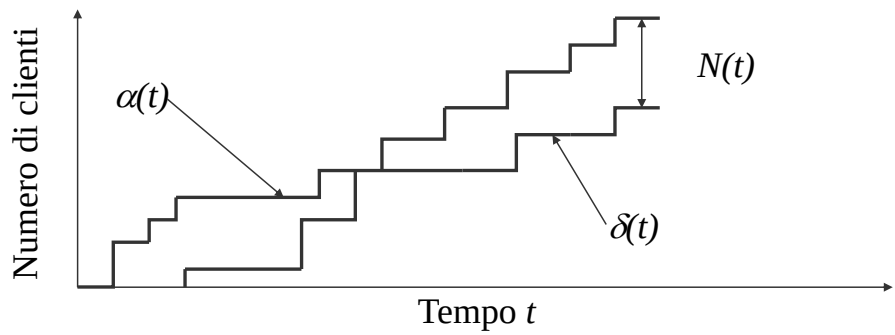
- A e B posono assumere i valori:
  - M esponenziale negativa o “Markoviana”
  - D deterministica o costante
  - E_i erlangiana con i stadi
  - H_i iper-esponenziale con i stadi
  - G generale

N.B.

Legge di Little

Ipotesi: **sistema senza perdite**

- $\alpha(t)$: numero di arrivi in [0,t]
- $\delta(t)$: numero di partenze in [0,t]
- $N(t)=\alpha(t)-\delta(t)$
- $\gamma(t)$: l'area tra le due curve rappresenta il tempo che tutti i clienti hanno trascorso nel sistema nell'intervallo [0,t]



Legge di Little

Definiamo:

- frequenza media di interarrivo in [0,t]: $\lambda_t=\alpha(t)/t$ ***(a)***
- tempo medio di sistema per ogni utente in [0,t]: $T_t=\gamma(t)/\alpha(t)$ (***b***)
- numero medio di clienti nel sistema in [0,t]: $N_t=\gamma(t)/t$

Quindi si ottiene:

- $N_t=\gamma(t)/t$; dalla (***a***) $N_t=\gamma(t)\ \lambda_t/\alpha(t)$; dalla (***b***) $N_t=T_t*\lambda_t$
- assumendo che esistano per il sistema i valori limite: $\lambda = \lim_{t\to\infty} \lambda_t\,,\ T = \lim_{t\to\infty} T_t$
- si ottiene:

Super N.B.

$$\overline{N} = \lambda T = \lambda W + \lambda \overline{x} = \overline{N}_q + \overline{N}_s$$

Risultato indipendente da A(t), B(t) e m

IPOTESI: ASSENZA DI PERDITE!

Fattore di utilizzazione

Definizione: ***rapporto*** tra la frequenza con la quale di 'lavoro' entra nel sistema e quella con la quale i serventi riescono a smaltirlo

Assumendo che la frequenza di erogazione del servizio sia indipendente da qualsiasi altro parametro del sistema, possiamo scrivere:

- caso 1 servente: $\rho = \lambda / \mu = \lambda \bar{x}$
- caso m serventi: $\rho = \lambda /(m\mu) = \lambda \bar{x} / m$

λ/μ, ossia il ritmo con cui il lavoro entra nel sistema, normalizzata alla capacità del singolo servente, è l'intensità di traffico (espressa in Erlang), vale $m\rho$

Fattore di utilizzazione

Se $0 \leq \rho < 1$:

- può essere interpretato come:
  - *frazione di serventi occupati*
- rappresenta la condizione di stabilità del sistema. Infatti, nel caso *G/G/m*, risulta quanto segue:
  - dato un intervallo τ sufficientemente grande, con probabilità tendente a 1 il numero di arrivi sarà pari a $\lambda\tau$
  - quindi un servente sarà occupato per un tempo pari a $\tau(1-p_0)$, definendo p_0 come la probabilità di trovare il server libero in un generico istante di tempo
  - il numero di utenti serviti in tale intervallo sarà $m(\tau-\tau p_0)\mu$
  - eguagliando il numero di arrivi con in numero di utenti serviti (condizione di stabilità), si ottiene:

$$\lambda\tau \cong m(\tau - \tau p_0)\mu \xrightarrow[\tau\to\infty]{} \rho = 1 - p_0$$

Concetto di Stato

- Intuitivamente, le prestazioni osservate da un utente generico (ossia che non gode di diritti di prelazione dei serventi) che entra in un sistema a coda dipendono da quanti utenti trova già presenti nel sistema.
- Le prestazioni, quindi, dipendono dalla storia più o meno recente del sistema.
- Ciò significa che quanto accaduto in passato ha lasciato traccia nel sistema, traccia che influenzerà il futuro del sistema stesso, e che chiameremo STATO del sistema.

Concetto di stato

La probabilità che il sistema si trovi in un generico stato *j* al tempo *t è uguale a:*

$$\Pi_j(t)=P[X(t)=j]$$

Sia la variabile t sia lo stato X(t) possono assumere valori in un insieme continuo o in un insieme discreto (finito o numerabile)

Noi considereremo processi continui nel tempo e discreti nelle realizzazioni

Chiaramente $\Sigma_{j=0,1\ldots}\Pi_j(t)=1$

Catene di Markov

Un processo aleatorio è detto Catena di Markov se lo spazio degli stati è discreto e gode della proprietà di Markov

Proprietà di Markov:

- dato un insieme di variabili aleatorie $\{X_n\}$, questo forma una catena di Markov se la probabilità di trovarsi in un tempo futuro in un determinato stato può essere espressa in funzione solo dello stato assunto al tempo corrente e non occorre specificare quali stati sono stati assunti in precedenza:
- lo stato attuale riassume tutta la storia del sistema
- la conoscenza più recente dello stato del sistema rende inutile la conoscenza degli stati assunti in precedenza
- la conoscenza del passato non ci consente di predire quanto tempo il processo debba rimanere nello stato in cui si trova
  - la distribuzione del tempo che il processo rimane in uno stato è "senza memoria", e nell'ipotesi d tempo continuo quest porta alla distribuzione esponenziale del tempo di permanenza nello stato.

N.B.

Catene di Markov

Formalmente:

- il processo aleatorio *X(t)* forma una catena di Markov tempo-continua se per tutti gli interi *n* e per una sequenza di istanti temporali $t_1<t_2<...<t_n<t_{n+1}$ risulta

$$P[X(t_{n+1})=j \mid X(t_n)=i_n, X(t_{n-1})=i_{n-1},..., X(t_1)=i_1] = P[X(t_{n+1})=j \mid X(t_n)=i_n]$$

Classificazione:

- Catene di Markov tempo continuo:
  - distribuzione del tempo di stato esponenziale $p_T(t)=\lambda e^{-\lambda t}$, $t\geq 0$
- Catene di Markov tempo discreto
  - distribuzione del tempo di stato geometrica $p_N(n)=(1-p)^{n-1}p$, $n\geq 0$

Catene di Markov

Probabilità di transizione da uno stato $i \to j$:

$$p_{ij}(s,t) = P[X(t) = j \mid X(s) = i] \quad \text{per } t \geq s$$

(A | C)

per passare dallo stato *i* all'istante *s* allo stato *j* all'istante $t>s$, il processo dovrà passare per uno stato intermedio *k* ad un certo istante intermedio *u:*

$$p_{ij}(s,t) = \sum_k P[X(t) = j, X(u) = k \mid X(s) = i] =$$

(P(A , B | C))

$$= \sum_k P[X(u) = k \mid X(s) = i] P[X(t) = j \mid X(s) = i, X(u) = k] =$$

(P(B | C) * P(A | C , B))

$$= \sum_k P[X(u) = k \mid X(s) = i] P[X(t) = j \mid X(u) = k] =$$

$$= \sum_k p_{ik}(s,u)\, p_{kj}(u,t)$$

Equazioni di Chapman-Kolmogorov

$$\sum_B P(A, B|C) = P(A|C) \qquad P(A|B,C)P(B|C) = \frac{P(A,B,C)}{P(B,C)} \frac{P(B,C)}{P(C)} = P(A,B|C)$$

Catene di Markov

$$p_{ij}(s,t) = \sum_k p_{ik}(s,u)\, p_{kj}(u,t) =$$

$$= \begin{bmatrix} p_{11}(s,u) & p_{12}(s,u)\ldots & p_{1k}(s,u) & \ldots \\ p_{21}(s,u) & p_{22}(s,u) & p_{2k}(s,u) & \ldots \\ & & & \\ p_{i1}(s,u) & p_{i2}(s,u) & p_{ik}(s,u) & \ldots \\ \vdots & \vdots & \vdots & \end{bmatrix} \begin{bmatrix} p_{11}(u,t) & p_{12}(u,t)\ldots & p_{1j}(u,t) & \ldots \\ p_{21}(u,t) & p_{22}(u,t) & p_{2j}(u,t) & \ldots \\ & & & \\ p_{i1}(u,t) & p_{i2}(u,t) & p_{kj}(u,t) & \ldots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

$$H(s,t) = H(s,u)H(u,t)$$

Equazioni di C-K in forma matriciale

$$H(s,t) = \begin{bmatrix} p_{11}(s,t) & p_{12}(s,t)\ldots & p_{1k}(s,t) & \ldots \\ p_{21}(s,t) & p_{22}(s,t) & p_{2k}(s,t) & \ldots \\ & & & \\ p_{i1}(s,t) & p_{i2}(s,t) & p_{ik}(s,t) & \ldots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

Catene di Markov

In forma matriciale:

$H(s,t)=[p_{ij}(s,t)]=H(s,u)H(u,t)$, con $H(t,t)=I$

definendo $P(t) = H(t, t+\Delta t)=[p_{ij}(t,t+\Delta t)]$ si ottiene:

$$H(s,t)-H(s,t-\Delta t)=H(s,t-\Delta t)H(t-\Delta t,t)-H(s,t-\Delta t)=$$
$$=H(s,t-\Delta t)P(t-\Delta t)-H(s,t-\Delta t)=H(s,t-\Delta t)(P(t-\Delta t)-I)$$

dividendo per Δt e facendo il limite tendente a zero:

$$\lim_{\Delta t\to 0}\frac{H(s,t)-H(s,t-\Delta t)}{\Delta t}=$$

$$\frac{\partial H(s,t)}{\partial t}=\lim_{\Delta t\to 0}\left(H(s,t-\Delta t)\frac{(P(t-\Delta t)-I)}{\Delta t}\right)=H(s,t)Q(t)$$

N.B.

$$Q(t)=\lim_{\Delta t\to 0}\frac{P(t-\Delta t)-I}{\Delta t}=[q_{ij}(t)],\ \text{con}\ q_{ij}(t)=\begin{cases}\lim_{\Delta t\to 0}\dfrac{p_{ii}(t-\Delta t,t)-1}{\Delta t} & i=j\\ \lim_{\Delta t\to 0}\dfrac{p_{ij}(t-\Delta t,t)}{\Delta t} & i\neq j\end{cases}$$

Catene di Markov

$$\frac{\partial H(s,t)}{\partial t} = H(s,t)Q(t)$$

Equazioni di C-K in forma differenziale

$$Q(t) = [q_{ij}(t)], \quad q_{ij}(t) = \begin{cases} \lim_{\Delta t \to 0} \frac{p_{ii}(t-\Delta t, t) - 1}{\Delta t} & \text{i} = \text{j} \\ \lim_{\Delta t \to 0} \frac{p_{ij}(t-\Delta t, t)}{\Delta t} & \text{i} \neq \text{j} \end{cases}$$

Q è nota come generatore infinitesimale or rate matrix del processo di Markov.

È evidente che

$$\sum_j q_{ij}(t) = 0 \quad \Rightarrow q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t) < 0$$

Catene di Markov

Pertanto:

$$Q(t) = \begin{bmatrix} -\sum_{j \neq 1} q_{1j}(t) & q_{12}(t) & \dots & q_{1N}(t) \\ q_{21}(t) & -\sum_{j \neq 2} q_{2j}(t) & \dots & q_{2N}(s,t) \\ \vdots & \vdots & & \vdots \\ q_{N1}(t) & q_{N2}(t) & \dots & -\sum_{j \neq N} q_{Nj}(t) \end{bmatrix}$$

Perché la matrice *Q(t)* è detta "rate" matrix?

Perché il valore degli elementi q_{ij} sono in relazione con la frequenza degli eventi che determinano l'evoluzione del processo di Markov process.

Catene di Markov

$$q_{ij}(t) = \begin{cases} \lim_{\Delta t \to 0} \dfrac{p_{ii}(t-\Delta t, t) - 1}{\Delta t} \\ \lim_{\Delta t \to 0} \dfrac{p_{ij}(t-\Delta t, t)}{\Delta t} \end{cases} = \begin{cases} \lim_{\Delta t \to 0} \dfrac{p_{ii}(t-\Delta t, t) - p_{ii}(t,t)}{\Delta t} & i = j \\ \lim_{\Delta t \to 0} \dfrac{p_{ij}(t-\Delta t, t) - p_{ij}(t,t)}{\Delta t} & i \neq j \end{cases}$$

$p_{ii}(t,t) = 1$

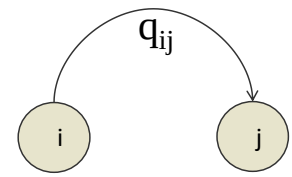
Quindi:

$$p_{ij}(t, t+\Delta t) \cong |q_{ij}(t)| \Delta t$$

$$1 - p_{ij}(0, t+\Delta t) = p_0(0, t+\Delta t) = p_0(0,t)(1 - p_{ij}(t, t+\Delta t)) = p_0(0,t)(1 - |q_{ij}(t)| \Delta t)$$

$$\lim_{\Delta t \to 0} \frac{p_0(0, t+\Delta t) - p_0(0,t)}{\Delta t} = \frac{dp_0(0,t)}{dt} = -|q_{ij}(t)| p_0(0,t)$$

$$p_0(0,t) = e^{-\int_0^t |q_{ij}(t)| dt} \qquad p_{ij}(0,t) = 1 - e^{-\int_0^t |q_{ij}(t)| dt} \qquad t \geq 0$$



if $q_{ij}(t) = q_{ij}$ $p_{ij}(0,t) = 1 - e^{-|q_{ij}|t}$, $\overline{T}_{ij} = \dfrac{1}{|q_{ij}|}$

se costanti, la distribuzione di prob di transizione di stato è un'exp

q_{ij} risulta essere uguale alla frequenza media di transizione di stato

Catene di Markov

Equazioni di Chapman-Kolmogorov:

- forma matriciale
  - in avanti $$\frac{\partial H(s,t)}{\partial t} = H(s,t)Q(t) \qquad s \le t$$
  - all'indietro $$\frac{\partial H(s,t)}{\partial s} = -Q(s)H(s,t) \qquad s \le t$$
- termine a termine
  - in avanti
  $$\frac{\partial p_{ij}(s,t)}{\partial t} = q_{jj}(t)p_{ij}(s,t) + \sum_{k \neq j} q_{kj}(t)p_{ik}(s,t) \quad con \quad p_{ij}(s,s) = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}$$
  - all'indietro
  $$\frac{\partial p_{ij}(s,t)}{\partial s} = -q_{ii}(t)p_{ij}(s,t) - \sum_{k \neq j} q_{ik}(s)p_{kj}(s,t) \quad con \quad p_{ij}(t,t) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Catene di Markov

Soluzione alle due equazioni:

$$H(s,t) = e^{\int_s^t Q(u)du}$$

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots$$

La probabilità di fare una transizione da $i \to j$ nell'intervallo $(t, t+\Delta T)$ è data da:

$$p_{ij}(t, t+\Delta t) = q_{ij}(t)\Delta t + o(\Delta t)$$

$$o(\Delta t) \Leftrightarrow \lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

La probabilità di uscire dallo stato *i-esimo* nell'intervallo $(t, t+\Delta T)$ è data da:

$$1 - p_{ii}(t, t+\Delta t) = \Delta t \sum_{j \neq i} q_{ij}(t) + o(\Delta t) = -q_{ii}(t)\Delta t + o(\Delta t)$$

credo abbia ripreso questa di slide 28

$$p_{ij}(t, t+\Delta t) \cong \left| q_{ij}(t) \right| \Delta t$$

Catene di Markov

Probabilità dello stato *j*: $\pi_j(t) = P[X(t)=j]$

È immediato scrivere che per t≥s

$$\pi_j(t) = \sum_i \pi_i(s)\, p_{ij}(s,t) = \underbrace{\begin{bmatrix}\pi_1(s) & \pi_2(s) & \pi_i(s) & \dots\end{bmatrix}}_{\pi^T(s)} \begin{bmatrix} p_{1j}(s,t) \\ p_{2j}(s,t) \\ \\ p_{ij}(s,t) \\ \vdots \end{bmatrix}$$

$$\pi^T(t) = \pi^T(s) H(s,t)$$

$$\frac{d\pi^T(t)}{dt} = \pi^T(s)\frac{\partial H(s,t)}{\partial t} = \pi^T(s) H(s,t) Q(t) = \pi^T(t) Q(t)$$

Forward Chapman-Kolmogorov equations

$$\frac{d\pi(s)}{ds} = Q(s)\pi(s)$$ Backward Chapman-Kolmogorov equations

Catene di Markov Omogenee

Una catena di Markov tempo-continua è detta omogenea se le probabilità di transizione fra due stati qualsiasi, in un dato intervallo di tempo, non dipende dall'istante di inizio di tale intervallo ma solo dalla sua durata:

Per catene di Markov omogenee

$$p_{ij}(s,t) = p_{ij}(\tau) \Rightarrow H(s,t) = H(\tau) = [p_{ij}(\tau)], \quad \tau = t - s$$
$$q_{ij}(t) = q_{ij} \; i,j = 1,2,\ldots \Rightarrow Q(t) = Q = [q_{ij}]$$

Le equazioni di Chapman-Kolmogorov diventano

$$\frac{d\pi^T(t)}{dt} = \pi^T(t)Q$$

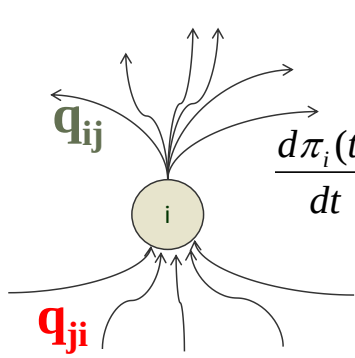
E' intuitivo il fatto che se la frequenza di transizione fra due stati in un certo intervallo di tempo non dipende dall'istante considerato, allora la frequenza di transizione di stato si mantiene costante nel tempo.

prima nota: versione prof; seconda nota: rimessa a posto da chatgpt

Catene di Markov Omogenee

Le equazioni di Chapman-Kolmogorov si possono scrivere facilmente mediante ispezione diretta:

Flusso di probabilità verso altri stati



$$\frac{d\pi_i(t)}{dt} = \pi^T(t)Q = \pi_i(t)\sum_{j\neq i} q_{ij} + \sum_{j\neq i}\pi_j(t)q_{ji}$$

Flusso di probabilità dagli altri stati

33

Ricorda come era fatta Q (togli la dipendenza dal tempo bc siamo in caso di OMOGENEITA')

$$Q(t) = \begin{bmatrix} -\sum_{j\neq 1} q_{1j}(t) & q_{12}(t) & \dots & q_{1N}(t) \\ q_{21}(t) & -\sum_{j\neq 2} q_{2j}(t) & \dots & q_{2N}(s,t) \\ \vdots & \vdots & & \vdots \\ q_{N1}(t) & q_{N2}(t) & \dots & -\sum_{j\neq N} q_{Nj}(t) \end{bmatrix}$$

Catene di Markov Omogenee

Catena di Markov irriducibile:

se
- ogni stato è raggiungibile da qualsiasi altro: $p_{ij}(t)>0$

$$\lim_{t\to\infty} p_{ij}(t) = p_j$$

Catena di Markov ergodica:
- le probabilità di stato convergono ad un valore limite, indipendente dalla distribuzione iniziale degli stati

$$\lim_{t\to\infty} \pi_j(t) = \pi_j = p_j$$

L'irriducibilità garantisce che tutti gli stati debbano avere probabilità positiva di essere visitati in qualche istante, l'ergodicità garantisce che il comportamento statistico della catena non dipende dallo stato di partenza. In questo modo il comportamento statistico di tutte le realizzazioni del processo di Markov è lo stesso, pertanto le statistiche temporali coincidono con quelle d'insieme.

Soluzione delle equazioni di Chapman Kolmogorov per catene omogenee.

$$\frac{d\pi(t)}{dt} = \pi(t)A$$ A matrice generica

Caso monidimensionale, A = a, scalare: $\pi(t) = \pi_0 e^{at}$

$$\frac{de^{at}}{dt} = e^{at}a = ae^{at}$$

$$e^{at} = 1 + at + \tfrac{1}{2!}(at)^2 + \cdots \tfrac{1}{k!}(at)^k + \cdots = \sum_{k=0}^{\infty} \frac{(at)^k}{k!}$$

$$k! = k \times (k-1) \times \cdots \times 2 \times 1$$

$$\frac{de^{at}}{dt} = \frac{d(1 + at + \tfrac{1}{2}(at)^2 + \cdots \tfrac{1}{k!}(at)^k + \cdots)}{dt}$$

$$= \frac{0 + a + \tfrac{2}{2}(at)a + \cdots \tfrac{k}{k!}(at)^{k-1}a + \cdots)}{dt} = ae^{at}$$

Soluzione delle equazioni di Chapman Kolmogorov per catene omogenee.

Caso multidimensionale: $\pi(t) = \pi_0 e^{At}$

L'esponenziale di matrice e^{At} è definito come:

$$e^{At} = I + At + \tfrac{1}{2!}A^2t^2 + \cdots \tfrac{1}{k!}A^kt^k + \cdots = \sum_{k=0}^{\infty} \frac{A^kt^k}{k!}$$

$$\frac{de^{At}}{dt} = \frac{de^{At}}{d(At)}\frac{d(At)}{dt} = Ae^{At} = e^{At}A$$

Soluzione delle equazioni di Chapman Kolmogorov per catene omogenee.

$$\frac{d}{dt}e^{At} = \frac{d}{dt}(I + At + \tfrac{1}{2!}A^2t^2 + \cdots \tfrac{1}{k!}A^kt^k + \cdots)$$

$$= 0 + A + \tfrac{1}{2}2tA^2 + \cdots + \tfrac{1}{k!}kt^{k-1}A^k + \cdots$$

$$= \mathrm{A} + \mathrm{A}^2t + \cdots + \tfrac{1}{(k-1)!}A^kt^{k-1} + \cdots$$

$$= \mathrm{A}(\mathrm{I} + \mathrm{At} + \cdots \tfrac{1}{(\mathrm{k\text{-}1})!}A^{k-1}t^{k-1} + \cdots)$$

$$= A\sum_{k=0}^{\infty}\frac{A^kt^k}{k!} = Ae^{At}$$

$$= (\mathrm{I} + \mathrm{At} + \cdots \tfrac{1}{(\mathrm{k\text{-}1})!}A^{k-1}t^{k-1} + \cdots)A$$

$$= (\sum_{k=0}^{\infty}\frac{A^kt^k}{k!})A = e^{At}A$$

Soluzione delle equazioni di Chapman Kolmogorov per catene omogenee.

$$\frac{d\pi(t)}{dt} = \pi(t)A \Rightarrow \pi(t) = \pi_0 e^{At}$$

$$s\pi(s) - \pi_0 = \pi(s)A$$

$$\pi(s) = \pi_0 (sI - A)^{-1}$$

$$\pi(t) = \pi_0 L^{-1}[(sI - A)^{-1}]$$

$$e^{At} = L^{-1}[(sI - A)^{-1}]$$

$$(sI - A)^{-1} = L[e^{At}] = L[I + At + \cdots + \frac{1}{k!}A^k t^k + \cdots]$$

$$= \frac{I}{s} + \frac{A}{s^2} + \frac{A^2}{s^3} + \cdots + \frac{A^k}{s^{k+1}} + \cdots = \sum_{k=0}^{\infty} \frac{A^k}{s^{k+1}}$$

Proprietà dell'esponenziale di matrice

"rivediamole"

$$e^{At}\,|_{t=0} = I \qquad [e^{At}]^{-1} = e^{-At}$$

$$e^{A(t_1+t_2)} = e^{At_1}e^{At_2} = e^{At_2}e^{At_1}$$

$$e^{A(t-t)} = e^{At}e^{-At} = e^{-At}e^{At} = I$$

$$e^{At}e^{Bt} = e^{(A+B)t}$$ solo se A e B commutano

Esempio

Si consideri la matrice sottostante.

$$\dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x \qquad \mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

Si ha che

$$\mathbf{A}^2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2^2 \end{pmatrix}, \quad \mathbf{A}^3 = \begin{pmatrix} 1 & 0 \\ 0 & 2^2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2^3 \end{pmatrix}, \ldots$$

In generale,

$$\mathbf{A}^n = \begin{pmatrix} 1 & 0 \\ 0 & 2^n \end{pmatrix}$$

Quindi

$$e^{\mathbf{A}t} = \sum_{n=0}^{\infty} \frac{\mathbf{A}^n t^n}{n!} = \sum_{n=0}^{\infty} \begin{pmatrix} 1/n! & 0 \\ 0 & 2^n/n! \end{pmatrix} t^n = \begin{pmatrix} e^t & 0 \\ 0 & e^{2t} \end{pmatrix}$$

Esempio

$$\dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x \qquad [sI - A] = \begin{bmatrix} s-1 & 0 \\ 0 & s-2 \end{bmatrix}$$

$$[sI - A]^{-1} = \begin{bmatrix} \dfrac{s-2}{(s-1)(s-2)} & 0 \\ 0 & \dfrac{s-1}{(s-1)(s-2)} \end{bmatrix}$$

$$\Phi(t) = L^{-1}\left\{[sI - A]^{-1}\right\} = \begin{bmatrix} e^{t} & 0 \\ 0 & e^{2t} \end{bmatrix}$$

Catene di Markov Omogenee in Equilibrio

Le probabilità limite (ossia in condizione di equilibrio statistico) sono le soluzioni del sistema lineare:

N.B.

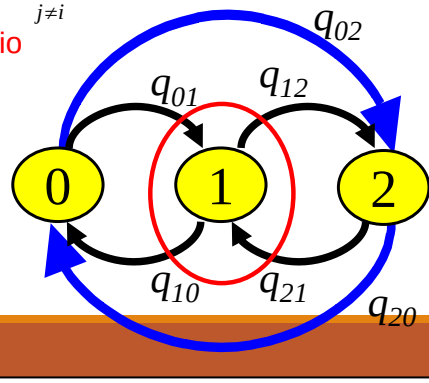
$$\begin{cases} q_{jj}\pi_j + \sum_{k\neq j} q_{kj}\pi_k = 0 \\ \sum_j \pi_j = 1 \end{cases}$$

oppure

$$\begin{cases} \pi Q = 0 \\ \sum_j \pi_j = 1 \end{cases}$$

$$q_{ii}(t) = -\sum_{j\neq i} q_{ij}(t) < 0$$

Esempio



$$\begin{cases} q_{00}\pi_0 + q_{10}\pi_1 + q_{20}\pi_2 = 0 \\ q_{11}\pi_1 + q_{01}\pi_0 + q_{21}\pi_2 = 0 \\ q_{22}\pi_2 + q_{02}\pi_0 + q_{12}\pi_1 = 0 \\ \pi_0 + \pi_1 + \pi_2 = 1 \end{cases}$$

$$\begin{cases} q_{00} = -q_{01} - q_{02} \\ q_{11} = -q_{10} - q_{12} \\ q_{22} = -q_{20} - q_{21} \end{cases}$$

Processi di nascita e morte

Catene di Markov in cui sono permesse, da un generico stato *j*, soltanto transizioni verso gli stati *j+1* (nascita) e *j-1* (morte), definendo

$$P_k(t)=p_k(t)=P[X(t)=k]$$

Consideriamo il caso di una catena di Markov omogenea; definiamo:

$$\begin{cases} \lambda_k = q_{k,k+1} \\ \mu_k = q_{k,k-1} \\ q_{k,k} = -(\lambda_k + \mu_k) \text{ da } \sum_j q_{kj} = 0 \\ q_{k,j} = 0 \text{ per } |k-j| > 1 \end{cases} \quad Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 \dots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 \dots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 \dots \\ \vdots & & & \end{bmatrix}$$

Processi di nascita e morte

La distribuzione di probabilità degli stati si ottiene risolvendo le seguenti equazioni:

$$\frac{d\pi^T(t)}{dt} = \pi^T(t)Q$$

$$\begin{cases} \sum_{k=0}^{\infty} P_k(t) = 1 \\ \frac{dP_k(t)}{dt} = -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) \\ \frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \\ Condizioni\ iniziali\ \ P_k(0), k = 0,1,... \end{cases}$$

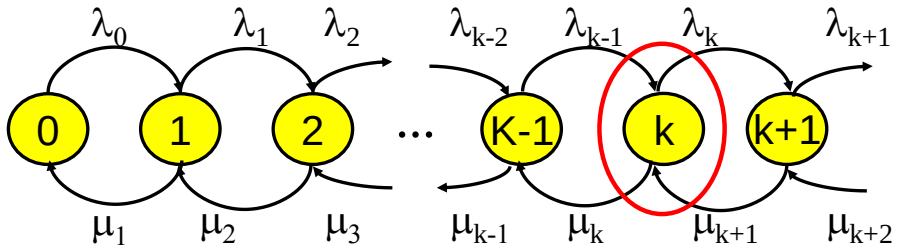
Processi di nascita e morte

… che si possono vedere anche per ispezione visiva, facendo il bilancio dei flussi di probabilità:

- Flusso entrante nello stato k: $\lambda_{k-1}P_{k-1}(t)+\mu_{k+1}P_{k+1}(t)$
- Flusso uscente dallo stato k: $-(\lambda_k+\mu_k)P_k(t)$

La differenza tra queste due quantità rappresenta il tasso di variazione di probabilità dello stato k:

$$\frac{dP_k(t)}{dt} = -(\lambda_k+\mu_k)P_k(t)+\lambda_{k-1}P_{k-1}(t)+\mu_{k+1}P_{k+1}(t)$$

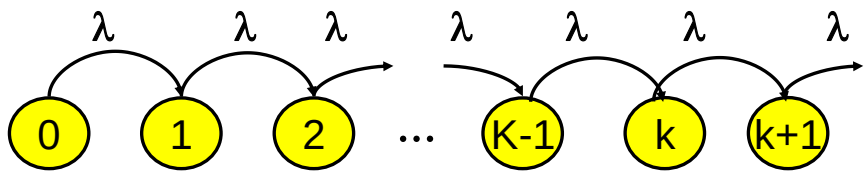


N.B.

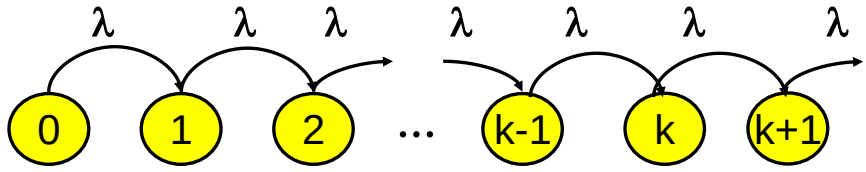
quindi questa è la formula generale della derivata rispetto al tempo dello stato k-esimo

Processi di Poisson

Caratteristiche: catena di Markov di pura nascita a tasso costante (λ)



Processi di Poisson



$$\begin{cases} \dfrac{dP_0(t)}{dt} = -\lambda P_0(t) \\ \dfrac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t), \qquad k \geq 1 \end{cases}$$

$P_0(0) = 1$ Condizioni iniziali

$$P_0(t) = e^{-\lambda t} \rightarrow \frac{dP_1(t)}{dt} = -\lambda P_1(t) + \lambda e^{-\lambda t} \qquad t \geq 0$$

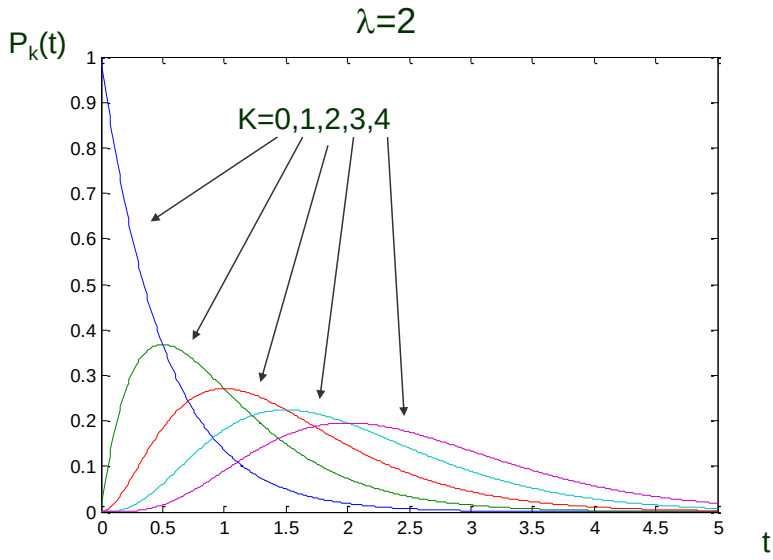
$$P_1(t) = \lambda t e^{-\lambda t} \rightarrow \frac{dP_2(t)}{dt} = -\lambda P_2(t) + \lambda^2 t e^{-\lambda t} \ldots$$

$$P_2(t) = \frac{(\lambda t)^2}{2} e^{-\lambda t} \rightarrow \ldots$$

$P_k(t) = \dfrac{(\lambda t)^k}{k!} e^{-\lambda t}$ $t \geq 0$ Probabilità di k arrivi in (0-t)

Avendo ipotizzato che P_0 (t)=1, ossia si cosiderino zero arrivi al tempo zero, P_k (t) rappresenta la probabilità che ci siano un numero k di arrivi nell'intevallo di tempo che va da 0 a t.

Processi di Poisson



Processi di Poisson

$$E[k]=\sum_{k=0}^{\infty} k\frac{(\lambda t)^k}{k!}e^{-\lambda t}=e^{-\lambda t}\sum_{k=0}^{\infty} k\frac{(\lambda t)^k}{k!}=e^{-\lambda t}\sum_{k=1}^{\infty}\frac{(\lambda t)^k}{(k-1)!}=$$

$$e^{-\lambda t}\lambda t\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}=e^{-\lambda t}\lambda t\sum_{j=0}^{\infty}\frac{(\lambda t)^j}{j!}=\lambda t$$

$$\sigma_k^2=E\left[(k-E[k])^2\right]=E\left[k^2-2kE[k]+E[k]^2\right]=*$$

$$=E\left[k^2+k-k-2kE[k]+E[k]^2\right]=E[k(k-1)]+E[k]-(E[k])^2$$

$$E[k(k-1)]=\sum_{k=0}^{\infty}k(k-1)\frac{(\lambda t)^k}{k!}e^{-\lambda t}=e^{-\lambda t}(\lambda t)^2\sum_{k=2}^{\infty}\frac{(\lambda t)^{k-2}}{(k-2)!}=(\lambda t)^2$$

$$\sigma_k^2=(\lambda t)^2+\lambda t-(\lambda t)^2=\lambda t$$

Valor medio = varianza

$*$

$$=E[K^2]-E[2KE[K]]+E[E[K]^2]=$$

$$=E[K^2]-2E[K]E[K]+E[K]^2=$$

SINCE $E[K]=\text{const}$

$$=E[K^2]-2E[K]^2+E[K]^2=$$

$$=E[K^2]-E[K]^2=E[K^2-K+K]-E[K]^2=$$

$$=E[K(K-1)]+E[K]-(E[K])^2$$

Processi di Poisson

statistica dei tempi di interarrivo

$$P\left(x(s,s+t)=k\right)=P_k(t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad k\geq 0,\ t\geq 0,\ \forall s$$

$$A(t)=p\left(\tilde{t}\leq t\right)=1-p\left(\tilde{t}>t\right)=1-P_0(t)$$

$$A(t)=1-e^{-\lambda t} \qquad t\geq 0$$

$$a(t)=\lambda e^{-\lambda t} \qquad t\geq 0$$

$$E(t)=\frac{1}{\lambda};\quad E(t^2)=\frac{2}{\lambda^2};\quad \sigma_t^2=\frac{1}{\lambda^2}$$

$$A(s)=\int_0^\infty \lambda e^{-\lambda t}e^{-st}dt=\frac{\lambda}{s+\lambda}$$

Proprietà "**memoryless**" della distribuzione esponenziale

se il tempo di permanenza in uno stato è una v.a. esponenziale, il tempo trascorso nello stato non è utilizzabile per predire quanto si resterà ancora nello stato stesso.

Processi di Poisson

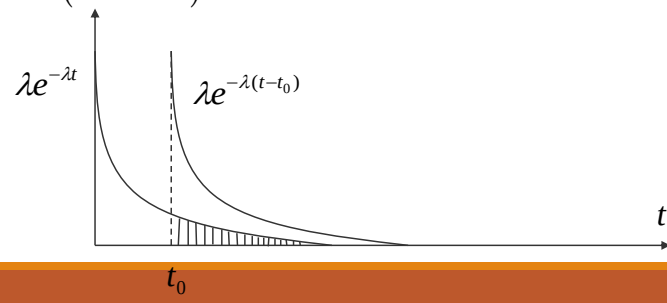
Proprietà "**memoryless**" della distribuzione esponenziale

Supponiamo di fissare un istante t=0 di riferimento in corrispondenza di un arrivo. Se al tempo t_0 non vi è stato nessun altro arrivo, ci chiediamo quale sia la probabilità che il prossimo arrivi si verifichi dopo t a partire da t_0.

$$P\left[\tilde{t} \le t+t_0 \middle| \tilde{t} > t_0\right] = \frac{P\left[t_0 < \tilde{t} \le t+t_0\right]}{P\left[\tilde{t} > t_0\right]} = \frac{P\left[\tilde{t} \le t+t_0\right] - P\left[\tilde{t} \le t_0\right]}{P\left[\tilde{t} > t_0\right]} =$$

$$= \frac{A(t+t_0) - A(t_0)}{1-A(t_0)} = \frac{1-e^{-\lambda(t+t_0)} - \left(1-e^{-\lambda(t_0)}\right)}{1-\left(1-e^{-\lambda(t_0)}\right)} = \frac{-e^{-\lambda(t+t_0)} + e^{-\lambda(t_0)}}{e^{-\lambda(t_0)}} =$$

$$= 1-e^{-\lambda(t)} = A(t)$$

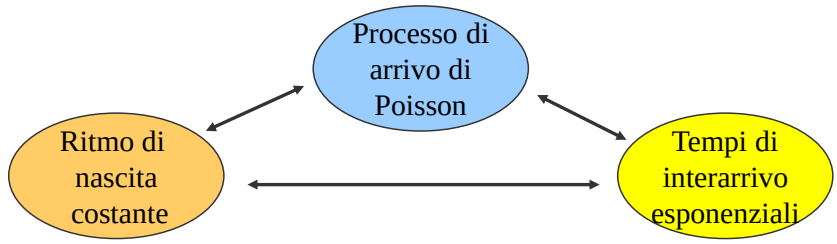


Processi di Poisson

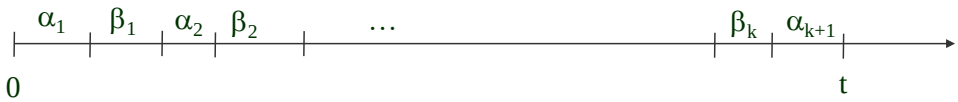
Caratteristiche: catena di Markov di pura nascita a ritmo costante (λ)

- $\lambda_k=\lambda$ $k=0,1,2,\ldots$
- $\mu_k=0$
- $E[K]=\lambda t$, $\sigma_K^2=\lambda t$

$$P_k(t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad k\geq 0,\ t\geq 0$$



Processi di Poisson



Ci siano k arrivi in [0-t] (evento condizionante B_k). Suddividiamo l'intervallo 0-t in intervalli generici di tipo α_i, senza arrivi, e in intervalli di tipo β_i, con un singolo arrivo (evento A_k)

$$P_1(\beta_i) = \lambda \beta_i e^{-\lambda \beta_i}$$

$$P_0(\alpha_i) = e^{-\lambda \alpha_i}$$

$$P(A_k | B_k) = \frac{e^{-\lambda\alpha_1}\lambda\beta_1 e^{-\lambda\beta_1} e^{-\lambda\alpha_2}\lambda\beta_2 e^{-\lambda\beta_2} \ldots e^{-\lambda\alpha_k}\lambda\beta_k e^{-\lambda\beta_k} e^{-\lambda\alpha_{k+1}}}{\frac{(\lambda t)^k}{k!} e^{-\lambda t}} =$$

$$= \frac{\lambda^k (\beta_1\beta_2 \ldots \beta_k) e^{-\lambda t}}{\frac{(\lambda t)^k}{k!} e^{-\lambda t}} = \frac{(\beta_1\beta_2 \ldots \beta_k) k!}{t^k}$$

Processi di Poisson

A questo punti si immagini di distribuire aleatoriamente k punti nell'intervallo 0-t con distribuzione uniforme. E' semplice ricavare che

$$P(A_k|B_k) = \left(\frac{\beta_1}{t}\frac{\beta_2}{t}\frac{\beta_3}{t}\ldots\frac{\beta_k}{t}\right)k!$$

Siccome le due probabilità coincidono, ne consegue che:

dati k arrivi in 0-t, se questo sono generati da un processo di Poisson, allora questi sono distribuiti uniformemente nell'intervallo 0-t.

Super NB.

Processi di Poisson

Si supponga di accumulare k arrivi di un processo di Poisson. A tal fine sarà necessario un tempo pari alla somma di k v.a. esponenziali, la cui funzione caratteristica sarà pari a:

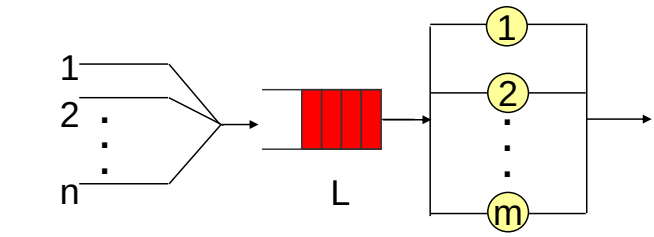
$$\Pi(s) = \left(\frac{\lambda}{s + \lambda} \right)^k$$

la cui anti-trasformata è pari a

$$f_X(x) \frac{\lambda(\lambda x)^{k-1}}{(k-1)!} e^{-\lambda x} \qquad x \geq 0$$

DISTRIBUZIONE DI ERLANG (k)

Sistemi di servizio



sorgenti di traffico *fila d'attesa* *servizio*

- Il sistema è descritto attraverso variabili aleatorie:
- k = numero di utenti nel sistema
- l = numero di utenti nella sola fila d'attesa
- h = numero di serventi contemporaneamente occupati
- x = tempo di servizio
- s = tempo di permanenza nel sistema (tempo di coda o di ritardo)
- w = tempo di permanenza nella fila d'attesa

Sistemi di servizio

La variabile aleatoria k è caratterizzata attraverso la sua probabilità limite

$\pi_k = p_k$=probabilità che in un generico istante di osservazione **in regime permanente** siano presenti k utenti (richieste di servizio) all'interno del sistema

Parametri prestazionali

- Probabilità di sistema bloccato (*m* serventi)

$$S_p = \Pr\{k = L + m\} = p_{L+m}$$

- Probabilità di rifiuto
  - Data una richiesta di servizio offerto (r.s.o.)

$$\Pi_p = \Pr\{\text{sistema bloccato/r.s.o.}\} = S_p \frac{\Pr\{\text{r.s.o./sistema bloccato}\}}{\Pr\{\text{r.s.o.}\}} = S_p$$

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

Parametri prestazionali

- Probabilità di servizio bloccato (*m* serventi)

$$S_r = \Pr\{k \geq m\}$$

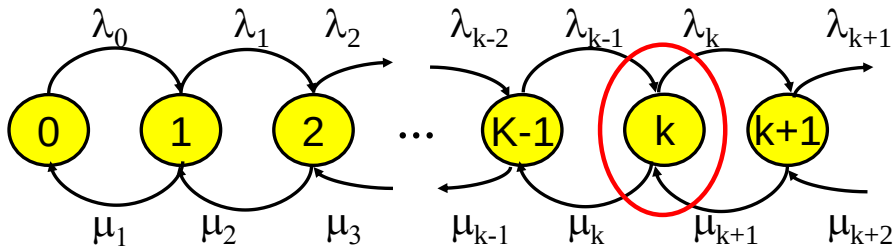
- Probabilità di ritardo
  - Data una richiesta di servizio accolta (r.s.a.)

$$\Pi_r = \Pr\{\text{servizio bloccato/r.s.a.}\} = S_r \frac{\Pr\{\text{r.s.a./servizio bloccato}\}}{\Pr\{\text{r.s.a.}\}} = S_r$$

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

Processi di nascita e morte in equilibrio statistico

$$\begin{cases} \sum_{k=0}^{\infty} P_k = 1 \\ 0 = -(\lambda_k + \mu_k)\mathrm{P_k} + \lambda_{k-1}\mathrm{P_{k-1}} + \mu_{k+1}\mathrm{P_{k+1}} \\ 0 = -\lambda_0\mathrm{P_0} + \mu_1\mathrm{P_1} \end{cases}$$



Processi di nascita e morte in equilibrio statistico

$$\mu_{k+1}P_{k+1} - \lambda_k P_k = \mu_k P_k - \lambda_{k-1}P_{k-1}$$

sia $\alpha_k = \mu_k P_k - \lambda_{k-1}P_{k-1}$

risulta

$$\alpha_k = \text{costante} = \mu_1 P_1 - \lambda_0 P_0 = 0, \qquad \text{quindi}$$

$$P_k = \frac{\lambda_{k-1}}{\mu_k} P_{k-1} \quad \Rightarrow \quad P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}$$

siccome

$$P_0 + P_0 \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = 1 \quad \Rightarrow \quad P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

Sistema a coda M/M/1/∞/∞

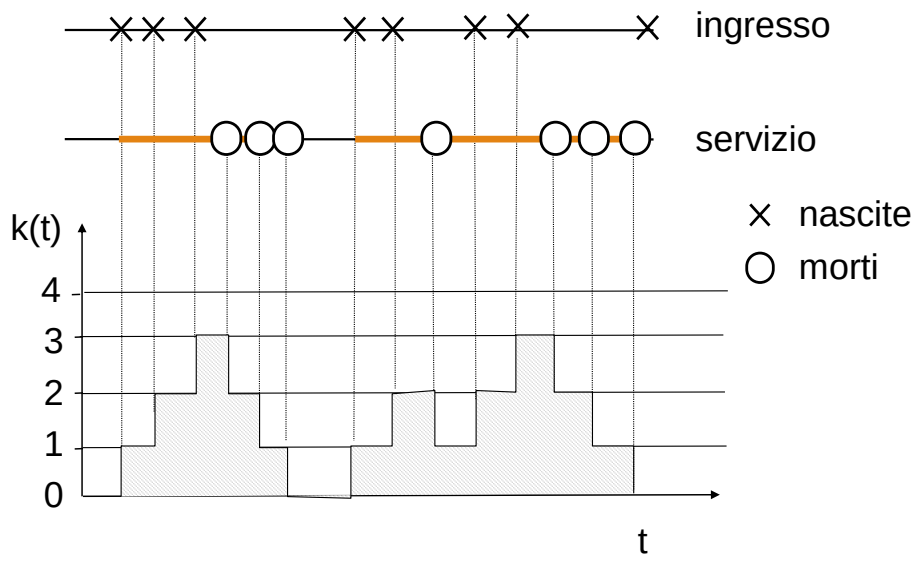
Ipotesi:

- tempi di interarrivo i.i.d. con distribuzione esponenziale negativa di parametro λ (ingresso di Poisson);
- tempi di servizio i.i.d. con distribuzione esponenziale negativa di parametro μ;
- processi di arrivo e di servizio statisticamente indipendenti.
- singolo servente;
- spazio infinito per la fila di attesa.

Il processo di coda K(t) è descrivibile mediante un processo di Markov di nascita e morte con spazio di stato {0,1,…}

Il processo di coda K(t) è ergodico se $\lambda/\mu<1$

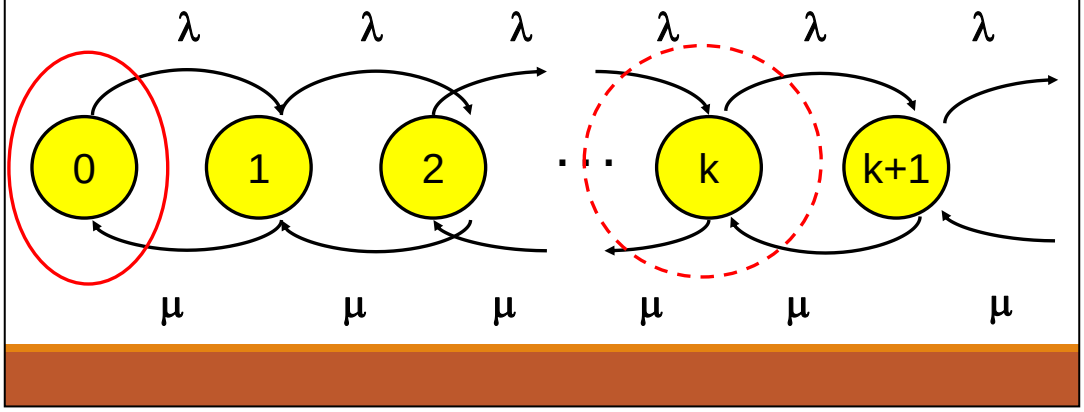
Evoluzione temporale



Frequenze di transizione di stato

$\lambda_k = \lambda$ per k ≥0 frequenza di nascita

$\mu_k = \mu$ per k ≥1 frequenza di morte



Probabilità limite di stato

Sostituendo λ e μ nella soluzione generale:

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{dove} \quad P_0 = \frac{1}{1+\sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

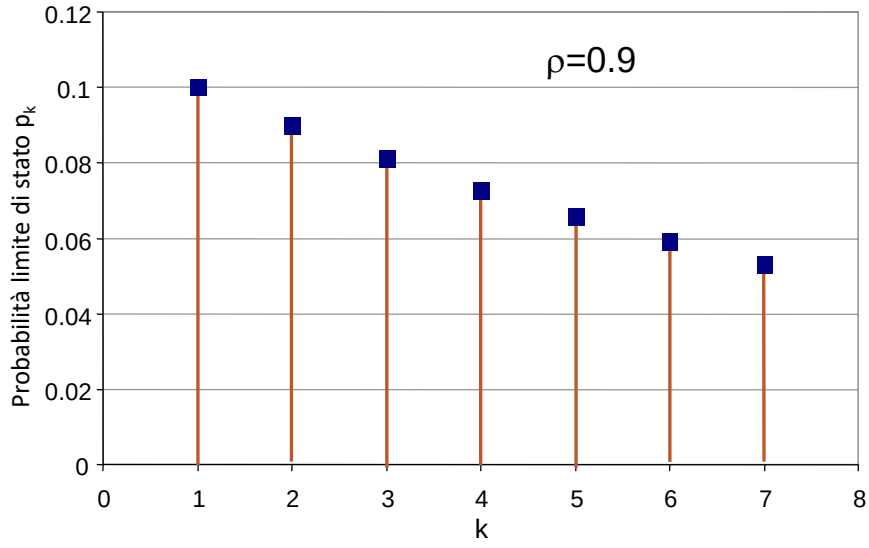
$$P_0 = \frac{1}{1+\sum_{k=1}^{\infty} \rho^k} = 1-\rho$$

dove $\rho = \lambda/\mu$, $\rho < 1$

$$P_k = (1-\rho)\rho^k$$

k=0,1,2, … (distribuzione geometrica)

Probabilità limite di stato



La distribuzione è di tipo geometrico con parametro ρ

Probabilità limite di stato

Il numero medio di utenti nel sistema è

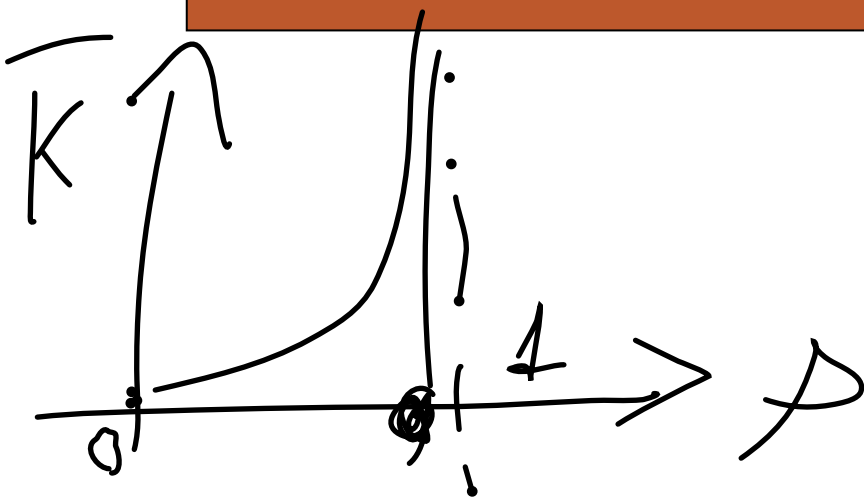
$$E[K] = \bar{k} = \sum_{k=0}^{\infty} k \cdot p_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k = (1-\rho)\sum_{k=0}^{\infty} k\rho^k =$$

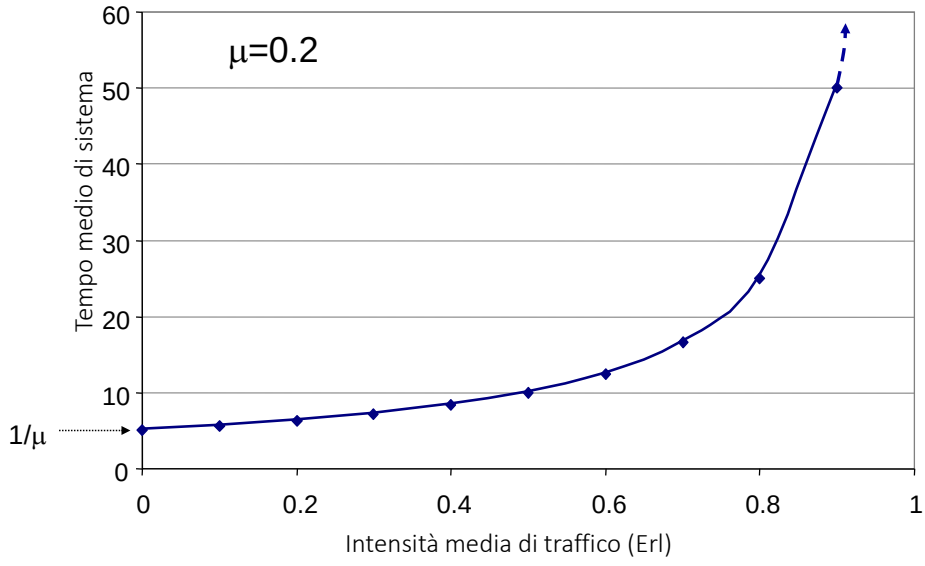
$$= (1-\rho)\rho\sum_{k=0}^{\infty} k\rho^{k-1} = (1-\rho)\rho\sum_{k=1}^{\infty} k\rho^{k-1} = (1-\rho)\rho\left(\frac{\partial}{\partial\rho}\sum_{k=0}^{\infty}\rho^k\right) =$$

$$= (1-\rho)\rho\left(\frac{\partial}{\partial\rho}\frac{1}{1-\rho}\right) = \frac{\rho}{1-\rho}$$

Il tempo di permanenza medio nel sistema è (Legge di Little)

$$T = \frac{\bar{k}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1/\mu}{(1-\rho)} = \frac{1}{\mu-\lambda}$$





Al crescere dell'intensità di traffico il tempo di coda tende all'infinito

Parametri prestazionali

In condizioni di equilibrio statistico l'intensità media di richieste smaltite A_s coincide con l'intensità di richieste di servizio offerte A_o

$$A_s = A_o = \rho = 1 - p_0$$

La probabilità di servizio bloccato S_r coincide con la probabilità di ritardo nel ricevere servizio Π_r

$$S_r = \Pi_r = (1-\rho)\sum_{k=1}^{\infty} \rho^k = \rho$$

ρ = prob. che il servente sia occupato = la percentuale temporale di occupazione del servente = la prob. che una richiesta in arrivo sia costretta ad attendere in coda

Distribuzioni in equilibrio statistico

l= lunghezza della fila d'attesa=numero di utenti nella fila d'attesa

$$Pr\{l=j\}=\begin{cases}(1-\rho)+\rho(1-\rho)=1-\rho^2 & j=0\\ (1-\rho)\cdot\rho^{j+1} & j\geq 1\end{cases}$$

$$\bar{l}=\frac{\rho^2}{1-\rho}$$

h=numero di serventi impegnati

$$Pr\{h=j\}=\begin{cases}1-\rho & j=0\\ \rho & j=1\end{cases}$$

$$\bar{h}=\rho$$

il numero medio di utenti all'interno del sistema è quindi

$$\bar{k}=\bar{l}+\bar{h}=\frac{\rho}{1-\rho}$$

Tempi di attesa in coda

Si supponga che un utente trovi, al suo arrivo, il sistema nello stato k, ossia vi sono altri k utenti presenti nel sistema (uno in servizio e k-1 nella fila di attesa). Nel caso di disciplina FIFO, l'utente, prima di essere servito dovrà attendere un tempo pari alla somma di k v.a. esponenziali. Questo avverrà con probabilità $\rho^K(1-\rho)$. In media si avrà che:

$$W(s)=\sum_{k=0}^{\infty}\left(\frac{\mu}{s+\mu}\right)^k\rho^k(1-\rho)=(1-\rho)\frac{1}{1-\frac{\mu}{s+\mu}\rho}=$$

$$=(1-\rho)\frac{s+\mu}{s+\mu-\mu\rho}=\frac{(1-\rho)(s+\mu+\lambda-\lambda)}{s+\mu(1-\rho)}=(1-\rho)+\frac{\lambda(1-\rho)}{s+\mu(1-\rho)}$$

$$w(t)=(1-\rho)\delta(t)+\lambda(1-\rho)e^{-\mu(1-\rho)t}$$

Tempi di attesa in coda

$$F_w(t) = Pr(w \le t) = 1 - \rho \cdot e^{-(1-\rho)\cdot\mu\cdot t}$$

$$W = \frac{\rho}{\mu} \cdot \frac{1}{1-\rho}$$

Detto inoltre w_r l' r-percentile del tempo di attesa (cioè quel valore che non è superato per una percentuale di tempo uguale a r%)

$$Pr(w \le w_r) = \frac{r}{100}$$

$$w_r = \frac{W}{\rho} ln\left(\frac{100\rho}{100-r}\right)$$

Nel tuo appunto, il **percentile** w_r indica il tempo di attesa tale che la probabilità che il tempo di attesa sia **maggiore** di w_r sia pari a una determinata percentuale r. In altre parole, w_r è il tempo di attesa per cui la probabilità che il tempo di attesa sia **minore o uguale** a w_r sia $r\%$. Quindi:

- Se vogliamo sapere, per esempio, quanto tempo è necessario affinché l'**80%** delle persone attenda **meno di quel tempo** (e quindi il 20% attenda **più di quel tempo**), stiamo cercando l'**80° percentile**.

Matematicamente, questo si traduce nel trovare il valore w_r tale che:

$$F_w(w_r) = r$$

Tempi di permanenza nel sistema

Nel caso di disciplina FIFO, l'utente, prima di essere servito dovrà attendere un tempo pari alla somma di k+1 v.a. esponenziali. Questo avverrà con probabilità $\rho^k(1-\rho)$. In media si avrà che:

$$S(s)=\sum_{k=0}^{\infty}\left(\frac{\mu}{s+\mu}\right)^{k+1}\rho^k(1-\rho)=\frac{\mu}{s+\mu}(1-\rho)\frac{1}{1-\frac{\mu}{s+\mu}\rho}=$$

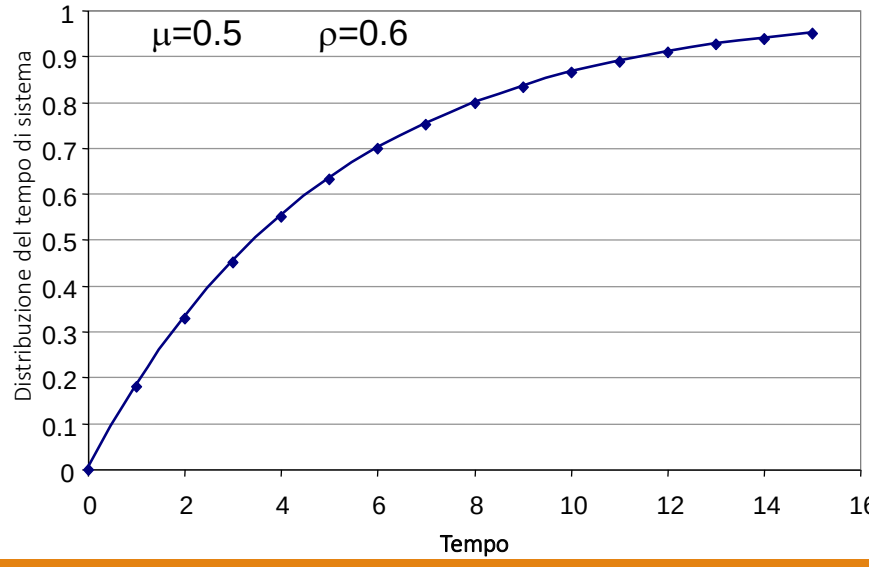
$$=\frac{\mu(1-\rho)}{s+\mu(1-\rho)} \qquad \Rightarrow \qquad s(t)=\mu(1-\rho)e^{-\mu(1-\rho)t}$$

$$F_s(t)=Pr(s\le t)=1-e^{-(1-\rho)\cdot\mu\cdot t} \qquad T=\frac{1}{\mu\cdot(1-\rho)}=\frac{1}{\mu-\lambda}=\overline{w}+\frac{1}{\mu}$$

detto inoltre s_r il percentile r% del tempo di coda

$$Pr(s\le s_r)=\frac{r}{100} \qquad s_r=T\frac{-ln(1-r/100)}{1-\rho}$$

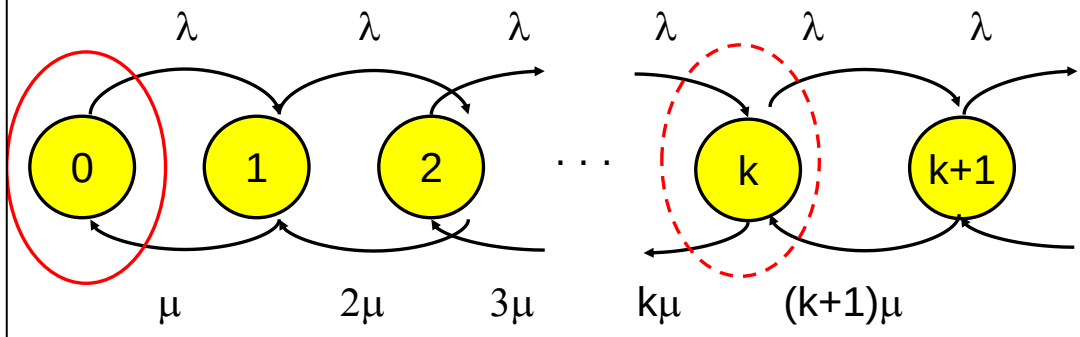
Tempi di permanenza nel sistema



Sistema a coda M/M/∞

$\lambda_k = \lambda$ per k ≥0 frequenza di nascita

$\mu_k = k\mu$ per k ≥0 frequenza di morte



$$\frac{\lambda}{\mu} < \infty \qquad P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{dove} \quad P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

Sistema a coda M/M/∞

$$P_0 = \frac{1}{1+\sum_{k=1}^{\infty}\prod_{i=0}^{k-1}\frac{\lambda}{(i+1)\mu}} = \frac{1}{1+\sum_{k=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^k\frac{1}{k!}} = e^{-\frac{\lambda}{\mu}}$$

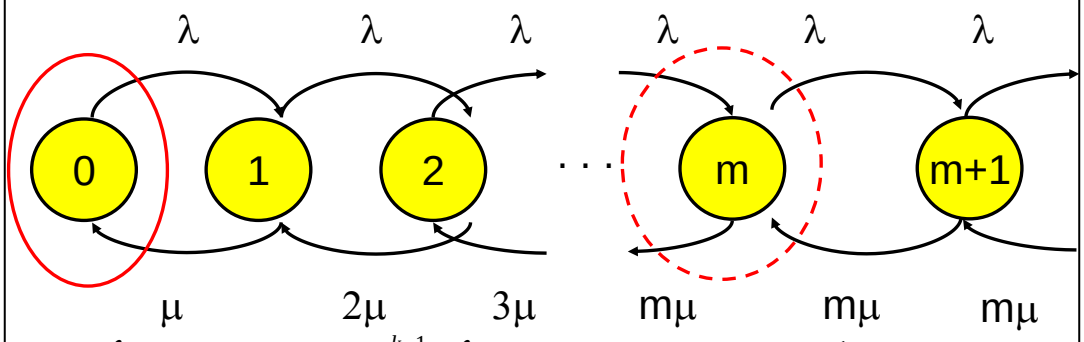
$$P_k = e^{-\frac{\lambda}{\mu}}\prod_{i=0}^{k-1}\frac{\lambda}{(i+1)\mu} = \left(\frac{\lambda}{\mu}\right)^k\frac{1}{k!}e^{-\frac{\lambda}{\mu}}, \qquad k = 0,1,2,\ldots$$

La distribuzione di probabilità degli stati è una distribuzione di Poisson valutata per t=1/μ

$$\overline{N} = \frac{\lambda}{\mu} \qquad T = \frac{1}{\mu}$$

Sistema a coda M/M/m

$$\lambda_k = \lambda \qquad \mu_k = \min(k\mu, m\mu) = \begin{cases} k\mu, & 0 < k < m \\ m\mu, & m \le k \end{cases}$$



$$\rho = \frac{\lambda}{m\mu} < 1 \qquad P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{dove} \quad P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

Sistema a coda M/M/m

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} = P_0 \frac{(m\rho)^k}{k!}, \qquad k \le m$$

$$P_k = P_0 \prod_{i=0}^{m-1} \frac{\lambda}{(i+1)\mu} \prod_{i=m}^{k-1} \frac{\lambda}{m\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m! m^{k-m}} = P_0 \frac{\rho^k m^m}{m!}, \qquad k \ge m.$$

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} = \frac{1}{1 + \sum_{k=1}^{m-1} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \sum_{k=m}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m! m^{k-m}}} =$$

$$= \frac{1}{1 + \sum_{k=1}^{m-1} \frac{(m\rho)^k}{k!} + \sum_{k=m}^{\infty} \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}}} = \frac{1}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \frac{1}{1-\rho}}$$

Sistema a coda M/M/m

Un utente che arriva in ingresso al sistema ha la necessità di accodarsi con probabilità pari a:

$$P[coda] = \sum_{k=m}^{\infty} P_k = \sum_{k=m}^{\infty} P_0 \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}} = P_0 \frac{(m\rho)^m}{m!} \frac{1}{1-\rho}$$

$$P[coda] = \frac{\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!}\frac{1}{1-\rho}}$$

FORMULA DI ELRLAG C, indicata come $C(m,\lambda/\mu)$

E' utilizzata per determinare la **probabilità di attesa** nell'accesso a una risorsa condivisa di *m* serventi disponibili. Ad esempio, può essere utilizzata nei call center per calcolare il numero di operatori necessari per gestire le chiamate entranti posto un certo livello di servizio.

Esempio

Si consideri un centralino telefonico operante ad attesa. Si assuma che:

- a un fascio di giunzioni all'uscita dell'autocommutatore sia offerto un traffico poissoniano entrante con intensità media di 25 Erl;
- tale fascio sia composto da 30 giunzioni;
- la durata di una conversazione telefonica sia distribuita con legge esponenziale negativa e con valore medio di 3 min.

Si determini la probabilità che una chiamata venga accodata

Modello M/M/m

La probabilità di entrare in coda (cioè di subire un ritardo) è data da:

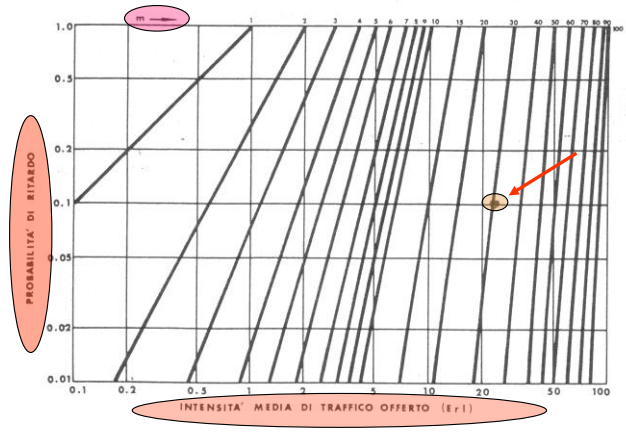
$$C\left(m, \frac{\lambda}{\mu}\right) = \frac{\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}}{\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!}\frac{1}{1-\rho}}$$

C-ERLANG

$m = 30$

Il traffico offerto A_0 è pari a 25 Erlang

$\rho = \frac{\lambda}{m\mu} = \frac{25}{30} = 0.83$ coefficiente di utilizzazione



$$C(30,25) \cong 0.1$$

Sistema a coda M/M/m/m/∞

Ipotesi:

- tempi di interarrivo i.i.d. con distribuzione esponenziale negativa (λ);
- tempi di servizio i.i.d. con distribuzione esponenziale negativa (μ);
- processi di arrivo e di servizio statisticamente indipendenti.
- *m* serventi, statisticamente identici ed indipendenti;
- capacità nulla della fila d'attesa.

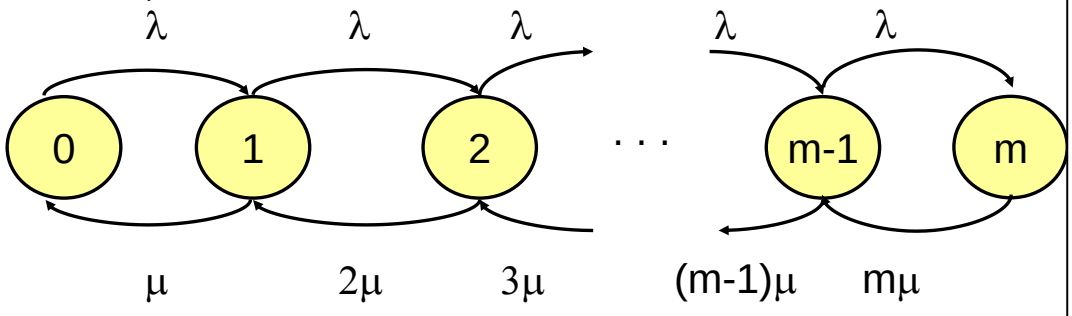
Il processo di coda è descrivibile mediante un processo di Markov di nascita e morte con spazio di stato {0,…, m}.

Il processo di coda è ergodico *per ogni valore positivo di λ e μ* (coda a perdita)

Frequenze di transizione di stato

$\lambda_k=\lambda$ per $0 \le k \le m-1$
frequenza di nascita

$\mu_k=k\mu$ per $1 \le k \le m$
frequenza di morte



$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{dove} \quad P_0 = \frac{1}{1+\sum_{k=1}^{\infty}\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}}$$

Probabilità limite di stato

Per l'equilibrio dei flussi si ha (formule generali):

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \quad \text{dove} \quad P_0 = \frac{1}{\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}$$

posto $A_0 = \lambda/\mu$: traffico offerto al sistema, risulta

$$P_k = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} = P_0 A_o^k \frac{1}{k!} \qquad 0 < k \leq m$$

$$P_0 = \frac{1}{\sum_{j=0}^{m} (A_o)^j \frac{1}{j!}}$$

$$P_k = \frac{A_o{}^k \frac{1}{k!}}{\sum_{j=0}^{m} A_o{}^j \frac{1}{j!}}$$

Probabilità di blocco di servizio

Nel caso di processo di ingresso di Poisson, dato che la probabilità di r.s.o. è indipendente dallo stato, si ha:

$$\Pi_p = S_p \frac{\lambda_m}{\Lambda_o} = S_p$$

Nel caso di sistema a coda M/M/m/m (per k=m)

$$\Pi_p = S_p = \frac{A_o{}^m \frac{1}{m!}}{\sum_{j=0}^{m} A_o{}^j \frac{1}{j!}}$$

FORMULA B
DI ERLANG

Super N.B.!

Formula B di Erlang

L'espressione della probabilità di sistema bloccato e di rifiuto per un sistema a coda M/M/m/m a perdita in senso stretto é denominata anche funzione di Erlang del 1° tipo di ordine m e di argomento A_o

Gode inoltre della proprietà di calcolo di tipo ricorsivo, infatti:

$$E_{1,m}(A_o) = \frac{\frac{A_o{}^m}{m!}}{\sum_{j=0}^{m} A_o{}^j \frac{1}{j!}} = \frac{A_o E_{1,m-1}(A_o)}{m + A_o E_{1,m-1}(A_o)}$$

- con il primo elemento pari a:

$$E_{1,1}(A_o) = \frac{A_o}{1 + A_o}$$

Formula B di Erlang

La grande importanza della formula B di Erlang risiede anche nel fatto che essa risulta valida per qualsiasi distribuzione dei tempi di servizio (resta necessaria l'ipotesi di i.i.d.).

In condizioni di equilibrio statistico la distribuzione del numero di utenti nel sistema è funzione del solo tempo medio di servizio $1/\mu$ e non della distribuzione del tempo di servizio stesso

Parametri prestazionali

Intensità media di traffico o «lavoro» smaltito A_s, che rappresenta il numero medio di serventi contemporaneamente occupati, dipende da A_o e dal numero di serventi m:

$$A_s = \sum_{k=1}^{m} kP_k = A_o\left[1 - E_{1,m}(A_o)\right]$$

Intensità media di traffico o «lavoro» rifiutato:

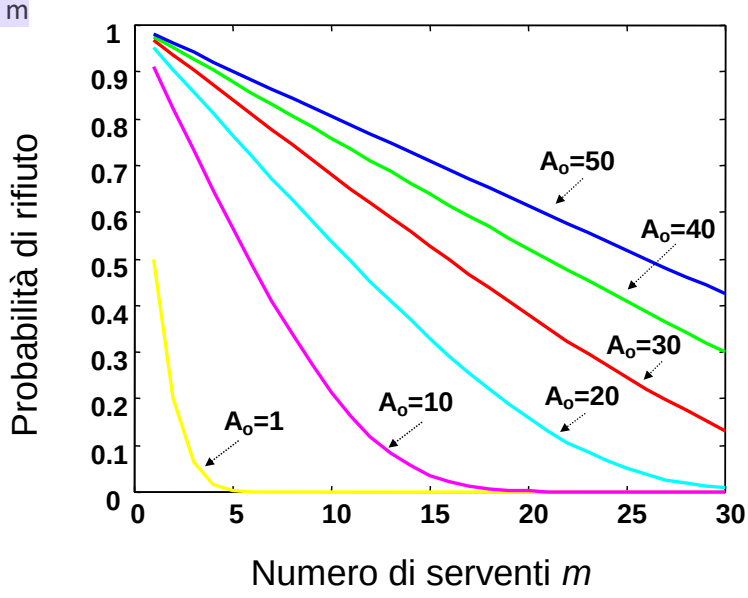
$$A_p = A_o - A_s = A_o E_{1,m}(A_o)$$

Coefficiente di utilizzazione del servente:

$$\rho = \frac{A_s}{m} = \frac{A_o}{m}\left[1 - E_{1,m}(A_o)\right]$$

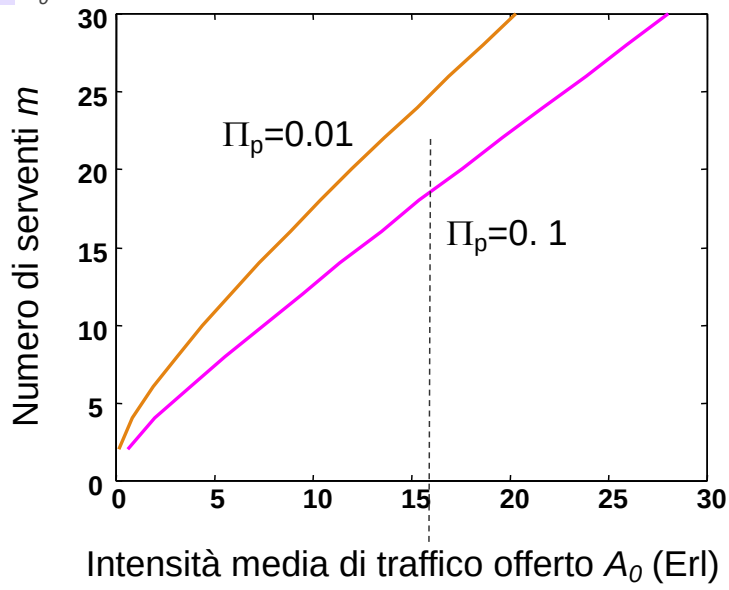
Probabilità di rifiuto in funzione di *m*

La probabilità di rifiuto, a parità di A_0, decresce al crescere del numero di serventi m

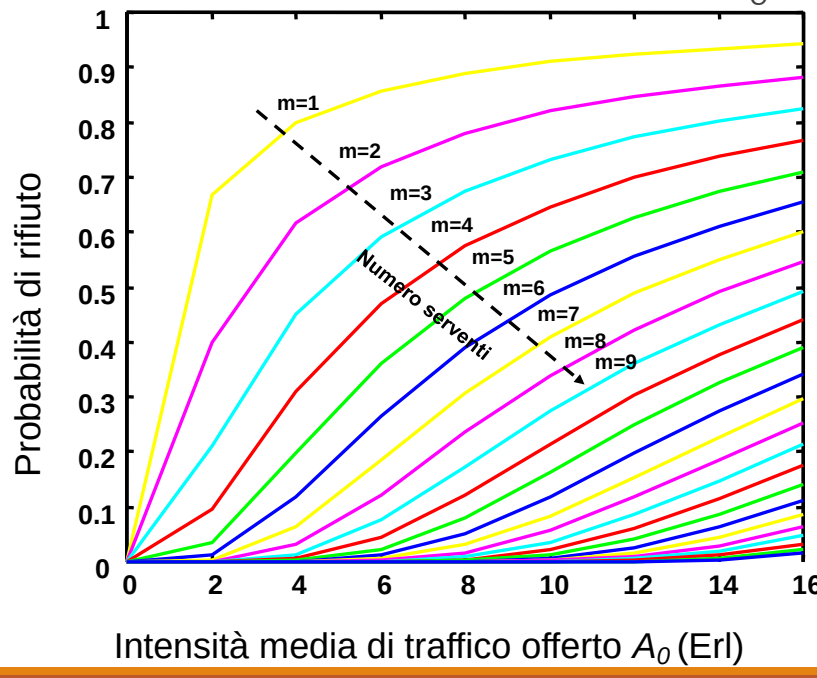


Dimensionamento di *m* in funzione di Π_p

La probabilità di rifiuto è, a parità di *m*, una funzione monotona crescente di A_0

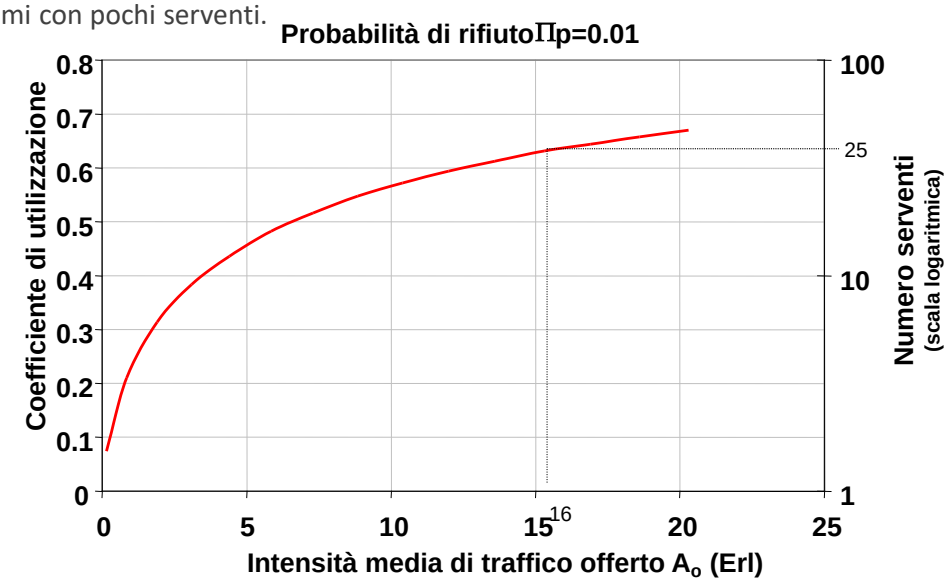


Probabilità di rifiuto in funzione di A_0



ρ in funzione di A_0

- A parità di congestione di chiamata, sistemi con elevato numero di serventi presentano, ***in condizioni di equilibrio statistico***, un rendimento ***MIGLIORE*** rispetto a sistemi con pochi serventi.



B di Erlang: dimensionamento del sistema

Dimensionamento del sistema: stimato il traffico offerto A_0 e fissato il valore massimo per la probabilità di congestione di chiamata Π_{max}, determinare m:

- trovare il più piccolo valore di *m* tale per cui

$$E_{1,m}(A_o) \leq \Pi_{max}$$

- tale valore può essere facilmente determinato per tentativi a partire da *m=1*
- il valore effettivo della congestione di chiamata potrà risultare inferiore a Π_{max}

Esempio

- Intensità di richieste a un server DHCP A_o=100 Erl
- Tale traffico è offerto ad un unico server in modo tale che la probabilità di rifiuto sia minore dell'1%

$$E_{1,m}(A_0) \leq 0.01 \quad \Rightarrow \quad \text{m=117 indirizzi IP necessari}$$

- Si supponga di ripartire tali richieste uniformemente su *n* subnet, con *n*=2, 4, 10 ,25, 50, 100
- Si può notare come all'aumentare di *n* aumenta il numero di indirizzi necessari e diminuisce il ρ di ogni singolo fascio

n	$A_{oi}=(A_o/n)$	m_i	$m=m_i*n$	Π_p	ρ
1	100	117	117	0.0098	0.8463
2	50	64	128	0.0084	0.7747
4	25	36	144	0.0080	0.6889
10	10	18	180	0.0071	0.5516
25	4	10	250	0.0053	0.3979
50	2	7	350	0.0034	0.2847
100	1	5	500	0.0031	0.1994

B di Erlang: valutazione delle prestazioni

Valutazione delle prestazioni: dato il numero dei serventi ed il traffico offerto, determinare la probabilità di congestione di chiamata:

- Occorre notare che solitamente è noto il traffico smaltito A_s^* e il numero di serventi m da cui si può stimare A_0 attraverso la relazione seguente

$$A_o\left[1 - E_{1,m}(A_o)\right] = A_s^*$$

- Una volta calcolato A_0 si calcola la probabilità di congestione di chiamata

$$\Pi_p = E_{1,m}(A_o)$$

Esempio (1/6)

- Si consideri un centralino telefonico automatico (PABX) di una grande azienda. Il centralino è collegato alla rete telefonica nazionale (RTN) tramite un certo numero di linee bidirezionali.
- Si consideri inoltre che:
  - nell'ora di punta gli utenti attestati al centralino formulano mediamente 140 chiamate dirette verso la RTN;
  - nell'ora di punta il numero di chiamate provenienti dalla RTN e dirette verso gli utenti del PABX è mediamente 180;
  - il flusso delle chiamate sia entranti sia uscenti è Poissoniano;
  - la distribuzione di probabilità delle durate delle conversazioni è di tipo esponenziale negativo con valor medio pari a 3 minuti;
  - la modularità delle linee è pari a 4, ovvero si possono inserire linee solo a gruppi di 4;
  - il PABX è del tipo a perdita pura.
- Si determini il numero di linee necessario a garantire un servizio con congestione di chiamata non superiore all'1%.
- Calcolare inoltre la frequenza massima delle chiamate consentita nell'ora di punta.

Per prima cosa calcoliamo Ao, aka traffico offerto

Esempio (2/6)

Il PABX può essere modellato con un sistema a coda del tipo *M/M/m/m* in cui *m* è il numero di linee tra PABX e RTN

Si calcola il traffico globale offerto. Questo è pari alla somma del traffico uscente

$$A_u = \frac{140}{60} 3 = 7\ Erl$$

+

e del traffico entrante

$$A_e = \frac{180}{60} 3 = 9\ Erl$$

quindi

$$A_o = A_u + A_e = 16\ Erl$$

Esempio (3/6)

Per calcolare il numero di linee necessario a garantire una probabilità di congestione di chiamata minore dello 0.01 si deve determinare il minimo valore di *m* tale che

$$E_{1,m}(A_o) \leq 0.01$$

Si ottiene in tal caso m=25

A causa del vincolo sulla modularità il numero di linee da inserire sarà pari quindi a m=28

Dato tale numero di linee la congestione di chiamata sarà notevolmente inferiore a quella richiesta infatti

$$\Pi_{p,effettivo} = E_{1,28}(16) = 0.0019$$

E con queste 28 linee, alla fine QUANTE CHIAMATE possiamo supportare, pur RISPETTANDO il vincolo di probabilità di blocco < 0.01? --> aka riusiamo Erland, ma stavolta è Ao l'incognita, quello MAX!

Esempio (4/6)

Per determinare la frequenza massima delle chiamate consentita nell'ora di punta si calcola prima il valore di $A_{0,max}$ tale che

$$E_{1,28}(A_{o\,max}) \leq 0.01$$

da cui si ricava $A_{0,max} = 18.64$

per cui

$$\lambda_{max} = A_{o\,max}\frac{60}{3} \cong 373\ chiamate/ora$$

Esempio(5/6)

Si consideri il PABX dimensionato con 28 linee bidirezionali che lo connettono alla Rete Telefonica Nazionale.

A distanza di tempo dalla sua installazione si vuole valutare la qualità di servizio offerta sapendo che a seguito di una campagna di misure si è riscontrato, nell'ora di punta, un valore di intensità media di traffico smaltito pari a circa 20.42 Erl.

Aka dobbiamo capire se persino nell'ora di punta si ha che

la prob di blocco è ancora RISPETTATA!

Esempio (6/6)

Dato il traffico smaltito misurato si può ricavare il traffico offerto al sistema risolvendo l'equazione

$$A_o\left(1-E_{1,28}\left(A_o\right)\right)=20.42$$

da cui si ha

$$A_o = 21\ Erl$$ --> e con questo traffico, come sono messa con la prob di blocco?

Per quanto riguarda il valore di congestione di chiamata, si ha

$$E_{1,28}\left(21\right)=0.0277$$ --> oh shit è maggiore! Dovrò intervenire

Il PABX non è più in grado di rispettare il vincolo sul grado di servizio. Le prestazioni sono variate, ad esempio, per un leggero incremento dell'utenza. Bisognerà quindi ridimensionare il numero di linee per riportare la probabilità di rifiuto sotto la soglia dello 0.01

Esempio (1/3)

Si considerino N terminali di utente che possiamo modellare come sorgenti di traffico dati. Ognuna emette traffico poissoniano con ritmo binario medio pari a λ bit/s e lunghezza dei pacchetti con distribuzione esponenziale negativa di media L.

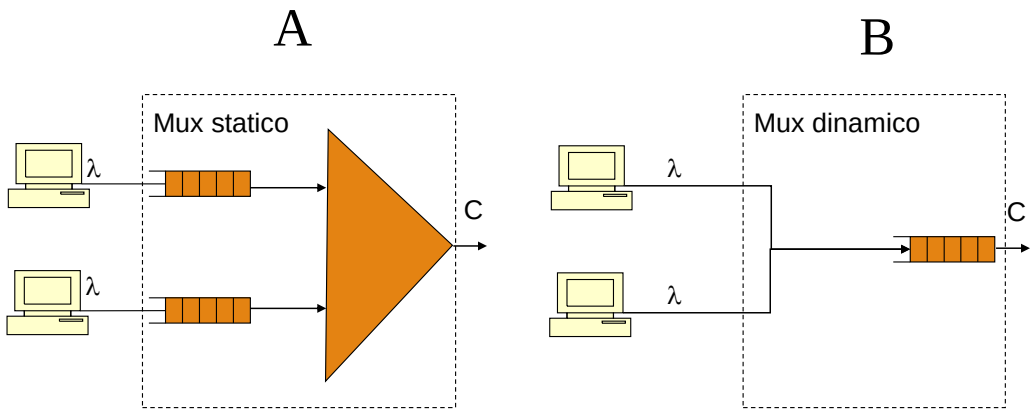
Il traffico prodotto è inoltrato verso un load balancer (multiplatore) con capacità di smaltimento complessiva pari a C.

Si considerino due tecniche di multiplazione, con $N^* \lambda < C$:

1) Multiplazione statica: ad ogni utente è assegnato un buffer di dimensione infinita ed una capacità pari a C/N
2) Multiplazione dinamica: tutta la capacità è dinamicamente condivisa tra tutte le sorgenti, che utilizzano un unico buffer di dimensione infinita

Si valuti e discuta la prestazioni delle due soluzioni in termini di coefficiente di utilizzazione della capacità C e del tempo medio di sistema T

Esempio (2/3)



Esempio (3/3)

- **Assunzioni: flussi statisticamente indipendenti**
- **Gestione delle code di tipo FIFO**

A - **N code M/M/1**

coefficiente di utilizzazione: $\rho = \frac{\frac{\lambda}{C}}{NL} = \frac{\lambda NL}{C}$

tempo medio di sistema: $T = \frac{\rho}{\lambda(1-\rho)}$ --> E' quello del sistema M/M/1, e io se conosco rho lo conosco! Lo avevamo visto l'altra volta!

qui affascio! ----> B - **1 coda M/M/1**

coefficiente di utilizzazione: $\rho = \frac{\frac{N\lambda}{C}}{L} = \frac{\lambda NL}{C}$

tempo medio di sistema: $T = \frac{\rho}{N\lambda(1-\rho)}$

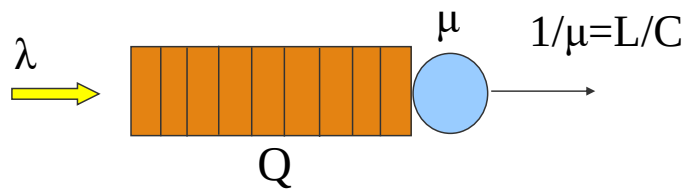
N.B.

Il GUADAGNO nel MULTIPLARE:

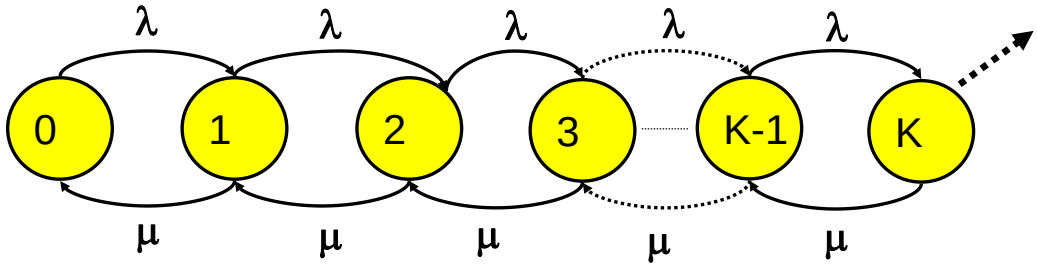
1) Nei sistemi orientati ALLA PERDITA (no coda) --> se affasciamo il traffico otteniamo un COEFFICIENTE DI UTILIZZAZIONE PIU' ALTO per ogni singolo servente!

2) Nei sistemi orientati AL RITARDO --> se affasciamo il traffico otteniamo dei TEMPI DI PERMANENZA di CODA MOLTO RIDOTTO, ridotto di quel fattore N!

Sistema a coda M/M/1/K



K=Q+1



$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{dove} \quad P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

Indicando con $A_0=\lambda/\mu$ il traffico offerto, si ricavano le probabilità limite di stato

$$P_k = P_0 A_0^k \quad \text{dove} \quad P_0 = \frac{1}{1+\sum_{i=1}^{K} A_0^i} = \frac{1}{1+\sum_{i=1}^{K} A_0^i} = \frac{1}{1+\left(\sum_{i=1}^{\infty} A_0^i - \sum_{i=K+1}^{\infty} A_0^i\right)} = \frac{1-A_0}{1-A_0^{K+1}}$$

$$= \frac{1}{1+\left(\frac{A_0}{1-A_0} - \frac{A_0^{K+1}}{1-A_0}\right)} = \frac{1-A_0}{1-A_0^{K+1}}$$

$$P_k = \begin{cases} \frac{1-A_0}{1-A_0^{K+1}} A_0^k & 0 \le k \le K \\ 0 & k > K \end{cases}$$

$$A_S = A_0(1-P_{rifiuto})$$

Ricorda:

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{dove} \quad P_0 = \frac{1}{1+\sum_{k=1}^{\infty}\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

Proprietà “PASTA”

Catena di Markov: stazionaria

- Il processo si è avviato dalla condizione statistica di stazionarietà, o
- Il processo dura per un tempo $t \to \infty$

- La probabilità che in un dato istante *t* il processo si trovi nello stato *i* è uguale alla probabilità stazionaria che

$$p_i = \lim_{t\to\infty} P\{N(t) = i\} = \lim_{t\to\infty} \frac{T_i(t)}{t}$$

T_i =tempo di permanenza del processo nello stato *i*

Domanda: Per un sistema a coda M/M/1, se t è il tempo di un arrivo, qual è la probabilità che $N(t)=i$?

- Risposta: Poisson Arrivals See Time Averages (*PASTA*).

Proprietà PASTA

Probabilità stazionarie:

$$p_n = \lim_{t\to\infty} P\{N(t) = n\}$$

Probabilità stazionarie in corrispondenza di un arrivo:

$$a_n = \lim_{t\to\infty} P\{N(t^-) = n \mid \text{arrival at } t\}$$

Ipotesi LAA (Lack of Anticipation): I futuri tempi di interarrivo e I tempi di servizio dei clienti arrivati precedentemente sono independenti

- Teorema: In un sistema a coda che soddisfa l'ipotesi LAA:
  1. Se il proceddo degli arrivi è di Poisson si ha che
     $$a_n = p_n, \quad n = 0,1,\ldots$$
  2. Quello di Poisson è il solo processo avente questa proprietà (condizione necessaria e sufficiente)

Proprietà PASTA

La proprietà PASTA si applica ad altri processi di arrivo?

Esempio:

Arrivi dererministici ogni 10 sec

Tempo di servizio deterministico di 9 sec

- In corrisponedenza di ogni arrivo il sistema è vuoto a_1=0
- Il tempo medio in cui un solo utente è presente nel sistema p_1=0.9

Le medie osservate dall'utente non sono necessariamente medie temporali del sistema

L'aleatorietà non è di aiuto, a meno che sia generata da un processo di Poisson!

Proprietà PASTA: dimostrazione

Sia $A(t,t+\delta)$, l'evento in cui vi sia un arrivo in $[t, t+\delta)$

Se che un utente arriva in t, la probabilità $a_n(t)$ di trovare il sistema nello stato n è data da

$$P\{N(t^-) = n \mid \text{arrival at t}\} = \lim_{\delta \to 0} P\{N(t^-) = n \mid A(t,t+\delta)\}$$

$A(t,t+\delta)$ è indipendente dallo stato del sistema prima di t, $N(t^-)$

- $N(t^-)$ è determinato dai tempi di arrivo $<t$, e dai corrispondenti tempi di servizio
- $A(t,t+\delta)$ è indipendente dallo stato del sistema, ossia dagli arrivi $<t$ [Poisson] e dai tempi di servizio degli utenti arrivati $<t$ [LAA]

$$a_n(t) = \lim_{\delta \to 0} P\{N(t^-) = n \mid A(t,t+\delta)\} = \lim_{\delta \to 0} \frac{P\{N(t^-) = n, A(t,t+\delta)\}}{P\{A(t,t+\delta)\}}$$

$$= \lim_{\delta \to 0} \frac{P\{N(t^-) = n\} P\{A(t,t+\delta)\}}{P\{A(t,t+\delta)\}} = P\{N(t^-) = n\}$$

$$a_n = \lim_{t \to \infty} a_n(t) = \lim_{t \to \infty} P\{N(t^-) = n\} = p_n$$

Esempi

Esempio 1: Arrivi non di Poisson

Tempi di interarrivo IID distribuiti uniformememente fra 2 and 4 sec

Tempi di servizio deterministici di 1 sec

- In corrispondenza di ogni arrivo il sistema è vuoto, quindi a_1 =0.
- λ=1/3, T=1 → N=Tλ=1/3 → p_0=2/3, p_1=1/3

Esempio 2: mancanza dell'ipotesi LAA

Arrivi di Poisson

Tempo di servizio del'utente *i*: $S_i = \alpha T_{i+1}$, $\alpha < 1$

- In corrispondenza di ogni arrivo il sistema è vuoto, a_1 =0.
- Il tempo medio in cui un solo utente è presente nel sistema $p_1 = \alpha$

Distribuzione dopo la partenza

$$d_n = \lim_{t\to\infty} P\{X(t^+) = n \mid \text{departure at } t\}$$

Probabilità di stato stazionarie dopo una partenza:

Usando la stessa proprietà di Markov:

- I limiti di a_n e d_n esistono e coincidono

$a_n = d_n$, $n = 0,1,\ldots$

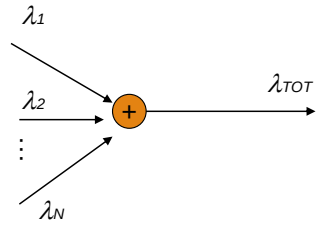
In condizioni stazionarie, il sitema appare stocasticamente identico agli utenti che arrivano e che lasciano il sistema.

Arrivi di Poisson + LAA: un utente che arriva e un utente che lascia il sistema vedono il sistema stesso con la stessa statistica osservabile in un tempo aleatorio.

Ancora sui processi di Poisson…

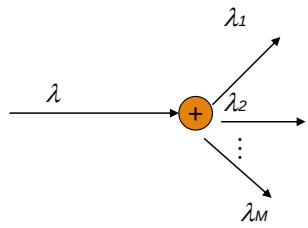
L'aggregazione di N processi di Posson indipendenti di parametro λ_i, i=1…N, è un processo di Poisson di parametro

$$\lambda_{TOT} = \sum_i \lambda_i$$



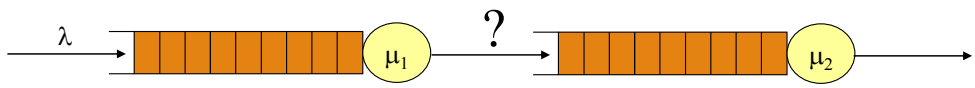
Ancora sui processi di Poisson…

La separazione statistica di un processo di Poisson di parametro λ con probabilità $p_1, p_2, \ldots, p_M$ genera M processi di Poisson di parametro $\lambda_i=\lambda p_i$, i=1,…,M.



Teorema di Burke

Date due code in "tandem", qual è la distribuzione dei tempi di interarrivo alla seconda coda, se la prima è una M/M/1?



Tale distribuzione sarà equivalente a quella dei tempi di interpartenza ($D(t)$) dalla coda 1

Teorema di Burke

Indichiamo con

- $D^*(s)$ la trasformata di Laplace di $D(t)$
- $B^*(s)$ la trasformata di Laplace di $B(t)$ (distribuzione dei tempi di servizio)

Quando un cliente parte da una coda (cioè ha ricevuto il suo servizio) può verificarsi uno solo dei seguenti eventi:

- un altro cliente è presente sulla coda 1 e sarà subito servito
- la coda 1 è vuota, quindi bisognerà attendere un tempo uguale alla somma di due contributi prima di un nuovo arrivo alla coda 2:
  - il tempo fino all'arrivo di un altro cliente
  - il suo tempo di servizio

Teorema di Burke

Nel primo caso risulta

- $D^*(s)\,|_{\text{coda non vuota}} = B^*(s)$

Nel secondo caso, ho la somma di due variabili aleatorie indipendenti, perciò la variabile aleatoria somma sarà la convoluzione delle pdf, quindi, nel dominio di Laplace, il prodotto delle trasformate delle distribuzioni

- $D^*(s)\,|_{\text{coda vuota}} = B^*(s)\ \lambda/(s+\lambda)$

Poiché il servente è di tipo esponenziale, si ha inoltre che

- $B^*(s) = \mu_I/(s+\mu_I)$

Teorema di Burke

Poiché la probabilità di avere il sistema non vuoto è pari $\rho = \lambda/\mu_1$, si ottiene che

- $D^*(s) = (1-\rho)\, D^*(s)|_{\text{coda vuota}} + \rho\, D^*(s)|_{\text{coda non vuota}}$

Sostituendo i valori precedenti si ottiene

- $D^*(s) = (1-\rho)\,(\lambda/(s+\lambda))\,(\mu_1/(s+\mu_1)) + \rho\,(\mu_1/(s+\mu_1))$

che risulta uguale a

- $D^*(s) = (\lambda/(s+\lambda)) = A^*(s) \Rightarrow D(t) = A(t) = 1-e^{-\lambda t}$ *per* $t \geq 0$

Quindi i tempi di interpartenza sono distribuiti esponenzialmente con lo stesso parametro dei tempi di interarrivo:

- la coda 2 può essere trattata come una M/M/1 indipendente dalla coda 1 !

Burke estende questo risultato alle code M/M/m

Teorema di Burke

Considerando le operazioni viste sui processi di Poisson, il Teorema di Burke implica che se si connettono dei server (cioè sistemi a coda) in modalità feed-forward allora ogni nodo della rete di code può essere esaminato SINGOLARMENTE come se fosse un sistema a coda indipendente.

Questo risultato è generalizzato dal Teorema di Jackson per le reti di code aperte

Teorema di Jackson

Si consideri una rete di code formata da K nodi (sistemi a coda) che soddisfano le seguenti tre condizioni:

- Ogni nodo contenga c_k serventi aventi tempo di servizio distribuiti esponenzialmente con parametro μ_k.
- Gli utenti provenienti dall'esterno giungono al generico nodo k secondo un processo di Poisson con parametro λ_k.
- Quando un utente è servito al nodo k è trasferito "istantanemente" al generico nodo j con probabilità p_{kj} oppure esce dalla rete con probabilità $1-\Sigma_j p_{kj}$

Teorema di Jackson

Il tasso degli arrivi al generico nodo k sarà:

$$\Lambda_k = \lambda_k + \sum_{j=1}^{K} p_{jk} \Lambda_j$$

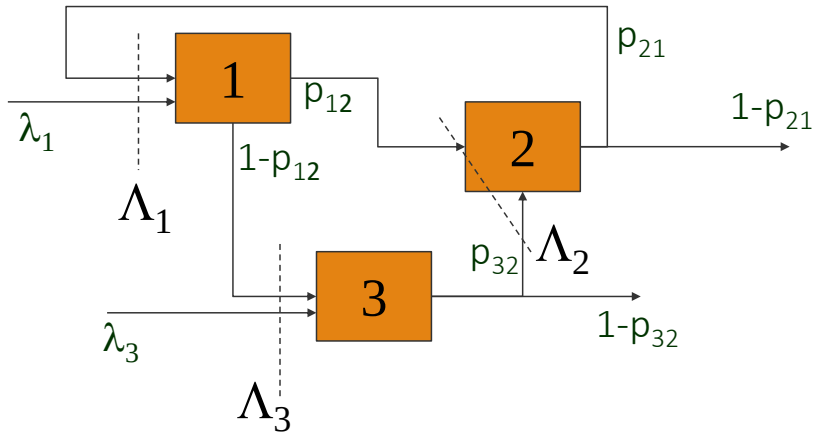
Indicando con $p(n_1,\ldots,n_K)$ la probabilità congiunta stazionaria di stato nei nodi, se

$$\Lambda_k < c_k \mu_k \qquad \forall k,$$

Allora $p(n_1,\ldots,n_K) = p_1(n_1)\, p_2(n_2) \ldots p_K(n_K)$

dove $p_i(n_i)$, i=1,…,K, è la probabilità stazionaria che il nodo i si trovi nello stato n_i, (vi siano ni utenti nel sistema a coda i), modellandolo come un sistema $M/M/c_i$ con processo degli arrivi di Poisson con parametro Λ_i e tasso di servizio μ_i

Teorema di Jackson



$$\begin{cases} \Lambda_1=\lambda_1+p_{21}\Lambda_2 \\ \Lambda_2=p_{12}\Lambda_1+p_{32}\Lambda_3 \\ \Lambda 3=\lambda_3+(1-p_{12})\Lambda_1 \end{cases} \Rightarrow \Lambda_1, \Lambda_2, \Lambda_3$$

Ipotesi di indipendenza di Kleinrock

- Tempi di interarrivo indipendenti alle varie code della rete

- Tempo di servizio di un generico pacchetto nelle varie code indipendente.
  - La lunghezza del pacchetto è random ogni volta che un pacchetto è trasmesso attraverso un collegamento.

- Tempo di servizio e tempi di interarrivo statisticamente indipendenti.

Le assunzioni sono state validate attraverso risultati sperimentali e simulazioni. In tal caso si ha che:

> La distribuzione stazionaria degli stati approssima quella descritta dal teorema di Jackson

L'approssimazione è accettabile quando:

- Il processo degli arrivi dai punti in ingresso alla rete è un processo di Poisson.
- Il tempo di trasmissione del pacchetto è una variabile aleatoria approssimabile mediante un'esponenziale.
- Molto flussi di pacchetti sono multiplati in ogni collegamento.
- La rete è densamente connessa
- Vale per un'intensità di traffico da moderata a pesante.

L'indipendenza in questione riguarda i tempi di servizio e i tempi di arrivo.