UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

CORSO DI LAUREA TRIENNALE IN INFORMATICA

TESI DI LAUREA TRIENNALE IN INFORMATICA

Reti tandem e analisi delle loro misure prestazionali

Relatore:
Ch.ma Prof.ssa
Amelia Giuseppina Nobile

Candidato:
Lorenzo Carpentieri
Matr.: 0512105205

ANNO ACCADEMICO 2019/2020

Indice

Capitolo 1

Introduzione

Il lavoro di tesi svolto, dal titolo *"Reti tandem e analisi delle loro misure prestazionali"*, ha lo scopo di analizzare reti di code tandem le cui risorse sono dei sistemi di servizio disposti in sequenza e descritti mediante processi stocastici di nascita morte.
L'elaborato è strutturato, principalmente, in tre sezioni in cui si spazia dalla trattazione teorica di sistemi di servizio e reti di code all'analisi dei loro parametri prestazionali.
Nella prima parte si illustra l'importanza dello studio della *teoria delle file di attesa* il cui obiettivo è 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. Si descrivono le componenti che caratterizzano i sistemi di servizio e si definiscono i meccanismi di arrivo e di partenza degli utenti. Un ampio settore è dedicato allo studio dei *processi stocastici* i quali costituiscono un potente mezzo per analizzare sistemi di servizio al variare del tempo. In particolare si evidenzia l'importanza dei *processi di Poisson* e *processi di nascita morte*. Si enunciano, inoltre, una serie di teoremi fondamentali nell'ambito della teoria delle code quali il *principio PASTA* e il *teorema di Burke*. Si analizzano, infine, dal punto di vista teorico le prestazioni di un sistema $M/M/1$ in condizioni di equilibrio statistico.
Nella seconda parte si introducono le conoscenze fondamentali necessarie allo studio della teoria delle reti di code il cui scopo è l'analisi di modelli matematici atti a descrivere un insieme di sistemi di servizio (detti risorse o nodi della rete), ognuno costituito da un centro di attesa e da un centro di servizio. Si effettua una classificazione delle reti di code (aperte, chiuse, miste) e si descrive, mediante una *struttura di interconnessione*, il modo in cui gli utenti passano da una risorsa all'altra. In particolar modo il focus si pone sullo studio di reti aperte acicliche con un cospicuo riferimento alle reti tandem, introdotte da Jackson nel 1954, le quali sono reti di code aperte acicliche costituite da un insieme finito di sistemi di servizio collegati in serie. Si enunciano, inoltre, una serie di risultati di notevole importanza per lo studio delle reti di code quali le *Leggi di Little locali e globali* e il *teorema di Jackson*. Un ampio spazio, infine, è dedicato allo studio del modello geometrico e binomiale negativo per le singole risorse della rete evidenziandone anche le loro differenze.
Nella terza parte si analizzano dettagliatamente le differenze tra reti tandem con distribuzione geometrica e reti tandem con distribuzione binomiale negativa al variare dei parametri di input. In particolare sono state esaminate reti tandem al crescere del numero di risorse e dei parametri β e ρ. I risultati ottenuti dagli studi effettuati sono resi evidenti mediante l'utilizzo di grafici realizzati in R nei quali si illustrano le relazioni sia tra media e varianza sia tra le distribuzioni di probabilità. Si illustra, inoltre, l'importanza della generazione di variabili aleatorie discrete necessarie alla simulazione di reti tandem caratterizzate da distribuzione geometrica e binomiale negativa. In particolar modo si focalizza l'attenzione sulla simulazione di variabili aleatorie geometriche e binomiali negative fornendo algoritmi per poterle generare. Si è utilizzato, infine, il

linguaggio R per ottenere la simulazione del numero medio di utenti presenti in reti tandem caratterizzate da distribuzione geometrica e binomiale negativa in condizioni di equilibrio statistico e si sono poi confrontati, mediante l'ausilio di grafici, i risultati teorici con quelli simulati.

Capitolo 2

Sistemi di servizio

2.1 Introduzione

Nello svolgimento delle attività quotidiane siamo continuamente soggetti alla necessità di richiedere l'erogazione di uno o più servizi a svariati enti con la possibilità che si crei una fila di attesa. Tutto ciò accade poiché fornire un servizio richiede del tempo e non sempre si è in grado di gestire le molteplici richieste dei clienti.
Lo scopo dell'ente è garantire un tempo di attesa in fila minimo per invogliare i clienti ad usufruire del servizio e aumentare il proprio guadagno.
Minimizzare il tempo di attesa in fila dei clienti, però, comporta un aumento delle spese dell'ente il quale dovrà investire le proprie risorse in impiegati o macchinari per poter incrementare la velocità di erogazione del servizio.
In tale contesto emerge l'importanza dello studio della *teoria delle file di attesa* che 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 i clienti in banca o in posta, le persone in attesa di un taxi o le chiamate ad un centralino telefonico.
I risultati forniti dalla teoria delle file di attesa trovano applicazione in numerosi campi: sistemi di elaborazione, sistemi di comunicazione e trasmissione dati, sistemi per la gestione di servizi pubblici e privati e tantissimi altri settori.
La teoria delle file di attesa è essenzialmente di natura probabilistica e permette di analizzare accuratamente le prestazioni di sistemi di servizio calcolando una serie di parametri al fine di stabilire la qualità del sistema e le misure da adottare per ottenere un miglioramento delle prestazioni minimizzando i costi.

2.2 Componenti di un sistema di servizio

Un generico sistema di servizio può essere schematizzato come illustrato nella Figura 2.1

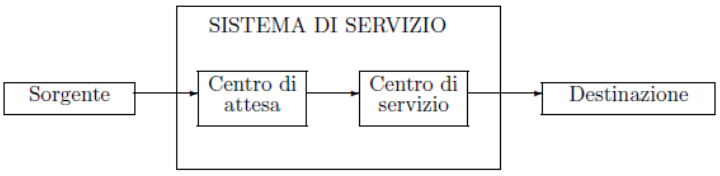


Figura 2.1 Rappresentazione di un sistema di servizio.

- **La sorgente** ossia i potenziali utenti che possono richiedere un servizio al sistema;
- **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 del sistema in cui sono erogati i servizi messi a disposizione;
- **La destinazione** ossia l'insieme delle richieste di servizio che essendo state soddisfatte lasciano i punti di servizio.

Sorgente
La sorgente è costituita dai potenziali utenti che possono richiedere uno o più servizi al sistema. Una sorgente è *finita* se il numero di utenti che possono accedere al servizio è ragionevole oppure *infinita* se il numero di utenti che possono accedere al servizio è molto grande.
Analizzare sistemi di servizio la cui sorgente è finita risulta più complesso rispetto a sistemi di servizio con sorgenti infinite in quanto avere una sorgente finita causa una variazione dei parametri di arrivo dovuta al numero di utenti nel sistema, infatti, se tutti i potenziali utenti sono già arrivati nel sistema i parametri di arrivo si annullano. Per tale motivo se la sorgente è finita ma ci sono molti potenziali utenti si assume che la sorgente sia infinita per semplificare la trattazione matematica.
Gli utenti possono anche provenire da distinte sorgenti. Gli utenti che provengono da una stessa sorgente sono tra loro indistinguibili ovvero gli utenti di una stessa sorgente sono uguali tra loro. 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.

Centro di attesa
L'accesso al centro di servizio è realizzato da un centro di attesa che può contenere un numero limitato o illimitato di utenti. Un centro di attesa può essere a capacità *nulla*, *finita* o *infinita*.
In un centro di attesa a capacità nulla il numero di utenti massimo che possono attendere in coda è nullo quindi un utente che arriva nel sistema o usufruisce immediatamente del servizio ed esce dal sistema oppure viene rifiutato.
Uno scenario sì fatto può essere descritto da un centralino telefonico che non ha la possibilità di avere chiamate in attesa. In tal caso un utente che effettua una chiamata potrebbe trovare la linea libera ed usufruire immediatamente del servizio oppure trovare la linea occupata ed essere rifiutato.
In un centro di attesa a capacità finita il numero di utenti massimo che possono attendere in coda è finito. Fin quando la coda non è satura gli utenti entrano nel sistema e restano in attesa del servizio. Quando la coda, invece, è satura i successivi utenti che cercano di accedere al sistema saranno rifiutati.
Un esempio che descrive perfettamente tale situazione è un centralino telefonico che ha la possibilità di avere un numero finito k di chiamate in attesa. In tal caso un utente che effettua una chiamata potrebbe trovare la linea libera ed usufruire immediatamente del servizio oppure trovare la linea occupata e restare in attesa, essendo il numero di utenti in attesa minore di k, oppure essere rifiutato, essendo il numero di utenti in attesa maggiore o uguale a k.
In un centro di attesa a capacità infinita non c'è limite al numero di utenti che possono attendere in coda. In tal caso tutti gli utenti che entrano nel sistema di servizio o usufruiscono immediatamente del servizio o restano in attesa.
Un esempio che descrive tale situazione è un centralino telefonico che ha la possibilità di avere un numero illimitato di chiamate in attesa. In tal caso un utente che effettua la chiamata potrebbe

trovare la linea libera ed usufruire immediatamente del servizio oppure trovare la linea occupata e restare in attesa.
I sistemi di servizio con centro di attesa a capacità nulla e finita non sono soggetti a congestione in quanto gli utenti in eccesso sono rifiutati. I sistemi di servizio con centro di attesa a capacità infinita, invece, sono soggetti a congestione nel momento in cui i servitori, coloro che offrono il servizio, non sono in grado di gestire il flusso di utenti che arriva nel sistema. Uno degli obiettivi della teoria delle file di attesa è proprio stabilire quali sono le condizioni necessarie affinché un sistema non si congestioni.
Solitamente i sistemi di servizio sono costituiti da un unico centro di attesa ma in alcuni casi è possibile avere sistemi con più centri di attesa e in tal caso si preferisce di parlare di *reti di code*.

Centro di servizio
Dal centro di attesa gli utenti accedono al centro di servizio che è composto da uno o più servitori. Il servitore è un'entità che si occupa di erogare i servizi offerti dal sistema. I servitori sono disposti in parallelo e nessun servitore può restare inattivo se sono presenti degli utenti da servire. Se tutti i servitori sono occupati allora l'utente che entra nel sistema sarà collocato nella fila di attesa se non è satura e attenderà che uno dei servitori si liberi.
Nell'analisi teorica di un sistema di servizio i servitori sono identici tra loro per poter semplificare il calcolo dei parametri prestazionali del sistema. Nella fase di simulazione, invece, si possono anche considerare servitori non identici per simulare al meglio il sistema reale di cui si analizzano le caratteristiche.

Destinazione
Dopo aver usufruito del servizio offerto l'utente lascia immediatamente il sistema. L'insieme delle richieste di servizio espletate sono instradate verso la destinazione.

Discipline di servizio
Le discipline di servizio consentono di definire le modalità con cui gli utenti transitano dal centro di attesa al centro di servizio. Esistono diversi tipi di discipline di servizio:

- **FCFS** (First come-first served): gli utenti sono serviti seguendo l'ordine di arrivo ovvero il primo che arriva è il primo ad essere servito.
- **LIFO** (Last in first out): l'ultimo utente che arriva è il primo ad essere servito.
- **SIRO** (Service in random order): gli utenti sono serviti in ordine casuale.
- **PRS** (Priority server): gli utenti sono serviti in base alla priorità. Ad esempio in un sistema che descrive i servizi erogati da un ospedale potrebbe essere applicata una disciplina PRS che garantisce ai pazienti in condizioni più gravi una maggiore probabilità di usufruire dei servizi richiesti.

2.3 Tempi di interarrivo e tempi di servizio

Per poter descrivere un generico sistema di servizio è necessario definire i meccanismi di arrivo e di servizio ovvero stabilire il tipo di distribuzione dei *tempi di interarrivo* e dei *tempi di servizio*.
I tempi di interarrivo rappresentano la lunghezza dell'intervallo di tempo tra un arrivo e l'arrivo successivo. I tempi di servizio rappresentano, invece, il tempo necessario per servire un utente per servitore.

Meccanismi di arrivo 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. Un esempio che descrive perfettamente un meccanismo di arrivo questo tipo è un sistema che rappresenta una catena di montaggio in cui gli arrivi avvengono ad intervalli regolari.
Si supponga che 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:

$$A(t) = P(T < t) = \begin{cases} 0, & \text{se } t \leq 1/\lambda \\ 1, & \text{se } t > 1/\lambda \end{cases} \tag{2.1}$$

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

$$E(T) = 1/\lambda, \quad Var(T) = 0 \tag{2.2}$$

Meccanismi di arrivo 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:

$$A(t) = P(T < t) = \begin{cases} 0, & \text{se } \ \mathrm{t} \leq \mathrm{a} \\ \dfrac{t-a}{b-a}, & \text{se a} < \mathrm{t} \leq \mathrm{b} \\ 1, & \text{se t} > \mathrm{b} \end{cases} \tag{2.3}$$

La densità di probabilità è:

$$a(t) = \begin{cases} \dfrac{1}{b-a}, & \text{se } \ \mathrm{a} < \mathrm{t} < \mathrm{b} \\ 0, & \text{altrimenti} \end{cases} \tag{2.4}$$

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

$$E(T) = \frac{a+b}{2}, \quad Var(T) = \frac{(b-a)^2}{12} \tag{2.5}$$

Meccanismi di arrivo di tipo M
Nel meccanismo degli arrivi di tipo M i tempi di interarrivo sono indipendenti e identicamente distribuiti con funzione di distribuzione esponenziale. 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:

$$A(t) = P(T < t) = \begin{cases} 0, & \text{se } t \leq 0 \\ \\ 1 - e^{-\lambda t}, & \text{se } t > 0 \end{cases} \tag{2.6}$$

La densità di probabilità è:

$$a(t) = \begin{cases} \lambda e^{-\lambda t}, & \text{se } t > 0 \\ \\ 0, & \text{altrimenti} \end{cases} \tag{2.7}$$

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

$$E(T) = \frac{1}{\lambda}, \quad Var(T) = \frac{1}{\lambda^2} \tag{2.8}$$

La funzione di distribuzione esponenziale è frequentemente utilizzata nella teoria delle file di attesa per le importanti proprietà di cui essa gode. La lettera M significa Markov ed indica la mancanza di memoria della distribuzione esponenziale ovvero:

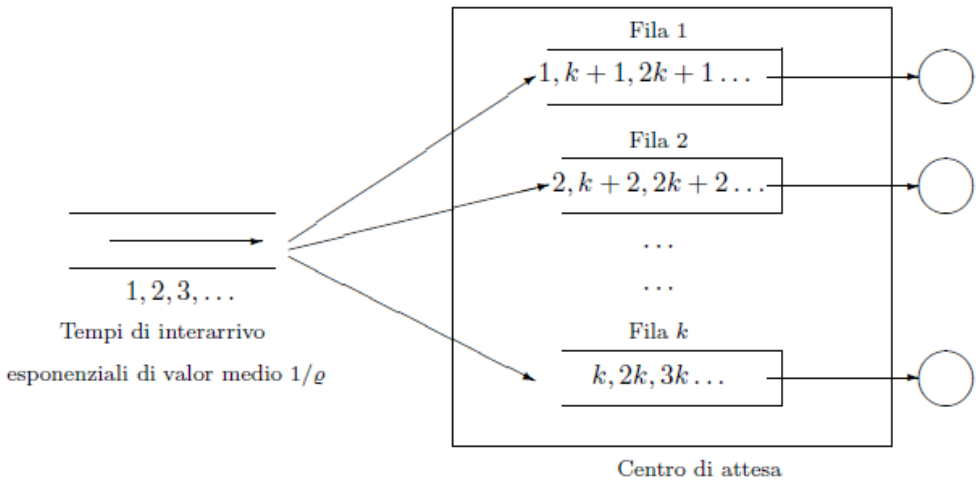
$$P(T > t + s | T > s) = P(T > t) \tag{2.9}$$

La formula (2.9) esprime 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. 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 si denota con T un generico tempo di interarrivo e con Z una variabile aleatoria che descrive il tempo di interarrivo residuo, ossia $Z = T$ - τ , se $t > 0$ si ha:

$$\begin{aligned} P(Z \leq t | Z > 0) = 1 - P(Z > t | Z > 0) = 1 - P(T - \tau > t | T > \tau) = 1 - P(T > t + \tau | T > \tau) \\ = 1 - P(T > t) = P(T \leq t) = 1 - e^{-\lambda t} \end{aligned} \tag{2.10}$$

, ossia una distribuzione esponenziale di parametro $1/\lambda$.
Da ciò si può stabilire che essendo il periodo di ozio (cioè intervallo di tempo in cui il sistema è vuoto) un tempo di interarrivo residuo anch'esso è distribuito esponenzialmente.
La presenza di periodi di ozio nel sistema garantisce che il sistema non si congestioni questo perché avere un intervallo di tempo in cui il sistema è vuoto sta ad indicare che i servitori sono in grado si smaltire il flusso di utenti che arrivano nel sistema.

Meccanismi di arrivo di tipo E_k
Si consideri il sistema di servizio, schematizzato in figura, 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 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,...); alla generica j–esima (j = 1, 2,..., 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,...).



Si supponga che i tempi di interarrivo degli utenti che accedono al sistema siano indipendenti ed esponenzialmente distribuiti con valore medio $1/\rho$. Si denoti 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 $1/\rho$. La somma di k variabili aleatorie indipendenti di tipo esponenziale con valore medio $1/\rho$ è distribuita con densità di probabilità di Erlang di ordine k, ossia:

$$a(t) = \begin{cases} \dfrac{\rho^k}{(k-1)!}\, e^{-\rho t}\, t^{(k-1)}, & \text{t} > 0 \\ 0, & \text{altrimenti} \end{cases} \tag{2.11}$$

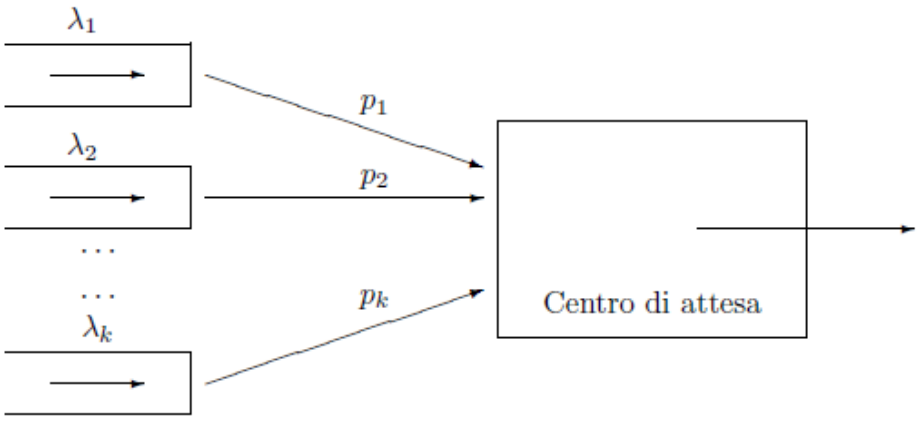
Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = E(T_1 + ... + T_\text{k}) = E(T_1) + ... + E(T_\text{k}) = \frac{k}{\rho}, \tag{2.12}$$

$$Var(T) = Var(T_1 + ... + T_\text{k}) = Var(T_1) + ... + Var(T_\text{k}) = \frac{k}{\rho^2} \tag{2.13}$$

Meccanismi di arrivo di tipo H_k
Si consideri il sistema di servizio, schematizzato in figura, che ha il compito di servire k differenti sorgenti.
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.



Si supponga che i tempi 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, . . . , 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, . . . , k) e ad avviare al centro di attesa la prima delle richieste di servizio relative alla sorgente selezionata. Si assuma che

$$p_j \geq 0 \quad (j = 1, 2, \ldots, k), \quad \sum_{j=1}^{k} p_j = 0 \tag{2.14}$$

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. Si denoti con T la variabile aleatoria che descrive la lunghezza dell'intervallo di tempo tra due arrivi successivi al centro di attesa del sistema. La funzione di distribuzione è

$$A(t) = \begin{cases} 1 - \sum_{j=1}^{k} p_j e^{-\lambda_j t}, & t > 0 \\ \\ 0, & \text{altrimenti} \end{cases} \tag{2.15}$$

La funzione di densità è

$$a(t) = \begin{cases} \sum_{j=1}^{k} p_j \lambda_j e^{-\lambda_j t}, & t > 0 \\ \\ 0, & \text{altrimenti} \end{cases} \tag{2.16}$$

Dalla (2.16) è possibile ricavare immediatamente il valore medio e la varianza del tempo di interarrivo, ossia

$$E(T) = \sum_{j=1}^{k} \frac{p_j}{\lambda_j}, \tag{2.17}$$

$$Var(T) = 2\sum_{j=1}^{k} \frac{p_j}{\lambda_j^2} - \left[\sum_{j=1}^{k} \frac{p_j}{\lambda_j}\right]^2 \tag{2.18}$$

Meccanismi 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

$$B(t) = P(S < t) = \begin{cases} 0, & \text{se } t \leq 1/\mu \\ 1, & \text{se } t > 1/\mu \end{cases} \tag{2.19}$$

Il valore medio e la varianza del tempo di servizio saranno rispettivamente:

$$E(S) = 1/\mu, \quad Var(S) = 0 \tag{2.20}$$

Meccanismi 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 si suppone che la variabile aleatoria S sia uniformemente distribuita in (a, b), allora la funzione di distribuzione è

$$B(t) = P(S < t) = \begin{cases} 0, & \text{se } \ \mathrm{t} \leq \mathrm{a} \\ \dfrac{t-a}{b-a}, & \text{se } \mathrm{a} < \mathrm{t} \leq \mathrm{b} \\ 1, & \text{se } \mathrm{t} > \mathrm{b} \end{cases} \tag{2.21}$$

La densità di probabilità è:

$$b(t) = \begin{cases} \dfrac{1}{b-a}, & \text{se } \ \mathrm{a} < \mathrm{t} < \mathrm{b} \\ 0, & \text{altrimenti} \end{cases} \tag{2.22}$$

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

$$E(S) = \frac{a+b}{2}, \quad Var(S) = \frac{(b-a)^2}{12} \tag{2.23}$$

Meccanismi 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

$$B(t) = P(S < t) = \begin{cases} 0, & \text{se } t \leq 0 \\ 1 - e^{-\mu t}, & \text{se } t > 0 \end{cases} \tag{2.24}$$

La densità di probabilità è:

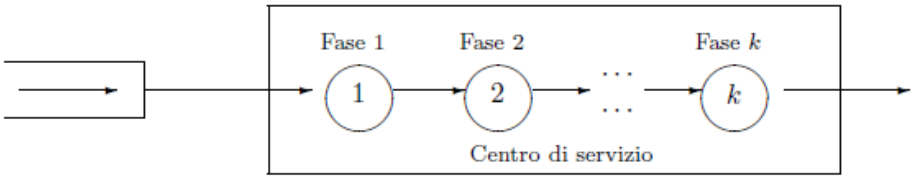
$$b(t) = \begin{cases} \mu e^{-\mu t}, & \text{se } t > 0 \\ 0, & \text{altrimenti} \end{cases} \tag{2.25}$$

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

$$E(T) = \frac{1}{\mu}, \quad Var(T) = \frac{1}{\mu^2} \tag{2.26}$$

La funzione di distribuzione esponenziale è frequentemente utilizzata nella teoria delle file di attesa per le importanti proprietà di cui essa gode.

Meccanismi di servizio di tipo E_k
Si consideri il sistema di servizio, schematizzato in figura, 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...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 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). Sia, inoltre, 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

$$S = S_1 + S_2 + ... + S_k$$

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

$$b(t) = \begin{cases} \dfrac{(k\mu)^k}{(k-1)!}\, e^{-k\mu t}\, t^{(k-1)}, & \text{t} > 0 \\ \\ 0, & \text{altrimenti} \end{cases} \tag{2.27}$$

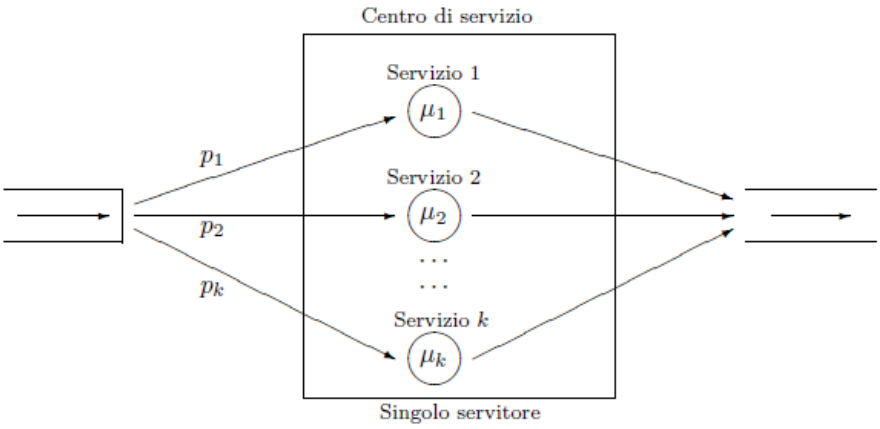
Il valore medio e la varianza del tempo di servizio sono rispettivamente:

$$E(S) = E(S_1 + ... + S_{\text{k}}) = E(S_1) + ... + E(S_{\text{k}}) = \frac{1}{\mu}, \tag{2.28}$$

$$Var(S) = Var(S_1 + ... + S_{\text{k}}) = Var(S_1) + ... + Var(S_{\text{k}}) = \frac{1}{k\mu^2} \tag{2.29}$$

Meccanismi di servizio di tipo $\mathbf{H_k}$

Si consideri il sistema di servizio, schematizzato in figura, consistente in un centro di servizio costituito da un unico servitore che provvede a fornire k tipi di differenti servizi.



Si supponga che la probabilità che l'utente richieda un servizio di tipo j sia p_j per ogni j=1,2,. . .,k dove

$$p_{\text{j}} \geq 0 \quad (j = 1, 2, \ldots, k), \quad \sum_{j=1}^{k} p_j = 0 \tag{2.30}$$

Si supponga, inoltre, che il j–esimo tipo di servizio (j=1,2...k) sia caratterizzato da una durata del servizio distribuita esponenzialmente con valore medio $1/\mu_j$. Un esempio tipico è quello di un impiegato incaricato per l'assistenza al pubblico che fornisce k tipi di informazioni all'ingresso di un grande ufficio. Si denoti 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. La funzione di distribuzione del tempo di servizio S è

$$B(t) = \begin{cases} 1 - \sum_{j=1}^{k} p_j e^{-\mu_j t}, & \text{t} > 0 \\ \\ 0, & \text{altrimenti} \end{cases} \tag{2.31}$$

La funzione di densità è

$$b(t) = \begin{cases} \sum_{j=1}^{k} p_j \mu_j e^{-\mu_j t}, & \text{t} > 0 \\ \\ 0, & \text{altrimenti} \end{cases} \tag{2.32}$$

Dalla (2.32) è possibile ricavare immediatamente il valore medio e la varianza del tempo di servizio, ossia

$$E(S) = \sum_{j=1}^{k} \frac{p_j}{\mu_j}, \tag{2.33}$$

$$Var(S) = 2\sum_{j=1}^{k} \frac{p_j}{\mu_j^2} - \left[\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right]^2 \tag{2.34}$$

2.4 Notazione di Kendall

La notazione di Kendall permette di descrivere in modo conciso le caratteristiche principali di un sistema di servizio. Una generica notazione di Kendall è costituita da sigle *A/B/s/K/m/Z* che identificano una specifica caratteristica del sistema in esame:

- **A**: definisce lo schema di arrivo ovvero descrive la distribuzione dei tempi di interarrivo che indicano il tempo che intercorre tra un arrivo e quello successivo;
- **B**: definisce lo schema di partenza ovvero descrive la distribuzione dei tempi di servizio che indicano il tempo necessario per servire un utente per servitore;
- **s**: definisce il numero di servitori in parallelo nel sistema;
- **K**: definisce la capacità del sistema ovvero il numero massimo di utenti che il sistema può ospitare. Ad esempio in un sistema con un servitore la cui capacità è k avremo che nella fila di attesa potranno esserci al massimo *k-1* utenti e un altro utente sarà in servizio;
- **m**: definisce la dimensione della popolazione, ovvero il numero di potenziali utenti;
- **Z**: definisce il tipo di disciplina di servizio utilizzata.

Le ultime 3 sigle della notazione di Kendall possono essere omesse e in tal caso assumono valori di default: in mancanza della sigla K si suppone che la capacità del sistema sia infinita, in mancanza della sigla m la sorgente sarà infinita e in mancanza della sigla Z la disciplina di servizio adottata sarà quella FCFS (First come-first served).

Capitolo 3

Processo di Poisson e processo di nascita morte

3.1 Introduzione ai processi stocastici

I *processi stocastici* costituiscono un potente mezzo per analizzare sistemi di servizio al variare del tempo.
Si definisce processo stocastico una famiglia di variabili aleatorie $\{X(t), t \in T \subset R^+\}$ dipendenti dal tempo, che assumono valori in un insieme detto spazio degli stati del processo.
Lo spazio degli stati e lo spazio della variabile tempo possono essere continui oppure discreti se assumono un numero finito o al più numerabile di valori.
In particolare, per l'analisi di sistemi di servizio siamo interessati ai *processi di Poisson* e *processi di nascita morte* i quali sono processi stocastici continui nel tempo e discreti nello spazio degli stati. Le realizzazioni di tali processi sono funzioni a gradino, ovvero funzioni costanti a tratti con salti diretti verso il basso o verso l'alto ogni volta che si verifica un cambiamento di stato.

3.2 Processo stocastico di Poisson

Un processo di Poisson, dal nome del matematico francese Siméon-Denis Poisson, è un processo stocastico che descrive il manifestarsi di eventi indipendenti l'uno dall'altro e che si verificano continuamente nel tempo.
Nell'ambito dei sistemi di servizio tale processo permette di rappresentare l'arrivo degli utenti al sistema.
Il processo di Poisson si rivela molto utile nella descrizione di alcuni fenomeni che evolvono nel tempo come l'arrivo di chiamate ad un call center e trova applicazione anche in numerosi ambiti scientifici ad esempio per descrivere l'attività spontanea di alcuni neuroni.
Sia N(t), t$\geq$0 il numero di arrivi nell'intervallo di tempo (0,t), N(t, t + Δt) il numero di arrivi nell'intervallo di tempo (t,t + Δt) e ρ la frequenza di arrivo nel sistema.
Si può stabilire che un processo stocastico $\{N(t), t \geq 0\}$ è detto di *Poisson* con parametro $\rho \geq 0$ se valgono le seguenti proposizioni:

1. N(0)=0, ossia il numero di arrivi nel sistema al tempo 0 è nullo cioè non possono verificarsi arrivi al tempo 0;

2. Il processo ha incrementi indipendenti e stazionari:

   - incrementi indipendenti, ovvero dati due intervalli di tempo qualsiasi I_1 e I_2 tali che $I_1 \cap I_2 = \emptyset$ (cioè i due intervalli non si sovrappongono) ciò che accade nell'intervallo I_1 non dipende da ciò che accade nell'intervallo I_2 e viceversa;

- incrementi stazionari, ovvero il numero di arrivi in un intervallo di tempo dipende solamente dalla lunghezza dell'intervallo di tempo considerato;

3. $P\{N(t,t+\Delta t)=1\}=\rho\Delta t+o(\Delta t)$, ossia la probabilità che si verifichi un arrivo nell'intervallo di tempo (t,t + Δt) è uguale a $\rho\Delta t+o(\Delta t)$ dove $o(\Delta t)$ è una quantità infinitesimale tendente allo 0 più rapidamente di Δt;

4. $P\{N(t,t+\Delta t)>1\}=o(\Delta t)$, ossia la probabilità che si verifichi più di un arrivo nell'intervallo di tempo (t,t + Δt) è pari a $o(\Delta t)$ cioè tendente allo 0 più rapidamente di Δt.

Dalla condizione 3 e 4 si può ricavare la probabilità che non si verifichino arrivi:
$P\{N(t,t+\Delta t)=0\}=1-P\{N(t,t+\Delta t)=1\}=1-\rho\Delta t+o(\Delta t)$.

Dall'ipotesi di indipendenza, inoltre, segue che:

- $P\{N(t,t+\Delta t)=1|N(t)=n\}=P\{N(t,t+\Delta t)=1\}=\rho\Delta t+o(\Delta t)$ essendo gli intervalli (0,t) e (t,t + Δt) disgiunti.
- $P\{N(t,t+\Delta t)>1|N(t)=n\}=P\{N(t,t+\Delta t)>1\}=o(\Delta t)$
- $P\{N(t,t+\Delta t)=0|N(t)=n\}=1-P\{N(t,t+\Delta t)=1\}=1-\rho\Delta t+o(\Delta t)$

Si denoti con $p_n(t)$ la probabilità che al tempo t ci siano n arrivi ossia $p_n(t)=P\{N(t)=n\}$ con $n=0,1,2...$
In un processi stocastico di Poisson si ha che:

$$p_n(t)=\frac{(\rho t)^n}{n!}e^{-\rho t}$$

ossia $p_n(t)$ è la distribuzione di probabilità di una variabile aleatoria di Poisson di parametro ρt.
Una proprietà importante del processo di Poisson è che il valore medio e la varianza assumono entrambi valori ρt:

$$E[N(t)]=\rho t,\ V[N(t)]=\rho t \tag{3.1}$$

Il coefficienti di variazione, invece, è:

$$C[N(t)]=\frac{\sqrt[2]{V[N(t)]}}{E[N(t)]}=\frac{1}{\sqrt[2]{\rho t}} \tag{3.2}$$

Si può notare che al crescere di t, ovvero per t $\to+\infty$, il coefficiente di variazione assume valore 0. Questo vuol dire che al crescere del tempo t il valore medio $E[N(t)]$ diventa sempre più significativo.
Una considerazione molto importante per quanto riguarda i processi di Poisson è che sono assimilabili a tempi di interarrivo indipendenti e identicamente distribuiti con densità esponenziale. Se si suppone che gli arrivi si verifichino ai tempi $T_1, T_1+T_2, T_1+T_2+T_3, ...$ dove il generico T_n denota la lunghezza dell'intervallo tra l'arrivo n-1-esimo e l'arrivo n-esimo si ha che:

$$f_{T_n(t)}=\begin{cases}\rho e^{-\rho t}, & \text{se } t>0\\ \\ 0, & \text{altrimenti}\end{cases} \tag{3.3}$$

Il valore medio e la varianza saranno:

$$E(T_n) = \frac{1}{\rho}, \quad Var(T_n) = \frac{1}{\rho^2} \tag{3.4}$$

Se si analizza, inoltre, l'intervallo di tempo compreso tra 0 e k si può considerare una somma di variabili aleatorie $T_1 + T_2 + T_3 + ... + T_k$ e in questo caso si ha una densità di probabilità di questo tipo:

$$f(t) = \frac{d}{dt}[P(T_1 + T_2 + ... + T_k < t)] = \begin{cases} \dfrac{\rho^k}{(k-1)!}\, e^{-\rho t}\, t^{(k-1)}, & \text{t} > 0 \\ 0, & \text{altrimenti} \end{cases} \tag{3.5}$$

che equivale ad una densità di Erlang di ordine k.

3.3 Processo stocastico di nascita morte

I processi di nascita morte, introdotti da Feller nel 1939, sono utilizzati per la costruzione di modelli di crescita della popolazione, sistemi di servizio, di epidemiologia e di molte aree di interesse sia teorico che applicativo.
Per poter introdurre i processi di nascita morte applicati ai sistemi di servizio è necessario definire le *frequenze di arrivo* e le *frequenze di partenza*:

- λ_i indica la frequenza di arrivo degli utenti nel sistema quando ci sono i utenti nel sistema;
- μ_i indica la frequenza di partenza degli utenti dal sistema quando ci sono i utenti nel sistema.

Definizione Dato $\{N(t), t \geq 0\}$ un processo stocastico avente spazio degli stati $0, 1, \ldots$ e continuo nel tempo. Si supponga che tale processo descriva un sistema che si trova nello stato E_n al tempo t se e solo se $N(t) = n$, ossia il sistema è nello stato E_n se e solo se il il numero di utenti nel sistema è esattamente n. Tale processo stocastico è detto di nascita morte se esistono $\lambda_n \quad (n = 0, 1, \ldots)$ e $\mu_n \quad (n = 1, 2, \ldots)$ tali che siano verificate le seguenti proposizioni:

1. In un intervallo di tempo di lunghezza Δt si possono verificare solo due cambiamenti di stato: dallo stato E_n allo stato E_{n+1} se si verifica un arrivo, dallo stato E_n a E_{n-1} se si verifica una partenza. Se $n = 0$ si può avere solo un arrivo poiché affinché si verifichi una partenza deve esserci almeno un utente nel sistema;
2. $P\{N(t, t + \Delta t) = n + 1 | N(t) = n\} = \lambda_n \Delta t + o(\Delta t)$ ossia la probabilità che nell'intervallo di tempo di lunghezza Δt si verifichi un cambiamento nel sistema dallo stato E_n allo stato E_{n+1} è pari al prodotto tra la frequenza di arrivo quando ci sono n utenti nel sistema e la lunghezza dell'intervallo di tempo considerato;
3. $P\{N(t, t+\Delta t) = n-1 | N(t) = n\} = \mu_n \Delta t + o(\Delta t)$ ossia la probabilità che nell'intervallo di tempo di lunghezza Δt si verifichi un cambiamento nel sistema dallo stato E_{n-1} allo stato E_n è pari al prodotto tra la frequenza di partenza quando ci sono n utenti nel sistema e la lunghezza dell'intervallo di tempo considerato;

4. $P\{N(t, t+\Delta t) = n \pm k | N(t) = n\} = o(\Delta t)$ ossia nell'intervallo di tempo di lunghezza Δt può verificarsi al più una transizione, infatti, la probabilità che si verifichi più di un arrivo o più di una partenza è pari a $o(\Delta t)$ che è una quantità infinitesimale tendente a 0 più velocemente di Δt.

Dalle assunzioni 2 e 3 si ottiene che la probabilità di non avere arrivi o partenze nel sistema nell'intervallo di tempo di lunghezza Δt è:

$$P\{N(t, t+\Delta t) = n | N(t) = n\} = \begin{cases} 1 - \lambda_n \Delta t + o(\Delta t), & \text{se } n = 0 \\ 1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t), & \text{se } n > 0 \end{cases}$$

Sia $p_n(t) = P\{N(t) = n\}$ per $n = 0, 1, 2, \ldots$ la probabilità che al tempo t ci siano n utenti nel sistema si ha che $p_0(t, t+\Delta t) = p_0(t)(1 - \lambda_0 \Delta t) + p_1(t)\mu_1 \Delta t + o(\Delta t)$ poiché per avere 0 utenti nel sistema nell'intervallo di tempo $(t, t+\Delta t)$ vuol dire che o al tempo t c'era un solo utente nel sistema e si è verificata una partenza oppure nel sistema c'erano 0 utenti e non si è verificato un arrivo.
Si ha, inoltre, che $p_n(t, t+\Delta t) = p_{n-1}(t)\lambda_{n-1}\Delta t + p_n(t)[1 - (\lambda_n + \mu_n)\Delta t] + p_{n+1}(t)\mu_{n+1}\Delta t + o(\Delta t)$ poiché per avere n utenti nel sistema nell'intervallo di tempo $(t, t+\Delta t)$ vuol dire che o c'erano n-1 utenti nel sistema e si è verificato un arrivo o c'erano n utenti nel sistema e non si è verificato né un arrivo né una partenza oppure c'erano n+1 utenti nel sistema e si è verificata una partenza.

Nella fase transiente risulta molto difficile calcolare la distribuzione di probabilità. Per questo motivo si ricorre alla simulazione per analizzare il sistema nella fase transiente e dal punto di vista teorico si analizza il sistema di servizio in condizioni di equilibrio statistico.
Sia

$$q_n = \lim_{t \to +\infty} p_n(t) \quad (n = 0, 1, 2, \ldots)$$

la probabilità che ci siano n utenti in condizioni di equilibrio statistico. Per ottenere la distribuzione di equilibrio si può costruire il *grafo di transizione* rappresentato nella Figura 3.1 ossia un grafo in cui ad ogni nodo è associato un numero che indica il numero di utenti nel sistema, ad ogni arco superiore è associata una frequenza di arrivo e ad ogni arco inferiore è associata una frequenza di partenza.

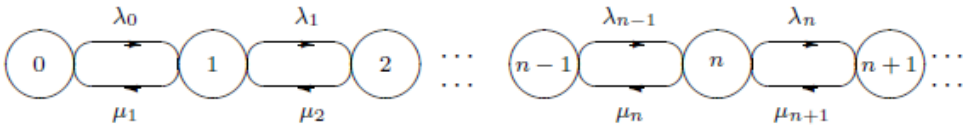


Figura 3.1 Rappresentazione di un grafo di transizione.

In condizioni di equilibrio statistico per il grafo di transizione vale *il principio di bilanciamento* che consente di uguagliare il flusso medio con cui il processo entra nello stato n con il flusso medio con cui il processo esce dallo stato n. Dall'applicazione del principio di bilanciamento si ricavano le equazioni di bilanciamento per il processo di nascita morte:

$$\mu_1 q_1 = \lambda_0 q_0$$
$$\lambda_{n-1} q_{n-1} + \mu_{n+1} q_{n+1} = (\lambda_n + \mu_n) q_n$$

Si dimostra che un processo di nascita morte raggiunge la condizione di equilibrio statistico se e solo se la serie:

$$1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n} < +\infty$$

converge ed in tal caso si ha che:

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

e

$$q_n = q_0 \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n}$$

Se, invece, la serie diverge allora non si raggiunge la condizione di equilibrio statico e la frequenza con cui gli utenti arrivano nel sistema sarà maggiore rispetto alla frequenza con cui gli utenti partono dal sistema. Questo indica che i servitori non sono in grado di smaltire il flusso di utenti che arrivano nel sistema.

3.4 Il principio PASTA

Sia $\{N(t), t \geq 0\}$ il processo stocastico che descrive il numero di utenti $N(t)$ nel sistema al tempo t e sia $p_n(t) = P\{N(t) = n\} \quad (n = 0, 1, 2, \dots)$ la probabilità che siano presenti n utenti nel sistema al tempo t.
Sia, inoltre, $q_n = \lim_{t \to +\infty} p_n(t) \quad (n = 0, 1, 2, \dots)$ la probabilità che ci siano n utenti nel sistema in condizioni di equilibrio statistico.
In ogni intervallo di tempo di lunghezza τ è possibile considerare il rapporto tra il tempo $t_n(\tau)$ che il processo spende nello stato n nell'intervallo di tempo $(0,\tau)$ e la lunghezza dell'intervallo τ considerato. In condizioni di equilibrio statistico al crescere di τ tale rapporto fornisce proprio la probabilità q_n e si ha:

$$q_n = \lim_{\tau \to +\infty} \frac{t_n(\tau)}{\tau} \quad (n = 0, 1, 2, \dots) \tag{3.6}$$

Si considerino gli eventi *A(n)* e *D(n)* così definiti:

- *A(n)* un utente in arrivo nel sistema trova esattamente n utenti nel sistema;
- *D(n)* un utente in partenza dal sistema lascia esattamente n utenti nel sistema;

e le rispettive probabilità:

- $\pi_n^{(a)} = P\{A(n)\}$, ossia la probabilità che un utente in arrivo trovi esattamente n utenti nel sistema;

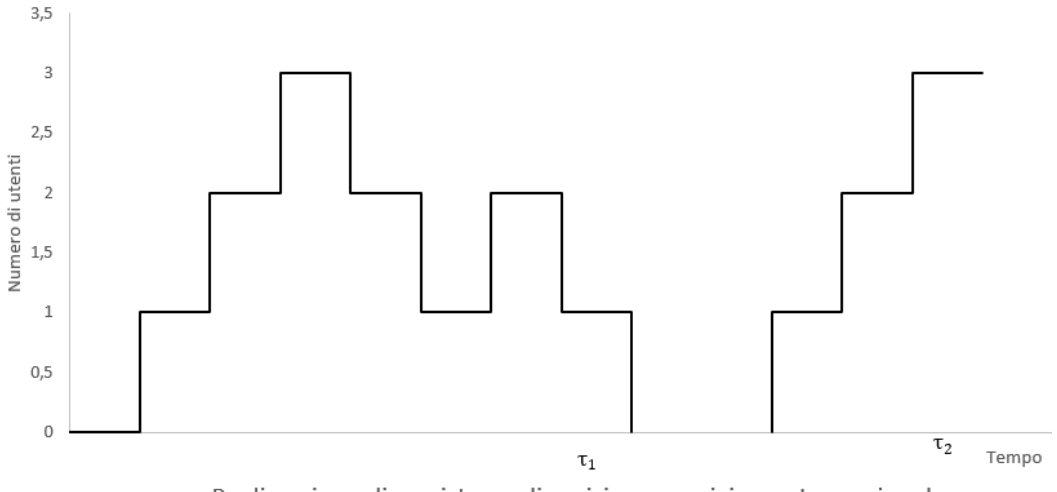
- $\pi_n^{(d)} = P\{D(n)\}$, ossia la probabilità che un utente in partenza dal sistema lasci esattamente n utenti nel sistema;

Le tre probabilità $q_n, \pi_n^{(a)}, \pi_n^{(d)}$ non sono sempre uguali per ogni $n = 0, 1, 2, \ldots$
Tuttavia sussiste il seguente risultato per sistemi di servizio in condizione di equilibrio statistico e in cui gli utenti arrivano e partono singolarmente dal sistema.
Teorema. In ogni sistema di servizio in cui gli utenti arrivano uno alla volta e gli utenti partono uno alla volta, nella situazione di equilibrio statistico si ha:

$$\pi_n^{(a)} = \pi_n^{(d)} \tag{3.7}$$

ossia la probabilità che un utente in arrivo nel sistema trovi esattamente n utenti nel sistema è uguale alla probabilità che un utente in partenza lasci esattamente n utenti nel sistema.

Dimostrazione
Si supponga che un utente in arrivo nel sistema trovi un numero di utenti pari a n. In tal caso poiché all'arrivo dell'utente il numero di utenti nel sistema è esattamente n è necessario aumentare il numero di utenti da n a $n+1$ in quanto l'arrivo dell'utente provoca un incremento di un'unità del numero di utenti.
Si supponga, inoltre, che un utente in partenza dal sistema lasci n utenti nel sistema. In tal caso poiché alla partenza dell'utente il numero di utenti nel sistema deve essere esattamente n è necessario diminuire il numero di utenti da $n+1$ a n in quanto la partenza dell'utente provoca un decremento di un'unità del numero di utenti.
Sia $n_A(\tau)$ il numero di transizioni dallo stato n allo stato $n+1$ nell'intervallo di tempo τ e sia $n_D(\tau)$ il numero di transizioni dallo stato $n+1$ allo stato n nell'intervallo di tempo τ. Si ha che nell'intervallo di tempo τ il numero di transizioni dallo stato n allo stato $n+1$ differisce al più di 1 dal numero di transizioni dallo stato $n+1$ allo stato n ossia $n_A(\tau) = n_D(\tau)$ oppure $n_A(\tau) = n_D(\tau) + 1$.
Si riporta di seguito un esempio di un grafico rappresentate un sistema di servizio in cui gli arrivi e le partenze avvengono singolarmente.
Si consideri l'intervallo di tempo τ_1, le transizioni dallo stato 1 allo stato 2 e le transizioni dallo stato 2 allo stato 1. In tal caso risulta che $n_A(\tau) = 2$ e $n_D(\tau) = 2$ ossia $n_A(\tau) = n_D(\tau)$.
Si consideri, invece, l'intervallo di tempo τ_2, le transizioni dallo stato 1 allo stato 2 e le transizioni dallo stato 2 allo stato 1. In tal caso risulta che $n_A(\tau) = 3$ e $n_D(\tau) = 2$ ossia $n_A(\tau) = n_D(\tau) + 1$.



Realizzazione di un sistema di sevizio con arrivi e partenze singole

Per ogni $n = 0, 1, 2, \ldots$ si ha che

$$\frac{n_A(\tau)}{\tau} = \frac{n_D(\tau)}{\tau}$$

oppure

$$\frac{n_A(\tau)}{\tau} = \frac{n_D(\tau)}{\tau} + \frac{1}{\tau}$$

Da cui si ottiene:

$$\lim_{\tau \to +\infty} \frac{n_A(\tau)}{\tau} = \lim_{\tau \to +\infty} \frac{n_D(\tau)}{\tau}$$

ossia per ogni $n = 0, 1, 2, 3, \ldots$ si ha che la frequenza con cui un utente in arrivo trova n utenti nel sistema è uguale alla frequenza con cui un utente in partenza lascia n utenti nel sistema. Ciò significa che in media gli utenti arrivano nel sistema allo stesso modo con cui gli utenti lasciano il sistema e quindi la probabilità che un utente in arrivo trovi n utenti nel sistema è uguale alla probabilità che un utente in partenza lasci n utenti nel sistema, ossia $\pi_n^{(a)} = \pi_n^{(d)}$.

C.V.D.

In condizione di equilibrio statistico la probabilità che un utenti in arrivo trovi n utenti nel sistema ($\pi_n^{(a)}$) è diversa dalla probabilità di avere n utenti nel sistema (q_n) ossia $\pi_n^{(a)} \neq q_n$. Un'importante eccezione, però, è data dal principio *Poisson Arrivals See Time Averages (principio PASTA).*

Teorema (Principio PASTA) In ogni sistema di servizio in cui gli arrivi degli utenti sono descritti da un processo di Poisson, nella condizione di equilibrio statistico la probabilità che ci siano n utenti nel sistema è uguale alla probabilità che un utente in arrivo trovi n utenti nel sistema, ossia:

$$q_n = \pi_n^{(a)} \quad (n = 0, 1, 2, \ldots)$$

Dimostrazione

Dalla (3.6) si ha che $q_n = \lim_{\tau \to +\infty} \frac{t_n(\tau)}{\tau}$ da cui si ottiene $t_n(\tau) \simeq q_n \tau$ per τ molto grande.
Sia λ la frequenza di arrivo indipendentemente dal numero di utenti nel sistema. Il numero medio di arrivi che trovano il sistema nello stato n nell'intervallo di tempo di lunghezza τ è $\lambda t_n(\tau) = \lambda q_n \tau$. Poiché per ipotesi gli arrivi sono descritti da un processo di Poisson risulta per la (3.1) che :

$$E[N(\tau)] = \lambda \tau$$

Da ciò si ottiene che:

$$\pi_n^{(a)} = \frac{\lambda \tau q_n}{\lambda \tau} = q_n$$

ossia la probabilità che ci siano n utenti nel sistema è uguale alla probabilità che un utente in arrivo trovi n utenti nel sistema.

C.V.D

Dal principio PASTA e dalla (3.7) segue che in ogni sistema di servizio in cui gli arrivi sono descritti da un processo di Poisson, in condizioni di equilibrio statistico $q_n = \pi_n^{(a)} = \pi_n^{(d)}$.
Si noti che quanto detto vale solo in sistemi di servizio con arrivi descritti da processi di Poisson in condizioni di equilibrio statistico, infatti, se si considera un sistema D/D/1 con queste caratteristiche:

- il sistema è vuoto al tempo t=0;
- gli arrivi si verificano agli istanti t=1, t=3, t=5 ...;
- il tempo di servizio è pari a 1.

In un sistema così definito ogni utente che arriva trova il sistema vuoto quindi la probabilità che un utente in arrivo trovi il sistema allo stato 0 è uguale a 1 mentre $q_n(t) = 1/2$ per ogni t.

3.5 Coefficiente di utilizzazione del sistema e throughput

Sistemi di servizio con singolo servitore
In un sistema di servizio con un solo servitore, nella condizioni di equilibrio statistico si denota con f il *coefficiente di utilizzazione del sistema* e si definisce come

$$f = E[N_s] = q_1 + q_2 + \ldots + q_n = P(N \geq 1) = 1 - P(N = 0)$$

Il *coefficiente di utilizzazione* in un sistema con singolo servitore indica la probabilità che ci sia almeno un utente nel sistema, ossia la probabilità che il centro di servizio sia occupato, e coincide con il numero medio di servitori occupati.
Sempre in condizioni di equilibrio statistico si indica con τ il *throughput* ovvero la capacità di trasporto o la capacità produttiva di un sistema di servizio con singolo servitore.
Il *throughput* indica la frequenza con cui gli utenti lasciano il sistema di servizio con successo ed è definito come prodotto tra la frequenza media di partenza e il numero medio di servitori occupati ossia:

$$\tau = \mu E[N_s] = \mu P(N \geq 1) = \mu[1 - P(N = 0)]$$

Sistemi di servizio con più servitori in parallelo
In un sistema di servizio con s servitori, nella condizione di equilibrio statistico si denota con f il *coefficiente di utilizzazione del sistema* e si definisce come

$$f = \frac{E[N_s]}{s}$$

dove il numero medio di servitori occupati è

$$E[N_s] = \sum_{n=0}^{s-1} nq_n + \sum_{n=s}^{\infty} nq_n$$

poiché quando il numero di utenti nel sistema è $n = 0, 1, 2, \ldots, s-1$ i servitori occupati saranno solamente n mentre quando $n > s - 1$ il numero di servitori occupati sarà sempre s.
In condizioni di equilibrio statistico il *throughput* si ottiene moltiplicando la frequenza media di arrivo per il numero medio di servitori occupati ossia:

$$\tau = \mu s f = \mu E[N_s] = \mu \left[\sum_{n=0}^{s-1} nq_n + \sum_{n=s}^{\infty} nq_n \right]$$

Il *throughput* quindi definisce la frequenza con cui gli utenti lasciano il sistema con successo in condizioni di equilibrio e con s servitori nel sistema.

3.6 Sistema di servizio M/M/1

Un sistema di servizio M/M/1 è costituito da un centro di attesa a capacità infinita e da un centro di servizio con un singolo servitore.
I tempi di interarrivo sono indipendenti ed identicamente distribuiti con funzione di distribuzione esponenziale e valore medio $\frac{1}{\lambda}$.
I tempi di servizio sono indipendenti ed identicamente distribuiti con funzione di distribuzione esponenziale e valore medio $\frac{1}{\mu}$.
La disciplina di servizio è FCFS (First come-first served).
Un sistema M/M/1 così definito può essere descritto utilizzando un processo di nascita morte caratterizzato dai seguenti parametri:

- $\lambda_n = \lambda$ ossia la frequenza di arrivo è sempre la stessa qualunque sia il numero di utenti nel sistema;
- $\mu_n = \mu$ ossia la frequenza di partenza è sempre la stessa qualunque sia il numero di utenti nel sistema.

Nota: le frequenze di arrivo e di partenza possono variare a seconda del tipo di sistema di servizio che si analizza. Ad esempio in un sistema M/M/1 con svendita in cui all'aumentare degli utenti nel sistema aumentano gli arrivi e anche la velocità con cui i servitori offrono il servizio i parametri saranno così definiti:

- $\lambda_n = (n+1)\lambda$ ossia la frequenza di arrivo varia all'aumentare del numero di utenti nel sistema;
- $\mu_n = n\mu$ ossia la frequenza di partenza varia all'aumentare del numero di utenti nel sistema.

Analisi del sistema M/M/1
L'analisi dei parametri prestazionali dal punto di vista teorico di un qualsiasi sistema di servizio è eseguita in condizione di equilibrio statistico per poter semplificare la trattazione matematica. Quindi il primo passo nell'analisi del sistema M/M/1 è stabilire sotto quali condizioni raggiunge una situazione di equilibrio statistico.
Si ricorda che per raggiungere la situazione di equilibrio statistico deve verificarsi la seguente condizione:

$$1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n} < +\infty$$

Nel sistema M/M/1 si ha che $1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n} = 1 + \sum_{n=1}^{\infty} \frac{\lambda^n}{\mu^n} = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = \sum_{n=0}^{\infty} \rho^n$ dove $\rho = \frac{\lambda}{\mu}$.

La $\sum_{n=0}^{\infty} \rho^n$ è una serie geometrica che converge a $(1-\rho)^{-1}$ solo se $\rho < 1$. Da ciò si ottiene che il sistema M/M/1 raggiunge l'equilibrio statistico quando $\rho < 1$ cioè quando la frequenza di arrivo è minore della frequenza di partenza.
In condizione di equilibrio statistico si possono calcolare la probabilità che ci siano 0 utenti nel sistema e la probabilità che ci siano n utenti nel sistema:

- $q_0 = \left(1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n}\right)^{-1} = 1 - \rho$;
- $q_n = q_0 \frac{\lambda_0 \dots \lambda_n - 1}{\mu_0 \dots \mu_n} = (1-\rho)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n$.

Da questi valori è possibile calcolare il numero medio di utenti e la varianza:

- $E[N] = \sum_{n=0}^{\infty} n q_n = \frac{\rho}{1-\rho}$;
- $Var[N] = \frac{\rho}{(1-\rho)^2}$.

In un sistema M/M/1 la frequenza media di arrivo per unità di tempo è $\lambda^* = \sum_{n=0}^{\infty} \lambda_n q_n = \lambda$ e la frequenza media di partenza per unità di tempo è $\mu^* = \frac{1}{1-q_0} \sum_{n=1}^{\infty} \mu_n q_n = \mu$.
In condizioni di equilibrio statistico le *Leggi di Little* costituiscono un potente mezzo per poter calcolare alcuni parametri prestazionali indipendentemente dal tipo di distribuzione dei tempi di interarrivo e di servizio, dal tipo di disciplina di servizio e dal numero di servitori nel sistema.
La prima legge di Little stabilisce che

$$E[N] = \lambda^* E[W]$$

ossia il numero medio di utenti nel sistema è uguale al prodotto tra la frequenza media di arrivo per unità di tempo e il tempo medio di attesa nel sistema.
La seconda legge di Little stabilisce che

$$E[N_q] = \lambda^* E[Q]$$

ossia il numero medio di utenti in coda è uguale al prodotto tra la frequenza media di arrivo per unità di tempo e il tempo medio di attesa in coda.
La terza legge di Little si ricava dalle due leggi precedenti

$$E[N_s] = E[N] - E[N_q] = \lambda^* E[S]$$

ossia il numero medio di utenti in servizio è uguale al prodotto tra la frequenza di arrivo e il tempo medio per servire un utente.
In un sistema M/M/1 di questo tipo il fattore di utilizzazione del sistema definito come $\frac{\lambda^*}{\mu^*}$ coincide con l'intensità di traffico del sistema definita come $\frac{\lambda^*}{s\mu^*}$ dove s indica il numero di servitori e in tal caso s=1.
Il coefficiente di utilizzazione f del sistema M/M/1 sarà:

$$f = E[N_s] = P(N \geq 1) = 1 - q_0 = \frac{\lambda}{\mu}$$

Il throughput τ del sistema M/M/1 sarà:

$$\tau = \mu E[N_s] = \mu P(N \geq 1) = \mu[1 - q_0)] = \lambda$$

ovvero in condizioni di equilibrio statistico la frequenza con cui gli utenti lasciano il sistema è uguale alla frequenza media con cui gli utenti arrivano.

Output del sistema M/M/1
Il problema dell'output di un sistema di servizio è stato considerato per la prima volta da Morse nel 1955. Egli osservò che il processo delle partenze da un sistema M/M/1 in condizioni di equilibrio è un processo di Poisson con lo stesso parametro del processo di Poisson di input. Quindi in un sistema con arrivi descritti da un processo di Poisson e tempi di servizio distribuiti esponenzialmente, anche il processo delle partenze dal sistema è un processo di Poisson, ovvero il processo delle partenze ha le stesse caratteristiche del processo degli arrivi, nonostante il servizio che è avvenuto e il relativo tempo. È necessario, inoltre, osservare che tale risultato è indipendente anche dalla disciplina della coda.
L'utilità di questo risultato all'interno della teoria delle code è evidente: se gli utenti che escono dal sistema M/M/1 entrano in un altro sistema a coda, in condizione di equilibrio statistico gli arrivi a questo secondo sistema saranno ancora descritti mediante un processo di Poisson. Quindi, assumendo che i tempi di servizio del secondo sistema siano distribuiti esponenzialmente, tale sistema si comporta come un sistema M/M/1 e può essere studiato indipendentemente dal primo sistema.

Teorema di Burke. In un sistema di servizio M/M/1 in condizioni di equilibrio statistico, le lunghezze degli intervalli di tempo tra le partenze (*tempi di interpartenze*) sono variabili aleatorie indipendenti ed esponenzialmente distribuite con valore medio $\frac{1}{\lambda}$ dove λ è il parametro del processo di Poisson di input. Quindi, il processo di output (delle partenze) di un sistema di servizio M/M/1 in condizioni di equilibrio statistico è un processo di Poisson di parametro λ.
Nota: il teorema di Burke può essere esteso anche a sistemi di servizio con più servitori M/M/s.

Capitolo 4

Reti di code

4.1 Introduzione

Nei capitoli precedenti abbiamo analizzato accuratamente le prestazioni di singoli sistemi di servizio calcolando una serie di parametri al fine di stabilire la qualità del sistema e le misure da adottare per ottenere un miglioramento delle prestazioni minimizzando i costi. Tuttavia, molto spesso i sistemi reali possono essere più complessi ed è quindi necessario descriverli utilizzando sistemi di servizio collegati tra loro, ossia *reti di code*.
La teoria delle reti di code riguarda la formulazione e l'analisi di modelli matematici atti a descrivere un insieme di sistemi di servizio (detti risorse o nodi della rete), ognuno costituito da un centro di attesa e da un centro di servizio. Si suppone che esista una struttura di interconnessione tra le risorse che consenta agli utenti di usufruire dei servizi offerti dalla rete passando da una risorsa all'altra.
Lo scopo è sempre quello di analizzare le prestazione della rete e di individuare, se necessario, idonee politiche atte a migliorarne le sue prestazioni.
Gli utenti in una rete di code possono essere job in computer-system, task in un sistema di elaborazione, pacchetti in un sistema di comunicazione, utenti di un centro commerciale ...
Ogni risorsa contiene un centro di attesa (buffer) e un centro di servizio (stazione) con uno o più servitori.
In una rete di code l'input (o gli input) di una risorsa può (possono) essere l'output (o gli output) di una o più risorse. Ad esempio, in Figura 4.1 è rappresentata una rete di code con due risorse collegate in serie in cui l'output della prima risorsa costituisce l'input della seconda risorsa.

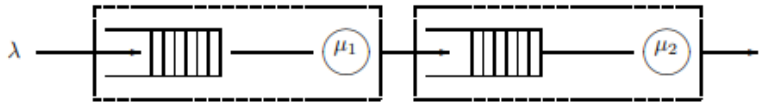


Figura 4.1 Rappresentazione di una rete di code con due risorse

Lo studio di reti di code è più complesso rispetto all'analisi di singoli sistemi di servizio in quanto è necessario tener conto di tutti i sistemi che compongono la rete per poter analizzare il flusso degli utenti.

4.2 Classificazione delle reti di code

Le reti di code possono essere classificate in diversi modi. Una classificazione delle reti di code si basa sulla possibilità di accettare utenti provenienti dall'esterno e distingue le reti in *aperte*, *chiuse* e *miste*.
Una rete di code *aperta* è caratterizzata da una o più sorgenti da cui gli utenti possono arrivare e da una o più destinazioni in cui confluiscono gli utenti quando abbandonano la rete.
In una rete di code *aperta* gli utenti possono arrivare dall'esterno e uscire all'esterno. Per tale motivo gli utenti presenti nella rete variano nel tempo, infatti, in ogni momento un utente dall'esterno può entrare nella rete oppure può uscire dalla rete.
Un utente che entra nella rete usufruisce delle varie risorse offerte passando da un sistema di servizio all'altro fino a quando non esce dalla rete.
La scelta da parte dell'utente della risorsa di cui usufruire può essere di due tipi:

- *deterministica*, ossia gli utenti che entrano nella rete attraversano le risorse seguendo un ordine prestabilito, quindi l'utente dopo aver usufruito del servizio accede ad una certa risorsa già prefissata. Questo tipo di scelta è utile nella descrizione di catene di montaggio in cui, ad esempio, un certo pezzo per essere prodotto deve attraversare risorse prefissate ognuna con un centro di servizio che esegue un'operazione (taglio, lavaggio,...);

- *aleatoria*, ossia gli utenti che entrano nella rete usufruiscono delle risorse senza seguire un ordine prestabilito, ma seguendo criteri probabilistici. Questo tipo di scelta è utile, ad esempio, quando l'instradamento verso una risorsa di un certo pezzo dipende dall'esito di un'operazione: se il pezzo risulta difettoso deve essere spedito ad una precedente risorsa per essere nuovamente lavorato. In generale è possibile stimare con una certa accuratezza la probabilità con cui questo accade.

Nelle reti di code *aperte* non c'è limite al numero di utenti presenti in un determinato istante di tempo.
Una rete di code *chiusa* è caratterizzata da un numero fissato e finito di utenti. In tale rete non possono verificarsi né arrivi di utenti dall'esterno né uscite di utenti verso l'esterno, quindi il numero di utenti non cambia nel corso del tempo e gli utenti circolano indefinitamente tra le risorse della rete. Tuttavia, nella realtà spesso avvengono sia ingressi sia uscite, nel senso che non appena un utente esce dal sistema è istantaneamente sostituito da un nuovo utente che entra nella rete.
Una rete di code *mista* è aperta per alcuni utenti e chiusa per altri. In tale rete possono verificarsi ingressi ed uscite dal sistema di alcuni tipi di utenti ed altri utenti, invece, circolano indefinitamente nella rete.
In reti di code aperte e miste gli utenti entrano dall'esterno in una qualsiasi delle risorse e una volta usufruito del servizio o si dirigono verso un'altra risorsa o entrano di nuovo nella stessa risorsa oppure escono dalla rete. In reti di code chiuse, invece, gli utenti o si trovano in una fila di attesa o in un centro di servizio e circolano indefinitamente nella rete usufruendo della varie risorse messe a disposizione.
Per le reti di code aperte e miste è necessario definire le caratteristiche delle sorgenti di input, ossia stabilire i potenziali utenti che possono accedere alla rete. Per le reti chiuse, invece, non è necessario definire la sorgente in quanto non possono verificarsi arrivi dall'esterno.
Per ogni rete di code è necessario definire le seguenti caratteristiche:

- il numero di risorse disponibili;

- la dimensione (nulla, finita, infinita) della fila di attesa di ogni risorsa;

- il processo degli arrivi di ogni risorsa, ossia definire per ogni risorsa la distribuzione dei tempi di interarrivo;
- il numero di servitori in ogni centro di servizio;
- la disciplina di servizio per ogni risorsa;
- la distribuzione dei tempi di servizio per ogni servitore;
- il processo di output da ogni risorsa verso le altre risorse o eventualmente verso l'esterno della rete.

Se la rete di code è chiusa, è necessario definire anche il numero di utenti che circolano nella rete. Se, invece, la rete è mista, è necessario definire il sottoinsieme di utenti per cui la rete è chiusa, ossia gli utenti che circolano indefinitamente nella rete di code.
Per alcune reti di code, infine, l'insieme di utenti è suddiviso in opportune classi contenenti utenti caratterizzati da caratteristiche simili e pertanto occorre classificare gli utenti in classi di priorità.

4.3 Struttura di interconnessione

Gli utenti all'interno di una rete di code passano da una risorsa all'altra usufruendo dei vari servizi messi a disposizione, per tale motivo è necessario definire una struttura, detta *struttura di interconnessione*, che rappresenta le possibili transizioni che gli utenti possono effettuare tra le risorse all'interno della rete.
Per poter definire una struttura di interconnessione di tipo probabilistico per una rete di code con R risorse è necessario costruire una matrice delle probabilità di dimensione $R \times (R+1)$:

$$D = \begin{pmatrix} p_{11} & \cdots & p_{1R} & p_{10} \\ \vdots & \ddots & \vdots & \vdots \\ p_{R1} & \cdots & p_{RR} & p_{R0} \end{pmatrix} \text{ffl}$$

dove p_{ij} $(i, j = 1, 2, \ldots, R)$ denota la probabilità che l'utente si diriga istantaneamente alla risorsa j dopo aver completato il servizio alla risorsa i-esima e p_{i0} $(i = 1, 2 \ldots, R)$ denota la probabilità che un utente esce istantaneamente dalla rete non appena ha completato il servizio nella risorsa i-esima.
Poiché risulta

$$p_{ij} \geq 0 \;\; (i, j = 1, 2, \ldots, R) \qquad p_{i0} \geq 0 \;\; (i = 1, 2 \ldots, R)$$

$$\sum_{j=1}^{R} p_{ij} + p_{i0} = 1 \quad (i = 1, 2, \ldots, R)$$

la matrice D è *stocastica*, ovvero la somma delle probabilità sulle righe è unitaria.
Dalla matrice D, eliminando l'ultima colonna relativa alle probabilità che un utente nella risorsa i-esima esca dalla rete, è possibile ricavare anche la matrice P di dimensione $R \times R$. La matrice P così costruita prende il nome di *matrice delle probabilità di switching* o *di instradamento*. La matrice P, a differenza della matrice D, è stocastica solo nel caso in cui la rete di code è chiusa in quanto la probabilità che un utente esca dalla rete è 0.
Una rete di code può essere rappresentata utilizzando un grafo in cui i nodi identificano le risorse della rete e gli archi che collegano le varie risorse indicano le possibili transizioni che gli utenti possono effettuare nella rete. In generale, si può pensare che ad una rete di code arrivino utenti dall'esterno in ciascun nodo della rete, così come da ciascun nodo un utente può lasciare la

rete. Ovvero, un utente entra nella rete in un certo nodo, attraversa alcuni nodi della rete e poi presso un nodo lascia la rete. Può anche accadere che gli utenti possano ritornare presso nodi già visitati.

4.4 Descrizione di una rete di code

Si consideri una rete di code costituita da R risorse numerate da 1 a R. Per ogni $i = 1, 2, \ldots, R$ sia $\{N_i(t), t \geq 0\}$ un processo stocastico unidimensionale, dove $N_i(t)$ indica il numero di utenti presenti nell'i-esima risorsa al tempo t. Una rete di code con R risorse può essere modellata con un processo stocastico R-dimensionale $\{N(t), t \geq 0\}$ con $N(t) = (N_1(t), N_2(t), \ldots, N_R(t))$. In ogni istante di tempo t la rete può trovarsi in uno stato specificato da una R-upla $(n_1, n_2, \ldots, n_R)$ dove n_i $(i = 1, 2, \ldots, R)$ rappresenta il numero di utenti presenti nella i-esima risorsa al tempo t. Inoltre, se la rete è chiusa e contiene K utenti è verificata la condizione $n_1 + n_2 + \ldots + n_R = K$. Si denoti la probabilità che al tempo t ci siano n_1 utenti nella risorsa 1, n_2 utenti nella risorsa 2, ..., n_R utenti nella risorsa R-esima come segue:

$$p(n_1, n_2, \ldots, n_R, t) = P\{N_1(t) = n_1, N_2(t) = n_2, \ldots, N_R(t) = n_R\}$$
$$(n_i = 0, 1, 2, \ldots; i = 1, 2, \ldots, R) \tag{4.1}$$

Da tali probabilità congiunte è possibile determinare le probabilità marginali

$$p_i(k, t) = P\{N_i(t) = k\} \quad (i = 1, 2, \ldots, R; k = 0, 1, 2, \ldots) \tag{4.2}$$

ossia la probabilità che al tempo t ci sono k utenti nella risorsa i-esima della rete.
Così come per singoli sistemi di servizio risulta difficile nella fase transiente calcolare le probabilità che al tempo t ci siano k utenti nel sistema, allo stesso modo è molto difficile calcolare nella fase transiente le probabilità congiunte (4.1) e marginali (4.2) nelle reti di code. Per tale motivo anche in questo caso si analizzano le reti di code in condizioni di equilibrio statistico.
Sia N_i la variabile aleatoria che descrive il numero di utenti nella risorsa $i-$esima in condizioni di equilibrio statistico.
Si denoti la probabilità che in condizioni di equilibrio statistico ci siano n_1 utenti nella risorsa 1, n_2 utenti nella risorsa 2, ..., n_R utenti nella risorsa R come segue

$$q(n_1, n_2, \ldots, n_R) = P\{N_1 = n_1, N_2 = n_2, \ldots, N_R = n_R\} = \lim_{t \to \infty} p(n_1, n_2, \ldots, n_R, t)$$
$$(n_i = 0, 1, 2, \ldots; i = 1, 2, \ldots, R)$$

Se tali limiti esistono e non dipendono dalle condizioni iniziali la rete raggiunge una situazione di equilibrio statistico e

$$q_i(k) = \lim_{t \to \infty} P\{N_i(t) = k\} \quad (k = 0, 1, \ldots)$$

denota la probabilità che siano presenti k utenti nella i-esima risorsa della rete in condizioni di equilibrio statistico. In particolare, se le variabili aleatorie $N_1, N_2, \ldots, N_R$ sono indipendenti, allora per ogni R-upla $(n_1, n_2, \ldots, n_R)$ la probabilità congiunta si fattorizza come il prodotto delle singole probabilità marginali, ovvero

$$q(n_1, n_2, \ldots, n_R) = q_1(n_1) q_2(n_2) \ldots q_R(n_R)$$

In una rete chiusa, invece, poiché le variabili aleatorie $N_1, N_2, \ldots, N_R$ non sono indipendenti la probabilità congiunta non si fattorizza come il prodotto delle singole probabilità marginali.
Per reti chiuse il numero di utenti è fissato

$$K = N_1 + N_2 + \ldots + N_R$$

mentre nel caso di reti aperte o miste il numero totale di utenti nella rete in condizioni di equilibrio statistico è descritto da una variabile aleatoria

$$N = N_1 + N_2 + \ldots + N_R$$

Da quest'ultima è possibile calcolare il numero medio di utenti nella rete e la varianza

$$E[N] = E[N_1 + N_2 + \ldots + N_R] = E[N_1] + E[N_2] + \ldots + E[N_R]$$

$$Var[N] = Var[N_1 + N_2 + \ldots + N_R] = \sum_{i=1}^{R} Var(N_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{R} Cov(N_i, N_j)$$

dove $Cov(N_i, N_j)$ indica la covarianza tra N_i e N_j.
In particolare, se le variabili aleatorie $N_1, N_2, \ldots, N_R$ sono indipendenti, la varianza del numero di utenti in una rete aperta o mista è uguale alla somma delle varianze del numero di utenti presenti nelle diverse risorse, ovvero

$$Var[N] = Var[N_1 + N_2 + \ldots + N_R] = Var[N_1] + Var[N_2] + \ldots + Var[N_R]$$

4.5 Leggi di Little locali e globali

In condizioni di equilibrio statistico siano:

- λ_i la frequenza media per unità di tempo degli arrivi nella i-esima risorsa dall'esterno della rete;
- α_i la frequenza media complessiva per unità di tempo degli arrivi nella i-esima risorsa considerando sia gli arrivi da altre risorse della rete sia gli arrivi provenienti dall'esterno;
- μ_i la frequenza media per unità di tempo delle partenze da un generico servitore nella i-esima risorsa.

L'*intensità di traffico* della risorsa i-esima sarà

$$a_i = \frac{\alpha_i}{\mu_i} \quad (i = 1, 2, \ldots, R)$$

che rappresenta l'intensità del lavoro svolto dalla risorsa i-esima in condizioni di equilibrio statistico.
Il *fattore di utilizzazione* della risorsa i-esima ρ_i è definito come il rapporto tra l'intensità di traffico e il numero di servitori

$$\rho_i = \frac{a_i}{s_i} = \frac{\alpha_i}{s_i \mu_i} \quad (i = 1, 2, \ldots, R)$$

che rappresenta l'intensità del lavoro di un generico servitore della i-esima risorsa nella situazione di equilibrio statistico.
Affinché il numero di utenti nella fila di attesa della i-esima risorsa non aumenti indefinitamente occorre che $\rho_i < 1$. Quindi, per una risorsa con capacità infinita, ρ_i può essere interpretato come una misura di congestione della risorsa i-esima della rete. Pertanto, nella situazione di equilibrio statistico, una condizione necessaria affinché la rete di code aperta o mista non sia congestionata è che $\rho_i < 1$ per ogni $i = 1, 2, \ldots, R$.
Sia W_i il tempo di attesa di un utente nella risorsa i-esima in condizioni di equilibrio statistico. Applicando la prima *legge di Little* alla risorsa i-esima si ottiene la *legge di Little locale*

$$E[N_i] = \alpha_i E[W_i] \quad (i = 1, 2, \ldots, R)$$

ossia il numero medio di utenti nella risorsa i-esima è dato dal prodotto tra la frequenza media complessiva per unità di tempo degli arrivi nella i-esima risorsa e dal tempo di attesa medio di un utente nella risorsa i-esima. Tale legge è chiamata locale poiché si riferisce ad una singola risorsa della rete.
In una rete aperta o mista, per quanto detto precedentemente, si ha che

$$E[N] = E[N_1 + N_2 + \ldots + N_R] = E[N_1] + E[N_2] + \ldots + E[N_R]$$

da cui si ottiene che

$$E[N] = \sum_{i=1}^{R} \alpha_i E[W_i]$$

essendo $E[N_i] = \alpha_i E[W_i]$ per la *legge locale di Little*.
Se la rete di code, inoltre, è aperta o mista, in condizioni di equilibrio statistico, la *legge di Little globale* per l'intera rete di code afferma che

$$E[N] = \Big[\sum_{i=1}^{R} \lambda_i\Big] E[W]$$

ovvero il numero medio di utenti nella rete è dato dal prodotto tra la somma delle frequenze di arrivo degli utenti dall'esterno nella risorsa i-esima e dal tempo di attesa medio degli utente nella rete.
Dalla legge di Little locale e globale, in reti aperte o miste, si può ricavare il tempo medio di attesa nella rete in condizioni di equilibrio statistico:

$$E[W] = \frac{E[N]}{\sum_{i=1}^{R} \lambda_i} = \frac{\sum_{i=1}^{R} \alpha_i E[W_i]}{\sum_{i=1}^{R} \lambda_i}$$

Capitolo 5

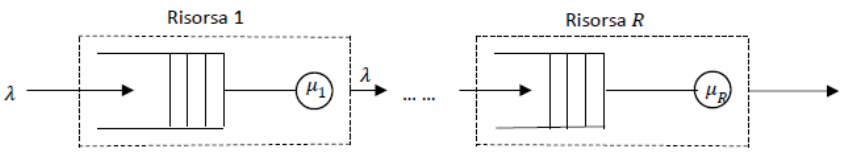
Reti tandem

5.1 Introduzione

Una rete di code aperta è detta *aciclica* se il grafo orientato associato alle risorse della rete è aciclico. In una rete di code di questo tipo gli utenti non possono usufruire più di una volta del servizio offerto da una risorsa, inoltre, alla prima risorsa è possibile accedere solo dall'esterno, mentre alle successive è possibile accedere sia dall'esterno che dalla precedente risorsa.
Le più semplici reti di code acicliche sono le reti tandem, introdotte da Jackson nel 1954, costituite da un insieme finito di sistemi di servizio collegati in serie. Le risorse di una rete tandem sono quindi organizzate in sequenza. Ogni risorsa è dotata di un centro di attesa e di un centro di servizio con uno o più servitori identici che lavorano in parallelo. Ogni utente che entra nella rete attraversa le singole risorse in sequenza ed è servito in ognuno dei centri di servizio delle varie risorse. Se il centro di servizio di una risorsa è occupato, ossia se tutti i servitori sono occupati, l'utente si accoda nel centro di attesa di quella risorsa.
In una rete tandem si assume che il centro di attesa di ogni singola risorsa sia a capacità infinita in maniera tale da contenere un numero qualsiasi di utenti in attesa.
Si suppone, inoltre, che la prima risorsa può ricevere arrivi soltanto dall'esterno, mentre nelle successive risorse il processo di input corrisponde al processo di output della risorsa precedente.
Lo scopo di questo capitolo è proprio l'analisi dei parametri prestazionali di reti tandem.

5.2 Reti tandem con R risorse

Si consideri una *rete tandem* con R risorse come rappresentato nella Figura 5.1.



Alla prima risorsa gli utenti arrivano secondo un processo di Poisson di parametro λ ed i tempi di servizio relativi alla i-esima risorsa sono indipendenti e distribuiti esponenzialmente con valore medio $1/\mu_i$ $(i = 1, 2, \ldots, R)$.
Dal teorema di Burke si evince che nella situazione di equilibrio statistico il processo delle partenze da ogni risorsa è un processo di Poisson di parametro λ. In condizioni di equilibrio statistico, quindi, l'input ad ogni risorsa della rete tandem è un processo di Poisson di parametro λ. La rete tandem con R risorse ognuna con un unico servitore può quindi essere rappresentata

con la notazione M/M/1 $\rightarrow$ M/M/1 $\rightarrow \ldots \rightarrow$ M/M/1, ossia ogni singola risorsa della rete può essere descritta con un sistema di servizio M/M/1.
Il sistema M/M/1 relativo alla i-esima risorsa è caratterizzato da intensità di traffico $\rho_i = \lambda/\mu_i$ e raggiunge una situazione di equilibrio statistico se $\rho_i < 1$.
Si ricordi che in condizioni di equilibrio statistico la probabilità di avere k utenti nella i-esima risorsa è

$$q_i(k) = P(N_i = k) = (1 - \rho_i)\rho_i^k \quad (k = 0, 1, 2, \ldots)$$

Per raggiungere la condizione di equilibrio statistico è necessario che tutte le risorse della rete non siano congestionate, ossia $\rho_i < 1$ per ogni $i = 1, 2, \ldots, R$.
Nella situazione di equilibrio statistico, la probabilità che nella rete siano presenti n_1 utenti nella prima risorsa, n_2 utenti nella seconda risorsa, ..., n_R utenti nella R-esima risorsa è

$$q(n_1, n_2, \ldots, n_R) = q_1(n_1)q_2(n_2)\ldots q_R(n_R) = (1 - \rho_1)\rho_1^{n_1}(1 - \rho_2)\rho_2^{n_2}\ldots(1 - \rho_R)\rho_R^{n_R}$$

ovvero la probabilità congiunta è il prodotto delle probabilità marginali. Quindi, nella situazione di equilibrio statistico le variabili aleatorie $N_1, N_2, \ldots, N_R$ sono indipendenti e le R risorse possono essere considerate come R sistemi M/M/1 indipendenti.
Nella situazione di equilibrio statistico il numero medio di utenti presenti nella risorsa i-esima è

$$E[N_i] = \frac{\rho_i}{1 - \rho_i} = \frac{\lambda}{\mu_i - \lambda} \quad (i = 1, 2, \ldots, R)$$

Sia N la variabile aleatoria che descrive il numero complessivo di utenti nella rete, in condizioni di equilibrio statistico, il numero medio di utenti nella rete è

$$E[N] = \sum_{i=1}^{R} E[N_i] = \sum_{i=1}^{R} \frac{\rho_i}{1 - \rho_i} = \lambda \sum_{i=1}^{R} \frac{1}{\mu_i - \lambda}$$

Dalla legge di Little locale, sempre in condizioni di equilibrio statistico, segue che il tempo medio di attesa nella risorsa è

$$E[W_i] = \frac{1}{\lambda}E[N_i] = \frac{1}{\mu_i - \lambda} \quad (i = 1, 2, \ldots, R)$$

Poiché ogni risorsa in una rete tandem è visitata una sola volta, in condizioni di equilibrio statistico, il tempo di attesa nella rete è

$$E[W_i] = \sum_{i=1}^{R} E[W_i] = \sum_{i=1}^{R} \frac{1}{\mu_i - \lambda}$$

La legge di Little globale per la rete di code tandem con R risorse, ognuna con singolo servitore, diventa

$$E[N] = \lambda \sum_{i=1}^{R} E[W_i] = \lambda E[W]$$

Se $\mu_1 = \mu_2 = \ldots = \mu_R$, ovvero $\rho_1 = \rho_2 = \ldots = \rho_r$, allora è possibile calcolare la distribuzione di probabilità della variabile aleatoria N.
Sia

$$D = \{(n_1, n_2, \ldots, n_R) | n_1 + n_2 + \ldots + n_R = n\}$$

si ha che

$$P(N=n) = P(N_1 + N_2 + \ldots + N_R = n) = \sum_D q(n_1, n_2, \ldots, n_R) = (1-\rho)^R \sum_D \rho^{n_1+n_2+\ldots+n_R}$$

$$= (1-\rho)^R \rho^n \sum_D 1$$

dove l'ultima sommatoria è la cardinalità dell'insieme D e, facendo ricorso al calcolo combinatorio, si può interpretare come il numero di modi di disporre n oggetti tra loro indistinguibili in R urne, ovvero come

$$\binom{n+R-1}{R-1}$$

e pertanto il numero complessivo di utenti nella rete è:

$$P(N=n) = \binom{n+R-1}{R-1}(1-\rho)^R \rho^n \quad (n=0,1,\ldots)$$

5.3 Modelli di rete di code in forma prodotto

Si consideri una rete tandem costituita da un numero fissato $R = r$ di risorse.
Le reti tandem di cui si desidera analizzare i parametri prestazionali hanno le seguenti caratteristiche:

1. un numero fissato r di risorse collegate in serie senza feedback (senza cicli), ciascuna con un'unica fila di attesa avente capacità illimitata;

2. singola classe di utenti e disciplina di servizio FIFO per ciascuna risorsa della rete;

3. la risorsa i-esima è costituita da un unico servitore caratterizzato da un tasso di servizio dipendente dallo stato della risorsa. Si indica, pertanto, con $\mu_i(n_i)$ la frequenza media di servizio del servitore della risorsa i-esima quando in essa sono presenti n_i utenti;

4. gli utenti dopo aver usufruito del servizio messo a disposizione dalla risorsa i-esima procedono istantaneamente nella $(i+1)$-esima risorsa $(i = 1, 2, \ldots, r-1)$;

5. gli utenti accedono alla prima risorsa dall'esterno della rete secondo un processo di Poisson di parametro λ quindi i tempi di interarrivo in tale risorsa sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$;

6. gli utenti dopo aver ricevuto servizio nella risorsa r-esima escono istantaneamente dalla rete.

Teorema di Jackson. Nelle ipotesi 1-6 sia $(n_1, n_2, \ldots, n_r)$ lo stato di una rete tandem con r risorse nella quale sono presenti n_i utenti nella i-esima risorsa con $(i = 1, 2, \ldots, r)$ e sia $q(n_1, n_2, \ldots, n_r)$ la probabilità che, in condizioni di equilibrio statistico, ci siano n_1 utenti nella risorsa 1, n_2 utenti nella risorsa 2, ..., n_r utenti nella risorsa r, ossia la probabilità che la rete si trovi nello stato $(n_1, n_2, \ldots, n_r)$ in condizioni di equilibrio statistico.
Se per ogni risorsa è verificata la condizione di equilibrio, ossia

$$1 + \sum_{n=1}^{\infty} \frac{\lambda^n}{\mu_i(1)\mu_i(2)\ldots\mu_i(n)} < \infty \quad (i = 1, 2, \ldots, r)$$

allora la probabilità $q(n_1, n_2, \ldots, n_r)$ assume la forma prodotto

$$q(n_1, n_2, \ldots, n_r) = \prod_{i=1}^{r} q_i(n_i) = q_1(n_1)q_2(n_2)\ldots q_r(n_r)$$

dove le probabilità marginali $q_i(m)$ di avere m utenti nella risorsa i-esima sono

$$q_i(0) = \left[1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{\mu_i(1)\mu_i(2)\ldots\mu_i(k)}\right]^{-1} \quad (i = 1, 2, \ldots, r; k = 1, 2, \ldots)$$

e

$$q_i(n) = q_i(0)\frac{\lambda^n}{\mu_i(1)\mu_i(2)\ldots\mu_i(n)} \quad (i = 1, 2, \ldots, r; n = 1, 2, \ldots)$$

Nel suo campo di applicazione, ovvero le reti di code aperte, il teorema di Jackson detiene un'importanza rilevante. Nel modello di Jackson, purché sia soddisfatta la condizione di equilibrio statistico, la distribuzione di probabilità congiunta è organizzata secondo una struttura ben precisa. Tale struttura, detta *struttura prodotto*, prevede che la distribuzione di probabilità congiunta $q(n_1, n_2, \ldots, n_r)$ risulti dal prodotto di distribuzioni di probabilità marginali $q_i(n_i)$. Questo risultato mostra che in condizioni di equilibrio statistico le singole risorse sono indipendenti. Determinare la probabilità congiunta permette quindi di calcolare a sua volta i parametri prestazionali relativi al comportamento della rete di code tandem aperta.
Nel teorema di Jackson per le reti tandem si assume che gli arrivi alla prima risorsa si verifichino secondo un processo di Poisson di frequenza λ e nella situazione di equilibrio anche la frequenza di partenza da ogni risorsa è uguale a λ indipendentemente dalle frequenza di servizio. Pertanto la frequenza di entrata nella successiva risorsa è sempre λ.

5.4 Modello esteso di tipo binomiale negativo

Si consideri una rete di Jakson costituita da $C_1, C_2, \ldots, C_r (r \leq m)$ risorse, in cui ogni C_i è modellata mediante processi di nascita morte con tassi di servizio dipendenti dallo stato del sistema, ossia la frequenza di servizio del servitore i-esimo dipende dallo stato della i-esima risorsa.
Si indichi con N_i la variabile aleatoria che, in condizioni di equilibrio statistico, rappresenta il numero di utenti nella risorsa $C_i (i = 1, 2, \ldots, r)$.
Un particolare modello per la generica risorsa i-esima della rete tandem è il *modello binomiale negativo* di cui, in seguito, ne sono riportate le caratteristiche fondamentali.

Modello esteso di tipo binomiale negativo per la risorsa i-esima
Gli arrivi alla risorsa i-esima sono definiti da un processo di Poisson di parametro λ e i tempi di servizio della risorsa C_i sono:

$$\mu_i(n) = \frac{\mu_i n}{\beta_i + n} \quad (n = 1, 2, \ldots; i = 1, 2, \ldots, r)$$

dove $\beta_i \geq 0$ e $\mu_i > 0$.
Si noti che $\mu_i(n) \leq \mu_i$ e $\lim_{n\to\infty} \mu_i(n) = \mu_i$.
Se il valore $\beta_i = 0$, inoltre, si ha che la risorsa i-esima è descritta da un sistema di servizio M/M/1 con frequenza di arrivo λ e frequenza di partenza $\mu_i (i = 1, 2, \ldots, r)$.
Per il modello esteso di tipo binomiale negativo si ha una condizione di equilibrio statistico per la risorsa C_i se $\rho_i = \lambda/\mu_i < 1 (i = 1, 2, \ldots, r)$ e la probabilità che ci siano n utenti in tale risorsa è

$$q_i(n) = P(N_i = n) = \frac{\rho_i^n}{n!}(\beta_i + 1)_n (1 - \rho_i)^{\beta_i + 1} \quad (n = 0, 1, \ldots) \tag{5.1}$$

dove $(\gamma)_n$ denota il *simbolo di Pochhammer* definito come segue:

$$(\gamma)_0 = 1$$

e

$$(\gamma)_n = \gamma(\gamma + 1) \ldots (\gamma + n - 1) \quad (n = 1, 2, \ldots)$$

Si osserva che:

- se $\beta_i = 0$, per quanto detto precedentemente, la risorsa C_i è un sistema di servizio M/M/1 e quindi in tale sistema la probabilità che ci siano n utenti in condizioni di equilibrio statistico corrisponde esattamente alla distribuzione geometrica del sistema di servizio M/M/1 in situazione di equilibrio, ossia

$$q_i(n) = (1 - \rho_i)\rho_i^n \quad (n = 0, 1, \ldots)$$

- Se $\beta_i = k - 1 (k = 2, 3, \ldots)$ si ha, invece, che

$$q_i(n) = \binom{n + k - 1}{k - 1} (1 - \rho_i)^k \rho_i^n \quad (n = 0, 1 \ldots)$$

La distribuzione di probabilità ottenuta è una generalizzazione della distribuzione binomiale negativa,

$$E[N_i] = \frac{\rho_i}{1 - \rho_i}(1 + \beta_i) \tag{5.2}$$

e

$$Var[N_i] = \frac{\rho_i}{(1-\rho_i)^2}(1+\beta_i) \tag{5.3}$$

che è una distribuzione binomiale negativa la quale descrive il numero di fallimenti n precedenti al k-esimo successo in una sequenza di lanci indipendenti di una moneta con probabilità di successo $1-\rho_i$ in ogni lancio.
La variabile aleatoria binomiale negativa può essere vista come la somma di k variabili aleatorie geometriche.
Per $\beta_i = 0$, invece, si ha

$$E[N_i] = \frac{\rho_i}{1-\rho_i}$$

e

$$Var[N_i] = \frac{\rho_i}{(1-\rho_i)^2}$$

Differenze tra il modello geometrico e il modello esteso di tipo binomiale negativo per la risorsa i-esima

Sia

- N^G la variabile aleatoria che descrive il numero di utenti in un sistema caratterizzato da distribuzione geometrica;
- N^B la variabile aleatoria che descrive il numero di utenti in un sistema caratterizzato da distribuzione binomiale negativa.

Si dimostra che sussistono le seguenti disuguaglianze:

$$E(N^G) \leq E(N^B)$$

e

$$Var(N^G) \leq Var(N^B)$$

Queste relazioni mostrano che il numero medio di utenti e la varianza sono inferiori nel modello geometrico rispetto al modello esteso di tipo binomiale negativo.
Lo scopo del successivo capitolo sarà proprio l'analisi delle differenze tra il modello geometrico e il modello binomiale negativo al variare dei parametri prestazionali.

5.5 Reti tandem con numero fissato di risorse

Si consideri una rete tandem che soddisfa le condizioni 1-6 e costituita da un numero fissato r ($r \leq m$) di risorse in cui ogni centro di servizio ha un unico servitore.
Sia N_i la variabile aleatoria che descrive il numero di utenti nella risorsa i-esima ($i = 1, 2, \ldots, r$) in condizioni di equilibrio statistico e sia

$$q_i(n) = P(N_i = n) \quad (n = 0, 1, \ldots)$$

la probabilità che, in condizioni di equilibrio statistico, ci siano n utenti nella risorsa i-esima.
Sia, inoltre, $(N_1, N_2, \ldots, N_r)$ il vettore aleatorio che rappresenta il numero di utenti nelle risorse della rete tandem e sia

$$q(n_1, n_2, \ldots, n_r) = P(N_1 = n_1, N_2 = n_2, \ldots, N_r = n_r)$$

la probabilità che, in condizioni di equilibrio statistico, siano presenti n_1 utenti nella risorsa 1, n_2 utente nella risorsa 2, ..., n_r utenti nella risorsa r.
Sia $M_r = N_1 + N_2 + \ldots + N_r$ il numero totale di utenti nella rete si ha che

$$P(M_r = n) = \sum_{n \in D} q_1(n_1) q_2(n_2) \ldots q_r(n_r) \quad (n = 0, 1, \ldots)$$

con

$$D = \{n = (n_1, n_2, \ldots, n_r) | n_i \geq 0 (i = 1, 2, \ldots, r), \sum_{i=1}^{r} n_i = n\}$$

Le notazioni sopra citate sono di fondamentale importanza per poter effettuare l'analisi di reti tandem con risorse caratterizzate da distribuzione binomiale negativa generalizzata (5.1).

5.6 Reti tandem con distribuzione binomiale negativa

Si consideri una rete tandem con r risorse caratterizzate da distribuzione binomiale negativa generalizzata (5.1).
Si supponga, inoltre, che $\max(\rho_1, \rho_2, \ldots, \rho_r) < 1$, ossia la rete tandem in esame non si congestiona, ed è quindi soddisfatta la condizione di equilibrio statistico. In una rete sì fatta, dalle formule (5.1) e (5.2), è possibile ricavare il valore medio e la varianza del numero complessivo di utenti presenti nella rete:

$$E(M_r) = \sum_{i=1}^{r} E(N_i) = \sum_{i=1}^{r} \frac{\rho_i}{1 - \rho_i} (1 + \beta_i)$$

e

$$Var(M_r) = \sum_{i=1}^{r} Var(N_i) = \sum_{i=1}^{r} \frac{\rho_i}{(1 - \rho_i)^2} (1 + \beta_i)$$

Se $\mu_1 = \mu_2 = \ldots = \mu_r$ è possibile calcolare la probabilità che il numero totale di utenti nella rete $M_r = N_1 + N_2 + \ldots + N_r$ sia esattamente n in condizioni di equilibrio statistico

$$P(M_r = n) = \frac{(\beta_1 + \beta_2 + \ldots + \beta_r + r)_n}{n!} (1 - \rho)^{\beta_1 + \beta_2 + \ldots + \beta_r + r} \rho^n \tag{5.4}$$

Capitolo 6

Analisi di reti tandem con distribuzione binomiale negativa al variare dei parametri

6.1 Introduzione

Fino ad ora sono stati definiti i concetti di base necessari per poter analizzare le prestazioni di singoli sistemi di servizio e reti di code.
In questo capitolo si utilizzeranno tali conoscenze per poter studiare reti tandem con numero fissato $R = r$ di risorse al variare dei parametri di input ponendo particolare attenzione alla differenza tra reti tandem con distribuzione geometrica e reti tandem con distribuzione binomiale negativa.
I risultati ottenuti saranno esplicitati mostrando diversi grafici definiti mediante il linguaggio R.
Si vuole, inoltre, per alcune reti tandem simulare il numero di utenti presenti nella rete in condizioni di equilibrio statistico e confrontare i risultati teorici con quelli simulati.

6.2 Analisi grafica delle differenze tra modello geometrico e modello esteso di tipo binomiale negativo

Si consideri una rete tandem costituita da $R = r$ risorse che soddisfa le proprietà (1-6) definite nel paragrafo 5.3. Si assuma, inoltre, che $\rho_1 = \rho_2 = \ldots = \rho_r = \rho$ e $\beta_1 = \beta_2 = \ldots = \beta_r = \beta$ al fine di semplificare la trattazione matematica e la rappresentazione grafica.
Si ricordi che per il modello geometrico e il modello esteso di tipo binomiale negativo il numero medio di utenti e la varianza sono:

- **modello geometrico**

$$E(M_r) = \sum_{i=1}^{r} E(N_i) = \sum_{i=1}^{r} \frac{\rho_i}{1-\rho_i}$$

e

$$Var(M_r) = \sum_{i=1}^{r} Var(N_i) = \sum_{i=1}^{r} \frac{\rho_i}{(1-\rho_i)^2}$$

ma poiché $\rho_1 = \rho_2 = \ldots = \rho_r = \rho$ si ha

$$E(M_r) = \frac{r\rho}{1-\rho}$$

e

$$Var(M_r) = \frac{r\rho}{(1-\rho)^2}$$

- **modello esteso di tipo binomiale negativo**

$$E(M_r) = \sum_{i=1}^{r} E(N_i) = \sum_{i=1}^{r} \frac{\rho_i}{1-\rho_i}(1+\beta_i)$$

e

$$Var(M_r) = \sum_{i=1}^{r} Var(N_i) = \sum_{i=1}^{r} \frac{\rho_i}{(1-\rho_i)^2}(1+\beta_i)$$

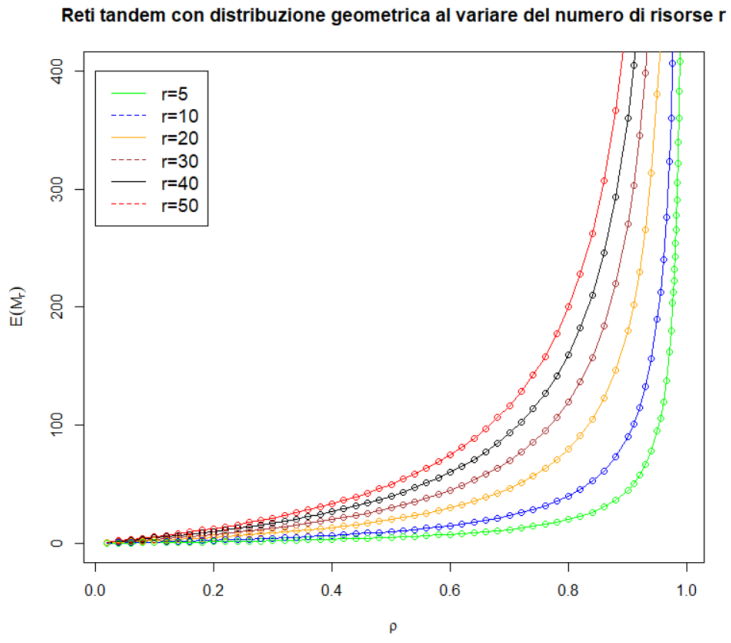
ma poiché $\rho_1 = \rho_2 = \ldots = \rho_r = \rho$ e $\beta_1 = \beta_2 = \ldots = \beta_r = \beta$ si ha

$$E(M_r) = \frac{r\rho}{1-\rho}(1+\beta)$$

e

$$Var(M_r) = \frac{r\rho}{(1-\rho)^2}(1+\beta)$$

Di seguito è riportato un grafico in cui si descrive il numero medio di utenti in reti tandem con distribuzione geometrica al variare del numero di risorse e di ρ



Si osserva che al crescere del numero di risorse r e di ρ il numero medio di utenti nella rete aumenta.

Analisi grafica della media degli utenti e varianza per il modello geometrico e binomiale negativo

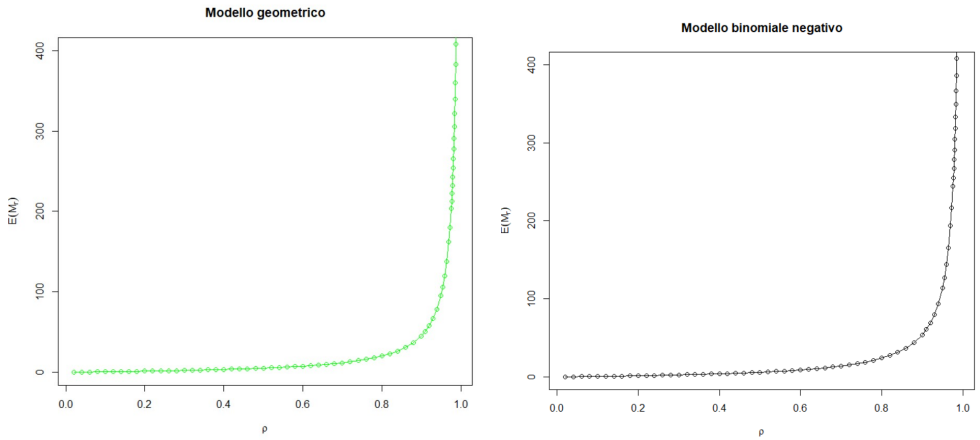
Si desidera mostrare, mediante rappresentazione grafica, che sussistono le seguenti relazione:

$$E(N^G) \leq E(N^B)$$

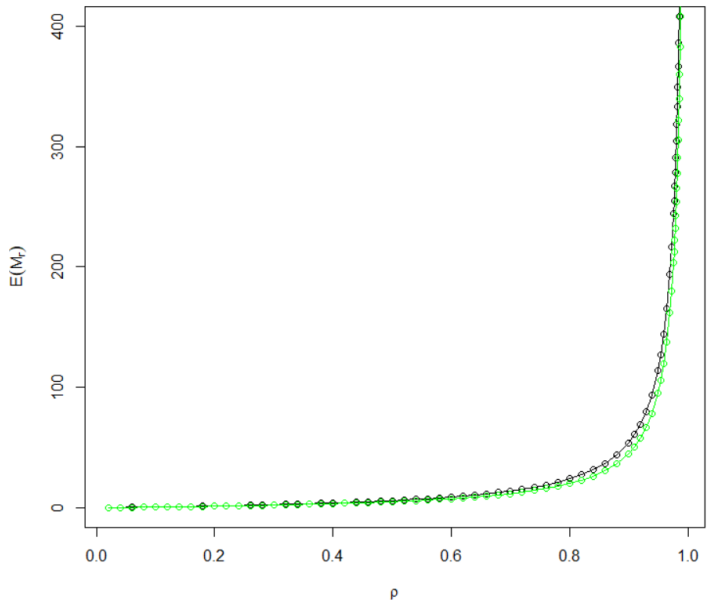
e

$$Var(N^G) \leq Var(N^B)$$

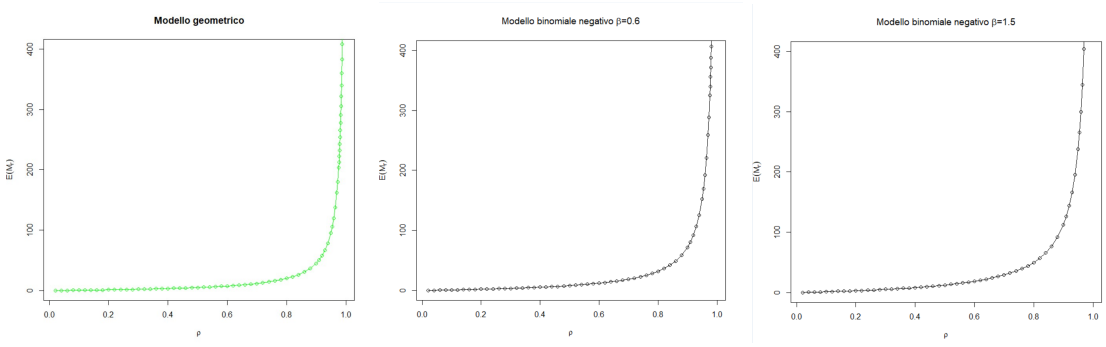
ossia che il numero medio di utenti e la varianza sono inferiori nel modello geometrico rispetto al modello esteso di tipo binomiale negativo.
Di seguito sono riportati dei grafici nei quali sono rese evidenti le differenze tra il modello geometrico e il modello esteso di tipo binomiale negativo:



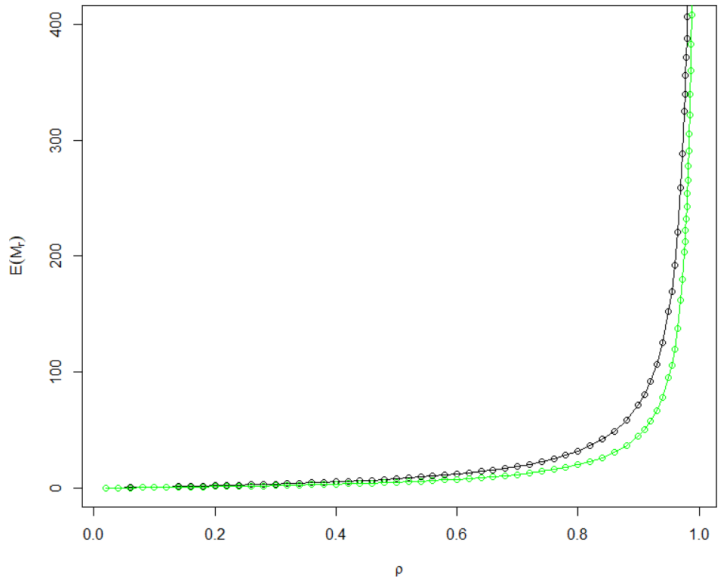
Confronto tra modello geometrico e modello binomiale negativo con $\beta = 0.2$



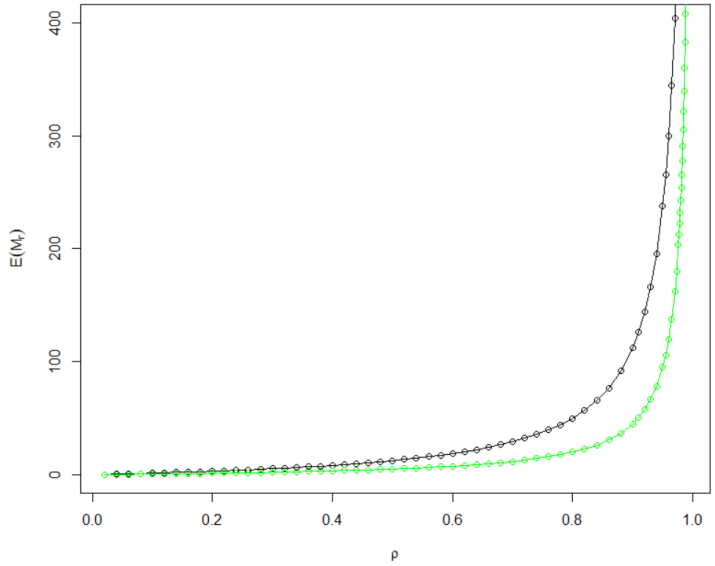
La linea verde descrive il numero medio di utenti al crescere di ρ in una rete tandem con 5 risorse e caratterizzata da distribuzione geometrica. La linea nera, invece, rappresenta il numero medio di utenti in una rete tandem con 5 risorse e caratterizzata da distribuzione binomiale negativa con $\beta = 0.2$.
Dalla sovrapposizione dei due grafi si evince che, anche per valori di β piccoli, il numero medio di utenti è inferiore nel modello geometrico rispetto al modello binomiale negativo.



Confronto tra modello geometrico e modello binomiale negativo con $\beta = 0.6$

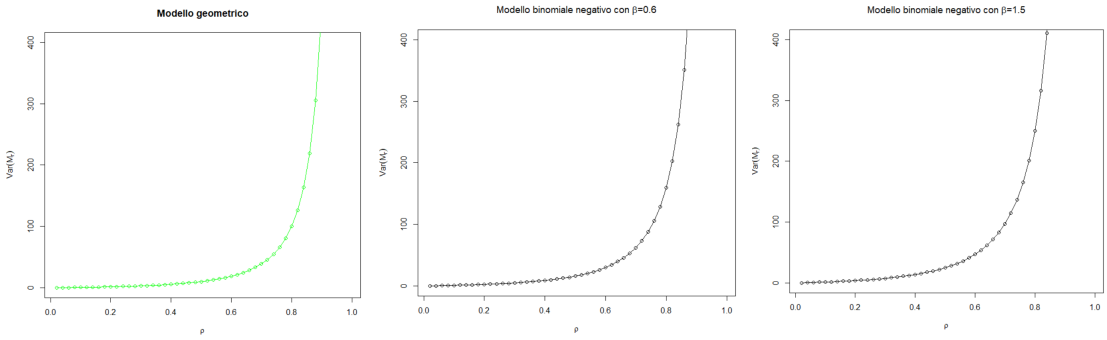


Confronto tra modello geometrico e modello binomiale negativo con $\beta = 1.5$

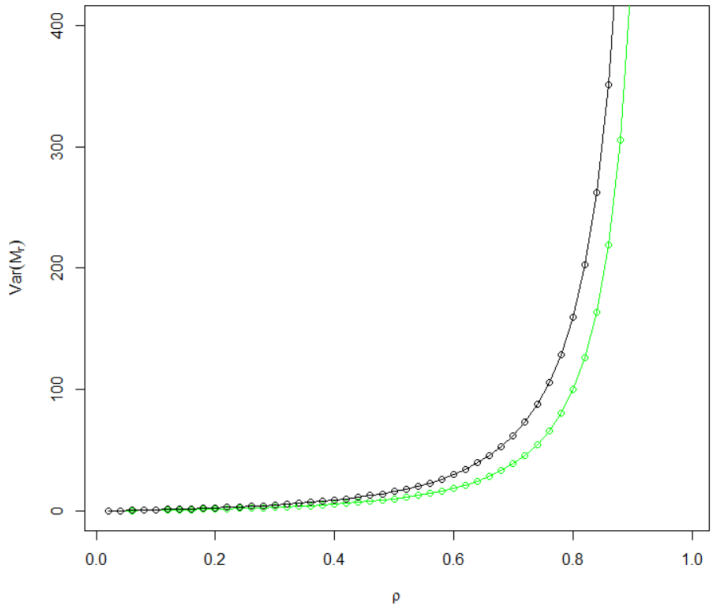


Dallo studio dei grafici riportati sopra si può facilmente notare che all'aumentare del parametro β risulta sempre più evidente che la media degli utenti nel modello geometrico risulta inferiore alla media degli utenti nel modello esteso di tipo binomiale negativo.

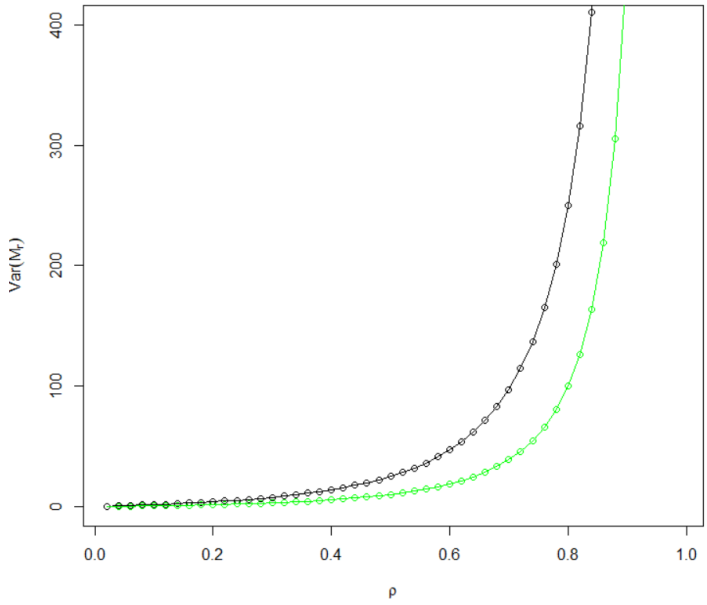
I risultati ottenuti per la media degli utenti si riscontrano anche per la varianza.
Di seguito, infatti, sono riportati dei grafici che mostrano tali risultati:



Confronto tra modello geometrico e modello binomiale negativo con $\beta = 0.6$



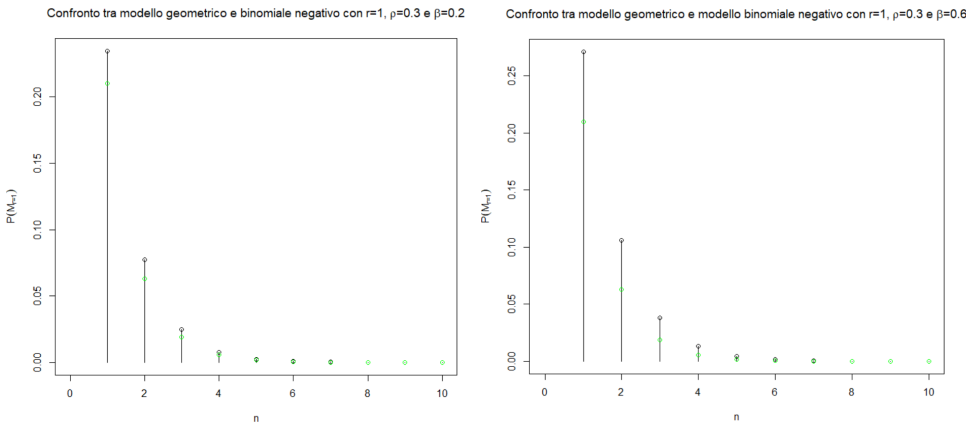
Confronto tra modello geometrico e modello binomiale negativo con $\beta = 1.5$

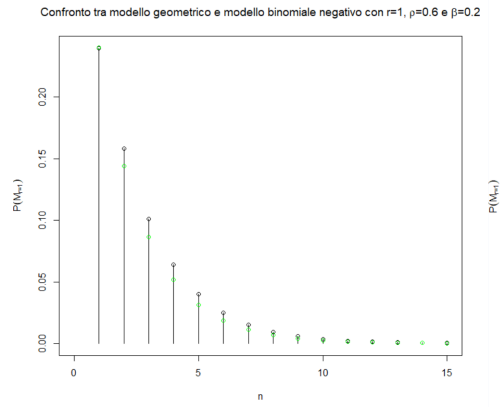


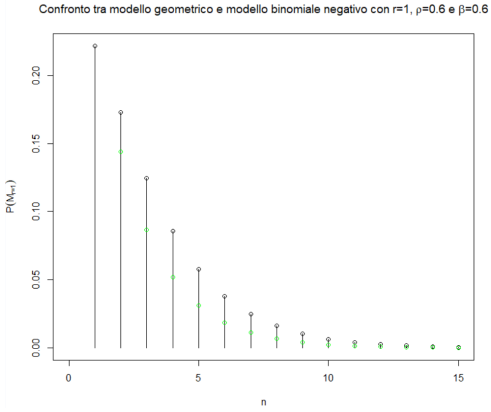
La linea verde descrive la varianza in una rete tandem con 5 risorse e caratterizzata da distribuzione geometrica al crescere di ρ. La linea nera, invece, rappresenta la varianza in una rete tandem con 5 risorse e caratterizzata da distribuzione binomiale negativa con $\beta = 0.6$ e $\beta = 1.5$. Dai grafici ottenuti si osserva che al crescere di β la varianza del modello geometrico è inferiore rispetto alla varianza del modello esteso di tipo binomiale negativo.

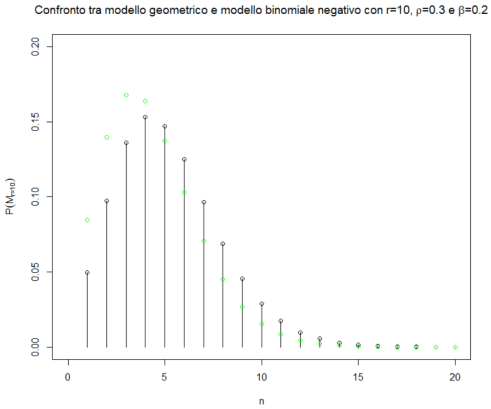
Analisi grafica della distribuzione di probabilità per il modello geometrico e binomiale negativo

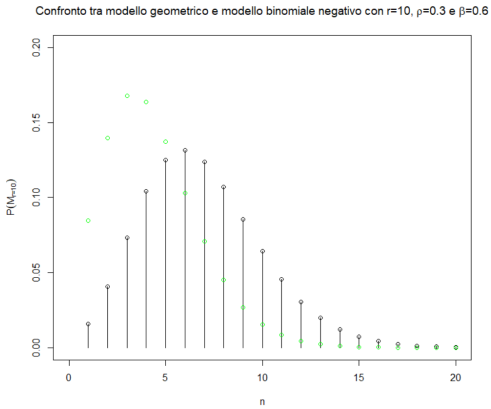
Un altro aspetto di notevole interesse è l'andamento grafico della distribuzione di probabilità al variare del numero di risorse r per i due modelli considerati.

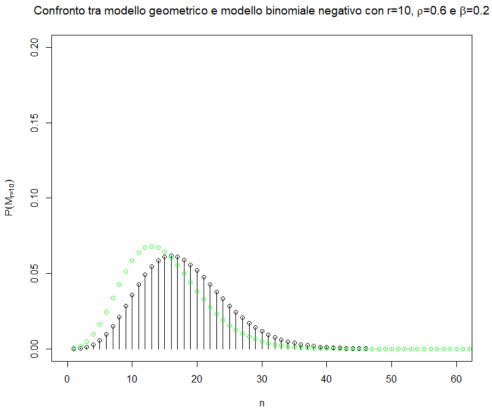


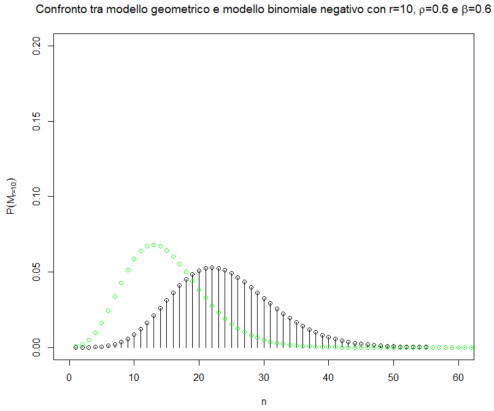

Confronto tra modello geometrico e modello binomiale negativo con r=1, ρ=0.6 e β=0.2
P(M=n)
n

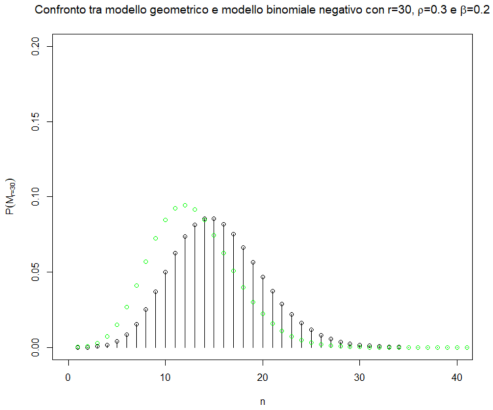

Confronto tra modello geometrico e modello binomiale negativo con r=1, ρ=0.6 e β=0.6
P(M=n)
n

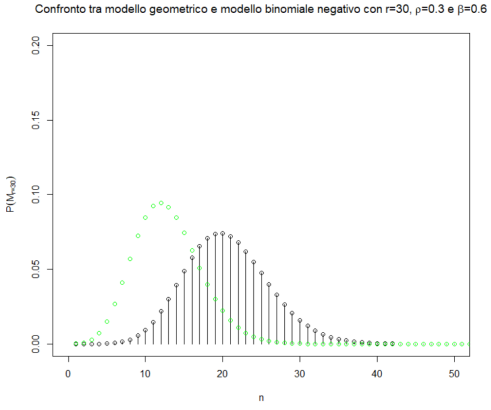

Confronto tra modello geometrico e modello binomiale negativo con r=10, ρ=0.3 e β=0.2
P(M=n)
n

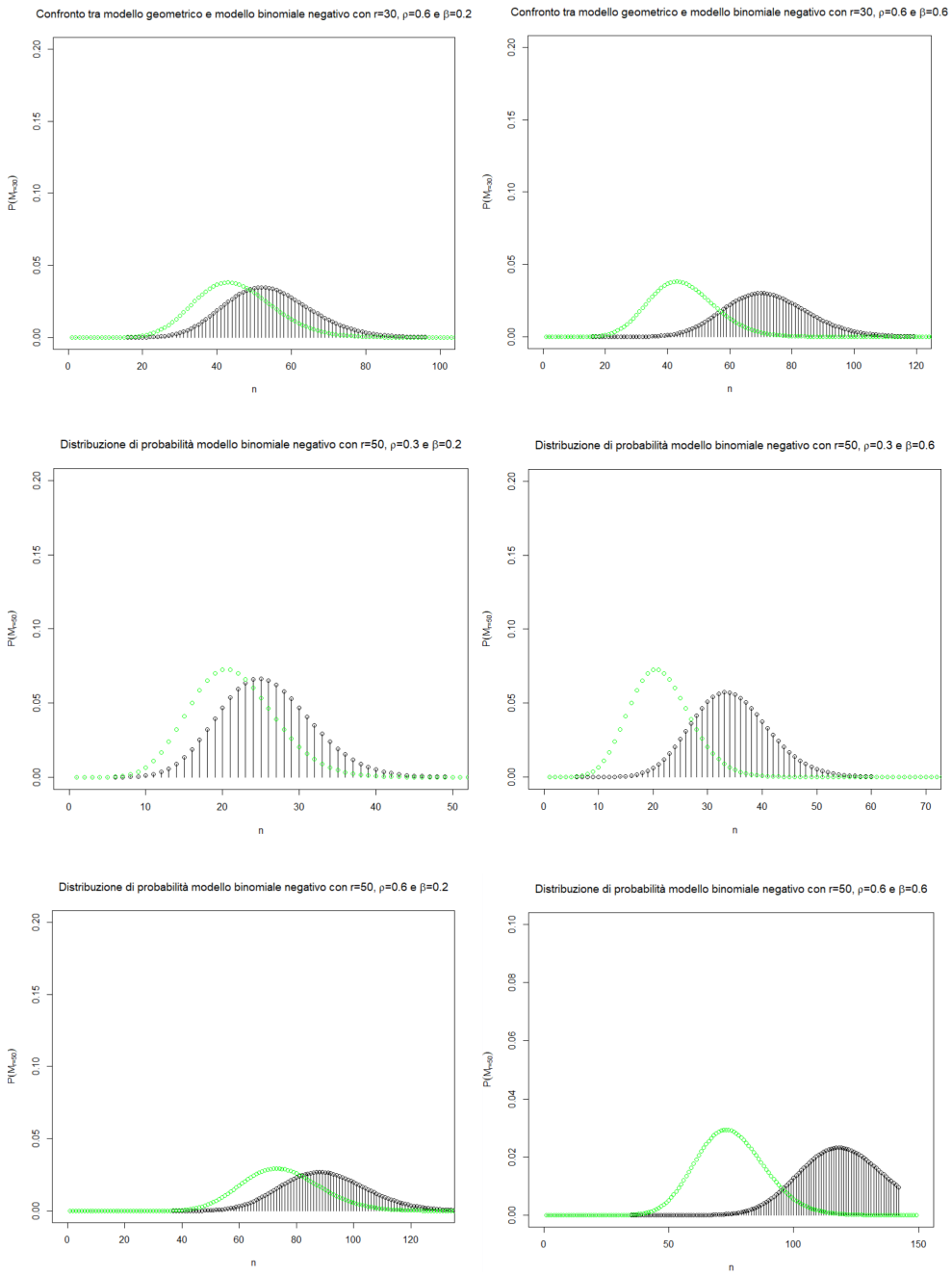

Confronto tra modello geometrico e modello binomiale negativo con r=10, ρ=0.3 e β=0.6
P(M=n)
n


Confronto tra modello geometrico e modello binomiale negativo con r=10, ρ=0.6 e β=0.2
P(M=n)
n


Confronto tra modello geometrico e modello binomiale negativo con r=10, ρ=0.6 e β=0.6
P(M=n)
n


Confronto tra modello geometrico e modello binomiale negativo con r=30, ρ=0.3 e β=0.2
P(M=n)
n


Confronto tra modello geometrico e modello binomiale negativo con r=30, ρ=0.3 e β=0.6
P(M=n)
n



I grafici mostrati sopra rappresentano in verde la distribuzione di probabilità del modello geometrico ed in nero la distribuzione di probabilità del modello esteso di tipo binomiale negativo. Dall'analisi grafica si evince che nel modello geometrico con $r = 1$ risorse la distribuzione di probabilità risulta assumere valori decrescenti in quanto il ridotto numero di risorse comporta un'affluenza via via decrescente. Al crescere del numero di risorse r, invece, il valore modale cresce. Questo si nota in tutti i casi in cui abbiamo aumentato il numero di risorse r.
È interessante notare anche come per valori elevati di r sia praticamente nulla la probabilità che si presenti un numero basso di utenti. Nel modello binomiale negativo l'andamento è molto simile al modello geometrico, tuttavia per valori bassi del numero di utenti n la probabilità risulta inferiore a quella ottenuta con il modello geometrico. Quindi, questo modello è più adatto a descrivere situazioni in cui l'affluenza degli utenti è maggiore.

6.3 Simulazione di variabili aleatorie discrete

Per poter simulare reti tandem caratterizzate da distribuzione geometrica e binomiale negativa è necessario generare variabili aleatorie discrete a partire da variabili aleatorie uniformemente distribuite nell'intervallo $(0, 1)$.
In un esperimento di simulazione, infatti, occorre spesso generare più sequenze indipendenti per rappresentare variabili aleatorie differenti. Ad esempio, nel caso della simulazione di un sistema di servizio è necessario generare due sequenze indipendenti per definire i tempi di interarrivo e di servizio.
Per ottenere sequenze uniformi in $(0, 1)$ indipendenti esistono diverse possibilità:

1. utilizzare costanti moltiplicative differenti nel metodo congruenziale moltiplicativo;
2. utilizzare semi differenti;
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 distinte sottosequenze da utilizzare per generare le differenti variabili aleatorie indipendenti.

Nei paragrafi successivi per la generazione di sequenze uniformi indipendenti sarà adottato il secondo metodo.

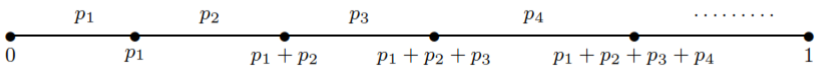
Variabili aleatorie discrete
Sia X un variabile aleatorie discreta che assume valori in un insieme finito o al più numerabile $S = \{x_1, x_2, \ldots\}$ e sia

$$p_j = P(X = x_j) \quad (j = 1, 2, \ldots) \tag{6.1}$$

la sua funzione di probabilità. Ovviamente si deve avere che

$$p_j \geq 0 \quad (j = 1, 2, \ldots), \qquad \sum_{j:x_j \in S} p_j = 1 \tag{6.2}$$

Per simulare la variabile aleatoria discreta X si suddivide l'intervallo $(0, 1)$ in tanti sottointervalli di ampiezza $p_1, p_2, \ldots$ in modo tale che $p_1 + p_2 + \ldots = 1$ come mostrato nella figura sottostante



Di seguito è riportato un metodo generale per la simulazione della variabile aleatoria discreta X:
Algoritmo
PASSO 1: generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
PASSO 2: Porre

$$X = \begin{cases} x_1, & 0 \leq U < p_1 \\ x_2, & p_1 \leq U < p_1 + p_2 \\ \ldots & \\ x_j, & \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \\ \ldots & \end{cases}$$

La variabile aleatoria discreta X generata con questo algoritmo ha la funzione di probabilità (6.1).
In particolare per la simulazione di reti tandem con distribuzione geometrica e binomiale negativa è necessario generare variabili aleatorie geometriche.

Simulazione di una variabile aleatoria geometrica
Sia X una variabile aleatoria con funzione di probabilità geometrica

$$p_j = P(X = x_j) = (1-p)^{j-1}p \quad (j = 1, 2, \ldots) \tag{6.3}$$

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:
PASSO 1: generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
PASSO 2: Porre

$$X = \begin{cases} x_1, & 0 \le U < p_1 \\ x_2, & p_1 \le U < p_1 + p_2 \\ \ldots & \\ x_j, & \sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i \\ \ldots & \end{cases}$$

Per poter semplificare l'algoritmo si può notare che dalla (6.3) 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:

$$\sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i \iff 1 - (1-p)^{j-1} \le U < 1 - (1-p)^j$$

$$\iff (1-p)^j < 1 - U \le (1-p)^{j-1}$$

$$\iff j\ln(1-p) < \ln(1-U) \le (j-1)\ln(1-p)$$

$$\iff \frac{\ln(1-U)}{\ln(1-p)} < j \le 1 + \frac{\ln(1-U)}{\ln(1-p)}$$

Per simulare una variabile aleatoria geometrica si può quindi utilizzare il seguente algoritmo:
Algoritmo
PASSO 1: generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
PASSO 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.

Simulazione di una variabile aleatoria binomiale negativa
Sia Y una variabile aleatoria binomiale negativa di parametro $\beta = k-1$ con distribuzione di probabilità

$$q_n = \frac{p^n}{n!}(k)_n(1-p)^k = p^n\binom{k+n-1}{n}(1-p)^k \quad (k=2,3,\ldots) \tag{6.4}$$

essendo

$$\frac{(k)_n}{n!} = \frac{k(k+1)\ldots(k+n-1)}{n!} = \frac{k(k+1)\ldots(k+n-1)(k-1)!}{(k-1)!n!} = \binom{k+n-1}{k-1} = \binom{k+n-1}{n}$$

La variabile aleatoria binomiale negativa Y esprime il numero di fallimenti n che precedono il successo k-esimo in una sequenza di lanci indipendenti di un moneta con probabilità di successo $1-p$ e probabilità di insuccesso p (successione di prove indipendenti di Bernoulli con probabilità di successo $1-p$ e probabilità di insuccesso p).
La variabile aleatoria binomiale negativa Y può essere considerata come somma di k variabili aleatorie geometriche. Quindi per simulare tale variabile si può utilizzare il seguente algoritmo:
Algoritmo
PASSO 1: generare k variabili aleatorie indipendenti $U_1,\ldots,U_k$ uniformemente distribuite nell'intervallo (0,1);
PASSO 2: Porre

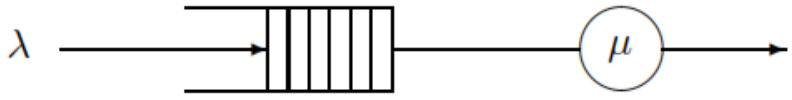
$$Y_i = \left\lceil \frac{\ln(1-U_i)}{\ln(p)} \right\rceil \quad (i=1,2,\ldots,k)$$

PASSO 3: Calcolare

$$Y = \sum_{i=1}^{k} Y_i$$

6.4 Simulazione del numero di utenti in un sistema di servizio M/M/1 con distribuzione geometrica

Si consideri un sistema di servizio M/M/1 con distribuzione geometrica caratterizzato da tempi di interarrivo e di servizio esponenziali con valore medio rispettivamente $1/\lambda$ e $1/\mu$ con un unico servitore e capacità del sistema infinita come riportato nella seguente figura:



Il sistema $M/M/1$ non si congestiona se $\rho = \lambda/\mu < 1$. Se si denota con N la variabile aleatoria che descrive il numero di utenti presenti nel sistema di servizio $M/M/1$ caratterizzato da distribuzione geometrica, in condizioni di equilibrio statistico, si ha:

$$q_n = P(N=n) = (1-\rho)\rho^n \quad (n=0,1,\ldots) \tag{6.5}$$

ossia una funzione di probabilità geometrica. Quindi per poter simulare la variabile aleatoria discreta N si può procedere in modo analogo alla generazione di una variabile aleatoria geometrica analizzata nella sezione 6.3.

Algoritmo
PASSO 1: generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0,1)$;
PASSO 2: Porre

$$N = \begin{cases} 0, & 0 \leq U < q_1 \\ 1, & q_1 \leq U < q_1 + q_2 \\ \dots & \\ j, & \sum_{i=1}^{j-1} q_i \leq U < \sum_{i=1}^{j} q_i \\ \dots & \end{cases}$$

Seguendo la stessa procedura adottata per la variabile aleatoria geometrica si può ricavare un algoritmo più semplice per simulare la variabile aleatoria discreta N. Dalla (6.5), infatti, si ha

$$\sum_{i=0}^{j-1} q_i = (1-\rho) \sum_{i=0}^{j-1} \rho^i = 1 - \rho^j$$

Occorre quindi porre $N = j$ se risulta

$$\sum_{i=0}^{j-1} q_i \leq U < \sum_{i=0}^{j} q_i \iff 1 - \rho^j \leq U < 1 - \rho^{j+1}$$

$$\iff \rho^{j+1} < 1 - U \leq \rho^j \iff (j+1)\ln(\rho) < \ln(1-U) \leq j \ln(\rho)$$

$$\iff \frac{\ln(1-U)}{\ln(\rho)} - 1 < j \leq \frac{\ln(1-U)}{\ln(\rho)},$$

dove l'ultima disuguaglianza segue poiché dividendo per $\ln(\rho) < 0$ occorre invertire i segni della disequazione essendo $\rho < 1$. 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
PASSO 1: generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0,1)$;
PASSO 2: Porre

$$N = \left\lceil \frac{\ln(1-U)}{\ln(\rho)} - 1 \right\rceil$$

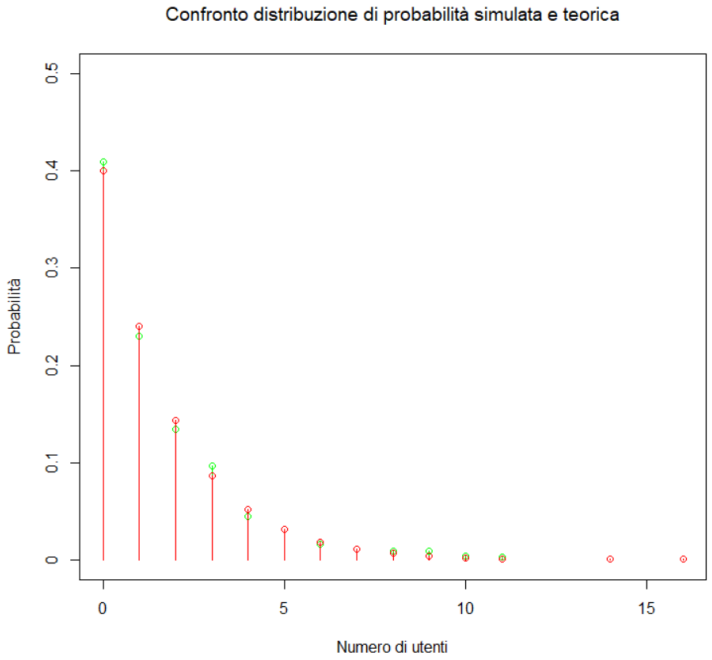
dove $\lceil x \rceil$ denota il più piccolo intero maggiore di x.
Si desidera ora, mediante una funzione in R, generare una sequenza che descriva il numero di utenti presenti nel sistema M/M/1 con distribuzione geometrica in condizioni di equilibrio statistico e confrontare la media campionaria ottenuta mediante la simulazione con il numero medio teorico $E(N) = \rho/(1-\rho)$.

```
> MMlqueue<-function(n,rho,seme){
+  set.seed(seme)
+  u<-runif(n)
+  w<-log(1-u)/log(rho)-1
+  N<-ceiling(w)
+  return(N)
+  }
> utenti <- MMlqueue(1000,0.6,3)
> mean(utenti)
[1] 1.553
```

Il valore della media teorica in questo caso è $E(N) = \rho/(1-\rho) = 0.6/(1-0.6) = 1.5$ che è molto vicino al risultato ottenuto dalla media simulata 1.553.
Applicando la funzione table, messa a disposizione da R, è possibile calcolare la frequenza assoluta del numero di utenti nella sequenza.
Il rapporto table(utenti)/length(utenti) fornisce, invece, la distribuzione di probabilità simulata del numero di utenti presenti nel sistema:

```
> table(utenti)
utenti
  0   1   2   3   4   5   6   7   8   9  10  11  14  16
409 230 134  97  45  31  16  11   9   9   4   3   1   1
> table(utenti)/length(utenti)
utenti
    0     1     2     3     4     5     6     7     8     9    10    11    14    16
0.409 0.230 0.134 0.097 0.045 0.031 0.016 0.011 0.009 0.009 0.004 0.003 0.001 0.001
```

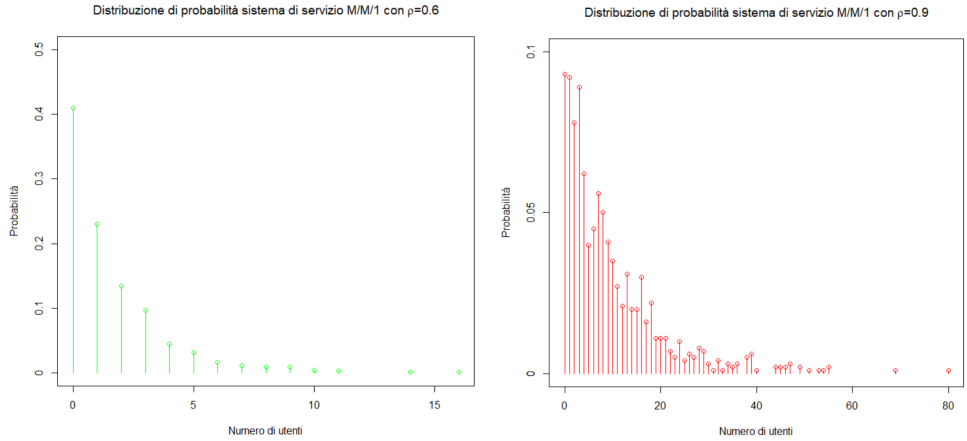
Le probabilità ottenute dalla simulazione sono molto vicine ai valori teorici. Tale risultato è reso evidente dal grafico riportato di seguito, nel quale in verde è rappresentata la distribuzione di probabilità simulata e in rosso quella teorica.

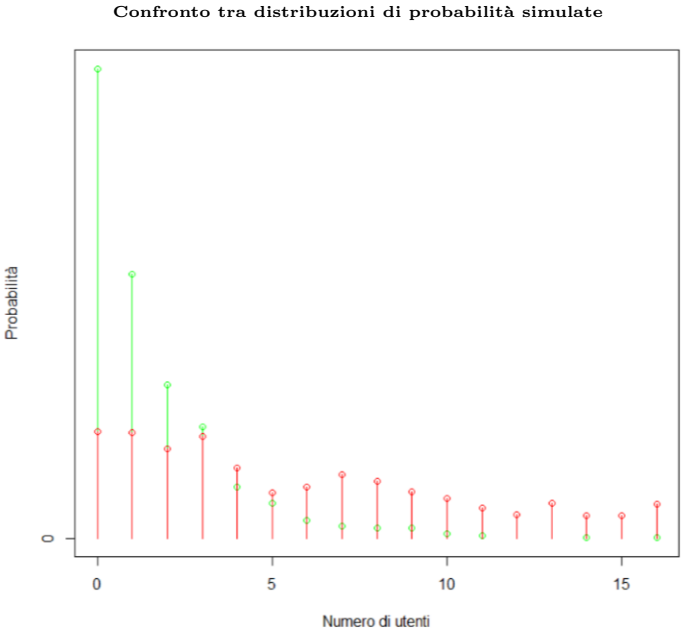


Si consideri ora un sistema di servizio $M/M/1$ con parametro $\rho = 0.9$.

```
> utenti<-MM1queue(1000,0.9,3)
> table(utenti)
utenti
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
93 92 78 89 62 40 45 56 50 41 35 27 21 31 20 20 30 16 22 11 11 11  7  5 10  4  6  5  8  7
30 31 32 33 34 35 36 38 39 40 44 45 46 47 49 51 53 54 55 69 80
 3  1  4  1  3  2  3  5  6  1  2  2  2  3  2  1  1  1  2  1  1
> table(utenti)/length(utenti)
utenti
    0     1     2     3     4     5     6     7     8     9    10    11    12    13    14
0.093 0.092 0.078 0.089 0.062 0.040 0.045 0.056 0.050 0.041 0.035 0.027 0.021 0.031 0.020
   15    16    17    18    19    20    21    22    23    24    25    26    27    28    29
0.020 0.030 0.016 0.022 0.011 0.011 0.011 0.007 0.005 0.010 0.004 0.006 0.005 0.008 0.007
   30    31    32    33    34    35    36    38    39    40    44    45    46    47    49
0.003 0.001 0.004 0.001 0.003 0.002 0.003 0.005 0.006 0.001 0.002 0.002 0.002 0.003 0.002
   51    53    54    55    69    80
0.001 0.001 0.001 0.002 0.001 0.001
```

Si può notare che, all'aumentare del parametro ρ, la probabilità di avere un numero di utenti nel sistema pari a 0 è sempre meno significativa, ovvero è più difficile che il sistema sia vuoto. Di seguito sono riportati dei grafici che mostrano l'andamento delle distribuzioni di probabilità simulate per $\rho = 0.6$ e per $\rho = 0.9$ considerando un campione di ampiezza 1000 e seme iniziale 3:





6.5 Simulazione del numero di utenti in un sistema di servizio con distribuzione binomiale negativa

Si consideri un sistema di servizio con distribuzione binomiale negativa di parametro $\beta = k - 1$ con un unico servitore e capacità del sistema infinita.
Un sistema sì fatto non si congestiona se $\rho = \lambda/\mu < 1$. Se si denota con N la variabile aleatoria che descrive il numero di utenti presenti nel sistema di servizio caratterizzato da distribuzione binomiale negativa con $\beta = k - 1$, in condizioni di equilibrio statistico, si ha:

$$q_n = \frac{\rho^n}{n!}(k)_n(1-\rho)^k = \rho^n \binom{k+n-1}{n}(1-\rho)^k \quad (k = 2, 3, \ldots)$$

ossia una distribuzione di probabilità binomiale negativa.
Si desidera ora, in condizioni di equilibrio statistico, generare una sequenza che descriva il numero di utenti presenti nel sistema con distribuzione binomiale negativa e confrontare la media campionaria e la varianza ottenute dalla simulazione con il numero medio teorico $E(N) = \dfrac{\rho}{(1-\rho)}(1+\beta)$

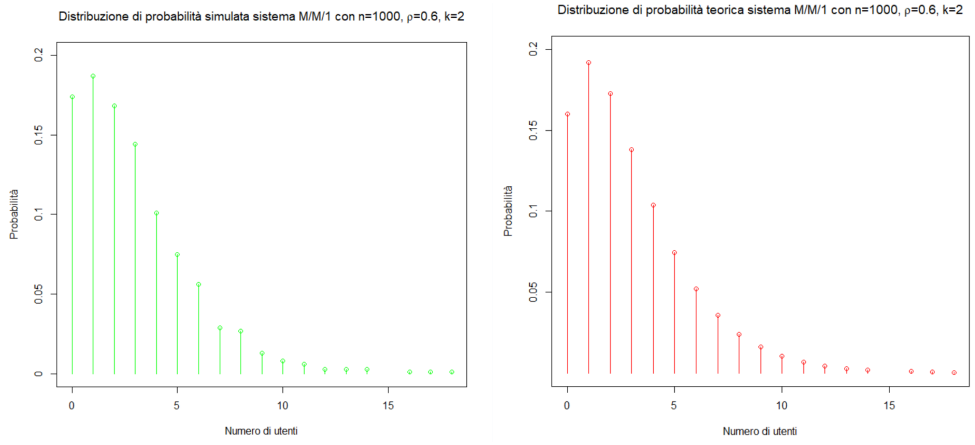
e la varianza $Var(N) = \frac{\rho}{(1-\rho)^2}(1+\beta)$.
Per quanto visto precedentemente la variabile aleatoria N può essere simulata come somma di k variabili aleatorie geometriche, ossia $N = N_1 + N_2 + \ldots + N_k$.

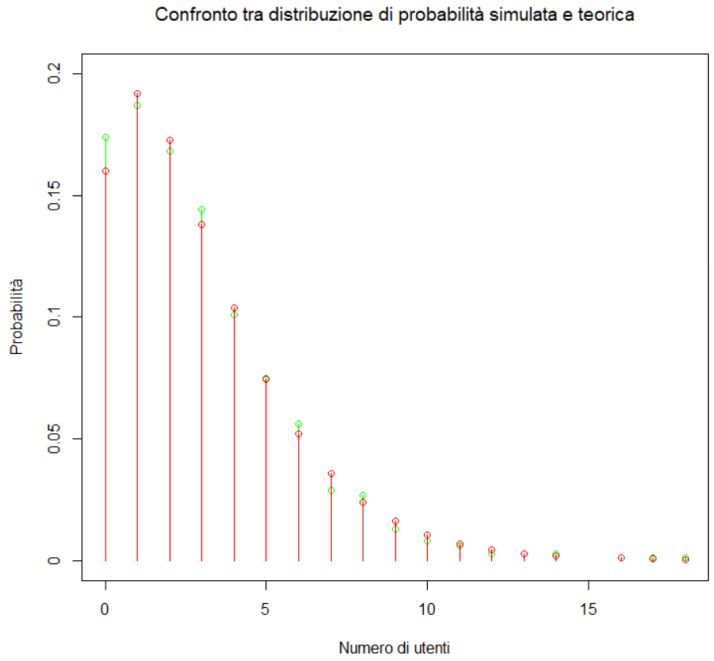
```
> BNqueue<-function(n,rho,k,seme){
+ N<-numeric(n)
+ for(j in 1:k){
+ set.seed(seme+j)
+ u<-runif(n)
+ w<-(log(1-u)/log(rho))-1
+ N<-N+ceiling(w)
+ }
+ return(N)
+ }
>
> utenti <- BNqueue(1000,0.6,2,3)
> mean(utenti)
[1] 2.92
> var(utenti)
[1] 7.266867
```

Per $k = 2, \beta = k - 1 = 1$ e $\rho = 0.6$ si ha che la media e la varianza simulate assumono un valore molto vicino rispettivamente alla media teorica $E(N) = \frac{\rho}{(1-\rho)}(1+\beta) = 3$ e alla varianza teorica $Var(N) = \frac{\rho}{(1-\rho)^2}(1+\beta) = 7.5$.
Utilizzando la funzione table, messa a disposizione da R, è possibile calcolare la frequenza assoluta del numero di utenti nella sequenza generata.
Il rapporto table(utenti)/length(utenti) fornisce, invece, la distribuzione di probabilità simulata del numero di utenti presenti nel sistema:

```
> utenti <- BNqueue(1000,0.6,2,3)
> distribuzione_n1000 <-table(utenti)/length(utenti)
> distribuzione_n1000
utenti
    0     1     2     3     4     5     6     7     8     9    10    11    12    13    14
0.174 0.187 0.168 0.144 0.101 0.075 0.056 0.029 0.027 0.013 0.008 0.006 0.003 0.003 0.003
   16    17    18
0.001 0.001 0.001
```

Di seguito sono riportati dei grafici in cui in verde si mostra la distribuzione di probabilità simulata e in rosso quella teorica:





Si può notare che, anche per sequenze non troppo grandi ($n = 1000$), la distribuzione di probabilità simulata si avvicina a quella teorica.
Incrementando, invece, il valore del parametro k la media e la varianza calcolate aumentano, infatti, si ha per $k = 5$ e $k = 10$:

```
> utenti <- BNqueue(1000,0.6,5,3)
> mean(utenti)
[1] 7.449
> var(utenti)
[1] 19.51692
>
> utenti <- BNqueue(1000,0.6,10,3)
> mean(utenti)
[1] 15.031
> var(utenti)
[1] 39.32737
```

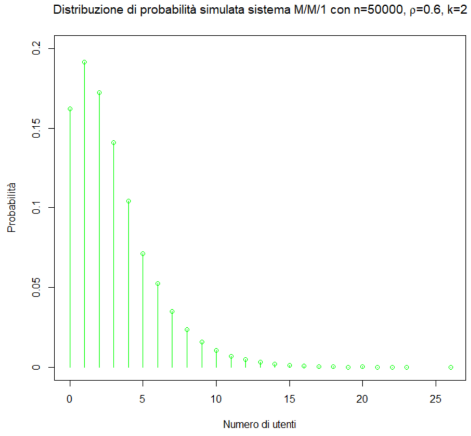
Applicando la funzione BNqueue per $n = 50000, k = 2, \beta = 1, \rho = 0.6$ si ha:

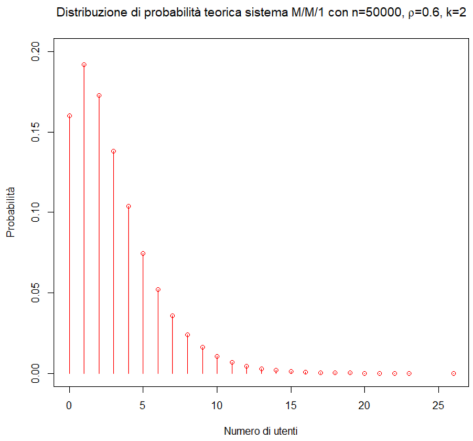
```
> utenti <- BNqueue(50000,0.6,2,3)
> mean(utenti)
[1] 2.99304
> var(utenti)
[1] 7.598904
```

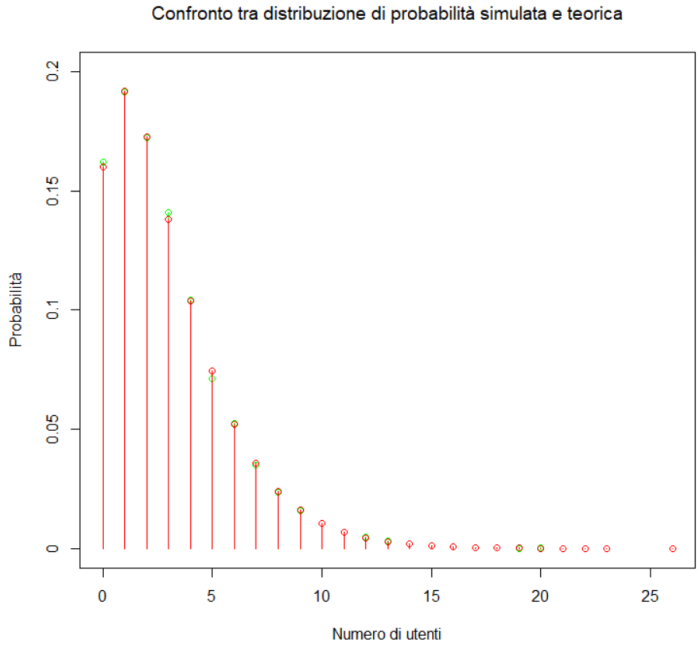
Da ciò si evince che, all'aumentare della lunghezza della sequenza generata, la media e la varianza simulate assumono valori sempre più vicini a quelli teorici, infatti, $E(N) = 3$ e $Var(N) = 7.5$.
Si desidera ora mostrare tale risultato anche per la distribuzione di probabilità simulata e teorica:

```
> utenti <- BNqueue(50000,0.6,2,3)
> distribuzione_n50000 <-table(utenti)/length(utenti)
> distribuzione_n50000
utenti
      0       1       2       3       4       5       6       7       8       9      10
0.16216 0.19140 0.17216 0.14074 0.10416 0.07130 0.05258 0.03488 0.02344 0.01572 0.01058
     11      12      13      14      15      16      17      18      19      20      21
0.00708 0.00494 0.00340 0.00216 0.00100 0.00080 0.00046 0.00040 0.00014 0.00028 0.00010
     22      23      26
0.00008 0.00002 0.00002
> distribuzioneDiProbabilitàTeorica_n50000 <- function(rho,k){
+ q<-numeric(25)
+ for(i in 0:24)
+ q[i+1]<-rho^(i)*choose(k+i-1,i)*(1-rho)^k
+ return(q)
+ }
> distribuzioneTeorica<-distribuzioneDiProbabilitàTeorica_n50000(0.6,2)
> round(distribuzioneTeorica,3)
 [1] 0.160 0.192 0.173 0.138 0.104 0.075 0.052 0.036 0.024 0.016 0.011 0.007 0.005 0.003
[15] 0.002 0.001 0.001 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
```

Di seguito sono mostrati dei grafici in cui in verde è rappresentata la distribuzione di probabilità simulata e in rosso quella teorica:



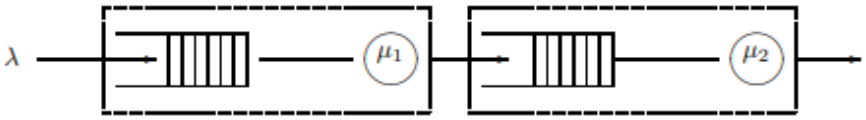




Chiaramente i grafici riportati sono quasi perfettamente sovrapposti in quanto i valori simulati sono molto vicini ai valori teorici.

6.6 Simulazione di una rete tandem con due risorse caratterizzate da distribuzione binomiale negativa

Si consideri una rete tandem costituita da due risorse caratterizzate da distribuzione binomiale negativa con singolo servitore come riportato in figura:



Tale rete raggiunge una condizione di equilibrio statistico se $\rho_1 = \lambda/\mu_1 < 1$ e $\rho_2 = \lambda/\mu_2 < 1$. Per semplificare la trattazione matematica si assuma che $\beta_1 = \beta_2 = k - 1$ e $\mu_1 = \mu_2 = \mu$ da cui segue che $\rho_1 = \rho_2 = \rho$.
Si denoti con $M_{r=2} = N_1 + N_2$ la variabile aleatoria che descrive il numero medio di utenti complessivo presenti nella rete tandem considerata dove N_1 e N_2 sono le variabili aleatorie che descrivono il numero di utenti rispettivamente nella prima e seconda risorsa.
Per quanto visto in precedenza, dalla formula (5.4), si ha:

$$P(M_{r=2} = n) = \frac{(2\beta + 2)_n}{n!}(1-\rho)^{2\beta+2}\rho^n = \frac{(2k)_n}{n!}(1-\rho)^{2k}\rho^n$$

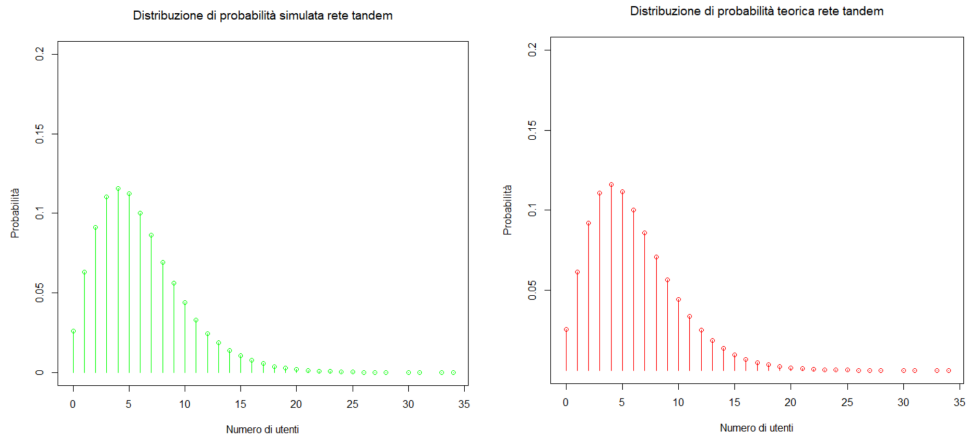
Si desidera ora, in condizioni di equilibrio statistico, generare una sequenza che descriva il numero di utenti presenti nella rete tandem analizzata e confrontare la media campionaria e la varianza ottenute dalla simulazione con il numero medio teorico $E(M_{r=2}) = \frac{r\rho}{1-\rho}(1+\beta) = \frac{2\rho}{1-\rho}k$ e la varianza $Var(M_{r=2}) = \frac{r\rho}{(1-\rho)^2}(1+\beta) = \frac{2\rho}{(1-\rho)^2}k$.

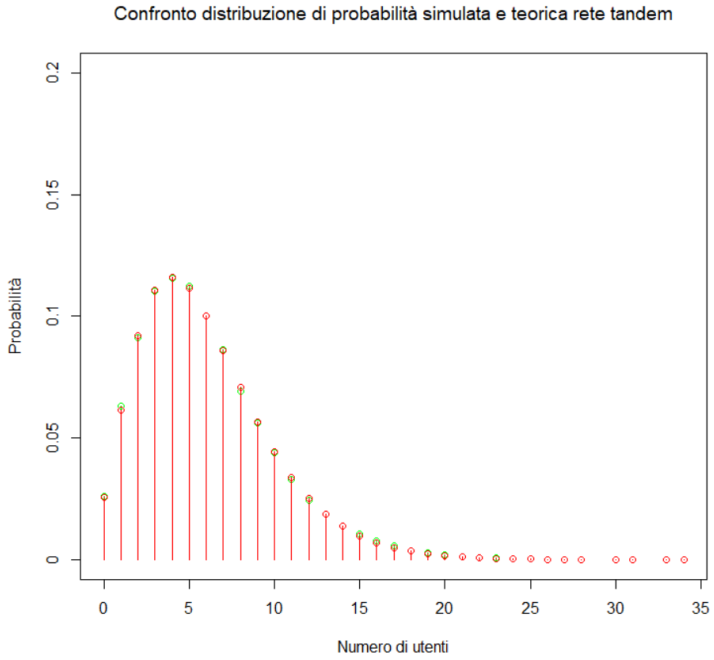
```
> #beta=k-1
> ReteTandem_BN <- function(n,r,rho,k,seme){
+ M<-numeric(n)
+ for(i in 1:r){
+ M<-M+BNqueue(n,rho,k,seme+(k*(i-1)))
+ }
+ return(M)
+ }
> utenti<-ReteTandem_BN(50000,2,0.6,2,3)
> mean(utenti)
[1] 6.0089
> var(utenti)
[1] 15.25885
```

I valori ottenuti dalla simulazione sono molto vicini alla media e alla varianza teorica, infatti, $E(M_{r=2}) = \frac{2\rho}{1-\rho}k = \frac{1.2}{0.4}2 = 6$ e $Var(M_{r=2}) = \frac{2\rho}{(1-\rho)^2}k = \frac{1.2}{(0.4)^2}2 = 15$.
Utilizzando la funzione table, messa a disposizione da R, è possibile calcolare la frequenza assoluta del numero di utenti nella sequenza generata.
Il rapporto table(utenti)/length(utenti) fornisce, invece, la distribuzione di probabilità simulata del numero di utenti presenti nella rete:

```
> table(utenti)
utenti
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
1294 3153 4557 5508 5780 5629 5011 4325 3462 2818 2203 1650 1213  942  687
  15   16   17   18   19   20   21   22   23   24   25   26   27   28   30
 534  378  276  175  132   92   55   42   34   22   10    8    3    2    2
  31   33   34
   1    1    1
> distribuzione <- table(utenti)/length(utenti)
> distribuzione
utenti
      0       1       2       3       4       5       6       7       8
0.02588 0.06306 0.09114 0.11016 0.11560 0.11258 0.10022 0.08650 0.06924
      9      10      11      12      13      14      15      16      17
0.05636 0.04406 0.03300 0.02426 0.01884 0.01374 0.01068 0.00756 0.00552
     18      19      20      21      22      23      24      25      26
0.00350 0.00264 0.00184 0.00110 0.00084 0.00068 0.00044 0.00020 0.00016
     27      28      30      31      33      34
0.00006 0.00004 0.00004 0.00002 0.00002 0.00002
```

Di seguito sono mostrati dei grafici in cui in verde è rappresentata la distribuzione di probabilità simulata e in rosso quella teorica:





Dalla sovrapposizione delle due distribuzioni si osserva, chiaramente, che i valori ottenuti dalla simulazioni sono molto vicini a quelli teorici.
Il procedimento finora descritto può essere esteso alla simulazione di una rete tandem con

un numero qualsiasi di risorse, anche caratterizzate da parametri differenti a seconda delle caratteristiche operative, preservando comunque la non congestione di ogni risorsa.

Conclusioni

In questa tesi sono state analizzate reti tandem le cui risorse sono modellate mediante processi di nascita morte.
In particolare sono stati considerati due differenti modelli, ossia il modello geometrico e binomiale negativo.
Nella tesi si confrontano reti tandem caratterizzate da distribuzione geometrica e binomiale negativa al variare dei parametri di input mostrando i risultati ottenuti mediante l'utilizzo di grafici generati in *R*.

Bibliografia

- Gross D., Harris C. M., Fundamental of Queueing Theory. Wiley, New York, third edition. (1998);
- Kleinrock L., Queueing Systems, vol. I, Theory. Wiley, New York, 1975;
- Kleinrock L., Queueing Systems, vol. II, Computer Applications. Wiley, New York, 1976;
- Ross S., Introduction to probability models. Academic Press, San Diego, eight edition (2003);
- Bolch G., Greiner S., de Meer H., Trivedi K. S., Queueing Networks and Markov Chains. Wiley-Interscience, New York, second edition (2006);
- Gelenbe E., Pujolle G., Introduction to Queueing Networks. Wiley, New York second edition (1987, 1988);
- Stewart W. J., Probability, Markov Chains, Queues, and Simulation. Princeton University Press, New Jersey, first edition (2009);
- Wolff R. W., Poisson Arrivals See Time Averages. Operations Research, Vol. 30, No. 2 (Mar.-Apr., 1982) pp. 223-231;
- Burke P. J., The Output of a Queueing System. Operations Research, Vol. 4, No. 6 (Dec., 1956), pp. 669-704;
- Burke P. J., The Output Process of a Stationary M/M/s Queueing System. The Annals of Mathematical Statistics, Vol. 39, No. 4 (Aug., 1968), pp. 1144-1152;
- Morse P. M., Operation Research. Communications on pure and applied mathematics, Vol. VIII, pp. 1-12 (1955);
- Giorno V., Nobile A. G., A Random tandem network with Queues Modeled as Birth-Death Processes. Springer International Publishing (Aug. 2018);
- Hansen Eldon R., A table of series and products. Prentice-Hall, Inc. (1975);
- Gradshteyn I. S., Ryzhik I. M., Table of Integrals, Series, and Products Academic Press, seventh edition;

- Abramowitz M., Stegun I. A., Handbook of Mathematical Functions. Dover Publications, Inc., New York;
- Chen H., Yao D. D., Fundamental of Queueing Networks. Performance, Asympotics and Optimization. Springer.