Capitolo 6: Teoria delle code ad applicazione alle reti di telecomunicazione

In questo capitolo, vedremo di analizzare in modo quantitativo e qualitativo i fenomeni di ritardo presenti in una rete di telecomunicazione. I parametri esaminati sono il ritardo medio che un pacchetto sperimenta nell'andare da una data sorgente ad una destinazione, ed il volume di traffico che la rete riesce a smaltire nell'unità di tempo (*throughput*).

La teoria delle code è uno strumento essenziale per:

- Caratterizzare le prestazioni di una rete di comunicazione;
- Dimensionare una rete o un apparato di comunicazione in modo da garantire le prestazioni desiderate in termini di ritardo medio.

6.1 Elementi di ritardo in una rete di telecomunicazione

Una rete di telecomunicazione consta di un insieme di nodi di smistamento del traffico (sotto forma di pacchetti di bit) inter-connessi da collegamenti. Il ritardo che subirà un generico pacchetto sarà dovuto alla somma dei ritardi accumulati su ogni link attraversato. Il ritardo associato ad ogni link, è costituito da quattro componenti:

1. ***Processing delay***: tempo che trascorre da quando il pacchetto è correttamente ricevuto al nodo di testa del link e quando esso viene assegnato alla coda di trasmissione di un link d'uscita.
2. ***Queueing delay***: tempo che trascorre tra l'istante in cui il pacchetto è assegnato ad una coda per la trasmissione e l'istante in cui il pacchetto inizia ad essere trasmesso.
3. ***Transmission delay***: tempo che trascorre tra l'istante in cui il primo e l'ultimo bit del pacchetto sono spediti.
4. ***Propagation delay***: tempo che impiega il segnale elettromagnetico a percorrere un collegamento. Questo tempo dipende dalle caratteristiche del mezzo trasmissivo, ed è proporzionale alla distanza che separa il *sender* dal *receiver.* Si ha ancora che tale ritardo risulta essere molto piccolo a meno del caso delle trasmissioni via satellite.

Da quanto detto sin ora, il nodo potrebbe essere schematizzato come in Figura 6.1, dal quale si deduce che il *processing delay* non è altro che la somma del tempo di attesa nella coda d'ingresso più il tempo di processamento nel server S_1, il *queueing delay* è il tempo di attesa nella coda d'uscita, mentre il *transmission delay* non è altro che il tempo di processamento o di servizio nel server corrispondente alla coda d'uscita.

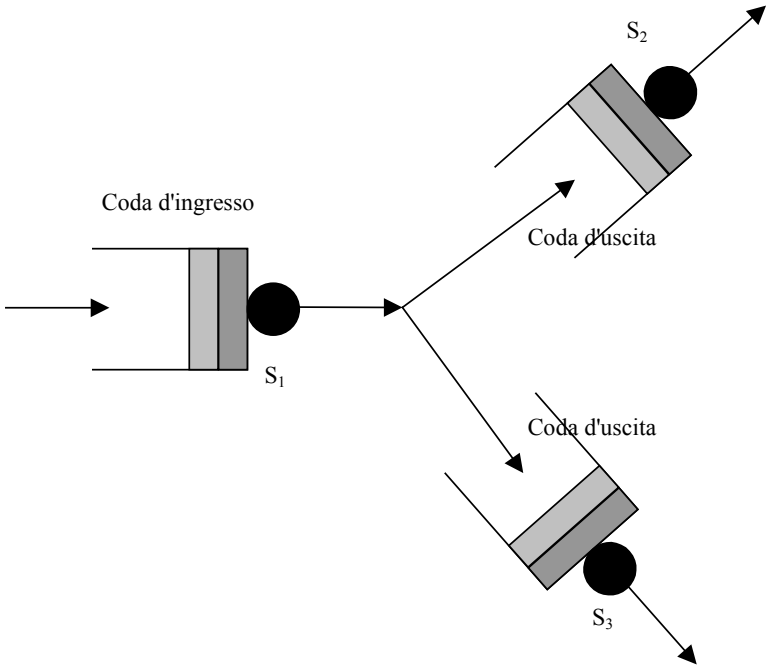


Figura 6.1: Schematizzazione delle code presenti nel generico nodo di una rete.

Quanto detto sin ora trascura la possibilità che un pacchetto debba essere ritrasmesso a causa di un errore o per qualche altra causa. In seguito, nel modellare il generico nodo, trascureremo sia il *propagation delay,* che il *processing delay,* in quanto il primo dipende dalle caratteristiche del mezzo e dalla distanza dei nodi connessi dal link, ma è del tutto indipendente dal traffico presente nella rete. Il processing delay, nelle reti a banda stretta, può essere trascurato, poiché incide poco sul ritardo totale. Esso rappresenta il tempo di lettura e di processamento del dato da parte del nodo di commutazione. Con l'ultima assunzione sparisce la coda d'ingresso ed il relativo servente dalla Figura 6.1.

6.1.1 Ritardo di trasmissione per traffico multiplexato

Abbiamo visto nel paragrafo precedente che uno degli addendi del ritardo di un pacchetto è costituito dal ritardo di trasmissione (*transmission delay*). Vedremo ora di valutare questo termine nel caso in cui la tecnica di multiplexing sia *TDM, FDM* o *Statistical Multiplexing.* Nel link possono essere trasmessi un certo numero di bit al secondo e questo valore, indicato tipicamente con C ed espresso in bps (bit/s), dipende sia dalle caratteristiche del mezzo fisico di cui il link è costituito, che dalle interfacce usate per la comunicazione.

6.1.2 Statistical Multiplexing

In questo caso i pacchetti che arrivano dalle varie sorgenti sono posti in un'unica coda e vengono serviti con una politica di tipo FIFO. L bits sia la *lunghezza di un generico pacchetto* e C bit/s la *capacità del canale*; poiché la risorsa trasmissiva è allocata interamente ad un singolo pacchetto alla volta si ha:

$$t_d = L / C \qquad (6.1)$$

6.1.3 Frequency Division Multiplexing (FDM)

Supponiamo di avere *m* sorgenti di traffico, le quali devono essere multiplexate mediante FDM sul nostro link. In tal caso, detta W la banda passante del canale, ad ogni stream di traffico verrà associato un canale avente banda circa pari a W/m. Detta C bit/s la capacità trasmissiva del canale, all'i-esimo traffic stream sarà associato un canale di C/m bit/s circa. Quindi, il tempo necessario a trasmettere un pacchetto lungo L bits è pari a:

$$t_d = mL/C \tag{6.2}$$

Osserviamo come questo tempo sia *m* volte più grande rispetto a quello relativo allo statistical multiplexing.

6.1.4 Time Division Multiplexing (TDM)

Supponendo di avere sempre *m* stream di traffico, dobbiamo distinguere il caso in cui la dimensione degli slot è piccola rispetto alla lunghezza del pacchetto, dal caso in cui slot e pacchetto hanno la stessa dimensione. Nel primo caso, per trasmettere un pacchetto di L bits, si ha lo stesso tempo di trasmissione dato dalla (6.2). Infatti anche in questo caso è come se ogni stream avesse associato un canale di capacità C/m. Nel caso in cui la dimensione del pacchetto e quella dello slot coincidono, vale la relazione (6.1) ma bisogna aspettare un tempo pari a:

$$(m - 1)\ L/C$$

prima di poter trasmettere un altro pacchetto appartenente allo stesso stream.

Dalle relazioni ricavate sin ora si nota che lo *statistical multiplexing* è quello che garantisce il trasmission delay più piccolo. Ciò è dovuto al fatto che le risorse allocate ai clienti dagli schemi di multiplazione TDM ed FDM vengono sprecate nel caso in cui una sorgente non ha da trasmettere. Resta adesso da calcolare il *queueing delay.* Questa grandezza è più difficile da calcolare poiché è un parametro statistico e per il suo studio faremo uso della teoria delle code.

6.2 Sistemi a Coda

Un sistema a coda si può genericamente definire come un sistema in cui vi sono degli utenti (*clienti*) che arrivano e che vogliono utilizzare una risorsa finita (*servente*). I clienti (*customers*) richiedenti un dato servizio, sono generati nel tempo da una *sorgente.* Questi clienti entrando nel *queueing system* e formano una coda (= lista d'attesa). A certi istanti, un membro della coda viene scelto come prossimo cliente da servire, secondo una certa politica nota come *disciplina della coda;* per esempio la disciplina potrebbe essere *First In First Ou*t (FIFO), *Last In First Ou*t (LIFO), etc. Il servizio richiesto dal cliente viene quindi svolto dal *servente* e il *customer* può uscire dal sistema a coda. Questo processo è rappresentato in Figura 6.2.

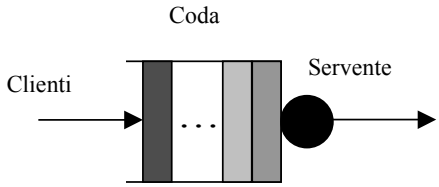


Figura 6.2: Schema di un sistema a coda.

Possono essere fatte svariate assunzioni sui vari elementi che costituiscono il queueing system. In generale, per caratterizzare un sistema a coda, deve essere specificata la *statistica dei tempi di interarrivo,* la *statistica dei tempi di servizio,* nonché la *disciplina usata per gestire la coda.*

In una rete di telecomunicazione i *customers* sono rappresentati dai pacchetti che arrivano e vengono assegnati ad un link per la trasmissione, mentre il *server* è rappresentato dalla linea di comunicazione. La *coda* corrisponde, invece, al buffer associato al link uscente dal nodo, tramite cui il pacchetto deve essere spedito.

Date le distribuzioni di probabilità dei tempi di interarrivo e dei tempi di servizio, il nostro obiettivo sarà quello di determinare le seguenti quantità:

- Il numero medio di clienti nel sistema;
- Il ritardo medio del generico cliente.

Per numero di clienti nel sistema si intende il numero di utenti presenti nella coda più il numero dei clienti che stanno usufruendo del servizio offerto dal sistema. Il ritardo di un cliente è costituito dal tempo di attesa in coda più il tempo di servizio.

Detta:

$p_n(t)$ = Probabilità che all'istante t vi siano n clienti nel sistema,

supposte note le informazioni statistiche necessarie per la determinazione delle probabilità $p_n(t)$ per ogni t, definito:

$N(t)$ = Numero medio di clienti nel sistema al tempo t,

si ha

$$E\{N(t)\} = N(t) = \sum_{n=0}^{+\infty} n \cdot p_n(t) \tag{6.3}$$

Osserviamo che sia $p_n(t)$ che $E\{N(t)\}$ dipendono dal tempo e dalla distribuzione delle probabilità al tempo $t = 0$, ossia, $\{ p_0(0) , p_1(0) , p_2(0) , \dots , p_n(0), \dots\}$. I sistemi con cui avremo a che fare saranno caratterizzati dal fatto di raggiungere una condizione di equilibrio, nel senso che:

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

$$N = \sum_{n=0}^{+\infty} np_n = \lim_{t\to\infty} \overline{N}(t) \tag{6.4}$$

dove p_n e N sono indipendenti dalla distribuzione delle probabilità iniziale.

Detta $N(t)$ una funzione di campionamento del numero dei clienti nel sistema, definiamo media temporale ditale funzione nell'intervallo [0, t] la grandezza:

$$N_t = \frac{1}{t}\int_0^t N(\tau)\, d\tau \tag{6.5}$$

Si definisce **ergodico,** un sistema per il quale vale la relazione:

$$\lim_{t\to\infty} N_t = \lim_{t\to\infty} \overline{N}(t) = N \tag{6.6}$$

Notiamo che se un sistema è ergodico, la media statistica e quella temporale coincidono.

Consideriamo ora il ritardo medio del generico cliente. Supposta nota la distribuzione di probabilità di ritardo di ciascun cliente, siamo in grado di calcolare il ritardo medio di ogni cliente. Sia $E\{T_k\}$ il ritardo medio del k-esimo cliente.

Nel caso in cui il sistema converga ad un valore stazionario per $k \to \infty$, si ha che il ritardo medio del generico cliente sarà:

$$T = \lim_{t\to\infty} \overline{T}_k \tag{6.7}$$

Se il sistema è ergodico si ha che:

$$T = \lim_{t\to\infty} \overline{T}_k = \lim_{t\to\infty} \frac{1}{k}\sum_{i=1}^{k} T_i \tag{6.8}$$

dove T_i rappresenta il ritardo dell'i-esimo cliente.

1.2.1 Applicazione dei sistemi a coda nelle reti

I sistemi a coda possono essere usati per modellare sia le reti a commutazione di pacchetto che le reti a commutazione di circuito. Nelle reti a commutazione di pacchetto i *clienti* sono i pacchetti da trasmettere. Supponendo che i pacchetti (= messaggi) abbiano dimensione variabile con media di L bit e che il canale trasmissivo abbia capacità di

trasmissione pari a C bit/s (singolo servente), il tempo medio di trasmissione di un pacchetto è dato da:

$$\frac{1}{\mu} = \frac{L}{C} \tag{6.9}$$

dove μ (rate medio di servizio espresso in pacchetti/secondo) è il numero medio di pacchetti trasmessi dal servente in un secondo. Detto λ il numero medio di arrivi in un secondo, si definisce **fattore di utilizzazione del servente** o intensità di traffico (misurata in Erlang):

$$\rho = \frac{\lambda}{\mu} \tag{6.10}$$

e sostituendo la (6.9) nella (6.10) si ha:

$$\rho = \frac{\lambda \cdot L}{C} \tag{6.11}$$

Il numeratore della (6.11) rappresenta il carico medio nella rete (λL [bit/s]), mentre il denominatore rappresenta la capacità di trasmissione della rete [bit/s]. Dunque il parametro ρ fornisce quantitativamente la misura di *quanto è caricato il sistema.* Se tale parametro è maggiore di 1, il sistema non riesce a smaltire il carico, poiché il numero medio di arrivi è superiore al numero medio di partenze (instabilità della relativa coda di attesa).

6.2 Teorema di Little

Il teorema di Little stabilisce che tra *N* e *T* intercorre una dipendenza lineare. Detta λ la costante di proporzionalità risulta:

$$N = \lambda T \tag{6.12}$$

dove:

λ **=** *tasso medio degli arrivi*

ed è dato dalla relazione:

$$\lambda = \lim_{t \to \infty} \frac{valore\ medio\ di\ arrivi\ in\ [0,t]}{t} \tag{6.13}$$

Definiamo rispettivamente:

- $\alpha(t)$ **=** Numero degli arrivi nell'intervallo [0, *t*],
- $\beta(t)$ **=** Numero delle partenze nell'intervallo [0, *t*].

Supposto $N(0) = 0$, dalla definizione di $\alpha(t)$ e $\beta(t)$ risulta chiaramente:

$$N(t) = \alpha(t) - \beta(t) \tag{6.14}$$

dove $N(t)$ indica il numero di clienti presenti nel sistema all'istante *t*. Indichiamo con t_i l'istante in cui l'i-esimo utente arriva nel sistema, mentre con T_i il tempo speso nel sistema dall' i-esimo utente. Consideriamo ora ad un istante *t*, l'area racchiusa tra le due curve di Figura 6.3 che, in virtù della (6.14), è pari a:

$$\int_0^t N(\tau)d\tau$$

Ma d'altro canto risulta:

$$\int_0^t N(\tau)d\tau = \sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i) \tag{6.15}$$

Dividendo ambo i membri della (6.15) per *t* si ottiene:

$$\frac{1}{t}\int_0^t N(\tau)d\tau = \frac{\alpha(t)}{t} \frac{\sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i)}{\alpha(t)} \tag{6.16}$$

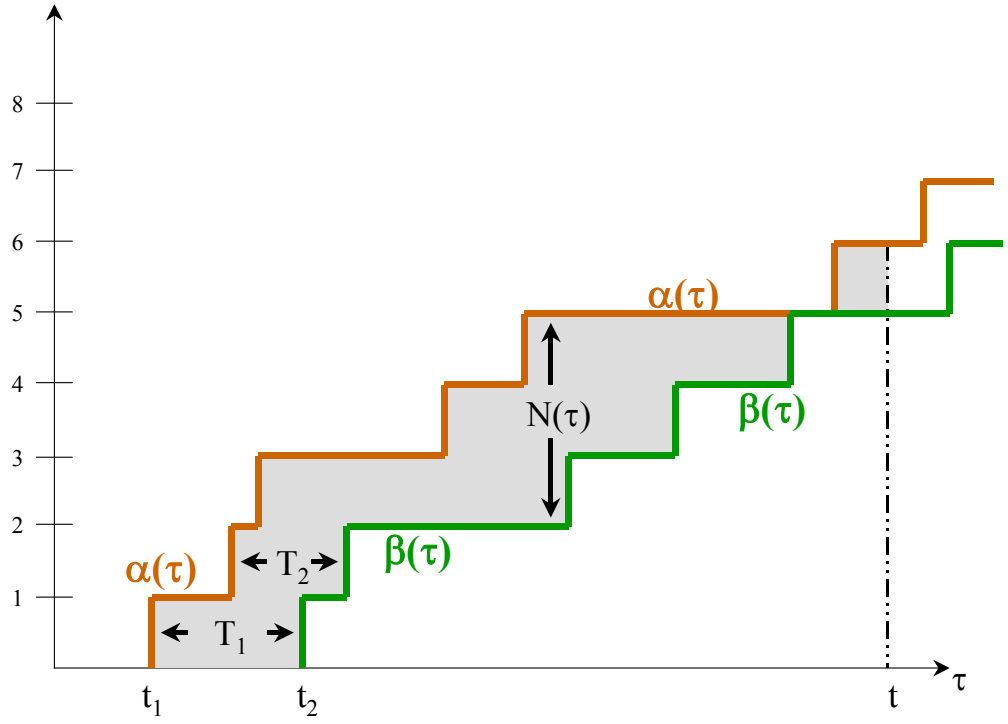


Figura 6.3: Schematizzazione degli arrivi e delle partenze in un sistema.

Osserviamo che il primo membro della (6.16) non è altro che la media temporale in [0, *t*] del numero di clienti presente nel sistema. D'altro canto si ha che:

$$\frac{\alpha(t)}{t} = \lambda_t$$

dove λ_t rappresenta la media temporale del tasso degli arrivi nell'intervallo [0, *t*]. Infine notiamo che risulta:

$$\frac{\sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i)}{\alpha(t)} = T_t$$

dove T_t rappresenta la media temporale del tempo che un cliente spende nel sistema nell'intervallo [0, *t*]. In virtù di queste osservazioni possiamo scrivere:

$$N_t = \lambda_t T_t \qquad (6.17)$$

Se supponiamo che si abbia (processo ergodico):

$$\lim_{t\to\infty} N_t = N$$
$$\lim_{t\to\infty} \lambda_t = \lambda$$
$$\lim_{t\to\infty} T_t = T$$

segue subito la formula di Little $N = \lambda T$.

E' importante notare che T_t include il tempo speso nel sistema da tutti i clienti arrivati tra 1 e $\beta(t)$, ma tralascia il tempo speso dai clienti ancora nel sistema all'istante *t*. Se si suppone che $N_t \to N < \infty$, (il che implica che tutti i clienti sono serviti in un tempo finito) l'effetto dovuto ai clienti presenti nel sistema all'istante *t* diviene via via trascurabile, ed al crescere di *t*, così che T_t può effettivamente essere interpretato come la media temporale del tempo di sistema.

L'importanza del teorema di Little deriva dalla sua generalità. Esso può essere applicato ad un qualsiasi sistema a coda che raggiunga una condizione di equilibrio statistico. La cosa importante nell'applicare il teorema di Little è quella di interpretare nel modo appropriato N, λ e T. Prendiamo in esame la parte di attesa della coda: la lunghezza media coda è data dalla relazione:

$$N_Q = \lambda W \qquad (6.18)$$

dove *W* è il tempo medio di attesa in coda.

Analogamente, applicando il teorema di Little nella parte di uscita del sistema a coda (= lato servente), si ha che il numero medio di pacchetti in trasmissione ρ è dato dalla frequenza di arrivo dei pacchetti λ per il tempo medio di trasmissione, $\overline{X} = \frac{1}{\mu}$:

$$\rho = \lambda \overline{X} \qquad (6.19)$$

Il parametro ρ è chiamato ***fattore di utilizzazione*** del sistema, perché è definito come il numero medio di pacchetti entranti nel servente per il tempo medio di trasmissione, ovvero rappresenta la porzione di tempo per il quale il sistema è occupato nella trasmissione di un pacchetto.

Gli **aspetti caratterizzanti** un sistema a coda sono:

- *Distribuzione del processo di arrivo delle richieste.* Si utilizza spesso anche la distribuzione del tempo che intercorre tra gli arrivi di richieste successive. Il numero medio o *tasso medio di arrivo delle richieste* in un secondo è generalmente indicato con λ,
- *Distribuzione della durata o lunghezza di un arrivo*,
- *Capacità del buffer per la memorizzazione delle richieste* in attesa di servizio ovvero la

capacità della coda,

- *Capacità di fornire il servizio* ovvero *numero di serventi*,
- *Tempo di servizio:* tempo che occorre al servente per soddisfare la richiesta in servizio. Può essere fisso o caratterizzato statisticamente.
- *Disciplina di coda.* La disciplina con cui si regola l'accesso al servizio può avere varie modalità:
  - Disciplina **FIFO** (*First In First Ou*t): la prima richiesta che raggiunge il sistema è la prima ad essere servita;
  - Disciplina **LIFO** (*Last In First Out*): l'ultima richiesta arrivata è la prima ad essere servita;
  - Disciplina **Random:** le richieste sono servite in modo casuale con distribuzione uniforme;
  - Disciplina con **priorità.** In questo caso si possono adottare due strategie:
    - **Priorità con interruzione di** servizio: all'arrivo di una richiesta con priorità maggiore rispetto a quella della richiesta attualmente servita, il servizio viene concesso immediatamente, togliendolo a quella che ne usufruiva.
    - **Priorità senza interruzione di** servizio: all'arrivo di una richiesta con priorità maggiore rispetto a quella della richiesta attualmente servita, il sistema termina il servizio in atto e successivamente serve la richiesta con priorità maggiore.

I parametri più importanti per lo studio e la valutazione della bontà di un sistema a coda sono:

- **Tempo di attesa in coda:** tempo che intercorre da quando una richiesta entra in coda a quando viene servita. Questo parametro è ovviamente influenzato dal tempo medio di servizio di ciascuna richiesta, dalla frequenza degli arrivi delle richieste e dalla lunghezza della coda;
- **Tempo trascorso nel sistema:** somma del tempo trascorso dalla richiesta all'interno della coda e del tempo di servizio a questa dedicato;
- **Numero di richieste all'interno del sistema:** costituto dalla somma del numero delle richieste presenti in coda e all'interno del/dei servente/i;
- **Tempo di occupato/libero di ogni servente.** E' intuitivo che aumentando il numero di serventi, diminuisce il numero di richieste presenti in coda. Ogni servente implica, però, una spesa che deve essere ripagata da un suo utilizzo efficiente e continuativo. Un numero troppo elevato di serventi utilizzati per una piccola frazione del tempo complessivo, può diminuire molto i tempi di coda, ma aumentare considerevolmente i costi diminuendo, inoltre, l'efficienza del sistema.

6.3 Nomenclatura per i sistemi a coda: la notazione di Kendall

La nomenclatura, introdotta da Kendall per identificare i vari tipi di sistemi a coda fa uso di *5* simboli separati dal carattere "/":

1. La *prima lettera* indica la natura del processo degli arrivi. I valori tipici sono:
   a. M: *memoryless,* indica che il processo degli arrivi è un processo di Poisson (distribuzione di probabilità dei tempi di interarrivo di tipo esponenziale e quindi senza memoria).

b. G: *general,* indica che il processo degli arrivi è caratterizzato da una distribuzione di probabilità generale. In questo caso serve conoscere i momenti del I° e del 2° ordine della distribuzione dei tempi di interarrivo.
c. D: *deterministic,* indica che il processo degli arrivi è deterministico e cioè con tempi di interarrivo costanti. Cioè i tempi di servizio sono costanti.

2. La *seconda lettera* indica la natura della distribuzione di probabilità dei *tempi di servizio.* I valori possibili, anche in questo caso, sono M, G, D e il significato è uguale a quello spiegato precedentemente con l'unica differenza che tali simboli si riferiscono alla distribuzione di probabilità del tempo di servizio di una richiesta.

3. Il terzo simbolo indica il numero di serventi del sistema a coda.

4. Il *quarto simbolo* indica il *numero massimo di clienti nel sistema.* Questo simbolo potrebbe non essere presente e per default è infinito.

5. Il *quinto simbolo* indica il *numero massimo di sorgenti.* Anche questo simbolo potrebbe non essere presente e per default è infinito. Ogni sorgente (nel nostro caso sorgente di pacchetti) può immettere un solo pacchetto alla volta, e potrà produrne un altro solo quando il precedente è stato spedito.

Un sistema a coda di tipo M/M/1 è dunque caratterizzato da un solo servente (terzo simbolo), i clienti arrivano secondo un processo di Poisson con tasso medio λ e la distribuzione dei tempi di servizio è esponenziale con valor medio $1/\mu$ sec. Il numero massimo di clienti nel sistema e il numero massimo di sorgenti attive nel sistema è infinito. I sistemi caratterizzati da processo di interarrivo e processo delle partenze a distribuzione esponenziale sono i più semplici da studiare e anche i più conservativi (come provato dal confronto con i sistemi M/G/1). I sistemi M/M/N possono essere studiati e risolti con la teoria delle *catene di Markov.* E' possibile, in particolare, calcolare la probabilità p_n che nel sistema vi siano *n* utenti e tramite essa è possibile ricavare il ***numero medio di utenti*** nel sistema (*N*). Si ha infatti che:

$$N=\sum_{n=0}^{\infty} n \cdot p_n \tag{6.20}$$

Sfruttando il teorema di Little è facile ricavare anche il ***tempo medio trascorso da un utente*** nel sistema, T:

$$\mathrm{T} = \frac{N}{\lambda} \tag{6.21}$$

In modo analogo è possibile ricavare il numero medio di utenti in coda (N_Q) ed il tempo medio di attesa in coda di un utente (W).

Qui di seguito si analizzeranno i sistemi a coda più importanti del tipo M/M/1, M/M/N, M/M/∞, M/M/1/N, M/M/N/N e M/G/1.

6.4 Catene di Markov

I processi di Markov rivestono una particolare importanza nella teoria delle reti di telecomunicazione. In questa sezione sono elencate le proprietà generali delle catene di Markov. Consideriamo lo stato $X(t)$ (= numero di messaggi in un buffer di trasmissione) di un sistema ad un certo istante t e supponiamo che $X(t)$ possa assumere valori discreti da un insieme numerabile $\{a_1, a_2, \ldots\}$.

L'occupazione degli stati nel tempo è un processo aleatorio che è molto importante per valutare il comportamento di un sistema di comunicazione. I processi di Markov possono essere divisi in due classi:

1. Processi di Markov tempo-discreto.
2. Processi di Markov tempo-continuo.

6.4.1 Catene di Markov tempo-continue e tempo-discrete

Il processo N(t) (utenti nel sistema all'istante t) può essere studiato facendo uso delle catene di Markov tempo continue, dove la variabile t assume valori continui. E' possibile anche adottare la teoria (più semplice) delle catene di Markov tempo-discrete (la variabile t è discreta) utilizzando questo semplice artificio: consideriamo gli istanti di tempo

$$0, \delta, 2\delta, \ldots, k\delta, \ldots$$

dove δ è un numero positivo *piccolo.* Indichiamo con:

N_k = numero di utenti nel sistema all'istante $k\delta$, $N(k\delta)$.

Poiché N(t) è una catena di Markov tempo continua e $N_k = N(k\delta)$, si vede che:

$$\{ N_k \mid k = 0, 1, 2, \ldots \}$$

è una catena di Markov tempo-discreta.

Detto questo, diamo una definizione più rigorosa alle catene di Markov tempo-discrete. Sia $\{X_n \mid n = 0, 1,\ldots\}$ un processo stocastico tempo discreto che assume valori interi non negativi. Gli stati in cui il processo può trovarsi sono: $i = 0, 1, \ldots$

Il processo è una **catena di Markov** se, c'è una probabilità fissa P_{ij} che il processo si troverà prossimamente nello stato i supponendo che si trovi nello stato j e tale probabilità è indipendente dalla storia che ha portato il processo nello stato i. Tale concetto è riassunto nelle equazioni sottostanti:

$$P_{ij} = P\{X_{n+1} = j \mid X_n = i \ , X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = P\{X_{n+1} = j \mid X_n = i\} \qquad \forall\, n > 0, i_{n-1}, \ldots, i_0, i, j \tag{6.22}$$

Le P_{ij} così definite sono dette *probabilità di transizione* dallo stato i allo stato j.

Ovviamente, essendo probabilità risulterà che:

$$P_{ij} \geq 0, \quad \sum_{j=0}^{\infty} P_{ij} = 1, \quad i = 0, 1, ... \tag{6.23}$$

Si definisce matrice delle probabilità di transizione:

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \dots \\ P_{10} & P_{11} & P_{12} & \dots \\ \dots & \dots & \dots & \dots \\ P_{q0} & P_{q1} & P_{q2} & \dots \end{bmatrix}$$

Possono essere definite anche le probabilità di transizione ad n passi:

$$P_{ij}^{n} = P\{X_{n+m} = j \mid X_m = i\} \quad n \geq 0, i \geq 0, j \geq 0$$

e può essere calcolata la matrice di transizione ad n passi P^n.

Date le probabilità di transizione ad n passi vale l'equazione di *Chapman-Kolmogorov:*

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^{n} \cdot P_{kj}^{m}, \qquad n, m \geq 0 \quad i, j \geq 0 \tag{6.24}$$

Introduciamo adesso alcune definizioni.

Si dice che due stati *i* e *j* **comunicano** tra loro se esistono due indici *n* e *n*' tali che:

$$P_{ij}^{n} > 0$$
$$P_{ji}^{n'} > 0$$

Se tutti gli stati comunicano fra loro, la catena di Markov si dice ***irriducibile.***

Una catena di Markov si dice aperiodica se esiste uno stato in cui è possibile ritornare solo dopo un numero di passi multiplo di d > 2.

Una distribuzione di probabilità $\{p_j \mid j \geq 1\}$ si dice essere una ***distribuzione stazionaria*** per la catena di Markov se:

$$p_j = \sum_{i=0}^{\infty} p_i \cdot P_{ij}, \quad j \geq 0 \tag{6.25}$$

Per catene di Markov **irriducibili e aperiodiche** si ha che:

$$p_j = \lim_{n\to\infty} P_{ij}^n \ , \quad j \geq 0 \tag{6.26}$$

dove p_j rappresenta la probabilità, a regime, che il sistema si trovi in quello stato; essa rappresenta dunque anche la porzione di tempo in cui il processo *visita* in media lo stato j; $1/p_j$ rappresenta il ***tempo medio di ricorrenza,*** ovvero il numero atteso di transizioni tra due successive visite dello stato j (se $p_j = 0$, il tempo medio di ricorrenza è infinito).

Si può inoltre dimostrare che in una catena di Markov irriducibile e aperiodica possono verificarsi due possibilità:

1. $p_j = 0$ per tutti gli stati $j \geq 0$. In questo caso la catena di Markov non ha distribuzione stazionaria (è il caso di un sistema M/M/1 in cui il tasso medio di arrivo è maggiore del tasso medio di servizio).
2. $p_j > 0$ per tutti gli stati $j \geq 0$. In questo caso la distribuzione di probabilità:

$$p_j = \sum_{i=0}^{\infty} p_i \cdot P_{ij} \ , \quad j \geq 0$$

è ***l'unica*** distribuzione stazionaria della catena.

La distribuzione stazionaria di una catena di Markov, se esiste, può essere calcolata attraverso le ***equazioni di bilanciamento globale.*** Esse derivano dalla (6.23). Si ha infatti che:

$$\sum_{i=0}^{\infty} P_{ij} = P_{jj} + \sum_{i=0,i\neq j}^{\infty} P_{ij} = 1 \ \Rightarrow \sum_{i=0,i\neq j}^{\infty} P_{ij} = 1 - P_{jj}$$

moltiplicando ambo i membri per p_j si ha:

$$p_j \cdot \sum_{i=0,i\neq j}^{\infty} P_{ji} = p_j - p_j \cdot P_{jj} \ .$$

Sfruttando la (6.25) si ha che:

$$p_j \cdot \sum_{i=0,i\neq j}^{\infty} P_{ji} = \sum_{i=0}^{\infty} p_i \cdot P_{ij} - p_j \cdot P_{jj} \Rightarrow p_j \cdot \sum_{i=0,i\neq j}^{\infty} P_{ji} = \sum_{i=0}^{\infty} p_i \cdot P_{ij} - (p_i \cdot P_{ij})_{i=j}$$

da cui si ottiene:

$$p_j \cdot \sum_{i=0,i\neq j}^{\infty} P_{ji} = \sum_{i=0,i\neq j}^{\infty} p_i \cdot P_{ij} \tag{6.27}$$

La (6.27) indica che in condizioni di equilibrio, la probabilità di una transizione in partenza dallo stato j eguaglia la probabilità di una transizione in arrivo allo stato j.

Generalizzando il discorso ad un insieme di stati S si ottengono le equazioni di bilanciamento globale:

$$\sum_{j \in S} p_j \sum_{j \notin S} P_{ji} = \sum_{j \notin S} p_i \sum_{j \in S} P_{ij} \tag{6.28}$$

La (6.28) indica che la probabilità che si abbia una transizione in partenza da S è pari alla probabilità che si abbia una transizione verso S.

6.4.2 Processi di nascita e morte

I processi di nascita e morte sono catene di Markov in cui due stati successivi differiscono solo di una unità, le transizioni dal generico stato k sono permesse solo verso gli stati adiacenti k + 1 e k - 1. Tali processi sono ideali per caratterizzare l'evolvere di una coda. In essa, infatti, gli utenti arrivano uno alla volta e si accodano per ricevere il servizio. Nella trattazione seguente si fa sempre riferimento alla trattazione tempo-discreta equivalente.

Condizione necessaria e sufficiente affinché la catena sia irriducibile è che:

$$P_{i,i+1} > 0 \text{ e } P_{i+1,i} > 0 \quad \text{per ogni i}$$

Considerando l'insieme di stati S = {0, 1, 2, ..., n}, le equazioni di bilanciamento parziali (6.28) danno:

$$p_n P_{n,n+1} = p_{n+1} P_{n+1,n} \quad n = 0, 1, \ldots \tag{6.29}$$

ovvero, la probabilità di una transizione dallo stato n allo stato n + 1 è pari alla probabilità di una transizione dallo stato n + 1 allo stato n.

Generalizzando la (6.29) si ottengono le ***equazioni di bilanciamento dettagliate:***

$$p_j \cdot P_{ji} = p_i \cdot P_{ij} \quad i, j \geq 0 \tag{6.30}$$

Queste equazioni permettono di calcolare facilmente la distribuzione stazionaria delle probabilità di stato $\{p_j \mid j \geq 0\}$. Osserviamo che ***non sempre*** valgono le equazioni di bilanciamento dettagliate per una data catena di Markov irriducibile e aperiodica. Un modo per verificare la loro validità è ipotizzarne la validità e tentare di risolvere il sistema che ne viene fuori per ottenere le probabilità pj con la condizione al contorno che:

$$\sum_j p_j = 1$$

Esistono due possibilità:

1. l'assunzione non è vera, ed il sistema di equazioni è inconsistente;
2. l'assunzione è vera, e la distribuzione di probabilità $\{p_j \mid j \geq 0\}$ trovata è l'unica distribuzione stazionaria del sistema (sicuramente essa soddisfa anche le equazioni di bilanciamento globali).

Valgono le equazioni di bilanciamento seguenti:

$$p_j \sum_{j \in S_j^m} P_{ji} = \sum_{j \in S_j^m} p_j \cdot P_{ji} \quad m = 1, 2, ..., k \tag{6.31}$$

Le (6.31) vengono dette ***equazioni di bilanciamento parziali.*** Si può dimostrare che se la $\{p_j \mid j \geq 0\}$ risolve un insieme di equazioni di bilanciamento parziali, allora risolve anche le equazioni di bilanciamento globali, e quindi è l'unica distribuzione stazionaria della catena di Markov irriducibile e aperiodica. E' quindi importante individuare il giusto insieme di equazioni parziali soddisfatte dalla distribuzione stazionaria per calcolare quest'ultima nel modo più semplice possibile.

Un metodo pratico per risolvere le catene di nascita-morte nel tempo-continuo è procedere come segue, facendo riferimento al generico processo nascita morte indicato in Figura 6.4 dove λ_i è il tasso medio di nascita dallo stato *i* e μ_i è il tasso medio di morte dallo stato *i*.

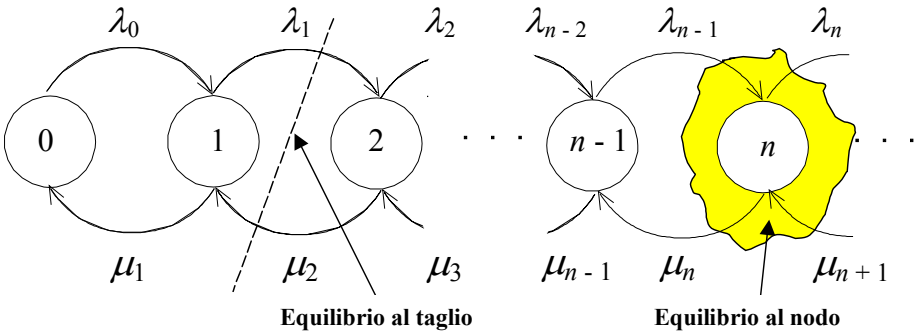


Figura 6.4: Generico processo di nascita morte.

Nell'ipotesi in cui la condizione "$\exists\ k\ tale\ che: \forall\ n \geq k,\ \lambda_n < \mu_n$" (***condizione di ergodicità***) è soddisfatta, esiste un regime stazionario per la catena. A regime le probabilità $P_n(t)$ (= probabilità di essere nello stato *n* al tempo *t*) non dipendono dal tempo $\Rightarrow dP_n(t)/dt = 0$ e $P_n(t) = P_n$. Allora valgono le seguenti equazioni di equilibrio ai nodi per caratterizzare le P_n:

$$\begin{aligned}
&\lambda_0 P_0 = \mu_1 P_1 \\
&(\lambda_n + \mu_n) P_n = \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1}, \quad \forall n > 1 \\
&\sum_{n=0}^{\infty} P_n = 1 \quad (normalizzazione)
\end{aligned}$$

La *n*-esima equazione può essere interpretata come ***la condizione di equilibrio*** tra il flusso entrante (= $\lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1}$) ed il flusso uscente (= $\lambda_n P_n + \mu_n P_n$) attorno una generica superficie che circonda lo stato *n*. In Figura 6.4 è rappresentato l'equilibrio attorno al generico nodo n-esimo.

Analogamente è possibile risolvere le probabilità P_n facendo gli ***equilibri ai tagli*** ($P_{n-1}\lambda_{n-1} = \mu_n P_n$) insieme alla condizione di normalizzazione. In Figura 6.4 è rappresentato l'equilibrio al taglio per $n = 2$.

Sistemi M/M/1

I sistemi M/M/1 sono, come preannunciato, i sistemi a coda più semplici da studiare. Essi sono processi nascita morte {N(*t*) | $t \geq 0$} (numero di clienti nel sistema all'istante *t*) in cui i tempi di interarrivo e di servizio sono distribuiti esponenzialmente, rispettivamente con tassi medi λ e μ. Nella Figura 6.5 è rappresentato in forma grafica il sistema M/M/1, in cui gli archi sono etichettati con i tassi medi di transizione da uno stato all'altro.

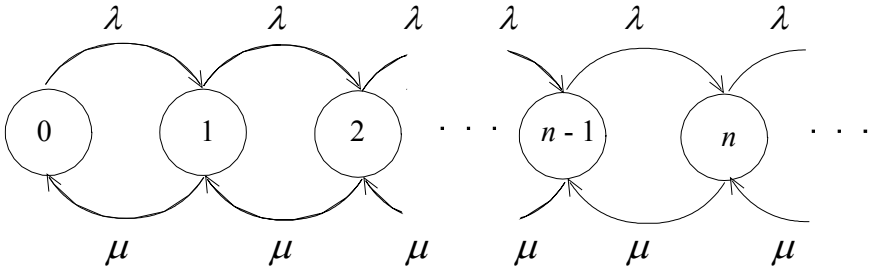


Figura 6.5: Rappresentazione in forma grafica di un sistema M/M/1.

La soluzione di questa catena è un caso particolare di quella in Figura 6.4.

Nel paragrafo precedente abbiamo visto come tale processo continuo può anche essere approssimato tramite una catena di Markov tempo-discreta $N_k = N(k\delta)$. $\{N_k \mid k = 0,1,2\}$ o può essere risolto imponendo gli equilibri sulla catena tempo-continua.

6.4.3 Relazione tra carico e throughput (in sistemi singolo servente)

Per un sistema single-server il *throughput* γ rappresenta il traffico smaltito nell'unità di tempo. Osserviamo che il throughput sarebbe uguale al rate medio di servizio (μ) se la coda non fosse mai vuota, per cui si ha che:

$$\gamma = \mu(1 - p_0) \tag{6.32}$$

Nei sistemi G/G/1, $\rho = (1 - p_0)$ è *l'intensità di traffico* che coincide con la probabilità che il servente sia occupato.

Nel caso ideale di sistema M/M/1, sostituendo a p_0 il valore ottenuto precedentemente, e cioè 1 - ρ ($\rho = \lambda/\mu$), si ha che:

$$\gamma = \mu(1 - (1 - \rho)) = \mu\rho = \lambda$$

Per sistemi con buffer infinito, infatti, tutti i clienti che entrano nel sistema, prima o poi verranno serviti, dunque il throughput è uguale alla frequenza degli arrivi. Il *throughput normalizzato γ/μ è il fattore di utilizzazione* ρ del servente ($\rho < 1$ Erlang per la stabilità in presenza di un unico servente).

6.5 Sistemi M/G/1

I sistemi M/G/1 sono sistemi a singolo servente in cui i clienti arrivano secondo un processo di Poisson con tasso medio λ e i tempi di servizio seguono una distribuzione generica (non necessariamente esponenziale come accadeva nei sistemi M/M/1).

Sia X_1 il tempo di servizio dell'i-esimo cliente. Assumiamo che le variabili casuali (X_1, X_2, …) siano identicamente distribuite e indipendenti dai tempi di interarrivo.

Indichiamo con:

$$\overline{X} = E\{X\} = \frac{1}{\mu}$$

il **tempo di servizio medio** (momento del primo ordine) e con:

$$\overline{X^2} = E\{X^2\}$$

il **momento del secondo ordine** del tempo di servizio.

Si può dimostrare che, per i sistemi M/G/1 vale la ***formula di Pollaczek-Khintchine:***

$$W = \frac{\lambda \cdot \overline{X^2}}{2(1-\rho)} \tag{633}$$

dove *W* è il tempo medio di attesa in coda, mentre risulta:

$$\rho = \frac{\lambda}{\mu} = \lambda \cdot \overline{X}$$

Il tempo medio di attesa nel sistema sarà pari alla somma del tempo medio speso nel servente più il tempo medio di attesa in coda dato dalla (6.33), dunque:

$$T = \overline{X} + \frac{\lambda \cdot \overline{X^2}}{2(1-\rho)} \tag{6.34}$$

E' facile adesso calcolare il numero medio di clienti in attesa in coda e nel sistema, applicando il teorema di Little, si ottiene che:

$$N_Q = \frac{\lambda^2 \cdot \overline{X^2}}{2(1-\rho)}$$

$$N = \rho + \frac{\lambda^2 \cdot \overline{X^2}}{2(1-\rho)}$$

Calcoliamo ora il tempo medio di attesa in coda per sistemi M/M/1 come caso particolare della formula (6.33). Ricordiamo che il momento del secondo ordine (o valore quadratico medio) è dato dalla somma della varianza più il valor medio al quadrato:

$$\overline{X^2} = \sigma^2 + \overline{X}^2$$

In una distribuzione esponenziale $\sigma^2 = 1/\mu^2$, dove $1/\mu$ è il valore medio, quindi:

$$\overline{X^2} = \frac{1}{\mu^2} + \frac{1}{\mu^2} = \frac{2}{\mu^2}$$

Sostituendo tale valore nella (6.33) si ha:

$$W = \frac{\rho}{\mu \cdot (1-\rho)} = \frac{\rho}{\mu - \lambda} \tag{6.35}$$

Questo risultato indica che la formula (6.33) per i sistemi M/G/1 vale anche per i sistemi M/M/1, che, in fondo, sono un caso particolare di sistema M/G/1.

Analizziamo adesso un altro sotto caso dei sistemi M/G/1: i sistemi M/D/1. In tali sistemi il tempo di servizio è ***deterministico.*** Un caso pratico per questo tipo di sistemi può essere dato dalle reti in cui la lunghezza dei pacchetti è costante, quindi il tempo di servizio è costante per tutti i pacchetti. La varianza del tempo di servizio è dunque nulla ($\sigma^2 = 0$) quindi:

$$\overline{X^2} = 0 + \frac{1}{\mu^2} = \frac{1}{\mu^2}$$

Dalla formula (6.33) si ottiene:

$$W = \frac{\rho}{2\mu \cdot (1-\rho)} \tag{6.36}$$

Nel caso M/D/1, si ha il valore minimo del momento del secondo ordine, e quindi anche W, T, N_Q, e N hanno il valore minimo. In particolare, W e N_Q sono la metà dei corrispondenti valori per sistemi M/M/l con uguale tasso medio di servizio e di arrivo.

I valori di T e N per i sistemi M/D/l sono la metà dei corrispondenti in M/M/1 se $\rho \approx 1$ Erlang e sono uguali ai corrispondenti M/M/1 per ρ piccolo. Ciò accade perché il tempo di servizio è circa lo stesso nei due casi e, per ρ piccolo, il tempo che incide di più è quello di servizio, mentre per ρ grande il termine più pesante è il tempo di attesa.

In genere, i valori di T e N per sistemi M/G/1 sono intermedi tra quelli di M/D/1 (che corrispondono al caso migliore) e quelli di M/M/1 (che corrispondono al caso peggiore).

Come conseguenza di quanto ricavato, osserviamo che, usare il multiplexing statistico suddividendo l'asse temporale in slots (in cui la durata dello slot coincide con il tempo di trasmissione di un pacchetto), implica il minimo tempo medio di attesa dei pacchetti in coda. Inoltre, bisogna evidenziare anche che, se per un dato sistema non è possibile conoscere il momento del primo e del secondo ordine del tempo di servizio, è giustificato l'utilizzo di un sistema M/M/1 come modello analitico, in quanto esso porta eventualmente al sovra-dimensionamento del sistema.

Dimostriamo adesso la formula di *Pollaczek-Khintchine* (PK), facendo riferimento al caso particolare della disciplina di servizio FIFO. Tale dimostrazione farà uso della definizione del ***tempo residuo di servizio****.*

Def:

> *Si definisce* ***tempo residuo di servizio*** *relativamente all'i-esimo cliente, il tempo (rimanente) necessario affinché l'utente in servizio all'arrivo del cliente i-esimo, esaurisca il servizio stesso.*

Siano:

- W_i : Tempo di attesa in coda dell'i-esimo cliente.
- R_i : Tempo di servizio residuo visto dall'i-esimo cliente. Cioè, se nel servente è presente il cliente j-esimo quando il cliente i arriva, con R_i indichiamo il tempo rimanente affinché il cliente j completi il servizio. Se non vi sono clienti nel sistema quando i arriva (cioè il sistema è vuoto), R_i sarà zero.
- X_i : Tempo di servizio dell'i-esimo cliente.
- N_i : Numero di clienti trovati in attesa in di servizio nella coda all'arrivo dell'i-esimo cliente.

Si ha che il tempo di attesa in coda per l'i-esimo cliente è pari al tempo di servizio residuo (del cliente già nel servente quando i arriva) più la somma dei tempi di servizio degli N_i utenti in coda prima dell'arrivo di i:

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j$$

Essendo le variabili N_i, X_{i-1}, …, X_{i-Ni} indipendenti, si ha che:

$$E\{W_i\} = E\{R_i\} + E\left\{ \sum_{j=i-N_j}^{i-1} E\{X_i\} \right\} = E\{R_i\} + \overline{X} \cdot E\{N_i\}$$

Facendo il limite ad ambo i membri per i che tende all'infinito si ha:

$$W = R + \frac{1}{\mu} N_Q \qquad (6.37)$$

dove R è il *tempo residuo medio* ed è definito come:

$$R = \lim_{i \to \infty} E\{R_i\}$$

Usando la formula di Little per la parte di attesa, si ha che: $N_Q = \lambda W$, sostituendo nella (6.37) si ha:

$$W = R + \frac{1}{\mu} \cdot \lambda \cdot W \Rightarrow W \cdot (1 - \rho) = R$$

da cui:

$$W = \frac{R}{(1-\rho)} \qquad (6.38)$$

Per ottenere dunque il tempo medio di attesa in coda bisogna trovare il tempo medio residuo e sostituirlo nella (6.38). E' possibile risolvere questo problema tramite il teorema della vita residua che collega la distribuzione del tempo residuo con la distribuzione degli intertempi di arrivo e con quella di servizio. E' infatti possibile dimostrare che vale la seguente formula che fa riferimento ai valori medi:

$$R = \frac{1}{2} \cdot \lambda \cdot \overline{X^2}$$

Sostituendo dunque nella (6.38) si ha:

$$W = \frac{\lambda \cdot \overline{X^2}}{2(1-\rho)}$$

che è proprio la ***formula di Pollaczen-Khintchine(P-K).***

Si noti che un sistema M/G/1 con $\rho < 1$ Erlang può presentare tempi di attesa infiniti se il momento del secondo ordine tende ad infinito (es. questo vale in certe condizioni ad esempio con la distribuzione di Pareto del tempo di servizio). Quello che succede in questo caso è che una piccola quantità di utenti hanno un tempo di servizio molto lungo. Durante questo ampio intervallo di tempo, un numero molto elevato di clienti arrivano nel sistema e vengono accodati subendo dunque un elevato ritardo.

6.6 Sistemi a coda con priorità

Consideriamo un sistema M/G/1 in cui i clienti sono divisi in classi di priorità decrescente. Supponiamo inoltre che le priorità vengano gestite senza *preemption.* Cioè al cliente sotto servizio è permesso di completare il servizio senza interruzione anche se arriva un cliente a più alta priorità. Una coda separata è mantenuta per ogni classe di priorità. Quando un server diventa disponibile, viene servito il primo cliente in attesa nella coda non vuota a più alta priorità.

Indichiamo con:

- λ_k : il tasso medio di arrivo degli utenti di classe k,
- $\overline{X_k}$ = 1 = $1/\mu_k$: il momento del primo ordine del tempo di servizio relativo alla classe k,
- $\overline{X_k^2}$: il momento del secondo ordine del tempo di servizio relativo alla classe k.

Calcoleremo adesso il tempo medio di attesa in coda per le varie classi di priorità.

Indichiamo con:

- $N_Q^{(k)}$: il numero medio di utenti nella coda di priorità k,
- W_k : il tempo medio di attesa nella coda con priorità k,
- $\rho_k = \lambda_k / \mu_k$: l'utilizzazione del sistema per la priorità k,
- R: il tempo residuo di servizio medio. Tale parametro non dipende dalla classe k perché stiamo supponendo che non ci sia preemption.

Per la classe di priorità più alta si ha che (dalla (6.37)):

$$W_1 = R + \frac{1}{\mu_1} N_Q^{(1)}$$

Ed usando il teorema di Little si ha:

$$W_1 = \frac{R}{(1-\rho_1)}$$

Per la seconda classe di priorità si ha:

$$W_2 = R + \frac{1}{\mu_1} N_Q^{(1)} + \frac{1}{\mu_2} N_Q^{(2)} + \frac{1}{\mu_1} \cdot \lambda_1 \cdot M_2$$

dove il quarto addendo del secondo membro rappresenta il ritardo aggiuntivo causato dai clienti con priorità più alta che arrivano quando il customer con priorità 2 è già in attesa in coda.

$$W_2 = R + \rho_1 \cdot W_1 + \rho_2 \cdot W_2 + \rho_1 \cdot W_2$$

$$W_2 = \frac{R + \rho_1 \cdot W_1}{1 - \rho_1 - \rho_2} = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

Analogamente si trova che:

$$W_k = \frac{R}{(1-\rho_1-...-\rho_{k-1})(1-\rho_1-...)} \tag{6.39}$$

Con un procedimento analogo al precedente si ottiene che:

$$R = \frac{1}{2} \cdot \lambda \cdot \overline{X^2} = \frac{1}{2} \cdot \left(\sum_{i=1}^{n} \lambda_i \right) \cdot \overline{X^2}$$

dove $\overline{X^2}$ è il momento del secondo ordine mediato su tutte le classi di priorità:

$$\overline{X^2} = \frac{\lambda_1}{\sum_{i=1}^{n} \lambda_i} \cdot \overline{X_1^2} + \frac{\lambda_2}{\sum_{i=1}^{n} \lambda_i} \cdot \overline{X_2^2} + ... + \frac{\lambda_n}{\sum_{i=1}^{n} \lambda_i} \cdot \overline{X_n^2}$$

Sostituendo nella (6.39) si ottiene:

$$W_2 = \frac{\sum_{i=1}^{n} \lambda_i \cdot \overline{X^2}}{2 \cdot (1-\rho_1-...-\rho_{k-1}) \cdot (1-\rho_1-...-\rho_k)} \tag{6.40}$$

$$T_k = \frac{1}{\mu_k} + W_k \tag{6.41}$$

Tali valori dipendono fortemente dalle distribuzioni dei tempi di servizio delle varie classi.

Si può facilmente dimostrare che il ritardo medio per cliente tende a ridursi quando si attribuisce priorità più alta ai clienti con tempi di servizio più brevi. Questo si traduce nelle reti a commutazione di pacchetto nell'attribuire priorità maggiore ai pacchetti di controllo che solitamente sono molto più brevi rispetto ai pacchetti dati.

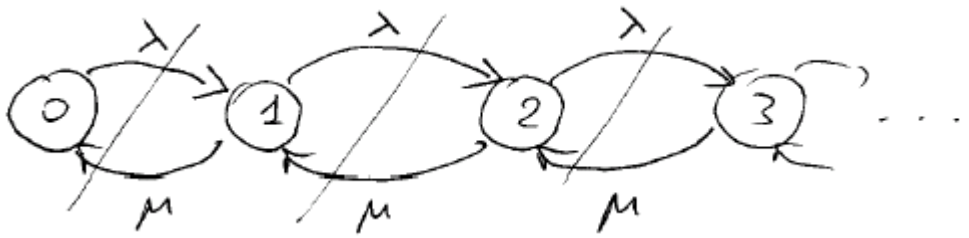
6.7 Esercizi risolti sulla teoria delle code

A completamento di questo studio sulla teoria delle code vediamo due tipologie di esercizi che sono esemplificativi per l'applicazione di questi metodi analitici.

Esercizio 1
Si consideri un multiplexer che raccoglie traffico formato da messaggi con intertempi di arrivo a distribuzione esponenziale. Il multiplexer è formato da un buffer e da una linea di trasmissione in uscita. Si consideri la seguente approssimazione: il tempo di trasmissione di un messaggio sulla linea è a distribuzione esponenziale con valore medio di E[x] = 10 ms. Da misure effettuate sullo stato del buffer sappiamo che la probabilità che il buffer sia vuoto è P_0 = 0.8. Ricavare il ritardo medio di trasmissione di un messaggio.

Soluzione
Il multiplexer è una coda con singolo servente. Il processo di arrivo è di Poisson con tasso medio λ da determinare. Il tempo medio di servizio è E[X] = 10 ms. Il tempo di servizio è a distribuzione esponenziale con tasso medio μ = 1/E[X].

Il sistema si modella come una coda M/M/1:



Siccome P_0 è positivo, il sistema è stabile perché è verificata la condizione di ergodicità.

Dalle equazioni ai tagli si ricava la probabilità degli stati:

$$P_o\lambda = P_1\mu \quad \Rightarrow \quad P_1 = \frac{\lambda}{\mu}P_o$$

$$P_1\lambda = P_2\mu \quad \Rightarrow \quad P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_o$$

$\rho = \frac{\lambda}{\mu}$ è l'intensità di traffico offerta al multiplexer in Erlang.

In generale $P_n = \left(\frac{\lambda}{\mu}\right)^n P_o = \rho^n P_o$

La probabilità di sistema vuoto si ricava con la condizione di normalizzazione:

$$P_0 = \frac{1}{1+\sum_{n=1}^{\infty} P_n / P_0} = \frac{1}{1+\sum_{n=1}^{\infty}\rho^n} = \frac{1}{\sum_{n=0}^{\infty}\rho^n} = 1-\rho = 1-\frac{\lambda}{\mu}$$

Siccome conosciamo $P_0 = 0.8$ e $\frac{1}{\mu} = \mathrm{E}[X] = 10\ \mathrm{ms}$, da questa formula si ricava il tasso medio di arrivo λ:

$$\mathrm{P}_0 = 1 - \frac{\lambda}{\mu} \Rightarrow 1 - 0.8 = \frac{\lambda}{\mu} \Rightarrow 1 - 0.8 = \lambda \cdot 10\ \mathrm{ms}$$

Quindi: $\lambda = \frac{1 - 0.8}{10} = 0.2 \quad \left[\frac{\text{messaggi}}{\text{ms}}\right]$

Il numero medio di richieste nel sistema è $N = \sum_{n=1}^{\infty} n\rho^n = \mathrm{P}'(z)\big|_{z=1}$, dove P(*z*) è la funzione generatrice del numero di messaggi nel sistema:

$$\mathrm{P}(z) = \sum_{n=0}^{\infty} z^n (1-\rho)\rho^n = \frac{1-\rho}{1-\rho \cdot z}$$

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

$$\mathrm{P}'(z)\big|_{z=1} = \frac{\rho(1-\rho)}{(1-\rho)^2} = \frac{\rho}{1-\rho} = N \Rightarrow N = \frac{0.2}{0.8} = 0.25\ \text{messaggi}$$

Il ritardo medio di messaggio è $T = \frac{N}{\lambda}$, in base al teorema di Little.

Quindi $T = 0.25\ \text{messaggi} \cdot \frac{1}{0.02}\ \frac{\text{ms}}{\text{messaggi}} = 12.5\ \text{ms} = \begin{matrix} 10\ \text{ms (tempo medio di servizio)} + \\ + 2.5\ \text{ms (tempo medio di attesa in coda)} \end{matrix}$.

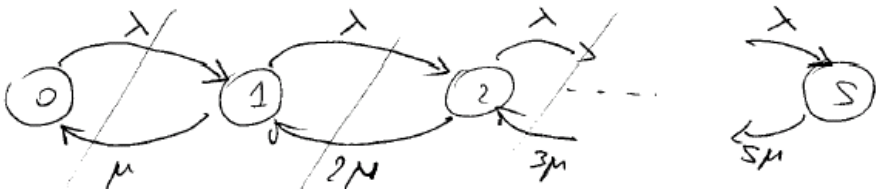
Esercizio 2

Si consideri un centralino telefonico che raccoglie le chiamate generate all'interno di una grande azienda in cui si hanno 1000 utenze telefoniche ciascuna generante un traffico di Poisson di 30 mErlang. Dimensionare il numero di linee telefoniche in uscita al centralino per collegarsi alla rete pubblica in modo da garantire una probabilità di blocco delle chiamate minore o uguale a 3% (si risolva formalmente il sistema in base al modello, ma si faccia uso della tabella allegata per trovare il risultato numerico). Cosa succede al numero di linee da prevedere in uscita per avere un blocco minore o uguale a 3% se il numero di utenti aumenta a 1300 ? Si confronti anche l'incremento percentuale di traffico offerto $\Delta\rho$% con l'incremento percentuale di linee ΔS% che ne risulta.

Soluzione

Visto che si tratta di 1000 utenti ciascuno generante un traffico di Poisson di 30 mErlang, si fa lo studio per numero infinito di utenti. Trattandosi di traffico telefonico, sappiamo che il modello per la durata di ogni chiamata è a distribuzione esponenziale con durata media di 3 minuti.

Per studiare il blocco nel sistema, applichiamo un modello M/M/S/S con S da determinare in modo da soddisfare i requisiti di blocco. Il modello è:



dove λ è il tasso di arrivo totale degli utenti, che si calcola così:

1) Ogni utente contribuisce un tasso di arrivo pari a $\frac{30 \cdot 10^{-3}\ \text{Erlang}}{3\ \text{min}} = 10^{-2}\ \frac{\text{chiamate}}{\text{min}}$
2) Il tasso totale è la somma dei tassi di utente: $\lambda = 1000 \cdot 10^{-2}\ \frac{\text{chiamate}}{\text{min}} = 10\ \frac{\text{chiamate}}{\text{min}}$

Inoltre:

$$\frac{1}{\mu} = 3\ \text{min}$$

La distribuzione di probabilità degli stati si determina scrivendo gli equilibri ai tagli:

$$P_o \lambda = P_1 \mu \quad \Rightarrow \quad P_1 = \frac{\lambda}{\mu} P_o$$

$$P_o \lambda = P_2 \cdot 2\mu \quad \Rightarrow \quad P_2 = \frac{\lambda}{2\mu} \cdot \frac{\lambda}{\mu} P_o$$

$$\rho = \frac{\lambda}{\mu} = 1000 \cdot 30 \cdot 10^{-3} = 30\ \text{Erlang}$$

(è l'intensità di traffico *offerta* al sistema)

In generale si ha: $P_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_o = \frac{\rho^n}{n!} P_o$

Ricaviamo P_0 con la condizione di normalizzazione:

$$P_0 = \frac{1}{1 + \sum_{n=1}^{S} P_n / P_0} = \frac{1}{\sum_{n=0}^{S} \frac{\rho^n}{n!}}$$

non conoscendo S non possiamo calcolare P_0.

Nessuna nuova chiamata è accettata quando il sistema ha tutti e S i serventi occupati. Pertanto la probabilità di blocco è la probabilità di avere il sistema nello stato S:

$$P_B \equiv P_S = \frac{\rho^S}{S! \sum_{n=0}^{S} \frac{\rho^n}{n!}}$$

(FORMULA ERLANG-B)

In questa formula imponiamo $P_S \leq 3\%$ e conosciamo ρ = 30 Erlang, possiamo quindi ricavare S. Per via numerica ciò è complesso e si fa uso della tabella ERLANG-B riprodotta qui di seguito. Nella tabella prendiamo la colonna per P_B = 3% e la scorriamo fino a prendere il valore di traffico immediatamente superiore ai 30 Erlang per fare un

dimensionamento cautelativo di *S*. Troviamo il valore di 30.53 Erlang a cui corrispondono *S* = 38 serventi.

Se il numero di utenti incrementa a 1300, il carico di traffico *offerto* al centralino è ρ = $1300 \cdot 30 \cdot 10^{-3}$ Erlang = 39 Erlang. Per dimensionare *S* usiamo di nuovo la tabella e troviamo che in corrispondenza a 39.06 Erlang si devono usare *S* = 47 serventi. L'incremento percentuale di serventi è ottenuto così:

$$38 + 38\frac{\Delta S\%}{100} = 47 \quad \Rightarrow \quad \Delta S\% = \left(\frac{47-38}{38}\right) \cdot 100 = 23.68\%$$

L'incremento percentuale di carico offerto al sistema è:

$$38 \text{ Erlang} + 38 \text{ Erlang} \cdot \frac{\Delta S\%}{100} = 39 \text{ Erlang} \quad \Rightarrow \quad \Delta S\% = \left(\frac{39-30}{30}\right) \cdot 100 = 30\%$$

Si noti l'effetto di multiplazione statistica del traffico che fa sì che $\Delta S\% < \Delta\rho\%$ pur mantenendo il requisito di $P_B \leq$ 3%.