Interfaces

In This Chapter

This chapter provides information about configuring chassis slots, cards, and ports. Topics in this chapter include:

Configuration Overview

NOTE: This document uses the term preprovisioning in the context of preparing or preconfiguring entities such as chassis slots, cards, input/output modules (IOMs)/Control Forwarding Module (CFM/IOM) cards and media dependent adapters (MDAs), media dependent adapters (MDAs), compact media adapters (CMAs), ports, and interfaces, prior to initialization. These entities can be installed but not enabled. When the entity is in a no shutdown state (administratively enabled), then the entity is considered to be provisioned.

Alcatel-Lucent routers provide the capability to configure chassis slots to accept specific line card and MDA types and set the relevant configurations before the equipment is actually installed. The preprovisioning ability allows you to plan your configurations as well as monitor and manage your router hardware inventory. Ports and interfaces can also be preprovisioned. When the functionality is needed, the card(s) can be inserted into the appropriate chassis slots when required.

The following sections are discussed.

- Chassis Slots and Cards on page 19
- MDAs on page 20
- Ports on page 35

Chassis Slots and Cards

To pre-provision a chassis slot, the line card type must be specified. System administrators or network operators can enter card type information for each slot, allowing a range of card types in particular slots. From the range of card types, a card and accompanying MDAs/CMAs are specified. When a card is installed in a slot and enabled, the system verifies that the installed card type matches the allowed card type. If the parameters do not match, the card remains off line. A preprovisioned slot can remain empty without conflicting with populated slots.

SR-7/SR-12 and ESS-7/ESS-12 systems accept Input/Output Modules (IOM) cards. These IOM cards have two slots which accept MDA modules. The SR-c12 and SR-c4 systems do not accept IOMs. SR-c12 and SR-c4 systems accept MDAs using an MDA Carrier Modules. SR-c12 and SR-c4 systems also accept Compact Media Modules (CMAs) directly without the need for MCMs. Refer to the appropriate system installation guide for more information.

MCMs

The following features are not applicable to the 7450-ESS even when in mixed mode.

An MCM (MDA Carrier Module) slot must be configured before an MDA (Media Dependant Adapter) can be provisioned. If you provision an MDA type before an MCM slot is configured, it is assumed you are provisioning a Compact Media Adapter (subscriber/SAP/spoke SDP). CMAs do not require MCM pre-configuration. Up to six MCMs may be provisioned on a 7750 SR-c12. Up to two MCMs may be provisioned on a on a 7710 SR-c4. Even numbered slots are invalid for MCM installation (MCMs physically span 2 slots; “mcm 1” spans slots 1 and 2)

Refer to the CMA Installation Guide(s) and MDA Installation Guide(s) for more information on the physical characteristics of each card.

MDAs

A chassis slot and card type must be specified and provisioned before an MDA can be preprovisioned. An MDA is provisioned when a type designated from the allowed MDA types is inserted. A preprovisioned MDA slot can remain empty without conflicting with populated slots.

Once installed and enabled, the system verifies that the installed MDA type matches the configured parameters. If the parameters do not match, the MDA remains offline.

A chassis slot, card type and MCM must be specified and provisioned before an MDA can be preprovisioned. An MDA is provisioned when a type designated from the allowed MDA type is inserted. A preprovisioned MDA slot can remain empty without conflicting with populated slots. Up to six MDAs may be provisioned on a 7750 SR-c12. Even numbered slots are invalid for MDA installation (MDAs physically span 2 slots; “mda 1” spans slots 1 and 2).

MDA output displays an “m” in the name of the card. The following displays a show card state command. In this example, an **m60-10/100eth-tx** MDA is installed in slot 1.

```
A:ALU-3>config>card# show card state
===============================================================================
Card State
===============================================================================
Slot/  Provisioned      Equipped         Admin Operational   Num   Num Comments
Id     Type             Type             State State         Ports MDA
-------------------------------------------------------------------------------
1      iom-xp           iom-xp           up    up                  12
1/1    mcm-xp           mcm-xp           up    up
1/3                     mcm-xp           up    unprovisioned
1/1    m60-10/100eth-tx m60-10/100eth-tx up    up
1/5    c8-10/100eth-tx  c8-10/100eth-tx  up    up
1/6                     c1-1gb-sfp       up    unprovisioned
1/7                     c8-chds1         up    unprovisioned
1/8                     c4-ds3           up    unprovisioned
```

```
1/9                    c8-10/100eth-tx  up    unprovisioned
1/10                   c1-1gb-sfp       up    unprovisioned
1/11                   c8-chds1         up    unprovisioned
1/12                   c4-ds3           up    unprovisioned
A      cfm-xp          cfm-xp           up    up                       Active
B      cfm-xp                           up    down                     Standby
===============================================================================
A:ALU-3>config>card#
```

Once installed and enabled, the system verifies that the installed MDA type matches the configured parameters. If the parameters do not match, the MDA remains offline.

CMAs

CMAs (Compact Media Adapter) are configured and provisioned in the same manner as MDAs (Media Dependent Adapter). 7750 SR-c12 and SR-c4 systems accept CMAs. Up to eight CMAs may be provisioned on a 7750 SR-c12, and up to 4 CMAs may be provisioned on an SR-c4. Up to four CMAs may be provisioned on a 7710 SR-c4. CMA output displays a “c” in the name of the card. The following displays **show card state** command output. In this example, a **c8-10/100eth-tx** CMA is installed in slot 5.

```
A:7750-3# show card state
===============================================================================================
Card State
===============================================================================================
Slot/  Provisioned       Equipped         Admin Operational    Num   Num Comments
ID     Type              Type             State State          Ports MDA
-----------------------------------------------------------------------------------------------
1      iom-xp            iom-xp             up    up                 12
1/5    c8-10/100eth-tx   c8-10/100eth-tx    up    up           8
1/6    c8-10/100eth-tx   c8-10/100eth-tx    up    up           8
1/7                      c8-chds1           up    unprovisioned
1/8                      c4-ds3             up    unprovisioned
1/9                      c8-10/100eth-tx    up    unprovisioned
1/10                     c1-1gb-sfp         up    unprovisioned
1/11                     c8-chds1           up    unprovisioned
1/12                     c4-ds3             up    unprovisioned
A      cfm-xp            cfm-xp             up    up                       Active
B      cfm-xp                               up    provisioned              Standby
===============================================================================================
A:7750-3#
```

```
A:7710-3# show card state
===============================================================================================
Card State
===============================================================================================
Slot/  Provisioned       Equipped         Admin Operational    Num   Num Comments
ID     Type              Type             State State          Ports MDA
-----------------------------------------------------------------------------------------------
1      iom-12g           iom-12g            up    up                 12
1/5    c8-10/100eth-tx   c8-10/100eth-tx    up    up           8
1/6    c8-10/100eth-tx   c8-10/100eth-tx    up    up           8
1/7                      c8-chds1           up    unprovisioned
1/8                      c4-ds3             up    unprovisioned
1/9                      c8-10/100eth-tx    up    unprovisioned
```

```
1/10                    c1-1gb-sfp          up    unprovisioned
1/11                    c8-chds1            up    unprovisioned
1/12                    c4-ds3              up    unprovisioned
A      cfm-12g          cfm-12g             up    up                       Active
B      cfm-12g                              up    provisioned              Standby
===============================================================================
A:7710-3#
```

A preprovisioned CMA slot can remain empty without conflicting with populated slots.

Once installed and enabled, the system verifies that the installed CMA type matches the configured parameters. If the parameters do not match, the CMA remains offline.

Note: On the E3 CMA, bit stuffing is not supported in G.751 framing mode. All of the 12 justification service bits and the 4 justification bits contain valid data on the transmitted signal. Incoming bitstreams should contain valid data in the 12 justification service bits and 4 justification bits, otherwise the link will not function.

Versatile Service Module (VSM)

The Versatile Service Module (VSM) is a module that allows operators to internally connect a VPLS or VLL service into an IES or IPVPN service. Each module is capable of 10 Gbps throughput.

This module is provisioned as a Cross Connect Adaptor (CCA). Unlike external port connections which utilize two TX-RX paths, a CCA interconnects the egress forwarding path on the IOM directly to the ingress forwarding path. This eliminates the need for the physical port MAC, PHY, cable and other MDA-specific components producing a less costly and more reliable adaptor. The complete 10G+ forwarding path is available allowing single conversations up to 10G.

Bandwidth is utilized in a more efficient manner than with externally cabled ports. Typically, the offered load presented to each side of the cross connect port pair is asymmetric in nature. When physical ports are used to cross connect services, each service is egress bandwidth limited to the link speed of the TX-RX path it is using. If one TX-RX path is under utilized, egress services on the other path cannot make use of the available bandwidth.

Since the CCA is forwarding all services over the same path, all the available bandwidth may be used. An example of this would be a two services connected over a CCA. Service A is a VPLS. Service B is an IES. There are two directions of traffic between the pair, A to B and B to A. Traffic in both directions travels across the CCA in the same path. The total bandwidth the CCA can forward is 10 Gbps. Therefore, A to B could consume 7 Gbps, and B to A could consume 3 Gbps. Any combination of services and traffic directions adding up to 10 Gbps can be supported on a single CCA.

The forwarding plane the CCA interconnects maintains the complete egress and ingress features of the services it is interconnecting. This includes the ability to remap QoS, enforce policing and shaping and provide ingress and egress accounting for each service.

In addition CCAs may be placed into Cross Connect Aggregation Groups (CCAGs). A CCAG provides a mechanism to aggregate multiple CCAs into a single forwarding group.

The CCAG uses conversation hashing to dynamically distribute cross connect traffic to the active CCAs in the aggregation group. In the event that an active CCA fails or is removed from the group, the conversation hashing function will redistribute the traffic over the remaining active CCAs within the group. The conversation hashing mechanism performed for a CCAG is identical to the hashing functions performed for Ethernet LAGs (Link Aggregation Groups).

The VSM module is not supported on 7750 SR-c12/c4 platforms.

Oversubscribed Ethernet MDAs

The 7750 SR and 7450 ESS support oversubscribed Ethernet MDAs. These have more bandwidth towards the user than the 10 Gbps capacity between the MDA and IOM.

A traffic management function is implemented on the MDA to control the data entering the IOM. This function consists of two parts:

- Rate limiting
- Packet classification and scheduling

Rate Limiting

The oversubscribed MDA/CMA limits the rate at which traffic can enter the MDA/CMA on a per port basis. If a port exceeds its configured limits then the excess traffic will be discarded, and 802.3x flow control frames (pause frames) are generated.

Packet Classification and Scheduling

The classification and scheduling function implemented on the oversubscribed MDA/CMA ensures that traffic is correctly prioritized when the bus from the MDA/CMA to the IOM is overcommitted. This could occur if the policing parameters configured are such that the sum of the traffic being admitted into the MDA/CMA is greater than 10 Gbps.

The classification function uses the bits set in the DSCP or Dot1p fields of the customer packets to perform classification. It can also identify locally addressed traffic arriving on network ports as Network Control packets. This classification on the oversubscribed MDA/CMA uses following rules:

- If the service QoS policy for the SAP (port or VLAN) uses the default classification policy, all traffic will be classified as Best Effort (be).
- If the service QoS policy for the SAP contains a Dot1p classification, the Dot1p field in the customer packets is used for classification on the MDA/CMA.
- If the service QoS policy for the SAP contains a DSCP classification, the DSCP field in the customer packets is used for classification on the MDA/CMA.
- If a mix of Dot1p and DSCP classification definitions are present in the service QoS policy then the field used to perform classification will be the type used for the highest priority definition. For example, if High Priority 1 is the highest priority definition and it specifies that the DSCP field should be used, then the DSCP field will be used for classification on the MDA/CMA and the Dot1p field ignored.

- If the service QoS policy for the SAP specifies IP or MAC filters for forwarding class identification, then traffic will be treated as Best Effort. Full MAC or IP classification is not possible on the MDA/CMA (but is possible on the IOM).
- The packet is classified into 16 classes. Typically, these are the eight forwarding classes and each packet is assigned one priority per forwarding class. After classification, the packet is offered to the queuing model. This queuing model is limited to three queues each having four thresholds. These thresholds define whether an incoming packet, after classification, is accepted in the queue or not. Table 1 displays typical mapping of classes onto queues/threshold.

Table 1: Typical Mapping Of Classes Onto Queues/Threshold

Counter	{Queue	Threshold	Traffic Class}
0	{2	3	"fc-nc / in-profile"}
1	{2	2	"fc-nc / out-profile"}
2	{2	1	"fc-h1 / in-profile"}
3	{2	0	"fc-h1 / out-profile"}
4	{1	3	"fc-ef / in-profile"}
5	{1	2	"fc-ef / out-profile"}
6	{1	1	"fc-h2 / in-profile"}
7	{1	0	"fc-h2 / out-profile"}
8	{0	3	"fc-l1 / in-profile"}
9	{0	3	"fc-l1 / out-profile"}
10	{0	2	"fc-af / in-profile"}
11	{0	2	"fc-af / out-profile"}
12	{0	1	"fc-l2 / in-profile"}
13	{0	1	"fc-l2 / out-profile"}
14	{0	0	"fc-be / in-profile"}
15	{0	0	"fc-be / out-profile"}

A counter is associated with each mapping. Note that the above is an example and is dependent on the type of classification (such as dscp-exp, dot1p, etc.). When the threshold of a particular class is reached, packets belonging to that class will not be accepted in the queue. The packets will be dropped and the associated counter will be incremented.

The scheduling of the three queues is done in a strict priority, highest priority basis is associated with queue 0. This means that scheduling is done at queue level, not on the class that resulted from the classification. As soon as a packet has been accepted by the queue there is no way to differentiate it from other packets in the same queue (for example, another classification result not exceeding its threshold). All packets queued in the same queue will have the same priority from a scheduling point of view.

Channelized MDA/CMA Support

Channelized DS-1/E-1 CMA

Each 8-port channelized DS-1/E-1 CMA supports channelization down to DS-0. Each 8-port channelized DS-1/E-1 CMA supports 64 channel groups. This CMA is not supported on the 7450-ESS.

Channelized DS-3/E-3 MDA

Each 4-port or 12-port channelized DS-3/E-3 media dependent adapter (MDA) supports channelization down to digital signal level 0 (DS-0) using a maximum of 8 or 24 (respectively) 1.0/2.3 coaxial connectors. Each port consists of one receive (RX) coaxial connector and one transmit (TX) coaxial connector.

Each physical DS-3 connection can support a full clear-channel DS-3, or it can be channelized into independent DS-1/E-1 data channels. Each DS1/E1 channel can then be further channelized down to DS-0s. E-3 ports do not support channelization. They only support clear channel operation.

Each DS-3/E-3 MDA supports 512 channels with DS-0 timeslots that are used in the DS-1/E-1 channel-group.

This MDA is not supported on the 7450-ESS.

Channelized CHOC-12/STM-4 MDA

Each 1-port channelized OC-12/STM-4 MDA supports channelization down to DS-0 and accepts one OC-12/STM-4 SFP small form factor pluggable (SFP) module. The same SFP optics used on Alcatel-Lucent's SONET/SDH cards can be used on the channelized OC-12/STM-4 MDA.

Each channelized OC-12/STM-4 supports 512 channels with DS-0 timeslots that are used in the DS-1/E-1 channel-group. DS-3 TDM channels can be further channelized to DS-1/E-1 channel groups. An E3 TDM channel cannot be channelized and can only be configured in clear channel operation.

Channelized CHOC-3/STM-1 MDA

Each 4-port channelized OC-3/STM-1 MDA supports channelization down to DS-0 and accepts one OC-3/STM-1 SFP small form factor pluggable (SFP) module. The same SFP optics used on Alcatel-Lucent's SONET/SDH cards can be used on the channelized OC-3/STM-1 MDA.

Each channelized OC-3/STM-1 supports 512 channels with DS-0 timeslots that are used in the DS-1 channel-group. DS-3 TDM channels can be further channelized to DS-1/E-1 channel groups. An E3 TDM channel cannot be channelized and can only be configured in clear channel operation.

This MDA is not supported on the 7450-ESS.

Channelized Any Service Any Port (ASAP) CHOC-3/STM-1

Each port for the channelized ASAP OC-3/STM-1 MDA supports channelization down to DS-0 and accepts one OC-3/STM-1 SFP small form factor pluggable (SFP) module. The same SFP optics used on Alcatel-Lucent's SONET/SDH MDAs can be used on the channelized ASAP OC-3/STM-1 MDA.

Each channelized OC-3/STM-1 supports up to 512 channels with DS-0 timeslots with per channel encapsulation configuration (for example, Frame Relay, PPP, cHDLC, ATM). DS-3 TDM channels can be further channelized to DS-1/E-1 channel groups. An E3 TDM channel cannot be channelized and can only be configured in clear channel operation. The MDA is based on a programmable data path architecture that enables enhanced L1 and L2 data path functionality, for example ATM TM features, MDA-based channel/port queuing, or multilink applications like Inverse ATM Multiplexing (IMA).

Channelized OC-12/STM-4 ASAP MDAs

The 4-port channelized OC-12/STM-4 variant of the ASAP MDAs have features and channelization options similar to the 4-port channelized OC-3/STM-1 ASAP MDA.

DS-3 TDM channels can be further channelized to DS-1/E-1 channel groups. An E-3 TDM channel cannot be channelized and can only be configured in clear channel operation.

Channelized DS-3/E-3 ASAP MDA (4-Port)

The 4-port MDA provides 4 ports configurable as DS-3 or E-3. The MDA has eight (8) 1.0/2.3 connectors and accepts up to eight (8) DS-3/E-3 coax patch cables.

Each physical DS-3 connection can support a full clear-channel DS-3, or it can be channelized into independent DS-1/E-1 data channels. Each DS-1/E-1 channel can then be further channelized down to DS-0s. E-3 ports do not support channelization, only clear channel operation.

Channelized DS-3/E-3 ASAP MDA (12-Port)

The 12-port MDA provides 12 ports configurable as DS-3 or E-3. The MDA has twenty-four (24) 1.0/2.3 connectors and accepts up to twenty-four (24) DS-3/E-3 coax patch cables.

Each physical DS-3 connection can support a full clear-channel DS-3, or it can be channelized into independent DS-1/E-1 data channels. Each DS-1/E-1 channel can then be further channelized down to DS-0s. E-3 ports do not support channelization, only clear channel operation.

Channelized OC-3/STM-1 Circuit Emulation Services (CES) CMA and MDA

The channelized OC-3/STM-1/OC-12/STM-4 CES MDAs (c1-choc3-ces-sfp / m1-choc3-ces-sfp, m4-choc3-ces-sfp, m1-choc12-ces-sfp) provide an industry leading consolidation for DS-1, E-1 and n*64kbps for CES. The CES MDAs are supported on IOM-2 and IOM-3XP in the 7750 SR.

The channelized OC-3/STM-1/OC-12/STM-4 CES CMA/MDAs support CES. Circuit emulation services are interoperable with the existing 7705 SAR and 7250 SAS circuit emulation services. They are also interoperable with the 1850 TSS-5 circuit emulation services.

Two modes of circuit emulation are supported, unstructured and structured. Unstructured mode is supported for DS-1 and E-1 channels as per RFC4553 (SAToP). Structured mode is supported for n*64 kbps circuits as per RFC 5086, *Structure-Aware Time Division Multiplexed (TDM) Circuit Emulation Service over Packet Switched Network (CESoPSN)*. In addition, DS-1, E-1 and n*64 kbps circuits are also supported as per MEF8, *Circuit Emulation Services over Ethernet (CESoETH)* (Oct 2004). TDM circuits are optionally encapsulated in MPLS or Ethernet as per the applicable standards.

All channels on the CES CMA/MDA are supported as circuits to be emulated across the packet network. This includes DS-1, E-1 and n*64 kbps channels. Structure agnostic mode is supported for DS-1 and E-1 channels. Structure aware mode is supported for n*64 kbps channel groups in DS-1 and E-1 carriers. N*64 kbps circuit emulation supports basic and Channel Associated

Signaling (CAS) options. CAS configuration must be identical for all channel groups on a given DS-1 or E-1.

Circuits encapsulated in MPLS will use circuit pipes (Cpipes) to connect to the far end circuit. Cpipes support either SAP-spoke SDP or SAP-SAP connections.

Circuits encapsulated in Ethernet can be selected as a SAP in Epipes. Circuits encapsulated in Ethernet can be either SAP-spoke SDP or SAP-SAP connections for all valid Epipe SAPs. An EC-ID and far-end destination MAC address must be configured for each circuit.

Each OC-3/STM-1 port can be independently configured to be loop-timed or node-timed. Each OC-3/STM-1 port can be configured to be a timing source for the node. Each DS-1 or E-1 channel can be independently configured to be loop-timed, node-timed, adaptive-timed, or differential-timed. One adaptive timed circuit is supported per CMA/MDA. The CES circuit configured for adaptive timing can be configured to be a timing source for the node. This is required to distribute network timing to network elements which only have packet connectivity to network.

On the 7750 SR-c12 CES CMA, a BITS port is also provided. The BITS port can be configured as one reference sources (ref1, ref2) in the system timing subsystem.

Network Interconnections

With the introduction of Alcatel-Lucent's 7750 SR, the SR-Series product family can fill the needs of smaller service providers as well as the more remote point of presence (PoPs) locations for larger service providers. To support the use of lower speed links as network links in the likelihood that lower speed circuits are used as network or backbone links, the 7750 SR-Series supports a DS-1/E-1/DS-3/E-3 port (ASAP MDAs) or channel and an MLPPP bundle (ASAP MDAs) as network ports to transport and forwarding of all service types. This feature allows service providers to use lower speed circuits to interconnect small PoPs and CoS that do not require large amounts of network/backbone bandwidth.

Digital Diagnostics Monitoring

Some Alcatel-Lucent SFPs, XFPs, QSFPs, CFPs and the MSA DWDM transponder have Digital Diagnostics Monitoring (DDM) capability where the transceiver module maintains information about its working status in device registers including:

- Temperature
- Supply voltage
- Transmit (TX) bias current
- TX output power
- Received (RX) optical power

For the case of QSFP and CFPs, DDM Temperature and Supply voltage is available only at the Module level (to be shown in Table 3.

The section called Statistics Collection on page 34 shows the following QSFP and CFP sample DDM and DDM Lane information:

The QSFP and CFPs, the number of lanes is indicated by DDM attribute “Number of Lanes: 4”.

Subsequently, each lane threshold and measured values are shown per lane.

If a given lane entry is not supported by the given QSFP or CFP specific model, then it will be shown as “-“ in the entry.

A sample QSFP and CFP lane information is provided below:

```
Transceiver Data
Transceiver Type   : QSFP+
Model Number       : 3HE06485AAAA01  ALU  IPUIBMY3AA
TX Laser Wavelength: 1310 nm                    Diag Capable     : yes
Number of Lanes    : 4
Connector Code     : LC                         Vendor OUI       : e4:25:e9
Manufacture date   : 2012/02/02                 Media            : Ethernet
Serial Number      : 12050188
Part Number        : DF40GELR411102A
Optical Compliance : 40GBASE-LR4
Link Length support: 10km for SMF
===============================================================================
Transceiver Digital Diagnostic Monitoring (DDM)
===============================================================================
                             Value High Alarm  High Warn   Low Warn  Low Alarm
-------------------------------------------------------------------------------
Temperature (C)              +35.6     +75.0      +70.0       +0.0       -5.0
Supply Voltage (V)            3.23      3.60       3.50       3.10       3.00
===============================================================================
===============================================================================
Transceiver Lane Digital Diagnostic Monitoring (DDM)
===============================================================================
                                High Alarm   High Warn    Low Warn   Low Alarm
```

```
Lane Tx Bias Current (mA)              78.0        75.0        25.0        20.0
Lane Rx Optical Pwr (avg dBm)          2.30        2.00      -11.02      -13.01
-------------------------------------------------------------------------------
Lane ID Temp(C)/Alm       Tx Bias(mA)/Alm   Tx Pwr(dBm)/Alm   Rx Pwr(dBm)/Alm
-------------------------------------------------------------------------------
    1              -              43.5                 -              0.42
    2              -              46.7                 -             -0.38
    3              -              37.3                 -              0.55
    4              -              42.0                 -             -0.52
===============================================================================
Transceiver Type   : CFP
Model Number       : 3HE04821ABAA01  ALU  IPUIBHJDAA
TX Laser Wavelength: 1294 nm                    Diag Capable     : yes
Number of Lanes    : 4
Connector Code     : LC                         Vendor OUI       : 00:90:65
Manufacture date   : 2011/02/11                 Media            : Ethernet
Serial Number      : C22CQYR
Part Number        : FTLC1181RDNL-A5
Optical Compliance : 100GBASE-LR4
Link Length support: 10km for SMF
===============================================================================
Transceiver Digital Diagnostic Monitoring (DDM)
===============================================================================
                             Value High Alarm  High Warn   Low Warn  Low Alarm
-------------------------------------------------------------------------------
Temperature (C)              +48.2     +70.0      +68.0       +2.0       +0.0
Supply Voltage (V)            3.24      3.46       3.43       3.17       3.13
===============================================================================
===============================================================================
Transceiver Lane Digital Diagnostic Monitoring (DDM)
===============================================================================
                                High Alarm   High Warn    Low Warn   Low Alarm
-------------------------------------------------------------------------------
Lane Temperature (C)                 +55.0       +53.0       +27.0       +25.0
Lane Tx Bias Current (mA)            120.0       115.0        35.0        30.0
Lane Tx Output Power (dBm)            4.50        4.00       -3.80       -4.30
Lane Rx Optical Pwr (avg dBm)         4.50        4.00      -13.00      -16.00
-------------------------------------------------------------------------------
Lane ID Temp(C)/Alm       Tx Bias(mA)/Alm   Tx Pwr(dBm)/Alm   Rx Pwr(dBm)/Alm
-------------------------------------------------------------------------------
    1          +47.6             59.2              0.30            -10.67
    2          +43.1             64.2              0.27            -10.31
    3          +47.7             56.2              0.38            -10.58
    4          +51.1             60.1              0.46            -10.37
===============================================================================
```

The transceiver is programmed with warning and alarm thresholds for low and high conditions that can generate system events. These thresholds are programmed by the transceiver manufacturer.

There are no CLI commands required for DDM operations, however, the **show>port** ***port-id*** **detail** command displays DDM information in the Transceiver Digital Diagnostics Monitoring output section.

DDM information is populated into the router's MIBs, so the DDM data can be retrieved by Network Management using SNMP. Also, RMON threshold monitoring can be configured for the

DDM MIB variables to set custom event thresholds if the factory-programmed thresholds are not at the desired levels.

The following are potential uses of the DDM data:

- Optics degradation monitoring — With the information returned by the DDM-capable optics module, degradation in optical performance can be monitored and trigger events based on custom or the factory-programmed warning and alarm thresholds.
- Link/router fault isolation — With the information returned by the DDM-capable optics module, any optical problem affecting a port can be quickly identified or eliminated as the potential problem source.

Supported real-time DDM features are summarized in Table 2.

Table 2: Real-Time DDM Information

Parameter	User Units	SFP/XFP Units	SFP	XFP	MSA DWDM
Temperature	Celsius	C	Supported	Supported	Supported
Supply Voltage	Volts	µV	Supported	Supported	Not supported
TX Bias Current	mA	µA	Supported	Supported	Supported
TX Output Power	dBm (converted from mW)	mW	Supported	Supported	Supported
RX Received Optical Power4	dBm (converted from dBm) (Avg Rx Power or OMA)	mW	Supported	Supported	Supported
AUX1	parameter dependent (embedded in transceiver)	-	Not supported	Supported	Not supported
AUX2	parameter dependent (embedded in transceiver)	-	Not supported	Supported	Not supported

The factory-programmed DDM alarms and warnings that are supported are summarized in Table 3.

Table 3: DDM Alarms and Warnings

Parameter	SFP/XFP Units	SFP	XFP	Required?	MSA DWDM
Temperature - High Alarm - Low Alarm - High Warning - Low Warning	C	Yes	Yes	Yes	Yes
Supply Voltage - High Alarm - Low Alarm - High Warning - Low Warning	µV	Yes	Yes	Yes	No
TX Bias Current - High Alarm - Low Alarm - High Warning - Low Warning	µA	Yes	Yes	Yes	Yes
TX Output Power - High Alarm - Low Alarm - High Warning - Low Warning	mW	Yes	Yes	Yes	Yes
RX Optical Power - High Alarm - Low Alarm - High Warning - Low Warning	mW	Yes	Yes	Yes	Yes
AUX1 - High Alarm - Low Alarm - High Warning - Low Warning	parameter dependent (embedded in transceiver)	No	Yes	Yes	No
AUX2 - High Alarm - Low Alarm - High Warning - Low Warning	parameter dependent (embedded in transceiver)	No	Yes	Yes	No

Alcatel-Lucent SFPs and XFPs

The availability of the DDM real-time information and warning/alarm status is based on the transceiver. It may or may not indicate that DDM is supported. Although some Alcatel-Lucent SFPs support DDM, Alcatel-Lucent has not required DDM support in releases prior to Release 6.0. Non-DDM and DDM-supported SFPs are distinguished by a specific ICS value.

For Alcatel-Lucent SFPs that do not indicate DDM support in the ICS value, DDM data is available although the accuracy of the information has not been validated or verified.

For non-Alcatel-Lucent transceivers, DDM information may be displayed, but Alcatel-Lucent is not responsible for formatting, accuracy, etc.

Statistics Collection

The DDM information and warnings/alarms are collected at one minute intervals, so the minimum resolution for any DDM events when correlating with other system events is one minute.

Note that in the Transceiver Digital Diagnostic Monitoring section of the **show port** *port-id* **detail** command output:

- If the present measured value is higher than the either or both High Alarm, High Warn thresholds; an exclamation mark "!" displays along with the threshold value.
- If the present measured value is lower than the either or both Low Alarm, Low Warn thresholds; an exclamation mark "!" displays along with the threshold value.

```
B:SR7-101# show port 2/1/6 detail
......
===============================================================================
Transceiver Digital Diagnostic Monitoring (DDM), Internally Calibrated
===============================================================================
                   Value High Alarm  High Warn   Low Warn  Low Alarm
-------------------------------------------------------------------------------
Temperature (C)       +33.0+98.0   +88.0      -43.0-45.0
Supply Voltage (V)       3.31 4.12    3.60       3.00 2.80
Tx Bias Current (mA)5.7 60.0    50.00.1  0.0
Tx Output Power (dBm)      -5.45 0.00   -2.00     -10.50    -12.50
Rx Optical Power (avg dBm)    -0.65-3.00!   -4.00!    -19.51    -20.51
===============================================================================
```

Ports

Port Types

Before a port can be configured, the slot must be provisioned with a card type and MDA type .

The Alcatel-Lucent routers support the following port types:

- Ethernet — Supported Ethernet port types include:
  - → Fast Ethernet (10/100BASE-T)
  - → Gigabit (1000BASE-T)
  - → 10Gigabit Ethernet (10GBASE-X) ports on an appropriate MDA.

  Router ports must be configured as either access, hybrid or network. The default is network.

  - → Access ports — Configured for customer facing traffic on which services are configured. If a Service Access Port (SAP) is to be configured on the port or channel, it must be configured as an access port or channel. When a port is configured for access mode, the appropriate encapsulation type must be configured to distinguish the services on the port or channel. Once a port has been configured for access mode, one or more services can be configured on the port or channel depending on the encapsulation value.
  - → Network ports — Configured for network facing traffic. These ports participate in the service provider transport or infrastructure network. Dot1q is supported on network ports.
  - → Hybrid ports — Configured for access and network facing traffic. While the default mode of an Ethernet port remains network, the mode of a port cannot be changed between the access/network/hybrid values unless the port is shut down and the configured SAPs and/or interfaces are deleted. Hybrid ports allow a single port to operate in both access and network modes. MTU of port in hybrid mode is the same as in network mode except for the 10/100 MDA. The default encap for hybrid port mode is dot1q; it also supports QinQ encapsulation on the port level. Null hybrid port mode is not supported. Hybrid mode on the is not supported.

    Once the port is changed to hybrid, the default MTU of the port is changed to match the value of 9212 bytes currently used in network mode (higher than an access port); this is to ensure that both SAP and network VLANs can be accommodated. The only exception is when the port is a 10/100 fast Ethernet. In those cases, the MTU in hybrid mode is set to 1522 bytes, which corresponds to the default access MTU with QinQ, which is larger than the network dot1q MTU or access dot1q MTU for this type of Ethernet port. The configuration of all parameters in access and network contexts will

continue to be done within the port using the same CLI hierarchy as in existing implementation. The difference is that a port configured in mode hybrid allows both ingress and egress contexts to be configured concurrently.

An Ethernet port configured in hybrid mode can have two values of encapsulation type: dot1q and QinQ. The NULL value is not supported since a single SAP is allowed, and can be achieved by configuring the port in the access mode, or a single network IP interface is allowed, which can be achieved by configuring the port in network mode. Hybrid mode can be enabled on a LAG port when the port is part of a single chassis LAG configuration. When the port is part of a multi-chassis LAG configuration, it can only be configured to access mode since MC-LAG is not supported on a network port and consequently is not supported on a hybrid port. The same restriction applies to a port that is part of an MC-Ring configuration.

For a hybrid port, the amount of the allocated port buffers in each of ingress and egress is split equally between network and access contexts using the following **config>port>hybrid-buffer-allocation>ing-weight access** *access-weight [0..100]* **network** *network-weight* [0..100] and **config>port>hybrid-buffer-allocation>egr-weight access** *access-weight* [0..100] **network** *network-weight* [0..100] commands.

Adapting the terminology in buffer-pools, the port's access active bandwidth and network active bandwidth in each ingress and egress are derived as follows (egress formulas shown only):

- total-hybrid-port-egress-weights = access-weight + network-weight
- hybrid-port-access-egress-factor = access-weight / total-hybrid-port-egress-weights
- hybrid-port-network-egress-factor = network-weight / total-hybrid-port-egress-weights
- port-access-active-egress-bandwidth = port-active-egress-bandwidth x
- hybrid-port-access-egress-factor
- port-network-active-egress-bandwidth = port-active-egress-bandwidth x
- hybrid-port-network-egress-factor

When a named pool policy is applied to the hybrid port's MDA or to the hybrid port, the port's fair share of total buffers available to the MDA is split into three parts: default pools, named pools local to the port, and named pools on the ports MDA. This allocation can be altered by entering the corresponding values in the **port-allocation-weights** parameter.

- SONET-SDH and TDM — Supported SONET-SDH and TDM port types include:
  - → n*DS-0 inside DS-1/E-1
  - → DS-1/E-1DS-3/E-3
  - → OC3/STM-1
  - → OC12/STM-4
  - → OC48/STM-16
  - → OC192/STM-64 SONET/SDH

→ OC768/STM-256

A SONET/SDH port/path or a TDM port/channel can be configured with the following encapsulations depending on the MDA type:

→ Frame Relay
→ PPP
→ cHDLC

- ATM

  Some MDAs support ATM encapsulation on SONET/SDH and TDM ports. The ATM cell format and can be configured for either UNI or NNI cell format. The format is configurable on a SONET/SDH or TDM port/channel path basis. All VCs on a path, channel or port must use the same cell format. The ATM cell mapping can also be configured on per-interface basis for either Direct or PLCP on some MDAs (for example ASAP MDA).

- Several Media Dependent Adapters (MDAs) support channelization down to the DS-0 level. ATM, Frame Relay, PPP, and cHDLC are supported encapsulations on channelized ports.
- Link Aggregation (LAG) — LAG can be used to group multiple ports into one logical link. The aggregation of multiple physical links allows for load sharing and offers seamless redundancy. If one of the links fails, traffic will be redistributed over the remaining links.
- Multilink Bundles — A multilink bundle is a collection of channels on channelized ports that physically reside on the same MDA. Multilink bundles are used by providers who offer either bandwidth-on-demand services or fractional bandwidth services (fraction of a DS-3/E-3 for example). Multilink bundles are supported over PPP channels (MLPPP) and ATM channels (IMA).
- APS — Automatic Protection Switching (APS) is a means to provide redundancy on SONET equipment to guard against linear unidirectional or bidirectional failures. The network elements (NEs) in a SONET/SDH network constantly monitor the health of the network. When a failure is detected, the network proceeds through a coordinated predefined sequence of steps to transfer (or switchover) live traffic to the backup facility (called protection facility.) This is done very quickly to minimize lost traffic. Traffic remains on the protection facility until the primary facility (called working facility) fault is cleared, at which time the traffic may optionally be reverted to the working facility.
- Bundle Protection Group (BPGrp) — A BPGrp is a collection of two bundles created on the APS Group port. Working bundle resides on the working circuit of the APS group, while protection bundle resides on the protection circuit of the APS group. APS protocol running on the circuits of the APS Group port monitors the health of the SONET/SDH line and based on it or administrative action moves user traffic from one bundle to another in the group as part of an APS switch.
- Cross connect adaptor (CCA) — A CCA on a VSM module interconnects the egress forwarding path on the IOM directly to the ingress forwarding path. This eliminates the

need for the physical port MAC, PHY, cable and other MDA-specific components producing a less costly and more reliable adapter.

- Optical Transport Network (OTN) — Including OTU2, OTU2e, and OTU3. OTU2 encapsulates 10-Gigabit Ethernet WAN and adds FEC (Forward Error Correction). OTU2e encapsulates 10-Gigabit Ethernet LAN and adds FEC (Forward Error Correction). OTU3 encapsulated OC768 and adds FEC.

Port Features

Port State and Operational State

There are two port attributes that are related and similar but have slightly different meanings: Port State and Operational State (or Operational Status).

The following descriptions are based on normal individual ports. Many of the same concepts apply to other objects that are modeled as ports in SR-OS such as PPP/IMA/MLFR multilink bundles or APS groups but the show output descriptions for these objects should be consulted for the details.

- Port State
  - → Displayed in port summaries such as **show port** or **show port 1/1**
  - → tmnxPortState in the TIMETRA-PORT-MIB
  - → Values: None, Ghost, Down (linkDown), Link Up, Up
- Operational State
  - → Displayed in the show output of a specific port such as **show port 2/1/3**
  - → tmnxPortOperStatus in the TIMETRA-PORT-MIB
  - → Values: Up (inService), Down (outOfService)

The behavior of Port State and Operational State are different for a port with link protocols configured (Eth OAM, Eth CFM or LACP for ethernet ports, LCP for PPP/POS ports). A port with link protocols configured will only transition to the **Up** Port State when the physical link is up and all the configured protocols are up. A port with no link protocols configured will transition from Down to Link Up and then to Up immediately once the physical link layer is up.

The SR OS linkDown and linkUp log events (events 2004 and 2005 in the SNMP application group) are associated with transitions of the port Operational State. Note that these events map to the RFC 2863, *The Interfaces Group MIB*, (which obsoletes RFC 2233, *The Interfaces Group MIB using SMIv2*) linkDown and linkUp traps as mentioned in the SNMPv2-MIB.

An Operational State of **Up** indicates that the port is ready to transmit service traffic (the port is physically up and any configured link protocols are up). The relationship between port Operational State and Port State in SR OS is shown in Table 4:

Table 4: Relationship of Port State and Oper State

	Operational State (Oper State or Oper Status) (as displayed in "show port x/y/z")	
Port State (as displayed in the **show port** summary)	For ports that have no link layer protocols configured	For ports that have link layer protocols configured (PPP, LACP, 802.3ah EFM, 802.1ag Eth-CFM)
Up	Up	Up
Link Up (indicates the physical link is ready)	Up	Down
Down	Down	Down

802.1x Network Access Control

The Alcatel-Lucent 7750 SR supports network access control of client devices (PCs, STBs, etc.) on an Ethernet network using the IEEE. 802.1x standard. 802.1x is known as Extensible Authentication Protocol (EAP) over a LAN network or EAPOL.

802.1x Modes

The Alcatel-Lucent7750 SR supports port-based network access control for Ethernet ports only. Every Ethernet port can be configured to operate in one of three different operation modes, controlled by the port-control parameter:

- **force-auth** — Disables 802.1x authentication and causes the port to transition to the authorized state without requiring any authentication exchange. The port transmits and receives normal traffic without requiring 802.1x-based host authentication. This is the default setting.
- **force-unauth** — Causes the port to remain in the unauthorized state, ignoring all attempts by the hosts to authenticate. The switch cannot provide authentication services to the host through the interface.
- **auto** — Enables 802.1x authentication. The port starts in the unauthorized state, allowing only EAPOL frames to be sent and received through the port. Both the router and the host can initiate an authentication procedure as described below. The port will remain in un-authorized state (no traffic except EAPOL frames is allowed) until the first client is authenticated successfully. After this, traffic is allowed on the port for all connected hosts.

802.1x Basics

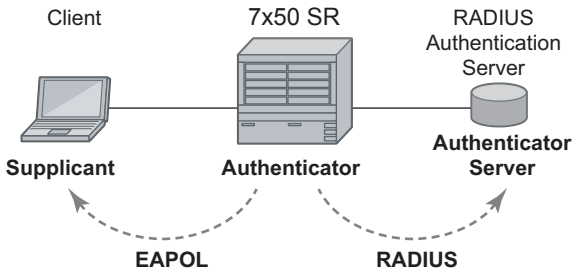

Figure 1: 802.1x Architecture

The IEEE 802.1x standard defines three participants in an authentication conversation (see Figure 1).

- The supplicant — This is the end-user device that requests access to the network.
- The authenticator — Controls access to the network. Both the supplicant and the authenticator are referred to as Port Authentication Entities (PAEs).
- The authentication server — Performs the actual processing of the user information.

The authentication exchange is carried out between the supplicant and the authentication server, the authenticator acts only as a bridge. The communication between the supplicant and the authenticator is done through the Extended Authentication Protocol (EAP) over LANs (EAPOL). On the back end, the communication between the authenticator and the authentication server is done with the RADIUS protocol. The authenticator is thus a RADIUS client, and the authentication server a RADIUS server.

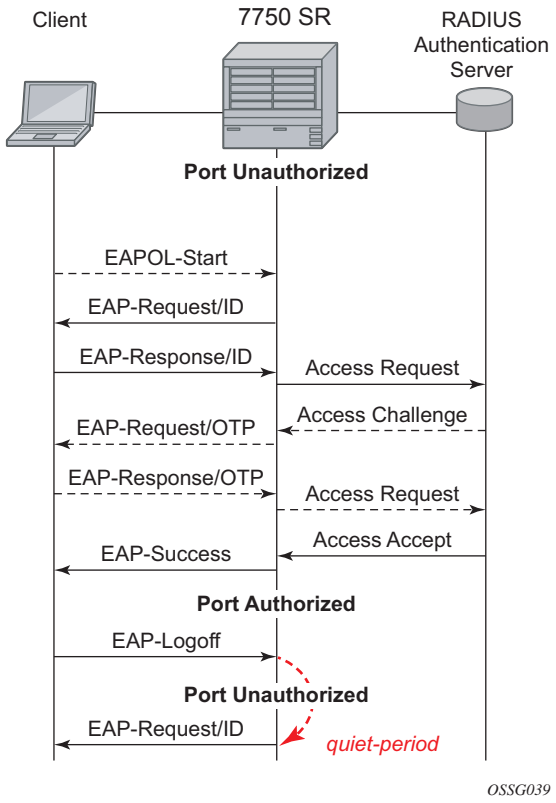

Figure 2: 802.1x Authentication Scenario

The messages involved in the authentication procedure are illustrated in Figure 2. The router will initiate the procedure when the Ethernet port becomes operationally up, by sending a special PDU called EAP-Request/ID to the client. The client can also initiate the exchange by sending an EAPOL-start PDU, if it doesn't receive the EAP-Request/ID frame during bootup. The client responds on the EAP-Request/ID with a EAP-Response/ID frame, containing its identity (typically username + password).

After receiving the EAP-Response/ID frame, the router will encapsulate the identity information into a RADIUS AccessRequest packet, and send it off to the configured RADIUS server.

The RADIUS server checks the supplied credentials, and if approved will return an Access Accept message to the router. The router notifies the client with an EAP-Success PDU and puts the port in authorized state.

802.1x Timers

The 802.1x authentication procedure is controlled by a number of configurable timers and scalars. There are two separate sets, one for the EAPOL message exchange and one for the RADIUS message exchange. See Figure 3 for an example of the timers.

EAPOL timers:

- **transit-period** — Indicates how many seconds the Authenticator will listen for an EAP-Response/ID frame. If the timer expires, a new EAP-Request/ID frame will be sent and the timer restarted. The default value is 60. The range is 1-3600 seconds.
- **supplicant-timeout** — This timer is started at the beginning of a new authentication procedure (transmission of first EAP-Request/ID frame). If the timer expires before an EAP-Response/ID frame is received, the 802.1x authentication session is considered as having failed. The default value is 30. The range is 1 — 300.
- **quiet-period** — Indicates number of seconds between authentication sessions It is started after logoff, after sending an EAP-Failure message or after expiry of the supplicant-timeout timer. The default value is 60. The range is 1 — 3600.

RADIUS timer and scaler:

- **max-auth-req** — Indicates the maximum number of times that the router will send an authentication request to the RADIUS server before the procedure is considered as having failed. The default value is value 2. The range is 1 — 10.
- **server-timeout** — Indicates how many seconds the authenticator will wait for a RADIUS response message. If the timer expires, the access request message is sent again, up to *max-auth-req* times. The default value is 60. The range is 1 — 3600 seconds.

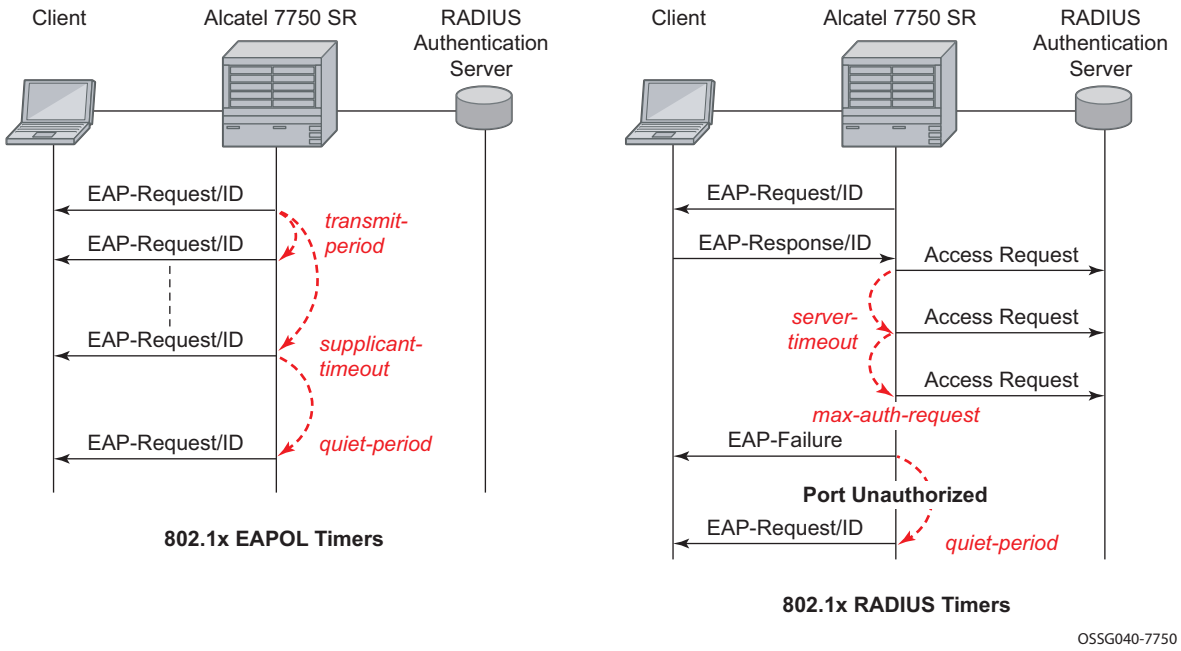

Figure 3: 802.1x EAPOL Timers (left) and RADIUS Timers (right)

The router can also be configured to periodically trigger the authentication procedure automatically. This is controlled by the enable re-authentication and reauth-period parameters. Reauth-period indicates the period in seconds (since the last time that the authorization state was confirmed) before a new authentication procedure is started. The range of reauth-period is 1 — 9000 seconds (the default is 3600 seconds, one hour). Note that the port stays in an authorized state during the re-authentication procedure.

802.1x Tunneling

Tunneling of untagged 802.1x frames received on a port is supported for both Epipe and VPLS service using either null or default SAPs (for example 1/1/1:*) when the port dot1x port-control is set to force-auth.

When tunneling is enabled on a port (using the command configure **port** *port-id* **ethernet dot1x tunneling**), untagged 802.1x frames are treated like user frames and are switched into Epipe or VPLS services which have a corresponding null SAP or default SAP on that port. In the case of a default SAP, it is possible that other non-default SAPs are also present on the port. Untagged 802.1x frames received on other service types, or on network ports, are dropped. This is supported on FP2 or higher hardware.

When tunneling is required, it is expected that it is enabled on all ports into which 802.1x frames are to be received. The configuration of dot1x must be configured consistently across all ports in LAG as this is not enforced by the system.

Note that 802.1x frames are treated like user frames, that is, tunneled, by default when received on a spoke or mesh SDP.

802.1x Configuration and Limitations

Configuration of 802.1x network access control on the router consists of two parts:

- Generic parameters, which are configured under **config>security>dot1x**
- Port-specific parameters, which are configured under **config>port>ethernet>dot1x**

801.x authentication:

- Provides access to the port for any device, even if only a single client has been authenticated.
- Can only be used to gain access to a pre-defined Service Access Point (SAP). It is not possible to dynamically select a service (such as a VPLS service) depending on the 802.1x authentication information.
- If 802.1x access control is enabled and a high rate of 802.1x frames are received on a port, that port will be blocked for a period of 5 minutes as a DOS protection mechanism.

SONET/SDH Port Attributes

One OC-3 / STM-1 port is supported on the CMA. One OC-3 / STM-1 port is supported on the MDA. The ports can be configured for either SONET or SDH operation. SONET ports are configured for channelized OC-3 operation. SDH ports can be configured for channelized STM-1 operation.

The port's transmit clock rate can be node or loop timed. The port's receive clock rate can be used as a synchronization source for the system. The Section Trace (C1) byte can be configured by the user to ensure proper physical cabling. The port can activate and deactivate local line and internal loopbacks.

All SONET/SDH line alarms are configurable to be either enabled (default) or disabled. Link hold timers can be configured in 100ms increments to control link up and link down indications. The line signal degradation bit error rate (ber-sd) threshold and the line signal failure bit error rate (ber-sf) threshold can be configured.

The CMAs and MDAs support all standard SR OC-3/STM-1 SFP optics including multi-mode, intermediate reach, and long reach. Single fiber mode is not supported.

The CMA contains 3 LEDs for power, status and link state of port #1. The MDA contains LEDs for power, status and one for each link state. The power LED is blue if power is connected and off if no power is present. The status LED is green when operationally up, amber when operationally down, off when administratively shutdown and blinking green during initialization. The link state LED is green when the link is established; amber when the link is down; and unlit when the port is shutdown.

SONET/ SDH Path Attributes

Any CES path can only be configured to operate in access mode. Each path has a configurable text description. The SONET/SDH signal label byte (C2) is configurable. The SONET/SDH path trace string (J1) is configurable. Payload scrambling can not be enabled on CES paths. The valid SONET and SDH path configurations are shown in Table 5.

Table 5: Valid SONET and SDH Path Configurations

Framing	Path Configuration Options Per Physical Port	Max Number of Paths Per Physical Port
SDH	STM1>AUG1>VC4>TUG3>TUG2>VC12> E1 STM1>AUG1>VC3>TUG2>VC12>E1	63 E1 or 512 n*64kbps
SONET	OC3>STS1 SPE>DS3>E1	

Table 5: Valid SONET and SDH Path Configurations

Framing	Path Configuration Options Per Physical Port	Max Number of Paths Per Physical Port
SONET	OC3>STS1 SPE>VT GROUP>VT1.5 SPE>DS1	84 DS1 or 512 n*64kbps
SONET	OC3>STS1 SPE>DS3	3 DS3
SONET	OC3>STS1 SPE>DS3>DS1	84 DS1, 63 E1 or 512 n*64kbps
SDH	STM1>AUG1>VC4>TUG3>TUG2>TU11> VC11>DS1 STM1>AUG1>VC3>TUG2>VC11>DS1	84 DS1 or 512 n*64kbps
SDH	STM1>AUG1>VC3>DS3>DS1	84 DS1, 63 E1 or 512 n*64kbps
SDH	STM1>AUG1>VC4>TUG3>VC3>E3 STM1>AUG1>VC3>E3	3 E3
SDH	STM1>AUG1>VC3>DS3	3 DS3
SDH	STM1>AUG1>VC3>DS3>E1	3 DS3

All SONET/SDH path alarms are configurable to be either enabled (the default) or disabled. The MTU size is configurable per path in the range of 512 to 2092. The path uses a default MTU size set to equal the largest possible CES packet size.

Load balancing options are not applicable to channelized CES paths.

Multilink Frame Relay

MLFR is a bundling capability allowing users to spray FR frame fragments over multiple T1/E1 links. This allows a dynamic provisioning of additional bandwidth by adding incremental bandwidth between T1/E1 and DS3/E3. A MLFR bundle increases fault tolerance and improves QoS characteristics since one single large frame of low priority cannot block a higher priority frame.

A MLFR supports up to eight (8) member links and a maximum of 128 bundles with up to 336 T1 / 252 E1 members links can be configured per MDA. NxDS0 circuits or higher speed circuits are not supported.

The MLFR implementation supports FRF.16.1 bundle link integrity protocol to verify serviceability of a member link.

MLFR Bundle Data Plane

FRF.16.1 reuses the UNI/NNI fragmentation procedures defined in FRF.12. Frames on all FR SAP on the MLFR bundle have the UNI/NNI fragmentation header added regardless if they are fragmented or not. A separate sequence number state machine is used for each FR SAP configured on the bundle. The fragmentation threshold is configurable in the range 128-512 bytes.

In order to provide priority based scheduling of the FR SAP fragments over the bundle links, the user configures a FR scheduling class for each FR SAP configured on the bundle. As in MC-MLPPP, four scheduling classes are supported.

A separate fragmentation context is used by each FR SAP. FR SAPs of the same scheduling class share the same egress FR scheduling class queue with fragments of each SAP packets stored contiguously. The fragments from each scheduling class queue are then sprayed over the member links. Furthermore, the user may select the option to not fragment but spray the FR frames with the fragmentation header included over the member links.

Received fragments over the member links are re-assembled on a per SAP basis to re-create the original FR frame.

A user is not allowed to add an FR SAP with FRF.12 e2e fragmentation enabled to an MLFR bundle. Conversely, the user cannot enable FRF.12 e2e fragmentation on an FR SAP configured on an MLFR bundle. If an FR frame with the e2e fragmentation header is received on a bundle, it is forwarded if the FR SAP is part of an Fpipe service. It will be discarded if the FR SAP is part of any other service.

Note that the operator must disable LMI before adding a link to an MLFR bundle. Also, the operator must shut down the bundle in order to change the value of the fragmentation threshold.

An FR SAP configured on an MLFR bundle can be part of a VLL, VPLS, IES, or VPRN service.

MLFR Bundle Link Integrity Protocol

FRF.16.1 defines a MLFR Bundle Link Integrity Protocol which verifies the serviceability of a member link. If a problem is found on the member link the link integrity protocol will identify the problem, flag the link as unusable, and adjust the Bundle's available bandwidth. For MLFR Bundles the link integrity protocol is always enabled.

For each member link of a bundle the link integrity protocol will do the following:

- Confirm frame processing capabilities of each member link.
- Verify membership of a link to a specific remote bundle.
- Report to the remote end of the member link the bundle to which the link belongs
- Detect loopbacks on the member link. This is always enabled on the 7750 SR. The near-end monitors the magic number Information Element (IE) sent by the far-end and if its value matches the one it transmitted in ten consecutive control messages, it sends a remove_link message to the far-end and brings the link down. The near-end will attempt to add the link until it succeeds.
- Estimate propagation delay on the member link. The differential delay is calculated as follows in the 7750 SR implementation. Every time the near-end sends an add_link or Hello message to the far-end, it includes the Timestamp Information Element (IE) with the local time the packet was sent. FRF16.1 standard requires that the remote equipment includes the timestamp IE and copies the received timestamp value unchanged if the sender included this IE. When the far-end node sends back the ACK for these messages, the near-end calculates the round trip time. The 7750 SR implementation maintains a history of the last "N" round-trip-times that were received. It takes the fastest of these samples for each member link to find out the member link with the fastest RTT. Then for each link it calculates the difference between the fastest links RTT, and the RTT for the current link. The user has the option to coordinate link removal between the local and remote equipment. Note, however, that in the 7750 implementation, the addition of a link will be hitless but the removing a link is not.

Specifically, the MLFR Bundle Link Integrity Protocol defines the following control messages:

- ADD_LINK
- ADD_LINK_ACK
- ADD_LINK_REJ
- HELLO

- HELLO_ACK
- REMOVE_LINK
- REMOVE_LINK_ACK

The control messages are encapsulated in a single-fragment frame where the C-bit, the B-bit, and the E-bit are all set. The details of the message format are given in FRF.16.1. Table 6 lists the user configured control parameters with values as specified in FRF.16.1.

Table 6: FRF.16.1 Values

Parameter	Default Value	Minimum Value	Maximum Value
Timer T_HELLO	10 seconds	1 second	180 seconds
Timer T_ACK	4 seconds	1 second	10
Count N_MAX_RETRY	2	1	5

T_HELLO Timer - this timer controls the rate at which hello messages are sent. Following a period of T_HELLO duration, a HELLO message is transmitted onto the Bundle Link.

Note that T_HELLO Timer is also used, during the Bundle Link adding process, as an additional delay before re-sending an ADD_LINK message to the peer Bundle Link when this peer Bundle Link does not answer as expected.

T_ACK Timer - this timer defines the maximum period to wait for a response to any message sent onto the Bundle Link before attempting to retransmit a message onto the Bundle Link.

N_RETRY - this counter specifies the number of times a retransmission onto a Bundle Link will be attempted before an error is declared and the appropriate action taken.

FRF.12 End-to-End Fragmentation

The user enables FRF.12 e2e fragmentation on a per FR SAP basis. A fragmentation header is added between the standard Q.922 header and the payload. This header consists of a 2-byte Network Layer Protocol ID (NLPID) of value 0xB1 to indicate e2e fragmentation payload and a 2-byte containing the Beginning bit (B-bit), the End-bit (E-bit), the Control bit (C-bit), and the Sequence Number field.

The following is the mode of operation for the fragmentation in the transmit direction of the FR SAP. Frames of all the FR SAP forwarding class queues are subject to fragmentation. The fragmentation header is, however, not included when the frame size is smaller than the user configured fragmentation size. The SAP transmits all fragments of a frame before sending the next full or fragmented frame. The fragmentation threshold is configurable in the range 128 — 512 bytes. In the receive direction, the SAP accepts a full frame interleaved with fragments of another frame to interoperate with other vendor implementations.

A FR SAP with FRF.12 e2e fragmentation enabled can be part of a VPLS service, an IES service, a VPRN service, an Ethernet VLL service, or an IP VLL service. This SAP cannot be part of a FR VLL service or an FRF.5 VLL service. However, fragmented frames received on such VLLs will be passed transparently as in current implementation.

SAP Fragment Interleaving Option

This option provides a different mode of operation for the fragmentation in the transmit direction of the FR SAP than in the default behavior of a FRF.12 end-to-end fragmentation. It allows for the interleaving of high-priority frames and fragments of low-priority frames.

When the interleave option is enabled, only frames of the FR SAP non expedited forwarding class queues are subject to fragmentation. The frames of the FR SAP expedited queues are interleaved, with no fragmentation header, among the fragmented frames. In effect, this provides a behavior like in MLPPP Link Fragment Interleaving (LFI). The receive direction of the FR SAP supports both modes of operation concurrently, for example, with and without fragment interleaving.

FRF.12 UNI/NNI Link Fragmentation

The user enables FRF.12 UNI/NNI link fragmentation on a per FR circuit basis. All FR SAPs configured on this circuit are subject to fragmentation. A fragmentation header is added on top of the standard Q.922 header. This header consists of 2 bytes containing the beginning bit (B-bit), the End-bit (E-bit), the Control bit (C-bit), and the sequence number field. The fragmentation header is included on frames of all SAPs regardless if the frame size is larger or not than the fragment size.

The FECN, BECN, and DE bits of all fragments of a given FR frame are set to the same value as the original frame. The FECN, BECN, and DE bits of a re-assembled frame are set to the logical OR of the corresponding bits on the constituent fragments.

The operator must delete all configured FR SAPs on a port before enabling or disabling FRF.12 UNI/NNI on that port. Also, the user must shut down the port in order to change the value of the fragmentation threshold.

A FR SAP on a FR circuit with FRF.12 UNI/NNI fragmentation enabled can be part of a VLL, VPLS, IES, or VPRN service.

QoS for a link with FRF.12 UNI/NNI fragmentation is the same as for a MLFR bundle. The FR class queue parameters and its scheduling parameters are configured by applying an egress QoS profile to an FRF.12 UNI/NNI port. The FR scheduling class ingress re-assembly timeout is not applicable to a FRF.12 UNI/NNI port.

MLFR/FRF.12 Support of APS, BFD, and Mirroring Features

The following APS support is provided:

- Single-chassis APS is supported on a SONET/SDH port with FRF.12 UNI/NNI fragmentation enabled on the port or on a constituent TDM circuit.
- Single-chassis APS is supported on a SONET/SDH port with FRF.12 e2e fragmentation enabled on one or more FR SAPs on the port or on a constituent TDM circuit.
- Single-chassis APS is not supported on a SONET/SDH port with MLFR bundles configured.
- Multi-chassis APS is not supported on a SONET/SDH port with FR encapsulation configured on the port or on a constituent TDM circuit.

The following BFD support is provided:

- BFD is supported on an IP interface configured over a FR SAP with e2e fragmentation enabled.
- BFD is supported on an IP interface configured over a FR SAP on a port or channel with UNI/NNI fragmentation enabled.
- BFD is not supported on an FR SAP configured on an MLFR bundle.

The following mirroring support is provided:

- Port mirroring and FR SAP mirroring on an MLFR bundle.
- IP mirroring for an FR SAP on an MLFR bundle.
- A mirror source can be an MLFR bundle or a FR SAP on an FR bundle.
- Mirror destinations must be FR SAPs and must not be part of an APS group or an MLFR bundle.

Multilink Point-to-Point Protocol (MLPPP)

Multilink point-to-point protocol is defined in the IETF RFC 1990, *The PPP Multilink Protocol (MP)*, and provides a way to distribute data across multiple links within an MLPPP bundle to achieve high bandwidth. MLPPP allows for a single frame to be fragmented and transmitted across multiple links. This allows for lower latency and also allows for a higher maximum receive unit (MRU).

MP is negotiated during the initial LCP option negotiations of a standard PPP session. A router indicates to its peer that it is willing to perform MLPPP by sending the MP option as part of the initial LCP option negotiation. This negotiation indicates the following:

1. The system offering the option is capable of combining multiple physical links into one logical link;
2. The system is capable of receiving upper layer protocol data units (PDU) fragmented using the MP header and reassembling the fragments back into the original PDU for processing;
3. The system is capable of receiving PDUs of size N octets where N is specified as part of the option even if N is larger than the maximum receive unit (MRU) for a single physical link.

Once MLPPP has been successfully negotiated, the sending system is free to send PDUs encapsulated and/or fragmented with the MP header.

MP introduces a new protocol type with a protocol ID (PID) of Ox003d. Figure 4 and Figure 5 show the MLPPP fragment frame structure. Framing to indicate the beginning and end of the

encapsulation is the same as that used by PPP, and described in PPP in HDLC-like framing [RFC 1662]. MP frames use the same HDLC address and control pair value as PPP, namely: Address - OxFF and Control - Ox03. The two octet protocol field is also structured the same as in PPP encapsulation. A summary of the MP encapsulation is shown in Figure 4.

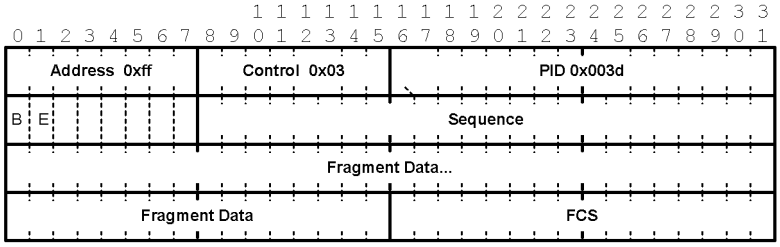

Figure 4: MLPPP 24-bit Fragment Format

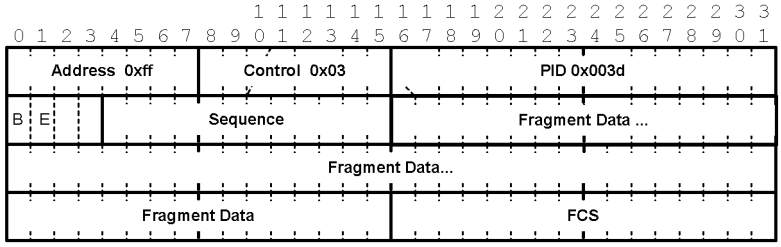

Figure 5: MLPPP 12-bit Fragment Format

The required and default format for MP is the 24-bit format. During the LCP state the 12-bit format can be negotiated. The SR-series routers can support and negotiate the alternate 12-bit frame format.

Protocol Field (PID)

The protocol field is two octets its value identifies the datagram encapsulated in the Information field of the packet. In the case of MP the PID also identifies the presence of a 4-octet MP header (or 2-octet, if negotiated).

A PID of Ox003d identifies the packet as MP data with an MP header.

The LCP packets and protocol states of the MLPPP session follow those defined by PPP in RFC 1661, *The Point-to-Point Protocol (PPP)*. The options used during the LCP state for creating an MLPPP NCP session are described below.

B & E Bits

The B&E bits are used to indicate the epoch of a packet. Ingress packets to the MLPPP process will have an MTU, which may or may not be larger than the MRRU of the MLPPP network. The B&E bits manage the fragmentation of ingress packets when it exceeds the MRRU.

The B-bit indicates the first (or beginning) packet of a given fragment. The E-bit indicates the last (or ending) packet of a fragment. If there is no fragmentation of the ingress packet both B&E bits are set true (=1).

Sequence Number

Sequence numbers can be either 12 or 24 bits long. The sequence number is zero for the first fragment on a newly constructed AVC bundle and increments by one for each fragment sent on that bundle. The receiver keeps track of the incoming sequence numbers on each link in a bundle and reconstructs the desired unbundled flow through processing of the received sequence numbers and B&E bits. For a detailed description of the algorithm refer to RFC 1990.

Information Field

The Information field is zero or more octets. The Information field contains the datagram for the protocol specified in the protocol field.

The MRRU will have the same default value as the MTU for PPP. The MRRU is always negotiated during LCP.

Padding

On transmission, the Information field of the ending fragment may be padded with an arbitrary number of octets up to the MRRU. It is the responsibility of each protocol to distinguish padding octets from real information. Padding must not be added to any but the last fragment (the E-bit set true).

FCS

The FCS field of each MP packet is inherited from the normal framing mechanism from the member link on which the packet is transmitted. There is no separate FCS applied to the reconstituted packet as a whole if transmitted in more than one fragment.

LCP

The Link Control Protocol (LCP) is used to establish the connection through an exchange of configure packets. This exchange is complete, and the LCP opened state entered, once a Configure-Ack packet has been both sent and received.

LCP allows for the negotiation of multiple options in a PPP session. MLPPP is somewhat different than PPP and therefore the following options are set for MLPPPP and not negotiated:

- → No async control character map
- → No link quality monitoring
- → No compound frames
- → No self-describing-padding

Any non-LCP packets received during this phase must be silently discarded.

Link Fragmentation and Interleaving Support

Link Fragmentation and Interleaving (LFI) provides the ability to interleave high priority traffic within a stream of fragmented lower priority traffic. This feature helps avoid excessive delays to high priority, delay-sensitive traffic over a low-speed link. This can occur if this traffic type shares a link with lower priority traffic that utilizes much larger frames. Without this ability, higher priority traffic must wait for the entire packet to be transmitted before being transmitted, which could result in a delay that is too large for the application to function properly

For example, if VoIP traffic is being sent over a DS-1 or fractional DS-1 which is also used for Best Effort Internet traffic, LFI could be used so the small (usually 64-128B) VoIP packets can be transmitted between the transmission of fragments from the lower priority traffic.

Figure 6 shows the sequence of events as low priority and high priority frames arrive and are handled by LFI.

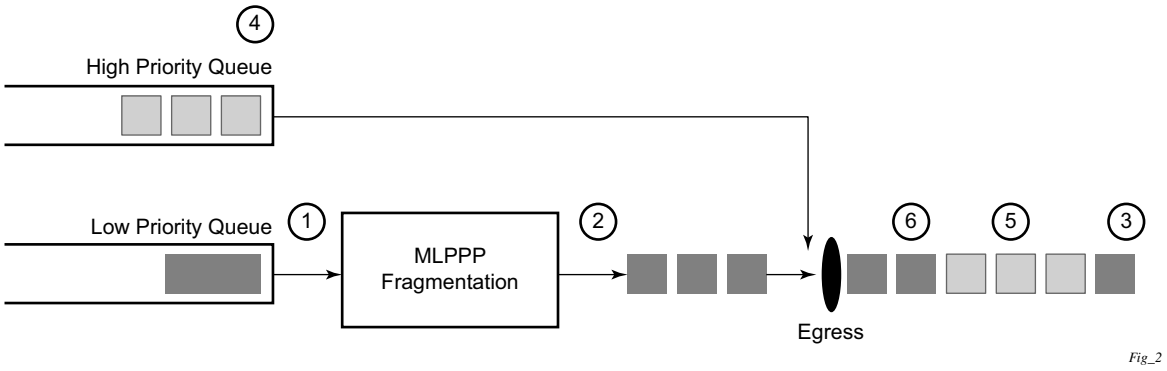

Figure 6: Frame Sequence of Events

1. A low priority frame arrives in the low priority queue. At this particular instant, there are no packets in the high priority queue so low priority frame is de-queued and passed to the fragmentation mechanism for MLPPP.
2. The original packet is divided into ‘n’ fragments based on the size of the packet and the fragment threshold configuration.
3. The fragments are then transmitted out the egress port.
4. After the transmission of the fragments has begun, high priority frames arrive in the high priority queue.
5. The transmission of the remaining fragments stops and the high priority packets are transmitted out the egress interface. Note that high priority packets are not fragmented.
6. When the high priority traffic is transmitted, the remaining lower priority fragments are then transmitted.

On the ingress side, LFI requires that the ingress port can receive non-fragmented packets within the fragment stream and pass these packets directly on to the forwarding engine and then continue with the reassembly process for the fragmented frames.

Multi-Class MLPPP

Multi-class MLPPP (MC-MLPPP) allows for the prioritization of multiple types of traffic flowing between the cell site routers and the mobile operator's aggregation routers. MC-MLPPP is an extension of the MLPPP standard which allows multiple classes of service to be transmitted over a MLPPP bundle. Originally (Figure 7), link fragmentation and interleaving (LFI) was added to MLPPP that allowed two classes, but in some applications, two classes of service can be insufficient.

The MLPPP header includes two class bits to allow for up to four classes of service (Figure 8). This enhancement to the MLPPP header format is detailed in RFC 2686, *The Multi-Class Extension to Multi-Link PPP*. This allows multiple classes of services over a single MLPPP connection and allows the highest priority traffic to be transmitted over the MLPPP bundle with minimal delay regardless of the order in which packets are received.

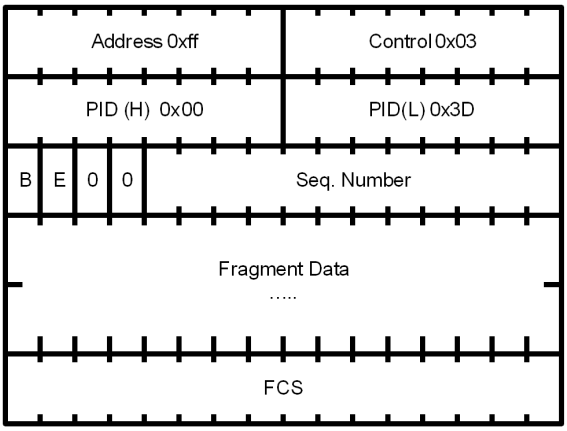

Figure 7: Original MLPPP Header Format

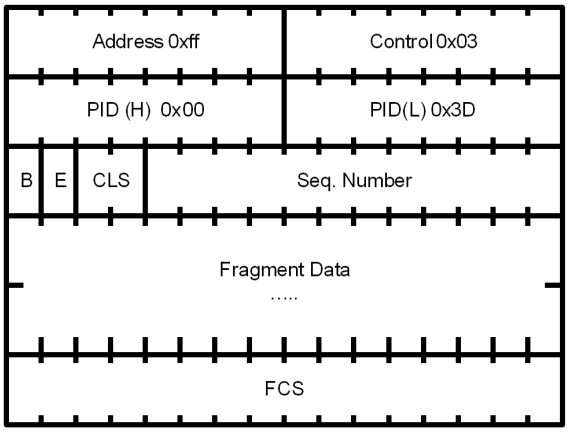

Figure 8: MC-MLPPP Short Sequence Header Format

The new MC-MLPPP header format uses the two (previously unused) bits before the sequence number as the class identifier. This allows four distinct classes of service to be identified into separate re-assembly contexts.

QoS in MC-MLPPP

If the user enables the multiclass option under an MLPPP bundle, the MDA egress data path provides a queue for each of the 4 classes of MLPPP. The user configures the required number of MLPPP classes to use on a bundle. The forwarding class of the packet, as determined by the ingress QoS classification, is used to determine the MLPPP class for the packet and hence which of the four egress MDA queues to store the packet. The mapping of forwarding class to MLPPP class is a function of the user configurable number of MLPPP classes. The default mapping for a 4-class, 3-class, and 2-class MLPPP bundle is shown in Table 7.

Table 7: Default Packet Forwarding Class to MLPPP Class Mapping

FC ID	FC Name	Scheduling Priority (Default)	MLPPP Class 4-class bundle	MLPPP Class 3-class bundle	MLPPP Class 2-class bundle
7	NC	Expedited	0	0	0
6	H1	Expedited	0	0	0
5	EF	Expedited	1	1	1
4	H2	Expedited	1	1	1
3	L1	Non-Expedited	2	2	1
2	AF	Non-Expedited	2	2	1
1	L2	Non-Expedited	3	2	1
0	BE	Non-Expedited	3	2	1

Table 8 shows a different mapping enabled when the user applies one of three pre-defined egress QoS profiles in the 4-class bundle configuration only.

Table 8: Packet Forwarding Class to MLPPP Class Mapping

FC ID	FC Name	Scheduling Priority (Default)	MLPPP Class (MLPPP Egress QoS profile 1, 2, and 3)
7	NC	Expedited	0
6	H1	Expedited	0
5	EF	Expedited	1
4	H2	Expedited	2
3	L1	Non-Expedited	2
2	AF	Non-Expedited	2
1	L2	Non-Expedited	2
0	BE	Non-Expedited	3

The MLPPP class queue parameters and its scheduling parameters are also configured by applying one of the three pre-defined egress QoS profiles to an MLPPP bundle.

Table 9 and Figure 9 provide the details of the class queue threshold parameters. Packets marked with a high drop precedence, such as out-of-profile, by the service or network ingress QoS policy will be discarded when any class queue reaches the OOP threshold. Packet with a low drop precedence marking, such as in-profile, will be discarded when any class queue reaches the max threshold.

Table 9: MLPPP Class Queue Threshold Parameters

	Class 0		Class 1		Class 2		Class 3	
Queue Threshold (in ms @ Available bundle rate)	Max	Oop	Max	Oop	Max	Oop	Max	Oop
2-Class Bundle Default Egress QoS Profile	250	125	750	375	N/A	N/A	N/A	N/A
3-Class Bundle Default Egress QoS Profile	50	25	200	100	750	375	N/A	N/A
4-Class Bundle Default Egress QoS Profile	10	5	50	25	150	75	750	375
4-Class Bundle Egress QoS Profile 1	25	12	5	3	200	100	1000	500
4-Class Bundle Egress QoS Profile 2	25	12	5	3	200	100	1000	500
4-Class Bundle Egress QoS Profile 3	25	12	5	3	200	100	1000	500

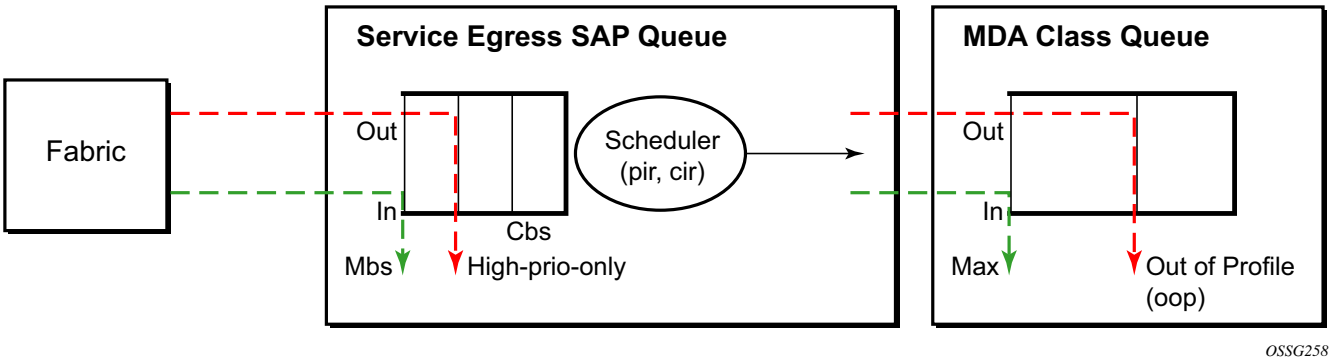

Figure 9: MLPPP Class Queue Thresholds for In-Profile and Out-of-Profile Packets

Table 10 and Figure 10 provide the details of the class queue scheduling parameters.

Table 10: MLPPP Class Queue Scheduling Parameters

		WRR Parameters		
4-class MLPPP Egress QoS Profile	MIR	W1	W2	W3
Profile 1	85%	<1%	66%	33%
Profile 2	90%	<1%	89%	10%
Profile 3	85%	<1%	87%	12%

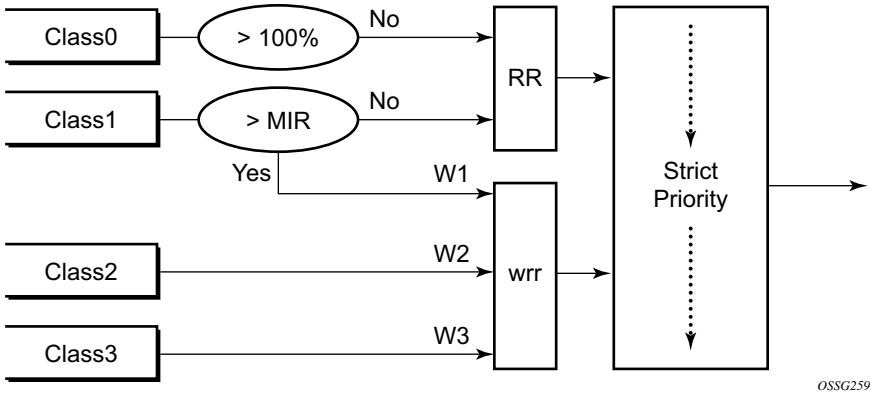

Figure 10: MLPPP Class Queue Scheduling Scheme

Note that all queue threshold and queue scheduling parameters are adjusted to the available bundle rate. If a member link goes down or a new member link is added to the bundle, the scheduling parameters MIR, W1, W2, W3, as well as the per class queue thresholds OOP and max are automatically adjusted to maintain the same values.

Class 0 queue is serviced at MLPPP at available bundle rate. Class 1 queue is guaranteed a minimum service rate but is allowed to share additional bandwidth with class 2 and 3 queues based on the configuration of WRR weight W1.

Class queues 2 and 3 can be given bandwidth guarantee by limiting MIR of class 1 queue to less than 100% and by setting the WRR weights W1, W2, and W3 to achieve the desired bandwidth distribution among all three class queues.

Note that there is one queue per bundle member link to carry link control packets, such as LCP: PPP, and which are serviced with strict priority over the 4 class queues (not shown).

In the default 2-class, 3-class, and 4-class egress QoS profile, the class queues are service with strict priority in ascending order of class number.

Ingress MLPPP Class Reassembly

For a MLPPP bundle with the multi-class option enabled, there is a default profile for setting the re-assembly timer value for each class. When the pre-defined MLPPP ingress QoS profile 1 is applied to a 4-class bundle, the values of the timers are modified as shown in Table 11.

Table 11: MLPPP Ingress QoS Profile: Reassembly Timers (msec)

	Class 0	Class 1	Class 2	Class 4
MLPPP ingress QoS default profile (2-Class bundle)	25ms	25ms	NA	NA
MLPPP ingress QoS default profile (3-Class bundle)	25ms	25ms	25ms	NA
MLPPP ingress QoS default profile (4-Class bundle)	25ms	25ms	100ms	1000ms
MLPPP ingress QoS profile 1 (4-class bundle)	10	10	100	1000

Configuring MC-MLPPP QoS Parameters

A 4-class MLPPP bundle can be configured with user-defined MLPPP QoS attributes. This feature cannot be used with MC-MLPPP bundles with fewer than 4 classes or with non-multiclass bundles.

The following describe the parameters and the configuration processes and rules

1. The user creates an ingress QoS profile in the **mlppp-profile-ingress** context, to configure a preferred value of the ingress per-class re-assembly timer. Ingress QoS profile 1 is reserved for the pre-defined profile with parameter values displayed in Table 11. The user is allowed to edit this profile and change parameter values. When a user creates a profile with a profile-id greater than 1, or performs the no option command on the parameter, the parameter's default value will always be the 1 in Table 11 for ingress QoS Profile #1 regardless of the parameter value the edited Profile 1 has at that point
2. The user creates an egress QoS profile in the **mlppp-profile-egress** context to configure preferred values for the per-class queue and queue scheduling parameters. The user can also configure system forwarding class mapping to the MLPPP classes. Egress QoS profiles 1, 2, and 3, are reserved for the pre-defined profiles with parameter values shown in Table 8, Table 9, or Table 10. Users can edit these profiles and change parameter values. When a user creates a profile with a profile-id higher than 3, or when the user specifies the no option command on the parameter, the default value will be the one shown in Table 8, Table 9, or Table 10 for the egress QoS Profile 1. This is regardless of the parameter value the edited profiles have at that point in time.
3. A maximum of 128 ingress and 128 egress QoS profiles can be created on the system.
4. The values of the ingress per-class re-assembly timer are configured in the ingress QoS profile.
5. The mapping of the system forwarding classes to the MLPPP Classes are configured in the egress QoS profile. There is a many-to-one relationship between the system FC and an MLPPP class. See Table 8 for the mapping when one of the three pre-defined 4-class egress QoS profiles is selected.
6. The maximum size for each MLPPP class queue in units of msec at the available bundle rate is configured in the egress QoS profile. This is referred to as max in Figure 9 and as max-queue-size in CLI. The out-of-profile threshold for an MLPPP class queue, referred to as oop in Figure 9, is not directly configurable and is set to 50% of the maximum queue size rounded up to the nearest higher integer value.
7. The MLPPP class queue scheduling parameters is configured in the egress QoS profile. The minimum information rate, referred to as **MIR** in Figure 10 and **mir** in CLI, applies to Class 1 queue only. The MIR parameter value is entered as a percentage of the available bundle rate. The WRR weight, referred to as W1, W2, and W3 in Figure 10 and weight in CLI, applies to class 1, class 2, and class 3 queues. Note that W1 in Figure 10 is not configurable and is internally set to a value of 1 such that Class 1 queue shares 1% of the available bundle rate when the sum of W1, W2, and W3 equals 100. W2 and W3 weights are integer values and are user configurable such that Class 2 queue shares (W2/

(W1 + W2 + W3)) and Class 3 queue shares (W3/(W1 + W2 + W3)) of the available bundle rate.

8. The user applies the ingress and egress QoS profiles to a 4-class MLPPP bundle for the configured QoS parameter values to take effect on the bundle.
9. The following operations require the bundles associated with a QoS profile to be shutdown to take effect.
   - → A change of the numbered ingress or egress QoS profile associated with a bundle.
   - → A change of the bundle associated ingress or egress QoS profile from default profile to a numbered profile and vice-versa.
10. The following operations can be performed without shutting down the associated bundles:
    - → Changes to any parameters in the ingress and egress QoS profiles.

The CLI commands for the creation of ingress and egress QoS profiles and configuration of the individual QoS parameters are described in the OS Quality of Service Guide.

Cisco HDLC

Cisco HDLC (cHDLC) is an encapsulation protocol for information transfer. It is a bit-oriented synchronous data-link layer protocol that specifies a data encapsulation method on synchronous serial links using frame characters and checksums.

cHDLC monitors line status on a serial interface by exchanging keepalive request messages with peer network devices. It also allows routers to discover IP addresses of neighbors by exchanging Serial Link Address Resolution Protocol (SLARP) (see SLARP on page 67) address-request and address-response messages with peer network devices.

The basic frame structure of a cHDLC frame is shown in Table 12. This frame structure is similar to PPP in an HDLC-link frame (RFC 1662, *PPP in HDLC-like Framing*). The differences to PPP in and HDLC-like frames are in the values used in the address, control, and protocol fields.

Table 12: cHDLC I-Frame

Flag	Address	Control	Protocol	Information Field	FCS
0x7E	0x0F/0x8F	0x00	—	—	16/32 bits

- Address field — The values of the address field include: 0x0F (unicast), 0x8F (broadcast).
- Control field — The control field is always set to value 0x00.
- Protocol field — The following values are supported for the protocol field:

Table 13: cHDLC Protocol Fields

Protocol	Field Value
IP	0x0800
Cisco SLARP	0x8035
ISO CLNP/ISO ES-IS DSAP/SSAP1	0xFEFE

- Information field — The length of the information field is in the range of 0 to 9Kbytes.
- FCS field — The FCS field can assume a 16-bit or 32-bit value. The default is 16-bits for ports with a speed equal to or lower than OC-3, and 32-bits for all other ports. The FCS for cHDLC is calculated in the same manner and same polynomial as PPP.

SLARP

An Alcatel-Lucent cHDLC interface will transmit a SLARP address resolution reply packet in response to a received SLARP address resolution request packet from peers. An Alcatel-Lucent cHDLC interface will not transmit SLARP address resolution request packets.

For the SLARP keepalive protocol, each system sends the other a keepalive packet at a user-configurable interval. The default interval is 10 seconds. Both systems must use the same interval to ensure reliable operation. Each system assigns sequence numbers to the keepalive packets it sends, starting with zero, independent of the other system. These sequence numbers are included in the keepalive packets sent to the other system. Also included in each keepalive packet is the sequence number of the last keepalive packet received from the other system, as assigned by the other system. This number is called the returned sequence number. Each system keeps track of the last returned sequence number it has received. Immediately before sending a keepalive packet, it compares the sequence number of the packet it is about to send with the returned sequence number in the last keepalive packet it has received. If the two differ by 3 or more, it considers the line to have failed, and will not route higher-level data across it until an acceptable keepalive response is received.

There is interaction between the SLARP address resolution protocol and the SLARP keepalive protocol. When one end of a serial line receives a SLARP address resolution request packet, it assumes that the other end has restarted its serial interface and resets its keepalive sequence numbers. In addition to responding to the address resolution request, it will act as if the other end had sent it a keepalive packet with a sequence number of zero, and a returned sequence number the same as the returned sequence number of the last real keepalive packet it received from the other end.

SONET/SDH Scrambling and C2-Byte

SONET/SDH scrambling and overhead for cHDLC follow the same rules used for POS (RFC 2615, *PPP over SONET/SDH*).

The two key SONET/SDH parameters are scrambling and signal-label (C2-byte). Scrambling is off by default. The default value of the C2-byte is 0xCF. These two parameters can be modified using the CLI. The other SONET overhead values (for example, j0) follow the same rules as the current POS implementation.

Timers

Cisco HDLC (cHDLC) has two timers associated with the protocol, the keepalive interval and the timeout interval. The keepalive interval is used to send periodic keepalive packets. The receiver process expects to receive a keepalive packet at the rate specified by the keepalive interval. The link is declared down if the receiver process does not receive a keepalive within the timeout interval. The link is declared up when the number of continual keepalive packets received equals the up-count.

It is recommended that the nodes at the two endpoints of the cHDLC link are provisioned with the same values.

Automatic Protection Switching (APS)

APS is designed to protect SONET/SDH equipment from linear unidirectional or bidirectional failures. The Network Elements (NEs) in a SONET/SDH network constantly monitor the health of the network. When a failure is detected, the network proceeds through a coordinated predefined sequence of steps to transfer (or switchover) live traffic to the backup facility (protection facility). This happens very quickly to minimize lost traffic. Traffic remains on the protection facility until the primary facility (working facility) fault is cleared, at which time the traffic may optionally be reverted to the working facility.

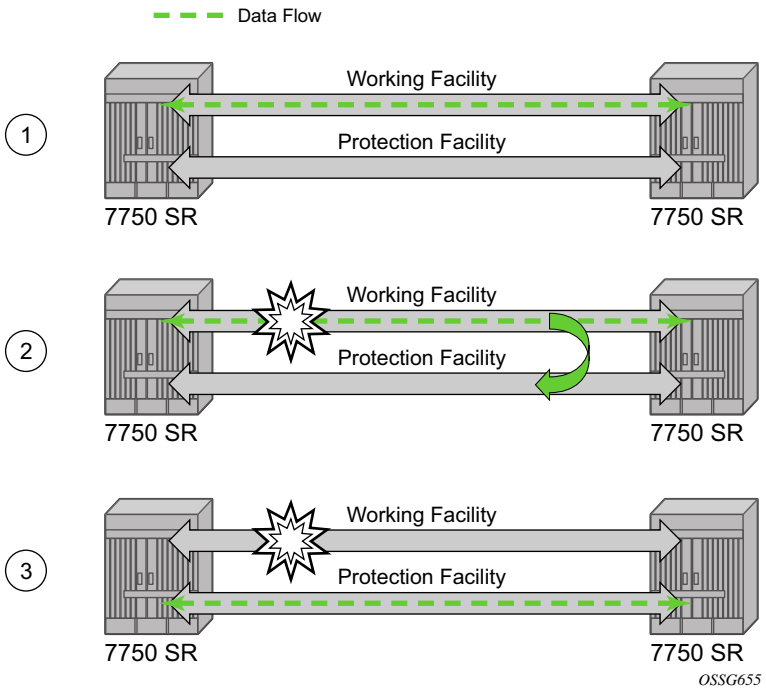

Figure 11: APS Protection (Single Chassis APS) and Switchover

Note that “facility” in the SR-OS context refers to the physical line (including intermediate transport/switching equipment) and directly attached line terminating hardware (SFP module, MDA and IOM). “Circuit” is also a term used for a link/facility (working-circuit).

A 1+1 APS group contains two circuits.

APS is configured on a port by port basis. If all ports on an MDA or IOM need to be protected then each port on the MDA or IOM must be individually added into an APS group.

Working and protection circuits can be connected to a variety of types of network elements (ADMs, DACSes, ATM switches, routers) and serve as an access or network port providing one or more services or network interfaces to the router. APS-protected SONET/SDH ports may be further channelized, and may contain bundled channels MLPPP or IMA Bundle Protection Groups). The ports may be one of a variety of encapsulation types as supported by the MDA including PPP, ATM, FR and more. For a definitive description of the MDAs, port types, switching modes, bundles and encapsulations supported with APS see APS Applicability, Restrictions and Interactions on page 88.

This section discusses the different APS architectures and their implementations.

Single Chassis and Multi-Chassis APS

APS can operate in a single chassis configuration (SC-APS) or in a multi-chassis configuration (MC-APS).

An SC-APS group can span multiple ports, MDAs or IOMs within a single node whereas as MC-APS can span two separate nodes.

Table 14: SC-APS versus MC-APS Protection

	Single Chassis APS	Multi-Chassis APS
Short form name	SC-APS	MC-APS
Link failure protection (including intermediate transmission equipment failure)	Yes	Yes
Optical/electrical module (SPF, XPF) failure protection	Yes	Yes
MDA failure protection	Yes	Yes
IOM failure protection	Yes	Yes
Node failure protection	No	Yes

The support of SC-APS and MC-APS depends on switching modes, MDAs, port types and encaps. For a definitive description of the MDAs, port types, switching modes, bundles and encapsulations supported with APS, see APS Applicability, Restrictions and Interactions on page 88.

APS on a Single Node (SC-APS)

In a single chassis APS both circuits of an APS group are terminated on the same node.

The working and protect lines of a single chassis APS group can be:

- Two ports on the same MDA.
- Two ports on different MDAs but on the same IOM.
- Two ports on different MDAs on two different IOMs (installed in different slots).

If the working and protection circuits are on the same MDA, protection is limited to the physical port and the media connecting the two devices. If the working and protection circuits are on different IOMs then protection extends to MDA or IOM failure. Figure 12 shows a configuration that provides protection against circuit, port, MDA or IOM failure on the 7750 SR connected to an Add-Drop-Multiplexer (ADM).

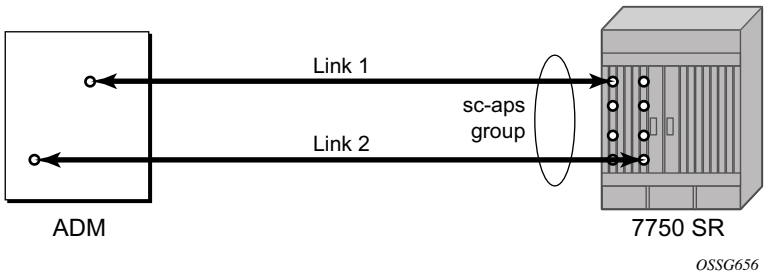

Figure 12: SC-APS Group with MDA and IOM Protection

APS Across Two Nodes (MC-APS)

Multi-Chassis APS functionality extends the protection offered by SC-APS to include protection against nodal (7750 SR) failure by configuring the working circuit of an APS group on one 7750 SR node while configuring the protect circuit of the same APS group on a different 7750 SR node.

These two nodes connect to each other with an IP link that is used to establish an MC-APS signalling path between the two 7750 SRs. Note that the working circuit and the protect circuit must have compatible configurations (such as the same speed, framing, and port-type). The relevant APS groups in both the working and protection routers must have same group ID, but they can have different names (for example, group port descriptions). Although the working and protection routers can be different platforms (7750 SR-7 and a 7750 SR-c12), switchover performance may be impacted so it is recommended to avoid a mix of platforms in the same MC-APS group where possible. The configuration consistency between the working circuit/router and

the protection circuit/router is not enforced by the 7750 SR. Service or network-specific configuration data is not signalled nor synchronized between the two service routers.

Signalling is provided using the direct connection between the two service routers. A heartbeat protocol can be used to add robustness to the interaction between the two routers. Signalling functionality includes support for:

- APS group matches between service routers.
- Verification that one side is configured as a working circuit and the other side is configured as the protect circuit. In case of a mismatch, a trap (incompatible neighbor) is generated.
- Change in working circuit status is sent from the working router to keep the protect router in sync.
- Protect router, based on K1/K2 byte data, member circuit status, and external request, selects the active circuit, and informs the working router to activate or de-activate the working circuit.

Note that external requests like lockout, force, and manual switches are allowed only on the APS group having the protection circuit.

The Figure 13 illustrates a Multi-Chassis APS group being used to protect against link, port, MDA, IOM or node failure.

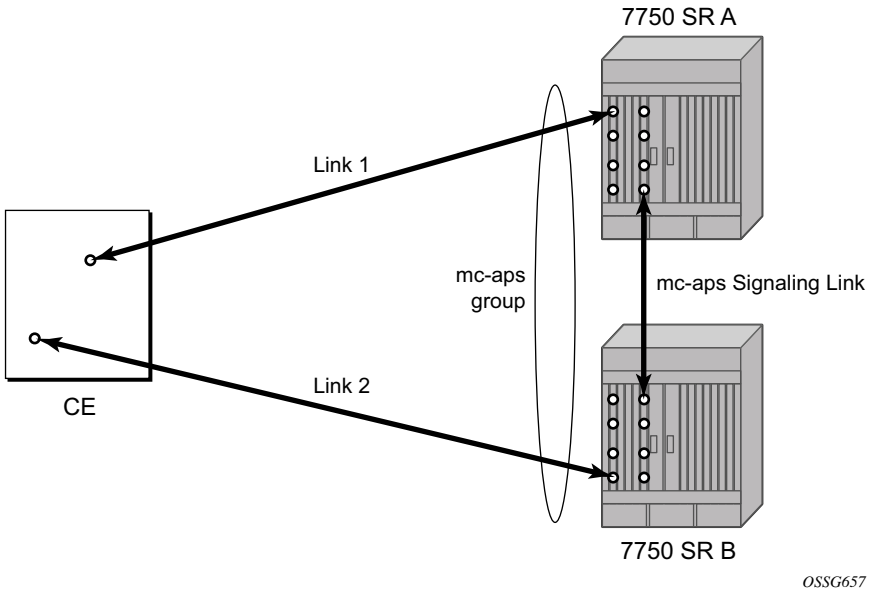

Figure 13: MC-APS Group Protects Against Node Failure

APS Switching Modes

APS behavior and operation differs based on the switching mode configured for the APS group. Several switching modes are supported in SR-OS.

The switching mode affects how the two directions of a link behave during failure scenarios and how APS tx operates.

Unidirectional / Bidirectional configuration must be the same at both sides of the APS group. The APS protocol (K byte messages) exchange switching mode information to ensure that both nodes can detect a configuration mismatch.

- If one end of an APS group is configured in a Unidirectional mode (Uni 1+1 Sig APS or Uni 1+1 Sig+Data APS) then the other end must also be configured in a Unidirectional mode (Uni 1+1 Sig+Data APS).
- If one end of an APS group is configured in a Bidirectional mode then the other end must also be configured in Bidirectional mode.

Table 15: APS Switching Modes

	Bidirectional 1+1 Signalling APS	Unidirectional 1+1 Signalling APS	Unidirectional 1+1 Signalling and Datapath APS
Short form name	Bidir 1+1 Sig APS	Uni 1+1 Sig APS	Uni 1+1 Sig+Data APS
CLI keyword	bi-directional	uni-directional	uni-1plus1
Interworks with a standards compliant APS implementation	Yes	Yes	Yes
Full 1+1 APS standards-based signalling	Yes	Yes	Yes
Data is transmitted simultaneously on both links/ circuits (1+1 Data)	No	No	Yes

The support of switching modes depends on SC-APS / MC-APS, MDAs, port types and encaps. For a definitive description of the MDAs, port types, switching modes, bundles and encapsulations supported with APS, see APS Applicability, Restrictions and Interactions on page 88.

Bidirectional 1+1 Signalling APS

In Bidir 1+1 Sig APS switching mode the Tx data is sent on the active link only (it is not bridged to both links simultaneously). 1+1 signalling, however, is used for full interoperability with signalling-compliant 1+1 architectures.

In the ingress direction (Rx), the decision to accept data from either the working or protection circuit is based on both locally detected failures/degradation and on what circuit the far-end is listening on (as indicated in the K bytes). If the far-end indicates that it has switched its active receiver, then the local SR-OS node will also switch its receiver (and Tx) to match the far-end. If the local Rx changes from one circuit to another it notifies the far end using the K bytes.

In the egress direction (Tx), the data is only transmitted on the active circuit. If the active Rx changes, then Tx will also change to the same circuit.

Bidirectional 1+1 Signalling APS ensures that both directions of active data flow (including both Rx) are using the same link/circuit (using the two directions of the same fiber pair) as required by the APS standards. If one end of the APS group changes the active receiver, it will signal the far end using the K bytes. The far end will then also change its receiver to listen on the same circuit.

Because the router transmits on active circuits only and keeps active TX and RX on the same port, both local and remote switches are required to restore the service.

The APS channel (bytes K1 and K2 in the SONET header – K bytes) is used to exchange requests and acknowledgments for protection switch actions. In Bidirectional 1+1 Signalling APS switching mode, the router sends correct status on the K bytes and requires the far-end to also correctly update/send the K-bytes to ensure that data is transmitted on the circuit on which the far-end has selected as its active receiver.

Line alarms are processed and generated independently on each physical circuit.

In Bidirectional 1+1 Signalling APS mode, the highest priority local request is compared to the remote request (received from the far end node using an APS command in the K bytes), and whichever has the greater priority is selected. The relative priority of all events that affect APS 1+1 protection is listed in the Table 16 on page 78 in descending order. The requests can be automatically initiated (such as signal failure or signal degrade), external (such as lockout, forced switch, request switch), and state requests (such as revert-time timers, etc.).

Unidirectional 1+1 Signalling APS

In Uni 1+1 Sig APS switching mode the Tx data is sent on the active link only (it is not bridged to both links simultaneously). 1+1 signalling, however, is used for full interoperability with signalling-compliant 1+1 architectures.

In the ingress direction (Rx), the decision to accept data from either the working or protection circuit is based on both locally detected failures/degradation and on what circuit the far-end is listening on (as indicated in the K bytes). Although it is not required in the APS standards, the SR-OS implementation of Unidirectional 1+1 Signalling APS uses standards based signaling to keep both the Rx and Tx on the same circuit / port. If the far-end indicates that it has switched its active receiver, then the local SR-OS node will also switch its receiver (and Tx) to match the far-end. If the local Rx changes from one circuit to another it notifies the far end using the K bytes.

In the egress direction (Tx), the data is only transmitted on the active circuit. If the active Rx changes, then Tx will also change to the same circuit.

Because the router transmits on active circuits only and keeps active TX and RX on the same port, both local and remote switches are required to restore the service. For a single failure a data outage is limited to a maximum of 100 milliseconds.

The APS channel (bytes K1 and K2 in the SONET header – K bytes) is used to exchange requests and acknowledgments for protection switch actions. In Unidirectional 1+1 Signalling APS switching mode, the router sends correct status on the K bytes and requires the far-end to also correctly update/send the K-bytes to ensure that data is transmitted on the circuit on which the far-end has selected as its active receiver.

Line alarms are processed and generated independently on each physical circuit.

In Unidirectional 1+1 Signalling APS switching mode:

- K-bytes are generated/transmitted based on local request/condition only (as required by the APS signalling).
- Local request priority is compliant to 1+1 U-APS specification.
- RX and TX are always forced on to the same (active) circuit (bi-directional). This has the following caveats:
  - → If an APS switch is performed due to a local condition, then the TX direction will be moved as well to the newly selected RX circuit (old inactive). The router will send LAIS on the old active TX circuit to force the remote end to APS switch to the newly active circuit. Note that some local request may not cause an APS switch when a remote condition prevents both RX and TX direction to be on the same circuit (for example an SD detected locally on a working circuit will not cause a switch if the protection circuit is locked out by the remote end).

→ If the remote end indicates an APS switch and the router can RX and TX on the circuit newly selected by the remote end, then the router will move its TX direction and will perform an APS switch of its RX direction (unless the router already TX and RX on the newly selected circuit).

→ If the remote end indicates an APS switch and the router cannot RX and TX on the circuit newly selected by the remote end (for example due to a higher priority local request, like a force request or manual request, etc.), then L-AIS are sent on the circuit newly selected by the remote end to force it back to the previously active circuit.

→ The sent L-AIS in the above cases can be either momentary or persistent. The persistent L-AIS is sent under the following conditions:

– On the protection circuit when the protection circuit is inactive and cannot be selected due to local SF or Lockout Request.

– On the working circuit as long as the working circuit remains inactive due to a local condition. The persistent L-AIS is sent to prevent revertive switching at the other end.

In all other cases a momentary L-AIS is sent. SR-OS provides debugging information that informs operators about the APS-induced L-AIS.

Unidirectional 1+1 Signalling and Datapath APS

Uni 1+1 Sig+Data APS supports unidirectional switching operations, 1+1 signaling and 1+1 data path.

In the ingress direction (Rx) switching is done based on local requests only as per the APS specifications. K-bytes are used to signal the far end the APS actions taken.

In the egress direction (Tx), the data is transmitted on both active and protecting circuits.

Each end of the APS group may be actively listening on a different circuit.

The APS channel (bytes K1 and K2 in the SONET header) is used to exchange APS protocol messages.

In Uni 1+1 Sig+Data APS a received L-RDI signal on the active circuit does not cause that circuit (port) to be placed out of service. The APS group can continue to use that circuit as the active receiver. This behavior is not configurable.

Uni 1+1 Sig+Data APS also supports configurable:

- Debounce timers for signal failure and degradation conditions
- Suppression of L-RDI alarm generation

APS Channel and SONET Header K Bytes

The APS channel (bytes K1 and K2 in the SONET header) is used to exchange APS protocol messages for all APS modes.

K1 Byte

The switch priority of a request is assigned as indicated by bits 1 through 4 of the K1 byte (as described in the rfc3498 APS-MIB).

Table 16: K1 Byte, Bits 1-4: Type of Request

Bit 1234	Condition
1111	Lockout of protection
1110	Force switch
1101	SF - High priority
1100	SF - Low priority
1011	SD - High priority
1010	SD - Low priority
1001	(not used)
1000	Manual switch
0111	(not used)
0110	Wait-to-restore
0101	(not used)
0100	Exercise
0011	(not used)
0010	Reverse request
0001	Do not revert
0000	No request

The channel requesting switch action is assigned by bits 5 through 8. When channel number 0 is selected, the condition bits show the received protection channel status. When channel number 1 is selected, the condition bits show the received working channel status. Channel values of 0 and 1 are supported.

Table 17 displays bits 5-8 of a K1 byte and K2 Bits 1-4 and the channel number code assignments.

Table 17: K1 Byte, Bits 5-8 (and K2 Bits 1-4), Channel Number Code Assignments

Channel Number Code	Channel and Notes
0	Null channel. SD and SF requests apply to conditions detected on the protection line. For 1+1 systems, Forced and Request Switch requests apply to the protection line. Only code 0 is used with Lockout of Protection request.
1 — 14	Working channel. Only code 1 applies in a 1+1 architecture. Codes 1 through n apply in a 1:n architecture. SD and SF conditions apply to the corresponding working lines.
15	Extra traffic channel. May exist only when provisioned in a 1:n architecture. Only No Request is used with code 15.

K2 Byte

The K2 byte is used to indicate the bridging actions performed at the line-terminating equipment (LTE), the provisioned architecture and mode of operation.

The bit assignment for the K2 byte is listed in Table 18.

Table 18: K2 Byte Functions

Bits 1-8		Function
1 — 4	Channel number. The 7750 SR supports only values of 0 and 1.	
5	0	Provisioned for 1+1 mode.
	1	Provisioned for 1:n mode.
6-8	111	Line AIS
	110	Line RDI
	101	Provisioned for bi-directional switching
	100	Provisioned for uni-directional switching
	011	(reserved for future use)
	010	(reserved for future use)
	001	(reserved for future use)
	000	(reserved for future use)

Differences in SONET/SDH Standards for K Bytes

SONET and SDH standards are slightly different with respect to the behavior of K1 and K2 Bytes.

Table 19 depicts the differences between the two standards.

Table 19: Differences Between SONET and SDH Standards

	SONET	SDH	Comments
SONET/SDH standards use different codes in the transmitted K1 byte (bits 1-4) to notify the far-end of a signal fail/signal degrade detection.	1100 for signal fail 1010 for signal degrade 1101 unused 1011 unused	1101 for signal fail 1011 for signal degrade 1100 unused 1010 unused	None
SONET systems signal the switching mode in bits 5-8 of the K2 byte whereas SDH systems do not signal at all.	101 for bi-dir 100 for uni-dir	Not used. 000 is signaled in bits 5 to 8 of K2 byte for both bi-directional as well as uni-directional switching.	SONET systems raise a mode mismatch alarm as soon as a mismatch in the TX and RX K2 byte (bits 5 to 8) is detected. SDH systems do not raise the mode mismatch alarm.

Failures Indicated by K Bytes

The following sections describe failures indicated by K bytes.

APS Protection Switching Byte Failure

An APS Protection Switching Byte (APS-PSB) failure indicates that the received K1 byte is either invalid or inconsistent. An invalid code defect occurs if the same K1 value is received for 3 consecutive frames (depending on the interface type (framer) used, the 7750 SR may not be able to strictly enforce the 3 frame check per GR-253 and G.783/G.841) and it is either an unused code or irrelevant for the specific switching operation. An inconsistent APS byte defect occurs when no three consecutive received K1 bytes of the last 12 frames are the same.

If the failure detected persists for 2.5 seconds, a Protection Switching Byte alarm is raised. When the failure is absent for 10 seconds, the alarm is cleared. This alarm can only be raised by the active port operating in bi-directional mode.

APS Channel Mismatch Failure

An APS channel mismatch failure (APS-CM) identifies that there is a channel mismatch between the transmitted K1 and the received K2 bytes. A defect is declared when the received K2 channel number differs from the transmitted K1 channel number for more than 50 ms after three identical K1 bytes are sent. The monitoring for this condition is continuous, not just when the transmitted value of K1 changes.

If the failure detected persists for 2.5 seconds, a channel mismatch failure alarm is raised. When the failure is absent for 10 seconds, the alarm is cleared. This alarm can only be raised by the active port operating in a bi-directional mode.

APS Mode Mismatch Failure

An APS mode mismatch failure (APS-MM) can occur for two reasons. The first is if the received K2 byte indicates that 1:N protection switching is being used by the far-end of the OC-N line, while the near end uses 1+1 protection switching. The second is if the received K2 byte indicates that uni-directional mode is being used by the far-end while the near-end uses bi-directional mode.

This defect is detected within 100 ms of receiving a K2 byte that indicates either of these conditions. If the failure detected persists for 2.5 seconds, a mode mismatch failure alarm is raised. However, it continues to monitor the received K2 byte, and should it ever indicate that the far-end has switched to a bi-directional mode the mode mismatch failure clearing process starts. When the failure is absent for 10 seconds, the alarm is cleared, and the configured mode of 1+1 bidirectional is used.

APS Far-End Protection Line Failure

An APS far-end protection line (APS-FEPL) failure corresponds to the receipt of a K1 byte in 3 consecutive frames that indicates a signal fail (SF) at the far end of the protection line. This forces the received signal to be selected from the working line.

If the failure detected persists for 2.5 seconds, a far-end protection line failure alarm is raised. When the failure is absent for 10 seconds, the alarm is cleared. This alarm can only be raised by the active port operating in a bi-directional mode.

Revertive Switching

The APS implementation also provides the revertive and non-revertive modes with non-revertive switching as the default option. In revertive switching, the activity is switched back to the working port after the working line has recovered from a failure (or the manual switch is cleared). In non-revertive switching, a switch to the protection line is maintained even after the working line has recovered from a failure (or if the manual switch is cleared).

A revert-time is defined for revertive switching so frequent automatic switches as a result of intermittent failures are prevented. A change in this value takes effect upon the next initiation of the wait to restore (WTR) timer. It does not modify the length of a WTR timer that has already been started. The WTR timer of a non-revertive switch can be assumed to be infinite.

In case of failure on both working and the protection line, the line that has less severe errors on the line will be active at any point in time. If there is signal degrade on both ports, the active port that failed last will stay active. When there is signal failure on both ports, the working port will always be active. The reason is that the signal failure on the protection line is of a higher priority than on the working line.

Bidirectional 1+1 Switchover Operation Example

Table 20 outlines the steps that a bi-directional protection switching process will go through during a typical automatic switchover.

Table 20: Actions for the Bi-directional Protection Switching Process

Status	APS Commands Sent in K1 and K2 Bytes on Protection Line		Action	
	B -> A	A -> B	At Site B	At Site A
No failure (Protection line is not in use)	No request	No request	No action	No action
Working line Degraded in direction A->B	SD on working channel 1	No request	Failure detected, notify A and switch to protection line.	No action
Site A receives SD failure condition	Same	Reverse request	No action	Remote failure detected, acknowledge and switch to protection line.
Site B receives Reverse request	Same	Same	No action	No action

Annex B (1+1 Optimized) Operation

Operation and behavior conferment with Annex B of ITU.T G.841 can be configured for an APS group.

Characteristics of this mode include are the following:

- Annex B operates in non-revertive bi-directional switching mode only as defined in G.841.
- Annex B in SR-OS operates with 1+1 signaling, but 1:1 data path where by data is transmitted on the active link only.
- K bytes are transmitted on both circuits.

Due to the request/reverse-request nature of an Annex B switchover, the data outage is longer than a typical (non Annex B single chassis) APS switchover. IMA bundles that are protected with Annex B APS have to resynchronize after a switchover. It is recommended to use maintenance commands (**tools>perform>aps…**) for planned switchovers (not MDA or IOM shutdown) to minimize the outage.

Annex B APS Outage Reduction Optimization

Typical standard Annex B behavior when a local SF is detected on the primary section (circuit), and this SF is the highest priority request on both the local side and from the remote side as per the APS specifications, is to send a request to the remote end and then wait until a reverse request is received before switching over to the secondary section. To reduce the recovery time for traffic, SR-OS will switch over to the secondary section immediately upon detecting the local SF on the primary section instead of waiting for the reverse request from the remote side. If the remote request is not received after a period of time then an “PSB Failure is declared” event is raised (Protection Switching Byte Failure – indicates an inconsistent or invalid Rx K1 Bytes), and the APS group on the local side switches back to the primary section.

When the remote side is in Lockout, and a local SF is detected then a reverse request will not be received by the local side. In this case, the traffic will no longer flow on the APS group since neither the primary nor secondary sections can carry traffic, and the outage reduction optimization will cause a temporary switchover from the primary to the secondary and then back again (which causes no additional outage or traffic issue since neither section is usable). If this temporary switchover is not desired then it is recommended to either perform Lockout from the 7x50 side, or to Lockout both sides, which will avoid the possibility of the temporary switchover.

Failures detected on the secondary section cause immediate switch over as per the Annex B specification. There is no outage reduction optimization in SR-OS for this case as it is not needed.

Some examples of events that can cause a local SF to be detected include: a cable being cut, laser transmitter or receiver failure, a port administratively “shutdown”, MDA failure or shutdown, IOM failure or shutdown.

Note: In Annex B operation, all switch requests are for a switch from the primary section to the secondary section. Once a switch request clears normally, traffic is maintained on the section to which it was switched by making that section the primary section. The primary section may be working circuit 1 or working circuit 2 at any particular moment.

Protection of Upper Layer Protocols and Services

APS prevents upper layer protocols and services from being affected by the failure of the active circuit.

The following example with figures and description illustrate how services are protected during a single-chassis APS switchover.

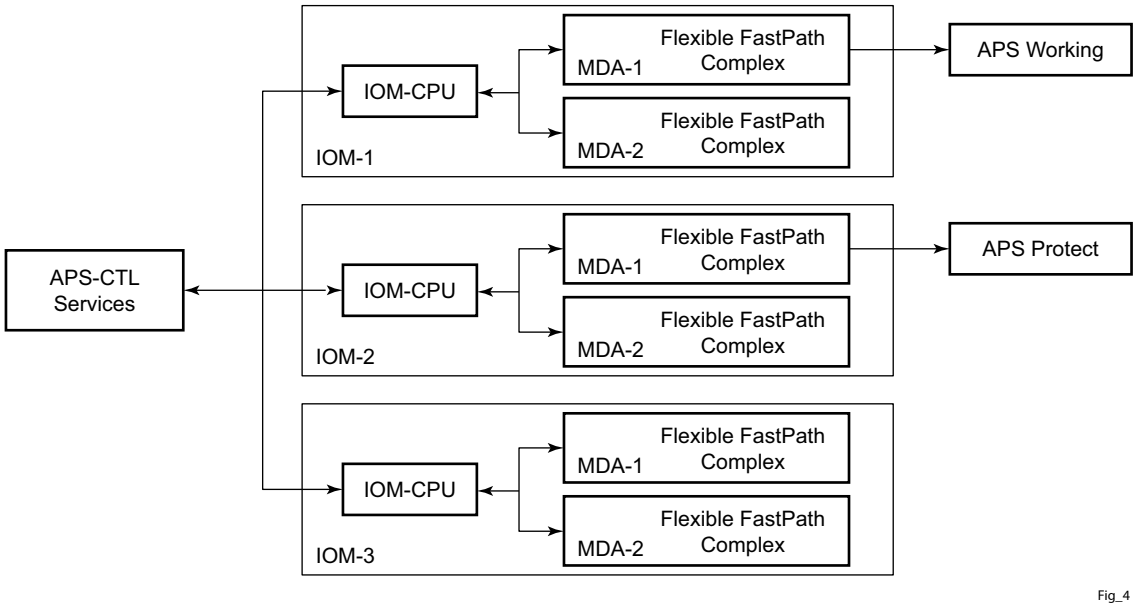

Figure 14: APS Working and Protection Circuit Example

Figure 14 is an example in which the APS working circuit is connected to IOM-1 / MDA-1 and the protection circuit is connected to IOM-2 / MDA-1. In this example, assume that the working circuit is currently used to transmit and receive data.

Switchover Process for Transmitted Data

For packets arriving on all interfaces that need to be transmitted over APS protected interfaces, the next hop associated with all these interfaces are programmed in all Flexible Fast-Path complexes

in each MDA with a logical next-hop index. This next hop-index identifies the actual next-hop information used to direct traffic to the APS working circuit on IOM-1 / MDA-1.

All Flexible Fast-Path complexes in each MDA are also programmed with next hop information used to direct traffic to the APS protect circuit on IOM-2/MDA-1. When the transmitted data needs to be switched from the working to the protect circuit, only the relevant next hop indexes need to be changed to the pre-programmed next-hop information for the protect circuit on IOM-2 / MDA-1.

Although the control CFM/CPM on the SF/CPM blade initiates the changeover between the working to protect circuit, the changeover is transparent to the upper layer protocols and service layers that the switchover occurs.

Physical link monitoring of the link is performed by the CPU on the relevant IOM for both working and protect circuits.

Switchover Process for Received Data

The Flexible Fast-Path complexes for both working and protect circuits are programmed to process ingress. The inactive (protect) circuit however is programmed to ignore all packet data. To perform the switchover from working circuit to the protect circuit the Flexible Fast-Path complex for the working circuit is set to ignore all data while the Flexible Fast-Path complex of the protect circuit will be changed to accept data.

The ADM or compatible head-end transmits a valid data signal to both the working and protection circuits. The signal on the protect line will be ignored until the working circuit fails or degrades to the degree that requires a switchover to the protect circuit. When the switchover occurs all services including all their QoS and filter policies are activated on the protection circuit.

APS User-Initiated Requests

The following sections describe APS user-initiated requests.

Lockout Protection

The lockout of protection disables the use of the protection line. Since the **tools>perform>aps>lockout** command has the highest priority, a failed working line using the protection line is switched back to itself even if it is in a fault condition. No switches to the protection line are allowed when locked out.

Request Switch of Active to Protection

The request or manual switch of active to protection command switches the active line to use the protection line unless a request of equal or higher priority is already in effect. If the active line is already on the protection line, no action takes place.

Request Switch of Active to Working

The request or manual switch of active to working command switches the active line back from the protection line to the working line unless a request of equal or higher priority is already in effect. If the active line is already on the working line, no action takes place.

Forced Switching of Active to Protection

The forced switch of active to protection command switches the active line to the protection line unless a request of equal or higher priority is already in effect. When the forced switch of working to protection command is in effect, it may be overridden either by a lockout of protection or by detecting a signal failure on the protection line. If the active line is already on the protection line, no action takes place.

Forced Switch of Active to Working

The forced switch of active to working command switches the active line back from the protection line to the working unless a request of equal or higher priority is already in effect.

Exercise Command

The exercise command is only supported in the bi-directional mode of the 1+1 architecture. The exercise command is specified in the **tools>perform>aps>force>exercise** context and exercises the protection line by sending an exercise request over the protection line to the tail-end and expecting a reverse request response back. The switch is not actually completed during the exercise routine.

APS and SNMP

SNMP Management of APS uses the APS-MIB (from rfc3498) and the TIMETRA-APS-MIB.

Table 21 shows the mapping between APS switching modes and MIB objects.

Table 21: Switching Mode to MIB Mapping

switching-mode	TIMETRA-APS-MIB tApsProtectionType	APS-MIB apsConfigDirection
Bidir 1+1 Sig APS (bi-directional)	onePlusOneSignalling (1)	bidirectional (2)
Uni 1+1 Sig APS (uni-directional)	onePlusOneSignalling (1)	unidirectional (1)
Uni 1+1 Sig+Data APS (uni-1plus1)	onePlusOne (2)	unidirectional (1)

apsConfigMode in the APS-MIB is set to onePlusOneOptimized for Annex B operation.

APS Applicability, Restrictions and Interactions

Note: The Release Notes for the relevant SR-OS release should be consulted for details about APS restrictions.

Table 22: Supported APS Mode Combinations

	Bidirectional 1+1 Signalling APS	Unidirectional 1+1 Signalling APS	Unidirectional 1+1 Signalling and Datapath APS
Single Chassis APS (SC-APS)	Supported	Supported	Supported (for 7750 SR-c4/ 12 platforms only)
Multi-Chassis APS (MC-APS)	Supported	Not supported	Not supported

APS and Bundles

Bundles (such as IMA and MLPPP) can be protected with APS through the use of Bundle Protection Groups (BPGRP). For APS-protected bundles, all members of a working bundle must reside on the working port of an APS group. Similarly all members of a protecting bundle must reside on the protecting circuit of that APS group.

IMA APS protection is supported only when the router is connected to another piece of equipment (possibly through an ADM) running a single IMA instance at the far end. By design, the IMA APS implementation is expected to keep the IMA protocol up as long as the far end device can tolerate some frame loss. Similarly, the PPP protocol state machine for PPP channels and MLPPP bundles remains UP when a switchover occurs between the working and protect circuits.

When APS protects IMA groups, IMA control cells, but not user traffic, are sent on the inactive circuit (as well as the active) to keep the IMA protocol up during an APS switch.

For details on MLFR/FRF.12 support with APS see the *MLFR/FRF.12 Support of APS, BFD, and Mirroring Features* section.

APS Switchover Impact on Statistics

All SAP-level statistics are retained with an APS switch. A SAP will reflect the data received regardless of the number of APS switches that has occurred. ATM statistics, however, are cleared after an APS switch. Thus, any ATM statistics viewed on an APS port are only the statistics since the current active member port became active.

Physical layer packet statistics on the APS group reflect what is currently on the active member port.

Port and path-level statistics follow the same behavior as described above.

Any SONET physical-layer statistics (for example, B1,B2,B3,...) on the APS port are only what is current on the active APS member port.

Supported APS MDA/Port Combinations

Table 23 displays examples of the port types that can be paired to provide APS protection. Both ports must be the same type and must be configured at the same speed.

Table 23: MDA/Port Type Pairing for APS

MDA Type	Unchannelized SONET/SDH (POS) For example: m16-oc12/3-sfp	ATM For example: m4-atmoc12/3-sfp	Circuit Emulation (CES) For example: m4-choc3-ces-sfp	Channelized Any Service Any Port (ASAP) For example: m1-choc12-as-sfp
Unchannelized SONET/SDH (POS) For example: m16-oc12/3-sfp	Supported			
ATM For example: m4-atmoc12/3-sfp		Supported		
Circuit Emulation (CES) For example: m4-choc3-ces-sfp			Supported	
Channelized Any Service Any Port (ASAP) For example: m1-choc12-as-sfp				Supported

For example, an APS group can be comprised of a pair of ports where each port is on one of the two following MDAs:

- m16-atmoc3-sfp
- m4-atmoc12/3-sfp (port in oc3 mode)

For example, an APS group can not be comprised of a pair of ports where one port is on an m16-oc12/3-sfp and the other port is on an m1-choc12-as-sfp.

APS Switchover During CFM/CPM Switchover

An APS switchover immediately before, during or immediately after a CFM/CPM switchover may cause a longer outage than normal.

Removing or Failure of a Protect MDA

The detection of a CMA/MDA removal or a CMA/MDA failure can take additional time. This can affect the APS switchover time upon the removal or failure of a protection CMA/MDA. If the removal is scheduled during maintenance, it is recommended that the port and/or protect circuit be shutdown first to initiate an APS switchover before the CMA/MDA maintenance is performed.

Mirroring Support

Mirroring parameters configured on a specific port or service, are maintained during an APS failover.

Sample APS Applications

The following sections provide sample APS application examples.

Sample APS Application: MLPPP with SC-APS and MC-APS on Channelized Interfaces

7750 and 7710 service routers support APS on channelized interfaces. This allows Alcatel-Lucent's service routers to be deployed as the radio access network (RAN) aggregation router which connects the base transceiver station (BTS) and the radio network controller (RNC).

Figure 15 displays an example of MLPPP termination on APS protected channelized OC-n/STM-n links. This example illustrates the following:

- SC-APS (the APS circuits terminate on the same node aggregation router A).
- APS protecting MLPPP bundles (bundles are between the BTS and aggregation router A, but APS operates on the SONET links between the DACS and the aggregation router).
- APS on channelized access interfaces (OC-3/OC-12 links)

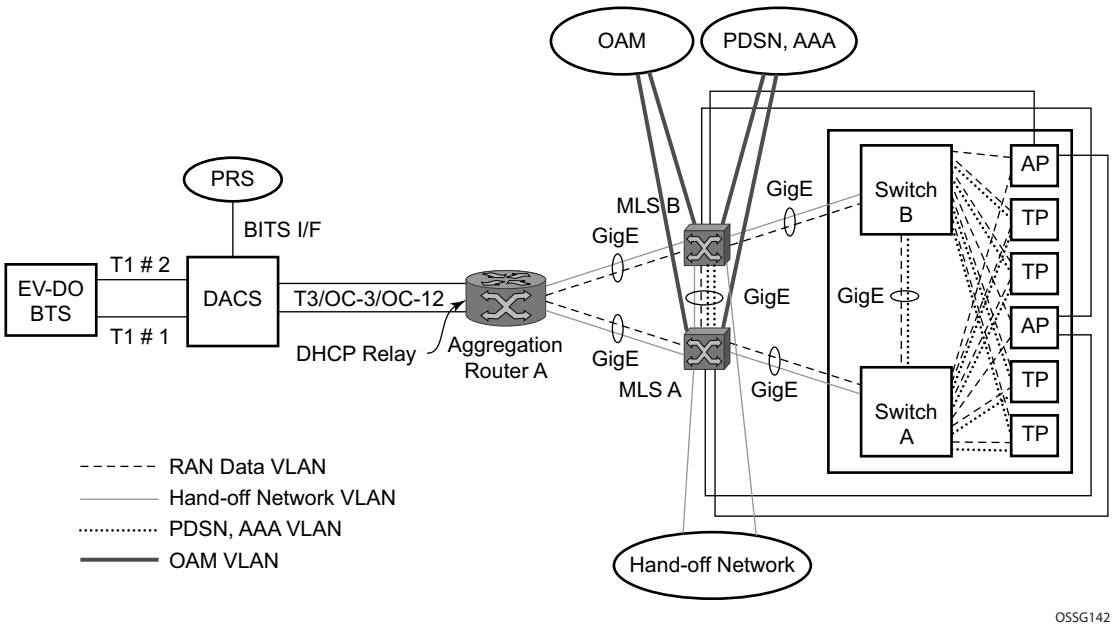

Figure 15: SC-APS MLPPP on Channelized Access Interfaces Example

Figure 16 depicts an APS group between a digital access cross-connect system (DACS) and a pair of aggregation routers. At one end of the APS group both circuits (OC-3/STM-1 and/or OC-12/STM-4 links) are terminated on the DACS and at the other end each circuit is terminated on a

different aggregation routers to provide protection against router failure. The MLPPP bundle operates between the BTS and the aggregation routers. At any one time only one of the two aggregation routers is actually terminating the MLPPP bundle (whichever aggregation router is processing the active APS circuit).

This example illustrates the following:

- MC-APS (the APS circuits terminate on different aggregation routers)
- APS protecting MLPPP bundles (bundles are between the BTS and the aggregation routers but APS operates on the SONET links between the DACS and the aggregation routers)
- APS on channelized access interfaces (OC-3/OC-12 links)

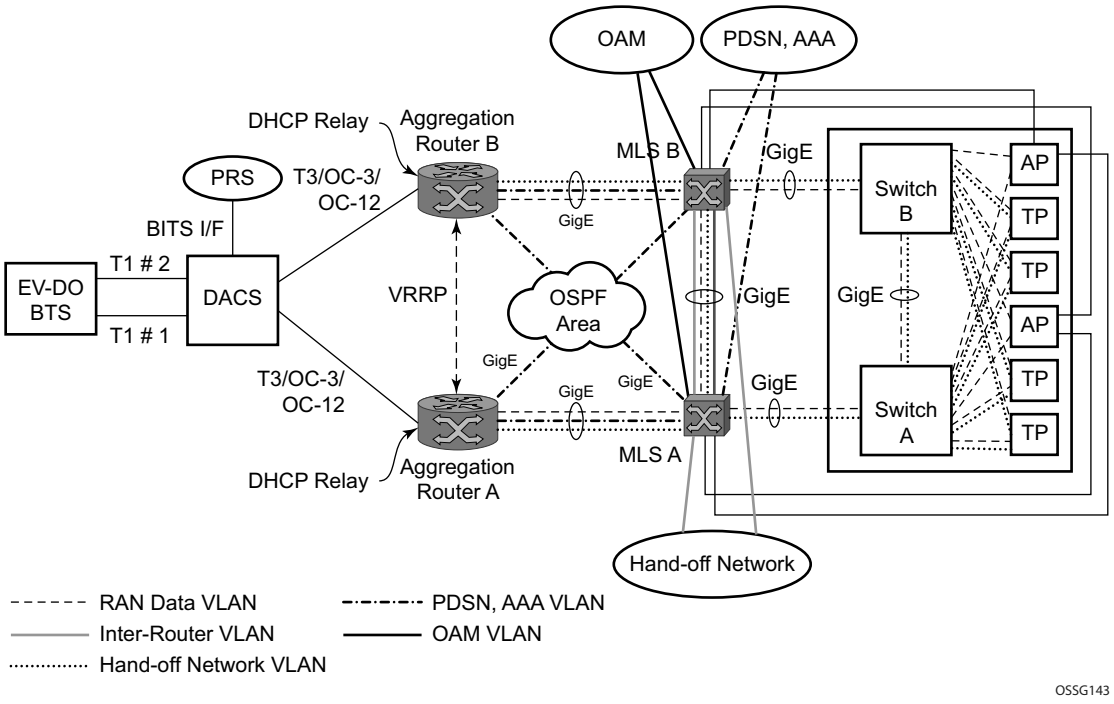

Figure 16: MC-APS MLPPP on Channelized Access Interfaces Example

Sample APS Application: MC-APS for ATM SAP with ATM VPLS Service

In Figure 17, service router A is connected to the ATM switch or 7670 through an OCx ATM 1 link. This link is configured as the working circuit. Service router B is connected to the same ATM switch or 7670 through an OCx ATM 2 link. This link is configured as the protection circuit.

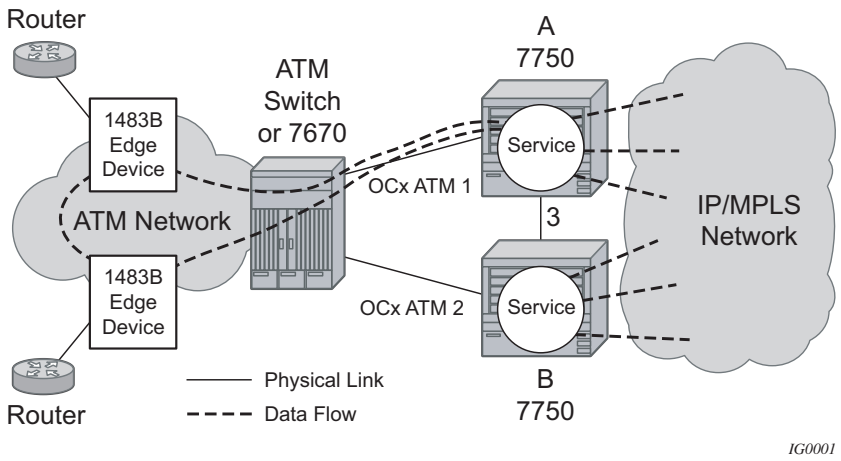

Figure 17: Multi-Chassis APS Application

Communication between service routers A and B is established through link 3. This link is for signalling. To guarantee optimum fail-over time between service routers A and B, link 3 must be a direct physical link between routers A and B.

Sample APS Application: MC-APS with VLL Redundancy

Support of MC-APS to ATM VLLs and Ethernet VLL with ATM SAPs allows MC-APS to operate with pseudowire redundancy in a similar manner that MC-LAG operates with pseudowire redundancy.

The combination of these features provides a solution for access node redundancy and network redundancy as shown in Figure 18.

MC-APS groups are configured as follows:

- MC-APS group between the MSAN on the left and Aggregation Nodes A & B
- MC-APS group between the MSAN on the right and Aggregation Nodes C & D

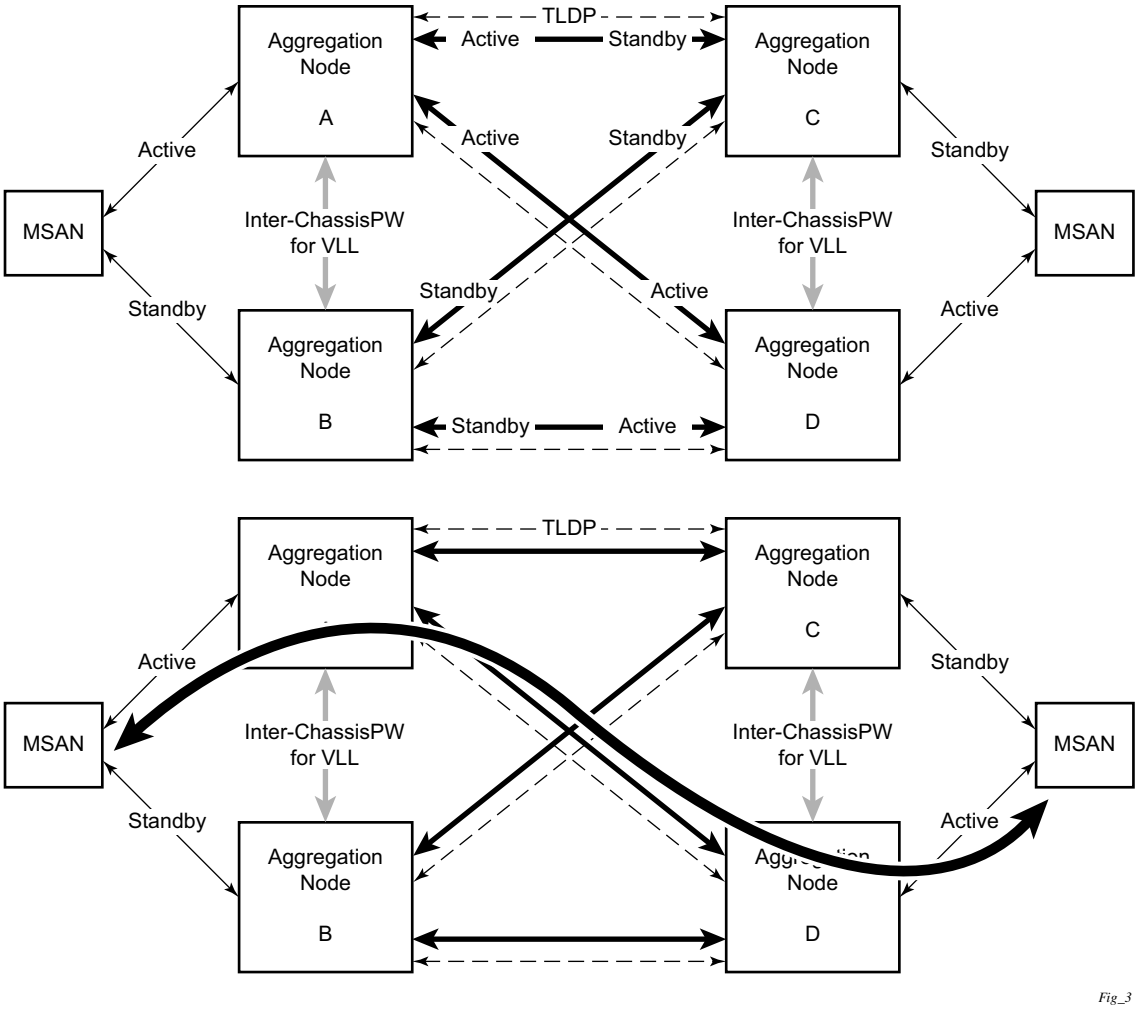

Figure 18: Access and Node and Network Resilience

An example of a customer application in the mobile market is displayed in Figure 19.

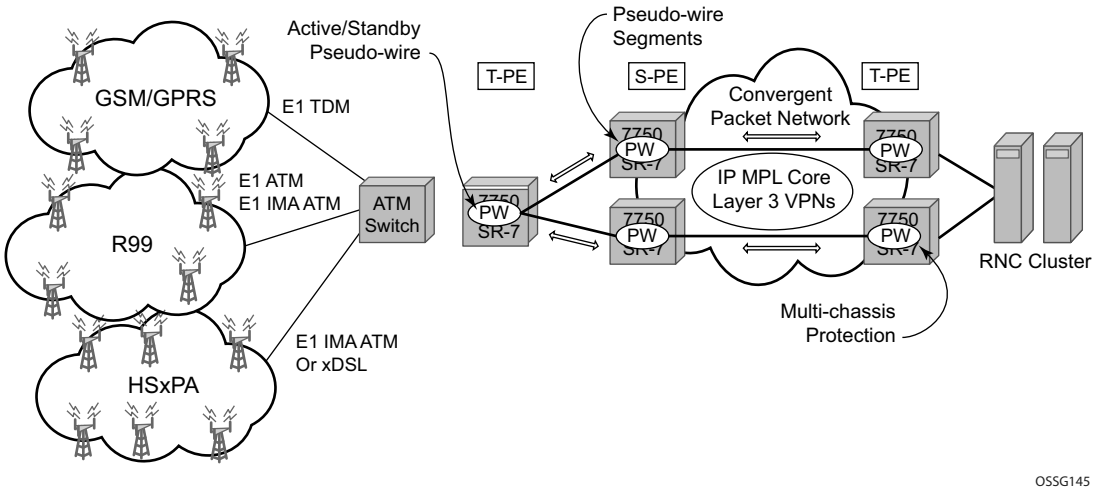

Figure 19: MC-APS with ATM VLL Redundancy

In the application show in Figure 19, 2G and 3G cell sites are aggregated into a Tier 2 or Tier 3 hub site before being backhauled to a Tier 1 site where the radio network controller (RNC) which terminates user calls is located. This application combines MC-APS on the RNC access side and pseudowire redundancy and pseudowire switching on the core network side. pseudowire switching is used in order to separate the routing domains between the access network and the core network.

Sample APS Application: RAN Aggregation with Microwave Radio Transport

Figure 20 displays a RAN aggregation network deployment example. In this example Uni-dir 1+1 Sig+Data APS is being used.

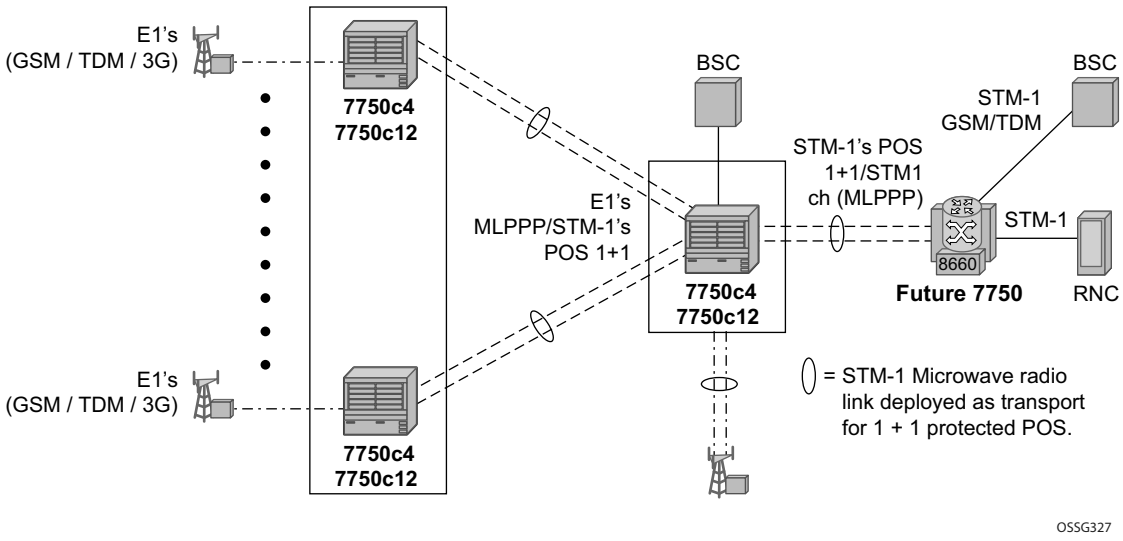

Figure 20: Mobile RAN with Microwave Transport Example

As depicted in Figure 20, some APS-protected interfaces may require microwave radio transport. Figure 21 depicts APS-protected links between two routers that use Microwave transport. The radio equipment acts as a SONET section/ SDH regenerator section equipment, yet it implements Unidirectional APS-like processing to provide equipment protection on the local/remote radio sites respectively.

The active RX line signal (switched independently from TX) is being transmitted over the radio link to the far end radio where the signal gets transmitted on both active and inactive circuits.

The radio reacts on APS triggered failures as detected by the segment termination function: LOS, LOF, manual APS commands, and optionally BER SF/SD. Since the radio does not terminate the SONET/SDH line layer, any line signaling (including Kbytes signaling for APS, line alarms like RDI/AIS) are not terminated by the radio and arrive at a far-end router.

Note that the far-end router can either send line alarms based on its active link status or based on physical circuit status (in which case for example, an L-RDI with a valid data will be received on the 77x0).

To facilitate a deployment such as shown in this example, some of following features of the 7750 SR-c12 routers are employed:

- Uni-dir 1+1 Sig+Data APS switching mode.
- Configurable L-RDI suppression.
- Active RX circuits are selected based on local conditions only. The SONET K Bytes are not needed to coordinate switch actions, but they are still used since they flow through and reach the far-end router.
- Ports are not failed on L-RDI, as L-RDI may be received on both ports momentarily, as a result of a local radio APS switch or, permanently as a result of a remote router APS switch (with remote radio selecting traffic from the TX line on the same port as failed RX line on the router).
- For some radio equipment, a radio can cause an APS switch resulting in the far end radio detecting radio alarm and generating L-AIS toward its locally attached router on both circuits. In some cases, that router also detects BER SD/BER SF conditions on both circuits as well. Therefore, to localize failure recovery, the 7750c12 can optionally debounce those alarms so a remote router does not invoke an APS switch on a local failure condition.

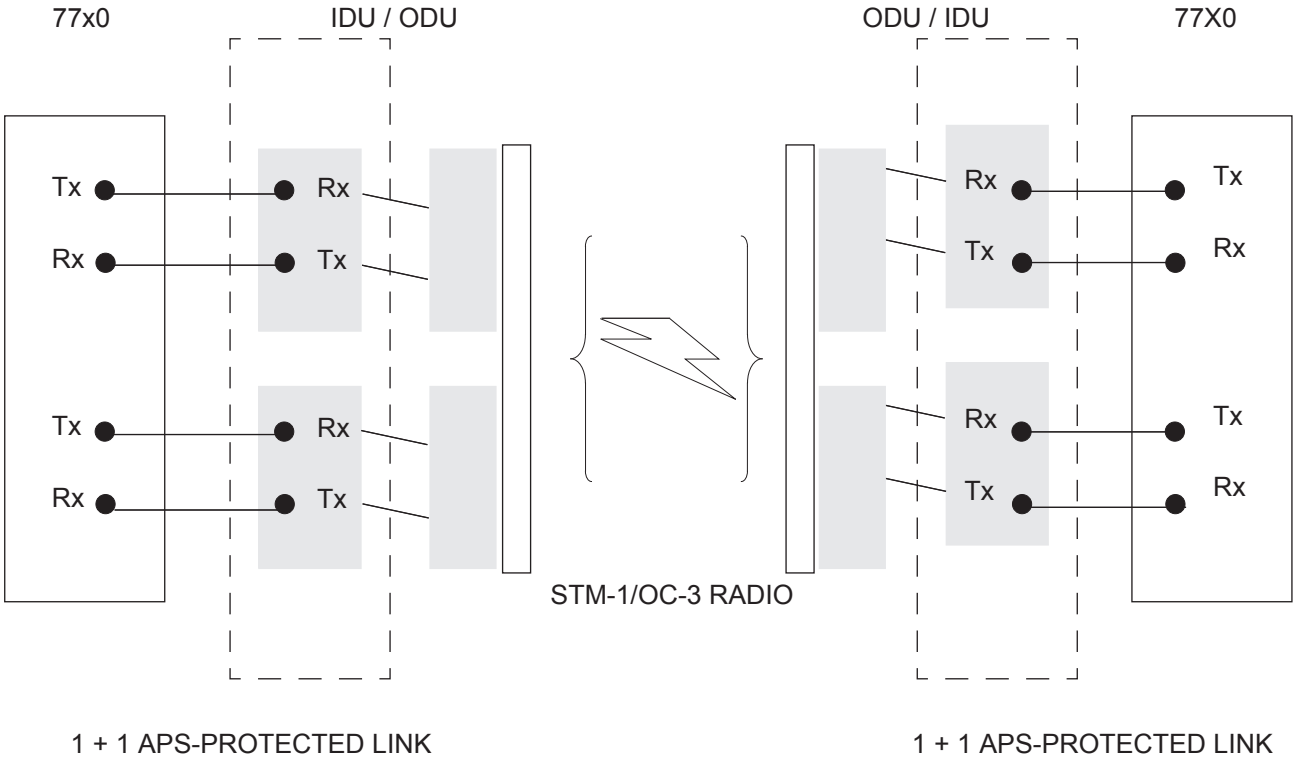

Figure 21: 1+1 APS Protected Microwave SDH Transport

Inverse Multiplexing Over ATM (IMA)

IMA is a cell based protocol where an ATM cell stream is inverse-multiplexed and de-multiplexed in a cyclical fashion among ATM-supporting channels to form a higher bandwidth logical link where the logical link concept is referred as an IMA group. By grouping channels into an IMA group, customers gain bandwidth management capability at in-between rates (for example, between E-1/DS-1 and E-3/DS-3 respectively) through addition/removal of channels to/from the IMA group.

In the ingress direction, traffic coming over multiple ATM channels configured as part of a single IMA group, is converted into a single ATM stream and passed for further processing to the ATM Layer where service-related functions, for example L2 TM, or feeding into a pseudowire are applied. In the egress direction, a single ATM stream (after service functions are applied) is distributed over all paths that are part of an IMA group after ATM layer processing takes place.

An IMA group interface compensates for differential delay and allows only for a minimal cell delay variation. The interface deals with links that are added, deleted or that fail. The higher layers see only an IMA group and not individual links, therefore service configuration and management is done using IMA groups, and not individual links that are part of it.

The IMA protocol uses an IMA frame as the unit of control. An IMA frame consists of a series of consecutive (128) cells. In addition to ATM cells received from the ATM layer, the IMA frame contains IMA OAM cells. Two types of cells are defined: IMA Control Protocol (ICP) cells and IMA filler cells. ICP cells carry information used by IMA protocol at both ends of an IMA group (for example IMA frame sequence number, link stuff indication, status and control indication, IMA ID, TX and RX test patters, version of the IMA protocol, etc.). A single ICP cell is inserted at the ICP cell offset position (the offset may be different on each link of the group) of each frame. Filler cells are used by the transmitting side to fill up each IMA frame in case there are not enough ATM stream cells from the ATM layer, so a continuous stream of cells is presented to the physical layer. Those cells are then discarded by the receiving end. IMA frames are transmitted simultaneously on all paths of an IMA group and when they are received out of sync at the other end of the IMA group link, the receiver compensates for differential link delays among all paths.

Inverse Multiplexing over ATM (IMA) Features

Hardware Applicability

IMA is supported on channelized ASAP MDAs.

Software Capabilities

Alcatel-Lucent's implementation supports IMA functionality as specified in ATM Forum's Inverse Multiplexing for ATM (IMA) Specification Version 1.1 (af-phy-0086.001, March 1999). The following details major functions

- TX Frame length — Only IMA specification default of 128 cells is supported.
- IMA version — Both versions 1.0 and 1.1 of IMA are supported. There is no support for automatically falling to version 1.0 if the far end advertises 1.0 support, and the local end is configured as 1.1. Due to potential protocol interoperability issues between IMA 1.0 implementations, it is recommended that IMA version 1.1 is used whenever possible.
- Alpha, beta, and gamma values supported are defaults required by the IMA specification (values of 2, 2, and 1 respectively).
- Clock mode — Only IMA specification default of common clock mode is supported (CTC).
- Timing reference link — The transmit timing reference link is chosen first among the active links in an IMA group. If none found, then it is chosen among the usable links or finally, among the unusable links.
- Cell Offset Configuration — The cell offsets for IMA links are not user configurable but internally assigned according to the recommended distribution described in the IMA spec.
- TX IMA ID — An internally assigned number equal to the IMA bundle number.
- Minimum Links — A configurable value is supported to control minimum member links required to be up for an IMA group to stay operationally up.
- Maximum Group Bandwidth — A configurable value is supported to specify maximum bandwidth available to services over an IMA group. The maximum may exceed the number of minimum/configured/active links allowing for overbooking of ATM shaped traffic.
- Symmetry mode — Only IMA specification default of symmetric operation and configuration is supported.
- Re-alignment — Errors that require a re-alignment of the link (missing or extra cells, corrupted frame sequence numbers), are dealt with by automatically resetting the IMA link upon detection of an error.

- Activation/Deactivation Link Delay Timers — Separate, configurable timers are supported defining the amount of delay between detection of LIF, LODS and RFI-IMA change and raising/clearing of a respective alarm to higher layers and reporting RXIFailed to the far end. This protocol dampening mechanism protects those higher layers from bouncing links.
- Differential delay — A configurable value of differential delay that will be tolerated among the members of the IMA group is supported. If a link exceeds the configured delay value, then LODS defect is declared and protocol management actions are initiated as required by the IMA protocol and as governed by Link Activation and Deactivation procedures. The differential delay of a link is calculated based on the difference between the frame sequence number received on the link and the frame sequence number received on the fastest link (a link on which the IMA frame was received first).
- Graceful link deletion — The option is supported for remotely originated requests only. To prevent data loss on services configured over an IMA group, it is recommended to initiate graceful deletion from the far end before a member link is deleted or a physical link is shutdown.
- IMA test pattern — Alcatel-Lucent's implementation supports test pattern procedures specified in the IMA specification. Test pattern procedures allow debugging of IMA group problems without affecting user data. Test pattern configurations are not preserved upon a router reboot.
- Statistics — Alcatel-Lucent's IMA implementation supports all standard-defined IMA group and IMA link status and statistics through proprietary TIMETRA-PORT-MIB. Display and monitoring of traffic related interface/SAP statistics is also available for IMA groups and services over IMA groups on par with physical ATM interfaces and services.
- Scaling — Up to 8 member links per IMA group, up to 128 groups per MDA and all DS-1/E-1 links configurable per MDA in all IMA groups per MDA are supported.

Ethernet Local Management Interface (E-LMI)

The Ethernet Local Management Interface (E-LMI) protocol is defined in Metro Ethernet Forum (MEF) technical specification MEF16. This specification largely based on Frame Relay - LMI defines the protocol and procedures that convey the information for auto-configuration of a CE device and provides the means for EVC status notification. MEF16 does not include link management functions like Frame Relay LMI does. In the Ethernet context that role is already accomplished with Clause 57 Ethernet OAM (formerly 802.3ah).

The SR OS currently implements the User Network Interface-Network (UNI-N) functions for status notification supported on Ethernet access ports with dot1q encapsulation type. Notification related to status change of the EVC and CE-VLAN ID to EVC mapping information is provided as a one to one between SAP and EVC.

The E-LMI frame encapsulation is based on IEEE 802.3 untagged MAC frame format using an ether-type of 0x88EE. The destination MAC address of the packet 01-80-C2-00-00-07 will be dropped by any 802.1d compliant bridge that does not support or have the E-LMI protocol enabled. This means the protocol cannot be tunneled.

Status information is sent from the UNI-N to the UNI-C, either because a status enquiry was received from the UNI-C or unsolicited. The Active and Not Active EVC status are supported. The Partially Active state is left for further study.

The bandwidth profile sub-information element associated with the EVC Status IE does not use information from the SAP QoS policy. A value of 0 is used in this release as MEF 16 indicates the bandwidth profile sub-IE is mandatory in the EVC Status IE. The EVC identifier is set to the description of the SAP and the UNI identifier is set to the description configured on the port. Further, the implementation associates each SAP with an EVC. Currently, support exists for CE-VLAN ID/EVC bundling mode.

As stated in the OAM Mapping section in the OAM and Diagnostics Guide, E-LMI the UNI-N can participates in the OAM fault propagation functions. This is a unidirectional update from the UNI-N to the UNI-C and interacting with service manager of VLL, VPLS, VPRN and IES services.

Link Layer Discovery Protocol (LLDP)

The IEEE 802.1ab Link Layer Discovery Protocol (LLDP) standard defines protocol and management elements that are suitable for advertising information to stations attached to the same IEEE 802 LAN (emulation) for the purpose of populating physical or logical topology and device discovery management information databases. The protocol facilitates the identification of stations connected by IEEE 802 LANs/MANs, their points of interconnection, and access points for management protocols.

Note that LAN emulation and logical topology wording is applicable to customer bridge scenarios (enterprise/carrier of carrier) connected to a provider network offering a transparent LAN emulation service to their customers. It helps the customer bridges detect misconnection by an intermediate provider by offering a view of the customer topology where the provider service is represented as a LAN interconnecting these customer bridges.

The IEEE 802.1ab standard defines a protocol that:

- Advertises connectivity and management information about the local station to adjacent stations on the same IEEE 802 LAN.
- Receives network management information from adjacent stations on the same IEEE 802 LAN.
- Operates with all IEEE 802 access protocols and network media.
- Establishes a network management information schema and object definitions that are suitable for storing connection information about adjacent stations.
- Provides compatibility with a number of MIBs as depicted in Figure 22.

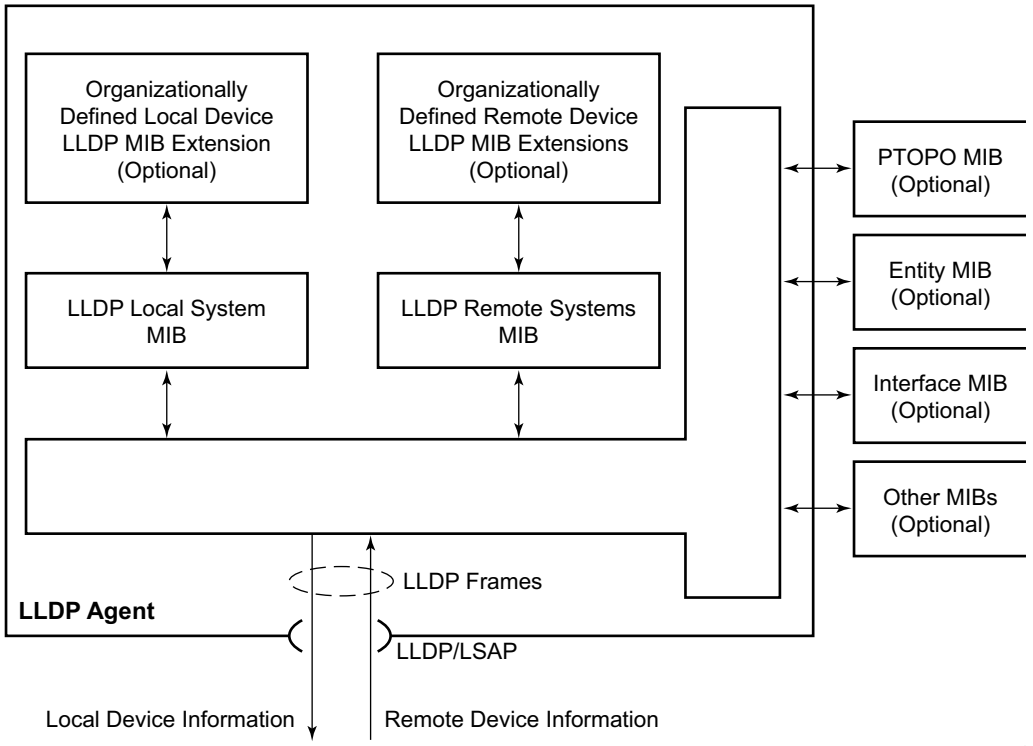

Figure 22: LLDP Internal Architecture for a Network Node

Network operators must be able to discover the topology information in order to detect and address network problems and inconsistencies in the configuration. Moreover, standard-based tools can address the complex network scenarios where multiple devices from different vendors are interconnected using Ethernet interfaces.

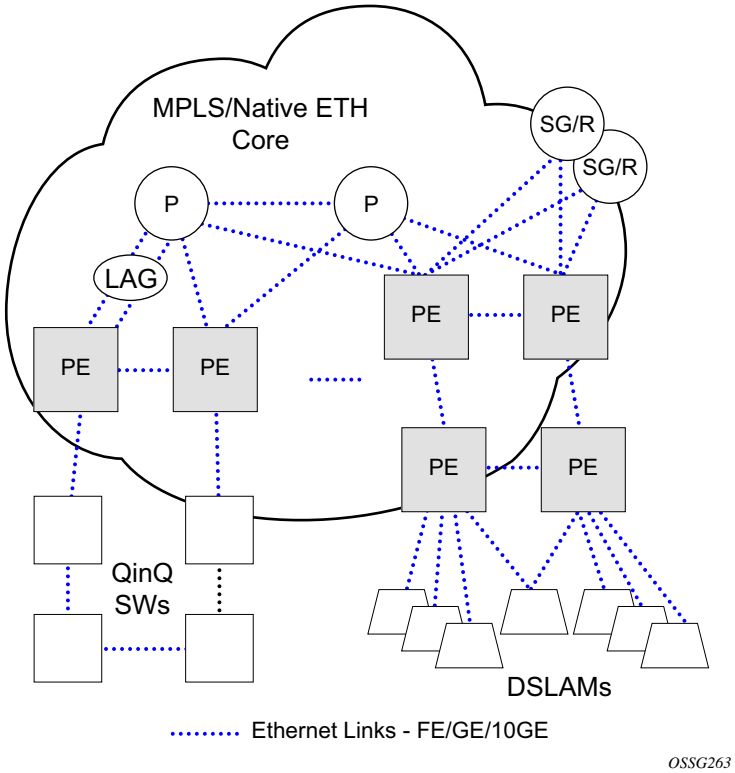

Figure 23: Generic Customer Use Case For LLDP

The example displayed in Figure 23 depicts a MPLS network that uses Ethernet interfaces in the core or as an access/handoff interfaces to connect to different kind of Ethernet enabled devices such as service gateway/routers, QinQ switches, DSLAMs or customer equipment.

IEEE 802.1ab LLDP running on each Ethernet interfaces in between all the above network elements may be used to discover the topology information.

Operators who are utilizing IOM3/IMM and above can tunnel the nearest-bridge at the port level using the **tunnel-nearest-bridge** command under the **config>port>ethernet>lldp>destmac** (nearest-bridge) hierarchy. The dest-mac nearest-bridge must be disabled for tunneling to occur.

LLDP Protocol Features

LLDP is an unidirectional protocol that uses the MAC layer to transmit specific information related to the capabilities and status of the local device. Separately from the transmit direction, the LLDP agent can also receive the same kind of information for a remote device which is stored in the related MIB(s).

LLDP itself does not contain a mechanism for soliciting specific information from other LLDP agents, nor does it provide a specific means of confirming the receipt of information. LLDP allows the transmitter and the receiver to be separately enabled, making it possible to configure an implementation so the local LLDP agent can either transmit only or receive only, or can transmit and receive LLDP information.

The information fields in each LLDP frame are contained in a LLDP Data Unit (LLDPDU) as a sequence of variable length information elements, that each include type, length, and value fields (known as TLVs), where:

- Type identifies what kind of information is being sent.
- Length indicates the length of the information string in octets.
- Value is the actual information that needs to be sent (for example, a binary bit map or an alphanumeric string that can contain one or more fields).

Each LLDPDU contains four mandatory TLVs and can contain optional TLVs as selected by network management:

- Chassis ID TLV
- Port ID TLV
- Time To Live TLV
- Zero or more optional TLVs, as allowed by the maximum size of the LLDPDU
- End Of LLDPDU TLV

The chassis ID and the port ID values are concatenated to form a logical identifier that is used by the recipient to identify the sending LLDP agent/port. Both the chassis ID and port ID values can be defined in a number of convenient forms. Once selected however, the chassis ID/port ID value combination remains the same as long as the particular port remains operable.

A non-zero value in the TTL field of the Time To Live TLV tells the receiving LLDP agent how long all information pertaining to this LLDPDU's identifier will be valid so that all the associated information can later be automatically discarded by the receiving LLDP agent if the sender fails to update it in a timely manner. A zero value indicates that any information pertaining to this LLDPDU's identifier is to be discarded immediately.

Note that a TTL value of zero can be used, for example, to signal that the sending port has initiated a port shutdown procedure. The End Of LLDPDU TLV marks the end of the LLDPDU.

The implementation defaults to setting the port-id field in the LLDP OAMPDU to **tx-local**. This encodes the port-id field as ifIndex (sub-type 7) of the associated port. This is required to support some releases of SAM. SAM may use the ifIndex value to properly build the Layer Two Topology Network Map. However, this numerical value is difficult to interpret or readily identify the LLDP peer when reading the CLI or MIB value without SAM. Including the **port-desc** option as part of the **tx-tlv** configuration allows an ALU remote peer supporting **port-desc** preferred display logic (11.0r1) to display the value in the port description TLV instead of the port-id field value. This does not change the encoding of the port-id field. That value continues to represent the ifIndex. In some environments, it may be important to select the specific port information that is carried in the port-id field. The operator has the ability to control the encoding of the port-id information and the associated subtype using the **port-id-subtype** option. Three options are supported for the port-id-subtype:

tx-if-alias — Transmit the ifAlias String (subtype 1) that describes the port as stored in the IF-MIB, either user configured description or the default entry (ie 10/100/Gig ethernet SFP)

tx-if-name — Transmits the ifName string (subtype 5) that describes the port as stored in the IF-MIB, ifName info.

tx-local — The interface ifIndex value (subtype 7)

IPv6 (address subtype 2) and IPv4 (address subtype 1) LLDP System Management addresses are supported.

LAG

Based on the IEEE 802.1ax standard (formerly 802.3ad), Link Aggregation Groups (LAGs) can be configured to increase the bandwidth available between two network devices, depending on the number of links installed. LAG also provides redundancy in the event that one or more links participating in the LAG fail. All physical links in a given LAG links combine to form one logical interface.

Packet sequencing must be maintained for any given session. The hashing algorithm deployed by Alcatel-Lucent routers is based on the type of traffic transported to ensure that all traffic in a flow remains in sequence while providing effective load sharing across the links in the LAG.

LAGs must be statically configured or formed dynamically with Link Aggregation Control Protocol (LACP). The optional marker protocol described in IEEE 802.1ax is not implemented. LAGs can be configured on network and access ports.

The LAG load sharing is executed in hardware, which provides line rate forwarding for all port types.

SR OS LAG implementation supports LAG that with all member ports of the same speed and LAG with mixed port-speed members (see later section for details).

SR OS LAG implementation is supported on access and network interfaces.

LACP

Under normal operation, all non-failing links in a given LAG will become active and traffic is load balanced across all active links. In some circumstances, however, this is not desirable. Instead, it desired that only some of the links are active (for example, all links on the same IOM) and the other links be kept in stand-by condition.

LACP enhancements allow active lag-member selection based on particular constrains. The mechanism is based on the IEEE 802.1ax standard so interoperability is ensured.

To use LACP on a given LAG, operator must enable LACP on the LAG including, if desired, selecting non-default LACP mode: active/passive and configuring administrative key to be used (**configure lag lacp**). IN addition an operator can configure desired LACP transmit interval (**configure lag lacp-xmit-interval**).

When LACP is enabled, an operator can see LACP changes through traps/log messages logged against the LAG. See TIMETRA-LAG-MIB.mib for more details.

LACP Multiplexing

The 7750 SR supports two modes of multiplexing RX/TX control for LACP: coupled and independent.

In coupled mode (default), both RX and TX are enabled or disabled at the same time whenever a port is added or removed from a LAG group.

In independent mode, RX is first enabled when a link state is UP. LACP sends an indication to the far-end that it is ready to receive traffic. Upon the reception of this indication, the far-end system can enable TX. Therefore, in independent RX/TX control, LACP adds a link into a LAG only when it detects that the other end is ready to receive traffic. This minimizes traffic loss that might occur in coupled mode if a port is added into a LAG before notifying the far-end system or before the far-end system is ready to receive traffic. Similarly, on link removals from LAG, LACP turns off the distributing and collecting bit and informs the far-end about the state change. This allows the far-end side to stop sending traffic as soon as possible.

Independent control provides for lossless operation for unicast traffic in most scenarios when adding new members to a LAG or when removing members from a LAG. It also reduces loss for multicast and broadcast traffic. When adding a port to LAG in a high scaled deployment, and that port is the first to be added to the LAG on that forwarding complex, it is recommended to first shut down the port, add the port to the LAG, and then re-enable the port after a short delay to allow for forwarding to be reprogrammed. This procedure minimizes outages.

Note that independent and coupled mode are interoperable (i.e. connected systems can have either mode set).

Active-Standby LAG Operation

Active/standby LAG is used to provide redundancy by logically dividing LAG into subgroups. The LAG is divided into subgroups by either assigning each LAG's ports to an explicit subgroup (1 by default), or by automatically grouping all LAG's ports residing on the same line card into a unique sub-group (auto-iom) or by automatically grouping all LAG's ports residing on the same MDA into a unique sub-group (auto-mda). When a LAG is divided into sub-groups, only a single sub-group is elected as active. Which sub-group is selected depends on selection criterion chosen.

The active/standby decision for LAG member links is a local decision driven by pre-configured selection-criteria. When LACP is configured, this decision was communicated to remote system using LACP signalling.

To allow non-LACP operation, an operator must disable LACP on a given LAG and select transmitter-driven standby signaling (configure lag standby-signaling power-off). As a consequence, the transmit laser will be switched off for all LAG members in standby mode. On switch over (active-links failed) the laser will be switched on all standby LAG members so they can become active.

When the power-off is selected as the standby-signaling, the selection-criteria **best-port** can be used.

It will not be possible to have an active LACP in power-off mode before the correct selection criteria is selected.

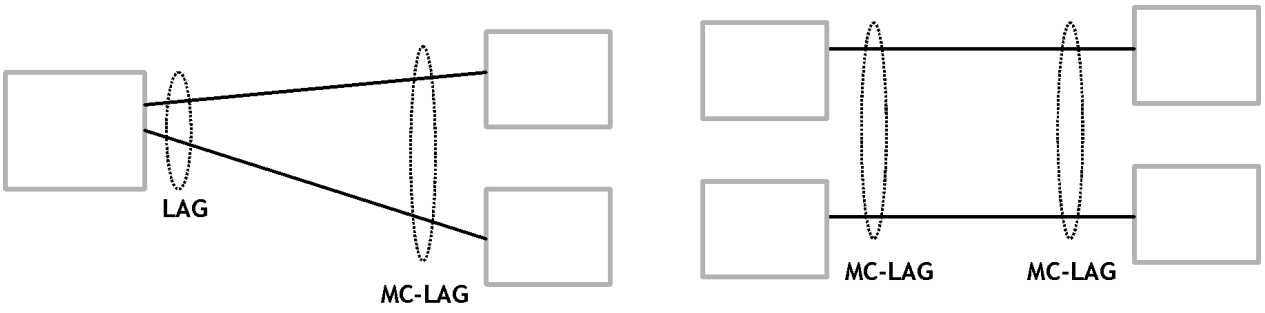

Figure 24: Active-Standby LAG Operation without Deployment Examples

Figure 24 depicts how LAG in Active/Standby mode can be deployed towards a DSLAM access using sub-groups with auto-iom sub-group selection. LAG links are divided into two sub-groups (one per line card).

In case of a link failure, Figure 25 and Figure 26, the switch over behavior ensures that all lag-members connected to the same IOM as failing link will become stand-by and lag-members

connected to other IOM will become active. This way, QoS enforcement constraints are respected, while the maximum of available links is utilized.

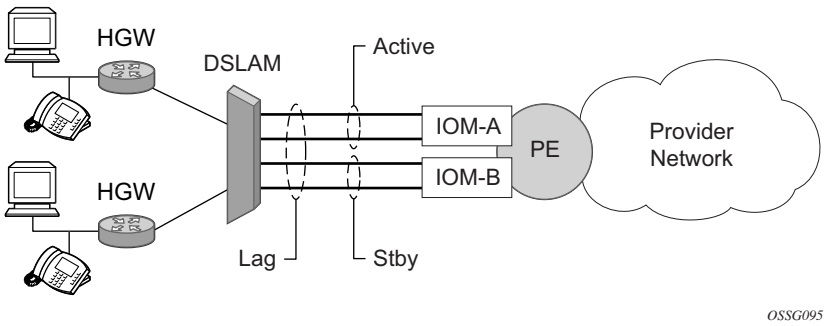

Figure 25: LAG on Access Interconnection

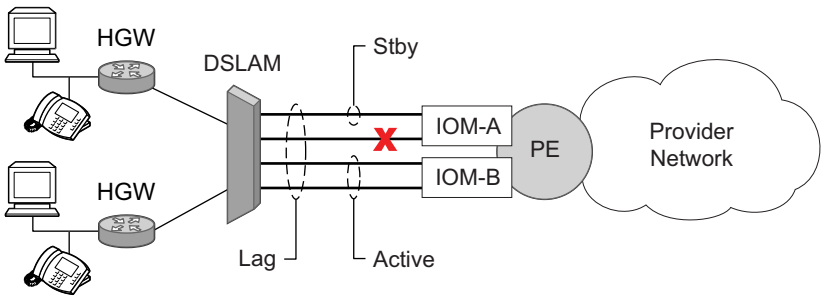

Figure 26: LAG on Access Failure Switchover

LAG on Access QoS Consideration

The following section describes various QoS related features applicable to LAG on access.

Adapt QoS Modes

Link Aggregation is supported on access side with access/hybrid ports. Similarly to LAG on network side, LAG on access is used to aggregate Ethernet ports into all active or active/standby LAG. The difference with LAG on networks lies in how the QoS/H-QoS is handled. Based on hashing configured, a given SAP's traffic can be sprayed on egress over multiple LAG ports or can always use a single port of a LAG. There are three user-selectable modes that allow operator to best adapt QoS configured to a LAG the SAPs are using:

1. adapt-qos distributed (default)

   In a distributed mode the SLA is divided among all line cards proportionally to the number of ports that exist on that line card for a given LAG. For example a 100Mbps PIR with 2 LAG links on IOM A and 3 LAG links on IOM B would result in IOM A getting 40 Mbps PIR and IOM B getting 60Mbps PIR. Thanks to such distribution, SLA can be enforced. The disadvantage is that a single flow is limited to IOM's share of the SLA. This mode of operation may also result in underrun due to a "hash error" (traffic not sprayed equally over each link). This mode is best suited for services that spray traffic over all links of a LAG.

2. adapt-qos link

   In a link mode the SLA is given to each and every port of a LAG. With the example above, each port would get 100 Mbps PIR. The advantage of this method is that a single flow can now achieve the full SLA. The disadvantage is that the overall SLA can be exceeded, if the flows span multiple ports. This mode is best suited for services that are guaranteed to hash to a single egress port.

3. adapt-qos port-fair

   Port-fair distributes the SLA across multiple line cards relative to the number of active LAG ports per card (in a similar way to distribute mode) with all LAG QoS objects parented to scheduler instances at the physical port level (in a similar way to link mode). This provides a fair distribution of bandwidth between cards and ports whilst ensuring that the port bandwidth is not exceeded. Optimal LAG utilization relies on an even hash spraying of traffic to maximize the use of the schedulers' and ports' bandwidth. With the example above, enabling port-fair would result in all five ports getting 20Mbps.

   When port-fair mode is enabled, per-Vport hashing is automatically disabled for subscriber traffic such that traffic sent to the Vport no longer uses the Vport as part of the hashing algorithm. Any QoS object for subscribers, and any QoS object for SAPs with explicitly configured hashing to a single egress LAG port, will be given the full bandwidth

configured for each object (in a similar way to link mode). A Vport used together with an egress port scheduler is supported with a LAG in port-fair mode, whereas it is not supported with a distribute mode LAG.

4. adapt-qos distributed include-egr-hash-cfg

   This mode can be considered a mix of link and distributed mode. The mode uses the configured hashing for LAG/SAP/service to choose either link or distributed adapt-qos modes. The mode allows:

   → SLA enforcement for SAPs that through configuration are guaranteed to hash to a single egress link using full QoS per port (as per link mode)

   → SLA enforcement for SAPs that hash to all LAG links proportional distribution of QoS SLA amongst the line cards (as per distributed mode)

   → SLA enforcement for multi service sites (MSS) that contain any SAPs regardless of their hash configuration using proportional distribution of QoS SLA amongst the line cards (as per distributed mode)

The following caveats apply to adapt-qos distributed include-egr-hash-cfg,

- The feature requires chassis mode D.
- LAG mode must be access or hybrid.
- The operator cannot change from **adapt-qos distribute include-egr-hash-cfg** to **adapt-qos distribute** when link-map-profiles or per-link-hash is configured.
- The operator cannot change from **adapt-qos link** to **adapt-qos distribute include-egr-hash-cfg** on a LAG with any configuration.
- Platforms supported except 7710 c12/c4, 7450 ESS-1

Table 24 shows examples of rate/BW distributions based on the **adapt-qos** mode used:

Table 24: Adapt QoS Bandwidth/Rate Distribution

	distribute	link	port-fair	distribute include-egr-hash-cfg
SAP Queues	% # local links[1]	100% rate	100% rate (SAP hash to one link) or %# all links[2] (SAP hash to all links)	100% rate (SAP hash to one link) or % # local links[a] (SAP hash to all links)

Table 24: Adapt QoS Bandwidth/Rate Distribution (Continued)

	distribute	link	port-fair	distribute include-egr-hash-cfg
SAP Scheduler	% # local links[a]	100% bandwidth	100% rate (SAP hash to one link) or %# all links[b] (SAP hash to all links)	100% bandwidth (SAP hash to a one link) or % # local links[a] (SAP hash to all links)
SAP MSS Scheduler	% # local links[a]	100% bandwidth	% # local links[a]	% # local links[a]

1.* % # local links = X * (number of local LAG members on a given line card/ total number of LAG members)
2.%# all links = X* (link speed)/(total LAG speed)

Per-fp-ing-queuing

Per-fp-ing-queuing optimization for LAG ports provides the ability to reduce the number of hardware queues assigned on each LAG SAP on ingress when the flag at LAG level is set for per-fp-ing-queuing.

When the feature is enabled in the **config>lag>access** context, the queue allocation for SAPs on a LAG will be optimized and only one queuing set per ingress forwarding path (FP) is allocated instead of one per port.

The following rules will apply for configuring the per-fp-ing-queuing at LAG level:

- To enable per-fp-ing-queuing, the LAG must be in access mode
- The LAG mode cannot be set to network mode when the feature is enabled
- Per-fp-ing-queuing can only be set if no port members exists in the LAG
- Per-fp-ing-queuing cannot be set if LAG's port-type is hsmda.

Per-fp-egr-queuing

Per-fp-egr-queuing optimization for LAG ports provides the ability to reduce the number of egress resources consumed by each SAP on a LAG, and by any encap groups that exist on those SAPs.

When the feature is enabled in the **config>lag>access** context, the queue and virtual scheduler allocation will be optimized. Only one queuing set and one H-QoS virtual scheduler tree per SAP/encap group will be allocated per egress forwarding path (FP) instead of one set per each port of the LAG. In case of a link failure/recovery, egress traffic uses failover queues while the queues are moved over to a newly active link.

Per-fp-egr-queuing can be enabled on existing LAG with services as long as the following conditions are met.

- The LAG's mode must be **access** or **hybrid.**
- The LAG's port-type must be **standard.**
- The LAG must have either **per-link-hash** enabled or all SAPs on the LAG must use **per-service-hashing** only and be of a type: VPLS SAP, i-VPLS SAP, or e-Pipe VLL or PBB SAP.
- The system must be, at minimum, in chassis mode **d** (**configure>system>chassis-mode**)

To disable per-fp-egr-queuing, all ports must first be removed from a given LAG.

Per-fp-sap-instance

Per-fp-sap-instance optimization for LAG ports provides the ability to reduce the number of SAP instance resources consumed by each SAP on a lag.

When the feature is enabled, in the config>lag>access context, a single SAP instance is allocated on ingress and on egress per each forwarding path instead of one per port. Thanks to an optimized resource allocation, the SAP scale on a line card will increase, if a LAG has more than one port on that line card. Because SAP instances are only allocated per forwarding path complex, h/w reprogramming must take place when as result of LAG links going down or up, a SAP is moved from one LAG port on a given line card to another port on a given line card within the same forwarding complex. This results in an increased data outage when compared to per-fp-sap-instance feature being disabled. During the reprogramming, failover queues are used when SAP queues are reprogrammed to a new port. Any traffic using failover queues will not be accounted for in SAPs statistics and will be processed at best-effort priority.

The following rules apply when configuring per-fp-sap-instance on a given LAG:

- Minimum chassis mode D is required.
- Per-fp-sap-ingress-queuing and per-fp-sap-egr-queuing must be enabled.
- The functionality can be enabled/disabled on LAG with no member ports only. Services can be configured.

Other caveats:

- SAP instance optimization applies to LAG-level. Whether a LAG is sub-divided into sub-groups or not, the resources are allocated per forwarding path for all complexes LAG's links are configured on (i.e. irrespective of whether a given sub-group a SAP is configured on uses that complex or not).
- Egress statistics continue to be returned per port when SAP instance optimization is enabled. If a LAG links are on a single forwarding complex, all ports but one will have no change in statistics for the last interval – unless a SAP moved between ports during the interval.
- Rollback that changes per-fp-sap-instance configuration is service impacting.

LAG and ECMP Hashing

When a requirement exists to increase the available bandwidth for a logical link that exceeds the physical bandwidth or add redundancy for a physical link, typically one of two methods is applied: equal cost multi-path (ECMP) or Link Aggregation (LAG). A system can deploy both at the same time using ECMP of two or more Link Aggregation Groups (LAG) and/or single links.

Different types of hashing algorithms can be employed to achieve one of the following objectives:

- ECMP and LAG load balancing should be influenced solely by the offered flow packet. This is referred to as *per-flow* hashing.
- ECMP and LAG load balancing should maintain consistent forwarding within a given service. This is achieved using *consistent per-service* hashing.
- LAG load balancing should maintain consistent forwarding on egress over a single LAG port for a specific network interface, SAP, etc. This is referred as *per link* hashing (including explicit per link hashing with LAG link map profiles). Note that if multiple ECMP paths use a LAG with per link hashing, the ECMP load balancing is done using either *per flow* or *consistent per service* hashing.

These hashing methods are described in the following subsections. Although multiple hashing options may be configured for a given flow at the same time, only one method will be selected to hash the traffic based on the following decreasing priority order:

For ECMP load balancing:

1. Consistent per service hashing
2. Per flow hashing

For LAG load balancing:

1. LAG link map profile
2. Per link hash
3. Consistent per service hashing
4. Per flow hashing

Per Flow Hashing

Per flow hashing uses information in a packet as an input to the hash function ensuring that any given flow maps to the same egress LAG port/ECMP path. Note that because the hash uses information in the packet, traffic for the same SAP/interface may be sprayed across different ports of a LAG or different ECMP paths. If this is not desired, other hashing methods outlined in this section can be used to change that behavior. Depending on the type of traffic that needs to be distributed into an ECMP and/or LAG, different variables are used as input to the hashing algorithm that determines the next hop selection. The following outlines default per flow hashing behavior for those different types of traffic:

- VPLS known unicast traffic is hashed based on the IP source and destination addresses for IP traffic, or the MAC source and destination addresses for non-IP traffic. The MAC SA/DA are hashed and then, if the Ethertype is IPv4 or IPv6, the hash is replaced with one based on the IP source address/destination address.
- VPLS multicast, broadcast and unknown unicast traffic.
  - → Traffic transmitted on SAPs is not sprayed on a per-frame basis, but instead the service ID is used to pick ECMP and LAG paths statically.
  - → Traffic transmitted on SDPs is hashed on a per packet basis in the same way as VPLS unicast traffic. However, per packet hashing is applicable only to the distribution of traffic over LAG ports, as the ECMP path is still chosen statically based on the service ID.

    Data is hashed twice to get the ECMP path. If LAG and ECMP are performed on the same frame, the data will be hashed again to get the LAG port (three hashes for LAG). However, if only LAG is performed, then hashing will only be performed twice to get the LAG port.
  - → Multicast traffic transmitted on SAPs with IGMP snooping enabled is load-balanced based on the internal multicast ID, which is unique for every (s,g) record. This way, multicast traffic pertaining to different streams is distributed across different LAG member ports.
  - → The hashing procedure that used to be applied for all VPLS BUM traffic would result in PBB BUM traffic being sent out on BVPLS SAP to follow only a single link when MMRP was not used. Therefore, in chassis mode D, traffic flooded out on egress BVPLS SAPs is now load spread using the algorithm described above for VPLS known unicast.
- Unicast IP traffic routed by a router is hashed using the IP SA/DA in the packet.
- MPLS packet hashing at an LSR is based on the whole label stack, along with the incoming port and system IP address. Note that the EXP/TTL information in each label is not included in the hash algorithm. This method is referred to as *Label-Only Hash* option and is enabled by default, or can be re-instated in CLI by entering the lbl-only keyword. A couple of options to further hash on the header of an IP packet in the payload of the MPLS packet are also provided.

- VLL traffic from a service access point is not sprayed on a per-packet basis, but as for VPLS flooded traffic, the service ID is used to pick one of the ECMP/LAG paths. The exception to this is when shared-queuing is configured on an e-pipe SAP, i-pipe SAP, or f-pipe SAP, or when H-POL is configured on an e-pipe SAP. In those cases, traffic spraying is the same as for VPLS known unicast traffic. Packets of the above VLL services received on a spoke-SDP are sprayed the same as for VPLS known unicast traffic.
- Note that a-pipe and c-pipe VLL packets are always sprayed based on the service-id in both directions.
- Multicast IP traffic is hashed based on an internal multicast ID, which is unique for every record similar to VPLS multicast traffic with IGMP snooping enabled.

In addition to the above outlined per-flow hashing inputs SROS supports multiple option to modify default hash inputs.

For all cases that involve per-packet hashing, the NPA produces a 20-bit result based on hashing the relevant packet data. This result is input to a modulo like calculation (divide by the number of routes in the ECMP and use the remainder) to determine the ECMP index.

If the ECMP index results in the selection of a LAG as the next hop, then the hash result is hashed again and the result of the second hash is input to the modulo like operation (divide by the number of ports in the LAG and use the remainder) to determine the LAG port selection.

Note however that when the ECMP set includes an IP interface configured on a spoke-SDP (IES/VPRN spoke interface), or a Routed VPLS interface, the unicast IP packets—which will be sprayed over this interface—will not be further sprayed over multiple RSVP LSPs (part of the same SDP), or multiple LDP FEC next-hops when available. In this case, a single RSVP LSP or LDP FEC next-hop will be selected based on a modulo operation of the service ID. The second round of the hash is exclusively used for LAG link selection. IP unicast packets from different IES/VPRN services or Routed VPLS services will be distributed across RSVP LSPs or LDP FEC next-hops based on the modulo operation of their respective service ID.

Changing Default Per Flow Hashing Inputs

For some traffic patterns or specific deployments, per-flow hashing is desired but the hashing result using default hash inputs as outlined above may not be produce a desired distribution. To alleviate this issue, SROS allows operators to modify default hash inputs as outlined in the following subsections.

LSR Hashing

The LSR hash routine operates on the label stack only. However, there is also the ability to hash on the IP header if a packet is IP. An LSR will consider a packet to be IP if the first nibble following the bottom of the label stack is either 4 (IPv4) or 6 (IPv6). This allows the user to include an IP

header in the hashing routine at an LSR for the purpose of spraying labeled IP packets over multiple equal cost paths in ECMP in an LDP LSP and/or over multiple links of a LAG group in all types of LSPs.

The user enables the LSR hashing on label stack and/or IP header by entering the following system-wide command: **config>system>load-balancing>lsr-load-balancing** [**lbl-only** | **lbl-ip** | **ip-only**]

By default, the 7x50 LSR falls back to the hashing on label stack only. This option is referred to as lbl-only and the user can revert to this behavior by entering one of the two commands:

config>system>load-balancing>lsr-load-balancing lbl-only

config>system>load-balancing>no lsr-load-balancing

The user can also selectively enable or disable the inclusion of label stack and IP header in the LSR hash routine on a specific network interface by entering the following command:

config>router>interface>load-balancing>lsr-load-balancing [lbl-only | lbl-ip | ip-only]

This provides some control to the user such that this feature is disabled if labeled packets received on a specific interface include non IP packets that can be confused by the hash routine for IP packets. These could be VLL and VPLS packets without a PW control word.

When the user performs the **no** form of this command on an interface, the interface inherits the system level configuration.

The default **lbl-only** hash option and the label-ip option with IPv4 payload is supported on all platforms and chassis modes. The **ip-only** option with both IPv4 and IPv6 payloads as well as the lbl-ip option with IPv6 payload are only supported on IP interfaces on IOM3/IMM ports.

LSR Default Hash Routine—Label-Only Hash Option

The following is the behavior of ECMP and LAG hashing at an LSR in the existing implementation. These are performed in two rounds.

First the ECMP hash. It consists of an initial hash based on the source port/system IP address. Each label in the stack is then hashed separately with the result of the previous hash, up to a maximum of five labels. The net result will be used to select which LDP FEC next-hop to send the packet to using a modulo operation of the net result with the number of next-hops. If there is a single next-hop for the LDP FEC, or if the packet is received on an RSVP LSP ILM, then a single next-hop exists.

This same net result will feed to a second round of hashing if there is LAG on the egress port where the selected LDP or RSVP LSP has its NHLFE programmed.

LSR Label-IP Hash Option Enabled

In the first hash round for ECMP, the algorithm will parse down the label stack and once it hits the bottom it checks the next nibble. If the nibble value is 4 then it will assume it is an IPv4 packet. If the nibble value is 6 then it will assume it is an IPv6 packet. In both cases, the result of the label hash is fed into another hash along with source and destination address fields in the IP packet header. Otherwise, it will just use the label stack hash already calculated for the ECMP path selection.

If there are more than five labels in the stack, then the algorithm will also use the result of the label hash for the ECMP path selection.

The second round of hashing for LAG re-uses the net result of the first round of hashing. This means IPv6 packets will continue to be hashed on label stack only.

LSR IP-Only Hash Option Enabled

This option behaves like the label-IP hash option except that when the algorithm reached the bottom of the label stack in the ECMP round and finds an IP packet, it throws the outcome of the label hash and only uses the source and destination address fields in the IP packet's header.

LSR Ethernet Encapsulated IP Hash only Option Enabled

This option behaves like LSR IP only hash except for how the IP SA/DA information is found. The following conditions are verified to find IP SA/DA for hash.

- Label stack must not exceed 3 labels deep
- After the bottom of the stack is reached, the hash algorithm verifies that what follows is Ethernet II untagged frame (by looking at the value of ethertype at the expected packet location whether it contains Ethernet encapsulated IPv4 (0x0800) or IPv6 (0x86DD) value.

  When the ethertype verification passes, the first nibble of the expected IP packet location is then verified to be 4 (IPv4) or 6 (IPv6).

L4 Load Balancing

Operator may enable L4 load balancing to include TCP/UDP source/destination port numbers in addition to source/destination IP addresses in per flow hashing of IP packets. By including the L4 information, a SA/DA default hash flow can be sub-divided into multiple finer-granularity flows if the ports used between a given SA/DA vary.

L4 load balancing can be enabled/disabled on system and interface levels. When enabled, the extra L4 port inputs apply to per-flow hashing for unicast IP traffic and multicast traffic (if **mc-enh-load-balancing** is enabled).

System IP Load Balancing

This enhancement adds an option to add the system IP address into the hash algorithm. This adds a per system variable so that traffic being forward through multiple routers with similar ECMP paths will have a lower chance of always using the same path to a given destination.

Currently, if multiple routers have the same set of ECMP next hops, traffic will use the same nexthop at every router hop. This can contribute to the unbalanced utilization of links. The new hash option avoids this issue.

This feature when enabled, enhances the default per-flow hashing algorithm described earlier. It however does not apply to services which packets are hashed based on service-id or when per service consistent hashing is enabled. This hash algorithm is only supported on IOM3-XPs/IMMs or later generations of hardware.The System IP load balancing can be enabled per-system only.

TEID Hash for GTP-Encapsulated Traffic

This options enables TEID hashing on L3 interfaces. The hash algorithm identifies GTP-C or GTP-U by looking at the UDP destination port (2123 or 2152) of an IP packet to be hashed. If the value of the port matches, the packet is assumed to be GTP-U/C. For GTPv1 packets TEID value from the expected header location is then included in hash. For GTPv2 packets the TEID flag value in the expected header is additionally checked to verify whether TEID is present. If TEID is present, it is included in hash algorithm inputs. TEID is used in addition to GTP tunnel IP hash inputs: SA/DA and SPort/DPort (if L4 load balancing is enabled). If a non-GTP packet is received on the GTP UDP ports above, the packets will be hashed as GTP**.**

Source-Only/Destination-Only Hash Inputs

This option allows an operator to only include source parameters or only include destination parameters in the hash for inputs that have source/destination context (such as IP address and L4 port). Parameters that do not have source/destination context (such as TEID or System IP for example) are also included in hash as per applicable hash configuration. The functionality allows, among others, to ensure that both upstream and downstream traffic hash to the same ECMP path/LAG port on system egress when traffic is sent to a hair-pinned appliance (by configuring source-only hash for incoming traffic on upstream interfaces and destination-only hash for incoming traffic on downstream interfaces).

Enhanced Multicast Load Balancing

Enhanced multicast load balancing allows operators to replace the default multicast per flow hash input (internal multicast ID) with information from the packet. When enabled, multicast traffic for Layer 3 services (such as IES, VPRN, r-VPLS) and ng-MVPN (multicast inside RSVP-TE, LDP LSPs) are hashed using information from the packet. Which inputs are chosen depends on which per flow hash inputs options are enabled based on the following:

- IP replication—The hash algorithm for multicast mimics unicast hash algorithm using SA/DA by default and optionally TCP/UDP ports (Layer 4 load balancing enabled) and/or system IP (System IP load balancing enabled) and/or source/destination parameters only (Source-only/Destination-only hash inputs).
- MPLS replication—The hash algorithm for multicast mimics unicast hash algorithm described in LSR Hashing on page 120.

NOTE: Enhanced multicast load balancing requires minimum chassis mode D. It is not supported with Layer 2 and ESM services. It is supported on 7750 SR (excluding c4/c12) platforms.

Security Parameter Index (SPI) Load Balancing

IPSec tunnelled traffic transported over LAG typically falls back to IP header hashing only. For example, in LTE deployments, TEID hashing cannot be performed because of encryption, and the system performs IP-only tunnel-level hashing. Because each SPI in the IPSec header identifies a unique SA, and thus flow, these flows can be hashed individually without impacting packet ordering. In this way, SPI load balancing provides a mechanism to improve the hashing performance of IPSec encrypted traffic.

SR OS allows enabling SPI hashing per L3 interface (this is the incoming interface for hash on system egress)/L2 VPLS service. When enabled, an SPI value from ESP/AH header is used in addition to any other IP hash input based on per-flow hash configuration: source/destination IPv6 addresses, L4 source/dest ports in case NAT traversal is required (l4-load-balancing is enabled). If the ESP/AH header is not present in a packet received on a given interface, the SPI will not be part of the hash inputs, and the packet is hashed as per other hashing configurations. SPI hashing is not used for fragmented traffic to ensure first and subsequent fragments use the same hash inputs.

SPI hashing is supported for IPv4 and IPv6 tunnel unicast traffic and for multicast traffic (mc-enh-load-balancing must be enabled) on all platforms and requires L3 interfaces or VPLS service interfaces with SPI hashing enabled to reside on IOM3-XP or newer line-cards.

Per Link Hashing

The hashing feature described in this section applies to traffic going over LAG and MC-LAG. Per link hashing ensures all data traffic on a given SAP or network interface uses a single LAG port on egress. Because all traffic for a given SAP/network interface egresses over a single port, QoS SLA enforcement for that SAP, network interface is no longer impacted by the property of LAG (distributing traffic over multiple links). Internally-generated, unique IDs are used to distribute SAPs/network interface over all active LAG ports. As ports go UP and DOWN, each SAP and network interface is automatically rehashed so all active LAG ports are always used.

The feature is best suited for deployments when SAPs/network interfaces on a given LAG have statistically similar BW requirements (since per SAP/network interface hash is used). If more control is required over which LAG ports SAPs/network interfaces egress on, a LAG link map profile feature described later in this guide may be used.

Per link hashing, can be enabled on a LAG as long as the following conditions are met:

- LAG **port-type** must be *standard*.
- LAG **access adapt-qos** must be *link* or *port-fair* (for LAGs in **mode** access or hybrid).
- System must be at minimum in chassis mode *d* (configure system chassis-mode)
- LAG mode is access/hybrid and the **access adapt-qos** mode is distribute **include-egr-hash-cfg**

Weighted per-link-hash

Weighted per-link-hash allows higher control in distribution of SAPs/interfaces/subscribers across LAG links when significant differences in SAPs/interfaces/subscribers bandwidth requirements could lead to an unbalanced distribution bandwidth utilization over LAG egress. The feature allows operators to configure for each SAPs/interfaces/subscribers on a LAG one of three 3 unique classes and a weight value to be used to when hashing this service/subscriber across the LAG links. SAPs/interfaces/subscribers are hashed to LAG links, such that within each class the total weight of all SAPs/interfaces/subscribers on each LAG link is as close as possible to each other.

Multiple classes allow grouping of SAPs/interfaces/subscribers by similar bandwidth class/type. For example a class can represent: voice – negligible bandwidth, Broadband – 10-100Mbps, Extreme Broadband – 300Mbps and above types of service. If a class and weight are not specified for a given service or subscriber, values of 1 and 1 are used respectively.

The following algorithm is used to hash SAPs/interfaces/subscribers to LAG egress links:

- TPSDA subscribers are hashed to a LAG link when subscribers are active, MSE SAPs/interfaces are hashed to a LAG link when configured

- For a new SAP/interface/subscriber to be hashed to an egress LAG link:
- Select active link with the smallest current weight for the SAP/network/subscriber class (lowest link id tie-breaker)
- On a LAG link failure:
  - → Only SAPs/interfaces/subscribers s on a failed link are rehashed over the remaining active links
  - → Processing order: Per class from lowest numerical, within each class per weight from highest numerical value
- LAG link recovery/new link added to a LAG
- auto-rebalance disabled: Existing SAPs/interfaces/subscribers remain on the currently active links, new SAPs/interfaces/subscribers naturally prefer the new link until balance reached.
- auto-rebalance is enabled: When a new port is added to a LAG a non-configurable 5 second rebalance timer is started. Upon timer expiry, all existing SAPs/interfaces/subscribers are rebalanced across all active LAG links minimizing the number of SAPs/interfaces/subscribers moved to achieve rebalance. The rebalance timer is restarted if a new link is added while the timer is running. If a port bounces 5 times within a 5 second interval, the port is quarantined for10 seconds. This behavior is not configurable.
- On a LAG start-up, the rebalance timer is always started irrespective of auto-rebalance configuration to avoid hashing SAPs/interfaces/subscribers to a LAG before ports have a chance to come UP.

Optionally an operator can use, a “tools perform lag load-balance” command to manually re-balance ALL weighted per-link-hashed SAPs/interfaces/subscribers on a LAG. The rebalance follows the algorithm as used on a link failure moving SAPs/interfaces/subscribers to different LAG links to minimize SAPs/interfaces/subscribers impacted.

An optional time-delay for off-peak rebalance can be specified. If LAG is moved from weighted per-link-hash while the load-balance is being time delayed, the time delay will be canceled and no rebalancing will happen. If LAG or its links change operational, administrative status, the time delay will not be impacted and will execute once the delay timer expires.

The following caveats exist:

- When weighted per-link-hash is deployed on a given LAG, no other methods of hash for subscribers/SAPs/interfaces on that LAG (like service hash or LAG link map profile) should be deployed, since the weighted hash is not able to account for load placed on LAG links by subscriber/SAPs/interfaces using the other hash methods.
- Weighted per-link-hash is not supported with mixed-speed LAGs and for network interfaces.
- For TPSDA model:
  - → only 1:1 (subscriber to SAP) model is supported and weight/class should not be enabled on a SAP.

The feature will not operate properly if those conditions are not met.

Explicit Per Link Hash Using LAG Link Mapping Profiles

The hashing feature described in this section applies to traffic going over LAG and MC-LAG. LAG link mapping profile feature gives operators full control of which links SAPs/network interface use on a LAG egress and how the traffic is rehashed on a LAG link failure. Some benefits that such functionality provides include:

- Ability to perform management level admission control onto LAG ports thus increasing overall LAG BW utilization and controlling LAG behavior on a port failure.
- Ability to strictly enforce QoS contract on egress for a SAP/network interface or a group of SAPs/network interfaces by forcing it/them to egress over a single port and using **access adapt-qos** link or port-fair mode.

To enable LAG Link Mapping Profile Feature on a given LAG, operators configure one or more of the available LAG link mapping profiles on the LAG and then assign that profile(s) to all or a subset of SAPs and network interfaces as needed. Enabling per LAG link Mapping Profile is allowed on a LAG with services configured, a small outage may take place as result of re-hashing SAP/network interface when a lag profile is assigned to it.

Each LAG link mapping profile allows operators to configure:

- Primary link—defines a port of the LAG to be used by a SAP/network interface when the port is UP. Note that a port cannot be removed from a LAG if it is part of any LAG link profile.
- Secondary link—defines a port of the LAG to be used by a SAP/network interface as a backup when the primary link is not available (not configured or down) and the secondary link is UP.
- Mode of operation when neither primary, nor secondary links are available (not configured or down):
  - **discard** – traffic for a given SAP/network interface will be dropped to protect other SAPs/network interfaces from being impacted by re-hashing these SAPs/network interfaces over remaining active LAG ports.

    Note: SAP/network interface status will not be affected when primary and secondary links are unavailable, unless an OAM mechanism that follows the data path hashing on egress is used and will cause a SAP/network interface to go down
  - **per-link-hash** – traffic for a given SAP/network interface will be re-hashed over remaining active ports of a LAG links using per-link-hashing algorithm. This behavior ensures SAP/network interfaces using this profile will be given available resources of other active LAG ports even if that means impacting other SAP/network interfaces on the LAG. The system will use the QoS configuration to provide fairness and priority if congestion is caused by the default-hash recovery.

LAG link mapping profiles, can be enabled on a LAG as long as the following conditions are met:

- LAG **port-type** must be *standard*.
- LAG **access adapt-qos** must be *link* or *port-fair* (for LAGs in **mode** access or hybrid)
- All ports of a LAG on a given router must belong to a single sub-group.
- System must be at minimum in chassis mode ***d*** (**configure system chassis-mode**)
- Access adapt-qos mode is distribute include-egr-hash-cfg.

LAG link mapping profile can co-exist with any-other hashing used over a given LAG (for example, per flow hashing or per-link-hashing). SAPs/network interfaces that have no link mapping profile configured will be subject to LAG hashing, while SAPs/network interfaces that have configured LAG profile assigned will be subject to LAG link mapping behavior, which is described above.

Consistent Per Service Hashing

The hashing feature described in this section applies to traffic going over LAG, Ethernet tunnels (eth-tunnel) in loadsharing mode, or CCAG load balancing for VSM redundancy. The feature does not apply to ECMP.

Per-service-hashing was introduced to ensure consistent forwarding of packets belonging to one service. The feature can be enabled using the [**no**] **per-service-hashing** configuration option under **config>service>epipe** and **config>service>vpls**, valid for Epipe, VPLS, PBB Epipe, IVPLS and BVPLS. Chassis mode D is required.

The following behavior applies to the usage of the [no] per-service-hashing option.

- The setting of the PBB Epipe/I-VPLS children dictates the hashing behavior of the traffic destined to or sourced from an Epipe/I-VPLS endpoint (PW/SAP).
- The setting of the B-VPLS parent dictates the hashing behavior only for transit traffic through the B-VPLS instance (not destined to or sourced from a local I-VPLS/Epipe children).

The following algorithm describes the hash-key used for hashing when the new option is enabled:

- If the packet is PBB encapsulated (contains an I-TAG ethertype) at the ingress side and enters a B-VPLS service, use the ISID value from the I-TAG. For PBB encapsulated traffic entering other service types, use the related service ID
- If the packet is not PBB encapsulated at the ingress side
  - → For regular (non-PBB) VPLS and EPIPE services, use the related service ID
  - → If the packet is originated from an ingress IVPLS or PBB Epipe SAP
    - – If there is an ISID configured use the related ISID value

  - If there is no ISID yet configured use the related service ID
- → For BVPLS transit traffic use the related flood list id
  - Transit traffic is the traffic going between BVPLS endpoints
  - An example of non-PBB transit traffic in BVPLS is the OAM traffic
- The above rules apply regardless of traffic type
  - → Unicast, BUM flooded without MMRP or with MMRP, IGMP snooped

Operators may sometimes require the capability to query the system for the link in a LAG or Ethernet tunnel that is currently assigned to a given service-id or ISID. This ability is provided using the **tools>dump>map-to-phy-port** {**ccag** *ccag-id* | **lag** *lag-id* | **eth-tunnel** *tunnel-index*} {**isid** *isid* [**end-isid** *isid*] | **service** *servid-id* | *svc-name* [**end-service** *service-id* | *syc-name*]} [**summary**] command.

A sample usage is as follows:

```
A:Dut-B# tools dump map-to-phy-port lag 11 service 1

ServiceId  ServiceName   ServiceType     Hashing                  Physical Link
---------- ------------- --------------  -----------------------  -------------
1                        i-vpls          per-service(if enabled)  3/2/8

A:Dut-B# tools dump map-to-phy-port lag 11 isid 1

ISID     Hashing                  Physical Link
-------- -----------------------  -------------
1        per-service(if enabled)  3/2/8

A:Dut-B# tools dump map-to-phy-port lag 11 isid 1 end-isid 4
ISID     Hashing                  Physical Link
-------- -----------------------  -------------
1        per-service(if enabled)  3/2/8
2        per-service(if enabled)  3/2/7
3        per-service(if enabled)  1/2/2
4        per-service(if enabled)  1/2/3
```

ESM – LAG Hashing per Vport

Background

Vport is a 7x50 BNG representation of a remote traffic aggregation point in the access network. It is a level in the hierarchical QoS model implemented within the 7x50 BNG that requires QoS treatment.

When 7x50 BNG is connected to access network via LAG, a VPort construct within the BNG is instantiated per member link on that LAG. Each instance of the Vport in such a configuration receives the entire amount of configured bandwidth. When traffic is sprayed in a per-subscriber

fashion over member links in an LAG without awareness of the Vport, it can lead to packet drops on one member link irrespective of the relative traffic priority on another LAG member link in the same Vport. The reason is that multiple Vport instances of the same Vport on different LAG member links are not aware of each other.

With a small number of subscribers per Vport and a great variation in bandwidth service offering per subscriber (from mbps to gbps), there is a great chance that the load distribution between the member links will be heavily unbalanced. For example, if the lag consists of two member links on the same IOM, three 1Gbps high priority subscribers can saturate the 2Gbps Vport bandwidth on one member link of the LAG. And all the while, twenty low priority 10Mbps subscribers that are using the other link are significantly under-utilizing available bandwidth on the corresponding Vport.

To remedy this situation, all traffic flowing through the same Vport must be hashed to a single LAG member link. This way, the traffic treatment will be controlled by a single Vport instance, and achieve a desired behavior where low priority 10Mbps subscribers traffic will be affected before any traffic from the high priority subscribers.

Hashing per Vport

Hashing traffic per Vport ensures that the traffic on the same PON (or DSLAM) traverse the same Vport, and therefore, it is the same member link that this Vport is associated with. The Vport instances of the same Vport on another member links are irrelevant for QoS treatment.

The Vport in 7x50 is referenced via inter-dest-string, which can be returned via RADIUS. For this reason, the terms hashing per inter-dest-string or hashing per Vport can be interchangeably used.

If the subscriber is associated with a Vport, hashing will be automatically performed per inter-dest-string. In case that no such association exists, hashing will default to per-subscriber hashing.

In certain cases, S-vlan tag can represent Vport. In such a case, per S-vlan hashing is desired. This can be implicitly achieved by the following configuration:

```
configure
  subscr-mgmt
    msap-policy <name>
      sub-sla-mgmt
       def-inter-dest-id use-top-queue

configure
  port <port-id>
    ethernet
      access
        egress
         vport <name>
           host-match dest <s-tag>
```

Through this CLI hierarchy, S-tag is implicitly associated with the inter-dest-string and consequently with the Vport.

Link Placement

This feature requires that all active member ports in a LAG reside on the same forwarding complex (IOM/IMM).

Multicast Consideration

Multicast traffic that is directly replicated per subscriber follows the same hashing algorithm as the rests of the subscribers (per inter-dest-string hashing).

Multicast traffic that is redirected to a regular Layer 3 interface outside of the ESM will be hashed per destination group (or IP address).

VPLS and Capture SAP Considerations

VPLS environment in conjunction with ESM allows hashing based on destination mac address. This is achieved through the following CLI hierarchy:

```
configure
  service vpls <vpls-id>
    sap lag-<id>
     sub-sla-mgmt
       mac-da-hashing
```

Note: This is only applicable to L2 ESM. In the case where this is configured AND Vport hashing is desired, the following order of evaluation will be executed:

1. Hashing based on subscriber-id or inter-dest-string
2. If configured, mac-da-hashing

Hashing per inter-dest-string will win if <Vport, subscriber> association is available at the same time as the mac-da-hashing is configured.

Mac-da-hashing mechanism cannot transition from capture SAP to a derived MSAP.

LSR Default Hash Routine— Label-Only Hash Option

The following is the behavior of ECMP and LAG hashing at an LSR in the existing implementation. These are performed in two rounds.

First the ECMP hash. It consists of an initial hash based on the source port/system IP address. Each label in the stack is then hashed separately with the result of the previous hash, up to a maximum of five labels. The net result will be used to select which LDP FEC next-hop to send the packet to using a modulo operation of the net result with the number of next-hops. If there is a single next-hop for the LDP FEC, or if the packet is received on an RSVP LSP ILM, then a single next-hop exists.

This same net result will feed to a second round of hashing if there is LAG on the egress port where the selected LDP or RSVP LSP has its NHLFE programmed.

LSR Label-IP Hash Option Enabled

In the first hash round for ECMP, the algorithm will parse down the label stack and once it hits the bottom it checks the next nibble. If the nibble value is 4 then it will assume it is an IPv4 packet. If the nibble value is 6 then it will assume it is an IPv6 packet. In both cases, the result of the label hash is fed into another hash along with source and destination address fields in the IP packet's header. Otherwise, it will just use the label stack hash already calculated for the ECMP path selection.

If there are more than five labels in the stack, then the algorithm will also use the result of the label hash for the ECMP path selection.

The second round of hashing for LAG re-uses the net result of the first round of hashing. This means IPv6 packets will continue to be hashed on label stack only.

LSR IP-Only Hash Option Enabled

This option behaves like the label-IP hash option except that when the algorithm reached the bottom of the label stack in the ECMP round and finds an IP packet, it throws the outcome of the label hash and only uses the source and destination address fields in the IP packet's header.

LAG Hold Down Timers

Operators can configure multiple hold down timers that allow control how quickly LAG responds to operational port state changes. The following timers are supported:

1. Port-level hold-time up/down timer
   This optional timer allows operator to control delay for adding/removing a port from LAG when the port comes UP/goes DOWN. Each LAG port runs the same value of the timer, configured on the primary LAG link. See Port Link Dampening description in Port Features section of this guide for more details on this timer.
2. Sub-group-level hold-time timer
   This optional timer allows operator to control delay for a switch to a new candidate sub-group selected by LAG sub-group selection algorithm from the current, operationally UP sub-group. The timer can also be configured to never expire, which prevents a switch from operationally up sub-group to a new candidate sub-group (manual switchover is possible using tools perform force lag command). Note that, if the port link dampening is deployed, the port level timer must expire before the sub-group-selection takes place and this timer is started. Sub-group-level hold-down timer is supported with LAGs running LACP only.
3. LAG-level hold-time down timer
   This optional timer allows operator to control delay for declaring a LAG operationally down when the available links fall below the required port/BW minimum. The timer is recommended for LAG connecting to MC-LAG systems. The timer prevents a LAG going down when MC-LAG switchover executes break-before-make switch. Note that, if the port link dampening is deployed, the port level timer must expire before the LAG operational status is processed and this timer is started.

BFD over LAG Links

The routersupports the application of BFD to monitor individual LAG link members to speed up the detection of link failures. When BFD is associated with an Ethernet LAG, BFD sessions are setup over each link member, and are referred to as micro-BFD sessions. A link is not operational in the associated LAG until the associated micro-BFD session is fully established. In addition, the link member is removed from the operational state in the LAG if the BFD session fails.

When configuring the local and remote IP address for the BFD over LAG link sessions, the **local-ip** parameter should always match an IP address associated with the IP interface to which this LAG is bound. In addition, the **remote-ip** parameter should match an IP address on the remote system and should also be in the same subnet as the **local-ip** address. If the LAG bundle is re-associated with a different IP interface, the **local-ip** and **remote-ip** parameters should be modified to match the new IP subnet.

Mixed Port-Speed LAG Support

SROS routers support mixing different speed member ports in a single LAG. The LAG must be configured explicitly to allow mixed port-speed operation through the port-weight-speed command. The port-weight-speed defines both the lowest port speed for a member port in that LAG and the type of higher speed ports allowed to be mixed in the same LAG. For example, port-weight-speed 10 defines the minimum member port speed of 10GE and allows addition of any port that has a speed, which is a multiple of 10GE as long as the mix is supported by a given release, refer to specific Release Notes. Any LAG can be configured to support mixed port-speed operation.

For mixed port-speed LAGs:

- Both LACP and non-LACP configurations are supported. With LACP enabled, LACP is unaware of physical port differences.
- QoS is distributed proportionally to port-speed, unless explicitly configured not to do so (see internal-scheduler-weight-mode)
- User data traffic is hashed proportionally to port speed when any per-flow hash is deployed.
- CPM-originated OAM control traffic that requires per LAG hashing is hashed per physical port.
- It is recommended operators use **weight-threshold** instead of **port-threshold** to control LAG operational status. For example, when 10GE and 100GE ports are mixed in a LAG, each 10GE port will have a weight of 1, while each 100GE port will have a weight of 10.

  Note that the weight-threshold can also be used for LAGs not in mixed port-speed mode

to allow common operational model (each port has a weight of 1 to mimic **port-threshold** and related configuration).

- Similarly to the above, it is recommended that operators use weight-based thresholds for other system configurations that react to operational change of LAG member ports, like MCAC (see **use-lag-port-weight**) and VRRP (see **weight-down**)
- When sub-groups are used, the following behavior should be noted for selection criteria:
  - → highest-count – continues to operate on physical link counts. Therefore, a sub-group with lower speed links will be selected even if its total bandwidth is lower. For example: a 4 * 10GE subgroup will be selected over a 100GE + 1 GE sub-group).
  - → highest-weight – continues to operate on operator-configured priorities. Therefore, it is expected that configured weights take into account the proportional bandwidth difference between member ports to achieve the desired behavior. For example, to favor sub-groups with higher bandwidth capacity but lower link count in a 10GE/100GE LAG, 100GE ports need to have their priority set to a value that is at least 10 times that of the 10GE ports priority value.
  - → best-port – continues to operate on operator-configured priorities. Therefore, it is expected that the configured weights will take into account proportional bandwidth difference between member ports to achieve the desired behavior.

Operators can add higher speed member ports to an existing LAG in service when all ports of the lag have the speed as selected by port-weight-speed or when port-weight-speed is disabled (non-mixed port-speed operation). To do so, first port-based thresholds related to that LAG should be switched to weight-based thresholds, and then port-speed-weight should be set to the port speed of the existing member ports. After that, operators can add higher speed ports adjusting weight-based thresholds as required.

Similarly, operators can disable mixed port-speed operation in service if all ports have the same port speed and port-weight-speed equals to member ports' speed. Note that weight-based thresholds may remain to be in use for the LAG.

Feature limitations:

- requires chassis mode D.
- supported on network, access, and hybrid mode LAGs, including MC-LAG.
- supported for standard-port LAGs and on 10GE WAN/100GE LAN port combinations.
- PIM lag-usage-optimization is not supported and must not be configured.
- LAG member links must have the default configuration for **config port ethernet egress-rate/ingress-rate**.
- not supported on 7450 ESS-6V and 7710 platforms.
- not supported for ESM
- not supported with weighted per-link-hash

Multi-Chassis LAG

This section describes the Multi-Chassis LAG (MC-LAG) concept. MC-LAG is an extension of a LAG concept that provides node-level redundancy in addition to link-level redundancy provided by “regular LAG”.

Typically, MC-LAG is deployed in a network-wide scenario providing redundant connection between different end points. The whole scenario is then built by combination of different mechanisms (for example, MC-LAG and redundant pseudowire to provide e2e redundant p2p connection or dual homing of DSLAMs in Layer 2/3 TPSDA).

Overview

Multi-chassis LAG is a method of providing redundant Layer 2/3 access connectivity that extends beyond link level protection by allowing two systems to share a common LAG end point.

The multi-service access node (MSAN) node is connected with multiple links towards a redundant pair of Layer 2/3 aggregation nodes such that both link and node level redundancy, are provided. By using a multi-chassis LAG protocol, the paired Layer 2/3 aggregation nodes (referred to as redundant-pair) appears to be a single node utilizing LACP towards the access node. The multi-chassis LAG protocol between redundant-pair ensures a synchronized forwarding plane to/from the access node and is used to synchronize the link state information between the redundant-pair nodes such that proper LACP messaging is provided to the access node from both redundant-pair nodes.

In order to ensure SLAs and deterministic forwarding characteristics between the access and the redundant-pair node, the multi-chassis LAG function provides an active/standby operation towards/from the access node. LACP is used to manage the available LAG links into active and standby states such that only links from 1 aggregation node are active at a time to/from the access node.

Alternatively, when access nodes does not support LACP, the **power-off** option can be used to enforce active/standby operation. In this case, the standby ports are **trx_disabled** (power off transmitter) to prevent usage of the lag member by the access-node.Characteristics related to MC are:

- Selection of the common system ID, system-priority and administrative-key are used in LACP messages so partner systems consider all links as the part of the same LAG.
- Extension of selection algorithm in order to allow selection of active sub-group.
  - → The sub-group definition in LAG context is still local to the single box, meaning that even if sub-groups configured on two different systems have the same sub-group-id they are still considered as two separate subgroups within given LAG.
  - → Multiple sub-groups per PE in a MC-LAG is supported.
  - → In case there is a tie in the selection algorithm, for example, two sub-groups with identical aggregate weight (or number of active links) the group which is local to the system with lower system LACP priority and LAG system ID is taken.
- Providing inter-chassis communication channel allows inter-chassis communication to support LACP on both system. This communication channel enables the following:
  - → Supports connections at the IP level which do not require a direct link between two nodes. The IP address configured at the neighbor system is one of the addresses of the system (interface or loop-back IP address).
  - → The communication protocol provides heartbeat mechanism to enhance robustness of the MC-LAG operation and detecting node failures.
  - → Support for operator actions on any node that force an operational change.

→ The LAG group-ids do not have to match between neighbor systems. At the same time, there can be multiple LAG groups between the same pair of neighbors.

→ Verification that the physical characteristics, such as speed and auto-negotiation is configured and initiates operator notifications (traps) if errors exist. Consistency of MC-LAG configuration (system-id, administrative-key and system-priority) is provided. Similarly, load-balancing mode of operation must be consistently configured on both nodes.

→ Traffic over the signalling link is encrypted using a user configurable message digest key.

- MC-LAG function provides active/stand-by status to other software applications in order to built a reliable solutions.

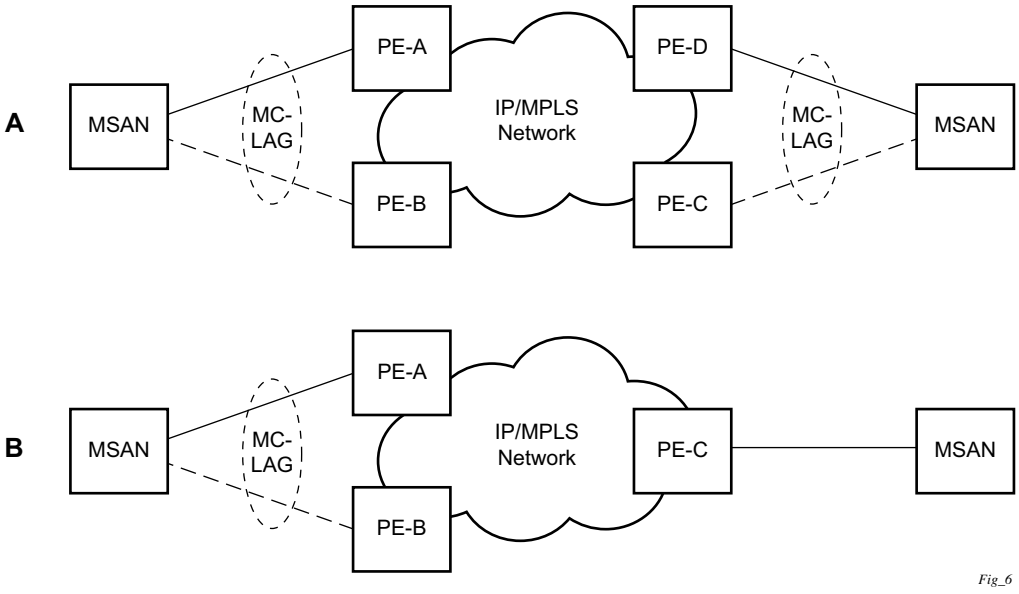

Figure 27: MC-LAG L2 Dual Homing to Remote PE Pairs

Figure 27 depicts different combinations of MC-LAG attachments supported. The supported configurations can be sub-divided into following sub-groups:

- Dual-homing to remote PE pairs
  - → both end-points attached with MC-LAG
  - → one end-point attached
- Dual-homing to local PE pair
  - → both end-points attached with MC-LAG
  - → one end-point attached with MC-LAG
  - → both end-points attached with MC-LAG to two overlapping pairs

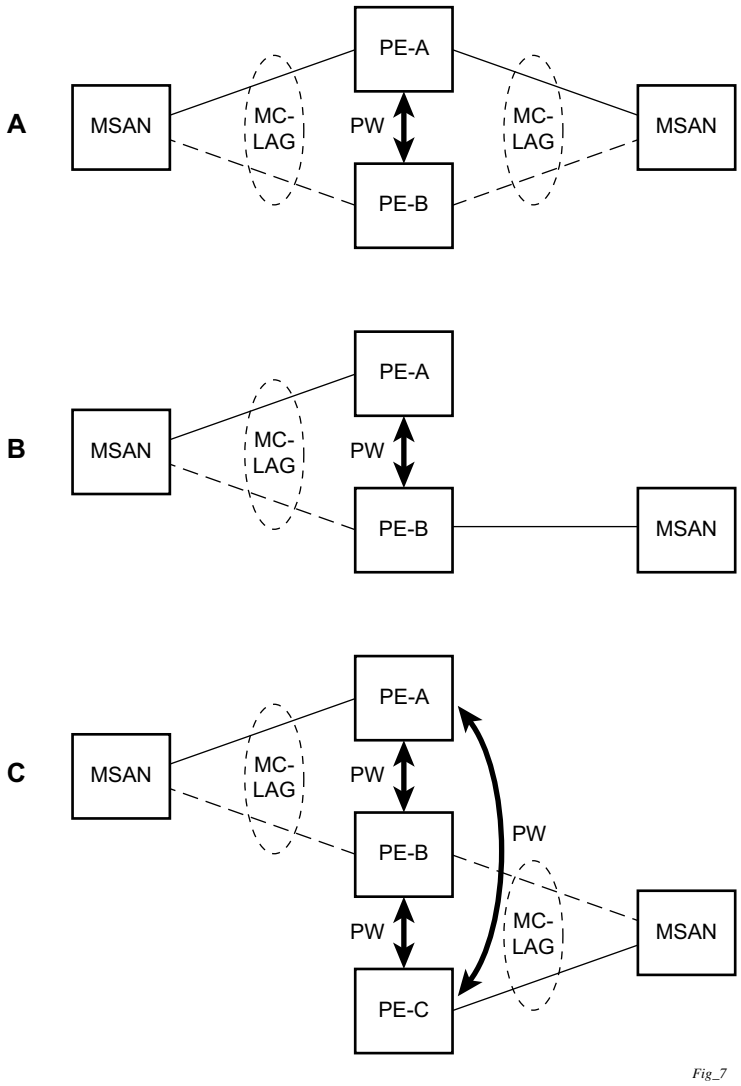

Figure 28: MC-LAG L2 Dual Homing to Local PE-Pairs

The forwarding behavior of the nodes abide by the following principles. Note that logical destination (actual forwarding decision) is primarily determined by the service (VPLS or VLL) and the principle below applies only if destination or source is based on MC-LAG:

- Packets received from the network will be forwarded to all local active links of the given destination-sap based on conversation hashing. In case there are no local active links, the packets will be cross-connected to inter-chassis pseudowire.
- Packets received from the MC-LAG sap will be forwarded to active destination pseudo-wire or active local links of destination-sap. In case there are no such objects available at the local node, the packets will be cross-connected to inter-chassis pseudowire.

MC-LAG and Subscriber Routed Redundancy Protocol (SRRP)

MC-LAG and SRRP enables dual-homed links from any IEEE 802.1ax (formerly 802.3ad) standards-based access device (for example, a IP DSLAM, Ethernet switch or a Video on Demand server) to multiple Layer 2/3 or Layer 3 aggregation nodes. In contrast with slow recovery mechanisms such as Spanning Tree, multi-chassis LAG provides synchronized and stateful redundancy for VPN services or triple play subscribers in the event of the access link or aggregation node failing, with zero impact to end users and their services.

Refer to the 7750 SR OS Triple Play Guide for information about SRRP.

Point-to-Point (p2p) Redundant Connection Across Layer 2/3 VPN Network

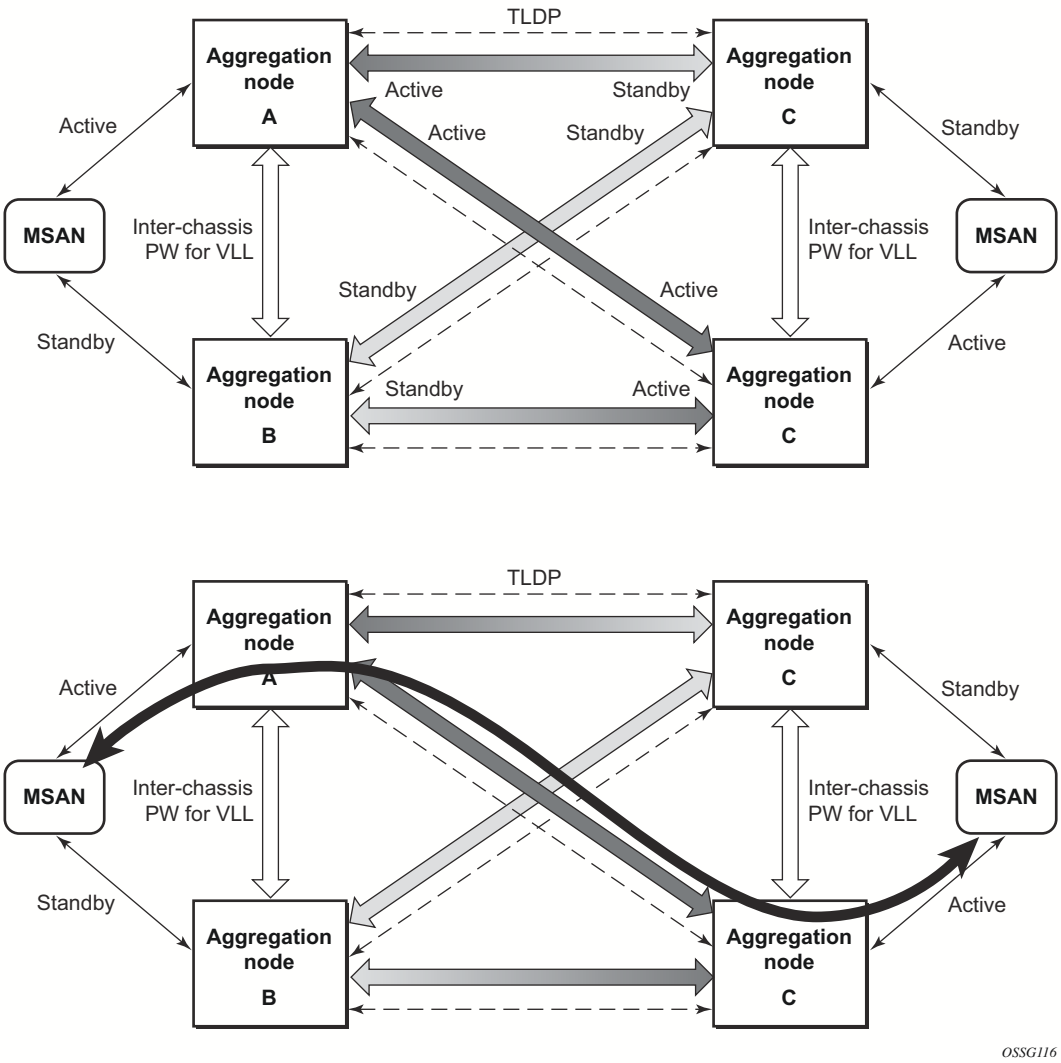

Figure 29: P2P Redundant Connection Through a Layer 2 VPN Network

Figure 29 shows the connection between two multi-service access nodes (MSANs) across network based on Layer 2/3 VPN pseudo-wires. The connection between MSAN and a pair of PE routers is realized by MC-LAG. From MSAN perspective, redundant pair of PE routers acts as a single partner in LACP negotiation. At any point in time, only one of the routers has an active link(s) in a given LAG. The status of LAG links is reflected in status signaling of pseudo-wires set between all

participating PEs. The combination of active and stand-by states across LAG links as well and pseudo-wires give only 1 unique path between pair of MSANs.

Note that the configuration in Figure 29 depicts one particular configuration of VLL connections based on MC-LAG, particularly the VLL connection where two ends (SAPs) are on two different redundant-pairs. In addition to this, other configurations are possible, such as:

- Both ends of the same VLL connections are local to the same redundant-pair.
- One end VLL endpoint is on a redundant-pair the other on single (local or remote) node.

DSLAM Dual Homing in Layer 2/3 TPSDA Model

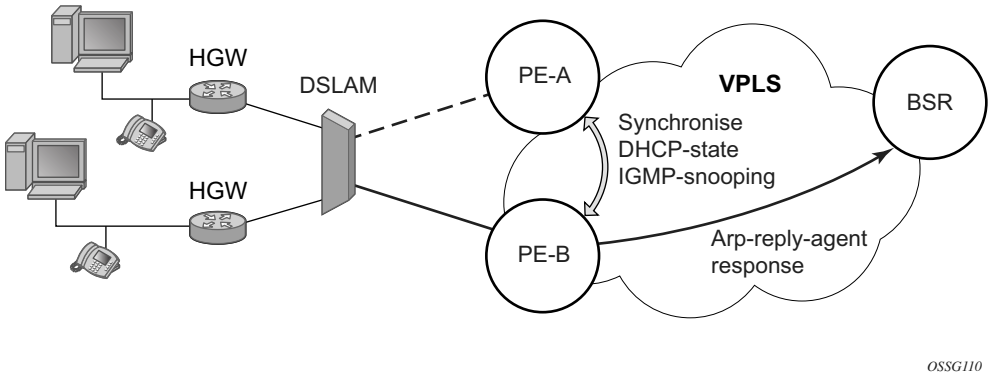

Figure 30: DSLAM Dual-Homing Using MC-LAG

Figure 30 illustrates a network configuration where DSLAM is dual homed to pair of redundant PEs by using MC-LAG. Inside the aggregation network redundant-pair of PEs is connecting to VPLS service which provides reliable connection to single or pair of Broadband Service Routers (BSRs).

MC-LAG and pseudo-wire connectivity, PE-A and PE-B implement enhanced subscriber management features based on DHCP-snooping and creating dynamic states for every subscriber-host. As in any point of time there is only one PE active, it is necessary to provide the mechanism for synchronizing subscriber-host state-information between active PE (where the state is learned) and stand-by PE. In addition, VPLS core must be aware of active PE in order to forward all subscriber traffic to a PE with an active LAG link. The mechanism for this synchronization is outside of the scope of this document.

G.8031 Protected Ethernet Tunnels

Alcatel-Lucent PBB implementation offers the capability to use core Ethernet tunnels compliant with ITU-T G.8031 specification to achieve 50 ms resiliency for failures in a native Ethernet backbone. For further information regarding Ethernet tunnels, see **G.8031 Protected Ethernet Tunnels** in the Services Guide.

G.8032 Protected Ethernet Rings

Ethernet ring protection switching offers ITU-T G.8032 specification compliance to achieve resiliency for Ethernet Layer 2 networks. Similar to G.8031 linear protection (also called Automatic Protection Switching (APS)), G.8032 (Eth-ring) is also built on Ethernet OAM and often referred to as Ring Automatic Protection Switching (R-APS).

For further information regarding Ethernet rings, see G.8032 Protected Ethernet Rings section in the Services Guide.

Ethernet Port Monitoring

Ethernet ports can record and recognize various medium statistics and errors. There are two main types of errors:

- Frame Based — Frame based errors are counted when the arriving frame has an error that means the frame is invalid. These types of errors are only detectable when frames are presents on the wire.
- Symbol Based — Symbol errors are invalidly encoded symbols on the physical medium. Symbols are always present on an active Ethernet port regardless of the presence of frames.

CRC-Monitor and Symbol-Monitor allows the operator to monitor ingress error conditions on the Ethernet medium and compare these error counts to the thresholds. CRC-Monitor monitors CRC errors. Symbol-Monitor monitors symbol errors. Symbol Error is not supported on all Ethernet ports. Crossing a signal degrade (SD) threshold will cause a log event to be raised. Crossing the configured signal failure (SF) threshold will cause the port to enter an operation state of down. The operator may consider the configuration of other protocols to convey the failure, through timeout conditions.

The error rates are in the form of M*10E-N. The operator has the ability to configure both the threshold (N) and a multiplier (M). By default if the multiplier is not configured the multiplier is 1. As an example, sd-threshold 3 would result in a signal degrade error rate of 1*10E-3 (one error per 1000). Changing the configuration to would sd-threshold 3 multiplier 5 result in a signal degrade rate of 5*10E-3 (5 errors per 1000). The signal degrade value must be a lower error rate than the signal failure threshold. This threshold can be used to provide notification that the port is operating in a degraded but not failed condition. These do not equate to a bit error rate (BER). CRC-Monitor provides a CRC error rate. Symbol-Monitor provides a symbol error rate.

The configured error thresholds are compared to the operator specified sliding window to determine if one or both of the thresholds have been crossed. Statistics are gathered every second. This means that every second the oldest statistics are dropped from the calculation. The default 10 second sliding window means that at the 11th second the oldest 1 second statistical data is dropped and the 11th second is included.

Symbol error crossing differs slightly from CRC based error crossing. The error threshold crossing is calculated based on the window size and the fixed number of symbols that will arrive (ingress) that port during that window. The following configuration is used to demonstrate this concept.

```
config>port>ethernet# info detail
----------------------------------------------
            symbol-monitor
                sd-threshold 5 multiplier 5
                sf-threshold 3 multiplier 5
                no shutdown
            exit

show port 2/1/2 ethernet
===============================================================================
Ethernet Interface
===============================================================================
Description        : 2/1/2
Interface          : 2/1/2                       Oper Speed       : N/A
Link-level         : Ethernet                    Config Speed     : 1 Gbps
Admin State        : down                        Oper Duplex      : N/A
Oper State         : down                        Config Duplex    : full
Physical Link      : No                          MTU              : 9212
Single Fiber Mode  : No                          Min Frame Length : 64 Bytes
IfIndex            : 69271552                    Hold time up     : 0 seconds
Last State Change  : 06/29/2014 05:04:12         Hold time down   : 0 seconds
Last Cleared Time  : N/A                         DDM Events       : Enabled
Phys State Chng Cnt: 0

Configured Mode    : network                     Encap Type       : null
Dot1Q Ethertype    : 0x8100                      QinQ Ethertype   : 0x8100
PBB Ethertype      : 0x88e7
Ing. Pool % Rate   : 100                         Egr. Pool % Rate : 100
Ing. Pool Policy   : n/a
Egr. Pool Policy   : n/a
Net. Egr. Queue Pol: default
Egr. Sched. Pol    : n/a
Auto-negotiate     : true                        MDI/MDX          : unknown
Oper Phy-tx-clock  : not-applicable
Accounting Policy  : None                        Collect-stats    : Disabled
Acct Plcy Eth Phys : None                        Collect Eth Phys : Disabled
Egress Rate        : Default                     Ingress Rate     : Default
Load-balance-algo  : Default                     LACP Tunnel      : Disabled

Down-when-looped   : Disabled                    Keep-alive       : 10
Loop Detected      : False                       Retry            : 120
Use Broadcast Addr : False

Sync. Status Msg.  : Disabled                    Rx Quality Level : N/A
Tx DUS/DNU         : Disabled                    Tx Quality Level : N/A
SSM Code Type      : sdh

Down On Int. Error : Disabled

CRC Mon SD Thresh  : Disabled                    CRC Mon Window   : 10 seconds
CRC Mon SF Thresh  : Disabled

Sym Mon SD Thresh  : 5*10E-5                     Sym Mon Window   : 10 seconds
Sym Mon SF Thresh  : 5*10E-3                     Tot Sym Mon Errs : 0

EFM OAM            : Disabled                    EFM OAM Link Mon : Disabled

Configured Address : 8c:90:d3:a0:c7:42
Hardware Address   : 8c:90:d3:a0:c7:42
```

```
Transceiver Data

Transceiver Status : not-equipped
===============================================================================
Traffic Statistics
===============================================================================
                                                    Input                Output
-------------------------------------------------------------------------------
Octets                                                  0                     0
Packets                                                 0                     0
Errors                                                  0                     0
===============================================================================
===============================================================================
Port Statistics
===============================================================================
                                                    Input                Output
-------------------------------------------------------------------------------
Unicast Packets                                         0                     0
Multicast Packets                                       0                     0
Broadcast Packets                                       0                     0
Discards                                                0                     0
Unknown Proto Discards                                  0
===============================================================================
===============================================================================
Ethernet-like Medium Statistics
===============================================================================
Alignment Errors :                  0  Sngl Collisions  :                    0
FCS Errors       :                  0  Mult Collisions  :                    0
SQE Test Errors  :                  0  Late Collisions  :                    0
CSE              :                  0  Excess Collisns  :                    0
Too long Frames  :                  0  Int MAC Tx Errs  :                    0
Symbol Errors    :                  0  Int MAC Rx Errs  :                    0
In Pause Frames  :                  0  Out Pause Frames :                    0
===============================================================================
```

The above configuration results in an SD threshold of 5*10E-5 (0.00005) and an SF threshold of 5*10E-3 (0.005) over the default 10 second window. If this port is a 1GbE port supporting symbol monitoring then the error rate is compared against 1,250,000,000 symbols (10 seconds worth of symbols on a 1GbE port 125,000,000). If the error count in the current 10 second sliding window is less than 62,500 then the error rate is below the signal degrade threshold and no action is taken. If the error count is between 62,501 and 6,250,000 then the error rate is above signal degrade but has not breached the signal failure signal threshold and a log event will be raised. If the error count is above 6,250,000 the signal failure threshold is crossed and the port will enter an operation state of down. Consider that this is a very simple example meant to demonstrate the function and not meant to be used as a guide for configuring the various thresholds and window times.

A port is not returned to service automatically when a port enters the failed condition as a result of crossing a signal failure threshold for both CRC-Monitor and Symbol-Monitor. Since the port is operationally down without a physical link error monitoring stops. The operator may enable the port using the **shutdown** and **no shutdown port** commands. Other port transition functions like clearing the MDA or slot, removing the cable, and other physical link transition functions.

802.3ah OAM

802.3ah Clause 57 (**efm-oam**) defines the Operations, Administration, and Maintenance (OAM) sub-layer, which provides mechanisms useful for monitoring link operation such as remote fault indication and remote loopback control. In general, OAM provides network operators the ability to monitor the health of the network and quickly determine the location of failing links or fault conditions. **efm-oam** described in this clause provides data link layer mechanisms that complement applications that may reside in higher layers.

OAM information is conveyed in slow protocol frames called OAM protocol data units (OAMPDUs). OAMPDUs contain the appropriate control and status information used to monitor, test and troubleshoot OAM-enabled links. OAMPDUs traverse a single link, being passed between peer OAM entities, and as such, are not forwarded by MAC clients (like bridges or switches).

The following **efm-oam** functions are supported:

- **efm-oam** capability discovery.
- Active and passive modes.
- Remote failure indication — Handling of critical link events (link fault, dying gasp, etc.)
- Loopback — A mechanism is provided to support a data link layer frame-level loopback mode. Both remote and local loopback modes are supported.
- **efm-oam** PDU tunneling.
- High resolution timer for **efm-oam** in 100ms interval (minimum).
- **efm-oam** kink monitoring

When the **efm-oam** protocol fails to negotiate a peer session or encounters a protocol failure following an established session the *Port State* will enter the *Link Up* condition. This port state is used by many protocols to indicate the port is administratively UP and there is physical connectivity but a protocol, such as **efm-oam**, has caused the ports operational state to enter a DOWN state. A reason code has been added to help discern if the **efm-oam** protocol is the underlying reason for the Link Up condition.

```
show port
===============================================================================
Ports on Slot 1
===============================================================================
Port        Admin Link Port    Cfg  Oper LAG/ Port Port Port   C/QS/S/XFP/
Id          State      State   MTU  MTU  Bndl Mode Encp Type   MDIMDX
-------------------------------------------------------------------------------
1/1/1       Down  No   Down    1578 1578    - netw null xcme
1/1/2       Down  No   Down    1578 1578    - netw null xcme
1/1/3       Up    Yes  Link Up 1522 1522    - accs qinq xcme
1/1/4       Down  No   Down    1578 1578    - netw null xcme
1/1/5       Down  No   Down    1578 1578    - netw null xcme
1/1/6       Down  No   Down    1578 1578    - netw null xcme

# show port 1/1/3
```

```
===============================================================================
Ethernet Interface
===============================================================================
Description        : 10/100/Gig Ethernet SFP
Interface          : 1/1/3                       Oper Speed       : N/A
Link-level         : Ethernet                    Config Speed     : 1 Gbps
Admin State        : up                          Oper Duplex      : N/A
Oper State         : down                        Config Duplex    : full
Reason Down        : efmOamDown
Physical Link      : Yes                         MTU              : 1522
Single Fiber Mode  : No                          Min Frame Length : 64 Bytes
IfIndex            : 35749888                    Hold time up     : 0 seconds
Last State Change  : 12/18/2012 15:58:29         Hold time down   : 0 seconds
Last Cleared Time  : N/A                         DDM Events       : Enabled
Phys State Chng Cnt: 1

Configured Mode    : access                      Encap Type       : QinQ
Dot1Q Ethertype    : 0x8100                      QinQ Ethertype   : 0x8100
PBB Ethertype      : 0x88e7
Ing. Pool % Rate   : 100                         Egr. Pool % Rate : 100
Ing. Pool Policy   : n/a
Egr. Pool Policy   : n/a
Net. Egr. Queue Pol: default
Egr. Sched. Pol    : n/a
Auto-negotiate     : true                        MDI/MDX          : unknown
Oper Phy-tx-clock  : not-applicable
Accounting Policy  : None                        Collect-stats    : Disabled
Acct Plcy Eth Phys : None                        Collect Eth Phys : Disabled
Egress Rate        : Default                     Ingress Rate     : Default
Load-balance-algo  : Default                     LACP Tunnel      : Disabled

Down-when-looped   : Disabled                    Keep-alive       : 10
Loop Detected      : False                       Retry            : 120
Use Broadcast Addr : False

Sync. Status Msg.  : Disabled                    Rx Quality Level : N/A
Tx DUS/DNU         : Disabled                    Tx Quality Level : N/A
SSM Code Type      : sdh

Down On Int. Error : Disabled

CRC Mon SD Thresh  : Disabled                    CRC Mon Window   : 10 seconds
CRC Mon SF Thresh  : Disabled

Configured Address : d8:ef:01:01:00:03
Hardware Address   : d8:ef:01:01:00:03
```

The operator also has the opportunity to decouple the **efm-oam** protocol from the port state and operational state. In cases where an operator wants to remove the protocol, monitor the protocol only, migrate, or make changes the **ignore-efm-state** can be configured in the **port>ethernet>efm-oam** context. When the **ignore-efm-state** command is configured on a port the protocol continues as normal. However, ANY failure in the protocol state machine (discovery, configuration, time-out, loops, etc.) will not impact the port on which the protocol is active and the optional ignore command is configured. There will only be a protocol warning message if there are issues with the protocol. The default behavior when this optional command is not configured

means the port state will be affected by any **efm-oam** protocol fault or clear conditions. Adding and removing this optional ignore command will immediately represent the *Port State* and *Oper State* based on the active configuration. For example, if the **ignore-efm-state** is configured on a port that is exhibiting a protocol error that protocol error does not affect the port state or operational state and there is no *Reason Down* code. If the **ignore-efm-state** is removed from a port with an existing **efm-oam** protocol error, the port will transition to *Link UP*, *Oper Down* with the reason code *efmOamDown*.

OAM Events

The Information OAMPDU is transmitted by each peer at the configured intervals. This OAMPDU performs keepalive and critical notification functions. Various local conditions are conveyed through the setting of the Flags field. The following Critical Link Event defined in IEEE 802.3 Section 57.2.10.1 are supported;

- Link Fault: The PHY has determined a fault has occurred in the receive direction of the local DTE
- Dying Gasp: An unrecoverable local failure condition has occurred
- Critical Event: An unspecified critical event has occurred

The local node can set an unset the various Flag fields based on the operational state of the port, shutdown or activation of the efm-oam protocol or locally raised events. These Flag fields maintain the setting for the continuance of a particular event. Changing port conditions, protocol state or operator intervention may impact the setting of these fields in the Information OAMPDU.

A peer processing the Information OAMPDU can take a configured action when one or more of these Flag fields are set. By default, receiving a set value for any of the Flag fields will cause the local port to enter the previous mentioned *Link Up* port state and an event will be logged. If this default behavior is not desired, the operator may choose to log the event without affecting the local port. This is configurable per Flag field using the options under **config>port>ethernet>efm-oam>peer-rdi-rx**.

Link Monitoring

The efm-oam protocol provides the ability to monitor the link for error conditions that may indicate the link is starting to degrade or has reached an error rate that exceeds acceptable threshold.

Link monitoring can be enabled for three types of frame errors; **errored-frame**, **errored-frame-period** and **errored-frame-seconds**. The **errored-frame** monitor is the number of frame errors compared to the threshold over a window of time. The **errored-frame-period** monitor is the number of frame errors compared to the threshold over a window of number of received packets. This window is checked once per second to see if the window parameter has been reached. The **errored-frame-seconds** monitor is the number of errored seconds compared to the threshold over a window of time. An errored second is any second with a single frame error.

An errored frame is counted when any frame is in error as determined by the Ethernet physical layer, including jabbers, fragments, FCS or CRC and runts. This excludes jumbo frames with a byte count higher than 9212, or any frame that is dropped by the phy layer prior to reaching the monitoring function.

Each frame error monitor functions independently of other monitors. Each of monitor configuration includes an optional signal degrade threshold **sd-threshold**, a signal failure threshold **sf-threshold**, a **window** and the ability to communicate failure events to the peer by setting a Flag field in the Information OAMPDU or the generation of the Event Notification OAMPDU, **event-notification**. The parameters are uniquely configurable for each monitor.

A degraded condition is raised when the configured signal degrade **sd-threshold** is reached. This provides a first level log only action indicating a link could become unstable. This event does not affect the port state. The critical failure condition is raised when the configured **sf-threshold** is reached. By default, reaching the signal failure threshold will cause the port to enter the *Link Up* condition unless the local signal failure **local-sf-action** has been modified to a **log-only** action. Signal degrade conditions for a monitor in signal failed state will be suppressed until the signal failure has been cleared.

The initial configuration or the modification of either of the threshold values will take affect in the current window. When a threshold value for a monitor is modified, all active local events for that specific monitor will be cleared. The modification of the threshold acts the same as the **clear** command described later in this section.

Notification to the peer is required to ensure the action taken by the local port detecting the error and its peer are synchronized. If peers do not take the same action then one port may remain fully operational while the other enters a non-operational state. These threshold crossing events do not shutdown the physical link or cause the protocol to enter a non-operational state. The protocol and network element configuration is required to ensure these asymmetrical states do not occur. There are two options for exchanging link and event information between peers; Information OAMPDU and the Event Notification OAMPDU.

As discussed earlier, the Information OAMPDU conveys link information using the Flags field; dying gasp, critical link and link fault. This method of communication has a number of significant advantages over the Event Notification OAMPDU. The Information OAMPDU is sent at every configured **transmit-interval**. This will allow the most recent information to be sent between peers, a critical requirement to avoid asymmetrical forwarding conditions. A second major advantage is interoperability with devices that do not support Link Monitoring and vendor interoperability. This is the lowest common denominator that offers a robust communication to convey link event information. Since the Information OAMPDU is already being sent to maintain the peering relationship this method of communication adds no additional overhead. The l**ocal-sf-action** options allow the dying gasp and critical event flags to be set in the Information OAMPDU when a signal failure threshold is reached. It is suggested that this be used in place of or in conjunction with Event Notification OAMPDU.

Event Notification OAMPDU provides a method to convey very specific information to a peer about various Link Events using Link Event TLVs. A unique Event Notification OAMPDU will be generated for each unique frame error event. The intension is to provide the peer with the Sequence Number, Event Type, Timestamp, and the local information that caused the generation of the OAMPDU; window, threshold, errors and error running total and event running total specific to the port.

- Sequence Number: The unique identification indicating a new event.
- Window: The size of the unique measurement period for the error type. The window is only checked at the end. There is not mid-window checking.
- Threshold: The value of the configured sf-threshold
- Errors: The errors counted in that specific window
- Error Running Total: The number of errors accumulated for that event type since monitoring started and the protocol and port have been operational or a reset function has occurred
- Event Running Total: The number of events accumulated for that event type since the monitoring started and the protocol and port have been operational

By default, the Event Notification OAMPDU is generated by the network element detecting the signal failure event. The Event Notification OAMPDU is sent only when the initial frame event occurs. No Event Notification OAMPDU is sent when the conditions clears. A port that has been operationally affected as a result of a Link Monitoring frame error event must be recovered manually. The typical recovery method is to shutdown the port and no shutdown the port. This will clear all events on the port. Any function that affects the port state, physical fiber pull, soft or hard reset functions, protocol restarts, etc will also clear the all local and remote events on the affected node experiencing the operation. None of these frame errors recovery actions will cause the generation of the Event Notification OAMPDU. If the chosen recovery action is not otherwise recognized by the peer and the Information OAMPDU Flag fields have not been configured to maintain the current event state, there is a high probability that the ports will have different forwarding states, notwithstanding any higher level protocol verification that may be in place.

A burst of between one and five Event Notification OAMPDU packets may be sent. By default, only a single Event Notification OAMPDU is generated, but this value can be changed under the **local-sf-action** context. An Event Notification OAMPDU will only be processed if the peer had previously advertised the EV capability. The EV capability is an indication the remote peer supports link monitoring and may send the Event Notification OAMPDU.

The network element receiving the Event Notification OAMPDU will use the values contained in the Link event TLVs to determine if the remote node has exceeded the failure threshold. The locally configured action will determine how and if the local port is affected. By default, processing of the Event Notification OAMPDU is log only and does not affect the port state. By default, processing of the Information OAMPDU Flag fields is port affecting. When Event Notification OAMPDU has been configured as port affecting on the receiving node, action is only taken when errors are equal to or above the threshold and the threshold value is not zero. No action is taken when the errors value is less than the threshold or the threshold is zero.

Symbol error, **errored-symbols**, monitoring is also supported but requires specific hardware revisions and the appropriate code release. The symbol monitor differs from than the frame error monitors. Symbols represent a constant load on the Ethernet wire whether service frames are present or not. This means the optional signal degrade threshold **sd-threshold** has an additional purpose when configured as part of the symbol error monitor. When the signal degrade threshold

is not configured, the symbol monitor acts similar to the frame error monitors, requiring manual intervention to clear a port that has been operationally affected by the monitor. When the optional signal degrade threshold is configured, it again represents the first level warning. However, it has an additional function as part of the symbol monitor. If a signal failure event has been raised, the configured signal degrade threshold becomes the equivalent to a lowering threshold. If a subsequent window does not reach the configured signal degrade threshold then the previous event will be cleared and the previously affected port will be returned to service without operator intervention. This return to service will automatically clear any previously set Information OAMPDU Flags fields set as a result of the signal failure threshold. The Event Notification OAMPDU will be generated with the symbol error Link TLV that contains an error count less than the threshold. This will indicate to the peer that initial problem has been resolved and the port should be returned to service.

The **errored-symbol** window is a measure of time that is automatically converted into the number of symbols for that specific medium for that period of time. The standard MIB entries "dot3OamErrSymPeriodWindowHi" and "dot3OamErrSymPeriodWindowLo" are marked as read-only instead of read-write. There is now way to directly configure these values. The configuration of the **window** will convert the time and program those two MIB values in an appropriate manner. Both the configured **window** and the number of symbols will be displayed under the **show port** *port-id* **ethernet efm-oam** command.

```
show port 1/1/1 ethernet efm-oam
===============================================================================
Ethernet Oam (802.3ah)
===============================================================================
Admin State        : up
Oper State         : link fault
Mode               : active
Pdu Size           : 1518
Config Revision    : 0
Function Support   : LB
Transmit Interval  : 1000 ms
Multiplier         : 5
Hold Time          : 0
Tunneling          : false
Loop Detected      : false
Grace Tx Enable    : true (inactive)

No Peer Information Available

Loopback State     : None
Loopback Ignore Rx : Ignore
Ignore Efm State   : false
Link Monitoring    : disabled

Peer RDI Rx
  Critical Event   : out-of-service
  Dying Gasp       : out-of-service
  Link Fault       : out-of-service
  Event Notify     : log-only

Local SF Action                        Discovery
  Event Burst      : 1                   Ad Link Mon Cap  : yes
```

```
  Port Action      : out-of-service
  Dying Gasp       : disabled
  Critical Event   : disabled

Errored Frame                           Errored Frame Period
  Enabled          : no                   Enabled          : no
  Event Notify     : enabled              Event Notify     : enabled
  SF Threshold     : 10                   SF Threshold     : 1
  SD Threshold     : disabled (0)         SD Threshold     : disabled (0)
  Window           : 10 ds                Window           : 1488095 frames

Errored Symbol Period                   Errored Frame Seconds Summary
  Enabled          : no                   Enabled          : no
  Event Notify     : enabled              Event Notify     : enabled
  SF Threshold     : 1                    SF Threshold     : 1
  SD Threshold     : disabled (0)         SD Threshold     : disabled (0)
  Window (time)    : 10 ds                Window           : 600 ds
  Window (symbols) : 125000000
===============================================================================
Active Failure Ethernet OAM Event Logs
===============================================================================
Number of Logs : 0
===============================================================================

===============================================================================
Ethernet Oam Statistics
===============================================================================
                                                    Input                Output
-------------------------------------------------------------------------------
Information                                             0                     0
Loopback Control                                        0                     0
Unique Event Notify                                     0                     0
Duplicate Event Notify                                  0                     0
Unsupported Codes                                       0                     0
Frames Lost                                                                   0
===============================================================================
```

A **clear** command "**clear port** *port-id* **ethernet efm-oam events [local | remote]**" has been added to clear port affecting events on the local node on which the command is issued. When the optional [**local | remote**] options are omitted, both local and remote events will be cleared for the specified port. This command is not specific to the link monitors as it clears all active events. When local events are cleared, all previously set Information OAMPDU Flag fields will be cleared regardless of the cause the event that set the Flag field.

In the case of symbol errors only, if Event Notification OAMPDU is enabled for symbol errors and a local symbol error signal failure event exists at the time of the clear, the Event Notification OAMPDU will be generate with an error count of zero and the threshold value reflecting the local signal failure threshold. The fact the error values is lower than threshold value indicates the local node is not in a signal failed state. The Event Notification OAMPDU is not generated in the case where the clear command is used to clear local frame error events. This is because frame error event monitors will only act on an Event Notification OAMPDU when the error value is higher than the threshold value, a lower value is ignored. As stated previously, there is no automatic return to service for frame errors.

If the clear command is used to clear remote events, events conveyed to the local node by the peer, no notification is generated to the peer to indicate a clear function has been performed. Since the Event Notification OAMPDU is only sent when the initial event was raised, there is no further Event Notification and blackholes can result. If the Information OAMPDU Flag fields are used to ensure a constant refresh of information, the remote error will be reinstated as soon as the next Information OAMPDU arrives with the appropriate Flag field set.

Local and remote efm-oam port events are stored in the efm-oam event logs. These logs maintain and display active and cleared signal failure degrade events. These events are interacting with the efm-oam protocol. This logging is different than the time stamped events for information logging purposes included with the system log. To view these events, the **event-log** option has been added to the s**how port** *port-id* **ethernet efm-oam** command. This includes the location, the event type, the counter information or the decoded Network Event TLV information, and if the port has been affected by this active event. A maximum of 12 port events will be retained. The first three indexes are reserved for the three Information Flag fields, dying gasp, critical link, and link fault. The other nine indexes will maintain the current state for the various error monitors in a most recent behavior and events can wrap the indexes, dropping the oldest event.

```
show port 1/2/1 ethernet efm-oam event-logs
===============================================================================
Active Failure Ethernet OAM Event Logs
===============================================================================
Log Index               : 4
Event Time Reference    : 0d 07:01:45
Location                : remote
Type                    : Errored Frame
Window                  : 50
Threshold               : 100
Value                   : 100
Running Total           : 100
Event Total             : 1
Port Affecting          : yes
-------------------------------------------------------------------------------
Number of Logs : 1
===============================================================================

===============================================================================
Active Degraded Ethernet OAM Event Logs
===============================================================================
Number of Logs : 0
===============================================================================

===============================================================================
Cleared Failure Ethernet OAM Event Logs
===============================================================================
Log Index               : 2
Event Time Reference    : 0d 06:59:08
Location                : remote
Type                    : Dying Gasp
Event Total             : 16
-------------------------------------------------------------------------------
Number of Logs : 1
===============================================================================
```

```
===============================================================================
Cleared Degraded Ethernet OAM Event Logs
===============================================================================
Number of Logs : 0
===============================================================================
```

SRoS supports the vendor specific soft reset graceful recovery of efm-oam through the configuration of **grace-tx-enable** under the **config>system>ethernet>efm-oam** and the **config>port>ethernet>efm-oam** contexts. This feature is not enabled by default. When this functionality is enabled the efm-oam protocol does not enter a non-operational state when both nodes understand the grace function. The ports associated with the hardware that has successfully executed the soft reset will clear all local and remote events. The peer that understands the graceful restart procedure for efm-oam will clear all remote events that it received from the peer that undergone the soft reset. The local events will not be cleared on the peer that has not undergone soft reset. Again, the Information OAMPDU Flag fields are critical in propagating the local event to the peer. Remember, the Event Notification OAMPDU will not be sent because it is only sent on the initial raise.

In mixed environments where Link Monitoring is supported on one peer but not the other the following behavior is normal, assuming the Information OAMPDU has been enabled to convey the monitor fault event. The arriving Flag field fault will trigger the efm-oam protocol on the receiving unsupportive node to move from operational to “send local and remote”. The protocol on the supportive node that set the Flag field to convey the fault will enter the “send local and remote ok” state. The supportive node will maintain the Flag field setting until the condition has cleared. The protocol will recover to the operational state once the original event has cleared; assuming no other fault on the port is preventing the negotiation from progressing. If both nodes were supportive of the Link Monitoring process, the protocol would remained operational.

In summary, Link monitors can be configured for frame and symbol monitors (specific hardware only). By default, Link Monitoring and all monitors are shutdown. When the Link Monitoring function is enabled, the capability (EV) will be advertised. When a monitor is enabled, a default window size and a default signal failure threshold are activated. The local action for a signal failure threshold event is to shutdown the local port. Notification will be sent to the peer using the Event Notification OAMPDU. By default, the remote peer will not take any port action for the Event Notification OAMPDU. The reception will only be logged. It is suggested the operator evaluate the various defaults and configure the **local-sf-action** to set one of the Flag fields in the Information OAMPDU using the **info-notifications** command options when fault notification to a peer is required. Vendor specific TLVs and vendors specific OAMPDUs are just that, specific to that vendor. Non-ALU vendor specific information will not be processed.

Capability Advertising

A supported capability, sometimes requiring activation, will be advertised to the peer. The EV capability is advertisement when Link Monitoring is active on the port. This can be disabled using

the optional command **no link-monitoring** under the **config>port>ethernet>efm-oam>discovery>advertise-capabilities**.

Remote Loopback

EFM OAM provides a link-layer frame loopback mode that can be remotely controlled.

To initiate remote loopback, the local EFM OAM client sends a loopback control OAM PDU by enabling the OAM remote-loopback command. After receiving the loopback control OAM PDU, the remote OAM client puts the remote port into local loopback mode.

To exit remote loopback, the local EFM OAM client sends a loopback control OAM PDU by disabling the OAM remote-loopback command. After receiving the loopback control OAM PDU, the remote OAM client puts the port back into normal forwarding mode.

Note that during remote loopback test operation, all frames except EFM OAM PDUs are dropped at the local port for the receive direction, where remote loopback is enabled. If local loopback is enabled, then all frames except EFM OAM PDUs are dropped at the local port for both the receive and transmit directions. This behavior may result in many protocols (such as STP or LAG) resetting their state machines.

Note that when a port is in loopback mode, service mirroring will not work if the port is a mirror-source or a mirror-destination.

802.3ah OAM PDU Tunneling for Epipe Service

The 7750 SR routers support 802.3ah. Customers who subscribe to Epipe service treat the Epipe as a wire, so they demand the ability to run 802.3ah between their devices which are located at each end of the Epipe.

Note: This feature only applies to port-based Epipe SAPs because 802.3ah runs at port level not VLAN level. Hence, such ports must be configured as null encapsulated SAPs.

When OAM PDU tunneling is enabled, 802.3ah OAM PDUs received at one end of an Epipe are forwarded through the Epipe. 802.3ah can run between devices that are located at each end of the Epipe. When OAM PDU tunneling is disabled (by default), OAM PDUs are dropped or processed locally according to the **efm-oam** configuration (**shutdown** or **no shutdown**).

Note that by enabling 802.3ah for a specific port and enabling OAM PDU tunneling for the same port are mutually exclusive. Enforcement is performed on the CLI level.

802.3ah Grace Announcement

A vendor-specific Grace TLV will be included in the Information PDU generated as part of the 802.3ah OAM protocol when a network element undergoes an ISSU function. Nodes that support the Soft Rest messaging functions will allow the local node to generate the grace TLV.

The grace TLV is used to inform a remote peer that the negotiated interval and multiplier should be ignored and the new 900s timeout interval should be used to timeout the session. The peer receiving the Grace TLV must be able to parse and process the vendor specific messaging.

The new command **grace-tx-enable** has been introduced to enable this functionality. This command exists at two levels of the hierarchy, system level and port level. By default this functionality is enabled on the port. At the system level this command defaults to disabled. In order to enable this functionality both the port and the system commands must be enabled. If either is not enabled then the combination will not allow those ports to generate the vendor specific Grace TLV. This functionality must be enabled at both the system and port level prior to the ISSU or soft reset function. If this is enabled during a soft reset or after the ISSU function is already in progress it will have no affect during that window. Both Passive and Active 802.3ah OAM peers can generate the Grace TVL as part of the informational PDU.

There is no command to enable this on the receiving node. As long as the receiver understands and can parse the Grace TLV it will enter the grace mode of operation.

The basic protocol flow below helps demonstrate the interaction between passive-active and active-active peer combinations supporting the Grace TLV. In the first diagram the passive node is entering an ISSU on a node that supports soft reset capabilities.

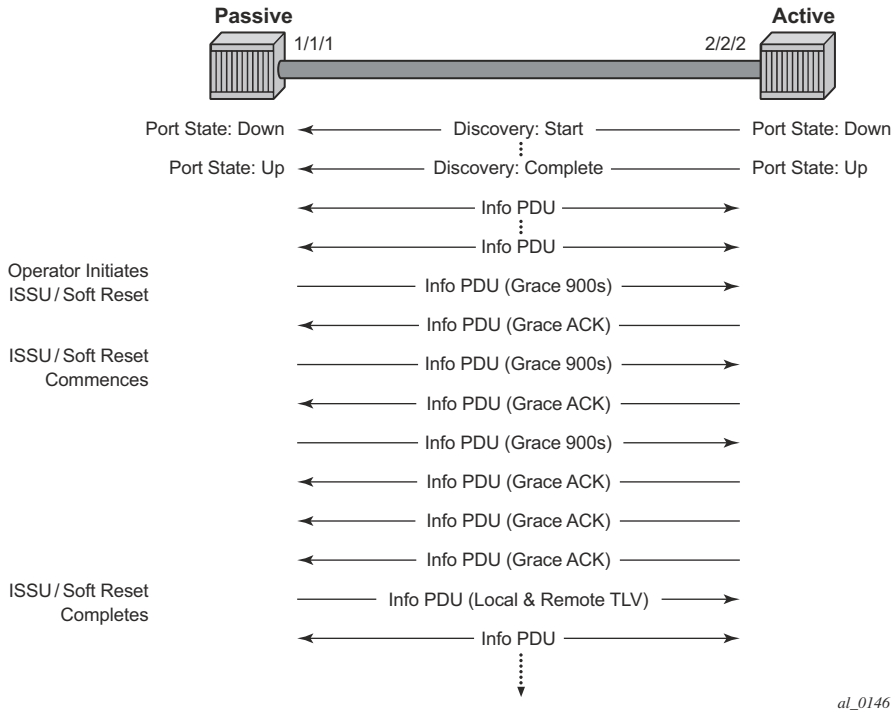

Figure 31: Grace TLV Passive Node with Soft Reset

In Figure 31 the Active node is experiencing the ISSU function on a node that supports soft reset capabilities.

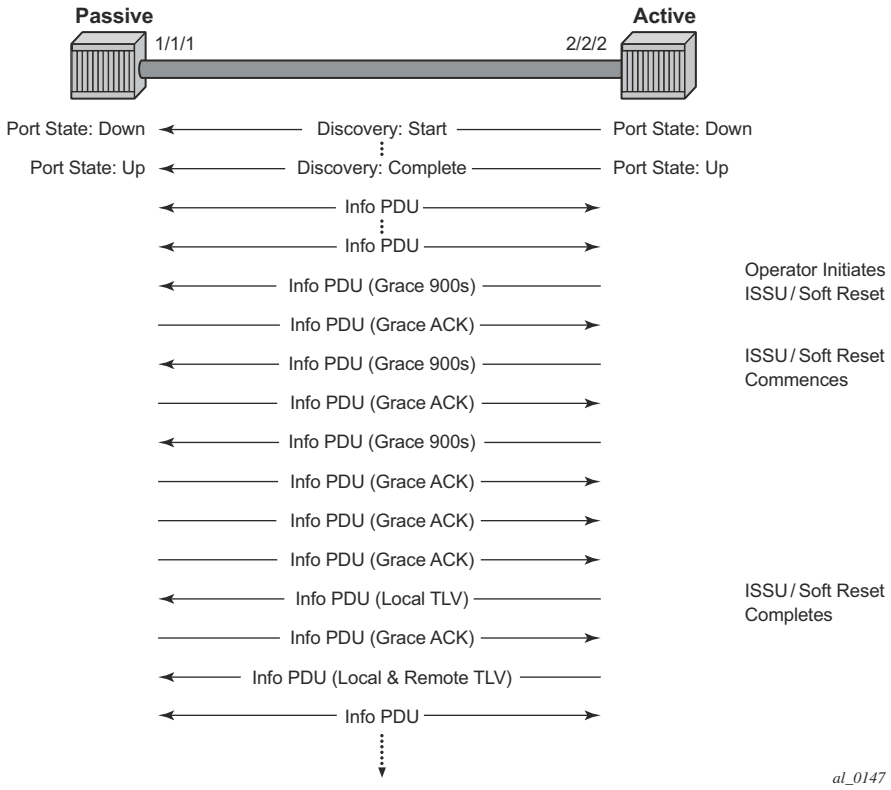

Figure 32: Grace TLV Active Node with Soft Reset

The difference between the two is subtle but important. When an active node performs this function it will generate an Informational TLV with the Local TLV following the successful soft reset. When it receives the Information PDU with the Grace Ack it will send its own Information PDU with both Local and Remote TLV completed. This will complete the protocol restart. When a passive node is reset the passive port will wait to receive the 802.3ah OAM protocol before sending its own Information PDU with both the Local and Remote TLV thus completing the protocol restart.

The renegotiation process allows the node which experienced the ISSU or soft reset to rebuild the session without having to restart the session from the discovery phase. This significantly reduces the impact of the native protocol on data forwarding.

Any situation that could cause the renegotiation to fail will force the protocol to revert to the discovery phase and fail the graceful restart. During a Major ISSU when the EFM OAM session is held operational by the Grace function, if the peer MAC address of the session changes, there will be no log event raised for the MAC address change.

This feature does not support the clearing of an IOM which does not trigger a soft reset. That is a forceful event that will not trigger this graceful protocol renegotiation.

A number of show commands have been enhanced to help operators determine the state of the802.3ah OAM Grace function and whether or not the peer is generating or receiving the Grace TLV.

The system level information can be viewed using the **show system info** command.

```
show system information
===============================================================================
System Information
===============================================================================
System Name            : ystem-name
System Type            : 7750 SR-12
System Version         : 11.0r4
System Contact         :
System Location        :
System Coordinates     :
System Active Slot     : A
System Up Time         : 62 days, 20:29:48.96 (hr:min:sec)

…snip…

EFM OAM Grace Tx Enable: False
===============================================================================
```

EFM OAM Grace Tx Enable:

False The system level functionality is not enabled. Grace will not be generated on any ports regardless of the state of the option on the individual ports

True The system level functionality is enabled and the determination of whether to send grace is base on the state of the option configured at the port level

Individual ports also contain information about the current port configuration and whether or not the Grace TLV is being sent or received.

Grace Tx Enable has two enable states with the current state in brackets to the right.

False The port level functionality is not enabled. Grace will not be generated on the port regardless of the state of the option at the system level.

True The port level functionality is enabled and the determination of whether to send grace is based on the state of the option configured at the system level

(inactive) Not currently sending Grace TLV

(active) Currently sending the Grace TLV as part of the Information PDU

Peer Grace Rx

False Not receiving Grace TLV from the peer

True Receiving Grace TLV from the peer

Port 1/2/1 is currently sending the Grace TLV and represents the node that is experiencing the ISSU function with soft reset support.

```
show port 1/2/1 ethernet efm-oam
===============================================================================
Ethernet Oam (802.3ah)
===============================================================================
Admin State        : up
Oper State         : operational
Mode               : active
Pdu Size           : 1514
Config Revision    : 0
Function Support   : LB
Transmit Interval  : 100 ms
Multiplier         : 2
Hold Time          : 0
Tunneling          : false
Loop Detected      : false
Grace Tx Enable    : true (active)

Peer Mac Address   : 00:16:4d:16:5e:40
Peer Vendor OUI    : 00:16:4d
Peer Vendor Info   : 00:00:00:00
Peer Mode          : active
Peer Pdu Size      : 1514
Peer Cfg Revision  : 0
Peer Support       : LB
Peer Grace Rx      : false

Loopback State     : None
Loopback Ignore Rx : Ignore
Ignore Efm State   : false
===============================================================================
Ethernet Oam Statistics
===============================================================================
                                                    Input                Output
-------------------------------------------------------------------------------
Information                                             0                   697
Loopback Control                                        0                     0
Unsupported Codes                                       0                     0
Frames Lost                                                                   0
===============================================================================
```

Port 3/2/1 is currently not sending the Grace TLV but is receiving the Grace TLV from its peer. This represents the peer node connected to the node that is experiencing the ISSU function with the soft reset support.

```
show port 3/2/1 ethernet efm-oam
===============================================================================
Ethernet Oam (802.3ah)
===============================================================================
Admin State        : up
Oper State         : operational
Mode               : active
Pdu Size           : 1514
Config Revision    : 0
Function Support   : LB
Transmit Interval  : 100 ms
Multiplier         : 2
Hold Time          : 0
Tunneling          : false
Loop Detected      : false
Grace Tx Enable    : true (inactive)

Peer Mac Address   : 00:16:4d:95:ea:2a
Peer Vendor OUI    : 00:16:4d
Peer Vendor Info   : 00:00:00:00
Peer Mode          : active
Peer Pdu Size      : 1514
Peer Cfg Revision  : 0
Peer Support       : LB
Peer Grace Rx      : true

Loopback State     : None
Loopback Ignore Rx : Ignore
Ignore Efm State   : false

===============================================================================
Ethernet Oam Statistics
===============================================================================
                                                    Input                Output
-------------------------------------------------------------------------------
Information                                         24488                 50984
Loopback Control                                     1784                  4859
Unsupported Codes                                       0                     0
Frames Lost                                                                   0
===============================================================================
```

MTU Configuration Guidelines

Observe the following general rules when planning your service and physical MTU configurations:

- The 7750 SR must contend with MTU limitations at many service points. The physical (access and network) port, service, and SDP MTU values must be individually defined.
- Identify the ports that will be designated as network ports intended to carry service traffic.
- MTU values should not be modified frequently.
- MTU values must conform to both of the following conditions:
  - → The service MTU must be less than or equal to the SDP path MTU.
  - → The service MTU must be less than or equal to the access port (SAP) MTU.

Default MTU Values

Table 25 displays the default MTU values which are dependent upon the (sub-) port type, mode, and encapsulation.

Table 25: MTU Default Values

Port Type	Mode	Encap Type	Default (bytes)
Ethernet	access	null	1514
Ethernet	access	dot1q	1518
Fast Ethernet	network	—	1514
Other Ethernet	network	—	9212*
SONET path or TDM channel	access	BCP-null	1518
SONET path or TDM channel	access	BCP-Dot1q	1522
SONET path or TDM channel	access	IPCP	1502
SONET path or TDM channel	network	—	9208
SONET path or TDM channel	access	frame-relay	1578
SONET path or TDM channel	access	atm	1524

*The default MTU for Ethernet ports other than Fast Ethernet is actually the lesser of 9212 and any MTU limitations imposed by hardware which is typically 16K.

Modifying MTU Defaults

MTU parameters should be modified on the service level as well as the port level.

- The service-level MTU parameters configure the service payload (Maximum Transmission Unit – MTU) in bytes for the service ID overriding the service-type default MTU.
- The port-level MTU parameters configure the maximum payload MTU size for an Ethernet port or SONET/SDH SONET path (sub-port) or TDM port/channel, or a channel that is part of a multilink bundle or LAG.

The default MTU values should be modified to ensure that packets are not dropped due to frame size limitations. The service MTU must be less than or equal to both the SAP port MTU and the SDP path MTU values. When an SDP is configured on a network port using default port MTU values, the operational path MTU can be less than the service MTU. In this case, enter the show service sdp command to check the operational state. If the operational state is down, then modify the MTU value accordingly.

Configuration Example

In order for the maximum length service frame to successfully travel from a local ingress SAP to a remote egress SAP, the MTU values configured on the local ingress SAP, the SDP (GRE or MPLS), and the egress SAP must be coordinated to accept the maximum frame size the service can forward. For example, the targeted MTU values to configure for a distributed Epipe service (ALA-A and ALA-B) are displayed in Figure 33.

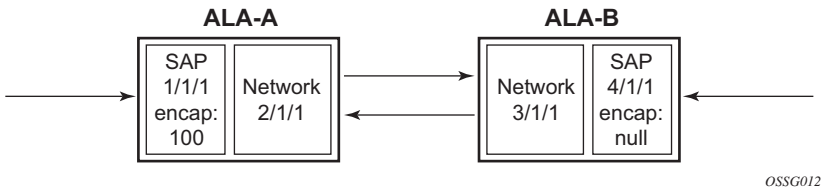

Figure 33: MTU Configuration Example

Table 26: MTU Configuration Example Values

	ALA-A		ALA-B	
	Access (SAP)	Network	Network	Access (SAP)

Table 26: MTU Configuration Example Values (Continued)

Port (slot/MDA/port)	1/1/1	2/1/1	3/1/1	4/1/1
Mode type	dot1q	network	network	null
MTU	1518	1556	1556	1514

Since ALA-A uses Dot1q encapsulation, the SAP MTU must be set to 1518 to be able to accept a 1514 byte service frame (see Table 25 for MTU default values). Each SDP MTU must be set to at least 1514 as well. If ALA-A's network port (2/1/1) is configured as an Ethernet port with a GRE SDP encapsulation type, then the MTU value of network ports 2/1/1 and 3/1/1 must *each* be at least 1556 bytes (1514 MTU + 28 GRE/Martini + 14 Ethernet). Finally, the MTU of ALA-B's SAP (access port 4/1/1) must be at least 1514, as it uses null encapsulation.

Deploying Preprovisioned Components

When a line card/CMA/MDAXCM/XMA is installed in a preprovisioned slot, the device detects discrepancies between the preprovisioned line card/CMA/MDAXCM/XMA type configurations and the types actually installed. Error messages display if there are inconsistencies and the card will not initialize.

When the proper preprovisioned line card/CMA/MDAXCM/XMA are installed into the appropriate chassis slot, alarm, status, and performance details will display.

Configuring SFM5-12e Fabric Speed

With the introduction of SFM5-12e and the mini-SFM5-12e, a new tools command (**set-fabric-speed**) was added to set the fabric operating speed. (tools command does not apply to SFM4-12e fabric-speed-a). 7750 SR-7 and 7750 SR-12 support f**abric-speed-b**.

fabric-speed-a

The 7750 SR-12e chassis defaults to the **fabric-speed-a** parameter when initially deployed with SFM5-12e. The **fabric-speed-a** parameter operates at 200GB per slot which permits a mixture of FP2/FP3 based cards to co-exist.

fabric-speed-b

The **fabric-speed-b** parameter enables the 7750 SR-12e to operate at up to 400 Gb/s, for which all cards in the 7750 SR-12e are required to be T3 based (FP3 IMM and/or IOM3-XP-C). The system will not support any FP2 based cards when the chassis is set to **fabric-speed-b**.

Configuration Process Overview

Figure 34 displays the process to provision chassis slots, line cards, MDAs, and ports.

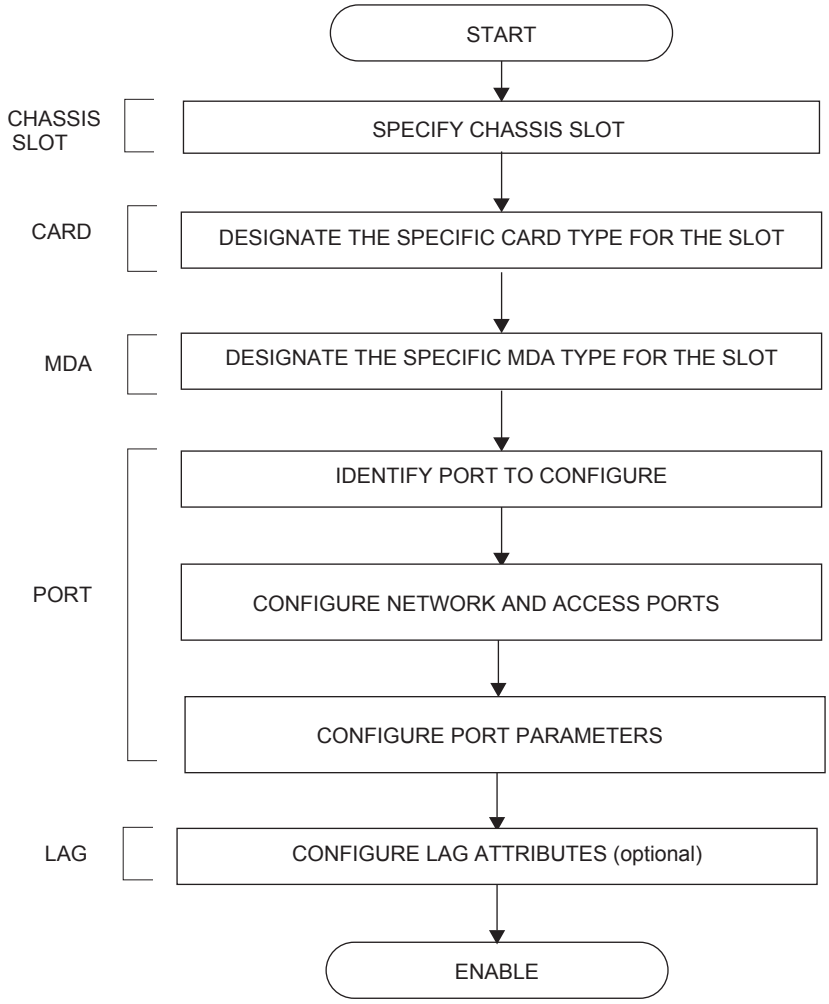

Figure 34: Slot, Card, MDA, and Port Configuration and Implementation Flow

Configuration Notes

The following information describes provisioning caveats:

- If a card or MDA type is installed in a slot provisioned for a different type, the card will not initialize.
- A card and MDA installed in an unprovisioned slot remain administratively and operationally down until the card type and MDA is specified.
- Ports cannot be provisioned until the slot, card and MDA type are specified.
- cHDLC does not support HDLC windowing features, nor other HDLC frame types such as S-frames.
- cHDLC operates in the HDLC Asynchronous Balanced Mode (ABM) of operation.
- APS configuration rules:
  - → A physical port (either working or protection) must be shutdown before it can be removed from an APS group port.
  - → For a single-chassis APS group, a working port must be added first. Then a protection port can be added or removed at any time.
  - → A protection port must be shutdown before being removed from an APS group.
  - → A path cannot be configured on a port before the port is added to an APS group.
  - → A working port cannot be removed from an APS group until the APS port path is removed.
  - → When ports are added to an APS group, all path-level configurations are available only on the APS port level and configuration on the physical member ports are blocked.
  - → For APS-protected bundles, all members of a working bundle must reside on the working port of an APS group. Similarly all members of a protecting bundle must reside on the protecting circuit of that APS group.