Reti di Telecomunicazioni 1

Corso "on-line" - AA2006/07
"Sistemi a coda" Blocco E1 v3

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

- **Risorse e sistemi di servizio**

Obiettivi

- Acquisire alcuni concetti di base relativi alle prestazioni di una rete di telecomunicazioni:
  - » efficienza
  - » utilizzazione
  - » ritardo
  - » perdita
- Essere in grado di valutare quantitativamente tali parametri in alcuni semplici casi (esercizi per l'esame…)

Iniziamo questa parte dedicata alle prestazioni delle reti di telecomunicazioni. L'obiettivo è quello di vedere quali sono i parametri prestazionali (es. "utilizzazione", "ritardo", "perdita") di una rete e come si possono valutare utilizzando dei "modelli".

Risorse di una Rete TLC

E' una *risorsa* tutto ciò che in senso logico permette l'attuarsi del servizio di trasferimento

- ogni singolo flusso informativo necessita di una certa quantità di *risorse* per essere adeguatamente (...) trasferito;
- le *risorse* hanno un costo, quindi sono disponibili in misura limitata;

Principali tipi di Risorse:

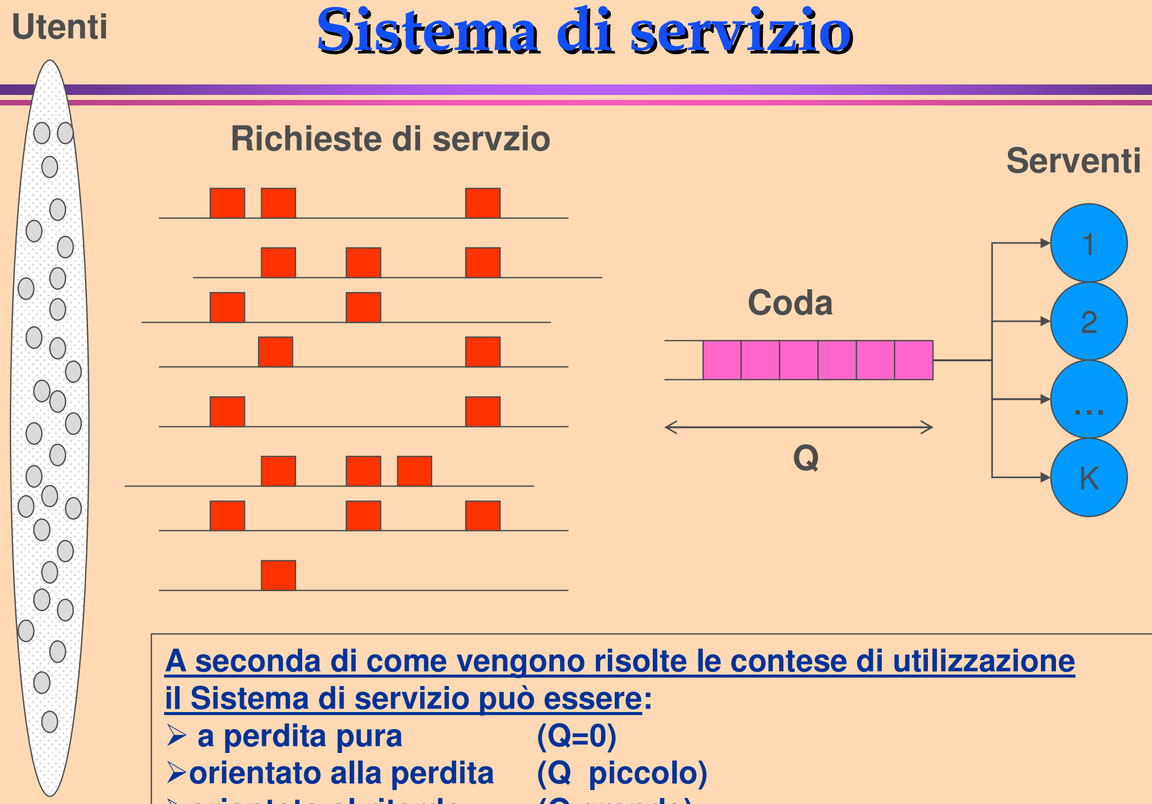
- **R. Trasmissive**:
  logicamente localizzate nei rami → Capacità Trasmissiva (*Banda*)
  (costo: dei cavi o delle frequenze)
- **R. di Memorizzazione**
  logicamente localizzate nei nodi → Quantità di Memoria (*Buffer*)
  (costo: delle memorie veloci)
- **R. di Elaborazione**
  logicamente localizzate nei nodi e in altri elementi di rete (*servers*)
  (costo: dei processori)

Abbiamo già introdotto in una lezione precedente il concetto di "risorsa" e visto come per realizzare un processo di telecomunicazioni una serie di "attività" hanno bisogno di queste risorse. Nel caso in cui le risorse siano condivise, le attività competono per le risorse e si può verificare il caso che una attività non trovi le risorse disponibili nel momento in cui gli servono.

L'interazione tra le attività e le risorse in una rete di telecomunicazioni può essere descritta da particolari modelli matematici chiamati "sistemi di servizio" o "sistemi a coda".

I sistemi a coda, di cui nella slide successiva vediamo una rappresentazione schematica, sono in realtà in grado di modellare anche altri fenomeni oltre a quelli relativi ai processi di comunicazione e trovano quindi la loro applicazione anche in altre discipline (es. ingegneria dei trasporti, organizzazione del lavoro, logistica, sistemi operativi…)

Sistema di servizio



A seconda di come vengono risolte le contese di utilizzazione
il Sistema di servizio può essere:

- a perdita pura (Q=0)
- orientato alla perdita (Q piccolo)
- orientato al ritardo (Q grande)
- a ritardo senza perdita (Q infinito o almeno uguale al numero di utenti)

Nella slide precedente è rappresentata a sinistra la "popolazione" di utenti, cioè gli utilizzatori potenziali del sistema. Per fare un esempio al di fuori del contesto delle telecomunicazioni, gli "utenti" possono modellare le automobili che arrivano ad un casello, se il modello di servizio rappresenta l'ingresso in una autostrada.

Gli utenti producono nel tempo delle "attività" o "richieste di servizio", che sono rappresentate dai quadrati rossi. Ogni linea orizzontale rappresenta l'evoluzione temporale delle richieste di servizio di un utente.

Le richieste di servizio arrivano al sistema di servizio vero e proprio, caratterizzato da un insieme di serventi (i 4 cerchi blu nello schema) ed eventualmente da una fila di attesa con una certa capacità Q.

La capacità Q della fila di attesa va messa in relazione con le strategie di risoluzione delle contese:

Se Q=0 il sistema è a "perdita pura": se non ci sono serventi disponibili l'attività viene scartata.

Se Q è relativamente piccola il sistema è "orientato alla perdita". Se Q è grande si ha un sistema "orientato al ritardo".

Se Q fosse infinita (impossibile nella pratica, ma vedremo che è una assunzione "comoda" nel modello matematico…) o comunque superiore o uguale al numero di utenti potenziali, si avrebbe un sistema "orientato al ritardo"

Modello del sistema

- Per valutare quantitativamente le prestazioni di una rete di telecomunicazioni è necessario rappresentare in modo astratto le sue funzionalità

Sorgenti di traffico → Sistema di servizio

- Tipici parametri prestazionali sono:
  - per sistemi a perdita
    - frequenza con cui le chiamate vengono rifiutate
    - carico smaltito dal sistema
    - durata dei periodi di congestione
  - per sistemi ad attesa
    - tempo di attesa
    - tempo di permanenza nel sistema
    - numero di richiesta inattesa

Nel “rappresentare in modo astratto le funzionalità della rete” bisogna definire un modello delle “sorgenti di traffico” che ci consenta di modellare la loro attività e l’interazione con le funzionalità della rete.

Una volta definito il modello, possiamo utilizzarlo per effettuare l’analisi prestazionale, che può essere fatta con due obiettivi: *valutazione* delle prestazioni di un dato sistema o *dimensionamento* di un nuovo sistema.

Nel primo caso si hanno a disposizione i dati strutturali del sistema (numero e capacità dei serventi, capacità della fila di attesa, …) e si vuole valutare quali sono le sue prestazioni, per una data intensità della richiesta di servizio da parte degli utenti, o al variare di tale intensità.

Nel secondo caso si parte da una stima della intensità della richiesta di servizio da parte degli utenti e da una serie di obiettivi prestazionali e si vuole arrivare a definire le caratteristiche strutturali del sistema in grado di raggiungere tali obiettivi prestazionali.

Obiettivi dell'analisi prestazionale

- Valutazione delle prestazioni

- Dimensionamento di un sistema

Analisi prestazionale di una rete

- Prestazioni

  Es.: Ritardo di trasferimento
  Probabilità di rifiuto di un servizio

- Costi

  Es.: Numero dei elementi di servizio
  Capacità degli elementi di servizio

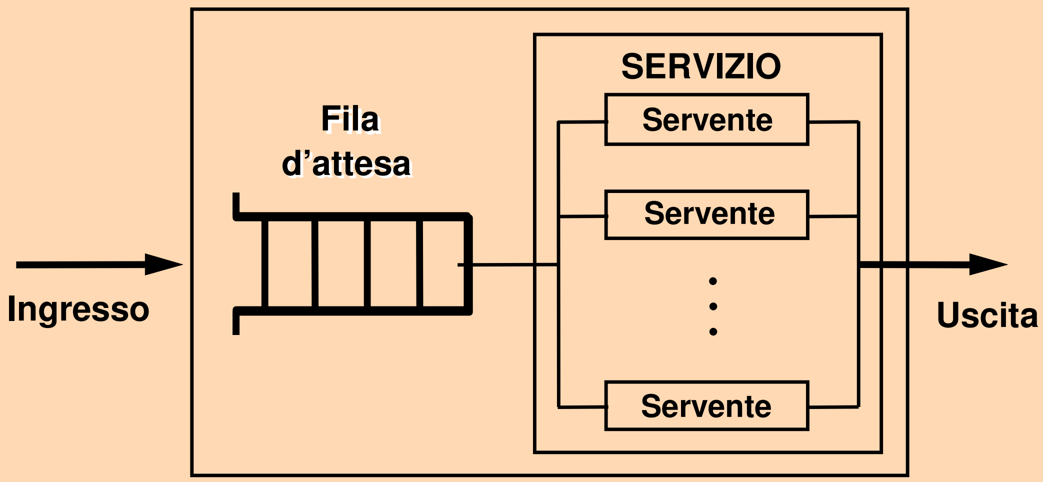
È importante ricordare che le prestazioni di un sistema dipendono dalle caratteristiche strutturali del sistema, ma che queste sono legate ai costi. In generale maggiore sarà la capacità del sistema, maggiori i suoi costi.

Per questo quando si affronta un problema di dimensionamento non si devono in genere tenere in conto solo le prestazioni del sistema ma anche i suoi costi.

Una formulazione di un problema di dimensionamento può quindi essere: realizzare un sistema a costo minimo, dato un "vincolo" sulle prestazioni. Se consideriamo il dimensionamento di un "call center" un esempio di "vincolo" sulle prestazioni può essere: "il ritardo medio degli utenti in attesa di risposta non può superare i 2 minuti".

Una diversa formulazione di un problema di dimensionamento può essere invece: dato un "budget" di spesa, massimizzare le prestazioni del sistema.

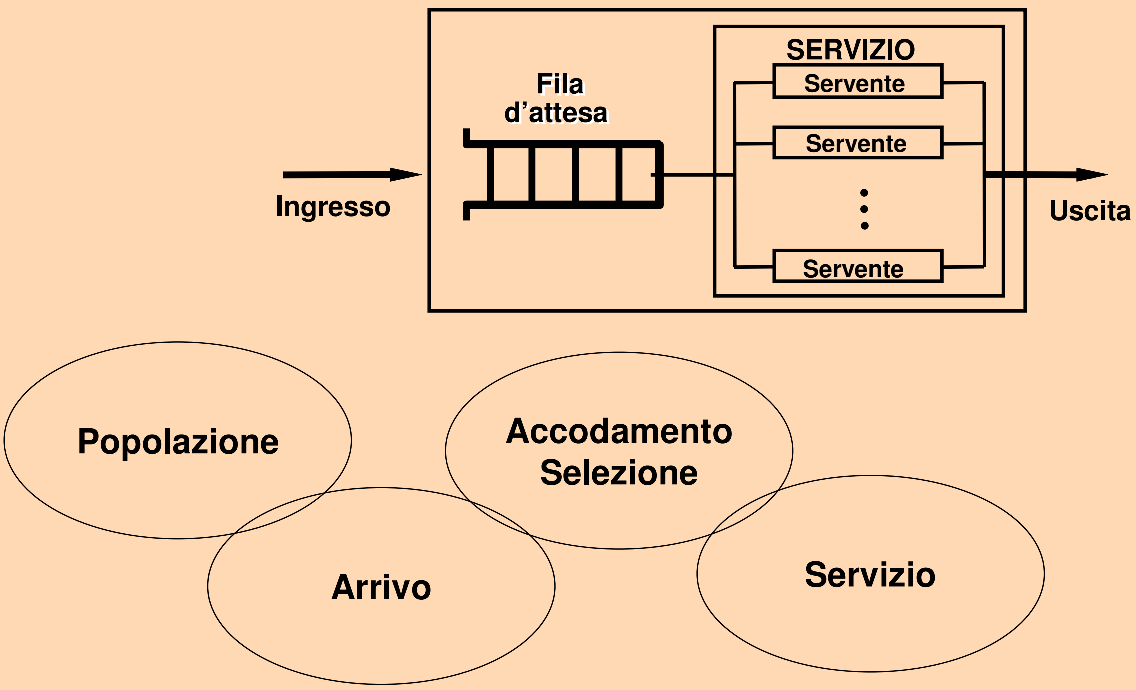
Sistemi a coda



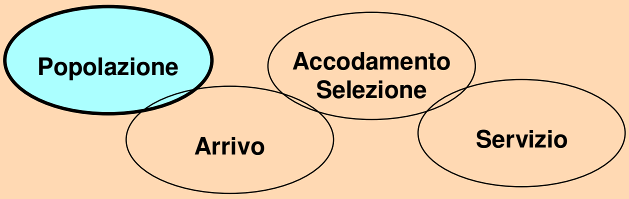
La slide precedente illustra il modello di sistema a coda che andremo a considerare. Le attività entrano dall'ingresso, vengono se necessario accodate nella fila di attesa (se questa non è nulla) e poi vengono servite da uno dei serventi (ci può anche essere un solo servente).

Nella slide successiva vediamo che il sistema a coda è caratterizzato da quattro elementi (Popolazione, Arrivo, Accodamento/Selezione, Servizio) che andiamo di seguito ad analizzare.

Caratterizzazione di un sistema a coda



Popolazione



- **Cardinalità della popolazione**
  - finita
  - infinita

La popolazione si caratterizza con la sua cardinalità ossia con il numero di utenti. La popolazione di un sistema "reale" sarà ovviamente sempre finita, ma risulta comodo nella trattazione del modello matematico utilizzare l'astrazione di "popolazione" infinita.

Il motivo della "comodità" della popolazione di utenti infinita si spiega così. Gli utenti nel modello possono essere o all'interno del sistema (nei serventi o in fila di attesa) oppure fuori in attesa di generare una richiesta di servizio.

Globalmente la quantità di richieste di servizio che arriveranno al sistema in un certo periodo di tempo dipende da quanti utenti potenziali sono fuori dal sistema (più utenti fuori dal sistema, più richieste arriveranno). Se il numero degli utenti è infinito, non conta quanti utenti sono nel sistema: la quantità di richieste sarà la stessa… Consideriamo il caso limite opposto: se esiste solo un utente di un servizio, nel periodo in cui l'utente sta ricevendo il servizio, non vi saranno altre richieste di servizio !

Nella pratica quindi l'astrazione di "popolazione infinita" fornisce una ottima rappresentazione della realtà quando il numero di utenti potenziali è grande rispetto al numero di utenti che possono essere contemporaneamente serviti dal sistema. Ad esempio se il sistema di servizio rappresenta il call center di un operatore cellulare, il numero di utenti che sono collegati al call center (es. dell'ordine dei 100 utenti) sarà sempre molto piccolo rispetto agli utenti dell'operatore (es. 10 milioni).

Accodamento

- **Dimensione della coda**
  - finita
  - infinita
- **Numero di code**

“Saltiamo” per il momento l’analisi dell’ “Arrivo” degli utenti ad un sistema di servizio e concentriamoci sull’accodamento.

La caratteristica principale qui è la dimensione della fila di attesa. Anche qui vale il discorso che la capacità di una fila di attesa di un sistema “reale” sarà sempre finita, ma risulta comodo nella trattazione matematica considerare la fila di attesa come infinita.

La comodità nasce qui dal fatto che nel caso di coda infinita non esistono eventi di perdita, cioè tutti gli utenti che entrano vengono accodati in attesa di essere serviti.

In generale vi può essere una sola coda o varie code in cui vengono accodati gli utenti.

Gli utenti potrebbero (vedi slide successiva) appartenere a classi diverse e quindi ricevere un trattamento diverso (ad esempio essere destinati in code diverse e il servizio potrebbe essere diverso per le varie code). Pensiamo a quello che succede in un ufficio postale dove vi sono le code riservate ai “correntisti” Banco Posta, o a quello che succede in un casello autostradale dove vi sono le code per i possessori di “ViaCard” o per il “Telepass”.

Nel fornire servizio agli utenti accodati nella fila di attesa il sistema può operare con diverse strategie, dette “discipline di coda”. La strategia più tipica è quella “Primo Arrivato Primo Servito” (PAPS) in inglese “First In First Out” (FIFO), che è la tipica strategia di accodamento degli utenti alle poste o alla banca…

Un’altra strategia (Shortest Job First) prevede che i servizi che richiedono un tempo minore siano serviti prima. Un esempio parziale di questa strategia è quello che avviene alla fila di attesa della cassa di un supermercato, dove se un cliente ha una cosa sola da pagare e il cliente precedente ha un carrello pieno ed è in giornata buona, viene fatto passare il cliente con una cosa sola…

I modelli che risolveremo matematicamente saranno esclusivamente modelli a fila di attesa singola gestiti con disciplina di servizio FIFO.

Selezione

- Disciplina di coda
  - Primo Arrivato Primo Servito (PAPS) (FIFO First In First Out)
  - Shortest Job First (SJF)
  - …
- Classi di priorità

Servizio

- Numero dei serventi
- Caratterizzazione del servizio

Infine (ma abbiamo lasciato in sospeso l' "Arrivo") vi è il Servizio, che si caratterizza anzitutto dal numero di serventi. Vi sono sistemi a servente singolo e sistemi con una molteplicità di serventi. Ad esempio il numero dei serventi in un casello autostradale è pari al numero di porte aperte, in un ufficio postale è pari al numero di sportelli aperti e così via.

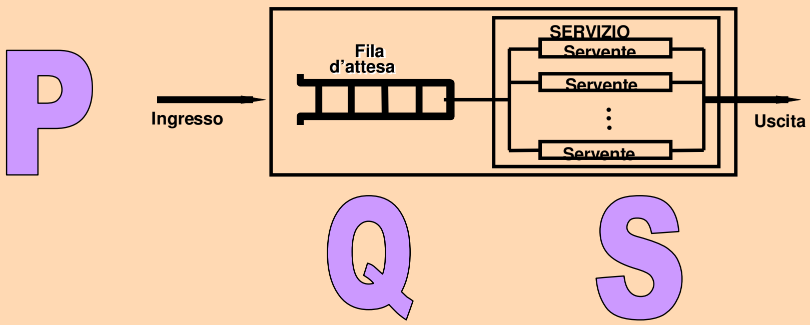
Il servizio si caratterizza ulteriormente con il "tempo di servizio", che può essere fisso ("deterministico") o variabile ("aleatorio"). Un esempio di sistema con tempo di servizio deterministico potrebbe essere un sistema di lavaggio automatico delle automobili che impieghi sempre lo stesso tempo per lavare una automobile.

In moltissimi altri sistemi il tempo di servizio è variabile ed è una variabile aleatoria cioè non è noto a priori. In questi casi si caratterizza il servizio mediante una analisi probabilistica, valutando il tempo medio di servizio, la sua varianza o, in modo più completo, la sua distribuzione di probabilità.

Più avanti torneremo su questa caratterizzazione del processo di servizio.

Si noti (come si vede anche dalla slide successiva) che vi possono essere più serventi. Questi verranno però considerati in questa modellazione tutti equivalente dal punto di vista del servizio che offrono. Anche l'accodamento viene quindi fatto in una unica coda e appena un qualunque servente è disponibile l'utente in testa alla fila di attesa si sposta nel servizi. Non viene quindi considerata (in questi modelli semplici) la possibilità di avere diverse classi di utenti che accedono a servizi tra loro differenziati.

Caratteristiche strutturali



P : cardinalità della popolazione

Q : dimensione della fila di attesa

S : numero di serventi

C=S+Q: capienza del sistema a coda

Esprimiamo quindi le caratteristiche strutturali di un sistema di servizio con i tre parametri C, Q, S.

Q è la dimensione della fila di attesa, S il numero di serventi, C la capacità totale del sistema, quindi C=Q+S.

La popolazione del sistema si esprime con il parametro P.

P, C, Q possono anche essere (nel modello matematico) infinite.

Una classificazione

P: cardinalità della popolazione
Q: dimensione della fila di attesa
S: numero di serventi

C=S+Q: capienza del sistema a coda

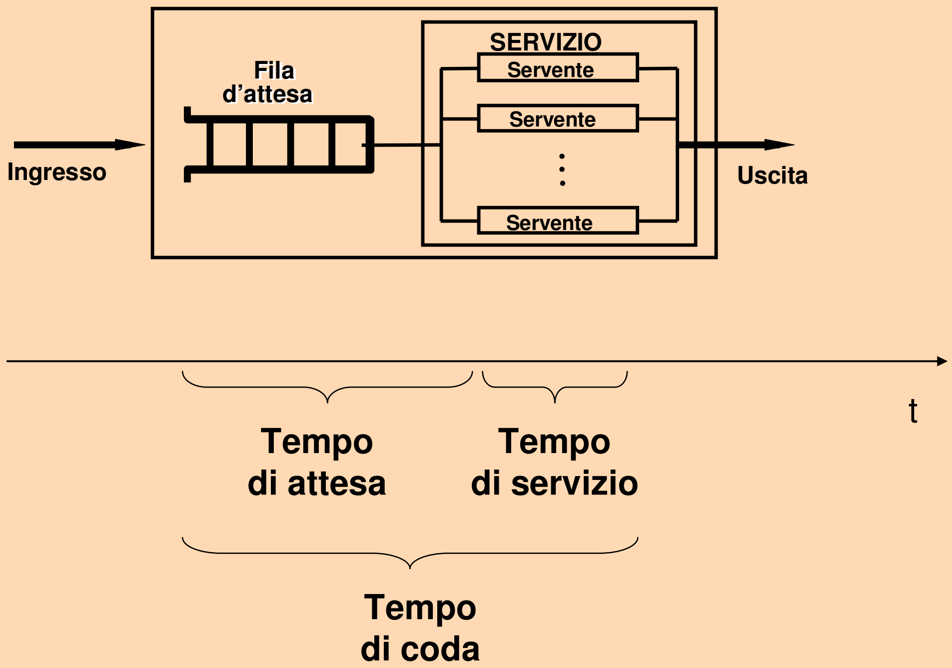
- Se *P*<=*C* e *Q*>0: sistema ad attesa (senza perdita)
- Se P>*C* e *Q*>0: sistema a perdita (con attesa)
- Se *P*>C e *Q*=0: sistema a perdita in senso stretto

Ovviamente se la capacità della coda è infinita $Q=\infty$ (o equivalentemente se la capacità del sistema è infinita $C = \infty$) allora il sistema è ad attesa (infatti $P<=C$ e $Q>0$). Quindi un sistema con fila di attesa non nulla può essere ad attesa o perché la capacità del sistema è maggiore della popolazione, o perché la capacità del sistema è infinita.

Un sistema è a perdita (con attesa) se la capacità del sistema è inferiore alla cardinalità della popolazione e la coda di attesa non è nulla.

Un sistema è a perdita in senso stretto se la capacità del sistema è inferiore alla cardinalità della popolazione e la coda di attesa è nulla.

Un parametro prestazionale: il tempo di coda



Un parametro prestazionale importante per un sistema di servizio è il “tempo di coda”, che (ATTENZIONE!) è definito come il tempo totale per ricevere servizio cioè la somma del “tempo di attesa” (speso nella fila di attesa) e del tempo effettivo di servizio speso presso un servente.

Per il tempo di coda, così come per il tempo di attesa (e il tempo di servizio che avevamo già discusso) si avrà una caratterizzazione probabilistica: valor medio, ove possibile varianza e nel caso più completo distribuzione di probabilità.

Un obiettivo dell’analisi di un modello di sistema di servizio è proprio fornire la caratterizzazione probabilistica del tempo di coda e del tempo di attesa, in modo quanto più completo e accurato possibile. In generale, non sempre è possibile ricavare la distribuzione di probabilità, in certi casi si può avere solo valor medio e varianza o addirittura solo il valor medio.

Arrivo

- Frequenza media …
- Varianza ...
- Processo di arrivo

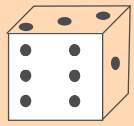
Passiamo quindi all'analisi dell'arrivo delle richieste di servizio ad un sistema a coda, che abbiamo lasciata per ultima. In generale le richieste di servizio arriveranno ad un sistema in modo aleatorio, non prevedibile a priori ma caratterizzabile in termini probabilistici.

La caratterizzazione più semplice è quella della intensità media di arrivo in un intervallo di tempo che, dato un intervallo di tempo, ci dice qual è il valore atteso o valor medio delle richieste di servizio che arriveranno. Si può poi avere una ulteriore caratterizzazione mediante la varianza di questo parametro.

Ulteriori approfondimenti di questa caratterizzaione ci portano a considerare la distanza temporale tra un arrivo e l'altro e la eventuale correlazione tra arrivi successivi. Si vuole cioè studiare nel modo più approfondito possibile il "*Processo di arrivo*". Prima di dare una pur semplificata trattazione di questo argomento, premettiamo alcuni richiami di teoria dei fenomeni aleatori.

- Richiami di teoria della probabilità

Variabili aleatorie e Processi Stocastici

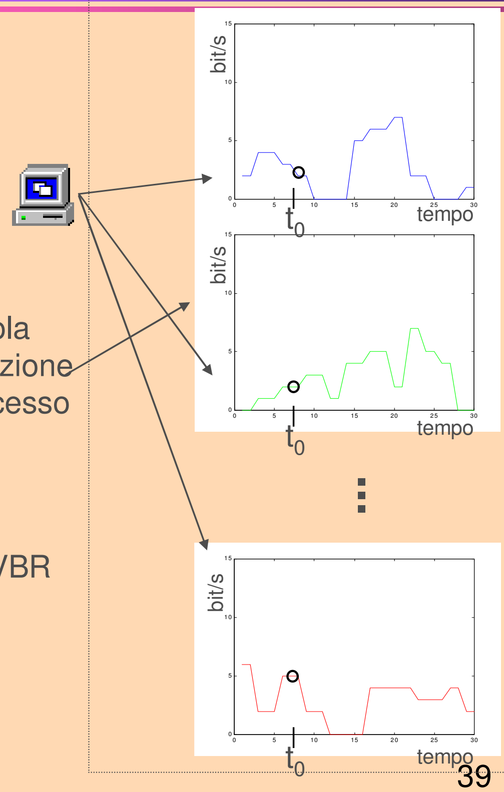


1 (prob. 1/6)
2 (prob. 1/6)
3 (prob. 1/6)
4 (prob. 1/6)
5 (prob. 1/6)
6 (prob. 1/6)

Singola realizzazione della variabile a.

Variabile Aleatoria

Processo aleatorio



Singola realizzazione del processo

Es. R(t) ritmo istantaneo di emissione di una sorgente VBR

Per ogni t_o fissato $R(t_o)$ è una variabile aleatoria

Un processo aleatorio o processo stocastico è una sequenza (continua o discreta) di variabili aleatorie.

La "realizzazione" di un processo aleatorio è una sequenza (continua o discreta) di realizzazioni delle variabili aleatorie che compongono il processo.

Ad esempio supponiamo che il processo aleatorio siano gli istanti temporali in cui arrivano le chiamate ad un call-center. Una realizzazione del processo è la "registrazione" degli orari di arrivo in una data giornata.

Richiami di teoria delle probabilità

- Sia T una variabile aleatoria, definiamo

  f_T la funzione di densità di probabilità di T

$$f_T(t) = P(t < T < t + dt)$$

  F_X la funzione di distribuzione cumulativa di X

$$F_T(t) = P(T < t)$$

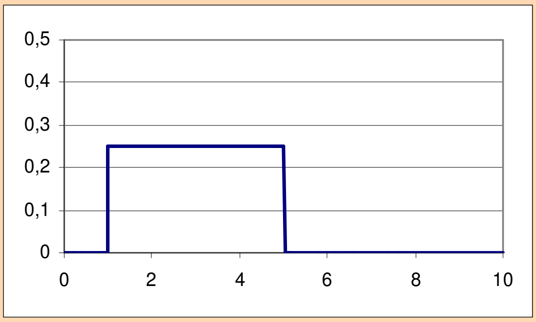
- Valgono le seguenti proprietà:

$$\int_{-\infty}^{+\infty} f_T(t)dt = 1 \qquad F(t) = \int_{-\infty}^{t} f_T(s)ds$$

Esempio: distribuzione uniforme

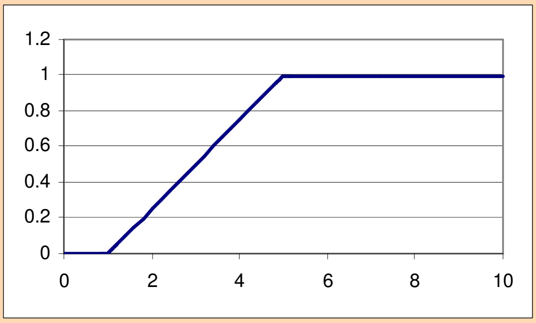
- Es.: T è una variabile aleatoria distribuita uniformemente tra 1 e 5

$$f_T(t) = \begin{cases} 0.25 & 1 \le t < 5 \\ 0 & t < 1, \quad t \ge 5 \end{cases}$$



$$\int_{-\infty}^{+\infty} f_T(t)dt = 1$$

$$F(t) = \int_{-\infty}^{t} f_T(s)ds$$

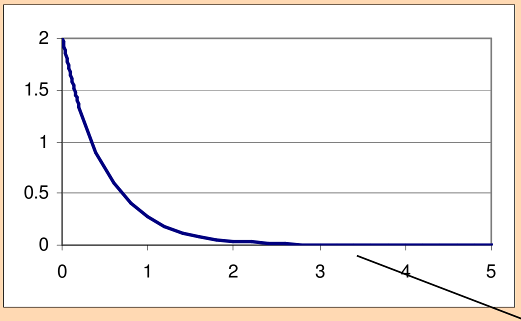


Es.: distribuzione esponenziale negativa

- T è una variabile aleatoria con distribuzione esponenziale negativa

$$f_T(t) = \lambda e^{-\lambda t} \qquad t > 0$$

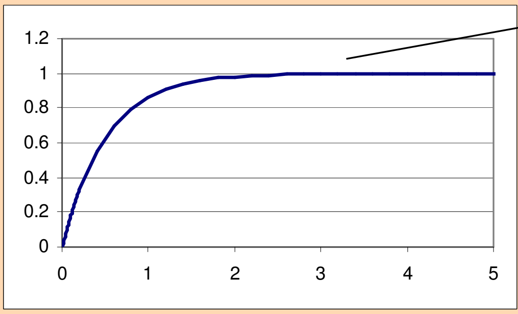
$$es. \quad \lambda = 2$$



$$\int_{-\infty}^{+\infty} f_T(t)dt = 1$$

Converge rapidamente… ma non tocca l’asintoto

$$F(t) = \int_{-\infty}^{t} f_T(s)ds = 1 - e^{-\lambda t}$$



La distribuzione esponenziale negativa non è limitata superiormente: una realizzazione cioè il valore assunto da una variabile aleatoria esponenziale negativa può essere arbitrariamente grande, anche se la probabilità di avere “grandi” valori è “piccola” (vedere l’asintoto orizzontale nella slide precedente).

Richiami di teoria delle probabilità

- Il valore atteso (o medio) di T $$E(T)=\int_{-\infty}^{+\infty} t\cdot f_T(t)dt$$
- La varianza di T $$Var(T)=\sigma_T^2=E(T^2)-(E(T))^2$$
- La deviazione standard di T $$\sigma_T=\sqrt{\sigma_T^2}$$
- Il coefficiente di variazione di T $$c_T=\frac{\sigma_T}{E(T)}$$

Valore medio: esempio

- Il valore atteso di una distribuzione esponenziale negativa

$$f_T(t)=\lambda e^{-\lambda t} \quad t>0$$

$$E(T)=\int_{-\infty}^{+\infty} t\cdot f_T(t)dt=\int_{0}^{+\infty} t\cdot \lambda e^{-\lambda t}dt=\frac{1}{\lambda}$$

- Varianza, deviazione standard, coeff. variazione di T

$$Var(T)=\sigma_T^2=\frac{1}{\lambda^2}; \quad \sigma_T=\frac{1}{\lambda}; \quad c_x=1$$

- Riprendiamo con l’analisi del processo di Arrivo e di Servizio

Caratterizzazione della domanda

La domanda è caratterizzata dalle <u>richieste di servizio</u> presentate dagli <u>utenti</u> del sistema

Consideriamo una sequenza degli istanti di richiesta di servizio:

a) 

t_i : istante di arrivo della i-esima richiesta di servizio

Gli istanti di richiesta di servizio sono distribuiti aleatoriamente sull’asse dei tempi e costituiscono un insieme numerabile

Il tempo di interarrivo

- **Definiamo l'i-esimo <u>tempo di interarrivo</u> τ_i : intervallo che intercorre tra l'istante di presentazione della richiesta immediatamente precedente e quello della richiesta considerata**

$$\tau_i = t_i - t_{i-1} \qquad i=1,2,\ldots$$

Il tempo di servizio

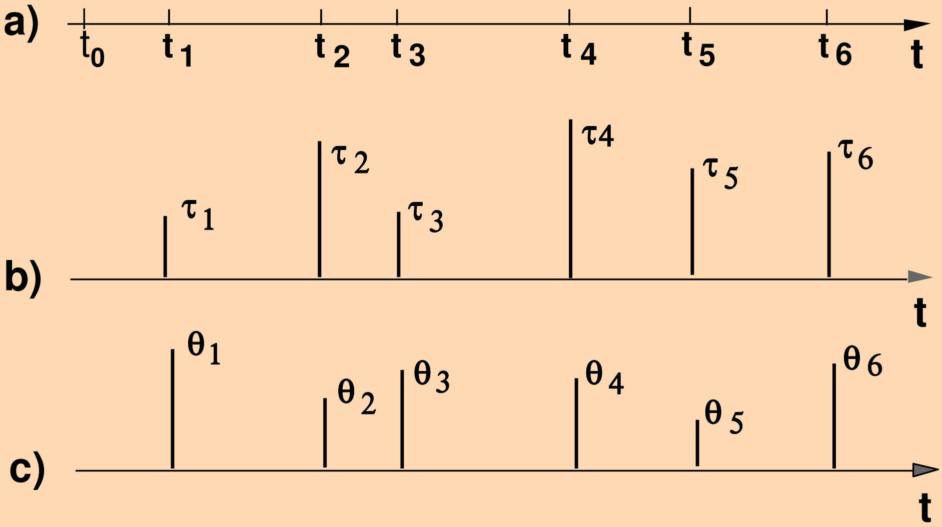
- ***Sia** L_i* **la quantità di lavoro che i serventi del sistema devono svolgere per soddisfare la richiesta *i*-esima**

- **Se supponiamo che la capacità *C* di ogni servente del sistema sia identica per servire la richiesta *i*-esima sarà necessario un tempo pari alla quantità di lavoro diviso la capacità.**

- **Defininiamo quindi l'*i*-esimo <u>tempo di servizio</u> θ_i : intervallo di tempo che un servente deve dedicare per soddisfare tale richiesta**

$$\theta_i = L_i / C \qquad i=1,2,\ldots$$

Caratterizzazione della domanda



t_i : istante di arrivo della i-esima richiesta di servizio

τ_i : tempo di inter-arrivo i-esimo = $t_i - t_{i-1}$

θ_i : tempo di servizio i-esimo

Caratterizzazione della domanda

Consideriamo le due sequenze $\{\tau_i\}$ e $\{\theta_i\}$

Ciascun elemento τ_i e θ_i è una variabile aleatoria

La sequenza dei tempi di inter-arrivo $\{\tau_i\}$ e quella dei tempi di servizio $\{\theta_i\}$ costituiscono la realizzazione di due processi stocastici:

il <u>processo di ingresso (o di arrivo)</u>

il <u>processo di servizio</u>

Per esempio il valore τ_1 cioè la realizzazione del primo tempo di interarrivo è la differenza t_1-t_0 tra le realizzazioni del tempo di arrivo della richiesta 0 e il tempo di arrivo della richiesta 1. Esiste un valore atteso di τ_1 che ci dice “in media” quanto vale τ_1 se osserviamo molte realizzazioni dello stesso processo.

Caratterizzazione della domanda

Normalmente si suppone che le variabili aleatorie componenti ognuno di questi due processi siano

- equi-distribuite
- statisticamente indipendenti

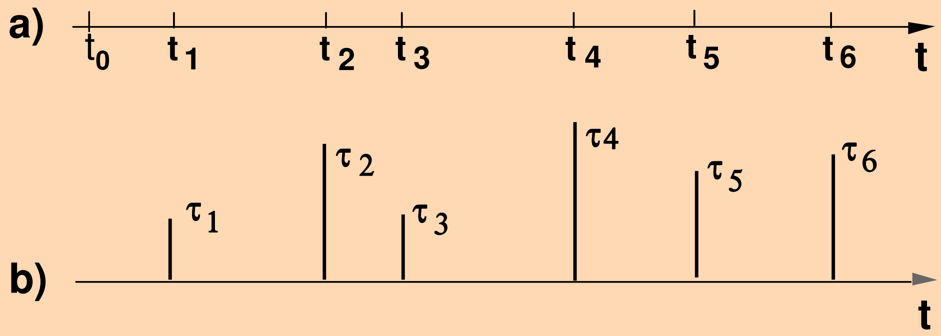
Inoltre, per ciò che riguarda la relazione tra i processi di ingresso e di servizio, assumeremo che tali processi siano statisticamente indipendenti

Il fatto che le variabili aleatorie siano equidistribuite significa ad esempio che il valore atteso di τ_1 è uguale al valore atteso di τ_2, di τ_3, di τ_4 e così via. In realtà vuol dire anche che la distribuzione di probabilità è uguale per tutte le variabili aleatorie componenti (e quindi oltre al valore atteso anche la varianza sarà uguale).

La indipendenza statistica tra le variabili aleatorie significa che il fatto che una certa τ_i abbia avuto un certo valore in una realizzazione non ha nessuna influenza sui valori delle altre variabili aleatorie τ_j della realizzazione.

Quanto detto per le τ_i (processo di arrivo) vale per le θ_i (processo di servizio). In più si aggiunge l'indipendenza statistica fra le τ_i e le θ_i.

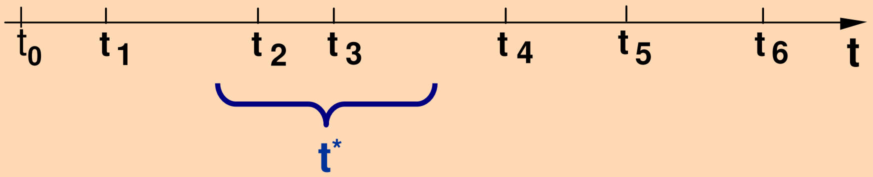
Caratterizzazione della domanda



- Consideriamo il processo di presentazione degli utenti. Abbiamo visto che una sua caratterizzazione è data dal processo dei tempi di interarrivo $\{\tau_i\}$.

Caratterizzazione della domanda

a)



- Possiamo alternativamente fissare un intervallo di dimensione t^* e chiederci quanti arrivi siano contenuti in tale intervallo. Sia $X(t^*)$ tale numero (<u>processo di conteggio</u>).

- Se $\lambda = 1 / T$ è la frequenza media di arrivo (T è il tempo medio di interarrivo), allora il valor medio di $X(t^*)$ vale:

$$E[X(t^*)] = \lambda \cdot t^*$$

Nella slide precedente abbiamo stabilito una relazione tra i valori medi del processo dei tempi di interarrivo e del processo di conteggio.

La relazione è ovvia: supponiamo che in media il tempo di interarrivo sia 0,2 secondi. Quindi la frequenza di interarrivo λ sarà 5, ossia 5 richieste al secondo. Se fisso un intervallo di 4 secondi, mi aspetto in media 5*4=20 richieste.

Caratterizzazione della domanda

- **La distribuzione di X(t*) cioè la P(X(t*)=k) dipende dalla distribuzione dei tempi di interarrivo τ_i**
- Se i tempi di interarrivo hanno distribuzione esponenziale negativa

$$f_T(t) = \lambda e^{-\lambda t} \quad t > 0$$

si dimostra che il processo di conteggio ha distribuzione:

$$P\left(X\left(t^*\right) = k\right) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

- Tale processo si chiama processo di poisson.

Bisogna ***solo*** sapere che un processo di poisson è un processo il cui tempo di interarrivo è esponenziale negativo e viceversa

NON useremo la distribuzione indicata sopra...

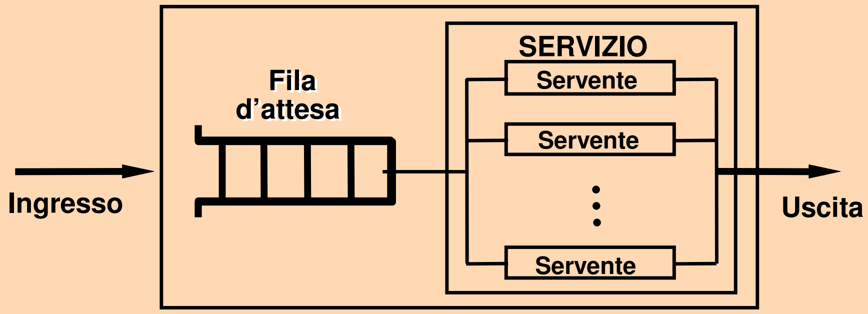
Nella slide precedente abbiamo invece stabilito una relazione tra le distribuzioni di probabilità del processo di interarrivo e del processo di conteggio, che vale quando il processo di interarrivo ha distribuzione esponenziale negativa.

La distribuzione esponenziale negativa verrà utilizzata spesso nella trattazione dei modelli di servizio, per diversi motivi di cui ne accenniamo due:

- un processo di arrivo ottenuto dalla "somma" di una infinità di eventi indipendenti (ad esempio un grande numero di utenti telefonici che generano chiamate) è un processo di poisson e quindi la distribuzione del tempo di interarrivo è esponenziale negativa

- sperimentalmente si è visto che la distribuzione della durata delle chiamate telefoniche è molto simile alla distribuzione esponenziale negativa (questa "somiglianza" è si ha in particolare per le chiamate telefoniche tra persone, mentre l'introduzione di fax, segreterie telefoniche e soprattutto l'uso del telefono per collegarsi ad internet via-modem modifica la distribuzione sperimentale allontanandola dalla esponenziale negativa)

Caratterizzazione di un sistema a coda



Caratterizzazione di un sistema a coda

- Sinteticamente, una coda è definita da:
  - » processo degli arrivi (d.d.p.) $A(t)$
  - » tempi di servizio (d.d.p.) $B(t)$
  - » numero di servitori S
  - » capacità del sistema C
  - » cardinalità della popolazione P
  - » disciplina di servizio

Notazione sintetica (Kendall)

A / B / S / C / P

A e B possono assumere i valori:

- » ***M*** **(Esponenziale negativa o “Markoviana”)**
- » *D* *(deterministica o costante)*
- » E_i *(Erlangiana di ordine i)*
- » *G* *(generale)*

S, C e P assumono valori numerici, quando C o P sono infiniti non si indicano

Attenzione: nel testo “Reti di telecomunicazione” sez. II.2.5 si segue la convenzione di indicare con il quarto parametro la capacità della fila di attesa Q invece della capacità del sistema C. Le due notazioni sono equivalenti, dato che C=Q+S.

Notazione sintetica (Kendall)

$$A / B / S / C / P$$

Esempi:

$M / M / 1$

$M / M / 1 / k$

$M / M / m$

$M / M / m / m$

$M / G / 1$

Noi tratteremo in particolare le code M/M/1 ossia con un solo servente, coda infinita e popolazione infinita e le code M/M/m/m ossia con m serventi e coda nulla (se ho m serventi e capacità del sistema m, vuol dire che la coda è nulla).

La prima coda rappresenterà per noi un multiplatore (nodo) a pacchetto, che trasferisce unità dati provenienti da più sorgenti in un unico canale trasmissivo. Il canale trasmissivo viene modellato dal servente, le richieste di servizio sono le unità dati da trasmettere, la fila di attesa rappresenta lo spazio di memoria in cui vengono memorizzate le unità dati nel multiplatore in attesa di essere trasmesse. Nella coda M/M/1 che studieremo si assume che la capacità di memoria sia infinita. La coda M/M/1/k (che non studieremo) invece può rappresentare un multiplatore a pacchetto "reale" in cui vi siano k-1 posti nella fila di attesa.

La coda M/M/m/m rappresenterà le chiamate telefoniche offerte ad un flusso aggregato che può trasportare m chiamate. Gli m canali disponibili in uscita sono rappresentati dai serventi. La coda di attesa è nulla in quanto le chiamate che arrivano al sistema nel momento in cui tutti i canali di uscita sono occupati vengono scartate.