Teoria delle code

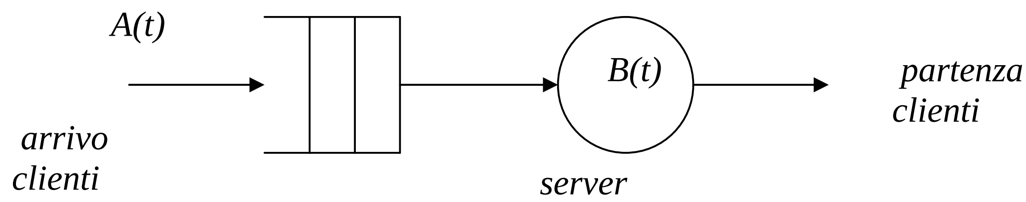
- Notazione, politica di servizio, misure di performance
- Legge di Little
- Code di servizio markoviane
- Reti di code markoviane:
  - *reti aperte*
  - *reti chiuse*
  - *reti in forma prodotto*

coda di servizio: cfr. *catena di Markov nascita/morte*

specifica completa di un modello di coda di servizio:

- **modelli stocastici dei processi di arrivo e di servizio**
- **specifica dei parametri strutturali** (capacità della coda, numero di server, …)
- **specifica delle politiche di servizio** (priorità dei clienti, condizioni per accettare/rifiutare clienti, …)

Modelli stocastici dei processi di arrivo e di servizio



$\Psi = \{a, p\}$

Evento arrivo $\leftrightarrow \{Y_1, Y_2, Y_3, \ldots\}$ Y_k: k-esimo *tempo di interarrivo* - tempo trascorso fra l'arrivo (k-1)-esimo e l'arrivo k-esimo
se $\{Y_k\}$ i.i.d. $\Rightarrow A(t) = P(Y \le t)$
la distribuzione di probabilità $A(t)$ *descrive completamente la sequenza dei tempi di interarrivo*
media della distribuzione $A(t)$: $E[Y] = 1/\lambda$ - λ:*frequenza media di arrivo dei clienti*

Evento partenza $\leftrightarrow \{Z_1, Z_2, Z_3, \ldots\}$ Z_k: k-esimo *tempo di servizio*
se $\{Z_k\}$ i.i.d. $\Rightarrow B(t) = P(Z \le t)$
la distribuzione di probabilità $B(t)$ *descrive completamente la sequenza dei tempi di servizio*
media della distribuzione $B(t)$: $E[Z] = 1/\mu$ - μ:*frequenza media di servizio*

Parametri strutturali

K : *capacità* della coda - K = 1, 2, …, ∞

m : *numero di server* - m = 1, 2, …, ∞

Nei casi visti finora (catene nascita/morte): K = ∞, m = 1

Politica di servizio

- **numero di classi di clienti** (sistemi a singola classe, sistemi a classi multiple)
- **politica di scheduling** (per sistemi a classi multiple, determina i criteri di decisione del prossimo cliente da servire)
- **disciplina di coda** (anche per sistemi a singola classe – descrive l'ordine con cui il server seleziona i clienti della stessa classe, es.: FIFO, LIFO, …)
- **politica di ammissione** (anche per code a capacità ∞)

Notazione di Kendall

A / B / *m* / K

distribuzione tempi di interevento

distribuzione tempi di servizio

numero di server

capacità della coda

convenzioni su A, B:

G : distribuzione generica (qualsiasi)
GI : distribuzione generica i.i.d.
D : distribuzione deterministica
M : distribuzione Markoviana ($\rightarrow$ esponenziale)

convenzione su K:

se $K = \infty$, si omette

Misure di prestazione di una coda di servizio

Y_k : tempo di interarrivo
Z_k : tempo di servizio

A_k : *tempo di arrivo*
D_k : *tempo di partenza*
W_k : *tempo di attesa*
S_k : *tempo nel sistema*

tutti riferiti al k-esimo cliente

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

All' istante *t* :

$X(t)$: *lunghezza della coda*
$U(t)$: *carico di lavoro*

$U(t)$: tempo necessario per svuotare la coda a partire dall'istante *t*

Spesso $P(W_k \leq t) \to_{k\to\infty} P(W \leq t)$

La variabile aleatoria W descrive il tempo di attesa di un utente *a regime*

E[W] = *tempo medio di attesa* a regime

Allo stesso modo, se $P(S_k \leq t) \to_{k\to\infty} P(S \leq t)$, allora:

E[S] = *tempo medio di permanenza nel sistema*

…

E[X] = *lunghezza media della coda*
E[U] = *carico di lavoro medio*

tutti per condizioni di regime
(cfr. $\pi_n = P(X \leq n)$)

Obiettivi nel progetto di una coda di servizio

- minimizzare il tempo di attesa medio a regime
- massimizzare l'utilizzazione del server (mantenere sempre la coda non vuota)

obiettivi contrastanti !!

compromesso fondamentale nel progetto e controllo di code di servizio:

- per mantere un server pienamente utilizzato, si devono tollerare lunghi tempi di attesa;
- per ottenere tempi di attesa contenuti, si deve tollerare che il server rimanga a periodi inutilizzato

Metodo di analisi: *si suppone che esista una situazione di regime*, e si considerano gli indici di prestazione: E[W] E[S] E[X]

da mantenere il più possibile piccoli

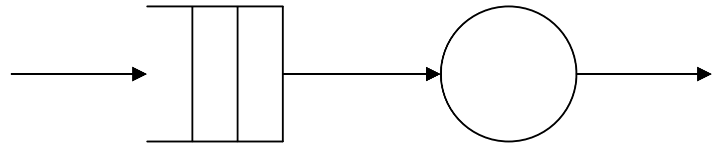
e inoltre i seguenti ulteriori indici:

utilizzazione: frazione del tempo in cui il server è impiegato

throughput: frequenza con cui i clienti lasciano il sistema dopo aver ricevuto servizio

intensità del traffico: rapporto fra la frequenza media di arrivo e la frequenza media di servizio

Esempio : G/G/1, FIFO



intensità del traffico: $\rho \equiv \dfrac{\lambda}{\mu}$ (se G/G/*m* : $\rho \equiv \dfrac{\lambda}{m\mu}$)

Poiché stiamo supponendo esistenza di regime, se G/G/1:

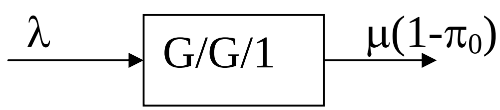
utilizzazione: $1 - \pi_0$

throughput: $\mu(1-\pi_0)$

\# medio di servizi nell'unità di tempo

frazione di tempo in cui il server è attivo

A regime, il flusso di clienti in ingresso ed in uscita dal sistema deve essere bilanciato:



$$\Rightarrow \lambda = \mu(1-\pi_0) \Rightarrow \rho = \frac{\lambda}{\mu} = 1-\pi_0$$

intensità del traffico $\equiv$ *utilizzazione del sistema*

$\pi_0 = 1$ $\rightarrow$ non ci sono arrivi ($\lambda = 0$)

$\pi_0 = 0$ $\rightarrow$ sistema permanentemente impiegato (in genere situazione instabile)

Dinamica di una coda di servizio

Arrivo del k-esimo cliente (disciplina: FIFO)

caso 1: sistema vuoto $\Rightarrow W_k = 0$; il sistema è vuoto se e solo se la (k-1)-esima partenza ha preceduto il k-esimo arrivo: $D_{k-1} \leq A_k$

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

caso 2: sistema non vuoto $\Rightarrow W_k > 0$; il k-esimo cliente attende la partenza del (k-1)-esimo cliente

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

Unendo i due casi: $W_k = \max \{ 0, D_{k-1} - A_k \}$

e ricordando che: $D_k = A_k + W_k + Z_k \quad \forall\, k$

$A_k - A_{k-1} = Y_k$

$\Rightarrow D_{k-1} - A_k = D_{k-1} - (A_{k-1} + Y_k) = W_{k-1} + Z_{k-1} - Y_k$

$$\mathbf{W_k = \max \{ 0, W_{k-1} + Z_{k-1} - Y_k \}}$$

equazione di Lindley

poiché $S_k = W_k + Z_k \quad \forall\, k,$ $\qquad \mathbf{S_k = max\ \{ 0, S_{k-1} - Y_k\} + Z_k}$

e infine: $D_k = A_k + W_k + Z_k = A_k + \max \{ 0, D_{k-1} - A_k \} + Z_k \Rightarrow \mathbf{D_k = max\ \{ A_k, D_{k-1} \} + Z_k}$

Le equazioni ricorsive per W_k, S_k e D_k sono del tutto generali ed indipendenti dalle distribuzioni di probabilità degli arrivi e delle partenze

Nell'equazione di Lindley, supponendo si abbia sempre $Y_k < W_{k-1} + Z_{k-1}$, si ha:

$$W_k = W_{k-1} + Z_{k-1} - Y_k$$

dinamica lineare!

I tempi di attesa avrebbero dinamica lineare se non per l'effetto di clienti il cui tempo di interarrivo è tale che $Y_k > W_{k-1} + Z_{k-1} = S_{k-1}$

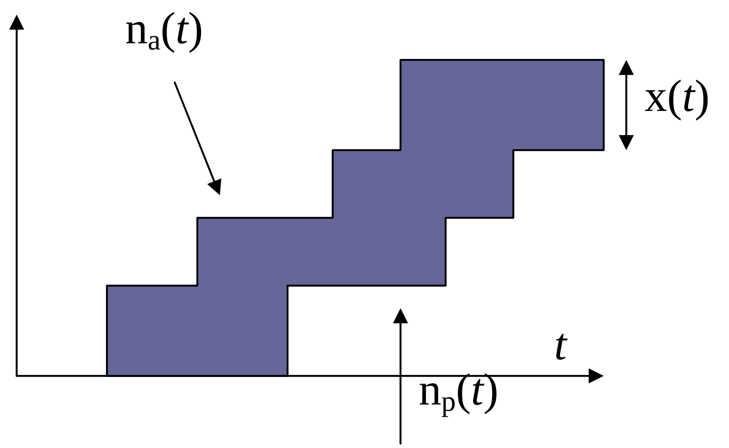
Legge di Little

coda G/G/1, disciplina FIFO

$\Psi = \{a, p\}$, $N_a(t)$, $N_p(t)$: *contatori*, coda inizialmente vuota

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

u(*t*): *ammontare totale di tempo speso da tutti i clienti all'interno del sistema*



$$\bar{s}(t) = \frac{u(t)}{n_a(t)}$$ tempo medio nel sistema per cliente

$$\bar{x}(t) = \frac{u(t)}{t}$$ lunghezza media della coda

$$\lambda(t) = \frac{n_a(t)}{t}$$ frequenza media di arrivo dei clienti

$$\Rightarrow \bar{x}(t) = \lambda(t)\bar{s}(t)$$ *relazione valida sempre su ogni realizzazione del processo* (movimento del sistema)

Ipotesi cruciali:

A – Esistono finiti i seguenti limiti:

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

B – tali limiti esistono (con lo stesso valore λ e $\bar{s}$) indipendentemente dalla particolare realizzazione del processo ($\Leftrightarrow$ tempo di arrivo e tempo nel sistema sono processi *ergodici*)

Allora:
anche la lunghezza della coda è un processo ergodico, e soddisfa la relazione:

$$\boxed{\mathbf{E[X] = \lambda\, E[S]}}$$

legge di Little

Osservazioni sulla legge di Little

- *indipendente dalla struttura di clock* associata alle partenze ed agli arrivi (dalle distribuzioni di probabilità di arrivi e partenze)
- *indipendente dalla politica/disciplina di servizio* (l'ipotesi FIFO non è mai stata usata)
- *valida per qualsiasi configurazione di code e server interconnessi*
- *valida solo a regime* (sincerarsi che i limiti siano stati raggiunti prima di invocare la legge)
- la frequenza di arrivi λ deve essere la *frequenza degli ingressi effettivi nel sistema*

Quando la coda è a capacità finita, e/o vi è controllo sulle ammissioni nel sistema, attenzione all'uso ed alle definizioni di λ

Code di servizio markoviane

Le misure di performance di una coda di servizio possono essere determinate dalla distribuzione di probabilità *a regime* della lunghezza della coda: $\pi_n = P(X = n), \qquad n = 0,1,2,...$

In generale una coda di servizio è un DES con: $\Sigma = \{0, 1, 2, \ldots\}$; $\Psi = \{a, p\}$; $X(t)$: stato : lunghezza della coda - *processo semi-Markov generalizzato*

Problema: dato un processo semi-Markov generalizzato, determinare, se esiste, la distribuzione di probabilità a regime dello stato

2 possibili metodi di soluzione:

1. fissare una specifica struttura di clock e derivare espressioni analitiche per π_n
2. attraverso simulazioni o osservazioni dirette di movimenti del sistema, stimare le probabilità a regime (es.: se il periodo di osservazione è T, ed il tempo trascorso nello stato *n* è T_n, $\Rightarrow \hat{\pi}_n = \dfrac{T_n}{T}$

Noi useremo l'approccio 1, ed in particolare consideremo **tempi di servizio e di interarrivo distribuiti esponenzialmente con parametri μ e λ rispettivamente $\Rightarrow$ catena di Markov nascita-morte**

Per le catene nascita-morte, con λ_k, μ_k frequenze di nascita e morte nello stato *k*, valgono le seguenti espressioni per le probabilità di stato a regime:

$$\boxed{\pi_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \pi_0 \,, \qquad n = 1,2,\ldots}$$

$$\boxed{\pi_0 = \frac{1}{1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j}}}$$

coda di servizio M/M/1

Arrivi markoviani; tempi di servizio markoviani; singolo server; (capacità infinita)

$$\lambda_n = \lambda \qquad \forall n = 0,1,2,\cdots \quad \mu_n = \mu \qquad \forall n = 1,2,3,\cdots$$

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

sotto la condizione $\frac{\lambda}{\mu} < 1$ *!!*

$\rho = \frac{\lambda}{\mu}$ *: intensità di traffico* $(0 \le \rho < 1)$

$\pi_0 = 1 - \rho$

$\pi_n = P(X = n) = \left(\frac{\lambda}{\mu}\right)^n \pi_0 = \left(\frac{\lambda}{\mu}\right)^n (1 - \rho) = (1 - \rho)\rho^n$ *distribuzione di probabilità a regime della lunghezza di coda per un sistema M/M/1*

La condizione $\rho < 1$ è anche chiamata *condizione di stabilità* per il sistema M/M/1

Per stabillità in questo contesto si intende l'esistenza di un regime

Sistemi M/M/1:

utilizzazione: $1-\pi_0=\rho$

throughput: $\mu(1-\pi_0)=\lambda$ (a regime arrivi e partenze si bilanciano)

se $\lambda>\mu$, "a regime" il throughput è μ, anche se $\nexists\pi_0$

lunghezza media della coda:

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

oss.: $\frac{d}{d\rho}\left(\sum_0^\infty\rho^n\right)=\frac{1}{\rho}\sum_0^\infty n\rho^n=\frac{d}{d\rho}\frac{1}{1-\rho}=\frac{1}{(1-\rho)^2}$

$$\boxed{\Rightarrow E[X]=\frac{\rho}{1-\rho}}$$

tempo medio nel sistema: dalla legge di Little: $E[X]=\lambda E[S]$

$\Rightarrow E[S]=\frac{1}{\lambda}\frac{\rho}{(1-\rho)}=\frac{1}{\mu}\frac{1}{(1-\rho)}$ per $\rho\to 0,\ E[S]\to 1/\mu$

tempo medio di attesa

$$E[S]=E[W]+E[Z]\Rightarrow E[W]=\frac{1}{\mu}\frac{1}{(1-\rho)}-\frac{1}{\mu}=\frac{\rho}{\mu(1-\rho)}$$

Esempio:

Dimensionare la frequenza di servizio μ di un sistema di elaborazione che riceve job con frequenza media λ = 1 job/s in maniera che il tempo medio nel sistema non ecceda a regime 0.5 s.

L'arrivo dei job forma un processo di Poisson, ed i tempi di servizio sono distribuiti esponenzialmente

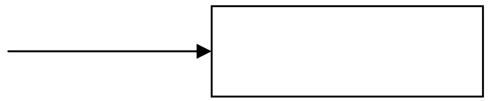
coda M/M/1

$$E[S] = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu\left(1 - \frac{1\,job/s}{\mu}\right)} = \frac{1}{\mu - 1} < 0.5s$$

$$\Rightarrow 0.5\mu > 1.5 \Rightarrow \mu > 3 \text{ job/s}$$

Esempio: linea di trasmissione con capacità c = 1200 b/s (c: velocità di trasmissione)

pacchetto di L bit



L: variabile aleatoria, distribuita esponenzialmente, E[L] = 600 b

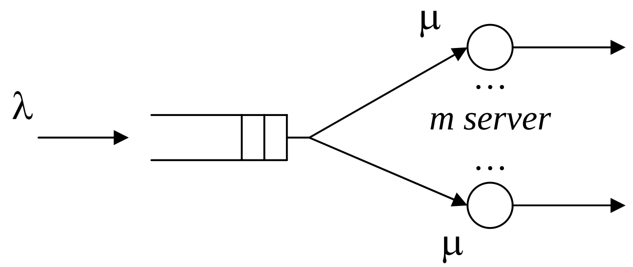
Determinare il numero medio massimo di messaggi in arrivo per unità di tempo tale da garantire un tempo di attesa medio per messaggio inferiore ad 1 s.

server: linea di trasmissione
frequenza di trasmissione messaggi μ: in media un messaggio richiede E[L]/c secondi $\Rightarrow \mu$ = c/E[L] = 2 messaggi/s

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

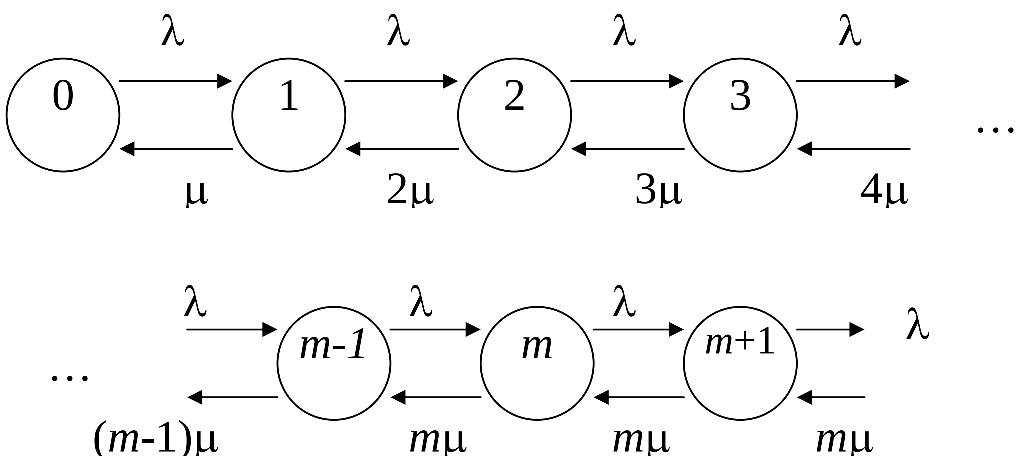
$$E[W] = \frac{\rho}{\mu(1-\rho)} < 1 \Rightarrow \lambda < 4/3 \Leftrightarrow \rho < 2/3$$

Sistemi M/M/m
(capacità della coda infinita)



L' effettiva frequenza di servizio varia con il numero di utenti presenti: se n clienti, $n < m$, allora la frequenza di servizio è $n\mu$

catena nascita-morte:



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

$$= \frac{\lambda^n}{n!\,\mu^n}$$

stabilità della coda ($\Leftrightarrow$ esistenza di π_0 a regime): $\dfrac{\lambda}{m\mu} < 1$

Sia $\rho = \dfrac{\lambda}{m\mu}$; allora $\displaystyle\sum_{n=m}^{\infty} \left(\frac{\lambda}{m\mu} \right)^{n-m+1} = \rho \sum_{k=0}^{\infty} \rho^k = \frac{\rho}{1-\rho}$

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

$$= \left(\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!} \frac{1}{1-\rho} \right)^{-1}$$

Determinazione di π_n:

caso 1: $n < m$:

$$\pi_n = \left(\frac{\lambda^n}{(\mu)(2\mu)\cdots(n\mu)} \right) \pi_0 = \frac{(m\rho)^n}{n!} \pi_0$$

caso 2: $n \geq m$:

$$\pi_n = \left(\frac{\lambda^{m-1}}{(\mu)(2\mu)\cdots(m-1)\mu} \right) \left(\frac{\lambda^{n-m+1}}{(m\mu)^{n-m+1}} \right) \pi_0 = \frac{(m\rho)^{m-1}}{(m-1)!} \rho^{n-m+1} \pi_0 = \frac{m^m}{m!} \rho^n \pi_0$$

In definitiva:

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

Utilizzazione

Sia B la variabile aleatoria che indica il numero dei server occupati - $B \in [0,1,\ldots,m]$

$$E[B] = \sum_{n=0}^{m-1} n\pi_n + mP(X \geq m); \qquad P(X \geq m) = \sum_{n=m}^{\infty} \pi_n = \sum_{n=m}^{\infty} \frac{m^m}{m!}\rho^n \pi_0 = \frac{m^m}{m!}\frac{\rho^m}{1-\rho}\pi_0$$

$$\Rightarrow E[B] = \sum_{n=0}^{m-1} n\frac{(m\rho)^n}{n!}\pi_0 + \frac{m(m\rho)^m}{m!}\frac{1}{1-\rho}\pi_0 = \left(\sum_{n=1}^{m-1}\frac{(m\rho)^n}{(n-1)!} + \frac{(m\rho)^m}{m!}\frac{m}{1-\rho}\right)\pi_0 =$$

$$= m\rho\left(\sum_{n=1}^{m-1}\frac{(m\rho)^{n-1}}{(n-1)!} + \frac{(m\rho)^{m-1}}{m!}\frac{m}{1-\rho}\right)\pi_0 = m\rho\left(\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \frac{(m\rho)^{m-1}}{m!}\frac{m}{1-\rho} - \frac{(m\rho)^{m-1}}{(m-1)!}\right)\pi_0 =$$

$$= m\rho\underbrace{\left(\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!}\frac{1}{1-\rho}\right)}\pi_0$$

π_0^{-1} !!!! (how convenient!) $\quad \Rightarrow E[B] = m\rho = \dfrac{\lambda}{\mu}$

ogni server nel sistema ha una utilizzazione pari a: $\dfrac{E[B]}{m} = \rho$

throughput: λ

(in condizioni di regime partenze ed arrivi si bilanciano)

lunghezza media della coda : $E[X] = \sum_{n=0}^{\infty} n\pi_n$

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

tempo medio nel sistema dalla legge di Little: $E[X] = \lambda E[S]$

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

Probabilità di accodamento

Probabilità che al suo arrivo il cliente non trovi nessun server libero e debba attendere nell'area di coda

Attenzione!

A_n = [il cliente in arrivo all'istante *t* trova X(*t*) = *n*]

B_n = [lo stato del sistema al generico istante *t* è X(*t*) = *n*]

A_n e B_n sono due eventi distinti – nel caso di A_n l' osservazione dello stato del sistema avviene in specifici istanti di tempo che dipendono dal processo di arrivo

Per processi di arrivo e di servizio generici, $P(A_n) \neq P(B_n)$

Poisson Arrival See Time Averages

Teorema: siano $\alpha_n(t) = P(A_n), \pi_n(t) = P(B_n)$; per una coda di servizio con processo di arrivo poissoniano indipendente dal processo di servizio, e con processo di servizio *generico* ed indipendente dal processo di arrivo, si ha: $\alpha_n(t) \equiv \pi_n(t)$

Inoltre, se tali probabilità esistono a regime, si ha:

$$\bar{\alpha}_n = \lim_{t\to\infty} \alpha_n(t)$$
$$\pi_n = \lim_{t\to\infty} \pi_n(t)$$
$$\bar{\alpha}_n \equiv \pi_n$$

Probabilità di accodamento

Sia P_Q la probabilità di accodamento *a regime*. Per la proprietà *PASTA*: $P_Q = P(X \geq m) = \sum_{n=m}^{\infty} \pi_n$

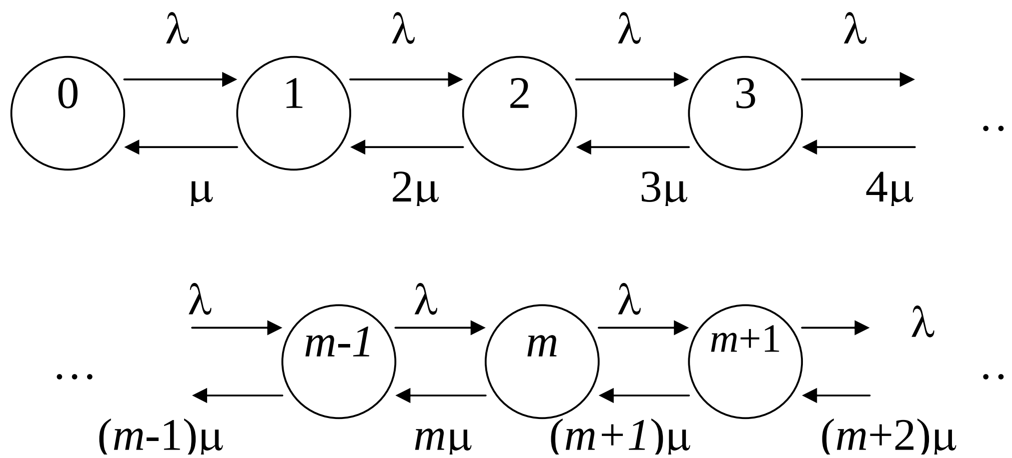
Questa grandezza è già stata valutata nel calcolo di E[B]:

$$\Rightarrow P_Q = \frac{(m\rho)^m}{m!} \frac{\pi_0}{1-\rho}$$

formula C di Erlang

λ : frequenza di arrivi di chiamate telefoniche
$1/\mu$: durata media di una chiamata
m : numero necessario di linee telefoniche per garantire $P_Q < p$

Sistemi M/M/∞



$$\pi_0 = \left(1 + \sum_{n=1}^{\infty} \frac{\lambda^n}{(\mu)(2\mu)\cdots(n\mu)}\right)^{-1} = \left(1 + \sum_{n=1}^{\infty} \frac{(\lambda/\mu)^n}{n!}\right)^{-1} = \frac{1}{e^{\lambda/\mu}}$$

se $\rho = \dfrac{\lambda}{\mu} < \infty$ (attenzione: $\rho \neq$ intensità di traffico!)

$$\Rightarrow \pi_0 = e^{-\rho}, \quad \pi_n = e^{-\rho}\frac{\rho^n}{n!}$$

distribuzione di Poisson con parametro ρ

utilizzazione: $1-\pi_0 = 1-e^{-\rho}$

throughput: λ

lunghezza media della coda: $\rho = \dfrac{\lambda}{\mu}$

(media della distribuzione poissoniana)

tempo medio nel sistema: dalla legge di Little: $E[X] = \lambda E[S]$

$$E[S] = \frac{\rho}{\lambda} = \frac{1}{\mu} = E[Z]$$

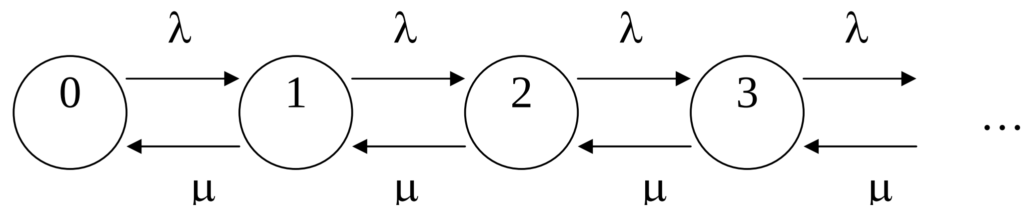
i clienti trovano sempre un server disponibile
e non devono mai aspettare

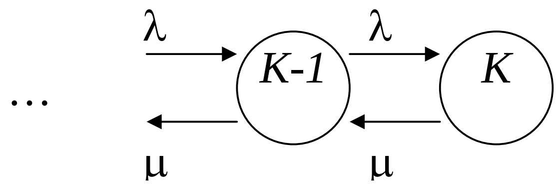
Il modello M/M/∞ è chiaramente non realistico. Tuttavia è utile nel modellare situazioni in cui i clienti non sono mai accodati

p.es., nei processi manifatturieri, il trasporto di pezzi e materiali con cinghie di trasmissione perpetue

La coda M/M/∞ serve come elemento di "ritardo puro"

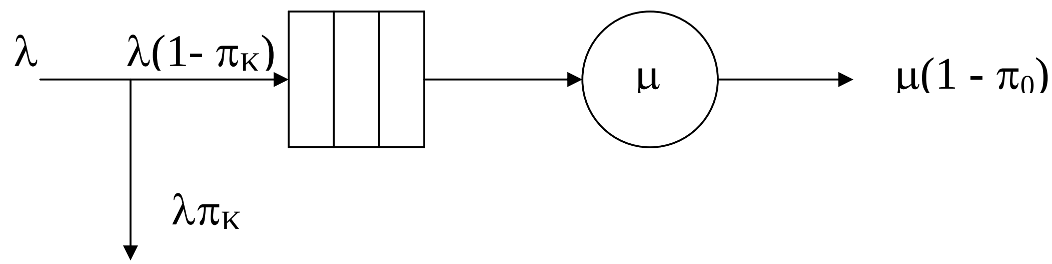
Sistemi M/M/1/K





fenomeno del ***blocking***

"sospendiamo" il processo (poissoniano) degli arrivi quando la coda ha lunghezza K. Lo riattiviamo non appena la coda ha lunghezza < K



$$\pi_0 = \left(\sum_{h=0}^{K} \left(\frac{\lambda}{\mu} \right)^h \right)^{-1} = \left(\frac{1-\rho^{K+1}}{1-\rho} \right)^{-1} = \frac{1-\rho}{1-\rho^{K+1}}$$

$$\pi_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n \qquad 0 < n \le K$$

non ci sono problemi di convergenza - $\rho = \lambda/\mu \neq$ intensità traffico (alcuni ingressi sono bloccati)

$$\textit{utilizzazione: } 1-\pi_0 = \rho\frac{1-\rho^K}{1-\rho^{K+1}}$$

se $\rho \to \infty$, utilizzazione $\to$ 1;

se $\rho < 1, K \to \infty$, utilizzazione $\to$ ρ (caso M/M/1)

$$\textit{throughput: } \mu(1-\pi_0) = \lambda\frac{1-\rho^K}{1-\rho^{K+1}}$$

minore del tasso di arrivi esterni, perché alcuni utenti trovano situazione di blocco

probabilità di blocco: per la proprietà *PASTA:*

$$P_B = \pi_K = (1-\rho)\frac{\rho^K}{1-\rho^{K+1}}$$

lunghezza media della coda

$$E[X] = \sum_{n=0}^{K} n\pi_n = \frac{1-\rho}{1-\rho^{K+1}} \underbrace{\sum_{n=0}^{K} n\rho^n}_{A}$$

$$A = \rho + 2\rho^2 + 3\rho^3 + \cdots + K\rho^K$$

$$A - \rho A = \rho + \rho^2 + \rho^3 + \cdots + \rho^K - K\rho^{K+1} = \sum_{h=1}^{K} \rho^h - K\rho^{K+1}$$

$$(1-\rho)A = \rho\frac{1-\rho^{K+1}}{1-\rho} - K\rho^{K+1} \Rightarrow A = \rho\frac{1-\rho^{K+1}}{(1-\rho)^2} - \frac{K\rho^{K+1}}{1-\rho}$$

$$\Rightarrow E[X] = \frac{\rho}{1-\rho^{K+1}}\left(\frac{1-\rho^K}{1-\rho} - K\rho^K\right)$$

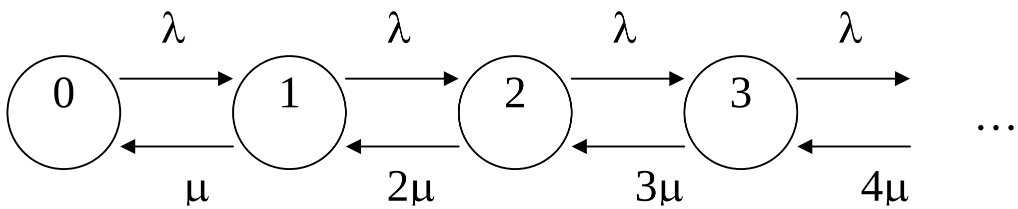
tempo medio nel sistema

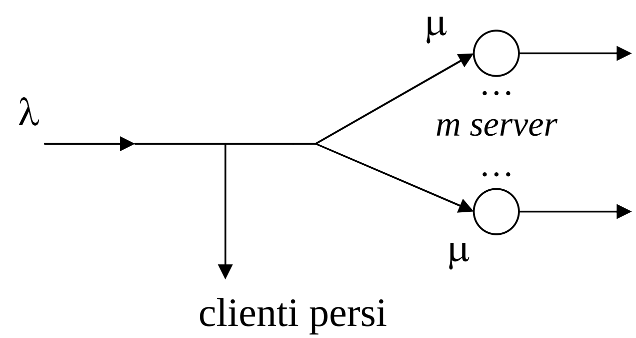
legge di Little … attenzione!! La frequenza effettiva degli ingressi è $\lambda(1-\pi_K)$

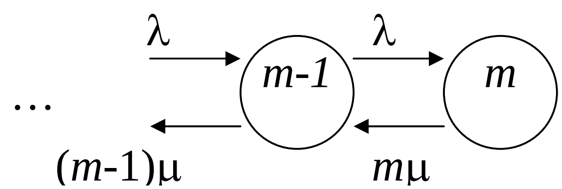
$$E[S] = \frac{E[X]}{\lambda(1-\pi_K)}$$

Sistemi M/M/m/m

m server identici senza spazio di attesa







$$\pi_0 = \left(1 + \sum_{n=1}^{m} \frac{\lambda^n}{(\mu)(2\mu)\cdots(n\mu)}\right)^{-1} = \left(1 + \sum_{n=1}^{m} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right)^{-1} = \left(\sum_{n=0}^{m} \frac{\rho^n}{n!}\right)^{-1}$$

$\rho = \dfrac{\lambda}{\mu} \neq$ intensità traffico !!

$$\pi_n = \frac{\rho^n}{n!}\pi_0 = \frac{\rho^n}{n!} \frac{1}{\sum_{j=0}^{m} \frac{\rho^j}{j!}}$$

Probabilità di blocco:

PASTA !!

$$P_B = \pi_m = \frac{(\rho^m / m!)}{\sum_{j=0}^{m} (\rho^j / j!)}$$

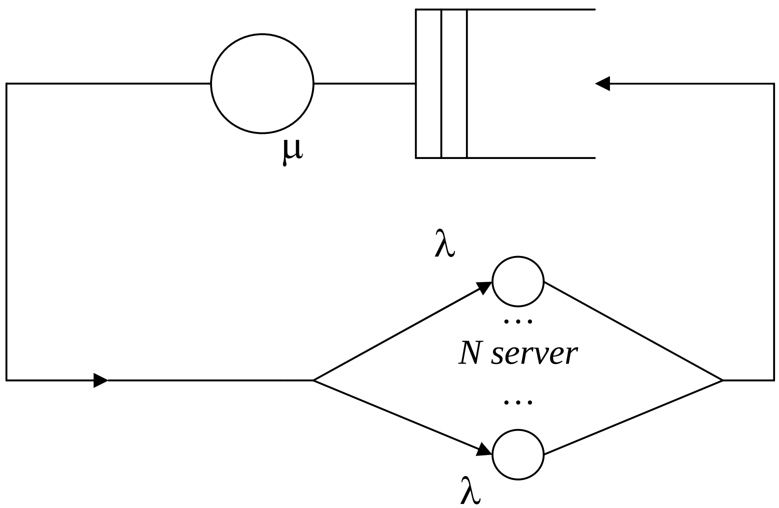
formula B di Erlang

frazione del tempo in cui, a regime, il sistema è pienamente occupato

In un sistema di telefonia, con frequenza media di chiamate λ, dimensionare il numero di linee in modo che la probabilità di perdita di una chiamata sia inferiore a una soglia prefissata

Sistemi M/M/1//N

singolo server, capacità coda ∞, *# clienti limitato* (N)



Il cliente viene servito (con tempo medio di servizio 1/μ) e successivamente ritorna nel sistema dopo un ritardo medio 1/λ

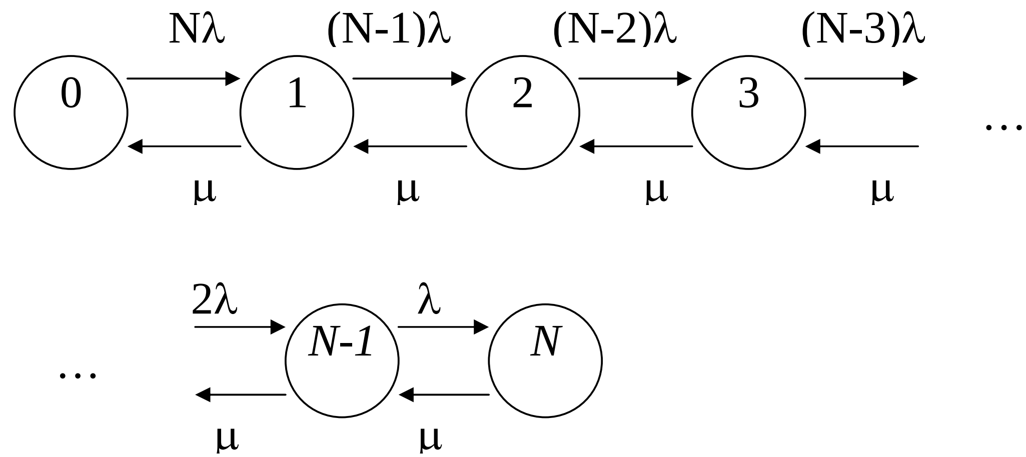
Esempio: sistema di elaborazione con N terminali, router di LAN per il servizio di N nodi locali

Non tutti i terminali devono *sempre* trasmettere/ricevere dati (ogni tanto l'operatore pensa/beve caffè/entrambi)

Stato per un sistema M/M/1//N

clienti nella coda (incluso il cliente "sotto processo"); Se vi sono *n* clienti in coda, N-*n* sono in attesa di tornare in coda → sovrapposizione di N-*n* processi di Poisson a tasso λ

Modello tramite catena nascita-morte



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

nessun problema di convergenza se $\rho = \lambda / \mu$

$$\pi_0 = \left(\sum_{n=0}^{N} \frac{N!}{(N-n)!} \rho^n \right)^{-1}, \ \pi_n = \frac{N!}{(N-n)!} \rho^n \pi_0$$

Utilizzazione: $1 - \pi_0$

Throughput: $\mu(1 - \pi_0)$

Tempo medio di risposta

R: tempo di risposta (tempo che il cliente spende nella coda e durante il servizio) - variabile aleatoria

E[R] = ???

Legge di Little, primo passo: $E[X] = \mu(1-\pi_0)E[R]$ (parte "superiore" dello schema)

Legge di Little, secondo passo: $E[N-X] = \mu(1-\pi_0)\dfrac{1}{\lambda}$

(parte "inferiore" dello schema: quando nella coda vi sono X utenti, sono in attesa di arrivo N-X utenti)

oss.: N: costante; X: variabile aleatoria

ora: $E[N-X] = E[N] - E[X] = N - E[X]$

$$\Rightarrow E[R] = \frac{N}{\mu(1-\pi_0)} - \frac{1}{\lambda}$$

Esempio:

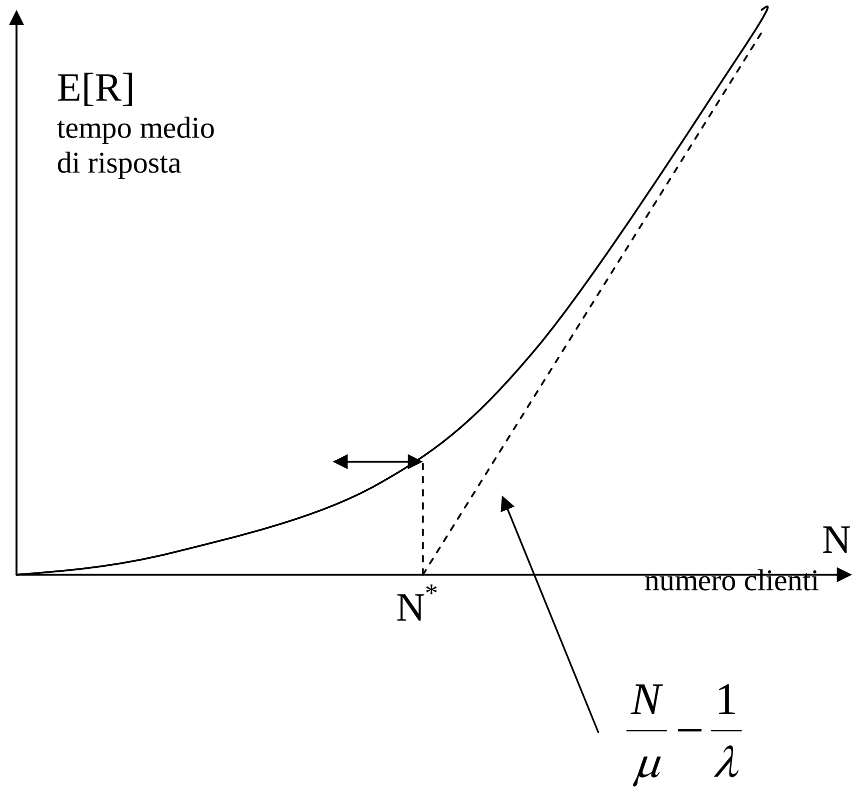
N: # terminali da acquistare;
μ: frequenza media di servizio dell' elaboratore;
1/λ : tempo medio di riflessione da parte di ogni utente.

Richiesta:
il tempo medio di risposta non deve eccedere β

Quanti terminali devo acquistare?

$$\frac{N}{\mu(1-\pi_0)} - \frac{1}{\lambda} < \beta$$

da cui N



per $N \to \infty$, $\pi_0 \to 0$

buon progetto $\Leftrightarrow$ *determinazione di* N^*