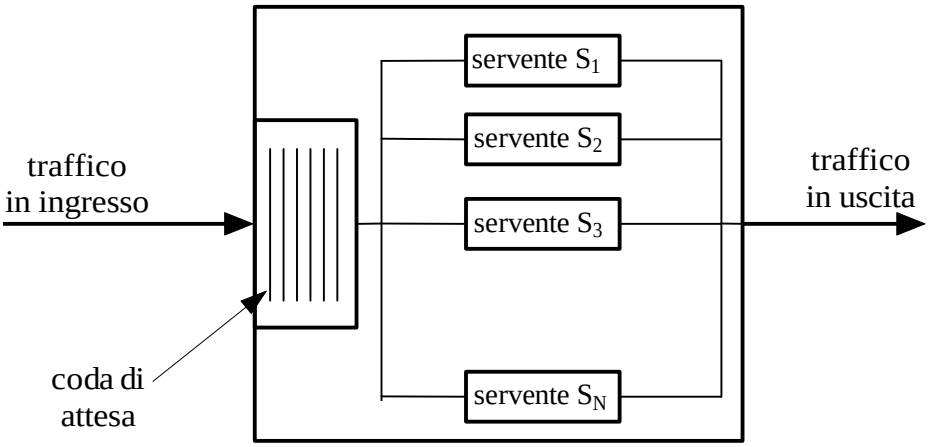
Appunti di Teoria dei Segnali

Capitolo 15 - I sistemi a coda

INTRODUZIONE

Per comprendere cosa sia un **sistema a coda**, pensiamo a quanto accade in un supermercato: ci sono una serie di clienti (che sono i cosiddetti "*utenti*"), i quali si mettono in fila (cioè in "*attesa di servizio*") allo scopo di passare da una delle casse (i cosiddetti "*serventi*") per pagare il conto (ossia ricevere il "*servizio*" desiderato).

Qualcosa di assolutamente analogo accade in un **sistema a coda**, che può essere schematizzato nel modo seguente:



Possiamo visualizzare il sistema a coda come un normale SISTEMA che, ricevendo "qualcosa" in ingresso, opera su questo "qualcosa" un certo numero di operazioni e genera una uscita. Nel caso del sistema a coda, l'ingresso, che prende il nome di **traffico in ingresso**, è costituito da tutte le richieste di servizio al sistema: ci sono cioè una serie di utenti che chiedono servizio al sistema e vengono da esso accettati (cioè *entrano* nel sistema). Il sistema dispone di un certo numero di componenti, che sono i cosiddetti "**serventi**", adibiti proprio a fornire i servizi richiesti. Dato che ciascun servente può servire una sola richiesta per volta, quindi 1 solo utente, è chiaro che *le richieste di servizio che possono essere soddisfatte contemporaneamente sono pari al numero di serventi.* Di conseguenza, se ci sono più richieste di quanti sono i serventi, le richieste in eccesso possono essere o respinte, nel qual caso si parla di **sistema con perdite**, oppure mantenute **in attesa** di essere servite, nel qual caso di parla di **sistema senza perdite** (o anche **sistema conservativo**). Una volta che un certo utente ha ricevuto il servizio richiesto, esso esce dal sistema e, insieme a tutti gli altri utenti che escono insieme a lui, forma il cosiddetto **traffico in uscita**.

Da questa descrizione, appare ovvio che, per definire in modo completo un sistema a coda, abbiamo bisogno di definire diversi parametri fondamentali:

- in primo luogo, dobbiamo conoscere il tipo di **traffico in ingresso**: nel caso più generale possibile, è ovvio che si tratterà di un processo stocastico, ma è anche possibile che invece si tratti di qualcosa di fisso e di determinato. Nel caso di un processo stocastico, servono le sue caratteristiche statistiche: il caso più frequenze è quello in cui tale traffico in ingresso è un processo stocastico di Poisson con intensità λ (dove λ rappresenta il numero medio di richieste di servizio nell'unità di tempo); per indicare questo, useremo, come simbolo formale che contraddistingue il traffico in ingresso, la lettera "**M**". Se, anziché avere un processo di Poisson, avessimo un processo deterministico, useremo la lettera "D"; se, infine, non avessimo né un processo di Poisson né un processo deterministico, useremo la lettera "G", per indicare che si tratta di una processo stocastico con distribuzione di probabilità nota ma generica;

Autore: Sandro Petrizzelli

- in secondo luogo, ci interessa conoscere il cosiddetto **tempo di servizio**, ossia *il tempo che ciascun servente impiega per fornire il servizio richiesto dall'utente;* a seconda delle caratteristiche del sistema, questo tempo di servizio potrà essere deterministico (ad esempio costante su un preciso valore) oppure aleatorio (cioè variabile di volta in volta con una precisa distribuzione di probabilità); noi assumeremo sempre che il tempo di servizio sia una variabile aleatoria; in particolare, il caso più frequente è quello di un tempo di servizio con distribuzione esponenziale: questo significa, come ben sappiamo, che si tratta di una variabile aleatoria senza memoria, ossia che, fissato un certo istante t, il tempo che ancora manca perché il servente termini il suo compito non dipende da quanto è successo prima dell'istante t. Quando il tempo di servizio ha distribuzione esponenziale, usiamo nuovamente la lettere "**M**"; quando invece questa variabile è deterministica, ossia conosciamo con precisione quanto essa vale, allora usiamo la lettere "D"; quando infine non conosciamo il suo valore, per cui è una variabile aleatoria propriamente detta, ma sappiamo anche che non ha distribuzione esponenziale, usiamo la lettera "D";

- ancora, un altro parametro fondamentale è ovviamente il **numero di serventi**, in quanto questa informazione ci serve a capire quanti utenti possono essere serviti contemporaneamente e quando è possibile che una richiesta di servizio non possa essere soddisfatta nel momento in cui arriva; si tratta di un valore deterministico, noto a priori;

- è importante anche conoscere la **capacità di memorizzazione** del sistema, ossia *il numero di utenti che possono essere contemporaneamente presenti nel sistema, siano essi sotto servizio o in attesa di servizio;* e' subito ovvio che il limite minimo di questa capacità è pari al numero di serventi (se così non fosse, ossia se la capacità di memorizzazione fosse inferiore al numero di serventi, noi avremmo sempre dei serventi inutilizzati e ciò non avrebbe alcun senso). In generale, quindi, la capacità di memorizzazione sarà pari o superiore (fino, teoricamente, ad ∞) rispetto al numero dei serventi; in particolare, dire che essa è superiore al numero di serventi significa dire che il sistema può mettere in attesa uno o più utenti (questi utenti sono all'interno del sistema ma non stanno ricevendo alcun servizio, in quanto sono in attesa di riceverlo);

- infine, un ultimo parametro importante, legato al traffico in ingresso, è il numero di utenti che "potenzialmente" possono chiedere il servizio al sistema a coda: per esempio, se noi consideriamo una *centrale telefonica* come un sistema a coda, è ovvio che questa centrale venga progettata e dimensionata non solo sulla base delle caratteristiche statistiche previste per il traffico in ingresso (ossia in base all'andamento delle richieste di servizio), ma anche in base al numero di utenti forniti di telefono, i quali sono perciò tutti potenziali utenti della centrale stessa. E' ovvio che un parametro legato a questo numero di potenziali utenti è il numero di serventi, che andrà opportunamente dimensionato perché la maggior parte delle richieste siano soddisfatte nel tempo minore possibile.

Vedremo più avanti come vengono specificati tutti questi parametri al fine di definire in modo completo un sistema a coda.

Da notare che gli ultimi due parametri (capacità di memorizzazione e potenziale numero di utenti complessivi) sono spesso talmente grandi da ritenerli ∞; in questi casi, essi non vengono specificati. Ad ogni modo, questo aspetto (più che altro formale) sarà chiaro più avanti.

LEGGE DI LITTLE

Prima di scendere nel dettaglio dell'esame dei principali sistema a coda, enunciamo e dimostriamo una importante legge legata ai sistemi in generale.

Consideriamo un generico sistema (non necessariamente un sistema a coda) caratterizzato da parametri assolutamente generici. Facciamo le seguenti posizioni:

- indichiamo con λ l'intensità del traffico in ingresso al sistema, il che significa che λ rappresenta il numero medio di utenti che chiedono servizio al sistema nell'unità di tempo;

- indichiamo inoltre con n il numero di utenti presenti contemporaneamente nel sistema, il che significa che tali utenti possono essere sia sotto servizio sia in attesa di servizio: in generale, si tratta di una variabile aleatoria, il che ci consente di dire che il suo valore medio $E[n]$ rappresenta il numero medio di utenti presenti complessivamente nel sistema;

- indichiamo infine con T il tempo di permanenza del generico utente nel sistema: si tratta cioè del **tempo di servizio**, cui si somma l'eventuale **tempo di attesa** (nel caso in cui sia previsto dal sistema). Anche in questo caso, abbiamo una variabile aleatoria, per cui il suo valore medio $E[T]$ rappresenta il tempo medio totale di permanenza degli utenti nel sistema.

Fatte queste premesse, la legge di Little afferma che queste tre grandezze sono legate dalla seguente relazione:

$$\boxed{E[n] = \lambda E[T]}$$

Vediamo di dare una giustificazione intuitiva di questo risultato.
Facciamo delle nuove posizioni:

- indichiamo con $A(t)$ il *numero di arrivi* (cioè di richieste di servizio) a partire dall'istante di osservazione, che supponiamo essere t=0, fino ad un generico tempo istante t; appare ovvio che questa quantità sia monotona crescente, in quanto aumenta di uno ad ogni arrivo e non c'è possibilità che diminuisca;
- indichiamo con $D(t)$ il numero di utenti serviti da t=0 fino all'istante t; anche questa quantità è monotona crescente;
- indichiamo con $N(t)$ il numero di utenti presenti nel sistema all'istante t; al contrario delle precedenti quantità, N(t) non necessariamente è monotona, in quanto dipende proprio da come variano A(t) e D(t).

Essendoci noi messi nell'ipotesi di un sistema a coda del tutto generico, queste tre quantità A(t), D(t) ed N(t) sono tutte dei processi stocastici. Ed è anche ovvio che esse siano legate dalla seguente relazione

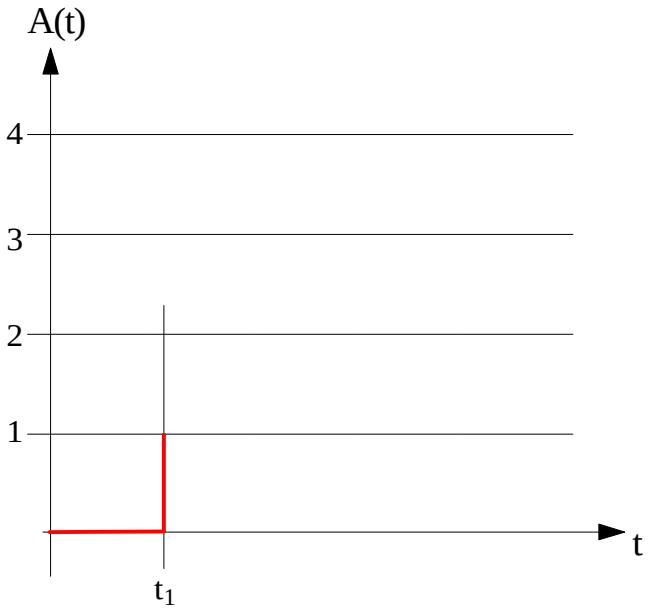
$$A(t) - D(t) = N(t)$$

Infatti, è chiaro che il numero di utenti N(t) presenti nel sistema all'istante t è pari al numero di richieste A(t) pervenute al sistema fino all'istante t, diminuito del numero di richieste soddisfatte D(t).

Quindi, N(t) risulta definito non appena definiamo A(t) e D(t); in altre parole, *se noi scegliamo una certa realizzazione del processo A(t) ed una certa*

realizzazione D(t), risulterà anche definita la corrispondente realizzazione di N(t).

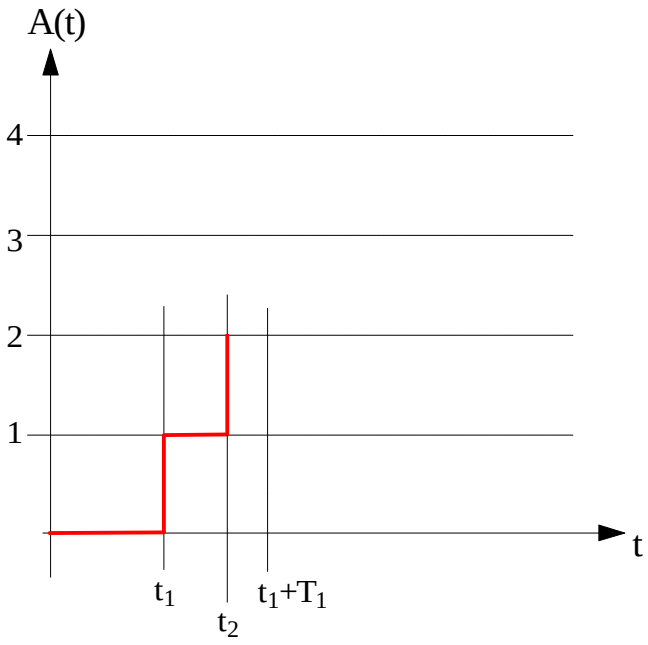
Supponiamo allora di considerare una particolare realizzazione del processo A(t) e ci aiutiamo con un grafico per descriverla. Riportiamo in ascisse il tempo t ed in ordinate il valore di A(t). Supponiamo che, inizialmente, il sistema non abbia ricevuto alcuna richiesta e che ad un certo istante t_1 giunga la prima richiesta di servizio al sistema: questo significa che il valore di A(t) passa da 0 ad 1 in corrispondenza dell'istante t_1:



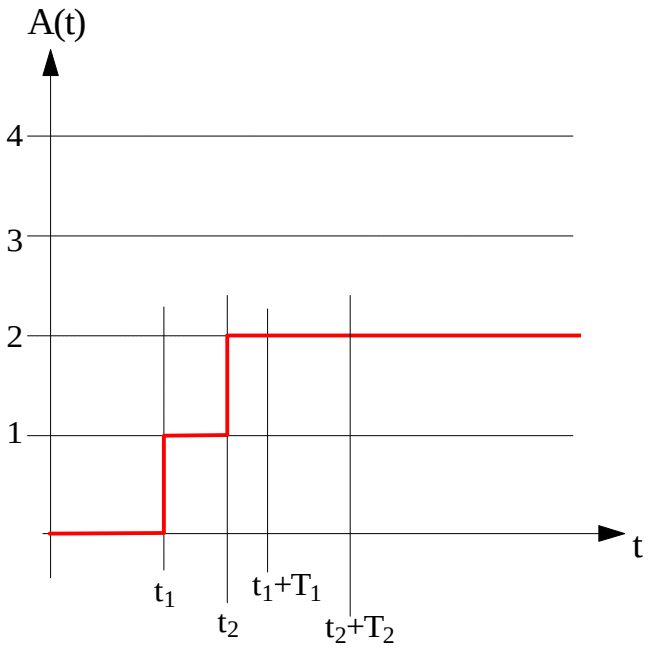
Dato che il sistema avrà evidentemente almeno 1 servente e dato che non c'era nessun utente già presente, la richiesta arrivata viene subito soddisfatta; indichiamo perciò con T_1 il tempo di permanenza di questo primo utente: è chiaro che, in questo caso, si tratta solo del tempo di servizio, in quanto non c'è nessuna attesa.

Se, durante e dopo questo intervallo di tempo T_1, non arrivano altre richieste di servizio, il valore di A(t) si mantiene costante, al fine proprio di indicare che, a partire dall'istante t_1, c'è stato un solo arrivo.

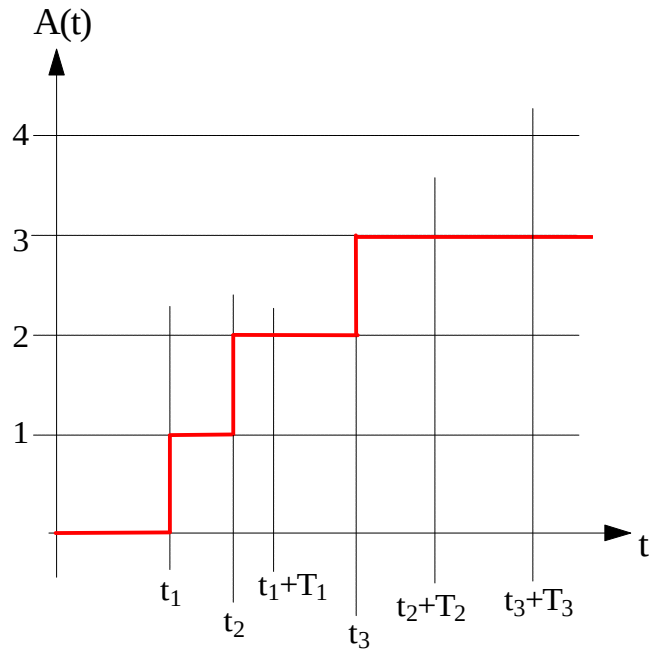
Supponiamo invece che, prima che sia passato l'intervallo di ampiezza T_1, quindi mentre ancora il primo utente è sotto servizio, arrivi una seconda richiesta: nell'ipotesi che il sistema sia senza perdite, questo secondo utente passerà un certo tempo T_2 nel sistema: se il sistema dispone di almeno 2 serventi, allora si tratta ancora una volta di semplice tempo di servizio; viceversa, se il servente è uno solo, avremo un tempo di attesa sommato poi al tempo di servizio. Ad ogni modo, in corrispondenza dell'istante t_2 di arrivo della seconda richiesta, il valore di A(t) passa ovviamente dal valore 1 al valore 2 (e questo anche se il sistema è con perdite, in quanto A(t) tiene conto solo del numero di richieste arrivate e non si preoccupa del fatto che una o più di esse sia stata respinta).



Qui possiamo ripetere lo stesso discorso di prima: nell’ipotesi che il tempo di permanenza di questo secondo utente sia T_2, se, durante o dopo questo intervallo di tempo T_2, non arrivano altre richieste di servizio, il valore di A(t) si mantiene costante:



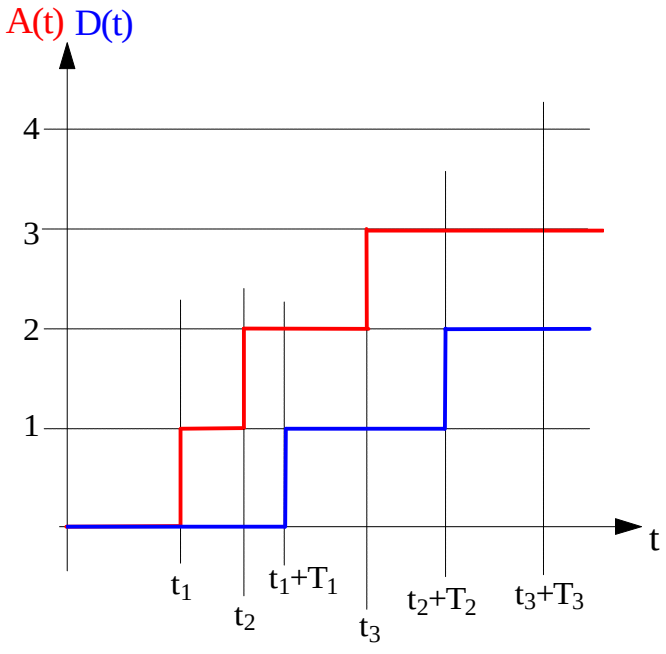
Al contrario, se, ad un certo istante t_3 (precedente o successivo l’istante t_2+T_2) arriva una terza richiesta, A(t) passa da 2 a 3 in corrispondenza di t_3 e così via, per cui si ottiene un grafico del tipo seguente:



In generale, man mano che arrivano altre richieste di servizio, A(t) aumenta, nel tempo, nel modo più o meno descritto nel grafico. Fissato un certo istante t, il valore di A(t) è evidentemente dato dal valore del gradino più alto.

Un'altra cosa da osservare è che, se noi, oltre alla realizzazione di A(t), conosciamo esattamente i valori dei tempi di permanenza T_1,T_2,T_3 e così via, siamo in grado di determinare anche la realizzazione di D(t): infatti, con riferimento al grafico di prima, è chiaro che il numero di utenti serviti fino all'istante t_1 è 0, in quanto proprio in t_1 è arrivata la prima richiesta di servizio e quindi prima di questo istante non può essere stato servito nessun utente; il numero di utenti serviti passa invece ad 1 in corrispondenza dell'istante t_1+T_1, quando cioè il primo utente esce del sistema; successivamente, D(t) passa a 2 all'istante t_2+T_2, quando anche il secondo utente è stato servito ed esce dal sistema; infine, D(t) passa a 3 in corrispondenza dell'istante t_3+T_3.

La realizzazione di D(t) è dunque quella disegnata in blu nel grafico seguente:



Ovviamente, come abbiamo detto prima, noti A(t) e D(t) in ogni istante t, è noto anche N(t), che è in ogni istante la differenza tra A(t) ed D(t): dal punto di vista grafico, N(t), fissato l'istante t, non è altro che la distanza tra la curva di A(t) e quella di D(t). Per esempio, all'istante t_2, N(t) vale 2, ossia ci sono due utenti contemporaneamente presenti nel sistema, mentre invece, all'istante t_3+T_3 e negli istanti successivi, N(t) vale 1.

A questo punto, detta appunto N(t) la realizzazione presa del processo stocastico che rappresenta il numero di utenti presenti contemporaneamente nel sistema all'istante t, è chiaro che N(t) è una funzione reale di variabile reale e quindi, come tale, essa possiede una "*media temporale*" (ossia il valore medio per unità di tempo): si tratta della quantità definita come

$$\langle N(t)\rangle_t = \frac{1}{t}\int_0^t N(\tau)d\tau$$

ed essa rappresenta dunque, relativamente alla realizzazione considerata, il numero medio di utenti presenti nel sistema all'istante t.

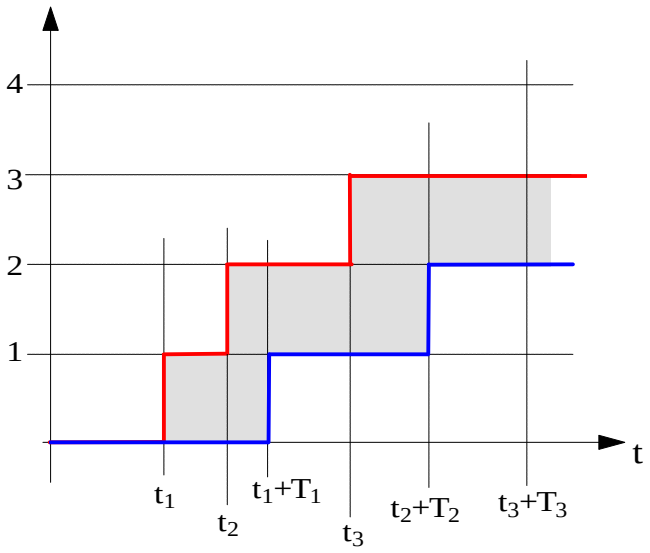
In modo analogo, se noi abbiamo indicato con λ l'intensità del traffico in ingresso e consideriamo una particolare realizzazione del processo stocastico che rappresenta tale traffico, possiamo indicare con $\langle\lambda\rangle_t$ la *media temporale del numero di utenti che chiedono il servizio al sistema*, ossia il numero medio di utenti che chiedono servizio nell'unità di tempo. E' ovvio che, essendo A(t) il numero di richieste di servizio pervenute fino all'istante t, sarà

$$\langle\lambda\rangle_t = \frac{A(t)}{t}$$

Da questa relazione si ricava che $t = \frac{A(t)}{\langle\lambda\rangle_t}$ e quindi anche, andando a sostituire nella espressione di $\langle N(t)\rangle_t$, che

$$\langle N(t)\rangle_t = \frac{\langle\lambda\rangle_t}{A(t)}\int_0^t N(\tau)d\tau$$

L'integrale che compare in questa formula non è altro che la somma, fino all'istante t, delle aree dei rettangoli che compaiono nel grafico di A(t) e D(t) messi insieme:



Tutti questi rettangoli hanno altezza unitaria e base pari ai rispettivi tempi di permanenza dei vari utenti, per cui la loro area è numericamente pari alla loro durata; inoltre, il numero di rettangoli, fino al generico istante t, è pari proprio ad A(t), per cui possiamo scrivere che

$$\langle N(t)\rangle_t = \frac{\langle\lambda\rangle_t}{A(t)}\int_0^t N(\tau)d\tau = \frac{\langle\lambda\rangle_t}{A(t)}\sum_{k=0}^{A(t)} T_k$$

Adesso, si osserva come il termine $\frac{1}{A(t)}\sum_{k=0}^{A(t)} T_k$ sia la somma dei tempi di permanenza, fino all'istante t, diviso per il numero totale di richieste di servizio fino all'istante t. Si tratta allora della *media temporale del tempo di permanenza*: posto allora

$$\langle T\rangle_t = \frac{1}{A(t)}\sum_{k=0}^{A(t)} T_k$$

possiamo scrivere che

$$\langle N(t)\rangle_t = \langle\lambda\rangle_t \langle T\rangle_t$$

A ben vedere, questa è proprio la **legge di Little**, applicata però ad una particolare realizzazione del traffico in ingresso e della permanenza nel sistema. Per passare da questa relazione a quella generale $E[n] = \lambda E[T]$, possiamo pensare di sfruttare il concetto di **ergodicità in media** dei processi stocastici: in particolare, è possibile dimostrare che *se il sistema a coda è stabile, risulta essere ergodico in media.*

Il fatto che sia ergodico in media significa che le medie temporali corrispondono alle medie di insieme con probabilità 1, il che quindi ci conferma la validità della legge di Little.

La cosa importante, che emerge dunque da questo discorso, è che la legge di Little vale solo per i sistemi a coda stabili: *dire che un sistema a coda è stabile equivale a dire che il traffico in uscita dal sistema è uguale al traffico in ingresso*[1].

Immaginando il traffico in uscita con un flusso di fluido che entra nel sistema e quello in uscita come un fluido che esce dal sistema, richiedere che il sistema sia stabile equivale a richiedere che il flusso in entrata sia pari a quello in uscita in ogni istante, ossia che non ci sia alcun accumulo di fluido all'interno del sistema.

FATTORE DI UTILIZZAZIONE

Possiamo definire un parametro che sia indice sia della stabilità del sistema sia anche della sua efficienza di funzionamento. Si definisce infatti **fattore di utilizzazione** il rapporto tra il numero medio di utenti che chiedono servizio al sistema nell'unità di tempo ed il numero medio di utenti che il sistema può servire nell'unità di tempo.

Per la simbologia adottata fino ad ora, il numero medio di utenti che chiedono servizio al sistema nell'unità di tempo corrisponde all'intensità λ del processo in ingresso; inoltre, se X è la variabile aleatoria corrispondente al tempo medio di servizio offerto dal sistema, allora la sua media E[X] è il

[1] Se il sistema non fosse stabile, ma instabile, potrebbe capitare una situazione del tipo seguente: supponiamo che il sistema riesca a servire, nell'unità di tempo, un numero medio di utenti pari a μ; se il numero medio λ di richieste di servizio nell'unità di tempo fosse maggiore di μ ed il sistema fosse instabile, il numero medio di utenti che permangono nel sistema tenderebbe a diventare infinito.

tempo medio di servizio e quindi il suo reciproco 1/E[X] è il numero medio di utenti che il sistema può servire nell'unità di tempo. Di conseguenza, il fattore di utilizzazione è

$$\rho = \frac{\lambda}{1/E[X]} = \lambda \cdot E[X]$$

In base a quanto detto poco fa, *dire che il sistema è* ***stabile*** *significa dire che deve necessariamente risultata* ***r<1****, in quanto solo questa condizione può consentire che il traffico medio in uscita sia pari a quello medio in ingresso.*

Esempio: sistema G/G/1/¥/¥

Possiamo fare immediatamente un esempio di applicazione della relazione $\rho=\lambda E[X]$. Consideriamo un sistema a coda di tipo molto generale e in particolare di tipo **G/G/1/∞/∞**: con questa simbologia, vogliamo indicare che il sistema ha un traffico in ingresso generico, un tempo di servizio generico, un solo servente, una capacità di memorizzazione infinita ed un numero infinito di potenziali utenti del sistema.

Vogliamo dimostrare che, per un siffatto sistema, il *fattore di utilizzazione* risulta essere

$$\boxed{\rho = 1 - p_0}$$

dove con p_0 indichiamo la probabilità che il sistema sia vuoto, ossia non ci sia alcun utente al suo interno (il sistema non fornisce alcun servizio).

Per fare questa dimostrazione, seguiamo una strada molto simile a quella seguita per dimostrare la legge di Little: consideriamo una particolare realizzazione del processo in ingresso, ricaviamo la tesi $\rho=1-p_0$ con riferimento a tale realizzazione e, infine, assumendo valida l'ergodicità del sistema, attribuiamo a tale tesi carattere del tutto generale.

Supponiamo che, durante un intervallo di tempo di durata τ, il traffico in ingresso sia caratterizzato da una intensità λ: essendo λ pari al numero medio di arrivi (o richieste di servizio) nell'unità di tempo, deduciamo che $\lambda\tau$ è il numero medio di arrivi nell'intervallo di durata τ.

Supponendo il sistema stabile ($\rho<1$), il numero medio di utenti in ingresso è pari al numero medio di utenti in uscita: quindi $\lambda\tau$ è anche il numero medio di utenti in uscita dal sistema nell'intervallo di durata τ.

Indichiamo ora con p_0 la probabilità che, in un istante generico, il sistema sia vuoto: con questa posizione, con riferimento ancora all'intervallo di tempo di durata τ, la quantità $(1-p_0)\tau$ rappresenta la frazione di tempo in cui il sistema non è vuoto; durante questo tempo, essendo presente nel sistema almeno un utente, l'unico servente presente starà erogando il proprio servizio, con un tempo medio di servizio pari a E[X]. Ciò significa che il numero medio di utenti in uscita è $\frac{(1-p_0)\tau}{E[X]}$.

D'altra parte, questo numero medio di utenti in uscita è stato prima identificato pari a $\lambda\tau$, per cui abbiamo l'uguaglianza

$$\frac{(1-p_0)\tau}{E[X]} = \lambda\tau$$

Da questa uguaglianza scaturisce la tesi che volevamo dimostrare: basta ricordarsi che $\rho=\lambda E[X]$.

Sistema a coda M/M/1

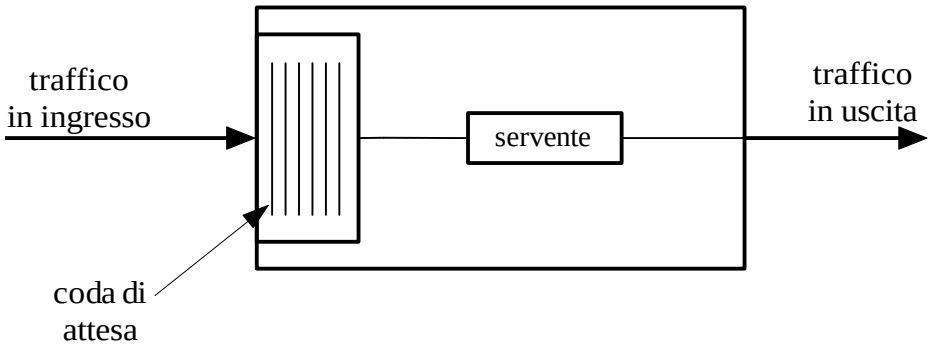
INTRODUZIONE: USO DELLE CATENE DI MARKOV

Passiamo adesso allo studio di alcuni dei più importanti sistemi a coda. In accordo a quanto abbiamo detto all'inizio, i parametri da specificare per definire in modo completo un sistema a coda sono i seguenti:

- tipo di traffico in ingresso (M per un processo di Poisson, D per un processo deterministico e G per un processo generico);
- distribuzione del tempo di servizio (M per la distribuzione esponenziale, D per una distribuzione deterministica e G per una distribuzione generica);
- numero di serventi (da 1 ad ∞);
- capacità di memorizzazione (da 1 ad ∞);
- utenti potenziali del sistema (da 1 ad ∞).

Quindi, per individuare il nostro sistema a coda, dobbiamo specificare, nell'ordine appena utilizzato, questi 5 parametri: per esempio, quando dire che si considera un sistema a coda di tipo M/M/1/7/∞ equivale a dire che il sistema riceve in ingresso un processo di Poisson (con una certa intensità λ), che il tempo di servizio è una variabile aleatoria con distribuzione esponenziale (con un certo parametro μ), che il sistema dispone di 1 solo servente, che il sistema è in grado di mantenere contemporaneamente dentro di sé 7 diversi utenti e che gli utenti che potenzialmente possono richiedere il servizio del sistema sono infiniti.

Il primo tipo di sistema che consideriamo è di tipo **M/M/1/∞/∞**, il che significa che il traffico in ingresso è un processo di Poisson (con intensità λ), che il tempo di servizio è una variabile aleatoria con distribuzione esponenziale (con parametro μ), che il sistema dispone di 1 solo servente, che il sistema è in grado di mantenere contemporaneamente dentro di sé ∞ utenti e che gli utenti potenziali sono ∞:



Talvolta, per semplicità, gli ultimi due parametri, essendo ∞, vengono omessi, per cui si parla semplicemente di sistema **M/M/1.**

La prima cosa da osservare riguarda la capacità di memorizzazione: dire che essa vale ∞ equivale a dire che il sistema può mantenere contemporaneamente dentro di sé infiniti utenti; in altre parole, nessuna eventuale richiesta di servizio viene rigettata dal sistema e, nel caso il servente sia già occupato, ciascuna richiesta viene posta in attesa per un certo tempo, che prende il nome di

tempo di attesa. Un sistema siffatto è un **sistema senza perdite** o **sistema conservativo**, il che equivale a dire che tutte le richieste di servizio vengono prima o poi soddisfatte.

Ovviamente, quando si modella un sistema reale tramite un sistema a coda, la capacità di memorizzazione non potrebbe essere ∞; d'altra parte, quando per esempio si stima che il numero di richieste è comunque basso o poco frequente, allora questa ipotesi di partenza diventa lecita oltre che conveniente (al fine di semplificare i calcoli).

La seconda osservazione riguarda invece gli "strumenti" che noi utilizziamo per studiare questo sistema. In particolare, ci chiediamo se è possibile studiare un sistema di questo tipo mediante una *catena di Markov a valori discreti*.

Prima ancora di verificare se questo sia possibile o meno, ricordiamo cos'è una catena di Markov: *una catena di Markov a valori discreti è un processo stocastico, tempo-discreto o tempo-continuo, a valori discreti, tale che, fissato un certo istante di osservazione t, l'evoluzione del processo successiva a tale istante dipende solo dallo "stato" del processo in t non dipende in alcun modo da quello che è successo negli istanti precedenti.*

I valori assumibili da parte del processo prendono spesso il nome di "*stati*" ed il processo stesso viene solitamente chiamato "*sistema*".

Per verificare se il nostro sistema a coda si possa considerare come una catena di Markov, devono essere dunque verificate due condizioni essenziali:

1. il sistema deve presentare solo un numero discreto di stati;
2. il sistema, fissato un certo istante di osservazione t, deve evolvere (passando in altri stati o rimanendo in quello iniziale) solo in base alla situazione in cui si trova nell'istante di osservazione e non in base a quello che è accaduto prima di tale istante.

La prima condizione è verificata banalmente se noi consideriamo, come "stati" del sistema a coda, il numero di utenti presenti in esso: in questo caso, infatti, potremo avere nel sistema 0 utenti, 1 utente, 2 utenti e così via fino teoricamente ad ∞ utenti. Quindi, d'ora in poi dire che il sistema si trova nello **stato k** equivale a dire che ci sono **k utenti presenti** al suo interno (siano essi sotto servizio e in attesa di servizio):

Inoltre, fare questa ipotesi di base significa anche che per "evoluzione" del sistema (cioè per passaggio del sistema in altri stati, incluso quello di partenza) noi intendiamo la comparsa e la scomparsa degli utenti, ossia la variazione del numero di utenti presenti all'interno del sistema stesso. Di conseguenza, la seconda condizione sarà verificata se noi dimostriamo che la variazione del numero di utenti presenti nel sistema, a partire dall'istante di osservazione, dipende solo dal numero di utenti in istante e non da quelli che erano presenti negli istanti precedenti.

Fissiamo dunque un istante generico di osservazione che indichiamo con t: in analogia a quanto fatto in precedenza, indichiamo con N(t) il numero di utenti presenti nel sistema all'istante t, con A(t) il numero di arrivi all'istante t e con D(t) il numero di utenti serviti sempre all'istante t. Dobbiamo dimostrare che N(t) dipende solo da t e non dagli istanti precedenti.

E' immediato dedurre che A(t) dipende solo da t e non dagli istanti precedenti: infatti, avendo supposto che il traffico in ingresso sia un processo di Poisson, sappiamo che il tempo di attesa , ossia il tempo che intercorre tra l'istante considerato e l'arrivo immediatamente successivo, ha distribuzione esponenziale, ossia è senza memoria, per cui A(t) non può che dipendere solo da t.

A questo, si aggiunge il fatto che anche D(t) dipende solo da t e non dagli istanti precedenti: il motivo sta nella ipotesi di partenza per cui il tempo di servizio, ossia il tempo necessario a ciascun servente per fornire il servizio richiesto, abbia distribuzione esponenziale; questa ipotesi fa sì che D(t) non dipenda dagli istanti precedenti t.

In conclusione, essendo N(t)=A(t)-D(t) e avendo dimostrato che A(t) e D(t) dipendono solo da t e non dagli istanti precedenti, è ovvio che lo stesso vale per N(t), per cui ***possiamo descrivere un sistema a coda di tipo M/M/1 mediante una catena di Markov.***

In effetti, questo risultato è generalizzabile: *un qualunque sistema a coda in cui il processo in ingresso sia un processo di Poisson ed il tempo di servizio abbia distribuzione esponenziale, è descrivile mediante una catena di Markov tempo-continua.*

Sulla base di ciò, possiamo sfruttare tutto quello che abbiamo detto circa le catene di Markov tempo-continue al fine di trarre una serie di importanti conclusioni circa i sistemi a coda che consideriamo.

FREQUENZE DI TRANSIZIONE DI STATO

I primi due concetti sulle catene di Markov tempo-continue che utilizziamo sono i seguenti:

- in primo luogo, abbiamo definito le cosiddette **probabilità asintotiche** come le probabilità di stato a regime: se $P(X(t)=j)$ è la probabilità che la catena si trovi nello stato j all'istante t, la corrispondente probabilità asintotica, ossia la probabilità che la catena, a regime (cioè per $t\to\infty$) si trovi nello stato j, è $p_j = \lim_{t\to\infty} P(X(t)=j)$;

- in secondo luogo, abbiamo definito le cosiddette **frequenze di transizione di stato**: la generica frequenza γ_{ij} è il numero di volte in cui il sistema passa dallo stato i allo stato j nell'unità di tempo.

Le probabilità asintotiche e le frequenze di transizione di stato sono legate dalla seguente relazione

$$p_j \sum_{i\neq j} \gamma_{ji} = \sum_{i\neq j} p_i \gamma_{ij} \qquad \forall j$$

dove j indica uno qualsiasi dei possibili stati della catena.

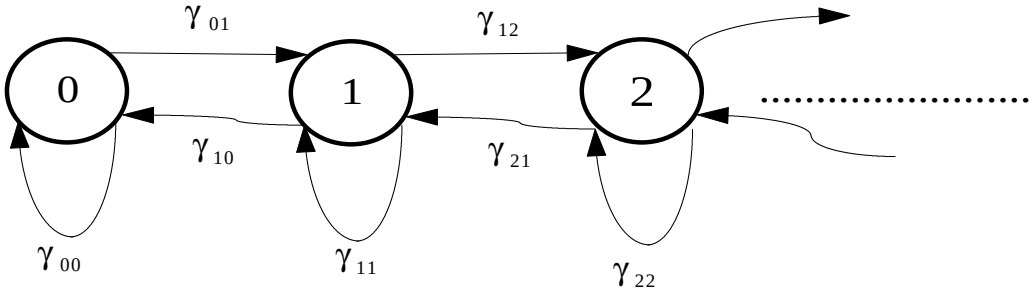
Vediamo allora come utilizzare questi concetti nel caso di un sistema a coda.

In primo luogo, quanti stati ha il nostro sistema? Considerando che si tratta di un sistema di tipo M/M/1/∞/∞, ossia di un sistema con capacità infinita di memorizzazione, e considerando che ciascuno stato corrisponde al numero di utenti presenti contemporaneamente nel sistema, è chiaro che il numero di stati è ∞: lo stato 0 corrisponde a 0 utenti presenti nel sistema lo stato 1 ad un solo utente e così via fino ad ∞.

La generica probabilità asintotica $p_j = \lim_{t\to\infty} P(X(t)=j)$ rappresenta la probabilità che, in condizioni di regime, ci siano j utenti nel sistema; la generica frequenza di transizione di stato γ_{ij} rappresenta invece il numero di volte in cui nel sistema, nell'unità di tempo, si passa da i utenti a j utenti presenti.

Diagramma delle frequenze di transizione di stato

Dalle catene di Markov possiamo anche prendere il modo grafico con cui abbiamo inteso rappresentare gli stati del sistema e le varie frequenze di transizione; si tratta di quello che abbiamo chiamato **diagramma delle frequenze di transizione di stato**, che qui riportiamo:



Ovviamente, il diagramma comprende in linea teorica infiniti stati e comprende anche le frequenze di transizione tra stati non adiacenti. Il motivo per cui queste ultime non sono state indicate sarà chiaro tra poco: faremo infatti vedere che, per il sistema considerato, esse sono nulle.

Determinazione delle frequenze di transizione di stato

Per determinare le frequenze di transizione di stato, note che siano le probabilità asintotiche, non dobbiamo far altro che risolvere il sistema rappresentato dall'equazione

$$p_j \sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} p_i \gamma_{ij} \qquad \forall j$$

Ovviamente, dato che j può assumere infiniti valori compresi tra 0 ed ∞, la risoluzione di questo sistema può essere effettuata solo a livello analitico e non a livello numerico (ossia col calcolatore). Possiamo tuttavia arrivare alla determinazione dei termini γ_{ij} usando un approccio intuitivo più che analitico.

Supponiamo ad esempio di volere $\gamma_{i,i+1}$, ossia la frequenza di transizione dallo stato i allo stato i+1. Intanto, quando il sistema passa dallo stato i allo stato i+1? Lo stato i+1 corrisponde a dire che c'è un utente in più nel sistema rispetto allo stato i, per cui il passaggio da i ad i+1, tenendo conto che nessuna richiesta viene mai respinta, avviene nel momento in cui arriva una nuova richiesta di servizio[2].

Indicato allora con A(t) il numero di arrivi in un intervallo di ampiezza t, ci chiediamo quanto valga $P(A(\delta)=1)$, ossia la probabilità che, in un intervallo di ampiezza δ generica, ci sia 1 solo arrivo.

Se il sistema è di tipo M/M/1/∞/∞, il traffico in ingresso è un processo di Poisson (di intensità λ), per cui, per valutare quella probabilità, possiamo usare la formula di Poisson:

$$P(A(\delta)=1) = \frac{\lambda\delta}{1!} e^{-\lambda\delta} = \lambda\delta e^{-\lambda\delta}$$

Se supponiamo che l'intervallo δ sia piccolo, possiamo sviluppare in serie il termine esponenziale:

$$P(A(\delta)=1) = \lambda\delta e^{-\lambda\delta} = \lambda\delta(1-\lambda\delta+o(\delta))$$

[2] Si tenga presente che, avendo il sistema una capacità di memorizzazione infinita, tutte le richieste di servizio vengono accettate, ossia si traducono in utenti che entrano nel sistema. Si tratta poi di vedere se la generica richiesta viene servita, nel qual caso il sistema era precedente vuoto, oppure se viene messa in attesa, nel qual caso il sistema comprendeva già almeno un utente.

da cui, quindi, assumendo che $\lambda^2\delta^2$ sia anch'esso un infinitesimo di ordine superiore, possiamo concludere che

$$P(A(\delta)=1)=\lambda\delta+o(\delta)$$

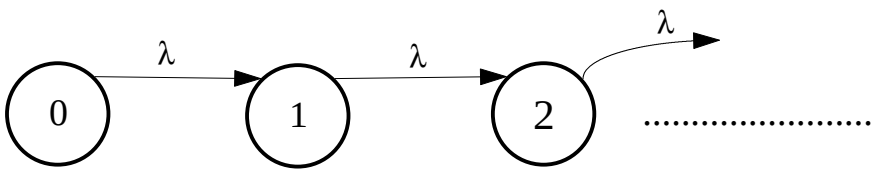
Per ottenere adesso la frequenza di transizione $\gamma_{i,i+1}$, non dobbiamo far altro che calcolare il limite, per $\delta\to 0$, di questa quantità:

$$\gamma_{i,i+1}=\lim_{\delta\to 0}P(A(\delta)=1)=\lim_{\delta\to 0}(\lambda\delta+o(\delta))=\lambda$$

La conclusione è dunque che

$$\boxed{\gamma_{i,i+1}=\lambda}$$

Questo risultato è molto importante se si considera che la frequenza di transizione $\gamma_{i,i+1}$ non dipende in alcun modo da quale sia lo stato i, il che significa che, *dati due stati adiacenti qualsiasi, la frequenza di transizione da uno al successivo è sempre pari a l.*



Vediamo adesso quanto vale $\gamma_{i,i+2}$, ossia la frequenza di transizione dallo stato i allo stato i+2. Intanto, il sistema passa dallo stato i allo stato i+2 quando arrivano contemporaneamente due diverse richieste di servizio. Di conseguenza, noi dobbiamo calcolare il limite, sempre per $\delta\to 0$, di $P(A(\delta)=2)$, che è la probabilità che, in un intervallo di ampiezza δ generica, ci siano 2 arrivi.

Usando ancora una volta la formula di Poisson, abbiamo che

$$P(A(\delta)=2)=\frac{(\lambda\delta)^2}{2!}e^{-2\lambda\delta}$$

Anche senza ricorrere allo sviluppo in serie dell'esponenziale, è chiaro che $P(A(\delta)=2)$ è la somma di infinitesimi tutti di ordine superiore, il che ci consente di concludere subito che

$$\gamma_{i,i+2}=0$$

Quindi, è nulla la frequenza di transizione da uno stato a due stati successivi. E' intuitivo accorgersi che valga il seguente risultato generale:

$$\boxed{\gamma_{i,i+k}=0 \qquad \forall k\geq 2}$$

Sono dunque nulle le frequenze di transizione tra uno stato generico i ed uno stato i+k ad esso successivo che però non sia ad esso consecutivo (cioè con $k\geq 2$).

La conclusione che traiamo da questi discorsi è la seguente: *in un sistema a coda di tipo M/M/M/¥/¥, il sistema passa da uno stato a quello successivo con frequenza di transizione pari a l (intensità del traffico in ingresso), mentre non è in grado di passare ad uno stato ad un altro che non sia appunto il successivo.*

Detto in termini di numero di utenti, il sistema, trovandosi, in un certo istante, con i utenti dentro di esso, può passare ad i+1 utenti con frequenza pari a λ, mentre non può assolutamente passare a i+2,i+3,.... utenti.

Adesso dobbiamo fare gli stessi ragionamenti per i passaggi di stato "all'indietro", ossia per i passaggi dallo stato generico i allo stato precedente i-1, a due precedenti i-2 e così via.

In particolare, cominciamo a valutare $\gamma_{i,i-1}$, ossia la frequenza di transizione dallo stato i allo stato precedente $i-1$. Intanto, il sistema passa dallo stato i allo stato i-1 quando esso termina di servire un utente, per cui quest'ultimo esce dal sistema e quindi decrementa di 1 il numero di utenti presenti complessivamente nel sistema stesso. Per analizzare questa situazione, indichiamo con τ *il tempo residuo di servizio da dedicare all'utente sotto servizio nell'istante considerato*: in altre parole, considerato un istante t e considerato un certo utente, τ indica quanto tempo tale utente deve rimanere ancora nel sistema perché possa uscirne. Allora, al fine di valutare $\gamma_{i,i-1}$, noi dobbiamo calcolare il limite, sempre per $\delta\to 0$, di $P(\tau \le \delta)$, che è la probabilità che ci voglia un intervallo di ampiezza δ generica perché l'utente considerato riceva il servizio e esca dal sistema.

Per calcolare $P(\tau \le \delta)$ sfruttiamo evidentemente il concetto di "tempo di servizio" del sistema, ossia il tempo di cui necessità ciascun servente del sistema per fornire il servizio richiesto. L'ipotesi che stiamo facendo è che il sistema a coda considerato abbia tempo di servizio con distribuzione esponenziale e parametro μ, il che significa che possiamo scrivere

$$P(\tau \le \delta) = 1 - e^{-\mu\delta}$$

dove ricordiamo che μ rappresenta la cosiddetta **frequenza di servizio** del generico servente (ossia il reciproco del **tempo medio di servizio**): si tratta in pratica del *numero medio di utenti serviti nell'unità di tempo*.

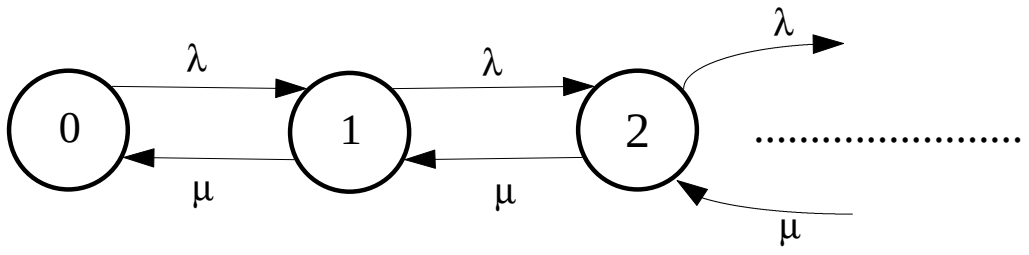
Sviluppando in serie quel termine esponenziale, abbiamo che

$$P(\tau \le \delta) = 1 - (1 - \mu\delta + o(\delta)) = \mu\delta + o(\delta)$$

da cui si deduce che $\lim_{\delta\to 0} P(\tau \le \delta) = \mu$, ossia quindi che

$$\boxed{\gamma_{i,i-1} = \mu}$$

Ancora una volta, abbiamo trovato una frequenza di transizione costante qualche che sia lo stato i di partenza: ciò significa che, *dati due stati adiacenti qualsiasi, la frequenza di transizione da uno al precedente è sempre pari a m.* Possiamo dunque ulteriormente perfezionare il diagramma delle frequenze di transizione:



Il passo successivo consiste nel calcolare $\gamma_{i,i-2}$, ossia la frequenza di transizione dallo stato i allo stato i-2: il sistema passa dallo stato i allo stato i-2 quando termina di servire due diversi utenti; tuttavia, il sistema ha un solo servente, che può servire un utente per volta, per cui è escluso che 2 o più utenti escano contemporaneamente dal sistema.

Deduciamo quindi anche in questo caso che

$$\boxed{\gamma_{i,i-k} = 0 \qquad \forall k \geq 2}$$

Abbiamo cioè trovato che le frequenze di transizioni all'indietro sono non nulle (ma pari a μ) solo tra stati adiacenti.

Mettendo insieme con quanto trovato prima circa le frequenze di transizione in avanti, possiamo concludere che *in un sistema a coda di tipo M/M/1/¥/¥, i passaggi di stato, sia in avanti sia indietro, sono consentiti SOLO tra stati adiacenti (con frequenza λ in avanti e μ indietro) .*

DETERMINAZIONE DELLE PROBABILITÀ ASINTOTICHE

Questo risultato, ossia la conoscenza di tutte le frequenze di transizione di stato, ci consente di calcolare le probabilità asintotiche del sistema: infatti abbiamo detto prima che tali probabilità sono legate alle frequenze di transizione dalla relazione

$$p_j \sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} p_i \gamma_{ij} \qquad \forall j$$

che rappresenta un sistema in un numero di equazioni pari al numero di stati possibili del sistema (che in questo caso sono ∞).

Nel caso del sistema a coda che stiamo considerando, abbiamo detto che

$$\gamma_{i,i+1} = \lambda$$
$$\gamma_{i,i-1} = \mu$$
$$\gamma_{i,i-k} = \gamma_{i,i+k} = 0 \qquad \forall k \geq 2$$

per cui quel sistema, al variare di j, si semplifica notevolmente.

Cominciamo dal caso in cui j=0: l'equazione è

$$p_0 \sum_{i \neq 0} \gamma_{0i} = \sum_{i \neq 0} p_i \gamma_{i0}$$

Al primo membro, l'unico valore di i per cui abbiamo una frequenza di transizione non nulla è i=1 e lo stesso vale per il secondo membro, per cui l'equazione è

$$p_0 \gamma_{01} = p_1 \gamma_{10}$$

Sappiamo poi che

$$\gamma_{i,i+1} = \lambda$$
$$\gamma_{i,i-1} = \mu$$

per cui $p_0\lambda = p_1\mu$.

Passiamo a j=1: l'equazione è

$$p_1 \sum_{i\neq 1} \gamma_{1i} = \sum_{i\neq 1} p_i \gamma_{i1}$$

Al primo membro, gli unici valori consentiti sono 0 e 2 e lo stesso vale per il secondo membro: abbiamo dunque che

$$p_1(\gamma_{10} + \gamma_{12}) = (p_0\gamma_{01} + p_2\gamma_{21})$$

e quindi anche che

$$p_1(\mu + \lambda) = (p_0\lambda + p_2\mu)$$

Vediamo ora per j=2: l'equazione è

$$p_2 \sum_{i\neq 2} \gamma_{2i} = \sum_{i\neq 2} p_i \gamma_{i2}$$

e essa diventa evidentemente

$$p_2(\gamma_{21} + \gamma_{23}) = (p_1\gamma_{12} + p_3\gamma_{32})$$

ossia anche

$$p_2(\mu + \lambda) = (p_1\lambda + p_3\mu)$$

Potremmo anche proseguire (in teoria fino a j=∞), ma le tre equazioni ricavate sono sufficienti per far vedere quello che ci interessa:

$$p_0\lambda = p_1\mu$$
$$p_1(\mu + \lambda) = (p_0\lambda + p_2\mu)$$
$$p_2(\mu + \lambda) = (p_1\lambda + p_3\mu)$$

Infatti, dalla prima equazione si ricava evidentemente che $p_1 = \frac{\lambda}{\mu} p_0$. Sostituendo questa nella seconda equazione e facendo qualche manipolazione algebrica, si ricava poi che

$$p_2 = \frac{\lambda}{\mu} p_1 = \left(\frac{\lambda}{\mu}\right)^2 p_0$$

Sostituendo questa nella terza equazione e facendo altre manipolazioni, si ricava infine che

$$p_3 = \frac{\lambda}{\mu} p_2 = \left(\frac{\lambda}{\mu}\right)^3 p_0$$

E' evidente, dunque, il seguente risultato fondamentale:

$$\boxed{p_k = \left(\frac{\lambda}{\mu}\right)^k p_0}$$

Questa relazione dice che *tutte le probabilità asintotiche dipendono, secondo il rapporto l/m elevato ad una opportuna potenza, dalla probabilità asintotica dello stato j=0.* Per calcolare questa probabilità asintotica, utilizziamo la **condizione di normalizzazione**: imponiamo cioè che risulti

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

Sostituendo la relazione trovata prima, risulta

$$\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i p_0 = 1$$

e da qui si ricava evidentemente che

$$p_0 = \frac{1}{\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i}$$

Nell'ipotesi che il rapporto λ/μ sia minore di 1, quella è semplicemente la somma della serie geometrica, per cui possiamo ulteriormente scrivere

$$p_0 = \frac{1}{\frac{1}{1-\frac{\lambda}{\mu}}} = 1-\frac{\lambda}{\mu}$$

A ben guardare, il rapporto λ/μ non è altro che il **fattore di utilizzazione** definito in precedenza: infatti, λ è il numero medio di utenti che chiedono servizio al sistema nell'unità di tempo, mentre μ è il numero medio di utenti serviti dal sistema nell'unità di tempo. Poniamo allora $\rho = \frac{\lambda}{\mu}$. A questo punto, ci ricordiamo che la condizione $\rho<1$ (cioè $\lambda<\mu$) equivale a richiedere un sistema stabile[3]. Possiamo concludere che, se il sistema considerato è stabile, le probabilità asintotiche valgono

$$\boxed{p_k = \rho^k(1-\rho)}$$

[3] Infatti, dire che $\mu>\lambda$ significa dire che il sistema può servire mediamente, nell'unità di tempo, più utenti di quanti chiedono servizio al sistema, il che è condizione necessaria affinché il traffico medio in uscita possa essere uguale al traffico medio in ingresso, ossia appunto alla stabilità del sistema. Se non fosse così, se cioè fosse $\lambda>\mu$, il sistema non riuscirebbe a rimanere stabile, in quanto il numero di utenti al suo interno si accumulerebbe indefinitamente, diventando perciò instabile.

Questa formula vale dunque se il sistema è stabile. In caso di sistema instabile, invece, la relazione da considerare è quella generica, ossia

$$p_0 = \frac{1}{\sum_{i=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^i}$$

e quindi

$$p_k = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{\sum_{i=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^i}$$

NUMERO MEDIO DI UTENTI PRESENTI NEL SISTEMA A REGIME

La conoscenza delle probabilità asintotiche ci consente di fare un ulteriore importante calcolo: se indichiamo con N la variabile aleatoria che indica il numero di utenti presenti complessivamente nel sistema in un certo istante, è chiaro che il suo valor medio $E[N]$ corrisponde al **numero medio di utenti presenti nel sistema** nel dato istante: applicando semplicemente la definizione di media di una variabile aleatoria, esso vale

$$E[N] = \sum_{i=0}^{\infty} i p_i$$

dove, dato che usiamo le probabilità asintotiche, ci stiamo ovviamente riferendo alla condizione di regime.

Sostituendo l'espressione ricavata prima per la generica probabilità asintotica nell'ipotesi in cui $\rho<1$, abbiamo dunque che

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

Essendo $\rho<1$, possiamo concludere che

$$E[N] = \sum_{i=0}^{\infty} i(\rho)^i(1-\rho) = (1-\rho)\frac{\rho}{(1-\rho)^2}$$

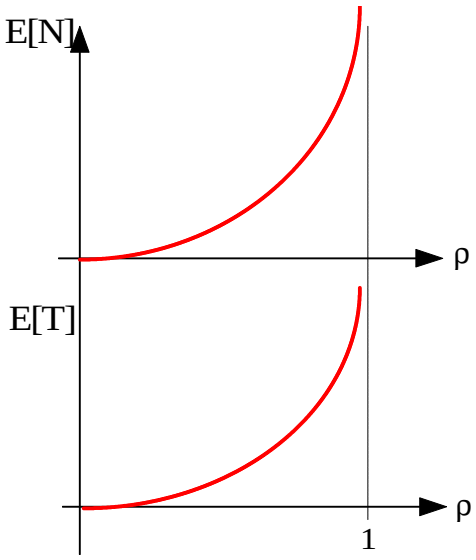
ossia

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

Possiamo a questo punto utilizzare la *legge di Little* $E[N] = \lambda E[T]$, in cui T rappresenta il tempo di permanenza totale di un utente nel sistema. Per quanto trovato poco fa su E[N], possiamo concludere che il **tempo medio di permanenza** del generico utente nel sistema vale

$$\boxed{E[T] = \frac{\rho/\lambda}{1-\rho} = \frac{1}{\mu-\lambda}}$$

Può essere interessante diagrammare sia E[N] sia E[T] in funzione di ρ, ovviamente facendo l'ipotesi che ρ vari tra 0 ed 1 (sistema stabile); in base alle relazioni appena ottenute, abbiamo grafici del tipo seguente:



L'andamento è chiaramente lo stesso in quanto le due quantità sono legate dal fattore λ.

Distribuzione del tempo di permanenza

In questo paragrafo, vogliamo continuare l'analisi del sistema M/M/1/∞/∞ e, in particolare, vogliamo ricavare la funzione densità di probabilità del tempo di permanenza T del generico utente nel sistema. Indichiamo tale funzione con $f_T(x)$.

Per determinare $f_T(x)$, dobbiamo intanto stabilire quale sia la cosiddetta *disciplina di coda* del sistema: per **disciplina di coda** si intende il criterio con cui si sceglie, tra gli utenti presenti nella coda di attesa, il prossimo da servire. Il caso più semplice, che viene adottato in tutti i sistemi a coda, è quello di una disciplina di tipo **FCFS** (che sta per *First Come First Served*): gli utenti vengono serviti nell'ordine con cui sono arrivati.

Premesso questo, consideriamo un generico utente che faccia richiesta di servizio al sistema; dato che il sistema ha una capacità di memorizzazione infinita, l'utente viene accettato nel sistema, per cui da qui parte il suo tempo di permanenza nel sistema; se il sistema è vuoto, allora l'utente viene immediatamente servito, per cui il tempo di permanenza si riduce al solo tempo di servizio; se, invece, nel sistema è già presente almeno un utente, allora il nuovo utente dovrà mettersi in attesa.

Supponiamo allora che il nuovo utente, entrando nel sistema, trovi altri k utenti, di cui uno è sotto servizio e gli altri k-1 sono in attesa (e saranno serviti prima di lui). Dato che il tempo di servizio è di tipo esponenziale ed il servizio al generico utente è indipendente dal servizio agli altri utenti, deduciamo che il tempo di permanenza (attesa+servizio) del nuovo utente sarà la somma dei tempi di servizio degli altri k utenti nonché del suo.

Tutti questi tempi di servizio hanno, per ipotesi, una distribuzione di tipo esponenziale con parametro μ, per cui la loro somma avrà una **distribuzione di Erlang di grado k+1 e con parametro μ** ([4]):

$$f_T(x) = \frac{\mu^{k+1}}{k!} x^k e^{-\mu x} \qquad x > 0$$

[4] Il risultato per cui la somma di variabili indipendenti esponenziali è una variabile di Erlang è stato dimostrato in precedenza

E' opportuno precisare che questo risultato è condizionato al fatto che ci fossero già k utenti nel sistema quando è arrivato l'utente da noi considerato. Quindi, la densità di probabilità da considerare è di tipo condizionato: se N_a è la variabile aleatoria che dà il numero di utenti già presenti nel sistema, dobbiamo scrivere, a rigore, che

$$f_T\left(x\middle|N_a = k\right) = \frac{\mu^{k+1}}{k!} x^k e^{-\mu x} \qquad x > 0$$

Per calcolare la densità di probabilità propriamente detta, dobbiamo considerare tutti i possibili valori di k (quindi da 0 a +∞) e le rispettive probabilità: dobbiamo cioè scrivere che

$$f_T(x) = \sum_{k=0}^{\infty} f_T\left(x\middle|N_a = k\right) \cdot P\left(N_a = k\right) = \sum_{k=0}^{\infty} \frac{\mu^{k+1}}{k!} x^k e^{-\mu x} \cdot P\left(N_a = k\right) = e^{-\mu x} \mu \sum_{k=0}^{\infty} \frac{\mu^k}{k!} x^k \cdot P\left(N_a = k\right)$$

Ci serve adesso conoscere $P\left(N_a = k\right)$, ossia la probabilità che il nuovo utente, entrando nel sistema, trovi già un numero k di altri utenti presenti. Non esiste, in proposito, un risultato generale. Tuttavia, nel caso del sistema M/M/1/∞/∞ che noi stiamo considerando, si può dimostrare che, essendo ∞ il numero di utenti potenziale per il sistema, $P\left(N_a = k\right)$ coincide proprio con la probabilità di stato $p_k = \rho^k(1-\rho)$. Sostituendo, abbiamo perciò che

$$f_T(x) = e^{-\mu x} \mu \sum_{k=0}^{\infty} \frac{\mu^k}{k!} x^k \cdot \rho^k (1-\rho) = e^{-\mu x} \mu (1-\rho) \sum_{k=0}^{\infty} \frac{(\mu \rho x)^k}{k!} = e^{-\mu x} \mu (1-\rho) \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} = e^{-\mu x} \mu (1-\rho) \cdot e^{\lambda x}$$

Scrivendo in maniera più opportuna quanto trovato, concludiamo che

$$\boxed{f_T(x) = (\mu - \lambda) e^{-(\mu - \lambda)x}}$$

Da questa espressione deduciamo che *il tempo totale di permanenza del generico utente nel sistema è ancora di tipo esponenziale, ma con parametro m-l.*

Da qui deduciamo anche che il tempo medio di permanenza è $\frac{1}{\mu - \lambda}$, così come avevamo trovato prima per altre vie.

CONDIZIONE DI STABILITÀ DEL SISTEMA

Sulla base dei risultati ottenuti, possiamo adesso controllare che sia verificata la condizione di **stabilità** del sistema, che corrisponde a dire che il traffico in uscita è uguale a quello in ingresso.

Se indichiamo con γ l'**intensità del traffico in uscita**, ossia il numero medio di utenti che escono dal sistema nell'unità di tempo, possiamo cioè far vedere che

$$\boxed{\gamma = \lambda}$$

dove λ è l'intensità del traffico in ingresso, ossia il numero medio di utenti che entrano nel sistema nell'unità di tempo.

Quanto vale γ? γ è pari al numero medio di utenti serviti dal sistema nell'unità di tempo, che abbiamo indicato con μ, per la probabilità che nel sistema ci sia almeno 1 utente: questa probabilità vale ovviamente $1-p_0$, per cui possiamo scrivere che

$$\gamma = \mu(1-p_0)$$

Sostituendo al posto di p_0 l'espressione trovata in precedenza nell'ipotesi che ρ<1, abbiamo che

$$\gamma = \mu(1-(1-\rho)) = \mu\sigma = \lambda$$

TEMPO MEDIO DI ATTESA E NUMERO MEDIO DI UTENTI IN ATTESA

Continuando nella ricerca e nella determinazione di parametri caratteristici relativi al sistema a coda in esame, proviamo adesso a valutare il **tempo medio di attesa** del generico utente nel sistema: si tratta cioè del tempo medio che ciascun utente deve attendere, a partire dal momento in cui fa richiesta di servizio, per ricevere effettivamente tale servizio.

Sappiamo già che questo tempo medio di attesa, che indichiamo con $E[W]$ (dove W è ovviamente la variabile aleatoria corrispondente al tempo di attesa) è legato al tempo medio di permanenza del sistema $E[T]$ ed al tempo medio di servizio, che abbiamo detto essere E[X]=1/μ, dalla relazione

$$E[T] = E[W] + E[X] = E[W] + \frac{1}{\mu}$$

Nei paragrafi precedenti abbiamo trovato che il tempo medio di permanenza vale $E[T] = \frac{1}{\mu-\lambda}$, per cui

$$E[W] = \frac{1}{\mu} - E[T] = \frac{1}{\mu} - \frac{1}{\mu-\lambda}$$

Oltre al tempo medio di attesa, possiamo anche calcolare quanto vale il **numero medio di utenti in attesa di servizio**. Se indichiamo con q la variabile aleatoria che fornisce il numero di utenti in attesa, ovviamente il suo valore medio E[q] sarà il numero medio di utenti in attesa di servizio.

Per calcolare questo valore medio, possiamo applicare nuovamente la *legge di Little* , ma non all'intero sistema a coda, bensì al *sottosistema* costituito solo dalla **coda di attesa**: non dimentichiamo, infatti, che la legge di Little è stata dimostrata per un sistema qualsiasi a patto che fosse stazionario e la coda di attesa rientra in tale ipotesi. Possiamo perciò scrivere che $E[q] = \lambda E[W]$, da cui si ricava evidentemente che

$$E[q] = \lambda\left(\frac{1}{\mu} - \frac{1}{\mu-\lambda}\right)$$

CASO PARTICOLARE: SISTEMA DI TIPO M/M/1/∞/∞ AD ARRIVI RALLENTATI

Consideriamo nuovamente un sistema a coda di tipo M/M/1/∞/∞, ossia un sistema in cui il traffico di ingresso è un processo di Poisson, il tempo di servizio è una variabile aleatoria di tipo esponenziale, c'è 1 solo servente, la capacità di memorizzazione è infinita (sistema senza perdite) e il numero di utenti potenziali del sistema è anch'esso ∞. Rispetto al caso considerato in precedenza, facciamo questa volta l'ipotesi che l'intensità del processo in ingresso non sia costante e pari a λ, ma sia funzione dello stato del sistema: in particolare, supponiamo che essa sia data dalla relazione

$$\lambda_k = \frac{\alpha}{k+1}$$

dove α è una costante reale mentre k individua lo stato del sistema (ossia il numero di utenti presenti nel sistema).

Un sistema siffatto viene detto **ad arrivi rallentati** in quanto è evidente che l'intensità del processo in ingresso diminuisce all'aumentare di k, ossia all'aumentare del numero di utenti già presenti nel sistema stesso[5].

Così come abbiamo fatto in precedenza, ci interessa valutare le frequenze di transizione di stato e le probabilità asintotiche.

Frequenze di transizione

Sappiamo già che, per un sistema di questo tipo, sono nulle le frequenze di transizione tra stati non adiacenti, mentre le frequenze di transizione di stato sono pari all'intensità del traffico in ingresso, quando si va in avanti, e a quello del traffico in uscita quando si va indietro.

La differenza con un sistema "normale" di tipo M/M/1/∞/∞ è che, mentre in quel caso l'intensità del traffico in ingresso è costante e pari a λ, in questo caso essa dipende dallo stato del sistema: abbiamo perciò che

$$\gamma_{01} = \lambda_0 = \frac{\alpha}{1} = \alpha$$

$$\gamma_{12} = \lambda_1 = \frac{\alpha}{1+1} = \frac{\alpha}{2}$$

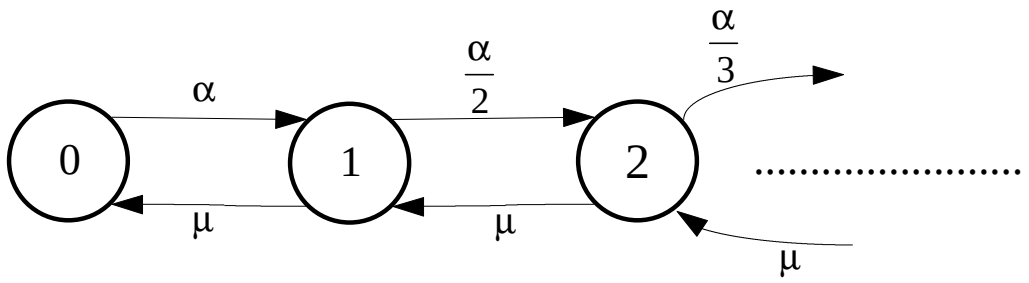
$$\gamma_{23} = \lambda_2 = \frac{\alpha}{2+1} = \frac{\alpha}{3}$$

....

$$\gamma_{kk+1} = \lambda_k = \frac{\alpha}{k+1}$$

[5] Un tipico sistema reale che possa funzionare in questo modo è il cosiddetto **nodo intermedio** in una rete di calcolatori: un nodo intermedio di una rete è sostanzialmente un computer che riceve dei dati da una serie di linee di ingresso e li instrada su una serie di linee di uscita; quando le linee di uscita sono in numero inferiore a quelle di ingresso, il nodo può risultare sovraccaricato, in quanto è possibile che i dati in ingresso arrivino in quantità superiore a quella che il nodo riesce a smaltire in uscita; allora, per ottimizzare il funzionamento ed evitare perdita di dati, il nodo segnala agli altri dispositivi cui è collegato di *rallentare* il flusso di informazioni verso di lui; si può allora pensare di *rallentare* il flusso proporzionalmente all quantità di dati che sono già presenti nel nodo e sono in attesa di essere instradati sulle linee di uscita.

Non cambia invece niente per le frequenze di transizione all'indietro, che rimangono tutte uguali a μ: infatti, quale che sia il numero di utenti presenti nel sistema (tranne ovviamente 0), noi abbiamo l' unico servente occupato, per cui il numero medio di utenti serviti dal sistema nell'unità di tempo è pari al numero medio di utenti serviti nell'unità di tempo dal servente stesso, ossia appunto μ.



Come si nota anche da questa rappresentazione grafica del sistema, ***il traffico in ingresso diminuisce all'aumentare del numero di utenti presenti.***

Note le frequenze di transizione di stato, ci andiamo a calcolare le probabilità asintotiche facendo ancora una volta uso della relazione

$$p_j \sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} p_i \gamma_{ij} \qquad \forall j$$

che rappresenta un sistema in un numero di equazione pari al numero di stati possibili del sistema (che in questo caso sono ∞).

Cominciamo dal caso in cui j=0: l'equazione è

$$p_0 \sum_{i \neq 0} \gamma_{0i} = \sum_{i \neq 0} p_i \gamma_{i0}$$

L'unico valore di i per cui abbiamo una frequenza di transizione non nulla è i=1, per cui l'equazione è $p_0 \gamma_{01} = p_1 \gamma_{10}$. Sappiamo poi che

$$\gamma_{i,i+1} = \frac{\alpha}{i+1}$$
$$\gamma_{i,i-1} = \mu$$

per cui $p_0 \alpha = p_1 \mu$.

Passiamo a j=1: l'equazione è

$$p_1 \sum_{i \neq 1} \gamma_{1i} = \sum_{i \neq 1} p_i \gamma_{i1}$$

Gli unici valori consentiti per l'indice i sono adesso 0 e 2: abbiamo dunque che

$$p_1 (\gamma_{10} + \gamma_{12}) = (p_0 \gamma_{01} + p_2 \gamma_{21})$$

e quindi anche che

$$p_1 \left(\mu + \frac{\alpha}{2} \right) = (p_0 \alpha + p_2 \mu)$$

Vediamo ora per j=2: l'equazione è

$$p_2 \sum_{i \neq 2} \gamma_{2i} = \sum_{i \neq 2} p_i \gamma_{i2}$$

e essa diventa evidentemente

$$p_2 (\gamma_{21} + \gamma_{23}) = (p_1 \gamma_{12} + p_3 \gamma_{32})$$

ossia anche

$$p_2 \left(\mu + \frac{\alpha}{3} \right) = \left(p_1 \frac{\alpha}{2} + p_3 \mu \right)$$

Potremmo anche proseguire (in teoria fino a j=∞), ma le tre equazioni ricavate

$$p_0 \alpha = p_1 \mu$$
$$p_1 \left(\mu + \frac{\alpha}{2} \right) = (p_0 \alpha + p_2 \mu)$$
$$p_2 \left(\mu + \frac{\alpha}{3} \right) = \left(p_1 \frac{\alpha}{2} + p_3 \mu \right)$$

sono sufficienti per far vedere quello che ci interessa. Infatti, dalla prima equazione si ricava evidentemente che $p_1 = \frac{\alpha}{\mu} p_0$: sostituendo questa nella seconda equazione e facendo qualche manipolazione algebrica, si ricava poi che

$$p_2 = \frac{\alpha^2}{2\mu^2} p_0$$

Sostituendo questa nella terza equazione e facendo altre manipolazioni, si ricava infine che

$$p_3 = \frac{\alpha^3}{6\mu^3} p_0$$

E' evidente, dunque, il seguente risultato fondamentale:

$$\boxed{p_k = \frac{1}{k!} \left(\frac{\alpha}{\mu} \right)^k p_0}$$

Passiamo al calcolo di p_0. Utilizziamo la *condizione di normalizzazione*, ossia imponiamo che risulti $\sum_{i=0}^{\infty} p_i = 1$: sostituendo la relazione trovata prima, risulta

$$\sum_{i=0}^{\infty} \frac{1}{k!} \left(\frac{\alpha}{\mu} \right)^i p_0 = 1$$

e da qui si ricava evidentemente che

$$p_0 = \frac{1}{\sum_{i=0}^{\infty} \frac{1}{i!}\left(\frac{\alpha}{\mu}\right)^i}$$

A prescindere dal valore del rapporto **α/μ**, la sommatoria che compare può essere risolta e si ricava in particolare che

$$p_0 = \frac{1}{e^{\frac{\alpha}{\mu}}} = e^{-\frac{\alpha}{\mu}}$$

Possiamo dunque concludere che le probabilità asintotiche per questo particolare tipo di sistema a coda valgono

$$\boxed{p_k = \frac{1}{k!}\left(\frac{\alpha}{\mu}\right)^k e^{-\frac{\alpha}{\mu}}}$$

Una cosa interessante che si nota è che quella formula non è altro che la formula di Poisson di parametro α/μ, dal che si deduce che *il numero medio di utenti presenti nel sistema nell'unità di tempo è pari a* a/m.

Da notare che, in questo caso particolare, non abbiamo vincoli sul valore di $\rho=\alpha/\mu$, al contrario di quanto visto nel caso precedente. Essendo μ un parametro caratteristico del sistema, non avere vincoli su ρ significa sostanzialmente non avere vincoli su α.

Note le probabilità di stato (asintotiche), possiamo anche calcolare il numero medio di utenti che chiedono servizio al sistema: come è ovvio che sia, questo numero (che indichiamo con λ) dipende dallo stato del sistema, in quanto è dato da

$$\lambda = \alpha p_0 + \frac{\alpha}{2}p_1 + \frac{\alpha}{3}p_2 + ... + \frac{\alpha}{N+1}p_N + ... = \sum_{k=0}^{\infty}\frac{\alpha}{k+1}p_k = \alpha\sum_{k=0}^{\infty}\frac{1}{k+1}p_k$$

Sostituendo l'espressione trovata prima per le probabilità di stato e facendo qualche manipolazione algebrica, otteniamo

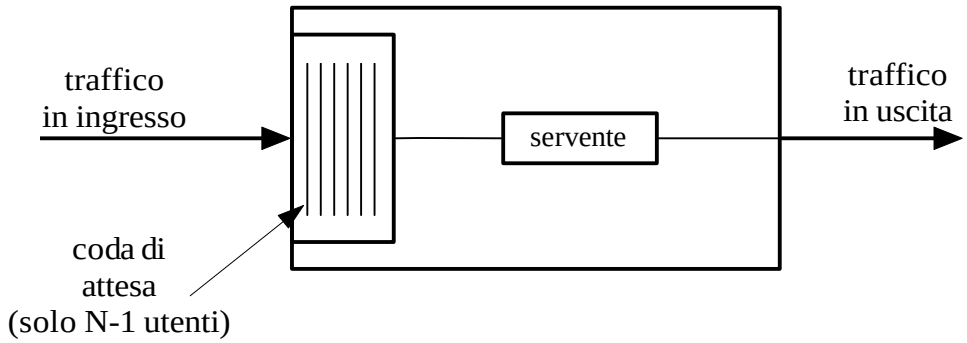
$$\lambda = \alpha\sum_{k=0}^{\infty}\frac{1}{k+1}\cdot\frac{1}{k!}\left(\frac{\alpha}{\mu}\right)^k e^{-\frac{\alpha}{\mu}} = \alpha e^{-\frac{\alpha}{\mu}}\sum_{k=0}^{\infty}\frac{1}{(k+1)!}\left(\frac{\alpha}{\mu}\right)^k = ... = \mu\left(1-e^{-\frac{\alpha}{\mu}}\right)$$

Sistemi a coda di tipo M/M/1/N/∞

INTRODUZIONE

Il secondo tipo di sistema a coda che consideriamo è il tipo **M/M/1/N/∞**, il che significa che si tratta di un sistema con le seguenti caratteristiche:

- il traffico in ingresso al sistema è un processo di Poisson che supponiamo abbia intensità λ (che rappresenta perciò il numero medio di utenti che entrano nel sistema nell'unità di tempo);
- il tempo di servizio è una variabile aleatoria con distribuzione esponenziale che supponiamo abbia parametro μ (che rappresenta quindi il numero medio di utenti serviti nell'unità di tempo);
- c'è 1 solo servente nel sistema;
- il sistema è con perdite, in quanto ha una capacità di memorizzazione (intesa come numero massimo di utenti che possono essere contemporaneamente presenti nel sistema, sia sotto servizio sia in attesa di servizio) finita e pari ad N;
- ci sono ∞ potenziali utenti del sistema.



La differenza con il sistema coda di tipo M/M/1/¥/¥ precedentemente esaminato è dunque nella capacità di memorizzazione, che da infinita è diventata adesso finita; questo fatto comporta una prima conseguenza fondamentale: mentre nel sistema di tipo M/M/1/∞/∞ ci potevano essere anche ∞ utenti presenti contemporaneamente, per cui avevamo ∞ possibili "stati" del sistema, dove ogni stato corrisponde appunto al numero di utenti presenti contemporaneamente, adesso gli stati sono diventati finiti e pari ad N+1:

stato 0 ⟷ 0 utenti presenti

stato 1 ⟷ 1 utente presente

stato 2 ⟷ 2 utenti presenti

….

stato N ⟷ N utenti presenti

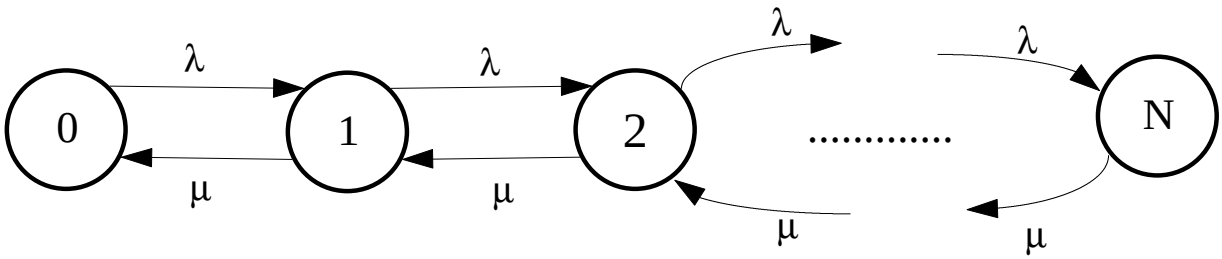
La seconda conseguenza, come detto anche prima, è il fatto che il sistema è con perdite: infatti, potendo conservare al suo interno al più N utenti (di cui ovviamente 1 sotto servizio, in quanto c'è 1 solo servente, e i rimanenti N-1 in attesa di servizio), ogni eventuale richiesta di servizio che pervenga quando ci sono già N utenti nel sistema, viene rigettata.

Ad ogni modo, trattandosi ancora una volta di un sistema in cui l'ingresso è un processo di Poisson e il tempo di servizio è una variabile con distribuzione esponenziale, possiamo studiarlo utilizzando una catena di Markov.

FREQUENZE DI TRANSIZIONE

Lo studio che dobbiamo fare è lo stesso condotto per il sistema di tipo M/M/1/∞/∞: dobbiamo cioè valutare le frequenze di transizione di stato e quindi le probabilità asintotiche.

Per quanto riguarda le frequenze di transizione, le conclusioni sono identiche a quelle trovate per il sistema M/M/1/∞/∞: essendo il traffico in ingresso di tipo poissoniano, le frequenze di transizione in avanti sono tutte pari a λ tra stati adiacenti e sono pari a 0 tra stati adiacenti, mentre quelle all'indietro sono tutte pari a μ tra stati adiacenti e tutte pari a 0 tra stati non adiacenti.



Il motivo per cui le frequenze $\gamma_{i,i-1}$ sono costanti e uguali a μ è lo stesso visto nel paragrafo precedente: quale che sia il numero di utenti presenti nel sistema (compreso ovviamente tra 1 e N), abbiamo l'unico servente occupato, per cui il numero medio di utenti serviti dal sistema nell'unità di tempo è pari al numero medio di utenti serviti nell'unità di tempo dal servente stesso, ossia appunto μ.

PROBABILITÀ ASINTOTICHE

Note le frequenze di transizione di stato, possiamo ancora una volta utilizzare l'**equazione di bilancio del flusso** relativa al generico stato j per determinare le probabilità asintotiche:

$$p_j \sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} p_i \gamma_{ij} \qquad \forall j$$

Come si vedrà dai calcoli, i risultati sono identici a quelli trovati per il sistema M/M/1/∞/∞, salvo il risultato conclusivo, che invece è diverso per via del fatto che, in questo caso, gli stati del sistema sono in numero finito.

Cominciamo da j=0: l'equazione è

$$p_0 \sum_{i \neq 0} \gamma_{0i} = \sum_{i \neq 0} p_i \gamma_{i0}$$

Al primo membro, l'unico valore di i per cui abbiamo una frequenza di transizione non nulla è i=1 e lo stesso vale per il secondo membro, per cui l'equazione è $p_0 \gamma_{01} = p_1 \gamma_{10}$.

Sappiamo poi che

$$\gamma_{i,i+1} = \lambda$$
$$\gamma_{i,i-1} = \mu$$

per cui $p_0\lambda = p_1\mu$.

Passiamo a j=1: l'equazione è

$$p_1 \sum_{i\neq 1} \gamma_{1i} = \sum_{i\neq 1} p_i \gamma_{i1}$$

Al primo membro, gli unici valori consentiti sono 0 e 2 e lo stesso vale per il secondo membro: abbiamo dunque che

$$p_1(\gamma_{10} + \gamma_{12}) = (p_0\gamma_{01} + p_2\gamma_{21})$$

e quindi anche che

$$p_1(\mu + \lambda) = (p_0\lambda + p_2\mu)$$

Vediamo ora per j=2: l'equazione è

$$p_2 \sum_{i\neq 2} \gamma_{2i} = \sum_{i\neq 2} p_i \gamma_{i2}$$

e essa diventa evidentemente

$$p_2(\gamma_{21} + \gamma_{23}) = (p_1\gamma_{12} + p_3\gamma_{32})$$

ossia anche

$$p_2(\mu + \lambda) = (p_1\lambda + p_3\mu)$$

Le tre equazioni ricavate (ci bastano queste) sono dunque

$$p_0\lambda = p_1\mu$$
$$p_1(\mu + \lambda) = (p_0\lambda + p_2\mu)$$
$$p_2(\mu + \lambda) = (p_1\lambda + p_3\mu)$$

Dalla prima equazione si ricava che $p_1 = \frac{\lambda}{\mu} p_0$; sostituendo nella seconda equazione e facendo qualche manipolazione algebrica, si ricava poi che

$$p_2 = \frac{\lambda}{\mu} p_1 = \left(\frac{\lambda}{\mu}\right)^2 p_0$$

Sostituendo questa nella terza equazione e facendo altre manipolazioni, si ricava infine che

$$p_3 = \frac{\lambda}{\mu} p_2 = \left(\frac{\lambda}{\mu}\right)^3 p_0$$

E' evidente, dunque, il seguente risultato fondamentale:

$$\boxed{p_k = \left(\frac{\lambda}{\mu}\right)^k p_0}$$

Come avevamo anticipato, questo è lo stesso risultato trovato nel caso del sistema M/M/1/∞/∞.

Per calcolare p_0 usiamo ancora una volta la *condizione di normalizzazione*: imponendo che $\sum_{i=0}^{N} p_i = 1$, otteniamo evidentemente che

$$p_0 = \frac{1}{\sum_{i=0}^{N}\left(\frac{\lambda}{\mu}\right)^i}$$

Dato che la sommatoria è estesa ad un numero finito di elementi, essa converge sempre (non abbiamo quindi vincoli sui $\rho=\lambda/\mu$), per cui non dobbiamo imporre vincoli particolari: per quanto riguarda la sua espressione in forma chiusa, basta osservare che si tratta della somma della serie geometrica, estesa però, anziché ad un numero infinito di termini, ad un numero finito N+1 di termini. Possiamo perciò scrivere che

$$p_0 = \frac{1}{\frac{1-\rho^{N+1}}{1-\rho}} = \frac{1-\rho}{1-\rho^{N+1}}$$

Possiamo perciò concludere che le probabilità asintotiche, a prescindere dal valore del fattore di utilizzabilità ρ, valgono

$$\boxed{p_k = \rho^k \frac{1-\rho}{1-\rho^{N+1}}}$$

VERIFICA DELLA STABILITÀ DEL SISTEMA

Verifichiamo anche in questo caso che il sistema a coda sia stabile: dobbiamo cioè verificare che il traffico in uscita sia uguale a quello in ingresso.

Indicata ancora una volta con γ l'intensità del traffico in uscita, dobbiamo vedere quanto essa vale: γ è pari al numero medio di utenti serviti dal sistema nell'unità di tempo, a patto però che ci siano questi utenti; allora, dato che il numero medio di utenti serviti nell'unità di tempo è μ e dato che la probabilità che nel sistema ci sia almeno 1 utente è $1-p_0$, possiamo scrivere che $\gamma = \mu(1-p_0)$ e quindi che

$$\gamma = \mu\left(1-\frac{1-\rho}{1-\rho^{N+1}}\right)$$

Fin qui niente di diverso dal sistema di tipo M/M/1/∞/∞. La differenza subentra invece adesso nella determinazione del traffico in ingresso: infatti, c'è da tenere conto del fatto che il sistema è con perdite, ossia della possibilità che una o più richieste di servizio vengano respinte.

Chiamiamo con p_B la cosiddetta **probabilità di blocco**, ossia la probabilità che una generica richiesta di servizio venga respinta (a causa del fatto che il servente è già occupato e ci sono già altri N-1 utenti in attesa di servizio). Quanto vale p_B? E' evidente che il blocco si verifica quando ci sono già N utenti nel sistema, per cui $p_B = p_N$ e quindi, in base a quanto trovato prima, abbiamo che

$$p_B = \rho^N \frac{1-\rho}{1-\rho^{N+1}}$$

Il numero medio di utenti accettati dal sistema nell'unità di tempo, ossia l'intensità del traffico in ingresso, è dunque pari al numero medio di richieste che arrivano nell'unità di tempo, cioè λ, per la probabilità che non ci sia nessun blocco, ossia $1-p_B$: quindi il traffico accettato in ingresso è

$$\lambda(1-p_B) = \lambda\left(1-\rho^N \frac{1-\rho}{1-\rho^{N+1}}\right)$$

Avendo detto che il traffico in uscita è invece

$$\gamma = \mu\left(1-\frac{1-\rho}{1-\rho^{N+1}}\right)$$

controlliamo che sia verificata la relazione

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

Portando μ al primo membro, quella diventa

$$\rho - \rho^{N+1}\frac{1-\rho}{1-\rho^{N+1}} = 1-\frac{1-\rho}{1-\rho^{N+1}}$$

Portando tutto sotto un unico denominatore, abbiamo inoltre che

$$\rho\left(1-\rho^{N+1}\right) - \rho^{N+1}(1-\rho) = 1-\rho^{N+1}-1+\rho$$

e questa è effettivamente una identità.

Notiamo una cosa: in base ai conti appena fatti, il sistema è sempre stabile per qualunque valore di ρ; è un risultato intuitivo, in quanto il sistema ha capacità di memorizzazione finita, per cui è escluso che il numero di utenti al suo interno aumenti indefinitamente. Questo, però, non significa che possiamo tollerare un qualsiasi valore di $\rho=\lambda/\mu$: infatti, se risultasse $\lambda>>\mu$, significherebbe che molte richieste di servizio vengono respinte, il che non è certo indice di efficienza da parte del sistema. E' quindi opportuno dimensionare il sistema in modo che ci sia il miglior bilanciamento possibile tra λ e μ.

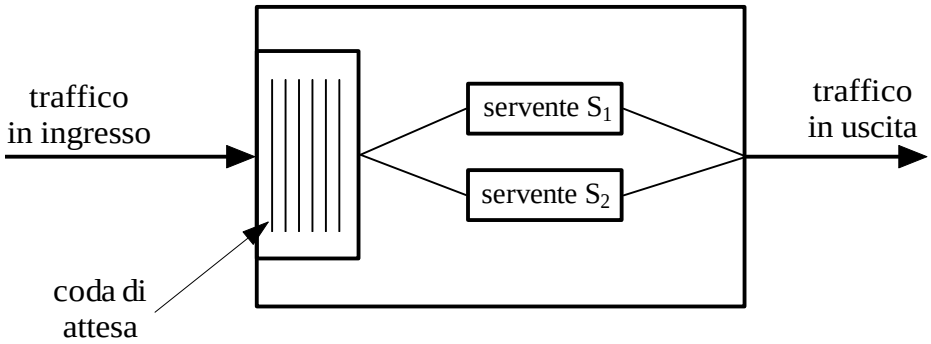
Sistemi a coda di tipo M/M/2/∞/∞

INTRODUZIONE

I due tipi di sistema a coda fino ad ora esaminati comprendevano entrambi 1 solo servente, mentre differivano per il valore della capacità di memorizzazione. Consideriamo adesso un nuovo tipo di sistema a coda, avente le seguenti caratteristiche:

- il traffico in ingresso è un processo di Poisson (con intensità λ, che indica il numero medio di richieste di servizio che giungono al sistema nell'unità di tempo);
- il tempo di servizio è una variabile aleatoria con distribuzione esponenziale (con un parametro μ, che indica il numero medio di utenti serviti dal sistema nell'unità di tempo),
- il sistema dispone di 2 serventi;
- il sistema è in grado di mantenere contemporaneamente dentro di sé ∞ utenti (ossia ha una capacità infinita di memorizzazione;
- gli utenti potenziali sono ∞.

Abbiamo dunque un sistema a coda di tipo **M/M/2/∞/∞**:



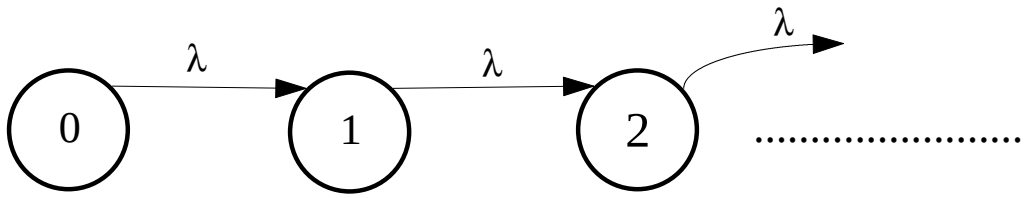
Così come abbiamo fatto per gli altri sistemi a coda esaminati, *lo possiamo studiare come se fosse una catena di Markov ad infiniti stati, dove ogni stato rappresenta il numero di utenti presenti contemporaneamente nel sistema.*

FREQUENZE DI TRANSIZIONE

La prima cosa che ci interessa sono le frequenze di transizione di stato: la generica di queste, indicata con γ_{ij} , rappresenta il numero di volte, nell'unità di tempo, in cui il sistema passa dallo stato i (corrispondente quindi ad i utenti presenti) allo stato j (corrispondente a j utenti presenti).

Per quanto riguarda le frequenze di transizione tra uno stato e quello immediatamente successivo, valgono le stesse considerazioni dei casi precedenti: essendo il traffico in ingresso un processo di Poisson (di intensità λ), indipendente dalle caratteristiche del sistema, si ha che

$$\gamma_{i,i+1} = \lambda$$



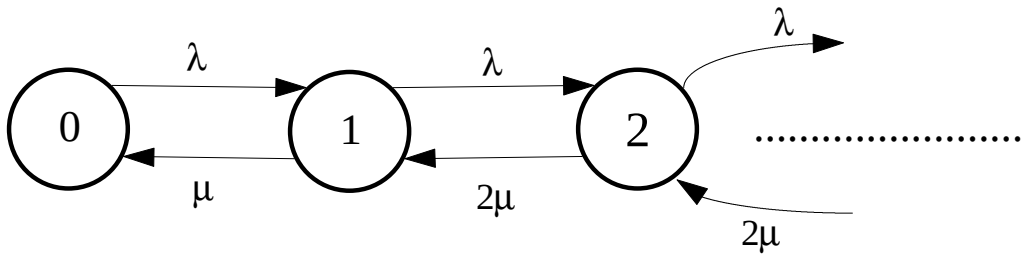
Valgono anche le stesse considerazioni per quanto riguarda le frequenze di transizione, sempre in avanti, ma tra stati non adiacenti:

$$\gamma_{i,i+k} = 0 \qquad \forall k \geq 2$$

La conclusione è dunque ancora una volta che *in un sistema a coda di tipo M/M/2/¥/¥, il sistema può passare da uno stato ad uno successivo solo a patto che sia quello adiacente; la frequenza di transizione vale sempre l (intensità del traffico in ingresso).*

Adesso dobbiamo fare gli stessi ragionamenti per i passaggi di stato "all'indietro", ossia per i passaggi dallo stato generico i allo stato precedente i-1, a due precedenti i-2 e così via.

In particolare, cominciamo a valutare $\gamma_{i,i-1}$, ossia la frequenza di transizione dallo stato i allo stato precedente $i-1$. Il sistema passa dallo stato i allo stato i-1 quando esso termina di servire un utente, per cui quest'ultimo esce dal sistema e quindi diminuisce di 1 il numero di utenti presenti complessivamente nel sistema stesso. Dato che il sistema dispone di 2 diversi serventi, abbiamo due diverse possibilità per questo passaggio di stato: quando c'è un solo utente nel sistema, valgono le stesse considerazioni fatte negli altri casi, in quanto il numero medio di utenti serviti dal sistema nell'unità di tempo, che è pari alla frequenza richiesta, è pari al numero medio di utenti serviti dall'unico servente in funzione e quindi vale μ; al contrario, quando ci sono almeno due utenti nel sistema, ossia quando il sistema si trova nello stato 2 o in uno successivo, allora la frequenza vale 2μ: infatti, in questo caso entrambi i serventi sono occupati, per cui il numero medio di utenti serviti dal sistema nell'unità di tempo è pari alla somma del numero medio di utenti serviti da ciascun servente nell'unità di tempo, ossia appunto 2μ.



Anche per i passaggi all'indietro, è facile verificare come non ci possano essere transizioni tra stati non adiacenti: è facile cioè trovare che

$$\gamma_{i,i-k} = 0 \qquad \forall k \geq 2$$

N.B. Le frequenze di transizione all'indietro prendono il nome di "**frequenze di servizio**": sono così chiamate perché il passaggio da uno stato ad uno precedente (adiacente o meno) corrisponde al fatto che uno o più utenti sono stati serviti e sono quindi usciti dal sistema. Le frequenze di transizione in avanti, invece, sono le "**frequenze di transizione**" propriamente dette.
Rispetto ai due sistemi a coda considerati in precedenza, questo sistema presenta sostanzialmente un aumento della frequenza di servizio, che raddoppia grazie alla presenza di 2 serventi anziché uno solo

LE PROBABILITÀ ASINTOTICHE

Note le frequenze di transizione, il passo successivo consiste nel determinare le probabilità asintotiche mediante la solita relazione

$$p_j \sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} p_i \gamma_{ij} \qquad \forall j$$

che rappresenta un sistema in un numero di equazione pari al numero di stati possibili del sistema (che in questo caso sono ∞).

Cominciamo dal caso in cui j=0: l'equazione è

$$p_0 \sum_{i \neq 0} \gamma_{0i} = \sum_{i \neq 0} p_i \gamma_{i0}$$

L'unico valore di i per cui abbiamo una frequenza di transizione non nulla è i=1, per cui l'equazione è

$$p_0 \gamma_{01} = p_1 \gamma_{10}$$

Sappiamo poi che $\begin{cases} \gamma_{01} = \lambda \\ \gamma_{10} = \mu \end{cases}$, per cui

$$p_0 \alpha = p_1 \mu$$

Passiamo a j=1: l'equazione è

$$p_1 \sum_{i \neq 1} \gamma_{1i} = \sum_{i \neq 1} p_i \gamma_{i1}$$

Gli unici valori consentiti per l'indice i sono adesso 0 e 2: abbiamo dunque che

$$p_1 (\gamma_{10} + \gamma_{12}) = (p_0 \gamma_{01} + p_2 \gamma_{21})$$

e quindi anche che

$$p_1 (\mu + \lambda) = (p_0 \lambda + 2 p_2 \mu)$$

Vediamo ora per j=2: l'equazione è

$$p_2 \sum_{i \neq 2} \gamma_{2i} = \sum_{i \neq 2} p_i \gamma_{i2}$$

e essa diventa evidentemente

$$p_2 (\gamma_{21} + \gamma_{23}) = (p_1 \gamma_{12} + p_3 \gamma_{32})$$

ossia anche

$$p_2 (2\mu + \lambda) = (p_1 \lambda + 2 p_3 \mu)$$

Da queste tre equazioni possiamo ricavare quello che ci interessa: le equazioni sono

$$p_0\lambda = p_1\mu$$
$$p_1(\mu+\lambda) = (p_0\lambda + 2p_2\mu)$$
$$p_2(2\mu+\lambda) = (p_1\lambda + 2p_3\mu)$$

Dalla prima si ricava evidentemente che $p_1 = \frac{\lambda}{\mu}p_0$. Sostituendo questa nella seconda equazione e facendo qualche manipolazione algebrica, si ricava poi che

$$p_2 = \frac{\lambda^2}{2\mu^2}p_0$$

Sostituendo questa nella terza equazione e facendo altre manipolazioni, si ricava infine che

$$p_3 = \frac{\lambda}{\mu}\left(\frac{\lambda}{2\mu}\right)^2 p_0$$

E' evidente, dunque, che il risultato generale è

$$p_k = \frac{\lambda}{\mu}\left(\frac{\lambda}{2\mu}\right)^{k-1} p_0$$

e può anche essere scritto nella forma

$$\boxed{p_k = 2\left(\frac{\lambda}{2\mu}\right)^k p_0}$$

Passiamo al calcolo di p_0. Utilizziamo la *condizione di normalizzazione*: imponiamo cioè che risulti

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

Ci conviene estrarre dalla sommatoria il termine per i=0:

$$p_0 + \sum_{i=1}^{\infty} p_i = 1$$

Sostituendo l'espressione trovata prima per la generica p_k, abbiamo allora che

$$p_0 + \sum_{i=1}^{\infty} 2\left(\frac{\lambda}{2\mu}\right)^i p_0 = 1$$

Da qui (oltre a comprendere perché abbiamo separato il termine p_0), si ricava evidentemente che

$$p_0 = \frac{1}{1+2\sum_{i=1}^{\infty}\left(\frac{\lambda}{2\mu}\right)^i}$$

Nella solita ipotesi che il rapporto **$\rho=\lambda/2\mu$** sia minore di 1, la sommatoria che compare è quella della serie geometrica convergente, per cui possiamo scrivere che

$$p_0 = \frac{1-\rho}{1+\rho}$$

Possiamo dunque concludere che le probabilità asintotiche per questo particolare tipo di sistema a coda, nell'ipotesi che $\rho=\lambda/(2\mu)<1$, valgono

$$\boxed{p_k = 2\rho^k \frac{1-\rho}{1+\rho}}$$

Note le probabilità di stato (asintotiche), possiamo calcolare alcuni interessanti parametri statistici sul sistema. Ad esempio, il numero medio E[N] di utenti presenti nel sistema sarà

$$E[N] = \sum_{k=0}^{\infty} k \cdot p_k = \sum_{k=0}^{\infty} k \cdot 2\rho^k \frac{1-\rho}{1+\rho} = 2\frac{1-\rho}{1+\rho}\sum_{k=0}^{\infty} k \cdot \rho^k = 2\frac{1-\rho}{1+\rho}\frac{\rho}{(1-\rho)^2} = 2\frac{\rho}{1-\rho^2}$$

Applicando inoltre la legge di Little, possiamo calcolare il tempo medio di permanenza E[T] del generico utente nel sistema: si ha che

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

DETERMINAZIONE DEL TRAFFICO IN USCITA

L'ultima cosa da fare è la determinazione del traffico in uscita, indicato fino ad ora con γ, e la verifica della condizione di stabilità secondo cui tale traffico risulta uguale a quello in ingresso.

Quanto vale il traffico in uscita? γ rappresenta il numero medio di utenti che escono dal sistema, nell'unità di tempo: esso corrisponde al numero medio di utenti serviti nell'unità di tempo moltiplicato per la probabilità che ci siano tali utenti; a differenza, però, degli altri casi, abbiamo questa volta due possibilità:

- quando c'è un solo utente nel sistema, allora il numero medio di utenti serviti nell'unità di tempo è μ, per cui il traffico in uscita risulta μp_1;
- al contrario, quando c'è più di un utente, il traffico in uscita è il massimo possibile e vale $2\mu(1-p_0-p_1)$, dove ovviamente $(1-p_0-p_1)$ è la probabilità che ci siano almeno due utenti nel sistema.

Possiamo dunque scrivere che

$$\gamma = 2\mu(1-p_0-p_1)+\mu p_1$$

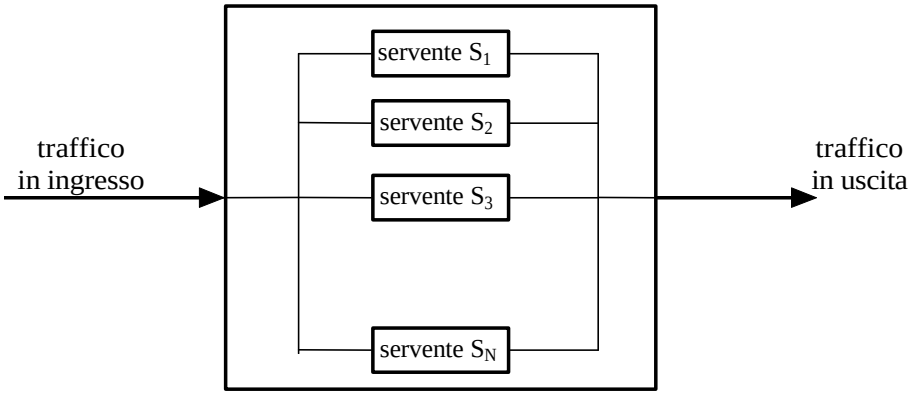
Sistemi a coda di tipo M/M/N/N/∞

INTRODUZIONE

Consideriamo adesso un nuovo tipo di sistema a coda, avente le seguenti caratteristiche:

- il traffico in ingresso è un processo di Poisson (con intensità λ, che indica il numero medio di richieste di servizio che giungono al sistema nell'unità di tempo);
- il tempo di servizio è una variabile aleatoria con distribuzione esponenziale (con un parametro μ, che indica il numero medio di utenti serviti dal sistema nell'unità di tempo),
- il sistema dispone di N serventi;
- il sistema è in grado di mantenere contemporaneamente dentro di sé N utenti (ossia ha una capacità di memorizzazione pari ad N);
- gli utenti potenziali sono ∞.

Abbiamo dunque un sistema a coda di tipo **M/M/N/N/∞**:



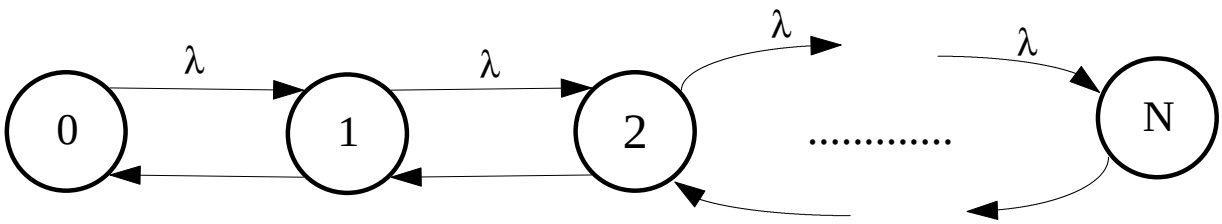
La prima cosa interessante da notare è che il numero di serventi è pari al numero massimo di utenti che possono trovarsi all'interno del sistema: ciò significa evidentemente che, all'interno del sistema, non ci sono utenti in attesa o anche, in altre parole, che, nel momento in cui un generico utente è ammesso nel sistema, questo accade perché c'è almeno un servente disponibile.

Si tratta dunque di un **sistema bloccante con perdite**, in quanto viene respinta ogni eventuale richiesta di servizio che giunge quando tutti i serventi sono occupati, e quindi anche di un sistema con numero di stati finito: in particolare, essendo N il numero massimo di utenti ammessi nel sistema, abbiamo N+1 stati, partendo dallo stato 0 (corrispondente a 0 utenti nel sistema) per finire allo stato N (corrispondente ad N utenti nel sistema, ossia quindi a tutti i serventi occupati).

FREQUENZE DI TRANSIZIONE

Nell'ipotesi che il traffico in ingresso sia un processo di Poisson, che sia indipendente dallo stato del sistema ed abbia intensità costante pari a λ, è immediato dedurre ancora una volta che

$$\gamma_{i,i+1} = \lambda$$
$$\gamma_{i,i+k} = 0 \qquad \forall k \geq 2$$



La conclusione è dunque ancora una volta che *in un sistema a coda di tipo M/M/N/N/¥, il sistema può passare da uno stato ad uno successivo solo a patto che sia quello adiacente; la frequenza di transizione vale sempre l (intensità del traffico in ingresso).*

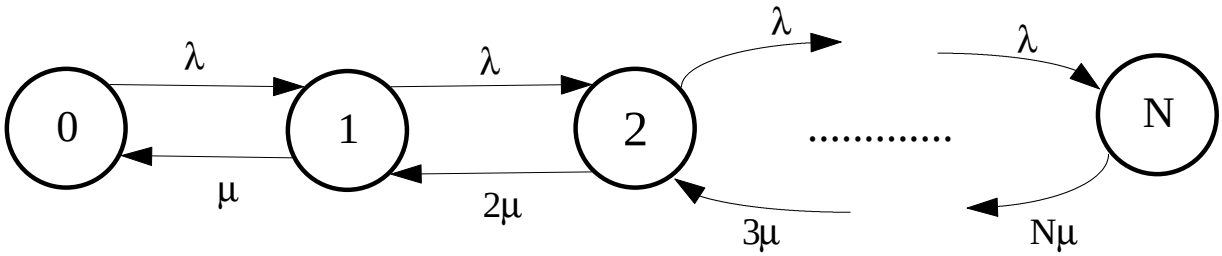
Per quanto riguarda le transizioni da uno stato ad uno precedente, è facile verificare che la frequenza di transizione aumenta all'aumentare del numero di utenti presenti; infatti, tenendo conto che la frequenza di transizione da uno stato al precedente è pari al numero medio di utenti serviti dal sistema nell'unità di tempo e tenendo conto che questo numero medio dipende da quanti serventi sono in funzione, ossia, in questo caso, da quanti utenti sono nel sistema, le possibilità sono le seguenti:

- se c'è 1 solo utente (stato 1), il numero medio di utenti serviti dal sistema nell'unità di tempo è pari al numero medio di utenti serviti dall'unico servente in funzione, ossia μ;
- se ci sono 2 utenti (stato 2), ossia se due serventi sono attivi, il numero medio di utenti serviti dal sistema nell'unità di tempo è pari alla somma del numero medio di utenti serviti dal primo servente più il numero medio di utenti serviti (sempre nell'unità di tempo) dall'altro servente, ossia 2μ;
- ...
- se ci sono N utenti (stato N), ossia se tutti gli N serventi sono attivi, il numero medio di utenti serviti dal sistema nell'unità di tempo è pari alla somma del numero medio di utenti serviti da ciascun servente nell'unità di tempo, ossia Nμ.

Possiamo dunque scrivere che

$$\gamma_{i,i-1} = i\mu$$

E' facile inoltre verificare che sono nulle anche in questo caso le frequenze di transizione tra stati non adiacenti.



LE PROBABILITÀ ASINTOTICHE

Note le frequenze di transizione, passiamo come al solito a determinare le probabilità asintotiche mediante la relazione

$$p_j \sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} p_i \gamma_{ij} \qquad \forall j$$

Per lo stato j=0, abbiamo dunque che

$$p_0 \sum_{i \neq 0} \gamma_{0i} = \sum_{i \neq 0} p_i \gamma_{i0}$$

L'unico valore di i per cui abbiamo una frequenza di transizione non nulla è i=1, per cui l'equazione è $p_0 \gamma_{01} = p_1 \gamma_{10}$. Sappiamo poi che $\begin{matrix} \gamma_{01} = \lambda \\ \gamma_{10} = \mu \end{matrix}$, per cui

$$p_0 \alpha = p_1 \mu$$

Passiamo a j=1: l'equazione è

$$p_1 \sum_{i \neq 1} \gamma_{1i} = \sum_{i \neq 1} p_i \gamma_{i1}$$

Gli unici valori consentiti per l'indice i sono adesso 0 e 2: abbiamo dunque che

$$p_1 (\gamma_{10} + \gamma_{12}) = (p_0 \gamma_{01} + p_2 \gamma_{21})$$

e quindi anche che

$$p_1 (\mu + \lambda) = (p_0 \lambda + 2 p_2 \mu)$$

Vediamo ora per j=2: l'equazione è

$$p_2 \sum_{i \neq 2} \gamma_{2i} = \sum_{i \neq 2} p_i \gamma_{i2}$$

e essa diventa evidentemente

$$p_2 (\gamma_{21} + \gamma_{23}) = (p_1 \gamma_{12} + p_3 \gamma_{32})$$

ossia anche

$$p_2 (2\mu + \lambda) = (p_1 \lambda + 3 p_3 \mu)$$

Da queste tre equazioni possiamo ricavare quello che ci interessa: le equazioni sono

$$\begin{aligned} p_0 \lambda &= p_1 \mu \\ p_1 (\mu + \lambda) &= (p_0 \lambda + 2 p_2 \mu) \\ p_2 (2\mu + \lambda) &= (p_1 \lambda + 3 p_3 \mu) \end{aligned}$$

Dalla prima si ricava evidentemente che

$$p_1 = \frac{\lambda}{\mu} p_0$$

Sostituendo questa nella seconda equazione e facendo qualche manipolazione algebrica, si ricava poi che $p_2 = \frac{\lambda^2}{2\mu^2} p_0$.

Sostituendo questa nella terza equazione e facendo altre manipolazioni, si ricava infine che

$$p_3 = \frac{1}{6}\left(\frac{\lambda}{\mu}\right)^3 p_0$$

E' evidente, dunque, che il risultato generale è

$$\boxed{p_k = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k p_0}$$

Passiamo al calcolo di p_0. Utilizziamo la condizione di normalizzazione: imponiamo cioè che risulti $\sum_{i=0}^{N} p_i = 1$, per cui

$$\sum_{i=0}^{N} \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i p_0 = 1$$

e da qui si ricava evidentemente che

$$\boxed{p_0 = \frac{1}{\sum_{i=0}^{N} \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i}}$$

Per la sommatoria a denominatore non esiste una espressione in forma chiusa, per cui la lasciamo così com'è.

PROBABILITÀ DI BLOCCO

Trattandosi di un sistema con perdite, è importante, anche ai fini della valutazione del traffico in uscita, determinare quella che abbiamo definito "*probabilità di blocco*", ossia la probabilità p_B che una richiesta di servizio venga respinta.

Come accennato all'inizio, questa probabilità è pari alla probabilità che ci siano N utenti già presenti nel sistema, in quanto questa situazione corrisponde a dire che tutti i serventi sono occupati. In base alle formule ricavate prima, possiamo dunque scrivere che

$$p_B = p_N = \frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N p_0$$

e anche, sostituendo l'espressione di p_0, che

$$p_B = \frac{\frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N}{\sum_{i=0}^{N}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i}$$

Questa formula prende il nome di "**formula di Erlang di tipo B (o con perdite) con parametri λ e μ**".

DETERMINAZIONE DEL TRAFFICO IN USCITA

L'ultima cosa da fare è determinare il traffico in uscita, indicato sempre con γ, e verificare la condizione di stabilità secondo cui tale traffico risulta uguale a quello in ingresso.

Quanto vale il traffico in uscita? Sappiamo ormai bene che γ rappresenta il numero medio di utenti che escono dal sistema nell'unità di tempo: ragionando ancora una volta sui serventi e sulle frequenze di servizio, abbiamo perciò che

$$\gamma = \mu p_1 + 2\mu p_2 + \ldots + N\mu p_N = \mu\sum_{k=1}^{N} k p_k$$

A ben guardare, quella sommatoria non è altro che il valor medio della variabile aleatoria N che indica il numero di utenti presenti nel sistema in un generico istante: quindi

$$\gamma = \mu E[N]$$

Verifichiamo adesso che il sistema sia stabile, ossia che il traffico in uscita appena determinato sia pari al traffico accettato in ingresso. Quanto vale il traffico accettato in ingresso? In modo analogo a quanto visto in precedenza, è pari al prodotto dell'intensità del traffico in ingresso λ per la probabilità $1\text{-}p_B$ che non ci sia blocco. Dobbiamo dunque verificare che

$$\mu\sum_{k=1}^{N} k p_k = \lambda(1 - p_B)$$

Cominciamo perciò a sostituire l'espressione di p_B prima ricavata:

$$\mu\sum_{k=1}^{N} k p_k = \lambda\left(1 - \frac{\frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N}{\sum_{k=0}^{N}\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}\right)$$

Ora sostituiamo l'espressione della generica p_K:

$$\mu\sum_{k=1}^{N} k\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k p_0 = \lambda\left(1-\frac{\frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N}{\sum_{k=0}^{N}\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}\right)$$

Infine sostituiamo l'espressione di p_0:

$$\mu\sum_{k=1}^{N}\frac{k\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}{\sum_{i=0}^{N}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i} = \lambda\left(1-\frac{\frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N}{\sum_{k=0}^{N}\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}\right)$$

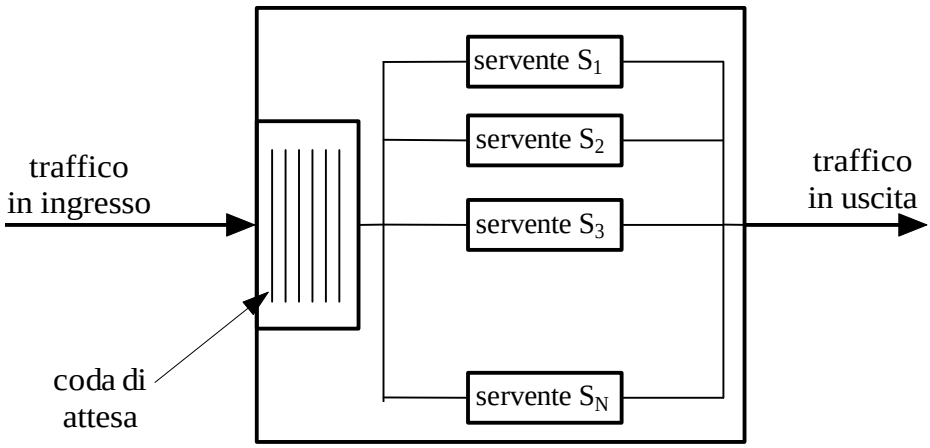
Adesso dobbiamo manipolare algebricamente questa relazione: ....facendolo, si trova che essa risulta essere una identità, da cui deduciamo che il sistema è stabile (come è ovvio che fosse, dato che la sua capacità di memorizzazione è finita, per cui non è possibile che il numero di utenti presenti nel sistema cresca indefinitamente.

Sistemi a coda di tipo M/M/N/∞/∞

INTRODUZIONE

Il sistema a coda che consideriamo adesso è il tipo **M/M/N/∞/∞**, il che significa che possiede le seguenti caratteristiche:

- il traffico in ingresso al sistema è un processo di Poisson che supponiamo abbia intensità λ (che rappresenta perciò il numero medio di utenti che entrano nel sistema nell'unità di tempo);
- il tempo di servizio è una variabile aleatoria con distribuzione esponenziale che supponiamo abbia parametro μ (che rappresenta quindi il numero medio di utenti serviti nell'unità di tempo);
- ci sono N serventi nel sistema;
- il sistema è senza perdite, in quanto ha una capacità di memorizzazione (intesa come numero massimo di utenti che possono essere contemporaneamente presenti nel sistema, sia sotto servizio sia in attesa di servizio) infinita;
- ci sono ∞ potenziali utenti del sistema.



Il fatto che il sistema abbia una capacità di memorizzazione infinita indica subito (oltre al fatto che tutte le richieste di servizio vengono comunque accettate) che si tratta di un sistema ad infiniti stati, se continuiamo ad individuare il generico stato in termini di numero di utenti presenti nel sistema.

FREQUENZE DI TRANSIZIONE

Per quanto riguarda le transizioni di stato in avanti, il risultato è ancora una volta quello per cui le frequenze di transizioni tra stati adiacenti valgono λ, mentre tutte le altre sono nulle:

$$\gamma_{i,i+1} = \lambda$$
$$\gamma_{i,i+k} = 0 \qquad \forall k \geq 2$$

Per quanto riguarda, invece, le frequenze di servizio, sono ancora nulle quelle tra stati non adiacenti, mentre sono particolari quelle tra stati adiacenti:

- se c'è 1 solo utente (stato 1), il numero medio di utenti serviti dal sistema nell'unità di tempo è pari al numero medio di utenti serviti dall'unico servente in funzione, ossia μ;
- se ci sono 2 utenti (stato 2), ossia se due serventi sono attivi, il numero medio di utenti serviti dal sistema nell'unità di tempo è pari alla somma del numero medio di utenti serviti dal primo servente più il numero medio di utenti serviti (sempre nell'unità di tempo) dall'altro servente, ossia 2μ;
- ....
- se ci sono N utenti (stato N), ossia se tutti gli N serventi sono attivi, il numero medio di utenti serviti dal sistema nell'unità di tempo è pari alla somma del numero medio di utenti serviti da ciascun servente nell'unità di tempo, ossia $N\mu$;
- se ci sono N+1 utenti o N+2 utenti,....e così via (ossia il sistema si trova in un generico stato n+1 con $n \geq N$), N di questi utenti sono sotto servizio (in quanto N sono i serventi), mentre quelli rimanenti sono in attesa di servizio (visto che il sistema ha capacità di memorizzazione infinita); quindi la situazione è la stessa dello stato N, ossia la frequenza di servizio vale $N\mu$.

Possiamo perciò concludere che

$$\boxed{\begin{aligned} \gamma_{i,i-1} &= i\mu \quad \forall i \le N \\ \gamma_{i,i-1} &= N\mu \quad \forall i > N \\ \gamma_{i,i-k} &= 0 \quad \forall k \ge 2 \end{aligned}}$$

PROBABILITÀ ASINTOTICHE

Note le frequenze di transizione, usiamo la solita equazione di bilancio del flusso per determinare le probabilità di stato a regime (o probabilità asintotiche):

$$p_j \sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} p_i \gamma_{ij} \qquad \forall j$$

Ovviamente, dato che le frequenze di servizio $\gamma_{i,i-1}$ differiscono a seconda che consideriamo stati precedenti allo stato N o stati successivi, dobbiamo distinguere i due casi.

Cominciamo dagli stati precedenti allo stato N. Per j=0 abbiamo

$$p_0 \sum_{i \neq 0} \gamma_{0i} = \sum_{i \neq 0} p_i \gamma_{i0}$$

L'unico valore di i per cui abbiamo una frequenza di transizione non nulla è i=1, per cui l'equazione è

$$p_0 \gamma_{01} = p_1 \gamma_{10}$$

Sappiamo poi che $\begin{matrix} \gamma_{01} = \lambda \\ \gamma_{10} = \mu \end{matrix}$, per cui

$$p_0 \alpha = p_1 \mu$$

Passiamo a j=1: l'equazione è

$$p_1 \sum_{i \neq 1} \gamma_{1i} = \sum_{i \neq 1} p_i \gamma_{i1}$$

Gli unici valori consentiti per l'indice i sono adesso 0 e 2: abbiamo dunque che

$$p_1 (\gamma_{10} + \gamma_{12}) = (p_0 \gamma_{01} + p_2 \gamma_{21})$$

e quindi anche che

$$p_1 (\mu + \lambda) = (p_0 \lambda + 2 p_2 \mu)$$

Vediamo ora per j=2: l'equazione è

$$p_2 \sum_{i \neq 2} \gamma_{2i} = \sum_{i \neq 2} p_i \gamma_{i2}$$

e essa diventa evidentemente

$$p_2(\gamma_{21}+\gamma_{23})=(p_1\gamma_{12}+p_3\gamma_{32})$$

ossia anche

$$p_2(2\mu+\lambda)=(p_1\lambda+3p_3\mu)$$

Dalle tre equazioni ottenute possiamo ricavare quello che ci interessa:

$$\begin{aligned} p_0\lambda &= p_1\mu \\ p_1(\mu+\lambda) &= (p_0\lambda+2p_2\mu) \\ p_2(2\mu+\lambda) &= (p_1\lambda+3p_3\mu) \end{aligned}$$

Dalla prima si ricava che $p_1=\frac{\lambda}{\mu}p_0$. Sostituendo nella seconda si ricava poi che $p_2=\frac{\lambda^2}{2\mu^2}p_0$. Sostituendo infine nella terza, si ha che

$$p_3=\frac{1}{6}\left(\frac{\lambda}{\mu}\right)^3 p_0$$

E' evidente, dunque, che il risultato generale è

$$p_k=\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k p_0$$

e questo può anche essere scritto nella forma

$$\boxed{p_k=\prod_{i=0}^{k-1}\frac{1}{(i+1)}\left(\frac{\lambda}{\mu}\right)p_0 \qquad k\le N}$$

Vediamo adesso come cambiano le cose per n≥N:....ciò che si trova è

$$\boxed{p_k=p_0\prod_{i=0}^{N-1}\frac{1}{(i+1)}\left(\frac{\lambda}{\mu}\right)\prod_{i=N}^{k-1}\frac{\lambda}{N\mu} \qquad k>N}$$

In entrambe queste formule, rimane ancora da determinare il valore di p_0 (probabilità che il sistema sia vuoto). Lo si ottiene imponendo ancora una volta la normalizzazione. Per imporre tale normalizzazione, però, conviene prima riscrivere la p_k, nei due casi, in modo più opportuno: in particolare, con qualche passaggio in più si trova, ponendo **$\rho=\lambda/N\mu$**, che

$$\begin{aligned} p_k &= \frac{1}{k!}(N\rho)^k p_0 \qquad k\le N \\ p_k &= \frac{1}{N!}(N\rho)^k p_0 \qquad k>N \end{aligned}$$

La condizione di normalizzazione si può scrivere come

$$1=\sum_{k=0}^{\infty}p_k=\sum_{k=0}^{N}p_k+\sum_{k=N+1}^{\infty}p_k$$

Sostituendo le due probabilità p_k trovate prima, si ottiene

$$\sum_{k=0}^{N}\frac{1}{k!}(N\rho)^k p_0+\sum_{k=N+1}^{\infty}\frac{1}{N!}(N\rho)^k p_0=1$$

Nella prima sommatoria, il termine per k=N si può estrarre e inglobare nella seconda, per cui

$$\sum_{k=0}^{N-1}\frac{1}{k!}(N\rho)^k p_0+\sum_{k=N}^{\infty}\frac{1}{N!}(N\rho)^k p_0=1$$

Esplicitando p_0 e facendo qualche altro passaggio, si conclude che

$$\sum_{k=0}^{N-1}\frac{1}{k!}(N\rho)^k+\frac{1}{N!}\sum_{k=N}^{\infty}(N\rho)^k=\frac{1}{p_0}\longrightarrow p_0=\frac{1}{\sum_{k=0}^{N-1}\frac{1}{k!}(N\rho)^k+\frac{1}{N!}\sum_{k=N}^{\infty}(N\rho)^k}=\ldots=\frac{1}{\sum_{k=0}^{N-1}\frac{1}{k!}(N\rho)^k+\frac{1}{N!}\frac{1}{1-\rho}(N\rho)^N}$$

PROBABILITÀ DI ATTESA E NUMERO MEDIO DI UTENTI IN CODA

Si definisce "**probabilità di attesa**" la probabilità che una richiesta di servizio venga posta in attesa: evidentemente, questo evento si verifica quando nel sistema tutti i serventi sono occupati; essendoci N serventi ed essendo infinita la capacità di memorizzazione, la probabilità di attesa è dunque la probabilità che ci siano almeno N utenti nel sistema, ossia un numero di utenti non inferiore al numero di serventi (il che corrisponde appunto a dire che tutti i serventi sono occupati).

E' facile allora valutare questa probabilità di attesa, che indichiamo con p_A: in base a quanto abbiamo appena detto, essa sarà

$$p_A=p_N+p_{N+1}+\ldots=\sum_{k=N}^{\infty}p_k$$

Ponendo $\rho=\frac{\lambda}{N\mu}$ e supponendo che sia $\rho<1$, ciò che si trova dopo aver sostituito l'espressione della generica probabilità di stato p_k per $k\geq N$, è

$$\boxed{p_A=\frac{(\rho N)^N}{(1-\rho)N!}p_0}$$

Questa prende il nome di "**formula di Erlang di tipo C con parametri λ e μ**" e si indica col simbolo $\mathbf{E_{2,N}(\lambda/\mu)}$:

$$p_A=E_{2,N}(\lambda/\mu)=\frac{(\rho N)^N}{(1-\rho)N!}p_0$$

Facciamo osservare che questo tipo di coda è spesso utilizzata per modellare le centrali telefoniche: infatti, quando gli utenti che usufruiscono della generica centrale non sono moltissimi e la centrale stessa dispone di opportune risorse di memoria, è possibile ipotizzare che essa abbia una capacità di memorizzazione infinita, intendendo semplicemente il fatto che essa non si trova praticamente mai a respingere delle richieste di servizio, ma tutt'al più a metterle in attesa. Il dimensionamento, allora, della centrale viene spesso effettuato partendo proprio dalla probabilità di attesa p_A: si fissa un valore massimo tollerabile per p_A e si procede al dimensionamento.

Possiamo inoltre valutare qualche altro parametro caratteristico del sistema. Valutiamo, ad esempio, il numero medio E[Q] di utenti che si trovano in coda (cioè in attesa di servizio): perché ci possano essere utenti in coda, è necessario che tutti i serventi siano occupati, ossia che il numero di utenti nel sistema sia almeno pari ad N. Se Q è la variabile che individua il numero di utenti in coda, il suo valore medio sarà, per definizione, dato dalla somma dei valori assumibili da Q stessa, pesati per le rispettive probabilità di essere assunti:

$$E[Q] = \sum_{i=1}^{\infty} i \cdot P(Q = i)$$

Dato che Q può assumere i valori 0 (nessun utente in coda, ossia non più di N utenti nel sistema), 1 (N+1 utenti nel sistema, di cui 1 in coda), 2 (N+2 utenti nel sistema, di cui 2 in coda), abbiamo che

$$E[Q] = 0 \cdot P(0 \le k \le N) + 1 \cdot P(k = N+1) + 2 \cdot P(k = N+2) + ...$$

dove k è il numero di utenti presenti complessivamente nel sistema.

In forma compatta, essendo nullo il primo termine di quella somma, possiamo scrivere che

$$E[Q] = \sum_{i=1}^{\infty} i \cdot P(Q = i) = \sum_{i=1}^{\infty} i \cdot p_{N+i}$$

Dobbiamo ora sostituire l'espressione delle probabilità di stato. Facendo la sostituzione ed i dovuti passaggi, il risultato finale è che

$$E[Q] = \frac{\rho}{1-\rho} E_{2,N}(\lambda/\mu) = \frac{\rho}{1-\rho} p_A$$

Come è intuitivo che accadesse, il risultato dipende strettamente dalla probabilità di attesa.

DETERMINAZIONE DEL TRAFFICO IN USCITA E VERIFICA DELLA STABILITÀ

Dato che il sistema ha capacità infinita di memorizzazione, per cui accetta tutte le richieste di servizio, è possibile che sia instabile. Supponiamo invece che sia stabile, il che equivale, come detto in precedenza, ad ipotizzare che risulti $\lambda/N\mu<1$.

Sotto questa condizione, data appunto la stabilità, il traffico in ingresso λ è pari al traffico in uscita γ. Calcoliamo allora esplicitamente questo traffico in uscita: in base agli stessi criteri adottati nei casi precedenti, abbiamo che

$$\gamma = 0 \cdot p_0 + \mu p_1 + 2\mu p_2 + 3\mu p_3 + ... + N\mu p_N + N\mu p_{N+1} + N\mu p_{N+2} + ...$$

Abbiamo evidentemente tenuto conto che il numero medio di utenti serviti dal sistema cresce fin quando il numero di utenti presenti nel sistema cresce ma non supera N; quando invece tale numero raggiunge e supera N, tutti i serventi sono occupati, per cui la frequenza di servizio rimane costante sul valore Nμ.

Scrivendo in forma più compatta quella somma, possiamo adottare due sommatorie, nel modo seguente:

$$\gamma = \sum_{k=1}^{N-1} k\mu \cdot p_k + \sum_{k=N}^{\infty} N\mu \cdot p_k = \mu \sum_{k=1}^{N-1} k \cdot p_k + N\mu \sum_{k=N}^{\infty} p_k = \mu \sum_{k=1}^{N-1} k \cdot p_k + N\mu p_A$$

Anche in questo caso è comparsa la probabilità di attesa.

Sistemi a coda di tipo M/M/∞/∞/∞

Un sistema a coda di tipo M/M/∞/∞/∞ ha le seguenti caratteristiche:

- il traffico in ingresso è un processo di Poisson che supponiamo abbia intensità λ (=numero medio di utenti che entrano nel sistema nell'unità di tempo);
- il tempo di servizio è una variabile aleatoria con distribuzione esponenziale che supponiamo abbia parametro μ (=numero medio di utenti serviti nell'unità di tempo);
- ci sono ∞ serventi nel sistema: questa è una idealizzazione, ma vale in questi casi pratici in cui il numero di serventi è sufficientemente alto da poter ritenere che ogni richiesta di servizio venga immediatamente soddisfatta o, ciò che è lo stesso, da poter ritenere che la probabilità di attesa in coda da parte di un utente sia piccolissima;
- il sistema è senza perdite, in quanto ha una capacità di memorizzazione (intesa come numero massimo di utenti che possono essere contemporaneamente presenti nel sistema, sia sotto servizio sia in attesa di servizio) infinita;
- ci sono ∞ potenziali utenti del sistema.

Si tratta di un sistema a infiniti stati caratterizzato dalle seguenti frequenze di transizione:

$$\begin{array}{ll} \gamma_{i,i+1} = \lambda & \forall i \\ \gamma_{i,i-1} = i\mu & \forall i \\ \gamma_{i,i-k} = 0 & \forall k \geq 2 \\ \gamma_{i,i+k} = 0 & \forall k \geq 2 \end{array}$$

Giustifichiamo ancora una volta il motivo per cui si hanno quelle particolari frequenze di servizio $\gamma_{i,i-1}$: è infatti intuitivo accorgersi che, se ci sono n utenti (stato n), con $1 \leq n \leq \infty$, n serventi sono attivi e i rimanenti sono inattivi, per cui il numero medio di utenti serviti dal sistema nell'unità di tempo è pari alla somma del numero medio di utenti serviti da ciascun servente nell'unità di tempo, ossia nμ. Questo spiega perché le frequenze di servizio vanno aumentando man mano che aumenta il numero di utenti.

Per comprendere ancora meglio il concetto, è sufficiente fare un confronto con quanto abbiamo trovato nel paragrafo precedente per i sistemi di tipo **M/M/N/∞/∞**: in quel caso, il numero di serventi è finito, per cui si arriva allo stato N a partire dal quale tutti i serventi sono occupati e quindi il numero medio di utenti serviti nell'unità di tempo raggiunge il suo valore massimo Nμ. In questo caso, invece, avendo ∞ serventi, c'è sempre qualche servente libero, per cui il numero medio di utenti serviti nell'unità di tempo non può che aumentare progressivamente all'aumentare del numero di utenti presenti.

Per quanto riguarda invece le probabilità asintotiche, si trova quanto segue:

$$p_k = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k e^{-\frac{\lambda}{\mu}}$$

Questo è lo stesso risultato trovato per i sistemi di tipo M/M/1/∞/∞ ad arrivi rallentati. Se K è la variabile che indica il numero di utenti nel sistema (per cui pk è la probabilità che ci siano K=k utenti), allora quella formula ci dice che K ha una distribuzione di Poisson con intensità λ/μ. Da qui scaturisce che il numero medio di utenti presenti nel sistema è E[K]=λ/μ.

Notiamo inoltre che essendo il sistema di tipo non bloccante e avendo a disposizione infiniti serventi, si tratta di un sistema sicuramente stabile, per il quale perciò il traffico in ingresso è sicuramente pari a quello in uscita. Possiamo rendercene conto facilmente, verificando che risulta λ=γ. Calcoliamo infatti il traffico in uscita γ:

$$\gamma = 0\cdot p_0 + \mu p_1 + 2\mu p_2 + 3\mu p_3 + \ldots + N\mu p_N + (N+1)\mu p_{N+1} + (N+2)\mu p_{N+2} + \ldots = \sum_{k=1}^{\infty} k\mu\cdot p_k = \mu\sum_{k=1}^{\infty} k\cdot p_k$$

Quella sommatoria non è altro che il valor medio di K, ossia il numero medio di utenti presenti nel sistema (tutti sotto servizio): abbiamo detto prima che E[K]=λ/μ, per cui deduciamo che effettivamente risulta γ/λ.

Segnaliamo infine che gli stessi risultati ricavati per questo tipo di sistema a coda valgono anche per sistemi di tipo **M/G/∞/∞/∞**, per le quali il tempo di servizio, pur essendo di tipo generico, abbiamo però un valor medio 1/μ.

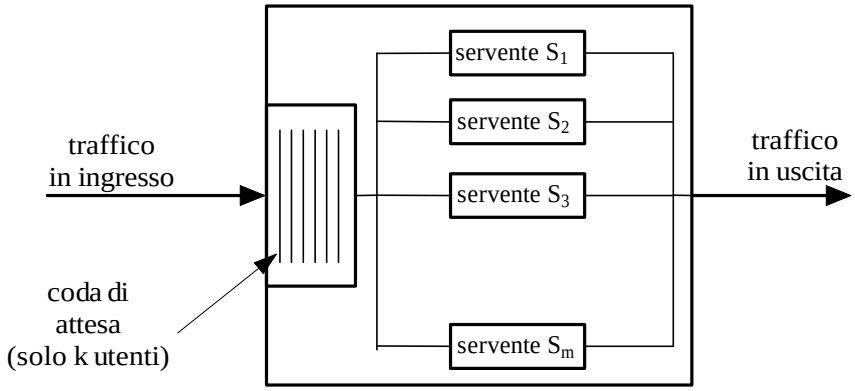
Sistemi a coda di tipo M/M/m/k/M

DESCRIZIONE

L'ultimo tipo di sistema a coda di cui ci occupiamo ha le seguenti caratteristiche:

- il traffico in ingresso al sistema è un processo di Poisson;
- il tempo di servizio è una variabile aleatoria con distribuzione esponenziale che supponiamo abbia parametro μ (=numero medio di utenti serviti nell'unità di tempo);
- ci sono m serventi nel sistema;
- il sistema è con perdite, in quanto ha una capacità di memorizzazione finita e pari k (ovviamente, sappiamo che deve risultare k≥m);

- ci sono M (numero finito e maggiore di m) potenziali utenti del sistema.



Per poter studiare il sistema e, in particolare, per stabilire se e come si possano usare ancora una volta le proprietà delle catene di Markov, dobbiamo dire qualcosa in più circa il processo degli arrivi (o richieste di servizio) nel sistema.

Osserviamo subito che, se tutti ed m i serventi sono occupati, nel sistema ci potranno essere al più K-m utenti in attesa; se sono proprio K-m, allora all'esterno del sistema abbiamo M-k utenti *esterni* al sistema[6]; questi utenti sono esterni o perché non hanno fatto alcuna richiesta di servizio oppure perché l'hanno fatta ma sono stati respinti (perché il sistema era già pieno).

In generale, il generico utente può trovarsi o all'interno del sistema o all'esterno del sistema. Dato che si presume che ogni utente, con una certa frequenza, faccia comunque delle richieste di servizio e che una parte di esse venga soddisfatta, possiamo ritenere che tale generico utente si trovi talvolta **libero** (cioè fuori dal sistema) e talvolta **occupato** (cioè nel sistema, sotto servizio o in coda).

Queste considerazioni ci servono per caratterizzare il processo degli arrivi al sistema, il quale processo è determinato proprio dal comportamento degli utenti. Facciamo allora l'ipotesi che tutti gli utenti abbiano lo stesso comportamento statistico; in particolare, se indichiamo con T_L la variabile aleatoria che indica il tempo durante il quale un utente è libero, assumiamo che T_L sia una variabile esponenziale con parametro λ, il che significa dire che

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

e quindi anche, ovviamente, che il tempo medio durante il quale il generico utente è libero valga $E[T_L]=1/\lambda$.

Con queste premesse, ci si rende conto facilmente che il sistema può ancora una volta essere modellato come una catena di Markov, nella solita ipotesi che lo stato del sistema corrisponda al numero di utenti presenti nel sistema stesso (per cui lo stato può assumere i valori da 0 a k).

FREQUENZE DI TRANSIZIONE DI STATO

Il passo successivo, così come fatto nei casi precedenti, consiste nel calcolare le frequenze di transizione di stato e, in particolare, quelle tra stati adiacenti, in quanto quelle tra stati non adiacenti sono ancora una volta nulle:

[6] Per utente *esterno* al sistema (o anche utente **libero**) intendiamo un utente che non sia né in coda né sotto servizio; al contrario, un utente *interno* al sistema (o anche utente **occupato**) è un utente che o sta ricevendo servizio oppure è in attesa di riceverlo

$$\gamma_{i,i-k} = 0 \quad \forall k \geq 2$$
$$\gamma_{i,i+k} = 0 \quad \forall k \geq 2$$

Per quanto riguarda le frequenze di servizio, non ci grossi problemi perché valgono le stesse considerazioni dei casi precedenti: essendo la capacità del sistema superiore al numero di serventi, possiamo avere utenti sotto servizio e utenti in coda; se ci sono utenti in coda, significa che tutti ed m i serventi sono occupati, per cui la massima frequenza di servizio è mμ. Possiamo perciò scrivere

$$\gamma_{i,i-1} = i\mu \qquad \forall i \leq m\text{-}1$$
$$\gamma_{i,i-1} = m\mu \qquad \forall i > m\text{-}1$$

Passiamo alle frequenze di transizione in avanti. Siamo costretti a ragionare in modo rigoroso: consideriamo ad esempio γ_{01}, ossia il numero di volte in cui il sistema passa dallo stato 0 allo stato 1 nell'unità di tempo: sappiamo di poter scrivere che

$$\gamma_{01} = \lim_{\delta \to 0} \frac{p_{01}(\delta)}{\delta}$$

Dobbiamo calcolare $p_{01}(\delta)$, ossia la probabilità che un utente chieda servizio al sistema in un intervallo di durata δ infinitesima. Essendoci M utenti potenziali, quella probabilità corrisponde alla probabilità che uno qualsiasi tra tali M utenti chieda servizio al sistema, il che significa scrivere che

$$p_{01}(\delta) = P\left(\begin{matrix} 1° \text{ utente} \\ \text{chiede servizio} \end{matrix} \cup \begin{matrix} 2° \text{ utente} \\ \text{chiede servizio} \end{matrix} \cup ..... \cup \begin{matrix} M° \text{ utente} \\ \text{chiede servizio} \end{matrix} \right)$$

Adesso possiamo sfruttare un noto risultato di probabilità: dati 3 eventi A,B e C, la probabilità dell'evento che si ottiene dalla loro unione è

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Applicando questa formula al nostro caso, possiamo sicuramente trascurare le probabilità incrociate (del tipo A∩B) e il termine finale (del tipo A∩B∩C) dato dall'intersezione dei vari eventi[7], in quanto si tratta senz'altro di infinitesimi di ordine superiore rispetto ai primi termini, ossia i termini

$$p_{01}(\delta) = P\left(\begin{matrix} 1° \text{ utente} \\ \text{chiede servizio} \end{matrix} \right) + P\left(\begin{matrix} 2° \text{ utente} \\ \text{chiede servizio} \end{matrix} \right) + .. + P\left(\begin{matrix} M° \text{ utente} \\ \text{chiede servizio} \end{matrix} \right) = \sum_{i=1}^{M} P\left(\begin{matrix} i° \text{ utente} \\ \text{chiede servizio} \end{matrix} \right)$$

Adesso, le M probabilità che qui compaiono sono identiche, in quanto identico (statisticamente) è il comportamento dei vari utenti: scriviamo perciò che

$$p_{01}(\delta) = M \cdot P\left(\begin{matrix} i° \text{ utente} \\ \text{chiede servizio} \end{matrix} \right) \qquad \forall i \in (1,2,...,M)$$

[7] Questa probabilità è un infinitesimo di ordine superiore per un semplice motivo: dato che l'evento per cui un utente chiede servizio è del tutto indipendente dall'evento per cui anche un altro utente chiede servizio, la probabilità congiunta è il prodotto delle singole probabilità; tale prodotto, rispetto al generica di tali probabilità, è sicuramente un infinitesimo di ordine superiore

L'ipotesi che stiamo facendo è inoltre quella per cui il generico utente i-simo chieda servizio in un intervallo di tempo di durata δ: avendo indicato con T_L il tempo durante il quale l'utente generico è libero, la probabilità che esso chieda servizio in un intervallo di durata δ equivale alla probabilità che $T_L \leq \delta$:: tenendo conto che T_L è di tipo esponenziale, abbiamo che

$$p_{01}(\delta) = M \cdot P\begin{pmatrix} \text{i° utente} \\ \text{chiede servizio} \end{pmatrix} = M \cdot P(T_L \leq \delta) = M \cdot \left(1 - e^{-\lambda\delta}\right) = M \cdot \left(1 - \left(1 - \lambda\delta + o(\delta)\right)\right) = M\lambda\delta$$

Concludendo, avendo detto che $\gamma_{01} = \lim_{\delta \to 0} \frac{p_{01}(\delta)}{\delta}$, deduciamo che **$\gamma_{01} = M\lambda$**.

Ripetendo il discorso per le altre $\gamma_{i,i+1}$, si trova che

$$\gamma_{i,i+1} = (M - i)\lambda \qquad \forall i$$

Possiamo perciò riepilogare i risultati circa le frequenze di transizione nel modo seguente:

$$\begin{aligned}
&\gamma_{i,i+1} = (M - i)\lambda && \forall i \\
&\gamma_{i,i-1} = i\mu && \forall i \leq m\text{-}1 \\
&\gamma_{i,i-1} = m\mu && \forall i > m\text{-}1 \\
&\gamma_{i,i-k} = 0 \quad \forall k \geq 2 && \\
&\gamma_{i,i+k} = 0 \quad \forall k \geq 2 &&
\end{aligned}$$

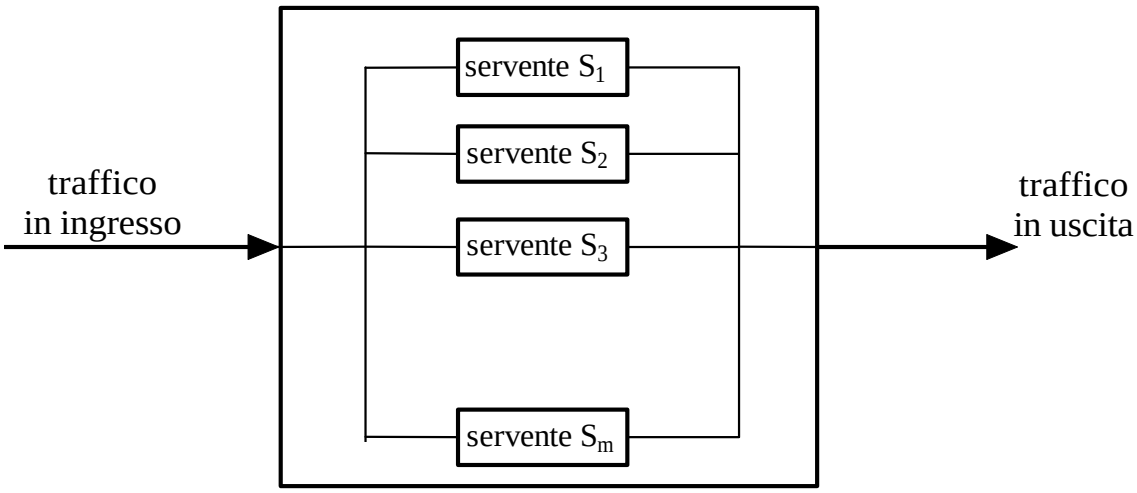
PROBABILITÀ ASINTOTICHE E PROBABILITÀ DI BLOCCO

Note le frequenze di transizione, dobbiamo applicare l'equazione di bilancio del flusso per ricavare le probabilità asintotiche. Si ricavano due distinte espressioni di tali probabilità, a seconda che si consideri un numero di utenti inferiore o superiore al numero *m* di serventi:

$$p_i = p_0 \binom{M}{i} \left(\frac{\lambda}{\mu}\right)^i \qquad i \leq m$$

$$p_i = p_0 \binom{M}{i} \left(\frac{\lambda}{\mu}\right)^i \frac{i!}{m!} m^{m-i} \qquad i \leq m$$

C'è evidentemente da determinare anche p_0, ma il calcolo (ottenuto imponendo la solita condizione di normalizzazione) è abbastanza complesso. Per semplicità, ci mettiamo allora in un caso particolare, ossia quello in cui il numero di serventi coincide con la capacità di memorizzazione del sistema: ciò significa che **m=k** (non c'è coda di attesa, per cui le richieste eccedenti vengono respinte):



In questa particolare situazione, si trova che la generica probabilità di stato, per i≤m=k, vale

$$p_i = p_0 \left(\frac{\lambda}{\mu}\right)^i \binom{M}{i}$$

Imponendo la normalizzazione, non ci sono molti conti da fare: si ha infatti che

$$\sum_{i=0}^{m} p_i = 1 \longrightarrow \sum_{i=0}^{m} p_0 \left(\frac{\lambda}{\mu}\right)^i \binom{M}{i} = 1 \longrightarrow p_0 = \frac{1}{\sum_{i=0}^{m} \left(\frac{\lambda}{\mu}\right)^i \binom{M}{i}} \longrightarrow \boxed{p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n \binom{M}{n}}{\sum_{i=0}^{m} \left(\frac{\lambda}{\mu}\right)^i \binom{M}{i}}}$$

Quella ottenuta è la cosiddetta "**formula di Engset**".

E' interessante calcolare anche la probabilità di blocco, ossia la probabilità che una richiesta di servizio venga respinta: questo avviene quando gli utenti nel sistema sono *m* (pari al numero di serventi), per cui coincide con p_m:

$$p_{BLOCCO} = p_m = \frac{\left(\frac{\lambda}{\mu}\right)^m \binom{M}{m}}{\sum_{i=0}^{m} \left(\frac{\lambda}{\mu}\right)^i \binom{M}{i}}$$

Autore: **SANDRO PETRIZZELLI**
e-mail: sandry@iol.it
sito personale: http://users.iol.it/sandry
succursale: http://digilander.iol.it/sandry1