TECNICHE DI SIMULAZIONE

Modelli a coda

Francesca Mazzia

Dipartimento di Matematica
Università di Bari

a.a. 2004/2005

Sistemi a coda

- Gli elementi chiave di un sistema a coda sono i clienti e i server.
- Il termine clienti può riferirsi a persone, macchine, camion, pazienti, aerei, e-mail, ordini, ..., qualsiasi cosa che richiede un servizio.
- il termine server si riferisce a addetti all'accoglienza, meccanici, medici, piste aereoportuali, CPU in un computer, ... ogni risorsa che effettua il servizio richiesto.
- Noi consideriamo un cliente che va dal server, analogamente può essere il server che va dal cliente.

Calling population

- L'insieme dei potenziali clienti si chiama "calling population", essa può essere sia finita che infinita.
- Per esempio consideriamo un insieme di 5 macchine che eseguono un lavoro.
- Dopo un certo intervallo di tempo una macchina si apre e l'addetto rimuove l'oggetto dalla macchina e ne inserisce un'altro.
- Le macchine sono i clienti che arrivano nell'istante in cui si aprono.
- L'addetto è il server che serve la macchina appena possibile.
- La "calling population" è finita e consiste delle 5 macchine.

Calling population

- In un sistema con un insieme grande di potenziali clienti, la "calling population" si considera infinita.
- Per tali sistemi questa assunzione è innocua e può semplificare il modello.
- Esempi sono i clienti di un ristorante, di una banca, ...
- Anche quando i possibili clienti sono in numero finito ma grande, si può sempre assumere che siano infiniti, a patto che il numero di clienti dentro il sistema si una piccola porzione del numero totale dei possibili clienti.

Calling population

- La differenza fra modelli con "calling population" finita o infinita è nella definizione del tasso di arrivo.
- In un modello con popolazione infinita, il tasso di arrivo (cioè il numero medio di arrivi per unità di tempo) non cambia a seconda del numero di clienti che hanno lasciato la popolazione e sono entrati nel sistema.
- Se il modo di arrivo è omogeneo nel tempo (non ci sono periodi di maggiore afflusso, il tasso di arrivo si assume costante.
- Per modelli con popolazione finita il tasso di arrivo dipende dal numero di clienti che sono nel sistema (in coda o sono al momento serviti).

Calling population

- Consideriamo l'esempio delle 5 macchine servite da un singolo addetto.
- Quando tutte e 5 sono chiuse l'addetto è libero e la probabilità che una macchina si apra è massima.
- Appena una macchina si apre il tasso di arrivo decresce.
- Se le macchine sono tutte aperte il tasso di arrivo è 0.

Capacità del sistema

- In alcuni sistemi a coda vi è un limite sul numero di clienti che può attendere in coda.
- per esempio in un autolavaggio vi può essere posto solo per 10 macchine.
- Un cliente che arriva e trova la coda piena, non entra nel sistema.
- In altri sistemi la coda può essere considerata infinita.
- Per esempio il botteghino di un teatro.
- Quando un sistema ha capacità limitate si fa una distinzione fra tasso di arrivo e tasso di arrivo effettivo (numero di clienti che arrivano e entrano nel sistema)

Il processo di arrivo

- Il processo di arrivo per modelli con popolazione infinita è caratterizzato in termini di tempi di interarrivo di clienti successivi.
- Gli arrivi possono verificarsi in tempi prefissati o casuali.
- Quando gli arrivi sono casuali, i tempi di interarrivo sono caratterizzati da una distribuzione di probabilità.
- Gli arrivi possono verificarsi uno alla volta o in gruppo.
- I gruppi possono essere con un numero di elementi costanti o casuali.

Il processo di arrivo

- Il più importante modello per gli arrivi casuali è il processo di Poisson.
- Se A_n rappresenta il tempo di interarrivo fra il cliente $n-1$ e il cliente n, allora A_n è distribuita esponenzialmente con media $1/\lambda$.
- Il tasso di arrivo è λ clienti per minuto.
- Il numero di arrivi in un intervallo di lunghezza t, $N(t)$, ha distribuzione di Poisson con media λt

Il processo di arrivo

- Il processo di Poisson è stato usato con successo come modello per simulare l'arrivo dei clienti in ristoranti, drive-in, banche, gli arrivi di chiamate telefoniche in un centralino, gli arrivi di richieste o ordini, gli arrivi di componenti rotte in un centro di riparazione.

Il processo di arrivo

- Una classe importante di arrivi è gli arrivi per appuntamento. Come gli arrivi dei pazienti in uno studio medico, o gli arrivi degli aerei in un aereoporto.
- In questo caso i tempi di interarrivo possono essere costanti con delle fluttuazioni casuali che tengono conto di eventuali anticipi o ritardi.

Il processo di arrivo

- Una terza situazione è quando si suppone che vi è sempre un cliente in coda.
- per esempio si può supporre che il cliente è del materiale grezzo per un prodotto, e che ne abbiamo a disposizione sempre una quantità sufficiente.

Il processo di arrivo

- Per modelli a popolazione finita il processo di arrivo è caratterizzato in modo completamente diverso.
- Un cliente si definisce in sospeso quando si trova fuori dal sistema ed è un membro della popolazione.
- Definiamo runtime per un dato cliente il tempo fra la partenza dal sistema a coda e l'arrivo successivo nella coda.

Il processo di arrivo

- Una applicazione importante dei modelli a popolazione finita è il problema di riparazione delle macchine.
- Le macchine sono i clienti e il runtime è chiamato tempo di guasto.
- Quando una macchina si guasta, arriva nel sistema a coda per essere riparata e rimane nel sistema fino a quando non è terminata la riparazione.
- I tempi di guasto per classi di macchine sono stati caratterizzati da distribuzioni esponenziali, di Weibull e gamma.

Comportamento della coda

- Il comportamento della coda si riferisce alle azioni dei clienti mentre aspettano in coda.
- Vi è la possibilità che i clienti lascino la coda se vedono che è troppo lunga.
- I clienti possono lasciare la coda se si accorgono che si muove troppo lentamente.
- possono spostarsi da una fila all'altra se pensano di avere scelto la fila più lenta.

Disciplina della coda

- La disciplina della coda si riferisce all'ordine in cui i clienti sono scelti dalla coda per essere serviti.
- FIFO: first in, first out.
- LIFO: last in first out.
- SIRO: servizio in ordine casuale.
- SPT: viene processato prima il servizio con tempo più breve
- PR: servizio per priorità.

Tempo di servizio

- I tempi di servizio di arrivi successivi sono denotati con S_1, S_2, ...
- possono essere di durata costante o casuale.
- se di durata casuale sono caratterizzati da una sequenza di variabili casuali indipendenti e distribuite nello stesso modo.
- Le distribuzioni esponenziale, gamma, di Weibull, lognormale e normale troncata, sono state usate con successo per modellare tempi di servizio.

Tempo di servizio

- Alcune volte i tempi di servizio possono essere distribuiti allo stesso modo per classi di clienti e in modo completamente diverso per altre.
- In alcuni casi il tempo di servizio dipende anche dall'ora del giorno, o dalla lunghezza della coda.

Sistemi a coda

- Un sistema a coda consiste in un numero di centri di servizio e di code interconnesse.
- Ogni centro di servizio è composto dallo stesso numero di server c che lavorano in parallelo: cioè quando arriva alla fine della coda il cliente viene servito dal primo server che si libera.
- Meccanismi di servizio paralleli sono o con un singolo server, o con molti server o con un numero di server illimitato (un sistema self-service è caratterizzato dall'avere un numero di server illimitato).

Notazioni

Una notazione per sistemi con server paralleli che viene comunemente usata è la seguente: $A/B/c/N/K$

- A : distribuzione dei tempi di interarrivo
- B: distribuzione dei tempi di servizio
- per A e B si usano, fra gli altri, i seguenti simboli:
  - M :esponenziale o Markov
  - D :costante o deterministica
  - E_k :Erlang di ordine k
  - G : arbitraria o generale
  - GI generale indipendente

Notazioni $A/B/c/N/K$

- c : numero di server in parallelo
- N : capacità del sistema
- K : dimensione della "calling population"
- Esempio: $M/M/1/\infty/\infty$ rappresenta un sistema con un singolo server con capacità illimitata, popolazione illimitata e distribuzione esponenziale dei tempi di interarrivo e dei tempi di servizio.
- Quando N e K sono infiniti si eliminano dalla notazione
- Esempio: $M/M/1/\infty/\infty$ corrisponde a $M/M/1$

Notazioni

- Assumiamo che ogni sistema ha una disciplina FIFO della coda.
- P_n probabilità stazionaria rispetto al tempo di avere n clienti nel sistema
- $P_n(t)$ probabilità di avere n clienti nel sistema al tempo t
- λ tasso di arrivo
- λ_e tasso di arrivo effettivo
- μ tasso di servizio di un server
- ρ utilizzo del server

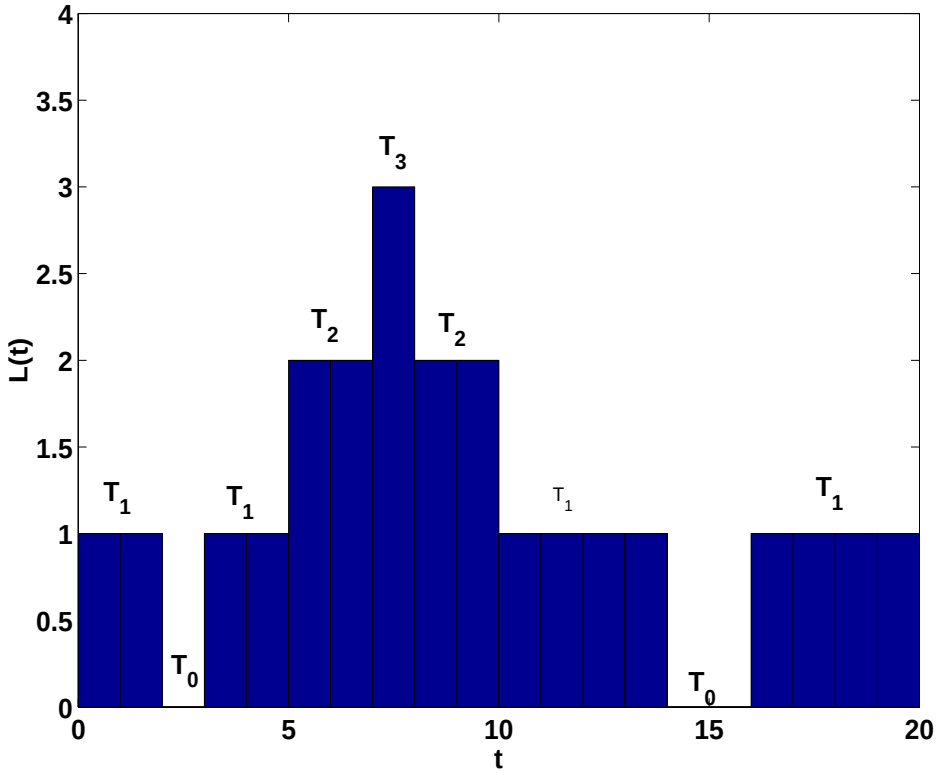
Notazioni

- A_n tempo di interarrivo fra il cliente $n-1$ e n
- S_n tempo di servizio del cliente n
- W_n tempo totale passato nel sistema dal cliente n
- W_n^Q tempo totale passato in coda dal cliente n
- $L(t)$ numero totale di clienti nel sistema al tempo t
- $L_Q(t)$ numero di clienti in coda al tempo t

Misure a lungo termine

- L numero medio rispetto al tempo di clienti nel sistema per tempi lunghi
- L_Q numero medio rispetto al tempo di clienti in coda per tempi lunghi
- w numero medio rispetto al tempo passato nel sistema per cliente per tempi lunghi
- w_Q numero medio rispetto al tempo passato nella coda per cliente per tempi lunghi

Misure a lungo termine



Sistema $G/G/c/N/K$

- studiamo un sistema a coda per un periodo di tempo T, $L(t)$ è il numero di clienti nel sistema al tempo t.
- Sia T_i il tempo totale in cui il sistema contiene i clienti in $[0, T]$

Misure a lungo termine

- Nell'esempio $T_0 = 3, T_1 = 12, T_2 = 4, T_3 = 1$.
- In generale $\sum_{i=0}^{\infty} T_i = T$.
- Il numero medio di clienti pesato rispetto al tempo è definito da:

$$\hat{L} = \frac{1}{T} \sum_{i=1}^{\infty} i T_i = \sum_{i=0}^{\infty} i \frac{T_i}{T}$$

- Nell'esempio
  $\hat{L} = [0(3) + 1(12) + 2(4) + 3(1)]/20 = 23/20 = 1.15$.

Misure a Lungo termine

- L'area totale sotto la funzione $L(t)$ può essere decomposta in rettangoli di altezza i e larghezza T_i.
- L'area totale è data da $\sum_{i=0}^{\infty} iT_i = \int_0^T L(t)dt$ quindi:

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int_0^T L(t)dt$$

- Questo è sempre vero per qualsiasi sistema a coda, indipendentemente dal numero di server, dalla disciplina della coda.

Misure a Lungo termine

- Per molti sistemi, quando T cresce $\hat{L}$ si avvicina ad un valore limite, che è chiamato numero medio rispetto al tempo, a lungo termine, di clienti nel sistema L.
- $\hat{L}$ è influenzato dalle condizioni iniziali e da T

Misure a lungo termine

L'equazione precedente è valida anche se si considerano dei sottosistemi del sistema a coda.

- $$\hat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int_0^T L_Q(t)dt$$
- $$\hat{L}_Q \to L_Q \text{ per } T \to \infty$$

Tempo medio passato nel sistema per cliente

- Se simuliamo un sistema a coda per un periodo di tempo T, possiamo memorizzare il tempo che ogni cliente passa nel sistema nel periodo $[0, T]$ (W_1, W_2, ...)
- Il tempo medio speso da un cliente nel sistema è dato da:

$$\hat{w} = \frac{1}{N} \sum_{i=1}^{N} W_i$$

- Per sistemi stabili si ha che $\hat{w} \rightarrow w$ con probabilità 1.
- w è il tempo medio passato nel sistema per tempi lunghi.

Tempo medio passato nel sistema per cliente

- Possiamo considerare anche solo il tempo passato in coda:

$$\hat{w}_q = \frac{1}{N} \sum_{i=1}^{N} W_i^Q$$

- $\hat{w}$ è influenzato dalla condizioni iniziali e da T.

Esempio

- Nel sistema della figura precedente, arrivano $N = 5$ clienti.
- Le informazioni sul primo e l'ultimo cliente si possono ricavare dal grafico. $W_1 = 2$ e $W_5 = 20 - 16 = 4$.
- Gli altri tempi si possono ricavare solo se si specifica ulterioremente il tipo di sistema.
- Supponiamo che il sistema è una coda con singolo server e disciplina FIFO.
- I clienti lasciano il sistema nello stesso ordine in cui sono arrivati.
- Quindi ogni salto in $L(t)$ rappresenta un arrivo.

Esempio

- Gli arrivi sono al tempo 0,3,5,7,16.
- Le partenze si verificano a: 2,8,10,14
- Con queste ipotesi si ha che
  $W_2 = 8 - 3 = 5, W_3 = 10 - 5 = 5, W_4 = 14 - 7 = 7.$
  $\hat{w} = (2 + 5 + 5 + 7 + 4)/5 = 23/5 = 4.6$

Legge di conservazione

- Il tasso di arrivo $\hat{\lambda} = N/T = 5/20 = 1/4$ clienti per unità di tempo.
- $\hat{L} = 1.15$ $\hat{w} = 4.6$, $\hat{L} = \hat{\lambda}\hat{w}$
- Questa relazione tra L, λ, w non è una coincidenza, ma si verifica per quasi tutti i sistemi a coda, indipendentemente dal numero di server e dalla disciplina della coda.
- Per $T \to \infty$ si ha $L = \lambda w$
- Questa si chiama equazione di conservazione.

Legge di conservazione

- Il numero medio di clienti nel sistema, in un tempo arbitrario, è uguale al numero medio di arrivi per unità di tempo moltiplicato il tempo medio passato dai clienti nel sistema

$$\sum_{i=1}^{N} W_i = \int_0^T L(t)dt$$

$$\hat{L} = \frac{1}{T}\int_0^T L(t)dt = \frac{N}{T}\frac{1}{N}\sum_{i=1}^{N} W_i = \hat{\lambda}\hat{w}$$

Utilizzo del server

- È la proporzione del tempo in cui un server è occupato.
- L'utilizzo del server nell'intervallo $[0, T]$ viene chiamato $\hat{\rho}$.
- L'utilizzo del server per tempi lunghi viene denotato da ρ.
- per sisteme stabili si ha che:

$$\hat{\rho} \to \rho \text{ per } T \to \infty$$

- Nell'esempio della figura, se il sistema ha un singolo server: $\hat{\rho} = (\text{ tempo in cui il server è occupato})/T = (T - T_0)/T = 17/20$

Utilizzo del server in $G/G/1$

- Consideriamo un sistema a coda con un singolo server, tasso medio di arrivo λ clienti per unità di tempo, tempo medio di servizio $E(S) = 1/\mu$ unità di tempo, capacità infinita e popolazione infinita.
- $E(S) = 1/\mu$ implica che quando il server è occupato vengono serviti μ clienti per unità di tempo, μ è chiamato tasso di servizio.
- Il server è un sottosistema che può essere considerato come un sistema, quindi vale la legge di conservazione.

Utilizzo del server in $G/G/1$

- Per sistemi stabili il tasso medio di arrivo al server, λ_s deve essere identico al tasso medio di arrivo nel sistema λ.
- λ_s deve essere $\leq \lambda$
- Se $\lambda_s < \lambda$ la lista di attesa tende a crescere in lunghezza con tasso medio di $\lambda - \lambda_s$ clienti per unità di tempo, quindi il sistema risulta instabile.

Utilizzo del server in $G/G/1$

- Per il sottosistema server il tempo medio del sistema è $w = E(S) = 1/\mu$.
- Il numero di clienti nel sottosistema è 0 o 1.
- Il numero medio
  $\hat{L}_s = \frac{1}{T}\int_0^T (L(t) - L_Q(t))dt = (T - T_0)/T$
- In generale il numero medio di clienti che sono serviti in un generico tempo è uguale all'utilizzo del server $\hat{L}_s = \hat{\rho}$.

Utilizzo del server in $G/G/1$

- Per $T \to \infty, \hat{L}_s = \hat{\rho} \to L_s = \rho$.
- L'equazione di conservazione per il sottosistema server diventa:

$$\rho = \lambda/\mu$$

- l'utilizzo del server per tempi lunghi è uguale al tasso medio di arrivo diviso il tempo di servizio.

Utilizzo del server in $G/G/1$

- Una coda con un singolo server per essere stabile deve avere $\lambda < \mu$ o $\rho = \lambda/\mu < 1$.
- Se il tasso di arrivo è maggiore del tasso di servizio, il sever rimarrà sempre più indietro, e dopo un certo tempo sarà sempre occupato, la coda tenderà a crescere con tasso medio $(\lambda - \mu)$ clienti per unità di tempo.
- le misure a lungo tempo hanno significato solo per sistemi stabili.
- per sistemi instabili l'uilizzo del server a lungo termine è 1, la lunghezza media della coda a lungo termine è infinita.

Utilizzo del server in $G/G/c$

- Analizziamo un sistema a coda con c server identici che lavorano in parallelo.
- Se un cliente arriva e trova più di un server libero, ne sceglie uno in modo casuale.
- Il tasso di arrivo è λ arrivi per unità di tempo da una popolazione infinita.
- Ogni server lavora con tasso di servizo μ clienti per unità di tempo.
- Il numero medio di server occupati è dato da $L_s = \lambda/\mu$, con $0 \leq L_s \leq c$.

Utilizzo del server in $G/G/c$

- L'utilizzo medio del server a lungo termine è definito da: $\rho = L_s/c = \lambda/(c\mu)$ con $0 \leq \rho < 1$
- il massimo tasso di servizio è $c\mu$.
- per sistemi stabili deve essere $\lambda < c\mu$.
- se $\lambda > c\mu$ la coda crescerà con tasso medio $(\lambda - c\mu)$ clienti per unità di tempo.

Esempio

- Clienti arrivano in modo casuale in un ufficio, con tasso $\lambda = 50$ clienti per ora. Ci sono 20 impiegati, i quali servono $\mu = 5$ clienti per ora in media.
- L'utilizzo medio stazionario di un server è $\rho = \lambda/(c\mu) = 0.5$
- il numero medio di server occupati è $L_s = \lambda/\mu = 10$.
- Per tempi lunghi un impiegato è occupato solo il 50 % del tempo.
- Il numero di server può essere diminuito?
- Per la stabilità è necessario che sia $c > 10$, quindi $c = 11, 12, 13..$
- $c = 11$ assicura solo la stabilità a lungo termine, un sistema stabile può avere code anche molto lunghe.

Utilizzo del server $G/G/1$

- La performance del sistema può variare molto per un dato valore ρ di utilizzo del server.
- Consideriamo una coda $G/G/1$ con tasso di arrivo λ, tasso di servizo μ, utilizzo del server $\rho = \lambda/\mu < 1$.
- Consideriamo il caso $D/D/1$ con tempi di servizio e di arrivo deterministici.
- I tempi di interarivo $A_1, A_2, A_3, ...$ sono uguali a $E(A) = 1/\lambda$, e i tempi di servizio $S_1, S_2, S_3, ...$ sono uguali a $E(S) = 1/\mu$.
- Assumiamo che un cliente arriva al tempo zero nel sistema vuoto, il sistema evolve in modo completamente deterministico.$L = \rho = \lambda/\mu$, $w = 1/\mu$, $w_q = L_q = 0$.

Esempio

- In generale è la variabilità dei tempi di interarrivo e dei tempi di servizio che fa fluttuare i tempi in cui i server sono occupati.

Esempio

- Consideriama lo studio di un medico, che prende un appuntamento ogni 10 minuti, e che fa visite di lunghezza 9 minuti con probablità 0.9 oppure 12 minuti con probabilità 0.1
- Gli arrivi sono deterministici, ma i tempi di servizio casuali con media e varianza data da:
  $E(S_i) = 9(0.9) + 12(0.1) = 9.3$ minuti,
  $V(S_i) = E(S_i^2) - E(S_i)^2 = 0.81$
- $\rho = \lambda/\mu = E(S)/E(A) = 9.3/10 = 0.93 < 1$, quindi il sistema è stabile.
- Il dottore sarà occupato il 93 % del tempo.

Esempio

- Se i pazienti richiedono solo 9 minuti di servizio non ci saranno code, la coda si forma appena arriva un paziente che richiede 12 minuti di servizio.
- Sia $S_1 = 9, S_2 = 12, S_3 = 9, S_4 = 9, ..$
- Il primo paziente entra nel sistema al tempo zero ed esce al tempo 9, il secondo entra a 10 ed esce a 22. Il terzo entra a 20 e deve aspettare in coda per 2 minuti, inizia il servizio a 22 e trmina a 31. Il quarto entra a 30, e deve aspettare in coda per 1 minuto, inizia il servizio a 31.

Costo

- Il costo può essere associato per esempio alla coda o ai server.
- Supponiamo che il sistema ha un costo per ogni cliente che attende in coda, per esempio 10 euro per ora per cliente.
- Se il cliente j passa W_j^Q ore in coda allora $\sum_{i=1}^{N} 10 * W_j^Q$ è il costo totale per N clienti.
- Il costo medio per cliente è:

$$\sum_{j=1}^{N} 10 W_j^Q / N = 10 \hat{w}_q$$

Costo

- Se $\hat{\lambda}$ clienti arrivano per ora in media, il costo medio per ora è:

$$(\hat{\lambda}\frac{\text{clienti}}{\text{ora}})(\frac{10\hat{w}_Q}{\text{clienti}}) = 10\hat{\lambda}\hat{w}_Q = 10\hat{L}_Q/\text{ora}$$

Costo

- Se un gruppo di c server in parallelo ha utilizzo ρ e ogni server ha un costo di 5 euro per ora, mentre è occupato, allora il costo totale per server è $5c\rho$ euro, $c\rho$ è il numero medio di server occupati.
- Se il costo è quando il server è libero allora il costo del server per ora è $5c(1-\rho)$.
- In molti sistemi questi costi sono combinati nei costi totali.
- In molti casi l'obiettivo è minimizzare i costi totali, variando i parametri sotto controllo del manager, come numero di server, tasso di arrivo, tasso di servizio o capacità del sistema.