NOTE DAL CORSO DI

SISTEMI AD EVENTI DISCRETI

ANNO ACCADEMICO 2019/20

10 dicembre 2019

CAPITOLO 4° : CODE E RETI DI CODE

Docente: Riccardo Minciardi

SOMMARIO

1 *Definizioni e proprietà generali*

1.1 I modelli

Una coda (singola) è un sistema costituito da uno o più servitori (identici) e da una fila di attesa per i clienti (Figura 1). I clienti si assumono tutti appartenenti alla stessa classe.

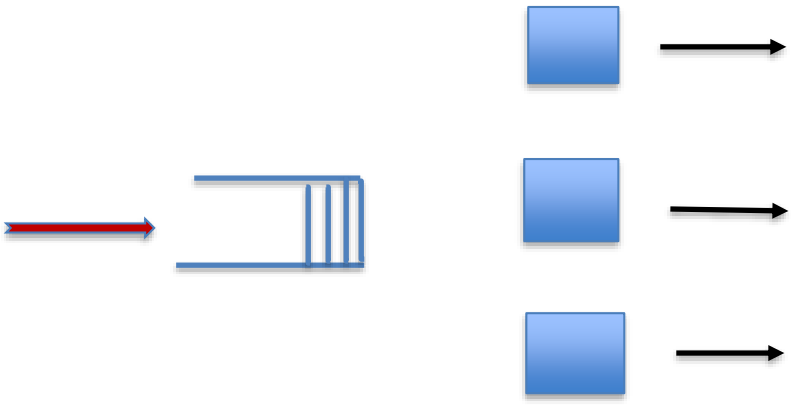

Figura 1: Rappresentazione schematica di una coda singola.

Una sistema a coda (o, più semplicemente, una coda) è caratterizzato da:

- una statistica del processo degli arrivi;
- una statistica del processo dei servizi;
- il numero dei servitori (tutti equivalenti);
- la dimensione del buffer in cui risiedono i clienti in attesa;
- la popolazione complessiva dei clienti;
- la disciplina (politica) di servizio.

Si suppone che l'assegnazione dei clienti ai diversi servitori segua procedure che non privilegino e penalizzino alcun servitore. Si suppone che il sistema operi in regime "work-conserving" (ovvero nessun servitore può essere inattivo se c'è qualche cliente in attesa del servizio). Si suppone inoltre che ogni cliente lasci immediatamente la coda, una volta che il suo servizio è stato completato.

Per individuare sinteticamente il modello di coda, si utilizza la notazione di Kendall

A/B/m/K/H

in cui:

A si riferisce alla statistica degli arrivi

B si riferisce alla statistica dei servizi

m specifica il numero dei servitori in parallelo

K definisce la dimensione del buffer

H definisce la popolazione complessiva dei clienti.

In genere, in questa notazione, non viene indicata la politica di servizio. Inoltre, quando K e H non sono specificati si sottintende che abbiano valore ∞. Per quanto riguarda le statistiche (degli arrivi e dei servizi) si usano le seguenti sigle:

- M $\rightarrow$ Markovian
- D $\rightarrow$ Deterministic
- N $\rightarrow$ Normal
- E $\rightarrow$ Erlangian
- G $\rightarrow$ General
- GI $\rightarrow$ General Independent

La statistica M è caratterizzata da un tempo di interarrivo o da un tempo di servizio con distribuzione esponenziale. Col termine GI si intende una distribuzione in cui i tempi, ad esempio di interarrivo, costituiscono una sequenza di variabili aleatorie indipendenti tutte distribuite in modo identico, secondo una pdf sulla quale non si fa alcuna assunzione. Anche M, D, N, ed E saranno intese come distribuzioni di sequenze di variabili aleatorie indipendenti e identicamente distribuite. Infine, quando non viene detto esplicitamente il contrario, il processo degli arrivi e quello dei servizi sono considerati indipendenti.

Lo studio dei sistemi a coda e delle reti di code è finalizzato alla determinazione delle proprietà statistiche di alcune grandezze di interesse. A questo proposito, occorre introdurre la seguente notazione:

- $L(t)$ = la lunghezza della coda all'istante di tempo t (equivale al numero dei clienti presenti nel sistema all'istante t, siano essi in servizio o in attesa);
- $L_w(t)$ = il numero di clienti in attesa all'istante t (equivale a $L(t)$ meno il numero di clienti in servizio all'istante t);
- $t_{q,i}$ è il tempo complessivo di permanenza nel sistema (ovvero tempo di attesa più tempo di servizio) dell'i-esimo cliente (l'ordine si riferisce all'arrivo nel sistema);
- $t_{w,i}$ è il tempo di attesa dell'i-esimo cliente;
- $t_{s,i}$ è il tempo di servizio dell'i-esimo cliente;
- a_i è l'istante di arrivo dell'i-esimo cliente;
- d_i è l'istante di partenza (dal sistema) dell'i-esimo cliente.

Ovviamente risulterà:

$t_{q,i} = t_{w,i} + t_{s,i}$

$t_{q,i} = d_i - a_i$

Per un sistema a coda con m servitori in parallelo si definisce coefficiente di carico ρ il rapporto

$$\rho = \frac{E[t_s]}{m\,E[t_a]}$$

Indicando con t_s la variabile aleatoria "tempo di servizio dei clienti" e con t_a la variabile aleatoria "tempo di interarrivo dei clienti". Ponendo $\lambda = 1/E[t_a]$ e $\mu = 1/E[t_s]$ (rispettivamente, frequenza degli arrivi e frequenza massima di servizio) si ha anche

$$\rho = \frac{\lambda}{m\,\mu} \qquad (1)$$

Per sistemi con buffer illimitati (in cui quindi i clienti che arrivano non possono mai essere rifiutati), condizione sufficiente di stabilità è che sia

$$0 \leq \rho < 1$$

(invece $\rho > 1$ è condizione sufficiente di instabilità).

Nel caso in cui $m > 1$, si è già detto che si suppone che le politiche di servizio non privilegino alcuno dei servitori. Possiamo quindi considerare la probablità $\tilde{\pi}_0$ che un generico servitore sia inattivo in un certo istante selezionato a caso, in condizioni di equilibrio stocastico.

Definiamo l'utilizzazione del servitore generico come $U = 1 - \tilde{\pi}_0$. Il throughput di un generico servitore (cioè il numero medio di clienti che escono dal servitore nell'unità di tempo) è evidentemente $(1 - \tilde{\pi}_0)\mu$.

Quindi il throughput complessivo del sistema è dato da $(1 - \tilde{\pi}_0)\mu m$.

Poiché si assume che il sistema si trovi in condizioni di equilibrio (stocastico), sarà $\lambda = (1 - \tilde{\pi}_0)\mu m$ e quindi si ha $\rho = 1 - \tilde{\pi}_0$, cioè ρ corrisponde alla sopra definita utilizzazione.

1.2 Le leggi fondamentali

Per code del tutto generali possono essere forniti solo pochi risultati. Due di questi sono la legge di Lindley e la legge di Little.

La legge di Lindley riguarda esclusivamente il transitorio. Per un sistema ad un solo servitore, governato da un politica FIFO, si può scrivere

$$d_k = max\{a_k, d_{k-1}\} + t_{s,k} \tag{2}$$

Assai più importante è la legge di Little, che si riferisce a sistemi a coda, in condizioni di equilibrio stocastico, con un generico numero di servitori. Tale legge definisce un legame fra tre quantità medie:

- $\overline{L}$: numero medio di clienti presenti nel sistema ($\overline{L} = E[L(t)]$);
- $\overline{t}_q$: tempo medio di permanenza nel sistema da parte dei clienti ($\overline{t}_q = E[t_{q,i}]$);
- λ: flusso medio o frequenza degli arrivi ($\lambda = 1/E[t_a]$).

Tale legame di esprime semplicemente come

$$\overline{L} = \lambda \overline{t}_q \tag{3}$$

e la sua validità può essere dimostrata in condizioni del tutto generali per quanto riguarda le statistiche dei processi degli arrivi e dei servizi, il numero di servitori, la politica di servizio, etc.

In realtà, la validità della legge di Little può essere dimostrata in riferimento a qualsiasi sistema (in equilibrio stocastico) in cui possa essere identificato un processo di arrivi (di clienti) e possa essere chiaramente individuato un sistema, distinto da quanto si trova all'esterno del sistema stesso. In particolare, la legge di Little vale anche per il sistema costituito dalla fila di attesa. In tal caso, la legge di Little si scrive, con ovvio significato dei simboli, come

$$\overline{L_w} = \lambda \bar{t}_w \qquad (4)$$

S noti che, sulla base delle (3),(4), e tenendo conto che ovviamente $\bar{t}_q = \bar{t}_w = +\bar{t}_s$, si ottiene

$\bar{L}/\lambda = \overline{L_w}/\lambda + 1/\mu$ ovvero $\bar{L} = \overline{L_w} + \lambda/\mu$ ovvero $\bar{L} = \overline{L_w} + m\rho$

2 *Code markoviane*

2.1 Generalità sulle code markoviane

Una coda markoviana è una coda in cui il processo degli arrivi e il processo dei servizi sono processi di Poisson. Quindi t_q e t_s sono variabili aleatorie distribuite in modo esponenziale. Un sistema a coda di questo genere corrisponde ad una catena di Markov a tempo continuo (CTMC) per cui è possibile determinare la distribuzione di probabilità a regime dello stato.

Prima di entrare nei dettagli per quanto riguarda i risultati che valgono per le code Markoviane, è opportuno menzionare una proprietà interessante dal punto di vista statistico.

Consideriamo le due seguenti distribuzioni di probabilità (in condizioni di equilibrio stocastico):

- $\pi_n = Pr\{\mathrm{L(t)} = \mathrm{n}\}$, *essendo t un istante di tempo generico*;
- $\alpha_n = Pr\{un\ cliente\ che\ arriva\ nel\ sistema\ trovi\ \mathrm{n}\ clienti\ nel\ sistema\}$.

Per sistemi a coda generali, le due distribuzioni $\pi_n(t)$ e $\alpha_n(t)$ sono in generali differenti. Sussiste invece la proprietà espressa dal seguente risultato, che viene fornito senza dimostrazione.

Risultato 1 *In un sistema a coda (in equilibrio stocastico) con arrivi rappresentati da un processo di Poisson, indipendentemente dal processo dei servizi, la probabilità che un cliente al momento del suo arrivo trovi n clienti nel sistema è uguale alla probabilità che, in un istante del tutto generico t, il numero di clienti presenti nel sistema sia pari a n. In altre parole*

$$\pi_n = \alpha_n \qquad \mathrm{n} = 0,1,2,3 \ldots$$

La proprietà definita nel precedente risultato è spesso indicata con il termine "*proprietà PASTA*" (Poisson Arrivals See Time Averages).

2.2 Coda M / M / 1

Si tratta del sistema a coda markoviano più semplice. Esso può essere modellato come un processo birth-death a tempo continuo (si veda il capitolo sulle Catene di Markov) in cui $\lambda_n = \lambda$, $n = 0,1,2,..$, e $\mu_n = \mu$, $n = 0,1,2,\ldots$.

Si assume $\rho = \lambda/\mu < 1$. Pertanto in base a quanto viene detto nel paragrafo 3.3 del capitolo sulle Catene di Markov, tutti gli stati della CTMC sono ricorrenti positivi. Esiste quindi una distribuzione di probabilità a regime dello stato data da

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n} \tag{5}$$

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0 \qquad n \geq 1 \tag{6}$$

Poiché $\lambda/\mu < 1$ si ha

$$\sum_{j=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} \tag{7}$$

e quindi le (5)-(6) si possono riscrivere, rispettivamente, come

$$\pi_0 = \frac{1}{1 + \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu} = 1 - \rho \tag{8}$$

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) = (\rho)^n (1 - \rho) \tag{9}$$

(si noti che la (9) è valida anche per $n = 0$, conglobando in tal modo la (8)).

E' possibile allora determinare $\overline{L}$ (lunghezza della coda media, ovvero numero medio di clienti nel sistema)

$$\overline{L} = \sum_{n=0}^{\infty} n\pi_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n \tag{10}$$

Osserviamo che

$$\frac{d}{d\rho}\sum_{n=0}^{\infty}\rho^n = \sum_{n=0}^{\infty} n\rho^{n-1} = \frac{1}{\rho}\sum_{n=0}^{\infty} n\rho^n \tag{11}$$

D'altra parte, dato che si suppone $\rho < 1$, si ha

$$\sum_{n=0}^{\infty}\rho^n = \frac{1}{1-\rho} \tag{12}$$

e quindi

$$\frac{d}{d\rho}\sum_{n=0}^{\infty}\rho^n = \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{1}{(1-\rho)^2} \tag{13}$$

Confrontando la (11) e la (13) si ha allora

$$\sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{(1-\rho)^2} \tag{14}$$

e quindi, sostituendo nella (10), si ottiene finalmente

$$\overline{L} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \tag{15}$$

Dalla legge di Little (3) si ha allora

$$\overline{t_q} = \frac{\overline{L}}{\lambda} = \frac{\frac{1}{\mu}}{1-\rho} = \frac{1}{\mu-\lambda} \tag{16}$$

Inoltre

$$\overline{t_w} = \overline{t_q} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)} \tag{17}$$

e, per la legge di Little (4),

$$\overline{L}_w = \lambda \overline{t}_w = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} \tag{18}$$

2.3 Coda M / M / m

Si tratta della generalizzazione, al caso di *m* servitori in parallelo, del sistema considerato al paragrafo precedente. Il sistema M / M / m può ancora essere modellato con una CTMC-BD, in cui però i rate di nascita e morte non sono indipendenti dallo stato. Semplici considerazioni mostrano infatti come in questo caso si abbia

$$\lambda_n = \lambda \qquad n = 01,2,\dots$$

$$\mu_n = \begin{cases} n\mu & n = 1,2,\dots,m-1 \\ m\mu & n \geq m \end{cases}$$

La dipendenza da *n* del death rate μ_n si spiega evidentemente con il fatto che più elevato è il numero di server attivi, più elevato deve essere μ_n, fino ad un valore di *n* pari ad *m*. Da $n = m$ in poi, il numero di server attivi rimane costante, e quindi anche μ_n deve rimanere costante.

Ricordando la condizione menzionata nel paragrafo 3.3 del capitolo sulle Catene di Markov, se supponiamo $\rho = \lambda/m\mu < 1$, esiste $\bar{\jmath}$ tale che $\lambda_j/\mu_j < 1$ per $j \geq \bar{\jmath}$ (basta prendere un qualunque $\bar{\jmath} \geq m$), e quindi tutti gli stati sono ricorrenti positivi. E' quindi possibile determinare la distribuzione di probabilità a regime dello stato, utilizzando le (31)-(32) del capitolo sulle Catene di Markov.

Dopo alcuni passaggi, si ottiene

$$\pi_0 = \left[1 + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!}\frac{1}{1-\rho}\right]^{-1} \tag{19}$$

$$\pi_n = \begin{cases} \pi_0 \dfrac{(m\rho)^n}{n!} & n = 1,2,\dots,m-1 \\ \pi_0 \dfrac{m^m}{m!}\rho^n & n = m, m+1, \dots \end{cases} \tag{20}$$

Una volta calcolata la distribuzione di probabilità a regime dello stato, è possibile determinare (i calcoli sono omessi per brevità) la lunghezza di coda media

$$\overline{L} = \sum_{n=0}^{\infty} n\pi_n = m\rho + \frac{(m\rho)^m}{m!}\frac{\rho}{(1-\rho)^2}\,\pi_0 \tag{21}$$

(naturalmente per $m = 1$ si ritrova il valore di $\overline{L}$ dato dalla (15)). Sulla base della legge di Little (3) è possibile determinare $\overline{t_q}$, che è dato da

$$\overline{t_q} = \frac{1}{\mu} + \frac{1}{\mu}\,\frac{(m\rho)^m}{m!}\,\frac{\pi_0}{m(1-\rho)^2} \tag{22}$$

Possono inoltre essere immediatamente determinati i valori di $\overline{t_w}$ e $\overline{L_w}$, che qui non vengono riportati per brevità.

Se indichiamo con *S* la variabile aleatoria che indica il numero dei servitori del sistema che sono attivi, si può determinare il suo valor medio come segue

$$E[S] = \sum_{n=0}^{m-1} n\pi_n + mPr\{L \geq m\} \tag{23}$$

dove

$$Pr\{L \geq m\} = \sum_{n=m}^{\infty} \pi_n = \sum_{n=m}^{\infty} \frac{m^m}{m!}\rho^n\,\pi_0 = \frac{m^m}{m!}\frac{\rho^m}{1-\rho}\,\pi_0 \tag{24}$$

Sostituendo nella (23), dopo alcuni passaggi si ottiene

$$E[S] = m\rho = \frac{\lambda}{\rho} \tag{25}$$

Si noti che la (24) esprime anche la cosiddetta "probabilità di blocco", ovvero la probabilità che un cliente, al momento del suo arrivo, trovi tutti i servitori occupati (questa affermazione trova la sua giustificazione nel Risultato 1).

2.4 Coda M / M / ∞

Si tratta del caso limite del sistema visto al paragrafo precedente. Serve a modellare una situazione in cui $t_q = t_s$ per ogni cliente, ovvero non vi è mai tempo di attesa, ovvero la capacità di servizio è illimitata.

Anche questo sistema può essere modellato come una CTMC-BD, in cui i rate di nascita e morte sono dati rispettivamente da

$$\lambda_n = \lambda \qquad n = 0,1,2,\dots$$
$$\mu_n = n\mu \qquad n = 1,2,\dots$$

In questo caso è evidente che, qualunque sia il valore di λ/μ, il sistema è sempre stabile, ovvero gli stati sono tutti ricorrenti positivi, e quindi esiste una distribuzione di probabilità a regime dello stato. Valutando tale distribuzione, sempre sulla base delle (31)-(32) del capitolo sulle Catene di Markov, si ottiene

$$\pi_0 = \left[1 + \sum_{n=1}^{\infty} \frac{\lambda^n}{\mu(2\mu)(3\mu)\dots(n\mu)}\right]^{-1} = \left[1 + \sum_{n=1}^{\infty} \frac{\left(\lambda/\mu\right)^n}{n!}\right]^{-1} = \left[e^{\frac{\lambda}{\mu}}\right]^{-1} = e^{-\rho} \qquad (26)$$

avendo posto $\rho = \lambda/\mu$ (in cui però il coefficiente ρ non ha il significato fisico di coefficiente di carico del sistema)

$$\pi_n = e^{-\rho}\frac{\rho^n}{n!} \qquad n \geq 1 \qquad (27)$$

(ma, data la (26), la (27) vale anche per $n = 0$).

La distribuzione di probabilità a regime dello stato è quindi una distribuzione di Poisson. Si ha quindi $\bar{L} = \rho = \lambda/\mu$. Ovviamente, per la legge di Little, $\bar{t_q} = \bar{L}/\lambda = 1/\mu = \bar{t_q}$, come ovviamente deve risultare.

Il modello M / M / ∞ può servire a modellare situazioni in cui il servizio è sostanzialmente un "ritardo puro", e non c'è nessuna linea di attesa (ad esempio, il trasporto di oggetti su un nastro trasportatore).

2.5 Coda M / M / 1 / K

Si tratta di una coda con buffer limitato. Si suppone che i clienti che arrivano quando il buffer è pieno siano semplicemente perduti. Ciò può essere modellato semplicemente imponendo che il processo degli arrivi si "azzeri" quando il buffer è pieno, e si "riaccenda" non appena nel buffer si libera qualche posizione.

Anche questo sistema può essere modellato come una CTMC-BD, in cui i rate di nascita e morte sono rispettivamente

$$\lambda_n = \begin{cases} \lambda & 0 \le n \le K \\ 0 & n = K \end{cases}$$

$$\mu_n = \mu \qquad n = 1,2,\dots,K$$

(*K* è il numero massimo di clienti presenti nel sistema).

Si noti che l'insieme dei valori possibili dello stato è, in questo caso, finito. Si noti anche che è possibile "riaccendere" semplicemente il processo degli arrivi grazie alla "proprietà memoryless" della distribuzione esponenziale del processo (markoviano) degli arrivi complessivo (comprendente cioè gli arrivi dei clienti accettati e quelli dei clienti rigettati); in altre parole, quando si riaccende il processo degli arrivi, qualunque sia il tempo trascorso dall'ultimo arrivo (accettato o rigettato), il tempo residuo che precede il prossimo arrivo ha sempre la medesima distribuzione esponenziale.

La condizione che assicura la positività ricorrente di tutti gli stati è qui banalmente soddisfatta in ogni caso (il sistema è ovviamente sempre "stabile"). Esiste quindi la distribuzione di probabilità regime dello stato, che può essere determinata nel modo consueto, cioè attraverso le (31)-(32) del capitolo sulle Catene di Markov.. In particolare, si ha

$$\pi_0 = \left[1 + \sum_{n=1}^{K} \left(\frac{\lambda}{\mu}\right)^n\right]^{-1} = \left[1 + \frac{\left(\frac{\lambda}{\mu}\right)\left(1 - \left(\frac{\lambda}{\mu}\right)^K\right)}{1 - \frac{\lambda}{\mu}}\right]^{-1}$$

Introducendo anche qui $\rho = \lambda/\mu$, si ha allora, dopo qualche passaggio,

$$\pi_0 = \frac{1-\rho}{1-\rho^{K+1}} \tag{28}$$

Si noti che qui ρ può essere ≥ 1, pur essendo il sistema stabile. In altre parole, anche in questo caso il coefficiente ρ non è una misura del grado di sollecitazione del sistema.

Inoltre, (conglobando la (28))

$$\pi_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n \qquad\qquad 0 \le n \le K \tag{29}$$

La lunghezza di coda media risulta in questo caso

$$\overline{L} = \sum_{n=0}^{K} n\pi_n = \frac{1-\rho}{1-\rho^{K+1}} \sum_{n=0}^{K} n\rho^n \tag{30}$$

Si osservi che

$$\sum_{n=0}^{K} n\rho^n - \rho\left(\sum_{n=0}^{K} n\rho^n\right) = (\rho + \rho^2 + \cdots + \rho^K) - K\rho^{K+1} = \frac{\rho(1-\rho^K)}{1-\rho} - K\rho^{K+1}$$

e quindi

$$\sum_{n=0}^{K} n\rho^n = \frac{\rho(1-\rho^K)}{(1-\rho)^2} - K\frac{\rho^{K+1}}{1-\rho}$$

per cui, sostituendo nella (30), si ha

$$\overline{L} = \frac{\rho}{1-\rho^{K+1}}\left[\frac{1-\rho^K}{1-\rho} - K\rho^K\right] \tag{31}$$

Per determinare la quantità $\bar{t}_q$ si può utilizzare la legge di Little. Essa però deve essere utilizzata qui riferendosi al flusso efficace in ingresso, cioè al flusso effettivamente entrante nel sistema. Il flusso efficace in ingresso può essere determinato come

$$\lambda_{eff} = \lambda\, Pr\{L < K\}$$

(anche questa relazione è giustificata dal Risultato 1). Pertanto, poiché

$$Pr\{L = K\} = \frac{1-\rho}{1-\rho^{K+1}}\rho^K \tag{32}$$

risulta

$$\lambda_{eff} = \lambda\left[1 - \frac{1-\rho}{1-\rho^{K+1}}\rho^K\right] = \lambda\frac{1-\rho^K}{1-\rho^{K+1}} \tag{33}$$

Ovviamente il throughput del sistema è uguale a λ_{eff}. L'utilizzazione del server è data da

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

Ovviamente il throughput può essere anche determinato anche come $\mu(1 - \pi_0)$, che darà ancora luogo a un risultato identico alla (33).

Dalla legge di Little, $\bar{t}_q$ può essere allora determinato come

$$\bar{t}_q = \frac{\overline{L}}{\lambda_{eff}} = \frac{\frac{\rho}{1-\rho^{K+1}}\left[\frac{1-\rho^K}{1-\rho} - K\rho^K\right]}{\lambda\frac{1-\rho^K}{1-\rho^{K+1}}} = \frac{1/\mu}{1-\rho^K}\left[\frac{1-\rho^K}{1-\rho} - K\rho^K\right] \tag{34}$$

Possono quindi essere determinati in maniera analoga $\bar{t}_w$ e $\bar{L}_w$.

2.6 Il Teorema di Burke

Si può dimostrare che in condizioni di equilibrio stocastico il processo di uscita di una coda M/M/1, o di una coda M/M/m, o di una coda M/M/∞, è ancora un processo di Poisson, con parametro pari al parametro del processo di ingresso. Questo risultato è fondamentale per quanto riguarda lo studio delle reti di code. Se infatti il processo di uscita di una coda M/M/, ad esempio, non potesse essere modellato come un processo di Poisson, due code in cascata NON potrebbero essere rappresentate come due code M/M/1, anche se l'ingresso alla prima coda fosse un processo di Poisson.

E' molto importante sottolineare il fatto che questo risultato NON vale per la coda M/M/1/K.

3 *Reti di code markoviane*

Questo paragrafo è dedicato alla presentazione dei risultati analitici che sussistono per due particolari modelli di reti di code: il modello di rete aperta (di Jackson) e il modello di rete chiusa (di Gordon-Newell). Entrambi i modelli sono caratterizzati dal fatto che i servitori delle macchine hanno tempi di servizio con distribuzione esponenziale. Inoltre, per il modello di rete aperta, anche i tempi di interarrivo dei clienti dall'esterno sono supposti caratterizzati da distribuzioni esponenziali. Tali assunzioni rendono poco realistica l'applicazione di questi modelli, e dei risultati ad essi relativi, in diversi ambiti applicativi come, ad esempio, quello dei processi manifatturieri. Tuttavia si può affermare che spesso l'impiego dei modelli in questione, anche se essi non risultano pienamente congruenti con le caratteristiche di un particolare sistema reale, costituisce un primo passo (computazionalmente assai efficiente, rispetto ad un approccio puramente simulativo) per il dimensionamento di massima di un sistema reale. In ogni caso, una volta determinato, dopo una serie di tentativi, tale dimensionamento di massima, deve comunque essere effettuata una verifica di tipo simulativo delle prestazioni ottenibili dal sistema reale, tenendo presenti le sue effettive caratteristiche.

3.1 Reti di code markoviane aperte

Il modello (di Jackson) che viene considerato in tal caso è definito nel modo seguente:

1. i clienti appartengono tutti alla medesima classe;
2. la rete di code è composta da N nodi ciascuno corrispondente ad una singola coda con un insieme di servitori identici in parallelo; m_i è il numero di servitori del nodo (macchina) i-esimo;
3. il tempo di servizio di ciascuno dei servitori del nodo i-esimo è una variabile aleatoria distribuita in modo esponenziale (con parametro μ_i);
4. in un certo numero di nodi della rete ha luogo il processo di arrivo dei clienti dall'esterno; ciascuno di tali processi è un processo di Poisson; il processo degli arrivi al nodo i-esimo è caratterizzato dal parametro λ_i;
5. dopo aver completato il servizio presso il nodo i-esimo, ciascun cliente può:
    - essere trasferito ad un altro nodo j (eventualmente $j = i$), con tempo di trasferimento nullo, con probabilità $r_{i,j}$;
    - uscire dal sistema, con probabilità $r_{i,0}$;

   è così definito il processo di instradamento; naturalmente risulterà

$$\sum_{j=1}^{N} r_{ij} + r_{i0} = 1 \qquad i = 1, \dots, N \tag{35}$$

6. i buffer delle linee di attesa ad ogni nodo hanno dimensione infinita;
7. tutti i processi stocastici (di arrivo, di servizio, di instradamento) corrispondono a sequenze di variabili aleatorie indipendenti e identicamente distribuite (i.i.d); i processi stocastici suddetti sono inoltre a due a due mutualmente indipendenti;
8. la popolazione complessiva dei clienti è infinita.

Avendo così definito il modello di rete di code è ovvio che il processo complessivo degli arrivi al nodo *i*-esimo (comprendendo quindi gli arrivi provenienti dall'esterno e quelli provenienti dall'interno) sia un processo stocastico caratterizzato da un rate di arrivi dato da

$$\Lambda_i = \lambda_i + \sum_{j=1}^{N} r_{ji} \widetilde{\Lambda}_j \tag{36}$$

essendo $\widetilde{\Lambda}_j$ il rate di uscite dal nodo *j*. In condizioni di equilibrio stocastico, il rate di arrivi in ingresso e quello in uscita saranno identici per ogni nodo. Si avrà in altre parole $\widetilde{\Lambda}_j = \Lambda_i$, $i = 1,\dots,N$. La (36) si potrà quindi scrivere per ogni *i*, determinando pertanto il sistema lineare di *N* equazioni in *N* incognite

$$\Lambda_i = \lambda_i + \sum_{j=1}^{N} r_{ji} \Lambda_j \qquad i=1,\dots N \tag{37}$$

che può essere risolto, allo scopo di ottenere i rate di arrivi (complessivi) per ogni nodo. Si noti che nella (37), per tutti i nodi in cui non vi sono arrivi dall'esterno, sarà semplicemente $\lambda_i = 0$.

In riferimento al modello di rete di code aperta sopra considerato, sussiste il seguente risultato, qui riportato senza dimostrazione.

Teorema 1 (di Jackson) *Si data una rete di code aperta corrispondente al modello precedentemente descritto. Si risolva allora il sistema* (37)*, determinando le quantità* Λ_i*,* $i = 1,\dots,N$*. Se risulta*

$$\frac{\Lambda_i}{m_i \mu_i} < 1 \qquad \forall i \tag{38}$$

allora il sistema può essere rappresentato con una CTMC irriducibile con tutti gli stati ricorrenti positivi. La distribuzione di probabilità a regime dello stato è esprimibile in forma prodotto come

$$\pi(n_1, n_2, \dots, n_N\ ) = \pi_1(n_1)\ \pi_2(n_2) \dots \pi_N(n_N) \tag{39}$$

essendo

$$\pi(n_1, n_2, \dots, n_N\ ) = Pr\{L_1 = n_1, \dots, L_N = n_N\} \tag{40}$$

(avendo chiamato L_i il numero di clienti presenti nel nodo i-esimo), ed essendo

$$\pi_i(n_i) = Pr\{L_i = n_i\}$$

Inoltre ciascuna delle distribuzioni $\pi_i(n_i)$ è identica alla distribuzione di probabilità dello stato a regime di una coda M / M / m_i, caratterizzata dai parametri Λ_i, μ_i, m_i. In altre parole

$$\pi_i(n_i) = \begin{cases} \pi_i(0)\dfrac{(m_i\rho_i)^{n_i}}{n_i!} & n_i = 1,2, \dots, m_i - 1 \\ \pi_i(0)\dfrac{{m_i}^{m_i}}{m_i!}{\rho_i}^{n_i} & n_i = m_i, m_i + 1, \dots \end{cases} \tag{41}$$

$$\pi_i(0) = \left[1 + \sum_{n_i=1}^{m_i-1} \frac{(m_i\rho_i)^n}{n_i!} + \frac{(m_i\rho_i)^{m_i}}{m_i!}\frac{1}{1-\rho_i}\right]^{-1} \tag{42}$$

dove $\rho_i = \Lambda_i/m_i\mu_i$.

Il significato del Teorema 1 è il seguente: in una rete di code aperta corrispondente al modello di Jackson, purché sia soddisfatta la condizione di stabilità (38), il sistema raggiunge una condizione di equilibrio stocastico in cui la distribuzione di probabilità congiunta è caratterizzata da una "struttura prodotto", cioè da una struttura tale da risultare prodotto di funzioni di una singola variabile aleatoria (associata ad un singolo nodo). Inoltre, tali funzioni corrispondono alle distribuzioni di probabilità marginali. Infne, ciascuna di tali distribuzioni di probabilità marginali corrisponde semplicemente alla distribuzione di probabilità a regime di una coda M / M / m_i.

Naturalmente la possibilità di determinare la probabilità congiunta consente la determinazione di diversi indici di prestazione caratterizzanti il comportamento della rete.

3.2 Reti di code markoviane chiuse

3. 2.1 *Il modello di Gordon e Newell e i risultati fondamentali*

Il modello a cui si fa riferimento in questo caso è identico a quello considerato al paragrafo 3.1, se non per le ipotesi seguenti:

1. $\lambda_i = 0 \ \forall i$, cioè non vi è alcun processo di arrivi di clienti dall'esterno; gli arrivi provengono tutti dai nodi della rete; inoltre, $r_{i,0} = 0 \ \forall i$, cioè i clienti non possono lasciare il sistema una volta completato il servizio su un certo nodo della rete; dal punto di vista pratico ciò vuol dire che vi è un numero costante di clienti nel sistema, e che quando un cliente completa il suo ciclo complessivo di servizi, esso viene sostituito da un altro cliente che invece deve ancora iniziare il suo ciclo;
2. la matrice di routing, il cui generico elemento è $r_{i,j}$, non può essere trasformata, attraverso semplici permutazioni di riga e di colonna, in una matrice avente la struttura

$$\begin{bmatrix} A & \emptyset \\ B & C \end{bmatrix}$$

essendo A e C sottomatrici quadrate.

Dal punto di vista pratico, la seconda ipotesi specifica che non è possibile isolare una parte del sistema in cui i clienti possono solo entrare, senza potere più uscire.

Sussiste allora il seguente risultato, fornito senza dimostrazione.

Teorema 2 (di Gordon e Newell) *In riferimento al modello visto al paragrafo 3.1, con le due ipotesi aggiuntive sopra riportate, si può dimostrare che il sistema può essere rappresentato come una CTMC irriducibile con uno spazio degli stati S finito. La cardinalità di tale spazio è data da*

$$|S| = \binom{N+K-1}{K} = \binom{N+K-1}{N-1} \tag{43}$$

dove K è il numero costante di clienti nella rete di code chiusa. Esiste quindi una distribuzione di probabilità a regime dello stato, ed essa è data da

$$\pi(n_1, n_2, \dots, n_N\ ) = C \prod_{i=1}^{N} \frac{{x_i}^{n_i}}{\beta_i(n_i)} \tag{44}$$

essendo le funzioni $\beta_i(n_i)$ *fornite da*

$$\beta_i(n_i) = \begin{cases} n_i! & n_i < m_i \\ m_i!\, {m_i}^{n_i - m_i} & n_i \geq m_i \end{cases} \tag{45}$$

e le costanti x_i *determinate risolvendo (a meno di una costante arbitraria) il sistema lineare omogeneo*

$$\mu_i x_i = \sum_{j=1}^{N} r_{ji}\, \mu_j x_j \qquad i=1,\dots,N \tag{46}$$

La costante C è una costante di normalizzazione che può essere determinata come

$$C = \frac{1}{\sum_{\underline{n} \epsilon S} \left(\prod_{i=1}^{N} \frac{{x_i}^{n_i}}{\beta_i(n_i)} \right)} \tag{47}$$

dove $\underline{n} = col(n_1, n_2, \dots, n_N\ )$. Si noti che la costante di normalizzazione ha la funzione di imporre che risulti

$\sum_{\underline{n} \epsilon S} \pi(n_1, n_2, \dots, n_N\ ) = 1$, indipendentemente dalla indeterminazione che caratterizza la risoluzione della (46). La determinazione di *C*, come prescritto dalla (47), richiede purtroppo la determinazione completa dello spazio degli stati *S*;, la cui cardinalità può essere piuttosto elevata, anche per *N* e *K* non necessariamente molto grandi.

Si noti ancora che la (44) esprime ancora una distribuzione di probabilità congiunta avente la forma prodotto. A differenza però di quanto vale per la (39), i fattori nella (44) non sono distribuzioni di probabilità marginali, ovvero le variabili aleatorie $n_1, n_2, \dots, n_N$ non possono più qui essere considerate come variabili aleatorie indipendenti (nella condizione di equilibrio stocastico). Tale considerazione risulta peraltro del tutto ovvia se teniamo presente che stiamo trattando il caso di una rete di code con un numero fisso di clienti.

La probabilità marginale $\pi_i(n_i)$ può essere calcolata dalla probabilità congiunta. Ad esempio, in una rete di code chiusa con 3 macchine e 4 clienti, risulta $\pi_1(2) = \pi(2,2,0) + \pi(2,1,1) + \pi(2,0,2)$. In generale

$$\pi_i(s) = \sum_{\underline{n} \in S_s^i} \pi(\underline{n})$$

essendo

$$S_s^i = \left\{ (n_1, n_2, \dots, n_N) : \sum_{l=1}^{N} n_l = K, \;\; n_i = s \right\}$$

La lunghezza di coda media del nodo *i*-esimo è data ovviamente da

$$\overline{L_i} = \sum_{s=0}^{K} s\pi_i(s)$$

Il coefficiente di carico relativo al nodo *i*-esimo è dato da

$$\rho_i = \frac{\Lambda_i}{m_i \mu_i} \tag{48}$$

ma esso corrisponde anche a $(1 - \tilde{\pi}_i(0))$, dove indichiamo con $\tilde{\pi}_i(0)$ la probabilità che un generico servitore nel nodo *i*-esimo sia inattivo in un istante selezionato a caso, in condizioni di equilibrio stocastico.

La quantità $\tilde{\pi}_i(0)$ può sempre essere ricavata dalla distribuzione marginale $\pi_i(n_i)$. Infatti, nel caso in cui sia m_i sia uguale a 1 si ha semplicemente $\tilde{\pi}_i(0) = \pi_i(0)$, nel caso in cui m_i sia uguale a 2 si ha $\tilde{\pi}_i(0) = \pi_i(0) + 0.5 \cdot \pi_i(1)$, e così via (si ricordi che si suppone che nessuno dei servitori in un nodo sia privilegiato (o penalizzato) dalla politica di assegnazione dei clienti in arrivo).

Una volta nota $\tilde{\pi}_i(0)$, si può trovare $\rho_i = 1 - \tilde{\pi}_i(0)$, e quindi, dalla (48), Λ_i. Dal punto di vista pratico, si noti che è sufficiente determinare una sola delle quantità Λ_i, $i = 1,\dots,N$, per determinarle tutte, sulla base della risoluzione del sistema lineare omogeneo

$$\Lambda_i = \sum_{j=1}^{N} r_{ji} \Lambda_j \qquad\qquad i = 1, \dots, N$$

che è risolubile a meno di una costante arbitraria.

Una volta determinato l'insieme delle Λ_i, $i = 1,\dots,N$, è immediato determinare il throughput dell'intero sistema, ovvero il flusso in uscita dal sistema dei clienti per cui "è stato completato il ciclo dei servizi", che coincide col flusso in ingresso dei clienti che devono ancora iniziare tale ciclo.

Il thouguput (denotato ora con X) può infatti essere determinato come la somma dei flussi in uscita dal sistema, ovvero come una combinazione lineare (i cui coefficienti dipendono dalla topologia della rete) dei flussi Λ_i che attraversano i vari nodi.

Una volta noto X, dalla legge di Little applicata all'intero sistema avremo

$$K = X\ \overline{FT}$$

che permette di ricavare $\overline{FT}$, ovvero il *flow time medio* (cioè il tempo medio complessivo di permanenza nel sistema per il generico cliente).

3. 2.2 *Il metodo di Denning e Buzen*

Si consideri ora una rete di code markoviane chiusa (che soddisfa il modello di Gordon e Newell) in cui si abbia $m_i = 1\ \forall i$. Da questa assunzione e dalla (45) segue che $\beta_i(n_i) = 1\ \forall i$. La distribuzione congiunta di probabilità data dalla (44) si può allora scrivere come

$$\pi(n_1, n_2, \dots, n_N\ ) = \frac{1}{G(K,N)} \prod_{i=1}^{N} {x_i}^{n_i} \tag{49}$$

avendo espresso in modo differente la costante di normalizzazione. Ovviamente adesso risulterà

$$G(K,N) = \sum_{\underline{n}\epsilon S} \left(\prod_{i=1}^{N} {x_i}^{n_i} \right) \tag{50}$$

Il teorema seguente (di cui non viene riportata xla dimostrazione) fornisce un metodo computazionalmente molto efficiente per determinare la costante $G(K,N)$.

Teorema 3 (di Denning e Buzen) *La costante di normalizzazione $G(K,N)$ può essere determinata attraverso il procedimento ricorsivo (in due dimensioni)*

$$g(k,i) = g(k,i-1) + x_i\, g(k-1,i) \tag{51}$$

$k=1,\dots, K$, $i=1,\dots, N$, inizializzato con

$$g(0,i) = 1 \qquad i = 1, \dots, N$$

$$g(k,0) = 0 \qquad k = 1, \dots K$$

Risulta allora

$$g(K,N) = G(K,N)$$

E inoltre

$a)$ $\qquad Pr\{n_i \geq s\} = {x_i}^s \ \frac{g(K-s,N)}{g(K,N)}$

$b)$ $\qquad Pr\{n_i = s\} = \ Pr\{n_i \geq s\} - Pr\{n_i \geq s+1\} = \frac{{x_i}^s g(K-s,N) - {x_i}^{s+1} g(K-s-1,N)}{g(K,N)}$

e quindi

$$Pr\{n_i = s\} = \frac{{x_i}^s}{g(K,N)}[g(K-s,N) - x_i \ g(K-s-1,N)]$$

c) (sulla base della b) $\qquad \mathrm{E}[\mathrm{n_i}] = \overline{\mathrm{L_i}} = \sum_{\mathrm{k=1}}^{\mathrm{K}} {x_i}^k \ \frac{g(K-k,N)}{g(K,N)}$

d) (segue dalla a), scritta in particolare per s=1)

$$Pr\{n_i \geq 1\} = x_i \quad \frac{g(K-1,N)}{g(K,N)}$$

e) infine, poiché $Pr\{n_i \geq 1\}$*=probabilità che la macchina i-esima sia attiva, e poichè*

$\Lambda_i = Pr\{n_i \geq 1\}\mu_i$

si ha

$$\Lambda_i = x_i \ \frac{g(K-1,N)}{g(K,N)} \ \mu_i \qquad (52)$$

Il teorema precedente fornisce evidentemente un metodo computazionalmente assai efficiente (con complessità computazionale polinomiale) per la determinazione della costante di normalizzazione *G*(*K,N*) e quindi per l'espressione della distribuzione congiunta di probabilità (49). Si noti che non è più necessario determinare tutti gli stati del sistema per ottenere la probabilità di un singolo stato.

E' evidente peraltro che, una volta calcolato il flusso Λ_i su una singola macchina M_i, posso trovare, tramite le equazioni di bilanciamento dei flussi, il flusso su ogni macchina. Dal flusso Λ_i sulla macchina generica, si piò poi determinare il coefficiente di utilizzazione ρ_i come

$$\rho_i = \frac{\Lambda_i}{\mu_i}$$

3. 2.3 *Macchina bottleneck*

Si assuma adesso (senza eccessiva perdita di generalità) che la macchina M_N sia sia l'ultima macchina visitata dai clienti, prima dell'uscita dal sistema. In alternativa, si può immaginare che la macchina M_N sia la macchina di carico/scarico. Si suppone che non ci siano "cicli di ritorno" su M_N.

Il flusso sulla macchina M_N è quindi uguale al throughput del sistema

$$X = \Lambda_N$$

Si possono definire allora le costanti $V_i, i = 1, \dots N$ attraverso le relazioni

$$\Lambda_i = V_i \Lambda_N \qquad i = 1, \dots, N$$

per cui ovviamente risulta $V_N = 1$. Tali costanti possono essere univocamente detrminate risolvendo il sistema (non omogeneo)

$$\begin{cases} V_i = \sum_{j=1}^{N} V_j r_{ji} \\ V_N = 1 \end{cases} \tag{53}$$

I valori delle costanti V_i, tutti positivi, possono essere minori, maggiori o uguali a 1. Il significato del generico coefficiente V_i è

$$V_i = E[numero\ di\ passaggi\ di\ un\ cliente\ generico\ attraverso\ il\ nodo\ i]$$

Si può notare come notare come, indipendentemente dal valore di *K* e dalla determinazione del throughput V_N, possa essere determinata a priori, sulla base dei soli parametri μ_i delle macchine e dei coefficienti di routing r_{ji}, la macchina (o, eventualmente, le macchine) con il maggior coefficiente di utilizzazione.

Tale macchina prende il nome di "macchina bottleneck" (ovvero "collo di bottiglia") e determina un limite fondamentale alle prestazioni del sistema (indipendentemente da *K,* cioè dalla numerosità dei clienti presenti nel sistema).

Si considerino infatti (si ricordi che siamo sempre nell'ipotesi $m_i = 1\ \forall i$) le disuguaglianze

$$\frac{\Lambda_i}{\mu_i} < 1 \qquad\qquad i = 1, \dots, N \tag{54}$$

che ovviamente saranno sempre soddisfatte, indipendentemente dal valore di *K*, in quanto una rete di code chiusa è intrinsecamente un sistema "stabile".

Le (54) si possono riscrivere come

$$\frac{V_i \Lambda_N}{\mu_i} < 1 \qquad\qquad i = 1, \dots, N \qquad\qquad (55)$$

in cui evidentemente i parametri μ_i e V_i sono noti a priori o determinabili sulla base di parametri noti a priori come i coefficienti di routing. Dalle (55) si ha allora

$$\Lambda_N < \frac{\mu_i}{V_i} \qquad\qquad i = 1, \dots, N$$

e quindi

$$\Lambda_N < \min_{i = 1, \dots N} \left\{ \frac{\mu_i}{V_i} \right\} \qquad\qquad (56)$$

La (56) determina il limite fondamentale per il throughput del sistema. La macchina i^* per cui

$$\frac{\mu_i^*}{V_i^*} = \min_{i = 1, \dots N} \left\{ \frac{\mu_i}{V_i} \right\}$$

si chiama macchina bottleneck (potrebbero anche essercene anche più di una).

Se supponiamo che tutte le caratteristiche del sistema considerato siano fissate (N, coefficienti della matrice di routing, parametri μ_i delle macchine), a parte K, il throughput Λ_N diventa una funzione solo di K. Si può provare che tale funzione è monotona crescente. Il valore asintotico (per K tendente ad infinito) di Λ_N è proprio μ_i^*/V_i^*.

La conoscenza questo valore asintotico è essenziale per determinare, in via preliminare, se una rete di code chiusa, con una predefinita gestione dei flussi, può raggiungere determinati livelli produttivi.

Si noti che, all'aumentare di K, aumenta il work-in-process del sistema (con conseguente aumento dei costi). La scelta di K potrebbe quindi essere effettuata cercando di massimizzare la seguente funzione obiettivo

$$F_b = c_1 \Lambda_N(K) - c_2 K$$

sotto il vincolo

$$\frac{K}{\Lambda_N(K)} \leq flow\ time\ medio \leq FT_{MAX}$$

dove c_1 e c_2 sono ovviamente opportuni coefficienti di guadagno e di costo e FT_{MAX} rappresenta un upper bound sul valore del flow time medio.

4 *Metodi approssimati per l'analisi di reti di code chiuse*

4.1 Il metodo della mean-value-analysis (MVA) nel caso monoclasse

Come si è visto, esistono un certo numero di risultati, piuttosto interessanti, per l'analisi delle reti di code chiuse markoviane. Evidentemente, nel caso in cui cade l'ipotesi di markovianità, tali risultati non sono più applicabili. Sono stati sviluppati allora, per lo studio delle reti di code chiuse non markoviane, metodi di analisi approssimati, che permettono di effettuare un'analisi preliminare delle prestazioni del sistema, che ovviamente dovrà essere successivamente validata da esperimenti simulativi.

Uno dei più utilizzati fra tali metodi prende il nome di "mean-value-analysis". Tale metodo approssimato (che qui viene presentato nella versione proposta da Schweitzer e Bard) si basa sulla stima (approssimata) del tempo medio di attesa del cliente generico al nodo *i*-esimo. Nell'ipotesi che tutti i nodi siano mono-server, tale tempo di attesa viene stimato come pari a

$$\bar{t}_i^w = \frac{K-1}{K}\bar{q}_i\bar{t}_i^s \tag{57}$$

essendo

- $\bar{t}_i^w$ il tempo medio di attesa al nodo *i*-esimo;
- $\bar{t}_i^s$ il tempo medio di servizio al nodo *i*-esimo;
- $\bar{q}_i$ la coda media al nodo *i*-esimo;
- K il numero complessivo (costante) di clienti nel sistema.

La scelta della stima data dalla (57) è motivata dall'assunzione che l'introduzione del termine $(K-1)/K$ sia appropriato per "ridurre" la lunghezza di coda media $\bar{q}_i$ di un fattore che tiene conto dell'attesa del servitore in una coda (FIFO) di cui anch'esso fa parte. Naturalmente, si tratta di un'approssimazione, totalmente euristica, che non può essere in alcun modo giustificata dal punto di vista analitico.

Se si accetta questa approssimazione diventa possibile stimare il tempo medio complessivo di permanenza nel nodo *i*-esimo (comprensivo cioè del tempo di attesa e del tempo di servizio) come

$$\bar{t}_i^q = \bar{t}_i^w + \bar{t}_i^s = \bar{t}_i^s\left[1 + \frac{K-1}{K}\bar{q}_i\right] \qquad i = 1, \dots, N \tag{58}$$

Supponiamo anche qui, per semplicità che il flusso (throughput) in uscita dal sistema corrisponda a Λ_N. Se allora V_i denota, come al solito, il "numero medio di passaggi" di un generico cliente per il nodo i-esimo (i valori di V_i, $i = 1,..,N$, possono essere determinati a priori come visto prima, cioè attraverso la risoluzione del sistema (53), il tempo complessivo di permanenza nel sistema $\overline{t}_{tot}^{q}$ può essere determinato come

$$\overline{t}_{tot}^{q} = \sum_{i=1}^{N} \overline{t}_i^{q} \; V_i \tag{59}$$

per cui, usando la legge di Little, il throughput del sistema può essere calcolato come

$$X = \frac{K}{\overline{t}_{tot}^{q}} = \frac{K}{\sum_{i=1}^{N} \overline{t}_i^{q}} \tag{60}$$

Noto X possono essere determinati i flussi Λ_i, $i = 1,..,N$, tramite la relazione $\Lambda_i = V_i \; X$, $i = 1,..,N$. Quindi, sempre applicando la legge di Little, si possono ottenere le code medie $\overline{q}_i$ come

$$\overline{q}_i = \Lambda_i \; \overline{t}_i^{q} \qquad\qquad i = 1, \dots, N \tag{61}$$

Sulla base di tali considerazioni, può essere strutturato un algoritmo iterativo per l'analisi (approssimata) delle prestazioni di una rete di code chiusa.

Algoritmo 1 (mean-value-analysis secondo Schweitzer-Bard)

1. *Si inizializza della stima delle code* $\overline{q}_i$, $i = 1,..,N$
2. *Si calcolano i tempi medi* $\overline{t}_i^{q}$, $i = 1,...,N$, *sulla base delle* $\overline{q}_i$, *per mezzo della* (58)
3. *Si calcola il throughput X per mezzo della* (60)
4. *Si calcolano nuove stime delle code* $\overline{q}_i^{\,new} = X \; V_i \overline{t}_i^{q}$, $i = 1,..,N$
5. *Se* $|\overline{q}_i^{\,new} - \overline{q}_i| < \varepsilon$, $\forall i = 1,..,N$, *stop; altrimenti si pone* $\overline{q}_i \leftarrow \overline{q}_i^{\,new}$, $i = 1,...,N$, *e si ritorna al passo 2*

Naturalmente non vi è alcuna garanzia a priori sulla convergenza dell'algoritmo, né vi è sul livello di approssimazione della stima degli indici di prestazione ottenuta nel caso di convergenza. Tuttavia, in letteratura sono riportate numerose applicazioni dell'algoritmo nelle quali la convergenza si raggiunge abbastanza rapidamente. Inoltre, il livello di approssimazione delle stime ottenute è generalmente buono, rispetto alle stime accurate ottenibili mediante (onerosi) esperimenti simulativi.

4.2 Estensione della mean-value-analysis (MVA) al caso multiclasse

E' possibile estendere l'applicazione del metodo approssimato della mean-value-analysis anche al caso di reti di code chiuse multiclasse (ovvero con diverse classi di clienti).

Siano allora P le classi di clienti nel sistema. Allora, per ogni $i = 1,..,N$ e per ogni $p = 1,..,P$, indicheremo con

- $\overline{q}_{i,p}$ il numero medio di clienti della classe p in coda al nodo i;
- $\overline{t}_{i,p}{}^{w}$ il tempo medio di attesa dei clienti della classe p al nodo i;
- $\overline{t}_{i,p}{}^{q}$ il tempo medio di permanenza dei clienti della classe p al nodo i;
- $\overline{t}_{i,p}{}^{s}$ il tempo medio di servizio per i clienti della classe p nel nodo i;
- $\overline{t}_{tot,p}$ il tempo medio complessivo di permanenza nel sistema per i clienti della classe p.

Naturalmente, in una rete di code chiusa multiclasse, sarà costante il numero complessivo di clienti nel sistema, ma saranno costanti anche i numeri di clienti delle diverse classi, singolarmente considerati. Sia $K(p)$ il numero costante di clienti della classe p, $p = 1,...,P$.

La stima del tempo medio $\overline{t}_{i,p}{}^{w}$ può essere ottenuta allora, in analogia con la (65), come

$$\overline{t}_{i,p}{}^{w} = \frac{K(p)-1}{K(p)}\overline{q}_{i,p}\overline{t}_{i,p}{}^{s} + \sum_{\substack{r=1 \\ r\neq p}}^{P} \overline{q}_{i,r}\,\overline{t}_{i,r}{}^{s} \qquad (62)$$

$i = 1,..,N, \qquad p = 1,..,P$

Naturalmente si otterrà

$$\overline{t}_{i,p}{}^{q} = \overline{t}_{i,p}{}^{w} + \overline{t}_{i,p}{}^{s} \qquad i = 1,..,N, \qquad p = 1,..,P$$

Inoltre avremo

$$\overline{t}_{tot,p} = \sum_{i=1}^{N} \overline{t}_{i,p}{}^{q}\, V_{i,p} \qquad i = 1,..,N, \qquad p = 1,..,P$$

dove $V_{i,p}$ indica il numero medio di passaggi per il nodo i, per i clienti della classe p, $i = 1,..,N$, $p = 1,..,P$. Naturalmente, le $V_{i,p}$ possono essere determinate attraverso la risoluzione del sistema (non omogeneo)

$$\begin{cases} V_{i,p} = \sum_{j=0}^{N} V_{j,p} r_{ji}^{p} & i = 1, \dots N \\ V_{N,p} = 1 \end{cases}$$

per ogni $p = 1,..,P$. Si noti che in una rete multiclasse anche i coefficienti di routing r_{ij}^{p} dipendono dalla classe p.

Noto $\overline{t}_{tot,p}$posso ricavare, tramite la legge di Little, il throughput X_p per i clienti della classe p

$$X_p = \frac{K(p)}{\overline{t}_{tot,p}}$$

per, p=1, ...,P, e quindi tutti i flussi ai singoli nodi

$$\Lambda_{i,p} = \mathrm{V}_{i,p} X_p \qquad i = 1,..,N, \qquad p = 1,..,P$$

sulla base dei quali, ancora utilizzando la legge di Little, possono essere stimate le code medie

$$\overline{q}_{i,p} = \Lambda_{i,p}\ \overline{t}_{i,p}^{q} \qquad i = 1, \dots, N \qquad p = 1, \dots, P$$

che possono essere confrontate con le stime iniziali. E' evidentemente possibile allora strutturare un algoritmo analogo a quello visto nel caso di reti monoclasse.

RIFERIMENTI BIBLIOGRAFICI

C. G. Cassandras, S. Lafortune, *Introduction to Discrete Event Systems*, 2nd ed., Springer, 2008

G. Bolche, S. Greiner, H. de Meer, K.S. Trivedi, *Queueing Networks and Markov Chains*, 2nd ed, J. Wiley, 2006.

J.P. Buzen, Computational Algorithms for Closed Queueing Networks with Exponential servers, Communications of the ACM, Vol 16, Nr. 9, September 1973, pp. 527-531.