



Politecnico di Milano

*Facoltà di Ingegneria
dell'Informazione*

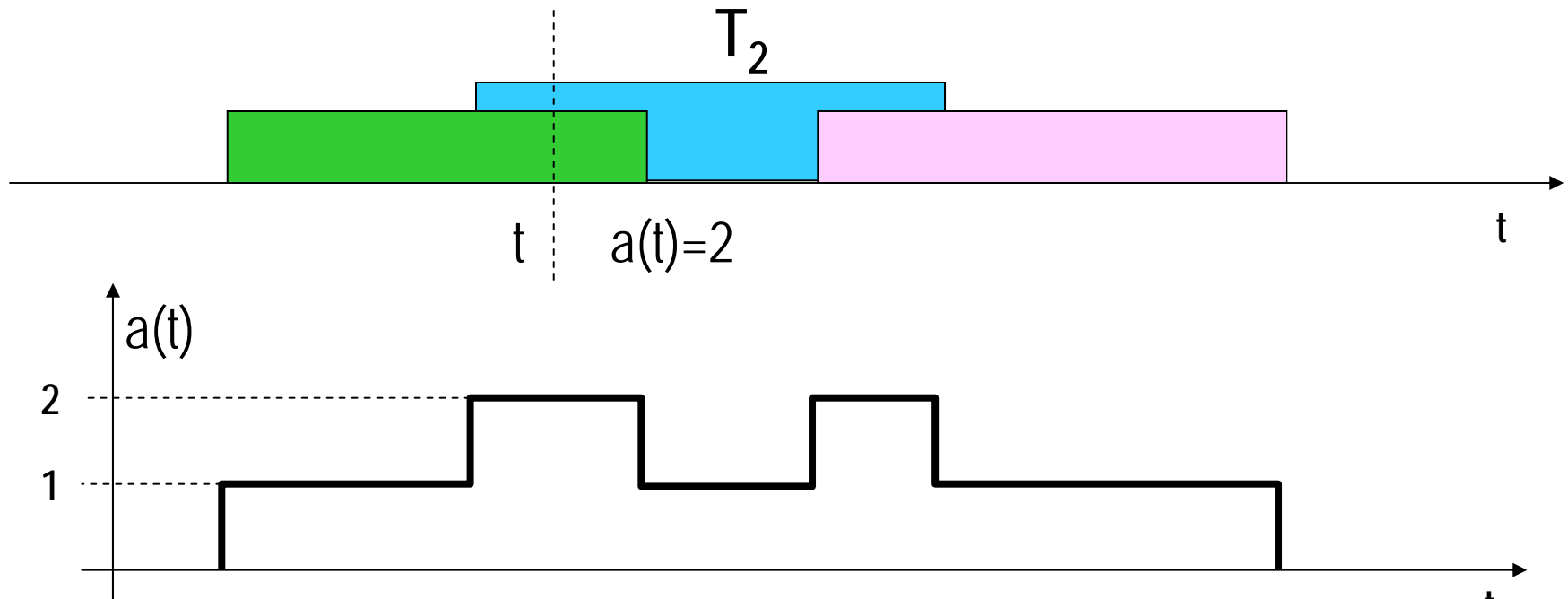
Infrastrutture e Protocolli per Internet

Prof. Antonio Capone

Teoria del Traffico

Teoria del traffico: il traffico istantaneo

- ◆ Il “traffico” istantaneo in t è il numero di chiamate (messaggi, pacchetti, ...) $a(t)$ in corso su un canale al tempo t

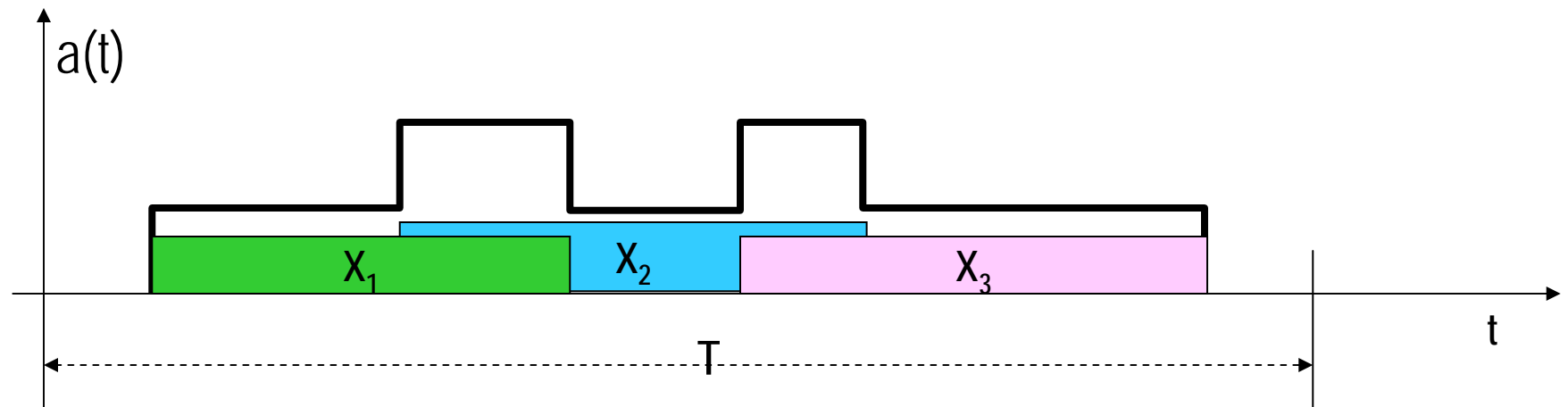


vedi corso di
“Processi stocastici e
sistemi a coda”

Teoria del traffico: Risultati sul traffico

Il traffico medio in T è

$$A(T) = 1/T \int_T a(t) dt$$

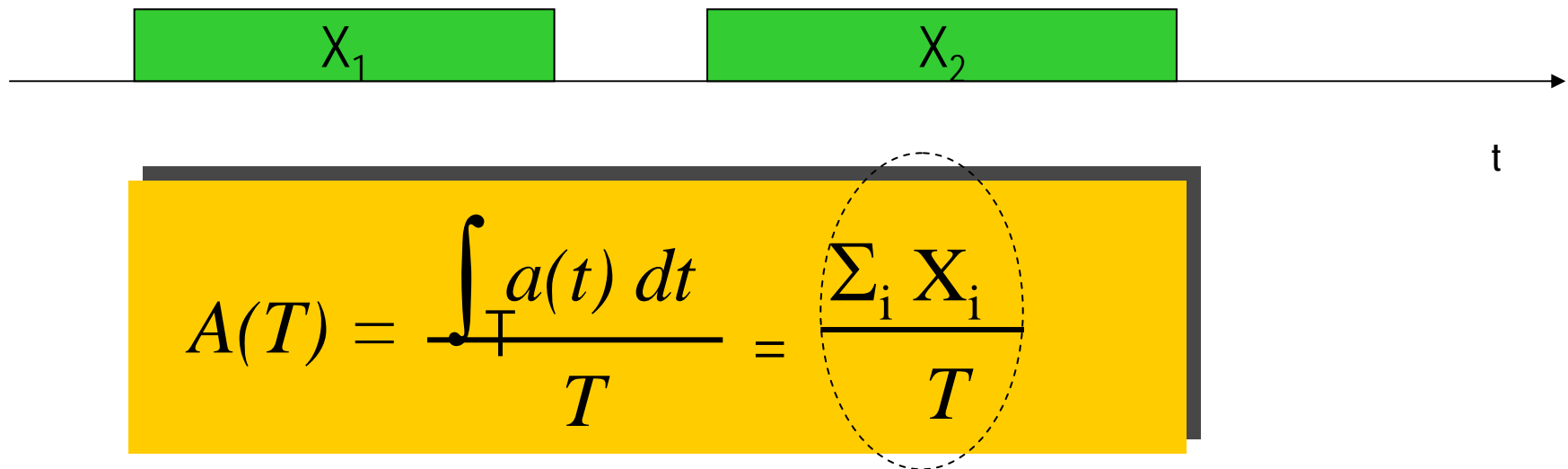


Risulta

$$\int_T a(t) dt = \sum_i X_i \quad \text{in } T$$

Teoria del traffico: Risultati sul traffico

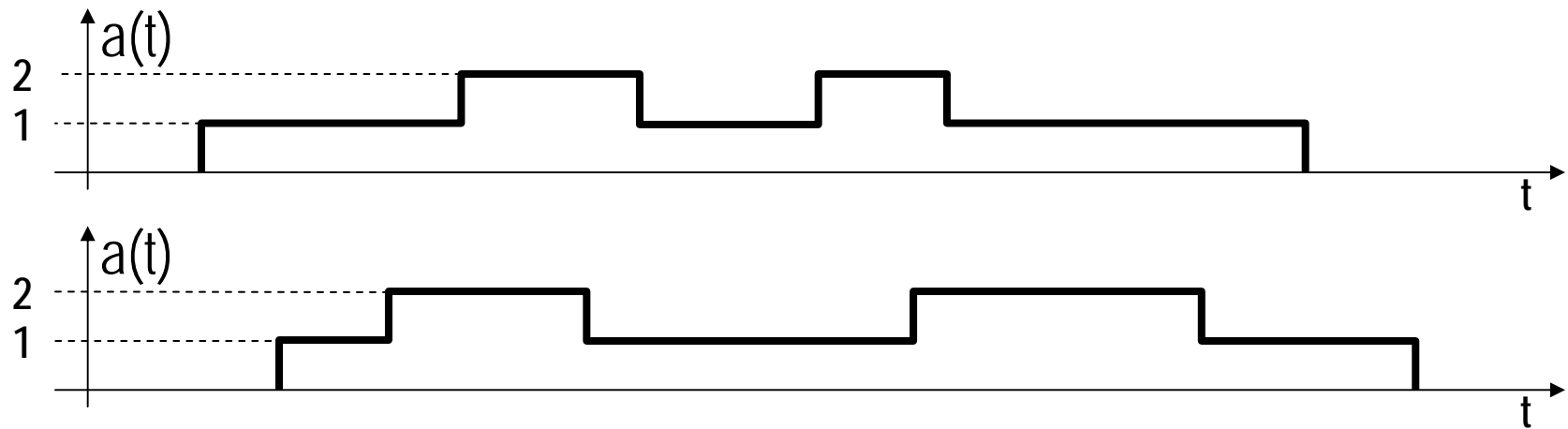
- ◆ Nel caso in cui le trasmissioni non possono sovrapporsi



- ◆ è la frazione di tempo in cui le trasmissioni sono attive

Teoria del traffico: Il Traffico

- ◆ In realtà il “traffico” istantaneo $a(t)$ è un processo casuale



- ◆ $A(T)$ è una variabile casuale
- ◆ di solito si considera la media d'insieme $E[A(T)]$

Teoria del traffico: Il Traffico

- ◆ In condizioni di stazionarietà le medie non dipendono da T

$$E[A(T)] = A$$

$$A = \lambda X$$

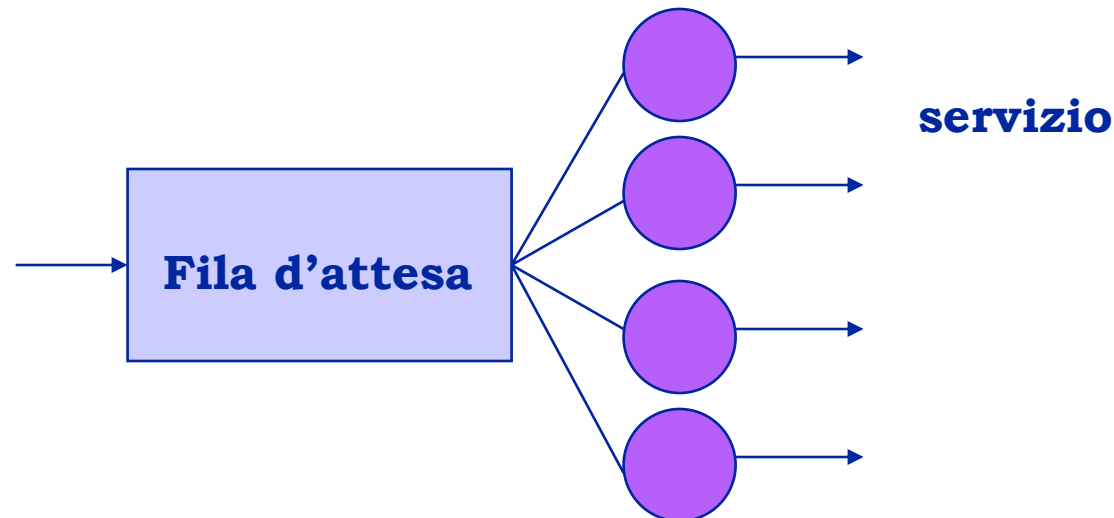
- ◆ A non ha dimensione
- ◆ Il traffico si misura in Erlang

Teoria del traffico: Efficienza

- ◆ Il traffico massimo smaltibile è un parametro importante
- ◆ Nel caso di singoli canali il massimo traffico consentito dai protocolli (da 0 a 1) riflette l'**efficienza** con cui i protocolli usano il canale

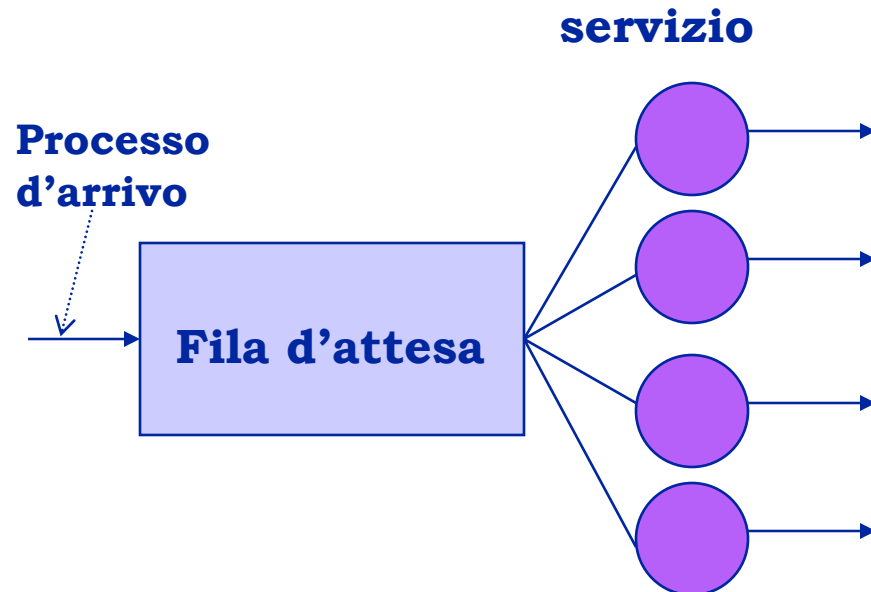
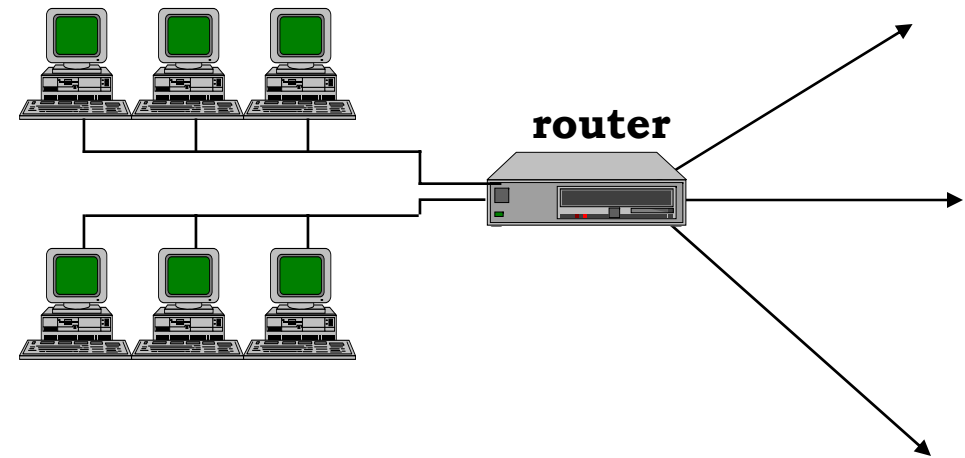
Teoria del traffico: Sistemi di servizio

- ◆ Ad un sistema di servizio arrivano richieste di servizio secondo un processo (puntuale) degli arrivi
- ◆ Ciascuna richiesta è caratterizzata da un tempo di servizio necessario ad uno dei serventi per soddisfarla
- ◆ è possibile la presenza di un sistema di attesa (o coda) dove le richieste attendono che un servente si liberi



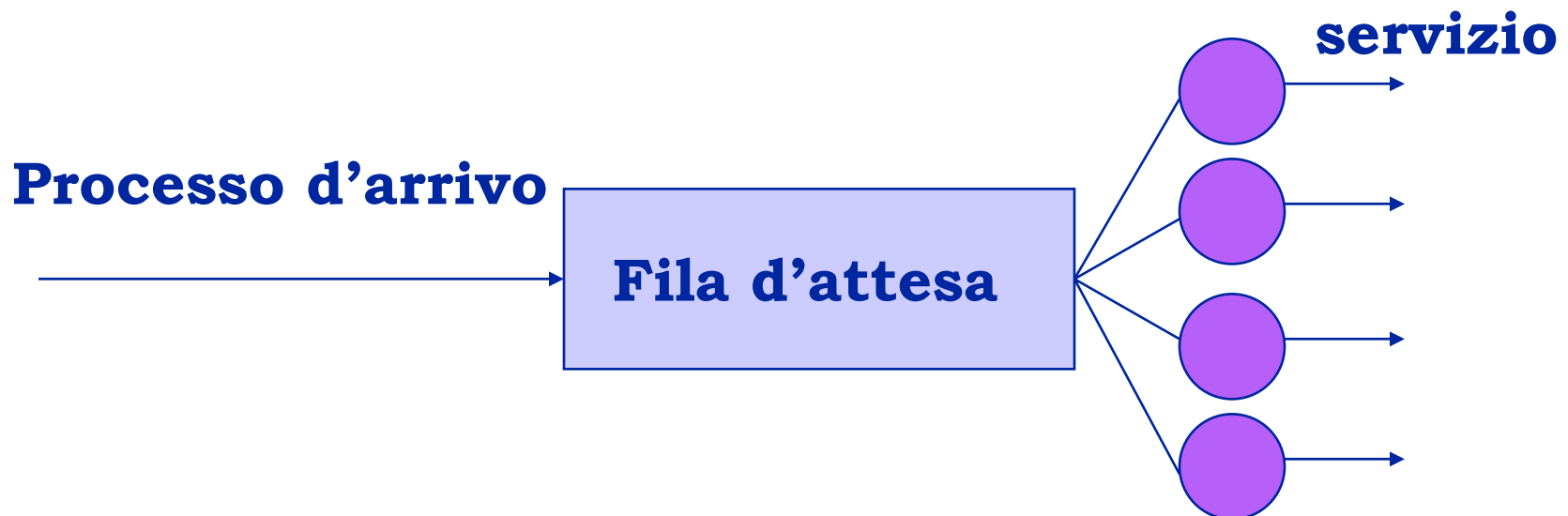
Sistemi ad attesa

- ◆ Nelle reti dati è molto comune la situazione in cui diversi flussi di pacchetti convergono in un dispositivo di instradamento (nodo, centrale, router, bridge, gateway, switch ...) e da questo si dividono in diversi flussi in uscita; i processi di arrivo sono causali ed è molto comune il caso in cui un pacchetto arriva al nodo quando altri pacchetti sono già in trasmissione sulla linea di uscita prescelta, per cui i pacchetti devono spesso attendere che il servente (trasmettitore) si liberi



Sistemi ad attesa

- ◆ **Un sistema ad attesa è un sistema che dispensa servizi, al quale giungono richieste di servizio (utenti) che vengono evase in base alla capacità del sistema**
- ◆ **Le richieste in attesa di servizio possono attendere in file d'attesa fino al momento dell'inizio del servizio, alla fine del quale abbandonano il sistema**



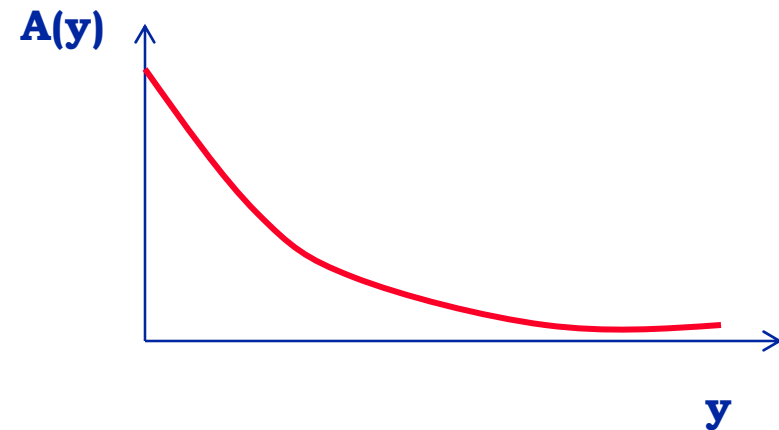
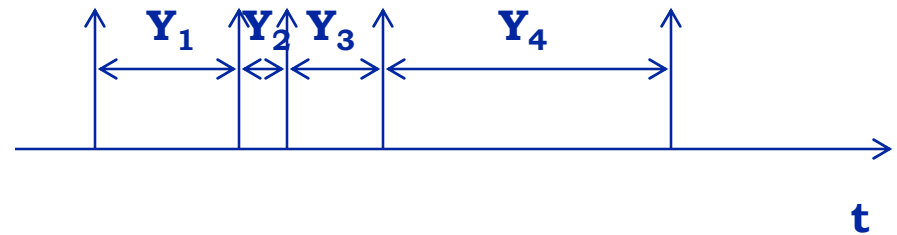
Sistemi ad attesa

◆ **I sistemi ad attesa sono caratterizzati dalle seguenti componenti**

- **PROCESSO DEGLI ARRIVI**
- **PROCESSO DI SERVIZIO**
- **MODALITÀ DI GESTIONE DEL SERVIZIO**
- **MODALITÀ DI GESTIONE DELLA FILA D'ATTESA**

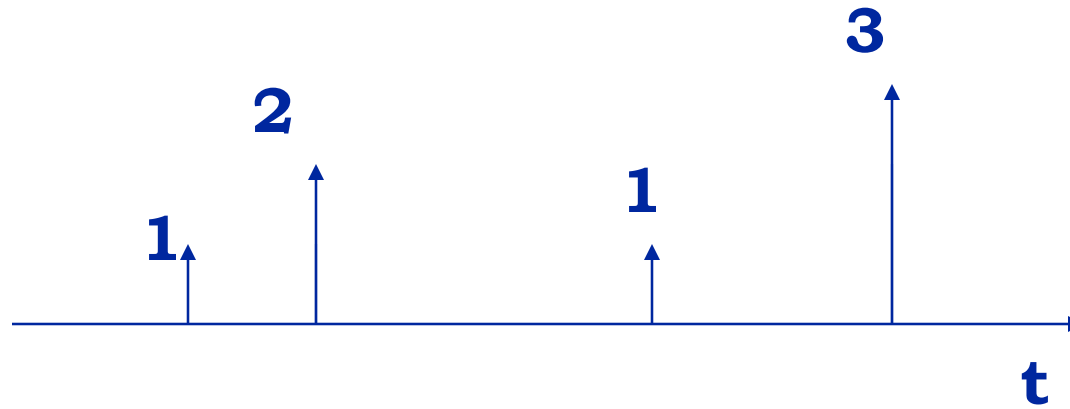
Processo degli arrivi

- ◆ Il processo degli arrivi è il processo secondo il quale gli utenti arrivano al sistema di servizio ad attesa
- ◆ Il processo di arrivo, di tipo puntuale $Y_1, Y_2, \dots, Y_n, \dots$ con densità di probabilità dell'intervallo Y di intercorrenza $a(y)$, funzione di distribuzione $A(y)$ e trasformata di Laplace $A^*(S)$
- ◆ Il tempo medio di interarrivo è $m_y = E[Y]$, per cui la frequenza media degli arrivi è pari a $\lambda = 1/m_y$



Processo degli arrivi

- ◆ Il processo degli arrivi può anche essere multiplo, nel senso che in un istante possono arrivare contemporaneamente più utenti al sistema
- ◆ In questo caso, è necessario caratterizzare il processo degli arrivi tramite la densità di probabilità del numero di arrivi contemporanei, $g(k)$, o tramite la associata trasformata Z , $G(z)$



Processo degli arrivi

- ◆ **Non tutti gli utenti che arrivano al sistema sono necessariamente accettati: alcuni possono essere rifiutati per vari motivi tra i quali, per esempio, la congestione del sistema**
- ◆ **λ_a è la frequenza di effettivo ingresso degli utenti nel sistema, λ_p è la frequenza di perdita di utenti**



Processo degli arrivi

- ◆ **Esiste l'ovvia relazione di congruenza $\lambda_o = \lambda_a + \lambda_p$**
- ◆ **λ_s è la frequenza di smaltimento degli utenti, e in condizioni normali si ha $\lambda_s = \lambda_a$**
- ◆ **In alcuni sistemi, come per esempio switch ATM che trasportano traffico IP, utenti (pacchetti IP) già in fila di attesa possono essere, in casi particolari, eliminati; per cui può accadere che λ_s è diverso da λ_a**



Richieste di servizio

◆ Sono i tempi necessari ad evadere il servizio per un servente che lavora a velocità costante

◆ Sono descritti dal processo $X_1, X_2, \dots, X_n, \dots$, a rinnovamento, con densità di probabilità $b(x)$, funzione di distribuzione $B(x)$ e trasformata di Laplace $B^*(s)$, e con valor medio $m_x = E[x]$

◆ Per esempio, nel caso di richieste di servizio esponenziali negative:

$$b(x) = \mu e^{-\mu x}$$

◆ Il valor medio del tempo di servizio e'

$$m_x = \int_0^{\infty} x \mu e^{-\mu x} dx = \frac{1}{\mu}$$

◆ La trasformata di Laplace di $b(x)$ e'

$$B^*(s) = \int_0^{\infty} \mu e^{-\mu x} e^{sx} dx = \frac{\mu}{s + \mu}$$

Processo di Poisson

- ◆ **Processi di servizio e di arrivo di tipo esponenziale negativo derivano tipicamente dai processi di Poisson**
- ◆ **Un processo di Poisson è definito nel seguente modo**
 - **ASSIOMA 1: La probabilità che si verifichi un evento nell'intervallo temporale Δt è pari a $\lambda\Delta t + o(\Delta t)$**
 - **ASSIOMA 2: La probabilità di più' di un evento in Δt è $o(\Delta t)$**
 - **ASSIOMA 3: Intervalli di tempo disgiunti sono statisticamente indipendenti**

Processo di Poisson

- ◆ La probabilità di k eventi di Poisson in un tempo T è pari a

$$P\{k, T\} = \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

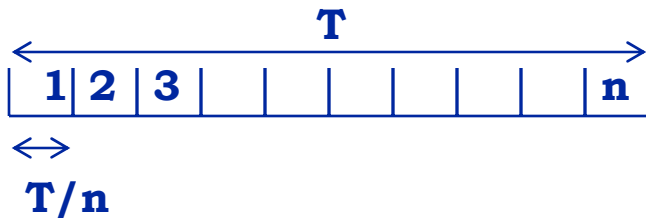
- ◆ Da cui

$$= \lim_{n \rightarrow \infty} \frac{n!}{(n-k)! k!} \frac{1}{\left(1 - \lambda \frac{T}{n}\right)^k} \left(\lambda \frac{T}{n}\right)^k \left(1 - \lambda \frac{T}{n}\right)^n$$

$$= \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-k+1)}{n^k} \frac{1}{k!} \frac{(\lambda T)^k}{1} e^{-\lambda T}$$

$$= \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

- ◆ Infatti, dividendo l'intervallo T in n intervallini



$$P\{k, T\} = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\lambda \frac{T}{n}\right)^k \left(1 - \lambda \frac{T}{n}\right)^{n-k}$$

Processo di Poisson

- ◆ Il processo di Poisson è senza memoria, cioè, la distribuzione del tempo di interarrivo condizionata ad un tempo t_0 già passato non varia; infatti, applicando la definizione di probabilità condizionata

$$P\{T \leq t + t_0 | T > t_0\} = \frac{P\{T \leq t + t_0, T > t_0\}}{P\{T > t_0\}}$$

- ◆ Che può essere scritta come

$$\frac{P\{t_0 < T \leq t + t_0\}}{1 - P\{T \leq t_0\}} = \frac{P\{T \leq t + t_0\} - P\{T \leq t_0\}}{1 - P\{T \leq t_0\}}$$

- ◆ Da cui

$$= \frac{1 - e^{-\lambda(t+t_0)} - 1 + e^{-\lambda t_0}}{e^{-\lambda t_0}}$$

- ◆ E infine

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

- ◆ La distribuzione dell'intervallo di tempo necessario per un nuovo evento quando si cominci ad osservare il processo all'istante t_0 è identica a quella dell'intervallo di tempo per un nuovo evento quando l'istante di osservazione iniziale è l'origine dei tempi

Modalità di gestione del servizio

◆ **Il servizio può essere impostato in varie modalità, tra cui, per esempio**

- **SINGOLO SERVITORE A VELOCITÀ COSTANTE O VARIABILE**
- **PIU' SERVITORI IN PARALLELO, CON ASSOCIATA UNA MODALITÀ DI SCELTA DEL SERVITORE DA PARTE DEGLI UTENTI**
- **ARTICOLAZIONI PIU' COMPLESSE**

Modalità di gestione della fila d'attesa

- ◆ **La fila d'attesa può avere capacità finita o infinita; nel caso di fila d'attesa finita alcuni utenti sono scartati perchè al loro arrivo non c'è posto in fila d'attesa**
- ◆ **Alcuni tipi di gestione della fila d'attesa sono:**
 - **FCFS (primo arrivato primo servito)**
 - **LCFS (ultimo arrivato primo servito)**
 - **RO (Random Order)**
 - **Politiche piu' complesse che presuppongono la divisione degli utenti in classi:**
 - **livelli di priorità, fair queuing, ecc.**

Modalità di gestione della fila d'attesa

- ◆ **Gli utenti possono essere suddivisi in classi di servizio, ognuna delle quali con diverse priorità; utenti a priorità più elevata sono serviti preferenzialmente, secondo diverse politiche**
- ◆ **Le politiche di servizio di utenti appartenenti a diverse classi di servizio sono utili al fine di assegnare un diverso grado di servizio alle varie classi**
- ◆ **D'altra parte, la gestione delle classi di servizio può essere molto complessa**

Sistemi ad attesa: grandezze di interesse

◆ Le principali grandezze d'interesse nei sistemi ad attesa sono

- il processo di occupazione, cioè il numero di utenti $N(t)$ nel sistema, al tempo t
- il processo di occupazione della fila d'attesa $N_c(t)$ (cioè il numero di utenti in fila d'attesa al tempo t)
- il processo $N_s(t) = N(t) - N_c(t)$ numero di utenti nel servizio al tempo t
- il processo V_i , tempo trascorso nel sistema dall'utente i
- il processo W_i , tempo trascorso in fila d'attesa dall'utente i (ovviamente, $V_i = W_i + X_i$)

Sistemi ad attesa: grandezze di interesse (continua)

- ◆ **Nel caso di fila d'attesa con capacità finita, interessa conoscere la congestione del sistema; esistono almeno due tipologie di congestione**
 - ***congestione temporale***: è la probabilità che $N(t)$ sia al suo valore massimo, per cui la congestione temporale è la probabilità che il sistema non sia in grado di ricevere ulteriori utenti
 - ***congestione di servizio***: è la probabilità che $N(t)$, campionato negli istanti di arrivo degli utenti, sia pari al suo valore massimo; la congestione di servizio è dunque la probabilità che un utente sia rifiutato al momento del suo arrivo
 - **N.B. La congestione di servizio e temporale NON coincidono sempre: questo avviene solo se il processo degli arrivi è senza memoria (Poissoniano ed indipendente dallo stato del sistema)**

Classificazione dei sistemi a coda

- ◆ **Esiste una classificazione standard dei sistemi a coda, detta notazione di Kendall**
- ◆ **In accordo a tale notazione, i sistemi a coda sono classificati come $X1/X2/m/c$, dove $X1$ indica il tipo di arrivi, $X2$ indica il tipo di servizi, m indica il numero di serventi (identici) in parallelo e c indica la capacità del sistema**
- ◆ **Per esempio, il classico sistema $M/M/1/C$ rappresenta una coda con interarrivi esponenziali (Markov), servizi esponenziali, un servente e capacità di contenere C utenti**

Classificazione dei sistemi a coda

- ◆ **I tipi di arrivi e servizi possono essere (tra le infinite possibilita')**
 - ***M* (Markov): interarrivi esponenziali, tra cui Poisson**
 - ***Ek*: Erlang-k**
 - ***D*: deterministici (utilizzabile nel caso di linee ATM, in quanto le celle hanno lunghezza, e quindi tempo di servizio, costante)**
 - **.....**
 - ***GI*: generali ed indipendenti dallo stato del sistema**

Probabilità di stato

- ◆ Un sistema a coda è caratterizzato dal processo $N(t)$, numero di utenti nel sistema
- ◆ Il processo $N(t)$ è tempo-continuo, cioè possono avvenire transizioni in ogni istante temporale
- ◆ D'altra parte $N(t)$ è discreto negli stati, nel senso che il numero di utenti nel sistema può essere solamente un numero intero
- ◆ La probabilità di stato $\pi_i(t)$ è definita come la probabilità che al tempo t ci siano i utenti nel sistema, ed è formalmente scritta come $\pi_i(t) = P[N(t) = i]$; il vettore $\Pi(t) = \pi_i(t)$ è il vettore delle probabilità di stato
- ◆ Un processo a stati discreti è anche detto *catena*

Probabilità di stato

- ◆ studiare il comportamento del vettore $\Pi(t) = \pi_i(t)$ nella sua evoluzione temporale vuol dire analizzare il comportamento in transitorio del sistema a coda
- ◆ di interesse (quando esiste) è studiare il

$$\lim_{t \rightarrow +\infty} \Pi(t)$$

- ◆ ovvero studiare il comportamento del sistema a regime

Catene di Markov

- ◆ Una catena è detta di Markov se vale la seguente proprietà

$$P\{X(t_n) = x_n | X(t_1) = x_1, X(t_2) = x_2, \dots, X(t_{n-1}) = x_{n-1}\} = P\{X(t_n) = x_n | X(t_{n-1}) = x_{n-1}\}$$

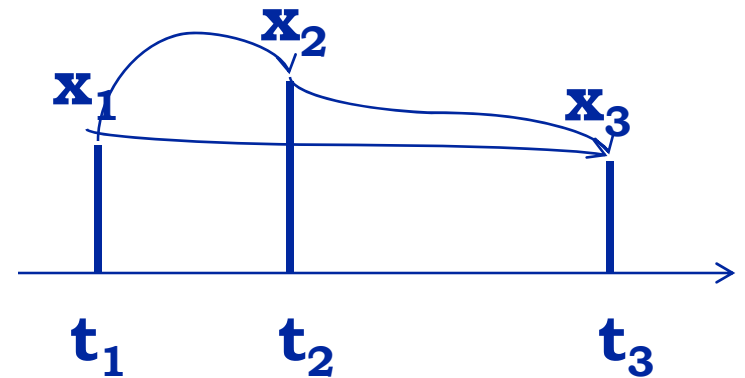
- ◆ Cioè, una volta che si conosca il processo all'istante t_{n-1} , la descrizione probabilistica del futuro (istante t_n) è indipendente da tutta la storia passata, ovvero dai valori che il processo ha assunto negli istanti precedenti a t_{n-1}
- ◆ Per questo motivo i processi di Markov vengono anche detti *senza memoria*

Catene di Markov

- ◆ Per i processi di Markov vale l'equazione di evoluzione di Chapman-Kolmogorov

$$P\{X(t_3) = x_3 | X(t_1) = x_1\} = \sum_{x_2 \in X} P\{X(t_3) = x_3 | X(t_2) = x_2\} P\{X(t_2) = x_2 | X(t_1) = x_1\}$$

- ◆ Che non è altro che una forma di applicazione del teorema della probabilità totale



La matrice stocastica

- ◆ Per descrivere l'evoluzione del processo, è di importanza fondamentale la matrice stocastica ad esso associata
- ◆ La matrice stocastica è definita come $P(t_1, t_2)$, ed il suo elemento $p_{jk}(t_1, t_2)$ è la probabilità di transizione dallo stato j (al tempo t_1) allo stato k (al tempo t_2):

$$p_{jk}(t_1, t_2) = P\{X(t_2) = k | X(t_1) = j\}$$

- ◆ La matrice P è detta stocastica perchè tutte le sue righe sommano a uno

Equazione di Chapman-Kolmogorov

- ◆ **L'equazione di Chapman-Kolmogorov può essere scritta in forma matriciale mediante la matrice stocastica delle probabilità di transizione:**

$$P\{X(t_3) = x_3 | X(t_1) = x_1\} = \sum_{x_2 \in X} P\{X(t_3) = x_3 | X(t_2) = x_2\} P\{X(t_2) = x_2 | X(t_1) = x_1\}$$

Può essere scritta anche come

$$P(t_1, t_3) = P(t_1, t_2) P(t_2, t_3)$$

Equazione di evoluzione

- ◆ **Applicando il teorema della probabilità totale si conclude facilmente che**

$$\pi_k(t_2) = \sum_i \pi_i(t_1) p_{ik}(t_1, t_2)$$

- ◆ **Che in notazione matriciale è scritto come**

$$\Pi(t_2) = \Pi(t_1) P(t_1, t_2)$$

- ◆ **Che è l'equazione di evoluzione del processo, dalla quale si può ricavare il vettore $\Pi(t)$, che descrive in modo completo il processo**

Catene tempo-omogenee

- ◆ **Le catene tempo-omogenee sono caratterizzate dall'aver un meccanismo di transizione che non varia nel tempo, cioè, la probabilità di transizione dallo stato i allo stato j non dipende dal tempo assoluto in cui questa transizione avviene, ma solo dalla durata della transizione stessa:**

$$\Pi(t_2) = \Pi(t_1)P(t_1, t_2)$$

che si può scrivere anche come

$$\Pi(t_0 + \Delta t) = \Pi(t_0)P(t_0, t_0 + \Delta t)$$

diventa

$$\Pi(\Delta t) = \Pi(0)P(\Delta t)$$

Catene tempo-omogenee

- ◆ **Nel caso di catene tempo-omogenee l'equazione di Chapman-Kolmogorov**

$$P(t_1, t_3) = P(t_1, t_2)P(t_2, t_3)$$

Può essere scritta come

$$P(\Delta t + \Delta \tau) = P(\Delta t)P(\Delta \tau)$$

Infatti, in una catena tempo-omogenea, per una transizione è importante solo la durata, e non il tempo assoluto di occorrenza

Equazioni risolutive

- ◆ Le equazioni esplicite per la risoluzione di una catena tempo omogenea sono ricavate così: per definizione

$$\frac{dP(t)}{dt} = \lim_{\Delta t \rightarrow 0} \frac{P(t + \Delta t) - P(t)}{\Delta t}$$

- ◆ E avvalendosi dell'equazione di Chapman-Kolmogorov

$$P(\Delta t + \Delta \tau) = P(\Delta t)P(\Delta \tau)$$

- ◆ Si ottiene

$$\frac{dP(t)}{dt} = \lim_{\Delta t \rightarrow 0} \frac{P(t)P(\Delta t) - P(t)}{\Delta t}$$

- ◆ E quindi, raccogliendo $P(t)$

$$\frac{dP(t)}{dt} = P(t) \lim_{\Delta t \rightarrow 0} \frac{P(\Delta t) - I}{\Delta t}$$

- ◆ In cui I è la matrice identità; per definizione si pone

$$Q = \lim_{\Delta t \rightarrow 0} \frac{P(\Delta t) - I}{\Delta t}$$

- ◆ E Q viene denominata matrice delle frequenze di transizione. L'equazione risolutiva diventa quindi

$$\frac{dP(t)}{dt} = P(t)Q$$

Equazioni risolutive

- ◆ Analogamente si può scrivere:

$$\frac{d\Pi(t)}{dt} = \Pi(t)Q$$

- ◆ Infatti:

$$\begin{aligned}\frac{d\Pi(t)}{dt} &= \lim_{\Delta t \rightarrow 0} \frac{\Pi(t + \Delta t) - \Pi(t)}{\Delta t} = \\ &= \lim_{\Delta t \rightarrow 0} \frac{\Pi(t)P(\Delta t) - \Pi(t)}{\Delta t} = \\ &= \Pi(t) \lim_{\Delta t \rightarrow 0} \frac{P(\Delta t) - I}{\Delta t} = \Pi(t)Q\end{aligned}$$

Equazioni risolutive

- ◆ L'equazione

$$\frac{d\Pi(t)}{dt} = \Pi(t)Q$$

- ◆ Permette di ricavare l'evoluzione transitoria del sistema; in generale interessa invece lo stato stazionario del sistema stesso, cioè'

$$\Pi = \lim_{t \rightarrow \infty} \Pi(t)$$

- ◆ Per ottenere il vettore di stato stazionario P si potrebbe risolvere l'equazione differenziale ed eseguire il limite per t tendente a infinito di $P(t)$ (ALTAMENTE SCONSIGLIATO!!!)

- ◆ Altrimenti, per ricavare il vettore di probabilità stazionario basta ricordare che in regime stazionario la derivata prima di $P(t)$ è identicamente nulla, per cui il problema è ridotto alla soluzione di un sistema algebrico lineare

$$\Pi Q = 0$$

- ◆ Resta solo da identificare esattamente la matrice Q

Equazioni risolutive

- ◆ **Q è definita come**

$$Q = \lim_{\Delta t \rightarrow 0} \frac{P(\Delta t) - I}{\Delta t}$$

- ◆ **e quindi**

$$\begin{cases} q_{ij} = \lim_{\Delta t \rightarrow 0} \frac{p_{ij}(\Delta t)}{\Delta t} \\ q_{ii} = \lim_{\Delta t \rightarrow 0} \frac{p_{ii}(\Delta t) - 1}{\Delta t} \end{cases}$$

- ◆ **Se si assume che i processi di arrivo e di servizio del sistema considerato sono di Poisson, si ricava immediatamente che**

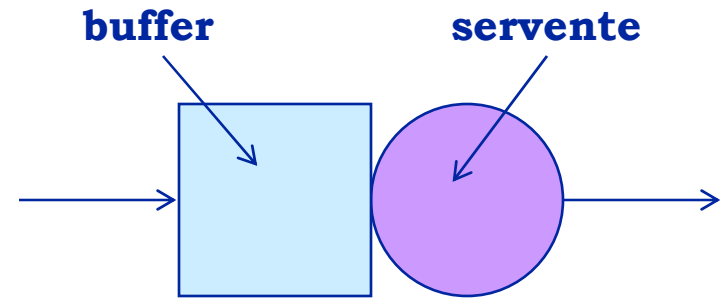
$$\begin{cases} q_{ij} = \lim_{\Delta t \rightarrow 0} \frac{\lambda_{ij} \Delta t + o(\Delta t)}{\Delta t} = \lambda_{ij} \\ q_{ii} = \lim_{\Delta t \rightarrow 0} \frac{1 - \sum_{j \neq i} p_{ij}(\Delta t) - 1}{\Delta t} = -\sum_{j \neq i} \lambda_{ij} \end{cases}$$

- ◆ **Da cui si ricavano i coefficienti del sistema lineare algebrico per ricavare il vettore di stato stazionario**
- ◆ **Si noti che la matrice Q è singolare, per cui per la risoluzione del sistema è necessario aggiungere l'equazione di congruenza**

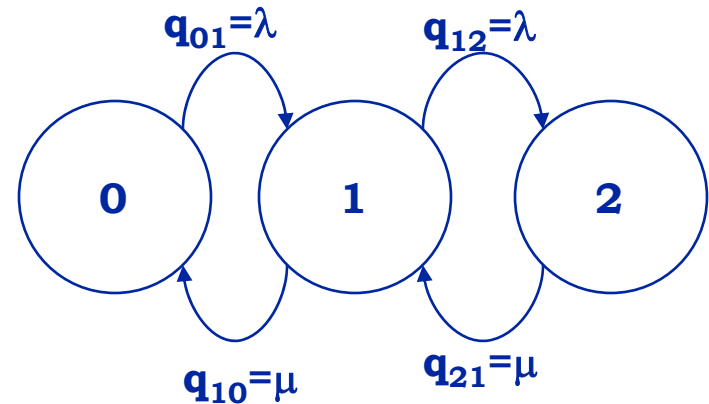
$$\sum_i \pi_i = 1$$

Esempio

- ◆ Ad una linea di trasmissione a velocità C [bit/s] arrivano, secondo un processo di Poisson a tasso λ , pacchetti di lunghezza distribuita in modo esponenziale negativo, con lunghezza media pari a l [bit]; la linea è dotata di un buffer in grado di memorizzare un pacchetto nell'eventualità in cui ci sia un nuovo arrivo mentre la linea è occupata nella trasmissione di un pacchetto precedentemente arrivato. Un pacchetto che arriva e trova il buffer occupato è perso



- ◆ Il sistema ha 3 stati (0, 1, 2 pacchetti nel sistema)



- ◆ In cui $\mu = C/l$

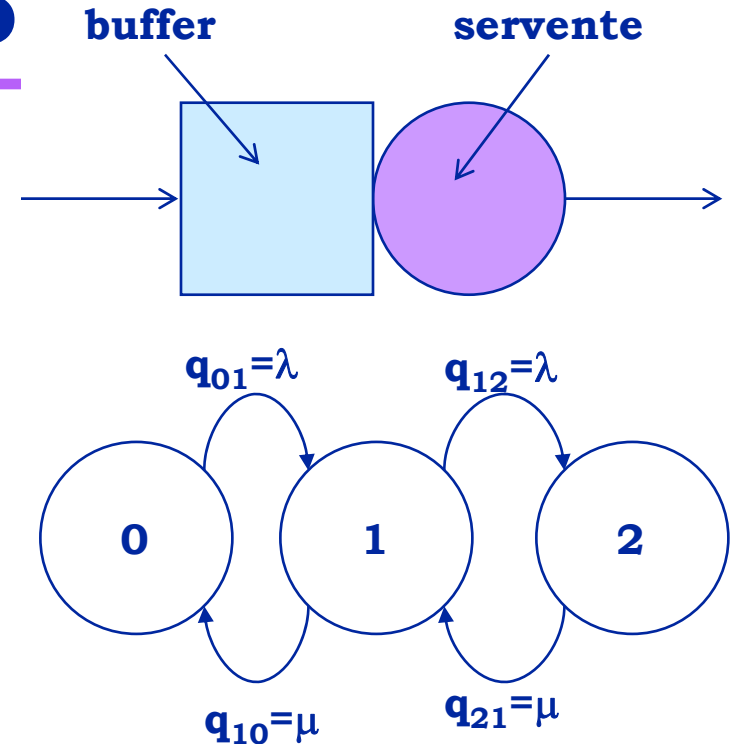
Esempio

◆ Il sistema risolutore è dato da

$$\begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \end{bmatrix}^T \begin{bmatrix} -\lambda & \lambda & 0 \\ \mu & -(\mu + \lambda) & \lambda \\ 0 & \mu & -\mu \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}^T$$

◆ Che è singolare, visto che le righe di Q sommano a 0, per cui bisogna aggiungere l'equazione di congruenza

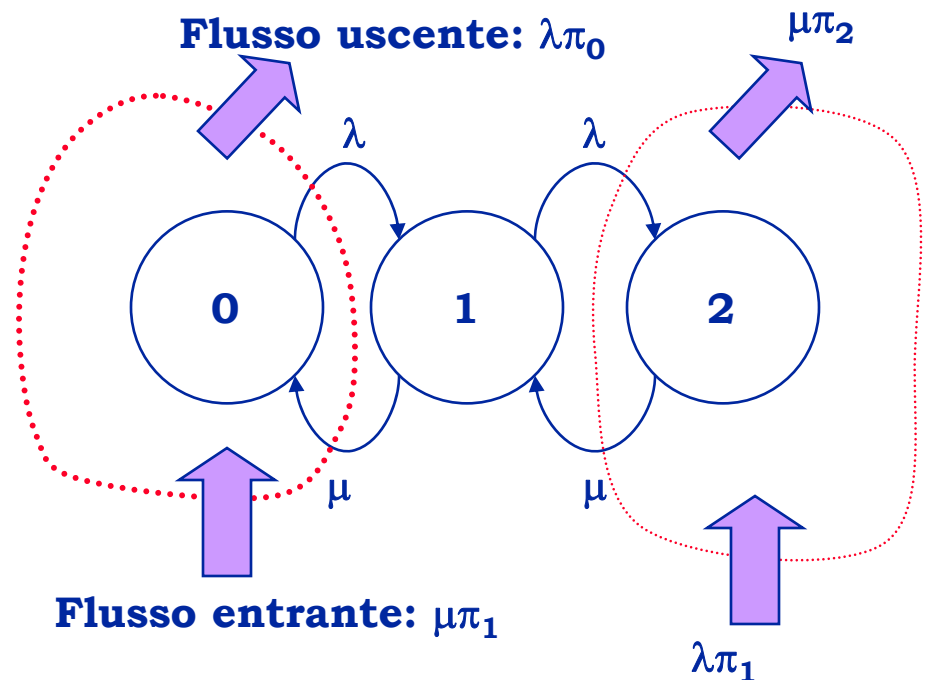
$$\pi_0 + \pi_1 + \pi_2 = 1$$



$$\begin{cases} \pi_0 + \pi_1 + \pi_2 = 1 \\ -\pi_0 \lambda + \pi_1 \mu = 0 \\ \pi_1 \lambda - \pi_2 \mu = 0 \end{cases}$$

Bilancio dei flussi

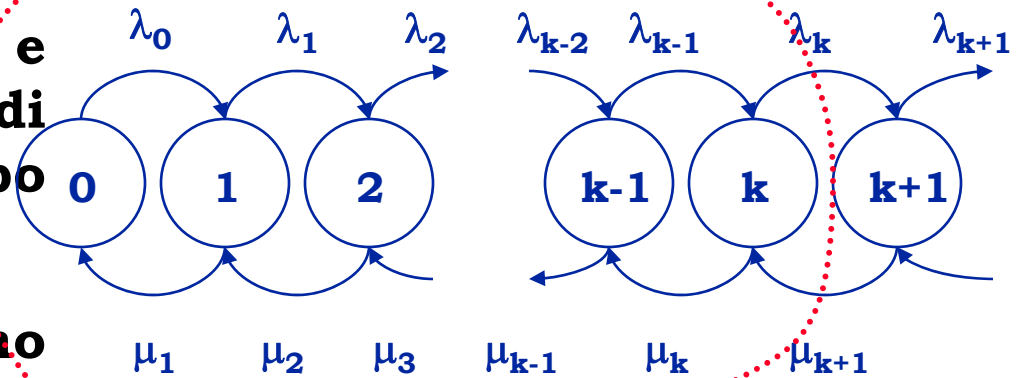
- ◆ Data una catena di Markov a N stati, si possono scrivere $N-1$ equazioni tramite il bilancio dei flussi
- ◆ Teorema: Data una qualsiasi superficie chiusa nella catena di Markov, in condizioni stazionarie il flusso di probabilità entrante è uguale al flusso di probabilità uscente
- ◆ Questa tecnica serve a scrivere il sistema di equazioni “*by inspection*”



$$\begin{cases} \lambda \pi_0 = \mu \pi_1 \\ \lambda \pi_1 = \mu \pi_2 \\ \pi_0 + \pi_1 + \pi_2 = 1 \end{cases}$$

Processi di nascita e morte

- ◆ Un processo di nascita e morte ha una catena di Markov associata del tipo mostrato in figura



- ◆ Molti sistemi a code sono rappresentabili tramite un processo di nascita e morte

- ◆ Tramite il bilancio dei flussi alla superficie indicata si ricava l'equazione ricorsiva

$$\pi_k \lambda_k = \pi_{k+1} \mu_{k+1}$$

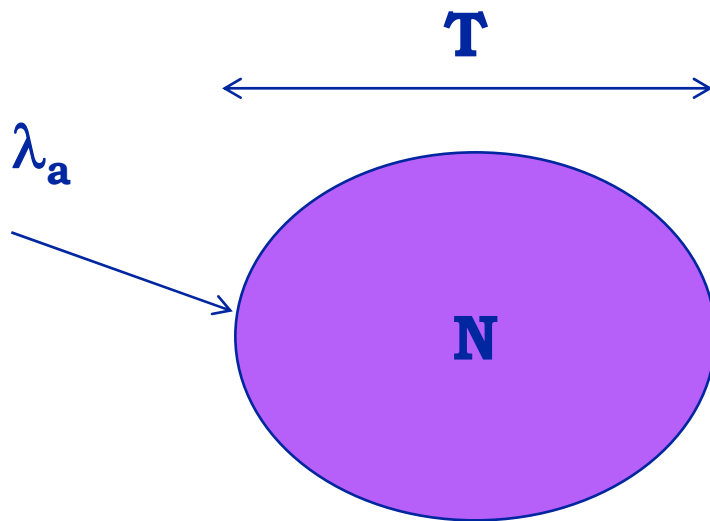
$$\left\{ \begin{array}{l} \pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \\ \pi_0 = \left[1 + \sum_{j=1}^{\Omega} \prod_{i=0}^{j-1} \frac{\lambda_i}{\mu_{i+1}} \right]^{-1} \end{array} \right.$$

- ◆ Da cui si ricava la soluzione

- ◆ Dove Ω è il numero di stati, finito o infinito

Il Little's result

- ◆ Il Little's result è molto importante e di validità molto generale
- ◆ Il Little's result mette in relazione il numero medio di utenti in un qualsiasi sottosistema ed il tempo medio speso dagli utenti nel sottosistema stesso



Delimitato un qualsiasi sottosistema in una superficie, detto λ_a il tasso medio di utenti accettati, cioè che entrano nella superficie, la relazione tra numero medio di utenti nella superficie, $E[N]$, e tempo medio speso dagli utenti nella superficie, $E[T]$, è data da

$$E[N] = \lambda_a E[T]$$

Il Little's result

- ◆ L'area della zona compresa tra le curve $N_a(\tau)$ e $N_p(\tau)$ può essere espressa in due modi:

$$\int_0^t N(\tau) d\tau = \sum_{i=1}^{N_p(t)} T_i + \sum_{i=N_p(t)+1}^{N_a(t)} (t - T_i)$$

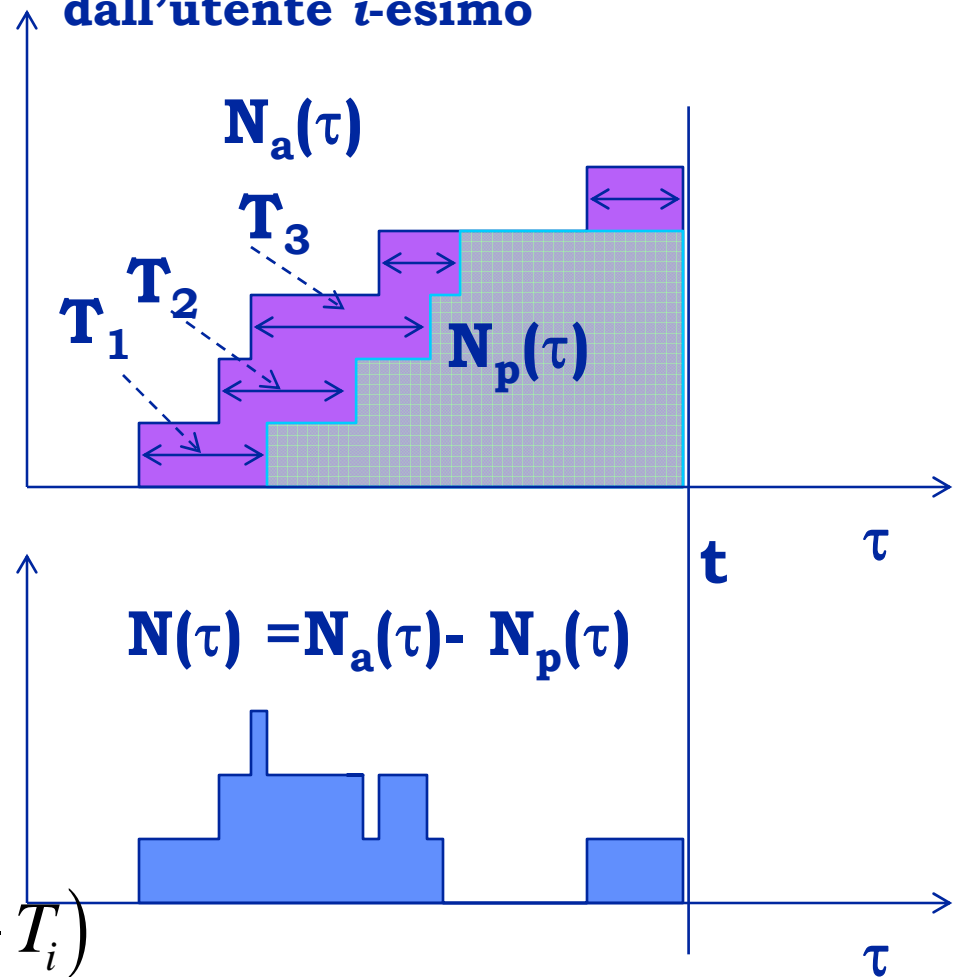
- ◆ E, dividendo per t :

$$\frac{\int_0^t N(\tau) d\tau}{t} = \frac{\sum_{i=1}^{N_p(t)} T_i + \sum_{i=N_p(t)+1}^{N_a(t)} (t - T_i)}{t}$$

- ◆ Da cui:

$$\frac{\int_0^t N(\tau) d\tau}{t} = \frac{N_a(t)}{t} \frac{\sum_{i=1}^{N_p(t)} T_i + \sum_{i=N_p(t)+1}^{N_a(t)} (t - T_i)}{N_a(t)}$$

T_i = tempo passato nel sistema dall'utente i -esimo



Il Little's result

◆ Gli ingredienti sono:

- media temporale del numero di utenti nel sistema nell'intervallo 0-t
- media temporale del tasso di arrivo in 0-t
- media temporale del tempo speso nel sistema dagli utenti in 0-t

$$\frac{\int_0^t N(\tau) d\tau}{t} = \frac{N_a(t)}{t} \frac{\sum_{i=1}^{N_p(t)} T_i + \sum_{i=N_p(t)+1}^{N_a(t)} (t - T_i)}{N_a(t)}$$

The diagram illustrates the derivation of Little's result. It shows three components from the list above, each with an arrow pointing to a part of the equation:

- The first arrow points from "media temporale del numero di utenti nel sistema nell'intervallo 0-t" to the first fraction $\frac{\int_0^t N(\tau) d\tau}{t}$.
- The second arrow points from "media temporale del tasso di arrivo in 0-t" to the second fraction $\frac{N_a(t)}{t}$.
- The third arrow points from "media temporale del tempo speso nel sistema dagli utenti in 0-t" to the third fraction $\frac{\sum_{i=1}^{N_p(t)} T_i + \sum_{i=N_p(t)+1}^{N_a(t)} (t - T_i)}{N_a(t)}$.

Il Little's result

- ◆ Prendendo il limite per t tendente a infinito dei vari componenti elencati si scrive

$$\lim_{t \rightarrow \infty} \frac{\int_0^t N(\tau) d\tau}{t} = \lim_{t \rightarrow \infty} \frac{N_a(t)}{t} \lim_{t \rightarrow \infty} \frac{\sum_{i=1}^{N_p(t)} T_i + \sum_{i=N_p(t)+1}^{N_a(t)} (t - T_i)}{N_a(t)}$$

- ◆ Da cui, sorvolando su alcune ipotesi sull'ergodicità del processo, si ottiene

$$E[N] = \lambda_a E[T]$$

Il Little's result

◆ In modo un po' piu' rigoroso:

◆ Per i processi rigenerativi ed ergodici vale:

$$E[N(t)] = \frac{E\left[\int_{ciclo} N(t) dt\right]}{E[C]}$$

◆ Dove C è il ciclo di rigenerazione del processo.

Il Little's result

◆ **Essendo**

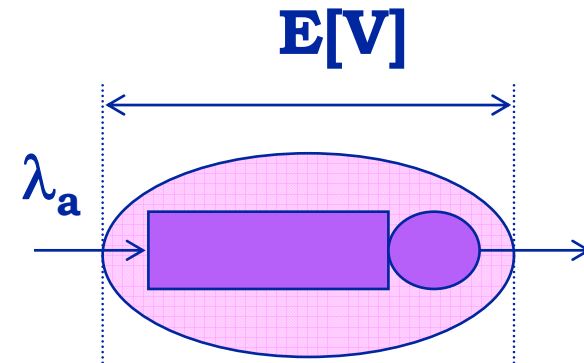
$$\int_{\text{ciclo}} N(t) dt = \sum_{i=1}^A T_i$$

◆ **con A arrivi nel ciclo, si ha:**

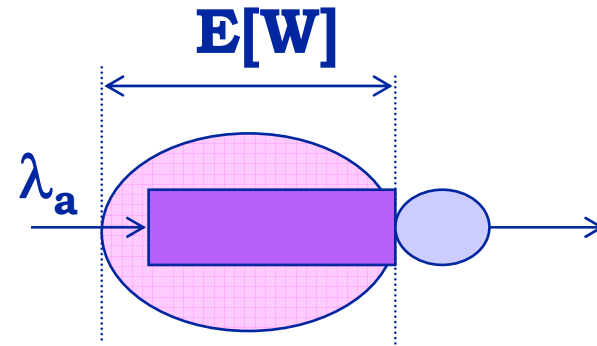
$$E[N(t)] = \frac{E\left[\int_{\text{ciclo}} N(t) dt\right]}{E[C]} = \frac{E\left[\sum_{i=1}^A T_i\right]}{E[C]} = \frac{E\left[\sum_{i=1}^A T_i\right]}{E[A]} \frac{E[A]}{E[C]} = E[T] \lambda_a$$

Il Little's result: esempio di applicazione

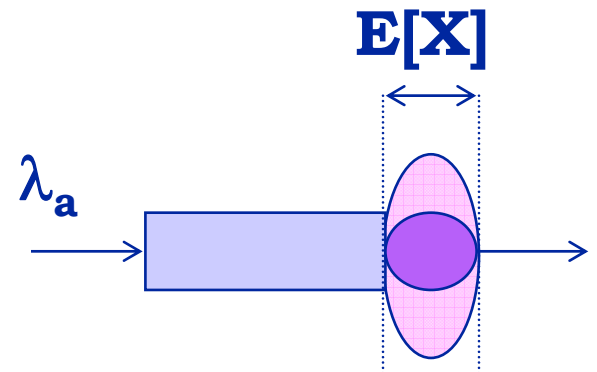
- ◆ Con riferimento all'intera coda: $E[N] = \lambda_a E[V]$



- ◆ Con riferimento alla sola fila di attesa : $E[N_c] = \lambda_a E[W]$

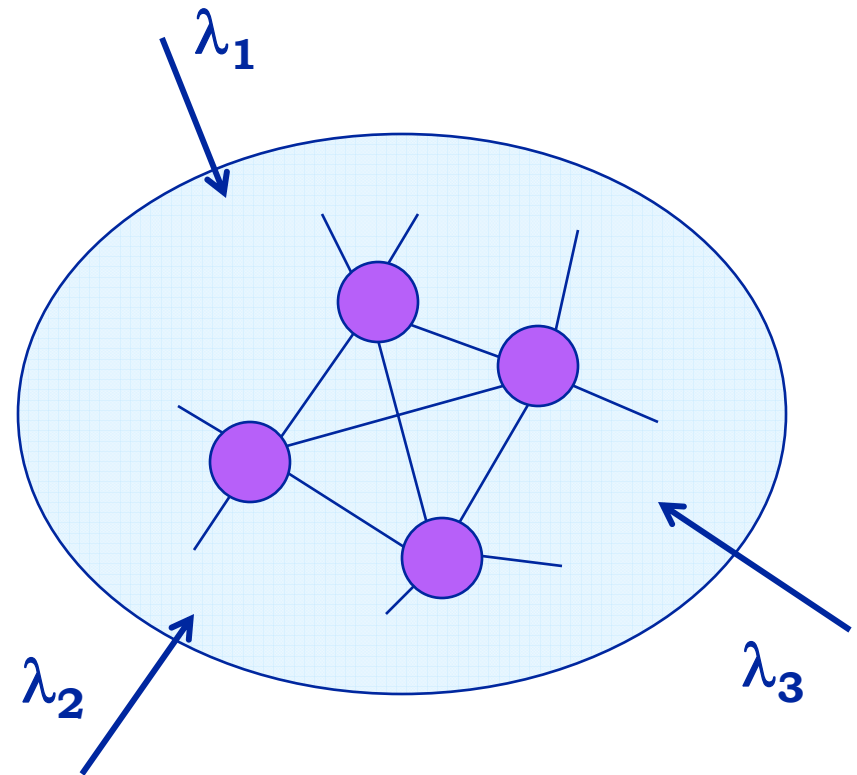


- ◆ Con riferimento alla solo servente:
 $E[S] = \rho = \lambda_a E[N_s] = \lambda_a m_x$



Il Little's result: esempio di applicazione

- ◆ In una rete a pacchetto, ci siano n flussi di traffico in ingresso, per un totale tasso di arrivo dato da $\Lambda = \lambda_1 + \lambda_2 + \dots + \lambda_n$
- ◆ Inoltre, sia N il numero medio totale di pacchetti nella rete
- ◆ Il tempo totale medio speso da un pacchetto generico all'interno della rete è $T=N/\Lambda$
- ◆ Indipendentemente da:
 - distribuzione delle lunghezze dei pacchetti
 - distribuzione degli interarrivi
 - metodi di instradamento



Fattore di utilizzo

- ◆ Il fattore di utilizzo (carico) ρ di un servente è definito come la frazione di tempo per cui il servente lavora
- ◆ Il carico è dunque uguale alla probabilità di trovare il servente occupato in un istante di ispezione casuale
- ◆ Dal Little's result, si ha che $\rho = \lambda_s m_x$
- ◆ Nel caso di m serventi identici, il carico del singolo servente è dato da $\rho = \lambda_s m_x / m$, in quanto ogni servente riceve una frazione $1/m$ dei clienti

Distribuzione all'ingresso e all'uscita

◆ è interessante a volte conoscere la distribuzione del numero di utenti nel sistema visti da un utente alla fine del suo servizio (distribuzione all'uscita)

◆ Detto $N^*(t_j)$ questo numero di utenti e definita r_i la $P[N^*(t_j)=i]$, si ha la seguente relazione

$$r_i = \frac{\lambda_i}{\lambda_a} \pi_i = \frac{\mu_{i+1}}{\lambda_a} \pi_{i+1}$$

◆ La distribuzione del numero di utenti nel sistema vista da un utente al momento del suo arrivo e'

$$q_i = \frac{\lambda_i}{\lambda_o} \pi_i$$

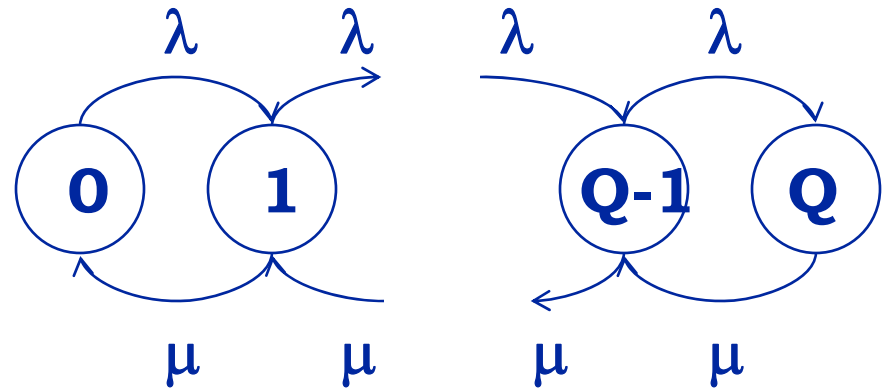
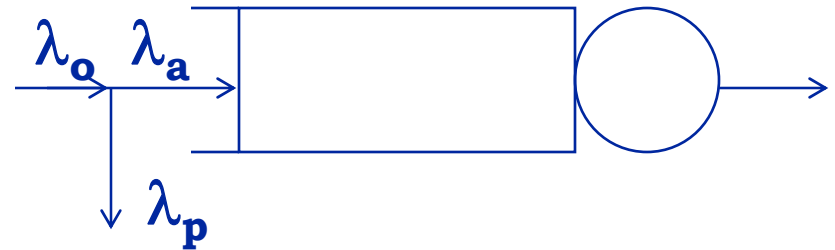
Il sistema M/M/1/Q

- ◆ Nel sistema M/M/1/Q i tempi di interarrivo ed i tempi di servizio sono esponenziali negativi:

$$b_Y(y) = \lambda e^{-\lambda y}$$

$$b_X(x) = \mu e^{-\mu x}$$

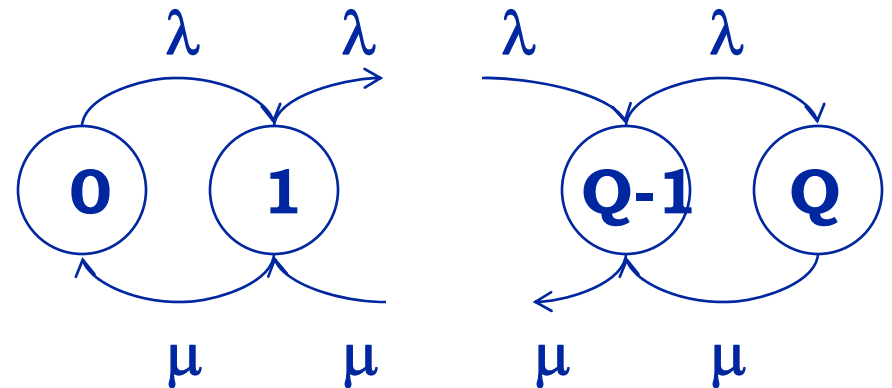
- ◆ Questo conduce alla catena di Markov rappresentata a lato
- ◆ Q è il massimo numero di utenti ammessi nel sistema; se un utente arriva quando ci sono già Q utenti nel sistema, viene respinto



Il sistema M/M/1/Q

- ◆ La catena di Markov a lato si risolve banalmente, come riportato a lato; il rapporto $\rho_o = \lambda/\mu$ è il carico offerto al sistema
- ◆ La probabilità dello stato i è infine data da

$$\pi_i = \frac{1 - \rho_o}{1 - \rho_o^{Q+1}} \rho_o^i$$



$$\pi_{i+1}\mu = \pi_i\lambda$$

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0$$

$$\sum_{i=0}^Q \pi_i = 1 \Rightarrow \pi_0 = \frac{1 - \rho_o}{1 - \rho_o^{Q+1}}$$

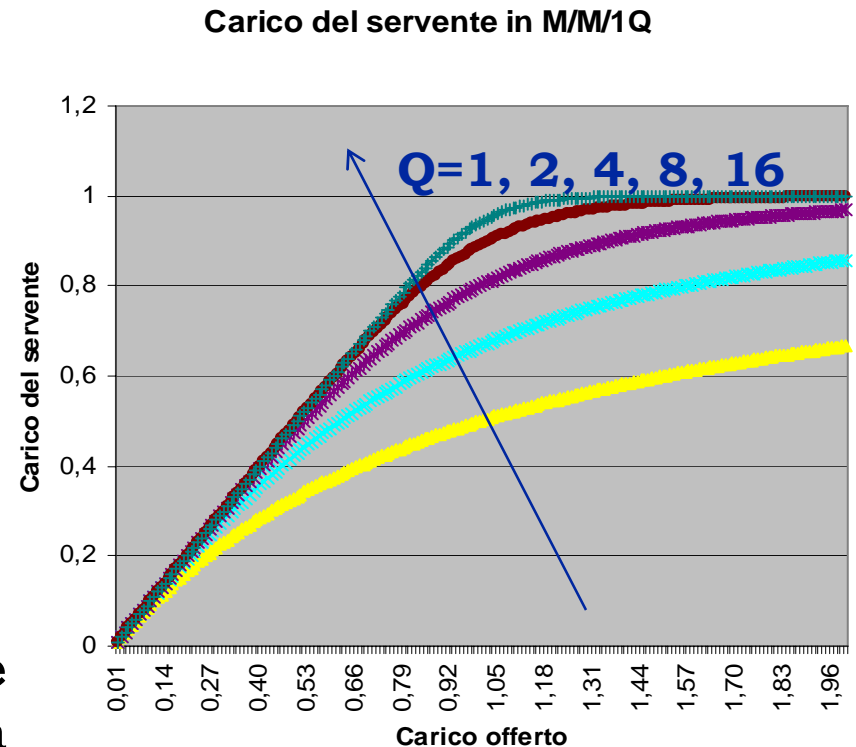
Il sistema M/M/1/Q

- ◆ Il fattore di utilizzo del servente ρ e', per definizione, la frazione di tempo in cui il servente è occupato, per cui può essere calcolato come $\rho = 1 - \pi_0$

- ◆ Dai precedenti risultati si ottiene

$$\rho = 1 - \frac{1 - \rho_0}{1 - \rho_0^{Q+1}}$$

- ◆ Il carico del servente è sempre inferiore a quello offerto; fino a carico offerto pari a 1, un sistema con elevata capacità (Q) può avere un carico del servente molto vicino al carico offerto

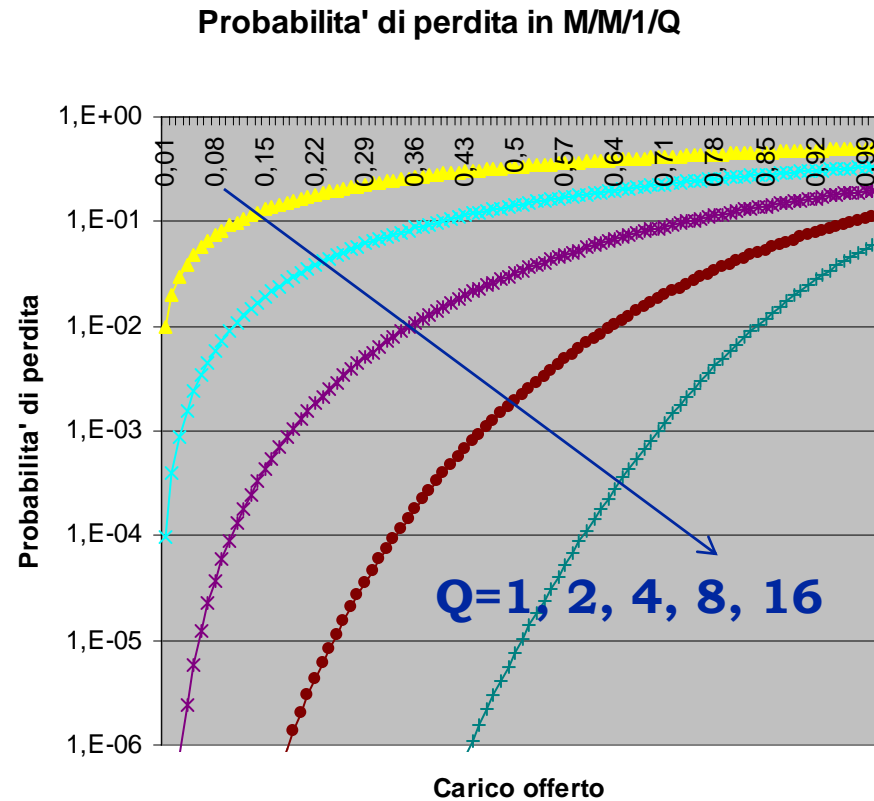


Il sistema M/M/1/Q

- ◆ La probabilità di perdita di utenti è definita come

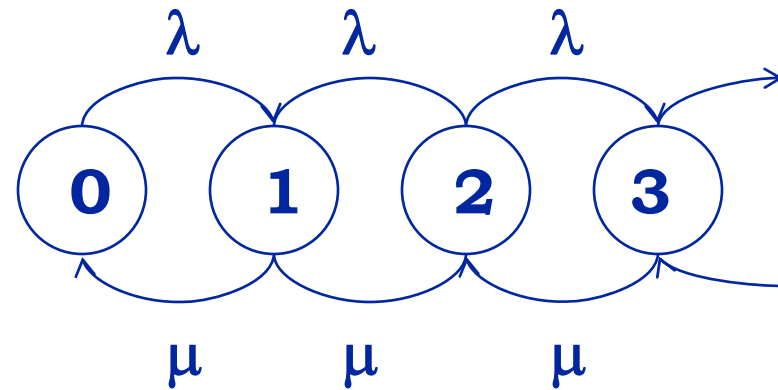
$$B = \frac{\lambda_p}{\lambda} = \frac{\pi_Q \lambda}{\lambda} = \pi_Q$$

- ◆ Ed è riportata nel grafico a lato



Il sistema M/M/1/∞

- ◆ In questo sistema i clienti non sono mai persi, in quanto la fila d'attesa ha capacità infinita
- ◆ Questo pone dei problemi sulla stabilità del sistema stesso: come si intuisce facilmente, se il tasso di arrivo è maggiore (o uguale) al tasso di servizio, il numero di utenti nel sistema tende a crescere senza limiti



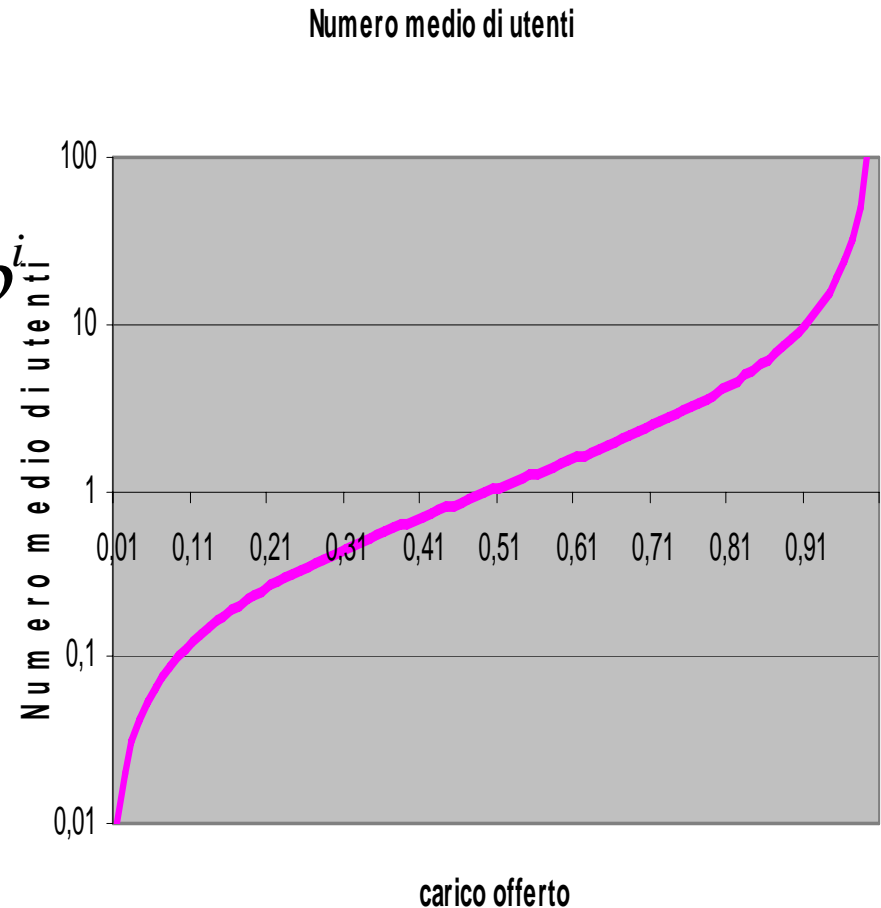
$$\pi_i = \pi_0 \rho^i = (1 - \rho) \rho^i$$

Il sistema M/M/1/∞

- ◆ Il numero medio di utenti nel sistema è dato da

$$E[N] = \sum_{i=0}^{\infty} i\pi_i = (1 - \rho) \sum_{i=0}^{\infty} i\rho^i$$

- ◆ Al tendere del carico a 1, il numero medio di utenti nel sistema tende a infinito
- ◆ Per questo motivo, il punto di lavoro di una coda va progettato per un carico sensibilmente inferiore a 1

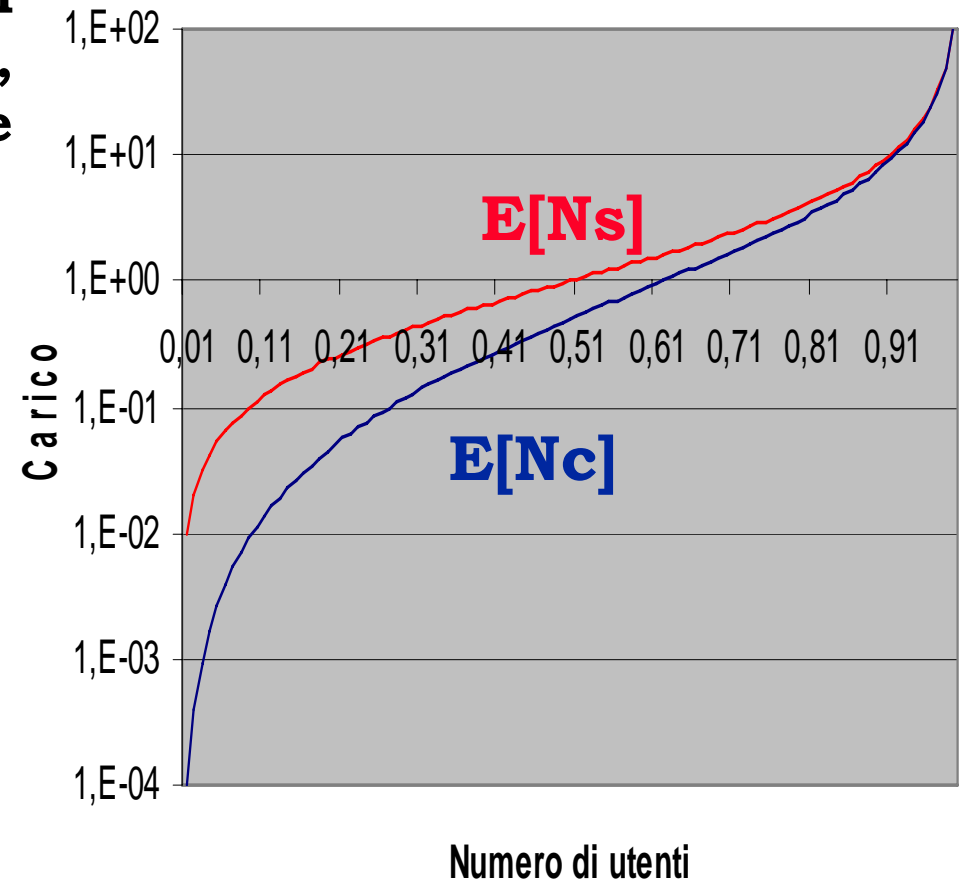


Il sistema M/M/1/∞

- ◆ Il numero medio di utenti nel servente è uguale a ρ , per cui, applicando l'ovvia equazione $E[N_s] = E[N_c] + \rho$ si ottiene

$$E[N_c] = \frac{\rho^2}{1-\rho}$$

- ◆ Ovviamente, anche il numero medio di utenti in fila d'attesa tende a infinito al tendere a uno del carico



Il sistema M/M/1/∞

- ◆ Il ritardo medio nel sistema, $E[V]$, si ricava tramite il Little's result:

$$E[V] = \frac{E[N]}{\lambda} = \frac{1}{\lambda} \frac{\rho}{1-\rho} = \frac{1}{\mu - \lambda}$$

- ◆ è molto importante notare che, al contrario del numero medio di utenti, il ritardo non dipende solo dal carico, ma anche dalla lunghezza media dei servizi ($1/\mu$)
- ◆ Questo significa che a parità di carico, una diversa lunghezza media dei servizi comporta differenti tempi di attesa

Il sistema M/M/1/∞

- ◆ Per esempio, consideriamo una linea a velocità C [bit/s], a cui sono offerti pacchetti con processo di arrivo Poissoniano e lunghezza dei pacchetti esponenziale di lunghezza media L [bit]
- ◆ Vogliamo individuare il ritardo di accodamento più trasmissione dei pacchetti
- ◆ Il ritardo è riportato a lato, tenendo conto del fatto che la capacità del servente, in servizi per unità di tempo, è in questo caso data dalla velocità della linea divisa per la lunghezza media L (in bit) dei pacchetti



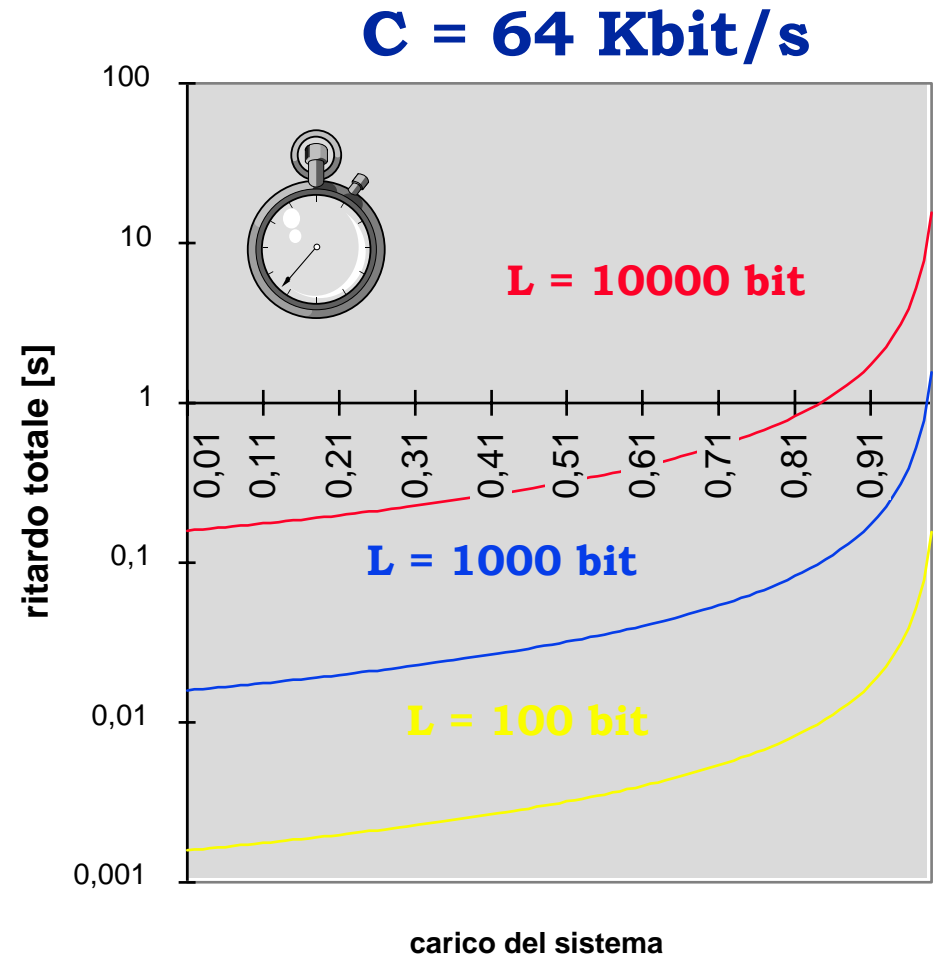
$$E[V] = \frac{1}{\mu - \lambda} = \frac{1}{\mu} \frac{1}{1 - \rho}$$

$$\mu = \frac{C}{L}$$

$$E[V] = \frac{1}{\frac{C}{L} - \lambda} = \frac{L}{C} \frac{1}{1 - \rho}$$

Il sistema $M/M/1/\infty$

- ◆ Come facilmente prevedibile, il ritardo tende a infinito se il carico del sistema tende a 1
- ◆ Inoltre, si nota che, a parità di carico, la lunghezza media dei pacchetti influisce notevolmente sul ritardo
- ◆ In particolare a lunghezza media maggiore corrisponde ritardo maggiore
- ◆ Questo è dovuto al fatto che il numero medio di pacchetti in fila d'attesa non dipende dalla lunghezza media dei pacchetti stessi



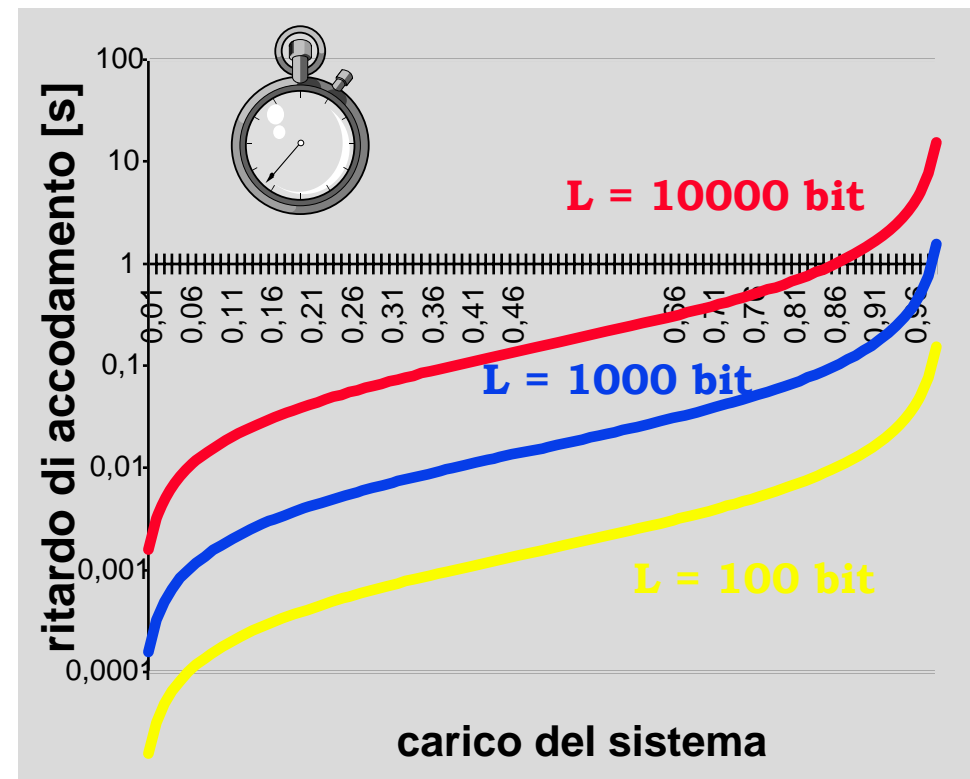
Il sistema M/M/1/∞

- ◆ Il ritardo medio in fila di attesa, $E[W]$, si ricava in modo analogo tramite il Little's result:

$$E[W] = \frac{E[N_c]}{\lambda} = \frac{1}{\lambda} \frac{\rho^2}{1-\rho} = \frac{\rho}{\mu - \lambda}$$

- ◆ Si nota che $E[W]$ tende a zero al tendere a zero del carico; questo non accade per $E[V]$, in quanto anche se il carico tende a zero, comunque in $E[V]$ c'è la componente di ritardo di trasmissione, L/C , che è costante

C = 64 Kbit/s



Il sistema M/M/1/∞

◆ Per esempio, se si vuole trasportare un flusso di λ pacchetti/s di lunghezza media pari a L bit con un totale ritardo medio di accodamento e trasmissione pari a T , si deve adottare una linea di capacità C pari a

$$C \geq L \left(\lambda + \frac{1}{T} \right)$$

◆ Per esempio, se $\lambda = 20$ [pacchetti/s], di lunghezza media pari a 1500 bit e si desidera un ritardo medio $T = 50$ ms, si ottiene $C > 60$ Kbit/s, per cui si deve adottare una linea da 64 Kbit/s (il ritardo effettivo risulta essere pari a 44 ms)

Il sistema M/M/1/∞

◆ Oltre alla media del ritardo di accodamento W e del ritardo di sistema V , è interessante conoscerne anche le distribuzioni (o le densità di probabilità)

◆ La densità di probabilità del tempo di sistema (accodamento più trasmissione) può essere calcolata condizionando al numero di utenti incontrati nel sistema da un nuovo utente

$$f_V(v) = \sum_{i=0}^{\infty} f_V(v|i)(1-\rho)\rho^i$$

◆ La densità di probabilità del tempo di sistema condizionata ad i utenti incontrati dal nuovo utente al suo arrivo è la convoluzione di $i+1$ esponenziali (i servizi più il servizio del nuovo utente), per cui:

$$f_V(v) = \sum_{i=0}^{\infty} \left(\mu e^{-\mu v} \right)^{\otimes(i+1)} (1-\rho)\rho^i$$

Il sistema M/M/1/∞

- ◆ La convoluzione di k esponenziali dà come risultato la distribuzione Erlang- k

$$\left(\mu e^{-\mu v}\right)^{(i+1)} = \frac{(\mu v)^i}{i!} \mu e^{-\mu v}$$

- ◆ Per cui si ottiene

$$\begin{aligned} f_V(v) &= \sum_{i=0}^{\infty} \frac{(\mu v)^i}{i!} \mu e^{-\mu v} (1-\rho) \rho^i = \\ &= (1-\rho) \mu e^{-\mu v} \sum_{i=0}^{\infty} \frac{(\mu v \rho)^i}{i!} \end{aligned}$$

- ◆ Per cui, infine:

$$f_V(v) = (1-\rho) \mu e^{-(1-\rho)\mu v}$$

- ◆ La densità di probabilità del ritardo di sistema è dunque un'esponenziale!

- ◆ La funzione di distribuzione di V è data da

$$F_V(v) = 1 - e^{-(1-\rho)\mu v}$$

Il sistema M/M/1/∞

- ◆ Per quanto riguarda il tempo di attesa W , in modo analogo si ricava che

$$f_W(w) = (1 - \rho)\delta(w) + \mu\rho(1 - \rho)e^{-(1-\rho)\mu w}$$

- ◆ $\delta(w)$ è la distribuzione delta di Dirac, ed il primo termine corrisponde all'evento in cui l'utente arriva nel sistema e lo trova vuoto (tempo di attesa nullo)
- ◆ La funzione di distribuzione di W è data da

$$F_W(w) = 1 - \rho e^{-(1-\rho)\mu w}$$

Il sistema M/M/1/∞

- ◆ Tramite le distribuzioni di V e W si può eseguire il progetto sui quantili del ritardo, piuttosto che sul ritardo medio
- ◆ Per esempio, è interessante valutare il ritardo V_r che non è superato per una percentuale di tempo uguale a $r/100$; V_r è definito dall'equazione

$$P\{V \leq V_r\} = \frac{r}{100}$$

- ◆ Dalla quale si ottiene

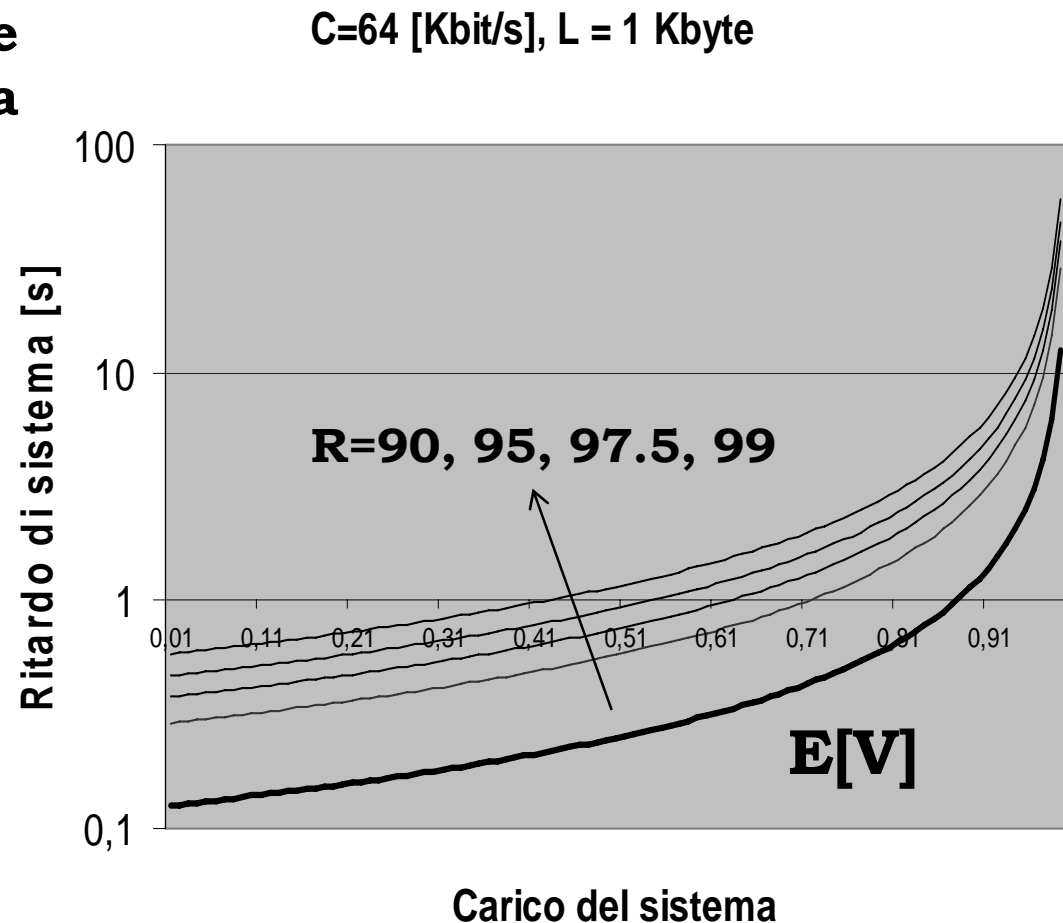
$$V_r = \frac{\log_e \left(\frac{100}{100 - r} \right)}{\mu(1 - \rho)}$$

- ◆ Tramite questa formula si può dimensionare una linea affinché il ritardo di sistema superi una soglia preassegnata con una probabilità preassegnata

Il sistema M/M/1/∞

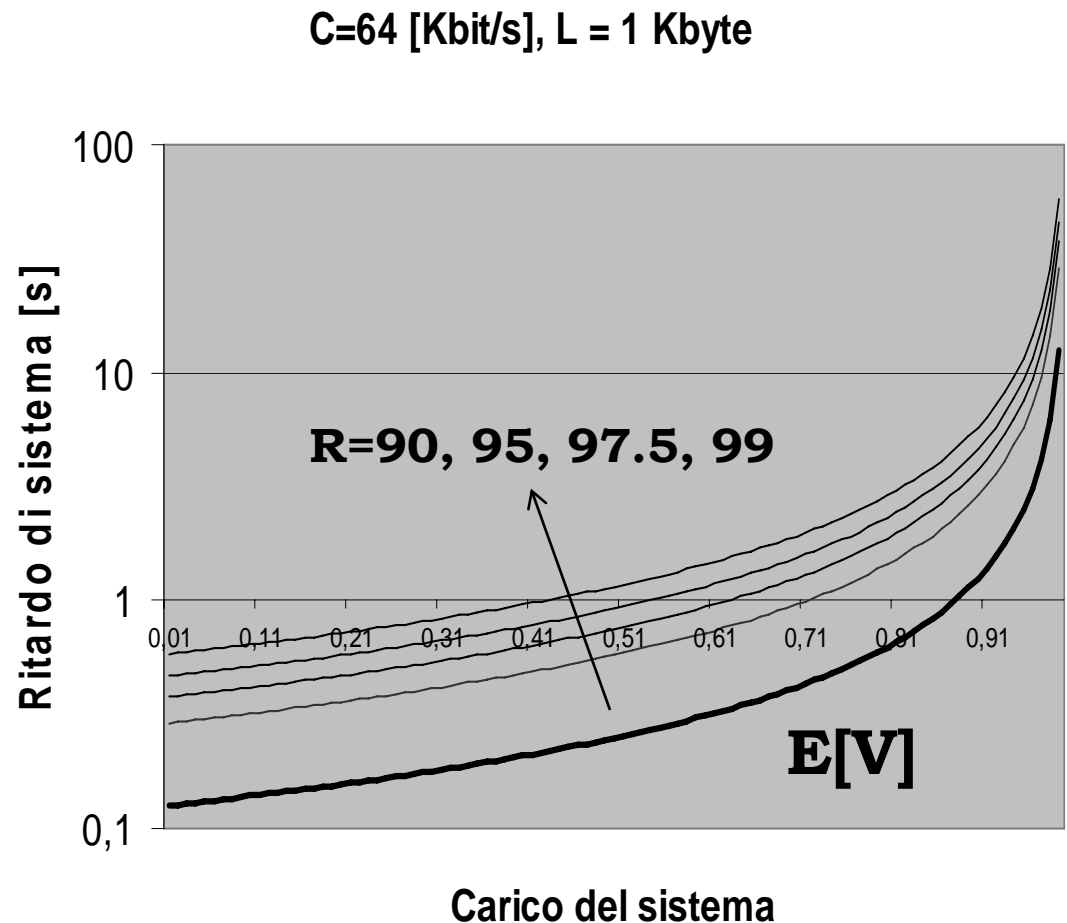
- ◆ Per esempio, nel caso di una linea a velocità $C = 64$ Kbit/s e pacchetti di lunghezza media $L = 1$ Kbyte si ottiene

$$V_r = \frac{\log_e \left(\frac{100}{100 - r} \right)}{8(1 - \rho)}$$



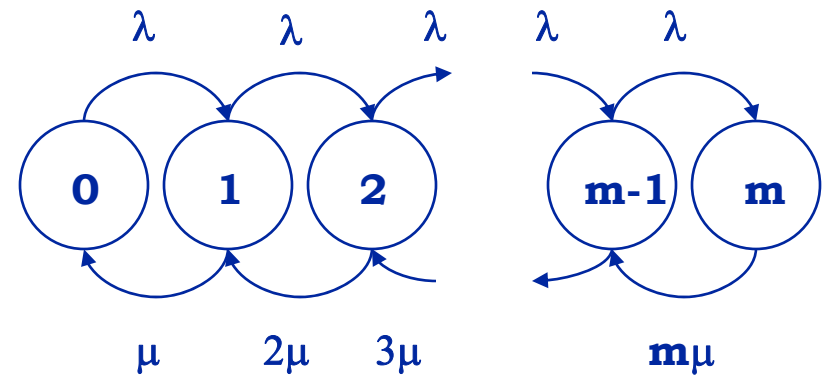
Il sistema $M/M/1/\infty$

- ◆ Per esempio, se si volesse progettare la linea a 64 Kbit/s per un ritardo medio di 1 s, si potrebbe accettare un carico pari a 0.88, per cui si potrebbero accettare 7.04 pacchetti/s
- ◆ Se invece si volesse un ritardo di un secondo, ma non superato nel 95% dei casi, si avrebbe un massimo carico ammissibile pari a 0.52 e dunque si potrebbero offrire al massimo 4.16 pacchetti/s



Sistema a pura perdita: M/M/m/0

- ◆ Un classico esempio di sistema a pura perdita è il fascio di giunzioni telefonico, in cui m circuiti sono disponibili per trasportare connessioni
- ◆ La catena associata (del tipo nascita e morte) è rappresentata a lato
- ◆ Tramite le equazioni risolutive generali dei processi di nascita e morte si ottiene la soluzione riportata a lato



$$p_k = \frac{A_o^k}{k! \sum_{j=0}^m \frac{A_o^j}{j!}}$$

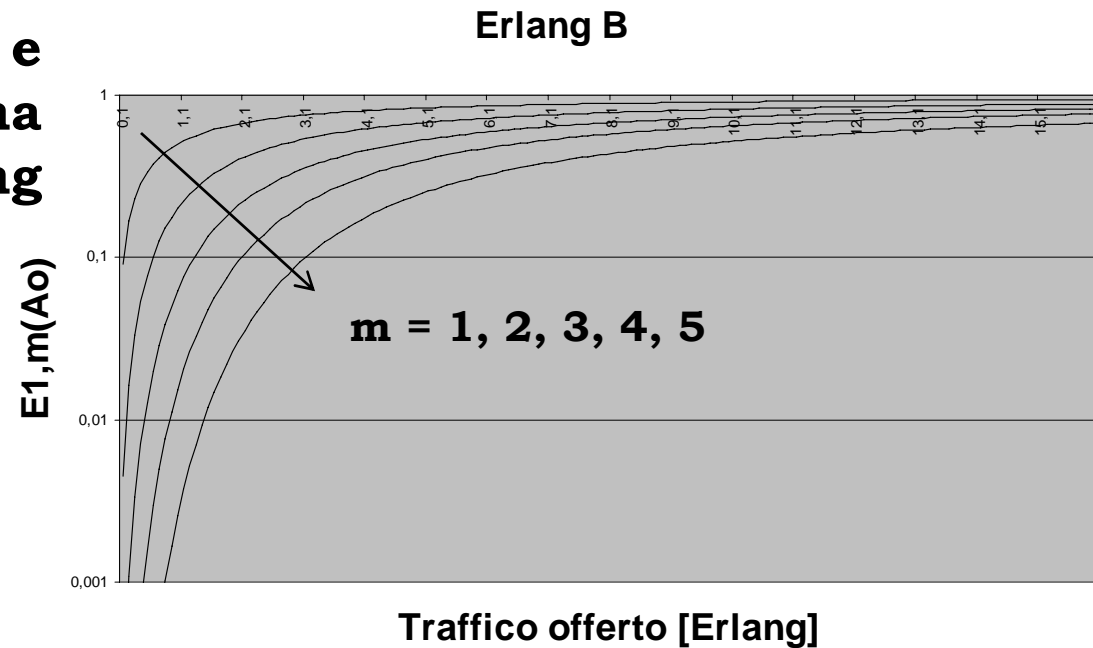
Sistema a pura perdita: M/M/m/0

- ◆ La probabilità di congestione e di perdita del sistema M/M/m/0 è la famosa Erlang B:

$$\Pi_p = S_p = \frac{\frac{A_o^m}{m!}}{\sum_{j=0}^m \frac{A_o^j}{j!}} = E_{1,m}(A_o)$$

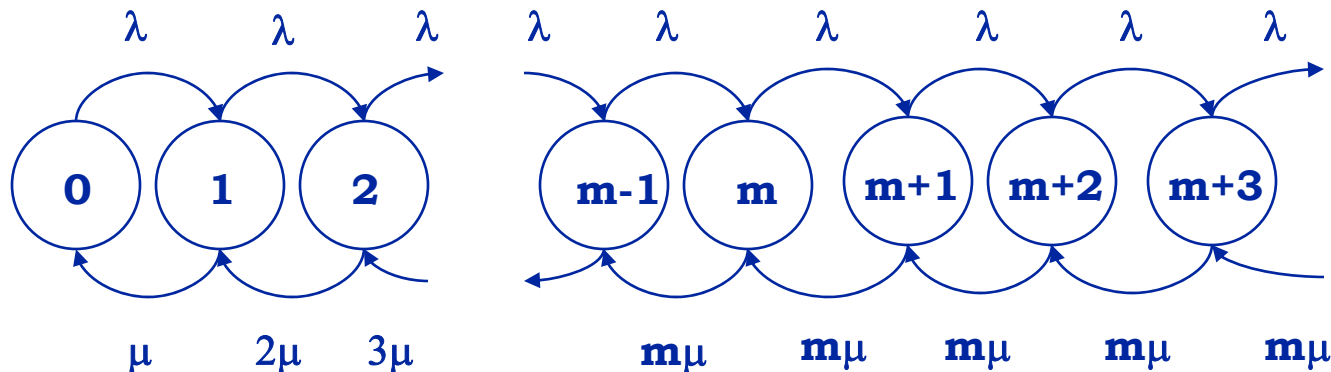
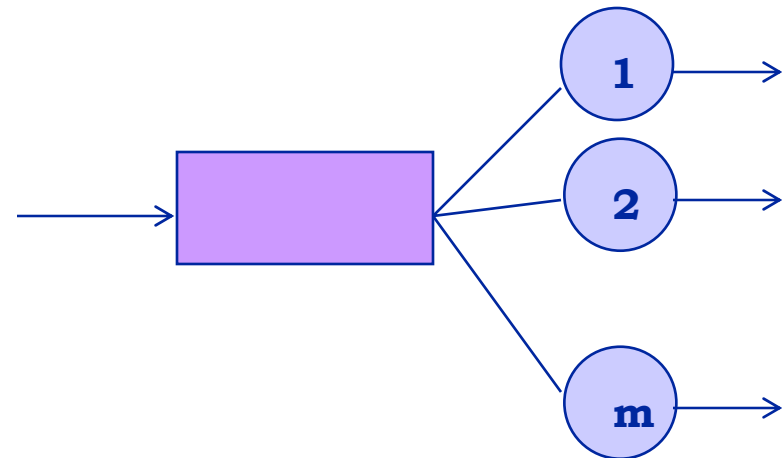
- ◆ Che si può facilmente porre in forma ricorsiva epr una computazione piu' efficiente

$$E_{1,m}(A_o) = \frac{A_o E_{1,m-1}(A_o)}{m + E_{1,m-1}(A_o)}$$



Sistema M/M/m/ ∞

- ◆ Questo sistema è detto a serventi multipli e senza perdita
- ◆ Infatti, ci sono m serventi e la fila d'attesa ha capacità infinita
- ◆ Un possibile sistema reale modellabile in questo modo è un fascio di giunzioni in cui, se una connessione trova tutti i circuiti occupati, viene posta in attesa invece che persa



Sistema M/M/m/∞

◆ Applicando le equazioni risolutive dei processi di nascita e morte si ottiene

$$p_k = \begin{cases} p_0 \frac{A_o^k}{k!}; & k \leq m \\ p_0 \frac{A_o^k}{k!} \frac{1}{m! m^{k-m}}; & k > m \end{cases}$$

$$p_0 = \left[\sum_{j=0}^{m-1} \frac{A_o^j}{j!} + \frac{A_o^m}{m!} \frac{m}{m - A_o} \right]^{-1}$$

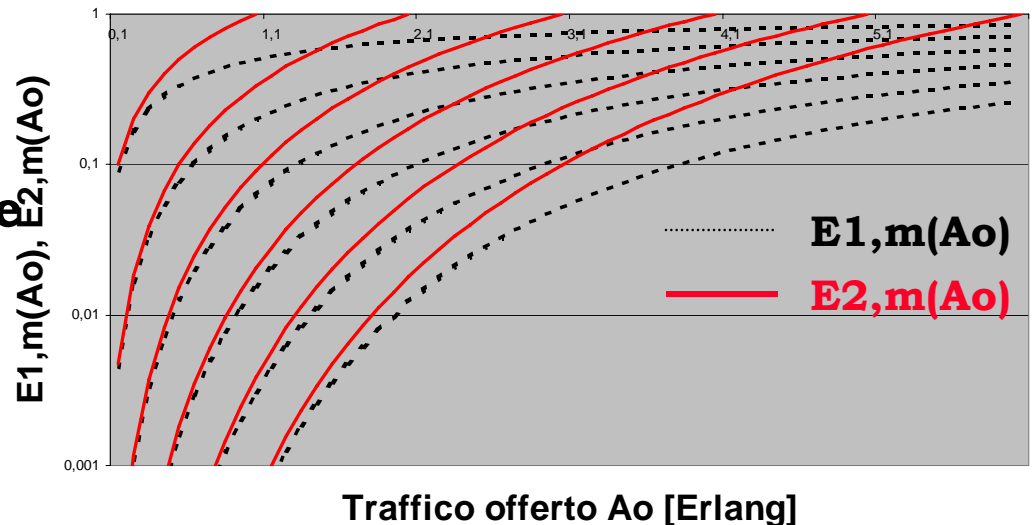
◆ E la probabilità che un utente venga ritardato è pari a

$$S_r = \Pi_r = \sum_{k=m}^{\infty} p_k = \frac{\frac{A_o^m}{m!} \frac{m}{m - A_o}}{\frac{A_o^m}{m!} \frac{m}{m - A_o} + \sum_{j=0}^{m-1} \frac{A_o^j}{j!}}$$

◆ Ed è detta Erlang C ($E_{2,m}(A_o)$); la Erlang C può essere espressa in forma ricorsiva

$$E_{2,m}(A_o) = \frac{E_{1,m}(A_o)}{1 - \frac{A_o}{m} [1 - E_{1,m}(A_o)]}$$

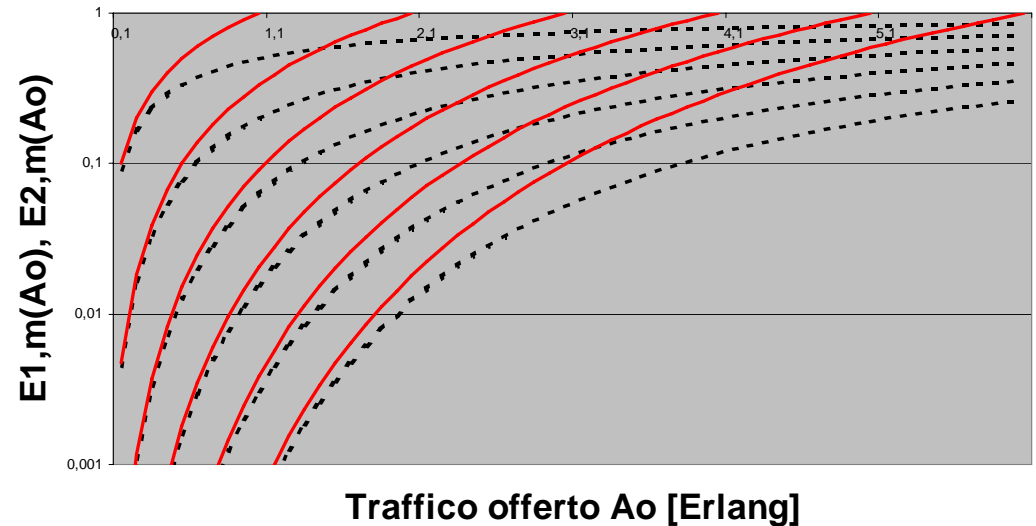
Erlang B ed Erlang C



Sistema $M/M/m/\infty$

- ◆ A parità di carico offerto e numero di giunzioni, la Erlang C è maggiore della Erlang B
- ◆ Questo è intuitivo, infatti nel sistema ad attesa il traffico smaltito è maggiore che in quello a perdita, per cui i server sono mediamente più occupati, per cui è più probabile che un nuovo utente al suo arrivo veda tutti i server impegnati
- ◆ La differenza tra la Erlang B e la Erlang C è molto significativa in regime di “alta perdita” e tende ad annullarsi in regime di “bassa perdita”

Erlang B ed Erlang C



- ◆ Anche questo è intuitivo: se la perdita del sistema $M/M/m/0$ è bassa allora il traffico smaltito è quasi uguale a quello offerto, per cui l'occupazione dei server è quasi uguale a quella del sistema ad attesa

Sistema M/M/m/∞

- ◆ Nel sistema ad attesa è interessante il tempo medio speso dagli utenti in attesa del servizio
- ◆ A tal fine, si calcola in primo luogo la lunghezza media della fila di attesa, e poi tramite il Little's result si ricava il ritardo

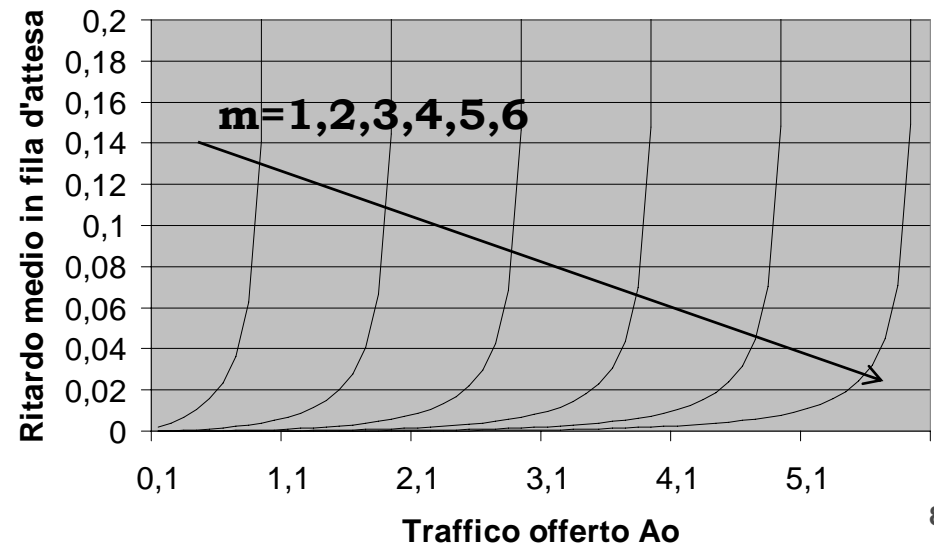
$$E[N_c] = \sum_{i=m}^{\infty} (i - m) p_i = \frac{A_o}{m - A_o} E_{2,m}(A_o)$$

$$E[N] = E[N_c] + E[N_s] = A_o \left[1 + \frac{E_{2,m}(A_o)}{m - A_o} \right]$$

- ◆ Da cui si ricavano i ritardi medi

$$E[W] = \frac{E[N_c]}{\lambda} = \frac{1}{\mu} \left[\frac{E_{2,m}(A_o)}{m - A_o} \right]$$

$$E[V] = \frac{E[N]}{\lambda} = \frac{1}{\mu} \left[1 + \frac{E_{2,m}(A_o)}{m - A_o} \right]$$



Sistema M/M/m/ ∞

- ◆ **La densità di probabilità del tempo trascorso in fila di attesa è data da**

$$f_W(w) = [1 - E_{2,m}(A_o)]\delta(w) + E_{2,m}(A_o)\mu(m - A_o)e^{-(m-A_o)\mu w}$$

- ◆ **E la relativa funzione di distribuzione e'**

$$F_W(w) = [1 - E_{2,m}(A_o)] - E_{2,m}(A_o)e^{-(m-A_o)\mu w}$$

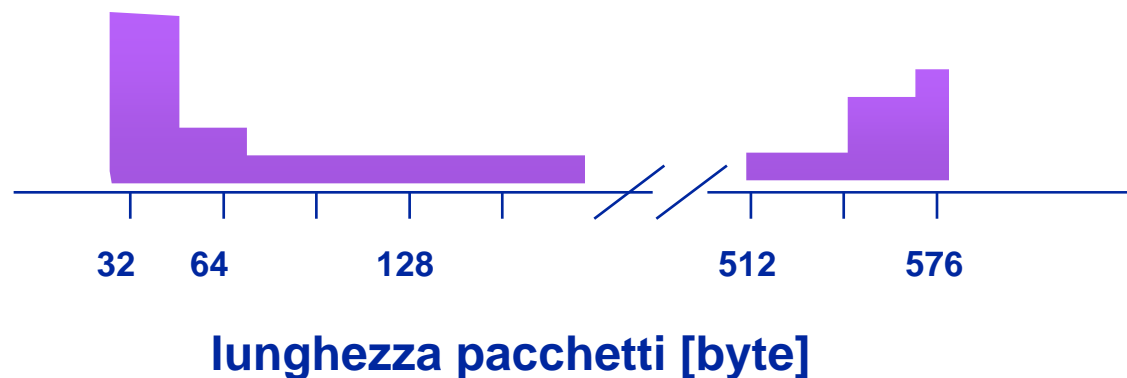
- ◆ **La distribuzione del ritardo del sistema M/M/1 già ricavata è un caso particolare, per $m=1$, della precedente espressione**

Sistema M/G/1/∞

- ◆ **Un caso estremamente interessante nella pratica è quello in cui la distribuzione dei tempi di servizio non è esponenziale negativa**
- ◆ **Questo corrisponde ad avere pacchetti la cui lunghezza non è esponenziale negativa, ma è distribuita in modo arbitrario (Generale)**
- ◆ **Il sistema corrispondente è il M/G, in cui gli arrivi sono ancora esponenziali negativi, ma i servizi no**
- ◆ **Il sistema è piu' complesso del M/M, in quanto non è piu' possibile scrivere una catena di Markov con variabile di stato uguale al numero di utenti nel sistema**
- ◆ **Questo è dovuto al fatto che i servizi non sono piu' "senza memoria"**

Sistema M/G/1/ ∞

- ◆ L'ipotesi sui tempi di interarrivo è soddisfacente in molti casi reali (per esempio, nel caso in cui molti flussi di traffico sono multiplati sullo stesso link).
- ◆ L'ipotesi sulla lunghezza dei pacchetti è frequentemente non valida; per esempio, la lunghezza dei pacchetti su una rete Ethernet ha una distribuzione di questo tipo (evidentemente non è esponenziale negativa)

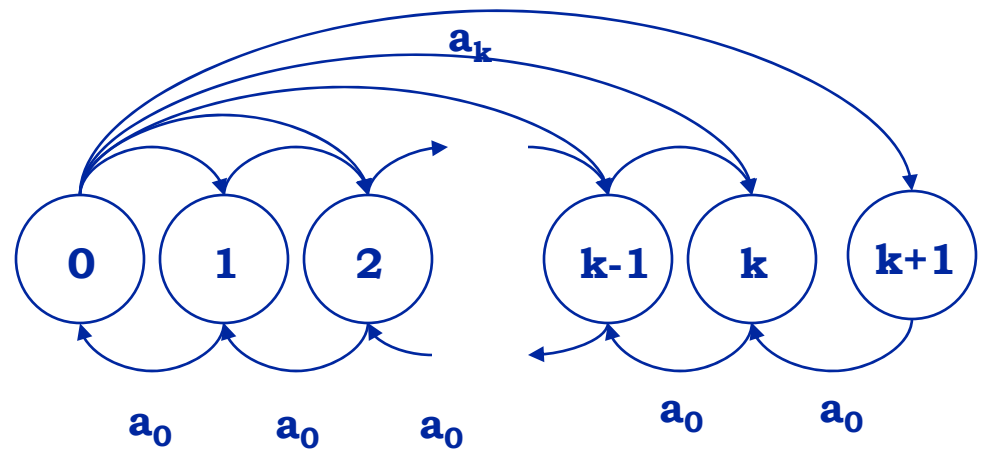


Sistema M/G/1/∞

- ◆ La variabile di stato “numero di utenti nel sistema” non è piu’ utilizzabile per costruire una catena di Markov
- ◆ La variabile “numero di utenti al momento della fine di un servizio” invece può essere utilizzata a questo scopo
- ◆ Il processo indicato è infatti Markoviano; si può capire anche intuitivamente, infatti, che alla fine di un servizio si perde memoria del passato del sistema, ed il numero di utenti rimasti alla fine del servizio è una descrizione

sufficiente per la descrizione dell’evoluzione futura del sistema

- ◆ In figura è rappresentata la catena di Markov segnalata



- ◆ Dove a_i è la probabilità di i nuovi arrivi durante un servizio

Sistema M/G/1/∞

- ◆ Il principale risultato sui sistemi M/G è l'insieme delle formule di Pollaczek-Khinchin (PK-formula), che danno numero medio di utenti e tempi medi di attesa

$$E[W] = \frac{\lambda E[X^2]}{2(1-\rho)}$$

$$E[V] = E[X] + \frac{\lambda E[X^2]}{2(1-\rho)}$$

$$E[N_c] = \frac{\lambda^2 E[X^2]}{2(1-\rho)}$$

$$E[N] = \rho + \frac{\lambda^2 E[X^2]}{2(1-\rho)}$$

- ◆ Quindi, per conoscere i tempi medi di attesa ed il numero medio di utenti nel sistema e nella fila di attesa, è sufficiente conoscere il primo e secondo momento della distribuzione dei tempi di servizio

Sistema M/G/1/∞

- ◆ Per esempio, nel caso di servizio esponenziale negativo il secondo momento (valor quadratico medio) del tempo di servizio è pari a

$$E[X^2] = \int_0^{\infty} t^2 \mu e^{-\mu t} dt = \frac{2}{\mu^2}$$

- ◆ E, andando a sostituire nelle formule precedenti si ottiene

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

- ◆ Come era già noto dall'analisi del sistema M/M/1

- ◆ Per esempio, nel caso di servizio deterministico (M/D/1), in cui il servizio dura un tempo costante pari a $(1/\mu)$, il secondo momento è ovviamente uguale a $(1/\mu)^2$, per cui il ritardo di accodamento è pari a

$$E[W] = \frac{\rho}{2\mu(1-\rho)}$$

- ◆ Cioè, è pari alla metà del tempo di attesa nel caso di servizio esponenziale

Sistema M/G/1/∞

- ◆ La distribuzione deterministica è caratterizzata dall'aver il minimo valore medio del quadrato del tempo di servizio rispetto a tutte le altre possibili distribuzioni
- ◆ Per questo motivo, il sistema M/D/1 ha una prestazione di lower-bound rispetto a tutti i possibili sistemi M/D/1
- ◆ è interessante notare che nel sistema M/D, $E[N_c]$ ed $E[W]$ sono esattamente la metà che nel sistema M/M
- ◆ Inoltre, $E[N]$ ed $E[V]$ sono circa la metà se il carico tende a 1, ma se il carico è leggero tendono ad essere uguali a quelli del sistema M/M
- ◆ Questo è perchè se il carico è leggero la maggior parte dell'attesa nel sistema è nel servizio, mentre a carico alto la maggior parte dell'attesa è in fila

Sistema M/G/1/∞

◆ Le distribuzioni dei tempi di attesa nel sistema M/G/1 non sono ricavabili in forma chiusa nel caso generale

◆ D'altra parte è possibile ricavare un'espressione analitica della trasformata di Laplace della distribuzione di questi tempi, anche se nel caso generale è invertibile solo numericamente

◆ La trasformata di Laplace del tempo di attesa in fila e'

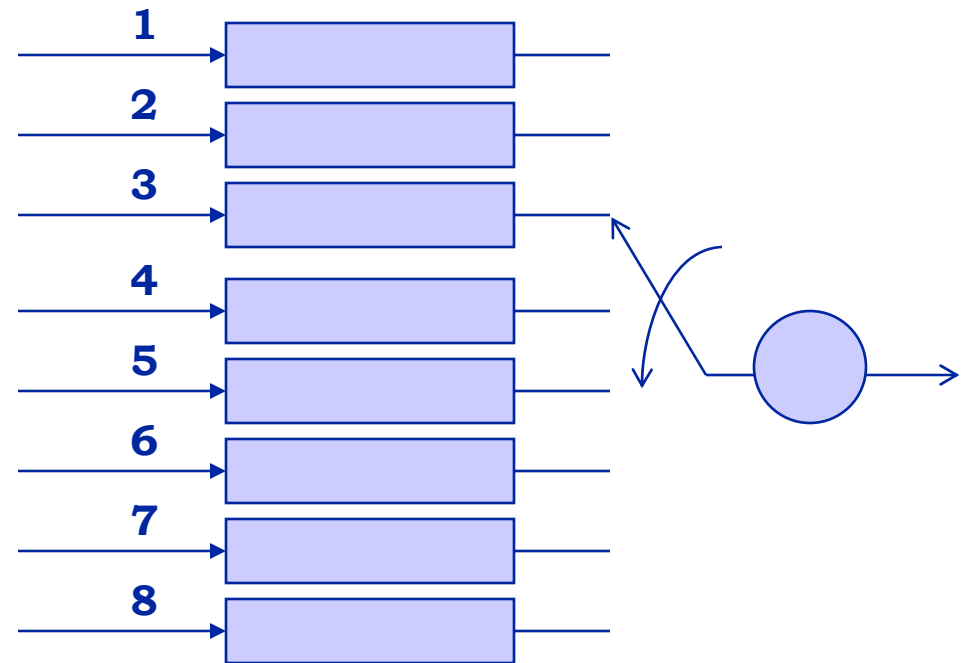
$$W^*(s) = \frac{s(1-\rho)}{s-\lambda + \lambda B^*(s)}$$

◆ E la trasformata di Laplace del tempo di sistema e'

$$V^*(s) = B^*(s) \frac{s(1-\rho)}{s-\lambda + \lambda B^*(s)}$$

Sistema M/G/1/∞ con prioritá'

- ◆ Gli utenti possono essere suddivisi in classi di servizio, ed ad ogni classe di servizio può essere assegnata una diversa prioritá'
- ◆ La classe 1 ha la prioritá piu' alta, la classe 2 è la seconda piu' alta e cosi' via
- ◆ Le prioritá', nel caso piu' semplice, possono essere non interruttive o interruttive
- ◆ Nel caso non interruttivo, si lascia terminare il servizio ad un utente anche se durante il servizio stesso arriva un utente di una classe piu' importante



- ◆ Nel caso interruttivo, un utente nel servizio è interrotto se arriva un utente di una classe piu' importante; il servizio interrotto è ripreso quando non ci sono piu' utenti di prioritá piu' elevata

Sistema M/G/1/∞ con prioritá'

- ◆ Detto λ_k il tasso di arrivo di utenti della classe k
- ◆ Detto $1/\mu_k$ il tempo medio di servizio di un utente di classe k
- ◆ Detto ρ_k il carico associato alla classe k
- ◆ Il tempo di attesa in fila degli utenti della classe di servizio k nel caso di prioritá non interrottiva è dato da

$$E[W_k] = \frac{1}{2} \frac{\sum_{i=1}^n \lambda_i E[X_i^2]}{(1 - \rho_1 - \dots - \rho_{k-1})(1 - \rho_1 - \dots - \rho_k)}$$

- ◆ Ed il tempo di sistema è dato da

$$E[V_k] = \frac{1}{\mu_k} + W_k$$

- ◆ è dunque chiaro che si può dimensionare il ritardo di una classe di servizio scegliendone opportunamente la prioritá'
- ◆ In generale, è bene dare precedenza ai servizi corti

Sistema M/G/1/∞ con priorità'

- ◆ Nel caso di priorità interrottiva, il tempo medio di sistema della classe 1 è dato da

$$E[V_1] = \frac{\frac{1 - \rho_1}{\mu_1} + R_1}{1 - \rho_1}$$

- ◆ Ed il tempo medio di sistema della classe $k > 1$ è dato da

$$E[V_k] = \frac{\frac{1 - \rho_1 - \dots - \rho_k}{\mu_k} + R_k}{(1 - \rho_1 - \dots - \rho_{k-1})(1 - \rho_1 - \dots - \rho_k)}$$

- ◆ Dove si ha

$$R_k = \frac{1}{2} \sum_{i=1}^k \lambda_i E[X_i^2]$$

Distribuzione del ritardo

- ◆ **In prima approssimazione, la probabilità che il ritardo di attesa in coda sia inferiore a w è data da**

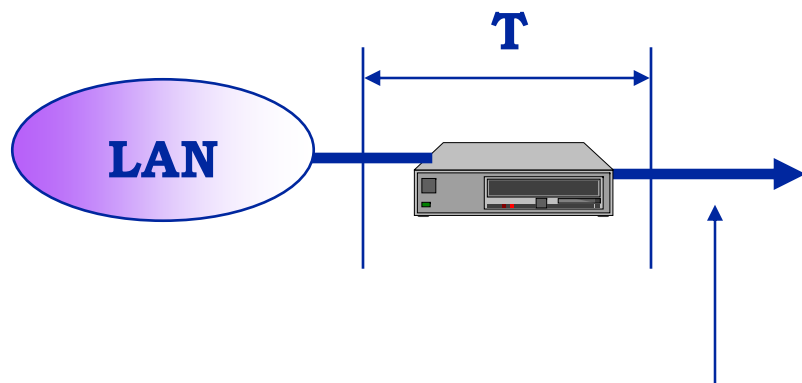
$$P\{W \leq w\} = 1 - \rho e^{-(1-\rho)\mu w}$$

- ◆ **Il vincolo sulla probabilità di non superamento di una soglia di ritardo è generalmente piu' stringente del vincolo sul ritardo medio**

Esempi numerici

Esempio: dimensionamento

- ◆ Si debba connettere una LAN ad una linea geografica. Si prevede un traffico di λ pacchetti/s, con lunghezza media dei pacchetti di L bit. Quale è la velocità di linea necessaria per garantire un ritardo medio di attraversamento del router minore di T secondi?



Linea a capacità C [bit/s]

Si adotta la formula

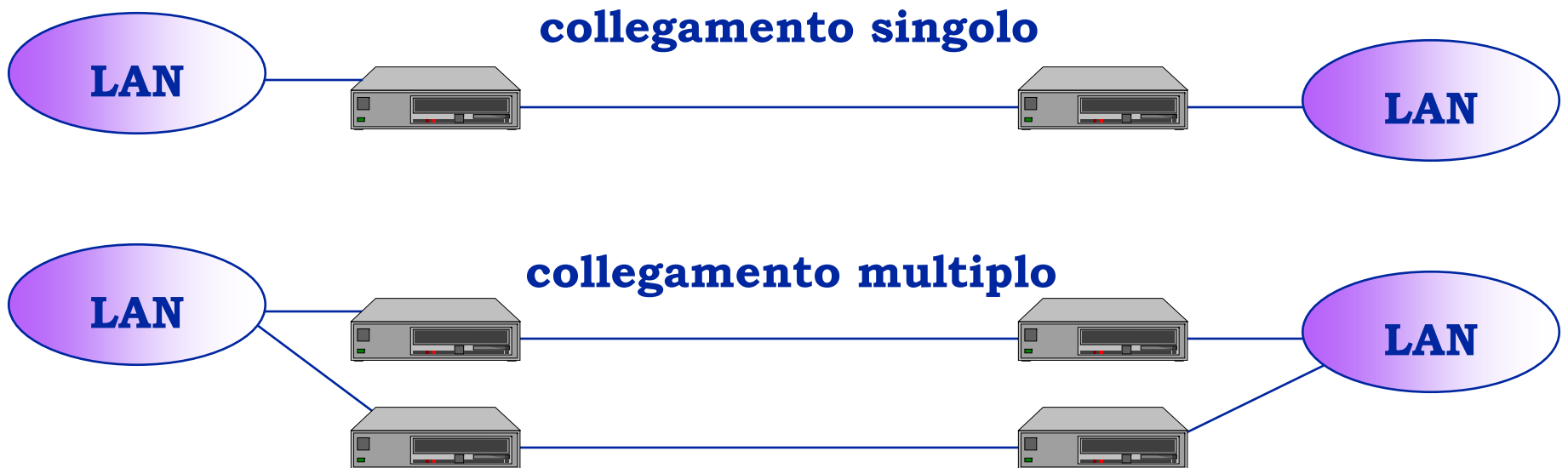
$$C \geq L \left(\lambda + \frac{1}{T} \right)$$

Esempio: dimensionamento

- ◆ Per esempio, se $\lambda = 20$ [pacchetti/s], di lunghezza media pari a 1500 bit e si desidera un ritardo medio $T = 50$ millisecondi, si ottiene $C > 60$ Kbit/s, per cui si deve adottare una linea da 64 Kbit/s (il ritardo effettivo risulta essere pari a 44 millisecondi)
- ◆ Se λ fosse pari a 30 [pacchetti/s], si avrebbe $C > 75000$ [bit/s]. Per realizzare tale collegamento si dovrebbero adottare due linee da 64 Kbit/s. D'altra parte, due linee in parallelo non danno una prestazione uguale ad una linea singola a pari capacità totale.

Esempio: collegamenti multipli

- ◆ Un collegamento punto-punto tra due LAN può essere realizzato in vari modi; una prima distinzione è:

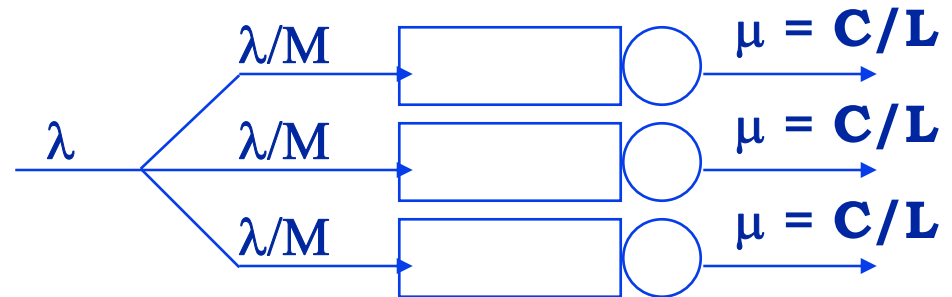


Collegamento singolo e multiplo: modelli

- ◆ I modelli a code dei due sistemi nel caso di link singolo a capacità C [bit/s] e M link paralleli, ciascuno a capacità C [bit/s]



collegamento singolo



collegamento multiplo

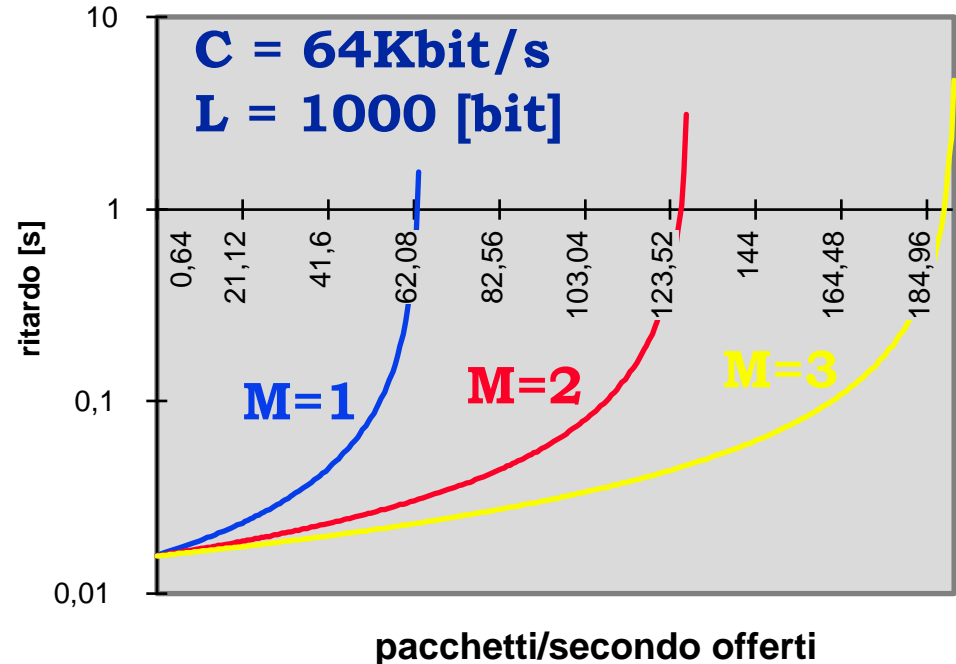
- ◆ Nel caso di collegamento multiplo, si assume che il carico sia equiripartito tra gli M link

Collegamento singolo e multiplo: prestazioni

◆ Collegamento singolo: $T = L / (C - L \lambda)$

◆ Collegamento multiplo (M link): $T = M L / (M C - L \lambda)$

Le prestazioni di M link in parallelo sono migliori, in quanto la capacità totale è pari a M volte la capacità del link singolo



Ottimizzazione di un collegamento punto-punto: ulteriore alternativa

- ◆ Il collegamento multiplo può essere realizzato in più modi:



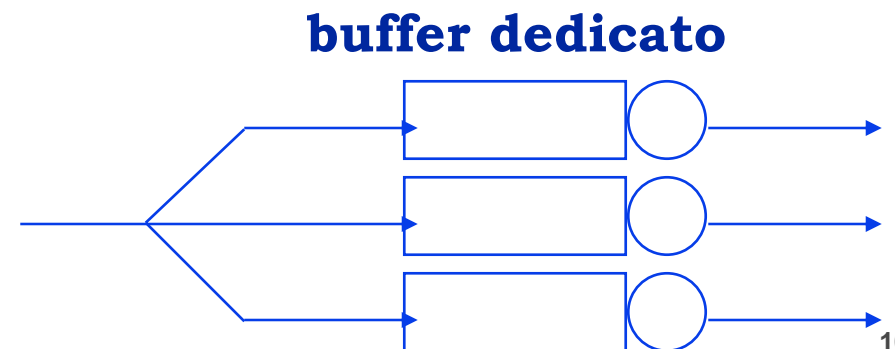
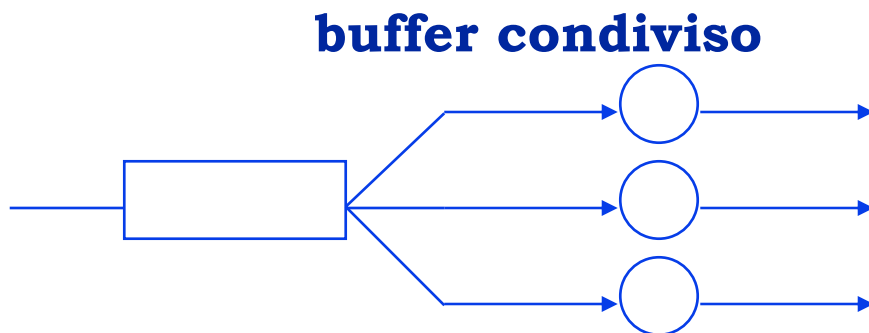
La seconda soluzione può essere molto più efficiente

Modelli nel caso di accodamento multiplo

- ◆ Nel caso di pool di linee in parallelo, la prestazione del sistema dipende dalla politica di accodamento adottata nel dispositivo di interconnessione

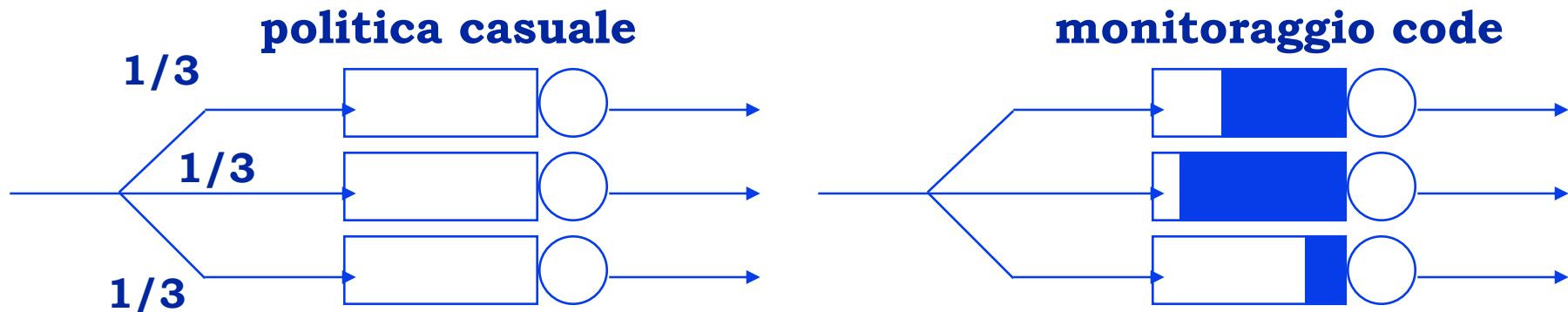


- ◆ Si può predisporre un unico buffer condiviso, oppure un buffer dedicato ad ogni linea



Buffer dedicato ad ogni linea

- ◆ In questo caso, le linee possono essere gestite secondo le politiche
 - *casuale*, secondo cui si sceglie a caso la linea su cui trasmettere un pacchetto.
 - *monitoraggio code*, secondo cui si trasmette un pacchetto sulla linea più scarica. Dal punto di vista prestazionale, il sistema a monitoraggio code è equivalente al sistema a buffer condiviso



Buffer dedicato, politica casuale: effetto del numero di linee

- ◆ **La prestazione è analoga a quella che si ottiene nel caso di coppie di dispositivi dedicati ad ogni link, cioè:**

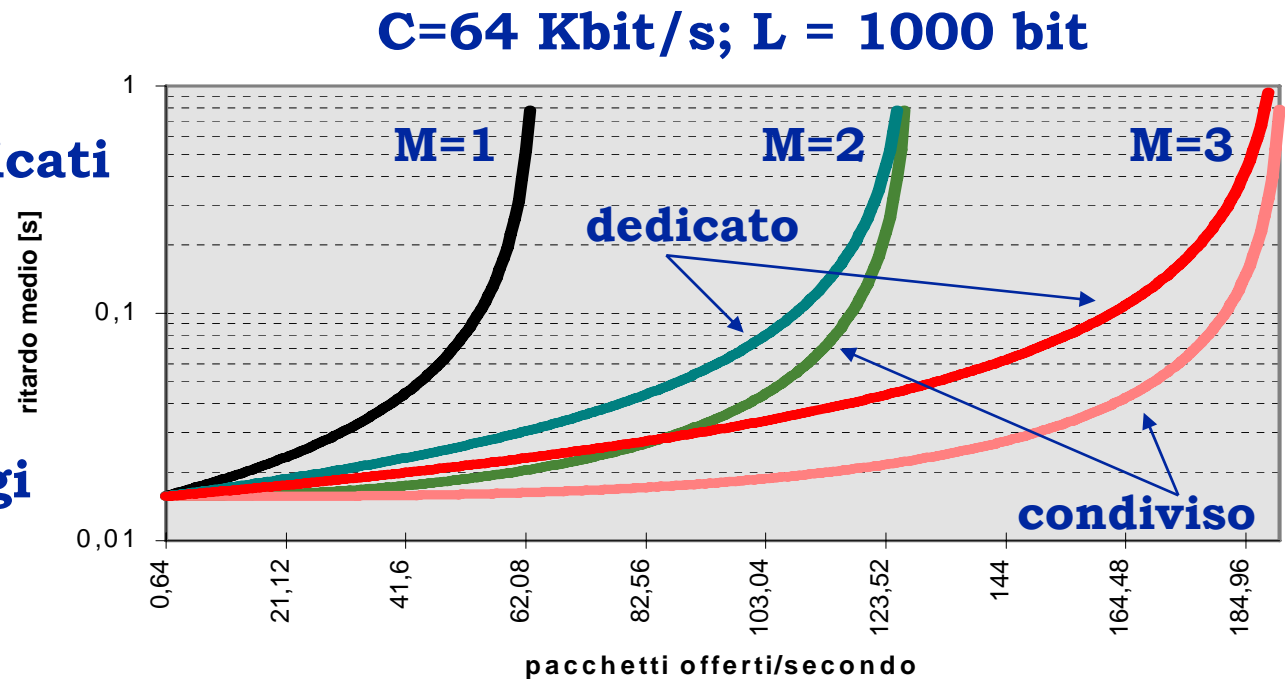
$$T = ML / (MC - \lambda L)$$

- ◆ **Questo accade in quanto a causa della politica casuale le linee in pool risultano essere praticamente indipendenti**

Buffer condiviso/Buffer dedicato con monitoraggio code

- ◆ La formula che dà la curva di prestazione è molto complessa

La scelta di buffer dedicati è penalizzante
La gestione delle linee di uscita deve essere intelligente per conseguire vantaggi

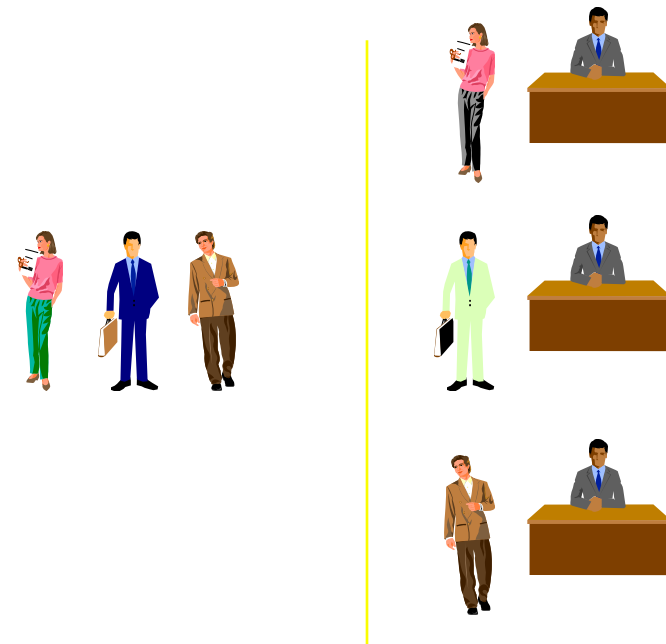


Buffer dedicato/condiviso, interpretazione fisica

buffer dedicati



buffer condiviso

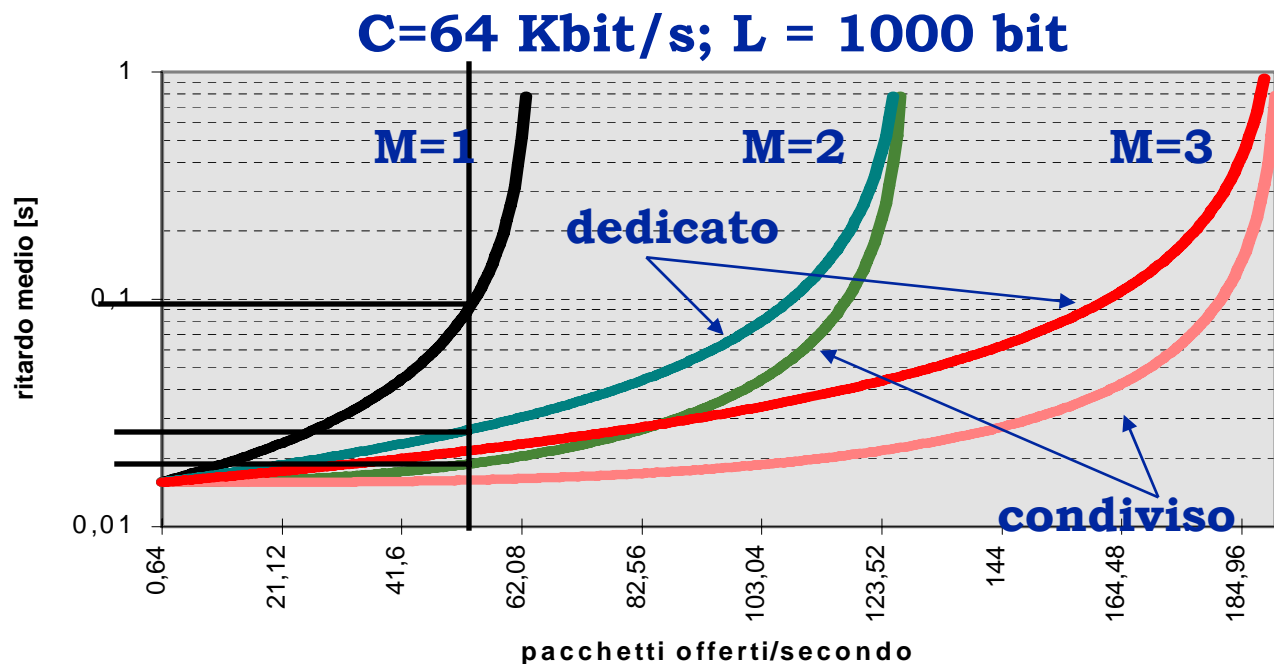


Per esempio, in alcune banche la fila di attesa è unica (condivisa) e l'utente in testa alla coda si reca dal primo servente che si libera. Questa strategia elimina il problema (presente nei buffer dedicati) dell'utente sfortunato che si trova davanti un cliente che ha bisogno di un servizio particolarmente lungo

Esempio di dimensionamento

- ◆ Si debbano instradare $\lambda=50$ [pacchetti/s] di lunghezza media pari a 1000 bit verso una entità remota. Per ottenere un ritardo medio inferiore a 50 millisecondi si deve avere una capacità C data da

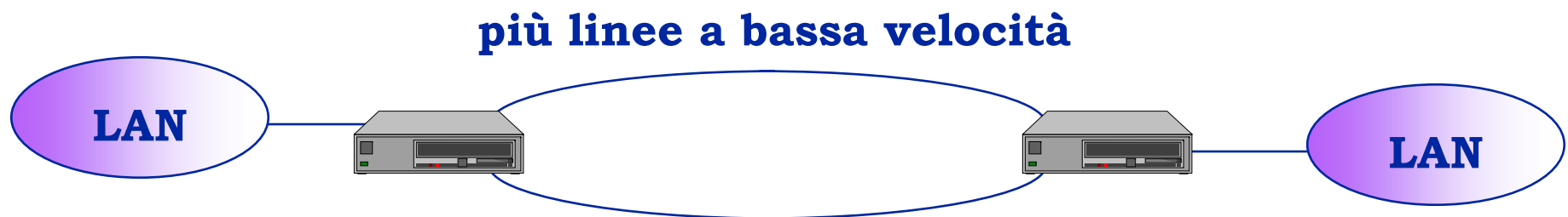
$$C \geq L \left(\lambda + \frac{1}{T} \right) = 70000 [\text{bit} / \text{s}]$$



Con due linee da 64 K si ottiene la prestazione desiderata. La soluzione a buffer dedicato e' peggiore di quella a buffer condiviso di circa 30%

Ottimizzazione collegamento punto-punto: linee ad alta velocità

- ◆ Si può aumentare la capacità totale del collegamento utilizzando una linea a capacità elevata, piuttosto che una molteplicità di linee a capacità minore
- ◆ Le prestazioni non sono le stesse

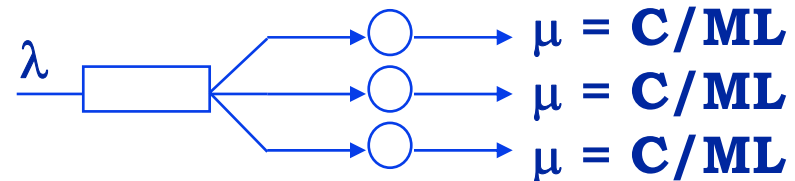


Collegamento punto-punto con linee ad alta velocità: modelli

linea ad alta velocità



più linee a bassa velocità,
buffer condiviso

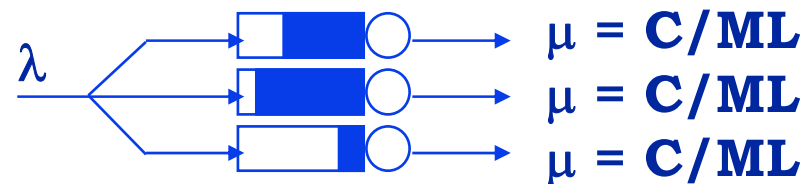


Nel caso di linea singola ad alta velocità, la capacità della linea è C [bit/s]

Nel caso di M linee multiple, ogni linea ha capacità pari a C/m

La capacità totale è la stessa

più linee a bassa velocità,
monitoraggio code



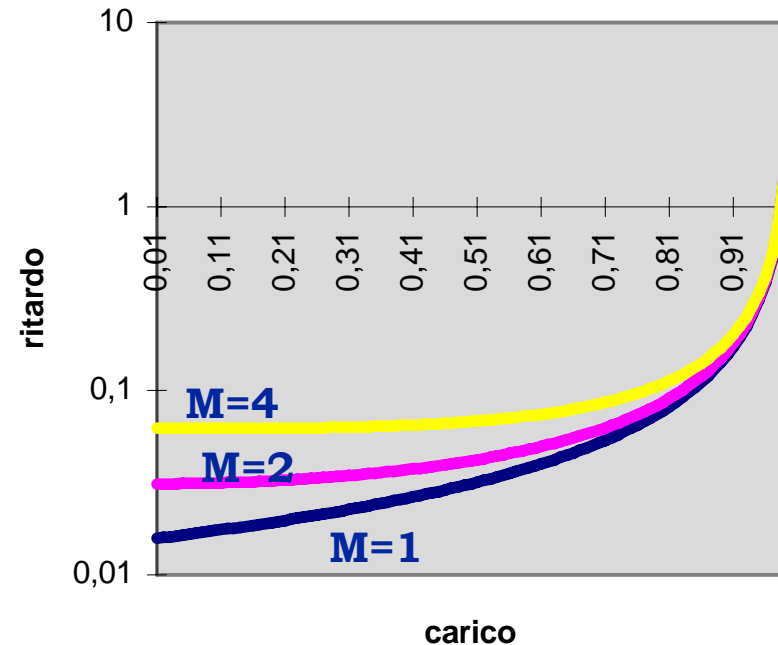
Confronto a parità di capacità totale

- ◆ **A parità di capacità totale, conviene adottare una singola linea piuttosto che molteplici linee a capacità più bassa**

La convenienza è notevole a basso carico, ad alto carico le prestazioni sono simili

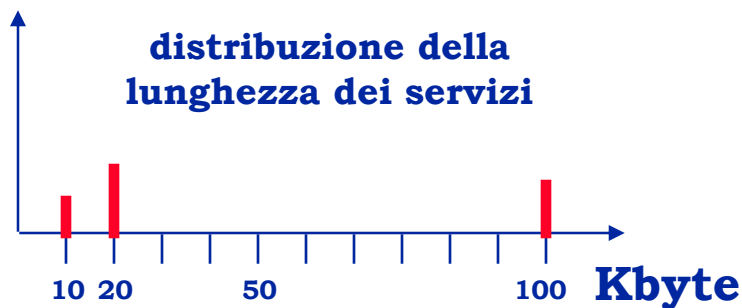
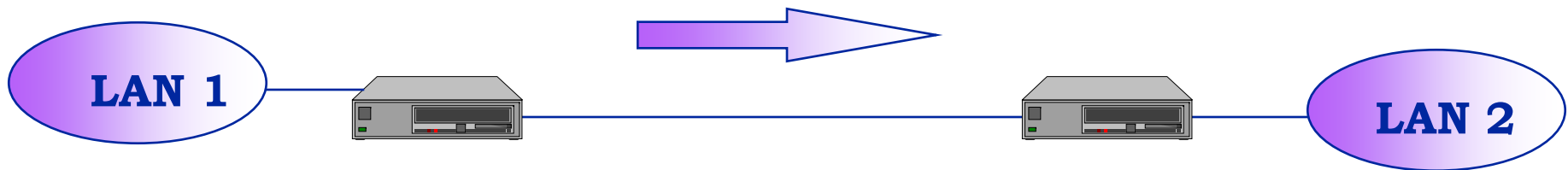
A basso carico, domina il ritardo di trasmissione, che è più elevato nel caso di molte linee in parallelo

C totale = 64 Kbit/s, L = 1000 bit



Distribuzione generale dei servizi

- ◆ Per tenere conto di richieste di servizio con distribuzione non esponenziale negativa sono necessari modelli più evoluti
- ◆ Per esempio, si consideri il caso in cui tra la LAN 1 e la LAN 2 si debbano trasmettere mediamente ogni minuto 2 file di 10 Kbyte, 4 da 20 Kbyte e 3 da 100 Kbyte



In questo caso la distribuzione della lunghezza dei servizi è molto diversa da una esponenziale negativa

Distribuzione generale dei servizi

- ◆ Il totale traffico offerto è pari a $(2 \times 10 + 4 \times 20 + 3 \times 100) \times 8000 / 60 = 53333$ bit/s, per cui si può adottare, come primo tentativo, una linea a 64 Kbit/s
- ◆ Il tasso di richiesta di servizi/s è $\lambda = (2 + 4 + 3) / 60 = 0,15$
- ◆ La lunghezza media in bit dei servizi è pari a $(2 \times 10 + 4 \times 20 + 3 \times 100) \times 8000 / 9 = 355555$ bit
- ◆ Se utilizzassimo il modello esponenziale, si avrebbe $\mu = 64000 / 355555 = 0,18$ servizi/s, per cui

$$T = 1 / (\mu - \lambda) = 33 \text{ [s]}$$



Distribuzione generale dei servizi

- ◆ Per tenere conto correttamente della distribuzione dei servizi, si deve considerare sia la durata media che la varianza degli stessi

dimensione servizio [bit]	durata [s]	frequenza relativa	durata quadratica [s ²]
80000	1,25	2/9 = 0,22	1,5625
160000	2,5	4/9 = 0,44	6,25
800000	12,5	3/9 = 0,33	156,25

Durata media servizi: $1,25 \times 0,22 + 2,5 \times 0,44 + 12,5 \times 0,33 = 5,5$ [s]

Durata quadratica media servizi: $1,5625 \times 0,22 + 6,25 \times 0,44 + 156,25 \times 0,33 = 54,66$ [s²]

Varianza della durata dei servizi: $\sigma^2 = 54,66 - 5,5 \times 5,5 = 24,41$ [s²]

Distribuzione generale dei servizi

- ◆ **A questo punto, dato il carico $\rho = 53333/64000 = 0,833$, dato il tasso medio di richiesta di servizi $\lambda = 0,15$ [servizi/s] e data la varianza della durata dei servizi $\sigma^2 = 24,41$ [s²] si può calcolare il tempo medio di attraversamento del sistema come**

$$T = \frac{\rho}{\lambda} + \frac{1}{\lambda} \frac{\rho^2 + \lambda^2 \sigma^2}{2(1-\rho)} = 286 \text{ [s]}$$

La differenza è notevole: il modello elementare dà un ritardo pari a 33 [s].

Esempio I

- ◆ In questo esempio, il modello semplice comporta un errore significativo, ma minore del caso precedente

Dlm. file [byte]	Frequenza [file/s]	Frequenza relativa	Tempo TX. [s]	Ritardo effettivo [s]
32	1,05	0,49	0,004	3,182953
128	0,11	0,05	0,016	Ritardo modello base [s]
512	0,53	0,25	0,064	0,664409
1024	0,32	0,15	0,128	
25000	0,11	0,05	3,125	
65000	0,02	0,01	8,125	
Capacita' link [bit/s]	64000	E[TX]	0,27434112	
Carico:	0,58709	E[TX * TX]	1,12242522	VAR[TX] 1,047162169

Esempio II

◆ In questo esempio, il modello base si comporta meglio

Dlm. file [byte]	Frequenza [file/s]	Frequenza relativa	Tempo TX. [s]	Ritardo effettivo [s]
32	10,00	0,33	0,004	0,430328
128	10,00	0,33	0,016	Ritardo modello base [s]
512	10,00	0,33	0,064	0,19093
1024	0,00	0,00	0,128	
25000	0,00	0,00	3,125	
65000	0,00	0,00	8,125	
Capacita' link [bit/s]	64000	$E[TX]$	0,02837643	
Carico:	0,851378	$E[TX * TX]$	0,00398219	$VAR[TX]$ 0,003176968

Esempio III

- ◆ In questo esempio (studiato appositamente) la differenza è drammatica

Dlm. file [byte]	Frequenza [file/s]	Frequenza relativa	Tempo TX. [s]	Ritardo effettivo [s]
32	80,00	1,00	0,004	10,5088
128	0,00	0,00	0,016	Ritardo modello base [s]
512	0,00	0,00	0,064	0,053536
1024	0,00	0,00	0,128	
25000	0,00	0,00	3,125	
65000	0,06	0,00	8,125	
Capacita' link [bit/s]	64000	$E[TX]$	0,01012731	
Carico:	0,810833	$E[TX * TX]$	0,04961036	$VAR[TX]$ 0,049507798

Esempio IV

◆ Il modello base non è sempre ottimista

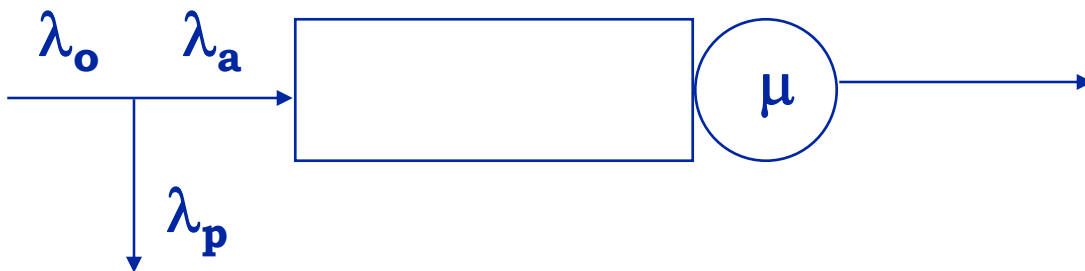
Dlm. file [byte]	Frequenza [file/s]	Frequenza relativa	Tempo TX. [s]	Ritardo effettivo [s]	
32	200,00	1,00	0,004	0,012	
128	0,00	0,00	0,016	Ritardo modello base [s]	
512	0,00	0,00	0,064	0,02	
1024	0,00	0,00	0,128		
25000	0,00	0,00	3,125		
65000	0,00	0,00	8,125		
Capacita' link [bit/s]	64000	$E[TX]$	0,004		
Carico:	0,800000001	$E[TX * TX]$	1,6E-05	$VAR[TX]$	3,78552E-11

Il problema della perdita di pacchetti

◆ Nei dispositivi reali la dimensione dei buffer è finita; questo può comportare perdita di pacchetti quando si esaurisce la memoria disponibile per accodare i pacchetti in attesa

◆ Parametri:

- λ_o : pacchetti/s offerti
- λ_p : pacchetti/s persi
- λ_a : pacchetti/s accettati



$$\lambda_o = \lambda_a + \lambda_p$$

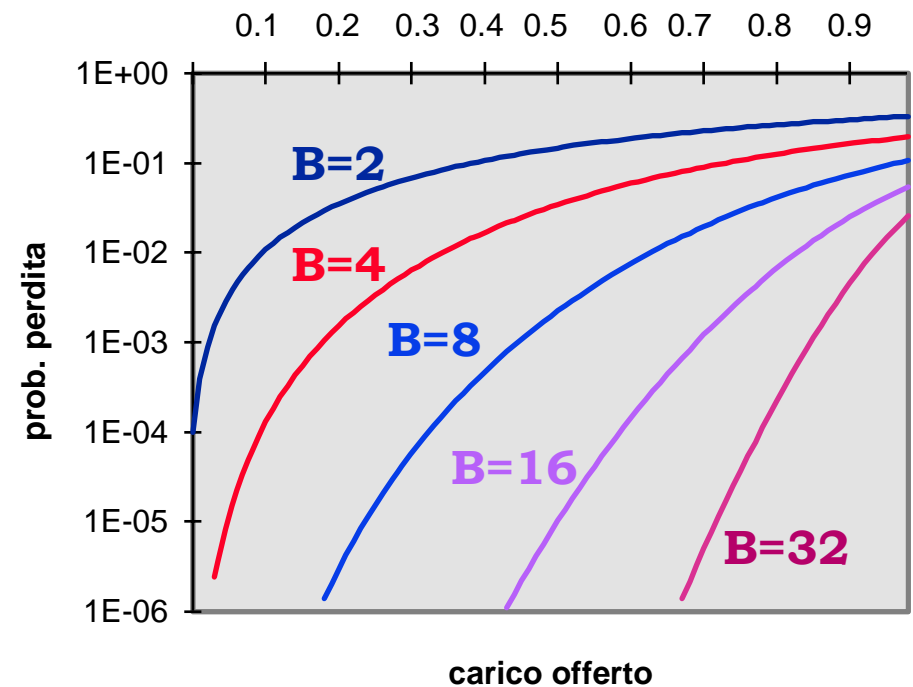
Il problema della perdita di pacchetti

- ◆ Nel caso in cui la perdita di pacchetti non sia limitata o addirittura eliminata tramite tecniche di controllo di flusso, si può in prima approssimazione assumere la seguente valutazione quantitativa della probabilità di perdita (B è il numero di pacchetti che il buffer può memorizzare)

$$\Pi = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{B+1}} \left(\frac{\lambda}{\mu}\right)^B$$

Il problema della perdita di pacchetti

- ◆ La probabilità di perdita diminuisce all'aumentare della dimensione del buffer
- ◆ La valutazione è ottimistica nel caso in cui gli arrivi dei pacchetti sono fortemente correlati (traffico a burst)



Utilizzo di linee commutate

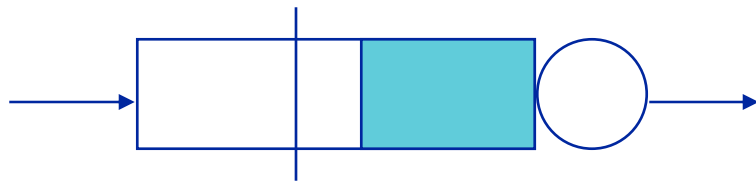
- ◆ Per migliorare le prestazioni, è possibile aggiungere al collegamento dedicato un collegamento alternativo commutato, da utilizzare nei momenti di congestione
- ◆ Le prestazioni migliorano; la contropartita è il costo della linea commutata, proporzionale al tempo di collegamento



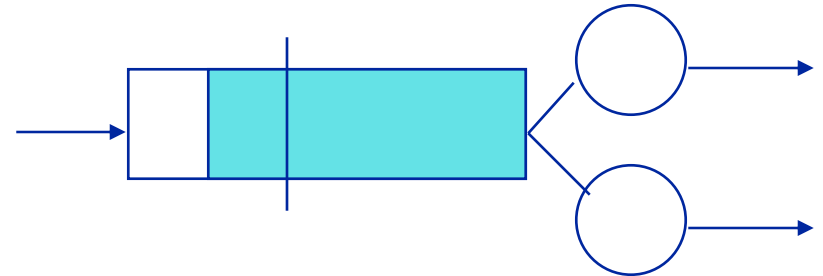
Utilizzo di linee commutate

- ◆ **Un possibile metodo per l'attivazione della linea commutata è il monitoraggio della coda associata: se l'occupazione della coda supera una soglia prefissata, si attiva la linea commutata**

Solo linea dedicata

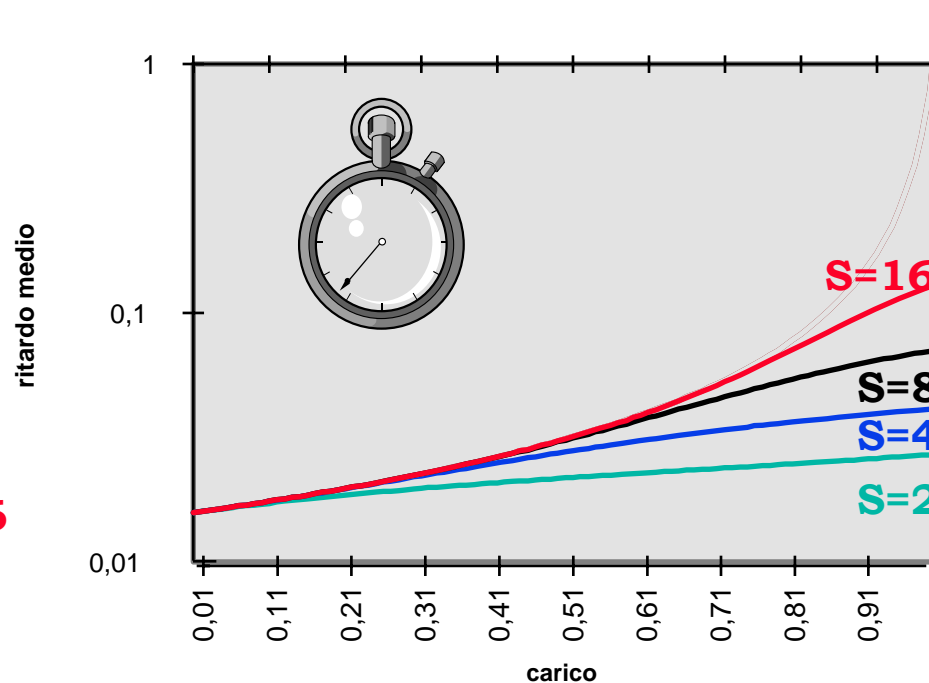
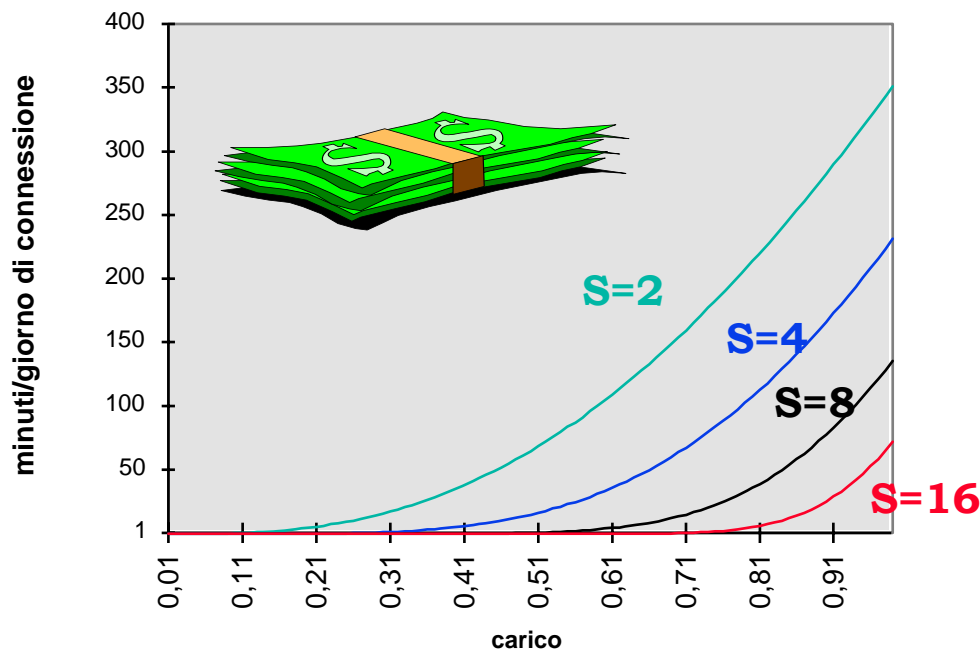


Linea dedicata e commutata



Utilizzo di linee commutate

- ◆ Se la soglia è piccola, le prestazioni sono notevolmente migliori, ma aumenta il tempo medio di connessione



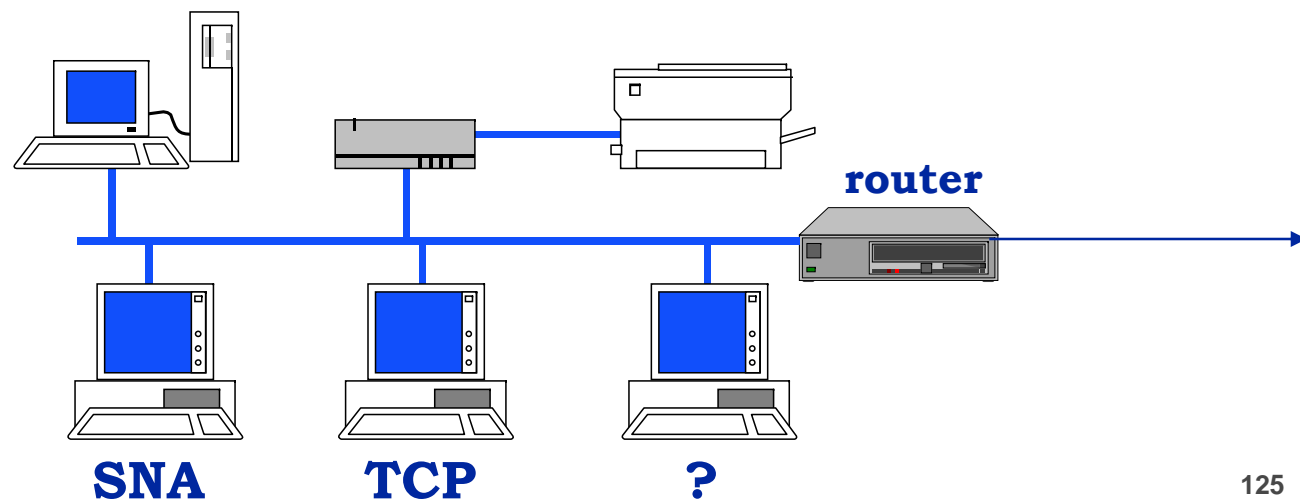
Utilizzo di linee commutate

- ◆ **Per gestire situazioni sistematiche di carico elevato la linea commutata non è la miglior scelta, a causa dei costi elevati**
- ◆ **La linea commutata va utilizzata per gestire variazioni transitorie del carico; per questo motivo la soglia di intervento deve essere alta**
- ◆ **Una soglia molto bassa è equivalente all'adozione di una linea dedicata (molto costosa)**

Sistemi multi-protocollo

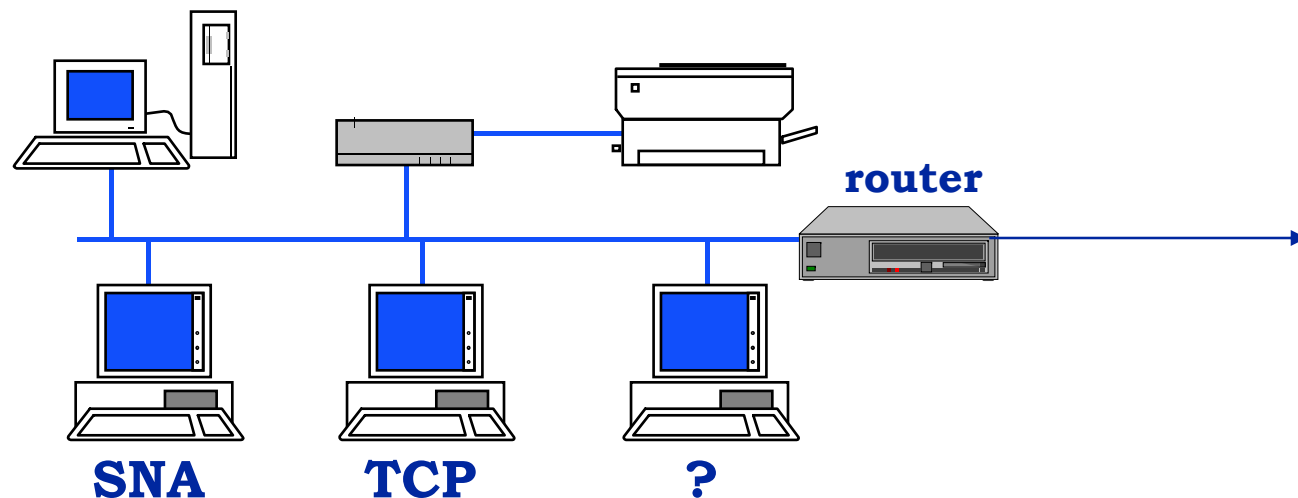
Sistemi multi-protocollo (I)

- ◆ In un sistema multi-protocollo esistono diversi tipi di pacchetti, con diverse distribuzioni della lunghezza e requisiti sul ritardo
- ◆ Per esempio, in SNA è fondamentale rispettare un rigido requisito sul ritardo
- ◆ Una tipica configurazione è quella di una LAN che ospita diversi protocolli



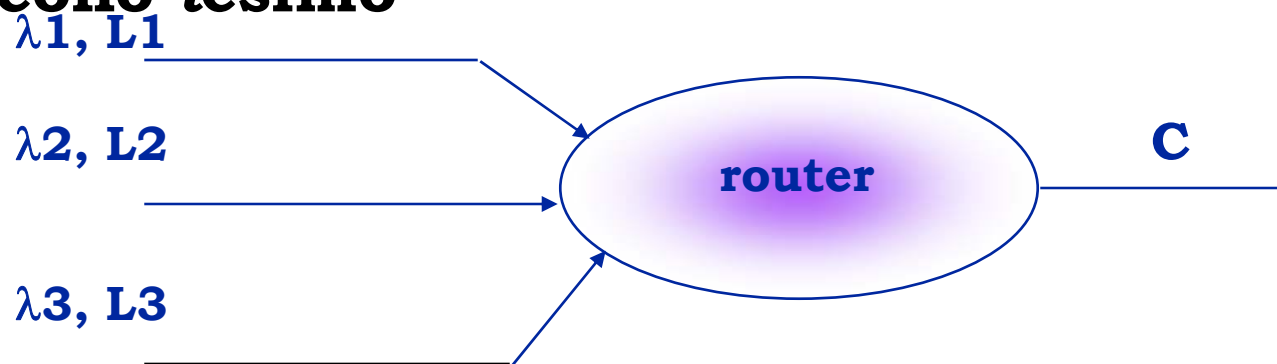
Sistemi multi-protocollo (II)

- ◆ Le prestazioni dei vari protocolli sono interdipendenti, in quanto attraversano lo stesso dispositivo di interconnessione remota
- ◆ Una possibile soluzione per migliorare le prestazioni è assegnare priorità diverse ai vari protocolli, in modo da favorire quelli piu' sensibili ai ritardi



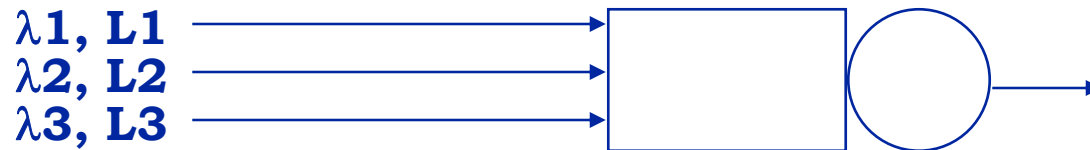
Sistemi multi-protocollo: parametri

- ◆ λ_i : frequenza media degli arrivi associata al protocollo *i*-esimo
- ◆ L_i : lunghezza media dei pacchetti del protocollo *i*-esimo
- ◆ Λ : frequenza media totale degli arrivi ($\Lambda = \lambda_1 + \lambda_2 + \dots$)
- ◆ C : capacità della linea di uscita
- ◆ $1/\mu_i$: tempo medio per la trasmissione di un pacchetto del protocollo *i*-esimo



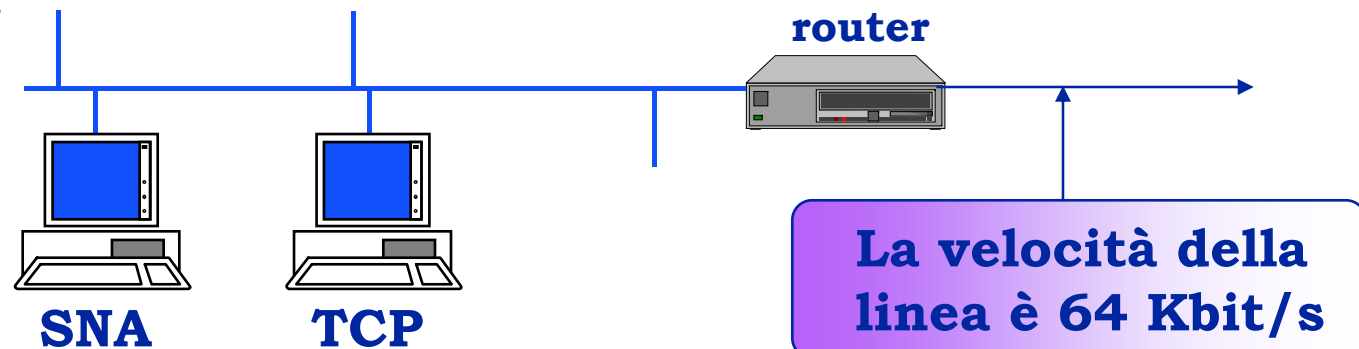
Sistemi multi-protocollo: trattamento senza priorità

- ◆ La capacità μ del servente è data da $\mu = (\lambda_1\mu_1 + \lambda_2\mu_2 + \dots)/\Lambda$
- ◆ Il carico del servente è $\rho = \Lambda/\mu$
- ◆ Il tempo medio di attesa in coda è dato (questa è una approssimazione) da $W = \rho/(\mu - \Lambda)$



Sistemi multi-protocollo con trattamento senza priorità: esempio (I)

- ◆ Si ha un traffico di tipo SNA di intensità media pari a 15 [pacchetti/s], e la lunghezza media dei pacchetti è pari a 2000 [bit]. Inoltre si ha un traffico di tipo TCP con intensità media pari a 20 [pacchetti/s], la lunghezza media dei pacchetti è 1500 [bit].
- ◆ Si vuole determinare il ritardo sperimentato dai pacchetti nel caso in cui il router li tratti indistintamente.



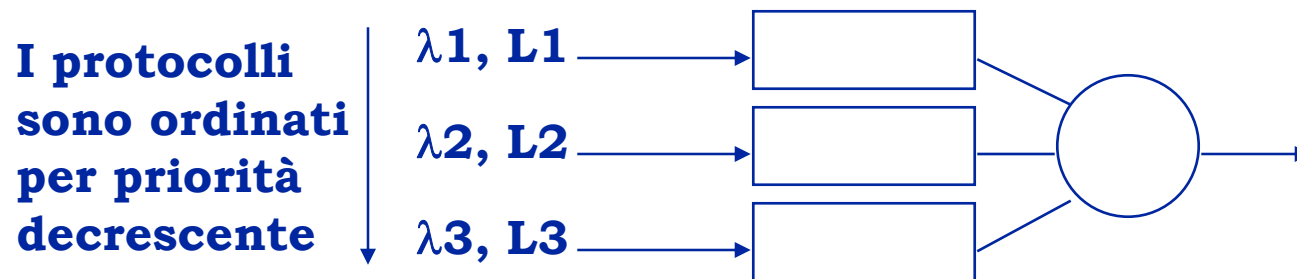
Sistemi multi-protocollo con trattamento senza priorità: esempio (II)

- ◆ Il tempo medio di trasmissione di un pacchetto SNA è pari a $1/\mu_1 = 2000/64000 = 0,03125$ [s]; il tempo medio di trasmissione di un pacchetto TCP è pari a $1/\mu_2 = 1500/64000 = 0,02344$. Il tempo medio generico di trasmissione di un pacchetto è pari a $1/\mu = \Lambda/(\lambda_1\mu_1 + \lambda_2\mu_2) = 0,02625$ [s].
- ◆ Il carico del router è pari a $\rho = \Lambda/\mu = 0,9188$.

Il ritardo di attesa in coda è dunque pari a
 $W = \rho/(\mu - \Lambda) = 0,2973$ [s]

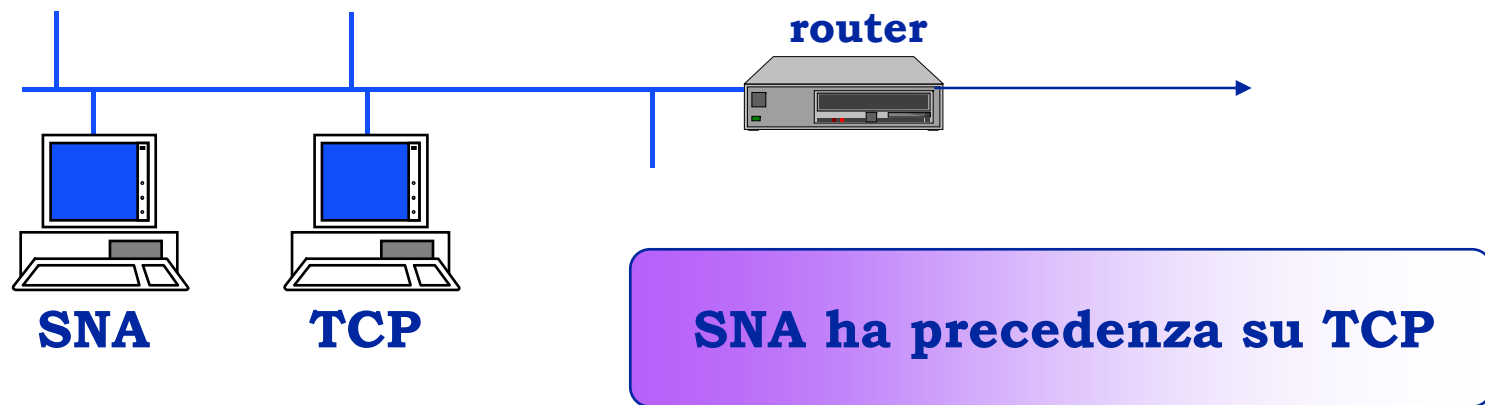
Sistemi multi-protocollo: trattamento con priorità

- ◆ Un pacchetto della classe i è servito solo se non ci sono in attesa pacchetti di classi inferiori
- ◆ Il carico della classe i è dato da $\rho_i = \lambda_i / \mu_i$
- ◆ Il tempo di attesa della classe i è dato da
 - $W_i = R / [(1 - \rho_1 - \dots - \rho_{i-1})(1 - \rho_1 - \dots - \rho_i)]$
 - in cui $R = \lambda_1 / \mu_1^2 + \lambda_2 / \mu_2^2 + \dots$



Sistemi multi-protocollo e trattamento con priorità: esempio (I)

- ◆ Nel caso dell'esempio precedente (SNA e TCP), si dia priorità ai pacchetti di SNA (SNA: protocollo 1, TCP: protocollo 2)
- ◆ I carichi dei due protocolli sono $\rho_1 = \lambda_1 / \mu_1 = 0,4688$ e $\rho_2 = \lambda_2 / \mu_2 = 0,4689$ rispettivamente.



Sistemi multi-protocollo e trattamento con priorità: esempio (II)

- ◆ In questo caso, $R = \lambda_1/\mu_1^2 + \lambda_2/\mu_2^2 = 0,02564$
- ◆ Il tempo di attesa in coda dei pacchetti SNA è pari a $W1 = R/(1-\rho_1) = 0,04828$ [s]
- ◆ Il tempo di attesa in coda dei pacchetti TCP è pari a $W1 = R/(1-\rho_1)(1-\rho_1-\rho_2) = 0,7735$ [s]

	Priorita': NO	Priorita': SI
SNA	0,2973	0,04828
TCP	0,2973	0,7735

Sistemi multi-protocollo e trattamento con priorità: esempio (III)

- ◆ In conclusione, il trattamento senza priorità potrebbe causare ritardi inaccettabili per certi protocolli come, per esempio, SNA
- ◆ Il trattamento con priorità può migliorare notevolmente le prestazioni dei protocolli a priorità più elevata
- ◆ Questo vantaggio è pagato con una peggiore prestazione dei protocolli meno critici

	Priorita': NO	Priorita': SI
SNA	0,2973	0,04828
TCP	0,2973	0,7735

Trattamento con priorità: esempio

- ◆ Si può facilmente calcolare il ritardo medio sperimentato da ogni protocollo con semplici calcoli

	Frequenza: [arrivi/s]	Lunghezza media pacchetti [bit]	1/mi	Carico	ritardo medio [s]
Protocollo 1	15,00	2000	0,03125	0,46875	0,04924977
Protocollo 2	20,00	1500	0,0234375	0,46875	0,787996324
Protocollo 3	3,00	850	0,01328125	0,039844	18,47715518
	Capacita' linea in [bit/s]	64000			
	R:	0,02616394			

I protocolli sono ordinati per priorit  decrescenti. Immettere per ogni protocollo la frequenza media di arrivo dei pacchetti [arrivi/s] e la lunghezza media dei pacchetti. Immettere inoltre la velocit  della linea in [bit/s].

Non immettere dati nelle celle grigie. Nella colonna di destra si ottengono i ritardi medi di accodamento in [s].

Esercizi

Esercizio 1

- ◆ **Un sistema d'attesa utilizza una coda con 1 posto e 2 serventi. Gli arrivi sono di Poisson con tasso $\lambda=2$ pac/s, i servizio hanno durata exp. neg. con media $1/\mu_a=0,5$ s per il primo servente $1/\mu_b= 1$ s per il secondo. Quando entrambi i serventi sono vuoti si utilizza il servente piu' veloce ($\mu_a > \mu_b$). Si calcoli**
- a) Il ritardo medio**
 - b) la probabilità di perdita**
 - c) il fattore di utilizzo dei due serventi**
 - d) il traffico smaltito dal servente a e dal servente b**

Esercizio 2

- ◆ Si consideri un sistema di trasmissione con capacità di memorizzazione pari a 2 pacchetti, che utilizza una linea (linea A) con velocità 9,6 Kb/s. Il traffico offerto è di Poisson con valore medio $\lambda=20$ pacc/s e lunghezza dei pacchetti v.c. esponenziale con valor medio $1/\mu=480$ bit. Quando ci sono più di 2 pacchetti nel sistema si attiva un'ulteriore linea (linea B) con velocità 4,8 Kb/s. Essa viene utilizzata fino a quando, al completamento di una trasmissione sulla linea B, non si trovino meno di 3 pacchetti nel sistema. Si calcoli
- Il traffico smaltito
 - l'utilizzo della linea A
 - Il traffico smaltito dalla linea B
 - l'occupazione media del buffer
 - il ritardo medio di attraversamento del sistema

Esercizio 3

- ◆ **Si consideri un sistema di trasmissione con capacità di memorizzazione pari a 3 pacchetti, che utilizza una linea con velocità $\mu=60$. Il traffico offerto è generato da 5 utenti che generano traffic di Poisson con valore medio $\gamma=10$ pacc/s se nessun loro pacchetto e' nel sistema e non generano traffico altrimenti.**
 - a) Si disegni la catena che descrive il sistema e la si risolva.**
 - b) Si calcoli il traffico medio offerto**
 - c) Il tempo medio nel sistema**
 - d) La probabilità di blocco**
 - e) La probabilità di rifiuto**