

# LƯU TRỮ DỮ LIỆU TRONG MẠNG HƯỚNG NỘI DUNG DỰA TRÊN DỮ LIỆU PHỔ BIẾN

Lê Phong Dữ<sup>1,2</sup>, Lê Tuấn Anh<sup>3</sup>, Nguyễn Đức Thái<sup>4</sup>

<sup>1</sup> Trường Đại học Lạc Hồng, <sup>2</sup> Trường Thực hành Sư phạm, Trường Đại học Trà Vinh,  
<sup>3</sup> Trường Đại học Thủ Dầu Một, <sup>4</sup> Trường Đại học Bách khoa, Đại học Quốc gia TP.HCM

**TÓM TẮT:** CCN (Content Centric Networking) được xem như là cấu trúc mạng Internet trong tương lai, trong đó dữ liệu được chuyển từ mô hình host - to - host sang mô hình truyền dữ liệu dựa trên tên nội dung. Lưu trữ dữ liệu tại các node là đặc điểm của CCN và hiệu quả của kiến trúc này phụ thuộc vào chính sách lưu trữ dữ liệu tại các node. Đã có nhiều giải pháp được đề xuất nhằm nâng cao hiệu suất của mạng CCN. Trong bài báo này, chúng tôi đề xuất một chính sách lưu trữ dữ liệu tại các node trong mạng CCN dựa vào dữ liệu phổ biến. Thông qua mô phỏng, chúng tôi đánh giá các tiêu chí về tỉ lệ tìm thấy dữ liệu và khoảng cách từ người dùng đến node tìm thấy dữ liệu là đạt kết quả tốt hơn chính sách lưu trữ mặc định của CCN, đồng thời đạt hiệu quả về tiết kiệm sử dụng tài nguyên trong mạng.

**Từ khóa:** CCN, Caching.

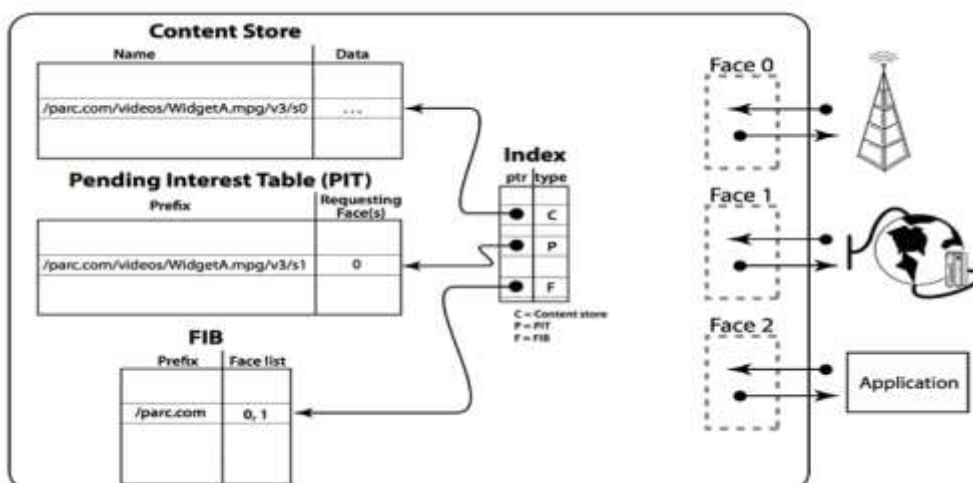
## I. GIỚI THIỆU

Internet đã phát triển nhanh các dịch vụ, dữ liệu trao đổi ngày càng lớn, xu hướng này tiếp tục tăng theo thời gian. Theo thống kê của Cisco VNI [1] số lượng IP toàn cầu tăng gấp tám lần trong năm năm qua, tốc độ tăng trung bình là 29% từ năm 2011-2016. Nhu cầu người dùng sử dụng ngày càng nhiều dịch vụ và truy cập lượng dữ liệu ngày càng lớn, đặt ra nhiều thách thức cho việc đáp ứng yêu cầu truyền dữ liệu trên mạng như: băng thông lớn, tốc độ bit cao, độ trễ thấp, bảo mật cao. Mô hình mạng truyền thống ngày càng quá tải so với những yêu cầu đặt ra. Việc cải tiến TCP/IP là một yêu cầu cấp thiết, cần đề xuất các giải pháp mới nhằm nâng cao hiệu quả sử dụng và hiệu suất mạng.

Xu hướng kết nối hiện nay hướng đến tiếp nhận và phổ biến dữ liệu ở nhiều nơi, nghĩa là mô hình kết nối Internet tập trung vào dữ liệu, không quan tâm nhiều đến vị trí vật lý mà dữ liệu được lưu trữ. Có nhiều nghiên cứu về mô hình kết nối dựa vào tên dữ liệu đã được đề xuất, trong đó Content Centric Networking hay mạng CCN được đề xuất bởi PARC (Palo Alto Research Center) được xem như là mạng Internet của tương lai.

CCN là một kiến trúc mạng mới với mục tiêu giải quyết những hạn chế của hệ thống mạng hiện tại. Lưu trữ dữ liệu trong mạng là đặc điểm của CCN, nội dung dữ liệu được phân thành các chunk [2], mỗi chunk được xác định bởi tên duy nhất để nhận và chuyển tiếp gói tin được sử dụng thay cho địa chỉ IP.

Mạng CCN tiết kiệm tài nguyên trong mạng được thực hiện bởi chính sách lưu trữ tại các node, mặc định là tất cả dữ liệu được lưu trữ tại các node mà dữ liệu đi qua. Mỗi node trong mạng CCN gồm ba cấu trúc dữ liệu chính [2]: Content Store (CS) là bộ nhớ lưu trữ dữ liệu, Forwarding Information Base (FIB) chứa một danh sách định tuyến dữ liệu và Pending Interest Table (PIT) lưu vết các gói tin đang chờ được phục vụ, nghĩa là đã được gửi yêu cầu qua mạng và chờ phản hồi từ mạng.



Hình 1. Cơ chế định tuyến-chuyển tiếp gói tin tại CCN node [2]

Có hai gói tin trong CCN, Interest là gói tin yêu cầu dữ liệu từ phía người dùng và Data mang dữ liệu được yêu cầu. Nguồn giữ Data mỗi khi nhận được các Interest có cùng tên dữ liệu yêu cầu thì tự động gửi Data cho người dùng theo cơ chế lưu vết đường đi của Interest trong CCN.

Khi một gói tin Interest đến node, CCN node sẽ tìm kiếm trong CS, nếu có tồn tại gói tin trong bộ nhớ sẽ gửi Data về cho người dùng đồng thời hủy bỏ gói Interest này. Nếu không tìm thấy dữ liệu có trong CS, node sẽ tìm trong danh sách các gói Interest đang chờ được phục vụ trong PIT. Nếu đã có trong danh sách đang chờ được phục vụ thì Interest sẽ bị hủy không được gửi đến node kế tiếp, ghi nhớ Interface gửi yêu cầu vào PIT. Nếu Interest không có trong danh sách đang chờ được phục vụ, node sẽ ghi lại thông tin vào bảng cần được phục vụ PIT, dò tìm các Interface trong bảng FIB để chuyển tiếp gói Interest.

Lưu trữ dữ liệu trong mạng là đặc điểm của CCN, vì thế có nhiều bản sao dữ liệu được lưu trữ trung gian tại các node trong mạng. Bất kỳ node nào cũng có thể trở thành server trả lời yêu cầu của người dùng, điều này giảm khả năng tắc nghẽn mạng, hạn chế truy cập dữ liệu tại server chính, thời gian phục vụ yêu cầu người dùng được cải thiện đáng kể.

Trong bài báo này, chúng tôi đề xuất chính sách lưu trữ dữ liệu tại các node trong mạng CCN dựa vào dữ liệu phổ biến. Cấu trúc bài báo gồm phần II trình bày các nghiên cứu liên quan; chính sách lưu trữ được đề xuất trình bày trong phần III; phần IV nội dung và tham số mô phỏng; đánh giá kết quả mô phỏng được trình bày ở phần V; phần kết luận trình bày trong phần VI.

## II. CÁC NGHIÊN CỨU LIÊN QUAN

Lưu trữ trong mạng không phải là chủ đề mới, đã có nhiều nghiên cứu liên quan đến chính sách lưu trữ dữ liệu và chính sách thay thế dữ liệu trên web caching. Đối với CCN, mỗi node đều có không gian bộ nhớ khác nhau nên việc thiết kế thuật toán tốt sẽ tiết kiệm và sử dụng hiệu quả tài nguyên mạng. Có nhiều chính sách lưu trữ được đề xuất để lưu trữ dữ liệu tại các node. Lưu trữ dữ liệu tại các node giúp giảm sử dụng băng thông, hạn chế truy cập dữ liệu tại server gốc, hạn chế tắc nghẽn mạng.

Thuật toán LCE [3][9] là thuật toán được định nghĩa mặc định cho mạng CCN. Dữ liệu sẽ được lưu trữ tại tất cả các node khi nó đi qua. Kết quả là nội dung dữ liệu được lưu trữ phổ biến trong mạng, gần với yêu cầu người dùng. Tuy nhiên, nó tốn không gian lưu trữ và dữ liệu bị trùng lặp, các node sẽ hết không gian lưu trữ trong thời gian ngắn và các chính sách thay thế dữ liệu sẽ phải thực hiện liên tục tại các node.

Trong khi thuật toán LCD [4][9] lưu trữ bản sao nội dung dữ liệu tại node mức  $i-1$ , khi tìm thấy dữ liệu tại một node thứ  $i$ . LCD đã giải quyết được trùng lặp dữ liệu cũng như tiết kiệm không gian bộ nhớ nhưng dữ liệu tại các node suy giảm. Một dạng khác của LCD là thuật toán MCD [4][9] cũng được đề xuất, trong thuật toán này, khi lưu bản sao dữ liệu tại node thứ  $i-1$  thì dữ liệu tại node thứ  $i$ , nơi tìm thấy dữ liệu sẽ bị xóa.

Một đề xuất khác là thuật toán dựa vào xác suất Prob cache [6][9], mỗi node sẽ lưu trữ một bản sao dữ liệu với xác suất  $p$  và không lưu với xác suất  $1-p$ . Khi xác suất là 1, thực hiện thuật toán LCE. Thuật toán này đã giải quyết được các vấn đề khó khăn của LCE và LCD gặp phải.

Bên cạnh đó, bộ nhớ lưu trữ tại các node có không gian bộ nhớ khác nhau, giới hạn về dung lượng, cần có các chính sách loại bỏ các dữ liệu không còn phù hợp, không còn được sử dụng để được thay thế bởi dữ liệu mới. Một số chính sách được sử dụng để thay thế dữ liệu như Least Recently Used (LRU) [5], Least Frequently Used (LFU) [7], Random (Rand) [8] hay First In First Out (FIFO).

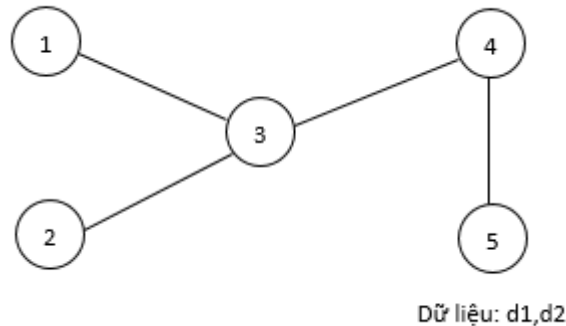
Lưu trữ trong mạng tồn tại hai dạng là on-path và off-path, các chính sách lưu trữ như LCE, LCD, MCD là lưu trữ dạng on-path, nghĩa là dữ liệu được lưu trữ trên đường chuyển về phía người dùng; trong khi off-path phải tính toán tối ưu cho việc thay thế dữ liệu tại các node. Chính sách lưu trữ dữ liệu off-path đạt hiệu năng mạng tốt hơn on-path, tuy nhiên nó tốn thời gian và phức tạp trong tính toán.

## III. LƯU TRỮ DỮ LIỆU DỰA VÀO NỘI DUNG PHỔ BIẾN

Để giải quyết những hạn chế của các chính sách lưu trữ dữ liệu mặc định của CCN và hạn chế việc thay thế dữ liệu liên tục tại các node, chúng tôi đề xuất một chính sách lưu trữ dữ liệu tại node dựa vào dữ liệu phổ biến, chúng tôi đặt tên là DP-CCN.

Dữ liệu phổ biến là dữ liệu được yêu cầu thường xuyên trong mạng. Mỗi node trong mạng CCN sẽ đếm dữ liệu được yêu cầu đi qua node. Mỗi dữ liệu yêu cầu đi qua node thì một danh sách chứa tên dữ liệu được yêu cầu và một biến chứa giá trị tương ứng của tên dữ liệu sẽ được tạo ra. Khi không tìm thấy dữ liệu trong CS thì giá trị của tên dữ liệu được yêu cầu sẽ được tăng lên 1, dữ liệu tiếp tục được chuyển đến node kế tiếp. Khi tìm thấy dữ liệu, dữ liệu trả về sẽ được lưu trữ tại node có giá trị bằng giá trị ngưỡng  $k$ . Biến chứa giá trị của dữ liệu sẽ được trả về giá trị 0.

Để mô tả thuật toán được rõ ràng, chúng tôi mô tả thông qua hình 2: Khi node 1 gửi yêu cầu dữ liệu d1, biến giá trị của dữ liệu d1 sẽ được tạo ra và tăng lên 1 ở các node 1, 3, 4 do không tìm thấy dữ liệu d1 trong CS được yêu cầu, tương tự node 2 cũng yêu cầu dữ liệu d1 thì biến giá trị của dữ liệu d1 sẽ tăng lên 1 ở các node 2, 3, 4. Như vậy giá trị của d1 ở node 3 và node 4 là 2. Dữ liệu d1 cũng được yêu cầu bởi node 3, giá trị của d1 tại node 4 là 3. Giả sử rằng ngưỡng để lưu trữ dữ liệu k=3 thì dữ liệu d1 sẽ được lưu tại node 4, đồng thời giá trị của d1 tại node 4 được thiết lập lại là 0.



Hình 2. Ví dụ mô hình DP-CCN

#### IV. NỘI DUNG VÀ THAM SỐ MÔ PHÒNG

Để đánh giá thuật toán đề xuất, chúng tôi sẽ tiến hành mô phỏng thuật toán trên môi trường mô phỏng ccnSim[11]. Mạng CCN được đề xuất phù hợp với mô hình mạng tự do hơn là mô hình mạng phân cấp như mô hình mạng hiện tại, vì thế chúng tôi sử dụng mô hình mạng tự do Abilene [11] có 11 node. Phân phối dữ liệu được sử dụng là phân phối Zipf với giá trị  $\alpha$  là 0.9.

Bên cạnh đó, CCN sử dụng các chính sách thay thế nội dung tại các node khi không gian lưu trữ tại các node không đủ để lưu dữ liệu mới. CCN đề xuất nhiều chính sách thay thế nội dung như Least Recently Used (LRU) [5], Least Frequently Used (LFU) [7], Random (Rand) [8] hay First In First Out (FIFO). Trong bài báo này, chúng tôi sử dụng chính sách thay thế dữ liệu là LRU, đây là chính sách thay thế dữ liệu lưu trữ tại node được sử dụng nhiều, nó sử dụng hàng đợi sắp xếp dữ liệu theo trình tự thời gian của lần truy cập trước đó. Khi không gian bộ nhớ node đầy, thuật toán xóa dữ liệu sau cùng và thêm dữ liệu mới vào đầu hàng đợi.

Bảng 1. Tham số mô phỏng

Tham số	Giá trị
Mô hình mạng	Abilene
Phân phối nội dung bằng Zipf	0,9
Thuật toán lưu trữ nội dung dữ liệu tại các node	LCE, DP-CCN
Thuật toán thay thế nội dung	LRU
Số chunk	$10^3$
Ngưỡng k	5

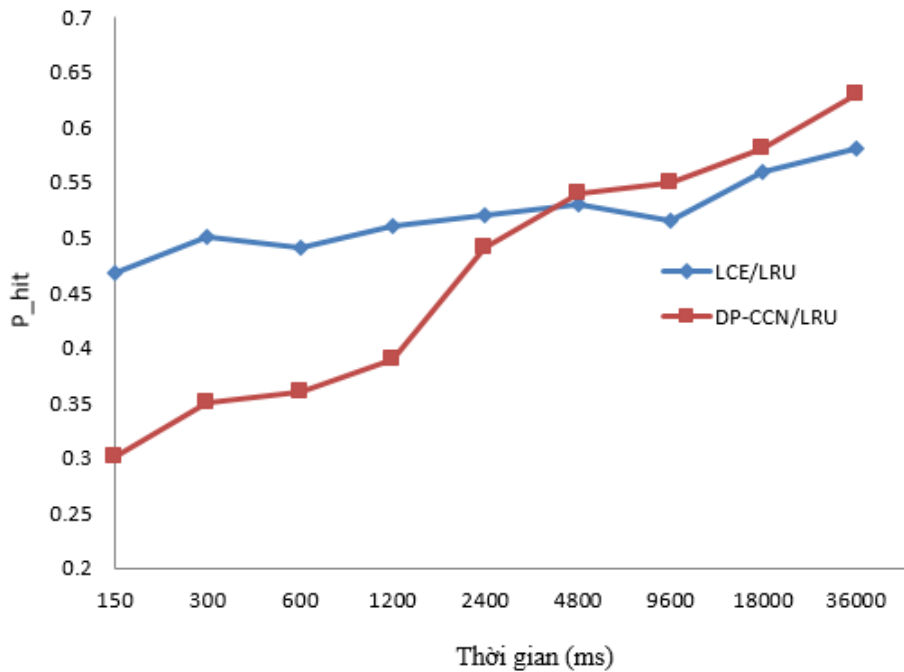
#### V. ĐÁNH GIÁ KẾT QUẢ CỦA THUẬT TOÁN DP-CCN

Trong phần này, sẽ trình bày kết quả mô phỏng dựa trên các tham số ở bảng 1. Chúng tôi đánh giá thuật toán DP-CCN thông qua các tiêu chí về xác suất tìm thấy dữ liệu và khoảng cách trung bình từ người dùng yêu cầu đến node tìm thấy dữ liệu.

Giá trị xác suất tìm thấy dữ liệu ( $P_{hit}$ ) để làm tham số so sánh. Xác suất tìm thấy dữ liệu được tính bởi công thức,

$$P_{hit} = \frac{\sum_{i=1}^n hit_i}{\sum_{i=1}^n miss_i + \sum_{i=1}^n hit_i},$$

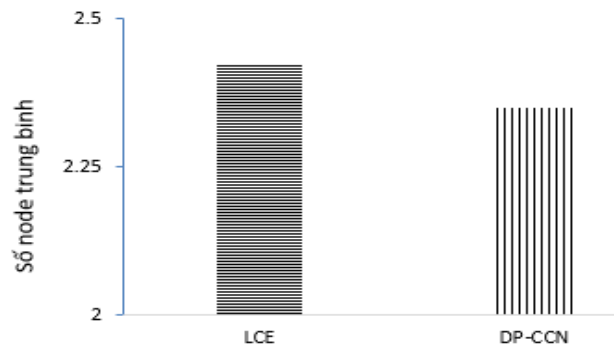
trong đó,  $hit_i$  là số Interest được gửi trong mạng CCN được tìm thấy dữ liệu tại node thứ  $i$ ,  $miss_i$  là số Interest của người dùng không tìm thấy nội dung dữ liệu tại node thứ  $i$ ,  $n$  là tổng số node trong mạng CCN.



**Hình 3.** Tỷ lệ tìm thấy dữ liệu giữa LCE và DP-CCN

Thuật toán DP-CCN cần thời gian để lưu trữ dữ liệu tại các node sau một khoản thời gian nhận yêu cầu từ người dùng. Kết quả mô phỏng như hình 3 cho thấy xác suất tìm thấy dữ liệu trong khoảng thời gian đầu thì thuật toán LCE đạt tỷ lệ cao hơn so với thuật toán DP-CCN. Tuy nhiên sau khoản thời gian 2400ms thì thuật toán DP-CCN đạt kết quả tốt hơn so với LCE. Thuật toán DP-CCN tính số lần được yêu cầu đi qua node, khi số lần được yêu cầu đạt ngưỡng là 5 thì dữ liệu sẽ lưu tại node. Thuật toán DP-CCN đã giải quyết được trùng lặp dữ liệu, tiết kiệm không gian bộ nhớ và hạn chế việc thay thế dữ liệu thường xuyên khi dữ liệu tại các node đầy.

Trong trường hợp xác định khoảng cách trung bình từ người dùng đến node đầu tiên tìm thấy dữ liệu được mô phỏng ở hình 4 cho thấy khoảng cách trung bình số node tìm thấy dữ liệu của DP-CCN ngắn hơn LCE khoảng 7.3%.



**Hình 4.** Số node trung bình tìm thấy dữ liệu

## VI. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất thuật toán DP-CCN, một chính sách lưu trữ dữ liệu tại các node trong mạng CCN. Thuật toán chúng tôi đề xuất đã khắc phục được những hạn chế của chính sách LCE trong mạng CCN, nó đã giải quyết được trùng lặp dữ liệu cũng như tiết kiệm không gian bộ nhớ thông qua kết quả mô phỏng. Tuy nhiên việc xác định giá trị ngưỡng k có ảnh hưởng đến kiến trúc mạng, do đó cần được nghiên cứu để tìm giá trị k cho phù hợp, đạt được hiệu suất tốt nhất trong nghiên cứu tiếp theo của chúng tôi.

## TÀI LIỆU THAM KHẢO

- [1] Cisco visual networking index: forecast and methodology: 2011-2015, 5/2012.
- [2] V. Jacobson, D. K. Smelters, I. D. Thornton, M. F. Plass, N. H. Briggs, R. L. Braynard, "Networking named content", ACM CoNEXT, Rome, Italy, December 2009.

- [3] G. Carofiglio, V. Gehlen, and D. Perino, "Experimental evaluation of memory management in content-centric networking", IEEE ICC, 2011, pp. 1-6.
- [4] K. Cho, M. Lee, K. Park, T. T. Kwon, Y. Choi, and S. Pack, "Wave: Popularity-based and collaborative in-network caching for content-oriented networks", IEEE INFOCOM WKSHPS, 2012, pp. 316-321.
- [5] D. He, W. K. Chai, and G. Pavlou. "Leveraging in-network caching for efficient content delivery in content-centric network", Proc. of London Communication Symposium 2011.
- [6] Psaras, Ioannis, Wei Koong Chai, and George Pavlou. "Probabilistic in-network caching for information-centric networks", Proceedings of the second edition of the ICN workshop on Information-centric networking. ACM, 2012.
- [7] S. Podlipnig and L. B" osz" ormenyi, "A survey of web cache replacement strategies," ACM Comput, vol. 35, no. 4, pp. 374-398, 2003.
- [8] N. Laoutaris, S. Syntila, and I. Stavrakakis, "Meta Algorithms for Hierarchical Web Caches", IEEE ICPC, 2004.
- [9] N Laoutaris, H Che, and I Stavrakakis, "The LCD interconnection of LRU caches and its analysis", Performance Evaluation, 63(7), 2006.
- [10] L. Breslau and et al., "Web caching and zipf like distributions: Evidence and implications," in In INFOCOM, 1999, pp. 126-134.
- [11] ccnSim homepage <http://www.infres.enst.fr/~drossi/ccnSim>, 04/2017.

## **CACHING POLICY FOR CONTENT CENTRIC NETWORKING BY POPULAR CONTENTS**

**Le Phong Du, Le Tuan Anh, Nguyen Duc Thai**

**ABSTRACTS:** *Content-centric networking (CCN) is considered to be the future Internet which moves from host-to-host model to transfer data model based on content name. Network data caching is a characteristic of CCN whose effectiveness depends on the content caching policies in the nodes. Several policies have been proposed to improve the caching performance in CCN. In this paper, we propose a policy for caching contents at nodes in the CCN based on popular contents. Our simulation evaluates cache hit ratio, and the distance from consumers to caching hit content in the first node to achieve better results than defaulting caching policy in CCN and saving resource in networking.*