La teoria delle code

Lezione 05

24 marzo 2015

Ing. Chiara Foglietta

foglietta.chiara@gmail.com

Sistemi di Controllo per l'Automazione Industriale
Ingegneria Gestionale
A.A. 2014 - 2015
Università degli Studi di Cassino e del Lazio Meridionale

Agenda

Introduzione

Modelli a code: specifiche

Prestazioni di un sistema a coda

Dinamiche di un sistema a coda

Legge di Little

Sistemi semplici a coda

Lo scopo principale è quello di presentare le idee e le tecniche principali che sono usate per analizzare semplici sistemi a coda. Per "semplici" sono associati a modelli di catene di nascita e morte. Molti dei sistemi più interessanti si possono ricondurre come casi speciali a questi modelli. Vengono anche forniti alcune estensioni necessarie per trattare situazioni più complesse che coinvolgono processi ad eventi che non soddisfano la proprietà di Markov.

La teoria delle code ha come scopo principale la determinazione delle prestazioni del sistema in particolari condizioni operative, piuttosto che la determinazione delle politiche operative per ottenere la migliore performance possibile. Per cui, lo scopo è stato quello di sviluppare strumenti descrittivi per studiare i sistemi a coda, piuttosto che strumenti prescrittivi per controllare i comportamento in un ambiente dinamico e incerto.

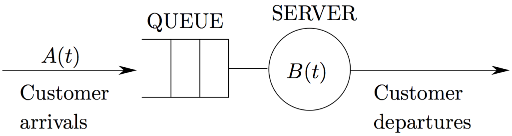
Specifiche dei modelli a code

Si comincia con le definizioni e le notazioni basilari necessarie per specificare un modello a code. Ci sono normalmente tre componenti:

1. Specifiche sul *modello statistico* per i processi di arrivo e di servizio
2. Specifiche sui *parametri strutturali* del sistema. Per esempio, la capacità di immagazzinamento della coda, il numero di serventi e così via.
3. Specifiche sulle *politiche operative* usate. Per esempio, le condizioni per cui gli utenti in arrivo vengono accettati, la priorità di alcuni utenti verso i serventi, ecc

Specifiche dei modelli a code

Si mostra un sistema a coda con infinito spazio di immagazzinamento per la coda e un singolo servente. Questo è un sistema a eventi discreti con insieme di eventi $E = \{a, d\}$, dove a è l'arrivo dell'utente e d è la partenza dopo il completamento di un servente.



Modelli stocastici per i processi di arrivo e di servizio

Nei sistemi a coda, si associano agli eventi di arrivo *a* una sequenza stocastica $\{Y_1, Y_2, \dots\}$ dove Y_k è il k-esimo tempo di inter-arrivo che è una variabile random tale che è il tempo passato tra l'arrivo $k-1$ e l'arrivo k, con $k = 1, 2, \dots$. Per semplicità, si pone che $Y_0 = 0$ e quindi Y_1 è la variabile random che descrive il tempo del primo arrivo.

In generale, il comportamento stocastico di questa sequenza richiede la specifica della probabilità congiunta di tutti gli eventi $[Y_k \leq t]$, $k = 1, 2, 1, \dots$. Nella maggior parte dei casi, si assume che la sequenza stocastica $\{Y_k\}$ è indipendente e identicamente distribuita. Quindi, una sola distribuzione di probabilità descrive completamente la sequenza dei tempi di inter-arrivo.

$$A(t) = P[Y \leq t]$$

In questo caso Y è spesso pensata come un generico tempo di inter-arrivo per cui non è necessario aggiungere l'indice k. Il valore medio della funzione di distribuzione $A(t)$, ossia $\mathbb{E}[Y]$, è particolarmente importante e di solito si indica come

$$\mathbb{E}[Y] \equiv \frac{1}{\lambda}$$

Per cui λ è la *frequenza media di arrivi* degli utenti.

Modelli stocastici per i processi di arrivo e di servizio

Similarmente, si associa all'evento *d* una sequenza stocastica $\{Z_1, Z_2, \dots\}$ dove Z_k è il k-esimo tempo di servizio , ossia, una variabile random tale che Z_k è il tempo richiesto per il k-esimo utente da essere servito, $k = 1, 2, \dots$.

Se si assume che la sequenza stocastica $\{Z_k\}$ è indipendente e identicamente distribuita, allora si definisce

$$B(t) = P\left[Z \leq t\right]$$

dove Z è il tempo di servizio generico. Quindi si può usare la seguente definizione per il valore medio di $B(t)$:

$$\mathbb{E}\left[Z\right] \equiv \frac{1}{\mu}$$

dove μ è la *frequenza media di servizio* nel nostro sistema.

Parametri strutturali

I tipici parametri strutturali di interessi in un sistema a coda sono:

1. La *capacità di immagazzinamento* della coda, di solito denotato da $K = 1, 2, \ldots$. Per convenzione, si include nella capacità anche lo spazio fornito per gli utenti in servizio
2. Il *numero di serventi*, di solito denotato con $m = 1, 2, \ldots$

Nel sistema presentato precedentemente si ha

1. $K = \infty$, ossia capacità di immagazzinamento infinito per gli utenti in coda
2. $m = 1$ ossia un singolo servente

Le politiche operative

Anche per un sistema semplice, ci sono diversi schemi che si possono adottare nel gestire il processo a coda.

- La distribuzione dei tempi di servizio è la stessa per tutti gli utenti in arrivo?
- Gli utenti sono differenziati sulla base di diverse classi di appartenenza, alcuni dei quali possono avere una maggiore priorità rispetto agli altri nella richiesta del servizio?
- Tutti gli utenti sono ammessi nel sistema?
- Agli utenti è consentito lasciare la coda prima di essere serviti, e, se si, sotto quali condizioni?
- Come il servente decide quale prossimo utente servire, se ce sono più di uno nella coda?
- Al servente è consentito fermare un utente in servizio per servire un utente a più alta priorità che è appena arrivato?

Le politiche operative I

Nel categorizzare le politiche operative, i problemi più comuni che devono essere considerati sono i seguenti:

1. Il numero di classi di utenti. Nel caso di sistema con una singola classe, tutti gli utenti hanno i medesimi requisiti di servizio e il servente tratta tutti nella medesima maniera. Questo significa che la distribuzione dei tempi di servizio è la medesima per tutti. Nel caso di un sistema con molteplici classi, gli utenti sono distinti in base ai requisiti di servizio e/o al modo in cui il servente li tratta.
2. Politiche di scheduling. In un sistema con molteplici classi, il servente deve decidere quale prossima classe di utente servire. Per esempio, il servente deve sempre dare priorità ad una particolare classe, o può fermare un utente durante il processo poichè è appena arrivato un utente a più alta priorità

Le politiche operative II

3. Discipline di coda. Una disciplina di coda descrive l'ordine con cui il servente seleziona gli utenti da processare, anche se c'è una sola classe. Per esempio, first come first served (FCFS), last come last served (LCLS), e ordine random
4. Politiche di ammissione. Anche se una coda ha capacità di immagazzinamento infinito, può essere desiderabile negare l'immissione ad alcuni degli utenti un arrivo. Nel caso di due classi di arrivi, per esempio, gli utenti a più alta priorità possono essere sempre ammessi, ma quelli con priorità più bassa possono essere ammessi solo se la coda è vuota o se è passato un po' di tempo da quando un utente è stato ammesso.

Le politiche operative

Nel nostro caso , si può assumere

- un sistema con una singola classe
- Tutti gli utenti che arrivano sono ammessi e serviti
- La disciplina di coda è first-com-first-served (FCFS)

La notazione $A/B/m/K$ I

Nella teoria delle code si impiega una particolare tipo di notazione per descrivere un sistema in maniera coincisa. La notazione è la seguente:

$$A/B/m/K$$

dove

- A è la distribuzione dei tempi di inter-arrivo
- B è la distribuzione dei tempi di servizio
- m è il numero di serventi presenti, $m = 1, 2, \ldots$
- K è la capacità di immagazzinamento della coda $K = 1, 2, \ldots$

La notazione $A/B/m/K$ II

Per cui il sistema a singolo servente e infinita capacità di immagazzinamento è descritto come $A/B/1/\infty$. Per semplificare la notazione, se K è omesso si intende come $K = \infty$. Quindi, nel nostro caso, si ha $A/B/1$.

La notazione $A/B/m/K$ III

C'è una notazione comune per rappresentare le distribuzioni A e B

- G sta per distribuzione *generale* quando null'altro è noto sul processo di arrivo/servizio
- GI sta per distribuzione *generale* in un processo di arrivo/servizio che è indipendente e identicamente distribuito
- D sta per caso *deterministico*, ossia i tempi di inter-arrivo/servizio sono fissati
- M sta per caso *markoviano*, ossia i tempi di inter-arrivo/servizio sono distribuiti esponenzialmente

La notazione $A/B/m/K$ IV

La distinzione tra G e GI non è sempre fatta esplicitamente. Di seguito alcuni esempi vengono riportati

$M/M/1$ Un sistema singolo servente con infinita capacità di immagazzinamento. I tempi di inter-arrivo e di servizio sono entrambi esponenzialmente distribuiti, ossia $A(t) = 1 - e^{-\lambda t}$, $B(t) = 1 - e^{-\mu t}$, per alcuni parametri positivi λ e μ

$M/M/1/K$ Un sistema singolo servente con capacità di immagazzinamento uguale a $K < \infty$, incluso lo spazio al servente. I tempo di inter-arrivo e di servizio sono entrambi esponenzialmente distribuiti

$M/G/2$ Un sistema con due serventi e infinita capacità di immagazzinamento. I tempi di inter-arrivo sono distribuiti esponenzialmente. I tempi di servizio hanno una distribuzione arbitraria generale.

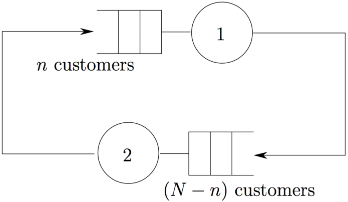
La notazione $A/B/m/K$ V

Si noti che questa notazione non specifica le politiche operative da usare. Non descrive il caso in cui due o più classi di utenti sono presenti, ognuna con differenti distribuzioni di tempi di inter-arrivo e di servizio. Quando si usa questa notazione, da sola, si tratta di un sistema a singola classe, nessun controllo di ammissione, e disciplina di coda FCFS.

Sistemi a coda aperti e chiusi I

Il sistema che abbiamo visto all'inizio è accessibile da qualsiasi utente dal mondo esterno che vuole essere servito. In altre parole, il sistema è aperto ad una popolazione infinita di utenti che possono richiedere il servizio in qualsiasi momento. Si riferisce a questo concetto come una *popolazione di utenti infiniti* o sistema a coda aperto.

Sistemi a coda aperti e chiusi II



In figura si mostra un sistema con due serventi, che è limitato ad una popolazione finita di utenti, di solito denotato con $N = 1, 2, \ldots$. In questo caso, un utente che completa il servizio al servente 1 ritorna sempre per un ulteriore servizio e non lascia mai il sistema. Si parla di *popolazione di utenti finita* o sistema a coda chiuso.

Sistemi a coda aperti e chiusi III

In alcuni casi, un singolo servente opera un sistema con una popolazione finita. Un tipico utente sperimenta un tempo di servizio con distribuzione B, poi va con un po' di ritardo la cui distribuzione è A prima di riunirsi alla coda. Per cui, A emula un processo di arrivo, ma il numero di utenti serviti rimane fissato ad N. Per rappresentare tali sistemi si estende la notazione $A/B/m/K$ con $A/B/m/K/N$ dove N rappresenta la popolazione di utenti. Come nel caso di capacità infinita, omettere N significa sistema aperto. Se $K = \infty$ ma $N < \infty$ allora si scrive $A/B/m//N$

Prestazioni di un sistema a coda I

Ci si concentri su un sistema a coda aperto a singolo servente. Sono già state definite due variabili di interesse random relativi all'utente k: il tempo di inter-arrivo Y_k e il tempo di servizio Z_k. Inoltre, si definiscono:

- A_k è il *tempo di arrivo* dell'utente k
- D_k è il *tempo di partenza* dell'utente k
- W_k è il *tempo di attesa* dell'utente k, dall'istante di arrivo fino all'inizio del servizio
- S_k è il *tempo di servizio* dell'utente k, dall'istante di arrivo fino alla partenza

Prestazioni di un sistema a coda II

Si noti che

$$S_k = D_k - A_k$$
$$S_k = W_k + Z_k$$

e quindi combinandoli si ottiene

$$D_k = A_k + W_k + Z_k$$

Inoltre, si definiscono le variabili random

- $X(t)$ è la lunghezza della coda all'istante $t, X(t) \in \{0, 1, 2, \dots\}$
- $U(t)$ è il carico di lavoro all'istante t, ossia la quantità di tempo richiesta per svuotare il sistema a t

Si noti che $X(t)$ è la quantità che normalmente si usa per descrivere lo stato di un sistema a coda temporale. Il carico di lavoro $U(t)$ si chiama anche lavoro non finito del sistema a t.

Prestazioni di un sistema a coda IV

Il comportamento stocastico della sequenza di tempo atteso $\{W_k\}$ fornisce un'importante informazione che riguarda le prestazioni del sistema. La funzione di distribuzione di probabilità di $\{W_k\}$, $P[W_k \leq t]$ generalmente dipende da k. Spesso si trova, per $k \to \infty$, che esiste una distribuzione stazionaria, $P[W \leq t]$, indipendente da k, tale che

$$\lim_{k\to\infty} P[W_k \leq t] = P[W \leq t]$$

25

Se questo limite esiste, la variabile W descrive il tempo di attesa di un tipico utente a regime. Intuitivamente, quando il sistema funziona per un periodo sufficientemente lungo, ogni nuovo utente ha un processo di attesa stocasticamente identico descritto da $P[W \leq t]$. Similarmente, se esiste una distribuzione stazionaria per la sequenza di tempo di sistema $\{S_k\}$ allora il suo valore medio $\mathbb{E}[S]$, è il tempo medio di sistema a regime.

Prestazioni di un sistema a coda VI

La medesima idea si può applicare ai processi stocastici $\{X(t)\}$ e $\{U(t)\}$. Se le distribuzioni stazionarie esistono per questi processi per $t \to \infty$, allora le variabili random X e U sono usate per descrivere la lunghezza della coda e il carico di lavoro del sistema a regime. Si usa la notazione $\pi_n, n = 0, 1, \ldots$ per indicare la probabilità stazionaria della lunghezza della coda, ossia

$$\pi_n = P\left[X = n\right], \qquad n = 0, 1, \ldots$$

Analogamente, $\mathbb{E}\left[X\right]$ è la lunghezza media della coda a regime, e $\mathbb{E}\left[U\right]$ è il carico di lavoro medio a regime.

Prestazioni di un sistema a coda VII

In generale, si vuole progettare un sistema a coda tale che un tipico utente a regime aspetta il minimo tempo possibile. Idealmente, vorremmo che ogni utente che arriva aspetta zero tempo. In un ambiente probabilistico questo è impossibile, poichè non possiamo fornire un numero infinito di serventi o un super-servente che opera in maniera infinitamente veloce. Per cui si considera il servente più veloce che ci possiamo permettere. Per giustificare il nostro investimento, si cerca di servire un numero massimo di utenti tenendo il servente più occupato possibile, cercando di massimizzare l'utilizzo del servente. Per fare ciò, si cerca di mantenere costantemente la coda non vuota; si cerca di avere sempre alcuni utenti da servire nel caso in cui i tempi di servizio siano brevi. Ma questo è contrario all'obiettivo di tenere il tempo di attesa basso per gli utenti.

Prestazioni di un sistema a coda VIII

Di solito si assume che esiste uno stato a regime che possa essere raggiunto e quindi ci si concentra sulle prestazioni di un sistema a coda a regime. Le misure delle prestazioni sono

- il tempo medio di attesa degli utenti a regime $\mathbb{E}[W]$
- il tempo medio di sistema degli utenti a regime $\mathbb{E}[S]$
- la lunghezza media della coda a regime $\mathbb{E}[X]$

che si vogliono mantenere più piccoli possibili

Prestazioni di un sistema a coda IX

Altre misure di prestazioni sono

- l'utilizzazione del sistema, ossia la frazione di tempo per cui il servente è occupato
- il volume di produzione (throughput) del sistema, ossia la frequenza per cui gli utenti lasciano il sistema dopo il servizio

che è desiderabile mantenere più alto possibile. Il volume di produzione non può essere superiore alla frequenza massima del servizio del servente. Se un sistema ha raggiunto il regime, gli utenti in ingresso e in uscita devono essere bilanciati e in questo caso il volume di produzione deve essere identico alla frequenza degli utenti in ingresso.

Si definisce *l'intensità del traffico* come segue

$$[\textit{traffic intensity}] \equiv \frac{[\textit{average arrival rate}]}{[\textit{average service rate}]}$$

Per un sistema a singolo servente, si usa la lettera ρ per l'intensità di traffico. Allora, per la definizione di λ e μ, si ha che

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

nel caso di *m* serventi, la frequenza di servizio medio diventa $m\mu$ e quindi

$$\rho = \frac{\lambda}{m\mu}$$

Prestazioni di un sistema a coda XI

L'utilizzazione di un servente è la frazione di tempo che è occupato. In un sistema a singolo servente a regime, la probabilità π_0 rappresenta la frazione di tempo che il sistema è vuoto e quindi il servente è in attesa. Per un servente a regime si ha che

$$[\textit{utilization}] \equiv [\textit{fraction of time server is busy}] \equiv 1 - \pi_0$$

Poichè un servente lavora alla frequenza μ e la frazione di tempo che è attualmente in funzione è $1 - \pi_0$, il volume di produzione di un sistema a singolo servente a regime è

$$\begin{aligned}[\textit{throughtput}] &\equiv [\textit{departure rate of customers after service}] \\ &= \mu(1 - \pi_0)\end{aligned}$$

Prestazioni di un sistema a coda XII

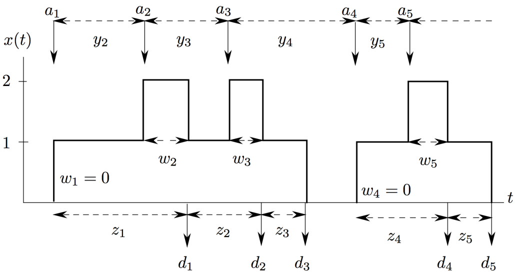
A regime, i flussi degli utenti in ingresso e in uscita dal sistema devono essere bilanciati, ossia

$$\lambda = \mu(1 - \pi_0)$$
$$\rho = (1 - \pi_0)$$

L'intensità del traffico, che è definita a partire dai parametri del servizio e le distribuzioni dei tempi di inter-arrivo, rappresenta anche l'utilizzazione del sistema. Questa relazione vale per qualsiasi sistema a singolo utente con capacità infinita. Se $\pi_0 = 1$, il sistema è permanentemente in attesa perchè non vi sono arrivi di utenti ($\lambda = 0$). Se $\pi_0 = 0$, il sistema è permanentemente occupato, generando instabilità, perchè la coda cresce infinitamente. I valori di ρ devono quindi essere $0 \leq \rho < 1$.

Dinamiche di un sistema a coda I

Si consideri il sistema a coda considerato all'inizio, che opera sulla base FCFS. Usando la notazione stabilita nel paragrafo precedente, un tipico cammino è mostrato nella figura.



Dinamiche di un sistema a coda II

Nell'esempio, quando arriva il primo utente all'istante a_1 trova una coda vuota. Durante l'intervallo che comincia a_1 e termina con d_3 il servente rimane occupato. Durante tale intervallo si parla di *periodo occupato* del sistema a coda. Durante l'intervallo che comincia con d_3 e termina con il prossimo arriva a_4, il servente rimane inattivo. Si parla di *periodo inattivo* del sistema. Un possibile modo di vedere questo sistema come una sequenza di cicli alternanti che consistono di periodi occupati e inattivi.

Dinamiche di un sistema a coda III

In generale quando il k-esimo utente arriva due casi sono possibili:

Caso 1. Il sistema è vuoto, e quindi $W_k = 0$. Il sistema può essere vuoto quando

$$D_{k-1} \leq A_k$$

ossia, il precedente utente parte prima che l'attuale utente arrivi. Quindi

$$D_{k-1} - A_k \leq 0 \Leftrightarrow W_k = 0$$

Questo è il caso per $w_4 = 0$, poichè $d_3 \leq a_4$

Dinamiche di un sistema a coda IV

Caso 2. Il sistema non è vuoto, quindi $W_k > 0$. In questo caso l'utente k-esimo è costretto ad aspettare fino a che il precedente $k-1$ utente esca. Quindi

$$D_{k-1} - A_k > 0 \Leftrightarrow W_k = D_{k-1} - A_k$$

Nella figura precedente questa situazione accade quando $w_2 = d_1 - a_2 > 0$

Dinamiche di un sistema a coda V

Combinando le due situazioni, si ottiene

$$W_k = \begin{cases} 0 & D_{k-1} - A_k \leq 0 \\ D_{k-1} - A_k & D_{k-1} - A_k > 0 \end{cases}$$

che può essere riscritta come

$$\begin{aligned} W_k &= \max\{0, D_{k-1} - A_k\} \\ W_k &= \max\{0, W_{k-1} + Z_{k-1} + A_{k-1} - A_k\} \\ W_k &= \max\{0, W_{k-1} + Z_{k-1} - Y_k\} \end{aligned}$$

L'equazione che lega i tempi di attesa W_k soddisfa una semplice equazione lineare. Per cui, le dinamiche dei tempi di attesa sono lineari eccetto per l'effetto di un utente con tempi di inter-arrivo tale che $Y_k > W_{k-1} + Z_{k-1} = S_{k-1}$. Tale equazione è detta equazione di Lindley.

Dinamiche di un sistema a coda VI

SI può anche scrivere tale equazione per i tempi di sistema, ossia

$$S_k = \max\{0, S_{k-1} - Y_k\} + Z_k$$

e per i tempi di partenza a partire da $W_k = D_k - A_k - Z_k$, si ottiene

$$D_k = \max\{A_k, D_{k-1}\} + Z_k$$

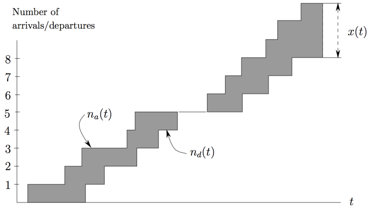
Dinamiche di un sistema a coda VII

Queste relazioni catturano le caratteristiche dinamiche essenziali di un sistema a coda. Queste relazioni sono molto generali, nel senso che possono essere applicati a qualsiasi sistema senza considerare le caratteristiche delle distribuzioni dei processi stocastici coinvolti.

La legge di Little I

Si consideri il sistema a coda con insieme di eventi $E = \{a, d\}$. Si consideri $N_a(t)$ e $N_d(t)$ che contano il numero di arrivi e di partenze, rispettivamente, nell'intervallo $(0, t]$. Si assuma che il sistema sia inizialmente vuoto, e che segue che la lunghezza della coda è

$$X(t) = N_a(t) - N_d(t)$$



La legge di Little II

Guardando l'area grigia, si noti che il rettangolo di altezza unitaria e lunghezza definita dal primo arrivo e dalla prima partenza sull'asse dei tempi rappresenta la quantità di tempo che il primo utente passa nel sistema. In generale, l'area grigia che consiste di tali rettangoli fino al tempo t rappresenta la quantità totale di tempo che tutti gli utenti hanno speso nel sistema fino a t. Quest'area è denotata come $u(t)$.

La legge di Little III

Dividendo quest'area per il numero totale di utenti che sono arrivati in $(0, t]$, $n_a(t)$, si ottiene il tempo medio di sistema per utente fino all'istante t, denotato da $\bar{s}(t)$

$$\bar{s}(t) = \frac{u(t)}{n_a(t)}$$

Dividendo l'area $u(t)$ per il tempo t, si ottiene il numero medio di utenti presenti nel sistema nell'intervallo $(0, t]$, ossia la lunghezza media della coda, $\bar{x}(t)$

$$\bar{x}(t) = \frac{u(t)}{t}$$

La legge di Little IV

Dividendo il numero totale di utenti che sono arrivati in $(0, t]$, $n_a(t)$ per t, si ottiene la frequenza di arrivi medi degli utenti, indicato da $\lambda(t)$

$$\lambda(t) = \frac{n_a(t)}{t}$$

Ottenendo $\bar{x}(t) = \lambda(t)\bar{s}(t)$

La legge di Little V

Si fanno anche due assunzioni fondamentali.

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

$$\lim_{t \to \infty} \bar{s}(t) = \bar{s}$$

Questi valori rappresentano la frequenza di arrivi a regime e il tempo di sistema. Se questi limiti esistono allora $\bar{x}(t)$ deve convergere ad un valore fissato $\bar{x}$. Quindi vale

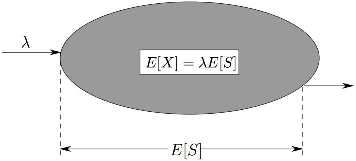
$$\bar{x} = \lambda \bar{s}$$

Assumendo che i limiti esistono per tutte le possibili traiettorie e con i medesimi valori a regime, allora $\bar{x}$ è la lunghezza media della coda $\mathbb{E}[X]$ e $\bar{s}$ è il tempo medio di sistema $\mathbb{E}[S]$ a regime. Quindi si ottiene

$$\mathbb{E}[X] = \lambda \mathbb{E}[S]$$

La legge di Little VI

Detta anche *legge di Little*. É un risultato importante in quanto è indipendente dalla struttura stocastica del sistema, ossia le distribuzioni di probabilità associate con gli eventi di arrivi e partenze. La legge di Little afferma che la lunghezza media della coda è proporzionale con il tempo medio del sistema, in cui la frequenza media degli arrivi λ è la costante di proporzionalità.



La legge di Little VII

È importante notare che tale legge è indipendente dalle politiche di funzionamento impiegate nel sistema a coda considerato. Inoltre, è valida per una configurazione arbitraria di code interconnesse e serventi. Si supponga di disegnare una linea intorno ad un parte o a tutto un sistema a coda, e sia λ la frequenza di arrivi totali di utenti che attraversano il confine verso il sistema. Allora, vale ancora con $\mathbb{E}[X]$ il numero totale medio di utenti, e $\mathbb{E}[S]$ il tempo totale di sistema medio, dal momento in cui un utente attraverso il confine verso il sistema fino a quando l'utente riattraverso il confine per lasciare il sistema.

La legge di Little VIII

Questa legge vale anche per una singola coda come: $\mathbb{E}[X_Q] = \lambda \mathbb{E}[W]$, dove $\mathbb{E}[X_Q]$ è il contenuto medio della coda senza servente e $\mathbb{E}[W]$ è il tempo di attesa medio.
Se il nostro sistema è definito solo da un singolo servente allora: $\mathbb{E}[X_S] = \lambda \mathbb{E}[Z]$, dove $\mathbb{E}[X_S]$ è lo stato medio del servente (tra 0 e 1) e $\mathbb{E}[Z]$ è il tempo di servizio medio.

Sistemi semplici a coda I

La maggior parte delle misure di prestazioni possono essere valutati dalla distribuzione di probabilità stazionaria della lunghezza della coda $\pi_n = P[X = n], n = 0, 1, \ldots$ del sistema. Data una specifica struttura stocastica, ossia alcune distribuzioni di probabilità particolari per gli eventi a e d, si cerca di ottenere le espressioni analitiche esplicite per le probabilità π_n stazionarie dello stato. In particolare si assume che le distribuzioni dei tempi di inter-arrivo e di servizio sono entrambi esponenziali con parametri λ e μ. La funzione di transizione dello stato per il sistema a coda è

$$f(x, a) = x + 1$$
$$f(x, d) = x - 1$$

Sistemi semplici a coda II

I processi di Poisson hanno alcune proprietà importanti. L'evento *utente che arriva all'istante t trova* $X(t) = n$ viene comparato con l'evento *stato del sistema all'istante t è* $X(t) = n$. Nel primo caso, l'osservazione dello stato del sistema, ossia la lunghezz della coda, avviene a specifici istanti temporali che dipendono dalla natura del processo di arrivo. Nel secondo caso, l'osservazione del sistema avviene ad istanti random. Tuttavia i due eventi sono uguali per i processi di arrivi poissoniani.

Sistemi semplici a coda III

Si usa la notazione standard $\pi_n(t)$ indica la probabilità che lo stato del sistema coincida con $X(t) = n$, e $\alpha_n(t)$ indica la probabilità che l'utente quando arriva trovi lo stato $X(t) = n$. Per un sistema a coda con processo di arrivo di Poisson indipendente dal processo di servizio, lo probabilità che l'arrivo di un utente trovi n utenti nel sistema è la stessa probabilità che lo stato del sistema sia n, ossia

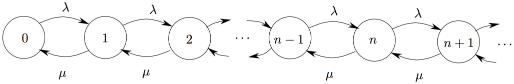
$$\alpha_n(t) = \pi_n(t)$$

Sistemi M/M/1 I

Si consideri un sistema singolo servente con capacità infinita e tempi di servizio e di inter-arrivo distribuiti esponenzialmente. Quindi

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

dove λ è la frequenza di arrivi di utenti e μ è la frequenza di servizio.



Sistemi M/M/1 II

Ne segue che

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

La somma del denominatore è una serie geometrica che converge per $\lambda/\mu < 1$. Sotto queste assunzioni, si ottiene

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

$$\pi_0 = 1 - \frac{\lambda}{\mu}$$

Sia $\rho = \frac{\lambda}{\mu}$ l'intensità del traffico allora, $\pi_0 = 1 - \rho$.

Sistemi M/M/1 III

La proprietà stazionaria è pari a

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n (1-\rho)$$
$$\pi_n = (1-\rho)\rho^n, \qquad n = 0, 1, \ldots$$

Avendo $0 \leq \rho \leq 1$, si ottiene che

$$S_1 = 1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n$$
$$S_2 = \sum_{n=1}^{\infty} \left(\frac{\mu}{\lambda}\right)^n \frac{1}{\lambda}$$

Si può notare come $S_1 < \infty$ e $S_2 = \infty$, per $\rho = \lambda/\mu < 1$, che sono dette condizioni di stabilità per un sistema $M/M/1$.

Sistemi M/M/1 IV

L'utilizzazione del sistema è pari a

$$\rho = 1 - \pi_0$$

Il volume di produzione è la frequenza di partenze dal servente, ed è pari a

$$\lambda = \mu(1 - \pi_0)$$

Come ci si aspetta da un sistema per cui ingressi e uscite sono bilanciate.
Per un sistema M/M/1 stabile, il volume di produzione è la frequenza di arriva λ.
Se $\lambda > \mu$, allora il volume di produzione è semplicemente μ, poichè il servente è costantemente in funzione alla frequenza μ.

Sistemi M/M/1 V

La lunghezza media della coda espressa da X è pari a

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} n\pi_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n$$

Si calcola la sommatoria

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

Sistemi M/M/1 VI

Usando la legge di Little, si ha

$$\frac{\rho}{1-\rho} = \lambda \mathbb{E}\left[S\right]$$

$$\mathbb{E}\left[S\right] = \frac{1}{\mu(1-\rho)}$$

Per $\rho \to 0$ si ha che $\mathbb{E}\left[S\right] \to 1/\mu$, che è il tempo medio di servizio.

Sistemi M/M/1 VII

A regime si che $\mathbb{E}[S] = \mathbb{E}[W] + \mathbb{E}[Z] = \mathbb{E}[W] + 1/\mu$. Per cui si ottiene

$$\mathbb{E}[W] = \frac{1}{\mu(1-\rho)} - \frac{1}{\mu}$$
$$\mathbb{E}[W] = \frac{\rho}{\mu(1-\rho)}$$

Esempio di M/M/1 I

In un sistema di comunicazione, la velocità di trasmissione è $C = 1200 bit/sec$. I messaggi in arrivo che devono essere trasmessi formano un processo Poisson. Un messaggio consiste di L bits, dove L è una variabile random. Si assume che L è esponenzialmente distribuito con media 600 bits. Il problema è determinare la frequenza di arrivi massimi che si possono sostenere per garantire il tempo di attesa medio di un messaggio minore di 1 secondo.

Esempio di M/M/1 II

Si comincia con il determinare la frequenza di servizio μ espresso in messaggio al secondo:

$$\mu = \frac{C}{L} = 2$$

$$\rho = \frac{\lambda}{2}$$

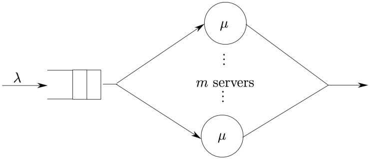
Quindi si può impone il vincolo sui tempi di attesa medi come

$$\mathbb{E}[W] = \frac{\lambda/2}{2(1-\lambda/2)} = \frac{\lambda}{4-2\lambda} < 1$$

Quindi $\lambda \leq 4/3$ messaggi/sec. Sotto queste condizioni, il sistema deve operare ad una frequenza di utilizzazione $\rho < 2/3$.

Sistemi M/M/m I

Questo è un sistema con capacità infinita con *m* serventi. Quando arriva un utente, un utente è servito da un servente disponibile. Se tutti i serventi sono occupati, l'utente è accodato fino a che la prossima partenza libera un servente. I temi di inter-arrivo sono esponenzialmente distribuiti con frequenza λ. I tempo di servizio per ogni servente sono esponenzialmente distribuiti con frequenza μ.



Sistemi M/M/m II

Si noti come la frequenza effettiva di servizio varia a seconda dello stato del sistema: se ci sono $n < m$ utenti presenti, allora ci sono n serventi occupati e la frequenza del servizio è $n\mu$; se, invece $n \geq m$ utenti sono presenti, la frequenza di servizio è fissata al suo massimo valore $m\mu$. In questo caso il modello è

$$\lambda_n = \lambda, \qquad n = 0, 1, \ldots$$

$$\mu_n = \begin{cases} n\mu, & 0 \leq n < m \\ m\mu, & n \geq m \end{cases}$$

Sistemi M/M/m III

La probabilità stazionaria π_0 può essere divisa in due parti: una per $n = m - 1$ e una per $n = m$

$$\pi_0 = \left[1 + \sum_{n=1}^{m-1} \frac{\lambda^n}{\mu(2\mu)\dots(n\mu)} + \frac{\lambda^{m-1}}{(m-1)!\mu^{m-1}} \sum_{n=m}^{\infty} \left(\frac{\lambda}{m\mu} \right)^{n-m+1} \right]^{-1}$$

Questa somma è una serie geometrica che converge per $\lambda / m\mu < 1$.

Sistemi M/M/m IV

Si impone che

$$\rho = \frac{\lambda}{m\mu}, \qquad 0 \leq \rho < 1$$

e quindi il risultato diventa

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

Sistemi M/M/m V

64

Manca anche il calcolo per π_n, $n = 1, 2, \ldots$.

$$\pi_n = \frac{(\lambda/\mu)^n}{n!}\pi_0, \qquad n = 1, 2, \ldots, m-1$$

$$\pi_n = \frac{m^m}{m!}\left(\frac{\lambda}{m\mu}\right)^n \pi_0, \qquad n = m, m+1, \ldots$$

Usando $\rho = \lambda/(m\mu)$, si ottiene

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

Sistemi M/M/m VI

Sia B la variabile random che denota il numero di serventi occupati, $B \in \{0, 1, \ldots, m\}$

$$\mathbb{E}[B] = \sum_{n=0}^{m-1} n\pi_n + mP[X \geq m]$$

La probabilità che almeno m utenti siano presenti, ossia $P[X \geq m]$, è data da

$$P[X \geq m] = \sum_{n=m}^{\infty} \pi_n = \sum_{n=m}^{\infty} \frac{m^m}{m!}\rho^n \pi_0 = \frac{m^m}{m!}\frac{\rho^m}{1-\rho}\pi_0$$

Dopo alcune analisi si ottiene

$$\mathbb{E}[B] = m\rho = \frac{\lambda}{\mu}$$

Sistemi M/M/m VII

Ogni servente ha un'utilizzazione pari a $\mathbb{E}[B]/m = \rho$.
Il volume di produzione del sistema è pari a λ, poichè a regime le frequenze degli arrivi e delle partenze devono essere bilanciate.

Sistemi M/M/m VIII

Per la legge di Little, si ha che

$$m\rho + \frac{(m\rho)^m}{m!} \frac{\rho}{(1-\rho)^2} \pi_0 = \lambda \mathbb{E}[S]$$

$$\mathbb{E}[S] = \frac{1}{\mu} + \frac{1}{\mu} \frac{(m\rho)^m}{m!} \frac{\pi_0}{m(1-\rho)^2}$$

Si noti che $\mathbb{E}[S] \to 1/\mu$ per $\rho \to 0$, poichè a bassi valori di traffico gli utenti in arrivo trovano sempre un servente disponibile e il tempo di sistema è limitato al tempo di servizio

Sistemi M/M/m IX

La probabilità che un utente in arrivo non trovi un servente libero, e quindi è costretto ad aspettare in coda, è indicata da P_Q

$$P_q = P\left[X \geq m\right] = \sum_{n=m}^{\infty} \pi_n = \frac{(m\rho)^m}{m!} \frac{\pi_0}{1-\rho}$$