Network Working Group	S. Amante
Internet-Draft	Level 3 Communications, LLC
Intended status: Informational	A. Atlas
Expires: August 23, 2008	BT
	A. Lange
	Alcatel-Lucent
	D. McPherson
	Arbor Networks, Inc.
	February 20, 2008

Operations and Maintenance Next Generation Requirements
draft-amante-oam-ng-requirements-01

Status of this Memo

Abstract

Current IP and MPLS OAM techniques need to be extended to permit operators to effectively diagnose load-balancing issues. Specifically, new ad-hoc OAM techniques are needed to diganose various link-bundling techniques, such as IP/MPLS Equal Cost Multi-Path (ECMP) and Link Aggregation Groups (LAG). In addition, these OAM tools should also be extended to permit performance monitoring over longer time durations.

This document defines requirements for the next generation of OAM solutions.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 (Bradner, S., “Key words for use in RFCs to Indicate Requirement Levels,” March 1997.) [RFC2119].

Table of Contents

---

1. Introduction

TOC

Current networks make extensive use of multiple network paths to create larger virtual links between network elements, in particular when a single physical-layer link has exceeded its carrying capacity and no larger bandwidth physical layer technologies exist. Operators use various link bundling techniques, such as Link Aggregation Groups (LAGs) and IP and MPLS Equal Cost Multi-Path (ECMP), to augment the capacity between network elements when physical link-layer capacity is exhausted. Existing troubleshooting tools, based on 'legacy' ping and traceroute, are insufficient to effectively examine the underlying component-links that traffic will use.

In addition, as more of the world's traffic converges around IP and MPLS based networks, service providers need to extract temporally aware traffic performance information.

This draft is NOT intended to address transport MPLS capabilities. Transport-oriented requirements would be complimentary to the requirements presented here.

1.1. Contributors

TOC

The following made vital contributions to this document:
Rajeev Manur, Force10 Networks, Inc.

---

2. Background

TOC

The use of Link Aggregate Groups (LAG's), Equal Cost Multi-Path (ECMP) or a combination of ECMP over LAG's is a common technique used to bond multiple parallel circuits or paths together to achieve the appearance of a larger aggregate link between two nodes. The advantage of these techniques, in particular LAG's, is a reduced number of routing and signaling protocol adjacencies between devices, reducing control plane processing overhead. A disadvantage of these techniques is an inability to determine the individual component-link used for traffic forwarding inside a LAG or ECMP path, specifically for a given microflow, between two devices using traditional traceroute or ping utilities.

A key problem related to LAG or ECMP paths is, due to inefficiencies in LAG or ECMP load-distribution algorithms, a particular component-link may experience congestion or a soft-failure, which would go unnoticed by NMS systems and, likely, IP/MPLS Control Plane protocols. The end result is performance degradation of a subset of end-user microflows that use the affected component-links between two adjacent devices.

What is needed by operators are the following. First, and the most immediate need, is a capability to determine the set of component-links used by individual network elements through which traceroute or ping messages are traversing. Second, a capability to specify an end-user's microflow, e.g.: a 5-tuple "flow" in the case of IP traffic, that will be used by intermediate devices to calculate the component-link or ECMP path used for that flow to allow periodic or perpetual performance monitoring. Ultimately, these capabilities are necessary to both determine and exercise the actual path that is/was used by an end-user's particular application through the network.

---

3. Use Cases

TOC

3.1. Types of Exercise Mechanisms

TOC

This memo classifies two types of ping and traceroute requests that are needed in modern networks where many inter-node links consist of LAG, ECMP or LAG over ECMP paths. First, a "traditional" or "legacy" traceroute and ping request where intermediate devices only understand how to use outer IP header information as the input to a LAG or ECMP hashing algorithm. This type of mechanism has limited utility insomuch as existing devices, interior to a Service Provider's network, only understand how to process limited information in traceroute or ping requests. Note that when operators originate traceroute and/or ping sessions from within their network, requests are sourced from devices, often routers, whose interfaces reside within their network.
On the other hand, a "next-generation" traceroute and ping request where intermediate devices understand new information likely contained in the payload of the traceroute and ping request, which can then be fed as input to the LAG or ECMP hashing algorithm. This would allow operators to, for example, specify the exact "tuple" used by customer traffic in order to properly exercise the LAG or ECMP paths used by a particular customer 'flow' through the network.

3.2. Scenario 1: Traceroute through Routed Hops

TOC

```
   I1: 10.1.1.1/30               I3: 10.5.1.1/30
+------+                       +------+                     +------+
|      |-- A1 ----------- A2 --|      |-- D1 ---------- D2 --|      |
|  R1  |-- B1 -- LAG-1 -- B2 --|  R2  |         LAG-2        |  R3  |
|      |-- C1 ----------- C2 --|      |-- E1 ---------- E2 --|      |
+------+                       +------+                     +------+
                10.1.1.2/30: I2                  10.5.1.2/30: I4
```

Note on figures: Figures 1 through 3 represent a piece of a network for illustrative purposes. In a real network, other nodes will be present.

Figure 1: Traceroute through Routed Hops

In the above example, the links A1-A2, B1-B2 and C1-C2 are grouped into a single LAG, called LAG-1, between nodes R1 and R2. Furthermore, D1-D2 and E1-E2 are grouped into a single LAG, called LAG-2, between nodes R2 and R3. I1 represents the IPv4 address 10.1.1.1/30 assigned to the LAG-1 interface on R1. I2 represents the IPv4 address 10.1.1.2/30 assigned to the LAG-1 interface on R2. I3 and I4 are the IP interfaces assigned to R2 and R3, respectively, on LAG-2. R1 and R2 will maintain a single set of routing and signaling protocol (e.g.: IS-IS, RSVP and/ or LDP), adjacencies over LAG-1, while R2 and R3 will maintain a single set of routing and signaling protocol adjacencies over LAG-2. Assuming the individual component link sizes between R1, R2 and R3 are 10 Gbps, the end result is that R1 and R2 believe they have a single 30 Gbps connection between them and R2 and R3 believe they have a 20 Gbps connection between them.

When performing a traceroute from R1 through R2 to R3, each router independently and automatically determines, through a proprietary LAG or ECMP load-distribution algorithm, the outgoing component-link inside a LAG or ECMP path to send out traceroute UDP probe packets. Unfortunately, the details of the specific component-links are not exposed to a user interface, which would allow operators to determine the exact physical path used by traceroute. Furthermore, those details cannot also be used as input to a 'ping' utility, (using ICMP echo-request and echo-reply messages [RFC792]), to test longer term performance of a specific physical path through the network. The end result is a network operator may believe that a given path between devices is behaving properly when, in fact, end-user traffic is traversing a different set of component-links and experiencing congestion or other link-layer forwarding problems.

3.3. Scenario 2: Traceroute through One Switched Hop

TOC

```
        I1: 10.1.1.1/30
    +------+                        +--------+                        +------+
    |      |-- A1 ----------- A2 --|        |-- C1 ----------- C2 --|      |
    |  R1  |         LAG-1          |  SW1   |-- D1 -- LAG-2 -- D2 --|  R2  |
    |      |-- B1 ----------- B2 --|        |-- E1 ----------- E2 --|      |
    +------+                        +--------+                        +------+
                                                           10.1.1.2/30: I2
```

Figure 2: Traceroute through One Switched Hop

In this scenario, links A1-A2 and B1-B2 are grouped into a single 20 Gbps LAG, called LAG-1, between nodes R1 and SW1. Furthermore, links C1-C2, D1-D2 and E1-E2 are also joined together into a single 30 Gbps LAG, called LAG-2, between nodes SW1 and R2. I1 represents the IPv4 address 10.1.1.1/30 assigned to the LAG-1 interface on R1. I2 represents the IPv4 address 10.1.1.2/30 assigned to the LAG-2 interface on R2. As in Scenario 1, R1 and R2 will maintain a single set of IP/MPLS routing and signaling protocol adjacencies over the LAG's through SW1.

As in scenario 1, each device along the path R1 to SW1 to R2, (or vice-versa), automatically and independently determines the outgoing component-link inside a LAG or ECMP "bundle" to send out traceroute UDP probe packets. Unfortunately, in this scenario if only the incoming component-link interface ID is displayed to an end-user or network operator, that will not reveal the entire physical path traversed from R1 through SW1 to R2. This scenario highlights the need to also show both the outgoing component-link interface ID on R1 and the incoming component-link interface ID on R2. With both of those pieces of information, and a priori knowledge that there is only one Layer-2 switch between R1 and R3, an operator can rely on a "legacy" traceroute implementation to determine the actual component-links that were used in a traceroute request.

If the operator does not have a priori knowledge that there is a Layer-2 switch between R1 and R2, it would be useful for R1 and R2 to include relevant Layer-2 information, learned from a Link-Layer Discovery Protocol, on both R1 and R3 in the traceroute reply. In this example, R1 would reply with its own outgoing component-link name, SW1's hostname and SW1's incoming component-link name. Furthermore, when R2 sends a traceroute reply it would respond with its own incoming component-link name, SW1's hostname and SW1's outgoing component-link name. This would immediately point out to an operator the presence of one, or more, Layer-2 switches in the middle of a Layer-3 path.

Ultimately, without specific component-link 'neighbor' information, such as from a Link-Layer Discovery Protocol, it will be difficult to rapidly determine the presence or absence of Layer-2 switches in the interior of a Layer-3 path.

It's also important to point out in this particular scenario that, at best, SW1 only understands how to parse information in the outer IP header of a legacy traceroute UDP probe, or other data packets, for input into its LAG hash algorithm, which ultimately determines the outgoing component-link it will use to send packets to R2. It would be highly desirable that SW1 was able to intercept and act upon data fields contained in "next-generation" traceroute and/or ping probe packets, so that operators could specify the actual 5-tuple "flow" to be input into SW1's LAG hash algorithm in order to exercise a specific component-link on SW1 outbound toward R3. If this approach is not used it would likely prevent operators from periodically or continuously exercising a specific set of component-links through a given edge-to-edge path on the network, such as through a proactive network monitoring system, as discussed in Section 4.1 of this document.

3.4. Scenario 3: Traceroute through Two, or More, Switched Hops

TOC

```
    I1: 10.1.1.1/30
+----+            +-----+            +-----+              +----+
|    |-A1-------A2-|     |-C1-------C2-|     |-E1---------E2-|    |
| R1 |    LAG-1    | SW1 |    LAG-2    | SW2 |-F1- LAG-3 -F2-| R2 |
|    |-B1-------B2-|     |-D1-------D2-|     |-G1---------G2-|    |
+----+            +-----+            +-----+              +----+
                                              10.1.1.2/30: I2
```

Figure 3: Traceroute through Two, or More Switched Hops

In this case, two Layer-2 switches are inserted in the path between Layer-3 nodes R1 and R2. LAG-1 and LAG-2 are each grouped together into their own 20 Gbps LAG. Furthermore, LAG-3, between nodes SW2 and R2, is joined together as a single 30 Gbps LAG. Finally, I1 represents the IPv4 address 10.1.1.1/30 assigned to the LAG-1 interface on R1; in addition, I2 denotes the IPv4 address 10.1.1.2/30 assigned to the LAG-2 interface on R2.

This scenario is common in Enterprise or DataCenter environments where R1 may be a router or server, SW1 a top-of-rack distribution switch, SW2 an aggregation switch and, finally, R2, which is a Layer-3 router typically providing WAN connectivity.

This particular case further highlights the need to automatically learn the presence of Layer-2 switches and, ideally, allow one to automatically exercise their LAG hash algorithms to fully qualify the exact set of component-links taken between two Layer-3 devices. In order to learn the presence of Layer-2 switches, it will be necessary for traceroute replies to also include relevant Layer-2 information, such as the next-hop device's hostname and incoming component-link name, from a Link-Layer Discovery Protocol. In the case of "legacy" traceroute, R1 would reply with its outgoing component-link name, plus two pieces of information learned from a Link-Layer Discovery Protocol: SW1's hostname and SW1's incoming component-link name. Furthermore, when the next traceroute UDP probe is sent to R2, it will reply with it's incoming component-link name, SW2's hostname and SW2's outgoing component-link name. Unfortunately, this only yields a partial solution, because it would not reveal the actual component-link used between SW1 and SW2, nor the presence of a third Layer-2 switch between SW1 and SW2. In this instance, an operator would want to use Layer-2 OAM tools in an attempt to identify and diagnose the particular component-link that is used between SW1 and SW2. Unfortunately, Layer-2 OAM tools do not have the ability to identify or troubleshoot component-links in a 802.3ad LAG. In addition, it is time consuming for operators to stop using Layer-2.5 (such as LSP-Ping or LSP-Trace) or Layer-3 ping/traceroute tools, login to R1 and R2 and use Layer-2 OAM tools to resume diagnosing the problem. Furthermore, due to the lack of an integrated toolset, it prevents operators from using an NMS to continuously monitor component-links on paths that go over one or more Layer-2 switches.

Instead, what is needed by operators is integrated Layer-2 and Layer-3 ping/traceroute tools, which allow for rapid and accurate diagnosis and troubleshooting of LAG/ECMP problems. Ultimately, if Layer-2 switches can intercept and act upon "next-generation" traceroute and ping requests, that would enable operators to specify the actual 5-tuple "flow" to be input into each Layer-2 switches' LAG hash algorithm. This would allow operators to periodically or continuously exercise a specific set of component-links over all Layer-2 and Layer-3 devices, all at the same time, along a complete edge-to-edge path through the network, as discussed in Section 4.1 of this document.

It should be noted that the above presumes intermediate Layer-2 switches are capable of intercepting and acting upon NG-OAM probe-requests, which may not be true initially in all environments. Therefore, this document requires all NG-OAM solutions to document how they will determine if intermediate Layer-2 switches are NG-OAM capable and communicating that back to the initiator of an NG-OAM request, in order that operators can tell if the complete path was properly exercised.

3.5. ECMP

TBD

3.6. Proxy Traceroute/Ping Functionality

TOC

To enable more rapid troubleshooting and diagnosis of problems related to LAG, ECMP and/or asymmetric paths in a large-scale network, it is useful to use "proxy" routers/hosts within a network that can initiate a traceroute or ping on behalf of a Network Monitoring System (NMS), such as via [PROXY-LSP-PING]. This is particularly valuable in the following scenarios:

* When troubleshooting problems related to asymmetric paths, it is useful to perform a traceroute and/or ping from a source to the destination as well as from the destination back to the source.
* Some IP/MPLS routers use 'input interface' as input into the LAG and/or ECMP hashing algorithm; therefore, quickly exercising the associated direction of a particular flow through the network is required.
* When narrowing a problem down to specific sequence of links within the network, it is useful to rapidly focus additional testing on suspicious segments, which are a subset of an overall edge-to-edge path.
* Periodic monitoring of a large-scale network composed of a multitude of LAG and/or ECMP paths. In order to divide up the periodic testing of a large set of component-links and paths while simultaneously providing timely results, it is useful to distribute testing out to the IP/MPLS routers in the network on or near the paths to be tested. (See Section 3.6 for more details).

In this scenario, there are three types of devices:
Initiator: The node which creates a proxy traceroute/ping request with: 1) a "5-tuple" to be used as input to a LAG and/or ECMP hashing algorithm; 2) the IP address of the Proxy IP/MPLS router that will initiate the ping/traceroute on behalf of the Initiator; and, 3) the IP address of the destination IP/MPLS router/host that will terminate this ping/traceroute request.
Proxy IP/MPLS Router: The node which receives a proxy traceroute/ping request from an Initiator. Once it has interpreted the proxy request, it initiates a proxy ping/traceroute request from itself toward the destination IP/MPLS router specified in the proxy ping/traceroute request.

Proxy Request Terminator: The node(s) which terminate a proxy traceroute/ping request received from the Proxy IP/MPLS Router. In the case of a proxy traceroute, intermediate nodes along the path to the final destination of proxy traceroute are considered "Intermediate Proxy Request Terminators".

A NG-OAM solution MUST support Proxy Traceroute/Ping Functionality. A NG-OAM solution MUST support replies from the Proxy Request Terminator (or Intermediate Proxy Request Terminators) being sent back to the Proxy IP/MPLS Router, before they are relayed back to the Initiator. The advantage of this approach is that replies should follow a symmetrical path back to the Initiator, which is useful if the NMS is behind a stateful firewall. On the other hand, an NG-OAM solution MAY support replies from the Proxy Request Terminator (or, Intermediate Proxy Request Terminators) directly back to the Initiator. The advantage of this scheme is that it does not rely on the Proxy IP/MPLS Router to cache or relay/reformat Proxy Reply Information, before replying back to the Initiator. This may be useful in situations where it's desirable to reduce the load on the Proxy IP/MPLS Router.

---

4. Performance Monitoring

TOC

4.1. Proactive Network Monitoring and Verification

TOC

There are two forms of Proactive Network Monitoring and Verification (PNMV): Perpetual and Periodic. In a Perpetual PNMV case, the nodes performing monitoring send OAM messages at a specific interval, and record the results on a perpetual basis. In the Periodic case, the messages are sent only on demand of an external system, such as an NMS, or an operator's command. These forms can be implementation cases of the same solution.

Today's solutions, such as ping, traceroute, and simulated user traffic between management nodes, can address the case when there is a single path between two endpoints. However, in large national and international networks, there will exist several routed hops for certain paths through the network. Furthermore, between each pair of IP/MPLS routers there will exist LAG's and/or ECMP paths. Unfortunately at present, Network Monitoring Systems (NMS) are unable to exercise the set of component-links through specific paths on the network. This would allow the NMS to identify and notify a Network Operations Center (NOC) to a soft-failure through one or more component-links on the network. The NOC could then proactively respond to the problem by, for example, quickly taking the affected component-link(s) out-of-service or, alternatively, administratively disabling the link bundle or ECMP path and allowing traffic to switch to another in-service path.

The challenge with monitoring a large set of LAG and/or ECMP paths in a network will be to find the right balance between monitoring all component-links in the network, minimizing the resource utilization (e.g.: CPU, memory, network I/O) on the NMS system(s) while simultaneously having a timely detection interval to allow for proactive notification of problems to the NOC. Therefore, a solution must be devised that allows an NMS to transmit multiple independent, concurrent LAG and/or ECMP path test queries into various points in the network. Within the network, Proxy IP/MPLS Routers will carry out the test queries and report back the test results to the NMS.

A NG-OAM solution SHOULD support the ability to do Proactive Perpetual Network Monitoring and Verification, again through the use of Proxy Traceroute/Ping Functionality described in Section 3.5. It should be noted that Perpetual PNMV may be more resource intensive on devices, which is why that requirement is relaxed compared to Periodic PNMV.

4.1.1. Proactive Periodic Network Monitoring and Verification TOC

Periodic network monitoring is often done in response to a suspected network event, or done as a sampled case of Perpetual network monitoring when Perpetual network monitoring cannot be scaled to the necessary level. Probes sent Periodically are often sent with a shorter inter-message interval, and often request more information than a test that runs on a Perpetual basis.

In order to perform periodic monitoring, the Initiator MUST send the Proxy IP/MPLS Router, the number and interval of the probe requests. For example, the Initiator may send the Proxy IP/MPLS Router a request to run 300 consecutive probes at an interval of 500 msec between probes.

4.1.2. Proactive Perpetual Network Monitoring and Verification TOC

Perpetual network monitoring is done consistently among a subset of end points in the total network. The subset, such as sample PoP router to sample PoP router, is selected to strike a balance between a good view of network performance and an unmaintainable set of messages.

In order to perform perpetual monitoring, the selected monitoring and monitored nodes must run the test, such as NG-Ping, at a set interval and collect and store the resulting statistics.

Network Performance Monitoring, as described in section 3.7, is as good example of the case where Perpetual PNMV is required.

An NG-OAM solution MUST offer the ability to change monitoring timing intervals. Values as low as 3.3 ms have been suggested, but are optional. Values down to 100 ms SHOULD be supported.

4.2. Network Performance Monitoring

TOC

Network Performance Monitoring (PM, or NPM) is the art and science of recording temporally aware network performance characteristics. A use case for the resulting statistics is for SLA verification, in addition to proactive maintenance.
Relevant PM characteristics are typically loss, latency and jitter. A PM solution MUST index these characteristics to time intervals. Knowing that 100 packets were lost, but not knowing when is not particularly actionable. The limits of existing tools and information often results in a NOC "clearing counters" then running a "fast ping" for an arbitrary length of time and hoping that the error occurs again.
Keeping all results of a Perpetual PNMV test is one possible solution, however this volume of information can be difficult to store or to sort through when a network event is occurring. A NG-OAM solution SHOULD provide easy-to-read, temporally-aware, statistic that allows an operator to easily assess the magnitude of the problem.
An example of this sort of statistic from the world of SONET/SDH transport is the errored second, and severely errored second.
The level of granularity of PM statistics gathering SHOULD be configurable.

5. Other Requirements

TOC

5.1. Intra-AS Requirements

TOC

The NG-OAM solution SHOULD use the same mechanism to address both the Intra-AS (this section) and Intra-AS (Section 5.2) requirements. An operator MUST be able to run a traceroute from one domain and through another. The amount of information this traceroute provides may differ depending on where the probe is originated, and what sort of authorization it possesses to access information in other domains.
Intra-AS requirements are applicable within an Autonomous System (AS), where all IP/MPLS devices are expected to be under a single administrative authority. Because devices are under a single administrative authority, copious diagnostic information that can be

returned to the Initiator of a ping/traceroute request. Ultimately, however, an NG-OAM solution MUST ensure that extensive Intra-AS diagnostic information is not leaked across the boundaries of the Autonomous System, since it would provide valuable network intelligence information. In addition, it is desirable if lightweight authentication and/or encryption techniques can be used to secure both probe requests and replies, in order to limit the effects of resource exhaustion on network elements that are processing probe request/replies.

The following is a brief summary of the minimal set of information that a NG-OAM solution is expected to address. NG-OAM solutions MAY capture additional information through, for example, experimental or vendor-specific objects specified in the NG OAM probe-request.

NG-OAM Probe Requests and Probe Replies MUST contain a "Query ID", generated by the Probe Initiator, that can be used to associate Probe Responses to Probe Requests.

Next-Gen Traceroute

* MUST work for IP and MPLS

* MUST be able to specify a 5-tuple IPv4 or IPv6 "flow" in a Probe Request

* MUST be able to specify whether the IPv4 packet is a first-fragment, or subsequent fragment, in order that intermediate devices can adjust their LAG/ECMP calculation appropriately.

* MUST be able to specify the MPLS label stack use to identify a "flow" across an MPLS-only portion of the network in a Probe Request.

* MUST be able to specify the Layer-2, (e.g.: Ethernet), header used in a Probe Request.

* MUST be able to specify a combination of label stack and IP 5-tuple, if both are used in the ECMP/LAG hash algorithm.

* MUST capture the following information in a Probe Reply:

  - The specific components of Layer-2, (e.g.: Ethernet), header, MPLS label stack and/or IP 5-tuple, that were used in the ECMP/LAG hash algorithm at this hop

  - Incoming Interface Name

  - Outgoing Interface Name

  - Number of component-links in a bundle

  - Size (Bandwidth) of individual component-links in a bundle

-Percent bandwidth utilization on interface(s)

-Remote Link-Layer neighbor name and interface name

*SHOULD be able to, on request of the source, to provide recent performance history of the incoming or outgoing link(s)

Next-Gen Ping

*MUST work for IP and MPLS

*MUST be able to specify a 5-tuple IPv4 or IPv6 "flow" in a Probe Request

*MUST be able to specify the MPLS label stack use to identify a "flow" across an MPLS-only portion of the network in a Probe Request.

*MUST be able to specify the Layer-2, (e.g.: Ethernet), header used in a Probe Request.

*MUST follow the regular data-plane path for forwarding within a network element

*MUST be able to test all links/paths concurrently, or serially, between two network elements when operators do not know a customer's "flow" information, which can be used as input to a LAG and/or ECMP hash calculation.

Proxy Traceroute

*All of the requirements mentioned above for "Next-Gen Traceroute", plus:

*The Initiator MUST be able to specify the number of Probe Requests.

*The Initiator MAY also specify the interval between Probe Requests, which the Proxy IP/MPLS Router is responsible for carrying out on the Initiator's behalf.

Proxy Ping

*All of the requirements mentioned above for "Next-Gen Ping", plus:

*The Initiator MUST be able to specify the number of Probe Requests and interval between Probe Requests, which the Proxy IP/MPLS Router is responsible for carrying out on the Initiator's behalf.

Next-Gen OAM Traceroute/Ping Probe Replies MUST capture error conditions that were encountered during an unsuccessful Probe Request. Those replies are expected to capture not only those conditions defined by classic [ICMP], (e.g: Destination Unreachable Type), but also new error conditions specific to NG-OAM solutions. In order to seamlessly accommodate future error conditions, NG-OAM solutions MUST use a TLV format for specifying error conditions in Probe Replies.
Intra-AS probe requests (and probe replies) MUST be easily identifiable in the data plane, in order that routers acting on NG-traceroute or NG-ping requests (or replies) can rapidly drop them in order to avoid resource exhaustion. NG-traceroute and NG-ping solutions MUST provide configurable methods to rate-limit the number of Intra-AS request (or reply) packets to prevent resource exhaustion.

5.2. Inter-AS Requirements

TOC

Inter-AS requirements are applicable across administrative domains, such as the Internet or, perhaps, several MPLS service providers delivering a single MPLS VPN solution. Because devices are not under a single administrative authority, only a limited amount of diagnostic information must be returned to the Initiator of a ping/traceroute request. This information is primarily useful in the context of helping the responsible party pinpoint the specific location of a problem. For example, Customer A may be experiencing packet loss in Service Provider A's network for his Internet service. The link between Customer A and Service Provider A consists of a ECMP path between SP A's ASBR and Customer A's ASBR. Customer A can perform a NG-traceroute through this ECMP path and provide the output of NG-traceroute to SP A's NOC in order to more rapidly identify the particular component-link, which is the causing a problem. Other examples where this is useful are: over Internet (IPv4 or IPv6) peering/transit links and within DataCenters from servers through to the DataCenter provider's ASBR attached to several SP's, where MPLS is not used.
Inter-AS probe requests (and probe replies) MUST be easily identifiable in the data plane, in order that routers acting on NG-traceroute or NG-ping requests (or replies) can rapidly drop them in order to avoid resource exhaustion. NG-traceroute and NG-ping solutions MUST provide configurable methods to rate-limit the number of Inter-AS request (or reply) packets to prevent resource exhaustion.
Next-Gen Traceroute

- MUST work for IP and MPLS

- MUST be able to specify a 5-tuple IPv4 or IPv6 "flow" in a Probe Request

- MUST be able to specify the MPLS label stack use to identify a "flow" across an MPLS-only portion of the network in a Probe Request.

- MUST be able to specify the Layer-2, (e.g.: Ethernet), header used in a Probe Request.

- MUST be able to specify a combination of label stack and IP 5-tuple, if both are used in the ECMP/LAG hash algorithm.

- MUST capture the following information in a Probe Reply:

  - Incoming Interface Name

  - Outgoing Interface Name

Next-Gen Ping

- MUST work for IP and MPLS

- MUST be able to specify a 5-tuple IPv4 or IPv6 "flow" in a Probe Request

- MUST be able to specify the MPLS label stack use to identify a "flow" across an MPLS-only portion of the network in a Probe Request.

- MUST be able to specify the Layer-2, (e.g.: Ethernet), header used in a Probe Request.

Proxy Ping/Traceroute requirements are not applicable to Inter-AS scenarios, since the risk of resource starvation is too large.

5.3. MTU considerations

TOC

Traceroute probes need to be kept to minimal size. Traceroute reply PDU's should be kept to 1500 Bytes in size in order to avoid the need for IP fragmentation. It is a safe assumption that operators have a minimum of 1500 Bytes for IP MTU, and often significantly larger. Optionally, path MTU discovery may be used to determine a minimum MTU. The MTU values MUST be configurable by the operator to adjust to unanticipated conditions. A Traceroute reply packet MAY span multiple packets.

TOC

5.4. Extensibility

It would be useful to allow for the "next-generation" traceroute and ping protocols to contain TLV's, in order that they may be easily extended in the future to account for additional capabilities, which may be developed at a later point in time.

5.5. Path Capabilities

TOC

In order to be certain that NG-ping or NG-traceroute will be able to properly exercise component-links in a LAG and/or ECMP path through the network, it is necessary to determine if all devices along a specific path are capable of supporting the requisite protocols and replying with appropriate results back to the originator of the NG-ping or NG-traceroute request. There are potentially two methods that can be employed to determine these capabilities: 1) path discovery; or, 2) encoding special/reserved codepoints into the packet header of NG-OAM request/reply packets. With the first method, the originating host/router could use a path discovery function to determine the capabilities and properties of intermediate and/or terminating devices prior to actually using NG-ping or NG-traceroute to test the data path. Once the originating host/router has learned the characteristics of intermediate and/or terminating devices, it could then originate a NG-ping/traceroute request using that information to exercise the actual data path.

The second method is likely to encode the NG OAM packets with specific values in the packet header of NG-OAM request/reply packets, (for example, via new ICMP type/codes or MPLS label values). In this approach, the originating host/router can simply launch a NG-ping/traceroute request allowing each intermediate and/or terminating device to independently determine if it's capable of supporting the NG-OAM request and, concurrently, exercising the component-links appropriate to the LAG and/or ECMP path.

Although the latter approach has the potential disadvantage that it may be more difficult to support on some existing hardware, this document recognizes that it is the superior approach of the two choices. If one depends on, for example, NG-traceroute to "discover" characteristics of a path before allowing one to ping, it creates a circular dependency. Specifically, in the case where one is doing perpetual pings and the underlying path changes for legitimate reasons, the NG-OAM would have to discover the change to the path, trigger a new NG-traceroute and then resume perpetual pings along the new path. Note that a change to the existing path could consist of any of the following: 1) a component-link in a LAG goes down, yet, the LAG itself remains operational, (e.g.: a 10x LAG goes to a 9x LAG), ultimately changing the result of LAG hashing algorithm; or, 2) the entire LAG and/or ECMP path goes down and data packets are routed along an alternate path. Ultimately, if each NG-OAM packet is a self-contained, autonomous OAM unit, then each intermediate and/or terminating device will act on it appropriately.

Therefore, this document specifies that a NG-OAM solution MUST support the second method, autonomous OAM units, outlined above. NG-OAM solutions MAY support the first method, to provide short-term NG OAM coverage with existing hardware.

5.6. Per Hop Behavior Modification

TOC

Modification of per-hop behavior in order to support NG-OAM is acceptable, but not required of NG-OAM solutions. This allows solutions where intermediate routers have to look at something new to determine if they are looking at an OAM packet, or to determine if they are they target or Proxy of a NG-OAM request.

---

6. IANA Considerations

TOC

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

---

7. Security Considerations

TOC

Devices MUST rate-limit the amount traceroute and/or ping traffic they process to avoid DoS attacks. Those rate-limits MUST be configurable to suit the appropriate environment in which they are deployed. An attacker must not be allowed to force an inordinate amount of traceroute and/or ping traffic down a single physical component-link causing congestion. Therefore, devices MUST rate-limit the amount of "external" traceroute and/or ping traffic through any specific component-link or set of component-links. Note, implementations SHOULD provide exceptions that to allow a network operators Intra-Domain traceroute and/or ping traffic, particularly for performance monitoring, to get through without interference by rate-limiters.

A lightweight authentication method SHOULD be provided by an NG-OAM solution. This mechanism can be used to defend against DoS or insertion attacks from other systems spoofing NG-OAM information. This can also be used in a reply message to defend against a "SLA Violation" attack where a malicious system could make it appear as if an operator's network has violated the SLA, when, in fact, they have not.

8. Acknowledgements

TOC

The authors would like to thank Nitin Bahadur, Ping Pan, Nasser El-Aawar, Dimitri Papadimitriou for their reviews and thoughtful feedback.

9. References

TOC

9.1. Informative References

TOC

[BFD-BASE]	"draft-ietf-bfd-base-07.txt - Bidirectional Forwarding Detection," January 2008.
[LLDP]	"IEEE Standard - 802.1AB-2005," May 2005.
[LMP]	"RFC 4204 - Link Management Protocol," October 2005.
[PROXY-LSP-PING]	George Swallow and Vanson Lim, "Proxy LSP Ping, draft-ietf-mpls-remote-lsp-ping-01.txt," November 2007.
[RSVP-DIAG]	"RFC 2745 - RSVP Diagnostic Messages," January 2000.

9.2. Normative References

TOC

[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).

9.3. References

TOC

[RFC 792]	"Internet Control Message Protocol," 2005.

Authors' Addresses

TOC

	Shane Amante
	Level 3 Communications, LLC
	1025 Eldorado Blvd

	Broomfield, CO 80021
Email:	shane.amante@level3.com
	Alia Atlas
	BT
Email:	alia.atlas@bt.com
	Andrew Lange
	Alcatel-Lucent
Email:	andrew.lange@alcatel-lucent.com
	Danny McPherson
	Arbor Networks, Inc.
Email:	danny@arbot.net

Full Copyright Statement

TOC

Intellectual Property