

RETI DI TLC

M. Ajmone Marsan

F. Neri

appunti dalle lezioni

con contributi di:

Andrea Bianco

Claudio Casetti

Renato Lo Cigno

Michela Meo

Antonio Nucci

3

Elementi di teoria delle code

Una coda è un sistema che viene rappresentato come nella figura 3.1, dove si distinguono gli *arrivi* dei clienti, una *fila d'attesa*, un *servizio*, e le *partenze* dei clienti.

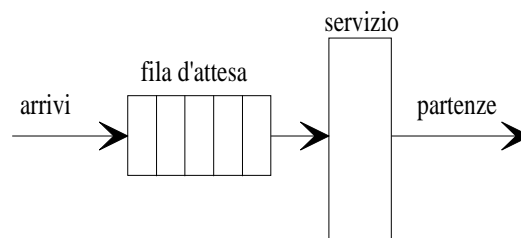


Figura 3.1. Una coda

La notazione usata per classificare i modelli a coda è quella proposta da Kendall. Essa comprende sei indicatori $A/B/c/d/e - x$ che hanno il seguente significato.

1. Con A si caratterizza il processo stocastico secondo cui si succedono gli arrivi dei clienti. In particolare, si classifica il tipo di distribuzione dei tempi tra due arrivi consecutivi. A può assumere uno dei valori M , $Geom$, D , E_n , H_n , G , i cui significati sono

M	-	processo degli arrivi di tipo markoviano in tempo continuo (tempi di interarrivo con densità di probabilità esponenziale), possibilmente (ma non necessariamente) arrivi secondo un processo di Poisson
$Geom$	-	processo degli arrivi di tipo markoviano in tempo discreto (tempi di interarrivo con densità di probabilità geometrica)
D	-	arrivi deterministici (tempi di interarrivo costanti)
E_n	-	processo degli arrivi con tempi di interarrivo con densità di probabilità Erlang a n stadi
H_n	-	processo degli arrivi con tempi di interarrivo con densità di probabilità iperesponenziale a n stadi
G	-	processo degli arrivi di tipo generale

I processi di arrivo appena elencati fanno riferimento normalmente ad arrivi di singoli clienti. Quando invece i clienti arrivano a gruppi (*batch arrivals* in inglese), si usa posporre (normalmente ad apice) all'indicatore del processo degli arrivi $[X]$, dove X è una variabile casuale che descrive il numero di clienti in ogni gruppo. Ad esempio $M^{[X]}$ indica interarrivi esponenziali di gruppi comprendenti un numero di clienti pari ad istanze della variabile casuale X .

2. Con B si descrive la caratteristica probabilistica dei *tempi di servizio*. Ogni cliente richiede al servitore un servizio la cui durata è una variabile casuale indipendente con densità di probabilità di vario tipo

M	-	densità di probabilità esponenziale
$Geom$	-	densità di probabilità geometrica
D	-	tempi di servizio costanti
E_n	-	densità di probabilità Erlang a n stadi
H_n	-	densità di probabilità iperesponenziale a n stadi
G	-	densità di probabilità di tipo generale

3. Con c si indica il numero (intero, possibilmente ∞) di *servitori* compresi nella stazione di servizio.
4. Con d si descrive la capacità della *fila di attesa*, cioè il massimo numero di clienti che possono trovare posto nella fila di attesa (escludendo quindi i clienti già in servizio). Si può avere $d = \infty$.
5. Con e si indica il numero di elementi nella popolazione da cui arrivano i clienti e quindi il massimo numero di clienti che possono arrivare alla coda. Anche in questo caso si può avere $e = \infty$.
6. Con x si indica la disciplina di coda, ovvero la regola che viene seguita per scegliere il prossimo cliente da servire tra quelli nella fila di attesa. Si possono avere diverse possibilità

FIFO (FCFS)	-	First In First Out (First Come First Served)
LIFO (LCFS)	-	Last In First Out (Last Come First Served)
RO	-	Random Order
RR	-	Round Robin
PS	-	Processor Sharing
Priorità		

Le discipline FIFO e LIFO specificano che il prossimo cliente da servire è rispettivamente il primo e l'ultimo nella fila di attesa.

La disciplina RO prevede che il prossimo cliente da servire venga estratto a sorte tra tutti quelli presenti nella fila di attesa.

Nel caso della disciplina RR l'unico servitore assegna ciclicamente a tutti i clienti nel sistema un intervallo di servizio costante di breve durata. Facendo tendere a zero la durata dell'intervallo di servizio si ottiene la configurazione PS (*Processor Sharing*) in cui, essendoci k clienti nel sistema, ognuno di essi riceve servizio con velocità uguale a $1/k$ -esimo della velocità caratteristica del servitore.

La disciplina con priorità prevede che la fila di attesa sia ordinata secondo la priorità dei clienti e non secondo l'ordine di arrivo, e che il prossimo cliente da servire sia quello che ha atteso più a lungo tra quelli in attesa con la priorità più alta.

I due parametri d ed e si indicano solo nel caso che siano diversi da ∞ . Il parametro x si indica solo se diverso da FIFO. Quindi

$$\begin{aligned} \cdot / \cdot / \cdot / \infty / \infty - FIFO &\Rightarrow \cdot / \cdot / \cdot \\ \cdot / \cdot / \cdot / k / \infty &\Rightarrow \cdot / \cdot / \cdot / k \\ \cdot / \cdot / \cdot / \infty / k &\Rightarrow \cdot / \cdot / \cdot / / k \end{aligned}$$

Con la notazione introdotta si possono avere un grande numero di sistemi diversi. Per esempio

$$\begin{aligned} M/M/1 \\ M/M/1 - LIFO \\ M/G/2/3/10 \\ G/M/1//6 \\ E_n/D/2/0 - RO \end{aligned}$$

3.1 La coda $M/M/1$

Con la notazione $M/M/1$ indichiamo una coda in cui gli arrivi si susseguono secondo un processo di Poisson (ciò deriva dal fatto che i tempi di interarrivo sono esponenziali, la capacità della fila di attesa è infinita e la popolazione dei clienti è infinita), i tempi di servizio hanno densità di probabilità esponenziale, si ha un solo servitore nel servizio, la fila di attesa ha dimensione illimitata, come pure illimitata è la popolazione da cui arrivano i clienti. La disciplina di coda è FIFO.

Supponiamo che il parametro del processo di Poisson degli arrivi sia λ e che il tempo medio di servizio sia pari a $1/\mu$ (con μ indichiamo la velocità di servizio). Come abbiamo già visto, scegliendo come stato della coda il numero totale di clienti o nella fila di attesa o in servizio, il modello a coda corrisponde ad una catena di Markov di nascita e morte tempo continua con velocità di nascita costante, pari a λ e velocità di morte costante pari a μ .

La condizione di ergodicità del modello markoviano ottenuto è data dalla relazione $\rho = \lambda/\mu < 1$. Quando tale relazione è soddisfatta si possono ricavare le probabilità di regime

$$\pi_i = (1 - \rho)\rho^i \quad i \geq 0 \quad (3.1)$$

che corrispondono alle probabilità di avere i clienti nella fila di attesa o in servizio in condizione di stato stazionario.

3.1.1 Numero medio di clienti nella coda a regime

Indichiamo con N la variabile casuale che indica il numero di clienti nel sistema (nella fila di attesa o in servizio) a regime. È possibile scrivere

$$\begin{aligned} E[N] &= \sum_{i=0}^{\infty} i \pi_i = \\ &= \sum_{i=0}^{\infty} i (1 - \rho)\rho^i = \\ &= \rho(1 - \rho) \sum_{i=0}^{\infty} i \rho^{i-1} = \\ &= \rho(1 - \rho) \frac{1}{(1 - \rho)^2} = \end{aligned}$$

$$= \frac{\rho}{1 - \rho}$$

3.1.2 Probabilità che il servitore sia occupato

La probabilità che a regime il servitore sia occupato si può calcolare come

$$\begin{aligned} P\{\text{servitore occupato}\} &= \sum_{j=1}^{\infty} \pi_j = \\ &= 1 - \pi_0 = \\ &= 1 - (1 - \rho) = \rho \end{aligned}$$

Si noti come tale probabilità sia pari all'intensità di traffico.

Questo risultato vale per qualsiasi sistema a servitore singolo e non solo per la coda $M/M/1$, tenendo conto che in generale la definizione dell'intensità di traffico è

$$\rho = E[\lambda]E[S] \quad (3.2)$$

dove $E[\lambda]$ è la velocità *media* di arrivo dei clienti, ed $E[S]$ è il tempo medio di servizio.

3.1.3 Numero medio di clienti nella fila di attesa

Il numero medio di clienti nella fila di attesa a regime è dato da

$$\begin{aligned} E[N_f] &= \sum_{i=1}^{\infty} (i - 1) \pi_i \\ &= \sum_{i=1}^{\infty} i \pi_i - \sum_{i=1}^{\infty} \pi_i \\ &= \sum_{i=1}^{\infty} i (1 - \rho) \rho^i - (1 - \pi_0) \\ &= \frac{\rho}{1 - \rho} - \rho \end{aligned}$$

e quindi

$$E[N_f] = \frac{\rho^2}{1 - \rho} \quad (3.3)$$

Si può ora notare che il numero totale di clienti nella coda è dato dalla somma del numero di clienti nella fila di attesa più il numero di clienti in servizio

$$N = N_f + N_s \quad (3.4)$$

e quindi, per la linearità della media

$$E[N] = E[N_f] + E[N_s] \quad (3.5)$$

La variabile casuale N_s può assumere i due valori 0 con probabilità π_0 e 1 con probabilità $(1 - \pi_0)$. Si può allora calcolare

$$E[N_s] = 0 \pi_0 + 1 (1 - \pi_0)$$

ottenendo

$$E[N_s] = \rho \quad (3.6)$$

Possiamo allora ricavare per via diversa il risultato dell'equazione (3.3)

$$E[N_f] = E[N] - E[N_s] = \frac{\rho}{1 - \rho} - \rho = \frac{\rho^2}{1 - \rho} \quad (3.7)$$

3.1.4 Tempo medio trascorso da un cliente nella coda

Dalla distribuzione di regime si possono calcolare altri parametri interessanti, come la media del tempo di permanenza di un cliente nel sistema (T) e del tempo trascorso da un cliente nella fila di attesa (T_f).

Si può scrivere che

$$T = T_f + T_s \quad (3.8)$$

dove T_s è il tempo medio di servizio; quindi

$$E[T] = E[T_f] + E[T_s] \quad (3.9)$$

È noto che $E[T_s] = 1/\mu$. Per calcolare $E[T_f]$, ricordando che la disciplina di coda è FIFO, si può scrivere

$$E[T_f] = \sum_{i=0}^{\infty} E[T_f \mid i] \pi_i^{(a)}$$

dove $E[T_f \mid i]$ rappresenta il tempo medio trascorso dal cliente nella fila di attesa condizionato dal fatto di trovare i clienti nel sistema all'arrivo e $\pi_i^{(a)}$ indica la probabilità che un cliente arrivando trovi i clienti nella coda. Degli i clienti, $(i - 1)$ sono nella fila di attesa e uno in servizio. Il tempo di servizio residuo del cliente in servizio nel momento di arrivo del cliente che stiamo considerando è indicato da $T_{sr}^{(0)}$. Osservando che $E[T_f \mid 0] = 0$, si può scrivere che per $i > 0$ vale

$$E[T_f \mid i] = E \left[T_{sr}^{(0)} + \sum_{j=1}^{i-1} T_s^{(j)} \right] = E[T_{sr}] + (i - 1)E[T_s] \quad (3.10)$$

e

$$E[T_f] = \sum_{i=1}^{\infty} E[T_{sr}] \pi_i^{(a)} + \sum_{i=1}^{\infty} (i - 1)E[T_s] \pi_i^{(a)} \quad (3.11)$$

L'assenza di memoria dei tempi di servizio implica che

$$E[T_{sr}] = \frac{1}{\mu} = E[T_s] \quad (3.12)$$

Per completare il calcolo è necessario saper calcolare $\pi_i^{(a)}$. $\pi_i^{(a)}$ in generale non coincide con π_i . Ad esempio, nel caso di tempi di interarrivo costanti pari a $1/\lambda$ e tempi di servizio pure costanti pari a $1/\mu < 1/\lambda$, si ha $\pi_1 = \lambda/\mu$ e $\pi_0 = 1 - \pi_1$, ma $\pi_0^{(a)}$ vale 0 (se il non c'erano clienti nel sistema al tempo 0). Per quanto riguarda invece un processo degli arrivi di tipo Poisson, un importante teorema recita che "Poisson arrivals see time averages" (PASTA); quindi la probabilità che un arrivo

trovi i clienti nel sistema è uguale alla probabilità che ci siano i clienti nel sistema in un istante qualsiasi ($\pi_i^{(a)} = \pi_i$).

Utilizzando tale teorema si trova che

$$\begin{aligned} E[T_f] &= \frac{1}{\mu} \sum_{i=1}^{\infty} i \pi_i \\ &= \frac{1}{\mu} \frac{\rho}{1-\rho} \end{aligned}$$

Il valor medio del tempo totale trascorso nella coda vale quindi

$$E[T] = E[T_f] + E[T_s] = \frac{1}{\mu} + \frac{1}{\mu} \frac{\rho}{1-\rho} = \frac{1}{\mu} \frac{1}{1-\rho} = \frac{1}{\mu - \lambda} \quad (3.13)$$

3.1.5 Il risultato di Little

Avendo trovato le espressioni dei numeri medi di clienti nella coda, nella fila di attesa ed in servizio ed i tempi medi relativi, possiamo osservare che

$$E[N] = \lambda E[T] \quad (3.14)$$

$$E[N_f] = \lambda E[T_f] \quad (3.15)$$

$$E[N_s] = \lambda E[T_s] \quad (3.16)$$

Queste tre equazioni sono casi particolari di uno dei risultati più generali e più potenti della teoria delle code. Tale risultato è il *teorema di Little*.

Il teorema dice che, dato un sistema stabile al quale arrivano clienti con una velocità media $E[\lambda]$ finita, per il quale il numero medio di clienti nel sistema $E[N]$ è finito, il tempo medio trascorso dai clienti nel sistema $E[T]$ è pari a

$$E[T] = \frac{E[N]}{E[\lambda]} \quad (3.17)$$

Intuitivamente è facile osservare che, per un sistema a coda deterministico in equilibrio, nel quale entra ed esce 1 cliente per unità di tempo (comunque definita) e nel quale si osserva sempre un cliente, il tempo di permanenza dei clienti nel sistema deve essere di 1 unità di tempo. Se raddoppia il tempo di permanenza, raddoppia anche il numero di clienti nel sistema. Se si aggiunge casualità al processo degli arrivi o ai tempi di servizio, ciò non altera le medie sul lungo periodo. Nel prossimo paragrafo si darà una giustificazione più formale, ma ancora basata su di un'analisi temporale, del risultato di Little, seguendo un approccio sovente chiamato "analisi operativa" del sistema a coda.

L'unico requisito per la validità del teorema di Little è che il sistema che si sta studiando sia in equilibrio e che $E[\lambda]$ e $E[N]$ siano finiti. Occorre inoltre prestare attenzione a calcolare le tre grandezze che compaiono nella relazione (3.17) sullo stesso flusso di clienti.

Il risultato di Little sarà usato spesso per trovare $E[T]$, perché $E[N]$ ed $E[\lambda]$ sono in genere più facili da calcolare direttamente che $E[T]$.

3.1.6 Analisi operativa di una coda a servitore singolo

Prendiamo in considerazione una coda con un unico servitore e impostiamo l'analisi sullo studio delle caratteristiche di una particolare realizzazione $N^{(i)}(t)$ del processo stocastico $N(t)$ corrispondente al numero di clienti nella coda.

Una particolare realizzazione del processo $N(t)$ che rappresenta l'andamento del numero di clienti nel sistema in funzione del tempo t è mostrata nella figura 3.2. Consideriamo l'intervallo di tempo τ compreso tra due istanti θ_0 e θ_1 tali che (considerando che la realizzazione sia continua a destra)

$$N^{(i)}(\theta_0^-) = N^{(i)}(\theta_1^-) = 0, \quad N^{(i)}(\theta_0) = N^{(i)}(\theta_1) = 1 \quad (3.18)$$

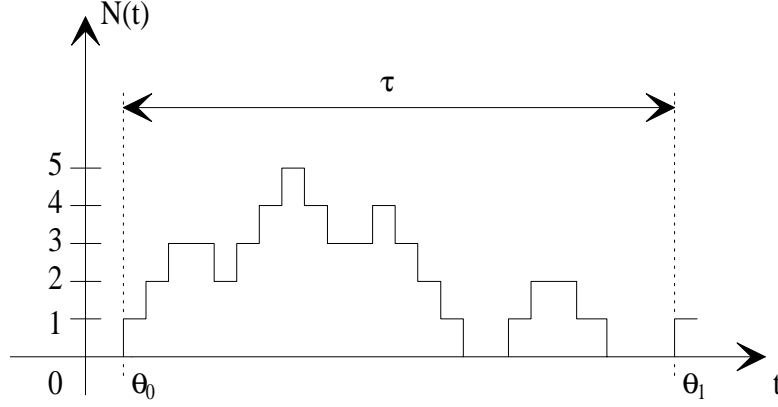


Figura 3.2. Una particolare realizzazione del processo di conteggio del numero di clienti in una coda

Per non appesantire la notazione, omettiamo nel seguito l'apice $^{(i)}$ che distingue la particolare realizzazione del processo presa in esame.

Definiamo

- A = numero totale di arrivi nell'intervallo τ
- D = numero totale di partenze nell'intervallo τ
- A_i = numero di arrivi che trovano i clienti nel sistema
- D_i = numero di partenze che lasciano $i - 1$ clienti nel sistema

Nell'intervallo τ si hanno A_i transizioni dallo stato i allo stato $i + 1$, e D_{i+1} transizioni dallo stato $i + 1$ allo stato i . Poiché lo stato all'inizio dell'intervallo è uguale allo stato alla fine dell'intervallo, il numero di volte che si è scesi dallo stato $i + 1$ deve essere uguale al numero di volte che si è saliti allo stato $i + 1$ ($D_{i+1} = A_i$).

Definiamo adesso

- τ_i = tempo totale trascorso nello stato i nell'intervallo τ
- p_i = frazione di tempo trascorsa nello stato i

e quindi

$$p_i = \frac{\tau_i}{\tau} \quad (3.19)$$

Notiamo che il seguente rapporto rimane costante per qualsiasi stato i

$$\frac{p_i}{\tau_i} = \frac{1}{\tau} \quad \forall i \quad (3.20)$$

Si possono allora scrivere le seguenti uguaglianze

$$A_i = D_{i+1} \quad (3.21)$$

$$\frac{1}{\tau} A_i = \frac{1}{\tau} D_{i+1} \quad (3.22)$$

$$\frac{p_i}{\tau_i} A_i = \frac{p_{i+1}}{\tau_{i+1}} D_{i+1} \quad (3.23)$$

$$p_i \frac{A_i}{\tau_i} = p_{i+1} \frac{D_{i+1}}{\tau_{i+1}} \quad (3.24)$$

Il rapporto

$$\frac{A_i}{\tau_i} = \lambda_i \quad (3.25)$$

rappresenta una velocità di nascita (arrivo) condizionata allo stato i .

Analogamente

$$\frac{D_{i+1}}{\tau_{i+1}} = \mu_{i+1} \quad (3.26)$$

rappresenta una velocità di morte (servizio) condizionata allo stato $i + 1$.

Possiamo quindi riscrivere la equazione (3.24) come

$$p_{i+1} = \frac{\lambda_i}{\mu_{i+1}} p_i \quad (3.27)$$

che definisce una relazione ricorrente che ammette la soluzione

$$p_i = p_0 \prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}} \quad (3.28)$$

Questa equazione è fondamentalmente identica a quella della catena di nascita e morte associata ad una coda $M/M/1$, anche se ha un significato diverso.

Proseguendo nello sviluppo si può ricavare una espressione analoga al risultato di Little.

Consideriamo la particolare realizzazione mostrata nella figura 3.2 del processo stocastico che descrive il numero di clienti nella coda a servitore singolo e definiamo

$$\begin{aligned} A(t) &= \text{numero di arrivi in } (\theta_0, \theta_0 + t) \\ D(t) &= \text{numero di partenze in } (\theta_0, \theta_0 + t) \end{aligned} \quad (3.29)$$

con la condizione iniziale $A(0^-) = D(0^-) = 0$.

Il numero di clienti presenti nel sistema al tempo $\theta_0 + t$ è dato dalla relazione

$$N(t) = A(t) - D(t) \quad (3.30)$$

Con le definizioni

$$\begin{aligned} a_j &= \text{istante di arrivo del cliente } j\text{-esimo} \\ d_j &= \text{istante di partenza del cliente } j\text{-esimo} \\ w_j &= \text{periodo di permanenza nella coda del cliente } j\text{-esimo} \end{aligned}$$

si ha

$$w_j = d_j - a_j \quad (3.31)$$

Definiamo il numero medio di clienti nel sistema e il tempo medio trascorso da un cliente nel sistema con le seguenti medie temporali

$$\bar{n} = \frac{1}{\tau} \int_{\theta_0}^{\theta_1} N(t) dt \quad (3.32)$$

$$\bar{\omega} = \frac{1}{A(\tau)} \sum_{j=1}^{A(\tau)} w_j \quad (3.33)$$

dove $A(\tau)$ è il numero totale di clienti arrivati (e partiti) nell'intervallo $\tau = \theta_1 - \theta_0$, di modo che $\bar{\omega}$ è la somma di tutti i periodi di permanenza dei singoli clienti nel sistema divisa per il numero di clienti.

Per ricavare il risultato di Little si deve definire la velocità media di arrivo dei clienti $\bar{\lambda}$

$$\bar{\lambda} = \frac{A(\tau)}{\tau} \quad (3.34)$$

ovvero il numero medio di clienti arrivati nell'unità di tempo.

Sostituendo $A(\tau) = \bar{\lambda}\tau$, nella (3.33) si ottiene

$$\bar{\omega} = \frac{1}{\bar{\lambda}\tau} \sum_{j=1}^{A(\tau)} w_j \quad (3.35)$$

Definiamo adesso la funzione $I_j(t)$ che indica la presenza del j -esimo cliente nel sistema

$$I_j(t) = \begin{cases} 1 & a_j < t \leq d_j \\ 0 & \text{altrimenti} \end{cases} \quad (3.36)$$

L'integrale di $I_j(t)$ nell'intervallo τ è uguale al periodo di permanenza del j -esimo cliente nel sistema

$$w_j = \int_{\theta_0}^{\theta_1} I_j(t) dt \quad (3.37)$$

Mediante la funzione $I_j(t)$ si può anche esprimere il numero di clienti nel sistema all'istante t come somma delle presenze in quell'istante

$$N(t) = \sum_{j=1}^{A(t)} I_j(t) \quad (3.38)$$

Al posto di $A(t)$ si può sostituire $A(\tau)$, in quanto la funzione di presenza dei clienti non ancora arrivati è zero

$$N(t) = \sum_{j=1}^{A(\tau)} I_j(t) \quad (3.39)$$

Possiamo ora riscrivere le (3.35) e (3.32) sostituendo le espressioni (3.37) e (3.39)

$$\bar{\omega} = \frac{1}{\bar{\lambda}\tau} \sum_{j=1}^{A(\tau)} \int_{\theta_0}^{\theta_1} I_j(t) dt \quad (3.40)$$

$$\bar{n} = \frac{1}{\tau} \int_{\theta_0}^{\theta_1} \sum_{j=1}^{A(\tau)} I_j(t) dt \quad (3.41)$$

Da queste due equazioni si vede immediatamente che

$$\bar{n} = \bar{w}\bar{\lambda} \quad (3.42)$$

Questo risultato è analogo al teorema di Little, pur essendo derivato con una analisi di tipo fenomenologico e non stocastico.

3.2 La coda $M/M/m$

La più semplice generalizzazione della coda $M/M/1$ consiste nel considerare un numero arbitrario di servitori (tipicamente maggiore di 1). Il modello che si ottiene è la coda $M/M/m$ mostrata nella figura 3.3.

Anche in questo caso il processo degli arrivi è Poisson con parametro λ e la distribuzione dei tempi di servizio è esponenziale con parametro μ . Come per la $M/M/1$, si suppone che la disciplina di coda sia FIFO.

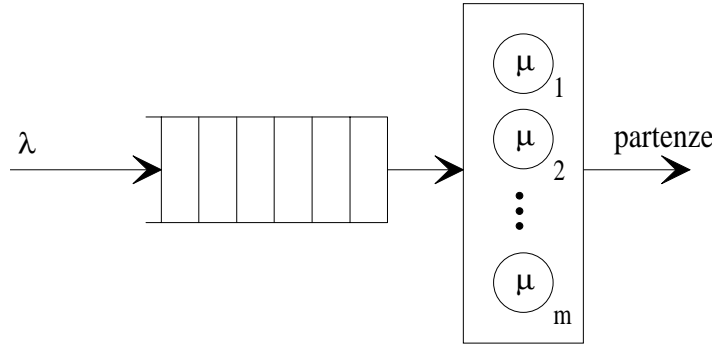


Figura 3.3. La coda $M/M/m$

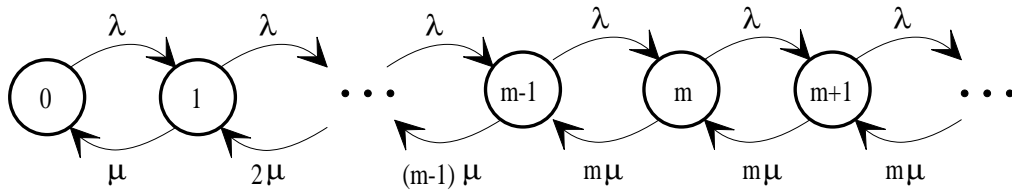


Figura 3.4. Diagramma delle velocità di transizione della catena di Markov tempo continuo associata alla coda $M/M/m$

Il diagramma delle velocità di transizione della catena di Markov N-M in tempo continuo associata alla coda $M/M/m$ è quello della figura 3.4. Anche in questo caso $S = \{0, 1, 2, \dots\}$.

La velocità di transizione dallo stato i allo stato $i - 1$ è pari a $i\mu$, per $i \leq m$, mentre è pari ad $m\mu$ per $i \geq m$. Questo deriva dal fatto che la diminuzione nel numero di clienti è prodotta dalla fine di uno tra i diversi servizi che procedono in parallelo. Il tempo che passa tra l'ingresso nello stato e la fine di servizio è quindi pari al minimo tra i tempi residui di servizio. Per la proprietà di assenza di memoria della distribuzione esponenziale dei tempi di servizio, il tempo residuo di servizio ha densità di probabilità uguale a quella di un intero tempo di servizio. Inoltre, l'indipendenza tra i tempi di servizio garantisce anche l'indipendenza dei tempi residui di servizio. È facile verificare che la distribuzione del minimo tra k tempi di servizio (o tempi residui di servizio) distribuiti esponenzialmente è quindi ancora esponenziale con parametro $k\mu$.

Dal bilanciamento dei flussi tra due stati qualsiasi adiacenti $i \leq m$ otteniamo

$$\pi_k = \frac{\lambda}{k\mu} \pi_{k-1} \quad k = 1, 2, \dots, m \quad (3.43)$$

da cui si ricava

$$\pi_k = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \pi_0 \quad k = 1, 2, \dots, m \quad (3.44)$$

Nella seconda parte della catena si ha invece

$$\pi_k = \frac{\lambda}{m\mu} \pi_{k-1} \quad k = m+1, m+2, \dots \quad (3.45)$$

che fornisce

$$\pi_{m+i} = \left(\frac{\lambda}{m\mu}\right)^i \pi_m \quad i \geq 1 \quad (3.46)$$

e quindi

$$\pi_j = \begin{cases} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} \pi_0 & j = 1, 2, \dots, m \\ \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!} \frac{1}{m^{j-m}} \pi_0 & j > m \end{cases} \quad (3.47)$$

Imponendo la condizione di normalizzazione, si ha

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^m \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} + \sum_{j=m+1}^{\infty} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!} \frac{1}{m^{j-m}}} \quad (3.48)$$

Il denominatore della (3.48) può essere riscritto come

$$1 + \sum_{j=1}^{m-1} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} + \sum_{j=m}^{\infty} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!} \frac{1}{m^{j-m}} \quad (3.49)$$

La seconda sommatoria converge a

$$\frac{m^m}{m!} \sum_{j=m}^{\infty} \left(\frac{\lambda}{m\mu}\right)^j = \frac{\left(\frac{\lambda}{m\mu}\right)^m}{\left(1 - \frac{\lambda}{m\mu}\right)} \frac{m^m}{m!} \quad (3.50)$$

sotto la condizione

$$\frac{\lambda}{m\mu} < 1 \quad (3.51)$$

che esprime la condizione di ergodicit  della catena.

La condizione sopra pu  essere riscritta come

$$\lambda \frac{1}{\mu} < m \quad (3.52)$$

Da questa riscrittura risulta evidente che la condizione dice che il prodotto della velocit  di arrivo λ per il tempo medio di servizio deve essere minore di m , ovvero della capacit  totale di servizio del sistema.

Il calcolo del numero medio di clienti in questo sistema risulta meno agevole del calcolo del tempo medio di permanenza nel sistema. Possiamo quindi ricavare direttamente il primo e poi ottenere il secondo dal teorema di Little.

Dal momento che un cliente   costretto ad attendere solo nel caso che tutti i servitori siano gi  occupati, si pu  scrivere

$$E[T] = E[T_s] + E[T_f] = \frac{1}{\mu} + \sum_{k=m}^{\infty} E[T_f \mid k] \pi_k^{(a)} \quad (3.53)$$

Osservando che il cliente considerato diventa primo in coda dopo che sono terminati $k - m$ servizi e entra in servizio dopo che   terminato un ulteriore servizio (cio  in totale dopo che sono terminati $k - m + 1$ servizi) ed utilizzando il teorema PASTA, si ricava

$$E[T] = \frac{1}{\mu} + \sum_{k=m}^{\infty} \frac{k - m + 1}{m\mu} \pi_k \quad (3.54)$$

Ora, dalla (3.46)

$$\begin{aligned} E[T] &= \frac{1}{\mu} + \sum_{k=m}^{\infty} \frac{k - m + 1}{m\mu} \pi_m \left(\frac{\lambda}{m\mu} \right)^{k-m} \\ &= \frac{1}{\mu} + \frac{\pi_m}{m\mu} \sum_{k=m}^{\infty} (k - m + 1) \left(\frac{\lambda}{m\mu} \right)^{k-m} \\ &= \frac{1}{\mu} + \frac{\pi_m}{m\mu} \sum_{i=1}^{\infty} i \left(\frac{\lambda}{m\mu} \right)^{i-1} \\ &= \frac{1}{\mu} + \frac{m\mu\pi_m}{(m\mu - \lambda)^2} \end{aligned}$$

E quindi, dal teorema di Little,

$$E[N] = \lambda E[T] = \frac{\lambda}{\mu} + \frac{m\lambda\mu\pi_m}{(m\mu - \lambda)^2} \quad (3.55)$$

3.2.1 Applicazione in ambiente telefonico

Il modello appena studiato può essere usato per descrivere in prima approssimazione il comportamento di una centrale telefonica

1. a cui fa capo una popolazione di utenti di grandi dimensioni,
2. che comprende m apparati per la gestione delle chiamate,
3. in cui le richieste che non possono essere immediatamente soddisfatte vengono inserite in una fila di attesa e soddisfatte appena possibile.

Nel modello $M/M/m$ i clienti rappresentano le chiamate ed i servitori rappresentano gli apparati della centrale.

Obiettivo di progetto di una centrale siffatta è scegliere il numero di apparati necessari per mantenere la probabilità che una chiamata venga posta in attesa entro limiti accettabili.

La probabilità che un cliente in arrivo venga posto nella fila di attesa vale

$$\begin{aligned}
 P\{\text{attesa}\} &= \sum_{k=m}^{\infty} \pi_k \\
 &= \sum_{k=m}^{\infty} \pi_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m!} \frac{m^m}{m^k} \\
 &= \frac{m^m}{m!} \pi_0 \sum_{k=m}^{\infty} \left(\frac{\lambda}{m\mu}\right)^k \\
 &= \pi_0 \frac{m^m}{m!} \frac{\left(\frac{\lambda}{m\mu}\right)^m}{1 - \frac{\lambda}{m\mu}} \\
 &= \pi_0 \frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{1 - \rho} \\
 &= \frac{\frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{1 - \rho}}{\sum_{k=0}^{m-1} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{1 - \rho}}
 \end{aligned}$$

dove $\rho = \frac{\lambda}{m\mu}$.

L'ultima espressione è una delle due formule di Erlang. Essa viene chiamata *Erlang 2* oppure *Erlang C* con parametri m e λ/μ e indicata con

$$E_{2,m}(\lambda/\mu) = C(m, \lambda/\mu) \quad (3.56)$$

3.3 Confronto tra code a servitore singolo e code a servitore multiplo

Vogliamo confrontare, a parità di velocità di arrivo dei clienti, pari richiesta di servizio e pari capacità complessiva di servizio tre sistemi a coda markoviani diversi, mostrati nelle figura 3.5, figura 3.6 e

figura 3.7.

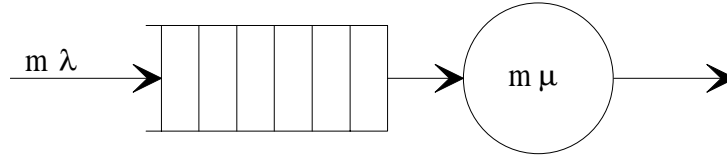


Figura 3.5. Sistema con unica fila di attesa e unico servitore

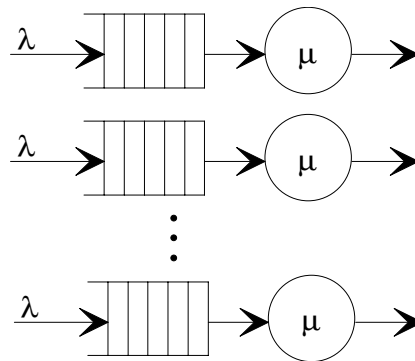


Figura 3.6. Sistema con m file di attesa separate ed m servitori

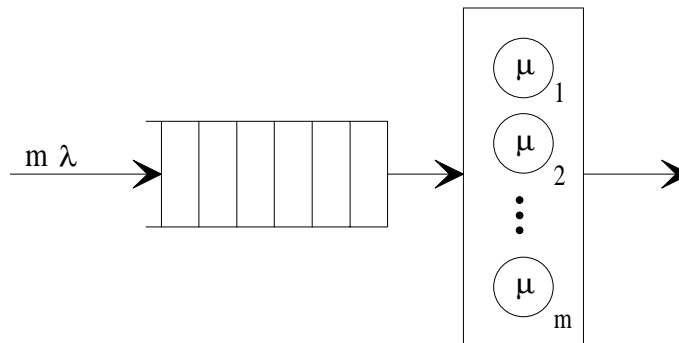


Figura 3.7. Sistema con unica fila di attesa ed m servitori

Vogliamo determinare quale dei tre sistemi dà le prestazioni migliori considerando come indice di qualità il tempo medio che un cliente trascorre nel sistema.

Conosciamo i risultati per i tre sistemi.

1. Nel primo caso il tempo medio di attesa è dato da

$$E[T]^{(1)} = \frac{1}{m\mu} \frac{1}{1 - \frac{m\lambda}{m\mu}}$$

$$= \frac{1}{m(\mu - \lambda)} \quad (3.57)$$

2. Nel secondo caso

$$\begin{aligned} E[T]^{(2)} &= \frac{1}{\mu} \frac{1}{1 - \frac{\lambda}{\mu}} \\ &= \frac{1}{\mu - \lambda} \end{aligned} \quad (3.58)$$

3. Per il terzo caso

$$E[T]^{(3)} = \frac{1}{\mu} + \frac{\mu\pi_m}{m(\mu - \lambda)^2} \quad (3.59)$$

Si può osservare che tra il primo caso e il secondo la differenza è un fattore moltiplicativo pari ad m .

$$E[T]^{(1)} = \frac{1}{m} E[T]^{(2)} \quad (3.60)$$

Il terzo caso dà risultati intermedi tra quelli dei primi due: va meglio del secondo ma peggio del primo.

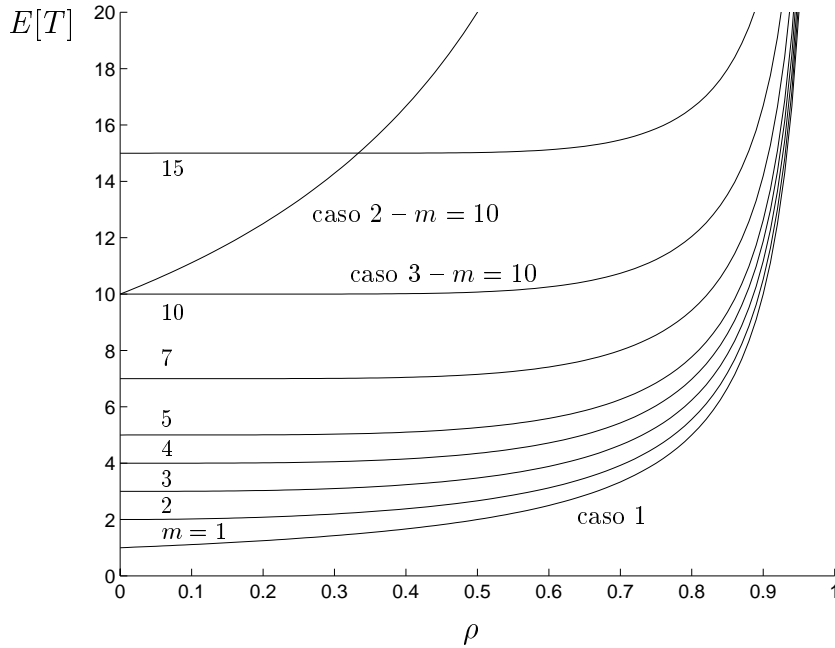


Figura 3.8. Grafico di $E[T]$ in funzione di ρ al variare di m

I risultati per i tre casi sono presentati nella figura 3.8, dove sono riportati i valori del ritardo medio per una coda $M/M/m$ in funzione del traffico $\rho = \lambda/(m\mu)$, per diversi valori di m e per capacità totale di servizio costante $m\mu = 1$.

Si può notare che, all'aumentare di m , le curve si appiattiscono per valori bassi di ρ , anche se per $\rho \rightarrow 1$ tutte le curve divergono.

Le curve danno direttamente i risultati per il terzo caso. Il primo caso è rappresentato dalla curva etichettata $m = 1$. I risultati per il secondo caso si ottengono moltiplicando per m i risultati della curva $m = 1$. Dal grafico si vede che il primo sistema è sempre il migliore, il secondo è il peggiore mentre il terzo dà risultati intermedi.

La giustificazione della differenza dei risultati sta nel fatto che il secondo sistema è tale per cui ci possono essere clienti in attesa in una coda anche quando esistono servitori disponibili (ad un'altra coda); ciò non può invece avvenire negli altri due casi. Inoltre, perché il terzo sistema possa impiegare tutta la sua capacità di servizio è necessario che nel sistema si trovino almeno m clienti; nel caso del primo sistema invece l'erogazione di servizio avviene sempre a piena capacità, anche in presenza di un solo cliente.

Come commento finale, possiamo notare che le considerazioni appena svolte dipendono dalla scelta dell'indice di qualità. Infatti è facile vedere dalla figura 3.8 che, considerando il tempo nella fila d'attesa T_f al posto del tempo di permanenza nel sistema T , valgono le disuguaglianze:

$$E[T_f]^{(3)} \leq E[T_f]^{(1)} \leq E[T_f]^{(2)} = \frac{1}{m}E[T_f]^{(1)}$$

Infatti il passaggio da $E[T]$ a $E[T_f]$ si ottiene sottraendo il tempo medio di servizio $E[t_s]$ (che è indipendente da ρ per ogni curva), in modo tale che le tre curve di $E[T_f]$ valgono 0 per $\rho = 0$.

3.4 La coda $M/M/\infty$

In una coda $M/M/\infty$ il numero di servitori è arbitrariamente grande e quindi è sempre disponibile un servitore per ogni cliente che arriva. Ciò comporta che

$$\begin{cases} \lambda_i = \lambda & i \geq 0 \\ \mu_i = i\mu & i \geq 1 \end{cases} \quad (3.61)$$

e quindi

$$\pi_k = \pi_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \quad k \geq 1 \quad (3.62)$$

infine

$$\begin{aligned} \pi_0 &= \frac{1}{1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}} \\ &= \frac{1}{\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}} \\ &= e^{-\lambda/\mu} \end{aligned} \quad (3.63)$$

Si noti che la sommatoria converge per qualsiasi valore finito di λ/μ e quindi la catena di Markov è sempre ergodica.

Sostituendo si ottiene:

$$\pi_k = \frac{(\lambda/\mu)^k}{k!} e^{-\lambda/\mu} \quad k \geq 0 \quad (3.64)$$

che costituisce una distribuzione di Poisson.

Il valor medio del numero di clienti presenti a regime nel sistema è dato da

$$\begin{aligned} E[N] &= E[\lambda] E[T] \\ &= \lambda \frac{1}{\mu} = \frac{\lambda}{\mu} \end{aligned} \quad (3.65)$$

poichè il tempo medio di permanenza nel sistema è ovviamente uguale al solo tempo medio di servizio.

3.5 La coda $M/M/m/0$

Consideriamo adesso una coda del tipo $M/M/m/0$. La differenza rispetto alla coda $M/M/m$ che abbiamo studiato in precedenza sta nel fatto che adesso non esiste alcuna fila di attesa: i clienti che arrivando trovano occupati tutti gli m servitori vengono persi, come indicato nella figura 3.9. Si noti che la perdita fa sì che il processo secondo cui i clienti entrano nella coda non è più un processo di Poisson.

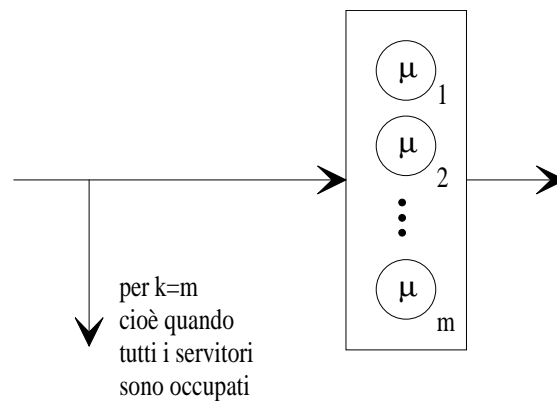


Figura 3.9. La coda $M/M/m/0$

Possiamo disegnare il diagramma delle velocità di transizione della catena di Markov associata a questa coda prendendo come stato il numero di clienti nel sistema, ottenendo il risultato della figura 3.10.

La catena di Markov così ottenuta è la versione troncata della catena associata ad una coda $M/M/m$, e precedentemente disegnata nella figura 3.4.

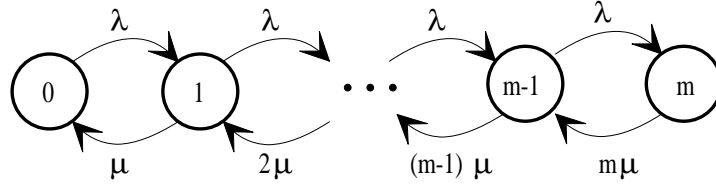


Figura 3.10. Diagramma delle velocità di transizione della catena di Markov associata alla coda $M/M/m/0$

Seguendo lo stesso procedimento già illustrato in precedenza si ricava

$$\pi_k = \pi_0 \left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!} \quad k = 1, 2, \dots, m \quad (3.66)$$

con

$$\pi_0 = \frac{1}{\sum_{k=0}^m \left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!}} \quad (3.67)$$

La catena di Markov è stazionaria ed irriducibile e quindi ergodica sotto la condizione

$$0 < \frac{\lambda}{\mu} < \infty \quad (3.68)$$

ovvero

$$\begin{cases} 0 < \mu < \infty \\ 0 < \lambda < \infty \end{cases} \quad (3.69)$$

Anche in questo caso il tempo medio di permanenza nel sistema è pari al tempo medio di servizio e quindi a μ^{-1} .

Per il calcolo del numero medio di clienti nella coda è più agevole utilizzare il teorema di Little piuttosto che seguire il procedimento diretto. A tal fine è necessario calcolare la velocità media di arrivo alla coda. Si può scrivere

$$E[\lambda] = \sum_{k=0}^{m-1} \lambda_k \pi_k = \lambda(1 - \pi_m) \quad (3.70)$$

Si ottiene quindi

$$E[N] = \frac{\lambda}{\mu}(1 - \pi_m) \quad (3.71)$$

3.5.1 Applicazione in ambiente telefonico

I risultati trovati possono essere utilizzati per lo studio di un sistema telefonico in cui le chiamate che trovano tutte le m apparecchiature di centrale occupate vengono perse (questo è il modo tipico di funzionamento dei sistemi telefonici elettromeccanici).

Per il progetto del sistema interessa conoscere la probabilità di perdita

$$\begin{aligned}
 P\{\text{perdita}\} &= \pi_m^{(a)} = \pi_m \\
 &= \frac{\left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!}}{\sum_{k=0}^m \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}} \\
 &= E_{1,m}(\lambda/\mu) \tag{3.72} \\
 &= B(m, \lambda/\mu) \tag{3.73}
 \end{aligned}$$

La formula ricavata è detta Erlang 1 o anche Erlang B.

3.6 La coda $M/M/1//N$

La coda $M/M/1//N$ differisce dalla coda $M/M/1$ per il fatto che la popolazione di clienti è finita (nonostante ciò la fila di attesa viene supposta illimitata; ciò è equivalente a supporre che la fila di attesa possa comprendere fino a $N - 1$ clienti).

Il processo degli arrivi non è più un processo di Poisson a tasso costante. Se così fosse succederebbe che, indipendentemente dal numero di clienti già presenti nella coda, ne possono arrivare sempre degli altri, ma ciò contrasta con l'ipotesi di popolazione finita.

Si suppone allora che il tasso di arrivo sia proporzionale al numero di clienti che non si trovano nella coda

$$\lambda_i = (N - i)\lambda \quad 0 \leq i \leq N \tag{3.74}$$

I servizi hanno densità di probabilità esponenziale negativa per tutti i clienti, con parametro μ , di modo che

$$\mu_i = \mu \quad i \geq 1 \tag{3.75}$$

La velocità media di arrivo al sistema è data da

$$\begin{aligned}
 E[\lambda] &= \sum_{i=0}^N \lambda_i \pi_i \\
 &= \sum_{i=0}^{N-1} (N - i) \lambda \pi_i \tag{3.76}
 \end{aligned}$$

La distribuzione a regime del numero di clienti nella coda si ricava nel solito modo

$$\begin{aligned}
 \pi_k &= \pi_0 \left(\frac{\lambda}{\mu}\right)^k [N(N-1) \cdots (N-k+1)] \\
 &= \pi_0 \left(\frac{\lambda}{\mu}\right)^k \frac{N!}{(N-k)!} \quad k = 1, 2, \dots, N \tag{3.77}
 \end{aligned}$$

con

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^N \left(\frac{\lambda}{\mu}\right)^j \frac{N!}{(N-j)!}} \tag{3.78}$$

La catena di Markov associata alla coda è sempre ergodica tranne nei casi degeneri.

3.7 Le distribuzioni Erlang e iperesponenziale

Prima di studiare code con tempi di servizio e tra gli arrivi con distribuzioni più generali, prenderemo in considerazione due particolari classi di distribuzioni non esponenziali:

1. E_n , distribuzione Erlang di ordine n
2. H_n , distribuzione iperesponenziale di ordine n

Il vantaggio derivante dall'utilizzare le distribuzioni E_n ed H_n per i tempi di servizio in una coda $M/G/1$ deriva dal fatto che esse permettono un'analisi di tipo markoviano del sistema.

3.7.1 La distribuzione E_n

Vediamo come è possibile generare una variabile casuale con densità di probabilità Erlang a partire da variabili casuali con densità di probabilità esponenziale negativa.

Siano date n variabili casuali statisticamente indipendenti ed identicamente distribuite ϵ_i con densità di probabilità

$$f_{\epsilon_i}(t) = n\mu e^{-n\mu t} u(t) \quad (3.79)$$

e quindi con

$$E[\epsilon_i] = \frac{1}{n\mu} \quad (3.80)$$

$$\sigma_{\epsilon_i}^2 = \frac{1}{(n\mu)^2} \quad (3.81)$$

$$C_{\epsilon_i}^2 \triangleq \frac{\sigma_{\epsilon_i}^2}{E[\epsilon_i]^2} = 1 \quad (3.82)$$

dove $E[\epsilon_i]$ è il valor medio di ϵ_i , $\sigma_{\epsilon_i}^2$ è la varianza e C_{ϵ_i} si chiama *coefficiente di variazione*, ed esprime la radice del rapporto tra la varianza ed il quadrato del valor medio.

Consideriamo ora la variabile casuale ϵ ottenuta come somma delle ϵ_i

$$\epsilon = \sum_{i=1}^n \epsilon_i \quad (3.83)$$

È possibile mostrare che la densità di probabilità della variabile casuale ϵ è una Erlang di ordine n a valor medio $1/\mu$. Tale ddp viene detta anche Erlang- n (E_n) ed ha espressione

$$f_{\epsilon}^{(n)}(t) = \frac{n\mu(n\mu t)^{n-1}}{(n-1)!} e^{-n\mu t} u(t) \quad (3.84)$$

Una variabile casuale con densità di probabilità E_n si può pensare che sia ottenuta come somma di n variabili casuali statisticamente indipendenti e identicamente distribuite con densità di probabilità esponenziale unilatera come in (3.79). La somma di variabili casuali indipendenti ha densità di probabilità pari alla convoluzione delle densità di probabilità. Per esempio

$$\begin{aligned} f_{\epsilon}^{(2)}(t) &= f_{\epsilon_1}(t) * f_{\epsilon_2}(t) = \\ &= u(t) \int_0^t n\mu e^{-n\mu\tau} n\mu e^{-n\mu(t-\tau)} d\tau = n\mu n\mu e^{-n\mu t} t u(t) \end{aligned}$$

Analogamente

$$\begin{aligned}
 f_{\epsilon}^{(3)}(t) &= f_{\epsilon_3}(t) * f_{\epsilon}^{(2)}(t) = \\
 &= u(t) \int_0^t n\mu \, n\mu \, e^{-n\mu\tau} \, \tau \, n\mu \, e^{-n\mu(t-\tau)} \, d\tau = \\
 &= \frac{n\mu(n\mu t)^2}{2} e^{-n\mu t} u(t)
 \end{aligned}$$

Si noti che per $n = 1$, E_n si riduce alla densità di probabilità esponenziale negativa con parametro μ .

Per la variabile casuale ϵ si ha che

$$E[\epsilon] = \frac{1}{\mu} \quad \forall n \quad (3.85)$$

$$\sigma_{\epsilon}^2 = \frac{1}{n\mu^2} \quad (3.86)$$

$$C_{\epsilon}^2 \triangleq \frac{\sigma_{\epsilon}^2}{E[\epsilon]^2} = \frac{1}{n} \quad (3.87)$$

Il coefficiente di variazione C è un parametro importante per la caratterizzazione di una densità di probabilità. Dai risultati ottenuti si può osservare che le densità di probabilità Erlang hanno coefficienti di variazione decrescenti al crescere di n . Ciò significa che per valori molto alti di n la variabile casuale ha dispersione dei valori molto piccola intorno al valor medio e quindi tende ad una costante quando n tende ad infinito.

Dalle considerazioni appena svolte consegue che un servitore con tempi di servizio distribuiti secondo una E_2 può essere rappresentato come nella figura 3.11. La figura evidenzia che il servizio è composto da due stadi separati in serie che hanno densità di probabilità esponenziale negativa. È da notare che i due stadi non sono due servitori, indipendenti perché non è possibile che sia in servizio più di un cliente alla volta.

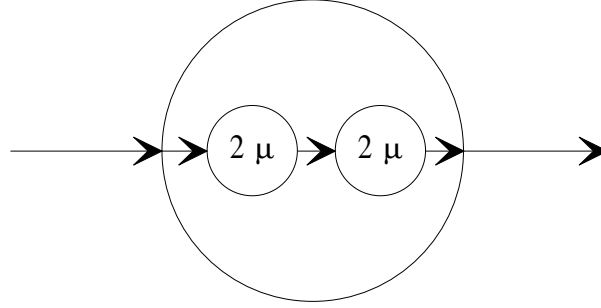


Figura 3.11. Servitore di una $M/E_2/1$

3.7.2 La distribuzione H_n

La densità di probabilità iperesponenziale di ordine n , H_n , è del tipo

$$f_{\eta}(t) = \sum_{i=1}^n \alpha_i \mu_i e^{-\mu_i t} u(t) \quad (3.88)$$

con

$$\sum_{i=1}^n \alpha_i = 1 \quad (3.89)$$

Una variabile casuale con densità di probabilità iperesponenziale può quindi essere considerata una somma pesata di variabili casuali con densità di probabilità esponenziale negativa.

Per i nostri scopi è sufficiente studiare la densità di probabilità H_2 , che è caratterizzata dal parametro α e dal rapporto μ_1/μ_2

$$f_\eta(t) = \alpha\mu_1 e^{-\mu_1 t} + (1 - \alpha)\mu_2 e^{-\mu_2 t} \quad (3.90)$$

Uno schema descrittivo del funzionamento del servitore in termini di stadi esponenziali di servizio può essere del tipo nella figura 3.12. Un cliente che arriva sceglie o l'uno o l'altro dei due stadi di servizio in parallelo, caratterizzati dai due parametri μ_1 e μ_2 .

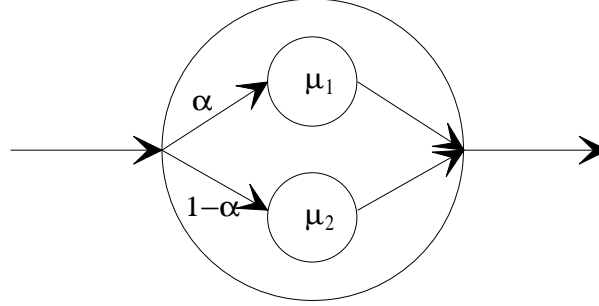


Figura 3.12. Servitore di una $M/H_2/1$

Il valor medio della variabile casuale vale

$$E[\eta] = \frac{\alpha}{\mu_1} + \frac{1 - \alpha}{\mu_2} \quad (3.91)$$

Il valor quadratico medio vale

$$E[\eta^2] = \frac{2\alpha}{\mu_1^2} + \frac{2(1 - \alpha)}{\mu_2^2} \quad (3.92)$$

e quindi la varianza

$$\sigma_\eta^2 = E[\eta^2] - E^2[\eta] \quad (3.93)$$

Per quanto riguarda il coefficiente di variazione, si ricava

$$C_\eta^2 = \frac{2 \left(\frac{\alpha}{\mu_1^2} + \frac{1-\alpha}{\mu_2^2} \right)}{\left(\frac{\alpha}{\mu_1} + \frac{1-\alpha}{\mu_2} \right)^2} - 1 \quad (3.94)$$

È possibile verificare che per le distribuzioni iperesponenziali C_η^2 assume valori maggiori dell'unità.

In pratica, conoscendo il valor medio ed il coefficiente di variazione desiderato per la distribuzione, si usa α per dimensionare C_η^2 mentre con il rapporto μ_1/μ_2 si dimensiona il valor medio.

3.7.3 La coda $M/E_2/1$

Nella coda $M/E_2/1$ gli arrivi seguono un processo di Poisson a parametro λ , mentre i tempi di servizio hanno densità di probabilità di tipo E_2 con valor medio μ^{-1} . Il servizio è quindi composto da due stadi esponenziali in serie come nella figura 3.13, e i due stadi hanno una velocità di servizio pari a 2μ .

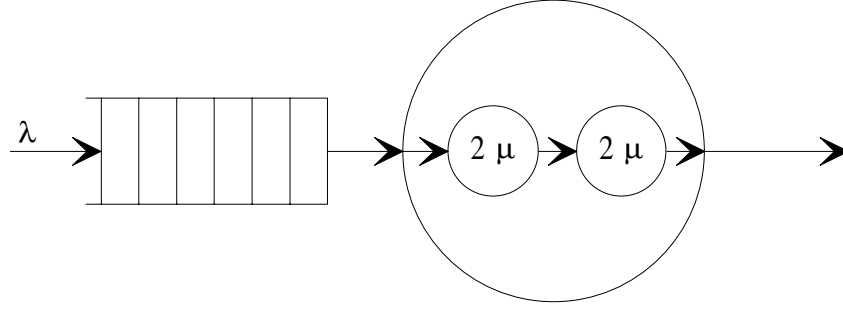


Figura 3.13. La coda $M/E_2/1$

Si desidera una definizione dello stato del sistema tale per cui sia possibile un'analisi di tipo markoviano. Se si sceglie come stato del sistema il numero di clienti nella coda o in servizio il comportamento non è di tipo markoviano a causa della memoria relativa ai tempi di servizio. Tale memoria è legata allo stadio attuale del servizio (primo o secondo).

Si può allora includere l'informazione che caratterizza la memoria del sistema nello stato, definendo lo stato mediante una coppia di valori che tengano conto sia del numero di clienti che si trovano o nella fila di attesa o in servizio sia dello stadio del servizio in corso. Così facendo si ha la possibilità di identificare la coda con un modello markoviano il cui diagramma delle velocità di transizione è presentato nella figura 3.14.

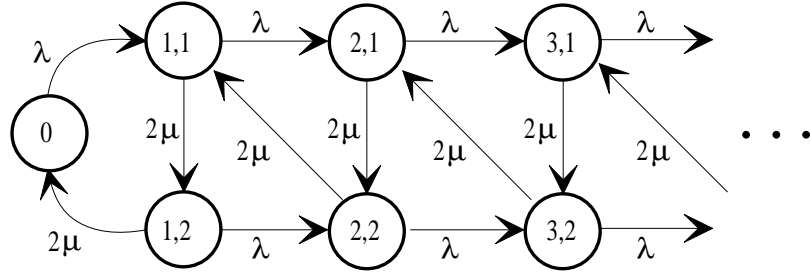


Figura 3.14. Catena di Markov associata alla coda $M/E_2/1$

Le equazioni che si derivano bilanciando i flussi entranti ed uscenti per ciascun stato sono

$$\begin{cases} \lambda \pi_0 &= 2\mu \pi_{1,2} \\ (\lambda + 2\mu) \pi_{1,1} &= \lambda \pi_0 + 2\mu \pi_{2,2} \\ (\lambda + 2\mu) \pi_{i,1} &= \lambda \pi_{i-1,1} + 2\mu \pi_{i+1,2} & i \geq 2 \\ (\lambda + 2\mu) \pi_{1,2} &= 2\mu \pi_{1,1} \\ (\lambda + 2\mu) \pi_{i,2} &= \lambda \pi_{i-1,2} + 2\mu \pi_{i,1} & i \geq 2 \end{cases} \quad (3.95)$$

Questo sistema di equazioni permette di calcolare la distribuzione di regime del sistema nel caso in cui la catena sia ergodica. È possibile verificare che la condizione di ergodicità dipende come al solito dal traffico

$$\frac{\lambda}{\mu} < 1 \quad (3.96)$$

Si noti che la catena non è più del tipo nascita-morte; ne consegue che non è possibile ricavare una semplice soluzione in forma chiusa.

La soluzione numerica del sistema di equazioni che definisce la distribuzione di regime è possibile nel caso in cui il numero degli stati sia finito. Ciò avviene sia nel caso di fila di attesa di dimensione finita sia nel caso di popolazione finita di utenti.

È relativamente facile studiare code $M/E_n/1/k$ o $M/E_n/1//N$. A titolo di esempio si riportano i diagrammi delle velocità di transizione per le code $M/E_2/1/3$ e $M/E_2/1//4$ nella figura 3.15 e nella figura 3.16.

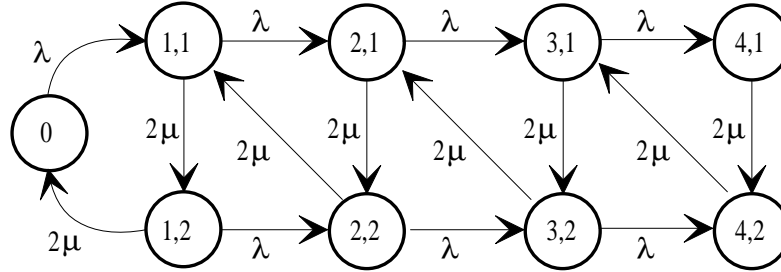


Figura 3.15. Catena di Markov associata ad una $M/E_2/1/3$

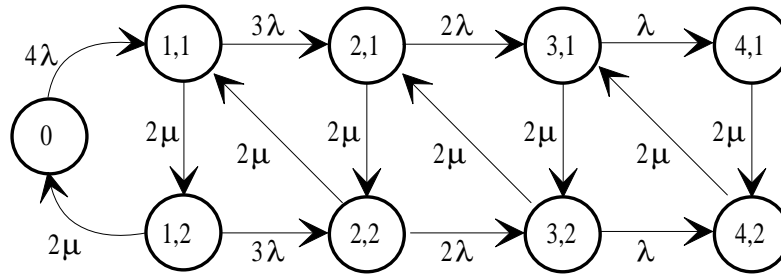


Figura 3.16. Catena di Markov associata ad una $M/E_2/1//4$

3.7.4 La coda $M/E_3/1$

Aggiungendo un terzo stadio al servizio otteniamo la coda $M/E_3/1$, il cui diagramma delle velocità di transizione è rappresentato nella figura 3.17.

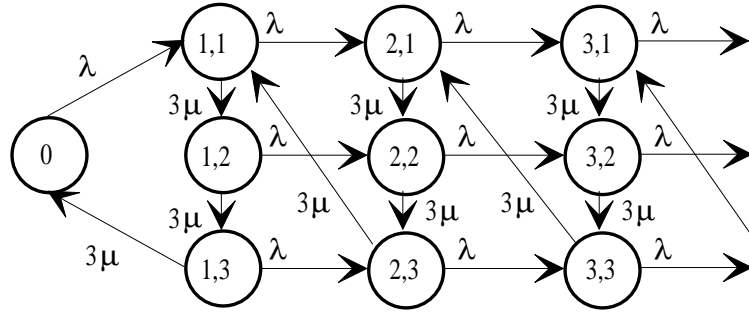


Figura 3.17. Catena di Markov associata ad una $M/E_3/1$

3.7.5 La coda $M/H_2/1$

Quando i tempi di servizio sono descritti mediante una variabile casuale con densità di probabilità iperesponenziale H_2 , è possibile rappresentare il servizio mediante 2 stadi in parallelo statisticamente indipendenti tra loro, ciascuno con tempi di servizio distribuiti secondo densità di probabilità esponenziali. La coda in questo caso è mostrata nella figura 3.18.

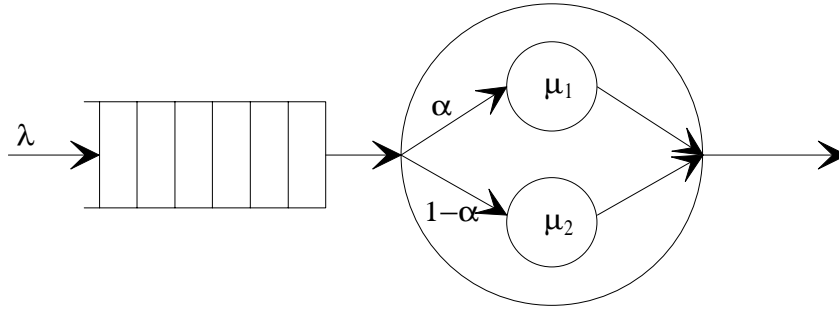


Figura 3.18. La coda $M/H_2/1$

Lo stato della catena di Markov che descrive il comportamento di questa coda tiene conto dei clienti in fila di attesa, nonché dello stadio di servizio che il cliente in servizio ha scelto.

Il diagramma delle velocità di transizione risultante è quello della figura 3.19.

3.7.6 Generalizzazioni

Generalizzando i risultati ricavati sopra è possibile risolvere code del tipo $M/G/m/k/n$ con $G \in \{E_n, H_n\}$.

Ulteriori estensioni permettono di risolvere anche code più complicate, del tipo $G/G/m/k/n$, sempre con $G \in \{E_n, H_n\}$. Infatti a una coda $E_2/M/1$, tale cioè per cui i tempi di interarrivo hanno distribuzione E_2 (quindi tra un arrivo e l'altro si interpongono due stadi esponenziali con velocità 2λ , come indicato nella figura 3.20) corrisponde una catena di Markov il cui diagramma delle velocità di transizione è riportato nella figura 3.21.

Lo strumento dell'espansione di una densità di probabilità in stadi esponenziali è molto potente e si può generalizzare a costruzioni miste serie-parallelo. È possibile verificare che tale espansione

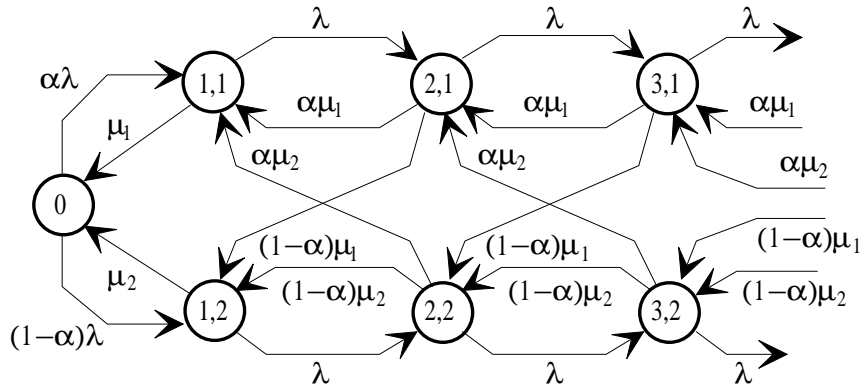


Figura 3.19. Catena di Markov associata ad una $M/H_2/1$

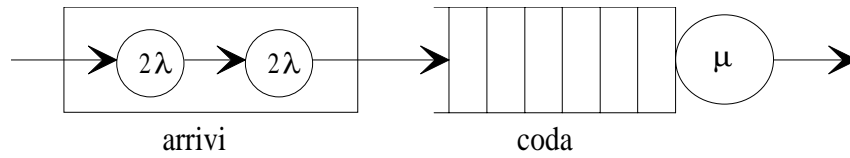


Figura 3.20. La coda $E_2/M/1$

in stadi esponenziali vale per qualsiasi densità di probabilità che abbia una funzione caratteristica razionale fratta.

3.8 La coda $M/G/1$

Una classe importante di modelli a coda fa riferimento a tempi di servizio con distribuzioni arbitrarie. Il più semplice sistema di questa classe è la coda $M/G/1$.

Nella figura 3.22 è illustrata una realizzazione del processo stocastico $N(t)$ che conta il numero di clienti nella coda. Tale realizzazione è simile a quella considerata nel caso della coda $M/M/1$,

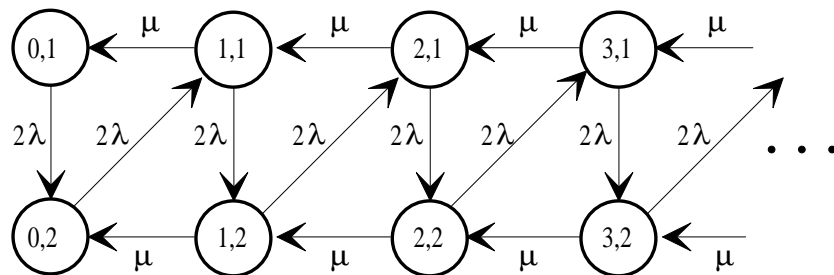


Figura 3.21. Catena di Markov associata ad una coda $E_2/M/1$

con l'unica differenza che adesso i tempi di servizio non sono più esponenzialmente distribuiti. Ciò comporta che $N(t)$ non sia più una catena di Markov a tempo continuo, a causa della memoria insita nella distribuzione dei tempi di servizio.

Il processo non è neppure semi-Markov, ma è ancora possibile individuare istanti di tempo in cui l'evoluzione futura del processo non dipende dal passato. Poiché la memoria è dovuta al tempo di servizio, questi istanti devono corrispondere con la fine o l'inizio di un servizio. Gli istanti che corrispondono alle partenze dei clienti dal sistema costituiscono quindi un insieme di istanti in cui non si ha memoria del passato.

Campionando lo stato del sistema in questi istanti di tempo si ottiene una catena di Markov a tempo discreto.

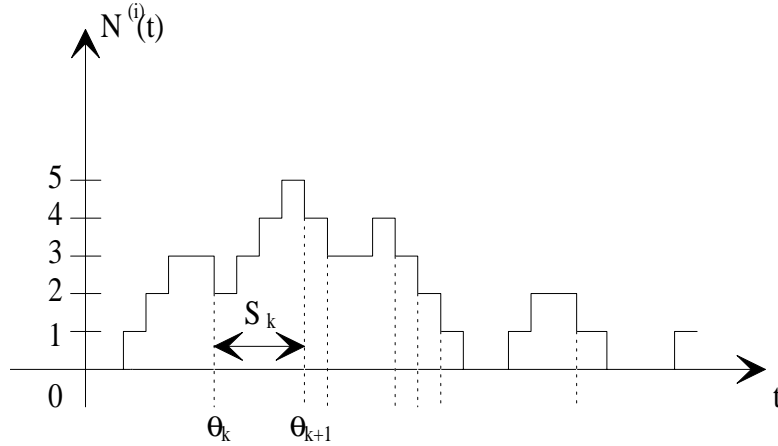


Figura 3.22. Particolare realizzazione del processo che conta il numero di clienti in una coda $M/G/1$

Usando le definizioni

$$Y_n \triangleq \begin{array}{l} \text{numero di clienti lasciati nel sistema} \\ \text{dall}'n\text{-esima partenza} \end{array} \quad (3.97)$$

$$Z_n \triangleq \text{numero di arrivi durante l}'n\text{-esimo servizio} \quad (3.98)$$

è possibile scrivere che

$$Y_{n+1} = \begin{cases} Y_n + Z_{n+1} - 1 & \text{se } Y_n > 0 \\ Z_{n+1} & \text{se } Y_n = 0 \end{cases} \quad (3.99)$$

La (3.99) può essere riscritta come

$$Y_{n+1} = Y_n + Z_{n+1} - u(Y_n) \quad (3.100)$$

Si noti che questa è una equazione operazionale derivata dall'osservazione della realizzazione del processo $N(t)$ ed è quindi valida per qualsiasi realizzazione del processo. Essa costituisce il punto di partenza per l'analisi della coda $M/G/1$.

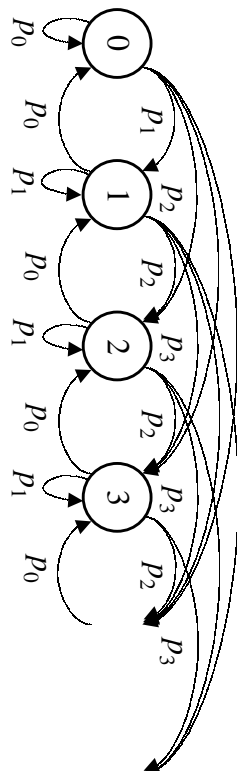


Figura 3.23. Catena interna per una coda $M/G/1$

La sequenza $\{Y_n, n = 0, 1, \dots\}$ è una catena di Markov a tempo discreto il cui diagramma delle probabilità di transizione è mostrato nella figura 3.23. Le sue probabilità di transizione sono

$$P\{Y_{n+1} = j \mid Y_n = i\} = P\{Z_{n+1} - u(Y_n) = j - i\} \quad (3.101)$$

e quindi

$$P\{Y_{n+1} = j \mid Y_n = i\} = \begin{cases} P\{Z_{n+1} - 1 = j - i\} & \text{se } i > 0 \\ P\{Z_{n+1} = j - i\} & \text{se } i = 0 \end{cases} \quad (3.102)$$

Ora, per il teorema della probabilità totale

$$P\{Z_n = k\} = \int_0^\infty P\{Z_n = k \mid S_n = \sigma\} f_{S_n}(\sigma) d\sigma \quad (3.103)$$

dove S_n è la durata dell' n -esimo servizio, mentre l'espressione di $P\{Z_n = k \mid S_n = \sigma\}$ è data dalla formula di Poisson, così che

$$P\{Z_n = k\} = \int_0^\infty \frac{(\lambda\sigma)^k}{k!} e^{-\lambda\sigma} f_{S_n}(\sigma) d\sigma \triangleq p_k \quad (3.104)$$

Per calcolare le p_k è necessario conoscere la densità di probabilità $f_{S_n}(\sigma)$. Si può però scrivere la matrice delle probabilità di transizione della catena di Markov a tempo discreto individuata

$$\mathbf{P} = \begin{bmatrix} p_0 & p_1 & p_2 & p_3 & \dots & p_k & \dots \\ p_0 & p_1 & p_2 & p_3 & \dots & p_k & \dots \\ 0 & p_0 & p_1 & p_2 & \dots & p_{k-1} & \dots \\ 0 & 0 & p_0 & p_1 & \dots & p_{k-2} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad (3.105)$$

La catena è sicuramente aperiodica perché gli elementi della diagonale principale della matrice \mathbf{P} sono in generale non nulli. La catena è anche omogenea ed irriducibile e quindi la distribuzione di regime esiste.

Tale distribuzione di regime risolve il problema relativo all'analisi della distribuzione del numero di clienti nella coda a regime, ma solo in istanti di tempi particolari, cioè in corrispondenza delle partenze dei clienti.

Normalmente interessa conoscere tale distribuzione in un istante qualsiasi. Si dimostra che, nel caso di code con variazioni unitarie del numero di clienti nel sistema (cioè nel caso di partenze e arrivi individuali), o di catene che godono di proprietà di reversibilità, la distribuzione ai tempi di

partenza è uguale alla distribuzione negli istanti di arrivo, a patto che sia verificata la condizione di equilibrio data da

$$\lambda E[S] < 1 \quad (3.106)$$

Inoltre, per il teorema PASTA, la distribuzione del numero di clienti nel sistema agli istanti di arrivo dei clienti coincide con la distribuzione ad istanti arbitrari.

Numero medio di clienti nella coda

Cerchiamo ora una espressione in forma chiusa per il numero medio di clienti in una coda $M/G/1$, $E[N]$, e per il tempo medio di permanenza di un cliente nella coda, $E[T]$.

Questi due parametri si possono ovviamente calcolare a partire dalla soluzione a regime della catena di Markov a tempo discreto vista in precedenza, ma si possono anche ricavare direttamente.

Prendendo la media della (3.100) otteniamo

$$E[Y_{n+1}] = E[Y_n] + E[Z_{n+1}] - E[u(Y_n)] \quad (3.107)$$

sappiamo inoltre che

$$\lim_{n \rightarrow \infty} E[Y_n] = \lim_{n \rightarrow \infty} E[Y_{n+1}] = E[Y] \quad (3.108)$$

e anche $E[Z_n] \rightarrow E[Z]$ e $E[u(Y_n)] \rightarrow E[u(Y)]$. Sostituendo otteniamo

$$E[Y] = E[Y] + E[Z] - E[u(Y)] \quad (3.109)$$

ovvero

$$E[Z] = E[u(Y)] \quad (3.110)$$

La funzione $u(Y)$ è definita da

$$u(Y) = \begin{cases} 0 & \text{coda vuota} \\ 1 & \text{coda non vuota} \end{cases} \quad (3.111)$$

In altri termini, $u(Y)$ indica se il servitore è occupato o meno, quindi possiamo scrivere che la probabilità che il servitore sia occupato è data da

$$P\{\text{servitore occupato}\} = E[u(Y)] = \rho = \lambda E[S] \quad (3.112)$$

Si noti che $E[u(Y)]$ rappresenta il carico del sistema a coda.

Per ricavare $E[Y]$, eleviamo al quadrato ambo i membri della (3.100)

$$Y_{n+1}^2 = Y_n^2 + Z_{n+1}^2 + u(Y_n) + 2Y_n Z_{n+1} - 2Y_n u(Y_n) - 2Z_{n+1} u(Y_n) \quad (3.113)$$

Osservando poi che $Y_n u(Y_n) = Y_n$, che Z_{n+1} e Y_n sono statisticamente indipendenti e prendendo la media della (3.113) si ha

$$\begin{aligned} E[Y_{n+1}^2] &= E[Y_n^2] + E[Z_{n+1}^2] + E[u(Y_n)] + 2E[Y_n]E[Z_{n+1}] - \\ &\quad - 2E[Y_n] - 2E[Z_{n+1}]E[u(Y_n)] \end{aligned} \quad (3.114)$$

Passando al limite per $n \rightarrow \infty$ e sostituendo il risultato della (3.110) si ottiene

$$0 = E[Z^2] + E[Z] + 2E[Y]E[Z] - 2E[Y] - 2E^2[Z] \quad (3.115)$$

$$2\{1 - E[Z]\}E[Y] = E[Z^2] + E[Z] - 2E^2[Z] \quad (3.116)$$

Infine, sommando e sottraendo $E[Z]$ a secondo membro e risolvendo per $E[Y]$ si ha

$$E[Y] = E[Z] + \frac{E[Z^2] - E[Z]}{2\{1 - E[Z]\}} \quad (3.117)$$

$E[Z]$ è già stata ricavata in precedenza come $E[Z] = \rho = \lambda E[S]$. Per ricavare una espressione in forma chiusa di $E[Y]$ è necessario ricavare ancora $E[Z^2]$. Si può scrivere

$$\begin{aligned} E[Z^2] &= \sum_{k=1}^{\infty} k^2 P\{Z = k\} \\ &= \sum_{k=1}^{\infty} k^2 \int_0^{\infty} \frac{(\lambda\sigma)^k}{k!} e^{-\lambda\sigma} f_S(\sigma) d\sigma \\ &= \int_0^{\infty} \sum_{k=1}^{\infty} k^2 \frac{(\lambda\sigma)^k}{k!} e^{-\lambda\sigma} f_S(\sigma) d\sigma \\ &= \int_0^{\infty} (\lambda^2 \sigma^2 + \lambda\sigma) f_S(\sigma) d\sigma \\ &= \lambda^2 E[S^2] + \lambda E[S] \end{aligned} \quad (3.118)$$

dato che il valor quadratico medio della distribuzione di Poisson con parametro α è pari a $\alpha^2 + \alpha$.

Sostituendo si ottiene

$$E[Y] = \rho + \frac{\lambda^2 E[S^2]}{2(1 - \rho)} \quad (3.119)$$

Ricordando che la distribuzione al momento di una partenza coincide con la distribuzione ad un istante arbitrario, quindi che $E[N] = E[Y]$, e avendo definito il coefficiente di variazione della distribuzione dei tempi di servizio C_S come

$$C_S^2 = \frac{E[S^2] - E^2[S]}{E^2[S]} \quad (3.120)$$

possiamo scrivere che

$$E[N] = \rho + \rho^2 \frac{1 + C_S^2}{2(1 - \rho)} \quad (3.121)$$

Quest'ultima è la formula di Pollaczek-Khintchin per il numero medio di clienti in una coda $M/G/1$.

Possiamo fare una verifica con il caso $M/M/1$, dove i tempi di servizio sono esponenzialmente distribuiti così che $C_S^2 = 1$. Sostituendo si ottiene

$$E[N] = \rho + \frac{\rho^2}{1 - \rho} = \frac{\rho}{1 - \rho} \quad (3.122)$$

che è l'espressione di $E[N]$ ricavata precedentemente per la $M/M/1$.

Usando la formula di Little si trova infine $E[T]$

$$E[T] = \frac{E[N]}{E[\lambda]} = \frac{E[Y]}{\lambda} \quad (3.123)$$

ovvero

$$E[T] = E[S] + \rho E[S] \frac{C_s^2 + 1}{2(1 - \rho)} \quad (3.124)$$

Quest'ultima è la formula di Pollaczek-Khintchin per il tempo medio di permanenza dei clienti in una coda $M/G/1$. Si noti che sia $E[N]$ sia $E[T]$ hanno una dipendenza quadratica dal coefficiente di variazione dei tempi di servizio a pari valori medi per tempi di servizio e di interarrivo. Quindi un incremento della “casualità” dei tempi di servizio porta ad un aumento della congestione, cioè di ritardi e numero di clienti in coda.

3.9 Reti di code

Una rete di code è un sistema che comprende più code interconnesse in modo arbitrario.

La più semplice rete di code è costituita da due code $M/M/1$ collegate in serie, come mostrato nella figura 3.24. Dall'esempio è immediato osservare che il processo degli arrivi alla seconda coda coincide con il processo delle partenze dalla prima coda. È quindi più corretto identificare la seconda coda con la notazione $\cdot/M/1$, per rendere esplicito il fatto che il processo degli arrivi per tale coda non può essere descritto indipendentemente dal resto della rete.

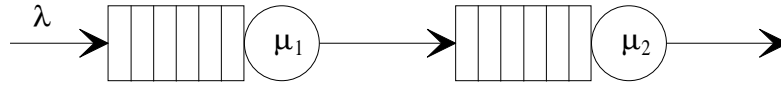


Figura 3.24. Due code $M/M/1$ in serie

Queste considerazioni rendono evidente l'importanza della caratterizzazione del processo delle partenze da una coda per lo studio delle reti di code.

3.9.1 Il teorema di Burke

Facendo sempre riferimento alla rete della figura 3.24, studiamo le caratteristiche probabilistiche del processo delle partenze dalla prima coda.

Consideriamo a tal fine una particolare realizzazione del processo $N(t)$, simile a quella nella figura 3.22, dove gli arrivi sono identificati dalle discontinuità con incremento di $N(t)$, mentre le partenze sono identificate dalle discontinuità con decremento di $N(t)$.

L'intervallo di tempo tra due partenze consecutive è rappresentato da una variabile casuale continua θ . Interessa calcolare la funzione distribuzione cumulativa di θ : $P\{\theta \leq t\}$.

Per svolgere i calcoli consideriamo separatamente i due casi in cui il numero di clienti lasciato nella coda dalla partenza all'inizio dell'intervallo sia uguale o diverso da zero. Applicando il teorema della probabilità totale si ha

$$\begin{aligned} P\{\theta \leq t\} &= P\{\theta \leq t \mid \text{la partenza lascia } 0 \text{ clienti}\} \pi_0^{(d)} \\ &\quad + P\{\theta \leq t \mid \text{la partenza lascia } \geq 1 \text{ clienti}\} (1 - \pi_0^{(d)}) \end{aligned}$$

dove $\pi_i^{(d)} = P\{\text{che una partenza lasci } i \text{ clienti nella coda}\}$.

Nel primo caso l'intervallo tra due partenze successive è pari alla somma del tempo residuo fino al prossimo arrivo più il tempo di servizio del cliente appena arrivato, come si vede dalla figura 3.25. Chiamando S il tempo di servizio di un cliente e τ_r il tempo residuo fino al prossimo arrivo, abbiamo $\theta = S + \tau_r$. Nel secondo caso si ha semplicemente $\theta = S$. Quindi

$$P\{\theta \leq t\} = P\{S + \tau_r \leq t\}\pi_0^{(d)} + P\{S \leq t\}(1 - \pi_0^{(d)})$$

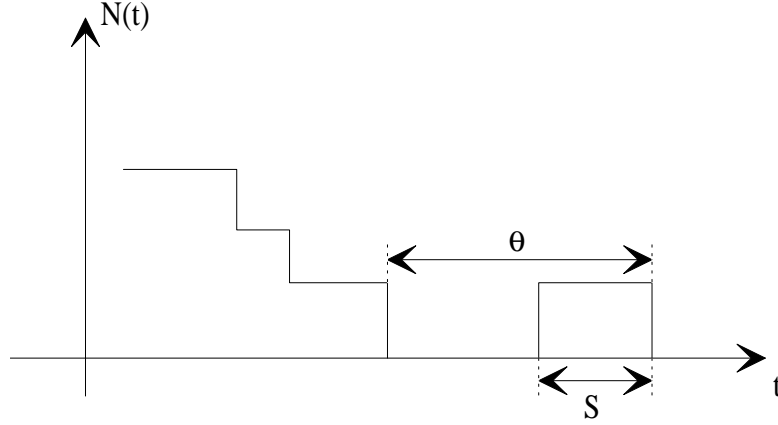


Figura 3.25. Tempo di interpartenza susseguente ad una partenza che lascia il sistema vuoto

Dal momento che i tempi di interarrivo sono distribuiti esponenzialmente, per la assenza di memoria della densità di probabilità esponenziale, anche il tempo residuo fino al prossimo arrivo è distribuito esponenzialmente. Le densità di probabilità delle due variabili casuali S e τ_r sono quindi le seguenti

$$f_s(t) = \mu e^{-\mu t} u(t) \quad (3.125)$$

$$f_{\tau_r}(t) = \lambda e^{-\lambda t} u(t) \quad (3.126)$$

Per l'indipendenza statistica tra le due variabili casuali S e τ_r

$$f_{S+\tau_r}(t) = f_S(t) * f_{\tau_r}(t) \quad (3.127)$$

dove l'asterisco indica convoluzione.

Possiamo allora scrivere

$$f_\theta(t) = \pi_0^{(d)} f_S(t) * f_{\tau_r}(t) + (1 - \pi_0^{(d)}) f_S(t) \quad (3.128)$$

Abbiamo già visto che nel caso di un processo degli arrivi di Poisson, la probabilità che un arrivo trovi i clienti nella coda è uguale alla probabilità che in un istante qualsiasi nella coda si trovino i clienti (teorema PASTA: $\pi_i^{(a)} = \pi_i$) e che, nel caso di code con variazioni unitarie del numero di clienti nel sistema (cioè nel caso di partenze e arrivi individuali), o di catene che godono di proprietà di reversibilità, la probabilità che una partenza lasci i clienti nella coda è uguale alla probabilità che un arrivo trovi i clienti nella coda ($\pi_i^{(a)} = \pi_i^{(d)}$). Quindi

$$\pi_0^{(d)} = 1 - \frac{\lambda}{\mu} \quad (3.129)$$

$$1 - \pi_0^{(d)} = \frac{\lambda}{\mu} \quad (3.130)$$

Otteniamo quindi

$$\begin{aligned}
f_\theta(t) &= \left(1 - \frac{\lambda}{\mu}\right) u(t) \int_0^t \lambda e^{-\lambda\tau} \mu e^{-\mu(t-\tau)} d\tau + \frac{\lambda}{\mu} \mu e^{-\mu t} u(t) = \\
&= \frac{\mu - \lambda}{\mu} \lambda \mu e^{-\mu t} u(t) \int_0^t e^{(\mu-\lambda)\tau} d\tau + \lambda e^{-\mu t} u(t) = \\
&= \lambda e^{-\mu t} \frac{\mu - \lambda}{\mu - \lambda} [e^{(\mu-\lambda)t} - 1] u(t) + \lambda e^{-\mu t} u(t) = \lambda e^{-\lambda t} u(t)
\end{aligned}$$

cioè che i tempi tra le partenze sono distribuiti esponenzialmente con parametro λ .

Per completare la dimostrazione che il processo delle partenze sia un processo di Poisson con tasso costante λ si deve ancora dimostrare l'indipendenza statistica tra due tempi di partenza consecutivi. Questa dimostrazione non viene qui riportata, ma è possibile con semplici considerazioni. Ne consegue che il processo delle partenze da una coda $M/M/1$ è un processo di Poisson a parametro costante λ .

Questo risultato si può generalizzare al caso in cui ci siano più servitori, e quindi a code $M/M/m$.

Il risultato generale è noto con il nome di **Teorema di Burke**:

Il processo delle partenze da una coda $M/M/m$ in condizioni di regime è un processo di Poisson con lo stesso parametro del processo in ingresso.

3.9.2 Due code $M/M/1$ in serie

Ritorniamo a studiare la rete di due code in serie della figura 3.24, definendo come stato del sistema la coppia di valori data dal numero di clienti in ciascuna coda: (k_1, k_2) .

Possiamo subito osservare che la velocità di arrivo ad entrambe le code ha valore λ . Quindi il sistema raggiunge una situazione di equilibrio se sono verificate entrambe le condizioni

$$\frac{\lambda}{\mu_1} < 1 \quad \frac{\lambda}{\mu_2} < 1 \quad (3.131)$$

cioè deve essere

$$\lambda < \min(\mu_1, \mu_2) \quad (3.132)$$

Lo scopo dello studio è calcolare la probabilità a regime di avere k_1 clienti alla prima coda e k_2 clienti alla seconda coda, π_{k_1, k_2} .

A tale risultato si può arrivare risolvendo direttamente la catena di Markov tempo continua che descrive l'evoluzione dello stato del sistema. Il diagramma delle velocità di transizione di tale catena è mostrato nella figura 3.26.

Dal diagramma delle velocità di transizione si ricavano le seguenti equazioni

$$\begin{cases}
\lambda \pi_{0,0} &= \mu_2 \pi_{0,1} \\
(\lambda + \mu_2) \pi_{0,k_2} &= \mu_1 \pi_{1,k_2-1} + \mu_2 \pi_{0,k_2+1} & k_2 \geq 1 \\
(\lambda + \mu_1) \pi_{k_1,0} &= \lambda \pi_{k_1-1,0} + \mu_2 \pi_{k_1,1} & k_1 \geq 1 \\
(\lambda + \mu_1 + \mu_2) \pi_{k_1,k_2} &= \lambda \pi_{k_1-1,k_2} + \mu_1 \pi_{k_1+1,k_2-1} + \\
&\quad + \mu_2 \pi_{k_1,k_2+1} & k_1, k_2 \geq 1
\end{cases} \quad (3.133)$$

Si può verificare facilmente per sostituzione che le equazioni trovate sono soddisfatte dalla soluzione

$$\pi_{k_1, k_2} = \left(1 - \frac{\lambda}{\mu_1}\right) \left(1 - \frac{\lambda}{\mu_2}\right) \left(\frac{\lambda}{\mu_1}\right)^{k_1} \left(\frac{\lambda}{\mu_2}\right)^{k_2} \quad (3.134)$$

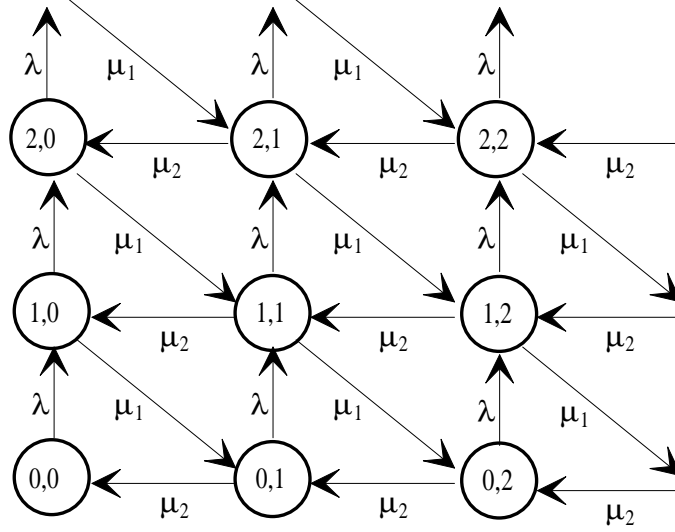


Figura 3.26. Diagramma delle velocità di transizione della catena di Markov tempo continua che modella due code $M/M/1$ in serie

Il medesimo risultato si ottiene osservando che, siccome il processo delle partenze dalla seconda coda è un processo di Poisson a velocità λ (per il teorema di Burke), la rete comprende due code $M/M/1$ che si comportano come se fossero in isolamento, di modo che la soluzione cercata può essere fattorizzata nelle soluzioni delle due code

$$\pi_{k_1, k_2} = \pi_{k_1} \pi_{k_2} \quad (3.135)$$

3.9.3 Reti di code acicliche

Il risultato trovato sopra per il sistema formato da due code in serie può essere generalizzato a reti con più di due code di tipo $M/M/m$, arrivando a caratterizzare una prima classe di reti per cui le probabilità di regime congiunte di avere k_i clienti alla coda i , con i che prende tutti i valori associati alle varie code, fattorizzano nel prodotto delle probabilità marginali alle varie code.

La definizione della classe di reti di code si basa sul fatto che i processi di arrivo alle varie code devono essere Poisson per potere studiare le code come se fossero in isolamento.

Il risultato generale cui si arriva è il seguente:

*tutte le reti di code che comprendono code (stazioni) del tipo $\cdot/M/m$, che ricevono arrivi sia da altre code della rete, sia dall'esterno secondo processi di Poisson a tasso costante e che non permettono a nessun cliente di ritornare ad una stazione già visitata (quindi senza cicli interni), sono risolubili in forma **prodotto**.*

L'assenza di cicli è essenziale per preservare la caratteristica di Poisson dei processi di arrivo. Infatti, anche per una coda $M/M/1$ con ritorno, come nella figura 3.27, il processo delle uscite non è più un processo di Poisson. Si ricordano due importanti proprietà dei processi di Poisson.

- L'unione di eventi di due processi di Poisson indipendenti con parametri λ_1 e λ_2 genera un processo di Poisson con parametro $\lambda_1 + \lambda_2$.

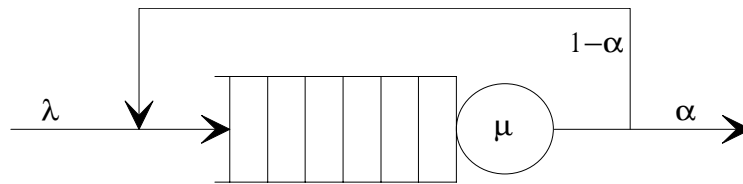


Figura 3.27. Coda con ritorno

- La suddivisione secondo scelte casuali indipendenti degli eventi di un processo di Poisson genera sottoprocessi di Poisson indipendenti. In particolare, se arrivi di clienti secondo un processo di Poisson a parametro λ vengono suddivisi tra due sistemi a coda inviando ogni cliente con probabilità α alla coda 1 e con probabilità $(1 - \alpha)$ alla coda 2, i processi di arrivo alle due code sono processi di Poisson tra di loro indipendenti (ma dipendenti dal processo originale) con parametri $\lambda\alpha$ e $\lambda(1 - \alpha)$.

Un esempio di rete di code aciclica che può essere risolta in forma prodotto è mostrato nella figura 3.28.

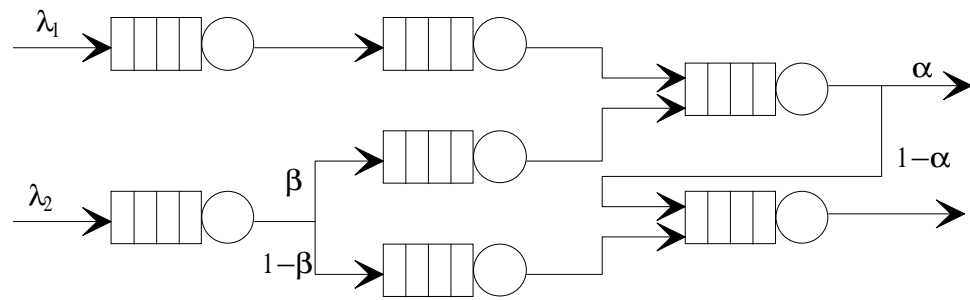


Figura 3.28. Esempio di rete di code aciclica

3.9.4 Reti di code di Jackson

La condizione di non avere cicli è assai restrittiva per la costruzione di modelli di sistemi di comunicazione. È stato però dimostrato che esistono anche reti con cicli che ammettono soluzione sotto forma prodotto.

Consideriamo la rete di due code $M/M/1$ in serie con ciclo mostrata nella figura 3.29.

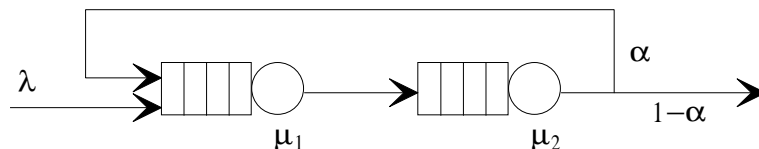


Figura 3.29. Due code $M/M/1$ collegate in serie e con ciclo

Il diagramma delle velocità di transizione della catena di Markov tempo continua derivata dal modello della figura 3.29 è mostrato nella figura 3.30.

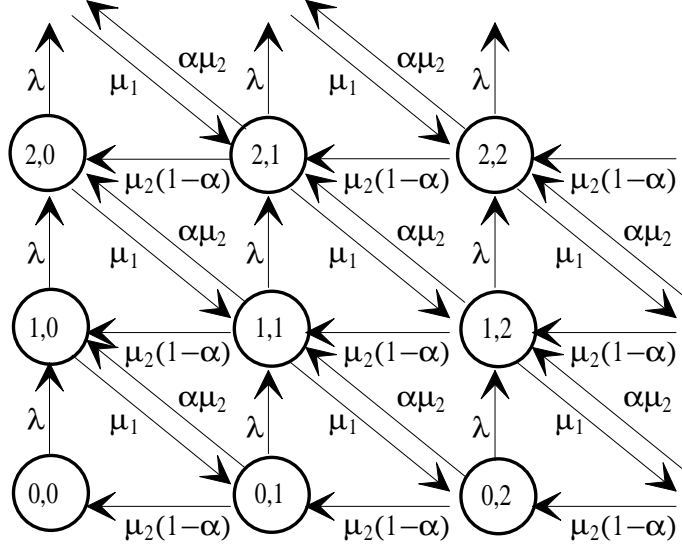


Figura 3.30. Diagramma delle velocità di transizione della catena di Markov tempo continua che modella due code $M/M/1$ in serie con ciclo

Dal diagramma delle velocità di transizione si ricavano le seguenti equazioni

$$\begin{cases} \lambda \pi_{0,0} &= \mu_2(1-\alpha) \pi_{0,1} \\ (\lambda + \mu_2) \pi_{0,k_2} &= \mu_1 \pi_{1,k_2-1} + \mu_2(1-\alpha) \pi_{0,k_2+1} & k_2 \geq 1 \\ (\lambda + \mu_1) \pi_{k_1,0} &= \lambda \pi_{k_1-1,0} + (1-\alpha) \mu_2 \pi_{k_1,1} + \alpha \mu_2 \pi_{k_1-1,1} & k_1 \geq 1 \\ (\lambda + \mu_1 + \mu_2) \pi_{k_1,k_2} &= \lambda \pi_{k_1-1,k_2} + \mu_1 \pi_{k_1+1,k_2-1} \\ &\quad + (1-\alpha) \mu_2 \pi_{k_1,k_2+1} + \alpha \mu_2 \pi_{k_1-1,k_2+1} & k_1, k_2 \geq 1 \end{cases} \quad (3.136)$$

Si può verificare facilmente per sostituzione che le equazioni trovate sono soddisfatte nuovamente dalla soluzione

$$\pi_{k_1,k_2} = \left(1 - \frac{\lambda_1}{\mu_1}\right) \left(1 - \frac{\lambda_2}{\mu_2}\right) \left(\frac{\lambda_1}{\mu_1}\right)^{k_1} \left(\frac{\lambda_2}{\mu_2}\right)^{k_2} \quad (3.137)$$

dove

$$\lambda_1 = \lambda + \alpha \lambda_2 \quad (3.138)$$

è la velocità totale di arrivo alla prima coda e

$$\lambda_2 = \lambda_1 \quad (3.139)$$

è la velocità totale di arrivo alla seconda coda. Si ricava

$$\lambda_1 = \lambda_2 = \frac{\lambda}{1-\alpha} \quad (3.140)$$

I primi risultati generali relativi a questa classe di reti sono stato ottenuti da Jackson¹. Jackson ha studiato reti di code con le seguenti caratteristiche:

- la rete comprende N code
- la coda i ha m_i servitori
- ogni cliente alla coda i richiede servizio per un tempo con densità di probabilità esponenziale con parametro μ_i , quindi ogni singola coda è del tipo $\cdot/M/m_i$
- γ_i indica la velocità del processo di Poisson secondo cui si susseguono gli arrivi dall'esterno alla coda i
- alla fine del servizio un cliente va dalla coda i alla coda j con probabilità $r_{i,j}$, $i, j = 1, 2, \dots, N$, oppure lascia la rete con probabilità

$$1 - \sum_{k=1}^N r_{i,k} \quad i, j = 1, 2, \dots, N \quad (3.141)$$

le probabilità $r_{i,j}$ sono fisse e sono indipendenti dagli altri parametri della rete

- ci possono essere cicli e quindi i processi di arrivo alle varie code non sono Poisson; definiamo le velocità di arrivo alla coda i come

$$\lambda_i = \gamma_i + \sum_{j=1}^N \lambda_j r_{j,i} \quad i = 1, 2, \dots, N \quad (3.142)$$

Si suppone di essere in condizione di stabilità

$$\lambda_i < m_i \mu_i \quad i = 1, 2, \dots, N \quad (3.143)$$

Definendo lo stato del sistema come

$$\mathbf{N} = (k_1, k_2, \dots, k_N) \quad (3.144)$$

e introducendo la notazione

$$\begin{aligned} \mathbf{N} &= (k_1, \dots, k_i, \dots, k_j, \dots, k_N) \\ \mathbf{N}_{i,0} &= (k_1, \dots, k_i + 1, \dots, k_j, \dots, k_N) \\ \mathbf{N}_{0,j} &= (k_1, \dots, k_i, \dots, k_j - 1, \dots, k_N) \\ \mathbf{N}_{i,j} &= (k_1, \dots, k_i + 1, \dots, k_j - 1, \dots, k_N) \end{aligned} \quad (3.145)$$

dove deve essere $k_j - 1 \geq 0$, è facile notare che sono possibili solo le seguenti transizioni.

1. $\mathbf{N}_{0,j} \rightarrow \mathbf{N}$:
un cliente arriva dal mondo esterno alla coda j (con velocità γ_j), così che il numero di clienti alla coda j passa da $k_j - 1$ a k_j .
2. $\mathbf{N}_{i,0} \rightarrow \mathbf{N}$:
un cliente lascia la stazione i uscendo dal sistema (con velocità $\alpha_i(k_i + 1)\mu_i$ e probabilità $r_{i,0} = 1 - \sum_{j=1}^N r_{i,j}$), così che il numero di clienti passa da $k_i + 1$ a k_i .

¹Jackson, J.R. "Jobshop-like Queueing Systems," *Management Science* 10, 1 (1963), 131-142.

3. $N_{i,j} \rightarrow N$:

un cliente lascia la stazione i per andare alla stazione j (con probabilità $r_{i,j}$ e velocità $\alpha_i(k_i + 1)\mu_i$), così che il numero di clienti della coda i passa da $k_i + 1$ a k_i e il numero dei clienti della coda j passa da $k_j - 1$ a k_j .

Il parametro $\alpha_i(k_i)$ è pari al minimo tra k_i (numero di clienti serviti $\leq m_i$) e m_i (numero massimo di clienti serviti alla coda i)

$$\alpha_i(k_i) = \min\{k_i, m_i\}$$

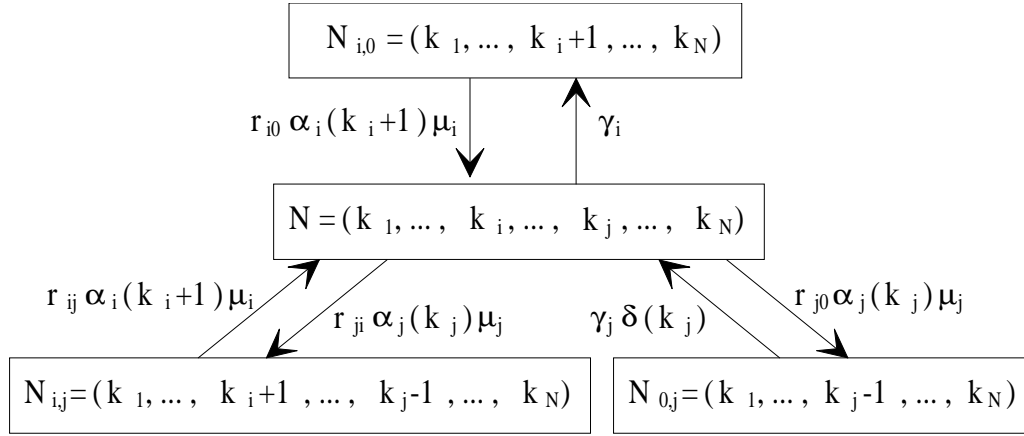


Figura 3.31. Bilanciamento dei flussi in una rete di Jackson

Bilanciando i flussi in ingresso ed in uscita dallo stato N possiamo scrivere (osservando la figura 3.31) il sistema di equazioni

$$\begin{aligned} \pi_N \left[\sum_{i=1}^N \gamma_i + \sum_{i=1}^N \alpha_i(k_i) \mu_i \right] = \\ = \sum_{i=1}^N \pi_{N_{i,0}} \alpha_i(k_i + 1) \mu_i r_{i,0} + \\ + \sum_{i=1}^N \pi_{N_{0,i}} \gamma_i \delta(k_i) + \sum_{i=1}^N \sum_{j=1}^N \pi_{N_{i,j}} \alpha_i(k_i + 1) \mu_i r_{i,j} \end{aligned} \quad (3.146)$$

dove al primo membro si ha il flusso uscente dallo stato N ottenuto come prodotto della probabilità dello stato per la somma delle velocità di uscita ed al secondo membro troviamo tre addendi che rappresentano il flusso entrante allo stato N . La funzione $\delta(k_i)$ vale 1 quando $k_i \geq 1$ e 0 altrimenti. Questa funzione ha il preciso compito di eliminare dalla sommatoria tutti gli addendi relativi a stazioni per cui non ci sono clienti in servizio.

Il sistema di equazioni lineari non è banale da risolvere in modo diretto. Jackson ha ipotizzato la soluzione che si otterrebbe se i processi di arrivo fossero Poisson (pur sapendo che in generale non è così)

$$\pi_N = \pi_{k_1, \dots, k_N} = \pi_{1, k_1} \pi_{2, k_2} \cdots \pi_{N, k_N} \quad (3.147)$$

dove

$$\pi_{i,k_j} = \begin{cases} \pi_{i,0} \left(\frac{\lambda_i}{\mu_i} \right)^{k_j} \frac{1}{k_j!} & 0 \leq k_j \leq m_i \\ \pi_{i,0} \left(\frac{\lambda_i}{\mu_i} \right)^{k_j} \frac{1}{m_i!} m_i^{m_i-k_j} & k_j \geq m_i \end{cases} \quad (3.148)$$

con la normalizzazione

$$\sum_{k_j=0}^{\infty} \pi_{i,k_j} = 1 \quad (3.149)$$

È facile verificare che questa soluzione soddisfa il sistema di equazioni lineari e quindi costituisce l'unica distribuzione di regime sotto la condizione

$$\frac{\lambda_i}{m_i \mu_i} < 1 \quad i = 1, \dots, N \quad (3.150)$$

3.9.5 Reti di Gordon e Newell

Le reti di Gordon e Newell, a differenza delle reti di Jackson non ammettono arrivi da e/o partenze verso l'esterno. Per questo motivo queste reti sono dette *chiuse*. L'assenza di arrivi e partenze fa sì che il numero totale di clienti nella rete rimanga costante.

Reti di questo tipo possono essere utili nella costruzione di modelli di sistemi con capacità finita, in cui si ammette un nuovo cliente solo quando si verifica una partenza dal sistema.

Ovviamente in reti di questo tipo i processi degli arrivi alle code non sono Poisson, altrimenti il numero dei clienti nella rete potrebbe crescere a valori arbitrariamente alti o decrescere fino a zero. La densità di probabilità dei tempi di servizio è sempre esponenziale. Lo spazio degli stati è finito poiché è finito il numero di clienti.

Come primo esempio consideriamo la rete di due code $\cdot/M/1$ in serie con ciclo mostrata nella figura 3.32 che è identica a quella nella figura 3.29 tranne che per l'assenza di arrivi dall'esterno.

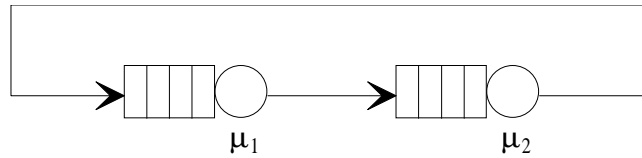


Figura 3.32. Due code $\cdot/M/1$ collegate in serie e con ciclo

Il diagramma delle velocità di transizione della catena di Markov tempo continua derivata dal modello della figura 3.32 è mostrato nella figura 3.33.

Dal diagramma delle velocità di transizione si ricavano le seguenti equazioni

$$\mu_1 \pi_{i,K-i} = \mu_2 \pi_{i-1,K-i+1} \quad (3.151)$$

da cui si ottiene

$$\pi_{i,K-i} = \pi_{0,K} \left(\frac{\mu_2}{\mu_1} \right)^i \quad (3.152)$$

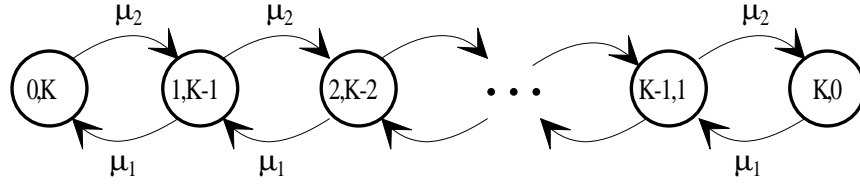


Figura 3.33. Diagramma delle velocità di transizione della catena di Markov tempo continua che modella due code $M/1$ in serie con ciclo

Si può verificare facilmente per sostituzione che le equazioni trovate sono soddisfatte dalla soluzione

$$\pi_{k_1, k_2} = \frac{1}{G} \left(\frac{\lambda_1}{\mu_1} \right)^{k_1} \left(\frac{\lambda_2}{\mu_2} \right)^{k_2} \quad (3.153)$$

con

$$\lambda_2 = \lambda_1 = \lambda \quad (3.154)$$

dove G è la costante di normalizzazione che garantisce che la somma di tutte le probabilità di regime associate agli stati raggiungibili sia pari ad uno.

Si può notare che anche in questo caso la soluzione è esprimibile in forma prodotto. Ora però il prodotto ha come fattori le distribuzioni marginali non normalizzate, essendo la normalizzazione demandata alla costante G .

Venendo ora al caso generale, se con k_i indichiamo il numero di clienti in attesa o in servizio alla coda i e con \hat{k} indichiamo il numero totale di clienti nella rete, per qualsiasi stato $N = (k_1, \dots, k_N)$ abbiamo

$$\sum_{i=1}^N k_i = \hat{k} \quad (3.155)$$

Gordon e Newell hanno dimostrato che queste reti ammettono soluzione in forma prodotto, poiché l'evoluzione del modello markoviano a loro associato è governata da un sottoinsieme delle equazioni ricavate in precedenza per le reti di Jackson.

Il procedimento per ricavare le equazioni in questo caso è molto simile a quello usato per le reti di Jackson.

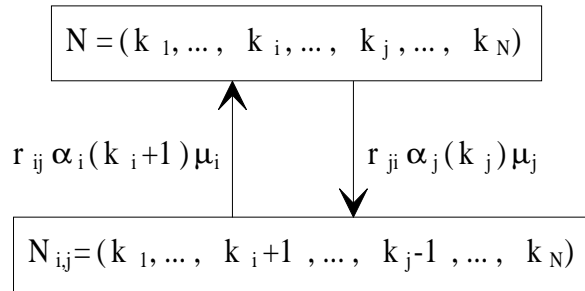


Figura 3.34. Bilanciamento dei flussi in una rete di Gordon e Newell

Come mostrato nella figura 3.34, l'unica transizione possibile è

$$N_{i,j} \rightarrow N$$

così che le equazioni risultanti sono

$$\pi_N \left[\sum_{i=1}^N \alpha_i(k_i) \mu_i \right] = \sum_{i=1}^N \sum_{j=1}^N \pi_{N_{i,j}} \alpha_i(k_i + 1) \mu_i r_{i,j} \quad (3.156)$$

e la soluzione cercata è

$$\pi_N = \frac{1}{G} \prod_{i=1}^N h_i(k_i) \quad (3.157)$$

dove la costante G vale

$$G = \sum_N \prod_{i=1}^N h_i(k_i) \quad (3.158)$$

Le funzioni $h_i(k_i)$ sono, a meno della normalizzazione, le probabilità di avere k_i clienti nella coda i studiata in isolamento, nel caso di un processo di arrivi avente parametro λ_i pari a

$$\lambda_i = \sum_{j=1}^N \lambda_j r_{j,i} \quad i = 1, \dots, N \quad (3.159)$$

Il sistema della (3.159), è omogeneo e quindi ammette sicuramente una soluzione tutta nulla (che non ci interessa); inoltre può ammettere infinite soluzioni che differiscono per una costante moltiplicativa. Per la soluzione della rete di code interessa una qualunque soluzione non nulla (la normalizzazione viene comunque garantita dalla costante G). Solitamente si sceglie $\lambda_1 = 1$ e si calcolano le altre λ_i in funzione di essa, così da avere la soluzione

$$\Lambda = (1, V_2, \dots, V_N) \quad (3.160)$$

dove $V_i = \lambda_i / \lambda_1$.

L'ergodicità dei modelli markoviani risultanti dalle reti di Gordon e Newell è garantita dal numero finito di stati.

Esempio

Si consideri la semplice rete di code chiusa mostrata nella figura 3.35. Chiamiamo coda 1 la coda di sinistra, coda 2 la coda in alto a destra e coda 3 la coda in basso a destra. I tempi di servizio alle tre code sono variabili casuali indipendenti ed esponenzialmente distribuite con parametro μ . Si supponga inoltre che $K = 3$ clienti siano presenti nella rete di code. Consideriamo il caso $\alpha = 1/4$.

La rete è di Gordon e Newell, per cui vale una soluzione in forma prodotto del tipo:

$$\pi_{k_1, k_2, k_3} = \frac{1}{G} \prod_{i=1}^3 h_i(k_i)$$

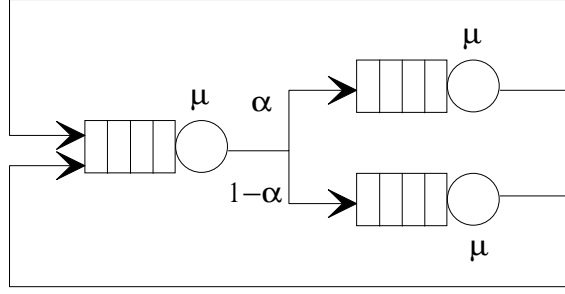


Figura 3.35. Semplice rete di Gordon e Newell

dove $k_1 + k_2 + k_3 = K = 3$ e $h_i(k_i) = (\lambda_i/\mu_i)^{k_i}$. Impostando il sistema (3.159) delle velocità di arrivo alle varie code si ottiene $\lambda_2 = \lambda_1/4$ e $\lambda_3 = 3\lambda_1/4$, da cui

$$\pi_{k_1, k_2, k_3} = \frac{1}{G} \left(\frac{\lambda_1}{\mu} \right)^{k_1} \left(\frac{\lambda_1}{4\mu} \right)^{k_2} \left(\frac{3\lambda_1}{4\mu} \right)^{k_3}$$

Inglobando μ e λ_1 nella costante di normalizzazione G , otteniamo

$$\pi_{k_1, k_2, k_3} = \frac{1}{G'} 1^{k_1} \left(\frac{1}{4} \right)^{k_2} \left(\frac{3}{4} \right)^{k_3}$$

La nuova costante di normalizzazione G' si trova imponendo la condizione

$$\sum_{k_1+k_2+k_3=3} \pi_{k_1, k_2, k_3} = 1$$

Gli stati possibili e le relative probabilità sono

(k_1, k_2, k_3)	(3,0,0)	(0,3,0)	(0,0,3)	(2,1,0)	(2,0,1)	(1,2,0)	(0,2,1)	(1,0,2)	(0,1,2)	(1,1,1)
$G' \pi_{k_1, k_2, k_3}$	1	$\left(\frac{1}{4}\right)^3$	$\left(\frac{3}{4}\right)^3$	$\frac{1}{4}$	$\frac{3}{4}$	$\left(\frac{1}{4}\right)^2$	$\left(\frac{1}{4}\right)^2 \frac{3}{4}$	$\left(\frac{3}{4}\right)^2$	$\frac{1}{4} \left(\frac{3}{4}\right)^2$	$\frac{1}{4} \frac{3}{4}$

Sommando tutte queste probabilità si ottiene

$$\sum_{k_1+k_2+k_3=3} \pi_{k_1, k_2, k_3} = \frac{55}{16} \frac{1}{G'} = 1$$

da cui

$$G' = \frac{55}{16} = 3.4375$$

e

$$\pi_{k_1, k_2, k_3} = \frac{16}{55} \left(\frac{1}{4} \right)^{k_2} \left(\frac{3}{4} \right)^{k_3}$$

Da questa distribuzione è possibile ricavare gli indici di prestazione di interesse. Per esempio il numero medio di clienti alle code vale

$$\begin{aligned}
E[N_1] &= 3\pi_{3,0,0} + 2(\pi_{2,1,0} + \pi_{2,0,1}) + (\pi_{1,1,1} + \pi_{1,2,0} + \pi_{1,0,2}) \\
&= \frac{16}{55} \left[3 \times 1 + 2 \times \left(\frac{1}{4} + \frac{3}{4} \right) + \left(\frac{3}{16} + \frac{1}{16} + \frac{9}{16} \right) \right] = \frac{93}{55} \approx 1.691 \\
E[N_2] &= 3\pi_{0,3,0} + 2(\pi_{1,2,0} + \pi_{0,2,1}) + (\pi_{1,1,1} + \pi_{2,1,0} + \pi_{0,1,2}) = \frac{27}{110} \approx 0.245 \\
E[N_3] &= 3\pi_{0,0,3} + 2(\pi_{0,1,2} + \pi_{1,0,2}) + (\pi_{1,1,1} + \pi_{0,2,1} + \pi_{2,0,1}) = \frac{117}{110} \approx 1.063
\end{aligned}$$

$E[N_3]$ può anche essere ricavato come $3 - E[N_1] - E[N_2]$.

Tali risultati rispettano il fatto che ogni cliente visita la coda 1, poi o la coda 2 o la coda 3, per poi tornare alla coda 1. I clienti quindi ciclano tra la coda 1 e il sottosistema composto dalle code 2 e 3. Visto che la capacità di erogare servizio è minore alla coda 1 rispetto al sottosistema composto dalle code 2 e 3, la fila d'attesa è più lunga alla coda 1 rispetto al numero totale di clienti alle code 2 e 3.

La velocità media di arrivo alla coda 1 coincide con la velocità media di partenza dalla coda, e vale

$$\begin{aligned}
\lambda_1 &= \mu (1 - P\{\text{coda vuota}\}) = \mu (1 - \pi_{0,3,0} - \pi_{0,0,3} - \pi_{0,2,1} - \pi_{0,1,2}) \\
&= \mu \left[1 - \frac{16}{55} \left(\frac{1}{64} + \frac{27}{64} + \frac{3}{64} + \frac{9}{64} \right) \right] = \mu \frac{9}{11}
\end{aligned}$$

Da λ_1 è possibile ricavare facilmente

$$\begin{aligned}
\lambda_2 &= \frac{\lambda_1}{4} = \mu \frac{9}{44} \\
\lambda_3 &= \frac{3\lambda_1}{4} = \mu \frac{27}{44}
\end{aligned}$$

Il tempo medio che intercorre tra due arrivi di clienti alla coda 1 vale $1/\lambda_1$, da cui si può dedurre che il tempo medio di ciclo di un generico cliente nella rete, cioè il tempo che intercorre tra due ingressi dello stesso cliente alla coda 1, vale $3/\lambda_1$, visto che i clienti si comportano tutti allo stesso modo.

Per calcolare $E[T_1]$, tempo medio di permanenza dei clienti alla coda 1, si può applicare il teorema di Little:

$$E[T_1] = \frac{E[N_1]}{\lambda_1} = \frac{31}{15\mu}$$

analogamente

$$\begin{aligned}
E[T_2] &= \frac{E[N_2]}{\lambda_2} = \frac{6}{5\mu} \\
E[T_3] &= \frac{E[N_3]}{\lambda_3} = \frac{26}{15\mu}
\end{aligned}$$

Il tempo di ciclo di un cliente è

$$\begin{aligned}
E[C] &= \frac{1}{4} (E[T_1] + E[T_2]) + \frac{3}{4} (E[T_1] + E[T_3]) \\
&= E[T_1] + \frac{1}{4} E[T_2] + \frac{3}{4} E[T_3] = \frac{11}{3\mu} = \frac{3}{\lambda_1}
\end{aligned}$$

come già ricavato in precedenza per altra via.

3.9.6 Reti BCMP

La ricerca di classi sempre più vaste di reti di code che ammettessero soluzione in forma prodotta ha portato al teorema BCMP (pubblicato nel 1975) che è stato esteso solo in misura marginale da ricerche successive. Il teorema e la classe di reti che esso definisce prendono il nome dai ricercatori che le hanno scoperte: Baskett, Chandy, Muntz e Palacios-Gomez.

Il teorema BCMP definisce una vasta classe di reti per cui la soluzione può essere espressa come prodotto di fattori che si possono calcolare a partire dall'analisi delle singole code in isolamento.

Le differenze sostanziali tra le reti BCMP e le reti considerate in precedenza si possono descrivere come segue.

- Classi di clienti:
fino ad ora non si era introdotta nessuna differenziazione tra i clienti; ora clienti di classi diverse possono avere comportamenti diversi per quanto riguarda il movimento da una coda alla successiva nella rete.
- Discipline di servizio:
l'unica disciplina di servizio ammessa nelle classi di reti precedenti era FCFS, adesso sono possibili altre discipline di servizio.
- Distribuzione dei tempi di servizio:
tutte le reti di code considerate ammettevano solo densità di probabilità esponenziali per i tempi di servizio; nelle reti BCMP sono permesse distribuzioni dei tempi di servizio diverse. In alcuni casi è ammessa una dipendenza della velocità di servizio dal numero di clienti presenti nella coda.

Il teorema BCMP considera reti con M code ed R classi di clienti. Oltre ad avere la possibilità di cambiare coda, al termine di un servizio i clienti possono cambiare classe. Indichiamo con $p_{i,r;j,s}$ la probabilità che un cliente di classe r alla coda i passi in classe s alla stazione j . La matrice

$$\mathbf{P} = [p_{i,r;j,s}] \quad (3.161)$$

è una matrice delle probabilità di transizione di una catena di Markov a tempo discreto, che ha come stato la coppia (stazione, classe).

Di solito la catena è riducibile in quanto non da ogni coppia stazione-classe si possono raggiungere tutte le altre code della rete in ogni possibile classe. È però possibile identificare nella catena U sottoclassi ergodiche (quindi irriducibili) che vengono chiamate

$$EC_1, EC_2, \dots, EC_U$$

Nel caso in cui i clienti non possono mai cambiare classe, le sottoclassi ergodiche della catena sono tante quante le classi, quindi $U = R$. Se le classi comunicano

$$U < R$$

Il numero di clienti di classe r alla coda i è indicato con $N_{i,r}$. Definiamo il vettore

$$\underline{N}_i = (N_{i,1}, N_{i,2}, \dots, N_{i,R}) \quad (3.162)$$

Le reti possono essere aperte, chiuse o anche miste. Nell'ultimo caso alcune classi sono aperte e altre sono chiuse.

Se la rete è chiusa si ha che

$$\sum_{i=1}^M \sum_{r \in EC_q} N_{i,r} = \hat{N}_q \quad (3.163)$$

Per reti aperte gli arrivi dall'esterno seguono processi di Poisson, o meglio processi con tempi di interarrivo distribuiti esponenzialmente, le cui velocità possono dipendere dal numero di clienti nella rete, o addirittura, se si ha un processo di arrivi per ogni sottoclasse ergodica, la sua velocità può dipendere dal numero di clienti nella sottoclasse.

Sono ammessi quattro tipi di code nelle reti BCMP.

1. Coda con servizi distribuiti esponenzialmente con la stessa media per tutte le classi; la velocità di servizio può dipendere dal numero totale di clienti nella coda. Disciplina di coda FCFS.
2. Coda con servitore unico e con disciplina PS. Sono possibili distribuzioni dei tempi di servizio diverse per classi diverse. Le distribuzioni dei tempi di servizio devono avere una funzione caratteristica razionale fratta.
3. Coda con servitore unico e con disciplina LCFS. Sono possibili distribuzioni dei tempi di servizio diverse per classi diverse. Le distribuzioni dei tempi di servizio devono avere una funzione caratteristica razionale fratta.
4. Coda con un numero di servitori sufficiente per allocare un servitore ad ogni cliente che arriva. Sono possibili distribuzioni dei tempi di servizio diverse per classi diverse. Le distribuzioni dei tempi di servizio devono avere una funzione caratteristica razionale fratta.

La definizione di stato della rete può essere anche molto complessa. Nei casi più semplici (reti di code comprendenti code ad infiniti servitori o con disciplina PS) lo stato della rete coincide con il vettore di vettori

$$\mathbf{N} = (\underline{N}_1, \underline{N}_2, \dots, \underline{N}_M) \quad (3.164)$$

Nel caso in cui siano presenti code con disciplina di servizio FCFS o LCFS è necessario, per ogni coda con disciplina FCFS o LCFS, esplicitare nella definizione di stato l'ordine con cui i clienti delle diverse classi sono accodati.

La distribuzione di regime per reti BCMP può essere espressa in forma prodotto

$$\pi(\mathbf{N}) = \frac{1}{G} \prod_{i=1}^M g_i(\underline{N}_i) \quad (3.165)$$

dove G è la costante di normalizzazione e le $g_i(\underline{N}_i)$ sono le densità di probabilità marginali non normalizzate, ottenute studiando le code in isolamento con velocità di arrivo date dal sistema di equazioni

$$\lambda_{i,r} = \gamma_{i,r} + \sum_{j=1}^M \sum_{s=1}^R \lambda_{j,s} p_{j,s;i,r} \quad i = 1, 2, \dots, M \quad r = 1, 2, \dots, R \quad (3.166)$$

dove $\gamma_{i,r}$ indica la velocità media di arrivo dall'esterno alla coda i in classe r .

Le condizioni di ergodicità, escludendo i casi degeneri (quali tempi medi di servizio non finiti), si riducono a verificare che il carico offerto da tutte le classi aperte ad ogni coda non ecceda la capacità

di erogare servizio. Per le code di tipo (4) questa condizione è sempre verificata. Per code degli altri tipi essa può essere espressa come:

$$\sum_{i=1}^{U_o} \frac{\lambda_{i,r}^*}{\mu_{i,r}^*} < 1 \quad i = 1, 2, \dots, M$$

dove U_o è il numero di sottoclassi ergodiche aperte, e

$$\lambda_{i,r}^* = \lim_{j \rightarrow \infty} \lambda_{i,r}(j) \quad \text{e} \quad \mu_{i,r}^* = \lim_{j \rightarrow \infty} \mu_{i,r}(j)$$

con $\lambda_{i,r}(j)$ e $\mu_{i,r}(j)$, rispettivamente, velocità media di arrivo e di servizio alla coda i per clienti di classe r quando alla coda i ci sono j clienti.

Nel caso in cui le velocità di arrivo e di servizio non dipendono dal numero di clienti nelle code e i clienti non possono cambiare classe, definendo con

$$N_i = \sum_{r=1}^R N_{i,r}$$

il numero totale di clienti alla coda i , le $g_i(\underline{N}_i)$ possono essere di tre tipi

$$g_i(\underline{N}_i) = \begin{cases} N_i! \prod_{r=1}^R \frac{1}{N_{i,r}!} \left(\frac{V_{i,r}}{\mu_i} \right)^{N_{i,r}} = \frac{N_i!}{\mu_i^{N_i}} \prod_{r=1}^R \frac{V_{i,r}^{N_{i,r}}}{N_{i,r}!} & \text{code di tipo (1)} \\ N_i! \prod_{r=1}^R \frac{1}{N_{i,r}!} \left(\frac{V_{i,r}}{\mu_{i,r}} \right)^{N_{i,r}} & \text{code di tipo (2) e (3)} \\ \prod_{r=1}^R \frac{1}{N_{i,r}!} \left(\frac{V_{i,r}}{\mu_{i,r}} \right)^{N_{i,r}} & \text{code di tipo (4)} \end{cases} \quad (3.167)$$

con

$$V_{i,r} = \begin{cases} \lambda_{i,r}/\lambda_{1,r} & r \text{ classe chiusa} \\ \lambda_{i,r} & r \text{ classe aperta} \end{cases}$$

Nel caso di disciplina di coda FCFS, quando i tempi di servizio hanno la stessa distribuzione esponenziale per tutti i clienti, la differenza tra le code del teorema BCMP e quelle di Jackson è data dal fatto che un cliente può cambiare classe da coda a coda, così che il percorso futuro può dipendere dalle stazioni visitate in precedenza.

Ovviamente le reti di Jackson e quelle di Gordon e Newell sono casi particolari delle reti del teorema BCMP.

Nel caso in cui si vogliano introdurre densità di probabilità dei tempi di servizio non esponenziali è necessario usare discipline di coda di tipo inusuale. Questo perché nel caso di densità di probabilità generali e disciplina FCFS la distribuzione del numero di clienti in coda non dipende solo dal valor medio del tempo di servizio. Particolari discipline (non FCFS) hanno invece l'effetto di cancellare l'impatto dei momenti di ordine superiore.

È possibile dare una spiegazione intuitiva della forma assunta dalle funzioni $g_i(\underline{N}_i)$, almeno nel caso FCFS [code di tipo (1)]. Avendo più classi di clienti, il termine $\rho_i^{N_i} = (\lambda_i/\mu_i)^{N_i}$ che compare, a meno delle solite costanti di normalizzazione, nelle espressioni della distribuzione di regime per le reti di Jackson e di Gordon e Newell, si decompone nel prodotto di termini $\hat{\rho}_{i,r} = V_{i,r}/\mu_{i,r} = V_{i,r}/\mu_i$.

Sapendo che alla coda ho N_i clienti in tutto e $N_{i,r}$ clienti di classe r , dobbiamo contare quante sono le configurazioni di coda che portano ad avere il carico $\hat{\rho}_{i,r}$ per tutti gli r . Tale enumerazione introduce i fattoriali che compaiono nella espressione di $g_i(\underline{N}_i)$. Infatti, i modi di avere $N_{i,1}$ clienti di classe 1 su N_i clienti sono

$$\binom{N_i}{N_{i,1}} = \binom{N_i}{N_i - N_{i,1}} = \binom{\sum_{r=1}^R N_{i,r}}{\sum_{r=2}^R N_{i,r}} = \frac{N_i!}{N_{i,1}!(N_{i,2} + N_{i,3} + \dots + N_{i,R})!}$$

Fissata la “posizione” dei clienti di classe 1 alla coda, possiamo contare i modi di avere $N_{i,2}$ clienti di classe 2 tra i rimanenti $N_{i,2} + N_{i,3} + \dots + N_{i,R} = \sum_{r=2}^R N_{i,r}$ clienti. Questi sono

$$\binom{\sum_{r=2}^R N_{i,r}}{\sum_{r=3}^R N_{i,r}} = \frac{(N_{i,2} + N_{i,3} + \dots + N_{i,R})!}{N_{i,2}!(N_{i,3} + N_{i,4} + \dots + N_{i,R})!}$$

Ripetendo questa enumerazione successivamente per le varie classi, otteniamo per la classe j , $j = 1, 2, \dots, R-1$, un fattore del tipo

$$\binom{\sum_{r=j}^R N_{i,r}}{\sum_{r=j+1}^R N_{i,r}} = \frac{(N_{i,j} + N_{i,j+1} + \dots + N_{i,R})!}{N_{i,j}!(N_{i,j+1} + N_{i,j+2} + \dots + N_{i,R})!}$$

Quando $j = R-1$ il fattore diventa

$$\frac{(N_{i,R-1} + N_{i,R})!}{N_{i,R-1}!N_{i,R}!}$$

Questi fattori devono comparire nella produttoria che esprime $g_i(\underline{N}_i)$ come “pesi” per i carichi di classe $\hat{\rho}_{i,r}$. Esplicitando ognuno di questi fattori come rapporti di fattoriali, e semplificando il fattoriale a numeratore di un fattore con l’equivalente fattoriale presente nel denominatore del fattore precedente, il prodotto di tutti questi “pesi” diventa

$$\frac{N_i!}{N_{i,1}!N_{i,2}!N_{i,3}!\dots N_{i,R}!}$$

come visto nella espressione (3.167) per le $g_i(\underline{N}_i)$.

Esempio

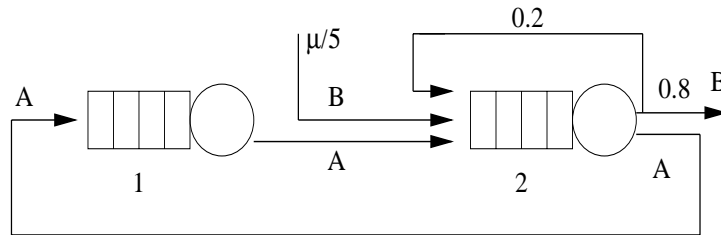


Figura 3.36. Rete BCMP mista

Si consideri la rete di code mostrata nella figura 3.36, composta da due code a servitore singolo con disciplina di servizio FCFS. Ad ogni coda i tempi di servizio sono indipendenti ed esponenzialmente distribuiti con parametro μ .

A tali code si presentano per ricevere servizio due tipi di clienti, che chiamiamo clienti di classe A e di classe B. I clienti di classe A sono due. Essi non abbandonano mai la rete di code: dopo aver ricevuto servizio alla coda 1, si spostano alla coda 2, per poi ritornare alla coda 1, e ripetere questo ciclo. I clienti di classe B arrivano dall'esterno alla coda 2 secondo un processo di Poisson con velocità $\mu/5$. Dopo aver ricevuto servizio alla coda 2, con probabilità 0.8 lasciano la rete di code, mentre con probabilità 0.2 ritornano alla coda 2.

I clienti non cambiano mai classe, quindi esistono $U = R = 2$ sottoclassi ergodiche EC_A e EC_B della catena corrispondente alle probabilità di transizione. La rete è mista: la classe B è aperta e la classe A è chiusa:

$$\sum_{i=1}^2 \sum_{r \in EC_A} N_{i,r} = \sum_{i=1}^2 N_{i,A} = \hat{N}_A = 2$$

Le velocità di arrivo alle singole code sono date dal sistema:

$$\begin{cases} \lambda_{1,A} = \lambda_{2,A} \\ \lambda_{1,B} = 0 \\ \lambda_{2,B} = \frac{\mu}{5} + 0.2\lambda_{2,B} \end{cases}$$

Da cui:

$$\begin{cases} \lambda_{1,A} = \lambda_{2,A} \\ \lambda_{1,B} = 0 \\ \lambda_{2,B} = \frac{\mu}{4} \end{cases} \quad \begin{cases} V_{1,A} = V_{2,A} = 1 \\ V_{1,B} = 0 \\ V_{2,B} = \frac{\mu}{4} \end{cases}$$

La rete risulta ergodica in quanto è chiusa rispetto ai clienti di classe A, mentre per i clienti di classe B, che affluiscono solo alla coda 2, si ha:

$$\frac{\lambda_{2,B}}{\mu} < 1$$

La rete soddisfa le condizioni imposte dal teorema BCMP, quindi la sua distribuzione di regime è in forma prodotto

$$\pi((N_{1,A}, N_{1,B}), (N_{2,A}, N_{2,B})) = \pi(N_{1,A}, (N_{2,A}, N_{2,B})) = \frac{1}{G} g_1(N_{1,A}, 0) g_2(N_{2,A}, N_{2,B})$$

Essendo entrambe le code di tipo (1),

$$g_i(\underline{N}_i) = \frac{N_i!}{\mu_i^{N_i}} \prod_{r=1}^2 \frac{V_{i,r}^{N_{i,r}}}{N_{i,r}!} \quad i = 1, 2$$

Per la coda 1

$$g_1(N_{1,A}, 0) = \left(\frac{V_{1,A}}{\mu_1} \right)^{N_{1,A}} = \left(\frac{1}{\mu} \right)^{N_{1,A}}$$

quindi le densità di probabilità marginali non normalizzate sono:

$$\begin{aligned} g_1(0,0) &= 1 \\ g_1(1,0) &= \frac{1}{\mu} \\ g_1(2,0) &= \left(\frac{1}{\mu} \right)^2 \\ g_1(j,i) &= 0 \quad \text{per } j > 2 \text{ oppure } i > 0 \end{aligned}$$

Per la coda 2

$$g_2(N_{2,A}, N_{2,B}) = \frac{(N_{2,A} + N_{2,B})!}{N_{2,A}! N_{2,B}!} \frac{V_{2,A}^{N_{2,A}} V_{2,B}^{N_{2,B}}}{\mu^{N_{2,A} + N_{2,B}}} = \frac{(N_{2,A} + N_{2,B})!}{N_{2,A}! N_{2,B}!} \frac{1}{\mu^{N_{2,A}} 4^{N_{2,B}}}$$

quindi le densità di probabilità marginali non normalizzate sono, per $i \geq 0$:

$$\begin{aligned} g_2(0, i) &= \left(\frac{1}{4}\right)^i \\ g_2(1, i) &= (i+1) \left(\frac{1}{\mu}\right) \left(\frac{1}{4}\right)^i \\ g_2(2, i) &= \frac{(i+1)(i+2)}{2} \left(\frac{1}{\mu}\right)^2 \left(\frac{1}{4}\right)^i \\ g_2(j, i) &= 0 \quad \text{per } j > 2 \end{aligned}$$

Sommando le probabilità di tutti gli stati della distribuzione di regime della rete BCMP, tenendo conto del vincolo $N_{2,A} = 2 - N_{1,A}$, si ricava la costante di normalizzazione G :

$$\begin{aligned} \frac{1}{G} \left[\sum_{i=0}^{\infty} g_1(2, 0) g_2(0, i) + \sum_{i=0}^{\infty} g_1(1, 0) g_2(1, i) + \sum_{i=0}^{\infty} g_1(0, 0) g_2(2, i) \right] &= 1 \\ \frac{1}{G} \left[\sum_{i=0}^{\infty} \left(\frac{1}{\mu}\right)^2 \left(\frac{1}{4}\right)^i + \sum_{i=0}^{\infty} (i+1) \left(\frac{1}{\mu}\right) \left(\frac{1}{4}\right)^i + \sum_{i=0}^{\infty} \frac{(i+1)(i+2)}{2} \left(\frac{1}{\mu}\right)^2 \left(\frac{1}{4}\right)^i \right] &= 1 \\ \frac{1}{G'} \left[\sum_{i=0}^{\infty} \left(\frac{1}{4}\right)^i + \sum_{i=0}^{\infty} (i+1) \left(\frac{1}{4}\right)^i + \sum_{i=0}^{\infty} \frac{(i+1)(i+2)}{2} \left(\frac{1}{4}\right)^i \right] &= 1 \end{aligned}$$

Da cui, risolvendo le sommatorie, si ricava:

$$\frac{1}{G'} = \frac{1}{G\mu^2} = \frac{27}{148}$$

Utilizziamo la notazione compatta $\pi(j, i)$ per indicare $\pi((2-j, 0), (j, i))$, probabilità a regime di avere j clienti di classe A e i clienti di classe B nella coda 2, $2-j$ clienti di classe A e 0 clienti di classe B nella coda 1. Possiamo scrivere:

$$\begin{aligned} \pi(0, i) &= \frac{27}{148} \left(\frac{1}{4}\right)^i \\ \pi(1, i) &= \frac{27}{148} (i+1) \left(\frac{1}{4}\right)^i \\ \pi(2, i) &= \frac{27}{148} \frac{(i+1)(i+2)}{2} \left(\frac{1}{4}\right)^i \\ \pi(j, i) &= 0 \quad \text{per } j > 2 \end{aligned}$$

Da queste probabilità possiamo ricavare gli indici di prestazione del sistema. Per esempio, per il numero medio di clienti alle due code, si ha:

$$E[N_1] = \frac{27}{148} \left[1 \times \sum_{i=0}^{\infty} (i+1) \left(\frac{1}{4}\right)^i + 2 \times \sum_{i=0}^{\infty} \left(\frac{1}{4}\right)^i \right] = \frac{30}{37} \approx 0.81$$

$$E[N_2] = \frac{27}{148} \left[\sum_{i=0}^{\infty} i \left(\frac{1}{4}\right)^i + \sum_{i=0}^{\infty} (i+1)^2 \left(\frac{1}{4}\right)^i + \sum_{i=0}^{\infty} \frac{(i+1)(i+2)^2}{2} \left(\frac{1}{4}\right)^i \right] = \frac{71}{37} \approx 1.92$$

Si ricordi che alla coda 1 abbiamo solo clienti di classe A. Quindi alla coda 2 abbiamo mediamente $2 - E[N_1] \approx 1.19$ clienti di classe A e $E[N_2] - 2 + E[N_1] \approx 0.73$ clienti di classe B.

Il tempo medio che intercorre tra due arrivi successivi di clienti alla coda 1, $E[I_A]$, si può ricavare dall'inverso della velocità di arrivo dei clienti di classe A (gli unici ad arrivare alla coda):

$$E[I_A] = \frac{1}{\lambda_{1,A}}$$

$\lambda_{1,A}$ a sua volta si ricava dalle probabilità marginali alla coda 1. Indicando con $\pi_A^{(1)}(i)$ la probabilità di avere i clienti di classe A alla coda 1 abbiamo:

$$\begin{aligned} \lambda_{1,A} &= \mu \left[\pi_A^{(1)}(1) + \pi_A^{(1)}(2) \right] \\ &= \mu \frac{27}{148} \left[\sum_{i=0}^{\infty} (i+1) \left(\frac{1}{4}\right)^i + \sum_{i=0}^{\infty} \left(\frac{1}{4}\right)^i \right] \\ &= \mu \frac{27}{148} \left[\frac{1}{(1-1/4)^2} + \frac{1}{1-1/4} \right] \\ &= \mu \frac{21}{37} \end{aligned}$$

Quindi si ha:

$$E[I_A] = \frac{37}{21\mu}$$

Si noti che il carico offerto alla coda 2 da clienti di classe A è $\frac{\lambda_{2,A}}{\mu} = \frac{\lambda_{1,A}}{\mu} = \frac{21}{37}$, mentre il carico offertole da clienti di classe B è $\frac{\lambda_{2,B}}{\mu} = \frac{1}{4}$. Il carico totale alla coda 2 è quindi $\rho_2 = \frac{21}{37} + \frac{1}{4} \approx 0.818$. Se la coda 2 fosse una M/M/1, il numero medio di clienti alla coda sarebbe pari a $\frac{\rho_2}{1-\rho_2} \approx 4.48$, valore molto maggiore di quello precedentemente calcolato per $E[N_2]$.

3.10 Ritardo in reti a commutazione di pacchetto

Consideriamo una rete di telecomunicazioni a commutazione di pacchetto con topologia molto semplice, mostrata nella figura 3.37. Nella rete si hanno tre nodi collegati da canali bidirezionali e due utenti chiamati A e B.

Internamente ad ogni nodo si hanno code in cui vengono inseriti i pacchetti in attesa di trasmissione.

Il ritardo subito da un pacchetto nel trasferimento dalla sorgente alla destinazione è l'intervallo che intercorre tra l'istante in cui il pacchetto entra nel nodo collegato alla sorgente e l'istante in cui il pacchetto giunge nel nodo al quale è collegato l'utente destinatario.

Si può esprimere quindi il ritardo di un pacchetto come somma dei tempi dovuti all'attesa in ciascuna coda di trasmissione, all'elaborazione in ciascun nodo, alla trasmissione e alla propagazione su ciascun canale attraversato.

Per poter sviluppare i calcoli introduciamo le seguenti ipotesi:

1. la disciplina di coda è di tipo FCFS per ogni coda,

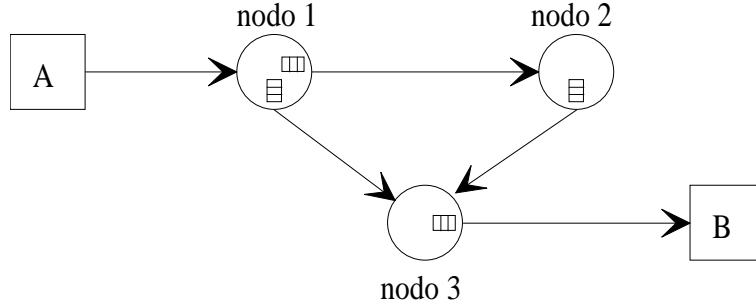


Figura 3.37. Rete a commutazione di pacchetto

2. l'instradamento dei pacchetti è di tipo *statico*, cioè le probabilità secondo cui un pacchetto passa da un nodo all'altro sono prefissate e non dipendono dal carico,
3. la disponibilità di memoria nei nodi per immagazzinare pacchetti è illimitata,
4. i canali sono privi di rumore,
5. i tempi di elaborazione e propagazione sono trascurabili,
6. il tempo di trasmissione dei pacchetti ha durata esponenziale,
7. gli arrivi dall'esterno seguono dei processi di Poisson.

Consideriamo una rete con

- N nodi,
- M canali; il canale i ha velocità pari a C_i [bit/s],
- pacchetti di lunghezza con densità di probabilità esponenziale e media $1/\mu$ [bit], di modo che tempo medio di trasmissione di un pacchetto sul canale i è dato da $1/(\mu C_i)$,
- traffico generato nel nodo j e destinato al nodo k di tipo Poisson con velocità $\gamma_{j,k}$ [pacchetti/s].

Si ricava che il traffico totale generato nel nodo j è Poisson con velocità

$$\gamma_j = \sum_{k=1}^N \gamma_{j,k} \quad (3.168)$$

e che il traffico totale offerto alla rete è Poisson con velocità

$$\gamma = \sum_{j=1}^N \gamma_j = \sum_{j=1}^N \sum_{k=1}^N \gamma_{j,k} \quad (3.169)$$

Interessa definire anche il traffico sul canale i , λ_i . Il traffico totale sugli M canali vale

$$\lambda = \sum_{i=1}^M \lambda_i \quad (3.170)$$

Poiché in generale i pacchetti possono attraversare più canali

$$\lambda \geq \gamma \quad (3.171)$$

La relazione che intercorre tra λ e γ dipende dalla politica di instradamento. Quando questa è nota, si possono calcolare le λ_i a partire dalle $\gamma_{j,k}$.

Il modello del sistema è una rete di code che differisce da una rete di Jackson per un solo aspetto: nel caso delle reti di Jackson il tempo di servizio di un cliente alle varie code è diverso, essendo ottenuto come istanza di variabili causali diverse. Nel caso delle reti a commutazione di pacchetto, il tempo di servizio coincide con il tempo di trasmissione e quindi dipende dalla lunghezza del pacchetto. Poiché questa è fissata, il tempo di servizio alle varie code dipende solo dalla velocità di trasmissione.

Per poter ottenere risultati in modo semplice L.Kleinrock suggerì di introdurre un'ulteriore **ipotesi di indipendenza** e così supporre che la lunghezza dei pacchetti vari da coda a coda. In questo modo si possono utilizzare i risultati disponibili per le reti di code di Jackson. L'impatto dell'ipotesi di indipendenza sui risultati dipende dalla topologia della rete. L'ipotesi è buona quando i nodi hanno molti ingressi e molte uscite; invece l'ipotesi porta ad errori consistenti quando i nodi hanno un solo ingresso e una sola uscita.

Con l'ipotesi di indipendenza possiamo affrontare il calcolo del ritardo medio in una rete di questo tipo, $E[T]$, usando il teorema di Little.

Sia $E[n]$ il numero medio di pacchetti nella rete. Si può scrivere

$$E[n] = \gamma E[T] \quad (3.172)$$

Poi

$$E[n] = \sum_{i=1}^M E[n_i] \quad (3.173)$$

dove con $E[n_i]$ si indica il numero medio di pacchetti accodati o in corso di trasmissione sul canale i .

Applicando nuovamente teorema di Little ricaviamo $E[n_i]$

$$E[n_i] = \lambda_i E[T_i] \quad (3.174)$$

dove $E[T_i]$ è il tempo medio trascorso nella fila di attesa o in trasmissione sul canale i . Sostituendo si ricava

$$E[T] = \frac{1}{\gamma} \sum_{i=1}^M \lambda_i E[T_i] \quad (3.175)$$

Per le ipotesi introdotte si può ricavare $E[T_i]$ studiando la coda i in isolamento

$$E[T_i] = \frac{\frac{1}{\mu C_i}}{1 - \frac{\lambda_i}{\mu C_i}} = \frac{1}{\mu C_i - \lambda_i}$$

Sostituendo infine si ricava

$$\begin{aligned} E[T] &= \frac{1}{\gamma} \sum_{i=1}^M \frac{\lambda_i}{\mu C_i - \lambda_i} \\ &= \frac{1}{\gamma} \sum_{i=1}^M \frac{f_i}{C_i - f_i} \end{aligned}$$

dove con f_i si indica il flusso di traffico sul canale i , definito come il prodotto dell'intensità del traffico che entra alla coda i per la lunghezza media dei pacchetti

$$f_i = \frac{\lambda_i}{\mu}$$

È possibile tenere conto dei tempi di elaborazione e propagazione introducendo nel modello a rete di code una stazione $(\cdot/M/\infty)$ per ogni canale e una stazione $(\cdot/M/\infty)$ all'ingresso di ogni nodo.

Supponendo che i tempi medi di servizio alle due stazioni siano indicati con $E[T_e]$ e $E[T_{pi}]$ si ottiene

$$E[T] = \left\{ \frac{1}{\gamma} \sum_{i=1}^M \lambda_i \left(\frac{1}{\mu C_i - \lambda_i} + E[T_e] + E[T_{pi}] \right) \right\} + E[T_e] \quad (3.176)$$

Il risultato è ancora esatto, grazie al teorema di BCMP (a meno che si desideri introdurre ritardi costanti).

Molto più critico è rimuovere l'ipotesi di distribuzione esponenziale per la durata della trasmissione di un pacchetto. In questo caso non vale più la soluzione in forma prodotto, pur mantenendo l'ipotesi di indipendenza. Si può comunque approssimare il risultato usando nella (3.176) la formula di Pollaczek-Khintchin

$$E[T_i] = \frac{1}{\mu C_i} \left[\frac{\frac{\lambda_i}{\mu C_i} (1 + C_i^2)}{2 \left(1 - \frac{\lambda_i}{\mu C_i} \right)} + 1 \right] \quad (3.177)$$

dove C_i è il coefficiente di variazione della distribuzione del tempo di trasmissione sul canale i .

Cerchiamo ora di ricavare un legame tra λ e γ . Indichiamo con la variabile $\nu_{j,k}$ il numero di canali attraversati dai pacchetti generati nel nodo j e destinati al nodo k . Se l'instradamento è fisso, allora $\nu_{j,k}$ è costante. Se invece l'instradamento è statico, ma suddiviso, allora $\nu_{j,k}$ è una variabile casuale. In questo caso calcoliamo il valor medio $E[\nu_{j,k}]$ e la media globale $E[\nu]$

$$E[\nu] = \sum_{j=1}^N \sum_{k=1}^N \frac{\gamma_{j,k}}{\gamma} E[\nu_{j,k}] \quad (3.178)$$

Sommando ora il traffico su tutti i canali della rete otteniamo

$$\sum_{i=1}^M \lambda_i = \lambda = \sum_{j=1}^N \sum_{k=1}^N \gamma_{j,k} E[\nu_{j,k}] = \gamma E[\nu]$$

quindi

$$\lambda = \gamma E[\nu] \quad (3.179)$$

Quindi anche

$$E[\nu] = \frac{1}{\gamma} \sum_{i=1}^M \lambda_i \quad (3.180)$$

essendo in generale $E[\nu] \geq 1$.

Supponiamo ora che il traffico nella rete sia basso così che i pacchetti non debbano attendere nel nodo e vengano subito trasmessi sul canale di uscita. Quindi

$$E[T_i] = \frac{1}{\mu C_i} \quad (3.181)$$

ovvero il ritardo medio alla coda i è dato solo dal tempo di trasmissione. Da ciò consegue che

$$E[T] = \frac{1}{\gamma} \sum_{i=1}^M \lambda_i \frac{1}{\mu C_i} \quad (3.182)$$

Se poi tutti i canali hanno la stessa velocità $C = C_i$, allora

$$E[T] = \frac{1}{\mu C} \frac{\sum_{i=1}^M \lambda_i}{\gamma} \quad (3.183)$$

e quindi

$$E[T] = \frac{1}{\mu C} E[\nu] \quad (3.184)$$

Quindi con traffico basso il ritardo totale è pari al prodotto del numero medio di canali attraversati per il tempo di trasmissione del pacchetto su un canale qualunque.