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

Tập 133, Số 2A, 2024, Tr. 63–82; DOI: 10.26459/hueunijtt.v133i2A.7492

MÔ HÌNH HÓA PHÂN CHIA TỶ LỆ KHỐI NSSI TRONG PHÂN CHIA MẠNG 5G DỰA TRÊN NGƯỠNG

Hoa Lý Cương[1,2], Đặng Thanh Chương[1*], Nguyễn Quang Hưng[1]

[1] Trường Đại học Khoa học, Đại học Huế, 77 Nguyễn Huệ, Thành phố Huế, Việt Nam
[2] Trường Đại học Cần Thơ, Thành phố Cần Thơ, Việt Nam

Tóm tắt. Trong mạng lõi 5G, các chức năng mạng có thể được phân chia tỉ lệ (scaling-out/in) một cách linh hoạt để điều chỉnh dung lượng cho các lát mạng (network slices). Quy trình phân chia tỉ lệ tự động nâng cao hiệu suất bằng cách mở rộng quy mô (scaling-out) các phiên bản chức năng mạng và giảm chi phí vận hành bằng cách giảm quy mô (scaling-in) các phiên bản. Các chức năng mạng 5G phải được triển khai/tắt đồng thời nhiều phiên bản NFIs, thường được xem xét trong một khối (block) các phiên bản NSSI. Vì vậy, việc thiết lập số lượng các phiên bản cần triển khai đồng thời, cũng như số lượng phiên bản cần tắt khi không có nhu cầu sử dụng sẽ ảnh hưởng đáng kể đến hiệu quả chi phí của hệ thống, cũng như đảm bảo chất lượng dịch vụ QoS. Ngoài ra, việc triển khai thiết lập các khối thường xuyên sẽ làm giảm đáng kể hiệu suất hệ thống (chi phí thiết lập), vì vậy vấn đề đặt trước (thiết lặp sẵn) một số khối nhất định cũng cần được xem xét. Trong bài viết này, chúng tôi sẽ mô hình hóa hệ thống tổng quát theo mô hình hàng đợi Markov cho bài toán phân chia tỉ lệ các khối phiên bản NSSI thuộc lát mạng trong mạng lõi 5G. Mô hình sẽ tích hợp 2 ngưỡng liên quan đến việc thiết lập/gỡ bỏ (scaling-out/scaling-in) khối NSSI dựa trên lưu lượng truy cập của người dùng đến hệ thống. Các kết quả phân tích số cho thấy mô hình phân chia tỉ lệ các khối NSSI dựa trên ngưỡng đề xuất trong bài báo này có thể giúp quản lý cấp phát tài nguyên hệ thống một cách hiệu quả.

Từ khóa. Phân chia mạng, Thiết lập khối NSSI, Scaling-out/Scaling-in, Mạng 5G, Mô hình Q-TS_NB

Threshold-based scaling model for NSSI blocks in 5G network slicing

Hoa Ly Cuong[1,2], Dang Thanh Chuong[1*], Nguyen Quang Hung[1]

[1] University of Sciences, Hue University, 77 Nguyen Hue St., Hue, Vietnam
[2] Can Tho University, Can Tho, Vietnam

Abstract. In the 5G core network, network functions may be dynamically scaled out and in to modify capacity for network slices. Automatic scaling enhances performance by scaling-out

** Liên hệ: dtchuong@hueuni.edu.vn*

Ngày nhận bài: 13–06–2024; Ngày nhận đăng: 03–07–2024

network function instances while decreasing operational expenses by scaling-in instances. 5G network functionalities must be deployed and stopped concurrently across numerous NFI instances, which are commonly treated as a block of NSSI instances. As a result, determining the number of instances that must be deployed concurrently, as well as the number of instances that must be shut off when no longer in use, will have a substantial impact on the system's cost efficiency and QoS service quality. Furthermore, using frequent block setup may drastically lower system performance (setup cost), hence the problem of reserving (presetting) specific blocks must be considered. In this article, we will use the Markov queuing model to represent the whole system for scaling NSSI instance blocks corresponding to network slices in the 5G core network. The model will include two thresholds for the setup/removal (scaling-out/scaling-in) of NSSI blocks based on system user traffic. The numerical analysis findings suggest that the scaling model of NSSI blocks based on the threshold presented in this study can efficiently control system resource allocation.

Keywords: Network Slicing, Block NSSI setup, Scaling-out/Scaling-in, 5G Network, Q-TS_NB (A Queueing Model for Thresholds-based Scaling of NSSI Block)

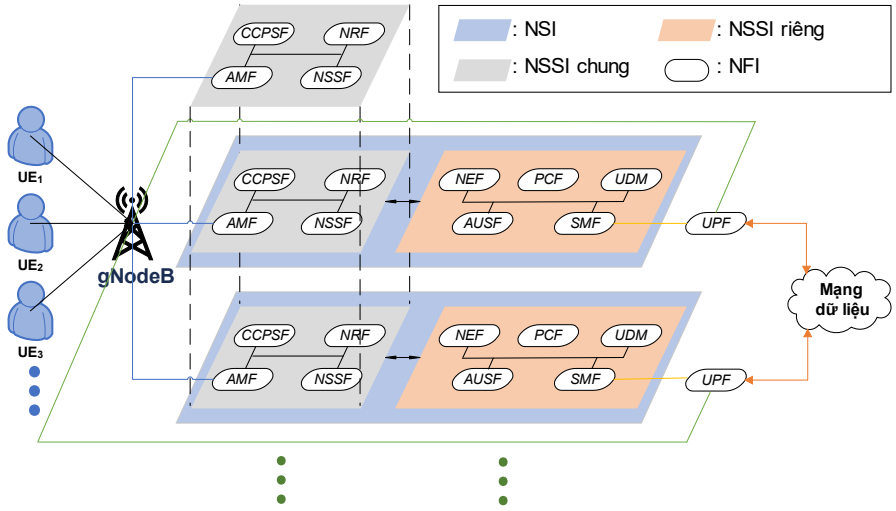
1 Giới thiệu

Từ 3GPP (3rd Generation Partnership Project) Release 15 (R15) [1-5], mạng băng thông rộng thế hệ thứ 5 (5G) được thiết kế không chỉ phục vụ cho con người mà còn cho tất cả mọi thứ, bao gồm cả các loại máy móc khác nhau. Hơn nữa, các hàm chức năng mạng (network functions) trong các mạng lõi 5G sẽ được triển khai không chỉ bởi các nhà điều hành mạng. Các doanh nghiệp, nhà máy hoặc cơ quan chính phủ cũng có thể xây dựng các mạng di động riêng của họ để tăng hiệu suất và tăng cường bảo mật. Khi người dùng triển khai một mạng di động riêng, các chức năng mạng về mặt phẳng điều khiển (control plane) có thể vẫn được cung cấp bởi các nhà điều hành. Người dùng chỉ sở hữu các chức năng mặt phẳng người dùng (user plane) theo yêu cầu của họ [1].

Chức năng ảo hoá mạng NFV (Network Function Virtualization) là mô hình kiến trúc mạng được thiết kế để ảo hóa các dịch vụ mạng thường chạy trên các thiết bị mạng chuyên dụng, riêng biệt. Với NFV, các chức năng như định tuyến, cân bằng tải và kiểm soát truy cập tường lửa được đóng gói dưới dạng các máy ảo VM (Virtual Machine) hoặc các quy trình được đóng gói được phân phối trên phần cứng thông dụng. Các chức năng mạng ảo VNF (Virtual Network Function) riêng lẻ là một thành phần thiết yếu của kiến trúc NFV.

Như được định nghĩa trong 3GPP TS 28.530 [5], một lát mạng (network slice) bao gồm một số phiên bản lát mạng (NSI - network slice instances), có thể được phân chia tỉ lệ (scaled out/in) (turned on/off - bật/tắt) để tự động điều chỉnh dung lượng của các chức năng mạng lõi (Hình 1).

Nghĩa là, các nhà khai thác có thể áp dụng các thuật toán chia tỷ lệ để quản lý các phiên bản lát mạng NSI một cách hợp lý [1].



Hình 1. Kiến trúc phân chia mạng dựa trên chuẩn 3GPP R15 [1]

Trong Hình 1, một lát mạng (slice network) bao gồm nhiều NSI, chẳng hạn như NSI_1 và NSI_2, trong đó một NSI có thể bao gồm nhiều phiên bản con lát mạng $NSSI$ (Network Slice Subnet Instance) hoặc một NSI chỉ gồm một NSSI duy nhất. Ví dụ: NSI1 bao gồm $NSSI_1$ và $NSSI$ chia sẻ chung. Một NSSI bao gồm nhiều phiên bản chức năng mạng NFI, về cơ bản chính là các VNFs. Ví dụ: NEF, PCF, UDM, $AUSF$, SMF và UPF đều là các VNF bên trong $NSSI_1$. Ngoài ra, các NSI có thể chia sẻ chung các chức năng của mặt phẳng điều khiển trên một $NSSI$ chung (ví dụ: $NSSI$ chung trong Hình 1 được chia sẻ giữa NSI_1 và NSI_2). $NSSI$ được chia sẻ chủ yếu bao gồm các chức năng mặt phẳng điều khiển mạng. Do hiện tượng tắc nghẽn (nút cổ chai - bottleneck) thường xảy ra ở mặt phẳng người dùng, nên các chức năng mạng mặt phẳng điều khiển không được triển khai thường xuyên thường được tích hợp lại và chia sẻ chung bởi các $NSSI$, chẳng hạn như $NSSI_1$ và $NSSI_2$. Khi thiết lập một kết nối, đầu tiên người dùng sẽ gửi yêu cầu xác thực ban đầu (initial authentication) đến mặt phẳng điều khiển. Sau khi thiết lập phiên thành công, dữ liệu có thể được truyền qua UPF mà không cần đi qua mặt phẳng điều khiển. Tuy nhiên, khi người dùng đột nhiên gửi một lượng lớn dữ liệu, nhà điều hành có thể tự động thiết lập nhiều UPF hơn nhằm giảm thiểu tắc nghẽn hệ thống.

Các chức năng mạng (NFI) trong mạng lõi 5G có thể được phân chia tỉ lệ (scaled out/in) một cách linh hoạt để điều chỉnh dung lượng cho các lát mạng (network slices). Quy trình phân chia tỉ lệ, nâng cao hiệu suất bằng cách mở rộng quy mô (scaling out) các phiên bản (instances)

và giảm chi phí vận hành bằng cách giảm quy mô (scaling in) các phiên bản. việc phân chia tỉ lệ tự động các NSSI trong mỗi *NSI* có thể mang lại sự linh hoạt, giảm chi phí vận hành và đáp ứng các yêu cầu về hiệu suất của người dùng. Quy trình phân chia tỉ lệ tự động của *NSSI* trong 3GPP tuân theo hai nguyên tắc [1]:

- *Thiết lập khối (Block setup):* Như đã nói ở trên, một lát mạng trong 5G có ba lớp khác nhau: *NSI*, *NSSI* và *(NFI - network function instance)*. Để triển khai các lát mạng, chúng ta thường cần xử lý một *NSSI*, không chỉ bao gồm một số *NFI* mà còn có thể được kết hợp với các *NSSI* khác để tạo thành một *NSI* mới. Nghĩa là, việc phân chia mạng được xác định trong các tiêu chuẩn 3GPP cần tính đến thành phần và sự kết hợp của từng lớp. Hơn nữa, theo 3GPP R15, các chức năng mạng 5G phụ thuộc nhiều vào nhau. Ví dụ, để triển khai UPF, mạng phải tính đến dung lượng hệ thống của SMF. Nếu số lượng UPF yêu cầu vượt quá tải của SMF, các SMF bổ sung cũng có thể được triển khai. Do đó, việc phân chia mạng 5G không chỉ xem xét kiến trúc phân lớp mà còn phân tích tác động của việc phân chia tỉ lệ nhiều *NFI* (multiple *NFIs*), được gọi là khối (block) trong bài báo này.
- *Đặt trước tài nguyên (Reservation):* Trong ETSI GS NFV-IFA 010 [4] và 3GPP 28.531 [5], các cơ quan tiêu chuẩn đề xuất khái niệm đặt trước (hay dự trữ, dành trước) tài nguyên, trong đó hệ thống bật trước (turn on) một số phiên bản mạng *NFI (network instances)* để phòng ngừa tình trạng lưu lượng yêu cầu truy cập tăng cao đột ngột.

Trong bài báo này, chúng tôi sẽ mô hình hóa bài toán phân chia tỉ lệ (thiết lập/gỡ bỏ) các khối NSSI (với nhiều VNFs tạo thành một khối *NSSI*) thuộc một *NSI* (triển khai thiết lập/gỡ bỏ nhiều khối NSSI) trong một lát mạng một cách linh động (dynamic) theo lưu lượng UE đến dựa trên các ngưỡng scaling out/in cho mạng lõi 5G bằng mô hình hàng đợi Markov [1, 6]. Chi tiết mô hình đề xuất sẽ được trình bày trong phần III của bài báo này.

Đóng góp chính của bài báo gồm:

(1). Xây dựng thuật toán phân chia tỉ lệ (mở rộng quy mô) nhằm thiết lập/gỡ bỏ khối NSSI trong phân chia mạng lõi 5G dựa trên lưu lượng yêu cầu đến.

(2). Xây dựng mô hình hàng đợi nhằm mô hình hóa bài toán phân chia tỉ lệ theo khối NSSI động theo các ngưỡng thiết lập và gỡ bỏ khối ứng với các NSI thuộc lát mạng trong mạng lõi 5G. Mô hình mới vì vậy sẽ có tên là Q-TS_NB (A Queueing Model for Thresholds-based Scaling of NSSI Block). Mô hình toán học của chúng tôi dựa trên chuẩn 3GPP và cho phép điều chỉnh các tham số như kích thước khối, quy mô hệ thống, tốc độ đến, tốc độ phục vụ, các tham số ngưỡng,… để đánh giá tác động của chúng đến hiệu suất của hệ thống. Vấn đề này được thể hiện trong phần phân tích kết quả của bài báo này.

Nội dung tiếp theo của bài báo bao gồm: Phần II khảo sát các tài liệu liên quan về phân chia tỉ lệ để quản lý tài nguyên trong lõi 5G ; mô hình đề xuất Q-TS_NB phân tích bài toán phân chia mạng trong mạng lõi 5G dựa trên kiểm soát theo các ngưỡng được trình bày trong Phần III; Phần IV trình bày kết quả phân tích dựa trên mô hình thuật toán đề xuất. Cuối cùng là phần kết luận và hướng phát triển của bài báo.

2 Các công trình nghiên cứu liên quan

Nhiều nghiên cứu ứng dụng mô hình hàng đợi với bài toán phân chia tỉ lệ tài nguyên trong mạng đã được đề xuất [1, 6, 7-8]. Các tác giả trong [7] xem xét hệ thống xếp hàng với hàng đợi và nhóm hàng đợi retrial của nó để mô hình hóa công nghệ phân chia mạng trong mạng không dây 5G. Trong công trình [8], các tác giả đề xuất ba số liệu (three metrics) có thể được sử dụng để đánh giá hiệu quả của phân chia mạng động. Một mặt, việc phân chia lại (re-slicing) có thể dẫn đến việc cấp phát tài nguyên thích ứng hơn cho các nhà khai thác mạng ảo (VNO - virtual network operators) khác nhau, nhưng có thể phát sinh chi phí báo hiệu. Mặt khác, số lượng phân chia lại không đủ có thể làm giảm đáng kể chất lượng dịch vụ cho người dùng VNO, nhưng giảm độ trễ tín hiệu. Các số liệu đề xuất trong [8] có thể được sử dụng để phân tích tác động nói trên. Theo đó, họ đã minh họa các số liệu bằng mô hình mô phỏng cho thuật toán phân chia mạng động đơn giản. Đồng thời, các tác giả trong [8] cũng đề xuất một cách tiếp cận sử dụng mô hình hàng đợi để phân tích việc phân chia mạng động cho 2 VNO.

Các chiến lược phân chia tỉ lệ tự động trong *NSSI* có thể mang lại sự linh hoạt, giảm chi phí vận hành và đáp ứng các yêu cầu về hiệu suất của người dùng. Tuy nhiên, việc phân chia tỉ lệ thường xuyên không chỉ làm giảm hiệu suất hệ thống mà còn dẫn đến chi phí tăng đáng kể, vì vậy các tác giả trong [1] xem xét tác động của việc đặt trước số phiên bản NFI để phòng ngừa tình trạng lưu lượng yêu cầu truy cập tăng cao đột ngột, đồng thời thiết kế thiết lập khối, tức là lập mô hình nhiều VNF dưới dạng một khối (multiple VNFs as a block) và xem xét việc triển khai nhiều khối theo chiến lược phân chia tỉ lệ theo ngưỡng. Với cách tiếp cận dựa trên ngưỡng, các tác giả trong [1] tận dụng cách thức ngưỡng đơn/kép để thiết kế cơ chế thiết lập khối, trong đó tối ưu hóa thời gian thiết lập và số lượng phiên bản bằng cách kiểm soát giá trị ngưỡng và xác định thời gian tắt và số lượng phiên bản dựa trên nhu cầu của người vận hành. Tuy nhiên việc thiết lập khối dựa trên ngưỡng m ứng với các yêu cầu UE trong hàng đợi có thể gặp vấn đề về độ trễ, điều này sẽ được chúng tôi khắc phục trong mô hình đề xuất của bài báo này tại phần tiếp theo.

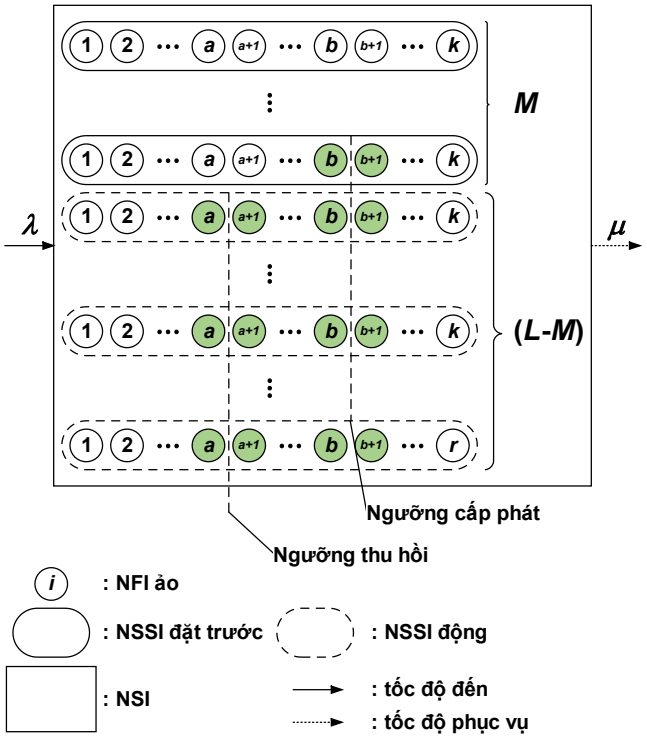
Trong 5G, UPF là chức năng mặt phẳng người dùng (user-plane), được sử dụng để truyền dữ liệu. Cụ thể, UPF chịu trách nhiệm duy trì bảng định tuyến và thiết lập các đường hầm GTP để chuyển tiếp các gói tin. Trước khi truyền, UPF giải mã các tiêu đề của gói và tìm kiếm các quy

tắc định tuyến (routing rules) trong tập dữ liệu của nó. Trong 3GPP R15, chức năng quản lý phiên (SMF - session management function) có thể kiểm soát/điều khiển nhiều UPF. Đối với các nhà khai thác (operators), làm thế nào để thiết lập một số lượng lớn UPF với chi phí thấp nhất đồng thời đáp ứng yêu cầu của người dùng luôn là một thách thức. Theo đó, các tác giả trong [6] đã đề xuất một mô hình hàng đợi cho hệ thống mặt phẳng dữ liệu UPF trong mạng lõi 5G dựa trên thuật toán chia tỷ lệ theo ngưỡng để quản lý các UPF một cách hợp lý. Ý tưởng phân chia tỉ lệ theo ngưỡng trong [6] cũng sẽ được chúng tôi áp dụng vào mô hình trong bài báo này.

3 Mô hình hàng đợi cho bài toán phân chia tỉ lệ các khối NSSI dựa trên ngưỡng trong phân chia mạng 5G

3.1 Mô hình hóa hệ thống phân chia tỉ lệ theo khối NSSI dựa trên ngưỡng

Mô hình hệ thống phân chia tỉ lệ các khối NSSI thuộc một NSI trong một lát mạng được mô hình hóa theo mô hình hàng đợi Markov $M/M/K/K$ được chỉ ra trong Hình 2.



Hình 2. Mô hình phân chia tỉ lệ theo khối NSSI dựa trên ngưỡng

Theo đó, giả thiết sẽ có K phiên bản NFIs ảo hóa trong một NSI được chia thành tối đa L khối NSSI, trong đó mỗi khối chứa k phiên bản (khối cuối cùng có thể có r phiên bản với $r \leq k$). Hệ thống sẽ khởi tạo (bật trước) tối thiểu M khối ($M < L$), các khối còn lại sẽ được thiết lập (scaling-out) hoặc tắt (scaling-in) theo thuật toán phân chia tỉ lệ (scaling) dựa trên ngưỡng (Thuật toán 1) tùy thuộc vào lưu lượng các UE yêu cầu đến [1,6]. Vấn đề đặt ra trong mô hình phân tích ở Hình 2, đó là có thể đạt hiệu quả chi phí sử dụng tài nguyên của của hệ thống bằng cách phân tích các vấn đề: *đặt trước*, thiết lập và gỡ bỏ khối động [1][6]. Các lát mạng (slice network) được triển khai với cấu trúc phân lớp trong đó nhiều NFI được thiết lập để tạo thành một $NSSI$ và nhiều $NSSI$ được triển khai để tạo thành một NSI [1]. Để đạt được mục tiêu này, chúng ta xem k phiên bản NFI là một *khối* (block) để mô tả mối quan hệ giữa các $NFIs$ và $NSSI$ và thiết lập các *khối* riêng biệt để mô tả mối quan hệ giữa $NSSI$ và NSI. Ngoài ra, việc triển khai thiết lập các khối thường xuyên sẽ làm giảm đáng kể hiệu suất hệ thống (chi phí thiết lập), vì vậy mô hình cũng thực hiện đặt trước (thiết lập sẵn) đặt trước tối thiểu M khối NSSI, đồng thời sử dụng các giá trị ngưỡng thiết lập và gỡ bỏ khối (T_1 và T_2 một cách tương ứng), phù hợp để đảm bảo việc cấp phát tài nguyên một cách hợp lý. Các thông số quan trọng trong mô hình được giới thiệu trong Bảng 1.

Bảng 1. Các ký hiệu được sử dụng trong mô hình thresholds-based scaling

Ký hiệu	Ý nghĩa
$I(t)$	Số lượng các yêu cầu UE đang được phục vụ bởi các NFI
$J(t)$	Số lượng khối NSSI (chứa các NFI ảo động) đang được triển khai tại thời điểm t
K	Tổng số phiên bản NFI ảo mà hệ thống cung cấp
k	Số lượng phiên bản NFI ảo trong một khối NSSI (kích thước khối)
r	Số lượng phiên bản NFI ảo trong khối NSSI cuối cùng (kích thước khối thứ L)
N_j	Số lượng phiên bản NFI ảo rỗi tối đa khi có j khối NSSI đã được triển khai ($M \leq j \leq L$) với điều kiện $N_j = \begin{cases} j \times k, (M \leq j < L) \\ (j-1) \times k + r, (j = L) \end{cases}$
M	Số khối NSSI (chứa các NFI ảo) được đặt trước.
L	Số khối NSSI tối đa (bao gồm cả các NFI ảo đặt trước và các NFI ảo động) có thể được khởi chạy của hệ thống ($L = K/k$).
λ	Tốc độ đến trung bình của các yêu cầu dịch vụ
μ	Tốc độ phục vụ một yêu cầu UE
T_1	Ngưỡng thiết lập (scaling-out) khối NSSI (chứa các NFI ảo động)
T_2	Ngưỡng tắt khối (scaling-in) NSSI (chứa các NFI ảo động)

Trong mô hình hệ thống này, chúng tôi sử dụng giá trị ngưỡng (T_1) để xác định khi nào (thời điểm) khối $NSSI$ cần được thiết lập theo các chiến lược tiết kiệm chi phí của nhà khai thác, gọi là ngưỡng cho quyết định chia tỷ lệ scaling-out [6]. Việc sử dụng ngưỡng T_1 ở đây khắc phục được vấn đề dùng ngưỡng m như mô hình trong [1], theo đó, trong [1], khi các yêu cầu đến mà tất cả n_0 NFI đã được cấp phát, UE phải được đưa vào hàng đợi, điều này có thể làm tăng độ trễ của các UE. Thay vào đó, mô hình của chúng tôi sử dụng ngưỡng T_1 cho phép thiết lập trước các khối $NSSI$ tùy theo lưu lượng vào, đảm bảo khi các UE đến luôn có sẵn tài nguyên (các phiên bản NFI) để cấp phát. Tương tự, giá trị ngưỡng tắt (gỡ bỏ) khối, còn gọi là ngưỡng cho quyết định chia tỉ lệ scaling-in, ký hiệu là T_2 được áp dụng khi hệ thống cần tắt một hoặc một vài khối NSSI đang rỗi trong trường tải hợp lượng đến thấp nhằm tránh lãng phí tài nguyên, giảm chi chí vận hành các khối [6].

3.2 Thuật toán phân chia tỉ lệ khối NSSI dựa trên ngưỡng

Theo phân tích ở trên, gọi $I(t)$ biểu thị số yêu cầu UE đang được phục vụ bởi các phiên bản $NFIs$ và $J(t)$, $(M \le J(t) \le L)$, là số khối $NSSI$ (chứa các NFI ảo động (theo Hình 2) đang được bật) tại thời điểm t. Số yêu cầu UE tối đa có thể được phục vụ tại thời điểm t là $N_{J(t)} = \left\lceil \frac{J(t)\times(L-1)}{L} \right\rceil \times k + \left\lfloor \frac{J(t)}{L} \right\rfloor \times r$. Trong mô hình phân tích này, việc thiết lập và tắt các khối được thực hiện một cách linh động theo lưu lượng UE đến dựa vào các ngưỡng T_1 và T_2. Cụ thể, mô hình sẽ thiết lập một khối gồm k phiên bản (scaling-out) khi số UE đang được phục vụ đạt đến giá trị $N_j - T_1 - 1$ theo ngưỡng T_1. Ngoài ra, mô hình cũng sẽ tắt một khối gồm k phiên bản khi có k phiên bản (trong 1 khối) đều ở trạng thái rỗi, tương ứng với trường hợp khi số UE đang được phục vụ đạt đến giá trị $N_j - T_2 + 1$ theo ngưỡng T_2. Việc kiểm soát các giá trị ngưỡng trong thiết lập và gỡ bỏ khối (T_1 và T_2 một cách tương ứng) có thể được thực hiện bằng cách áp dụng các thuật toán chia tỷ lệ (scaling algorithm) để quản lý các phiên bản lát mạng $NSSI$ một cách hợp lý (Thuật toán 1).

Thuật toán 1: phân chia tỉ lệ các khối NSSI theo ngưỡng [6]:

Input: các giá trị M, L, k, λ, μ, T_1, T_2,

Output: $I(t), J(t)$

Ý tưởng:

1. While $I(t) \le (L-1) \times k + r$
2. $\quad \left\lceil \frac{J(t)\times(L-1)}{L} \right\rceil \times k + \left\lfloor \frac{J(t)}{L} \right\rfloor \times r \rightarrow N_{J(t)}$
3. $\quad$ <Khi có UE đến>
   $\quad$ {

4. $I(t) \rightarrow I(t) + 1$
5. If $I(t) = N_{J(t)} - T_1 - 1$ and $J(t) < L$
6. $J(t) \rightarrow J(t) + 1$; // *scaling-out*

}

7. <Khi có UE rời đi>

{

8. $I(t) \rightarrow I(t) - 1$
9. If $I(t) = N_{J(t)} - T_2 + 1$ and $J(t) \geq M + 1$
10. $J(t) \rightarrow J(t) - 1$; // *scaling-in*

}

11. Return $I(t), J(t)$

Ở đây thuật toán có độ phức tạp $0(k \times L)$.

3.3 Mô hình Q-TS_NB

Mô hình Q-TS_NB có thể được mô hình hóa hoạt động quản lý tài nguyên ứng với một lát mạng NSI là một chuỗi Markov hai chiều với thời gian liên tục CTMC $\{(I(t), J(t)), t \geq 0\}$, tốc độ của các yêu cầu UE đến được giả thiết theo phân phối Poisson λ và khoảng thời gian phục vụ theo phân phối hàm mũ $\frac{1}{\mu}$ (μ là tốc độ phục vụ). Ta có không gian trạng thái $S = \{(i, j) | i \in I, j \in J\}$ với các bước chuyển trạng thái tương ứng trong mô hình có thể được mô tả theo Thuật toán 1, như sau [6]:

- Chuyển từ trạng thái (i, j) sang $(i + 1, j)$: khi có yêu cầu UE mới đến và hệ thống có thể phục vụ được (cấp phát 1 phiên bản NFI), và số lượng khối NSSI không điều chỉnh tăng, với một trong các các điều kiện như sau:
  - $0 \leq i \leq N_M - T_1 - 2, j = M,$
  - $N_j - T_2 < i \leq N_j - T_1 - 2, M < j < L,$
  - $N_L - T_2 < i \leq N_L - 1, j = L.$
- Chuyển từ trạng thái (i, j) sang $(i + 1, j + 1)$: khi có yêu cầu UE mới đến và hệ thống có thể phục vụ được (cấp phát 1 phiên bản NFI), đồng thời khởi chạy (bật) khối $NSSI$ mới dựa trên quy tắc hoạt động ở trên (scaling-out). Lúc này $i = N_j - T_1 - 1, M \leq j < L$.

- Chuyển từ trạng thái (i,j) sang trạng thái $(i-1,j)$: khi có một người dùng đã hoàn thành và rời đi (trả lại 1 phiên bản NFI), số khối $NSSI$ đã bật vẫn được giữ nguyên, với một trong các điều kiện như sau:
  - $1 \le i < N_M - T_1, j = M,$
  - $N_j - T_2 + 2 \le i < N_j - T_1, M < j < L,$
  - $N_L - T_2 + 2 \le i \le N_L, j = L.$
- Chuyển từ trạng thái (i,j) sang trạng thái $(i-1,j-1)$: khi có một người dùng đã hoàn thành và rời đi, đồng thời hệ thống cũng chấm dứt một khối NSSI dựa trên quy tắc hoạt động ở trên (scaling-in). Lúc này $i \le N_j - T_2 + 1$, $M < j \le L$. Với trường hợp chọn $T_2 = k$, có thể viết lại $i \le N_j - k + 1$ hay $i \le k \cdot (j-1) + 1$, với $M < j \le L$.

Khi đó, không gian trạng thái S của quá trình Markov CTMC $\{(I(t), J(t)), t \ge 0\}$ được biểu thị bằng [6]:

$$S = \{(i,M): 0 \le i \le N_M - T_1 - 1\} \cup \{(i,j): N_j - T_2 + 1 \le i \le N_j - T_1 - 1, M < j < L\} \cup \{(i,L): N_L - T_2 + 1 \le i \le N_L\}$$

Số các trạng thái là:

$$N_M - T_1 + (T_2 - T_1 - 1)(L - M - 1) + T_2 + k - r = M \times k + T_2 - T_1 + (T_2 - T_1 - 1)(L - M - 1) + k - r$$

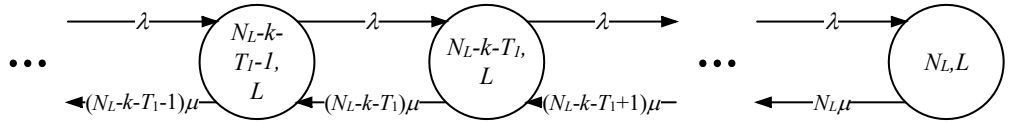
Các trạng thái (states) khi có j khối $NSSI$ được khởi chạy được gọi là trạng thái cấp độ (level) j. Nếu với j, $M < j < L$, các khối NSSI được triển khai, do quy tắc hoạt động, sẽ có hai trạng thái đặc biệt xảy ra:

- Trạng thái $(N_j + k - T_2, j)$ có thể đạt được từ trạng thái $(N_j + k - T_2 + 1, j + 1)$ $\big((N_j + k - T_2 + 1, j + 1) \to (N_j + k - T_2, j)\big)$ do sự rời đi của một UE và hành động scaling-in được thực hiện để tắt một khối NSSI.
- Trạng thái $(N_j - k - T_1, j)$ là kết quả hành động scaling-out từ trạng thái trước đó là $(N_j - k - T_1 - 1, j - 1)$ $\big((N_j - k - T_1 - 1, j - 1) \to (N_j - k - T_1, j)\big)$ khi một yêu cầu UE đến (và khởi tạo 1 khối NSSI).

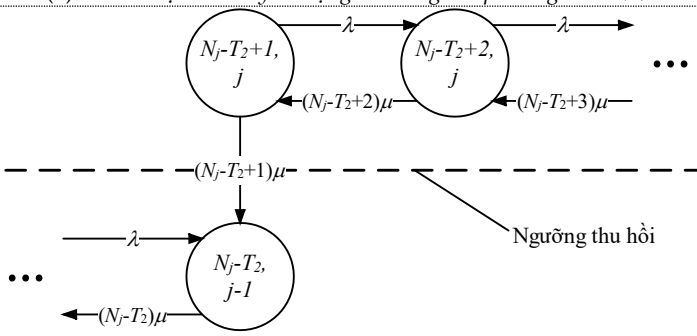
Ở đây, có hai trường hợp cần phân biệt trên dựa theo mối quan hệ giữa $N_j + k - T_2$ và $N_j - k - T_1$. Trong trường hợp đầu tiên $N_j + k - T_2 \ge N_j - k - T_1$ (tức là $T_2 - T_1 \le 2k$). Trường hợp thứ 2 tương ứng với $T_2 - T_1 > 2k$. Trong phạm vi bài báo này, chúng tôi chỉ xét trường hợp $N_j + k - T_2 \ge N_j - k - T_1$, hay $T_2 - T_1 \le 2k$. Trường hợp còn lại đã được chứng minh như trong [6].

Lược đồ trạng thái và hệ phương trình trạng thái cân bằng.

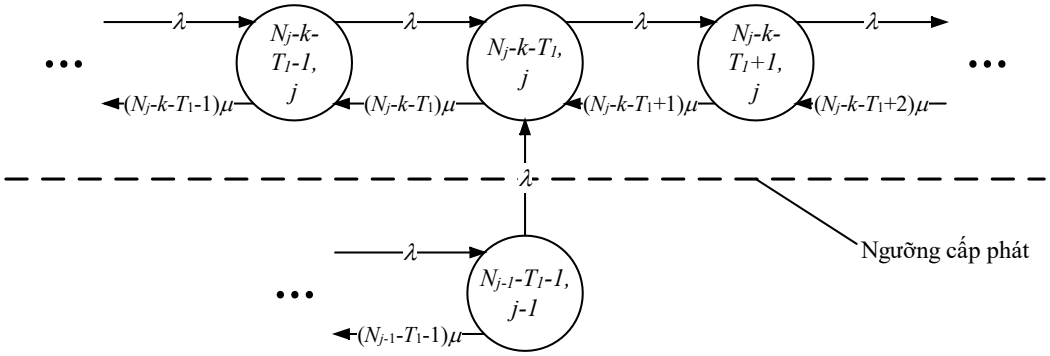
Lược đồ trạng thái của mô hình Q-TS_NB được mô tả ở Hình 3, bao gồm 6 nhóm chuyển trạng thái.



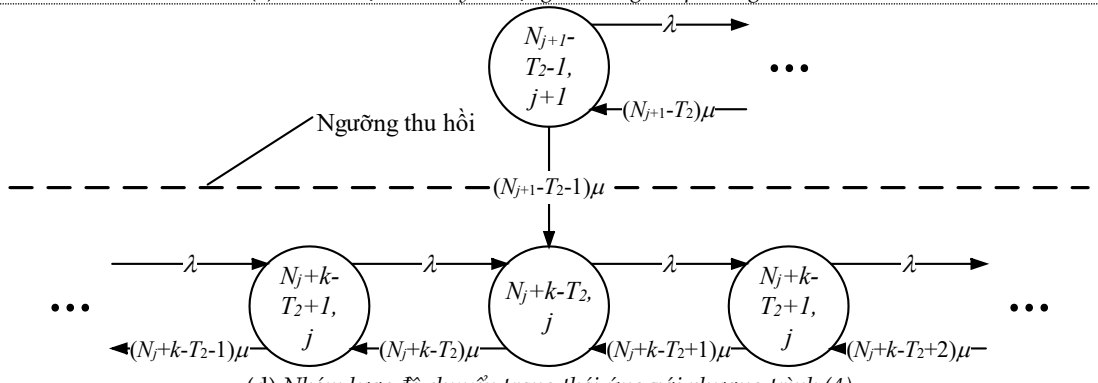
(a) *Nhóm lược đồ chuyển trạng thái ứng với phương trình (1)*



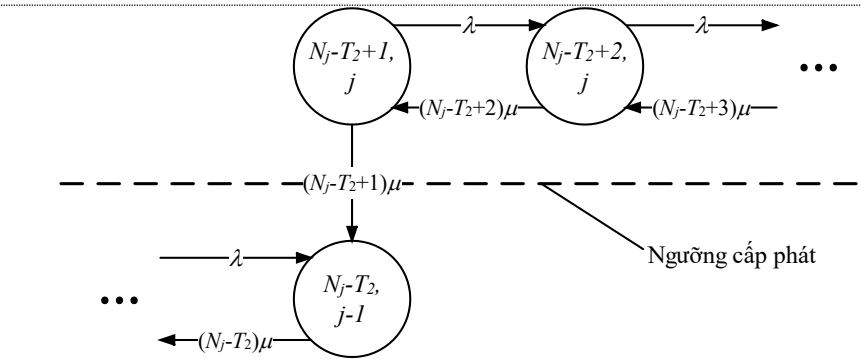
(b) *Nhóm lược đồ chuyển trạng thái ứng với phương trình (2)*



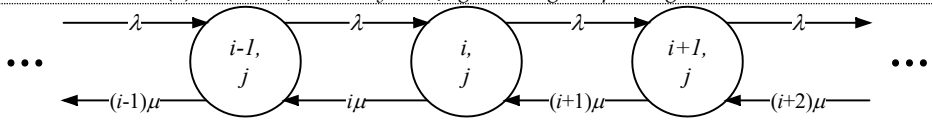
(c) *Nhóm lược đồ chuyển trạng thái ứng với phương trình (3)*



(d) *Nhóm lược đồ chuyển trạng thái ứng với phương trình (4)*



(e) *Nhóm lược đồ chuyển trạng thái ứng với phương trình (5)*



(f) *Nhóm lược đồ chuyển trạng thái ứng với phương trình (6)*

Hình 3. Lược đồ chuyển trạng thái ứng với mô hình Q-TS_NB

Hệ phương trình trạng thái cân bằng ứng với các nhóm lược đồ chuyển trạng thái ở Hình 3 có thể được xây dựng như sau:

$p_{i,L} i\mu = p_{i-1,L}\lambda, (N_L - k - T_1 < i \le N_L)$	(1)
$p_{N_j-T_2+1,j}\left[\lambda + (N_j - T_2 + 1)\mu\right] = p_{N_j-T_2+2,j}(N_j - T_2 + 2)\mu, (M < j \le L)$	(2)
$p_{N_j-k-T_1,j}\left[\lambda + (N_j - k - T_1)\mu\right] = p_{N_j-k-T_1-1,j-1}\lambda + $ $+p_{N_j-k-T_1-1,j}\lambda + p_{N_j-k-T_1+1,j}(N_j - k - T_1 + 1)\mu, (M \le j < L)$	(3)
$p_{N_j+k-T_2,j}\left[\lambda + (N_j + k - T_2)\mu\right] = p_{N_j+k-T_2,j}\lambda + $ $+p_{N_j+k-T_2+1,j}(N_j + k - T_2 + 1)\mu + p_{N_j+k-T_2+1,j+1}(N_j + k - T_2 + 1)\mu, (M < j \le L)$	(4)
$p_{N_j-T_1-1,j}\left[\lambda + (N_j - T_1 - 1)\mu\right] = p_{N_j-T_1-2,j}\lambda, (M \le j < L)$	(5)
$p_{i,j}(\lambda + i\mu) = p_{i-1,j}\lambda + p_{i+1,j}(i+1)\mu,$ $\left(i \notin \begin{Bmatrix}(N_j - T_2 + 1) \cup (N_j - k - T_1) \cup \\ \cup (N_j + k - T_2) \cup (N_j - T_1 - 1)\end{Bmatrix} \middle\| M \le j < L\right)$	(6)

Hệ phương trình từ (1) đến (6) có thể giải bằng cách đặt $p'_{i,j} = \frac{p_{i,j}}{p_{N_L,L}}$ thì khi này $p'_{N_L,L} = 1$ và các phương trình từ (1) đến (6) được viết lại như sau:

$p'_{i,L} i\mu = p'_{i-1,L}\lambda, (N_L - k - T_1 < i \le N_L)$	(7)
$p'_{N_j-T_2+1,j}\left[\lambda + (N_j - T_2 + 1)\mu\right] = p'_{N_j-T_2+2,j}(N_j - T_2 + 2)\mu, (M < j \le L)$	(8)
$p'_{N_j-k-T_1,j}\left[\lambda + (N_j - k - T_1)\mu\right] = p'_{N_j-k-T_1-1,j-1}\lambda + $ $+p'_{N_j-k-T_1-1,j}\lambda + p'_{N_j-k-T_1+1,j}(N_j - k - T_1 + 1)\mu, (M \le j < L)$	(9)
$p'_{N_j+k-T_2,j}\left[\lambda + (N_j + k - T_2)\mu\right] = p'_{N_j+k-T_2,j}\lambda + $ $+p'_{N_j+k-T_2+1,j}(N_j + k - T_2 + 1)\mu + p'_{N_j+k-T_2+1,j+1}(N_j + k - T_2 + 1)\mu, (M < j \le L)$	(10)
$p'_{N_j-T_1-1,j}\left[\lambda + (N_j - T_1 - 1)\mu\right] = p'_{N_j-T_1-2,j}\lambda, (M \le j < L)$	(11)
$p'_{i,j}(\lambda + i\mu) = p'_{i-1,j}\lambda + p'_{i+1,j}(i+1)\mu,$	(12)

$$\left(i \notin \begin{Bmatrix}(N_j - T_2 + 1) \cup (N_j - k - T_1) \cup \\ \cup (N_j + k - T_2) \cup (N_j - T_1 - 1)\end{Bmatrix} \middle| M \le j < L\right)$$

Khi này hệ phương trình (7)-(12) có thể giải bằng cách hồi quy lùi do đã có một giá trị $p'_{N_L,L} = 1$ xác định [11]. Sau khi xác định được các $p'_{i,j}$, dựa vào điều kiện chuẩn hóa $\sum_{(i,j)\in S} p_{i,j} = 1$ chúng ta tính được:

$$p_{N_L,L} = \frac{1}{\sum_{(i,j)\in S} p'_{i,j}}$$

Từ đây ta dễ dàng suy ra các giá trị xác suất $p_{i,j}$ dựa vào công thức $p_{i,j} = p'_{i,j} p_{N_L,L}$.

Các thông số độ đo hiệu năng

Mô hình phân tích ở đây được đánh giá thông qua một số thông số độ đo như sau [6]:

- Số block NSSI trung bình đang được triển khai (bao gồm số block đang được sử dụng và số block rỗi):

$$V_d = \sum_{(i,j)\in S} j p_{i,j} \quad (13)$$

- Số block NSSI trung bình đang được sử dụng (bận):

$$V_b = \sum_{(i,j)\in S} \left\lceil \frac{i}{k} \right\rceil p_{i,j} \quad (14)$$

- Số block NSSI trung bình đang rỗi:

$$V_{id} = V_d - V_b \quad (15)$$

- Hiệu suất sử dụng tài nguyên của hệ thống (tỉ lệ giữa tài nguyên được sử dụng và tài nguyên được cấp phát):

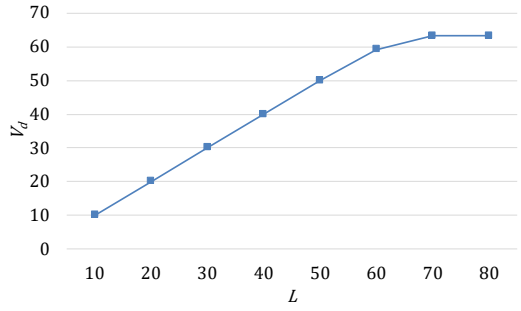
$$U = \sum_{(i,j)\in S} \frac{i}{jk} p_{i,j} \quad (16)$$

4 Phân tích kết quả

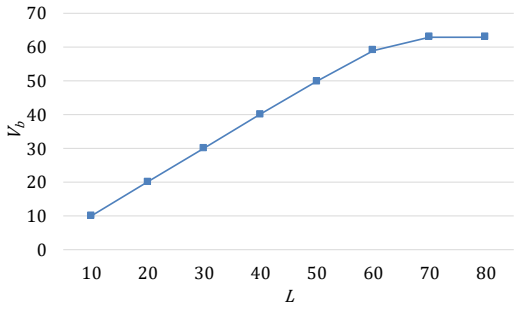
Chúng tôi đánh giá hiệu năng của mô hình theo các thông số độ đo (13) – (16) ở trên khi thay đổi các giá trị ngưỡng phân chia tỉ lệ, các giá trị thông số của mô hình như L, M, r, thông lượng λ/μ.

Cố định các ngưỡng $T_1 = 1$, $T_2 = 11$, thông lượng $\lambda/\mu = 500$, khi này số khối NSSI trung bình và số khối NSSI trung bình đang bận tăng dần (Hình 4. .a) và b)), trong đó số khối NSSI trung bình đang bận bằng khoảng 99,5% trở lên so với L khi L tăng từ 10 đến 60. Trường hợp L

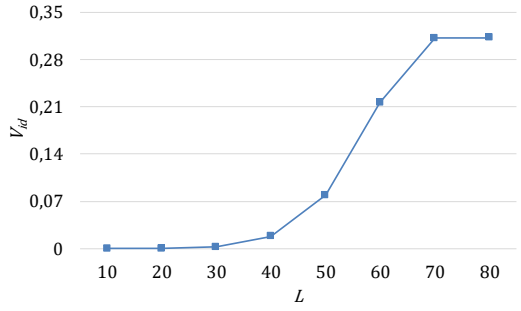
khoảng từ 70 trở lên thì khi đó số lượng NSSI trung bình và số lượng NSSI trung bình đang bận dao động khoảng 63 (Hình 4. .a) và b)). Số khối NSSI trung bình nhàn rỗi từ 0.29×10^{-7} tăng đến 0.31, trong khi đó hiệu năng chỉ giảm từ 0.998 xuống 0.988 (giảm 0.95%) (Hình 4. .c) và d)). Điều này phù hợp với lý thuyết do khi L tăng thì hiển nhiên với cùng tốc độ đến $\lambda/\mu = 500$, số lượng NSSI trung bình nhàn rỗi sẽ tăng lên.



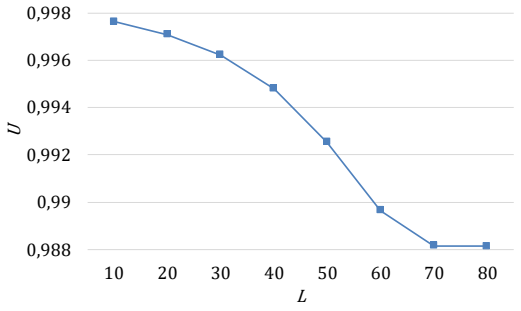
a) Số khối NSSIs trung bình V_d vs L.



b) Số khối NSSIs đang bận trung bình V_b vs L.



c) Số khối NSSIs rỗi trung bình V_{id} vs L.

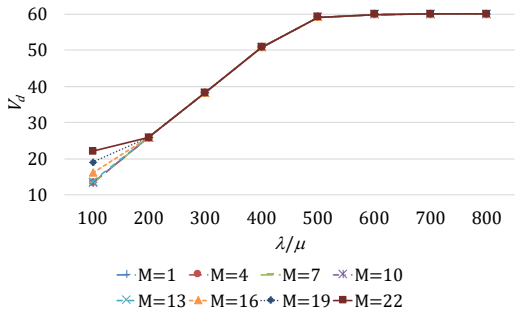


d) Hiệu suất U vs L.

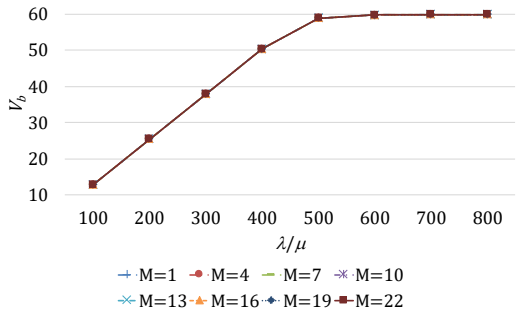
Hình 4. Các kết quả khi $k = r = 8$, $\lambda/\mu = 500$, $T_1 = 1$, $T_2 = 11$ khi thay đổi L.

Với $L = 60$ thì khi này chúng ta thấy rằng khi thông lượng λ/μ còn chưa đủ lớn ($\lambda/\mu < 200$) thì khi này số khối NSSI trung bình, số khối NSSI trung bình nhàn rỗi và hiệu năng có thay đổi đáng kể (số khối NSSI trung bình, số khối NSSI trung bình nhàn rỗi chênh lệch nhau 8.75 và hiệu năng chênh lệch nhau 0.38) (Hình 5.a), c) và d)), nhưng với $\lambda/\mu \geq 200$ thì khi này các thông số độ đo hầu như không thay đổi. Đối với số lượng NSSI trung bình đang bận thì không thay đổi theo các giá trị M như trong Hình 5.b). Điều này có thể là do hệ thống đặt sẵn một số số khối NSSI trung bình nhưng khi này do thông lượng λ/μ lúc đầu đến còn thấp nên khi này số khối NSSI trung bình nhàn rỗi sẽ tăng cao do đối với trường hợp số khối M đặt trước nhiều. Số khối NSSI trung bình nhàn rỗi càng cao đối với M càng lớn khi thông lượng còn thấp là do khi này hệ thống

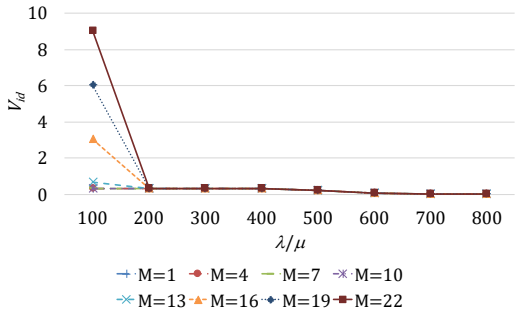
chỉ phục vụ một lượng yêu cầu như nhau nên số khối NSSI trung bình nhàn rỗi sẽ tăng do số khối NSSI đặt trước dư thừa (Hình 5.c)) và dẫn đến hiệu năng sẽ giảm do tài nguyên rỗi tăng nhiều hơn (Hình 5.d)).



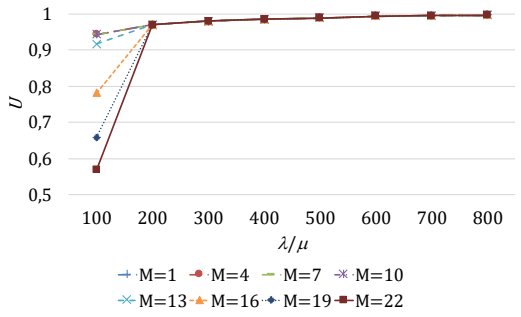
a) Số khối NSSIs trung bình V_a *vs* M.



b) Số khối NSSIs đang bận trung bình V_b *vs* M.



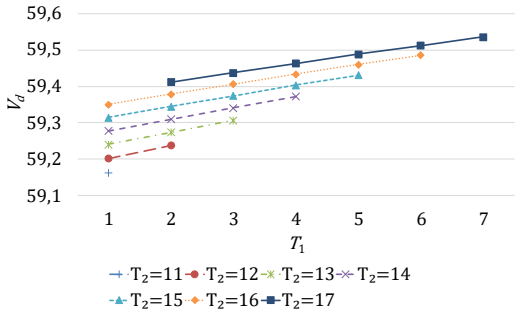
c) Số khối NSSIs rỗi trung bình V_{id} vs M.



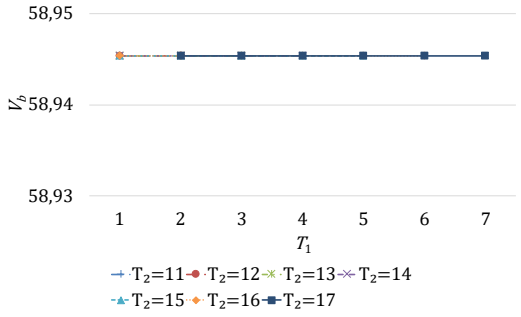
d) Hiệu suất U vs M.

Hình 5. Các kết quả khi $k = r = 8$, $L = 60$, $T_1 = 1$, $T_2 = 11$ theo M khi thay đổi thông lượng λ/μ

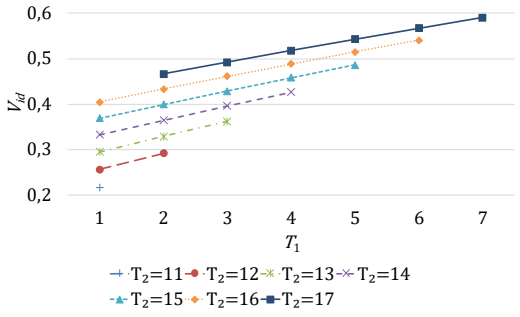
Khi thay đổi các ngưỡng T_1 và T_2, chúng ta nhận thấy rằng $T_1 = 7$ và $T_2 = 17$ sẽ cho kết quả số khối NSSI trung bình cao nhất và với $T_1 = 1$ và $T_2 = 11$ sẽ cho kết quả thấp nhất (Hình 6.a)). Số khối NSSI trung bình đang bận khi này sẽ ổn định và hầu như không có sự thay đổi đối với các trường hợp (Hình 6.b)). Mặt khác, khi này $T_1 = 1$ và $T_2 = 11$ cho kết quả số khối NSSI trung bình nhàn rỗi thấp nhất với 0.22 và đồng thời khi đó hiệu năng cũng đạt cao nhất với 0.99 (Hình 6.c) và d)). Điều này được giải thích là do khi này ngưỡng cấp phát $N_j - T_1 - 1$ là lớn nhất và thu hồi $N_j - T_2 + 1$ cũng lớn nhất dẫn đến số khối NSSI trung bình nhàn rỗi thấp nhất và hiệu năng là cao nhất.



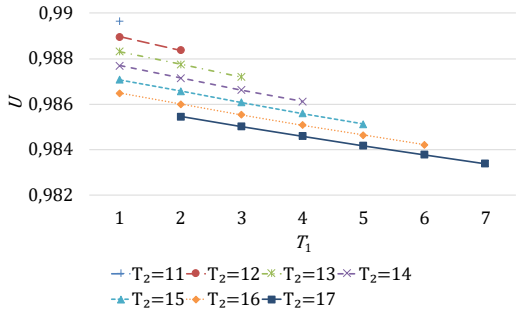
a) Số khối NSSIs trung bình V_d vs T_1 và T_2.



b) Số khối NSSIs đang bận trung bình V_b vs T_1 và T_2.



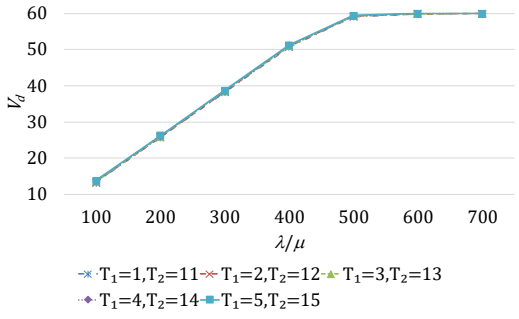
c) Số khối NSSIs rỗi trung bình V_{id} vs T_1 và T_2.



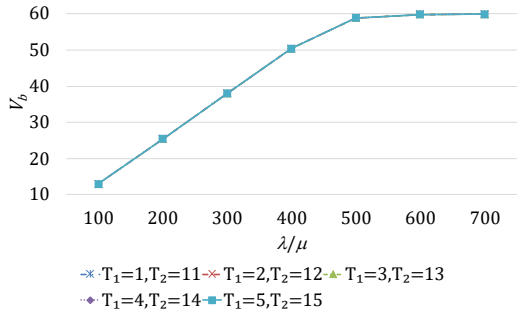
d) Hiệu suất U vs T_1 và T_2.

Hình 6. Các kết quả với $k = r = 8$, $\lambda/\mu = 500$, $M = 1$, $L = 60$ khi thay đổi T_1 và T_2.

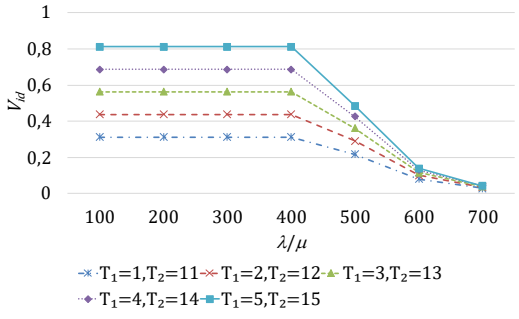
Với $L = 60$ và $k = r = 8$ và thay đổi thông lượng λ/μ từ 100 đến 700 thì lúc này số lượng NSSI trung bình và số lượng NSSI trung bình đang bận tăng dần và tiệm cận 60 (Hình 7). Khi này số lượng NSSI trung bình nhàn rỗi và hiệu năng đối với trường hợp $T_1 = 1$ và $T_2 = 11$ cũng là trường hợp tốt nhất theo từng giá trị thông lượng λ/μ do khi này ngưỡng cấp phát và thu hồi là lớn nhất (tốt nhất) do khi này ngưỡng cấp phát $N_j - T_1 - 1$ là lớn nhất và thu hồi $N_j - T_2 + 1$ cũng lớn nhất dẫn đến số khối NSSI trung bình nhàn rỗi thấp nhất và hiệu năng là cao nhất.



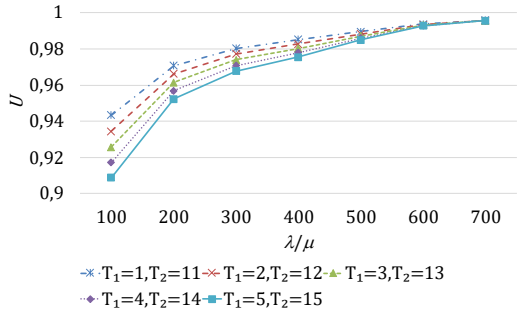
a) Số khối NSSIs trung bình V_d vs λ/μ.



b) Số khối NSSIs đang bận trung bình V_b vs λ/μ.



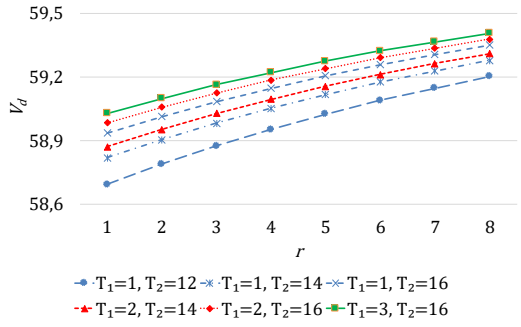
c) Số khối NSSIs rỗi trung bình V_{id} vs λ/μ.



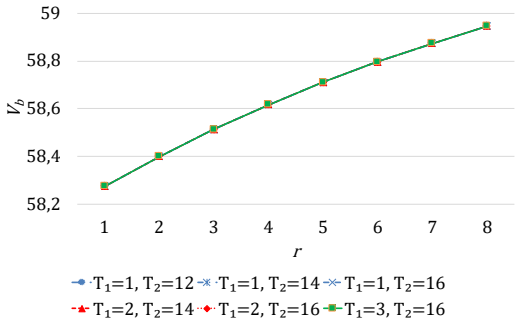
d) Hiệu suất U vs λ/μ.

Hình 7. Các kết quả khi $k = r = 8$, $L = 60$, $M = 1$, $T_2 = 11$ theo thông lượng λ/μ

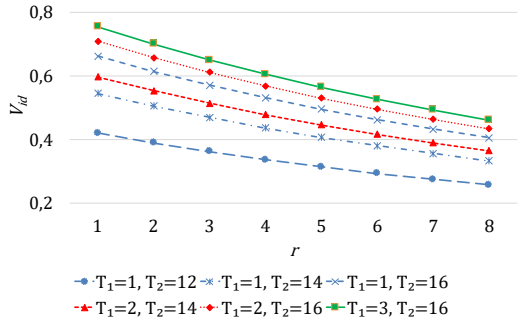
Khi số lượng NFI trong NSSI cuối r thay đổi $(1 \le r \le k)$, $k = 8$ và $T_1 = 1$, $T_1 = 2$ hoặc $T_1 = 3$ và $T_2 = 12$, $T_2 = 14$ hoặc $T_2 = 16$, với cùng T_2 thì khi này T_1 nào nhỏ hơn sẽ cho số lượng NSSI trung bình nhàn rỗi và hiệu năng tốt hơn (Hình 8) do khi này ngưỡng cấp phát $N_j - T_1 - 1$ là lớn nhất và thu hồi $N_j - T_2 + 1$ cũng lớn nhất dẫn đến số khối NSSI trung bình nhàn rỗi thấp nhất và hiệu năng là cao nhất. Đồng thời khi xét r thay đổi chúng ta cũng nhận thấy rằng số NSSI trung bình nhàn rỗi sẽ tỷ lệ nghịch và hiệu năng tỷ lệ thuận khi tăng r từ 1 đến k. Điều này được giải thích là do nếu $r < k$ thì khi này hệ thống sẽ phải bật thêm NSSI cuối nếu một vài yêu cầu NFI tăng thêm so với trường hợp $r = k$ phải yêu cầu tất cả NFI cuối đến thì mới bật lên.



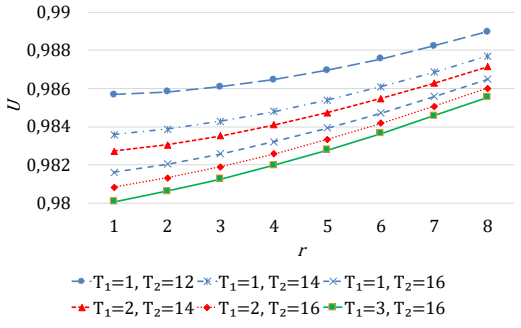
a) Số khối NSSIs trung bình V_d vs r.



b) Số khối NSSIs đang bận trung bình V_b vs r.



c) Số khối NSSIs rỗi trung bình V_{id} vs r.



d) Hiệu suất U vs r.

Hình 8. Các kết quả khi $k = 8$, $L = 60$, $M = 1$, $\lambda/\mu = 500$ theo r

5 Kết luận

Trong bài báo này, chúng tôi mô hình hóa vấn đề phân chia mạng dựa theo 3GPP R15 với việc mô hình hóa bài toán phân chia tỉ lệ theo các khối NSSI (bao gồm một số các phiên bản chức năng mạng) cho mạng lõi 5G. Mô hình đề xuất tích hợp các ngưỡng để điều khiển, kiểm soát việc phân chia tỉ lệ các khối NSSI, cụ thể với thiết lập/tắt các khối một cách linh động theo lưu lượng đến, hỗ trợ cấp phát tài nguyên một cách hiệu quả, chi phí hợp lý, bằng việc đưa vào các ngưỡng thiết lập (T_1) và tắt (T_2) khối NSSI theo Thuật toán 1. Chúng tôi đã cung cấp một giải pháp hiệu quả để tính toán các xác suất trạng thái cân bằng và từ đó tính toán các thông số đo hiệu suất của mô hình. Kết quả số cho thấy mô hình phân chia tỉ lệ dựa trên ngưỡng có thể tự động điều chỉnh số lượng khối NSSI để đáp ứng với sự thay đổi của tải lưu lượng, tiết kiệm mức tiêu thụ tài nguyên và duy trì mức sử dụng cao tài nguyên được yêu cầu. Mô hình đề xuất của chúng tôi trong bài báo này cũng xem xét, đánh giá vấn đề đặt trước (thiết lập sẵn) một giá trị M khối NSSI

nhất định để có thể luôn sẵn sàng phục vụ người dùng. Kết quả cho thấy, khi tải thấp, việc thiết lập M lớn sẽ dẫn đến hiệu suất sử dụng tài nguyên giảm, ngược lại khi tải cao, nhờ điều khiển và kiểm soát scaling một cách hiệu quả, nên hiệu suất hệ thống không bị ảnh hưởng đáng kể bởi M. Trong bài báo, giả định về quá trình đến Poisson là một phương pháp phổ biến để làm cho mô hình hàng đợi có thể xử lý được về mặt toán học. Trong tương lai, chúng tôi sẽ xem xét việc áp dụng các quá trình ngẫu nhiên khác để mô hình hóa quá trình đến non-Poisson, cũng như áp dụng các giải pháp dựa trên học tăng cường ([9-10]) vào mô hình để thu được kết quả tối ưu hơn.

Acknowledgments. This work was supported by the Ministry of Education and Training (Vietnam) for the development of Science and Technology under grant number B2023-DHH-17.

Tài liệu tham khảo

1. Cheng-Ying Hsieh, Tuan Phung-Duc, Yi Ren, Jyh-Cheng Chen (2022), "Design and Analysis of Dynamic Block-setup Reservation Algorithm for 5G Network Slicing", IEEE Transactions on Mobile Computing, pp. 1536-1233. DOI 10.1109/TMC.2022.3169034
2. 5GPP, "View on 5G Architecture," White paper 22.891,5GPPP Architecture Working Group, 07 2018. Version 14.2.0.
3. 3GPP, "5G; Study on scenarios and requirements for next generation access technologies," Technical Specification (TS) 38.913, 3rd Generation Partnership Project (3GPP), 07 2020. Version 16.0.0 Release 16.
4. 3GPP, "Technical Specification Group Services and System As- pects; System architecture for the 5G System (5GS); Stage 2," Technical Specification (TS) 23.501, 3rd Generation Partnership Project (3GPP), 12 2020. Version 16.7.0.
5. 3GPP, 3rd Generation Partnership Project "The Mobile Broadband Standard", Available: https://www.3gpp.org/.
6. C. Rotter and T. V. Do, "A queueing model for threshold-based scaling of UPF instances in 5G core," IEEE Access, vol. 9, pp. 81443–81453, 2021.
7. K. Yves Adou and Ekaterina V. Markova, "Methods for Analyzing Slicing Technology in 5G Wireless Network Described as Queueing System with Unlimited Buffer and Retrial Group", A. Dudin et al. (Eds.): ITMM 2020, CCIS 1391, pp. 264–278, 2021. https://doi.org/10.1007/978-3-030-72247-0_20.
8. Irina Kochetkova, Anastasia Vlaskina, Sofia Burtseva, Valeria Savich, and Jiri Hosek, "Analyzing the Effectiveness of Dynamic Network Slicing Procedure in 5G Network by Queuing and Simulation Models", O. Galinina et al. (Eds.): NEW2AN 2020/ruSMART 2020, LNCS 12525, pp. 71–85, 2020. https://doi.org/10.1007/978-3-030-65726-0_7.
9. Nguyen HT, Tien Van Do, Hegyi A, Rotter C. An Approach to Apply Reinforcement Learning for a VNF Scaling Problem. 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN). 2019 Feb 1. doi: 10.1109/ICIN.2019.8685866.
10. Nguyen, Hai T., Do, TV, Rotter, "Scaling UPF Instances in 5G/6G Core with Deep Reinforcement Learning", IEEE ACCESS 9 pp. 165892-165906. , 15 p. (2021).

11. Shah, K, Sinha, B. Theory of Optimal Designs. Springer New York; 2012.