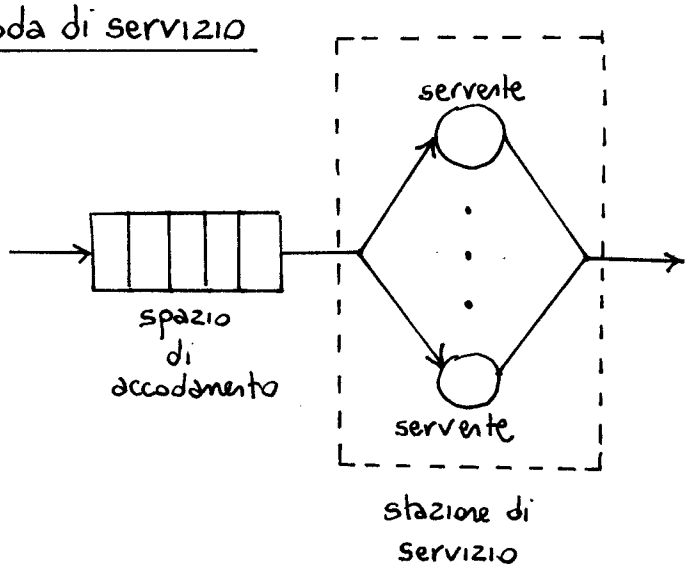
TEORIA DELLE CODE

Lezione del 26 novembre 2008

- coda di servizio

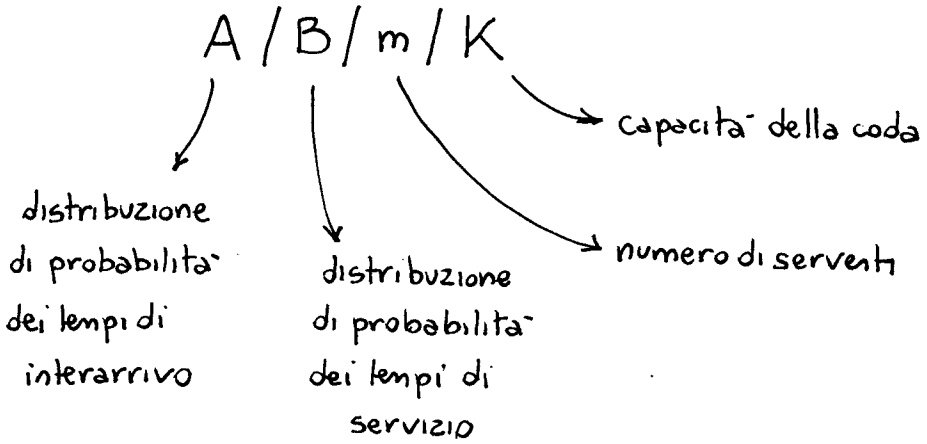


- rete di code: interconnessione di un certo numero di code di servizio

Una coda di servizio è specificata da:

- parametri strutturali
  - capacità della coda
  - numero di serventi
  - ...
- politiche di servizio
  - tipi di clienti serviti
  - disciplina della coda / priorità tra clienti
  - condizioni per accettare / rifiutare clienti
  - ...
- modelli stocastici dei processi di arrivo e servizio
  - distribuzione di probabilità dei tempi di interarrivo
  - distribuzione di probabilità dei tempi di servizio
  - ...

NOTAZIONE DI KENDALL (versione base)



- Convenzioni per A e B:

  G: distribuzione generica

  D: distribuzione deterministica

  U: distribuzione uniforme

  M: distribuzione esponenziale

- I serventi si sottintendono tutti uguali.

- Convenzioni per K:

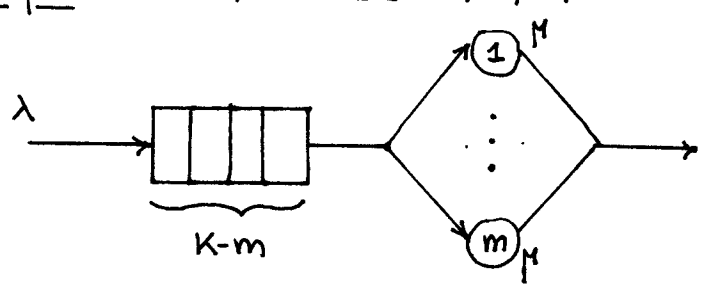
  - K rappresenta il numero massimo di clienti nel sistema, incluso i serventi. Quindi $K \geq m$.

  - K-m è la capacità dello spazio di accodamento

  - Se $K = \infty$, si omette

- Si sottintende una disciplina della coda di tipo FIFO (first-in first-out).

Esempio: coda di servizio M/M/m/K



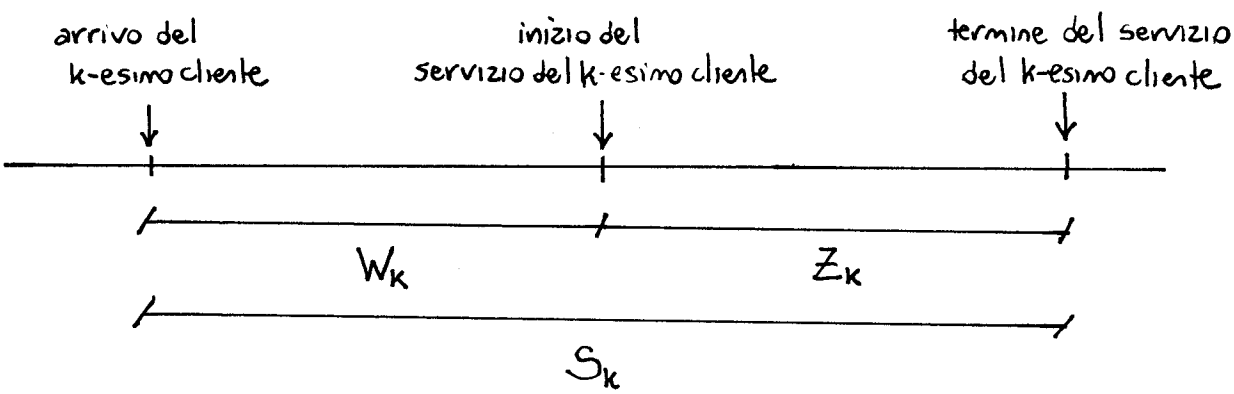
λ: tasso degli arrivi

μ: tasso di servizio

NOTA: Una coda di servizio (così come una rete di code) può essere modellata mediante un automa a stati. Quindi l'analisi di una coda di servizio (o di una rete di code) può essere svolta utilizzando i concetti e gli strumenti studiati nel caso degli automi a stati.

Per esempio, nel caso di tempi di interarrivo e tempi di servizio con distribuzione di probabilità esponenziale, il sistema di servizio è modellabile come un automa a stati stocastico con temporizzazione di Poisson, che sappiamo ammette una rappresentazione come catena di Markov a tempo continuo. Questo ci fornisce gli strumenti per svolgere, per esempio, l'analisi a regime del sistema di servizio.

Sia k il contatore del numero di clienti effettivamente ammessi nella coda di servizio. Consideriamo il k-esimo cliente.



W_k: tempo di attesa del k-esimo cliente (≡ tempo speso nello spazio di accodamento)

Z_k: tempo di servizio del k-esimo cliente (≡ tempo speso nel servente)

S_k: tempo di soggiorno del k-esimo cliente nel sistema

$$\boxed{S_k = W_k + Z_k}$$

S_k è una variabile aleatoria la cui distribuzione, in generale, dipende dall'indice k.

ESEMPIO

coda di servizio G/G/1/1

Supponiamo che la coda sia inizialmente vuota.

Dunque il primo cliente che arriva va direttamente in servizio (il suo tempo di attesa è nullo), da cui segue:

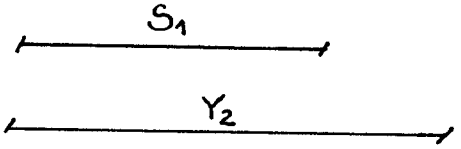
$$S_1 = Z_1$$

Chiamiamo Y_2 il tempo di interarrivo tra il primo e il secondo cliente.

Quando il secondo cliente arriva, possono verificarsi due casi:

caso 1

Il primo cliente ha già terminato il proprio servizio.



NOTA: $S_1 - Y_2 < 0$

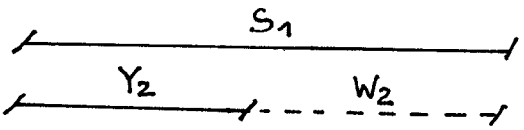
Dunque il secondo cliente trova il servente libero, e va direttamente in servizio.

$$S_2 = Z_2 \qquad (*)$$

caso 2

Il primo cliente non ha ancora terminato il proprio servizio.

=> Il secondo cliente deve mettersi in attesa.



Il tempo di attesa del secondo cliente è $W_2 = S_1 - Y_2$. Dunque:

$$S_2 = W_2 + Z_2 = (S_1 - Y_2) + Z_2 \qquad (**)$$

(*) e (**) possono essere compattate in un'unica espressione:

$$S_2 = \max\{0, S_1 - Y_2\} + Z_2$$

Risulta chiaro che S_1 e S_2 hanno, in generale, distribuzioni diverse.

~ o ~

Dall'esempio precedente appare che calcolare la distribuzione di probabilità di S_k per k generico può essere un lavoro arduo.

Quindi saremmo interessati all'esistenza di una <u>distribuzione stazionaria</u> per S_k al crescere di k, in modo da eliminare la dipendenza da k.

Se esiste una variabile aleatoria S tale che

$$\lim_{k\to\infty} P(S_k \le t) = P(S \le t) \quad \forall t$$

allora la variabile aleatoria S descrive il <u>tempo di soggiorno</u> di un generico cliente nel sistema <u>in condizioni stazionarie</u> (<u>a regime</u>).

Lo stesso ragionamento si può ripetere per il <u>tempo di attesa</u> a regime W:

$$\lim_{k\to\infty} P(W_k \le t) = P(W \le t) \quad \forall t.$$

Sia $t \ge 0$ la variabile temporale (tempo continuo).

$X(t)$: lunghezza della coda all'istante t.

$X(t)$ è una variabile aleatoria la cui distribuzione, in generale, dipende dall'istante temporale t.

Anche in questo caso, se esiste una variabile aleatoria X tale che

$$\lim_{t\to\infty} P(X(t)\leq x) = P(X\leq x) \qquad \forall x$$

allora la variabile aleatoria X descrive la lunghezza della coda in condizioni stazionarie (a regime).

Come per le catene di Markov, usiamo la notazione

$$\Pi_i = P(X=i) \qquad i=0,1,\dots,K$$

Per un singolo servente, si definisce UTILIZZAZIONE U la frazione di tempo (la probabilità) a regime che il servente è occupato.

ESEMPIO - Nel caso di coda di servizio con un unico servente, risulta

$$U = 1-\Pi_0$$

sotto l'ipotesi di esistenza della situazione di regime.

OBIETTIVI NEL PROGETTO DI UNA CODA DI SERVIZIO

Consideriamo la situazione a regime.

- dal punto di vista dell'utente:
  minimizzare il tempo medio di attesa $\longrightarrow$ $\min E[W]$
- dal punto di vista del gestore del servizio:
  massimizzare l'utilizzazione dei serventi $\longrightarrow$ $\max U$

Sono obiettivi contrastanti!

Infatti, per mantenere i serventi pienamente utilizzati, si devono tollerare lunghi tempi di attesa. Viceversa, per ottenere tempi di attesa contenuti, si deve tollerare che i serventi rimangano spesso inutilizzati.

$\Rightarrow$ necessità di un compromesso

Caratterizzazione della situazione a regime

λ_{eff} : frequenza media degli arrivi effettivi nel sistema

μ_{eff} : frequenza media delle partenze dal sistema

A regime, intuitivamente non ci deve essere né accumulo né svuotamento di clienti, in media, nel sistema. Quindi deve valere l'equazione di bilanciamento del flusso:

$$\boxed{\lambda_{eff} = \mu_{eff}}$$

Esempio

coda di servizio G/G/1

Ipotizziamo l'esistenza della situazione a regime.

λ: frequenza media degli arrivi

Tutti gli arrivi sono accettati (capacità infinita)

$\Rightarrow \lambda_{eff} = \lambda$

μ: frequenza media dei servizi in condizioni di pieno utilizzo del servente

$$\Rightarrow \mu_{eff} = \mu \cdot U = \mu(1-\pi_0)$$

μ: numero medio di servizi nell'unità di tempo

U: utilizzazione ≡ frazione di tempo che il servente è occupato

$$\Rightarrow \boxed{\lambda = \mu(1-\pi_0)} \Rightarrow \frac{\lambda}{\mu} = 1-\pi_0 < 1$$

condizione necessaria per l'esistenza della situazione a regime.