

GIẢM TẢI ĐA MỤC TIÊU TRONG ĐIỆN TOÁN BIÊN DI ĐỘNG

Nguyễn Thanh Trà

Khoa Viễn thông, Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Điện toán biên di động (Mobile edge Computing) là một trong các trụ cột quan trọng của mạng 5G nhằm cung cấp một tiếp cận điều khiển phân tán tương thích với các nhu cầu dịch vụ mới nổi. Một trong các ưu điểm vượt trội của điện toán biên nhờ vào quá trình điều phối hợp lý nhu cầu tính toán của các thiết bị người dùng tới các thiết bị cục bộ và máy chủ biên mạng. Bên cạnh các lợi điểm của quá trình giảm tải (offloading), thách thức lớn nhất của tiếp cận này là vấn đề lựa chọn cơ chế giảm tải tối ưu thích hợp trong bài toán tối ưu đa mục tiêu do các ràng buộc đặt bài toán này thuộc dạng NP-Hard. Bài báo này trình bày một cơ chế giảm tải cân bằng giữa năng lượng tiêu thụ và yêu cầu trễ của thiết bị người dùng. Kết quả mô phỏng số với các giả thiết về kênh truyền vô tuyến thực tiễn gồm kết nối MIMO và 802.11p cùng với các mẫu lưu lượng đầu vào khác nhau cho thấy các tác động tới hiệu năng của cơ chế giảm tải từng phần phân mà các giải pháp khác chưa đề cập.

Từ khóa: Điện toán biên di động, giảm tải, tối ưu đa mục tiêu, MIMO.

I. GIỚI THIỆU

Sự phát triển mạnh mẽ của các giải pháp Internet vạn vật trong thập kỷ qua đã và đang tạo ra các áp lực lên hạ tầng mạng truyền thông. Bên cạnh sự gia tăng về số lượng dữ liệu và thiết bị, các yêu cầu khắt khe hơn về chất lượng dịch vụ của các ứng dụng cũng đã được đặt ra [1] [2]. Vì vậy, nhu cầu xử lý tính toán tại gần thiết bị người sử dụng UE (User Equipment) đã đem đến giải pháp công nghệ điện toán biên EC (Edge Computing) [3]. Giải pháp điện toán biên được đề xuất nhằm vượt qua các giới hạn năng lực tính toán của thiết bị người dùng và giảm thiểu thời gian trễ so với điện toán đám mây hay điện toán sương mù. Các nhiệm vụ hoặc thông tin cần xử lý cục bộ sẽ được tính toán phân tán tại biên mạng, trong khi nhu cầu xử lý toàn cục sẽ được gửi lên đám mây. Đặc biệt, cùng với các công nghệ mạng định nghĩa bằng phần mềm SDN (Software Defined Network) và ảo hóa chức năng mạng NFV (Network Function Virtualization), điện toán biên di động đóng vai trò trụ cột trong thể hệ mạng 5G và phía sau [4].

Bên cạnh các ưu điểm của điện toán biên, một số các thách thức liên quan tới bảo mật và hiệu năng mạng cũng đã được chỉ ra. Do kiến trúc điều khiển phân tán và giới hạn tài nguyên của các thiết bị biên, hiệu năng của điện toán biên phụ thuộc vào rất nhiều yếu tố ràng buộc như:

quyết định giảm tải, phân bổ tài nguyên tính toán giữa các thành phần tham gia, hay tính động của các tham số theo kích bản thực tế. Các điều kiện thực rất đa dạng và tạo ra các giới hạn lên các giải pháp nâng cao hiệu năng và bảo mật của điện toán biên.

Đặc biệt, công việc giảm tải là một quá trình phức tạp và có thể bị ảnh hưởng bởi một số yếu tố khác nhau. Cụ thể, quá trình này liên quan đến việc phân vùng ứng dụng, ra quyết định giảm tải và thực thi tác vụ phân tán [5]. Vì vậy, các nghiên cứu và triển khai kích bản điện toán biên đã và đang tiếp tục hướng tới việc tìm kiếm các giải pháp giảm tải hiệu quả nhằm tiếp cận gần nhất tới các điều kiện thực tiễn [6].

Một trong các trở ngại chính liên quan tới chiến lược giảm tải xuất phát từ độ phức tạp của bài toán tối ưu. Các tiếp cận tìm đường biên giải tích thông qua các phương pháp quy hoạch toán học thường chỉ đưa đến các giới hạn biên với thời gian tính toán lớn [7]. Một trong các trở ngại thách thức chính liên quan tới các ràng buộc phi tuyến và trái ngược nhau về lợi ích. Mặt khác, Tiếp cận heuristic đem lại một số hiệu quả nhất định trong tính toán nhưng cũng hiệu quả phụ thuộc mạnh vào các giả thiết đầu vào cùng như thông tin dữ liệu lịch sử [6] [7]. Vì vậy, các nỗ lực tìm kiếm các giải pháp giảm tải hiệu quả trong các kích bản điển hình vẫn đang thu hút nhiều nghiên cứu gần đây.

Do sự phức tạp của các tính toán giảm tải tối ưu đa mục tiêu mà tính chất đường truyền 5G chưa được đề cập tới trong các nghiên cứu trước đây. Vì vậy, nhằm bù vào sự thiếu hụt đó để hướng tới các kích bản ứng dụng hữu ích, bài báo này trình bày một cơ chế giảm tải tối ưu hai mục tiêu đồng thời gồm năng lượng tiêu hao của thiết bị người dùng UE và đáp ứng trễ cho các phiên giảm tải. Đặc biệt, lược đồ sử dụng UE lân cận để trợ giúp một phần khối lượng tính toán nhằm đạt được cân bằng năng lượng tổng thể các UE sẽ mở ra các giải pháp phân cụm hiệu quả. Đây là một tiếp cận mà các nghiên cứu trước đây chưa đề cập tới. Cơ chế giảm tải từng phần nhiệm vụ được mô phỏng số để tìm ra các khoảng giá trị hiệu quả đối với một số mô hình lưu lượng đầu vào của các dịch vụ IoT (Internet of Things) điển hình.

Bài báo có bố cục như sau. Mục tiếp theo sẽ trình bày về các nghiên cứu liên quan tới quyết định giảm tải của các đề xuất gần đây cho điện toán biên. Trong đó chỉ rõ các điểm mạnh và điểm yếu của các đề xuất và những vấn đề chưa đặt ra trong nghiên cứu. Mô hình tính toán giảm tải được đề xuất trong mục tiếp theo trên cơ sở bài toán tối ưu với các ràng buộc về kênh truyền thông vô tuyến. Kết quả mô phỏng số cho một số trường hợp lưu lượng điển hình được chỉ ra nhằm chỉ ra các điểm mạnh của đề xuất. Mục cuối cùng tóm tắt lại các kết quả đạt được của nghiên cứu và hướng phát triển tiếp theo.

Tác giả liên hệ: Nguyễn Thanh Trà

Email: trant@ptit.edu.vn

Đến tòa soạn: 9/2021, chỉnh sửa: 10/2021, chấp nhận đăng: 10/2021.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Vấn đề giảm tải trong điện toán biên di động MEC (Mobile Edge Computing) được tiếp cận từ nhiều góc độ khác nhau. Trong đó, các thiết bị người sử dụng UE được trang bị nhiều kết nối vô tuyến nhằm tổ chức liên kết thành kiến trúc mạng biên mềm dẻo. Dưới góc độ kích bản ứng dụng vào các dịch vụ IoT, các lược đồ giảm tải ứng dụng vào trong các hình thức như: kết nối từ 01 UE tới 01 máy chủ MEC, 01 UE tới nhiều máy chủ MEC, nhiều UE tới 01 máy chủ MEC và nhiều UE kết nối tới nhiều máy chủ MEC. Mặt khác, các mục tiêu tối ưu cũng rất khác biệt và hướng theo nhu cầu ứng dụng điện toán biên như: độ trễ, năng lượng, băng thông/ tần số, cân bằng tải hay giá triển khai. Trong đó, các quyết định giảm tải được phân chia thành hai chiến lược gồm giảm tải hoàn toàn (còn gọi là giảm tải nhị phân) và giảm tải từng phần [8].

Với tiếp cận giảm tải nhị phân, một khối nhiệm vụ của UE lựa chọn thực thi trên chính UE hoặc trên máy chủ biên. Trên thực tế, các quyết định dạng này với một hàm mục tiêu thì có thể đạt được tối ưu toàn cục trong một số điều kiện nhất định. Thông thường, mục tiêu tiết kiệm năng lượng của UE sẽ được tối đa hóa nếu không đề cập tới các vấn đề khác như năng lượng tiêu tán khác như công suất truyền dẫn. Mặc dù vậy, với số lượng UE lớn và các ràng buộc bổ sung cũng dẫn tới độ phức tạp của thuật toán cao tới 2^N , với N là số lượng các thiết bị người dùng trong mạng. Hơn nữa, sự cân bằng tải của các UE sẽ không được tính đến khi các thực thi tính toán được thực hiện trên từng phần tử cục bộ và tiếp cận này làm giảm thời gian sống của toàn mạng [9] [10].

Giảm tải từng phần thường có độ phức tạp cao hơn cho các quyết định khi đối diện với nhiều điều kiện ràng buộc hơn, nhưng mang lại nhiều lợi ích như khả năng xử lý song song và phân biệt chất lượng dịch vụ. Trong đó, các tác vụ có thể phân chia theo lượng dữ liệu, theo loại dữ liệu hay theo tính chất dữ liệu. Nhờ vậy, khả năng hỗ trợ chất lượng dịch vụ cho các UE cũng sẽ được mở rộng. Các giải pháp đề xuất quyết định giảm tải rất đa dạng tùy thuộc vào kịch bản ứng dụng, mục tiêu tối ưu, phương pháp áp dụng hay mô hình chuyển động. Hướng về mục tiêu đảm bảo an toàn và hiệu quả năng lượng cho thiết bị di động, trong [11] đề xuất lược đồ giảm tải cho các UAV (Unmanned Aerial Vehicle). Nhóm tác giả đã đề xuất chiến lược phân cụm nhằm cải thiện năng lượng và độ tin cậy cho các thiết bị IoT. Trong đề xuất này, các giao tiếp ngang hàng giữa các thiết bị người dùng không được đề cập. Nhằm cân bằng tải tính toán trong giới hạn trễ, các tác giả trong [12] sử dụng hàm giá (Cost) có tổng trọng số bằng 1 để xác định mức ưu tiên trong quyết định giảm tải. Các UE được phân cụm dựa trên thuật toán di truyền GA (Genetic Algorithm) để chọn UE chịu trách nhiệm giảm tải. Đây là một lược đồ quyết định rất đáng chú ý do khả năng thích ứng với các yêu cầu đa dạng của tác vụ tính toán. Tuy nhiên, các tham số đường truyền vô tuyến không đề cập chi tiết ở đây và mô hình phân cụm chỉ thích hợp với mật độ rất cao.

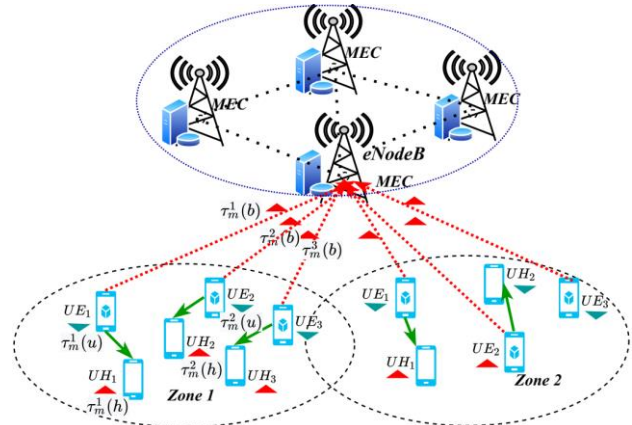
Nhằm tối ưu năng lượng cho các UE, các tác giả trong [13] sử dụng phương pháp quy hoạch tuyến tính để tìm kiếm các tỷ lệ giảm tải hợp lý. Giải pháp lựa chọn nút UE lân cận mang tới sự mềm dẻo trong truyền thông ngang hàng của cả mô hình tĩnh và động. Tuy nhiên, bên cạnh các tác động cụ thể của đường truyền vô tuyến chưa được

đề cập thì các mẫu lưu lượng đầu vào còn rất hạn chế so với điều kiện thực.

Vì vậy, trong nghiên cứu này, một lược đồ giảm tải hướng tới mục tiêu cân bằng năng lượng và trễ tính toán được đề xuất trong các điều kiện kênh truyền vô tuyến thực tiễn để bù đắp vào các thiếu hụt trong các nghiên cứu trên.

III. MÔ HÌNH ĐỀ XUẤT

A. Mô hình mạng và giả thiết chung



Hình 1. Mô hình giảm tải điện toán biên

Một mô hình giảm tải đề xuất được trình bày trên hình 1. Trong đó, tập các thiết bị người dùng UE được phân vùng theo dung lượng cung cấp từ eNodeB, thiết bị lựa chọn giảm tải D2D được ký hiệu là UH (User Helper). Các vùng (Zone 1, Zone 2, ..., Zone Z) được xác định theo vùng phủ của các trạm gốc (eNodeB) hay các chiến lược phân cụm. Để không mất tính tổng quát, trong mô hình này sử dụng vùng phủ như một tham số cơ bản cho bài toán phân cụm. Các tham số đường truyền vô tuyến 5G sẽ được giả thiết với các điều kiện thông dụng và phổ biến nhất theo các công nghệ hiện hành.

Mỗi UE có một tập M tác vụ cần được tính toán, ta đặt ký hiệu $\Gamma = \{\tau_1, \tau_2, \dots, \tau_M\}$, $m \in M$. Trong một nhóm các thiết bị người dùng, các phiên kết nối đồng thời tương ứng với nhiệm vụ thứ m được đánh số theo chỉ số tương ứng. Ví dụ, ký hiệu là τ_m^1 cho biết số thứ tự thuộc phiên thứ nhất của nhiệm vụ thứ m trong luồng tác vụ yêu cầu đầu vào. Mỗi tác vụ trên một node UE thứ i gồm là một cặp tham số yêu cầu gồm độ dài dữ liệu cần xử lý và thời gian yêu cầu, $m(i) \leq l_i, t_i^* > 0$, l_i là độ dài dữ liệu đầu vào tính theo đơn vị bit.

Mỗi một tác vụ τ_m của một UE được tính toán tại chính UE, nút hỗ trợ UH và máy chủ MEC, tương ứng $\tau_m(u)$, $\tau_m(h)$ và $\tau_m(b)$. Ta có,

$$\tau_m(UE) = \tau_m(u) + \tau_m(h) + \tau_m(b) \quad (1)$$

Công thức trên biểu diễn tiếp cận giảm tải từng phần thông qua thiết bị hỗ trợ UH. Để đảm bảo tính tổng quát, UH và UE có các thông số tương tự nhau và thay đổi vai trò theo tiêu chí cân bằng năng lượng tiêu hao giữa các nút.

B. Mô hình kênh truyền thông

Giả thiết các vùng phủ của 5G tạo thành các vùng zone và các thiết bị UE truy nhập tới máy chủ MEC theo cơ chế đa truy nhập phân chia theo tần số trực giao OFDMA (Orthogonal Frequency Division Multiple Access), tốc độ truyền được tính theo định lý Shannon-Hartley theorem như dưới đây.

$$r_{i,b}^{ub} = \frac{W_{b_j}}{N} \log_2 \left(1 + \frac{p_{ub} \cdot h_{ub}, b_j}{\omega_0 + \sum_{k=1, k \neq i}^N p_{ub} \cdot h_{ub}, b_k} \right) \quad (2)$$

Trong đó, W_{b_j} là độ rộng băng tần cấp cho các vùng với N thiết bị UE, p_{ub} và h_{ub} là công suất phát của UE và độ lợi kênh. ω_0 là nhiễu nền.

BẢNG 1. CÁC KÝ HIỆU VÀ THAM SỐ

$l_{i,u}$	Lượng bit cần xử lý tại UE của nhiệm vụ i
$l_{i,h}$	Lượng bit cần xử lý tại UH của nhiệm vụ i
$l_{i,b}$	Lượng bit cần xử lý tại eNodeB của nhiệm vụ i
C	CPU Cycles require to process 1 bit
f_u	Tần số CPU tại UE và UH
F_b	Tần số CPU tại eNodeB
k	Hằng số điện dung
$r_{i,b}^{ub}$	Tốc độ truyền dẫn từ UE tới eNodeB (5G)
$r_{i,h}^{uh}$	Tốc độ truyền dẫn từ UE tới eNodeB (802.11p)
p_{ub}	Công suất truyền từ UE tới eNodeB

C. Mô hình xử lý tác vụ

Với mỗi tác vụ τ_m ta có $l_i(m)$ bit cần xử lý và được phân bổ tại UE, UH và MEC.

$$l_i(m) = l_{i,u}(m) + l_{i,h}(m) + l_{i,b}(m) \quad (3)$$

Nếu nhiệm vụ chỉ được xử lý tại UE, năng lượng tiêu hao và trễ được tính đơn giản thông qua tốc độ xử lý của CPU.

$$E_{i,u}^{comp} = l_{i,u} \cdot C \cdot k \cdot f_u^2 \quad (4)$$

$$t_{i,u}^{comp} = \frac{l_{i,u} \cdot C}{f_u} \quad (5)$$

Trong đó, $E_{i,u}^{comp}$ là năng lượng tiêu hao tính toán cho $l_{i,u}$ bit, f_u là tần số của CPU, k là hằng số điện dung, C là chu kỳ CPU cần thiết để xử lý tác vụ tại các bộ vi xử lý [14]. Các bộ tham số giả thiết được trình bày trên bảng 1.

Với giả thiết UE và UH thực hiện đóng vai trò luân phiên nhau trong các phiên truyền thông, năng lượng và trễ xử lý tại UH được tính tương đương như trên công thức (4), (5) trên đây.

Khi xử lý một phần tác vụ tại máy chủ điện toán biên ở eNodeB và một phần tại thiết bị trợ giúp UH, thời gian xử lý được tính toán bằng thời gian truyền dẫn và thời

gian xử lý tác vụ. Giả thiết tốc độ xử lý tại MEC cao hơn nhiều so với tại UE, UH và ta có thể bỏ qua, cũng như thời gian nhận thông tin kết quả từ các phần tử mạng này. Ta có, trễ truyền thông từ UE tới eNodeB và từ UE tới UH được tính như sau:

$$t_{i,u}^{ub} = \frac{l_{i,b}}{r_{i,b}^{ub}} \quad (6)$$

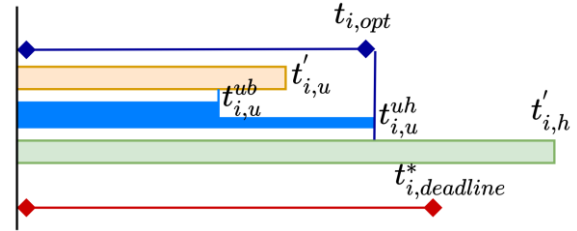
$$t_{i,u}^{uh} = \frac{l_{i,u}}{r_{i,u}^{uh}} \quad (7)$$

Năng lượng tiêu hao dành cho truyền dẫn phụ thuộc vào số lượng bit truyền và công suất truyền theo bit lên eNodeB và UH tương ứng là p_{ub} và p_{uh} . Ta có,

$$E_{i,b}^{trans} = \frac{l_{i,b}}{r_{i,b}^{ub}} \times p_{ub} \quad (8)$$

$$E_{i,h}^{trans} = \frac{l_{i,u}}{r_{i,u}^{uh}} \times p_{uh} \quad (9)$$

Các ràng buộc trễ của việc thực hiện các tác vụ theo yêu cầu được biểu diễn trên hình 2 dưới đây. Với tác vụ τ_m bất kỳ, thiết bị UE có thể thực hiện giảm tải hoàn toàn lên MEC (full offloading), mức năng lượng tiêu hao là lớn nhất đối với UE, nhưng đánh đổi bởi thời gian xử lý tác vụ thấp nhất, $t_{i,u}^{ub}$. Trái ngược với giải pháp trên, việc xử lý tại UH có độ trễ lớn nhất do trễ đường truyền và trễ xử lý tại UH, $t_{i,u}^{uh}$. Vì vậy, chiến lược giảm tải từng phần rất quan trọng đối với các mục tiêu tổng quát.



Hình 2. Các ràng buộc trễ

Thời gian trễ đối với tác vụ giảm tải sang UH được tính như sau:

$$t_{i,uh} = t_{i,u}^{uh} + t_{i,u}^{comp} = \frac{l_{i,u}}{r_{i,u}^{uh}} + \frac{l_{i,u} \cdot C}{f_u} \quad (10)$$

D. Mô hình hóa bài toán

Tiếp cận cân bằng trễ và năng lượng tiêu hao đã được đưa ra trong [15] [16]. Tiếp cận này cho thấy sự thích ứng tốt đối với các ứng dụng IoT đa mục tiêu với khả năng mở rộng hiệu quả. Vì vậy, trong nghiên cứu này sẽ sử dụng tiếp cận trên để tính hàm giá (Costs) cho từng phiên i thực hiện tác vụ τ_m với các trọng số điều chỉnh như dưới đây.

$$c_i^m = \alpha \cdot E_i + \beta \cdot T_i, \alpha + \beta = 1 \quad (11)$$

Như vậy, từ mục tiêu trên công thức (11) ta đi tìm một lời giải tối ưu của các phiên kết nối đảm bảo điều kiện ràng buộc trễ.

$$t_{i,uh} \leq t_i^* \quad (12)$$

Hàm Cost cho tổng số các phiên i đồng thời trong một vùng phủ của eNodeB được tính như dưới đây.

$$C_M \sum_{i \in M} c_i \quad (13)$$

Mục tiêu của bài toán là tối thiểu đồng thời năng lượng và trễ cho các phiên làm việc. Ta biểu diễn qua bài toán tối ưu như sau:

$$\min_M C_M = \min_M \sum_{i=1}^M \alpha E_i + \beta T_i \quad (14)$$

Ràng buộc

$$C1: l_i = l_{i,u} + l_{i,h} + l_{i,b}$$

$$C2: u_i \in \{1, 2, \dots, n\}, \forall u_i \in N$$

$$C2: \tau_i \in \{1, 2, \dots, m\}, \forall m \in M$$

$$C3: t_{i,uh} \leq t_i^*$$

$$C4: t_{i,ub} \leq t_i^*$$

Các ràng buộc của bài toán như sau: C1 là ràng buộc về phân chia từng phần của tác vụ với tổng số phần phân chia bằng tổng tác vụ cần xử lý; ràng buộc C1 chỉ ra giới hạn các UE mà một eNodeB có thể cung cấp trong vùng phủ. Dung lượng và tốc độ liên kết phụ thuộc vào số lượng node trong vùng phủ qua điều kiện C2; ràng buộc C3 giới hạn số các nhiệm vụ đồng thời trong phiên; hai ràng buộc C3 và C4 là đáp ứng giới hạn trễ của các tác vụ.

IV. MÔ PHỎNG VÀ THẢO LUẬN

Để thực hiện tìm kiếm lời giải tối ưu cho (14), ta đưa các tham số truyền thông vô tuyến và các điều kiện biên của kênh kết nối. Với mô hình đề xuất, truyền thông vô tuyến từ eNodeB tới các UE theo đặc tính của kênh truyền MIMO 5G, truyền thông giữa UE và UH theo chuẩn 802.11p. Các giả thiết về kênh truyền MIMO được trình bày dưới đây là các tiêu chuẩn tổng quát cho các kênh truyền được sử dụng trong các nghiên cứu.

Hệ thống hoạt động trong chế độ song công TDD (Time Division Duplex), với khung truyền dẫn vô tuyến gồm pilot và dữ liệu. Kênh truyền được coi là đối xứng hoàn hảo và tuân theo mô hình fading Rician không tương quan không gian. Tốc độ kênh truyền dẫn từ UE tới MEC trong công thức (2) phụ thuộc vào tỷ số công suất tín hiệu trên tổng công suất nhiễu trong các điều kiện thực của 5G MIMO.

Với giả thiết các máy chủ MEC có khả năng phục vụ được mô hình hóa qua hàng đợi M/M/1 [16]. Các thiết bị đầu cuối có khả năng chọn được máy chủ thích hợp và tạo cụm giảm tải theo nguyên tắc xác suất và thể hiện khái quát trên hình 1. Ta xác định các yếu tố tác động kênh như dưới đây.

Một khung truyền dẫn $\tau_{i,u}^{ub}(i)$ từ User tới MEC gồm hai phần huấn luyện $\tau_{i,pilot}^i$ và ước lượng hệ số kênh truyền $\tau_{i,u}^{ub}$;

$$\tau_{i,u}^{ub}(i) = \tau_{i,pilot}^i + \tau_{est}^i \quad (15)$$

Kênh truyền tuân theo mô hình fading Rician không tương quan về không dây. Hệ số suy hao truyền theo kênh được tính theo tầm nhìn thẳng LoS và theo không có tầm nhìn thẳng NLoS là.

$$\beta_{u,los}^{rician} = \sqrt{\frac{R_{rician}}{R_{rician} + 1}} \times \beta_u; \beta_{u,nlos}^{rician} = \sqrt{\frac{1}{R_{rician} + 1}} \times \beta_u \quad (16)$$

Mảng antenna giả thiết được phân bố tuyến tính đều ULA (Uniform Linear Array), vector hệ số kênh truyền phạm vi nhỏ tuân theo phân bố Gaussian chuẩn tắc, thiết bị người dùng sử dụng kỹ thuật ước lượng kênh theo lỗi bình phương trung bình tối thiểu MMSE (Minimum mean square error), thì hệ số kênh truyền được tính như công thức dưới đây.

$$\tilde{h}_u = g_u + \frac{\sqrt{p_{ub}} \beta_{u,nlos}^{rician}}{p_{ub} \tau_{pilot}^i + \delta^2} (y_{ub} - \sqrt{p_{ub}} \tau_{pilot}^i g_u) \quad (17)$$

Trong đó, g_u là vector hệ số kênh truyền ứng với tầm nhìn thẳng LoS; y_{ub} là mức thu tín hiệu tại eNodeB;

p_{ud} là công suất truyền từ thiết bị người dùng UE, khung thời gian huấn luyện kênh được tính từ công thức (15); hệ số suy hao được tính theo công thức (16).

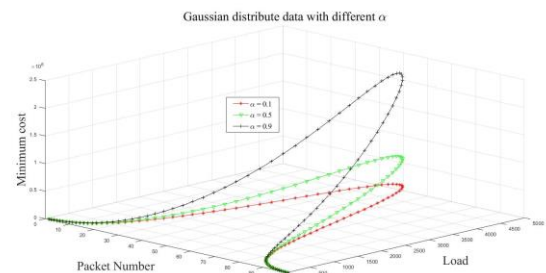
Các thành phần không trực giao được coi là nhiễu trên kênh truyền tác động tới tốc độ của user tới MEC. Từ đó, ta thu được thành phần gây nhiễu lên công thức (2), như dưới đây.

$$\eta_{u,b}^i = \frac{p_{ub} (\beta_{u,los}^{rician} + \beta_{u,nlos}^{rician})^2}{(p_{ub} \beta_{u,los}^{rician} + \sigma^2) (\beta_{u,los}^{rician} + \beta_{u,nlos}^{rician}) + p_{ub} \beta_{u,los}^{rician} \beta_{u,nlos}^{rician}} \quad (18)$$

Trong đó, σ^2 là phương sai của các chuỗi tạp âm trắng cộng và p_{ub} là công suất truyền liên quan tới năng lượng tại công thức (8). Các tham số sử dụng trong mô phỏng được trình bày tại bảng 2.

Trên cơ sở kịch bản và các tham số đầu vào được giả thiết, chúng tôi khảo sát hàm giá (cost) theo công thức (14) với các trọng số cân bằng thay đổi. Đồng thời, các mẫu nhiệm vụ đầu vào được thay đổi bởi các hàm phân bố khác nhau gồm (phân bố chuẩn, phân bố mũ).

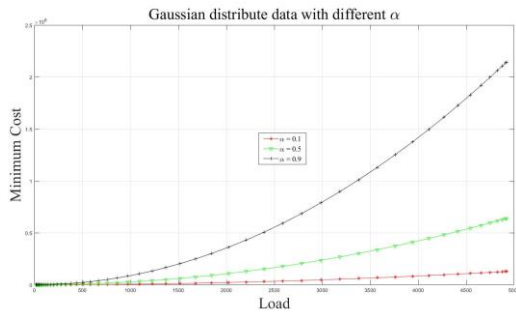
Kết quả mô phỏng trên hình 3 dưới đây chỉ ra sự biến động của hàm C_M với sự cân bằng của các ưu tiên về trễ và năng lượng cho thiết bị người dùng UE.



Hình 3. Sự biến thiên hàm Cost với phân bố gauss

Một trong các nhiệm vụ cơ bản của các bộ cảm biến là thu thập các thông tin định kỳ với nhiều mẫu lưu lượng khác nhau [21]. Các thông tin này tạo thành các tiến trình phân bố Gaussian, và tổng của các phân bố này tạo ra nhu

cầu đầu vào của nhiệm vụ như được mô phỏng trên hình 3.



Hình 4. Hàm Cost với trọng số ưu tiên và phân bố gauss

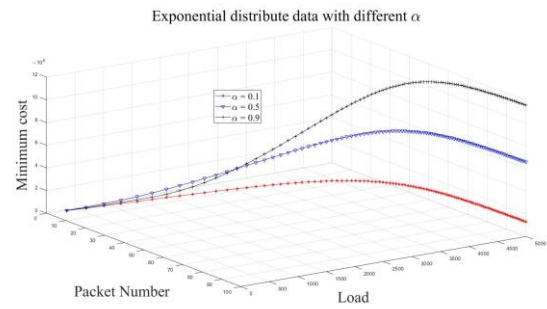
Với các mức ưu tiên khác nhau tại công thức (14), cơ chế giảm tải cho thấy sự thích ứng với các điều kiện đầu vào và chỉ ra được mức hiệu dụng tối đa của quá trình giảm tải.

Mặc dù hàm giá tăng theo mức tải của nhiệm vụ đầu vào như trên hình 4. Nhưng sự cân bằng trọng số giữa năng lượng và trễ tại các mức cân bằng cho giá trị hàm giá hướng về giá trị trung bình. Điều kiện ưu tiên về trễ yêu cầu kết thúc của nhiệm vụ có thể thay đổi và tác động đến độ dốc của các đường cong hàm giá.

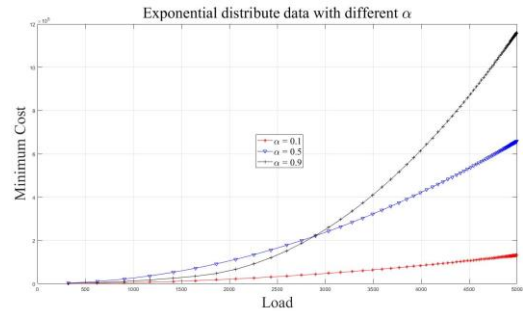
BẢNG 2. TRÌNH BÀY CÁC GIÁ TRỊ CỦA THAM SỐ MÔ PHỎNG SỐ.

Vùng phủ	300 m
Công suất phát p_{ub}	0.5 W [18]
Kích cỡ nhiệm vụ τ_m	[100- 1000] Kb [14]
Hàm phân bố task	Uniform, normal, exponent
Trọng số λ, β	0.5, 0.5
Hằng số điện dung	10^{-28}
Tốc độ truyền UE-UH	2Mbps
Tốc độ tính tại UE, UH	2.85×10^5 cycles/s
Hệ số tổn hao	β_u tính theo thực nghiệm [19]
Khoảng cách anten	$d=0.5$ m
Tần số sóng mang	3.5 GHz
Băng thông	10MHz-25MHz
Công suất phát p_{ud}	24dBm-46dBm [20]
Tạp âm N_0	-174dBm/Hz

Mẫu lưu lượng đầu vào của UE có thể biến động theo các dịch vụ ứng dụng và tuân thủ theo hàm mũ. Các điều kiện đầu vào điển hình đã được chứng minh trong [22]. Hình 5 và hình 6 dưới đây chỉ ra các khảo sát về biến thiên hàm giá với các nhiệm vụ đầu vào theo phân bố mũ. Với trọng số ưu tiên năng lượng, giá trị tối thiểu của hàm giá có độ dốc lớn theo độ dài nhiệm vụ. Vì vậy, cho dù mỗi quan hệ rõ giữa năng lượng và độ trễ không rõ ràng thì trọng số cân bằng đem lại mức thích ứng tốt hơn đối với máy chủ MEC.



Hình 5. Hàm Cost với đầu vào phân bố mũ



Hình 6. Hàm Cost thay đổi theo trọng số ưu tiên

V. KẾT LUẬN

Điện toán biên hứa hẹn đem lại rất nhiều ứng dụng hiệu quả trong tiếp cận IoT và 5G hiện nay. Tuy nhiên, việc tìm kiếm giải pháp giảm tải đa mục tiêu với các mô hình truyền thông thực tiễn gặp rất nhiều thách thức. Bài báo này đề xuất một mô hình giảm tải cân bằng năng lượng và trễ với sự hỗ trợ của các kết nối D2D tới thiết bị lân cận. Với các bộ tham số kênh truyền thực tế của kết nối 5G MIMO, cơ chế giảm tải từng phần đã được đề xuất cùng với các mô phỏng chứng minh về tính thích ứng với các phân bố đầu vào IoT điển hình. Xu hướng tương thích của hàm hiệu dụng có thể sử dụng để dự báo và hoạch định các mức suy hao năng lượng và thời gian sống của các thiết bị người dùng hiệu quả. Các quá trình phân nhóm người dùng và chiến lược lựa chọn máy chủ MEC bằng tiếp cận tính toán thông minh sẽ là nhiệm vụ tiếp theo của nghiên cứu.

LỜI CẢM ƠN

Tác giả đề tài cảm ơn Học viện Công nghệ Bưu Chính Viễn thông đã tài trợ cho nghiên cứu này theo đề tài mã số 06-HV-2021-VT1.

TÀI LIỆU THAM KHẢO

- [1] Kumar, Sachin, Prayag Tiwari, and Mikhail Zymbler. "Internet of Things is a revolutionary approach for future technology enhancement: a review." *Journal of Big data* 6.1 (2019): 1-21.
- [2] Khanna, Abhishek, and Sanmeet Kaur. "Internet of Things (IoT), applications and challenges: A comprehensive review." *Wireless Personal Communications* 114 (2020): 1687-1762.
- [3] Atieh, Ali T. "The Next Generation Cloud technologies: A Review On Distributed Cloud, Fog And Edge Computing and Their Opportunities and Challenges." *ResearchBerg Review of Science and Technology* 1, no. 1 (2021): 1-15.
- [4] De Donno, Michele, Koen Tange, and Nicola Dragoni. "Foundations and evolution of modern computing paradigms: Cloud, iot, edge, and fog." *Ieee Access* 7 (2019): 150936-150948.
- [5] Blanco, Bego, Jose Oscar Fajardo, Ioannis Giannoulakis, Emmanouil Kafetzakis, Shuping Peng, Jordi Pérez-

- Romero, Irena Trajkovska et al. "Technology pillars in the architecture of future 5G mobile networks: NFV, MEC and SDN." *Computer Standards & Interfaces* 54 (2017): 216-228.
- [6] Lin, Hai, Sherali Zeadally, Zhihong Chen, Houda Labiod, and Lusheng Wang. "A survey on computation offloading modeling for edge computing." *Journal of Network and Computer Applications* (2020): 102781.
- [7] Mach, Pavel, and Zdenek Becvar. "Mobile edge computing: A survey on architecture and computation offloading." *IEEE Communications Surveys & Tutorials* 19, no. 3 (2017): 1628-1656.
- [8] Zhou, Wen, Ling Xing, Junjuan Xia, Lisheng Fan, and Arumugam Nallanathan. "Dynamic Computation Offloading for MIMO Mobile Edge Computing Systems with Energy Harvesting." *IEEE Transactions on Vehicular Technology* 70, no. 5 (2021): 5172-5177.
- [9] Sufyan, Farhan, and Amit Banerjee. "Computation Offloading for Distributed Mobile Edge Computing Network: A Multiobjective Approach." *IEEE Access* 8 (2020): 149915-149930.
- [10] Zhou, Wen, Ling Xing, Junjuan Xia, Lisheng Fan, and Arumugam Nallanathan. "Dynamic Computation Offloading for MIMO Mobile Edge Computing Systems with Energy Harvesting." *IEEE Transactions on Vehicular Technology* 70, no. 5 (2021): 5172-5177.
- [11] Yu, Zhe, Yanmin Gong, Shimin Gong, and Yuanxiong Guo. "Joint task offloading and resource allocation in UAV-enabled mobile edge computing." *IEEE Internet of Things Journal* 7, no. 4 (2020): 3147-3159.
- [12] Liao, Zhuofan, Jingsheng Peng, Bing Xiong, and Jiawei Huang. "Adaptive offloading in mobile-edge computing for ultra-dense cellular networks based on genetic algorithm." *Journal of Cloud Computing* 10, no. 1 (2021): 1-16.
- [13] Saleem, Umber, Yu Liu, Sobia Jangsher, Xiaoming Tao, and Yong Li. "Latency minimization for D2D-enabled partial computation offloading in mobile edge computing." *IEEE Transactions on Vehicular Technology* 69, no. 4 (2020): 4472-4486.
- [14] Tran, Tuyen X., and Dario Pompili. "Joint task offloading and resource allocation for multi-server mobile-edge computing networks." *IEEE Transactions on Vehicular Technology* 68, no. 1 (2018): 856-868.
- [15] Zhang, Guanglin, Wenqian Zhang, Yu Cao, Demin Li, and Lin Wang. "Energy-delay tradeoff for dynamic offloading in mobile-edge computing system with energy harvesting devices." *IEEE Transactions on Industrial Informatics* 14, no. 10 (2018): 4642-4655.
- [16] Yuchong, Luo, Wu Jigang, Wu Yalan, and Chen Long. "Task scheduling in mobile edge computing with stochastic requests and m/m/1 servers." In *2019 IEEE 21st International Conference on High Performance Computing and Communications (HPCC/SmartCity/DSS)*, pp. 2379-2382. IEEE, 2019.
- [17] Minh HT, Tra NT and Thu HT. "An offload scheme for energy optimization in mobile edge computing systems." *Journal of Science and Technology on Information and Communications* 1, no. 3 (2020): 40-43.
- [18] Yang, Yibo, Yongkui Ma, Wei Xiang, Xuemai Gu, and Honglin Zhao. "Joint optimization of energy consumption and packet scheduling for mobile edge computing in cyber-physical networks." *IEEE Access* 6 (2018): 15576-15586.
- [19] Chu, Xiaoli, David Lopez-Perez, Yang Yang, and Fredrik Gunnarsson, eds. *Heterogeneous cellular networks: theory, simulation and deployment*. Cambridge University Press, 2013.
- [20] Papadopoulos, Haralabos, Chenwei Wang, Ozgun Bursalioglu, Xiaolin Hou, and Yoshihisa Kishiyama. "Massive MIMO technologies and challenges towards 5G." *IEICE Transactions on Communications* 99, no. 3 (2016): 602-621.
- [21] Gupta, Vinay, Sendil Kumar Devar, N. Hari Kumar, and Kala Praveen Bagadi. "Modelling of IoT traffic and its impact on LoRaWAN." In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pp. 1-6. IEEE, 2017.
- [22] Gharbieh, Mohammad, Hesham ElSawy, Hong-Chuan Yang, Ahmed Bader, and Mohamed-Slim Alouini. "Spatiotemporal model for uplink IoT traffic: Scheduling and random access paradox." *IEEE Transactions on Wireless Communications* 17, no. 12 (2018): 8357-8372.

THE MULTI-PURPOSES OFFLOADING SCHEME IN MOBILE EDGE COMPUTING

Abstract: Mobile edge computing is one of the key pillars of 5G network technologies to provide a distributed control approach compatible with emerging service needs. One of the outstanding advantages of edge computing comes from the streamlined process of coordinating the computing needs of user devices to local devices and edge servers. However, the biggest challenge of this approach is the problem of choosing an appropriate offloading mechanism due to the constraints imposed on multi-objective optimization problem is NP-Hard. This paper presents a offloading mechanism that balances the energy consumption and the delay requirements of the user device. The numerical simulation results with realistic radio channel assumptions including MIMO and 802.11p connections with different input traffic patterns show the impacts of the partial offload mechanism which was not mentioned by previous study.

Keywords: Mobile Edge Computing, offloading, multi-purposes optimization, MIMO.



Nguyen Thanh Tra received the Diploma of Engineer in Electronic Engineering and Telecommunication Technologies from Hanoi University of Science and Technology (HUST), the Master of Engineering in Electronics and Telecommunications

from Posts and Telecommunications Institute of Technology (PTIT) in 2000 and 2006, respectively. Currently, she is a main lecturer at the Telecommunication Faculty 1 of PTIT. She interested in optimization, security and network performance fields.