Peer-to-Peer Infrastructure: A Survey of Research on the Application-Layer Traffic Optimization Problem and the Need for Layer Cooperation

Enrico Marocco
Telecom Italia Lab
Email: enrico.marocco@telecomitalia.it

Vijay K. Gurbani, Volker Hilt, Ivica Rimac,
and Marco Tomsu
Bell Laboratories, Alcatel-Lucent
Email: {vkg,volkerh,rimac}@bell-labs.com
and marco.tomsu@alcatel-lucent.com

***Abstract*—A significant part of the Internet traffic today is generated by peer-to-peer (P2P) applications used traditionally for file-sharing, and more recently for real-time communications and live media streaming. Such applications discover a route to each other through an overlay network with little knowledge of the underlying network topology. As a result, they may choose peers based on information deduced from empirical measurements, which can lead to suboptimal choices. We refer to this as the Application Layer Traffic Optimization (ALTO) problem and present a survey of existing literature. We summarize and compare existing approaches, identify open research issues and argue for the need of layer cooperation as a solution to the ALTO problem . Finally, we examine the role of the Internet Engineering Task Force (IETF) [1] in standardizing specific protocols related to this problem.**

I. Introduction and problem statement

A significant part of today's Internet traffic is generated by peer-to-peer (P2P) applications, used originally for file sharing, and more recently for realtime multimedia communications and live media streaming. P2P applications are posing serious challenges to the Internet infrastructure; by some estimates, P2P systems are so popular that they make up anywhere between 50% to 85% of the entire Internet traffic [1–6].

P2P systems ensure that popular content is replicated at multiple instances in the overlay. But perhaps ironically, a peer searching for that content may ignore the topology of the latent overlay network and instead select among available instances based on information it deduces from empirical measurements, which, in some particular situations may lead to suboptimal choices. For example, a shorter round-trip time estimation is not indicative of the bandwidth and reliability of the underlying links, which have more of an influence than delay for large file transfer P2P applications.

Most distributed hash tables (DHT) – the data structure that imposes a specific ordering for P2P overlays – use greedy forwarding algorithms to reach their destination, making locally optimal decisions that may not turn to be globally optimized [7]. This naturally leads to the Application-Layer Traffic Optimization (ALTO) problem [8]: how to best provide the topology of the underlying network while at the same time allowing the requesting node to use such information to effectively reach the node on which the content resides. Thus, it would appear that P2P networks with their application layer routing strategies based on overlay topologies are in direct competition against the Internet routing and topology.

One way to solve the ALTO problem is to build distributed application-level services for location and path selection [9–14], in order to enable peers to estimate their position in the network and to efficiently select their neighbors. Similar solutions have been embedded into P2P applications such as Azureus [15]. A slightly different approach is to have the Internet service provider (ISP) take a pro-active role in the routing of P2P application traffic; the means by which this can be achieved have been proposed [16–18]. There is an intrinsic struggle between the layers – P2P overlay and network underlay – when performing the same service (routing), however there are strategies to mitigate this dichotomy [19, 20]. Our position in this paper is that solutions to the ALTO problem will be best achieved by enabling communications between the P2P application layer and the network layer.

The rest of this paper is structured as follows: section II surveys the existing literature on topology estimation and layer interactions. Section III makes a case for our position on the need for layer cooperation. Section IV details the open research issues that will need to be addressed for layer cooperation, and section V concludes the paper by examining the role that IETF can play in fostering protocols and solutions for these issues.

II. Survey of existing literature

Gummadi et al. [7] compare popular DHT algorithms and besides analyzing their resilience, provide an accurate evaluation of how well the logical overlay topology maps on the physical network layer. In their paper, relying only on measurements independently performed by overlay nodes without the support of additional location information provided by external entities, they demonstrate that the most efficient algorithms in terms of resilience and proximity performance

[1] We point out that these are our views as long-time members and contributors to the IETF and should not be construed as being endorsed by the IETF.

are those based on the simplest geometric concept (i.e. the ring geometry, rather than hypercubes, tree structures and butterfly networks).

Regardless of the geometrical properties of the DHTs involved, interactions between application-layer overlays and the underlying networks are a rich area of investigation. The available literature in this field can be taxonomixed in two categories: using application-level techniques to estimate topology and using an infrastructure of some sort.

A. Application-Level Topology Estimation

In order to provide P2P overlays with topology information essential for optimizing node selection, different systems have been proposed.

Estimating network topology information on the application level has been an area of active research. Early work on network distance estimation based on clustering by Francis et al. [9] was followed by the introduction of network coordinate systems such as GNP by Ng et al. [10]. Network coordinate systems embed the network topology in a low-dimensional coordinate space and enable network distance estimations based on vector distance. Vivaldi [11] and PIC [12] propose distributed network coordinate systems that do not need landmarks for coordinate calculation. Vivaldi is now being used in the popular P2P application Azureus [15] and studies indicate that it scales well to very large networks [21].

Coordinate systems require the embedding of the Internet topology into a coordinate system. This is not always possible without errors, which impacts the accuracy of distance estimations. For example, it has proved to be difficult to embed the triangular inequalities found in Internet path distances [24]. Thus, Meridian [13] abandons the generality of network coordinate systems and provides specific distance evaluation services. The Ono project [22] take a different approach and uses network measurements from content-distribution network (CDN) like Akamai to find nearby peers [23]. Used as a plugin to the Azureus BitTorrent client, Ono provides 31% average download rate improvement.

Most of the work on estimating topology information focuses on predicting network distance in terms of latency and does not provide estimates for other metrics such as throughput. However, for many P2P applications throughput is often more important than latency. iPlane [14] aims at creating an atlas of the Internet using measurements that contains information about latency, bandwidth, capacity and loss rates.

To determine features of the topology, network measurement tools, e.g., based on packet dispersion techniques (packet pairs and packet trains) as described by Dovrolis et al. in [25] can be used. Moreover, methods of active network probing or passive traffic monitoring can also generate network statistics relating indirectly to performance attributes that cannot be directly measured but need to be inferred. An extensive study of such techniques that are summarized under the notion of network tomography has been provided by Coates et al. [26].

B. Topology Estimation through Layer Cooperation

Instead of estimating topology information on the application level through distributed measurements, this information could be provided by the entities running the physical networks – usually ISPs or network operators. In facts, they have full knowledge of the topology of the networks they administer and, in order to avoid congestion on critical links, are interested in helping applications to optimize the traffic they generate. The remainder of this section briefly describes three recently proposed solutions that follow such an approach to address the ALTO problem; we consider this a good example of what could be standardized by the IETF.

1) P4P Architecture: The architecture proposed by Xie et al. [16] have been adopted by the DCIA P4P working group [27], an open group established by ISPs, P2P software distributors and technology researchers with the dual goal of defining mechanisms to accelerate content distribution and optimize utilization of network resources.

The main role in the P4P architecture is played by servers called “iTrackers”, deployed by network providers and accessed by P2P applications (or, in general, by elements of the P2P system) in order to make optimal decisions when selecting a peer to connect. An iTracker may offer three interfaces:

- Info: Allows P2P elements (e.g. peers or trackers) to get opaque information associated to an IP address. Such information is kept opaque to hide the actual network topology, but can be used to compute the network distance between IP addresses.
- Policy: Allows P2P elements to obtain policies and guidelines of the network, which specify how a network provider would like its networks to be utilized at a high level, regardless of P2P applications.
- Capability: Allows P2P elements to request network providers’ capabilities.

The P4P architecture is under evaluation with simulations, experiments on the PlanetLab distributed testbed and with field tests with real users. Initial simulations and PlanetLab experiments results [27] indicate that improvements in BitTorrent download completion time and link utilization in the range of 50-70% are possible. Results observed in field tests conducted with a modified version of the software used by the Pando content delivery network [28] show improvements in download rate by 23% and a significant drop in data delivery average hop count (from 5.5 to 0.89) in certain scenarios.

2) Oracle-based ISP-P2P Collaboration: In the general solution proposed by Aggarwal et al. [17,29], network providers host servers, called “oracles”, that help P2P users choose optimal neighbours.

The mechanism is fairly simple: a P2P user sends the list of potential peers to the oracle hosted by its ISP, which ranks such a list based on its local policies. For instance, the ISP can prefer peers within its network, to prevent traffic from leaving its network; further, it can pick higher bandwidth links, or peers that are geographically closer. Once the application has obtained an ordered list, it is up to it to establish connections with a number of peers it can individually choose, but it has enough information to perform an optimal choice.

Such a solution has been evaluated with simulations and experiments run on the PlanetLab testbed and the results show both improvements in content download time and a reduction of overall P2P traffic, even when only a subset of the applications actually query the oracle to make their decisions.

3) ISP-Driven Informed Path Selection (IDIPS) Service: The IDIPS solution [18] was presented during the SHIM6 session of the 71st IETF meeting. It is essentially a modified version of the solution described in section II-B2, extended to accept lists of source addresses other than destinations in order to function also as a back end for protocols like SHIM6 and LISP (which aim at optimizing path selection at the network layer). An evaluation performed on IDIPS shows that costs for both providing and accessing the service are negligible [30].

III. The case for Layer Cooperation as a solution to the ALTO problem

The application-level techniques described in Section II-A provide tools for peer-to-peer applications to estimate parameters of the underlying network topology. Although these techniques can improve application performance, there are fundamental limitations of what can be achieved by operating only on the application level.

Topology estimation techniques use abstractions of the network topology which often hide features that would be of interest to the application. Network coordinate systems, for example, are unable to detect overlay paths shorter than the direct path in the Internet topology. However, these paths frequently exist in the Internet [24]. Similarly, application-level techniques may not accurately estimate topologies with multipath routing.

When using network coordinates to estimate topology information the underlying assumption is that distance in terms of latency determines performance. However, for file sharing and content distribution applications there is more to performance than just the network latency between nodes. The utility of a long-lived data transfer is determined by the throughput of the underlying TCP protocol, which depends on the round-trip time as well as the loss rate experienced on the corresponding path [31]. Hence, these applications benefit from a richer set of topology information that goes beyond latency including loss rate, capacity, available bandwidth.

Some of the topology estimation techniques used by peer-to-peer applications need time to converge to a result. For example, current BitTorrent clients implement local, passive traffic measurements and a tit-for-tat bandwidth reciprocity mechanism to optimize peering selection at a local level. Peers eventually settle on a set of neighbors that maximizes their download rate but because peers cannot reason about the value of neighbors without actively exchanging data with them and the number of concurrent data transfers is limited (typically to 5-7), convergence is delayed and easily can be sub-optimal.

Skype's P2P VoIP application chooses a relay node in cases where two peers are behind NATs and cannot connect directly. Ren et al. [32] measured that the relay selection mechanism of Skype is (1) not able to discover the best possible relay nodes in terms of minimum RTT (2) requires a long setup and stabilization time, which degrades the end user experience (3) is creating a non-negligible amount of overhead traffic due to probing a large number of nodes. They further showed that the quality of the relay paths could be improved when the underlying network AS topology is considered.

Some features of the network topology are hard to infer through application-level techniques and it may not be possible to infer them at all. An example for such a features are service provider policies and preferences such as the state and cost associated with interdomain peering and transit links. Another example is the traffic engineering policy of a service provider, which may counteract the routing objective of the overlay network leading to a poor overall performance [19].

Finally, application-level techniques often require applications to perform measurements on the topology. These measurements create traffic overhead, in particular, if measurements are performed individually by all applications interested in estimating topology.

Given these problems of application-level topology estimation techniques we argue that a better solution involves the cooperation between network and application layer.

IV. Open research issues

We believe that there are sizable open research issues to tackle in an infrastructure-based approach to traffic optimization. The following is not an exhaustive list, but a representative sample of the pertinent issues.

Co-ordinate estimation or path latencies? Despite the many solutions that have been proposed for providing applications with topology information in a fully distributed manner, there is currently an ongoing debate in the research community whether such solutions should focus on estimating nodes' coordinates or path latencies. Such a debate has recently been fed by studies showing that the triangle inequality on which coordinate systems are based is often proved false in the Internet [24]. Proposed systems following both approaches – in particular, Vivaldi [11] and PIC [12] following the former, Meridian [13] and iPlane [14] the latter – have been simulated, implemented and studied in real-world trials, each one showing different points of strength and weaknesses. Concentrated work will be needed to determine which of the two solutions will be conducive to the ALTO problem.

Malicious nodes. Another open issue common in most distributed environments consisting of a large number of peers is the resistance against malicious nodes. Security mechanisms to identify misbehavior are based on triangle inequality checks [12], which however tend to fail and thus return false positives in presence of measurement inaccuracies induced, for example, by traffic fluctuations that occur quite often in large networks [21, 24]. Beyond the issue of using triangle inequality checks, authoritatively authenticating the identity of an oracle, and preventing an oracle from attacks are also important. Exploration of existing techniques – such as public key infrastructure or identity-based encryption for authenticating the identity and the use of secure multi-party computation techniques to prevent an oracle from collusion attacks – need to be studied for judicious use in ALTO-type of solutions.

Information integrity. Similarly, even in controlled architectures deployed by network operators where system elements may be authenticated [16–18], it is still possible that the information returned to applications is deliberately altered, for example, assigning higher priority to cheap (monetary-wise) links instead of neutrally applying proximity criteria. What are the effects of such deliberate alterations if multiple peers collude to determine a different route to the target, one that is not provided by an oracle? Similarly, what are the consequences if an oracle targets a particular node in another AS by redirecting an inordinate number querying peers to it causing, essentially, a DDoS attack on the node? Furthermore, does an oracle broadcast or multi-cast a response to a query? If so, techniques to protect the confidentiality of the multi-cast stream will need to be investigated to thwart “free riding” peers.

Simulate or build? Much debate in the P2P research community clusters around the simulate or build question. Undoubtedly, it is hard to foresee how proposed systems would perform in the Internet. Simulations and testbed emulations are in most cases the only options available on benchmarking the performance of the system. However these have often proved to be inadequate – in at least one particular case [21], they have only provided a rough optimistic approximations of what would be measured in the real world. Even using near-realistic testbeds such as PlanetLab do not suffice for certain aspects of quantifying P2P traffic: more often, these testbeds do not take in account the user component, which is crucial for file-sharing P2P systems. After all, a P2P system depends on the choices and interests of its users to fetch, store, and disseminate content and it is hard to simulate a sizable user population with varying tastes to authoritatively observe the behavior of a P2P network. New techniques in simulation or testbed usage would need to be investigated.

Richness of topological information. Many systems already use RTT to account for delay when establishing connections with peers (e.g., CAN, Bamboo). An operator can provide not only the the delay metric but other metrics that the peer cannot figure out on its own. These metrics may include the characteristics of the access links to other peers, bandwidth available to peers (based on operator’s engineering of its network), network policies, and preferences such as state and cost associated with intradomain peering links, and so on. Exactly what kinds of metrics can an operator provide to stabilize the network throughput will also need to be investigated.

Applicability of ALTO to centralized or semi-centralized services. The Joost Video-on-Demand Service uses P2P technology to distribute streaming video at a bit rate of about 600 kbit/s and higher. In their experimental analysis, Lei et al. [33] conclude that the system is heavily based on a media server infrastructure – in particular for channels with lower popularity – and that a geographical distance based on address prefix analysis is considered during the server selection. They show that the peer selection process today is unlikely based on topology locality. Instead the peer’s capacity influences the the creation of the peer lists similar to BitTorrent: low capacity peers connect mostly with other low capacity peers to avoid wasting the high capacity peers bandwidth. It remains to be seen whether an ALTO-type of solution can be conducive to a hybrid media-server assisted P2P system.

V. Role of the IETF

We believe that the IETF can and should play an important role in designing specific protocols and mechanisms for an effective solution to address the Application-Layer Traffic Optimization (ALTO) problem [8]. The IETF is recognized for its high quality standards and is thus the best candidate to foster the wide adoption the effectiveness of an ALTO solution.

As mentioned previously, such a solution should enable cross-layer cooperation, allowing communications between applications and network elements aware of the underlying network topology. In particular, the IETF should specify the following:

- a lookup mechanism to be used by applications to discover the appropriate network elements to query in order to obtain topology information they need for ALTO;
- a protocol to be used in communications between applications and those network elements.

It is conceivable that P2P users may not be comfortable with operator intervention to provide topology information. To eliminate this intervention, alternative schemes to estimate topological distance can be used. For instance, Ono uses client redirections generated by Akamai CDN servers as an approximation for estimating distance to peers; Vivaldi, GNP and PIC use synthetic coordinate systems. A network service provided by a neutral third-party could make the collected topological information available to other peers without the cooperation of the ISP. [2] The protocols specified by the IETF should work uniformly, irrespective of querying an operator-provided resource or a neutral third-party resource.

Acknowledgment

The authors thank Rohan Mahy and Emil Ivov for their support and for providing useful comments during the writing of the problem statement. Discussions within Telecom Italia, Alcatel-Lucent, TU-Berlin and the P4P working group have also contributed substantially to the contents of the paper.

References

[1] M. Meeker and D. Joseph. *The State of the Internet, Part 3.* Web 2.0 Conference, November 2006; available online at http://www.morganstanley.com/institutional/techresearch/pdfs/Webtwopto2006.pdf.

[2] It is likely in the future that protocols for implementing application-level systems like those described in section II-A will also need to be standardized. However, currently these techniques are still considered research topics and therefore too early to be engineered in the IETF.

[2] T. Karagiannis, A. Broido, N. Brownlee, K.C. Claffy, and M. Faloutsos. *Is P2P dying or just hiding?* In proceedings of the IEEE Globecom Conference, 2004.

[3] Lightreading. *Controlling P2P traffic.* Available online at http://www.lightreading.com/document.asp?site=lightreading&doc_id=44435&page_number=3

[4] linuxReviews.org, *Peer to peer network traffic may account for up to 85% of Internet's bandwidth usage.* Linux Review, November 2004. Available online at http://linuxreviews.org/news/2004/11/05_p2p/

[5] A. Parker, *The true picture of peer-to-peer filesharing.* Available online at http://www.cachelogic.com.

[6] J. Glasner, *P2P fuels global bandwidth binge.* Wired Magazine, April 15, 2005. Available online at http://www.wired.com/techbiz/media/news/2005/04/67202.

[7] K. Gummadi, R. Gummadi, S. Gribble, S. Ratnasamy, S. Shenker, and I. Stoica. *The impact of DHT routing geometry on resilience and proximity.* In Proceedings of ACM SIGCOMM, August 2003.

[8] E. Marocco, and V. Gurbani. *Application-Layer Traffic Optimization (ALTO) Problem Statement.* Internet-Draft draft-marocco-alto-problem-statement-01 (Work in Progress), April 2008.

[9] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang. *IDMaps: A global Internet host distance estimation service.* In IEEE/ACM Transactions on Networking, October 2001.

[10] T. S. E. Ng, and H. Zhang. *Predicting internet network distance with coordinates-based approaches.* In Proceedings of IEEE INFOCOM, June 2002.

[11] F. Dabek, R. Cox, F. Kaashoek, and R. Morris. *Vivaldi: A Decentralized Network Coordinate System.* In Proceedings of ACM SIGCOMM, August 2004.

[12] M. Costa, M. Castro, A. Rowstron, and P. Key. *PIC: Practical Internet coordinates for distance estimation.* In International Conference on Distributed Systems, Tokyo, Japan, March 2004.

[13] B. Wong, A. Slivkins, and E. G. Sirer. *Meridian: A lightweight network location service without virtual coordinates*. In Proceedings of ACM SIGCOMM, August 2005.

[14] H. V. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and A. Venkataramani. *iPlane: an information plane for distributed services.* In Proceedings of OSDI, November 2006.

[15] Azureus BitTorrent Client, http://www.azureus.com/.

[16] H. Xie, A. Krishnamurthy, A. Silberschatz, Y. R. and Yang. *P4P: Explicit Communications for Cooperative Control Between P2P and Network Providers.* http://www.dcia.info/documents/P4P_Overview.pdf.

[17] V. Aggarwal, A. Feldmann, and C. Scheidler. *Can ISPs and P2P systems co-operate for improved performance?* In ACM SIGCOMM Computer Communications Review (CCR), 37:3, pp. 29-40, July 2007.

[18] Saucez, D., Donnet, B., and O. Bonaventure. *IDIPS : ISP-Driven Informed Path Selection.* Internet-Draft draft-saucez-idips-00 (Work in Progress), February 2008.

[19] S. Seetharaman, V. Hilt, M. Hofmann, M. Ammar. *Preemptive Strategies to Improve Routing Performance of Native and Overlay Layers.* In Proceedings of IEEE INFOCOM, May 2007.

[20] V. Pappas, V. Hilt, M. Hofmann. *Coordinate-Based Routing for Overlay Networks*. In Proceedings of IEEE ICCCN, August 2007.

[21] J. Ledlie, P. Gardner, and M. Seltzer. *Network Coordinates in the Wild.* Proceedings of NSDI, Cambridge, MA, April 2007.

[22] Northwestern University Ono Project, http://www.aqualab.cs.northwestern.edu/projects/Ono.html

[23] A.-J. Su, D. R. Choffnes, A. Kuzmanovic, and F. E. Bustamante, *Drafting behind Akamai (travelocity-based detouring).* In ACM SIGCOMM Computer Communication Review, 36(4), pp. 435–446, 2006.

[24] G. Wang, B. Zhang, T. S. E. Ng. *Towards Network Triangle Inequality Violation Aware Distributed Systems.* IMC'07, October 24-26, 2007, San Diego, California, USA.

[25] C. Dovrolis, P. Ramanathan, and D. Moore, *What do packet dispersion techniques measure?* In Proceedings of IEEE INFOCOM, pp. 905–914, 2001.

[26] M. Coates, A. Hero, R. Nowak, and B. Yu, *Internet Tomography.* IEEE Signal Processing Magazine 19(3), pp. 47–65, 2002.

[27] DCIA P4P Working group, http://www.dcia.info/activities/#P4P.

[28] OpenP4P Web Page, http://openp4p.net/front/fieldtests.

[29] V. Aggarwal, O. Akonjang, and A. Feldmann. *Improving User and ISP Experience through ISP-aided P2P Locality.* In Proceedings of IEEE Global Internet Symposium, April 2008.

[30] D. Saucez, B. Donnet, and O. Bonaventure. *Implementation and Preliminary Evaluation of an ISP-Driven Informed Path Selection.* In Proceedings of ACM CoNEXT, December 2007.

[31] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. *Modeling TCP throughput: A simple model and its empirical validation.* In Computer Communications Review, 28(4), 1998.

[32] S. Ren, L. Guo, and X. Zhang, *ASAP: An AS-aware peer-relay protocol for high quality VoIP.* In proceedings of IEEE International Conference on Distributed Computing Systems, pp. 70-79, 2006.

[33] J. Lei, L. Shi, and X. Fu, *An experimental analysis of Joost peer-to-peer VoD service.* Technischer Bericht des Instituts für Informatik, Georg-August-Universität Göttingen, October 2007.