SCIENCE CHINA
Information Sciences

P2P traffic optimization

GuoQiang ZHANG[1,2,*], MingDong TANG[3], SuQi CHENG[2,4], GuoQing ZHANG[2], HaiBin SONG[5], JiGuang CAO[6] and Jing YANG[7]

Citation: SCIENCE CHINA Information Sciences **55**, 1475 (2012); doi: 10.1007/s11432-011-4464-8

View online: https://engine.scichina.com/doi/10.1007/s11432-011-4464-8

View Table of Contents: https://engine.scichina.com/publisher/scp/journal/SCIS/55/7

Published by the Science China Press

SCIENCE CHINA
Information Sciences

• REVIEW •

July 2012 Vol. 55 No. 7: 1475–1492
doi: 10.1007/s11432-011-4464-8

P2P traffic optimization

ZHANG GuoQiang[1,2*], TANG MingDong[3], CHENG SuQi[2,4], ZHANG GuoQing[2], SONG HaiBin[5], CAO JiGuang[6] & YANG Jing[7]

[1]*School of Computer Science and Technology, Nanjing Normal University, Nanjing* 210046*, China;*
[2]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan* 411201*, China;*
[4]*Graduate University of Chinese Academy of Sciences, Beijing* 100190*, China;*
[5]*Nanjing R&D Center, Huawei Technologies, Nanjing* 210012*, China;*
[6]*China Academy of Telecommunication Research of MIIT, Beijing* 100191*, China;*
[7]*China Mobile Research Institute, Beijing* 100053*, China*

Received December 8, 2010; accepted April 21, 2011; published online December 21, 2011

Abstract Peer-to-peer (P2P) based content distribution systems have emerged as the main form for content distribution on the Internet, which can greatly reduce the distribution cost of content providers and improve the overall system scalability. However, the mismatch between the overlay and underlay networks causes large volume of redundant traffic, which intensifies the tension between P2P content providers and ISPs. Therefore, how to efficiently use network resources to reduce the traffic burden on the ISPs is crucial for the sustainable development of P2P systems. This paper surveys the state-of-art P2P traffic optimization technologies from three perspectives: P2P cache, locality-awareness and data scheduling. Technological details, comparison between these technologies and their applicabilities are presented, followed by a discussion of the issues that remain to be addressed and the direction of future content distribution research.

Keywords P2P traffic localization, P2P traffic optimization, P2P cache, locality-awareness, network coding, content-centric network

Citation Zhang G Q, Tang M D, Cheng S Q, et al. P2P traffic optimization. Sci China Inf Sci, 2012, 55: 1475–1492, doi: 10.1007/s11432-011-4464-8

1 Introduction

Peer-to-peer based content distribution systems, e.g., Kazaa, Gnutella, Emule, BitTorrent, PPLive, and PPStream, have attracted increasing popularity among Internet users in recent years. Compared with traditional client-server or CDN modes, P2P based content distribution significantly reduces the distribution cost of content providers, enhances the system scalability, and improves the adaptability to network dynamics of these systems. P2P systems achieve high system scalability by taking full advantage of the bandwidth resources of end users. In these systems, the system throughput increases with the growth of user population.

*Corresponding author (email: guoqiang@ict.ac.cn)

However, the ever-increasing tremendous traffic generated by P2P applications restricts its further development. P2P traffic has already surpassed web traffic, becoming the dominating contributor to the overall network traffic. Moreover, its growth rate is not expected to be slowed down [1–3]. Presently, end users and content providers benefit quite a lot from the prevalence of P2P applications. End users can experience fast downloading speed, and content providers can reduce their distribution cost. The only victim is ISP. Large volume of P2P traffic makes it hard for ISPs to do effective traffic control and management. Many traffic engineering techniques are of little use in coping with P2P traffic. P2P traffic typically crosses network boundaries multiple times, which increases the expenses ISPs have to pay. Even for ISPs with peering relationships, P2P traffic could result in traffic imbalance between them, and disrupt the peering agreement [4, 5]. In addition, sharp rise of P2P traffic (especially cross-domain P2P traffic) consumes large portion of network bandwidth, which severely affects the normal operation of other services. Though ISPs can balance their investment/revenue by improving users' access fee, it faces the potential of losing users.

One cause of the tremendous P2P traffic is the mismatch between the overly topology and the underlay topology, which results in inefficient use of network resources. The tension between P2P content providers and ISPs can only be mitigated by optimizing the resource usage. So far, the solutions developed can be categorized into three stages. At the earliest stage, ISPs use DPI to identify the P2P traffic, and unilaterally throttle, shape or rate limit it. In response, P2P content providers use dynamic port and message encryption to hide their traffic. This leads to a vicious cycle, which is dramatically termed as a war between the two sides in the IETF. In the second stage, both sides independently adopt some positive approaches to make better use of network resources. P2P content providers rely on reverse-engineering techniques to infer the underlying network topology and state, and build overlay topologies that match the underlying ones accordingly. ISPs can localize the traffic by caching the content, or, they can influence the overlay topology setup by deployment of proxy trackers. Nevertheless, these non-cooperative approaches have their inherent limitations. First, reverse engineering alone cannot acquire accurate and fine-grained network topology and state information, while ISPs are at the right position of offering these information. Second, deploying caches violates the economic rules. In fact, ISPs are reluctant to deploy caches because they cannot get revenue from doing it. Caches are considered to be the distribution cost transition from content providers to ISPs. Finally, there are copyright issues of caching content. After recognizing these limitations, recently, P2P content providers and ISPs try to optimize the network resource usage through cooperation. Under this architecture, ISPs provide network topology and state information as services, and P2P content providers use these services to optimize the topologies and data scheduling. In this way, ISPs can offer diverse services by controlling the granularity of exposed network topology and state information, achieve privacy protection by anonymizing the information, and most importantly, get revenue by service provision. P2P content providers can rely on the services to optimize its resource scheduling, which reduces the probe overhead and avoids the potential of being blocked by ISPs.

P2P traffic includes the control flow and data flow, among which real data transmission dominates [6]. So, this paper focuses on optimizing real data transmission, and will review the technologies from three aspects: P2P cache, locality-awareness and data scheduling.

The remaining of this paper is organized as follows: Section 2 briefly reviews the architecture and principle of P2P based content distribution systems. Section 3 discusses P2P cached based traffic optimization techniques. Section 4 discusses locality-awareness based P2P traffic optimization techniques, and section 5 discusses the impact of data scheduling on P2P traffic localization. Finally, section 6 concludes the work and points out future research directions.

2 Introduction to P2P content distribution systems

There are many ways to classify P2P content distribution systems. According to whether the content to be distributed has real-time requirement, they can be classified into real time streaming systems (e.g., PPLive and PPStream) and non-real-time file distribution systems (e.g., Gnutella, Kazaa, and BitTor-

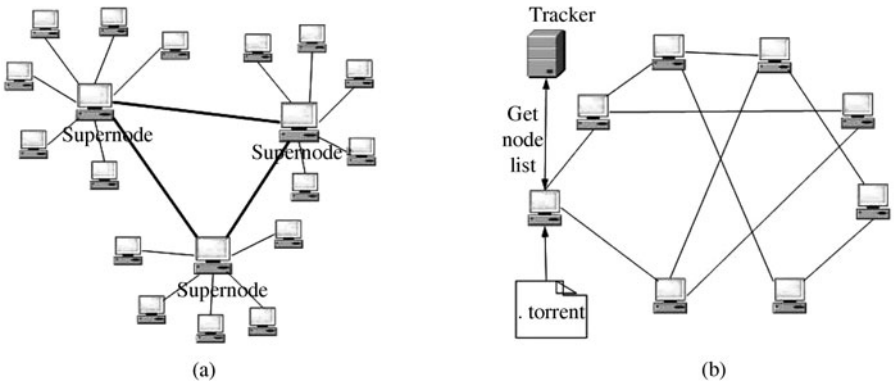

Figure 1 (a) Hybrid structure of Kazaa; (b) illustration of the BitTorrent system.

rent). According to the overlay network structure, they can be classified into structured P2P systems and unstructured P2P systems. Structured P2P systems are established on the distributed hashtable (DHT) technology, e.g., Chord [7], Pastry [8]. These systems have low search cost but high topology maintenance overhead, and they cannot efficiently support keyword based fuzzy search. Presently, most P2P traffic comes from unstructured P2P systems, so this paper only focuses on this kind of content distribution systems. According to different technologies and the emerging time, unstructured P2P systems can be further classified into three generations. Napster is the representative of the first generation, which includes a centralized index server. The second generation includes fully distributed Gnutella, hybrid Gnutella and Kazaa. BitTorrent and PPLive are typical representatives of the third generation. In the following, we will use Kazaa, BitTorrent and PPLive as examples to briefly analyze the architecture and principle of P2P content distribution systems, and present technology distinctions between them. These distinctions determine the applicability of different P2P traffic optimization technologies.

2.1 Kazaa

Kazaa is a representative of the hybrid P2P system. There are two kinds of node in these systems: ordinary node and supernode. Ordinary nodes only establish connections with supernodes, and supernodes index content for ordinary nodes. Supernodes interconnect with each other to form a cooperative network, as is shown in Figure 1(a). In Kazaa, a file is identified by its content, which means two files with same content will have same identifier. Kazaa client establishes persistent connection with the supernode. When it wants to download a file, it first sends a query to its supernode, then this supernode cooperates with other supernodes to complete the query and returns the query results to the client. The requesting node then sets up connections with some nodes in the result list, requesting different parts of the file, until the whole file is downloaded. Once the user has downloaded the entire file, it will update its state to its supernode to share this file.

2.2 BitTorrent

BitTorrent is an open protocol which is supported by various systems, such as BitComet, Vuze, and Xunlei. These systems are the most popular P2P file sharing systems today, and are contributing the largest amount of traffic. A BT-based system consists of three entities: a torrent file, a Tracker and the participating nodes. A torrent file contains the metadata of the file to be shared, such as file length, hash value, name and tracker's address. A tracker maintains the information of participating nodes.

Figure 1(b) exemplifies a BT system. If a user wants to share a file, it needs to generate a torrent file and publish it by Web or other means at first. When a user node wants to download a file, it first obtains the torrent file, then interacts with the tracker to get necessary information to join the cooperative network, and finally finishes the downloads through this cooperative network. In BT, the original file is partitioned into blocks of fixed length. A peer periodically exchanges data block availability information with its neighbors on the cooperative network. A node does not have to wait for the file to be completely downloaded before it can offer upload service to other nodes.

2.3 PPLive

PPLive is a P2P-based media streaming system that gains popularity all over the world. The data transmission of PPLive is similar to BitTorrent. The main differences lie in the data scheduling algorithms. In PPLive, each data block has an attribute which specifies the presentation time. If a data block is received after its presentation time, it is useless to the node. A common approach is to use a sliding window to decide the useful blocks. A peer will not request data blocks outside the window. Another difference between PPLive and BT is that PPLive does not need to store the whole file, but only maintains a small buffer to store the data blocks near the play point. When the slide window proceeds, old buffered data will be discarded.

2.4 Comparison and summary

Though the three systems are all based on P2P, there are significant differences. In BT or PPLive, nodes only have to download a fraction of the whole file before they can offer upload service to other nodes, while in Kazaa, a node can only provide service to other nodes when it has the entire file. In BT or PPLive, the overlay network is both the carrier of control information flow and real data flow, whereas in Kazaa, the overlay is only the carrier of control information. In Kazaa, real data downloading does not rely on the overlay network, and there is no concept of cooperative network in the downloading process. In the following, we will see that this native distinction determines the applicability of different P2P traffic optimization technologies.

3 P2P cache

With the increase of P2P traffic, caching the traffic becomes ISPs' first choice. However, compared with traditional Web applications, P2P applications differ in several aspects, e.g., traffic characteristics, size of the transmitted object, transmission mode, and object popularity. Moreover, P2P caching and Web caching also differ in their objectives, which leads to different cache algorithm design and performance evaluation. For example, Web caching typically use object hit ratio and average response time to evaluate the performance of a cache algorithm, whereas P2P cache uses byte hit ratio as the performance metric. Due to the above distinctions, it is not appropriate to directly apply web cache algorithms to P2P systems. As a consequence, it is necessary to develop highly efficient caching algorithms specialized for P2P applications. This section first surveys the research efforts of P2P traffic characterization and modeling, particularly its distinctions from web applications, then discusses different caching algorithms, and finally analyzes the limitations of the present cache technologies and proposes future research directions.

3.1 P2P traffic characterization

P2P traffic differs in several aspects with Web traffic. Traffic characteristics of the P2P systems have been widely studied in several papers [6, 9–12]. The main distinctions between P2P and web traffic is summarized in Table 1.

The differences in traffic characteristics affect the cache design. For example, object size and object popularity are main sources that affect the performance of a caching algorithm, while protocol openness affects the complexity and scalability of the cache design. In the following, we discuss some of the main factors that affect the cache performance.

• Object popularity. Object popularity is a major factor that affects the cache performance. It is confirmed in several work [6, 9, 12] that object popularity in the P2P system does not follow zipf's law, but has a very flat head in the probability distribution curve. It is proposed to use the mandelbrot-zipf distribution to model object popularity in the P2P system [12]. In this distribution, if objects are ordered by their popularity in decreasing order, the access frequency of the ith popular object obeys the following

Table 1 Distinctions of the traffic characteristics between P2P and Web

	P2P	Web
Object size	approximately three classes: small file below 10M, median file of several hundreds Megbytes, and large files over 1G [9, 10, 12].	in general are very small
Popularity	does not follow zipf's law; the popularity of the most popular object is far lower than what the zipf's law predicts [6, 9, 12].	follows zipf's law [13]
Variance of popularity	popularity can experience sudden change; objects can become popular in one night, and also lose popularity in short time [10, 12].	popularity does not vary much
Object mutability	immutable	more and more mutable objects
User access time	mostly once [9]	same object can be accessed by a user many times.
Number of sessions downloading an object	possibly dozens simultaneously	one or limited number
Session persistency	can last for hours	typically end in several seconds
Protocol openness	large number of private protocols	standard HTTP protocol
Port	varies across networks and possibly clients	single well known port(80)

distribution: $p(i) = \frac{K}{(i+q)^{\alpha}}$, where $K = 1/\Sigma_{i=1}^{N} \frac{1}{(i+q)^{\alpha}}$. So, q is an important parameter that implicitly determines the cache performance. The larger q is, the less gain of the cache.

• Object size. Object size is also an important factor that influences the cache algorithm. The fact that there are multiple kinds of traffic load in P2P systems has important effect on the cache algorithm. For example, audio files typically have higher request frequency than video files, hence, LFU-based cache replacement algorithms would be biased against large objects. On the contrary, using object size as the replacement criterion, e.g., small objects are evicted first, would be biased against smaller objects. In the worst case, it may have to evict dozens of popular MP3 files to make sufficient space to accommodate a not-so-popular large video object.

• Locality-awareness of P2P node. One of the prerequisites that P2P cache is effective is that P2P network does not adopt or only has very weak locality awareness. So, even the requested object has multiple copies within the same network, the requesting node cannot know this, and thus will request the object from the outside network. Present P2P systems often satisfy this prerequisite, which has been confirmed in [9, 10]. They discover that, if caches are placed at network boundaries in present P2P systems, then the theoretic byte hit ratio can be as high as 86% and 67%, which means the absence or weakness of locality-awareness. This observation provides practical feasibility of applying caching in present P2P systems.

3.2 P2P cache algorithms

Designing of cache algorithm is closely related to the cache objectives and traffic characteristics. In Web cache design, a primary objective is to reduce user perceived access latency, so Web cache replacement algorithms may contain considerations for the extra cost incurred upon cache missing. Whereas the main objective of P2P cache is to optimize the bandwidth usage, hence, improving byte hit ratio (not object hit ratio) has higher priority [11, 12].

Two things lie at the heart of a cache algorithm: 1) whether to cache the entire object or a part of it? 2) the cache replacement algorithm.

In Web cache, since objects are typically very small, it is commonplace to cache the entire object. However, since objects can be very large in P2P systems, caching the entire object may result in only very limited number of objects being cached, which can degrade the cache performance. A better way is to partition an object into blocks or ranges, and let the cache system to only cache part of the object.

Another key factor is the cache replacement algorithm. Traditional cache replacement algorithms include LRU (least recently used), LFU (least frequently used), MINS (minimal sized object first), MAXS

Table 2 Optimal cache replacement algorithms under the different combinations of caching mode and replacement granularity policy

	Full P2P caching	Partial P2P caching
File-based policy	LSB-F	LSB-F
Range-based policy	MINS-R	LRU-R

(maximal sized object first), and the GDS (greedy-dual-size) algorithm that is specialized for Web cache [14], however, all these algorithms are not specialized for P2P systems.

Ref. [11] first proposed the LSB (least-sent-byte) algorithm, a cache replacement algorithm specialized for the P2P. When a cache replacement is required, the object that serves the least bytes is evicted. This work also proposes two working modes of the P2P cache, called full P2P caching and partial P2P caching respectively. In both modes, the cache records and stores one or several ranges of an object. The two modes differ in the action taken when the cached ranges do not cover the range requested by the user. If in this case, the cache does not serve the request, it is called full P2P caching mode. Otherwise, if the cache can negotiate with the user to serve a subrange of the requested range, it is called partial P2P caching mode. The cache replacement can occur at two granularities: file based and range based. Table 2 summarizes the optimal cache replacement algorithms under different combinations of caching mode and replacement granularity proposed in [11].

Statistical analysis shows that in P2P systems, popular objects gain popularity in a relatively short timescale reaching their peak in about 5–10 weeks, and then the popularity drops dramatically [12]. A sixfold decrease in popularity can be observed within 5–10 weeks. This means that if we cache objects according to their access frequencies, then these objects will stay in the cache long after they lose their popularity. On the other hand, the popularity of an object is not very short-lived, indicating that incremental partial caching of objects is beneficial because it allows the cache to build the object popularity profiles with sufficient time. In this way, the cache can identify which objects are truly popular, and incrementally increase its cached ratio until the entire object has been cached. Based on this idea, a proportional partial caching algorithm is proposed [12]. Unlike LSB, this algorithm partitions each object into fixed length blocks, and then increases the number of blocks of popular objects in the cache based on access history. Simulation results show that proportional partial caching is superior than other caching algorithms, including LSB.

3.3 Problems with P2P cache

Though from ISPs' point of view, P2P cache is the first choice to optimize P2P traffic, but it also has the following problems:

• Effectiveness. It has already been mentioned before that P2P cache is only effective when P2P nodes have very weak locality-awareness capability. However, just as section 4 will describe, with the increasing support of locality knowledge, the effectiveness of P2P cache will decrease.

• Scalability. The presence of various protocols, some of which are even proprietary, restricts the scalability of P2P cache system. To cache the traffic of a new P2P application, the cache system has to understand its protocol details, which increases the implementation complexity, and reduces the system scalability.

On the other hand, either P2P cache or Web cache does not allow applications to access and manage the cache, making it impossible to design customized cache strategy for a particular application. To address this issue, IETF established the DECADE working group [15, 16] in 2010, which tries to provision public in-network cache, separate the control plane from the data plane in P2P applications, and allow applications to autonomously use and manage their caches through open protocols. In this way, it forms a manageable and controllable content distribution platform, which completely solves the scalability issue of P2P cache.

• Copyright issue. Though the copyright issue also exists in the Web cache, this problem is more salient in P2P systems because these systems are often used to distribute media files that are more sensitive to copyright issues. P2P cache systems should be careful to avoid violating any copyright protection acts

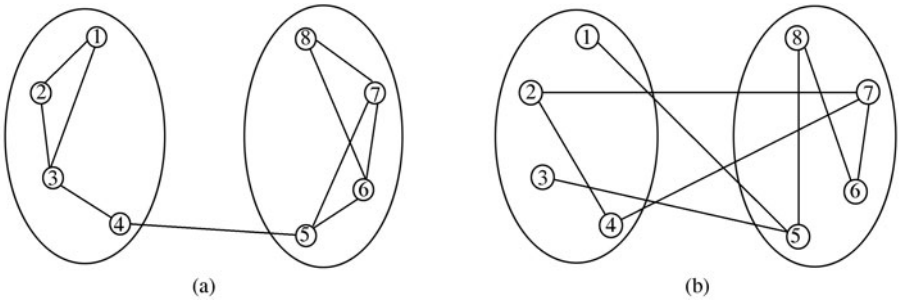

Figure 2 Two forms of P2P overlay construction. (a) Locality knowledge assisted P2P overlay construction; (b) random P2P overlay construction.

in providing the cache service. It should be noticed that there is still no explicit laws in China. Across the world, the only explicit one is Digital Millennium Copyright Act (DMCA) in USA. Hence, it leaves a large gray zone in the digital copyright protection of cache systems [17].

• Cost issue. P2P cache violates the operating mode of the network. ISPs deploy P2P cache in a passive way, which cannot completely solve the conflicts between ISPs and P2P content providers. Under the DECADE architecture, ISPs treat network storage as an open content distribution platform capable of creating values for them, which provides a precondition for solving the conflict.

4 Locality-awareness

P2P network is an overlay network built over the underlying network. The extent to which it matches the underlying network topology determines the usage behavior of the underlying network resource. In early P2P networks, peers randomly select neighbors, resulting in severe mismatch between the overlay and the underlay network topologies. This not only reduces application performance, but also places heavy traffic burden on the underlay network. Topology mismatch brings two problems. First, from the content aspect, nodes located in proximity have higher interest similarity [18, 19], but topology mismatch makes it impossible to take advantage of this. For example, in Figure 2(b), nodes 2, 3, 4 have higher probability than nodes 5, 6, 7, 8 to contain contents of interest to node 1, but topology mismatch makes the search overhead very high for node 1 to discover nodes 2, 3, and 4. In the extreme case, node 1 will not be able to discover 2, 3, 4. Second, topology mismatch increases transmission cost. In the previous example, even all nodes have the content that is of interest to node 1, in the transmission mode similar to BT and PPLive, node 1 can only download data blocks from its neighbors (node 5). However, node 5 can be at the far end of another continent. In [20], it is pointed out that only 2%–5% overlay connections in Gnutella network are within same autonomous systems, but at the same time, over 40% Gnutella nodes are in the largest 10 ASes, which means the Guntella network does not have or only have very weak locality awareness capability. Figure 2(a) illustrates the locality-aware overlay network topology, which is highly consistent with the underlying network. In this case, nodes that are proximate on the underlying network are clustered on the overlay network, making optimization of network resources possible.

Locality-aware overlay network construction involves two aspects: how to obtain and provision locality information, and how to use these information to assist the overlay topology construction. P2P systems only require the relative distance between nodes, not the absolute locations of nodes. Retrieving and provisioning relative locality information has been extensively studied in the past few years, mainly including real-time measurement based techniques and non real-time techniques.

4.1 Reverse-engineering the locality knowledge

4.1.1 *Real-time measurement based locality knowledge retrieval*

In P2P systems, peers can obtain delays or network distances between nodes based on Ping or Traceroute. For instance, TopBT proposed to translate the traceroute result into router and AS hops, and then use them to optimize the neighbor selection [21]. Ref. [22] proposed LTM for a node to optimize neighbor

selection by locally broadcasting application-specific TTL-2 detector packets. Ref. [23] proposed a dynamic landmark based approach which does not need dedicated landmark servers, but relies on the nodes in the overlay network as landmarks, and classifies nodes accordingly. A common disadvantage of these approaches is that locality information inferred by one application cannot be shared by another. Hence, provisioning of an infrastructure that can provide locality information for different applications gradually becomes a common recognition. These infrastructures can be broadly classified into non landmark based and landmark based ones.

(1) Non-landmark based locality-aware techniques. Non-landmark based techniques can rely on newly built dedicated infrastructures, e.g., Tracers in the IDMaps [24] approach and landmark servers in the Binning approach [25]. Also, these techniques can rely on infrastructures already available on the Internet, e.g., DNS and CDN.

In IDMaps [24], there are two concepts: Tracer and AP(Address Prefix). Tracers are dedicated probing servers distributed across the whole Internet. A Tracer periodically measures the delays to other Tracers and to the APs near it. The delay between two arbitrary IP addresses is considered to be the sum of the delays between the IP addresses' associated APs and their respective nearest Tracers, plus the delay between the two Tracers.

Binning [25] is a landmark based approach. Denote landmarks as $L_1, L_2, \ldots, L_k$. When P joins the system, it first measures the delays $l_1, l_2, \ldots, l_k$ to these k landmarks. Listing the delays in increasing order results in a permutation of the landmarks $L'_1, L'_2, \ldots, L'_k$, which can be considered as a classifier of the nodes. If P_1 and P_2 result in the same permutation, then P_1 and P_2 can be attributed to the same class. Based on this, one can also get fine-grained classification by partitioning the absolute delay values into different grades.

King [26] proposes to predict the delays using the present DNS system. Its advantage is not having to deploy additional measurement platform, and not having to make any modifications to the present protocol stack. Given two arbitrary nodes A and B, King first finds their authoritative name servers NS(A) and NS(B) near the two nodes respectively, then obtains the delay between NS(A) and NS(B) by DNS's recursive lookup, and uses this delay as an estimate of the delay between A and B. However, not all authoritative name servers are close to the targeted nodes, hence, this approach can incur non-negligible errors. Turbo King [27] improves on this issue, reduces the error range, and overcomes the large-scale DNS cache pollution problem incurred by King.

In [28, 29], it is proposed to reuse the CDN information to estimate the proximity between two nodes. CDN uses dynamic DNS redirection technique to provide low latency mirror servers for clients, which makes it possible to use the CDN infrastructure to achieve optimized neighbor selection. This approach assumes that two nodes with similar redirection behavior have high probability to be close to each other, e.g., located in the same ISP.

Other techniques are summarized in [30, 31], which will not be detailed here.

(2) Network coordinate based locality-aware techniques. The basic idea of network coordinate approaches is to embed all nodes in the network into a virtual space according to the measured network distances (delay, bandwidth and packet drop rate), and assign a virtual coordinate to each node. Then, network distance between any two nodes can be calculated by the nodes' virtual coordinates, e.g., according to the Euclidean distance between two coordinates. One advantage of the network coordinate system is the elimination of the additional probing overhead for nodes to determine the delays in between, which significantly increases the system's scalability.

GNP (global network positioning) [32] is the earliest network coordinate system based on centralized landmarks. In GNP, a group of dedicated landmarks measure latencies to each other, resulting in a latency matrix. Based on this matrix, GNP first models the network coordinate computation as a multi-dimensional global optimization problem, and assigns each landmark an appropriate coordinate through the solving of the optimization problem. When the coordinates of all landmarks have been identified, an ordinary node H only needs to measure the latencies to these N landmarks to identify its coordinate.

However, GNP has the following problems: 1) high computation complexity; 2) slow convergence, and the possibility of converging to local optimum; 3) still high measurement cost. Several approaches

are proposed to address these problems. ICS [33] and virtual landmarks [34] use relative coordinate to compute network coordinate. In ICS and virtual landmark, ordinary nodes first acquire the latency matrix of the landmarks and represent it as vectors. The dimensionality of the vectors equals to the number of landmarks. Then, the principle component analysis (PCA) technique is used to project the latency space into a new uncorrelated and orthogonal Cartesian coordinate system of much smaller dimensions. Through coordinate transformation, the coordinates of nodes can be calculated by linear transformation, whose efficiency is much higher than the non linear iterative method adopted by GNP. Regarding the second problem, ref. [35] proposes to use network geolocation to optimize the initial network coordinate values, and thus speed up the convergence of network coordinate calculation, as well as improve the accuracy. In order to completely avoid the probing overhead between ordinary nodes and landmarks, ref. [36] proposes an approach that passively monitors the source address and TTL of each passing packet on each landmark, infers the hops between the monitoring landmark to the observed source address, and uses this information as input to calculate the coordinates. In this way, it significantly reduces the probing overhead.

In all these approaches, landmarks are static and the number is fixed, which raises the problem of poor scalability and single point of failure. If some landmark fails or reboots, then all coordinates should be recalculated. Distributed landmark approaches try to avoid the problems of centralized landmarks. Typically, any node whose coordinate has been computed can become landmark. Lighthouse [37] is a representative of the distributed landmark approach. By introducing the concept of local coordinate, Lighthouse allows each node to arbitrarily probe $K+1$ (K is the dimension of the coordinate) landmarks, compute its local base by the Gram-Schmidt process, and finally calculate the global coordinate of the new node by the transition matrix between the local base and global base. Since this method only needs to measure the delays to part of landmarks, it avoids the problem of single-point-of-failure and improves the system scalability. Besides Lighthouse, there are other distributed landmarks based approaches, such as PIC [38] and IDES [39]. We will not detail these approaches here.

In addition, there are other approaches that model the coordinate calculation by physical processes, whose objectives are to improve the convergence speed and accuracy of coordinates. Vivaldi [40] is a representative one. It translates the square-error between two nodes as spring forces and simulates the movements of nodes under the spring forces. Physical process based approaches also include [41] and [42]. The common advantage of these approaches is that by including the network coordinate adjusting process, these approaches actively adapt to network dynamics, and improve the adaptability, robustness and accuracy of network coordinates. The concerns of these approaches are the convergence, stability, and the additional computation and communication overhead.

4.1.2 *Non-real-time locality knowledge retrieval*

(1) IP prefix matching. In most cases, IP address to some extent implies the locality information of a node, which can be used to determine the relative closeness of the node with other nodes[1)]. Present Internet is based on prefix-based routing. Each IP address corresponds to one or more routable IP address prefixes. However, a node cannot know the exact IP prefix length without the support of the routing information. Traditional globally routable unicast IP addresses can be classified into three classes: A, B, and C. A node can simply determine the prefix length of an IP address to be 8, 16 or 24, according to the class of the IP address. However, due to the prevalence of subneting and CIDR, the above approach will often incur errors.

A more reasonable way is to utilize the routing table information [43]. There are two kinds of routing tables: inter-domain BGP routing tables, and intra-domain routing tables. Typically, intra-domain routing tables have more specific routing entries for their responsible routing domains. However, ISPs are often unwilling to publicize their intra-domain routing tables because this involves their privacy. In comparison, there are more channels to obtain inter-domain routing tables, of which the Routeviews project [44] and RIPE's RIS project [45] are the two most prominent ones. They collect the BGP views

1) Here we assume the observed IP address is the real IP address of the node. We do not consider the occasion when proxies are used, in which the observed IP address is the address of the proxy rather than the real node.

of many autonomous systems by establishing peering sessions with different BGP routers. By merging multiple BGP views, it is possible to obtain more fine-grained IP prefix set, and identify whether two IP addresses are within the same network by the longest prefix matching.

One disadvantage of the above approach is that it relies on the IP prefix information collected in prior. When network address assignment changes, it is unable to make in-time adjustment. Ref. [46] proposes a self organizing distributed prefix matching technique. Each node performs a hash function on its own IP netmask, or k bits of the IP address, and then stores its IP address on the node that is responsible for this hashed key on a DHT system. In this way, a new node can easily find those nodes who have the same IP prefix. The node can use a bottom-up approach, which looks up from long prefixes, and then gradually shortens the prefix length until sufficient neighbors are found. The advantage of this approach is that it is fully self organizing and self adaptive, not relying on a preprocessed set of IP prefixes. The disadvantage is that it needs to maintain a DHT.

(2) IP-TO-AS mapping. A coarse-grained approach to test whether two nodes are close to each other is to map IP addresses to AS numbers, so that each node can determine whether other nodes are located within its own autonomous system. This information, although coarse-grained, can be used to optimize the inter-domain traffic. Mapping the IP addresses to AS numbers also relies on the BGP routing tables. Apart from the IP prefixes, BGP routing tables also store the AS path that an IP prefix advertisement traverses before it arrives at this BGP router. The first AS that an IP prefix traverses is called its origin AS. Multiple BGP views can be synthesized to build the mapping table between IP prefixes and their origin ASes. Given an IP address, the longest prefix matching can be used to find its corresponding AS number.

4.1.3 *Comparison*

Table 3 makes a comparison between different locality information retrieval technologies. Here, the usage is classified into distance estimate and closeness estimate. Distance estimate means the technology is able to estimate the network distance between two nodes, whereas closeness estimate means the technology is only able to estimate the relative closeness of two nodes. Any method that provides distance estimate capability can be used to estimate the closeness between nodes.

4.2 Locality service provision

Reverse-engineering based approaches cannot guarantee the accuracy of locality information. In fact, ISPs are at the right position of providing network topology and state information. Recently, some work suggest ISPs to provide locality information service. Since this involves the cooperation between ISPs and P2P content providers, providing loosely coupled, scalable, secure, general and flexible standard service interface is the basic requirement of these approaches. Oracle and P4P are two such representative work.

Oracle [4] is the P2P traffic optimization scheme proposed by Deutsche Telekom Laboratories, which provides network information service for P2P applications by ISPs. Oracle collects the topology information of ISP networks, such as AS number, AS connectivity and more fine-grained topology information, e.g., POP or city level topology information, and then provides peer matching suggestions to P2P applications. When a P2P client needs to choose neighbors or download data, it submits a group of candidate nodes to the Oracle. Oracle orders the candidate nodes according to the network topology information, and assists the client to choose optimized nodes. Since Oracle only provides ordering service, it will not leak much information. In [47], the authors make improvement on the Oracle to provide network-layer service, intending to choose the optimal source and destination addresses for two nodes to establish connection under the multi-homing environment.

P4P [5] is a P2P traffic optimization architecture proposed by Yale University. It proposes to optimize P2P traffic and performance by cooperation and communication between ISPs and P2P applications. The idea of P4P is similar to Oracle, but the technology is more comprehensive and complicated. P4P collects more comprehensive network information, such as link cost, ISP policy. It proposes to deploy the iTracker devices on ISP networks to collect ISPs' information, and provide interfaces for interaction with

Table 3 A brief comparison between different locality information retrieving technologies

	Real-time	Rely on third party	Applicability	Accuracy	Usage
IDMaps	yes	yes	IP address	depends on the Tracer number and deployment	distance estimate
Binning	yes	yes	node	depends on the number of landmarks	closeness estimate
mOverlay	yes	no	node	depends on the reasonability of group assignment criteria	closeness estimate
King	yes	yes	node	depends on the distance between the node and its authoritative name server	distance estimate
CDN	yes	yes	node	depends on the accuracy of CDN's redirection	closeness estimate
network coordinate	yes	yes	node	depends on the algorithm of network coordinate computation	distance estimate
IP Prefix Matching	no	no	IP address	depends on the granularity of prefixes and native geolocation property of the IP address	closeness estimate
IP-TO-AS	no	no	IP address	low	closeness estimate

P2P applications. These interfaces include: static network policy, P4P distances that reflect network policy and state, network capability (for example, cache). They provide support for other applications to obtain ISPs' network information, but meanwhile protect their privacy, hence it allows the two sides to jointly optimize their respective network performance.

Since locality information provision involves cooperation between different ISPs and applications, providing standard service interface is the prerequisite for these services to be widely adopted. In this respect, the China Communications Standards Association (CCSA) began to standardize the "Technical framework for carrier network-aware P2P traffic optimization"[48] since 2008, which is jointly proposed by the Institute of Computing Technology, Chinese Academy of Sciences, China Academy of Telecommunication Research of MIIT, and major ISPs in China. This framework defines a network matching service that assists the P2P applications for decision making. Figure 3 is the basic architecture of the network matching service, which includes four functional entities: carrier information provider, network matching server, network matching service discovery server and network matching service client.

Network matching service client can be ordinary P2P host or P2P index server, which is authorized to access some network matching servers. Network matching service clients can utilize the network matching service to optimize the selection of resource provisioning nodes, P2P overlay topology, and relay nodes and cache nodes. Network matching server is the entity that is implemented and deployed by ISPs to provide network matching service. Carrier network information provider provides description of the current carrier network status, including network topology, network policy, network capability and network dynamics.

Each network matching server maintains one or more views of the Internet. Each view is constructed from the collected carrier network information, which reflects the network positions of hosts and distances between network positions. A view can be represented by a table. Each entry $d(LID_i, LID_j)$ of the table represents the network distance between network position LID_i of the resource owner to the network position LID_j of the requesting node. This network distance is a composite result of synthesizing the carrier network status and network policies. When a client requests service, the network matching server chooses a view according to the policy constraint, uses this view to map P2P nodes to network position identifiers, and chooses resource owners that have short distances to the requesting node.

Similarly, IETF's ALTO (Application Layer Traffic Optimization) working group is also making the ALTO standards [49, 50], which lets ISPs or third parties to provide ALTO services, e.g., network status information and network cost information. Among the services provided by ALTO, ALTO map service is the most important one. The Map Service provides batch information to ALTO Clients in the form of

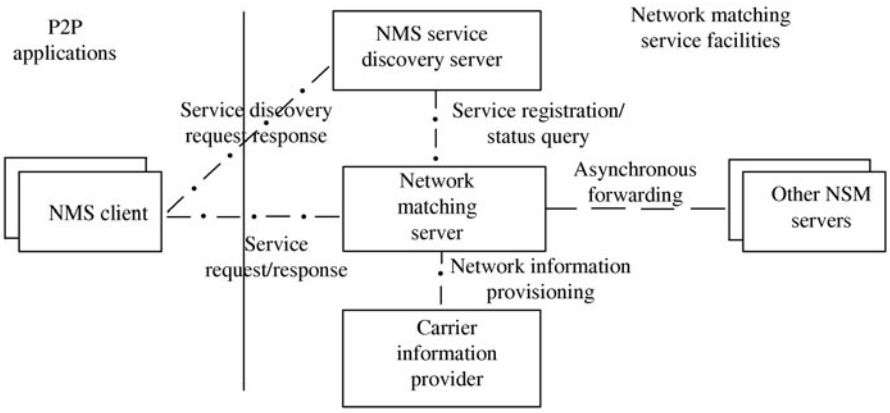

Figure 3 Basic architecture of the network matching service.

a Network Map and Cost Map. The Network Map provides the full set of Network Location groupings defined by the ALTO Server and the endpoints contained with each grouping. Each group is identified by a network location independent PID defined by the provider to hide network details. The Cost Map provides costs between the defined groupings. Cost can be delay, bandwidth or other link attributes measurable by ISPs. Besides, ALTO also provides other services such as map filtering service, endpoint property service and endpoint cost service.

4.3 Locality-aware topology construction

The main goal of obtaining locality information is to construct topology-aware overlay topology. In a tracker-based system such as BT and PPLive, the tracker maintains the peers that participate in the current session. When a node tries to join the system or wants to replace its neighbors, it will send request to the tracker for candidate neighbor set [29, 46, 51]. Tracker can utilize the locality information to assign a group of candidates close to the requesting node. P2P node can also self optimize its overlay topology at the runtime based on the probed information, e.g., optimizing the connections based on two hop latency probing [22]. ISPs can assist the applications to build locality-aware topology through the deployment of Proxy-tracker. Proxy-tracker is deployed at network boundaries, which maintains the information of nodes participating in a P2P session within the network. Proxy-tracker intercepts the candidate neighbor response message sent from the P2P application tracker to the requesting peer, modifies the response message by replacing some candidates that are far from the requesting node with some nodes that are within the same network as the requesting node, and consequently assists the requesting node to transparently build locality-aware topology [18, 51]. ISPs can also provide redirection servers to realize locality-aware overlay network construction. A redirection server maintains the information of the participating nodes and the objects they own within the domain. It intercepts the user's connection or data request, and performs redirection. For example, for those P2P networks with supernodes, the connection requests from supernodes can be redirected to supernodes within the same domain [9, 52].

4.4 Summary

Building locality-aware overlay topology that matches the underlying network topology is the main research direction of P2P traffic optimization. Although reverse-engineering based approaches are extensively used today, they share the following two problems:

• No accuracy guarantees. The accuracy and granularity of the locality information are limited, whether the approach is IP prefix matching based or active probing based.

• High overhead. Active probing approaches introduce additional probing overhead, whereas IP prefix matching based approaches incur preprocessing overhead. In order to keep the IP prefixes up-to-date, the preprocessing should be done periodically. Even the self organized IP prefix matching method introduces additional DHT.

Locality knowledge provision that bases on cooperation between ISPs and P2P content providers is the right direction, however, it requires a standardized interface between two sides. Though there are some attempts both in the IETF and CCSA, there is long time to go for both sides to accept the standards.

In addition, the effectiveness of most researches of locality-aware technologies are evaluated by simulation, lacking experimental evaluation in real network environment. In real network environment, the prevalence of NAT, firewalls and VPN may have significant impact on the effectiveness and usage of locality information.

5 P2P data scheduling algorithms

Data scheduling are typically studied from the perspective of improving application-layer performance, without consideration of the impact on underlying ISPs. Here when we say data scheduling algorithms, we are talking about the third generation P2P applications, i.e., the original file is split into multiple data blocks, and the participating nodes form a cooperative network to distribute the file. In this kind of systems, each node only establishes connections with a few number of nodes, which are called its neighbors. The data scheduling decision is made on the local information of a node. Often, it is restricted that a node can only make the decision based on its own data block information and its neighbors' data block information. Typical data scheduling algorithms include:

• Sequential scheduling: a node chooses the block with smallest sequence number among all the blocks that it does not have for downloading. Apparently, this scheduling algorithm has poor performance, because peers often are unable to cooperate. This scheduling algorithm is sometimes applied to video streaming that has real time or near real time requirement.

• Random scheduling: a node randomly chooses a block that it does not have for downloading. This algorithm greatly improves the data usability between neighboring peers, and hence enhances the utilization of the bandwidth between neighbors.

• Local rarest first scheduling: LRF is an algorithm introduced by BitTorrent, which aims to improve the balance of data blocks among the network. In LRF, a node chooses a block with the lowest local frequency among all the blocks that it does not have. The local rarest first policy greatly improves the throughput of the system.

• Global rarest first: this is a theoretically optimal scheduling algorithm which assumes that each node knows the whole network status. Each time, a node chooses among the candidate blocks a block that has the lowest frequency globally.

• Network coding: network coding [53] is a new data communication paradigm that not only allows intermediate nodes to store, replicate and forward data, but also to perform arbitrary coding operations. In network coding based scheduling, blocks being transmitted are no longer the original blocks, but coded blocks out of several original blocks. Linear network coding [54, 55] and random network coding [56] have paved the way for the practicality of network coding. In P2P systems that apply random linear network coding [57–59], when a node receives a data block request, it randomly selects m available blocks $b_1, b_2, \ldots, b_m$, generates m random local encoding coefficients $c_1, c_2, \ldots, c_m$, and produces a coded block $b = c_1 \cdot b_1 + c_2 \cdot b_2 + \cdots + c_m \cdot b_m$. Also, it calculates the global coding coefficient g as follows:

$$g_b = \begin{pmatrix} c_1 & c_2 & \cdots & c_m \end{pmatrix} \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ g_{m1} & g_{m2} & \cdots & g_{mn}, \end{pmatrix},$$

where $(g_{i1}, g_{i2}, \ldots, g_{in})$ is the embedded global coding coefficient of block b_i. The node then encapsulates b and its associated global coding coefficient g, and sends the encapsulated packet to the requesting node. When a node receives n linearly independent blocks, it can recover the original data blocks.

• Erasure code: erasure code can be considered as a special case of network coding. In erasure code, only the source node can encode the blocks, while all other nodes are not allowed to do so.

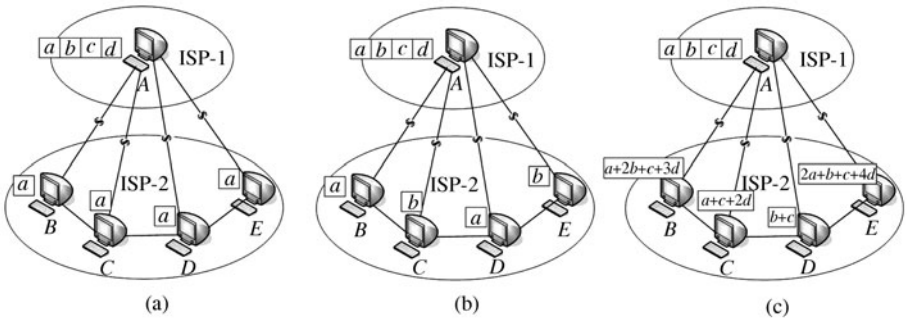

Figure 4 Impacts of data scheduling on P2P traffic optimization. (a) Random scheduling; (b) local rarest first scheduling; (c)network coding based scheduling.

This section focuses on the impacts of three data scheduling algorithms on P2P traffic localization: random scheduling, local rarest first scheduling, and network coding based scheduling. It should be noted that data scheduling algorithms alone cannot optimize P2P traffic, only when locality-aware knowledge is integrated can they be effective for P2P traffic localization. Figure 4 shows how these three data scheduling algorithms influence P2P traffic. Assume initially only A contains the data blocks a, b, c, d, which is to be distributed to nodes B, C, D, E. Figure 4(a) uses the random scheduling. In this case, it is possible that B, C, D, E all request the same block from node A, say a. This makes connections between B, C, D, E cannot be utilized in the subsequent rounds, and data still has to be transmitted through the inter-domain links. Though $B - E$ may request distinct blocks from A, the probability is only $4!/4^4 = 9.375\%$. Figure 4(b) uses the local rarest first policy. However, even this sophisticated data scheduling can result in unbalanced data distribution. For example, B first requests block a from A, after which C makes a data request to A. According to the local rarest first policy, C can request block b from A. After that, D can request block a from A, and E can request block b from A. Then, B, C, D, E contain two distinct blocks, a and b. After another round of data transmission, the intra-domain links between $B - E$ cannot be utilized, so new data has to be fetched from A. Figure 4(c) uses network coding. In this case, A sends randomly coded blocks to B, C, D, E. When the finite field is sufficiently large, e.g., $F(2^8)$ or $F(2^{16})$, there is high probability that the four coded blocks sent to B, C, D, E are linearly independent [56]. Thus, B, C, D, E can use intra-domain links to obtain the necessary blocks to recover the original blocks, and make optimized utilization of network resources.

Ref. [51] compares random scheduling and local rarest first scheduling. Results show that supported by the same locality information, local rarest first based data scheduling can achieve 80%–90% reduction of inter-domain P2P traffic. However, even under the strongest locality-aware setting, the inter-domain traffic redundancy of the local rarest first scheduling is still above 3, which means on average a block has to travel into the same domain for more than 3 times. Ref. [60] proposes to use random network coding to achieve globally optimized network resource utilization for P2P streaming. Our work compares the impact of local rarest first scheduling and network coding based scheduling on P2P traffic optimization, supported by the same locality knowledge [61]. Results show that network coding based scheduling can halve the inter-domain traffic redundancy of the local rarest first scheduling.

6 Discussion and conclusion

6.1 Comparison

Although P2P cache, locality-awareness and data scheduling are all beneficial for P2P traffic optimization, they differ in several aspects: applicability, independency and implementing party. Table 4 summarizes these distinctions.

6.2 Conclusion and future work

P2P content distribution systems have gained increasing popularity among content providers and end users. However, they have significant impacts on the underlying Internet infrastructure and business

Table 4 Comparison between different P2P traffic optimization technologies

	P2P cache	Locality awareness	Data scheduling
Applicability	applies to all applications, but the effectiveness decreases as the locality-awareness increases	all P2P applications	applies only to applications that have a cooperative downloading network such as BT and PPLive
Independency	can be used independently	can be used independently	used in conjunction with locality-awareness
Implementing party	ISPs	ISPs or P2P content providers	P2P content providers

models, bringing serious challenges for ISPs. Hence, optimizing P2P traffic is a prerequisite for the healthy, benign and sustainable development of P2P applications. This paper surveys the P2P traffic optimization technologies from three perspectives: P2P cache, locality-awareness and data scheduling. The distinct differences of traffic characteristics and caching objectives between P2P applications and Web applications require the design of caching algorithms specialized for P2P cache to improve the byte hit ratio of the caching system. The developing trend of locality-awareness is for ISPs and P2P content providers to cooperate for joint optimization of the underlying network resource utilization. This requires a loosely coupled and scalable standard interface between the two sides. Optimized data scheduling algorithm coupled with locality information can further improve the utilization of network resources, pushing the capability of P2P traffic optimization to its limit.

However, though P2P traffic optimization has gained extensive research, there is still a wide gap between theoretical results and its wide adoption. On one hand, the effectiveness of most theoretical results is based on simulation studies, lacking large-scale experimental evaluation on real Internet. Under real Internet environment, the prevalence of NAT and firewalls may significantly reduce the effectiveness of locality awareness technologies. Hence, there is an urgent need to evaluate the effectiveness of locality-aware technologies on P2P traffic optimization through large-scale experimental test on real Internet. On the other hand, if technology is feasible, how to build a suitable business model for ISPs, P2P content providers and end users is critical to push forward the technology [62].

Additionally, non-optimized network resource usage to some degree originates from the mismatch between the end-to-end communication design of the Internet and the present content-centric user requirement. Essentially, either P2P or CDN targets at improving the efficiency of content distribution at the application layer, however, under the present architecture, content retrieval includes name resolution and addressing. In order to make better use of network resources, these systems all sense network status from the application layer, and achieve intelligent resource location, which is called network-aware application. However, this approach has the following disadvantages: 1) intelligent resource location based on middle boxes such as DNS or ALTO introduces additional resolution delay; 2) there is a problem of function duplication between the location function of the middle box and the routing function of the network layer; 3) it is possible that the delayed update of the mapping between names and locations can cause resolution failure. Hence, the present end-to-end transmission paradigm is challenged by some researchers who suggest content-centric networking [63–65]. This new network architecture eliminates the need to translate names into addresses by the adoption of route-by-name paradigm. In essence, content-centric networking makes three modifications to the present Internet: 1) transforms the major communication mode from unicast to anycast; 2) transforms the data transmission mode from sender driven to receiver driven; and 3) uses in-network caching to absorb network traffic. In principle, content-centric networking can gracefully fit the users' access demand for contents, i.e., content will be asynchronously accessed for multiple times. It allows the user to find the best content provisioning location through routing, and achieve optimal resource utilization. However, how to achieve incremental deployment of content-centric networking on the present Internet and realize smooth transition are of great challenges.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 61100178), National Basic Research Program of China (Grant No. 2012CB315802), Startup Foundation of Nanjing Normal University (Grant No. 2011119XGQ0248), Beijing Natural Science Foundation (Grant No. 4112057), and National Science and Technology Major Project (Grant No. 2011ZX03002-002-03).

References

1 Karagiannis T, Broido A, Brownlee N, et al. Is P2P dying or just hiding? In: Proc IEEE Global Telecommunications Conf (GLOBECOM'04), Dallas, TX, 2004. 1532–1538
2 Ledlie J, Gardner P, Seltzer M. Network coordinates in the wide. In: Proceedings of NSDI 2007, Cambridge, MA, USA, 2007
3 Gurbani V K, Hilt V, Rimac I, et al. A survey of research on the application-layer traffic optimization problem and the need for layer corporation. IEEE Commun Mag, 2009, 47: 107–112
4 Aggarwal V, Feldmann A, Scheideler C. Can ISPs and P2P users cooperate for improved performance? ACM SIGCOMM Comput Commun Rev, 2007, 37: 31–40
5 Xie H Y, Yang Y R, Krishnamurthy A, et al. P4P: Provider portal for applications. In: Proc ACM SIGCOMM 2008, Seattle, WA, USA, 2008
6 Sen S, Wang J. Analyzing peer-to-peer traffic across large networks. IEEE/ACM Trans Netw, 2004, 12: 219–232
7 Stoica I, Morris R, Karger D, et al. Chord: A scalable peer-to-peer lookup service for Internet Applications. In: Proc ACM SIGCOMM'01, San Diego, CA, USA, 2001
8 Rowstron A, Druschel P. Pastry: scalable, decentralized object location and routing for large-scale peer-to-peer systems. In: IFIP/ACM International Conference on Distributed Systems Platforms, Heidelberg, Germany, 2001
9 Gummadi K, Saroiu S, Gribble S, et al. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In: Proc 19th ACM Symp Operating Systems Principles (SOSP'03), Bolton Landing, NY, 2003. 314–329
10 Leibowitz N, Bergman A, Ben-shaul R, et al. Are file swapping networks cacheable? Characterizing P2P traffic. In: Proc 7th Intl Workshop on Web Content Caching and Distribution (WCW'02), Boulder, CO, 2002
11 Wierzbicki A, Leibowitz N, Ripeanu M, et al. Cache replacement policies revisited: the case of P2P traffic. In: Proc 4th Intl Workshop on Global and Peer-to-Peer Computing (GP2P'04), Chicago, IL, 2004. 182–189
12 Hefeeda M, Saleh O. Traffic modeling and proportional partial caching for peer-to-peer systems. IEEE/ACM Trans Netw, 2008, 16: 1447–1460
13 Breslau L, Cao P, Fan L, et al. Web caching and Zipf-like distributions: evidence and implications. In: Proc IEEE INFOCOM, New York, NY, USA, 1999
14 Cherkasova L. Improving WWW proxies performance with greedy-dualSize frequency caching policy. HP Laboratories Report No. HPL-98-69R1, April, 1998
15 DECADE working group, https://datatracker.ietf.org/wg/decade, 2010
16 Song H, Zong N, Yang Y, et al. Decoupled application data ENROUTE (DECADE) problem statement. draft-ietf-decade-problem-statement-00.txt, Aug, 2010
17 Zhou R. P2P traffic optimization based P2P cache (in Chinese). Telecommun Netw Tech, 2009, 1: 11–15
18 Karagiannis T, Rodriguez P, Papagiannaki K. Should internet service providers fear peer-assisted content distribution? In: Proc 5th ACM SIGCOMM Conf Internet Measurement (IMC'05), Berkeley, CA, 2005. 63–76
19 Rasti A, Stutzbach D, Rejaie R. On the long-term evolution of the two-tier Guntella overlay. In: Global Internet, Barcelona, Spain, 2006
20 Ripeanu M, Foster L, Iamnitchi A. Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. IEEE Internet Comput J (Special Issue on Peer-to-Peer Networking), 2002, 6: 50–57
21 Ren S, Tan E, Luo T, et al. TopBT: A topology-aware and infrastructure-independent BitTorrent client. In: Proc IEEE INFOCOM, San Diego CA, USA, 2010
22 Liu Y H, Xiao L, Liu X M, et al. Location awareness in unstructured peer-to-peer systems. IEEE Trans Parall Distr Syst, 2005, 6: 163–174
23 Zhang X Y, Zhang Q, Zhang Z S, et al. A construction of locality-aware overlay networks: mOverlay and its performance. IEEE J Select Areas Commun, 2004, 22: 18–28
24 Francis P, Jamin S, Jin C, et al. IDMaps: A global internet host distance estimation service. IEEE/ACM Trans Netw, 2001, 9: 525–540
25 Ratnasamy S, Handley M, Karp R, et al. Topologically aware overlay construction and server selection. In: Proc IEEE INFOCOM, New York, NY, USA, 2002

26 Gummadi K P, Saroiu S, Gribble S D. King: Estimating latency between arbitrary internet end hosts. In: Proc Internet Measurement Conference (IMC), Marseille, France, 2002

27 Leonard D, Loguinov D. Turbo king: Framework for large-scale internet delay measurements. In: Proc IEEE INFOCOM'08, Phoenix, AZ, USA, 2008

28 Su A J, Choffnes D, Bustamante F E, et al. Relative network positioning via CDN redirections. In: Proc IEEE ICDCS'08, Beijing, China, 2008

29 Choffnes D R, Bustamante F E. Taming the torrent: A practical approach to reducing cross-ISP traffic in peer-to-peer systems. In: Proc ACM SIGCOMM, Seattle, WA, USA, 2008

30 Wang Y J, Li X Y. Network distance predication technology research (in Chinese with English Abstract). J Software, 2009, 20: 1574–1590

31 Xing C Y, Chen M. Techniques for network distance predication (in Chinese with English abstract). J Software, 2009, 20: 2470–2482

32 Ng E, Zhang H. A network positioning system for the Internet. In: USENIX Conference, Boston, MA, 2004

33 Lim H, Hou J C, Choi C H. Constructing internet coordinate system based on delay measurement. IEEE/ACM Trans Netw, 2005, 13: 513–525

34 Tang L, Crovella M. Virtual landmarks for the Internet. In: Internet Measurement Conference, Miami, Florida, USA, 2003

35 Agarwal S, Lorch J R. Matchmaking for online games and other latency-sensitive P2P systems. In: Proc ACM SIGCOMM, Barcelona, Spain, 2009

36 Eriksson B, Barford P, Nowak R. Estimating hop distance between arbitrary host pairs. In: Proc IEEE INFOCOM, Rio de Janeiro, Brazil, 2009

37 Pias M, Crowcroft J, Wilbur S, et al. Lighthouses for scalable distributed location. In: IPTPS'03, Berkeley, CA, USA, 2003

38 Costa M, Castro M, Rowstron A, et al. PIC: Practical internet coordinates for distance estimation. In: Conf Distributed Systems, Tokyo, Japan, 2004

39 Mao Y, Saul L K, Smith J M. IDES: An internet distance estimation service for large networks. IEEE J Select Areas Commun, 2006, 24: 2273–2284

40 Dabek F, Cox R, Kaashoek F, et al. Vivaldi: A decentralized network coordinate system. In: Proc ACM SIGCOMM, Portland, OR, 2004

41 Shavitt Y, Tankel T. Big-bang simulation for embedding network distances in Euclidean space. IEEE/ACM Trans Netw, 2004, 12: 993–1006

42 Lehman L, Lerman S. PCoord: Network position estimation using peer-to-peer Measurements. In: Proc of the 3rd IEEE International Symposium on Network Computing and Applications, Cambridge, MA, USA, 2004

43 Krishnamurthy B, Wang J. On network-aware clustering of web clients. In: Proc ACM SIGCOMM'00, Stockholm, Sweden, 2000

44 Routeviews project. http://www.routeviews.org

45 RIPE project. http://www.ripe.net/np/ris/

46 Cramer C, Kutzner K, Fuhrmann T. Bootstrapping locality-aware P2P networks. In: Proc 12th Intl Conf Networks (ICON'04), Singapore, 2004. 1: 357–361

47 Saucez D, Donnet B, Bonaventure O. Implementation and preliminary evaluation of an ISP-driven informed path selection. In: Proc ACM CoNEXT, New York, NY, USA, 2007

48 YD/T 2146-2010. Technical framework for carrier network-aware P2P traffic optimization

49 Seedorf J, Burger E. Application-layer traffic optimization (ALTO) problem statement. RFC 5693, IETF, Oct, 2009

50 Alimi R, Penno R, Yang Y. ALTO protocol. draft-ietf-alto-protocol-06.txt, Oct, 2010

51 Bindal R, Cao P, Chan W, et al. Improving traffic locality in BitTorrent via Biased neighbor selection. In: Proc IEEE ICDCS, Lisboa, Portugal, 2006

52 Horovitz S, Dolev D. LiteLoad: Content unaware routing for localizing P2P protocols. In: IPDPS, Shanghai, China, 2008

53 Ahlswede R, Cai N, Li S R, et al. Network information flow. IEEE Trans Inf Theory, 2000, 46: 1204–1216

54 Li S R, Yeung R W, Cai N. Linear network coding. IEEE Trans Inf Theory, 2003, 49: 371–381

55 Koetter R, Medard M. An algebraic approach to network coding. IEEE/ACM Trans Netw, 2003, 11: 782–795

56 Ho T, Koetter R, Medard M, et al. The benefits of coding over routing in a randomized setting. In: Proc of International Symposium on Information Theory, Yokohama, Japan, 2003

57 Gkanisidis C, Rodnguez P R. Network coding for large scale content distribution. In: Proc IEEE INFOCOM, Miami, FL, USA, 2005

58 Wang M, Li B. Lava: A reality check of network coding in peer-to-peer live streaming. In: Proc IEEE INFOCOM, Anchorage, Alaska, USA, 2007

59 Lei Y C, Cheng S, Wu C L, et al. P2P content distribution with network coding (in Chinese with English abstract). J Comput R&D, 2009, 46: 108–119
60 Tomozei D C, Massoulié L. Flow control for cost-efficient peer-to-peer streaming. In: Proc IEEE INFOCOM, San Diego, CA, USA, 2010
61 Zhang G Q, Zhang G Q, Cheng S Q. LANC: locality-aware network coding for better P2P traffic localization. Comput Netw, 2011, 55: 1242–1256
62 Tang M D, Zhang G Q, Yang J, et al. A survey of P2P traffic optimization technologies (in Chinese). Telecommun Network Tech, 2009, 1: 1–7
63 Koponen T, Chawla M, Chun B G, et al. A data-oriented (and beyond) network architecture. In: Proc ACM SIGCOMM, Kyoto, Japan, 2007
64 Jacobson V, Smetters D K, Thornton J D, et al. Networking named content. In: Proc CoNEXT'09, Rome, Italy, 2009
65 Paul S, Pan J, Jain R. Architecture for the future networks and the next generation Internet: A survey. Comput Commun, 2011, 34: 2–42