

Queueing System with Potential for Recruiting Secondary Servers

Luca Lombardo

Seminario per Metodi Numerici per Catene di Markov

Struttura del seminario

- 1 Introduzione
- 2 Modello Matematico
- 3 Studio del modello di coda in stato stazionario
 - QDB
 - Condizione di ergodicità del processo QBD
 - Calcolo della distribuzione stazionaria del processo QBD
 - Approccio G1/M/1
- 4 Risultati Numerici
- 5 Conclusioni

Queueing Theory

- I modelli di coda sono strumenti matematici utilizzati per rappresentare sistemi di risorse che devono essere utilizzati da diversi utenti, chiamati "clienti".
- La terminologia deriva da applicazioni come gli sportelli bancari o i caselli autostradali, dove i clienti effettivamente si mettono in coda finché non vengono serviti da un dipendente.

Queueing Theory

Code semplici

- Le code semplici consistono di un solo server che attende un solo cliente alla volta, in ordine di arrivo
- Si assume che il tempo sia discretizzato in intervalli di lunghezza fissa
- Un numero casuale di clienti si unisce al sistema durante ogni intervallo
- Il server rimuove un cliente dalla coda alla fine di ogni intervallo, se presente

Queueing Theory

Sia α_n il numero di nuovi arrivi durante l'intervallo $[n-1, n)$ e X_n il numero di clienti nel sistema al tempo n , abbiamo:

$$X_{n+1} = \begin{cases} X_n + \alpha_{n+1} - 1 & \text{se } X_n + \alpha_{n+1} \geq 1 \\ 0 & \text{se } X_n + \alpha_{n+1} = 0 \end{cases}$$

Se α_n è una collezione di variabili casuali indipendenti, allora X_{n+1} è condizionalmente indipendente da X_0, \dots, X_{n-1} se X_n è noto. Se, inoltre, le α_n sono identicamente distribuite, allora X_n è omogenea.

Queueing Theory

Lo spazio degli stati è \mathbb{N} e la matrice di transizione è

$$P = \begin{pmatrix} q_0 + q_1 & q_2 & q_3 & q_4 & \dots \\ q_0 & q_1 & q_2 & q_3 & \ddots \\ \vdots & q_0 & q_1 & q_2 & \ddots \\ 0 & & \ddots & \ddots & \ddots \end{pmatrix}$$

q_i è probabilità $P[\alpha = i]$ che i nuovi clienti che entrino in coda durante un intervallo di un'unità di tempo

α denota ognuna delle possibili distribuzioni di α_n identicamente distribuite.

Queueing Theory

Le catene di Markov aventi matrice di transizione della forma

$$P = \begin{pmatrix} B_1 & B_2 & B_3 & B_4 & \dots \\ A_0 & A_1 & A_2 & A_3 & \ddots \\ & A_0 & A_1 & A_2 & \ddots \\ 0 & & \ddots & \ddots & \ddots \end{pmatrix}$$

dove $A_i, B_{i+1}, i \geq 0$ sono matrici non negative di dimensione $k \times k$, sono dette M/G/1-type Markov Chains

Obiettivi del paper

Nuovo approccio per migliorare i modelli di coda utilizzando server secondari temporanei reclutati tra

- Server secondari disponibili solo temporaneamente e servono gruppi di diversa dimensione
- Dopo aver servito un gruppo, i server secondari lasciano il sistema

Obiettivi del paper

Il paper analizza anche altri approcci di modelli di coda con server secondari presenti in letteratura, ma si concentra su due caratteristiche che sono intrinseche in alcuni sistemi del mondo reale e non sono state studiate in passato:

- Server secondari assegnati ad un gruppo e offrono i servizi uno alla volta
- Con probabilità, un cliente servito da un server secondario diventa insoddisfatto e torna al sistema principale per ottenere un nuovo servizio

Markovian arrival process (MAP)

- Si considera un sistema di coda a singolo server con arrivi secondo un processo di arrivo markoviano (MAP) con matrici di parametro di ordine m .
- Il MAP generalizza processi puntiformi noti come Poisson, Poisson interrotto e rinnovamenti di tipo fase.
- MAP è ideale per situazioni in cui può essere presente una correlazione nei tempi tra gli arrivi.
- L'uso di MAP semplifica l'analisi e ne rende più facile la comprensione grazie alla notazione semplice e all'utilizzo del formalismo delle matrici.

Caratterizzazione del MAP

- Il generatore irriducibile del processo di arrivo markoviano (MAP) è dato dalla somma delle matrici di parametro D_0 e D_1 di ordine m .
- La matrice D_0 governa le transizioni del generatore sottostante che non producono arrivi, mentre la matrice D_1 governa quelle transizioni corrispondenti agli arrivi nel sistema.

L'invariante di probabilità δ soddisfa l'equazione

$$\delta(D_0 + D_1) = \mathbf{0} \quad \delta e = 1$$

Proprietà del MAP

Rate medio di arrivi (λ)

$$\lambda = \delta D_1 e$$

Varianza degli tempi interni di arrivo (σ^2)

$$\sigma^2 = \frac{2}{\lambda} \delta (-D_0)^{-1} e - \frac{1}{\lambda^2}$$

Correlazione (ρ_c) tra due successivi tempi interni di arrivo

$$\rho_c = \frac{\lambda \delta (-D_0)^{-1} D_1 (-D_0)^{-1} e - 1}{2 \lambda \delta (-D_0)^{-1} e - 1}$$

Modello di coda con server principale e secondario

Il sistema ha un singolo server che offre servizi in modo FCFS.

- Il server principale offre servizi esponenziali con parametro μ_1 .
- Con probabilità p , un cliente servito può essere reclutato per diventare un server secondario, che offre servizi ai clienti in attesa del sistema.
- I tempi di servizio del server secondario sono esponenziali con parametro μ_2 .

Achtung!

Un cliente insoddisfatto dal servizio ricevuto dal server secondario potrebbe richiedere di essere servito di nuovo con probabilità v .

Modello di coda con server principale e secondario

Il sistema può avere al massimo due server in qualsiasi momento.

- Il server secondario sarà assegnato a un gruppo di i clienti dove $i = \min\{\text{numero nella coda}, L\}$, ed L è un pre-determinato positivo finito intero.
- I clienti insoddisfatti sono reinseriti nel sistema. Quando il server secondario ha finito di servire tutti i clienti assegnati, viene rilasciato dal sistema.

Edge case

Il caso in cui $\nu = 1$ non è interessante poiché ogni cliente servito da un server secondario viene reinserito nel sistema e l'assunzione di server secondari rallenta solo il sistema nell'offrire servizi.

Struttura del sistema

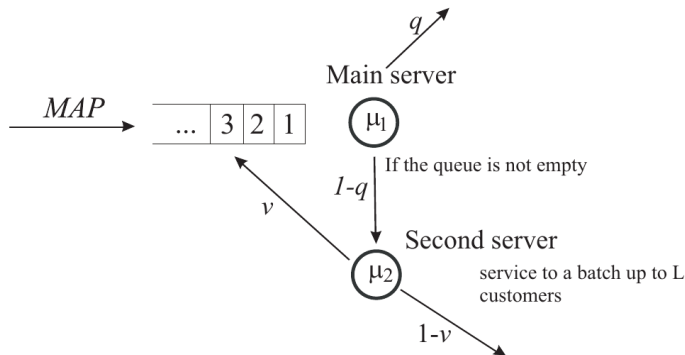


Figure: Immagine da [1]

Due approcci possibili

QDB

Primo processo che analizzeremo in questa sezione: un caso particolare delle catene di markov a tempo continuo (CTMC)

GI/M/1

Secondo processo che analizzeremo nella sezione successiva

Descrizione del processo QBD che governa il sistema e il suo generatore

Al tempo $t \geq 0$, indichiamo:

- $i_t \geq 0$ il numero di clienti nel sistema
- $n_t \in \{0, \dots, \min(i_t, L)\}$ il numero di clienti in servizio al server secondario
- $\xi_t = 1, \dots, m$ lo stato del processo sottostante del MAP che descrive gli arrivi dei clienti

Allora, il processo stocastico $\{\zeta_t = (i_t, n_t, \xi_t), t \geq 0\}$ che descrive il comportamento del modello in esame è un CTMC regolare e irriducibile.

Generatore infinitesimale del processo QBD

Enumerando gli stati del CTMC, $\{\zeta_t, t \geq 0\}$, in ordine lessicografico e indicando con i il livello, per $i \geq 0$, l'insieme di stati come

$$\{(i, n, k) : 0 \leq n \leq \min(i, L), 1 \leq k \leq m\}$$

il generatore (infinitesimale), Q , di questo CTMC è dato dal seguente teorema

Descrizione del processo QBD che governa il sistema e il suo generatore

Theorem

Il generatore infinitesimale Q del processo stocastico CTMC $\{\zeta_t, t \geq 0\}$ ha una struttura a blocchi tridiagonale come segue:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & 0 & \dots & 0 & 0 & 0 & 0 & \dots \\ 0 & Q_{2,1} & Q_{2,2} & Q_{2,3} & \dots & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & 0 & \dots & Q_{L,L-1} & Q_{L,L} & Q^+ & 0 & \ddots \\ 0 & 0 & 0 & 0 & \dots & 0 & Q^- & Q^0 & Q^+ & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & Q^- & Q^0 & Q^+ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Descrizione del processo QBD che governa il sistema e il suo generatore

Dove i blocchi $Q_{i,i}$ sono definiti come segue:

$$Q_{0,0} = D_0,$$

$$Q_{i,i} = I_{i+1} \otimes D_0 + v\mu_2 E_i^- \otimes I_m - (\mu_1 \hat{I}_i + \mu_2(I_{i+1} - \bar{I}_i)) \otimes I_m, 1 \leq i \leq L,$$

$$Q_{i,i} = Q^0 = I_{L+1} \otimes D_0 + v\mu_2 E_L^- \otimes I_m - (\mu_1 I_{L+1} + \mu_2(I_{L+1} - \bar{I}_L)) \otimes I_m, i > L,$$

$$Q_{i,i+1} = E_i^+ \otimes D_1, 0 \leq i \leq L-1,$$

$$Q_{i,i+1} = Q^+ = I_{L+1} \otimes D_1, i \geq L,$$

$$Q_{1,0} = (1-v)\mu_2 \tilde{E}_1^- \otimes I_m + \mu_1 I_1^- \otimes I_m, 1 \leq i \leq L,$$

$$Q_{i,i-1} = (1-v)\mu_2 \tilde{E}_i^- \otimes I_m + q\mu_1 I_i^- \otimes I_m + (1-q)\mu_1 I_i^+ \otimes I_m, 1 \leq i \leq L,$$

$$Q_{i,i-1} = Q^- = (1-v)\mu_2 E_L^- \otimes I_m + q\mu_1 I_{(L+1)m} + (1-q)\mu_1 I^+ \otimes I_m, i > L.$$

Condizione di ergodicità del processo QBD

Theorem

Il processo stocastico CTMC $\{\zeta_t, t \geq 0\}$ è ergodico se e solo se vale la seguente disuguaglianza:

$$\lambda < \mu_1 + \mu_2(1 - \nu) \frac{L(1 - q)\mu_1}{L(1 - q)\mu_1 + \mu_2}$$

Dimostrazione del teorema

Dimostrazione

Il criterio per l'ergodicità del QBD con il generatore di forma data come nel teorema precedente soddisfa l'ineguaglianza:

$$yQ^-e > yQ^+e$$

dove il vettore y è l'unica soluzione del sistema

$$y(Q^- + Q^0 + Q^+) = \mathbf{0}, \quad ye = 1$$

Dimostrazione

Si può inoltre verificare facilmente che

$$Q^- + Q^0 + Q^+ = I_{L+1} \otimes (D_0 + D_1) + S \otimes I_m$$

dove

$$S = \begin{pmatrix} -\mu_1(1-q) & 0 & 0 & \dots & 0\mu_1(1-q) & \\ \mu_2 & -\mu_2 & 0 & \dots & 0 & 0 \\ 0 & \mu_2 & -\mu_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \mu_2 & -\mu_2 \end{pmatrix}$$

Dimostrazione

dove usando le regole del mixed product per il prodotto di Kronecker, e ricordando che

$$\delta(D_0 + D_1) = 0, \quad \delta e = 1$$

si verifica facilmente caratteristiche

$$y = x \otimes \delta$$

dove x è soluzione del sistema

$$xS = 0, \quad xe = 1$$

Dimostrazione

per sostituzione diretta, si verifica facilmente che le componenti del vettore $x = (x_0, x_1, \dots, x_L)$, corrispondenti alle uniche soluzioni del sistema visto prima, sono date da

$$x_0 = \frac{\mu_2}{L(1-q)\mu_1 + \mu_2}, \quad x_i = \frac{\mu_1(1-q)}{L(1-q)\mu_1 + \mu_2}, \quad i = 1, \dots, L$$

La tesi segue dalle equazioni viste in precedenza assieme alla definizione di λ . □

Osservazioni sulla dimostrazione

Osservazione 1

- La condizione di ergodicità richiede che il tasso di arrivo dei clienti per unità di tempo debba essere inferiore al tasso di servizio che i clienti ricevono per unità di tempo quando il sistema è sovraccarico.
- Il tasso di servizio medio totale nel modello di coda è dato dalla somma del tasso di servizio fornito dal server principale e del tasso di servizio fornito dal server secondario.

Possiamo esprimere il tasso di servizio medio totale come segue:

$$\mu = \mu_1 + \mu_2(1 - v) \frac{L(1 - q)\mu_1}{L(1 - q)\mu_1 + \mu_2}$$

Osservazioni sulla dimostrazione

Osservazione 2

Calcoliamo la probabilità x_0 che il secondo server non sia presente nel sistema in un qualsiasi momento in cui il sistema è sovraccarico.

- Quando il sistema attiva un server secondario la durata media del server secondario continuamente presente nel sistema è data da $\frac{L}{\mu_2}$. Pertanto, abbiamo:

$$x_0 = \frac{\frac{1}{\mu_1(1-q)}}{\frac{1}{\mu_1(1-q)} + \frac{L}{\mu_2}} = \frac{\mu_2}{L(1-q)\mu_1 + \mu_2}$$

Probabilità stazionarie

Sotto l'assunzione che la condizione di ergodicità sia valida, esistono le seguenti probabilità stazionarie degli stati del CTMC $\{\zeta_t, t \geq 0\}$:

$$\pi(i, n, \zeta) = \lim_{t \rightarrow \infty} P\{i_t = i, n_t = n, \zeta_t = \zeta\}, \quad i \geq 0$$

Consideriamo i vettori riga delle probabilità di stato stazionario π_i come segue

$$\pi(i, n) = (\pi(i, n, 1), \dots, \pi(i, n, m))$$

$$\pi_i = (\pi(i, 0), \dots, \pi(i, \min\{i, L\})), \quad i \geq 0$$

Probabilità stazionarie

Sappiamo che i vettori di probabilità stazionari $\pi_i, i \geq 0$, soddisfano il sistema di equazioni algebriche lineari (equazioni di equilibrio):

$$(\pi_0, \pi_1, \pi, \dots)Q = 0 \quad (\pi_0, \pi_1, \pi, \dots)e = 1$$

dove Q è la matrice di transizione del CTMC $\{\zeta_t, t \geq 0\}$ e e è il vettore colonna di tutti gli elementi 1

Algoritmo per risolvere il sistema di equazioni di equilibrio

Vediamo un algoritmo per risolvere il sistema infinito di equazioni di equilibrio che sfrutta la struttura tridiagonale a blocchi ma dipendente dal livello del generatore per i livelli minori di $L + 1$.

Algoritmo per risolvere il sistema di equazioni di equilibrio

Theorem

I vettori $\pi_i, i \geq 0$, sono trovati come soluzione del sistema di equazioni algebriche lineari:

$$\pi_i = \alpha_i \left(\sum_{l=0}^{\infty} \alpha_l e \right)^{-1}, \quad i \geq 0$$

dove il vettore α_0 è calcolato come l'unica soluzione del sistema di equazioni

$$\alpha_0 (Q_{0,0} + Q_{0,1} G_0) = 0, \quad \alpha_0 e = 1$$

ed i vettori $\alpha_i, i \geq 1$, sono definiti come

$$\alpha_i = \alpha_0 \prod_{l=1}^i R_l, \quad i \geq 1$$

Algoritmo per risolvere il sistema di equazioni di equilibrio

Theorem

Altrimenti tramite la formula ricorsiva

$$\alpha_i = \alpha_{i-1} R_i, \quad i \geq 1$$

dove

$$R = \begin{cases} -Q_{i-1,i}(Q_{i,i} + Q_{i,i+1}G_i)^{-1}Q & 1 \leq i \leq L-1 \\ -Q_{L-1,L}(Q_{L,L} + Q^+G)^{-1} & i = L \\ -Q^+(Q^0 + Q^+G)^{-1} = R & i > L \end{cases}$$

Algoritmo per risolvere il sistema di equazioni di equilibrio

Theorem

Le matrici stocastiche G_i sono calcolate utilizzando la seguente formula ricorsiva all'indietro:

$$G_L = G$$

$$G_{L-1} = -(Q_{L,L} + Q^+ G_L)^{-1} Q_{L,L-1}$$

$$G_i = -(Q_{i+1,i+1} + Q_{i+1,i+2} G_{i+1})^{-1} Q_{i+1,i}, \quad i = L-2, L-3, \dots, 0$$

dove la matrice G è la minima soluzione non negativa dell'equazione quadratica matriciale

$$Q^+ G^2 + Q^0 G + Q^- = 0$$

Algoritmo per risolvere il sistema di equazioni di equilibrio

- L'algoritmo proposto è una modifica dell'algoritmo per il calcolo della distribuzione stazionaria del CTMC asintoticamente quasi-Toeplitz.
- Utilizzando la ricorsione di vettori anziché quella di matrici si ha una significativa riduzione della memoria del computer e del tempo di esecuzione.
- Le inverse delle matrici utilizzate nell'algoritmo sono sub-generatori irriducibili e semi-stabili, il che rende stabile l'implementazione numerica dell'algoritmo.

Code di tipo GI/M/1

Definiamo come prima cosa lo spazio degli stati Ω del CTMC come:

$$\Omega = \{(i, j, k) : i \geq 0, 0 \leq j \leq K, 1 \leq k \leq m\}$$

Definiamo il livello

$$\mathbf{i} = \{(i, j, k) : 0 \leq j \leq L, 1 \leq k \leq m\} = \{(\mathbf{i}, 0), \dots, (\mathbf{i}, L)\}, i \geq 0$$

- il livello (\mathbf{i}, j) indica che il server principale è occupato, ci sono $i - 1$ clienti in attesa nella coda principale; il server secondario è occupato e il processo di arrivo si trova in varie fasi
- Il livello $(0, 0)$ corrisponde al sistema inattivo con il processo MAP in una delle m fasi.

Il generatore del CTMC

Il generatore \tilde{Q} della CTMC che governa il sistema in studio è:

$$\tilde{Q} = \begin{pmatrix} B_0 & A_0 & & & & & & & & \\ B_1 & A_1 & A_0 & & & & & & & \\ B_2 & A_2 & A_1 & A_0 & & & & & & \\ \vdots & & \ddots & \ddots & \ddots & & & & & \\ B_L & & & & & A_2 & A_1 & A_0 & & \\ B_{L+1} & & & & & A_2 & A_1 & A_0 & & \\ & A_{L+2} & & & & & & A_2 & A_1 & A_0 \\ & & A_{L+2} & & & & & A_2 & A_1 & A_0 \\ & & & \ddots & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

Il generatore del CTMC

Dove abbiamo:

$$B_0 = \begin{pmatrix} D_0 & & & & \\ \tilde{v}\mu_2 I & D_0 - \mu_2 I & & & \\ & \tilde{v}\mu_2 I & D_0 - \mu_2 I & & \\ & & \ddots & \ddots & \\ & & & \tilde{v}\mu_2 I & D_0 - \mu_2 I \end{pmatrix}$$

Il generatore del CTMC

Dove abbiamo:

$$A_0 = \begin{pmatrix} D_1 & & & & \\ v\mu_2 I & D_1 & & & \\ & v\mu_2 I & D_1 & & \\ & & \ddots & \ddots & \\ & & & v\mu_2 I & D_1 \end{pmatrix}$$

$$A_1 = B_0 - \mu_1 I$$

$$A_2 = \mu_1 \Delta(q, 1, \dots, 1)$$

$$B_1 = \mu_1 I$$

$$B_r = \rho \mu_1 (e_r^T \otimes e(L+1)) \quad 2 \leq r \leq L+1$$

$$A_{L+2} = B_{L+1}$$

Proprietà delle queue di tipo GI/M/1

Proprietà 1

Sia

$$\tilde{y} = (\tilde{y}_0, \dots, \tilde{y}_L)$$

il vettore invariante di $A = \sum_{i=0}^{L+2} A_i$. Allora:

$$\tilde{y}_0 = \delta(\mu_2 I - D_0 - D_1)[\mu_2 U + L\rho\mu_1 I - D_0 - D_1]^{-1}$$

$$\tilde{y}_r = \rho\mu_1\pi_0(\mu_2 I - D_0 - D_1)^{-1}, \quad 1 \leq r \leq L$$

Proprietà delle queue di tipo G1/M/1

Proprietà 2

La condizione di stabilità

$$\tilde{y}A_0e < \tilde{y} \sum_{i=1}^{L+2} (i-1)A_i e$$

si riduce a:

$$\lambda < \mu_1 + \mu_2(1-\nu) \frac{L(1-q)\mu_1}{L(1-q)\mu_1 + \mu_2}$$

Proprietà delle queue di tipo GI/M/1

Proprietà 3

Data R la matrice di rate, soddisfa l'equazione matriciale non lineare data da:

$$R^{L+2}A_{L+2} + R^2A_2 + RA_1 + A_0 = 0$$

Proprietà delle queue di tipo GI/M/1

Proprietà 4

Indicando con $\tilde{\pi}$ il vettore di probabilità in stato stazionario del generatore \tilde{Q} come visto prima, otteniamo qui la soluzione matrice-geometrica classica:

$$\tilde{\pi}_i = \tilde{\pi}_0 R^i, \quad i \geq 1$$

dove $\tilde{\pi}_0$ è ottenuto risolvendo il seguente sistema di equazioni lineari:

$$\tilde{\pi}_0 \left[\sum_{i=0}^{L+1} R^i B_i \right] = 0, \quad \tilde{\pi}_0 e = 1$$

Introduzione ai risultati numerici

Vedremo 3 elementi illustrativi utilizzando 5 processi di arrivo. In particolare prendiamo i 5 MAP come

ERL

Erlang di ordine 5 con parametro 2.5 in ciascuno dei 5 stati. Notare che qui abbiamo $\lambda = 0.5$, $\sigma = 0.899427$ e $\rho_c = 0$.

EXP

Un esponenziale con una frequenza di 0.5. Notare che qui abbiamo $\lambda = 0.5$, $\sigma = 2$ e $\rho_c = 0$.

HEX

Distribuzione iper-esponenziale con una probabilità di mixing data da $(0.5, 0.3, 0.15, 0.04, 0.01)$ con i corrispondenti tassi della distribuzione esponenziale pari a $(1.09, 0.545, 0.2725, 0.13625, 0.068125)$. Qui abbiamo $\lambda = 0.5, \sigma = 3.3942$ e $\rho_c = 0$.

NCR

MAP negativamente correlato, con matrici di rappresentazione:

$$D_0 = \begin{pmatrix} -1.125 & 0.125 & 0 & 0 & 0 \\ 0 & -1.125 & 0.125 & 0 & 0 \\ 0 & 0 & -1.125 & 0.125 & 0 \\ 0 & 0 & 0 & -0.125 & 0 \\ 0 & 0 & 0 & 0 & -2.25 \end{pmatrix}$$

$$D_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.01125 & 0 & 0 & 0 & 1.11375 \\ 2.2275 & 0 & 0 & 0 & 0.0225 \end{pmatrix}$$

dove abbiamo $\lambda = 0.5$, $\sigma = 2.02454$ e $\rho_c = -0.57855$

PCR

MAP positivamente correlato, con matrici di rappresentazione:

$$D_0 = \begin{pmatrix} -1.125 & 0.125 & 0 & 0 & 0 \\ 0 & -1.125 & 0.125 & 0 & 0 \\ 0 & 0 & -1.125 & 0.125 & 0 \\ 0 & 0 & 0 & -0.125 & 0 \\ 0 & 0 & 0 & 0 & -2.25 \end{pmatrix}$$

$$D_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1.11375 & 0 & 0 & 0 & 0.01125 \\ 0.0225 & 0 & 0 & 0 & 2.2275 \end{pmatrix}$$

dove abbiamo $\lambda = 0.5$, $\sigma = 2.02454$ e $\rho_c = -0.57855$

Introduzione ai risultati numerici

Osservazioni

- Le cinque MAP sopra riportate sono qualitativamente diverse.
- Il processo di arrivo denominato PCR è ideale per situazioni di arrivi altamente irregolari.
- La correlazione positiva nel processo PCR ha un impatto significativo e la variabilità nei tempi tra gli arrivi è stata ben documentata in letteratura.

Primo esempio illustrativo

Fissiamo $\mu_1 = 1$, $\mu_2 = 0.5$, $q = 0.5$, e $\nu = 0.4$, e variamo L da 1 a 30.

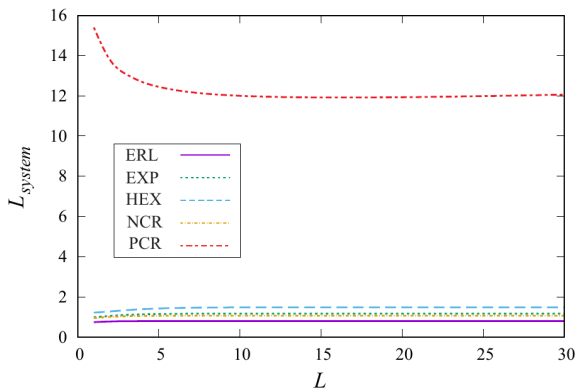


Figure: Impatto di L sul numero medio di clienti nel sistema L_{system} per diversi MAPs

Primo esempio illustrativo

 L_{sec}

Definiamo L_{sec} come il numero medio di clienti nel sistema con server secondari ad un momento arbitrario come:

$$L_{\text{sec}} = \sum_{i=1}^{\infty} \sum_{n=1}^{\min\{i, L\}} n \pi(i, n) e$$

Primo esempio illustrativo

- La figura mostra che PCR ha un alto numero medio di clienti nel sistema rispetto ad altri processi di arrivo.
- L aumenta il numero medio di clienti nel sistema per i primi quattro MAP, ma per PCR il trend non è crescente a causa della correlazione positiva.
- L'alta L aumenta la probabilità di avere più clienti nel sistema, soprattutto per i primi quattro MAP.
- Tuttavia, per gli arrivi PCR, L diminuisce il numero medio di clienti nel sistema perché i server secondari aiutano a ripulire la coda.

Primo esempio illustrativo

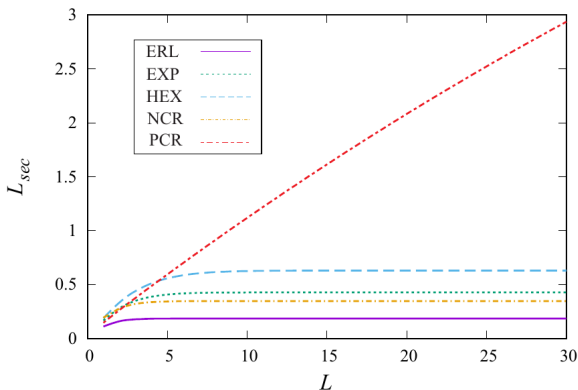


Figure: Dipendenza del numero medio di clienti con il server secondario L_{sec} al variare di L per diversi MAPs

Primo esempio illustrativo

- L_{sec} aumenta all'aumentare di L , come previsto.
- Il valore di L_{sec} è elevato per PCR e solo per valori piccoli di L è inferiore per ERL-NCR.
- L'alta irregolarità degli arrivi nel processo PCR causa la fame del sistema, mentre una maggiore varianza implica un grande numero di clienti nel sistema per ERL-HEX.

Primo esempio illustrativo

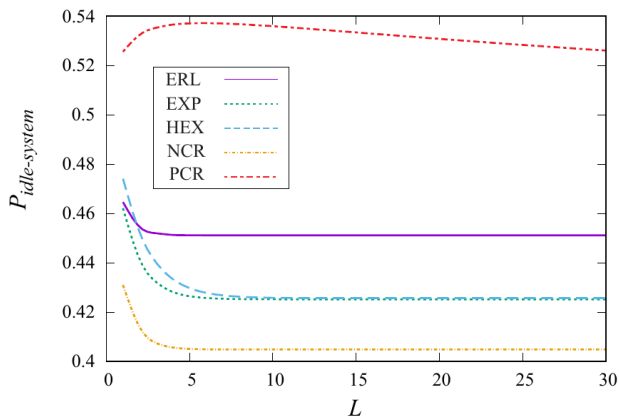


Figure: Dipendenza della probabilità $P_{idle-system}$ rispetto ad L che il sistema sia in idle ad un momento arbitrario, per diversi MAPs

Primo esempio illustrativo

$P_{\text{idle-system}}$

Definiamo la probabilità che il sistema sia in equilibrio ad un momento arbitrario come:

$$P_{\text{idle-system}} = \pi_0 e$$

- Esiste una grande differenza nella misura a seconda dei diversi MAPs utilizzati.
- Il valore ottimale di L dipende dall'obiettivo: ad esempio, per il processo di arrivo PCR, il valore ottimale di L è 16 se si cerca di minimizzare L_{system} , ma è 6 se si massimizza $P_{\text{idle-system}}$.

Primo esempio illustrativo

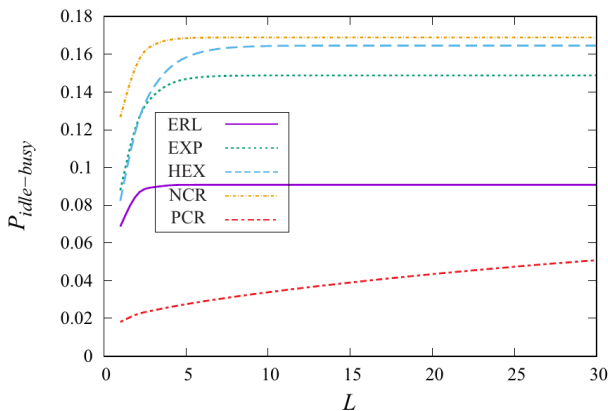


Figure: Dipendenza della probabilità $P_{\text{idle-busy}}$ rispetto ad L che il main server sia in idle quando il server secondario è in occupato, per diversi $M\Delta P_c$

Primo esempio illustrativo

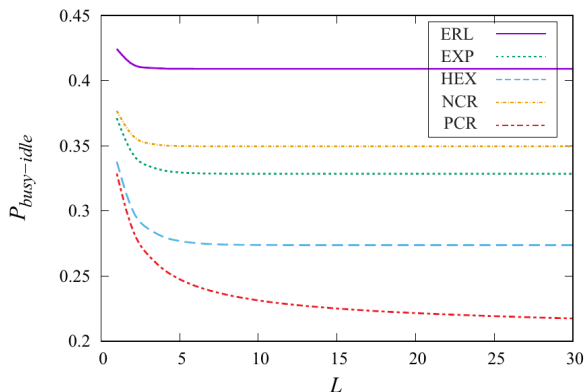


Figure: Dipendenza della probabilità $P_{\text{busy-idle}}$ rispetto ad L che il main server sia occupato quando il server secondario è in idle, per diversi MAPs

Primo esempio illustrativo

$P_{\text{idle-busy}}$

Definiamo la probabilità che il main server sia in idle quando il server secondario è occupato come:

$$P_{\text{idle-busy}} = \sum_{n=1}^L \pi(n, n) e$$

$P_{\text{busy-idle}}$

Definiamo la probabilità che il main server sia occupato quando il server secondario è in idle come:

$$P_{\text{busy-idle}} = \sum_{i=0}^{\infty} \pi(i, 0) e$$

Secondo esempio illustrativo

- L'obiettivo è valutare l'impatto dei parametri q e ν sulla prestazione del sistema.
- Fissiamo il valore di L a 10 e i tassi di servizio μ_1 e μ_2 a 1 e 0.5.
- Si variano i valori di q e ν da 0 a 1 con passo 0.05 e si analizza l'impatto sulle misure di prestazione del sistema.

In questo esempio ci concentriamo sul processo di arrivo PCR, la cui scelta è basata sul comportamento di questo processo sulle misure evidenziato nel primo esempio illustrativo

Secondo esempio illustrativo

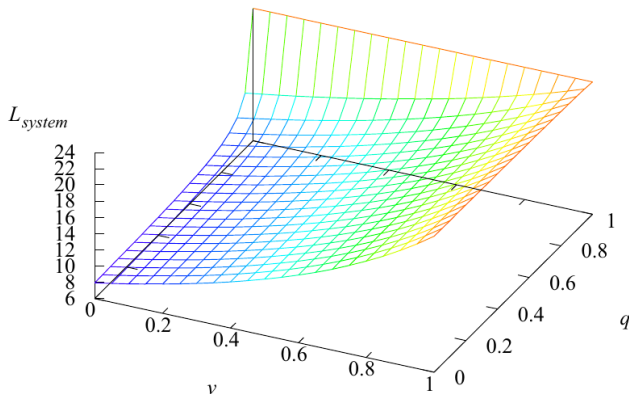


Figure: Dipendenza del numero medio di clienti nel sistema L_{system} rispetto a q e v

Secondo esempio illustrativo

- L'analisi mostra che il valore di L_{system} è minimo a 7.9328 quando q e ν sono entrambi uguali a 0.
- Aumentando q o ν , il valore di L_{system} aumenta, con un aumento più veloce quando uno o entrambi si avvicinano a 1.
- Quando $q = 1$, il sistema diventa un modello MAP/M/1 classico e il valore di L_{system} diventa 22.30425 per tutti i valori di ν .
- L'uso di un server secondario riduce il numero medio di clienti nel sistema di oltre il 40%, e il punto di interruzione per il modello classico è $\nu^* \sim 0.985$.

Secondo esempio illustrativo

- Si aumenta λ del 50% a 0.75 per testare l'importo della riduzione del numero medio di clienti nel sistema.
- Mantenendo gli altri parametri costanti, si ottiene una riduzione superiore al 52,8%.
- Ciò suggerisce che l'aggiunta di un server secondario beneficia notevolmente l'aumento del carico del sistema anche con un tasso di insoddisfazione del cliente del 50%.

Secondo esempio illustrativo

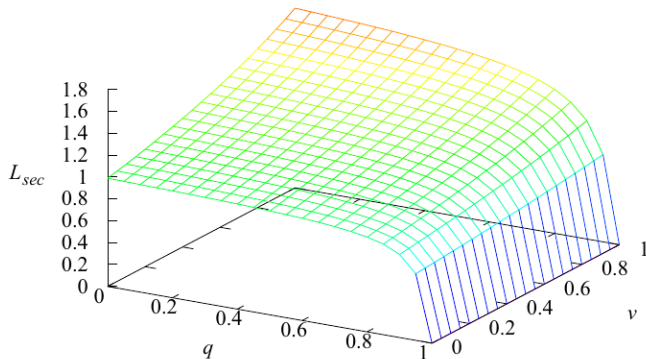


Figure: Dipendenza del numero medio di clienti nel sistema L_{sec} rispetto a q e v con $\lambda = 0.75$

Secondo esempio illustrativo

- La figura mostra come il numero medio di clienti con il server secondario L_{sec} dipenda dai parametri q e ν .
- L_{sec} diminuisce significativamente quando q si avvicina a 1 e quando i clienti sono raramente reclutati per diventare server secondari.
- L_{sec} ha il valore massimo quando q è uguale a zero e ν è vicino a 1, ma questo può creare ulteriore lavoro per il sistema e riflettersi negativamente sulla fornitura di servizi.

Secondo esempio illustrativo

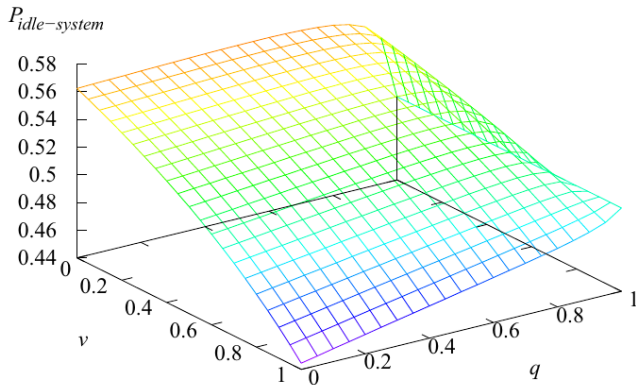


Figure: Dipendenza della probabilità $P_{idle-system}$ che il sistema sia in idle ad un momento arbitrario rispetto a q e v .

Secondo esempio illustrativo

- $P_{\text{idle-system}}$ ha il valore minimo di 0.4445 quando $\nu = 1$ e $q = 0$, il che è intuitivo poiché servire nuovamente i clienti dopo aver passato attraverso un server secondario mette un carico sul sistema.
- $P_{\text{idle-system}}$ aumenta quando q aumenta e/o ν diminuisce, con un valore massimo di 0.5652 ottenuto quando $q = 0.65$ e $\nu = 0$.
- Nel sistema MAP/M/1 classico corrispondente, questa misura è $P_{\text{idle-system}} = 0.5$.

Terzo esempio illustrativo

- In questo esempio, si analizza l'impatto della variazione dei tassi di servizio μ_1 e μ_2 quando tutti gli altri parametri sono fissati.
- I parametri fissati sono $L = 10$, $q = 0.5$, $\nu = 0.4$, e $\lambda = 0.5$.
- I tassi μ_1 e μ_2 vengono variati da 0.25 a 2.0 con incrementi di 0.05, ma per soddisfare la condizione di ergodicità, il valore di μ_2 viene limitato quando μ_1 è piccolo.
- Solo per $\mu_1 \geq 0.4$, il valore di μ_2 può essere variato da 0.25, come originariamente indicato

Terzo esempio illustrativo

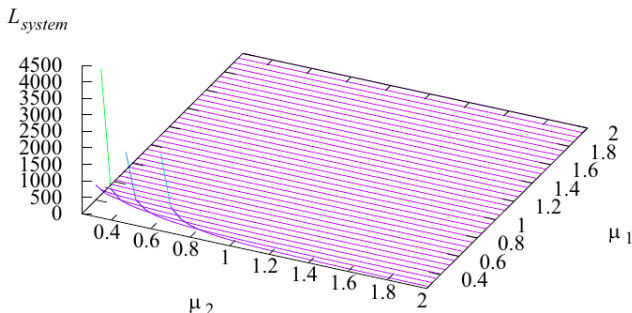


Figure: Dipendenza del numero medio di clienti nel sistema L_{system} rispetto a μ_1 e μ_2

Terzo esempio illustrativo

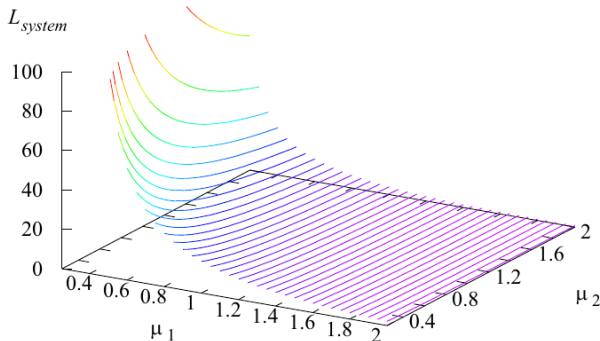


Figure: Dipendenza del numero medio di clienti nel sistema L_{system} rispetto a μ_1 e μ_2

Terzo esempio illustrativo

- La condizione di ergodicità limita il valore di μ_2 per valori piccoli di μ_1 .
- Gran parte della superficie mostrata nella prima figura è piatta a causa della violazione della condizione di ergodicità.
- La tendenza generale è che L_{system} diminuisce quando aumenta μ_1 o μ_2 (per valori non piccoli di μ_1).

Terzo esempio illustrativo

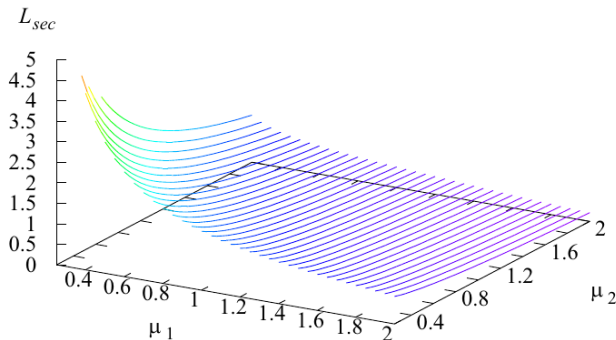


Figure: Dipendenza del numero medio di clienti con server secondario L_{sec} rispetto a μ_1 e μ_2

Terzo esempio illustrativo

- Il valore massimo di L_{sec} è di circa 5 quando μ_1 e μ_2 sono piccoli.
- Con un aumento di μ_1 e μ_2 , il valore di L_{sec} diminuisce come ci si aspetterebbe.
- Per valori piccoli di μ_1 , la diminuzione è significativa all'aumentare di μ_2 ; per valori più grandi di μ_1 , notiamo un tasso insignificante di diminuzione in L_{sec} con un aumento di μ_2 .

Conclusioni

- Il sistema di coda analizzato prevede la possibilità di reclutare un cliente già servito come server secondario per aiutare il server principale.
- Il processo di arrivo dei clienti è stato modellizzato utilizzando un processo di punto Markoviano versatile, MAP.
- È stata considerata la possibilità che i clienti insoddisfatti con il servizio fornito dal server secondario possano ritornare nel sistema.
- L'analisi dello stato stazionario della catena di Markov multidimensionale ha permesso di ottenere risultati numerici utili per prendere decisioni manageriali.

Generalizzazione del modello

- Si può effettuare il servizio fornito dal server secondario in gruppi.
- Si può rilassare l'ipotesi di avere solo un server secondario e vedere l'impatto dell'aumento a 2.
- Si possono incorporare diverse politiche di controllo, come la possibilità di reclutare molti server secondari con due tipi di clienti