Raffaele Pesenti

dispense di

TEORIA DELLE CODE O FILE D'ATTESA

Indice

1 INTRODUZIONE ... 3

1.1 INTRODUZIONE E PREREQUISITI ... 3
1.2 SCOPI E OBIETTIVI ... 4
1.3 TESTI ... 4
1.4 DOMANDE ED ESERCIZI ... 5

2 FONDAMENTI ... 6

2.1 RICHIAMI DI PROBABILITÀ ... 6
2.2 IL SISTEMA CODA E LE SUE COMPONENTI. ... 6
2.3 LA NOTAZIONE DI KENDALL ... 9
2.4 LO STATO DI UNA CODA ... 9
2.5 DOMANDE ED ESERCIZI ... 9

3 LE PROBLEMATICHE DI INTERESSE ... 11

3.1 LE PROBLEMATICHE D'INTERESSE ... 11
3.2 IL CASO DETERMINISTICO D/D/1 ... 12
3.3 DOMANDE ED ESERCIZI ... 13

4 IL RUOLO DELLE DISTRIBUZIONI ESPONENZIALE E DI POISSON ... 14

4.1 LA DISTRIBUZIONE ESPONENZIALE ... 14
4.2 IL PROCESSO DI POISSON ... 16
4.3 DOMANDE ED ESERCIZI ... 17

5 IL PROCESSO NASCITE - MORTI ... 18

5.1 IL PROCESSO NASCITE - MORTI ... 18
5.2 DOMANDE ED ESERCIZI ... 19

6 LA CODA M/M/1 ... 21

6.1 LA CODA M / M / 1 ... 21
6.2 FORMULA DI LITTLE ... 22
6.3 L'INFLUENZA DEL FATTORE DI UTILIZZAZIONE ... 23
6.4 L'INTERTEMPO TRA DUE PARTENZE ... 24
6.5 DOMANDE ED ESERCIZI ... 24

7 ALTRE CODE POSSONIANE ... 25

7.1 ALTRE CODE POISSONIANE (M / M / …) ... 25
7.2 M / M / S ... 25
7.3 M / M / 1 / K ... 26
7.4 M / M / 1 / N / N ... 27
7.5 DOMANDE ED ESERCIZI ... 27

8 ALCUNE CODE NON POISSONIANE ... 28

8.1 ALCUNE CODE NON POISSONIANE ... 28
8.2 M / G / 1 ... 28
8.3 M / D / 1 ... 28
8.4 M / E_K / 1 ... 29
8.5 M / H_R / 1 ... 30
8.6 DOMANDE ED ESERCIZI ... 30

INDICE ANALITICO ... 31

1 Introduzione

1.1 Introduzione e prerequisiti
1.2 Scopi e obiettivi
1.3 Testi
1.4 Domande ed esercizi

1.1 Introduzione e prerequisiti

Questo testo raccoglie e commenta le lezioni su Teoria delle Code svolte dal Prof. Walter Ukovich all'interno del corso di Matematica III.

La **Teoria delle Code** (o file d'attesa) si propone di sviluppare modelli per lo studio dei fenomeni d'attesa che si possono manifestare in presenza di una domanda di un servizio. Quando la domanda stessa e/o la capacità di erogazione del servizio sono soggetti ad aleatorietà, si possono infatti verificare situazioni temporanee in cui chi fornisce il servizio non ha la possibilità di soddisfare immediatamente le richieste.

La stessa esperienza comune suggerisce che i campi di applicazione della teoria delle code sono estremamente numerosi. Ad esempio sono in generale soggetti ad attese:
- i clienti in banca o in posta;
- le persone in attesa di un taxi o comunque le chiamate ad un servizio di radiotaxi;
- le automobili ad un incrocio;
- gli aerei in attesa di decollare o di atterrare;
- le parti in attesa di essere lavorati;
- le macchine in avaria in un'officina;
- i progetti di legge in parlamento;
- . . .

ne consegue che i risultati ottenibili della Teoria delle Code trovano applicazione, ad esempio, nei:
- sistemi di elaborazione;
- sistemi di comunicazione / trasmissione dati;
- sistemi di trasporto;
- sistemi flessibili di lavorazione;
- . . .

In particolare i concetti fondamentali della Teoria delle Code venivano formalizzati nel 1917 da Erlang proprio per applicarli nel settore telefonico.

La Teoria delle Code può trovare applicazione non solo nel settore industriale e dei servizi, ma può portare immediati vantaggi anche alla qualità della vita delle persone comuni. In questo contesto una delle ultime applicazioni osservate nell'esperienza quotidiana da parte dell'estensore di queste pagine è stata all'interno di un ospedale. La macchina distributrice dei biglietti che davano diritto alle visite specialistiche ambulatoriali segnava, oltre al numero di prenotazione, anche una ragionevole stima per difetto del tempo si sarebbe dovuto attendere prima di essere visitati. In base a tale informazione il paziente poteva decidere, ad esempio, di andare a prendere un caffè o svolgere alcune commissioni prima di ripresentarsi, evitando così di perdere tempo inutilmente.

Nelle pagine seguenti si assume che il lettore abbia una preparazione di base, molto elementare, sulla Teoria della Probabilità e sui Modelli Stocastici. In particolare si considereranno noti i seguenti concetti, anche se alcuni di essi verranno brevemente riproposti nei capitoli successivi:
- probabilità;
- leggi elementari della probabilità;
- variabili aleatorie;
- speranza matematica (proprietà);
- processi stocastici;
- distribuzione esponenziale.
- processo di Poisson.

1.2 Scopi e obiettivi

Lo **scopo** di questo testo è di introdurre alcuni dei concetti base che permettono lo studio dei fenomeni d'attesa. In particolare tra questi ultimi sono analizzati solo quelli che hanno una certa diffusione nel modo reale, ma che al contempo sono facilmente trattabili dal punto di vista matematico. A tal fine in questo testo sono proposti alcuni modelli elementari. Con l'ausilio di tali modelli elementari sono poi derivati dei risultati generali. Infine vengono proposti anche alcuni cenni su modelli più complessi.

Lo studente dovrebbe cercare di fare proprio il metodo di ragionamento utilizzato nella Teoria delle Code, piuttosto che memorizzare meccanicamente concetti e risultati. Per questo motivo è opportuno che al termine di ogni capitolo lo studente verifichi mentalmente la comprensione di quanto presentato e cerchi di rispondere alle domande e agli esercizi proposti.

Gli **obiettivi** del corso sono quelli di mettere lo studente in condizione di:
- scegliere la tipologia di modelli più adatti a diverse situazioni reali;
- individuare e calcolare i parametri che misurano le prestazioni di ciascuna situazione;
- giustificare formalmente l'espressione di tali parametri in casi molto semplici;
- risolvere semplici problemi di progetto;
- usare criticamente modelli nuovi, non trattati esplicitamente nelle lezioni.

1.3 Testi

Testo di riferimento

Hillier, Lieberman:
Introduction to Operations Research,
quinta edizione, McGraw-Hill (1990)
capitoli 16 (Queueing Theory) e 17 (The Application of Queueing Theory)

Altri testi di consultazione

- Kleinrock: Queueing Systems, Wiley 1975.
- Cooper: Queueing Theory, in: Heyman, Sobel (eds.): Handbooks in Operations Research and Management Science, Vol. 2: Stochastic Models, Elsevier 1990.
- Walrand: Queueing Networks, ibidem.

Il testo di Hillier e Lieberman è introduttivo e di facile lettura. Il testo di Kleinrock è un classico nel settore, è di livello intermedio ma molto ben scritto, per questo pur non essendo recente riceve ancora notevole attenzione. Gli ultimi due riferimenti sono articoli di rassegna inseriti nel secondo volume della serie degli Handbooks in Operations Research and Management Science. Questa serie contiene lavori in generale molto belli, ma essi devono però essere considerati più come strumento di consultazione che come testo di studio.

I risultati più recenti riguardanti la Teoria delle Code e le sue applicazioni sono pubblicati su riviste scientifiche. Tali riviste, generalmente disponibili nelle biblioteche delle Università italiane, afferiscono ai settori di: Ricerca Operativa, Controlli Automatici, Reti di Comunicazioni, Informatica (Sistemi Operativi) e Ingegneria dei Trasporti solo per citarne alcuni.

1.4 Domande ed esercizi

1. Enunciare quale è l'oggetto di studio della Teoria delle Code.

2. Elencare alcuni fenomeni di attesa di cui si ha avuto esperienza recentemente; valutare il costo economico pagato singolarmente e socialmente da quanti coinvolti in tali fenomeni; proporre azioni che conducano alla riduzione di tali costi a parità di servizio erogato.

3. Un adulto medio in età lavorativa produce un reddito di almeno 100 milioni di Lire all'anno, stimando che in Italia esistano almeno 20 milioni di persone in tale condizione e generalizzando l'esperienza personale, valutare il costo sociale complessivo dovuto ai fenomeni di attesa in cui tipicamente incorrono i cittadini del nostro paese.

4. Le industrie comprano le materie prime e i semilavorati necessari alla produzione dei beni finali prendendo i soldi occorrenti a prestito dalle banche. Tenendo presente gli attuali tassi di interesse e sapendo che tipicamente in un ciclo produttivo i beni passano almeno l'80% del tempo in attesa di essere lavorati, valutare il costo di tali attese.

5. Il dipendente dello stato, quando esegue una missione per motivi di servizio, di solito anticipa i soldi necessari e viene risarcito a viaggio concluso. L'estensore di questo testo ha svolto una missione negli Stati Uniti spendendo circa Lit.3.000.000. Per essere rimborsato deve necessariamente presentare un documento fornito dal Consolato Italiano a Los Angeles. Attualmente egli è in attesa di questo documento da più di un mese. Valutare economicamente il danno subito a causa della lentezza burocratica del Consolato in questione, nell'ipotesi che il denaro, se fosse disponibile, verrebbe investito in buoni del tesoro.

6. Commentare se, essendo il quinto in coda, è preferibile sapere che tipicamente ogni cliente è servito in un tempo che varia in modo assolutamente imprevedibile tra dieci minuti e la mezz'ora oppure sapere che ogni cliente richiede esattamente 25 minuti. Assumere che si possa conservare il proprio turno anche se ci si allontana. Dedurre alcune conclusioni circa gli indici che devono essere utilizzati per valutare la qualità di un servizio dal punto di vista del tempo necessario ad erogarlo. [*Oltre che al valore medio possono essere opportune delle indicazioni sulla variabilità del tempo di servizio*].

2 Fondamenti

2.1 Richiami di probabilità
2.2 Il sistema coda e le sue componenti
2.3 La notazione di Kendall
2.4 Lo stato di una coda
2.5 Domande ed esercizi

Nel primo paragrafo di questo capitolo sono introdotti alcuni concetti relativi alla teoria della probabilità. Questi concetti sono ripresi nel secondo paragrafo, quando viene definito il sistema coda e le sue componenti.

2.1 Richiami di probabilità

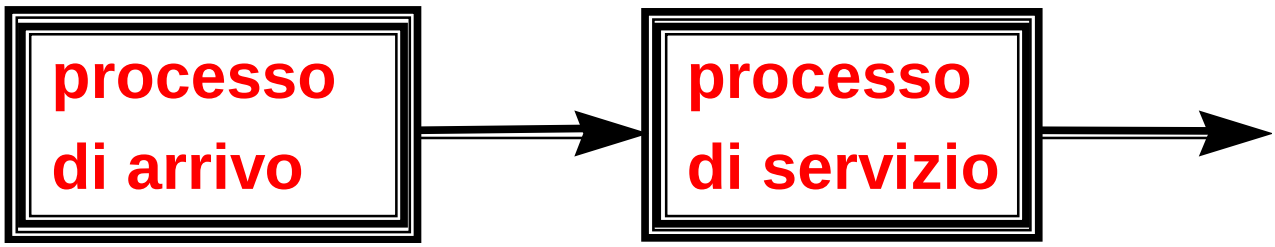
Una **variabile aleatoria** discreta X è un'entità che può assumere un numero discreto (finito o infinito) di valori x_i. Ogni valore x_i ha probabilità di occorrenza $P(X=x_i)=p_i$, dove $\Sigma_i\, p_i = 1$.

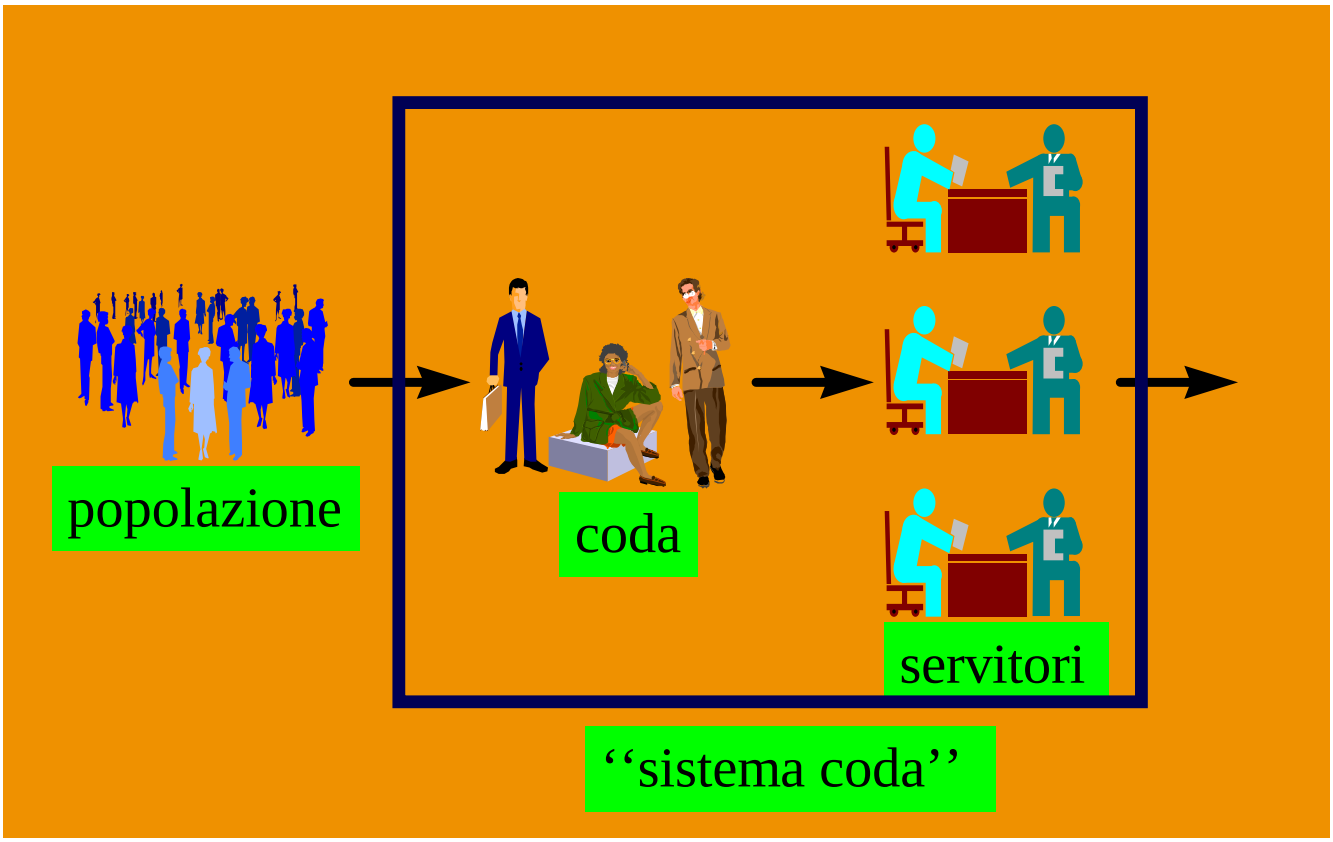
Una variabile aleatoria continua Y è un'entità che può assumere valori y in un sottoinsieme S della retta reale composto da uno o più intervalli. Ogni intervallo infinitesimo di ampiezza dy ha probabilità $P(y<Y\leq y+dy)=p(y)dy$ che Y assuma un valore all'interno di esso. La funzione p(y) è detta **funzione di densità** ed è tale che $\int_S p(y)dy = 1$.

Un **processo stocastico** è una variabile aleatoria i cui valori e le relative probabilità sono funzioni del tempo. L'evoluzione temporale di un processo stocastico può avvenire in modo continuo o discreto. Ad esempio il numero di persone presenti in una coda tipicamente varia ad istanti discreti all'occorrenza di determinati eventi. Viceversa il tempo che deve trascorrere fino all'arrivo di un dato cliente è un valore che varia con continuità.

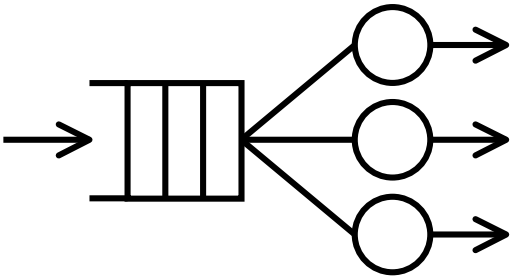
2.2 Il sistema coda e le sue componenti.

Dal punto di vista fisico un **sistema coda** è un sistema composto da un insieme non vuoto di **servitori,** capaci di fornire un servizio imprecisato, e da un insieme non vuoto di **aree di attesa** (**buffer**) capaci di accogliere i **clienti** che non possono essere serviti immediatamente.
I clienti che non trovano un servitore libero al loro arrivo si dispongono in modo ordinato, cioè in **coda,** e vengono serviti in accordo a determinate **discipline di servizio.**
Dal punto di vista dinamico una coda è costituita essenzialmente da due processi stocastici: il **processo d'arrivo** dei clienti e il **processo di servizio.**





RAPPRESENTAZIONE SCHEMATICA DI UNA CODA (1 BUFFER, 3 SERVER)



Gli elementi che permettono di definire completamente il fenomeno d’attesa sono quindi:

- la popolazione dei clienti
- il processo d'arrivo
- la coda (in senso stretto)
- i servitori
- il processo di servizio
- la disciplina di servizio.

La **popolazione** è l’insieme dei potenziali clienti, ovvero l’insieme da cui arrivano i clienti e a cui tornano dopo essere stati serviti. Essa può essere finita o infinita. Nel primo caso le modalità di arrivo dei clienti dipendono dal numero di loro correntemente nel sistema. Una tipica situazione in cui si può ritenere che i clienti provengano da una popolazione finita è quando essi devono presentarsi forniti di (o contenuti in) determinate strutture disponibili in numero limitato; ad

esempio in ambiente manifatturiero spesso le parti per essere lavorate devono essere poste su opportuni pallet.
I clienti di una stessa popolazione sono tra loro indistinguibili. Di conseguenza si suppone essi provengano da popolazioni differenti ogniqualvolta in cui debbano essere distinti, ad esempio per livello di priorità o tipo di servizio richiesto.

Il **processo d'arrivo,** che descrive il modo secondo cui i clienti si presentano, è in generale un processo stocastico. Esso è definito in termini della distribuzione dell'**intertempo d'arrivo**, cioè dell'intervallo di tempo che intercorre tra l'arrivo di due clienti successivi.
Per ottenere modelli analiticamente trattabili di solito si assume che sia il processo di arrivo che quello di servizio siano **stazionari**, ovvero che le loro proprietà statistiche non varino nel tempo. Tale assunzione in certi ambiti può essere molto limitativa, infatti l'esperienza comune suggerisce che ad esempio il processo di arrivo dei clienti ad una banca varia durante le ore della giornata.

La **coda** (in senso stretto) è formata dai clienti presenti nel buffer in attesa di essere serviti.
La capacità del buffer può essere infinita o finita. Nel secondo caso essa limita di conseguenza la **capacità del sistema,** cioè il numero dei clienti in attesa nel buffer più quelli che correntemente sono serviti. I clienti che arrivano dopo che sia saturata quest'ultima capacità sono respinti. Ad esempio ha capacità di sistema limitata un centralino telefonico che può tenere in attesa solo un numero finito di chiamate. In assenza di centralino la dimensione della coda è addirittura zero, di conseguenza una chiamata o è servita immediatamente o è rifiutata.

I **servitori** sono in numero noto e costante fissato a livello di progetto. Usualmente essi hanno caratteristiche identiche, possono sempre lavorare in parallelo, viceversa non possono mai rimanere inattivi in presenza di clienti in coda. Anche se vi sono di più servitori in una coda in generale si assume l'esistenza di un unico buffer comune, quando infatti ogni servitore ha il suo buffer separato si preferisce pensare ad un insieme di code. Può però essere comodo introdurre, almeno logicamente, più buffer in presenza di clienti provenienti da popolazioni diverse.

Il **processo dei servizi** descrive il modo secondo cui ciascun servitore eroga il servizio, in particolare definisce la durata dello stesso ed è di solito un processo stocastico. Esso è definito in termini delle distribuzioni dei **tempi di servizio** dei diversi servitori**.** Il processo dei servizi è alimentato dal processo d'arrivo. Conseguentemente il processo d'arrivo è indipendente e condiziona il processo dei servizi. Un cliente, infatti, può essere servito solo se è già arrivato. Quando non c'è nessuno, il servitore è inattivo e quindi non può avvantaggiarsi in vista d'impegni futuri. In altre parole un servitore non può servire in anticipo clienti non ancora arrivati. Non può esistere una coda negativa.

La **disciplina di servizio** specifica quale sarà il prossimo cliente servito fra quelli in attesa al momento in cui si libera un servitore. Le discipline di servizio usualmente considerate, poiché sia molto comuni nella realtà che matematicamente trattabili, sono: servizio in ordine di arrivo FCFS (first-come first-served) o FIFO (first-in first-out), servizio in ordine inverso di arrivo LCFS (last-come first-served) o LIFO (last-in first-out), servizio in ordine casuale SIRO (service in random order), servizio basato su classi di priorità (vedi centri di emergenza quali il pronto soccorso).

2.3 La notazione di Kendall

Tutti gli elementi che definiscono una coda sono evidenziati nella notazione A/B/c/K/m/Z detta di Kendall, dove le lettere rispettivamente indicano:

- A: la distribuzione degli intertempi d'arrivo;
- B: la distribuzione dei tempi di servizio;
- c: il numero di servitori;
- K: la capacità del sistema (default: infinita);
- m: la dimensione della popolazione (default: infinito);
- Z: la disciplina di servizio (default: FCFS);

In particolare ad A e B possono essere sostituite le seguenti lettere:
M : distribuzione esponenziale (Markoviana)
D : distribuzione costante (Degenere)
E_k : distribuzione di Erlang di ordine k
G : distribuzione generica
GI : distribuzione generica di eventi indipendenti (per gli arrivi)

Ad esempio M/M/1 sta per M/M/1/∞/∞/FCFS coda con processo degli arrivi e dei servizi markoviani, con un servitore, con capacità del sistema (e quindi del buffer) infinita e con arrivi provenienti da una popolazione infinita che vengono serviti su base FCFS.

2.4 Lo stato di una coda

Lo stato di un sistema dinamico in un dato istante temporale rappresenta l'insieme informativo minimo che permette di conoscere l'evoluzione futura del sistema stesso, una volta note le realizzazioni dei fenomeni stocastici cui è soggetto.
Lo stato di una coda è dato dal numero di clienti presenti nel sistema, dal tempo trascorso dall'ultimo arrivo dell'ultimo cliente, infine, per ogni servitore, da un valore binario indicante se sta correntemente fornendo un servizio ed in questo ultimo caso anche dal tempo trascorso dall'inizio del servizio.

2.5 Domande ed esercizi

1. Individuare le varie componenti di un sistema coda negli esempi riportati nel Capitolo 1.

2. Elencare alcuni casi reali di file di attesa in cui la disciplina di servizio è differente da FCFS, evidenziare i motivi che giustificano le scelte alternative. [*Pronto soccorso, gestione ordini di produzione, gestione tasks in sistemi operativi*]

3. Si discutano qualitativamente le prestazioni ottenibili dalle seguenti possibili organizzazioni delle file di attesa in una banca:
   - un buffer per ogni sportello con sportelli specializzati secondo operazioni, ma con alcuni sportelli uguali;
   - un buffer per ogni sportello con sportelli universali;
   - buffer unico su sportelli universali.

[*il secondo caso è dal punto di vista del cliente il peggiore poiché la varianza dei tempi di servizio dei clienti davanti a lui è massima. Il terzo caso è socialmente ottimo. Il primo caso minimizza la varianza dei tempi di servizio*....]

4. Discutere in che cosa si distingue una coda in una banca da una coda in un supermercato. *[quale sistema ha un evoluzione più prevedibile?*]

5. Volendo impedire che si formino code eccessivamente lunghe si può intervenire sulla struttura fisica del sistema (e.g., aggiungendo servitori) oppure sui processi stocastici che lo descrivono (i.e., arrivo clienti, velocità servizi). Discutere come, in base a queste considerazioni, le società autostradali possono intervenire per ridurre ingorghi.

6. Determinare le caratteristiche delle seguenti code e definire i possibili campi di applicazione: M/M/1, M/M/s, M/M/1/K, M/M/1/N/N, M/D/1.

3 Le problematiche di interesse

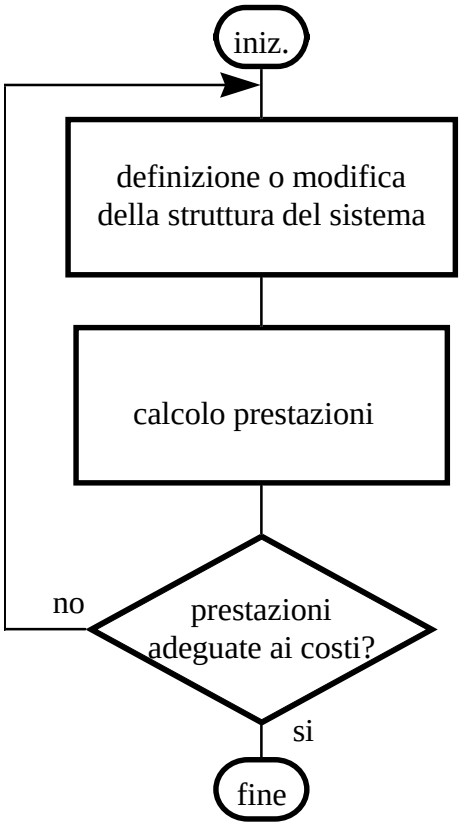
3.1 Le problematiche di interesse
3.2 Il caso deterministico
3.3 Domande ed esercizi

3.1 Le problematiche d'interesse

Qualunque sia il sistema fisico considerato le problematiche di interesse generalmente riguardano i **costi** (o i **profitti**) di tipo economico coinvolti. I costi sono di solito suddivisi tra **variabili,** ovvero funzione di almeno una delle grandezze che caratterizzano la dinamica del sistema, e **fissi**, ovvero indipendenti dalla dinamica osservata e generalmente funzione della sola struttura fisica del sistema. In una coda si possono ritenere sempre presenti almeno i costi variabili legati al tempo d'attesa dei clienti e i costi fissi legati al numero dei servitori disponibili. I differenti attori coinvolti nel sistema ovviamente considerano questi costi con enfasi diversa. I clienti ritengono fondamentale la riduzione dei tempi d'attesa, mentre il gestore del sistema è probabilmente interessato al massimo sfruttamento delle risorse (servitori) pur cercando di rispettare le esigenze dei clienti. In questo contesto la Teoria delle Code individua alcuni indici di prestazione direttamente legati ai costi che, quando valgono alcune ipotesi, sono facilmente calcolabili:

- L_s: numero medio di clienti nel sistema (sia in attesa di servizio e che riceventi servizio);
- L_q: numero medio di clienti in attesa di servizio;
- W_s: tempo di attesa medio dei clienti nel sistema (sia in attesa di servizio e che riceventi servizio);
- W_q: tempo d'attesa medio dei clienti prima di essere serviti;
- p_n: probabilità che vi siano a regime n clienti nel sistema;
- ρ: fattore di utilizzazione dei servitori (rapporto tra tempo impiegato in servizio e tempo disponibile complessivo).

I valori che sono assunti dagli indici sopraenunciati dipendono parametricamente dalla struttura della coda e dal tasso di arrivo dei clienti. Il progettista di sistema dovrebbe quindi essere capace di determinare le caratteristiche della coda, ad esempio il numero di servitori e la loro velocità di servizio, in modo da soddisfare le specifiche. In particolare, una volta che siano fissati dei valori (o degli intervalli) per gli indici prestazione e che sia noto il tasso di arrivo dei clienti, il progettista deve minimizzare i costi di realizzazione
Per la difficoltà dei calcoli coinvolti però la Teoria delle Code in generale fornisce solo **modelli descrittivi**, ovvero modelli che permettono di valutare le prestazioni del sistema a fronte d'ipotesi sulla sua struttura, ma che non risolvono direttamente problemi di progetto come invece fanno invece i **modelli normativi**. Di conseguenza il progetto di una coda di solito avviene per tentativi (vedi diagramma di flusso). Sono formulate delle ipotesi sulla struttura del sistema, si valutano le prestazioni corrispondenti, se le prestazioni sono adeguate ai costi la struttura è accettata altrimenti si torna a modificare la struttura e si itera.



La fase di calcolo delle prestazioni avviene attraverso l'utilizzo di formule matematiche chiuse quando esse sono note oppure, per sistemi particolarmente complessi che includano ad esempio più code, attraverso la realizzazione di esperimenti simulativi o l'utilizzo di metodi approssimati. In ogni caso il problema consiste nel capire in che direzione devono orientarsi le modifiche al sistema di tentativo. Solo in questo modo all'iterazione successiva si può definire una struttura migliore.

Alcune considerazioni qualitative possono comunque essere fatte circa le relazioni esistenti tra gli indici di prestazione indicati, i parametri del sistema e gli intertempi d'arrivo dei clienti. Si consideri ad esempio un sistema in cui vi sia un unico buffer infinito, i clienti siano serviti su base FCFS e ogni servitore soddisfi una richiesta alla volta, con una velocità di servizio indipendente sia dalla presenza degli altri servitori che dal numero dei clienti. In questo caso L_s, L_q, W_s e W_q crescono/decrescono al diminuire/aumentare/ degli intertempi d'arrivo dei clienti, al diminuire/aumentare del numero o della velocità dei servitori. Allo stesso modo si comporta il numero medio di clienti serviti da ogni servitore, ammesso che il sistema riesce a mantenersi **stabile**, ovvero che il tempo di attesa dei clienti non cresce all'infinito. In questo contesto si osservi che un eventuale servitore "pigro" diminuirebbe la propria velocità di servizio per apparire molto impegnato. In questo modo però egli danneggerebbe i clienti che dovrebbero attendere più a lungo. Per questo motivo un eventuale incentivo ai servitori non dovrebbe essere basato sul solo loro fattore di utilizzo, bensì, a parità di altre condizioni, sul numero medio di clienti serviti nell'unità di tempo.

3.2 Il caso deterministico D/D/1

Si possono facilmente determinare le prestazioni del sistema quando gli istanti d'arrivo dei clienti ed i tempi di espletamento dei servizi richiesti sono noti a priori senza incertezza.

Se la disciplina di servizio è FCFS, per ogni cliente, l'istante di uscita dal sistema è dato dalla somma del suo tempo di servizio e del massimo tra il suo istante d'arrivo e l'istante di uscita del cliente precedente.

Siano dati i valori:
- a(i): istante d'arrivo del cliente i
- s(i): durata del servizio del cliente i

e le variabili:
- x(i): istante d'uscita dal sistema del cliente i
- w(i): tempo d'attesa del cliente i
- n(t): numero di persone nel sistema all'istante t

Posto x(0) = 0, si ottiene

$$x(i) = s(i) + \max\{x(i-1), a(i)\} \quad i = 1,2,3,...$$

e

$$w(i) = x(i) - s(i) - a(i) \quad i = 1,2,3,...$$

Quindi, per calcolare il numero di clienti nel sistema all'istante t, basta contare il numero di valori di clienti i per cui $a(i) \leq t < x(i)$, dal momento che un cliente è nel sistema nell'istante in cui entra, non vi è più nell'istante in cui esce.

Il caso totalmente deterministico è però difficile che occorra nella realtà. In genere gli arrivi dei clienti e la durata dei servizi sono affetti da incertezza, quindi sono modellati come **processi stocastici**, come sarà descritto nei capitoli successivi.

3.3 Domande ed esercizi

1. Individuare i costi fissi e i costi variabili in sistemi che contengano file di attesa. Valutare economicamente ognuno di tali costi. [*es., quanto si sarebbe disposti a pagare pur di non dovere mettersi in coda? quanto costa un nuovo impiegato?*]

2. Individuare i costi fissi e i costi variabili e le loro interdipendenze nei sistemi indicati al Capitolo 1.

3. Fissare gli appuntamenti dei clienti di un medico in modo che ogni paziente arrivi dieci minuti del suo turno. Si supponga di sapere con buona approssimazione quanto il dottore impiega a visitare ognuno dei pazienti e che essi debbano essere visitati in ordine di prenotazione.

Paziente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Tempo visita	15	10	25	15	20	10	15	30	10	5	15	25	30	30	10	20

4 Il ruolo delle distribuzioni esponenziale e di Poisson

4.1 La distribuzione esponenziale
4.2 Il processo di Poisson
4.3 Domande ed esercizi

4.1 La distribuzione esponenziale

Nei casi pratici si possono trovare code con intertempi d'arrivo dei clienti e tempi di servizio soggetti a distribuzioni probabilistiche di quasi qualunque tipo. Tra le tante, la distribuzione esponenziale è forse quella che trova maggiore applicazione e che inoltre presenta migliore trattabilità dal punto di vista matematico.

Una variabile aleatoria (v.a.) X ha **distribuzione esponenziale** con parametro $\lambda > 0$ quando la sua densità p(x) è:

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

I tempi intercorrenti tra due eventi successivi relativi allo stesso processo (e.g., arrivo di clienti oppure inizio e fine di un servizio) possono essere modellati come una v.a. esponenziale se soddisfano le seguenti condizioni:

- la probabilità un evento occorra in un intervallo di tempo infinitesimo dx è proporzionale a dx, con λ come costante di proporzionalità, ovvero
$$P(x < X \leq x + dx) = \lambda dx$$
- la probabilità di avere più di un evento in un intervallo di tempo infinitesimo dx è nulla;
- la probabilità che il prossimo evento ritardi oltre un dato limite non dipende da quanto tempo si è verificato l'evento precedente.

Il processo deve avere quindi memoria (proprietà markoviana), ovvero

$$P(X > x+u; X > u) = P(X > x)$$

che implica

$$P(X > x+u) = P(X > x)P(X > u)\,.$$

Solo una v.a. esponenziale soddisfa tali condizioni, infatti

$$\begin{aligned} P(X \leq x+dx) &= 1 - P(X > x+dx) = 1 - P(X > x)P(X > dx) = \\ &= 1 - P(X > x)(1 - P(X \leq dx)) = 1 - P(X > x)(1 - \lambda dx) = \\ &= P(X \leq x) + (1 - P(X \leq x))\lambda dx \end{aligned}$$

allora, posto $P(x) = P(X \leq x)$, si giunge a

$$dP(x) = P(x) + (1 - P(x))\lambda dx - P(x) = (1 - P(x))\lambda dx$$

che è un'equazione differenziale la cui unica soluzione che soddisfa la condizione

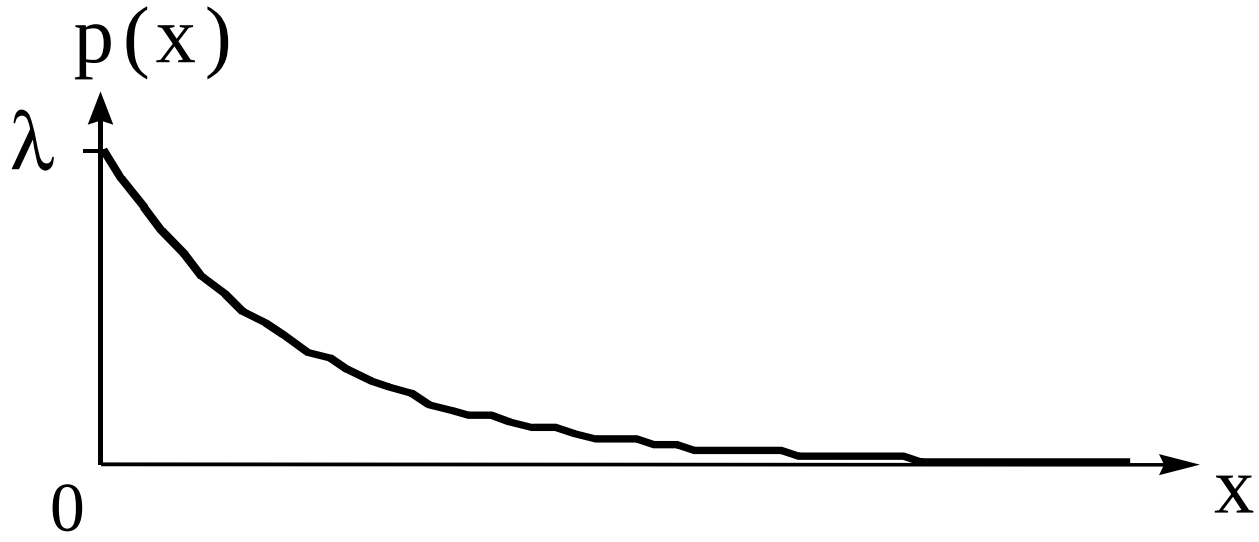
$$\lim_{x\to\infty} P(x) = 1$$

risulta essere

$$P(x) = \begin{cases} 1-e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad \text{ovvero} \qquad p(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

La speranza matematica della variabile aleatoria X con distribuzione esponenziale è $E\{X\}=1/\lambda$ e la sua varianza è $1/\lambda^2$. Il parametro λ è l'inverso del valore atteso del tempo che intercorre tra l'arrivo di due clienti successivi e può essere interpretato come il tasso medio d'arrivo di clienti per unità di tempo.

LA DISTRIBUZIONE ESPONENZIALE



La distribuzione esponenziale è una funzione strettamente decrescente, quindi i valori più piccoli sono più probabili. Si hanno spesso valori inferiori alla media e qualche volta valori molto superiori.

La mancanza di memoria della distribuzione esponenziale rende la stessa ragionevole per modellare gli intertempi d'arrivo che non siano correlati, cioè tali per cui l'arrivo di un cliente non favorisca o sfavorisca altri arrivi. La stessa proprietà giustifica l'uso della distribuzione in presenza di tempi di servizio che riguardino prestazioni poco omogenee, ad esempio i servizi di pronto soccorso o la durata conversazioni e la lunghezza di messaggi. Viceversa la distribuzione esponenziale non deve essere usata per modellare produzioni industriali identiche, a meno che non si considerino come clienti gli ordini da eseguire e solo nel caso in cui questi possano avere dimensione variabile.

Collegata a questa proprietà di mancanza di memoria è il cosiddetto il **paradosso del tempo di servizio** residuo. Se T è il tempo medio di servizio, un nuovo cliente che arrivi in modo completamente casuale, quando il servitore è occupato, deve comunque aspettare in media un

tempo T (non T/2, come si sarebbe tentati di pensare) prima che il servitore termini il servizio in corso.

Le variabili aleatorie esponenziali godono, infine, di un'ulteriore proprietà:

Proprietà
Il minimo di variabili aleatorie esponenziali indipendenti è ancora una variabile aleatoria esponenziale.

Se si considerano eventi di tipo diverso, ciascuno con intertempo di occorrenza esponenziale: allora l'intertempo tra eventi di tipo qualsiasi è ancora esponenziale, con parametro pari alla somma dei parametri.

4.2 Il processo di Poisson

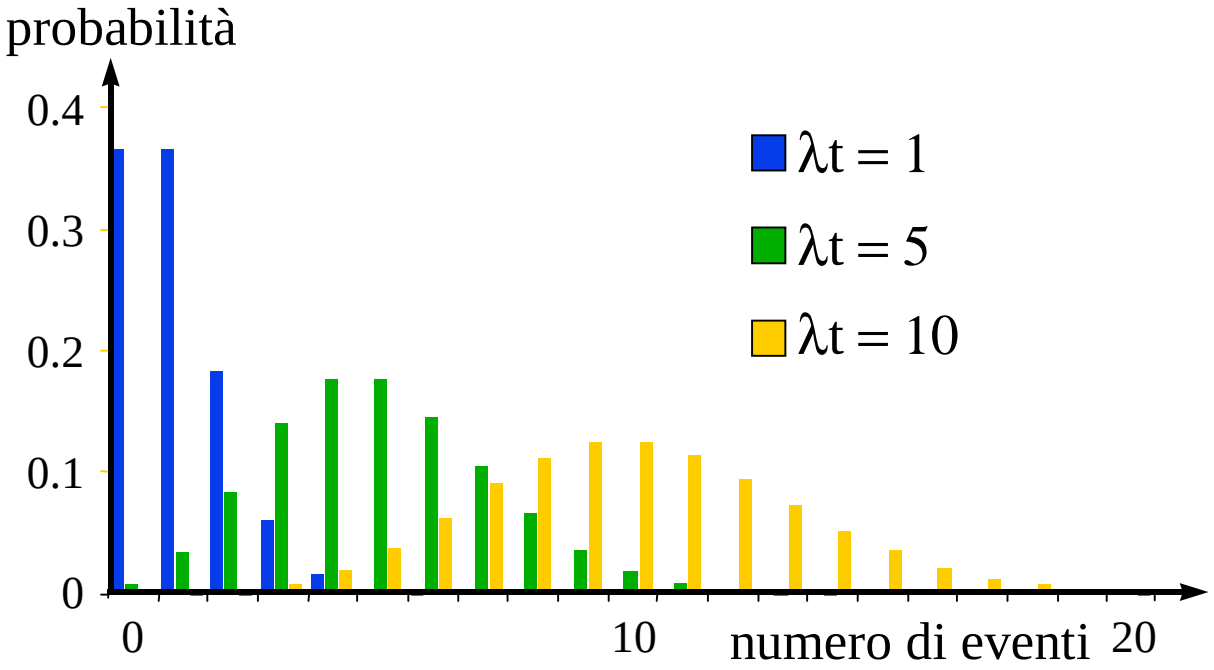
Quando gli intertempi sono esponenziali il numero di eventi N(t) che si verifica in un dato tempo t è un processo di Poisson:

$$P\{N(t) = n\} = [\ (\lambda t)^n e^{-\lambda t}\ ] / n!.$$

Il processo di Poisson N(t) ha valore atteso $E\{N(t)\} = \lambda t$, dove λ esprime il numero medio di eventi nell'unità di tempo, cioè la frequenza media.

La distribuzione di Poisson assume al variare del valore λt le forme presenti nella figura successiva

LA DISTRIBUZIONE DI POISSON



Ai processi di Poisson si generalizzano le proprietà delle v.a. esponenziali. In particolare:

- λ dt rappresenta la probabilità di occorrenza di un evento in un intervallo di tempo infinitesimo dt;
- se si hanno eventi di tipo diverso i, i=1,2,...,n, e il loro processo di accadimento globale è poissoniano con parametro λ, se inoltre ogni evento ha probabilità fissa p_i di essere di tipo i ($\Sigma_i p_i=1$), allora ciascun tipo di eventi i è di per sé poissoniano, con parametro $\lambda_i = \lambda p_i$.

4.3 Domande ed esercizi

1. Dire se può essere modellato con una v.a. esponenziale l'intertempo di arrivo di clienti ha media 2 minuti e deviazione standard di 4 minuti.

2. Dire se può essere modellato come un processo di Poisson l'arrivo degli ospiti ad una festa.

3. Dire se può essere modellato come un processo di Poisson il numero di particelle emesse da un materiale radioattivo.

4. Dall'osservazione empirica dei tempi di lavorazione da parte di una cella di lavorazione si sono tratti i seguenti dati sperimentali:

n. osservazioni	37	21	20	9	6	2	3	1	0	1
interv. tempi servizio	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	9-10	>10

   Dire se i tempi di lavorazione possono essere ragionevolmente modellati con una v.a. esponenziale. In caso positivo indicare il valore del parametro λ. [*usare il test del* χ^2]

5. Ripetere l'esercizio precedente con la seguente serie di dati

n. osservazioni	5	2	6	2	1	3	0	1	0	0
interv. tempi servizio	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	9-10	>10

6. I dati in entrambi gli esercizi precedenti sono stati ottenuti da v.a. esponenziali. Commentare le differenze nei risultati e dedurne delle conseguenze pratiche.

7. I clienti di una coda arrivano seguendo un processo di Poisson. Se in media arrivano 3 clienti ogni 10 minuti, determinare dopo quanto tempo la probabilità che siano arrivati almeno 6 clienti è maggiore di 0.5, oppure di 0.95, oppure di 0.99.

5 Il processo nascite - morti

5.1 Il processo nascite - morti
5.2 Domande ed esercizi

5.1 il processo nascite - morti

Il **processo nascite-morti** è un processo stocastico utile per studiare le code. Esso rappresenta il numero di elementi N(t) di una popolazione che può aumentare, per effetto di una nascita, o diminuire, per effetto di una morte, di un'unità alla volta.

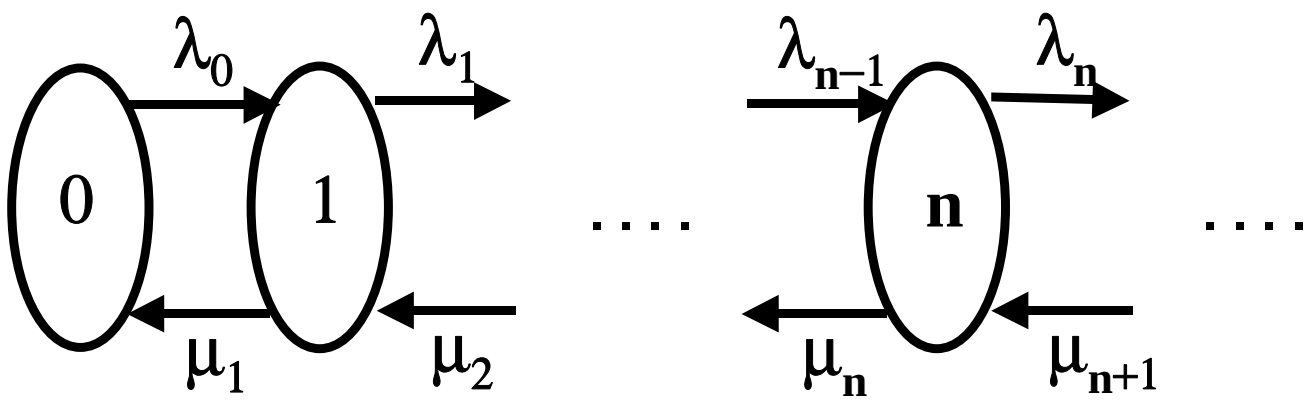
In modo formale il processo nascite-morti assume che, ad ogni generico istante t, possa avvenire un solo evento (di nascita o di morte); inoltre che, data una popolazione di numerosità N(t)=n, l'intervallo di tempo fino alla prossima nascita sia una v.a. esponenziale con parametro λ_n, mentre l'intervallo di tempo fino alla prossima morte sia una v.a. esponenziale con parametro μ_n. In questo contesto i parametri λ_n e μ_n possono essere interpretati come rispettivamente il tasso medio di nascita e di morte quando di una popolazione è composta di n individui.

Un particolare processo nascite-morti, dove avvengono solo nascite, è quello di Poisson. In questo caso $\lambda_n=\lambda$ e $\mu_n=0$, n=0,1,2,…. Con un processo nascite-morti si può descrivere anche il numero di clienti in una coda. In questo caso valgono le seguenti relazioni:

arrivo = nascita
uscita = morte

Il processo nascite-morti è usualmente rappresentato in modo grafico come indicato nella figura seguente. Gli ovali rappresentano lo stato (ovvero solo la numerosità della popolazione dato che le v.a. esponenziali sono senza memoria); i coefficienti associati alle frecce esprimono invece il tasso di probabilità di transizione da uno stato all'altro.

PROCESSO NASCITE MORTI



Per trovare il valore $p_n(t)$ della probabilità che al tempo t il processo nascite-morti si trovi nello stato n, ovvero la probabilità che al tempo t siano in vita n persone, si può fare ricorso alla soluzione di un sistema di equazioni differenziali. Infatti la probabilità $p_n(t + dt)$ che al tempo t+dt ci siano n persone è data dalla somma dei seguenti termini:

- la probabilità $p_n(t)$ che in t ci siano n persone per la probabilità $(1 - \lambda_n - \mu_n)\, dt$ che nell'intervallo di tempo tra t e t+dt non sia avvenuta nè una nascita nè una morte;
- la probabilità $p_{n-1}(t)$ che in t ci siano n-1 persone per la probabilità $\lambda_{n-1} dt$ che nell'intervallo di tempo tra t e t+dt sia avvenuta una nascita;
- la probabilità $p_{n+1}(t)$ che in t ci siano n+1 persone per la probabilità $\mu_{n+1} dt$ che nell'intervallo di tempo tra t e t+dt sia avvenuta una morte.

Si ottiene quindi il seguente sistema di equazioni differenziali:

$$p_0(t + dt) = p_0(t)(1 - \lambda_0)\, dt + p_1(t)\, \mu_1\, dt$$

$$p_n(t + dt) = p_n(t)(1 - \lambda_n - \mu_n)\, dt + p_{n-1}(t)\, \lambda_{n-1}\, dt + p_{n+1}(t)\, \mu_{n-1}\, dt \quad n=1,2,....$$

Per t che tende all'infinito, se il tasso delle morti complessivamente supera il tasso delle nascite, il processo diventa stazionario, ovvero le sue proprietà statistiche non variano più nel tempo e quindi $p_n(t)=p_n$, per ogni tempo t. In quest'ipotesi il sistema di equazioni differenziali diventa un sistema di equazioni lineari omogeneo con soluzione:

$$p_n = C_n\, p_0 \qquad n = 1,2,...$$

dove

$$C_n = (\lambda_{n-1}\, \lambda_{n-2} ... \lambda_0) / (\mu_n\, \mu_{n-1} ... \mu_1).$$

Osservando in fine che i termini p_n rappresentano delle probabilità e che quindi

$$\Sigma_n\, p_n = 1$$

si ottiene che

$$p_0 = 1 / (1 + \Sigma_n\, C_n).$$

Da questi risultati si possono ricavare le distribuzioni di probabilità di tutte le code poissoniane (M/M/…)

5.2 Domande ed esercizi

1. Elencare alcuni processi nascite-morti ed determinare i valori dei parametri λ_n, μ_n.

2. Le code M/M/… possono essere rappresentate come processi nascite-morti. Determinare i valori dei parametri λ_n, μ_n nel caso M/M/1, M/M/s, M/M/1/K, M/M/1/N/N, M/M/s/N/N.

3. Calcolare i valori di p_0 e di p_n nei seguenti processi nascite morti:
   - $\lambda_n=1$, $\mu_n=n$;

- $\lambda_0=0,\ \mu_1=0, \lambda_n=1,\ \mu_{n+1}=2$, per $n>0$.
  [*considerare come stato 0 la situazione i cui è presente un solo individuo*]

4. Determinare lo stato più probabile e il numero medio di individui in vita nel caso dei seguenti processi di nascita-morte
   - $\lambda_n=1,\ \mu_n=1.1$;
   - $\lambda_n=1$, $\mu_{n+1}=1/(n+1)$, per $n=1,...,5$; $\lambda_n=0$, $\mu_{n+1}=0$, per $n>5$.

6 La coda M/M/1

6.1 La coda M / M / 1
6.2 La formula di Little
6.3 L'influenza del fattore di utilizzazione
6.4 L'intertempo tra due partenze
6.5 Domande ed esercizi

6.1 La coda M / M / 1

Una coda M/M/1 è fisicamente composta da un buffer e da un solo servitore; in essa l'intertempo tra due arrivi successivi e il tempo di servizio sono due variabili aleatorie markoviane, cioè con distribuzione esponenziale. Il tasso medio di interarrivo e il tasso medio di servizio sono usualmente indicati con λ e μ. La coda M/M/1 può essere considerata un processo nascite - morti con

$$\lambda_n = \lambda \quad \mu_n = \mu.$$

L'arrivo di un nuovo cliente in coda può, infatti, essere interpretato come una nascita; viceversa la fine di un servizio, quindi l'uscita di un cliente dal sistema, come una morte.
Di conseguenza

$$p_n = \rho^n p_0 \qquad n = 1,2,...$$

dove il **fattore di utilizzazione** $\rho = (\lambda / \mu)$ esprime il rapporto tra il tempo medio di servizio e il tempo medio tra due arrivi.
Dato che vale la seguente condizione

$$p_0 = 1 / (1 + \sum_n C_n) = 1 / (1 + \sum_n \rho^n),$$

non dovrebbe stupire che p_0 esiste se e solo se $\rho < 1$, ovvero se in media il sistema ha la potenzialità a servire i clienti più velocemente di quanto essi arrivino. La condizione $\rho < 1$ è detta di **stabilità**. Infatti lo stato stazionario non può essere raggiunto e la coda cresce all'infinito qualora essa non occorra.

Per $\rho < 1$ si verifica che $1 + \sum_n \rho^n = 1 + \rho/1- \rho$, e di conseguenza

$$p_0 = 1- \rho \qquad e \qquad p_n = \rho^n (1- \rho).$$

Dalle condizioni precedenti si può esprimere $\rho = 1-p_0$, quindi ρ può essere interpretato anche come il tasso di occupazione del servitore, ovvero la frazione di tempo in cui il servitore lavora, ovvero la probabilità che ci sia almeno un cliente nel sistema oppure, infine, come il numero medio di ingressi durante un servizio.
Una volta note le probabilità p_n possono essere calcolati i valori delle altre grandezze d'interesse. In particolare il numero medio di clienti nel sistema è

$$L_s = E \{ n \} = \sum_n n p_n = \rho / (1 - \rho),$$

con varianza

$$\sigma^2_{L_s} = E\{(n - L_s)^2\} = \rho/(1-\rho)^2$$

Il numero medio di clienti in attesa è invece

$$L_q = L_s - [\text{n. medio di clienti correntemente serviti}] = L_s - \rho = \rho^2/(1-\rho);$$

dove L_q può anche essere dedotto nel seguente modo

$$L_q = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n = L_s - (1-p_0) = L_s - \rho$$

6.2 Formula di Little

Se una coda è stabile, qualunque essa sia, in media devono uscire dal sistema tanti clienti quanti entrano. Per una coda M/M/1 il tasso d'uscita dal sistema è quindi λ e non μ. Si può dedurre di conseguenza che il tempo media di attesa dei clienti nel sistema è

$$W_S = L_S / \lambda,$$

Questa formula, detta **formula di Little,** vale, come già evidenziato, per qualunque sistema in equilibrio e si enuncia affermando che: il numero medio di elementi presenti nel sistema è eguale al tempo medio di permanenza nel sistema per il tasso d'ingresso.

Applicando la formula di Little al caso M/M/1 si ottiene che il tempo d'attesa nel sistema è:

$$W_S = 1/(\mu - \lambda)$$

e che tempo medio d'attesa in coda è:

$$W_q = W_S - (1/\mu) = \lambda/(\mu(\mu - \lambda)).$$

La **formula di Little** può essere **generalizzata** in modo da considerare i momenti del secondo ordine del tempo d'attesa nel sistema e del numero medio di clienti

$$E\{L_s^2 - L_s\} = \lambda^2 E\{W_s^2\}$$

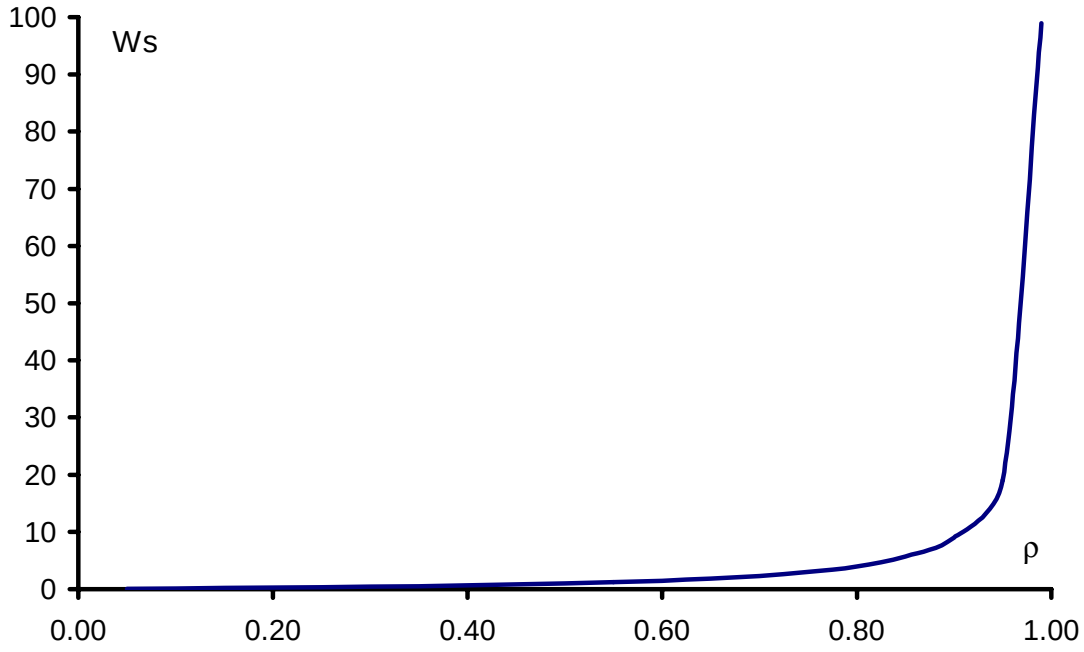
Più in generale, per i momenti di ordine k, essa diventa:

$$E\{L_s(L_s-1)(L_s-2)\cdots(L_s-k+1)\} = \lambda^k E\{W_s^k\}.$$

6.3 L'influenza del fattore di utilizzazione

Nei paragrafi precedenti si è osservato che ci deve essere una probabilità $p_0 = 1- \rho$ non nulla che il servitore sia inattivo per assicurare la stabilità del sistema. In particolare al crescere di ρ aumenta l'occupazione del servitore e quindi la permanenza media e il numero medio dei clienti nel sistema, nonché il numero medio e il tempo medio dei clienti in attesa.

TEMPO MEDIO DI ATTESA NEL SISTEMA AL VARIARE DEL FATTORE DI UTILIZZAZIONE



Un incremento del valore di ρ non introduce però solo svantaggi. Se ρ aumenta a causa di un maggiore arrivo di clienti, si ha corrispondentemente una maggiore utilizzazione delle risorse disponibili. Viceversa, se l'aumento di ρ è dovuto all'utilizzo di servitori meno veloci, dovrebbero diminuire conseguentemente i costi di acquisizione degli stessi. In un problema di progetto si deve quindi fissare ρ cercando un giusto compromesso tra costi, prestazioni (qualità) e utilizzazione delle risorse.

Un altro fattore molto importante che è influenzato da ρ è il tempo di **raggiungimento del regime**, ovvero il tempo dopo il quale le statistiche che descrivono il comportamento medio del sistema non variano più in modo significativo e quindi il processo che descrive la dinamica della coda può essere ritenuto praticamente stazionario. In questo contesto si ricordi che le formule presentate in questo capitolo ed in quelli precedenti si basano sull'assunzione che il sistema abbia raggiunto il regime. In particolare con $\rho \leq 0.7$ il regime è raggiunto dopo meno di un migliaio di clienti, con $\rho \leq 0.85$ il regime è raggiunto dopo una decina di migliaia di clienti, con $\rho > 0.95$ il regime è raggiunto dopo diversi milioni di clienti.
Alla luce delle considerazioni precedenti è legittimo chiedersi se, per tassi di utilizzazione vicini all'unità, i risultati ottenuti abbiano interesse pratico. Infatti pochi sistemi mantengono caratteristiche costanti per tempi così lunghi sia per quanto riguarda l'arrivo dei clienti che il servizio agli stessi. Si deve notare però che, se al tempo iniziale la coda era vuota, le grandezze

calcolate in una situazione di regime forniscono in generale dei limiti superiori per i valori che saranno assunti dalle stesse grandezze nella fase di **transitorio**.

6.4 L'intertempo tra due partenze

Le code M/M/1 godono di un'altra interessante caratteristica. Gli intertempi tra la fine di servizi successivi possono essere descritti come v.a. esponenziali con parametro λ coincidente in valore con quello del processo degli arrivi.
Questa caratteristica peculiare permette di studiare facilmente catene di code M/M/1, ma anche reti più complesse (dette **reti di Jackson**). Nelle reti di Jackson ad ogni coda è associato un insieme di probabilità tempo invarianti, una per ogni altra coda del sistema e una per l'universo esterno. In base a tali probabilità ogni cliente, una volta terminato il servizio in una coda, è indirizzato o fuori dal sistema o verso un'altra coda. Si dimostra che in questo modo la generica coda i-ma osserva un processo d'arrivo di clienti poissoniano di parametro

$$\lambda_i = \lambda_{0i} + \sum_j p_{ji} \lambda_j$$

dove λ_{0i} è il tasso di arrivi alla coda dei clienti che provengono dall'esterno del sistema, mentre p_{ji} è la probabilità che un cliente in uscita dalla coda j-ma sia indirizzato verso la coda i-ma, infine λ_j è il tasso complessivo di arrivo dei clienti alla coda j-ma.
Le proprietà del processo di uscita dei clienti di una coda M/M/1 possono essere anche utilizzate per verificare la correttezza di software di simulazione per sistemi di code. Infatti, se i numeri casuali che devono simulare gli intertempi di arrivo dei clienti e i tempi di servizio non sono generati correttamente, si osserva quasi sempre che il processo di uscita da una coda M/M/1 non è poissoniano.

6.5 Domande ed esercizi

1. Verificare se è vero che $W_q = L_q / \mu$

2. Determinare di quanto aumenta il tempo medio di permanenza nel sistema se il tasso d'arrivo aumenta del 10%:
   con $\rho = 0{,}5$;
   con $\rho = 0{,}7$;
   con $\rho = 0{,}9$.
   Osservare come il fattore di utilizzazione ρ esprime anche la ''stabilità'' o ''robustezza'' del sistema, cioè la sua capacità di sopportare variazioni dei tempi di arrivo (e di servizio)

3. Determinare se e di quanto funziona meglio un sistema in cui tutti i clienti vengono serviti da un unico servitore veloce oppure un sistema in cui i clienti sono ripartiti tra n code con servitori n volte più lenti. [*il numero di utenti di ogni coda è uguale a quello che si ottiene con una coda unica, quindi ...*]

7 Altre code possoniane

7.1 Altre code possoniane
7.2 M/M/s
7.3 M/M/1/K
7.4 M/M/1/N/N
7.5 Domande ed esercizi

7.1 Altre code poissoniane (M / M / …)

Il processo nascite-morti permette di studiare anche altre code poissoniane. In tutti i casi l'arrivo di un cliente può essere considerato come una nascita e il completamento di un servizio, e quindi l'abbandono del sistema da parte di un cliente, come una morte.
Date le relazioni

$$p_n = C_n p_0 \qquad n = 1,2,...$$

con

$$p_0 = 1 / (1 + \Sigma_n C_n),$$
$$C_n = (\lambda_{n-1} \lambda_{n-2} ... \lambda_0) / (\mu_n \mu_{n-1} ... \mu_1)$$

caso per caso, si determinano i parametri λ_n e μ_n, quindi si valuta C_n e e le altre grandezze.

7.2 M / M / s

La coda M/M/s ha s servitori in parallelo ciascuno con tasso di servizio $1/\mu$. Di conseguenza

$$\lambda_n = \lambda,$$

e, per le proprietà dell'esponenziale,

$$\mu_n = n \mu \qquad \text{per} \qquad 1 \le n \le s$$
$$\mu_n = s \mu \qquad \text{per} \qquad n > s$$

Il fattore di utilizzazione vale $\rho = \lambda / (s \mu)$ e quindi

$$p_n = \begin{cases} p_0 \dfrac{(s\rho)^n}{n!} & 1 \le n < s \\ p_0 \dfrac{s^s \rho^n}{n!} & n \ge s \end{cases} \qquad p_0 = 1 / \left[\sum_{k=0}^{s-1} \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!(1-\rho)} \right]$$

Se la condizione di stabilità $\rho < 1$ è rispettata si ottiene

$$L_q = \frac{(s\rho)^{c+1}}{s^2 (s-1)! (1-\rho)^2} p_0$$

e quindi si possono calcolare

$$W_q = L_q / \lambda$$
$$W_s = W_q + 1 / \mu$$
$$L_s = L_q + \lambda / \mu.$$

Un caso estremo di coda M/M/s è quello in cui vi sono infiniti servitori M/M/∞. Tale situazione si verifica nei self-service in cui ogni cliente serve se stesso.
Si può verificare che

$$L_s = \lambda / \mu.$$
$$W_s = 1 / \mu.$$
$$W_q = L_q = 0.$$

7.3 M / M / 1 / K

La coda M/M/1/K ha capacità finita K, ovvero nel sistema non possono essere presenti più di K clienti, quindi

$$\lambda_n = \lambda \quad \text{per} \quad 0 \le n < K$$
$$\lambda_n = 0 \quad \text{per} \quad n \ge K$$
$$\mu_n = \mu \quad \text{per} \quad 1 \le n \le K$$
$$\mu_n = 0 \quad \text{per} \quad n > K.$$

La coda M/M/1/K è sempre stabile per definizione.
Applicando le solite formule dei processi nascite-morti si ottiene

$$L_s = \rho / (1 - \rho) - (K + 1) \rho^{K+1} / (1 - \rho^{K+1})$$

da cui

$$L_q = L - (1 - p_0)$$
$$W_s = L / \lambda'$$
$$W_q = L_q / \lambda'$$

con $\lambda' = \lambda (1 - p_K)$ dove λ' è detto tasso d'ingresso.

7.4 M / M / 1 / N / N

La coda M/M/1/N/N ha capacità finita N, ma anche popolazione finita N, quindi il tasso di arrivo per ciascun cliente è

$$\lambda_n = (N - n) \lambda \qquad \text{per} \qquad 0 \le n \le N$$
$$\lambda_n = 0 \qquad \text{per} \qquad n \ge K$$
$$\mu_n = \mu$$

Anche questa coda è sempre stabile, per essa si ricava

$$L_s = N - (1 - p_0) \mu / \lambda$$

da cui

$$L_q = L - (1 - p_0)$$
$$W_s = L / \lambda'$$
$$W_q = L_q / \lambda'$$

con $\lambda' = \lambda (N - L)$.

7.5 Domande ed esercizi

1. Discutere se coda M/M/∞ può modellare una stazione di benzina con servizio self-service. Se non è possibile suggerire un modello alternativo.

2. Discutere se coda M/M/∞ può modellare il metodo di acquisto in uso nei supermercati.

3. Determinare se e di quanto funziona meglio un sistema in cui tutti i clienti vengono serviti da un unico servitore veloce oppure un sistema in cui i clienti sono ripartiti tra 2 servitori 2 volte più lenti.

4. Dati s servitori determinare se conviene organizzare i clienti in un'unica coda, o in s code distinte.
   In particolare valutare numerica la differenza dei tempi di attesa per s=2 e ρ=0.8.
   [*il numero medio di persone nel sistema è sempre maggiore nel caso centralizzato. il tempo di permananza nel sistema è sempre inferiore nel caso centralizzato*].

5. Una cella di lavorazione all'interno di un impianto produttivo ha un buffer finito che contiene al massimo n parti in attesa di lavorazione. Determinare il tasso di inattività della cella al variare del potenziale tasso di interarrivo λ delle parti. Assumere due possibilità: che la popolazione dei pezzi sia finita e uguale a n, che la popolazione dei pezzi sia infinita.
   Supponendo che per la cella si debba ammortare il valore di 100 Mlit. l'anno, determinare al variare di λ e di n il costo della non completa saturazione della cella.

8 Alcune code non poissoniane

8.1 Alcune code non poissoniane
8.2 M/G/1
8.3 M/D/1
8.4 $M/E_k/1$
8.5 $M/H_r/1$
8.6 Domande ed esercizi

8.1 Alcune code non poissoniane

Nei modelli non poissoniani almeno uno tra gli intertempi d'arrivo e i tempi di servizio non è una v.a. esponenziale. In particolare poiché il modellamento con v.a. esponenziali è più spesso non accettabile per i tempi di servizio che non per gli intertempi d'arrivo, si limita l'analisi ai casi M/G/1.

8.2 M / G / 1

La coda M/G/1 ha arrivi poissoniani, ma tempi di servizio qualunque, purché indipendenti e omogenei (con la stessa distribuzione), con media $1 / \mu$ e varianza σ^2 note.
Anche in questo caso la condizione di stazionarietà è $\rho = \lambda / \mu < 1$

Si dimostra che:

$$L_q = (\lambda^2 \sigma^2 + \rho^2) / (2 (1 - \rho))$$

(formula di **Pollaczek-Khintchine**) dove σ^2 è la varianza del tempo di servizio.
A partire da L_q si possono derivare le altre grandezze di interesse nel modo usuale

$$L_s = L_q + \rho$$
$$W_q = L_q / \lambda$$
$$W_s = W_q + 1 / \mu$$

Si osservi che L_q cresce con σ e quindi un servitore regolare ha prestazioni migliori.

8.3 M / D / 1

La coda M/D/1 con arrivi poissoniani e tempo di servizio costante è un caso particolare di M/G/1 con $\sigma=0$, dove la formula di Pollaczek - Khintchine si riduce a:

$$L_q = \rho^2 / (2 (1 - \rho)).$$

Il numero medio dei clienti in attesa di servizio è per una coda M/D/1 la metà che per M/M/1. Infatti la varianza del tempo di servizio è 0 per M/D/1 mentre è $1/\mu^2$ per M/M/1.

8.4 $M / E_k / 1$

La coda $M/E_k/1$ è utilizzata per modellare casi intermedi in cui, oltre che la media e la varianza, è nota anche la forma della distribuzione degli intertempi di servizio. E_k indica che i tempi di servizio sono v.a. con **distribuzione di Erlang** di ordine k

$$f(t) = (k\mu)^k t^{k-1} e^{-k\mu t} / (k-1)! \quad \text{per} \quad t \geq 0$$

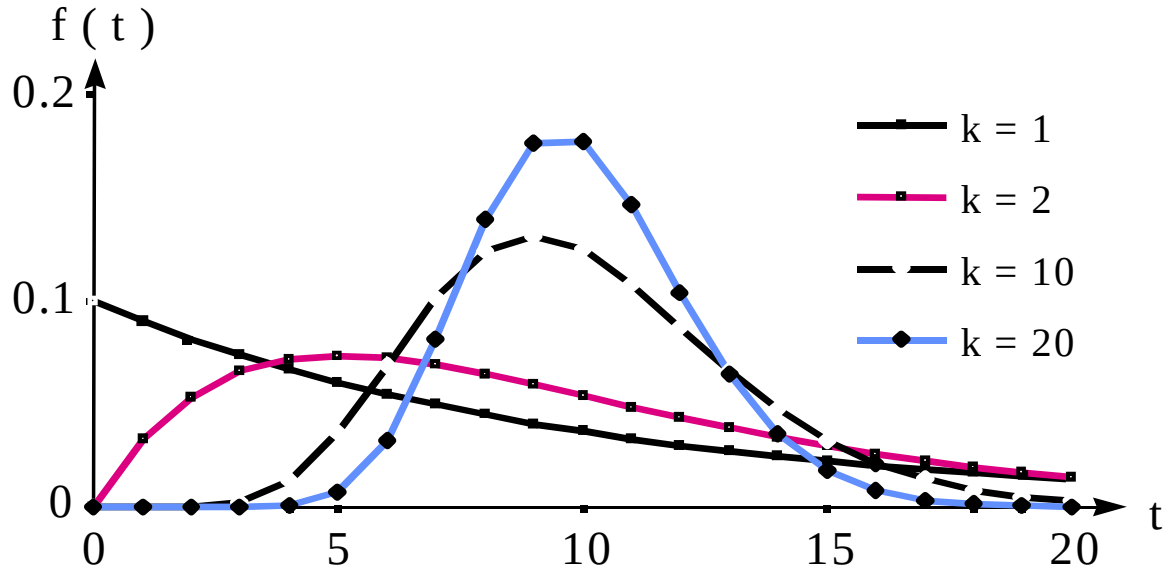
dove k è un intero positivo ed è detto **fattore di forma**.
La distribuzione di Erlang di ordine k ha media $1 / \mu$ e varianza $1 / k\mu^2$. E_k è quindi una variabile aleatoria nonnegativa che dipende da due parametri: μ e k dove μ determina la media k determina la varianza.

Le funzioni di Erlang godono della seguenta proprietà:

Proprietà
La somma di k variabili aleatorie indipendenti esponenziali ciascuna con media $1 / k\mu$:
$T = T_1 + T_2 + ... + T_k$
è una v.a. con distribuzione di Erlang di ordine k e parametri μ e k.

La precedente proprietà implica che per k che tende all'infinito la distribuzione di Erlang tende alla a diventare la distribuzione normale.

LA DISTRIBUZIONE DI ERLANG ($\mu = 0.1$)



La stessa proprietà implica che la distribuzione di Erlang può essere interpretata come la distribuzione del tempo di servizio di un sistema in cui vi siano k servitori esponenziali in serie, in cui però il primo servitore non può iniziare un nuovo servizio se l'ultimo non ha concluso il proprio.

Per la coda $M/E_k/1$ si ricava che

$$L_q = ((1 + k) \lambda^2) / (\mu (\mu - \lambda))$$

da cui

$$L_s = L_q + \rho$$
$$W_q = L_q / \lambda$$
$$W_s = W_q + 1 / \mu$$

8.5 $M / H_R / 1$

La coda $M/ H_R/1$ è utilizzata quando le varianze dei tempi di servizio sono maggiori di $1/\mu^2$.
H_R indica che i tempi di servizio sono v.a. con **distribuzione iperesponenziale** di ordine R:

$$f (t) = \Sigma^R \alpha_i \mu_i \exp(- \mu_i t) \quad \text{per} \quad t \geq 0$$

La distribuzione iperesponenziale può essere interpretata come la distribuzione del tempo di servizio di un sistema in cui vi siano R servitori esponenziali con prestazioni differenti. Il cliente sceglie con probabilità α_i servitore l'i-mo con la condizione che

$$\alpha_1 + \alpha_2 + ... + \alpha_R = 1$$

e che un cliente non può iniziare ad essere servito prima che il cliente che lo precedeva non sia uscito dal sistema. In altre parole i servitori sono in parallelo ma non possono lavorare contemporaneamente.

Per la coda $M/ H_R/1$ si ricava che la varianza del tempo di servizio è:

$$\sigma^2 = 2\sum_{i=1}^{R} \frac{\alpha_i}{\mu_i^2} - \left(\sum_{i=1}^{R} \frac{\alpha_i}{\mu_i} \right)^2 .$$

Sostituendo tale valore nella formula di Pollaczek-Khintchine si ottiene L_q e conseguentemente si possono derivare L_s, W_q, e W_s.

8.6 Domande ed esercizi

1. Determinare di quanto aumenta il tempo medio di permanenza nel sistema se la deviazione standard del tempo di servizio varia del 10%, per $\sigma=1/\lambda$, per $\sigma=0.5/\lambda$, per $\sigma=2/\lambda$, e per $\rho = 0,5$, $\rho = 0,7$, $\rho = 0,9$.

2. Determinare la differenza dei valori delle grandezze caratteristiche tra una coda M/M/2 e una coda $M/H_2/1$.

Indice analitico

aree di attesa; 6
buffer; 6; 8; 9; 12
capacità del sistema; 8; 9
clienti; 3; 6; 7; 11; 12; 13; 14; 18; 21; 22; 23; 24; 26; 27; 29
coda; 6; 7; 8; 9; 10; 11; 18; 21; 23; 24; 25; 26; 27; 28
coda stabile; 12; 21; 22; 26; 27
code non poissoniane; 28
costi; 5; 11; 23
costi fissi; 11
costi variabili; 11
disciplina di servizio; 6; 7; 8; 9; 13
discipline su priorità; 8
distribuzione di Erlang; 9; 29
distribuzione esponenziale; 4; 9; 14; 15; 21
distribuzione iperesponenziale; 30
Erlang; 3; 29
fattore di forma; 29
fattore di utilizzazione; 11; 21; 23; 24; 25
FCFS; 8; 9; 12; 13
FIFO; 8
file di attesa. *Vedi* Code
formula di Little; 22
formula di Little generalizzata; 22
funzione di densità; 6
incentivo ai servitori; 12
intertempo d'arrivo; 8; 9; 14; 15
intertempo tra partenze; 24
LCFS; 8
LIFO; 8
L_q; 11
L_s; 11
M / D / 1; 28
M / E_k / 1; 29
M / G / 1; 28
M / H_R / 1; 30
M / M / 1; 21
M / M / 1 / K; 26
M / M / 1 / N / N; 27
M / M / s; 25
M/M/∞; 26
mancanza di memoria; 15
modelli descrittivi; 11
modelli normativi; 11
notazione di Kendall; 9
paradosso del tempo di servizio residuo; 15
p_n; 11
Pollaczek-Khintchine; 28; 30
popolazione; 7; 9; 18; 27
processi stocastici; 4; 6; 13
processo d'arrivo; 6; 7; 8; 24
processo di Poisson; 14; 16; 18
processo di servizio; 6; 7; 8
processo nascite-morti; 18; 19; 25
processo stazionario; 8
processo stocastico; 6; 8; 18
progetto di una coda; 11
proprietà v.a. esponenziali; 4; 8; 14; 15; 16; 19; 25; 29
proprietà markoviana; 9
raggiungimento del regime; 23
reti di Jackson; 24
servitori; 6; 7; 8; 9; 11; 12; 23; 24; 25; 26; 27
simulazione; 24
SIRO; 8
sistema coda; 6; 9
stabilità; 21
stato di una coda; 5; 9; 18; 19; 21
tasso medio d'arrivo di clienti; 15
tempi di servizio; 8; 9; 14; 15; 24; 28
Teoria delle Code; 3; 4; 5; 11
testi di consultazione; 4
testo di riferimento; 4
variabile aleatoria; 6; 14; 29
variabili aleatorie; 4; 16; 21; 29
W_q; 11
W_s; 11
ρ; 11