Code con priorità

prof.ssa Carla De Francesco

Università degli Studi di Padova
Dipartimento di Matematica “Tullio Levi-Civita”

Ottimizzazione Stocastica

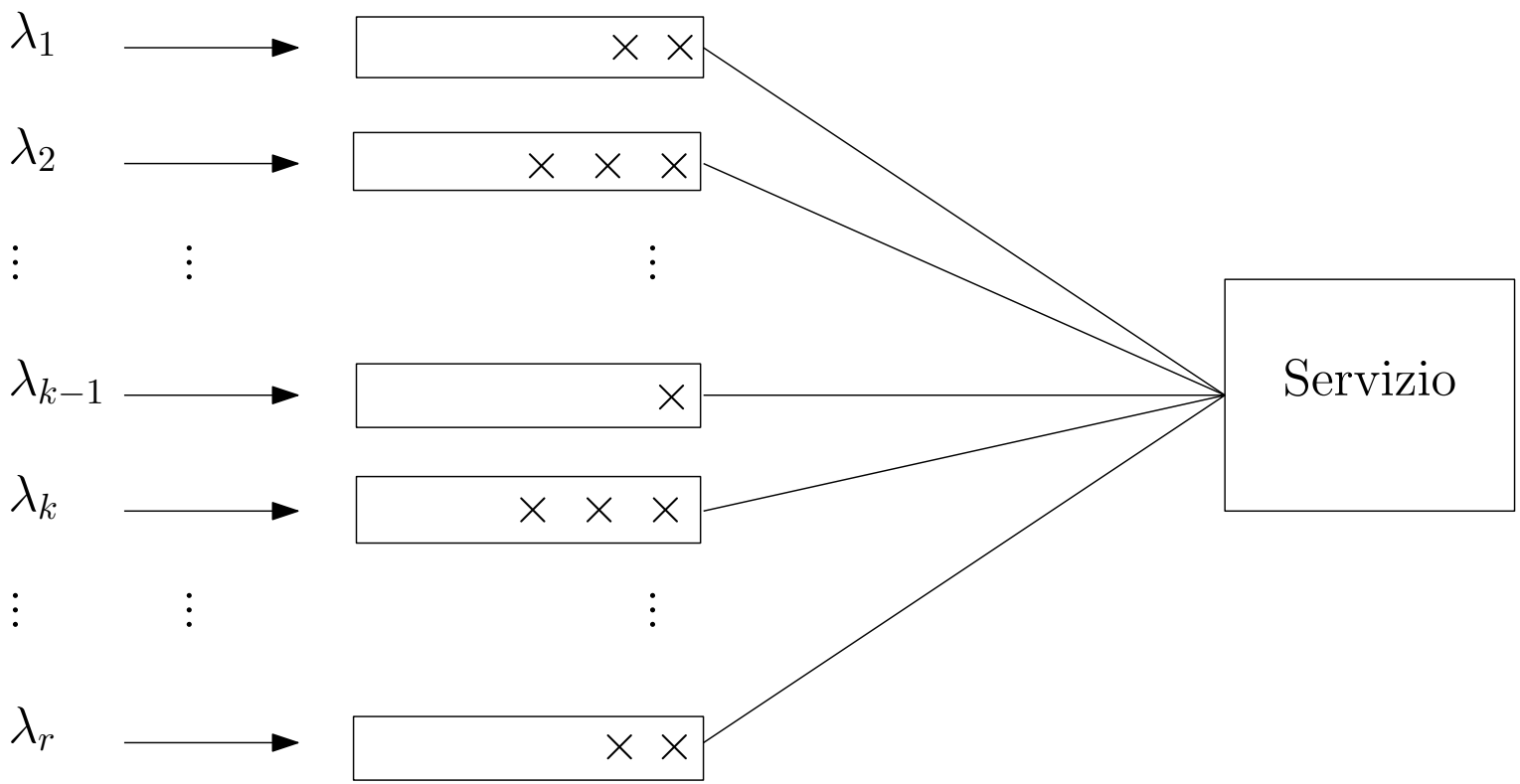
Osservazioni

1. Nei modelli visti finora, che importanza ha avuto la disciplina di coda? Quando è stato necessario assumere FIFO?
2. Nei modelli visti finora, se si considerano discipline della coda che non dipendono dai tempi di servizio (FIFO, LIFO, SIRO):
   - il diagramma degli stati e delle transizioni non dipende dalla particolare disciplina scelta
   - le distribuzioni di N e N_q sono invarianti rispetto alla disciplina
   - L, L_q, $\overline{W}$, $\overline{W}_q$ sono invarianti (per $M/G/1$, il nostro calcolo di $\overline{W}$ utilizza FIFO, ma la formula finale ottenuta è generale)
   - cambiano, invece, le distribuzioni di W e di W_q
   - $Var(W_{FIFO}) < Var(W_{SIRO}) < Var(W_{LIFO})$
3. Consideriamo sistemi di code in cui i clienti sono suddivisi in classi con diversa priorità:
   - priorità *preemptive* / *nonpreemptive*
   - *preemptive-resume* / *preemptive-repeat*

Sistema M/G/1 a priorità *nonpreemptive*

- r classi di clienti: la classe 1 ha priorità più alta, ..., la classe r priorità più bassa
- priorità *nonpreemptive*
- disciplina di tipo FIFO all'interno di ciascuna classe
- processo degli arrivi di Poisson per ogni classe k, tasso λ_k
- tempi di servizio generali:
  v.a. S_k per classe k, con distribuzione $f_{s_k}(s)$,
  media $E[S_k] = 1/\mu_k$, momento di secondo grado $E[S_k^2]$ noti
- capacità infinita della coda per ogni classe
- definiamo $\rho_k = \lambda_k/\mu_k$
- il sistema ha un tasso di utilizzo $\rho = \rho_1 + \rho_2 + \ldots + \rho_r$
- supponiamo per ora $\rho < 1$, quindi tutte le classi raggiungono lo stato stazionario

Sistema di code con *r* classi di priorità



M/G/1 a priorità *nonpreemptive*: tempo medio atteso in coda da un cliente di classe k

In condizioni stazionarie,

$$\overline{W}_{qk} = \overline{W}_0 + \sum_{i=1}^{k} \frac{1}{\mu_i} \cdot L_{qi} + \sum_{i=1}^{k-1} \frac{1}{\mu_i} \cdot M_i \qquad (1)$$

dove:

$\overline{W}_0$ = tempo rimanente per completare il servizio del cliente che sta occupando il servente quando arriva il nuovo cliente di classe k

L_{qi} = numero medio di clienti di classe i che sono già in coda nell'istante in cui arriva il nuovo cliente di classe k

M_i = numero medio di clienti di classe i che arrivano mentre il nuovo cliente di classe k è in coda

... calcoli (1)

Dall'incidenza casuale, supponendo noto che nell'istante di arrivo del nuovo cliente il servente stia servendo un cliente di classe i:

$$(\overline{W}_0|i) = \frac{E[S_i^2]}{2E[S_i]} = \frac{\mu_i E[S_i^2]}{2}$$

In condizioni stazionarie, la probabilità che il servente sia occupato con un cliente di classe i è ρ_i:

$$\overline{W}_0 = \sum_{i=1}^{r} \rho_i \cdot (\overline{W}_0|i) = \sum_{i=1}^{r} \frac{\rho_i \cdot \mu_i E[S_i^2]}{2} = \sum_{i=1}^{r} \frac{\lambda_i \cdot E[S_i^2]}{2} \quad (2)$$

$$L_{qi} = \lambda_i \cdot \overline{W}_{qi} \quad (3)$$

$$M_i = \lambda_i \cdot \overline{W}_{qk} \quad (4)$$

... calcoli (2)

Usando (2), (3) e (4), la (1) diventa:

$$\overline{W}_{qk} = \overline{W}_0 + \sum_{i=1}^{k} \rho_i \overline{W}_{qi} + \overline{W}_{qk} \sum_{i=1}^{k-1} \rho_i = \overline{W}_0 + \sum_{i=1}^{k-1} \rho_i \overline{W}_{qi} + \overline{W}_{qk} \sum_{i=1}^{k} \rho_i$$

e ricavando $\overline{W}_{qk}$:

$$\overline{W}_{qk} = \frac{\overline{W}_0 + \sum_{i=1}^{k-1} \rho_i \overline{W}_{qi}}{1 - \sum_{i=1}^{k} \rho_i} \qquad per \quad k = 1, \ldots r \qquad (5)$$

Risolvendo (5) ricorsivamente per $k = 1$, $k = 2$, ... si ottiene:

$$\overline{W}_{qk} = \frac{\overline{W}_0}{(1 - a_{k-1})(1 - a_k)} \qquad per \quad k = 1, \ldots r$$

dove $a_0 = 0$ e $a_k = \sum_{i=1}^{k} \rho_i$

M/G/1 a priorità *nonpreemptive*: tempo medio atteso in coda da un cliente di classe k

Alla fine, il tempo medio di attesa in coda di un cliente di classe k in condizioni stazionarie è dato dalla

$$\overline{W}_{qk} = \frac{\overline{W}_0}{(1-a_{k-1})(1-a_k)}$$

dove $a_0 = 0, \qquad a_k = \sum_{i=1}^{k} \rho_i$

$$\overline{W}_0 = \frac{1}{2}\sum_{i=1}^{r} \lambda_i \cdot E[S_i^2]$$

- Dipende solo da media e varianza di S_i, non da $f_{s_i}(s)$
- La condizione $\rho < 1$ assicura che tutte le classi di clienti raggiungano lo stato stazionario
- Se invece $\rho_1 + \rho_2 + \ldots + \rho_p < 1 \leq \rho_1 + \rho_2 + \ldots + \rho_{p+1}$ per qualche p, solo le classi $1, \ldots, p$ raggiungono lo stato stazionario (la formula per $\overline{W}_{qk}$ va modificata di conseguenza)

M/G/1 a priorità *nonpreemptive*: formula di Little

La formula di Little è ancora applicabile, ma facendo riferimento ad ogni singola classe.

Numero medio di clienti della classe k in coda =

$$L_{qk} = \lambda_k \overline{W}_{qk}$$

Tempo medio trascorso nel sistema da un cliente di classe k =

$$\overline{W}_k = \overline{W}_{qk} + \frac{1}{\mu_k}$$

M/G/1 a priorità *nonpreemptive*: costo minimo del tempo trascorso nel sistema (1)

c_k = costo per ogni unità di tempo che un cliente di classe k spende nel sistema

Problema: assegnare le priorità in modo da minimizzare il costo atteso C (per unità di tempo) del tempo totale che tutti i clienti spendono nel sistema, calcolato dalla

$$C = \sum_{i=1}^{r} c_i \cdot L_i = \sum_{i=1}^{r} c_i \cdot \rho_i + \sum_{i=1}^{r} c_i \cdot \lambda_i \cdot \overline{W}_{qi}$$

ottenuta sapendo che $L_i = \lambda_i \overline{W}_i = \lambda_i(\overline{W}_{qi} + 1/\mu_i)$

Soluzione:
Per ogni classe calcolare il rapporto $f_k = c_k / E[S_k] = c_k \mu_k$
Dare priorità più alta alle classi con f_k più alto

M/G/1 a priorità *nonpreemptive*: costo minimo del tempo trascorso nel sistema (2)

Teorema (senza dim):
Per minimizzare il costo atteso C bisogna assegnare le priorità secondo i rapporti f_k: più alto il rapporto, più alta la priorità della classe.

Intuitivamente, questo significa servire prima i clienti della classe con c_k più alto e con tempo medio di servizio più basso.

Corollario:
Per minimizzare il tempo totale trascorso nel sistema da tutti i clienti ($c_k = 1$ per ogni k), assegnare le priorità secondo i tempi medi di servizio, per ogni classe di clienti: più basso il tempo medio di servizio, più alta la priorità della classe:
SEPT = disciplina di coda *Shortest Expected Time First*

M/M/1 a priorità *preemptive*

- M/M/1 con r classi di clienti a priorità *preemptive*
- tutti i clienti hanno lo stesso tasso di servizio μ
- *preemptive-resume* = *preemptive-repeat* per la mancanza di memoria della distribuzione esponenziale dei tempi di servizio

Si può dimostrare che il tempo medio trascorso nel sistema da un cliente di classe k è dato dalla formula

$$\overline{W}_k = \frac{1/\mu}{(1-a_{k-1})(1-a_k)} \qquad \textit{per} \quad k = 1, \ldots, r$$

dove $a_k = \sum_{i=1}^{k} \rho_i$ e $a_0 = 0$