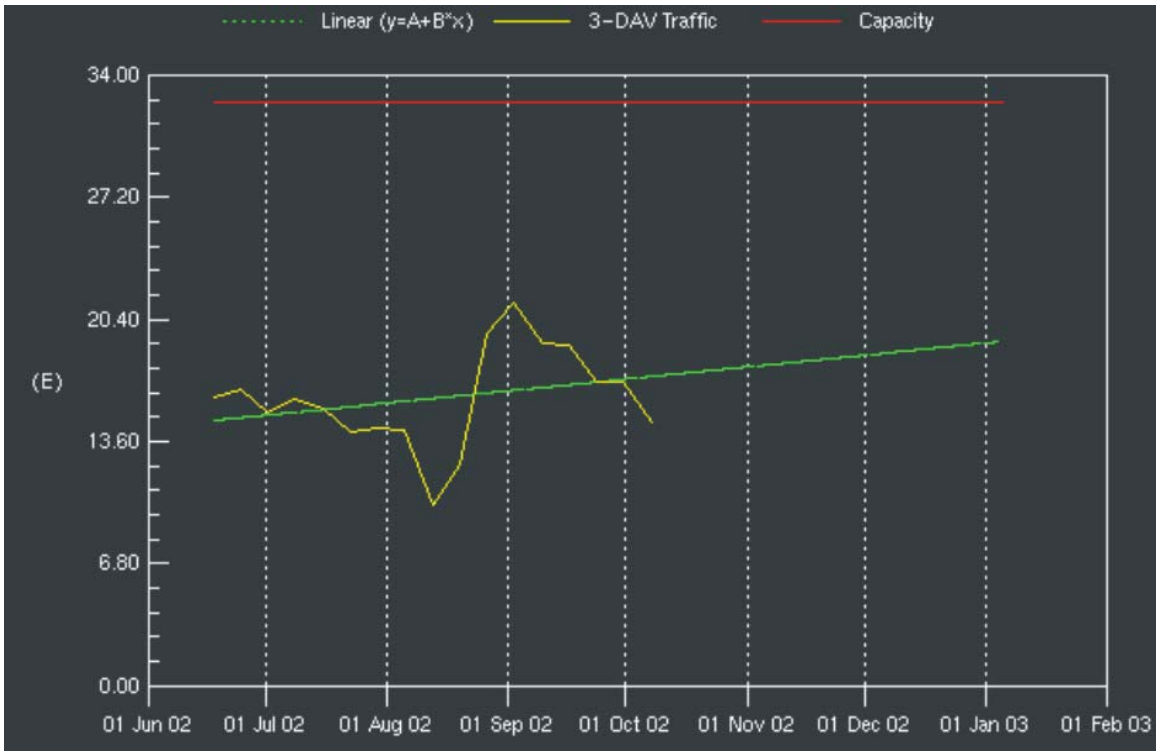
Ciro Carbone

TEORIA DEL TRAFFICO TELEFONICO



Elementi di base per l'analisi di GoS e QoS per le reti telefoniche a commutazione di circuito

Creare una qualsiasi rete telefonica non richiede semplicemente l'installazione degli apparati e le fasi di marketing per la ricerca della clientela. Una rete dovrà essere gestita (*network management*) ed ottimizzata (*performance management*) , affinché possa essere efficiente nel tempo e soddisfacente per i clienti. Se, infatti, si provvedesse ad installare le sole apparecchiature, come si farebbe poi a mantenerla ed a conoscere se, col trascorrere del tempo, un apparato (o un insieme di essi) è malfunzionante e si rende necessaria una riparazione o sostituzione o riconfigurazione ? Ed ancora, come si farebbe a conoscere se la performance di una parte o dell'intera rete è quella desiderata dall'operatore e dal cliente? E' quindi necessario che tutte le apparecchiature del network, oltre a svolgere la propria funzione, provvedano ad inviare uno "status" aggiornato sulla propria "salute" e sulle proprie modalità di funzionamento ad un sistema di database centralizzato. Questo sistema centralizzato, che raccoglie dati di allarmistica e di performance è spesso chiamato NMS (*Network Menagement System*). La centralizzazione consente ai tecnici preposti di avere il controllo reale circa il funzionamento delle singole apparecchiature e dell'intera rete nel suo complesso al fine di decidere le opportune misure per la riparazione, il miglioramento (*improvement*) e la caccia ai guasti (*trouble-shooting*). Ecco perché in una rete telefonica gli apparati, oltre a svolgere la propria attività principale preposta, cioè quella di garantire il servizio di traffico telefonico, avranno parallelamente anche la funzione di comunicare a chi è addetto alla manutenzione ed alla gestione tutta una serie di informazioni supplementari sulle proprie modalità di funzionamento. Gli apparati trasmettono queste informazioni grazie a delle operazioni semplici: il conteggio di tutti gli eventi che accadono in essi. Già nelle vecchie reti telefoniche analogiche, avere queste informazioni supplementari era una necessità. Il conteggio degli eventi avveniva tramite contatori (*counters*) implementati nell'hardware degli apparati. Diversamente, oggi, è il software (*SW*) che svolge per queste operazioni e di fatto il conteggio degli eventi avviene grazie a dei blocchi di programmi SW (*program blocks*) che "vigilano" costantemente su tutto ciò che avviene al loro interno mentre essi continuano a svolgere la specifica funzione di servizio al traffico telefonico. Saranno i numeri prodotti dai contatori SW a costituire una mole di dati a noi utili per percepire la reale bontà di funzionamento del network e quindi la qualità del servizio (*QoS*, cioè *Quality of Service*), il giusto dimensionamento (*GoS* = *Grade of Service*) e per stabilire, in base ai targets qualitativi scelti dal network-operator, quali dovranno essere gli indicatori più importanti da assurgere sempre e costantemente come riferimento (*KPIs* = *Key Performance Indicators*) (**figura 1** e **figura2**).

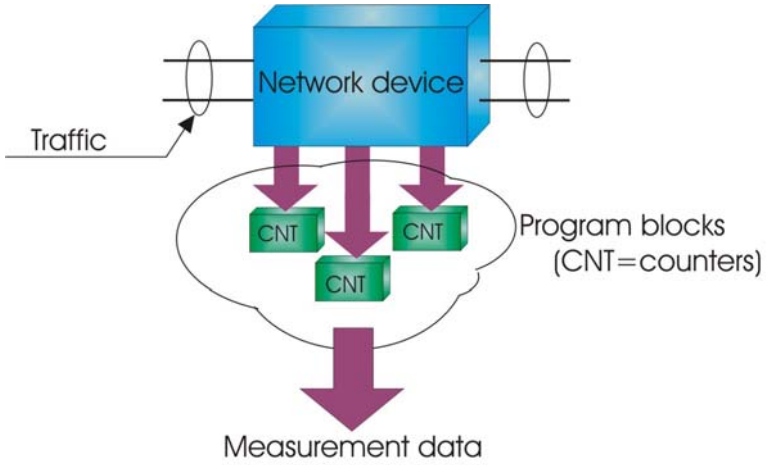


Figura1

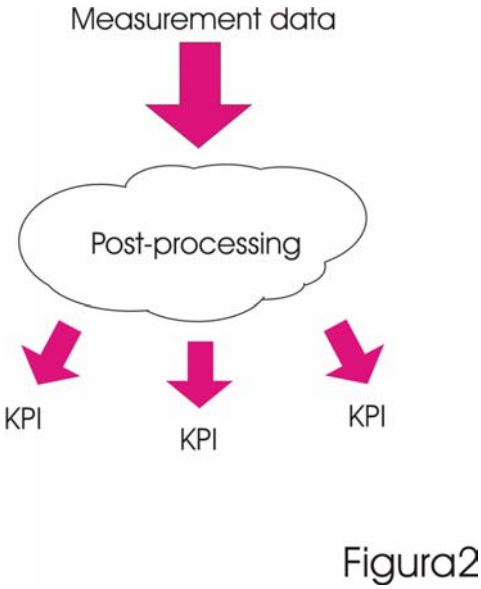


Figura2

CONCETTI BASILARI DELLE MISURE DI TRAFFICO TELEFONICO DI UNA RETE RADIOMOBILE

- ***Le unità di misura del traffico telefonico: ERLANG e CCS***

Con la nascita delle reti telefoniche nacque anche l'esigenza di poter misurare la quantità del traffico. Questo sia per un fattore meramente tecnico, sia per avere un riferimento durante il dimensionamento progettuale delle linee e delle apparecchiature. Non solo chi costruisce apparecchiature per reti telefoniche ma anche chi le gestisce deve avere la possibilità di quantificare il volume di traffico che investe i propri dispositivi e le proprie linee. Ci sono due unità di misura del traffico telefonico maggiormente utilizzate:

- l' *Erlang* (o in forma abbreviata *Erl*), prevalentemente utilizzato in Europa e nel mondo
- il *CCS* prettamente americano

La prima è oggi sicuramente la più utilizzata in assoluto ed è nata in Europa, mentre il *CCS* americano è stata, sino a pochi anni fa, l'unità di misura del traffico usato prevalentemente dalla statunitense **AT&T** e dalla **Bell Canada**.
Strettamente parlando e nella sua definizione originaria un *Erlang* (1 *Erl*) rappresenta il traffico generato da una conversazione telefonica tra due utenti (*voice-path*) per un tempo continuo e mai interrotto di un'ora.
Più in generale 1 *Erl* è anche il volume di traffico cumulativo prodotto da tanti utenti telefonici in un'ora.
Per esempio se un gruppo di utenti fa 30 chiamate in un'ora e la durata media di ogni chiamata è di quattro minuti, allora il volume di traffico prodotto in *Erl* equivale a:

minuti di traffico in un'ora: 30 x 4 = 120 minuti
ore di traffico: 120/60 = 2 ore di traffico continuato
figura di traffico o volume: **2 Erl**

Nel nostro esempio trenta utenti contribuiscono, con chiamate di quattro minuti ciascuna, a creare un volume di traffico pari a 2 *Erl* ma lo stesso volume di traffico potrebbe essere prodotto anche e più semplicemente da due utenti che contemporaneamente impegnano due canali telefonici per un'ora conversando con altri due corrispondenti utenti all'altro capo.

In una rete radiomobile, una BTS con due trasmettitori di 8 canali (o Time-Slots *TS*) avrà un totale di 16 *TS* disponibili. Escludendo un *TS* riservato alla segnalazione BCCH ed uno alla segnalazione SDCCH rimangono 14 *TS* usufruibili per il traffico telefonico e, quindi, 14 *Erl* è il limite massimo di volume di traffico sopportabile dalla BTS. Se sulla BTS è implementato l'*Half Rate* (cioè ogni canale può servire due chiamate, poiché è stata dimezzata la larghezza di banda di ogni utente) il massimo traffico sopportabile raddoppierà a 28 *Erl*. Continuando con gli esempi se in una linea "viaggia" un flusso telefonico PCM da 2Mbit/sec. sono disponibili un totale di 30 canali per il traffico telefonico: il massimo volume di traffico sopportabile da questa linea è di 30 *Erl*.
In America l'unità di misura del traffico è il *CCS* (*Centi-Call Second,* ovvero cento secondi di chiamata). Nell'unità di misura americana il traffico di una chiamata non viene riferito nell'intervallo temporale di un'ora ma di cento secondi. Quindi, il traffico di una chiamata di un'ora, cioè di 1 *Erl*, viene diviso in frazioni di 100 secondi ottenendo il *Centi-Call*. In un'ora di traffico telefonico avremo quindi 36 *Centi-call*:

- ✓ 1 *Erlang* = 36 *CCS*
- ✓ 1 *CCS* = 0,0278 *Erlangs*

- ❑ ***Traffico offerto e smaltito. L'assunzione di ergodicità del traffico ed il processo di Poisson***

Nell'ingegneria del traffico uno dei maggiori argomenti di studio è il rapporto *traffico smaltito/traffico offerto*. Innanzitutto, per comprendere bene la complessità e l'importanza del rapporto *traffico smaltito/traffico offerto* in un qualsiasi sistema telefonico è necessario proseguire per piccoli passi.
Un qualsiasi utente che vuole effettuare una chiamata telefonica, nel momento in cui lo farà, richiederà per sé una risorse fisica alla rete. Un network-operator dovrà, quindi, garantire la quantità di risorse fisiche affinché si soddisfino gli utenti che ne faccino richiesta.
Ma se si osserva come vengono impiegati oggi i mezzi di comunicazione nelle relazioni sociali ci si rende conto che la richiesta di uso del servizio telefonico, non è determinabile a priori, poiché ha caratteristiche del tutto arbitrarie secondo le necessità proprie dell'utenza. Questa variabilità intrinseca di un servizio di telecomunicazioni richiederebbe un'elevatissima disponibilità di risorse (al limite infinita), in contrasto con la logica e realistica esigenza di economicità e di ottimizzazione.
Tuttavia, come è possibile conciliare l'esigenza di un servizio telefonico efficiente che deve garantire la disponibilità di risorse agli utenti con l'arbitrarietà e l'indeterminabilità della richiesta?
E' questo il primo vero grande problema da risolvere nel progetto di una rete di telecomunicazioni: il suo dimensionamento, ovvero il *GoS* (*Grade of Service* = Grado di Servizio) da offrire ai clienti.
Partiamo dal concetto di *traffico offerto*: volendo a priori semplificare il concetto di *traffico offerto*, nella **figura 3**, supponiamo che un certo numero di utenti di una determinata area geografica voglia fare uso del servizio telefonico effettuando delle chiamate ed impegnando, quindi, delle risorse fisiche della rete messa a disposizione. Il *traffico offerto* coincide sostanzialmente con le esigenze degli utenti telefonici ed è così chiamato perché è la sorgente di traffico telefonico offerta alle nostre risorse, cioè alla nostra rete.

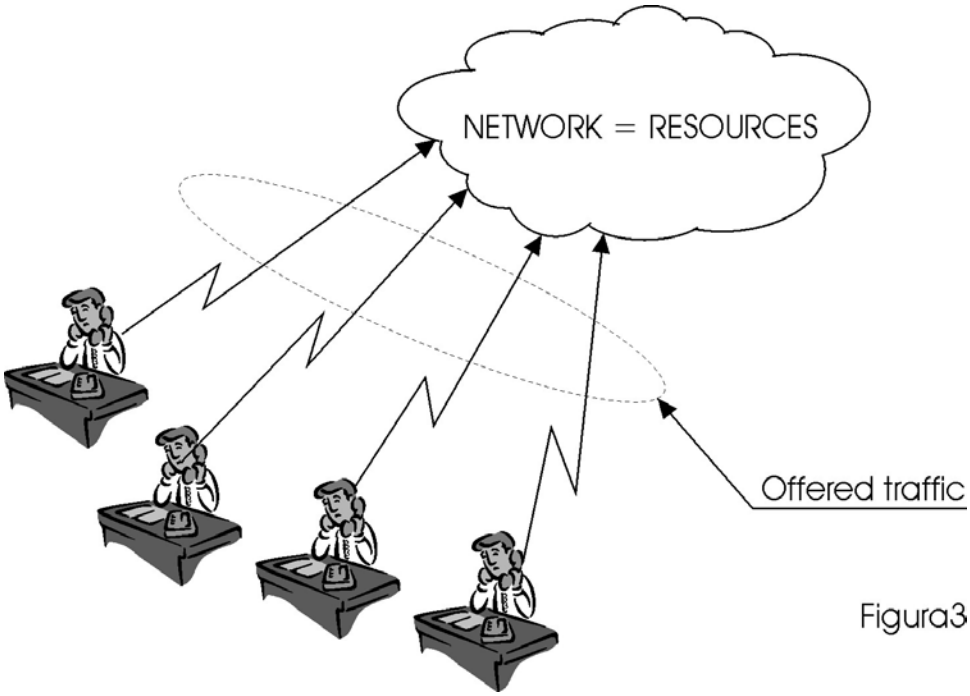


Figura3

E' chiaro che ogni network-operator desidererebbe poter garantire il miglior *GoS* e teoricamente poter garantire, anche nei momenti più critici, tante risorse quante ne sono realmente richieste dall'utenza ma ciò, come già detto, è praticamente impossibile. Ne consegue che in un sistema di telecomunicazioni potrà sempre verificarsi una condizione di congestione. Ad esempio, se nell'area di copertura di una BTS radiomobile sono disponibili 80 canali telefonici per il traffico ed, a causa di un evento straordinario, 100 utenti decidono insieme di effettuare chiamate, 20 di questi saranno esclusi (in genere quelli che hanno richiesto il servizio più tardivamente). In tal caso si genererà una condizione di congestione ed i 20 utenti non potranno essere soddisfatti.
Proprio nel tentativo di studiare l'arbitrarietà del *traffico offerto*, l'ingegneria del traffico si è sempre basata sugli studi di matematica probabilistica a partire da quelli del francese Siméon Denis Poisson. Durante il progetto di una rete, proprio a causa dell'imprevedibilità del *traffico offerto*, si conviene nell'utilizzo di un modello di matematica probabilistica in grado di semplificarene la natura: è questo il modello del *traffico poissoniano*. Con l'adozione del modello poissoniano si è reso possibile lo studio analitico del traffico nelle telecomunicazioni poiché Poisson rese adattabile ad un sistema ergodico e stazionario l'aleatorietà di molti fenomeni casuali.

Siméon Denis Poisson nacque nel 1781 a Pithiviers in Francia. E' stato uno dei più grandi studiosi di fisica e matematica della storia. Si narra che a causa della sua natura maldestra fu penalizzato nel mestiere della medicina e della chirurgia verso cui il padre cercò di indirizzarlo sin da fanciullo. Sconsigliato dallo zio medico e tornato a casa poté, però, dare libero sfogo alla sua indole: lo studio della matematica. Ammesso all'**Ecole Polytechnique** a Parigi, ebbe come insegnanti insigni studiosi come Lagrance e Laplace. Questi ultimi da subito notarono in Siméon Denis un vero talento. La sua fu una carriera repentina ed impressionante, specie se si considera che nei primi momenti universitari ebbe grandi difficoltà nell'approccio agli studi matematici, dovute alla carenza di basi propedeutiche. Dopo il conseguimento del titolo accademico e circondato già da una rinomata fama fu assistente per quattro anni del celebre Jean Baptiste Fourier. Quando nel 1808 Fourier fu allontanato a Grenoble da Napoleone per le sue idee politiche, egli ne ereditò la cattedra. Successivamente divenne il primo professore di meccanica alla **Sorbona**. Tantissimi sono i suoi scritti ed appunti. Deve la sua fama ai contributi teoretici nel campo dell'elettricità, del magnetismo e della matematica pura, sebbene si fosse occupato anche di calcolo delle variazioni, astronomia, geometria differenziale e teorie delle probabilità. Nell'ingegneria del traffico delle telecomunicazioni il suo tributo è dato dalle teorie delle probabilità ed, in special modo, al suo celebre schema di distribuzione binominale statistica, oggi noto come "distribuzione di Poisson" o "processo poissoniano". In questo campo lo studioso francese osservò tutti quei fenomeni aleatori (eventi) che nascono e si presentano in modo indipendente e del tutto casuale (*Poisson Arrivals*) e la sua curiosità matematica lo spinse a cercare un modello analitico che gli desse la possibilità di calcolare la quantità di probabilità che si potesse verificare un determinato numero di eventi in un dato intervallo di osservazione. Morì a Parigi nel 1840.

Per illustrare brevemente il processo di Poisson, partiamo dal presupposto che il matematico francese studiò tutti quei fenomeni aleatori in cui ci sono eventi che nascono secondo una logica del tutto casuale e con totale indipendenza reciproca (*Poisson Arrivals*). Proprio per la casualità con cui nascono, per *Poisson*

Arrivals possiamo immaginare le chiamate telefoniche che partono da una determinata area geografica o le persone che salgono su un mezzo pubblico o ancora fare riferimento alle automobili che sfrecciano su una autostrada mentre vengono osservate da un ipotetico osservatore.
Innanzitutto è necessario premettere che la descrizione del "processo poissoniano" qui di seguito riportata non si attiene ad una trattazione matematica rigorosa, bensì è sviluppata per essere quanto più possibilmente intuitiva. Inoltre, utilizziamo costantemente l'esempio telefonico che costituisce il campo concreto di applicazione della teoria matematica in esame.
Nel "processo di Poisson" l'analisi delle probabilità di accadimento degli eventi presuppone innanzitutto che siano scontate tre condizioni essenziali:

1) che sia molto alto il numero di eventi che con probabilità possano accadere (ovvero, in analogia col mondo telefonico, che sia molto elevato il numero di utenti che può originare una chiamata);

2) che il verificarsi di un evento non condizioni la stabilità del sistema (ovvero che una chiamata telefonica di un utente impatti minimamente sulle risorse del sistema telefonico oppure, ancora, poter dire che il sistema telefonico è talmente capiente per cui la chiamata di un utente non potrà congestionarlo)

3) che il verificarsi degli eventi siano casuali ed incondizionati da fattori esterni, cioè del tutto indipendenti gli uni dagli altri (ovvero che la decisione di un utente di generare una chiamata non condizioni l'uso del telefono per altri utenti)

In realtà la terza condizione poco si addice al mondo telefonico dove la decisione di un utente di chiamare condizionerà inevitabilmente l'altro utente chiamato. Pur tuttavia, se consideriamo grande il sistema telefonico ed alto il numero di utenti la interdipendenza telefonica può divenire trascurabile e quindi, con buona approssimazione, si può "esportare" il processo poissoniano anche per l'analisi della generazione del traffico telefonico offerto.
Supponiamo che in un intervallo di tempo Δt pari a 5 minuti partano 30 chiamate telefoniche generate da un certo numero di utenti imprecisato e distribuite nel tempo in modo del tutto illogico ed aleatorio (**figura 4**).

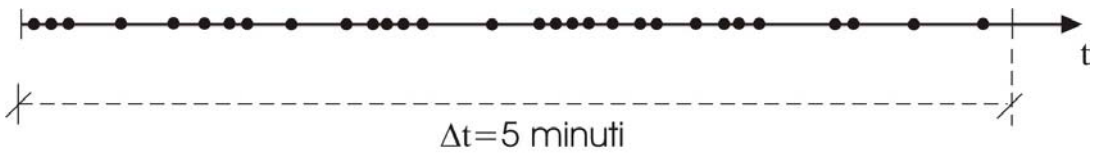


Figura4

La frequenza media di partenza delle chiamate (frequenza media di arrivo) è quindi di 6 chiamate al minuto o di 0,1 chiamate al secondo.

$$\bar{\lambda} = \frac{30}{300} = 0.1 \ \textit{calls/second} \qquad \text{oppure} \qquad \bar{\lambda} = \frac{30}{5} = 6 \ \ \textit{calls/min.} \qquad 1.1$$

Per la validità del "processo poissoniano" è indispensabile che la frequenza media di arrivo λ sia considerata costante nel tempo. Constatata la frequenza media di accadimento degli eventi (nel nostro caso di partenze di chiamate telefoniche), Poisson studiò come in ogni intervallo (di un minuto, nell'esempio) si potesse distribuire la probabilità di accadimento di K eventi. Chiaramente la probabilità più elevata la si ottiene se in un minuto accadessero 6 eventi (K=6) oppure 0,1 eventi in un secondo. Al contrario la probabilità diminuirebbe man mano che ci si allontana dalla frequenza media λ. La formula per il calcolo della probabilità (formula di Poisson) scoperta dal matematico francese fu, infatti:

$$P_K(t) = \frac{(\lambda \cdot t)^K}{K!} \cdot e^{-(\lambda \cdot t)} \qquad 1.2$$

in cui λ è la frequenza media di accadimento degli eventi, K è il numero degli eventi nell'intervallo di tempo *t* (nel nostro esempio di 1 minuto) e P è la probabilità che K eventi possano verificarsi in *t*.

Nel caso preso in esame noteremo che con K=6 avremo, naturalmente, la probabilità più alta:

$$P_6(60\,\text{sec.}) = \frac{(0.1\cdot 60)^6}{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}\cdot e^{-(6\cdot 60)} = 0{,}16 \qquad \text{o, in percentuale:} \quad P_6(60\,\text{sec.}) = 0{,}16\cdot 100 = 16\%$$

Se ripetessimo il calcolo per K=4 e per K=10 otterremo rispettivamente 0,133 (13,3%) e 0,041 (4,1%). Pertanto la distribuzione binominale delle probabilità per K eventi (K=1,2,3….. *n*), e con λ pari a 0.1, è approssimativamente quella visibile in **figura 5**.

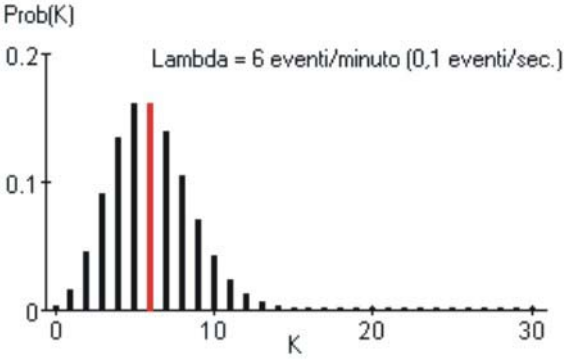


Figura5

Tuttavia la formula di Poisson non consente una intuibilità mentale della distribuzione probabilistica che descrive. Differentemente una comprensione più intuitiva la si può meglio ottenere se si analizza un particolare caso di probabilità: la probabilità che non si verifichi alcun evento nell'intervallo di osservazione, ovvero che K sia uguale a zero. Se poniamo K=0, la formula di Poisson diviene:

$$P_K(t) = e^{-(\lambda\cdot t)} \qquad 1.3$$

In tal caso la formula è divenuta un'equazione esponenziale negativa e conoscendo la costante λ è possibile con facilità tracciarene l'andamento grafico:

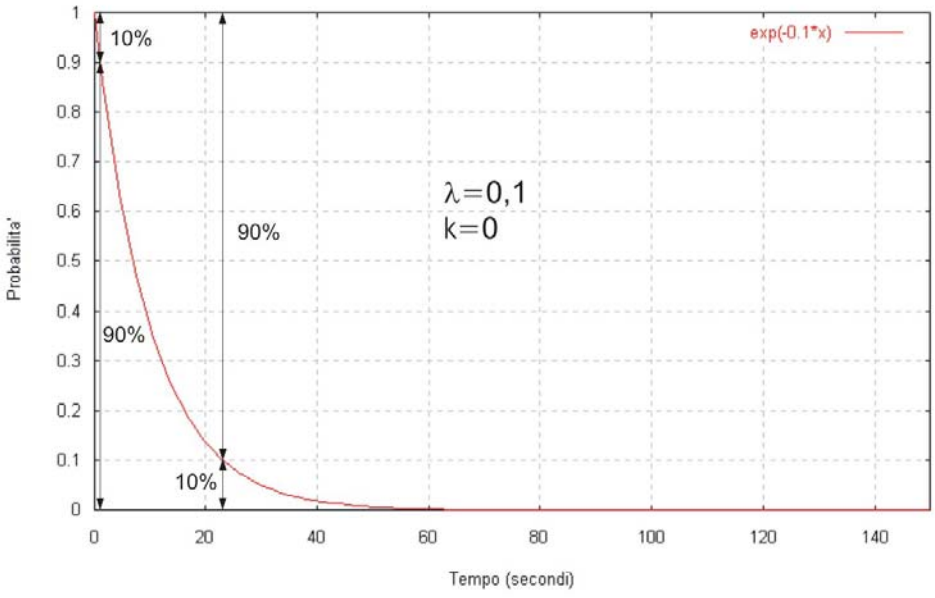


Figura6

Questo grafico esponenziale negativo offre, dato λ=0.1, la distribuzione della probabilità che non si verifichino eventi (K=0) in funzione del tempo di osservazione. E' ora possibile ragionare intuitivamente su questo grafico per conoscere le due più importanti proprietà del "processo di Poisson": la distribuzione esponenziale negativa delle probabilità dei tempi di interarrivo (il tempo di interarrivo è il tempo trascorso tra un evento e il successivo) e l'assenza di memoria.
Se si osserva il grafico esponenziale negativo ci si accorge che con un intervallo di osservazione di un secondo c'è un 90% di probabilità che non accada alcun evento, mentre in un intervallo di osservazione di circa 23 secondi la percentuale di non assistere ad alcun evento scende vertiginosamente al 10%. Intuitivamente è come dire che in un intervallo di un secondo c'è solo un 10% di probabilità che due eventi possano verificarsi con un tempo di interarrivo minore di un secondo e che allargando lo spazio temporale di osservazione a circa 23 secondi c'è un 90% di probabilità che due eventi si verifichino distanziati con un tempo di interarrivo minore di 23 secondi. Ciò significa che i tempi di interarrivo sono distribuiti in funzione dell'intervallo di osservazione con una probabilità che ha un andamento esponenziale comunque dipeso da λ (vedi 1.3). La seconda proprietà del *traffico poissoniano* è l'assenza di memoria. Proprio per la peculiarità dell'andamento esponenziale, aumentando il tempo di osservazione, si nota dal grafico che la curva tende a zero con andamento rettilineo cosicché la probabilità di eventi a 60 secondi è pressoché la stessa anche a 80, a 140 e proseguendo oltre (in teoria all'infinito). Infatti, supponiamo che il tempo di osservazione è di 150 secondi e che al sessantesimo secondo non si sia ancora verificato alcun evento. Chiunque sarebbe portato a supporre che dal sessantesimo secondo in poi la probabilità di avere almeno un evento aumenterebbe. Al contrario la curva esponenziale conferma, invece, che nei tempi successivi al sessantesimo secondo la probabilità non varia per cui nel processo poissoniano tutto quello che è accaduto in precedenza (che non si sia o si sia verificato un evento) non condiziona il futuro, ovvero che in qualsiasi istante di osservazione la probabilità di evento non è condizionato da una memoria storica.
E' con questo modello matematico che si è potuto studiare analiticamente un fenomeno probabilistico come lo è nella realtà dei casi il traffico telefonico offerto. Durante il progetto e l'analisi di una rete telefonica, il traffico offerto verrà sempre paragonato ad un *traffico poissoniano*.

- ***Concetto di congestione (Blocking)***

Nella **figura 4** su una nostra singola risorsa di rete nascono (iniziano) e muoiono (terminano) tre chiamate telefoniche ma distribuite nel tempo in modo tale da non accavallarsi.

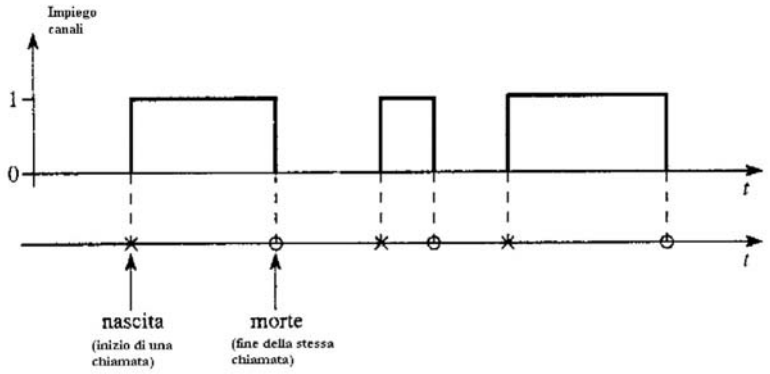


Figura7

In **figura 5** si immagina la distribuzione del traffico offerto da quattro sorgenti "poissoniane" e che impattono contemporaneamente su quattro risorse della rete telefonica.

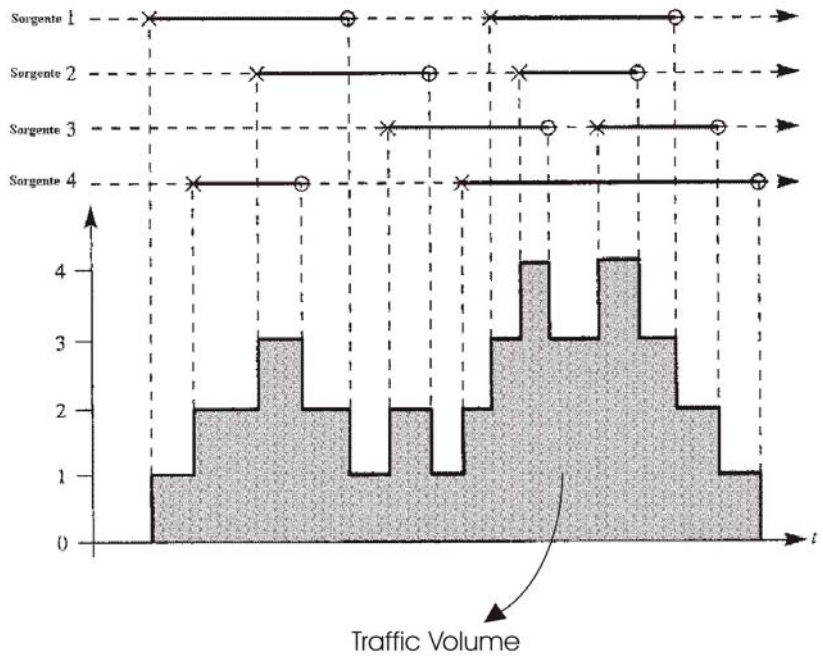


Figura8

Come intuibile, la criticità nel dimensionare le risorse di rete deriva dall'aleatorietà di due variabili:

- l'imprevedibilità dell'istante d'arrivo di una chiamata telefonica
- il tempo per cui ciascuna chiamata impegnerà la risorsa (*tempo di servizio*)

Come già analizzato in precedenza l'arbitrarietà degli arrivi delle chiamate telefoniche viene esemplificata nel modello matematico-probabilistico del processo poissoniano. Per quanto concerne il *tempo di servizio*, anch'esso essendo arbitrario può essere considerato come un processo poissoniano a se stante con distribuzione delle probabilità di morte di tipo esponenziale negativa.
In conclusione nello studio di una rete, sia l'arbitrarietà dei tempi di interarrivo tra le chiamate, sia la durata dei *tempi di servizio* delle chiamate vengono considerati processi poissoniani.
In **figura 6** pur disponendo di tre canali telefonici, gli arrivi delle cinque sorgenti si distribuiscono in modo da non oberare le risorse di rete: nel caso, tutto il *traffico offerto* può essere considerato *traffico smaltito* dalla rete.

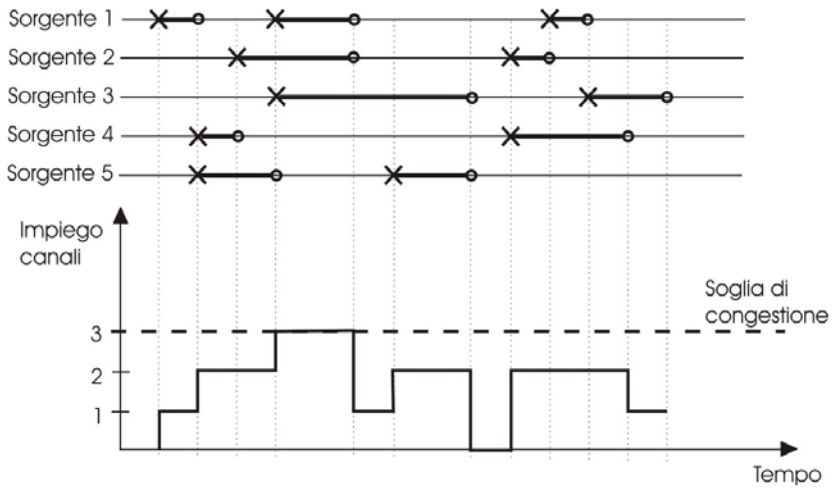


Figura9

In **figura 7** i *poisson arrivals* delle cinque sorgenti producono un sovraccarico delle risorse della rete, cosicché la prima chiamata sulla sorgente 4 e la terza chiamata sulla sorgente 1 trovano una condizione di congestione (*Blocking*). In questo caso il traffico che sarebbe stato generato da queste due chiamate è chiamato traffico bloccato.

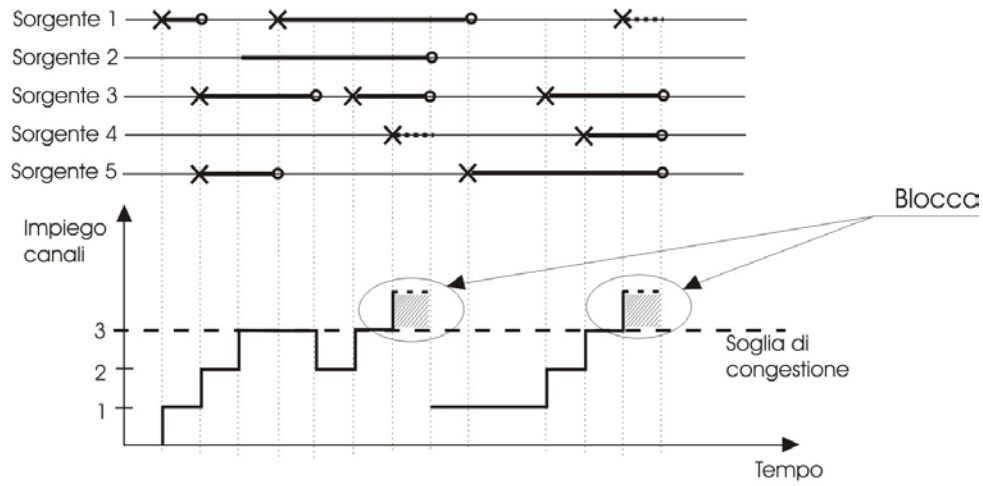


Figura10

Il *Blocking* è anche tecnicamente denominato ***UTRNG***, acronimo di *User Traffic channels Request Not Garanted* (canali di traffico richiesti dall'utente e non garantiti).

- ***Traffico di punta (Busy Hour)***

Come già evidenziato una caratteristica del traffico telefonico è la aleatorietà che lo rende imprevedibile a causa della natura casuale degli arrivi (*poisson arrivals*) e dei tempi di impegno dei canali. In **figura 8** il grafico mostra l'andamento del traffico in Erlangs nell'arco delle ventiquattro ore di una giornata di una ipotetica cella BTS siglata CE51622.

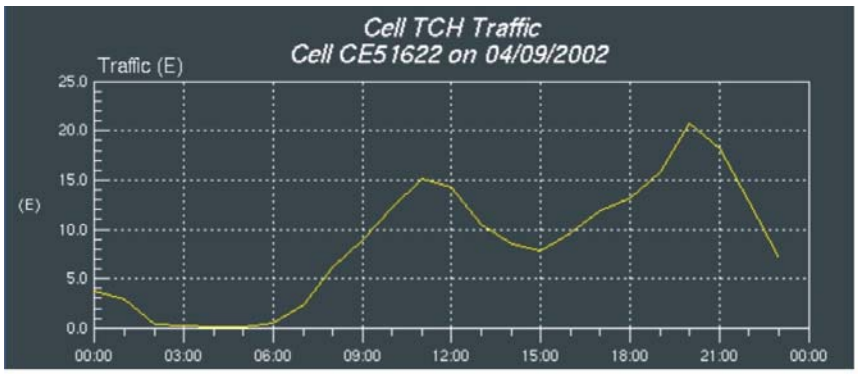


Figura11

Ciò che si nota è che il traffico è tutt'altro che stazionario: presenta un picco in corrispondenza delle ore 20, mentre dalle 2 alle 6 di mattina è pressoché nullo. Il picco delle ore 20, in corrispondenza del quale si registra il maggior traffico, è detto "ora di punta" o *Busy Hour*.
Poiché il dimensionamento deve garantire una buona disponibilità di servizio, anche in condizioni di forte carico, sembra logico avere come riferimento il *traffico offerto* nell' "ora di punta". Chiaramente la *busy hour* è differente da giorno a giorno: può cambiare sia l'ora (un giorno può localizzarsi alle 20, come in **figura 8**, il giorno seguente alle 22) che il carico di *traffico offerto*. Proprio a causa della variabilità della *busy hour* si rende necessario mediare su più giorni per ottenere un traffico di punta medio. Il CCITT (*Comitato Consultivo Internazionale Telegrafico e Telefonico*), che oggi ha cambiato denominazione in ITU (*International Telecommunication Union*), ha stabilito nella famiglia di raccomandazioni E.500 le procedure standard per mediare il traffico di *busy hour*. Si stabilisce di far partire una campagna di misure per analizzare il *traffico offerto* in un intervallo di 5, 10 o 30 giorni consecutivi (a seconda del tipo di carico e delle risorse da mettere a disposizione) e che, si presuppone, siano quelli con maggior traffico nell'arco dell'anno. Il traffico dovrà essere misurato ogni quarto d'ora (avremo, quindi, 96 misure di traffico giornaliero). Dopo avere eseguito la campagna di misure, con i dati alla mano, si calcola il profilo del "giorno medio". In sostanza viene effettuata la media tra il quarto d'ora corrispondente di tutti i giorni, ovvero si mediano i valori di traffico del primo quarto d'ora di tutti i giorni presi in esame, poi del secondo, poi del terzo e così via fino al novantaseiesimo.

$$M_i = \frac{\sum_{j=1}^{g} m_{qj}}{g}$$

dove Mi è il valore di traffico del quarto d'ora del "giorno medio", g è il numero totale dei giorni presi in esame (5,10,30), m_{qj} è il *traffico smaltito* nel quarto d'ora q del giorno j. Il risultato è quello di ottenere 96 valori Mi $\left(M_1, M_2, M_3 .....M_{96}\right)$ che rappresentano le misure di traffico medio di ogni quarto d'ora del "giorno medio". Tra questi 96 dati si ricerca la *TCBH* (*Time Consistent Busy Hour*) cioè i 4 quarti d'ora consecutivi i cui valori di traffico sommati tra di loro danno il risultato più alto.
Nella famiglia di raccomandazioni E.500 esistono anche altri sistemi di calcolo delle *busy hours* come *ADPH* (*Average Daily Peak Hour*) e la *FDMH* (*Fixed Daily Measurement Hour*).

- ***I modelli analitici di Erlang per la stima del GoS: alcuni esempi di calcolo***

Agner Krarup Erlang fu un pioniere negli studi del traffico telefonico. Nacque nel 1878 in Danimarca e lavorò per vent'anni, fino alla morte sopraggiunta nel 1929, nella **Copenhagen Telephone Company**. Nel 1909 pubblicò il suo primo importante lavoro "*La teoria delle probabilità e le conversazioni telefoniche*" guadagnandosi riconoscimenti prestigiosi in tutto il mondo e l'uso dei suoi modelli matematici presso la **General Post Office** britannica. Dieci anni dopo la sua morte, nel 1940, l' Erlang divenne, in sua memoria, l'unità di misura del traffico telefonico più nota nel mondo delle telecomunicazioni. Nel 1917, basandosi sugli studi di distribuzione binominale probabilistica di Siméon-Denis Poisson, iniziò a studiare il comportamento tipo degli abitanti di un ipotetico villaggio, inerentemente all'uso che essi potessero fare del telefono. Gli abitanti usano il telefono oltre che per chiamarsi tra di loro anche per chiamare altri utenti fuori dal villaggio ed il matematico danese desiderava dimensionare quanti doppini diretti al di fuori potessero essere necessari agli abitanti per scongiurare il più possibile l'insorgere di problemi di congestionamento. Suppose, quindi, che il villaggio avesse un certo numero di doppini telefonici diretti verso il mondo esterno e che noi chiameremo "*N*". Egli, però, non poteva sapere quando essi chiamassero all'esterno ne quanto fossero lunghe queste loro telefonate. Suppose, quindi, che ci fosse una media "*m*" di chiamate che partissero per minuto secondo il modello stocastico del processo poissoniano. Il suo intento era quello di stimare la quantità di chiamanti che avrebbe trovato le linee occupate nel tentativo di chiamare al di fuori del villaggio. Alla fine trovò in una formula la risposta al suo quesito:

$$B = \frac{m^N}{N!}\left[\sum_{x=0}^{N}\frac{m^x}{x!}\right]^{-1}$$

Quest'equazione è oggi a noi nota come *formula di Erlang*. Conoscendo il numero delle linee telefoniche verso il mondo esterno "*N*" ed una stima di quante chiamate partono per minuto da una determinata area, con questa formula, è possibile calcolare statisticamente quanti utenti troveranno le linee occupate. Di conseguenza è possibile utilizzare la formula per dimensionare il numero adeguato di canali telefonici (e quindi di linee, apparati e risosrse) di una determinata area geografica. Sebbene il modello di Erlang può sembrare banale, la complessa matematica delle reti telefoniche odierna è ancora basata su questi studi. La sua formula e le successive ricerche da lui condotte costituiscono un importantissimo contributo alla telefonia ed in particolare alle teorie di accodamento telefonico (*queueing techniques*).

Gli studi di A.K.Erlang hanno portato all'adozione di alcuni modelli analitici per il dimensionamento delle reti (*GoS*, ovvero *Grade of Service*) che sono:

- modello A o "*A Erlang*"
- modello B o *"B Erlang"*
- modello B esteso o *"B Extended Erlang"*
- modello C o *"C Erlang"*

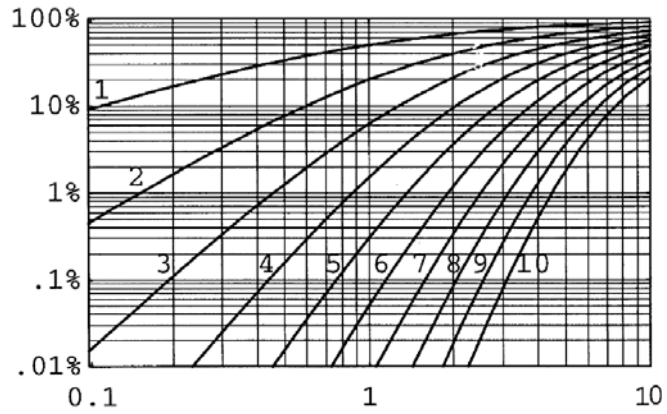
Il modello *"B di Erlang"* è una formula molto utilizzata per il *GoS* ovvero il dimensionamento delle risorse dei sistemi di telecomunicazioni (cioè il numero di linee o di ponti radio e la capacità degli apparati) quando si conosce il massimo valore di traffico in *Erl* nell'ora di punta (*Busy Hour*). Si da per scontato che se il numero di chiamanti dovesse superare la capacità del sistema, dimensionato con il modello *"B di Erlang"*, le chiamate in eccesso incontreranno il segnale di occupato (nel GSM il classico squillo tritonale) e per questi utenti non ci sarà nulla fare se non riprovare più tardi 9tutto il traffico bloccato è traffico perduto). Il modello *"B di Erlang"* altro non è che l'applicazione pura della *formula di Erlang*. Con esso si calcola il numero di canali telefonici da mettere a disposizione conoscendo il volume del traffico stimato in *Busy Hour* e la percentuale di chiamate bloccate che si è disposti ad accettare. Esistono in commercio, o distribuiti gratuitamente, piccoli programmi per computer in grado di calcolare velocemente, con la formula *"B di Erlang"*, il numero dei circuiti richiesti. Nonostante ciò, riproponiamo per intero la formula:

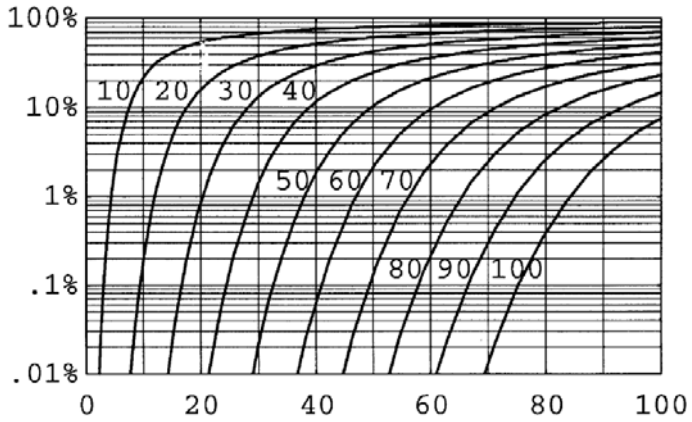
$$B = \frac{\dfrac{m^N}{N!}}{\displaystyle\sum_{x=0}^{N} \frac{m^x}{x!}}$$

in cui *B* rappresenterà la percentuale delle chiamate bloccate, *N* il numero di canali telefonici a disposizione, *m* la stima del volume di *traffico offerto* in *Erl*.
Volendo presentare un esempio pratico ed esemplificativo dell'utilizzo formula della *"B di Erlang"* supponiamo di avere a disposizione 4 canali telefonici e che, da una stima effettuata, il volume di *traffico offerto* è di 2 *Erl*. Con questa formula possiamo calcolare la percentuale del traffico che con molta probabilità incontrerà un blocco.

$$B = \frac{\dfrac{2^4}{4 \cdot 3 \cdot 2 \cdot 1}}{1 + \dfrac{2}{1} + \dfrac{2^2}{2 \cdot 1} + \dfrac{2^3}{3 \cdot 2 \cdot 1} + \dfrac{2^4}{4 \cdot 3 \cdot 2 \cdot 1}} = 0{,}095 \quad \text{che in percentuale equivale a} \quad B\% = 0{,}095 \cdot 100 = 9{,}5$$

Il 9,5% dei chiamati avrà la possibilità di trovare tutti i 4 canali telefonici occupati.
Tuttavia oggi nessuno si sognerebbe di calcolare manualmente la *"B di Erlang"* vista la disponibilità di calcolatori, fogli elettronici, tabelle e grafici come questi presenti di seguito.





L’asse orizzontale è quello del traffico offerto in *Erl*, quello verticale della percentuale di *Blocking*, mentre la famiglia di curve porta accanto il numero di canali *N*. Anche sul grafico è possibile constatare l’esattezza del calcolo effettuato: con 2 *Erl* e 4 canali la percentuale di *Blocking* è quasi pari al 10%. Come visibile dal grafico, aumentando i canali da 4 a 5 la percentuale di *Blocking* diminuirebbe al 4%.
Nell’uso della formula *B erlang* si è dato per scontato che tutto il traffico bloccato è anche traffico perso; ciò significa che tutti quegli utenti i quali, nel tentativo di impegnare risorse per una chiamata telefonica, hanno incontrato il blocco (segnale di occupato) non vengono ripresentati al sistema in modo automatico ma il loro tentativo è da considerarsi fallito. A questi utenti non rimane altro che riprovare in un secondo momento se lo volessero. Tuttavia oggi la tecnologia delle centrali permette di mettere in attesa gli utenti che per motivi di congestione incontrano un blocco. E’ il caso dei customer-care (o call-centre) composti da un certo numero di agents (operatori) preposti a ricevere un vasto numero di chiamate nell’unità di tempo. Quando in una centrale telefonica per customer-care il volume di traffico entrante supera la soglia di disponibilità degli agents, la centrale stessa provvede a mettere automaticamente in attesa le richieste di traffico in eccedenza in modo da ripresentarle al sistema in un secondo momento: in tal caso il traffico bloccato a causa della congestione non coincide col traffico perso. Generalmente, gli utenti in blocco vengono messi in attesa secondo uno schema di priorità basato su un modello di accodamento di tipo FIFO (First-In-First-Out).
Nel caso si dovesse dimensionare un sistema di telecomunicazioni simile, cioè in grado di ripresentare automaticamente, dopo un tempo di attesa e con priorità FIFO, gli utenti bloccati, è chiaro che la semplice formula *B Erlang* non servirebbe. Subentrano, in un caso simile, la formula *B extended Erlang* o la *C Erlang*.
La *B extended Erlang*, nel suo calcolo, tiene conto che una certa percentuale di traffico bloccato (chiamato *recall factor*) deve essere ripresentata al sistema, mentre la *C Erlang* prevede che tutto il traffico bloccato viene interalmente ripresentato. Nel dimensionare un sistema telefonico per call-centre, verrà sicuramente utilizzata la *C Erlang* e proprio per il suo uso frequente vale la pena soffermarsi su una sua breve descrizione.
In sostanza, come con la *B Erlang* è possibile calcolare la percentuale probabilistica di chiamate bloccate, similmente con la *C Erlang* si calcola la percentuale probabilistica di chiamate che non avendo una risposta immediata dagli agents del call-centre verranno messe in attesa, magari con l’intrattenimento di un motivo musicale pre-registrato.

L'uso della *C Erlang* prevede che si conoscano preliminarmente le stime di 4 variabili:

- la frequenza media di arrivo delle chiamate;
- la durata media di ogni chiamata;
- il numero di agents pronti a rispondere alle chiamate;
- il livello di servizio che si desidera raggiungere, ovvero entro quanto tempo vorremmo che la chiamata in attesa abbia una risposta da un agent

Di seguito si propone un semplice esempio di calcolo per rendere esaustivo l'uso della formula. Chiaramente per dimensionare un call-centre o per un uso professionale è opportuno utilizzare direttamente uno dei tanti calcolatori disponibili in commercio poiché la presenza di calcoli fattoriali e numeri elevati di agents renderebbero il calcolo manuale enormemente laborioso. Per tale motivo nella nostra esemplificazione faremo riferimento ad un call-centre di soli 4 operatori, con una media di 20 chiamate in arrivo ogni ora e con un tempo medio di durata di circa 5 minuti per ciascuna chiamata. Inoltre, il target desiderato è che ogni chiamata non rimanga in attesa per più di 15 secondi prima di avere la risposta da un agent.
Il primo passo prevede il calcolo del volume di *traffico offerto* in *Erlangs*:

- minuti di traffico in un'ora: 20 x 5 = 100 minuti
- ore di traffico continuato: 100 / 60 = 1,67 ore
- ✓ volume di traffico (*m*): 1,67 *Erlang*

A tal punto si calcola la probabilità del traffico che viene messo in attesa con la formula *C Erlang*:

$$C = \frac{\frac{m^N}{N!} \cdot \frac{N}{N-m}}{\left(\sum_{X=0}^{N-1} \frac{m^X}{X!}\right) + \left(\frac{m^N}{N!} \cdot \frac{N}{N-m}\right)}$$

in cui *m* è il volume di traffico, *N* il numero degli operatori disponibili.
Con i dati del nostro esempio avremo:

$$C = \frac{\frac{1{,}67^4}{4 \cdot 3 \cdot 2 \cdot 1} \cdot \frac{4}{4-1{,}67}}{\left(1 + 1{,}67 + \frac{1{,}67^2}{2 \cdot 1} + \frac{1{,}67^3}{3 \cdot 2 \cdot 1}\right) + \left(\frac{1{,}67^4}{4 \cdot 3 \cdot 2 \cdot 1} \cdot \frac{4}{4-1{,}67}\right)} = 0{,}10$$

Ciò equivale a dire che il 10% del traffico entrante andrebbe in attesa.
Proseguendo è ora possibile calcolare il tempo medio di attesa o *ASA* (*Average Speed of Answer*), cioè il tempo che mediamente l'utente in attesa attenderà prima di ottenere la risposta da un operatore:

$$T_{ASA} = \frac{C \cdot T_S}{N \cdot \left(1 - \frac{m}{N}\right)}$$

in cui *Ts* è il tempo stimato della durata di ogni chiamata (nel nostro esempio di 5 minuti, ovvero 300 secondi).

Quindi:

$$T_{ASA} = \frac{0{,}10 \cdot 300}{4 \cdot \left(1 - \frac{1{,}67}{4}\right)} \cong 13\, secondi$$

L'ultimo passo è quello di calcolare il livello di servizio garantito. Con quest'ultimo calcolo appureremo qual è la percentuale di utenza che con molta probabilità otterrà la risposta da un operatore prima del tempo target di 15 secondi che ci si era imposti:

$$SVLV = 1 - \left(C \cdot e^{-(N-m)\cdot\frac{t}{T_S}} \right) = 1 - \left(0{,}10 \cdot e^{-(4-1{,}67)\cdot\frac{15}{300}} \right) \cong 0{,}9$$ o in percentuale il 90%.

Il novanta percento del traffico messo in attesa potrebbe ottenere una risposta prima del tempo target di 15 secondi.

- ***I vari tipi di attempts in una rete radiomobile ed il concetto di dropped-call***

Nelle normali reti telefoniche con apparecchi fissi l'unico tipo di tentativo di accesso alla rete (*attempt*) da parte dell'utente è quello vincolante dall'assegnazione fisica del proprio doppino telefonico. Per questo motivo tutto il *traffico offerto* alla rete dal bacino di clientela è traffico di tipo poissoniano (vedi **figura 3**). Differentemente, nella nuova telefonia mobile è possibile accedere col proprio apparecchio telefonico (*mobile station*) ovunque ci si trovi purché esiste la copertura radio che garantisce l'accesso ai servizi. Proprio per la peculiare capacità di gestione della mobilità, le reti radiomobili cellulari danno la possibilità agli utenti che sono impegnati in conversazione o nell'uso di qualsiasi servizio offerto di potersi muovere in piena libertà all'interno degli spazi geografici di copertura radio.
Supponiamo che l'utente in chiamata, movendosi, si allontana da una stazione radio base (BTS) che gli garantisce la copertura radio fino al punto che non è più in grado di servirlo adeguatamente. La rete radiomobile, che effettua costantemente misurazioni della qualità del collegamento con il *mobile station*, provvede automaticamente ad assegnarli una nuova BTS, che per la sua collocazione geografica è più idonea a servirlo, facendolo "saltare" dalla vecchia BTS alla nuova. In sostanza la rete non fa altro che liberare la vecchia risorsa fisica che l'utente occupava sulla precedente BTS ed assegnargli una nuova risorsa sulla nuova BTS. Questo "salto", che avviene con continuità in modo tale che l'utente non percepisca il distacco dalla vecchia risorsa e l'assegnazione della nuova, è chiamato *handover*.
Per poter garantire con l'*handover* la mobilità dell'utente è intuibile che si rende necessario collocare una serie di BTS le une vicine alle altre ed adiacenti tra loro (**figura 12**). L'*handover* è perciò chiamato *handover* adiacente o sinteticamente *handover adj*.

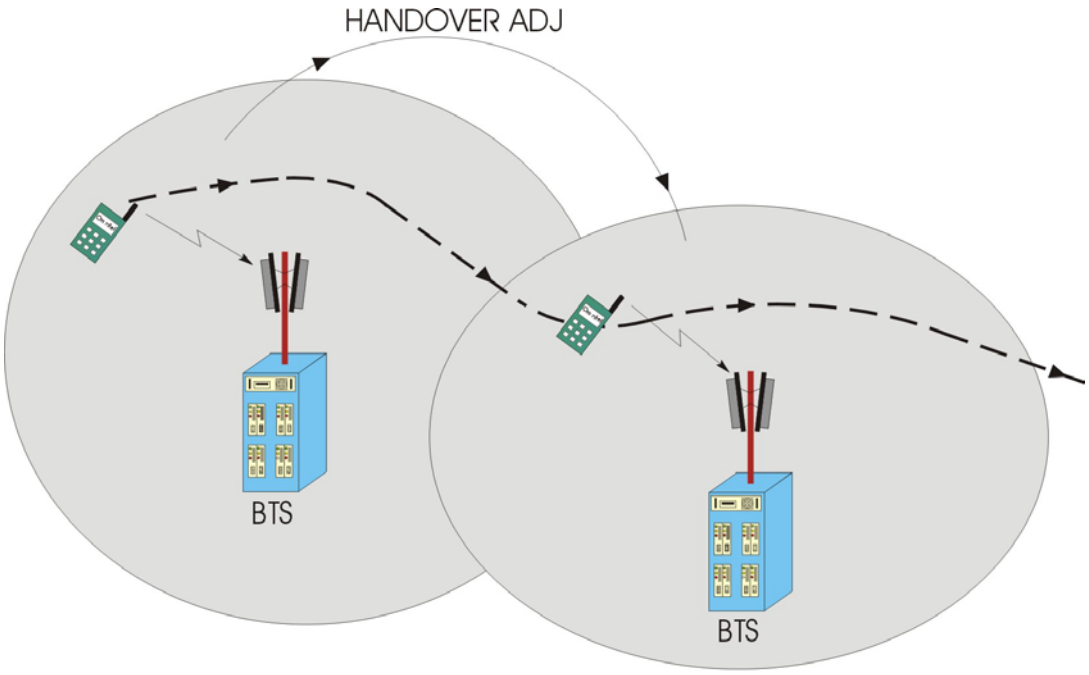


Figura12

Un *handover* può avvenire anche quando l'utente non è in mobilità ma se per mutevoli condizioni della qualità di servizio di radio-copertura la rete "decide", sulla base delle sue misurazioni, che è più opportuno che egli "salti" su di un'altra BTS.

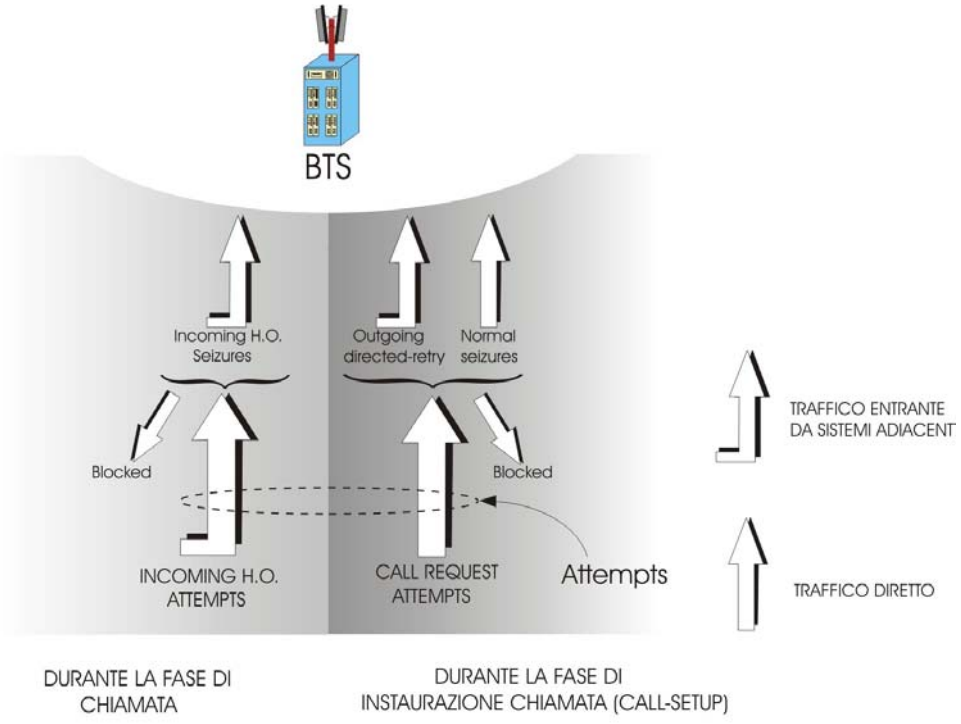


Figura13

E' inoltre possibile che un utente, sotto la copertura di una BTS, nel richiedere una per instaurare una chiamata (*call request*), venga smistato ad una BTS adiacente alla prima, poiché la prima non ha risorse da

cedergli. In tal caso, pur non avendo risorse sulle altre BTS, viene scongiurato il blocco nell'eventualità si dovessero trovare risorse disponibili sulle altre BTS adiacenti. Questo particolare handover in fase di instaurazione di chiamata è chiamato *directed-retry*.
Quindi, in una rete radiomobile si verificano ben due tipologie di *attempts*:

- I tentativi di accesso alla rete durante la fase di chiamata (*call-request attempt*)
- I tentativi di accesso alla rete durante la fase di chiamata per questioni di mobilità o di copertura radio (*incoming handover attempt*)

Se non si verifica congestione, i due tipi di *attempt* diventano accessi effettivi e prendono il nome di *seizure* (cattura).
A sua volta tra i *call-request attempts* possiamo discernere l'accesso normale (*normal seizure*, ovvero cattura semplice) che parte da un utente "accampato" sotto la BTS e l'accesso per re-indirizzamento diretto, cioè per *directed-retry*, proveniente da BTS adiacenti che non hanno risorse da assegnare.
Allo stesso modo i tentativi di accesso durante la fase di chiamata per *handover*, se riescono ad ottenere l'assegnazione del canale, diventano *incoming handover seizures* (catture per *handover* entranti). La **figura 13** ha lo scopo di rendere più esaustivo quanto scritto.
Solo i *normal seizures* costituiscono un traffico poissoniano.
E' quindi evidente che, a differenza delle reti telefoniche fisse, il *traffico offerto* alle BTS di una rete radiomobile non è solo e semplicemente un traffico poissoniano. Ciò vuol dire che la progettazione della rete radiomobile, se operata tramite le classiche formule di Erlang, diviene sicuramente un dimensionamento più approssimativo. Tuttavia, modelli matematici successivi a quelli di Erlang, come il teorema di Wilkinson del 1959 o di Friedericks degli anni '80, pur derivati dalle formule di Erlang, cercano di rendere meno approssimativo il dimensionamento progettuale di reti con sorgenti di traffico miste non "poissoniane".
Un ultimo concetto basilare per poter procedere allo studio del *performance management* e dei *KPI* è la *dropped-call* (chiamata caduta). Il numero di *dropped-calls* è sicuramente uno dei parametri più interessanti per "testare" l'efficienza della rete. Per *dropped-call* si intende una chiamata telefonica che cade improvvisamente senza che termini correttamente (cioè senza che avvenga la procedura di segnalazione per il rilascio della risorsa da ambo i lati). *Dropped-calls* possono verificarsi per problemi relativi alla rete di qualsiasi entità e natura: per problemi di qualità radio, di segnalazione, di trasmissione e di commutazione.
Nell'analisi delle chiamate cadute, per discriminare eventuali problemi della rete è necessario prestare molta attenzione, poiché una *dropped* può originarsi anche per problemi estranei alla rete ma dipesi, invece, dal malfunzionamento di una *mobile station* o a causa di comportamenti strani ed inusuali degli utenti (ad esempio le batterie scollegate dal telefonino durante la chiamata o un improvviso esaurimento della carica).
Un analisi delle *dropped-calls* e delle relative cause per discriminarne la natura è un argomento vasto che aprirebbe un capitolo a parte e che rientra prettamente nella sfera della *QoS*.