Sally Floyd/ACIRI
Steve Bellovin/AT&T
John Ioannidis/AT&T
Kireeti Kompella/Juniper
Ratul Mahajan/UW
Vern Paxson/ACIRI

July, 2001
Expires: January, 2002

Pushback Messages for Controlling Aggregates in the Network

Status of this Memo

Abstract

1. Introduction

Pushback [MB01] is designed to detect and control high bandwidth aggregates in the network. An aggregate is a collection of packets with a common property. For instance, with the destination prefix as the common property, all packets with a matching prefix define an aggregate. During a time of severe congestion from a flash crowd or from a denial of service (DoS) attack, a router might enforce a rate-limit on the traffic aggregate responsible for the congestion. In addition, the congested router could ask adjacent upstream routers to limit the amount of traffic they send for that aggregate. This upstream rate-limiting is called pushback and can be recursively propagated to routers further upstream. It serves to spatially isolate the traffic aggregate so that other traffic sharing the same downstream links is not impaired by the aggregate.

By imposing only a rate limit, rather than a complete blockage, of the aggregate, pushback aims to minimize "collateral damage" suffered by the non-hostile traffic matching the aggregate during a DoS attack. In general, the hope is that during a DoS attack, pushback will propagate sufficiently far in the network so that non-hostile traffic fits within the rate limit imposed on its specific path to the destination, and accordingly does not have its performance limited.

This document specifies messages passed between cooperating routers. It does not address procedures in routers for identifying aggregates to be rate-limited, or for determining the rate-limits for those identified aggregates. The goal is to specify an experimental standard for pushback messages so that we can learn from the experimental use of pushback. We expect that the specifications for pushback messages will evolve over time, as we gain more experience with their use.

There are two main pushback messages - REQUEST and STATUS. Pushback REQUEST messages are sent to upstream routers asking them to rate-limit the aggregate. Such a request for rate-limiting is only advisory; the upstream router is not compelled to follow the request. As part of rate-limiting on behalf of the downstream router, the upstream router sends periodic STATUS messages to the downstream router. The STATUS messages report the arrival rate for that aggregate at the upstream router, and enable the congested router to take decisions regarding the continuance of pushback. In addition to REQUEST and STATUS messages, REFRESH messages reinforce the soft-state rate-limiting, and CANCEL messages terminate it.

Pushback messages can be used in two ways. In one pushback type, pushback messages are used to request upstream rate-limiting for the specified aggregate. In a second pushback type (DUMMY_PROP), pushback messages are used simply to get information about the arrival rates of an aggregate at upstream routers.

A pushback in progress can be visualized as a tree (or, with multipath routing, possibly a graph), where the congested router initiating the pushback is the root. The parent of a router in the tree is the downstream router from which it got the pushback REQUEST. Routers that do not propagate pushback further are leaves of the tree.

The following sections specify the format for pushback messages and the timing of REFRESH and STATUS messages. This document also specifies the procedures for propagating pushback REQUEST and REFRESH messages further upstream, and for merging the resulting STATUS messages from upstream routers.

2. The Common Header

All pushback messages have the following fields prepended in the header.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Version  |AdF|    Msg Type   |   Rate-Limiting Session ID    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Pushback Initiating Router's IP Address            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Sender's IP Address                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The first field specifies the version of the pushback protocol the sender speaks. The protocol described in this document is Version=0.

"AdF" specifies the type of address family used. Currently defined values are IPv4=0 and IPv6=1. Other values are reserved for future definition.

The message type is one of REQUEST (= 0), REFRESH (= 1), STATUS (= 2), or CANCEL (= 3). The fields following the common header are dependent on the type of the message.

The Rate-limiting Session ID (RLSID) is generated by the congested router initiating pushback. It MUST be unique among all current rate-limiting sessions initiated by this router. The RLSID combined with the IP address of the congested router defines a pushback session over the whole network. A router receives both these fields from the downstream router requesting pushback. These fields enable the routers to map incoming messages to the appropriate rate-limiting session. A router MAY use its different addresses when initiating different pushback sessions.

Note that if the router's address reflects a private addressing realm, then it MUST be altered upon crossing into a different addressing realm. Ideally this transformation uses a new address unique to the router; if not available, then the address of the router propagating pushback (by sending the message) into the different realm is used.

The sender's IP address has been included in the pushback message, making message interpretation independent of the IP source address field. This eliminates any confusion regarding which interface's address is included in that field (that is, whether it is the IP address of the message-sending interface, or of some other interface at the router). It also helps when pushback messages traverse between routers that are not directly connected. If a sender sends pushback messages to two peers in two different addressing realms, so that the sender doesn't have a unique address to send to both peers, then the sender will use different values for the sender's IP address in the two messages.

Both the IP addresses will be 128 bit fields for IPv6.

2. The Pushback REQUEST Message

Pushback requests (type REQUEST=0) are sent upstream when a router wants the aggregate to be rate-limited upstream. The fields in a pushback REQUEST, in addition to the common header, are shown below.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PType |SRMode |   Max Depth   | Depth in Tree |   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Bandwidth Limit                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Expiration Time                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Status Frequency                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                    Congestion Signature                       |
         ..............................................
         ..............................................
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

"PType" denotes the type of pushback and determines the upstream router's behavior in various ways. PType=0 (HI_DROP_PROP) requests the upstream router to propagate pushback if the restricted aggregate suffers a high drop rate due to the restriction. (This is the usual mode of pushback.)

PType=1 (ALWAYS_PROP) requests that the upstream router propagate pushback irrespective of the drop rate experienced by the aggregate. It would typically be used when the aggregate is known with high confidence to be malicious.

PType=2 (DUMMY_PROP) indicates no actual rate-limiting should take place---the downstream router is just interested in the arrival rate estimate of this aggregate. The extent of propagation of these pushback messages is controlled by the congested router using "Max Depth" (explained below) to determine the arrival rate of an aggregate several hops upstream.

Other values of PType are reserved for future definition.

The pushback requester can specify the mode in which it wants feedback with the "SRMode" (Status Reporting Mode) field. SRMode=0 (COMPACT) specifies the feedback should be just the total arrival-rate estimate of the aggregate. SRMode=1 (CLOSEST) specifies the feedback should include per-router feedback for upstream routers, and if there is not room for all of them, then those closest (lowest hop count) should be preferred. SRMode=2 (FURTHEST) is similar to CLOSEST, but prefers routers further away from the congested router. SRMode=3 (SAMPLE) specifies that instead a pseudo-random subset of the upstream routers should be included. These last two modes facilitate different forms of mapping to aid with tracing back the attack sources. Other values are reserved for future definition.

Using "Max Depth" the congested router can control the maximum number of hops pushback will propagate to. The special value of 255 indicates unrestricted propagation; 254 indicates that pushback should be propagated up to, but not across, the AS boundary.

The depth in the tree is the distance of the requesting node from the root of the pushback tree. The depth of the root is zero, and a child's depth is one more than the depth of its parent. Depth information is useful in setting timers for sending feedback. If a router receives overlapping pushback requests from multiple peers, its depth is one more than the minimum depth of its parents.

The bandwidth limit is a single precision (32-bit) floating point number in IEEE format, as described in [SPG97]. It expresses the rate in bytes per second. It is only a requested upper bound for the bandwidth to be given to the aggregate. If congested, the upstream router could send less than that upper bound. In addition, the upstream router is not *required* to observe the requested bound.

The expiration time is the time period after which the pushback request expires if no REFRESH messages arrive. The status frequency gives the time after which the upstream routers should send STATUS messages downstream. Both the times are specified in milliseconds, and represented using 32-bit integers in network byte order.

2.0.1 The Congestion Signature

The congestion signature is the description of the aggregate that is to be rate-limited. Its specification is a major open question. As an example, the attack signature might consist of the destination prefix(es) characterizing the aggregate. At the other extreme, we could allow arbitrary ACLs of fields in the packet header. For this first experimental use of pushback, we will limit the congestion signature to depend only on the source and destination IP addresses in the packet header. This excludes the use of aggregates defined by the protocol field in the IP header and the port numbers in the transport header; an example is an aggregate consisting of all DNS traffic.

Congestion signatures are specified as TLVs (type-length-value):

```
 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      |     Length    |          Value ....            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             ...............................................
             .....................Value.....................
             ...............................................
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |              .... Value and final Padding ....                |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The Length field includes the Type and Length fields, as well as the length of the Value and any Padding.

Values are padded up to 32-bit alignment. If, after doing so, more data remains in the datagram, then it's interpreted as another TLV.

Type=0 (SRC_PREFIX) indicates a source address prefix. Its Length is 4 bytes plus the length of an address in the format specified by AdF above. The first octet of Value (bits 16 through 23 in the first line of the above figure) are reserved. The second octet gives the prefix length, in the range 1-32 (IPv4) or 1-128 (IPv6). Other values are reserved.

Type=1 (DST_PREFIX) is the same but for destination address prefix.

When all prefixes in the list are of the same type, the congestion signature describes packets that have the corresponding field (source or destination) matching one of the prefixes in the list. In presence of both source and destination prefixes, packets belonging to the aggregate are those destined for one of the destination prefixes *and* coming from one of the source prefixes.

2.1 Propagating a Pushback REQUEST Message

When propagating a pushback request upstream, the router MUST insert the correct depth information, which is one more than the depth of its parent(s).

In addition, the destination prefixes in the congestion signature MUST be checked to see whether they have to be *narrowed*, to restrict the rate-limiting only to traffic headed for the downstream router that requested pushback, as follows. Suppose the congested router X identifies a certain aggregate A with destination prefix 128.95/16. X will ask its upstream router Y (among others) to rate-limit traffic from aggregate A (128.95/16). However, Y cannot use the same specification directly because while Y could be forwarding 128.95.1/24 to X, it might not be forwarding the rest of 128.95/16 to X. If Y (and routers upstream of Y) started rate-limiting all of 128.95/16, the network would drop traffic which would not have reached the congested router X.

To avoid this unnecessary packet-dropping, it is important that Y look at its routing table to find prefixes within 128.95/16 that are forwarded to X. Y has to check all extensions of the given prefix in the routing table.

The issue of narrowing the congestion signature occurs when a pushback request is propagated upstream by a router (thus becoming a non-leaf in the tree), or when the pushback request is passed from the output interface to an input interface at a router. The above algorithm for narrowing the congestion signature works only for congestion signatures with a destination address component in them. It cannot be applied to other signatures, pure source-based ones, for instance. We do not deal with the issue of narrowing non-destination-based signatures in this document except noting that it can be done given the right routing information at the upstream router.

A router could receive requests from different downstream routers with overlapping congestion signatures. Future work might address the possibility of merging two different rate-limiting sessions in this case.

2.2 Pushback REFRESH Messages

Pushback REFRESH messages are initiated by the congested router that started the pushback, if it wants the pushback to continue. For uninterrupted rate-limiting, these messages should be generated before the rate-limiting expires at the upstream routers

The REFRESH message is identical to the REQUEST message, so that if the upstream router has crashed in the meanwhile, the state can be reestablished. However, the message type is set to REFRESH so that, if state already exists, it is matched against the RLSID and router address fields so that the receiving router does not have to go through the process of setting up state from scratch.

REFRESH messages can change any field specified earlier in the pushback REQUEST. On receiving the pushback REFRESH message the upstream routers update the expiration time for the rate-limit session and the limit imposed on the aggregate, and set the timer for the STATUS message. Non-leaf routers in the pushback tree SHOULD send REFRESH messages further upstream after dividing the rate limit among upstream neighbors. If the aggregate specification has changed, the router MUST check if the new aggregate needs to be narrowed, using the process described above, before propagating the pushback REFRESH.

2.3 Pushback CANCEL Messages

The pushback CANCEL message is sent upstream to stop rate-limiting the aggregate. It SHOULD be propagated upstream by routers that have propagated pushback requests (non-leaf routers in the pushback tree).

The CANCEL message has no fields beyond those present in the common header.

3. Pushback STATUS Messages

Upstream routers that receive a pushback REQUEST send pushback STATUS messages to the router from whom they got their REQUEST. The additional fields in the STATUS message are:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Arrival Rate Estimate                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| SRMode| Rsrvd |     Height    |            NumElem            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          <Router ID>                          |
|                         <Router Info>                         |
|                    <Arrival Rate at Router>                   |
         ...............................................
         ................................................
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The arrival rate estimate is a single precision floating-point number in IEEE format, as described in [SPG97]. It expresses the arrival rate of the aggregate in bytes per second if there was no rate-limiting upstream of the STATUS sender. The arrival rate for the first STATUS message is computed over the interval since the receipt of the pushback REQUEST. For the subsequent messages, it is computed over the interval since the last STATUS message.

The SRMode field specifies the mode in which status is reported. It is the same as that in the pushback REQUEST message. The supported modes and their semantics are described in Section 2.

"Height" denotes the height of the sender in the pushback tree. It is zero for leaf nodes, and one more than the maximum height among children for non-leaf nodes. This field tells the receiver how far pushback has propagated upstream of it.

For status modes that return a list of routers, NumElem gives the number of listed routers. Then, for each router

* Router ID gives an address identifying the router
* Router Info gives information associated with the router. Currently, this consists of:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|S|               Reserved                  |   Depth in Tree   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

S is a bit that if set means that the entry is Stale, meaning that no new STATUS message has been received from the router since the last time the sending router sent a STATUS message.

Depth in Tree reflects the depth of the router in the pushback tree.

3.1 Aggregating STATUS Messages

A parent node uses the STATUS messages from its children to construct the STATUS message it sends downstream. The merger technique used depends on the mode of the STATUS message. In COMPACT mode, the parent node adds the arrival rate estimates received from the children and its estimate from upstream routers that were not sent pushback messages. In other modes, the lists obtained from the various children are merged, and appended with the parent node's own estimate, as indicated in the previous section. The height is one more than the maximum height recieved from the children.

3.2 Timing of STATUS Messages

The status frequency specified in the pushback REQUEST and REFRESH messages is the rate at which the originating router would like to receive status reports. Since the upstream routers are at different distances from the root, the timer values they set have to be different. In particular, routers further away from the root should set smaller timer values because they get messages late and their STATUS messages take time getting to the root node.

On receiving a REQUEST or REFRESH message, the routers set a timer to send the STATUS message. The value of this timer is the status frequency minus the _depth_ * _k_, where _depth_ is the router's depth in the pushback tree, and _k_ is a constant that signifies the maximum round trip time for a message over a pushback tree edge (including message processing time). _k_ should be configured to some comfortable upper bound like 100 ms (it is same for all the routers in the pushback tree). For satellite hops or other links with round-trip times greater than the configured value _k_, the consequences will simply be stale STATUS messages.

Setting timers in this fashion means that parents are likely to obtain fresh STATUS messages from their children before their own STATUS message timer goes off. This in turn means that fresh STATUS messages are sent further downstream after aggregation. If a parent router's timer fires before it has received STATUS message from one of its children, it MUST send its own STATUS message downstream using the last value received from this child or its own estimate, and, if including an individual rate report for this child, marking it with S=1 to indicate it is Stale.

The status timer is set again immediately after sending the STATUS messages downstream. The value of the timer is the same for all the routers in this case, and is equal to the status frequency, since the required offset has already been achieved. If a router receives a REFRESH message before its status timer expires, new timers are set as described above.

A small jitter can be applied to status timers so that the downstream router recieves STATUS messages from its children at different times.

In some cases, the original sender of the pushback REQUEST might want some variation in the status timers to provide some degree of protection against gaming adversaries that try to time their bursts to avoid detection. This variation could be achieved by the original sender by making changes to the Status Frequency specified in the pushback REFRESH messages.

4. Authentication for Pushback Messages

Pushback messages require some form of authentication, even if the pushback messages are between adjacent routers. However, this document currently does not specify the form of authentication to be used.

5. Messages between Routers and Local Agents

Some routers might send packet headers from a sample of the traffic to an agent for outboard processing, and receive control messages back from the agent about identified aggregates to be rate-limited. The router and local agent will also exchange control messages, for example, to control the sampling at the router. The formats for these messages will probably be addressed in a separate document.

Because this is a purely local conversation between a router and an attached local agent, it is not necessary that a router and its attached local agent follow the protocol suggested in that document.

6. Messages exchanged with the NOC

In some cases the NOC (Network Operations Center) will want to have final approval before an aggregate is rate-limited. Thus, one category of pushback messages will be the messages exchanged with the NOC. This draft currently does not specify these messages.

Conclusions

Acknowledgements

There is a list of people who can be either co-authors, or can be acknowledged in this section. So far, this list includes the following. The pushback authors: Ratul Mahajan, Steven M. Bellovin, Sally Floyd, John Ioannidis, Vern Paxson, and Scott Shenker. From Juniper: Kireeti Kompella. From Cisco: Barbara Fraser, David Meyer. Other: Randy Bush.

References

[MB01] Ratul Mahajan, Steven M. Bellovin, Sally Floyd, John Ioannidis, Vern Paxson, and Scott Shenker, Controlling High Bandwidth Aggregates in the Network, February 2001. URL: "http://www.aciri.org/pushback/".

[SPG97] S. Shenker, C. Partridge, R. Guerin. Specification of Guaranteed Quality of Service. RFC 2212. September 1997.

Security Considerations

We will eventually address the potential DoS features and security vulnerabilities of pushback in detail here.

IANA Considerations

AUTHORS' ADDRESSES

Sally Floyd
Phone: +1 510 666 2989
ACIRI
Email: floyd@aciri.org
URL: http://www.aciri.org/floyd/

Steve Bellovin
Phone: +1.973.360.8656
AT&T Labs - Research
Email: smb@research.att.com

John Ioannidis
Phone: +1.973.360.7012
AT&T Labs - Research
Email: ji@research.att.com

Kireeti Kompella
Juniper Networks
1994 N. Mathilda Ave
Sunnyvale, CA 94089
Email: kireeti@juniper.net

Ratul Mahajan
Phone: +1 206 616 1853
Univerity of Washington
Email: ratul@cs.washington.edu
URL: http://www.cs.washington.edu/homes/ratul/

Vern Paxson
Phone: +1 510 666 2882
ACIRI
Email: vern@aciri.org
URL: http://www.aciri.org/vern/

This draft was created in July 2001.
It expires January 2002.