# Note sul sistema a coda $\mathcal{M}/\mathcal{G}/1$

Franco Callegati
Walter Cerroni
Giorgio Corazza

Maggio 2007

# Indice

# Capitolo 1

# Il sistema a coda $\mathcal{M}/\mathcal{G}/1$

## 1.1 Alcuni risultati per il sistema $\mathcal{G}/\mathcal{G}/1$

Prima di procedere nella trattazione analitica dei sistemi $\mathcal{M}/\mathcal{G}/1$, è utile soffermarsi su alcuni risultati generali relativi ai sistemi a singolo servitore.

L'unica ipotesi che viene fatta sul sistema a coda è che i tempi di servizio dei clienti siano indipendenti e ugualmente distribuiti (anche independent identical distributed o i.i.d.), ossia che la statistica del tempo di servizio non vari da servizio a servizio.

**Definizione**
Osservando il sistema a coda per un periodo $T$ siano:

- $B(T)$ la quantità di tempo per cui il servitore risulta impegnato e $\bar{B}(T)$ il suo valore medio;
- $I(T)$ la quantità di tempo per cui il servitore non viene utilizzato e $\bar{I}(T)$ il suo valore medio.

Ovviamente:

$$B(T) + I(T) = T$$

Qualora il sistema sia ergodico, il servitore sarà impegnato in media per una porzione di $T$ pari a $\bar{B}(T) = T(1-P_0)$ e sarà libero per la rimanente parte di $T$ pari a $\bar{I}(T) = TP_0$.

Il numero medio di utenti serviti in $T$ risulta quindi $T(1-P_0)\mu$. Questo valore, in condizioni di equilibrio, deve essere uguale al numero medio di utenti giunti nello stesso tempo che è ovviamente pari a $\lambda T$, perciò:

$$\lambda T = T(1-P_0)\mu$$

e quindi:

$$P_0 = 1 - \frac{\lambda}{\mu}$$

Ricordando che $A_0 = \frac{\lambda}{\mu}$ e che per un sistema singolo servitore e senza perdita $A_s = A_0 = \rho$ se ne deduce che:

$$P_0 = 1 - A_0 = 1 - \rho$$

Quindi se ne conclude che per tutti i sistemi a singolo servitore ergodici la probabilità di avere il servitore libero è la medesima e vale

$$P_0 = 1 - \rho \tag{1.1}$$

Si noti che, per un sistema ergodico, risulta anche

$$\rho = \frac{\bar{B}(T)}{T} = \frac{\bar{B}(T)}{\bar{B}(T) + \bar{I}(T)} = 1 - P_0 \tag{1.2}$$

che risulta essere un modo alternativo per ricavare la (1.1).

**Definizione**
In un generico sistema a coda a singolo servitore si definisce

**periodo di attività** o *busy period* il generico periodo di tempo $T_{on}$ che inizia quando il servitore inattivo, a fronte dell'arrivo di un nuovo cliente, inizia a servirlo e termina quando il servitore ritorna inattivo;

**periodo di inattività** o *idle period* il periodo di tempo $T_{off}$ per il quale il servitore non fa nulla e resta in attesa di clienti.

Qualora il sistema sia ergodico esisteranno finiti i valori medi $\bar{T}_{on}$ e $\bar{T}_{off}$. Noti questi valori, un modo alternativo di calcolare l'utilizzazione del servitore è quello di considerare il tempo suddiviso in cicli composti da un tempo di attività e da un tempo di inattività. Per un sistema ergodico l'utilizzazione del servitore sarà quindi il rapporto fra la durata del periodo di attività $T_{on}$ e la durata totale del ciclo che vale $T_{on} + T_{off}$, da cui

$$\rho = \frac{\bar{T}_{on}}{\bar{T}_{off} + \bar{T}_{on}} \tag{1.3}$$

Ricordando l'equazione (1.2) ne consegue infine che:

$$\frac{\bar{T}_{off}}{\bar{T}_{on}} = \frac{\bar{I}}{\bar{B}} = \frac{P_0}{1 - P_0}$$

## 1.2 Il sistema $\mathcal{M}/\mathcal{G}/1$

Il sistema a coda $\mathcal{M}/\mathcal{G}/1$ è un sistema a singolo servitore con coda infinita tale che

- il processo degli arrivi è di Poisson;

- il processo di servizio ha distribuzione di probabilità generale.

Per il tempo di servizio, indicato come di consueto con $\vartheta$, ipotizziamo sia nota la densità di probabilità $f_\vartheta(t)$, per cui sono anche noti i vari momenti. In virtù di questo nel seguito ipotizzeremo che esistano finiti almeno i primi due momenti. Utilizzeremo nei calcoli che seguono il valor medio $E[\vartheta] = \bar{\vartheta}$ e la varianza $\sigma_\vartheta^2 = E[\vartheta^2] - E[\vartheta]^2$.

Anche per il sistema $\mathcal{M}/\mathcal{G}/1$ risulta che $P_0 = 1 - \rho$. Quindi la probabilità che il servitore sia occupato vale $1 - P_0 = \rho$. Per la proprietà PASTA, essendo gli arrivi di Poisson, questa risulta anche essere la probabilità di blocco, intesa come probabilità che un cliente venga accodato. Ne risulta quindi che per il sistema $\mathcal{M}/\mathcal{G}/1$

$$\pi_r = \rho \tag{1.4}$$

In un sistema con arrivi poissoniani l'idle period equivale, per la proprietà di assenza di memoria dell'esponenziale, ad un intervallo interarrivo, per cui ha distribuzione esponenziale e durata media $\bar{T}_{off} = \frac{1}{\lambda}$. Ricordando poi la (1.3)

$$\rho = \frac{\bar{T}_{on}}{\bar{T}_{off} + \bar{T}_{on}} = \frac{\lambda}{\mu}$$

da cui consegue $\mu\bar{T}_{on} = 1 + \lambda\bar{T}_{on}$ e quindi:

$$\bar{T}_{on} = \frac{1}{\mu - \lambda}$$

### 1.2.1 Tempo di servizio residuo

In un sistema a coda $\mathcal{M}/\mathcal{G}/1$ un nuovo utente, che arrivi al sistema all'istante $t_0$, trova il servitore occupato con probabilità $\rho$. In questo caso ovviamente il servitore è occupato da un utente che richiede uno specifico tempo di servizio $\vartheta_s$. Essendo tale servizio generalmente già in corso chiameremo $\zeta$ il tempo di servizio *residuo* del quale vogliamo calcolare la statistica.

In generale $t_0$ si posiziona all'interno di un tempo di servizio $\vartheta_s$ con probabilità tanto maggiore quanto maggiore è $\vartheta_s$. Perciò la densità di probabilità di $\vartheta_s$ risulta [1]

$$f_{\vartheta_s}(s) = Cs f_\vartheta(s)$$

dove $C$ è una costante di normalizzazione. Imponendo $\int_0^\infty f_{\vartheta_s}(s)ds = 1$ si ottiene $C = 1/\bar{\vartheta}$ per cui

$$f_{\vartheta_s}(s) = \frac{s f_\vartheta(s)}{\bar{\vartheta}} \qquad s \geq 0$$

È ora possibile calcolare la densità di probabilità di $\zeta$ condizionata ad un certo valore di $\vartheta_s$, considerando che $t_0$ si posiziona in modo casuale all'interno del tempo di servizio e

[1] Si noti che, per rendere più semplice la trattazione che segue, per questa densità di probabilità utilizzeremo il simbolo $s$ per indicare la variabile indipendente, anziché il consueto simbolo $t$.

quindi può assumere qualunque valore fra 0 e $\vartheta_s$ con probabilità uniforme. Ne consegue che

$$f_{\zeta|\vartheta_s}(t,s) = \begin{cases} \frac{1}{s} & 0 < t \leq s \\ 0 & \text{altrove} \end{cases}$$

Quindi

$$f_{\zeta,\vartheta_s}(t,s) = f_{\vartheta_s}(s) f_{\zeta|\vartheta_s}(t,s) = \frac{s f_\vartheta(s)}{\bar{\vartheta} s} = \frac{f_\vartheta(s)}{\bar{\vartheta}} \qquad t > 0 \text{ e } t \leq s$$

La densità di probabilità di $\zeta$ si può ottenere integrando $f_{\zeta,\vartheta_s}(t,s)$ per tutti i valori possibili di $\vartheta_s$. Tenendo conto che $\zeta \leq \vartheta_s$ in quanto il tempo di servizio residuo non può essere maggiore del tempo di servizio totale, si ottiene che:

$$f_\zeta(t) = \int_{-\infty}^{\infty} f_{\zeta,\vartheta_s}(t,s)ds = \int_t^{\infty} \frac{f_\vartheta(s)}{\bar{\vartheta}} ds = \begin{cases} \frac{1-F_\vartheta(t)}{\bar{\vartheta}} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{1.5}$$

Dalla densità di probabilità di $\zeta$ è ora possibile ricavare il relativo valor medio.

$$\bar{\zeta} = \int_0^{\infty} t f_\zeta(t)dt = \int_0^{\infty} t \frac{1-F_\vartheta(t)}{\bar{\vartheta}} dt = \frac{1}{\bar{\vartheta}} \int_0^{\infty} t(1-F_\vartheta(t))dt = \\ \frac{1}{\bar{\vartheta}} \left[ \left| \frac{t^2}{2}(1-F_\vartheta(t)) \right|_0^{\infty} + \int_0^{\infty} \frac{t^2}{2} f_\vartheta(t)dt \right] = \frac{E[\vartheta^2]}{2\bar{\vartheta}} = \frac{\sigma_\vartheta^2 + \bar{\vartheta}^2}{2\bar{\vartheta}} \tag{1.6}$$

Più in generale è possibile mostrare che il momento di ordine $k$ di $\zeta$ è esprimibile in funzione di $\bar{\vartheta}$ e del momento di ordine $k+1$ di $\vartheta$, come segue:

$$E[\zeta^k] = \int_0^{\infty} t^k f_\zeta(t)dt = \int_0^{\infty} t^k \frac{1-F_\vartheta(t)}{\bar{\vartheta}} dt = \frac{1}{\bar{\vartheta}} \int_0^{\infty} t^k(1-F_\vartheta(t))dt = \\ \frac{1}{\bar{\vartheta}} \left[ \left| \frac{t^{k+1}}{k+1}(1-F_\vartheta(t)) \right|_0^{\infty} + \int_0^{\infty} \frac{t^{k+1}}{k+1} f_\vartheta(t)dt \right] = \frac{E[\vartheta^{k+1}]}{(k+1)\bar{\vartheta}} \tag{1.7}$$

## 1.3 Traffico e tempi medi di permanenza

In questa sezione vedremo di formalizzare il modello matematico con cui sia possibile studiare il sistema $\mathcal{M}/\mathcal{G}/1$ ed in particolare di concentrarci sulla valutazione delle grandezze medie tipicamente di interesse.

### 1.3.1 La formula di Pollaczek-Khintchine

Ipotizzando un sistema a coda con politica di servizio di tipo FIFO, calcoliamo inizialmente il valor medio del tempo di attesa in coda.

Ciò si può fare tenendo presente quanto segue, con riferimento ad un generico utente in arrivo al sistema che dovrà

- attendere per un tempo $T'$ che termini di essere servito il cliente attualmente in servizio, il quale terrà occupato il servitore per un tempo di servizio residuo $\zeta$;
- attendere per un tempo $T''$ che vengano serviti tutti i $k_c$ clienti in attesa prima di lui.

Pertanto

$$\eta = T' + T''$$

Ovviamente, passando ai valori medi,

$$\bar{\eta} = E[T'] + E[T'']$$

in cui

$$E[T'] = \rho\bar{\zeta} \tag{1.8}$$

e

$$E[T''] = k_c\bar{\vartheta} = A_c\bar{\vartheta} = \lambda\bar{\eta}\bar{\vartheta} = \rho\bar{\eta} \tag{1.9}$$

La prima equazione si giustifica tenendo presente che, all'arrivo di un nuovo cliente, questi troverà il servitore occupato con probabilità $1 - P_0 = \rho$, nel qual caso dovrà attendere un tempo di servizio residuo medio, mentre troverà il servitore libero con probabilità $P_0 = 1 - \rho$ nel quale non dovrà attendere. Pertanto risulta

$$E[T'] = \rho\bar{\zeta} \tag{1.10}$$

La seconda semplicemente tenendo conto che il numero di utenti mediamente presenti in coda altro non è che il traffico in coda. Ne consegue che

$$\bar{\eta} = \rho\bar{\zeta} + \rho\bar{\eta} \tag{1.11}$$

da cui si ottiene

$$\bar{\eta} = \bar{\zeta}\frac{\rho}{1-\rho} = \frac{\rho E[\vartheta^2]}{2(1-\rho)\bar{\vartheta}} = \frac{\lambda E[\vartheta^2]}{2(1-\rho)} \tag{1.12}$$

L'espressione $\bar{\eta} = \bar{\zeta}\frac{\rho}{1-\rho}$ mostra come il tempo medio di attesa in coda dipenda dal tempo di servizio residuo medio (dovuto al fatto che quando si va in coda si deve per prima cosa attendere che il servitore termini il servizio in corso), opportunamente pesato da un'espressione che dipende dal traffico offerto $\rho$.

Dal precedente risultato è possibile calcolare tutti i restanti valori medi delle grandezze caratteristiche del comportamento del sistema. Risulta infatti:

$$\bar{\delta} = \bar{\vartheta} + \bar{\eta} = \bar{\vartheta} + \frac{\lambda E[\vartheta^2]}{2(1-\rho)} = \bar{\vartheta} + \frac{\rho^2 + \lambda^2\sigma_\vartheta^2}{2\lambda(1-\rho)} = \frac{2\rho - \rho^2 + \lambda^2\sigma_\vartheta^2}{2\lambda(1-\rho)} \tag{1.13}$$

da cui

$$A = \frac{2\rho - \rho^2 + \lambda^2\sigma_\vartheta^2}{2(1-\rho)} = \rho + \frac{\rho^2 + \lambda^2\sigma_\vartheta^2}{2(1-\rho)} \tag{1.14}$$

e infine

$$A_c = \frac{\lambda^2 E[\vartheta^2]}{2(1-\rho)} = \frac{\rho^2 + \lambda^2 \sigma_\vartheta^2}{2(1-\rho)} \tag{1.15}$$

La (1.14) è chiamata *formula del valor medio di Pollaczek-Khintchine* (formula PK) ed il suo andamento in funzione di $\sigma_\vartheta^2$ è mostrato in figura 1.1 per $\lambda = 1$ e diversi valori di $\rho$.

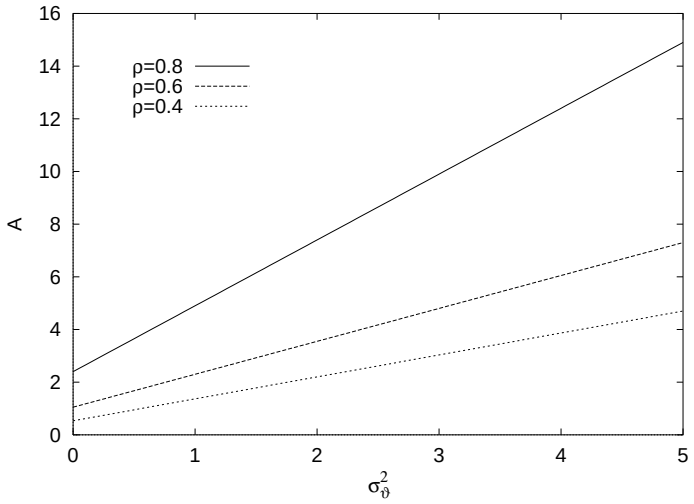


Figura 1.1: $\mathcal{M}/\mathcal{G}/1$: $A$ in funzione di $\sigma_\vartheta^2$ per $\lambda = 1$ e diversi valori di $\rho$.

### 1.3.2 La catena di Markov nascosta e le probabilità di stato della coda $\mathcal{M}/\mathcal{G}/1$

**Definizione**
Unitamente alle quantità già note, le principali variabili che verranno utilizzate nel seguito sono:

$t_n$ : istante in cui il cliente $n$-esimo lascia il sistema;

$k(t_n^+) = k_n$ numero di clienti nel sistema immediatamente dopo l'istante $t_n$;

$a_n$ : numero di clienti arrivati al sistema durante il servizio del cliente $n$-esimo;

$\bar{a}$ : numero medio di clienti arrivati al sistema durante un generico tempo di servizio, quando il sistema si trova in equilibrio statistico (non dipende dall'istante di tempo considerato);

$\sigma_a^2$ : varianza di $a$.

La descrizione statistica del numero di arrivi nel tempo di servizio $n$-esimo dipende dalla statistica degli arrivi e da quella dei tempi di servizio. Fissata la durata del tempo di servizio $\vartheta$, la probabilità $\alpha_{j/\vartheta}$ di avere $j$ arrivi durante uno specifico tempo di servizio è calcolabile con formula di Poisson

$$\alpha_{j/\vartheta} = \Pr(a = j \mid \vartheta) = P(j, \vartheta) = \frac{(\lambda\vartheta)^j}{j!} e^{-\lambda\vartheta} \tag{1.16}$$

da cui, decondizionando rispetto a $\vartheta$,

$$\alpha_j = \int_0^\infty P(j,t) f_\vartheta(t) dt = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} f_\vartheta(t) dt \tag{1.17}$$

Quest'ultimo integrale è solitamente non risolvibile in forma chiusa e quindi richiede il ricorso al calcolo numerico.

È invece possibile calcolare il valor medio e la varianza di $a$, sfruttando le conoscenze sul processo di Poisson:

$$E[a \mid \vartheta] = \lambda\vartheta$$

da cui si ottiene il valor medio non condizionato

$$E[a] = \bar{a} = \int_0^\infty E[a \mid t] f_\vartheta(t) dt = \lambda\bar{\vartheta} = \rho \tag{1.18}$$

Inoltre

$$E[a^2 \mid \vartheta] = \sigma^2_{a|\vartheta} + E[a \mid \vartheta]^2 = \lambda\vartheta + (\lambda\vartheta)^2$$

da cui

$$E[a^2] = \int_0^\infty E[a^2 \mid t] f_\vartheta(t) dt = \lambda\bar{\vartheta} + \lambda^2 E[\vartheta^2] = \lambda\bar{\vartheta} + \lambda^2(\sigma^2_\vartheta + \bar{\vartheta}^2) = \rho + \lambda^2\sigma^2_\vartheta + \rho^2 \tag{1.19}$$

Per valutare il comportamento, almeno in termini medi, del sistema $\mathcal{M}/\mathcal{G}/1$, per il momento si consideri come grandezza rappresentativa dello stato del sistema non più il numero di utenti presenti al suo interno in un istante generico, bensì $k(t_n^+) = k_n$ ossia il numero di clienti nel sistema immediatamente dopo la partenza del cliente $n$-esimo. Per quanto riguarda l'evoluzione del sistema, lo stato $k_{n+1}$ dipende solamente dallo stato $k_n$ e dal numero di arrivi che si sono avuti nell'intervallo di tempo compreso fra $t_n$ e $t_{n+1}$, ossia $a_{n+1}$.

Pertanto il sistema si può descrivere come catena di Markov poiché lo stato futuro dipende solamente dallo stato presente. In generale ciò non significa che la coda $\mathcal{M}/\mathcal{G}/1$ sia una catena di Markov, ma solamente che il sistema la cui evoluzione è descritta tramite gli stati $k_n$ è una catena di Markov. Tale catena di Markov viene solitamente chiamata *catena di Markov nascosta* o *embedded Markov chain*. La relazione che lega $k_{n+1}$ e $k_n$ si può ricavare con il ragionamento che segue.

Se il sistema si svuota all'istante $t_n$, cioè $k_n = 0$, prima o poi arriverà il cliente $n+1$-esimo che verrà servito immediatamente e, una volta terminato il servizio, lascerà

il sistema con un numero di clienti pari a quanti ne sono arrivati durante il suo tempo di servizio, cioè $a_{n+1}$.

Nel caso in cui, invece, il sistema non si svuoti con la partenza dell'$n$-esimo cliente, cioè $k_n > 0$, al termine del servizio dell'$n+1$-esimo cliente il numero di utenti nel sistema sarà pari a quelli che c'erano prima, meno il cliente appena uscito, più tutti gli arrivi avvenuti durante il suo servizio. In formule:

$$k_{n+1} = \begin{cases} a_{n+1} & k_n = 0 \\ k_n - 1 + a_{n+1} & k_n > 0 \end{cases} \tag{1.20}$$

Definita la funzione *gradino discreta* come segue:

$$u(k_n) = \begin{cases} 1 & k_n > 0 \\ 0 & k_n = 0 \end{cases}$$

l'equazione (1.20) si può riscrivere come:

$$k_{n+1} = k_n - u(k_n) + a_{n+1} \tag{1.21}$$

Si può dimostrare che, se $\rho < 1$, la catena di Markov nascosta la cui evoluzione è descritta dall'equazione (1.21) è ergodica e che quindi esistono le probabilità di stato all'equilibrio

$$\Pi = \{\pi_0, \pi_1, \ldots, \pi_k, \ldots\}, \quad \text{dove} \;\; \pi_k = \lim_{n\to\infty} \Pr\{k_n = k\}$$

Per il calcolo di $\Pi$ è necessario determinare le probabilità di transizione. Per $k_n \neq 0$

$$P_{ij}(n, n+1) = \Pr\{k_{n+1} = j \mid k_n = i\} = \Pr\{a_{n+1} = j - i + 1\}$$

in cui le transizioni possibili sono quelle per cui $j \geq i - 1$. All'equilibrio:

$$P_{ij} = \lim_{n\to\infty} P_{ij}(n, n+1) = \lim_{n\to\infty} \Pr\{a_{n+1} = j - i + 1\} = \Pr\{a = j - i + 1\} = \alpha_{j-i+1}$$

che indica la probabilità di avere $j - i + 1$ arrivi in un generico tempo di servizio.

Per $k_n = 0$ le probabilità di transizione sono:

$$P_{0j}(n, n+1) = \Pr\{a_{n+1} = j\}$$

e

$$P_{0j} = \lim_{n\to\infty} P_{0j}(n, n+1) = \alpha_j$$

valide per $j \geq 0$. Le $\pi_k$ si possono quindi ricavare risolvendo il sistema lineare:

$$\Pi = \Pi\mathcal{P}$$

dove $\Pi$ è il vettore delle probabilità di stato e $\mathcal{P}$ è la matrice delle probabilità di transizione. In questo caso si ha:

$$\mathcal{P} = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \ldots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \ldots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \ldots \\ 0 & 0 & \alpha_0 & \alpha_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1.22}$$

da cui

$$\pi_j = \pi_0 \alpha_j + \sum_{i=1}^{j+1} \pi_i \alpha_{j-i+1} \quad j = 0, 1, \ldots, \infty \tag{1.23}$$

Questo sistema di equazioni non è facilmente risolvibile. Può essere utile formulare la soluzione utilizzando metodi trasformazionali. A questo proposito va ricordato che:

- data una variabile aleatoria $x$ con densità di probabilità $f_x(t)$ e distribuzione di probabilità $F_x(t)$, $\mathcal{L}_x(s) = \int_0^\infty e^{-st} f_x(t)dt$ è la trasformata di Laplace[2] della densità di probabilità e $\Phi_x(s) = \int_0^\infty e^{-st} F_x(t)dt = \frac{\mathcal{L}_x(s)}{s}$ la trasformata della distribuzione di probablità ;[3]
- data una serie di valori $x_i$, la trasformata $z$ della serie, detta $X(z)$ è definita come: $X(z) = \sum_{i=0}^\infty z^i x_i$.[4]
- presa la convoluzione discreta di due sequenze di numeri $x_i$ e $y_i$, $c_i = x_i \star y_i = \sum_{j=0}^i x_j y_{i-j}$, la trasformata $z$ della convoluzione è uguale al prodotto delle trasformate $X(z)$ e $Y(z)$: $C(z) = \sum_{i=0}^\infty c_i z^i = X(z)Y(z)$

Cercheremo ora la soluzione del problema della determinazione delle $\pi_k$, cercando di determinare in forma chiusa la relativa trasformata $z$:

$$\Pi(z) = \sum_{k=0}^\infty z^k \pi_k$$

Innanzitutto calcoliamo:

$$\begin{aligned}\alpha(z) = \sum_{j=0}^\infty z^j \alpha_j &= \sum_{j=0}^\infty z^j \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} f_\vartheta(t)dt = \int_0^\infty \sum_{j=0}^\infty \left( z^j \frac{(\lambda t)^j}{j!} \right) e^{-\lambda t} f_\vartheta(t)dt \\ &= \int_0^\infty e^{z\lambda t} e^{-\lambda t} f_\vartheta(t)dt = \int_0^\infty e^{-\lambda(1-z)t} f_\vartheta(t)dt = \mathcal{L}_\vartheta(\lambda - \lambda z) \end{aligned} \tag{1.24}$$

Ora, moltiplicando l'equazione (1.23) per $z^j$ e sommando per $j = 0, 1, \ldots, \infty$

$$\begin{aligned}\Pi(z) = \sum_{j=0}^\infty \pi_j z^j &= \sum_{j=0}^\infty \left[ \pi_0 \alpha_j z^j + \frac{1}{z} \sum_{i=1}^{j+1} \pi_i \alpha_{j-i+1} z^{j+1} \right] \\ &= \sum_{j=0}^\infty \left[ \pi_0 \alpha_j z^j + \frac{1}{z} \sum_{i=0}^{j+1} \pi_i \alpha_{j-i+1} z^{j+1} - \frac{\pi_0 \alpha_{j+1} z^{j+1}}{z} \right] \end{aligned} \tag{1.25}$$

[2]Si noti che le densità e le distribuzioni di probabilità di interesse per questo corso sono nulle per valori negativi della variabile aleatoria per cui scrivere $\int_0^\infty$ equivale a scrivere $\int_{-\infty}^\infty$.

[3]Nel calcolo delle probabilità è spesso utilizzata anche la funzione detta *generatrice dei momenti* definita come $\mathcal{M}_x(s) = \int_0^\infty e^{st} f_x(t)dt$ la cui proprietà fondamentale è che $\frac{d^n}{ds^n}\mathcal{M}_x(s)\big|_{s=0} = E[x^n]$. Evidentemente si ha che $\mathcal{M}_x(s) = \mathcal{L}_x(-s)$.

[4]La trasformata $z$ è invertibile con la seguente operazione di antitrasformazione, che ci riconsegna la serie di numeri originali $x_i = \frac{1}{i!}\frac{d^i X(z)}{dz^i}\big|_{z=0} = X^{(i)}(0)/i!$

Il primo termine tra parentesi quadre dà luogo semplicemente a $\pi_0\alpha(z)$, mentre per il secondo si può scrivere:

$$\sum_{j=0}^{\infty}\sum_{i=0}^{j+1}\pi_i\alpha_{j-i+1}z^{j+1} = \sum_{j=1}^{\infty}\sum_{i=0}^{j}\pi_i\alpha_{j-i}z^j = \sum_{j=0}^{\infty}\sum_{i=0}^{j}\pi_i\alpha_{j-i}z^j - \pi_0\alpha_0 = \alpha(z)\Pi(z) - \pi_0\alpha_0$$

Inoltre:

$$\sum_{j=0}^{\infty}\frac{\pi_0\alpha_{j+1}z^{j+1}}{z} = \sum_{j=0}^{\infty}\frac{\pi_0\alpha_j z^j}{z} - \frac{\pi_0\alpha_0}{z} = \frac{\pi_0\alpha(z)}{z} - \frac{\pi_0\alpha_0}{z}$$

Ne consegue:

$$\Pi(z) = \pi_0\alpha(z) + \frac{\alpha(z)\Pi(z)}{z} - \frac{\pi_0\alpha(z)}{z} \tag{1.26}$$

Dall'equazione (1.26), ricordando che $\pi_0 = 1 - \rho$, si può ricavare il risultato finale per la trasformata $z$ di $P_k$:

$$\Pi(z) = \frac{(1-\rho)(1-z)\alpha(z)}{\alpha(z) - z} = \frac{(1-\rho)(1-z)\mathcal{L}_\vartheta(\lambda - \lambda z))}{\mathcal{L}_\vartheta(\lambda - \lambda z) - z} \tag{1.27}$$

Questa espressione può essere utilizzata per ricavare le $\pi_k$ qualora sia nota $\mathcal{L}_\vartheta(s)$, sia in forma chiusa, sia in forma numerica.

### 1.3.3 Le probabilità di stato ad un istante qualunque

È possibile dimostrare che le probabilità di stato agli istanti $t_n$ sono anche le probabilità di stato del sistema in un istante qualunque.

**Definizione**
Si consideri un intervallo di vita del sistema $T$. Si definiscono:

$a_k(T)$ il numero di arrivi in corrispodenza del fatto che vi siano $k$ utenti nel sistema (cioè un contatore che incrementa ogni volta che un arrivo porta lo stato da $k$ a $k+1$);

$d_k(T)$ il numero di partenze che lasciano $k$ utenti nel sistema (cioè un contatore che incrementa ogni volta che una partenza porta lo stato da $k+1$ a $k$);

$a(T) = \sum_{k=0}^{\infty} a_k(T)$ il numero totale di arrivi in $T$;

$d(T) = \sum_{k=0}^{\infty} d_k(T)$ il numero totale di partenze in $T$;

$\pi_k^a = \lim_{T\to\infty} \frac{a_k(T)}{a(T)}$ le probabilità di stato agli istanti immediatamente precedenti all'arrivo di un cliente;

$\pi_k^d = \lim_{T\to\infty} \frac{d_k(T)}{d(T)}$ le probabilità di stato agli istanti immediatamente successivi alla partenza di un cliente dal sistema.

Poiché non ci possono essere arrivi né partenze contemporanei, lo stato $k$ può essere raggiunto solamente tramite una nascita dallo stato $k-1$ o una morte dallo stato $k+1$. Ciò vuol dire che lo stato del sistema cambia di una unità alla volta e vale:

$$|a_k(T) - d_k(T)| \leq 1$$

che afferma in pratica che nel periodo $T$ il numero di arrivi e di partenze differisce al più di una unità. Per dimostrare che questa disequazione è vera, si considerino le seguenti possibilità:

- il sistema è partito da uno stato $p \leq k$ e si troverà dopo $T$ nello stato $q > k$: ci dovrà sempre essere una transizione da $k$ a $k+1$ in più rispetto a quelle da $k+1$ a $k$, quindi $a_k(T) = d_k(T) + 1$;
- il sistema è partito da uno stato $p > k$ e si troverà dopo $T$ nello stato $q \leq k$: è il duale del caso precedente, per cui $a_k(T) = d_k(T) - 1$;
- il sistema è partito nello stato $p > k$ e si troverà dopo $T$ nello stato $q > k$: per ogni transizione da $k+1$ a $k$ ce ne sarà sempre una in direzione opposta, quindi $a_k(T) = d_k(T)$;
- il sistema è partito nello stato $p \leq k$ e si troverà dopo $T$ nello stato $q \leq k$: il duale del precedente, per cui si ha ancora $a_k(T) = d_k(T)$.

Inoltre vale l'ovvia relazione:

$$k(T) = k(0) + a(T) - d(T)$$

Procedendo al limite per $T \to \infty$:

$$\pi_k^d = \lim_{T\to\infty} \frac{d_k(T)}{d(T)} = \lim_{T\to\infty} \frac{a_k(T) + d_k(T) - a_k(T)}{a(T) + k(0) - k(T)} = \lim_{T\to\infty} \frac{a_k(T)}{a(T)} = \pi_k^a \tag{1.28}$$

poiché $d_k(T) - a_k(T)$ è limitato ad 1 e sia $k(T)$ sia $k(0)$ sono finiti se il sistema è stabile.

Tenendo conto che il processo degli arrivi è assolutamente indipendente dallo stato del sistema, gli istanti degli arrivi sono istanti casuali (proprietà PASTA). Per questo le $\pi_k$ agli istanti di arrivo devono essere le stesse che le $\pi_k$ prese in un istante qualunque lungo l'asse dei tempi. Perciò le probabilità di stato in un istante qualunque sono uguali alle probabilità di stato negli istanti delle partenze.

## 1.4 Distribuzione del tempo di attesa

Utilizzando i risultati precedenti è abbastanza immediato ricavare la trasformata del tempo speso complessivamente nel sistema $\delta$ e del tempo speso in coda $\eta$, qualora la politica di scheduling sia di tipo FIFO.

Si ipotizzi di considerare il cliente $n$-esimo. Esso spenderà nel sistema il tempo $\delta_n$, in servizio il tempo $\vartheta_n$ ed in coda il tempo $\eta_n$. Inoltre immaginiamo che lasci il sistema all'istante $t_n$. Sia $b_n$ il numero di utenti che giungono al sistema durante $\delta_n$. Si noti che

$$b_n = k(t_n^+) = k_n \tag{1.29}$$

ossia il numero di arrivi in $\delta_n$ altro non è che lo stato del sistema quando l'utente $n$-esimo lascia il sistema stesso. La ragione di questo è facilmente comprensibile notando che, in un sistema FIFO, a $t_n^+$ sono stati serviti tutti gli utenti presenti nel sistema all'arrivo del cliente $n$-esimo più il cliente $n$-esimo stesso. Rimangono quindi nel sistema tanti utenti quanti ne sono arrivati durante la permanenza del cliente $n$-esimo e quindi durante $\delta_n$. Ne consegue che

$$\beta_k = \Pr\{b_n = k\} = \pi_k \tag{1.30}$$

A questo punto calcoliamo

$$\beta_{k/\delta} = \Pr\{b_n = k \mid \delta\} = \frac{(\lambda\delta)^k}{k!} e^{-\lambda\delta}$$

conseguentemente

$$\beta_k = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-\lambda t} f_\delta(t) dt$$

per cui, con calcoli simili all'equazione (1.24), ricordando la (1.29),

$$\beta(z) = \Pi(z) = \mathcal{L}_\delta(\lambda(1-z))$$

In conclusione risulta

$$\mathcal{L}_\delta(\lambda(1-z)) = \frac{(1-\rho)(1-z)\mathcal{L}_\vartheta(\lambda(1-z))}{\mathcal{L}_\vartheta(\lambda(1-z)) - z} \tag{1.31}$$

da cui, ponendo $s = \lambda(1-z)$ e pertanto $z = 1 - \frac{s}{\lambda}$,

$$\mathcal{L}_\delta(s) = \mathcal{L}_\vartheta(s) \frac{s(1-\rho)}{s - \lambda + \lambda\mathcal{L}_\vartheta(s)} \tag{1.32}$$

Dal momento che $\delta = \vartheta + \eta$ e che $\vartheta$ e $\eta$ sono indipendenti, risulta $f_\delta(t) = f_\vartheta(t) \star f_\eta(t)$ e quindi $\mathcal{L}_\delta(s) = \mathcal{L}_\vartheta(s)\mathcal{L}_\eta(s)$, per cui

$$\mathcal{L}_\eta(s) = \frac{s(1-\rho)}{s - \lambda + \lambda\mathcal{L}_\vartheta(s)} \tag{1.33}$$

E' possibile esprimere la $\mathcal{L}_\eta(s)$ in funzione della trasformata di Laplace della densità di probabilità del tempo residuo (1.5), che può essere riscritta come

$$f_\zeta(t) = \frac{1 - F_\vartheta(t)}{\bar{\vartheta}}\, u(t)$$

dove $u(t)$ è la funzione *gradino*. La trasformata di Laplace vale

$$\mathcal{L}_\zeta(s) = \frac{1}{\bar{\vartheta}}\left[\frac{1}{s} - \frac{\mathcal{L}_\vartheta(s)}{s}\right] = \frac{1-\mathcal{L}_\vartheta(s)}{s\bar{\vartheta}} \tag{1.34}$$

da cui risulta, sostituendo $\mathcal{L}_\vartheta(s)$ nella (1.33),

$$\mathcal{L}_\eta(s) = \frac{1-\rho}{1-\rho\,\mathcal{L}_\zeta(s)}$$

Le trasformate di Laplace delle distribuzioni cumulative di probabilità sono poi ovviamente espresse da

$$\Phi_\delta(s) = \frac{\mathcal{L}_\delta(s)}{s} = \mathcal{L}_\vartheta(s)\frac{(1-\rho)}{s-\lambda+\lambda\,\mathcal{L}_\vartheta(s)} \tag{1.35}$$

e

$$\Phi_\eta(s) = \frac{\mathcal{L}_\eta(s)}{s} = \frac{(1-\rho)}{s-\lambda+\lambda\,\mathcal{L}_\vartheta(s)} \tag{1.36}$$

Le equazioni (1.35) e (1.36) sono particolarmente utili qualora si voglia calcolare la probabilità che questi tempi eccedano un particolare limite. Ad esempio

$$\Pr\{\eta > t_0\} = 1 - F_\eta(t_0) = 1 - \int_{-\infty}^{\infty}\Phi_\eta(s)e^{st_0}ds = 1 - \int_{-\infty}^{\infty}\frac{(1-\rho)}{s-\lambda+\lambda\,\mathcal{L}_\vartheta(s)}e^{st_0}ds$$

## 1.5 Alcuni casi particolari

In questo paragrafo i risultati appena ottenuti sono applicati ad alcuni casi particolari di code a singolo servitore con arrivi poissoniani: in particolare si considerano tempi di servizio esponenziali (coda $\mathcal{M}/\mathcal{M}/1$), deterministici (coda $\mathcal{M}/\mathcal{D}/1$) ed erlangiani (coda $\mathcal{M}/\mathcal{E}_n/1$).

### 1.5.1 Il sistema $\mathcal{M}/\mathcal{M}/1$

Il tempo di servizio esponenziale presenta valor medio $\bar{\vartheta} = 1/\mu$ e varianza $\sigma_\vartheta^2 = 1/\mu^2$. Applicando la formula PK, si ottiene:

$$A = \frac{2\rho - \rho^2 + \lambda^2/\mu^2}{2(1-\rho)} = \frac{\rho}{1-\rho}$$

Per le probabilità di stato si ottiene:

$$\mathcal{L}_\vartheta(s) = \int_0^\infty \mu e^{-\mu t}e^{-st}dt = \frac{\mu}{\mu+s}$$

e quindi

$$\Pi(z) = \frac{(1-\rho)(1-z)\frac{\mu}{\mu+\lambda-\lambda z}}{\frac{\mu}{\mu+\lambda-\lambda z} - z} = \frac{(1-\rho)(1-z)}{1-z-\rho z(1-z)} = \frac{(1-\rho)(1-z)}{(1-z)(1-\rho z)} = \frac{1-\rho}{1-\rho z} \tag{1.37}$$

Le probabilità di stato quindi sono

$$\begin{aligned}
\pi_0 &= \Pi(0) = 1-\rho \\
\pi_1 &= \Pi'(0) = (1-\rho)\rho \\
\pi_2 &= \frac{\Pi''(0)}{2} = (1-\rho)\rho^2 \\
&\vdots \\
\pi_k &= \frac{\Pi^{(k)}(0)}{k!} = (1-\rho)\rho^k
\end{aligned}$$

Inoltre

$$\begin{aligned}
\mathcal{L}_\delta(s) = \mathcal{L}_\vartheta(s)\frac{s(1-\rho)}{s-\lambda+\lambda\mathcal{L}_\vartheta(s)} &= \frac{\mu}{\mu+s}\left(\frac{s(1-\rho)}{s-\lambda+\frac{\lambda\mu}{\mu+s}}\right) \\
&= \frac{s\mu(1-\rho)}{s^2-\lambda s+s\mu-\lambda\mu+\lambda\mu} = \frac{\mu(1-\rho)}{s+\mu(1-\rho)} \quad (1.38)
\end{aligned}$$

e

$$\begin{aligned}
\mathcal{L}_\eta(s) = \frac{s(1-\rho)}{s-\lambda+\frac{\lambda\mu}{\mu+s}} &= \frac{(1-\rho)(\mu+s)}{s+\mu-\lambda} \\
&= (1-\rho)+\frac{\lambda(1-\rho)}{s+\mu(1-\rho)} = (1-\rho)+\rho\frac{\mu(1-\rho)}{s+\mu(1-\rho)} \quad (1.39)
\end{aligned}$$

da cui si ricava che

$$f_\delta(t) = \mu(1-\rho)e^{-\mu(1-\rho)t} = (\mu-\lambda)e^{-(\mu-\lambda)t} \quad t \geq 0$$

e

$$f_\eta(t) = (1-\rho)\delta(t) + \rho(\mu-\lambda)e^{-(\mu-\lambda)t} \quad t \geq 0$$

### 1.5.2 Il sistema $\mathcal{M}/\mathcal{D}/1$

In questo caso il tempo di servizio è costante per cui $\vartheta = \bar{\vartheta} = D$ e $\sigma_\vartheta^2 = 0$. Dalla formula PK consegue che:

$$A = \frac{2\rho-\rho^2}{2(1-\rho)} = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)}$$

che mostra come il sistema a coda $\mathcal{M}/\mathcal{D}/1$ sia mediamente meno pieno del sistema $\mathcal{M}/\mathcal{M}/1$. Per quanto riguarda le probabilità di stato, ricordando che se $D$ è il valore del tempo di servizio la relativa densità è $f_\vartheta(t) = \delta(t-D)$, si ha:

$$\mathcal{L}_\vartheta(s) = e^{-Ds}$$

Quindi la trasformata $z$ delle probabilità di stato risulta:

$$\Pi(z) = \frac{(1-\rho)(1-z)e^{-\rho(1-z)}}{e^{-\rho(1-z)} - z} = \frac{(1-\rho)(1-z)}{1 - ze^{\rho(1-z)}} \tag{1.40}$$

che può essere anti-trasformata ottenendo le relative probabilità di stato:

$$\pi_k = \begin{cases} 1-\rho & k = 0 \\ (1-\rho)(e^{\rho} - 1) & k = 1 \\ (1-\rho)\sum_{i=1}^{k}(-1)^{k-i}e^{i\rho}\left[\frac{(i\rho)^{k-i}}{(k-i)!} + \frac{(i\rho)^{k-i-1}}{(k-i-1)!}\right] & k > 1 \end{cases} \tag{1.41}$$

Vale la pena notare che in questo caso particolare la densità di probabilità del tempo speso in coda e complessivamente nel sistema si può calcolare direttamente dalle probabilità di stato. Per un sistema FIFO il cliente che trova $k$ clienti nel sistema al suo arrivo sperimenta un tempo di permanenza che è composto da $k$ tempi di servizio (derivanti dal servizio dei $k-1$ utenti che lo precedono oltre al suo) più il tempo di servizio residuo dell'utente in servizio al suo arrivo.

### 1.5.3 Il sistema $\mathcal{M}/\mathcal{E}_r/1$

Si consideri un sistema a coda in cui il tempo di servizio $\vartheta$ sia una variabile aleatoria data dalla somma di $r$ variabili aleatorie esponenziali indipendenti aventi tutte lo stesso valor medio $\bar{\vartheta}_0$. Una variabile aleatoria di questo tipo ha una distribuzione di probabilità detta *erlangiana di ordine* $r$, con valor medio $\bar{\vartheta} = r\bar{\vartheta}_0 = r/\mu_0 = 1/\mu$ e varianza $\sigma^2_{\vartheta} = r\sigma^2_{\vartheta_0} = r/\mu_0^2 = 1/r\mu^2$. Le funzioni densità e distribuzione di probabilità erlangiane sono rispettivamente:

$$f_{r,\vartheta}(t) = \frac{(r\mu t)^{r-1}}{(r-1)!} r\mu e^{-r\mu t}$$

e

$$F_{r,\vartheta}(t) = 1 - e^{-r\mu t}\sum_{j=0}^{r-1}\frac{(r\mu t)^j}{j!}$$

In figura 1.2 sono mostrati gli andamenti della $f_{r,\vartheta}(t)$ per diversi valori di $r$ e per $\mu = 1$. Ovviamente $f_{1,\vartheta}(t)$ è la densità di probabilità esponenziale con parametro $\mu$, mentre dalla figura è evidente che per $r \to \infty$ il tempo di servizio diventa deterministico e vale $\vartheta = 1/\mu$. Di conseguenza ci si aspetta che tutti i risultati ottenuti per la coda $\mathcal{M}/\mathcal{E}_r/1$, si riducano a quelli della coda $\mathcal{M}/\mathcal{M}/1$ per $r = 1$ e a quelli della $\mathcal{M}/\mathcal{D}/1$ per $r \to \infty$. Dalla formula PK si ottiene:

$$A_r = \frac{2\rho - \rho^2 + \lambda^2/r\mu^2}{2(1-\rho)} = \frac{\rho}{1-\rho} - \frac{\rho^2(1-1/r)}{2(1-\rho)}$$

Come previsto, si ha (figura 1.3)

$$A_1 = \frac{\rho}{1-\rho}$$

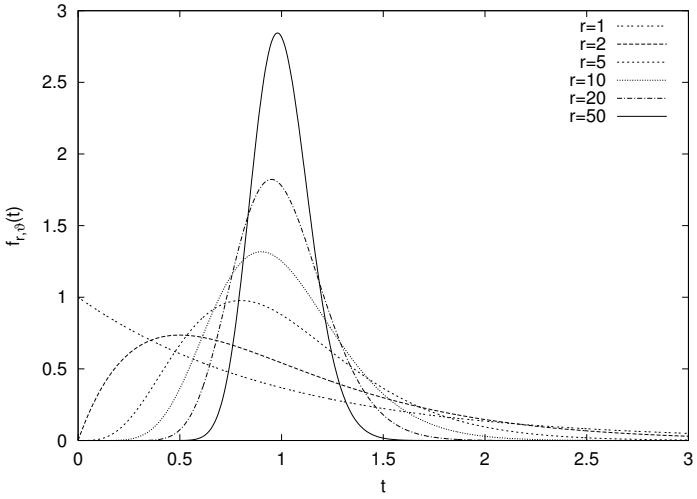


Figura 1.2: Densità di probabilità erlangiane per diversi valori di $r$ calcolate per $\mu = 1$.

e

$$\lim_{r\to\infty} A_r = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)}$$

Applicando la (1.17) si ottiene:

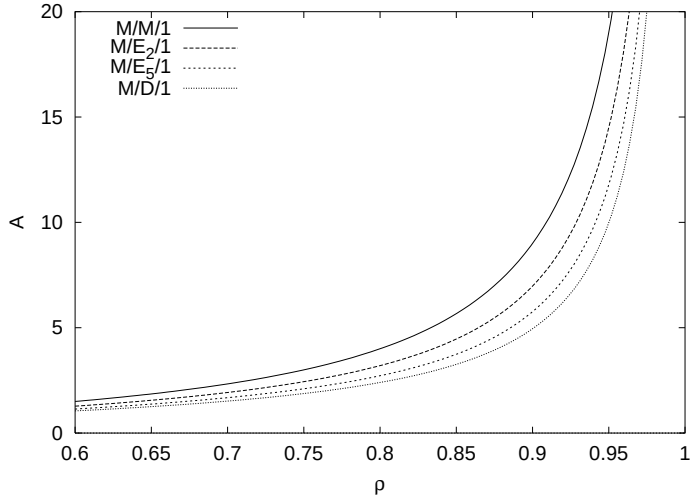


Figura 1.3: Confronto di $A$ in funzione di $\rho$ nei casi $\mathcal{M}/\mathcal{M}/1$, $\mathcal{M}/\mathcal{E}_r/1$ e $\mathcal{M}/\mathcal{D}/1$.

$$\mathcal{L}_{r,\vartheta}(s) = \left(\frac{r\mu}{s+r\mu}\right)^r$$

e quindi

$$\Pi_r(z) = \frac{(1-\rho)(1-z)}{1 - z\left[1 + \frac{(1-z)\rho}{r}\right]^r} \tag{1.42}$$

ottenendo, come previsto:

$$\Pi_1(z) = \frac{1-\rho}{1-\rho z}$$

e

$$\lim_{r\to\infty} \Pi_r(z) = \frac{(1-\rho)(1-z)}{1 - ze^{\rho(1-z)}}$$

In figura 1.4 è rappresentato il tempo medio speso nel sistema $\bar{\delta}_r = A_r/\lambda$ al variare di $r$ confrontandolo con i casi $\mathcal{M}/\mathcal{M}/1$ e $\mathcal{M}/\mathcal{D}/1$.

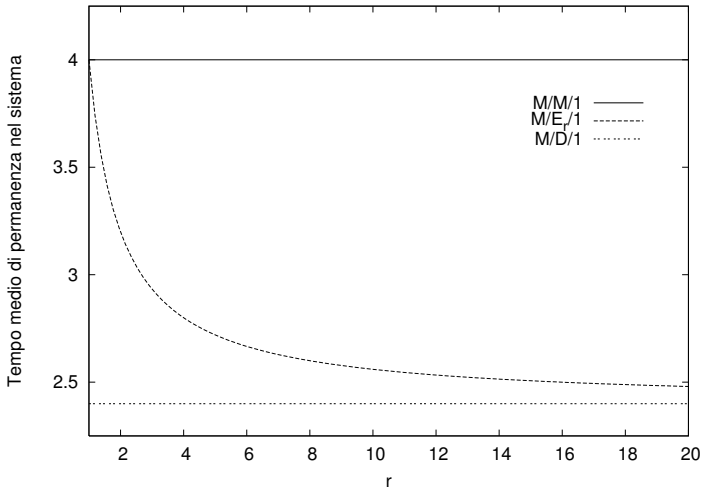


Figura 1.4: Tempo medio speso nel sistema $\mathcal{M}/\mathcal{E}_r/1$ con $\lambda = 1$ e $\rho = 0.8$ al variare di $r$ e confronto con i casi $\mathcal{M}/\mathcal{M}/1$ e $\mathcal{M}/\mathcal{D}/1$.

# Capitolo 2

# Discipline di coda alternative per sistemi $\mathcal{M}/\mathcal{G}/1$

In questo capitolo verranno presentati i principali risultati relativi a sistemi a coda $\mathcal{M}/\mathcal{G}/1$ in cui la politica di accodamento non è più di tipo FIFO. Nel seguito si farà riferimento alle seguenti discipline di accodamento:

- con priorità di tipo non-preemptive
- shortest job next (SJN)
- con priorità di tipo preemptive

## 2.1 Sistemi a coda con priorità non-preemptive

Si vuole studiare un sistema a coda a singolo servitore in cui:

- gli arrivi si compongono di una serie di flussi di traffico poissoniano che numeriamo con l'indice $r = 1, \ldots, R$, per cui la frequenza media di arrivo per ogni flusso sarà $\lambda_r$;
- il tempo di servizio dei clienti è di tipo generale, può essere diverso da classe a classe e, per ogni classe, è rappresentato dalla variabile aleatoria $\vartheta_r$ con valore medio $\bar{\vartheta}_r = 1/\mu_r$;
- la politica di scheduling funziona con priorità:
  - ad ogni flusso di traffico viene assegnata una priorità prestabilita (supponiamo che il flusso di indice minore abbia priorità più elevata);
  - i clienti di un flusso che trovino altri clienti nel sistema vengono serviti dopo i clienti aventi priorità maggiore e prima dei clienti aventi priorità inferiore;

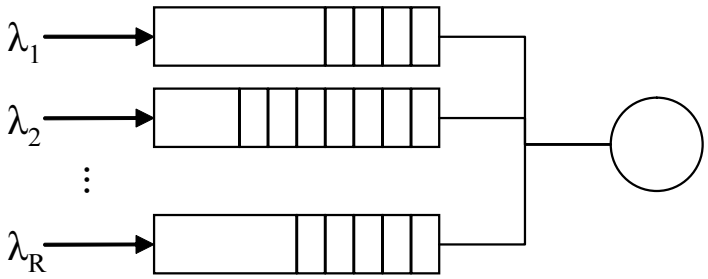


Figura 2.1: Schema di un sistema a coda con priorità.

  - qualora durante l'attesa si verifichi l'arrivo di clienti di priorità maggiore rispetto a quello considerato, questi vengono serviti prima;

- la politica di priorità non influenza le modalità con cui viene realizzato il servizio, cioè non si interrompe il servizio del cliente attualmente servito anche se si verifica l'arrivo di un cliente avente priorità più elevata (politica *non-preemptive*).

Dal punto di vista logico, un sistema a coda di questo tipo può essere schematizzato come in figura 2.1: ogni flusso ha una sua coda dedicata e il servitore preleva un cliente dalla coda $r$ solo se le code da 1 a $r-1$ sono vuote.

Si vuole studiare il tempo medio di attesa per i clienti delle varie classi di priorità. Si può ragionare a questo scopo in modo analogo, anche se leggermente più complesso, a quanto fatto nella sezione 1.3.1. Infatti quando un cliente di classe $r$ arriva al sistema, prima di essere servito deve attendere che:

- termini di essere servito il cliente attualmente in servizio;
- vengano serviti tutti i clienti in attesa appartenenti alle classi di priorità $i$ con $i \leq r$;
- vengano serviti tutti i clienti appartenenti alle classi di priorità $i$ con $i < r$ che arrivano durante il suo tempo di attesa.

Questa considerazione si può porre in formula come segue:

$$\eta_r = T' + \sum_{i=1}^{r} T_i'' + \sum_{i=1}^{r-1} T_i''' \tag{2.1}$$

dove $T'$ è il tempo di servizio residuo del cliente in servizio, $T_i''$ il tempo necessario a servire tutti i clienti con priorità $i$ già presenti nel sistema e $T_i'''$ il tempo necessario a servire tutti i clienti con priorità $i$ giunti durante l'attesa del cliente considerato.

Il valore medio del tempo di attesa in coda per il cliente della classe $r$, che chiameremo $\bar{\eta}_r$, si può calcolare come somma dei valori medi delle quantità presenti a destra del segno di uguale nella (2.1). Questi valori medi vengono calcolati come segue:

- $E[T']$ è il valor medio del tempo residuo di servizio del cliente attualmente servito. Nel caso in cui in servizio ci sia un cliente di classe $i$, dalla (1.6) si ottiene $\bar{\zeta}_i = E[\vartheta_i^2]/2\bar{\vartheta}_i$. Inoltre, la probabilità che un cliente di classe $i$ sia attualmente in servizio è pari al traffico offerto di classe $i$, $\rho_i = \lambda_i/\mu_i$, per cui:

$$E[T'] = \sum_{i=1}^{R} \rho_i \frac{E[\vartheta_i^2]}{2\bar{\vartheta}_i} = \frac{1}{2}\sum_{i=1}^{R} \lambda_i E[\vartheta_i^2]$$

- $E[T_i'']$ si può calcolare tenendo presente che il numero medio di clienti di classe $i$ presenti in coda è dato dal teorema di Little e vale $\lambda_i\bar{\eta}_i$. Poiché poi ogni utente di tipo $i$ richiede mediamente un tempo pari a $\bar{\vartheta}_i$ per essere servito, ne risulta

$$E[T_i''] = \lambda_i\bar{\eta}_i\bar{\vartheta}_i = \rho_i\bar{\eta}_i$$

- $E[T_i''']$ si può calcolare tenendo presente che il numero medio di clienti di classe $i$ che giungono al sistema durante il tempo di attesa del cliente considerato è pari a $\lambda_i\bar{\eta}_r$ e quindi

$$E[T_i'''] = \lambda_i\bar{\eta}_r\bar{\vartheta}_i = \rho_i\bar{\eta}_r$$

In conclusione, si ottiene

$$\bar{\eta}_r = \frac{1}{2}\sum_{i=1}^{R} \lambda_i E[\vartheta_i^2] + \sum_{i=1}^{r} \rho_i\bar{\eta}_i + \sum_{i=1}^{r-1} \rho_i\bar{\eta}_r = \frac{1}{2}\sum_{i=1}^{R} \lambda_i E[\vartheta_i^2] + \sum_{i=1}^{r-1} \rho_i\bar{\eta}_i + \sum_{i=1}^{r} \rho_i\bar{\eta}_r$$

che risolta in $\bar{\eta}_r$ da come risultato

$$\bar{\eta}_r = \frac{\frac{1}{2}\sum_{i=1}^{R} \lambda_i E[\vartheta_i^2] + \sum_{i=1}^{r-1} \rho_i\bar{\eta}_i}{1 - \sum_{i=1}^{r} \rho_i} \tag{2.2}$$

Da questa possiamo ricavare il tempo medio complessivo speso nel sistema $\bar{\delta}_r = \bar{\vartheta}_r + \bar{\eta}_r$. La formula (2.2) può essere riscritta in una forma iterativa adatta al calcolo numerico. A questo scopo usiamo la notazione $S_r = \sum_{i=1}^{r} \rho_i$ e $\bar{\vartheta}_P = E[T']$. Infatti

$$\bar{\eta}_1 = \frac{\bar{\vartheta}_P}{1 - S_1}$$

$$\bar{\eta}_2 = \frac{\bar{\vartheta}_P + \rho_1\bar{\eta}_1}{1 - S_2} = \frac{\bar{\vartheta}_P}{(1 - S_1)(1 - S_2)}$$

e, in generale,

$$\bar{\eta}_r = \frac{\bar{\vartheta}_P}{(1 - S_{r-1})(1 - S_r)} \tag{2.3}$$

che è detta *formula di Cobham*.

La validità della formula di Cobham si può dimostrare per induzione: supponendo che essa valga per $r = j$, occorre dimostrare che vale anche per $r = j + 1$. Dalla (2.2) scritta per $r = j$ si ottiene

$$\sum_{i=1}^{j-1} \rho_i \bar{\eta}_i = (1 - S_j)\bar{\eta}_j - \bar{\vartheta}_P$$

e, sostituendo la precedente nella (2.2) scritta per $r = j + 1$, si ha

$$\bar{\eta}_{j+1} = \frac{\bar{\vartheta}_P + \sum_{i=1}^{j} \rho_i \bar{\eta}_i}{1 - S_{j+1}} = \frac{\bar{\vartheta}_P + \sum_{i=1}^{j-1} \rho_i \bar{\eta}_i + \rho_j \bar{\eta}_j}{1 - S_{j+1}} = \frac{\bar{\vartheta}_P + (1 - S_j)\bar{\eta}_j - \bar{\vartheta}_P + \rho_j \bar{\eta}_j}{1 - S_{j+1}}$$

Dall'ipotesi di induzione si ha quindi

$$\bar{\eta}_{j+1} = \frac{1 - S_j + \rho_j}{1 - S_{j+1}} \bar{\eta}_j = \frac{1 - S_{j-1}}{1 - S_{j+1}} \frac{\bar{\vartheta}_P}{(1 - S_{j-1})(1 - S_j)} = \frac{\bar{\vartheta}_P}{(1 - S_j)(1 - S_{j+1})}$$

### 2.1.1 Esempio per due classi di priorità

Nel caso di due classi di priorità è abbastanza semplice ricavare che:

$$\bar{\eta}_1 = \frac{\bar{\vartheta}_P}{1 - \rho_1}$$

e

$$\bar{\eta}_2 = \frac{\bar{\vartheta}_P}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

Supponendo di avere un sistema con traffico offerto complessivo $\rho = \rho_1 + \rho_2 = 0.8$ e di variare la percentuale di traffico a priorità 1 da 0 al 100%, si può scrivere $\rho_1 = x\rho$ e $\rho_2 = (1 - x)\rho$ con $x \in [0 : 1]$. Inoltre, si assume che le due classi di priorità abbiano la stessa distribuzione dei tempi di servizio e, per semplicità, si ipotizza $\bar{\vartheta} = 1$ per ciascuna delle due classi. Di conseguenza, $\lambda_1 = 0.8x$ e $\lambda_2 = 0.8(1 - x)$. Ne consegue che

$$\bar{\vartheta}_P = \frac{1}{2}(0.8x + 0.8(1 - x))E[\vartheta^2] = 0.4E[\vartheta^2]$$

e quindi che

$$\bar{\eta}_1 = \frac{0.4E[\vartheta^2]}{1 - 0.8x}$$

$$\bar{\eta}_2 = \frac{0.4E[\vartheta^2]}{(1 - 0.8x)(1 - 0.8)}$$

La figura 2.2 mostra l'andamento dei tempi di attesa al variare della percentuale di traffico ad alta priorità $x$, per un sistema $\mathcal{M}/\mathcal{M}/1$, in cui $E[\vartheta^2] = 2$, confrontato con il tempo di attesa della coda senza priorità e stesso traffico totale. Si noti come per $x = 0$ e per $x = 1$ i tempi di attesa per la sola classe rimasta abbiano lo stesso valore del caso senza priorità, come lecito aspettarsi.

Inoltre si noti come, per $x$ piccolo, si abbia un sostanziale miglioramento del tempo di attesa per il traffico prioritario, senza intaccare in modo significativo le prestazioni per il traffico non prioritario.

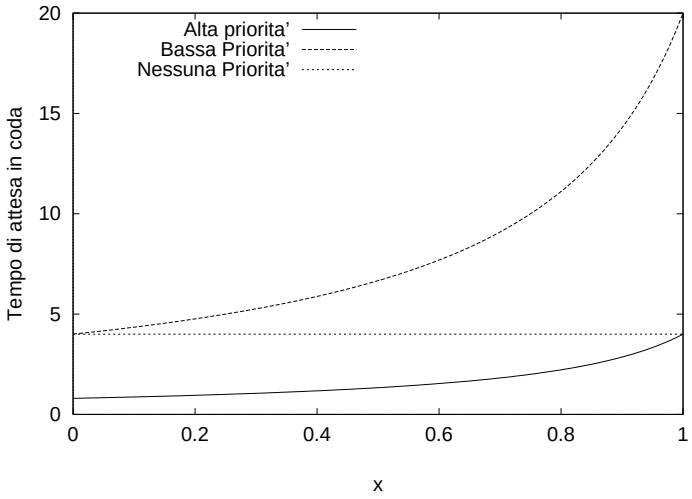


Figura 2.2: $\mathcal{M}/\mathcal{M}/1$ con due classi di priorità e traffico offerto complessivo $\rho = 0.8$: tempo di attesa in coda per le due classi in funzione della percentuale di traffico ad alta priorità $x$ e confronto con il caso senza priorità.

### 2.1.2 La legge di conservazione di Kleinrock

La reciproca influenza dei tempi di attesa in coda di flussi diversi, come mostrato nell'esempio di figura 2.2, si verifica per un numero generico $R$ di classi ed è ovviamente conseguenza del fatto che i clienti meno prioritari devono aspettare che vengano prima serviti tutti i clienti a maggiore priorità, anche se arrivati dopo. Si noti comunque che la politica di accodamento a priorità di tipo non-preemptive non fa altro che modificare l'ordine con cui si servono i clienti in attesa, ma non modifica il servizio in sé. Più in generale, se si fa l'ipotesi che il servizio sia indipendente dalla disciplina di coda, è possibile stabilire la relazione che intercorre tra i diversi tempi di attesa in coda delle $R$ classi di clienti.

A questo scopo, si definisce *funzione lavoro rimanente* la funzione $U(t)$ che indica il tempo richiesto per smaltire i clienti presenti nel sistema all'istante $t$, cioè il tempo che sarebbe necessario per svuotare il sistema a partire dall'istante $t$ se non ci fossero altri arrivi successivi. La funzione $U(t)$ è sempre non negativa, decresce linearmente con pendenza $-1$ quando è positiva e presenta una discontinuità in corrispondenza dell'istante di arrivo di ciascun cliente, con un salto di entità pari al tempo di servizio richiesto dal nuovo cliente. Un esempio è mostrato in figura 2.3, in cui sono indicate le discontinuità in corrispondenza degli arrivi negli istanti $t_1, t_2, t_3, t_4$ e $t_5$.

Per definizione, $U(t)$ è data dalla somma del tempo residuo di servizio dell'utente eventualmente in servizio all'istante $t$ più la somma dei tempi di servizio dei clienti eventualmente presenti in coda. Nel caso di arrivi composti da $R$ flussi di traffico, usando gli

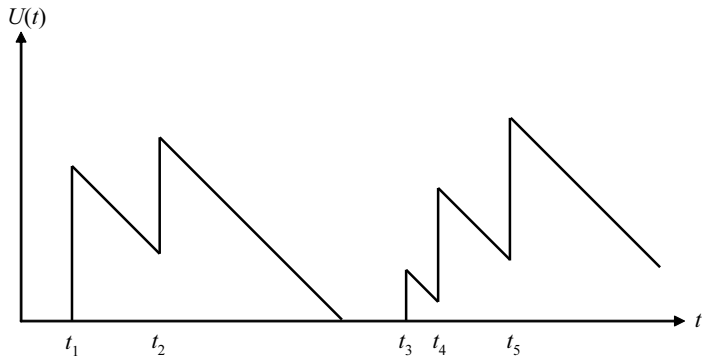


Figura 2.3: Andamento della funzione lavoro rimanente.

stessi simboli adottati nella (2.1), si ha

$$U(t) = T' + \sum_{i=1}^{R} T_i''$$

da cui, calcolando il valor medio, si ottiene

$$\bar{U} = E[U(t)] = E[T'] + \sum_{i=1}^{R} E[T_i''] = \bar{\vartheta}_P + \sum_{i=1}^{R} \rho_i \bar{\eta}_i \tag{2.4}$$

Se vale l'ipotesi che il tempo di servizio è indipendente dalla disciplina di coda, come nel caso di accodamento con priorità non-preemptive, allora l'andamento di $U(t)$ è anch'esso indipendente dalla disciplina di coda e il suo valor medio può essere calcolato semplicemente considerando il caso di coda $\mathcal{M}/\mathcal{G}/1$ con disciplina FIFO soggetta allo stesso carico di quella con priorità, cioè tale per cui

$$\lambda = \sum_{i=1}^{R} \lambda_i \qquad \bar{\vartheta} = \sum_{i=1}^{R} \frac{\lambda_i}{\lambda} \bar{\vartheta}_i \qquad \rho = \sum_{i=1}^{R} \rho_i \qquad E[\vartheta^2] = \sum_{i=1}^{R} \frac{\lambda_i}{\lambda} E[\vartheta_i^2]$$

Nel caso di coda FIFO, per la proprietà PASTA il tempo medio necessario per smaltire tutti gli utenti presenti nel sistema è pari al tempo medio di attesa in coda di un utente generico dato dalla (1.12), cioè

$$\bar{U} = \bar{\eta} = \frac{\lambda E[\vartheta^2]}{2(1-\rho)} = \frac{\frac{1}{2}\sum_{i=1}^{R} \lambda_i E[\vartheta_i^2]}{1-\rho} = \frac{\bar{\vartheta}_P}{1-\rho} \tag{2.5}$$

Confrontando la (2.4) e la (2.5) si ricava

$$\sum_{i=1}^{R} \rho_i \bar{\eta}_i = \bar{\vartheta}_P \frac{\rho}{1-\rho} \tag{2.6}$$

che è detta *legge di conservazione di Kleinrock*. Il secondo membro della (2.6) è costante al variare della disciplina di coda, il che dimostra l'invarianza della somma pesata dei tempi medi di attesa in coda delle varie classi al variare della politica di accodamento. Se ne deduce, quindi, che la riduzione del tempo medio di attesa in coda di una classe di utenti ha ripercussioni su quello degli altri nei limiti della legge di conservazione.

Applicando la legge di conservazione all'esempio del paragrafo 2.1.1 si ottiene

$$\rho_1\bar{\eta}_1 + \rho_2\bar{\eta}_2 = 1.6E[\vartheta^2]$$

che è indipendente dalla disciplina di coda.

## 2.2 Scheduling con politica Shortest Job Next (SJN)

In questa sezione si intende studiare un sistema a coda $\mathcal{M}/\mathcal{G}/1$ in cui la politica di scheduling favorisca i clienti con il più piccolo tempo di servizio. Ad esempio, in un sistema di commutazione o multiplazione a pacchetto, si può ipotizzare che il sistema di trasmissione favorisca i pacchetti brevi rispetto a quelli lunghi, selezionandoli prioritariamente dalla coda di attesa.

È facile comprendere come questo sistema sia sostanzialmente analogo ad un sistema con priorità non-preemptive, in cui la priorità viene assegnata non in modo discreto per classe di priorità, ma in modo *continuo* sulla base della lunghezza del tempo di servizio. Di conseguenza, è possibile ragionare in modo assolutamente analogo a quanto fatto nel paragrafo 2.1, considerando la classe di priorità $r$ composta dai clienti tali per cui $r \leq \vartheta < r + dr$. Il traffico per ciascuna classe di priorità è facilmente calcolabile tenendo conto che

$$\lambda_r = \lambda \Pr\{r \leq \vartheta < r + dr\} = \lambda f_\vartheta(r)dr$$

per cui, tenendo conto che per la classe di priorità $r$ vale $\vartheta_r = r$,

$$\rho_r = \lambda_r\vartheta_r = \lambda r f_\vartheta(r)dr$$

È possibile scrivere una formula del tutto analoga alla (2.1):

$$\eta_r = T' + \int_0^r T_t''dt + \int_0^r T_t'''dt$$

Il tempo di servizio residuo medio risulta pari a quello di una coda $\mathcal{M}/\mathcal{G}/1$ espresso dalla (1.10), per cui

$$E[T'] = \bar{\vartheta}_P = \rho\bar{\zeta} = \rho\frac{E[\vartheta^2]}{2\bar{\vartheta}}$$

Analogamente a quanto fatto nella sezione 2.1, definita la somma del traffico delle classi di priorità fino alla $r$ come

$$S_r = \int_0^r \lambda t f_\vartheta(t)dt$$

si può pervenire ad un'equazione analoga alla formula di Cobham

$$\bar{\eta}_r = \frac{\bar{\vartheta}_P}{(1-S_r)^2} = \frac{\lambda E[\vartheta^2]}{2(1-S_r)^2} \tag{2.7}$$

Il tempo medio di attesa in coda per un utente generico si ottiene mediando $\bar{\eta}_r$ su tutti i possibili valori del tempo di servizio, per cui risulta

$$\bar{\eta} = \int_0^\infty \bar{\eta}_r f_\vartheta(r) dr$$

Anche in questo caso, essendo il sistema di tipo non-preemptive, vale una legge di conservazione analoga alla (2.6)

$$\int_0^\infty \rho_r \bar{\eta}_r = \int_0^\infty \lambda r \bar{\eta}_r f_\vartheta(r) dr = \vartheta_P \frac{\rho}{1-\rho}$$

## 2.3 Sistemi a coda con priorità preemptive

Analogamente a quanto fatto nel paragrafo 2.1, si consideri un sistema a coda a singolo servitore soggetto ad un traffico composto da $R$ classi di utenti. Tutte le ipotesi restano invariate, tranne per il fatto che il servizio del cliente attualmente servito viene interrotto nel caso in cui si verifichi l'arrivo di un cliente avente priorità maggiore, che viene quindi servito immediatamente (politica *preemptive*).

In un sistema del genere non vale più l'ipotesi di conservazione del lavoro rimanente fatta nel paragrafo 2.1.2, poiché il tempo di servizio dipende dalla politica di accodamento. Infatti, se un utente in servizio viene interrotto a causa dell'arrivo di un altro più prioritario, il suo servizio viene temporaneamente sospeso per poi essere ripreso successivamente. Il risultato è che il tempo di servizio si dilata di una quantità che dipende dagli eventuali arrivi di utenti prioritari e, quindi, da come questi vengono accodati nel sistema.

Per studiare un sistema siffatto si deve innanzitutto specificare cosa avvenga dei clienti il cui servizio viene interrotto. A seconda del tipo di sistema che si sta descrivendo, si può assumere che il servizio interrotto venga ripreso ricominciando dall'inizio, con un nuovo tempo di servizio uguale a quello iniziale (modalità *preemptive repeat identical*) oppure diverso ma comunque generato dalla distribuzione dei tempi di servizio (modalità *preemptive repeat different*). In realtà, la scelta più efficiente è quella di riprendere il servizio dal punto in cui era stato interrotto (modalità *preemptive resume*).

In questo caso quando un cliente di classe $r$ arriva al sistema, prima di iniziare ad essere servito deve attendere che:

- termini di essere servito il cliente attualmente in servizio solo se di priorità $i \leq r$;
- vengano serviti tutti i clienti in attesa appartenenti alle classi di priorità $i$ con $i \leq r$;

- vengano serviti tutti i clienti appartenenti alle classi di priorità $i$ con $i < r$ che arrivano durante il suo tempo di attesa.

Inoltre, una volta che il cliente è entrato in servizio, dovrà interrompersi ed attendere che vengano serviti anche tutti i clienti appartenenti alle classi di priorità $i$ con $i < r$ che arrivano durante il suo tempo di servizio.

Per il calcolo del tempo di attesa in coda dell'utente generico di classe $r$ si può ragionare in maniera simile al caso non-preemptive:

$$\eta_r = T'_r + \sum_{i=1}^{r} T''_i + \sum_{i=1}^{r-1} T'''_i \tag{2.8}$$

i cui termini hanno lo stesso significato della (2.1). Si noti che il tempo residuo di servizio $T'_r$ è diverso da quello del caso non-preemptive e dipende da $r$, poiché $T'_r$ è nullo se in servizio c'è un utente di classe superiore a $r$. Di conseguenza si ha

$$E[T'_r] = \bar{\vartheta}_{Pr} = \frac{1}{2} \sum_{i=1}^{r} \lambda_i E[\vartheta_i^2]$$

Le due sommatorie presenti nella (2.8) sono esattamente le stesse del caso non-preemptive, per cui è possibile scrivere

$$\bar{\eta}_r = \frac{\bar{\vartheta}_{Pr}}{(1 - S_{r-1})(1 - S_r)}$$

Nel caso di clienti a massima priorità ($r = 1$), è immediato verificare che il tempo di attesa in coda è pari a quello che si otterrebbe se il sistema fosse soggetto al solo carico $\rho_1$, cioè

$$\bar{\eta}_1 = \frac{\lambda_1 E[\vartheta_1^2]}{2(1 - \rho_1)}$$

Di conseguenza, la modalità preemptive ha il vantaggio di fornire prestazioni ottimali ai clienti più privilegiati, come se fossero gli unici ad utilizzare il servitore. Questo ovviamente a scapito delle classi meno prioritarie che aspettano di più.

A questo punto resta da stabilire quanto vale la dilatazione del tempo di servizio medio $\bar{\vartheta}_r$ dovuta alle interruzioni del servizio. Per questo occorre tenere presente che, durante il servizio medio di un cliente di classe $r$, il numero medio di arrivi di utenti di classe $i < r$ vale $\lambda_i \bar{\vartheta}_r$ e che ciascuno di essi verrà servito per un tempo medio pari a $\bar{\vartheta}_i$. Quindi le interruzioni del servizio dell'utente di classe $r$ producono un'attesa aggiuntiva media pari a

$$\sum_{i=1}^{r-1} \lambda_i \bar{\vartheta}_r \bar{\vartheta}_i = \sum_{i=1}^{r-1} \rho_i \bar{\vartheta}_r = S_{r-1} \bar{\vartheta}_r$$

Ma anche durante questo periodo, in cui il servizio dell'utente di classe $r$ è sospeso, potrebbero arrivare dei clienti con priorità $i < r$ che dovranno essere serviti prima, allungando ulteriormente il periodo di sospensione di una quantità media pari a $S_{r-1}^2 \bar{\vartheta}_r$,

durante la quale altri arrivi prioritari potrebbero verificarsi. Iterando il ragionamento, si comprende facilmente come il tempo medio necessario per completare il servizio di un utente di classe $r$ valga

$$\bar{\vartheta}_{Cr} = \bar{\vartheta}_r + S_{r-1}\bar{\vartheta}_r + S_{r-1}^2\bar{\vartheta}_r + S_{r-1}^3\bar{\vartheta}_r + \ldots = \bar{\vartheta}_r \sum_{i=0}^{\infty} S_{r-1}^i = \frac{\bar{\vartheta}_r}{1 - S_{r-1}}$$

da cui si evince che il tempo medio di servizio subisce una dilatazione di un fattore $(1 - S_{r-1})^{-1}$ e che la politica preemptive produce un attesa aggiuntiva pari a

$$\bar{\vartheta}_{Cr} - \bar{\vartheta}_r = \frac{S_{r-1}}{1 - S_{r-1}}\ \bar{\vartheta}_r$$

In conclusione, il tempo medio speso all'interno del sistema dall'utente di classe $r$ vale

$$\bar{\delta}_r = \bar{\vartheta}_r + \frac{S_{r-1}}{1 - S_{r-1}}\ \bar{\vartheta}_r + \frac{\bar{\vartheta}_{Pr}}{(1 - S_{r-1})(1 - S_r)}$$

# Appendice A

# Derivazione classica della formula di Pollaczek-Khintchine

In questa appendice si riporta il metodo tradizionalmente usato in letteratura per ricavare la formula (1.14). Si prenda il valor medio dell'equazione (1.21):

$$E[k_{n+1}] = E[k_n] - E[u(k_n)] + E[a_{n+1}]$$

In equilibrio $E[k_{n+1}] = E[k_n]$ per cui, ricordando la (1.18), si ottiene

$$E[a] = \bar{a} = E[u(k_n)] = \rho$$

relazione che si può anche ottenere direttamente come segue:

$$E[u(k)] = \sum_{k=0}^{\infty} P_k u(k) = \sum_{k=1}^{\infty} P_k = 1 - P_0 = \rho$$

Ora si prenda il valor medio del quadrato dell'equazione (1.21):

$$E[k_{n+1}^2] = E[k_n^2] + E[u(k_n)^2] + E[a_{n+1}^2] - 2E[k_n u(k_n)] + 2E[k_n a_{n+1}] - 2E[u(k_n)a_{n+1}]$$

Di nuovo all'equilibrio $E[k_{n+1}^2] = E[k_n^2]$. Rimangono da calcolare i termini rimanenti:

1. $E[u(k_n)^2]$: poiché $u(k)^2 = u(k)$ se ne deduce $E[u(k_n)^2] = E[u(k_n)] = \rho$;
2. $E[a_{n+1}^2]$: in condizioni di equilibrio e ricordando la (1.19)

$$E[a^2] = \rho + \lambda^2 \sigma_\vartheta^2 + \rho^2$$

3. $E[k_n u(k_n)]$: poiché $ku(k) = k, \forall k \geq 0$, all'equilibrio è semplicemente $E[k_n u(k_n)] = E[k_n] = A$, cioè il numero medio di utenti nel sistema;
4. $E[k_n a_{n+1}]$: per questo termine, tenendo presente che gli arrivi al sistema sono indipendenti dallo stato del sistema stesso, si deduce che

$$E[k_n a_{n+1}] = E[k_n]E[a_{n+1}] = A\rho$$

5. $E[u(k_n)a_{n+1}]$: per le stesse ragioni del punto precedente si ottiene

$$E[u(k_n)a_{n+1}] = E[u(k_n)]E[a_{n+1}] = \rho^2$$

In conclusione si ottiene l'equazione

$$0 = \rho + \rho + \lambda^2\sigma_\vartheta^2 + \rho^2 - 2A + 2A\rho - 2\rho^2$$

che, risolta per $A$, fornisce come risultato la formula di Pollaczek-Khintchine:

$$A = \frac{2\rho - \rho^2 + \lambda^2\sigma_\vartheta^2}{2(1-\rho)}$$