Split Multi-link Trunking (SMLT)

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Abstract

This document provides a description of Split Multi-Link Trunking (SMLT). SMLT is an architecture for reliable and redundant Layer 2 networks, which allows the full usage of all provisioned links that are dual-homing into a pair of aggregation devices in a MAC-bridged network. Multi-Link Trunking (MLT) is a link-aggregation mechanism similar to IEEE 802.3ad without the LACP (Link Aggregation Control Protocol). Typical redundant Ethernet networks consist of wiring closet (edge) switches dual-homed to network center aggregation (core) switches in a building or campus. Similarly, in carrier and ISP environments, CPE devices may be dual-homed to two SMLT aggregation switches at a Service Provider’s point of presence (POP). Traditionally, when a loop is created in the Layer 2 network by dual-homing devices, Spanning Tree Protocol must block one of the redundant network paths. SMLT avoids this by allowing all aggregated paths in a dual-homed switch configuration to be active and forwarding traffic simultaneously as well as providing very fast traffic fail-over in the event of a link failure.

Table of Content:

Status of this Memo..................................................1
Abstract...........................................................1
1. Conventions used in this document............................2
  2. Introduction....................................................2
  2.1 Reasons to deploy SMLT....................................3
  3. Definitions....................................................3

Lapuh, et. al           Informational            [Page 1]
4. How does SMLT work in a L2 network?............................4
4.1 IST Protocol...................................................6
5. Problems Solved................................................8
5.1 Layer-2 Traffic Load Sharing................................8
5.2 SMLT Configurations...........................................8
5.3 No single point of failure....................................9
6. Failure Scenarios.............................................9
6.1 Loss of SMLT link...........................................9
6.2 Loss of SMLT aggregation switch............................10
6.3 Loss of IST Link..............................................10
6.4 Loss of multiple SMLT aggregation switches in different SMLT aggregation switch pairs........................................10
6.5 Loss of all IST Links between an SMLT aggregation switch pair.10
7. SMLT’s relation with Spanning Tree/ Rapid Spanning Tree...11
8. Security Considerations........................................12
9. References.....................................................12
10. Acknowledgments.............................................12
11. Author’s Addresses........................................12
Full Copyright Statement........................................13

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 [2].

2. Introduction

This document provides a description of Split Multi-Link Trunking (SMLT) SMLT is an architecture for reliable and redundant Layer 2 networks, which allows the full usage of all provisioned paths and bandwidth in a MAC bridged network. Multi-Link Trunking (MLT) is a link-aggregation mechanism similar to IEEE 802.3ad without the LACP (Link Aggregation Control Protocol). Typical redundant Ethernet networks consist of wiring closet (edge) switches dual-homed to network center aggregation (core) switches in a building or campus. Similarly, in carrier and ISP environments, CPE devices may be dual-homed to two SMLT aggregation switches at a Service Provider’s point of presence (POP). Traditionally, when a loop is created in the Layer 2 network by dual-homing devices, Spanning Tree Protocol's point of view. SMLT avoids this by allowing all aggregation paths in a dual-homed switch configuration to be active and forwarding traffic simultaneously as well as providing very fast traffic fail-over in the event of a link failure. This is accomplished by implementing a method that allows two SMLT aggregation switches appear as a single
path aggregation end point device to dual-homed edge switches. The SMLT aggregation switches make use of an Inter-Switch Trunk (IST) between them over which they exchange information, permitting rapid fault detection and forwarding path modification. Although SMLT is primarily designed for Layer 2 Ethernet networks, it also provides benefits for Layer 3 networks as well.

2.1 Reasons to deploy SMLT

As networks grow ever more critical there is a demand for multiple paths from all wiring closet switches into the core of the network in order to eliminate all single points of failure, such that in most failure cases there is no permanent loss of connectivity and the recovery from most failures is in the sub second area. Moreover, it is highly desirable that the elimination of single points of failure does not result in unused capacity (often very costly) and perhaps more importantly, that rerouting around failures is as fast as possible.

Further requirements that the solution is simple to implement and as transparent as possible and interoperable with the majority of existing wiring closet/CPE/edge devices, are key advantage that SMLT presents compared to previous attempts to solve this problem.

3. Definitions

Before describing how SMLT works in detail it is necessary to define various terms.

**SMLT Client Switch**
A switch located at the edge of the network. End stations typically connect directly (or indirectly through a repeater) to a wiring closet switch.

**SMLT Aggregation Switch**
A switch that connects to multiple SMLT client switches, edge switches or CPE devices. Such SMLT Aggregation Switches may be owned by the customer in the customer's enterprise network or may be owned by Service Provider. SMLT applies to these different deployment scenarios.

**IST (Inter Switch Trunk)**
IST consists of one or more parallel point-to-point links that connect two SMLT aggregation switches together. The two SMLT aggregation switches utilize this channel to share information so that they may operate as a single logical switch.

**MLT (Multi-Link Trunk)**
MLT is a method of link aggregation that allows multiple Ethernet trunks to be aggregated together in order to provide a single logical trunk. An MLT provides the combined bandwidth of the
multiple links, as well as the physical layer protection against the failure of any single link.

A Multi-Link Trunk can be any kind of link aggregation mechanism. There are currently several methods used in the industry such as, but not restricted to IEEE 802.3ad, etc. The current proposal is interoperable with 802.3ad in static mode (Link aggregation control protocol LACP disabled). There is however no restriction to SMLT that would preclude it to be extended to support LACP of 802.3ad. This is currently not part of this proposal.

SMLT (Split Multi-Link Trunk)
A multi-link trunk with connection across an aggregated set of ports, where each link of the multi-link trunk connects a pair of ports on different devices (e.g. SMLT client switch and SMLT Aggregation Switch), such that one end of the trunk is split between two SMLT aggregation switches.

SMLT Client
A switch located at the edge of the network, such as in a wiring closet or CPE. An SMLT Client switch must be able to perform link aggregation (such as with MLT or some other compatible method) but does not require any SMLT intelligence.

4. How does SMLT work in a L2 network?

![Figure 1: Reference Network for S-MLT](image-url)

Figure 1 illustrates a configuration that includes a pair of SMLT Aggregation switches E and F, and four separate SMLT client switches.
A, B, C, and D. At least B and C are MLT-compatible devices. SMLT client switches B and C are connected to the SMLT aggregation switches via multi-link trunks that are split between the two SMLT aggregation switches. For example, SMLT client switch B may use two parallel links for its connection to E, and two additional parallel links for its connection to F.

SMLT client switch C may have only a single link to both E and F. As shown in the illustration, Switch A is also configured for MLT, but the MLT terminates on only one switch in the network core. Switch D has a single connection to the core. Although both switch A and switch D could also be configured to terminate across both of the SMLT aggregation switches using SMLT, neither switch will benefit from the advantages of SMLT in the configuration shown.

As seen in this example, implementation of SMLT only requires two SMLT capable aggregation switches. Those switches must be connected via an IST (Inter Switch Trunk.)

The SMLT aggregation switches use IST for three purposes:

- Firstly, they use it to confirm that each switch is alive as well as exchange learned MAC address information; therefore this link must be reliable and not exhibit a single point of failure itself.

- Secondly, it is used for the forwarding of flooded packets or packets destined for non-SMLT connected switches or servers physically connected to the other SMLT aggregation switch.

- Thirdly it is used to forward traffic in case of a SMLT link failure.

If all devices that are connected to the SMLT aggregation switches are dual-homed (like switches B and C) the traffic on the IST only consists of the IST control packet exchanges in normal operation. If devices are single homed like A and D in average 50% of the traffic will use the IST link in normal operation.

In case of a SMLT link failure the IST link is used as a backup link to forward traffic to the destinations. These requirements dictate that the IST should preferably be but are not restricted to a multi-gigabit MLT with connections between both SMLT aggregation switches in order to ensure that there is no single point of failure in the IST.

In case of an IST failure and all devices are dual-homed no data traffic will be lost and the network remains intact.

The SMLT client switches are dual-homed to the two SMLT aggregation switches but they require no knowledge of whether they are connected...
to a single switch or to two switches. SMLT Intelligence is required only on the SMLT aggregation switches. Logically, they appear as a single switch to the edge switches.

Figure 1 also includes end stations connected to each of the switches: a, b1, b2, c1, c2, and d may be hosts; e and f may be hosts, servers or routers.

SMLT client switches B and C may use any method for determining which link of their multi-link trunk connections to use for forwarding a packet, so long as the same link is used for a given flow. This requirement ensures that there will be no out-of-sequence packets between any pair of communicating devices. SMLT aggregation switches will always send traffic direct to a SMLT client switch and only use the IST for traffic that they cannot forward another more direct way. The examples below will explain the process.

Traffic from a to b1 and/or b2 — assuming a and b1/b2 are communicating via layer 2 — goes from switch A to switch E and then forwarded up its direct link to switch B. Traffic coming down from b1 or b2 to a is sent by switch B on one of its MLT ports. Since it does not know that the MLT is anything special it may send traffic from b1 to a on the link to switch E and the traffic from b2 to a on the link to switch F. In the case of traffic from b1, switch E just forwards the traffic directly to switch A, while traffic from b2, which arrived at switch F, is forwarded across the IST to switch E and then to switch A.

Traffic from b1/b2 to c1/c2 will always be sent by switch B down its MLT to the core. No matter which switch (E or F) it arrives at it will be sent directly to switch C through the local link. This is the reason why dual-homing all client switches to the SMLT aggregation pair, will reduce the amount of possible traffic load on the IST link. A single IST failure (all SMLT links active) in this scenario will also not cause traffic interruption. This minimizes the risk of network downtime even further.

Traffic from a to d and vice versa is forwarded across the IST because it is the shortest path, but this is treated purely as a standard link with no account taken of SMLT and the fact that it is also an IST.

Finally traffic from f to c1/c2 will be sent out direct from switch F. Return traffic from c1/c2 to f will be passed across the IST if switch C sends it down the link to the switch E.

4.1 IST Protocol
The IST link connecting two SMLT aggregation switches in Figure 1 runs a protocol that allows the following messages to be exchanged between the two SMLT aggregation switches:

IST Hello
Each SMLT aggregation switch periodically transmits and listens for IST Hello messages on its IST ports where IST ports are those ports that connect IST link. These messages indicate the following:
- SMLT aggregation switch port from which the IST Hello message is being sent is set to type IST
- Whether the sending port is receiving IST Hello messages from the other end
- The expected time interval between IST Hello messages that are received from the other end.

SMLT Status
Each SMLT aggregation switch periodically reports SMLT Status to the other SMLT aggregation switch. This may include the following:
- SMLT ID - this value is assigned to a SMLT by an SMLT aggregation switch when the SMLT is configured on the SMLT aggregation switch. It provides a reference to each SMLT client switch on an SMLT aggregation switch.
- SMLT Status - Status of SMLT can be either 1) in service (one or more of the SMLT links on an SMLT aggregation switch are operating), or 2) out of service (all members of the SMLT links on an SMLT aggregation switch are not operating).

Learned or Migrated MAC addresses
When an SMLT aggregation switch learns a new MAC address on any of its ports, it notifies the other SMLT aggregation switch about the MAC address value and the port type of the port at which the address was learned. When address is learned against a SMLT port, SMLT ID is also passed.

MAC address aged out
When the age expires for a MAC address assigned to a non-SMLT port, the SMLT aggregation switch deletes the MAC address record and sends a message to the other SMLT aggregation switch to report the event. The other SMLT aggregation switch deletes its record.

SMLT address aged out
When the age expires for a MAC address assigned to a SMLT port, the SMLT aggregation switch does not delete the MAC address record. It does mark the address as having expired locally, sends a message to the other SMLT aggregation switch and waits to receive confirmation from the other switch. When the other switch receives the message, it marks the address as having expired remotely, but waits to delete the MAC address record only after it expires locally.

Local Bridge ID
This message is sent by an SMLT aggregation switch to notify the other switch of the value of its own Bridge Identifier. This permits
each SMLT aggregation switch to compare its ID with the other switch's Bridge ID, so they may determine who will be Root Bridge, and the values of BPDU parameters used on the SMLTs.

The above messages can be mapped to any existing protocols e.g. 802.3ad’s LACP etc.

5. Problems Solved

5.1 Layer-2 Traffic Load Sharing

Load Sharing from the SMLT client perspective, is achieved by the MLT path selection algorithm used on the edge switch. Usually this is done on a SRC/DST MAC address basis but other techniques can be used.

Load sharing from the SMLT aggregation switch perspective, is achieved by sending all traffic destined to the SMLT client switch directly and not over the IST trunk. The IST trunk is normally not used to send/receive traffic to/from a SMLT dual-homed SMLT client. Traffic received on the IST by an SMLT aggregation switch is not forwarded on SMLT links because the other SMLT aggregation switch will have performed that job, thus eliminating the possibility of a loop in the network. The only exception to this rule is, if the SMLT links on the peer aggregation switch are down, then traffic received over IST will be forwarded to the corresponding SMLT on this switch.

5.2 SMLT Configurations

SMLT may be used within the core of the network. It is also possible to configure SMLT groups in a square or full mesh scenario, but in this case both sides of the link are configured for SMLT.

SMLT Square with two SMLT aggregation pairs facing each other

```
+-----+    +-----+
| E   |----| F   |
+-----+    +-----+
      ||          |
      ||          |
      ||          |
      ||          |
      +---+      +---+
| B   |------| C   |
      +---+      +---+
```
SMLT full mesh with two SMLT aggregation pairs facing each other

```
+-----+    +-----+
|  E  |----|  F  |
+-----+    +-----+
  \   /   
  |  /    |
  | /     |
  |/      |
  +---+    +---+
  | B |------| C |
  +---+    +---+
```

This is possible because there is no state information passed across the MLT link and thus both ends believe that the other end is a single switch. The result of this is that there has been no loop introduced into that network and any of the core switches or any of the connecting links between them may fail and the network will rapidly recover.

It is furthermore possible to scale SMLT groups to achieve hierarchical network designs by connecting SMLT groups together thus allowing building redundant loop free L2 domains without Spanning Tree and still fully using all network links.

5.3 No single point of failure

Any single link or either SMLT aggregation switch can fail and recovery will take place in less than 1 second.

Note that this number is conservative depending on the implementation and the link loss detection mechanisms network might experience loss for less than 1 second. See the analysis below for further details.

6. Failure Scenarios

6.1 Loss of SMLT link

The SMLT client switch detects link failure and sends traffic on the other SMLT link(s) just as is done with standard MLT. Detection and fail-over is dependent on how quickly the device can detect link failures.
If the link is not the only one between the SMLT client and SMLT aggregation switches in question then the SMLT aggregation switch also just uses standard MLT detection and rerouting to move traffic to the remaining links. If the link is the only one to the SMLT aggregation switch then on failure detection the switch informs the other SMLT aggregation switch of SMLT trunk loss. The other SMLT aggregation switch then treats the SMLT trunk as a regular MLT trunk. If the link is reestablished, the SMLT aggregation switches detect this and move the trunk back to regular SMLT operation.

6.2 Loss of SMLT aggregation switch

The SMLT client switch detects link failure and sends traffic on the other SMLT link(s) just as with standard MLT. The operational SMLT aggregation switch detects loss of partner (IST and keep alive packets lost) and changes all the SMLT trunks to regular MLT trunks. If the partner returns, the operational SMLT aggregation switch detects this (IST becomes active) and moves the trunks back to regular SMLT operation once full connectivity is reestablished.

6.3 Loss of IST Link

The SMLT client switches do not detect a failure and communicate as usual.

In normal use, there will be more than one link in the IST (as it may itself be an aggregated link. Thus IST traffic resumes over the remaining links in the IST.

6.4 Loss of multiple SMLT aggregation switches in different SMLT aggregation switch pairs

Note that in this case one may have exceeded the goal of providing connectivity only after a single failure since for this to happen multiple failures must occur.

The SMLT client switches do not detect a failure and communicate as usual.

Since each SMLT aggregation switch pair is a separate entity, each is unaffected by failures elsewhere. Thus connectivity is unaffected, although, since available bandwidth is drastically reduced, packet loss and increased latency may occur.

6.5 Loss of all IST Links between an SMLT aggregation switch pair
Note that in this case one may have exceeded the goal of providing connectivity only after a single failure since for this to happen multiple failures must occur.

In the event that all links in the IST fail, the SMLT aggregation switches do not see each other anymore (keep alive lost) and both assume that their partner is dead. However for the most part there are no ill effects in the network if all SMLT client switches are dual homed to the SMLT aggregation switches.

7. SMLT’s relation with Spanning Tree/ Rapid Spanning Tree

SMLT is an architecture that fits between the MAC control layer and the 802.1D/w on top of 802.3ad.

```
+------------------+
| 802.1D/w         |
+------------------+

+------------------+
|       SMLT       |
+------------------+

+------------------+
| 802.3ad          |
+------------------+
```

Therefore the (Rapid) Spanning Tree Protocol can be supported on top of SMLT path aggregation architecture.

Link Failures will therefore not anymore trigger any STP reconvergence because the logical link remains still intact as long as one SMLT link out of a group is active.

SMLT as an underlying path aggregation architecture underneath a (Rapid) Spanning Tree design has following advantages:

SMLT does not have any blocking links. Therefore all configured bandwidth is available for traffic forwarding.

SMLT’s IST protocol is used only on a set of two switches, the SMLT aggregation pair. The protocol, therefore, does not have any inherent delays because the two neighbors are directly connected.

SMLT convergence targets are sub-second in every failure scenario. There is no root bridge election and therefore long re-election is not an issue.
SMLT link failures don’t generate TCNs as long as one logical link out of an SMLT group is still active – therefore flooding is only limited to the SMLT aggregation pair.

8. Security Considerations

This document does not introduce any new security issues. From a customer point of view, the SMLT looks to be the same as IEEE 802.3ad link aggregation. From the Service Provider point of view, no security issues are expected since the IST communication occurs between SMLT aggregation switches located inside the same autonomous SP network. When the two SMLT switches are located in two different autonomous networks, there may be some security issues, e.g. sharing of bridging information. However, it is not expected that the SMLT aggregation switches will be deployed in two SP networks.

9. References


10. Acknowledgments

The authors would like to thank Wassim Tawbi and Yili Zhao for their contributions and furthering the content of SMLT.

11. Author’s Addresses

Roger Lapuh
Nortel Networks
Wilstrasse 11 Building U95
Switzerland 8610
Phone: +1 (408) 495 1599
Email: rlapuh@nortelnetworks.com

Dinesh Mohan
Nortel Networks
P O Box 3511 Station C
Ottawa ON K1Y 4H7 Canada