Path Precedence Discovery   December 1996

Geoff Huston
Telstra
gih@telstra.net

Marshall T. Rose
Dover Beach Consulting, Inc.
mrose@dbc.mtview.ca.us

Abstract

This memo describes a technique for dynamically discovering the maximum precedence (MPrec) of an arbitrary internet path. It specifies a small change to the way that routers generate one type of ICMP messages. For a path that passes through a router which does not implement this change, this technique may not necessarily discover the correct Path MPrec. In this case, an application which desires to make use of a non-zero precedence should degrade gracefully.

Table of Contents

1. (A Lengthy) Introduction

Today's Internet is composed of many Internet Access Providers (IAPs), connected in a disorganized graph. By and large, each IAP provisions its own network independently of its neighbor IAPs. This leads to an Internet mesh of widely-varying characteristics.

When an Internet consumer (henceforth "consumer") negotiates with an IAP to provision service, the consumer enters into an agreement which specifies a expected quality-of-service (eQoS) enjoyed by packets which travel between the consumer's point of attachment with the IAP (henceforth "consumer attachment") and anywhere throughout the IAP's network. eQoS includes, but is not limited, to characteristics such as throughput, latency, loss, availability and so on. Accordingly, if two consumers subscribe to the same IAP, then they may determine eQoS for the packets they exchange based on the "intersection" of their service agreement parameters with that IAP. The eQoS is, of course, determined in an out-of-band fashion. For example, two consumers may choose to agree to subscribe to the same IAP for an identical level of service.

However, if two consumers subscribe to different IAPs, then it becomes more difficult to determine the eQoS for the packets they exchange. Simply put, there is no mechanism for communicating the eQoS exhibited by the two IAP networks when they exchange traffic. Further, if the two IAP networks do not share an interconnection point, but rather rely on the services of one or more transit networks, then there is no a priori predictability to the eQoS for the packets exchanged by the two consumers.

A fully-general solution to this problem is far beyond the scope of this memo. However, any problem can be solved...if the problem can be made small enough!

In this context, observe that what is lacking in today's Internet is a simple mechanism whereby IAPs may provision service which differentiates traffic based on precedence (that is, packets with a higher precedence value receive preferential treatment). Although RFC 791, the "Internet Protocol" document, provides for a three-bit precedence field, the operational Internet lacks an mechanism whereby IAPs can treat traffic preferentially. The reason, of course, is that there is no mechanism available which allow a consumer to expect that asking for a higher precedence results in preferential behavior. As a consequence, there is no incentive for IAPs to provide preferential behavior on a cost-differential basis.

To solve this much smaller problem, IAPs must take two actions..lp First, the IAP's routers must be configured to preferentially handle packets based on their precedence. There are three aspects to this:

- precedence-ordered queue service (c.f., Section 5.3.3.1 of RFC 1812, the "Requirements for IP Version 4 Routers" document), which (among other things) causes a router to order the forwarding process and output interface queues based on highest precedence;

- precedence-based congestion control (c.f., Section 5.3.6 of RFC 1812), which causes a router to drop packets based on lowest precedence; and,

- link layer priority features (c.f., Section 5.3.3.2 of RFC 1812), which causes a router to select service levels of the lower layers to provide preferential treatment.

Second, each router on the IAP's side of the consumer/IAP attachment must discard packets higher than the maximum precedence (MPrec) for that consumer site, and return an ICMP destination unreachable message with a code indicating the consumer's MPrec. Naturally, one would expect that the MPrec for the consumer site would be a new variable added into the service agreement between the IAP and the consumer. The default MPrec, of course, is zero, which is the common practice in today's Internet.

If the consumer opts to negotiate for a non-zero MPrec, then it must have an expectation that the packets it sends with non-zero precedence will be honored along the path from its IAP attachment to the destination. Further, the consumer probably also wants an assurance that the return traffic can also enjoy the same level of precedence. This memo describes the protocol used in order probe an internet path with respect to the precedence it supports.

However, before specifying the protocol, it is necessary to discuss some aspects of provisioning a precedence-based facility in an IAP.

2. RSVP Considered Harmful?

In a word, no. Having said that, we must be careful to describe why the solution described in this memo is appropriate for today's Internet environment, and why the RSVP-solution space is premature for today's market.

An Internet network can be considered as a collection of switches and interconnecting bandwidth. When provisioning a network to support a single grade of service, the IAP must deploy adequate switches and bandwidth to accomodate the offered load with an "acceptably low level" of transit loss.

Although one could view an "acceptably low level" as no loss, the TCP flow control algorithm uses packet loss as one threshold signal when searching for a dynamic level of peak transmission throughput. As such, it is perhaps more appropriate to define the acceptable level as one which avoids periods of degenerative congestion-induced protocol collapse.

When the network is placed under load stress the IAP has two options:

- augment the bandwidth and basic switching resources to a level commensurate to the increase in load; or,

- increase the complexity of the switching algorithm being used.

In the latter case, the increased complexity of switching takes the form of algorithms which perform precedence-ordered queue service. By applying these algorithms, network performance is selectively downgraded, allowing a category of network traffic greater access to network resources.

2.1. Investment Economics

Ultimately, the choice between these options is based on investment economics.

If the IAP purchases additional bandwidth outright, the IAP is making a capital investment with a relatively long investment life. Instead, if the IAP leases additional bandwidth, then given adequate supply, the IAP is encountering a recurrent cost, scaling at a rate commensurate with the growth in traffic volume. Both of these activities can be undertaken with reasonable financial certainty given basic soundness in the provider's business structure.

Alternatively, if the IAP increases the level of switching complexity, the IAP is encountering a capital cost with a relatively short investment life cycle. Further, as the same switching algorithm must be applied across higher traffic levels within a constant timeframe, this capital cost increases as a function of transmission speed.

To reduce the impact of the scaling of the capital cost of switching complexity, network state models, such as RSVP, were developed to scaling the switching complexity in the face of increasing traffic levels.

2.2. The Cost of Switching and Bandwidth

Since increasing the complexity of switching does not increase the absolute level of traffic carried by the network, the IAP must apply differential charges to high precedence traffic, in order to generate a financial return on the investment in more complex switching systems. In contrast, increasing bandwidth allows greater traffic volumes, which result into an increased revenue stream, which offsets the cost of the additional bandwidth.

As such, the costs of additional switching complexity and bandwidth must be measured against the difference in revenue streams.

In general, switching complexity is less expensive than bandwidth only when bandwidth is very expensive, e.g., in international traffic circuits, and in the context of an under-developed communications infrastructure. Although significant, these environments are not dominant parts of the Internet infrastructure. As a result, the end-to-end environment is heterogenous, wherein:

- some providers will continue to offer a single grade of service, using augmentation of bandwidth and single service level switching complexity to service their traffic load; whilst

- other providers will adopt service class structures as a means of management of the congestion impact of imposed traffic.

2.3. QoS and Critical Mass

From the consumer's perspective, service quality results from end-to-end behavior. Contracting for a particular eQoS is irrelevant unless that service is supported across the path taken by the traffic. Indeed, without a level of transitive bilateral, or multilateral, agreements on policy for precedence there is no economic motivation for an Internet Transit Provider (ITP) to honor a precedence request.

This leads to a situation where the investment in switching complexity as a unilateral decision by one IAP yields no visible enhancement in end-to-end QoS, even when both end subscribers have subscribed to an enhanced service. For precedence to function in a useful fashion across a multi-provider Internet there is a requirement for a critical number of IAPs and ITPs to adhere to a common structure for honoring precedence requests.

We claim, in an heterogenous environment, that it is not uniformly economically attractive for network providers to multilaterally subscribe to the implementation of a definition of an end-to-end network state, e.g., RSVP, to support defined QoS measures. Instead, we claim that it is viable for network providers to:

- elect to honor a common semantic structure which allows consumers to make precedence requests; and,

- to give the consumer the capability to probe for precedence capability for the path taken by its traffic.

This approach allows a graduated imposition of QoS across the Internet, and allows the consumer the option to precedence request in those situations where no appreciable benefit is derived.

3. The Path Precedence (PPrec) Discovery Protocol

In brief:

- A consumer's host sends a packet with the desired precedence.

- If any of the routers along the packet's path are configured to administratively disallow sending packets with that precedence, the router discards the packet and returns an ICMP Destination Unreachable message with a (new) code meaning "precedence not allowed".

- Upon receipt of such a "precedence not allowed" message, the host takes reacts based on the requirements of its application. For example, an application might view a particular precedence as a mandatory requirement, and opt not to communicate at that time. Alternatively, it may direct the host to send packets with a lesser precedence.

- The PPrec discovery process ends when the host's estimate of the PPrec is low enough that its packets can be delivered without being discarded for administrative reasons.

- Changes in the routing topology may reduce the PPrec, resulting in packets being discarded and "precedence not allowed" messages being returned. A host should react accordingly. Similarly, changes in the routing topology may (silently) increase the PPrec. To probe for this, a host may infrequently generate packets with a higher precedence, providing such actions are not destructive to the host application (i.e., receiving a subsequent "precedence not allowed" message will not, by itself, abort a TCP connection).

3.1. Host specification

When a host receives a "precedence not allowed" message, it MUST reduce its estimate of the MPrec for the relevant path, based on the value of the Next-Hop Maximum Precedence field in the message (c.f., Section

3.2). No further specification is placed upon the hosts behavior, as different applications may have different requirements, and since different implementation architectures may favor different strategies.

It is required that after receiving a "precedence not allowed" message, a host MUST attempt to avoid eliciting more such messages in the near future, by reducing the precedence of the packets that it is sending along the path.

Hosts using PPrec Discovery MUST detect decreases in MPrec as fast as possible. Hosts MAY detect increases in PPrec, but because doing so requires sending packets larger than the current estimated PPrec, and because the likelihood is that the PPrec will not have increased, this MUST be done at infrequent intervals. An attempt to detect an increase (by sending a packet larger than the current estimate) MUST NOT be done less than 5 minutes after a "precedence not allowed" message has been received for the given destination, or less than 1 minute after a previous, successful attempted increase. It is RECOMMENDED that these timers be set at twice their minimum values (10 minutes and 2 minutes, respectively).

A host MUST not increase its estimate of the PPrec in response to the contents of a "precedence not allowed" message. A message purporting to announce an increase in the PPrec might be a stale packet that has been floating around in the Internet, a false packet injected as part of a denial-of-service attack, or the result of having multiple paths to the destination.

3.2. Router specification

When a router is administratively configured to discard a packet because it exceeds the precedence allowed for the packet's source, the router MUST return an ICMP Destination Unreable message to the source of the packet, with the code indicating "precedence not allowed" [[value TBD]]. The router MUST include the maximum precedence allowed to the packet's source in the Next-hop Maximum Precedence (NH-MPrec) field:

```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |   Type = 3    |  Code = TBD   |           Checksum            |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |           unused = 0                          |    NH-MPrec   |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |      Internet Header + 64 bits of Original Datagram Data      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
   The value carried in the 8-bit Next-Hop MPrec field is:
      Bits 0-2:  Maximum precedence allowed over the next-hop for the
                 packet's source.
      Bit  3-7:  Reserved for Future Use.
```

3.3. Determining if the Discovery Algorithm is Available

RFC 1812 makes no mention of administrative controls of precedence-based routing, other than to say that that there must be a way to disable such mechanisms. As such, there is no "standard" administrative behavior for today's routers when they encounter packets with a non-zero precedence field.

It is RECOMMENDED that routers be administratively configured to always generate a "precedence not allowed" message when receiving a packet with a precedence value of 7 (all-ones). This allows a sophisticated host to send to probe for the existence of the PPrec Discovery algorithm by sending packets to the destination with this all-ones value. (Of course, this doesn't guarantee that algorithm is available at all routers along the path, but it does provide a good initial estimate.)

4. Implementation Issues

The issues in handing PPrec Discovery are similar to those associated with PMTU Discovery. Accordingly, the reader is directed to Section 6 of RFC 1191.

5. Security considerations

This PPrec Discovery mechanism makes possible a denial-of-service attack, in which a third-party sends a false "precedence not allowed" message indicates a Next-hop MPrec much smaller than reality. This may cause an application which requires a higher PPrec to cease its efforts to communicate.

A third-party party could also cause problems if it could stop a host from receiving legitimate "precedence not allowed" messages, but in this case there are simpler denial-of-service attacks available.

6. Acknowledgements

This proposal is inspired by RFC 1191, the "Path MTU Discovery" for IPv4 document. All good ideas contained herein have been borrowed freely from other sources, whilst all bad ideas contained herein are wholly new.

7. Authors' Address

Geoff Huston
Telstra
5/490 Northbourne Ave
Dickson ACT 2609
UA

Tel: +1 61 208 1908
Fax: +1 61 248 6165
E-Mail: gih@telstra.net

Marshall T. Rose
Dover Beach Consulting, Inc.
11975 El Camino Real
Suite 200
San Diego, CA 92130
US

Tel: +1 619 793 2700
Fax: +1 619 793 2950
E-Mail: mrose@dbc.mtview.ca.us