Sistemi Discreti

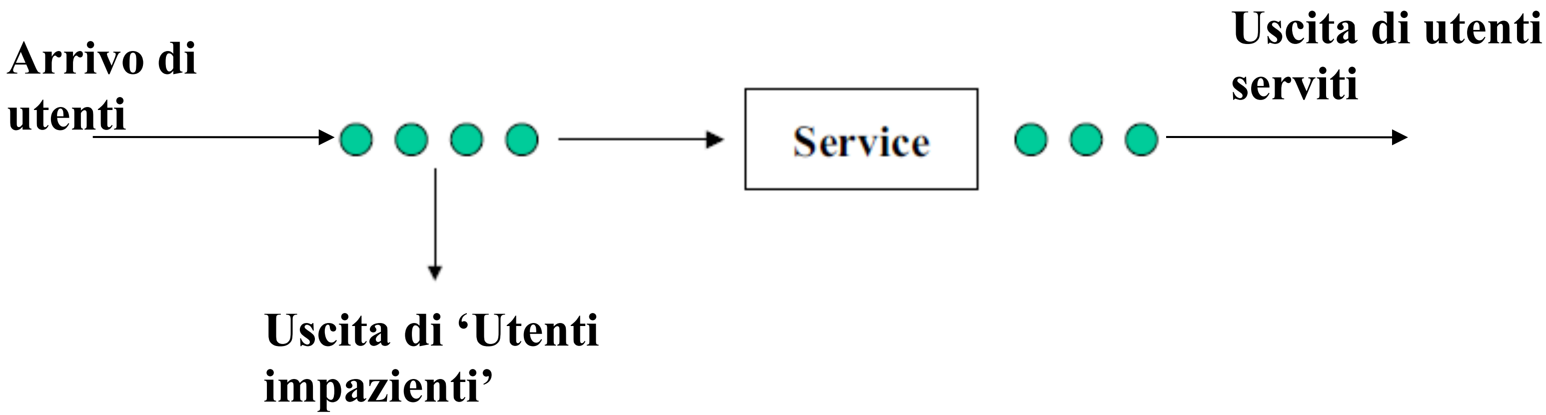
- Reti di Petri Stocastiche
- Automi stocastici
- Code e Reti di Code
- Algebra di processi

Code

- Introduzione
- Classificazione dei sistemi a coda
- Legge di Little
- Sistemi a coda singola
- Reti di Code

Introduzione

Definizione di un sistema a coda



- Un sistema a coda può essere definito nel seguente modo:
  ”utenti arrivano e richiedono un servizio, attendono in coda se il servizio non è disponibile, escono dopo aver ottenuto il servizio "

- Il termine ”utente" può far riferimento a persone, prodotti, macchine, ...

Storia dei sistemi a coda

- I modelli a coda (teoria delle code) furono sviluppati per predire il comportamento di sistemi soggetti a richieste random
- Il primo problema studiato faceva riferimento a traffico telefonico (Erlang, "the theory of probabilities and telephone conversations ", 1909)
- Erlang notò che un sistema telefonico poteva essere modellato con un sorgente di arrivi poissoniana e tempi di servizi esponenziali
- Altri contributi da Pollaczek, Kolmogorov, Khintchine, ….

Interests of queueing systems

Teoria delle code trova applicazioni in:

– Controllo del traffico (communication networks, air traffic, …)

– Sistemi di elaborazione

– Sistemi …

Classificazione dei sistemi a coda

Characteristics of simple queueing systems

Criteri che ci consentono di definire i sistemi a coda:

- Processo di arrivo degli utenti
- Tempi e tipologia dei servizi
- Disciplina di servizio
- Capacità del sistema
- Numero di serventi

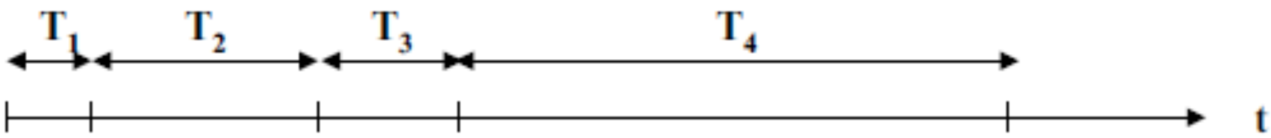
Notazione di Kendall

Notazione per i sistemi a coda:

T/X/C/K/P/Z con:

– T: distribuzione di probabilità dei tempi di interarrivo;
– X: distribuzione di probabilità del tempo di servizio;
– C: numero di serventi;
– K: capacità della coda;
– P: popolazione;
– Z: disciplina di servizio.

Processo di Arrivo
T/X/C/K/P/Z



- T può assumere uno dei seguenti valori:

– M : esponenziale (markoviano)
– G : distribuzione generale
– D : deterministica
– Ek : distribuzione di Erlang (k)
– …

- Nel caso in cui gli utenti arrivino in gruppi, si utilizza la notazione T[X] dove X è una variabile casuale che indica il numero di utenti ad ogni arrivo

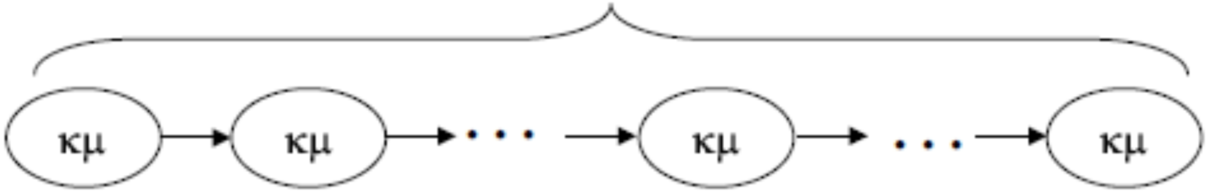
– P{X=k} = P{k utenti arrivano allo stesso istante di tempo}

- Alcuni utenti possono essere esclusi dal sistema se la lunghezza della coda supera uno specifico threshold

Tempi di servizio
T/X/C/K/P/Z

- X può assumere i seguenti valori:
  - M : esponenziale (markoviana)
  - G : Distribuzione generale
  - D : deterministica
  - Ek : distribuzione di Erlang (k)
  - …

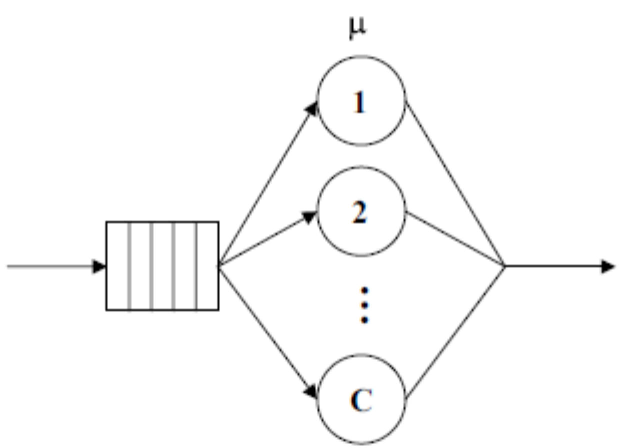
k stage esponenziali con parametro kμ

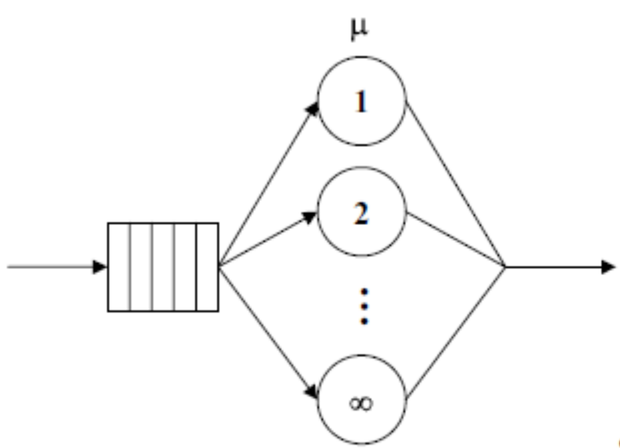


Erlang distribution E_k with parameter μ

Numero di serventi
T/X/C/K/P/Z

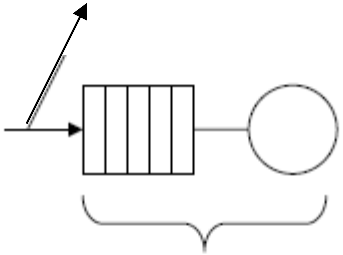
Nei sistemi a coda più semplici i serventi sono identici





Capacità della coda
T/X/C/**K**/P/Z

In caso di coda piena l'utente in arrivo viene perso



Capacity K

Popolazione
T/X/C/K/**P**/Z

La popolazione può essere finita o infinita

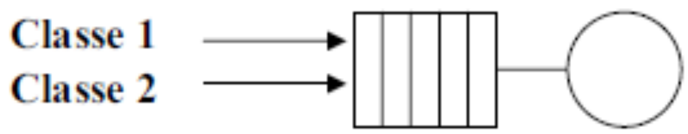
Nel caso di popolazione finita il tasso di arrivo è funzione del numero di utenti nel sistema:
$\boldsymbol{\lambda(n)}$.

Disciplina di servizio
T/X/C/K/P/Z

Z può assumere uno dei seguenti valori:

- FCFS or FIFO : First Come First Served
- LCFS or LIFO : Last Come First Served
- RANDOM : servizio assegnato in modo casuale.
- HL (Hold On Line) : quando un utente ‘importante’ arriva, viene messo in testa alla coda
- PR (Preemption/prelazione) : quando un utente ‘importante’ arriva viene servito immediatamente e l’utente attualmente in servizio viene riassegnato alla coda
- PS (Processor Sharing): tutti gli utenti sono serviti ‘simultaneamente’ con un tasso inversamente proporzionale al numero di utenti.
- GD (General Discipline)

Classi di utenti



Un sistema a coda può offrire servizio a diverse clasi di utenti caratterizzati da:

- Differenti processi di arrivo
- Differenti tempi di servizio (distribuzioni)
- Costi differenti
- Priorità basate sulla classe

Notazione semplificata

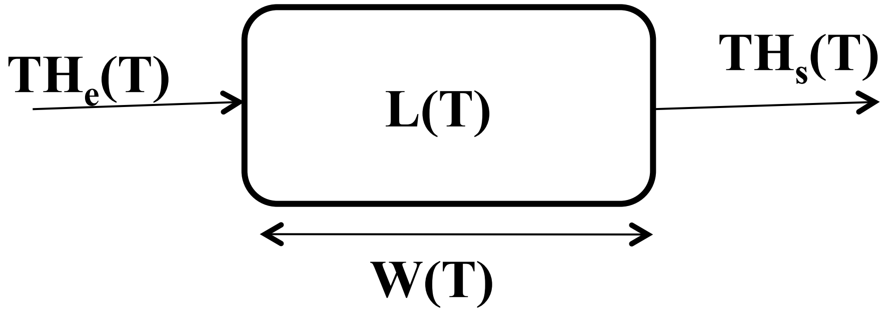
Useremo la notazione semplificataT/X/C nel caso in cui:

- La capacità è infinita
- La popolazione è infinita
- La disciplina di servizio è FIFO

- Quindi T/X/C = T/X/C/∞/∞/FIFO

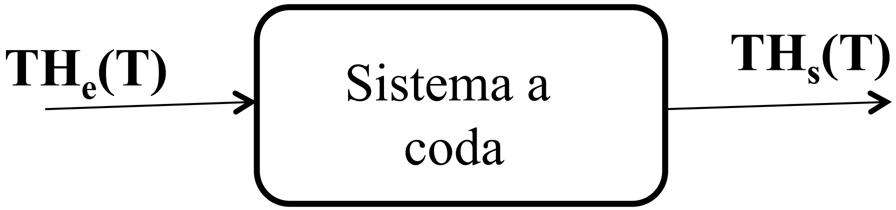
Little's law

Analisi transiente



- A(T) : **numero di utenti arrivati nell'intervallo** [0,T]
- D(T) : **numero di utenti usciti nell'intervallo** [0,T]
- $TH_e(T) = A(T)/T$: **tasso di arrivi nell'intervallo** [0,T]
- $TH_s(T) = D(T)/T$: **tasso di serviti nell'intervallo** [0,T]
- L(T) : **numero medio di utenti nel sistema in** [0,T]
- W_k: **tempo di soggiorno** del k-th utente nel sistema
- $W(T) = \frac{1}{A(T)} \sum_{k=1}^{A(T)} W_k$ **tempo medio di soggiorno** in [0,T]

Stabilità del sistema a coda



Difinizione : Un sistema a coda è detto stabile se il numero di utenti nel sistema rimane finito.

Proprietà di un sistama stabile:

$$\lim_{T \to \infty} TH_e(T) = \lim_{T \to \infty} TH_s(T)$$

$$\lim_{T \to \infty} \frac{D(T)}{A(T)} = 1$$

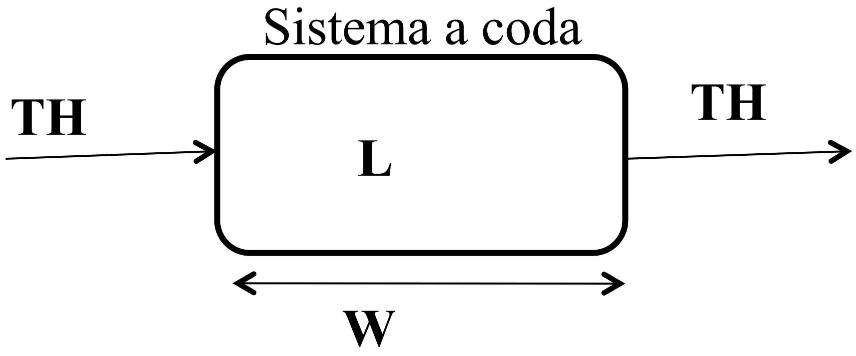
Little's law

Per un sistema a coda stabile,

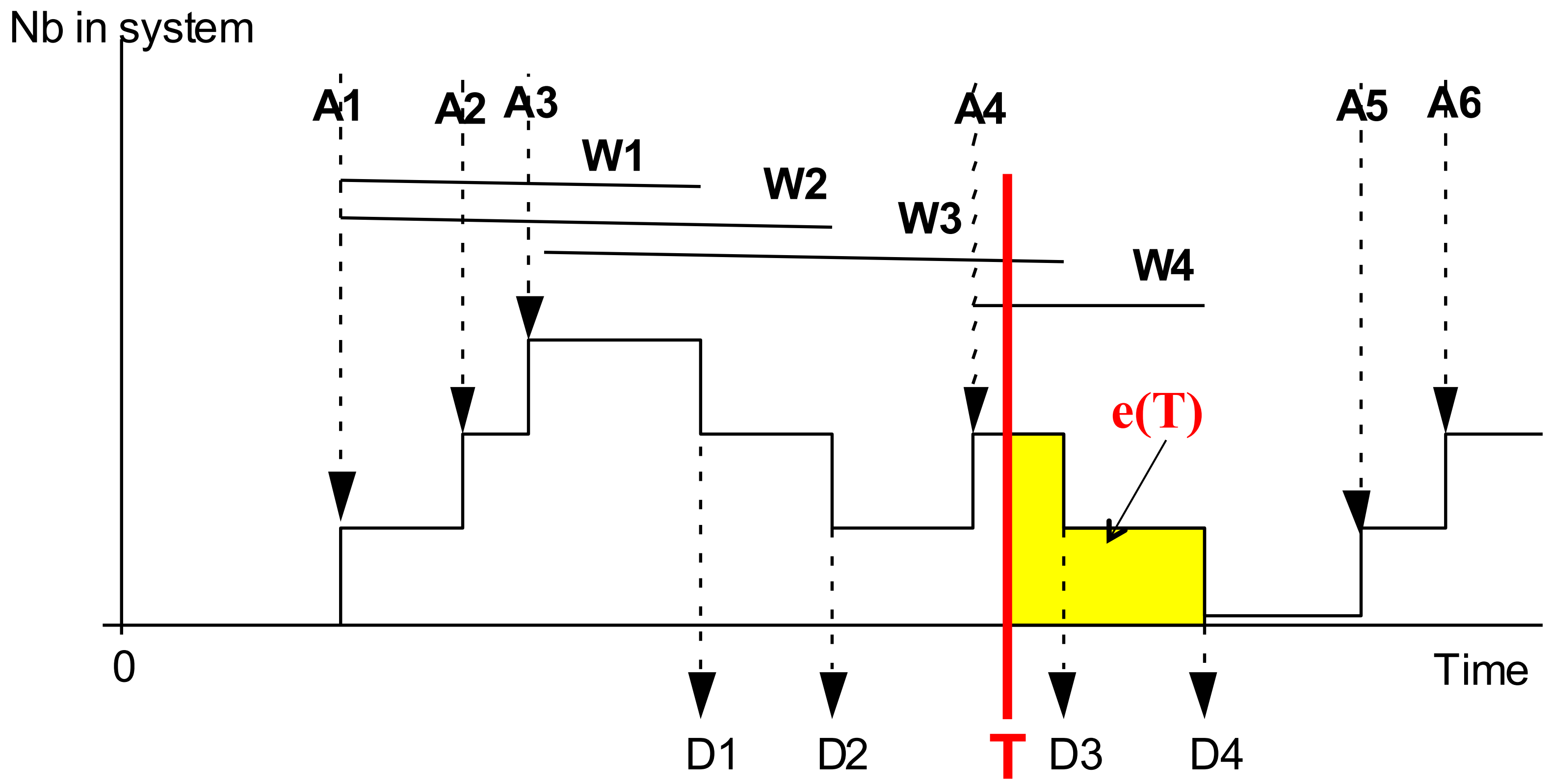
$$\mathbf{L = TH \times W}$$

Dove:

- L : numero medio di utenti nel sistema
- W : tempo medio di risposta del sistema
- TH : throughput (medio) del sistema



dimostrazione



Dimostrazione

$$R(T)TH(T) = \left(\frac{1}{A(T)}\sum_{k=1}^{A(T)} R_k\right)\left(\frac{D(T)}{T}\right) = \left(\frac{D(T)}{A(T)}\right)\left(\frac{1}{T}\sum_{k=1}^{A(T)} R_k\right)$$

$$\frac{1}{T}\sum_{k=1}^{A(T)} R_k = \frac{1}{T}\sum_{k=1}^{A(T)}\int_{t=0}^{T} 1(k \text{ a } t)\,dt + \frac{1}{T}\sum_{k=A(T)-N(T)+1}^{A(T)} r_k(T)$$

$$= Q(T) + \frac{1}{T}e(T)$$

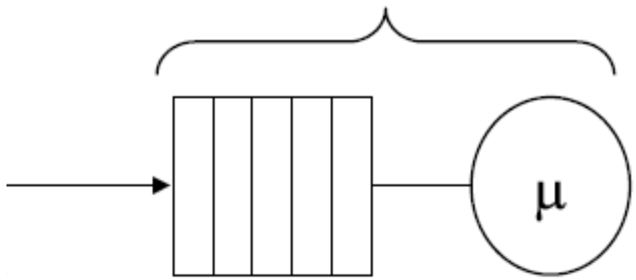
dove N(T) è il numero di utenti al tempo T, e(T) **tempo totale restante nel sistema degli utenti presenti al tempo** T.

per T che va all'infinito, la stabilità implica la dimostarzione.

Sistemi a coda singola

Coda M/M/1

N(t) : numero di utenti nel sistema



λ

Arrivi: Poisson, ovvero tempi di interarrivo esponenziali

Distribuzione dei tempi di servizio esponenziale

Condizioni di stabilità per M/M/1

coda M/M/1 è stabile iff

$$\lambda < \mu$$

Equivale a:

$$\rho < 1$$

Dove:

- $\rho = \lambda/\mu$ fattore di utilizzo o intensità el trafffico

- Notare che il numero di utenti nel sistema non è limitato ein caso di sistema non stabile…..

Misure di prestazioni per code M/M/1

Ls = numero di utenti nel sistema= $\rho/(1-\rho) = \lambda/(\mu-\lambda)$

Ws = tempo di soggiorno nel sistema= $1/(1-\rho)\mu = 1/(\mu-\lambda)$

Lq = lunghezza della coda= $\lambda^2/(\mu-\lambda)\mu = Ls - \rho$

Wq = tempo medio in coda= $\lambda/(\mu-\lambda)\mu = Ws - 1/\mu$

TH = rate di uscita= λ

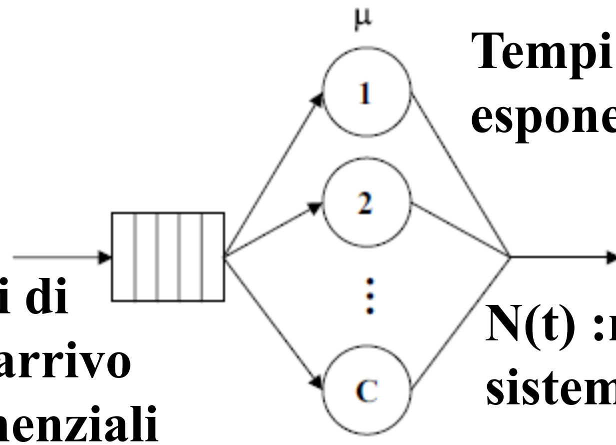
Fattore di utilizzo del servente= ρ

Probalilità che il sistema non contenga utenti = $P_0 = 1 - \rho$

$P\{n > k\}$ = Probabilità di avere più di k utenti nel sistema = ρ^{k+1}

Coda M/M/C

Λ
tempi di interarrivo esponenziali



Tempi di servizio esponenziali

N(t) :numero di uenti nel sistema

Condizione di stabilità: $\lambda < c\mu$.

Misure di prestazioni per coda M/M/C

$\rho = \lambda/\mu$

$\pi_n = \rho^n/n!\ \pi_0,\ \forall\ 0 < n \leq C$

$$\pi_{n+C} = \left(\frac{\rho}{C}\right)^n \pi_C, \forall n \geq 0$$

$$\pi_0 = \left(\sum_{n=0}^{C-1} \frac{\rho^n}{n!} + \frac{\rho^C}{C!\left(1-\rho/C\right)}\right)^{-1}$$

Misure di Prestazioni per code M/M/C

Ls = Numero di utenti nel sistema
= Lq + ρ

Ws = tempo di soggiorno nel sistema
= Wq + 1/μ

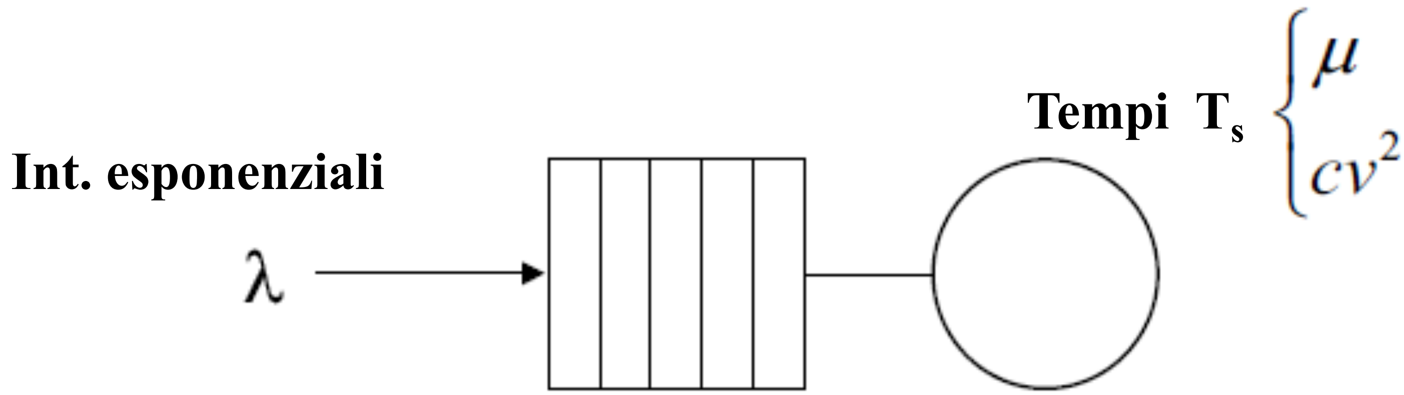
Lq = lunghezza media della coda
=

$$\frac{\rho/C}{\left(1-\rho/C\right)^{2}}\pi_C$$

Wq = tempo medio di attesa
= Lq / λ

π = numero medio di serventi occupati, π = ρ

Coda M/G/1



$$
\begin{cases}
\mu = \dfrac{1}{E[T_s]} \\
cv^2 = \dfrac{VAR[T_s]}{E[T_s]^2}
\end{cases}
$$

Coda M/G/1 : Pollaczek-Khinchin formula

- Pollaczek-Khinchin formula - PK formula-

$$Ls = \frac{\rho}{1-\rho} + \frac{\rho^2}{2(1-\rho)}\left(cv^2 - 1\right)$$

- Utlizzando la formula PK possiamo derivare altre misure di prestazione quali: Ws, Lq, Wq.

- Dalla formula PK formula osserviamo che…..

Coda G/G/1

- **Tempi di interarrivo A_n** tra n and n+1 :

$$E[A_n] = 1/\lambda$$

$$\sigma_A^2 = Var(A_n)$$

- **Tempi di servizio T_n** of dell'utente n :

$$E[T_n] = 1/\mu$$

$$\sigma_T^2 = Var(T_n)$$

- **Tempo di attesa W_n** dell'utente n nella coda (**Lindley equation**)

$$W_{n+1} = \max\{0, W_n + T_n - A_n\}$$

Coda G/G/1

- Bounds sul tempo di attesa

$$\frac{\lambda\sigma_T^2 - \frac{1}{\mu}(2-\rho)}{2(1-\rho)} \le E[W] \le \frac{\lambda(\sigma_A^2 + \sigma_T^2)}{2(1-\rho)}$$

- Waiting time (approssimazione) (**Kingman's equation** or VUT equation)

$$E[W] \approx \frac{(\sigma_A^2 + \sigma_T^2)}{2}\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu}$$

Variabilità **U**tilizzo **T**empo

Queueing networks