Reti di Telecomunicazioni 1

Corso "on-line" - AA2006/07
***"Sistemi a coda" Blocco E3** v3*

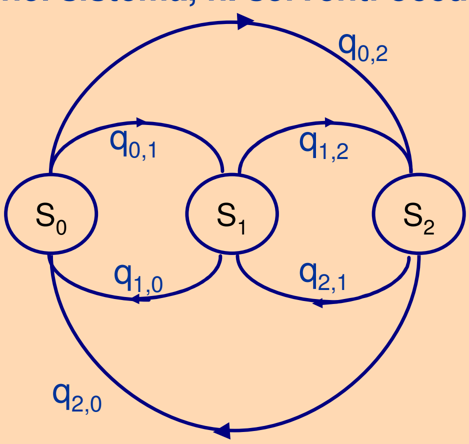
Ing. Stefano Salsano
e-mail: stefano.salsano@uniroma2.it

- **Risoluzione di un sistema a coda – le catene di Markov**

Le catene di Markov sono uno strumento matematico che ci consente di "risolvere" un sistema a coda, cioè determinare alcuni parametri prestazionali (es. ritardo di attraversamento, numero medio di utenti, probabilità di perdita), date le caratteristiche dei processi di ingresso e di servizio e le caratteristiche "strutturali" del sistema (es. numero di serventi, dimensione della fila di attesa…).

Sistemi a Coda e catene di Markov

- I sistemi a coda sono modellabili con catene di Markov tempo continuo
- La catena è formata da stati (S_i) ed archi di transizione fra stati
- Lo stato rappresenta una "situazione" in cui si trova il sistema. Lo stato è associato al valore di un insieme di parametri di sistema; per es. [n. richieste di servizio nel sistema, n. serventi occupati]



Supponiamo di osservare una mosca in un barattolo chiuso, sul fondo del barattolo vi è del miele, sul tappo (all'interno) vi è della marmellata. La mosca staziona o sul fondo o sul tappo e di tanto in tanto "salta" dal tappo al barattolo e viceversa. Supponiamo che il "salto" avvenga in modo istantaneo.

Diciamo quindi che il sistema può trovarsi in due stati:

S0: "mosca sul fondo"

S1: "mosca sul tappo"

e che sono possibili le transizioni di stato:

"fondo->tappo"

"tappo->fondo"

Osservando la mosca per un certo tempo, potremo valutare la frazione del tempo in cui è stata sul fondo e la frazione del tempo in cui è stata sul tappo.

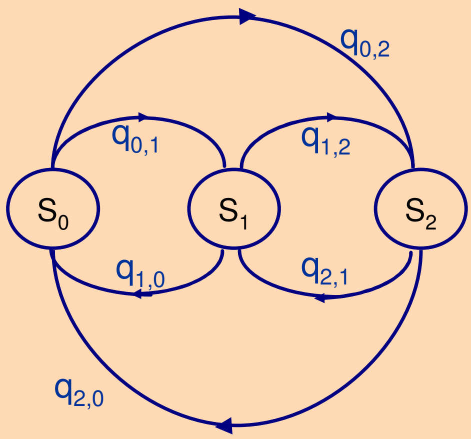
Più in astratto, definiamo la "probabilità di stato" come la probabilità di trovare la mosca sul fondo o sul tappo. Per fare questo dobbiamo supporre che il sistema sia "in equilibrio" cioè che queste probabilità di stato non variano. Ad esempio se il miele diminuisse, la mosca si comporterà diversamente e il sistema non può essere considerato in equilibrio.

P("fondo") = Probabilità che la mosca stia sul fondo

P("tappo") = Probabilità che la mosca stia sul tappo

Catene di Markov

- L'evoluzione del sistema si mappa su una sequenza di salti fra gli stati
- Assumendo che il sistema al tempo t_0 si trovi nello stato i-esimo, la probabilità che in dt sec. sia avvenuta la transizione S_i->S_k è definita come $q_{i,k} \cdot dt$, dove $q_{i,k}$ è la <u>frequenza di transizione</u> dallo i a quello k



Definiamo il concetto di frequenza di transizione di stato aiutandoci con l'esempio della mosca. Avremo due frequenze di transizione di stato:

$q_{fondo,tappo}$ corrispondente alla transizione "fondo->tappo"

$q_{tappo,fondo}$ corrispondente alla transizione "tappo->fondo"

Supponiamo che la mosca arrivi sul fondo del barattolo e misuriamo il tempo in cui rimane, prima di ripartire verso il tappo. Se il tempo medio di permanenza è 10 secondi, vuol dire che, <u>quando la mosca è sul fondo</u>, in media una volta ogni 10 secondi la mosca salta dal fondo verso il tappo. La frequenza di transizione di stato $q_{fondo,tappo}$ è 1/10 [s^{-1}]. Una mosca "iperattiva" potrebbe invece rimanere sul fondo in media 0,5 secondi (5 decimi di secondo). Vuol dire che, <u>quando la mosca è sul fondo</u>, in media due volte al secondo salta verso il tappo: la sua frequenza di transizione è 2 [s^{-1}].

Attenzione ad una cosa importante: se osserviamo il sistema e valutiamo la frequenza con cui avvengono effettivamente le transizioni dal fondo verso il tappo, questa NON corrisponde alla frequenza di transizione di stato !

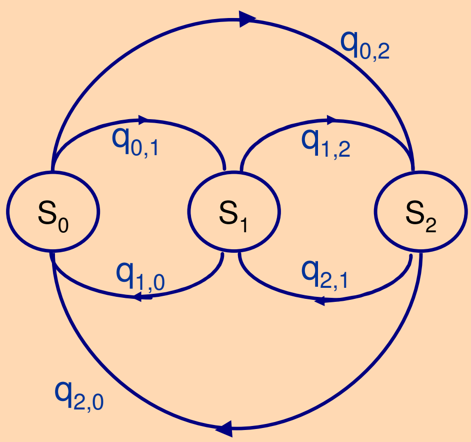
Ad esempio consideriamo la mosca che aveva frequenza di transizione "fondo-tappo" di 1/10 [s^{-1}]. Osserviamo il sistema per T=1000 secondi, potremmo aspettarci di osservare N= $q_{fondo,tappo}$ * T = 1/10 * 1000 = 100 transizioni.

In realtà commettiamo un grave errore, perché la frequenza di transizione di stato $q_{fondo,tappo}$ si riferisce agli intervalli in cui la mosca si trova sul fondo. Quindi ad esempio se la mosca si trova sul fondo per il 40% del tempo, cioè P("fondo")=0,4 allora il numero di osservazioni in T=1000 secondi sarà:

N= $q_{fondo,tappo}$ * T*P("fondo") = 1/10 * 1000 * 0,4 = 40 transizioni

Catene di Markov

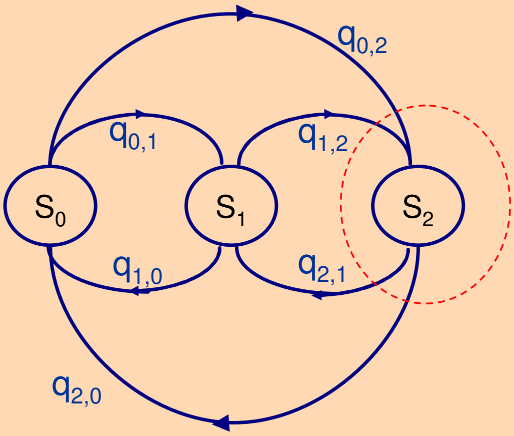
- **Note le frequenze di transizione è possibile derivare le probabilità di stato π_k = P{S_k}, "risolvendo" la catena di Markov. Il relativo procedimento viene analizzato nelle slides successive.**



Nei casi di nostro interesse le catene di Markov modellano i sistemi a coda. La conoscenza di queste probabilità di stato sarà il punto di partenza per derivare i parametri prestazionali di interesse quali il traffico smaltito, la prob. di rifiuto, il tempo di coda ecc.

Catene di Markov

- Si definisce "flusso proveniente dallo stato i-esimo diretto verso lo stato k-esimo" come il prodotto $P\{S_k\} \cdot q_{k,i}$
- Per il calcolo delle probabilità di stato si utilizza la legge di conservazione dei flussi : all'equilibrio statistico l'ammontare del flusso entrante in uno stato equivale a quello uscente



Esempio relativo allo stato S_2:

Flusso entrante = $P\{S_0\} \cdot q_{0,2} + P\{S_1\} \cdot q_{1,2}$

Flusso uscente = $P\{S_2\} \cdot q_{2,0} + P\{S_2\} \cdot q_{2,1}$

Senza alcuna pretesa di dare una dimostrazione formale, facciamo delle considerazioni intuitive sulla legge di conservazione del flusso: avere equilibrio statistico vuol dire che la probabilità di trovare il sistema nei vari stati non varia nel tempo. Quindi il valore atteso dei salti in ingresso ad uno stato S_i deve coincidere con il valore atteso dei salti in uscita da S_i . Se così non fosse la probabilità di trovare il sistema nello stato S_i aumenterebbe o diminuirebbe.

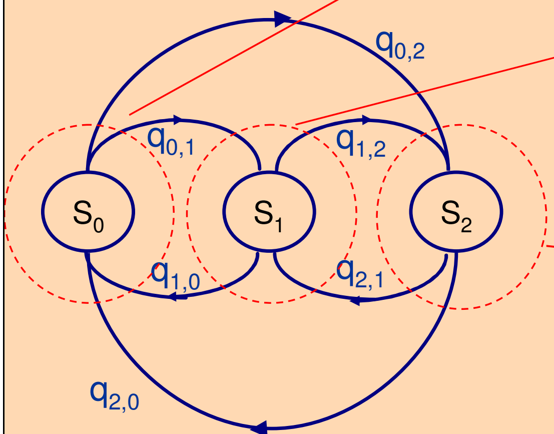
Consideriamo l'esempio della mosca: osservando il sistema per un certo tempo T il numero di salti che ci si aspetta che la mosca faccia da fondo a tappo deve coincidere con quelli da tappo a fondo. Quindi:

$$q_{fondo,tappo} * T*P(\text{“fondo”}) = q_{tappo,fondo} * T*P(\text{“tappo”})$$

ossia: $$q_{fondo,tappo} * P(\text{“fondo”}) = q_{tappo,fondo} * P(\text{“tappo”})$$

Catene di Markov

- Applicando il principio di conservazione dei flussi ad ogni stato, otteniamo tante equazioni quanti sono gli stati.



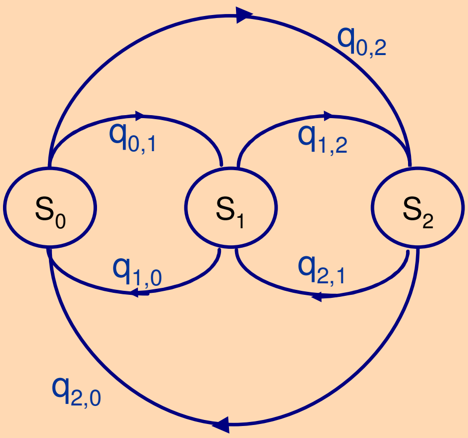
$$P\{S_1\}\cdot q_{1,0} + P\{S_2\}\cdot q_{2,0} = P\{S_0\}\cdot q_{0,1} + P\{S_0\}\cdot q_{0,2}$$

$$P\{S_0\}\cdot q_{0,1} + P\{S_2\}\cdot q_{2,1} = P\{S_1\}\cdot q_{1,0} + P\{S_1\}\cdot q_{1,2}$$

$$P\{S_1\}\cdot q_{1,2} + P\{S_0\}\cdot q_{0,2} = P\{S_2\}\cdot q_{2,0} + P\{S_2\}\cdot q_{2,1}$$

Catene di Markov

- **Se il sistema ha N stati si ha un sistema lineare di N equazioni ed N incognite che però ha determinante nullo.**
- **A questo punto si elimina una equazione e si aggiunge la legge di conservazione (la somma delle prob. è uguale a 1). Quindi si risolve il sistema per le incognite $P\{S_k\}$.**

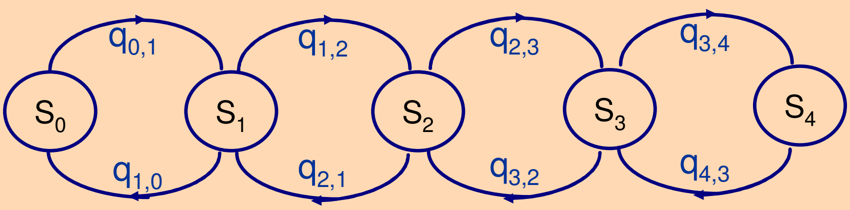


$$P\{S_1\}\cdot q_{1,0} + P\{S_2\}\cdot q_{2,0} = P\{S_0\}\cdot q_{0,1} + P\{S_0\}\cdot q_{0,2}$$
$$P\{S_0\}\cdot q_{0,1} + P\{S_2\}\cdot q_{2,1} = P\{S_1\}\cdot q_{1,0} + P\{S_1\}\cdot q_{1,2}$$
$$\cancel{P\{S_1\}\cdot q_{1,2} + P\{S_0\}\cdot q_{0,2} = P\{S_2\}\cdot q_{2,0} + P\{S_2\}\cdot q_{2,1}}$$

$$P\{S_0\} + P\{S_1\} + P\{S_2\} = 1$$

Catene di Markov di nascita e morte

- **Un tipo particolare di catena di Markov in cui sono ammesse solo transizioni fra stati adiacenti si dice catena di “nascita e morte”.**
- **Ad esempio, qui sotto è rappresentata una catena di nascita e morte a 5 stati**



Sistemi a coda Markoviani – La coda M/M/1

Quanto visto finora riguarda le catene di Markov, che come abbiamo detto sono uno strumento matematico che può essere utilizzato per descrivere diversi tipi di sistemi in vari campi di applicazione.

Ritorniamo ora ai nostri "sistemi di servizio" o "sistemi a coda" e vediamo che relazione c'è tra questi e le catene di Markov.

Sotto certe condizioni relative ai processi di arrivo e di servizio il sistema a coda può essere modellato da una catena di Markov e si può parlare di un "sistema a coda Markoviano"

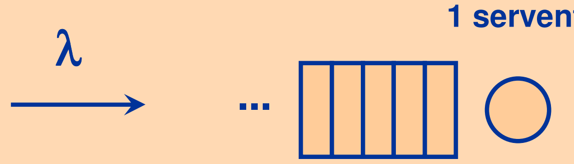
Si noti quindi che se tali condizioni non sono verificate il sistema a coda non può essere risolto da una catena di Markov.

Sistemi a coda Markoviani

- I sistemi a coda Markoviani hanno le seguenti caratteristiche:
  - Processo d'ingresso esponenziale negativo con frequenza media di interarrivo pari a λ
  - Processo di servizio esponenziale negativo con frequenza media di terminazione di servizio μ.
    - La frequenza media di terminazione di servizio è uguale all'inverso del tempo medio di servizio
    - Popolazione d'utenti infinita
- <u>SOTTO QUESTE IPOTESI,</u> si può dimostrare che:
  - Associando lo stato al <u>numero di richieste di servizio presenti nel sistema</u>, si può costruire una catena di Markov di nascita e morte, le cui frequenze di transizione sono:
    - $q_{i,i+1} = \lambda$
    - $q_{i,i-1} = \mu$

Coda M/M/1

1 servente



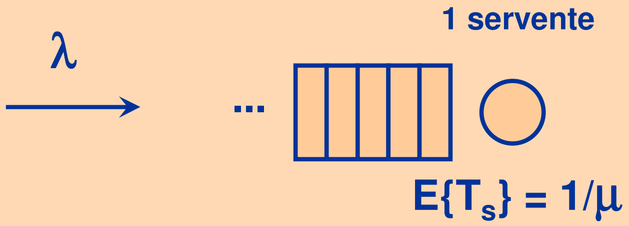
$E\{T_s\} = 1/\mu$

- Consideriamo una coda con 1 servente e capacità della fila di attesa infinita
- L'interarrivo degli utenti abbia distribuzione esponenziale negativa con ritmo medio λ (s^{-1}).
- Il tempo di servizio abbia distribuzione esponenziale negativa con ritmo medio μ (s^{-1}),
  ossia tempo medio di servizio $E\{T_s\}=1/\mu$ (s)
- Notazione di Kendall: M/M/1/∞/∞ o semplicemente M/M/1

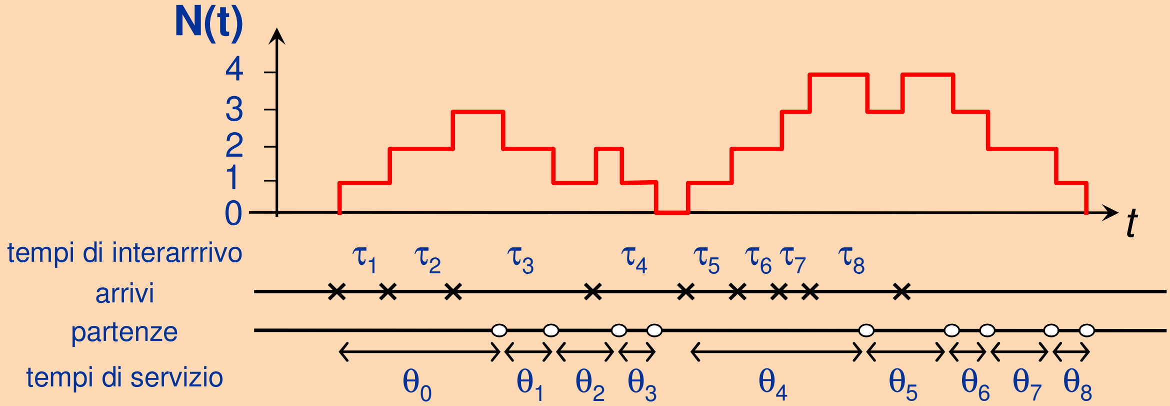
Il primo tipo di sistema a coda che andiamo a considerare è riportato nella slide precedente ed è indicato nella notazione di Kendall con M/M/1.

Questo tipo di coda viene utilizzato per modellare il servizio offerto da un multiplatore a pacchetto. Le unità informative (pacchetti) che arrivano al multiplatore sono gli utenti del sistema a coda, la linea di uscita sulla quale vengono trasmessi i pacchetti uno alla volta è il singolo servente. La memoria del multiplatore in cui vengono accodati i pacchetti in attesa di essere trasferiti corrisponde alla fila di attesa. Nella modellazione con il sistema M/M/1 si suppone che la dimensione della memoria sia infinita, trascurando così gli eventi di perdita dei pacchetti.

Coda M/M/1: evoluzione



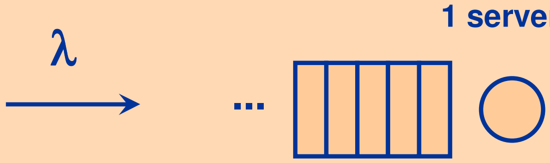
- Lo stato è rappresentato dal numero di utenti nel sistema N(t)



Nella slide precedente si mostra l'evoluzione di un sistema M/M/1. Il tempo è rappresentato sull'asse delle ascisse, in ordinata è rappresentato lo stato del sistema ossia il numero di utenti nel sistema: 0 = sistema vuoto, 1 = un utente in servizio, >=2 un utente in servizio e gli altri nella fila di attesa.

Sotto all'asse delle ascisse, in corrispondenza con le transazioni di stato sono riportati gli eventi che le hanno generate (arrivo di un utente nel sistema, terminazione di un servizio cioè partenza dell'utente dal sistema). Si riportano anche i tempi di interarrivo τ_i e i tempi di servizio θ_i.

Coda M/M/1: traffico offerto



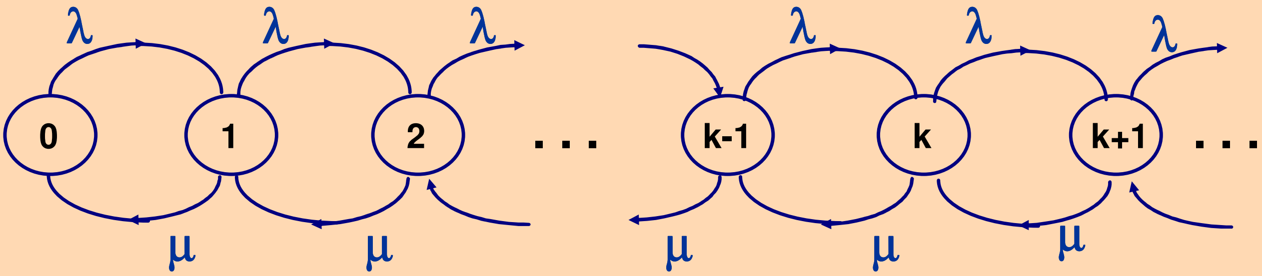
$$E\{T_s\} = 1/\mu$$

- Applicando la definizione di traffico offerto A_o alla coda M/M/1 abbiamo:

$$A_o = \lambda \cdot E\{T_s\} = \lambda/\mu$$

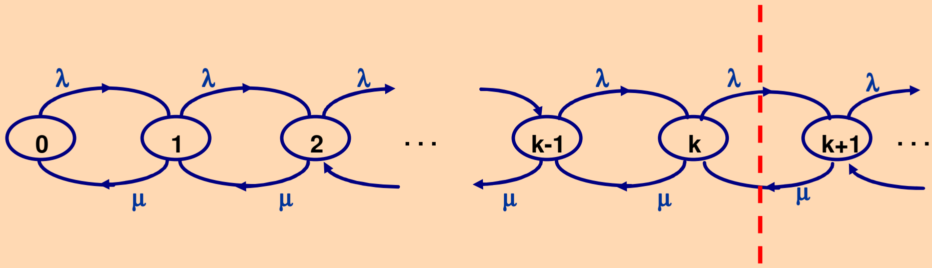
Coda M/M/1: spazio degli stati

- Supponiamo che il sistema raggiunga uno stato stazionario e cerchiamo la distribuzione del numero di utenti all'interno del sistema cioè la probabilità di stato P{N(t)=k}
- Rappresentiamo il sistema graficamente e consideriamo le transizioni tra uno stato e l'altro



- Ad ogni stato k è associata una probabilità π_k = P{N(t)=k}

Coda M/M/1: distribuzione di π_k



- Imponiamo una condizione di equilibrio su ogni "taglio":

$$\mu\pi_{k+1} = \lambda\pi_k \qquad k \geq 0$$

$$\pi_{k+1} = \frac{\lambda}{\mu}\pi_k$$

Nella slide precedente si mostra come ricavare le equazioni per risolvere la catena di Markov applicando il principio di conservazione del flusso. Si noti che viene usata la condizione di equilibrio dei flussi in modo diverso da quanto visto in precedenza, dove veniva applicata considerando il flussi in entrata e in uscita rispetto ad una linea chiusa che "circondava" ciascuno stato. In realtà il principio di conservazione del flusso si applica anche a qualunque linea chiusa tracciata sullo spazio degli stati, e in generale a qualunque "taglio" che partizioni in due parti distinte l'insieme degli stati.

Coda M/M/1 : distribuzione di π_k

$$\text{Ponendo} \quad \rho = \frac{\lambda}{\mu} \quad \text{si} \quad \text{ha :}$$

$$\pi_{k+1} = \rho \pi_k, \quad \pi_k = \rho^k \pi_0$$

queste corrispondono a k equazioni

Si impone quindi che la somma delle probabilità π_k sia 1:

$$1 = \sum_{i=0}^{\infty} \pi_i$$

Coda M/M/1 : distribuzione di π_k

Si "risolve" quindi il sistema sostituendo le k equazioni in quella relativa alla somma delle probabilità di stato.

$$1 = \sum_{i=0}^{\infty} \pi_i = \sum_{i=0}^{\infty} \pi_0 \rho^i = \pi_0 \sum_{i=0}^{\infty} \rho^i = \pi_0 \frac{1}{1-\rho}$$

$$\pi_0 = 1 - \rho$$

$$\pi_k = \rho^k (1-\rho)$$

$$\rho < 1$$

I passaggi mostrati nella slide precedente sono tutti passaggi algebrici, si noti però che si è utilizzata la formula per la sommatoria di ρ^i (per i che va da 0 a infinito) che vale solo se $\rho < 1$, altrimenti la serie diverge.

Coda M/M/1: stabilità

- Abbiamo quindi la distribuzione di probabilità del numero di utenti nel sistema:

$$\pi_k = P\{N(t) = k\} = \rho^k (1-\rho)$$

$$se \quad \rho < 1$$

- La condizione su <u>ρ<1</u> garantisce la convergenza della sommatoria delle probabilità di stato e quindi la <u>stabilità del sistema M/M/1</u>
La condizione di stabilità può anche essere espressa come:

$$\rho < 1 \Leftrightarrow \lambda < \mu$$

Abbiamo quindi "risolto" la catena di Marcov e ricavato le probabilità di stato sotto la condizione "di stabilità" $\rho < 1$. Questa condizione corrisponde a $\lambda < \mu$ cioè la frequenza media di arrivo degli utenti deve essere minore della frequenza di terminazione di servizio.

Consideriamo ad esempio un sistema di servizio costituito da un casello autostradale con una singola postazione. Se è previsto il transito di 300 automobili in un'ora e per effettuare il pagamento servono in media 30 secondi, possiamo immediatamente vedere che il sistema non è stabile. La frequenza media di terminazione di servizio infatti sarà di 3600/30 = 120 [h^{-1}] cioè la "capacità" del casello è di 120 automobili l'ora. Qundi $\lambda = 300$ [h^{-1}], $\mu = 120$ [h^{-1}] e la condizione $\lambda < \mu$ non è rispettata.

In un sistema non stabile come questo appena visto la occupazione della fila di attesa tenderà a crescere indefinitamente. Questo significa che le probabilità di stato non si potranno "stabilizzare": osservando il sistema dopo un certo tempo aumenterà la probabilità di trovare il sistema negli stati con più utenti nella fila di attesa e diminuirà la probabilità di trovare il sistema negli stati con meno utenti nella fila di attesa.

Dal punto di vista matematico questo corrisponde a non poter trovare delle probabilità di stato "finite" che mantengono in equilibrio il sistema (la sommatoria diverge).

Osserviamo infine che in un sistema non stabile non si conserva il flusso degli utenti: nell'esempio visto prima arrivano al casello 300 macchine l'ora, mentre ne escono 120. In queste condizioni NON si applica la legge di Little, che ha bisogno della condizione di equilibrio statistico.

Coda M/M/1: carico=utilizzazione

- Il carico del sistema è definito come il rapporto fra traffico offerto (A_o) e numero di serventi
- L'utilizzazione del sistema è definita come il rapporto fra traffico smaltito (A_s) e numero di serventi
- In un sistema senza perdita stabile, come l'M/M/1 :
  - » $A_o = A_s$
  - » carico=utilizzazione

Avevamo già definito nella lezione precedente il carico e la utilizzazione di un sistema, vediamo ora che in un sistema stabile senza perdite il traffico smaltito corrisponde a quello offerto. Il carico quindi coinciderà con l'utilizzazione.

Coda M/M/1: traffico smaltito

- Calcolo del traffico smaltito con le probabilità di stato :
  - » A_s = numero medio di serventi occupati (S)

$$\begin{aligned} A_s = E(S) \quad &= 0 \cdot P\{N(t)=0\} + 1 \cdot P\{N(t) \neq 0\} \\ &= P\{N(t) \neq 0\} = \\ &= (1-\pi_0) = (1-(1-\rho)) = \\ &= \rho = \lambda/\mu \end{aligned}$$

- Si ritrova quindi per A_s la stessa espressione per A_o che avevamo dato alla slide 23!

Il traffico smaltito è per definizione il numero medio di utenti che stanno ricevendo servizio. Si osservi che la notazione che avevamo usato nella lezione scorsa è: $A_s=E\{X_s(t)\}$, nella slide precedente usiamo E(S) con lo stesso identico significato.

Nel caso della M/M/1, $X_s(t)$ può valere 0 o 1, a seconda che l'unico servente nell'istante t sia libero o occupato. Il servente sarà libero se non ci sono utenti nel sistema (stato S_0) , mentre sarà occupato se c'è almeno un utente nel sistema (in tutti gli altri stati: S_1, S_2, S_3 …). Come si vede dai passaggi riportati nella slide precedente, quindi E(S) cioè $E\{X_s(t)\}$ corrisponde alla probabilità che il sistema stia in uno stato diverso da S_0.

Coda M/M/1: traffico smaltito

- Si può valutare la stessa quantità, utilizzando la legge di Little:

  consideriamo solo il sottosistema formato dal servente
  per la stabilità, $\lambda_o = \lambda_s = \lambda$
  Quindi, numero medio di serventi occupati è:

  $$E(S)= \lambda_s \cdot E(T_s) = \lambda \cdot 1/\mu = \rho$$

- Dato che il servente è unico, E(S) deve essere minore di 1, confermando la condizione di stabilità.

Nella slide precedente mostriamo come si poteva valutare il numero medio di serventi occupati cioè il traffico smaltito nella M/M/1 anche senza utilizzare le probabilità di stato. Si noti come la legge di Little sia stata applicata non a tutto il sistema, ma solo alla parte rappresentata dal servente.

Mostriamo qui sotto infatti come si può applicare la legge di Little a tutto il sistema, al servente, alla sola fila di attesa.

$E(N) = \lambda \cdot E(T)$	**tutto il sistema**
$E(S) = \lambda \cdot E(T_s)$	**solo il servente**
$E(Q) = \lambda \cdot E(T_{attesa})$	**solo la fila di attesa**

Dove E(T) è il tempo medio di attraversamento dell'intero sistema coda (cioè il tempo di coda), E(Q) il numero medio di utenti nella fila di attesa, $E(T_{attesa})$ il tempo medio speso da un utente nella fila di attesa.

Passiamo ora alla valutazione del numero medio di utenti nella sistema, che sfrutta le probabilità di stato ottenute dalla risoluzione della catena di Markov.

Nella derivazione si sfruttano calcoli algebrici e la formula per la sommatoria di $i \cdot \rho^i$ (per i che va da 0 a infinito).

Coda M/M/1: numero medio di utenti

- Valutiamo il numero medio di utenti nel sistema:

$$
\begin{aligned}
E(N) &= \sum_{k=0}^{\infty} k\pi_k = \sum_{k=0}^{\infty} k\rho^k (1-\rho) = \\
&= (1-\rho)\sum_{k=0}^{\infty} k\rho^k = (1-\rho)\frac{\rho}{(1-\rho)^2} = \\
&= \frac{\rho}{1-\rho}
\end{aligned}
$$

Il numero medio di utenti nella coda dipende da ρ cioè dall'utilizzazione del sistema.

Coda M/M/1: tempo di coda

- Valutiamo il tempo medio di coda (attesa + servizio) utilizzando la legge di Little applicata all'intero sistema:

$$E(N) = \lambda E(T)$$

$$E(T) = \frac{E(N)}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu}\frac{1}{(1-\rho)} = E(T_s)\frac{1}{(1-\rho)}$$

$$= \frac{1}{\mu - \lambda}$$

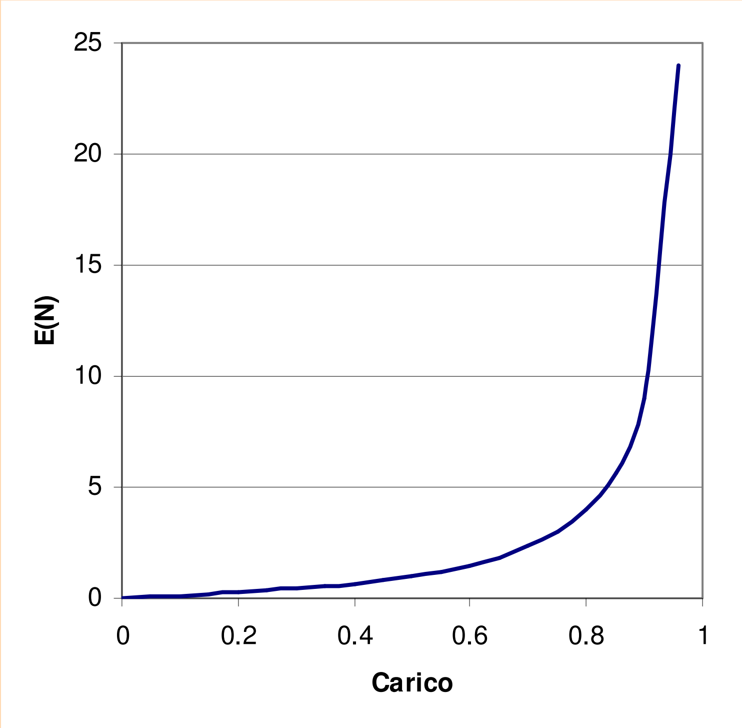
Abbiamo ricavato due espressioni equivalenti per il tempo medio di coda della M/M/1.

Nella prima espressione si vede come il tempo medio di coda è pari al prodotto del tempo di servizio per un fattore che dipende dall'utilizzazione ρ.

Nella seconda si vede come il tempo medio di coda è inversamente proporzionale alla differenza tra la frequenza di terminazione di servizio e la frequenza di arrivo degli utenti al sistema,

E(N) in funzione del carico

$$E(N)=\frac{\rho}{1-\rho}$$

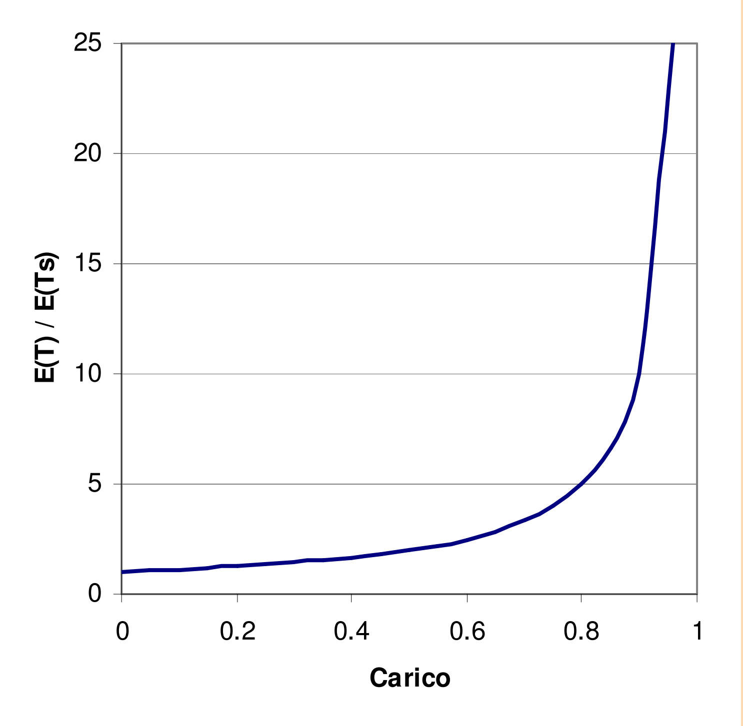


La slide precedente riporta il numero medio di utenti nel sistema in funzione del carico ρ (o dell'utilizzazione, visto che coincidono).

Se il carico è vicino allo 0, il sistema sarà vuoto per la maggior parte del tempo ed E(N) sarà anch'esso vicino a 0. Al crescere del carico, il numero medio di utenti cresce inizialmente in modo "contenuto", fino a che il carico arriva a valori intorno a 0,6. La curva prende quindi a salire in modo drastico, fino a divergere quando il carico tende a 1.

E(T) in funzione del carico

$$E(T) = E(T_s)\frac{1}{1-\rho}$$



La slide precedente riporta invece il tempo medio di coda (attesa + servizio) in funzione del carico ρ (o dell'utilizzazione, visto che coincidono).

In particolare viene graficato il rapporto tra il tempo medio di coda e il tempo medio di servizio. Se il carico è vicino allo 0, tale rapporto vale 1 e il tempo di coda è uguale al tempo di servizio. Infatti gli utenti non troveranno altri utenti nel sistema, ma riceveranno immediatamente servizio. Analogamente a quanto visto per il numero medio di utenti, al crescere del carico il tempo medio di coda cresce in modo "contenuto", fino a che il carico arriva a valori intorno a 0,6. La curva prende quindi a salire in modo drastico, fino a divergere quando il carico tende a 1.

Le tre slide seguenti NON fanno parte del programma di esame, sono facoltative.

Si mostra come per una coda M/M/1 esiste una espressione che si ricava analiticamente della distribuzione del tempo di coda. Abbiamo quindi una informazione molto più completa delle prestazioni che ricevono gli utenti rispetto alla conoscenza del semplice valore medio.

Distribuzione del ritardo in una M/M/1

- È possibile valutare analiticamente la distribuzione (densità di probabilità) dei tempi di attraversamento di una coda M/M/1
- Omettiamo tale calcolo, che utilizza la trasformata di Laplace della densità di probabilità del tempo di servizio, e riportiamo il risultato:

$$f_T(t) = (1-\rho)\mu \cdot e^{-(1-\rho)\mu \cdot t}$$

- Il tempo di coda ha una distribuzione esponenziale negativa con parametro $(1-\rho)\mu$. Come già visto il tempo medio di coda è $1 / (\mu(1-\rho))$.

Distribuzione del ritardo in una M/M/1

- La distribuzione di probabilità del tempo di coda T è quindi:

$$F_T(t) = 1 - e^{-(1-\rho)\mu \cdot t}$$

- Definiamo percentile r del ritardo il ritardo non superato da una frazione dei pacchetti maggiore di 1-r%, ossia il valore d_r tale che:

$$P\{T < d_r\} = r/100 \Rightarrow P\{T > d_r\} = 1 - r/100$$

Distribuzione del ritardo in una M/M/1

- Dalla distribuzione otteniamo la seguente relazione:

$$F_{\Delta_t}(t) = P\{\Delta < t\} = 1 - e^{-(1-\rho)\mu \cdot t}$$

$$F_{\Delta_t}(t) < r/100 \Rightarrow 1 - e^{-(1-\rho)\mu \cdot t} < r/100$$

Da cui ricaviamo d_r:

$$1 - e^{-(1-\rho)\mu \cdot d_r} = r/100$$

$$e^{-(1-\rho)\mu \cdot d_r} = 1 - r/100$$

$$(1-\rho)\mu \cdot d_r = -\log\left(1 - r/100\right)$$

$$d_r = \frac{1}{\mu} \frac{-\log\left(1 - r/100\right)}{(1-\rho)}$$

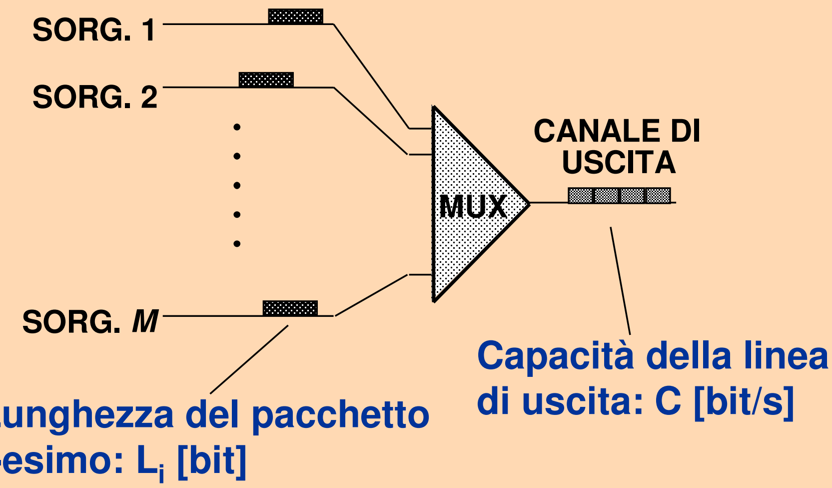
(Da qui riprende il programma di esame…)

Discutiamo infine come la coda M/M/1 possa essere utilizzata per modellare un multiplatore a pacchetto.

Nella slide successiva si vede come la trasmissione dei pacchetti sulla linea di uscita viene modellata dal processo di servizio. Se la linea di uscita ha capacità C [bit/s], il tempo di servizio dipende dalla lunghezza del pacchetto da trasmettere in modo direttamente proporzionale. Allo stesso modo il tempo medio di servizio sarà direttamente proporzionale alla lunghezza media dei pacchetti da trasmettere. Più in generale la distribuzione del tempo di servizio corrisponderà alla distribuzione della lunghezza dei pacchetti.

Modello del multiplatore a pacchetto

Multiplatore statistico



- Il tempo di servizio del pacchetto i-esimo $T_s(i)$ sarà:
  $T_s(i)= L_i/C$ [s]
- Il tempo di medio di servizio $E\{T_s\}$ sarà:
  $E\{T_s\} = E\{L\}/C$ [s]

Modello del multiplatore a pacchetto

- Il tempo di servizio del pacchetto i-esimo $T_s(i)$ sarà:

  $T_s(i) = L_i/C$ [s]
- Il tempo di medio di servizio $E\{T_s\}$ sarà:

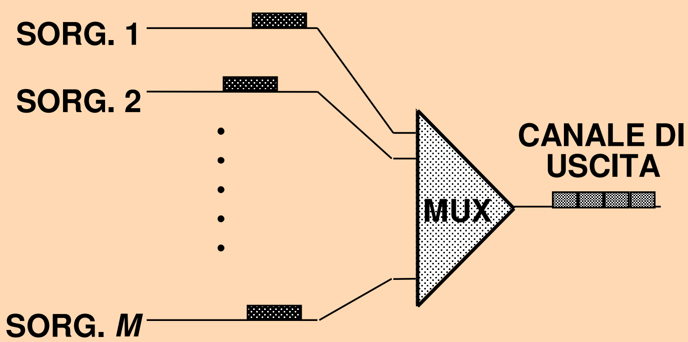
  $E\{T_s\} = E\{L\}/C$ [s]
- Dove E(L) è la lunghezza media dei pacchetti. Infatti se n è il numero di pacchetti, si ha che

$$E\{T_s\} = \frac{\sum_{i=1}^{n} T_s(i)}{n} = \frac{\sum_{i=1}^{n} L_i / C}{n} = \frac{1}{C}\frac{\sum_{i=1}^{n} L_i}{n} = \frac{1}{C}E\{L\}$$

Ipotesi per applicare M/M/1 al multiplatore

Multiplatore

statistico



- I flussi di pacchetti prodotti dalle sorgenti sono rappresentabili mediante processi di Poisson;
- I flussi di pacchetti emessi dalle sorgenti sono indipendenti tra loro;
- Le lunghezze dei pacchetti hanno distribuzione esponenziale negativa e sono indipendenti tra loro;
- Il processo di ingresso complessivo è indipendente dal processo di servizio

→ Modello M/M/1 adeguato nel caso di *M*>>1 (ipotesi di Kleinrock).

La slide precedente elenca un insieme di condizioni sotto cui la coda M/M/1 rappresenta un modello realistico del multiplatore a pacchetto.

In pratica si tratta di capire se e come:
1) il processo di arrivo dei pacchetti possa essere assimilato ad un processo con distribuzione esponenziale negativa
2) il processo di servizio possa essere assimilato ad un processo con distribuzione esponenziale negativa
3) il processo di arrivo e di servizio possano essere considerati statisticamente indipendenti

L'ipotesi di avere molte sorgenti di ingresso indipendenti (M>>1) aiuta in particolare a realizzare i punti 1) e 3). Infatti una caratteristica di un processo di arrivo esponenziale negativo è di poter avere diversi arrivi successivi con tempi di interarrivo molto piccoli. Se si hanno poche sorgenti c'è invece un limite a questo, perché per ciascuna sorgente dopo che un pacchetto è arrivato al multiplatore, prima che arrivi un altro pacchetto deve trascorrere un tempo pari alla lunghezza del nuovo pacchetto divisa per la capacità della linea di ingresso. Se si considera una singola sorgente c'è una correlazione tra processo di arrivo e di servizio, perché se le lunghezze dei pacchetti sono più lunghe (e quindi i tempi di servizio maggiori) anche i tempi di interarrivo sono maggiori.

Per quanto riguarda il punto 2) cioè i tempi di servizio esponenziali negativi, questo corrisponde a richiedere che la lunghezza dei pacchetti abbia una ampia variabilità, da molto corta a molto lunga.

Le lunghezze dei pacchetti tipiche delle varie tecnologie di rete sono spesso meno “variabili”, essendo caratterizzate da una lunghezza minima e massima. L’assunzione di tempi di servizio esponenziali negativi è quindi spesso lontana dalla realtà. Si tenga però presente che a parità di tempo medio di servizio ad una variabilità minore corrispondono “prestazioni” migliori (minor numero di utenti in coda, minor tempo di coda). L’assunzione di tempi di servizio esponenziali negativi può quindi essere utile in certi casi per avere una sovrastima di questi parametri (numero di utenti in coda, tempo di coda).