Internet Draft
Document: draft-allan-mpls-oam-frmwk-05.txt
Category: Informational
Expires: April 2004

David Allan(editor)
Nortel Networks

October 2003

A Framework for MPLS Data Plane OAM

Status of this Memo

Copyright Notice

Abstract

This Internet draft discusses many of the issues associated with data plane OAM for MPLS. The goal being to provide a comprehensive framework for developing tools capable of performing "in service" maintenance of LSPs. Included in this discussion is some of the implications of MPLS architecture on the ability to support fault, diagnostic and performance management OAM applications, a summary of currently specified OAM mechanisms, and a framework whereby collectively this MPLS-OAM toolset can address all aspects of the MPLS architecture.

Sub-IP ID Summary

(This section to be removed before publication.)

WHERE DOES IT FIT IN THE PICTURE OF THE SUB-IP WORK

Fits in the MPLS box.

WHY IS IT TARGETED AT THIS WG

MPLS WG has added requirements, framework and mechanisms for OAM to its charter. This draft is a candidate framework document.

JUSTIFICATION

The WG should consider this document, as it discusses the design aspects of error detection and measurement for packet based MPLS LSPs.

Table of Contents

1. Conventions used in this document

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 [1].

The term MPLS "level" nominally refers to the MPLS stack level inclusive of reserved labels. In this document the term "level" is used exclusive of reserved labels, therefore the term "level" is more precisely analogous to a specific MPLS subnetwork layer instance.

2. Changes since the last version (to be removed on publication)

Section 11 recast from being a discussion of potential mechanisms, to being a survey of the defined mechanisms.

Section 14 added which provides a survey of MPLS OAM mechanisms defined in both the IETF and the ITU-T.

Reference to [CHANG] draft and discussion of reverse notification tree removed.

Reference to [HEINANEN] on directory based LDP VPNs removed.

Reference to [HARRISON-REQ] and [HARRISON-MECH] replaced with Y.1710 and Y.1711 respectively.

[MARTINI] reference updated.

3. Contributors

Mina Azad
Azad-Mohtaj Consulting Phone: 1-613-722-0878
Ottawa, Ontario, CANADA Email: mohtaj@rogers.com

Jerry Ash
AT&T
Room D5-2A01

200 Laurel Avenue Phone: +1 732-420-4578
Middletown, NJ 07748, USA Email: gash@att.com

Neil Harrison
BT Global Services Email: neil.2.Harrison@bt.com

Sanford Goldfless
192 Fuller St Phone: 617-738-1754
Brookline MA 02446 Email: sandy9@rcn.com

Eric Davalo
Maple Optical Systems

3200 North First Street Phone: 408 545 3110
San Jose CA 95134 Email: edavalo@mapleoptical.com

Arun Punj
Marconi Communications
1000 Marconi Drive,
Warrandale - PA - 15086 Email: Arun.Punj@marconi.com

Marcus Brunner
Network Laboratories - NEC Europe Ltd.
Adenauerplatz 6 Phone: +49 (0)6221/ 9051129
D-69115 Heidelberg, Germany Email: brunner@ccrle.nec.de

Chou Lan Pok
SBC Technology Resources, Inc.
4698 Willow Road, Phone: +1925-598-1229
Pleasanton, CA 94583 Email: pok@tri.sbc.com

Wesam Alanqar
Sprint
9300, Metcalf Ave, Phone: +1-913-534-5623
Overland Park, KS 66212 Email : wesam.alanqar@mail.sprint.com

M. Akber Qureshi
Lucent Technologies
101 Crawfords Corner Road Phone: +1 732 949 4791
Holmdel, NJ 07733 Email: mqureshi@lucent.com

Don Fedyk
Nortel Networks
600 Technology Park Phone: +1 978 288 3041
Billerica MA 01821 EMail: dwfedyk@nortelnetworks.com

4. Requirements

MPLS data-plane OAM specific requirements and a summary of requirements that have appeared in numerous PPVPN, PWE3, and MPLS documents appear in [Y1710] and [MPLSREQS]. This Internet draft discusses the implications of extending OAM across the MPLS architecture, and adds additional data-plane OAM requirements and capabilities for managing multi-provider networks. This document also broadens the scope of the requirements discussion in identifying where certain OAM applications simply cannot be implemented without modifications to current practice/architecture.

Finally this draft offers a survey of the currently standardized or about to be standardized tools.

5. Domain Concepts

MPLS introduces a richness in layering which renders traditional definitions of 'domain' inadequate. In particular, it is noted that

MPLS has no fixed layered hierarchy (this is a unique property that no other technology has offered before).

A provider may have MPLS peer providers, use MPLS transit from serving providers (and require MPLS or non-MPLS client transport), and offer MPLS transit to MPLS or non-MPLS clients). Further, the same provider may use a hierarchy of LSPs within their own network. This Internet Draft defines the concept of an "Operations Domain" (to cover OAM capabilities operated by a single provider) that may only be a portion of the end-to-end LSP. Operations Domain functions are an interdependent mix of control-plane, data-plane (a.k.a. user-plane), and management-plane functions.

An LSP "of level m" may span numerous Operational Domains. The data plane of the LSP is a contiguous entity consisting of data plane portions of traversing operational domains. The control and management planes of these operational domains may be disjoint. The goal is to provide OAM functionality for each LSP independent of the LSP creation mechanism or payload.

It is possible to have a hierarchy of operators (e.g. carriers of carriers), where overlay Operational Domains are opaque to the serving Domain. Therefore it is required that each LSP Operational Domain implement its own OAM functionality, and the OAM applications are confined to the Operational Domains traversed at level "m".

Note that this concept has subtle differences with concepts of horizontal and vertical hierarchy as defined in [HIERARCHY].

Vertical hierarchy usually refers to networking layer boundaries distinguished by technology. An operational domain may refer to an operator specific hierarchical subset of the LSP levels within the MPLS network and/or a horizontal partitioning within a specific LSP. Similarly there is a further way to consider the concept of operational domain and horizontal hierarchy. An operational domain may be hierarchically partitioned (e.g. OSPF "areas") but may be operationally integrated and contiguous.

6. OAM Applications

The purpose of having data plane LSP specific OAM transactions is to support useful OAM applications. Examples of such applications include:

Fault management

- On demand verification: the ability to perform connectivity tests that exercise the specific LSP and the provisioning at the ingress and egress. On demand suggests that verification may be performed on an ad-hoc basis.

- Fault detection: Operators cannot expect customers to act as fault detectors, and so the ability to perform automated detection of the failure of a specific LSP is a "must have" feature (although when

one reviews the section on LSP creation above, one realizes it will not be ubiquitously used). Some MPLS deployment scenarios may not have a control plane or may have LSP processing components not in common with the control plane, so fault detection procedures may need to be augmented with LSP specific methods.

- Fault sectionalization: The ability to efficiently determine where a failure has occurred in an LSP. Sectionalization must be able to be performed from an arbitrary LSR along the path of the LSP.

- Fault Propagation: specific MPLS deployment scenarios may not have a control-plane to propagate LSP failure information. Fault propagation has numerous forms and there are variations depending on whether the failure is in the serving layer/level or :
i) Northbound from the failed level to the management plane.
ii) Within the failed level.
iii) From the failed level to its clients.
iv) Within the client level to the LSP ingress and egress either via the user or control planes.
And in all cases it is the termination of a layer that performs the function.

Performance management

- The ability to determine whether an LSP meets certain goals with respect to latency, packet loss etc.
- The ability to collect information to facilitate network engineering decisions.

Of the above applications, verification, detection and sectionalization explicitly need to exercise all components of the forwarding path of the target LSP, otherwise there will be failure scenarios that cannot be detected or properly sectionalized. These applications cannot be supported properly if there are differences in handling between user traffic and OAM probes at intermediate LSRs.

A separate and useful classification of the applications outlined above is to distinguish the difference between monitoring applications and diagnosis. Monitoring applications are typically unattended in operation, collect operational statistics, and upon detection of problems, must provide sufficient information to permit precise diagnosis of the problem to be performed and frequently some form of automated network response to problems. Diagnosis applications are typically attended in operation and must be able to authoritatively locate and isolate faults. The security implications of this distinction is discussed in the security considerations section.

7. Deployment Scenarios

At the present time there are a number of MPLS deployment scenarios each with a number of subtleties from a data plane OAM perspective. Each can be viewed as a characteristic of an operational domain:

The sparse model: This can be in conjunction with control plane signaling (e.g. MPLS based traffic engineering applied to an IP network) or with simple provisioned LSPs (no control plane signaling). The key feature being that the MPLS operational domain will not have any-to-any connectivity at the MPLS layer due to the sparse use of LSPs to augment the served layer connectivity. This has operational and scalability implications as OAM connectivity must be explicitly added to the model, or the operator may be obliged to depend on "layer violations" embedded in OAM mechanisms which are strictly only relevant to a different higher layer network (e.g. [ICMP]) to generate a return path.

The ubiquitous model: This model generally combines MPLS, integrated routing and control to produce universal any-to-any connectivity within an operational domain. This may be combined with a hierarchy of LSPs to modify the topology presented to the client layer. This offers providers the option of utilizing the resources inherent to all planes of the Operational Domain in designing OAM functionality.

These two models of MPLS connectivity can be stacked or concatenated to support numerous configurations of peering and overlay networking arrangements between providers and users. A direct inference being that an operational domain will not necessarily have knowledge of the domains above and/or below it, and in the general case far less knowledge of (and certainly less control over) its peer domains. OAM applications for LSPs of a specific level are confined to an operational domain and its data plane peers.

More recently there is a tendency to overlay a L2 or L3 VPN service level on the data-plane of an operational domain, with it's own identifiers and addressing, while tunneling control information across the control plane of the operational domain using BGP-4 [2547][KOMPELLA] or extended LDP discovery [MARTINI]. From a data plane OAM perspective, we would consider this to be a separate operational domain, and anticipate that it is only a matter of time before such service levels evolve to span multiple operational domains (for example, an L2 or L3 VPN that spans multiple providers, or the introduction of tandem points at the data plane of the service level).

8. MPLS architecture implications for OAM

8.1 Topology variations within an MPLS level

There are a number of topology variations in the MPLS architecture that have OAM implications. These are:

- Uni-directional and bi-directional LSPs. A uni-directional LSP only provides connectivity in one direction, and if return path

connectivity exists, it is an attribute of the operational domain (e.g. signaling, management or client layers), and not a unique attribute of the LSP. Bi-directional LSPs or specific return path (e.g. [DUBE]) have inherent symmetrical connectivity as an attribute of the LSP.

- Multipoint-to-point (mp2p) LSPs are where a single LSP uses "merge" LSR transfer functions to provide connectivity between multiple ingress LSRs and one egress LSR (sufficient information being present in the payload to permit higher layer demultiplexing

at the egress). There are a number of problems inherent to mp2p topological constructs that cannot be addressed by traditional p2p mechanisms. One issue being that for some OAM applications (e.g. data plane fault propagation) OAM flows may require visibility at merge-points to limit the impact of partial failures or congestion.

"Best effort" mp2p LSPs may have fairness issues with some packet schedulers. This may complicate obtaining consistent measurements under congestion conditions. Explicitly routed mp2p LSPs with associated resource reservations are significantly more complex to engineer. The resource reservations required will be cumulative at merge points (as will jitter), and the ability to provide differentiated handling for specific ingresses is lost once any merge point is crossed. One opinion would be that the complexity and difficulty in the configuration/maintenance of ER-mp2p LSPs significantly outweighs the scalability benefits, and would not likely be deployed.

- Penultimate Hop label Popping (PHP), is an optimization in the architecture in which the last LSR prior to the egress removes the redundant current MPLS label from the label stack. Therefore the ability to infer LSP specific context (OAM and other) is lost prior to reaching the final destination.

MPLS does not provide for protocol multiplexing via payload identification (with the exception of the explicit IPv4 and IPv6 labels). PHP requires that the final hop have a common protocol payload (typically IP) or is able to map to lower layer protocol multiplexing capability (e.g. PPP Protocol Field or Ethernet ethertype) as the ability to infer payload type from LSP label is lost.

Another scenario where PHP is employed is when the egress LSR is not actually MPLS data plane capable. This has data-plane OAM implications in that any MPLS specific flows need to terminate at the PHP LSR. This requires that the PHP LSR proxies OAM functions on behalf of the egress LSR. This will introduce complexity when any type of consequent actions such as layer interworking of fault notification is required.

- E-LSPs [MPLSDIFF] in which a single LSP supports multiple queuing disciplines to support multiple QoS behavior aggregates. Ability to

perform OAM performance functions on a "per behavior aggregate" basis is critical to managing E-LSPs.

- Management plane provisioned LSPs, vs. control plane signaled

LSPs. In many scenarios associated with a control plane, the topology of the LSP varies over time. This can be due to many reasons, implicit routing, dynamic set up of local repair tunnels etc. etc.

- The potential existence of multiple LSPs between an ingress and an egress LSR. This can be for many reasons, L-LSPs, equal cost multipath routing etc. etc.

- The potential existence of multiple next hop label forwarding entries (NHLFEs) for a single incoming label. This is the scenario whereby the incoming label map (ILM) for an incoming label switch hop (LSH) maps to an inverse multiplex of NHLFEs which may be re-merged into a common egress or have multiple egress points. The mechanism for selecting the NHLFE to use may be proprietary and is performed on a packet by packet basis. Some implementations hash both the label stack and any IP payload source and destination addresses in order to preserve flow ordering while achieving good fan out. However this means that predictability of any nested LSPs degrades in the presence of problems.

OAM tools not specifically aware of this construct may produce random results (insufficient frequency of failure to trigger threshold detection), or pathologically may only test a subset of the NHLFEs impacting both the detection and diagnosis of defects. Similarly performance monitoring is impacted as packets in flight cannot accurately be accounted for. The ramifications are comprehensively discussed in [ALLAN].

- Use of per-platform label space. A per-platform label has significance at a nodal level and not just an interface level. Some of the more interesting applications being the ability to create unsignalled facility backup LSPs in "bypass tunnels" [SWALLOW]. Traffic arriving on multiple interfaces and/or LSP tunnels may use a common per-platform label and will have a common ILM and NHLFEs. This can have implications similar to mp2p and PHP depending on how it is used; LSP origin information is not conserved when multiple sources use a common label.

- p2mp and mp2mp LSPs (a.k.a. MPLS Multicast) is for further study. At the present time what placeholders exist in the architecture for multicast treat it as a separate protocol from "unicast" MPLS (with the exception of ATM variations of MPLS).

These topological variations introduce complexity when attempting to instrument OAM applications within a specific MPLS level such as performance management, fault detection, fault isolation/diagnosis, fault handling (e.g. consequent actions taken to avoid raising unnecessary alarms in client layers) and fault notification.

8.1.1 Implications for Fault Management

mp2p, E-LSPs and PHP have implications for fault management, specifically if an LSR is required to have knowledge of both the ingress LSR and the specific LSP that an OAM message arrived on, or is expected to have knowledge of, and maintain state about the set of ingress LSRs for an LSP. OAM messaging needs mechanisms to distinguish both the ingress LSR and the specific LSP. (This ability is expressed on these terms as LSPs are typically not given globally unique identifiers, more frequently some locally administered LSR-ID is used).

Connectivity verification requires testing of connectivity between all possible ingress/egress combinations. Frequently it will not be possible to infer the ingress LSR and specific LSP directly at the egress as such information may be lost at merge points in mp2p LSPs or due to PHP. This is true for both OAM messaging, and normal data plane payloads. There may be numerous reasons why an ingress-egress pair may have a plurality of LSPs between them, so the ability to distinguish the source and purpose of specific probes beyond mere knowledge of the originating LSR is required.

The ability to distinguish the ingress can be achieved via modifying the OAM protocol to carry such information, or may be achieved via modifications to operational procedures such as overlaying p2p connectivity.

8.1.2 Implications for Performance Management

Many performance management functions can be performed by obtaining and comparing measurements taken at different points in the network. Comparing ingress and egress statistics being the simplest example (but is usually restricted to within a single domain). The key issue is ensuring that "apples-to-apples" comparison of measurements is possible. This means that all measurement points need to be able to similarly classify the traffic and performance they are measuring, and that the measurements are synchronized in time and compensate for traffic in flight between the measurement points.

For example, a relatively simple technique for establishing key performance metrics would be to compare what was sent with what was received. For example in the PPP line quality monitoring (LQM) function the ingress periodically sends statistics to the egress for comparison subject to the same queuing discipline as the data plane traffic, such that traffic in flight is properly accounted for. (Note that re-ordering will introduce errors but is not expected to be frequent.)

It is important to distinguish, and be able to measure, what constitutes the up and down states of an LSP. This needs to be standardized so that there is unified treatment. A key observation here is that QoS metrics (like loss, errored packets, delay, etc)

are only relevant to when the LSP is in the up-state; and so any collection of QoS measurements is suspended when the LSP enters the down-state. This requires specification of the state transitions to achieve measurement consistency, and is a pre-requisite to QoS assessment. This is a particularly important metric to operators, since customers will be expecting operators to be able to offer both QoS and availability SLAs, and so these must be differentiated and uniquely measurable.

A simple ingress/egress comparison is not always possible, there is no ability to similarly classify what is being measured at the ingress and egress of an LSP. mp2p LSPs and PHP do not have a 1:1 relationship between the ingress and the egress. LSPs containing ILMs that map to multiple NHLFEs introduce measurement inaccuracy as not all packets share a common queuing discipline and where this results in multiple egress points from the network, there is an inability to synchronize measurements. Partial failure of an mp2p LSP (incl. ECMP) will result in the inability to successfully collect statistics

So, in addition to having to define up/down-state transitions, for successful PM the 1:1 relationship needs to be restored by either:

- The mp2p/PHP LSP is modeled as one LSP for measurement. This means that measurements performed at ingress points need to be synchronized and adjusted for common LSP segments such that the results are all presented to the egress simultaneously (again correcting for traffic in flight), a technique dependent on such a high degree of synchronization would be impossible to perfect, and prone to a degree of error.

- The mp2p/PHP LSP is modeled as a collection of "ingress" LSPs for measurement. This means that the egress needs to be able to maintain statistics by ingress and appropriately classify traffic measurements.

Neither of the above is achievable at the present time without modifying existing operational procedures. The first approach involves treating the mp2p/PHP LSP as an aggregate, and as such it can partially fail and degrade. This complicates the establishment of performance metrics and specifying recovery procedures on errors.

The second approach requires decomposing the mp2p/PHP LSP such that both payload and OAM traffic can be demultiplexed at the egress and correctly associated with "per-ingress" state. The ability to demultiplex both OAM and payload implies a common wrapper, and the net effect would be to overlay p2p connectivity on top of the merge/PHP based transport level.

The existence of E-LSPs adds a wrinkle to the problem of measurement synchronization. An E-LSP may implement multiple diffserv PHBs and incorporate multiple queuing disciplines. An aggregate measurement for the entire LSP sent from ingress to egress would frequently have

a small margin of error when compared with an aggregate measurement taken at the egress. Separate measurement comparisons for each supported EXP code point would be required to eliminate the error.

The situation is slightly different for p2p LSPs containing ILMs that map to multiple NHLFEs. If all the NHLFEs are merged back into a single entity prior to the egress, there will inherently be a degree of measurement error that modifications to operational procedure cannot correct. However there is no guarantee that this will be the case, and any individual ingress measurement may be compared with only one of several egress measurement points (either random or pathological).

8.2 LSP Creation Method

The ability to usefully audit the constituent components of an LSP is dependent on the technique used to create the LSP.Presently defined are provisioning, LDP, CR_LDP, RSVP-TE, and BGP.

LSP creation techniques that are currently defined fall at two relative extremes:

At one extreme is explicitly routed point-to-point connection between fixed ingress and egress points in the network. Explicitly routed (ER) LSPs (today created via provisioning, CR-LDP, RSVP-TE or BGP) have a significant degree of testability as the path across the network and the egress point is fixed and knowable to a testing entity. Similarly explicit pairwise and stateful testing/measurement relationships can be set up (e.g. connectivity verification) and strict criteria for failure established.

In the middle is static mp2p constructs typically signaled via BGP (e.g. RFC 2547).

At the other extreme is when LSP construction is topology driven (such as dynamic "shortest path first" routing combined with LDP),

whereby the details of path construction between the ingress and egress points in the network will vary over time and may involve several stages of multiplexing with traffic from other sources. The details of path construction at any given instant are not necessarily knowable to an auditing entity so any attempt to interpret the results of an audit may generate spurious results. Further, the MPLS network may only be a portion of the operational domain, and the egress point from the network for an FEC may vary over time.

The testable unit in an LDP network is the FEC not the LSP, and the potential existence of a many to many relationship of ingress and egress points limits the testability of the FEC, or at least may limit the frequency of using such tests.

The connectivity instantiated in a specific LSP created by a topology driven control plane signaling mechanism will recover from

many defects in the network. The quality of recovery typically being a function of how the network is engineered.

Problems are typically detected by having MPLS connectivity fate share with the constituent physical links and routing adjacencies, and topology driven path re-arrangement will restore the connectivity (with some interruption and other side effects occurring between the initial failure and re-convergence of the network). However exclusive dependence on fate sharing for failure detection means that LSP components may have unique failure modes from which the network will not recover and can only be diagnosed reactively.

As can be inferred from the above, what is required for topology driven LSPs is a test mechanism that audits forwarding policy as this is the metric by which some aspects of network performance can be measured.

8.3 Lack of Fixed Hierarchy

MPLS supports an arbitrary hierarchy in the form of label stacking. This is a facility that can be leveraged for OAM purposes. As an example, the section on implications for performance management has already outlined how p2p topology for PM can be overlaid on an arbitrary merged topology to add manageability of services. Similarly functions requiring sectionalization of an LSP or ability to isolate partial failure of a complex construct can be achieved by constructing the LSP as an overlay upon a concatenation of operationally significant shorter LSPs. By operationally significant

we would refer to LSPs that spanned useful portions of the whole construct (e.g. a branch of an mp2p LSP, or bypassed LSRs that did not have OAM capability).

This could simplify the instrumentation of level specific OAM by ensuring only e2e functions were required (as opposed to functions originating or terminating at arbitrary points in the network), while driving up the complexity of LSP establishment due to the resultant inter-level configuration issues when creating multi-level constructs with the desired manageability.

8.4 Use of time to live (TTL)

Experience within the IP world has suggested that TTL was a serendipitous feature that can be similarly leveraged by MPLS.

However in the MPLS world, TTL suffers from inconsistent implementation depending on the link layer technology spanned by the target LSP. The existence of non-TTL capable links (e.g. MPLS/ATM) has impact on the utility of using TTL to augment the MPLS OAM toolkit. For example, use of TTL as an aid in fault sectionalization can only isolate a fault to the granularity of a non-TTL capable span of LSH or LSP segments.

There are other variations in TTL handling that suggest interpreting results of TTL based tests may be problematic. As outlined in [TTL] there are two models of TTL handling with different implications:

- the uniform model, in which decrement of TTL is independent of the MPLS level. At the ingress point to an MPLS level, the current TTL is copied into the new top label, and at egress is copied back to the revealed top level.

- the pipe and short pipe models, whereby MPLS tunnels (aka LSPs) are used to hide the intermediate MPLS nodes between LSP ingress and egress from a TTL perspective.

The uniform model originates with preserving IP TTL semantics when IP traffic transits an MPLS subnetwork. The uniform model will reduce the resource consumption of routing loops, but in a correctly operating network may lead to premature discard of packets outside the operational domain they originated from (due to the existence of an arbitrary number of serving MPLS levels). Similarly when a routing loop occurs, determining the MPLS level that is the source of the problem will be difficult as there is no method to correlate it with the level where the exhaustion event occurred.

The pipe model is more consistent with the operational domain model in that TTL exhaustion will only occur at a specified level and the initial values used at LSP ingress are more likely to be reflective of detecting what would genuinely constitute a routing loop.

A reasonable expectation is that the uniform model would not be used outside of an operational domain.

A separate issue is that it is also possible that an LSR may decrement TTL by an amount other than one as a matter of policy. This means that the results obtained via any tools that use TTL exhaustion will require some interpretation.

8.5 State Association

The design of OAM flows in MPLS levels that multiplex traffic from multiple sources together may introduce implementation complexity where the flows are processed. The receiver of the OAM message will need to extract information from the packet to identify the LSP and associate it with ingress and LSP specific state. If the ingress/LSP identifier in the packet is not administered by the processing node, it will be unable to optimize the implementation of the state association mechanism and will be required to perform some sort of table search.

If the identifier is administered by the processing node and that node is not the originator of the probe, some mechanism will be required to distribute this information uniquely to each probe originator.

8.6 Alarm Management

MPLS permits layers of different operational behaviors to recurse. When the alarm management paradigms differ they may not be reconcilable. For example, an LDP network has no ability to perform alarm suppression directly within the dataplane for e2e tools either used within the LDP layer or overlaid on an LDP layer that are impacted by a failure. The LDP network will recover, but the node that could report the failure may not directly participate in the recovery, therefore data plane alarm suppression mechanisms cannot be synchronized with service restoration.

8.7 Other Design Issues

It is desirable to make the data plane OAM implementations independent of LSP specifics. It would be desirable to have common mechanisms across p2p and mp2p LSPs, PHP or no-PHP and independent

of payload and the method of LSP creation in order to minimize overall complexity. The OAM application originator should not need (as far as is practical) any knowledge of the details of LSP construction.

PM may require that instrumentation of many OAM applications is only possible for p2p LSPs and therefore would only be possible for a select group of MPLS levels (e.g. overlaid service labels as per [KOMPELLA] or [MARTINI]).

Fault management must be applicable across the spectrum of all label levels and LSR transfer functions.

Finally, the possibility of re-ordering of OAM messaging must be considered. The design of OAM applications and messaging must be tolerant of out of order delivery and some degree of packet loss. For some applications the originator/termination will require a means to uniquely correlate requests with probe responses (including responses to mis-directed probes) or verify in sequence receipt.

9. Ease of Implementation

Complex functions are typically require software implementation and are not capable of handling line rate messaging. Implementations defend themselves via rate-limiting or similar load management techniques to avoid vulnerabilities to DOS attacks or simple mis-use by incompetent craftspersons. In many cases, the complexity of adding strong authentication as defense against DOS attacks may be less onerous than promiscuous processing of complex probes.

Probes supporting monitoring applications gain the most benefit when they can run at line rate such that there are no concerns about processing capacity at the processing network elements. Such tests will generate predicable results (or at least not have results

degraded when network elements are under stress) and automated procedures can be designed around such mechanisms. MP2P LSPs are an exemplary case where egress processing of probes may be required to support probes from an arbitrary number of unsynchronized sources.

Messaging mechanisms to perform diagnostic tests (once a fault has been authoritatively established) tend to be more complex and software intensive. Diagnostic tests are frequently used by craftspersons, and can be more tolerant of things like discard due to rate limiting.

10.OAM Messaging

OAM should be decoupled from user behavior to ensure consistent OAM functional behavior (under any traffic conditions) and avoid the use of customers as guinea pigs.

At the specific LSP level, support of OAM applications require messages that flow between three entities, the LSP ingress, the intervening network and the LSP egress. As an LSP is unidirectional, it should be self evident that OAM applications that require feedback in the reverse direction will have such communication occur either at the specific LSP level, or some data plane LSP level in the operational domain, or one of the other planes (control or management) of the operational domain.

The set of possible individual transactions (plus examples of their utility) is as follows:

LSP specific data-plane transactions:
- ingress to egress
  applicability: verification, fault detection, performance management
- ingress to network
  message will terminate at an intermediate LSR traversed by the LSP.
  Applicability: sectionalization from source
- network to egress
  message is inserted into the LSP at an intermediate node and terminates at the LSP egress LSR.
  Applicability: sectionalization from arbitrary point in an LSP.
- Network to network
  Applicability: sectionalization from arbitrary point in an LSP.

Feedback transactions
- egress to ingress
  applicability: verification, fault detection.
- egress to network
  flow originates at the LSP egress and terminates at an
  intermediate node traversed by the LSP.

Applicability: sectionalization from arbitrary point in an LSP.
- network to ingress
  flow will originate at an intermediate LSR traversed by

the LSP and terminate at the LSP source.
Applicability: sectionalization from ingress.

- network to network
  Applicability: sectionalization from arbitrary point in an LSP.

11. Distinguishing OAM data plane flows

MPLS provides several mechanisms for distinguishing OAM data plane flows.

11.1 RFC 3429 "OAM Alert Label"

RFC 3429 [3429] defines the OAM alert label which identifies that the payload is a Y.1711 PDU. The OAM alert label may be used for p2p LSPs that do not encounter lower layer ECMP, and for Y.17fec-cv PDUs.

11.2 VCCV

[VCCV] provides procedures for PEs to negotiate an OAM protocol to be multiplexed with payload over a PW, and defines a bit in the PW header which indicates when the PW PDU contains OAM flows or payload flows. The purpose is to carry IP based OAM protocols (LSP-PING, ICMP etc.) opaque to any ECMP mechanisms

11.3 PW PID

[ARCH] defines a PW PID which permits OAM protocols to be multiplexed with a PW in a form whereby they self identify to the far end PE. This can be used to transport Y.1711 or Y.17fec-cv PDUs opaquely over an ECMP infrastructure such that they properly fate share with the PW.

12. The OAM Return Path

The ability to use OAM applications such as single-ended monitoring of both directions from one end, or to support applications such as protection switching in a 1/N:M case, requires a return path to the LSP ingress. This enhances the scalability and reliability of some OAM applications as data plane OAM can function as a closed system. A specific example being use of a loopback where the only place state and timing need be maintained is at the loopback originator.

This requires a return path to complete the loop between the "target LSP" and the OAM application originator. This will permit reliable transaction flows to be implemented that impose minimal state on the network.

For the few OAM applications that require a return path, the return path can be tolerant of being topologically disjoint with the target LSP (providing the differential delays are small, ie <<1s), reachability of the application originator being the only hard requirement. Similarly, different OAM applications will have different return path requirements, and a hybrid of using all the planes of the operational domain (according to the application) may be significantly simpler and more operationally tractable than significant modifications to current usage to fill in connectivity gaps at the specific label level.

This is a key point, LSPs are currently by definition uni-directional (bi-directional to date being a construct of multiple uni-directional LSPs), so for any non-ubiquitous deployment of MPLS connectivity, some modification of operational procedure to provide for OAM messaging will be required for the few applications that need it. Strict symmetry of connectivity at a specific label level is not guaranteed.

In any type of sparse usage scenario (e.g. provisioned LSPs or use exclusively for TE) there will not be an inherent any-to-any connectivity in the data plane, and there may not be a control plane signaling system.

In an implicit MPLS topology (e.g. LDP DU), any to any connectivity will typically exist, or will be easily available with minor alterations to operational procedure (LSRs advertise selves as FECs). This would continue to be true for an integrated model in which TE and an implicit topology were combined.

In any type of multi-provider MPLS topology, the scenario is more complex, as for numerous reasons a provider may not wish to provision/advertise external connectivity to their LSRs. Similarly, for security reasons, providers may wish to apply some degree of policy or filtering of OAM traffic at operational domain boundaries.

Data plane OAM messaging should be designed to leverage as much "free connectivity" as can be obtained in the network, while ensuring the constructs have sufficient extensibility to ensure the corner cases are covered.

Within the operational domain of a single provider, it is relatively easy to envision that a combination of data-plane, and control plane functionality will ensure that a data-plane return path is frequently available (although it may be topologically disjoint from the target LSP). This is less so for inter provider scenarios. Here there are a number of potential obstacles such as:
- disjoint control plane

- disjoint addressing plan
- requirements for policy enforcement and security
- impacts to scalability of ubiquitous visibility of individual LSRs across multiple operational domains.

There are a number of approaches to providing inter-domain OAM connectivity, the following is a brief commentary on each:

1) Reverse Notification Tree (a.k.a using bi-directional LSP)
In this method, each LSP has a dedicated reverse path - i.e. the reverse path is established and associated with the LSP at the LSP setup time. This requires binding the reverse path to each LSR that is traversed by the LSP. This method is not scaleable, as it requires doubling the number of LSPs in the network. Moreover each reverse path requires its own OAM.

2) Global OAM capability
Similar to IP v4 to IP v6 migration methodology, this method proposes use of a global operations domain with control-plane, data-plane, and management-plane that interact with control-plane, data-plane, and management-plane of individual operations domains. This method requires commitment and buy-in from all network operators.

3) Inter-domain OAM gateway
This method proposes use of a gateway like functions at LSRs that are at operations domain boundaries. OAM gateway like functions includes capabilities to correlate OAM information from one operations domain to another and permit inter-carrier sectionalization problems to be resolved.

Specification of an inter-domain OAM gateway capability would appear to be the most realistic solution.

13. Use of Hierarchy to Simplify OAM

MPLS hierarchy provides a mechanism to address a number of OAM issues.

Section 5 outlined domain concepts that nominally would require intermediate nodes to inspect and possibly process OAM PDUs. MPLS does not currently have this capability. However frequently an operational domain is self contained and may easily be instantiated as a distinct MPLS layer which transports the domain spanning MPLS client. This permits the domain specific components of the LSP to be uniquely instrumented using end to end tools and provides security benefits in that the provider specific components of the domain are logically isolated from the clients.

Section 7.1 outlined some of the impacts of MPLS topological constructs that multiplexed traffic from multiple sources together. Section 7.5 identified additional complexity modifying protocols to address state mapping for OAM purposes could entail. The key issue identified is that for fault management, OAM protocol design would permit mp2p and PHP to be addressed (but at a specific implementation cost), but this is not possible for performance management, in particular if ingress specific traffic counts are required.

Rather than attempting OAM protocol design to address what by definition will be an incomplete solution, it would be useful to define a common mechanism to demultiplex both MPLS level payload and OAM flows. The common mechanism ideally would be in the form of a wrapper that included an egress administered ingress identifier.

One instantiation of such a wrapper would be a p2p MPLS label. The mechanisms exist for label distribution (in the form of extended LDP discovery), and LSPs are already passively instrumented (e.g. packet and byte counts). Similar benefits are obtained when the implementation is extended into the use of probe messages. State association at the egress becomes simple in that the state is associated directly with the incoming label (and can be obtained by augmenting the ILM lookup).

The use of p2p overlays is one method of instrumenting mp2p and PHP LSPs that addresses all the issues outlined in section 7. It also significantly simplifies OAM protocol design and implementation.

14. Current Tools and Applicability

A number of OAM tools have been specified by both the IETF and the ITU-T.

14.1 LSP-PING (MPLS WG)

LSP-PING is designed to be retrofitted to existing deployed networks and to exercise all functionality currently deployed. In order to do so, the design trade off is that detection or diagnosis of a problem may take an arbitrary number of transactions.

Protocol complexity is tolerated as initial implementations will be in software. Protocol complexity manifests itself in the form of TLV encoding of key information (FEC stack elements, and downstream LSR label map. Future functionality may be added to the protocol via the definition of additional Type-Length-Value (TLV)

information elements.

Aspects of the protocol design would permit a sparse subset to be handled in hardware (exact pattern match on the PDU). For example, in a VPN application, pinging a PE is facilitated by limiting the number of FECs at any level in the stack to one. Presumably an implementation of probe handling that matched on a ping of the PE loopback address could be optimized for that specific case.

LSP-PING permits a uni-directional path to be tested from a single point, but depends on a reliable return path in order to propagate the test results back to the originating LSR. Therefore the protocol is designed to tolerate degrees of ambiguity in individual test results. Failure of an individual ping response may be due to any of several causes:

- Forwarding path failure (including partial failure of ECMP

  or other load balancing constructs)
- Return path failure
- Port rate limiting at the egress
- Port rate limiting at the ping origin
- Congestive loss in the network

And to deal with this ping supports several features to allow ambiguity to be eliminated via having the ingress perform variations of the original transaction:

- Probe sequencing to permit both ingress and egress to detect gaps in probe sequences.
- Return path may be specified permitting data plane and control plane problems to be distinguished.
- Destination address may be manipulated to exercise payload sensitive ECMP implementations

LSP-PING generally assumes PHP at the egress and that any specific LSP binding at the egress point of probe processing may not exist. From the perspective of reliable fault detection this is a minor issue as the use of a non-routable destination address limits any untested modes of failure. However this does alter the granularity of useful verification, as probe contents must be checked with the set of FECs associated with the LSR, rather than simply the set specifically associated with the LSP of interest. When testing a label stack for a VPN PE, the number of individual transactions required may be quite large as the number of FEC elements supported by the PE can be considerable.

LSP-PING permits a label stack. For PW and VPN application, PHP may be employed by the PE such that PWs and VPN labels may not be

directly tested (hence the FEC stack to permit transport or PSN probes to proxy verification for the transported application).

LSP-PING has a traceroute mode that can extract a significant amount of information w.r.t. network configuration. Specifically all details of path construction for a given FEC (note that LSP-PING will most likely need to be augmented with authentication and authorization capability in the long term).

Modes of use for LSP ping are being defined [LSR-TEST] that leverage TTL decrement to bound the scope of any individual test.

14.2 Y.1711 (ITU-T SG13/Q3)

Y.1711 is focused on fault/alarm management and availability measurement for P2P LSPs. The major design objective of Y.1711 as it currently stands is automatic defect detection and handling. A secondary goal is to be able to measure availability. It trades precision in fault isolation in return for simplified defect detection/handling capability (frequently referred to as "bounded detection time"). Y.1711 PDUs have a small number of fixed fields in order to minimize parsing and processing overhead.

Message processing is primarily performed at the egress such that for uni-directional LSPs, there is minimal ambiguity in detecting failure. This is also required to take the appropriate consequent actions, eg to inform higher layer clients of lower layer failures and thus avoid generating alarm storms in inappropriate places, or to suppress traffic if a security compromise is indicated (ie traffic arriving from the wrong source).

Probe processing provides a simple "pass/fail" indication and sufficient information to permit a craftsperson to initiate diagnosis. It is dependent on other tools to perform specific diagnosis and isolation of problems.

Y.1711 is not designed to extract information from the network as to configuration and layout of network components. It does not currently define any path tracing functionality and only operates on LSP endpoints.

A corollary of the above, is that only LSP end points have any role in OAM processing, and the Y.1711 PDUs pass transparently through intermediate nodes.

Y.1711 depends on some degree of ubiquitous deployment at the edge

to maximize coverage of fault detection.

Y.1711 is primarily focused on tunnel end points. However core LSRs may add significant value by implementing a specific subset of Y.1711: FDI generation for P2P LSPs to provide alarm suppression and fault notification to the edge devices when failures in the core occur.

14.2.1 Connectivity Verification (CV) PDU

The CV PDU is used as a heartbeat mechanism to verify connectivity between the LSP ingress and egress. Frequent injection of CV probes is a prerequisite for consistent/deterministic defect detection/handling and availability measurement. Injection of CV probes into LSPs from multiple sources (MP2P possibly with ECMP) is assumed to result in arrival rates at the LSP egress bursting at line rate.

14.2.2 Fast-Failure-Detection (FFD) PDU

The FFD PDU also provides a heartbeat mechanism similar to CV PDU but at a much faster rate. Y.1711 suggests that a LSP can be provisioned either with CV PDU or FFD PDU. CV PDU provides failure detection in order of 3 seconds whereas FFD PDU when provisioned can improve the failure detection time to 100 msecs range. FFD PDU can be selectively provisioned on LSPs requiring fast failure detection.

14.1.3 Forward and Backward Defect Indication (FDI & BDI)

The CV probe is augmented with defect notification PDUs, FDI for the forward direction, and BDI for the reverse direction. These are used for alarm suppression and control of performance measurement functions. BDI has limited applicability given that most LSPs are uni-directional, however it is very useful for interworking OAM with bi-directional PW clients (e.g. ATM).

14.3 Y.17fec-cv (ITU-T SG13/Q3)

A slightly more sophisticated probe type based upon Y.1711 protocol mechanisms is the Forwarding Equivalence Class Connectivity Verification (FEC-CV) PDU. FEC-CV, can carry aggregated LSP information (in the form of a bloom filter) such that a significant amount of configuration information can be verified in a single transaction. This is generally in the form of FEC information that functions as a functional description of the LSP. Simple boolean

operations on the bloom filter at the LSP egress can be used to detect misbranching while being tolerant of inbound filtering and other artifacts of network operations. The PDU can adapt to new applications via defining new coding rules for the FEC information, but no not require any changes to the actual PDU processing.

Y.17fec-cv is designed to complement existing link and node failure detection mechanisms by filling a fault detection gap in the MPLS OAM toolset as part of an overall operational framework. Unlike the Y.1711 CV or LSP-PING, it is not a self contained mechanism for detection of all faults or performing availability assessment.

15. Security Considerations

Support for intra-provider data plane OAM messaging does not introduce any new security concerns to the MPLS architecture. Though it does actually address some that already exist, i.e. through rigorous defect handling operator's can offer their customers a greater degree of integrity protection that their traffic will not be misdelivered (for example by being able to detect leaking LSP traffic from a VPN).

Support for inter-provider data plane OAM messaging introduces a number of security concerns as by definition, portions of LSPs will not be in trusted space, the provider has no control over who may inject traffic into the LSP. This creates opportunity for malicious or poorly behaved users to disrupt network operations. Attempts to introduce filtering on target LSP OAM flows may be problematic if flows are not visible to intermediate LSRs. However it may be possible to interdict flows on the return path between providers (as faithfulness to the forwarding path is not a return path requirement) to mitigate aspects of this vulnerability.

OAM tools may permit unauthorized or malicious users to extract significant amounts of information about network configuration. This would be especially true of IP based tools as in many network configurations, MPLS does not typically extend to untrusted hosts, but IP does. This suggests that tools used for problem diagnosis or which by design are capable of extracting significant amounts of information will require authentication and authorization of the originator. This may impact the scalability of such tools when employed for monitoring instead of diagnosis.

16. A summary of what can be achieved.

This draft identifies useful MPLS OAM capability that potentially could be provided via data plane OAM functions. In particular with respect to automatic fault detection and failure handling.

This draft suggests that it may be possible to provide this capability for any level in the label stack either by instrumenting that level, or instrumenting an overlay and provides an overview of the tools available to do so.

This draft also identifies that many aspects of performance management are intractable for some MPLS topological constructs. Any type of comparative measurement between an ingress and the egress of an LSP requires a 1:1 cardinality, or the ability of the egress to uniquely determine the ingress for each measured unit of communication, something that LSP merge, PHP and possible use of per platform label space at the measured LSP level undermine. Again a potential solution is to instrument a p2p overlay where such detailed measurements are required, and otherwise unavailable.

17. References

[ALLAN] Allan, D., "Guidelines for MPLS Load Balancing", draft-allan-mpls-loadbal-05.txt, IETF work in progress, October 2003

[ARCH] Bryant et.al. "PWE3 Architecture", draft-ietf-pwe3-arch-06.txt, IETF work in progress, October 2003

[DUBE] Dube, R., Costa, M. "Bi-directional LSPs for classical MPLS", draft-dube-bidirectional-lsp-00.txt, IETF work in progress, July 2002

[HIERARCHY] Lai et.al. "Network Hierarchy and Multilayer Survivability", draft-ietf-tewg-restore-hierarchy-00.txt, IETF Work in Progress, September 2001

[ICMP] Bonica et. al. "ICMP Extensions for MultiProtocol Label Switching", draft-ietf-mpls-icmp-02.txt, IETF Work in Progress, August 2000.

[KOMPELLA] Kompella et.al. "MPLS-based Layer 2 VPNs", draft-kompella-mpls-l2vpn-02.txt, IETF Work in Progress,

December 2000

[LSP-PING] Pan et.al. "Detecting Data Plane Liveliness in MPLS", draft-ietf-mpls-lsp-ping-03, IETF work in progress, June 2003

[LSR-TEST] Swallow et.al., "Label Switching Router Self-Test",

draft-ietf-mpls-lsr-self-test-00.txt, IETF Work in Progress, October 2003

[MARTINI]Martini et.al. "Pseudowire Setup and Maintenance using LDP", draft-ietf-pwe3-control-protocol-04.txt, IETF Work in Progress, October 2003

[MPLSDIFF] Le Faucheur et.al. "MPLS Support of Differentiated Services", IETF RFC 3270, May 2002

[MPLSREQS] Nadeau et.al., "OAM Requirements for MPLS Networks", draft-ietf-mpls-oam-requirements-01.txt, June 2003

[2547] Rosen, E. Rekhter, Y., "BGP/MPLS VPNs", IETF RFC 2547, March 1999

[SWALLOW] Swallow, G. and Goguen, R., "RSVP Label Allocation for Backup Tunnels", draft-swallow-rsvp-bypass-label-01.txt, November 2000

[TTL] Agarwal, P., and Akyol, B., "TTL Processing in MPLS Networks", IETF RFC 3443, January 2003

[VCCV] Nadeau et.al., "Pseudo Wire (PW) Virtual Circuit Connection Verification (VCCV)", draft-ietf-pwe3-vccv-00.txt, July 2003

[Y1710] ITU-T Recommendation Y.1710(2002), "Requirements for OAM Functionality for MPLS Networks"

[Y1711] ITU-T Recommendation Y.1711(2002), "OAM Mechanism for MPLS Networks"

[Y17FECCV] ITU-T Draft Recommendation Y.17fec-cv, "Misbranching Detection in MPLS Networks", Temporary Document TD25rev1 (WP3/13), July 2003

18. Editor's Address

David Allan
Nortel Networks
3500 Carling Ave.
Ottawa, Ontario, CANADA

Phone: 1-613-763-6362
Email: dallan@nortelnetworks.com