Queueing System with Potential for Recruiting Secondary Servers

Luca Lombardo

Contents

1 Introduzione

1.1 Queueing Theory

I modelli di coda sono utilizzati per rappresentare sistemi di risorse, tradizionalmente chiamati "server", che devono essere utilizzati da diversi utenti, chiamati "clienti". La terminologia deriva da applicazioni come gli sportelli bancari, le reception degli hotel, i caselli autostradali, e così via, dove i clienti effettivamente si mettono in coda finché non vengono serviti da un dipendente.

Le code semplici consistono di un solo server che attende un solo cliente alla volta, in ordine di arrivo, con l'aggiunta dell'assunzione che i clienti siano indefinitamente pazienti. Si assume che il tempo sia discretizzato in intervalli di lunghezza fissa, che un numero casuale di clienti si unisca al sistema durante ogni intervallo e che il server rimuova un cliente dalla coda alla fine di ogni intervallo, se presente. Definendo α_n come il numero di nuovi arrivi durante l'intervallo $[n-1, n)$ e X_n come il numero di clienti nel sistema al tempo n, abbiamo

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

Se α_n è una collezione di variabili casuali indipendenti, allora X_{n+1} è condizionalmente indipendente da $X_0, \ldots, X_{n-1}$ se X_n è noto. Se, inoltre, le α_n sono identicamente distribuite, allora X_n è omogenea. Lo spazio degli stati è $\mathbb{N}$ e la matrice di transizione è

$$P = \begin{pmatrix} q_0 + q_1 & q_2 & q_3 & q_4 & \cdots \\ q_0 & q_1 & q_2 & q_3 & \ddots \\ \vdots & q_0 & q_1 & q_2 & \ddots \\ 0 & & \ddots & \ddots & \ddots \end{pmatrix} \tag{2}$$

dove q_i è la probabilità $P[\alpha = i]$ che i nuovi clienti che entrino in coda durante un intervallo di un'unutà di tempo, mentre α denota ognuna delle possibili distribuzioni di α_n identicamente distribuite.

1.2 Obiettivi del paper

Il paper presenta un nuovo approccio per migliorare i modelli di coda con l'utilizzo di server secondari temporanei, reclutati tra i clienti stessi. Questi server secondari sono disponubili solo temporaneamente e fornuranno servizi in gruppi di diverse dimensionu. Dopo aver servito esattamente un gruppo, i server secondari lasceranno il sistema, permettendo ai clienti di proseguire le loro attività senza essere trattenuti. Il contributo principale del paper è l'introduzione del concetto di reclutamento di server secondari da parte dei clienti, in modo da aiutare il sistema. I risultati numerici indicano che il modello proposto funziona meglio del modello di coda classico corrispondente. Questo può aiutare i responsabili del sistema a reclutare server secondari quando necessario per migliorare le prestazionu del sistema.

Il paper si concentra in particolare sulle due seguenti caratteristiche che sono intrinseche in alcuni sistemi del mondo reale e non sono state studiate in passato: (i) un server secondario verrà assegnato ad un gruppo (che non supererà una soglia finita prestabilita); questo server offrirà i servizi uno alla volta; e una volta che tutti i clienti assegnati sono stati serviti, il server secondario lascerà anche il sistema; e (ii) con una certa probabilità, un cliente servito da un server secondario diventa insoddisfatto e quindi torna al sistema principale per ottenere un nuovo servizio.

2 Modello Matematico

Consideriamo un sistema di coda a singolo server in cui gli arrivi avvengono secondo un processo di arrivo markoviano (MAP) con matrici di parametro (D0, D1) di ordine m. Un Processo di Arrivo Markoviano (MAP) è un processo stocastico usato per descrivere il comportamento degli arrivi in un sistema di coda. In un sistema di coda, gli arrivi rappresentano le richieste o le entità che si presentano al sistema per essere servite.

Un MAP è caratterizzato dalla sua distribuzione di probabilità di interarrivo, che descrive il tempo tra due arrivi consecutivi, e dalla sua distribuzione di probabilità di dimensione, che descrive il numero di entità che arrivano contemporaneamente.

Un MAP può essere definito come un processo di Markov a tempi continui, dove la probabilità di transizione dipende solo dallo stato corrente del sistema e non dalla sua storia passata. In altre parole, la probabilità di passare da uno stato ad un altro dipende solo dallo stato attuale e dal tempo trascorso da quando lo stato attuale è stato raggiunto.Il MAP è ideale in situazioni in cui può essere presente una correlazione nei tempi tra gli arrivi.

2.1 Markovian arrival process (MAP)

Il generatore irriducibile del MAP è dato da $D_0 + D_1$. Sia δ il vettore invariante tale che

$$\delta(D_0 + D_1) = \mathbf{0}, \qquad \delta e = 1 \tag{3}$$

La matrice D_0 governa le transizioni corrispondenti al generatore sottostante che non produce arrivi ed è una matrice non-singolare con elementi diagonali negativi e elementi non-diagonali non negativi. Mentre la matrice D_1 governa quelle transizioni corrispondenti agli arrivi nel sistema ed è una matrice non-negativa. Andiamo a vedere adesso 3 proprietà fondamentali di un MAP.

Rate medio di arrivi

$$\lambda = \delta D_1 e$$

rappresenta il numero medio di arrivi che si verificano nell'unità di tempo. Se si considera il caso in cui il processo è Poissoniano, allora il tasso di arrivo è costante; tuttavia, in un processo MAP, il tasso di arrivo può variare nel tempo a seconda delle condizioni del sistema.

Varianza dei tempi di arrivo

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

rappresenta la variazione dei tempi di interarrivo rispetto alla media. In altre parole, la varianza misura quanto i tempi di interarrivo si discostano dalla media. Se la varianza è bassa, i tempi di interarrivo tendono ad essere più uniformi e costanti, mentre se la varianza è alta, i tempi di interarrivo tendono ad essere più irregolari e meno prevedibili

Correlazione tra due successivi tempi di arrivo

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

indica quanto il tempo di arrivo successivo dipende dal tempo di arrivo precedente. In altre parole, la correlazione misura il grado di dipendenza tra due eventi successivi. Se la correlazione è alta, il tempo di arrivo successivo dipende fortemente dal tempo di arrivo precedente; se la correlazione è bassa o nulla, il tempo di arrivo successivo è indipendente dal tempo di arrivo precedente. La correlazione è spesso misurata attraverso il coefficiente di correlazione ρ_c, che varia tra -1 e 1, dove i valori positivi indicano una correlazione positiva (cioè, l'aumento del tempo di arrivo precedente corrisponde all'aumento del tempo di arrivo successivo) e i valori negativi indicano una correlazione negativa (cioè, l'aumento del tempo di arrivo precedente corrisponde alla diminuzione del tempo di arrivo successivo). Una correlazione pari a zero indica che non c'è dipendenza tra i tempi di arrivo successivi.

Il sistema ha un singolo server che offre servizi in modo FCFS. Questo server sarà chiamato server principale. I tempi di servizio sono esponenziali con parametro μ_1. Con probabilità $p, 0 \leq p \leq 1$, un cliente servito può essere reclutato (o optato dal punto di vista del cliente servito) per servire altri clienti in attesa nel sistema (assumendo che la dimensione della coda sia positiva) a condizione che non ci sia già un altro server secondario che sta servendo. Un tale server è chiamato server secondario. In altre parole, una reclutamento avviene solo quando c'è almeno un cliente in attesa nella coda e quando non c'è altro server secondario presente nel sistema. Pertanto, il sistema può avere al massimo due server in qualsiasi momento. Si noti che con probabilità $q = 1 - p$, il cliente servito, che può diventare il server secondario, non accetta di farlo e lascia il sistema. Quando viene reclutato un server secondario, il server verrà assegnato a un gruppo di, diciamo, i clienti, dove $i = \min\{\text{numero nella coda}, L\}$, dove L è un pre-determinato positivo finito intero. Il server secondario offrirà servizi ai clienti del gruppo uno alla volta e i tempi di servizio sono distribuiti in modo esponenziale con parametro μ_2. Un cliente che riceve un servizio da un server secondario potrebbe non essere soddisfatto del servizio ricevuto e richiedere di essere servito di nuovo con probabilità v, $0 \leq v \leq 1$, e con probabilità $\bar{v} = 1 - v$ lascerà il sistema. I clienti insoddisfatti sono reinseriti nel sistema. Una volta che il server secondario ha finito di servire tutti i clienti assegnati, il sistema rilascerà questo server. Si noti che prendendo $p = 0$ (in questo caso v non ha alcun ruolo e può essere ignorato), otteniamo il modello di coda a singolo server classico. Questo caso viene utilizzato solo come verifica dell'accuratezza nei calcoli numerici e non è altrimenti interessante. Il caso in cui $v = 1$ non è interessante poiché in questo caso ogni cliente servito da un server secondario viene reinserito nel sistema e l'assunzione di server secondari rallenta solo il sistema nell'offrire servizi.

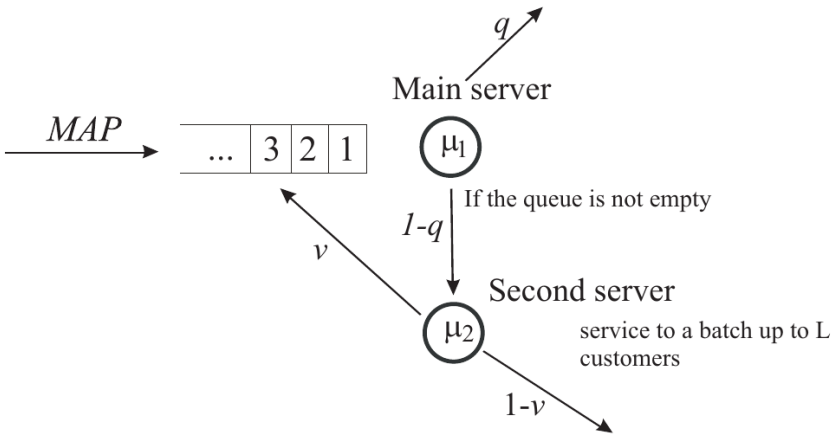

Figure 1: Structure of the system under study

3 QBD Approach to the Steady State Analysis

Analizzeremo il modello di coda in studio in stato stazionario. L'analisi può essere effettuata tramite il processo QBD o tramite un tipo GI/M/1. In questa sezione, adotteremo l'approccio QBD, mentre nella prossima sezione evidenzieremo brevemente l'altro approccio. Come è noto, il processo QBD è un caso particolare della catena di Markov a tempo continuo (CTMC). *Concetto dato per buono*

3.1 Description of the QBD Process Governing the System and Its Generator

CTMC: Ricordiamo velocemente la definizione di Catena di Markov a tempo continuo (CTMC). Il processo stocastico $\{X(t) : t \in \mathbb{R}^+\}$ in uno spazio di stati al più numberabile E è un processo di Markov omogeneo se

$$P[X(t+s) = j \mid X(u) : 0 \le u \le t] = P[X(t+s) = j \mid X(t)]$$

e se

$$P[X(t+s) = j \mid X(t) = i] = P[X(s) = j \mid X(0) = i]$$

per tutti gli stati $i, j \in E$, per tutti i tempi $t \ge 0$ e per tutti gli intervalli di tempo $s \ge 0$.

Un *quasi-death-birth process* (QBD) è un caso particolare di una catena di Markov a tempo continuo (CTMC) che può essere utilizzato per modellare certi tipi di sistemi di coda. Ci sono due tipi di eventi che possono verificarsi: eventi di morte e eventi di nascita.

- Un evento di morte avviene quando un cliente lascia il sistema (i.e finisce di essere servito e se ne va)
- Un evento di nascita avviene quando un nuovo cliente entra nel sistema

Un modo per creare questo tipo di processi è quello di "combinare" le strutture delle markov chains di tipo M/G/1 e G/M/1, imponendo le restrizioni di entrambe, in modo che il processo possa cambiare di un livello alla volta, diventando così skip-free in entrambe le direzioni. Si potrebbe pensare al QBD come un semplice lista lineare in evoluzione: ogni livello è un nodo nella lista ed il processo è autorizzato a muoversi da un nodo ad uno dei suoi due vicini. Queste catene di Markov sono chiamate processi quasi-nascita-e-morte (QBD) e la loro matrice di transizione è

$$P = \begin{pmatrix} B_0 & A_1 & & & 0 \\ A_{-1} & A_0 & A_1 & & \\ & A_{-1} & A_0 & A_1 & \\ & & A_{-1} & A_0 & \ddots \\ 0 & & & \ddots & \ddots \end{pmatrix}, \quad A_{-1}, A_0, A_1, \in \mathbb{R}^{m\times m}, \quad B_0 \in \mathbb{R}^{m\times m}$$

dove gli elementi sono matrici non negativa tali che $A_{-1} + A_0 + A_1$ e $B_0 + B_1$ sono stocastiche. Quindi il generatore infinitesimale di un processo QBD è una matrice tridiagonale a blocchi infinita che descrive la probabilità di transizione del sistema da uno stato i ad uno stato j, in un dato istante di tempo t, attraverso un evento infinitesimo. Osserviamo che un QBD può essere visto come una catena di Markov di tipo M/G/1 ma anche come una catena di Markov di tipo G/M/1. Vediamo come sono fatti gli elementi di questo generatore infinitesimale.

Elementi Diagonali Per tutte le $i \neq j$, $Q_{i,j}$ è il rate istantaneo di transizione dallo stato i allo stato j. Ovvero, $Q_{i,j}h$ è la probabilità condizionale che $X(t+h) = j$, data la condizione che $X(t) = i$, per un intervallo di tempo h abbastanza piccolo. $Q_{i,j}$ è non negativo e strettamente positivo se è possibile spostarsi da i a j in un solo salto.

Elementi non Diagonali Gli elementi diagonali sono tali che

$$Q_{i,i} = -\sum_{j \in E, j \neq i} Q_{i,j}$$

Possiamo vederla in questo modo: il processo rimane in ogni stato durante un intervallo di tempo esponenzialmente distribuito, con parametro $q_i = -Q_{i,i}$, per ogni i, prima d.i saltare allo stato successivo. Se $Q_{i,i} = 0$, allora $Q_{i,j} = 0$ per tutti j, ciò significa che i è uno stato assorbente: una volta raggiunto, il processo non cambia più e rimane in quel stato per sempre.

Supponiamo adesso che, al tempo $t \geq 0$, indichiamo:

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

Allora, il processo stocastico $\{\zeta_t = (i_t, n_t, \xi_t), , t \geq 0\}$ che descrive il comportamento del modello in esame è una CTMC regolare e irriducibile. Una catena regolare garantisce che il processo raggiungerà uno stato stazionario, cioè un equilibrio nel lungo periodo. Invece, l'irriducibilità garantisce che il processo non si dividerà in due o più sottoprocessi che non comunicano tra loro.

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.

Teorema 3.1. *Il generatore infinitesimale Q del processo stocastico CTMC* $\{\zeta_t, , t \geq 0\}$ *ha una struttura a blocchi tridiagonale*

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

Dove i blocchi $Q_{i,i}$ *non nulli sono definiti come segue*

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

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

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

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

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

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

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

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

Dove si usa la seguente notazione:

- O and I are, respectively, zero and identity matrices of appropriate dimensions as indicated in the suffix;
- $\otimes$ indicates the Kronecker product of matrices (see, e.g., [51–54]);

- E_l^+ is a matrix of dimension $(l+1) \times (l+2)$ with $(E_l^+)_{k,k} = 1$, $0 \leq k \leq l$, and all other entries are zero;
- E_l^- is a square matrix of dimension $l+1$ with $(E_l^-)_{k,k-1} = 1$, $1 \leq k \leq l$, and all other entries are zero;
- $\hat{I}_l$ is a square matrix of dimension $l+1$ with $(\hat{I}_l)_{k,k} = 1$, $0 \leq k \leq l-1$, and all other entries are zero;
- $\bar{I}_l$ is a square matrix of dimension $l+1$ with $(\bar{I}_l)_{0,0} = 1$, and all other entries are zero;
- $\tilde{E}_l^-$ is a matrix of dimension $(l+1) \times l$ with $(\tilde{E}_l^-)_{k,k-1} = 1$, $1 \leq k \leq l$, and all other entries are zero;
- I_l^- is the matrix of dimension $(l+1) \times l$ with $(I_l^-)_{k,k} = 1$, $0 \leq k \leq l-1$, and all other entries are zero;
- I_l^+ is the matrix of dimension $(l+1) \times l$ with $(I_l^+)_{0,l-1} = 1$, $(I_l^+)_{k,k} = 1$, $1 \leq k \leq l-1$, and all other entries are zero;
- I^+ is the matrix of dimension $(L+1) \times (L+1)$ with $(I^+)_{k,k} = 1$, $1 \leq k \leq L$, $(I^+)_{0,L} = 1$, and all other entries are zero.

Proof. Immediata □

3.2 Ergodicity Condition of the QBD Process

Erogidicità a parole In un processo ergodico la sua distribuzione di probabilità si stabilisce su un valore costante a lungo termine, indipendentemente dalle condizioni iniziali. Nel caso di un QBD, la condizione di ergodicità è importante perché determina se il sistema raggiungerà uno stato stazionario o meno. Se il processo è ergodico, allora esiste un unico stato stazionario e il sistema raggiungerà questo stato a lungo termine. D'altra parte, se il processo non è ergodico, allora non esiste uno stato stazionario e il sistema non raggiungerà mai una distribuzione di probabilità costante a lungo termine. Le condizioni di ergodicità, in generale, sono le seguenti:

- Il processo deve essere irriducibile. Ciò significa che ogni stato del processo può essere raggiunto da ogni altro stato con una certa probabilità in un numero finito di passi.
- Il processo deve essere aperiodico. Ciò significa che non esiste un ciclo di stati che il processo attraversa regolarmente.

Formalismo Siamo interessati a studiare l'ergodicità, la probabilità $\pi(j)$ di essere nello stato j dopo un lungo periodo di tempo, indipendentemente dallo stato iniziale i. Formalmente in una catena di Markov si dice *ergodica* se il limite

$$\pi(j) = \lim_{n\to\infty} \mathbb{P}_i\{X_n = j\}$$

esiste per ogni stato j e non dipende dallo stato iniziale i. Questo può anche essere scritto come

$$\pi(j) = \lim_{n\to\infty} (P^n)_{i,j}$$

Andiamo a vedere come si può esprimere questa proprietà nel nostro caso di studio

Teorema 3.2. *Il processo stocastico CTMC $\{\zeta_t, , t \geq 0\}$ è ergodico se e solo se vale la seguente disuguaglianza:*

$$\lambda < \mu_1 + \mu_2(1-\nu)\frac{L(1-q)\mu_1}{L(1-q)\mu_1 + \mu_2} \tag{4}$$

Proof. È noto, grazie all'approccio matriciale-geometrico di Neuts, che il criterio per l'ergodicità del QBD con il generatore di forma vista prima, coincide con la soddisfazione dell'ineguaglianza:

$$yQ^- e > yQ^+ e \tag{5}$$

dove il vettore y è l'unica soluzione del sistema

$$y(Q^- + Q^0 + Q^+) = \mathbf{0}, \qquad ye = 1 \tag{6}$$

Si può inoltre verificare facilmente che

$$Q^- + Q^0 + Q^+ = I_{L+1} \otimes (D_0 + D_1) + S \otimes I_m \tag{7}$$

dove Usando la regole del mixed product per il prodotto di Kronecker, ed usando 3 si verifica che la soluzione del sistema di equazioni lineari è

$$y = x \otimes \delta \tag{8}$$

dove δ è come definita in 3 ed x è la soluzione del sistema

$$xS = 0, \qquad xe = 1 \tag{9}$$

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

per sostituzione diretta, si verifica facilmente che le componenti del vettore $x = (x_0, x_1, ..., x_L)$, corrispondenti alle uniche soluzioni del sistema 9, sono date da

$$x_0 = \frac{\mu_2}{L(1-q)\mu_1 + \mu_2}, \qquad x_i = \frac{\mu_1(1-q)}{L(1-q)\mu_1 + \mu_2}, \qquad i = 1, ..., L \tag{10}$$

La tesi segue delle equazioni 5, 7 e 10 assieme alla definizione di λ. □

Osservazione 3.3. *La condizione di stabilità data nell'Equazione 5 può essere intuitivamente spiegata nel seguente modo. In generale, 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 (nel senso che il numero di clienti presenti nel sistema è molto grande). Qui, il tasso di arrivo dei clienti è λ per unità di tempo. Il tasso di servizio dei clienti quando il sistema è sovraccarico è la somma di μ_1 (il tasso di servizio per unità di tempo fornito dal server principale) e il tasso di servizio (per unità di tempo) fornito dal server secondario. Quest'ultimo tasso di servizio è* 0 *quando il server secondario non è presente nel sistema, il che avviene con probabilità x_0. Quando il server secondario è presente nel sistema, che avviene con probabilità* $(1-x_0)$*, i clienti ricevono il servizio e lasciano il sistema ad un tasso di $\mu_2(1-\nu)$ per unità di tempo. Pertanto, il tasso di servizio medio totale è dato da:*

$$\mu = \mu_1 + \mu_2(1-\nu)\frac{L(1-q)\mu_1}{L(1-q)\mu_1 + \mu_2} \tag{11}$$

da cui segue che la condizione di ergodicità vista in 5

Osservazione 3.4. *La probabilità, x_0, che il secondo server non sia presente nel sistema in un qualsiasi momento in cui il sistema è sovraccarico può essere facilmente calcolata dalla seguente considerazione. Si considerino i periodi in cui il server secondario non è presente nel sistema (ovvero, la durata media di questo periodo è $\frac{1}{\rho\mu_1}$) alternati ai periodi in cui il server secondario è presente nel sistema. Quando il sistema attiva un server secondario (quando il sistema è sovraccarico, il server secondario viene assegnato a gestire L per i servizi), la durata media del server secondario continuamente presente nel sistema è data da $\frac{L}{\mu_2}$. Pertanto, abbiamo:*

$$x_0 = \frac{\frac{1}{\mu_1(1-q)}}{\frac{1}{\mu_1(1-q)} + \frac{L}{\mu_2}} = \frac{\mu_2}{L(1-q)\mu_1 + \mu_2} \tag{12}$$

che corrisponde all'espressione 10 vista in precedenza.

3.3 Probabilità Stazionaria

Lo stato stazionario di una catena di Markov a tempo continuo (CTMC) è un punto di equilibrio a lungo termine, in cui la distribuzione di probabilità della catena non cambia nel tempo. In altre parole, quando una CTMC raggiunge lo stato stazionario, la probabilità di trovarsi in ciascuno dei suoi stati rimane costante nel tempo, anche se la CTMC continua a muoversi da uno stato all'altro. Sotto l'assunzione che

la condizione di ergodicità data dalla relazione 5 sia valida, esistono le seguenti probabilità stazionarie degli stati del CTMC $\zeta_t, t \geq 0$:

$$\pi(i,n,\zeta) = \lim_{t\to\infty} P\{i_t = i, n_t = n, \zeta_t = \zeta\},\ i \geq 0, \quad n \in \{0,1,\ldots,\min\{i,L\}\}, \quad \zeta \in \{0,\ldots,n\} \tag{13}$$

La prima formula rappresenta la probabilità che il processo rimanga nello stato (i,n,ξ) per un tempo sufficientemente grande. Qui, i rappresenta lo stato del processo principale, n rappresenta il numero di clienti nel server secondario, e ξ rappresenta lo stato del sistema ausiliario. La probabilità stazionaria $\pi(i,n,\xi)$ è definita come il limite della probabilità che il processo si trovi nello stato (i,n,ξ) al tempo t, quando t tende all'infinito. In altre parole, se il processo rimane in funzione per un tempo sufficientemente grande, la probabilità di trovarlo in uno stato particolare diventa indipendente dal tempo trascorso. Questa probabilità stazionaria rappresenta una distribuzione di equilibrio del processo. In questo caso, visto che la condizione di ergodicità è valida, il processo ha un'unica distribuzione di equilibrio, rappresentata dalle probabilità stazionarie degli stati del processo principale, che sono i vettori riga π_i

Andiamo a definire i vettori riga delle probabilità di stato stazionario π_i come segue: definiamo $\pi(i,n) = (\pi(i,n,1),...,\pi(i,n,m))$ e allora otteniamo

$$\pi_i = (\pi(i,0),...,\pi(i,\min\{i,L\})), \quad i \geq 0 \tag{14}$$

Il vettore riga π_i rappresenta le probabilità stazionarie degli stati del processo principale, sotto l'assunzione che la condizione di ergodicità sia valida. In particolare, π_i è un vettore riga di lunghezza $\min i, L+1$, dove L rappresenta il numero massimo di clienti che possono essere presenti nel server secondario. Ogni elemento del vettore π_i rappresenta la probabilità stazionaria che il processo principale si trovi in uno stato specifico, ovvero con n clienti nel sistema, dove n varia da 0 a $\min i, L$. In altre parole, il vettore π_i descrive la distribuzione di equilibrio dei clienti nel sistema in base allo stato del processo principale. Questa distribuzione di equilibrio è determinata dalle probabilità stazionarie degli stati del processo principale

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

$$(\pi_0, \pi_1, \pi_,\ldots)Q = 0 \qquad (\pi_0, \pi_1, \pi_,\ldots)e = 1 \tag{15}$$

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.

Cosa vuol dire? In particolare, la prima equazione richiede che il prodotto tra il vettore di probabilità stazionarie e la matrice di transizione sia uguale al vettore nullo. Questo significa che, una volta raggiunto l'equilibrio, il processo rimarrà in equilibrio per sempre, poiché il prodotto di una matrice di transizione per un vettore di probabilità stazionarie restituisce un nuovo vettore di probabilità stazionarie. La seconda equazione richiede che la somma degli elementi del vettore di probabilità stazionarie sia uguale a 1, ovvero la somma delle probabilità degli stati deve essere uguale a 1. Questo significa che il vettore di probabilità stazionarie rappresenta una distribuzione di probabilità valida sullo spazio degli stati del CTMC.

Come trovare le soluzioni? Per trovare la soluzione del problema del calcolo della distribuzione di equilibrio in un processo QBD, esistono due casi principali da considerare:

- Quando le transizioni della QBD non dipendono dal livello in cui ci si trova
- Quando le transizioni della QBD dipendono dal livello in cui ci si trova

Nel primo caso, le probabilità stazionarie vengono trovate grazie alla forma matriciale geometrica. In particolare, la matrice di transizione Q del CTMC ζ_t può essere riscritta come $Q = I + P\tilde{Q}$, dove P è la matrice di transizione della catena di Markov di livello n_t e $\tilde{Q}$ è la matrice di transizione della catena di Markov di taglia ξ_t. In questo caso, la distribuzione di equilibrio π_i può essere calcolata come $\pi_i = \pi_0 Q^i$, dove π_0 è il vettore iniziale di probabilità di stato. Nel secondo caso, le probabilità stazionarie dei livelli di confine, in cui le transizioni del QBD dipendono dal livello, vengono direttamente trovate come soluzione del sistema di equazioni lineari algebriche (equazioni di equilibrio) rappresentato da $(\pi_0, \pi_1, \pi_2, \ldots)Q = 0$ e $(\pi_0, \pi_1, \pi_2, \ldots)e = 1$, dove Q è la matrice di transizione del CTMC ζ_t e e è il vettore colonna di tutti gli elementi 1. Tuttavia, se il numero di livelli di confine è elevato (cosa che avviene nel modello considerato se L è grande), questo sistema può diventare molto grande e quindi può essere difficile da risolvere in modo efficiente.

Approccio matriciale-geometrico di Neuts : L'approccio matriciale-geometrico di Neuts è un metodo utilizzato per analizzare i processi di nascita e morte a coda (QBD) mediante l'utilizzo di strumenti matematici e geometrici. Il metodo consiste nel rappresentare il sistema QBD come una successione di sottosistemi di dimensione ridotta, ognuno dei quali rappresenta uno stato di una coda o di una fase del sistema. Ogni sottosistema è rappresentato da una matrice di transizione, che descrive le probabilità di transizione da uno stato del sottosistema a un altro. Le matrici di transizione di ogni sottosistema sono quindi utilizzate per costruire una matrice di transizione globale, che rappresenta l'intero sistema QBD. La matrice globale è quindi utilizzata per calcolare le proprietà del sistema, come la distribuzione stazionaria, il tempo medio di soggiorno e la probabilità di coda. L'approccio matriciale-geometrico di Neuts utilizza inoltre una rappresentazione geometrica del sistema QBD, chiamata "diagramma di regime di equilibrio". Questo diagramma rappresenta il sistema come un insieme di reticoli, dove ogni reticolo rappresenta uno stato della coda o della fase. La dimensione dei reticoli rappresenta la dimensione del sottosistema associato allo stato corrispondente. Il diagramma di regime di equilibrio consente di visualizzare in modo intuitivo la struttura del sistema e di identificare le possibili traiettorie di transizione tra i vari stati del sistema. Inoltre, le proprietà del sistema, come la distribuzione stazionaria e il tempo medio di soggiorno, possono essere calcolate utilizzando metodi geometrici come il teorema di Jackson. L'approccio matriciale-geometrico di Neuts è un metodo potente per l'analisi dei sistemi QBD, in quanto consente di rappresentare il sistema in modo intuitivo e di calcolare le proprietà del sistema in modo efficiente utilizzando strumenti matematici e geometrici.

Stati del QBD In un processo di Markov con matrice di transizione QBD (Quasi-Birth-Death), le transizioni possono essere classificate in due categorie: dipendenti dal livello o indipendenti dal livello. Le transizioni indipendenti dal livello sono quelle che avvengono con la stessa probabilità in ogni stato del processo, senza dipendere dal livello corrente. Queste transizioni sono rappresentate nella matrice di transizione da diagonali e sottodiagonali costanti, che riflettono il fatto che la probabilità di passare da uno stato all'altro non dipende dallo stato corrente. Le transizioni dipendenti dal livello, al contrario, sono quelle che avvengono con probabilità differenti a seconda del livello in cui ci si trova. In altre parole, la probabilità di una transizione dipendente dal livello dipende dallo stato corrente. Queste transizioni sono rappresentate nella matrice di transizione da diagonali e sottodiagonali che variano con il livello, in modo da riflettere la dipendenza della probabilità di transizione dallo stato corrente. In sintesi, quando

una matrice di transizione QBD ha transizioni indipendenti dal livello, la probabilità di transizione da uno stato all'altro non dipende dallo stato corrente. Quando, invece, le transizioni dipendono dal livello, la probabilità di transizione da uno stato all'altro dipende dallo stato corrente e quindi dal livello corrente del processo di Markov.

Livello di confine Un livello di confine è un livello in cui le transizioni dipendono dal livello stesso. In altre parole, le probabilità di transizione da un livello di confine a un altro livello dipendono dal livello di partenza. Un esempio concreto di processo QBD con livelli di confine è un sistema di code in cui gli arrivi sono rappresentati come processi di Poisson e i servizi sono rappresentati come processi deterministici. In questo caso, i livelli di confine rappresentano il numero di clienti nel sistema che corrispondono a un cambio di regime, ad esempio, quando un nuovo server viene aperto o chiuso. Nel nostro caso, viene detto che per i livelli di confine, le probabilità stazionarie non possono essere trovate in modo diretto attraverso la matrice di transizione, ma devono essere calcolate come soluzione di un sistema di equazioni lineari algebriche. In particolare, viene sottolineato che se il numero di livelli di confine è grande, la soluzione di questo sistema può diventare complessa e richiedere l'utilizzo di algoritmi appositi per risolverlo in modo efficiente.

Gli autori presentano quindi un algoritmo che sfrutta la struttura tridiagonale a blocchi, ma dipendente dal livello, del generatore per i livelli inferiori a L+1, in modo da risolvere più efficientemente il sistema di equazioni lineari algebriche quando il numero di livelli di confine è elevato.

Teorema 3.5. *I vettori $\pi_i, i \geq 0$, sono trovati come soluzione del sistema di equazioni algebriche lineari:*

$$\pi_i = \alpha_i \Big(\sum_{l=0}^{\infty} \alpha_l e \Big)^{-1}, \qquad i \geq 0 \tag{16}$$

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

$$\alpha_0 (Q_{0,0} + Q_{0,1} G_0) = 0, \qquad \alpha_0 e = 1 \tag{17}$$

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

$$\alpha_i = \alpha_0 \prod_{l=1}^{i} R_l, \qquad i \geq 1 \tag{18}$$

o tramite la formula ricorsiva

$$\alpha_i = \alpha_{i-1} R_i, \qquad i \geq 1 \tag{19}$$

dove

$$R = \begin{cases} -Q_{i-1,i}(Q_{i,i} + Q_{i,i+1} G_i)^{-1} Q & 1 \leq i \leq L-1 \\ -Q_{L-1,L}(Q_{L,L} + Q^+ G)^{-1} & i = L \\ -Q^+ (Q^0 + Q^+ G)^{-1} = R & i > L \end{cases} \tag{20}$$

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

$$\begin{aligned} G_L &= G \\ G_L - 1 &= -(Q_{L,L} + Q^+ G_L)^{-1} Q_{L,L-1} \\ G_i &= -(Q_{i+1,i+1} + Q_{i+1,i+2} G_{i+1})^{-1} Q_{i+1,i}, \qquad i = L-2, L-3, \ldots, 0 \end{aligned} \tag{21}$$

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

$$Q^+ G^2 + Q^0 G + Q^- = 0 \tag{22}$$

Proof. non fornita □

Questo algoritmo è una modifica efficace dell'algoritmo per il calcolo della distribuzione stazionaria del CTMC asintoticamente quasi-Toeplitz. Utilizzando la ricorsione di vettori come indicato nell'Equazione 19 invece della ricorsione di matrici corrispondente, si ottiene una significativa riduzione della memoria del computer richiesta e del tempo di esecuzione. L'esistenza delle inverse delle matrici (tutte subgeneratori irriducibili) che appaiono nell'algoritmo sopra segue immediatamente dal teorema di O. Tausska. Inoltre, queste matrici sono semi-stabili (e quindi le inverse dei negativi di queste matrici sono non negative), risultando nella produzione di procedure ricorsive stabili nell'implementazione numerica dell'algoritmo.

Matrice stocastica Una matrice stocastica è una matrice quadrata in cui ogni elemento è non negativo e la somma degli elementi di ogni riga è uguale a 1. In altre parole, una matrice stocastica è una matrice di probabilità, in cui ogni elemento rappresenta la probabilità di transizione da uno stato a un altro in un processo stocastico. Ad esempio, in un processo di Markov a tempo discreto, la matrice di transizione è una matrice stocastica, in cui ogni elemento rappresenta la probabilità di transizione da uno stato a un altro nello spazio degli stati del processo. Le matrici stocastiche hanno alcune proprietà importanti, come la conservazione della probabilità, la reversibilità e la convergenza alla distribuzione stazionaria. La conservazione della probabilità implica che la somma degli elementi di ogni riga della matrice è uguale a 1, il che significa che la probabilità totale di transizione da uno stato a un altro è 1. La reversibilità implica che il processo stocastico può essere eseguito in modo indifferente in avanti o all'indietro nel tempo, e la convergenza alla distribuzione stazionaria implica che il processo raggiunge uno stato di equilibrio a lungo termine, in cui la distribuzione di probabilità del processo non cambia nel tempo.

4 Risultati numerici

In questa sezione, forniamo alcuni esempi illustrativi utilizzando cinque diversi processi di arrivo. Di questi cinque, tre sono processi di rinnovo e due sono processi correlati. In particolare, prendiamo i cinque MAP come:

- **ERL**: Questo è un 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**: Questo è un esponenziale con una frequenza di 0.5. Notare che qui abbiamo $\lambda = 0.5, \sigma = 2$ e $\rho_c = 0$.
- **HEX**: Questa è una 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 $\rho_c = 0$.

I due processi correlati, negativo e positivo, sono i seguenti:

- **NCR**: Questo è un MAP negativamente correlato, con matrici di rappresentazione date da: dove

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

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

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

- **PCR**: Questo è un MAP positivamente correlato, con matrici di rappresentazione date da: dove abbiamo $\lambda = 0.5, \sigma = 2.02454$ e $\rho_c = 0.57855$.

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

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

Guardando le cinque MAP sopra riportate, è evidente che sono tutte qualitativamente diverse. È importante sottolineare che il processo di arrivo denominato PCR è ideale per situazioni in cui gli arrivi dei clienti sono altamente irregolari, con periodi alternati di congestione e di scarsità del sistema. Tali arrivi sono comuni nella pratica, specialmente nelle telecomunicazioni e nelle industrie dei servizi. Invece, il processo di arrivo HEX è noto per presentare un comportamento irregolare simile: che gli arrivi con tempi tra di essi più brevi sono separati da tempi più lunghi. Tuttavia, la differenza tra questi due processi sta nella correlazione positiva presente nel processo PCR. Discutiamo tre esempi numerici rappresentativi e illustrativi per evidenziare la natura qualitativa del modello in studio.

4.1 Primo esempio numerico

Qui discutiamo l'impatto del parametro L su alcune misure di performance del sistema selezionate per tutte e cinque le MAP. Innanzitutto, fissiamo $\mu_1 = 1$, $\mu_2 = 0.5$, $q = 0.5$, e $\nu = 0.4$, e variamo L da 1 a 30.

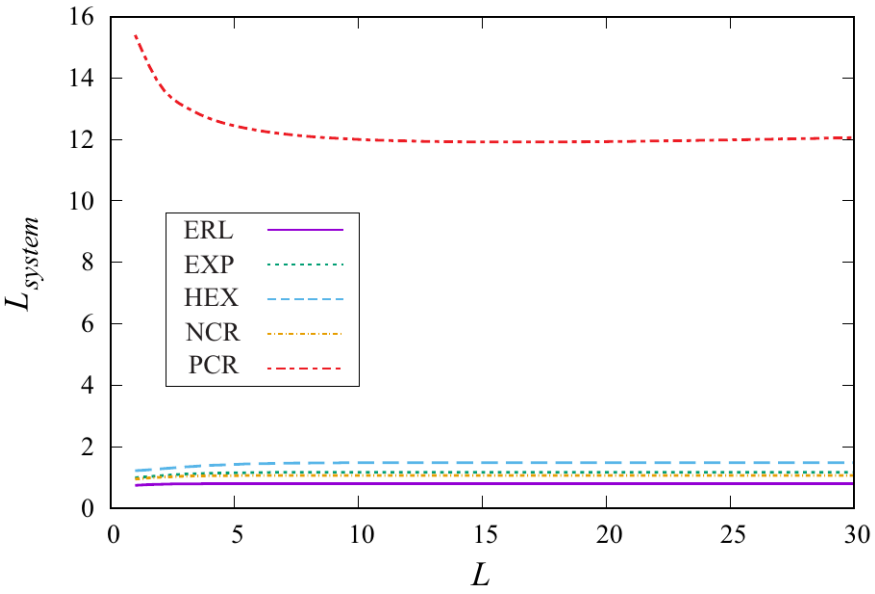

Figure 2: Impact of L on the average number of customers in the system L_{system} for different MAPs.

La Figura 2 illustra chiaramente l'effetto dell'irregolarità nel processo di arrivo PCR. Il numero medio di clienti nel sistema nel caso di PCR è molte volte maggiore rispetto agli altri MAPs. Per i primi quattro MAPs, la misura L_{system} è una funzione non decrescente di L, mentre per PCR si osserva un trend non crescente. Inoltre, un valore elevato di L indica che quando un server secondario viene reclutato, verranno assegnati più clienti e, a causa della lentezza del server secondario (rispetto al server principale), c'è una alta probabilità, soprattutto per i casi dei primi quattro MAPs, che il sistema abbia in media più clienti nel sistema.

Tuttavia, per quanto riguarda gli arrivi PCR, osserviamo un trend interessante ma opposto, ovvero un trend decrescente. Questo può essere intuitivamente spiegato come segue. Innanzitutto, si osserva che il sistema L ha un valore massimo quando $L = 1$, il che può essere spiegato utilizzando il fatto che, quando $L = 1$, i server secondari lasciano il sistema dopo aver servito un solo cliente; con una probabilità del solo 0.5 di essere reclutati, la coda tende ad accumularsi rapidamente. Aumentando L, i server secondari sono maggiormente coinvolti nella pulizia della coda, soprattutto quando gli arrivi avvengono a sprazzi,

e quindi L_{system} diminuisce. Raggiunge un valore minimo quando $L = 16$ e poi inizia ad aumentare a causa della mancata possibilità di essere serviti dal server principale.

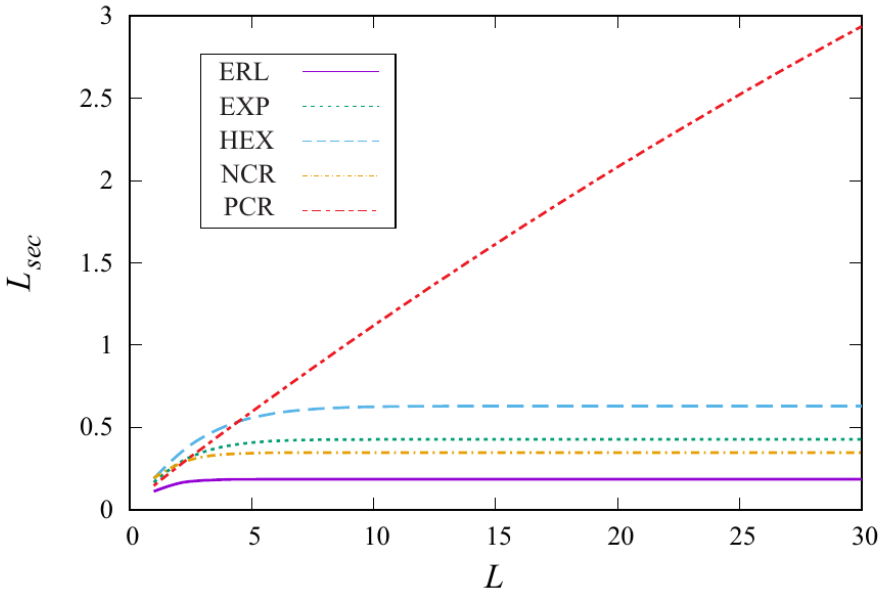

Figure 3: Dependence of the average number of customers with the secondary server L sec on the parameter L for different MAPs.

La Figura 3mostra il comportamento del numero medio di clienti con il server secondario L_{sec}. Come ci si può aspettare, si nota che L_{sec} aumenta quando L aumenta. Come nella figura precedente, il valore di L_{sec} è, in generale, elevato nel caso del PCR. Solo per valori piccoli di L, questo valore è più piccolo per l'ERL-NCR. Ciò può essere spiegato dall'elevata irregolarità degli arrivi osservata nel processo PCR, che provoca la fame del sistema, durante la quale solo il server principale è occupato per la maggior parte del tempo nell'offrire servizi. Tra l'ERL-HEX, si conferma l'effetto noto che una maggiore varianza implica un gran numero di clienti nel sistema.

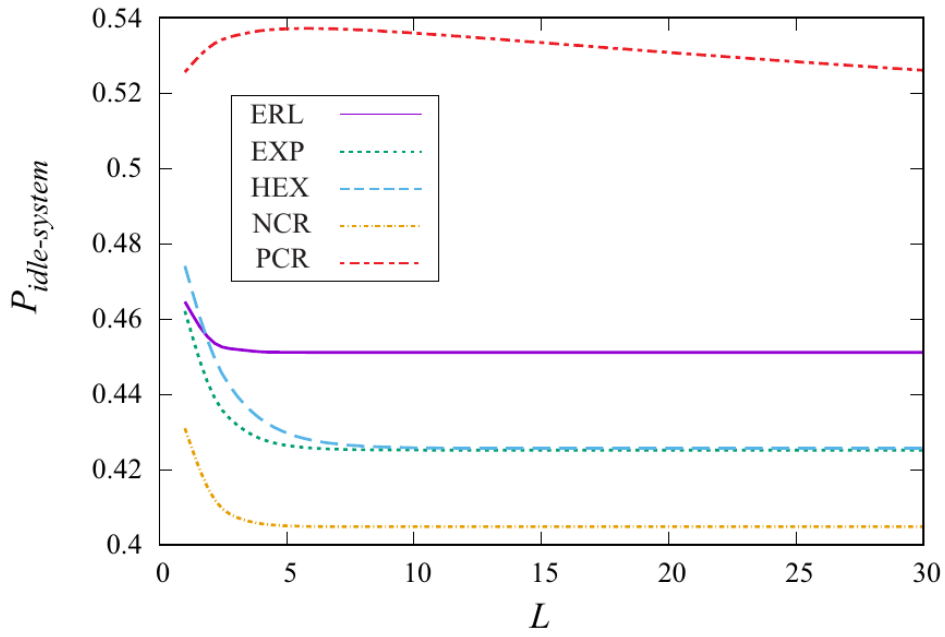

Figure 4: Dependence of the probability $P_{\text{idle-system}}$ that the system is idle at an arbitrary moment on the parameter L for different MAPs

La Figura 4 illustra il comportamento della probabilità, $P_{\text{idle-system}}$, che il sistema sia inattivo in un momento arbitrario. Questa figura corrisponde alla Figura 2 su due aspetti. Il primo è la grande discrepanza nella misura quando viene confrontata tra i vari MAPs. Quando si cerca di trovare un valore ottimale di L, è evidente che conta quale misura viene scelta come funzione obiettivo e il tipo di MAPs utilizzato quando tutti gli altri parametri sono fissati. Ad esempio, se si considera il processo di arrivo PCR, il valore ottimale di L è 16 se si cerca di minimizzare L_{system}. Tuttavia, se la misura $P_{\text{idle-system}}$ è l'obiettivo del problema di ottimizzazione, allora L =6 produce il valore più grande per questa misura.

Le Figure 5 e 6 illustrano il comportamento delle probabilità $P_{\text{idle-busy}}$ e $P_{\text{busy-idle}}$, che corrispondono rispettivamente al momento in cui il server principale è inattivo con il server secondario occupato, e al momento in cui il server principale è occupato con il server secondario inattivo, in un momento arbitrario. Mentre la prima probabilità aumenta all'aumentare di L, la seconda probabilità diminuisce.

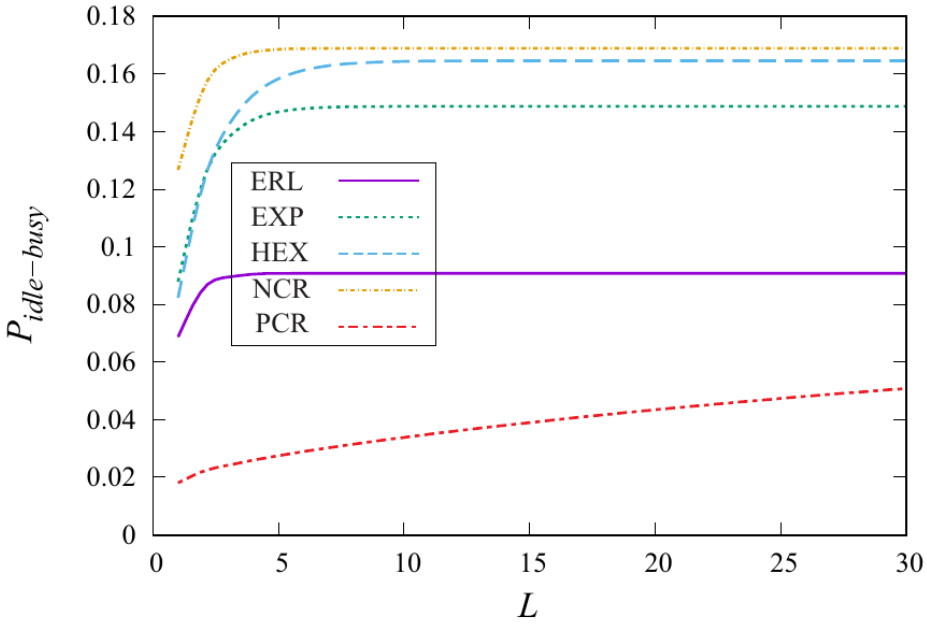

Figure 5: Dependence of the probability $P_{\text{idle}-\text{busy}}$ that the main server is idle while the secondary server is busy on the parameter L for different MAPs

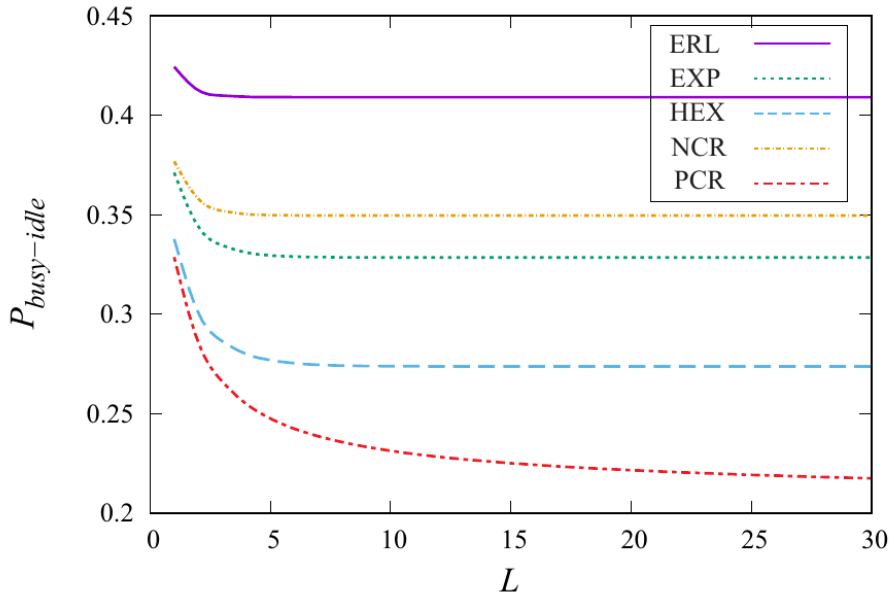

Figure 6: Dependence of the probability $P_{\text{busy}-\text{idle}}$ that the main server is busy while the secondary server is idle on the parameter L for different MAPs.

4.2 Secondo esempio numerico

Lo scopo di questo esempio è di indagare l'impatto dei parametri q (la probabilità che un cliente servito si rifiuti di agire come server secondario) e ν (questa è la probabilità che un cliente servito da un server secondario sia insoddisfatto e torni al sistema). Fissiamo il valore di L a 10 (punto medio tra i due valori ottimali menzionati nel primo esempio). Fissiamo anche i tassi di servizio come $\mu_1 = 1$ e $\mu_2 = 0.5$ e indaghiamo la dipendenza di diverse misure di prestazione dalle probabilità q e ν. Variamo i valori di queste probabilità da 0 a 1 con passo 0.05. Si noti che il valore $q = 1$ corrisponde al classico sistema MAP/M/1 con il tasso di servizio μ_1. .

In questo esempio ci concentriamo sul processo PCR, essendosi dimostrato un caso particolare nel primo esempio illustrativo. Dalla Figura 7, che mostra la dipendenza del numero medio di clienti nel sistema L_{system} dai parametri q e ν, deduciamo diverse osservazioni interessanti.

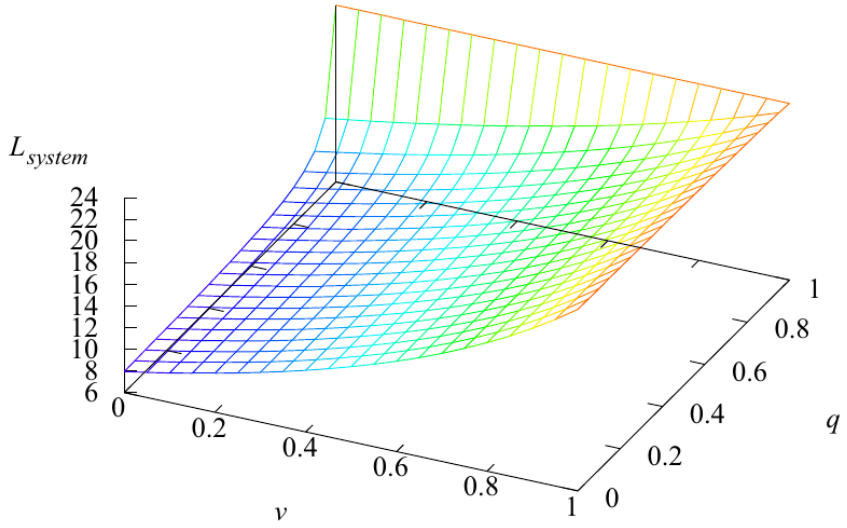

Figure 7: Dependence of the average number of customers in the system L_{system} on the parameters q and ν.

Il valore di L_{system} è minimo quando il cliente servito è sempre disponibile per essere reclutato (quando il sistema ne ha bisogno) e quando il cliente che riceve il servizio da un server secondario è sempre soddisfatto. Il valore minimo si ottiene quando $q = 0$ e $\nu = 0$. Questa misura aumenta quando aumenta q o ν, e il tasso di aumento diventa più elevato quando uno o entrambi si avvicinano al valore 1. Quando $q = 1$, il sistema si trasforma nel corrispondente modello di coda MAP/M/1 classico e in un sistema senza l'uso del server secondario, e $L_{\text{system}} \sim 22$ per tutti i valori di ν (come è evidente). Quando $q = 0$, che corrisponde al caso in cui un cliente servito viene sempre reclutato (quando necessario), anche quando la probabilità di insoddisfazione è alta ($\nu = 0,5$), il valore di L_{system} è circa 13. Pertanto, l'uso di un server secondario riduce essenzialmente il numero medio di clienti nel sistema di più del 40%. Inoltre, gli autori mostrano che esiste un valore di ν tale per cui il modello classico di coda è migliore del modello proposto qui. Questo valore è molto alto, arrivando alla conclusione che la percentuale di insoddisfazione deve essere superiore al 98,5% affinché il modello classico funzioni meglio.

Per testare ulteriormente l'importo della riduzione nel numero medio, aumentano λ del 50% a $\lambda = 0.75$. Mantenendo tutti gli altri parametri (ad eccezione della normalizzazione dei parametri del processo di arrivo per ottenere questo specifico λ) gli stessi, ottenendo una percentuale di riduzione superi-

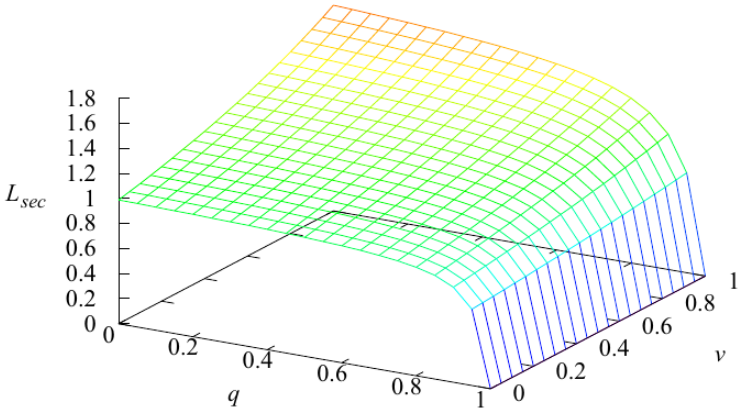

Figure 8: Dependence of the average number of customers with the secondary server L_{sec} on the parameters q and v

ore al 52,8%. Pertanto, un aumento del carico del sistema beneficerà notevolmente dell'avere un server secondario per aiutare il sistema anche con un tasso di insoddisfazione del cliente del 50% con questo server secondario.

La figura 8 mostra invece la dipendenza del numero medio di clienti con il server secondario L_{sec} dai parametri q e v. Questa probabilità 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, ovvero tutti i clienti vengono reclutati (quando necessario) per diventare server secondari, e quando v è vicino a 1. Ovviamente, in quest'ultimo caso, quasi tutti i clienti serviti da un server secondario devono essere rimandati al sistema a causa della loro insoddisfazione.

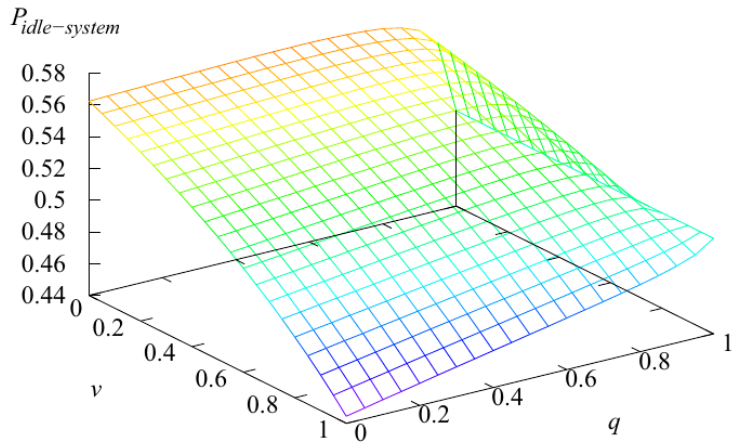

Figure 9: Dependence of the probability $P_{\text{idle-system}}$ that the system is idle at an arbitrary moment on the parameters q and v.

Nella Figura 9, è mostrato il comportamento della probabilità $P_{\text{idle-system}}$ che il sistema è inattivo in un momento arbitrario in funzione di q e v. Questa probabilità ha valore minimo quando $v = 1$ e $q = 0$, il che è intuitivamente chiaro, poiché dover servire nuovamente i clienti dopo aver passato attraverso un server secondario mette un carico sul sistema. La probabilità che $P_{\text{idle-system}}$ aumenta quando q

aumenta e/o ν diminuisce: il valore massimo di questa probabilità si ottiene quando $q = 0.65$ e $\nu = 0$.

ALTRE COSE DA DIRE SU QUESTO ESEMPIO NUMERICO CHE SECONDO ME SI POSSONO SALTARE. TORNARE DOPO

4.3 Esempio numerico 3

In questo ultimo esempio, analizziamo l'impatto della variazione dei tassi di servizio μ_1 e μ_2 quando tutti gli altri parametri sono fissati. A tal fine, fissiamo $L = 10$, $q = 0.5$, $\nu = 0.4$ e $\lambda = 0.5$. I tassi μ_1 e μ_2 vengono variati da 0.25 a 2.0 con incrementi di 0.05. È importante menzionare che, per soddisfare la condizione di ergodicità (vedi Equazione 5), limitiamo ulteriormente il valore di μ_2 quando μ_1 è piccolo.

Nelle Figure 10 e 11 viene evidenziata la dipendenza della misura L_{system} da μ_1 e μ_2. Nella Figura 10, gran parte della superficie che mostra la dipendenza appare piatta. Ciò è dovuto al fatto che, per molte combinazioni dei valori dei parametri con un piccolo tasso μ_1, la condizione di ergodicità viene violata e la misura L_{system} diventa molto grande. Invece, se andiamo a zoommare, nella Figura 11, la dipendenza di L_{system} da μ_1 e μ_2 è rappresentata solo per valori non piccoli di μ_1. Chiaramente, si può notare una tendenza decrescente, poiché L_{system} diminuisce rapidamente quando μ_1 aumenta per μ_2 fissato e viceversa.

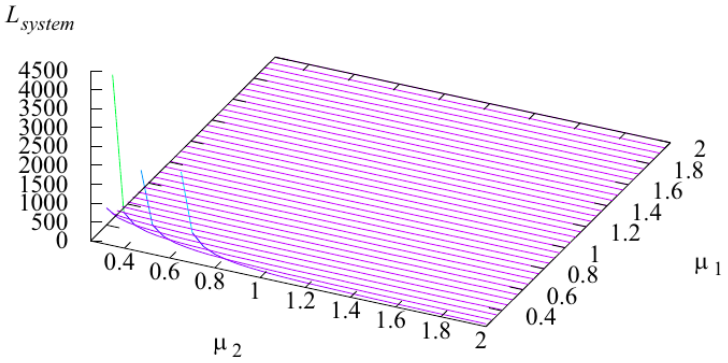

Figure 10: Dependence of the average number of customers in the system L_{system} on the parameters μ_1 and μ_2

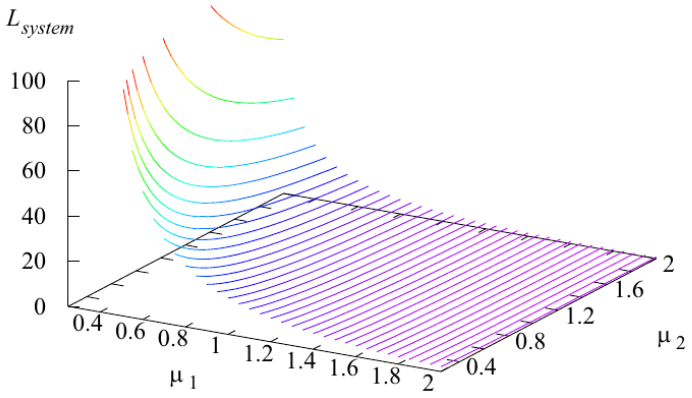

Figure 11: Dependence of the average number of customers in the system L_{system} on the parameters μ_1 and μ_2.

La Figura 12 mostra il comportamento del numero medio di clienti con il server secondario L_{sec}. Il valore di L_{sec} è massimo quando μ_1 e μ_2 sono piccoli. Questo è intuitivamente chiaro poiché per valori piccoli di μ_1 e μ_2: la condizione di ergodicità è vicina a essere violata, causando un alto tasso di reclutamento per i server secondari che, molto probabilmente prima di lasciare il sistema, serviranno

un gruppo di dimensione $L = 10$. 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.

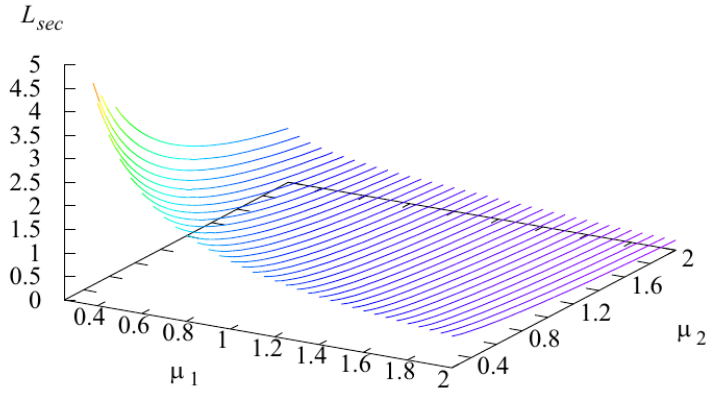

Figure 12: Dependence of the average number of customers with the secondary server L sec on the parameters μ_1 and μ_2.

ALTRE COSE DA DIRE SU QUESTO ESEMPIO NUMERICO CHE SECONDO ME SI POSSONO SALTARE. TORNARE DOPO

5 Conclusioni

In questo articolo, è stato analizzato un sistema di coda in cui c'è la possibilità di reclutare un cliente già servito come server secondario per aiutare il server principale assegnando un gruppo di clienti in attesa. Il processo di arrivo è stato modellizzato utilizzando un processo di punto Markoviano versatile, MAP. È stata presa in considerazione la possibilità di insoddisfazione dei clienti con il servizio fornito dal server secondario, causando il ritorno di quei clienti nel sistema. È stata implementata l'analisi dello stato stazionario della catena di Markov multidimensionale che descrive il comportamento del sistema e sono stati presentati risultati numerici illustrativi potenzialmente utili per prendere decisioni manageriali. Il modello studiato in questo articolo può essere generalizzato in diversi modi. Ad esempio,

1. il servizio fornito dal server secondario può essere effettuato in gruppi;
2. rilassare l'ipotesi di avere solo un server secondario a più di uno e vedere l'impatto dell'aumento a, diciamo, 2;
3. utilizzare servizi di tipo fase-possibilmente con rappresentazioni diverse per il server principale e secondario;
4. incorporare l'impazienza dei clienti sia nei buffer principali che secondari;
5. implementare un processo di reclutamento in base alla lunghezza della coda osservata basato su una politica di controllo di tipo soglia;
6. consentire arrivi di gruppo; e infine
7. incorporare la possibilità di reclutare molti server secondari con due tipi di clienti in modo che solo un tipo possa agire come server secondario.