The 2ACT Model-based Evaluation for In-network Caching Mechanism

Ke Xu*, Min Zhu*, Ning Wang†, Song Lin*, Haiyang Wang‡, Tong Li*
*Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, P.R.China
Email: {xuke,zhumin}@csnet1.cs.tsinghua.edu.cn, {linsong1984,litong662008}@gmail.com
†Centre for Communication Systems Research, University of Surrey, United Kingdom. Email: n.wang@surrey.ac.uk
†School of Computing Science, Simon Fraser University, Canada. Email: hwa17@cs.sfu.ca

***Abstract*—With the popularity of information and content items that can be cached within ISP networks, developing high-quality and efficient content distribution approaches has become an important task in future internet architecture design. As one of the main techniques of content distribution, in-network caching mechanism has attracted attention from both academia and industry. However, the general evaluation model of in-network caching is seldom discussed. The trade-off between economic cost and the deployment of in-network caching still remains largely unclear, especially for heterogeneous applications. We take a first yet important step towards the design of a better evaluation model based on the Application Adaptation CapaciTy (2ACT) of the architecture to quantify the trade-off in this paper. Based on our evaluation model, we further clarify the deployment requirements for the in-network caching mechanism. Based on our findings, ISPs and users can make their own choice according to their application scenarios.**

I. Introduction

With more than 40 years of development, the Internet has effectively become a distributed repository of massive data and digital media content. In the Visual Networking Index (VNI) Forecast (2011-2016) report[1] published in 2012, Cisco predicted that the sum of all kinds of video traffic would constitute 86% of global consumer Internet traffic by 2016. The increasing volume of application data and content that can be cached on the Internet like videos has triggered the reconsideration of the fundamental communication model of Internet infrastructure.

Up to now, numerous research papers [4] and influential approaches [1], [6] have been published, offering a wide variety of methods to cater for the explosion of application types such as video, file sharing, cloud computing, etc.. Many innovation projects also have been working on the design of content-oriented communication, such as Content-Centric Networking (CCN) [6], (Named Data Networking) NDN[2], etc.. How to distribute the heterogeneous data in a more efficient way is the main focus of these works. Since users and applications normally do not care about the location of the required data, they need not to obtain information from a specific endpoint attached to the Internet. Furthermore, the evolution of the hardware and the reduction at the cost of storage and CPU processing make in-network storage and data replication possible at large scale [2]. This provides the opportunity for in-network caching design, which allows the content to be cached anywhere on the delivery path.

Existing studies show that the in-network caching mechanism can provide better performance in data transfer with lower delays and higher accuracy [7]. But this does not mean it is superior in all cases. A recent study from [8] demonstrates that the performance gain in CCN has close relationship with the popularity of the content, the size of router-cache and the placement of the content. This result is also applicable to in-network caching mechanism. Furthermore, the new communication style of in-network caching mechanism, which is changed from the traditional type of "Where" to the new type of "What", also raises new challenges for supporting interactive applications and host-oriented services as voice and video conferencing [5].

It is easy to see that most of the protocols and mechanisms in the architecture can only serve some types of applications well; this is also true to in-network caching mechanism. Furthermore, simply satisfying the performance requirement of the application is not sufficient for the evolution of the mechanism and the architecture. The new mechanism and architecture should be both performance effective and cost-aware. It is obvious that the advancement of in-network caching mechanism under different scenarios requires further investigations. We begin our elaboration with the following discussions:

(1) How to evaluate the sustainable development capability and competitiveness of a certain kind of new technology and infrastructure like in-network caching mechanism?

(2) How does composition of different types of application data affect the deployment of the in-network caching mechanism in the architecture?

(3) How do economic factors affect the development of in-network caching mechanism in the Internet business market?

This research is supported by New generation broadband wireless mobile communication network of the National Science and Technology Major Projects (2012ZX03005001), NSFC Project (61170292), 973 Project of China (2009CB320501, 2012CB315803), 863 Project of China (2013AA013302), The National Science and Technology support Program (2011BAK08B05-02) and EU MARIE CURIE ACTIONS EVANS (PIRSES-GA-2010-269323).

[1]http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html

[2]http://www.named-data.net/.

The existing models that study the adoption of potential technologies are generally focusing on their utility [9], [10]. They construct the model according to the economic profit and the cost of the users and ISPs, which cannot reflect the variation tendency of the architectures with the change of the composition of heterogeneous application data types.

In this article, we attempt to explore the development trend of some kind of mechanism or architecture from the application perspective, in particular, the application data composition. We believe that the ability of a new mechanism or architecture to keep vitality and to maintain sustainable development depends on its Application Adaptation Capaciity (2ACT), i.e., whether it can provide better performance under different application scenarios and application data compositions, and whether it is less costly to support the application requirements. The former reflects the function and the performance of the architecture, which is called services adaptability of the architecture in this paper. The latter reflects the complexity of the architecture, and it is called economic adaptability of the architecture in this paper.

Based on the principle mentioned above, we first construct a 2ACT evaluation model, which can be used to measure the development trends of architectures under different application data compositions and economic cost scenarios. Then, taking in-network caching mechanism as an example, we use our proposed evaluation model to analyze its development trends. The results of our experiment show that the in-network caching mechanism does not perform efficiently in any case. In this context, ISP customers can make their own choice based on our distinct findings.

II. Methodology

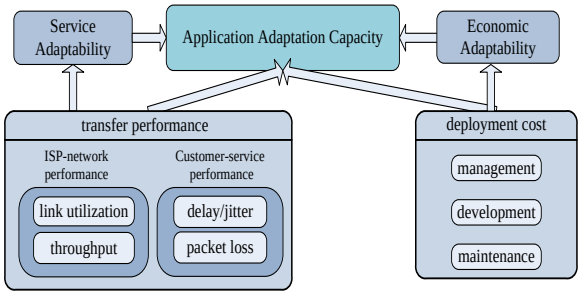

Fig. 1. The 2ACT Model framework.

The network applications deployed on the Internet, which provide great essential services for human beings, are one of the most important reasons of the popularity of the Internet. The motivation of the architecture innovation and development comes from the needs for new applications. The emergence of various applications such as Email, Web, and P2P continues to promote the popularity of the Internet to a new high. Therefore, it can be envisaged that the smooth development and long-term success of network architecture have close relationship with its ability to meet the application demands, which is called *Application Adaptation CapaciTy* (2ACT) in this article.

We believe that the evaluation for the development capacity of the architecture should start with two aspects as shown in Figure 1: Service Adaptability and Economic Adaptability.

(1) Service Adaptability

The most important function of the host-based Internet is to provide users with all kinds of services through different applications. Since the basic way of the network is to deliver the application data or packets to destination nodes in the network, we reckon that the data forwarding capacity of the architecture directly reflects its ability to provide Internet-based services.

It is easy to see that different data delivery techniques and the resulting routing protocols will incur different data transmission performance, and then provide different service experiences in the architecture. To ISP, it appears as network performance such as link utilization, throughput, etc.. To customers, it denotes service performance which mainly refers to the quality of service (QoS) assurance, with the key metrics of delay/jitter and packet loss. We believe that the data transmission performance, reflecting the ability of the architecture to satisfy heterogeneous service requirements, is the most important factor that influences the sustainable development of the architecture. We call it the *service adaptability* of the architecture and it is the first application adaptability element in 2ACT model (Figure 1).

The service adaptability represents the technicality of the architecture. Intuitively, the technology of the architecture is more advanced when it provides better services, meaning that it is more efficient in data transfer with shorter delays and less resource utilization.

(2) Economic Adaptability

The slow deployment of end-to-end QoS and IPv6 shows that the success of a new technology relies not only on the superiority of the technology itself, but also some other objective factors [10], such as the profit of the ISP, the cost of the development, the management and the maintenance. The development of Internet technologies has witnessed the failures of some proposed technologies caused by their costly implementation and deployment.

In fact, both the development and management cost and the architecture maintenance cost are the results of supportting the applications. The reason is that there are some high-level technical and economic questions that need to be addressed before and after the applications are deployed on the architecture. For example, what modifications the architecture needs to support all types of popular applications and how much does it cost for the modification?

We define the ability of introducing as little as possible overhead for application deployment, maintenance and management as the *economic adaptability* of the architecture in this article, which is also the second factor of 2ACT model that impacts the sustainable development of the architecture.

III. The Construction of 2ACT Model

As we mentioned in the last section, if the architecture can support current and potential future applications sufficiently

well, which means performing functionally without too high overhead costs, we believe that the architecture has sufficient potential to keep sustainability. Following this principle, we construct a 2ACT model in this section, which can be used to evaluate the architecture's sustainable development capability in the future. It can also be used to evaluate the adoption of other emerging networking paradigms since they all have the ability to support the applications on the Internet. This model also provides useful hints for users and ISPs to make a choice in accordance with their application scenarios.

A. Service Adaptability Model

As shown in Figure 2, there are numerous application-layer mechanisms, such as FTP, HTTP, P2P etc., over the Internet to provide all kinds of services such as file transfer, video/audio, email, etc.. All the data generated by these applications can be viewed as groups of packets delivered in the network through the transport protocols and routing protocols. This confirms our discussion in the last section, different data transmission performance represents different service adaptability.

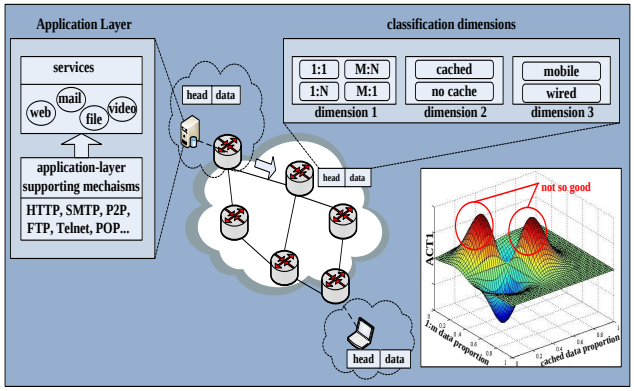

Fig. 2. The application data transfer model of the architecture.

According to different classification dimensions, the application data can be divided into different types. For example, there are three classification dimensions in Figure 2. Depending on the way the data is generated, the data can be categorized into two types: those can be cached (cacheable) and cannot be cached (non-cacheable). Data in VoIP telephony and some of Instant Message belong to non-cacheable data. Most of the VOD traffic and file data are cacheable for serving future incoming requests of the same content item. On the other hand, based on the number of senders and receivers involved in the communication, we can classify the application data into the following types: 1-to-1, 1-to-N, M-to-1 and M-to-N. Moreover, we can also classify the application data as data applied in the mobile network circumstance and that in the wired network circumstance according to its application environment. Users and ISPs can take new classification method according to their own demands of course.

As users may have different application demands under different circumstances, the proportion of application data in the network will also be different with the change in times or regions. Therefore, we can model the service adaptability of the architecture as:

$$U_s = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}(Perf)_{ij}$$
$$s.t. \sum_{j=1}^{n} p_{ij} = 1 \quad (1)$$

where m denotes the number of application classification dimensions while n denotes the number of application types in a particular classification dimension. The parameter p_{ij} represents the proportion of application data under classification i and application type j in the network.

The parameter $(Perf)_{ij}$ indicates the application data transmission performance of the architecture. It can be represented by some performance parameter such as latency, bandwidth, link utilization etc. Since the in-network caching mechanism can efficiently reduce the average router hop count in data delivery, we use the product of the average router hop count H_{ij} and unit data transmission performance in each hop $(PHop)_{ij}$ to indicate it in this article, which means that:

$$U_s = \frac{1}{m}\sum_{i=0}^{m}\sum_{j=0}^{n} p_{ij}(PHop)_{ij}H_{ij}$$
$$s.t. \sum_{j=0}^{n} p_{ij} = 1 \quad (2)$$

According to Formula (2), we can see that the smaller the average router hop count and the unit data transmission performance value are, the better service adaptability of the architecture will be. As the subfigure of Figure 2 shows, if U_s can not keep low with the change of data traffic proportion, we can not say that the service adaptability of the architecture is superior. But good service adaptability does not necessarily mean that the development tendency of the architecture is better because it also depends on its economic adaptability.

B. Economic Adaptability Model

Regardless of any distinct feature the architecture will have in the future, it should be built based on the current architecture for the sake of incremental migration. To deploy the corresponding mechanisms or protocols that support some kind of service in the architecture, it needs to pay for the development of a new network equipment or the upgrading of the existing network products. As in-network caching mechanism requires the architecture to have a certain level of ability to cache content, related costs will be incurred to upgrade or update current routers for the architecture. For users, costs that are too high are likely to increase the price to access the appropriate services in the Internet market. For network equipment providers and network operators, the increasing cost would definitely affect their profits, ultimately, as well as their decision-making on the choice of infrastructure. As a result, these types of deployment costs for improvement

would bring about important effect on the development of the architecture. Here we denote the cost as C_{dep}.

Besides the deployment costs mentioned above, the architecture incurs maintenance and management costs to support all kinds of applications. Here we denote these costs as C_{mt}.

Summing up the above two types of costs, we can construct the economic adaptability model as $U_e = C_{dep} + C_{mt}$.

Intuitively, a more complex technology implemented in the architecture requires higher costs for maintenance and management because it may consume more resources.

C. The 2ACT Model

Building on the service adaptability model and the economic adaptability model, we construct our 2ACT model as the following:

$$U = \frac{\alpha}{m}\sum_{i=0}^{m}\sum_{j=0}^{n} p_{ij}(PHop)_{ij}H_{ij} + \beta U_e$$
$$s.t. \sum_{j=0}^{n} p_{ij} = 1, \alpha + \beta = 1 \quad (3)$$

where the parameters α and β are used to weight the relative impact of service adaptability and economic adaptability on the development of the architecture. The economic adaptability function can be treated as the constraint of service adaptability. Since Lagrange function in optimization will get the same kind of equation as in Formula (3), to figure out the relationship between the performance and the cost of the architecture, we organize them in one equation and suppose they have linear relation.

In this formula, we can see, intuitively, that the 2ACT value U is lower with stronger service adaptability (smaller router hop count and unit data transmission performance) and superior economic adaptability (lower cost for deployment and management).

IV. The 2ACT Model-based Evaluation for In-network Caching Mechanism

How to deliver data to its interested recipients in an efficient way is one of the most fundamental issues in the content distribution and the design of the architecture. In-network caching mechanism is an efficient method in theory with copies of data replicated at multiple nodes in the network, and the increase of information and content items that can be cached in the network.

The average accessing time to data in the architecture is expected to be faster with the support of in-network caching mechanism which applies the principle of proximity. But in-network caching is only a performance enhancing mechanism with the cost of data duplicating and storage overhead, and it is only efficient in dealing with cacheable data. Although the amount of the cacheable data in the network such as video is increasing recently, it is hard to say that the in-network caching mechanism will be suitable without any limitation.

In this section, we perform some analysis to investigate the relationship between the composition of application data on the Internet, the economic cost, and the deployment of in-network caching mechanism in the architecture. To find the effectiveness of the in-network caching mechanism in the architecture, we take a current network, which has not deployed the in-network caching mechanism, to make a comparison. We call the architecture with data caching capability *cache network* and the network without memory on data passing by *regular network* in this article.

Considering the difference in various network scenarios, we select two types of Internet backbone, the experimental system PlanetLab[3] and the Internet2 Abilene network[4], to conduct our analysis.

A. The Average Router Hop Count for Non-cacheable Data Delivery

We have mentioned earlier that there are many types of application data carried over the Internet. For non-cacheable data, regardless the transport and network protocols, it needs to be transmitted from the source to the destination. That means the average router hop count for delivering non-cacheable data is identical in both the *cache network* and the *regular network* (we ignore the difference brought by other protocols such as routing protocol here to simplify the analysis). We use the average router hop count value getting from PlanetLab and Abilene network to represent this.

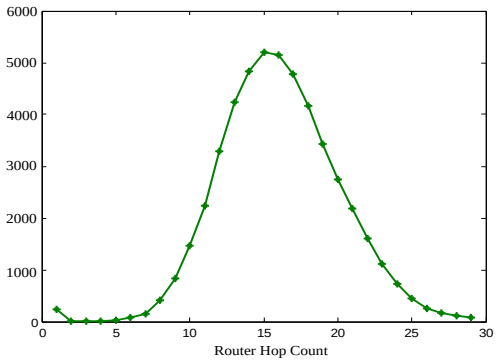

Fig. 3. Statistics in PlanetLab.

We randomly selected 130 source nodes and 580 destination nodes across the globe in the PlanetLab. This number of nodes can well capture the feature of real-world networks. Then we used traceroute to calculate the router hop count between any source node and destination node, whose distribution is reported in Figure 3. From the statistic results in Figure 3, we obtained the average router hop count in the PlanetLab, 16, which represents the actual end-to-end average router hop counts on the Internet. We also make similar calculation in the Abilene network, and get its average router hop count 5. It denotes the average router hop of a single backbone network.

[3]http://www.planet-lab.org/
[4]http://www.internet2.edu/network/

B. The Average Router Hop Count for Cacheable Data Delivery

In-network caching mechanisms mainly deal with the transfer of cacheable application data. Caches close to the interested recipients can reduce the transmission hop count of this kind of traffic in the *cache network*, while such traffic still needs to be delivered through the full path from the source to the destination in the *regular network*. Therefore, the average router hop count for the cacheable data delivery is expected to be much smaller in the *cache network*.

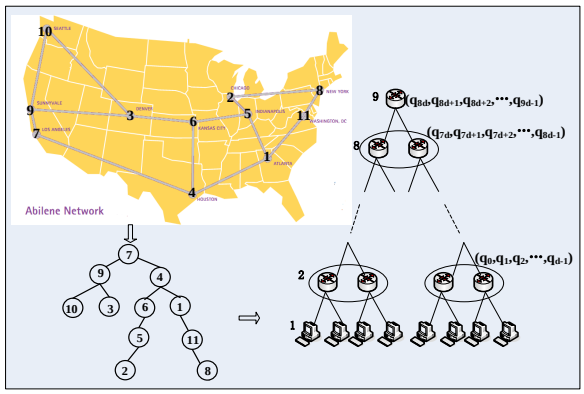

Fig. 4. The average router hop count for cacheable data delivery.

For quantitative comparison, we construct a tree topology as each receiver can make a spanning tree according to the hierarchy and the logicality relationship between the routers, which makes the router at its first hop the root node. For example, suppose that a user who connects the router #7 in Figure 4 needs some content, a spanning tree rooted at that node can be constructed, as shown in Figure 4.

As it is very difficult to compute the router hop value according to all kinds of random spanning trees, we just simplify it to a full K-ary tree because we only want to know the relative relationship between the average router hop count for cacheable data delivery in the cache network and in the regular network. In addition, we also make the following assumptions:

(1) The height of the tree is n. The leaf nodes of the tree represent "communication endpoints" such as the end users and servers, and the intermediate nodes represent network elements. Therefore, the number of leaf nodes is K^n, and the average hop count between leaf nodes is $\sum_{i=1}^{n} 2i(K-1)K^{i-1}/(K^n-1)$. After calculating the average hop between the leaf nodes under different K and n, we assume a full binary tree as shown in Figure 4, whose height is 9, to be used to compute the average router hop for cacheable data delivery, as its average end-to-end hop count is 16. It is right the same as that in PlanetLab.

(2) There are m different cacheable items, which are equally partitioned into k classes of popularity, and cached in the router of the binary tree according to Zipf popularity distribution [3]. I.e., $q_k = C/k^{\alpha}, \sum q_k = 1$, where q_k represents the popularity of the items in class k. The more popular the content, the closer it is to the leaf nodes [8]. So, we get $C = (\sum_{k=1}^{k'} 1/k^{\alpha})^{-1}$.

(3) The average size of each class is P, and each router in the network has a cache with storage size S, then each router in the network can store d classes of items and d is equal to S/P.

(4) The probability for the end users to get the content items from the routers at the same level in the tree is equal, then, we get the popularity distribution of data cached in router at level j as $A_j = \sum_{s=d(j-1)}^{d*j-1} q_s$ as shown in Figure 4.

According to the binary tree, we get the average router hop count for cacheable data delivery in in-network caching mechanism based network under Zipf distribution:

$$H' = \frac{k-1}{k^n-1} \sum_{i=1}^{n} K^{i-1} (\sum_{j=0}^{i} jA_j + 2i(1 - \sum_{j=1}^{i} A_j)).$$

Suppose that α is 1.2, we obtain the average router hop count for cacheable data delivery in *cache network* under Zipf distribution as 7, while the average router hop count for non-cacheable data delivery in *cache network* is 16. After we change the value of α, we get the similar value.

C. Experiment Analysis

Since the packet forwarding in *cache network* relies on the content, the delivery delay or resource utilization in hop in *cache network* may be larger than the IP-based packet forwarding in *regular network*, so does the economic cost. So, we set the unit data transfer performance in hop and the economic cost in *cache network* is r times and t times higher than that in the *regular network* respectively. Assume that $\alpha = \beta = 1$ in both networks, the proportion of the non-cacheable traffic in the network is p_1, and the proportion of cacheable traffic is p_2 ($p_1 + p_2 = 1$), we make the following analysis:

(1) The 2ACT of in-network caching mechanism under different data delivery performance

According to Formula (3) and the average router hop count of Abilene network (it is 5), we get:

$$\begin{aligned} U_{regular} &= 5(PHop)_{regular} + U_e \\ U_{cache} &= (5p_1 + H'p_2)r(PHop)_{regular} + tU_e \end{aligned} \quad (4)$$

when $t = 1$, the value of U_e and $(PHop)_{regular}$ will not affect the difference between $U_{regular}$ and U_{cache}. With the parameter r changes, we obtain different results in Abilenet network circumstance as shown in Figure 5.

As shown in Figure 5, the change of traffic proportion has no effect on *regular network* because it has not deployed the in-network caching mechanism, and the 2ACT of *cache network* becomes much better with the increase of the cacheable traffic proportion. When the cacheable traffic reaches a lower bound of 22% and the data transmission performance value in *cache network* is lower than 1.1 times of that in the *regular network*, the in-network caching mechanism performs well. We can also see from Figure 5 that, under some specific data transmission performance, i.e., the parameter r being determinant,

the development potential of *cache network* becomes not so attractive once its economic cost exceeds the shadow area produced by the U_e curve between the *regular network* and the *cache network* in Figure 5.

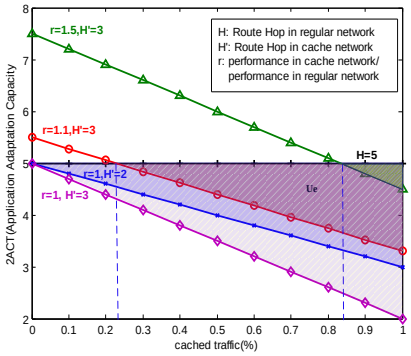

Fig. 5. The 2ACT of the network under different unit data delivery performance.

(2) The 2ACT of in-network caching mechanism under different economic cost.

In order to figure out the relationship between the economic cost and the development of in-network caching mechanism, we further analyze the results coming from the PlanetLab. Assume that $r = 1$, according to Formula (3), we have:

$$\begin{aligned} U_{regular} &= 16(PHop)_{regular} + U_e \\ U_{cache} &= (16p_1 + 7p_2)(PHop)_{regular} + tU_e \end{aligned} \tag{5}$$

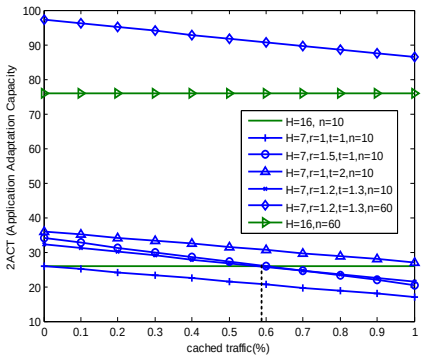

Fig. 6. The 2ACT of the network with different economic cost.

We make some experiments to test the change of 2ACT in different networks with the change of the unit data transfer performance and the ratio of U_e and $(PHop)_{regular}$, which is denoted by parameter n. We get the result as shown in Figure 6. We can see that the difference of the 2ACT between the *regular network* and the *cache network* increases with the increase of parameter r,t and n. When the economic cost and the unit data transfer performance value in the *cache network* is more than 1.2 times and 1.3 times of that in the *regular network* respectively, and the ratio of economic cost and the unit data transfer performance is more than 10 times, the value of 2ACT U_{cache} in the *cache network* is larger than 1. This means the in-network caching mechanism is not suitable for deployment, even if the cacheable traffic in the network has more than 55%.

V. CONCLUSION AND FUTURE WORK

With the fast growing amount of multimedia and video traffic on the Internet, the in-network caching mechanism, the core technology of the content distribution network and the future architecture, has attracted more and more attentions from both academia and industry. However, the advantage of the in-network caching mechanism has not been evaluated in a quantitative manner yet. In this article, we propose a holistic 2ACT model to find the relationship between the proportion of application data, the economic cost and the development of in-network caching mechanism based architecture. The results show that there are some distinct constraints and limitations for in-network caching mechanisms to pose real significant effects on data delivery across the Internet, for example, when the following conditions are satisfied at the same time: 1)The cacheable traffic proportion in the network should reach a lower bound of 22%. 2)The economic cost of deploying and managing the in-network-caching-based network is not higher than that of the regular network without the data cache capacity. 3)The router-level hop count for cacheable data delivery must be less than half of the regular network level. 4)The value of data transfer performance in router hop should be less than 1.1 times of that in regular network.

REFERENCES

[1] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman. A survey of information-centric networking (draft). In *Information-Centric Networking*, 2011.

[2] A. Anand, A. Gupta, A. Akella, S. Seshan, and S. Shenker. Packet caches on routers:the implications of universal redundant traffic elimination. In *ACM SIGCOMM*, October 2008.

[3] L. Breslau, P. Cao, L. Fan, G.Philips, and S. Shenker. Web caching and zipf-like distribution: Evidence and implications. In *IEEE INFOCOM*, March 1999.

[4] J. Choi, J. Han, E. Cho, T. Kwon, and Y.Choi. A survey on content-oriented networking for efficient content delivery. *IEEE Communications Magazine*, 49(3):121–127, March 2011.

[5] V. Jacobson, D. Smetters, N. H. Briggs, M. F. Plass, P. Stewart, J. D. Thornton, and R. L. Braynard. Voccn: Voice-over content centric networks. In *the 2009 Workshop on ReARCH*, December 2009.

[6] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard. Networking named content. In *the 5th international CoNEXT*, December 2009.

[7] S. Paul, R. Yates, D. Raychaudhuri, and J. Kurose. The cache-and-forward network architecture for efficient mobile content delivery services in the future internet. In *First ITU-T Kaleidoscope Academic Conference on Innovations in NGN: Future Networkand Services*, May 2008.

[8] I. Psaras, R. G. Clegg, R. Landa, W. K. Chai, and G. Pavlou. Modeling and evaluation of ccn-caching trees. In *the 10th international IFIP TC6 conference on Networking*, May 2011.

[9] S. Sen, Y. Jin, R. Guerin, and K. Hosanagar. Modeling the dynamics of network technology adoption and the role of converters. *IEEE/ACM Transactions on Networking*, 18(6):1793–1805, December 2010.

[10] T. A. Trinh, L. Gyarmati, and G. Sallai. Migrating to ipv6: A game-theoretic perspective. In *2010 IEEE 35th Conference on Local Computer Networks (LCN)*, October 2010.