Appunti di Sistemi Operativi

Enzo Mumolo

e-mail address :mumolo@units.it
web address :www.units.it/mumolo

Indice

0.1 Introduzione

In questo capitolo sono riportati alcuni richiami su concetti fondamentali sulla analisi analitica dei sistemi di coda d'attesa per applicazioni nella analisi e dimensionamento di alcuni aspetti dei Sistemi Operativi.

0.2 Richiami sulle variabili aleatorie

Il discorso sulle variabili aleatorie parte dal concetto di esperimento casuale: ossia un esperimento dal quale si ottiene in risultato *incerto* (ovvero un risultato *non* regolato da relazioni deterministiche, quindi non é possibile calcolare il risultato ma quest'ultimo soddisfa a delle caratterisiche medie). Il risultato dell'esperimento ha valori nell'insieme $\omega \in \Omega$ con Ω spazio dei campioni.

Esempi:

- Se l'esperimento casuale é il lancio di un dado allora $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Se l'esperimento casuale é il lancio di una moneta allora $\Omega = \{T, C\}$
- Se l'esperimento casuale é la misurazione della temperatura allora $\Omega \subset \mathbb{R}$

Nel nostro caso, al fine di semplificare il discorso, possiamo sempre far riferimento al campo reale in quanto nulla ci vieta di considerare una variabile $X(\omega) : \Omega \to \mathbb{R}$ che una variabile che prende i valori da Ω e fornisce valori in $\mathbb{R}$ (Es.: lanciando un dado l'evento $5 \in \mathbb{N}$ diventa $5 \in \mathbb{R}$). Cosí facendo X é una variabile aleatoria reale; in queste note si fará sempre riferimento a variabili aleatorie appartenenti ad un sottoinsieme di $\mathbb{R}$.

Supponiamo che una variabile aleatoria X sia il risultato di una serie di esperimenti casuali e abbia valori compresi tra 0 e 4. Gli esperimenti casuali forniscono ad esempio i seguenti risultati: 1 3 4 2 1 3 3 1 2 2 3 2 0 2 1 2, cioé vengono fatti 16 esperimenti, dai quali il numero 3 esce quattro volte, il numero 2 esce sei volte e cosí via. Ció vuol dire che il numero 3 é associato alla probabilitá 4/16, il numero 2 alla probabilitá 6/16 e cos'ıvia. Questo puó essere visualizzato con l'istogramma delle probabilitá di fig. 2.1. L'altezza di ciascun rettangolo é la probabilitá associata alla variabile aleatoria X. Rappresentando $x = 0$ con l'intervallo da -0.5 a 0.5, $x = 1$ con l'intervallo da 0.5 a 1 e cosí via, é possibile associare le probabilitá alle aree dei rettangoli. Ci sono diversi vantaggi nell'associare l'area dell'istogramma - o distribuzione delle probabilitá - alla probabilitá. Per esempio, quando si approssima l'istogramma con curve continue, la probabilitá puó essere calcolata mediante integrali, che comportano una maggiore semplicitá degli sviluppi analitici.

0.2.1 Probabilitá condizionata

Denotiamo con $P(A)$ e $P(B)$ le probabilitá degli eventi A e B. Se l'avverarsi di A é condizionato all'avverarsi di B, si parla di probabilitá di A condizionata a B, e si indica con

$$P(A \mid B)$$

Nel caso di indipendenza statistica si ha che

$$P(A \cap B) = P(A)P(B)$$

Nel caso invece in cui ci sia dipendenza si ha che

$$P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

da cui segue la nota **Formula di Byes**

$$P(A) = \frac{P(A \mid B)}{P(B \mid A)} P(B)$$

Parlando dell'insieme dei numeri reali, l'evento $P(X = x)$ normalmente ha una probabilitá di valore infinitesimo; per ovviare a questo problema ha piú senso parlare della probabilitá che $X \leq x$: questo evento

ha cosí una probabilitá che non é piú infinitesimale. Questo evento viene chiamato funzione **cumulativa di distribuzione di probabilitá** e si denota con:

$$\mathrm{F}(X) = \mathrm{Prob}\{X \leq x\}$$

Ora diremo che la variabile aleatoria X ha **densit** $f(x)$ se

$$\mathrm{F}(X) = \int_{-\infty}^{x} f(\xi)\, d\xi$$

Naturalmente queste due funzioni sono legate dalla relazione:

$$f(x) = \frac{\partial \mathrm{F}(X)}{\partial x}$$

A questo punto possiamo valutare:

$$\mathrm{Prob}\{x_1 \leq X \leq x_2\} = \int_{x_1}^{x_2} f(\xi)\, d\xi$$

Ne consegue che $\int_{-\infty}^{+\infty} f(\xi)\, d\xi = 1$.

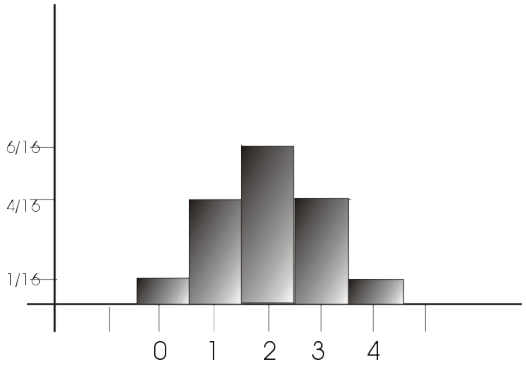


Figura 0.1 La probabilit associata all'area...

Possiamo anche vedere che:

$$\int_{x_1}^{x_2} f(\xi)\, d\xi = \int_{-\infty}^{x_2} f(\xi)\, d\xi - \int_{-\infty}^{x_1} f(\xi)\, d\xi = \mathrm{F}(X_2) - \mathrm{F}(X_1)$$

Per descrivere meglio queste funzioni si utilizzano dei momenti statistici; la descrizione statistica delle variabili aleatorie (definita dai momenti) serve a mettere in luce alcune loro caratteristiche particolari che possono essere utilizzate in questo contesto.

Definizione 1 *Si definisce* ***momento del k-esimo ordine*** *attorno alla variabile w il valore*

$$\int_{-\infty}^{+\infty} (x - w)^k f(x)\, dx$$

A questo punto siamo pronti per definire altri due concetti fondamentali.

Definizione 2 *Si definisce* ***speranza matematica*** *della variabile X il valore*

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \lambda_x$$

La speranza matematica altro non é il momento del primo ordine attorno allo zero. Appare subito chiaro, mettendo la definizione della speranza matematica in questa forma: $\mathrm{E}(X) = \frac{\int_{-\infty}^{+\infty} x\, f(x)\, dx}{\int_{-\infty}^{+\infty} f(x)\, dx}$, che il parametro λ_x puó essere interpretato come *centro di gravitá* o *baricentro* della distribuzione (vedi figura 2.2).

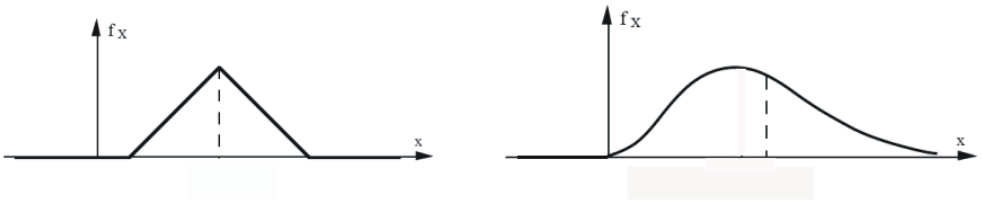


Figura 0.2 Due esempi di baricentro.

Definizione 3 *Si definisce* ***momento del secondo*** *ordine attorno a* λ_x *il valore*

$$\mathrm{var}^2(X) = \int_{-\infty}^{+\infty} (x - \lambda_x)^2 f(x)\, dx = \sigma_X^2$$

La radice quadrata di questa quantitá é lo **scarto quadratico medio**, σ_X.

Questi due momenti sono i piú semplici e i piú usati per descrivere queste distribuzioni e sono quelli che bastano per delineare moltissime caratteristiche delle variabili aleatorie.

Un primo esempio riguarda il noto significato fisico dello scarto quadratico medio di una distribuzione, cioé quello di dispersione della probabilitá intorno alla media. In altri termini possiamo dire che piú lo scarto quadratico medio é piccolo, piú la densitá é concentrata intorno alla media. Questo puó essere subito visto mediante il seguente Teorema, che vale per qualsiasi tipo di distribuzione.

Teorema di Bienaym-Chebyshev 0.2.1

$$\mathrm{Prob}\{\mid x - \lambda_x \mid \geq k\sigma\} \leq \frac{1}{k^2}$$

Distinguiamo i due casi:

- Se $\mid x - \lambda_x \mid > 0 \Rightarrow x - \lambda_x \geq k\sigma \Rightarrow x \geq \lambda_x + k\sigma$
- Se $\mid x - \lambda_x \mid < 0 \Rightarrow -(x - \lambda_x) \geq k\sigma \Rightarrow x \leq \lambda_x - k\sigma$

Questo significa che se rappresentiamo la funzione di distribuzione secondo una curva qualsiasi (vedi figura 2.3) allora la somma delle aree tratteggiate (le 'code' della distribuzione) é minore o uguale a $1/k^2$. Infatti, se moltiplichiamo la relazione presentata nella enunciazione del Teorema per -1 e se sommiamo 1 ad entrambi i termini, otteniamo

$$\mathrm{Prob}\{\mid x - \lambda_x \mid < k\sigma\} > \frac{k^2 - 1}{k^2}$$

ovvero $\mathrm{Prob}\{\lambda_x - k\sigma < x < \lambda_x + k\sigma\} > \frac{k^2-1}{k^2}$. Cioé, l'area della zona centrale in figura 2.3 é maggiore di una costante. Questo vuol dire che se σ diminuisce, l'altezza della curva aumenta. Quindi il significato dello scarto quadratico medio é quello di dispersione intorno alla media.

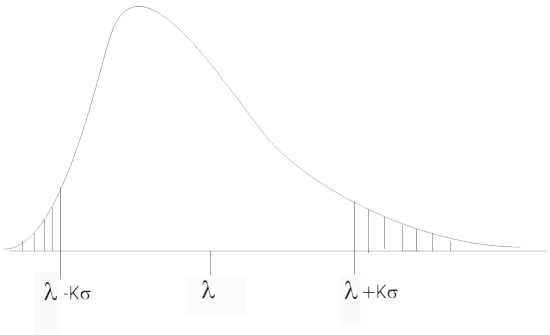


Figura 0.3 Esempio di distribuzione e diseguaglianza di Chebyshev

Ora, tornando alla varianza:

$$\begin{aligned}\mathrm{var}^2(X) &= \sigma^2(X) = \int_{-\infty}^{+\infty}(x^2+\lambda_x^2-2x\lambda_x)f(x)\,dx \\ &= \int_{-\infty}^{+\infty}x^2\,f(x)\,dx+\lambda_x^2\underbrace{\int_{-\infty}^{+\infty}f(x)\,dx}_{=1}-2\lambda_x\underbrace{\int_{-\infty}^{+\infty}x\,f(x)\,dx}_{=\lambda_x} \\ &= \int_{-\infty}^{+\infty}x^2\,f(x)\,dx+\lambda_x^2-2\lambda_x^2 = \mathrm{E}(X^2)-\mathrm{E}^2(X)\end{aligned}$$

ovvero: $\boxed{\mathrm{var}^2(X)=\mathrm{E}(X^2)-\mathrm{E}^2(X)}$

A questo punto complichiamo le cose; introduciamo un'altra variabile:

$$\mathrm{Prob}\{X\le x, Y\le y\} = \mathrm{F}(X,Y) = \int_{-\infty}^{x}\int_{-\infty}^{y}f(\xi_1,\xi_2)\,d\xi_1\,d\xi_2$$

Questa é un'estensione del caso a una sola variabile, di conseguenza le considerazioni fatte in precedenza valgono anche in questo caso: per esempio esiste un valor medio attorno allo zero, ovvero esiste una

$$\mathrm{E}(\phi(x,y)) = \iint_{-\infty}^{+\infty}\phi(x,y)f(x,y)\,dx\,dy$$

cosí come esiste una varianza:

$$\sigma^2_{\phi(x,y)} = \iint_{-\infty}^{+\infty}[\phi(x,y)-\mathrm{E}(\phi(x,y))]^2(f(x,y)\,dx\,dy$$

Il fatto di avere a che fare con due variabili aleatorie introduce due possibilitá che riguardano la dipendenza o l'indipendenza tra le due variabili; per parlare di indipendenza bisogna anzitutto introdurre delle descrizioni statistiche di due variabili aleatorie; una descrizione usuale é quella della **covarianza**:

$$\mathrm{cov}(X,Y) = \iint_{-\infty}^{+\infty}(x-\lambda_x)(t-\lambda_y)f(x,y)\,dx\,dy$$

Sviluppando i conti troviamo che:

$$\begin{aligned}\mathrm{cov}(X,Y) &= \iint_{-\infty}^{+\infty}(x-\lambda_x)(t-\lambda_y)f(x,y)\,dx\,dy \\ &= \iint_{-\infty}^{+\infty}(xy-x\lambda_y-y\lambda_x+\lambda_x\lambda_y)f(x,y)\,dx\,dy \\ &= \iint_{-\infty}^{+\infty}xyf(x,y)\,dx\,dy-\lambda_y\iint_{-\infty}^{+\infty}xf(x,y)\,dx\,dy+ \\ &\quad -\lambda_x\iint_{-\infty}^{+\infty}yf(x,y)\,dx\,dy+\lambda_x\lambda_y\underbrace{\iint_{-\infty}^{+\infty}f(x,y)\,dx\,dy}_{=1} \\ &= \mathrm{E}(X,Y)-\lambda_y\int_{-\infty}^{+\infty}x\underbrace{\left(\int_{-\infty}^{+\infty}f(x,y)\,dy\right)}_{h(x)=densità\,marginale}dx-\lambda_x\underbrace{\int_{-\infty}^{+\infty}yg(x)\,dy}_{\lambda_y}+\mathrm{E}(X)\mathrm{E}(Y) \\ &= \mathrm{E}(X,Y)-\mathrm{E}(X)\mathrm{E}(Y)\end{aligned}$$

Ovvero possiamo sottolineare la seguente

Propriet 1 $\mathrm{cov}(X,Y) = \mathrm{E}(X,Y)-\mathrm{E}(X)\mathrm{E}(Y)$

A questo punto diamo una definizione di variabili aleatorie indipendenti.

Definizione 4 *Due variabili aleatorie sono* ***indipendenti*** *se le loro distribuzioni sono fattorizzabili:* $f(x,y) = h(x)g(y)$.

Abbiamo visto che $\mathrm{cov}(X,Y) = \mathrm{E}(X,Y) - \mathrm{E}(X)\mathrm{E}(Y)$. Ora

$$\mathrm{E}(X,Y) = \iint_{-\infty}^{+\infty} xy\, f(x,y)\, dx\, dy$$

Se le due variabili aleatorie sono indipendenti si ha che

$$\begin{aligned} \mathrm{E}(X,Y) &= \iint_{-\infty}^{+\infty} xy\, g(x)h(y)\, dx\, dy \\ &= \int_{-\infty}^{+\infty} x\, g(x)\, dx \cdot \int_{-\infty}^{+\infty} y\, h(y)\, dy \\ &= \mathrm{E}(X) \cdot \mathrm{E}(Y) \end{aligned}$$

Di conseguenza abbiamo verificato la seguente proprietá:

Propriet 2 *Per variabili indipendenti si ha che* $\boxed{\mathrm{cov}(X,Y) = 0}$

Esempio. Nel caso seguente le due variabili x e y sono indipendenti perché:

$$f(x,y) = e^{-(x+y)} = e^{-x} \cdot e^{-y}$$

La covarianza in generale vale zero se le variabili sono indipendenti. In realtá la covarianza é una quantitá importante per misurare il livello di indipendenza di due variabili. Il particolare la **correlazione** normalizza questa misura:

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$$

Il suo valore é compreso tra -1 e +1. Per variabili indipendenti, la correlazione é nulla. Se tuttavia la correlazione nulla, non necessariamente vero che le variabli sono indipendenti. Infatti per corr = 0 le due variabili sono non correlate, mentre per corr = 1 o corr = -1 sono assolutamente dipendenti (dipendenza lineare positiva o negativa).

Finora abbiamo analizzato il caso in cui $\phi(x,y) = (x - \lambda_x)(y - \lambda_y)$, che in qualche modo un momento statistico che ha delle relazioni con la varianza. interessante ora vedere cosa succede se cambio la funzione $\phi(x,y)$; per esempio se

$$\phi(x,y) = x + y$$

quanto vale $\mathrm{E}(x+y)$?

$$\begin{aligned} \mathrm{E}(X,Y) &= \iint_{-\infty}^{+\infty} (x+y)\, f(x,y)\, dx\, dy \\ &= \int_{-\infty}^{+\infty} x\, f(x,y)\, dx \cdot \int_{-\infty}^{+\infty} y\, f(x,y)\, dy \\ &= \mathrm{E}(X) + \mathrm{E}(Y) \end{aligned}$$

In generale vale la seguente propriet:

Propriet 3 *La speranza matematica della combinazione lineare di due variabili aleatorie é data dalla combinazione lineare delle rispettive speranze matematiche:*

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$

Ora che abbiamo calcolato la $\mathrm{E}(x+y)$ vediamo come risulta la $\sigma^2(x+y)$ (ovvero la varianza della somma di due variabili aleatorie).

$$\begin{aligned}
\sigma^2(x+y) &= \iint_{-\infty}^{+\infty} \Big[(x+y) - \mathrm{E}(x+y)\Big]^2 f(x,y)\,dxdy \\
&= \iint_{-\infty}^{+\infty} \Big[x+y - \Big(\mathrm{E}(X) + \mathrm{E}(Y)\Big)\Big]^2 f(x,y)\,dxdy \\
&= \iint_{-\infty}^{+\infty} \Big[\Big(x - \mathrm{E}(X)\Big) + \Big(y - \mathrm{E}(Y)\Big)\Big]^2 f(x,y)\,dxdy \\
&= \iint_{-\infty}^{+\infty} \Big(x - \mathrm{E}(X)\Big)^2 f(x,y)\,dxdy + \iint_{-\infty}^{+\infty} \Big(y - \mathrm{E}(Y)\Big)^2 f(x,y)\,dxdy + \\
&\quad + 2\iint_{-\infty}^{+\infty} \Big(x - \mathrm{E}(X)\Big)\Big(y - \mathrm{E}(Y)\Big) f(x,y)\,dxdy
\end{aligned}$$

ovvero

$$\begin{aligned}
\sigma^2(x+y) &= \sigma_x^2 + \sigma_y^2 + 2\iint_{-\infty}^{+\infty} (x-\lambda_x)(y-\lambda_y)\, f(x,y)\,dxdy \\
&= \sigma_x^2 + \sigma_y^2 + 2\,\mathrm{cov}(X,Y)
\end{aligned}$$

Nel caso in cui le due variabili siano indipendenti la varianza della somma uguale alla somma delle varianze (infatti in questo caso si annulla il termine legato alla covarianza: $2\,\mathrm{cov}(X,Y)$).

A questo punto facciamo qualche esempio di alcune distribuzioni usate in pratica.

0.2.2 La distribuzione esponenziale

La distribuzione esponenziale ha densitá:

$$f(x) = \lambda e^{-\lambda x}$$

Di conseguenza

$$\begin{aligned}
\mathrm{F}(X) &= \int_{-\infty}^{x} f(\xi)\,d\xi \overset{(*)}{=} \int_{0}^{x} f(\xi)\,d\xi = \int_{0}^{x} \lambda e^{-\lambda x}\,d\xi = x\left[-\frac{e^{-\lambda}}{x}\right]_0^x \\
&= -e^{-\lambda x} + 1 = 1 - e^{-\lambda x}
\end{aligned}$$

Ovvero $\boxed{\mathrm{F}(X) = 1 - e^{-\lambda x}}$

(*) Ipotesi legata all'utilizzo pratico di questa distribuzione: la distribuzione esponenziale viene in genere usata per descrivere un tempo; il tempo pertanto non puó essere $-\infty$, ma viene preso $x = 0$ come istante iniziale.

Con questa distribuzione si descrive generalmente il tempo tra due arrivi consecutivi, ovvero il tempo di interarrivo. Si puó vedere che se gli interarrivi sono esponenziali, allora la probabilitá di avere n arrivi in un tempo t é una variabile aleatoria discreta distribuita secondo una distribuzione di Poisson, che verrá richiamata tra breve. Per ora importante ricordare il seguente Teorema.

Teorema 0.2.1 *La distribuzione esponenziale é l'unica distribuzione continua che gode della proprietá markoviana, ovvero la distribuzione dipende solo dall'istante iniziale.*

Dim. Dimostriamo nelle prossime righe la proprietá Markoviana. Supponiamo che la distribuzione degli interarrivi sia esponenziale con frequenza λ. Dopo aver aspettato l'arrivo aspettando t secondi dopo l'ultimo arrivo, ci si chiede qual'é la probabilitá che l'arrivo cada nei prossimi x secondi, cio qual'é la probabilitá che in prossimo interarrivo la variabile aleatoria T sia minore di $t+x$. Se gli interarrivi sono esponenziali:

$$\mathrm{Prob}\{T \le t+x \mid T > t\} = \frac{\mathrm{Prob}\{t < T < t+x\}}{\mathrm{Prob}\{T > t\}} = \frac{\int_t^{t+x} \lambda e^{-\lambda\xi}\,d\xi}{\int_t^{\infty} \lambda e^{-\lambda\xi}\,d\xi} = 1 - e^{-\lambda x}$$

che é indipendente dal tempo d'attesa t ed uguale alla distribuzione iniziale.

0.2.3 La distribuzione di Poisson

Per quanto riguarda gli arrivi consideriamo la distribuzione discreta di **Poisson** che descrive la probabilitá di avere n arrivi all'istante t nel seguente modo:

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

Questa distribuzione si dimostra essere, in questo contesto, un compromesso molto ragionevole tra semplicitá analitica e rigore descrittivo grazie a due caratteristiche importanti (come mostra la figura 2.5a):

1. per t piccoli si hanno pochi arrivi;
2. la distribuzione presenta un certo massimo, ma per n che tende a infinito il valore tende via via a diminuire.

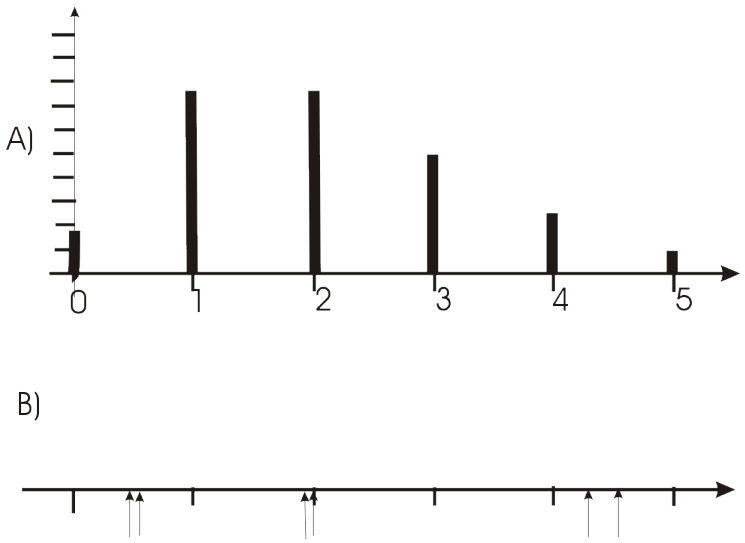


Figura 0.4 A) Possibile distribuzione di Poisson con $\lambda = 1$ [arrivi/s], t=2 [s] **B)** corrispondenti interarrivi

Associata a questa probabilit viene introdotto l'evento **interrarrivo** che descrive la distanza τ tra due arrivi; l'interrarrivo é chiaramente una variabile aleatoria, una quantitá non certa che puó essere vista come risultato di un esperimento casuale: come tale avrá una distribuzione definita da[1]

$$\mathrm{F}(t) = \mathrm{Prob}\,(T \leq t) = 1 - e^{-\lambda t}$$

e una certa densitá

$$f(t) = \lambda e^{-\lambda t}$$

che risulta *semplice* come forma analitica e gode della proprietá markoviana.

Definiamo ora il tempo medio di interarrivo

$$\mathrm{E}(T) = \frac{1}{\lambda}$$

essendo λ la frequenza degli arrivi, misurata in arrivi al secondo.

Notiamo a questo punto due proprietá piuttosto importanti nella pratica.

Propriet 4 *La confluenza di n arrivi Poissoniani con frequenze d'arrivo $\lambda_1, \lambda_2, \ldots, \lambda_n$ é ancora Poissoniana con frequenza $\sum_{i=1}^{n} \lambda_i$.*

Dim. *La situazione visualizzata nella seguente figura:*

La probabilitá di avere 1 arrivo in Δt dopo la confluenza é data dall'unione dei seguenti eventi:

[1] Nota: in questo contesto T é la variabile aleatoria, τ un caso particolare.

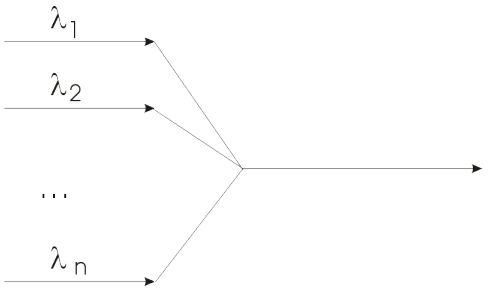


Figura 0.5 confluenza di n arrivi Poissoniani

$$\begin{aligned} prob(1arrivoin\Delta t) &= prob(1arrivoin\Delta tnelprimoramo) \cup \\ &\cup\ prob(1arrivoin\Delta tnelsecondoramo \cup \\ &\cup\ \ldots \cup prob(1arrivoin\Delta tnelramon - esimo = \\ &= \lambda_1\Delta t + \lambda_2\Delta t + \ldots + \lambda_n\Delta t = \sum_{i=1}^{n} \lambda_i \Delta t \end{aligned}$$

dove si é usata l'approssimazione $e^{-x} = 1 - x$ che vale per x piccoli (vedi oltre).
D'altra parte

$$\begin{aligned} prob(0arriviin\Delta t) &= prob(0arriviin\Delta tnelprimoramo) \cap \\ &\cap\ prob(0arriviin\Delta tnelsecondoramo) \cap \\ &\cap\ \ldots \cap prob(0arriviin\Delta tnelramon - esimo) = \\ &= (1 - \lambda_1\Delta t)(1 - \lambda_2\Delta t)\ldots(1 - \lambda_n\Delta t) = 1 - \sum_{i=1}^{n} \lambda_i \Delta t \end{aligned}$$

In modo analogo si puó mostrare che la decomposizione di un processo Poissoniano con frequenza λ in n rami d luogo a n processi Poissoniani con frequenza $p_1\lambda$ per il primo ramo, $p_2\lambda$ per il secondo ramo e $p_n\lambda$ per il ramo n-esimo, dove $p_1, p_2, ...p_n$ sono le probabilitá che l'arrivo sia assegnato rispettivamente al ramo 1, 2, ... n.

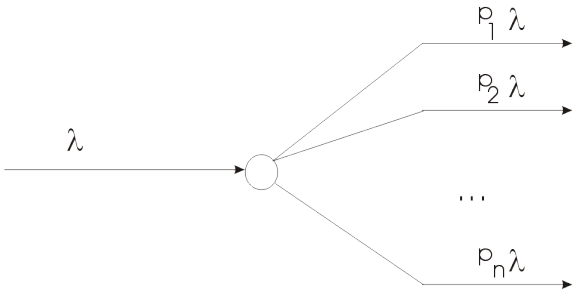


Figura 0.6 decomposizione di n arrivi Poissoniani

Basta notare come la probabilit di avere 1 arrivo sul primo ramo in Δt é data dalla probabilit di avere un arrivo sul ramo principale in Δt e che l'arrivo sia assegnato al primo ramo, cioé $prob(1arrivosulprimoramo) = \lambda\Delta t p_1$.

0.2.4 La distribuzione gaussiana

La densitá di questa distribuzione la seguente:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\lambda x}{\sigma}\right)^2}$$

Questa distribuzione é molto importante per l'esistenza di una proprietá espressa dal seguente

Teorema del limite centrale 0.2.1 *Siano $X_1, X_2, ...X_n$ delle variabili aleatorie indipendenti con la stessa distribuzione. Siano λ e σ^2 rispettivamente la media e la varianza della sequenza. Allora per n che tende a infinito (nella pratica per n sufficientemente grande) $S = \sum_i X_i$ é una variabile aleatoria con distribuzione gaussiana, con media $n\lambda$ e varianza $n\sigma^2$. Inoltre $S_0 = \frac{S-n\lambda}{n\sigma}$ é una variabile aleatoria gaussiana con media nulla e varianza unitaria (variabile aleatoria normalizzata).*

Purtroppo questa funzione non é integrabile, quindi non ammette la distribuzione in forma chiusa, l'unico modo effettuare una tabulazione.

Le due distribuzioni appena viste sono due distribuzioni di variabili aleatorie continue. Un caso importante parlando di variabili aleatorie discrete é la distribuzione geometrica.

0.2.5 La distribuzione geometrica

La distribuzione geometrica si descrive con

$$\text{Prob}(X = n) = P_n = \rho^n(1-\rho) \qquad 0 < \rho < 1$$

Di questa variabile é interessante calcolare la speranza matematica

$$\begin{aligned}
\text{E}(n) &= \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\rho\sum_{n=0}^{\infty} n\rho^{n-1} \\
&= (1-\rho)\rho\sum_{n=0}^{\infty}\frac{\partial\rho^n}{\partial\rho} = (1-\rho)\rho\frac{\partial}{\partial\rho}\underbrace{\sum_{n=0}^{\infty}\rho^n}_{=1/(1-\rho)} \\
&= (1-\rho)\rho\frac{\partial}{\partial\rho}\left[\frac{1}{1-\rho}\right] = (1-\rho)\rho\frac{1}{(1-\rho)^2} \\
&= \frac{\rho}{1-\rho}
\end{aligned}$$

che rappresenta il valor medio di una distribuzione geometrica.

0.3 Applicazione delle distribuzioni statistiche

0.3.1 Modello a coda semplice: M/M/1

Prendiamo in considerazione un modello estremamente semplificato: un sistema a coda semplice come quello mostrato in figura 2.4; questo modello potrebbe rappresentare un sistema operativo dedicato, ovvero (per i nostri scopi) un sistema operativo senza disco (*diskless*) e privo di interazioni con l'utente.

Per decidere quale distribuzione utilizzare per modellizzare il sistema bisogna saper conciliare due esigenze tra loro contrastanti:

1. dovendo fare calcoli analitici con un sistemi sempre piú complessi bisogna utilizzare delle distribuzioni che risultino analiticamente accessibili;
2. il modello deve comunque essere il piú aderente possibile alla realtá.

Tornando al modello iniziale, per semplicitá assumiamo che sia gli arrivi che i tempi di esecuzione siano variabili aleatorie distribuite esponenzialmente: facendo cosí si giunge a quella che viene chiamata **coda M/M/1** (M sta per Markov, 1 sta per una CPU).

Lo scopo é quello di vedere (note le caratteristiche statistiche degli arrivi e dei servizi) qual'é il numero medio di processi e il tempo di attesa medio. Per giungere a queste conclusioni serve un'altra proprietá:

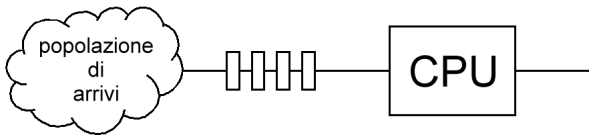


Figura 0.7 Gli arrivi e i tempi di servizio sono variabili aleatorie.

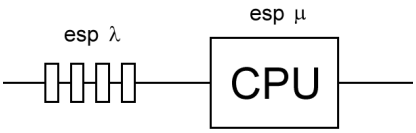


Figura 0.8 Coda M/M/1

Formula di Little 0.3.1 *Il numero medio di processi nel sistema operativo é uguale alla frequenza degli arrivi moltiplicata per il tempo di attesa medio:*

$$\overline{n} = \lambda w$$

NB: Questa é una proprietá che non dipende dalla particolare distribuzione (o dal particolare sistema a coda).

Da $\overline{n} = \lambda w$ si puó ricavare immediatamente il tempo di attesa medio:

$$w = \frac{1}{\lambda}\overline{n}$$

Ora per calcolare il numero medio di processi bisogna considerare le proprietá di stazionarietá: si sfrutterá in particolare il fatto che le derivate temporali sono nulle (e per questo motivo si fará ricorso ai limiti).

Stazionariet: Non potendo fare delle stime sul *bootstrap* si considera il sistema quando a regime; questa un'enorme semplificazione: al boot si potrebbe infatti avere una grande quantitá di processi che influenzano l'andamento futuro del sistema.

I modelli che verranno usati saranno basati sul paragone di un sistema a coda con un sistema a fluido (figura 2.7)

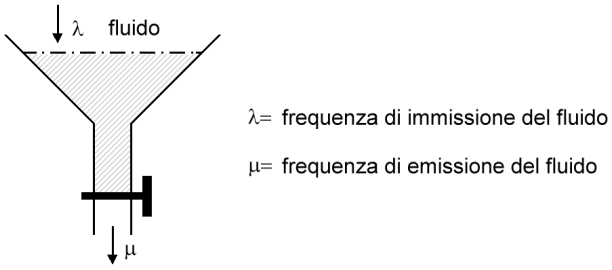


Figura 0.9

A differenza del sistema a code (discreto) questo é un sistema *continuo*: quest'approssimazione consente di utilizzare degli strumenti matematici che danno alcune informazioni non calcolabili nel caso discreto.

Definiamo il **coefficiente di utilizzazione** del sistema coda

$$\textit{coeff. utilizzazione} = \frac{\mathrm{E}(CPU)}{\mathrm{E}(\textit{interarrivi})} = \frac{\textit{tempo medio di CPU}}{\textit{tempo medio di interarrivi}} = \boxed{\rho < 1}$$

$\rho < 1$: ovvero la frequenza di CPU dev'essere superiore alla frequenza con la quale arrivano i nuovi procassi: se cosí non fosse la lunghezza della coda tenderebbe a crescere (la CPU non riuscirebbe a smaltire i processi) fino a diventare "infinita".

NB: Il nostro modello al momento non ha limiti per quanto riguarda l'arrivo dei processi in coda. Nella pratica questo non succede in quanto un sistema reale ha delle caratteristiche di finitezza (ovvero la coda avrá sempre una lunghezza finita).

Domanda: Qual'é la probabilitá che ci siano n processi nel sistema all'istante $t + \Delta t$?
Prima di tutto definiamo:

- $P_{nE}(t) = \text{Prob}\,(n \textit{ processi eseguiti in } t)$
- $P_{nA}(t) = \text{Prob}\,(n \textit{ processi arrivati in } t)$

Quindi

$$\begin{aligned} P_n(t+\Delta t) &= P_n(t) \cap P_{0A}(\Delta t) \cap P_{0E}(\Delta t) \\ &\cup P_n(t) \cap P_{1A}(\Delta t) \cap P_{1E}(\Delta t) \\ &\cup P_{n+1}(t) \cap P_{0A}(\Delta t) \cap P_{1E}(\Delta t) \\ &\cup P_{n-1}(t) \cap P_{1A}(\Delta t) \cap P_{0E}(\Delta t) \\ &\cup ecc\ldots \end{aligned} \tag{1}$$

Si potrebbe continuare, ma gli eventuali termini aggiuntivi sono termini infinitesimali (verrá fatto un processo di limite per $\Delta t \to \infty$)

Ipotesi: arrivi ed esecuzioni in CPU sono eventi indipendenti.

- $\text{Prob}\,(n \textit{ processi in } t) = \dfrac{(\lambda t)^n}{n!}\, e^{-\lambda t}$
- $\text{Prob}\,(0 \textit{ processi arrivati in } \Delta t) = \dfrac{(\lambda \Delta t)^0}{0!}\, e^{-\lambda \Delta t} = e^{-\lambda \Delta t}$
- $\text{Prob}\,(1 \textit{ processo arrivato in } \Delta t) = \dfrac{(\lambda \Delta t)^1}{1!}\, e^{-\lambda \Delta t} = \lambda \Delta t\, e^{-\lambda \Delta t}$

A questo punto semplifichiamo le cose considerando lo sviluppo in serie di Taylor attorno all'origine:

$$f(t) = f(0) + f'(0)t + f''(0)t^2 + \ldots$$

se considero $f(t) = e^{-\lambda \Delta t}$ allora ho che $f' = -\lambda e^{-\lambda \Delta t}$ e $f'' = \lambda^2 e^{-\lambda \Delta t}$ e di conseguenza si ha che

$$e^{-\lambda \Delta t} = 1 - \lambda \Delta t + \lambda^2 \Delta t^2 + \ldots$$

ora facendo tendere $\Delta t \to 0$ é possibile troncare al secondo termine, e quindi

$$\boxed{e^{-\lambda \Delta t} = 1 - \lambda \Delta t}$$

A questo punto si ha che

- $\text{Prob}\,(0 \textit{ processi arrivati in } \Delta t) = 1 - \lambda \Delta t$
- $\text{Prob}\,(1 \textit{ processo arrivato in } \Delta t) = \lambda \Delta t\,(1 - \lambda \Delta t) = \lambda \Delta t$

quindi si sostituiscono i valori appena trovati nella 1 specificando con λ la frequenza di arrivo dei processi e con μ quella di esecuzione:

$$\begin{aligned} P_n(t+\Delta t) &= P_n(t) \cap (1-\lambda \Delta t)(1-\mu \Delta t) \\ &\cup P_n(t) \cap \lambda \Delta t \cap \mu \Delta t \\ &\cup P_{n+1}(t) \cap (1-\lambda \Delta t) \cap \lambda \Delta t \\ &\cup P_{n-1}(t) \cap \lambda \Delta t \cap (1-\mu \Delta t) \end{aligned}$$

Ora si utilizza l'ipotesi di indipendenza: se le probabilitá sono indipendenti allora le intersezioni diventano prodotti e le unioni diventano somme. Trascurando tutto ció che svolgendo i conti diventa un infinitesimo di ordine superiore si ottiene

$$P_n(t+\Delta t) = P_n(t)[1 - (\lambda + \mu)\Delta t] + P_{n+1}(t)\mu \Delta t + P_{n-1}(t)\lambda \Delta t$$

da questa relazione si puó ricavare un rapporto incrementale:

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = P_{n+1}(t)\mu - P_n(t)(\lambda+\mu) + P_{n-1}(t)\lambda$$

facendo il limite per $\Delta t \to 0$ si ottiene

$$P_n'(t) \quad = \quad \mu P_{n+1}(t) - (\lambda+\mu) + \lambda P_{n-1}(t) \stackrel{\downarrow}{=} 0 \tag{2}$$

derivata che viene posta a zero in quanto ci interessa una soluzione stazionaria (non un transitorio).

Ora il coefficiente di utilizzazione é dato da

$$\rho = \frac{\textit{tempo attesa unità centrale}}{\textit{media interarrivi}} = \frac{\frac{1}{\mu}}{\frac{1}{\lambda}} = \frac{\lambda}{\mu}$$

dividendo la (2) per μ si giunge alla relazione:

$$P_{n+1}(t) - (1+\rho)P_n(t) + \rho P_{n-1}(t) = 0$$

Con questa relazione possibile definire delle iterazioni (n dev'essere ≥ 1)

$$\begin{array}{ll} n=1 & P_2(t) = (1+\rho)P_1(t) - \rho P_0(t) \\ n=2 & P_3(t) = (1+\rho)P_2(t) - \rho P_1(t) \\ n=3 & \qquad ecc\ldots \end{array} \tag{3}$$

In queste formule compare il termine $P_0(t)$ che bisogna calcolare.

Domanda: Quanto vale $P_0(t)$?

$$\begin{array}{rcl} P_0(t+\Delta t) & = & P_0(t) \cap P_{0A}(\Delta t) \ {}^{(*)} \\ & \cup & P_1(t) \cap P_{1E}(\Delta t) \cap P_{0A}(\Delta t) \end{array}$$

(*) vengono trascurati i termini che non hanno senso: avendo 0 processi nel sistema non ci possono essere processi in via di esecuzione.

$$\begin{array}{rcl} P_0(t+\Delta t) & = & P_0(t)(1-\lambda\Delta t) + P_1(t)\mu\Delta t \\ & = & P_0(t) - \lambda\Delta t P_0(t) + \mu P_1(t)\Delta t \end{array}$$

Come prima ci si conduce al rapporto incrementale, quindi si fa il limite per $\Delta t \to 0$ e si ricava

$$P_0'(t) = \mu P_1(t) - \lambda P_0(t) = 0$$

da cui segue immediatamente

$$P_1(t) = \rho P_0(t)$$

valida *solo* nel caso in cui ci siano 0 processi all'interno del sistema.

Tornando alle equazioni (3) troviamo che

$$\begin{array}{ll} n=1 & \rho^2 P_0(t) \\ n=2 & \rho^3 P_0(t) \end{array}$$

In generale $\boxed{P_n(t) = \rho^n P_0(t)}$

Da note proprietá statistiche si osserva che

$$\begin{array}{rcl} \displaystyle\sum_{n=0}^{\infty} P_n(t) = 1 & \Rightarrow & \displaystyle\sum_{n=0}^{\infty} \rho^n P_0(t) = 1 \\ & \Rightarrow & \displaystyle P_0(t) \sum_{n=0}^{\infty} \rho^n \ \stackrel{(*)}{=} \ P_0(t)\frac{1}{1-\rho} = 1 \\ & \Rightarrow & P_0(t) = 1-\rho \end{array}$$

(•) la sommatoria in questione rappresenta una serie geometrica di ragione $\rho < 1$

A questo punto é possibile dare una stima alla probabilitá che all'istante t ci siano n processi nel sistema: $P_n(t) = \rho^n(1-\rho)$ ovvero é possibile stimare il numero medio di processi in questo tipo di sistema operativo, da

$$\mathrm{E}(n) = \sum_{n=0}^{\infty} nP_n(t) = \sum_{n=0}^{\infty} n\rho^n P_0(t) = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\frac{\rho}{(1-\rho)^2}$$

si ricava che $\mathrm{E}(n) = \frac{\rho}{(1-\rho)}$ A questo punto é possibile calcolare il tempo d'attesa medio dei processi in questo sistema; da

$$w = \frac{1}{\lambda}\,\frac{\rho}{1-\rho} = \frac{1}{\lambda}\,\frac{\frac{\lambda}{\mu}}{1-\rho}$$

si ottiene $w = \frac{1}{\mu(1-\rho)}$ **tempo medio di attesa** dei processi (espresso in secondi).

Applicazione: questa formula viene tipicamente usata nei problemi di dimensionamento del tipo: qual'é la potenza della CPU necessaria per far girare una data applicazione al fine di riuscire a smaltire i processi in coda?

0.3.2 Modello di coda ciclica

Prendiamo ora in considerazione il modello di coda ciclica mostrato in figura 2.8: per il momento trascuriamo la possibilitá di terminare qualche processo o di introdurne di nuovi.

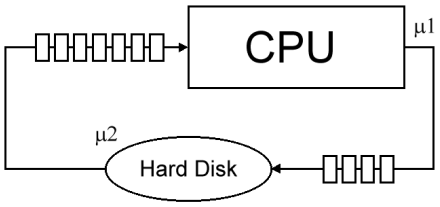


Figura 0.10 Sistema di coda ciclica. Il sistema é chiuso nel senso che c'é un numero fisso di processi. μ_1 e μ_2 rappresentano le frequenze di esecuzione e di servizio disco.

Caratteristica: all'interno di questo sistema ci sono un numero costante (J) di processi: di conseguenza non si presenta il problema della coda che puó raggiungere una lunghezza infinita.

In generale avevamo definito

$$\rho = \frac{\mathrm{E}(CPU)}{\mathrm{E}(interarrivi)} = \frac{tempo\ medio\ di\ CPU}{tempo\ medio\ di\ interarrivi}$$

e si aveva posto $\rho < 1$ per motivi di stabilitá. In questo contesto ρ *non limitato* da questa quantitá: in linea teorica puó assumere anche valori maggiori dell'unitá.

Quello che ci interessa calcolare di questo modello é

1. il numero medio di processi in coda;
2. il tempo di attesa medio dei processi in coda;
3. il coefficiente di utilizzazione della CPU;

   Quest'ultimo indica la probabilitá che l'unitá centrale sia attiva (utilizzata): é auspicabile che il sistema sia progettato in modo che l'utilizzo della CPU tenda al 100%.

Naturalmente esistono diversi modi per calcolare queste quantitá, ciascuno con un diverso livello di precisione.

Numero di processi e tempo di attesa medi

Per calcolare i primi due parametri si usa lo stesso metodo usato in precedenza nel caso del sistema a coda semplice, ovvero, considerando μ_1 la frequenza delle esecuzioni dei processi da parte della CPU e μ_2 la frequenza delle richieste servite dal disco, i passi da fare sono:

- Si calcola $P_n(t+\Delta t)$, ovvero la probabilit di avere n processi nel sistema all'istante $t+\Delta t$; si ricava

$$P_n(t+\Delta t) = P_n(t)[1-(\mu_1+\mu_2)\Delta t] + P_{n+1}(t)\mu_1\Delta t + P_{n-1}(t)\mu_2\Delta t$$

- Da questa espressione si trova il rapporto incrementale:

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \mu_1 P_{n+1}(t) - P_n(t)(\mu_1+\mu_2) + \mu_2 P_{n-1}(t)$$

- Quindi si ricava la derivata facendo il limite per $\Delta t \to 0$ e, considerando il sistema a regime (quindi una situazione stazionaria), la si pone uguale a zero:

$$P_n'(t) = \mu_1 P_{n+1}(t) - P_n(t)(\mu_1+\mu_2) + \mu_2 P_{n-1}(t) = 0 \tag{4}$$

- Da questa relazione é possibile ricavare un'iterazione, a condizione di avere un punto iniziale da cui partire; si va pertanto a cercare $P_0(t+\Delta t)$, per poi ricavare un ulteriore rapporto incrementale ($\Rightarrow P_0'(t) = \mu_1 P_1(t) - \mu_2 P_0(t)$) da annullare per la stazionarietá e quindi trovare

$$P_0(t) = \frac{\mu_1}{\mu_2} P_1(t) = \frac{1}{\rho} P_1(t) \;\Rightarrow\; P_1(t) = \rho P_0(t) \tag{5}$$

- Dalla (4) e dalla (5) si ottiene la seguente relazione generale

$$\boxed{P_n(t) = \rho^n P_0(t)}$$

  che descrive la probabilitá di avere n processi presenti in un sistema a coda ciclica.

Quest'ultima relazione é esattamente uguale a quella del caso M/M/1 (ovvio, perché anche questo sistema puó essere visto come un sistema a coda M/M/1 in cui gli arrivi sono rappresentati dal servizio di disco).

Attenzione: Le cose cambiano grazie a una differenza sostanziale; qui il numero di processi é limitato!

Quello che a questo punto anche qui bisogna fare é valutare $P_0(t)$. Come fatto in precedenza (vedi pagina 12) si puó dire che

$$\sum_{n=0}^{J} P_n(t) = 1 \Rightarrow \sum_{n=0}^{J} \rho^n P_0(t) = 1 \Rightarrow P_0(t) = \left(\sum_{n=0}^{J} \rho^n\right)^{-1}$$

L'ultimo termine di ques'equazione non é piú una serie geometrica, bensí una somma geometrica e, considerando che ρ non é piú < 1 si ha:

$$\sum_{n=0}^{J} \rho^n = \frac{1-\rho^{J+1}}{1-\rho}$$

e di conseguenza

$$\boxed{P_0(t) = \frac{1-\rho}{1-\rho^{J+1}}}$$

Quindi generalizzando si trova che in un sistema a coda la probabilitá di avere n processi in coda all'istante t pari a: $P_n(t) = \rho^n \frac{1-\rho}{1-\rho^{J+1}}$. Ora é bene ricordare il nostro obiettivo: quello che si sta cercando é

1. il numero medio di processi in coda;
2. il tempo di attesa medio dei processi in coda;

Per quanto riguarda il primo punto, ricordiamo che qui siamo in un contesto di distribuzione discreta, quindi si puó dire che

da cui deriva che $\mathrm{E}(n) = \frac{\rho}{1-\rho} \; \frac{J\,\rho^{J+1} - (J+1)\,\rho^J + 1}{1-\rho^{J+1}}$

Ora calcoliamo il secondo punto, ovvero il tempo di attesa medio: la formula di Little dice che il numero medio di processi nel sistema dato dalle due quantitá frequenza di arrivo e tempo di attesa medio, ovvero $\overline{n} = \lambda w$ che nel nostro caso si traduce con

$$\mathrm{E}(n) = \lambda w$$

da questa si ricava

$$w = \frac{1}{\mathit{freq.\ arrivi\ in\ coda}} \cdot \mathrm{E}(n)$$

Dunque manca da calcolare la frequenza con cui arrivano i processi in coda.

Facciamo un passo avanti: ora non trascuriamo piú la possibilitá di terminare qualche processo o di introdurne di nuovi: anche con questa modifica il numero di processi nel sistema non cambia col tempo in quanto viene imposto il vincolo che un nuovo processo non possa entrare prima che un altro non sia terminato.

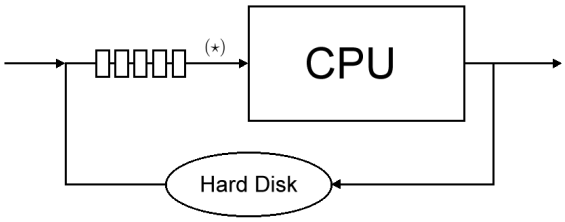


Figura 0.11

La frequenza di partenza dei processi dalla coda (vedi figura 0.11 ($\star$)) é data dalla probabilitá di avere dei processi in coda $\cap$ la probabilitá di avere un'esecuzione da parte dell'unitá centrale, ovvero

$$\begin{aligned}
\mathit{freq.\ partenza\ processi\ da\ coda} &= \mathrm{Prob}(n\ \mathit{processi\ in\ coda}\ \neq 0)\,\mu_1 \\
&= (1 - P_0(t))\,\mu_1 = \left(1 - \frac{1-\rho}{1-\rho^{J+1}}\right)\mu_1 \\
&= \frac{1-\rho^{J+1} - 1 + \rho}{1-\rho^{J+1}} \cdot \mu_1 \\
&= \rho\,\frac{1-\rho^J}{1-\rho^{J+1}} \cdot \mu_1
\end{aligned}$$

se si considerano le condizioni di stazionarietá, ovvero la situazione in cui a regime le dimensioni delle singole code sono uguali, allora deve valere:

$$\mathit{freq.\ arrivi} = \mathit{freq.\ partenze}$$

A questo punto abbiamo tutti i dati per poter ripartire dalla formula di Little:

$$w = \frac{1}{\mathit{freq.\ arrivi\ in\ coda}} \cdot \mathrm{E}(n) = \frac{1}{\rho\mu_1} \cdot \frac{1-\rho^{J+1}}{1-\rho^J} \cdot \frac{\rho}{1-\rho} \cdot \frac{J\,\rho^{J+1} - (J+1)\,\rho^J + 1}{1-\rho^{J+1}}$$

quindi il tempo di attesa medio in coda ciclica si puó esprimere come: $w = \frac{1}{\mu_1(1-\rho)} \cdot \frac{J\,\rho^{J+1} - (J+1)\,\rho^J + 1}{1-\rho^J}$

Ora resta solo da determinare il coefficiente di utilizzazione della CPU.
Qual'é la probabilitá che la CPU sia attiva? Quest'ultimo é un parametro non immediato da calcolare, peró di importanza fondamentale; con il tipo di modello in questione é possibile rispondere in modo approssimativo.

Supponiamo di avere J molto alto, ovvero supponiamo di avere un elevato livello di multiprogrammazione. J non puó tendere a infinito (altrimenti $P_0(t) \to (1-\rho)$) ma dev'essere sufficientemente grande da fare in modo che la probabilitá che la CPU sia attiva equivalga alla probabilitá di avere dei processi in coda (ovvero se J elevato trascuro il singolo processo in esecuzione rispetto ai J processi che sono in coda).

$$\text{Prob(CPU attiva)}\Big]_{J\uparrow} = \text{Prob(n processi in coda CPU} \neq 0)$$

con questa approssimazione, chiamato μ il coefficiente di utilizzazione della CPU, si ha che

$$\mu\Big]_{J\uparrow} = 1 - P_0(t) = 1 - \frac{1-\rho}{1-\rho^{J+1}}$$

ovvero $\mu\Big]_{J\uparrow} = \rho \cdot \frac{1-\rho^J}{1-\rho^{J+1}}$

Discussione dei risultati

Per quanto riguarda l'approssimazione é importantissimo il modo in cui si valuta questo coefficiente; sono state fatte delle analisi sperimentali e si é visto che questa approssimazione é insufficiente, ovvero ha un grande margine d'errore. A causa di questo margine si é costretti a passare al **sistema continuo** (che rappresenta un'altra approssimazione): qui siamo in presenza di un processo stocastico discreto rappresentato dal numero di arrivi di processi in coda e dal numero di processi usciti dalla coda, processi che possiamo indicare con $\alpha(t)$ e $\delta(t)$:

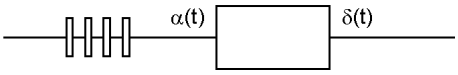


Figura 0.12

In questo modello non verranno usate le statistiche usate fino a questo momento (μ_1, μ_2 sono statistiche) ma verrá utilizzato un numero in funzione del tempo: questo permetterá di andare ad indagare in modo piú puntuale e in particolare consentirá di vedere quello che succede come andamento temporale (quindi ad esempio i transitori...). I procssi stocastici in gioco sono del tipo:

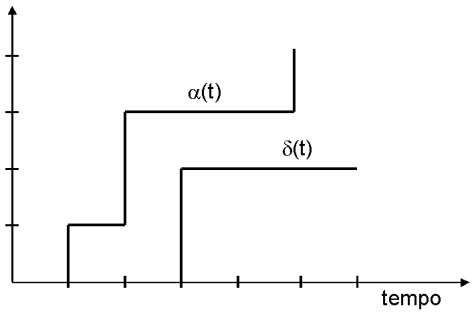


Figura 0.13

Introduciamo ora la variabile $N(t)$ che descrive il numero di processi nel sistema coda ed é dato da:

$$N(t) = \alpha(t) - \delta(t)$$

A questo punto, se ci si trova in condizioni di alto carico della coda, i due numeri $\alpha(t)$ e $\delta(t)$ sono molto piú elevati, quindi le discontinuitá sono molto piccole rispetto il sistema nel suo complesso e di conseguenza i grafici tendono ad assomigliare sempre di piú a curve continue.

Approssimazione: a questo punto non consideriamo piú un $\alpha(t)$ e un $\delta(t)$, bensí dei valori medi: $\overline{\alpha}(t)$ e $\overline{\delta}(t)$ che rappresentano le medie temporali degli arrivi e delle partenze. Ovviamente si ha:

$$N(t) = \overline{\alpha}(t) - \overline{\delta}(t)$$

Osservazione: In questo caso le medie tendono (sempre in condizioni di alto traffico) al valore medio, ma *localmente* i valori possono discostarsi di molto...

In sintesi: si considerano processi stocastici continui che tendono (come limite estremo) a quelli discreti.

Il fatto di passare da un sistema stocastico discreto ad uno continuo se da un lato introduce un'ulteriore approssimazione (un modello continuo diverso da modello reale che discreto), dall'altro comporta il vantaggio di poter rappresentare le code d'attesa con un sistema a fluido continuo.

0.4 Modello a fluido continuo

L'approssimazione consiste nel descrivere il sistema a coda d'atesa con un modello a fluido continuo. Riprendendo quanto detto in precedenza si ha

$$N(t) = \overline{\alpha}(t) - \overline{\delta}(t)$$

con $N(t)$ numero di processi nel sistema coda, $\overline{\alpha}(t)$ e $\overline{\delta}(t)$ che rappresentano le medie temporali degli arrivi e delle partenze.

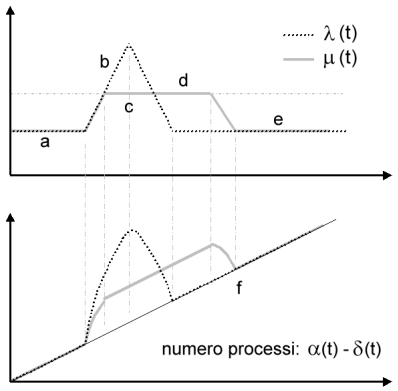
A questo punto vengono introdotte due nuove variabili

che rappresentano rispettivamente la frequenza degli arrivi e la frequenza delle partenze.

A questo punto possiamo dire che

$$\overline{\alpha}(t) = \overline{\alpha}(0) + \int_0^t \lambda(\xi)\, d\xi \qquad \overline{\delta}(t) = \overline{\delta}(0) + \int_0^t \mu(\xi)\, d\xi$$

Diamo ora un'occhiata qualitativa a cosa puó succedere a un sistema di questo tipo. Possiamo rappresentare ad esempio una variazione di $\lambda(t)$ come rappresentato in figura:



,sopra:frequenze di arrivo e partenza, sotto: numero di arrivi e partenze] Analizziamo la figura accanto: **a.** $\lambda = \mu$; **b.** eccesso di arrivi che il sistema non riesce a smaltire; **c.** saturazione dei processi in esecuzione; **d.** gli arrivi sono ritornati a regime ma la CPU sta ancora lavorando a pieno carico; **e.** la CPU riuscita a smaltire i processi in coda (le due aree si equivalgono); **f.** le due curve (frequenza d'ingresso e d'uscita) tornano a sovrapporsi.
La cosa che emerge da questa rappresentazione é che comunque viene fatta l'ipotesi che il numero di processo nel sistema sia la differenza tra la media degli arrivi e la media ddelle partenze.

Calcoliamo ora la probabilitá che il numero di arrivi all'istante t sia almeno n:

$$\text{Prob}[\,\alpha(t) \geq n\,] = \text{Prob}[\,T_n \leq t\,]$$

con T_n istante (*assoluto*, non interarrivo) di arrivo dell'n-esimo processo.

Se consideriamo gli interarrivi t_i indipendenti, allora

$$T_n = \sum_{i=1}^{n} t_i$$

e applicando il teorema del limite centrale si puó dire che l'istante assoluto di arrivo é distribuito *normalmente* (ovvero la distribuzione degli istanti di arrivo una variabile gaussiana). A questo punto anche la distribuzione del numero di arrivi nell'istante t una variabile gaussiana. Quindi, riassumendo:

- $\alpha(t)$, $\overline{\alpha}(t)$ sono delle variabilio gaussiane;
- quindi anche $\delta(t)$, $\overline{\delta}(t)$ e di conseguenza $N(t)$ sono variabili aleatorie gaussiane.

Si puó dunque dire che $\overline{\alpha}(t)$ e $\overline{\delta}(t)$ sono *normali* con la media e con la varianza che le caratterizzano:

$$\overline{\alpha}(t) = N(\lambda_{\overline{\alpha}},\, \sigma^2_{\overline{\alpha}}) \qquad \overline{\delta}(t) = N(\lambda_{\overline{\delta}},\, \sigma^2_{\overline{\delta}})$$

Per capire quanto valgono queste variabili torniamo al sistema di coda ciclica mostrato nella figura successiva; si puó dimostrare che:

$$\boxed{\lambda_{\overline{\alpha}} \approx \mu_2} \qquad \boxed{\lambda_{\overline{\delta}} \approx \mu_1}$$

rappresentano delle buone approssimazioni.

Allo stesso modo si puó ricavare σ in funzione di λ:

$$\sigma^2_{\overline{\alpha}} = f(\lambda_{\overline{\alpha}}) \qquad \sigma^2_{\overline{\delta}} = f(\lambda_{\overline{\delta}})$$

A questo punto, come annunciato in precedenza, $N(t) = \overline{\alpha}(t) - \overline{\delta}(t)$ é una variabile aleatoria gaussiana con un certo valor medio $\mu = \mu_2 - \mu_1$ e una certa $\sigma^2 = f(\sigma^2_{\overline{\alpha}},\, \sigma^2_{\overline{\delta}})$.

Osservazione: Essendo una variabile continua possiamo introdurre l'uso di strumenti differenziali. Osserviamo come l'approssimazione con un sistema continuo venga compensata dal fatto che la funzione continua venga espressa in funzione di media e di varianza, che descrivono molto meglio la variabile aleatoria di quanto possa farlo la sola frequenza.

Osservazione: Poniamo in evidenza una stranezza: nel sistema a coda semplice era possibile ricavare un fattore di traffico:

$$\rho = \frac{\mathrm{E}(CPU)}{\mathrm{E}(disco)} = \frac{\mu_2}{\mu_1} < 1$$

Ora, nel sistema a coda ciclica (in cui i processi sono al piú J) ρ puó essere anche > 1: al massimo si satureranno alcune code ma non potrá succedere nulla di... irreparabile!

Tornando sulle relazioni del valor medio (denotando con $\mu_{N(t)}$ una media) si ha:

$$\mu_{N(t)} = \mu_2 - \mu_1 = \mu_1 \left(\frac{\mu_2}{N} - 1 \right) = \mu_1(\rho - 1)$$

In realtá vediamo che é buona norma avere anche in questo caso $\rho < 1$:infatti per $\rho < 1$ abbiamo che $\mu_{N(t)}$ é un valore negativo (questo deriva dall'approssimazione che abbiamo fatto). Questo problema viene mitigato introducendo alcune condizioni.

Condizioni di riflessione: $\boxed{N(t) \geq 0}$

Queste condizioni stanno a significare che il valore effettivo $N(t)$ é positivo: ovvero la distribuzione gaussiana ha un valore medio μ negativo ma a noi interessa solo (vedi grafico) il quadrante positivo.

I casi che tratteremo avranno tipicamente un valore $\mu < 0$ anche se abbastanza vicino all'origine.

Introduciamo a questo punto la funzione distribuzione:

$$\mathrm{F}(x,t) = \mathrm{Prob}[N(t) \leq x] = \mathrm{Prob}\{\text{n processi nel sistema in T sia} < x\}$$

Possono a questo punto seguire delle considerazioni molto complesse che si possono riassumere intuitivamente nel seguente modo: consideriamo una funzione definita come segue

$$\psi(x) = \frac{\partial \mathrm{F}}{\partial t}$$

questa funzione ora la sviluppiamo in serie di Taylor troncando al secondo termine:

$$\begin{aligned}\psi(x) &= \mathrm{F}'(0)\cdot\frac{\partial \mathrm{F}}{\partial x}+\mathrm{F}''(0)\cdot\frac{\partial^2 \mathrm{F}}{\partial x^2} \\ &\quad \text{si pu dimostrare che } \begin{cases} \mathrm{F}'(0)=-\mu \\ \mathrm{F}''(0)=\frac{\sigma^2}{2}\end{cases} \\ &= -\mu\cdot\frac{\partial \mathrm{F}}{\partial x}+\frac{\sigma^2}{2}\cdot\frac{\partial^2 \mathrm{F}}{\partial x^2}\end{aligned}$$

giungendo cosí all'**equazione di diffusione**:
$\frac{\partial \mathrm{F}}{\partial t}=-\mu\cdot\frac{\partial \mathrm{F}}{\partial x}+\frac{\sigma^2}{2}\cdot\frac{\partial^2 \mathrm{F}}{\partial x^2}$
Questa é un'equazione differenziale che descrive un determinato fenomeno fisico; in questo contesto viene utilizzato per descrivere un sistema di code d'attesa: qui infatti μ e σ^2 sono proprio le μ e σ^2 di $N(t)$.

Quest'equazione viene qui usata in un contesto di stazionarietá rispetto al tempo, per cui avremo

$$-\mu\cdot\frac{\partial \mathrm{F}}{\partial x}+\frac{\sigma^2}{2}\cdot\frac{\partial^2 \mathrm{F}}{\partial x^2}=0$$

Questa é un'equazione differenziale la cui soluzione

$$\mathrm{F}(x)=A(1-Be^{\frac{2\mu}{\sigma^2}x}) \qquad \text{con } A \text{ e } B \text{ costanti}$$

Questa soluzione descrive $\mathrm{Prob}[N\le x]$ (trascurando i transitori, ovvero prendendo la situazione a regime).

Il problema ora é calcolare le costanti A e B; queste si trovano con le condizioni al contorno considerando due punti limite come $x=0$ e $x=J$:

$$\mathrm{F}(J)=1 \qquad \mathrm{F}(0)\ge 0$$

con $\mathrm{F}(J)$ probabilitá che il numero di processi nel sistema sia $\le J$.

Dalla prima condizione si ricava

$$\mathrm{F}(J)=A(1-Be^{\frac{2\mu}{\sigma^2}J}) \quad \Rightarrow \quad \boxed{A=\frac{1}{1-Be^{\frac{2\mu}{\sigma^2}J}}}$$

quindi ottengo che la $\mathrm{F}(x)$ si puó rappresentare come

$$\mathrm{F}(x)=\frac{1-Be^{\frac{2\mu}{\sigma^2}x}}{1-Be^{\frac{2\mu}{\sigma^2}J}}$$

Lavorando invece sulla seconda condizione al contorno si vede che

$$\mathrm{F}(0)\ge 0 \quad \Rightarrow \quad \mathrm{Prob}[N\le 0]=\mathrm{Prob}[N=0]\ge 0$$

non potendo essere N segativo.

Osservazione: Ora abbiamo che

$$F(0)=\frac{1-B}{1-Be^{\frac{2\mu}{\sigma^2}J}} \tag{6}$$

un'espressione ancora troppo complicata al fine di ricavare B. A questo punto introduciamo un'ulteriore:

Ipotesi: $J\to\infty$, ovvero ipotizziamo un livello di multiprogrammazione molto elevato. Non é l'unica soluzione, ma quella che ci interessa maggiormente. Dalla (6) si possono ricavare diverse soluzioni in base al livello di multiprogrammazione. Nella pratica conviene analizzare il sistema nei casi piú critici, come quello di un livello di multiprogrammazione elevato.

Ora, ricordando che $\mu < 0$, si ha che

$$J \to \infty \quad \Rightarrow \quad \mathrm{F}(0) = 1 - B$$

Quanto vale questo $1 - B$? In realtá si puó far ricorso ad un altro modello, ovvero a quello della coda ciclica dove la probabilitá di avere i processi in coda era data da (vedi pagina 14):

$$P_i(t) = \rho^i \cdot \frac{1-\rho}{1-\rho^{J+1}}$$

A questo punto si pone

$$\mathrm{F}(0) = P_0(t) = \rho^0 \cdot \frac{1-\rho}{1-\rho^{J+1}}$$

con, ricordiamolo, $\rho < 1$ e con $J \to \infty$; quindi

$$\lim_{J\to\infty} P_0(t) = \lim_{J\to\infty} \frac{1-\rho}{1-\rho^{J+1}} = 1 - \rho$$

da cui si ricava finalmente che $\boxed{B = \rho}$

A questo punto si giunge alla soluzione corretta:

$\mathrm{F}(x) = \frac{1-\rho \cdot e^{\frac{2\mu}{\sigma^2}x}}{1-\rho \cdot e^{\frac{2\mu}{\sigma^2}J}}$

Quello che ora rimane da determinare il coefficiente di utilizzazione della CPU:

$$\mu_{CPU} = \text{Prob[ci sia qualche processo in coda]} = \text{Prob}[N \neq 0] = 1 - \mathrm{F}(0)$$

e dato che

$$\mathrm{F}(0) = \frac{1-\rho}{1-\rho\, e^{\frac{2\mu}{\sigma^2}J}}$$

allora

$$1 - \mathrm{F}(0) = \frac{1 - \rho \cdot e^{\frac{2\mu}{\sigma^2}J} - 1 + \rho}{1 - \rho \cdot e^{\frac{2\mu}{\sigma^2}J}}$$

da cui si ricava il **coefficiente di utilizzazione**:

$\mu_{CPU} = \frac{\rho-\rho \cdot e^{\frac{2\mu}{\sigma^2}J}}{1-\rho \cdot e^{\frac{2\mu}{\sigma^2}J}}$

in questa formula il coefficiente di utilizzazione definito in termini di valor medio e di varianza di N. Questa una soluzione che si accompagna a quella vista in precedenza che teneva conto solo delle valutazioni della probabilitá. Si era visto che

$$\mu = 1 - P_0 = 1 - \frac{1-\rho}{1-\rho^{J+1}}$$

da cui si era giunti alla soluzione precedente:

$\mu_{CPU} = \frac{\rho-\rho^{J+1}}{1-\rho^{J+1}}$

Esiste una somiglianza formale tra questi due risultati. Anche se piú "rudimentale" in molti casi, per motivi di semplicitá, conviene usare quest'ultima.

0.5 Esempi

In questa sezione vengono descritte alcune applicazioni pratiche della teoria delle code nella analisi e dimensionamento dei sistemi operativi.

- Esempio 1

  Si consideri un S.O. non pre-emptive, senza iterazioni di I/O. Supponendo che il tempo medio di esecuzione dei processi sia di 0.2 secondi, trovare la frequenza massima di arrivo dei processi tale che il tempo di attesa nel sistema sia minore di 0.5 secondi.

  Sol. Il modello e' quello della coda M/M/1. Il tempo d'attesa nel sistema e' $w = \frac{1}{\mu(1-\rho)}$, dove μ e' la frequenza di elaborazione, $\mu = \frac{1}{t_e} = 5$, dove t_e e' il tempo medio di elaborazione. Inoltre λ e' la frequenza di arrivo, e ρ e' il coefficiente di traffico, ed e' uguale a $\rho = \frac{\lambda}{\mu}$. Dunque: $w = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\mu\rho} = \frac{1}{\mu-\lambda} < 0.5$. Quindi $\mu - \lambda > 1/0.5$ ovvero $\lambda < \mu - 2$. Quindi $\lambda < 3$.

- Esempio 2

  Si consideri un S.O. non pre-emptive, senza iterazioni di I/O. Si supponga che da alcune misure risulti che il tempo medio di esecuzione dei processi sia pari a 100/(frequenza clock CPU in MHz). Se il tempo medio di interarrivo e' di 3 secondi, qual'e' il clock minimo della CPU tale che il tempo d'attesa nel sistema sia minore di 0.5 secondi?

  Sol. Il modello e' quello della coda M/M/1. Il tempo d'attesa nel sistema e' $w = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda} < 0.5$. Cioe' $\mu - \lambda > 2$, cioe' $\mu > 2 + \lambda$. Visto che $\lambda = 1/3$, $\mu > 7/3$. Dato che $\mu = clock/100$, ne risulta che $clock > 233$.

- Esempio 3

  Si consideri un S.O. di tipo batch, con un dispositivo di I/O. Supponendo che la CPU lavori al ritmo di 3 elaborazioni al secondo e l'unita' di I/O soddisfi 2 richieste al secondo, qual'e' il livello (minimo) di multiprogrammazione tale che l'utilizzazione della CPU sia maggiore del 60% ?

  Sol. Il modello e' quello della coda ciclica. Dato che la utilizzazione della CPU, U, e': $U = \rho\frac{1-\rho^N}{1-\rho^{N+1}}$, dove N il numero di processi concorrenti, ovvero il livello di multiprogrammazione, dai dati del problema ottengo $\rho = 2/3$, cioe' $U = \frac{2}{3}\frac{1-\frac{2}{3}^N}{1-\frac{2}{3}^{N+1}}$. Provando con alcuni N, si vede che il minimo N che da' una utilizzazione della CPU maggiore del 60% e' $N = 4$.

- Esempio 4

  Un grosso Centro di Elaborazione Dati funziona nel seguente modo: a. Riceve da rete un blocco di processi da eseguire da parte di N utenti b. Una volta ricevuti, esegue gli N processi in multiprogrammazione c. Raccoglie i risultati e li invia agli utenti, assieme alla fattura con il costo del servizio. Sapendo che: - il tempo medio per ricevere un blocco di N richieste $T_1 = 500$ secondi - il tempo medio richiesto da un processo $T_2 = 10$ secondi - il costo fisso del centro $CS = 100$ euro/secondo - il costo del tempo d'attesa di un utente CU = 50 euro/secondo si determini il numero ottimale di utenti del blocco, cioe' il numero N che minimizza il costo totale per utente.

  [suggerimento: il costo totale per utente (tempo totale per eseguire i programmi)(costo totale al secondo)/N cio $(T_1 + NT_2)(CS + NCU)/N$

  Sol. Come suggerito, il costo totale C dato da $C = \frac{(T1+NT2)(CS+NCU)}{N}$. Volendo minimizzare C, possiamo annullare la derivata prima di C rispetto a N. Cioe':
  $\frac{dC}{dN} = \frac{(T1CU+T2CS+2NT2CU)N-(T1+NT2)(CS+NCU)}{N_2} = 0$ Cio : $T_1CUN + T_2CSN + 2N_2T_2CU = T_1CS + T_1CUN + T_2CSN + N_2T_2CU$ Ovvero : $N_2T_2CU = T_1CS$ cio $N = SQRT(T1CS/T_2CU) = SQRT(500x100/10x50) = 10$ Cio, il costo totale per utente viene minimizzato se ci sono 10 utenti del centro di elaborazione.

- Esempio 5

  Un Sistema Operativo rappresentabile nel seguente modo:

  La frequenza di elaborazione della 2 CPU é di 1 processo/s. La probabilitá che ci sia 1 processo nella coda della 2 CPU é del 25%. Sapendo che la prima CPU sia utilizzata al 50%, determinare il numero di processi in attesa della prima CPU. [ricorda: la probabilit di avere n processi in coda $\rho^n(1-\rho)$]

Figura 0.14 Il Sistema Operativo dell'esempio

Sol. Con i dati del problema, basta avere le informazioni sulla prima CPU. Le informazioni sulla seconda CPU sono irrilevanti. Infatti: utilizzazione CPU $= \rho = 0,5$. Visto che $E[N] = \sum_n np(n) = \frac{\rho}{(1-\rho)}$, il numero medio di processi in attesa della prima CPU pari a 1.

- Esempio 6

  Un Sistema Operativo possiede una CPU (che esegue in media di 50 processi al secondo) ed un disco (in media esegue 40 richieste al secondo). Sapendo che un processo ha uno spazio di indirizzamento medio di 500KB, che il Sistema Operativo non usa memoria virtuale, e che la CPU utilizzata al 60%, determinare la minima quantit di memoria di cui il Sistema Operativo deve disporre. [ricorda: la probabilit di avere n processi in coda in un sistema a coda ciclica $\rho^n \frac{1-\rho}{1-\rho^{N+1}}$ dove ρ (frequenza esecuzione disco)/(frequenza elaborazione CPU)]

  Sol. Visto che $\rho_n(t) = \rho^n \frac{1-\rho}{1-\rho^{N+1}}$, e che il coefficiente di utilizzazione $U = 1-p_0(t)=1-\frac{1-\rho}{1-\rho^{N+1}}=\rho\frac{1-\rho^N}{1-\rho^{N+1}}$. Quindi, visto che $\rho = 40/50 = 0.8$, possiamo calcolare il coefficiente di utilizzazione: $U = 0.8\frac{1-0.8^N}{1-0.8^{N+1}} = 0.6$. Sviluppando questa relazione si ha che $0.5 = 0.8^{N+1}$ da cui, facendo il logaritmo dei due membri, $N = \ln 0.5/\ln 0.8 - 1$, quindi N=2. Cioé ci sono due processi nel sistema operativo. La memoria minima richiesta quindi é di 1 Mbyte.

- Esempio 7

  In una biblioteca c'é un calcolatore dotato di un unico terminale mediante il quale gli utenti fanno ricerche sulla disponibilit e collocazione dei libri. Se un nuovo utente arriva per consultare il terminale in media ogni 3 minuti, ed il tempo medio di attesa che gli utenti aspettano prima di usare il terminale di 8 minuti, qual il numero medio di utenti che aspetta in coda?

  Sol. Secondo Little, NumeroMedioDiUtentiInCoda = freq.arrivo*TempoAttesaInCoda cio $n = \lambda * W = 1/3 * 8 = 2.66$ utenti medi in attesa.

- Esempio 8

  I processi che arrivano ad un sistema operativo provengono da varie sorgenti, tutte poissoniane, come si vede in figura:

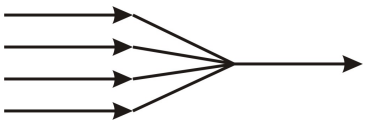


Figura 0.15 Il meccanismo degli arrivi

  I tempi medi di interarrivo sono di 10, 5, 3.33 e 2.5 secondi rispettivamente per le varie sorgenti. Qual'é la probabilitá che nel processo degli arrivi risultante arrivino 5 processi in 10 secondi?

  Sol. Le varie sorgenti hanno evidentemente le seguenti frequenze d'arrivo ($\lambda = 1/T$): rispettivamente $\lambda_1 = 0.1, \lambda_2 = 0.2, \lambda_3 = 0.3, \lambda_4 = 0.4$. La confluenza é sempre di Poisson con $\lambda = \sum_n \lambda_i$, cioé $\lambda = 1$. In accordo alla distribuzione di Poisson, la probabilitá che ci siano 5 processi arrivati in 10 secondi é pari a $(\lambda t)^5 e^{-\lambda t}/5!$ con t=10 s. E' cioé pari a $10^5 e^{-10}/120 = 0.037$.

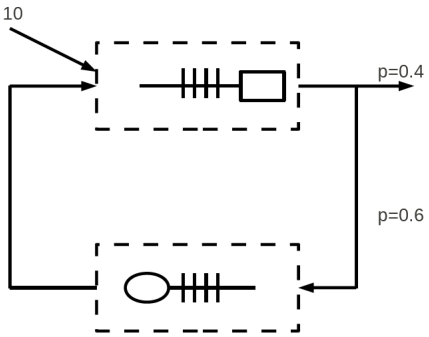
- Esempio 9

  Un sistema operativo é modellabile come una coda ciclica. Il tempo medio di CPU di 2 secondi, quello di I/O di 3 secondi. Il numero di processi nel sottosistema CPU, N(t), é una variabile aleatoria con media pari a -1/6 e varianza pari a 1. Usando l'approssimazione del fluido continuo, determinare il numero minimo di processi nel sistema tale che il coefficiente di utilizzazione della CPU sia maggiore dell'80%.

Sol. Si vuole cioé trovare il valore di J tale che U(J)>0.8. Dato che $U = \rho(1 - e^{-2mJ/var})/(1 - \rho e^{-2mJ/var})$, con i dati del problema ($\rho = 2/3, \mu = -1/6 evar = 1$) si ha che $(1-e^{-J/3})/(1.5-e^{-J/3}) > 0.8$. Questa disuguaglianza porta a $-1 > e^{-J/3}$ che non ha soluzioni nel campo reale. In effetti con i valori assegnati non si troverá mai una probabilitá maggiore di 0.8. Infatti, per J che tende a infinito, il coefficiente di utilizzazione tende a 2/3=0.6666.

- Esempio 10

  Si consideri un sistema di calcolo modellabile come in figura:



  Il disco ha una frequenza media di risposta μ_2 pari a 18.18 richieste/secondo e la CPU ha una frequenza media di esecuzione μ_1 pari a 33.3 esecuzioni/secondo. Arrivano processi dall'esterno con una frequenza di 10 processi al secondo e la probabilitá d'uscita dal sistema é di 0.4. Calcolare il numero medio di processi nell'intero sistema.

  Sol. In caso di code non pesantemente caricate, e in regime di stazionarietá, la frequenza di ingresso alla coda CPU, λ_{cpu} é pari alla frequenza d'uscita. Le richieste al disco arrivano con frequenza $\lambda_{disco} = 0.6 \cdot \lambda_{cpu}$. Si puó allora scrivere $\lambda_{cpu} = 10 + 0.6 \cdot \lambda_{cpu}$. Dalla relazione si ricava $\lambda_{cpu} = 25$ processi/secondo e $\lambda_{disco} = 25 \cdot 06 = 15$ richieste/secondo. Allora $\rho_{CPU} = 25/33.3 = 0.75$ e $\rho_{disco} = 15/18.18 = 0.825$, da cui il numero medio di processi nel sottosistema CPU $= \rho_{cpu}/1 - \rho_{cpu} = 3$ mentre in quello del disco é 4.714. Quindi il numero medio di processi nell'intero sistema é di 7.714 processi.

- Esempio 11

  Si consideri un sistema di calcolo modellabile come in figura:

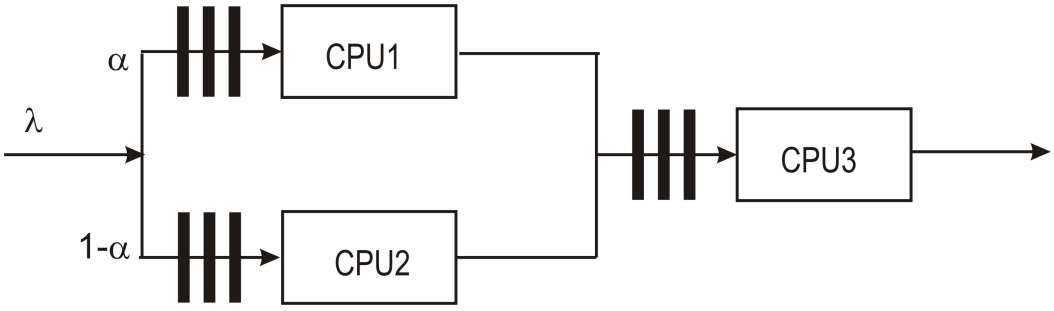


  Figura 0.16 Il Sistema Operativo dell'esempio

  I processi arrivano al sistema e si biforcano con probabilitá α e $1 - \alpha$ sui due sottosistemi di CPU1 e CPU2. Quando i processi terminano l'elaborazione sulle due CPU, confluiscono sulla terza CPU per l'elaborazione finale. Sapendo che la frequenza d'arrivo al sistema é di 1.5 processi al secondo e che la frequenza di esecuzione della terza CPU é di 2 processi al secondo, e che l'utilizzazione della CPU1 e CPU2 sono del 90% e 80% rispettivamente, calcolare l'utilizzazione di CPU3 nel caso che $\alpha = 0.5$.

  Sol. Se $\lambda = 1.5$, per $\alpha = 0.5$ si ha che la frequenza d'arrivo sulle due code é di 0.75. Visto che $U_1 = \rho_1 = \lambda_1/\mu_1 = 0.9$ si ha che la frequenza di esecuzione della prima CPU é di 0.8333 processi al secondo. Ugualmente, da $U_2 = \rho_2 = \lambda_2/\mu_2 = 0.8$ si ha che la frequenza di esecuzione della

seconda CPU é di 0.9374 processi al secondo. La frequenza di arrivo nella terza coda é la somma della frequenza di esecuzione delle prime due CPU, cioé $\lambda_3 = \mu_1 + \mu_2$, cio $\lambda_3 = 1.7708$. Da cui si ha che $U_3 = \lambda_3/\mu_3 = 0.8854$ cio 88.54 %

- esempio 12

  Un utente di un sistema operativo scrive una applicazione che risponde al paradigma del produttore/consumatore. L'applicazione composta da due processi, uno dei quali - il produttore - scrive continuamente messaggi in un buffer di 100 byte. Si consideri che:

  1. Ogni messaggio richiede 5 byte
  2. Il buffer di 100 byte non circolare: dopo che il buffer si riempie i messaggi ulteriormente prodotti non sovrascrivono altri messaggi, semplicemente vengono persi se non vengono letti.

  L'altro processo - il consumatore - legge il buffer con degli intertempi aleatori distribuiti secondo una distribuzione esponenziale con frequenza pari a 2 [letture/s]. Tenendo conto dei valori medi, trovare la frequenza massima alla quale il processo produttore scrive messaggi in modo tale che non venga perso nessun messaggio.

  Sol. Il problema si puó descrivere con una coda M/M/1 con $\mu = 2$ e λ da determinare in modo tale che la lunghezza della coda non superi mai 100 byte. Questo vuol dire che, considerando i valori medi e chiamando n il numero medio di messaggi in coda, $5 * n < 100$ ovvero $5\frac{\rho^2}{1-\rho} < 100$ ovvero $5\rho^2 + 100\rho - 100 < 0$ che ha due soluzioni: $\rho_1 = 0,954$ e $\rho_2 = -20,954$. Dal che risulta che $\rho < 0,954$ ovvero, visto che $\rho = \lambda/\mu = \lambda/2$, la frequenza media con la quale il processo scrittore scrive messaggi deve essere minore di 1,9 messaggi al secondo.

0.6 Un modello a code d'attesa della multiprogrammazione (con code caricate)

L'ipotesi é di avere un sistema chiuso, cioé con un numero costante di processi, pari a N. I processi sono distriobuiti fra le M code d'attesa, che fanno riferimento ad una CPU e $M-1$ dispositivi di I/O. Tutte le distribuzioni sono esponenziali. Il sistema é visualizzato nella seguente figura, dove le probabilitá p_i sono delle probabilitá costanti di arrivo sulle varie code e μ_1, μ_2, ... μ_M sono le frequenze relative alla CPU e ai dispositivi di I/O rispettivamente.

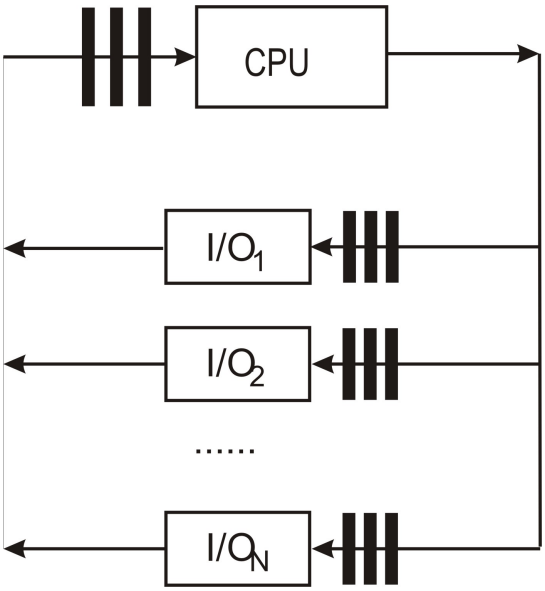


Figura 0.17 Rappresentazione del sistema operativo multiprogrammato in esame.

In queste ipotesi semplificativi si puó dimostrare con una tecnca simile a quella utilizzata precedentemente, che la probabilitá che ci siano n_1 processi nella prima coda, n_2 processi nella seconda coda e cos via data dalla seguente relazione:

$$p(n_1, n_2, \ldots, n_M) = \frac{1}{G(N)} \prod_{i=1}^{M} X_i^{n_i} \tag{7}$$

dove $G(N)$ una costante relativa ad un sistema con N processi e M code. Inoltre, si dimostra che $X_1 = 1$ e $X_i = \frac{\mu_1 p_i}{\mu_i}$.

Naturalmente in un sistema con N processi, $\sum_{i=1}^{M} n_i = N$.

0.6.1 Primo problema

In quanti modi si possono distribuire N processi su M code? Questo é ovviamente un problema combinatorio. Per trovare la soluzione e per visualizzare il problema, é bene enumerare alcune possibili casi semplici. Se nel sistema ci sono due code e due processi, cioé una CPU e un dispositivo di I/O, ci sono solo tre possibilitá, cioé il caso di un processo nella coda CPU ed un processo nella coda di I/O, il caso in cui i due processi sono entrambi nella coda CPU e il caso in cui i due processi sono entrambi nella coda di I/O. Analogamente si puó vedere che se ci sono tre processi, ci sono 4 possibili distribuzioni sulle due code, e se ci sono 4 processi le possibilitá salgono a 5. Se ci sono 3 code nel sistema, cioé una CPU e due I/O, allora le possibilitá sono visualizzate nella seguente tabella per 2, 3 e 4 processi.

	2processi	3 processi	4 processi
nr. code	1 2 3	1 2 3	1 2 3
nr. proc.	0 0 2	0 0 3	0 0 4
	0 2 0	0 3 0	0 4 0

2 0 0	3 0 0	0 2 2
0 1 1	1 1 1	0 3 1
1 0 1	1 0 2	0 1 3
1 1 0	1 2 0	1 0 3
	2 1 0	1 3 0
	2 0 1	1 1 2
	0 2 1	2 2 1
	0 1 2	2 1 1
		2 0 2
		2 2 0
		3 0 1
		3 1 0
		4 0 0

Le sequenze $(2,2) \to 3, (2,3) \to 4, (2,4) \to 5, (3,2) \to 6, (3,3) \to 10, (3,4) \to 15$ possono essere ottenute con la formula delle combinazioni di N-M-1 elementi su N. Quindi: N processi si possono distribuire su M code in $\begin{pmatrix} M+N-1 \\ N \end{pmatrix}$ modi. Sia S l'insieme delle possibili combinazioni $(n_1, n_2, \ldots, n_M)$; per esempio per N=2, M=2, $S = (02), (20), (11)$. Allora $\sum_S p(n_1, n_2, \ldots, n_M) = 1$. Data la relazione precedente, si ha:

$$G(N) = \sum_S \prod_{i=1}^{M} X_i^{n_i} \tag{8}$$

0.6.2 Secondo problema

Come calcolare la costante G(N)? J.P.Buzen propone una soluzione ricorsiva a questo problema, introducendo la funzione ausiliaria $g(n,m) = \sum_F \prod_{i=1}^{m} X_i^{n_i}$ che é riferita ad un sistema con m code e n processi, cioé un sistema per il quale ci sono $\begin{pmatrix} m+n-1 \\ n \end{pmatrix}$ possibili combinazioni dei processi sulle code. Data la definizione, $G(N) = g(N,M)$. Per arrivare alla ricorsione si divide la sommatoria in due termini, uno per le configurazioni di $(n_1, n_2, \ldots, n_M)$ per cui $n_m = 0$ e l'altro per le configurazioni per cui $n_m > 0$. Facendo riferimento alla tabella precedente, se ho 3 code e 2 processi, e per quanto riguarda la costante G(N) ci corrisponde a :

$$G(N) = \sum_S \prod_{i=1}^{M} X_i^{n_i} = \sum_{(020),(200),(110)} \prod_{i=1}^{M} X_i^{n_i} + \sum_{(002),(011),(101)} \prod_{i=1}^{M} X_i^{n_i} \tag{9}$$

Dato che nel primo insieme $n_m = 0$, la produttoria del primo termine della sommatoria puó essere limitata a (m-1), quindi la prima sommatoria descrive la costante in un sistema con $m-1$ code. Per quanto riguarda la seconda sommatoria, mettiamo a fattore comune X_m. L'esponente di X_m diventa allora n_{m-1} e quindi la sommatoria descrive un sistema con $n-1$ processi perché la somma degli esponenti é $n-1$. In definitiva:

$$g(n,m) = g(n,m-1) + X_m g(n-1,m) \tag{10}$$

Inoltre $g(n,1) = \sum_{F,m=1} \prod_{i=1}^{m} X_i^{n_i} = X_1^{n_1}$ per tutti gli N da 0 a N e, dalla ricorsione, g(0,m)=1 per tutti gli m da 1 a M. La ricorsione consente di calcolare $g(N,M)$ in NM passi, partendo da $(1,2)$ in avanti. Un esempio di determinazione dato nella seguente tabella.

```
g(1,2)=g(1,1)+X2*g(0,2); g(1,3)=g(1,2)+X3*g(0,3); ...
g(1,M)=g(1,M-1)+XM*g(0,M) g(2,2)=g(2,1)+X2*g(1,2);
g(2,3)=g(2,2)+X3*g(1,3); ... g(2,M)=g(2,M-1)+XM*g(1,M)
g(3,2)=g(3,1)+X2*g(2,2); g(3,3)=g(3,2)+X3*g(2,3); ...
g(3,M)=g(3,M-1)+XM*g(2,M) ... g(N,2)=g(N,1)+X2*g(N-1,2);
g(N,3)=g(N,2)+X3*g(N-1,3); ... g(N,M)=g(N,M-1)+XM*g(N-1,M)
```

alla fine $G(N) = g(N,M)$. Da osservare che $g(0,M) = G(0)$, $g(1,M) = G(1)$, $g(2,M) = G(2)$,..., $g(N,M) = G(N)$.

0.6.3 Terzo problema

Un'altra cosa interessante riguarda la determinazione della probabilitá $p\{n_i \geq k\}$.

Chiaramente $p(n_i \geq k) = \sum_{S \cap n_i \geq k} p(n_1, n_2, \ldots n_M) = \sum_{S \cap n_i \geq k} \frac{1}{G(N)} \prod_{n_i}^{M} X_i^{n_i} = \frac{1}{G(N)} \sum_{S \cap n_i \geq k} \prod_{n_i}^{M} X_i^{n_i} = \frac{X_i^k}{G(N)} \sum X_1^{n_1} X_2^{n_2} \ldots X_i^{n_i - k} \ldots X_M^{n_M} = X_i^k \frac{G(N-k)}{G(N)}$.

Ora, si osservi che $p(n_i \geq 1)$ il coefficiente di utilizzazione della coda i. Quindi i coefficienti di utilizzazione (cioé la probabilitá che stiano eseguendo o processando richieste) delle varie code sono: $U_{CPU} = \frac{G(N-1}{G(N)}, U_{I/O_1} = X_2 \frac{G(N-1}{G(N)}, \ldots, U_{I/O_{M-1}} = X_M \frac{G(N-1}{G(N)}$ dove $X_i = \frac{\mu_1 p_i}{\mu_i)}$.

0.6.4 Un esempio applicativo

Si consideri un sistema operativo multiprogrammato con una CPU, e due dispositivi di I/O. Il numero di processi concorrenti nel sistema due. Misure sul sistema evidenziano che: $\mu_1 = 100, \mu_2 = 25, \mu_3 = 40, p_1 = 0.1, p_2 = 0.2 e p_3 = 0.7$. Ci si chiede: se si vuole aumentare l'utilizzazione della CPU di almeno 10% in termini assoluti, é sufficiente aumentare la memoria in modo che il numero di processi concorrenti raddoppi? (da 2 passa a 4)

Con i dati si ottiene: $X_1 = 1, X_2 = 0.8, X_3 = 1.75$. Inoltre $G(0) = 1, G(1) = 3.55, G(2) = 8.65$. Quindi l'utilizzazione della CPU $U = G(1)/G(2) = 41\%$. Se i processi concorrenti passano da 2 a quattro, si ottiene: $G(3) = 18, G(4) = 35$. Quindi il nuovo coefficiente di utilizzazione della CPU $U = 52\%$. Quindi aumenta piú del 10% come si voleva.