Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
April 2013, 20(1): 99–105
www.sciencedirect.com/science/journal/10058885
The Journal of China Universities of Posts and Telecommunications
http://jcupt.xsw.bupt.cn

Modeling in-network caching and bandwidth sharing performance in information-centric networking

WANG Guo-qing, HUANG Tao (✉), LIU Jiang, CHEN Jian-ya, LIU Yun-jie

1. Key Laboratory of Universal Wireless Communications of Ministry of Education, Beijing University of Posts and Telecommunications Beijing 100876, China
2. School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract

Information-centric networking (ICN) proposes a content-centric paradigm which has some attractive advantages, such as network load reduction, low dissemination latency, and energy efficiency. In this paper, based on the analytical model of ICN with receiver-driven transport protocol employing least-recently used (LRU) replacement policy, we derive expressions to compute the average content delivery time of the requests' arrival sequence of a single cache, and then we extend the expressions to a cascade of caches' scenario. From the expressions, we know the quantitative relationship among the delivery time, cache size and bandwidth. Our results, analyzing the trade-offs between performance and resources in ICN, can be used as a guide to design ICN and to evaluation its performance.

Keywords ICN, LRU, miss probability, content delivery time

1 Introduction

The Internet is originally designed as a communication substrate enabling the delivery of data between pairs of end-hosts. With the exponential growth of digital information diffusing over the Internet, this Internet architecture based on the end-point appears to be unsuited to deal with the trends.

In order to solve the problem about the explosion of traffic, in the last years, many significant research projects have been funded focusing on the definition of novel architectures for the future Internet, both in Europe (publish-subscribe Internet routing paradigm (PSIRP) [1], architecture and design for the future Internet (4WARD) [2–3], publish subscribe internet technology(PURSUIT) and scalable and adaptable internet solutions (SAIL) and in the US (content-centric networking (CCN) [4] and data-oriented network architecture (DONA) [5]). Although these approaches differ with respect to their specific architectures, they all have reached a common view that content is the first class network citizen in the next-generation internet.

CCN proposes architecture in the computer networks centered with the content that the user needs, and packets of content can be transparently cached in the scattered routers. To develop a new Internet foundation and design a CCN environment, ICN [6] has emerged recently. The most important factors of such a network are the availability of the embedded storage and receiver-driven transport, the interaction between which impacts the performance of the whole system.

To quantify potential benefits and to guide optimized protocol design in ICN, it is necessary to understand the transport and storage issues. In this paper, we focus on such a system. Content items are permanently stored in one repository (or server) in the form of chunks. Every router in the network is equipped with a caching storage, and a check keeps track of pending queries. Users can retrieve chunks by sending requests. If a request reaches a router which does not store the content requested, we call the request miss at the router, and then it will be forwarded to the upstream routers; if the requested content is cached

Received date: 13-12-2012
Corresponding author: HUANG Tao, E-mail: htao@bupt.edu.cn
DOI: 10.1016/S1005-8885(13)60035-7

in the router, the request will hit it, and the content will be sent back following with the coming path of the request, and cached in the routers along the path to the user employing LRU replacement policy.

In the setting of network caching, there have been researched about modeling content-level cache dynamics before, and most of them related to a single cache scenario under LRU cache replacement policy. In Ref. [7], Jelenkovic gave an asymptotic characterization of the LRU miss probabilities both in the light-tailed and in the heavy-tailed case for a large number of contents. In 2008, he and Kang studied the miss sequence of LRU cache replacement policy, and provided an analytical characterization of the miss probability under Poisson's assumptions of content requests' arrivals [8]. Recently, Tofis et al. have derived an expression to compute the total average network queuing delay in CCN [9] by combinatorial analysis in LRU caching scheme, which has exponential complexity. Carofiglio et al [10–12] have, firstly, developed an analytical model and derived closed-form expressions for the average content delivery time and miss probabilities based on network level and chunk-based system as described above. Their goal is to get the average content delivery time by computing the miss probability. However, they only study the miss sequence of the requests, and ignore that the residual virtual round trip time (RVRTT) also impacts on the miss probability. In the ICN system, when a request for a chunk (we denoted it by A for convenience) misses, a following request for the same chunk A will also miss in two cases:

1) The arrival time interval between the two requests for chunk A is greater than RVRTT, and the sum of different chunks arrived is larger than the cache size.

2) The arrival time interval between the two requests for chunk A is less than RVRTT.

In the research of LRU caching scheme, they have only considered the first case. As we known, RVRTT is decided by the distribution of resources in ICN and it is greater than zero. So, the second case will happen and it cannot be ignored in ICN although RVRTT is small sometimes. To the best of our knowledge, our paper is the first attempt to considering both cases to compute the miss probability in ICN with the LRU replacement policy. In this paper, from a practical perspective, we take attention to the sequence of the arrival time of the requests rather than the sequence of the miss time of the requests as in Ref. [10], and develop a Markov chain to derive the miss probability in steady states, thus we get the average content delivery time. In the process of proof, we find that the miss probability is influenced by the RVRTT which decides the average content delivery time directly, which means miss probabilities and the average content delivery time influence each other, and we get them by iterating.

2 Model description

We consider a set of M different content items equally partitioned in K classes, and each class has a popularity of $q_k, k = 1, 2, ..., K$. We assume a Zipf popularity distribution, $q_k = c/k^{\alpha}$ with parameter $\alpha > 0$ and $c = 1\Big/\sum_{k=1}^{K}\left(1/k^{\alpha}\right)$. Content items are segmented into chunks and have different sizes with geometric distribution: σ denotes the average content size in terms of number of chunks (chunks are fixed size of P bit). Each node in the network has a cache size of x chunks. The repository has an infinite cache size, as it stores all content items. We focus on a topology as shown in Fig. 1, assuming the path is the best routing for the users to get the required contents.

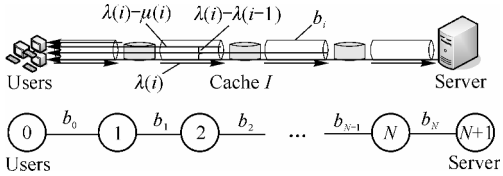


Fig. 1 Linear topologies with bandwidth-limited down-link

The request arrival process is modeled through a Markov modulated rate process (MMRP) [10] where requests for content items in class k are generated according to a Poisson process of intensity $\lambda_k = \lambda q_k$, and the content being requested is uniformly chosen among the m different content items in class k (i.e. each class has the same number of content items $m = M/K$, and the items in the same class has the same requested rates). The content is split into chunks, and a content request coincides with the request of the first chunk of the content. A receiver can download data from multiple nodes (routers or server) and share bandwidth with other concurrent transfers. We consider that bandwidth is fairly shared among each transfer in the max-min sense (Fig. 1). We assume that all the significant source of delay is generated at the bottleneck link encountered along the path, and consider any constant propagation delay negligible either in the upstream or downstream. Therefore, the round trip

delay of a chunk from the node 0 (i.e. users) to node $(i+1), R(i)$ is,

$$R(i)=PD_{\text{up}}+PD_{\text{down}}+\frac{P}{\gamma(i)}\approx\frac{P}{\gamma(i)};\quad i=0,1,...,N \tag{1}$$

where the nodes 0 and $(N+1)$ denote the users and the server respectively, and $\gamma(i)$ is the bandwidth assigned to node $(i+1)$ in bit/s, D_{up} and D_{down} are propagation delay time, P is the chunk size. As the trip from the node 0 to the node I is shorter than that to the node $(i+1)$, so we have $R(i-1)\leqslant R(i),\ i=1,2,...,N$. We define virtual round trip time (VRTT) of a content chunk in class k as Δ_k, and it means the average time that elapses from the dispatch of a chunk request to the chunk reception in steady states [10],

$$\Delta_k=\Delta_{\text{VRTT}_k}=\sum_{i=0}^{N}R(i)(1-p_k(i+1))\prod_{j=1}^{i}p_k(j);\quad k=1,2,...,K \tag{2}$$

where $p_k(i)$ denotes the miss probability for class k at node i. Similarly, we define the RVRTT at node i as

$$\Delta_k(i)=\Delta_{\text{RVRTT}_k(i)}=\sum_{j=i}^{N}(R(j)-R(i-1))(1-p_k(j+1))\prod_{l=i+1}^{j}p_k(l);\quad i=1,2,...,N \tag{3}$$

$\Delta_k(i)$ represents the RVRTT that one would have if node i would be the requester, and make $\Delta_k(0)=\Delta_k$.

From the above descriptions, the average delay time of content k is the product of σ and Δ_k (the number of chunks that a content having been split multiplies Δ_k). Now, the problem will turn to calculate the miss probability $p_k(i)$. Thus, we deduce the expressions of miss probability following.

3 Single cache model

3.1 Theoretical analysis

For simplicity, let us study the single cache model with $N=1,\ \sigma=1,\ m=1$ firstly, under the assumptions and notations in Sect. 2.

Proposition 1 Given a MMRP request arrival process as described in Sect. 2 with content size of 1 chunk (i.e. $\sigma=1$), and each class has 1 content (i.e. $m=1$), then the stationary miss probability for contents of class k, p_k, is given by

$$p_k\equiv p_k(1)\sim \mathrm{e}^{-\lambda_k(gx^{\alpha}-\Delta_k(1))} \tag{4}$$

for large x, where $1/g=\lambda c\Gamma(1-1/\alpha)^{\alpha}$ and e, $\Gamma()$ is the natural base number and gamma function respectively.

Proof As described in Sect. 2, contents are requested at time sequence $\{\tau_n\}_{n\geqslant 0}$, with increments $\{\tau_{n+1}-\tau_n\}_{n\geqslant 0}$, $\tau_0=0$, following exponential distribution of rate λ. For a certain content k (or class k, as $m=1$), it is requested at time sequence $\{\tau_n^k\}_{n\geqslant 0}$, which is a subsequence of $\{\tau_n\}_{n\geqslant 0}$, with exponential distribution of rate λ_k. We assume the state is 1 if there is a content k in the cache, otherwise, the state is 2 when a request for content k arrives. Or, we can define

$$X(\tau_n^k)=\begin{cases}1; & \text{hit at time } \tau_n^k\\ 2; & \text{miss at time } \tau_n^k\end{cases} \tag{5}$$

Clearly, the sequence $\{\tau_n^k\}_{n\geqslant 0}$ is a homogeneous Markov chain [8], and the arriving sequence of content in class k is showed by Fig. 2. We can notate $\{\tau_{n+1}^k-\tau_n^k\}_{n\geqslant 0}$ as τ^k, and

$$\Pr(X(\tau_n^k)=j\,|\,X(\tau_{n-1}^k)=i)=p_{ij}^k;\quad i,j=1,2 \tag{6}$$

then we obtain the transition probability matrix of the Markov chain

$$\boldsymbol{P}^k=\begin{bmatrix}p_{11}^k & p_{12}^k\\ p_{21}^k & p_{22}^k\end{bmatrix}$$

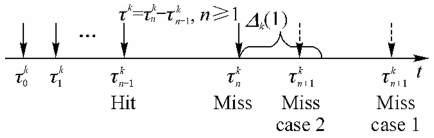


Fig. 2 The arrival time sequence of class k and the two miss cases

As $p_{ij}^k>0, i,j=1,2$, there is a unique $\pi^k=(\pi_1^k,\pi_2^k)$, s.t. $\pi^k\boldsymbol{P}^k=\pi^k$. Solve the equation, and with normalization $\pi_1^k+\pi_2^k=1$, we have the miss probability

$$p_k=\pi_2^k=\frac{p_{12}^k}{1-(p_{22}^k-p_{12}^k)} \tag{7}$$

then, we define $H^k(u,t)$ as the number of requests for content k in open interval (u,t), $H^k(u,t)\sim$ Poisson$(\lambda_k(t-u))$, and $B^k(u,t)=1_{H^k(u,t)>0}$ the Bernoulli variable associated to the event that at least one content k is requested in the open interval (u,t) with $\Pr[B^k(u,t)=1]=1-\mathrm{e}^{-\lambda_k(t-u)}$. $S(u,t)$ denotes the number of different contents requested in (u,t), and $S(u,t)=\sum_{k=1}^{K}B^k(u,t)$.

As proved in Ref. [10], when $x\to\infty$, we get

$$\left.\begin{aligned} p_{12}^k &= \Pr[S(\tau^k)\geqslant x] \sim \Pr[\tau^k \geqslant gx^\alpha] = \mathrm{e}^{-\lambda_k g x^\alpha} \\ \Pr(S(\tau^k)&\geqslant x, \tau^k < \Delta_k(1)) \sim \Pr(gx^\alpha \leqslant \tau^k < \Delta_k(1)) \end{aligned}\right\} \quad (8)$$

where $1/g = \lim_{t\to\infty} E[S(0,t)]^\alpha / t = \lambda c \Gamma\left[1-(1/\alpha)\right]^\alpha$, $\alpha > 1$, and $\lim_{t\to\infty} E[S(0,t)]/t$ denotes the average number of different contents requested per unit time.

$$\begin{aligned} p_{22}^k - p_{12}^k &= \Pr(\tau^k < \Delta_k(1)) + \Pr(S(\tau^k)\geqslant x, \tau^k \geqslant \Delta_k(1)) - \\ &\Pr(S(\tau^k)\geqslant x) \sim \Pr(\tau^k < \Delta_k(1)) = 1 - \mathrm{e}^{-\lambda_k \Delta_k(1)}; \\ &x \to \infty \end{aligned} \quad (9)$$

put Eqs. (8) and (9) into Eq. (7), we have

$$p_k = \pi_2^k = \frac{p_{12}^k}{1-(p_{22}^k - p_{12}^k)} = \mathrm{e}^{-\lambda_k (g x^\alpha - \Delta_k(1))}$$

The expressions Eqs. (1), (3) and (5) give the relationship between the miss probability and RVRTT, both of which are decided by the content popularity, the cache size and bandwidth.

Now let us extend the results above to a more general situation that content size is geometrically distributed with the average of σ chunks, and $m = M/K$, just as the assumptions described in Sect. 2.

Proposition 2 Given a MMRP request arrival process as described in Sect. 2 with intensity λ, popularity distribution $q_k = c/k^\alpha, \alpha > 1, c > 0$, and average content size σ, $x > 0$ is the cache size in number of chunks, then the stationary miss probability for chunks of class k, p_k, is given by

$$p_k \equiv p_k(1) \sim \mathrm{e}^{-\lambda_k (g x^\alpha - \Delta_k(1))/m} \quad (10)$$

for large x, where $1/g = \lim_{t\to\infty} (1/t) E[S(0,t)]^\alpha = \lambda c \sigma^\alpha m^{\alpha-1} \cdot \Gamma\left[1-(1/\alpha)\right]^\alpha$, $\alpha > 1$.

The proof is similar to the proof of Proposition 1, because each class has the same number of contents with the same popularity, and g is given in [10].

In ICN delivery systems, a user issues a request which is forwarded through the network until the given content is found. To avoid requests' flooding, CCN sets up a pending interest table (PIT) [4] in each router which can prevent the dispatch of new requests for the same content if the former request has not returned. We assume the effective timescale of aggregation for requests of content k is $\Delta_k(1)$. In order to compute the filtering miss rate in the situation, we give the following lemma which is proved in Ref. [10].

Lemma 1 Given the content request process defined in Sect. 2, the timescale of aggregation is $\Delta_k(1)$, then the filtering probability associated to class k at the first node is,

$$p_k^{\text{filter}}(1) = \frac{1-a_k}{1-\left(1-\frac{1}{\sigma}\right)a_k} \quad (11)$$

with $a_k = \mathrm{e}^{-\Delta_k(1)\lambda_k}$.

Thanks to Lemma1, we get the filtering miss rate easily.

Proposition 3 Given a MMRP request arrival process as described in Sect. 2 with intensity λ, popularity distribution $q_k = c/k^\alpha, \alpha > 1, c > 0$, and average content size σ, $x > 0$ is the cache size in number of chunks, and the timescale of aggregation is $\Delta_k(1)$, then the stationary miss rate of class k with filtered, μ_k^f, is given by

$$\mu_k^f = \mu_k^f(1) = (1 - p_k^{\text{filter}}(1)) p_k \lambda_k \quad (12)$$

In the expression, $(1-p_k^{\text{filter}}(1))\lambda_k$ can be looked as the request rate at node 1 with filtering, and μ_k^f represents the request rate with filtering at the following upstream node.

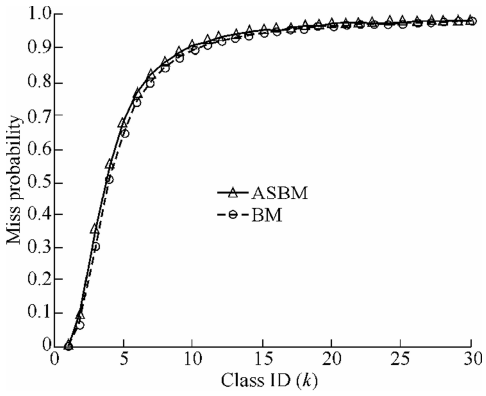
3.2 Numerical results

Now we consider a population of $M = 20\,000$ content items, organized in $K = 400$ classes of decreasing popularity, each one with $m = 50$ items. Content popularity follows Zipf distribution, i.e. content items in class $k (k = 1, 2, ..., K)$ are requested with probability $q_k = c/k^\alpha, c > 0$, with $\alpha = 2$ (the value of α has no impact on the problems we care if $\alpha > 1$). As the contents in the same class have the same popularities, a given content in class k is requested with probability q_k/m. We propose content items are split into chunks of 10 KB each, i.e. $P = 10$ KB, and their sizes are geometrically distributed with average 690 chunks (6.9 MB).

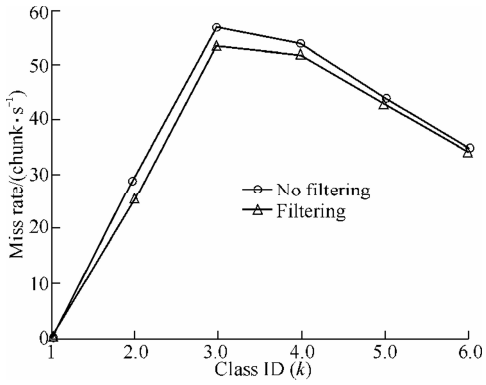
Users generate content requests according to a Poisson process of intensity $\lambda = 4$ content/s, and the chunk transmission window size is W=1. We suppose a cache of size $x = 200\,000$ chunks (2 GB) which implements the LRU replacement policy.

In Fig. 3(a), we show the difference of the miss probability between the arrival sequence based model (ASBM) and the base model (BM) in Ref. [10]. Results show that the miss probabilities of our new model are larger than that of the base model. The reason is that our new model considers one more miss case besides the base model, which corresponds with the actual condition better. We let $\boldsymbol{B}$=(300,30) Mbit/s in Fig. 3(b) which reports the difference of miss rate between filtering and no filtering of the 6 most popular classes. It shows that the miss rate is decreased with filtering. We consider $\Delta = 2$ s in

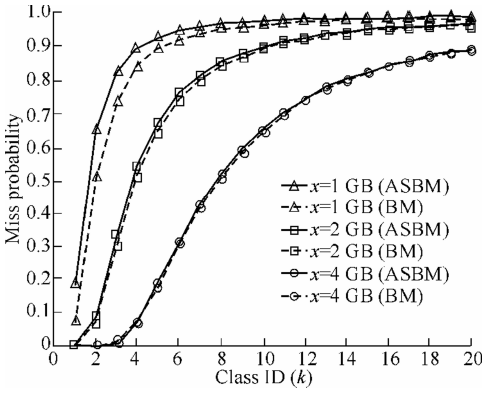
Fig. 3(c) which illustrates the impact of cache size on the miss probability. As expected, miss probability decreases as cache size increases. Meanwhile, the miss probability of the new model is also larger, which also corresponds with the actual condition. Moreover, the gap of miss probability between the two models is increasing as the cache size decreasing, which means the effect of RVRTT in the new model becomes clearer.



(a) Miss probability as a function of class popularity



(b) Miss rate with and without request filtering



(c) Miss probability for different cache sizes

Fig. 3 Numerical results of one routers

4 A cascade of *N* caches

4.1 Theoretical analysis

The output of the first node can be well approximated by a renewal process [10]. Thus a miss process at the *i*th node, which constitutes the input process for caches at the $(i+1)$ th node, has an intensity

$$\mu(i)=\lambda(i)\Sigma_k p_k(i)q_k(i);\quad i=0,1,...,N+1 \tag{13}$$

where $\mu(0)=\lambda(1)=\lambda$, and $\mu(N+1)=0$.

As illustrated in Fig. 1, down-link $i\geqslant 0$ is shared by all routers $j, j>i$. We denote the traffic load at link *i* as ρ_i,

$$\rho_i=\frac{\sigma P}{b_i}\sum_{j=i}^{N}(\mu(j)-\mu(j+1))=\frac{\sigma P}{b_i}\sum_{k=1}^{K}\lambda_k(i)p_k(i) \tag{14}$$

We assume ρ_i satisfy the stability condition $\rho_i<1,\ \forall i=0,1,...,N$. Under the assumptions in Sect. 2, we have the round trip delay of a chunk between users and node $(i+1), R(i)$,

$$R(i)\approx\frac{P}{\gamma(i)}=\frac{P}{\wedge_{j\ni i}b_j(1-\rho_j)} \tag{15}$$

where $j\ni i$ denotes the set of links from the user to the $(i+1)$ th node along the path that goes up to the repository, and '$\wedge$' denotes the minimum operator, as the delay is decided by the bottleneck bandwidth [12]. Thus, we give the expressions of the miss probability following.

Proposition 4 Given a cascade of *N* caches as in Fig. 1 and a MMRP content request process with rate $\lambda(i)$ and the average delivery time for a chunk of content items in popularity class *k* at node $i, i\geqslant 1$, then it holds

$$\lg p_k(i)=\frac{gx^{\alpha}-\Delta_k(i)}{gx^{\alpha}-\Delta_k(1)}\prod_{l=1}^{i-1}p_k(l)\lg p_k(1);\quad i=1,2,...,N \tag{16}$$

where $p_k(1)$ is given by Eq. (10), and $p_k(N+1)=0$.

From the Eq. (16) we know that the miss probabilities at the upstream routers are decided by the downstream routers, and then they are all decided by the miss probability at the first router. Thus, we can use the same method to derive miss probability functions in any hierarchical networks easily.

Now let us study the miss probability with filtering. The effective timescale of aggregation for requests of class *k* at node *i* is $\Delta_k(i)$, for simplicity, we only consider that the requests are aggregated only at the first cache, the request process is not filtered at the following nodes, as only the first router (i.e. the node 1) in our model will receive the same content requests before the requested content coming back. Then, we have

Proposition 5 Given a cascade of *N* caches as in Fig. 1, a MMRP content request arrival process as described in Sect. 2, and a timescale aggregation for content requests is $\Delta_k(1)$, then it holds

$$\lg p_k^f(i) = (1 - p_k^{\text{filter}}(1)) \prod_{l=2}^{i-1} \frac{p_k^f(l)}{p_k(l)} \frac{gx^\alpha - \Delta_k^f(i)}{gx^\alpha - \Delta_k(i)} \lg p_k(i);$$
$$i = 2,3,...,N \quad (17)$$

where $\Delta_k^f(i) = \sum_{j=i}^{N} (R(j) - R(i-1))(1 - p_k^f(j+1)) \prod_{l=i+1}^{j} p_k^f(l)$, $i = 0,1,...,N$.

Due to lack of space we do not proof the Propositions 4–5 here.

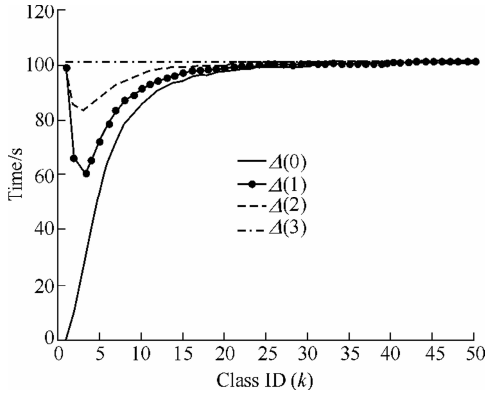
The Eq. (17) shows that the miss probabilities at the upstream routers are decided by the downstream routers, and then they are all decided by the miss probability at the first router with filtering.

The Propositions 4–5 give the miss probability of content k at node i without and with filtering, which can be used in Eqs. (2) and (3) to calculate the delay of fetching a chunk of users. From the expressions we can get that the delay time is decided by the cache size and the bandwidth.

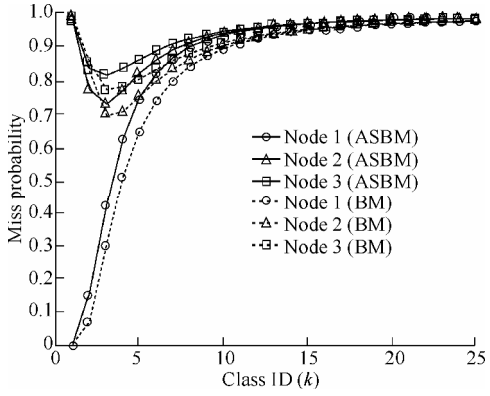
4.2 Numerical results

Let us now consider $N = 3$, $\rho \to 1-$ (the heavier of the traffic load, the larger of the delay, then the case 2 will happen more likely), $\lambda = 4$ content/s, other parameters are all the same as in Sect. 3.

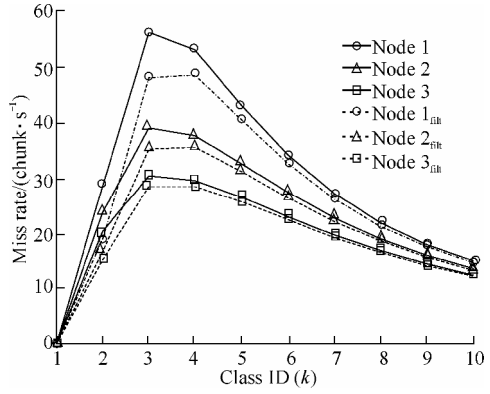
In Fig. 4(a), we show the delay of a contents' chunk in class k at node i, $\Delta_k(i), i = 0,1,2,3$ as a function of class popularity. We find that the requests of the more popular classes are more likely to be satisfied at the first three nodes. The class which has minimum $\Delta(i)$ implicates most of the contents in the class satisfied by node $(i+1)$ $(i = 0,1,2)$. $\Delta(3)$ is a horizon because all the miss requests from the node 3 can be satisfied by the repository which stores all the contents. These are consistent with the Fig. 4(b), which compares the miss probabilities at different nodes as a function of classes of ASBM and bm. The real lines of ASBM are higher than that the dotted lines. The reason is that our new model considers one more miss case besides the base model, which appears frequently when the traffic load is heavy. The lines of node 2 and 3 which decreases at the first three classes and then increases indicates that contents of class 1 are not cached in node 2 and 3, as class 1 serviced by node 1 are not so popular at node 2 and 3. We can also notice that the miss probability of the contents in class ID ($k > 10$) is very great, and this is the flaw of the caching scheme which is one of our future work. Fig. 4(c) illustrates that the miss rates decrease when requests are filtered at each one node and the capacity of links is $\boldsymbol{B} = (300,200,100,10)$ Mbit/s.



(a) The delivery time of a content chunk as a function of class popularity



(b) Miss probability at different nodes without filtering



(c) Miss rate with and without filtering

Fig. 4 Numerical results of a cascade of routers

5 Performance trade-offs

Bandwidth and storage capacity are the most critical resources in the information-centric architecture, and represent the network cost for an operator to deploy this kind of distribution infrastructures. We always expect to derive a direct relationship between resources and

performance. From above, we derive the average content delivery time experienced by end-users

$$T_k(0)=\sigma\Delta_k=\sigma\sum_{i=0}^{N}R(i)(1-p_k(i+1))\prod_{j=1}^{i}p_k(j) \tag{18}$$

and the average content delivery time at node i

$$T_k(i)=\sigma\Delta_k(i)=\sigma\sum_{j=i}^{N}(R(j)-R(i-1))(1-p_k(i+1))\prod_{l=i+1}^{j}p_k(l);\quad i=1,2,...,N \tag{19}$$

We let $\boldsymbol{B}_1=(100,70,40,8)$ Mbit/s, $\boldsymbol{B}_2=(200,100,50,9)$ Mbit/s, $\boldsymbol{B}_3=(300,200,100,10)$ Mbit/s in Fig. 5 which shows the impact of bandwidth on the delivery time. They are all consistent with our intuition that the delivery time decreases with the increasing of the bandwidth. The figure also shows the trade-offs between cache size and bandwidth. Intersection1 denotes that {x=1 GB, $\boldsymbol{B}_1$} has the same performance of {x=2 GB, $\boldsymbol{B}_1$} for the contents in class 11, and Intersection2 denotes that {x =2 GB, $\boldsymbol{B}_3$} has the same performance of {x=4 GB, $\boldsymbol{B}_1$} for the contents in class 20.

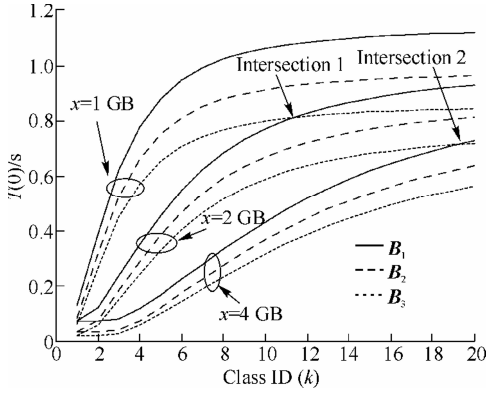


Fig. 5 Average delivery time with different bandwidths and cache sizes

6 Conclusions

As the content is considered as the center of the next-generation network, the trade-offs between performance and resources is an orientation to design the protocols of the future Internet. Based on ICN, we give a new way to compute the miss probabilities. We also derive the expressions of miss rate with filtering. For the average content delivery time and the miss probabilities at different nodes influence each other, we iterate the expressions and get the steady solution. Finally, we analyze the trade-offs between resources and performance briefly. Our results provide a more exact quantitative method to compute the content delivery time, which is a foundation to optimize resources' allocation of the network. We also provide a research method to analysis other similar networks.

ICN is treated as a new paradigm in the future Internet, a comprehensive study of which still lacks, and key features are poorly understood. In the future, we can study more complex network with multiple-paths to route content requests. Other possible future research directions are the study of different cache replacement policies and optimizing the locations of storage and bandwidth.

Acknowledgements

This work was supported by the National Basic Research Program of China (2012CB315801, 2011CB302901), and the Fundamental Research Funds for the Central Universities (2011RC0118).

References

1. Jokela P, Zahemszky A, Rothenberg C E, et al. LIPSIN: line speed publish/subscribe inter-networking. Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'09), Aug 17–21, 2009, Barcelona, Spain. New York, NY, USA: ACM, 2009: 195–206
2. Ahlgren B, D'Ambrosio M, Dannewitz C, et al. D–6.2 second NetInf architecture description. The FP7 4WARD Project. 2010
3. Ahlgren B, D'Ambrosio M, Marchisio M, et al. Design considerations for a network of information. Proceedings of the 4th ACM International Conference on Emerging Networking Experiments and Technologies (CoNext'08), Dec 9–12, 2008, Madrid, Spain. New York, NY, USA: ACM, 2008
4. Jacobson V, Smetters D K, Thornton J D, et al. Networking named content. Proceedings of the 5th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT'09), Dec 1–4, 2009, Rome, Italy. New York, NY, USA: ACM, 2009
5. Koponen T, Chawla M, Chun B G, et al. A data-oriented (and beyond) network architecture. Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'07), Aug 27–31, 2007, Kyoto, Japan. New York, NY, USA: ACM, 2007: 181–192
6. Ahlgren B, Dannewitz C, Imbrenda C, et al. A survey of information-centric networking (draft). Ahlgren B, Karl H, Kutscher D, et al. Information-centric Networking. Dagstuhl, German: Schloss Dagstuhl, 2011
7. Jelenkovic P R. Asymptotic approximation of the move-to-front search cost distribution and least recently-used caching fault probabilities. The Annals of Applied Probability, 1999, 9(2): 430–464
8. Jelenkovic P R, Kang X. Characterizing the miss sequence of the LRU cache. ACM SIGMETRICS, 2008, 36(2): 119–121
9. Tofis Y, Psaras I, Pavlou G. Modeling queuing delays in content-centric networks. London, UK: Univerity College London, 2011
10. Carofiglio G, Gallo M, Muscariello L, et al. Modeling data transfer in content-centric networking (extended version). Technical Report. http://perso.rd.francetelecom.fr/muscariello/report\\-itc- transport.pdf
11. Carofiglio G, Gallo M, Muscariello L, et al. Modeling data transfer in content centric networking. Proceedings of the 23rd International Teletraffic Congress (ITC'11), Sep 6–9, 2.011, San Francisco, CA, USA. Piscataway, NJ, USA: IEEE, 2011: 111–118
12. Muscariello L, Carofiglio G, Gallo M, et al. Bandwidth and storage sharing performance in information centric networking. Proceedings of the ACM SIGCOMM Workshop on Information-centric Networking (ICN'11), New York, NY, USA: ACM, 2011: 26–31

(Editor: WANG Xu-ying)