Tạp chí Khoa học Đại học Huế: Kỹ thuật và Công nghệ; pISSN 2588-1175 | eISSN 2615-9732

Tập 131, Số 2A, 2022, Tr. 21–37; DOI: 10.26459/hueunijtt.v131i2A.6617

MÔ HÌNH HÀNG ĐỢI RETRIAL TRONG HỆ THỐNG TRUNG TÂM CUỘC GỌI VỚI YẾU TỐ CÂN BẰNG TRỘN CUỘC GỌI

Đặng Thanh Chương[1], Hoàng Đình Long[2], Hoa Lý Cương[1], Nguyễn Đặng Duy Trinh[3], Nguyễn Thanh Sơn[1]

[1] Trường Đại học Khoa học, Đại học Huế, 77 Nguyễn Huệ, Huế, Việt Nam
[2] Trường Đại học Sư phạm, Đại học Huế, 34 Lê Lợi, Huế, Việt Nam
[3] Trường Cao đẳng Y tế Huế, 1 Nguyễn Trường Tộ, Huế, Việt Nam

Tóm tắt. Trong các hệ thống trung tâm cuộc gọi, sự trộn lẫn lộn các cuộc gọi (call blending) là sự pha trộn các hoạt động của các cuộc gọi đến và các cuộc gọi đi ra. Mô hình hàng đợi retrial có thể được sử dụng nhằm phân tích các hoạt động của hệ thống trung tâm cuộc gọi với sự kết hợp của các cuộc gọi đến và các cuộc gọi đi. Các cuộc gọi đến khi chưa thể được phục vụ sẽ được lưu trữ trong bộ đệm (gọi là orbit), nơi mà các cuộc gọi này sau đó sẽ được gửi đến lại máy chủ sau một khoảng thời gian ngẫu nhiên. Ngay khi máy chủ chuyển sang chế độ rỗi, nó sẽ thực hiện cuộc gọi đi. Mô hình phân tích trong bài báo này sử dụng một hệ thống hàng đợi retrial với kích thước orbit giới hạn cho hệ thống trung tâm cuộc gọi với truyền thông hai chiều (mối tương quan giữa các cuộc gọi đến và cuộc gọi đi) và tốc độ retrial không thay đổi trong cả hai trường hợp đơn và đa máy chủ. Bài báo sử dụng phương pháp phân tích quá trình giả sinh tử theo ma trận sinh Q để tính các xác suất trạng thái cân bằng đối với mô hình phân tích để đạt được kết quả mong muốn.

Từ khóa: hàng đợi retrial, hệ thống trung tâm cuộc gọi, quá trình giả sinh tử

Retrial queueing analysis model in call center system with balanced mixing calls

Dang Thanh Chuong[1], Hoang Dinh Long[2], Hoa Ly Cuong[1], Nguyen Dang Duy Trinh[3], Nguyen Thanh Son[1]

[1] University of Sciences, Hue University, 77 Nguyen Hue St., Hue, Vietnam
[2] University of Education, Hue University, 34 Le Loi St., Hue, Vietnam
[3] Hue Medical College, 1 Nguyen Truong To St., Hue, Vietnam

Abstract. In call center systems, call blending is the mixing of activities of incoming calls and outgoing calls. This study presents a retrial queue model with a constant retrial rate for incoming calls to analyze the call center system with a combination of call activities of

** Liên hệ: dtchuong@hueuni.edu.vn*

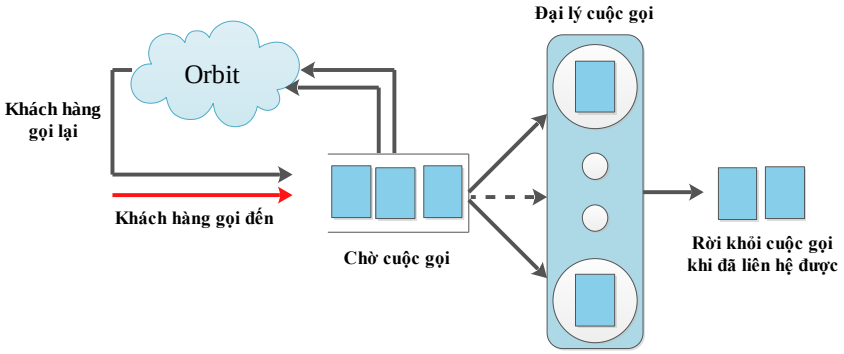
Nhận bài: 25–11–2021; Ngày nhận đăng: 21–02–2022

incoming calls and outgoing calls. Incoming calls that cannot be serviced are stored in a buffer, where the tasks are sent back to the server after a random time. As soon as the server goes idle, it makes an outgoing call. The analyzed model was used as a retrial queue system for a call center system with 2-way communication (correlation between incoming and outgoing calls) and a constant retrial rate in both single and multi-server scenarios. The quasi-birth-and-death process according to the generation matrix Q was used to calculate the equilibrium probabilities for the analytical model to achieve the desired results.

Keywords: retrial queueing, call center, quasi-birth-and-death process

1 Giới thiệu

Hàng đợi retrial đã được tập trung nghiên cứu trong nhiều năm gần đây. Nó cung cấp mô hình để đánh giá hiệu năng của các trung tâm cuộc gọi (call center), mạng máy tính và các hệ thống truyền thông. Đặc tính quan trọng của hàng đợi retrial là các cuộc gọi, mà chưa thể được phục vụ tại thời điểm đến hệ thống sẽ được đưa vào một vùng đệm, gọi là orbit và sẽ yêu cầu phục vụ lại (retrial) sau một vài đơn vị thời gian ngẫu nhiên. Vì vậy, việc phân tích đối với hàng đợi retrial thường là phức tạp và khó khăn hơn so với mô hình tương tự mà không xét yếu tố retrial và thường chỉ có thể thu được kết quả rõ ràng trong một số trường hợp đặc biệt [1–9].



Hình 1. Mô hình hàng đợi retrial cho hệ thống trung tâm cuộc gọi

Một trung tâm cuộc gọi có vai trò quan trọng đối với một công ty vì nó cung cấp một kênh để khách hàng liên hệ với công ty. Trong một cuộc gọi trung tâm, các đại lý cuộc gọi (còn được gọi là các tổng đài viên) là những người trả lời các cuộc gọi từ khách hàng. Khi một khách hàng thực hiện một cuộc gọi điện thoại, nếu có một tổng đài viên cuộc gọi ở trạng thái rỗi, khách hàng ngay lập tức được các đại lý cuộc gọi trả lời. Nếu tất cả các tổng đài viên đang ở trạng thái bận, khách hàng có thể nghe phản hồi từ hệ thống là hiện tại các tổng đài viên cuộc gọi đang bận, vui lòng chờ trong giây lát. Tại thời điểm này, khách hàng có thể cúp điện thoại ngay lập tức hoặc

tiếp tục giữ máy và chờ (trong orbit). Trong trường hợp lần trước chưa liên hệ được, khách hàng có thể thử lại lần tiếp theo sau một khoảng thời gian ngẫu nhiên (quay lại yêu cầu phục từ orbit). Khách hàng chờ đợi một tổng đài viên cuộc gọi ở trạng thái rỗi để có thể liên hệ. Những khách hàng này cũng có thể gọi điện sau nếu không đủ kiên nhẫn chờ đợi.

Nội dung của bài báo là tìm hiểu mô hình hàng đợi retrial áp dụng cho hệ thống cuộc gọi trung tâm (Hình 1). Một đặc điểm của cuộc gọi trung tâm là thông thường nó có thể chỉ xử lý với lưu lượng các cuộc gọi đến (inbound traffic) hoặc chỉ với lưu lượng các cuộc gọi đi (outbound traffic). Các cuộc gọi trung tâm, khi đó, sẽ được gán nhãn một cách tương ứng là inbound và outbound. Một số mô hình hàng đợi retrial với cuộc gọi trung tâm trong các tài liệu phân tích trước đây là một chiều (gọi đến hoặc gọi đi) [3, 4]. Tuy nhiên, thông thường, thì hệ thống cuộc gọi trung tâm sẽ xử lý đồng thời các cuộc gọi đến và cuộc gọi đi [1, 2, 5–9]. Theo đó, các cuộc gọi đến sẽ được chỉ định đến một tổng đài viên (operator) bởi một bộ phân phối cuộc gọi tự động (ACD - automatic call distributor). Đối với các cuộc gọi đi, cuộc gọi có thể được khởi tạo bởi ACD (một cách tự động) hoặc gởi các tổng đài viên (thủ công).

Nguyên tắc kết hợp các cuộc gọi (call blending) với truyền thông hai chiều, cho phép phục vụ hai mục đích sau:

Thứ nhất, nó có thể được thêm vào các nhiệm vụ thông thường. Điều này tương ứng với một thói quen phổ biến trong các trung tâm cuộc gọi thuộc kiểu inbound, nơi các tổng đài viên có thể sử dụng thời gian nhàn rỗi của họ để thực hiện các cuộc gọi đi thứ cấp, có thể không khẩn cấp. Ở đây, việc kết hợp (pha trộn) các cuộc gọi phục vụ nhằm tăng hiệu suất tổng thể bằng cách tăng hiệu suất làm việc của các tổng đài viên có khả năng thông qua chính sách kiểm soát [5–7].

Thứ hai, nó có thể xảy ra như một phần không thể thiếu của các tác vụ được thực hiện tại trung tâm cuộc gọi. Trong trường hợp này, các cuộc gọi đến cũng như các cuộc gọi đi là các yếu tố quan trọng của dịch vụ được cung cấp và cả hai cần được thực hiện một cách thường xuyên. Điều này xảy ra khi các nhiệm vụ yêu cầu các cuộc gọi tiếp theo theo cả hai hướng.

Cả hai trường hợp trên có thể được mô hình hóa với hàng đợi retrial hỗ trợ truyền thông hai chiều. Đã có một vài một công trình liên quan đã được đề xuất. Phung Duc trong [5, 6] giả định một mô hình có tốc độ retrial cổ điển cho các cuộc gọi đến. Một lựa chọn như vậy dẫn đến một mô hình phù hợp với trường hợp thứ nhất. Tuy nhiên, lưu ý rằng, ở quy mô thời gian ngắn, hoạt động của các cuộc gọi đi có thể bị tắc nghẽn với xác suất lớn trong các giai đoạn lưu lượng các cuộc gọi đến tăng cao, như khi tốc độ retrial tăng tuyến tính với số lượng cuộc gọi đến trong orbit. Một cách tương ứng, việc cân bằng sự kết hợp trong thời gian ngắn sẽ bị bị ảnh hưởng nặng nề bởi sự biến đổi của tải lưu lượng các cuộc gọi đến. Ngược lại, trong trường hợp thứ hai, hoạt động cuộc gọi đi vẫn tiếp tục thường xuyên ngay cả khi có nhiều cuộc gọi đến đang chờ được

phục vụ. Bằng cách giả sử tốc độ retrial là không đổi, các cuộc gọi đi vẫn được khởi tạo trong khoảng thời gian ngắn (bằng ACD hoặc bởi các tổng đài viên), ngay cả khi số lượng cuộc gọi đến trong orbit là cao. Một cách tương ứng, việc cân bằng sự kết hợp trong thời gian ngắn sẽ ít chịu ảnh hưởng của sự biến đổi của tải lưu lượng các cuộc gọi đến.

Ý tưởng của mô hình trong bài báo này phân tích sự kết hợp hàng đợi retrial cho hệ thống trung tâm cuộc gọi với truyền thông hai chiều (mối tương quan giữa các cuộc gọi đến và cuộc gọi đi) và tốc độ retrial không thay đổi trong cả hai trường hợp đơn và đa máy chủ [1]. Cuộc gọi đến có thể tìm máy chủ rỗi (tổng đài viên) để được nhận phục vụ ngay lập tức. Ngay khi máy chủ chuyển sang chế độ rỗi, nó sẽ thực hiện cuộc gọi đi. Điểm khác biệt của mô hình trong bài báo này với mô hình trong [1] là chúng tôi xét với kích thước orbit giới hạn, từ đó chúng tôi đưa ra phương pháp nhằm phân tích quá trình giả sinh tử theo ma trận sinh Q để tính các xác suất trạng thái cân bằng đối với mô hình phân tích để đạt được kết quả mong muốn [10]. Nội dung tiếp theo của bài báo gồm: phần 2 giới thiệu các mô hình chúng tôi phân tích với truyền thông hai chiều. Kết quả phân tích thông qua các đồ thị về những thay đổi của xác suất tắc nghẽn chuyển biến theo mật độ luồng được trình bày ở phần 3. Cuối cùng là phần kết luận.

2 Mô hình hàng đợi retrial cho hệ thống trung tâm cuộc gọi với truyền thông hai chiều

2.1 Một số thông số giả thiết của mô hình

Mô hình phân tích ở đây dựa trên một số giả thiết sau:

Các cuộc gọi đến (ban đầu) máy chủ (hoặc tổng đài viên) yêu cầu được phục vụ theo quá trình Poisson với tốc độ đến trung bình là λ. Cuộc gọi đến có thể tìm máy chủ rỗi để được nhận phục vụ ngay lập tức. Trong trường hợp máy chủ bận, cuộc gọi sẽ được đưa vào orbit. Trong orbit, chính sách tốc độ retrial không thay đổi (hằng số) sẽ được áp dụng. Tức là, tốc độ của cuộc gọi đến từ orbit (tốc độ retrial) là $\mu(1-\delta_{0,n})$ với n là số khách hàng trong orbit; ở đây $\delta_{i,j}$ (Kronecker delta) là hàm hai biến các số nguyên không âm, được xác định như sau [1]:

$$\delta_{i,j} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Điều này trái ngược với trường hợp được phân tích trong một số tài liệu trước đây [3, 4], với tốc độ retrial cổ điển là $n\mu$, phụ thuộc vào số lượng khách hàng trong orbit, là n. Như đã đề cập ở trên, tốc độ retrial không đổi (constant retrial rate) xảy ra khi khách hàng trong orbit được tổ chức theo hàng đợi kiểu FCFS và chỉ khách hàng ở đầu hàng đợi mới có thể yêu cầu dịch vụ.

Ngoài ra, khi máy chủ chuyển sang chế độ rỗi, nó sẽ thực hiện cuộc gọi đi theo phân phối mũ với tốc độ là α.

Thời gian phục vụ của các cuộc gọi đến và cuộc gọi đi cũng theo phân phối mũ với tốc độ phục vụ là ν_1 và ν_2.

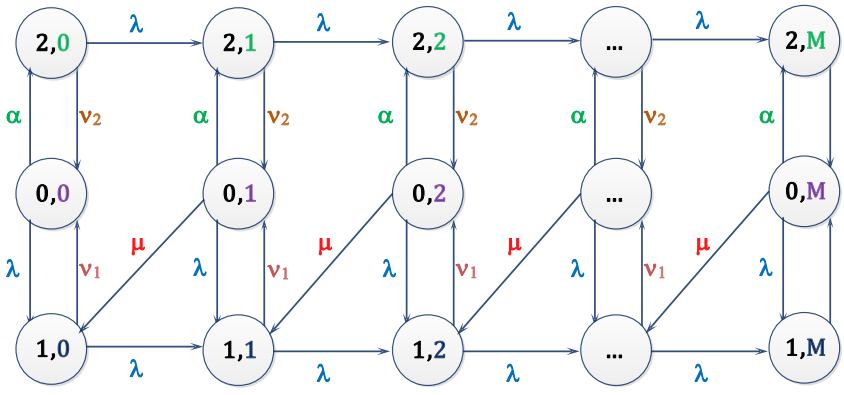
2.2 Mô hình phân tích với đơn máy chủ

Mô hình phân tích ở đây tương tự mô hình trong [1] nhưng với một điểm khác biệt. Trong mô hình này của bài báo, chúng tôi xét với trường hợp *orbit* có kích thước giới hạn ($0 \le N(t) \le M$).

Đặt $S(t)$ ($t \ge 0$), xác định trạng thái của máy chủ tại thời điểm t. Theo đó, có thể định nghĩa trạng thái của máy chủ theo các giá trị như sau 1:

$$S(t) = \begin{cases} 0, & \text{máy chủ ở trạng thái rỗi,} \\ 1, & \text{máy chủ đang phục vụ một cuộc gọi đến,} \\ 2, & \text{máy chủ đang phục vụ một cuộc gọi đi.} \end{cases}$$

Đặt $N(t)$ ($t \ge 0$) là số cuộc gọi của khách hàng trong orbit tại thời điểm $t \ge 0$. Mô hình khi đó được mô hình hóa như là mô hình Markov hai chiều: $\{X(t) = (S(t), N(t)), t \ge 0\}$, với không gian trạng thái có dạng: $\mathcal{S} = \{0, 1, 2\} \times \{0, 1, 2, \dots, \mathrm{M}\}$. Lược đồ chuyển trạng thái cụ thể được trình bày trên Hình 2. Giả thiết hệ thống là ở trạng thái ổn định, tức là luôn tồn tại các xác suất trạng thái cân bằng. Điều kiện cần và đủ để mô hình có trạng thái ổn định là $\lambda < \nu_1$ sẽ được sử dụng trong các phân tích sau.



Hình 2. Lược đồ chuyển trạng thái của mô hình đơn máy chủ

Đặt $\pi_{i,j} = P[S(t) = i, N(t) = j]$, là các xác suất trạng thái cân bằng của hệ thống tại các trạng thái (i, j). Khi đó, hệ phương trình trạng thái cân bằng được xây dựng như sau [1]:

$$\left(\lambda + \alpha + \mu\left(1 - \delta_{0,\mathrm{j}}\right)\right)\pi_{0,j} = \nu_1\pi_{1,j} + \nu_2\pi_{2,j} \quad (1)$$

$$(\lambda + \nu_1)\pi_{1,j} = \lambda\pi_{0,j} + \mu\pi_{0,j+1} + \lambda\pi_{1,j-1} \quad (2)$$

$$(\lambda + \nu_2)\pi_{2,j} = \alpha\pi_{0,j} + \lambda\pi_{2,j-1} \quad (3)$$

với $\delta_{0,\mathrm{j}} = 1$ khi $\mathrm{j} = 0$ và bằng 0 trong trường hợp ngược lại; $\pi_{\mathrm{i},-1} = 0$ với $\mathrm{i} \in \{1, 2\}$ [1]. Việc tính các các giá trị $\pi_{i,j}$ sẽ được chúng tôi thực hiện bằng cách giải các phương trình (1)–(3) bằng phương pháp phân tích quá trình giả sinh tử theo ma trận sinh Q [5]. Kết quả phân tích sẽ được trình bày trong phần tiếp theo.

Trong bài báo này, khác với trong [1], chúng tôi giả thiết số cuộc gọi (khách hàng) trong orbit là giới hạn (bằng M). Vì vậy, mô hình có thể được phân tích theo quá trình giả sinh tử với các ma trận chuyển trạng thái tổng quát và được mô tả trên Hình 2. Đây cũng chính là điểm khác biệt của mô hình trong bài báo này với mô hình trong [1].

Dựa vào ma trận sinh Q của quá trình giả sinh tử QBD, các ma trận chuyển trạng thái A_j, B_j và C_j (mỗi ma trận đều có kích thước 3×3) mô tả các bước chuyển trạng thái ứng với lược đồ trên Hình 1 như sau [10]:

(a). $A_j(i,k)$: Là việc chuyển từ trạng thái (i,j) tới trạng thái (k,j) (với $0 \le j \le M; 0 \le i,k \le 2$) do máy chủ rỗi hoặc đang phục vụ một cuộc gọi đến hoặc đang phục vụ một cuộc gọi đi. Ma trận A_j có kích thước 3×3 với các phần tử $A_j(i,k)$.

$$A_j = \begin{pmatrix} 0 & \lambda & \alpha \\ \nu_1 & 0 & 0 \\ \nu_2 & 0 & 0 \end{pmatrix}, (1 \le j \le M).$$

(b). $B_j(i,k)$: Biểu thị cho một bước nhảy (lên) từ trạng thái (i,j) tới trạng thái $(k,j+1)$ (với $0 \le j \le M-1; 0 \le i,k \le 2$) do một yêu cầu đến từ cuộc gọi đến (cuộc gọi đến ban đầu hoặc cuộc gọi quay lại từ orbit), nhưng máy chủ đang bận (phục vụ cuộc gọi đến hoặc cuộc gọi đi). Ma trận B_j hay B (do j là mức độc lập) có kích thước 3×3 với các phần tử $B_j(i,k)$.

$$B_j = B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, (0 \le j \le M-1).$$

(c). $C_j(i,k)$*: Biểu thị cho một bước nhảy (xuống) từ trạng thái* (i,j) *tới trạng thái* $(k,j-1)$ *(*$1 \le j \le M; 0 \le i,k \le 2$*) do có cuộc gọi quay lại từ orbit sau một khoảng thời gian và được phục vụ do máy chủ đang rỗi. Ma trận* C_k *có kích thước* 3×3 *với các phần tử* $C_j(i,k)$*.*

$$C_j = \begin{pmatrix} 0 & \mu & 0 \\ 0 & 0 & 0 \\ 0 & \nu_2 & 0 \end{pmatrix}, (1 \le j \le M).$$

Khi đó, ma trận sinh Q của quá trình giả sinh tử QBD được biểu diễn qua ma trận chuyển trạng thái con ở trên như sau:

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{A_0} & \boldsymbol{B_0} & & & \\ \hline \boldsymbol{C_1} & \boldsymbol{A_1} & \boldsymbol{B_1} & & \\ \hline & \boldsymbol{C_1} & \boldsymbol{A_2} & \boldsymbol{\dots} & \\ \hline & & \boldsymbol{\dots} & \boldsymbol{\dots} & \boldsymbol{B_j} \\ \hline & & & \boldsymbol{C_j} & \boldsymbol{A_j} \\ \hline \end{array}$$

Với các giá trị trên đường chéo ma trận Q:

$Q(i,i) = -(tổng\ các\ phần\ tử\ trên\ dòng\ i)$, được tính như sau:

$$Q(i,i) = -\sum_{i,j=0;j\neq i}^{n_s-1} Q_{i,j} \qquad \text{tương ứng với } \sum_{i,j=0}^{n_s-1} Q_{ij} = 0$$

trong đó $n_s = 3 * (M+1)$.

Với $M = 4, (i = 0,1,2; j = 0,1,2,3,4)$, ma trận Q có dạng như sau:

(i,j)	0,0	1,0	2,0	0,1	1,1	2,1	0,2	1,2	2,2	0,3	1,3	2,3	0,4	1,4	2,4
0,0	(1)	$\boldsymbol{\lambda}$	$\boldsymbol{\alpha}$	**0**	**0**	**0**	0	0	0	0	0	0	0	0	0
1,0	$\boldsymbol{\nu_1}$	(2)	**0**	**0**	$\boldsymbol{\lambda}$	**0**	0	0	0	0	0	0	0	0	0
2,0	$\boldsymbol{\nu_2}$	**0**	(3)	**0**	**0**	$\boldsymbol{\lambda}$	0	0	0	0	0	0	0	0	0
0,1	**0**	$\boldsymbol{\mu}$	**0**	(4)	$\boldsymbol{\lambda}$	$\boldsymbol{\alpha}$	**0**	**0**	**0**	0	0	0	0	0	0
1,1	**0**	**0**	**0**	$\boldsymbol{\nu_1}$	(5)	**0**	**0**	$\boldsymbol{\lambda}$	**0**	0	0	0	0	0	0
2,1	**0**	**0**	**0**	$\boldsymbol{\nu_2}$	0	(6)	**0**	**0**	$\boldsymbol{\lambda}$	0	0	0	0	0	0
0,2	0	0	0	**0**	$\boldsymbol{\mu}$	**0**	(7)	$\boldsymbol{\lambda}$	$\boldsymbol{\alpha}$	**0**	**0**	**0**	0	0	0
1,2	0	0	0	**0**	**0**	**0**	$\boldsymbol{\nu_1}$	(8)	**0**	**0**	$\boldsymbol{\lambda}$	**0**	0	0	0

2,2	0	0	0	**0**	**0**	**0**	$\boldsymbol{\nu_2}$	0	(9)	**0**	**0**	$\boldsymbol{\lambda}$	0	0	0
0,3	0	0	0	0	0	0	**0**	$\boldsymbol{\mu}$	**0**	(10)	$\boldsymbol{\lambda}$	$\boldsymbol{\alpha}$	**0**	**0**	**0**
1,3	0	0	0	0	0	0	**0**	**0**	**0**	$\boldsymbol{\nu_1}$	(11)	**0**	**0**	$\boldsymbol{\lambda}$	**0**
2,3	0	0	0	0	0	0	**0**	**0**	**0**	$\boldsymbol{\nu_2}$	**0**	(12)	**0**	**0**	$\boldsymbol{\lambda}$
0,4	0	0	0	0	0	0	0	0	0	**0**	$\boldsymbol{\mu}$	**0**	(13)	$\boldsymbol{\lambda}$	$\boldsymbol{\alpha}$
1,4	0	0	0	0	0	0	0	0	0	**0**	**0**	**0**	$\boldsymbol{\nu_1}$	(14)	**0**
2,4	0	0	0	0	0	0	0	0	0	**0**	**0**	**0**	$\boldsymbol{\nu_2}$	**0**	(15)

Giá trị của các ô trên đường chéo của ma trận Q lần lượt được xác định như sau:

(1) $-(\lambda + \alpha)$

(2) $-(\lambda + \nu_1)$

(3) $-(\lambda + \nu_2)$

(4) $-(\lambda + \alpha + \mu)$

(5) $-(\lambda + \nu_1)$

(6) $-(\lambda + \nu_2)$

(7) $-(\lambda + \alpha + \mu)$

(8) $-(\lambda + \nu_1)$

(9) $-(\lambda + \nu_2)$

(10) $-(\lambda + \alpha + \mu)$

(11) $-(\lambda + \nu_1)$

(12) $-(\lambda + \nu_2)$

(13) $-(\lambda + \alpha + \mu)$

(14) $-\nu_1$

(15) $-\nu_2$

2.3 Mô hình phân tích đa máy chủ

Mô hình ở đây sẽ mở rộng mô hình đơn máy chủ ở trên với trường hợp mô hình đa máy chủ với kích thước orbit giới hạn. Theo đó, mô hình phân tích ở đây sẽ được mô hình hóa theo mô hình hàng đợi retrial $M/M/c/\mathrm{c}+M$ $(c>1, M>1)$ với các tham số tương tự như phần trên, trong đó mỗi một máy chủ (trong đa máy chủ) sẽ hoạt động tương tự như trường hợp đơn máy chủ [1–3, 5].

Gọi $S_1(t)$, $S_2(t)$ là số lượng máy chủ đang hoạt động (bận) để phục vụ các cuộc gọi đến và các cuộc gọi đi và $N(t)$ là số lượng cuộc gọi trong *orbit* tại thời điểm t. Khi đó $X(t)=\{S_1(t), S_2(t), N(t); t\geq 0\}$ có tập không gian trạng thái $\mathcal{S}=\{(i,j,k); i=\overline{0,c}, j=\overline{0,c-i}, k=\overline{0,M}\}$.

Khi đó ma trận sinh Q sẽ có dạng như sau [1]:

$$Q=\begin{pmatrix} A^0 & B & O & O & O & O \\ C & A & B & O & O & O \\ O & C & A & \ddots & O & O \\ O & O & C & \ddots & B & O \\ O & O & O & \ddots & A & B \\ \underbrace{O \quad O \quad O \quad O \quad C \quad A}_{M+1} \end{pmatrix}$$

trong đó các ma trận A^0, A, B và C là các ma trận:

$$A=\begin{pmatrix} A_{0,1} & A_{0,0} & O & O & \cdots & O \\ A_{1,2} & A_{1,1} & A_{1,0} & O & \ddots & O \\ O & A_{2,2} & A_{2,1} & \ddots & O & \cdots \\ O & O & A_{3,2} & \ddots & A_{c-2,0} & O \\ \vdots & \ddots & \ddots & \ddots & A_{c-1,1} & A_{c-1,0} \\ O & O & \cdots & O & A_{c,2} & A_{c,1} \end{pmatrix} \quad A^0=\begin{pmatrix} A^0_{0,1} & A_{0,0} & O & O & O & O \\ A_{1,2} & A^0_{1,1} & A_{1,0} & O & \ddots & O \\ O & A_{2,2} & A^0_{2,1} & \ddots & \ddots & \cdots \\ O & O & A_{3,2} & \ddots & A_{c-2,2} & O \\ \vdots & O & \ddots & \ddots & A^0_{c-1,1} & A_{c-1,0} \\ O & O & \cdots & O & A_{c,2} & A^0_{c,1} \end{pmatrix}$$

$$C=\begin{pmatrix} \mathbf{0} & C_0 & O & O & O & O \\ O & \mathbf{0} & C_1 & O & O & O \\ O & O & \mathbf{0} & \ddots & O & O \\ O & O & O & \ddots & C_{c-2} & O \\ O & O & O & O & \mathbf{0} & C_{c-1} \\ O & O & O & O & O & \mathbf{0} \end{pmatrix} \quad B=\begin{pmatrix} B_0 & O & O & O & O & O \\ O & B_1 & O & O & O & O \\ O & O & B_2 & O & O & O \\ O & O & O & \ddots & O & O \\ O & O & O & O & B_{c-1} & O \\ O & O & O & O & O & B_c \end{pmatrix}$$

trong đó C_i và $A_{i,0}$ có kích thước $(c-i+1)\times(c-i)$, B_i, $A_{i,1}$ và $A^0_{i,1}$ có kích thước $(c-i+1)\times(c-i+1)$, $A_{i,2}$ có kích thước $(c-i+1)\times(c-i+2)$.

$$C_i = \begin{pmatrix} \mu & & & \\ & \mu & & \\ & & \ddots & \\ \underbrace{\qquad\qquad}_{c-i} & & & \mu \\ 0 & 0 & 0 & 0 \end{pmatrix}, (i = \overline{0, c-1}) \qquad B_i = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & \ddots & \\ \underbrace{\qquad\qquad}_{c-i} & & & 0 \\ & & & \lambda \end{pmatrix}, (i = \overline{0, c})$$

$$A_{i,2} = \begin{pmatrix} i\nu_1 & & & 0 \\ & i\nu_1 & & 0 \\ & & \ddots & 0 \\ \underbrace{\qquad\qquad}_{c-i+1} & & & i\nu_1 \; 0 \end{pmatrix}, (i = \overline{1, c}) \qquad A_{i,1} = \begin{pmatrix} -\gamma_{i,0} & (c-i)\alpha & & & \\ \nu_2 & -\gamma_{i,1} & (c-i-1)\alpha & & \\ & 2\nu_2 & -\gamma_{i,2} & \ddots & \\ & & 3\nu_2 & \ddots & \alpha \\ \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{c-i+1} & & & \ddots & -\gamma_{i,c-i} \end{pmatrix}, (i = \overline{0, c})$$

trong đó $\gamma_{i,j} = \lambda + \mu + i\nu_1 + j\nu_2 + (c-i-j)\alpha$.

$$A_{i,1}^0 = \begin{pmatrix} -\gamma_{i,0}^0 & (c-i)\alpha & & & \\ \nu_2 & -\gamma_{i,1}^0 & (c-i-1)\alpha & & \\ & 2\nu_2 & -\gamma_{i,2}^0 & \ddots & \\ & & 3\nu_2 & \ddots & \alpha \\ \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{c-i+1} & & & \ddots & -\gamma_{i,c-i}^0 \end{pmatrix}, (i = \overline{0, c})$$

trong đó $\gamma_{i,j}^0 = \lambda + i\nu_1 + j\nu_2 + (c-i-j)\alpha$.

$$A_{i,0} = \begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \ddots & \\ \underbrace{\qquad\qquad}_{c-i} & & & \lambda \\ 0 & 0 & 0 & 0 \end{pmatrix}, (i = \overline{0, c-1})$$

Xét ma trận $P = \mathrm{C} + A + \mathrm{B}$ là ma trận sinh của quá trình Markov $C(t) = \{S_1(t), S_2(t); t \geq 0\}$ với tập không gian trạng thái $\mathcal{V} = \{(i,j); i = \overline{0,c}, j = \overline{0, c-i}\}$. Cần lưu ý rằng chuỗi Markov này tương ứng với hệ thống tổn thất với giao tiếp hai chiều, trong đó tốc độ đến là $(\lambda + \mu)$ và các thông số khác không thay đổi. Gọi $\pi_{i,j} = \lim_{t\to\infty} P[S_1(t) = i, S_2(t) = j]$ với $(i,j) \in \mathcal{V}$ và $\boldsymbol{\pi_i} = (\pi_{i,0}, \pi_{i,1}, \dots, \pi_{i,c-i}), (i = \overline{0,c})$. Đặt $\pi = (\pi_0, \pi_1, \dots, \pi_c)$ là phân phối dừng của $C(t), (t \geq 0)$ và nó là nghiệm duy nhất của hệ phương trình [1]:

$$\pi P = \underbrace{(0,0,\dots,0)}_{\frac{(c+1)\times(c+2)}{2}} \tag{4}$$

$$\sum_{i=0}^{c}\sum_{j=0}^{c-i}\pi_{i,j} = 1 \tag{5}$$

Khi đó xác suất tắc nghẽn PB của hệ thống tổn thất được xác định:

$$PB = \sum_{i=0}^{c}\pi_{i,c-i} \tag{6}$$

2.4 Minh họa ma trận Q

Xét trường hợp đơn giản với $c = 2$ và $M = 2$, khi đó A^0, A, B và C là các ma trận:

$$C_0 = \begin{pmatrix}\mu & 0\\ 0 & \mu\\ 0 & 0\end{pmatrix},\ C_1 = \begin{pmatrix}\mu\\ 0\end{pmatrix},\ B_0 = \begin{pmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & \lambda\end{pmatrix},\ B_1 = \begin{pmatrix}0 & 0\\ 0 & \lambda\end{pmatrix},\ B_2 = (\lambda),$$

$$A_{1,2} = \begin{pmatrix}\nu_1 & 0 & 0\\ 0 & \nu_1 & 0\end{pmatrix},\ A_{2,2} = (2\nu_1 \quad 0),\ A_{0,0} = \begin{pmatrix}\lambda & 0\\ 0 & \lambda\\ 0 & 0\end{pmatrix},\ A_{1,0} = \begin{pmatrix}\lambda\\ 0\end{pmatrix},$$

$$A_{0,1}^0 = \begin{pmatrix}-(\lambda+2\alpha) & 2\alpha & 0\\ \nu_2 & -(\lambda+\nu_2+\alpha) & \alpha\\ 0 & 2\nu_2 & -(\lambda+2\nu_2)\end{pmatrix},\ A_{1,1}^0 = \begin{pmatrix}-(\lambda+\nu_1+\alpha) & \alpha\\ \nu_2 & -(\lambda+\nu_1+\nu_2)\end{pmatrix}$$

$$A_{2,1}^0 = (-(\lambda+2\nu_1))$$

$$A_{0,1} = \begin{pmatrix}-(\lambda+\mu+2\alpha) & 2\alpha & 0\\ \nu_2 & -(\lambda+\mu+\nu_2+\alpha) & \alpha\\ 0 & 2\nu_2 & -(\lambda+2\nu_2)\end{pmatrix},$$

$$A_{1,1} = \begin{pmatrix}-(\lambda+\mu+\nu_1+\alpha) & \alpha\\ \nu_2 & -(\lambda+\nu_1+\nu_2)\end{pmatrix},\ A_{2,1} = \big(-(\lambda+2\nu_1)\big),$$

Khi đó:

$$A^0 = \begin{pmatrix}A_{0,1}^0 & A_{0,0} & \mathbf{0}\\ A_{1,2} & A_{1,1}^0 & A_{1,0}\\ \mathbf{0} & A_{2,2} & A_{2,1}^0\end{pmatrix}$$

$$= \begin{pmatrix} -(\lambda+2\alpha) & 2\alpha & 0 & \lambda & 0 & 0 \\ \nu_2 & -(\lambda+\nu_2+\alpha) & \alpha & 0 & \lambda & 0 \\ 0 & 2\nu_2 & -(\lambda+2\nu_2) & 0 & 0 & 0 \\ \nu_1 & 0 & 0 & -(\lambda+\nu_1+\alpha) & \alpha & \lambda \\ 0 & \nu_1 & 0 & \nu_2 & -(\lambda+\nu_1+\nu_2) & 0 \\ 0 & 0 & 0 & 2\nu_1 & 0 & -(\lambda+2\nu_1) \end{pmatrix},$$

$$A = \begin{pmatrix} A_{0,1} & A_{0,0} & \mathbf{0} \\ A_{1,2} & A_{1,1} & A_{1,0} \\ \mathbf{0} & A_{2,2} & A_{2,1} \end{pmatrix}$$

=

$$\begin{pmatrix} -(\lambda+\mu+2\alpha) & 2\alpha & 0 & \lambda & 0 & 0 \\ \nu_2 & -(\lambda+\mu+\nu_2+\alpha) & \alpha & 0 & \lambda & 0 \\ 0 & 2\nu_2 & -(\lambda+2\nu_2) & 0 & 0 & 0 \\ \nu_1 & 0 & 0 & -(\lambda+\mu+\nu_1+\alpha) & \alpha & \lambda \\ 0 & \nu_1 & 0 & \nu_2 & -(\lambda+\nu_1+\nu_2) & 0 \\ 0 & 0 & 0 & 2\nu_1 & 0 & -(\lambda+2\nu_1) \end{pmatrix}$$

,

$$C = \begin{pmatrix} \mathbf{0} & C_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & C_1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} B_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & B_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & B_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda \end{pmatrix}.$$

và

$$P = A + C + B$$

$$= \begin{pmatrix} -(\lambda+\mu+2\alpha) & 2\alpha & 0 & \lambda+\mu & 0 & 0 \\ \nu_2 & -(\lambda+\mu+\nu_2+\alpha) & \alpha & 0 & \lambda+\mu & 0 \\ 0 & 2\nu_2 & -2\nu_2 & 0 & 0 & 0 \\ \nu_1 & 0 & 0 & -(\lambda+\mu+\nu_1+\alpha) & \alpha & \lambda+\mu \\ 0 & \nu_1 & 0 & \nu_2 & -(\nu_1+\nu_2) & 0 \\ 0 & 0 & 0 & 2\nu_1 & 0 & -2\nu_1 \end{pmatrix}$$

có các vectơ xác suất mức $\pi_0 = (\pi_{0,0}, \pi_{0,1}, \pi_{0,2})$, $\pi_1 = (\pi_{1,0}, \pi_{1,1})$ và $\pi_2 = (\pi_{2,0})$. Vectơ xác suất $\pi = (\pi_0, \pi_1, \pi_2)$ là nghiệm của hệ phương trình sau:

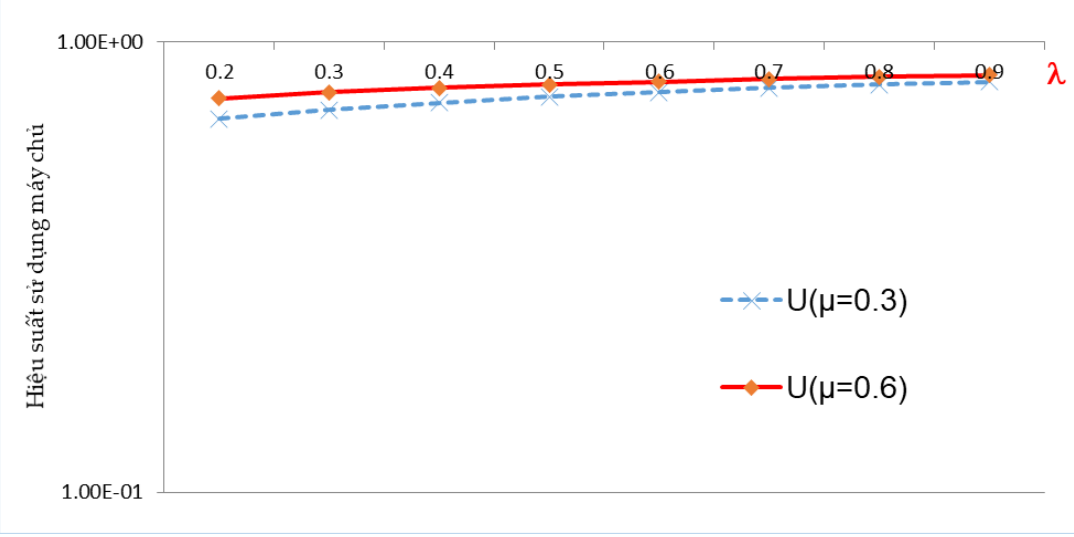
$$\begin{cases} \pi P = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0) \\ \pi(1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1)^T = 1 \end{cases}$$

3 Phân tích kết quả

Dựa theo các mô hình đã phân tích ở trên, trong phần này, chúng tôi tập trung vào phần đánh giá hiệu năng hệ thống với các thông số thay đổi để nhấn mạnh tính hiệu quả của các mô hình mà chúng tôi đã phân tích thông qua phần mềm tính toán Mathematica của Wolfram Research. Dựa theo kết quả phân tích từ các lược đồ trạng thái, chúng tôi sử dụng phương pháp phân tích quá trình giả sinh tử theo ma trận sinh Q để giải hệ phương trình tuyến tính (1)–(3) và (4), từ đó tính được các xác suất trạng thái cân bằng $\pi_{i,j}$, tính các thông số độ đo (hiệu suất sử dụng máy chủ, xác suất tắc nghẽn), sau đó tiến hành mô tả về mặt đồ thị sự biến thiên của các giá trị phụ thuộc vào lưu lượng mạng.

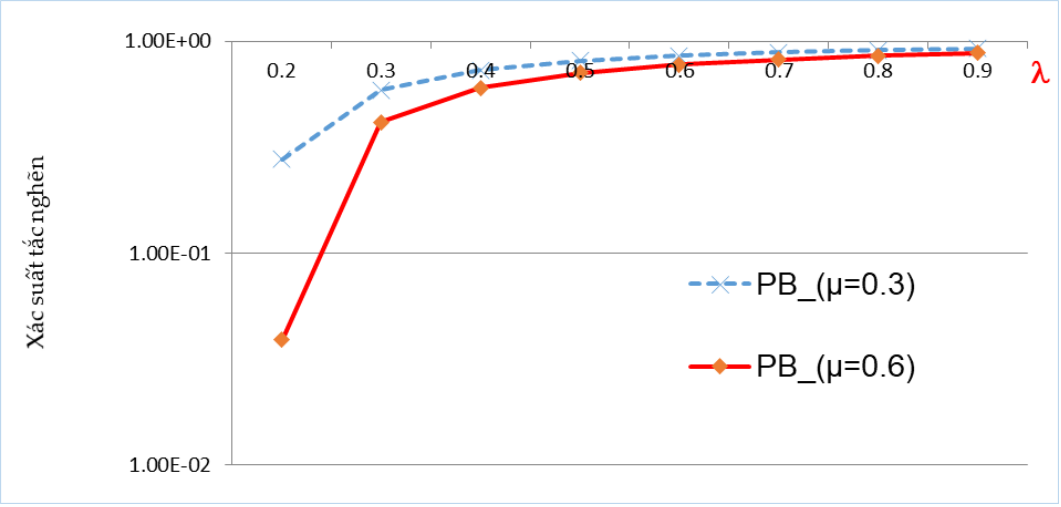
Xét với trường hợp đơn máy chủ

Kết quả trên Hình 3 biểu thị hiệu suất sử dụng server tăng khi thay đổi giá trị tốc độ retrial μ từ 0,3 lên 0,6. Điều này phù hợp với lý thuyết; đó là khi μ tăng, số cuộc gọi retrial tăng, dẫn đến khả năng server được sử dụng tăng; tức là hiệu suất sử dụng server cũng tăng.



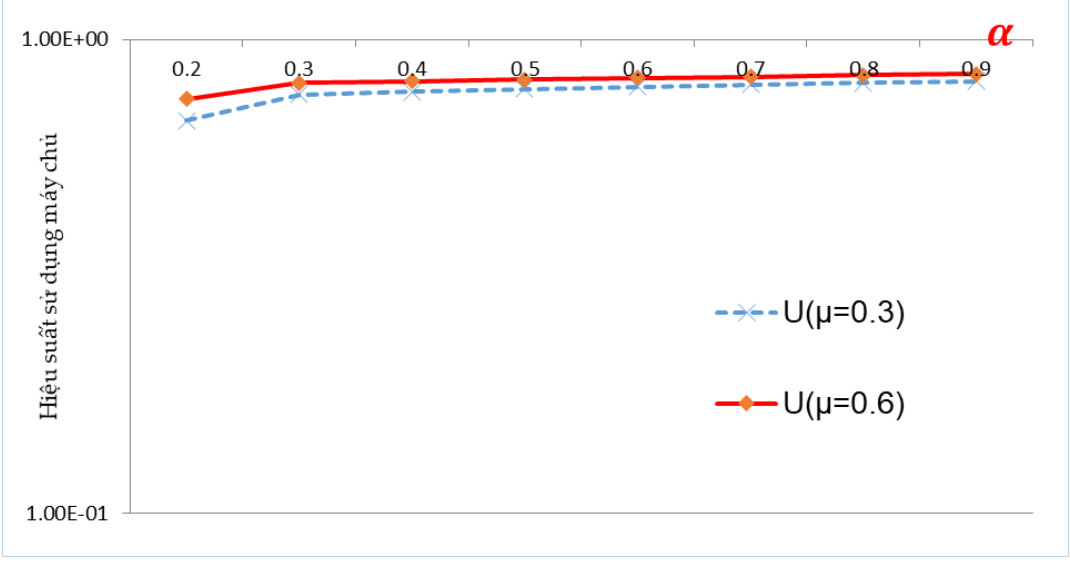
Hình 3. Hiệu suất sử dụng máy chủ theo $\mu = 0{,}3$ và $\mu = 0{,}6$ vs λ

Kết quả trên Hình 4 biểu thị xác suất tắc nghẽn khi thay đổi giá trị tốc độ retrial μ từ 0,3 lên 0,6. Theo đó, khi tốc độ cuộc gọi retrial μ tăng, tức là cho phép số cuộc gọi sẽ được thực hiện lại tăng, nên xác suất tắc nghẽn giảm. Điều này cũng cho thấy ưu điểm khi xét yếu tố retrial trong hệ thống cuộc gọi trung tâm.

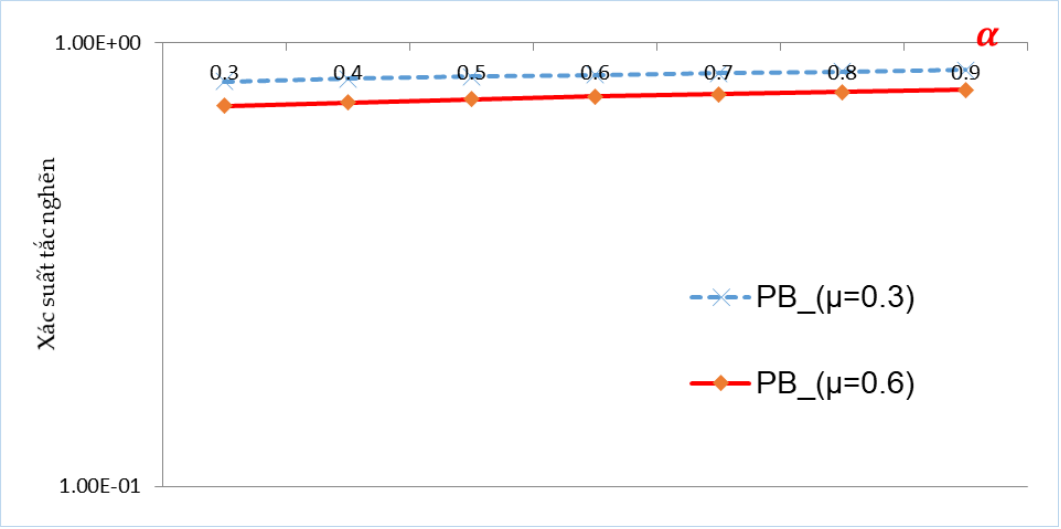


Hình 4. Xác suất tắc nghẽn theo $\mu = 0{,}3$ và $\mu = 0{,}6$ vs λ

Tương tự với các kết quả trên Hình 3 và Hình 4, chúng tôi cũng phân tích trong trường hợp giữ nguyên tốc độ đến trung bình của các cuộc gọi đến (λ) và thay đổi tốc độ trung bình của các cuộc gọi đi (α). Kết quả được trình bày trên Hình 5 và Hình 6 và đều cho thấy phù hợp với phân tích lý thuyết như trên Hình 3 và Hình 4.

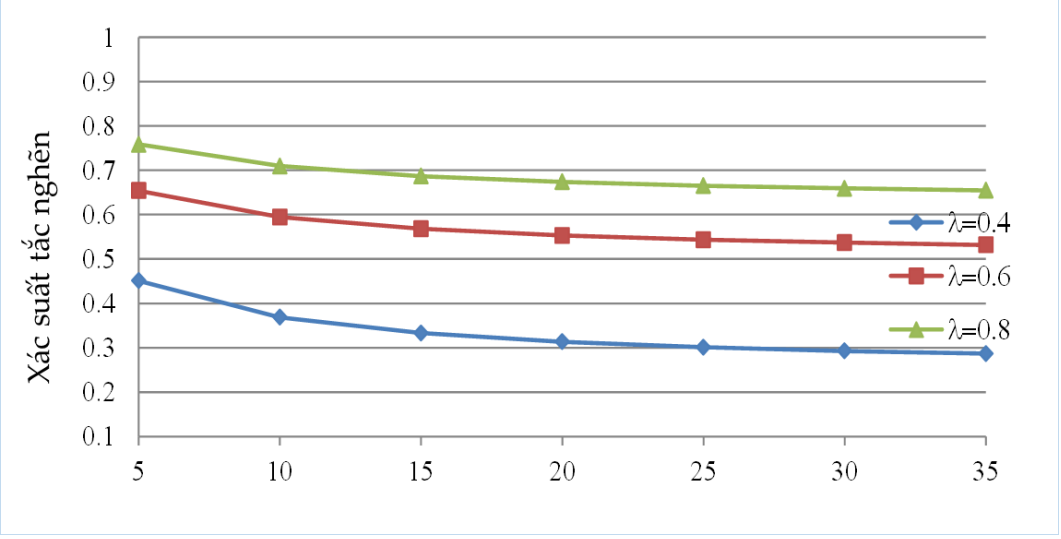


Hình 5. Hiệu suất sử dụng máy chủ theo $\mu = 0{,}3$ và $\mu = 0{,}6$ vs α



Hình 6. Xác suất tắc nghẽn theo $\mu = 0{,}3$ và $\mu = 0{,}6$ *vs* α

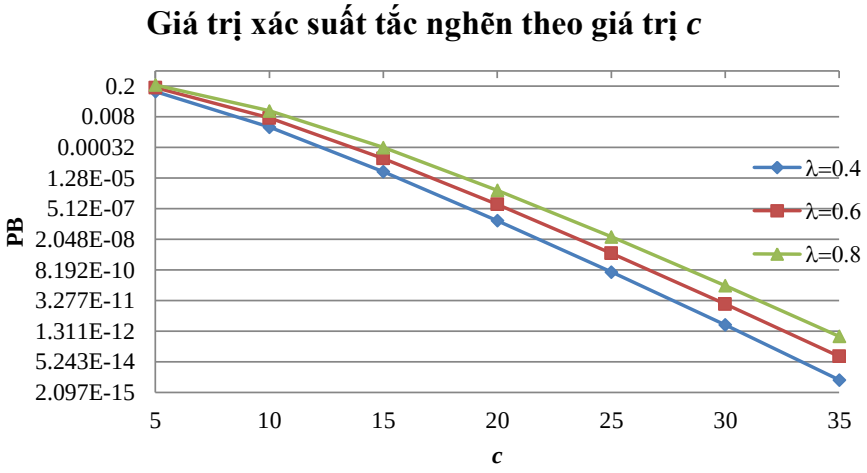
Hình 7 cho thấy kết quả xác suất tắc nghẽn theo ba giá trị λ là 0,4, 0,6 và 0,8. Trong trường hợp này, xác suất tắc nghẽn chưa bị ảnh hưởng khi giá trị M thay đổi do hệ thống đang xét chỉ có một máy chủ.



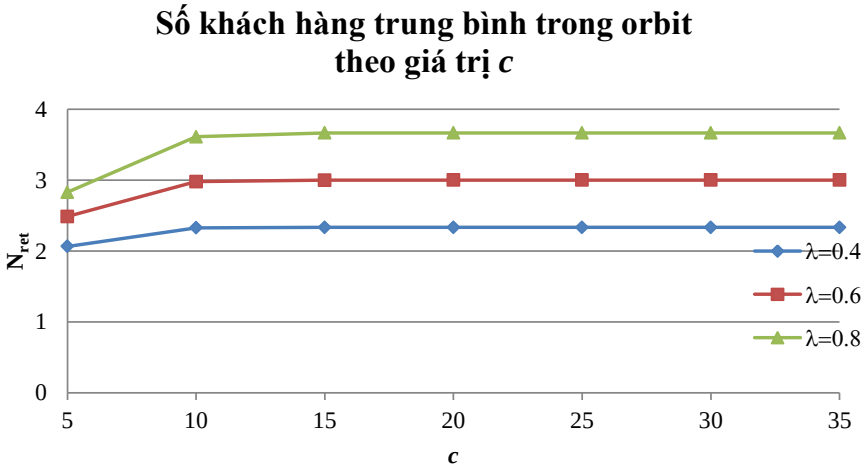
Hình 7. Xác suất tắc nghẽn theo M *vs* λ

Xét với trường hợp đa máy chủ

Khác với trường hợp chỉ có một máy chủ, giá trị xác suất tắc nghẽn có sự thay đổi lớn khi tăng số lượng máy chủ lên (Hình 8) do số lượng cuộc gọi trung bình trong orbit chỉ dao động trong khoảng hai đến bốn cuộc gọi (Hình 9, với các giá trị λ là 0,4, 0,6 và 0,8) và ít dẫn đến tắc nghẽn hệ thống.



Hình 8. Xác suất tắc nghẽn PB theo giá trị c và λ



Hình 9. Số khách hàng trung bình trong orbit N_{ret} theo giá trị c và λ

4 Kết luận

Chúng tôi đã giới thiệu bài toán mô hình hàng đợi retrial trong hệ thống trung tâm cuộc gọi với sự kết hợp các hoạt động của các cuộc gọi đến và các cuộc gọi đi trong trường hợp kích thước orbit giới hạn. Mô hình cũng xem xét với cả hai trường hợp là chỉ có một máy chủ (phân tích trạng thái máy chủ) và đa máy chủ ($c \geq 2$). Việc áp dụng phương pháp phân tích quá trình giả sinh tử theo ma trận sinh Q để tính các xác suất trạng thái cân bằng đối với mô hình cũng được xem là kết quả đạt được của bài báo. Kết quả phân tích cho thấy hiệu quả của mô hình với trường hợp nhiều máy chủ qua kết quả cải thiện xác suất tắc nghẽn, cùng thông số độ đo là số khách hàng trung bình trong orbit ổn định (phụ thuộc vào tốc độ đến của khách hàng λ).

Tài liệu tham khảo

1. Tuan Phung-Duc, Wouter Rogiest, Yutaka Takahashi, Herwig Brunee, (2014), "Retrial queues with balanced call blending: analysis of single-server and multiserver model", Ann Oper Res, DOI 10.1007/s10479-014-1598-2.
2. Artalejo, et al (2013). "Single server retrial queues with two-way communication". Applied Mathematical Modelling 37(4), 1811-1822
3. Mag. DI Dr. Christian Dombacher (9125296), Nikolaus Lenaugasse 8, A-2232 Deutsch-Wagram (2010), "Queueing Models for Call Centres", Thesis of Author and extend existing works from Zeltyn/Mandelbaum (Technion Israel), Stolletz (TU Clausthal), Whitt (AT&T Labs) and Koole (VU University Amsterdam).
4. Dequan Yue, Chunyan Li and Wuyi Yue (2016), "Performance Analysis and Optimization of a Queueing Model for a Multi-skill Call Center in M-Design", Advances in Intelligent Systems and Computing 383, Springer. [DOI: 10.1007/978-3-319-22267-7_14].
5. Tuan Phung-Duc, Ken'ichi Kawanishi (2014), "Performance analysis of call centers with abandonment, retrial and after-call work", Performance Evaluation 80, 43–62, 2014.
6. Hiroyuki Sakurai and Tuan Phung-Duc (2014), "Two Way Communication Retrial Queues with Multiple Types of Outgoing Calls", TOP, 2015, Volume 23, Number 2, Page 466 – 492.
7. S. Ding, M. Remerova, R.D. van der Mei, B. Zwart, "Fluid approximation of a call center model with redials and reconnects", Performance Evaluation, Volume 92, July 2015, Pages 24-39.
8. Chia-Jung Chang, Fu-Min Chang, Jau-Chuan Ke (2019), "Optimal power consumption analysis of a load-dependent server activation policy for a data service center", Computers & Industrial Engineering, Volume 130, Pages 745-756.
9. Dmitry Efrosinin, Anastasia Winkler (2011), "Queueing system with a constant retrial rate, non-reliable server and threshold-based recovery", European Journal of Operational Research, Volume 210, Issue 3, Pages 594-605.
10. Tien Van Do, Ram Chakka (2010). "An efficient method to compute the rate matrix for retrial queues with large number of servers", Applied Mathematics Letters 23, 638-643.