UNIVERSITÀ DEGLI STUDI DI PADOVA

Facoltà di Ingegneria

Corso di laurea in Ingegneria Gestionale

TESI DI LAUREA

RETI DI CODE

Relatore: Giorgio Romanin Jacur

Laureando: Matteo Chiarello

ANNO ACCADEMICO 2010 – 2011

Indice

CAPITOLO 4 – Reti di code

CAPITOLO 5 – Reti di code chiuse

Introduzione

La teoria delle code è stata sviluppata per fornire modelli capaci di prevedere il comportamento di sistemi che tentano di fornire un servizio per richieste, che arrivano in modo aleatorio. Proprio per questo motivo i primi problemi studiati con questi modelli sono stati relativi al traffico telefonico.

Fondatore della teoria delle code può essere considerato il matematico danese A. K. Erlang (Denmark 1878-1929), che nel 1909 scrisse una pubblicazione dal titolo “La teoria delle probabilità e le conversazioni telefoniche”. Egli osservò che, in generale, un centralino telefonico è un sistema avente richieste di servizio aleatorie di tipo poissoniano (S. D. Poisson, France 1781-1840), tempi di attesa per la connessione aleatori di tipo esponenziale o costante, una o più linee di comunicazione a disposizione.

I risultati di Erlang sono stati generalizzati alla fine degli anni ’20 da E. C. Molina e T. C. Fry.

E’ però agli inizi degli anni ’30 che l’austriaco F. Pollaczek (Austria 1892-1981) ed i russi A. N. Kolmogorov (Russia 1903-1987) e A. Y. Khinchin (Russia 1894-1959) proposero nuovi fondamentali risultati relativi ai sistemi con servizi aleatori generici.

A partire dagli anni ’50, grazie alle sue notevoli applicazioni economiche, lo studio della teoria delle code ebbe un notevole impulso nell’ambito della probabilità, della ricerca operativa, della management science e dell’ingegneria industriale. Esempi sono i problemi di gestione del traffico (veicoli, aerei, persone, trasmissioni in generale), scheduling (pazienti negli ospedali, lavori in catene di montaggio, programmi su un computer), ed organizzazione di servizi (banche, uffici postali, parchi di divertimento, ristoranti fast-food).

Fu proprio agli inizi degli anni ’50 che D. G. Kendall (England 1918-2007) propose una notazione standard per la descrizione dei sistemi di code, mentre agli inizi degli anni ’60 l’americano John D. C. Little dimostrò la fondamentale proprietà relativa ad un sistema di code stabilizzato.

CAPITOLO 1
Il sistema

1. La coda in un sistema

La teoria delle code ha avuto impulso all'inizio del secolo dall'emergente settore telefonico. Le sue origini vengono fatte risalire al 1909 quando l'ingegnere danese Agner Krarup Erlang pubblicò un articolo intitolato "The theory of probability and telephone conversations" relativo alle attese nelle chiamate telefoniche.

Il problema affrontato è quello dell'analisi del comportamento di un sistema isolato con capacità limitata in grado di offrire i suoi servizi ad un certo numero di utenti che entrano in conflitto tra loro per poterlo utilizzare. Rappresentano sistemi le macchine operatrici di un'officina e gli utenti i pezzi da lavorare, l'unità centrale di un calcolatore e gli utenti i programmi da elaborare, una linea telefonica (o un telefono) e gli utenti gli utilizzatori, un ufficio amministrativo e gli utenti le pratiche in attesa di essere completate.

L'analisi del processo, con la valutazione dei tempi d'attesa, delle lunghezze delle code, del tasso dei servizi resi, ha come obiettivo la configurazione del sistema.

Accanto al progetto, come vedremo più in particolare nel capitolo successivo, i modelli ottenuti ci permettono di stabilire politiche di controllo del processo.

2. Generalità: il sistema

La struttura di quell'entità che chiameremo *Sistema* è invariabilmente costituita da:

- Un processo di arrivi di utenti;
- Una coda di utenti in attesa;
- Un certo numero di serventi, caratterizzati da un loro processo di servizio;
- Una politica di coda, con cui gli utenti in attesa vengono estratti ed avviati al primo servente disponibile.

Il funzionamento ad eventi discreti di un sistema è il seguente:

- Un nuovo utente raggiunge il sistema;
- Se nessuno dei serventi è libero si pone in attesa altrimenti è immediatamente avviato ad uno dei serventi ed inizia il servizio;

- Appena un servente si libera, uno degli utenti in attesa, scelto secondo la politica adottata, passa dalla coda al servente libero ed inizia il servizio;
- Dopo un tempo di servizio l'utente servito lascia il sistema.

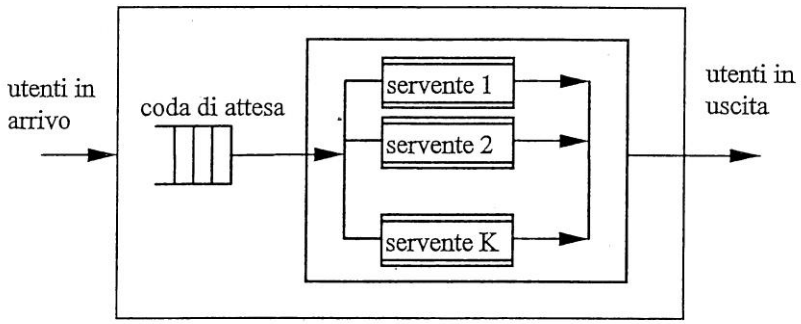


Fig. 1.1 Schema di un sistema

La struttura dei modelli è definita da:

- Stato del sistema: lo stato del sistema è dato dal numero di utenti presenti in coda e da quelli attualmente in servizio.
- Modalità degli arrivi: il processo degli arrivi al servizio è caratterizzato dagli intervalli di tempo che intercorrono tra due arrivi successivi di utenti al sistema di servizio. Questo intervallo è detto *tempo di interarrivo*.
  Il tempo di interarrivo può essere una variabile deterministica o casuale.
- Modalità di servizio: è il periodo di tempo necessario per servire un utente; è detto *tempo di servizio*.
  Può essere anch'esso una variabile deterministica o casuale.
- Numero di serventi: una risorsa è caratterizzata dal numero dei serventi, uno o più di uno.
- Capacità della coda: la coda d'attesa di una risorsa può avere capacità infinita oppure finita. In questo secondo caso è indicato il numero massimo di utenti che possono attendere in coda.
- Disciplina della coda: ogni risorsa adotta una politica con cui gli utenti in attesa sono avviati ai servizi. Può essere ad esempio FIFO (First In-First Out), LIFO (Last In-First Out).
- Dimensione della popolazione: una risorsa può essere visitata da utenti di una popolazione finita oppure infinita.

- Notazione di Kendall

  Nel 1953 David George Kendall introdusse la notazione A/B/C, successivamente estesa in notazione a sei campi del tipo 1/2/3/4/5/6 al fine di sintetizzare le caratteristiche di una risorsa:

  arrivi/servizi/serventi/capacità/disciplina/popolazione

  Es. M / G / 1 / ∞ / FIFO / P

  Gli arrivi ad una risorsa possono avvenire secondo un processo di Poisson (tempi di interarrivo variabili casuali indipendenti con distribuzione esponenziale), si parla di arrivi Markoviani e si indicano con la lettera M; arrivi sempre con tempi di interarrivo variabili casuali indipendenti di distribuzione generica si indicano con la lettera G; oppure arrivi con intertempi deterministici, cioè tempi di interarrivo costanti, si indicano con la lettera D.

  I tempi di servizio possono essere variabili casuali indipendenti con distribuzione esponenziale, si parla di processo Markoviano e si indica con la lettera M; tempi di servizio variabili casuali indipendenti, di distribuzione generica si indicano con la lettera G; oppure tempi di servizio grandezze deterministiche, si indicano con la lettera D.

  L'ultimo campo infine indica la dimensione della popolazione.

  Se capacità della risorsa e dimensione della popolazione sono infinite e la disciplina della coda FIFO, gli ultimi tre campi possono essere omessi.

Ovviamente, solo se entrambi, arrivi e servizi, sono Markoviani la risorsa dà origine ad una catena di Markov, altrimenti il processo è semi-Markov.
Dunque i modelli più studiati sono i sistemi M/M/s/k (con s e k che possono assumere diversi valori), perché la distribuzione esponenziale degli intertempi di arrivo e di servizio permette di studiare i sistemi a coda come dei processi di nascita e morte.

Una risorsa è caratterizzata dallo stato *X(t)*, cioè il numero di utenti presenti nella risorsa all'istante *t* (numero degli utenti in attesa più quelli in servizio).
Per descrivere le sue prestazioni introduciamo la seguente notazione:

- $\pi_n(t) \in [0 \div 1]$, probabilità che lo stato del sistema di servizio all'istante *t* sia pari a *n*.

- $\prod(z,t) \coloneqq \sum_{n=0}^{\infty} \pi_n(t) \cdot z^{-n}$, funzione generatrice delle probabilità di stato, ricordando che $\prod(1,t) = 1$.
- *N(t)*, valore medio del numero di utenti presenti nella risorsa all'istante *t*.
- *L(t)*, valore medio del numero di utenti in coda all'istante *t*.
- $\lambda(t)$, valore medio del numero di arrivi nell'unità di tempo all'istante *t*, detto anche *tasso di arrivo*.
- $\mu(t)$, valore medio del numero di servizi effettuati nell'unità di tempo all'istante *t*, detto anche *tasso di servizio*; $\frac{1}{\mu(t)}$ è il *tempo medio di servizio*.
- $\rho(t) = \frac{\lambda(t)}{\mu(t)}$, *intensità di traffico* all'istante *t*.
- $\Delta T_c(t)$, valore medio del tempo speso da un utente in coda all'istante *t*.
- $\Delta T(t)$, valore medio del tempo totale speso da un utente nella risorsa sino all'istante *t*.

Fra le grandezze precedenti sono verificate le seguenti relazioni.
Il valore medio degli utenti nella risorsa è dato dalla definizione:

$$N(t) = \sum_{n=0}^{\infty} n \cdot \pi_n(t)$$

che può essere, inoltre, ricavato attraverso la funzione generatrice di probabilità:

$$N(t) = \left. \frac{d \prod(z,t)}{dz} \right|_{z=1} .$$

Quando vi sono *m* serventi nella risorsa, la lunghezza media della coda dipende da *m*:

$$l(t) = \begin{cases} 0 & se\ i \leq m \\ i - m & se\ i > m \end{cases}$$

$$\tilde{\pi}_l(t) = \begin{cases} \sum_{i=0}^{m} \pi_i(t) & se\ l = 0 \\ \pi_{m+l}(t) & se\ l > 0 \end{cases}$$

$$L(t) = \sum_{l=0}^{\infty} i \cdot \tilde{\pi}_l(t) \ .$$

In condizioni stazionarie, il tempo medio totale trascorso da un utente in una risorsa è dato dalla somma del tempo medio trascorso in coda e del tempo medio di servizio:

$$\Delta T = \Delta T_c + \frac{1}{\mu} \ .$$

3. Legge di Little

Prima di iniziare ad analizzare in dettaglio la casistica di alcuni esempi principali di code, citiamo un risultato fondamentale, che si applica indipendentemente dalle proprietà di Markovianità di una risorsa, ed è noto come la legge di Little.

Questa legge, formulata nel 1961 dal professore del MIT Sloan School of Management John D.C. Little, pone in relazione il tempo medio necessario ad un utente per attraversare una risorsa ed il numero medio di utenti presenti nella risorsa; formalizza l'idea intuitiva che maggiore è la coda più lungo sarà il tempo di attesa.

Questa legge è utilizzata nella gestione degli impianti industriali per stabilire il tempo di attraversamento, il ritmo produttivo o il materiale che è in lavorazione durante il tempo di attraversamento in un sistema produttivo.

Poniamoci di fronte ad una risorsa e consideriamo i due processi di conteggio del numero di arrivi e del numero di partenze *a(t),p(t).* Osserviamo una realizzazione di questi processi durante un periodo di tempo *(0,t).*

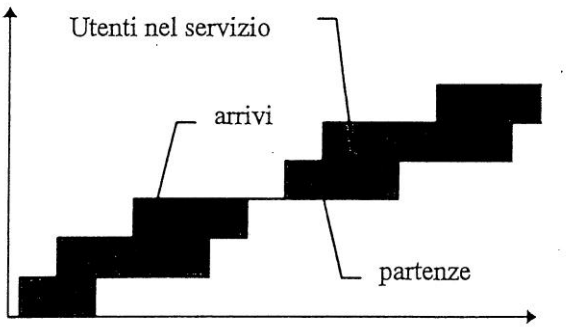


Fig. 1.2 Processi di conteggio di arrivi e partenze in una risorsa

Indichiamo con:
a(t) il numero di arrivi sino al tempo *t;*
p(t) il numero di partenze sino al tempo *t;*
N(t) il numero medio di utenti in coda al tempo *t.*
Calcoliamo l'area compresa fra le due curve dei conteggi degli arrivi e delle partenze; questa rappresenta il numero totale di utenti x tempo speso nella risorsa durante l'intervallo di osservazione:

$$\gamma(t) = \int_0^t \big(a(\tau) - p(\tau)\big) \cdot d\tau \tag{1.1}$$

Quindi, dividendo l'integrale per *t,* si ha la media temporale del numero di utenti nella risorsa osservati durante l'intervallo *t:*

$$\widehat{N}(t) = \frac{\gamma(t)}{t} \tag{1.2}$$

dividendo l'integrale, invece, per il numero degli arrivi, si ha la media temporale del tempo speso da ciascun utente nella risorsa:

$$\Delta\widehat{T}(t) = \frac{\gamma(t)}{a(t)} \tag{1.3}$$

Ricavando $\gamma(t)$ dall'equazione (1.3) e sostituendola nell'equazione (1.2),ricordando che:

$$\hat{\lambda}(t) = \frac{a(t)}{t}$$

è il tasso medio degli arrivi sino a *t,* si hanno le seguenti due relazioni che legano ai tempi medi d'attesa il numero medio di utenti nella risorsa:

$$\widehat{N}(t) = \hat{\lambda}(t) \cdot \Delta\widehat{T}(t)$$

oppure in coda:

$$\widehat{L}(t) = \hat{\lambda}(t) \cdot \delta\widehat{T}_c(t).$$

Se, in condizioni stazionarie, esistono i seguenti tre limiti:

$$\hat{\lambda} = \lim_{t\to\infty} \hat{\lambda}(t)$$

$$\Delta\hat{T} = \lim_{t\to\infty} \Delta\hat{T}(t)$$

$$\hat{\mu} = \lim_{t\to\infty} \hat{\mu}(t)$$

le seguenti tre relazioni sono verificate per la particolare realizzazione:

$$\hat{N} = \hat{\lambda} \cdot \Delta\hat{T}$$

$$\Delta\hat{T} = \Delta\hat{T}_c + \frac{1}{\hat{\mu}}$$

$$\hat{N} = \hat{L} + \hat{\rho}\ .$$

Infine, se ipotizziamo che i limiti precedenti esistano per qualsiasi realizzazione dei due processi degli arrivi e delle partenze, cioè in altre parole assumiamo che i processi degli arrivi e della coda siano ergodici[1], si ha il risultato noto come *Legge di Little:*

$$N = \lambda \cdot \Delta T$$

$$L = \lambda \cdot \Delta T_c$$

indipendentemente dal tipo di distribuzione dei processi d'arrivo e di servizio.

[1] Si dice di un sistema o processo in cui la media temporale delle grandezze che lo descrivono coincide con una opportuna media presa su un insieme di stati possibili del sistema stesso.

CAPITOLO 2
La coda

1. La coda M/M/1

Il primo esempio, il più semplice, di coda è rappresentato da una risorsa avente un solo servente, con distribuzioni esponenziali per i tempi d'interarrivo degli utenti e per i tempi di servizio, e capacità infinita per la coda d'attesa.
Si vede immediatamente che questa coda è una catena di Markov ed è modellata da un processo di nascita e morte a tempo continuo. Lo stato è rappresentato dal numero di utenti presenti nella risorsa. In generale i tassi di nascita e di morte possono essere funzioni dello stato.

$$\lambda_i\,,\ \mu_i\,,\quad \rho_i = \frac{\lambda_i}{\mu_i}\,,\quad i = 1,2,\cdots\cdots.$$

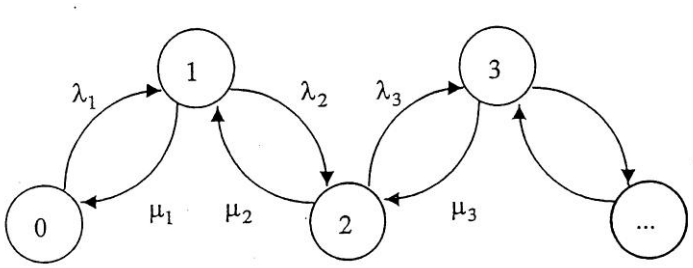


Fig. 2.1 La coda M/M/1

Le probabilità di stato a regime risultano:

$$\pi_i = \rho_1 \cdots \rho_i \cdot \pi_0. \tag{2.1}$$

Particolarmente semplice è il caso in cui tutti i coefficienti sono uguali tra loro, indipendentemente dallo stato; il coefficiente di traffico risulta:

$$\rho = \frac{\lambda}{\mu}\,.$$

La probabilità di stato (2.1) diventa:

$$\pi_i = \rho^i \cdot \pi_0$$

da cui discende la condizione di ergodicità $\rho < 1$ e la probabilità a regime che il sistema sia vuoto:

$$\sum_{i=0}^{\infty} \rho^i \cdot \pi_0 = \frac{1}{1-\rho}\pi_0 = 1\,; \quad \pi_0 = 1-\rho\,.$$

In conclusione, quindi, la densità di probabilità di stato di una coda M/M/1 a parametri costanti è la serie geometrica.

1.1.Fattore di utilizzo e tasso di uscita

Il fattore di utilizzo, dato dalla probabilità che la risorsa non sia vuota, coincide nella coda M/M/1 con l'intensità del traffico:

$$1-\pi_0 = \rho = \frac{\lambda}{\mu}\,.$$

Il tasso di attraversamento è dato alternativamente dal flusso degli arrivi o delle partenze per la somma delle corrispondenti probabilità, e vale:

$$\mu(1-\pi_0) = \lambda\,.$$

1.2.Numero medio di utenti e tempo medio di attesa

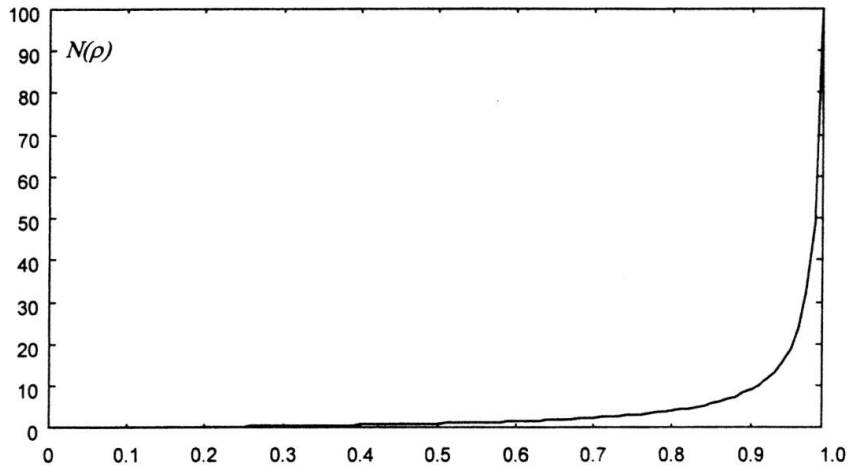


Fig. 2.2 Andamento del numero medio degli utenti in funzione dell'utilizzo

Il numero degli utenti nella risorsa si calcola dalla funzione generatrice di probabilità:

$$\prod(z) = \sum_{i=0}^{\infty} \pi_i \cdot z^{-i} = (1-\rho) \cdot \sum_{i=0}^{\infty} \rho^i \cdot z^{-i} = (1-\rho) \cdot \frac{1}{1-\rho \cdot z^{-1}} = (1-\rho) \cdot \frac{z}{z-\rho}$$

$$\frac{d\prod(z)}{dz} = -\frac{\rho \cdot (1-\rho)}{(z-\rho)^2}$$

$$N = -\frac{d\prod(z)}{dz}\bigg|_{z=1} = \frac{\rho}{1-\rho}$$

e da questo si ricava il tempo medio di attesa:

$$\Delta T = \frac{N}{\lambda} = \frac{1}{\mu \cdot (1-\rho)} \ .$$

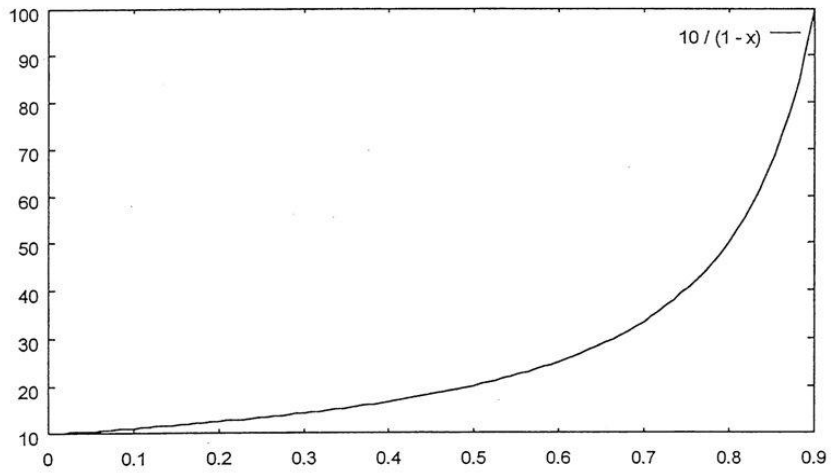


Fig. 2.3 Andamento del tempo d'attraversamento funzione dell'utilizzo

Gli andamenti di $N(\rho)$ e $\Delta T(\rho)$ al variare del fattore di utilizzo sono contenuti nelle figure 2.2 e 2.3.

2. La coda M/M/m

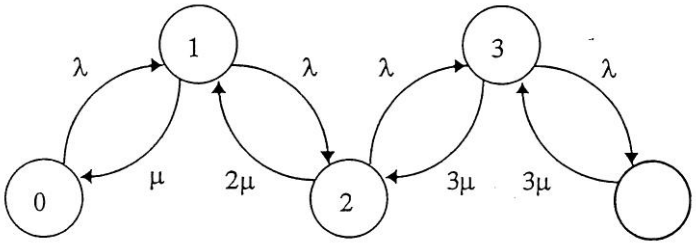


Fig. 2.4 Esempio di una risorsa con 3 serventi

L'estensione di una coda con un solo servente al caso di *m* serventi si modella assegnando un tasso medio di servizio funzione dello stato. Infatti se sono presenti nella risorsa più utenti, più di questi utenti, sino al raggiungimento del numero di serventi, sono contemporaneamente in servizio.
Sono quindi attivati contemporaneamente più eventi di morte, che nel caso Markoviano significa un tasso di morte proporzionale al numero degli eventi.

$$\mu_i = \begin{cases} i \cdot \mu & i < m \\ m \cdot \mu & i \geq m \end{cases}$$

Definiamo i seguenti coefficienti per l'intensità di traffico:

$$\rho_i = \frac{\lambda}{i \cdot \mu} \quad i \leq m\,; \qquad \rho = \frac{\lambda}{m \cdot \mu} < 1\ .$$

La probabilità di stato si calcola come segue:

$$\pi_i = \prod_{j=1}^{i} \rho_j \cdot \pi_0$$

$$\pi_i = \begin{cases} \frac{(m \cdot \rho)^i}{i!} \pi_0 & i < m \\ \frac{m^m}{m!} \rho^i \cdot \pi_0 & i \geq m \end{cases} \tag{2.2}$$

$$\pi_0 = \left(1 + \sum_{i=1}^{m-1} \frac{(m\cdot\rho)^i}{i!} + \frac{(m\cdot\rho)^m}{m!}\sum_{i=0}^{\infty}\rho^i\right)^{-1}$$
$$= \left(1 + \sum_{i=1}^{m-1} \frac{(m\cdot\rho)^i}{i!} + \frac{(m\cdot\rho)^m}{m!}\frac{1}{1-\rho}\right)^{-1}. \quad (2.3)$$

2.1.Fattore di utilizzo e tasso di uscita

Indichiamo con $B \in \{0,1,\cdots,m\}$ la variabile casuale che rappresenta il numero di serventi occupati. Il suo valor medio risulta:

$$E(B) = \sum_{i=0}^{m-1} i\cdot\pi_i + m\cdot Pr\{X \geq m\} \quad (2.4)$$

con:

$$Pr\{X \geq m\} = \frac{m^m}{m!}\sum_{i=m}^{\infty}\rho^i = \frac{m^m}{m!}\frac{\rho^m}{1-\rho}\,\pi_0$$

calcolata dalle probabilità di stato (2.2) e (2.3), che sostituita nella (2.4) porta a:

$$E(B) = m\cdot\rho\cdot\left(1 + \sum_{i=1}^{m-1}\frac{(m\cdot\rho)^{i-1}}{(i-1)!} + \frac{(m\cdot\rho)^{m-1}}{m!}\frac{m}{1-\rho}\right)\pi_0$$

$$E(B) = m\cdot\rho\cdot\left(1 + \sum_{j=1}^{m-1}\frac{(m\cdot\rho)^{j}}{j!} - \frac{(m\cdot\rho)^{m-1}}{(m-1)!} + \frac{(m\cdot\rho)^{m-1}}{m!}\frac{m}{1-\rho}\right)\pi_0$$

$$E(B) = m\cdot\rho\cdot\left(1 + \sum_{j=1}^{m-1}\frac{(m\cdot\rho)^{j}}{j!} + \frac{(m\cdot\rho)^{m}}{m!}\frac{1}{1-\rho}\right)\pi_0$$

da cui il risultato finale risulta:

$$E(B) = m\cdot\rho = \frac{\lambda}{\mu}.$$

Ne segue che ciascun servente ha un utilizzo:

$$\frac{E(B)}{m} = \rho \,.$$

Il tasso di uscita è ancora λ, in quanto in una coda stabile tassi di ingresso ed uscita devono essere identici.

2.2.Numero medio di utenti e tempo medio di attesa

Omettendo gli sviluppi il numero medio di utenti nella risorsa è:

$$N = m \cdot \rho + \frac{(m \cdot \rho)^m}{m!} \frac{1}{(1-\rho)^2} \pi_0$$

mentre il tempo medio di attesa:

$$\Delta T = \frac{1}{\mu} + \frac{1}{\mu} \frac{(m \cdot \rho)^m}{m!} \frac{1}{m(1-\rho)^2} \pi_0 \,.$$

3. La coda M/M/∞

Poiché in questa coda nessun utente deve attendere, se non il tempo di servizio, il comportamento è quello di un puro ritardo pari al tempo di servizio. Nel grafo del processo di nascita e morte il tasso di morte cresce proporzionalmente allo stato.
Si definisce come al solito il fattore di traffico:

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

La probabilità di stato risulta:

$$\pi_i = \frac{\rho^i}{i!} \pi_0$$

da cui la condizione di densità di probabilità porta allo sviluppo in serie di una funzione esponenziale:

$$\sum_{i=0}^{\infty} \frac{\rho^i}{i!} \pi_0 = 1$$

$$\pi_0 = e^{-\rho} .$$

3.1.Fattore di utilizzo e tasso di uscita

Il fattore di utilizzo in questa coda è:

$$1 - \pi_0 = 1 - e^{-\rho}$$

ed il tasso di uscita λ .

3.2.Numero medio di utenti e tempo medio di attesa

Il tempo medio di attesa è il reciproco del tasso di servizio:

$$\Delta T = \frac{1}{\mu}$$

ed il numero medio di utenti nella risorsa per la legge di Little vale:

$$N = \frac{\lambda}{\mu} = \rho .$$

4. La coda M/M/1/K

Questa coda ha una capacità finita del magazzino d'attesa. Quindi quando la capacità massima è stata raggiunta gli arrivi successivi sono scartati.

La condizione di densità di probabilità porta a:

$$\sum_{i=0}^{K} \rho^i \cdot \pi_0 = 1$$

che con alcuni passaggi fornisce il risultato:

$$\sum_{i=0}^{K} \rho^i + \sum_{i=K+1}^{\infty} \rho^i = \frac{1}{1-\rho} \; ;$$

$$\sum_{i=K+1}^{\infty} \rho^i = \frac{\rho^{K+1}}{1-\rho} \; ;$$

$$\pi_0 = \frac{1-\rho}{1-\rho^{K+1}} .$$

La probabilità di trovare la coda piena è data da:

$$\pi_K = 1 - \rho \frac{\rho^K}{1-\rho^{K+1}} . \tag{2.5}$$

4.1.Fattore di utilizzo e tasso di uscita

Il fattore di utilizzo per questa coda è dato da:

$$1 - \pi_0 = \rho \frac{(1-\rho^K)}{1-\rho^{K+1}} .$$

Il tasso di uscita dalla risorsa, che corrisponde al tasso effettivo d'attraversamento da parte degli utenti, quindi, non è uguale al tasso di ingresso e si ricava dalle probabilità di stato:

$$\gamma = \lambda \cdot \sum_{i=0}^{K-1} \pi_i = \lambda \cdot (1 - \pi_K) = \mu \cdot \sum_{i=1}^{K} \pi_i = \mu \cdot (1 - \pi_0)$$

mentre il tasso dei pezzi scartati è:

$$\gamma_{scarto} = \lambda \cdot \pi_K .$$

4.2.Numero medio di utenti e tempo medio di attesa

La traccia per il calcolo della lunghezza media della coda è la seguente:

$$\sum_{i=0}^{K} \rho^i z^{-1} = \sum_{i=0}^{\infty} \rho^i z^{-i} - \sum_{i=K+1}^{\infty} \rho^i z^{-i} = \frac{z}{z-\rho} - \frac{\rho^{K+1} z^{K+2}}{z-\rho}$$

che porta a:

$$N = \frac{\rho}{1-\rho^{K+1}} \left(\frac{1-\rho^K}{1-\rho} - K \cdot \rho^K \right).$$

Il tempo medio di attesa deriva dalla legge di Little:

$$N = \lambda \cdot (1 - \pi_K) \Delta T.$$

5. La coda M/M/1/M

Consideriamo adesso una coda con un servente, dove la popolazione non è più infinita ma limitata a M. Gli utenti possono essere nella risorsa, oppure al di fuori e, quando fuori, ciascuno in uno stato di arrivo rappresentato da un tempo di attesa casuale con distribuzione esponenziale di parametro λ. Il grafo delle transizioni di stato a dimensioni finite è rappresentato nella figura seguente:

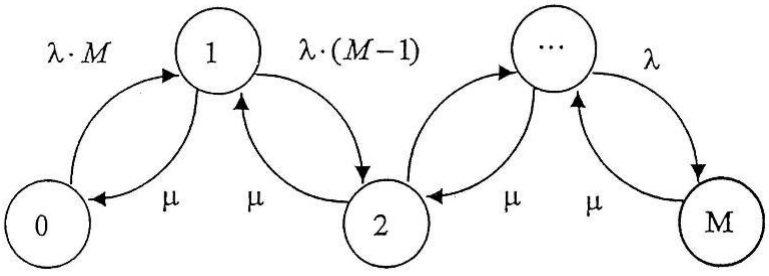


Fig. 2.5 La coda M/M/M

dove i tassi di nascita ad i diversi stati sono dati da:

$$\lambda_i = \begin{cases} \lambda \cdot (M - i + 1) & 1 \le i \le M \\ 0 & i = M + 1 \end{cases}.$$

Le probabilità di stato a regime risultano:

$$\pi_i = \begin{cases} \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{M!}{(M-i)!} \cdot \pi_0 & 0 < i \leq M \\ 0 & i > M \end{cases}$$

La probabilità di essere nello stato 0 è inoltre:

$$\pi_0 = \left[\sum_{i=0}^{M} \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{M!}{(M-i)!}\right]^{-1}.$$

Si potrebbe continuare la trattazione e sviluppare tutte le possibili variazioni sul tema della coda classica modellata da processi di nascita e morte.
In sostanza, le probabilità di stato di tutte queste variazioni godono della proprietà di forma prodotto data dall'equazione (2.1).

CAPITOLO 3
Oltre i processi di nascita e morte

Appena ci si allontana dal modello di nascita e morte, determinare le prestazioni di una coda diventa estremamente laborioso, al punto che spesso conviene ricorrere alla simulazione.
Esistono alcune eccezioni, tuttavia, che permettono ancora di essere risolte in modo analitico. Sono, queste, code in cui i tempi di interarrivo ed i tempi di servizio hanno distribuzioni ottenute come combinazioni lineari di funzioni esponenziali.
Verranno introdotte perché aiuteranno a comprendere alcune particolarità di comportamento delle reti di code.
Questi risultati si fanno risalire alle idee di Erlang all'inizio del secolo.

1. Metodo degli stadi

Questa tecnica di studio tratta particolari strutture degli intertempi fra due eventi, ottenute come interconnessione di stadi ciascuno dei quali ha distribuzione esponenziale.
L'interesse verso queste distribuzioni per descrivere processi di arrivo e processi di servizio di una coda, da utilizzarsi in alternativa alla distribuzione esponenziale, è dovuta al fatto che queste offrono una maggiore libertà nell'assegnazione del rapporto tra media e varianza. Infatti nella distribuzione esponenziale, media e deviazione standard hanno rapporto 1.

1.1.Stadi in serie

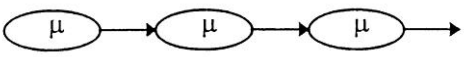


Fig. 3.1 Intertempo fra due eventi come cascata di tempi con distribuzione esponenziale.

In una struttura in serie il tempo di vita di un evento si costruisce attraversando *n* stadi in serie, ciascuno caratterizzato da una distribuzione esponenziale.

La funzione caratteristica e la corrispondente funzione di densità di probabilità del tempo di servizio sono quelle della distribuzione Gamma:

$$\Phi(s) = \frac{\mu^n}{(s+\mu)^n} \ , \quad d(t) = \mu \cdot \frac{(\mu \cdot t)^{n-1}}{(n-1)!} \cdot e^{-\mu \cdot t} \ . \tag{3.1}$$

Se al tasso di servizio μ si sostituisce il valore $n \cdot \mu$, in modo da mantenere inalterato, al variare di n, il valore medio del tempo di servizio $1/\mu$, si ha la densità di probabilità:

$$d(t) = n \cdot \mu \cdot \frac{(n \cdot \mu \cdot t)^{n-1}}{(n-1)!} \cdot e^{-n \cdot \mu \cdot t} \tag{3.2}$$

che prende il nome di *densità di probabilità di Erlang ad n stadi*. Si può verificare che al crescere di *n* questa funzione di probabilità ha una varianza che tende a 0.
Attraverso la funzione caratteristica si calcolano immediatamente media e varianza della distribuzione, che valgono:

$$\mu_X = \frac{1}{\mu} \ , \qquad {\sigma^2}_X = \frac{1}{n \cdot \mu^2} \ . \tag{3.3}$$

Questa distribuzione può essere impiegata per rappresentare fenomeni la cui incertezza è minore di quella della distribuzione esponenziale.

1.2.Stadi in parallelo

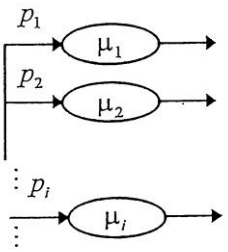


Fig. 3.2 L'intertempo ha, con probabilità p_i, una fra N funzioni di distribuzione esponenziale differenti

In una struttura in parallelo il tempo di vita di un evento si costruisce attraversando uno fra n stadi possibili in parallelo, scelti con probabilità p_i.

La funzione caratteristica e la conseguente densità di probabilità sono:

$$\Phi(s) = \sum_i p_i \cdot \frac{\mu_i}{s+\mu_i}, \quad d(t) = \sum_i p_i \cdot \mu_i e^{-\mu_i \cdot t}, \quad \sum_i p_i = 1. \tag{3.4}$$

Questa è la funzione di densità di probabilità *iperesponenziale.*
Media e varianza per questa distribuzione valgono:

$$\mu_X = \sum_i \frac{p_i}{\mu_i}, \quad {\sigma^2}_X = 2\sum_i \frac{p_i}{{\mu_i}^2} - \left(\sum_i \frac{p_i}{\mu_i}\right)^2. \tag{3.5}$$

Si può mostrare che in questa distribuzione la varianza è sempre maggiore di quella della distribuzione esponenziale, a parità di media. Può essere impiegata per rappresentare fenomeni di cui si sa che l'incertezza è maggiore di quella della distribuzione esponenziale.

1.3.Stadi in forma diretta

Più in generale si possono realizzare strutture ottenute combinando le due precedenti nella forma serie-parallelo o parallelo-serie.
Fra queste, di un certo interesse è quella in cui il tempo di vita si costruisce attraversando in alternativa con probabilità $(1-p)\cdot p'$ uno fra n stadi in parallelo, ciascuno costituito da i stadi in serie, che prende il nome di *connessione in forma diretta.*

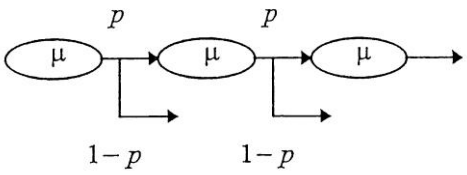


Fig. 3.3 L'intertempo ha una struttura in forma diretta

La funzione caratteristica e la relativa densità di probabilità sono in questo caso:

$$\Phi(s) = \sum_{i=1,n} (1-p)\cdot p^{i-1} \frac{\mu^i}{(s+\mu)^i}, \quad d(t) = \sum_{i=1,n} (1-p)\cdot p^{i-1} \cdot \mu \cdot \frac{(\mu\cdot t)^{i-1}}{(i-1)!} \cdot e^{-\mu\cdot t}. \tag{3.6}$$

CAPITOLO 4
Reti di code

Normalmente una risorsa non è utilizzata in modo isolato; più facilmente diverse risorse sono interconnesse fra loro per costruire un unico sistema. Un esempio è offerto da un'officina meccanica dove un pezzo deve eseguire nell'ordine una sequenza d'operazioni utilizzando una serie di macchine. Entrando in officina, il pezzo si pone in attesa di fronte alla prima macchina, quando ha terminato prosegue immediatamente verso la seconda, e così fino al completamento quando esce dal sistema.
In alcuni casi la stessa operazione può essere effettuata su più di una macchina in alternativa; in questo caso, al termine dell'operazione precedente, è necessario decidere verso quale delle possibili macchine la prossima operazione dovrà essere effettuata. Si parla in questo caso di *instradamento* o *routing.*
Una volta effettuata una operazione, dopo un controllo di qualità, può essere necessario ritornare su una delle macchine precedentemente visitate per ripetere una delle operazioni già fatte; si hanno in questo caso *anelli di ricircolo* nel percorso dei pezzi.
Gli utenti di un sistema possono essere della stessa classe oppure di classi diverse, e ciascuna classe può avere un suo programma di visite alle risorse diverso dalle altre. Si può anche verificare il caso che sulla stessa risorsa utenti di classi diverse abbiano tempi di servizio diversi tra loro. In un sistema con più classi di utenti in attesa della stessa risorsa, alcune classi possono essere servite in modo privilegiato rispetto ad altre; si parla in questo caso di *sequenziamento* o *schedulazione.*

Il sistema si può rappresentare come una rete di cui le singole risorse sono i nodi e dove i rami indicano i flussi di utenti da una risorsa all'altra; si parla allora di una *rete di code.*
Fra le classificazioni che distinguono una rete di code, la principale è sicuramente quella di *rete di code aperta* e *rete di code chiusa.* Nel primo caso il numero di utenti della rete non è prefissato, vi saranno arrivi e partenze da e verso l'esterno. E' questo il caso, per esempio, di una rete telefonica, dove il numero degli utenti varia nel tempo.

Nel secondo caso il numero degli utenti nella rete è fisso, non vi sono ingressi ed uscite dall'esterno, e gli utenti presenti continuano a circolare fra le risorse. Un'officina dove i pezzi da lavorare devono preventivamente essere montati su attrezzaggi offre un esempio di quest'ultimo tipo di rete. Il numero degli attrezzaggi presenti in officina è fissato; soltanto quando un pezzo ha finito il suo ciclo di operazioni, è stato scaricato ed ha liberato un attrezzaggio, un nuovo pezzo può entrare per prendere il suo posto. Poiché i pezzi possono circolare nella rete soltanto sui loro attrezzaggi, il numero di utenti rimane costante.

Studiare una rete di code significa definire e studiare il comportamento del suo stato. Lo stato di una rete di code è dato dall'unione degli stati di ciascun nodo, rappresentato, per code Markoviane, dal vettore di variabili casuali discrete del numero di utenti presenti presso ciascuna risorsa:

$$\boldsymbol{X} = [X_1, X_2, \cdots, X_M]$$

con probabilità di stato:

$$\pi(n_1, \cdots, n_M) = Pr\{X_1 = n_1, \cdots, X_M = n_M\}\,.$$

Per le reti di code, come nel caso di code isolate, è interessante stabilire quelle condizioni che permettono di offrire soluzioni analitiche al problema. Data la complessità del modello dinamico di una rete di code, ci si accontenta normalmente della sua soluzione in condizioni stazionarie.

L'idea di base per una soluzione semplice della rete è che ciascuna risorsa continui a comportarsi come una coda Markoviana, isolata ed indipendente dalle altre, e che la probabilità congiunta degli stati della rete sia il prodotto delle probabilità marginali di ciascuna delle risorse che compongono la rete: condizione che prende il nome di proprietà della *forma prodotto.* Questa proprietà è verificata per alcune classi di reti di code. La classe più importante è certamente costituita dalle reti di code Markoviane, dove gli arrivi alla rete sono processi di Poisson, i servizi forniti hanno tempi con distribuzione esponenziale e gli instradamenti sono casuali.

Nello studio di reti di code Markoviane gli aspetti da considerare sono tre: la natura del processo di uscita da una risorsa (che diventerà l'ingresso della risorsa successiva), la composizione di processi che confluiscono da nodi diversi della rete verso la stessa risorsa, e l'instradamento che scompone un processo in tanti processi separati.

Vedremo che in larga misura questi processi continuano ad essere processi di Poisson e quindi il risultato della forma prodotto è in un certo senso scontato.
La condizione di Markovianità delle code, però, non è l'unica a garantire la forma prodotto. In presenza di ricircoli, ad esempio, i processi all'interno della rete non sono più processi di Poisson e tuttavia la rete continua a comportarsi come una rete di code Markoviane.
.

1. Introduzione alle caratteristiche dei processi all'interno della rete

1.1.Processo in uscita da una coda Markoviana

Non deve sorprendere che il processo in uscita da una coda Markoviana ha ancora la proprietà di assenza di memoria. Un primo risultato rappresentativo è dovuto a *Burke* ed è contenuto nell'omonimo teorema.

Teorema 4.1: Teorema di Burke

In una coda M/M/1 ergodica a regime il processo in uscita è un processo di Poisson indipendente, che ha come parametro il parametro del processo di ingresso.

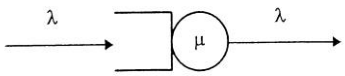


Fig. 4.1 Coda M/M/1

Dimostrazione:

Che i tassi di ingresso ed uscita debbano essere uguali è immediato, in quanto se la frequenza degli arrivi e quella delle partenze fossero differenti il numero medio di utenti in coda non sarebbe stazionario.
Consideriamo ora in dettaglio due eventi successivi in uscita. Possiamo trovarci in due condizioni possibili:

- In corrispondenza del primo evento la risorsa entra nello stato 0; allora il prossimo evento ci sarà soltanto dopo che, al prossimo arrivo, sarà trascorso un tempo di servizio.

La variabile casuale dell'intervallo di tempo fra due uscite sarà la somma di un tempo d'interarrivo e di un tempo di servizio, con funzione caratteristica:

$$\Phi_1(s) = \frac{\lambda}{s+\lambda} \cdot \frac{\mu}{s+\mu} . \tag{4.1}$$

- In corrispondenza del primo evento il servizio non entra nello stato 0; allora il prossimo evento avverrà dopo soltanto un tempo di servizio. La variabile casuale dell'intertempo corrisponde ad un tempo di servizio, con funzione caratteristica:

$$\Phi_2(s) = \frac{\mu}{s+\mu} . \tag{4.2}$$

La probabilità che un utente arrivi e trovi la risorsa vuota si verifica facilmente essere $\pi_0 = 1 - \rho$, che è quindi anche la probabilità di avere un intertempo con funzione caratteristica (4.1). Altrimenti si ha il secondo caso con probabilità $1 - \pi_0 = \rho$.
La funzione generatrice di probabilità degli intertempi del processo di uscita sarà quindi una media pesata tra (4.1) e (4.2):

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

che porta come risultato ancora ad una distribuzione esponenziale, con parametro il tasso di arrivo.

1.2.La composizione di due processi di Poisson

Consideriamo due processi di Poisson indipendenti, P_1 con parametro λ_1 e P_2 con parametro λ_2 , che si compongono per dare origine ad un unico processo di arrivo. Consideriamo il tempo d'interarrivo fra due eventi successivi.
Iniziando l'osservazione dall'ultimo evento accaduto, il prossimo evento apparterrà a quello dei due processi con il minimo residuo di vita.

Per la proprietà di assenza di memoria i residui di vita dei due processi, indipendentemente dal tempo già trascorso, avranno probabilità:

$$Pr\{X_1 \leq t\} = 1 - e^{-\lambda_1 \cdot t} \quad e \quad Pr\{X_2 \leq t\} = 1 - e^{-\lambda_2 \cdot t}$$

quindi la probabilità dell'intertempo fra due eventi successivi sarà uguale a quella del minimo fra i due tempi:

$$Pr\{X \leq t\} = 1 - Pr\{X > t\} = 1 - Pr\{min_{i=1,2} X_i > t\}.$$

E' facile verificare che l'evento $\{min_{i=1,2} X_i > t\}$ è equivalente all'evento $\{X_1 > t \cap X_2 > t\}$ e, poiché i due processi sono indipendenti, la probabilità degli intertempi risulta il prodotto delle probabilità:

$$Pr\{X \leq t\} = 1 - \prod_{i=1,2} Pr\{X_i > t\} = 1 - \prod_{i=1,2} e^{-\lambda_i \cdot t} = 1 - e^{-(\lambda_1 + \lambda_2) \cdot t}$$

risultato che dimostra che la composizione dei due processi è a sua volta un processo di Poisson con parametro la somma dei parametri dei due processi componenti.
L'estensione al caso di n processi è immediata.

1.3.Instradamento casuale

Un punto chiave per garantire la forma prodotto nelle reti di code è che gli instradamenti siano casuali. Questo significa che appena un utente ha terminato un servizio su una risorsa è immediatamente avviato alla prossima risorsa mediante l'estrazione di una variabile casuale indipendente.
Consideriamo il processo di Poisson in uscita da una coda Markoviana di parametro λ , ed assumiamo che, con probabilità r_i $(\sum_i r_i = 1)$, ogni utente in uscita è instradato in modo casuale verso uno degli I possibili percorsi alternativi. Consideriamo l'i-esimo percorso. Se ad un certo evento l'ultimo utente è stato instradato lungo quel percorso, la probabilità che lo sia anche il prossimo è r_i, che lo sia quello dopo è $r_i \cdot (1 - r_i)$, e così via con una probabilità che ha l'andamento di una serie geometrica. Quindi il tempo d'interarrivo lungo il percorso i-esimo corrisponde con probabilità $r_i \cdot (1 - r_i)^{k-1}$ alla somma di k tempi d'interarrivo del processo di Poisson;

Cioè la sua funzione caratteristica risulta:

$$\Phi(s) = r_i \cdot \sum_{k=1}^{\infty} (1 - r_i)^{k-1} \cdot \left(\frac{\lambda}{s+\lambda}\right)^k$$

$$= r_i \cdot \frac{\lambda}{s+\lambda} \cdot \sum_{k=0}^{\infty} (1 - r_i)^k \cdot \left(\frac{\lambda}{s+\lambda}\right)^k = r_i \cdot \frac{\lambda}{s+\lambda} \cdot \frac{1}{1 - (1 - r_i) \cdot \frac{\lambda}{s+\lambda}} = \frac{r_i \cdot \lambda}{s + r_i \cdot \lambda} .$$

Questo rappresenta ancora un processo di Poisson con tasso di arrivo $r_i \cdot \lambda$, che è la frazione r_i del tasso di uscita della risorsa precedente.

1.4.Reti con ricircolo

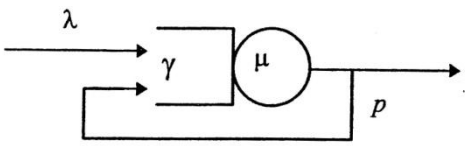


Fig. 4.2 Coda con ricircolo

Nel caso di ricircolo, i processi che si compongono all'ingresso non sono più indipendenti.

Trovandoci in assenza di memoria è indifferente, agli effetti delle prestazioni, se l'utente che ricircola si pone nuovamente in coda o accede immediatamente al servente.

Immaginiamo di trovarci in questa seconda ipotesi; ciò permette di considerare il tempo complessivo che un utente tiene occupato un servente ininterrottamente come soluzione di un processo con un numero infinito di stadi in forma diretta (si veda paragrafo 1 del cap. 3).

Si verifica facilmente che questa è la distribuzione esponenziale di parametro $(1 - p) \cdot \mu$.

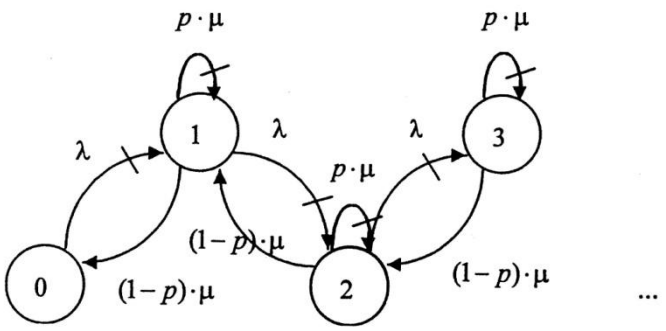


Fig. 4.3 Coda con ricircolo: M/M/1 con $\rho = \frac{\lambda}{(1-p)\cdot\mu}$

Quindi il sistema si comporta verso l'esterno ancora come una coda Markoviana ed il processo in uscita alla risorsa, ovviamente se questa è ergodica, è un processo di Poisson con parametro λ.

Il processo complessivo in ingresso, però, è la composizione degli arrivi dall'esterno e del ricircolo, e non è più un processo di Poisson.

Esempio 4.1: Processo di ingresso in una coda con ricircolo

Osserviamo innanzitutto, nella figura 4.3, che gli eventi che rappresentano un nuovo arrivo in coda sono quelli in corrispondenza delle transizioni che hanno una barretta trasversale.

Sappiamo che il tempo che intercorre fra i due eventi consecutivi di ingresso ed uscita da uno stato (tempo di permanenza nello stato) ha distribuzione esponenziale con parametro la somma delle frequenze di transizione da quello stato (frequenza totale), e che la probabilità di scattare di ciascuna transizione è il rapporto fra la frequenza di quella transizione e la frequenza totale.

Quindi se un evento di arrivo porta nello stato 1, il tasso di uscita da questo stato è $\lambda + \mu$, e l'intertempo al prossimo evento di ingresso (non necessariamente un arrivo dall'esterno) avrà:

- con probabilità $\frac{p\cdot\mu+\lambda}{\lambda+\mu}$

  distribuzione esponenziale di parametro $\lambda + \mu$ (eventi che portano un nuovo utente in coda)

- e con probabilità $\frac{(1-p)\cdot\mu}{\lambda+\mu}$

  distribuzione somma di due variabili esponenziali indipendenti di parametri rispettivamente $\lambda+\mu$ e λ (evento che non porta nessun utente in coda; quindi, dopo aver atteso nello stato è necessario attendere un nuovo arrivo dall'esterno).

La funzione caratteristica dell'intertempo fra due arrivi in coda all'entrata nello stato 1 sarà perciò:

$$\Phi_1(s) = \frac{p\cdot\mu+\lambda}{\lambda+\mu}\cdot\frac{\lambda+\mu}{s+\lambda+\mu} + \frac{(1-p)\cdot\mu}{\lambda+\mu}\cdot\frac{\lambda+\mu}{s+\lambda+\mu}\cdot\frac{\lambda}{s+\lambda}$$

che scriveremo anche, come media pesata fra due funzioni caratteristiche $F(s)$ e $G(s)$, nella forma seguente:

$$\Phi_1(s) = (1-b)\cdot F(s) + b\cdot F(s)\cdot G(s).$$

Con un ragionamento perfettamente analogo, la funzione caratteristica dell'intertempo fra due eventi di arrivo in coda all'entrata nello stato 2 sarà:

$$\Phi_2(s) = (1-b)\cdot F(s) + b\cdot F(s)\cdot \Phi_1(s)$$

ed, in generale, all'entrata nello stato i:

$$\Phi_i(s) = (1-b)\cdot F(s) + b\cdot F(s)\cdot \Phi_{i-1}(s)\,.$$

Ricordiamo che gli eventi di entrata citati prima si verificano quando si è rispettivamente negli stati $0,1,\cdots,i-1$, le cui probabilità, quale processo di nascita e morte, sono $1-\rho, (1-\rho)\cdot\rho,\cdots,(1-\rho)\cdot\rho^{i-1}$.

La funzione caratteristica cercata è complessivamente data dalla seguente serie:

$$\Phi(s) = (1-\rho)\cdot\sum_{i=0}^{\infty}\rho^i\cdot\Phi_{i+1}(s) =$$

$$= (1-\rho)\cdot\sum_{i=0}^{\infty}\rho^i\cdot\big((1-b)\cdot F(s) + b\cdot F(s)\cdot\Phi_i(s)\big)$$

che, sviluppata, porta a:

$$\Phi(s) = (1-\rho) \cdot \sum_{i=0}^{\infty} \rho^i \cdot \left((1-b) \cdot F(s) \cdot \sum_{j=0}^{i} (b \cdot F(s))^j + b \cdot F(s) \cdot \big(b \cdot F(s)\big)^i \cdot G(s) \right).$$

Utilizzando opportunamente le proprietà della serie geometrica, l'espressione precedente può essere riscritta come:

$$\Phi(s) = (1-\rho) \cdot \left((1-b) \cdot F(s) \cdot \sum_{i=0}^{\infty} \rho^i + b \cdot F(s) \cdot G(s) \right) \cdot \sum_{i=0}^{\infty} (\rho \cdot b \cdot F(s))^i$$

ed in conclusione ricondotta a:

$$\Phi(s) = (1-\rho) \cdot \left((1-b) \cdot \frac{1}{1-\rho} \cdot F(s) + b \cdot F(s) \cdot G(s) \right) \cdot \frac{1}{1-\rho \cdot b \cdot F(s)}$$

che, fatte le opportune sostituzioni e semplificazioni, diventa:

$$\Phi(s) = \frac{p \cdot \mu}{\mu - \lambda} \cdot \frac{\mu}{s+\mu} + \frac{(1-p) \cdot \mu - \lambda}{\mu - \lambda} \cdot \frac{\lambda}{s+\lambda}.$$

Questo risultato dimostra che il processo di ingresso alla coda ha intertempi con distribuzione *iperesponenziale.*

2. Reti di code aperte

Consideriamo ora una rete di code Markoviane aperte.

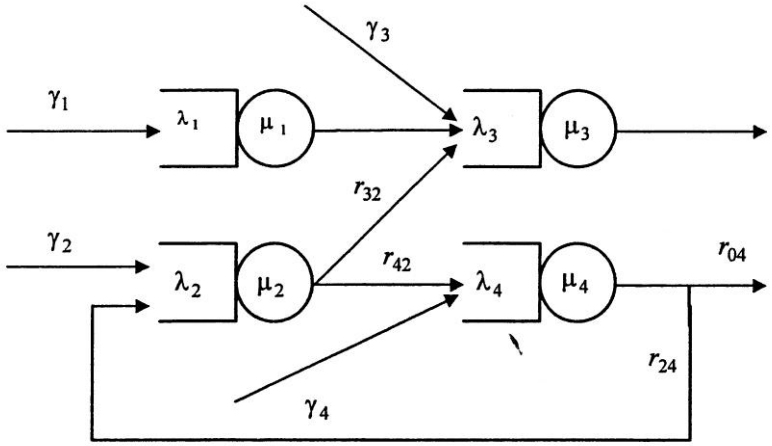


Fig. 4.4 Rete di code aperte

Inizialmente ipotizziamo che vi sia un'unica classe di utenti. I processi di ingresso sono processi di Poisson indipendentemente, le code sono M/M/m (in particolare $m = 1$), la disciplina di coda se non indicato diversamente è FIFO, gli instradamenti sono casuali, con probabilità (coefficienti di routing) r_{ij} assegnati, indipendenti dallo stato.

Questa rete, che prende il nome di rete di code Markoviane, è stato dimostrato da *Jackson* godere della proprietà della forma prodotto anche se, in presenza di ricircoli, i processi in ingresso alle code non sono propriamente di Poisson. Il suo comportamento a regime può essere studiato considerando i differenti nodi della rete isolatamente e la funzione di probabilità è ottenuta come prodotto delle funzioni di probabilità marginali del numero di utenti presso ciascun nodo.
Considerata la rete a regime ed indicati con γ_i i tassi medi dei flussi in ingresso alla rete dall'esterno verso il nodo i-esimo, con λ_i i tassi medi di attraversamento del nodo i-esimo e con r_{ij} i coefficienti d' instradamento dal nodo j al nodo i, si dimostra con una tecnica di bilanciamento dei flussi che i processi in ingresso a ciascun nodo hanno tassi dati da:

$$\lambda_i = \gamma_i + \sum_j r_{ij} \cdot \lambda_j$$

oppure, unendo le equazioni per tutti i nodi, in forma di matrice:

$$\begin{gathered}\boldsymbol{\lambda} = \boldsymbol{\gamma} + \mathbf{R} \cdot \boldsymbol{\lambda} \\ [\mathbf{I} - \mathbf{R}] \cdot \boldsymbol{\lambda} = \boldsymbol{\gamma} \\ \mathbf{R} = \{r_{ij}\}\end{gathered} \tag{4.3}$$

dove i vettori hanno la solita interpretazione ed $\mathbf{R}$ è la matrice delle probabilità d'*instradamento* (*routing*). La soluzione di questo sistema lineare di equazioni, che esiste poiché, essendovi flussi da e verso l'esterno, la matrice $\mathbf{R}$ non ha autovalori uguali ad 1, fornisce i tassi di ingresso a ciascun servizio:

$$\boldsymbol{\lambda} = [\mathbf{I} - \mathbf{R}]^{-1} \cdot \boldsymbol{\gamma}\,. \tag{4.4}$$

La rete si dice *ergodica* quando ciascuna coda della rete è ergodica; è necessario quindi che:

$$\frac{\lambda_i}{m_i \cdot \mu_i} < 1, \quad \forall i \tag{4.5}$$

dove λ_i è la componente del vettore soluzione di (4.4), μ_i è il tasso di servizio e m_i il numero di serventi relativi alla coda i-esima.
Il numero medio di utenti nella rete è pari alla somma del numero medio di utenti presenti in ciascun nodo $\sum N_i$ ed, attraverso la legge di Little, il tempo medio di permanenza nella rete risulta:

$$\Delta T_{tot} = \frac{\sum_i N_i}{\sum_i \gamma_i}\,.$$

2.1.Legge di Little nel grande

Si noti che la legge di Little, già vista in relazione ad una singola coda, si applica anche per un'intera sottorete o rete di code, nel senso che, isolato un qualsiasi sott'insieme di una rete ergodica, il numero medio di utenti presenti in quel sott'insieme è dato dal prodotto del tempo medio di permanenza degli utenti al suo interno per la somma di tutti i flussi entranti.

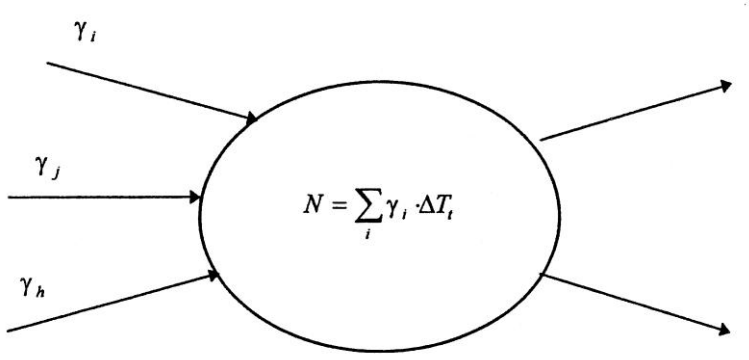


Fig. 4.5 Legge di Little nel grande

Esempio 4.2: Flussi medi di utenti in una rete di code aperte

I flussi di utenti della rete di figura 4.4 sono dati da:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -r_{24} \\ -1 & -r_{32} & 1 & 0 \\ 0 & -r_{42} & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix}$$

I tassi di utilizzo delle stazioni, nell'ipotesi che siano monoserventi è:

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{\mu_1} & 0 & 0 & 0 \\ 0 & \frac{1}{\mu_2(1-r_{24}r_{42})} & 0 & \frac{r_{24}}{\mu_2(1-r_{24}r_{42})} \\ \frac{1}{\mu_3} & \frac{r_{32}}{\mu_3(1-r_{24}r_{42})} & \frac{1}{\mu_3} & \frac{r_{32}r_{24}}{\mu_3(1-r_{24}r_{42})} \\ 0 & \frac{r_{42}}{\mu_4(1-r_{24}r_{42})} & 0 & \frac{1}{\mu_4(1-r_{24}r_{42})} \end{bmatrix} \cdot \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix}.$$

Teorema 4.2: Teorema di Jackson. *Forma prodotto nelle reti di code aperte*
In una rete di code Markoviane ergodiche con le condizioni poste all'inizio, indicati con n_i il numero di utenti presenti presso la risorsa i-esima, a regime la funzione densità di probabilità congiunta della distribuzione degli utenti presso i diversi nodi è il prodotto delle funzioni di densità di probabilità marginali del numero di utenti presso ciascun nodo:

$$\pi(n_1, n_2, \cdots, n_N) = \pi_1(n_1) \cdot \pi_2(n_2) \cdots \pi_N(n_N)$$

dove le probabilità marginali sono calcolate con la teoria classica delle code, assumendo come tassi di arrivo a ciascuna coda i valori forniti da (4.4).

Va osservato, a conclusione, che due aspetti sono critici per l'esistenza della forma prodotto: gli instradamenti casuali ed indipendenti dallo stato e le code a capacità infinita. In altri termini, la presenza di una politica d' instradamento deterministico, come ad esempio quando ogni utente in uscita da un servizio è indirizzato fra i servizi in alternativa a quello con la coda minore, fa cadere la forma prodotto. Lo stesso avviene quando il riempimento di una coda di capacità finita causa il bloccaggio dei nodi all'origine, impedendo lo scarico degli utenti che hanno completato il servizio.

Esempio 4.3: Rete di code aperte
Consideriamo la seguente rete di code aperte, in cui il secondo nodo è un biservente.

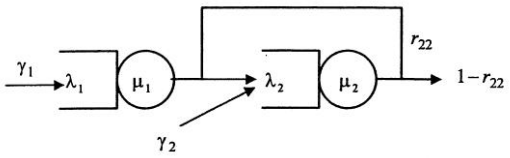


Fig. 4.6 Rete di code aperte

I tassi dei flussi d'attraversamento di ciascun nodo sono:

$$\begin{bmatrix} 1 & 0 \\ -1 & 1 - r_{22} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$$

$$\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \begin{bmatrix}\gamma_1\\ \frac{1}{1-r_{22}}\cdot(\gamma_1+\gamma_2)\end{bmatrix}$$

da cui i coefficienti di traffico risultano:

$$\begin{bmatrix}\rho_1\\ \rho_2\end{bmatrix} = \begin{bmatrix}\frac{\gamma_1}{\mu_1}\\ \frac{\gamma_1+\gamma_2}{(1-r_{22})\cdot\mu_1}\end{bmatrix}.$$

La probabilità di stato della rete è:

$$\pi(n_1, n_2, \cdots, n_N) = (1-\rho_1)\cdot\rho^{n_1}\cdot\frac{2-\rho_2}{2+\rho_2}\cdot\frac{{\rho_2}^{n_2}}{m}, \quad m = \begin{cases}1 & se\ n_2 < 2\\ 2 & se\ n_2 \geq 2\end{cases}.$$

Da questi dati è possibile calcolare il numero medio di utenti nella rete ed il tempo medio di attraversamento:

$$N_{tot} = \frac{\rho_1}{1-\rho_1} + \rho_2 + \frac{2\cdot{\rho_2}^2}{4-{\rho_2}^2}$$

$$\Delta T_{tot} = \frac{1}{\gamma_1+\gamma_2}\cdot\left(\frac{\rho_1}{1-\rho_1} + \rho_2 + \frac{2\cdot{\rho_2}^2}{4-{\rho_2}^2}\right).$$

Il risultato di Jackson si estende immediatamente al caso di *classi multiple di utenti*, dove ciascuna classe ha proprie probabilità d'instradamento, con i seguenti vincoli sui nodi:

- M/M/1 monoserventi con tassi di servizio uguali per tutte le classi;
- M/M/∞ multiserventi con un numero infinito di serventi (ritardi puri), con tempi di servizio possibilmente diversi per ciascuna classe; non sono permessi invece multiserventi con numero finito di serventi.

CAPITOLO 5
Reti di code chiuse

In una rete di code chiusa non esiste alcun flusso di utenti da e verso l'esterno della rete.

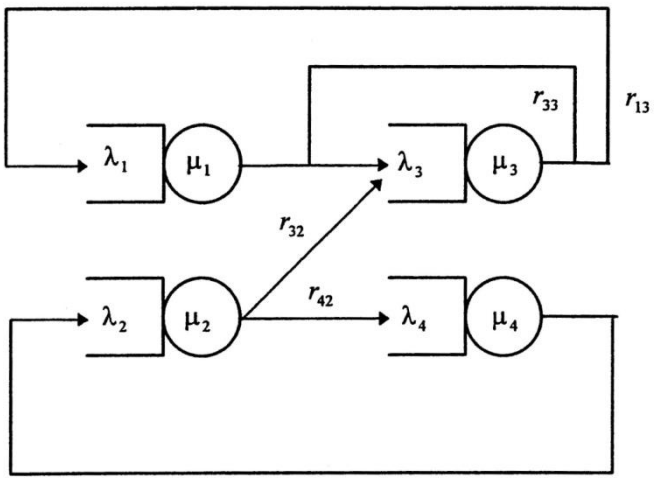


Fig. 5.1 Rete di code chiuse

Queste reti sono caratterizzate, quindi, dall'avere un numero fisso di utenti e quindi un numero finito di stati, che dipende dal numero dei nodi e degli utenti presenti. Infatti, se indichiamo con K il numero di utenti e con N il numero dei nodi, il numero degli stati possibili del processo è uguale al coefficiente binomiale:

$$\binom{N+K-1}{N-1} = \binom{N+K-1}{K} = \frac{(N+K-1)!}{K!\,(N-1)!}.$$

La struttura del grafo dell'automa (da non confondere con il grafo della rete di code) di una rete di code chiuse è particolare, in quanto da uno stato $(k_1,\cdots,k_j\cdots,k_i,\cdots,k_N)$, all'uscita di un utente da uno dei nodi j $(j = 1,2,\dots,N)$ e suo ingresso in uno dei nodi successivi i, è possibile portarsi soltanto in una serie di stati contigui caratterizzati dalla

perdita di un utente in j ed il suo acquisto in i $(k_1,\cdots,k_j-1\cdots,k_i+1,\cdots,k_N)$, con tasso di transizione $r_{ij}\cdot\mu_j$.

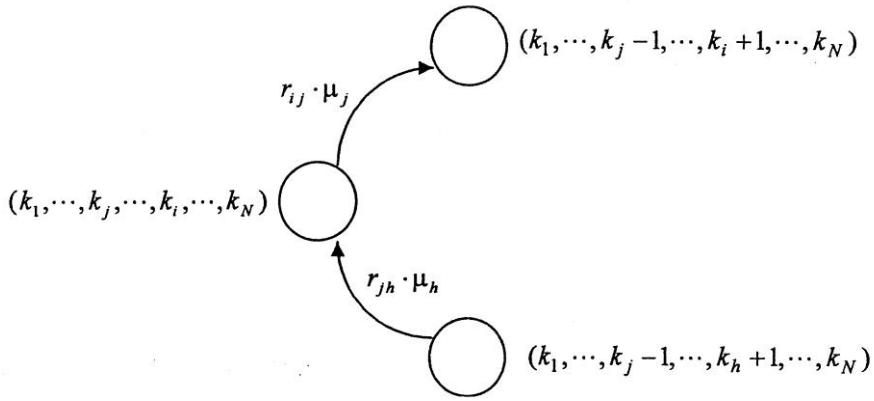


Fig. 5.2 Struttura dell'automa di una rete di code chiuse

Esempio 5.1: Tre code in cascata con due utenti

Quest'esempio mette in mostra la struttura dell'automa associato ad una rete di code chiuse.

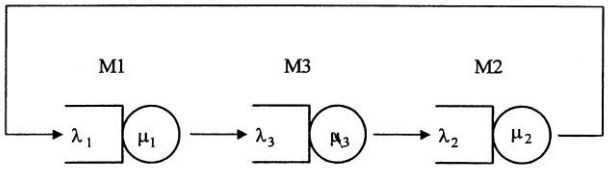


Fig. 5.3 Esempio di rete di code chiuse

La figura 5.4 rappresenta il grafo dell'automa associato quando vi sono due utenti nella rete, che ha $\binom{N+K-1}{N-1}=6$ stati (numero di utenti nelle stazioni M1,M2,M3).

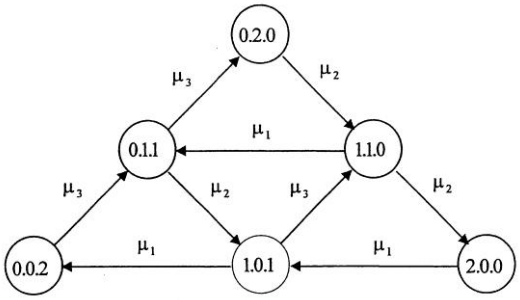


Fig. 5.4 Automa associato ad una rete di code chiuse

1. Bilanciamento dei flussi in reti di code chiuse

E' possibile determinare le condizioni di equilibrio mediante il bilanciamento dei flussi. Considerato allora un generico stato $(k_1,\cdots,k_j\cdots,k_i,\cdots,k_N)$, il flusso totale uscente da questo stato è dato dalla somma dei contributi di tutti i nodi che hanno numero di utenti diverso da zero:

$$\sum_{i=1;\, k_i>0}^{N} \mu_i \cdot \pi\left(k_1,\cdots,k_j\cdots,k_i,\cdots,k_N\right) \tag{5.1}$$

mentre il flusso entrante dagli stati contigui è:

$$\sum_{j=1;\, k_j>0}^{N} \sum_{i=0}^{N} r_{ji} \cdot \mu_i \cdot \pi\left(k_1,\cdots,k_j-1\cdots,k_i+1,\cdots,k_N\right). \tag{5.2}$$

Questo risultato offre un primo rudimentale metodo per risolvere la rete.
Per il principio del bilanciamento dei flussi, le due espressioni (5.1) e (5.2) devono essere uguali.

Esempio 5.2: Tre code in cascata con due utenti, risolto attraverso il bilanciamento dei flussi

Ricaviamo le probabilità di stato dell'esempio di figura 5.4:

$$\pi(1.0.1)=\frac{\mu_1}{\mu_3}\pi(2.0.0),\qquad \pi(1.1.0)=\frac{\mu_1}{\mu_2}\pi(2.0.0),\qquad \pi(0.1.1)=\frac{{\mu_1}^2}{\mu_2\mu_3}\pi(2.0.0),$$

$$\pi(0.0.2)=\frac{{\mu_1}^2}{{\mu_3}^2}\pi(2.0.0),\qquad \pi(0.2.0)=\frac{{\mu_1}^2}{{\mu_2}^2}\pi(2.0.0)$$

$$\pi(2.0.0)=\left(1+\frac{\mu_1}{\mu_3}+\frac{\mu_1}{\mu_2}+\frac{{\mu_1}^2}{\mu_2\mu_3}+\frac{{\mu_1}^2}{{\mu_3}^2}+\frac{{\mu_1}^2}{{\mu_2}^2}\right)^{-1}.$$

Per conoscere i flussi di attraversamento di ciascun nodo, uguali tra loro in questo caso, è sufficiente calcolare il flusso uscente da una delle stazioni scelta a piacere, ad esempio la terza:

$$\lambda = \mu_3 \cdot \left(\frac{\mu_1}{\mu_3} + \frac{{\mu_1}^2}{\mu_2 \mu_3} + \frac{{\mu_1}^2}{{\mu_3}^2} \right) \cdot \pi(2.0.0) \, .$$

2. Forma prodotto in reti di code chiuse

Anche per le reti di code chiuse, sotto condizioni analoghe a quelle delle reti di code aperte (non è necessario qui la specifica sui flussi dall'esterno), è verificata la proprietà della *forma prodotto*. Questo risultato è dovuto a *Gordon* e *Newell*, che offrono una variante del teorema di *Jackson* per le reti di code chiuse.

I tassi medi di ingresso (e quindi di uscita) alle macchine si calcolano, come per le reti di code aperte (4.3), dal seguente sistema lineare di equazioni, che in questo caso è omogeneo:

$$\lambda_i = \sum_{j=1}^{N} r_{ij} \cdot \lambda_j$$

$$[\mathbf{I} - \mathbf{R}] \cdot \boldsymbol{\lambda} = \mathbf{0} \, .$$

Il sistema di equazioni omogeneo ha matrice dei coefficienti singolare, in quanto non essendovi flussi da e verso l'esterno la matrice $\mathbf{R}$ ha un auto valore uguale ad 1. La soluzione, ovviamente a meno di un fattore di proporzionalità h, fornisce i tassi di ingresso a ciascun nodo e con loro le intensità di traffico dei nodi risultano:

$$\rho_i = \frac{\lambda_i \cdot h}{\mu_i} \, . \tag{5.3}$$

Dai risultati dei processi di nascita e morte, le probabilità marginali per ogni coda, a meno di un comune fattore di proporzionalità, sono dati dalla teoria classica delle code, ad esempio per code monoservente:

$$\pi_i(k_i) = {\rho_i}^{k_i}$$

oppure per i multiserventi:

$$\pi_i(k_i) = \frac{{\rho_i}^{k_i}}{1 \cdot 2 \cdot \cdots}.$$

La densità di probabilità congiunta in forma prodotto assume, perciò, (considerando solo il caso monoservente) la seguente forma:

$$\pi(k_1, k_2, \cdots, k_N) = \prod_{i=1}^{N} \pi_i = \frac{1}{G} \prod_{i=1}^{N} {\rho_i}^{k_i}$$

dove la costante di normalizzazione G è calcolata in modo da garantire che le espressioni precedenti siano consistenti con la definizione di densità di probabilità, cioè la somma di tutte le probabilità di stato uguale a 1:

$$\frac{1}{G} \cdot \sum_{k_1, k_2, \cdots \epsilon K} \prod_{i=1}^{N} {\rho_i}^{k_i} = 1 \rightarrow G = \sum_{k_1, k_2, \cdots \epsilon K} \prod_{i=1}^{N} {\rho_i}^{k_i} \ . \tag{5.4}$$

Note le probabilità di stato, la risoluzione del grado di indeterminazione che ancora sussiste per i flussi d'attraversamento delle stazioni, rappresentato dal parametro h, e dei relativi fattori di traffico (5.3), si ha valutando il flusso di uscita da una qualsiasi stazione come somma dei flussi di uscita da tutti gli stati del grafo per cui quella stazione non è vuota, ad esempio per l'i-esima stazione monoservente:

$$\lambda_i = \sum_{n_1, \cdots n_i, \cdots = 0, K;\ n_i \neq 0} \pi(n_1, \cdots, n_i, \cdots) \cdot \mu_i. \tag{5.5}$$

Esempio 5.3: Rete di code chiuse risolta con il bilanciamento dei flussi

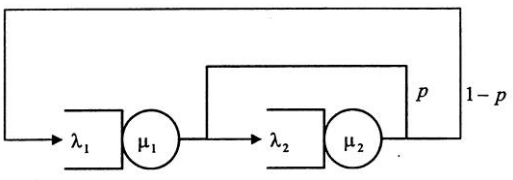


Fig. 5.5 Rete di code chiuse con ricircolo un numero indefinito di volte

Consideriamo una rete costituita da due code monoservente in cascata, la seconda delle quali può fallire un numero indefinito di volte l'operazione e con probabilità p deve ripeterla.
La macchina a stati che descrive questa rete nel caso di un singolo utente è la seguente:

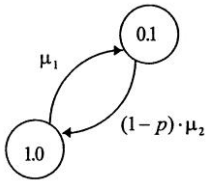


Fig. 5.6 Grafo dell'automa della rete precedente

Le probabilità di stato a regime sono:

$$\mu_1 \cdot \pi_{1.0} = (1-p) \cdot \mu_2 \cdot \pi_{0.1}$$

$$\pi_{0.1} = \frac{\mu_1}{(1-p) \cdot \mu_2} \pi_{1.0}$$

$$\pi_{1.0} = \frac{(1-p) \cdot \mu_2}{(1-p) \cdot \mu_2 + \mu_1}$$

$$\pi_{0.1} = \frac{\mu_1}{(1-p) \cdot \mu_2 + \mu_{12}}$$

Il tasso di circolazione nella rete è:

$$\lambda = \lambda_1 = \frac{(1-p) \cdot \mu_2 \cdot \mu_1}{(1-p) \cdot \mu_2 + \mu_1}.$$

Esempio 5.4: Rete di code chiuse risolta utilizzando la forma prodotto
La struttura della rete è la stessa dell'esempio precedente; la seconda coda è, però, un biservente e nella rete sono presenti rispettivamente 2 e 3 utenti.

Il vettore dei tassi di ingresso ai diversi nodi è:

$$\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \begin{bmatrix}1\\ \frac{1}{1-p}\end{bmatrix} \cdot h\,.$$

Le probabilità di stato marginali, a meno di una costante di proporzionalità sono, per la prima coda:

$$\pi_1(0) = 1, \qquad \pi_1(1) = \frac{h}{\mu_1}, \qquad \pi_1(2) = \left(\frac{h}{\mu_1}\right)^2, \qquad \pi_1(3) = \left(\frac{h}{\mu_1}\right)^3,$$

e per la seconda coda:

$$\pi_2(0) = 1, \qquad \pi_2(1) = \frac{h}{(1-p)\cdot\mu_2},$$

$$\pi_2(2) = \frac{1}{2}\left(\frac{h}{(1-p)\cdot\mu_2}\right)^2, \qquad \pi_2(3) = \frac{1}{4}\left(\frac{h}{(1-p)\cdot\mu_2}\right)^3.$$

Risolviamo prima il caso con due utenti. Le probabilità congiunte sono:

$$\pi(2.0) = \frac{1}{G}\left(\frac{h}{\mu_1}\right)^2, \qquad \pi(1.1) = \frac{1}{G}\frac{h}{\mu_1}\frac{h}{(1-p)\cdot\mu_2},$$

$$\pi(0.2) = \frac{1}{G}\frac{1}{2}\left(\frac{h}{(1-p)\cdot\mu_2}\right)^2.$$

Il calcolo della costante di normalizzazione porta a:

$$\frac{h^2}{G}\left[\left(\frac{1}{\mu_1}\right)^2 + \frac{1}{(1-p)\mu_1\mu_2} + \frac{1}{2}\left(\frac{1}{(1-p)\mu_2}\right)^2\right] = 1$$

$$\frac{h^2}{G} = \frac{2(1-p)^2{\mu_1}^2{\mu_2}^2}{2(1-p)^2{\mu_2}^2 + 2(1-p)\mu_1\mu_2 + {\mu_1}^2}.$$

La grandezza che fornisce le prestazioni della rete è principalmente il tasso di circolazione, che in questo caso coincide con il tasso di ingresso (uscita) alla prima stazione:

$$\lambda(2) = \lambda_1 = \big(\pi(2.0) + \pi(1.1)\big) \cdot \mu_1$$

$$\lambda(2) = \frac{2(1-p)^2 {\mu_1}^2 {\mu_2}^2}{2(1-p)^2 {\mu_2}^2 + 2(1-p)\mu_1\mu_2 + {\mu_1}^2}\left[\left(\frac{1}{\mu_1}\right)^2 + \frac{1}{(1-p)\mu_1\mu_2}\right] \cdot \mu_1 =$$

$$= \frac{2(1-p)\mu_1\mu_2[(1-p)\mu_2 + \mu_1]}{2(1-p)^2 {\mu_2}^2 + 2(1-p)\mu_1\mu_2 + {\mu_1}^2}.$$

Assegnamo ai parametri i seguenti valori numerici:

$$p = 0.7, \qquad \mu_1 = 5.0, \qquad \mu_2 = 3.0\,.$$

Le probabilità di stato sono:

$$\pi(2.0) = 0.0455, \qquad \pi(1.1) = 0.2527, \qquad \pi(0.2) = 0.702\,.$$

Il tasso di circolazione risulta:

$$\lambda(2) = 1.49\,.$$

Il numero medio di utenti alle singole stazioni:

$$N_1 = 0.3437, \qquad N_2 = 1.6563\,.$$

Esempio 5.5: Rete di code chiuse con tre utenti risolta utilizzando la forma prodotto

L'esempio precedente è trattato ora con tre utenti.

Le probabilità congiunte degli stati sono:

$$\pi(3.0) = \frac{1}{G}\left(\frac{h}{\mu_1}\right)^3, \qquad \pi(2.1) = \frac{1}{G}\left(\frac{h}{\mu_1}\right)^2 \frac{h}{(1-p)\cdot\mu_2},$$

$$\pi(1.2) = \frac{1}{G}\frac{h}{\mu_1}\frac{1}{2}\left(\frac{h}{(1-p)\cdot\mu_2}\right)^2, \qquad \pi(0.3) = \frac{1}{G}\frac{1}{4}\left(\frac{h}{(1-p)\cdot\mu_2}\right)^3.$$

Il calcolo della costante di normalizzazione porta a:

$$\frac{h^3}{G}\left[\left(\frac{1}{\mu_1}\right)^3 + \frac{1}{(1-p){\mu_1}^2\mu_2} + \frac{1}{2\mu_1}\left(\frac{1}{(1-p)\mu_2}\right)^2 + \frac{1}{4}\left(\frac{1}{(1-p)\mu_2}\right)^3\right] = 1$$

$$\frac{h^3}{G} = \frac{4(1-p)^3{\mu_1}^3{\mu_2}^3}{4(1-p)^3{\mu_2}^3 + 4\mu_1(1-p)^2{\mu_2}^2 + 2{\mu_1}^2(1-p)\mu_2 + {\mu_1}^3}.$$

La grandezza che fornisce le prestazioni della rete è principalmente il tasso di circolazione, che in questo caso coincide con il tasso di ingresso (uscita) alla prima stazione:

$$\lambda(3) = \lambda_1 = \big(\pi(3.0) + \pi(2.1) + \pi(1.2)\big)\cdot\mu_1$$

$$\lambda(3) = \frac{h^3}{G}\left[\left(\frac{1}{\mu_1}\right)^3 + \frac{1}{(1-p){\mu_1}^2\mu_2} + \frac{1}{2\mu_1}\left(\frac{1}{(1-p)\mu_2}\right)^2\right]\mu_1$$

$$\lambda(3) = \frac{4(1-p)^3{\mu_1}^3{\mu_2}^3}{4(1-p)^3{\mu_2}^3 + 4\mu_1(1-p)^2{\mu_2}^2 + 2{\mu_1}^2(1-p)\mu_2 + {\mu_1}^3}$$

$$\left[\left(\frac{1}{\mu_1}\right)^3 + \frac{1}{(1-p){\mu_1}^2\mu_2} + \frac{1}{2\mu_1}\left(\frac{1}{(1-p)\mu_2}\right)^2\right]\mu_1 =$$

$$= \frac{2(1-p)\mu_1\mu_2[2(1-p)^2{\mu_2}^2 + 2\mu_1(1-p)\mu_2 + {\mu_1}^2]}{4(1-p)^3{\mu_2}^3 + 4\mu_1(1-p)^2{\mu_2}^2 + 2{\mu_1}^2(1-p)\mu_2 + {\mu_1}^3}.$$

Con gli stessi parametri dell'esempio precedente, il tasso di circolazione risulta ora:

$$\lambda(3) = 1.69\,.$$

Se confrontiamo questo risultato con il precedente, osserviamo un fenomeno generale: la produttività della rete aumenta in modo monotono al crescere della dimensione della popolazione.

Questo è dovuto al fatto che con un maggior numero di utenti aumenta l'utilizzo delle risorse. Se tracciassimo un grafico del tasso di circolazione, funzione della dimensione della popolazione, vedremmo che questo grafico ha un andamento asintotico, tende cioè ad un valore massimo che identifica la produttività idealmente ottenibile dalle risorse disponibili avendo una dimensione di popolazione infinita. Poiché le risorse in un sistema hanno sempre un costo, l'analisi dell'andamento asintotico del grafico permette una scelta ottima della dimensione della popolazione.

Come per le reti di code aperte anche per quelle chiuse la proprietà di forma prodotto si estende a classi multiple di utenti, limitatamente però al caso di stazioni monoservente con tassi di servizio uguali per ciascuna classe.
La dimensione dello stato cresce però così rapidamente da rendere la tecnica basata su questa proprietà non utilizzabile in pratica.

3. Reti BCMP

Sino a questo punto ci siamo limitati a trattare reti di code Markoviane. Abbiamo anticipato, però, che non è una prerogativa delle sole code Markoviane garantire reti con densità di probabilità congiunta in forma prodotto. La famiglia più ampia di reti con le precedenti proprietà oggi conosciuta è stata caratterizzata da *Basket*, *Chandy*, *Muntz* e *Palacios*, che, dall'acronimo delle iniziali dei nomi degli autori, è indicata come famiglia di reti BCMP.

Teorema 5.1: Reti di code BCMP

Hanno densità di probabilità congiunta in forma prodotto reti di code aperte, chiuse o miste (aperte con ricircolo), in presenza di classi multiple di utenti con le condizioni elencate. Gli instradamenti sono probabilistici ed indipendenti dallo stato; gli arrivi dall'esterno, nel caso di reti di code aperte, sono processi di *Poisson*, con tassi istantanei che possono dipendere dalla dimensione della popolazione nella rete;

Le code che costituiscono la rete sono dei tipi seguenti:

1. Monoserventi in presenza di classi multiple di utenti, tutte con la medesima distribuzione esponenziale dei tempi di servizio, e disciplina di code FIFO;
2. Multiserventi in presenza di una sola classe di utenti, con tempi di servizio distribuiti esponenzialmente, e disciplina di coda FIFO;
3. Multiserventi con infiniti serventi in presenza di classi multiple di utenti, con distribuzioni dei tempi di servizio diverse fra loro, aventi funzioni caratteristiche razionali;
4. Monoserventi in presenza di classi multiple di utenti, con distribuzioni dei tempi di servizio diverse tra loro, aventi funzioni caratteristiche razionali, con politiche di coda ad interruzione "*pre-emption*" del tipo a distribuzione uniforme del servente "*processor sharing"*, oppure ultimo arrivato primo servito "Last Come First Served" LCFS.

La condizione che la funzione caratteristica dei tempi di servizio sia razionale non è molto restrittiva.
Le politiche di coda con interruzione prevedono che un utente venga servito in tempi parziali successivi: nella disciplina "processor sharing" gli utenti presenti si distribuiscono uniformemente a turno il servente, ciascuno con periodi di servizio che tendono a zero; nella disciplina LCFS l'ultimo utente arrivato interrompe l'utente in servizio, sino a completare il suo servizio o essere a sua volta interrotto da un nuovo arrivato. Queste politiche di coda trovano applicazione soprattutto nell'assegnazione dell'unità centrale ai diversi programmi in attesa nei sistemi di elaborazione.

4. Fattori di visita e tassi di circolazione

Abbiamo visto che reti di code aperte e reti di code chiuse possono essere trattate sostanzialmente applicando le stesse tecniche di soluzione, basate sul concetto di bilanciamento dei flussi oppure sfruttando la proprietà della forma prodotto. Appare chiaro, però, che al crescere delle dimensioni, mentre le reti di code aperte richiedono la soluzione di un sistema lineare di equazioni che, sia pure grande, è sempre alla portata delle moderne tecniche di calcolo numerico, le reti di code chiuse assumono rapidamente, al crescere del numero delle stazioni e degli utenti, dimensioni proibitive

dovute all'esplosione in modo combinatorio del numero degli stati dell'automa che rappresenta. E' necessario perciò per queste reti cambiare l'ottica con cui si osserva il fenomeno.

Prima il punto di partenza era costituito dal calcolo delle probabilità di stato, e quindi da queste erano derivate tutte le statistiche significative del processo. Si ricordino, per questo, le formule (5.4) e (5.5).

Adesso, grazie alla forma chiusa della rete, è possibile portare in primo piano i flussi medi di circolazione all'interno della rete e tutte le statistiche d'interesse del problema sono poi calcolate direttamente da questi, senza necessità di conoscere le probabilità di stato.

Per entrare nella nuova ottica osserviamo che gli utenti presenti in una rete di code chiuse compiono ciclicamente visita ai nodi della rete; quindi definiamo con λ il *tasso medio di circolazione* come il numero di cicli effettuati nell'unità di tempo dalla totalità degli utenti. Se la popolazione ha dimensione K, il *tempo medio per compiere un ciclo* da parte di un utente è invece K/λ, ed è definito come il rapporto fra il numero di utenti della rete ed il tasso di circolazione.

Associato al tasso di circolazione, esisterà un tasso d'attraversamento di ciascuna risorsa della rete λ_i e da questo si può definire il *fattore di visita* come il numero di volte che una risorsa è visitata per ciclo. E' questo il rapporto fra il flusso d'attraversamento di una risorsa ed il tasso di circolazione della rete $v_i = \lambda_i/\lambda$.

Per stabilire in modo non ambiguo queste grandezze è necessario fare riferimento alle specifiche del problema. In molte circostanze il completamento di un ciclo nella rete è identificato dall'uscita dalla stazione terminale o, nel caso di più stazioni in parallelo, dall'ingresso nella prima stazione del ciclo. Questa stazione è indicata come stazione di riferimento della rete e per questa si assume convenzionalmente fattore di visita uguale ad uno, da cui risultano i fattori di visita di tutte le altre stazioni. In alcuni casi i fattori di visita si ricavano dai coefficienti d'instradamento della rete, in altri la situazione è invertita, in quanto il problema specifica i fattori di visita, mentre i coefficienti d'instradamento non sono conosciuti.

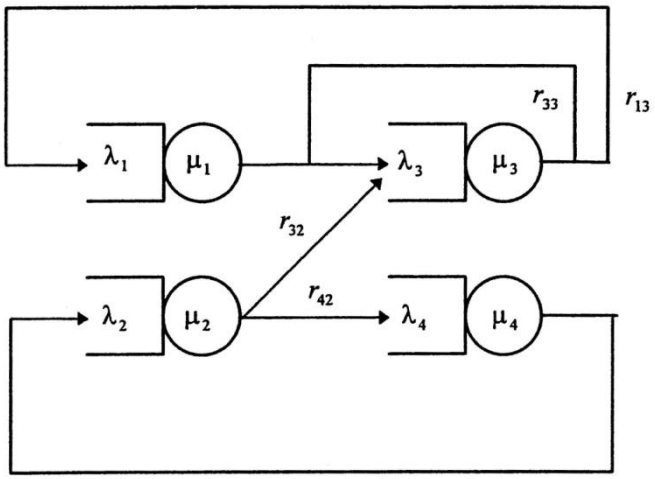


Fig. 5.7 Rete di code chiuse e tassi di visita

Esempio 5.6: Tassi di visita dalle probabilità d'instradamento

Nell'esempio seguente un ciclo nella rete si completa tutte le volte che un utente ritorna alla stazione 1, che ha qui il ruolo di stazione di riferimento.

Il tasso di circolazione coincide perciò con il tasso di ingresso nella stazione 1, mentre il bilanciamento dei flussi delle diverse stazioni, dall'analisi dei coefficienti d'instradamento, porta ai fattori di visita:

$$\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{bmatrix} = \begin{bmatrix}\nu_1\\ \nu_2\\ \nu_3\end{bmatrix} \cdot \lambda = \begin{bmatrix}1\\ \frac{r_{21}}{p}\\ r_{31}\end{bmatrix} \cdot \lambda \,.$$

Esempio 5.7: Probabilità d'instradamento dai fattori di visita

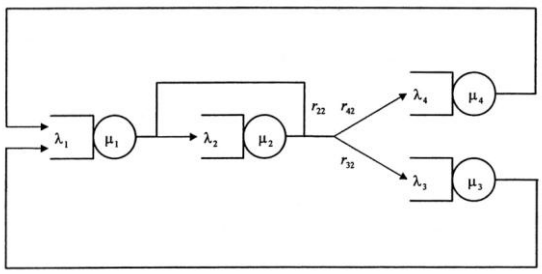


Fig. 5.8 Esempio di rete di code chiuse

In molti casi, in pratica, il processo è invertito rispetto al caso precedente, in quanto è direttamente specificato il ciclo dei nodi visitati con i loro fattori di visita; da questi si risale ai coefficienti d'instradamento. Cerchiamo di risalire alla struttura della rete ed analizziamo i problemi che questi casi comportano.
La seguente specifica:

$$[S_1, 2 \cdot S_2, 0.5 \cdot S_3 | 0.5 \cdot S_4] \tag{5.6}$$

indica che in un ciclo gli utenti devono visitare una volta S_1, due volte S_2 ed in alternativa, con uguale probabilità, S_3 oppure S_4.
Una struttura di rete, rappresentativa della specifica precedente, è quella di figura 5.8. Le probabilità d'instradamento si ricavano dai fattori di visita, osservando che il vettore dei fattori di visita, così come il vettore dei tassi d'attraversamento dei nodi, è una soluzione del sistema omogeneo di equazioni dato dal bilanciamento dei flussi:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0.5 \\ 0.5 \end{bmatrix} \cdot \lambda$$

$$[\mathbf{I} - \mathbf{R}] \cdot \boldsymbol{\lambda} = \begin{bmatrix} 1 & 0 & -1 & -1 \\ -1 & 1-r_{22} & 0 & 0 \\ 0 & -r_{32} & 1 & 0 \\ 0 & -r_{42} & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 0.5 \\ 0.5 \end{bmatrix} \cdot \boldsymbol{\lambda} = 0$$

da cui si ottiene $r_{22} = 0.5,\ r_{32} = 0.25,\ r_{42} = 0.25$.

Attenzione però, gli strumenti che abbiamo a disposizione permettono di risolvere soltanto reti di code Markoviane con la proprietà della forma prodotto. Questo modello prevede che gli utenti si muovano lungo la rete con instradamento casuale, di cui i coefficienti di routing rappresentano le probabilità. Quindi i risultati che si ottengono si riferiscono al caso in cui la specifica del ciclo di visite rappresentato dalla (5.6) è soddisfatta da ciascun utente soltanto in media, ma non in modo puntuale per ciascun suo ciclo di lavorazione.

Se con le specifiche precedenti si intendeva, invece, che in modo deterministico in un ciclo ogni utente deve visitare esattamente due volte la stazione 2, il modello offerto dalla forma prodotto non rappresenta più il problema.

Esempio 5.8: Rete non in forma prodotto

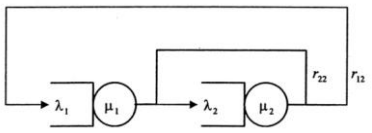


Fig. 5.9 Rete di code chiuse non in forma prodotto

Consideriamo per necessità un caso semplice, ottenuto modificando leggermente l'esempio 5.3 precedente. Una rete è costituita da due utenti e due nodi in cascata, che gli utenti visitano ciclicamente; all'uscita dal secondo nodo, però, con probabilità p ogni utente può fallire l'operazione e deve ripeterla, ma questo soltanto la prima volta di ogni ciclo.

E' immediato verificare che il tasso di visita al secondo nodo è $1+p$, cui corrispondono le seguenti percentuali dei tassi dei flussi secondo ciascun percorso (non possiamo più chiamarle probabilità d'instradamento):

$$r_{12} = \frac{1}{1+p}\,, \quad r_{22} = \frac{p}{1+p}\,.$$

La rete, però, con le specifiche precedenti perde la proprietà della forma prodotto, in quanto l'instradamento non è casuale e per ogni utente uscente dal secondo nodo si deve mantenere memoria se è al primo oppure al secondo passaggio.

Il modello classico di una rete di code Markoviane non rappresenta correttamente il problema.

In questo caso si può trovare una catena di Markov che rappresenta esattamente il problema, anche se non una nuova rete di code Markoviane, scegliendo un nuovo modello con un numero di stati maggiori del precedente.

Infatti va osservato che se il processo è rappresentabile mediante una catena di Markov, non significa che è automaticamente associato ad una rete di code Markoviane con proprietà della forma prodotto.
Questo risultato è interessante, perché è un fatto abbastanza generale poter approssimare un processo arbitrariamente complesso mediante una catena di Markov, a condizione di far crescere sufficientemente la dimensione dello spazio degli stati del modello. I risultati delle code risolte con il modello degli stadi sono un esempio.

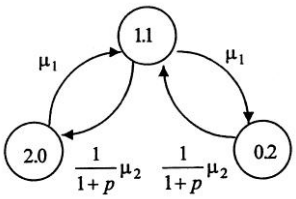


Fig. 5.10 Grafo di transizione dello stato nell'ipotesi della forma prodotto

$p = 0.7, \qquad \mu_1 = 5, \qquad \mu_2 = 3$

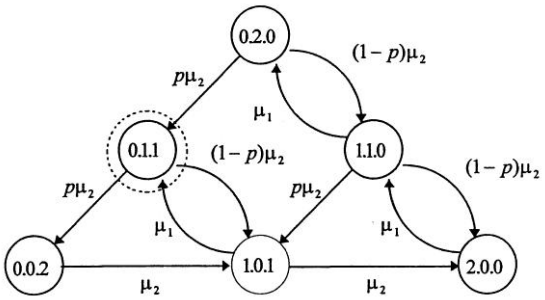
Già sappiamo che una rete di code Markoviane chiuse è rappresentabile mediante una catena di Markov di dimensioni finite. Quindi, se la figura precedente (5.9) rappresentasse una rete di code con instradamenti casuali che soddisfa le condizioni di forma prodotto, si avrebbe con due utenti il grafo di transizione degli stati di figura 5.10, dove gli stati rappresentano il numero di pezzi presenti in ciascuna stazione.
Con un semplice bilanciamento dei flussi si hanno, da questo grafo, le seguenti probabilità di stato e tasso di circolazione:

$$\pi(2.0) = 0.084, \qquad \pi(1.1) = 0.2388, \qquad \pi(0.2) = 0.6768$$
$$\lambda(2) = 1.61\,.$$

Le specifiche del problema, con instradamenti deterministici, non permettono invece di utilizzare la forma prodotto. Il grafo che rappresenta il sistema non è più quello della figura 5.10; il processo è tuttavia Markoviano, ed esistono ben due grafi differenti che modellano correttamente le specifiche del problema e sono ottenuti aumentando la dimensione dello stato per rappresentare

rispettivamente il numero di utenti nella stazione uno, e nella stazione due al primo ed al secondo passaggio.

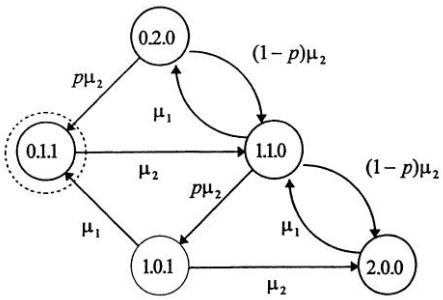
L'esistenza di due grafi è giustificata dal fatto che non è dichiarato dalle specifiche se, terminato il servizio, in caso di ricircolo l'utente viene rimesso in coda o accede immediatamente al servente. Più precisamente, il primo grafo rappresenta la situazione in cui la ripetizione dell'operazione sulla seconda macchina è sempre servita per ultima, ponendo l'utente in coda (stato evidenziato).



$$\pi(2.0.0) = 0.095, \quad \pi(1.1.0) = 0.0592, \quad \pi(0.2.0) = 0.141$$
$$\pi(1.0.1) = 0.140, \quad \pi(0.1.1) = 0.33, \quad \pi(0.0.2) = 0.23$$
$$\lambda(2) = 1.47\ .$$

Questo secondo grafo, invece, rappresenta la specifica che la ripetizione dell'operazione sulla seconda stazione è sempre servita per prima (stato evidenziato), e corrisponde ad un servente con tempo di servizio che ha distribuzione di probabilità risultante da una struttura a stadi in forma diretta.

Come si vede, i tassi di circolazione sono differenti nei tre casi.



$$\pi(2.0.0) = 0.055, \quad \pi(1.1.0) = 0.164, \quad \pi(0.2.0) = 0.39$$
$$\pi(1.0.1) = 0.043, \quad \pi(0.1.1) = 0.34$$
$$\lambda(2) = 1.31\ .$$

5. Analisi del valor medio

Sino a questo punto abbiamo visto risolvere una rete di code attraverso il calcolo di tutte le probabilità di stato, ed è questa una notevole quantità di informazione utilizzata soltanto in minima misura per determinare alcune prestazioni della rete, quali ad esempio il tasso di circolazione medio. Quando soltanto prestazioni medie sono d'interesse, esiste un metodo molto efficiente per calcolarle, che non passa attraverso le probabilità di stato, noto come analisi del valor medio (*Mean Value Analysis MVA).* Questo metodo, su cui si basano tutti gli algoritmi numerici moderni e che si applica sempre a reti di code con la proprietà della forma prodotto, è stato scoperto da *Reiser*. Analizziamolo nei due casi: quando nella rete è presente una sola classe di utenti e quando invece vi sono diverse classi di utenti.

5.1.Singola classe di utenti

Il primo caso considerato è quando nella rete c'è un'unica classe di utenti, in altre parole tutti gli utenti seguono lo stesso percorso.

Il principio alla base di questa tecnica è concettualmente molto semplice e risponde ad una visione soggettiva di una coda. Si immagini di essere uno degli utenti, e ci si domandi qual è il numero medio di altri utenti che, nel muoversi lungo la rete, si incontra innanzi a se ad ogni nodo (che non è il numero medio degli utenti nella risorsa fornito dalla teoria delle code). Se si sa rispondere a questa domanda, è possibile immediatamente conoscere il tempo medio di attesa presso ogni nodo; quindi la somma dei tempi medi di attesa pesata per i fattori di visita fornisce il tempo ciclo medio, ed il suo reciproco moltiplicato per il numero di utenti il tasso di circolazione.

Reiser e *Lavenberg* hanno formalizzato questa domanda e fornito la soluzione.

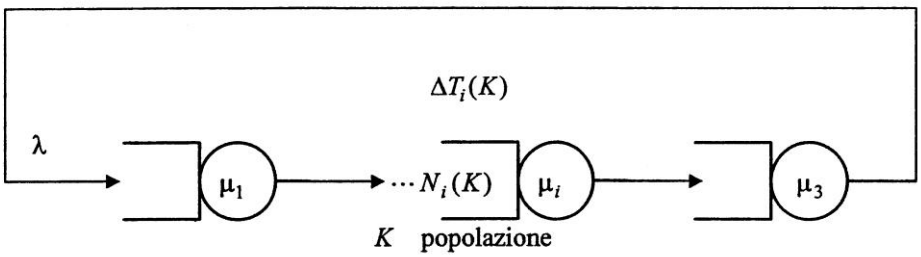


Fig. 5.11 Rete di code chiuse

Teorema 5.2: Teorema di Reiser. Analisi del valor medio di una rete con una sola classe di utenti

In una rete di code chiuse come in figura 5.11, con risorse M/M/1, e popolazione di k utenti, un utente in arrivo ad una coda incontra una lunghezza media della coda, intesa come elementi in attesa più elementi in servizio, pari alla lunghezza media della coda che esisterebbe con una popolazione $k-1$.

Indicato con $N_i(k-1)$ il numero medio di utenti presenti presso il nodo i-esimo di una rete con $k-1$ utenti, il tempo medio di attesa ad un nodo per un utente di popolazione k è pari al tempo necessario a smaltire la coda che lo precede più il tempo necessario per il proprio servizio. In questa trattazione per semplicità di notazione è usato il tempo medio $\tau = {}^{1}/_{\mu}$ anziché il tasso medio di servizio.

$$\Delta T_i(k) = \tau_i \cdot \big(1 + N_i(k-1)\big) \tag{5.7}$$

la frequenza di circolazione media risulta allora:

$$\lambda(k) = \frac{k}{\sum_i \Delta T_i(k)}\,. \tag{5.8}$$

Applicando la legge di Little è possibile quindi conoscere il numero medio di utenti presente in una risorsa di una rete di popolazione k:

$$N_i(k) = \lambda(k) \cdot \Delta T_i(k). \tag{5.9}$$

Dal teorema di Reiser deriva un procedimento iterativo che partendo da $k = 1$, raggiunge la popolazione desiderata $k = K$.

Dal caso elementare di un anello di macchine, si può immediatamente passare al caso più generale di una rete arbitraria determinando, dai coefficienti d'instradamento e dalle specifiche del problema, i fattori di visita.

Detti ν_i i fattori di visita, le formule (5.8) e (5.9) precedenti si modificano nel modo seguente:

$$\lambda(k) = \frac{k}{\sum_i \nu_i \cdot \Delta T_i(k)}$$

$$N_i(k) = \nu_i \cdot \lambda(k) \cdot \Delta T_i(k).$$

Esempio 5.9: Una rete costituita da tre nodi in cascata con due utenti risolta con analisi del valor medio

L'esempio è lo stesso della figura 5.3 che ora è risolto mediante l'analisi del valor medio.

$k = 1$

$$\lambda(1) = \frac{1}{\tau_1 + \tau_2 + \tau_3} \text{ tasso di circolazione;}$$

$$N_1(1) = \frac{\tau_1}{\tau_1 + \tau_2 + \tau_3}$$

$$N_2(1) = \frac{\tau_2}{\tau_1 + \tau_2 + \tau_3} \text{ numero medio utenti in coda;}$$

$$N_3(1) = \frac{\tau_3}{\tau_1 + \tau_2 + \tau_3}$$

$k = 2$

$$\Delta T_1(1) = \tau_1 \cdot \left(1 + \frac{\tau_1}{\tau_1 + \tau_2 + \tau_3}\right)$$

$$\Delta T_2(1) = \tau_2 \cdot \left(1 + \frac{\tau_2}{\tau_1 + \tau_2 + \tau_3}\right) \text{ tempi medi d'attraversamento}$$

$$\Delta T_3(1) = \tau_3 \cdot \left(1 + \frac{\tau_3}{\tau_1 + \tau_2 + \tau_3}\right)$$

e tasso di circolazione:

$$\lambda(2) = \frac{2(\tau_1 + \tau_2 + \tau_3)}{\tau_1(2\tau_1 + \tau_2 + \tau_3) + \tau_2(\tau_1 + 2\tau_2 + \tau_3) + \tau_3(\tau_1 + \tau_2 + 2\tau_3)}.$$

5.2.Stazioni multiserventi

Come abbiamo già visto, reti di code chiuse o aperte Markoviane, almeno in presenza di un'unica classe di utenti, possono comprendere risorse multiserventi mantenendo la proprietà della forma prodotto. Per le reti di code chiuse la tecnica risolutiva dell'analisi del valor medio, poiché non utilizza l'intera informazione sullo stato del sistema, non consente questa estensione. Poiché, però, questa tecnica è numericamente molto più efficiente delle altre, è utile disporre anche per questo algoritmo di una soluzione approssimata del problema multiservente.

Ricordiamo la logica con cui opera l'analisi del valor medio, che si basa sulla valutazione soggettiva da parte di uno specifico utente della coda che deve subire nell'attraversare la rete. Allora, un'idea intuitiva, in presenza di risorse multiservente , è che l'utente che sopraggiunge incontra soltanto una frazione degli utenti effettivamente in attesa, pari al rapporto del loro numero totale per il numero dei serventi. Questa idea risulta ragionevole, ma non completamente corretta.

Indicando, per una popolazione k della rete, con π_i la probabilità che l'utente in arrivo trovi i $(i = 0,1,2,\cdots,k-1)$ altri utenti già nella risorsa, analizziamo il caso di una risorsa biservente; l'estensione al caso generale sarà poi immediata. Quindi compiliamo la seguente tabella che indica, per ogni numero di utenti presenti, il numero di utenti che effettivamente fanno attendere il nuovo arrivato. Il risultato fa uso della proprietà di assenza di memoria della distribuzione esponenziale:

utenti presenti	0	1	2	3	4
probabilità	π_0	π_1	π_2	π_3	π_4
utenti subiti	0	0	1	1	2

Il numero medio di utenti presenti nella risorsa, all'arrivo di un nuovo utente, quale ci fornisce l'analisi del valor medio è per definizione:

$$N = \sum_{i=0}^{K-1} i \cdot \pi_i$$

mentre il numero medio di utenti effettivamente subiti dal nuovo arrivato è invece:

$$\tilde{N} = \pi_2 + \pi_3 + 2 \cdot (\pi_4 + \pi_5) + \cdots.$$

Se, come suggerisce l'intuizione, si fosse assunto come numero di utenti subiti il rapporto del numero medio dei presenti per il numero di serventi (due nel caso del biservente), si sarebbe avuto:

$$\overline{N} = \frac{N}{2} = 0.5 \cdot \pi_1 + \pi_2 + 1.5 \cdot \pi_3 + \cdots$$

Che, confrontato con il valore effettivo, porta a:

$$\widetilde{N} = \overline{N} - 0.5 \cdot \sum_{i=0}^{K-1} \pi_{1+2 \cdot i} \, .$$

Le due grandezze $\widetilde{N}$ e $\overline{N}$ si discostano di una quantità che è sicuramente inferiore a 0.5; per questo si può concludere che, definendo $\underline{N} = \overline{N} - 0.5$, si ha:

$$\underline{N} \leq \widetilde{N} \leq \overline{N} \, .$$

In conclusione, non è possibile determinare il numero medio di utenti cha sarà causa di attesa per un utente in arrivo, ma sarà possibile calcolare un suo limite inferiore e superiore.
La formula del tempo medio d'attraversamento del teorema di *Reiser* diventa:

$$\Delta T_{t_i}(k) = \frac{1}{\mu_i} \cdot \left(1 + \widetilde{N}_i(k-1)\right)$$

dove, in mancanza del valore esatto, $\widetilde{N}$ può essere sostituito da uno dei suoi due limiti, inferiore o superiore, calcolabili nel corso dello sviluppo dell'algoritmo.

Esempio 5.10: Analisi del valor medio di rete con multi servente

L'esempio 5.5 è riproposto usando l'analisi del valor medio.

In particolare trattiamo il multiservente.

I tassi di visita sono:

$$\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \begin{bmatrix}1\\ \frac{1}{1-p}\end{bmatrix} \cdot \lambda\,.$$

Si noti come il vettore dei fattori di visita è soluzione dell'equazione omogenea dei coefficienti d'instradamento normalizzando ad 1 la componente del vettore associato al nodo di riferimento (quello per specifica con fattore di visita uguale ad 1).

$k = 1$

$$\lambda(1) = \frac{1}{\frac{1}{\mu_1} + \frac{1}{(1-p)\mu_2}} = \frac{(1-p)\mu_1\mu_2}{(1-p)\mu_2 + \mu_1}$$

$$N_1(1) = \frac{(1-p)\mu_2}{(1-p)\mu_2 + \mu_1}\,, \qquad N_2(1) = \frac{\mu_1}{(1-p)\mu_2 + \mu_1}$$

$k = 2$

$$\Delta T_1(2) = \frac{1}{\mu_1}\left(1 + \frac{(1-p)\mu_2}{(1-p)\mu_2 + \mu_1}\right) = \frac{2(1-p)\mu_2 + \mu_1}{\mu_1[(1-p)\mu_2 + \mu_1]}\,, \qquad \Delta T_2(2) = \frac{1}{\mu_2}.$$

Si noti che il secondo nodo è un biservente, quindi con una popolazione di due utenti vi sarà sempre un servente libero; $\lambda(2)$ sarà perciò:

$$\lambda(2) = \frac{2}{\frac{2(1-p)\mu_2 + \mu_1}{\mu_1[(1-p)\mu_2 + \mu_1]} + \frac{1}{(1-p)\mu_2}}$$

$$= \frac{2(1-p)\mu_2\mu_1[(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}.$$

$k = 3$

$$\Delta T_1(3) = \frac{1}{\mu_1}\left(1 + \frac{2(1-p)\mu_2[2(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}\right),$$

$$\Delta \overline{T}_2(3) = \frac{1}{\mu_2}\left(1 + \frac{\mu_1[(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}\right)$$

$$\Delta \underline{T}_2(3) = \frac{1}{\mu_2}\left(0.5 + \frac{\mu_1[(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}\right)$$

$$\underline{\lambda}(3) = \frac{3}{\frac{1}{\mu_1}\left(1 + \frac{2(1-p)\mu_2[2(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}\right) + \cdots}$$
$$\cdots + \frac{1}{(1-p)\mu_2}\left(1 + \frac{\mu_1[(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}\right)$$
$$\overline{\lambda}(3) = \frac{3}{\frac{1}{\mu_1}\left(1 + \frac{2(1-p)\mu_2[2(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}\right) + \cdots}$$
$$\cdots + \frac{1}{(1-p)\mu_2}\left(0.5 + \frac{\mu_1[(1-p)\mu_2 + \mu_1]}{[2(1-p)\mu_2 + \mu_1](1-p)\mu_2 + \mu_1[(1-p)\mu_2 + \mu_1]}\right)$$

Assegniamo ai parametri del modello i valori numerici usati nell'esempio 5.4.

I tassi di circolazione trovati per una popolazione di due utenti sono identici nelle due trattazioni, né poteva essere differente poiché con due soli utenti una coda con due serventi fornisce nell'analisi del valor medio un risultato esatto. Questo non è più vero con popolazione superiore a due.

La tecnica dell'analisi del valor medio fornisce risultati approssimati ed in particolare permette di valutare limiti inferiori e superiori. Con una popolazione di tre utenti, troviamo con l'analisi del valor medio i seguenti limiti superiore ed inferiore che approssimano il tasso di ricircolo esatto:

$$\underline{\lambda}(3) = 1.3 < 1.69 < \overline{\lambda}(3) = 1.72\,.$$

Nel caso di classi multiple di utenti la popolazione della rete è definita attraverso un vettore e nella notazione adottata l'indice della classe appare come apice
$\mathbf{K} = (K^1, \cdots, K^R)$ con $\sum_{r=1,R} K^r = K$.

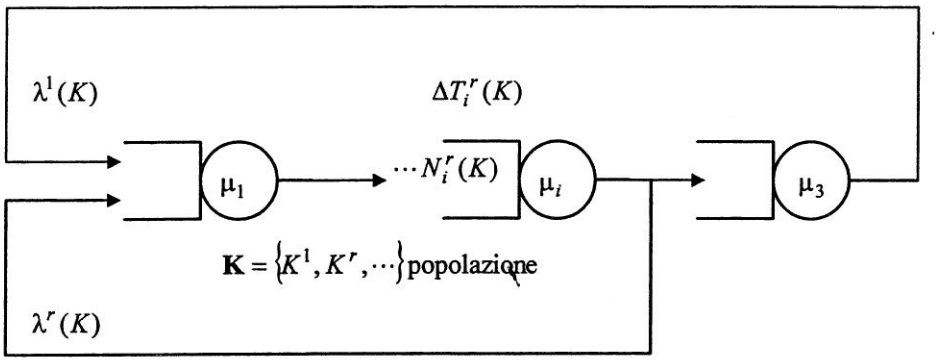


Fig. 5.12 Rete di code chiuse con classi multiple di utenti

Indichiamo con $\boldsymbol{k}$ il vettore del numero dei componenti di ciascuna classe e con $\boldsymbol{e}_r = (0,\cdots,1,\cdots,0)^T$ il versore con un solo 1 in posizione r-esima.
Consideriamo due tipi di code M/M/1 e M/M/∞, con le ipotesi solite che garantiscono la forma prodotto; nei monoserventi tutte le classi di utenti hanno sulla stessa risorsa gli stessi tempi di servizio.
Il teorema di Reiser è così aggiornato.

Teorema 5.3: Teorema di Reiser. Analisi del valor medio di una rete con classi di utenti multiple

Il tempo d'attraversamento della stazione i da parte di un elemento di classe r in una popolazione $\boldsymbol{k}$ risulta dalla lunghezza media della coda della rete che si avrebbe eliminando un elemento di classe r dalla popolazione.

$$\Delta T_i^r(\boldsymbol{k}) = \begin{cases} \tau_i^r \ se\ puro\ ritardo \\ \tau_i \cdot \left[1 + \sum_j N_i^j(\boldsymbol{k} - \boldsymbol{e_r})\right] \end{cases} \tag{5.10}$$

indicato con $Q(r)$ l'insieme delle stazioni visitate dalla classe r, la frequenza di circolazione di quella classe risulta:

$$\lambda^r(\boldsymbol{k}) = \frac{k^r}{\sum_{i\in Q(r)} v_i^r \cdot \Delta T_i^r(\boldsymbol{k})}$$

ed il numero di utenti in coda:

$$n_i^r(\boldsymbol{k}) = v_i^r \cdot \lambda^r(\boldsymbol{k}) \cdot \Delta T_i^r(\boldsymbol{k}).$$

CAPITOLO 6
Controllo di una rete di code

Per poter risolvere in modo analitico una rete di code sono state fatte una serie di ipotesi semplificative, le più critiche delle quali sono che la politica d'instradamento degli utenti è probabilistica e la disciplina di sequenziamento di una coda è FIFO.
Nella pratica queste condizioni non sono rispettate. Piuttosto, in fase operativa sono applicate politiche di controllo in tempo reale ottenute in catena chiusa come una funzione dei valori attuali misurati dello stato del sistema ed atte ad ottimizzare le prestazioni. In generale un sistema controllato perde la proprietà della forma prodotto, e quindi quelle caratteristiche che ne permettono di valutare le prestazioni in modo elementare.
Qui vogliamo analizzare alcuni casi particolari d'interesse pratico che servono a mettere in evidenza gli aspetti caratteristici di questi tipi di controllo, e tuttavia dove, con artifici o approssimazioni, le proprietà della forma prodotto possono ancora essere utilizzate.

1. Bilanciamento del carico di servizi in parallelo

Il caso più interessante è quando la politica d'instradamento dei pezzi alle macchine (*routing*) non è probabilistica ma deterministica e funzione dello stato.

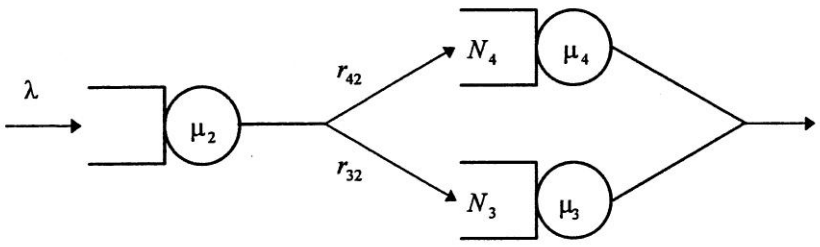


Fig. 6.1 Instradamento intelligente

Definizione 6.1: Politica di bilanciamento del carico

Quando più di una risorsa può eseguire un'operazione (stazioni in parallelo), ogni pezzo uscente dalla stazione precedente è avviato alla stazione

successiva, non in modo casuale come vorrebbe la teoria delle reti di code Markoviane, ma scegliendo la stazione meno carica, intesa come stazione che ha in quell'istante la coda minore. Questa azione prende il nome di politica di bilanciamento del carico.

Risultato:
Un primo risultato è che i carichi delle risorse in parallelo risultano bilanciati, e tutte esprimeranno la medesima lunghezza media delle code. Inoltre, con questa politica d'instradamento l'insieme delle risorse in parallelo si comporta approssimativamente come un'unica risorsa multiservente, che può quindi sostituire l'insieme in un modello modificato.
Un multiservente offre prestazioni sempre migliori di una serie di code in parallelo.
Le prestazioni ottenibili con una politica di bilanciamento, in termini di tempo medio di attesa da parte degli utenti e quindi di numero medio di utenti in attesa, sono certamente migliori di quelle offerte da un insieme di risorse con instradamento casuale ed hanno come limite inferiore quelle di un multiservente equivalente. Questo risultato si deduce dalle seguenti due osservazioni:

- La probabilità che un utente sia avviato ad una coda che non è la coda minima è, come per il multiservente, zero.
- La probabilità di avere un utente in coda ad una risorsa, quando almeno un'altra risorsa è libera, a differenza del multiservente, non è zero. Quindi, il multiservente rappresenta il limite inferiore delle prestazioni raggiungibili da questa politica di bilanciamento.

Esempio 6.1: Stazioni in parallelo e multiserventi in una rete di code aperte
Quest'esempio è utilizzato per mettere in evidenza come una risorsa multiservente è sempre più efficiente di una corrispondente serie di risorse in parallelo quando si adotta una politica d'instradamento casuale.
Confrontiamo le seguenti due reti di code aperte costituite rispettivamente da tre stazioni M/M/1, la prima, e da una stazione M/M/1 ed una M/M/2, la seconda.

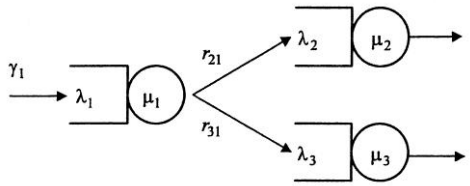


Fig. 6.2 Instradamento intelligente

$\mu_1 = 5, \qquad \mu_2 = \mu_3 = 2.5, \qquad \gamma_1 = 4$

Consideriamo due casi. Nel primo caso i coefficienti di routing sono uguali così come i coefficienti di traffico delle due macchine in parallelo, il numero medio di utenti nella rete ed il tempo medio di attraversamento è dato da:

$$r_{21} = r_{31} = 0.5$$

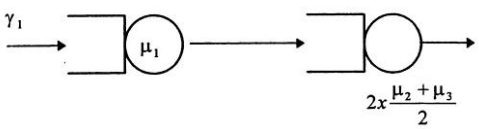
$$N_1 = \frac{\gamma_1}{\mu_1 - \gamma_1} = 4, \qquad N_2 = \frac{r_{21}\gamma_1}{\mu_2 - r_{21}\gamma_1} = 4, \qquad N_3 = \frac{r_{31}\gamma_1}{\mu_3 - r_{21}\gamma_1} = 4,$$

$$N_{tot} = N_1 + N_2 + N_3 = 12, \qquad \Delta T_{tot} = \frac{12}{4} = 3.$$

Nel secondo caso i coefficienti di routing non sono uguali, i coefficienti di traffico delle due macchine in parallelo non sono uguali, il numero medio di utenti nella rete ed il tempo medio di attraversamento è dato da:

$$r_{21} = 0.4, \qquad r_{31} = 0.6,$$
$$N_2 = 1.8, \qquad N_3 = 24,$$
$$N_{tot} = 29.8, \qquad \Delta T_{tot} = 7.45.$$

Supponiamo ora di sostituire alla politica d'instradamento probabilistico una politica intelligente di bilanciamento dei carichi.
Il nuovo modello diventa:



$$N_2 = 4.42,$$
$$N_{tot} = N_1 + N_2 = 8.42 \Delta T_{tot} = 2.1.$$

Come si può vedere, le due risorse isolate hanno globalmente una lunghezza media delle code maggiore anche quando il loro carico è bilanciato, rispetto al caso del biservente. La situazione è ulteriormente peggiorata se il carico delle macchine non è bilanciato come nel secondo caso.

Esempio 6.2: Stazioni in parallelo e multiserventi in una rete di code chiuse

Le stesse due configurazioni di macchine dell'esempio precedente sono utilizzate in una rete di code chiuse, con una popolazione di due utenti.

- Prima configurazione – tre macchine

Tasso di circolazione:

$$\lambda(1) = \frac{1}{\tau_1 + r_{21} \cdot \tau_2 + r_{31} \cdot \tau_3}$$

Numero medio utenti in coda:

$$N_1(1) = \frac{\tau_1}{\tau_1 + r_{21} \cdot \tau_2 + r_{31} \cdot \tau_3}$$
$$N_2(1) = \frac{r_{21} \cdot \tau_2}{\tau_1 + r_{21} \cdot \tau_2 + r_{31} \cdot \tau_3}$$
$$N_3(1) = \frac{r_{31} \cdot \tau_3}{\tau_1 + r_{21} \cdot \tau_2 + r_{31} \cdot \tau_3}$$

Tempi medi d'attraversamento:

$$\Delta T_1(2) = \tau_1 \cdot \left(1 + \frac{\tau_1}{\tau_1 + r_{21} \cdot \tau_2 + r_{21} \cdot \tau_3}\right)$$
$$\Delta T_2(2) = \tau_2 \cdot \left(1 + \frac{r_{21} \cdot \tau_2}{\tau_1 + r_{21} \cdot \tau_2 + r_{21} \cdot \tau_3}\right)$$
$$\Delta T_3(2) = \tau_3 \cdot \left(1 + \frac{r_{31} \cdot \tau_3}{\tau_1 + r_{21} \cdot \tau_2 + r_{21} \cdot \tau_3}\right)$$

Tasso di circolazione:

$$\lambda(2) = \frac{2(\tau_1 + r_{21} \cdot \tau_2 + r_{31} \cdot \tau_3)}{\tau_1(2\tau_1 + r_{21} \cdot \tau_2 + r_{31} \cdot \tau_3) + r_{21} \cdot \tau_2(\tau_1 + 2r_{21} \cdot \tau_2 + \tau_3) + \cdots} \quad (6.1)$$
$$\cdots r_{31} \cdot \tau_3(\tau_1 + \tau_2 + 2r_{31} \cdot \tau_3)$$

- Seconda configurazione – due macchine, un multiservente

Tasso di circolazione:

$$\lambda(1) = \frac{1}{\tau_1 + \tau_2}$$

Numero medio di utenti per coda:

$$N_1(1) = \frac{\tau_1}{\tau_1 + \tau_2}$$
$$N_2(1) = \frac{\tau_2}{\tau_1 + \tau_2}$$

$k = 2$

Tempi medi d'attraversamento:

$$\Delta T_1(1) = \tau_1 \cdot \left(1 + \frac{\tau_1}{\tau_1 + \tau_2}\right)$$

$$\Delta T_2(1) = \tau_2 \ \ perchè\ biservente$$

Tasso di circolazione:

$$\lambda(2) = \frac{2(\tau_1 + \tau_2)}{\tau_1(2\tau_1 + \tau_2) + \tau_2(\tau_1 + \tau_2)} \tag{6.2}$$

Per evidenziare il risultato, consideriamo come al solito le due macchine in parallelo, identiche e con carico bilanciato $r_{21} = r_{31} = 0.5,\ \mu_2 = \mu_3 = 2.5$ e riscriviamo il tasso di circolazione (6.1):

$$\lambda(2) = \frac{2(\tau_1 + \tau_2)}{\tau_1(2\tau_1 + \tau_2) + \tau_2(\tau_1 + 2 \cdot \tau_2)} \tag{6.3}$$

e confrontiamolo con il tasso di circolazione (6.2) in presenza del multiservente. Si vede che in quest'ultimo caso il tasso di circolazione è sempre maggiore del precedente.

Il risultato visto negli esempio precedenti è molto generale. Risponde al principio che mantenere gli utenti uniti in un'unica coda e decidere la loro assegnazione ad un servente soltanto nel momento in cui l'utente è avviato al servizio, permette una maggiore utilizzazione delle risorse, rispetto al caso in cui la decisione è anticipata al momento in cui l'utente raggiunge la coda.

2. Controllo del rapporto dei tassi di circolazione

Nel caso di una rete di code chiuse con classi multiple di utenti, solitamente il rapporto tra le frequenze di circolazione delle diverse classi è un dato di progetto. Indichiamo con λ_{rif} un tasso di riferimento e con $\boldsymbol{m} = (m^1, m^2, \cdots, m^R)^T$ il vettore dei rapporti ("mix") fra i tassi di circolazione di ciascuna classe ed un tasso di riferimento, nel senso che ciascuna classe r ha tasso di circolazione $\lambda^r = m^r \cdot \lambda_{rif}$. Solitamente si sceglie come tasso di riferimento $\lambda_{rif} = \sum_r \lambda^r$ che rappresenta il tasso totale di circolazione della rete.

Il controllo del rapporto fra i tassi di circolazione si potrebbe ottenere adottando su uno o più nodi del sistema, invece della classica disciplina FIFO, una disciplina di coda a priorità, funzione della classe di utente. Questo però, farebbe perdere la proprietà di forma prodotto della rete; ricorriamo perciò al seguente artificio.

Il controllo del rapporto dei tassi di circolazione è ottenuto, anziché con la modifica della disciplina di coda delle risorse, cambiando la struttura logica della rete di code controllata, aggiungendo per ciascuna classe di utenti un nodo fittizio con un puro ritardo di valore medio assegnato (in altre parole introducendo una coda M/M/∞ per ogni classe) e lasciando inalterate le politiche FIFO delle risorse della rete.

Queste stazioni hanno esclusivamente la funzione di introdurre un termine di controllo sul tempo di ingresso degli utenti in coda alla prima stazione della rete. Questo si traduce in una modifica dell'indice di ingresso in coda, rispetto al valore naturale che la rete avrebbe assegnato a ciascun utente, offrendo di fatto un indice di priorità differente per ciascuna classe. Nella realizzazione pratica, il sistema informativo che controlla la rete otterrà lo stesso risultato, modificando artificiosamente con un termine di ritardo (diverso per ciascuna classe di utenti) i tempi effettivi di arrivo in coda di ciascun utente ed utilizzando questi tempi corretti nelle politiche FIFO di sequenziamento delle risorse della rete.

I parametri di progetto del controllo sono rappresentati dal vettore dei tempi medi di servizio per queste stazioni fittizie di controllo $\boldsymbol{\tau} = \{\tau^r\}$, con l'avvertenza di assegnare sempre il valore zero al minore dei tempi di ritardo (non è di alcuna utilità ritardare tutte le classi di utenti insieme).

Poiché l'algoritmo di analisi del valor medio è molto veloce, è possibile impiegarlo in modo iterativo per risolvere il problema di determinare i parametri di progetto del controllo per imporre i rapporti dei tassi di circolazione desiderati.

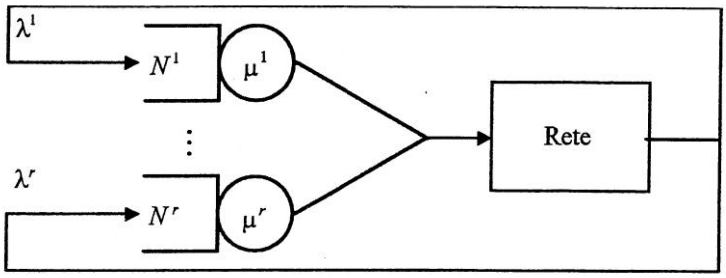


Fig. 6.3 Il controllo dei rapporti dei tassi di circolazione

Proponiamo per questo il seguente algoritmo:

Algoritmo 6.1: Controllo dei rapporti dei tassi di circolazione in una rete di code chiuse

Sia $\boldsymbol{m}_{ass}$ il vettore dei rapporti dei tassi di circolazione, assegnato, e $[\boldsymbol{\tau}]_j, [\boldsymbol{\lambda}]_j$ e $[\boldsymbol{m}]_j$ i vettori dei tempi medi di ritardo delle stazioni di controllo, dei tassi di circolazione e del rapporto fra i tassi di circolazione risultanti all'iterazione j di soluzione dell'algoritmo di analisi del valor medio.

Ad ogni iterazione i nuovi tempi di ritardo sono calcolati dai risultati dell'iterazione precedente nel modo seguente:

$$[\boldsymbol{m}]_j = \frac{1}{\sum_{\mathrm{r}}[\lambda^{\mathrm{r}}]_{\mathrm{j}}}[\boldsymbol{\lambda}]_j$$

$$[\boldsymbol{\tau}]_{j+1} = b \cdot [\boldsymbol{\tau}]_j + k \cdot \left(\boldsymbol{m}_{ass} - [\boldsymbol{m}]_j\right) - \boldsymbol{e} \cdot \tau_{min}$$

dove $0 < b < 1, \quad k < 0$ sono parametri da scegliere per controllare la stabilità e la velocità di convergenza delle iterazioni, $\boldsymbol{e}$ è un vettore di tutti 1 e τ_{min} è la componente minima del vettore $[\boldsymbol{\tau}]_{j+1}$ (quest'ultimo elemento è inserito nell'algoritmo per garantire che tutti i ritardi siano sempre $\tau_i \geq 0$). L'algoritmo esprime buone caratteristiche di convergenza:

$$(1-b) \cdot [\boldsymbol{\tau}]_\infty + \boldsymbol{e} \cdot \tau_{min} = k \cdot (\boldsymbol{m}_{ass} - [\boldsymbol{m}]_\infty)$$

$$\boldsymbol{m}_{ass} - [\boldsymbol{m}]_\infty = \frac{(1-b)}{k} \cdot [\boldsymbol{\tau}]_\infty + \frac{1}{k}\boldsymbol{e} \cdot \tau_{min} \,.$$

Quindi se l'algoritmo converge, lo scarto fra $[\boldsymbol{m}]_\infty$ e $\boldsymbol{m}_{ass}$ potrà essere reso arbitrariamente piccolo facendo crescere il parametro k.

Attraverso i tempi di ritardo delle stazioni di controllo è possibile assegnare i rapporti dei tassi di circolazione della rete, in altre parole intervenire sui valori relativi, non potendo, invece, imporre il tasso totale, cioè i valori assoluti. Questo tasso totale è un indice della produttività della rete e dipende dai tempi di servizio delle stazioni e dalle dimensioni della popolazione di utenti.

Se una configurazione di rete non è in grado di garantire il livello di produttività desiderato, è possibile intervenire sulla dimensione della popolazione. Già sappiamo, infatti, che al crescere della popolazione i tassi di circolazione aumentano verso un valore asintotico. Se questo non è ancora sufficiente, non resta che cambiare la configurazione della rete: aggiungendo risorse in parallelo, oppure sostituendole con altre con tempi di servizio minori.

3. Controllo del numero di utenti

Popolazione in reti di code chiuse

In una rete di code chiuse le dimensioni della popolazione nella rete è, al pari della sua configurazione, un dato di progetto che deve essere impiegato per controllare sia l'utilizzo delle risorse sia i rapporti fra i tassi di circolazione di ciascuna classe.

Infatti la popolazione totale definisce il grado di utilizzo delle risorse e di conseguenza il tasso totale di circolazione nella rete, mentre la distribuzione del numero degli utenti fra le diverse classi influenza i rapporti fra i tassi di circolazione

(è intuitivo che una classe con popolazione più numerosa ottiene un maggior utilizzo della risorse a scapito delle altre e questo si traduce per quella classe in un rapporto dei tassi di circolazione più favorevole).
In questo caso cerchiamo di imporre i rapporti fra i tassi di circolazione delle diverse classi di utenti intervenendo, anziché con un controllo in tempo reale, dimensionando il numero di utenti nella rete. Facciamo uso di una tecnica ricorsiva basata sull'algoritmo di analisi del valor medio. Ancora una volta $\boldsymbol{m}_{ass}$ è il vettore dei rapporti fra i tassi di circolazione, assegnato. Si risolve iterativamente l'algoritmo *MVA* partendo da una popolazione iniziale $\boldsymbol{k} = [1,1,1,\cdots]$ con un utente per ciascuna classe; quindi, ad ogni iterazione si aggiunge alla configurazione dell'iterazione precedente un utente della classe r-esima per cui è massimo lo scarto fra i rapporti assegnati e quelli trovati:

$$r = arg\left(\max_i\left(m^i{}_{ass} - \left[m^i\right]_j\right)\right).$$

Al procedere delle iterazioni la differenza fra i due vettori, misurata come lo scarto massimo fra le loro componenti, si riduce progressivamente.
In corrispondenza al crescere della dimensione della popolazione, vi sarà una crescita del tasso totale di circolazione in maniera monotona verso un valore asintotico. L'algoritmo è arrestato quando questo tasso totale raggiunge il ginocchio del suo andamento asintotico.

Popolazione in reti di code aperte

Sino ad ora abbiamo analizzato primariamente reti di code chiuse. Sono quelle, infatti, più interessanti nelle applicazioni di controllo. Vediamo ora come alcuni problemi, sempre relativi a classi di utenti diversi, in reti di code aperte si possono ricondurre a loro volta a reti di code chiuse.
In una rete di code aperte il numero medio di utenti presenti nella rete per ogni classe dipende dai rispettivi tassi medi di arrivo $\boldsymbol{\gamma} = \{\gamma^r\}$. Nei problemi classici di reti di code, pensiamo alla telefonia, l'arrivo degli utenti è un dato del problema da identificarsi nel comportamento naturale del fenomeno. Esistono situazioni, invece, dove questo tasso di ingresso può essere regolato, o se vogliamo progettato, nel senso di non accettare utenti nella rete oltre ad una certa dimensione.

Pensiamo ad un'officina di lavorazione. Abbiamo visto officine che a motivo del sistema di movimentazione automatico, o a motivo degli attrezzaggi hanno un numero di utenti costante. Per queste, una rete di code chiuse è il modello ideale.
Esistono peraltro officine dove non esiste questo vincolo, e pur tuttavia per evitare un eccessivo ed inutile affollamento si vuole regolare l'accesso alla rete in funzione della classe di utenti. Questo problema può essere formulato come il progetto del vettore dei tassi di ingresso $\boldsymbol{\gamma}$ in modo che i rapporti fra i tassi di circolazione delle diverse classi di utenti siano assegnati ed il numero medio totale di utenti nella rete non superi una soglia.
Un modello appropriato per rappresentare questo problema è ancora una rete di code chiuse.
Si costruisce una rete di code chiuse in cui la rete di code aperte originaria è inserita in un anello con una serie di stazioni di controllo, dove i flussi degli utenti sono fatti ricircolare.
La struttura del controllo è costituita da una stazione di ritardo per ciascuna classe di utenti, per permettere di assegnare i rapporti dei tassi d'attraversamento, e da una stazione comune a tutte le classi che offre un grado di libertà per mantenere indipendente il controllo del numero complessivo di utenti presenti nella parte di rete d'interesse.
Ancora una volta queste stazioni di controllo sono puramente virtuali, con l'unico scopo di fornire un modello che permetta di progettare i tassi di ingresso alla rete reale e rappresenta la politica con cui in tempo reale un nuovo utente di una certa classe sarà immesso nella rete appena nuovo spazio si rende disponibile.
I parametri di progetto sono i ritardi di ciascuna stazione di controllo e le dimensioni della popolazione di ciascuna classe.

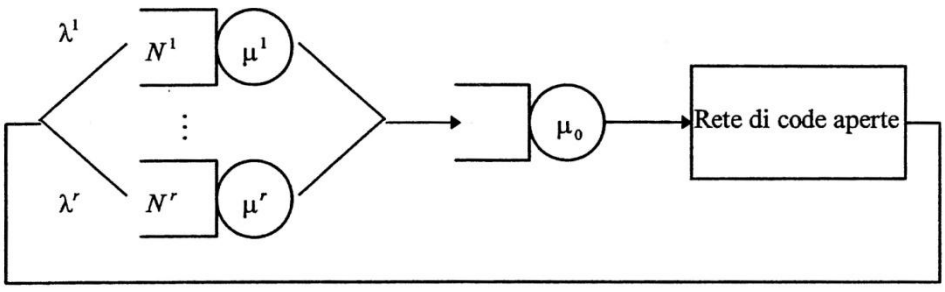


Fig. 6.4 Controllo del numero di utenti e dei rapporti fra i flussi di classi di utenti in una rete di code aperte

Gli elementi controllati sono i rapporti fra i tassi di circolazione, il tasso totale ed il numero medio di utenti nella rete di code aperta.
I parametri di controllo rappresentati dai tempi di ritardo e dalle dimensioni della popolazione servono per garantire i rapporti fra i tassi ed il tasso totale di ingresso alla rete. La stazione comune di controllo offre, invece, un grado di libertà aggiuntivo per gestire la dimensione della popolazione media nella rete d'interesse.

Conclusioni

Data l'ampiezza dell'argomento trattato, nella stesura della tesi è stato fondamentale individuare gli aspetti importanti e suddividere conseguentemente l'intero lavoro in capitoli, il più possibile completi e chiari.
Un ampio preambolo è stato doveroso, riguardante i principali aspetti della teoria delle code, materia tanto diffusa nella vita di tutti i giorni quanto specifica e complessa; a tale aspetto sono stati riservati i primi due capitoli.
Nello specifico, il capitolo uno si è concentrato nel Sistema – Risorsa, con i suoi aspetti principali e le leggi che lo regolano, mentre il capitolo due si è focalizzato sui principali esempi di modelli di code.
Il terzo capitolo ha avuto una trattazione a se stante, in quanto esce dal modello di nascita e morte che caratterizza l'intero lavoro; affrontare anche questi aspetti è risultato fondamentale in quanto aiutano a comprendere alcune particolarità comportamentali delle reti di code.
Al corpo centrale della tesi sono stati dedicati il capitolo quattro e cinque. Le reti di code si suddividono, come abbiamo visto, in reti di code aperte, trattate nel quarto capitolo insieme alle reti con ricircolo, altro tipo fondamentale di rete di code, e reti di code chiuse, alle quali è stato riservato il quinto capitolo, con una trattazione autonoma a causa delle loro caratteristiche, che portano ad efficienti algoritmi numerici di soluzione, detti di analisi del valor medio, a loro volta oggetto di approfondimento.
La trattazione è stata conclusa con l'ultimo capitolo, il sesto, riguardante i principali aspetti del controllo in una rete di code.
L'aspetto più interessante di questo approfondimento si è rivelato essere l'affrontare un aspetto banale nel suo genere, che ogni persona comune può riscontrare durante una semplice attività quotidiana come quella di essere in coda ad uno sportello, però sotto l'aspetto matematico-scientifico-informatico, che probabilmente è il principio alla base della ingegnerizzazione nello studio dei fenomeni. L'ampiezza degli argomenti ha concesso di spaziare nella trattazione, anche tramite l'uso di esempi esplicativi.
La polivalenza dei principi base, poi, ha permesso di pensare ai fenomeni ed alle materie più svariati per 'entrare nella mentalità' dell'argomento, dall'informatica con l'esempio dell'unità centrale con i programmi da elaborare, alle telecomunicazioni con la linea telefonica ed i suoi utilizzatori, fino all'ingegneria gestionale con la gestione dei più svariati principi aziendali, dal flusso delle risorse alla gestione degli ordini.

Bibliografia

Fonte principale per la redazione della tesi è stato il libro di testo:

- Carlucci D., Menga G., 1998. *Teoria dei sistemi ad eventi discreti*. Prima edizione. Torino: UTET Libreria S.r.l., pp. 227-305.

Ulteriori riferimenti:

- Kleinrock L., 1975, *Queueing* Systems, John Weley & Sons, Inc., New York, pp. 87-147.

- Kobayashi M., 1979, *Modeling and Analysis: an Introduction to System Performance Evaluation Methodology,* Addison-Wesley, Reading, Mass.

- Little J. D. C., 1961, *A Proof of* $L = \lambda W$*,* "Operations Research", Vol. 9, pp. 383-387.

- Stidham S. Jr., 1974, *A Last Word on* $L = \lambda W$*,* "Operations Research", Vol. 22, pp.417-421.

- Burke P. J., 1956, *The Output of a Queueing System,* "Operations Research", Vol. 4, pp. 699-704.

- Jackson J. R., 1963, *Jobshop-like Queueing Systems,* "Management Science", Vol. 10, pp. 131-142.

- Gordon W. J., Newell G. F., 1967, *Closed Queueing Systems,* "Operations Research", Vol. 15, pp. 254-265.

- Kleinrock L., 1975, *Queueing Systems,* John Wiley & Sons, Inc., New York, pp.147-164.

- Baskett F., Chandy K. M., Muntz R. R., Palacios F. G., 1975, *Open, Closed and Mixed Networks with Different Classes of Customers,* "Journal of ACM, Vol. 22, pp. 248-260.

- Reiser M., 1980, Lavenberg S. S., *Mean Value Analysis of Closed Multichain Queueing Networks,* "Journal of ACM", Vol. 27, pp. 313-322.

- Menga G., Bruno G., Conterno R., Actis Dato M., 1984, *Modeling FMS by Closed Queueing Network Analysis Methods*, "IEEE Transactions on CHMT, Vol. 7, No. 3.

- Conterno R., 1985, Produttività e Pianificazione nella fabbrica Automatica, Tesi di dottorato, Dip. Automatica e Informatica Politecnico di Torino.

- Walrand J., 1988, *An Introduction to Queueing Systems,* Prentice Hall, Englewood Cliffs, NJ.

- G. Giambene, "Queuing Theory and Telecommunications: Networks and Applications", Springer, May 2005.

Fonti internet di supporto:

- WIKIPEDIA, l'enciclopedia libera. *Teoria delle code* [online]. Disponibile su <http://it.wikipedia.org/wiki/Teoria_delle_code>[Data di accesso 16/07/2011].

- RAFFAELE PRESENTI. *Teoria delle code o file d'attesa* [online]. Disponibile su <ftp://docenti.ing.units.it/arc_stud/Pesenti/Nettuno/CodeDispense.pdf> [Data di accesso 18/07/2011].

- M. STRANO. *Teoria delle code* [online]. Disponibile su <*http*://webuser.unicas.it/dweb/gestione/download.php?id=1433>[Data di accesso 31/07/2011].

- *Teoria della probabilità e Teoria delle code* [online]. Disponibile su <http://www.dia.uniroma3.it/~adacher/automazione1/TeoriaCode.pdf> [Data di accesso 31/07/2011].

- SIMONA SACONE. *Teoria delle code e delle reti di code* [online]. Disponibile su <http://www.dist.unige.it/simona/CorsoAI1/cap7.pdf> [Data di accesso 30/07/2011].

- INGEGNERIA DEL TELETRAFFICO, 2006. *Reti di code* [online]. Disponibile su <wwwtlc.iet.unipi.it/teaching/teletraffico/2005-06/docs/lez14-05.pdf> [Data di accesso 20/08/2011].

- NELLO SCARABOTTOLO, 2008. *Reti di code per l'analisi dei sistemi di calcolo* [online]. Disponibile su <www.dti.unimi.it/scarabottolo/reti/Reti%20di%20code.pdf> [Data di accesso 20/08/2011].

- S. BALSAMO. *Modelli a rete di code* [online]. Disponibile su <www.dsi.unive.it/~balsamo/disp.pdf/Cap4.pdf> [Data di accesso 20/07/2011].