Corso di
SIMULAZIONE

Prof.ssa Amelia Giuseppina Nobile[1]

(a.a. 2011/2012)

5 dicembre 2011

[1]Facoltà di Scienze Matematiche Fisiche e Naturali, Università di Salerno

Capitolo 1

Sistemi di servizio

1.1 Introduzione

Un'area di grande interesse dell'informatica, dell'ingegneria e della matematica applicata è la *teoria delle file di attesa*, detta anche *teoria delle code*. La teoria delle file di attesa si propone di formulare e analizzare modelli matematici e di simulazione atti a descrivere sistemi reali in cui il generico utente richiede un particolare servizio e deve attendere in qualche tipo di coda (o fila di attesa) se il servitore non è immediatamente disponibile. Tipici esempi in cui si presentano file di attesa sono le chiamate ad un centralino telefonico, gli utenti in banca, alla posta o in un ospedale, i clienti in mense, in supermarket o in ristoranti, le persone in attesa di un taxi, le automobili ad un incrocio, gli aerei in attesa di decollare o di atterrare in un aeroporto, i pezzi in attesa di essere lavorati, le macchine in avaria in un'officina, ...

I risultati forniti dalla teoria delle file di attesa trovano applicazione in numerosi campi: sistemi di elaborazione, sistemi di comunicazione e di trasmissione dati, sistemi di trasporto, sistemi di produzione industriale, sistemi per la gestione di servizi pubblici e privati, ...

La teoria delle file di attesa è essenzialmente di natura probabilistica e fornisce una descrizione dei cambiamenti di stato nella lunghezza delle code, del tempo di permanenza di un utente nella fila di attesa, del tempo di attesa di un utente nel sistema, del periodo di occupazione e del periodo di ozio di un centro di servizio, ...

L'analisi effettuata mediante la teoria delle file di attesa si propone di stabilire la tipologia dei modelli più adatti a descrivere sistemi di servizio reali fornendo idonee misure di prestazione e individuando gli opportuni cambiamenti da apportare a tali sistemi per migliorare, se necessario, le loro prestazioni.

Un generico sistema di servizio può essere schematizzato come illustrato nella Figura 1.1 nella quale sono rappresentati:

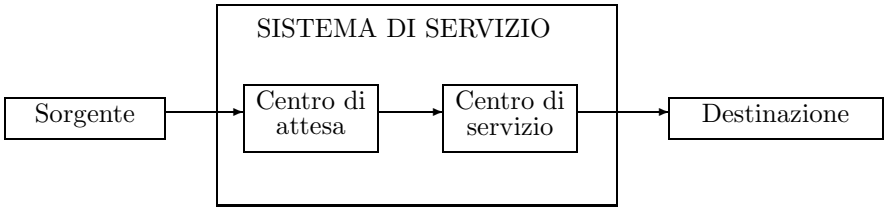

Figura 1.1: Rappresentazione di un sistema di servizio.

- *la sorgente*, ossia l'insieme delle richieste di servizio che si possono presentare al sistema di servizio;
- *il centro di attesa*, ossia l'insieme delle richieste di servizio che, non potendo essere immediatamente soddisfatte, restano in attesa di poter essere prese in considerazione;
- *il centro di servizio*, ossia l'insieme dei punti nei quali viene soddisfatta la richiesta;
- *la destinazione*, ossia l'insieme delle richieste di servizio che, essendo state soddisfatte, lasciano i punti di servizio.

Sorgente
La sorgente (o popolazione) contiene i potenziali utenti del sistema di servizio, ossia l'insieme da cui arrivano gli utenti. Essa può essere finita o infinita. Un sistema di servizio la cui sorgente è infinita è più facile da descrivere matematicamente di un sistema con sorgente finita. Ciò è dovuto alla circostanza che nel caso di sorgente finita il numero degli utenti nel sistema influenza i parametri di arrivo; infatti, se tutti i potenziali utenti sono già arrivati nel sistema i parametri di arrivo sono nulli. Spesso se la sorgente è finita e contiene numerosi potenziali utenti, si assume che sia infinita per rendere più semplice la trattazione matematica. Gli utenti possono anche provenire da diverse distinte sorgenti. Gli utenti che provengono da una stessa sorgente sono tra loro indistinguibili. Si suppone invece che esistano diverse sorgenti quando si desidera distinguere gli utenti per qualche ragione, ad esempio a causa di differenti livelli di priorità oppure a causa di differenti provenienze geografiche.

Gli utenti provenienti da una sorgente si inseriscono in un sistema di servizio per ricevere un determinato servizio. Il termine *utente* è inteso in senso generico: può essere un messaggio che deve essere trasmesso, una richiesta di servizio I/O, un programma che richiede servizi di CPU in un sistema multiprogrammato, ...

Centro di attesa
L'accesso ad un sistema di servizio può essere realizzato attraverso un *centro di attesa* (buffer) che può avere la possibilità di contenere un numero limitato o illimitato di utenti. La capacità del centro di attesa può essere quindi finita o infinita. Se il sistema possiede un centro di attesa a capacità limitata, il numero

degli utenti in attesa non può superare un certo limite caratteristico del sistema di servizio e pertanto una richiesta di servizio che si presenta quando il centro di attesa è saturo viene respinta. Un esempio di centro di attesa con capacità limitata è un centralino telefonico che può avere in attesa soltanto un numero finito di chiamate. Esistono sistemi, noti in letteratura come *loss systems*, che hanno un centro di attesa a capacità nulla; in essi se un utente arriva quando tutti i servitori sono occupati, la sua richiesta di servizio è respinta. Un esempio di centro di attesa con capacità nulla è un centralino telefonico in cui una chiamata in arrivo è accettata immediatamente oppure è rifiutata. Se, invece, il sistema possiede un centro di attesa a capacità illimitata nessuna richiesta di servizio viene perduta per quanto lunga possa essere la durata dell'attesa (a meno che gli individui in attesa decidano di allontanarsi spontaneamente dal sistema).

È evidente che non sempre le richieste di servizio entrano in attesa per poter essere soddisfatte; il fenomeno dell'attesa si presenta soltanto quando il sistema di servizio non ha risorse immediatamente disponibili per soddisfare le richieste.

Centro di servizio

Superato il centro di attesa gli utenti accedono al centro di servizio che può consistere di uno o più servitori. Il *servitore* è un'entità in grado di eseguire il servizio richiesto dall'utente. Ovviamente un sistema con più servitori può fornire simultaneamente servizio a più utenti. I servitori hanno caratteristiche identiche, lavorano in parallelo e non possono rimanere inattivi in presenza di utenti nella fila di attesa. Se tutti i servitori nel centro di servizio sono occupati l'utente, quando si inserisce nel sistema, deve mettersi in fila di attesa finché non si liberi uno dei servitori.

In un sistema di servizio si suppone che esista un unico centro di attesa anche in presenza di uno o più servitori che lavorano in parallelo. Quando ogni singolo servitore è dotato di un proprio centro di attesa (buffer) si preferisce parlare di una *rete di code* piuttosto che di un unico sistema di servizio.

Destinazione

Ogni utente lascia istantaneamente il sistema di servizio dopo aver completato il suo servizio. L'insieme delle richieste di servizio espletate sono instradate verso la destinazione.

Capacità del sistema

Con il termine *capacità del sistema* si intende il numero massimo di utenti (inclusi quelli in servizio) che possono essere contenuti nel sistema.

Disciplina di servizio

Il complesso di regole secondo le quali gli utenti in attesa passano dal centro di attesa al centro di servizio è detto *disciplina di servizio*. Essa specifica quale sarà il prossimo utente tra quelli in attesa che accede al centro di servizio non appena si libera uno dei servitori. La disciplina di servizio più comune è la disciplina *FIFO* (first-in, first-out) secondo la quale il primo arrivato è il primo ad essere servito. Esiste anche la disciplina di servizio *LIFO* (last-in, first-out) secondo la quale l'ultimo arrivato è il primo ad essere servito. Un'altra importante disciplina è la *SIRO* (service in random order) con la quale ogni utente nel centro di attesa ha la stessa probabilità di essere selezionato per il

servizio. La disciplina *PRI* (priority service) invece prevede che alcuni utenti abbiano un trattamento privilegiato; gli utenti sono in tal caso suddivisi in classi di priorità ed il sistema di coda attua una politica preferenziale nei riguardi di alcune classi di utenti.

In un sistema di servizio gli utenti ritengono fondamentale la riduzione dei tempi di attesa, mentre il gestore del sistema è solitamente interessato al massimo sfruttamento delle risorse (servitori) pur cercando di rispettare le esigenze degli utenti. Il progettista di un sistema di servizio deve quindi essere in grado in base alla struttura del sistema, alla frequenza di arrivo degli utenti, al numero di servitori e alla loro velocità di servizio di analizzare le prestazioni del sistema e di apportare, se necessario, opportuni cambiamenti alla struttura stessa.

Di fondamentale importanza per la descrizione di un sistema di servizio sono i meccanismi degli arrivi e delle partenze.

1.2 Meccanismo degli arrivi

Per descrivere il meccanismo degli arrivi occorre conoscere la funzione di distribuzione delle variabili aleatorie $T_1, T_2, \ldots$, descriventi i *tempi di interarrivo*. Il generico T_i rappresenta la lunghezza dell'intervallo di tempo che intercorre tra l'arrivo $(i-1)$-esimo e l'arrivo i-esimo.

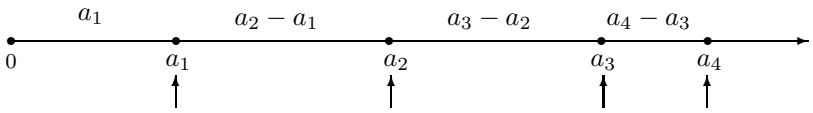

Figura 1.2: Una rappresentazione degli intervalli di interarrivo.

Ad esempio, in Figura 1.2 gli istanti di arrivo degli utenti sono denotati con $a_1, a_2, a_3, \ldots$ e gli intervalli di interarrivo hanno rispettive lunghezza a_1, $a_2 - a_1$, $a_3 - a_2$, ...

Di particolare importanza sono alcune caratteristiche numeriche di tali variabili aleatorie, quali i valori medi $E(T_i)$ e le varianze $\mathrm{Var}(T_i)$ dei tempi di interarrivo T_i $(i = 1, 2, \ldots)$.

Spesso si suppone che $T_1, T_2, \ldots$ sia una successione di variabili aleatorie indipendenti e identicamente distribuite (*iid*). In tal caso, se si denota con T una generica di tali variabili aleatorie, occorre specificare la sua funzione di distribuzione e la sua densità di probabilità.

In letteratura la funzione di distribuzione di T è solitamente denotata con $A(t) = P(T < t)$ e la sua densità di probabilità con $a(t)$. Alcune delle notazioni più frequentemente utilizzate per i tempi di interarrivo sono le seguenti:

D - tempi di interarrivo *iid* con funzione di distribuzione deterministica,

U - tempi di interarrivo *iid* con funzione di distribuzione uniforme,

M - tempi di interarrivo *iid* con funzione di distribuzione esponenziale,

E_k - tempi di interarrivo *iid* con funzione di distribuzione di Erlang di ordine k,

H_k - tempi di interarrivo *iid* con funzione di distribuzione iperesponenziale di ordine k,

GI - tempi di interarrivo *iid* con funzione di distribuzione generale.

Analizziamo ora più in dettaglio i vari meccanismi degli arrivi.

1.2.1 Meccanismo degli arrivi di tipo D

Il meccanismo degli arrivi più semplice che si possa immaginare è quello regolare (deterministico); esso è caratterizzato da una cadenza temporale costante degli arrivi. Supponiamo che un generico intervallo di interarrivo sia di lunghezza fissa $1/\lambda$. La lunghezza di tale intervallo può essere quindi descritta da una variabile aleatoria T degenere la cui funzione di distribuzione è

$$A(t) = P(T < t) = \begin{cases} 0, & t \leq 1/\lambda \\ 1, & t > 1/\lambda\,. \end{cases} \tag{1.1}$$

Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = \frac{1}{\lambda}\,, \qquad \mathrm{Var}(T) = 0. \tag{1.2}$$

Meccanismi degli arrivi deterministici si possono presentare quando si considerano sistemi di servizio in cascata che prevedono due o più posti di lavoro nei quali l'uscita di un posto di lavoro costituisce l'ingresso per il successivo. Se il tempo di servizio del precedente posto di lavoro è costante, allora la distribuzione degli intervalli di interarrivo per il successivo posto di lavoro è quella deterministica. Occorre sottolineare che, eccetto nel caso di una catena di montaggio, nella realtà raramente si incontrano meccanismi degli arrivi regolari.

1.2.2 Meccanismo degli arrivi di tipo U

Nel meccanismo degli arrivi di tipo U i tempi di interarrivo sono indipendenti e identicamente distribuiti con funzione di distribuzione uniforme. Sia T una variabile aleatoria uniformemente distribuita nell'intervallo (a, b), descrivente la lunghezza di un generico tempo di interarrivo uniforme. La sua funzione di distribuzione è

$$A(t) = P(T < t) = \begin{cases} 0, & t \leq a \\ \dfrac{t-a}{b-a}\,, & a < t \leq b \\ 1, & t > b \end{cases} \tag{1.3}$$

e quindi la densità di probabilità è:

$$a(t) = \frac{dA(t)}{dt} = \begin{cases} \dfrac{1}{b-a}\,, & a < t < b \\ 0, & \text{altrimenti.} \end{cases} \tag{1.4}$$

In Figura 1.3 è rappresentata la funzione di distribuzione (1.3) e la densità di probabilità (1.4) della variabile aleatoria T uniformemente distribuita nell'intervallo (a, b).

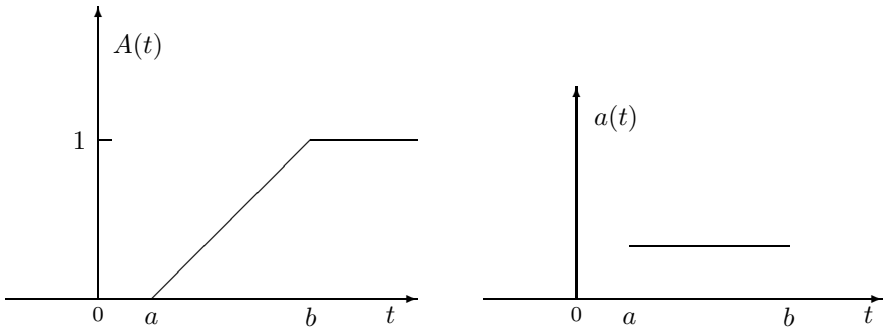

Figura 1.3: Funzione di distribuzione e densità di probabilità della variabile aleatoria T uniformemente distribuita nell'intervallo (a, b).

Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = \frac{a+b}{2}\,, \qquad \mathrm{Var}(T) = \frac{(b-a)^2}{12}\,\cdot \tag{1.5}$$

Se si desidera che il tempo medio di interarrivo sia $E(T) = 1/\lambda$ (come nel caso deterministico), basta scegliere $a = 0$ e $b = 2/\lambda$. La variabile aleatoria T è allora uniformemente distribuita nell'intervallo $(0, 2/\lambda)$ con densità di probabilità:

$$a(t) = \begin{cases} \dfrac{\lambda}{2}\,, & 0 < t < \dfrac{2}{\lambda} \\ 0, & \text{altrimenti.} \end{cases} \tag{1.6}$$

1.2.3 Meccanismo degli arrivi di tipo M

Nel meccanismo degli arrivi di tipo M i tempi di interarrivo sono indipendenti e identicamente distribuiti con funzione di distribuzione esponenziale. La lettera M significa Markov ed indica che il processo degli arrivi è casuale. La funzione di distribuzione esponenziale è frequentemente utilizzata nella teoria delle file di attesa per le importanti proprietà di cui essa gode.

Sia T una variabile aleatoria esponenzialmente distribuita con valore medio $1/\lambda$, descrivente la lunghezza di un generico tempo di interarrivo esponenziale. La sua funzione di distribuzione è:

$$A(t) = P(T < t) = \begin{cases} 0, & t \le 0 \\ 1 - e^{-\lambda\, t}, & t > 0 \end{cases} \tag{1.7}$$

e quindi la sua densità di probabilità è:

$$a(t) = \frac{dA(t)}{dt} = \begin{cases} \lambda\, e^{-\lambda\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.8}$$

In Figura 1.4 è rappresentata la funzione di distribuzione (1.7) e la densità di probabilità (1.8) della variabile aleatoria T esponenzialmente distribuita con valore medio $1/\lambda$.

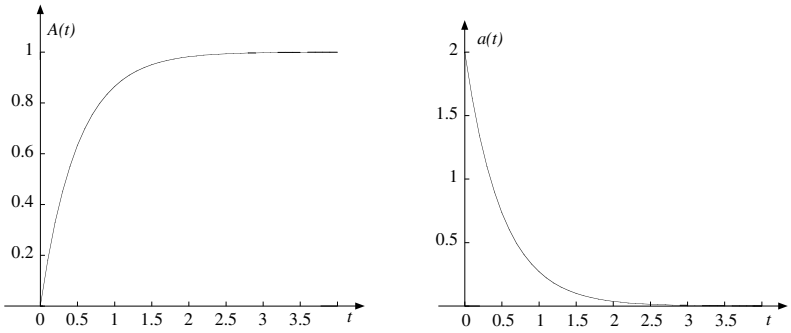

Figura 1.4: Funzione di distribuzione e densità di probabilità della variabile aleatoria T esponenzialmente distribuita con valore medio $1/\lambda$, con $\lambda = 2$.

La densità esponenziale è una funzione strettamente decrescente e quindi i valori più piccoli sono i più probabili. Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = \frac{1}{\lambda}\,, \qquad \mathrm{Var}(T) = \frac{1}{\lambda^2}\cdot \tag{1.9}$$

Il parametro λ è l'inverso del tempo medio di interarrivo, ossia del tempo medio che intercorre tra l'arrivo di due utenti successivi, e può essere interpretato come la *frequenza media di arrivo degli utenti per unità di tempo.*

La funzione di distribuzione esponenziale riveste notevole importanza sia teorica che applicativa. Interviene spesso quando si considerano sistemi di servizio in cui si suppone che i tempi di interarrivo degli utenti (oppure i tempi di servizio) siano distribuiti esponenzialmente con un certo parametro. Interviene anche quando si considera la durata di vita, ovviamente aleatoria, di un componente elettronico o di una macchina.

Se si denota con T la lunghezza di un generico intervallo di interarrivo, la variabile aleatoria esponenziale gode (come accade per la distribuzione geometrica nel caso discreto) dell'importante proprietà

$$P(T > t + s \mid T > s) = P(T > t) \qquad (t > 0, s > 0), \tag{1.10}$$

che esprime il fatto che la probabilità condizionata che il tempo di interarrivo sia maggiore di $t+s$ dato che tale tempo è maggiore di s non dipende da quanto si è già atteso, ossia da s. Questa circostanza esprime la *mancanza di memoria* della funzione di distribuzione esponenziale. La validità della (1.10) può essere

facilmente dimostrata. Infatti, risulta:

$$P(T > t+s \mid T > s) = \frac{P(T > t+s,\ T > s)}{P(T > s)} = \frac{P(T > t+s)}{P(T > s)}$$
$$= \frac{e^{-\lambda\,(t+s)}}{e^{-\lambda\,s}} = e^{-\lambda\,t} = P(T > t).$$

Per la mancanza di memoria della funzione di distribuzione esponenziale, si ha inoltre che il *tempo di interarrivo residuo* ha la stessa distribuzione del tempo di interarrivo. Infatti, se denotiamo con T un generico tempo di interarrivo e con Z una variabile aleatoria che descrive il tempo di interarrivo residuo, ossia $Z = T - \tau$, se $t > 0$ si ha

$$P(Z \le t \mid Z > 0) = 1 - P(Z > t \mid Z > 0) = 1 - P(T - \tau > t \mid T > \tau)$$
$$= 1 - P(T > t + \tau \mid T > \tau) = 1 - e^{-\lambda\,t},$$

ossia Z è distribuita esponenzialmente con valore medio $1/\lambda$.

1.2.4 Meccanismo degli arrivi di tipo E_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.5, in cui l'ingresso al sistema è unico ed esiste un distributore che assegna ordinatamente a ciascuno delle k file di attesa gli arrivi. Alla prima fila di attesa sono così assegnati il primo arrivo, il $(k{+}1)$–esimo arrivo, il $(2k{+}1)$–esimo arrivo ed in generale il $(i\,k+1)$–esimo arrivo $(i = 0, 1, \ldots)$; alla generica j–esima $(j = 1, 2, \ldots, k)$ fila di attesa viene assegnato il j–esimo arrivo, il $(k+j)$–esimo arrivo ed in generale il $(i\,k+j)$–esimo arrivo $(i = 0, 1, \ldots)$.

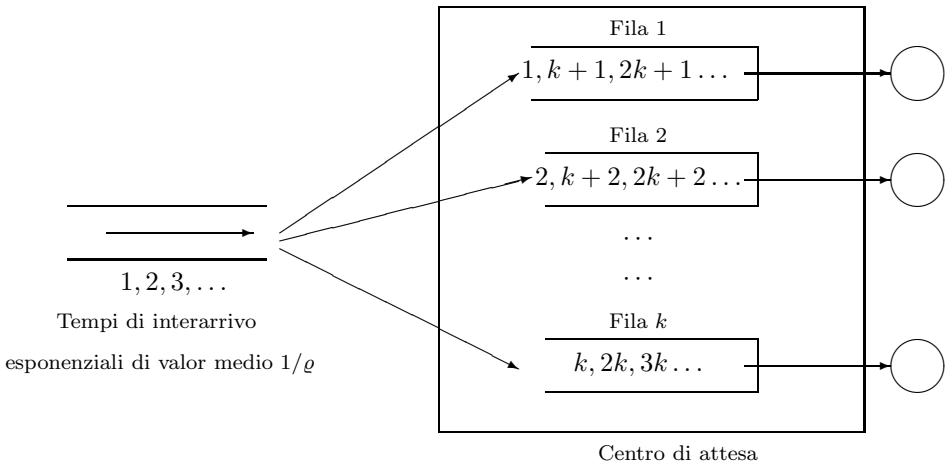

Figura 1.5: Tempi di interarrivo di tipo E_k in ognuna delle k file di attesa.

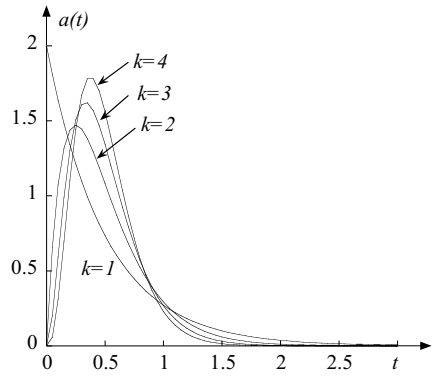

Figura 1.6: Densità (1.13) di Erlang di ordine k con $\lambda = 2$ e per $k = 1, 2, 3, 4$.

Supponiamo che i tempi di interarrivo degli utenti che accedono al sistema siano indipendenti ed esponenzialmente distribuiti con valore medio $1/\varrho$. Denotiamo con T la lunghezza dell'intervallo di tempo che intercorre tra due arrivi in una generica delle k file di attesa. Poiché tra un arrivo ed il successivo in una delle k file di attesa intercorrono k intervalli di interarrivo esponenziali, la variabile aleatoria T può essere vista come la somma di k variabili aleatorie $T_1, T_2, \ldots, T_k$ indipendenti, ognuna distribuita esponenzialmente con valore medio $1/\varrho$. La somma di k variabili aleatorie indipendenti di tipo esponenziale con valore medio $1/\varrho$, è distribuita con densità di probabilità di Erlang di ordine k, ossia

$$a(t) = \begin{cases} \dfrac{\varrho^k}{(k-1)!}\, e^{-\varrho t}\, t^{k-1}, & t > 0 \\ 0, & t \leq 0. \end{cases} \tag{1.11}$$

Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = E(T_1 + T_2 + \ldots + T_k) = E(T_1) + E(T_2) + \ldots + E(T_k) = \frac{k}{\varrho}\,, \tag{1.12}$$

$$\mathrm{Var}(T) = \mathrm{Var}(T_1 + T_2 + \ldots + T_k) = \mathrm{Var}(T_1) + \mathrm{Var}(T_2) + \ldots + \mathrm{Var}(T_k) = \frac{k}{\varrho^2}\cdot$$

Se si desidera che il tempo medio di interarrivo sia $E(T) = 1/\lambda$ (come nei casi deterministico e esponenziale), basta scegliere $\varrho = k\,\lambda$. La variabile aleatoria T è allora caratterizzata dalla seguente densità di probabilità di Erlang di ordine k:

$$a(t) = \begin{cases} \dfrac{(k\,\lambda)^k}{(k-1)!}\, e^{-k\,\lambda t}\, t^{k-1}, & t > 0 \\ 0, & t \leq 0. \end{cases} \tag{1.13}$$

e quindi il valore medio e la varianza del tempo di interarrivo sono $E(T) = 1/\lambda$ e $\mathrm{Var}(T) = 1/(k\,\lambda^2)$. Se si pone $k = 1$ nella (1.13) si ottiene la densità esponenziale (1.8). In Figura 1.6 è rappresentata la densità di probabilità (1.13) della variabile aleatoria T con densità di Erlang di ordine k con $\lambda = 2$ e per $k = 1, 2, 3, 4$.

1.2.5 Meccanismo degli arrivi di tipo H_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.7, che ha il compito di servire k differenti sorgenti. Ricordiamo che i potenziali utenti sono suddivisi in k diverse sorgenti a causa di differenti livelli di priorità loro assegnati oppure a causa di loro diverse provenienze geografiche. Supponiamo che i tempi

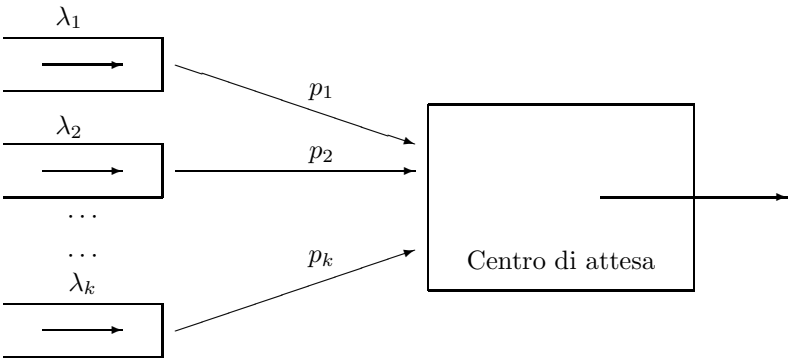

Figura 1.7: Tempi di interarrivo di tipo H_k nella fila di attesa.

di interarrivo degli utenti che accedono alla sorgente j–esima siano descritti da variabili aleatorie indipendenti e distribuite esponenzialmente con parametro λ_j $(j = 1, 2, \ldots, k)$. Il centro di attesa è provvisto di un ingresso unico che provvede a scegliere con probabilità p_j la sorgente j–esima $(j = 1, 2, \ldots, k)$ ed ad avviare al centro di attesa la prima delle richieste di servizio relative alla sorgente selezionata. Assumiamo che

$$p_j \geq 0 \qquad (j = 1, 2, \ldots, k), \qquad \sum_{j=1}^{k} p_j = 1. \tag{1.14}$$

La scelta delle probabilità $p_1, p_2, \ldots, p_k$ dipende dalla priorità assegnata agli utenti delle varie sorgenti oppure dal numero di potenziali utenti provenienti da diverse località geografiche che accedono al centro di servizio.

Denotiamo con T la variabile aleatoria che descrive la lunghezza dell'intervallo di tempo tra due arrivi successivi al centro di attesa del sistema. Inoltre, denotiamo T_j la variabile aleatoria che descrive la lunghezza dell'intervallo di

interarrivo degli utenti nella sorgente j–esima e con A_j l'evento "*è stata scelta la sorgente j–esima*" ($j = 1, 2, \ldots, k$). Si nota immediatamente che l'evento $\{T < t\}$ può essere così rappresentato:

$$\{T < t\} = \bigcup_{j=1}^{k} \Big[A_j \cap \{T < t\} \Big]. \tag{1.15}$$

Infatti, l'evento $\{T < t\}$ si realizza se si verifica uno qualunque dei k eventi incompatibili $A_1, A_2, \ldots, A_k$ ed inoltre $T < t$. Pertanto se $t > 0$, la probabilità della realizzazione dell'evento $\{T < t\}$ è data dalla somma delle probabilità p_j (associata all'evento A_j) per la probabilità dell'evento $\{T_j < t\}$, ossia

$$\begin{aligned} A(t) = P(T < t) &= \sum_{j=1}^{k} P\Big[A_j \cap \{T < t\} \Big] = \sum_{j=1}^{k} P(A_j)\, P(T < t \mid A_j) \\ &= \sum_{j=1}^{k} p_j\, P(T_j < t) = \sum_{j=1}^{k} p_j \left(1 - e^{-\lambda_j\, t} \right). \end{aligned}$$

Ricordando la (1.14) segue immediatamente che la funzione di distribuzione del tempo di interarrivo T è

$$A(t) = \begin{cases} 0, & t \le 0 \\ 1 - \displaystyle\sum_{j=1}^{k} p_j\, e^{-\lambda_j\, t}, & t > 0 \end{cases} \tag{1.16}$$

e pertanto la sua densità di probabilità è

$$a(t) = \begin{cases} \displaystyle\sum_{j=1}^{k} p_j\, \lambda_j\, e^{-\lambda_j\, t}, & t > 0 \\ 0, & \text{altrimenti,} \end{cases} \tag{1.17}$$

ossia una densità iperesponenziale di ordine k relativa ai tempi di interarrivo. Ponendo $k = 1$, oppure $\lambda_1 = \lambda_2 = \ldots = \lambda_k = \lambda$, nella (1.17) si ottiene la densità esponenziale (1.8).
Dalla (1.17) è possibile ricavare immediatamente il valore medio e la varianza del tempo di interarrivo, ossia

$$\begin{aligned} E(T) &= \sum_{j=1}^{k} \frac{p_j}{\lambda_j}\,, \\ \mathrm{Var}(T) &= E(T^2) - [E(T)]^2 = 2 \sum_{j=1}^{k} \frac{p_j}{\lambda_j^2} - \left[\sum_{j=1}^{k} \frac{p_j}{\lambda_j} \right]^2. \end{aligned} \tag{1.18}$$

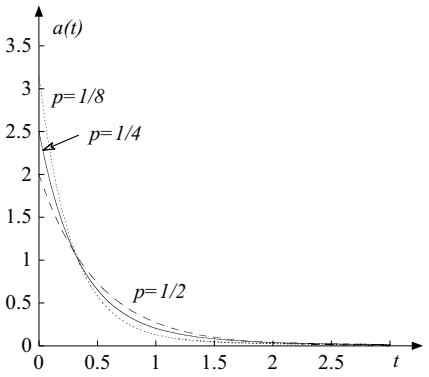

Figura 1.8: Densità iperesponenziale di ordine k con $\lambda = 2$, $k = 2$ e con $p = 1/2$, $p = 1/4$ e $p = 1/8$.

Se si desidera che il tempo medio di interarrivo sia $E(T) = 1/\lambda$, basta scegliere $\lambda_j = k\, p_j\, \lambda$ $(j = 1, 2, \ldots, k)$. La variabile aleatoria T è allora caratterizzata dalla seguente densità di probabilità iperesponenziale di ordine k:

$$a(t) = \begin{cases} k\,\lambda \sum_{j=1}^{k} p_j^2\, e^{-k\, p_j \lambda\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.19}$$

In Figura 1.8 è rappresentata sia la densità di probabilità (1.19) della variabile aleatoria T con densità iperesponenziale di ordine k con $\lambda = 2$, $k = 2$ e con $p = 1/2$, $p = 1/4$ e $p = 1/8$.

1.2.6 Meccanismo degli arrivi di tipo GI

Nel meccanismo degli arrivi di tipo GI si suppone che i tempi di interarrivo siano indipendenti e identicamente distribuiti con distribuzione di tipo generale. Occorre ricercare caratteristiche generali del sistema di servizio per una qualsiasi distribuzione dei tempi di interarrivo. Ovviamente quando si specifica il tipo di distribuzione (D, U, M, E_k, H_k,...) è possibile ottenere maggiori informazioni sull'evoluzione del sistema considerato.

1.3 Meccanismo di servizio

Per descrivere il meccanismo di servizio occorre conoscere la funzione di distribuzione delle variabili aleatorie $S_1, S_2, \ldots$, rappresentanti i tempi di servizio per ognuno degli utenti. Il generico S_i descrive la lunghezza dell'intervallo

di tempo occorrente per servire l'utente i-esimo ($i = 1, 2, \ldots$). Di particolare importanza sono alcune caratteristiche numeriche di tali variabili aleatorie, quali i valori medi $E(S_i)$, le varianze $\mathrm{Var}(S_i)$ e i coefficienti di variazione $C(S_i) = \sqrt{\mathrm{Var}(S_i)}/E(S_i)$ dei tempi di servizio S_i ($i = 1, 2, \ldots$).

Spesso si suppone che $S_1, S_2, \ldots$ sia una successione di variabili aleatorie indipendenti e identicamente distribuite. In tal caso se si denota con S una generica di tali variabili aleatorie, occorre specificare la sua funzione di distribuzione e la sua densità di probabilità. In letteratura la funzione di distribuzione di S viene solitamente denotata con $B(t) = P(S < t)$ e la sua densità di probabilità con $b(t)$. Alcune delle notazioni più frequentemente utilizzate per i tempi di servizio sono le seguenti:

D - tempi di servizio *iid* con funzione di distribuzione deterministica,

U - tempi di servizio *iid* con funzione di distribuzione uniforme,

M - tempi di servizio *iid* con funzione di distribuzione esponenziale,

E_k - tempi di servizio *iid* con funzione di distribuzione di Erlang di ordine k,

H_k - tempi di servizio *iid* con funzione di distribuzione iperesponenziale di ordine k,

G - tempi di servizio *iid* con funzione di distribuzione generale.

Analizziamo ora in dettaglio i vari meccanismi di servizio.

1.3.1 Meccanismo di servizio di tipo D

Il meccanismo di servizio più semplice che si possa immaginare è quello regolare; esso è caratterizzato da una cadenza temporale costante del servizio. Se si suppone quindi che il generico tempo di servizio sia di lunghezza fissa $1/\mu$, allora tale tempo può essere descritto da una variabile aleatoria S degenere la cui funzione di distribuzione è

$$B(t) = P(S < t) = \begin{cases} 0, & t \leq 1/\mu \\ 1, & t > 1/\mu. \end{cases} \tag{1.20}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{1}{\mu}, \qquad \mathrm{Var}(S) = 0, \qquad C(S) = 0. \tag{1.21}$$

Meccanismi di servizio di tipo D si possono presentare, ad esempio, in catene di montaggio in cui il tempo di produzione di un certo pezzo può ritenersi costante.

1.3.2 Meccanismo di servizio di tipo U

Nel meccanismo di servizio di tipo U i tempi di servizio sono indipendenti e identicamente distribuiti con funzione di distribuzione uniforme. Se supponiamo che la variabile aleatoria S sia uniformemente distribuita in (a, b), allora la funzione di distribuzione è

$$B(t) = P(S < t) = \begin{cases} 0, & t \le a \\ \dfrac{t-a}{b-a}, & a < t \le b \\ 1, & t > b \end{cases} \tag{1.22}$$

e quindi la sua densità di probabilità è

$$b(t) = \frac{dB(t)}{dt} = \begin{cases} \dfrac{1}{b-a}, & a < t < b \\ 0, & \text{altrimenti.} \end{cases} \tag{1.23}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{a+b}{2}, \qquad \mathrm{Var}(S) = \frac{(b-a)^2}{12}, \qquad C(S) = \frac{b-a}{\sqrt{3}\,(a+b)}\cdot \tag{1.24}$$

Si nota che il coefficiente di variazione $C(S)$ è sempre minore dell'unità.
Se si desidera che il tempo medio di servizio sia $1/\mu$ (come nel caso deterministico), basta scegliere $a = 0$ e $b = 2/\mu$. La variabile aleatoria S è allora uniformemente distribuita in $(0, 2/\mu)$ con densità di probabilità:

$$b(t) = \begin{cases} \dfrac{\mu}{2}, & 0 < t < \dfrac{2}{\mu} \\ 0, & \text{altrimenti.} \end{cases} \tag{1.25}$$

1.3.3 Meccanismo di servizio di tipo M

Nel meccanismo di servizio di tipo M i tempi di servizio sono indipendenti e identicamente distribuiti con funzione di distribuzione esponenziale. La lettera M significa Markov a causa della mancanza di memoria della funzione di distribuzione esponenziale. Sia S una variabile aleatoria esponenzialmente distribuita con valore medio $1/\mu$. La sua funzione di distribuzione è

$$B(t) = P(S < t) = \begin{cases} 0, & t \le 0 \\ 1 - e^{-\mu\, t}, & t > 0 \end{cases} \tag{1.26}$$

e quindi la sua densità di probabilità è

$$b(t) = \frac{dB(t)}{dt} = \begin{cases} \mu\, e^{-\mu\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.27}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{1}{\mu}\,, \qquad \mathrm{Var}(S) = \frac{1}{\mu^2}\,, \qquad C(S) = 1. \tag{1.28}$$

Il parametro μ è l'inverso del tempo medio di servizio, ossia del tempo medio necessario per servire un utente, e può essere interpretato come la frequenza media di partenza degli utenti per unità di tempo.

Per la mancanza di memoria della funzione di distribuzione esponenziale, si ha che il *tempo di servizio residuo* ha la stessa distribuzione del tempo di servizio.

La funzione di distribuzione esponenziale gode inoltre di un'altra interessante proprietà: il minimo di k variabili aleatorie $S_1, S_2, \ldots, S_k$ indipendenti e distribuite esponenzialmente con rispettivi valori medi $1/\mu_1, 1/\mu_2, \ldots, 1/\mu_k$ è ancora distribuito esponenzialmente con valore medio $1/(\mu_1 + \mu_2 + \ldots + \mu_k)$. Infatti, se si denota con

$$S = \min(S_1, S_2, \ldots, S_k),$$

allora quando $t > 0$ si ha:

$$\begin{aligned} P(S > t) &= P\{\min(S_1, S_2, \ldots, S_k) > t\} = P\{S_1 > t, S_2 > t, \ldots, S_k > t\} \\ &= P(S_1 > t)\, P(S_2 > t) \cdots P(S_k > t) = e^{-\mu_1 t} e^{-\mu_2 t} \cdots e^{-\mu_k t}, \end{aligned}$$

avendo utilizzato l'indipendenza delle variabili aleatorie e la loro distribuzione esponenziale. Quindi si ha

$$P(S < t) = \begin{cases} 1 - e^{-(\mu_1+\mu_2+\ldots+\mu_k)\, t} & t > 0 \\ \\ 0, & \text{altrimenti,} \end{cases}$$

ossia S è distribuita esponenzialmente con valore medio $1/(\mu_1 + \mu_2 + \ldots + \mu_k)$. Questa proprietà si rivela particolarmente utile quando si considera un sistema di servizio con k servitori identici che lavorano in parallelo i cui tempi di servizio sono distribuiti esponenzialmente con valore medio $1/\mu$. Se tutti i servitori sono occupati, il prossimo utente in fila di attesa per accedere al servizio dovrà attendere il minimo dei tempi residui di servizio dei k servitori. Tale tempo è quindi distribuito esponenzialmente con valore medio $1/(k\,\mu)$.

1.3.4 Meccanismo di servizio di tipo E_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.9, in cui il centro di servizio consiste di k identiche ed indipendenti fasi. Il tempo di servizio di una generica fase j ($j = 1, 2, \ldots, k$) è descritto da una variabile aleatoria esponenziale di valore medio $1/(k\,\mu)$. Un esempio tipico è quello di un centro di assistenza automobilistico che prevede varie operazioni elementari sulle auto (rifornimento, cambio dell'olio, controllo acqua, ingrassaggio, ...). Sia S una

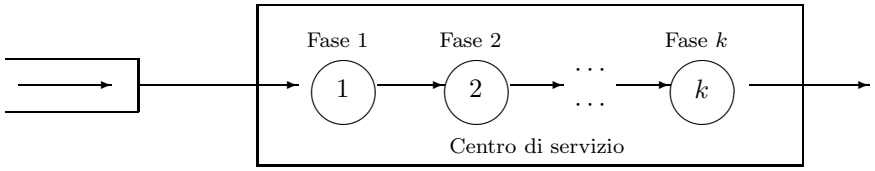

Figura 1.9: Tempi di servizio di tipo E_k nel caso in cui il centro di servizio prevede k successive fasi.

variabile aleatoria che descrive il tempo di servizio di un utente (ossia il tempo misurato dall'istante in cui l'utente entra nella prima fase fino a quando esce dalla k–esima fase). Inoltre sia S_j la variabile aleatoria che descrive il tempo di servizio alla stazione j–esima (ossia il tempo misurato dall'istante in cui l'utente entra nella fase j–esima fino a quando ne esce). Si nota che

$$S = S_1 + S_2 + \ldots + S_k, \tag{1.29}$$

ossia S è la somma di k variabili aleatorie indipendenti, ognuna distribuita esponenzialmente con valore medio $(k\mu)^{-1}$. Pertanto S è caratterizzata da una densità di Erlang di ordine k:

$$b(t) = \begin{cases} \dfrac{(k\,\mu)^k}{(k-1)!}\, e^{-k\,\mu\,t}\; t^{k-1}, & t > 0 \\ 0, & t \leq 0. \end{cases} \tag{1.30}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{1}{\mu}\,, \qquad \mathrm{Var}(S) = \frac{1}{k\,\mu^2}\,, \qquad C(S) = \frac{1}{\sqrt{k}}\cdot \tag{1.31}$$

Se si pone $k = 1$ nella (1.30) si riottiene la densità esponenziale (1.27). Inoltre, quando $k \to +\infty$ si nota che il coefficiente di variazione in (1.31) tende a zero come nel caso deterministico.
Si può pertanto affermare che quando viene applicata una distribuzione di Erlang di ordine k alla durata globale del servizio, ciò equivale ad immaginare il funzionamento del servizio organizzato in k fasi successive, a ciascuna delle quali è adibito un operatore specializzato il cui servizio ha una durata distribuita esponenzialmente; viceversa, quando un servizio viene svolto in k fasi successive ciascuna delle quali caratterizzata da una stessa distribuzione esponenziale, la durata globale del servizio si distribuisce secondo una distribuzione di Erlang di ordine k.

1.3.5 Meccanismo di servizio di tipo H_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.10, consistente in un centro di servizio costituito da un unico servitore che provvede a fornire k tipi di differenti servizi.

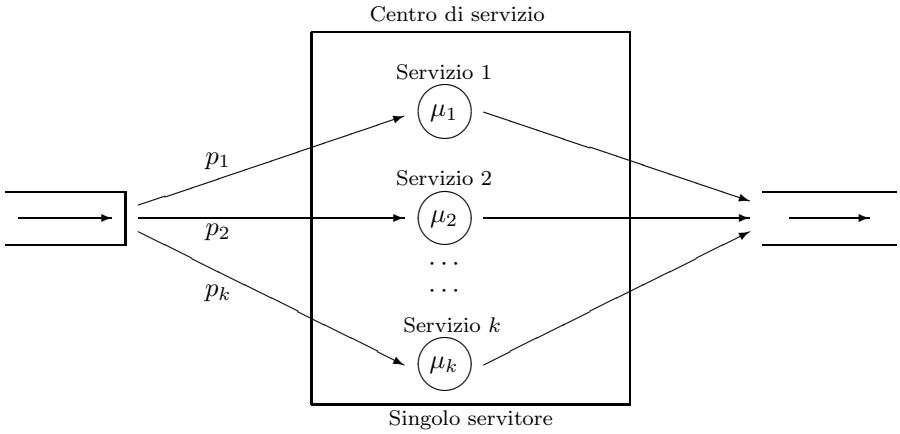

Figura 1.10: Tempi di servizio di tipo H_k per un singolo servitore che offre k differenti servizi.

Supponiamo che la probabilità che l'utente richieda un servizio di tipo j sia p_j per ogni $j = 1, 2, \ldots, k$, dove

$$p_j \geq 0 \qquad (j = 1, 2, \ldots, k), \qquad \sum_{j=1}^{k} p_j = 1. \tag{1.32}$$

Inoltre supponiamo che il j–esimo tipo di servizio ($j = 1, 2, \ldots, k$) sia caratterizzato da una durata del servizio distribuita esponenzialmente con valore medio $1/\mu_j$. Un esempio tipico è quello di un incaricato per l'assistenza al pubblico che fornisce k tipi di informazioni all'ingresso di un grande ufficio.

Denotiamo con S la variabile aleatoria che descrive il tempo di servizio, ossia il tempo necessario per soddisfare un tipo qualsiasi di richiesta fatta dall'utente. Inoltre, denotiamo S_j la variabile aleatoria che descrive il tempo di servizio degli utenti che richiedono il tipo j–esimo di servizio e con B_j l'evento "*è stata scelto dall'utente il j–esimo tipo di servizio*" ($j = 1, 2 \ldots, k$). Si nota immediatamente che l'evento $\{S < t\}$ può essere così scritto

$$\{S < t\} = \bigcup_{j=1}^{k} \Big[B_j \cap \{S < t\}\Big]. \tag{1.33}$$

Infatti, l'evento $\{S < t\}$ si realizza se si verifica uno qualunque dei k eventi incompatibili $B_1, B_2, \ldots, B_k$ ed inoltre $S < t$. Pertanto se $t > 0$, la probabilità della realizzazione dell'evento $\{S < t\}$ è quindi data dalla somma delle probabilità p_j (associata all'evento B_j) per la probabilità dell'evento $\{S_j < t\}$,

ossia

$$B(t) = P(S < t) = \sum_{j=1}^{k} P\Big[B_j \cap \{S < t\}\Big] = \sum_{j=1}^{k} P(B_j)\, P(S < t \mid B_j)$$
$$= \sum_{j=1}^{k} p_j\, P(S_j < t) = \sum_{j=1}^{k} p_j \left[1 - e^{-\mu_j\, t}\right].$$

Segue immediatamente che la funzione di distribuzione del tempo di servizio S è

$$B(t) = \begin{cases} 0, & t \le 0 \\ 1 - \displaystyle\sum_{j=1}^{k} p_j\, e^{-\mu_j\, t}, & t > 0 \end{cases} \tag{1.34}$$

e quindi la sua densità di probabilità è

$$b(t) = \begin{cases} \displaystyle\sum_{j=1}^{k} p_j\, \mu_j\, e^{-\mu_j\, t}, & t > 0 \\ 0, & \text{altrimenti,} \end{cases} \tag{1.35}$$

ossia una densità iperesponenziale di ordine k relativa alla durata del servizio. Ponendo $k = 1$ oppure $\mu_1 = \mu_2 \ldots = \mu_k = \mu$ nella (1.35) si ottiene la densità esponenziale (1.27).
Dalla (1.35) è possibile ricavare immediatamente il valore medio e la varianza del tempo di servizio, ossia

$$\begin{aligned} E(S) &= \sum_{j=1}^{k} \frac{p_j}{\mu_j}\,, \\ \mathrm{Var}(S) &= E(S^2) - [E(S)]^2 = 2 \sum_{j=1}^{k} \frac{p_j}{\mu_j^2} - \left[\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right]^2. \end{aligned} \tag{1.36}$$

Il coefficiente di variazione è quindi:

$$C(S) = \sqrt{\frac{2 \displaystyle\sum_{j=1}^{k} \frac{p_j}{\mu_j^2}}{\left[\displaystyle\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right]^2} - 1}\quad . \tag{1.37}$$

Proposizione 1.1 *$C(S) \ge 1$, con l'uguaglianza se e solo se $\mu_1 = \mu_2 = \ldots = \mu_k$.*

Dimostrazione Osserviamo in primo luogo che dalla seconda delle (1.36) segue

$$\mathrm{Var}(S) - [E(S)]^2 = 2\left[\sum_{j=1}^{k} \frac{p_j}{\mu_j^2} - \left(\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right)^2\right]. \tag{1.38}$$

Vogliamo mostrare che $C(S) \geq 1$, ossia che $\mathrm{Var}(S) - [E(S)]^2 \geq 0$. La disuguaglianza di Cauchy afferma che

$$\left(\sum_{j=1}^{k} x_j\, y_j\right)^2 \leq \left(\sum_{j=1}^{k} x_j^2\right)\left(\sum_{j=1}^{k} y_j^2\right), \tag{1.39}$$

con l'uguaglianza se e solo se esistono due numeri reali a e b non entrambi nulli tali che $a\, x_j + b\, y_j = 0$ $(j = 1, 2, \ldots, k)$. Ponendo $x_j = \sqrt{p_j}$, $y_j = \sqrt{p_j}/\mu_j$ per ogni $j = 1, 2, \ldots, k$, la disuguaglianza di Cauchy diventa

$$\left[\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right]^2 \leq \sum_{j=1}^{k} p_j \sum_{j=1}^{k} \frac{p_j}{\mu_j^2} = \sum_{j=1}^{k} \frac{p_j}{\mu_j^2} \tag{1.40}$$

e l'uguaglianza vale se e solo se esistono due numeri reali a e b non entrambi nulli tali che $a\sqrt{p_j} + b\sqrt{p_j}/\mu_j = 0$ per ogni $j = 1, 2, \ldots, k$, ossia se e solo se $\mu_1 = \mu_2 = \ldots = \mu_k$.
Facendo uso della (1.40) in (1.38) segue che $\mathrm{Var}(S) - [E(S)]^2 \geq 0$ e quindi $C(S) \geq 1$ con l'uguaglianza se e solo se $\mu_1 = \mu_2 = \ldots = \mu_k$. □

Se si desidera che il tempo medio di servizio sia $E(S) = 1/\mu$, basta scegliere $\mu_j = k\, p_j\, \mu$ $(j = 1, 2, \ldots, k)$. La variabile aleatoria S è allora caratterizzata dalla seguente densità di probabilità iperesponenziale di ordine k:

$$b(t) = \begin{cases} k\,\mu \displaystyle\sum_{j=1}^{k} p_j^2\, e^{-k\, p_j \mu\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.41}$$

1.3.6 Meccanismo di servizio di tipo G

Nel meccanismo di servizio di tipo G si suppone che i tempi di servizio siano indipendenti e identicamente distribuiti con distribuzione di tipo generale. Occorre ricercare caratteristiche generali del sistema di servizio per una qualsiasi distribuzione dei tempi di servizio. Ovviamente quando si specifica il tipo di distribuzione (D, U, M, E_k, H_k,...) è possibile ottenere maggiori informazioni sull'evoluzione del sistema in esame.

Occorre sottolineare che per una variabile aleatoria S, con $E(S) \neq 0$, il coefficiente di variazione

$$C(S) = \frac{\sqrt{\mathrm{Var}(S)}}{E(S)} \tag{1.42}$$

è un utile parametro per misurare il carattere della distribuzione di probabilità usata. Infatti, se S è deterministica $C(S) = 0$, se S è uniforme in $(0, 2/\mu)$ si ha $C(S) = 1/\sqrt{3}$, se S è esponenziale $C(S) = 1$, se S è caratterizzata da una distribuzione di Erlang di ordine k allora $C(S) = 1/\sqrt{k} \leq 1$ ed infine se S è caratterizzata da distribuzione iperesponenziale di ordine k allora $C(S) \geq 1$.

1.4 Notazioni nella teoria delle file di attesa

Per descrivere i sistemi di servizio si utilizza una speciale terminologia, dovuta a Kendall, ossia

$$A/B/s/K/m/Z \tag{1.43}$$

dove

A - descrive la distribuzione dei tempi di interarrivo,

B - la distribuzione dei tempi di servizio per ognuno dei servitori,

s - il numero di servitori (che lavorano in parallelo),

K - la capacità del sistema (ossia il massimo numero di utenti che possono essere presenti nel sistema inclusi quelli in servizio),

m - il numero di potenziali utenti nella sorgente,

Z - la disciplina di servizio.

Spesso si adopera la notazione abbreviata

$$A/B/s, \tag{1.44}$$

intendendo che non ci sono limitazioni alla lunghezza della fila di attesa, la sorgente è infinita e la disciplina di servizio è quella *FIFO*. I simboli scelti da Kendall e tradizionalmente usati per A e B sono

D - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione deterministica,

U - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione uniforme,

M - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione esponenziale,

E_k - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione di Erlang di ordine k,

H_k - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione iperesponenziale di ordine k,

GI - tempi di interarrivo *iid* con funzione di distribuzione generale,

G - tempi di servizio per servitore *iid* con funzione di distribuzione generale.

La notazione di Kendall $D/D/1$ significa che i tempi di interarrivo sono indipendenti e della stessa lunghezza, i tempi di servizio sono anche indipendenti e della stessa lunghezza, esiste un unico servitore, la sorgente è infinita, la capacità del sistema è infinita e la disciplina di servizio è quella *FIFO*. Invece, la notazione di Kendall $M/E_5/4/16/\infty/SIRO$ significa che i tempi di interarrivo sono indipendenti e identicamente distribuiti con legge esponenziale, i tempi di servizio sono indipendenti e identicamente distribuiti con legge di Erlang di ordine 5 per ognuno dei 4 servitori disponibili, la capacità del sistema è 16 (4 in servizio e 12 in fila di attesa), il numero di potenziali utenti nella sorgente è infinito e la disciplina di servizio è "service in random order". Inoltre, la notazione di Kendall $M/H_3/1/10$ mostra che i tempi di interarrivo sono indipendenti e identicamente distribuiti con legge esponenziale, i tempi di servizio sono indipendenti e identicamente distribuiti con legge iperesponenziale di ordine 3 (l'unico servitore offre tre differenti tipi di servizio), la capacità del sistema è 10 (massimo 9 utenti in fila di attesa e uno in servizio), il numero di potenziali utenti della sorgente è infinito e la disciplina di servizio è quella FIFO.

Quando si suppone che un sistema di servizio sia $M/G/s$, con tempi di interarrivo esponenziali e tempi di servizio generali per ognuno degli s servitori, si intende determinare delle relazioni valide per qualsiasi sistema di servizio di questo tipo. Pertanto tali relazioni debbono sussistere anche per il sistema di servizio $M/M/s$. Comunque, se si analizza il sistema di servizio $M/M/s$ si riescono ad ottenere maggiori informazioni rispetto al sistema di servizio generale $M/G/s$.

Occorre sottolineare che per quanto la notazione di Kendall sia molto utilizzata in letteratura, essa non permette di descrivere tutte le situazioni possibili, come ad esempio quello in cui si verificano arrivi degli utenti in gruppo e non singolarmente.

Capitolo 2

Analisi del sistema

2.1 Introduzione

In questo capitolo introdurremo le grandezze e i parametri prestazionali di maggiore interesse nell'analisi dei sistemi di servizio.

Il numero di utenti presenti in un sistema di servizio può essere descritto da un processo stocastico $\{N(t), t \geq 0\}$ discreto nello spazio degli stati e continuo nel tempo. Per ogni fissato istante di tempo t, $N(t)$ è una variabile aleatoria descrivente il numero di utenti presenti nel sistema al tempo t. Tale variabile aleatoria assume valori in un insieme finito $\{0, 1, \ldots, k\}$ se la capacità del sistema di servizio è k, mentre assume valori nell'insieme numerabile $\{0, 1, \ldots\}$ se la capacità del sistema di servizio è infinita.

2.2 Alcune misure prestazionali

L'indagine effettuata tramite la teoria delle file di attesa permette di descrivere:

- **lo stato del sistema** Se si denota con $\{N(t),\, t \geq 0\}$ il processo stocastico che descrive il numero $N(t)$ di utenti presenti nel sistema al tempo t occorre, se possibile, determinare

$$p_n(t) = P\{N(t) = n\} \qquad (n = 0, 1, \ldots), \tag{2.1}$$

ossia la probabilità che siano presenti n utenti nel sistema al tempo t ed alcune caratteristiche quali il valore medio $E[N(t)]$ e la varianza $Var[N(t)]$ del numero di utenti presenti nel sistema ad ogni fissato istante di tempo t. Occorre inoltre stabilire se il sistema raggiunge una *situazione di equilibrio statistico* ed in tal caso occorre calcolare la distribuzione di equilibrio

$$q_n = \lim_{t \to +\infty} p_n(t) \qquad (n = 0, 1, \ldots) \tag{2.2}$$

ed alcune caratteristiche quali il valore medio e la varianza del numero di utenti presenti nel sistema nella situazione di equilibrio statistico.

- **il tempo di permanenza nella fila di attesa di un utente** Il tempo di permanenza nella fila di attesa di un utente è una variabile aleatoria che descrive il tempo che un utente deve attendere in fila di attesa prima di essere servito.

- **il tempo di attesa di un utente nel sistema** Il tempo di attesa di un utente è una variabile aleatoria che descrive il tempo che un utente spende nel sistema, ossia il tempo che un utente deve attendere in fila di attesa più il suo tempo di servizio.

- **il periodo di occupazione** Il periodo di occupazione è una variabile aleatoria che descrive la lunghezza dell'intervallo di tempo che inizia con l'arrivo di un utente che trova l'unico servitore libero e continua fino a quando il servitore è per la prima volta nuovamente libero. Nel caso di più servitori il periodo di occupazione descrive la lunghezza l'intervallo di tempo che inizia con l'arrivo di un utente che trova tutti i servitori liberi e continua fino a che tutti i servitori sono per la prima volta nuovamente liberi. Il periodo di occupazione (*busy period*) quindi descrive la lunghezza dell'intervallo di tempo in cui il centro di servizio non è disponibile.

- **il tempo di ozio** Il tempo di ozio (*idle period*), detto anche periodo di soggiorno nello stato 0, è una variabile aleatoria che descrive la lunghezza dell'intervallo di tempo in cui il centro di servizio è inutilizzato.

Una tipica *realizzazione* di un sistema di servizio è una funzione a gradini, ossia una funzione costante a tratti con salti diretti verso il basso o verso l'alto ogni volta che accade un evento. Nella Figura 2.1 è rappresentata una realizzazione di un sistema di servizio con singolo servitore e disciplina FIFO e sono indicati gli istanti di arrivo $a_1, a_2, \ldots$ degli utenti, gli istanti di partenza $u_1, u_2, \ldots$ degli utenti, lo stato dei sistema $N(t)$ ai vari istanti di tempo t, i tempi di attesa nel sistema $W_1, W_2, \ldots$ degli utenti, i tempi di interarrivo $T_1, T_2, \ldots$ degli utenti, i tempi di servizio $S_1, S_2, \ldots$ per servitore degli utenti, i periodi di occupazione $B_1, B_2, \ldots$ del centro di servizio e i tempi di ozio $O_1, O_2, \ldots$ del centro di servizio.

Le notazioni fondamentali utilizzate nella teoria delle file di attesa sono le seguenti:

$N(t)$ - variabile aleatoria che descrive il numero di utenti presenti nel sistema (inclusi quelli in servizio) al tempo t;

$p_n(t)$ - probabilità che al tempo t siano presenti nel sistema n utenti (inclusi quelli in servizio);

N - variabile aleatoria che descrive il numero di utenti presenti nel sistema (inclusi quelli in servizio) nella situazione di equilibrio statistico;

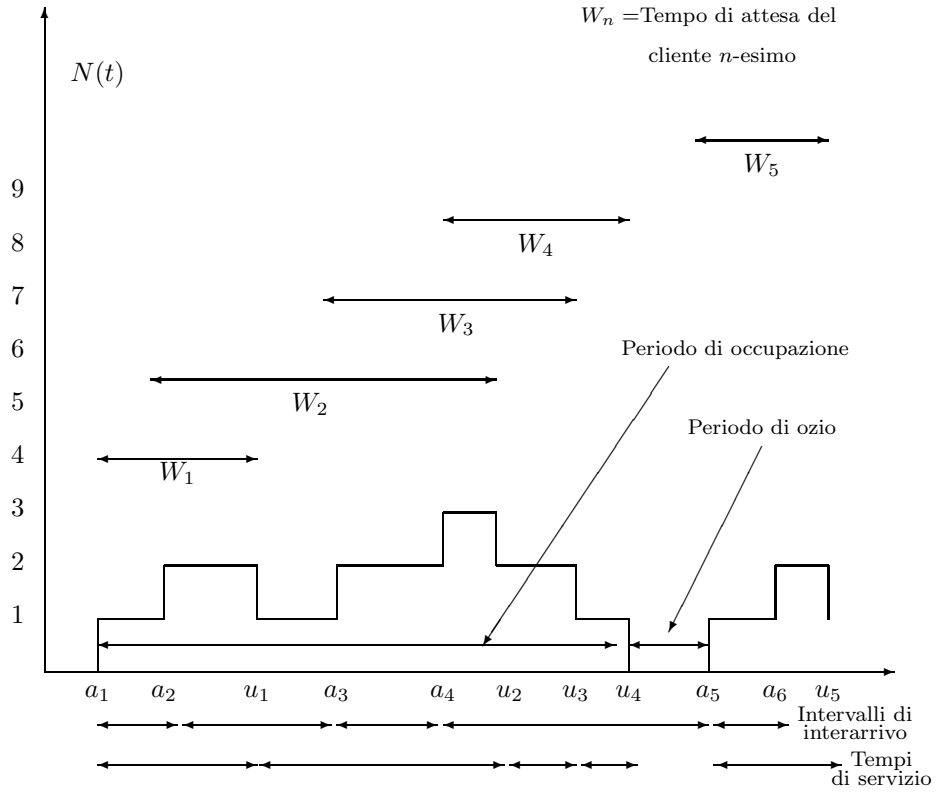

Figura 2.1: Una tipica realizzazione di un sistema di servizio.

q_n - probabilità che siano presenti nel sistema n utenti (inclusi quelli in servizio) nella situazione di equilibrio statistico;

$N_q(t)$ - variabile aleatoria che descrive il numero di utenti presenti nella fila di attesa al tempo t;

N_q - variabile aleatoria che descrive il numero di utenti presenti nella fila di attesa nella situazione di equilibrio statistico;

$N_s(t)$ - variabile aleatoria che descrive il numero di utenti in servizio al tempo t;

N_s - variabile aleatoria che descrive il numero di utenti in servizio nella situazione di equilibrio statistico;

T - variabile aleatoria che descrive il generico tempo di interarrivo nell'ipotesi in cui i tempi di interarrivo siano *iid*;

S - variabile aleatoria che descrive il generico tempo necessario ad un servitore per servire un utente nell'ipotesi in cui i tempi di servizio siano *iid*;

W - variabile aleatoria che descrive il tempo di attesa di un utente nel sistema incluso il suo tempo di servizio;

Q - variabile aleatoria che descrive il tempo che un utente spende nella fila di attesa prima di essere servito;

B - variabile aleatoria che descrive il periodo di occupazione del centro di servizio, ossia il periodo di tempo in cui esiste almeno un utente e quindi il centro di servizio è occupato da almeno un utente.

I - variabile aleatoria che descrive il tempo di ozio del centro di servizio, ossia il periodo di tempo in cui non ci sono utenti nel sistema ed il centro di servizio è inoperoso.

Un sistema di servizio alterna sempre periodi di occupazione e periodi di ozio. Come si evince dalla Figura 2.1 un periodo di ozio si presenta quando l'istante di partenza u_i dell'utente i–esimo è immediatamente seguito dall'istante a_{i+1} di arrivo dell'utente $(i+1)$–esimo creando il periodo di ozio (u_i, a_{i+1}). La Figura 2.2 mostra che i periodi di ozio possono essere riguardati come *tempi residui degli intervalli di interarrivo*.

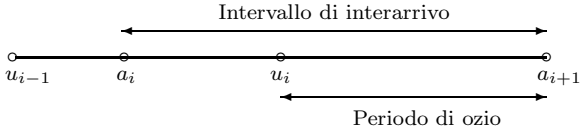

Figura 2.2: Un periodo di ozio del centro di servizio.

La Figura 2.3 mostra che sussiste la seguente relazione:

$$N(t) = N_q(t) + N_s(t) \qquad (t \geq 0), \tag{2.3}$$

ossia il numero di utenti presenti al tempo t è uguale alla somma del numero di utenti presenti nella fila di attesa al tempo t e del numero di utenti in servizio nello stesso istante di tempo. Nella situazione di equilibrio statistico, se esiste, si ha quindi:

$$N = N_q + N_s. \tag{2.4}$$

Sussiste inoltre la seguente relazione:

$$W = Q + S, \tag{2.5}$$

ossia il tempo di attesa di un utente nel sistema è uguale alla somma del suo tempo di attesa in coda Q e del suo tempo di servizio S.

Come suggerisce la notazione di Kendall, per valutare le misure prestazionali di un sistema di servizio occorre assumere che siano note alcune proprietà del

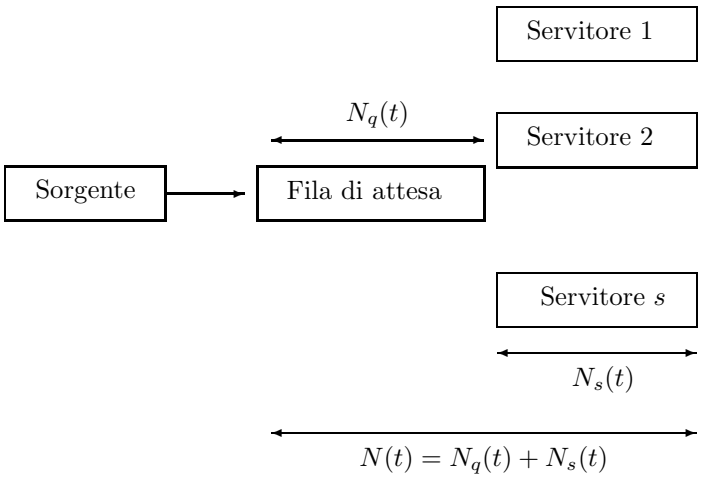

Figura 2.3: Numero di utenti presenti in un sistema di servizio al tempo t.

sistema stesso. Spesso si assume che siano note le distribuzioni dei tempi di interarrivo e dei tempi di servizio per ognuno dei servitori.

Se si denota con

λ^* : *frequenza media di arrivo per unità di tempo*

μ^* : *frequenza media di partenza da un generico servitore per unità di tempo*,

una misura prestazionale fondamentale in un sistema di servizio è rappresentata dall'*intensità di traffico* a, così definita

$$a = \frac{\lambda^*}{\mu^*}. \tag{2.6}$$

Tale coefficiente rappresenta *l'intensità del lavoro che svolge il sistema di servizio nella situazione di equilibrio statistico.*

Il rapporto tra la frequenza media degli arrivi e la frequenza totale delle partenze che il sistema di servizio può realizzare lavorando a pieno regime, ossia il rapporto tra l'intensità di traffico ed il numero di servitori presenti nel centro di servizio

$$\varrho^* = \frac{\lambda^*}{s\,\mu^*} \tag{2.7}$$

prende il nome di *fattore di utilizzazione del sistema.* Tale coefficiente rappresenta *l'intensità del lavoro di ognuno dei servitori nella situazione di equilibrio statistico.*

Quando più ϱ^* si avvicina all'unità tanto più il sistema tende ad avere tutti i posti di lavoro occupati con il rischio di entrare in congestione permanente ($\varrho^* = 1$). Quando $\varrho^* \geq 1$, ossia quando nell'unità di tempo la frequenza media

degli arrivi è maggiore o uguale della frequenza media delle partenze, il sistema non raggiunge una situazione di equilibrio statistico e la lunghezza della fila di attesa tende ad aumentare indefinitamente. Quindi, in un sistema di servizio a capacità infinita ϱ^* può essere interpretata come una *misura di congestione* del sistema.

2.3 Leggi di Little

Nella teoria delle file di attesa esistono delle relazioni che valgono sotto condizioni abbastanza generali. Tra queste relazioni rivestono un ruolo fondamentale le *leggi di Little*. Esse si applicano ad un qualsiasi *sistema di servizio in condizioni di equilibrio statistico*.

Se si denota con λ^* la frequenza media di arrivo nel sistema per unità di tempo, con $E(N)$ il numero medio di utenti nel sistema e con $E(W)$ il tempo medio di attesa di un utente nel sistema, la *prima legge di Little* afferma che

$$E(N) = \lambda^* E(W), \tag{2.8}$$

ossia il numero medio di utenti nel sistema è uguale al prodotto della frequenza media di arrivo nel sistema per unità di tempo e del tempo medio di attesa di un utente nel sistema.
La *seconda legge di Little* si applica alla fila di attesa e afferma che

$$E(N_q) = \lambda^* E(Q), \tag{2.9}$$

ossia il numero medio di utenti nella fila di attesa è uguale al prodotto della frequenza media di arrivo nel sistema per unità di tempo e del tempo medio di permanenza di un utente nel centro di attesa.
La legge di Little può essere formalizzata anche per il centro di servizio. Infatti, sottraendo membro a membro i termini delle relazioni (2.8) e (2.9), si ottiene

$$E(N) - E(N_q) = \lambda^* [E(W) - E(Q)].$$

Essendo $E(N) = E(N_q) + E(N_s)$ e $E(W) = E(Q) + E(S)$, la *terza legge di Little* afferma che

$$E(N_s) = \lambda^* E(S), \tag{2.10}$$

ossia il numero medio di utenti in servizio è uguale al prodotto della frequenza media di arrivo e del tempo medio di servizio.
Le tre leggi di Little rivestono notevole importanza nelle teoria delle file di attesa poiché esse non dipendono dalla distribuzione dei tempi di interarrivo e dei tempi di servizio, dal numero di servitori nel sistema e dalla disciplina di servizio.

Una dimostrazione rigorosa delle relazioni (2.8) e (2.9) è stata fornita da Little (1961) e perciò tali relazioni sono note come *formule di Little*. La dimostrazione rigorosa di queste relazioni si rivela complessa dal punto di vista

matematico; esistono comunque varie dimostrazioni di tipo euristico molto più semplici.

Una dimostrazione più semplice, che descriveremo nel seguito, delle relazioni (2.8) e (2.9) è stata fornita da Eilon (1969). Essa non dipende *(i)* dalla distribuzione dei tempi di interarrivo e dei tempi di servizio, *(ii)* dal numero di servitori nel sistema e *(iii)* dalla disciplina di servizio.

2.3.1 Formula di Little per l'intero sistema

Dimostriamo la prima formula di Little, ossia la (2.8).

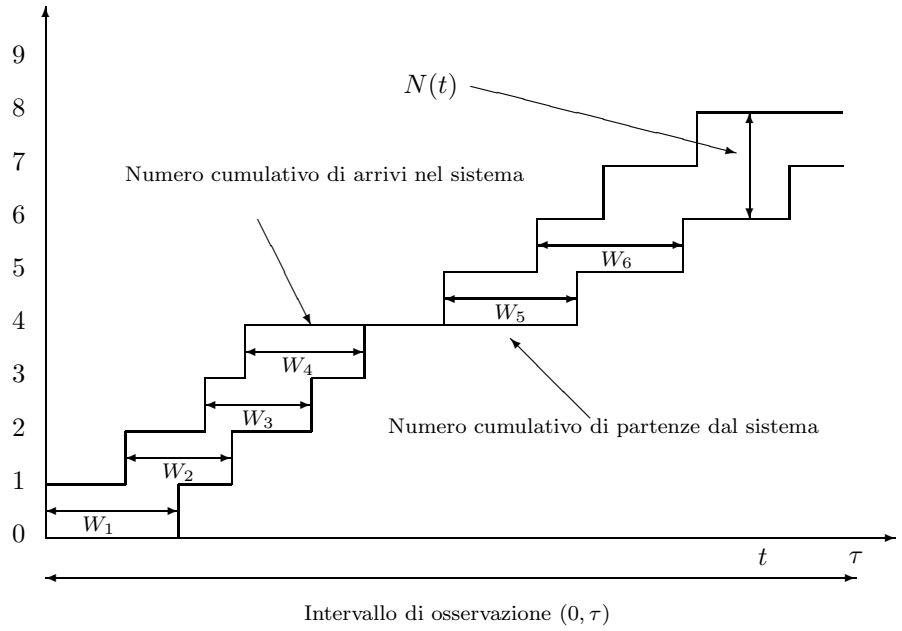

Figura 2.4: Numero cumulativo di arrivi e numero cumulativo di partenze dal sistema di servizio.

Nella Figura 2.4 la linea superiore descrive il *numero cumulativo di arrivi* e la linea inferiore il *numero cumulativo di partenze dal sistema*. La distanza verticale tra due linee fornisce il numero di utenti $N(t)$ presenti nel sistema ad un fissato istante t, mentre la distanza orizzontale denota il tempo di attesa nel sistema (tempo di permanenza in fila di attesa più tempo di servizio). Supponiamo che il sistema sia stato in funzione per un certo tempo e che successivamente abbia raggiunto una situazione di equilibrio statistico. Consideriamo un intervallo di tempo $(0, \tau)$ che può includere nessuno, uno o più periodi di occupazione. Denotiamo con

- $N_a(\tau)$ numero totale di arrivi durante l'intervallo $(0, \tau)$

- $\overline{\lambda}(\tau)$ frequenza media di arrivo per unità di tempo nell'intervallo $(0, \tau)$. Si nota che

$$\overline{\lambda}(\tau) = \frac{N_a(\tau)}{\tau}\,, \tag{2.11}$$

ossia frequenza media di arrivo per unità di tempo nell'intervallo $(0, \tau)$ è data dal rapporto tra il numero totale di arrivi durante l'intervallo $(0, \tau)$ e la lunghezza dell'intervallo.

Siano inoltre

- $W_c(\tau)$ tempo totale di attesa (tempo cumulativo di attesa) nel sistema di tutti gli utenti che arrivano nell'intervallo $(0, \tau)$.

È evidente che

$$W_c(\tau) = W_1 + W_2 + \ldots,$$

ossia il tempo totale di attesa nel sistema è la somma dei tempi di attesa dei vari utenti che sono arrivati nell'intervallo $(0, \tau)$. Osservando la Figura 2.4 si nota anche che

$$W_c(\tau) = 1 \times W_1 + 1 \times W_2 + \ldots = \int_0^\tau N(t)\, dt,$$

ossia $W_c(\tau)$ descrive l'area compresa tra le due linee nell'intervallo $(0, \tau)$. Indichiamo inoltre con

- $\overline{W}(\tau)$ media dei tempi di attesa nel sistema degli utenti arrivati durante l'intervallo $(0, \tau)$. Risulta che

$$\overline{W}(\tau) = \frac{W_c(\tau)}{N_a(\tau)} = \frac{1}{N_a(\tau)} \int_0^\tau N(t)\, dt, \tag{2.12}$$

ossia la media del tempo di attesa nel sistema degli utenti arrivati durante l'intervallo $(0, \tau)$ è uguale al rapporto tra il tempo totale di attesa nel sistema di tutti gli utenti che sono arrivati nell'intervallo $(0, \tau)$ ed il numero totale di arrivi in tale intervallo. Denotiamo infine con

- $\overline{N}(\tau)$ media per unità di tempo del numero di utenti nel sistema nell'intervallo $(0, \tau)$.

Si nota che

$$\overline{N}(\tau) = \frac{1}{\tau} \int_0^\tau N(t)\, dt = \frac{W_c(\tau)}{\tau}\,, \tag{2.13}$$

ossia la media per unità di tempo del numero di utenti nel sistema è uguale al rapporto tra tempo totale di attesa nel sistema di tutti i utenti che arrivano nell'intervallo $(0, \tau)$ e la lunghezza di tale intervallo.
Dalle relazioni (2.11), (2.12) e (2.13) segue che

$$\overline{N}(\tau) = \frac{W_c(\tau)}{\tau} = \frac{W_c(\tau)}{N_a(\tau)}\, \frac{N_a(\tau)}{\tau} = \overline{W}(\tau)\, \overline{\lambda}(\tau),$$

ossia

$$\overline{N}(\tau) = \overline{\lambda}(\tau)\, \overline{W}(\tau). \tag{2.14}$$

Supponiamo che quando $\tau \to +\infty$ esistano finiti i limiti di $\overline{\lambda}(\tau)$ e di $\overline{W}(\tau)$:

$$\lambda^* = \lim_{\tau \to +\infty} \overline{\lambda}(\tau)\,, \qquad E(W) = \lim_{\tau \to +\infty} \overline{W}(\tau). \tag{2.15}$$

In tali ipotesi, dalla (2.14) segue che esiste finito anche il limite di $\overline{N}(\tau)$ quando $\tau \to +\infty$ e risulta

$$E(N) = \lim_{\tau \to +\infty} \overline{N}(\tau).$$

La prima formula di Little, ossia la (2.8), segue quindi immediatamente dalla (2.14) procedendo al limite per $\tau \to +\infty$.

2.3.2 Formula di Little per la fila di attesa

Dimostriamo ora la seconda formula di Little per la fila di attesa, ossia la (2.9).

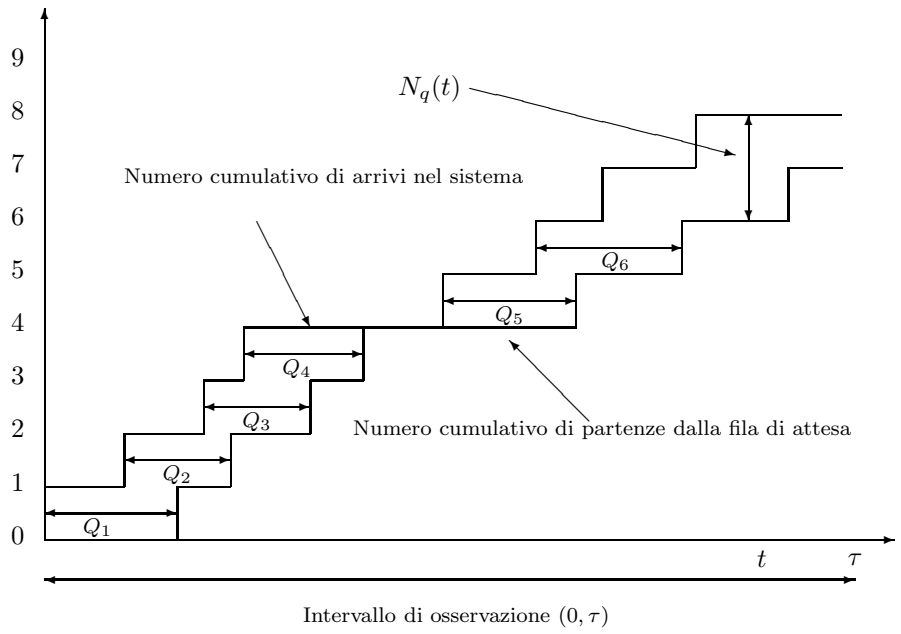

Figura 2.5: Numero cumulativo di arrivi e numero cumulativo di partenze dalla fila di attesa.

Nella Figura 2.5 la linea superiore descrive il *numero cumulativo di arrivi* e la linea inferiore il *numero cumulativo di partenze dalla fila di attesa*. La distanza verticale tra due linee fornisce il numero di utenti $N_Q(t)$ presenti nella fila di

attesa ad un fissato istante t, mentre la distanza orizzontale denota il tempo di permanenza nella fila di attesa. Supponiamo che il sistema sia stato in funzione per un certo tempo e che successivamente abbia raggiunto una situazione di equilibrio statistico. Consideriamo un intervallo di tempo $(0, \tau)$ che può includere nessuno o più periodi di occupazione. Denotiamo nuovamente con $N_a(\tau)$ il numero totale di arrivi durante l'intervallo $(0, \tau)$ e con $\overline{\lambda}(\tau)$ la frequenza media di arrivo per unità di tempo nell'intervallo $(0, \tau)$. Risulta nuovamente essere valida la (2.11). Indichiamo inoltre con

- $Q_c(\tau)$ tempo totale di permanenza nella fila di attesa di tutti gli utenti che arrivano nell'intervallo $(0, \tau)$;

  Si nota che
  $$Q_c(\tau) = Q_1 + Q_2 + \ldots,$$
  ossia il tempo totale di permanenza nella fila di attesa è la somma dei tempi di permanenza nella fila di attesa dei vari utenti che sono arrivati nell'intervallo $(0, \tau)$. Osservando la figura si nota che
  $$Q_c(\tau) = 1 \times Q_1 + 1 \times Q_2 + \ldots = \int_0^\tau N_q(t)\, dt$$
  ossia $Q_c(\tau)$ descrive l'area compresa tra le due linee nell'intervallo $(0, \tau)$. Indichiamo inoltre con

- $\overline{Q}(\tau)$ la media dei tempi di permanenza nella fila di attesa degli utenti arrivati durante l'intervallo $(0, \tau)$. Si nota che
  $$\overline{Q}(\tau) = \frac{Q_c(\tau)}{N_a(\tau)}, \tag{2.16}$$
  ossia la media del tempo di permanenza nel sistema degli utenti arrivati durante l'intervallo $(0, \tau)$ è uguale al rapporto tra il tempo totale di permanenza nella fila di attesa di tutti gli utenti che sono arrivati nell'intervallo $(0, \tau)$ ed il numero totale di arrivi in tale intervallo. Denotiamo infine con

- $\overline{N}_q(\tau)$ la media per unità di tempo del numero di utenti nella fila di attesa nell'intervallo $(0, \tau)$.

È evidente che
$$\overline{N}_q(\tau) = \frac{1}{\tau} \int_0^\tau N_q(t)\, dt = \frac{Q_c(\tau)}{\tau}, \tag{2.17}$$
ossia la media per unità di tempo del numero di utenti nella fila di attesa è uguale al rapporto tra tempo totale di permanenza nella fila di attesa di tutti gli utenti che arrivano nell'intervallo $(0, \tau)$ e la lunghezza di tale intervallo. Dalle relazioni (2.11), (2.16) e (2.17) segue che
$$\overline{N}_q(\tau) = \frac{Q_c(\tau)}{\tau} = \frac{Q_c(\tau)}{N_a(\tau)}\, \frac{N_a(\tau)}{\tau} = \overline{Q}(\tau)\, \overline{\lambda}(\tau),$$

ossia

$$\overline{N}_q(\tau) = \overline{\lambda}(\tau)\ \overline{Q}(\tau). \tag{2.18}$$

Supponiamo che quando $\tau \to +\infty$ esistano finiti i limiti di $\overline{\lambda}(\tau)$ e di $\overline{Q}(\tau)$:

$$\lambda^* = \lim_{\tau\to+\infty} \overline{\lambda}(\tau)\,, \qquad E(Q) = \lim_{\tau\to+\infty} \overline{Q}(\tau). \tag{2.19}$$

In tali ipotesi, dalla (2.18) segue che esiste finito anche il limite di $\overline{N}_q(\tau)$ quando $\tau \to +\infty$ e risulta

$$E(N_q) = \lim_{\tau\to+\infty} \overline{N}(\tau).$$

La seconda formula di Little, ossia la (2.9), segue quindi immediatamente dalla (2.18) procedendo al limite per $\tau \to +\infty$.

2.4 Periodi di occupazione e di ozio

Consideriamo un sistema di servizio con un unico servitore. Tale sistema alterna periodi di ozio (quando non ci sono utenti nel sistema e quindi il servitore è libero) e periodi di occupazione (quando esiste almeno un utente nel sistema ed il servitore è occupato).

Denotiamo con $I_1, I_2, \ldots$ e $B_1, B_2, \ldots$ rispettivamente le lunghezze dei periodi di ozio e dei periodi di occupazione. Nella situazione di equilibrio statistico la probabilità che il sistema sia asintoticamente vuoto può essere così calcolata:

$$q_0 = \lim_{n\to\infty} \frac{I_1 + I_2 + \ldots + I_n}{I_1 + I_2 + \ldots + I_n + B_1 + B_2 + \ldots + B_n}\,. \tag{2.20}$$

Se le successioni $I_1, I_2, \ldots$ e $B_1, B_2, \ldots$ sono entrambe costituite da variabili aleatorie indipendenti e identicamente distribuite, allora dividendo il numeratore ed il denominatore del rapporto in (2.20) per n ed applicando la legge dei grandi numeri si ha

$$\begin{aligned} q_0 &= \lim_{n\to\infty} \frac{(I_1 + I_2 + \ldots + I_n)/n}{(I_1 + I_2 + \ldots + I_n)/n + (B_1 + B_2 + \ldots + B_n)/n} \\ &= \frac{E(I)}{E(I) + E(B)}\,, \end{aligned} \tag{2.21}$$

ossia la probabilità che il sistema sia asintoticamente vuoto è uguale al rapporto tra il tempo medio di ozio e la somma del tempo medio di ozio e del tempo medio di occupazione del servitore. Dalla (2.21) si ricava:

$$1 - q_0 = \frac{E(B)}{E(I) + E(B)}, \tag{2.22}$$

ossia la probabilità che nel sistema sia presente almeno un utente è uguale al rapporto tra il tempo medio di occupazione e la somma del tempo medio di ozio

e del tempo medio di occupazione del servitore. La (2.22) può anche essere così scritta

$$E(B) = \frac{(1 - q_0)\, E(I)}{q_0}\,, \tag{2.23}$$

Come mostra la Figura 2.2 i periodi di ozio del servitore possono essere riguardati come tempi residui dei tempi di interarrivo.

Se si ipotizza che i tempi di interarrivo sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$, allora ricordando la proprietà di mancanza di memoria della distribuzione esponenziale si ricava che anche i tempi residui sono caratterizzati dalla stessa distribuzione esponenziale. In tale ipotesi la densità di probabilità del periodo di ozio è quindi:

$$f_I(t) = \begin{cases} \lambda\, e^{-\lambda\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{2.24}$$

Essendo $E(I) = 1/\lambda$, dalla relazione (2.23) si ottiene $E(B) = (1 - q_0)/(\lambda\, q_0)$.

In conclusione, nella situazione di equilibrio statistico, se i tempi di interarrivo sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$, allora il tempo medio di ozio e il tempo medio di occupazione del servitore sono rispettivamente:

$$E(I) = \frac{1}{\lambda}, \qquad E(B) = \frac{1 - q_0}{\lambda\, q_0}. \tag{2.25}$$

Capitolo 3

Processi di nascita morte

3.1 Introduzione

In questo capitolo siamo interessati ad introdurre alcuni processi stocastici discreti nello spazio degli stati e continui nel tempo, ossia il processo di Poisson e i processi di nascita morte. Per i processi di nascita morte individueremo le condizioni affinché si raggiunga l'equilibrio statistico e determineremo la distribuzione di equilibrio. I processi di nascita morte saranno utilizzati nel prossimo capitolo per analizzare alcuni sistemi di servizio determinando i loro principali parametri prestazionali.

Sia $\{N(t),\ t \geq 0\}$ un processo stocastico continuo nel tempo e discreto nello spazio degli stati. Per ogni fissato $t \geq 0$, $N(t)$ è una variabile aleatoria discreta che assume un numero finito o al più numerabile di valori. Le realizzazioni di tale processo sono funzioni a gradino, ossia funzioni costanti a tratti con salti diretti verso il basso o verso l'alto ogni volta che si verifica un cambiamento di stato. Le situazioni fisiche da cui tali processi sorgono sono quelle in cui lo stato del sistema è caratterizzato da un numero intero di particelle (o individui, utenti) e i cambiamenti di stato rappresentano l'addizione o la sottrazione di particelle dal sistema in vario modo: nascite, morti, immigrazioni, emigrazioni,. . ..

3.2 Processo stocastico di Poisson

Il più semplice processo stocastico continuo nel tempo e discreto nello spazio degli stati è il *processo di Poisson*. Tale processo si rivela utile nella descrizione di alcuni fenomeni che evolvono nel tempo quali l'arrivo di chiamate ad un centralino telefonico, l'attività spontanea di certi neuroni, l'emissione di particelle da una sorgente radioattiva, . . .

Supponiamo di indicare con $N(t) \quad (t \geq 0)$ il numero di arrivi (ad esempio, chiamate che si presentano ad un centralino telefonico) nell'intervallo di tempo

$(0,t)$ e con $N(t,t+\Delta t)$ il numero di arrivi nell'intervallo $(t,t+\Delta t)$. Ovviamente risulta $N(t,t+\Delta t) = N(t+\Delta t) - N(t)$.

Definizione 3.1 *Un processo stocastico* $\{N(t),\ t \geq 0\}$ *è detto di Poisson con parametro* ϱ $(\varrho > 0)$ *se si ha:*

(i) $N(0) = 0$,

(ii) il processo ha incrementi indipendenti e stazionari,

(iii) $P\big\{N(t,t+\Delta t) = 1\big\} = \varrho\,\Delta t + o(\Delta t)$,

(iv) $P\big\{N(t,t+\Delta t) > 1\big\} = o(\Delta t)$,
dove $o(\Delta t)$ *è un infinitesimo di ordine superiore rispetto a* Δt *e* ϱ *denota il parametro di arrivo con dimensione fisiche* $[tempo]^{-1}$.

La condizione *(i)* significa assumere che fino al tempo $t = 0$ non si sono verificati eventi. La condizione *(ii)* assicura che gli eventi che si verificano in intervalli di tempo disgiunti, ossia che non si sovrappongono, sono stocasticamente indipendenti (il processo ha incrementi indipendenti) e inoltre la distribuzione del numero di eventi che si verificano in ogni intervallo di tempo dipende soltanto dalla lunghezza dell'intervallo considerato (il processo ha incrementi stazionari). Le condizioni *(iii)* e *(iv)* invece assicurano che

$$P\big\{N(t,t+\Delta t) = 0\big\} = 1 - \varrho\,\Delta t + o(\Delta t).$$

Inoltre, la condizione *(iv)* mostra che in un piccolo intervallo di tempo $(t,t+\Delta t)$ gli eventi si verificano al più singolarmente. Dalla condizione *(ii)* di indipendenza scaturisce che

$$\begin{aligned}
&P\big\{N(t,t+\Delta t) = 1 | N(t) = n\big\} = \varrho\,\Delta t + o(\Delta t) \qquad (n = 0,1,\ldots),\\
&P\big\{N(t,t+\Delta t) = 0 | N(t) = n\big\} = 1 - \varrho\,\Delta t + o(\Delta t) \qquad (n = 0,1,\ldots),\\
&P\big\{N(t,t+\Delta t) > 1 | N(t) = n\big\} = o(\Delta t) \qquad (n = 0,1,\ldots).
\end{aligned}$$

Denotiamo con

$$p_n(t) = P\big\{N(t) = n\big\} \qquad (n = 0,1,\ldots)$$

la probabilità che sia n il numero di arrivi fino al tempo t, ossia nell'intervallo di tempo $(0,t)$.

Proposizione 3.1 *Per un processo stocastico di Poisson* $\{N(t),\ t \geq 0\}$ *di parametro* $\varrho > 0$, *si ha*

$$p_n(t) = \frac{(\varrho\,t)^n}{n!}\, e^{-\varrho\,t} \qquad (n = 0,1,\ldots), \tag{3.1}$$

ossia per ogni fissato t *si ottiene una funzione di probabilità di Poisson di parametro* $\varrho\,t$.

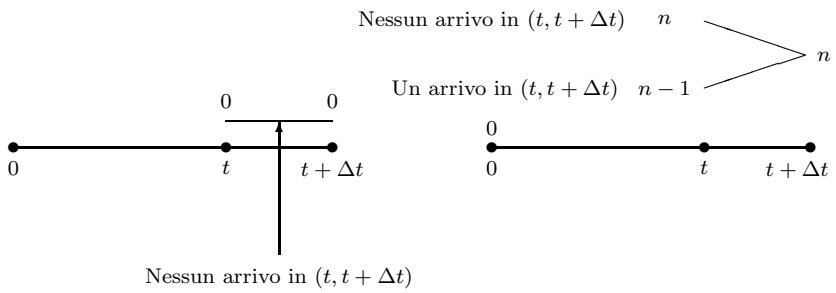

Figura 3.1: Cambiamenti di stato in $(t, t+\Delta t)$ nel processo di Poisson.

Dimostrazione Come si evince dalla Figura 3.1, si nota che sussistono le seguenti identità:

$$\begin{aligned}
p_0(t+\Delta t) &= P\big\{N(t+\Delta t)=0\big\} = P\big\{N(t)=0, N(t,t+\Delta t)=0\big\} \\
&= p_0(t)\ (1-\varrho\ \Delta t) + o(\Delta t) \\
p_n(t+\Delta t) &= P\big\{N(t+\Delta t)=n\big\} = P\big\{N(t)=n, N(t,t+\Delta t)=0\big\} \\
&\qquad + P\big\{N(t)=n-1, N(t,t+\Delta t)=1\big\} + o(\Delta t) \\
&= p_n(t)\ (1-\varrho\ \Delta t) + p_{n-1}(t)\ \varrho\ \Delta t + o(\Delta t) \qquad (n=1,2,\ldots),
\end{aligned}$$

ossia

$$\begin{aligned}
&\frac{p_0(t+\Delta t)-p_0(t)}{\Delta t} = -\varrho\ p_0(t) + \frac{o(\Delta t)}{\Delta t}, \\
&\frac{p_n(t+\Delta t)-p_n(t)}{\Delta t} = -\varrho\ p_n(t) + \varrho\ p_{n-1}(t) + \frac{o(\Delta t)}{\Delta t} \qquad (n=1,2,\ldots).
\end{aligned}$$

Procedendo al limite quando $\Delta t \to 0$ si ricava:

$$\begin{aligned}
&\frac{dp_0(t)}{dt} = -\varrho\ p_0(t), \\
&\frac{dp_n(t)}{dt} = -\varrho\ p_n(t) + \varrho\ p_{n-1}(t) \qquad (n=1,2,\ldots).
\end{aligned} \tag{3.2}$$

Abbiamo ottenuto un sistema di equazioni differenziali e alle differenze del primo ordine in n che, per l'ipotesi *(i)*, deve essere risolto con le condizioni iniziali

$$p_n(0) = \begin{cases} 1, & n=0 \\ 0, & n=1,2,\ldots \end{cases} \tag{3.3}$$

Consideriamo la funzione generatrice di probabilità

$$G(z,t) = \sum_{n=0}^{+\infty} z^n\ p_n(t). \tag{3.4}$$

Moltiplicando ambo i membri della seconda delle (3.2) per z^n e sommando su $n = 1, 2, \ldots$ si ha:

$$\frac{\partial}{\partial t}\Big[G(z,t) - p_0(t)\Big] = -\varrho\,\Big[G(z,t) - p_0(t)\Big] + \varrho\, z\, G(z,t)$$

da cui, utilizzando la prima delle (3.2), si ottiene

$$\frac{\partial G(z,t)}{\partial t} = \varrho\,(z-1)\,G(z,t). \tag{3.5}$$

Ricordando (3.3) e (3.4) segue che l'equazione (3.5) deve essere risolta con la condizione iniziale:

$$G(z,0) = \sum_{n=0}^{+\infty} z^n\, p_n(0) = 1. \tag{3.6}$$

La soluzione della (3.5), con la condizione iniziale (3.6), è:

$$G(z,t) = \exp\{\varrho\, t\,(z-1)\} = e^{-\varrho\, t}\; e^{\varrho\, t\, z}. \tag{3.7}$$

Espandendo $e^{\varrho\, t\, z}$ in serie di potenze di z, la (3.7) diventa

$$G(z,t) = e^{-\varrho\, t} \sum_{n=0}^{+\infty} \frac{(\varrho\, t)^n}{n!}\; z^n. \tag{3.8}$$

Uguagliando uguali potenze di z in (3.4) e (3.8) segue immediatamente la (3.1). Per ogni fissato t abbiamo quindi ottenuto una funzione di probabilità di Poisson di parametro $\varrho\, t$. □

Nella Figura 3.2 dall'alto verso il basso sono riportate le probabilità $p_0(t)$, $p_1(t)$, $p_2(t)$, $p_3(t)$, $p_4(t)$ e $p_5(t)$ in funzione di $\varrho\, t$. Si nota che mentre $p_0(t)$ è una funzione decrescente in $\varrho\, t$, le probabilità $p_n(t) \quad (n = 1, 2, \ldots)$ presentano un punto di massimo quando $\varrho\, t = n$. Il valore medio e la varianza del numero di arrivi nell'intervallo $(0, t)$ sono:

$$E[N(t)] = \sum_{n=1}^{+\infty} n\, p_n(t) = \varrho\, t, \qquad \mathrm{Var}[N(t)] = \varrho\, t. \tag{3.9}$$

Inoltre, il coefficiente di variazione è:

$$C[N(t)] = \frac{\sqrt{\mathrm{Var}[N(t)]}}{E[N(t)]} = \frac{1}{\sqrt{\varrho\, t}}\,.$$

Si nota che $\lim_{t \to +\infty} C[N(t)] = 0$, che evidenzia che al crescere del tempo il numero medio di arrivi diventa sempre più significativo.

Un'importante proprietà del processo stocastico di Poisson di parametro ϱ è che *i tempi di interarrivo*, ossia le lunghezze degli intervalli tra due arrivi successivi, *sono indipendenti e identicamente distribuiti con densità esponenziale*

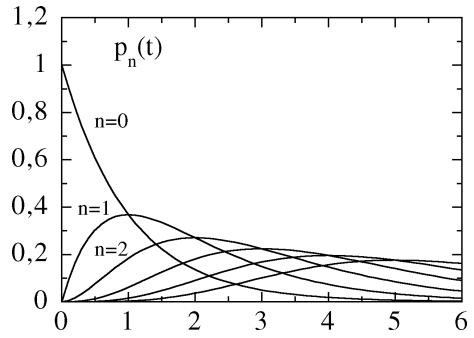

Figura 3.2: Le probabilità $p_n(t)$ $(n = 0, 1, \ldots, 5)$ del processo di Poisson in funzione di $\varrho\, t$.

di parametro ϱ. Quindi, se supponiamo che gli arrivi si verifichino ai tempi T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, $\cdots$, dove T_n denota la lunghezza dell'intervallo aleatorio di tempo tra l'evento $(n-1)$–esimo e l'evento n–esimo, si ha:

$$f_{T_n}(t) = \begin{cases} \varrho\, e^{-\varrho\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Il valore medio e la varianza dei tempi di interarrivo sono rispettivamente:

$$E(T_n) = \frac{1}{\varrho}\,, \qquad \mathrm{Var}(T_n) = \frac{1}{\varrho^2} \qquad (n = 1, 2, \ldots)$$

e il coefficiente di variazione è quindi unitario.

Inoltre, in un processo stocastico di Poisson di parametro $\varrho > 0$, la variabile aleatoria $T_1 + T_2 + \ldots + T_k$, che descrive la lunghezza l'intervallo di tempo fino all'arrivo k–esimo, è caratterizzata da densità di probabilità

$$f(t) = \frac{dP(T_1 + T_2 + \ldots + T_k < t)}{dt} = \begin{cases} \dfrac{\varrho^k}{(k-1)!}\, e^{-\varrho t}\, t^{k-1}, & t > 0 \\ 0, & \text{altrimenti,} \end{cases}$$

ossia una densità di Erlang di ordine k e di parametro ϱ.

Esempio 3.1 Supponiamo che gli arrivi ad un sistema di servizio si verifichino in accordo ad un processo di Poisson e che il tempo medio tra due arrivi successivi sia di 25 secondi. Determinare la funzione di distribuzione $P(T < t)$ dei tempi di interarrivo e la probabilità che nell'intervallo $(0, t)$ si siano verificati n arrivi, misurando il tempo in minuti. Calcolare infine la media e la varianza del numero di arrivi alla fine della prima ora.
Si nota che

$$\frac{1}{\lambda} = 25 \text{ secondi} = 25\,\frac{1}{60} \text{ minuti} = \frac{5}{12} \text{ minuti},$$

da cui segue che la frequenza di arrivo al minuto è $\lambda = 12/5 = 2.4$. La funzione di distribuzione dei tempi di interarrivo è

$$P(T < t) = 1 - e^{-2.4\,t} \qquad (t > 0),$$

con il tempo t misurato in minuti. La probabilità che si verifichino n arrivi nell'intervallo $(0, t)$, con t misurato in minuti, è

$$p_n(t) = P\{N(t) = n\} = \frac{(2.4\,t)^n}{n!}\, e^{-2.4\,t} \qquad (n = 0, 1, \ldots).$$

Per calcolare la media e la varianza del numero di arrivi alla fine della prima ora, basta porre $t = 60$ minuti e ricordare la (3.9):

$$E[N(60)] = \mathrm{Var}[N(60)] = 2.4 \cdot 60 = 144,$$

ossia si hanno in media 144 arrivi dopo la prima ora con una deviazione standard di 12. ♢

Il processo stocastico di Poisson è di fondamentale importanza nella costruzione di vari modelli probabilistici atti a descrivere fenomeni in cui lo stato del sistema è caratterizzato da un numero intero di individui e in cui i cambiamenti di stato rappresentano l'addizione o la sottrazione di individui dal sistema in vari modi: nascite, morti, immigrazioni, emigrazioni, arrivi e partenze degli utenti da un sistema di servizio,. . ..

3.3 Processi stocastici di nascita morte

I più noti processi stocastici discreti nello spazio degli stati e continui nel tempo sono i processi di nascita–morte, introdotti da Feller nel 1939. Essi sono utilizzati per costruire modelli di crescita di popolazione, di sistemi di servizio, di epidemiologia e di molte aree di interesse sia teorico che applicativo.

Con riferimento ai sistemi di servizio, è spesso ragionevole supporre che sia il parametro di arrivo dell'utente che entra nel sistema sia il parametro di partenza dell'utente che esce dal sistema (essendo stato servito) dipendano dal numero degli utenti presenti nel sistema. Spesso si suppone che il tempo T_n che intercorre tra l'arrivo $(n-1)$–esimo e l'arrivo n–esimo $(n = 0, 1, \ldots)$ sia distribuito esponenzialmente con valore medio $1/\lambda_n$ e che il tempo S_n occorrente per servire l'utente n–esimo $(n = 1, 2, \ldots)$ sia anche distribuito esponenzialmente con valore medio $1/\mu_n$; inoltre, spesso si assume che sia i tempi di interarrivo $T_1, T_2, \ldots$ sia i tempi di servizio $S_1, S_2, \ldots$ siano indipendenti tra loro. Gli insiemi $\{\lambda_n,\, n = 0, 1, \ldots\}$ e $\{\mu_n,\, n = 1, 2, \ldots\}$ contengono rispettivamente i parametri di arrivo (di nascita) e i parametri di partenza (di morte).

Definiamo ora un processo nascita-morte facendo riferimento principalmente alla teoria delle delle file di attesa.

Definizione 3.2 *Sia* $\{N(t),\, t \geq 0\}$ *un processo stocastico avente spazio degli stati* $0, 1, 2, \ldots$. *Supponiamo che questo processo descriva un sistema che si trova*

nello stato E_n al tempo t se e solo se $N(t) = n$, ossia se il numero di utenti (individui) presenti al tempo t è n. Tale processo stocastico è detto di nascita–morte se esistono dei parametri di arrivo (di nascita) $\{\lambda_n,\, n = 0, 1, \ldots\}$ e di partenza (di morte) $\{\mu_n,\, n = 1, 2, \ldots\}$ tali da soddisfare i seguenti postulati:

(i) In un piccolo intervallo di tempo Δt si possono avere cambiamenti di stato soltanto dallo stato E_n allo stato E_{n+1} oppure dallo stato E_n allo stato E_{n-1} se $n \geq 1$, mentre se $n = 0$ si può avere un cambiamento di stato soltanto dallo stato E_0 allo stato E_1.

(ii) Se al tempo t il sistema è nello stato E_n, la probabilità che nell'intervallo di tempo $(t, t+\Delta t)$ avvenga una transizione dallo stato E_n allo stato E_{n+1} è $\lambda_n \Delta t + o(\Delta t)$, mentre la probabilità che nell'intervallo di tempo $(t, t+\Delta t)$ avvenga una transizione dallo stato E_n allo stato E_{n-1} è $\mu_n \Delta t + o(\Delta t)$.

(iii) Se al tempo t il sistema è nello stato E_n, la probabilità che nell'intervallo di tempo $(t, t + \Delta t)$ avvenga più di una transizione è $o(\Delta t)$.

Il postulato *(i)* mostra che se nel sistema è presente almeno un utente in un piccolo intervallo di tempo può verificarsi al più una transizione (un arrivo oppure una partenza), mentre se il sistema è vuoto non vi possono essere uscite dal sistema. Il postulato *(ii)* fornisce le probabilità di transizione, ossia le probabilità di arrivo o di partenza in un piccolo intervallo di tempo $(t, t + \Delta t)$ quando il numero degli utenti nel sistema al tempo t è n. L'ultimo postulato mostra che la probabilità che in un piccolo intervallo di tempo si verifichi più di una transizione è trascurabile.

Se indichiamo con $N(t)$ $(t \geq 0)$ il numero degli utenti presenti nel sistema al tempo t e con $N(t, t + \Delta t)$ il numero degli utenti che arrivano nell'intervallo $(t, t + \Delta t]$, le ipotesi precedenti possono essere così formulate:

$$
\begin{aligned}
&P\big\{N(t, t+\Delta t) = 1 | N(t) = n\big\} = \lambda_n\, \Delta t + o(\Delta t) \quad (n = 0, 1, \ldots),\\
&P\big\{N(t, t+\Delta t) = -1 | N(t) = n\big\} = \mu_n\, \Delta t + o(\Delta t) \quad (n = 1, 2, \ldots),\\
&P\big\{N(t, t+\Delta t) = 0 | N(t) = n\big\} = \begin{cases} 1 - \lambda_0\, \Delta t + o(\Delta t), & n = 0\\ 1 - (\lambda_n + \mu_n)\, \Delta t + o(\Delta t), & n = 1, 2, \ldots, \end{cases}
\end{aligned}
\tag{3.10}
$$

dove $o(\Delta t)$ è un infinitesimo di ordine superiore rispetto a Δt, $\lambda_0, \lambda_1, \ldots$ denotano i parametri di arrivo e $\mu_1, \mu_2, \ldots$ denotano i parametri di partenza. Tali parametri di arrivo e di partenza hanno dimensioni fisiche $[tempo]^{-1}$. Dalle (3.10) segue immediatamente che la probabilità che in un piccolo intervallo di tempo si verifichi più di una transizione è trascurabile. Denotiamo con

$$p_n(t) = P\big\{N(t) = n\big\} \qquad (n = 0, 1, \ldots),$$

la probabilità che sia n il numero degli utenti presenti nel sistema al tempo t. Come si evince dalla Figura 3.3, si ha:

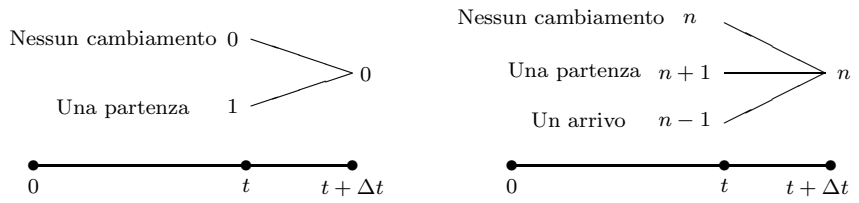

Figura 3.3: Cambiamenti di stato in $(t, t+\Delta t)$ nel processo di nascita morte.

$$p_0(t+\Delta t) = P\big\{N(t+\Delta t)=0\big\} = p_0(t)\,[1-\lambda_0\,\Delta t] + p_1(t)\,\mu_1\,\Delta t + o(\Delta t),$$
$$p_n(t+\Delta t) = P\big\{N(t+\Delta t)=n\big\} = p_{n-1}(t)\,\lambda_{n-1}\,\Delta\, t + p_n(t)\,[1-(\lambda_n+\mu_n)\,\Delta t] + p_{n+1}(t)\,\mu_{n+1}\,\Delta t + o(\Delta t) \qquad (n=1,2,\ldots),$$

ossia

$$\frac{p_0(t+\Delta t)-p_0(t)}{\Delta t} = -\lambda_0\,p_0(t) + \mu_1\,p_1(t) + \frac{o(\Delta t)}{\Delta t},$$
$$\frac{p_n(t+\Delta t)-p_n(t)}{\Delta t} = \lambda_{n-1}\,p_{n-1}(t) - (\lambda_n+\mu_n)\,p_n(t) + \mu_{n+1}\,p_{n+1}(t) + \frac{o(\Delta t)}{\Delta t} \qquad (n=1,2,\ldots).$$

Procedendo al limite quando $\Delta t \to 0$ si ottiene il seguente sistema di equazioni differenziali e alle differenze del secondo ordine in n:

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\lambda_0\,p_0(t) + \mu_1\,p_1(t), \\ \frac{dp_n(t)}{dt} &= \lambda_{n-1}\,p_{n-1}(t) - (\lambda_n+\mu_n)\,p_n(t) + \mu_{n+1}\,p_{n+1}(t) \qquad (n=1,2,\ldots). \end{aligned} \tag{3.11}$$

Se supponiamo che inizialmente nel sistema di servizio siano presenti i utenti, ossia $P\{N(0)=i\}=1$, occorre risolvere il sistema (3.11) con le condizioni iniziali:

$$p_n(0) = \begin{cases} 1, & n=i \\ 0, & n\neq i. \end{cases} \tag{3.12}$$

Esempio 3.2 Supponiamo che

$$\begin{aligned} \lambda_n &= \varrho \qquad n=0,1,2,\ldots \\ \mu_n &= 0 \qquad n=1,2,\ldots \end{aligned}$$

e inoltre

$$p_n(0) = \begin{cases} 1, & n=0 \\ 0, & n=1,2,\ldots. \end{cases}$$

In questo caso il processo di nascita morte si riduce a un processo di Poisson di parametro ϱ. Abbiamo precedentemente mostrato che $\{p_0(t), p_1(t), \ldots\}$ è una distribuzione di Poisson di parametro $\varrho\, t$ per ogni $t \geq 0$. Il processo di Poisson può essere allora visto come un processo di pura nascita. ♢

In un processo stocastico di nascita–morte per determinare le probabilità di avere n utenti ($n = 0, 1, \ldots$) nel sistema al tempo t ($t > 0$) occorre quindi risolvere il sistema di equazioni differenziali (3.11) con le condizioni iniziali (3.12). Si può dimostrare che in condizioni abbastanza generali questo sistema di equazioni ammette un'unica soluzione. Comunque, eccetto in casi particolarmente semplici, la distribuzione $\{p_0(t), p_1(t), \ldots\}$ è difficile da determinare teoricamente risolvendo il sistema (3.11). Risulta anche notevolmente complesso calcolare la soluzione del sistema (3.11) ricorrendo a metodi numerici. Nella maggior parte dei casi per ottenere informazioni sul comportamento del processo di nascita–morte nel transiente occorre ricorrere a metodi di simulazione.

3.4 Equilibrio statistico

La conoscenza delle probabilità $\{p_0(t), p_1(t), \ldots\}$ di un processo di nascita-morte permette di descrivere il comportamento del sistema. Come già precedentemente sottolineato è in generale molto difficile calcolare la distribuzione di probabilità nel transiente, ossia per ogni fissato istante di tempo t. Vogliamo quindi determinare delle condizioni sui parametri di arrivo $\{\lambda_n,\ n = 0, 1, \ldots\}$ e di partenza $\{\mu_n,\ n = 1, 2, \ldots\}$ che conducano ad una situazione di equilibrio statistico del sistema. A tal fine denotiamo con

$$q_n = \lim_{t \to +\infty} p_n(t) \qquad (n = 0, 1, \ldots) \tag{3.13}$$

la probabilità di avere n utenti nel sistema nella situazione di equilibrio statistico.

Se i limiti nella (3.13) *esistono e non dipendono dalle condizioni iniziali* per ogni $n = 0, 1, \ldots$, diremo che *il sistema raggiunge una situazione di equilibrio statistico* descritta dalla distribuzione di equilibrio $\{q_0, q_1, \ldots\}$.

Osserviamo che se i limiti in (3.13) esistono, allora al crescere del tempo $dp_n(t)/dt$ tende a zero per ogni $n = 0, 1, \ldots$ Il sistema (3.11) quindi diventa:

$$\begin{aligned} &-\lambda_0\ q_0 + \mu_1\ q_1 = 0 \\ &\lambda_{n-1}\ q_{n-1} - (\lambda_n + \mu_n)\ q_n + \mu_{n+1}\ q_{n+1} = 0 \qquad (n = 1, 2, \ldots). \end{aligned} \tag{3.14}$$

Le equazioni alle differenze (3.14) possono essere ricavate in modo alternativo costruendo il *grafo di transizione*, illustrato in Figura 3.4. In questo grafo ogni stato E_n è rappresentato da un cerchietto (nodo) etichettato con il numero n. Gli archi che collegano i nodi mostrano quali sono le possibili transizioni di stato e sono etichettati con i parametri di transizione (ossia i parametri di arrivo o di servizio).

Se un processo nascita–morte ha raggiunto la situazione di equilibrio statistico, allora per ogni stato E_n del sistema ($n = 0, 1, \ldots$) assumiamo che valga il *principio di bilanciamento* che afferma: "*il flusso medio che entra nel nodo n deve uguagliare il flusso medio che esce da tale nodo*". Le equazioni che esprimono tale principio sono dette *equazioni di bilanciamento*.

Ricaviamo ora il sistema di equazioni alle differenze (3.14) utilizzando il principio di bilanciamento. Nella situazione di equilibrio statistico il grafo di transizione di un processo di nascita–morte è il seguente: Per il nodo E_0 si ha

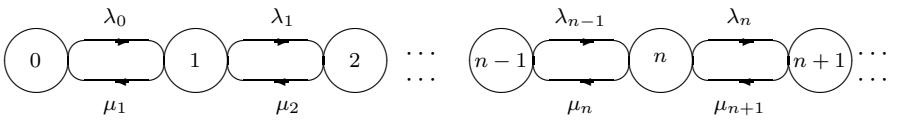

Figura 3.4: Grafo di transizione di un processo di nascita–morte in condizioni di equilibrio statistico.

che il flusso medio entrante è $\mu_1\, q_1$ e il flusso medio uscente è $\lambda_0\, q_0$; pertanto per il principio di bilanciamento si deve avere

$$\mu_1\, q_1 = \lambda_0\, q_0. \tag{3.15}$$

Invece, per un generico nodo $E_n \quad (n = 1, 2, \ldots)$ si ha che il flusso medio entrante è $\lambda_{n-1}\, q_{n-1} + \mu_{n+1}\, q_{n+1}$ e il flusso medio uscente è $\lambda_n\, q_n + \mu_n\, q_n$; pertanto per il principio di bilanciamento si deve avere

$$\lambda_{n-1}\, q_{n-1} + \mu_{n+1}\, q_{n+1} = \lambda_n\, q_n + \mu_n\, q_n \qquad (n = 1, 2, \ldots). \tag{3.16}$$

Le equazioni (3.15) e (3.16) corrispondono a quelle del sistema (3.14). Ci proponiamo ora di determinare la soluzione del sistema (3.14).

Proposizione 3.2 *Un processo nascita–morte ammette una distribuzione di equilibrio statistico* $\{q_0, q_1, \ldots\}$ *se e solo se*

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} < +\infty \tag{3.17}$$

e si ha

$$\begin{aligned} q_0 &= \left[1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}\right]^{-1}, \\ q_n &= q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \qquad (n = 1, 2, \ldots). \end{aligned} \tag{3.18}$$

Dimostrazione Se si pone

$$g_n = \lambda_{n-1}\, q_{n-1} - \mu_n\, q_n, \qquad (n = 1, 2, \ldots), \tag{3.19}$$

la seconda delle (3.14) si può così scrivere:

$$g_{n+1} = g_n \qquad (n = 2, 3, \ldots),$$

ossia un'equazione alle differenze lineare del primo ordine la cui soluzione è una costante reale c. Si deve quindi avere

$$g_n = \lambda_{n-1}\, q_{n-1} - \mu_n\, q_n = c \qquad (n = 2, 3, \ldots). \tag{3.20}$$

Imponendo che la (3.20) sia soddisfatta anche per $n = 1$ si ottiene

$$\lambda_0\, q_0 - \mu_1\, q_1 = c,$$

da cui, per la prima delle (3.14), si ha $c = 0$. Ponendo nella (3.20) $c = 0$ si ottiene:

$$q_n = \frac{\lambda_{n-1}}{\mu_n}\, q_{n-1} = \frac{\lambda_{n-1}\lambda_{n-2}}{\mu_n\mu_{n-1}}\, q_{n-2} = \ldots\ldots = \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}\, q_0. \tag{3.21}$$

La probabilità q_0 può essere determinata imponendo che l'insieme $\{q_0, q_1, \ldots\}$ sia una distribuzione di probabilità, ossia

$$q_n \geq 0 \qquad (n = 0, 1, \ldots), \qquad\qquad \sum_{n=0}^{+\infty} q_n = 1. \tag{3.22}$$

Facendo uso di (3.21) nella seconda delle (3.22) si ricava:

$$1 = \sum_{n=0}^{+\infty} q_n = q_0 + q_0 \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n} = q_0\left[1 + \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}\right],$$

ossia

$$q_0 = \left[1 + \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}\right]^{-1}.$$

Si nota che affinché il processo nascita–morte ammetta una distribuzione di equilibrio statistico $\{q_0, q_1, \ldots\}$ è necessario che la serie

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}$$

converga. In un processo nascita–morte si può dimostrare che tale condizione è anche sufficiente. □

Se la serie (3.17) diverge, ciò indica che il sistema di servizio è *instabile* nel senso che in media gli arrivi si verificano più frequentemente delle partenze. Pertanto la (3.17) è chiamata "condizione di equilibrio" o "condizione di stabilità" di un processo di nascita morte.

In un processo di nascita morte in condizioni di equilibrio statistico la frequenza media di arrivo per unità di tempo può essere così definita:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n\, q_n, \tag{3.23}$$

ossia è la media pesata delle frequenze di arrivo λ_n $(n = 0, 1, \ldots)$. Inoltre, la frequenza media di partenza per unità di tempo per ognuno dei servitori è data da:

$$\mu^* = \frac{1}{1 - q_0} \sum_{n=1}^{+\infty} \mu_n \, q_n = \sum_{n=1}^{+\infty} \mu_n \, \frac{q_n}{1 - q_0} \,, \tag{3.24}$$

ossia è la media pesata delle frequenze di partenza μ_n $(n = 1, 2, \ldots)$. Quindi, λ^* e μ^* si possono rispettivamente interpretare come i valori medi dei parametri di arrivo $\lambda_0, \lambda_1, \ldots$ e di partenza $\mu_1, \mu_2, \ldots$, ottenuti utilizzando la distribuzione di equilibrio $\{q_0, q_1, \ldots\}$.

L'intensità di traffico, ossia l'intensità di lavoro che svolge il sistema di servizio nella situazione di equilibrio statistico, è $a = \lambda^*/\mu^*$. Se si denota con s il numero di servitori presenti nel sistema di servizio, il fattore di utilizzazione del sistema, ossia l'intensità di lavoro per servitore nella situazione di equilibrio statistico, è quindi $\varrho^* = a/s = \lambda^*/(s\,\mu^*)$.

Capitolo 4

Modelli con singolo servitore

4.1 Introduzione

In questo capitolo analizzeremo i principali sistemi di servizio con singolo servitore, ossia i sistemi $M/M/1$, $M/M/1/K$ e $M/G/1$ e confronteremo il sistema $M/M/1$ con un sistema di servizio adattivo di interesse nella teoria delle file di attesa. Lo scopo è quello di individuare i principali parametri prestazionali dei vari sistemi e di analizzarne il comportamento in situazioni di equilibrio.

4.2 Sistema di servizio $M/M/1$

Supponiamo che gli utenti arrivino ad un sistema di servizio secondo un processo di Poisson di parametro λ. I tempi tra due successivi arrivi (tempi di interarrivo) sono quindi indipendenti e esponenzialmente distribuiti con valore medio $1/\lambda$. Il sistema ha un unico servitore e utilizza la disciplina di servizio FIFO. La capacità del sistema è infinita. Dopo l'arrivo, ogni utente è immediatamente servito se il servitore è libero, mentre se il servitore è occupato l'utente si mette in fila di attesa. Quando il servitore termina di servire un utente, si ha una partenza dal sistema e un nuovo utente nella fila di attesa, se ne esiste almeno uno presente, può accedere al servizio. Supponiamo che i tempi di servizio degli utenti siano indipendenti e esponenzialmente distribuiti con valore medio $1/\mu$.

Tale sistema, illustrato in Figura 4.1, è noto in letteratura come *sistema di servizio* $M/M/1$ (o equivalentemente come *sistema di servizio* $M/M/1/\infty$*)*. La prima M significa che i tempi di interarrivo sono indipendenti e distribuiti esponenzialmente (proprietà di Markov legata alla mancanza di memoria della funzione di distribuzione esponenziale); la seconda M significa che i tempi di

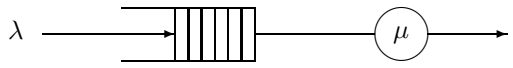

Figura 4.1: Sistema di servizio $M/M/1$.

servizio sono indipendenti e distribuiti esponenzialmente; il simbolo 1 si riferisce all'unico servitore disponibile, il simbolo ∞ indica che la capacità del sistema è illimitata.

Sia $N(t)$ il numero di utenti presenti nel sistema al tempo t. Il processo stocastico $\{N(t),\, t \geq 0\}$ può essere descritto facendo ricorso ad un processo di nascita–morte caratterizzato da parametri

$$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad\qquad \mu_n = \mu \quad (n = 1, 2, \ldots). \tag{4.1}$$

Vogliamo ora vedere sotto quali condizioni il sistema $M/M/1$ raggiunge una situazione di equilibrio statistico. Facendo uso di (4.1) in (3.17) e ponendo $\varrho = \lambda/\mu$, si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \sum_{n=0}^{+\infty} \Big(\frac{\lambda}{\mu}\Big)^n = \sum_{n=0}^{+\infty} \varrho^n,$$

ossia una serie geometrica che converge a $(1-\varrho)^{-1}$ se e solo se la ragione $\varrho = \lambda/\mu < 1$. Quindi, il sistema di servizio $M/M/1$ raggiunge una situazione di equilibrio statistico se e solo se $\varrho = \lambda/\mu < 1$ e si ha:

$$\begin{aligned} q_0 &= P(N = 0) = 1 - \varrho, \\ q_n &= P(N = n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = (1 - \varrho)\, \varrho^n \qquad (n = 1, 2, \ldots). \end{aligned}$$

Proposizione 4.1 *Il sistema di servizio $M/M/1$ raggiunge una situazione di equilibrio statistico se e solo se $\varrho = \lambda/\mu < 1$ e si ha:*

$$q_n = P(N = n) = (1 - \varrho)\, \varrho^n \qquad (n = 0, 1, \ldots), \tag{4.2}$$

ossia una funzione di probabilità geometrica di parametro ϱ.

Nella situazione di equilibrio statistico il valore medio, il momento del secondo ordine e la varianza del numero di utenti presenti nel sistema sono quindi:

$$\begin{aligned} E(N) &= \sum_{n=0}^{\infty} n\, q_n = (1 - \varrho) \sum_{n=1}^{\infty} n\, \varrho^n = \frac{\varrho}{1 - \varrho}, \\ E(N^2) &= \sum_{n=0}^{\infty} n^2\, q_n = (1 - \varrho) \sum_{n=1}^{\infty} n^2\, \varrho^n = \frac{\varrho\,(1 + \varrho)}{(1 - \varrho)^2}, \\ \mathrm{Var}(N) &= E(N^2) - [E(N)]^2 = \frac{\varrho}{(1 - \varrho)^2}, \end{aligned} \tag{4.3}$$

dove si è fatto uso delle seguenti identità:

$$\sum_{n=1}^{\infty} n\, z^n = \frac{z}{(1-z)^2}, \qquad \sum_{n=1}^{\infty} n^2\, z^n = \frac{z\,(1+z)}{(1-z)^3}\,. \qquad (\mid z \mid < 1)$$

Se $\varrho \geq 1$ il sistema di servizio è *instabile* e il numero di utenti nel sistema è destinato a crescere indefinitamente.

Se denotiamo con T la variabile aleatoria che descrive un generico tempo di interarrivo e con S la variabile aleatoria che descrive un generico tempo di servizio, per le ipotesi fatte sul sistema di servizio $M/M/1$ si ha:

$$E(T) = \frac{1}{\lambda}, \qquad E(S) = \frac{1}{\mu}, \qquad \varrho = \frac{E(S)}{E(T)} = \frac{\lambda}{\mu}\,.$$

La condizione $\varrho < 1$ è quindi equivalente a richiedere che $E(S) < E(T)$, ossia il sistema raggiunge una situazione di *equilibrio statistico* se e solo se *il tempo medio di servizio è minore del tempo medio di interarrivo.*

Nella Tabella 4.1 sono indicati il valore medio, la varianza e il coefficiente di variazione del numero di utenti presenti nel sistema nella situazione di equilibrio statistico per alcune scelte dell'intensità di traffico ϱ. Si nota che il valore medio e la varianza del numero di utenti sono funzioni crescenti in ϱ e tendono all'infinito quando $\varrho \to 1$.

ϱ	$E(N)$	$\mathrm{Var}(N)$	$C(N)$
0.1	0.11111	0.12346	3.1623
0.2	0.25000	0.31250	2.2361
0.3	0.42857	0.61224	1.8257
0.4	0.66667	1.1111	1.5811
0.5	1.0000	2.0000	1.4142
0.6	1.5000	3.7500	1.2910
0.7	2.3333	7.7778	1.1952
0.8	4.0000	20.000	1.1180
0.9	9.0000	90.000	1.0541
0.99	99.000	9900.0	1.0050
0.999	999.01	9.9903×10^5	1.0005

Tabella 4.1: Valore medio, varianza e coefficiente di variazione del numero di utenti nel sistema $M/M/1$ in condizioni di equilibrio per alcune scelte di ϱ $(0 < \varrho < 1)$.

Facendo uso di (3.23) e (3.24), la frequenza media di arrivo e la frequenza media di partenza per unità di tempo nel sistema $M/M/1$ sono:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n\, q_n = \lambda, \qquad \mu^* = \frac{1}{1-q_0} \sum_{n=1}^{+\infty} \mu_n\, q_n = \mu.$$

Il *fattore di utilizzazione* del sistema $M/M/1$ (che coincide con l'intensità di traffico) è quindi $\varrho^* = \lambda/\mu = \varrho$ e fornisce l'intensità di lavoro svolta dal sistema di servizio nella situazione di equilibrio. Infatti, il fattore di utilizzazione

del sistema $M/M/1$ coincide con *la probabilità di avere almeno un utente nel sistema*, ossia con

$$P(N \geq 1) = 1 - q_0 = \varrho.$$

Inoltre, se $\varrho < 1$, dalla (4.2) segue che la probabilità che in condizioni di equilibrio siano presenti nel sistema un numero maggiore o uguale a k di utenti è:

$$P(N \geq k) = \sum_{n=k}^{+\infty} q_k = (1-\varrho) \sum_{n=k}^{+\infty} \varrho^n = (1-\varrho)\,\varrho^k \sum_{n=k}^{+\infty} \varrho^{n-k} = \varrho^k.$$

Dalla prima legge di Little ricaviamo che nella situazione di equilibrio statistico il tempo medio di attesa di un utente nel sistema è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{\varrho}{\lambda\,(1-\varrho)} = \frac{1}{\mu\,(1-\varrho)} = \frac{1}{\mu - \lambda}.$$

Il tempo medio di permanenza di un utente nella fila di attesa è:

$$E(Q) = E(W) - E(S) = \frac{1}{\mu\,(1-\varrho)} - \frac{1}{\mu} = \frac{\varrho}{\mu\,(1-\varrho)} = \frac{\varrho}{\mu - \lambda}\,.$$

Dalla seconda legge di Little segue che nella situazione di equilibrio statistico il numero medio di utenti nella fila di attesa è:

$$E(N_q) = \lambda^*\,E(Q) = \lambda\,\frac{\varrho}{\mu\,(1-\varrho)} = \frac{\varrho^2}{1-\varrho}.$$

Si noti che $E(N_q)$ si può anche valutare nel seguente modo:

$$E(N_q) = \sum_{n=1}^{\infty} (n-1)\,q_n = E(N) - (1-q_0) = \frac{\varrho}{1-\varrho} - \varrho = \frac{\varrho^2}{1-\varrho}.$$

La terza legge di Little mostra infine che il numero medio di utenti in servizio

$$E(N_s) = \lambda^*\,E(S) = \frac{\lambda}{\mu} = \varrho$$

è uguale all'intensità di traffico del centro di servizio, che coincide anche con il *fattore di utilizzazione del sistema*.

Nel sistema $M/M/1$ i periodi di ozio del servitore, che possono essere riguardati come tempi residui dei tempi di interarrivo, sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$, ossia con densità

$$f_I(t) = \begin{cases} \lambda\,e^{-\lambda\,t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Ricordando la (2.25), risulta che

$$E(B) = \frac{(1-q_0)\,E(I)}{q_0} = \frac{\varrho}{\lambda\,(1-\varrho)} = \frac{1}{\mu - \lambda}$$

Nella situazione di equilibrio statistico, per il sistema $M/M/1$ risulta che il tempo medio di occupazione del servitore coincide con il tempo medio di attesa nel sistema, ossia $E(B) = E(W)$.

Nella Tabella 4.2 sono riportati i principali parametri prestazionali del sistema $M/M/1$.

$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad \mu_n = \mu \quad (n = 1, 2, \ldots)$
$\varrho = \lambda/\mu < 1 \qquad$ (condizione di equilibrio statistico)
$q_n = P(N = n) = (1 - \varrho)\,\varrho^n \qquad (n = 0, 1, \ldots)$
$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad \varrho^* = \dfrac{\lambda}{\mu} = 1 - q_0 = P(N \geq 1)$
$E(N) = \dfrac{\varrho}{1-\varrho}, \qquad E(W) = \dfrac{1}{\mu - \lambda}$
$E(N_q) = \dfrac{\varrho^2}{1-\varrho}, \qquad E(Q) = \dfrac{\varrho}{\mu - \lambda}$
$E(N_s) = \dfrac{\lambda}{\mu} = \varrho, \qquad E(S) = \dfrac{1}{\mu}$
$E(I) = \dfrac{1}{\lambda}, \qquad E(B) = \dfrac{1}{\mu - \lambda}$

Tabella 4.2: Parametri prestazionali del sistema di servizio $M/M/1$.

Esempio 4.1 Supponiamo di considerare una linea di comunicazione che può trasmettere con una frequenza media di 2000 bits al secondo. Tale linea è utilizzata per trasmettere messaggi di 8 bits. Assumiamo che la frequenza media di arrivo è di 12000 messaggi al minuto. Se un sistema di servizio $M/M/1$ modella la linea di comunicazione considerata, dire se il sistema raggiunge una situazione di equilibrio e calcolare i principali parametri prestazionali del sistema.

In questo caso risulta

$$\lambda = \frac{12000}{60} = 200 \text{ messaggi al secondo},$$
$$\mu = \frac{2000}{8} = 250 \text{ messaggi al secondo},$$

e quindi il fattore di utilizzazione del sistema è:

$$\varrho = \frac{\lambda}{\mu} = \frac{200}{250} = \frac{4}{5} = 0.8 < 1.$$

In condizioni di equilibrio statistico il numero medio di messaggi nel sistema e in attesa di essere trasmessi sono:

$$E(N) = \frac{\varrho}{1-\varrho} = 4 \text{ messaggi}, \qquad E(N_q) = \frac{\varrho^2}{1-\varrho} = \frac{16}{5} = 3.2 \text{ messaggi}.$$

Inoltre, il tempo medio di attesa nel sistema e nella fila di attesa sono:

$$E(W) = \frac{E(N)}{\lambda} = \frac{4}{200} = 0,02 \text{ secondi}, \quad E(Q) = \frac{E(N_q)}{\lambda} = \frac{3.2}{200} = 0,016 \text{ secondi},$$

e il numero medio di messaggi in trasmissione e il tempo medio di trasmissione sono:

$$E(N_s) = \varrho = 0.8 \text{ messaggi}, \qquad E(S) = \frac{1}{\mu} = 0.004.$$

La probabilità che nel sistema siano presenti un numero maggiore o uguale a k messaggi è $P(N \geq k) = \varrho^k$, da cui segue che

$$P(N \geq 1) = \varrho = 0.8, \qquad P(N \geq 2) = \varrho^2 = 0.64, \qquad P(N \geq 3) = \varrho^3 = 0.512,$$
$$P(N \geq 4) = \varrho^4 = 0.4096, \qquad P(N \geq 5) = \varrho^5 = 0.32768, \ldots\ldots$$

$\diamondsuit$

4.3 Sistema di servizio con svendita

Consideriamo un sistema di servizio a capacità infinita con unica fila di attesa e singolo servitore in cui gli utenti sono attratti da una lunga coda e il servitore accelera il suo servizio all'aumentare della lunghezza della coda.

Un sistema di servizio di questo tipo può essere descritto mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri:

$$\begin{aligned} \lambda_n &= \lambda\,(n+1) \qquad (n = 0, 1, \ldots) \\ \mu_n &= \mu\, n \qquad (n = 1, 2, \ldots). \end{aligned} \tag{4.4}$$

Tale modello si presta a descrivere alcune situazioni reali quali una grossa svendita in cui gli utenti sono attratti dalla lunga coda e il servitore cerca di accelerare il servizio per vendere la maggior parte della merce in deposito. Vogliamo ora vedere in quali condizioni tale sistema raggiunge una situazione di equilibrio statistico. Facendo uso di (4.4) in (3.17) e denotando con $\varrho = \lambda/\mu$, si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{+\infty} \frac{\lambda\, 2\lambda \cdots n\,\lambda}{\mu\, 2\,\mu \cdots n\,\mu} = \sum_{n=0}^{+\infty} \Big(\frac{\lambda}{\mu}\Big)^n = \sum_{n=0}^{+\infty} \varrho^n,$$

ossia una serie geometrica che converge se e solo se la ragione $\varrho = \lambda/\mu < 1$, fornendo come somma $(1-\varrho)^{-1}$.

Proposizione 4.2 *Il sistema di servizio con svendita raggiunge una situazione di equilibrio statistico se e solo se* $\varrho = \lambda/\mu < 1$ *e si ha:*

$$q_n = P(N = n) = (1 - \varrho)\,\varrho^n \qquad (n = 0, 1, \ldots).$$

Si nota che il sistema di servizio con svendita ammette la stessa distribuzione di equilibrio del sistema $M/M/1$, ossia una distribuzione geometrica di parametro ϱ. Nella situazione di equilibrio statistico, il valore medio e la varianza del numero di utenti presenti nel sistema sono quindi forniti in (4.3), ossia sono gli stessi del sistema $M/M/1$.

I modelli $M/M/1$ e quello con svendita, anche se caratterizzati dalla stessa distribuzione di equilibrio, sono fondamentalmente diversi nella fase transiente e inoltre hanno alcuni parametri prestazionali differenti. Infatti, nel sistema di servizio con svendita le frequenza medie di arrivo e di partenza per unità di tempo sono differenti da quelle del sistema $M/M/1$ e risulta:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n\, q_n = \lambda \sum_{n=0}^{+\infty} (n+1)\, q_n = \lambda\,[E(N) + 1] = \lambda \left(\frac{\varrho}{1-\varrho} + 1\right) = \frac{\lambda}{1-\varrho}$$

$$\mu^* = \frac{1}{1-q_0} \sum_{n=1}^{+\infty} \mu_n\, q_n = \frac{\mu}{1-q_0} \sum_{n=1}^{+\infty} n\, q_n = \frac{\mu}{1-q_0}\, E(N) = \frac{\mu}{\varrho}\, \frac{\varrho}{1-\varrho} = \frac{\mu}{1-\varrho},$$

e quindi il fattore di utilizzazione del sistema (che coincide con l'intensità di traffico) è $\varrho^* = \lambda/\mu$. Si nota che $\varrho^* = P(N \geq 1) = 1 - q_0 = \varrho$.

Dalla prima legge di Little ricaviamo che nella situazione di equilibrio statistico il tempo medio di attesa di un utente nel sistema è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{1-\varrho}{\lambda}\, \frac{\varrho}{1-\varrho} = \frac{1}{\mu}\,.$$

Il numero medio di utenti in fila di attesa può essere così ottenuto:

$$E(N_q) = \sum_{n=1}^{+\infty} (n-1)\, q_n = E(N) - (1 - q_0) = \frac{\varrho^2}{1-\varrho}\,,$$

che coincide con quello del sistema di servizio $M/M/1$. Dalla seconda legge di Little si ottiene:

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{\varrho^2}{1-\varrho}\, \frac{1-\varrho}{\lambda} = \frac{\varrho}{\mu}.$$

Il numero medio di utenti in servizio è

$$E(N_s) = E(N) - E(N_q) = \frac{\varrho}{1-\varrho} - \frac{\varrho^2}{1-\varrho} = \varrho,$$

che coincide con la probabilità che nel sistema sia presente almeno un utente. Dalla terza legge di Little si ricava quindi

$$E(S) = \frac{E(N_s)}{\lambda^*} = \varrho\, \frac{1-\varrho}{\lambda} = \frac{1-\varrho}{\mu},$$

$$\lambda_n = \lambda\,(n+1) \quad (n = 0, 1, \ldots), \qquad \mu_n = \mu\, n \quad (n = 1, 2, \ldots)$$

$$\varrho = \lambda/\mu < 1 \qquad \text{(condizione di equilibrio statistico)}$$

$$q_n = P(N = n) = (1 - \varrho)\,\varrho^n \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \frac{\lambda}{1-\varrho}, \qquad \mu^* = \frac{\mu}{1-\varrho}, \qquad \varrho^* = \frac{\lambda}{\mu} = 1 - q_0 = P(N \geq 1)$$

$$E(N) = \frac{\varrho}{1-\varrho}\,, \qquad E(W) = \frac{1}{\mu}$$

$$E(N_q) = \frac{\varrho^2}{1-\varrho}\,, \qquad E(Q) = \frac{\varrho}{\mu}$$

$$E(N_s) = \frac{\lambda}{\mu} = \varrho, \qquad E(S) = \frac{1-\varrho}{\mu}$$

Tabella 4.3: Parametri prestazionali del sistema di servizio con svendita e unico servitore

ossia $E(S) = 1/\mu^*$. Nella Tabella 4.3 sono riportati i principali parametri prestazionali del sistema di servizio con svendita e unico servitore.

Come si evince dalle Tabelle 4.2 e 4.3, entrambi i sistemi di servizio $M/M/1$ e con svendita sono caratterizzati dalla stessa distribuzione di equilibrio, dallo stesso numero medio di utenti nel sistema, nella fila di attesa e nel centro di servizio e dalla stessa intensità di traffico; risultano invece differenti la frequenza media di arrivo e di partenza per unità di tempo, i tempi medi di attesa degli utenti nel sistema, i tempi medi di permanenza nella fila di attesa e i tempi medi di servizio.

4.4 Sistema di servizio $M/M/1/1$

Il sistema $M/M/1/1$ descrive un centralino telefonico con un'unica linea disponibile, con fila di attesa ha capacità nulla in cui le chiamate che arrivano e trovano il centralino occupato sono perse. Supponiamo che gli utenti arrivino al sistema di servizio secondo un processo di Poisson di parametro λ. Se nel sistema è già presente un utente (in servizio) il nuovo utente in arrivo non può accedere al sistema. Quando il servitore termina di servire un utente, tale utente lascia il sistema e il nuovo utente in arrivo può usufruire del servizio. Supponiamo che i tempi di servizio degli utenti siano indipendenti e esponenzialmente distribuiti con valore medio $1/\mu$. Tale sistema, illustrato in Figura 4.2, è noto in letteratura come *sistema di servizio* $M/M/1/1$.

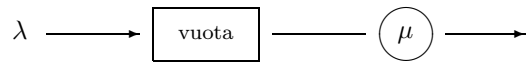

Figura 4.2: Sistema di servizio $M/M/1/1$.

La prima M significa che i tempi di interarrivo sono distribuiti esponenzialmente; la seconda M significa che i tempi di servizio sono anche distribuiti esponenzialmente; il simbolo 1 si riferisce all'unico servitore e l'ultimo simbolo indica che la capacità del sistema è unitaria. Nel sistema $M/M/1/1$ sono nulli sia il numero medio di utenti nella fila di attesa sia il tempo medio di permanenza nella fila di attesa. Inoltre, il tempo medio di attesa di un utente coincide con il tempo medio di servizio, ossia con $E(W) = E(S) = 1/\mu$ ed il numero medio di utenti nel sistema $E(N) = E(N_s)$.

Denotiamo con $N(t)$ il numero di utenti presenti nel sistema $M/M/1/1$ al tempo t. Il processo stocastico $\{N(t),\, t \geq 0\}$ può essere descritto mediante un processo di nascita–morte caratterizzato da parametri:

$$\lambda_n = \begin{cases} \lambda, & n = 0 \\ 0, & n = 1, 2, \ldots \end{cases} \qquad \mu_n = \begin{cases} \mu, & n = 1 \\ 0, & n = 2, 3, \ldots \end{cases} \tag{4.5}$$

Poiché il sistema di servizio $M/M/1/1$ è a capacità finita, raggiunge sempre una situazione di equilibrio statistico. Vogliamo ora determinare tale distribuzione. Facendo uso di (4.5) in (3.17) e ponendo $\varrho = \lambda/\mu$ si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \frac{\lambda}{\mu} = 1 + \varrho$$

Proposizione 4.3 *Per il sistema $M/M/1/1$ in condizioni di equilibrio statistico si ha*

$$q_0 = P(N = 0) = \frac{1}{1+\varrho}, \qquad q_1 = P(N = 1) = \frac{\varrho}{1+\varrho}$$

e

$$E(N) = E(N_s) = q_1 = \frac{\varrho}{1+\varrho}\cdot$$

Inoltre, la frequenza media di servizio è $\mu^* = \mu$ e dalla terza legge di Little si ricava

$$\lambda^* = \frac{E(N_s)}{E(S)} = \frac{\varrho}{1+\varrho}\,\mu = \frac{\lambda}{1+\varrho} = \lambda\, q_0,$$

che mostra che la frequenza media di arrivo è il prodotto della probabilità che un utente in arrivo non trovi utenti nel sistema e della frequenza di arrivo λ al sistema. Quindi, l'intensità di traffico, che coincide con il fattore di utilizzazione del sistema, è

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{\varrho}{1+\varrho} = 1 - q_0$$

che coincide con la probabilità che il sistema sia occupato. I risultati per il sistema $M/M/1/1$ sono riportati in Tabella 4.4.

$$\lambda_n = \begin{cases} \lambda, & n = 0 \\ 0, & n = 1, 2, \ldots \end{cases} \qquad \mu_n = \begin{cases} \mu, & n = 1 \\ 0, & n = 2, 3, \ldots \end{cases}$$

$$\varrho = \lambda/\mu < +\infty$$

$$q_0 = P(N = 0) = \frac{1}{1+\varrho}, \qquad q_1 = P(N = 1) = \frac{\varrho}{1+\varrho}$$

$$\lambda^* = \frac{\lambda}{1+\varrho} \qquad \mu^* = \mu \qquad \varrho^* = \frac{\varrho}{1+\varrho}$$

$$E(N) = \frac{\varrho}{1+\varrho}, \qquad E(W) = \frac{1}{\mu}$$

$$E(N_q) = 0, \qquad E(Q) = 0, \quad E(N_s) = \frac{\varrho}{1+\varrho} = \varrho^*, \quad E(S) = \frac{1}{\mu}$$

Tabella 4.4: Parametri prestazionali del sistema di servizio $M/M/1/1$

4.5 Sistema di servizio $M/M/1/K$

Supponiamo che gli utenti arrivino ad un sistema di servizio secondo un processo di Poisson di parametro λ. Dopo l'arrivo, ogni utente è immediatamente servito se il servitore è libero, mentre se il servitore è occupato l'utente si mette in fila di attesa se il numero di utenti presenti nel sistema è minore di K, mentre se K utenti sono già presenti nel sistema un nuovo utente in arrivo non può accedere al sistema. Quando il servitore termina di servire un utente, tale utente lascia il sistema e il nuovo utente nella fila di attesa (se ne esiste almeno uno presente) può usufruire del servizio. Supponiamo che i tempi di servizio degli utenti siano indipendenti e esponenzialmente distribuiti con valore medio $1/\mu$. In questo sistema la fila di attesa è limitata, ossia al più K utenti (incluso quello in servizio) possono essere presenti nel sistema e la disciplina di servizio è quella FIFO.

Tale sistema, illustrato in Figura 4.3, è noto in letteratura come *sistema di servizio* $M/M/1/K$.

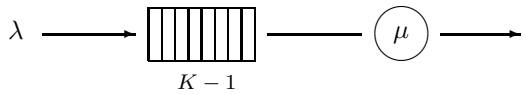

Figura 4.3: Sistema di servizio $M/M/1/K$.

La prima M significa che i tempi di interarrivo sono distribuiti esponenzialmente; la seconda M significa che i tempi di servizio sono anche distribuiti esponenzialmente; il simbolo 1 si riferisce all'unico servitore e il simbolo K indi-

ca la capacità del sistema. In particolare, se si pone $K = 1$, si ottiene il sistema di servizio $M/M/1/1$.

Denotiamo con $N(t)$ il numero di utenti presenti nel sistema $M/M/1/K$ al tempo t. Il processo stocastico $\{N(t),\, t \geq 0\}$ può essere descritto mediante un processo di nascita–morte caratterizzato da parametri:

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \ldots, K-1 \\ 0, & n = K, K+1, \ldots \end{cases}$$
$$\mu_n = \begin{cases} \mu, & n = 1, 2, \ldots, K \\ 0, & n = K+1, K+2, \ldots \end{cases} \tag{4.6}$$

Poiché il sistema di servizio $M/M/1/K$ è a capacità finita, raggiunge sempre una situazione di equilibrio statistico. Vogliamo ora determinare tale distribuzione. Facendo uso di (4.6) in (3.17) e ponendo $\varrho = \lambda/\mu$ si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \sum_{n=0}^{K} \Big(\frac{\lambda}{\mu}\Big)^n = \sum_{n=0}^{K} \varrho^n = \begin{cases} \dfrac{1-\varrho^{K+1}}{1-\varrho}, & \varrho \neq 1 \\ K+1, & \varrho = 1. \end{cases}$$

Proposizione 4.4 *Per il sistema $M/M/1/K$ in condizioni di equilibrio statistico si ha*

$$q_0 = P(N=0) = \begin{cases} \dfrac{1-\varrho}{1-\varrho^{K+1}}, & \varrho \neq 1 \\ \dfrac{1}{K+1}, & \varrho = 1 \end{cases}$$

e per $n = 1, 2, \ldots, K$ risulta: (4.7)

$$q_n = P(N=n) = q_0 \, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \begin{cases} \dfrac{1-\varrho}{1-\varrho^{K+1}}\, \varrho^n, & \varrho \neq 1 \\ \dfrac{1}{K+1}, & \varrho = 1. \end{cases}$$

A differenza del sistema $M/M/1$, il sistema $M/M/1/K$ raggiunge una situazione di equilibrio anche quando $\varrho \geq 1$. In particolare, quando $\varrho = 1$ la distribuzione di equilibrio è quella equiprobabile, ossia una distribuzione che assegna uguali probabilità a tutti i $K+1$ stati del sistema di servizio. Se $\varrho \neq 1$, dalle (4.7) si ricava

$$\frac{q_n}{q_{n-1}} = \varrho \qquad (n = 1, 2, \ldots)$$

Tale relazione mostra che se $\varrho < 1$ si ha $q_n < q_{n-1}$, ossia $q_0 > q_1 > \ldots > q_K$; se invece $\varrho > 1$ si ha $q_n > q_{n-1}$, ossia $q_K > q_{K-1} > \ldots > q_0$. Quindi, se $\varrho < 1$ è più probabile trovare il sistema $M/M/1/K$ vuoto, mentre se $\varrho > 1$ è più probabile trovare il sistema $M/M/1/K$ saturo. Per evitare la congestione di un sistema di servizio (come nel sistema $M/M/1$ quando $\varrho \geq 1$) non è quindi conveniente ridurre la capacità del sistema. Infatti, se $\varrho > 1$, essendo più probabile trovare il sistema $M/M/1/K$ saturo, si impedisce a molti utenti di accedere al sistema di servizio.

Proposizione 4.5 *In condizioni di equilibrio statistico il numero medio di utenti presenti nel sistema $M/M/1/K$ è:*

$$E(N) = \begin{cases} \dfrac{\varrho}{1-\varrho} - \dfrac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{2}\,, & \varrho = 1. \end{cases} \tag{4.8}$$

Dimostrazione Se $\varrho = 1$ si ha:

$$E(N) = \sum_{n=1}^{K} n\, q_n = \frac{1}{K+1} \sum_{n=1}^{K} n = \frac{1}{K+1}\,\frac{K\,(K+1)}{2} = \frac{K}{2}\,,$$

mentre se $\varrho \neq 1$ risulta

$$E(N) = \sum_{n=1}^{K} n\, q_n = \frac{1-\varrho}{1-\varrho^{K+1}} \sum_{n=1}^{K} n\,\varrho^n = \frac{\varrho}{1-\varrho} - \frac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\,,$$

dove si è fatto uso dell'identità:

$$\sum_{n=1}^{K} n\, z^n = \begin{cases} \dfrac{z\,(1-z^{K+1}) - (K+1)\,z^{K+1}(1-z)}{(1-z)^2}, & z \neq 1 \\ \dfrac{K\,(K+1)}{2}, & z = 1. \end{cases}$$

□

La (4.8) mostra che se $\varrho < 1$ il numero medio di utenti presenti nel sistema $M/M/1/K$ è inferiore al numero medio $E(N) = \varrho/(1-\varrho)$ di utenti presenti nel sistema $M/M/1$.

La probabilità che un utente in arrivo sia rifiutato è $P(N = k) = q_k$, mentre

$$P(N < K) = 1 - q_K = \begin{cases} \dfrac{1-\varrho^K}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{K+1}\,, & \varrho = 1. \end{cases}$$

fornisce la *probabilità che il sistema di servizio non sia saturo*, ossia è la probabilità che altri utenti possano accedere al sistema.

Nel sistema $M/M/1/K$ le frequenze medie di arrivo e di partenza per unità di tempo sono:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n q_n = \lambda \sum_{n=0}^{K-1} q_n = \lambda\,(1-q_K) = \begin{cases} \dfrac{\lambda\,(1-\varrho^K)}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{\lambda\,K}{K+1}\,, & \varrho = 1, \end{cases}$$

$$\mu^* = \frac{1}{1-q_0} \sum_{n=1}^{+\infty} \mu_n\, q_n = \frac{\mu}{1-q_0} \sum_{n=1}^{K} q_n = \frac{\mu}{1-q_0}\,(1-q_0) = \mu.$$

Si nota inoltre che la frequenza media di partenza per unità di tempo è la stessa del sistema $M/M/1$, mentre la frequenza media di arrivo per unità di tempo dipende dalla capacità K del sistema e differisce da quella del modello $M/M/1$. Il fattore di utilizzazione del sistema, coincidente con l'intensità di traffico, è:

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{\lambda\,(1-q_K)}{\mu} = \varrho\,(1-q_K) = \begin{cases} \varrho\,\dfrac{1-\varrho^K}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{K+1}\,, & \varrho = 1. \end{cases}$$

Si nota immediatamente che

$$\varrho^* = 1 - q_0,$$

ossia il *fattore di utilizzazione coincide con la probabilità che nel sistema sia presente almeno un utente*, ossia $\varrho^* = P(N \geq 1)$.

Proposizione 4.6 *Nella situazione di equilibrio statistico il tempo medio di attesa di un utente nel sistema $M/M/1/K$ è quindi:*

$$E(W) = \frac{E(N)}{\lambda^*} = \begin{cases} \dfrac{1}{\mu}\Big[\dfrac{1}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\Big], & \varrho \neq 1 \\ \dfrac{K+1}{2\,\mu}\,, & \varrho = 1. \end{cases} \tag{4.9}$$

Dimostrazione Dalla prima legge di Little ricaviamo il tempo medio di attesa di un utente nel sistema. Se $\varrho \neq 1$ si ha:

$$\begin{aligned} E(W) = \frac{E(N)}{\lambda^*} &= \frac{1-\varrho^{K+1}}{\lambda\,(1-\varrho^K)}\Big[\frac{\varrho}{1-\varrho} - \frac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\Big] \\ &= \frac{1}{\mu\,(1-\varrho^K)}\Big[\frac{1-\varrho^{K+1}}{1-\varrho} - (K+1)\,\varrho^K\Big] \\ &= \frac{1}{\mu\,(1-\varrho^K)}\,\frac{1-\varrho^K - K\,\varrho^K\,(1-\varrho)}{1-\varrho} \\ &= \frac{1}{\mu}\Big[\frac{1}{1-\varrho} - \frac{K\,\varrho^K}{1-\varrho^K}\Big], \end{aligned}$$

mentre se $\varrho = 1$ si ottiene:

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{K}{2}\,\frac{K+1}{\lambda\,K} = \frac{K+1}{2\,\lambda} = \frac{K+1}{2\,\mu}\,.$$

□

Se $\varrho < 1$, dalla (4.9) segue che il tempo medio di attesa nel sistema $M/M/1/K$ è inferiore al tempo medio di attesa $E(W) = 1/[\mu(1-\varrho)]$ nel sistema $M/M/1$.

Proposizione 4.7 *Nella situazione di equilibrio statistico il numero medio di utenti in fila di attesa nel sistema $M/M/1/K$ è:*

$$E(N_q) = \begin{cases} \dfrac{\varrho\,(1-\varrho^K)}{1-\varrho^{K+1}}\,\Big[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\Big], & \varrho \neq 1 \\ \dfrac{K\,(K-1)}{2\,(K+1)}, & \varrho = 1. \end{cases} \tag{4.10}$$

Dimostrazione Il numero medio di utenti in fila di attesa è:

$$E(N_q) = \sum_{n=1}^{K}(n-1)\,q_n = \sum_{n=1}^{K} n\,q_n - \sum_{n=1}^{K} q_n = E(N) - (1-q_0) \tag{4.11}$$

Dalla (4.11), se $\varrho = 1$ si ha

$$E(N_q) = \frac{K}{2} - \frac{K}{K+1} = \frac{K\,(K-1)}{2\,(K+1)}\,,$$

mentre se $\varrho \neq 1$ risulta:

$$E(N_q) = \frac{\varrho}{1-\varrho} - \frac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}} - \frac{\varrho\,(1-\varrho^k)}{1-\varrho^{k+1}} = \frac{\varrho}{1-\varrho} - \frac{\varrho\,(K\,\varrho^k+1)}{1-\varrho^{k+1}}$$
$$= \frac{\varrho}{1-\varrho^{k+1}}\,\Big(\frac{1-\varrho^{k+1}}{1-\varrho} - K\,\varrho^k - 1\Big) = \frac{\varrho\,(1-\varrho^K)}{1-\varrho^{K+1}}\,\Big[\frac{\varrho}{1-\varrho} - \frac{K\,\varrho^K}{1-\varrho^K}\Big].$$

□

Se $\varrho < 1$, dalla (4.10), il numero medio di utenti nella fila di attesa del sistema $M/M/1/K$ è inferiore al numero medio $E(N_q) = \varrho^2/(1-\varrho)$ di utenti nella fila di attesa del sistema $M/M/1$.

Dalla seconda legge di Little scaturisce che il tempo medio di permanenza di un utente nella fila di attesa è

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \begin{cases} \dfrac{1}{\mu}\,\Big[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\Big], & \varrho \neq 1 \\ \dfrac{K-1}{2\,\mu}, & \varrho = 1. \end{cases} \tag{4.12}$$

Se $\varrho < 1$, dalla (4.12) segue che il tempo medio di permanenza nella fila di attesa nel sistema $M/M/1/K$ è inferiore al tempo medio di permanenza nella fila di attesa $E(Q) = \varrho/[\mu(1-\varrho)]$ nel sistema $M/M/1$.

Nella situazione di equilibrio si può valutare anche il numero medio di utenti in servizio, ossia

$$E(N_s) = E(N) - E(N_q) = \begin{cases} \dfrac{\varrho\,(1-\varrho^K)}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{K+1}\,, & \varrho = 1, \end{cases}$$

che coincide con l'intensità di traffico e anche con il fattore di utilizzazione ϱ^* del sistema. Se $\varrho < 1$ si ha anche che il numero medio di utenti in servizio nel sistema $M/M/1/K$ è inferiore al numero medio $E(N_s) = \varrho$ di utenti in servizio nel sistema $M/M/1$. Dalla terza legge di Little segue anche l'identità $E(S) = E(N_s)/\lambda^* = 1/\mu$.

Nella Tabella 4.5 sono riportati i principali parametri prestazionali del sistema di servizio $M/M/1/K$.

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \ldots, K-1 \\ 0, & n = K, K+1, \ldots \end{cases} \qquad \mu_n = \begin{cases} \mu, & n = 1, 2, \ldots, K \\ 0, & n = K+1, K+2, \ldots \end{cases}$$

$$\varrho = \lambda/\mu < +\infty$$

$$q_n = P(N = n) = \begin{cases} \dfrac{1-\varrho}{1-\varrho^{K+1}}\,\varrho^n\,, & \varrho \neq 1 \\ 1/(K+1)\,, & \varrho = 1 \end{cases} \qquad (n = 0, 1, \ldots, K)$$

$$\lambda^* = \lambda\,(1 - q_K) = \begin{cases} \dfrac{\lambda\,(1-\varrho^K)}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \lambda\,K/(K+1)\,, & \varrho = 1 \end{cases}, \qquad \mu^* = \mu \qquad \varrho^* = \varrho\,(1 - q_K)$$

$$E(N) = \begin{cases} \dfrac{\varrho}{1-\varrho} - \dfrac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ K/2\,, & \varrho = 1 \end{cases}, \quad E(W) = \begin{cases} \dfrac{1}{\mu}\left[\dfrac{1}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\right], & \varrho \neq 1 \\ (K+1)/(2\,\mu)\,, & \varrho = 1 \end{cases}$$

$$E(N_q) = \begin{cases} \varrho^*\left[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\right], & \varrho \neq 1 \\ K\,(K-1)/[2\,(K+1)]\,, & \varrho = 1 \end{cases}, \quad E(Q) = \begin{cases} \dfrac{1}{\mu}\left[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\right], & \varrho \neq 1 \\ (K-1)/(2\,\mu)\,, & \varrho = 1 \end{cases}$$

$$E(N_s) = \varrho\,(1 - q_K) = \varrho^*, \qquad E(S) = 1/\mu$$

Tabella 4.5: Parametri prestazionali del sistema di servizio $M/M/1/K$

Esempio 4.2 Si consideri un centralino telefonico con un'unica linea disponibile che consenta l'attesa di due chiamate. Ulteriori chiamate quando nel sistema sono presenti tre chiamate (una in servizio e due in attesa) saranno rifiutate. Si supponga che la frequenza media di arrivo è di 120 chiamate all'ora e che la durata media di una telefonata è di 20 secondi. Se un sistema di servizio $M/M/1/3$ modella il centralino telefonico considerato, calcolare i principali parametri prestazionali del sistema.

In questo caso risulta

$$\lambda = \frac{120}{60} = 2 \text{ chiamate al minuto}, \qquad \mu = 60\,\frac{1}{20} = 3 \text{ chiamate al minuto},$$

e quindi $\varrho = \lambda/\mu = 2/3$. La distribuzione di equilibrio risulta essere:

$$q_n = P(N = n) = \frac{1-\varrho}{1-\varrho^4}\,\varrho^n = \frac{1-2/3}{1-(2/3)^4}\left(\frac{2}{3}\right)^n = \frac{27}{65}\left(\frac{2}{3}\right)^n,$$

da cui segue che

$$q_0 = \frac{27}{65}, \qquad q_1 = \frac{18}{65}, \qquad q_2 = \frac{12}{65}, \qquad q_3 = \frac{8}{65}.$$

La frequenze medie di arrivo e di partenza sono:

$$\lambda^* = \lambda\,(1 - q_3) = 2\left(1 - \frac{8}{65}\right) = \frac{114}{65} = 1.754 \text{ chiamate al minuto,}$$
$$\mu^* = \mu = 3 \text{ chiamate al minuto.}$$

e quindi il fattore di utilizzazione del sistema è

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{114}{65} \cdot \frac{1}{3} = \frac{38}{65} = 0.585,$$

che coincide con $E(N_s)$. Il numero medio di chiamate presenti nel centralino telefonico è

$$E(N) = \frac{\varrho}{1-\varrho} - \frac{4\,\varrho^4}{1-\varrho^4} = \frac{2/3}{1-2/3} - \frac{4\,(2/3)^4}{1-(2/3)^4} = 2 - \frac{64}{65} = \frac{66}{65} = 1.015 \text{ chiamate}$$

e il tempo medio di attesa nel centralino (permanenza in fila di attesa e in servizio) è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{66}{65} \cdot \frac{65}{114} = \frac{33}{57} = 0.579 \text{ minuti.}$$

Inoltre il numero medio di chiamate in fila di attesa è:

$$E(N_q) = E(N) - E(N_s) = \frac{66}{65} - \frac{38}{65} = \frac{28}{65} = 0.431 \text{ chiamate}$$

e quindi il tempo medio di permanenza in fila di attesa è

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{28}{65} \cdot \frac{65}{114} = \frac{14}{57} = 0.245 \text{ minuti.}$$

♢

4.6 Sistema di servizio $M/G/1$

Lo stato di un sistema di servizio di tipo $M/M/1$ può essere modellato utilizzando un processo stocastico di nascita morte $\{N(t), t \geq 0\}$, dove $N(t)$ rappresenta il numero di utenti presenti nel sistema al tempo t. Per studiare tale sistema ci siamo avvalsi soprattutto dell'ipotesi che i tempi di interarrivo ed anche quelli di servizio sono distribuiti esponenzialmente. Per la proprietà di Markov, legata all'assenza di memoria della densità di probabilità esponenziale, i tempi di interarrivo e di servizio residui sono distribuiti con la stessa legge di probabilità dei rispettivi tempi di interarrivo e di servizio. Quindi nel sistema $M/M/1$ non è necessario avere memoria sia del tempo trascorso dall'ultimo arrivo quando

un nuovo cliente arriva nel sistema sia del tempo di servizio già speso dal cliente che sta ricevendo attualmente il servizio.

A differenza del sistema di servizio $M/M/1$, il sistema di servizio $M/G/1$ deve essere modellato con un processo stocastico non–markoviano. Infatti per poter descrivere lo stato del sistema di servizio $M/G/1$ occorre specificare per ogni istante temporale t non soltanto il numero $N(t)$ di utenti presenti nel sistema al tempo t, ma anche il tempo di servizio $Y(t)$ che ha già ricevuto il cliente attualmente in servizio.

Vogliamo ora analizzare il sistema $M/G/1$ in cui gli arrivi si verificano secondo un processo di Poisson di parametro λ ed in cui i tempi di servizio sono indipendenti ed identicamente distribuiti con funzione di distribuzione di tipo generale, esiste un unico servitore e la capacità del sistema è infinita. Supponiamo che $E(S) = 1/\mu$.

Analogamente al sistema $M/M/1$, il sistema $M/G/1$ raggiunge una situazione di equilibrio statistico se $\varrho = \lambda/\mu < 1$ e la probabilità che nel sistema non siano presenti utenti è ancora $q_0 = 1 - \varrho$. Inoltre, il numero medio di utenti nel sistema in condizioni di equilibrio statistico è:

$$E(N) = \varrho + \frac{\varrho^2\,[1 + C^2(S)]}{2\,(1 - \varrho)} \qquad (\varrho < 1) \tag{4.13}$$

dove $C(S)$ denota il coefficiente di variazione della variabile aleatoria S. La (4.13) è detta *formula di Pollaczek–Khintchine*.

In particolare, se si considera il sistema $M/M/1$, allora $C(S) = 1$ e la (4.13) diventa $E(N) = \varrho/(1 - \varrho)$, che corrisponde all'espressione direttamente calcolata per il sistema $M/M/1$. Invece se si considera il sistema $M/D/1$, allora $C(S) = 0$ e la (4.13) diventa $E(N) = \varrho + \varrho^2/[2(1 - \varrho)]$.

Vogliamo ora determinare gli altri parametri prestazionali del sistema di servizio $M/G/1$. Utilizzando la prima legge di Little risulta:

$$E(W) = \frac{E(N)}{\lambda} = \frac{1}{\mu} + \frac{\varrho\,[1 + C^2(S)]}{2\,\mu\,(1 - \varrho)}$$

Pertanto

$$E(Q) = E(W) - E(S) = \frac{\varrho\,[1 + C^2(S)]}{2\,\mu\,(1 - \varrho)},$$

da cui applicando la seconda legge di Little segue che

$$E(N_q) = \lambda\,E(Q) = \frac{\varrho^2\,[1 + C^2(S)]}{2\,(1 - \varrho)}.$$

Inoltre si ha:

$$E(N_s) = E(N) - E(N_q) = \varrho,$$

che coincide con l'intensità di traffico del sistema di servizio $M/G/1$.

Notiamo ora che nel modello $M/G/1$ i periodi di ozio, che possono essere riguardati come intervalli residui degli intervalli di interarrivo, sono indipendenti

e distribuiti esponenzialmente con valore medio $1/\lambda$. Pertanto, si ha:

$$E(I) = \frac{1}{\lambda}, \qquad E(B) = \frac{(1-q_0)\,E(I)}{q_0} = \frac{\varrho}{\lambda\,(1-\varrho)} = \frac{1}{\mu-\lambda}.$$

Quindi, il tempo medio di ozio e di occupazione coincidono con quelli ottenuto per il modello $M/M/1$.

Occorre infine osservare che fissati il tempo medio di interarrivo $E(T) = 1/\lambda$ ed il tempo medio di servizio $E(S) = 1/\mu$, all'aumentare del coefficiente di variazione $C(S)$ aumenta il tempo medio di permanenza nella fila di attesa, il tempo medio di attesa nel sistema, il numero medio di utenti nella fila di attesa ed il numero medio di utenti nel sistema. Pertanto, a parità di tempo medio di interarrivo $E(T) = 1/\lambda$ e di tempo medio di servizio $E(S) = 1/\mu$, il sistema $M/D/1$ (con coefficiente di variazione $C(S) = 0$) ha parametri prestazionali migliori rispetto al sistema $M/E_k/1$ (con coefficiente di variazione $C(S) = 1/\sqrt{k}$ che a sua volta ha parametri prestazionali migliori rispetto al sistema $M/M/1$ (con coefficiente di variazione $C(S) = 1$). Le prestazioni peggiori si manifestano nel sistema $M/H_k/1$, che ha coefficiente di variazione $C(S) \geq 1$.

Capitolo 5

Modelli con più servitori

5.1 Introduzione

In questo capitolo analizzeremo i principali sistemi di servizio con più servitori che lavorano in parallelo, ossia i sistemi $M/M/2$, $M/M/s$, $M/M/s/s$ e $M/M/\infty$, oltre ad altri sistemi di servizio di tipo adattivo di utilità nella teoria delle file di attesa. Lo scopo è quello di determinare i principali parametri prestazionali dei vari sistemi di servizio e di individuare, se necessario, idonee politiche atte ad evitare la congestione del sistema.

5.2 Sistema di servizio $M/M/2$

Un ovvio rimedio per un sistema di servizio che presenta congestione consiste nell'aumentare il numero di servitori. Consideriamo quindi un sistema di servizio a capacità infinita con un'unica fila di attesa e due servitori identici che lavorano in parallelo, rappresentato in Figura 5.1.

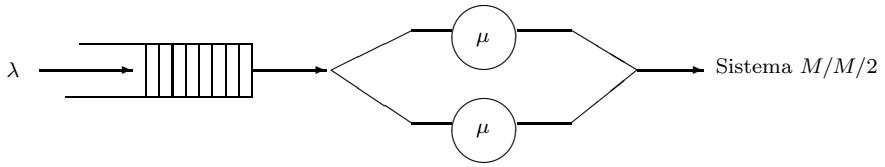

Figura 5.1: Il sistema di servizio $M/M/2$.

Supponiamo che i tempi di interarrivo siano indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$ e che i tempi di servizio per ognuno dei due

servitori siano indipendenti e distribuiti esponenzialmente con valore medio $1/\mu$. Se un utente arriva e trova entrambi i servitori occupati si mette in fila di attesa, se trova un servitore occupato e l'altro libero sceglie il servitore libero e se entrambi i servitori sono liberi sceglie a caso uno di essi per essere servito. Tale sistema di servizio, noto in letteratura come $M/M/s$, è descrivibile mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri:

$$\begin{aligned} &\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \\ &\mu_n = \mu\, \min(n, 2) = \begin{cases} \mu & n = 1 \\ 2\,\mu, & n = 2, 3, \ldots \end{cases} \end{aligned} \tag{5.1}$$

Vogliamo vedere in quali condizioni il sistema $M/M/2$ raggiunge una situazione di equilibrio statistico. Facendo uso di (5.1) in (3.17) si ha:

$$\begin{aligned} 1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} &= 1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu\,(2\,\mu)} + \frac{\lambda^3}{\mu\,(2\,\mu)^2} + \frac{\lambda^4}{\mu\,(2\,\mu)^3} + \ldots \\ &= 1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu\,(2\,\mu)} \sum_{k=0}^{+\infty} \Big(\frac{\lambda}{2\,\mu}\Big)^k. \end{aligned}$$

Se si pone

$$\varrho_2 = \frac{\lambda}{2\,\mu},$$

si nota che la serie converge se e solo se $\varrho_2 < 1$ e si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + 2\,\varrho_2 + 2\,\varrho_2^2\, \frac{1}{1 - \varrho_2} = \frac{1 + \varrho_2}{1 - \varrho_2}.$$

Proposizione 5.1 *Se $\varrho_2 = \lambda/(2\,\mu) < 1$ il sistema $M/M/2$ raggiunge una situazione di equilibrio e risulta:*

$$\begin{aligned} q_0 &= P(N = 0) = \frac{1 - \varrho_2}{1 + \varrho_2}\,, \\ q_n &= P(N = n) = 2\,\frac{1 - \varrho_2}{1 + \varrho_2}\,\varrho_2^n \quad (n = 1, 2, \ldots). \end{aligned} \tag{5.2}$$

La probabilità che in condizioni di equilibrio un utente in arrivo debba attendere in coda è

$$P(N \geq 2) = \sum_{n=2}^{+\infty} q_n = 2\,\frac{1 - \varrho_2}{1 + \varrho_2} \sum_{n=2}^{+\infty} \varrho_2^n = 2\,\frac{1 - \varrho_2}{1 + \varrho_2}\,\varrho_2^2 \sum_{n=2}^{+\infty} \varrho_2^{n-2} = \frac{2\,\varrho_2^2}{1 + \varrho_2}.$$

Tale probabilità è nota come "*formula C di Erlang*" ed è indicata con $C[2, \lambda/\mu]$. Per il sistema $M/M/2$ in condizioni di equilibrio si ha che la media e la varianza

$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad \mu_1 = \mu, \qquad \mu_n = 2\,\mu \quad (n = 2, 3, \ldots)$

$\varrho_2 = \dfrac{\lambda}{2\,\mu} < 1 \qquad$ (condizione di equilibrio statistico)

$q_0 = \dfrac{1-\varrho_2}{1+\varrho_2}\,, \qquad q_n = 2\,\dfrac{1-\varrho_2}{1+\varrho_2}\,\varrho_2^n \qquad (n = 1, 2, \ldots)$

$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad a = \dfrac{\lambda^*}{\mu^*} = \dfrac{\lambda}{\mu}\,, \qquad \varrho^* = \dfrac{a}{2} = \dfrac{\lambda}{2\,\mu} = \varrho_2$

$C[2, \lambda/\mu] = P(N \geq 2) = \dfrac{2\,\varrho_2^2}{1+\varrho_2} \quad$ (formula C di Erlang)

$E(N) = \dfrac{2\,\varrho_2}{1-\varrho_2^2}, \qquad E(W) = \dfrac{1}{\mu\,(1-\varrho_2^2)}$

$E(N_q) = \dfrac{2\,\varrho_2^3}{1-\varrho_2^2}\,, \qquad E(Q) = \dfrac{\varrho_2^2}{\mu\,(1-\varrho_2^2)}$

$E(N_s) = \dfrac{\lambda}{\mu}, \qquad E(S) = \dfrac{1}{\mu}$

Tabella 5.1: Parametri prestazionali del sistema di servizio $M/M/2$.

del numero di utenti nel sistema è

$$E(N) = \sum_{n=1}^{+\infty} n\,q_n = 2\,\frac{1-\varrho_2}{1+\varrho_2} \sum_{n=1}^{+\infty} n\,\varrho_2^n = \frac{2\,\varrho_2}{1-\varrho_2^2},$$
$$\mathrm{Var}(N) = \frac{2\,\varrho_2\,(1+\varrho_2^2)}{(1-\varrho_2^2)^2},$$

e quindi applicando la prima legge di Little risulta:

$$E(W) = \frac{E(N)}{\lambda} = \frac{1}{\mu\,(1-\varrho_2^2)}.$$

Poichè il tempo medio di servizio per servitore è $E(S) = 1/\mu$, segue che

$$E(Q) = E(W) - E(S) = \frac{1}{\mu\,(1-\varrho_2^2)} - \frac{1}{\mu} = \frac{\varrho_2^2}{\mu\,(1-\varrho_2^2)},$$

da cui applicando la seconda legge di Little si ricava:

$$E(N_q) = \lambda\,E(Q) = \frac{2\,\varrho_2^3}{1-\varrho_2^2}.$$

Infine si può notare che il numero medio di utenti in servizio

$$E(N_s) = E(N) - E(Nq) = \frac{2\,\varrho_2}{1-\varrho_2^2} - \frac{2\,\varrho_2^3}{1-\varrho_2^2} = \frac{\lambda}{\mu}$$

coincide con l'intensità di traffico, ossia con l'intensità di lavoro svolta dal centro di servizio. I parametri prestazionali del sistema di servizio $M(M/2$ sono stati elencati in Tabella 5.1.

5.2.1 Confronti tra i sistemi $M/M/1$ e $M/M/2$

Una domanda che ci si può porre è la seguente: *a parità del processo degli arrivi è più conveniente scegliere un sistema con due servitori ognuno dei quali ha tempo medio di servizio pari a* $1/\mu$ *oppure un sistema con unico servitore avente tempo medio di servizio dimezzato rispetto al tempo medio di servizio dei due servitori del precedente sistema.*
In Figura 5.2 sono illustrati i due sistemi $M/M/1$ e $M/M/2$ considerati; il sistema $M/M/1$ ha un unico servitore doppiamente più veloce rispetto ad ognuno dei singoli servitori del sistema $M/M/2$.

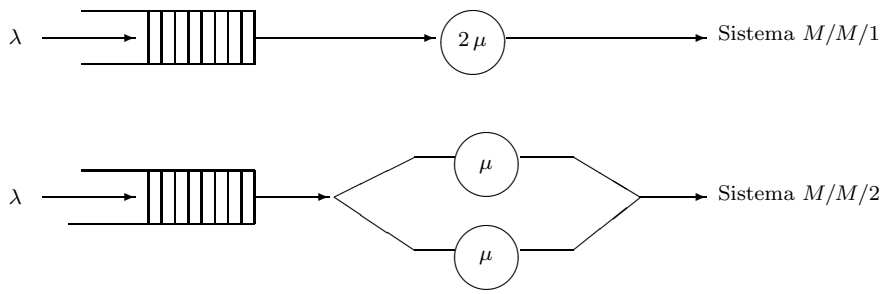

Figura 5.2: Due particolari sistemi di servizio $M/M/1$ e $M/M/2$.

Per rispondere alla precedente domanda consideriamo in primo luogo un sistema di servizio $M/M/1$ con parametri di arrivo $\lambda_n = \lambda \quad (n = 0, 1, \ldots)$ e con parametri di partenza $\mu_n = 2\,\mu \quad (n = 1, 2, \ldots)$; abbiamo precedentemente mostrato che se $\varrho_2 = \lambda/(2\,\mu) < 1$ esiste la distribuzione di equilibrio $(q_0^{(1)}, q_1^{(1)}, \ldots)$ e si ha

$$q_n^{(1)} = (1 - \varrho_2)\,\varrho_2^n \qquad (n = 0, 1, \ldots).$$

Consideriamo poi un sistema di servizio $M/M/2$ con parametri di arrivo $\lambda_n = \lambda$ $(n = 0, 1, \ldots)$ e con parametri di partenza $\mu_1 = \mu$, $\mu_n = 2\,\mu \quad (n = 2, 3, \ldots)$ che ammette la distribuzione di equilibrio $(q_0^{(2)}, q_1^{(2)}, \ldots)$ fornita in (5.2). Quando $\varrho_2 = \lambda/(2\,\mu) < 1$ si ha:

$$\frac{q_n^{(1)}}{q_n^{(2)}} = \begin{cases} 1 + \varrho_2, & n = 0 \\ \dfrac{1 + \varrho_2}{2}, & n = 1, 2, \ldots \end{cases} \tag{5.3}$$

Si nota immediatamente che

$$q_0^{(1)} > q_0^{(2)}; \qquad q_n^{(1)} < q_n^{(2)} \quad (n = 1, 2, \ldots). \tag{5.4}$$

Nella situazione di equilibrio statistico è quindi più probabile trovare il sistema vuoto nel sistema $M/M/1$ piuttosto che nel sistema $M/M/2$ e inoltre è meno probabile trovare nel sistema $M/M/1$ un numero n di utenti ($n = 1, 2, \ldots$) piuttosto che nel sistema $M/M/2$. Facendo uso delle (5.3) si nota anche che quando ϱ_2 si approssima all'unità si ha che $q_0^{(1)} \sim 2\, q_0^{(2)}$ mentre $q_n^{(1)} \sim q_n^{(2)}$ per $n = 1, 2, \ldots$. Ciò significa che se i sistemi sono molto utilizzati, mantenendosi però in condizioni di non congestione, la probabilità di avere il sistema vuoto nel sistema $M/M/1$ è approssimativamente il doppio di quella del sistema $M/M/2$ mentre la probabilità di avere un certo numero di utenti è approssimativamente uguale nei due sistemi.

$M/M/1$	$M/M/2$
$\lambda_n = \lambda \quad (n = 0, 1, \ldots)$ $\mu_n = 2\,\mu \quad (n = 1, 2, \ldots)$	$\lambda_n = \lambda \quad (n = 0, 1, \ldots)$ $\mu_1 = \mu, \quad \mu_n = 2\,\mu \quad (n = 2, 3, \ldots)$
$\varrho_2 = \lambda/(2\,\mu) < 1$ $\lambda^* = \lambda, \quad \mu^* = 2\,\mu, \quad \varrho^* = \dfrac{\lambda}{2\,\mu}$	$\varrho_2 = \lambda/(2\,\mu) < 1$ $\lambda^* = \lambda, \quad \mu^* = \mu, \quad \varrho^* = \dfrac{\lambda}{2\,\mu}$
$E(T) = \dfrac{1}{\lambda}, \qquad E(S) = \dfrac{1}{2\,\mu}$	$E(T) = \dfrac{1}{\lambda}, \qquad E(S) = \dfrac{1}{\mu}$
$q_n = (1 - \varrho_2)\,\varrho_2^n \quad (n = 0, 1, \ldots)$	$q_0 = \dfrac{1 - \varrho_2}{1 + \varrho_2}, \quad q_n = 2\,\dfrac{1 - \varrho_2}{1 + \varrho_2}\,\varrho_2^n$ $(n = 1, 2, \ldots)$
$E(N) = \dfrac{\varrho_2}{1 - \varrho_2}$	$E(N) = \dfrac{2\,\varrho_2}{1 - \varrho_2^2}$
$\mathrm{Var}(N) = \dfrac{\varrho_2}{(1 - \varrho_2)^2}$	$\mathrm{Var}(N) = \dfrac{2\,\varrho_2\,(1 + \varrho_2^2)}{(1 - \varrho_2^2)^2}$

Tabella 5.2: Confronto tra i parametri prestazionali di un sistema $M/M/2$ e di un sistema $M/M/1$ il cui tempo medio di servizio è dimezzato rispetto ai tempi medi di servizio dei singoli servitori del sistema $M/M/2$.

Per comprendere meglio le differenze tra i due sistemi è anche utile analizzare il valore medio, la varianza e il coefficiente di variazione del numero di utenti. Nella situazione di equilibrio statistico, se denotiamo con $E(N^{(s)})$ e $\mathrm{Var}(N^{(s)})$ rispettivamente il valore medio e la varianza del numero di utenti presenti nel sistema $M/M/s$ ($s = 1, 2$) si può mostrare che

$$\begin{aligned} E(N^{(1)}) &= \frac{\varrho_2}{1 - \varrho_2}, \qquad \mathrm{Var}(N^{(1)}) = \frac{\varrho_2}{(1 - \varrho_2)^2} \\ E(N^{(2)}) &= \frac{2\,\varrho_2}{1 - \varrho_2^2}, \qquad \mathrm{Var}(N^{(2)}) = \frac{2\,\varrho_2\,(1 + \varrho_2^2)}{(1 - \varrho_2^2)^2}. \end{aligned} \tag{5.5}$$

Quindi, se $\varrho_2 < 1$ si ha:

$$E(N^{(1)}) < E(N^{(2)}), \qquad \mathrm{Var}(N^{(1)}) < \mathrm{Var}(N^{(2)}).$$

Tali disuguaglianze mostrano che in media si hanno meno utenti nel sistema $M/M/1$ piuttosto che nel sistema $M/M/2$ con una dispersione dal valore medio inferiore. Inoltre, dalle (5.5) segue:

$$\frac{E(N^{(1)})}{E(N^{(2)})} = \frac{1+\varrho_2}{2}, \qquad \frac{\mathrm{Var}(N^{(1)})}{\mathrm{Var}(N^{(2)})} = \frac{(1+\varrho_2)^2}{2\,(1+\varrho_2^2)}$$

che mostrano che quando ϱ_2 è prossimo all'unità si ha $E(N^{(1)}) \sim E(N^{(2)})$ e $\mathrm{Var}(N^{(1)}) \sim \mathrm{Var}(N^{(2)})$. Si nota quindi che se i due sistemi di servizio sono molto utilizzati (ϱ_2 prossimo a uno), in condizioni di non congestione il numero medio di utenti sarà approssimativamente lo stesso in entrambi i sistemi. In conclusione, tra i due sistemi considerati, *il sistema $M/M/1$ è il più efficiente.*

La Tabella 5.2 riassume i principali risultati ottenuti dal confronto tra i sistemi $M/M/1$ e $M/M/2$, con il primo sistema avente tempo medio di servizio dimezzato rispetto ai tempi medi di servizio dei singoli servitori del sistema $M/M/2$.

Esempio 5.1 Ci proponiamo di confrontare i due sistemi di servizio illustrati Figura 5.3. Il primo sistema consiste di due sistemi di servizio $M/M/1$ indipendenti ognuno con fattore di utilizzazione del sistema $\varrho^* = \lambda/(2\,\mu) = \varrho$. Il secondo sistema consiste di un sistema di servizio $M/M/2$ con fattore di utilizzazione del sistema $\varrho^* = \lambda/(2\,\mu) = \varrho_2$, essendo presenti due servitori. In condizioni di equilibrio statistico, si desidera stabilire quale dei due sistemi sia più efficiente in base al tempo medio di attesa degli utenti nel sistema. Affinché i due sistemi

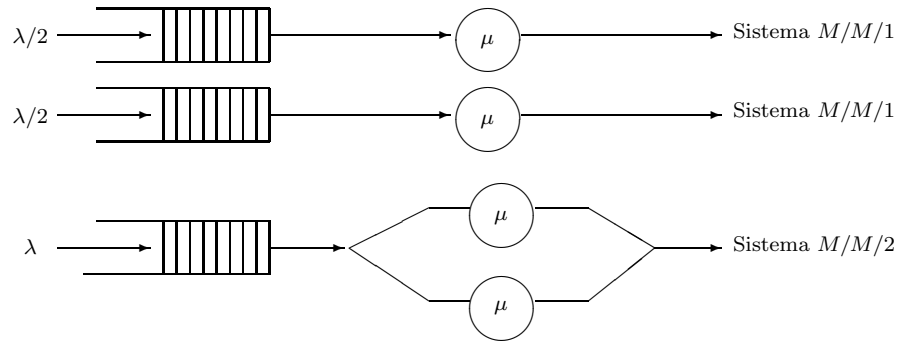

Figura 5.3: Confronto tra il sistema di servizio $M/M/2$ e un sistema di servizio costituito da due sistemi $M/M/1$ indipendenti.

raggiungano una situazione di equilibrio statistico occorre che $\varrho^* = \lambda/(2\,\mu) < 1$. Denotiamo con $W^{(1)}$ il tempo di attesa di un utente nel primo sistema e con A_i

l'evento "*è stato scelto il sistema* $M/M/1$ *i-esimo*" $(i = 1, 2)$. Dalla Tabella 4.2 segue:

$$E(W^{(1)}) = \frac{1}{2}\,E(W^{(1)}|A_1) + \frac{1}{2}\,E(W^{(1)}|A_2) = \frac{1}{2}\,\frac{1}{\mu - \lambda/2} + \frac{1}{2}\,\frac{1}{\mu - \lambda/2} = \frac{2}{2\,\mu - \lambda}.$$

Denotiamo poi con $W^{(2)}$ il tempo di attesa di un utente nel secondo sistema $M/M/2$. Dalla Tabella 5.1 segue:

$$E(W^{(2)}) = \frac{1}{\mu\,(1 - \varrho_2^2)} = \frac{1}{\mu\left[1 - \left(\frac{\lambda}{2\,\mu}\right)^2\right]} = \frac{4\,\mu}{4\,\mu^2 - \lambda^2}.$$

Si nota che

$$E(W^{(2)}) = \frac{4\,\mu}{(2\,\mu - \lambda)\,(2\,\mu + \lambda)} = \frac{2\,\mu}{2\,\mu + \lambda}\,E(W^{(1)}) < E(W^{(1)}).$$

Quindi il tempo medio di attesa di un utente nel sistema di servizio $M/M/2$ è inferiore al tempo medio di attesa di un utente nel sistema costituito da due sistemi $M/M/1$ indipendenti. Inoltre, se denotiamo con $N^{(1)}$ il numero di utenti presenti nel primo sistema e con q_j la probabilità di avere j utenti in uno dei due sistemi $M/M/1$ risulta che

$$\begin{aligned} q_n^{(1)} = P(N^{(1)} = n) &= \sum_{k=0}^{n} q_k\,q_{n-k} = \sum_{k=0}^{n}\left[(1 - \varrho_2)\varrho_2^k\right]\left[(1 - \varrho_2)\varrho_2^{n-k}\right] \\ &= (1 - \varrho_2)^2 \sum_{k=0}^{n} \varrho_2^n = (n+1)(1 - \varrho_2)^2\,\varrho_2^n \qquad (n = 0, 1, \ldots). \end{aligned}$$

Utilizzando la prima legge di Little si ha inoltre:

$$E(N^{(1)}) = \lambda\,E(W^{(1)}) = \frac{2\,\lambda}{2\,\mu - \lambda} = \frac{2\varrho_2}{1 - \varrho_2}.$$

Sia ora $N^{(2)}$ il numero di utenti presenti nel sistema $M/M/2$. Dalla Tabella 5.1 segue che $E(N^{(2)}) = 2\varrho_2/(1 - \varrho_2^2)$, che mostra che il numero medio di utenti nel sistema di servizio $M/M/2$ è inferiore al numero medio di utenti nel sistema costituito da due sistemi $M/M/1$ indipendenti.

In conclusione, *è più efficiente il sistema di servizio* $M/M/2$. Ciò è dovuto alla circostanza che un utente in arrivo nel primo sistema sceglie a caso una delle due file non tenendo conto del numero di utenti già presenti nei due sistemi $M/M/1$ indipendenti. ♢

5.3 Sistema di servizio $M/M/s$

Consideriamo ora un sistema di servizio a capacità infinita con un'unica fila di attesa e s servitori identici illustrato in Figura 5.4.

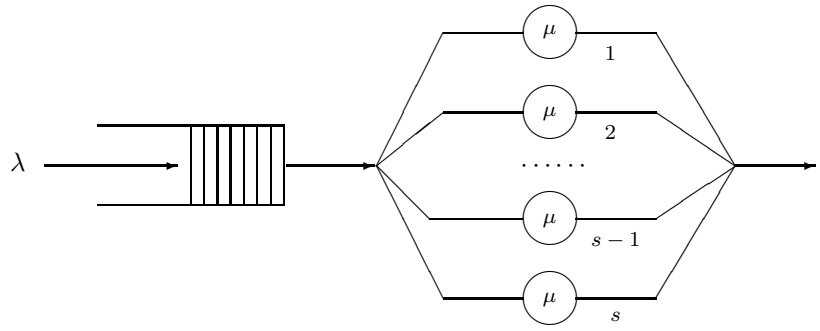

Figura 5.4: Sistema di servizio $M/M/s$.

Supponiamo che i tempi di interarrivo siano indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$ e che i tempi di servizio per ogni servitore siano indipendenti e distribuiti esponenzialmente con valore medio $1/\mu$. Se un utente arriva e trova tutti i servitori occupati si mette in fila di attesa, mentre se trova dei servitori liberi sceglie a caso uno di essi per essere servito. Tale sistema di servizio, noto in letteratura come $M/M/s$, è descrivibile mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri:

$$\begin{aligned} &\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \\ &\mu_n = \mu \, \min(n, s) = \begin{cases} n\,\mu & n = 1, 2, \ldots, s \\ s\,\mu, & n = s+1, s+2, \ldots \end{cases} \end{aligned} \tag{5.6}$$

In particolare, se $s = 1$ si ha il sistema di servizio $M/M/1$ con unico servitore mentre se $s = 2$ si ottiene il sistema di servizio $M/M/2$ con due servitori.

In primo luogo vediamo in quali condizioni tale sistema raggiunge una situazione di equilibrio statistico. Facendo uso di (5.6) in (3.17) si ha:

$$\begin{aligned} 1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} &= 1 + \sum_{n=1}^{s-1} \frac{1}{n!} \Big(\frac{\lambda}{\mu}\Big)^n + \frac{\lambda^s}{s!\,\mu^s} \sum_{k=0}^{+\infty} \Big(\frac{\lambda}{s\,\mu}\Big)^k \\ &= \sum_{n=0}^{s-1} \frac{s^n}{n!} \Big(\frac{\lambda}{s\,\mu}\Big)^n + \frac{s^s}{s!} \Big(\frac{\lambda}{s\,\mu}\Big)^s \sum_{k=0}^{+\infty} \Big(\frac{\lambda}{s\,\mu}\Big)^k . \end{aligned}$$

Se si pone

$$\varrho_s = \frac{\lambda}{s\,\mu}\,, \tag{5.7}$$

si nota che la serie converge se e solo se $\varrho_s < 1$ e si ha

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{s^s}{s!} \frac{\varrho_s^s}{1 - \varrho_s}$$

$$= \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)}\,.$$

Proposizione 5.2 *Il sistema di servizio $M/M/s$ raggiunge quindi una situazione di equilibrio statistico se e solo se $\varrho_s = \lambda/(s\,\mu) < 1$ e risulta:*

$$q_0 = P(N=0) = \Big[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)}\Big]^{-1}, \tag{5.8}$$

$$q_n = P(N=n) = q_0\,\frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n} = \begin{cases} q_0\,\dfrac{(\lambda/\mu)^n}{n!}, & n = 1,2,\ldots,s \\ q_0\,\dfrac{(\lambda/\mu)^n}{s^{n-s}\,s!}, & n = s+1, s+2, \ldots. \end{cases}$$

Si noti che ponendo $s = 2$ nella (5.8) si ottiene la (5.2) per il sistema $M/M/2$.

Il parametro $\varrho_s = \lambda/(s\,\mu)$ fornisce una misura di congestione del sistema $M/M/s$. Infatti, se $\varrho_s \geq 1$ il sistema di servizio è *instabile* nel senso che il numero di utenti in fila di attesa è destinato a crescere indefinitamente.

Per decidere il numero di servitori necessari e sufficienti affinché il sistema sia stabile occorre osservare il rapporto λ/μ determinando l'intero positivo s tale che

$$s - 1 \leq \frac{\lambda}{\mu} < s. \tag{5.9}$$

Se $s - 1 \leq \varrho < s$, allora s servitori sono necessari per raggiungere la situazione di equilibrio statistico poiché in tal caso $\varrho_s = \lambda/(s\mu) < 1$ e inoltre sono anche sufficienti nel senso che sarebbe poco economico considerare più di s servitori.

Poiché nel sistema $M/M/s$ il tempo medio di interarrivo è $E(T) = 1/\lambda$ e il tempo medio di servizio per ognuno dei servitori è $E(S) = 1/\mu$, si ha

$$\lambda^* = \frac{1}{E(T)} = \lambda, \qquad \mu^* = \frac{1}{E(S)} = \mu.$$

Quindi l'intensità di traffico relativa al centro di servizio è

$$a = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu}$$

e il fattore di utilizzazione del sistema è:

$$\varrho^* = \frac{\lambda^*}{s\,\mu^*} = \frac{\lambda}{s\,\mu} = \varrho_s\,.$$

Un parametro molto importante per il sistema di servizio $M/M/s$ è rappresentato dalla *probabilità che un utente in arrivo debba attendere nella fila di attesa prima di ricevere il servizio.* È evidente che ciò si verifica se e solo se vi sono

almeno s utenti già presenti nel sistema. Tale probabilità, detta *formula C di Erlang*, può essere così ottenuta:

$$C[s,\lambda/\mu] = P(N \geq s) = \sum_{n=s}^{+\infty} q_n = q_0 \sum_{n=s}^{+\infty} \frac{(\lambda/\mu)^n}{s^{n-s}s!} = \frac{q_0}{s!}\left(\frac{\lambda}{\mu}\right)^s \sum_{n=s}^{+\infty} \varrho_s^{n-s}$$
$$= q_0 \, \frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)}. \tag{5.10}$$

Sostituendo q_0, la (5.10) può essere così riscritta:

$$C[s,\lambda/\mu] = \frac{\dfrac{(\lambda/\mu)^s}{s!}}{\dfrac{(\lambda/\mu)^s}{s!} + (1-\varrho_s)\displaystyle\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!}}.$$

Come vedremo nel seguito i principali parametri prestazionali del sistema $M/M/s$ coinvolgono la formula C di Erlang. Si può infatti mostrare che

$$E(N) = \sum_{n=1}^{+\infty} n\, q_n = \frac{\lambda}{\mu} + \frac{\varrho_s}{1-\varrho_s}\, C[s,\lambda/\mu]\,. \tag{5.11}$$

Dalla prima legge di Little possiamo ricavare il tempo medio di attesa nel sistema

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{1}{\mu} + \frac{1}{s\,\mu}\,\frac{(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\, q_0 = \frac{1}{\mu} + \frac{1}{s\,\mu\,(1-\varrho_s)}\, C[s,\lambda/\mu]. \tag{5.12}$$

Inoltre, in condizioni di equilibrio statistico il tempo medio di permanenza nella fila di attesa è

$$E(Q) = E(W) - E(S) = E(W) - \frac{1}{\mu} = \frac{1}{s\,\mu}\,\frac{(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\, q_0$$
$$= \frac{1}{s\,\mu\,(1-\varrho_s)}\, C[s,\lambda/\mu], \tag{5.13}$$

e dalla seconda legge di Little segue che il numero medio di utenti nella fila di attesa è:

$$E(N_q) = \sum_{n=s}^{+\infty}(n-s)\, q_n = \lambda^*\, E(Q) = \frac{\lambda}{s\,\mu}\,\frac{(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\, q_0$$
$$= \frac{\varrho_s\,(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\, q_0 = \frac{\lambda}{s\,\mu\,(1-\varrho_s)}\, C[s,\lambda/\mu] = \frac{\varrho_s}{1-\varrho_s}\, C[s,\lambda/\mu]. \tag{5.14}$$

Infine, il numero medio di utenti in servizio (o equivalentemente il numero medio di servitori occupati)

$$E(N_s) = \lambda^*\, E(S) = \frac{\lambda}{\mu} < s.$$

$$\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \qquad \mu_n = \begin{cases} n\,\mu & n = 1, 2, \ldots, s-1 \\ s\,\mu, & n = s, s+1, \ldots \end{cases}$$

$$\varrho_s = \frac{\lambda}{s\,\mu} < 1 \qquad \text{(condizione di equilibrio statistico)}$$

$$q_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)} \right]^{-1},$$

$$q_n = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \begin{cases} q_0\, \dfrac{(\lambda/\mu)^n}{n!}, & n = 1, 2, \ldots, s \\ q_0\, \dfrac{(\lambda/\mu)^n}{s^{n-s}\, s!}, & n = s+1, s+2, \ldots \end{cases}$$

$$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad a = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu}, \qquad \varrho^* = \frac{a}{s} = \frac{\lambda}{s\,\mu} = \varrho_s$$

$$C[s, \lambda/\mu] = P(N \geq s) = q_0\, \frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)} \qquad \text{(formula C di Erlang)}$$

$$E(N) = \frac{\lambda}{\mu} + \frac{\varrho_s}{1-\varrho_s}\, C[s, \lambda/\mu], \qquad E(W) = \frac{1}{\mu} + \frac{C[s, \lambda/\mu]}{s\,\mu\,(1-\varrho_s)}$$

$$E(N_q) = \frac{\varrho_s}{1-\varrho_s}\, C[s, \lambda/\mu], \qquad E(Q) = \frac{1}{s\,\mu\,(1-\varrho_s)}\, C[s, \lambda/\mu]$$

$$E(N_s) = \frac{\lambda}{\mu}, \qquad E(S) = \frac{1}{\mu}$$

Tabella 5.3: Parametri prestazionali del sistema di servizio $M/M/s$

coincide con l'intensità di traffico del centro di servizio.

Nella Tabella 5.3 sono riportati i principali parametri prestazionali del sistema $M/M/s$.

Risultati per il sistema M/M/1 e confronti

Supponiamo ora che $s = 1$, ossia riconsideriamo il sistema $M/M/1$. Per $s = 1$ si ha $C[s, \lambda/\mu] = C[1, \lambda/\mu] = \varrho = \lambda/\mu$, che coincide con la probabilità che almeno un utente sia presente nel sistema. Inoltre, per $s = 1$ la (5.11) fornisce il numero medio di utenti presenti nel sistema $M/M/1$, ossia $E(N) = \varrho/(1-\varrho)$, la (5.12) fornisce il tempo medio di attesa nel sistema per il modello $M/M/1$, ossia $E(W) = 1/[\mu\,(1-\varrho)]$, la (5.13) fornisce il tempo di permanenza nella fila di attesa per il sistema $M/M/1$, ossia $E(Q) = \varrho/[\mu\,(1-\varrho)]$, e la (5.14) permette di ottenere il numero medio di utenti presenti nella fila di attesa del sistema $M/M/1$, ossia $E(N_q) = \varrho^2/(1-\varrho)$.

Il confronto tra i sistemi $M/M/1$ e $M/M/2$ può anche essere esteso al confronto tra i sistemi $M/M/1$ e $M/M/s$. Infatti, si può dimostrare che *a parità del processo degli arrivi è meno conveniente scegliere un sistema con s servitori ognuno dei quali ha tempo medio di servizio pari a* $1/\mu$ *rispetto ad un sistema*

$M/M/1$ avente tempo medio di servizio pari a $1/(s\,\mu)$, ossia un sistema con un servitore s volte più veloce rispetto a ciascuno dei singoli servitori del sistema $M/M/s$.

Esempio 5.2 Consideriamo un bar in cui gli utenti arrivano secondo un processo di Poisson con frequenza media di 2 utenti al minuto. Supponiamo che i tempi di servizio di ogni servitore siano distribuiti esponenzialmente con tempo medio di servizio di 40 secondi. Se un sistema di servizio $M/M/s$ modella gli utenti del bar, determinare il numero di servitori necessari e sufficienti affinché il sistema non si congestioni e calcolare i principali parametri prestazionali del sistema.

In questo caso risulta

$$\lambda = 2 \text{ utenti al minuto}, \qquad \mu = \frac{60}{40} = \frac{3}{2} \text{ utenti al minuto},$$

e quindi l'intensità di traffico è $a = \lambda/\mu = 4/3$. Dalla (5.9) segue che il numero di servitori necessari e sufficienti per evitare la congestione del bar è $s = 2$. Valutiamo ora i principali parametri prestazionali del sistema $M/M/2$ con $\lambda = 2$ e $\mu = 3/2$ utilizzando la Tabella 5.1. Il fattore di utilizzazione del sistema è $\varrho^* = \lambda/(2\mu) = 2/3 = \varrho_2$. Nella situazione di equilibrio statistico statistico, la distribuzione di equilibrio è

$$q_0 = \frac{1-\varrho_2}{1+\varrho_2} = \frac{1}{5} = 0.2, \qquad q_n = 2\,\frac{1-\varrho_2}{1+\varrho_2}\,\varrho_2^n = \frac{2}{5}\left(\frac{2}{3}\right)^n \quad (n = 1, 2, \ldots).$$

La probabilità che entrambi i servitori siano occupati è fornita dalla formula C di Erlang, ossia:

$$C[2, \lambda/\mu] = P(N \geq 2) = \frac{2\,\varrho_2^2}{1+\varrho_2} = \frac{8}{15} = 0.533.$$

Il numero medio di utenti nel sistema e nella fila di attesa sono:

$$E(N) = \frac{2\,\varrho_2}{1-\varrho_2^2} = \frac{12}{5} \text{ utenti}, \quad E(N_q) = \frac{2\,\varrho_2^3}{1-\varrho_2^2} = \frac{16}{15} \text{ utenti}.$$

Utilizzando le leggi di Little si ottengono i tempi medi di attesa nel sistema e in fila di attesa:

$$E(W) = \frac{E(N)}{\lambda} = \frac{6}{5} = 1.2 \text{ minuti} \qquad E(Q) = \frac{E(N_q)}{\lambda} = \frac{8}{15} = 0.533 \text{ minuti}.$$

Ovviamente il numero medio di utenti in servizio è

$$E(N_s) = E(N) - E(N_q) = \frac{\lambda}{\mu} = \frac{4}{3} = 1.333 \text{ utenti},$$

e il tempo medio di servizio per servitore è:

$$E(S) = E(W) - E(Q) = \frac{2}{3} = 0.667 \text{ minuti}.$$

♢

5.4 Sistema di servizio $M/M/s/s$

Consideriamo un sistema di servizio con un'unica fila di attesa, s servitori, capacità finita s, rappresentato in Figura 5.5.

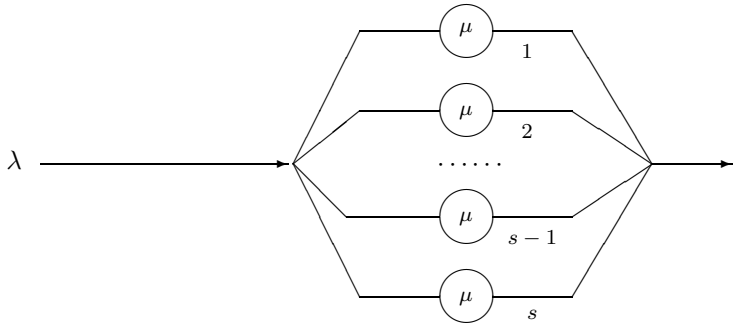

Figura 5.5: Sistema di servizio $M/M/s/s$.

La prima M significa che gli arrivi si verificano secondo un processo di Poisson di parametro λ. La seconda M significa che i tempi di servizio per ognuno dei servitori sono indipendenti ed esponenzialmente distribuiti con valore medio $1/\mu$. Se si verifica un arrivo quando tutti gli s servitori sono occupati la richiesta di servizio è rifiutata e quindi non ha effetto sul sistema. Se invece un utente in arrivo trova dei servitori liberi sceglie a caso uno di essi per essere servito.

Questo sistema di servizio è stato inizialmente proposto dal matematico e statistico danese *Agner Krarup Erlang* come modello per analizzare il comportamento di un centralino telefonico caratterizzato da s linee disponibili. Le chiamate che arrivano e trovano tutte le s linee occupate sono rifiutate e quindi vengono perdute.

Nel 1908 Erlang venne assunto dalla compagnia telefonica di Copenaghen e nel 1909 pubblicò il suo lavoro "*The theory of probability and telephone conversations*", che costituisce il primo studio dettagliato sul traffico telefonico. Nel 1917 tale studioso introdusse la variabile aleatoria di Erlang e le due formule (B e C) usate quasi subito da tutte le altre compagnie telefoniche nella progettazione dei centralini telefonici per prevenire sovraccarichi della linea o per calcolare il numero di linee e di personale necessari.

Denotiamo con $N(t)$ il numero di utenti presenti nel sistema $M/M/s/s$ al tempo t. Il processo stocastico $\{N(t), t \geq 0\}$ è descrivibile mediante un processo di nascita–morte caratterizzato da parametri

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \ldots, s-1 \\ 0, & n = s, s+1, \ldots \end{cases} \tag{5.15}$$

$$\mu_n = \begin{cases} n\,\mu, & n = 1, 2, \ldots, s \\ 0, & n = s+1, s+2, \ldots \end{cases}$$

Poiché il sistema di servizio $M/M/s/s$ è a capacità finita, raggiunge sempre una situazione di equilibrio statistico. Vogliamo ora determinare tale distribuzione. Facendo uso di (5.15) in (3.17) e ponendo $\varrho_s = \lambda/(s\mu)$ si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{s} \frac{\lambda^n}{\mu^n\, n!} = \sum_{n=0}^{s} \frac{(s\,\varrho_s)^n}{n!}.$$

Proposizione 5.3 *In condizioni di equilibrio statistico per il sistema $M/M/s/s$ si ha:*

$$\begin{aligned} q_0 &= P(N=0) = \Big[\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!}\Big]^{-1} \\ q_n &= P(N=n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \frac{(\lambda/\mu)^n}{n!}\, q_0\,, \qquad (n = 1, 2, \ldots, s). \end{aligned} \tag{5.16}$$

Tale distribuzione di probabilità è detta *distribuzione di Poisson troncata* di parametro λ/μ.

La probabilità che un utente in arrivo venga rifiutato può essere così ottenuta:

$$B[s, \lambda/\mu] = P(N=s) = q_s = \frac{\dfrac{(\lambda/\mu)^s}{s!}}{\displaystyle\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!}} \tag{5.17}$$

ed è detta *formula B di Erlang*. Come vedremo nel seguito i principali parametri prestazionali del sistema $M/M/s/s$ coinvolgono la formula B di Erlang. In primo luogo, si nota che

$$q_n = \frac{s!}{n!}\,(s\,\varrho_s)^{n-s}\, B[s, \lambda/\mu] \qquad (n = 0, 1, \ldots, s).$$

La frequenza media di arrivo e la frequenza media di partenza sono:

$$\begin{aligned} \lambda^* &= \sum_{n=0}^{s-1} \lambda_n\, q_n = \lambda \sum_{n=0}^{s-1} q_n = \lambda\,(1 - q_s) = \lambda\,\{1 - B[s, \lambda/\mu]\}, \\ \mu^* &= \mu. \end{aligned}$$

Si nota che λ^* è il prodotto della frequenza di arrivo λ e della probabilità che l'utente in arrivo non venga rifiutato. Pertanto l'intensità di traffico relativa al centro di servizio è

$$a = \frac{\lambda^*}{\mu^*} = \frac{\lambda\,(1 - q_s)}{\mu} = \frac{\lambda}{\mu}\,\big\{1 - B[s, \lambda/\mu]\big\}$$

e quindi il fattore di utilizzazione del sistema è

$$\varrho^* = \frac{\lambda^*}{s\,\mu^*} = \frac{\lambda\,(1-q_s)}{s\mu} = \frac{\lambda}{s\,\mu}\,\{1 - B[s, \lambda/\mu]\}.$$

Nella situazione di equilibrio statistico il numero medio di utenti presenti nella fila di attesa ed il tempo medio di permanenza in coda sono nulli, ossia $E(N_q) = 0$ e $E(Q) = 0$. Poiché il tempo di attesa nel sistema coincide con il tempo di servizio segue che la variabile aleatoria W è distribuita esponenzialmente con valore medio $E(W) = 1/\mu$ e $E(N) = E(N_s)$.

Quindi il numero medio di utenti nel sistema, che coincide con il numero medio di clienti in servizio (numero medio di servitori occupati) è

$$E(N) = E(N_s) = \lambda^*\,E(W) = \frac{\lambda\left\{1 - B[s, \lambda/\mu]\right\}}{\mu}.$$

Si nota nuovamente che il numero medio di clienti nel sistema (ossia il numero medio di servitori occupati) coincide con l'intensità di traffico relativa al centro di servizio.

I principali risultati ottenuti per il modello $M/M/s/s$ sono elencati in Tabella 7.

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \ldots, s-1 \\ 0, & n = s, s+1, \ldots \end{cases} \qquad \mu_n = \begin{cases} n\,\mu, & n = 1, 2, \ldots, s \\ 0, & n = s+1, s+2, \ldots \end{cases}$$

$$q_0 = \left[\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!}\right]^{-1} \qquad q_n = \frac{(\lambda/\mu)^n}{n!}\,q_0\,, \qquad (n = 1, 2, \ldots, s)$$

$$\lambda^* = \lambda\,\{1 - B[s, \lambda/\mu]\}, \qquad \mu^* = \mu,$$

$$a = \frac{\lambda}{\mu}\,\{1 - B[s, \lambda/\mu]\}, \qquad \varrho^* = \frac{\lambda}{s\,\mu}\,\{1 - B[s, \lambda/\mu]\}$$

$$B[s, \lambda/\mu] = q_s = \frac{(\lambda/\mu)^s/s!}{\sum_{n=0}^{s} (\lambda/\mu)^n/n!} \qquad \text{(formula B di Erlang)}$$

$$E(N_q) = 0, \qquad E(Q) = 0$$

$$E(N) = E(N_s) = \frac{\lambda}{\mu}\,\{1 - B[s, \lambda/\mu]\}$$

$$E(W) = E(S) = \frac{1}{\mu}$$

Tabella 5.4: Parametri prestazionali del sistema di servizio $M/M/s/s$.

Esempio 5.3 Si consideri un centralino telefonico con due linee disponibili che non consenta l'attesa di nessuna chiamata. Si supponga che arrivino con una

frequenza media di 2 chiamate al minuto e che la durata media di una telefonata è di 40 secondi. Se il sistema di servizio $M/M/2/2$ modella il centralino considerato, calcolare i principali parametri prestazionali del sistema.

In questo caso risulta

$$\lambda = 2 \text{ telefonate al minuto}, \qquad \mu = \frac{60}{40} = \frac{3}{2} \text{ telefonate al minuto},$$

e quindi l'intensità di traffico è $a = \lambda/\mu = 4/3$. La probabilità che in condizioni di equilibrio si abbiano k telefonate ($k = 0, 1, 2$) è:

$$\begin{aligned} q_0 &= \Big[1 + \frac{(\lambda/\mu)}{1!} + \frac{(\lambda/\mu)^2}{2!}\Big]^{-1} = \Big[1 + \frac{4}{3} + \frac{8}{9}\Big]^{-1} = \frac{9}{29}, \\ q_1 &= \frac{\lambda}{\mu}\, q_0 = \frac{4}{3} \cdot \frac{9}{29} = \frac{12}{29}, \\ q_2 &= \frac{1}{2}\Big(\frac{\lambda}{\mu}\Big)^2 q_0 = \frac{16}{18} \cdot \frac{9}{29} = \frac{8}{29}\,. \end{aligned}$$

La formula B di Erlang fornisce la probabilità che un utente in arrivo venga rifiutato:

$$B[s, \lambda/\mu] = B[2, 4/3] = q_2 = \frac{8}{29} = 0.2759.$$

Il coefficiente di utilizzazione del sistema è quindi:

$$\varrho^* = \varrho_s\{1 - B[s, \lambda/\mu]\} = \frac{2}{3} \cdot \frac{21}{29} = \frac{14}{29} = 0.4828.$$

In condizioni di equilibrio, il numero medio di linee occupate e la durata di ogni telefonata sono:

$$\begin{aligned} E(N) &= \frac{\lambda}{\mu}\,\{1 - B[s, \lambda/\mu]\} = \frac{4}{3}\Big(1 - \frac{8}{29}\Big) = \frac{4}{3} \cdot \frac{21}{29} = \frac{28}{29} = 0.9656 \text{ telefonate}, \\ E(W) &= \frac{E(N)}{\lambda} = \frac{2}{3} \cdot \frac{28}{29} = \frac{56}{87} = 0.6436 \text{ minuti}. \end{aligned}$$

Si nota che in media è occupata una sola linea telefonica. ♢

5.5 Sistema di servizio $M/M/\infty$

Consideriamo un sistema di servizio a capacità infinita con un'unica fila di attesa e infiniti servitori. Supponiamo che i tempi di interarrivo siano indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$ e che i tempi di servizio per ognuno dei servitori siano indipendenti e distribuiti esponenzialmente con valore medio $1/\mu$. Poiché esistono infiniti servitori un utente che arriva può essere immediatamente servito, ossia il suo tempo di attesa nel sistema è uguale al suo tempo di servizio. Tale sistema è noto in letteratura come sistema $M/M/\infty$.

Il sistema $M/M/\infty$ è descrivibile mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri

$$\begin{aligned} \lambda_n &= \lambda \qquad (n = 0, 1, \ldots) \\ \mu_n &= n\,\mu \qquad (n = 1, 2, \ldots). \end{aligned} \tag{5.18}$$

Tale sistema costituisce una buona approssimazione per molti sistemi reali del tipo self–service, quali grandi parcheggi, cinema, supermarket, ...; in tali casi si può ipotizzare che un utente in arrivo sia immediatamente servito.

Facendo uso di (5.18) in (3.17) si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{+\infty} \frac{1}{n!} \Big(\frac{\lambda}{\mu}\Big)^n = \exp\Big\{\frac{\lambda}{\mu}\Big\}.$$

Essendo la serie sempre convergente, il sistema $M/M/\infty$ raggiunge sempre una situazione di equilibrio statistico e si ha

$$\begin{aligned} q_0 &= P(N = 0) = \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \\ q_n &= P(N = n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \frac{1}{n!} \Big(\frac{\lambda}{\mu}\Big)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \qquad (n = 1, 2, \ldots). \end{aligned}$$

Proposizione 5.4 *Per il sistema $M/M/\infty$, in condizioni di equilibrio si ha:*

$$q_n = P(N = n) = \frac{1}{n!} \Big(\frac{\lambda}{\mu}\Big)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \qquad (n = 1, 2, \ldots), \tag{5.19}$$

ossia una funzione di probabilità di Poisson di parametro λ/μ.

Il valore medio e la varianza del numero di utenti nel sistema sono:

$$E(N) = \frac{\lambda}{\mu}\,, \qquad \mathrm{Var}(N) = \frac{\lambda}{\mu}\cdot$$

Poiché nel sistema $M/M/\infty$ il tempo medio di interarrivo è $E(T) = 1/\lambda$ e il tempo medio di servizio per ognuno dei servitori è $E(S) = 1/\mu$ si ha

$$\lambda^* = \frac{1}{E(T)} = \lambda, \qquad \mu^* = \frac{1}{E(S)} = \mu.$$

L'intensità di traffico relativa al centro di servizio è quindi:

$$a = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu} < +\infty$$

e il fattore di utilizzazione del sistema è $\varrho^* = 0$.

Nella situazione di equilibrio statistico il numero medio di utenti presenti nella fila di attesa e il tempo medio di permanenza in coda sono nulli, ossia

$E(N_q) = 0$ e $E(Q) = 0$. Poiché il tempo di attesa nel sistema coincide con il tempo di servizio, la variabile aleatoria W è distribuita esponenzialmente con valore medio $E(W) = 1/\mu$. Pertanto la densità di probabilità di W è:

$$f_W(t) = \begin{cases} \mu\, e^{-\mu\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Il numero medio di utenti nel sistema, coincidente con il numero medio di utenti in servizio (numero medio di servitori occupati), è ottenibile dalle leggi di Little:

$$E(N) = E(N_s) = \lambda^*\, E(W) = \lambda^*\, E(S) = \frac{\lambda}{\mu}\,.$$

Si nota nuovamente che il numero medio di utenti nel sistema (ossia il numero medio di servitori occupati) coincide con l'intensità di traffico relativa al centro di servizio.

Nel sistema $M/M/\infty$ i periodi di ozio del centro di servizio, che possono essere riguardati come tempi residui dei tempi di interarrivo, sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$. Pertanto la densità di probabilità della variabile aleatoria I descrivente un periodo di ozio è:

$$f_I(t) = \begin{cases} \lambda\, e^{-\lambda\, t}, & t > 0 \\ 0, & \text{altrimenti} \end{cases}$$

e quindi

$$E(B) = \frac{(1 - q_0)\, E(I)}{q_0} = \frac{(1 - q_0)}{\lambda\, q_0} = \frac{1 - e^{-\lambda/\mu}}{\lambda\, e^{-\lambda/\mu}} = \frac{1}{\lambda}\left(e^{\lambda/\mu} - 1\right).$$

Si nota che $E(B) > E(I)$, ossia il tempo medio di occupazione è maggiore del tempo medio di ozio del servitore.

I principali parametri prestazionali del sistema $M/M/\infty$ sono indicati in Tabella 5.5.

Esempio 5.4 Si desidera modellare il numero di utenti connessi simultaneamente ad una rete con un sistema $M/M/\infty$. Gli utenti accedono alla rete secondo un processo di Poisson con una frequenza media di 500 utenti all'ora e restano connessi alla rete in media per 20 minuti. Determinare i parametri prestazionali del sistema.
In questo caso

$$\lambda = 500 \text{ utenti all'ora}, \qquad \mu = \frac{60}{20} = 3 \text{ utenti all'ora},$$

e quindi l'intensità di traffico è $a = \lambda/\mu = 500/3$. Dalla (5.9) segue che il numero di servitori necessari e sufficienti in un sistema $M/M/s$ per evitare la congestione è $s = 167$, ossia occorrono un numero molto elevato di servitori. Per descrivere la rete si può quindi utilizzare un sistema $M/M/\infty$, con $\lambda = 500$ e $\mu = 3$. Dalla Tabella 5.4 segue che la distribuzione di equilibrio è

$$q_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\} = \frac{1}{n!}\left(\frac{500}{3}\right)^n \exp\left\{-\frac{500}{3}\right\} \qquad (n = 0, 1, \ldots).$$

$$\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \qquad \mu_n = n\,\mu \qquad (n = 1, 2, \ldots)$$

$$q_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad a = \frac{\lambda}{\mu}, \qquad \varrho^* = 0$$

$$E(N) = \frac{\lambda}{\mu}, \qquad E(W) = \frac{1}{\mu}$$

$$E(N_q) = 0, \qquad E(Q) = 0$$

$$E(N_s) = E(N) = \frac{\lambda}{\mu}, \qquad E(S) = \frac{1}{\mu}$$

$$E(I) = \frac{1}{\lambda}, \qquad E(B) = \frac{1}{\lambda}\left[\exp\left\{\frac{\lambda}{\mu}\right\} - 1\right]$$

Tabella 5.5: Parametri prestazionali del sistema di servizio $M/M/\infty$.

Il numero medio di connessioni è $E(N) = E(N_s) = \lambda/\mu = 500/3 = 166,667$ e il tempo medio di connessione è $E(W) = E(S) = 1/\mu = 1/3$ di ora, ossia 20 minuti. ♢

5.6 Sistema con accelerazione del servizio

Il processo di nascita morte caratterizzato da parametri (5.18), ossia da parametri $\lambda_n = \lambda$ $(n = 0, 1, \ldots)$ e $\mu_n = n\,\mu$ $(n = 1, \ldots)$, si può anche interpretare come descrivente un sistema di servizio a capacità infinita con un'unica fila di attesa e un unico servitore che accelera il suo servizio all'aumentare della lunghezza della coda in maniera tale da soddisfare tutte le richieste degli utenti.

Il sistema di servizio con accelerazione del servizio ammette la stessa distribuzione di equilibrio del sistema $M/M/\infty$, ossia una distribuzione di Poisson di parametro λ/μ, fornita in (5.19). Il valore medio e la varianza del numero di utenti nel sistema sono quindi:

$$E(N) = \frac{\lambda}{\mu}, \qquad \mathrm{Var}(N) = \frac{\lambda}{\mu}\cdot$$

Anche se il sistema $M/M/\infty$ e quello con accelerazione del servizio possiedono la stessa distribuzione di equilibrio, hanno alcuni parametri prestazionali diversi. Infatti, per il sistema con accelerazione del servizio, la frequenza media di arrivo per unità di tempo è uguale a quella del sistema $M/M/\infty$, ossia $\lambda^* = \lambda$, mentre

la frequenza media di partenza per unità di tempo è:

$$\mu^* = \frac{1}{1-q_0}\sum_{n=1}^{+\infty}\mu_n\, q_n = \frac{\mu}{1-q_0}\sum_{n=1}^{+\infty} n\, q_n = \frac{\mu}{1-q_0}\,\frac{\lambda}{\mu} = \frac{\lambda}{1-\exp\{-\lambda/\mu\}}.$$

L'intensità di traffico, che coincide con il fattore di utilizzazione del sistema, è quindi:

$$a = \varrho^* = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\} < 1.$$

Dalla prima legge di Little segue che il tempo medio di attesa nel sistema è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{1}{\mu}.$$

Il numero medio di utenti in fila di attesa è

$$E(N_q) = \sum_{n=1}^{+\infty}(n-1)\, q_n = E(N) - (1-q_0) = \frac{\lambda}{\mu} - 1 + \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

e quindi dalla seconda legge di Little si ottiene il tempo medio di permanenza nella fila di attesa:

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{1}{\mu} - \frac{1}{\lambda}\Big[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\Big].$$

Il numero medio di utenti in servizio è

$$E(N_s) = E(N) - E(N_q) = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

e coincide con l'intensità di traffico. Pertanto, dalla terza legge di Little otteniamo:

$$E(S) = \frac{E(N_s)}{\lambda^*} = \frac{1}{\lambda}\Big[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\Big].$$

I principali parametri prestazionali del sistema con accelerazione del servizio sono indicati in Tabella 5.6.

Si nota che alcuni parametri prestazionali del sistema $M/M/\infty$ e di quello con accelerazione del servizio sono differenti, anche se entrambi i modelli sono caratterizzati dalla stessa distribuzione di equilibrio; ciò dipende dal fatto che nel modello con accelerazione del servizio è previsto un unico servitore, mentre nel modello $M/M/\infty$ si considerano infiniti servitori. Quindi, prima di determinare i parametri prestazionali di un sistema di servizio occorre sempre precisare il numero di servitori.

5.7 Sistema di servizio con scoraggiamento

Consideriamo un sistema di servizio a capacità infinita con unica fila di attesa. Supponiamo che gli utenti siano scoraggiati da una lunga coda e che il tempo di

$$\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \qquad \mu_n = n\,\mu \qquad (n = 1, 2, \ldots)$$

$$q_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \lambda, \qquad \mu^* = \frac{\lambda}{1 - \exp\{-\lambda/\mu\}}, \qquad \varrho^* = 1 - \exp\left\{-\frac{\lambda}{\mu}\right\} < 1$$

$$E(N) = \frac{\lambda}{\mu}, \qquad E(W) = \frac{1}{\mu},$$

$$E(N_q) = \frac{\lambda}{\mu} - 1 + \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad E(Q) = \frac{1}{\mu} - \frac{1}{\lambda}\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right]$$

$$E(N_s) = 1 - \exp\left\{-\frac{\lambda}{\mu}\right\} = \varrho^*, \qquad E(S) = \frac{E(N_s)}{\lambda^*} = \frac{1}{\lambda}\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right]$$

Tabella 5.6: Parametri prestazionali del sistema con unico servitore che velocizza il suo servizio.

servizio del generico utente sia distribuito esponenzialmente con valore medio $1/\mu$.

Tale sistema può essere descritto mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri

$$\begin{aligned} \lambda_n &= \frac{\lambda}{n+1} \qquad (n = 0, 1, \ldots) \\ \mu_n &= \mu \qquad (n = 1, 2, \ldots). \end{aligned} \tag{5.20}$$

Vogliamo ora vedere in quali condizioni tale sistema raggiunge una situazione di equilibrio statistico. Facendo uso di (5.20) in (3.17) si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{+\infty} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n = \exp\left\{\frac{\lambda}{\mu}\right\}.$$

Poiché tale serie è sempre convergente, il sistema considerato raggiunge sempre una situazione di equilibrio statistico e risulta

$$q_0 = \exp\left\{-\frac{\lambda}{\mu}\right\},$$

$$q_n = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\} \qquad (n = 1, 2, \ldots).$$

Si è ottenuta la stessa distribuzione di equilibrio (5.19) del sistema $M/M/\infty$, ossia una distribuzione di Poisson di parametro λ/μ. Pertanto, nella situazione

di equilibrio statistico, il valore medio e la varianza e il coefficiente di variazione del numero di utenti presenti nel sistema sono

$$E(N) = \frac{\lambda}{\mu}\,, \qquad \mathrm{Var}(N) = \frac{\lambda}{\mu}\cdot$$

$$\lambda_n = \frac{\lambda}{n+1} \qquad (n = 0, 1, \ldots) \qquad \mu_n = \mu \qquad (n = 1, 2, \ldots)$$

$$q_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \mu\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right], \qquad \mu^* = \mu, \qquad \varrho^* = 1 - \exp\left\{-\frac{\lambda}{\mu}\right\}$$

$$E(N) = \frac{\lambda}{\mu}\,, \qquad E(W) = \frac{\lambda}{\mu^2}\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right]^{-1}$$

$$E(N_q) = \frac{\lambda}{\mu} - 1 + \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad E(Q) = \frac{\lambda}{\mu^2}\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right]^{-1} - \frac{1}{\mu}$$

$$E(N_s) = 1 - \exp\left\{-\frac{\lambda}{\mu}\right\} = \varrho^*, \qquad E(S) = \frac{1}{\mu}\cdot$$

Tabella 5.7: Parametri prestazionali del sistema di servizio con scoraggiamento degli utenti e unico servitore.

Anche se il sistema $M/M/\infty$ e quello con scoraggiamento possiedono la stessa distribuzione di equilibrio, essi sono fondamentalmente diversi sia nell'evoluzione transiente sia relativamente ai parametri prestazionali. Infatti, si ha:

$$\begin{aligned}
\lambda^* &= \sum_{n=0}^{+\infty} \lambda_n\, q_n = \sum_{n=0}^{+\infty} \frac{\lambda}{n+1}\,\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\} \\
&= \mu\,\exp\left\{-\frac{\lambda}{\mu}\right\} \sum_{n=0}^{+\infty} \frac{1}{(n+1)!}\left(\frac{\lambda}{\mu}\right)^{n+1} = \mu\,\exp\left\{-\frac{\lambda}{\mu}\right\} \sum_{k=1}^{+\infty} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \\
&= \mu\,\exp\left\{-\frac{\lambda}{\mu}\right\}\left[\exp\left\{\frac{\lambda}{\mu}\right\} - 1\right] = \mu\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right], \\
\mu^* &= \mu,
\end{aligned}$$

da cui l'intensità di traffico, che coincide con il fattore di utilizzazione del sistema, è

$$\varrho^* = \frac{\lambda^*}{\mu^*} = 1 - \exp\left\{-\frac{\lambda}{\mu}\right\}.$$

Dalla prima legge di Little si ricava:

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{\lambda}{\mu^2}\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right]^{-1}.$$

Osserviamo inoltre che Il numero medio di utenti in fila di attesa è

$$E(N_q) = \sum_{n=1}^{+\infty}(n-1)\,q_n = E(N) - (1-q_0) = \frac{\lambda}{\mu} - 1 + \exp\Big\{-\frac{\lambda}{\mu}\Big\},$$

da cui, facendo ricorso alla seconda legge di Little, si ottiene il tempo medio di permanenza in fila di attesa:

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{\lambda}{\mu^2}\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right]^{-1} - \frac{1}{\mu}$$

Infine, il numero medio di utenti in servizio è dato da

$$E(N_s) = E(N) - E(N_q) = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

che coincide con l'intensità di traffico. Dalla terza legge di Little segue infine che

$$E(S) = \frac{E(N_s)}{\lambda^*} = \frac{1}{\mu}.$$

I principali parametri prestazionali del sistema di servizio con scoraggiamento degli utenti e unico servitore sono forniti in Tabella 5.7.

Si nota che il modello con accelerazione del servizio e quello con scoraggiamento hanno la stessa intensità di traffico, anche se le frequenze medie λ^* e μ^* sono differenti. Poiché le leggi di Little coinvolgono le frequenze medie di arrivo, i due sistemi hanno differenti parametri prestazionali

In conclusione, la distribuzione di equilibrio non è sufficiente per comprendere tutte le caratteristiche di un sistema di servizio. Dalle Tabelle 5.4, 5.5 e 5.7 si nota che il sistema $M/M/\infty$, il sistema con accelerazione del servizio e quello con scoraggiamento, anche se caratterizzati dalla stessa distribuzione di equilibrio, hanno parametri prestazionali fondamentalmente diversi. Ciò è dovuto al numero di servitori (infiniti nel modello $M/M/\infty$ e unico nei modelli con accelerazione del servizio e con scoraggiamento degli utenti) e a differenti frequenze medie di arrivo e di partenza per unità di tempo.

Capitolo 6

Simulazione

6.1 Introduzione alla simulazione

La teoria delle file di attesa ricorre a *modelli probabilistici* per ottenere i parametri prestazionali del sistema di servizio. Spesso si rivela difficile studiare tali modelli sia per le caratteristiche delle distribuzioni degli intervalli di interarrivo e dei tempi di servizio e sia per natura della disciplina di servizio. Inoltre, con i modelli probabilistici si riesce ad analizzare il sistema in condizioni di equilibrio statistico, mentre spesso si è interessati al comportamento del sistema di servizio anche nella sua fase transiente.

Per superare tali difficoltà si ricorre spesso a modelli di sistemi di servizio che utilizzano *tecniche di simulazione*. La simulazione consiste nel *riprodurre al computer il comportamento del sistema in esame*. Si basa sulla definizione di un modello, detto modello di simulazione, che descriva l'evoluzione del sistema nel tempo. Mediante la simulazione è possibile osservare il *comportamento dinamico del sistema* fornendo informazioni sulle sue prestazioni. Infatti, la simulazione permette di ottenere *stime* (ad esempio, *medie e varianze campionarie*) del tempo di permanenza nella fila di attesa, del tempo di attesa nel sistema, del numero di utenti in fila di attesa e nel sistema, del tempo di ozio e di occupazione del centro di servizio.

Entrambi gli approcci (probabilistico e di simulazione) richiedono l'utilizzazione di un modello che permetta in base ai parametri di input di ottenere gli indici di prestazione del sistema.

Con un *modello probabilistico* di un sistema di servizio ci si prefigge di ottenere *soluzioni analitiche* e pertanto le ipotesi di base del modello sono semplificate in maniera tale da superare le difficoltà di natura matematica. Nel modello probabilistico si ottengono delle formule che permettono di esprimere gli indici di prestazione del sistema in funzione dei parametri di input (tempi medi di interarrivo, tempi medi di servizio, numero di servitori, ...) e delle caratteristiche

del sistema. Tali formule sono utilizzabili in maniera veloce ed efficiente per un ampio intervallo di valori dei parametri di input e permettono di interpretare qualitativamente il comportamento del sistema e di individuare le condizioni sui parametri che garantiscono il raggiungimento dell'equilibrio statistico.

In un *modello di simulazione*, invece, si possono includere un maggiore numero di caratteristiche significative, spesso rispecchiando meglio il comportamento del sistema reale. La simulazione comunque necessita di *lunghi periodi di esecuzione* per ottenere stime degli indici di prestazione e richiede anche indagini approfondite per l'individuazione delle condizioni che garantiscono il raggiungimento dell'equilibrio statistico.

Non è sempre semplice decidere se utilizzare modelli probabilistici o di simulazione per analizzare un sistema di servizio. I modelli di simulazione si rivelano particolarmente utili nello studio di sistemi di servizio complessi. Il loro utilizzo deve basarsi su due elementi essenziali: "*adeguatezza*" e "*semplicità d'uso*". Uno stesso sistema può essere descritto utilizzando diversi tipi di modelli; pertanto, un modello è tanto più adeguato alla descrizione del sistema di servizio reale quanto meglio rappresenta gli aspetti del sistema che sono di interesse per chi sta effettuando lo studio. La simulazione è quindi un metodo alternativo per la descrizione di un sistema di servizio e può essere utile se riesce a fornire soluzioni significative e di facile interpretazione da parte di coloro che saranno addetti all'utilizzazione del modello ad un costo competitivo rispetto ad altre tecniche di natura probabilistica.

Quando si decide di adottare tecniche di simulazione occorre pianificare l'esperimento in più fasi successive:

(i) formulazione del problema e del modello di simulazione;

(ii) acquisizione dei dati del sistema reale;

(iii) stima e verifica dei parametri e delle caratteristiche operative del sistema reale;

(iv) formulazione del programma di simulazione;

(v) progettazione degli esperimenti e analisi dei risultati.

(i) **Formulazione del problema e del modello di simulazione**
Formulare il problema significa *fissare gli obiettivi di studio* e *stabilire dei criteri per esaminare le soluzioni al problema*. Gli *obiettivi* di un esperimento di simulazione sono essenzialmente di due tipi:

(a) studio del problema di dimensionamento

(b) studio degli effetti del cambiamento dei parametri o delle caratteristiche funzionali sul comportamento del sistema

Relativamente al punto *(a)* ci si pone il problema di *stabilire il numero di centri di servizio in parallelo necessari per ottenere migliori prestazioni* oppure di

individuare come organizzare in modo efficiente il servizio in più fasi successive in una catena di produzione.

Con riferimento al punto *(b)* si desidera stabilire gli effetti dovuti al cambiamento dei parametri, quali i tempi medi di interarrivo degli utenti o i tempi medi di servizio per ognuno dei servitori, oppure se l'introduzione di una nuova apparecchiatura in una catena di produzione può permettere di migliorare i tempi di produzione.

Occorre infine individuare un modello di simulazione atto a descrivere in modo astratto il sistema reale, precisando i *parametri di input*, individuando le *grandezze ritenute significative per la descrizione del sistema* e specificando *gli indici di prestazione di interesse.*

(ii) **Acquisizione dei dati del sistema reale**

L'acquisizione dei dati consiste nell'effettuare un'*analisi preliminare sul sistema reale per rilevare i dati* su cui si baserà la scelta del modello, la stima dei parametri e delle caratteristiche operative e anche la decisione sui componenti e sulle variabili da introdurre nel modello di simulazione. Occorrerà, ad esempio, effettuare opportuni campionamenti sui tempi di interarrivo degli autoveicoli ai caselli autostradali in varie ore della giornata, oppure sui tempi di produzione di particolari articoli in un'industria.

(iii) **Stima e verifica dei parametri e delle caratteristiche operative del sistema reale**

Questo passo è inteso a trasferire nel modello i parametri e le caratteristiche operative del sistema reale, stimate sulla base dei dati raccolti nel punto *(ii)*. Essa si traduce nell'applicare delle *tecniche statistiche per la stima del valore medio e della varianza* di una distribuzione di probabilità di tipo noto e talora di distribuzioni di probabilità non note. Spesso interessa anche stabilire quale sia il *tipo di distribuzione di probabilità più adeguata* da utilizzare.

Una volta che, sulla base dei dati raccolti, si è formulato un modello e si sono scelti i parametri e le caratteristiche funzionali è necessaria una valutazione della sua adeguatezza per descrivere il sistema reale prima di procedere alla costruzione del simulatore. Quindi, costruito il modello è opportuno effettuare degli opportuni *test di verifica di ipotesi statistiche sia sui parametri sia sulle distribuzioni di probabilità.* Ad esempio, se i tempi di interarrivo sono simulati mediante un opportuno generatore pseudocasuale, la valutazione richiederà l'analisi delle *proprietà statistiche del generatore prescelto* per stabilire se esso ha prodotto sequenze pseudocasuali adeguate al tipo di distribuzione di probabilità e ai suoi parametri.

(iv) **Formulazione del programma di simulazione**

Questo passo consiste nella traduzione del modello di simulazione in un modello interpretabile dall'elaboratore. Esso comprende la *stesura di un programma di simulazione* che descriva la successione logica delle operazioni necessarie per analizzare il modello nella fase transiente (per produrre la storia del sistema), la scelta del linguaggio di programmazione e la scelta dello stato iniziale, ossia dei valori da assegnare inizialmente alle variabili del programma.

Occorre anche verificare se il simulatore si comporta come previsto e rifletta la situazione reale. Infatti, nella simulazione possono essere intervenuti errori di programmazione, errori di arrotondamento, problemi di convergenza, ...

(v) **Progettazione degli esperimenti e analisi dei risultati**
In tale passo occorre *progettare gli esperimenti da effettuare*. In particolare, occorre stabilire *quante esecuzioni sono necessarie per ogni esperimento* e *come effettuare le misure al termine della simulazione*. Aumentando il numero di esecuzioni aumenta la significatività statistica dei risultati, ma aumenta anche considerevolmente il costo stesso della simulazione. Infine, occorre effettuare un'analisi delle misure ottenute mediante gli esperimenti di simulazione in maniera da comprendere il comportamento del sistema reale.

Si nota che nella simulazione gioca un ruolo essenziale *l'inferenza statistica*, con particolare riguardo alla stima dei parametri e alla verifica delle ipotesi di cui parleremo nel Capitolo 9.

6.2 Classificazione dei simulatori

Esistono vari modi di classificare i simulatori.

Una *prima classificazione* distingue i simulatori in *statici* e *dinamici*:
• Nei *simulatori statici* la variabile temporale non gioca alcun ruolo e lo scopo fondamentale è quello di determinare alcune caratteristiche utilizzando prove ripetute indipendenti. Tipici esempi sono i simulatori che utilizzano il metodo di Monte Carlo.
• Nei *simulatori dinamici* si descrive l'evoluzione temporale del modello e quindi il tempo diventa la variabile principale. Lo scopo della simulazione è raccogliere dati statistici su processi che evolvono nel tempo. A differenza delle prove ripetute viene a mancare l'indipendenza e occorre prendere in esame la correlazione delle osservazioni. Inoltre, il processo osservato può raggiungere o meno una situazione di equilibrio statistico. Poiché a priori non si hanno informazioni sull'evoluzione del sistema, occorre analizzare la fase transiente per decidere se si raggiungerà una situazione di stabilità.

Una *seconda classificazione* distingue i simulatori in *deterministici* e *casuali*:
• I *simulatori deterministici* sono basati su modelli la cui evoluzione è univocamente determinata una volta fissati i parametri di input. Un esempio tipico è un sistema di servizio $D/D/1$ con tempi di interarrivo e di servizio deterministici; infatti, assegnati i tempi di interarrivo e di servizio, di durata costante, l'evoluzione del sistema è completamente specificata.
• I *simulatori casuali* sono basati su modelli che includono variabili aleatorie o processi stocastici e necessitano della generazione di variabili aleatorie; l'evoluzione del modello dipende dai parametri di input e dalla generazione di una o più variabili aleatorie. Ad esempio, nel sistema di servizio $D/M/1$ i tempi di interarrivo sono di durata costante, mentre i tempi di servizio debbono essere generati tramite la simulazione di una variabile aleatoria esponenziale.

Una *terza classificazione* distingue i simulatori in *sincroni* e *asincroni* in base alle modalità con cui viene descritto il trascorrere del tempo:

• Nei *simulatori sincroni* il tempo di simulazione viene suddiviso in tanti intervalli di uguale ampiezza. All'inizio o alla fine di ciascuno di questi intervalli

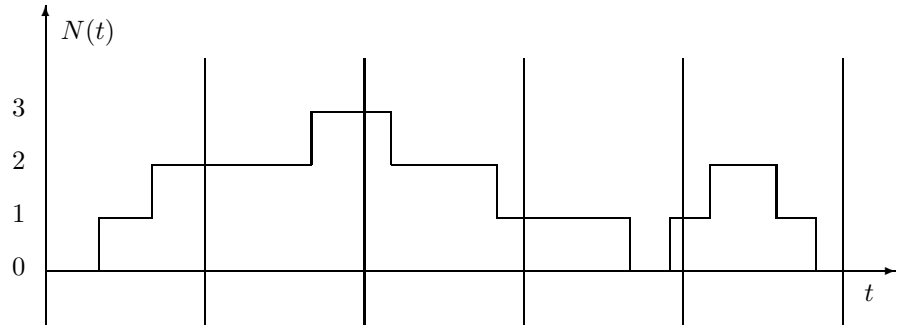

Figura 6.1: Effetto della discretizzazione con passo grande in un simulatore sincrono.

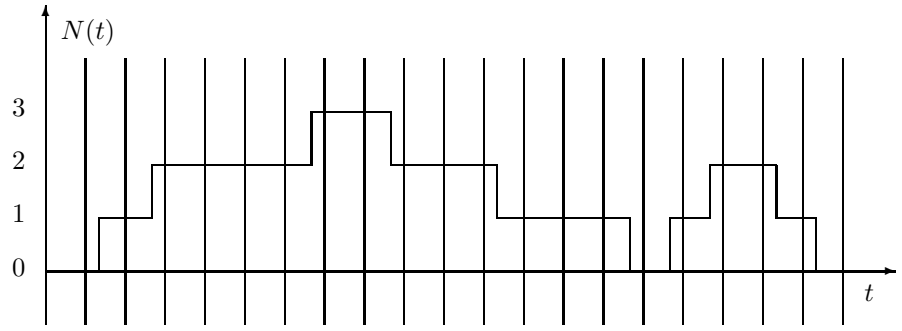

Figura 6.2: Effetto della discretizzazione con passo piccolo in un simulatore sincrono.

si determinano gli eventuali cambiamenti di stato del sistema. In tali simulatori è importante *scegliere accuratamente il passo di discretizzazione* dell'asse temporale. Infatti se si sceglie un *passo molto piccolo* si rischia di aggiornare troppo spesso e inutilmente le variabili, la durata della simulazione diventa elevata e aumenta il costo computazionale. Invece, un *passo di discretizzazione molto grande* può fornire una rappresentazione imprecisa (grossolana) del comportamento del sistema. Se gli eventi si verificano su una scala temporale molto diversa dalla suddivisione temporale in intervalli di uguale ampiezza effettuata con un simulatore sincrono, si potrebbe verificare una *perdita di precisione* poiché si rischia di trascurare comportamenti particolari del sistema. Inoltre in un simulatore sincrono si potrebbe avere una *perdita di efficienza* in presenza di dinamiche temporali poco uniformi, come ad esempio quando si hanno periodi in cui si verificano un gran numero di variazioni dello stato del sistema alternati a periodi in cui si verificano poche variazioni. La perdita di precisione e

di efficienza di un simulatore sincrono comportano che le misure prestazionali del sistema potrebbero risultare imprecise. In Figura 6.1 e in Figura 6.2 sono visualizzate due differenti scelte del passo di discretizzazione. Nella Figura 6.1 il passo di discretizzazione è grande e, osservando il sistema solo all'inizio o alla fine di ciascuno intervallo della suddivisione, si nota che alcuni cambiamenti di stato del processo $N(t)$ non possono essere osservati. Invece nella Figura 6.2 il passo di discretizzazione è piccolo, la simulazione è più precisa ma inefficace, poiché si osserva lo stato del sistema anche in istanti di tempo in cui non si sono verificati cambiamenti di stato.

• Nei *simulatori asincroni* il tempo di simulazione viene aggiornato in modo irregolare in base ai cambiamenti di stato del sistema. Nella Figura 6.3 questi cambiamenti sono indicati con le frecce riportate sull'asse dei tempi. I simulatori asincroni si basano sul principio che lo stato del sistema rimane invariato tra gli istanti successivi di cambiamento di stato del processo e quindi non occorre osservare il sistema in questi periodi di tempo.

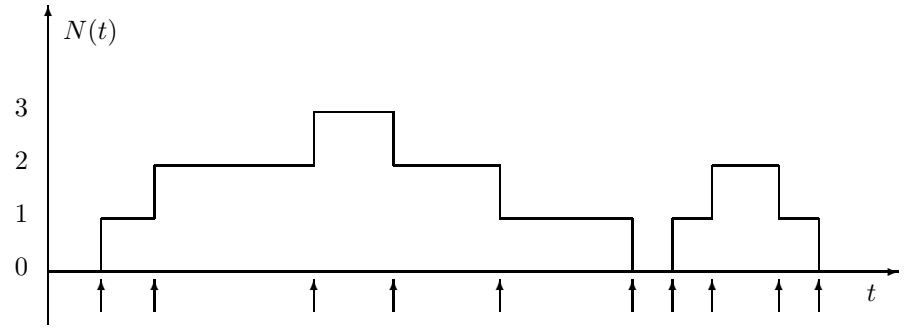

Figura 6.3: Effetto della discretizzazione in un simulatore asincrono.

Una *quarta classificazione* distingue i simulatori in *orientati agli eventi* (*ad eventi discreti*) e *orientati ai processi* in base alle modalità di funzionamento:

• Nei *simulatori ad eventi discreti* (*orientati agli eventi*) lo scorrere del tempo è strettamente legato al verificarsi di cambiamenti di stato (detti *eventi*) del sistema. Si definisce *evento* ogni possibile situazione che porta ad un cambiamento del valore delle variabili di stato che descrivono il comportamento del sistema. Gli eventi non avvengono in modo continuo ma soltanto negli istanti temporali in cui si verificano cambiamento di stato del sistema. Il simulatore ad eventi discreti quindi salta i periodi in cui non si verificano cambiamenti di stato poiché tali periodi non sono significativi per la descrizione dell'evoluzione del sistema. Nei simulatori ad eventi discreti i tempi in cui si verificano gli eventi debbono essere mantenuti in ordine crescente e gli eventi debbono essere realizzati in sequenza.

In un simulatore ad eventi discreti occorre:

(i) definire i tipi di eventi che si possono verificare;

(ii) definire per ogni evento le modifiche da apportare allo stato che descrive il comportamento del sistema;

(iii) definire una struttura dati (calendario) che permetta di ordinare gli eventi sulla base del loro istante di occorrenza e che raccolga le informazioni relative a tali eventi;

(iv) definire la fase di inizializzazione delle variabili;

(v) scorrere il calendario e ogni volta che si incontra un evento eseguire le modifiche delle variabili di stato corrispondenti a quell'evento;

(vi) valutare i parametri prestazionali del sistema.

• Nei *simulatori orientati ai processi* il sistema è descritto in termini di processi (piuttosto che di eventi) che sono eseguiti in parallelo e che interagiscono tra loro scambiandosi informazioni.

Si nota che per *simulare un sistema di servizio* occorre scegliere un *simulatore dinamico, asincrono e ad eventi discreti*. In esso si possono identificare due tipi di eventi che causano cambiamenti di stato, ossia gli arrivi e le partenze degli utenti dal sistema.

6.3 Metodo di Monte Carlo

Nei *simulatori statici* si ricorre spesso al *metodo di Monte Carlo*. Il metodo di Monte Carlo è un procedimento matematico applicabile in diversi campi scientifici e si rivela spesso utile per risolvere vari tipi di problemi matematici: valutazione di integrali unidimensionali e multidimensionali, sistemi di equazioni lineari, inversioni di matrici, problemi di natura statistica, ...

Lo studioso von Neumann nel 1944 chiamò metodo di Monte Carlo *un procedimento statistico basato sull'utilizzazione di numeri casuali*, intendendo con questo nome riferirsi alla capitale del Principato di Monaco. Più precisamente si riferiva alle roulette presenti nei casinò come semplici congegni per la generazione di numeri casuali.

In generale, con il termine *metodo di Monte Carlo* si intende *rappresentare la soluzione di un problema come un parametro non noto (di una ipotetica popolazione) e si cerca di stimare tale parametro utilizzando un campione (estratto dalla popolazione) ottenuto mediante sequenze di numeri casuali.*

Applichiamo ora il metodo di Monte Carlo per *calcolare l'area sottesa da una curva*, che rappresenta il parametro non noto da stimare di una ipotetica popolazione. Sia $f(x)$ una funzione positiva definita nell'intervallo (a, b) e sia

$$J = \int_a^b f(x)\, dx, \tag{6.1}$$

ossia J è l'area sottesa dalla curva $f(x)$ nell'intervallo (a, b).

Considereremo prima una procedura numerica per calcolare l'integrale definito unidimensionale (6.1) e successivamente utilizzeremo due differenti metodi di Monte Carlo per calcolare lo stesso integrale.

6.3.1 Metodo numerico

Suddividiamo l'intervallo (a, b) in N sottointervalli di ampiezza Δ. Quindi $b - a = N\,\Delta$. Poniamo inoltre

$$x_0 = a, \qquad x_i = a + i\,\Delta = a + i\,\frac{b-a}{N} \qquad (i = 1, 2, \ldots, N) \tag{6.2}$$

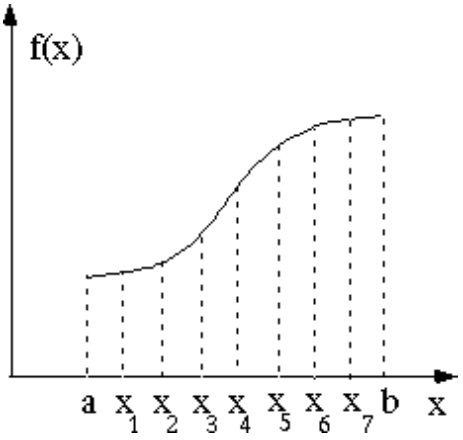

Figura 6.4: Metodo numerico per il calcolo di un integrale in (a, b).

Si nota che $f(x_i)\,\Delta$ rappresenta l'area del rettangolo di altezza $f(x_i)$ e di base Δ. Un'approssimazione numerica per J è data da

$$\widehat{J} = \sum_{i=1}^{N} f(x_i)\,\Delta = \sum_{i=1}^{N} f(x_i)\,\frac{b-a}{N} = (b-a)\,\Big[\frac{1}{N}\,\sum_{i=1}^{N} f(x_i)\Big]. \tag{6.3}$$

Tale approssimazione numerica corrisponde al calcolo della media aritmetica di $f(x_1), f(x_2), \ldots, f(x_N)$ moltiplicata per l'ampiezza dell'intervallo di integrazione $b - a$.

6.3.2 Primo metodo di Monte Carlo

Tale metodo consiste nel generare N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$ e nel trasformarle in variabili aleatorie ancora indipendenti e uniformemente distribuite nell'intervallo di integrazione (a, b) mediante la trasformazione

$$X_i = a + (b-a)\,U_i \qquad (i = 1, 2, \ldots, N). \tag{6.4}$$

A differenza del metodo numerico ora gli N punti non vengono più scelti deterministicamente e equidistanziati di Δ nell'intervallo (a, b) ma *scelti in maniera probabilistica*, ossia si assume che siano *indipendenti e uniformemente distribuiti*

nell'intervallo (a, b). Uno stimatore che utilizza il metodo di Monte Carlo per valutare J è

$$\widehat{J} = (b-a)\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big]. \tag{6.5}$$

Vogliamo ora dimostrare che tale stimatore gode di importanti proprietà statistiche (*correttezza* e *consistenza*), ossia:

$$E(\widehat{J}) = J, \qquad \lim_{N\to+\infty} \mathrm{Var}(\widehat{J}) = 0. \tag{6.6}$$

Infatti, poiché risulta

$$E[f(X_i)] = \int_a^b f(x)\,\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b f(x)\,dx = \frac{J}{b-a}\,,$$
$$E[f^2(X_i)] = \int_a^b f^2(x)\,\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b f^2(x)\,dx,$$

si ha

$$E(\widehat{J}) = E\Big\{(b-a)\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big]\Big\} = \frac{b-a}{N}\sum_{i=1}^{N} E[f(X_i)] = J,$$
$$\mathrm{Var}(\widehat{J}) = \mathrm{Var}\Big\{(b-a)\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big]\Big\} = \frac{(b-a)^2}{N^2}\sum_{i=1}^{N}\mathrm{Var}[f(X_i)],$$

da cui si ottengono le (6.6). I passi dell'algoritmo del primo metodo di Monte Carlo per valutare J sono quindi i seguenti:

Algoritmo
STEP 1: generare una sequenza di N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: generare una sequenza di N variabili aleatorie $X_1, X_2, \ldots, X_N$ indipendenti e uniformemente distribuite nell'intervallo (a, b) tramite la trasformazione

$$X_i = a + (b-a)\,U_i \qquad (i = 1, 2, \ldots, N)$$

STEP 3: stimare J utilizzando lo stimatore:

$$\widehat{J} = (b-a)\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big].$$

6.3.3 Secondo metodo di Monte Carlo: successo e insuccesso

Sia $f(x)$ una funzione definita nell'intervallo (a, b) tale che $0 \le f(x) \le c$. Il metodo consiste nel generare indipendentemente N coppie di numeri casuali (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) uniformemente distribuiti nel rettangolo $R =$

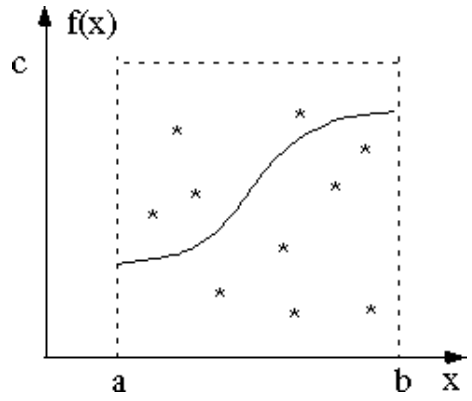

Figura 6.5: Secondo metodo di Monte Carlo per il calcolo di un integrale in (a, b).

$\{(x, y) : \ a < x < b, \ 0 < y < c\}$. La funzione densità di probabilità di ogni singola coppia (X, Y) è uniforme nel rettangolo, ossia

$$g_{XY}(x, y) = \begin{cases} \dfrac{1}{c\,(b-a)}, & (x, y) \in R \\ 0, & \text{altrimenti.} \end{cases} \tag{6.7}$$

Le variabili aleatorie X e Y sono tra loro indipendenti e caratterizzate rispettivamente da densità marginali

$$g_X(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{altrimenti,} \end{cases} \qquad g_Y(y) = \begin{cases} \dfrac{1}{c}, & 0 < y < c \\ 0, & \text{altrimenti.} \end{cases}$$

Geometricamente è ovvio che se N_S di questi punti sono sotto la curva $y = f(x)$, sussiste l'approssimazione:

$$\frac{N_S}{N} \simeq \frac{\displaystyle\int_a^b f(x)\,dx}{c\,(b-a)},$$

che ci conduce a considerare come stimatore per l'integrale (6.1):

$$\widehat{J} = c\,(b-a)\,\frac{N_S}{N}\,. \tag{6.8}$$

Vogliamo ora dimostrare che tale stimatore gode delle proprietà (6.6). Osserviamo in primo luogo che ciascuna delle N prove costituisce una prova di Bernoulli in cui le probabilità di successo e insuccesso sono rispettivamente

$$p = \frac{\displaystyle\int_a^b f(x)\,dx}{c\,(b-a)}, \qquad q = 1 - \frac{\displaystyle\int_a^b f(x)\,dx}{c\,(b-a)}. \tag{6.9}$$

Essendo $E(N_S) = N\,p$ e $\mathrm{Var}(N_S) = N\,p\,q$, si nota che

$$E(\widehat{J}) = E\Big[c\,(b-a)\,\frac{N_S}{N}\Big] = \frac{c\,(b-a)}{N}\,E(N_S) = \int_a^b f(x)\,dx = J$$
$$\mathrm{Var}(\widehat{J}) = \mathrm{Var}\Big[c(b-a)\,\frac{N_S}{N}\Big] = \frac{c^2(b-a)^2}{N^2}\,\mathrm{Var}(N_S),$$

da cui si ottengono le (6.6). I passi dell'algoritmo del secondo metodo di Monte Carlo per valutare J sono quindi i seguenti:

Algoritmo
STEP 1: generare indipendentemente due sequenze $U_1, U_2, \ldots, U_N$ e $V_1, V_2, \ldots, V_N$, ognuna costituita da N variabili aleatorie indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: generare una sequenza di N variabili aleatorie $X_1, X_2, \ldots, X_N$ uniformemente distribuite nell'intervallo (a, b) e inoltre generare una sequenza di N variabili aleatorie $Y_1, Y_2, \ldots, Y_N$ uniformemente distribuite nell'intervallo $(0, c)$ mediante le trasformazioni

$$X_i = a + (b-a)\,U_i, \qquad Y_i = c\,V_i \qquad (i = 1, 2, \ldots, N) \tag{6.10}$$

STEP 3: porre inizialmente la variabile aleatoria contatore $N_S = 0$; per ognuna delle coppie (X_i, Y_i) $(i = 1, 2, \ldots, N)$ determinate con il generatore uniforme controllare se $Y_i < f(X_i)$; se tale condizione è soddisfatta allora la variabile aleatoria contatore N_S è incrementata di un'unità mentre se tale condizione non è soddisfatta N_S non viene incrementata;
STEP 4: stimare J utilizzando lo stimatore:

$$\widehat{J} = c\,(b-a)\,\frac{N_S}{N}\,.$$

6.3.4 Integrali su domini infiniti: metodo di Monte Carlo

Finora ci siamo occupati della valutazione di integrali definiti su un dominio finito. Desideriamo ora calcolare con il metodo di Monte Carlo il seguente integrale:

$$J = \int_0^{+\infty} f(x)\,dx, \tag{6.11}$$

Effettuiamo in (6.11) il cambiamento di variabile $y = (x+1)^{-1}$, ossia $x = (1-y)/y$. Allora si ha:

$$J = \int_0^{+\infty} f(x)\,dx = \int_0^1 f\Big(\frac{1-y}{y}\Big)\,\frac{1}{y^2}\,dy = \int_0^1 h(y)\,dy, \tag{6.12}$$

dove si è posto

$$h(y) = f\Big(\frac{1-y}{y}\Big)\,\frac{1}{y^2} \tag{6.13}$$

Abbiamo ricondotto la valutazione di J alla valutazione di un integrale nell'intervallo $(0,1)$ del tipo (6.1). Applicando quindi il primo metodo di Monte Carlo si ha:

$$\widehat{J} = \frac{1}{N}\sum_{i=1}^{N} h(U_i) = \frac{1}{N}\sum_{i=1}^{N} f\Big(\frac{1-U_i}{U_i}\Big)\,\frac{1}{U_i^2}. \tag{6.14}$$

Vogliamo dimostrare che sussistono ancora le (6.6). Infatti, ricordando la (6.13) si ha

$$E[h(U_i)] = \int_0^1 h(y)\,dy = \int_0^{+\infty} f(x)\,dx = J\,,$$
$$E[h^2(U_i)] = \int_0^1 h^2(y)\,dy = \int_0^{+\infty} f^2(x)\,dx,$$

da cui si ottiene:

$$E(\widehat{J}) = E\bigg[\frac{1}{N}\sum_{i=1}^{N} h(U_i)\bigg] = \frac{1}{N}\sum_{i=1}^{N} E[h(U_i)] = J,$$
$$\mathrm{Var}(\widehat{J}) = \mathrm{Var}\bigg[\frac{1}{N}\sum_{i=1}^{N} h(U_i)\bigg] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}[h(U_i)],$$

da cui si ottengono le (6.6). I passi dell'algoritmo del primo metodo di Monte Carlo per calcolare J in (6.11) sono quindi i seguenti:

Algoritmo
STEP 1: generare una sequenza di N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0,1)$;
STEP 2: stimare J utilizzando lo stimatore:

$$\widehat{J} = \frac{1}{N}\sum_{i=1}^{N} f\Big(\frac{1-U_i}{U_i}\Big)\,\frac{1}{U_i^2}.$$

Nel caso di integrali multidimensionali il metodo di Monte Carlo si rivela spesso competitivo e più efficace rispetto ad altri metodi puramente numerici.

Si nota che nelle procedure del metodo di Monte Carlo precedentemente utilizzate per il calcolo di integrali unidimensionali intervengono due importanti punti:

(i) generazione di variabili aleatorie uniformemente distribuite nell'intervallo $(0,1)$;

(ii) generazione, a partire dal punto *(i)*, di altre variabili aleatorie con opportune distribuzioni di probabilità.

In entrambi i punti l'affidabilità dei risultati che si ottengono si basa principalmente sulla qualità della sorgente dei numeri casuali e sulla scelta di un algoritmo computazionale efficiente. Le problematiche indicate nei punti *(i)* e *(ii)* intervengono anche nella *simulazione dei sistemi di servizio.*

6.4 Simulazione di un sistema di servizio

Vogliamo ora occuparci della simulazione di un sistema di servizio singolo servitore, singola fila di attesa, a capacità infinita, con disciplina di servizio FIFO in cui i tempi di interarrivo e i tempi di servizio hanno una distribuzione di probabilità di tipo generale (deterministica, uniforme, esponenziale, di Erlang, iperesponenziale, ...). In Figura 6.6 è mostrata una tipica realizzazione di un tale sistema di servizio.

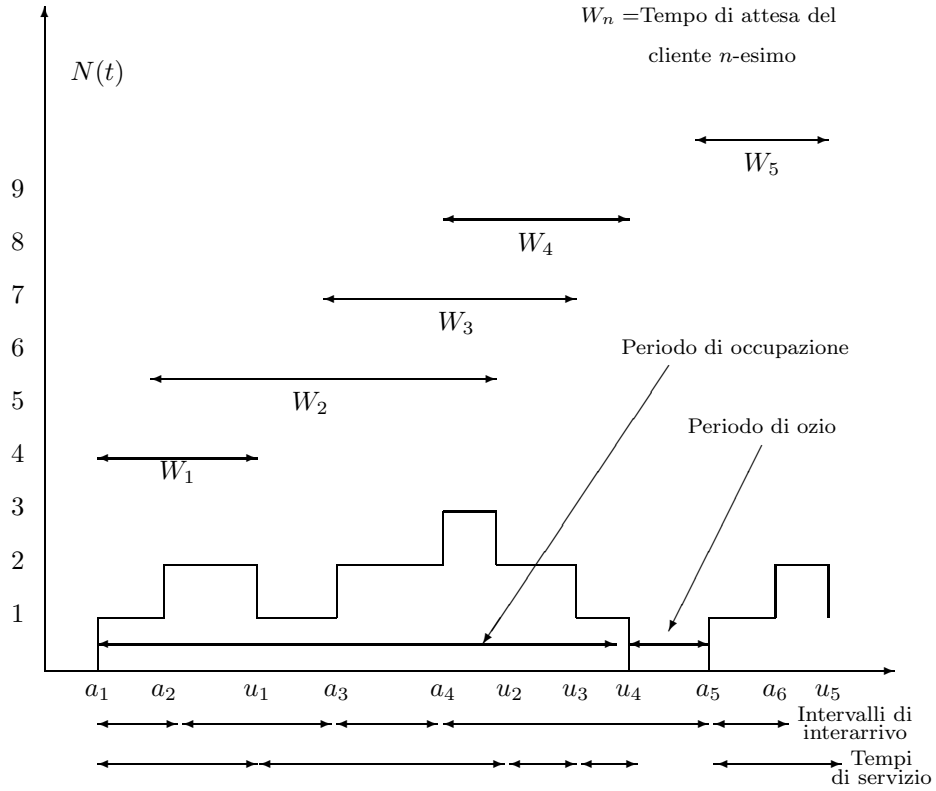

Figura 6.6: Una tipica realizzazione di un sistema di servizio.

Supponiamo che i cambiamenti di stato siano istantanei ed avvengano in corrispondenza di eventi. Nel caso di un sistema di servizio con unico servitore si possono identificare due tipi di eventi:

- l'arrivo di un utente nel sistema che causa l'incremento del numero di utenti nel sistema;

- la fine del servizio di un utente con la conseguente uscita dal sistema e l'accesso al servizio di un altro utente (se ne esistono altri in fila di attesa); ciò causa il decremento del numero di utenti nel sistema.

Scopo della simulazione è quello di ricostruire l'evoluzione nel tempo del comportamento del sistema. Le componenti fondamentali della simulazione del sistema di servizio sono:

- la simulazione dei tempi di interarrivo con specificata distribuzione;

- la simulazione dei tempi di servizio con specificata distribuzione;

- il meccanismo con cui la simulazione ha termine.

La scelta della distribuzione teorica dei tempi di interarrivo e di servizio che meglio approssima il comportamento del sistema reale avviene calcolando le frequenze, le medie e le deviazioni standard ricavate da un'analisi storica del sistema reale e confrontandole con quelle teoriche ipotizzate. Il *test del chi–quadrato*, di cui parleremo nel Capitolo 9, mostrerà poi se è possibile sostituire la distribuzione empirica con quella teorica.

Riferendoci alla Figura 6.6, utilizziamo le seguenti notazioni:

- $T(k)$ tempo di interarrivo k–esimo, ossia l'intervallo di tempo intercorrente tra il $(k-1)$-esimo arrivo e il k-esimo arrivo al sistema di servizio ($k = 1, 2, \ldots$).

- $S(k)$ tempo di servizio del k-esimo utente ($k = 1, 2, \ldots$).

- $A(k)$ istante di arrivo del k-esimo utente nel sistema, ossia l'istante di tempo in cui effettivamente arriva il k-esimo utente. Sussiste la relazione ricorsiva:

$$\begin{aligned} A(1) &= T(1) \\ A(k) &= T(1) + T(2) + \ldots + T(k-1) + T(k) = A(k-1) + T(k) \quad (k = 2, 3, \ldots). \end{aligned} \tag{6.15}$$

- $U(k)$ istante di uscita del k-esimo utente dal sistema, ossia l'istante di tempo in cui il k-esimo utente completa il servizio e lascia il sistema di servizio ($k = 1, 2, \ldots$). In un sistema di servizio con unico servitore e disciplina di servizio FIFO, l'istante di partenza del k-esimo utente può essere ricavato utilizzando la relazione ricorsiva:

$$\begin{aligned} U(1) &= A(1) + S(1) \\ U(k) &= \max\big\{A(k), U(k-1)\big\} + S(k) \qquad (k = 2, 3, \ldots). \end{aligned} \tag{6.16}$$

Tale formula ricorsiva è detta "legge di Lindley". Si nota che se $k = 2, 3, \ldots$

$$U(k) = \begin{cases} A(k) + S(k), & A(k) \geq U(k-1) \\ U(k-1) + S(k), & A(k) < U(k-1) \end{cases}$$

Ciò significa che se l'utente $(k-1)$-esimo esce dal sistema prima o nello stesso istante dell'entrata dell'utente k-esimo, allora questo ultimo utente entra immediatamente in servizio e il suo tempo di uscita è uguale al suo

tempo di arrivo più il suo tempo di servizio. Se, invece, l'utente $(k-1)$-esimo esce dal sistema dopo l'entrata dell'utente k-esimo, allora questo ultimo utente non può entrare immediatamente in servizio (deve attendere nella fila di attesa) e quindi il suo tempo di uscita è uguale al tempo di uscita dell'utente (k-1)-esimo più il tempo di servizio dell'utente k-esimo.

- $W(k)$ tempo di attesa nel sistema di servizio del k-esimo utente ($k = 1, 2, \ldots$). Si deve avere

$$W(1) = S(1)$$
$$W(k) = U(k) - A(k) \qquad (k = 2, 3, \ldots).$$

- $Q(k)$ tempo di permanenza nella fila di attesa del k-esimo utente ($k = 1, 2, \ldots$). Si deve avere

$$Q(1) = 0$$
$$Q(k) = W(k) - S(k) \qquad (k = 2, 3, \ldots).$$

- $O(k)$ tempo di ozio del centro di servizio fino all'istante di arrivo del k-esimo utente ($k = 1, 2, \ldots$). Si deve avere

$$O(k) = \begin{cases} 0, & U(k-1) \geq A(k) \\ A(k) - U(k-1), & U(k-1) < A(k). \end{cases} \tag{6.17}$$

La sequenza $T(1), T(2), \ldots$ dei tempi di interarrivo è ottenuta simulando una variabile aleatoria con la distribuzione di probabilità desiderata (uniforme, esponenziale, di Erlang, iperesponenziale). Analogamente, la sequenza $S(1), S(2), \ldots$ dei tempi di servizio è anch'essa ottenuta simulando una variabile aleatoria con la desiderata distribuzione di probabilità (uniforme, esponenziale, di Erlang, iperesponenziale).

Supponiamo di iniziare la simulazione al tempo $t = 0$. Il modello di simulazione reagisce agli eventi mettendo in calendario nuovi (futuri) eventi. Avendo identificato come eventi possibili l'arrivo di un nuovo utente e la partenza di un utente, ossia la fine di un servizio, il simulatore viene inizializzato ponendo sia la variabile tempo sia le variabili di stato a 0 e inserendo nel calendario degli eventi un arrivo per ogni istante $A(k)$ e una partenza per ogni istante $U(k)$, come mostrato in Figura 6.6. Ha poi inizio la simulazione vera e propria. Il tempo viene aggiornato al valore $A(1)$, istante in cui si verifica il primo arrivo. Il primo utente, che arriva al tempo $A(1) = T(1)$, entra direttamente in servizio e quindi il suo tempo di permanenza nella fila di attesa è nullo inoltre, il numero di utenti è incrementato di un'unità e assume quindi il valore 1. Sapendo che il primo utente entra in servizio all'istante $A(1)$, un evento di tipo partenza è inserito nel calendario degli eventi all'istante $U(1) = A(1) + S(1) = T(1) + S(1)$.

La simulazione procede analizzando il prossimo evento presente nel calendario degli eventi, ossia il secondo arrivo e confrontando il tempo del secondo arrivo con il tempo di uscita del primo utente.

Il secondo utente arriverà al tempo $A(2) = T(1) + T(2)$. Distinguiamo tre casi: *(1)* $U(1) > T(1) + T(2) = A(2)$, *(2)* $U(1) < A(2)$ e *(3)* $U(1) = A(2)$

(1) $U(1) > A(2)$

In tal caso l'istante di uscita del primo utente è maggiore del tempo di arrivo del secondo utente. Il secondo utente non può quindi essere immediatamente servito e deve attendere nella fila di attesa (di lunghezza 1). Risulta quindi che

$$Q(2) = U(1) - A(2), \qquad W(2) = Q(2) + S(2), \qquad O(2) = 0.$$

(2) $U(1) < A(2)$

In tal caso l'istante di uscita dal sistema del primo utente è minore del tempo di arrivo del secondo utente. Il primo utente lascia quindi il sistema prima dell'arrivo del secondo utente e il centro di servizio resta inutilizzato fino all'arrivo del secondo utente. Pertanto si ha:

$$Q(2) = 0, \qquad W(2) = Q(2) + S(2) = S(2), \qquad O(2) = A(2) - U(1).$$

(3) $U(1) = A(2)$

In tal caso il tempo di uscita dal sistema del primo utente coincide con il tempo di arrivo del secondo utente. Il primo utente lascia quindi il sistema quando arriva il secondo utente e quindi si ottiene:

$$Q(2) = 0, \qquad W(2) = Q(2) + S(2) = S(2), \qquad O(2) = A(2) - U(1) = 0.$$

Inoltre, l'istante di uscita dal sistema del secondo utente può essere così calcolato:

$$U(2) = \max\big\{A(2), U(1)\big\} + S(2).$$

Tale procedura può essere iterata per $k = 3, 4, \ldots$ Si nota che occorre ordinare gli eventi in un calendario seguendo l'ordine cronologico del verificarsi degli eventi registrando l'istante di tempo in cui si verifica l'evento e un identificativo del tipo di evento per distinguere se l'evento è un arrivo o una partenza.

Esempio 6.1 Consideriamo un sistema di servizio singolo servitore, singola fila di attesa, a capacità infinita, con disciplina di servizio FIFO.

Supponiamo di aver generato le seguenti due sequenze di tempi di interarrivo e di servizio, misurati in minuti:

k	1	2	3	4	5	6	7	8	9	10	11
$T(k)$	1.73	1.35	0.71	0.62	14.28	0.70	15.52	3.15	0.76	1.00	0.50
$S(k)$	2.90	1.76	3.39	4.52	4.46	4.36	2.07	3.36	2.37	5.38	0.50

Tabella 6.1: Sequenze dei tempi di interarrivo e di servizio.

Scegliamo $t_c = 40$ minuti come istante di tempo oltre il quale non è più concesso agli utenti di accedere al sistema.

Nella Tabella 6.2 sono elencati i tempi di arrivo dei vari utenti utilizzando la (6.15).

$A(1) = T(1) = 1.73$
$A(2) = A(1) + T(2) = 1.73 + 1.35 = 3.08$
$A(3) = A(2) + T(3) = 3.08 + 0.71 = 3.79$
$A(4) = A(3) + T(4) = 3.79 + 0.62 = 4.41$
$A(5) = A(4) + T(5) = 4.41 + 14.28 = 18.69$
$A(6) = A(5) + T(6) = 18.69 + 0.70 = 19.39$
$A(7) = A(6) + T(7) = 19.39 + 15.52 = 34.91$
$A(8) = A(7) + T(8) = 34.91 + 3.15 = 38.06$
$A(9) = A(8) + T(9) = 38.06 + 0.76 = 38.82$
$A(10) = A(9) + T(10) = 38.82 + 1.00 = 39.82$
$A(11) = A(10) + T(11) = 39.82 + 0.5 = 40.32$

Tabella 6.2: Tempi di arrivo.

Osservando la Tabella 6.2 si nota che fino al tempo $t_c = 40$ minuti soltanto 10 utenti possono accedere al sistema di servizio. La media del tempo di interarrivo di questi dieci utenti è:

$$\overline{T} = \frac{1}{10}\sum_{k=1}^{10} T(k) = 3.982 \text{ minuti.}$$

Nella Tabella 6.3 sono invece elencati i tempi di uscita dei vari utenti utilizzando la (6.16). Inoltre, ordinando in ordine crescente gli istanti di arrivo e di partenza

$U(1) = A(1) + S(1) = 1.73 + 2.90 = 4.63$
$U(2) = \max\{A(2), U(1)\} + S(2) = 4.63 + 1.76 = 6.39$
$U(3) = \max\{A(3), U(2)\} + S(3) = 6.39 + 3.39 = 9.78$
$U(4) = \max\{A(4), U(3)\} + S(4) = 9.78 + 4.52 = 14.3$
$U(5) = \max\{A(5), U(4)\} + S(5) = 18.69 + 4.46 = 23.15$
$U(6) = \max\{A(6), U(5)\} + S(6) = 23.15 + 4.36 = 27.51$
$U(7) = \max\{A(7), U(6)\} + S(7) = 34.91 + 2.07 = 37$
$U(8) = \max\{A(8), U(7)\} + S(8) = 38.06 + 3.36 = 41.42$
$U(9) = \max\{A(9), U(8)\} + S(9) = 41.42 + 2.37 = 43.79$
$U(10) = \max\{A(10), U(9)\} + S(10) = 43.79 + 5.38 = 49.17$

Tabella 6.3: Tempi di uscita.

è possibile ricostruire la realizzazione, illustrata in Figura 6.7, del numero di utenti presenti nel sistema di servizio.

I risultati della simulazione permettono di ottenere una stima dei tempi di attesa nel sistema e di permanenza in fila di attesa dei dieci utenti, ossia $W(i) = U(i) - A(i)$ e $Q(i) = W(i) - S(i)$ $(i = 1, 2, \ldots, 10)$, come mostrato in Tabella 6.4.

Facendo uso dei risultati di Tabella 6.4 in (6.19) si ricavano le medie campionarie del tempo di servizio, di attesa nel sistema e di permanenza in fila di attesa di un utente:

$$\overline{S} = \frac{1}{10}\sum_{k=1}^{10} S(k) = 3.457 \text{ minuti,}$$

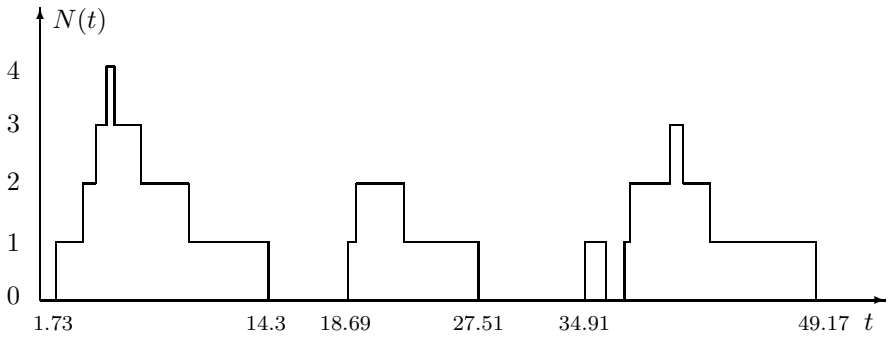

Figura 6.7: Una realizzazione del numero di utenti presenti nel sistema di servizio con singolo servitore.

$S(1) = 2.90$	$W(1) = U(1) - A(1) = 2.90$	$Q(1) = W(1) - S(1) = 0.00$
$S(2) = 1.76$	$W(1) = U(2) - A(2) = 3.31$	$Q(2) = W(2) - S(2) = 2.55$
$S(3) = 3.39$	$W(3) = U(3) - A(3) = 5.99$	$Q(3) = W(3) - S(3) = 2.60$
$S(4) = 4.52$	$W(4) = U(4) - A(4) = 9.89$	$Q(4) = W(4) - S(4) = 5.37$
$S(5) = 4.46$	$W(5) = U(5) - A(5) = 14.46$	$Q(5) = W(5) - S(5) = 10.00$
$S(6) = 4.36$	$W(6) = U(6) - A(6) = 8.12$	$Q(6) = W(6) - S(6) = 3.76$
$S(7) = 2.07$	$W(7) = U(7) - A(7) = 2.09$	$Q(7) = W(7) - S(7) = 0.02$
$S(8) = 3.36$	$W(8) = U(8) - A(8) = 3.36$	$Q(8) = W(8) - S(8) = 0.00$
$S(9) = 2.37$	$W(9) = U(9) - A(9) = 4.97$	$Q(9) = W(9) - S(9) = 2.60$
$S(10) = 5.38$	$W(10) = U(10) - A(10) = 9.35$	$Q(10) = W(10) - S(10) = 3.97$

Tabella 6.4: Tempi di servizio, di attesa nel sistema e di permanenza in fila di attesa degli utenti.

$$\overline{Q} = \frac{1}{10}\sum_{k=1}^{10} Q(k) = 2.987 \text{ minuti},$$

$$\overline{W} = \frac{1}{10}\sum_{k=1}^{10} W(k) = 6.444 \text{ minuti}.$$

♢

Per simulare il sistema di servizio occorre una variabile che rappresenti il tempo e una struttura che raccolga le informazioni relative agli eventi. Questa struttura è detta *calendario degli eventi*. Il calendario degli eventi è realizzato tramite una lista ordinata in base al campo che contiene l'istante di tempo in cui gli eventi si verificano. Tale lista deve contenere per ogni tempo di osservazione t (dove t è l'istante di tempo in cui si verifica un arrivo o una partenza) il numero di utenti presenti nel sistema, il numero cumulativo di arrivi, il numero cumulativo di partenze, i tempi del prossimo arrivo dopo t e i tempi completamento di servizio dell'utente attualmente in servizio.

Occorrono inoltre delle variabili che consentano di trarre informazioni complessive sulla simulazione, ossia le variabili di input, di stato e di output.

Variabili di input

- i parametri necessari per definire la distribuzione dei tempi di interarrivo degli utenti;

- i parametri necessari per definire la distribuzione dei tempi di servizio;

- il dato in base al quale sarà definita la fine del processo di simulazione (numero massimo di utenti da servire, tempo massimo di simulazione, ...)

Supponiamo di denotare con t_c l'istante dopo il quale non sia più consentito agli utenti in arrivo di accedere al sistema, sebbene dopo tale istante il servitore dovrà completare il servizio di tutti gli utenti presenti nel sistema al tempo t_c. Questa situazione si verifica frequentemente nei sistemi di servizio (banche, uffici postali, centri ambulatoriali, ...) in cui è previsto un orario giornaliero di apertura e di chiusura al pubblico.

Variabili di stato

- istanti di arrivo di ogni utente;

- istanti di partenza di ogni utente;

- numero di utenti nel sistema;

- numero di utenti in fila di attesa;

- numero cumulativo di arrivi;

- numero cumulativo di partenze;

- istante t_f di fine della simulazione, ossia l'istante di tempo in cui l'ultimo utente lascia il sistema (nel sistema non sono più presenti utenti).

Variabili di output
Occorre ricavare delle stime delle seguenti variabili:

- fattore di utilizzazione del sistema;

- numero medio di utenti in fila di attesa;

- numero medio di utenti nel sistema;

- tempo medio di permanenza nella fila di attesa;

- tempo medio di attesa di un utente nel sistema.

Nella formulazione dell'algoritmo generale per la simulazione di un sistema di servizio occorre distinguere sulla base dello stato i seguenti casi:

- l'arrivo di un utente quando il sistema di servizio è vuoto (inizio immediato del servizio);

- l'arrivo di un utente quando il sistema di servizio non è vuoto;

- la fine del servizio con la fila di attesa vuota;

- la fine del servizio con la fila di attesa non vuota (inizio immediato del servizio del successivo utente).

Algoritmo generale per simulare il sistema di servizio con unico servitore

Inizializzazione:
Si inizializzano a zero le seguenti variabili:

- il tempo di sistema t,

- il numero cumulativo di arrivi N_A,

- il numero cumulativo di partenze N_U,

- il numero di utenti presenti nel sistema n,

- il numero di utenti nella fila di attesa n_q.

Questi stati debbono essere aggiornati in modo appropriato al verificarsi di ogni evento. Successivamente occorre muoversi lungo l'asse temporale fino ad incontrare il prossimo evento.

Dopo la fase di inizializzazione, si fa partire il processo di simulazione generando il primo evento, ossia un arrivo, innescando la procedura arrivo. Occorre quindi generare un tempo di interarrivo con la distribuzione di probabilità desiderata.

Procedura arrivo

In generale, la procedura arrivo si innesca quando il tempo del prossimo arrivo dopo il tempo di sistema t è minore o uguale al tempo di completamento del servizio dell'utente attualmente in servizio ed inoltre il tempo del prossimo arrivo dopo t (minimo tra i due tempi) è minore o uguale a t_c.

1. aggiornare il tempo di sistema t all'attuale istante di arrivo;

2. registrare il tempo di arrivo dell'attuale utente;

3. incrementare di uno il numero cumulativo degli arrivi;

4. incrementare di uno il numero di utenti nel sistema;

5. se l'utente che arriva trova il servitore libero, ossia l'utente arrivato è l'unico nel sistema:

    - registrare l'istante attuale come inizio del servizio dell'utente;

- generare un tempo di servizio con la distribuzione di probabilità desiderata e calcolare l'istante di completamento del servizio (registrandolo come evento futuro). La nuova partenza avverrà al tempo $t+S$, dove S è il tempo di servizio (generato mediante simulazione).

altrimenti se l'utente che arriva non trova il servitore libero

- inserire l'utente in coda secondo la disciplina di servizio e incrementare di uno il numero di utenti in coda.

6. generare un tempo di interarrivo con la distribuzione di probabilità desiderata e calcolare l'istante del prossimo arrivo (registrandolo come evento futuro). Il nuovo arrivo avverrà al tempo $t+T$, dove T è un tempo di interarrivo (generato mediante simulazione).

Si nota che se il servitore è libero, l'utente appena arrivato non aspetta in coda e uscirà dal sistema dopo un tempo pari al suo tempo di servizio. Quando viene scandito un evento arrivo e il sistema è vuoto viene inserito un nuovo evento partenza nel calendario ad un tempo pari al precedente tempo più il valore del tempo di servizio.

Procedura partenza

In generale, la procedura di partenza si innesca quando il tempo del prossimo arrivo dopo il tempo di sistema t è maggiore del tempo di completamento del servizio dell'utente attualmente in servizio ed, inoltre, questo tempo di completamento del servizio (minimo tra i due tempi) è minore o uguale a t_c

1. aggiornare il tempo t di sistema all'attuale istante di partenza;
2. registrare il tempo di fine servizio per l'attuale utente;
3. incrementare di uno il numero cumulativo di utenti serviti;
4. decrementare di uno il numero di utenti nel sistema;
5. se l'utente in partenza lascia altri utenti in fila di attesa, ossia il sistema non è vuoto

   - prelevare un utente dalla fila di attesa secondo la disciplina di servizio;
   - generare il tempo di servizio con la distribuzione di probabilità desiderata e calcolare l'istante di partenza (registrandolo come evento futuro). La nuova partenza avverrà al tempo $t+S$, dove S è il tempo di servizio (generato mediante simulazione)..

   altrimenti se il sistema è vuoto

   - porre il prossimo istante di partenza uguale ad infinito.

Quando viene scandito un evento fine servizio con coda non vuota viene inserito un nuovo evento fine del servizio nel calendario degli eventi al un tempo pari al precedente tempo più il valore del tempo di servizio.

Procedura di terminazione

La procedura di terminazione si innesca quando il minimo tra il tempo del prossimo arrivo dopo il tempo di sistema t ed il tempo di completamento del servizio dell'utente attualmente in servizio dopo il tempo di sistema t sono maggiori di t_c. In questo caso non possono più accedere nuovi utenti nel sistema ed occorre fornire il servizio agli utenti già entrati nel sistema.

1. - innescare ripetutamente la procedura di partenza finché nel sistema non sono più presenti utenti, nel qual caso terminare la simulazione e registrare nella variabile t_f l'istante di tempo in cui l'ultimo utente lascia il sistema.

Procedura statistica

Al termine della simulazione abbiamo ottenuto N_A (numero totale degli arrivi) che sarà anche uguale a N_U (numero totale delle partenze) ed inoltre anche t_f, ossia il tempo finale di simulazione corrispondente all'istante in cui l'ultimo utente lascia il sistema. Dalle sequenze dei tempi di interarrivo $T(1), T(2), \ldots, T(N_A)$ e di servizio $S(1), S(2), \ldots, S(N_A)$ si possono immediatamente ricavare i tempi di interarrivo, i tempi di servizio, i tempi di permanenza in coda e nel sistema, gli istanti di arrivo e di partenza e alcune caratteristiche statistiche quali medie campionarie e varianze campionarie. Le stime della media dei tempi di interarrivo e di servizio sono:

$$\overline{T} = \frac{1}{N_A} \sum_{k=1}^{N_A} T(k), \qquad \overline{S} = \frac{1}{N_A} \sum_{k=1}^{N_A} S(k). \tag{6.18}$$

Inoltre, avendo ottenuto tramite le (6.15) i tempi di arrivo $A(1), A(2), \ldots, A(N_A)$ degli utenti nel sistema e tramite la (6.16) i tempi di uscita $U(1), U(2), \ldots, U(N_A)$ degli utenti dal centro di servizio, allora le differenze $W(k) = U(k) - A(k)$ $(k = 1, 2 \ldots, N_A)$ rappresentano i tempi di attesa dei vari utenti nel sistema e le differenze $Q(k) = W(k) - S(k)$ $(k = 1, 2 \ldots, N_A)$ forniscono i tempi di permanenza dei vari utenti nella fila di attesa. Le stime della media dei tempi di attesa nel sistema e nella fila di attesa sono:

$$\overline{W} = \frac{1}{N_A} \sum_{k=1}^{N_A} W(k), \qquad \overline{Q} = \frac{1}{N_A} \sum_{k=1}^{N_A} Q(k) \tag{6.19}$$

I risultati della simulazione permettono anche di ottenere il numero di utenti $N(t)$ presenti nel sistema al tempo t a partire dalle coppie (n, t), dove t denota l'istante di tempo in cui si è verificato un evento (arrivo o partenza) e dove n denota il numero di utenti presenti nel sistema al tempo t. Una stima della probabilità di avere k utenti nel sistema nell'intervallo $(0, t_f)$ può essere così

ottenuta:

$$\widehat{q}_k = \frac{\text{tempo trascorso nello stato } k}{t_f} \qquad (k = 0, 1, \ldots, N_A), \tag{6.20}$$

da cui è possibile ricavare una stima della media e della varianza del numero di utenti presenti nel sistema nell'intervallo $(0, t_f)$:

$$\overline{N} = \sum_{i=1}^{N_A} i\, \widehat{q}_i, \qquad \overline{N}_q = \sum_{i=2}^{N_A} (i-1)\, \widehat{q}_i \tag{6.21}$$

Esempio 6.2 Riconsideriamo il sistema di servizio descritto nell'Esempio 6.1. Utilizziamo ora l'algoritmo generale per la simulazione del sistema di servizio con unico servitore. I risultati della simulazione sono elencati in Tabella 6.5, in cui t indica il tempo di sistema, N_A il numero cumulativo di arrivi, N_U il numero di completamenti di servizio, n il numero di utenti nel sistema, n_q il numero di utenti in fila di attesa, I_s il tempo di inizio del servizio, $S(k)$ il tempo di servizio dell'utente k-esimo, F_s l'istante di completamento del servizio (evento futuro), $T(k)$ il tempo di interarrivo intercorrente tra il $(k-1)$-esimo e il k-esimo utente e P_a l'istante del prossimo arrivo (evento futuro).

t	N_A	N_U	n	n_q	I_s	$S(k)$	F_s	$T(k)$	P_a
0	0	0	0	0	-	-	-	1.73	-
1.73	1	0	1	0	1.73	2.90	4.63	1.35	3.08
3.08	2	0	2	1	-	-	-	0.71	3.79
3.79	3	0	3	2	-	-	-	0.62	4.41
4.41	4	0	4	3	-	-	-	14.28	18.69
4.63	4	1	3	2	-	1.76	6.39	-	-
6.39	4	2	2	1	-	3.39	9.78	-	-
9.78	4	3	1	0	-	4.52	14.3	-	-
14.3	4	4	0	0	-	-	∞	-	-
18.69	5	4	1	0	18.69	4.46	23.15	0.70	19.39
19.39	6	4	2	1	-	-	-	15.52	34.91
23.15	6	5	1	0	-	4.36	27.51	-	-
27.51	6	6	0	0	-	-	∞	-	-
34.91	7	6	1	0	34.91	2.07	37.00	3.15	38.06
37.00	7	7	0	0	-	-	∞	-	-
38.06	8	7	1	0	38.06	3.36	41.42	0.76	38.82
38.82	9	7	2	1	-	-	-	1.00	39.82
39.82	10	7	3	2	-	-	-	0.5	$40.32 > t_c$
41.42	10	8	2	1	-	2.37	43.79	-	-
43.79	10	9	1	0	-	5.38	49.17	-	-
49.17	10	10	0	0	-	-	∞	-	-

Tabella 6.5: Esempio di funzionamento del simulatore del sistema di servizio con singolo servitore con soltanto dieci arrivi.

Osservando il tempo di sistema $t = 39.82$ della Tabella 6.5 emerge che il tempo del prossimo arrivo $P_a = 40.32$ è maggiore del tempo $t_c = 40$ di chiusura

dell'accesso degli utenti al sistema. In questo istante di tempo sono arrivati 10 utenti, di cui soltanto 7 sono già stati serviti. Pertanto a partire da questo tempo di sistema occorre servire gli utenti ancora presenti nel sistema. Inoltre, dalla Tabella 6.5 emerge anche che l'istante di tempo in cui l'ultimo utente lascia il sistema è $t_f = 49.17$ minuti.

Dalla Tabella 6.5 si ricava:

$$N(t) = \begin{cases} 0, & 0 \le t < 1.73 \\ 1, & 1.73 \le t < 3.08 \\ 2, & 3.08 \le t < 3.79 \\ 3, & 3.79 \le t < 4.41 \\ 4, & 4.41 \le t < 4.63 \\ 3, & 4.63 \le t < 6.39 \\ 2, & 6.39 \le t < 9.78 \\ 1, & 9.78 \le t < 14.3 \\ 0, & 14.3 \le t < 18.69 \\ 1, & 18.69 \le t < 19.39 \\ 2, & 19.39 \le t < 23.15 \\ 1, & 23.15 \le t < 27.51 \\ 0, & 27.51 \le t < 34.91 \\ 1, & 34.91 \le t < 37.00 \\ 0, & 37.00 \le t < 38.06 \\ 1, & 38.06 \le t < 38.82 \\ 2, & 38.82 \le t < 39.82 \\ 3, & 39.82 \le t < 41.42 \\ 2, & 41.42 \le t < 43.79 \\ 1, & 43.79 \le t < 49.17 \\ 0, & t = 49.17, \end{cases} \qquad N_q(t) = \begin{cases} 0, & 0 \le t < 3.08 \\ 1, & 3.08 \le t < 3.79 \\ 2, & 3.79 \le t < 4.41 \\ 3, & 4.41 \le t < 4.63 \\ 2, & 4.63 \le t < 6.39 \\ 1, & 6.39 \le t < 9.78 \\ 0, & 9.78 \le t < 19.39 \\ 1, & 19.39 \le t < 23.15 \\ 0, & 23.15 \le t < 38.82 \\ 1, & 38.82 \le t < 39.82 \\ 2, & 39.82 \le t < 41.42 \\ 1, & 41.42 \le t < 43.79 \\ 0, & 43.79 \le t < 49.17 \\ 0, & t = 49.17, \end{cases} \tag{6.22}$$

da cui è possibile ricostruire la realizzazione, illustrata in Figura 6.7, del numero di utenti presenti nel sistema di servizio con singolo servitore.

Inoltre, a partire dalla Tabella 6.22 è anche possibile ricostruire una realizzazione, illustrata in Figura 6.8, del numero di utenti presenti in fila di attesa.

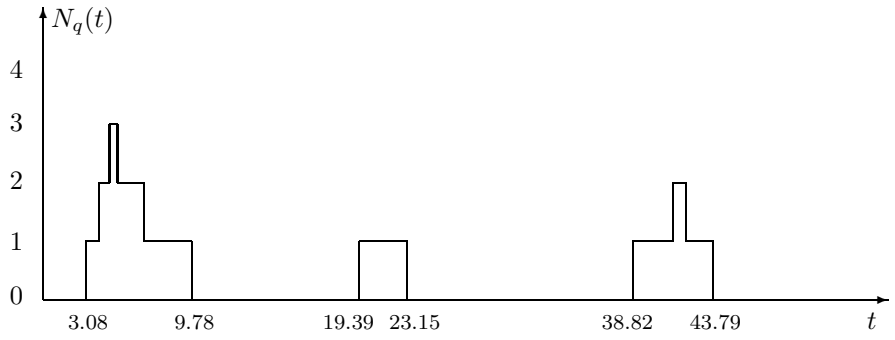

Figura 6.8: Una realizzazione del numero di utenti presenti in fila di attesa.

Facendo uso di (6.22) in (6.20) si ottiene poi una stima delle probabilità di

avere k utenti nel sistema nell'intervallo $(0, 49.17)$:

$$\widehat{q}_0 = \frac{1.73 + 4.39 + 7.4 + 1.06}{49.17} = \frac{14.58}{49.17},$$
$$\widehat{q}_1 = \frac{1.35 + 4.52 + 0.7 + 4.36 + 2.09 + 0.76 + 5.38}{49.17} = \frac{19.16}{49.17},$$
$$\widehat{q}_2 = \frac{0.71 + 3.39 + 3.76 + 1.00 + 2.37}{49.17} = \frac{11.23}{49.17},$$
$$\widehat{q}_3 = \frac{0.62 + 1.76 + 1.6}{49.17} = \frac{3.98}{49.17},$$
$$\widehat{q}_4 = \frac{0.22}{49.17}.$$

Utilizzando poi la (6.21) con $N_A = 10$ si ricava una stima del numero medio di utenti nel sistema nell'intervallo $(0, 49.17)$:

$$\overline{N} = \left[1 \cdot \frac{19.16}{49.17} + 2 \cdot \frac{11.23}{49.17} + 3 \cdot \frac{3.98}{49.17} + 4 \cdot \frac{0.22}{49.17}\right] = \frac{54.44}{49.17} = 1.107.$$

Ricordando la (6.22) si ottiene anche una stima delle probabilità di avere k utenti nella fila di attesa nell'intervallo $(0, 49.17)$:

$$\widetilde{q}_0 = \frac{3.08 + 9.61 + 15.67 + 5.38}{49.17} = \frac{33.74}{49.17},$$
$$\widetilde{q}_1 = \frac{0.71 + 3.39 + 3.76 + 1.00 + 2.37}{49.17} = \frac{11.23}{49.17},$$
$$\widetilde{q}_2 = \frac{0.62 + 1.76 + 1.6}{49.17} = \frac{3.98}{49.17},$$
$$\widetilde{q}_3 = \frac{0.22}{49.17},$$

da cui utilizzando la (6.21) con $N_A = 10$ si ricava una stima del numero medio di utenti in fila di attesa nell'intervallo $(0, 49.17)$:

$$\overline{N}_q = \left[1 \cdot \frac{11.23}{49.17} + 2 \cdot \frac{3.98}{49.17} + 3 \cdot \frac{0.22}{49.17}\right] = \frac{19.85}{49.17} = 0.4037.$$

Una stima del fattore di utilizzazione del sistema è quindi $\overline{N}_s = \overline{N} - \overline{N}_q = 0.7035$. ♢

Nel Capitolo 7 ci occuperemo della generazione di variabili aleatorie uniformemente distribuite nell'intervallo $(0, 1)$ individuando algoritmi computazionalmente efficienti. Successivamente, nel Capitolo 8 ci interesseremo della generazione di variabili aleatorie con un qualsiasi tipo di funzione di distribuzione. Abbiamo infatti visto che nella simulazione dei sistemi di servizio occorre generare variabili aleatorie continue utili a descrivere i tempi di interarrivo e i tempi di servizio. Spesso occorre anche simulare variabili aleatorie discrete utili, ad esempio, a descrivere il numero di telefonate in arrivo ad un centralino telefonico.

Capitolo 7

Generatori uniformi

7.1 Introduzione

Nella simulazione giocano un ruolo fondamentale i *numeri casuali* uniformemente distribuiti in un fissato intervallo. In passato sono state proposte varie tecniche per generare numeri casuali. Una consisteva nel dotare l'elaboratore di una speciale apparecchiatura capace di generare numeri casuali sfruttando qualche particolare fenomeno fisico allo scopo di *produrre tabelle di numeri casuali* che gli studiosi potessero poi utilizzare nelle loro simulazioni. Furono così costruite delle *macchine* sia di *tipo meccanico* sia di *tipo elettronico* che presentavano però l'inconveniente di imporre una *laboriosa manutenzione* per garantire l'efficienza di apparecchiature, spesso delicate, atte a generare sequenze casuali, che risultavano talvolta poco adatte alle applicazioni desiderate.

Il metodo più classico per ottenere sequenze casuali uniformemente distribuite è il *metodo dell'urna* in cui dischetti numerati sono messi in un'urna e mescolati prima di ogni estrazione; ogni dischetto estratto, dopo essere stato letto, è rimesso nell'urna. Questo metodo ha discrete caratteristiche random ma le sequenze ottenute, come tutte quelle effettivamente basate su un processo di successive estrazioni con rimpiazzamento, presentano l'inconveniente di essere *non ripetibili*, a meno che non si proceda ad una laboriosa ed ingombrante *registrazione di tutti i valori estratti in una tabella*, che può essere inserita in un elaboratore e successivamente utilizzata per generare sequenze casuali. Ogni metodo che porta alla costruzione di tabelle di numeri casuali, anche se permette di riottenere gli stessi risultati utilizzando gli stessi numeri, è *costoso* sia per lo *spazio di memoria occupato dalla tabella*, sia per il *tempo richiesto per accedervi* e sia poiché per analizzare determinati problemi potrebbero essere necessari molti più numeri di quelli presenti nelle tabelle.

Per superare gli inconvenienti finora discussi sono stati sviluppati metodi che permettono di ottenere dagli elaboratori sequenze di numeri casuali attraverso il

ripetuto uso di un *meccanismo algebrico deterministico* (ossia mediante opportune formule di ricorrenza) che presentano il vantaggio, rispetto alle altre tecniche già citate, di essere facilmente implementabili con algoritmi estremamente veloci computazionalmente, di richiedere poco spazio di memoria, di permettere di riprodurre sequenze identiche a quelle già utilizzate in modo da riottenere gli stessi risultati.

Storicamente i primi a proporre una tecnica di questo tipo furono von Neumann e Metropolis nel 1946 con il *metodo del centro del quadrato*. La procedura utilizzata per generare una sequenza $\alpha_0, \alpha_1, \ldots$ è la seguente. Un arbitrario numero positivo α_0, detto seme (o valore iniziale), è scelto come input per generare il processo ricorrente. Tale numero α_0, rappresentato con $2k$ digits, è elevato al quadrato. Si produce così un numero α_0^2 di $4k$ digits (inserendo, se necessario, degli 0 alla sinistra del numero per formare esattamente $4k$ digits) dal quale viene estratto un numero α_1 costituito dai $2k$ digits centrali di α_0^2 (includendo quindi i bits da $k+1$ a $3k$), che a sua volta viene elevato al quadrato per generare α_2. Si prosegue poi in questo modo per generare gli altri numeri della sequenza.

Esempio 7.1 Sia $\alpha_0 = 8234$. Con il metodo del centro del quadrato si ottiene la sequenza

$$\begin{aligned}
&\alpha_0 = 8234 \\
&\alpha_0^2 = 67798756 \Longrightarrow \alpha_1 = 7987 \\
&\alpha_1^2 = 63792169 \Longrightarrow \alpha_2 = 7921 \\
&\alpha_2^2 = 62742241 \Longrightarrow \alpha_3 = 7422 \\
&\alpha_3^2 = 55086084 \Longrightarrow \alpha_4 = 0860 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

♢

Il metodo del centro del quadrato ha scarse qualità statistiche; inoltre, il valore iniziale deve essere scelto accuratamente, come mostrato nei due esempi seguenti.

Esempio 7.2 Sia $\alpha_0 = 7182$. Con il metodo del centro del quadrato si ottiene la sequenza

$$\begin{aligned}
&\alpha_0 = 7182 \\
&\alpha_0^2 = 51581124 \Longrightarrow \alpha_1 = 5811 \\
&\alpha_1^2 = 33767721 \Longrightarrow \alpha_2 = 7677 \\
&\alpha_2^2 = 58936329 \Longrightarrow \alpha_3 = 9363 \\
&\alpha_3^2 = 87665769 \Longrightarrow \alpha_4 = 6657 \\
&\alpha_4^2 = 44315649 \Longrightarrow \alpha_5 = 3156 \\
&\alpha_5^2 = 09960336 \Longrightarrow \alpha_6 = 9603 \\
&\alpha_6^2 = 92217609 \Longrightarrow \alpha_7 = 2176 \\
&\alpha_7^2 = 04734976 \Longrightarrow \alpha_8 = 7349
\end{aligned}$$

$$\alpha_8^2 = 54007801 \Longrightarrow \alpha_9 = 0078$$
$$\alpha_9^2 = 00006084 \Longrightarrow \alpha_{10} = 0060$$
$$\alpha_{10}^2 = 00003600 \Longrightarrow \alpha_{11} = 0036$$
$$\alpha_{11}^2 = 00001296 \Longrightarrow \alpha_{12} = 0012$$
$$\alpha_{12}^2 = 00000144 \Longrightarrow \alpha_{13} = 0001$$
$$\alpha_{13}^2 = 00000001 \Longrightarrow \alpha_{14} = 0000$$
$$\alpha_{14}^2 = 00000000 \Longrightarrow \alpha_{15} = 0000$$

Si nota che $\alpha_{14} = \alpha_{15} = \ldots = 0$. ♢

Esempio 7.3 Sia $\alpha_0 = 99$. Con il metodo del centro del quadrato si ottiene la sequenza

$$\alpha_0 = 99$$
$$\alpha_0^2 = 9801 \Longrightarrow \alpha_1 = 80$$
$$\alpha_1^2 = 6400 \Longrightarrow \alpha_2 = 40$$
$$\alpha_2^2 = 1600 \Longrightarrow \alpha_3 = 60$$
$$\alpha_3^2 = 3600 \Longrightarrow \alpha_4 = 60$$

Si nota che $\alpha_3 = \alpha_4 = \ldots = 60$. ♢

Il metodo del centro del quadrato agli inizi conseguì un discreto successo ma ben presto non venne più utilizzato principalmente per tre motivi:

(a) Relativa lentezza

Il metodo si rivelava lento computazionalmente nella generazione di sequenze di numeri sia a causa dell'operazione di esponenziazione sia per la successiva operazione di selezione delle cifre centrali del numero ottenuto al passo precedente.

(b) Difficoltà analitiche

Queste consistono soprattutto nel determinare la lunghezza del ciclo della sequenza casuale ottenuta a partire da un generico valore iniziale fino al valore per il quale la sequenza inizia a ripetersi.

(c) Insoddisfacente comportamento statistico delle sequenze ottenute.

I numeri generati con questo metodo non sono uniformemente distribuiti ed inoltre viene meno l'indipendenza statistica tra gli elementi della sequenza.

Il metodo del centro del quadrato fu, quindi, abbandonato e ad esso seguirono i *metodi congruenti*. Attualmente proprio questi metodi e loro varianti sono spesso utilizzati nelle applicazioni.

In generale, un processo è veramente casuale se le predizioni sul suo comportamento futuro non possono essere migliorate dalla conoscenza del comportamento passato. Nell'adottare meccanismi algebrici deterministici per la generazione di sequenze casuali, si può incorrere nell'inconveniente di venire meno all'indipendenza statistica tra gli elementi della sequenza stessa. In processi di simulazione il termine casuale è quindi generalmente sostituito con il termine *pseudocasuale*.

Un metodo per la generazione di sequenze di numeri pseudocasuali con distribuzione uniforme è accettabile se soddisfa ai seguenti requisiti:

(i) i numeri debbono essere statisticamente uniformemente distribuiti nella sequenza;

(ii) i numeri debbono essere statisticamente indipendenti nella sequenza;

(iii) la sequenza deve essere riproducibile;

(iv) la sequenza deve poter avere un ciclo di lunghezza abbastanza grande;

(v) il metodo deve poter essere eseguito dall'elaboratore con rapidità e deve occupare poco spazio di memoria.

7.2 Metodo congruenziale moltiplicativo

I generatori congruenziali moltiplicativi producono sequenze $\{x_n,\ n = 0, 1, 2, \ldots\}$ come segue:

(i) fissare un intero positivo m detto *modulo* del generatore

(ii) scegliere degli interi positivi a e x_0 minori del modulo m; x_0 ($x_0 \neq 0$) è detto *valore iniziale* o *seme* e la costante a ($a \neq 0$) è detta *costante moltiplicativa* oppure *moltiplicatore*.

(iii) generare x_n mediante la relazione di congruenza lineare

$$x_{n+1} \equiv a\, x_n \pmod{m} \tag{7.1}$$

che si legge x_{n+1} *è congruente ad* $a\, x_n$ *modulo* m.

La procedura inizia con un valore iniziale x_0 che deve essere diverso da zero. Per determinare gli elementi della sequenza $\{x_n,\ n = 1, 2, \ldots\}$ occorre assegnare a x_{n+1} il resto r (con $0 \leq r \leq m-1$) della divisione di $a\, x_n$ per il modulo m.

La relazione di ricorrenza (7.1) è analoga all'equazione alle differenze del primo ordine $x_{n+1} = a\, x_n$ che ammette come soluzione

$$x_n = x_0\, a^n.$$

La relazione di congruenza lineare (7.1) può quindi essere così riscritta

$$x_n \equiv x_0\, a^n \pmod{m}. \tag{7.2}$$

Esempio 7.4 Sia $m = 32 = 2^5$, $x_0 = 1$ e $a = 3$. La relazione (7.1) diventa

$$x_{n+1} \equiv 3\,x_n \pmod{2^5}$$

e conduce alla sequenza

$$\begin{aligned}
&x_0 = 1\\
&x_1 \equiv 3\,x_0 = 3 \pmod{32} \Longrightarrow x_1 = 3\\
&x_2 \equiv 3\,x_1 = 9 \pmod{32} \Longrightarrow x_2 = 9\\
&x_3 \equiv 3\,x_2 = 27 \pmod{32} \Longrightarrow x_3 = 27\\
&x_4 \equiv 3\,x_3 = 81 \pmod{32} \Longrightarrow x_4 = 17\\
&x_5 \equiv 3\,x_4 = 51 \pmod{32} \Longrightarrow x_5 = 19\\
&x_6 \equiv 3\,x_5 = 57 \pmod{32} \Longrightarrow x_6 = 25\\
&x_7 \equiv 3\,x_6 = 75 \pmod{32} \Longrightarrow x_7 = 11\\
&x_8 \equiv 3\,x_7 = 33 \pmod{32} \Longrightarrow x_8 = 1.
\end{aligned}$$

Si nota che a partire da x_8 la sequenza 1, 3, 9, 27, 17, 19, 25, 11 comincia a ripetersi. ♢

Il più piccolo intero p tale che

$$x_0 = x_p \tag{7.3}$$

è detto *periodo fondamentale della sequenza*, ossia il periodo fondamentale rappresenta la lunghezza del ciclo della sequenza a partire da una generico valore iniziale fino al valore per il quale la sequenza inizia a ripetersi. Poiché i valori generati dalla (7.1) sono sempre minori del modulo m, è chiaro che tra due valori identici non possono presentarsi più di m valori diversi; quindi *la lunghezza del periodo fondamentale non può essere superiore a* m.

Il *vantaggio del generatore congruente moltiplicativo* è l'estrema velocità di generazione dei numeri. Invece, *alcuni svantaggi* sono:

(a) la sequenza generata è periodica di periodo al più uguale a m;

(b) ogni valore della sequenza è completamente determinato dai tre parametri a, x_0 e m;

(c) esiste una correlazione tra i valori successivi della sequenza.

Come si vedrà nel seguito di questo capitolo occorre effettuare una scelta accurata dei parametri a, x_0 e m del generatore congruente moltiplicativo.

7.2.1 Scelta del modulo come potenza di 2

Sebbene non esistano restrizioni sulla scelta del modulo m, ai fini dell'implementazione su elaboratori binari una possibile scelta del modulo è $m = 2^b$ per rendere più veloce la generazione di ogni numero x_{n+1} della sequenza a partire

dal numero x_n usando la relazione congruente (7.1). Con questa scelta la (7.2) diventa

$$x_{n+1} \equiv a\, x_n \pmod{2^b}. \tag{7.4}$$

Vogliamo ora mostrare che nella relazione di ricorrenza (7.4) è conveniente scegliere i parametri a e x_0 nel seguente modo:

(a) $a = 3, 5, \ldots, 2^b - 1$

(b) $x_0 = 1, 3, 5, \ldots, 2^b - 1$

(a) **Scelta della costante moltiplicativa** a

Occorre scegliere $a \neq 1$; infatti se $a = 1$ si ripeterebbe sempre lo stesso valore iniziale della sequenza generata da (7.4).

Esempio 7.5 Sia $m = 2^5$, $x_0 = 1$ e $a = 1$. La relazione (7.1) diventa

$$x_{n+1} \equiv x_n \pmod{2^5}$$

e conduce alla sequenza $x_0 = x_1 = x_2 = \ldots = 1$. ♢

Se a è pari, ossia $a = 2k$, l'equazione (7.4) diventa

$$x_{n+1} \equiv 2\, k\, x_n \pmod{2^b},$$

ossia per la (7.2):

$$x_n \equiv x_0\, (2\, k)^n \pmod{2^b}.$$

Quando $n = b$ tale relazione di congruenza conduce a

$$x_b \equiv x_0\, 2^b\, k^b \pmod{2^b},$$

da cui si deduce che $x_b = 0$; tutti i numeri successivamente generati sono quindi nulli, ossia $x_{b+1} = x_{b+2} = \ldots = 0$. Quindi se a è pari, a partire da $n = b$ viene meno l'indipendenza statistica dei valori generati.

Esempio 7.6 Sia $m = 2^5$, $x_0 = 1$ e $a = 2$. La relazione (7.1) diventa

$$x_{n+1} \equiv 2\, x_n \pmod{2^5}$$

e conduce alla sequenza

$$x_0 = 1,\ x_1 = 2,\ x_2 = 4,\ x_3 = 8,\ x_4 = 16,\ x_5 = 0,\ x_6 = 0, \ldots$$

♢

Occorre quindi scegliere a intero positivo dispari diverso da 1 e minore del modulo $m = 2^b$.

(b) **Scelta del valore iniziale** x_0

Se x_0 è pari, ossia $x_0 = 2\, k$, l'equazione (7.2) diventa

$$x_n \equiv 2\, k\, a^n \pmod{2^b}$$

L'operazione di determinare il resto della divisione di $2\,k\,a^n$ per 2^b conduce allo stesso risultato di dividere $k\,a^n$ per 2^{b-1} ossia

$$\frac{2\,k\,a^n}{2^b} = \frac{k\,a^n}{2^{b-1}}.$$

L'equazione di congruenza $x_n \equiv 2\,k\,a^n \pmod{2^b}$ è equivalente a calcolare $x_n = 2\,y_n$, dove $y_n \equiv k\,a^n \pmod{2^{b-1}}$. Scegliere quindi x_0 pari, ossia $x_0 = 2\,k$, è equivalente a ridurre il modulo m da 2^b ad almeno 2^{b-1}. Ovviamente se il resto della divisione è pari si può iterare il ragionamento ottenendo un modulo ancora più piccolo. Poiché il modulo stabilisce il limite superiore della lunghezza del periodo, essere riusciti a ridurre il modulo corrisponde ad aver considerato sequenze il cui periodo si discosta di molto dal limite superiore $m = 2^b$.

Esempio 7.7 Consideriamo le relazioni di congruenza

$$x_{n+1} \equiv 3\,x_n \pmod{2^5}, \qquad y_{n+1} \equiv 3\,y_n \pmod{2^4}$$

e sia $x_0 = 2$ e $y_0 = 1$. Quindi si ha

$$x_n \equiv 2\,3^n \pmod{2^5}, \qquad y_n \equiv 3^n \pmod{2^4},$$

da cui si ottengono rispettivamente le seguenti sequenze:

$$x_0 = 2,\ x_1 = 6,\ x_2 = 18,\ x_3 = 22,\ x_4 = 2, \ldots$$
$$y_0 = 1,\ y_1 = 3,\ y_2 = 9,\ \ y_3 = 11,\ y_4 = 1, \ldots$$

Essendo x_0 pari, si ha $x_n = 2\,y_n$ $(n = 0, 1, \ldots)$. In entrambe le sequenze il periodo fondamentale è 4; tale periodo si discosta molto da $m = 2^5$. ♢

Occorre quindi scegliere x_0 dispari.

(c) **Altre considerazioni sulla costante moltiplicativa** a

Il modulo m stabilisce soltanto il limite superiore della lunghezza del periodo, che è invece fortemente influenzato dal valore del moltiplicatore a. Tale valore deve essere scelto in maniera tale che il periodo non si discosti molto dal limite superiore 2^b, ma anche in modo tale da avere una sequenza con i requisiti di casualità richiesti.

Esempio 7.8 Sia $m = 2^5$, $x_0 = 1$ con $a = 3, 5, 7, \ldots, 31$. La relazione (7.1) diventa

$$x_{n+1} \equiv a\,x_n \pmod{2^5}$$

e conduce alle sequenze indicate nella Tabella 7.1:

Osservando la Tabella 7.1 si può notare che

(a) il massimo periodo è 8 e si ottiene in otto casi su 15; in quattro casi il periodo è 4 e in tre casi il periodo è 2.

a	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31
x_0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
x_1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31
x_2	9	25	17	17	25	9	1	1	9	25	17	17	25	9	1
x_3	27	29	23	25	19	21			11	13	7	9	3	5	
x_4	17	17	1	1	17	17			17	17	1	1	17	17	
x_5	19	21			27	29			3	5			11	13	
x_6	25	9			9	25			25	9			9	25	
x_7	11	13			3	5			27	29			19	21	
x_8	1	1			1	1			1	1			1	1	
p	8	8	4	4	8	8	2	2	8	8	4	4	8	8	2

Tabella 7.1: Sequenze prodotte con il generatore congruente moltiplicativo per $m = 2^5$, $x_0 = 1$ e $a = 3, 5, 7, \ldots, 31$.

(b) la sequenza ottenuta per $a = 11$ è l'immagine speculare di quella ottenuta per $a = 3$; analogamente $a = 5$ e $a = 13$, $a = 19$ e $a = 27$, $a = 21$ e $a = 29$ sono immagini speculari.

(c) le sequenze aventi periodo maggiore che rispecchiano meglio la distribuzione uniforme sono quelle ottenute scegliendo $a = 5, 13, 21, 29$, ossia $1, 5, 9, 13, 17, 21, 25, 29$ (tutti i numeri sono distanziati di 4). Inoltre le sequenze ottenute scegliendo $a = 3, 11, 19, 27$ (aventi anch'esse periodo 8), ossia $1, 3, 9, 11, 17, 19, 25, 27$ sono più raggruppate e quindi rispecchiano meno la distribuzione uniforme.

$\diamondsuit$

Se si utilizza il generatore congruente moltiplicativo (7.1), per determinare una costante moltiplicativa a^* diversa da a che genera la *stessa sequenza in ordine speculare* (inverso) occorre scegliere

$$a^* \equiv a^{p-1} \pmod{m},$$

dove p è il periodo fondamentale della sequenza (lunghezza del ciclo). Infatti, scegliendo $m = 2^5$, dalla Tabella 7.1 si nota che se $a = 3$ allora $a^* = 11$, se $a = 5$ allora $a^* = 13$, se $a = 19$ allora $a^* = 27$ e se $a = 21$ allora $a^* = 29$.

Inoltre, nel generatore congruente moltiplicativo (7.1) per determinare una *sequenza antitetica* $x_1^*, x_2^*, \ldots$, ossia tale che $x_n^* = m - x_n$ $(n = 1, 2, \ldots)$, basta scegliere come seme della nuova sequenza $x_0^* = m - x_0$. Ad esempio, scegliendo $a = 3$, $m = 2^5$, $x_0 = 1$ e $x_0^* = m - x_0 = 31$, i due generatori moltiplicativi seguenti

$$x_{n+1} \equiv 3\, x_n \pmod{2^5}, \qquad x_{n+1}^* \equiv 3\, x_n^* \pmod{2^5}$$

producono rispettivamente le sequenze:

$$x_0 = 1,\ x_1 = 3,\ x_2 = 9,\ x_3 = 27,\ x_4 = 17,\ x_5 = 19,\ x_6 = 25,\ x_7 = 11$$
$$x_0^* = 31,\ x_1^* = 29,\ x_2^* = 23,\ x_3^* = 5,\ x_4^* = 15,\ x_5^* = 13,\ x_6^* = 7,\ x_7^* = 21$$

e si nota che $x_n + x_n^* = 2^5$ $(n = 0, 1, \ldots)$.

Per scegliere opportunamente a in maniera tale da ottenere un periodo massimo viene in aiuto la teoria dei numeri e si può dimostrare il seguente risultato.

Proposizione 7.1
(a) Sia $m = 2^b$ con $b \geq 4$. Se si scelgono i parametri a e x_0 del generatore congruente moltiplicativo minori del modulo e tali che

(i) x_0 positivo dispari,

(ii) $a = 8\,n + 3$ oppure $a = 8\,n + 5$, dove n è un qualsiasi intero non negativo,

si ottiene il periodo massimo 2^{b-2}.

(b) Sia $m = 10^b$ con $b \geq 5$. Se si scelgono i parametri del generatore congruente moltiplicativo minori del modulo e tali che che

(i) x_0 positivo dispari e non divisibile per 5,

(ii) $a = 200\,n \pm z$ dove z può assumere uno dei seguenti 32 valori 3, 11, 13, 19, 21, 27, 29, 37, 53, 59, 61, 67, 69, 77, 83, 91, 109, 117, 123, 131, 133, 139, 141, 147, 163, 171, 173, 179, 181, 187, 189, 197,

si ottiene il periodo massimo $5 \cdot 10^{b-2}$.

Esempio 7.9 Riconsideriamo la Tabella 7.1 in cui $m = 2^5$, $x_0 = 1$, $a = 3, 5, 7, \ldots, 31$. Le condizioni della Proposizione 7.1 sono soddisfatte. Infatti $x_0 = 1$ è dispari e se si sceglie $a = 8n+3$ oppure $a = 8n+5$, ossia $a = 3, 11, 19, 27$ oppure $a = 5, 13, 21, 29$, si ottiene il periodo è massimo $2^3 = 8$. ♢

Essendo il periodo del moltiplicatore sempre minore del modulo m, con il metodo congruenziale moltiplicativo non tutti i numeri sono presenti nel periodo. Il metodo presenta nell'intervallo $[0, m-1)$ delle zone vuote, ossia zone in cui non si presentano numeri. La distribuzione e la distanza tra le zone vuote varia a seconda del seme e del moltiplicatore.

Per elaboratori binari i più comuni valori del modulo m sono 2^{31} e 2^{35}.

Per un elaboratore binario con parola macchina di 32 bit un algoritmo introdotto nel 1960 dall'IBM (detto RANDU) era caratterizzato da $m = 2^{31} = 2.147.483.648$ e $a = 65539$; in tal caso poiché $a = 8 \cdot 8192 + 3 = 2^{16} + 3$ il periodo è 2^{29}. Knuth (1981) ha dimostrato che tale generatore non ha buone proprietà di casualità. L'algoritmo SIMULA invece utilizza $m = 2^{35}$, $a = 5^{13} = 1.220.703.125$; in tal caso poiché $a = 8 \cdot 152587890 + 5$ il periodo è 2^{33}. Park e Miller nel 1988 hanno mostrato che anche tale generatore non ha buone proprietà statistiche.

Il passaggio dalla sequenza pseudocasuale $x_0, x_1, \ldots$ (con $0 \leq x_n < m$) ad una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n < 1$) può essere effettuata ponendo

$$u_n = \frac{x_n}{m} \qquad (n = 0, 1, \ldots). \tag{7.5}$$

Analogamente, se si desidera passare dalla sequenza pseudocasuale $x_0, x_1, \ldots$ (con $0 \leq x_n \leq m-1$) ad una sequenza pseudo–casuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n \leq 1$) basterà invece porre

$$u_n = \frac{x_n}{m-1} \qquad (n = 0, 1, \ldots). \tag{7.6}$$

Le (7.5) e (7.6) possono quindi essere utilizzate per generare numeri uniformemente distribuiti nell'intervallo $[0, 1)$ e $[0, 1]$, rispettivamente.

I passi dell'algoritmo per generare una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n < 1$) utilizzando il metodo congruenziale moltiplicativo con $m = 2^b$ sono quindi i seguenti:

Algoritmo
STEP 1: fornire in input x_0, a e b tali da soddisfare le ipotesi della Proposizione 7.1
STEP 2: per ogni $n = 1, 2, \ldots, 2^{b-2} - 1$ calcolare

$$x_n \equiv a\, x_{n-1} \pmod{2^b}$$

STEP 3: per ogni $n = 0, 1, \ldots, 2^{b-2} - 1$ calcolare

$$u_n = x_n \cdot 2^{-b}.$$

7.2.2 Scelta del modulo come numero primo

Nel generatore congruente moltiplicativo (7.1) si può scegliere m come numero primo. Si può dimostrare il seguente risultato.

Proposizione 7.2 *Se si scelgono i parametri a e m del generatore congruente moltiplicativo (7.1) tali che*

(i) m è un numero primo (ad esempio $m = 2^{31} - 1$);

(ii) a (minore del modulo) è un elemento primitivo modulo m

si ottiene il periodo massimo $m - 1$.

Il numero *a è un elemento primitivo modulo m se il più piccolo numero intero s per il quale $(a^s - 1)$ è divisibile per m è proprio $s = m - 1$*. Quindi, a è un elemento primitivo modulo m se $a^s - 1$ è un multiplo di m per $s = m - 1$ ma non lo è per valori interi s più piccoli di $m - 1$.

Il generatore congruente moltiplicativo utilizzato nella Proposizione 7.2, genera nel periodo ogni intero compreso nell'intervallo $[1, m - 1)$ prima che la sequenza cominci nuovamente a ripetersi.

Esempio 7.10 Consideriamo il generatore congruente moltiplicativo

$$x_{n+1} \equiv 5\, x_n \pmod{7}$$

Tale generatore soddisfa le ipotesi della Proposizione 7.2. Infatti, $m = 7$ è un numero primo ed inoltre $a = 5$ è un elemento primitivo modulo 7, poiché il più piccolo intero s tale che $a^s - 1 = 5^s - 1$ è divisibile per 7 è proprio $s = 6$. Infatti, si nota che

$$a^s - 1 = 5^s - 1 = \begin{cases} 4, & s = 1 \\ 24, & s = 2 \\ 124, & s = 3 \\ 624, & s = 4 \\ 3124, & s = 5 \\ 15624, & s = 6 \end{cases}$$

e risulta $15624 = 2232 \cdot 7$. Il generatore ha quindi un periodo massimo pari a $m - 1 = 6$.

x_0	x_1	x_2	x_3	x_4	x_5	x_6	p
1	5	4	6	2	3	1	6
2	3	1	5	4	6	2	6
3	1	5	4	6	2	3	6
4	6	2	3	1	5	4	6
5	4	6	2	3	1	5	6
6	2	3	1	5	4	6	6

Tabella 7.2: Sequenze prodotte con il generatore congruente moltiplicativo per $m = 7$, $a = 5$ e $x_0 = 1, 2, 3, 4, 5, 6$.

Nella Tabella 7.2 sono visualizzate tutte le sequenze generate per $m = 7$, $a = 5$, $x_0 = 1, 2, 3, 4, 5, 6$ e il loro rispettivo periodo $p = m - 1 = 6$. ♢

Esempio 7.11 Consideriamo il generatore congruente moltiplicativo

$$x_{n+1} \equiv 3\, x_n \pmod{31}$$

Tale generatore soddisfa le ipotesi della Proposizione 7.2. Infatti, $m = 31$ è un numero primo ed inoltre $a = 3$ è un elemento primitivo modulo 31, poiché il più piccolo intero s tale che $a^s - 1 = 3^s - 1$ è divisibile per 31 è proprio $s = 30$. Infatti, le radici primitive di 31 sono $a = 3, 11, 12, 13, 17, 21, 22, 24$.

Partendo con un seme $x_0 = 1$ si ottiene la sequenza in Tabella 7.3 il cui periodo è 30. ♢

I passi dell'algoritmo per generare una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n < 1$) utilizzando il metodo congruenziale moltiplicativo con $m = 2^b - 1$ sono quindi i seguenti:

Algoritmo
STEP 1: fornire in input x_0, a e m tali da soddisfare le ipotesi della Proposizione 7.2
STEP 2: per ogni $n = 1, 2, \ldots, m - 2$ calcolare

$$x_n \equiv a\, x_{n-1} \pmod{m}$$

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	3	9	27	19	26	16	17	20	29	25
	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
	13	8	24	10	30	28	22	4	12	5
	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}
	15	14	11	2	6	18	23	7	21	1

Tabella 7.3: Sequenza prodotta con il generatore congruente moltiplicativo $x_{n+1} \equiv 3\,x_n \pmod{31}$ con $x_0 = 1$.

STEP 3: per ogni $n = 0, 1, \ldots, m-2$ calcolare

$$u_n = x_n/m.$$

Una scelta spesso utilizzata dei parametri del generatore congruente moltiplicativo con m numero primo è $m = 2^{31} - 1 = 2.147.483.647$, $a = 7^5 = 16807$ oppure $a = 630360016$ oppure $a = 742938285$ oppure $a = 397204094$. Nella Tabella 7.4 sono riportati alcuni semi che conducono, almeno nella fase iniziale, a differenti sequenze di numeri pseudocasuali per il generatore congruente moltiplicativo

$$x_{n+1} \equiv 7^5\,x_n \pmod{2^{31}-1}. \tag{7.7}$$

Tale generatore ha periodo massimo $2^{31} - 2$.

	Seme Iniziale		Seme Iniziale
1	748932582	16	1651217741
2	1985072130	17	909094944
3	1631331038	18	2095891343
4	67377721	19	203905359
5	366304404	20	2001697019
6	1094585182	21	431442774
7	1767585417	22	1659181395
8	1980520317	23	400219676
9	392682216	24	1904711401
10	64298628	25	263704907
11	250756106	26	350425820
12	1025663860	27	873344587
13	186056398	28	1416387147
14	522237216	29	1881263549
15	213453332	30	1456845529

Tabella 7.4: Semi da utilizzare per il generatore congruente moltiplicativo in (7.7).

Il generatore congruente moltiplicativo (7.7) è stato estensivamente analizzato nella letteratura e fornisce uno standard minimo qualitativo che deve

possedere un generatore di numeri pseudocasuali. Alcuni difetti di questo generatore sono che spesso numeri piccoli tendono ad essere seguiti da numeri grandi e che esiste una correlazione tra numeri successivi, come è possibile evidenziare disponendo i numeri presi a coppie in un grafico bidimensionale.

A partire dal generatore congruente moltiplicativo (7.7) sono stati introdotti in letteratura generatori sempre più complessi in grado da soddisfare vari tipi di test statistici. Molti di questi generatori sono ottenuti combinando due o più generatori congruenziali moltiplicativi in modo tale da ottenere periodi più lunghi.

Dalla sequenza prodotta con il generatore congruente moltiplicativo otteniamo tramite la (7.5) una sequenza di numeri $u_0, u_1, \ldots$ in $[0, 1)$. Suddividiamo tale sequenza in sottosequenze consecutive di n numeri che sono utilizzate come coordinate di punti di un cubo n–dimensionale di lato unitario. Se i numeri fossero non correlati tenderebbero a coprire tutto il cubo n–dimensionale. In realtà, Marsaglia nel 1968 ha dimostrato che tali punti cadono al più in $(n!\, m)^{1/n}$ iperpiani $(k-1)$–dimensionali paralleli. Se, ad esempio, si considera una sequenza suddivisa in sottosequenze di tre numeri consecutivi che sono utilizzati come punti di un cubo di lato unitario, si nota che tali punti cadono al più in $(3!\, m)^{1/3}$ piani paralleli; se si sceglie $m = 2^{15}$, tali punti cadono al più in 58 piani paralleli.

Nella Tabella 7.5 per $m = 2^{31}$ è indicato il valore approssimato $(n!\, m)^{1/n}$, ossia la maggiorazione di Marsaglia per il numero di distinti iperpiani paralleli per alcune scelte di n.

n	$(n!\, m)^{1/n}$
1	2^{31}
2	2^{16}
3	2344
4	476
5	191
6	107

Tabella 7.5: Per $m = 2^{31}$ è indicata la maggiorazione di Marsaglia per il numero di distinti iperpiani paralleli per alcune scelte di n.

In alcune applicazioni, come ad esempio nel caso del metodo di Monte Carlo per il calcolo di integrali unidimensionali e multidimensionali, un piccolo numero di iperpiani paralleli può fornire risultati della simulazione non accettabili poiché il generatore non si comporterebbe in maniera simile ad un generatore perfettamente casuale.

Occorre sottolineare che il numero effettivo di iperpiani paralleli è, alcune volte, di gran lunga inferiore alla maggiorazione di Marsaglia $(n!\, m)^{1/n}$. Un tipico esempio è fornito dall'algoritmo RANDU dell'IBM basato sul generatore confluente moltiplicativo con $m = 2^{31}$ e $a = 65539$; se infatti si scelgono sottosequenze consecutive di $n = 3$ numeri si può dimostrare che tutti i punti cadono in soltanto 15 piani paralleli di un cubo unitario. Tale algoritmo (alcune volte

ancora oggi utilizzzato) è stato definito da Knuth "really horrible" per le sue insufficienti proprietà statistiche.

7.3 Altri tipi di generatori congruenti

Esistono altri tipi di generatori, ossia *congruenti moltiplicativi misti*, *congruenti additivi*, *di Fibonacci*,...

I *generatori congruenti moltiplicativi misti,* introdotti da Lehmer nel 1951, producono sequenze $\{x_n,\ n = 0, 1, 2, \ldots\}$ come segue:

(i) fissare un intero positivo m detto *modulo* del generatore

(ii) scegliere degli interi positivi a, c e x_0 minori del modulo m; x_0 è detto *valore iniziale* o *seme*, la costante a ($a \neq 0$) è detta *moltiplicatore* e la costante c è detta *incremento*.

(iii) generare x_n dalla relazione di congruenza lineare

$$x_{n+1} \equiv a\, x_n + c \pmod{m} \tag{7.8}$$

che si legge x_{n+1} *è congruente ad* $a\, x_n + c$ *modulo* m.

Se $c = 0$ si ottiene il *generatore congruente moltiplicativo* descritto nel paragrafo precedente, mentre se $c > 0$ si ottiene il *generatore congruente moltiplicativo misto*.

La procedura inizia con un valore iniziale x_0; se $c > 0$ il seme x_0 può anche essere scelto nullo. Per determinare gli elementi della sequenza $\{x_n,\ n = 1, 2, \ldots\}$ occorre assegnare a x_{n+1} il resto r (con $0 \leq r \leq m-1$) della divisione di $a\, x_n + c$ per il modulo m.

La relazione di ricorrenza (7.8) è analoga all'equazione alle differenze del primo ordine $x_{n+1} = a\, x_n + c$ che ammette come soluzione

$$x_n = \begin{cases} x_0\, a^n + c\, \dfrac{a^n - 1}{a - 1}, & a \neq 1 \\ x_0 + a\, c, & a = 1. \end{cases}$$

Se $a \neq 1$ la (7.8) può essere scritta come

$$x_n \equiv x_0\, a^n + \frac{a^n - 1}{a - 1}\, c \pmod{m},$$

mentre se $a = 1$ si ha:

$$x_n \equiv x_0 + a\, c \pmod{m}.$$

La scelta dei parametri a, c, x_0 e m è importante per determinare la bontà del generatore considerato. Anche per il generatore congruente moltiplicativo misto si riesce a dimostrare un risultato analogo a quello espresso nella Proposizione 7.1 per il generatore congruente moltiplicativo.

Proposizione 7.3 *Sia $m = 2^b$ con $b \geq 2$. Se si scelgono i parametri a e c del generatore congruente moltiplicativo misto (7.8) minori del modulo e tali che*

(i) c intero positivo dispari,

(ii) $a = 8n + 1$ oppure $a = 8n + 5$ dove n è un qualsiasi intero non negativo (o equivalentemente $a = 4n + 1$),

si ottiene il periodo massimo 2^b.

Esempio 7.12 Sia $m = 2^4$, $x_0 = 7$ e $c = 3$. La relazione (7.8) diventa

$$x_n \equiv a\, x_{n-1} + 3 \pmod{2^4}.$$

a	1	5	9	13
x_0	7	7	7	7
x_1	10	6	2	14
x_2	13	1	5	9
x_3	0	8	0	8
x_4	3	11	3	11
x_5	6	10	14	2
x_6	9	5	1	13
x_7	12	12	12	12
x_8	15	15	15	15
x_9	2	14	10	6
x_{10}	5	9	13	1
x_{11}	8	0	8	0
x_{12}	11	3	11	3
x_{13}	14	2	6	10
x_{14}	1	13	9	5
x_{15}	4	4	4	4
x_{16}	7	7	7	7
p	16	16	16	16

Tabella 7.6: Sequenze prodotte con il generatore congruente moltiplicativo misto per $m = 2^4$, $x_0 = 7$, $c = 3$ e $a = 1, 5, 9, 13$.

Dalla Tabella 7.6 si nota che se $a = 1, 5, 9, 13$ il periodo p è $m = 2^4 = 16$. $\diamondsuit$

I passi dell'algoritmo per generare una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n < 1$) utilizzando il metodo congruenziale moltiplicativo misto con $m = 2^b$ sono quindi i seguenti:

Algoritmo
STEP 1: fornire in input x_0, a, c e b tali da soddisfare le ipotesi del Teorema 7.3

STEP 2: per ogni $n = 1, 2, \ldots, 2^b - 1$ calcolare

$$x_n \equiv a\, x_{n-1} + c \pmod{m},$$

STEP 3: per ogni $n = 0, 1, \ldots, 2^b - 1$ calcolare

$$u_n = x_n \cdot 2^{-b}.$$

Dalle Proposizioni 7.1 e 7.3 emerge che se $m = 2^b$ il generatore congruente moltiplicativo ha un periodo massimo 2^{b-2} che è pari ad $1/4$ del periodo massimo 2^b del generatore congruente moltiplicativo misto. Occorre però sottolineare che il generatore congruente moltiplicativo è solitamente preferito a quello misto per la maggiore casualità con cui spesso si presentano i numeri della sequenza.

Alcune scelte dei parametri m, a e c del generatore congruente moltiplicativo misto sono $m = 2^{31}$, $a = 314159269$, $c = 453806245$ oppure $m = 2^{35}$, $a = 5^{15} = 30517578125$, $c = 1$.

I generatori congruenti moltiplicativi e quelli congruenti moltiplicativi misti precedentemente discussi sono casi particolari del *generatore congruente additivo*

$$x_{n+1} \equiv a_0\, x_n + a_1\, x_{n-1} + \ldots + a_r\, x_{n-r} + c \pmod{m} \qquad (n = r, r+1, \ldots), \tag{7.9}$$

che richiede la conoscenza di $r + 1$ valori iniziali $x_0, x_1, \ldots, x_r$, di $r + 1$ costanti moltiplicative $a_0, a_1, \ldots, a_r$ e di una costante additiva c.

Un particolare generatore congruente additivo è il *generatore di Fibonacci*:

$$x_{n+1} \equiv x_n + x_{n-1} \pmod{m} \qquad (n = 1, 2, \ldots), \tag{7.10}$$

che richiede la conoscenza di soltanto due valori iniziali. Tale generatore, *di interesse storico*, è detto di Fibonacci per la sua similarità con la successione dei numeri di Fibonacci.

x_0	1	x_7	21	x_{14}	610
x_1	1	x_8	34	x_{15}	987
x_2	2	x_9	55	x_{16}	597
x_3	3	x_{10}	89	x_{17}	584
x_4	5	x_{11}	144	x_{18}	181
x_5	8	x_{12}	233	x_{19}	765
x_6	13	x_{13}	377	x_{20}	946

Tabella 7.7: Sequenza prodotta con il generatore di Fibonacci per $m = 1000$ e $x_0 = x_1 = 1$.

Se, ad esempio, si sceglie $m = 1000$ e $x_0 = x_1 = 1$, i primi valori della sequenza generata con il metodo di Fibonacci sono riportati in Tabella 7.7.

Spesso le sequenze prodotte con il generatore di Fibonacci non sono dotate di buone qualità statistiche poiché presentano una forte correlazione tra i numeri della sequenza.

Un altro tipo di generatore congruente additivo è il seguente

$$x_{n+1} \equiv x_n + x_{n-r} \pmod{m} \qquad (n = r, r+1, \ldots), \tag{7.11}$$

che richiede la conoscenza di $r+1$ valori iniziali; tale generatore fornisce sequenze tanto più soddisfacenti dal punto di vista statistico quanto più grande si sceglie il parametro r.

In conclusione, le caratteristiche che deve avere un buon generatore di sequenze pseudocasuali sono:

- *ripetibilità*, che garantisce la possibilità di ripetere più volte lo stesso esperimento di simulazione;
- *soddisfacimento di test statistici*, in maniera da verificare che il generatore sia abbastanza simile ad un generatore di numeri perfettamente casuali
- *semplicità e rapidità di utilizzazzione*, in maniera da risultare efficiente computazionalmente
- *periodo lungo*, in maniera tale da poter disporre di sequenze lunghe di numeri pseudocasuali;
- *portabilità*, in maniera tale da rendere l'implementazione del generatore indipendente dalla piattaforma.

Capitolo 8

Generatori non uniformi

8.1 Introduzione

Nel Capitolo 7 abbiamo considerato alcuni metodi per costruire generatori uniformi nell'intervallo $(0, 1)$. Vogliamo ora introdurre delle tecniche che permettano di ottenere variabili aleatorie con una funzione di distribuzione diversa da quella uniforme a partire da variabili aleatorie con distribuzione uniforme nell'intervallo $(0, 1)$.

In un esperimento di simulazione occorre spesso generare più sequenze indipendenti per rappresentare variabili aleatorie differenti. Ad esempio, nel caso della simulazione di un sistema di servizio occorre generare due sequenze (esponenziali, di Erlang, iperesponenziali, ...) per rappresentare i tempi di interarrivo e di servizio e le due sequenze generate debbono essere indipendenti. Per ottenere sequenze uniformi in $(0, 1)$ indipendenti esistono le seguenti possibilità:

1) utilizzare *costanti moltiplicative* (moltiplicatori) *differenti* nel metodo congruenziale moltiplicativo;

2) utilizzare *semi iniziali* (valori iniziali) *differenti* nel metodo congruenziale moltiplicativo;

3) utilizzare una sola sequenza (generata con un unico seme iniziale) per ottenere istanze di una variabile aleatoria uniforme in $(0, 1)$ e successivamente *partizionare la sequenza generata in più distinte sottosequenze* da utilizzare per generare le differenti variabili aleatorie indipendenti.

Nel seguito considereremo separatamente la simulazione di variabili aleatorie continue e di variabili aleatorie discrete fornendo sia metodi generali sia metodi specifici per la simulazione di particolari variabili aleatorie.

8.2 Variabili aleatorie continue

I metodi maggiormente utilizzati nella pratica per simulare variabili aleatorie continue sono due: metodo di *inversione della funzione di distribuzione* e *metodo di reiezione*.
Nel seguito denoteremo con U una variabile aleatoria uniformemente distribuita nell'intervallo $(0,1)$ e con

$$F_U(u) = P(U < u) = \begin{cases} 0, & u \leq 0 \\ u, & 0 < u \leq 1 \\ 1, & u > 1, \end{cases} \tag{8.1}$$

la sua funzione di distribuzione.

8.2.1 Metodo di inversione della funzione di distribuzione

Desideriamo simulare una variabile aleatoria continua X con funzione di distribuzione $F_X(x)$. Il metodo di inversione della funzione di distribuzione si basa sulla seguente proposizione:

Proposizione 8.1 *Sia U una variabile aleatoria uniformemente distribuita nell'intervallo $(0,1)$. Definiamo una variabile aleatoria X continua tramite la trasformazione*

$$X = F^{-1}(U), \tag{8.2}$$

essendo $F(x)$ una funzione di distribuzione invertibile. La funzione di distribuzione della variabile aleatoria X è data da

$$F_X(x) = \begin{cases} 0, & x \leq F^{-1}(0) \\ F(x), & F^{-1}(0) < x \leq F^{-1}(1) \\ 1, & x > F^{-1}(1). \end{cases} \tag{8.3}$$

Dimostrazione Poiché $F(x)$ è una funzione di distribuzione, essa è non decrescente. Ricordando quindi la (8.2) e facendo uso di (8.1) si ha

$$\begin{aligned} F_X(x) = P(X < x) &= P\big[F^{-1}(U) < x\big] = P\big[U < F(x)\big] \\ &= \begin{cases} 0, & F(x) \leq 0 \\ F(x), & 0 < F(x) \leq 1 \\ 1, & F(x) > 1, \end{cases} \end{aligned}$$

da cui segue direttamente la (8.3). □

La Proposizione 8.1 mostra che è possibile simulare una variabile aleatoria X continua caratterizzata da funzione di distribuzione $F(x)$ invertibile simulando una variabile aleatoria U uniformemente distribuita in $(0,1)$ e ponendo $X = F^{-1}(U)$. L'algoritmo per simulare la variabile aleatoria X con il metodo di inversione della funzione di distribuzione è quindi il seguente:

Algoritmo

STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
STEP 2: Porre

$$X = F^{-1}(U).$$

Per generare una sequenza di valori reali di una data variabile aleatoria X con funzione di distribuzione $F(x)$ si procede quindi nel seguente modo. Si genera un reale u_i uniformemente distribuito nell'intervallo $(0, 1)$, si pone $u_i = F(x_i)$ e si ricava tramite l'applicazione della funzione inversa F^{-1} ad entrambi i membri di tale equazione il valore $x_i = F^{-1}(u_i)$ corrispondente al numero u_i generato. Iterando questo metodo ai vari elementi della sequenza $u_0, u_1, \ldots$ si ottiene la sequenza $x_0, x_1, \ldots$ di numeri reali che costituiscono osservazioni di una variabile aleatoria X con funzione di distribuzione $F(x)$.

Problema 8.1 Simulazione di una variabile aleatoria distribuita uniformemente in (a, b)
Consideriamo una variabile aleatoria X uniformemente distribuita nell'intervallo (a, b) e sia

$$F_X(x) = P(X < x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a} & a < x \le b \\ 1, & x > b \end{cases}$$

la sua funzione di distribuzione. Applicando il metodo di inversione della funzione di distribuzione possiamo scrivere

$$U = F(X) = \frac{X-a}{b-a},$$

da cui segue che

$$X = a + (b-a)\,U,$$

dove U è una variabile aleatoria uniformemente distribuita nell'intervallo $(0, 1)$.

Il *metodo di inversione della funzione di distribuzione* allora simula la variabile aleatoria X uniformemente distribuita in (a, b) nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
STEP 2: Porre

$$X = a + (b-a)\,U.$$

A partire dalla sequenza $u_0, u_1, \ldots$ di numeri uniformemente distribuiti in $(0, 1)$ possiamo quindi ottenere la sequenza $x_0, x_1, \ldots$ di numeri uniformemente distribuiti in (a, b) tramite la relazione

$$x_i = a + (b-a)\,u_i \qquad (i = 0, 1, \ldots).$$

□

Problema 8.2 Simulazione di una variabile aleatoria esponenziale
Consideriamo una variabile aleatoria X esponenzialmente distribuita con valore medio $1/\lambda$ e sia

$$F_X(x) = P(X < x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0 \end{cases}$$

la sua funzione di distribuzione. Applicando il metodo di inversione della funzione di distribuzione si ha:

$$U = F(X) = 1 - e^{-\lambda X},$$

ossia

$$e^{-\lambda X} = 1 - U,$$

da cui segue

$$X = -\frac{1}{\lambda} \ln(1 - U),$$

dove U è una variabile aleatoria uniformemente distribuita nell'intervallo $(0, 1)$.

Il *metodo di inversione della funzione di distribuzione* quindi simula la variabile aleatoria X esponenzialmente distribuita con valore medio $1/\lambda$ nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
STEP 2: Porre

$$X = -\frac{1}{\lambda} \ln(1 - U).$$

A partire dalla sequenza $u_0, u_1, \ldots$ di numeri uniformemente distribuiti in $(0, 1)$, possiamo quindi ottenere la sequenza $x_0, x_1, \ldots$ di numeri esponenzialmente distribuiti con valore medio $1/\lambda$ tramite la relazione

$$x_i = -\frac{1}{\lambda} \ln(1 - u_i) \qquad (i = 0, 1, \ldots).$$

Se U è uniformemente distribuita in $(0, 1)$, anche la variabile aleatoria $1-U$ è uniformemente distribuita in $(0, 1)$. Quindi X può essere anche espressa tramite la relazione $X = -(1/\lambda) \ln U$. □

Esempio 8.1 Supponiamo di voler simulare il sistema $M/M/1$ utilizzando l'algoritmo descritto nel Capitolo 7. Supponiamo che i tempi di interarrivo sono esponenzialmente distribuiti con valore medio $1/\lambda$ e i tempi di servizio sono esponenzialmente distribuiti con valore medio $1/\mu$. In questo caso occorre generare due sequenze indipendenti di numeri uniformi in $(0, 1)$, ossia $u_0, u_1, \ldots$ e $v_0, v_1, \ldots$. I tempi di interarrivo t_i possono essere generati con $t_i = -(1/\lambda)\ln(1 - u_i)$ e i tempi di servizio possono essere generati con $s_i = -(1/\mu)\ln(1 - v_i)$ ♢

Esempio 8.2 Supponiamo di voler simulare il sistema $M/U/1$ utilizzando l'algoritmo descritto nel Capitolo 7. Supponiamo che i tempi di interarrivo sono esponenzialmente distribuiti con valore medio $1/\lambda$ e i tempi di servizio sono uniformemente distribuiti nell'intervallo $(0, 2/\mu)$. In questo caso occorre generare due sequenze indipendenti di numeri uniformi in $(0,1)$, ossia $u_0, u_1, \ldots$ e $v_0, v_1, \ldots$. I tempi di interarrivo t_i possono essere generati con $t_i = -(1/\lambda)\ln(1-u_i)$ e i tempi di servizio possono essere generati con $s_i = (2/\mu)\, v_i$. ♢

Problema 8.3 Simulazione di una variabile aleatoria distribuita secondo Rayleigh
Consideriamo una variabile aleatoria X distribuita secondo Rayleigh e siano

$$f_X(x) = \begin{cases} x\, e^{-x^2/2}, & x > 0 \\ 0, & \text{altrimenti,} \end{cases} \qquad F_X(x) = P(X < x) = \begin{cases} 0, & x \le 0 \\ 1 - e^{-x^2/2}, & x > 0 \end{cases}$$

la sua densità di probabilità e la sua funzione di distribuzione, rispettivamente. Applicando il metodo di inversione della funzione di distribuzione possiamo scrivere

$$U = F(X) = 1 - e^{-X^2/2},$$

ossia

$$e^{-X^2/2} = 1 - U,$$

da cui segue

$$X = \sqrt{-2\,\ln(1-U)},$$

dove U è una variabile aleatoria uniformemente distribuita nell'intervallo $(0,1)$.

Il *metodo di inversione della funzione di distribuzione* quindi simula la variabile aleatoria X distribuita secondo Rayleigh nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0,1)$;
STEP 2: Porre

$$X = \sqrt{-2\,\ln(1-U)},$$

A partire dalla sequenza $u_0, u_1, \ldots$ di numeri uniformemente distribuiti in $(0,1)$, possiamo quindi ottenere la sequenza $x_0, x_1, \ldots$ di numeri distribuiti secondo Rayleigh tramite la relazione

$$x_i = \sqrt{-2\,\ln(1-u_i)} \qquad (i = 0, 1, \ldots).$$

□

8.2.2 Metodo di reiezione

Il metodo di inversione della funzione di distribuzione trova un limite preciso nel suo campo di applicazione in presenza di variabili aleatorie X per le quali risulti molto difficile, o addirittura impossibile, esprimere analiticamente la funzione

di distribuzione inversa. Un tipico esempio di funzione di distribuzione difficile da invertire è quella della variabile aleatoria normale standard la cui funzione di distribuzione è

$$F(x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{x} e^{-z^2}\, dz \qquad (x \in \mathbb{R}).$$

In casi come questo si utilizza spesso il metodo di reiezione che ora descriviamo.

Sia X la variabile aleatoria che si desidera simulare caratterizzata da densità di probabilità $f(x)$. Supponiamo di disporre di un metodo che permetta di simulare un'altra variabile aleatoria Y caratterizzata da densità di probabilità $g(y)$. Il *metodo di reiezione* allora simula la variabile aleatoria X nel seguente modo:

Algoritmo
STEP 1: Generare indipendentemente una variabile aleatoria Y avente densità $g(y)$ e una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
STEP 2: Sia c una costante reale positiva scelta in modo tale che

$$\frac{f(y)}{c\,g(y)} \leq 1 \tag{8.4}$$

per ogni y tale che $f(y) > 0$ e $g(y) > 0$. Se risulta

$$U < \frac{f(Y)}{c\,g(Y)} \tag{8.5}$$

porre $X = Y$, altrimenti ritornare al passo *1.*

Proposizione 8.2 *La variabile aleatoria continua X generata con il metodo di reiezione è caratterizzata da densità di probabilità $f(x)$.*

Dimostrazione Vogliamo dimostrare che la variabile aleatoria X che si desidera simulare con il metodo di reiezione è caratterizzata da densità di probabilità $f(x)$. Dal passo 2 dell'algoritmo segue che

$$\begin{aligned} P(X < x) = P\Big\{ Y < x \,\Big|\, U < \frac{f(Y)}{c\,g(Y)} \Big\} &= \frac{P\Big\{ Y < x,\, U < \dfrac{f(Y)}{c\,g(Y)} \Big\}}{P\Big\{ U < \dfrac{f(Y)}{c\,g(Y)} \Big\}} \\ &= \frac{\displaystyle\int_{-\infty}^{\infty} P\Big\{ Y < x,\, U < \frac{f(Y)}{c\,g(Y)} \,\Big|\, Y = y \Big\}\, g(y)\, dy}{P\Big\{ U < \dfrac{f(Y)}{c\,g(Y)} \Big\}} \\ &= \frac{\displaystyle\int_{-\infty}^{x} P\Big\{ U < \frac{f(y)}{c\,g(y)} \Big\}\, g(y)\, dy}{P\Big\{ U < \dfrac{f(Y)}{c\,g(Y)} \Big\}}. \end{aligned} \tag{8.6}$$

Poiché U è uniformemente distribuita in $(0,1)$, facendo uso di (8.1) in (8.6) si ricava:

$$P(X<x)=\frac{\displaystyle\int_{-\infty}^{x}\frac{f(y)}{c\,g(y)}\,g(y)\,dy}{P\Big\{U<\dfrac{f(Y)}{c\,g(Y)}\Big\}}=\frac{\displaystyle\int_{-\infty}^{x}f(y)\,dy}{c\,P\Big\{U<\dfrac{f(Y)}{c\,g(Y)}\Big\}}. \tag{8.7}$$

Procedendo al limite quando x tende all'infinito nella (8.7), si ottiene:

$$1=\frac{\displaystyle\int_{-\infty}^{\infty}f(y)\,dy}{c\,P\Big\{U<\dfrac{f(Y)}{c\,g(Y)}\Big\}}=\frac{1}{c\,P\Big\{U<\dfrac{f(Y)}{c\,g(Y)}\Big\}},$$

ossia

$$P\Big\{U<\frac{f(Y)}{c\,g(Y)}\Big\}=\frac{1}{c}\,. \tag{8.8}$$

Sostituendo (8.8) in (8.7) si ha:

$$P(X<x)=\frac{\displaystyle\int_{-\infty}^{x}f(y)\,dy}{c\,P\Big\{U<\dfrac{f(Y)}{c\,g(Y)}\Big\}}=\int_{-\infty}^{x}f(y)\,dy,$$

ossia $P(X<x)$ coincide con la funzione di distribuzione di una variabile aleatoria continua di densità $f(x)$. La variabile aleatoria X generata con il metodo di reiezione è quindi caratterizzata da densità di probabilità $f(x)$. □

Problema 8.4 Simulazione del valore assoluto di una variabile aleatoria normale standard
Sia Z una variabile aleatoria di densità normale standard

$$f_Z(z)=\frac{1}{\sqrt{2\,\pi}}\,\exp\Big\{-\frac{z^2}{2}\Big\}\qquad(z\in\mathbb{R}). \tag{8.9}$$

La variabile aleatoria $X=|Z|$ è quindi caratterizzata da densità di probabilità:

$$f(x)=\begin{cases}\dfrac{2}{\sqrt{2\,\pi}}\,\exp\Big\{-\dfrac{x^2}{2}\Big\}, & x>0\\ 0, & \text{altrimenti.}\end{cases}$$

Vogliamo simulare la variabile aleatoria X utilizzando il metodo di reiezione. Nell'algoritmo di reiezione scegliamo una variabile aleatoria Y esponenzialmente distribuita di valore medio unitario, ossia di densità di probabilità:

$$g(y)=\begin{cases}e^{-y}, & y>0.\\ 0, & \text{altrimenti.}\end{cases}$$

Per applicare il metodo di reiezione occorre determinare una costante c che soddisfi la condizione (8.4), ossia tale che per ogni $y > 0$ risulti $f(y)/g(y) \leq c$. Se $y > 0$ si ha

$$\frac{f(y)}{g(y)} = \frac{2}{\sqrt{2\,\pi}}\,\exp\Big\{-\frac{y^2}{2}+y\Big\} = \sqrt{\frac{2e}{\pi}}\,\exp\Big\{-\frac{(y-1)^2}{2}\Big\}.$$

Poiché

$$\frac{f(y)}{g(y)} \leq \sqrt{\frac{2e}{\pi}},$$

affinché la (8.4) sia soddisfatta per ogni $y > 0$ basterà scegliere nel metodo di reiezione

$$c = \sqrt{\frac{2e}{\pi}}\,.$$

Pertanto, se $y > 0$ si ha

$$\frac{f(y)}{c\,g(y)} = \exp\Big\{-\frac{(y-1)^2}{2}\Big\}\,.$$

L'algoritmo per simulare la variabile aleatoria $X = |Z|$, con Z avente distribuzione normale standard, mediante il metodo di reiezione è quindi il seguente:

Algoritmo
STEP 1: Generare le variabili aleatorie indipendenti Y esponenzialmente distribuita di valore medio unitario e U uniformemente distribuita in $(0, 1)$;
STEP 2: Se risulta

$$U < \exp\Big\{-\frac{(Y-1)^2}{2}\Big\}$$

porre $X = Y$, altrimenti ritornare al passo *1.* □

Problema 8.5 Simulazione di una variabile aleatoria normale standard
La simulazione della variabile aleatoria $X = |Z|$ discussa nel Problema 8.4 permette di simulare una variabile aleatoria Z di densità normale standard (8.9) ponendo $Z = X$ oppure $Z = -X$ con uguale probabilità. Ad esempio, si può porre

$$Z = \begin{cases} -X, & 0 \leq V < 1/2 \\ X, & 1/2 \leq V < 1, \end{cases}$$

essendo U una variabile aleatoria uniformemente distribuita in $(0, 1)$. Infatti risulta

$$\begin{aligned} P(Z < z) &= P\Big(Z < z, 0 \leq V < \frac{1}{2}\Big) + P\Big(Z < z, \frac{1}{2} \leq V < 1\Big) \\ &= \frac{1}{2}\,P\Big(Z < z \mid 0 \leq V < \frac{1}{2}\Big) + \frac{1}{2}\,P\Big(Z < z \mid \frac{1}{2} \leq V < 1\Big) \\ &= \frac{1}{2}\,P(-X < z) + \frac{1}{2}\,P(X < z) = \frac{1}{2}\,P(X > -z) + \frac{1}{2}\,P(X < z), \end{aligned}$$

da cui derivando rispetto a z si ottiene la densità di probabilità (8.9).

L'algoritmo per simulare la variabile aleatoria Z con distribuzione normale standard è quindi il seguente:

Algoritmo
STEP 1: Generare le variabili aleatorie indipendenti Y esponenzialmente distribuita di valore medio unitario e U uniformemente distribuita in $(0,1)$;
STEP 2: Se risulta

$$U < \exp\Big\{-\frac{(Y-1)^2}{2}\Big\}$$

porre $X = Y$, altrimenti ritornare al passo *1.*
STEP 3: Generare una variabile aleatoria V uniformemente distribuita in $(0,1)$ e porre

$$Z = \begin{cases} X, & 0 \le V < 1/2 \\ -X, & 1/2 \le V < 1. \end{cases}$$

□

Problema 8.6 Simulazione di una variabile aleatoria normale
Sia T una variabile aleatoria distribuita normalmente con valore medio μ e varianza σ^2, caratterizzata da densità di probabilità

$$f_T(t) = \frac{1}{\sigma\sqrt{2\,\pi}}\,\exp\Big\{-\frac{(t-\mu)^2}{2\,\sigma^2}\Big\} \qquad (t \in \mathbb{R}).$$

Se si considera il cambiamento di variabile

$$Z = \frac{T-\mu}{\sigma}\,,$$

si ottiene una variabile aleatoria Z con distribuzione normale standard (di valore medio nullo e varianza unitaria).

Ricordando i Problemi 8.4 e 8.5, l'algoritmo per simulare una variabile aleatoria T distribuita normalmente con valore medio μ e varianza σ^2, è il seguente:

Algoritmo
STEP 1: Generare le variabili aleatorie indipendenti Y esponenzialmente distribuita di valore medio unitario e U uniformemente distribuita in $(0,1)$;
STEP 2: Se risulta

$$U < \exp\Big\{-\frac{(Y-1)^2}{2}\Big\}$$

porre $X = Y$, altrimenti ritornare al passo *1.*
STEP 3: Generare una variabile aleatoria V uniformemente distribuita in $(0,1)$ e porre

$$Z = \begin{cases} X, & 0 \le V < 1/2 \\ -X, & 1/2 \le V < 1. \end{cases}$$

STEP 4: Porre

$$T = \mu + \sigma\,Z.$$

□

8.3 Particolari variabili aleatorie continue

Nel Paragrafo 8.2 abbiamo illustrato due metodi per la simulazione di variabili aleatorie continue a partire da variabili aleatorie con distribuzione uniforme in $(0, 1)$. Per alcuni tipi di variabili aleatorie continue esistono idonei algoritmi che risultano *più efficienti computazionalmente* di quelli ottenibili con il metodo di inversione della funzione di distribuzione o con il metodo di reiezione.

Vogliamo ora analizzare metodi specifici per simulare alcuni tipi di variabili aleatorie continue, ossia *normale*, *di Erlang* e *ipergeometrica*.

Problema 8.7 Simulazione di una variabile aleatoria normale standard
Esistono vari metodi per generare una variabile aleatoria con distribuzione normale standard oltre a quello discusso nel Problema 8.5. Uno di questi si basa sul teorema centrale di convergenza. Tale teorema afferma che se $X_1, X_2, \ldots$ è una successione di variabili aleatorie indipendenti e identicamente distribuite, con valore medio $E(X_i) = \mu$ finito e varianza $\mathrm{Var}(X_i) = \sigma^2$ finita, allora

$$\lim_{N \to +\infty} P\left(\frac{X_1 + X_2 + \ldots + X_N - N\,\mu}{\sigma\,\sqrt{N}} < x \right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\,\pi}} \exp\left\{ -\frac{z^2}{2} \right\} dz. \tag{8.10}$$

Se si considera una sequenza di variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$, ognuna caratterizzata da valore medio $E(U_i) = 1/2$ e varianza $\mathrm{Var}(U_i) = 1/12$, dal teorema centrale di convergenza segue che per N abbastanza grande la funzione di distribuzione della variabile aleatoria

$$Z = \frac{U_1 + U_2 + \ldots + U_N - N/2}{\sqrt{N/12}} \tag{8.11}$$

è approssimativamente quella di una variabile aleatoria normale standard. Un metodo per simulare una variabile aleatoria con distribuzione normale standard è quindi il seguente:

Algoritmo
STEP 1: Generare N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: Valutare Z tramite la relazione

$$Z = \frac{U_1 + U_2 + \ldots + U_N - N/2}{\sqrt{N/12}}.$$

Solitamente si sceglie $N \geq 12$ poiché sperimentalmente si è visto che già 12 variabili aleatorie con distribuzione uniforme sono sufficienti per simulare una variabile aleatoria con distribuzione normale standard. In particolare, scegliendo $N = 12$ l'algoritmo di simulazione della variabile aleatoria Z con distribuzione normale standard diventa il seguente.

Algoritmo

STEP 1: Generare 12 variabili aleatorie $U_1, U_2, \ldots, U_{12}$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: Valutare Z tramite la relazione

$$Z = U_1 + U_2 + \ldots + U_{12} - 6.$$

Occorre osservare che la variabile aleatoria $U_1 + U_2 + \ldots + U_{12} - 6$ assume valori nell'intervallo $(-6, 6)$. Questo intervallo è sufficiente per la generazione di una variabile aleatoria con distribuzione normale standard; infatti, ricordando che $P(-z < Z < z) = 2\,P(Z < z) - 1$, segue che già scegliendo $z = 3$ si ottiene $P(-3 < Z < 3) = 2\,P(Z < 3) - 1 \simeq 0.9974$, il che mostra che la probabilità che la variabile aleatoria Z assuma valori nell'intervallo $(-3, 3)$ è già prossima all'unità.

Un altro metodo per simulare una variabile aleatoria con distribuzione normale standard è il *metodo di Box–Muller* che permette di ottenere una coppia di valori della variabile aleatoria normale standard sfruttando ogni volta la simulazione di due variabili aleatorie uniformi in $(0, 1)$. Il metodo di Box-Muller per simulare una variabile aleatoria con distribuzione normale standard può essere sintetizzato nel seguente algoritmo:

Algoritmo
STEP 1: Generare due variabili aleatorie U_1 e U_2 indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: Porre

$$X = \sqrt{-2\,\ln(1 - U_1)}\,\cos(2\,\pi\,U_2)$$
$$Y = \sqrt{-2\,\ln(1 - U_1)}\,\sin(2\,\pi\,U_2).$$

Come mostrato nel Problema 8.3 la variabile aleatoria $\sqrt{-2\,\ln(1 - U_1)}$ utilizzata nell'algoritmo di Box-Muller è caratterizzata da una funzione di distribuzione di Rayleigh. □

Problema 8.8 Simulazione di una variabile aleatoria di Erlang
Sia Y una variabile aleatoria con densità di Erlang di ordine k, ossia

$$f_Y(x) = \begin{cases} \dfrac{\lambda^k}{(k-1)!}\,e^{-\lambda x}\,x^{k-1}, & x > 0 \\ 0, & x \le 0, \end{cases} \tag{8.12}$$

dove $\lambda > 0$. La funzione di distribuzione della variabile aleatoria Y

$$F_Y(x) = P(Y_k < x) = \begin{cases} 0, & x \le 0 \\ 1 - e^{-\lambda\,t}\,\displaystyle\sum_{i=0}^{k-1} \frac{(\lambda\,x)^i}{i!}, & x > 0, \end{cases} \tag{8.13}$$

è difficile da invertire; il metodo di inversione della funzione di distribuzione quindi non è utilizzabile. Inoltre, il metodo di reiezione, anche se applicabile

(scegliendo, ad esempio, la variabile aleatoria Y esponenzialmente distribuita con valore medio coincidente con quello della variabile aleatoria di Erlang Y, ossia k/λ) è di difficile utilizzazione pratica.

Come abbiamo precedentemente visto se $X_1, X_2, \ldots, X_k$ sono variabili aleatorie indipendenti e esponenzialmente distribuite con valore medio $1/\lambda$, allora la variabile aleatoria $Y = X_1 + X_2 + \ldots + X_k$ è caratterizzata da una densità di Erlang di ordine k. I passi fondamentali dell'algoritmo per simulare Y sono quindi i seguenti:

Algoritmo
STEP 1: Generare k variabili aleatorie indipendenti $X_1, X_2, \ldots, X_k$ esponenzialmente distribuite di valore medio $1/\lambda$;
STEP 2: Calcolare

$$Y = \sum_{i=1}^{k} X_i.$$

Ricordando il Problema 8.2, nel precedente algoritmo si può porre $X_i = -\ln(1-U_i)/\lambda$ $(i = 1, 2, \ldots, k)$, con $U_1, U_2, \ldots, U_k$ uniformemente distribuite in $(0, 1)$.

L'algoritmo per simulare una variabile aleatoria di Erlang di ordine k può quindi anche essere così formulato:

Algoritmo
STEP 1: Generare k variabili aleatorie indipendenti $U_1, U_2, \ldots, U_k$ uniformemente distribuite in $(0, 1)$;
STEP 2: Porre

$$X_i = -\frac{1}{\lambda}\ln(1-U_i) \qquad (i = 1, 2, \ldots, k).$$

STEP 3: Calcolare

$$Y = \sum_{i=1}^{k} X_i = -\frac{1}{\lambda}\sum_{i=1}^{k}\ln(1-U_i) = -\frac{1}{\lambda}\ln\Big[\prod_{i=1}^{k}(1-U_i)\Big].$$

□

Esempio 8.3 Supponiamo di voler simulare il sistema $M/E_2/1$ utilizzando l'algoritmo descritto nel Capitolo 7. I tempi di interarrivo sono esponenzialmente distribuiti con valore medio $1/\lambda$ e i tempi di servizio sono distribuiti secondo Erlang di ordine 2, ossia il servizio è organizzato in due fasi successive indipendenti, ognuna distribuita esponenzialmente con valore medio $1/(2\mu)$. In tal caso occorre generare tre sequenze indipendenti di numeri uniformi in $(0, 1)$, ossia $u_0, u_1, \ldots$, $v_0, v_1, \ldots$ e $z_1, z_2, \ldots$. I tempi di interarrivo t_i possono essere generati con $t_i = -(1/\lambda)\ln(1-z_i)$ e i tempi di servizio con $s_i = -[1/(2\mu)]\ln(1-u_i) - [1/(2\mu)]\ln(1-v_i)$. ♢

8.3.1 Metodo composto

Assumiamo che la densità di probabilità di una variabile aleatoria continua X si possa porre nella forma

$$f_X(x) = \sum_{j=1}^{k} p_j \, g_j(x) \tag{8.14}$$

dove $p_1, p_2, \ldots, p_k$ soddisfano le condizioni

$$p_j \geq 0 \qquad (j = 1, 2, \ldots, k), \qquad \sum_{j=1}^{k} p_j = 1 \tag{8.15}$$

e dove $g_1(x), g_2(x), \ldots, g_k(x)$ sono le densità di probabilità delle variabili aleatorie continue $Z_1, Z_2, \ldots, Z_k$. Supponiamo di disporre di un metodo che permetta di simulare le k variabili aleatorie $Z_1, Z_2, \ldots, Z_k$. Il *metodo composto* simula la variabile aleatoria X nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$X = \begin{cases} Z_1, & 0 \leq U < p_1 \\ Z_2, & p_1 \leq U < p_1 + p_2 \\ \cdots\cdots\cdots & \\ Z_j, & \displaystyle\sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots & \\ Z_k, & \displaystyle\sum_{i=1}^{k-1} p_i \leq U < \sum_{i=1}^{k} p_i = 1 \end{cases}$$

Proposizione 8.3 *La variabile aleatoria continua X generata con il metodo composto ha densità di probabilità (8.14).*

Dimostrazione Osserviamo che

$$\begin{aligned} P(X < x) &= \sum_{j=1}^{k} P\Big(X < x, \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \Big) \\ &= \sum_{j=1}^{k} P\Big(\sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \Big) \, P\Big(X < x \,\Big|\, \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \Big). \end{aligned}$$

dove si è posto $\sum_{i=1}^{j-1} p_i = 0$ per $j = 1$. Poiché risulta

$$P\Big(X < x \,\Big|\, \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \Big) = P(Z_j < x) \qquad (j = 1, 2, \ldots, k),$$

segue che

$$P(X < x) = \sum_{j=1}^{k} p_j \, P(Z_j < x).$$

Derivando ambo i membri rispetto a x, segue la (8.14). La variabile aleatoria X è quindi caratterizzata da densità di probabilità (8.14). □

Problema 8.9 Simulazione di una variabile aleatoria iperesponenziale Sia X una variabile aleatoria caratterizzata da una densità iperesponenziale di ordine k, ossia

$$f_X(x) = \begin{cases} \displaystyle\sum_{j=1}^{k} p_j \, \lambda_j \, e^{-\lambda_j x}, & x > 0 \\ 0, & \text{altrimenti,} \end{cases} \tag{8.16}$$

dove $p_1, p_2, \ldots, p_k$ soddisfano le (8.15).

Osserviamo che la densità iperesponenziale si presenta come una *combinazione lineare di funzioni densità esponenziali*. Per simulare X si può quindi utilizzare il *metodo composto*. In questo caso la variabili aleatoria Z_j del metodo composto è caratterizzata da densità esponenziale di valore medio $1/\lambda_j$ $(j = 1, 2, \ldots, k)$. Ricordando la simulazione di una variabile aleatoria esponenziale fornita nel Problema 8.2, il metodo composto simula la variabile aleatoria X iperesponenziale nel seguente modo:

Algoritmo
STEP 1: Generare $k + 1$ variabili aleatorie $V_1, V_2, \ldots, V_k, U$ indipendenti e uniformemente distribuite in $(0, 1)$;
STEP 2: Porre

$$X = \begin{cases} -\ln(1 - V_1)/\lambda_1, & 0 \le U < p_1 \\ -\ln(1 - V_2)/\lambda_2, & p_1 \le U < p_1 + p_2 \\ \cdots\cdots\cdots \\ -\ln(1 - V_j)/\lambda_j, & \displaystyle\sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots \\ -\ln(1 - V_k)/\lambda_k, & \displaystyle\sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i. \end{cases}$$

□

Esempio 8.4 Supponiamo di voler simulare il sistema di servizio $U/H_2/1$ utilizzando l'algoritmo descritto nel Capitolo 7. I tempi di interarrivo sono distribuiti uniformementi in $(0, 2/\lambda)$ e i tempi di servizio sono distribuiti con densità iperesponenziale di ordine 2, ossia $b(t) = p_1\mu_1 e^{-\mu_1 t} + p_2\mu_2 e^{-\mu_2 t}$, con $p_1 + p_2 = 1$. Abbiamo un unico servitore che fornisce due differenti servizi distribuiti esponenzialmente con valori medi $1/\mu_1$ e $1/\mu_2$; l'utente sceglie il primo servizio con

probabilità p ed il secondo servizio con probabilità $1-p$. Occorre generare quattro sequenze indipendenti uniformemente distribuite in $(0,1)$, ossia $u_0, u_1, \ldots$, $v_0, v_1, \ldots$ e $h_1, h_2, \ldots$, $k_1, k_2, \ldots$ I tempi di interarrivo sono invece generati con $t_i = (2/\lambda)u_i$. I tempi di servizio sono generati come $s_i = -(1/\mu_1)\ln(1-h_i)$ se $0 \le v_i < p$ mentre sono simulati come $s_i = -(1/\mu_2)\ln(1-k_i)$ se $p \le v_i < 1$. $\diamondsuit$

8.4 Variabili aleatorie discrete

Sia X una variabile aleatoria discreta che assume valori in un insieme finito o al più numerabile $S = \{x_1, x_2, \ldots\}$ e sia

$$p_j = P(X = x_j) \qquad (j = 1, 2, \ldots) \tag{8.17}$$

la sua funzione di probabilità. Ovviamente si deve avere che

$$p_j \ge 0 \qquad (j = 1, 2, \ldots), \qquad \sum_{j:\, x_j \in S} p_j = 1. \tag{8.18}$$

Un metodo generale per simulare la variabile aleatoria discreta X è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;
STEP 2: Porre

$$X = \begin{cases} x_1, & 0 \le U < p_1 \\ x_2, & p_1 \le U < p_1 + p_2 \\ \cdots\cdots\cdots & \\ x_j, & \displaystyle\sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots & \end{cases}$$

Proposizione 8.4 *La variabile aleatoria discreta X generata con tale algoritmo ha funzione di probabilità (8.17).*

Dimostrazione Per ogni $j = 1, 2, \ldots$ si ha

$$P(X = x_j) = \sum_{k:x_k \in S} P\Big(X = x_j,\, \sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big)$$
$$= \sum_{k:x_k \in S} P\Big(\sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big)\, P\Big(X = x_j \,\Big|\, \sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big)$$

dove si è posto $\sum_{i=1}^{k-1} p_i = 0$ se $k = 1$. Poiché

$$P\Big(X = x_j \,\Big|\, \sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big) = \begin{cases} 1, & k = j \\ 0, & \text{altrimenti,} \end{cases}$$

segue che

$$P(X = x_j) = P\Big(\sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i\Big) = \sum_{i=1}^{j} p_i - \sum_{i=1}^{j-1} p_i = p_j.$$

La variabile aleatoria X generata con il precedente algoritmo ha quindi funzione di probabilità (8.17). □

L'algoritmo per simulare una variabile aleatoria discreta con funzione di probabilità (8.17) è quindi il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Se $0 \le U < p_1$ porre $X = x_1$ e terminare;
Se $U < p_1 + p_2$ porre $X = x_2$ e terminare;
Se $U < p_1 + p_2 + p_3$ porre $X = x_3$ e terminare;
. . . . . .
. . . . . .
Vediamo ora come generare alcune particolari variabili aleatorie discrete.

Problema 8.10 Simulazione del numero di utenti presenti in un sistema di servizio $M/M/1$ in condizioni di equilibrio statistico
Se si denota con N il numero di utenti presenti nel sistema di servizio $M/M/1$ in condizioni di equilibrio statistico, si ha:

$$q_n = P(N = n) = \varrho^n (1 - \varrho) \qquad (n = 0, 1, \ldots) \tag{8.19}$$

con $\varrho = \lambda/\mu < 1$. Un metodo generale per simulare N è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$N = \begin{cases} 0, & 0 \le U < q_0 \\ 1, & q_0 \le U < q_0 + q_1 \\ \vdots\vdots\vdots & \\ j, & \sum_{i=0}^{j-1} q_i \le U < \sum_{i=0}^{j} q_i \\ \vdots\vdots\vdots & \end{cases}$$

Poiché da (8.19) segue che

$$\sum_{i=0}^{j-1} q_j = (1 - \varrho) \sum_{i=0}^{j-1} \varrho^j = 1 - \varrho^j,$$

occorre porre $N = j$ se risulta

$$\sum_{i=0}^{j-1} q_i \le U < \sum_{i=0}^{j} q_i \Longleftrightarrow 1 - \varrho^j \le U < 1 - \varrho^{j+1}$$

$$\Longleftrightarrow \varrho^{j+1} < 1-U \leq \varrho^{j} \Longleftrightarrow (j+1)\ln\varrho < \ln(1-U) \leq j\ln\varrho$$
$$\Longleftrightarrow \frac{\ln(1-U)}{\ln\,\varrho} - 1 < j \leq \frac{\ln(1-U)}{\ln\,\varrho},$$

dove l'ultima disuguaglianza segue poiché $\ln\varrho < 0$ essendo $\varrho < 1$. Quindi l'algoritmo per simulare la variabile aleatoria N, che descrive il numero di utenti presenti nel sistema $M/M/1$ in condizioni di equilibrio statistico, è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;
STEP 2: Porre

$$N = \left\lceil \frac{\ln(1-U)}{\ln\,\varrho} - 1 \right\rceil,$$

dove $\lceil x \rceil$ denota il più piccolo intero maggiore di x.

□

Problema 8.11 Simulazione di una variabile aleatoria geometrica
Sia X una variabile aleatoria con funzione di probabilità geometrica

$$p_j = P(X=j) = (1-p)^{j-1}\,p \qquad (j=1,2,\ldots). \tag{8.20}$$

La variabile aleatoria geometrica X permette di descrivere il tempo di attesa per ottenere il primo successo in una successione di prove indipendenti di Bernoulli in cui la probabilità di successo è p e la probabilità di insuccesso è $1-p$. Un metodo generale per simulare X è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;
STEP 2: Porre

$$X = \begin{cases} 1, & 0 \leq U < p_1 \\ 2, & p_1 \leq U < p_1 + p_2 \\ \cdots\cdots\cdots & \\ j, & \displaystyle\sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots & \end{cases}$$

dove $\sum_{i=1}^{j-1} p_i = 0$ se $j=1$. Dalla (8.20) risulta

$$\sum_{i=1}^{j} p_i = \sum_{i=1}^{j} (1-p)^{i-1}\,p = p\sum_{k=0}^{j-1}(1-p)^k = 1-(1-p)^j.$$

e quindi occorre porre $X=j$ se e solo se:

$$\sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \Longleftrightarrow 1-(1-p)^{j-1} \leq U < 1-(1-p)^j$$
$$\Longleftrightarrow (1-p)^j < 1-U \leq (1-p)^{j-1}$$

$$\begin{aligned}&\Longleftrightarrow j\ln(1-p) < \ln(1-U) \leq (j-1)\ln(1-p)\\ &\Longleftrightarrow \frac{\ln(1-U)}{\ln(1-p)} < j \leq 1 + \frac{\ln(1-U)}{\ln(1-p)}\,.\end{aligned}$$

Per simulare una variabile aleatoria geometrica possiamo quindi utilizzare il seguente algoritmo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;
STEP 2: Porre

$$X = \left\lceil \frac{\ln(1-U)}{\ln(1-p)} \right\rceil.$$

dove $\lceil x \rceil$ denota il più piccolo intero maggiore di x.

□

Problema 8.12 Simulazione di una variabile aleatoria di Bernoulli
Sia X una variabile aleatoria di Bernoulli caratterizzata da funzione di probabilità

$$P(X=0) = 1-p, \qquad P(X=1) = p,$$

con $0 < p < 1$, dove p denota la probablità di successo e $1-p$ la probabilità di insuccesso. Un metodo generale per simulare X è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;
STEP 2: Porre

$$X = \begin{cases} 0, & 0 \leq U < 1-p \\ 1, & 1-p \leq U < 1. \end{cases}$$

□

Problema 8.13 Simulazione di una variabile aleatoria binomiale di parametri (n,p)
Sia X una variabile aleatoria binomiale di parametri (n,p) caratterizzata da funzione di probabilità

$$P(X=j) = \binom{n}{j} p^j (1-p)^{n-j} \qquad (j = 1,2,\ldots,n). \tag{8.21}$$

La variabile aleatoria binomiale X di parametri (n,p) permette di descrivere il numero di successi ottenuti in n prove indipendenti di Bernoulli in cui la probabilità di successo è p e la probabilità di insuccesso $1-p$.

Una variabile aleatoria binomiale di parametri (n,p) può essere facilmente simulata ricordando che essa si può esprimere come la somma di n variabili aleatorie indipendenti $X_1, X_2, \ldots, X_n$ distribuite secondo Bernoulli, ossia tali che

$$P(X_i = 0) = 1-p, \qquad P(X_i = 1) = p \qquad (i = 1,2,\ldots,n).$$

Per simulare una variabile aleatoria binomiale di parametri (n, p) possiamo quindi considerare il seguente algoritmo:

Algoritmo
STEP 1: Generare n variabili aleatorie indipendenti $U_1, U_2, \ldots, U_n$ uniformemente distribuite in $(0, 1)$;
STEP 2: Porre

$$X_i = \begin{cases} 0, & 0 \le U_i < 1-p \\ 1, & 1-p \le U_i < 1 \end{cases} \qquad (i = 1, 2, \ldots, n)$$

STEP 3: Valutare

$$X = \sum_{i=1}^{n} X_i.$$

Una difficoltà del precedente algoritmo è che richiede la generazione di n variabili aleatorie indipendenti e uniformemente distribuite. Esiste un differente algoritmo basato sulla generazione di un'unica variabile aleatoria uniformemente distribuita. In primo luogo notiamo che se U_k è una variabile aleatoria uniformemente distribuita in $(0, 1)$, allora la variabile aleatoria

$$U_{k+1} = \begin{cases} \dfrac{U_k}{1-p}, & 0 \le U_k < 1-p \\ \dfrac{U_k - (1-p)}{p}, & 1-p \le U_k < 1 \end{cases}$$

gode della seguente proprietà

$$P(U_{k+1} < 1-p) = 1-p, \qquad P(U_{k+1} \ge 1-p) = 1 - P(U_{k+1} < p) = p.$$

Infatti, si ha:

$$\begin{aligned} P(U_{k+1} < 1-p) &= P(U_{k+1} < 1-p,\, U_k < 1-p) + P(U_{k+1} < 1-p,\, U_k \ge 1-p) \\ &= P(U_k < 1-p)\, P(U_{k+1} < 1-p \mid U_k < 1-p) \\ &\quad + P(U_k \ge 1-p)\, P(U_{k+1} < 1-p \mid U_k \ge 1-p) \\ &= (1-p)\, P(U_{k+1} < 1-p \mid U_k < 1-p) + p\, P(U_{k+1} < 1-p \mid U_k \ge 1-p) \\ &= (1-p)^2 + p\,(1-p) = 1-p \end{aligned}$$

avendosi:

$$\begin{aligned} P(U_{k+1} < 1-p \mid U_k < 1-p) &= P\Big(\frac{U_k}{1-p} < 1-p \Big| U_k < 1-p\Big) \\ &= P(U_k < (1-p)^2 \mid U_k < 1-p) = \frac{P(U_k < (1-p)^2, U_k < 1-p)}{P(U_k < 1-p)} \\ &= \frac{P(U_k < (1-p)^2)}{P(U_k < 1-p)} = \frac{(1-p)^2}{1-p} = 1-p, \\ P(U_{k+1} < 1-p \mid U_k \ge 1-p) &= P\Big(\frac{U_k - (1-p)}{p} < 1-p \Big| U_k \ge 1-p\Big) \end{aligned}$$

$$= P(U_k < p(1-p) + 1 - p \mid U_k \geq 1-p) = \frac{P(U_k < 1-p^2, U_k \geq 1-p)}{P(U_k \geq 1-p)}$$
$$= \frac{P(1-p \leq U_k < 1-p^2)}{P(U_k \geq 1-p)} = \frac{1-p^2-(1-p)}{p} = 1-p.$$

Quindi se U_k è uniformemente distribuita in $(0,1)$ anche U_{k+1} è uniformemente distribuita in $(0,1)$. Inoltre, si può dimostrare che U_k e U_{k+1} sono variabili aleatorie indipendenti.

L'algoritmo per simulare una variabile aleatoria binomiale di parametri (n,p) usando la generazione di un'unica variabile aleatoria uniformemente distribuita in $(0,1)$ può essere così formulato:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0,1)$;
STEP 2: Per $k = 1, 2, \ldots, n$ procedere come segue:

- se $U \leq 1-p$ porre $X_k = 0$ e aggiornare $U \leftarrow U/(1-p)$.

- se $U \geq 1-p$ porre $X_k = 1$ e aggiornare $U \leftarrow [U-(1-p)]/p$.

Si generano così le variabili aleatorie $X_1, X_2, \ldots, X_n$;
STEP 3: Valutare

$$X = \sum_{i=1}^{n} X_i$$

□

Problema 8.14 Simulazione di una variabile aleatoria di Poisson
Sia X una variabile aleatoria di Poisson di parametro λ caratterizzata da funzione di probabilità

$$p_j = P(X = j) = \frac{\lambda^j}{j!}\, e^{-\lambda} \qquad (j = 0, 1, \ldots) \tag{8.22}$$

Si noti che le probabilità (8.22) possono essere calcolate tramite la procedura iterativa:

$$\begin{aligned} p_0 &= e^{-\lambda} \\ p_{j+1} &= \frac{\lambda^{j+1}}{(j+1)!}\, e^{-\lambda} = \frac{\lambda}{j+1}\, p_j \qquad (j = 0, 1, \ldots). \end{aligned} \tag{8.23}$$

Un metodo per simulare la variabile aleatoria di Poisson X è quindi il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;

STEP 2: Porre

$$X = \begin{cases} 0, & 0 \le U < p_0 \\ 1, & p_0 \le U < p_0 + p_1 \\ \cdots\cdots\cdots & \\ n, & \sum_{i=0}^{n-1} p_i \le U < \sum_{i=0}^{n} p_i \\ \cdots\cdots\cdots & \end{cases}$$

dove le p_i sono valutate tramite la procedura iterativa (8.23). □

8.5 Processo di Poisson

Per simulare un processo stocastico di Poisson $\{N(t),\ t \ge 0\}$ di parametro ϱ, dove $N(t)$ descrive il numero di chiamate che arrivano ad un centralino telefonico fino al tempo t, si può utilizzare l'algoritmi descritto nel precedente paragrafo. Ricordando infatti che per ogni $t > 0$ la funzione di probabilità di $N(t)$, fornita in (3.1), è di Poisson con valore medio $\varrho\, t$, per simulare $N(t)$ basta porre nel precedente algoritmo $\lambda = \varrho\, t$.

Denotiamo con T_s la variabile aleatoria che descrive il *tempo del primo arrivo dopo* s. Come si evince dalla Figura 8.1 l'intervallo di tempo che intercorre tra s e il tempo T_s del primo arrivo dopo s è un intervallo residuo di interarrivo. Poiché il processo degli arrivi è di Poisson, l'intervallo residuo di interarrivo ha la stessa distribuzione esponenziale dell'intervallo di interarrivo.

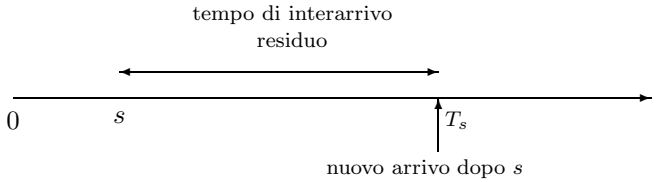

Figura 8.1: Tempo del primo arrivo dopo s per il processo di Poisson.

La variabile aleatoria T_s può quindi essere così simulata:

◇ **Algoritmo per la generazione dei tempi di arrivo di un processo di Poisson**

Step 1: Porre $t = s$;

Step 2: Generare una variabile aleatoria V uniformemente distribuita nell'intervallo $(0, 1)$;

Step 3: Porre

$$T_s = s - \frac{1}{\varrho} \ln(1 - V).$$

Utilizzando iterativamente il precedente algoritmo possiamo simulare gli istanti di tempo $t = a_1, a_2, \ldots$ in cui si verificano gli arrivi del processo di Poisson, che corrispondono ai cambiamenti di stato del sistema. Descriviamo quindi l'evoluzione del processo soltanto in tali istanti di tempo.

Algoritmo
STEP 1: Per ogni fissato istante di tempo t (ottenuto tramite l'algoritmo di generazione dei tempi di arrivo), generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$N(t) = \begin{cases} 0, & 0 \le U < p_0(t) \\ 1, & p_0(t) \le U < p_0(t) + p_1(t) \\ \cdots\cdots\cdots & \\ n, & \displaystyle\sum_{i=0}^{n-1} p_i(t) \le U < \sum_{i=0}^{n} p_i(t) \\ \cdots\cdots\cdots & \end{cases}$$

dove le $p_i(t)$ sono valutate tramite la procedura iterativa

$$p_0(t) = e^{-\varrho\, t}$$

$$p_{j+1}(t) = \frac{(\varrho\, t)^{j+1}}{(j+1)!}\, e^{-\varrho\, t} = \frac{\varrho\, t}{j+1}\, p_j(t) \qquad (j = 0, 1, \ldots).$$