Network Working Group J. Peterson
Internet-Draft NeuStar
Intended status: Standards Track A. Cooper
Expires: August 27, 2009 Center for Democracy & Technology

February 23, 2009

Report from the IETF workshop on P2P Infrastructure, May 28, 2008
draft-p2pi-cooper-workshop-report-01

Status of this Memo

Copyright Notice

Abstract

This document reports the outcome of a workshop organized by the

Real-time Applications and Infrastructure Area Directors of the IETF to discuss network delay and congestion issues resulting from increased P2P traffic volumes. The workshop was held on May 28, 2008 at MIT in Cambridge, MA, USA. The goals of the workshop were twofold: to understand the technical problems ISPs and end users are experiencing as a result of high volumes of P2P traffic, and to begin to understand how the IETF may be helpful in addressing these problems. Gaining an understanding of where in the IETF this work might be pursued and how to extract out feasible work items were highlighted as important tasks in pursuit of the latter goal. The workshop was very well attended and produced several work items that have since been taken up by members of the IETF community.

Table of Contents

1. Introduction

Increasingly, large ISPs are encountering issues with P2P traffic. The transfer of static, delay-tolerant data between nodes on the Internet is a well-understood problem, but traditional management of fairness at the transport level is under strain from applications designed to achieve the best end-user transfer rates. At peak times this results in networks running near absolute capacity, causing all traffic to incur delays; the applications that bear the brunt of this additional latency are real-time applications like VoIP and Internet gaming. To explore how IETF standards work could be useful in addressing these issues, the Real-time Applications and Infrastructure Area Directors organized a "P2P Infrastructure" workshop and invited contributions from subject matter experts in the problem and solution spaces.

The goals of the workshop were twofold: to understand the technical problems ISPs and end users are experiencing as a result of high volumes of P2P traffic, and to begin to understand how the IETF may be helpful in addressing these problems. Gaining an understanding of where in the IETF this work might be pursued and how to extract out feasible work items were highlighted as important tasks in pursuit of the latter goal. The workshop's focus was on engineering solutions that promise some imminent benefit to the Internet as a whole, as opposed to longer-term research or closed properietary solutions. While public policy must inform work in this space, crafting or debating public policy was outside the scope of the workshop.

Position papers were solicited in the weeks prior to the workshop, and a limited number of speakers were invited to present their views at the workshop based on these submissions. This report is a summary of all participants' contributions. The program committee and participant list are attached in Appendix A and Appendix B, respectively. The agenda of the workshop can be found in Appendix C.
A link to the presentations given at the workshop and the position papers submitted prior to the workshop is in Appendix D.

The workshop showcased the IETF community's recognition of the impact of P2P and other high-volume applications on the Internet as a whole.
Participants welcomed the opportunity to discuss potential standardization work that network operators, applications providers, and end users would all find mutually beneficial. Two transport-related work items gained significant traction: designing a protocol for very deferential end-to-end congestion control for delay-

tolerant
   applications, and producing an informational document about the
   reasoning behind and effects of applications opening multiple
   transport connections at once. A seperate area of interest that
   emerged at the workshop focused on improving peer selection by
having

networks make more information available to applications. Finally, presenters also covered traditional approaches to multiple service-tier queuing such a diffserv.

2. Scoping of the Problem and Solution Spaces

The genesis for the P2PI workshop grew in large part out of specific pain points that ISPs are experiencing as a result of high volumes of P2P traffic. However, several workshop participants felt that the IETF should approach a more general space of problems, of which P2P-related congestion may be merely one instance.

For example, high-volume applications besides P2P, whether they already exist or have yet to be developed, could cause congestion issues similar to those caused by P2P. And while much attention has been paid to congestion on access links, increased traffic volumes could impact other parts of the network. Although the workshop focused primarily on the specific causes and effects of current P2P traffic volumes, it may be useful in the future for the IETF to consider how to pursue solutions to these larger problems.

Obtaining more data about Internet congestion may also be a helpful step before the IETF pursues solutions. This data collection could focus on where in the network congestion is occurring, its duration and frequency, its effects, and its root causes. Although individual service providers expressed interest in sharing congestion data, strategies for reliably and regularly obtaining and disseminating such data on a broad scale remain elusive.

3. Service Provider Perspective

To help participants gain a fuller understanding of one specific network operator view of P2P-induced congestion, Jason Livingood and Rich Woundy provided an overview of Comcast's network and approach to management of P2P traffic.

3.1. DOCSIS Architecture and Upstream Contention

In the Data Over Cable Service Interface Specification (DOCSIS) architecture [DOCSIS] used for many cable systems, there may be a single Cable Modem Termination System (CMTS) serving hundreds or thousands of residential cable customers. Each CMTS has multiple DOCSIS domains, each of which typically has a single downstream link and a number of upstream links. Each CMTS is connected through a hybrid fiber-coaxial (HFC) network to subscribers' cable modems.

The limiting resource in this architecture is usually bandwidth, so bandwidth is typically the measure used for capacity planning. As with all networks, congestion manifests itself when instantaneous load exceeds available capacity.

In the upstream direction, any cable modem connected to a CMTS can make a request to the CMTS to transmit, and requests are randomized to minimize collisions. With many cable modems issuing requests at once, the requests may collide, resulting in delays. DOCSIS does not specify a size for cable modem buffers, but buffer delays of one to four seconds have been observed with various cable modems from different vendors.

Once the CMTS has granted a cable modem the ability to transmit its data PDU, the modem can piggyback its next request on top of that data PDU. In situations with a lot of upstream traffic, piggybacking happens more often, which sends heavy upstream users to the front of the CMTS queue, ahead of interactive but less-upstream-intensive applications. For example, if the CMTS is granting requests approximately every one to three milliseconds, then a cable modem transmitting data for a service like VoIP with a packetization delay of 20-30 milliseconds may get into contention with another modem on the same CMTS that is constantly transmitting upstream and piggybacking each new request. This may explain how heavy upstream users ultimately dominate the pipe over more interactive applications. Consequentially, it is imperative that assessments of the problem space, and potential solutions, are mindful of the influence that specific layer-2 networks may exert on the behavior of Internet traffic, especially when considering the alleviation of congestion in an access network.

3.2. TCP Flow Fairness and Service Flows

How TCP flow fairness applies to upstream requests to the CMTS is an open question. A CMTS sees many service flows, each of which could be a single TCP flow or many TCP flows (or UDP). The CMTS is not aware of the source or destination IP address of a packet until it has already been transmitted upstream, so those cannot be used to impose flow fairness.

A particular cable modem can have multiple service flows defined. For example, a modem that is also a VoIP endpoint can provision a service flow for VoIP that would allow VoIP traffic to avoid the upstream request process to the CMTS (and thereby avoid contention with other modems). The service flow would have upstream capacity provisioned for it. The modem would have a separate service flow for best efforts traffic. Some ISPs provision such a flow for their own VoIP offerings; others allow subscribers to pay extra to have

particular traffic assigned to a provisioned service flow.

It may also be possible for an ISP to provision such a flow on the fly when it recognizes the need for it. DiffServ [RFC2475] bits set by the customer premises equipment could be used to classify flows, for example.

3.3. Service Provider Responses

Starting in 2005, ISP customers were increasingly complaining about the performance of delay-sensitive traffic (VoIP and gaming), due in part to the issues arising out of the DOCSIS architecture as described above. At the same time, ISPs were seeing heavy growth in P2P traffic, and an increasing correlation between high levels of P2P activity and packet loss.

In responding to this situation, cable ISPs have several avenues to pursue. The newest generation of the DOCSIS specification, DOCSIS 3.0, enables faster transfer rates than most cable systems currently support. While the rollout of DOCSIS 3.0 will provide additional capacity, it will likely not obviate the need for congestion management in an environment where client software is designed to maximize bandwidth consumption regardless of available capacity.

Congestion management can take many forms; Jason and Rich explained the new protocol-agnostic approach that Comcast is currently trialing. Prior to these trials, all traffic was marked as "best efforts." During the trials, all traffic is re-classified as "priority." When a CMTS is approaching peak congestion on a particular upstream or downstream port (the "Near Congestion State"), some subscribers will have traffic re-classified as "best efforts." The threshold for determining when a CMTS port is in Near Congestion State and the number of minutes it remains in this state are both parameters being explored during the trials. To re-classify upstream traffic, a new default DOCSIS service flow is used that has the same provisioned bandwidth as the "priority" stream, but is treated with lower priority.

The subscribers whose traffic is re-marked will be selected by determining whether they have temporarily entered a "Long Duration Bulk Consumption State." This state is achieved by consuming a certain amount of bandwidth over a certain period of minutes (both tweakable parameters being explored during the trials). These thresholds will depend on the subscriber's service tier -- subscribers who pay for more bandwidth will have higher thresholds. The re-marking will not distinguish between multiple users of the same subscriber connection, so one family member's P2P usage could cause another family member's Web browsing traffic to be lowered in

priority. There is no current mechanism for users to determine that their traffic has been re-marked.

By temporarily reducing the traffic priority of subscribers who have been consuming bandwidth in bulk for lengthy periods, this congestion management technique aims to preserve a good user experience for subscribers with burstier traffic patterns, including those using real-time applications. As compared to an approach that reduces particular subscribers' bandwidth during periods of congestion, this technique eliminates the ability for applications to set their own priority levels, but it also avoids the negative connotations that some users may associate with bandwidth reductions.

This approach involves many tweakable parameters. A large part of the trial process is aimed at determining the best settings for these parameters, but there may also be opportunities to work with the research community to identify the best way to adjust the thresholds necessary to optimize the performance of the management technique.

4. Application Provider Perspective

Stanislav Shalunov provided an overview of BitTorrent's view of the impact of increased P2P traffic volumes and potential mitigations. The impact is described here; his proposed solutions (comprising the bulk of his talk) are addressed in the appropriate subsections of Section 5.

As uptake in P2P usage has grown, so has end-user latency. For example, a user whose uplink capacity is 250-500 Kbps and whose modem buffer has a capacity of 32-64 Kbps may easily fill the buffer (unless the modem uses AQM, which is uncommon). This can result in delay on the order of seconds, with disastrous effects on application performance. On a cable system with shared capacity between neighbors, one neighbor could saturate the buffer and affect the latency of another neighbor's traffic.

5. Potential Solution Areas

The submissions received in advance of the workshop covered a broad array of work addressing specific aspects of P2P traffic volume and other related issues. Solution suggestions generally fell into one or more of three topic areas: improving peer selection, new approaches to congestion control, and quality of service mechanisms. The workshop discussions and outcomes in each area are described below.

5.1. Improving Peer Selection: Information Sharing, Localization, and Caches

Peer selection is an integral factor in determining the efficiency of P2P networks from both the ISP and the P2P client point of view. How peers are selected will determine both network load and client performance.

The way that P2P clients select peers today varies from protocol to protocol and client to client, but as a general matter peers are largely oblivious to routing-level and network topology information. This results in P2P topologies that are agnostic of underlay topologies and constraints.

Approaches to closing this gap generally involve an entity that has knowledge of network topology, costs, or constraints (e.g., an ISP) making some of this information available to P2P clients or trackers. This information may be used to localize traffic based on some metric of locality, or otherwise alter peer selection decisions based on the provided network information (hereafter referred to simply as "localization"). One special case of this kind of approach would help peers find caches containing the content they seek.

Any alteration to current peer selection algorithms will have engineering trade-offs. BitTorrent, for example, used randomized peer selection by design. Choosing peers randomly out of a large selection helps to average out problems among peers, and it allows for connections to good peers that may be far away. Randomized peer selection also supports "rarest first" piece selection, which allows swarms to continue even when the original seed disappears and distributes pieces so that more peers are likely to have pieces of interest to other peers. Any move away from randomized selection would have to take these factors into account.

Although localization has the potential to improve peer selection, the incentives for both parties to the information exchange are complex. ISPs may want to move traffic off of their own networks, which could motivate them to provide information to peers that has the opposite effect of what the peers would expect. Likewise, peers will want the use of the information they receive to result in performance improvements; otherwise, they have no incentive to consult with the network before selecting peers. Even when both parties find the information sharing to be beneficial, user experiences will not necessarily be uniform depending on the scope of the information provided and the peer's location. Localization information could form one component of a peer selection decision,

but it will likely need to be balanced against other factors.

Workshop participants discussed both current research efforts in this area and how IETF standards work may be useful in furthering the general concept of improved peer selection. Those discussions are summarized below.

5.1.1. Leveraging AS Numbers

One simple way to potentially make peer selection more efficient would be for a peer to prefer peers within its own AS. Transfers between peers within the same AS may be faster on some networks, although more data is needed to determine the extent of the potential improvement. On mobile networks, for example, the utility of AS numbers is limited since they do not correlate to geographic location. Peers may also see improvements by connecting to other peers within a specific set of ASes or IP prefixes provided by their ISPs. Some ISPs may have an incentive to expose this granularity of information because it will potentially reduce their transit costs.

A case study was conducted with the four most popular BitTorrent torrents to determine what the effect of localizing to an AS might be. The swarm sizes for the torrents were 9984, 3944, 2561, and 2023, with the size distributions appearing to be polynomial. With more than 20 peers in a single AS, peers within an AS could trade only with each other, avoiding interdomain traffic. More than half (57%) of peers in the four swarms were in ASes like this. Thus, in these cases connecting to peers within an AS could reduce transit traffic by at least 57%. If the ASes have asymmetric upload and download links, however, the resulting user experience may deteriorate since each peer's download speed would be limited by slower upload speeds.

With the largest swarm size at 9984, the probability of two peers being in the same neighborhood is too low to make localization to the neighborhood level worthwhile. Attempting a simple localization scheme, such as the AS localization described above, and determining its effectiveness likely makes more sense as a first step.

5.1.2. P4P: Provider Portal for P2P Applications

The P4P project [P4P] aims to design a framework to enable cooperation between "providers" and applications (including P2P), where providers may be ISPs, content distribution networks, or caching services. In this architecture, each provider can communicate information to P2P clients through a portal known as an iTracker. An iTracker could be identified through a DNS SRV record (perhaps with its own new record type), a whois look-up, or through a trusted third party.

An iTracker has different interfaces for different types of information that the provider may want to share. The core interface allows the provider to express the "virtual cost" of its intradomain or interdomain links. Virtual cost may reflect any kind of provider preferences, and may be based on the provider's choice of metrics, including utilization, transit costs, or geography. It is up to the provider to decide how dynamic it wants to be in updating its virtual cost determinations.

In tests of this framework, two parallel swarms were created with approximately the same number of clients and similar geographical and network distributions, both sharing the same file. One of the swarms used the P4P framework, with the ISP's network topology map as input to its iTracker, and the other swarm used traditional peer selection. The swarm without P4P saw 98% of traffic to and from peers external to the ISP, whereas with P4P that number was 50%. Download completion times for the P4P-enabled swarm improved approximately 20% on average.

5.1.3. Multi-Layer Tracker-Based Architecture

The multi-layer tracker-based P2P scheme described at the workshop is a generic example of an architecture that demonstrates how localization may be useful in principle.

In a traditional tracker-based P2P system, trackers provide clients with information about seeds and peers where clients can find the content they seek. A multi-layered tracker architecture incorporates additional "local" trackers that provide the same information, but only for content located within their own local network scope. Client queries are re-directed from the global tracker to the appropriate local trackers. Local trackers may also exist on multiple levels, in which case queries would be further re-directed. This sort of architecture could also serve hybrid P2P/content delivery networks, where the global tracker functions as both a tracker and a content server, and local trackers track locally provisioned caches in addition to seeds and peers.

One challenge in this architecture is determining what "local" means for trackers, seeds and peers. Locality could be dependent on traffic conditions, load balancing, static topology, policy or some other metric. These same considerations would also be crucial for determining appropriate cache placement in a hybrid network.

This architecture presents in the abstract the problem of re-

directing from a global entity to a local entity. Client queries need to find their way to the appropriate local tracker. This can be accomplished through an off-path, explicit mechanism where local

trackers register with the global tracker in advance, or through an on-path approach where the network proxies P2P requests. The off-path tracker format approach is preferable for performance and reliability reasons.

Inasmuch as the multi-layer scheme might require ISPs to aid peers in finding the optimal paths to unauthorized copies of copyrighted content, ISPs may be concerned about the legal liability of participating.

5.1.4. ISP-Aided Neighbor Selection

ISPs have a lot of knowledge about their networks: everything from the bandwidth, geography, and service class of particular nodes to overarching routing policies, OSPF and BGP metrics, and distances to peering points. The ISP-aided neighbor selection service described below seeks to leverage this knowledge without requiring ISPs to reveal any information that could not already be discerned through reverse-engineering by client applications.

The service consists of an "oracle" hosted by an ISP. The oracle receives a list of IP addresses from a network node, sorts the list according to its own confidential criteria, and returns the sorted list to the node. The peer ranking provided by the oracle could be viewed as a special case of the virtual cost interface described in the previous section.

This service could be used by P2P clients or trackers, or any other application that would benefit from learning its ISP's connection preferences. The oracle could be run as a web server or UDP service at a known location (perhaps similar to BIND).

For interdomain ranking, an ISP could rank its own peers first, or it could base its ranking on the AS number of the IPs in the provided list. Another option would be for multiple ISPs to work together to have their oracles exchange lists with each other.

The main challenge in implementing the oracle service is scalability. If peers need to communicate to the oracle the IP address of every peer they know, the size of oracle requests may be inordinately large. Additionally, today's largest swarms approach 10000 peers, and with every peer requesting a sorted list, oracle request volume will swell. With the growth of business models dependent upon P2P for distribution of content, swarms in the future may be far larger, further exacerbating the problem. Potential mitigations include having trackers instead of peers issue oracle requests, and using other peers' sorted lists as input rather than always using an unsorted list.

On the other hand, this approach is advantageous from a legal liability perspective, because it does not require ISPs to have any knowledge of where particular content might be located, or any role in directing peers to particular content.

5.1.5. Caches

Deploying caches as peers in P2P networks was suggested as a component of multiple different proposals put forth at the workshop. Caches may help to ease network load by reducing the need for peers to upload to each other and by localizing traffic.

The two main concerns about P2P caches relate to network capacity and legal liability. For caches to be useful, they will likely need to be large (one suggestion was that a 1 TB cache could service 30% of requests within a single AS, and a 100 TB cache could service 80% of requests). Large caches will require sizable bandwidth in order to avoid contention among peers. Caches would not be usefully placed within an HFC network on a cable system, for example.

The legal liability attached to hosting a P2P cache likely reduces the incentives to do so. Even under legal regimes where liability for caching may be unclear, ISPs and others may view hosting a cache as too great of a legal risk to be worthwhile.

5.1.6. Potential IETF Work

Much of the localization work discussed at the workshop is still in its initial stages, and many questions remain about the value that localization provides for varying network and overlay architectures. More data is needed to evaluate the effects on both traffic load and client performance. Understanding swarm distributions is important; swarms with long tails may not particularly benefit from localization.

Against this backdrop, the key task for the IETF as identified at the workshop is to pinpoint incrementally beneficial work items in the the spaces discussed above. In the future it may be possible to standardize entire P2P mechanisms, but as a starting point it makes more sense to single out core manageable components for standardization. The focus should be on items that are not so specific to one ISP or P2P network that standardization is rendered useless. Ideally, any mechanisms resulting from this work might apply to future applications that exhibit the same bandwidth-intensive properties as today's P2P file-sharing.

In considering any of these items, it will be necessary to ensure that the information exchanged by networks and applications does not

harm any of the parties involved. Not every piece of information exchanged with be beneficial or verifiable, and this fact must be recognized and accounted for. Solutions that leave applications or networks worse off than they already are today will not gain any traction.

It should also not be assumed that a particular party will be best suited to provide a particular kind of information. For example, an ISP may not know what the connection costs are in other ISPs' networks, whereas an overlay network that receives cost information from several ISPs may have a better handle on this kind of data. Standardization of information sharing should not assume the identity of particular parties doing the sharing.

The list of potential work items discussed at the workshop is provided below. Workshop participants showed particular interest in pursuing the first three items further.

5.1.6.1. AS Numbers

P2P clients are currently reliant on IP-to-AS mapping tables when they want to determine AS numbers. Providing a standard, easier way for clients to obtain this information may help to make peer selection more efficient on certain networks.

5.1.6.2. Querying for Preferred Peers

In situations where a peer or tracker can make requests in real time to a service that expresses its ISP's peering preferences, standardizing a format for requests and responses may be useful. The focus would be on the communication of the information, not on the criteria used to decide preferences. The information provided to peers would have to be crafted to ensure that it protects the privacy of other peers and safeguards proprietary network information.

5.1.6.3. Local Tracker, iTracker, Oracle, or Cache Discovery

With the deployment of trackers, iTrackers, oracles or other mechanisms that provide some information specific to a node's locality, nodes will need a way to find these resources. One task for the IETF could be to explore a way to do discovery, potentially by leveraging an existing discovery protocol (DNS, DHCP, anycast, etc.). Depending on the resource to be discovered, discovery may require only a simple look-up, or it may require a more complex determination of which resource is "closest" to the node issuing the request.

5.1.6.4. ISP Account Usage Information

Where ISP subscribers are bound by network usage policies or volume-based quotas, it may be useful to have a standard way of communicating the subscriber's current usage status. This would be similar to information about how many minutes of cell phone airtime are left in a subscriber's billing cycle. Applications could use this information to make decisions about when and how to transfer data. One challenge in implementing such a standard would be support for potentially limitless different ISP business models. The level of granularity that an ISP is able to provide may also be constrained depending on the pricing model and how dynamic the information is expected to be.

5.1.6.5. Tracker Formats

A multi-layered tracker approach could potentially be aided by a standard tracker format for re-directing from a global tracker to a local tracker. While the extent to which existing trackers will be willing to consult with other trackers is unclear, the re-direction format may have an analog in another context -- many HTTP servers build their own indexes of mirror information for a similar purpose, though these are not standardized. If the two problem spaces prove to be similar enough, there may be room to standardize a format across both.

5.2. New Approaches to Congestion Control

One recent informal survey presented at the workshop found that ISPs perceive traffic volumes from heavy users to be a problem, but no single congestion management tool has been put to wide use. Within developer and research communities, congestion issues raised by increased P2P traffic volumes have spurred new thinking about congestion control mechanisms at both the transport layer and the application layer. The subsections below explore some of these new ideas and highlight areas where IETF work may be appropriate.

5.2.1. End-to-End Congestion Control

As noted previously, uptake in P2P usage can result in perceptible end-user latency on the order of seconds for interactive applications. One approach to resolving this "RTT in seconds" problem would be for P2P clients to implement better congestion control that keeps the bottleneck full while yielding to keep the delay of competing traffic low. Such an algorithm has been implemented in BitTorrent's client by continuously sampling one-way delay (separating propagation from queuing delay) and targeting a small queuing delay value. This essentially approximates a scavenger

service class in an end-to-end congestion control mechanism by by forcing bulk, elastic traffic to yield to competitors under congestion.

In a similar vein, the P4P framework supports a component that allows applications to mark traffic as "bulk data" (not time sensitive). Applications adjust their behavior according to the feedback they receive from such markings.

Experimenting with the standardization of these kinds of techniques or any congestion control framework with design goals that differ from those of TCP may be helpful work for the IETF to pursue.

5.2.2. Weighted Congestion Control

Congestion control has typically been implemented at a protocol level, as a optional, cooperative effort between endpoints experiencing congestion, but in looking for a long-term approach to congestion control, we may need a more rigorous way for available bandwidth to be allocated by and between the hosts using a network. The idea behind weighted congestion control is to allow hosts to give more weight to interactive applications during times of congestion.

Comparing such an approach with DiffServ showcases its strengths and weaknesses. Unlike DiffServ, weighted congestion control could be implemented on hosts with a simple extension to socket APIs (although consensus among OSes would be necessary for portability). Control resides with the host, whereas even when DiffServ APIs are available, it is difficult for a host to know that the network is complying with its classifications. With weighted congestion control, hosts need some disincentive to setting their weights at maximum levels, whereas DiffServ was not designed for individual users to employ. Both approaches must rely on traffic senders to set policies, meaning that the congestion issues stemming from P2P use on the receiver side are not aided by either mechanism. With DiffServ, a light user may waste his or her priority connecting to a heavy user on another network, which is not a problem with host-controlled weighting.

Weighted congestion control is just one example of a generalized set of features that characterize useful approaches to congestion control. These characteristics include full user control of priorities within a user's own scope, and no possibility of interpreting ISP behavior as discriminatory. The former means that ISPs should not override user decisions arbitrarily (though this

does
   not preclude an ISP from offering prioritization as an option). The
   latter means that the metric for decision-making needs to obviate
   suspicion of ISP motivations.

One metric that meets these criteria is a harm (cost) metric, where cost is equal to the amount of data that was not served to its destination. Using this metric, cost is greatest when traffic peaks are greatest. It allows for a policy of not sending too much data during times of congestion, without specifying exactly how much is too much. The cost metric could be used either for a DiffServ approach or for weighted congestion control.

One important limitation on ISPs from a congestion control perspective is that they do not have a window into congestion on other ISPs' networks. Solving this problem requires a separate mechanism to express congestion across domains.

One potential avenue for the IETF or IRTF to pursue would be to establish a long-term design team to assess congestion problems in general and the long-term effects of any proposed quick fixes. These issues are not necessarily confined to P2P and should be viewed in the broader context of massive increases in bandwidth use.

5.3. Quality of Service

Although ISPs have implemented a wide variety of short-term approaches to dealing with congestion, several of these may not be viable in the long term. For example, some ISPs have found that using deep packet inspection to change the delivery characteristics of certain traffic at times of congestion is more cost effective than adding additional bandwidth. Over time, this approach could lead to a cat-and-mouse game where applications providers continually adapt to avoid being correctly classified by DPI equipment. Similarly, ISPs implementing traffic analysis to identify P2P traffic may find that in the long run the overhead required of an effective classification scheme will be excessive.

Quality of service (QoS) arrangements may be more suitable in the long term. One approach that distinguishes certain classes of traffic during times of congestion was described in Section 3.3. A standardized mechanism that may be useful for implementing QoS is DiffServ Code Points (DSCP) [RFC2474].

With DSCP, devices at the edge of the network mark packets with the service level they should receive. Nodes within the network do not need to remember anything about service flows, and applications do not need to request a particular service level. Users effectively avoid self-interference through service classification.

Although DSCP may have many uses, perhaps the most relevant to the P2P congestion issue is its ability to facilitate usage-based charging. User pricing agreements that charge a premium for real-

time traffic and best effort traffic could potentially shape user behavior, resulting in reduced congestion (although ISPs would need a mechanism to mitigate the risk of charging subscribers for things like unintentional malware downloads or DoS attacks). DSCP could also be used to limit a user's supply of high-priority bandwidth, resulting in a similar effect.

Equipment to support DSCP is already available. Although there has been some concern about a perceived lack of DSCP deployment, it is widely used by enterprise customers, and growth has been strong due to uptake in VoIP at the enterprise level.

However, DSCP still faces deployment hurdles on many networks. Perhaps the largest barrier of all to wide deployment is the lack of uniform code points to be used across networks. For example, the latest Windows Vista API marks voice traffic as CS7, above the priority reserve for router traffic. To properly take advantage of this change, every switch will need to re-mark all traffic. In addition, disparate ISPs are currently without a means of verifying each others' markings, and thus may be unwilling to trust the markings they receive.

6. Applications Opening Multiple TCP Connections

The workshop discussions about P2P congestion spurred a related discussion about applications (P2P or otherwise) that open multiple TCP connections. With multiple users sharing one link, TCP flow fairness gives users with multiple open connections a larger proportion of bandwidth. Since some P2P protocols use multiple open connections for a single file transfer, and users often pursue multiple transfers at once, this can cause a P2P user to have many more open connections at once than other users on the same link. The same is true for users of other applications that open multiple connections. A single user with multiple open connections is not necessarily a problem on its face, but the fact that fairness is determined per flow rather than per user leaves that impression. Workshop participants thought it may be useful for the IETF to provide some information about such situations.

7. Costs and Congestion

Workshop participants expressed diverging opinions about how much the cost of transferring data -- as experienced by ISPs, and by extension, by their subscribers -- should factor into IETF thinking on P2P traffic issues.

On one hand, bandwidth costs may be significant, even when viewed in isolation from congestion issues. Some estimates put the total cost of shipping 1 GB between $0.10 and $2. The cost of transit bandwidth in markets where subscribers are charged flat rates appears to have leveled off and may no longer be getting cheaper. Thus, it may be reasonable to expect more service providers to move to volume-based pricing (where they have not already done so) as a means to control congestion and increase revenues. This is only feasible if bandwidth consumption is visible to end users, which argues for some mechanism of exposing quotas and usage to applications. However, expressing cost information may be outside of the technical purview of the IETF.

On the other hand, congestion can be viewed merely as a manifestation of cost. An ISP that invests in capacity could be considered to be paying to relieve congestion. Or, if subscribers are charged for congesting the network, then cost and congestion could be viewed as one and the same. The distinction between them may thus be artificial.

Workshop participants felt that the issues highlighted here may be useful fodder for IRTF work.

8. Next Steps

The IETF community recognizes the significance of both growing P2P traffic volumes and increased network load at large. The importance of addressing the impact of high-volume, delay-tolerant data transfer on end user experiences was highly apparent at the workshop.

At the conclusion of the workshop and in the days following, it became clear that the largest areas of interest fell into two categories: transport-related issues and improved peer selection.

8.1. Transport Issues

Two main transport-related work items evolved out of the workshop. The first was the creation of a standardized delay-based end-to-end congestion control mechanism that applications such as P2P clients could use to reduce their own impact on interactive applications in use on shared links (as described in Section 5.2.1). The second was an informational document that describes the current practice of P2P applications opening multiple transport connections and makes recommendations about the best practices for doing so (as discussed in Section 6).

8.2. Improved Peer Selection

Participants expressed strong interest in further pursuing the range of concepts described in Section 5.1 that support mechanisms for information sharing between networks and applications to help improve peer selection. Adding to the appeal of this topic is its potential utility for applications other than P2P that may also benefit from information about the network. Because the scope of potential solutions discussed at the workshop was broad, extracting out the most feasible pieces to pursue is the necessary first step.

9. Security Considerations

The workshop discussions covered a range of potential engineering activities, each with its own security considerations. For example, if networks are to provide preference or topology information to applications, the applications may desire some means of verifying the authenticity of the information. As the IETF community begins to pursue specific avenues arising out of this workshop, addressing relevant security requirements will be crucial.

10. Acknowledgements

The IETF would like to thank MIT, which hosted the workshop, and all those people at MIT and elsewhere who assisted with the organization and logistics of the workshop.

The IETF is grateful to the program committee (listed in Appendix A) for their time and energy in organizing the workshop, reviewing the position papers, and crafting an event of value for all participants.
The IETF would also like to thank the scribes, Spencer Dawkins and Alissa Cooper, who diligently recorded the proceedings during the workshop.

A special thanks to all the participants in the workshop (listed in Appendix B), who took the time, came to the workshop to participate in the discussions, and who put in the effort to make this workshop a success. The IETF especially appreciates the effort of those that prepared and made presentations at the workshop.

11. Informative References

[DOCSIS] CableLabs, "DOCSIS Specifications - DOCSIS 2.0 Interface", 2008, <http://www.cablemodem.com/specifications/

specifications20.html>.

[P4P] Xie, H., Yang, Y., Krishnamurthy, A., and A. Silberschatz, "P4P: Provider Portal for Applications", August 2008, <htt p://uwnews.org/relatedcontent/2008/August/ rc_parentID43281_thisID43282.pdf>.

[RFC2475] Carlson, M., Weiss, W., Blake, S., Wang, Z., Black, D., and E. Davies, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2475, December 1998.

Appendix A. Program Committee

Dave Clark, MIT
Lars Eggert, TSV AD
Cullen Jennings, RAI AD
John Morris, Center for Democracy and Technology
Jon Peterson, RAI AD
Danny Weitzner, MIT

Appendix B. Workshop Participants

Vinay Aggarwal, Deutsche Telekom Labs, TU Berlin
Marvin Ammori, Free Press
Loa Andersson, Acreo AB
Jari Arkko, Ericsson
Alan Arolovitch, PeerApp
Timothy Balcer
Mary Barnes, Nortel
Colby Barth, Cisco Systems
John Barlett, NetForecast
Salman Baset, Columbia University
Chris Bastian, Comcast
Matthew Bell, Charter Communications
Donald Bowman, Sandvine Inc.
Scott Bradner, Harvard University
Bob Briscoe, British Telecom
David Bryan, SIPeerior Technologies
Rex Bullinger, National Cable & Telecommunications Association

Gonzalo Camarillo, Ericsson
Mary-Luc Champel, Thomson
William Check, NCTA
Alissa Cooper, Center for Democracy and Technology
Patrick Crowley, Washington University
Leslie Daigle, Internet Society
Spencer Dawkins
John Dickinson, Bright House Networks
Lisa Dusseault, CommerceNet
Lars Eggert, Nokia Research Center
Joe Godas, Cablevision
Vernon Groves, Microsoft
Daniel Grunberg, Immedia Semiconductor
Carmen Guerrero, University Carlos III Madrid
Vijay Gurbani, Bell Laboratories/Alcatel-Lucent
William Hawkins III, ITT
Volker Hilt, Bell Labs, Alcatel-Lucent
Russell Housley, Vigil Security, LLC
Robert Jackson, Camiant
Cullen Jennings, Cisco Systems
Paul Jessop, RIAA
XingFeng Jiang, Huawei
Michael Kelsen, Time Warner Cable
Tom Klieber, Comcast
Eric Klinker, BitTorrent Inc.
Umesh Krishnaswamy
Gregory Lebovitz, Juniper
Erran Li, Bell-Labs
Jason Livingood, Comcast
Andrew Malis, Verizon
Enrico Marocco, Telecom Italia Lab
Marcin Matuszewski, Nokia
Danny McPherson, Arbor Networks, Inc.
Michael Merritt, AT&T
Lyle Moore, Bell Canada
John Morris, Center for Democracy and Technology
Jean-Francois Mule, Cablelabs
David Oran, Cisco Systems
Reinaldo Penno, Juniper Networks
Jon Peterson, NeuStar
Howard Pfeffer, Time Warner Cable
Laird Popkin, Pando Networks
Stefano Previdi, Cisco systems
Satish Putta

Eric Pescorla
Benny Rodrig, Avaya
Damien Saucez, UCLouvain (UCL)
Henning Schulzrinne, Columbia University
Michael Sheehan, Juniper Networks
Don Shulzrinne, Immedia Semiconductor
David Sohn, Center for Democracy and Technology
Martin Stiemerling, NEC
Clint Summers, Cox Communications
Robert Topolski
Mark Townsley, Cisco Systems
Yushun Wang, Microsoft
Hao Wang, Yale University
Ye Wang, Yale University
David Ward, Cisco
Nicholas Weaver, ICSI
Daniel Weitzner, MIT
Magnus Westerlund, Ericsson
Thomas Woo, Bell Labs
Steve Worona, EDUCAUSE
Richard Woundy, Comcast
Haiyong Xie
Richard Yang, Yale University

Appendix C. Workshop Agenda

1. Welcome/Note Well/Intro Slides
2. Service Provider Perspective (Comcast)
3. Application Designer Perspective (BitTorrent)
4. Lightning Talks & General Discussion
5. Localization and Caches
6. New Approaches to Congestion
7. Quality of Service
8. Conclusions & Wrap-Up

Appendix D. Slides and Position Papers

Slides and position papers are available at: http://trac.tools.ietf.org/area/rai/trac/wiki/PeerToPeerInfrastructure

Authors' Addresses

Jon Peterson
NeuStar
USA

Email: jon.peterson@neustar.biz

Alissa Cooper
Center for Democracy & Technology
1634 Eye St. NW, Suite 1100
Washington, DC 20006
USA

Email: acooper@cdt.org