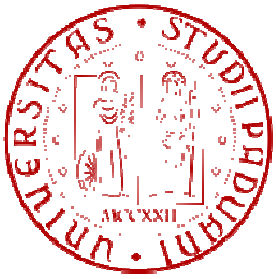


Università degli studi di Padova

Dipartimento di Tecnica e Gestione dei Sistemi Industriali

Corso di Laurea Triennale in Ingegneria Gestionale

METODI DI DIMENSIONAMENTO DI UN BUFFER INTEROPERAZIONALE: TEORIA DELLE CODE E SIMULAZIONE MONTECARLO

RELATORE: CH.MO PROF. GAMBERI MAURO

LAUREANDO: SPAGNOLO IVAN

ANNO ACCADEMICO: 2012-13

INDICE

INTRODUZIONE

La produzione industriale può essere realizzata con diverse modalità a seconda del posizionamento della propria azienda nel grafico prodotti-quantità. Per le aziende che realizzano un mix produttivo limitato ma con elevati volumi unitari l' organizzazione più adatta è quella "per prodotto", ovvero si posizionano in serie le varie stazioni di lavoro in modo tale da rispettare il ciclo tecnologico del prodotto che si realizza (i pezzi vengono lavorati sequenzialmente passando da una stazione alla successiva).

Si realizza in tal modo la cosiddetta produzione in linea; le aziende che operano in questo modo possono poi decidere se utilizzare una linea sincrona o una linea asincrona.

Nella linea asincrona tra una stazione di lavoro e l'altra è presente un punto di accumulo dei pezzi, che prende il nome di **magazzino interoperazionale o buffer**.

L' inserimento del buffer permette di disaccoppiare le fasi operative in quanto l' interruzione del lavoro di una qualsiasi stazione non costringe a fermare tutta la linea (la stazione a monte di quella guasta accumula i pezzi nel buffer, quella a valle li preleva). Si devono però sostenere i costi relativi agli immobilizzi di materiale (work in process, wip), i costi di gestione dei buffer stessi ed il fatto che questi occupano spazio all' interno del capannone.

L' alternativa è la linea sincrona, priva di buffer interoperazionali. Il vantaggio è che si evitano i costi a cui si è appena fatto riferimento, ma si è costretti a sovradimensionare la capacità produttiva perché ciascuna stazione di lavoro risente delle inefficienze delle stazioni a valle (se si ferma una stazione, tutta la linea è costretta ad interrompere il lavoro), quindi aumentano gli investimenti in macchinari.

Appare quindi chiaro che, qualora si voglia utilizzare la linea asincrona, risulta fondamentale **dimensionare in maniera corretta il buffer interoperazionale**; ed è proprio questo l' obiettivo della tesi.

Nel seguente elaborato verranno infatti trattate due diverse metodologie per provvedere a questo scopo: nella prima parte verrà seguito un approccio analitico, nella seconda un approccio simulativo.

Verrà approfondita la **teoria delle code o delle file d'attesa**, che è un argomento molto vasto e che può essere applicato a qualsiasi ambito, purchè ci sia un fenomeno in cui si formino delle "code" (cioè dei clienti si mettono in fila in attesa di un servizio). Nel primo capitolo verrà presentato questo fenomeno in termini del tutto generali specificando quelli che sono i componenti e gli obiettivi della teoria delle code e la notazione da utilizzare.

Il secondo capitolo tratta i principali fondamenti statistici della teoria, che è infatti basata sulla statistica e sulla probabilità.
Nel terzo capitolo verrà presentata dal punto di vista analitico la teoria vera e propria, concentrandosi però sulle code che sono di interesse ai fini di questo elaborato, fornendo solo qualche accenno sulle altre.
Nel quarto capitolo verrà chiarito il concetto di rete di code, cioè l' interconnessione di più file d' attesa, sempre restando focalizzati solamente sugli aspetti più utili allo scopo specificato.
Il quinto capitolo invece tratta l' approccio simulativo, approfondendo la **metodologia di simulazione MonteCarlo**. Anche questa teoria è applicabile ai più svariati ambiti grazie alla propria versatilità (con tale metodo si può simulare infatti qualsiasi fenomeno aleatorio), ma dopo aver presentato in termini generali l' argomento ed aver chiarito i fondamenti statistici su cui essa si basa , verrà spiegato come applicarla al dimensionamento del buffer interoperazionale.
Infine nel sesto capitolo verranno utilizzate entrambe le tecniche applicandole ad un semplice esempio.

CAPITOLO 1: Introduzione alla Teoria delle Code

1.1: Concetti di base

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 verificare situazioni temporanee in cui chi fornisce il servizio non ha la possibilità di soddisfare immediatamente le richieste.
I campi di applicazione della teoria delle code sono pertanto estremamente numerosi; ad esempio sono in generale soggetti ad attese:

- i clienti in banca o in posta;
- le automobili ad un incrocio;
- gli aerei in attesa di decollare o di atterrare;
- le parti in attesa di essere lavorate;
- …

In particolare la teoria delle code è una tecnica avanzata che può rappresentare un notevole supporto alle decisioni anche nell'ambito della progettazione di un impianto industriale. Infatti una "coda" si presenta tutte le volte che dei "clienti"(non necessariamente persone fisiche) richiedono un "servizio" da parte di stazioni di erogazione. Dal punto di vista fisico un **sistema coda** è un sistema composto da un insieme non vuoto di **servitori,** capaci di fornire un certo servizio, 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.** 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 (cioè arrivi e uscite dei clienti). Viceversa il tempo che deve trascorrere fino all'arrivo di un dato cliente è un valore che varia con continuità.

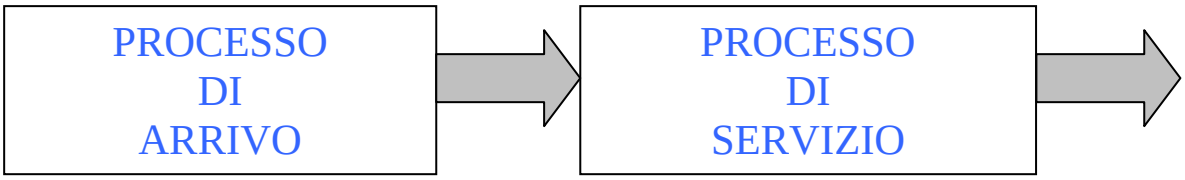


Fig. n.1.1.1: Rappresentazione schematica

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

- la popolazione dei clienti
- il processo di 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, e la teoria delle code è pesantemente influenzata da eventuali limitazioni sul numero di clienti che possono potenzialmente richiedere il servizio: quando questo valore è comparabile col numero di stazioni di servizio vengono meno le ipotesi che permettono una buona applicabilità.
Di conseguenza la teoria delle code può offrire il suo contributo nel caso di popolazione potenziale molto più numerosa delle stazioni di servizio; questo viene tradotto nella pratica con la definizione "popolazione infinita". Inoltre i clienti di una stessa popolazione sono tra loro indistinguibili; ogni qualvolta ci sia la necessità di distinguerli, si considerano appartenenti a popolazioni differenti.

Il **processo di 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 di 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. Si pensi per esempio a limitazioni di carattere spaziale (luoghi di ridotte dimensioni). In queste condizioni la teoria delle code peggiora notevolmente le proprie performance; nella pratica si richiede l'ipotesi di "capacità infinita" riguardo all'accettabilità di clienti.

I **servitori** sono in numero noto e costante fissato a livello di progetto. Solitamente essi hanno caratteristiche identiche, possono sempre lavorare in parallelo e non possono mai rimanere inattivi in presenza di clienti in coda. Inoltre anche se vi sono più servitori in una coda in generale si assume l'esistenza di un unico buffer comune (come esempio basti pensare a quanto avviene in ambito industriale: c'è un unico buffer di una particolare materia prima che va ad alimentare un centro di lavoro con più macchine).

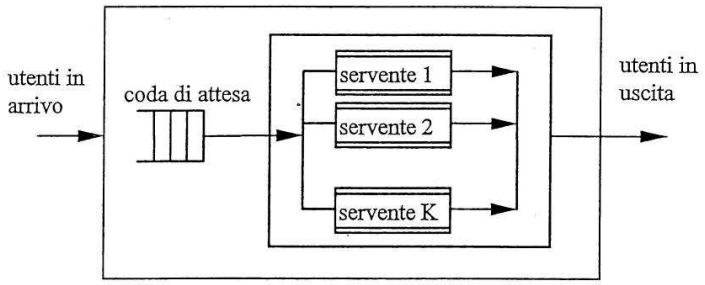


Figura n.1.1.2: Schema di una coda con buffer comune e servitori multipli

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; il processo d'arrivo è quindi 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** rappresenta la modalità di soddisfacimento delle richieste dei clienti che richiedono il servizio. 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 (lastcome 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).

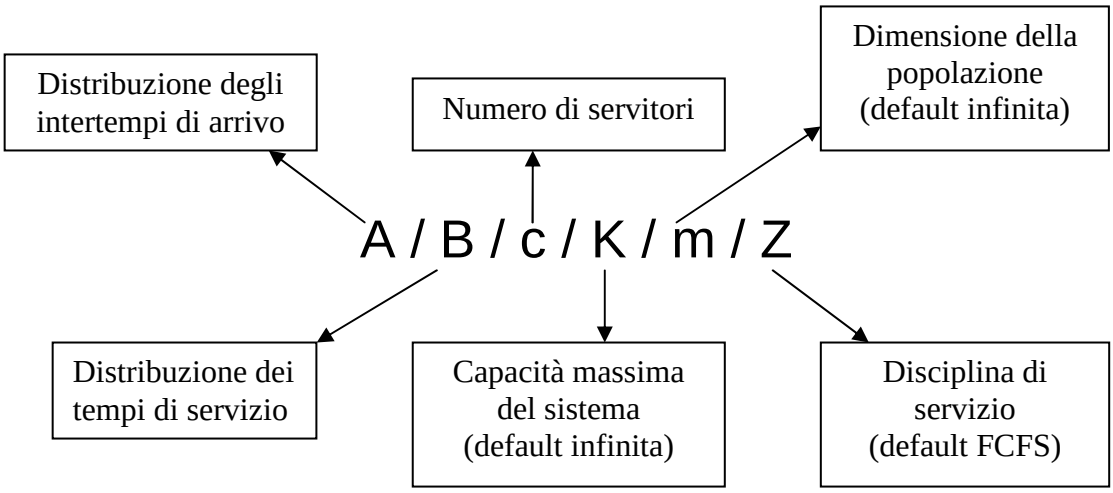
1.2: Obiettivi della teoria

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. 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 considerano questi costi con enfasi diversa (i clienti ritengono fondamentale la riduzione dei tempi d'attesa, mentre il gestore del sistema è probabilmente più 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, sotto alcune ipotesi, sono facilmente calcolabili in quanto dipendono parametricamente dalla struttura della coda:

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

1.3: Notazione

Nel 1953 David George Kendall introdusse la seguente notazione al fine di sintetizzare le caratteristiche di una coda:



Esempio: **M/M/1** sta per **M/M/1/∞/∞/FCFS:** coda con processo degli arrivi e dei servizi markoviani, un servitore, capacità del sistema infinita e arrivi provenienti da una popolazione infinita che vengono serviti con criterio FCFS.

CAPITOLO 2: Richiami di probabilità e statistica

2.1: Variabili aleatorie e probabilità

Una **variabile aleatoria discreta X** è un'entità che può assumere un numero discreto (finito o infinito) di valori x_i.

Una **variabile aleatoria continua Y** è un'entità che può assumere valori y_i in un sottoinsieme S della retta reale composto da uno o più intervalli.

La **funzione densità di probabilità f(x)** rappresenta la probabilità che la variabile aleatoria X assuma il valore x_i in conseguenza dell'accadere di un certo evento elementare. Ogni valore x_i ha probabilità di occorrenza $P(X=x_i) = p_i$, dove $\sum_i p_i = 1$ nel caso discreto; p(y) è invece tale che $\int_S p(y)dy = 1$ nel caso continuo.

La **funzione distribuzione di probabilità F(x)** esprime la probabilità che la variabile aleatoria X assuma un valore minore o uguale a x_i ed è definita come $F(x_i) = P(X \leq x_i)$.

Tale funzione è monotona non decrescente e tale che $0 \leq F(x_i) \leq 1$, con $F(-\infty) = 0$ e $F(+\infty) = 1$

Per x_1 e x_2 tali che $x_1 \leq x_2$, si ha che $F(x_1) \leq F(x_2)$ e $F(x_1) - F(x_2) = P(x_1 \leq X \leq x_2)$.

La funzione di distribuzione di una variabile aleatoria continua è continua e derivabile. La funzione di distribuzione di una variabile aleatoria discreta presenta invece un andamento a "scalinata", con punti di discontinuità in corrispondenza dei valori x_i considerati e con "salti" pari alle probabilità corrispondenti.

2.2: 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 (in seguito v.a.) X ha **distribuzione esponenziale** con parametro $\lambda > 0$ quando la sua densità di probabilità f(x) è:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

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 di arrivo dei clienti.

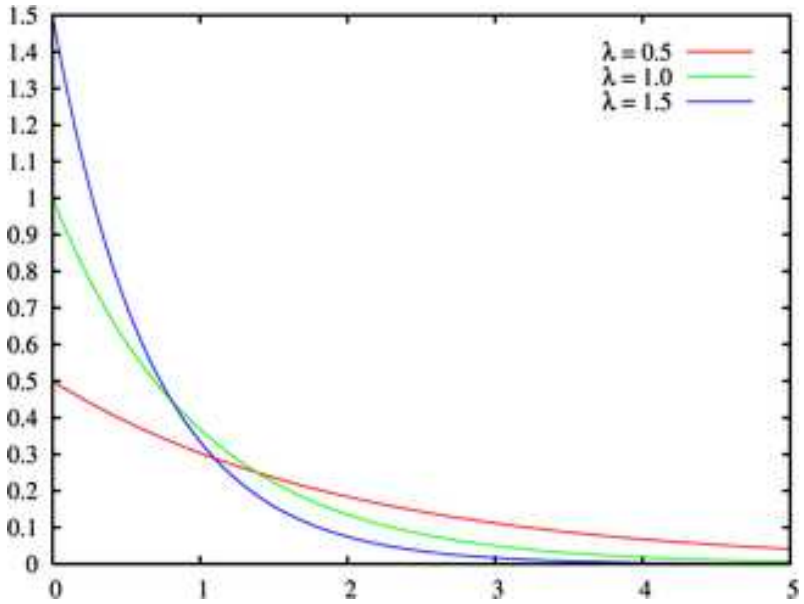


Fig n.2.2.1: Distribuzione esponenziale

I tempi intercorrenti tra due eventi successivi relativi allo stesso processo (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 che 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 non deve quindi avere memoria, ovvero deve godere della proprietà markoviana (che prende il nome dal matematico russo Andrej Andreevič Markov che per primo ne sviluppò la teoria). Matematicamente:

$P(X > x + u \mid X > u) = P(X > x)$ in cui il primo membro rappresenta la situazione iniziale; noto che la variabile aleatoria assume valori maggiori di u (il tempo d'attesa è più di u), la probabilità che essa ritardi ancora di un tempo x (cioè si realizzi al minimo al tempo x + u) è la stessa di quella che la variabile si realizzi al minimo al tempo x rispetto all'inizio della misurazione; in altre parole quello che e' successo in precedenza non incide in alcun modo su quello che possiamo aspettarci per il futuro. Supponiamo ad esempio che X sia il tempo di vita di una macchina soggetta a guasti e supponiamo che essa non sia guastata fino al tempo u. La proprieta' di non-memoria afferma che, supposto che la macchina si arrivata al tempo u senza guasti, la probabilita' che la macchina non si guasti per un ulteriore intervallo di tempo pari a x (calcolato a partire da u), non dipende dal tempo u ovvero da quanto tempo è trascorso dall' ultimo guasto verificatosi.

Una v.a. esponenziale soddisfa tutte queste condizioni; inoltre la mancanza di memoria 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 distribuzione esponenziale è una funzione strettamente decrescente, quindi i valori più piccoli sono più probabili; il valore atteso(o speranza matematica) è $E\{X\}=1/\lambda$ e la varianza è $\sigma\{X\} = 1/\lambda^2$.

2.3: Distribuzione e processo di Poisson

Una variabile aleatoria X ha una distribuzione di Poisson se assume valori in N con probabilità: $P(n) = e^{-\lambda}\frac{\lambda^n}{n!}$ per ogni $n \in \mathbb{N}$.
dove λ è il numero medio di eventi per intervallo di tempo, mentre n è il numero di eventi per intervallo di tempo (lo stesso col quale si misura λ) di cui si vuole la probabilità.
Il valore atteso e la varianza valgono entrambi λ.
Le distribuzioni di Poisson ed esponenziale sono tra loro collegate in quanto se la distribuzione di Poisson di parametro λ descrive il numero di eventi in un intervallo di tempo, il tempo di attesa tra due eventi successivi è descritto dalla distribuzione esponenziale di parametro λ; proprio per questo trovano applicazione nello studio dei cosiddetti processi di Poisson, rappresentabili come eventi casuali, indipendenti fra loro e distribuiti uniformemente nel tempo.
Quando gli intertempi sono esponenziali la probabilità che si verifichino n arrivi in un intervallo di tempo pari a t è calcolabile come $P\{N(t) = n\} = [(\lambda t)^n e^{-\lambda t}] / n!$
Il numero N(t) segue il cosiddetto processo di Poisson; esso ha valore atteso $E\{N(t)\} = \lambda t$.
Ai processi di Poisson si generalizzano le proprietà delle v.a. esponenziali; inoltre:
- λdt rappresenta la probabilità di occorrenza di un evento in un intervallo di tempo infinitesimo dt;
- la probabilità che in una certa area di opportunità (intervallo di tempo continuo nel quale un evento può verificarsi più volte) l'evento di interesse si verifichi più di una volta diminuisce al diminuire dell'area di opportunità stessa;
- la probabilità che in una certa area di opportunità si osservi un certo evento è la stessa in tutte le varie aree;

- il numero di volte in cui un evento si realizza in una certa area di opportunità è indipendente dal numero di volte in cui un evento si è verificato in un' altra area.

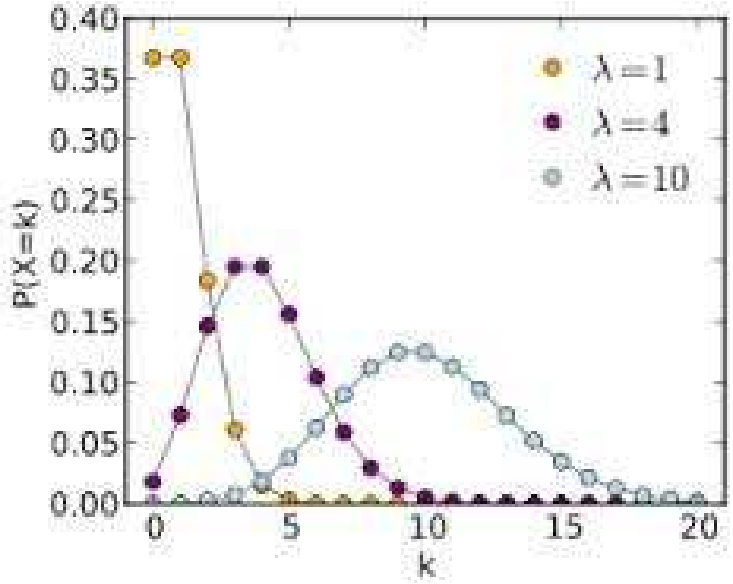


Fig n.2.3.1: Distribuzione di Poisson

2.4: Processo nascite-morti

La maggior parte dei modelli elementari di coda considera che le entrate e le uscite del sistema si verifichino secondo un processo di nascita e morte; il termine nascita si riferisce all'arrivo di una nuova unità e il termine morte alla partenza di un'unità servita. Esso rappresenta il numero di elementi N(t) di una popolazione ed assume che, ad ogni generico istante t, possa avvenire un solo evento (di nascita o di morte) e che, data una popolazione di numerosità N(t) = n, l'intervallo di tempo fino alla prossima nascita sia una variabile aleatoria esponenziale con parametro λ_n, mentre l'intervallo di tempo fino alla prossima morte sia una variabile aleatoria 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 (coefficiente di natalità) e di morte (coefficiente di mortalità) quando la popolazione è composta di n individui. Graficamente:

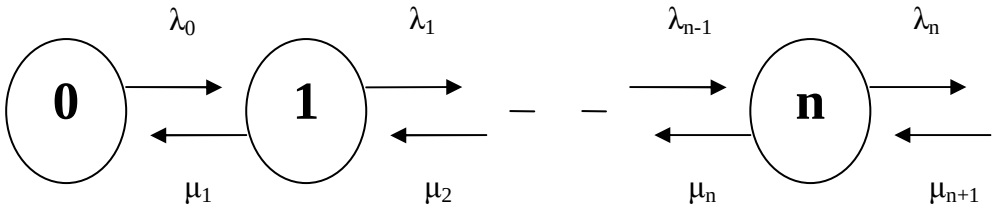


Fig. n.2.4.1: Rappresentazione grafica del processo

Gli ovali rappresentano lo stato del sistema (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.
Per ogni stato n si definisce probabilità di stato la probabilità che il numero di utenti nel sistema all'istante t, N(t), sia uguale a n; in generale le probabilità di stato sono funzione dell'istante di tempo considerato, ma se il sistema raggiunge uno stato stazionario esse diventano indipendenti dall'istante t.
Un particolare processo nascite-morti, dove avvengono solo nascite, è quello di Poisson.
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:

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

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

Per t che tende all'infinito, se il tasso delle morti complessivamente supera il tasso delle nascite ($\lambda < \mu$, condizione di stazionarietà), il processo diventa stazionario, ovvero le sue probabilità di stato diventano costanti indipendenti dal tempo e quindi $p_n(t) = p_n$, per ogni tempo t. Esiste quindi una soluzione stazionaria se esistono finiti i seguenti limiti:

$$\lim_{t\to\infty} p_n(t) = p_n \qquad \text{per } n=0,1,2,\ldots$$

Si noti che assumere che esista la distribuzione stazionaria non è un' ipotesi limitativa, in quanto in molti casi pratici, è sufficiente limitarsi a valutare la distribuzione stazionaria p_n piuttosto che la probabilità $p_n(t)$, assumendo che il sistema è stato in funzione per un tempo sufficientemente grande per raggiungere tale condizione di equilibrio.
Sotto questa condizione il sistema di equazioni differenziali diventa un sistema di equazioni lineari omogeneo; per determinare la soluzione stazionaria infatti basta prendere il limite per $t \to \infty$ di ambo i membri di ciascuna delle equazioni differenziali; sapendo che

$\lim_{t\to\infty} p_n(t) = p_n$, si impone che $\lim_{t\to\infty} \frac{d\, p_n(t)}{dt} = 0$

(in quanto per la stazionarietà per $t\to\infty$ le $p_n(t)$ assumono valori costanti p_n), ricordando inoltre che $\sum_n p_n(t) = 1$.

Sviluppando i calcoli si ottiene quindi il sistema algebrico:

$\lambda_{n-1}\, p_{n-1} - (\lambda_n + \mu_n) p_n - \mu_{n+1}\, p_{n+1} = 0$ per $n > 0$

$\mu_1\, p_1 - \lambda_0\, p_0 = 0$

Ciascuna equazione del sistema rappresenta, per ogni stato, una equazione di bilancio, ovvero uguaglia il "flusso entrante" e il "flusso uscente". Più precisamente, le equazioni di bilancio esprimono l'uguaglianza tra il tasso con cui il processo lascia uno stato e il tasso con cui il processo entra nello stato. Consideriamo ad esempio lo stato n = 0: quando il sistema è allo stato 0, esso può lasciare questo stato solo a causa di una nascita (perchè il sistema è vuoto), quindi si ha che il processo lascia lo stato 0 con tasso $\lambda_0\, p_0$. D'altra parte, lo stato 0 può essere raggiunto a partire dallo stato 1 solo con una morte e quindi si ha che il tasso con il quale il sistema entra nello stato 0 è $\mu_1\, p_1$ e quindi si ottiene l' equazione di bilancio, da cui si può ricavare: $p_1 = (\lambda_0 / \mu_1)\, p_0$.

Il ragionamento si ripete per gli stati successivi allo stato 0; si ottiene quindi che la soluzione stazionaria del processo nascite-morti è

$p_n = C_n\, p_0$ $\quad n = 1,2,...$ $\quad$ dove $\quad C_n = (\lambda_{n-1}\lambda_{n-2} \dots \lambda_0) / (\mu_n\, \mu_{n-1} \dots \mu_1)$.

Osservando infine che i termini p_n rappresentano delle probabilità e che quindi deve valere

$p_0 + \sum_{n=1}^{\infty} p_n = 1$, si ottiene (sostituendo l' espressione appena ricavata per p_n) che:

$p_0 = 1 - \sum_{n=1}^{\infty} C_n\, p_o$ $\quad$ da cui $\quad p_0 = 1 / (1 + \sum_{n=1}^{\infty} C_n)$.

Risulta fondamentale che $\sum_{n=1}^{\infty} C_n < \infty$, che rappresenta la condizione di esistenza dello stato stazionario: se infatti tale sommatoria fosse infinita, allora si avrebbe $p_n = 0$ per ogni n finito, e quindi lo stato del processo crescerebbe indefinitamente e non si raggiungerebbe mai l'equilibrio.

Da questi risultati si possono ricavare le distribuzioni di probabilità di tutte le code poissoniane.

CAPITOLO 3: Trattazione delle code

3.1: Coda M/M/1

Questo è l'esempio più semplice di coda, fisicamente formata da un buffer e da un solo servitore, con distribuzioni esponenziali per i tempi d'interarrivo dei clienti e per i tempi di servizio, e capacità infinita per la coda d'attesa. In generale i tassi di nascita e di morte possono essere funzioni dello stato e sono usualmente indicati con λ_i e μ_i; particolarmente semplice è il caso in cui tutti i coefficienti sono uguali tra loro, indipendentemente dallo stato, per cui risulta che la coda M/M/1 può essere modellata come un processo nascite-morti con $\lambda_n = \lambda$ e $\mu_n = \mu$.

Conseguentemente le probabilità di stato risultano $p_n = \rho^n p_0$, n=1,2,… dove **$\rho = \lambda/\mu$** è il **fattore di utilizzazione** ed esprime il rapporto tra il tempo medio tra due arrivi ed il tempo medio di servizio. Dato che vale la condizione $p_0 = 1 / (1 + \sum_{n=1}^{\infty} C_n) = 1 / (1 + \sum_{n=1}^{\infty} \rho^n)$, è chiaro che p_0 esiste se e solo se $\rho < 1$ (e quindi $\lambda < \mu$), 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, come visto in precedenza, lo stato stazionario non può essere raggiunto e la coda cresce all'infinito qualora essa non occorra.

Per $\rho < 1$ si verifica che: $1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \rho^n p_o = p_o \sum_{n=0}^{\infty} \rho^n = p_o / (1 - \rho)$ da cui si ricava che

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

Dalle relazioni precedenti si evince che ρ può essere interpretato anche come il tasso di occupazione del servitore (cioè 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:

$$L_s = E\{n\} = \sum_{n=0}^{\infty} n\, p_n = \sum_{n=0}^{\infty} n\, \rho^n (1- \rho) = (1- \rho) \sum_{n=0}^{\infty} n\, \rho^{n-1} \rho = (1- \rho)\, \rho \sum_{n=0}^{\infty} \frac{d}{dp}(p^n) =$$

$$= (1- \rho)\, \rho \frac{d}{dp} \left[\sum_{n=0}^{\infty} p^n \right] = (1- \rho)\, \rho \frac{d}{dp} \left[1 / (1- \rho) \right] = - (1- \rho)\, \rho / (1 - \rho)^2 (-1) = \mathbf{\rho / (1 - \rho)}$$

Il numero medio di clienti in attesa vale: $L_q = L_s$ - [n. medio di clienti correntemente serviti] = $L_s - \rho = \rho / (1 - \rho) - \rho =$ **$\rho^2 / (1 - \rho) = \lambda / (\mu - \lambda)$**

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 vale quindi λ; si può dedurre di conseguenza che il tempo media di attesa dei clienti nel sistema è $Ws = Ls / \lambda$.
Questa formula, detta **formula di Little**, vale per qualunque sistema in equilibrio e si enuncia affermando che il numero medio di utenti presenti nel sistema è eguale al tempo medio di permanenza nel sistema per il tasso d'ingresso (indipendentemente da qualunque ipotesi sul processo degli arrivi, sul processo dei servizi, sui servitori ecc.).
Applicando la formula di Little alla coda M/M/1 si ottiene che il tempo di attesa medio nel sistema è:
Ws = $\lambda / (\mu - \lambda) \lambda$ **$= 1 / (\mu - \lambda)$**
Mentre il tempo medio d'attesa in coda è:
Wq = $Ws - (1 / \mu) = 1 / (\mu - \lambda) - (1 / \mu) =$ **$\lambda /(\mu (\mu - \lambda))$.**

Si può vedere come il parametro $\mu - \lambda$ rappresenti, in termini medi, la differenza fra la capacità di erogazione del servizio e quella di "consumo" dello stesso da parte dei clienti.
Si è osservato inoltre 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.
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.

3.2: Coda M/M/s

Questo tipo di coda è identica alla precedente a parte il fatto che sono presenti s servitori in parallelo, ciascuno con tasso di servizio 1/ μ (uguale per tutti). Di conseguenza, sono attivati contemporaneamente più eventi di morte, che nel caso markoviano significa un tasso di morte proporzionale al numero degli eventi (cioè la velocità media alla quale le unità servite abbandonano il sistema dipende dallo stato dello stesso), ovvero:

$$\begin{cases} \mu_n = n\,\mu & \text{per } 1 \le n \le s; \\ \mu_n = s\,\mu & \text{per } n > s. \end{cases}$$

Il fattore di utilizzazione vale $\rho = \lambda / (s\mu)$ per $n > s$; se la condizione di stabilità per avere una soluzione stazionaria $\rho < 1$ (e quindi $\lambda < s\mu$) è rispettata si ottiene, a partire da

$p_0 = 1 / (1 + \sum_{n=1}^{\infty} \rho^n) = 1 / \sum_{n=0}^{\infty} \rho^n$ si ottiene dopo aver sviluppato alcuni calcoli (omessi):

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

Il numero medio di servitori occupati vale $s\rho = \lambda / \mu$ e ciascun servitore ha quindi un utilizzo pari a ρ. Il numero medio di clienti presenti nel sistema vale:

$\mathbf{L_s} = \sum_{n=0}^{s-1} (n-s)\, p_n = \mathbf{s\,\rho} + \dfrac{\mathbf{(s\,\rho)^s}}{\mathbf{s!}} \dfrac{\mathbf{p_o}}{\mathbf{(1-\rho)^2}}$ dove il termine $s\rho$ è proprio il numero medio di servitori occupati mentre il secondo termine rappresenta il numero medio di clienti in coda $\mathbf{L_q}$ (infatti $L_s = L_q + \lambda / \mu$).

Per ricavare i rimanenti indici di prestazione si applica la formula di Little (che si ricorda essere valida per un qualsiasi sistema stabile):

$\mathbf{W_q} = L_q / \lambda$; $\mathbf{W_s} = W_q + 1 / \mu$

3.3: Cenni ad altre code

Sino ad ora si è fatto riferimento ad intertempi d' arrivo e tempi di servizio esponenziali, ma non sempre questa ipotesi è verificata; spesso almeno uno dei due tempi segue un' altra distribuzione (in gran parte dei casi è il tempo di servizio che non è una v.a. esponenziale, per cui in seguito si manterrà l' ipotesi di arrivi di Poisson). Le code che possiedono queste caratteristiche si dicono non poissoniane, e nel seguito saranno esposti i risultati principali.

- Coda M/G/1: questa coda ha ancora intertempi di arrivo esponenziali ma tempi di servizio qualunque, purchè indipendenti ed omogenei, e con una distribuzione di media $1 / \mu$ e varianza σ^2 note. Se la condizione di stazionarietà $\rho < 1$ è rispettata vale la **formula di Pollaczek-Khintchine**:

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

  A partire da questa formula gli altri parametri si ricavano nel modo usuale con la formula di Little.

- Coda M/D/1: in questo caso gli arrivi sono poissoniani ma i tempi di servizio sono costanti (deterministici); quindi $\sigma^2 = 0$ e la formula di Pollaczek-Khintchine si riduce a

  $\mathbf{L_q = \rho^2 / (2(1- \rho))}$

- Coda $M/E_k/1$: gli arrivi sono poissoniani mentre i tempi di servizio seguono la distribuzione di Erlang di ordine k:

  $f(t) = [(k \mu)^k t^{k-1} e^{- k\mu t}] / (k-1)!$ dove k è un intero positivo detto fattore di forma.

  Tale distribuzione ha media $1 / \mu$ e varianza $1 / k\mu^2$, e gode della proprietà che la somma di k variabili aleatorie indipendenti esponenziali ciascuna con media $1 / k\mu$ 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 a diventare la distribuzione normale.

  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.

  Si ricava che $\mathbf{L_q = ((1 + k)\lambda^2) / (\mu (\mu - \lambda))}$.

CAPITOLO 4: Reti di code

4.1: Introduzione

Solitamente una risorsa non è utilizzata in modo isolato; più spesso diverse risorse sono interconnesse fra loro per costruire un unico sistema, che si può rappresentare come una rete in cui le singole risorse sono i nodi e dove i rami indicano i flussi di utenti da una risorsa all'altra. Come esempio si pensi al caso di una linea produttiva, formata da una serie di stazioni di lavoro. Entrando nella prima stazione della linea, il pezzo si pone in attesa di fronte alla prima macchina, quando ha terminato la lavorazione prosegue immediatamente verso la seconda, e così via fino al completamento del ciclo tecnologico quando il prodotto finito esce dal sistema.

Questo tipo di struttura costituisce una **rete di code**.

Una rete di code è quindi l'interconnessione di un insieme di M stazioni (indicate con j = 1,2,…,M), ciascuna costituita da uno o più serventi e da una coda. Tutte le grandezze riferite alla stazione j-esima si indicano con il pedice j: ad esempio, con s_j si indica il numero di serventi presente nella stazione j.

Come detto, in generale, i clienti in uscita da una stazione non tornano subito nella popolazione di appartenenza ma si dispongono invece in coda ad un' altra stazione, scelta eventualmente con criteri probabilistici. Se sono possibili ingressi e uscite dall' esterno, la rete si dice aperta; se invece ciò non è possibile e quindi il numero di clienti che fluiscono nella rete rimane costante, allora la rete si dice chiusa.

Studiare una rete di code significa definire e studiare il suo stato, che è dato dall' unione degli stati di ciascun centro di servizio o nodo, rappresentato, per code markoviane, dal solo vettore di variabili casuali discrete del numero di utenti presenti presso ciascuna risorsa (per la proprietà di mancanza di memoria).

In questa tesi per semplicità verranno trattate solamente le reti di code aperte, in quanto è la classe di reti di code che meglio rappresenta una linea produttiva.

4.2: Reti aperte

Una rete di code aperta è un insieme di M stazioni interconnesse da un sistema di trasporto, nel quale possono avvenire ingressi di clienti nel sistema, uscite dal sistema, o trasferimenti di clienti da una stazione ad un' altra. La scelta della stazione dove il cliente si pone può essere stocastica o deterministica. Un esempio di scelta deterministica si ha quando un pezzo deve seguire una successione prefissata di macchine, ognuna delle quali è preposta ad eseguire una particolare operazione (taglio, fresatura, tornitura…); in altre parole, dopo aver subito una certa operazione, il pezzo ne dovrà subire un' altra, determinata con certezza. Nel caso stocastico invece, al termine di un' operazione, la prossima lavorazione può essere effettuata su più di una macchina in alternativa ed è quindi necessario decidere quale (si parla in questo caso di instradamento o routing). Si pensi ad esempio al caso in cui l' instradamento del pezzo al termine di un' operazione dipende dall' esito della stessa: se un prodotto passa attraverso una stazione di collaudo o di controllo qualità, può essere rispedito indietro se risultasse difettoso (creando così dei ricircoli di pezzi all'interno della rete) oppure può procedere con le successive operazioni se invece non presentasse alcun problema. Tipicamente si riesce a stimare con buona precisione le probabilità che si verifichi un caso o l'altro.
Per le reti di code, come nel caso di code isolate, è importante stabilire le condizioni che permettono di ottenere soluzioni analitiche al problema. Data la complessità del modello dinamico di una rete di code, ci si ferma normalmente alla sua soluzione stazionaria.
La condizione di base per trovare una soluzione stazionaria della rete è che ciascuna risorsa continui a comportarsi come una coda Markoviana, isolata ed indipendente dalle altre, e che la probabilità congiunta degli stati della rete sia il prodotto delle probabilità di ciascuna delle risorse che la compongono (si parla di soluzione in forma prodotto). Questa proprietà è verificata per alcune classi di reti di code. La classe più importante è certamente costituita dalle reti di code Markoviane, dove gli arrivi alla rete sono processi di Poisson, i servizi forniti hanno tempi con distribuzione esponenziale e gli instradamenti sono casuali.
Quindi nella trattazione che segue si suppone per semplicità che un solo tipo di pezzo viene lavorato nel sistema; inoltre si farà riferimento a:

- intertempi di arrivo a ciascuna stazione j esponenziali con parametro λ_j;
- tempi di servizio di ciascun servente nella stazione j anch'essi distribuiti esponenzialmente con parametro μ_j;
- code con disciplina di servizio FIFO;
- buffer di ingresso con dimensioni illimitate per ciascuna stazione j;

- sistema con topologia probabilistica, ovvero un utente si muove tra i nodi in accordo ad una distribuzione di probabilità: un cliente in uscita dalla stazione j ha probabilità p_{jk} di essere instradato alla stazione k e probabilità p_{jo} di uscire dal sistema.

La rete di code che possiede queste caratteristiche è abbastanza generale: arrivi dall'esterno possono giungere ad una qualsiasi stazione così come i clienti possono uscire dal sistema da una qualunque stazione; inoltre sono ammessi ricircoli di materiale nella rete. In ogni caso i pezzi che arrivano ad un centro di lavoro vengono tutti trattati allo stesso modo, cioè con tempo di servizio $1/\mu_j$, indipendentemente dal fatto che essi visitino la stazione j per la prima volta o no.
E' comunque da sottolineare che la validità modellistica di questa classe di reti di code dipende in maniera fondamentale dalle ipotesi sopraelencate ed in particolare da tempi di servizio ed intertempi d' arrivo esponenziali; spesso invece i tempi di lavoro nelle stazioni si possono considerare con buona approssimazione deterministici. Questo richiederebbe di costruire le reti con code diverse da quelle poissoniane (cioè M/M/1, M/M/s ecc.) con conseguente aumento della complessità di risoluzione del modello. Comunque, sono stati effettuati studi a riguardo per definire delle distribuzioni esponenziali "equivalenti" a reti con distribuzioni più generali (Yao e Buzacott,1986) e la maggior precisione ottenuta nei risultati non era tale da giustificare lo sforzo computazionale aggiuntivo.

4.3: Concetti preliminari

I concetti qui presentati sono fondamentali per poter studiare le reti di code aperte.

1. **Legame distribuzione esponenziale – distribuzione di Poisson**: come detto in precedenza (paragrafo 2.3), se il numero di arrivi in un intervallo di tempo di lunghezza t ha distribuzione poissoniana, il tempo di interarrivo alla stazione ha distribuzione esponenziale.

2. **Composizione di più processi di Poisson**: la somma di variabili aleatorie poissoniane di parametro $\lambda_1, \lambda_2,..., \lambda_n$ rispettivamente è una variabile aleatoria poissoniana di parametro $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.

Per semplicità si dimostra il caso n=2. Si considerino due processi di Poisson indipendenti, P_1 con parametro λ_1 e P_2 con parametro λ_2 , che si compongono per dare origine ad un unico processo di arrivo. Iniziando l' osservazione dall' ultimo evento accaduto, il prossimo evento apparterrà al processo con il minimo residuo di vita.

Per la proprietà di assenza di memoria i residui di vita dei due processi, indipendentemente dal tempo già trascorso, avranno probabilità:

$$P(X_1 \leq t) = 1 - e^{-\lambda 1 t} \qquad \text{e} \qquad P(X_2 \leq t) = 1 - e^{-\lambda 2 t}$$

quindi la probabilità dell'intertempo fra due eventi successivi sarà uguale a quella del minimo fra i due tempi:

$$P(X \leq t) = 1 - (\min_{i=1,2} X_i > t)$$

Poichè i due processi sono indipendenti la probabilità degli intertempi risulta il prodotto delle probabilità:

$$P(X \leq t) = 1 - \prod_{i=1,2} P(X_i > t) = 1 - \prod_{i=1,2} e^{-\lambda i t} = 1 - e^{-(\lambda 1 + \lambda 2) t}$$

L' estensione al caso di n processi è analoga.

3. **Teorema di Burke**: Il processo delle uscite da un sistema M/M/s con coda di capacità infinita è un processo poissoniano di parametro λ uguale a quello che caratterizza il processo degli arrivi a tale sistema.

Ciò significa che il processo di uscita ha le stesse caratteristiche del processo di entrata.

Si dimostra per semplicità il caso s = 1. Innanzitutto è chiaro che i tassi di ingresso ed uscita devono essere uguali, in quanto se la frequenza degli arrivi λ e quella delle partenze μ fossero differenti il numero medio di utenti in coda non sarebbe stazionario. Si consideri ora un cliente che esce dal sistema all' istante generico t. Si possono a questo punto verificare due

situazioni: nella prima c' è almeno un altro cliente presente nel sistema, per cui il prossimo evento in uscita avverrà soltanto dopo un tempo di servizio, quindi la densità di probabilità del tempo di inter-uscita è:

$$p(t_{usc}) = p(t_{ser}) = \mu\, e^{-\mu tser}$$

Nella seconda situazione invece il sistema é vuoto, ovvero entra nello stato 0; in tal caso per il prossimo evento in uscita si dovrà attendere l' arrivo di un nuovo cliente e quindi l' espletamento del servizio. Il tempo di inter-uscita sarà quindi pari alla somma delle due variabili aleatorie indipendenti t_{arr} e t_{ser} (vedere composizione processi Poisson):

$p(t_{usc}) = p(tarr) * p(t_{ser}) = \lambda\, e^{-\lambda tarr} * \mu\, e^{-\mu tser}$

Il secondo caso si presenta con probabilità $p_o = 1 - \rho$, quindi il primo caso si presenta con probabilità ρ. Applicando il teorema della probabilità totale (dato un insieme A_i finito o numerabile di eventi a due a due incompatibili, la probabilità dell'unione di tutti gli eventi è uguale alla somma delle probabilità degli eventi) si ottiene la tesi.

4.4: Reti di code in forma prodotto

Si consideri una rete con M nodi e con tutte le caratteristiche specificate in precedenza e sia γ_i il parametro del processo di Poisson di arrivo dall'esterno al nodo i. Applicando il teorema di Burke e la proprietà di composizione di processi Poissoniani indipendenti (si veda il paragrafo 4.3) si ricava immediatamente che il parametro del processo di Poisson totale in ingresso al nodo i può essere ottenuto risolvendo il sistema:

$$\lambda_i = \gamma_i + \sum_{j=1}^{i-1} \lambda_j\, p_{ji} \qquad 1 \le i \le M$$

Sotto le ipotesi fatte, definendo $n = (n_1, n_2, \ldots, n_M)$ lo stato del sistema, si associa alla rete un processo stocastico Markoviano con distribuzione stazionaria di stato $\pi(n_1, n_2, \ldots, n_M)$. Se vale la condizione di stazionarietà, possiamo scrivere:

$$\pi(n_1, n_2, \ldots, n_M) = \prod_{i=1}^{M} \mathrm{Prob}_i\,\{n_i\}$$

dove $\mathrm{Prob}_i\,\{n_i\} = \rho_i^{ni}(1-\rho_i)$, $\quad n_i \ge 0$, $\quad \rho_i = \lambda_i/\mu_i < 1$, $\quad 1 \le i \le M$.

Questa classe di reti è un esempio di modelli di reti in forma prodotto.

La distribuzione congiunta $\pi(n_1, n_2, \ldots, n_M)$ è esprimibile come prodotto delle distribuzioni relative ai singoli nodi analizzati come sistemi isolati di tipo M/M/1, dove il nodo i è un sistema M/M/1 con parametri λ_i e μ_i. Da tale distribuzione di stato si ricavano gli altri indici

di prestazione (tutto ciò si può estendere a reti di code con centri di servizio esponenziali e serventi multipli s, analizzando ogni centro come un sistema M/M/s isolato).
Questo fondamentale risultato venne formulato nel 1957 attraverso il **teorema di Jackson**, basilare per poter risolvere le reti di code aperte:
"Se è soddisfatta la condizione di stazionarietà $\rho_i = \lambda_i / m_i\mu_i < 1$, $1 \leq i \leq M$, allora la distribuzione stazionaria di stato di una rete di Jackson si esprime come:

$$\pi(n_1, n_2,\ldots, n_M) = \prod_{i=1}^{M} \pi_i (n_i)$$

dove:

$\pi_i (n_i) = \pi_i (0) \ (m_i \rho_i)^n / n!$ per $1 \leq n \leq m_i$

$\pi_i (n_i) = \pi_i (0) \ m_i^{mi} \rho_i^{\ n} / n!$ per $n > m_i$

e $\pi_i (0)$ è ottenuto dalla condizione $\sum_i \pi_i (n_i) = 1$.

La dimostrazione è omessa in quanto non significativa ai fini di codesta tesi; verrà invece analizzata l' utilità applicativa di questo teorema.
Tale utilità è principalmente rappresentata dal fatto che i termini $\pi_i(n_i)$ che compaiono nell'espressione della probabilità di stato $\pi(n_1, n_2,\ldots, n_M)$ sono esattamente gli stessi che esprimono la probabilità di avere n_j clienti nella stazione M_j considerata singolarmente (con parametro di arrivo λj e parametro di servizio μ_j).
L' espressione di $\pi(n_1, n_2,\ldots, n_M)$ costituisce quella che, come spiegato in precedenza, prende il nome di soluzione in forma prodotto, in quanto la probabilità di stato si esprime come prodotto delle probabilità che le singole stazioni si trovino nei rispettivi stati n_j indipendentemente da ciò che avviene nelle altre stazioni.
Questo implica che l' algoritmo risolutivo di questa classe di reti, dette appunto reti di Jackson, si sviluppi in tre passi:

- determinare le frequenze effettive di arrivo alle stazioni tramite:

  $$\lambda_i = \gamma_i + \sum_{j=1}^{i-1} \lambda_j \, p_{ji} \qquad 1 \leq i \leq M$$

- analizzare ogni stazione indipendentemente dalle altre e verificare che per ciascuna valga la condizione di stazionarietà

  $$\rho_i = \lambda_i / m_i\mu_i < 1 \qquad 1 \leq i \leq M$$

- aggregare i risultati per avere indicazioni sul sistema complessivo a partire dalla probabilità di stato congiunta

$$\pi(n_1, n_2, \ldots, n_M) = \prod_{i=1}^{M} \pi_i(n_i)$$

Come conseguenza a quanto detto fino ad ora, è chiaro che il fatto che le stazioni si comportino come centri di servizio indipendenti facilita il calcolo dei parametri di maggior interesse relativi al funzionamento globale del sistema.

Innanzitutto il numero medio di clienti nel sistema $\mathbf{L_s} = N$ si può calcolare sommando i valori N_j relativi alle singole stazioni: $\mathbf{L_s} = N = \sum_{j=1}^{M} N_j$

Il tempo medio di attraversamento del sistema $\mathbf{W_s}$ si può calcolare applicando la formula di Little (che risulta valida non solo per una singola coda ma anche per una rete di code, nel senso che il numero medio di utenti presenti nella rete è sempre dato dal prodotto del tempo medio di permanenza degli utenti al suo interno per la somma di tutti i flussi entranti):

$\mathbf{W_s} = Ls / \lambda = N / \lambda$ in cui $\lambda = \sum_{j=1}^{M} \lambda_j$

CAPITOLO 5: Approccio simulativo

5.1: Introduzione

Oltre all' approccio analitico trattato fino a questo punto, ovvero utilizzando la teoria delle code, esiste anche un approccio simulativo che permette di poter risolvere questi tipi di problema. Tale approccio può essere condotto attraverso la cosiddetta simulazione numerica ed in particolare attraverso il metodo Monte Carlo.

La metodologia Monte Carlo, dovuta al francese George-Louis Leclerc Buffon e poi ulteriormente sviluppata da John Von Neumann ed Enrico Fermi, è una tecnica nata per approssimare la soluzione di un problema, creando delle distribuzioni statistiche a partire da numeri generati a caso (random), in modo da imitare l' aleatorietà insita nel problema originale che si sta studiando. Da questo emerge quindi la particolare utilità del metodo nelle situazioni che richiedono modelli di tipo stocastico, ed in particolare nel caso di problemi decisionali complessi comprendenti numerose variabili aleatorie, che possono essere anche dipendenti l'una dall'altra (in tali casi i procedimenti analitici sono quasi sempre non applicabili). Un pregio di questa tecnica è che consente di ottenere una descrizione statistica completa, ancorché approssimata, del problema oggetto dello studio, dalla quale poi si possono ricavare media, varianza ed altri indicatori sintetici utili all' analista.

Con i numeri casuali è possibile rappresentare un qualsiasi fenomeno aleatorio, anche quando esso non rientri in una legge statistica conosciuta. Partendo da rilevazioni sperimentali su un campione sufficientemente consistente degli oggetti generanti i valori delle variabili aleatorie in ingresso al modello, si può associare dei valori "random" ai valori di tali variabili e quindi studiarne gli effetti sul nostro sistema. Questa regola di transizione fra il piano della realtà e il piano dei numeri casuali è il vero cuore del metodo, in quanto è fondamentale che la generazione casuale dei valori delle variabili di ingresso riproduca in maniera accurata l' andamento probabilistico del fenomeno in esame.

Per condurre quindi questo tipo di analisi, il primo passo consiste nel costruire un modello analitico che rappresenti la reale situazione decisionale (ciò può essere anche molto complesso). Si procede poi con la ricostruzione della distribuzione di probabilità di ciascuna variabile aleatoria, sulla base di dati soggettivi o storici; proprio a partire da queste distribuzioni viene generato in modo casuale un insieme di valori di ingresso, che poi viene utilizzato per determinare il risultato di un esperimento (simulazione) del modello.

Dalla ripetizione di questo processo per un elevato numero di volte è possibile ottenere una serie di risultati degli esperimenti, espressi tramite i valori assunti dall'indicatore prescelto come variabile di uscita del modello; da tale serie si possono ricostruire le frequenze relative dei valori di uscita e quindi la densità di probabilità dell' indicatore, ottenendo in tal modo una descrizione statistica completa, sebbene approssimata, del problema in esame.

5.2: Fondamenti statistici del metodo

Sia X una variabile aleatoria discreta che può assumere le modalità x_i, ed f(x) la sua densità di probabilità. Noto un evento x^*, $f(x^*)$ rappresenta la sua probabilità di accadimento; invece

$F(x^*) = \int_{-\infty}^{x^*} f(t)dt$ rappresenta la probabilità cumulata, ovvero la probabilità che accada x* o uno degli eventi che lo precedono. F(x) viene anche detta funzione integrale; essa presenta un asintoto a $+\infty$ e tale asintoto vale 1. Essendo una funzione integrale, per definizione essa è invertibile sempre in quanto monotona non decrescente. Allora se si genera un numero random compreso tra 0 ed 1 (tali numeri sono uniformemente distribuiti tra 0 ed 1 e ciascuno ha la stessa probabilità di essere generato di tutti gli altri), invertendo la funzione integrale si ottiene $x^* = F^{-1}(RND)$ ovvero un evento x^* che rispetta sia l'aleatorietà che la distribuzione di probabilità di partenza (si sfrutta la Probability Integral Transormation, proprietà che permette di convertire una variabile aleatoria con qualsiasi distribuzione continua in una variabile aleatoria con una distribuzione uniforme).

5.3: Dimensionamento del buffer

Quando si dimensiona un buffer, si deve fare in modo che esso abbia una capienza sufficiente alle necessità, ovvero deve avere una dimensione tale da poter sopportare una oscillazione del proprio livello di riempimento compresa tra il valore massimo ed il valore minimo raggiungibili. In tal modo si evita che, in certe situazioni quali guasti ed interruzioni nel lavoro di una stazione lungo la linea produttiva, il buffer si trovi ad essere pieno o vuoto costringendo così l' intera linea ad arrestarsi. Si rendono cioè indipendenti le stazioni una dall'altra, nel senso che si assicura che ciscuna continui a lavorare (prelevando o accumulando i pezzi nel buffer) a prescindere da quello che fanno le altre (**linea asincrona**). Quindi, detta $\mathbf{Q_{buffer}}$ la dimensione del buffer, vale: $\mathbf{Q_{buffer}} = Liv_{max} - Liv_{min}$

Si noti come il livello di riempimento di partenza del buffer non abbia alcuna rilevanza nel dimensionamento, in quanto ciò che conta è solamente la massima oscillazione che esso può subire.

La determinazione di tale oscillazione si effettua a partire dalla simulazione del comportamento delle stazioni di lavoro immediatamente a monte e a valle del buffer, che si effettua applicando il metodo MonteCarlo.

Si può rappresentare graficamente il problema in questo modo:

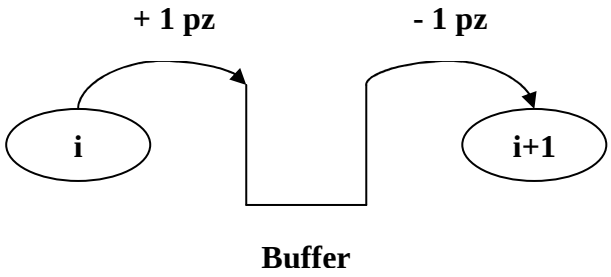


Fig. n.5.3.1: Rappresentazione grafica

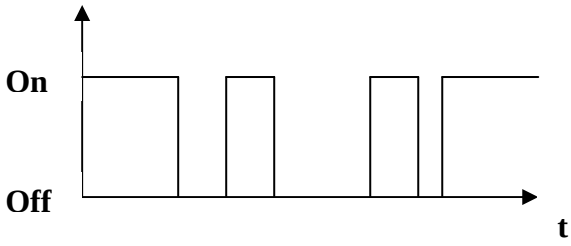


Fig. n.5.3.2: Comportamento operatore **i** a monte del buffer

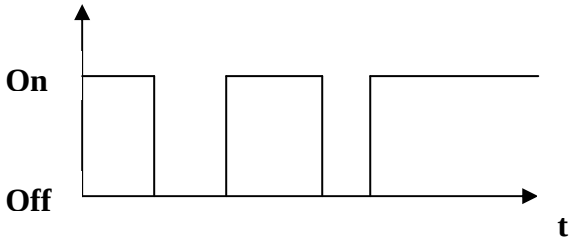


Fig. n.5.3.3: Comportamento operatore **i + 1** a valle del buffer

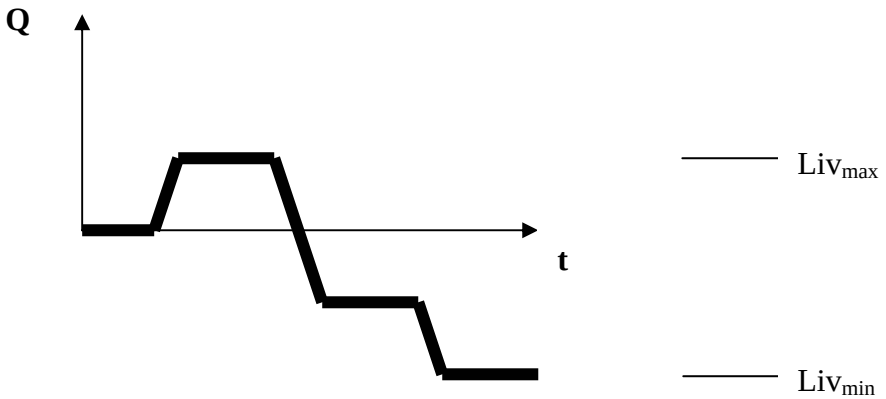


Fig. n.5.3.4: Livello di riempimento del buffer

La stazione di lavoro **i** a monte del buffer accumula pezzi in esso; la stazione di lavoro **i+1** a valle preleva pezzi da esso. Nelle figure **n.5.3.2** e **n.5.3.3** sono rappresentati i comportamenti delle rispettive stazioni; On significa che esse stanno lavorando regolarmente, mentre Off significa che per un qualche motivo il loro lavoro si è interrotto (guasti, inefficienze dell' operatore ecc.).
Tali comportamenti sono soggetti ad aleatorietà, in quanto non è possibile prevedere quando la stazione sta lavorando e quando no; proprio per questo si utilizza il metodo MonteCarlo.
Il grafico di figura **n.5.3.4** è quello da cui si ricava la dimensione corretta del buffer e si ottiene dalla sovrapposizione dei due grafici precedenti: quando entrambe le stazioni di lavoro sono attive, il livello di riempimento rimane costante; quando lavora solo la stazione i, il livello cresce; quando invece lavora solamente la stazione i+1, il livello cala.
Si effettua una simulazione ottenendo come risultato $Q_{buffer} = Liv_{max} - Liv_{min}$ dove il livello massimo e minimo fanno riferimento alla singola simulazione; si effettua quindi un gran numero di esperimenti e si dimensionerà il buffer sulla base della massima Q_{buffer} ottenuta.
Per poter utilizzare questo metodo ci si avvale del supporto di un calcolatore, in particolare si può utilizzare un foglio elettronico.

CAPITOLO 6: Esempio applicativo

6.1: Introduzione all' esempio

Si prenda come esempio il seguente tratto di un impianto manufatturiero costituito da due centri di lavorazione in serie, M_1 ed M_2 (in entrambe lavora un solo operaio). Tutti i pezzi grezzi che arrivano devono subire una lavorazione su entrambe le macchine; i dati sui processi di arrivo e di servizio sono riportati in figura n.6.1.1. L' obiettivo è quello di dimensionare correttamente il buffer interoperazionale tra le due stazioni; ciò verrà fatto seguendo l'approccio analitico prima e poi quello simulativo.

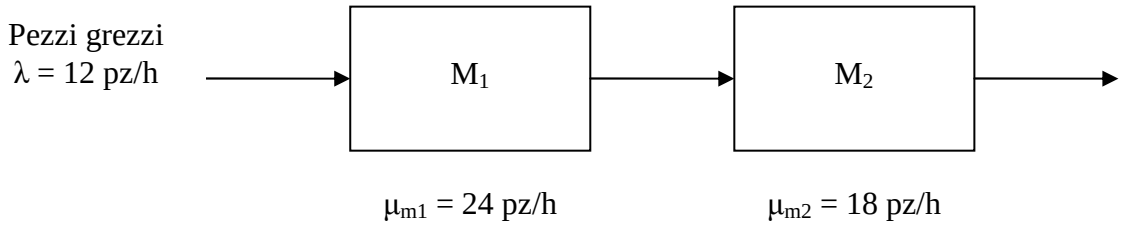


Fig. n.6.1.1: Schema dell' impianto

6.2: Dimensionamento con approccio analitico

Per poter utilizzare la teoria delle code, si deve assumere che gli intertempi di arrivo siano distribuiti esponenzialmente, così come deve essere per i tempi di servizio sia di M_1 che di M_2; ciascuna stazione poi è dotata di un buffer in ingresso illimitato e nell' impianto viene lavorato un solo tipo di prodotto.

Questo è un semplice esempio di sistema seriale, poiché l' uscita da una stazione costituisce l' ingresso nella successiva (nella terminologia relativa alla teoria delle code vengono anche detti sistemi tandem).

Innanzitutto siccome il processo degli arrivi a M_1 è poissoniano di parametro $\lambda = 12$ pz/h, anche quello degli arrivi a M_2 è poissoniano con lo stesso parametro, come ci assicura il teorema di Burke (ciò ha anche una giustificazione logica in quanto basta pensare che tutti i pezzi che entrano nell' impianto devono essere lavorati in entrambi i centri senza perdite né aggiunte).

Le capacità di servizio delle due stazioni sono invece diverse, rispettivamente $\mu_{m1} = 24$ pz/h

e μ_{m2} = 18 pz/h. Verifichiamo subito che $\lambda / \mu_{m1} < 1$ e $\lambda / \mu_{m2} < 1$; si ricorda che queste sono due condizioni fondamentali per la stazionarietà della coda ovvero per poter ottenere una distribuzione stazionaria di probabilità. Il fattore di utilizzazione dei servitori vale rispettivamente:

$\boldsymbol{\rho_{m1}}$ = 12/24 = 0,5 e $\boldsymbol{\rho_{m2}}$ = 12/18 = 0,666667.

Si è già visto che, sotto le ipotesi assunte, si possono studiare i due processi M_1 ed M_2 separatamente ed indipendentemente uno dall' altro (come corollario del teorema di Jackson). Allora il numero medio di clienti nel sistema (sia in attesa di servizio che riceventi servizio) $L_s = \rho / (1 - \rho)$ vale rispettivamente:

$\boldsymbol{L_{sm1}}$ = 0,5 / (1 – 0,5) = 1 pz e

$\boldsymbol{L_{sm2}}$ = 0,666667 / (1 – 0,666667) = 2 pz

e nel complesso sommando i valori relativi alle singole stazioni si ha $\boldsymbol{L_{stot}}$ = 1 + 0,5 = 1,5 pz. Conseguentemente il numero medio di clienti in attesa di servizio $L_q = L_s - \rho$ vale rispettivamente:

$\boldsymbol{L_{qm1}}$ = 1 – 0,5 = 0,5 pz e

$\boldsymbol{L_{qm2}}$ = 2 - 0,666667 = 1,333333 pz

A partire da questi dati basta applicare la legge di Little per poter ricavare il tempo di attesa medio dei clienti nel sistema (sia in attesa di servizio che riceventi servizio) $Ws = 1 / (\mu - \lambda) = L_s / \lambda$:

$\boldsymbol{W_{sm1}}$ = 1 / (24 − 12) = 1 / 12 = 0,083333 h e

$\boldsymbol{W_{sm2}}$ = 1 / (18 − 12) = 2 / 12 = 0,166667 h

Per poter ricavare ora la giusta dimensione del buffer interoperazionale, si deve considerare il risultato principale fra quelli ricavati, ovvero il numero medio dei clienti in coda in attesa di servizio tra le due stazioni (cioè a monte di M_2) $\boldsymbol{L_{qm2}}$ = 1,333333 pz. Non basta però semplicemente prendere tale numero come capienza del buffer, perché, essendo un dato medio, nel 50% dei casi si avranno meno pezzi in coda, nel restante 50% dei casi invece ce ne saranno di più. Quindi è necessario cautelarsi con un coefficiente ζ.

Prendiamo ad esempio ζ = 1,5; risulterà che la dimensione corretta del buffer è 1,333333 * 1,5 = 2 pz.

Prendendo invece un coefficiente maggiore (ci si cautela di più), ζ = 2, si ottiene 2,666667, ovvero una dimensione del buffer pari a 3.

6.3: Dimensionamento con approccio simulativo

Per risolvere il problema con questo approccio ci si avvale di un foglio elettronico (Excel), di cui ho riportato una parte nella figura sottostante.

Anche in questo caso si devono assumere tempi di servizio esponenziali; siccome il processo degli arrivi a M_1 è poissoniano di parametro $\lambda = 12$ pz/h, anche quello degli arrivi a M_2 è poissoniano con lo stesso parametro, come ci assicura il teorema di Burke. Ciò significa che entrambe le stazioni lavorano ad un tasso di servizio pari a 12 pz/h = 0,2 pz/min; quindi il tempo di lavoro di M_1 ed M_2 ha media 5 min/pz ed è distribuito esponenzialmente con parametro $\lambda = 0,2$ pz/min.

Si procede prendendo alcuni valori per il tempo di lavorazione attorno al valore medio 5 min e si calcola la loro probabilità secondo la distribuzione esponenziale di parametro $\lambda = 0,2$ ed anche la probabilità cumulata (prime due tabelle di fig. n.6.3.1). Proprio ai valori di probabilità cumulata si associano gli intervalli di valori "random". Ad esempio se si prende come primo valore del tempo di lavoro 4 minuti ed esso ha probabilità cumulata 0,5506, allora a tale tempo si associa l' intervallo di valori random da 0 a 5505; il ragionamento va ripetuto con tutti gli altri valori del tempo di servizio.

Si considera poi una sequenza di pezzi (nell' esempio 90, ma in figura è riportata la sequenza solo fino a 10), e per ogni pezzo si genera un numero casuale sia per M_1 sia per M_2, a cui corrisponderanno dei tempi di lavoro secondo gli intervalli stabiliti in precedenza (si è in tal modo simulato il comportamento delle due stazioni, riportato nelle colonne 4 e 5 della terza tabella in figura n.6.3.1). Il livello di riempimento del buffer durante la sequenza dei pezzi si ricava arrotondando per eccesso il risultato della formula [sum(Tm1) – sum(Tm2)] / Tmedio; a partire da ciò non resta altro che fare la differenza tra il livello massimo ed il livello minimo di riempimento raggiunti nella singola simulazione che si sta svolgendo, ripetere l' esperimento per un gran numero di volte e prendere la massima Q_{buffer} o la media Q_{buffer} ottenuta.

In questo caso la massima dimensione ottenuta è 10, ma la dimensione media calcolata su 50 simulazioni è 5.

Tm1[min]	f(T)	CUM_F
4	0,089866	0,550671
4,5	0,081314	0,59343
5	0,073576	0,632121
5,5	0,066574	0,667129
6	0,060239	0,698806
6,5	0,054506	0,727468
7	0,049319	0,753403

Tm2[min]	f(T)	CUM_F
4	0,0898658	0,550671
4,5	0,0813139	0,5934303
5	0,0735759	0,6321206
5,5	0,0665742	0,6671289
6	0,0602388	0,6988058
6,5	0,0545064	0,7274682
7	0,0493194	0,753403

Qbuffer
5

Pz	RND_m1	RND_m2	Tm1[min]	Tm2 [min]	Sum(Tm1)	Sum(Tm2)		Buffer
1	2800	5363	4	4	4	4	0	0
2	1070	6021	4	5	8	9	0,2	1
3	1811	9037	4	7	12	16	0,8	1
4	8934	9443	7	7	19	23	0,8	1
5	2702	8408	4	7	23	30	1,4	2
6	6300	2155	5	4	28	34	1,2	2
7	9491	6445	7	5,5	35	39,5	0,9	1
8	9859	5834	7	4,5	42	44	0,4	1
9	9184	8621	7	7	49	51	0,4	1
10	4510	394	4	4	53	55	0,4	1

Fig. n.6.3.1: parte del foglio elettronico utilizzato per la simulazione MonteCarlo

CONCLUSIONI

Come già accennato nell' introduzione, data l'ampiezza dell'argomento "teoria delle code", nella stesura della tesi è stato fondamentale individuare gli aspetti principali coerenti con l' obiettivo dell' elaborato e suddividere conseguentemente l'intero lavoro in sezioni che entrino in maniera crescente nello specifico della teoria, dal primo al quarto capitolo.
Una ampia presentazione è stata doverosa, riguardante i principali componenti della teoria delle code, materia tanto diffusa nella vita di tutti i giorni quanto specifica e complessa.
La trattazione del metodo MonteCarlo invece è stata portata avanti entrando nello specifico fin da subito, nonostante anche in questo caso sia stata presentata la sua applicabilità in svariati contesti.

Confrontando i due approcci di dimensionamento del buffer, si può capire come la teoria delle code sia un metodo molto efficace per ottenere una descrizione in termini medi del comportamento del sistema in esame; essendo un metodo analitico però richiede una serie di ipotesi piuttosto forti che devono essere verificate affinché i risultati che si ottengono si possano considerare validi.
Una volta che si è certi di questo, si devono semplicemente sostituire i dati del sistema nelle formule degli indici di prestazione e si arriva velocemente al risultato cercato.
Per quanto riguarda la metodologia MonteCarlo invece, essendo un metodo di simulazione numerica non richiede particolari ipotesi ed anzi si può utilizzare anche non sapendo quale sia la distribuzione di probabilità delle variabili in ingresso al modello,purchè si abbiano a disposizione dei dati storici o soggettivi su di esse.
Occorre poi effettuare un gran numero di simulazioni per ottenere un risultato attendibile ma questo grazie ai computer e ai software disponibili richiede pochissimo sforzo.

Una azienda che intenda quindi dimensionare i buffer interoperazionali lungo la propria linea di produzione può avvalersi di entrambi i metodi; si potrebbero poi ampliare i risultati ottenuti dalla teoria delle code andando a modificare le assunzioni del modello, come ad esempio sostituire le distribuzioni degli intertempi di arrivo e dei tempi di servizio, ma ciò andrebbe a modificare drasticamente lo sviluppo della soluzione analitica e quindi non ho ritenuto opportuno sviluppare tali temi all' interno di questa tesi.

BIBLIOGRAFIA

Libri di riferimento:

- William G. Sullivan, Elin M.Wicks, James T. Luxhoj, 2006, “Economia applicata all’ ingegneria”, prima edizione italiana, Pearson Education Italia srl

- D.M. Levine, T.C. Krehbiel, M.L. Berenson, 2006, “Statistica”, seconda edizione, Apogeo

- U. Narayan Bhat, 2008, “An Introduction to Queueing Theory: Modeling and Analysis in Applications”, Birkhäuser Boston

- A. Pareschi, A. Persona, E. Ferrari, 2002, “Logistica integrata e flessibile. Per i sistemi produttivi dell'industria e del terziario”, Esculapio

Fonti internet di supporto:

- R. Pesenti, “Teoria delle code o file d’attesa” [online]. Disponibile su <ftp://docenti.ing.units.it/arc_stud/Pesenti/Nettuno/CodeDispense.pdf> [Data di accesso 05/07/13].

- A. Agnetis, “Introduzione alle reti di code nei sistemi manifatturieri” [online]. Disponibile su <http://www.dia.uniroma3.it/~adacher/automazione1/RetiCode.pdf>
[Data di accesso 18/07/13].

- “Teoria della probabilità e teoria delle code” [online]. Disponibile su <www.dia.uniroma3.it/~adacher/automazione1/**TeoriaCode**.pdf>
[Data di accesso 02/07/13].

- S. Balsamo, “Capitolo 4 - Modelli a rete di code” [online]. Disponibile su <www.dsi.unive.it/~balsamo/disp.pdf/Cap4.pdf> [Data di accesso 22/08/2013].

- S. Balsamo, “Capitolo 3 - Modelli a singola coda” [online]. Disponibile su <www.dsi.unive.it/~balsamo/disp.pdf/Cap3.pdf> [Data di accesso 12/09/2013].

- “La teorie delle code” [online]. Disponibile su <https://www.universibo.**unibo**.it/file/1915/download/> [Data di accesso 02/07/2013].

- “Programmazione e controllo della produzione” [online]. Disponibile su <http://dma.dima.uniroma1.it:8080/users/boschetto/PCPM/Lucidi/PCP%201011%2003%20Analisi%20dei%20flussi.pdf> [Data di accesso 18/07/2013].

- F. Lo Piccolo, “Reti di Telecomunicazioni - sistemi a coda M/M/1” [online]. Disponibile su <http://www.uniroma2.it/didattica/frs/deposito/sistemi-a-coda-MM1.pdf> [Data di accesso 11/09/2013].

- “Processi di nascita e morte” [online]. Disponibile su <http://www.dis.uniroma1.it/~roma/didattica/SSS09-10/parteC.pdf> [Data di accesso 11/09/2013].