# Reti di Calcolatori

Teoria delle Code

# La Teoria delle Code

La teoria delle code studia I fenomeni di attesa che si possono verificare quando si richiede un servizio. Rappresenta un formalismo matematico che trova una naturale applicazione nelle reti di calcolatori, nei sistemi operativi e/o nella programmazione, ma anche in vari contesti applicative della vita quotidiana.

La formulazione del primo problema di teoria delle code risale al 1908, quando A.K. Erlang ritenuto il fondatore della teoria delle code, voleva studiare come dimensionare centrali telefoniche allo scopo di mantenere ad un valore ragionevolmente basso il numero delle chiamate che non potevano essere connesse (chiamate perse) perché il centralino era occupato.

# Sistema di Servizio

Un Sistema di Servizio rappresenta una struttura a cui arrivano casualmente dei client che richiedono lo svolgimento di una operazione o servizio da parte di uno o più servente. Ogni servente può svolgere l'operazione per un solo cliente per volta.

Quando il numero di clienti eccede quello dei serventi, allora è necessario accordarli, e quindi tale Sistema dispone di uno o più aree di attesa o buffer, a cui è associate una politica di scelta di quali client accodati servire.

Lo scopo della teoria della code è quello di valutare alcune misure di prestazione sulle quali basarsi per dimensionare il sistema di servizio, come la lunghezza media della coda, il numero medio di utenti presenti nel sistema, la durata media del tempo passato nella coda.

# Componenti di un Sistema di Servizio

- Popolazione - I potenziali clienti in arrivo al Sistema, che sono indistinguibili e caratterizzati dalla dimensione (finita o infinita), il numero totale dei distinti potenziali clienti che richiedono un servizio.
- Numero di serventi – è possibile avere più serventi che lavorano in serie o in parallelo e il cui numero è indicato con s.
- Schema di arrivo - il modo (deterministico o aleatorio) secondo il quale i clienti si presentano a richiedere il servizio, e viene definito in termini di intervalli di tempo tra due arrivi successivi di clienti nel sistema (tempo di inter-arrivo). $t_i^a$ rappresenta il tempo che intercorre tra l'arrivo del cliente (i-1)-esimo e il cliente i-esimo. Degli intertempi di arrivo si suppone nota la distribuzione di probabilità.

# Componenti di un Sistema di Servizio

- Schema di servizio - il modo (deterministico o aleatorio) secondo il quale i serventi svolgono un servizio ai clienti, $t_i^s$ rappresenta il tempo che intercorre per svolgere un servizio da parte di un servente al cliente i-esimo.

- Capacità del sistema - numero massimo di utenti che possono essere contemporaneamente nel sistema, comprendendo sia gli utenti in attesa in coda, sia quelli che stanno fruendo del servizio.

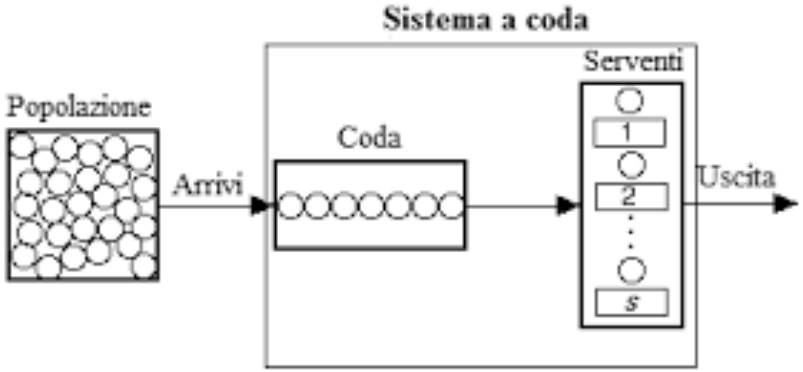
- Disciplina della coda - l'ordine rispetto al quale gli utenti vengono serviti:
  - FIFO ("first in first out") utenti serviti nell'ordine in cui arrivano;
  - LIFO ("last in first out") servire per primo l'ultimo cliente arrivato;
  - SIRO ("service in random order") servire gli utenti scegliendoli a caso;
  - criteri di priorità (PRI).

# Componenti di un Sistema di Servizio

- Dimensione della coda – numero di utenti (finito o infinito) che possono trovare posto nella coda. $t_i^q$ rappresenta il tempo passato in coda dall'i-esimo utente nella coda prima di iniziare ad usufruire del servizio richiesto.

I clienti che richiedono un servizio sono generati nel tempo dalla popolazione, entrano nel sistema e raggiungono la coda. Ad un certo momento, un cliente viene selezionato secondo la disciplina della coda. Il servizio richiesto è effettuato da un servente e successivamente a ciò il cliente lascia il sistema. Quindi se indichiamo con $t_i^w$ il tempo passato complessivamente dall'i-esimo utente nel sistema si ha

$$t_i^w = t_i^q + t_i^s$$




Si assumono $t_i^a$ e $t_i^s$ indipendenti e identicamente distribuiti.

# Notazione di Kendall

Nel 1953 David George Kendall introdusse una notazione capace di poter indicare in modo unitario tutti gli elementi caratteristici di un sistema a coda:

A/B/s/c/p/Z

- A rappresenta lo schema di arrivo, ovvero la distribuzione di probabilità degli intertempi di arrivo;
- B rappresenta lo schema di servizio, ovvero la distribuzione di probabilità dei tempi di servizio;
- s rappresenta il numero di serventi;
- c rappresenta la capacità del sistema, se non indicata si assume infinita;
- p rappresenta la dimensione della popolazione , se non indicata si assume infinita;
- Z rappresenta la disciplina della coda, se non indicata si assume FIFO.

# Notazione di Kendall

Le componenti s, c e p sono numeri interi non negativi.

Le distribuzioni di probabilità dello schema di arrivo e di servizio più frequentemente assunte sono la distribuzione esponenziale, la distribuzione costante (degenere) o tempi deterministici, la distribuzione di Erlang di ordine k.

Queste vengono indicate per le componenti A e B, nel seguente modo:

- M indica la distribuzione esponenziale o Markoviana;
- D indica la distribuzione costante (degenere) o tempi deterministici;
- $E_k$ indica la distribuzione di Erlang di ordine k;
- G indica una distribuzione generica che, per quanto riguarda gli intertempi di arrivo, può essere sostituita dalla sigla GI ad indicare un distribuzione generica di eventi indipendenti.

# Definizioni e Notazioni standard

- $\lambda$ indica la frequenza media degli arrivi dei clienti nel sistema, ovvero il numero medio di arrivi di utenti nell'unità di tempo. Dati i tempi di inter-arrivo, si ha $E(t_i^a)$ come media dei tempi di interarrivo ho

$$\lambda = \frac{1}{E(t_i^a)}$$

- $\mu$ indica la velocità di servizio, ovvero il numero medio di utenti per il quali è espletato il servizio nell'unità di tempo. Come prima,

$$\mu = \frac{1}{E(t_i^s)}$$

- Sia $\lambda$ che $\mu$ potrebbero non essere costanti al variare del numero di utenti presenti nel sistema. In questi casi, se k denota il numero di utenti presenti nel sistema, verranno denotate con $\lambda_k$ sia $\mu_k$.

# Definizioni e Notazioni standard

- $\varrho$ indica il fattore di utilizzazione dei serventi, ovvero il rapporto tra la frequenza media degli arrivi e la velocità del servizio moltiplicata per il numero dei serventi:

$$\varrho = \frac{\lambda}{s\mu}$$

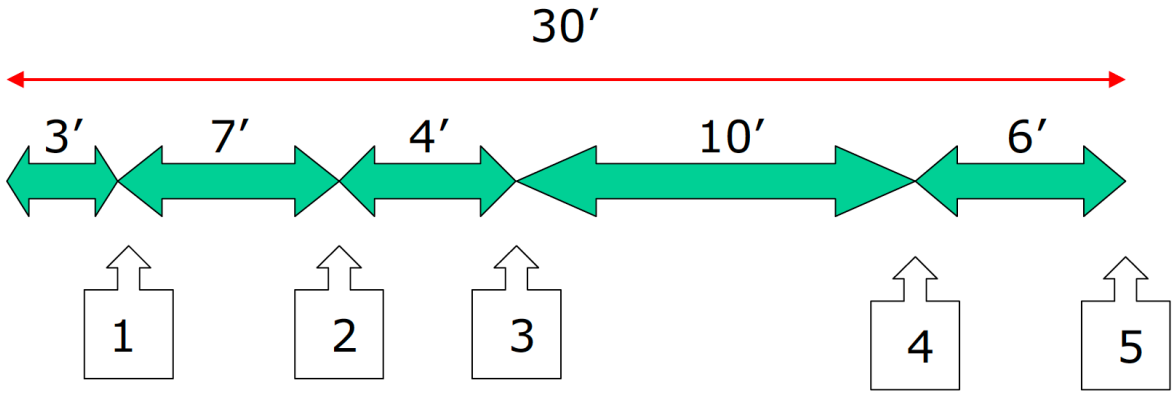
Rappresenta la frazione di tempo che i serventi sono occupati.

- Lo stato di un sistema a coda al tempo t o n(t) è il numero di clienti presenti nel sistema ed è quindi dato dalla somma del numero dei clienti che sono nella fila di attesa e il numero dei serventi attivi.

- $n^q(t)$ è la lunghezza della coda al tempo t:

$$n^q(t) = \begin{cases} 0 & se\ n(t) \leq s \\ n(t) - s & se\ n(t) > s \end{cases}$$

# Esempio




Il numero medio di arrivi di utenti nell'unità di tempo è di 5 in 30 minuti, ovvero 1/6, quindi $\lambda$ = 0.1666 utenti al minuto. Analogamente se il numero medio di utenti per il quali è espletato il servizio è pari a 4 al minuto, ovvero $\mu$ = 4, allora il tempo medio di servizio è 1/4 di minuto.

# Misure di Prestazione

La teoria delle code mira a studiare il sistema di servizio quando ha raggiunto una situazione di regime e ciò avviene quando il sistema è stato in funzione per un tempo sufficientemente grande.

- Quando lo stato del sistema sarà fortemente influenzato dallo stato iniziale e dal tempo che è trascorso dall'attivazione, si dice che è in condizioni transitorie.

- Trascorso un tempo sufficientemente grande, il sistema diviene indipendente dallo stato iniziale e si dice che il sistema ha raggiunto condizioni stazionarie o di equilibrio (steady–state).

La teoria delle code analizza principalmente sistemi in condizioni di stazionarietà.

# Misure di Prestazione

Per effettuare l'analisi di un sistema in condizioni di stazionarietà si fa uso delle seguenti quantità fondamentali:

- $p_k$: probabilità che k utenti siano presenti nel sistema. Questa può dipendere dal tempo e quindi la si indica come $p_k$(t). La maggior parte dei sistemi a coda di interesse raggiungono una situazione di equilibrio, indipendentemente dalla stato iniziale:

$$\lim_{t\to\infty} p_k(t) = p_k,\ \text{k = 0, 1, ...,}$$

- $N$ : numero medio degli utenti nel sistema:

$$E\big(n(t)\big) = \sum_{k=0}^{\infty} k p_k(t) \Rightarrow N = \lim_{t\to\infty} E\big(n(t)\big) = \sum_{k=0}^{\infty} k p_k$$

- $N^q$: numero medio degli utenti nella coda:

$$E\big(n^q(t)\big) = \sum_{k=0}^{\infty} (k-s) p_k(t) \Rightarrow N = \lim_{t\to\infty} E\big(n(t)\big) = \sum_{k=0}^{\infty} (k-s) p_k$$

# Misure di Prestazione

- $T$ : tempo medio passato da un utente nel sistema: $T = \lim_{t \to \infty} E(t_i^w)$
- $T^q$ : tempo medio passato da un utente nella coda: $T^q = \lim_{t \to \infty} E(t_i^q)$

Esistono importanti relazioni che legano le precedenti 4 quantità in un sistema a coda in condizioni stazionarie. La prima relazione è il cosiddetto Teorema di Little:

- In un sistema a coda in condizioni stazionarie vale $N = \lambda T$, ovvero il numero medio di clienti in un sistema è pari al tempo medio di permanenza nel sistema per il tasso medio di arrivo.

La formula di Little ha una validità generale ed è indipendente dalle distribuzioni di probabilità dei tempi di inter-arrivo e dei tempi di servizio, dalla disciplina del servizio, è valida per i sistemi stazionari, e la frequenza considerata deve essere la frequenza degli ingressi effettivi nel sistema.

# Misure di Prestazione

Sostituendo al numero atteso di clienti nel sistema la lunghezza media della coda e al tempo di permanenza medio nel sistema il tempo di permanenza medio nella coda, otteniamo la stessa relazione riferibile però alla coda:

$$N^q = \lambda T^q$$

Una ulteriore importante relazione lega il valore atteso del tempo passato da un utente nel sistema e il valore atteso del tempo passato nella coda:

$$T = T^q + \frac{1}{\mu}$$

Le tre relazioni analizzate rappresentano le relazioni fondamentali tra le 4 grandezze e sono del tutto generali, cioè valgono in qualsiasi tipologia di sistema di code. Sono basilari in quanto una volta nota una delle grandezze si possono facilmente determinare le altre tre.

# Processi Stocastici

I processi stocastici sono modelli matematici adatti a studiare l'andamento di fenomeni che seguono leggi casuali o probabilistiche. Il risultato dell'osservazione ad ogni istante non sia un dato certo e scevro da errori ad esso connessi, ma avrà sempre un'incertezza correlata.

I processi stocastici che caratterizzano un sistema coda sono due:

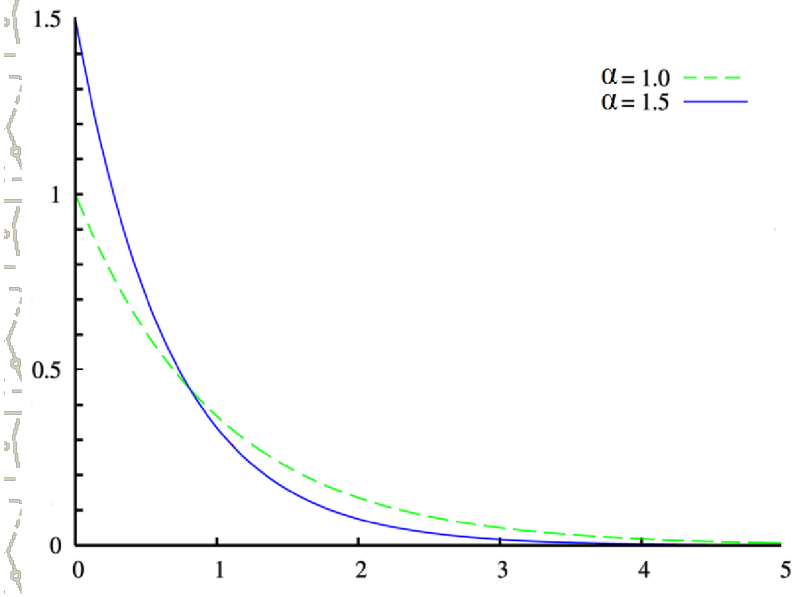
1. Il processo degli arrivi: questo è caratterizzato da una distribuzione di probabilità dei tempi di interarrivo, ovvero i tempi tra gli arrivi di due clienti successivi. Al fine di renderli facilmente studiabili, nella teoria delle code, essi vengono considerati stazionari, ovvero le loro proprietà statistiche (come il tempo medio di interarrivo) non variano nel tempo.
2. Il processo di servizio: una distribuzione di probabilità descrive il tempo impiegato da ogni operatore per soddisfare le richieste del generico cliente

# Distribuzione Esponenziale

Nella teoria della probabilità, una variabile aleatoria o stocastica è una variabile che può assumere valori diversi in dipendenza da qualche fenomeno aleatorio.

Una variabile aleatoria continua $T$ si dice distribuita esponenzialmente con parametro $\alpha$ quando la funzione di densità è data da

$$f(t) = \begin{cases} \alpha e^{-\alpha t} & se\ t \geq 0 \\ 0 & se\ t < 0 \end{cases}$$




Il parametro $\alpha$ è l'inverso del valore atteso che intercorre tra l'arrivo di due clienti successivi e può essere interpretato come il tasso medio di arrivo dei clienti.

$$E(T) = \frac{1}{\alpha} \quad Var(T) = \frac{1}{\alpha^2}$$

# Distribuzione Esponenziale

È maggiore la probabilità che T assuma un valore piccolo vicino allo zero piuttosto che valori vicini a $E(T)$.

- Se T rappresenta il tempo di servizio, in caso di servizio identico per ogni cliente i tempi di servizio tenderanno ad essere sempre costanti e vicini al valore atteso $E(T)$, la distribuzione esponenziale non rappresenta una buona approssimazione. Se invece il servizio differisce da cliente a cliente una distribuzione esponenziale dei tempi di servizio sembra ragionevole per esprimere questa situazione.
- Se T rappresenta il tempo di inter-arrivo tale proprietà esclude che i clienti rinuncino ad entrare nel sistema vedendo entrare un altro cliente prima di loro; ciò è in sintonia con la casualità che deve essere espressa nel processo degli arrivi, pertanto è una buona assunzione per rappresentare un modello realistico.

La distribuzione di densità del tempo di attesa per l'evento è sempre la stessa, indipendentemente dal tempo già passato (Δt): in altre parole il processo dimentica il passato perché senza memoria.

# Distribuzione Esponenziale

- Per i tempi di servizio, se il servizio varia da cliente a cliente questa proprietà è ottimale per rappresentare questa situazione; al contrario nei casi in cui il servizio è tendenzialmente costante, non è ottimale l'uso di tale distribuzione.
- Questa proprietà, invece, rende la distribuzione esponenziale adatta a modellare i tempi di inter-arrivo, purché essi non sono correlati, cioè quei casi in cui l'arrivo di un cliente sfavorisce altri arrivi.

Se il tempo tra due occorrenze successive di un determinato evento è distribuito esponenzialmente con parametro $\alpha$ allora il numero di eventi che si verificano in un dato tempo è un processo di Poisson con parametro $\alpha t$.

Un processo aleatorio $\{N(t), t \geq 0\}$ viene chiamato contatore se $N(t)$ rappresenta il numero totale di eventi che si sono verificati entro l'istante $t$. Un processo di Poisson è un particolare processo contatore con $N(0) = 0$, incrementi indipendenti e numero di eventi in un qualsiasi intervallo di durata $t$ è una distribuzione di Poisson con media $\lambda t$ per ogni $s, t \geq 0$.

# Distribuzione Esponenziale

Quando gli intertempi sono distribuiti esponenzialmente il numero di eventi che si verifica in un dato tempo t è un processo di Poisson.

- Assumiamo che i tempi di servizio hanno una distribuzione esponenziale con parametro μ e sia $N(t)$ il numero di servizi completati da un singolo servente continuamente occupato in un tempo t con $\alpha = \mu$: nelle code con più serventi, $N(t)$ è definito come il numero di servizi completati in un tempo t da n serventi occupati, con $\alpha = n\mu$.
- Quando i tempi di interarrivo hanno una distribuzione esponenziale con parametro α, $N(t)$ è il numero di arrivi nel tempo t, dove $\alpha = \lambda$ è il tasso medio di arrivo. Per questo motivo si deduce che gli arrivi si verificano secondo un processo di Poisson con parametro λ.

# Test del Chi-Quadro

Uno degli aspetti più complessi della selezione del modello di code per rappresentare la realtà risiede nella determinazione delle distribuzioni di probabilità adatte per i tempi di inter-arrivo e per i tempi di servizio.

Un metodo efficiente è quello di raccogliere dati statistici riguardanti il numero di arrivi in intervalli di tempo costanti ed i tempi di servizio. Dopo aver scelto una certa distribuzione per queste due variabili si può usare il test del chi-quadro per verificare se è valida l'ipotesi che i dati osservati nel sistema appartengono alla distribuzione scelta.

- Avendo a disposizione n osservazioni reali di un fenomeno le dobbiamo dividere in k classi, in base ai valori che possono assumere. Per ciascuna classe individuiamo la frequenza reale, cioè il numero di volte per cui l'osservazione reale ha assunto un valore di quella classe.
- Si usa il modello scelto per stimare le frequenze previste.

# Test del Chi-Quadro

- Si effettua un confronto come segue:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i^{reale} - f_i^{stima})^2}{f_i^{stima}}$$

Più grande è il valore di $\chi^2$ maggiore è la discrepanza tra le frequenze osservate e quelle teoriche, peggiore sarà il modello scelto.

# Proprietà PASTA

Supponiamo che un sistema di code sia caratterizzato da arrivi che seguono un processo di Poisson, esso gode di una importante proprietà detta Poisson Arrivals See Times Average (PASTA): gli utenti che arrivano nel sistema di code trovano, in media, nel sistema la stessa situazione che vedrebbe un osservatore esterno al sistema che osserva il sistema in un momento arbitrario nel tempo.

Indicando con $a_k(t)$ la probabilità che un utente che arriva nel sistema al tempo t trova il sistema allo stato k (cioè con k utenti presenti) e con $p_k(t)$ alla probabilità che il sistema sia allo stato k al tempo t, si ha:

$$a_k(t) = p_k(t)$$

In condizioni di stazionarietà, imponendo il limite su t si ha $a_k = p_k$.

# Proprietà PASTA

In maniera analoga si può studiare la distribuzione dopo la partenza di un cliente dal sistema dopo che ha usufruito del servizio definendo, indicando con $d_k(t)$ la probabilità che un utente che esce al tempo t lascia il sistema nello stato k, e, in condizioni di stazionarietà:

$$\lim_{t \to \infty} d_k(t) = d_k$$

Per un qualsiasi sistema di code a cosa singola, non necessariamente con arrivi poissoniani, vale l'uguaglianza

$$a_k = d_k$$

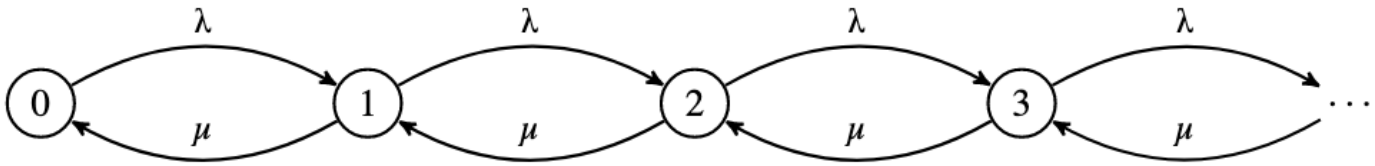
purché i clienti arrivano al sistema uno alla volta e sono serviti uno alla volta. Quando gli arrivi sono poissoniani vale la proprietà PASTA e quindi

$p_k = a_k = d_k$, ovvero,

sia un cliente che arriva, sia un cliente che parte da un sistema in condizioni di stazionarietà, vede un sistema che è statisticamente equivalente ad un sistema visto da un osservatore che osserva il sistema dall'esterno in un arbitrario istante di tempo.

# Code M/M/1

È fisicamente composta da un buffer e da un solo servitore; in essa il tempo di inter-arrivo tra due arrivi successivi e il tempo di servizio sono due variabili aleatorie markoviane, cioè con distribuzione esponenziale, così che lo stato della coda è descritta come un processo di Markov a salti:




La condizione dello stato stazionario si ha quando $0 < \varrho < 1$.

In un tale sistema si ha che il numero medio di clienti nel sistema è pari a:

$$N = E\big(n(t)\big) = \frac{\varrho}{1-\varrho} = \frac{\lambda}{\mu-\lambda}$$

Mentre il numero medio di utenti in attesa è pari a:

$$N^q = E\big(n^q(t)\big) = \frac{\varrho}{\mu(1-\varrho)} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

# Code M/M/1

Applicando il teorema Little, si ha:

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

$$T^q = \frac{N^q}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)}$$

Il tempo di permanenza nel sistema $t^w$ è distribuito esponenzialmente con parametro $\mu - \lambda$:

$$P(t^w > t) = e^{-(\mu-\lambda)t}, t \geq 0$$

Per la distribuzione del tempo di attesa si ha che:

$$P(t^q = 0) = 1 - \varrho$$
$$P(t^q > t) = \varrho e^{-(\mu-\lambda)t}, t \geq 0$$

# Esempio di M/M/1

In un sistema di comunicazione, la velocità di trasmissione è C = 1200bit/sec. I messaggi in arrivo che devono essere trasmessi formano un processo Poisson. Un messaggio consiste di L bits, dove L è una variabile random. Si assume che L è esponenzialmente distribuito con media 600 bits. Il problema è determinare la frequenza di arrivi massimi che si possono sostenere per garantire il tempo di attesa medio di un messaggio minore di 1 secondo.

# Esempio di M/M/1

Si comincia con il determinare la frequenza di servizio μ espresso in messaggio al secondo, come rapporto tra la velocità di trasmissione e il numero di bit nel messaggio.

$$\mu = \frac{C}{L} = 2 \quad \rho = \frac{\lambda}{2}$$

Il vincolo sui tempi di attesa può essere formulato come segue:

$$T^q = \frac{\lambda}{2(1-\lambda)} = \frac{\lambda}{2-2\lambda} < 1$$

Quindi ne consegue che $\lambda \leq 2/3$ messaggi al secondo. Sotto queste condizioni il sistema deve operare ad una frequenza di utilizzazione pari a $\varrho < 1/3$

# Code M/M/s

Ipotizzianmo che s > 1 e $\varrho = \frac{\lambda}{s\mu} < 1$ per l'esistenza dello stato stazionario, ecco le formule per le quattro grandezze fondamentali:

$$N^q = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{\varrho}{(1-\varrho)^2} p_0, \text{ con } p_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{1}{1-\varrho}}$$

$$T^q = \frac{N^q}{\lambda}$$

$$T = T^q + \frac{1}{\mu}$$

$$N = \lambda T = N^q + \frac{\lambda}{\mu}$$

# Esempio di Code M/M/s

Una stazione ha due thread ugualmente efficienti, ciascuno dei quali è in grado di servire, in media, 60 messaggi al minuto e i tempi di servizio sono distribuiti esponenzialmente. Le richieste di invio arrivano alla stazione secondo un processo di Poisson, con frequenza di 100 al minuto.

Determinare:

1. Il numero medio di richieste in attesa di essere servite;
2. Il tempo medio di attesa prima di essere serviti;
3. Se utilizzando un terzo thread è possibile dimezzare il tempo medio di attesa in coda.

# Esempio di Code M/M/s

Si tratta di un modello di code M/M/2. Assumendo come unità di tempo il minuto, si ha $\lambda = 100$ e $\mu = 60$. Inoltre, poiché risulta $\varrho = \frac{\lambda}{2\mu} = \frac{5}{6} < 1$, la condizione per l'esistenza della distribuzione stazionaria è verificata.

1. Per calcolare il numero medio di clienti in attesa di essere serviti $N^q$ abbiamo bisogno di calcolare $p_0$. Applicando la formula, si ottiene $p_0 = 1/11$, quindi $N^q = \frac{125}{33} = 3.78$.
2. $T^q = \frac{N^q}{\lambda} = 0.0378$ minuti
3. Con un terzo thread, la stazione diventa di tipo M/M/3 con $\varrho = 5/9$, e si ottiene $p_0 = 0.173$, $N^q = 0.374$ e $T^q = 0.00374$ e quindi il tempo di attesa medio di attesa in coda è ridotto a circa un decimo del precedente.

# Code M/M/s/K

Ad ogni utente che arriva nel sistema quando la coda è piena è negato l'accesso definitivamente (cliente perso). Questa situazione è nota con il nome di balking.

Consideriamo prima il caso di un singolo servente, si ha che con $\varrho \neq 1$

$$p_n = \begin{cases} \varrho^n p_0 = \varrho^n \left( \dfrac{1-\varrho}{1-\varrho^{K+1}} \right) & per\ n = 0, 1, \dots, K \\ 0 & per\ n > K \end{cases}$$

Per $\varrho = 1$ si ha che $p_n = \frac{1}{K+1}$. In queste condizioni:

$$N = \frac{\varrho}{1-\varrho} - \frac{(K+1)\varrho^{K+1}}{1-\varrho^{K+1}}$$

E $N^q = N\ - (1 - p_o)$, mentre per le altre grandezze si applica il teorema di Little ricordando che $\lambda_n$ non è costante.

# Code M/M/s/K

La probabilità $p_K$ è detta fattore di perdita, che va perduta perché non entra nel sistema. Definendo $\epsilon = 1 - p_K$ la frequenza media effettiva diventa:

$$\bar{\lambda} = \varepsilon\lambda, con\ \varepsilon < 1$$

Per la verifica della stazionarietà non è richiesto che $\varrho = \frac{\lambda}{\mu} < 1$.

# Esempio di Code M/M/s/K

Una stazione trasmittente ha un unico buffer e servente, e le richieste di trasmissione arrivano con un processo di Poisson a frequenza di 10 al minuto. Il tempo necessario alla trasmissione è distribuito esponenzialmente con valore medio pari a 2 secondi e il buffer può contenere 4 richieste in attesa, e non oltre. Le richieste che non trovano spazio nel buffer sono scartate.

Determinare:

1. Il numero medio di richieste nella stazione;
2. La probabilità che una richiesta possa essere soddisfatta;
3. Il tempo medio di attesa prima di avere la trasmissione.

# Esempio di Code M/M/s/K

La stazione di trasmissione è un sistema a coda M/M/1/4, assumendo come unità di tempo il minuto si ha che applicando le formule $\lambda = 10$, $\mu = 30$, $\varrho = 1/3$ e K = 4.

1. N = 0.4793;
2. $p_4 = \varrho^4 p_o = 1/3^4 \, p_o . p_o = \frac{1-\frac{1}{3}}{1-\frac{1}{3}^5} = 0.6694$, quindi $p_4 = 0.008264$.
3. $\bar{\lambda} = (1 - p_4)\lambda = 9.9173$ per avere $T = \frac{N}{\bar{\lambda}} = 0.04840$, da cui $T^q = T - \frac{1}{\mu} = 0.015$, ovvero il tempo medio di attesa è 54 secondi.

# Reti di Code

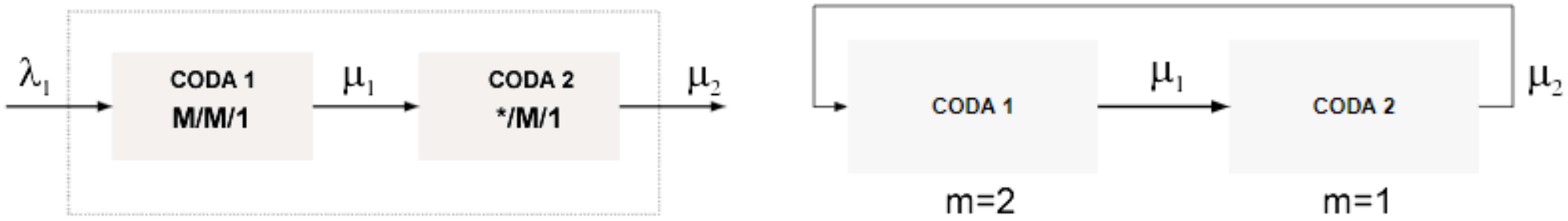
La trattazione della teoria delle code effettuata fino ad ora ha sempre considerato singoli sistemi a coda. Tuttavia, molto spesso i sistemi reali si possono presentare sotto forma di più sistemi a coda connessi fra di loro e si parla quindi di reti di code.

La trattazione della teoria delle code effettuata fino ad ora ha sempre considerato singoli sistemi a coda. Tuttavia, molto spesso i sistemi reali si possono presentare sotto forma di più sistemi a coda connessi fra di loro e si parla quindi di reti di code.

Le reti di code si dividono in

- reti aperte in cui sono possibili ingressi di utenti alla rete dall'esterno e uscite degli utenti dalla rete verso l'esterno;
- reti chiuse in cui il numero degli utenti all'interno della rete `e fissato e gli utenti circolano all'interno della rete senza che ci sia possibilità di ingressi dall'esterno o uscite verso l'esterno.

# Reti di Code




Per caratterizzare completamente una rete di code devono essere assegnati:

1. la topologia della rete;
2. le distribuzioni di probabilità dei tempi di inter-arrivo degli utenti presso i nodi che prevedono ingresso di utenti;
3. le distribuzioni di probabilità dei tempi di servizio presso ciascun nodo costituente la rete
4. le regole di istradamento.

# Teorema di Burke

Si consideri un sistema M/M/s con $s \geq 1$ con frequenza media di arrivo pari a $\lambda$ in condizioni di stazionarietà. Allora

1. il processo della partenze dal sistema è un processo di Poisson di parametro $\lambda$;
2. ad ogni istante di tempo t, il numero di utenti presenti nel sistema e indipendente dalla sequenza dei tempi di partenza prima di t.

Se gli utenti che escono dal un sistema M/M/s entrano in un altro sistema a coda, gli arrivi a questo secondo sistema saranno ancora Poissoniani.

In virtu' del teorema di Burke, possiamo collegare in una rete sistemi a coda M/M/s e, purché non ci siano cicli, ovvero purché gli utenti non possano rivisitare nodi già precedentemente visitati (reti feedforward), si può analizzare la rete decomponendola nodo a nodo. L'assenza di cicli è necessaria altrimenti si potrebbe perdere la natura Poissoniana dei flussi in ingresso ai nodi.

# Serie di Code

La più semplice rete di code che si può costruire consiste nell'avere un numero fissato (m) di sistemi a coda in serie, in cui non ci siano limiti sulla capacità della coda di ogni singolo sistema componente la rete.

In condizioni stazionarie, per il Teorema di Burke, ciascuno dei sistemi ha arrivi Poissoniani con parametro $\lambda$, e quindi i sistemi possono essere analizzati come tanti sistemi M/M/si isolati.

Il caso più semplice corrisponde ad avere due sistemi in serie (sistema tandem) con singolo servente. In questo caso è molto semplice dimostrare che la probabilità congiunta che $n_1$ utenti sono nel primo sistema e $n_2$ utenti nel secondo sistema è data dal prodotto delle singole probabilità:

$$P(n_1, n_2) = p_{n_1} p_{n_2} = \varrho_1^{n_1}(1-\varrho_1)\varrho_2^{n_2}(1-\varrho_2)$$

con $\varrho_1 = \frac{\lambda}{\mu_1} < 1$ e $\varrho_2 = \frac{\lambda}{\mu_2} < 1$.

# Serie di Code

Il tempo medio totale di permanenza nell'intero sistema, ovvero nella rete, e il numero medio di utenti presenti nella rete si possono calcolare semplicemente sommando le corrispondenti quantità calcolate in riferimento ai singoli sistemi.

Si osservi che le considerazioni fino ad ora riportate non valgono se la capacità dei singoli sistemi a coda componenti la rete fosse finita.

Una tipologia di reti di code molto studiate e che continua a prevedere l'utilizzo del modello M/M/s sono le cosiddette reti di Jackson aperte. A differenza dei Rete di sistemi a coda in serie, gli utenti visitano i nodi in un ordine qualsiasi e in ogni Jackson nodo ci possono essere utenti che arrivano sia dall'esterno sia da altri nodi.

Teorema di Jackson. Sia data una rete aperta di Jackson composta da m nodi ciascuno dei quali con $\mu_j$ e a singolo servente. Allora la distribuzione di probabilità congiunta che la rete di trovi allo stato $n = (n_1; \ldots ; n_m)$ si fattorizza nel prodotto delle distribuzioni di probabilità marginali.

# Serie di Code

In virtu' del Teorema di Jackson, per studiare una rete di code aperta di Jackson nel caso mono-servente è sufficiente determinare le frequenze medie effettive degli arrivi a ciascun nodo j date da $\lambda_j$ e analizzare indipendente ogni singolo nodo.

La condizione sotto la quale una rete di code ammette distribuzione stazionaria è che la “capacità del servizio” di ogni singolo nodo sia strettamente maggiore della frequenza media effettiva degli arrivi.

# Esempio Tandem di Code

Supponiamo di avere un sistema formato da due stazioni, una di trasmissione e una di ricezione, monoserventi in serie. Le richieste arrivano alla prima stazione secondo un processo di Poisson di parametro $\lambda$ = 10. I tempi di servizio dei serventi sono distribuiti esponenzialmente con $\mu_1$ = 12 e $\mu_1$ = 15.

Calcolare il tempo di permanenza nel sistema e il numero di richieste nel sistema.

# Esempio Tandem di Code

Per quanto esposto, gli arrivi alla seconda stazione sono Poissoniani di parametro $\lambda$ = 10 e l'analisi può essere condotta studiando singolarmente i due sistemi M/M/1. Poiché risulta

$$\varrho_1 = \frac{5}{6} < 1$$

$$\varrho_2 = \frac{2}{3} < 1$$

esiste una distribuzione stazionaria della rete. Analizzando i due sistemi singolarmente si ha

$$T_1 = \frac{1}{\mu_1 - \lambda} = 1/2\,, \quad T_2 = \frac{1}{\mu_2 - \lambda} = 1/5$$

$$N_1 = \lambda T_1 = 5, N_2 = \lambda T_2 = 2$$

Per quanto riguarda la rete si ha N = N1 + N2 = 7 richieste e T = T1 + T2 = 7/10 minuti, ovvero 42 secondi.