Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 09-10/8/2018
DOI: 10.15625/vap.2018.00036

MỘT CHÍNH SÁCH LƯU TRỮ DỮ LIỆU TRONG MẠNG HƯỚNG NỘI DUNG

Lê Phong Dũ[1,2], Lê Tuấn Anh[3], Nguyễn Đức Thái[4]
[1] Trường Đại học Lạc Hồng
[2] Trường Thực hành Sư Phạm, Trường Đại học Trà Vinh
[3] Trường Đại học Thủ Dầu Một
[4] 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. Mỗi node mạng CCN có một bộ nhớ để lưu trữ lại các dữ liệu trong mạng, các dữ liệu lưu trữ được quyết định bởi chính sách lưu trữ, dung lượng bộ nhớ của các node CCN là giới hạn, do đó các chính sách lưu trữ và chính sách thay thế dữ liệu trên mỗi node CCN cần có được nghiên cứu và cải tiến. 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 xác suất khả năng lưu trữ của tất cả các node và thứ tự của node. Thông qua mô phỏng, chúng tôi đánh giá các tiêu chí về tỉ lệ tìm thấy dữ liệu và khoảng cách trung bình từ người dùng đến node tìm thấy dữ liệu là đạt kết quả tốt hơn chính sách lưu trữ LCE và LCD.*

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 8 lần trong năm năm qua, tốc độ tăng trung bình là 21 % từ năm 2016-2021. 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 trên mạng như: băng thông yêu cầu ngày càng cao, thời gian truyền dữ liệu có độ trễ lớn và chiếm dụng đường truyền lâu khi người dùng truy cập đến một máy chủ ở cách rất xa vị trí của người dùng, về độ tin cậy, về khả năng mở rộng và tính bảo mật dữ liệu,.... 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. Để phù hợp với xu thế ngày nay và thay thế kiến trúc mạng TCP/IP, 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, rất thích hợp cho các dịch vụ cung cấp dữ liệu mà điển hình đáng kể là video sẽ tăng lên 81 % đến 2021 trên tổng số dữ liệu [1] được yêu cầu trên mạng.

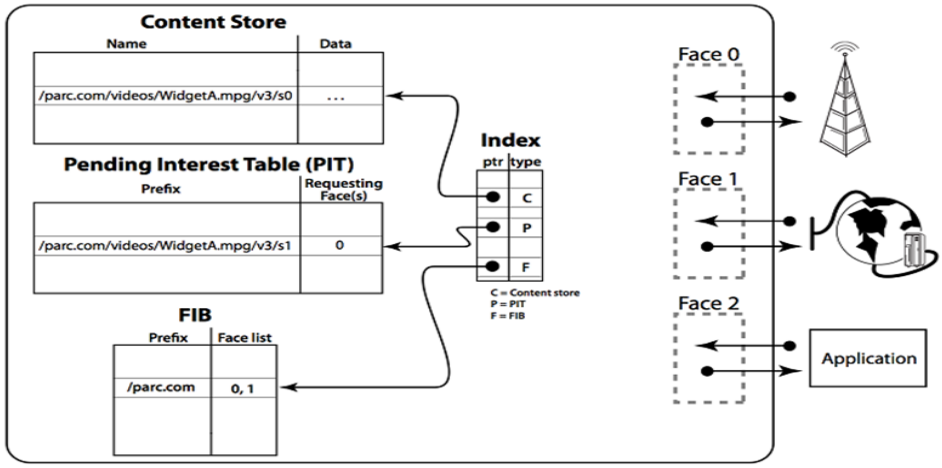
Mỗi node mạng CCN có một bộ nhớ dùng để lưu các dữ liệu được quyết định bởi *chính lưu trữ dữ liệu*, các dữ liệu sẽ được thay thế bởi các chính sách thay thế khi bộ nhớ đầy. Các gói tin dữ liệu sau khi đi qua mỗi node mạng thì nó sẽ được lưu lại trong các node nhằm tái sử dụng lại các gói tin dữ liệu để đáp ứng cho những người dùng khác có nhu cầu tương tự. Lưu trữ dữ liệu trong mạng là đặc điểm của CCN, 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.

Mỗi node trong mạng CCN gồm ba cấu trúc dữ liệu chính [2]: Content Store (CS) là bộ nhớ lưu trữ dữ liệu, tương tự như bộ nhớ đệm của một bộ định tuyến IP, là một trong những chức năng cốt lõi trong CCN; Forwarding Information Base (FIB) chứa một danh sách định tuyến dữ liệu, được dùng để chuyển tiếp các gói tin tin dữ liệu đến các nguồn dữ liệu để so khớp; 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.

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, 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 tắt nghẽn mạng, hạn chế truy cập dữ liệu tại server chính, thời gian phục vụ yêu cầu người dùng được cải thiện đáng kể.



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

Trong bài báo này, chúng tôi đề xuất chính sách lưu trữ dữ liệu tại các node trong mạng CCN dựa vào xác suất khả năng lưu trữ của tất cả các node và thứ tự của node. Cấu trúc bài báo gồm: Phần I giới thiệu tổng quan; 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. Thực tế, với mỗi khối dữ liệu trong mạng luôn có độ phổ biến nhất định (theo thời gian), các khối dữ liệu chỉ được sử dụng 1 lần (trong một khoản thời gian đủ dài nhất định) thường chiếm đến 45 % của các yêu cầu và chiếm đến 75 % tổng số các khối dữ liệu mạng [7], do đó các khối dữ liệu này có độ phổ biến thấp nhất. Phần 25 % còn lại là các khối dữ liệu có thể được yêu cầu để sử dụng nhiều lần, tức là nó có độ phổ biến cao hơn

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], [8] 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. Vấn đề của LCE chính là quá thường xuyên thay thế dữ liệu, do chính sách lưu trữ LCE không xét đến độ phổ biến của dữ liệu được lưu trữ. Một hạn chế khác của LCE là các dữ liệu trong mạng trùng lặp quá nhiều ở các node mạng lân cận.

Trong khi thuật toán LCD [4], [8] 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, dữ liệu phổ biến sẽ dần chuyển ra các node biên mạng gần với người dùng, trong khi các dữ liệu ít phổ biến thì ngược lại.

Một dạng khác của LCD là thuật toán MCD [4], [8] 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], [8], 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.

Một đề xuất khác là lưu trữ dữ liệu dựa vào dữ liệu phổ biến trong mạng [10]. 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. Tuy nhiên dữ liệu phổ biến sẽ giảm sau mỗi lần khởi tạo lại giá trị.

Lưu trữ trong mạng CCN 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 khí 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.

Để lưu trữ dữ liệu tại các node hiệu quả, chính sách thay thế nội dung dữ liệu đóng vai trò rất quan trọng vì bộ nhớ lưu trữ tại các node có không gian bộ nhớ khác nhau, giới hạn về dung lượng, không thể lưu trữ tất cả thông tin,

cần có các chính sách loại bỏ các dữ liệu không còn phù hợp để lưu thêm 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) [7] hay First In First Out (FIFO).

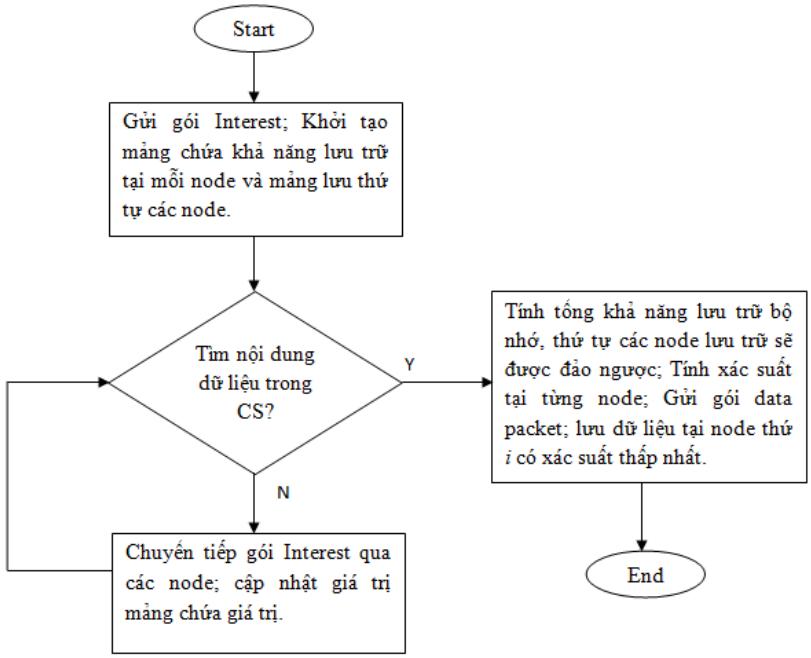
III. CHÍNH SÁCH LƯU TRỮ DỮ LIỆU DỰA TRÊN XÁC SUẤT

Để giải quyết những hạn chế của các chính sách lưu trữ dữ liệu 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 xác suất khả năng lưu trữ của tất cả các node và thứ tự của node, chúng tôi đặt tên là PCCN.

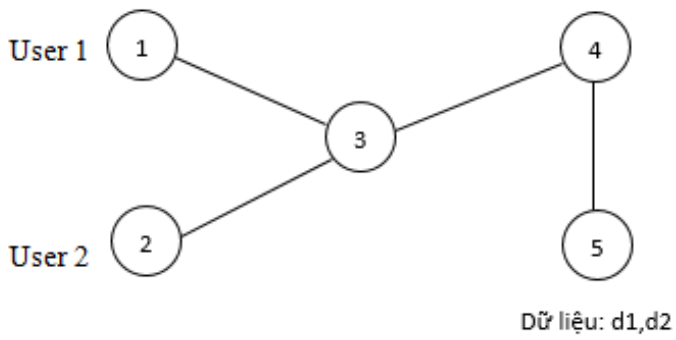
Mỗi node trong CCN có dung lượng bộ nhớ lưu trữ giới hạn nhất định. Khi yêu cầu nội dung dữ liệu, người dùng gửi gói Interest, nội dung dữ liệu phù hợp phản hồi yêu cầu bằng cách trả lời gói Data. Trong gói Interest chúng tôi sử dụng một mảng lưu trữ dung lượng bộ nhớ tại mỗi node thứ *i* đi qua và một mảng lưu thứ tự các node từ người dùng đến node tìm thấy nội dung. Khi gặp node chứa nội dung dữ liệu, tính tổng dung lượng bộ nhớ của tất cả các node trên đường mà gói tin Interest đã đi qua. Khi đó, thứ tự vị trí các node đi qua sẽ được đảo theo chiều ngược lại. Tại thời điểm gói Data gửi dữ liệu về phía người dùng, xác suất tại các node thứ *i* mà gói Interest đã đi qua theo công thức:

$$p(i) = \frac{c_i}{c_{total} * h}$$

trong đó, C_i là dung lượng bộ nhớ tại node thứ *i*, C_{total} là tổng dung lượng tất cả các node, *h* là giá trị vị trị tại node thứ *i* đã được đảo ngược. Dữ liệu được lưu tại node thứ *i* có xác suất thấp nhất.



Hình 2. Thuật toán PCCN



Hình 3. Ví dụ mô hình PCCN

Giả sử User 1 yêu cầu nội dung d1, gói Interest sẽ đi qua các node 1, 3, 4 và tìm thấy nội dung dữ liệu d1 theo yêu cầu tại node thứ 5, một mảng sẽ lưu dung lượng bộ nhớ của các node 1, 3, 4 và 1 mảng lưu thứ tự của các node trên đường đi node 1=1, node 3= 2, node 4 =3. Đến node thứ 5 tìm thấy nội dung dữ liệu d1, tổng dung lượng tất cả các

node 1, 3, 4 đồng thời đảo ngược mảng tứ tự, tức node 4=1, node 3=2, node 1=3. Xác suất tại các node lần lượt được tính như sau: $p(4) = \frac{C_4}{C_{total*1}}$, $p(3) = \frac{C_3}{C_{total*2}}$, $p(1) = \frac{C_1}{C_{total*3}}$. Dữ liệu sẽ được lưu tạo node có xác suất thấp nhất.

IV. THAM SỐ MÔ PHỎNG

Để đánh giá thuật toán đề xuất, chúng tôi sẽ tiến hành mô phỏng thuật toán đề xuất bằng ngôn ngữ lập trình C++. Mạng CCN được đề xuất phù hợp với mô hình mạng tự do hơn là mô hình mạng phân cấp, vì thế chúng tôi sử dụng mô hình mạng tự do Abilene [11] có 11 node, các bộ nhớ lưu trữ tại các node được cấp phát ngẫu nhiên trong mạng. Phân phối dữ liệu được sử dụng là phân phối Zipf [9] với giá trị α 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 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, LCD, PCCN
Thuật toán thay thế nội dung	LRU
Số chunk	10^3

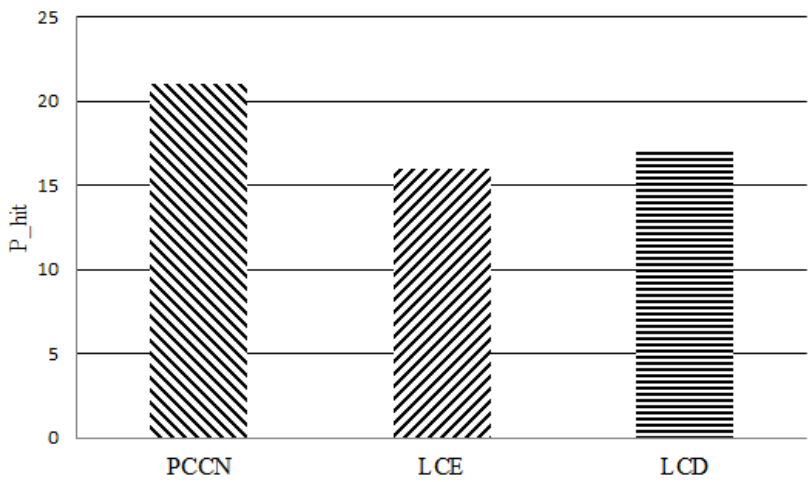
V. ĐÁNH GIÁ KẾT QUẢ CỦA THUẬT TOÁN PCCN

Trong phần này, sẽ trình bày kết quả mô phỏng dựa trên các tham số ở bảng 1. Chúng tôi đánh giá thuật toán PCCN 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:

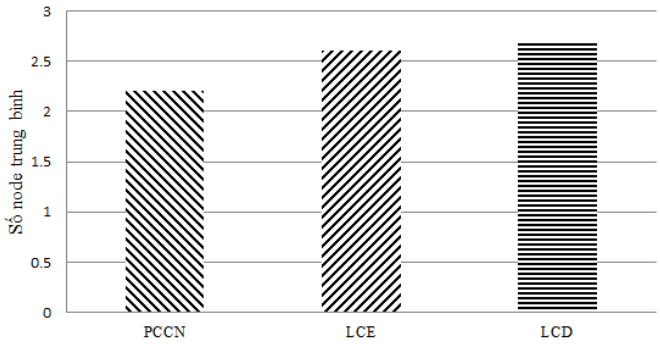
$$p\_hit = \frac{\sum_{i=1}^{n} hit_i}{\sum_{i=1}^{n} miss_i + \sum_{i=1}^{n} hit_i},$$

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. Kết quả mô phỏng hình 4 cho thấy thuật toán PCCN xác suất tìm thấy nội dung cao hơn chính sách lưu trữ LCE và LCD. PCCN đạt tỉ lệ cao hơn LCE là 4 % và 3 % so với LCD. Thuật toán PCCN có xu hướng lưu trữ dữ liệu tại các node dần ra ngoài các node ở biên nhiều hơn so với các node bên trong mạng, gần với người dùng.



Hình 4. Tỉ lệ tìm thấy dữ liệu giữa LCE, LCD và PCCN

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 5 cho thấy khoảng cách trung bình số node tìm thấy dữ liệu của PCCN ngắn hơn LCE và LCD, những nội dung phổ biến gần với người dùng hơn, những nội dung ít phổ biến sẽ được lưu trữ ở node bên trong mạng.



Hình 5. Số node trung bình tìm thấy dữ liệu

VI. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất thuật toán PCCN, một chính sách lưu trữ dữ liệu tại các node trong mạng CCN. Thuật toán chúng tôi cải tiến đạt kết quả tốt chính sách lưu trữ LCE và LCD trong CCN, nó đã giải quyết được trùng lặp dữ liệu cũng như tiết kiệm không gian bộ nhớ tại các node bằng cách tính xác suất và kết quả mô phỏng cho thấy đạt kết quả tốt hơn trong tìm thấy nội dung dữ liệu, các dữ liệu phổ biến có xu hướng lưu lưu trữ tại các node gần đến người dùng hơn.

VII. TÀI LIỆU THAM KHẢO

[1] Cisco visual networking index: forecast and methodology: 2016-2021, 9/2017.

[2] V. Jacobson, D. K. Smelters, 1. 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] N. Laoutaris, S. Syntila, and I. Stavrakakis. "Meta Algorithms for Hierarchical Web Caches". IEEE ICPCC, 2004.

[8] N Laoutaris, H Che, and I Stavrakakis. "The LCD interconnection of LRU caches and its analysis". Performance Evaluation, 63(7), 2006.

[9] L. Breslau and et al.. "Web caching and zipf like distributions: Evidence and implications" in In INFOCOM, 1999, pp. 126-134.

[10] Le Phong Du, Le Tuan Anh, Nguyen Duc Thai. "Lưu trữ dữ liệu trong mạng hướng nội dung dựa vào dữ liệu phổ biến". Kỷ yếu hội thảo Fair 2017.

[11] ccnSim homepage http://www.infres.enst.fr/~drossi/ccnSim, 04/2017.

A CACHING POLICY IN CONTENT CENTRIC NETWORKING

Le Phong Du, Le Tuan Anh, Nguyen Duc Thai

ABSTRACT: *Content Centric Networking (CCN) is considered as the future Internet structure which data is transferred from the host-to-host model to the content-based transmission model. Each node in CCN network has a memory to store the data in the network, The storage data is determined by the caching policies, The memory capacity of the CCN node is limited, so the caching policies and replacement policies on each node need to be researched and improved. There are many solutions proposed to improve the efficiency of CCN. In this paper, we propose a policy of caching data at nodes in an CCN based on the storage probabilities of all nodes and the order of nodes. Through simulation, we evaluated the cache hit ratio and average distance from users to node finding the data to achieve better results than LCE and LCD caching policies.*

Từ khóa: *CCN, Caching.*